BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 15 2013, pages 1843–1850
doi:10.1093/bioinformatics/btt308

Genome analysis

Advance Access publication June 4, 2013

Identification of deleterious synonymous variants in human genomes

Orion J. Buske[1,*], AshokKumar Manickaraj[2], Seema Mital[2], Peter N. Ray[3,4] and Michael Brudno[1,2,5]

[1]Department of Computer Science, University of Toronto, Toronto, ON M5S 3H5, [2]Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, ON M5G 1L7, [3]Department of Paediatric Laboratory Medicine, Hospital for Sick Children, Toronto, ON M5G 1X8, [4]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8 and [5]Donnelly Centre and the Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, Canada

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: The prioritization and identification of disease-causing mutations is one of the most significant challenges in medical genomics. Currently available methods address this problem for non-synonymous single nucleotide variants (SNVs) and variation in promoters/enhancers; however, recent research has implicated synonymous (silent) exonic mutations in a number of disorders.

Results: We have curated 33 such variants from literature and developed the Silent Variant Analyzer (SilVA), a machine-learning approach to separate these from among a large set of rare polymorphisms. We evaluate SilVA's performance on *in silico* 'infection' experiments, in which we implant known disease-causing mutations into a human genome, and show that for 15 of 33 disorders, we rank the implanted mutation among the top five most deleterious ones. Furthermore, we apply the SilVA method to two additional datasets: synonymous variants associated with Meckel syndrome, and a collection of silent variants clinically observed and stratified by a molecular diagnostics laboratory, and show that SilVA is able to accurately predict the harmfulness of silent variants in these datasets.

Availability: SilVA is open source and is freely available from the project website: http://compbio.cs.toronto.edu/silva

Contact: silva-snv@cs.toronto.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on January 22, 2013; revised on April 8, 2013; accepted on May 25, 2013

1 INTRODUCTION

The realization of the medical advantages of the personal genome remains limited by our inability to identify the disease-causing variation from the millions of non-functional (neutral) single nucleotide, structural and copy number variants, which are present in each individual's genome. Despite the successes of using genome sequencing to identify disease-causing mutations in individuals with Mendelian disorders (Majewski *et al.*, 2011b), and cohorts of individuals with more common genetic disorders such as autism (O'Roak *et al.*, 2011), the prioritization of variants based on their involvement in disorders remains a significant challenge (Cooper and Shendure, 2011). Methods for the identification of disease-causing mutations typically use one of two complementary approaches: statistical association between a variant and a disorder, or the prioritization of all genomic variants found in a genome based on their possible functional effect.

In the statistical association approach, individuals with the disorder (cases) are genotyped in parallel with matched controls, and statistical tests are then used to identify variants that are overrepresented in cases as compared with controls. Although such tests have shown promise for identifying genes involved in some common disorders, including autism (Wang *et al.*, 2009) and type 2 diabetes (Frayling, 2007), these tests are not applicable to rare genetic disorders, where unrelated individuals may all be affected because of different (personal) variants within the same gene or pathway. Approaches like the Cohort Allelic Sums Test (CAST) (Morgenthaler and Thilly, 2007) and Combined Multivariate and Collapsing (CMC) method (Li and Leal, 2008) aggregate the rare variants seen within a gene or a pathway to mitigate this, but the applicability of association-based methods remains extremely limited for small cohorts.

The alternative approach of prioritizing disease-causing single nucleotide variants (SNVs) based on their population frequency, conservation and the type of change (radical versus conservative amino acid change, introduction of a stop codon, etc.) has been extremely effective at identifying causal non-synonymous mutations in a number of Mendelian disorders, including Charcot–Marie–Tooth neuropathy (Lupski *et al.*, 2010), Hajdu–Cheney syndrome (Majewski *et al.*, 2011a) and Miller syndrome (Ng *et al.*, 2009). In this approach, the variants identified in the genome are filtered to those with low allele frequencies (common variants are unlikely to cause rare disorders) and are sorted based on a 'harmfulness' score generated by tools such as PolyPhen, SIFT or PANTHER (Adzhubei *et al.*, 2010; Ng and Henikoff, 2003; Thomas *et al.*, 2003). Although some of these functional variants may not be harmful, functionality is typically used as a proxy for harmfulness within such tools, and we use the terms interchangeably. Most recently, the SNV prioritization and association-based approaches have been combined within the VAAST method (Yandell *et al.*, 2011).

*To whom correspondence should be addressed.

Tools for the prioritization of harmful non-synonymous variants typically consider multiple features that may affect the functioning of the protein, including the level of conservation of the changed residue, the severity of the amino acid change (a change from a hydrophobic to a hydrophilic residue is more likely to be harmful than a change within one of these groups), the location of the variant relative to functional regions of the protein, such as active sites, and the likelihood that the mutation would affect protein secondary or tertiary structure. These features are then combined using either heuristic weights (Ramensky *et al*., 2002) or more rigorous machine-learning frameworks (Adzhubei *et al*., 2010; Thomas *et al*., 2003) to identify SNVs likely to have functional effects. All of the features can contribute to the overall success of the prioritization; however, the conservation of the amino acid across evolution clearly has the strongest effect, and some argue it may be sufficient on its own (Cooper and Shendure, 2011).

Although tools have been developed for the prioritization of non-synonymous SNVs, and to a lesser extent copy-number variation (Hehir-Kwa *et al*., 2010), currently there are no methodologies for prioritizing functional synonymous SNVs. Most pipelines for identification of disease-causing mutations filter out synonymous SNVs at the earliest stages, concentrating on amino acid altering and regulatory variation. However, there is growing evidence that synonymous SNVs affect protein splicing, expression and ultimately function, and some of these SNVs contribute to disease (see reviews: Cartegni *et al*., 2002; Chamary *et al*., 2006; Sauna and Kimchi-Sarfaty, 2011). A synonymous SNV may contribute to a phenotype in several ways, including by changing the splicing pattern, the folding energy and the structure of the pre-mRNA and the ultimate fold of the protein by altering translation dynamics. Splice changes are perhaps the best-studied effect of functional synonymous SNVs (Cartegni *et al*., 2002). The creation or modification of a splice donor or acceptor site, or the binding site of a splicing enhancer, silencer or regulator can lead to intron inclusion or alternative splicing of the exon, and a drastically different protein product (Drögemüller *et al*., 2011). Synonymous substitutions that change a common codon to a rare one, or vice versa, can also result in a different protein by affecting translational efficiency, as is the case with a mutation in the CFTR gene associated with cystic fibrosis (Bartoszewski *et al*., 2010). Additionally, synonymous mutations have been shown to change the expression (Kudla *et al*., 2009) and function (Cortazzo *et al*., 2002; Komar *et al*., 1999) of proteins in *Escherichia coli* and play a role in substrate specificity (Kimchi-Sarfaty *et al*., 2007) and cancer outcomes (Ho *et al*., 2011) in humans, though the later claim has been controversial (Renneville *et al*., 2011).

Several previous computational approaches have looked at variation that does not alter the coding sequence, including methods that evaluate the changes in RNA folding energies and ensembles (Halvorsen *et al*., 2010; Salari *et al*., 2013; Waldispühl and Ponty, 2011) and studies that aim to identify alternative splicing genome-wide by analyzing exonic splicing enhancers and silencers (Barash *et al*., 2010a, b). However, to our knowledge, no current method combines multiple genomic features to identify ‘silent’ genetic variants with functional effects. Toward this end, we developed the Silent Variant Analyzer (SilVA), a random forest-based prioritization method for synonymous variants in the human genome. Our method considers multiple features based on sequence conservation, splice factor motifs, splice donor/acceptor sites, RNA folding energy, codon usage and CpG content. We use a custom-curated dataset of 33 rare synonymous disease-causing variants to train and evaluate the overall efficacy of SilVA, as well as two additional datasets for independent validation, showing that SilVA is able to accurately predict the harmfulness of silent variants in these datasets.

2 METHODS

2.1 Datasets

One of the challenges in investigating synonymous disease-causing variants is the relatively small number of known examples. We have curated from literature a dataset of 33 rare (allele frequency <5%) synonymous variants according to strict criteria: they must have been implicated in a disorder and experimentally validated to affect splicing, transcript abundance, mRNA stability or translational efficiency (Supplementary Table S1). For training and benchmarking negative controls, we used all rare synonymous variants from an individual in the 1000 Genomes Project (NA10851) (Durbin *et al*., 2010). We identified 758 variants with minor allele frequencies (MAF) <5%. For case studies and validation, we trained SilVA on the NA10851 variants, but used the 746 variants in another 1000 Genomes Project individual (NA07048) during testing. Fifty-nine variants were shared by both NA10851 and NA07048.

After developing and benchmarking the SilVA method, we obtained two further validation datasets (Table 1). The first contained seven synonymous variants found in Meckel syndrome families (Khaddour *et al*., 2007). Four of these variants were reported to be novel, of which two (MKS1: E139E, TMEM67: A813A) were suspected to cause a Meckel syndrome phenotype. The other three variants were predicted to be benign polymorphisms, with minor allele frequencies of 1–7%.

We also obtained a dataset of 12 synonymous mutations encountered by the Molecular Diagnostic Laboratory at the Hospital for Sick Children (HSC; Toronto, Canada). Of these 12 variants, six were determined to be pathogenic by a molecular diagnostician, whereas the remaining six were believed to be benign polymorphisms. Of the six pathogenic ones, two were already in our training data, whereas the other four were novel.

2.2 Features

We annotate each variant with 26 features across six categories: (i) conservation, (ii) codon usage, (iii) sequence features (CpG and relative mRNA position), (iv) exon splicing enhancer and suppressor (ESE/ESS) motifs, (v) splice site motifs and (vi) pre-mRNA folding energy (Table 2).

The GERP++ score is used to measure the evolutionary conservation at the mutation position (Davydov *et al*., 2010). Relative synonymous codon usage (RSCU) (Sharp and Li, 1987) features are calculated using codon frequencies in the Codon Usage Database (Nakamura *et al*., 2000). Splicing regulatory features include the SR-protein motifs for SF2/ASF, SC35, SRp40 and SRp50, scored using ESE Finder 3.0 with default thresholds (Smith *et al*., 2006), the FAS-hex3 hexamer dataset from FAS-ESS, used for the ESS6 features (Wang *et al*., 2004) and PESX enhancer and suppressor octamers, used for the pESE and pESS features (Zhang *et al*., 2005). The splice site motif strength features (MES) are calculated using MaxEntScan (Eng *et al*., 2004). The change in free energy from pre-mRNA folding ($\Delta\Delta G$) features are calculated with UNAFold 3.8 (Markham and Zuker, 2008), and the ensemble diversity (ΔD) features are calculated with ViennaRNA 2.1.1 (Lorenz *et al*., 2011).

Table 1. Independent validation dataset consisting of seven variants (two putatively pathogenic, five putatively benign) associated with Meckel syndrome and twelve variants (six pathogenic, affecting splicing, and six polymorphic) encountered by the Molecular Diagnostic Laboratory at the HSC (Toronto, Canada)

Gene	Mutation	Rank	Score	Description [MAF]
Meckel syndrome				
TMEM67	A813A,G > A	1	0.737	novel, putatively pathogenic
MKS1	E139E,G > A	1	0.705	novel, putatively pathogenic
MKS1	L557L,G > C	277.5	0.020	polymorphic [0.06]
TMEM67	D799D,T > C	311	0.015	polymorphic [0.01]
TMEM67	C62C,T > C	356	0.011	novel, putatively benign
TMEM67	T964T,A > C	447.5	0.006	polymorphic [0.07]
TMEM67	A984A,A > G	722	0.000	novel, putatively benign
Molecular Diagnostics Laboratory at the HSC				
TP53	T125T,G > A	1	0.795	pathogenic, in training data
ACVRL1	P459P,G > C	1	0.794	pathogenic
FGFR2	A344A,G > A	1	0.762	pathogenic, in training data
CFTR	E528E,G > A	1	0.524	pathogenic, exon skipped
PKP2	G828G,C > T	29	0.153	pathogenic, cryptic splicing
IDS	G374G,C > T	73	0.083	pathogenic, cryptic splicing
TP53	L257L,C > T	106	0.065	polymorphic, novel
FGFR2	V232V,A > G	169	0.042	polymorphic [0.18]
CFTR	T854T,T > G	329.5	0.014	polymorphic [0.44]
CDKN1C	E236E,G > A	435	0.006	polymorphic [0.02]
IDS	T146T,C > T	501.5	0.004	polymorphic [0.24]
TP53	P36P,G > A	638.5	0.001	polymorphic [0.01]

Note: Of the eight pathogenic variants, two (those in TP53 and FGFR2) were already included in our training data. We used SilVA to rank each variant relative to all (746) rare putatively neutral synonymous variants in a 1000 Genomes Project individual not used during model development or training (NA07048). The SilVA method ranked all pathogenic variants higher than all polymorphic variants. Moreover, we ranked 4/6 new pathogenic and putatively pathogenic variants as more harmful than any control variant (a rank of 1). For all listed mutations, the third codon position is affected.

Before training, we preprocess each of the features to have zero mean and unit variance.

Table 2. Synonymous variants were annotated with a diverse set of 26 features spanning six distinct categories of information relevant to assessing the harmfulness of SNVs

Feature	Description
Conservation	
GERP++	Conservation at the mutation position
Codon usage bias	
RSCU	RSCU of new codon
$\lvert\Delta\text{RSCU}\rvert$	Change in RSCU caused by mutation
Sequence features	
CpG?	Does the mutation change a CpG?
CpG_{exon}	Observed/expected CpG content of exon
f_{pre}	Relative distance to end of pre-mRNA
f_{post}	Relative distance to end of mature mRNA
Exon splice enhancer/suppressor motifs	
SR−	SR-protein motifs lost
SR+	SR-protein motifs gained
FAS6−	Hexamer splice suppressor motifs lost
FAS6+	Hexamer splice suppressor motifs gained
PESE−	Octamer splice enhancer motifs lost
PESE+	Octamer splice enhancer motifs gained
PESS−	Octamer splice suppressor motifs lost
PESS+	Octamer splice suppressor motifs gained
Splice site motifs	
MES	Max splice site score
$\lvert\Delta\text{MES}\rvert$	Max change in splice site score
ΔMES+	Max splice site score increase
ΔMES−	Max splice site score decrease
MES-MC?	Did strongest site change?
MES-CS?	Is a cryptic site now strongest?
MES-KM?	Did a known site change most?
Pre-mRNA folding free energy	
$\Delta\Delta G_{\text{pre, 50}}$	Folding energy change, pre-mRNA, 50 bp window
$\Delta\Delta G_{\text{post, 50}}$	Folding energy change, mature mRNA, 50 bp window
$\Delta D_{\text{pre, 50}}$	Ensemble diversity change, pre-mRNA, 50 bp window
$\Delta D_{\text{post, 50}}$	Ensemble diversity change, mature mRNA, 50 bp window

2.3 Selection of the random forest model

We compared the ability of five different methods, the GERP++ score and four machine-learning models, to identify the deleterious synonymous variants. These methods are:

(1) Sort by GERP++ conservation score. Mutations at more conserved residues are ranked higher.

(2) Fisher's linear discriminant (FLD). The variants are ranked by the one-dimensional projected value.

(3) Support vector machine (SVMmap), using the nu-SVR regression mode of the lib-svm toolkit, version 3.11 (Chang and Lin, 2011). We then sort variants by the regression score.

(4) Neural network (NNet), with a single fully connected hidden layer of five hidden units, using the PyBrain Python package, version 0.3 (Schaul *et al.*, 2010). We activate the trained network on the test SNVs and use the value at the output node to prioritize them.

(5) Random forest (Forest), with 1001 trees and the default number of variables used for each split (the square root of the total number of variables), using the randomForest R package, version 4.6–6. Variants are ranked by the number of votes, with the most popular variants ranked highest.

For all of these methods, we trained predictive models using both 50/50 splits and leave-one-out cross-validation. For 50/50 splits, we trained each model on half of the positive and negative examples (~17 known deleterious, 379 presumed benign or control), and then ranked the remaining (16 known deleterious and 379 control variants). We excluded from training any positive examples that occurred within the same gene as any of the positive test mutations. Each method was then evaluated according to the quality of the topmost predictions. We aggregated the results across 50 iterations of training and testing, each time with a new random subset of deleterious and control variants. As shown in Figure 1A and Supplementary Figure S1, the random forest method outperforms the other methods and has more than three times the true positive rate (at a false positive cut-off of 1%) as simply using the GERP++ score.

To compare the prioritization performance of the five methods in a more realistic scenario, we performed *in silico* 'infection' experiments (leave-one-out cross-validation). In each experiment, we held out one of the 33 deleterious variants and half of the control variants, and then used each model to rank the held-out variant against the set of control variants. As in the 50/50 split, we excluded from training any positive examples that occurred within the same gene as the held-out variant. We repeated this process 10 times with different random subsets of

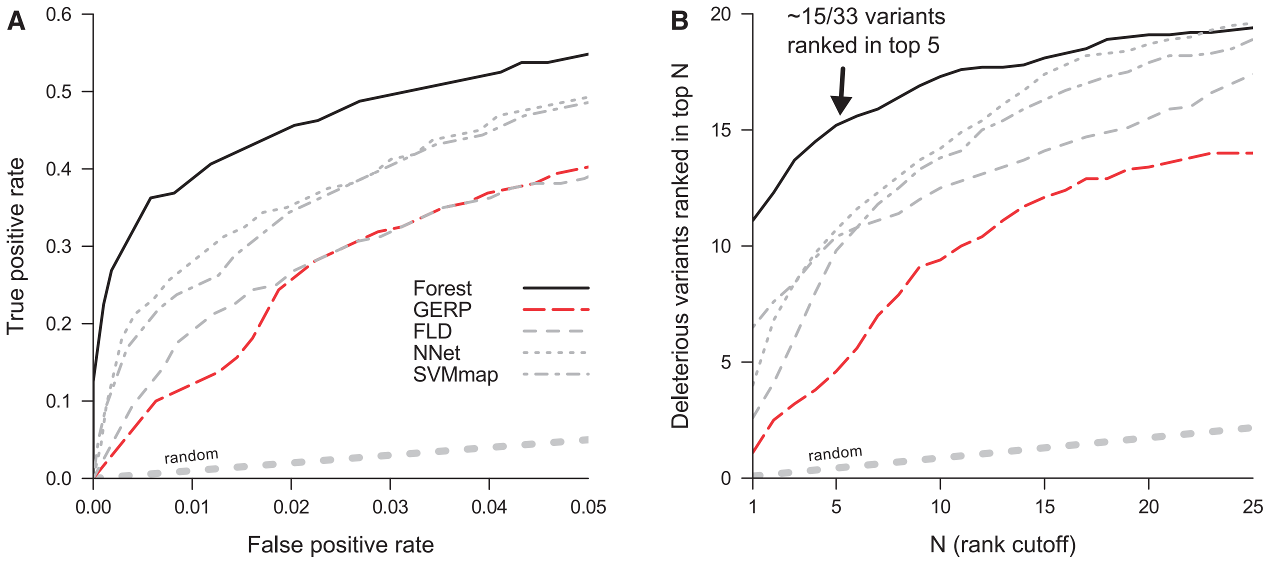

Fig. 1. Plots comparing the performance of several machine-learning methods (Forest: random forest; SVMmap: support vector machine; FLD: Fisher's linear discriminant; NNet: neural network) and the GERP++ score at classifying harmful synonymous variants. (**A**) This plot is the bottom-left region on a receiver operating characteristic (ROC) curve. Curves were averaged over 50 training iterations, with half of the positive and negative examples used for testing. The performance of a random ordering appears at the bottom. Random forest is able to rank more of the held-out positive examples highly than any of the other methods at low false-positive thresholds. (**B**) Performance on simulated 'infection' (leave-one-out cross-validation) experiments using 33 deleterious variants and 758 rare (putatively neutral) variants from 1000 Genomes Project individual NA10851. If we consider the causal variant to have been found if it is ranked in the top five, random forest succeeds on an average of 15.2 deleterious variants (versus 10.7 for NNet, 10.4 for SVMmap, 9.8 for FLD and 4.6 for GERP++)

control variants and averaged the results. The prioritization performance of the five methods is compared in Figure 1B and Supplementary Figure S1. The random forest method achieved the best performance, ranking the deleterious variant in the top five most harmful variants for ~15 of 33 deleterious variants (compared with ~5 for the GERP++ score by itself).

3 RESULTS

3.1 SilVA

To enable the automated prioritization of harmful synonymous variants for medical sequencing projects, we developed SilVA. For projects in which candidate non-synonymous variants cannot be found, we offer SilVA as an effective method for prioritizing the large set of synonymous variants that might normally be ignored. SilVA takes a list of variants in VCF format and orders them by a computed harmfulness score. We then expect a geneticist to evaluate the top several candidates based on a review of the literature and potential functional effects. We designed SilVA with this approach in mind, and SilVA is able to rank harmful synonymous variants within the top five genome-wide > 45% of the time.

SilVA is implemented using a combination of Python, Bash and R and is freely available from http://compbio.cs.toronto.edu/silva.

3.2 The SilVA score and classifier

To score the potential harmfulness of the variants, SilVA first annotates each variant with 26 features organized into six categories (Fig. 2A and Methods). SilVA then scores variants using a random forest model trained on 33 synonymous harmful mutations that we have identified from the literature (see Methods). The SilVA score corresponds to the fraction of trees in the random forest that predict the mutation to be harmful and is significantly higher in harmful synonymous variants than in control variants, even for just variants near splice sites.

To evaluate the ability of the SilVA to differentiate between harmful and benign variants, we measured the mean SilVA scores for our harmful mutation dataset and common polymorphisms (MAF >5%), which are unlikely to be harmful. We computed the scores of the 33 harmful variants using leave-one-out cross-validation and the scores of all common polymorphisms from the 1000 Genomes Project (May 2011, phase 1, release v2) and found the harmful variants to have a significantly higher mean score (0.322 versus 0.031, Student's *t*-test: $P \leq 1.8 \times 10^{-7}$). Further, we still achieve significance when comparing against rare synonymous variants from a healthy individual (0.322 versus 0.031, $P \leq 1.9 \times 10^{-7}$, 1000 Genomes Project individual NA07048), and even when we focus on just variants within three residues of a splice site (0.544 versus 0.153, $P \leq 3.8 \times 10^{-6}$). Thus, the SilVA score is an effective tool for prioritizing synonymous variants.

SilVA classifies variants as likely benign, potentially pathogenic or likely pathogenic based on their score to aid interpretation. These score thresholds (of 0.27 and 0.485) correspond to true positive rates of 52 and 33%, and false positive rates of <1% and 0.1%, respectively, when ranking all common polymorphisms from the 1000 Genomes Project. Because we expect harmful synonymous variants to be extremely rare, we do not intend SilVA to be used in the same way as typical non-synonymous

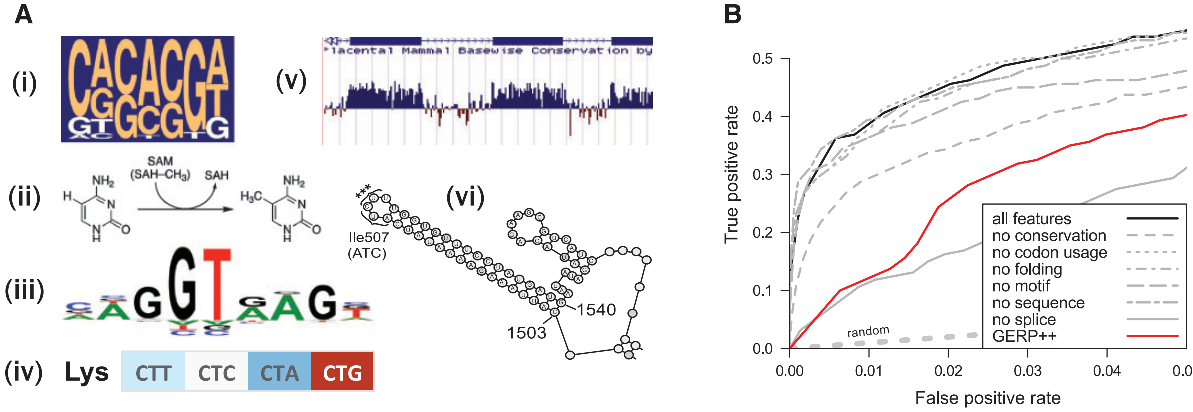

Fig. 2. (**A**) Illustrations of the various feature categories used within SilVA to predict the harmfulness of synonymous variants: (i) exon splicing enhancer/suppressor motifs (SF2/ASF shown), (ii) sequence features (CpG and relative position in the mRNA), (iii) splice site motifs, (iv) codon usage bias, (v) conservation and (vi) RNA folding (image adapted from Bartoszewski *et al*., 2010). (**B**) The bottom-left region of an ROC curve comparing the performance of the SilVA method with groups of features removed. We aggregated the results across 50 iterations of training and testing, each time with a random half of the positive and negative examples used for training and the remainder for testing. For comparison, we also show the performances of sorting by the GERP++ score and a random ordering

harmfulness prediction tools and thus focus on ranking variants instead of classifying them, though we also report classification results for each dataset.

3.3 Feature performance

To better understand the relative contributions of the features used within SilVA and to explore the relative importance of each category of feature, we compared SilVA's cross-validation performance leaving out different classes of features from the analysis (Fig. 2B). Removing features related to codon usage, mRNA folding, splicing enhancer and suppressor motifs and sequence (CpG, relative position in mRNA) does not substantially affect performance. Removing either splice site features or conservation (GERP++), however, causes SilVA's performance to drop substantially, with splice site features appearing to be more informative than conservation for harmfulness prediction.

3.4 Disease SNV identification

To further assess the performance of the SilVA method, we performed *in silico* 'infection' experiments, where we add a known deleterious variant to half of the variants in a human genome (1000 Genomes Project individual, NA10851) and train SilVA on the remaining variants. This leave-one-out cross-validation method allows us to estimate the number of disorders for which our method is able to rank the deleterious variant among the top few variants genome-wide. As shown in Figure 3, the SilVA method is able to consistently rank 14–17 of the 33 deleterious variants within the top five variants (46% on average). This suggests that in a large fraction of cases, SilVA is able to effectively prioritize harmful synonymous variation in human genomes. Of the 33 deleterious variants, 11 were classified as likely pathogenic, 6 as potentially pathogenic and 16 as likely benign. For comparison, of the rare synonymous variants across 82 CEU 1000 Genomes Project individuals, an average of <1 variant per genome was classified as likely pathogenic, 7 as potentially pathogenic and 727 as likely benign. Variants that were mistakenly classified as benign tended to be far from splice sites and affect protein production by disrupting ESE/ESS motifs or translational dynamics. Though we have features that attempt to capture these mechanisms, the machine-learning algorithms did not find these specific features to be informative.

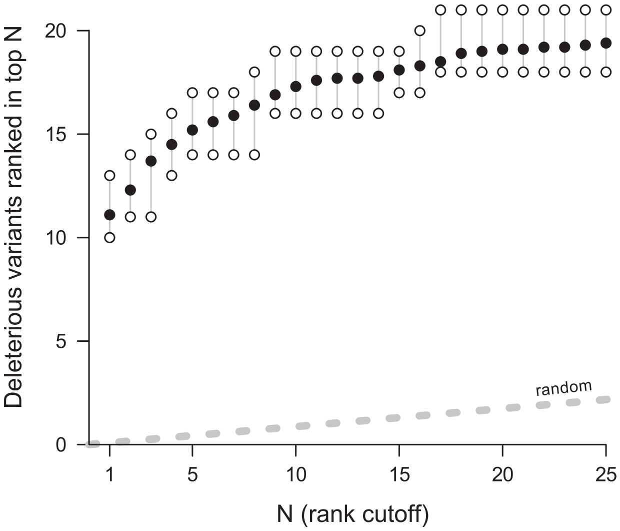

Fig. 3. Performance of the SilVA method on 'infection' (leave-one-out cross-validation) simulations. In each simulation, half (379) of the negative examples are held out, along with one positive example. This is repeated 10 times for each of the 33 positive examples. Plotted is the minimum, maximum and mean number of held-out positive examples that are ranked within the top N variants

3.5 Validation on independent SNV sets

In addition to the cross-validation experiments described above, we used two smaller independent datasets to validate the performance of SilVA.

3.5.1 Meckel syndrome variants First, we used SilVA to predict the harmfulness of a collection of synonymous SNVs reported by Khaddour *et al.* (2007) across many cases of Meckel syndrome, a rare fatal developmental disorder of the nervous system, kidney, liver and lungs. Khaddour *et al.* describe seven synonymous mutations in the MKS1 and TMEM67 (MKS3) genes, of which four are novel and three are known polymorphisms (minor allele frequencies of 1–7%). Two of the novel mutations are suspected of causing Meckel syndrome through the disruption of splice donor motifs. These variants were not included in our training dataset because they did not meet our criterion of experimental validation.

As controls, we used all (746) rare synonymous variants in a 1000 Genomes Project individual not used for training or benchmarking (NA07048). In agreement with the literature, SilVA ranks the two suspected harmful mutations (MKS1:E139E, G > A; TMEM67:A813A, G > A) higher than every control variant (a rank of 1) and none of the remaining five mutations within even the top 250 variants.

3.5.2 Variants from HSC's Molecular Diagnostics Laboratory The Molecular Diagnostics Laboratory at the HSC conducts Sanger sequencing for gene panels in patients with suspected genetic disorders. Each variant is analyzed by a molecular diagnostician, who classifies it as benign or harmful based on an interpretation of its likely molecular effect and a literature review. The Molecular Diagnostic Laboratory provided us with six pathogenic synonymous variants and six benign polymorphisms identified during their analyses, with two of the pathogenic variants already appearing in our training data. Similar to our analysis of the Meckel variants, we implanted the 10 remaining variants in a 1000 Genomes individual (NA07048). As shown in Table 1, SilVA ranks all pathogenic variants higher than all polymorphic variants, with two of the four new pathogenic variants (and both of the ones in the training data) ranking higher than any control variants (a rank of 1).

3.6 Genome-wide comparison of polymorphisms and random synonymous substitutions

The rate of synonymous substitutions is widely used as a proxy for the neutral mutation rate, including for the purposes of identifying selection on a gene (*e.g.* McDonald and Kreitman, 1991). However, a number of mechanisms have been studied, including those discussed in this article, through which synonymous substitutions can exert a phenotypic effect and thus be selected against. Previously, there have been several attempts to understand the fraction of synonymous sites that are under constraint, and the strength of selection at these sites both overall (see review: Chamary *et al.*, 2006) and at specific locations such as exon splicing enhancers (Parmley *et al.*, 2006). The heterogeneity of both the genome and even individual genes, and substantial methodological differences, have resulted in widely varying estimates, with some suggesting that up to 39% of synonymous substitutions are under selection (Hellmann *et al.*, 2003). Simultaneously, all such studies have used comparison of multiple mammalian genomes, and not analysis of human polymorphisms; constraint observable from human polymorphisms would represent generally stronger selection, due to the small human effective population size (Ne ≈104) (Tenesa *et al.*, 2007).

Recently, Salari *et al.* (2013) compared the effects of common human polymorphisms and random mutations on RNA structural ensembles, and found significant evidence of ensemble-stabilizing selection. If a significant fraction of human synonymous sites are under constraint and the SilVA score reflects this, we would expect to see a difference in SilVA scores between common synonymous polymorphisms and a matched set of random mutations. We test this hypothesis by applying SilVA to each synonymous SNP in NA10851 (9596 variants with allele frequencies of 5–95%) and a matched random synonymous site within the same gene. The matched set of random mutations was controlled for creation or destruction of CpG dinucleotides and splice site proximity (the random mutation created/destroyed a CpG site only if the synonymous variant did, and whether or not the mutation was within three bases of an exon boundary). We then compared the distribution of SilVA scores for the two datasets. Although the mean observed scores for polymorphisms and random mutations were similar (0.031 and 0.034, respectively), the difference in the means is highly statistically significant due to the large number of datapoints (Student's *t*-test, paired, $P \leq 2.4 \times 10^{-6}$). The overall higher scores of random mutations suggest that factors beyond CpG and exon boundaries impose purifying selection at synonymous sites of the human genome that is statistically significant.

To further quantify this constraint, we measured the difference in the number of random mutations and true polymorphisms (Fig. 4) above a certain SilVA score. This difference can be interpreted as the number of mutations 'rejected' during evolution as being unfit, and represents synonymous sites under constraint (Cooper *et al.*, 2005). At a SilVA threshold of 0.005, we observe 626 more random mutations (6531) than true polymorphisms

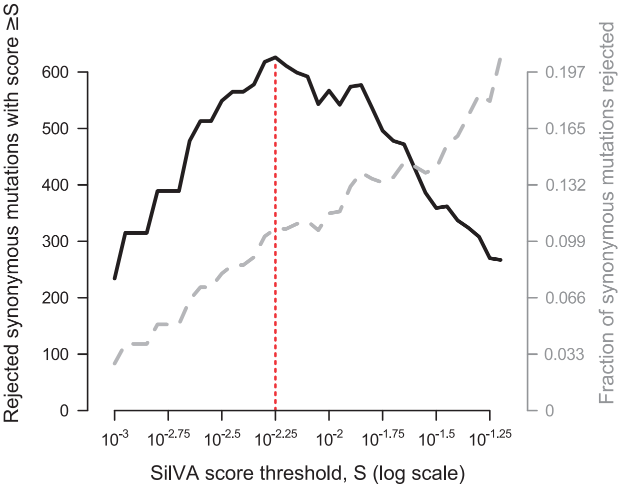

Fig. 4. The number and fraction of synonymous mutations that are rejected at various SilVA score thresholds. The largest number of rejected mutations occurs at a SilVA threshold of $10^{-2.25}$, marked with a dashed vertical line, where 626 more random mutations pass the threshold than actual polymorphisms, corresponding to a 10.6% rejection rate at this threshold

(5905). Thus, we estimate that 6.5% of potential synonymous substitutions (626/9596 SNPs) have been rejected since human divergence due to constraint beyond just CpGs and splice sites. Note that this is a conservative estimate, as there are likely additional functional features in the genome that the SilVA score is not modeling.

4 DISCUSSION

Current technologies are able to sequence a human genome relatively cheaply and quickly, but the key bottleneck is the interpretation of the variants to identify those that are most likely to be related to an observed phenotype or a disorder. The automated prioritization of deleterious variants is an important step toward the realization of genomic medicine. Although methods for this task have been designed for non-synonymous coding variation, many other types of variants are currently not prioritized, and are typically ignored in genome analysis pipelines. The growing evidence that some of the 'silent' variation has important functional roles implies that effective prioritization is necessary to explore such variation in medical sequencing studies.

In this article we present the first method for the prioritization of disease-causing synonymous SNVs based on a number of features, including sequence conservation, splice sites, splice-regulatory motifs, codon frequency, CpG content and RNA secondary structure energy. We have curated 33 such disease-related variants, and evaluated several machine-learning approaches for prioritizing these from among a set of rare putatively neutral SNVs. Our results indicate that splicing information and sequence conservation are currently the two most informative features for identifying deleterious synonymous variants, and the performance degrades without either of these features. The random forest method outperforms other statistical learning methods at prioritizing disease-causing SNVs, and yields variant scores that are significantly higher in known harmful variants than in control variants. When a deleterious SNV is added to a human genome, this method ranks the deleterious SNV among the top five candidates for 15 of the 33 diseases and is able to identify harmful variants in independent validation sets. Together, these findings indicate that automated methods for identification of deleterious synonymous SNVs can be useful in parallel with methods that prioritize other types of genomic variation for the analysis of full human genomes.

Our method's performance is currently limited by the small number of training examples, but as more examples of pathogenic synonymous variants are found, the model predictions will only improve. Additional features can also be developed to better address the mechanisms by which synonymous mutations can effect change, such as better measures of exon strength, the effect on RNA folding ensembles and the significance of changes to exon splicing enhancer and suppressor motifs.

As approaches for the stratification of different types of variants are developed, one natural extension of the current article and the work on non-synonymous variation would be to design an approach that would consider both types of variants, and attempt to classify deleterious synonymous and non-synonymous variants together.

ACKNOWLEDGEMENT

We would like to thank Aziz Mezlini and Anna Goldenberg for their assistance with the manuscript.

Funding: National Science and Engineering Research Council of Canada (NSERC, to M.B.) and Ontario Graduate Scholarship (OGS, to O.J.B.). M.B. is an Alfred P. Sloan fellow.

Conflict of Interest: none declared.

REFERENCES

Adzhubei,I. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Barash,Y. *et al.* (2010a) Deciphering the splicing code. *Nature*, **465**, 53–59.

Barash,Y. *et al.* (2010b) Model-based detection of alternative splicing signals. *Bioinformatics*, **26**, i325–i333.

Bartoszewski,R. *et al.* (2010) A synonymous single nucleotide polymorphism in δF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J. Biol. Chem.*, **285**, 28741–28748.

Cartegni,L. *et al.* (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.

Chamary,J. *et al.* (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98–108.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27:27.

Cooper,G. and Shendure,J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–640.

Cooper,G. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

Cortazzo,P. *et al.* (2002) Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, **293**, 537–541.

Davydov,E.V. *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.

Drögemüller,C. *et al.* (2011) An unusual splice defect in the mitofusin 2 gene (MFN2) is associated with degenerative axonopathy in Tyrolean grey cattle. *PloS One*, **6**, e18931.

Durbin,R. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Eng,L. *et al.* (2004) Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. *Hum. Mutat.*, **23**, 67–76.

Frayling,T. (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet.*, **8**, 657–662.

Halvorsen,M. *et al.* (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.*, **6**, e1001074.

Hehir-Kwa,J. *et al.* (2010) Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS Comput. Biol.*, **6**, e1000752.

Hellmann,I. *et al.* (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.*, **13**, 831–837.

Ho,P. *et al.* (2011) WT1 synonymous single nucleotide polymorphism rs16754 correlates with higher mRNA expression and predicts significantly improved outcome in favorable-risk pediatric acute myeloid leukemia: a report from the Children's Oncology Group. *J. Clin. Oncol.*, **29**, 704.

Khaddour,R. *et al.* (2007) Spectrum of MKS1 and MKS3 mutations in Meckel syndrome: a genotype-phenotype correlation. *Hum. Mutat.*, **28**, 523–524.

Kimchi-Sarfaty,C. *et al.* (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, **315**, 525–528.

Komar,A. *et al.* (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.*, **462**, 387–391.

Kudla,G. *et al.* (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–258.

Li,B. and Leal,S. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Lupski,J. *et al.* (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.*, **362**, 1181–1191.

Majewski,J. *et al.* (2011a) Mutations in NOTCH2 in families with Hajdu-Cheney syndrome. *Hum. Mutat*., **32**, 1114–1117.

Majewski,J. *et al.* (2011b) What can exome sequencing do for you? *J. Med. Genet*., **48**, 580–589.

Markham,N.R. and Zuker,M. (2008) UNAFold: Software for nucleic acid folding and hybridization. *Methods in Molecular Biology*, **453**, 3–31.

McDonald,J. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.

Morgenthaler,S. and Thilly,W. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res./Fundam. Mol. Mech. Mutagen*., **615**, 28–56.

Nakamura,Y. *et al.* (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*., **28**, 292.

Ng,P. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.

Ng,S. *et al.* (2009) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet*., **42**, 30–35.

O'Roak,B. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet*., **43**, 585–589.

Parmley,J. *et al.* (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol*., **23**, 301–309.

Ramensky,V. *et al.* (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.

Renneville,A. *et al.* (2011) Wilms' tumor 1 single-nucleotide polymorphism rs16754 does not predict clinical outcome in adult acute myeloid leukemia. *Leukemia*, **25**, 1918–1921.

Salari,R. *et al.* (2013) Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. *Nucleic Acids Res*., **41**, 44–53.

Sauna,Z. and Kimchi-Sarfaty,C. (2011) Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet*., **12**, 683–691.

Schaul,T. *et al.* (2010) PyBrain. *J. Mach. Learn. Res*., **11**, 743–746.

Sharp,P.M. and Li,W.H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*., **15**, 1281–1295.

Smith,P.J. *et al.* (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet*., **15**, 2490–2508.

Tenesa,A. *et al.* (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res*., **17**, 520–526.

Thomas,P. *et al.* (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*., **13**, 2129–2141.

Waldispühl,J. and Ponty,Y. (2011) An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *Journal of Computational Biology*, **18**, 1465–1479.

Wang,K. *et al.* (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, **459**, 528–533.

Wang,Z. *et al.* (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.

Yandell,M. *et al.* (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res*., **21**, 1529–1542.

Zhang,X.H. *et al.* (2005) Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol*., **25**, 7323–7332.