BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 22 2013, pages 2953–2954
doi:10.1093/bioinformatics/btt507

Databases and ontologies Advance Access publication September 3, 2013

M2SG: mapping human disease-related genetic variants to protein sequences and genomic loci

Renkai Ji[1,†], Qian Cong[1,†], Wenlin Li[1] and Nick V. Grishin[1,2,*]

[1]Departments of biophysics and biochemistry, University of Texas Southwestern Medical Center and [2]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050, USA

Associate Editor: Igor Jurisica

ABSTRACT

Summary: Online Mendelian Inheritance in Man (OMIM) is a manually curated compendium of human genetic variants and the corresponding phenotypes, mostly human diseases. Instead of directly documenting the native sequences for gene entries, OMIM links its entries to protein and DNA sequences in other databases. However, because of the existence of gene isoforms and errors in OMIM records, mapping a specific OMIM mutation to its corresponding protein sequence is not trivial. Combining computer programs and extensive manual curation of OMIM full-text descriptions and original literature, we mapped 98% of OMIM amino acid substitutions (AASs) and all SwissProt Variant (SwissVar) disease-related AASs to reference sequences and confidently mapped 99.96% of all AASs to the genomic loci. Based on the results, we developed an online database and interactive web server (M2SG) to (i) retrieve the mapped OMIM and SwissVar variants for a given protein sequence; and (ii) obtain related proteins and mutations for an input disease phenotype. This database will be useful for analyzing sequences, understanding the effect of mutations, identifying important genetic variations and designing experiments on a protein of interest.

Availability and implementation: The database and web server are freely available at http://prodata.swmed.edu/M2S/mut2seq.cgi.

Contact: grishin@chop.swmed.edu

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

Received on April 23, 2013; revised on August 13, 2013; accepted on August 27, 2013

1 INTRODUCTION

Online Mendelian Inheritance in Man (OMIM) (McKusick, 2007) consists of full-text overviews of phenotypes, especially human diseases, and the corresponding genetic variants including substitutions, deletions, insertions and intervening sequences. The information in OMIM is derived from literature and documented manually by human curators. OMIM is a valuable database to associate phenotypes and human diseases with particular genes; however, it is not a reliable resource to assign these phenotypes to certain mutations in the corresponding proteins. A previous report (Li *et al*., 2012) suggested that >20% of amino acid substitutions (AASs) in OMIM cannot be mapped to the canonical sequences in SwissProt (i.e. either the mutation position or the residue type is inconsistent with the SwissProt sequence) (Apweiler *et al*., 2004). This discrepancy mainly results from OMIM mutations being derived from literature, where authors report different gene isoforms. In addition, errors in the manual OMIM records are another obstacle for confident mapping.

Previous studies attempted to map the OMIM AASs onto SwissProt sequences (Martin, 2005; Peterson *et al*., 2010; Yip *et al*., 2004). However, the mutations that cannot be confidently mapped were simply excluded or ignored. In 2005, Martin designed an automated procedure to maintain a validated mapping between OMIM AASs and SwissProt entries. However, the performance was not sufficient. In all, 20% of all the entries fall into their class C, meaning not all the associated mutations could be mapped. In addition, the simple offset approach they applied could cause a considerable rate of false-positives.

Combining automatic tools and manual curation, we mapped 98% of OMIM AASs and all disease-related genetic variants in SwissProt Variant (SwissVar) to SwissProt reference sequences (RSs) and compiled a database, M2SG (Mapping mutations to Sequence and Gene). (A detailed comparison between our work and previous studies is in Supplementary Table S1). In all, 99.96% of these mutations are mapped to the genomic loci and 99.2% can be attributed to either single nucleotide polymorphisms (SNPs) in dbSNP database or putative SNPs, which validated our mapping at the protein level. In addition, we provide a user-friendly interface to search the database with various queries. We expect it to be a useful resource for understanding the effects of mutations and experimental work on the disease-related proteins.

2 METHODS

2.1 Mapping mutations to RS

OMIM and SwissProt databases [March, 2012] were obtained as flatfiles, supplemented by querying their web interfaces. The cross-references between SwissProt and OMIM entries were obtained by the cross-links in databases. The canonical sequences of the SwissProt entries were used as RSs. AASs in the SwissVar database can be directly mapped to RS, whereas OMIM AASs were processed hierarchically by the following methods:

Direct mapping: If all 'native' residues in OMIM mutations from one OMIM entry match the residues in the RS at the positions indicated in the OMIM records, we consider the mutations to be validated directly.

BLAST mapping: Due to the presence of multiple gene isoforms and errors in coding region prediction of genes, the original studies cited in OMIM might have referred to an alternative native sequence (ANS). Presumably the ANS should be highly similar to our RS, and thus we identified them by BLAST (*E*-value = 0.001, Altschul *et al*., 1997) against

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

the non-redundant database from the RS, filtered by higher than 98% sequence identity computed ignoring gaps to improve detection of exon-skipped isoforms. If all the mutations of an OMIM entry can be mapped to the ANS, this ANS was used as an intermediate to map the mutations to RS and correct the mutation positions to match those in the RS.

Offset mapping: In many cases, the ANSs used in the original study were not present in NR but could be deduced by applying an offset to the residue numbers in the RS. This strategy is equivalent to truncating or extending the open reading frames of the encoding gene for the RS. The offset that maximized the number of mutations that could be correctly mapped was tested. If these offset mutations could be cross-validated by records in the SwissVar database or they corresponded to all the mutations from one OMIM entry containing at least three mutations, the deduced offset was considered valid and used to map these mutations.

Manual mapping: We manually checked the OMIM mutation description, information in SwissProt and the original literature for mutations that could not be mapped by the automatic methods. A considerable amount of errors in OMIM mutation record were detected, and several types of errors are exemplified in Supplementary Figure S1.

2.2 Mapping mutations to genomic loci and SNPs

A portion of OMIM and SwissVar entries have links to the dbSNP database (Sherry *et al.*, 2001), which provides the genomic loci and corresponding SNPs. For those without such links, we applied the PICMI server (Le Pera *et al.*, 2010) to map the mutations to the genomic loci and putative SNPs. For cases where PICMI failed, we combined computational and manual approaches to align our RSs to the corresponding protein sequences in the Ensembl (Flicek *et al.*, 2013) or CCDS (Pruitt *et al.*, 2009) database, which were used as intermediates to map mutations to genomic loci and deduce SNPs that could cause these AASs in the proteins.

3 RESULTS

We mapped >98% of OMIM AASs, and the number of mutations mapped by each method is shown in Figure 1A. Although the majority of OMIM mutations can be confidently mapped by computer programs, manual curation was applied to a considerable portion (18.0%). A small portion of OMIM mutations (2.7%) could not be mapped due to the lack of evidence even in the cited literature (Supplementary Table S2) or missing cross-links between OMIM and SwissProt entries. To make the mutations consistent with the RS, 3096 out of 13 221 OMIM mutation records were modified. Most of these modifications involved a shift in the residue number and 4% corrected other OMIM annotation errors (Supplementary Table S3).

The current M2SG database includes 12 855 AASs from 2292 OMIM entries and 21 405 disease-related AASs from 1727 proteins in SwissVar. OMIM and SwissVar AASs overlap partially, and our mapping revealed 33 redundant AASs in OMIM, resulting in a total of 2315 proteins with 26 851 AASs (Fig. 1B). Interestingly, 596 out of 36 942 functionally neutral mutations in SwissVar are disease causing according to OMIM (Supplementary Table S4), indicating possible errors in either database.

In all, 99.96% of the mutations in M2SG are mapped to the genomic loci and 99.2% can be attributed to either SNPs in the dbSNP or other single nucleotide changes in the genomic loci (Fig. 1C), which validated mutations in M2SG. In contrast, original OMIM mutation records yielded a 24% failure rate for this genomic loci and SNP mapping (Fig. 1D).

Our results are presented as an online database with a user-friendly interface. The database can be queried by protein sequence, original OMIM entry number, UniProt accession number, protein name, gene name or disease phenotype (Supplementary Fig. S2). The returned web page contains the most relevant protein and correctly mapped mutations (Supplementary Fig. S3).

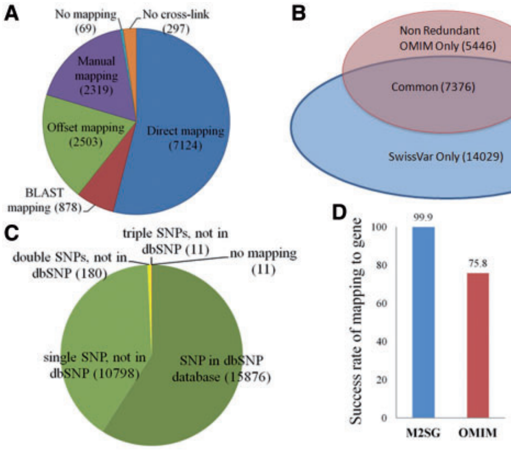

Fig. 1. (**A**) Mapping OMIM mutations to protein sequences done by different methods; (**B**) constitution of the M2SG database; (**C**) mapping mutations in M2SG to recorded or putative SNPs; and (**D**) success rate of mapping to genomic loci and SNPs from M2SG and OMIM mutations

ACKNOWLEDGEMENTS

The authors thank Xiaoming Zhang for help with literature curation and Dr. Zbyszek Otwinowski for useful suggestions.

Funding: National Institutes of Health GM094575 and the Welch Foundation I-1505 to NVG. Qian Cong is a Howard Hughes Medical Institute International Student Research fellow.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.

Le Pera,L. *et al.* (2010) PICMI: mapping point mutations on genomes. *Bioinformatics*, **26**, 2904–2905.

Li,Z. *et al.* (2012) An examination of the OMIM database for associating mutation to a consensus reference sequence. *Protein Cell*, **3**, 198–203.

Martin,A.C.R. (2005) http://www.bioinf.org.uk/omim/ (9 September 2013, date last accessed).

McKusick,V.A. (2007) Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

Peterson,T.A. *et al.* (2010) DMDM: domain mapping of disease mutations. *Bioinformatics*, **26**, 2458–2459.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

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