MutationTaster2 -
mutation prediction for the deep sequencing age

Supplement

Jana Marie Schwarz[1,2], David N. Cooper[3], Markus Schuelke[1,2] & Dominik Seelow[1,2]

[1] Department of Neuropediatrics, Charité – Universitätsmedizin Berlin, Berlin, Germany; [2] NeuroCure Clinical Research Center, Charité – Universitätsmedizin Berlin, Berlin, Germany; [3] Institute of Medical Genetics, Cardiff University, Cardiff, UK.

Correspondence should be addressed to D.S. (dominik.seelow@charite.de)

Nature Methods: doi:10.1038/nmeth.2890

Supplementary Tables

Supplementary Table 1. Polymorphisms and disease mutations used to train and cross-validate the classification models.

Model	Polymorphisms [n]	Disease mutations [n]
without_aae	6,807,269	122,238
simple_aae	20,967	151,542
complex_aae	2,340	123,213

Note that owing to the existence of multiple transcripts, several predictions for the same alterations may occur.

Supplementary Table 2. Results of 5-fold cross-validation of all three classification models.

Model	n	NPV negative prediction value	PPV positive prediction value	Sensitivity	Specificity	Accuracy
without_aae	4,000	0.957 [0.003]	0.888 [0.006]	0.954 [0.004]	0.895 [0.005]	0.922 [0.004]
simple_aae	4,000	0.877 [0.008]	0.895 [0.005]	0.879 [0.007]	0.893 [0.004]	0.886 [0.004]
complex_aae	400	0.869 [0.032]	0.944 [0.004]	0.879 [0.026]	0.939 [0.005]	0.907 [0.017]

This table depicts the results of the cross-validation of the three different models of MutationTaster2 (standard deviation in brackets), either for non-coding/synonymous variants (*without_aae*); non-synonymous variants (*simple_aae*), and for variants leading to greater changes in the amino acid sequence (*complex_aae*). Please note that the actual performance of MutationTaster2 is even better because confirmed polymorphisms and known disease mutations are automatically detected and categorized.

Supplementary Table 3. Application of MutationTaster2 to a real exome

	Homozygous non-synonymous, synonymous, and non-coding variants				**Only homozygous variants leading to single amino acid substitutions**			
Case	**MutationTaster2**	**PPH**	**SIFT**	**PROVEAN**	**MutationTaster2**	**PPH**	**SIFT**	**PROVEAN**
	all predictions				all predictions			
FP	103	464	353	462	6	376	295	331
TN	7,714	943	6,623	6,436	2,771	776	2,482	2,446
FPR	1.3%	33.0%	5.1%	6.7%	0.2%	32.6%	10.6%	11.9%
	variants analyzed by all tools				variants analyzed by all tools			
FP	7	401	286	308	6	376	274	290
TN	1,224	830	945	923	1,146	776	878	862
FPR	0.6%	32.6%	23.2%	25.0%	0.5%	32.6%	23.8%	25.2%

FP, false positives; TN, true negatives; FPR, false positive rate = FP / (FP + TN)

All predictions that assumed pathogenicity (including "possibly damaging" and "probably damaging") were counted as false positives – see 'Application to a real exome' in the Supplementary Methods below; more details are provided on our website (http://www.mutationtaster.org/info/statistics.html).

Supplementary Methods

Additional and improved tests

Regulatory features

We have integrated regulatory features from Ensembl[1]. These include data from the ENCODE project[2] and JASPAR[3], such as DNaseI hypersensitive-, histone modification- and transcription factor-binding sites, and 'promoter-associated' regulatory features. Whenever a variant is located within a regulatory feature, the potential malfunction of the regulatory element is included in the prediction.

phyloP/phastCons scores

phyloP[4] and phastCons[5] are methods designed to determine the degree of evolutionary conservation of nucleotides in the genome. For their integration into MutationTaster2, we used precomputed scores based on the multiple alignment of 46 vertebrate species[6].

Integration of data from 1000G, NCBI ClinVar and HGMD Public

We have integrated all publicly available SNPs and InDels from the 1000G, known disease-variants from ClinVar and HGMD Public. MutationTaster2 provides information about genotypes found in the 1000G, and disease alleles and clinical phenotypes present in ClinVar. Alterations found more than four times in the homozygous state in the 1000G or in HapMap[7], are automatically regarded as neutral. Variants marked as *probable-pathogenic* or *pathogenic* in ClinVar are automatically predicted to be disease-causing and the disease phenotype is displayed. The data from HGMD Public include rather more disease mutations but only their position and HGMD ID are provided, not the corresponding alleles. Hence, we cannot use these data for automatic predictions, but instead display a hyperlink to the mutation in HGMD Public, where users can look up alleles and the respective disease phenotypes.

Splicing

In the first release of MutationTaster, all changes of mRNA splicing predicted by NNSplice[8], were taken into account by the Bayes classifier. To reduce the number of false positive predictions, we refined our model: MutationTaster2 now determines the position of a predicted splice site change relative to annotated intron-exon borders and only the loss or decreased strength of a splice site at an intron-exon border is considered. New splice sites in the variant sequence are only included if their confidence score exceeds 0.3. For increased likelihoods of cryptic splice sites also found in the wild-type sequence, we additionally require an increase of at least 10%. MutationTaster2 also detects sequence changes within 2 bp of an intron-exon junction and considers these as constituting the loss of a splice site.

Analysis of variants spanning intron-exon borders

MutationTaster2 can now handle such variants that span intron-exon borders. These are very likely to perturb normal splicing and hence have considerable pathogenic potential.

Improvements in speed

We have substantially increased the speed of MutationTaster2 by caching BLASTP results from protein conservation analysis and by implementing our own function to search for amino acid substitutions or InDels in the altered amino acid sequence instead of using the Needleman-Wunsch algorithm[9]. Moreover, the alignment employed for conservation analysis at the nucleotide level is now optional. As before, external software used by MutationTaster2 and input/output files are stored on a RAM disk for further performance gains. A single analysis now takes less than 0.10 seconds on average.

New services

Single queries by chromosomal position

The old version of MutationTaster required the specification of a transcript and transcript-based positions for each query. Although this proved useful for Sanger sequencing, NGS variants usually refer to chromosomal positions. We therefore developed an additional web interface for single queries using chromosomal positions. MutationTaster2 automatically maps the variant to all suitable genes and transcripts at the respective position and analyzes the variant in all of them. The Results page features a table summarizing the predictions for all transcripts and detailed results for each transcript.

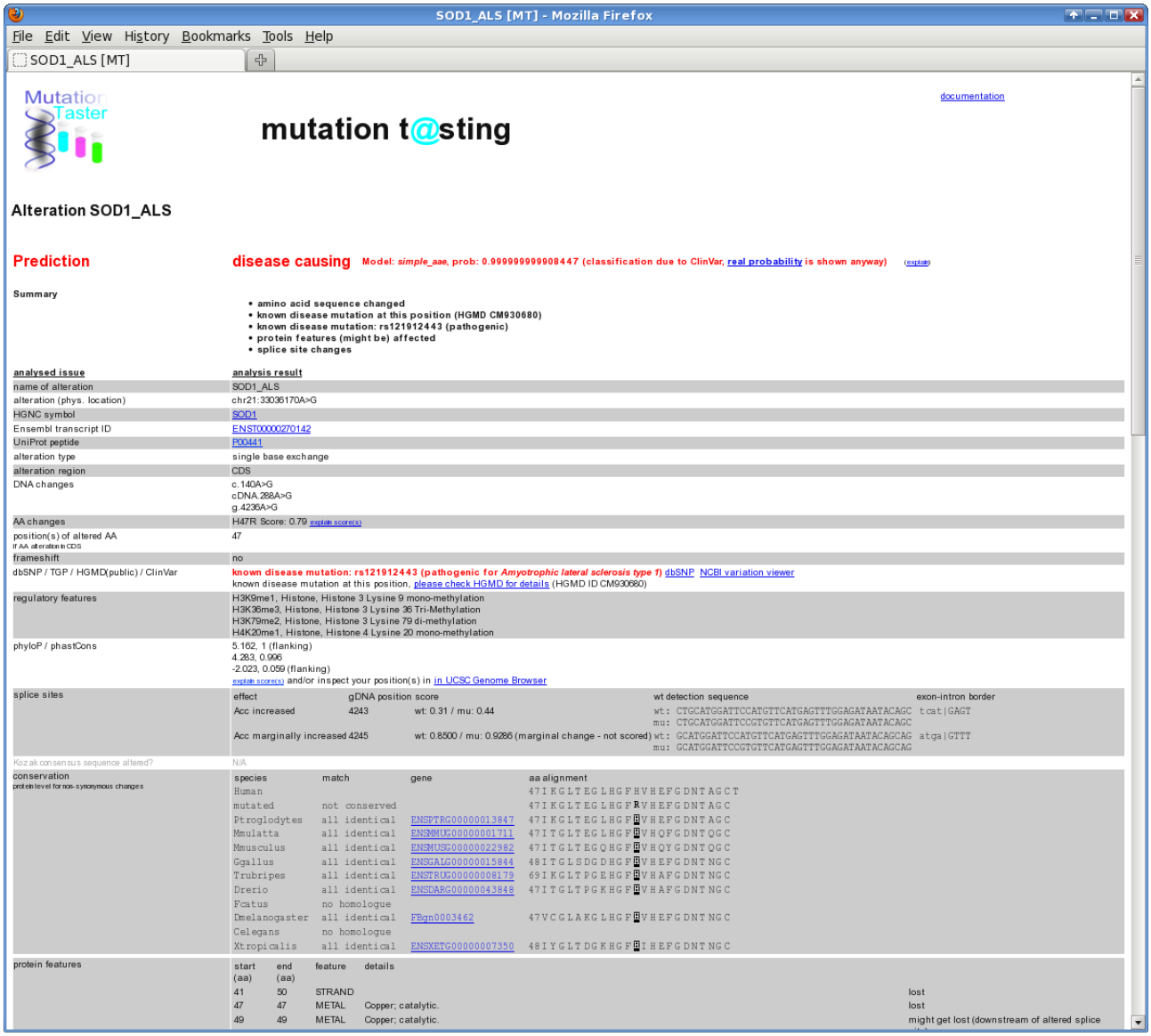
Automatic processing of VCF files

For the rapid and user-friendly analysis of NGS results with MutationTaster2, we created a dedicated query engine. Users can now upload entire VCF files – which is the consensus output format of most NGS pipelines. They can adjust several parameters, such as confining consideration to homozygous variants or certain regions, as well as filtering for known polymorphisms. Once the VCF file has been uploaded, job scheduler software processes the queries in a highly parallel fashion (500,000 alterations/hour). If so desired, users can be notified by e-mail when the process is complete. The results can also be filtered, prioritized and inspected in a web browser or downloaded. We integrated our candidate gene search engine, GeneDistiller[10], to let users determine the most likely candidate genes among the potentially deleterious variants.

Refined results page

We reformatted the MutationTaster2 results page, which now lists the most relevant information on top. We also provide the mutation nomenclature according to the recommendations of the Human Genome Variation Society[11]. The results of splice site analysis are now presented in a more intuitive format (**Figure A**).

Figure A. MutationTaster2 results



The results page of MutationTaster2 for a pathogenic *SOD1* mutation that causes *Amyotrophic lateral sclerosis* type 1 (OMIM #105400). At the top, the prediction 'disease causing' is displayed, together with the advice that it is an automatic prediction because the variant is a known disease mutation. Below, gene-specific information and the results of the different tests performed by MutationTaster2 are presented.

This mutation is one of the examples on our website (http://www.mutationtaster.org/examples/example3.html).

Training of the classifier

MutationTaster2 uses a Bayes classifier to generate predictions. Because alterations with different effects on the protein sequence require different tests, we employ three different classification models, designed for alterations that (1) lead to single amino acid substitutions (*simple_aae),* (2) involve more than one amino acid *(complex_aae*), or (3) are non-coding/synonymous (*without_aae)*.

Data sets for training and validation

The Bayes classifier was trained and tested with single base exchanges and short InDels, comprising >6,000,000 validated polymorphisms from the 1000G and, with kind permission from BIOBASE GmbH, >100,000 known disease mutations from HGMD Professional.

Variants were only included when the homozygous genotype was found in at least 20 individuals in the 1000G (polymorphisms) dataset or when they were listed as a 'disease mutation' in HGMD Professional (disease mutations). Variants that appeared in both groups were discarded. The composition of the data sets is shown in **Supplementary Table 1**. More detailed data along with the models we used can be found on our website: http://www.mutationtaster.org/info/statistics.html.

Cross-validation

Although the Bayes classifier does not remember the combination of test results for single cases and is therefore less prone to overfitting than other approaches, we performed a 5-fold cross-validation on each of our three models (**Supplementary Table 2**). For this purpose, we randomly extracted 2,000 polymorphisms and 2,000 disease mutations from the respective data set of each model and trained the classifier with the remaining variants. Variants that could be mapped to more than one transcript were used only once. To remove the prior caused by different frequencies of disease mutations or polymorphisms, variants of the less frequent type were used several times during training to obtain equal frequencies. We performed five validation cycles for each model. Owing to the low number of confirmed polymorphisms in the *complex_aae* model, we restricted the number of test cases in this model to 400.

We were able to improve the accuracy in all three classification models, with a slight increase in the *simple_aae* model (MutationTaster 87.2%, MutationTaster2 88.6%), and substantial changes in the *without_aae* model from 82.7% to 92.2% and the *complex_aae* model from 79.3% to 90.7%. Test cases and the predictions made by the different tools can be found on our website: http://www.mutationtaster.org/info/statistics.html.

Comparison of MutationTaster2 with similar tools

Comparison with similar tools

We compared the predictions of the web-version of MutationTaster2, MutationTaster1, SIFT, PROVEAN[12], and both PolyPhen-2 models (HumDiv and HumVar) on the same 1,100 common polymorphisms and 1,100 known disease mutations. Since PPH2 and SIFT can only handle single amino acid substitutions, we confined the comparison to non-synonymous single base exchanges using a test set of 1,800 known disease mutations from HGMD Professional and 1,800 putatively harmless polymorphisms from the 1000G. We submitted the test set via the online batch query services of the individual tools by entering the DNA change and the chromosomal position and parsed the results to extract the prediction corresponding to the amino acid exchange defined in the test set. Of the 3,600 single base exchanges from the test set, only 2,381 could be analyzed by all tools and at the same time resulted in the amino acid substitution specified in the test set – mainly because MutationTaster1 uses Ensembl build 59 with different transcripts (without MutationTaster1, the overlap was 2,957 cases). We randomly selected 1,100 disease mutations and 1,100 polymorphisms for the comparison. If several results were obtained for the given amino acid exchange, we used the first result. The statistics are shown in **Table 1** (main text).

In the case of MutationTaster, we did not consider the automatic prediction (as this would have resulted in an accuracy of 100% for 1000G variants and ClinVar mutations) but rather the probability (displayed to the right of the prediction). The performance of MutationTaster2 on real-life data, where known polymorphisms and many disease mutations are recognized, is therefore even better than the accuracies given here (see 'Application to a real exome' below).

More detailed data can be found on our website: http://www.mutationtaster.org/info/statistics.html.

Application to a real exome

In order to evaluate the performance of MutationTaster2 in a real-world example, especially the false positive rate (FPR), we exported all exonic variants found in one 1000G sample to MutationTaster2, PolyPhen-2, SIFT, and PROVEAN (see **Supplementary Table 3**). In the first step, the variants were extracted from the BAM alignment file of sample HG00377 using samtools/bcftools.

A list of all variants within exons (± 10 flanking bases) was obtained with a Perl script. This list was then exported to MutationTaster2's Query Engine and to the web services of PolyPhen-2, and SIFT/PROVEAN. The results obtained from the different tools were written to a database table; in case of more than one prediction for a variant (due to multiple transcripts), the most deleterious score was used. From this database table, we extracted all predictions for homozygous variants with a coverage of 10 or higher.

All predictions that assumed pathogenicity (including "possibly damaging" and “probably damaging”) were counted as false positives. Please note that in this real-life example, MutationTaster2's automatic classification routines were used.

Detailed results of all tests along with the procedures can be found on our website: http://www.mutationtaster.org/info/statistics.html.

Definitions

TP, true positive; TN, true negative; FP, false positive; FN, false negative

accuracy = (TP + TN) / (TP + TN + FP + FN)

sensitivity = TP / (TP + FN)

specificity = TN / (TN + FP)

NPV, negative prediction value = TN / (TN + FN)

PPV, positive prediction value = TP / (TP + FP)

FPR, false positive rate = FP / (FP + TN)

Implementation

MutationTaster2 runs on a 24-CPU system with 72 GB RAM under Linux (Fedora Core 14). Data are stored in a PostgreSQL 8.4 database. MutationTaster2 is programmed in Perl (version 5.12) and runs under mod_perl2 in an Apache 2.2 web server. All user interfaces are written in HTML (with JavaScript and AJAX functions) and are developed for the Firefox browser under Linux, Mac OS and Microsoft Windows and are regularly tested with Google Chrome, Microsoft Internet Explorer and Safari. We employ TORQUE (version 2.4) as the job scheduling system. External software used by MutationTaster2 (NNSplice, bl2seq, polyadq) runs on a RAM disk.

References

1. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Research* **40,** D84–D90 (2011).
2. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
3. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* **38,** D105–D110 (2009).
4. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20,** 110–121 (2009).
5. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
6. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41,** D64–D69 (2013).
7. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
8. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in Genie. *J. Comput. Biol.* **4,** 311–323 (1997).
9. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).
10. Seelow, D., Schwarz, J. M. & Schuelke, M. GeneDistiller--distilling candidate genes from linkage intervals. *PLoS ONE* **3,** e3874 (2008).
11. Taschner, P. E. M. & den Dunnen, J. T. Describing structural changes by extending HGVS sequence variation nomenclature. *Human Mutation* **32,** 507–511 (2011).
12. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7,** e46688 (2012).