BIOINFORMATICS REVIEW Vol. 27 no. 4 2011, pages 441–448
doi:10.1093/bioinformatics/btq695

Genome analysis Advance Access publication December 15, 2010

Using bioinformatics to predict the functional impact of SNVs

Melissa S. Cline[1] and Rachel Karchin[2,*]

[1]Department of Molecular Cell and Developmental Biology and Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA and [2]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The past decade has seen the introduction of fast and relatively inexpensive methods to detect genetic variation across the genome and exponential growth in the number of known single nucleotide variants (SNVs). There is increasing interest in bioinformatics approaches to identify variants that are functionally important from millions of candidate variants. Here, we describe the essential components of bionformatics tools that predict functional SNVs.

Results: Bioinformatics tools have great potential to identify functional SNVs, but the black box nature of many tools can be a pitfall for researchers. Understanding the underlying methods, assumptions and biases of these tools is essential to their intelligent application.

Contact: karchin@jhu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 20, 2010 ; revised on November 21, 2010; accepted on December 12, 2010

1 INTRODUCTION

The most common form of intra-species variation is the single nucleotide variant (SNV) (Durbin *et al.*, 2010). Recent sequencing of whole human genomes has identified the number of SNVs in each individual to be in the range of 3–5 million (Ahn *et al.*, 2009; Levy *et al.*, 2007; Schuster *et al.*, 2010; Wang *et al.*, 2008; Wheeler *et al.*, 2008). Common SNVs have recently been a focus of population genetics and medical research. With the advent of high-density SNV microarrays, researchers can do fast genome-wide genotyping of study populations. This technology has been used to trace the ancestry of modern humans and migration patterns of ancient humans (Gutenkunst *et al.*, 2009; Li *et al.*, 2008). SNVs are now the most commonly used markers in case–control association studies. Genome-wide association studies (GWAS) have identified over 3300 common SNVs with statistically significant correlations to disease phenotypes (http://www.genome.gov/gwastudies/ October 14, 2010). However, these correlations do not necessarily imply causality, and researchers have found that most significant SNVs identified by GWAS do not contribute biologically to the disease phenotype of interest (Cantor *et al.*, 2010). While most SNVs are likely to be functionally neutral (McClellan and King, 2010), some SNVs have a functional impact and thus directly contribute to disease susceptibilities and drug sensitivities. Discovering these *functional SNVs* is one of the main goals of modern genetics and genomics studies.

From a biological perspective, there are numerous ways that a SNV can have a functional impact. A SNV may impact the transcriptional machinery of a cell if it occurs in a region of DNA that contains signals recognized by transcription factors and/or transactivational enhancers. If it occurs at a splice site or a site where exonic or intronic splicing enhancers or repressors bind, it may result in alternative or aberrant splice isoforms of a transcribed gene. It can interfere with the cell's translational machinery. SNVs that result in altered protein sequence may interfere with protein folding, localization, stability, binding or catalysis.

Bioinformatics tools to identify important SNVs must consider these possible impacts indirectly, through 'proxies'. A bench biologist interested in whether a SNV of interest impacts the transcription of a gene might perform site-directed mutagenesis on genomic DNA, transfect mutated DNA into cell culture and use readouts of the gene's transcriptional activity to measure changes with respect to wild type. In contrast, a bioinformatics-based approach typically involves computational analysis of the DNA sequence surrounding the SNV, possibly supplemented with information from published bench experiments. Does the DNA contain motifs associated with binding of transcriptipon factors? Can it be aligned to equivalent DNA sequence in other species? Does statistical analysis of the alignment suggest that signatures of evolutionary selective pressure are present in the region?

Most bioinformatics work on SNV function prediction has focused on SNVs in protein-coding exons that change an amino acid residue in the gene's protein product. It has been argued that these SNVs (known as cSNVs or nsSNVs) have the most relevance to human health (Botstein and Risch, 2003), but growing knowledge of the functional importance of non-coding DNA and RNA-coding genes has brought this assumption into question (Altshuler *et al.*, 2008). Most SNVs found to be differentially distributed in cases and controls at genome-wide significance level in GWAS are not in protein exons. In fact, most SNVs do not occur in protein-coding regions. New methods are being developed to predict which of these putative regulatory SNVs (known as rSNVs) may be consequential. Undoubtedly, as scientific knowledge about functional regions in mammalian and other genomes grows, our ability to enumerate and predict the possible impacts of SNVs will grow as well.

2 CLASSICAL APPROACHES

The functional assessment of single nucleotide polymorphism (SNP) prediction is a recent problem, yet some approaches have already

*To whom correspondence should be addressed.

become mainstream. Technically speaking, a SNP has referred to a point mutation having minor allele frequency greater than some threshold (typically 1–5%), distinguishing them from *rare variants*. There is a move in the genetics and cancer fields to use the broader term SNV instead of SNP, which encompasses both common and rare variation.

2.1 Impact of cSNVs

A cSNV can have negligible impact on normal protein activity or interfere with normal activity in mild to severe fashion. In some cases, cSNVs have major impact on genetic disease susceptibility (sickle cell anemia), common disease risk (Alzheimers) and drug sensitivities (warfarin) (McClellan and King, 2010).

2.1.1 Properties of amino acid residue substitution To a first approximation, these effects can be predicted by properties of amino acid substitution.

Many amino acid substitution metrics are based on estimates of the expected evolutionary distance between each possible amino acid pair. PAM matrices (Dayhoff and Schwartz, 1978) approximate this distance according to the distribution of amino acids at equivalent protein positions in closely related species. BLOSUM matrices (Henikoff and Henikoff, 1992) include distantly related species but only consider highly conserved protein regions. Both approaches use the raw mutation rates to compute a score for each amino acid substitution, which quantifies the likelihood that it was produced by an evolutionary trajectory over time rather than by chance. The intuition is that substitutions that are more consistent with evolutionary trends (*conservative* substitutions) are less likely to be disruptive while substitutions that are less consistent with evolution (*non-conservative* substitutions) are more likely to be deleterious.

Another approach is to consider how amino acid biophysical properties differ. Changes in volume, hydrophobicity, net charge, packing density and solvent accessibility have been shown to correlate with the functional impact of cSNVs (Wang and Moult, 2001).

The Grantham distance (Grantham, 1974) combines both biophysical properties and evolutionary distance. Grantham hypothesized that the biophysical importance of amino acid substitution could be quantified in a 3D space, whose coordinate axes are side chain composition, polarity and volume. Thus, each amino acid substitution can be represented by a Euclidean distance in this space (Supplementary Equation 1). To ascertain the relative importance of these three properties, he parameterized Euclidean distance by assigning a weight to each term. The weights were fit to amino acid substitution mutation rates as estimated by McLachlan (1971).

While PAM and BLOSUM likelihood scores, changes in biophysical properties and/or large Grantham distances may be predictive of the impact of populations of SNVs, these metric are not sufficient to make accurate predictions about individual cSNVs reliably (Karchin *et al.*, 2005; Ng and Henikoff, 2001). For example, a change from valine to isoleucine is conservative (positive), when measured by the common PAM250 and BLOSUM62 matrices, has no charge change and small Grantham distance. Yet, there are valine to isoleucine cSNVs that have been shown to be deleterious and clinically important. One instance is the SNV rs28928891, an isoleucine to leucine change at codon 322 of HOXD13, which causes brachydactyly and results in very short fingers and toes (Johnson *et al.*, 2003).

2.1.2 The evolutionary history of an amino acid position BLOSUM and PAM scores estimate an evolutionary distance separating a pair of amino acids. Ng and co-workers suggested that the importance of this distance depends on the position where an amino acid substitution occurs (Ng and Henikoff, 2001). For example, arginine to cysteine substitution has a non-conservative BLOSUM score, but this substitution should be treated differently at a position invariant for arginine versus a position where arginine and cysteine are both seen in protein family members.

The intuition behind considering the composition of amino acids at equivalent positions in a protein family is that functionally or structurally important positions will not tolerate a variety of amino acids. Equivalent positions are identified by building a multiple sequence alignment of related protein sequences and all residues that occur in a column of the alignment are assumed to be at equivalent positions. For example, if some alignment column consists entirely of tryptophans, then these tryptophans may be required for the members of the protein family to form a stable structure. During the course of evolution, nucleotides in the tryptophan codon may have been mutated, and organisms may have been born with other amino acids at this position. If no such amino acid variant can be observed today, then the tryptophan is important for the position, the protein and organisms that use the protein in their cellular machinery. This represents evidence of *purifying evolutionary selection* for tryptophan. Thus, positions where there is high conservation are likely to contain specific amino acid residues that are important for normal protein activity.

A variety of *scores* have been designed to quantify this intuition. One approach is to compute the frequency of the most common amino acid in an alignment column. Alternatively, *Shannon entropy* (Supplementary Equation 2) quantifies how surprised we are by the distribution of all amino acids in a column. Highly conserved columns present few surprises while columns of little conservation present more surprises (Schneider, 1997). *Relative entropy* refines this concept by comparing Shannon entropy of a column with the Shannon entropy of the amino acid background distribution.

Many scores incorporate both properties of amino acid substitution and evolutionary conservation. The SIFT score (Ng and Henikoff, 2003) is a weighted average of the frequency with which a variant amino acid residue appears in the multiple alignment column, and an estimate of unobserved frequencies *via* Dirichlet mixture pseudocounts (Sjander *et al.*, 1996). The PSIC score (Sunyaev *et al.*, 1999) considers the difference between the likelihood of the reference and variant amino acid at an alignment column, after encoding of the alignment in a position-specific scoring matrix (PSSM) (Gribskov *et al.*, 1987) (Supplementary Methods). The AGVGD score (Tavtigian *et al.*, 2006) is a position-specific variant of the Grantham distance, where the Grantham variation *GV* is computed by replacing each pair of components representing composition, polarity and charge with the maximum and minimum value in the alignment column. The Grantham deviation measures the extent that a variant amino acid deviates from the range of variation seen in the column, an estimate of its violation of evolutionary constraints on the protein position. The MAPP score uses a statistical summary of an alignment column

by constructing a phylogenetic tree and weighting each sequence by tree topology and branch lengths (Stone and Sidow, 2005). The mean and variance of amino acid physiochemical properties in the summarized column are used to estimate position-specific constraints on amino acid substitution in a biologically meaningful way. Finally, a principal components analysis (PCA) transforms the properties into decorrelated components, which are used to generate an integrated score that measures constraint violations with respect to all of the amino acid properties.

2.1.3 Sequence–function relationships There are core bioinformatics web services that contain information about sequence–function relationships and these can be useful in assessing whether a cSNV is functional. The best known is the UniProtKB database (Bairoch *et al.*, 2005), which maintains a feature table for each curated protein sequence to annotate regions and specific positions of interest, based on published studies. These features can be used to identify whether a cSNV occurs at a location that may be particularly sensitive to amino acid substitution, including DNA binding regions, biologically important sequence motifs, active site residues, metal-binding sites, sites of post-translational modifications and lipid-binding sites. In some cases, the results of mutational functional testing at a position are included as a feature, e.g. serine to alanine substitution at codon 362 of TP53 is annotated as abolishing binding to interaction partner USP7, as described in two referenced papers.

2.1.4 Structure–function relationships If a cSNV can be mapped onto an experimentally determined protein structure or a high-quality homology model, it is possible to compute a number of properties that are useful in predicting its functional impact. Solvent accessibility of an amino acid is one of the strongest predictors of cSNVs with functional impact, as substitutions in the hydrophobic core of soluble proteins can disrupt thermodynamic stability. Structural modeling of mutant proteins can be used to assess whether cSNVs induce backbone strain, lead to overpacking, occur in cavities, or impact key pairwise residue interactions (Yue *et al.*, 2005). Many X-ray crystal structures have been co-crystallized with interacting protein partners and/or small molecule, nucleotide and peptide ligands. Thus, the ability to locate a cSNV on a protein structure also makes it possible to assess, in some cases, whether the change occurs at or near a binding or catalytic site or at a domain–domain interface in a protein complex. Even in the absence of co-crystallized structures, electrostatic analysis of the surface of a protein structure or model can reveal highly charged patches whose disruption by a cSNV may impact binding interactions.

2.2 Impact of rSNVs

A SNV can have regulatory impact in many ways, from disrupting chromatin structure (McDaniell *et al.*, 2010) to causing the loss of a post-translational modification site (Li *et al.*, 2010). In many cases, we do not understand the regulatory processes well enough to predict if any given SNV will be consequential. Yet, there are areas where even if our knowledge is incomplete, we are prepared to make well-educated guesses. These areas include transcription, pre-mRNA splicing, premature termination codons, microRNA binding and post-translational modification.

2.2.1 Transcription A SNV can interfere with regulation by disrupting a transcription factor binding site. In the human genome, most annotated transcription factor binding sites (TFBSs) are derived through PSSMs, from sources such as TRANSFAC (Matys *et al.*, 2006) or JASPAR (Portales-Casamar *et al.*, 2010). Each PSSM describes a statistical profile of the sequences bound by a given transcription factor, according to experimental evidence from sources such as ChIP-seq [reviewed in (Hudson and Snyder, 2006)]. PSSMs assume that each position is independent. While questionable, this assumption allows researchers to estimate TFBS profiles from limited sets of observations. Within a TFBS, some positions may be highly conserved, reflecting that the transcription factor requires a specific nucleotide for proper binding, while other positions may be more divergent. This is captured by the relative entropy of each column in the PSSM. A SNV that alters a conserved position is more likely to be deleterious than one that alters a divergent position. A number of transcription factors also function as chromatin modification factors, such as CTCF (McDaniell *et al.*, 2010). SNVs in such binding sites of such factors can be especially consequential because of the potential to regulate not simply individual genes but also larger regions of chromatin.

The gain or loss of a predicted TFBS might not be consequential, for several reasons. First, PSSMs tend to be low specificity predictors, because transcription factor binding motifs are short and degenerate. *De novo* TFBS searches for binding sites tend to yield many hits with no clear regulatory role, and consequently researchers often limit TFBS prediction to the promoter region. However, recent experimental evidence indicates that most transcription factor binding in humans is outside of promoter regions [reviewed in (Farnham, 2009)]. Some of these TFBSs may have no regulatory importance, which is another reason why a TFBS gain or loss might not be consequential. Identifying the important TFBSs is an open research question, but one clear lesson has emerged: TFBSs are not likely to be occupied if they are not in regions of open chromatin [reviewed in (Berger, 2007)]. The ENCODE data in the UCSC Genome Browser details the regions of open chromatin for many common cell types (Raney *et al.*, 2010).

2.2.2 Pre-mRNA splicing SNVs that affect splicing tend to be consequential. Of disease-associated single point mutations, an estimated 15% are SNVs that alter splice sites in particular (Cartegni *et al.*, 2002; Krawczak *et al.*, 1992), and an estimated 62% alter splicing in general (Lopez-Bigas *et al.*, 2005). SNVs can alter splicing by disrupting an existing splice site, creating a *de novo* splice site inside an exon, and adding or removing splicing regulatory motifs such as exon-splicing enhancers (ESEs) and silencers (ESSs). Misspliced mRNAs are often degraded with little or no translation through nonsense-mediated decay;when translated to protein, they often yield unstable proteins that perform no catalytic function [reviewed in (Garcia-Blanco *et al.*, 2004)]. SNVs in exons can also introduce premature termination codons, which yield an mRNA transcript that is tagged for nonsense-mediated decay and produces little or no protein.

There are three categories of splice sites in the human genome. *Canonical* splice sites are flanked by the dinucleotides GT and AG, and represent more than 98% of human splice sites. *Non-canonical* splice sites are flanked by the dinucleotides GC and AG, and represent approximately 0.5% of human splice sites. *Non-consensus* splice sites are flanked with other dinucleotides, each of which represents less than 0.05% of human splice sites. These dinucleotides are contained within conserved motifs of eight to

twelve nucleotides in length. Canonical and non-canonical splice sites can be recognized with high accuracy by PSSMs (Burset *et al.*, 2000). Non-consensus splice sites are not amenable to modeling, due to their low frequency.

Any SNV close to a splice site has the potential to disrupt splicing, and splice site analysis is a key component in SNV function prediction. Common approaches include flagging SNVs if they map to one of the splice site dinucleotides, and assessing the impact of SNV through PSSMs [reviewed in (Mooney *et al.*, 2010)]. Methods that limit analysis to the splice site dinucleotides will miss SNVs that disrupt splicing at other conserved splice site positions. PSSMs are effective at capturing the importance of each splice site position and the likelihood of each nucleotide at each position. What they do not capture is the interdependence between certain positions in the splice site, where a SNV in one position is balanced by a compensatory SNV in a second position. These positions are not always adjacent (Yeo and Burge, 2004). Consequently, changes in PSSM score are not always correlated with changes in splice site strength, especially for splice sites that match the PSSM weakly.

Strong splice sites are necessary but not sufficient for correct splicing. Correct splicing also depends on exonic splicing regulatory (ESR) motifs including ESEs and ESSs. Predicted ESEs represent binding targets of SR proteins, a family of splicing enhancers and experimental evidence has shown that motifs targeted by the SR protein SFRS1 are enriched for disease-associated variants (Sanford *et al.*, 2009). Predicted ESSs represent targets of the hnRNP family of splicing silencers. Both ESEs and ESSs are predicted according to profiles of binding sites obtained through SELEX experiments (Klug and Famulok, 1994). Some ESR predictors apply PSSMs estimated from ESE or ESS binding site sequences. Others look for specific short sequences that are enriched in observed binding sites, or are overrepresented (in the case of ESEs) or underrepresented (in the case of ESSs) in exons compared with introns or other non-spliced sequences. To underscore their importance, ESEs cover roughly half of all exonic nucleotides in the human genome [reviewed in (Chasin, 2007)]. As this abundance might suggest, the gain or loss of any single ESR is not necessarily consequential. The SNVs that are more likely to alter splicing include those that introduce ESSs, those that impact a larger number of ESRs (considering that one SNV can impact several overlapping binding sites), and those in weakly defined exons with fewer ESEs and more ESSs (Woolfe *et al.*, 2010). The Skippy web server (http://research.nhgri.nih.gov/skippy) estimates the likelihood that a SNV will affect splicing according to these factors.

ESRs can also be found in intronic regions proximal to exons, where they appear to have a regulatory role that depends on their location relative to the exon and the splice sites. Besides ESEs and ESSs, there are many intronic splicing regulatory enhancers, especially in conserved regions proximal to alternative exons (Yeo *et al.*, 2007). Analysis of ISREs does not enter into many SNV analysis pipelines yet, because most ISREs are not yet well-characterized and the ISRE detection methods are still maturing. Yet ISRE detection is an active research area, and will probably become a component of SNV analysis in the future.

Finally, SNVs that introduce premature termination codons (PTCs) have a high likelihood of being consequential. In mammals, mRNA transcripts that contain termination codons (UAG, UAA or UGA) more than about 50 nt upstream of the last splice site are flagged for nonsense-mediated decay. While there are exceptions, these transcripts are generally decayed prematurely with little or no translation [reviewed in (Holbrook *et al.*, 2004)].

2.2.3 MicroRNA binding One more way in which SNVs can impact regulation is by altering microRNA (miRNA) binding sites [reviewed in (Mishra and Bertino, 2009)]. In animals, miRNAs are short (21–25 nt) non-coding RNAs that regulate as many as 30% of all human genes (Lewis *et al.*, 2005). miRNAs can perform a diversity of functions but are most often implicated in mRNA silencing through translational repression or mRNA cleavage [reviewed in (Mattick and Makunin, 2005)]. They generally bind to the 3′ UTR of target mRNAs. The overall binding may be weak, with limited Watson–Crick base pairing, but strong binding at nucleotides 2–7 the 5′ end of the miRNA is vital for proper target recognition. This would suggest that SNVs in this *seed region* would have a greater impact than SNVs elsewhere in the binding site, but the situation is not quite that simple.

There is a spectrum of target sites. 5′*-dominant* seed sites show perfect complementarity at the 5′ seed region but very poor complementarity at the 3′ end, while 3′ *compensatory sites* balance weaker pairing at the 5′ end (with seeds as small as 4–6 bp or 7–8 bp with bulges) with stronger pairing at the 3′ end (Brennecke *et al.*, 2005). Stronger overall pairing is associated with stronger silencing. Consequently, the impact of a SNV in a binding site depends on the entire binding site.

Because miRNA targets are difficult to determine experimentally, most miRNA binding sites are predicted computationally. There are now a variety of miRNA binding predictors [reviewed in (Rajewsky, 2006)]. While details differ, all predictors share some fundamental ideas. They all search conserved regions of the 3′ UTR of the mRNA for regions of strong or perfect complementarity to the seed region of the miRNA. Candidate matches are extended to longer, imperfect matches, and the thermodynamic stability of the mRNA–miRNA complex is evaluated. Some predictors consider additional features to increase specificity, such as the solvent accessibility of the binding site or clusters of putative binding sites in the same UTR. Yet, the shortness and limited complementarity of the binding sites make this prediction task difficult at best, with an estimated specificity of 50% (Rajewsky, 2006). Different predictive methods tend to yield contrasting sets of predictions, reflecting two things. First, when evaluating a weak signal, even small algorithmic differences can yield significant contrasts in the output. Second, our knowledge on miRNAs has been evolving rapidly, and different predictors reflect different stages and aspects of this knowledge. For example, a 3′-compensatory target site might not be detected by a miRNA target predictor, which requires a strong 5′ seed match. Thus, while it might seem practical to increase prediction confidence by applying different miRNA binding site predictors to the same region and the same SNV, such a strategy will not necessarily lead to a better prediction.

2.2.4 Altering post-translational modification sites The activity of almost every eukaryotic protein is modulated through post-translational modification (PTM). There are many forms of PTM [Uniprot currently lists over 400 (Bairoch *et al.*, 2005)], including many that are recently discovered [reviewed in (Witze *et al.*, 2007)]. An estimated 5% of disease-associated mutations result in the loss of a PTM site (Li *et al.*, 2010), which is almost certainly an

underestimate due to the many types of PTM sites that are not yet well-characterized.

There are two ways of estimating large-scale PTM sites: experimental measurement and prediction. Recent advances in mass spectrometry proteomics have dramatically improved the ability to detect PTM sites [reviewed in (Seo and Lee, 2004; Witze *et al.*, 2007)], but in spite of these advances, the vast majority of PTM annotations in dbPTM (Lee *et al.*, 2006) come from prediction. Yet due to the diversity of PTMs, each form of PTM must be predicted by a specialized predictor. Often, subsets of PTMs must be predicted separately. For example, in mammals there are many families of kinases, each of which bind to distinct sets of sequences, so kinase binding prediction is best addressed with a cadre of predictors, each of which predicts the binding of one specific kinase family [reviewed in (Juncker *et al.*, 2009)]. While this sort of specialization allows for a more precise prediction, it also increases the risk of overestimation as the number of training examples per example decreases (Blom *et al.*, 2004). Thus, when evaluating PTM site annotations, the onus is on the researcher to understand how the prediction was derived and the limitations of the method.

2.3 Bioinformatics predictors under the hood

2.3.1 Single versus multiple feature strategies While some cSNV and rSNV function predictors rely on only a single feature, an increasingly popular approach is to apply *data integration* strategies, which combine multiple features into a single predictive score. If a pair of features is highly correlated, there is probably not much to gain by using both. However, features that are uncorrelated or independent are expected to increase predictor accuracy when combined. Thus, it makes sense to design a predictor by identifying a set of highly informative and independent features.

To ascertain which features are useful, most researchers take an empirical approach. Good features are those which can be applied to discriminate between functional SNVs and neutral SNVs that have no functional impact. Intelligent feature selection requires that we have a collection of SNVs from both categories and that the collection is representative of the variety of SNVs within both categories. The assumption is that if the collection is large enough, it will be sufficiently representative of this variety. Importantly, if this collection is too small, not representative of the larger population of functional and neutral SNVs, or contaminated by mislabeled SNVs, we will not be able to ascertain whether selected features are useful.

2.3.2 Benchmark sets Researchers have used both functional assay results and data-mining approaches to assemble collections of functional and neutral SNVs. SNVs with known disease associations are often used as the functional SNVs. Such SNVs can been collected from curated databases such as SwissProt Variants (Yip *et al.*, 2004), OMIM (Hamosh *et al.*, 2005), HGMD (Cooper *et al.*, 2006) or obtained from clinical or functional studies. Almost all of these disease associations are with respect to Mendelian/monogenic diseases. The SwissProt variants also include cSNVs with no disease impact, which can be used as the neutral class. Alternatively, researchers have used common, high-frequency SNVs [from HapMap (Consortium, 2003)] as the neutral class (Schwarz *et al.*, 2010; Woolfe *et al.*, 2010) or generated artificial cSNVs from the equivalent amino acids in closely related species (Sunyaev *et al.*, 2001). Many early cSNV predictors took a different approach, and used the results of saturation mutagenesis experiments in bacterial and viral proteins as benchmarks (Chasman and Adams, 2001; Ng and Henikoff, 2001). For example, 4000 *Escherichia coli* lac repressor amino acid substitution mutants had been tested for their ability to repress β-galactosidase activity and respond to inducer, with a colorimetric assay (Markiewicz *et al.*, 1994). This yielded a set of amino acid substitutions that had no impact on function (neutrals) and a set of 'functional' substitutions that impaired function. Finally, some researchers use all available information about possible deleterious/neutral amino acid substitutions (human disease databases, systematic functional assays of mutation in single proteins) to construct large benchmark sets (Bromberg and Rost, 2007). Use of any of these benchmark sets assumes debatable cause-and-effect relationships between SNVs and phenotypic effects. The disease-based benchmarks assume a simple relationship between SNV and phenotype, when the reality is likely more complex, with the SNV as one contributing factor. The experimental benchmarks implicitly assume that a measured impact on a single molecular function is directly coupled with disease. Researchers should remain aware of the underlying uncertainties inherent in these assumptions.

2.3.3 Supervised learning Perhaps the most data-driven approach to SNV function prediction is to use a collection of functional and neutral SNVs both to assess the most relevant predictive features and to train a classification algorithm. In this case, the benchmark set is also used as a *training set*. The most commonly used classifiers are supervised statistical learning algorithms, in which each SNV is represented by multiple features and a *class label*. These algorithms detect patterns associated with each class and learn a decision rule, which is subsequently applied to SNVs whose class membership is unknown. Examples of such algorithms are classification trees, support vector machines, random forests, Bayesian networks and logistic regression [reviewed in (Kotsiantis, 2007)]. Supervised learners are successful if the decision rule yielded from the training phase is *generalizable* or able to correctly predict the class membership of examples (i.e. SNVs) that were not in the training set. Thus, they are evaluated on estimates of *generalization error*, not on estimates of misclassification during the training phase. In practice, such evaluations are done with *cross-validation* approaches, in which the training set is randomly split into n partitions. The experiment is repeated n times with each possible combination of $n-1$ partitions used for training and the remaining partition for testing.

Importantly, as with selection of predictive features, the relevance of the decision rule yielded by a supervised learning algorithm depends critically on the collection of SNVs used for training.

3 BUYER BEWARE

Researchers who would like to use bioinformatics methods to assess SNVs of interest are faced with a wide range of possibly useful methods. It makes sense to approach these methods with a critical perspective to decide which, if any, is a good choice for your particular purpose (Fig. 1).

Numerous bioinformatics SNV function predictors are available via web interfaces. These services are often easy to use, and it is tempting to apply them as black boxes, without a solid understanding of what they can and cannot deliver. This strategy is dangerous from both a practical and scientific perspective. Before using a

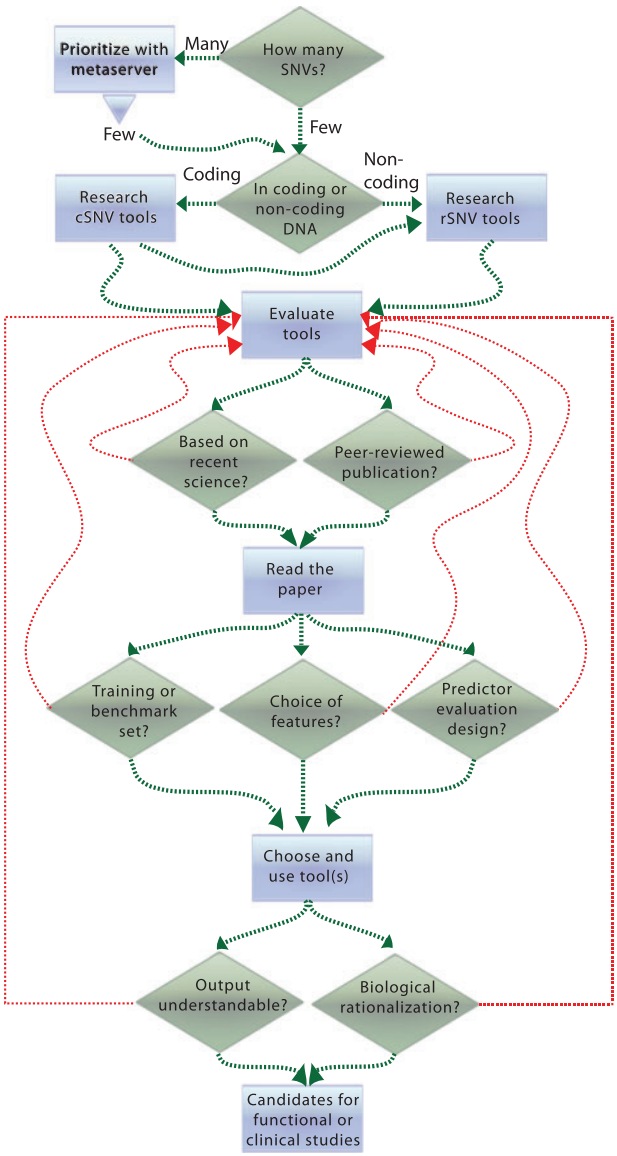

Fig. 1. Flow chart for informed use of SNV function prediction tools. Black box metaservers can be useful for narrowing down a large number of SNVs to a tractable set. SNVs in coding regions may have regulatory impact and should be assessed by both cSNV and rSNV tools. Once a tractable set of SNVs is selected, tools should be carefully evaluated and used only if they meet the criteria shown. After a tool is applied, users should be careful that they both understand the tool's output and can rationalize for themselves why SNVs are predicted to be functional or not. Finally, results can be used as candidates for planned clinical studies or functional testing.

web interface, we suggest that you study the underlying method(s) carefully.

Any method offered by a web service should be the subject of one or more peer-reviewed scientific publications. These publications will explain the scientific assumptions made by the method, its predictive features and algorithmic strategy, and assessment of its performance, most often compared to the performance of similar methods. They should also provide details about how predictive features were calculated and the training/benchmark set that was used for method assessment.

What does it mean when a publication claims that the SNV function predictor designed by the authors performs well compared to competing methods? The first thing to consider is the data used to develop the method, because this choice can bias the method in many subtle ways. For example, some ESE predictors were developed based on exons with weak splice sites, assuming that weak splice sites are balanced by a greater density of splicing enhancers. Arguably, these methods are biased toward alternative exons, which tend to have weaker splice sites, and might not be effective at evaluating SNVs in constitutive exons (Chasin, 2007). Similarly, any method that uses protein sequence alignments to evaluate cSNVs will be biased by the sequences that are included in the alignment. Many methods use alignments of all available sequences in large protein databases that can be inferred to have an evolutionary relationship to the SNV-containing sequence. To effectively identify biologically important conservation signals requires an appropriate sample of both closely and distantly related sequences, which may or may not be present in these databases for the sequence of interest. The best distribution of phylogenetic distances within such an alignment is not well understood. Yet, if the alignment contains sequences that are too distantly related to the SNV-containing sequence, it may contain residues that would be deleterious if substituted into the equivalent positions in human. If it contains only sequences that are too closely related to the SNV-containing sequence, almost all positions will appear to be conserved, not just those which are functionally important. Some methods use the contrast in conservation between closely related sequences and more distant homologs to identify important residues positions (Reva *et al.*, 2007), an approach which also requires good sampling of related proteins. The SIFT alignment algorithm is designed to avoid this pitfall by selecting a diverse set of homologs. However, even this method is not foolproof. For example, if the protein sequence database used to construct the alignment contains a large number of paralogs from a particular species, SIFT's output can still be biased toward that species.

The next consideration is the statistic(s) used to measure SNV function prediction performance. Statistics such as accuracy or error are based on the fraction of misclassified example SNVs in the benchmark set (which contains SNVs for whom the correct class is presumably known). These statistics will yield overly optimistic (or pessimistic) results if one class predominates in the benchmark set. One can achieve 80% accuracy in a benchmark set that contains 80% functional SNVs (the positive class), with a decision rule that always predicts the positive class. Yet, this decision rule has no scientific utility. Statistics that quantify the tradeoff between precision (fraction of correctly identified positives with respect to all predicted positives) and recall (fraction of correctly identified positives with respect to all positives in the benchmark) (Buckland and Gey, 1994) (e.g. the balanced F-measure) (Supplementary Equation 3) and the Matthews correlation coefficient (Supplementary Equation 4), are more desirable. One of the best performance measures is the AUC statistic (or area under the ROC curve), which is invariant to class distribution (Fawcett and Flach, 2005).

However, these statistics only have meaning with respect to the actual benchmark sets used. For example, to what extent does prediction performance on a benchmark set of loss-of-function versus neutral amino acid substitutions in a single bacterial

protein generalize to its utility in predicting SNVs in a variety of genes/proteins involved in complex human diseases? What if the benchmark set consists of Mendelian disease mutations and common SNVs? Does it make sense to combine results from functional assays of proteins in yeast and/or cell culture with disease-associated SNVs found in the literature? To our knowledge, these scientific questions have not been tested. However, before using a method you should understand the assumptions made by its choice of a benchmark set.

If a prediction method involves feature selection and/or supervised machine learning, *information leak* introduced during testing may yield an overly optimistic evaluation of performance. Accurate performance estimates require clean separation between the examples used for feature selection, classifier training, and classifier testing. Whenever an example is used for more than one of these purposes, information leak may occur. Information leak may also occur when there are dependencies between the examples used in feature selection, training and testing. For example, the dataset used to train and benchmark SNAP (Bromberg and Rost, 2007) contains two amino acid substitutions at codon 180 of the T2E5 gene in *E.coli*: alanine to glutamic acid (A180E) and alanine to lysine (A180K). If this data was split randomly into feature selection, training and test partitions, A180E might end up in the training partition and A180K in the test partition, yielding information leak. Before accepting the prediction performance claimed by authors of a method, you should look at its training/benchmark set and understand how this set was split up to assess performance.

Another problem involves implicit bias when comparing competing prediction algorithms. Unbiased comparison requires that both algorithms be trained and/or benchmarked on the same dataset. Before accepting claims of superior performance to a competing method, always check and see whether a method's authors assess themselves on examples used for training and how cross-validation was performed. Was the competing method originally trained/assessed on a different set of examples, but now evaluated with respect to the authors' training/benchmark set? In all fairness, the authors may have had no practical alternative but to take this approach. However, it calls into question claims that one method is superior to others.

Ideally, a user will download the software that implements a method, rather than use the web site, for a whitebox approach. Descriptions of a method in the literature may contain errors.

4 CONCLUSION

In an age of personal genomics, SNV prioritization has become essential. This task can now be facilitated by SNV *meta-servers* (Fig. 1), which are essentially black boxes to automate execution of multiple assessments, one for each of the many factors that can lead to a functional SNV. However, anyone who uses black box SNV assessment tools should understand how the assessments are made, and what their limitations are.

One limitation shared by all of the prediction methods described in this review is that prediction relies on knowledge, and our knowledge is incomplete. For example, current ESS predictors are based on only one family of splicing factors (hnRNP family proteins), but there is no reason to believe that these proteins represent the complete universe of ESSs. Happily, our knowledge is likely to expand thanks to new technologies. ChIP-seq technology represents a significant advance in evaluating the impact of SNVs on DNA binding, because there is no requirement of knowing the precise binding sequences in advance. This technology will facilitate research on the impact of SNVs on DNA methylation, chromatin and transcription factor binding, and RNA processing (through CLIP-seq). Additionally, the task of SNV assessment is likely to benefit from research currently underway on assessing the consequence of gains or losses of TFBSs, predicting ISREs, and further characterizing microRNA binding. Chromosomal segmentation may soon offer new insights on the medical relevance of intergenic SNVs, by distinguishing active intergenic chromatin from large-scale repressed chromatin (Ernst and Kellis, 2010). A community-wide effort CAGI (critical assessment of genomic Interpretation) has been created to compare analytical methods to assess genomic variation through blind prediction, modeled after the CASP contests and should provide excellent information on what types of methods are effective in practice (http://genomecommons.org/cagi/).

ACKNOWLEDGEMENT

We thank Laurence Meyer for insightful discussions and critical review of the manuscript.

Funding: This research was made possible with US Government support under and awarded by National Human Genome Research Institute (grants 5P41HG002371-09 and 5U41HG004568-02); Muscular Dystrophy Association (grant 135140 for M.S.C.); National Science Foundation CAREER award DBI 0845275 to R.K.

Conflict of Interest: none declared.

REFERENCES

Ahn,S.M. *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.*, **19**, 1622–1629.

Altshuler,D. *et al.* (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412.

Blom,N. *et al.* (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33**, 228–237.

Brennecke,J. *et al.* (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.

Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.

Buckland,M. and Gey,F. (1994) The relationship between recall and precision. *J. Am. Soc. Inf. Sci.*, **45**, 12–19.

Burset,M. *et al.* (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.

Cantor,R.M. *et al.* (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, **86**, 6–22.

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

Chasin,L.A. (2007) Searching for splicing motifs. *Adv. Exp. Med. Biol.*, **623**, 85–106.

Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.

Consortium,I.H. (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Cooper,D.N. *et al.* (2006) The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr. Protoc. Bioinformatics*, **Chapter 1**, Unit 1.13.

Dayhoff,M.O. and Schwartz,R.M. (1978) Chapter 22: a model of evolutionary change in proteins. In Dayhoff,M.O. (ed), *In Atlas of Protein Sequence and Structure*. Vol. 5, Supplement 1.

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

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.

Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.

Fawcett,T. and Flach,P.A. (2005) A response to Webb and Tings; on the application of ROC analysis to predict classification performance under varying class distributions. *Mach. Learn.*, **58**, 33–38.

Garcia-Blanco,M.A. *et al.* (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol.*, **22**, 535–546.

Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

Gribskov,M. *et al.* (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Gutenkunst,R.N. *et al.* (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, **5**, e1000695.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Holbrook,J.A. *et al.* (2004) Nonsense-mediated decay approaches the clinic. *Nat. Genet.*, **36**, 801–808.

Hudson,M.E. and Snyder,M. (2006) High-throughput methods of regulatory element discovery. *Biotechniques*, **41**, 673, 675, 677.

Johnson,D. *et al.* (2003) Missense mutations in the homeodomain of HOXD13 are associated with brachydactyly types D and E. *Am. J. Hum. Genet.*, **72**, 984–997.

Juncker,A.S. *et al.* (2009) Sequence-based feature prediction and annotation of proteins. *Genome Biol.*, **10**, 206.

Karchin,R. *et al.* (2005) Improving functional annotation of non-synonomous SNPs with information theory. *Pac. Symp. Biocomput.*, 397–408.

Klug,S.J. and Famulok,M. (1994) All you wanted to know about SELEX. *Mol. Biol. Rep.*, **20**, 97–107.

Kotsiantis,S.B. (2007) Supervised machine learning: a review of classification techniques. *Informatica*, **31**.

Krawczak,M. *et al.* (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.

Lee,T.Y., *et al.* (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.

Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Li,J.Z. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

Li,S. *et al.* (2010) Loss of post-translational modification sites in disease. *Pac. Symp. Biocomput.*, 337–347.

Lopez-Bigas,N. *et al.* (2005) Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.*, **579**, 1900–1903.

Markiewicz,P. *et al.* (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.*, **240**, 421–433.

Mattick,J.S. and Makunin,I.V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet.*, **14 Spec No 1**, R121–R132.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

McClellan,J. and King,M.C (2010) Genetic heterogeneity in human disease. *Cell*, **141**, 210–217.

McDaniell,R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

McLachlan,A.D. (1971) Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551 . *J. Mol. Biol.*, **61**, 409–424.

Mishra,P.J. and Bertino,J.R. (2009) MicroRNA polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine. *Pharmacogenomics*, **10**, 399–416.

Mooney,S.D. *et al.* (2010) Bioinformatic tools for identifying disease gene and SNP candidates. *Methods Mol. Biol.*, **628**, 307–319.

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

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

Portales-Casamar,E. *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38**(Suppl), S8–S13.

Raney,B.J. *et al.* (2010) ENCODE whole-genome data in the UCSC genome browser. *Nucleic Acids Res.*, **38** (Suppl. 1), D613–D619.

Reva,B. *et al.* (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.*, **8**, R232.

Sanford,J.R. *et al.* (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.

Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.

Schuster,S.C. *et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature*, **463**, 943–947.

Schwarz,J.M. *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.

Seo,J. and Lee,K.J. (2004) Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J. Biochem. Mol. Biol.*, **37**, 35–44.

Sjander,K. *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.

Stone,E.A. and Sidow,A. (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, **15**, 978–986.

Sunyaev,S.R. *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.

Sunyaev,S. *et al.* (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.

Tavtigian,S.V. *et al.* (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.*, **43**, 295–305.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.

Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.

Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Witze,E.S. *et al.* (2007) Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods*, **4**, 798–806.

Woolfe,A. *et al.* (2010) Genomic features defining exonic variants that modulate splicing. *Genome Biol.*, **11**, R20.

Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.

Yeo,G.W. *et al.* (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet.*, **3**, e85.

Yip,Y.L. *et al.* (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.*, **23**, 464–470.

Yue,P. *et al.* (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.*, **353**, 459–473.