Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm

Prateek Kumar[1], Steven Henikoff[2,3] & Pauline C Ng[1,3]

[1]Department of Genomic Medicine, J. Craig Venter Institute, San Diego, California, USA. [2]Basic Sciences Division, Howard Hughes Medical Institute, Seattle, Washington, USA. [3]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. Correspondence should be addressed to P.C.N. (png@jcvi.org).

Published online 25 June 2009; doi:10.1038/nprot.2009.86

The effect of genetic mutation on phenotype is of significant interest in genetics. The type of genetic mutation that causes a single amino acid substitution (AAS) in a protein sequence is called a non-synonymous single nucleotide polymorphism (nsSNP). An nsSNP could potentially affect the function of the protein, subsequently altering the carrier's phenotype. This protocol describes the use of the 'Sorting Tolerant From Intolerant' (SIFT) algorithm in predicting whether an AAS affects protein function. To assess the effect of a substitution, SIFT assumes that important positions in a protein sequence have been conserved throughout evolution and therefore substitutions at these positions may affect protein function. Thus, by using sequence homology, SIFT predicts the effects of all possible substitutions at each position in the protein sequence. The protocol typically takes 5–20 min, depending on the input. SIFT is available as an online tool (http://sift.jcvi.org).

INTRODUCTION

A non-synonymous single nucleotide polymorphism (nsSNP) occurring in a coding gene may cause an amino acid substitution in the corresponding protein product, thus affecting the phenotype of the host organism. Non-synonymous variants constitute more than 50% of the mutations known to be involved in human inherited diseases[1]. This demonstrates the important role of the non-synonymous variation in human health and the strong effects it can have on an organism's phenotype. With ~122,000 human nsSNPs in single nucleotide polymorphism database (dbSNP), a database of genetic variation hosted by the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/projects/SNP/), there is a significant need to characterize nsSNPs, with respect to their effect on the corresponding protein function. Prioritization of amino acid substitutions is also useful in conducting mutagenesis experiments and targeted gene re-sequencing studies. Although 'Sorting Tolerant From Intolerant' (SIFT) has been primarily applied to human polymorphisms, it can be applied to any organism because it is based on the principles of protein evolution. It has been applied to bacteria, plants and other animals[2,3]. We have also noticed that SIFT and other similar prediction tools have been used for clinical diagnosis. This is not recommended[4]; SIFT is intended to guide future experiments and not intended for direct use in a clinical setting, because *in silico* predictions are not a substitute for laboratory experiments.

SIFT is a multi-step algorithm that uses a sequence homology-based approach to classify amino acid substitutions (AASs)[3,5]. The SIFT algorithm was developed by our group at the Fred Hutchinson Cancer Center, and the SIFT server was recently transferred to the J. Craig Venter Institute. The underlying premise for this technique is based on the evolutionary conservation of the amino acids within protein families. Highly conserved positions tend to be intolerant to substitution, whereas those with a low degree of conservation tolerate most substitutions.

Algorithm

For a given protein sequence, SIFT compiles a dataset of functionally related protein sequences by searching a protein database using the PSI-BLAST algorithm[6]. It then builds an alignment from the homologous sequences with the query sequence. In the second step of the algorithm, SIFT scans each position in the alignment and calculates the probabilities for all possible 20 amino acids at that position. These probabilities are normalized by the probability of the most frequent amino acid and are recorded in a scaled probability matrix. SIFT predicts a substitution to affect protein function if the scaled probability, also termed the SIFT score, lies below a certain threshold value. Generally, a highly conserved position is intolerant to most substitutions, whereas a poorly conserved position can tolerate most substitutions. A schematic for the SIFT algorithm is depicted in **Figure 1**.

SIFT also provides a measure of confidence in the prediction. There can be cases in which there is too little sequence diversity in the set of aligned sequences to derive a good substitution model. For example, in the case of a viral protein that has been sequenced many times over, the top hits returned by a BLAST search are likely to be nearly identical to each other. This might lead to many positions in the alignment appearing as highly conserved. In such a case, the algorithm may predict neutral substitutions as deleterious, thereby increasing the false positive error[7]. To assess confidence in the prediction, SIFT calculates a conservation value at each position in the alignment. The conservation value for a position ranges from zero, when all 20 amino acids are observed at that position, to $\log_2 20$ (=4.32), when only one amino acid is observed at that position[5]. To maintain the optimum diversity within the selected sequences, SIFT ensures that the final set of aligned sequences has a median conservation value of ~3.0. If the set of sequences used for prediction are too conserved (median conservation value >3.25), then a low-confidence warning is issued.

One limitation of the SIFT algorithm is that it does not use the protein structure to assess the effect of an AAS. However, addition of a protein structure marginally improves the performance of AAS prediction algorithms[8].

Accuracy of prediction

When applied to a dataset of mutations found in affected individuals with disease, SIFT correctly predicted 69% of the substitutions

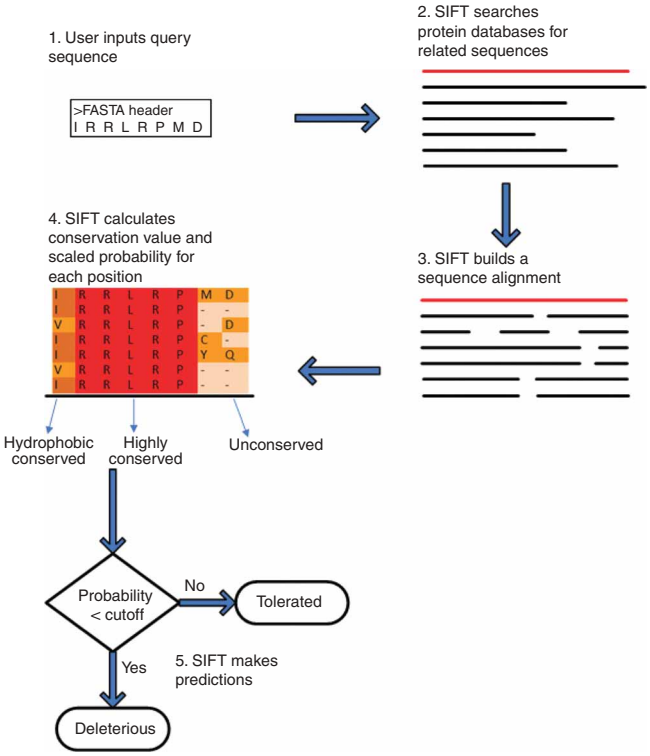

Figure 1 | 'Sorting Tolerant From Intolerant' (SIFT) algorithm flowchart for scoring individual amino acid substitutions (AASs). The process flow is described for the amino acid sequence 'IRRLRPMD'.

associated with the disease to affect protein function[5]. When applied to a second dataset consisting of nsSNPs in healthy humans[9], SIFT predicted only 19% of variants to affect protein function. Thus, most of the variants in normal individuals are predicted to be tolerated. The performance of SIFT was also analyzed for nsSNPs in the dbSNP database. **Table 1** summarizes some SIFT predictions for substitutions that were found to be associated with the disease[5]. One of the proteins shown in **Table 1** is the melanocyte-stimulating hormone receptor (MSHR), which is associated with an increased risk of cutaneous malignant melanoma (CMM)[10]. All three AASs in the MSHR gene are known to increase the risk of CMM, and were correctly predicted by SIFT to affect protein function as they were found to be at highly conserved positions (R151C, R160W and D294H). SIFT also predicted the substitutions of L60V and R163Q to be neutral, and this is supported by experimental evidence[10]. Thus, the ability of SIFT to distinguish between neutral and deleterious substitutions supports the use of SIFT as a prediction tool.

Alternative methodologies

There are alternative prediction tools that use a combination of methods based on sequence homology, protein structure information and physicochemical properties of amino acids for prediction[11]. Prediction methods based on protein structure alone are limited in scope to proteins with known structural homologs; thus, structure-based prediction is often an added feature of a sequence-based method. For methods that rely on sequence homology, it is important to select the optimal set of related sequences, which are distantly related orthologs[11]. **Table 2** summarizes the attributes of

TABLE 1 | A sample of SIFT predictions for nonsysnonymous variants from dbSNP

Protein	Prediction supported by evidence?	Substitution	SIFT Prediction	Evidence
Melanocyte stimulating hormone receptor	Yes	R151C	Damaging	Increased risk of cutaneous malignant melanoma (CMM)[10]
		R160W	Damaging	Doubles risk of CMM[10]
		D294H	Damaging	Doubles risk of CMM[10]
	No	D84E	Damaging	Not associated with CMM[10,16]
Fibrillin	Yes	D1113G	Damaging	In Marfan patient[17]
		C2038Y	Damaging	In Marfan patient (P. Oefner, personal communication)
		C2053F	Damaging	In Marfan patient (P. Oefner, personal communication)
		C2500S	Damaging	In Marfan patient (P. Oefner, personal communication)
		G2514R	Damaging	In Marfan patient (P. Oefner, personal communication)
		C2170F	Damaging	In Marfan patient (P. Oefner, personal communication)
		P1148A	Tolerated	In control individual[17]
		V2018I	Tolerated	In control individual (P. Oefner, personal communication)
		Y2113F	Tolerated	In control individual (P. Oefner, personal communication)
		D2329E	Tolerated	In control individual (P. Oefner, personal communication)
		P2278S	Tolerated	In control individual (P. Oefner, personal communication)
	No	S1077P	Tolerated	In Marfan patient[17]
		V1667I	Tolerated	In Marfan patient[17]
		N1341S	Tolerated	In Marfan patient[17]
		N1282S	Tolerated	In Marfan patient[17]

CMM, cutaneous malignant melanoma; dbSNP, single nucleotide polymorphism database; SIFT, sorting tolerant from intolerant. Many of the substitutions predicted to be damaging were shown to be involved in disease or found in disease patients. This is a partial list reproduced with permission from Genome Research (2002) (ref. 5).

TABLE 2 | Some alternative methodologies available as online tools

Method	Algorithm
SIFT (http://sift.jcvi.org)	SIFT uses sequence homology; scores are calculated using position-specific scoring matrices with Dirichlet priors
Polyphen[12] (http://genetics.bwh.harvard.edu/pph/)	Polyphen uses sequence conservation, structure and SWISS-PROT annotation
PMUT[13] (http://mmb2.pcb.ub.es:8080/PMut/)	PMUT provides prediction by neural networks, which use internal databases, secondary structure prediction and sequence conservation
SNPs3D[18] (http://www.snps3d.org/)	SNPs3D is based on support vector machine that uses structural or sequence conservation features
PantherPSEC[19] (http://www.pantherdb.org/tools/csnpScoreForm.jsp)	Panther PSEC uses sequence homology; scores are calculated using PANTHER Hidden Markov model families
MAPP[14] (http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html)	MAPP considers the physicochemical variation present in a column of a protein sequence alignment to predict the effect of all possible amino acid substitutions on protein function
Align-GVGD[15] (http://agvgd.iarc.fr/agvgd_input.php)	Align-GVGD combines the biophysical characteristics of amino acids and protein multiple sequence alignments

some of the alternative methodologies used for predicting the effects of AASs. Among the methods listed in **Table 2**, prediction of functional effect of human nsSNPs (Polyphen)[12] and PMUT[13] use a combination of structural and sequence homology, whereas MAPP[14] and Align-GVGD[15] use sequence homology along with the physicochemical properties of amino acids for AAS prediction. The user is advised to choose the tool suitable for the type and size of input data. In the absence of a standardized performance test to measure the accuracy of these tools, we do not recommend any one tool based on accuracy of prediction.

SIFT toolbox

Before describing the different options in the SIFT toolbox, it is important to describe some of the standard identifiers and representations that will be referred to throughout this protocol.

(1) *NCBI protein identifiers*: There are two types of protein identifiers supported by SIFT: NCBI Reference protein sequence (RefSeq) identifiers and NCBI GI numbers. RefSeq identifiers are alphanumeric and start with the prefix NP for known proteins or XP for predicted proteins, whereas NCBI GI numbers are numeric identifiers that are serially assigned to sequences processed by NCBI. To obtain these identifiers, the user can visit the NCBI website and search the protein database for their protein of choice. For example, the user may enter 'BRCA2 human' as a protein of interest at the NCBI protein search page. From the results, the user may choose the RefSeq tab on the top filter menu to obtain RefSeq hits. The final line of each hit returned is an identifier string. For 'BRCA2 human', NCBI returns gi|119395734|ref|NP_000050.2|[119395734], in which the numeric sub-string following 'gi' is the GI number (119395734) and the alphanumeric substring following 'ref' is the versioned RefSeq identifier (NP_000050.2). Unversioned RefSeq identifiers are required as an input to SIFT, i.e., NP_000050 instead of NP_000050.2. Please note that a single protein sequence may be denoted by both the GI number and the RefSeq identifier and either of these identifiers may be used as an input.

(2) *Amino acid substitutions*: The standard format for specifying an AAS is (original amino acid)(position)(substituted amino acid). For example, a substitution at position 10 from alanine to cysteine should be represented as A10C.

(3) *dbSNP reference SNP identifiers (rs ID or RefSNP)*: These identifiers represent a SNP in the NCBI dbSNP database. They have a prefix 'rs' followed by a number, e.g., rs1098. To obtain these identifiers, the user can visit the dbSNP database at http://www.ncbi.nlm.nih.gov/projects/SNP/.

The SIFT toolbox offers six separate functionalities for users to submit their input data. Two of them are batch tools that allow predictions for multiple proteins and their substitutions and the remaining four offer detailed predictions for a single protein for all or selected substitutions. To avoid repetition, we describe five user input options. The omitted option is similar to the Tool E described below, except that the omitted option allows users to input related homologs without aligning them first.

The choice of the suitable functionality depends on the kind of analysis required and the type of input data. These functionalities are depicted in **Figure 2**. For its batch tools, SIFT uses a database of

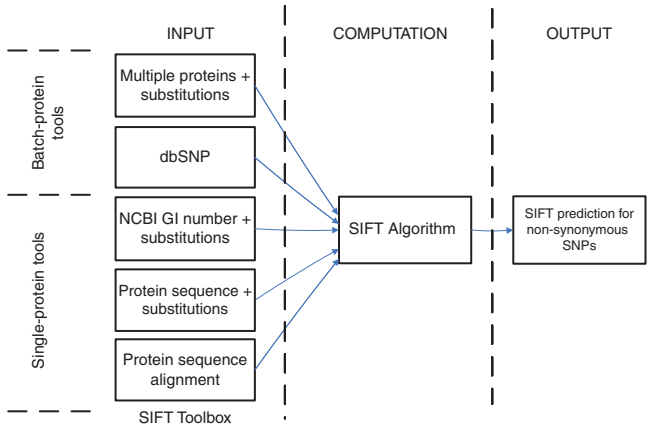

Figure 2 | Basic functionalities and flow of 'Sorting Tolerant From Intolerant' (SIFT).

TABLE 3 | Functionality, input and output of each of the tools in SIFT toolbox

Tool	Type of analysis	Input data type	Expected output	● TIMING
Tool A protein identifiers	Batch analysis of multiple proteins and their amino acid substitutions	List of protein identifiers (RefSeq identifiers or NCBI GI numbers) and list of amino acid substitutions	Predictions for user-specified substitutions in tabular format. If no substitutions are specified, predictions are returned for dbSNP substitutions for that protein	5–10 min for a list of 1,000 protein identifiers
Tool B dbSNP identifiers	Batch analysis of multiple nsSNPs (rs IDs) in NCBI's dbSNP database	List of dbSNP rs IDs	Predictions for user-specified nsSNPs (rs IDs) in tabular format	2–3 min for a list of 1,000 rs Ids
Tool C GI number	Detailed analysis for all substitutions in the user-specified protein	NCBI GI number or RefSeq identifier, selected amino acid substitutions	All predictions file, scaled probability matrix file and predictions file for selected amino acid substitutions	2–3 min
Tool D protein sequence	Detailed analysis for all substitutions in the user-specified protein	Protein sequence in FASTA format, selected amino acid substitutions	All predictions file, scaled probability matrix file and predictions file for selected amino acid substitutions	Up to 20 min
Tool E query sequence with aligned sequences	Detailed analysis for all substitutions in the user-specified protein	Multiple protein sequence alignment in FASTA format, selected amino acid substitutions	All predictions file, scaled probability matrix file and predictions file for selected amino acid substitutions	Less than a minute

pre-computed results for all nsSNPs in the current version of dbSNP database (Build 129). **Table 3** briefly describes the functionality, input and output of each of the tools in the SIFT toolbox. The user can supply input by either uploading text files or by pasting data into the suitable text boxes provided. For each of the tools, the user may obtain result files via email by providing the email address in the suitable textbox.

Tool A: Protein IDs with substitutions. This is a batch protein tool that should be used if the user wants a batch analysis of multiple proteins with predictions of nsSNPs from dbSNP or AASs of their choice. This tool may take 5–10 min to run for an input list of 1,000 protein identifiers. A screenshot of this functionality is shown in **Figure 3**. When using this tool it is possible for the user to set two parameters.

(i) *Sequences to include (BEST HITS or ALL HITS)*: This parameter determines the properties of the set of related sequences that SIFT obtains through a BLAST search. The default choice for this parameter is BEST HITS. The BEST HITS option aims for functional similarity among the sequences. These sequences are the top hits from each organism represented in the BLAST results and are assumed to be orthologs. The ALL HITS option simply fetches the top 100 hits ordered by their score after running BLAST with default parameters. This option should be chosen if the BEST HITS option yields a selection of sequences not diverse enough for prediction. In such a case, SIFT gives a warning about low confidence in the prediction, which is discussed in the TROUBLESHOOTING section of this protocol.

(ii) *Identity percentage*: This parameter allows the user to mask out sequences that are highly identical to the query sequence.

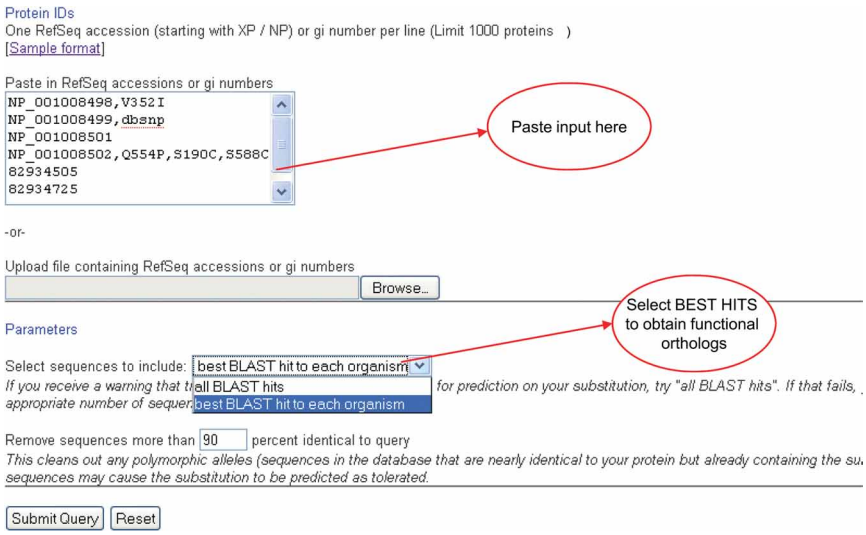

Figure 3 | Input screen for the 'Sorting Tolerant From Intolerant' (SIFT) batch tool that takes as an input a list of protein identifiers with corresponding amino acid substitutions (AASs). Shown here are the input text boxes and the optional parameters section. http://sift.jcvi.org/www/SIFT_pid_subst_all_submit.html.

These sequences may already contain the nsSNP to be analyzed because of the sequencing of polymorphic alleles. The default value for this parameter is 90%. We have rarely needed to change this parameter and recommend using the default value.

Tool B: dbSNP identifiers. This is a batch protein tool that should be used if the user is interested in predictions for known SNPs contained in the dbSNP database. This tool may take 2–3 min to run for an input list of 1,000 rs IDs.

Tool C: GI number. This is a single protein tool that accepts a GI number or a RefSeq identifier as an input and allows the user to obtain detailed predictions for the corresponding protein. SIFT obtains related proteins by using NCBI's pre-computed BLAST Link tool, which makes this option much faster than using a protein sequence as input (Tool D described in the following section). The user should choose this option if he/she is interested in obtaining predictions for all of the possible AASs at every position in the protein sequence. It is also possible to specify AASs of interest and to modify the parameters as described in Tool A.

Tool D: Protein sequence. This is a single protein tool that should be used if the user wishes to analyze a specific protein sequence. However, this is a time-consuming option and may take up to 20 min to run as it requires a BLAST search to compile a set of related proteins. It is advisable to use Tool C if the GI number or the RefSeq identifier of the protein is known. A screenshot of this functionality is shown in **Figure 4**. In this option the FASTA format of the protein sequence is required. To obtain this the user may search either the UniProt protein database (http://www.uniprot.org/) or NCBI's protein database (http://www.ncbi.nlm.nih.gov/). The FASTA sequence requires a header line and an amino acid sequence. An example of the FASTA format is shown below.

```
>gi|440979|gb|AAB28880.1| Oct-1B [Homo sapiens]
SPILAIFPSPTSLALQTPPPTTQQP
```

The header is the first line that starts with the character '>' and contains the relevant sequence identifier and name. The amino-acid sequence starts on the second line. For further information about the FASTA format, please visit the NCBI website. SIFT runs a BLAST search for the input query sequence to obtain related homologs. When using this tool a user can set three parameters.

(i) *Database to search (UniProt SwissProt, UniProt TrEMBL or NCBI nr)*: The default choice of database is UniProt-TrEMBL; this is a large high-quality database. Additional options include the UniProt-SwissProt, which is a small but high-quality database and the NCBI's non-redundant (nr) database, which is the largest database. Although the selection of a smaller database would decrease the time taken for the BLAST search, in some cases, it may be appropriate to choose a larger database such as NCBI's 'nr' database to ensure sufficient diversity among the set of related sequences. A larger database should be chosen when SIFT finds the median sequence information value to be above 3.25 and displays a warning message as discussed in the TROUBLESHOOTING section of this protocol.

(ii) *Median conservation of sequences*: The conservation value for a position in the alignment ranges from zero, when all 20 substitutions are observed at that position, to $\log_2 20$ (=4.32), when only one amino acid is observed at that position[5]. Thus, the median of the conservation values at all positions in the alignment is a measure of the diversity of the set of sequences. The default value for this parameter is 3.0. We have optimized this parameter and recommend keeping the default value.

(iii) *Identity percentage*: Please refer to the parameters section of Tool A for information on setting this parameter.

Tool E: Protein sequence alignment. This is a single protein tool and should be used when the user has a multiple alignment of the query sequence with homologous sequences. As the quality of the substitution model depends on the quality of the related orthologous sequences obtained, this tool can be used by a more advanced user who may want to use a manually refined set of related sequences. This tool typically takes less than 1 min to run. A manually refined alignment of orthologous sequences that share the same function usually yields the best results. The aligned sequences can be obtained either by searching the NCBI protein database using BLAST or by using BLink (if you have the protein identifier in addition to the sequence) to obtain NCBI's pre-computed BLAST results for your protein. To use BLink, the user should place the protein identifier (GI number or RefSeq identifier) at the end of the URL (in place of proteinID) http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?pid=proteinID and get the results. Once homologous sequences have been retrieved by either BLAST or BLink, the user can use online tools such as ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) to align the sequences. The final input to SIFT is a list of sequences in the multi-FASTA format, a sample of which is shown below.

```
>QUERY
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT

>P27757 Envelope glycoprotein gp160 OS=Simian
ELRLRYCAPAGYALLKCNDWDYAGFKPECSNVSVVHCTTLMNTTVTTGLLLNGSYSENRT

>P27977 Envelope glycoprotein gp160 OS=Simian
ELRLRYCAPAGFALLKCNDYDYAGFKTNCSNVSVVHCTNLINTTVTTGLLLNGSYSENRT

>P32536 Envelope glycoprotein gp160 OS=Human
XMRFRYCAPPGFALLRCNDTNYSGFEPNCSKVVAATCTRMMETQTSTWFGFNGTRAENRT

>P20872 Envelope glycoprotein gp160 OS=Human
XMRFRYCAPPGFALLRCNDTNYSGFEPNCSKVVAATCTRMMETQTSTWFGFNGTRAENRT
```

MATERIALS

EQUIPMENT

Software: Internet browser (such as Mozilla Firefox, Internet Explorer, etc.) and a text editor for creating list files and/or comma-separated files.
Files: A text file of one of the following: a list of protein IDs, a protein sequence alignment, a protein sequence in FASTA format or a list of NCBI SNP IDs (rs IDs). For the first three files listed above, a list of substitutions of interest in the corresponding protein can also be submitted.

PROCEDURE

Accessing the SIFT web application

1| Make sure your computer is connected to the Internet and has an Internet browser installed and running. Visit the SIFT website (http://sift.jcvi.org). The homepage consists of the SIFT toolbox in the left column with six functionalities to choose from. The right column contains basic introduction about SIFT along with the link to download the source code and links referencing the literature about SIFT.

Preparing and submitting the input data

2| SIFT offers different user input options in its toolbox. The option selected depends on your choice of analysis or the availability of starting data. Use option A to analyze multiple protein IDs with substitutions, option B to analyze multiple dbSNP identifiers, option C to analyze a single GI number or RefSeq Identifier, option D to analyze a protein sequence and option E to analyze a query protein sequence along with aligned sequences.

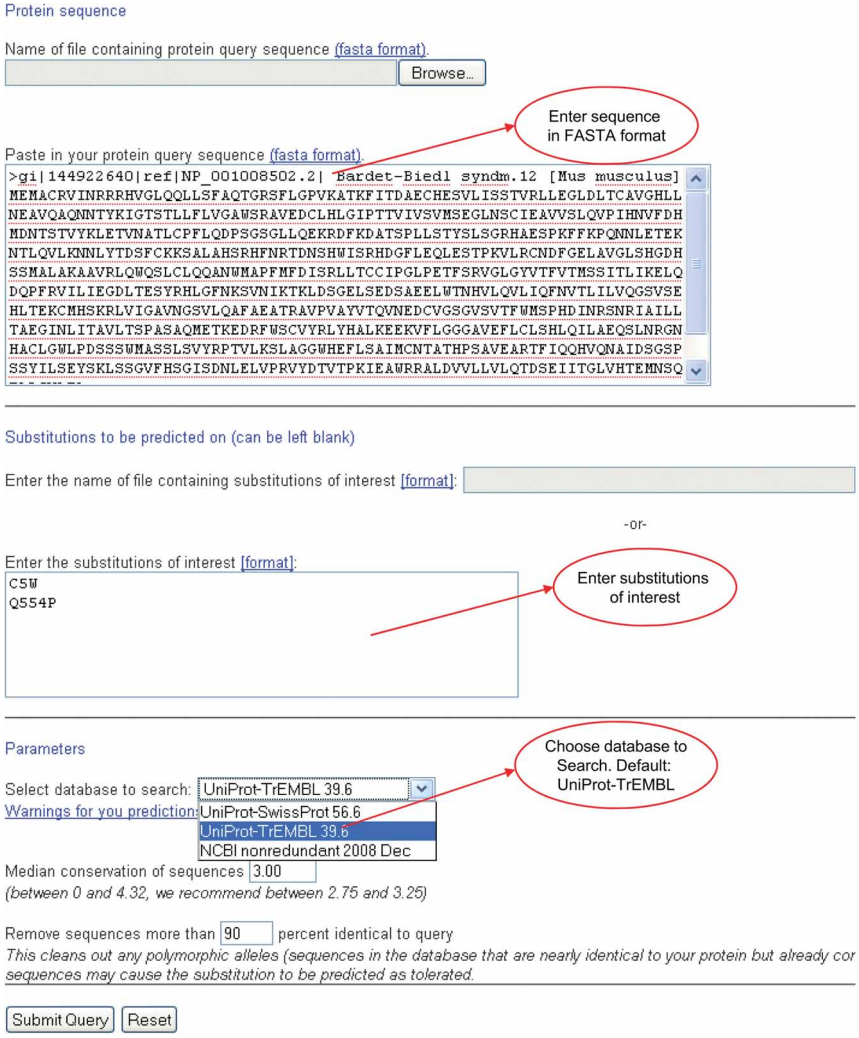

Figure 4 | Input screen for a single protein that takes as an input the protein sequence along with corresponding amino acid substitutions (AASs). Shown here are the input text boxes and the optional parameter selection. http://sift.jcvi.org/www/SIFT_seq_submit2.html.

(A) Protein IDs with substitutions

(i) Obtain the list of identifiers belonging to the proteins that require prediction analysis, e.g., NCBI reference sequence (RefSeq) identifiers or NCBI GI numbers. The input list may contain a combination of RefSeq identifiers and GI numbers. Once compiled, the list may look like this:

NP_001008567
NP_001008589
82934725
82934691

▲ CRITICAL STEP Do not include version numbers in the RefSeq identifiers in the input data.

(ii) Obtain a list of AASs for each corresponding protein identifier chosen in Step 2 option A (i) above (optional).

(iii) Compile the final list for submission by appending each protein identifier in the list from Step 2 option A (i) with the AASs obtained in Step 2 option A (ii). Add the reserved word 'dbsnp' or leave it blank to obtain SIFT predictions for all of the AASs that are in NCBI's dbSNP database. An example of a final list for submission is shown below.

NP_001008499,dbsnp
NP_001008501
NP_001008502,Q554P,S190C,S588C

▲ CRITICAL STEP Make sure that the entries in the list are comma separated.

(iv) Copy and paste the input data into the relevant textbox or upload an input file. Sample input data may be obtained by clicking the link 'Sample Format', which is above the input textbox for protein identifiers.

(v) In the drop down menu for the parameter 'sequences to include', choose between BEST HITS or ALL HITS. The default choice for this parameter is 'BEST HITS' (optional).
(vi) Choose a value for the parameter 'identity percentage'; this allows the masking of sequences that are highly identical to the query sequence. The recommended and default value for this parameter is 90% (optional).
(vii) Enter your email address on this page if you wish the results to be emailed to you. Click the 'submit query' button and wait for the results.
? TROUBLESHOOTING

(B) dbSNP Identifiers

(i) Prepare a list of dbSNP reference SNP ID (rsID or RefSNP) containing one entry per line. Refer to the SIFT Toolbox section of the Introduction for how to obtain rsIDs.
rs1098
rs1258
rs1303
rs1320
rs1337
(ii) Copy and paste the list of nsSNPs into the relevant textbox or upload it as an input file. Sample input data may be obtained by clicking the link 'Sample Format', which is above the textbox for dbSNP identifier input.
(iii) Enter your email address on this page if you wish the results to be emailed to you. Click the 'submit query' button and wait for the results.

(C) GI Number

(i) Obtain the NCBI GI number or the RefSeq identifier associated with the protein of interest. Refer to the SIFT Toolbox section of the Introduction for how to obtain NCBI GI numbers or Refseq identifiers.
(ii) Obtain a list of AASs of interest (Optional).
(iii) Enter the GI number or the RefSeq identifier of the protein of interest in the text box titled 'GI Number'.
(iv) Paste a list of AASs prepared as described in Step 2 option A (ii) in the suitable textbox (optional). Alternatively, you may upload the text file containing the list of AASs.
(v) Choose the parameters as described in Step 2 option A (v) and option A (vi) (optional).
(vi) Enter your email address on this page if you wish the results to be emailed to you. Click the 'submit query' button and wait for the results.
? TROUBLESHOOTING

(D) Protein sequence

(i) Obtain the FASTA sequence of the protein of interest from either the UniProt protein database or NCBI's protein database.
(ii) Obtain a list of AASs of interest (optional).
(iii) Paste the protein sequence in FASTA format into the suitable textbox. Alternatively, you may upload the FASTA file.
(iv) Paste the AASs of interest in the suitable text box or upload the text file containing substitutions (optional).
(v) Select the option for the parameter 'Database to search' from the drop down menu (default is UniProt-TrEMBL) (optional).
(vi) Choose a value for the parameter 'Median conservation of sequences' (the recommended and default value is 3.00) (optional).
(vii) Choose a value for the parameter 'identity percentage' as explained in Step 2 option A (vi) (optional).
(viii) Enter your email address on this page if you wish the results to be emailed to you. Click the 'submit query' button and wait for the results.
? TROUBLESHOOTING

(E) Protein sequence along with aligned sequences

(i) Create a multi-FASTA file with your query sequence at the top of the file, followed by aligned homologous sequences. For further instructions, refer to the SIFT Toolbox section in the Introduction.
(ii) Obtain a list of AASs of interest (optional).
(iii) Upload or paste the aligned multi-FASTA file prepared in Step 2 option E (i) in the suitable text box.
(iv) Upload or paste the amino acid substitutions list into the suitable text box (Optional).
(v) Enter your email address on this page if you wish the results to be e-mailed to you. Click the 'submit query' button and wait for the results.

● **TIMING**

Each of the six options described in Step 2 of the PROCEDURE section typically takes less than 20 min to run.
More specifically,
Step 2A: Submission of protein identifiers with their substitutions takes about 5–10 min for a list of 1,000 proteins.

Protein ID	Substitution	dbSNP ID	Prediction	Score	Median Info	Number of Seqs at position
NP_001008498	V352I	rs27180025	TOLERATED	1.00	1.38	99
NP_001008499	Y219C	rs33644739	DAMAGING	0.00	1.90	98
NP_001008501	C192Y	rs32419104	TOLERATED	0.25	1.46	84
82934505	Q23L	rs13484479	DAMAGING *Warning! Low confidence.	0.00	4.32	1
82934725	K52E	rs26925385	TOLERATED	1.00	4.32	2
82934691	P52R	rs26956085	TOLERATED	1.00	4.23	101

* Low confidence means that the protein alignment does not have enough sequence diversity. Because the position artifically appears to be conserved, an amino acid may incorrectly

Click here to download the following table in tab separated format. You can open it using excel with delimiter set as TAB

If you received a warning that the sequences were not diverse enough, you can have SIFT choose more diverse sequences here.

Figure 5 | Output table of batch tool that uses the input type of protein identifiers with corresponding substitutions.

Step 2B: Submission of dbSNP RefSNP IDs takes 2–3 min for a list of 1,000 rsIDs.
Step 2C: Submission of GI number takes 2–3 min to run.
Step 2D: Submission of a protein sequence may take up to 20 min depending on the time it takes to run the BLAST search. Longer protein sequences may take longer to process.
Step 2E: Submission of a protein sequence alignment takes less than 1 min to run.

? TROUBLESHOOTING

Step 2 option A and option C: If you receive the warning 'the median sequence information is greater than 3.25 for your amino substitution', then this implies that there is too little diversity in the protein sequence alignment. Choose the ALL HITS option in the parameters section. This option fetches the top 100 hits ordered by their score after running BLAST with default parameters.

Step 2 option D: A prediction for an AAS may have a low-confidence warning. If you receive the warning that 'the median sequence information is greater than 3.25 for your amino substitution', then this implies that there is too little diversity in the protein sequence alignment. This might lead to many positions in the sequence appearing as highly conserved, which increases the false-positive error where neutral substitutions are predicted to be deleterious[5]. In such a case, choose the largest (nr) database in the parameters section of your submission page. This may remedy the problem of low diversity if additional homologous sequences are found.

Step 2 option D: Processing of the protein sequence query as described in Step 2 option D may take about 20 min. A link is provided that leads to the results page, which will be automatically refreshed once the results are available. If you get a time-out message, then use a smaller (UniProt-SwissProt) database to search from.

You may also refer to the SIFT help page: http://sift.jcvi.org/www/SIFT_help.html
Or contact us with your specific problem at http://sift.jcvi.org/sift-bin/contact.pl

ANTICIPATED RESULTS

As mentioned in the procedure, the results of the analysis can either be obtained via email or can be viewed online. The results for batch tools are presented in tabular format, whereas those for the single protein tools are in the form of html and tab-separated text files.

SIFT results table for batch tools: as shown in **Figure 5**, the columns in this table are described below.

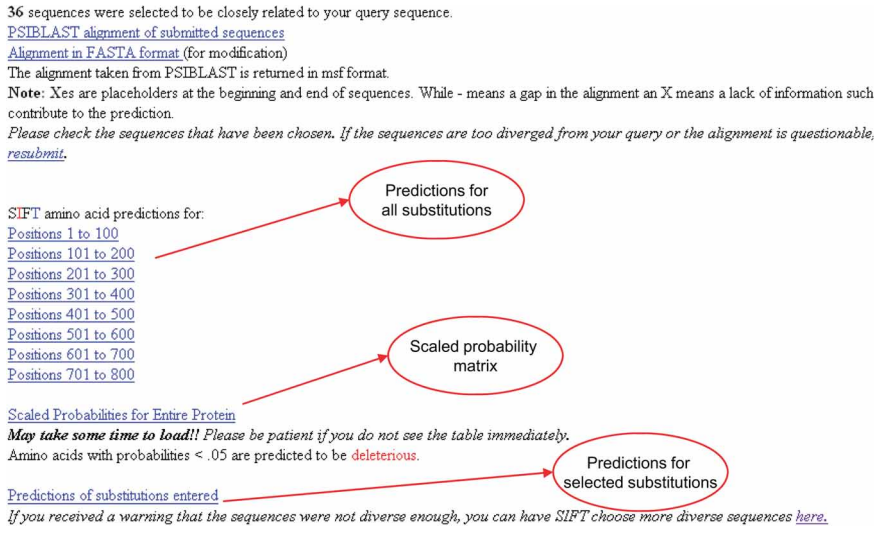

Figure 6 | Results summary screen for a single protein. This page contains links to predictions for all substitutions (**Fig. 7**), the scaled probability matrix (**Fig. 8**) and predictions for user-submitted substitutions (**Fig. 9**).

Predict not tolerated	Position	Seq Rep	Predict tolerated
d c g w h n e s p r k q y t	3M	0.60	a f v i ML
w	4A	0.60	c f m y i h v l p r q t n k s e GDA
w m h f	5C	0.80	y i q r e l k d p n v g t a S C
w c m f	6R	0.80	y i h v p l d g n q e t a k S R
w d h q p n c e r g k s	7V	0.80	y t a m F l I V
w d h g	8I	0.80	n c r q p e y k s f m t Av LI
	9N	0.80	w c h p f y Mi q r v g e d t a k s l N
c w m f i d	10R	0.80	p v y g s l t n a e q Hk R
c w f m y i v d h p g l	11R	0.80	n s t a e k Q R
c w f m y i v d h p g l	12R	0.80	n s t a e k Q R
y w v t s r q p n m l k i g f e d c a	13H	0.80	H

Figure 7 | Output prediction for all substitutions for a single protein. Shown here are amino acid positions 3–13 of a protein. Capital letters indicate amino acids appearing in the protein sequence alignment used for prediction whereas lower case letters indicate amino acids not observed in the protein alignment. The numbers in the 'Seq Rep' column represent the fraction of sequences in the alignment that have an amino acid at the corresponding position.

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1M 0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2E 0.25	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3M 0.50	0.07	0.02	0.02	0.03	0.12	0.02	0.02	0.24	0.03	1.00	0.63	0.02	0.03	0.03	0.03	0.03	0.05	0.17	0.02	0.04
4A 0.50	1.00	0.05	0.13	0.17	0.04	0.68	0.04	0.05	0.16	0.09	0.04	0.12	0.14	0.10	0.10	0.30	0.15	0.11	0.01	0.05
5C 0.75	0.59	1.00	0.17	0.15	0.09	0.33	0.08	0.11	0.17	0.15	0.06	0.19	0.18	0.12	0.12	0.93	0.44	0.21	0.03	0.11
6R 0.75	0.37	0.04	0.24	0.36	0.06	0.23	0.11	0.10	0.58	0.17	0.06	0.26	0.15	0.29	1.00	0.63	0.33	0.15	0.02	0.09
7V 0.75	0.10	0.03	0.02	0.03	0.22	0.04	0.03	0.99	0.04	0.48	0.10	0.03	0.03	0.03	0.04	0.05	0.09	1.00	0.02	0.09
8I 0.75	0.44	0.07	0.07	0.13	0.21	0.07	0.07	0.86	0.13	1.00	0.21	0.08	0.09	0.10	0.10	0.13	0.24	0.85	0.04	0.12
9N 0.75	0.90	0.15	0.73	0.85	0.41	0.60	0.37	0.54	0.94	1.00	0.36	0.95	0.36	0.59	0.66	0.93	0.84	0.64	0.10	0.46
10R 0.75	0.14	0.02	0.08	0.16	0.07	0.10	0.38	0.06	0.46	0.12	0.04	0.13	0.08	0.19	1.00	0.13	0.13	0.09	0.02	0.11
11R 0.75	0.10	0.01	0.06	0.13	0.02	0.06	0.05	0.03	0.33	0.07	0.02	0.08	0.05	0.46	1.00	0.09	0.08	0.05	0.01	0.03
12R 0.75	0.10	0.01	0.06	0.13	0.02	0.06	0.05	0.03	0.33	0.07	0.02	0.08	0.05	0.46	1.00	0.09	0.08	0.05	0.01	0.03
13H 0.75	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 8 | Output scaled probability matrix for a single protein. Shown here are the results for amino acid positions 3–13 of a protein. The threshold score for intolerance of a substitution is 0.05. Rows correspond to positions in the reference protein and columns correspond to all possible 20 amino acids. The header column on the left contains the position, reference amino acid and the fraction of sequences in the alignment that have an amino acid at the corresponding position. Each element of the matrix represents the scaled probability of substitution at each position. Substitutions predicted to be intolerant are highlighted in red.

(a) *Protein ID*: Protein identifiers entered by the user.
(b) *Substitution*: Amino acid substitutions entered by the user. If substitutions are not specified by the user, then this column displays substitutions from the NCBI dbSNP database for the corresponding protein.
(c) *dbSNP ID*: dbSNP rsID for the substitution entered by the user. If the user-specified substitution does not correspond to a SNP in the dbSNP, then this column is left blank for the corresponding protein.
(d) SIFT prediction, which can be either 'DAMAGING' or 'TOLERATED'. 'DAMAGING' means that the substitution is predicted to affect protein function and 'TOLERATED' means that the substitution is predicted to be functionally neutral.
(e) *Score*: The SIFT prediction score ranges from 0 to 1, and is the scaled probability of an AAS being tolerated. AASs with scores that fall below 0.05 are predicted to affect protein function.
(f) *Median Info*: The conservation value for a position ranges from zero, when all 20 amino acids are observed at a position in a protein sequence alignment, to $\log_2 20 (=4.32)$, when only one amino acid is observed at that position[5]. The 'median info' is the median of conservation values in the alignment at the position of the substitution. If a substitution is predicted to be damaging with a median conservation value greater than 3.25, SIFT warns about the lack of diversity of the chosen sequences. A low-confidence warning can be seen in **Figure 5**. If this warning is displayed, then it is advisable to attempt the analysis again choosing a larger database, such as the NCBI non-redundant database, as discussed in the TROUBLESHOOTING section.
(g) *Number of Sequences at position*: The number of sequences used in the alignment that has an amino acid (not an insertion/deletion) at the corresponding position. A sufficient number of sequences used in the alignment are typically greater than 10. A high number of sequences represented at a position improve the confidence of the prediction at that position.

SIFT Results for single protein tools: There are three kinds of output files that are generated for single protein tools. A summary results page is shown in **Figure 6** and contains the links to the output files, which are described below.

(a) Predictions for all 20 possible amino-acid changes at each position of the protein, as shown in **Figure 7**. Capital letters indicate amino acids appearing in the protein sequence alignment that was used for prediction, whereas lower case letters indicate amino acids not observed in the protein alignment. The numbers in the 'Seq Rep' column correspond to the percentage of sequences in the alignment, which have an amino acid (not an insertion/deletion) at the corresponding position. A higher fraction of sequences in the alignment represented at a position improves the confidence of prediction at that position. In **Figure 7**, it can be observed that (a) position 13 is highly conserved and tolerates only the amino acid H and (b) position 3 tolerates A, F, V, I, M and L, all non-polar hydrophobic amino acids.
(b) A scaled probability matrix for the query protein sequence as depicted in **Figure 8**. The rows of the matrix represent positions of the protein sequence and the columns represent all 20 amino acids. The matrix contains the SIFT scores of all substitutions at each position of the sequence. The scores below 0.05 are predicted to affect protein function.
(c) Predictions for the AASs submitted by the user, as shown in **Figure 9**. The predictions and quantitative scores obtained may help investigators in prioritizing AASs in targeted gene re-sequencing or mutagenesis studies.

SIFT predictions for selected substitutions

```
Substitution at pos 5 from C to W is predicted to AFFECT PROTEIN FUNCTION with a score of 0.03.
    Median sequence conservation: 2.39
    Sequences represented at this position:3

Substitution at pos 554 from Q to P is predicted to be TOLERATED with a score of 0.33.
    Median sequence conservation: 3.39
    Sequences represented at this position:3
```

Figure 9 | Predictions on selected substitutions in a single protein.

ACKNOWLEDGMENTS Development of the SIFT server at the J. Craig Venter Institute was funded by the National Human Genome Research Institute (R01 HG004701-01).

Published online at http://www.natureprotocols.com
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Krawczak, M. *et al.* Human gene mutation database-a biomedical information and research resource. *Hum. Mutat.* **15**, 45–51 (2000).
2. Henikoff, S. & Comai, L. Single-nucleotide mutations for plant functional genomics. *Annu. Rev. Plant. Biol.* **54**, 375–401 (2003).
3. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
4. Tchernitchko, D., Goossens, M. & Wajcman, H. *In silico* prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. *Clin. Chem.* **50**, 1974–8 (2004).
5. Ng, P.C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12**, 436–446 (2002).
6. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
7. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
8. Saunders, C.T. & Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891–901 (2002).
9. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
10. Palmer, J.S. *et al.* Melanocortin-1 receptor polymorphisms and risk of melanoma: Is the association explained solely by pigmentation phenotype? *Am. J. Hum. Genet.* **66**, 176–186 (2000).
11. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genom.* **7**, 61–80 (2006).
12. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
13. Ferrer-Costa, C. *et al.* PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* **21**, 3176–3178 (2005).
14. Stone, E.A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
15. Mathe, E. Computational approaches for predicting the biological effect of p53 missense mutation: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* **34**, 1317–1325 (2006).
16. Healy, E. *et al.* Skin type, melanoma, and melanocortin 1 receptor variants. *J. Invest. Dermatol.* **112**, 512–513 (1999).
17. Liu, W. *et al.* Denaturing HPLC-identified novel FBN1 mutations, polymorphisms, and sequence variants in Marfan syndrome and related connective tissue disorder. *Genet. Test.* **1**, 237–242 (1997/1998).
18. Yue, P.E., Melamud, M. & Moult, J. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7**, 166 (2006).
19. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–41 (2003).