W154–W161 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 30 April 2020*
doi: 10.1093/nar/gkaa288

LIST-S2: taxonomy based sorting of deleterious missense mutations across species

Nawar Malhis [1,*], Matthew Jacobson[1], Steven J. M. Jones[2,3] and Jörg Gsponer[1,4,*]

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [2]Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada, [3]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada and [4]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

Received January 27, 2020; Revised April 05, 2020; Editorial Decision April 14, 2020; Accepted April 19, 2020

ABSTRACT

The separation of deleterious from benign mutations remains a key challenge in the interpretation of genomic data. Computational methods used to sort mutations based on their potential deleteriousness rely largely on conservation measures derived from sequence alignments. Here, we introduce LIST-S2, a successor to our previously developed approach LIST, which aims to exploit local sequence identity and taxonomy distances in quantifying the conservation of human protein sequences. Unlike its predecessor, LIST-S2 is not limited to human sequences but can assess conservation and make predictions for sequences from any organism. Moreover, we provide a web-tool and downloadable software to compute and visualize the deleteriousness of mutations in user-provided sequences. This web-tool contains an HTML interface and a RESTful API to submit and manage sequences as well as a browsable set of precomputed predictions for a large number of UniProtKB protein sequences of common taxa. LIST-S2 is available at: https://list-s2.msl.ubc.ca/

INTRODUCTION

High-throughput sequencing technologies enable the affordable mapping of mutations present in any species' population as well as within individuals affected by disease. Separating the large number of neutral mutations from those that underlie deleterious phenotypes is an important bottleneck to overcome in order to effectively use the growing wealth of available sequencing data. Numerous computational methods have been developed to predict the deleteriousness of mutations in coding regions by quantifying the evolutionary constraints on affected residues, i.e. conservation methods, (e.g. SIFT (1), PROVEAN (2), phyloP (3), GERP++ (4), SiPhy (5), PhastCons (6) and EVmutation (7)) or by combining conservation with features derived from functional genomic and gene annotation data, i.e. ensemble methods, (e.g., PolyPhen-2 (8), CADD (9), Eigen (10), DANN (11) and fitCons (12)). Classically, conservation measures derived from sequence alignments are either based on variant frequencies (13) (e.g. SIFT, PROVEAN, EVmutation) or the phylogenetic relationships among preselected subsets of species (e.g. phyloP, GERP++, SiPhy, PhastCons). Recently, we introduced a new framework for evolutionary conservation with measures that exploit local sequence identity and taxonomy distances across species (14). These measures are based on the assumption that variations observed in homologs from closely related species are more significant in assessing conservation compared to those in distantly related species. We used the new conservation measures to create a method (LIST) for predicting the deleteriousness of human coding mutations which is centred on two features: Local Identity and Shared Taxa. LIST comfortably outperforms methods that rely on existing conservation measures.

Here we introduce LIST-S2, an updated method for predicting the deleteriousness of mutations along with an accompanying web-tool, API and software. Unlike its predecessor (LIST), LIST-S2 is not limited to human mutations and is capable of making predictions for sequences from any species (suffix S2). Results show that LIST-S2 substantially outperforms comparable methods that rely solely on conservation, independent of the dataset used (Tables 1 and 2, Supplementary Tables S1–S4, Supplementary note 1). In addition, while restricted solely to conservation measures, LIST-S2 still outperforms ensemble methods that combine conservation with features derived from functional genomics studies and/or gene annotations (Supplementary Tables S1–S4, Supplementary note 1).

MATERIALS AND METHODS

LIST-S2, similarly to LIST, is assembled hierarchically from three modules. The position mutation module (PMM)

*To whom correspondence should be addressed. Tel: +1 604 822 4838; Fax: +1 604 822 2114; Email: nmalhis@msl.ubc.ca
Correspondence may also be addressed to Jörg Gsponer. Tel: +1 604 827 4754; Fax: +1 604 822 2114; Email: gsponer@msl.ubc.ca

Table 1. Optimization and benchmarking data sets used

		Benign		Deleterious	
Set ID	Protein set	Definition	Count	Definition	Count
OP1	A	ExAC $\geq$ 0.5%	24 096	0.015% $\leq$ ExAC $\leq$ 0.03%	48 142
OP2	A	ExAC $\geq$ 1%	18 109	ClinVar pathogenic annotation & ExAC > 0	2146
BE1	B	ExAC $\geq$ 1%	10 971	ClinVar pathogenic annotation & ExAC > 0	1684
BE2	B	gnomAD $\geq$ 1%	9578	UniProt missense pathogenic annotation	7414
BE3	B	gnomAD $\geq$ 1%	9578	UniProt missense pathogenic cancer annotation	1147
BE4	B	HumVar benign	7552	HumVar disease	7624

Table 2. The AUC values contrasting LIST-S2 performance to LIST and other conservation tools using four benchmarking datasets, BE1, BE2, BE3 and BE4. The highest AUCs achieved are highlighted in bold

Deleterious count	**1684**	**7414**	**1147**	**7624**
Benign count	**10 971**	**9578**	**9578**	**7552**
Datasets	**BE1 (ClinVar & ExAC) vs. ExAC**	**BE2 UniProt vs. gnomAD**	**BE3 Cancer vs. gnomAD**	**BE4 Humvar**
LIST-S2	0.880	0.919	**0.897**	0.895
LIST	**0.885**	**0.922**	0.885	**0.900**
SIFT	0.817	0.883	0.820	0.880
PROVEAN	0.816	0.879	0.821	0.880
phyloP_V	0.814	0.880	0.816	0.856
SiPhy	0.805	0.853	0.790	0.825
SIFT4G	0.798	0.859	0.792	0.857
GERP++_RS	0.774	0.823	0.795	0.793
phastCons_V	0.771	0.825	0.768	0.797
phyloP_M	0.744	0.804	0.772	0.780
phyloP_P	0.716	0.765	0.745	0.737
phastCons_M	0.716	0.792	0.791	0.749
phastCons_P	0.707	0.777	0.799	0.730

estimates deleteriousness of a specific mutation by determining whether an amino acid matching the mutation occurs in a homolog of a species closely or distantly related to the species for which we are making the prediction. The second module, the position module (PM; subdivided in PM1 and PM2i), assesses how vulnerable a sequence position is to variations. Specifically, it determines whether amino acid variations in this position occur in species closely or distantly related to the species of interest. The mutation module (MM) finally assesses the likelihood of changing the reference to the mutant amino acid.

The PMM and PM comprise the core of this predictor, contributing most to its prediction performance. Both modules exploit local sequence identity (LI) and shared taxa (ST), two terms that we define as follows. Given a query sequence aligned to a database sequence, we define the local identity (LI) at position τ as the number of residues in the database sequence that are identical to those of the query in a window of 11 residues centred at τ, excluding the residue at τ. We define the shared taxa (ST) of the database sequence as the number of taxonomy tree edges that are shared between the species from which the database and query sequences originate. In PMM and PM, local identity (LI) and shared taxa (ST) are exploited in the conservation measures, variant shared taxa (VST) and shared taxa profile (STP), that we previously introduced in (14).

The purpose of VST is to have a measure of how close in the taxonomy tree is a species that has a homolog (high LI) with an amino acid at position τ matching the mutation of interest. The closer the species that has this matching amino acid, the higher the number of taxonomy tree edges that the species share (high ST) and thus the higher the VST value for the given amino acid. Specifically, VST_τ is defined as a vector of $N = 20$ such that for each possible amino acid variation (v) at position τ $VST_{\tau,v}$ is the ST of the database sequence with the highest LI and the amino acid v at τ, given the restriction that τ is not located within two residues of indels. If the highest LI is shared by several database sequences, we select the ST of the sequence with the highest segment identity (SI), where SI is the number of residues that are identical between the query and the continually aligned segment of the database sequence that harbours amino acid variation (v) at position τ in the blastp output. If multiple database sequences share the same highest SI (as well as highest LI), then we select the highest ST. Previous analysis has shown that mutations with higher $VST_{\tau,v}$ values (i.e. observed in sequences from closely related species) are less likely to be deleterious when compared to those with lower $VST_{\tau,v}$ values, and that the average $VST_{\tau,v}$ for all 19 possible variations is higher for benign positions (14). Figure 1A,C provides two examples that illustrate this difference of $VST_{\tau,v}$ for benign and deleterious mutations. The mutation of histidine (H) to arginine (R) at position 317 in the protein MEGF11 (UniProtKB: A6BM72) is frequent (benign). Accordingly, the VST value for R at this position is high (Figure 1A), i.e. R exists in closely related species. By contrast, the mutation of lysine (K) to glutamine (Q) at position 666 in protein SF3B1 (UniProtKB: O75533) has been shown to be deleterious. The VST value for glutamine at this position is low (Figure 1C), indicating that glutamine is only found at this position in species that share few taxonomy tree edges with the query's species, thus are far away in the taxonomy tree.

The purpose of STP is to measure whether sequences across the ST spectrum with an amino acid at position τ that differs from the one present in the query have 'strong' homologs (high LI) or only 'weak' homologs (low LI) to the query sequence. Query sequence positions where variations are observed in strong homologs (higher LI) from closely related species (higher ST) are likely to have higher tolerance to mutations compared to other locations. Specifically, we define STP_τ as a vector of size ST_max such that for each possible ST value s, $0 < s \leq$ ST_max, $STP_{\tau,s}$ is the highest LI of the set of database sequences with ST = s that do

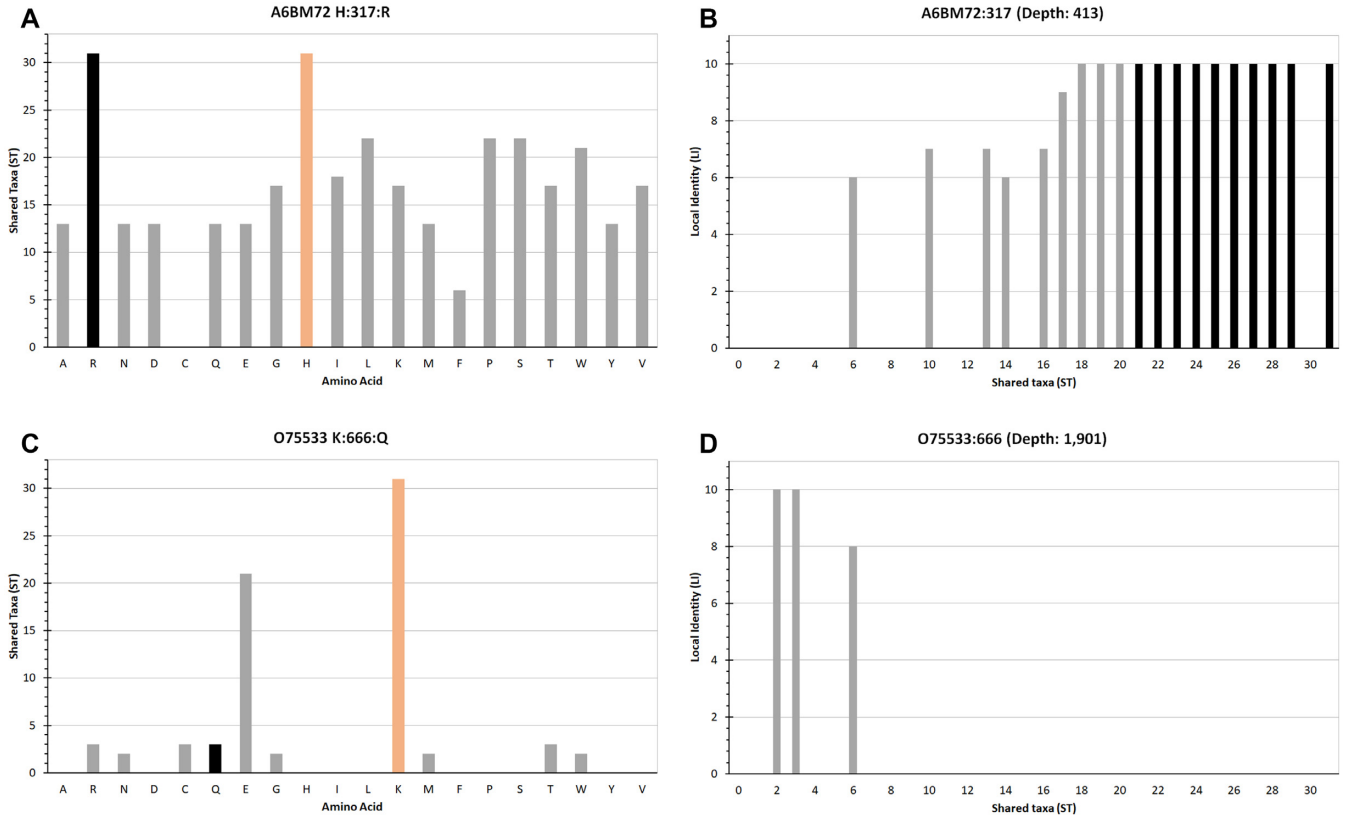

Figure 1. Contrast in VST and STP vectors for benign and deleterious mutations. Shown are the VST (**A**) and STP vectors (**B**) for the frequent (benign) human mutation A6BM72 H:317:R, as well as the VST (**C**) and STP vectors (**D**) for the human deleterious mutation O75533 K:666:Q, which has been linked to malignant melanoma of the skin (CMM). In (A) and (C), the ST value of the mutant residue ($VST_{\tau,v=m}$) are shown in black and that of the reference residue in orange. In (B) and (C), LI values at ST higher than two thirds the ST_max are in black.

not have an amino acid matching that of the query at τ. ST_max is defined as the query's species taxonomy lineage size. Our previous analysis has revealed that benign mutations have higher average STP values for higher shared taxa when compared to deleterious mutations. Figures 1B,D illustrate this observation for the two mutated positions in proteins MEGF11 and SF3B1. The position of the benign mutation in protein MEGF11 (Figures 1B) has high LI values for high STs, which shows that this position is variable in sequences of species close to the query's species (high STs). The opposite is observed for the position with a deleterious mutation in Figures 1D.

VST_τ and STP_τ are used in the position mutation module (PMM) and the position module (PM) to calculate mutation scores as follows:

The PMM module scores mutations m at τ based on the formula:

$$PMM_{\tau,m} = \begin{cases} 1 - \frac{VST_{\tau,v=m}}{ST\_max}, & LI \geq \alpha \\ 1, & LI < \alpha \end{cases} \quad (1)$$

The minimum local identity cut-off (α) used in LIST-S2 for PMM is 4, which maximizes PMM's AUC value on the OP1 dataset (see below). Higher $VST_{\tau,v=m}$, usually observed for benign mutations (Figure 1A), result in lower $PMM_{\tau,m}$ deleteriousness scores when compared to deleterious mutations that generally have low $VST_{\tau,v=m}$ values (Figure 1C).

PM module scores are assembled from the two subcomponents PM1 and PM2i. PM1 is derived from the VST_τ vector by averaging the $PMM_{\tau,m}$ values for all possible mutations m at τ:

$$PM1_\tau = \frac{\sum_{m \neq ref}^{20} PMM_{\tau,m}}{19} \quad (2)$$

PM2i is derived from the average $STP_{\tau,s}$:

$$PM2i_\tau = 1 - \frac{3 * \sum_{s=s1}^{ST\_max} L_{\tau,s}}{ST\_max} \quad (3)$$

Where s1 = ST_max * 2/3, is learned using random search to maximize the AUC of PM2i on OP1, and:

$$L_{\tau,s} = \begin{cases} STP_{\tau,s}, & STP_{\tau,s} \geq \beta \\ 0, & STP_{\tau,s} < \beta \end{cases} \quad (4)$$

The minimum local identity ($STP_{\tau,s}$) cut-off (β) used in LIST-S2 for PM2i is 6. This cut-off maximizes PM2i's AUC value on the OP1 dataset.

The PMM and PM modules are complemented by the mutation module (MM) to assess the general amino acid swap-ability between the reference (r) and the mutant (m) residue and is calculated as:

$$MM_m = \frac{1}{AASM_{r,m}} \quad (5)$$

Where the amino acid swap-ability matrix is defined as:

$$\mathrm{AASM}_{r,m} = \frac{\mathrm{FP}_{r,m}}{(\mathrm{FP}_{r,m} + \mathrm{RP}_{r,m})} \quad (6)$$

The rare probability matrix (RP) and the frequent probability matrix (FP) are defined as:

$$\mathrm{RP}_{r,m} = \frac{\mathrm{RC}_{r,m}}{\sum_{m_j, m_j \neq r}^{19} \mathrm{RC}_{r,m_j}} \quad (7)$$

$$\mathrm{FP}_{r,m} = \frac{\mathrm{FC}_{r,m}}{\sum_{m_j, m_j \neq r}^{19} \mathrm{FC}_{r,m_j}} \quad (8)$$

where $\mathrm{RC}_{r,m}$ ($\mathrm{FC}_{r,m}$) is the count of rare (frequent) mutations in the OP1 dataset (see below).

From the above we see that LIST-S2′ scores are derived from four features: the VST of the mutant amino acid used by the PMM module, the average VST of all 19 possible variations at the mutation position used by PM1, the average values of the top one third of the STP vector used by PM2i, and the general amino acid swap-ability matrix AASM used by the MM module. Bayes rule is used to incorporate weighted scores of PM1 and PM2i into a PM score, and then weighted scores of PM and PMM into a CORE score (Figure 2), and weighted CORE scores are combined with weighted MM scores into the final LIST-S2 scores as explained in Figure 2, and supplementary notes S2 and S3. Supplementary Figure S1a,b shows the discriminative power of each of the LIST-S2 sub-modules, as well as the CORE, in AUC values.

Previous analysis revealed that the scores of each of the three modules used in LIST need to be rescaled to compensate for the alignment depth before hierarchical integration (14). In supplementary note S2, we describe the details of the optimization of LIST-S2′ hierarchical structure, including the steps taken to compensate for alignment depth and to determine Bayes rule input weights. Moreover, scores have been fitted to target distributions during this integration and for the final output, which we describe in detail in supplementary note S3.

Data sets

We used two sets of human allele frequencies: the Exome Aggregation Consortium (15) (ExAC) based on 60,706 individuals and the Genome Aggregation Database v2.1.1 (gnomAD) based on 118 479 cancer-free individuals (http://gnomad.broadinstitute.org/). We also used three sets of deleterious mutations: UniProt missense mutations that have been associated with diseases (UPPath, N = 19,744), UniProt mutations that have been associated with cancer (CANCER, $N = 2647$), and ClinVar germline missense mutations that are also observed at least once in the ExAC data (CEPath, $N = 4070$). We also used CD-HIT (16) to divide the SwissProt human protein sequences (release 2017_07) at random into two equal sets, such that sequences in A have <50% identity to those in B. Mutations that map to proteins in set A are used for optimization while those that map to proteins in set B are used for benchmarking.

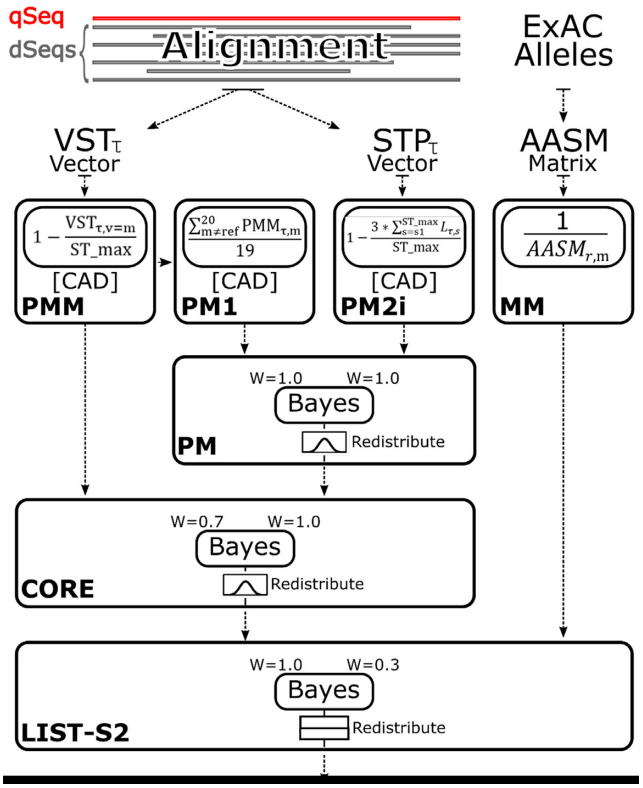

Figure 2. Architecture of LIST-S2. For each position τ in the query sequence, a variant shared taxa VST_τ vector and a shared taxa profile STP_τ vector are constructed. The scores of PMM and PM1 are derived from VST_τ, and the PM2i score is derived from STP_τ. The three scores that are derived from sequence alignments (PMM, PM1 and PM2i) are rescaled to compensate for alignment depth [CAD]. The PM score is assembled from the weighted scores of PM1 (weight 1.0) and PM2i (weight 1.0) using Bayes rule and then redistributed to fit a normal distribution. The CORE score is computed by joining the weighted scores of PMM (weight 0.7) with PM (weight 1.0) and then redistributing the outcome to fit a normal distribution. The amino acid swap-ability matrix (AASM) is derived from ExAC mutations. MM scores are the inverse of the AASM values for the reference (r) and the mutant (m) amino acids. The final LIST-S2 score is computed by joining the weighted scores of CORE (weight 1.0) with MM (weight 0.3) and then redistributing it to fit a uniform distribution. For a detailed description of compensating for alignment depth [CAD] and score weights see supplementary note S2, and for redistributing scores to fit a target distribution see supplementary note S3.

Optimization data. Two datasets that map to proteins in set A were used for optimization. In the OP1 dataset, mutations in ExAC with allele frequency in the range of 0.015% to 0.03% were considered deleterious (positive class), counting 48 142 and mutations with allele frequency ≥0.5% were considered benign (negative class), counting 24 096. The OP2 dataset includes the CEPath (see above) deleterious mutations, counting 2146, and frequent ExAC mutations (frequency ≥ 1%) as benign class, counting 18 109.

Benchmarking data. Mutations that map to proteins in set B and have precomputed scores for 24 predictors in dbNSFP4.0a (17) (namely: SIFT (1), SIFT4G (18), PROVEAN (2), phyloP (3) X 3, SiPhy (5), phastCons (6) X 3, GERP++ (4), Eigen (10), CADD (9), DANN (11),

PolyPhen-2 (8), FATHMM-MKL (19), PrimateAI (20), MutationTast (21), MutationAssessor (22), MPC (23) and fitCons (12) X 4) are used for benchmarking. Four benchmarking data sets were used: BE1 includes frequent ExAC mutations (frequency ≥ 1%) that are not in CEPath, counting 10 971, as the benign negative class and CEPath mutations, counting 1684, define the deleterious positive class. In BE2, benign mutations are those with an allele frequency ≥1% in gnomAD that are not part of UPPath (see above), counting 9578, and deleterious mutations are those that are part of UPPath, counting 7414. In BE3, benign mutations are those with allele frequency ≥1% in gnomAD that are not part of UPPath, counting 9578, while deleterious are those defined in the CANCER set (see above), counting 1147. Finally, BE4 is the subset of HumVar mutations provided by PolyPhen-2, counting 7624 deleterious and 7552 benign mutation, where the deleterious class includes all mutations associated with diseases or a loss of activity/function, excluding those associated with cancer, and the benign class includes those mutations that are frequent (allele frequency ≥ 1%) (Table 1). The precomputed ranked scores of all tools used in the benchmarking, except those from EVmutations and LIST, were collected from dbNSFP4.0a. Precomputed LIST scores were collected from (24), precomputed LIST-S2 scores used can be found at (25), and EVmutation scores were collected from https://marks.hms.harvard.edu/evmutation/human_proteins.html. Table 1 summarizes the optimization and benchmarking datasets used.

RESULTS AND DISCUSSION

Benchmarking

Using the four benchmarking datasets described above (BE1-4), we contrasted the performance of LIST-S2 to that of LIST and other conservation methods. Our analysis (Table 2, Supplementary Tables S1–S5) reveals that LIST-S2 performs similar to LIST and substantially better than leading alternatives. Our analysis using the BE1 dataset indicates that, compared to other conservation methods, LIST-S2 provides a higher sensitivity for any specificity value (Figure 3A) and a higher precision for any recall (Figure 3B). In addition, LIST-S2 outperforms ensemble tools (Figure 3C, Supplementary Tables S1–S4). EVmutation and LRT have specific alignment requirements and, thus, score considerably lower numbers of mutations. LIST-S2 also achieves higher AUCs for the subset of mutations scored by EVmutation and LRT (Supplementary Table S1–S4). LIST-S2 also does better than all other tested methods when evaluating mutations in protein parts predicted to be disordered (IDRs) by ESpritz (26) or IUPred (27) (Supplementary Tables S1–S4).

Interpreting LIST-S2 scores

LIST-S2 scores are redistributed to fit a uniform distribution (learned from OP2), i.e. approximately rank scores. These scores reflect the deleteriousness of mutations in tissues where the mutated proteins are transcribed.

Understanding germline mutations. Germline mutations exists in all cells, however, the deleteriousness effect of germline deleterious mutations is limited to the tissue(s) that transcribe the proteins harbouring these mutations and are involved in the pathways related to the deleteriousness of these mutations. Consequently, to reduce the number of candidate mutations, studies identifying the underlying genetic causes of specific disorders often prioritize genes linked to these disorders. For instance, to identify mutations linked to autism spectrum disorder (ASD), the authors of (28) prioritized mutations observed in genes known to have links to neurodevelopmental conditions. Thus, LIST-S2 scores should be used and interpreted in a specific context to optimize their usefulness.

Understanding cancer somatic mutations. In most cancer applications, we start with a set of somatic mutations observed in a tumour, and we need to identify driver mutations that propel the proliferation of cancer from those that have no effect on the cancer proliferation also known as passenger mutations. We note that the lack of effect on cancer proliferation for a large fraction of passenger mutations is likely the result of being in genes that are not transcribed or have no role in the cancer cell and therefore are not subject to any selective constraint. Studies show that closed chromatin marks are found to be associated with regions of high cancer somatic mutation density compared to regions with high gene activity and open chromatin (29). As a result, many passenger mutations that are in closed chromatin regions have the characteristics of deleterious mutations that would likely generate a deleterious effect in normal tissues and thus should be scored by LIST-S2 and other tools as deleterious overshadowing or obfuscating driver variations that are actually contributing to the growth of tumours. Compared to conservation and ensemble methods, LIST-S2 achieves substantially higher performance in separating variations associated with cancer from those that are frequent (benign) (Figure 3D, Supplementary Table S3). However, since passenger mutations are not necessarily benign, the search for drivers needs to be limited to genes known to be associated with oncogenesis. Thus, LIST-S2 can be used to predict the deleteriousness of mutations in cancer driver genes (e.g. the 299 identified by (30)) that are also expressed in the cancer tissue, since mutations with higher LIST-S2 scores are more likely to alter the functionality of these driver genes which make them candidate driver mutations.

SERVER DESCRIPTION

Input

Protein sequences with identifier headers in fasta format. Optional OX = n can be included in the header to explicitly identify the sequence Taxa ID as n, otherwise, the Taxa ID of the sequence with the highest alignment (Bits score) will be used as the query Taxa ID.

Output

LIST-S2 computes the potential deleteriousness of every possible amino acid mutation of the input protein sequence. Scores reflect the deleteriousness of each possible mutation

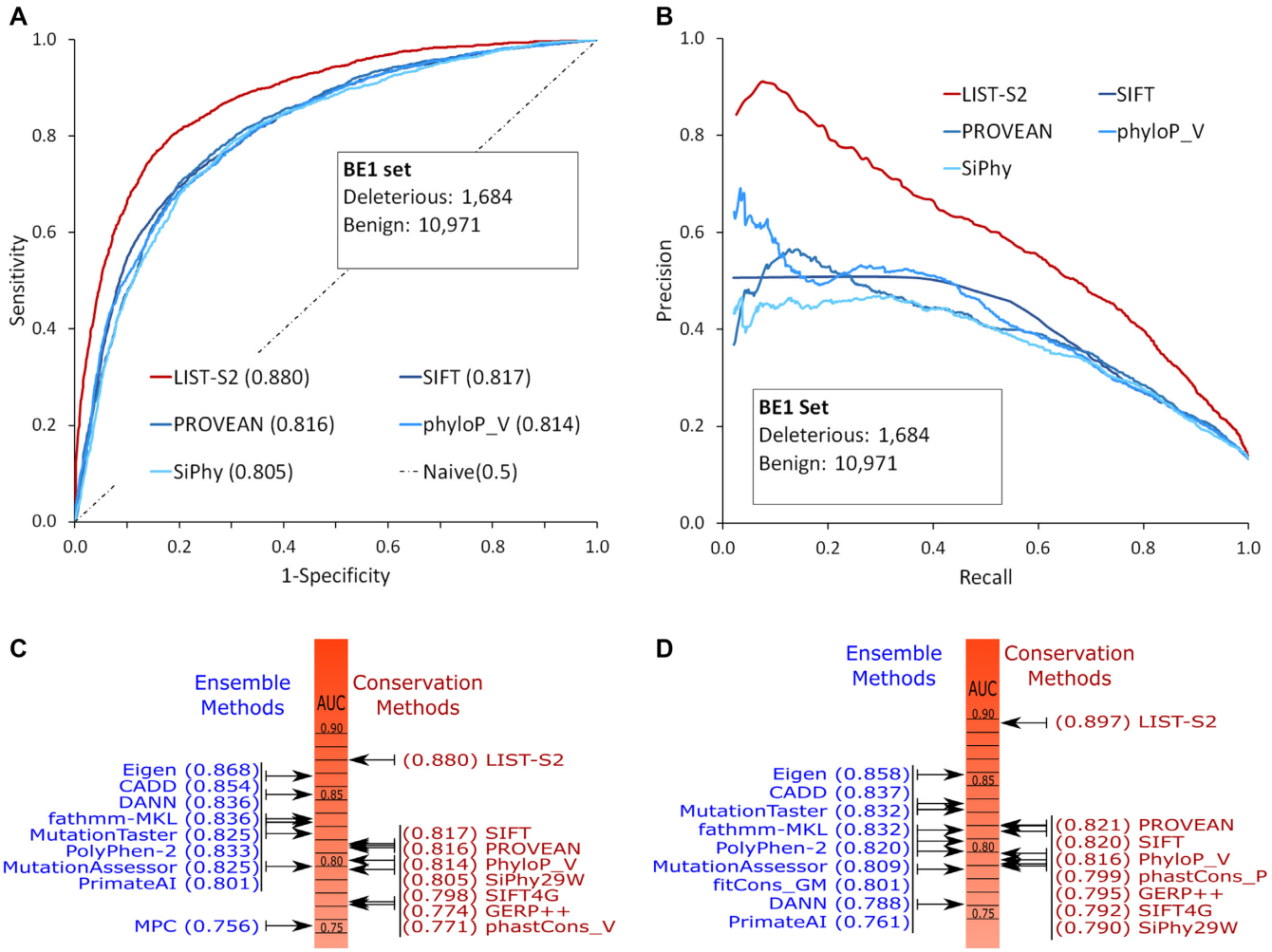

Figure 3. Contrasting the performance of LIST-S2 to other methods in separating human germline deleterious. (**A**) The ROC curves contrasting LIST-S2 to the four best performing conservation methods using the BE1 dataset: SIFT, PROVEAN, PhyloP_V (based on 100 Vertebrata species) and SiPhy. AUC values are provided for each method in parentheses. (**B**) The precision recall curves contrasting LIST-S2 to the same four methods. (**C**) Comparison of AUC values for LIST-S2 and common conservation as well as ensemble methods. (**D**) The AUC values in separating Cancer from frequent mutations (BE3 dataset) for LIST-S2 and common conservation as well as ensemble methods.

at each protein position, where higher scores imply higher potential deleteriousness. We also provide the alignment depth and average deleteriousness of all mutations at each position as a general conservation score. Results can either be visualized in the form of a heatmap (Figure 4) or downloaded as a text file.

Usage example

Figure 4 shows the heatmap output for the Putative POU domain, class 5, transcription factor 1B protein (Q06416).

The LIST-S2 web-tool is comprised of four main components

The processing queue along with its two user-facing clients (HTML interface and RESTful API) and a browsable set of precomputed predictions (Figure 5). Each client acts as a lightweight interface to the processing queue, which is the component actively processing user-provided sequences. This microservice-like design allows flexibility and robustness in server distribution and scaling.

Processing queue

The processing queue accepts 'job requests' from a client to process user-provided sequences and is only accessible via the provided HTML interface and API. In order to prevent any one client or user from overwhelming the LIST-S2 instances it uses a prioritized, tiered system of queues, one for each combination of client and user. The processing queue then distributes the incoming jobs to be processed by LIST-S2.

HTML interface

One of two user-facing clients of the processing queue. This client (https://list-s2.msl.ubc.ca/) provides a web-interface for submitting/managing jobs to be processed by LIST-S2

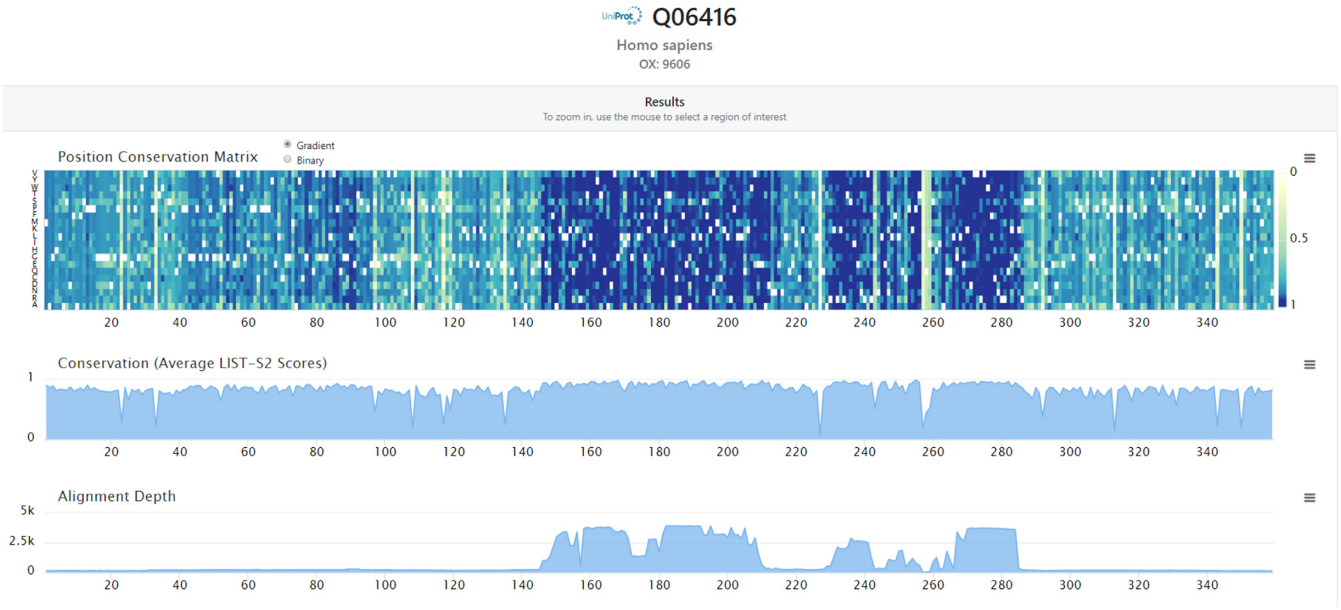

Figure 4. Example of a prediction score visualization. Top, a heatmap matrix for the potential deleteriousness of every possible amino acid mutation. Middle, position average conservation of all possible mutations at that position. Bottom, alignment depth at each position.

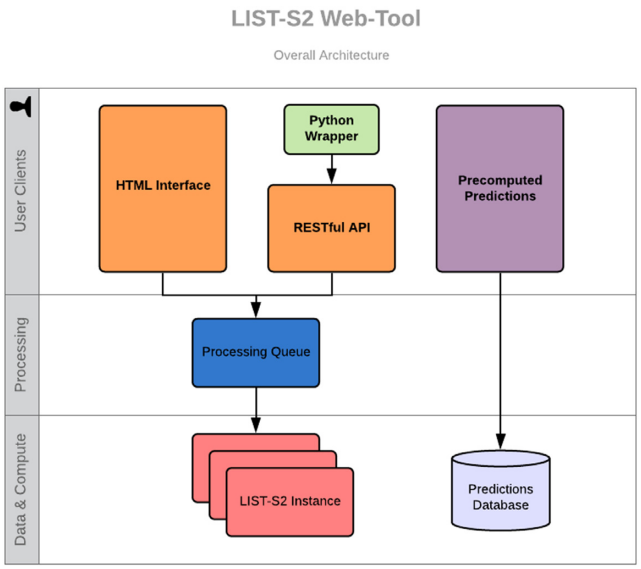

Figure 5. LIST-S2 Web-tool architecture. User-provided sequences can be submitted through either the HTML or RESTful interface. These jobs can then be viewed and managed by interacting with either the provided website or through API calls (optional python wrapper also provided). Jobs submitted in this way are queued and prioritized by the Processing Queue before being distributed to a LIST-S2 instance for results to be computed. Alternatively, precomputed results can be viewed through the Precomputed Predictions interface which houses previously computed results on UniprotKB protein sequences.

as well as interactive visualizations (Figures 4 and 5) of the resulting scores.

RESTful API

The second of the two user-facing clients of the processing queue. This client (https://list-s2-api.msl.ubc.ca/) provides a set of REST endpoints to programmatically submit, retrieve and delete jobs from the processing queue. We provide a python wrapper to help simplify the use of the API from both command-line and scripts.

Precomputed predictions

An interface (https://precomputed.list-s2.msl.ubc.ca/) for browsing and visualizing precomputed predictions of a large number of UniprotKB protein sequences of common taxa. Currently contains predictions for over 700 000 sequence.

FINAL REMARKS

The LIST-S2 web server introduced here enables users to compute and visualize the deleteriousness of mutations in protein sequences of any organism. We validated its performance on human mutations but future studies are required to validate performance on mutations in other species for which benchmarking data is currently sparse or not available. Moreover, we provide the downloadable software together with the associate preformatted UniProt TrEMBL/SwissProt database release 2019_02 so that our measures and prediction scores can more easily be combined with orthogonal data to improve predictions further.

DATA AVAILABILITY

We provide a web-tool (https://list-s2.msl.ubc.ca/) which includes an HTML interface to compute and visualize the deleteriousness of mutations in user-provided sequences and a RESTful API to submit and manage sequences as well as a browsable set of precomputed predictions for a large number of UniprotKB protein sequences of common taxa. We also provided a downloadable software (https://github.com/NawarMalhis/LIST-S2).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank prof. Paul Pavlidis for his comments on this manuscript**.**

FUNDING

Canadian Institutes of Health Research (CIHR); Natural Sciences and Engineering Research Council of Canada (NSERC); Genome Canada and Genome BC [175REG to J.G.]; Michael Smith Foundation for Health Research (MSFHR) [Grant ID 7081 to J.G.]. Funding for open access charge: CIHR; NSERC; Genome Canada and Genome BC [175REG to J.G.]; MSFHR [Grant ID 7081 to J.G.].
Conflict of interest statement. None declared.

REFERENCES

1. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
2. Choi,Y., Sims,G.E., Murphy,S., Miller,J.R. and Chan,A.P. (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, **7**, e46688.
3. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
4. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
5. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–62.
6. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
7. Hopf,T.A., Ingraham,J.B., Poelwijk,F.J., Scharfe,C.P., Springer,M., Sander,C. and Marks,D.S. (2017) Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, **35**, 128–135.
8. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0720s76.
9. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
10. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
11. Quang,D., Chen,Y. and Xie,X. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.
12. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
13. Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
14. Malhis,N., Jones,S.J.M. and Gsponer,J. (2019) Improved measures for evolutionary conservation that exploit taxonomy distances. *Nat. Commun.*, **10**, 1556.
15. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
16. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
17. Liu,X., Jian,X. and Boerwinkle,E. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.
18. Vaser,R., Adusumalli,S., Leng,S.N., Sikic,M. and Ng,P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.
19. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
20. Sundaram,L., Gao,H., Padigepati,S.R., McRae,J.F., Li,Y., Kosmicki,J.A., Fritzilas,N., Hakenberg,J., Dutta,A., Shon,J. *et al.* (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.*, **50**, 1161–1170.
21. Schwarz,J.M., Rodelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
22. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
23. Samocha,K.E., Kosmicki,J.A., Karczewski,K.J., O'Donnell-Luria,A.H., Pierce-Hoffman,E., MacArthur,D.G., Neale,B.M. and Daly,M.J. (2017) Regional missense constraint improves variant deleteriousness prediction. bioRxiv doi: https://doi.org/10.1101/148353, 12 June 2017, preprint: not peer reviewed.
24. Malhis,N. (2018) LIST: Deleteriousness levels of amino acid variations. UBC Research Data doi: http://dx.doi.org/10.14288/1.0373460, 31 August 2018, preprint: not peer reviewed.
25. Malhis,N. (2019) LIST-S2: pre-computed deleteriousness of all possible mutations in human (OX = 9606) protein sequences. UBC Research Data doi: http://dx.doi.org/10.14288/1.0379831, 14 July 2019, preprint: not peer reviewed.
26. Walsh,I., Martin,A.J., Di Domenico,T. and Tosatto,S.C. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
27. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
28. Callaghan,D.B., Rogic,S., Tan,P.P.C., Calli,K., Qiao,Y., Baldwin,R., Jacobson,M., Belmadani,M., Holmes,N., Yu,C. *et al.* (2019) Whole genome sequencing and variant discovery in the ASPIRE autism spectrum disorder cohort. *Clin. Genet.*, **96**, 199–206.
29. Polak,P., Karlic,R., Koren,A., Thurman,R., Sandstrom,R., Lawrence,M., Reynolds,A., Rynes,E., Vlahovicek,K., Stamatoyannopoulos,J.A. *et al.* (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, **518**, 360–364.
30. Bailey,M.H., Tokheim,C., Porta-Pardo,E., Sengupta,S., Bertrand,D., Weerasinghe,A., Colaprico,A., Wendl,M.C., Kim,J., Reardon,B. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.