Automated validation of genetic variants from large databases: ensuring that variant references refer to the same genomic locations

Mark Y. Tong\textsuperscript{1}, Christopher A. Cassa\textsuperscript{2} and Isaac S. Kohane\textsuperscript{1,2,*}

\textsuperscript{1}CBMI, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115 and \textsuperscript{2}Children’s Hospital Informatics Program, Children’s Hospital Boston, 1 Autumn Street, #721, Boston, MA 02215-5362, USA

Received on August 19, 2010; revised on December 14, 2010; accepted on January 15, 2011

ABSTRACT

Summary: Accurate annotations of genomic variants are necessary to achieve full-genome clinical interpretations that are scientifically sound and medically relevant. Many disease associations, especially those reported before the completion of the HGP, are limited in applicability because of potential inconsistencies with our current standards for genomic coordinates, nomenclature and gene structure. In an effort to validate and link variants from the medical genetics literature to an unambiguous reference for each variant, we developed a software pipeline and reviewed 68,641 single amino acid mutations from Online Mendelian Inheritance in Man (OMIM), Human Gene Mutation Database (HGMD) and dbSNP. The frequency of unresolved mutation annotations varied widely among the databases, ranging from 4 to 23%. A taxonomy of primary causes for unresolved mutations was produced.

Availability: This program is freely available from the web site (http://safegene.hms.harvard.edu/aa2nt/).

Contact: mt153@hms.harvard.edu; mark_tong2009@yahoo.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Large numbers of genetic variants from medical and genetics publications have been compiled in databases, including the Online Mendelian Inheritance in Man (OMIM), the Human Gene Mutation Database (HGMD), among others. For example, the HGMD (Stenson \textit{et al}., 2009) has curated 100,329 disease-associated genetic variants in its current release (March 2010), and OMIM has described 20,068 variants as of June 2010 (Amberger \textit{et al}., 2009). These disease-associated variants are valuable in the understanding, prevention and diagnosis of human disease. With the imminent reduction to practice of whole-genome interpretation (Ashley \textit{et al}., 2010; Ormond \textit{et al}., 2010), an overview of the accuracy of these databases is important in understanding how much quality improvement work remains to make these prior genome-wide annotations clinically useful. We focus here on the syntactic accuracy of the annotations which are an important but small step toward assessing their clinical validity (Kohane \textit{et al}., 2006).

To this end, we developed a software module, aa2nt, which provides basic validation of single amino acid changes using information from current databases, derives the corresponding DNA change from an amino acid change and generates Human Genome Variation Society (HGVS)-recommended names (Supplementary Fig. S1). We applied aa2nt to a selected set of variants from three commonly used databases (OMIM, HGMD and dbSNP) to evaluate whether we could correctly resolve the locations of variants in current annotation databases. We validated 66,638 single nucleotide mutations from OMIM, HGMD and dbSNP and obtained a passing rate ranging from 77 to 96%.

2 METHODS

The following algorithm was used to validate and map amino acid substitutions caused by a single nucleotide change:

1. Required input data: gene symbol, codon number, wild-type amino acid, variant amino acid.


2. For each cDNA transcript obtained in step 1.2, generates the cDNA codon sequence corresponding to the codon number of amino acid change, and translate it to the corresponding amino acid.

3. Compare the obtained amino acid to the ‘wild-type’ reference amino acid at that position.

3.1 If identical, it is validated.

3.2 Otherwise, test if the gene has a signal peptide (http://www.signalpeptide.de), which could alter the codon numbering. If yes, adjust the codon number with signal peptide added.

4. Identify all possible single nucleotide changes from the reference cDNA sequence to all the possible genetic codons of the variant amino acid.

5. Generate tuple of HGVS name(s) of DNA and protein changes.

A detailed flow chart illustrating this process with sample data can be found in the Supplementary Materials (Supplementary Figure S1).
Table 1. Summary of validated mutation annotationsa,b

<table>
<thead>
<tr>
<th>Database</th>
<th>OMIM</th>
<th>HGMD (I)</th>
<th>HGMD (II)</th>
<th>dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passed</td>
<td>7722 (76.8%)</td>
<td>47260 (81.2%)</td>
<td>55115 (95.8%)</td>
<td>2310 (87.3%)</td>
</tr>
<tr>
<td>Unresolved</td>
<td>2332</td>
<td>10922</td>
<td>2364</td>
<td>336</td>
</tr>
<tr>
<td>Total</td>
<td>10054</td>
<td>58182</td>
<td>57479</td>
<td>2646</td>
</tr>
</tbody>
</table>

aTwo sets of codon numbers were used for HGMD data: original (I) and HGVS form (II).

3 RESULTS

3.1 Validation of variants by codon number and amino acid substitutions

We selected 10054 single amino acid substitution variants from OMIM (Supplementary Table S1), 58182 variants from HGMD, 57479 variants from HGMD where the HGVS codon number is available, 2646 variants from table OmimVarLocusIdSNP in dbSNP (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/database/organism_data/OmimVarLocusIdSNP/bcp.gz, Supplementary Table S2) as input files for the validation program. The results of validations are summarized in Table 1.

3.2 Validating program performance using a gold-standard dataset

To evaluate the accuracy of base changes predicted from the specified amino acid change, we selected 5959 mutations in OMIM which have details about the specific base change involved in the HGMD. 5113 (86%) of 5959 variants mapped to a single nucleotide change identical to one described in HGMD. The remaining 846 variants mapped to more than one possible codon. If the highest frequency codon is used based on the frequency table (Nakamura et al., 2000), 5586 (94%) of the predicted codons agree with the official gene symbol, 219 passed validation.

We would like to thank Dr Vincent Fusaro and Dr Joon Lee for their assistance in the development of the matching algorithm.

ACKNOWLEDGEMENTS

We would like to thank Dr Vincent Fusaro and Dr Joon Lee for their assistance in the development of the matching algorithm.

Funding: This research was supported by grant 1-RC1-LM010470-01 from the National Library of Medicine and by training grant 5T32HD040128 from the National Institute of Child Health and Human Development (Dr Cassa).

Conflict of Interest: none declared.
REFERENCES


