MutationTaster evaluates disease-causing potential of sequence alterations

To the Editor: Identification of pathogenic DNA sequence alterations in patients with inherited diseases is one of the main tasks of human genetics. Next-generation sequencing (NGS) techniques enable sequencing of hundreds of candidate genes, whole linkage intervals or the entire exome. This inevitably leads to the detection of vast numbers of alterations, all of which have to be tested for their disease-causing potential. A recent study revealed more than 3.5 million alterations in the whole genome of a single individual, roughly corresponding to 1,000 alterations per mega–base pair[1].

Automated pre-evaluation of sequence variations can help to direct the subsequent in-depth analysis to the most promising candidates, hence saving time and resources. However, the currently available evaluation tools predict only the outcome of amino-acid exchanges and cannot process thousands of queries in a reasonable time.

To meet the challenges of handling high-throughput sequencing data, we developed MutationTaster, a free, web-based application for rapid evaluation of the disease-causing potential of DNA sequence alterations. MutationTaster integrates information from different biomedical databases and uses established analysis tools (**Supplementary Methods**). Analyses comprise evolutionary conservation, splice-site changes, loss of protein features and changes that might affect the amount of mRNA. Test results are then evaluated by a naive Bayes classifier[2], which predicts the disease potential. A typical query is completed in less than 0.3 seconds.

Depending on the nature of the alteration, MutationTaster chooses between three different prediction models, which are either aimed at 'silent' synonymous or intronic alterations (without_aae), at alterations affecting a single amino acid (simple_aae) or at alterations causing complex changes in the amino acid sequence (complex_aae).

To train the classifier, we generated a dataset with all available and suitable common polymorphisms and known disease-causing mutations extracted from common databases and the literature. We cross-validated the classifier five times including all three prediction models and obtained an overall accuracy of 91.1 ± 0.1%. We also compared MutationTaster with similar applications (Panther[3], Pmut[4], PolyPhen and PolyPhen-2 (ref. 5) and 'screening for non-acceptable polymorphisms' (SNAP)[6]) and analyzed the identical 1,000 disease-linked mutations and 1,000 polymorphisms with all programs. For this comparison, we used only alterations causing single amino acid exchanges. MutationTaster performed best in terms of accuracy and speed (**Table 1**). A description of all training and validation procedures and detailed statistics are available in **Supplementary Methods**.

MutationTaster can be used via an intuitive web interface to analyze single mutations as well as in batch mode. To streamline and to standardize the analysis of NGS data, we provide Perl scripts that can process data from all major platforms (Roche 454, Illumina Genome Analyzer and ABI SOLiD). MutationTaster hence allows the efficient filtering of NGS data for alterations with high disease-causing potential (see **Supplementary Methods** for an example).

Present limitations of the software comprise its inability to analyze insertion-deletions greater than 12 base pairs and alterations spanning an intron-exon border. Also, analysis of non-exonic alterations is restricted to Kozak consensus sequence, splice sites and poly(A) signal. We will add tests for other sequence motifs in the near future. MutationTaster is available at http://www.mutationtaster.org/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This project was supported by the Deutsche Forschungsgemeinschaft via the NeuroCure Cluster of Excellence, Exc 257, and the Collaborative Research Center 665 TP C4. M.S. and D.S. are members of the German network for mitochondrial disorders (mitoNET, 01GM0862), funded by the German ministry of education and research (BMBF). We thank H. Peters for providing

Table 1 | Comparison of MutationTaster with other prediction tools

Software	Number of identifications[a]					Percentage[b]		Accuracy (%)[c]			Speed (s)[d]
	TP	TN	FP	FN	ND	PPV	NPV	Predicted cases	Common cases	All cases	
MutationTaster	859	855	145	141	0	85.6	85.8	85.7	86.1	85.7	0.3
PolyPhen	728	789	206	272	5	77.9	74.4	76.0	75.8	75.8	5
PolyPhen-2 (Var model)	773	666	211	134	216	78.6	83.2	80.7	81.9	72.0	>20
PolyPhen-2 (Div model)	776	655	222	131	216	77.8	83.3	80.2	81.3	71.5	>20
SNAP	789	403	362	185	261	68.5	68.5	68.5	68.3	59.6	>20
Panther	510	196	503	181	610	50.3	52.0	50.8	53.4	35.3	>20
Pmut	581	720	270	418	11	68.3	63.3	65.4	62.0	65.0	>20

[a] TP, true positive; TN, true negative; FP, false positive; FN, false negative; and ND, not determined. [b] PPV (positive prediction value) = TP / (TP + FP) and NPV (negative prediction value) = TN / (TN + FN). [c] Accuracy for predicted cases = (TP + TN) / (TP + TN + FP + FN); accuracy for common cases = (TP + TN) / (TP + TN + FP + FN), and only those cases predicted by all tools were included); and accuracy for all cases = (TP + TN) / (TP + TN + FP + FN + ND). Owing to the limitation of the other prediction tools to single amino acid exchanges, MutationTaster's accuracy shown in this table only refers to the single amino acid exchange prediction model and should not be mistaken for the overall accuracy of 91.1%, which includes all three models. [d] Mean time needed for one prediction via the web interface for the same five alterations.
We analyzed the identical 2,000 variations using each tool.

mutation data, M. Zhang and M. Reese for allowing us to integrate polyadq and NNSplice, E. Lüdeking for proofreading the manuscript, and all beta users whose valuable recommendations guided the development of MutationTaster.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jana Marie Schwarz[1,2], Christian Rödelsperger[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, "Charité–Universitätsmedizin Berlin", Berlin, Germany.
e-mail: dominik.seelow@charite.de

1. Wheeler, D.A. *et al. Nature* **452**, 872–876 (2008).
2. Hand, D.J. & Yu, K.M. *Int. Stat. Rev.* **69**, 385–398 (2001).
3. Mi, H. *et al. Nucleic Acids Res.* **33**, D284–D288 (2005).
4. Ferrer-Costa, C. *et al. Bioinformatics.* **21**, 3176–3178 (2005).
5. Adzhubei, I.A. *et al. Nat. Methods* **7**, 248–249 (2010).
6. Bromberg, Y. & Rost, B. *Nucleic Acids Res.* **35**, 3823–3835 (2007).

mrsFAST: a cache-oblivious algorithm for short-read mapping

To the Editor: In addition to single-nucleotide variations and small insertions-deletions (indels), larger-sized structural variations (for example, insertions, deletions, inversions, segmental duplications and copy-number polymorphisms) contribute to human genetic diversity. In almost all recent structural variation discovery (SVD) studies, short reads from a donor genome have been mapped to a reference genome as a first step. The accuracy of such an SVD study is directly correlated to the accuracy of this mapping step, which also provides the main computational bottleneck of the SVD study.

Next-generation sequencing technologies provide increasingly longer reads (currently ~400 base pairs (bp) for the Roche 454 platform and 2 ×100 bp for the Illumina platform). However, even with the increased read lengths, ambiguity in read mapping remains a problem. A human genome resequencing study[1] using 36-bp reads has reported, on average, 1,628 mapping locations per read within two mismatches and indels. In our study, on a set of one million 36-bp reads from a Yoruban individual (NA18507), we observed an average of 1,486 mapping locations within two mismatches and indels and 1,411 mapping locations when we allowed only two mismatches but no indels. The mapping multiplicity only reduced to 615 locations for 50-bp reads within three mismatches, 185 locations for 75-bp reads within four mismatches and 140 locations for 100-bp reads within six mismatches (**Table 1**).

As structural variants are typically observed in repeat regions, it is critical to consider all possible mapping locations for each read. To address this need, recently developed mapping tools such as Maq[2], Bowtie[3], RazerS[4] and Burrows-Wheeler alignment (BWA)[5] have options to report read multiplicities, but they do not capture all possible mapping locations. Bowtie and BWA use the Ferragina-Manzini index[6] (FMI), which is designed (and works effectively) for finding exact matches. Extending the FMI to handle mismatches or indels is only achieved by heuristic generalizations. As the read length and the corresponding number of mismatches and indels to be tolerated increase, these methods deteriorate exponentially in terms of speed and/or accuracy.

We developed 'micro-read (substitutions only) fast alignment and search tool' (mrsFAST), a cache-oblivious short read mapping algorithm that rapidly finds all mapping locations of a collection of short reads from a donor genome in the reference genome within a user-specified number of mismatches through indexing both the reference genome and the short reads, and executing a simple cache-oblivious, all-to-all list comparison algorithm (**Supplementary Note**). We also developed mrFAST-CO, a version of mrsFAST that can handle indels and substitutions (equivalently, mrFAST-CO is a doubly indexed, cache-oblivious version of previously developed mrFAST[1], a simple 'seed and extend'–type mapping algorithm).

Like mrFAST, mrsFAST and mrFAST-CO are seed-and-extend algorithms. Such algorithms work by first placing a *k*-mer (seed) from a read by interrogating the index (in the form of a hash table for all *k*-mers and their respective loci) of the reference genome and then extending them by allowing at most a user-specified number of mismatches or indels. During the execution of the algorithms, the operating system copies the information related to the seed locations from the main memory to the much faster levels of cache memory, and the extension step is performed using the information stored in the cache. In a naive execution (in comparison to a cache-oblivious execution) of such a seed-and-extend algorithm, the seed mapping locations to be compared to the read would be streamed through the cache. As cache capacity is very

Table 1 | Speed and accuracy of mapping

Read length (errors):	36 bp (2 errors)			50 bp (3 errors)			75 bp (4 errors)			100 bp (6 errors)		
Algorithm	Time (h:min)[a]	Reads mapped (%)[b]	Locations reported (millions)[c]	Time (h:min)[a]	Reads mapped (%)[b]	Locations reported (millions)[c]	Time (h:min)[a]	Reads mapped (%)[b]	Locations reported (millions)[c]	Time (h:min)[a]	Reads mapped (%)[b]	Locations reported (millions)[c]
Bowtie	5:14	91.65	1,404	3:13	92.73	610	NA	NA	NA	NA	NA	NA
BWA	3:10	92.05	1,581	10:23	93.38	729	59:35	90.16	212	67:38	87.91	42
Maq	6:45	90.91	1,609	10:05	89.25	458	NA	NA	NA	NA	NA	NA
mrFAST-CO	6:12	92.18	1,486	9:21	93.39	663	11:32	90.22	193	17:54	88.55	155
mrsFAST	2:00	91.79	1,411	1:55	92.91	613	2:00	89.35	177	2:49	87.27	138
RazerS[d]	10:17	91.79	<100	12:17	92.91	<100	12:00	89.35	<100	25:10	87.27	<100
BWA[e]	0:10	92.05	<1	0:15	93.38	<1	0:25	90.16	<1	7:04	87.91	<1

We mapped one million reads of indicated read lengths and within the given number of errors, to the human reference genome HG18 build 36 by indicated algorithms. All rows (except the last two) denote the time needed to report all mapping locations. Because of its high memory requirement, we could not run RazerS for read multiplicities >100. Note that in some columns the total number of mapping locations is higher for Maq or BWA than for mrsFAST or mrFAST-CO because Maq often returns mapping locations with an error rate higher than the user-specified rate and BWA returns certain mapping locations multiple times. NA, not applicable.

[a]Time required for mapping (on a single personal computer). [b]Percentage of the reads mapped. [c]Total map locations reported (in millions). [d]Maximum multiplicity, 100. [e]Single location.