BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 6 2011, pages 891–893
doi:10.1093/bioinformatics/btr029

Databases and ontologies

Advance Access publication January 22, 2011

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

Mark Y. Tong[1], Christopher A. Cassa[2] and Isaac S. Kohane[1,2,*]

[1]CBMI, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115 and [2]Children's Hospital Informatics Program, Children's Hospital Boston, 1 Autumn Street, #721, Boston, MA 02215-5362, USA

Associate Editor: Jonathan Wren

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*.

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

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 *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 *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 *et al*., 2010; Ormond *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 *et al*., 2006).

*To whom correspondence should be addressed.

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.
   - 1.1 Replace gene synonyms with official gene symbols by querying data available in Entrez Gene (ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz).
   - 1.2 Retrieve available human cDNA sequences from RefSeq (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.gbff.gz) for the given gene.
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 codon 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 annotations[a,b]

Database	OMIM	HGMD (I)	HGMD (II)	dbSNP
Passed	7722 (76.8%)	47 260 (81.2%)	55 115 (95.8%)	2310 (87.3%)
Unresolved	2332	10 922	2364	336
Total	10 054	58 182	57 479	2646

[a]Two sets of codon numbers were used for HGMD data: original (I) and HGVS form (II).
[b]Versions: OMIM: 2010; HGMD: professional version, 2010 (2); dbSNP: 2010; HG18.

3 RESULTS

3.1 Validation of variants by codon number and amino acid substitutions

We selected 10 054 single amino acid substitution variants from OMIM (Supplementary Table S1), 58 182 variants from HGMD, 57 479 variants from HGMD where the HGVS codon number is available, 2646 variants from table OmimVarLocusIdSNP in dbSNP (ftp://ftp.ncbi.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 mutant codon sequence in HGMD. The aa2nt module does not predict codon change(s) if more than one nucleotide is required to make the prediction.

3.3 Evaluation of major categories of unresolved annotations

We analyzed 2332 annotations from the OMIM database which did not achieve proper resolution using the aa2nt test pipeline, and grouped them into the following categories:

- *Amino acid assignment problems*: the annotated amino acid was not present at the described location in any of the known gene product isoforms. Of the 2044 unresolved variants, we checked if the gene products contained a signal peptide. Of 2044, 950 did have a signal peptide sequence and 528 of the annotations passed a second round validation after re-indexing the codon number with the signal peptide added.
- *Non-standard gene symbols*: 258 variants in 75 genes used gene aliases instead of official gene symbols. After replacing the alias with an official gene symbol, 219 passed validation.
- *Codon number greater than protein length*: 27 variants belong to this class. An example is PTEN, which encodes a protein of 403 amino acids. Therefore, HIS861ASP (OMIM 601728) would be invalid.
- *Genomic coordinates unavailable*: there were three examples where genomic coordinates were unavailable: gene ImmunoGlobin Heavy constant Mu (IGHM) is an official gene symbol, but is not in the University of California at Santa Cruz (UCSC) gene annotation database table reflink (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refLink.txt.gz). Gene OA1 and KCNJ18 (OMIM: 613236) were also missing coordinate information.
- *Naming of DNA changes*: in one mutation, OMIM ID: 600946.0005 (GAA180GAG), the codon change was used to number the cDNA change. This is inconsistent with the HGVS suggestion, and it is a GAG to GAA change, not GAA to GAG (Berg *et al.*, 1992).

In this first step of assessing the syntactic validity of the largest publicly available mutation annotation databases, we found that the majority of the annotations were accurate. Nonetheless, in aggregate there were several thousand mutation annotations that did not pass a simple syntactic verification procedure even after allowances were made for isoforms, signal peptide sequence and the use of gene symbol aliases rather than the standard nomenclature. There are other potential explanations for mis-numbered sequences, including other propeptides that might be cleaved during the post-translational process. This may explain the difference between the 44% of variants (422 of 950) that did contain a signal peptide in their sequence that still did not pass resolution using aa2nt even when it was considered.

We have made available a list of the variants that did not resolve using aa2nt to enable a community review and manual annotation process. These data are available using a web application at http://safegene.hms.harvard.edu/zak/unresolvedOmimVariants.jsp.

Many of these difficulties are the residue of early discovery work prior to standardization—it is unsurprising that there is difficulty in resolving non-synonymous from OMIM, as the database hosts historical discoveries from the literature. Other variants that did not achieve resolution appear to potentially be the result of some form of data transcription, transfer or copying error. These syntactic errors fall well short of the clinical requirements for accurate interpretation of human variants. As we approach whole genome clinical interpretation, it seems that there is an increasing common interest and public good in ensuring that all previous and new annotation data be vetted automatically by a suite of tools such as aa2nt with a standard resolution procedure for those annotations that do not pass this validation process. Indeed, such a pipeline appears to be an essential component to the Genome Commons that some have envisaged (Brenner, 2007; Field *et al*., 2009), as well as a valuable addition to the process of mutation finding through text mining (Horn *et al*., 2004; Kuipers *et al*., 2010).

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

Amberger,J. *et al.* (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.

Ashley,E.A. *et al.* (2010) Clinical assessment incorporating a personal genome. *Lancet*, **375**, 1525–1535.

Berg,M.A. *et al.* (1992) Mutation creating a new splice site in the growth hormone receptor genes of 37 Ecuadorean patients with Laron syndrome. *Hum. Mutat.*, **1**, 24–32.

Brenner,S.E. (2007) Common sense for our genomes. *Nature*, **449**, 1915–1916.

Field,D. *et al.* (2009) 'Omics data sharing. *Science*, **326**, 234.

Horn,F. *et al*. (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, **20**, 557–568.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kohane,I.S. *et al.* (2006) The incidentalome: a threat to genomic medicine. *JAMA*, **296**, 212–215.

Kuipers,R. *et al*. (2010) Novel tools for extraction and validation of disease-related mutations applied to fabry disease. *Hum. Mutat.*, **31**, 1026–1032.

Nakamura,Y. *et al.* (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.*, **28**, 292.

Ormond,K.E. *et al.* (2010) Challenges in the clinical application of whole-genome sequencing. *Lancet*, **375**, 1749–1751.

Stenson,P.D. *et al.* (2009) The Human Gene Mutation Database: 2008 update. *Genome Med.*, **1**, 13.