Protein Cell 2012, 3(3): 198–203
DOI 10.1007/s13238-012-2037-2

Protein & Cell

COMMUNICATION

An examination of the OMIM database for associating mutation to a consensus reference sequence

Zuofeng Li[1] ✉, Beili Ying[2], Xingnan Liu[3], Xiaoyan Zhang[3], Hong Yu[4]

[1] Shanghai Center for Bioinformation Technology, Shanghai 200235, China
[2] School of Life Science, Fudan University, Shanghai 200433, China
[3] School of Life Science and Technology, Tongji University, Shanghai 200295, China
[4] University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA
✉ Correspondence: lizuofeng@gmail.com
Received February 24, 2012 Accepted March 19, 2012

ABSTRACT

Gene mutation (e.g. substitution, insertion and deletion) and related phenotype information are important biomedical knowledge. Many biomedical databases (e.g. OMIM) incorporate such data. However, few studies have examined the quality of this data. In the current study, we examined the quality of protein single-point mutations in the OMIM and identified whether the corresponding reference sequences align with the mutation positions. Our results show that close to 20% of mutation data cannot be mapped to a single reference sequence. The failed mappings are caused by position conflict, site shifting (peptide, N-terminal methionine) and other types of data error. We propose a preliminary model to resolve such inconsistency in the OMIM database.

KEYWORDS single-point mutation, OMIM, reference sequence, data quality

INTRODUCTION

Mutation data (i.e. substitution, insertion and deletion) is an important part of biomedical knowledge, and most mutation genotypes and their corresponding phenotypes are reported in the literature individually which makes it a tedious job to search and retrieve the effects of a gene variant (Li et al., 2011). There are some online resources such as the Online Mendelian Inheritance in Man (OMIM), Human Gene Mutation Database (HGMD) and Locus-specific Database (LSDB) (Horaitis et al., 2007; George et al., 2008). In order to understand the genetic and biochemical mechanisms through mutations, biomedical scientists need to associate mutations with their actual sequences. However, such associations are difficult to access due to the following reasons. First, most existing databases were generated manually and were frequently built within a specific biomedical sub-domain. Second, it has also been reported that there are substantial inconsistencies among these databases with respect to mutation data (George et al., 2008).

For biomedical scientists, manually identifying a mutation of a gene can be tedious and sometimes error prone. For example, several important amino acid residues (Y697, Y706, Y721, Y807 and Y559) for the protein CSF-1R are validated by a single-point mutation experiment in the following excerpt:

"*Several tyrosine autophosphorylation sites have been mapped in the CSF-1R, including Tyr 697, Tyr 706, and Tyr 721 in the so-called kinase insert (KI) region that divides the catalytic domain, Tyr 807 in the activation loop of the catalytic domain, and Tyr 559 in the juxtamembrane region.*" (Lee and States, 2000)

Another scientist interested in studying the function of the protein CSF-1R might like to search on whether additional mutations of CSF-1R have been studied and finds another article that states:

"*Interestingly, tyrosine 561 lies in the juxtamembrane region of the CSF-1 receptor, as does tyrosine 579 in the PDGF receptor, and there is some similarity between the two sequences.*" (Alonso et al., 1995)

In both articles, the authors seem to describe the same

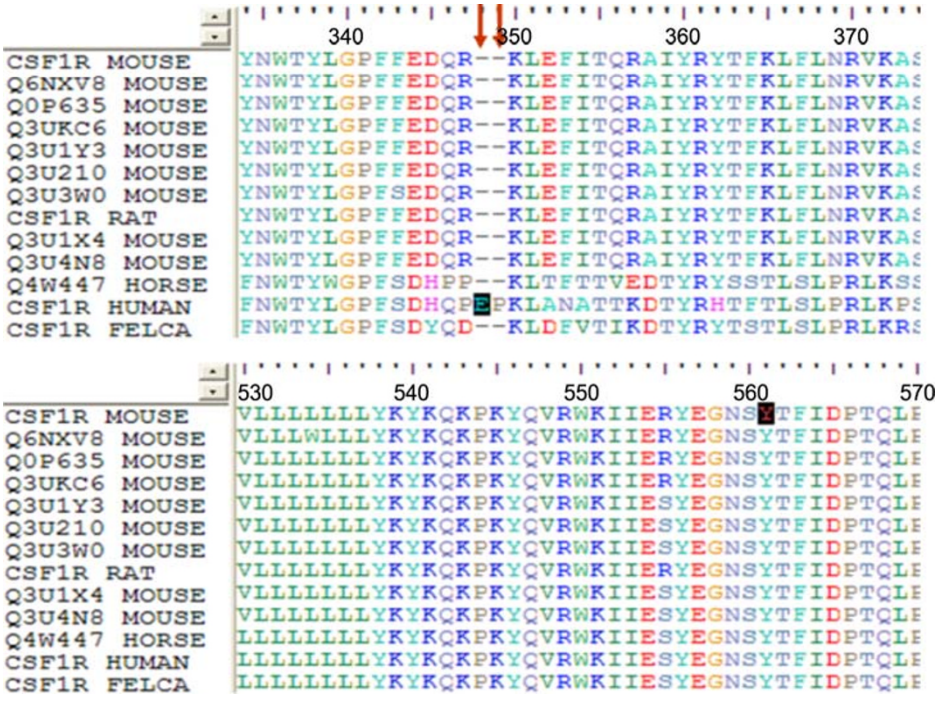

Figure 1. Sequence alignment CSF-1R. The two insertion events are marked by red arrow at positions 348 and 349.

protein residue in the juxtamembrane region of protein CSF-1R (Y559 in Lee's paper and Y561 in Alonso's paper). They may also be referring to the SwissProt sequence P09581 (CSF1R_MOUSE) because most of the tyrosine sites match except Y561.

After tracing the revision history of P09581 using the UNISAVE from EBI (Leinonen et al., 2006), we found no change in the 559 or 561 sites in 116 versions. We then retrieved sequences from different species. After alignment, we found that two amino acid insertions in the human sequence at position 348. From this alignment, we can say that Y561 and Y559 refer to the same functional site, which is conserved between human and mouse (Fig. 1). However, how can a biologist know that Y559 is the same as Y561? We propose to associate each mutation data to its original sequence so that the problem created by "different sites" will be resolved through the sequence alignment. Known as reference sequence analysis, this analysis process is a common method that biologists use for tracing mutation sequences.

One of many biological databases that provide links to reference sequences, the OMIM database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) is an expert-annotated database that organizes genes and genetic disorders. Each OMIM entry has a full-text summary of a genetically determined phenotype and/or gene (Hamosh et al., 2005). The OMIM Entrez database entry contains established gene locus and phenotypic descriptions. Although the OMIM is a non-sequence-based information resource, many important sequence-relating links are provided, such as locus-specific databases. The data can be of tremendous value for clinical genomics researchers, physicians and patients.

Although it is not included in each record, an important part of many OMIM records is the ALLELIC VARIANTS (AV) section, which primarily describes disease-producing mutations (Xi et al., 2009). The Entrez Programming Utilities (eUtils) provide a stable Entrez query and database system interface (Wheeler et al., 2007). The AV information is deposited in a local SQlite database. A Perl script is used to regularly update the database. One subsection of the AV record in XML format is named AV_TEXT which provides gene names, mutation information (position and amino or nucleotide acid). All of the mutation data are extracted from this section.

In this study, we evaluate the success rate for identifying the reference sequence for each of the mutations deposited in the AV section of the OMIM database through the RefSeq link and sequence tracing approach.

RESULTS

Mutation data collection and sequence mapping

The point mutation data is the major part of OMIM allelic variance data. In total, there are 2329 OMIM entries comprising a total of 17,337 allelic variance records. After parsing for point mutation, there are 1939 OMIM entries with 10,766 position and amino acid pairs (PAAPs). Compared to other types of mutations, such as insertion and deletion, point mu-

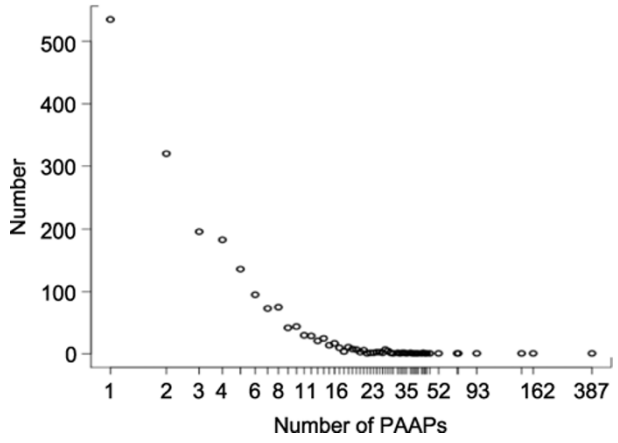

Figure 2. Number of OMIM entries as a function of the number of position and amino acid pairs (PAAPs).

tation data represent a major part of the OMIM database.

Most of the OMIM entries have few mutation data. Ninety-six percent of the 1939 OMIM entries incorporated 20 or fewer PAAPs. The maximum PAAP number is 387, which corresponds to Hemoglobin B (OMIM# 141900). Two OMIM entries, Hemoglobin A1 (OMIM# 141800) and Hemophilia A (OMIM# 306700), incorporate more than 100 point mutations. The distribution of PAAPs number of OMIM entry is shown in Fig. 2.

In this work, we filtered out some conflict positions. There are 25 allelic variances in which there is more than one type of wild amino acid residue in the same position, such as in OMIM# 602421.

*602421 CFTR
.0022 CYSTIC FIBROSIS [CFTR, TRP1282TER]
.0129 CYSTIC FIBROSIS [CFTR, HIS1282TER]

We call these mutation conflict positions, which may be caused by existence of polymorphism. In this study, we just selected one of them for analysis.

Moreover, by using mapping files from the sequence database, not all of the OMIM entries with allelic variance could be mapped to a sequence. There are seven such instances in GenBank and eight in UniProtKB. However, there is no overlap between them.

Table 1 The OMIM IDs without sequence link*

Database	OMIM IDs
GenBank	516001 516002 516003 516004 516030 516070 608620
UniProtKB	142858 300757 611770 612373 612676 612719 612724 612732

* these ids are added into inconsistent group

Reference sequence analysis

We ran the OMIM point mutation data against GenBank and UniProtKB separately with or without sequence tracing with the aid of the eUtils and dbFetch tools. For all of the OMIM entries with the same number of PAAPs, we calculated the total number and the ratio of mapped entries found by *RefSeq-link* or *Sequence tracing* method.

For the RefSeq-link approach to the Genbank database, with the increase of the number of PAAPs, the average number of mapped sequences decreased rapidly, especially when the number was greater than ten (Fig. 3A). At the same time, the ratio of mapped entries also decreased gradually (Fig. 3B).

When the number of PAAPs was greater than 40, we could not find consistent reference sequence for most of the OMIM entries. However, there are three exceptions: Phenylalanine hydroxylase (OMIM:612349), Dystrophin (OMIM:300377), and HEMOGLOBIN--ALPHA LOCUS 1 (OMIM:141800), which have 43, 45 and 96 point mutation data, respectively. For the Sequence tracing approach taken with the GenBank database, the results are better, but not significantly (Supplementary data 1). Only one entry for ANKYRIN 2 (OMIM: 106410) was mapped with the aid of sequence tracing. The mapped GenBank entry is CAA40279.1.

Compared to the old version with the BLAST2 program, the newest version of CAA40279 has an insertion of 33 amino acid residues at 1042 (Tatusova and Madden, 1999),

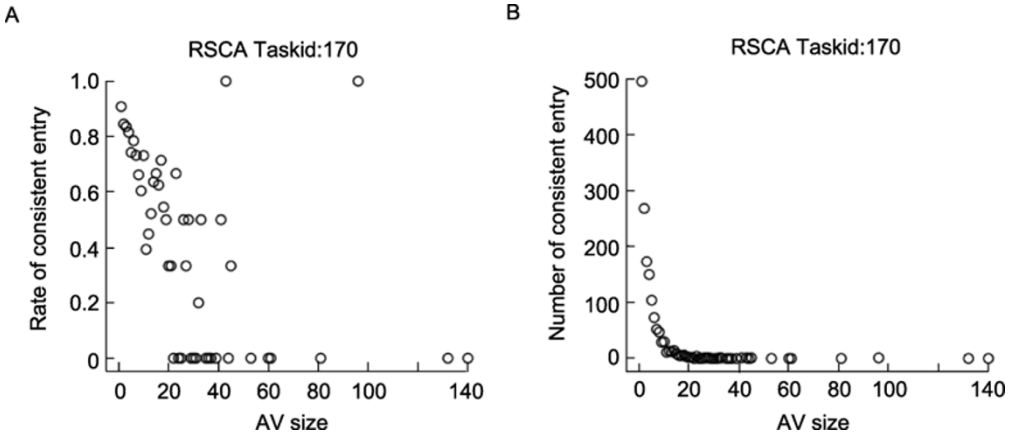

Figure 3. RSCA result for GenBank database. (A and B) RefSeq-link approach.

Table 2 Reference sequence mapping results

	GenBank		UniProtKB	
	RefSeq	Tracing	RefSeq	Tracing
Mapped	1515**	1516	1373**	1515
In-consistency	417	416	558	416
Total	1932		1931	

** $p < 0.001$.

which influences all of the existing mutations. The CAA40279.1 entry was created in 1998 and replaced by CAA40279.2 in 2008. However, all of the mutation reports were published before 2005.

For the RefSeq-link and Sequence tracing approach to the UniProtKB database, the trend is similar with that of GenBank (Supplementary data 2).

The number of mapped and inconsistent OMIM entries of each analysis are shown in Table 2. For the RefSeq-link approach to the UniProtKB database, the results are very significantly different from that of GenBank ($p < 0.001$).

Totally, based on results of the Sequence tracing approach for both databases, there are 1483 OMIM entries that were mapped in both databases. There are 33 and 32 entries that could only be mapped in GenBank or UniProtKB separately. Therefore, totally 1548 entries (79.8%) were mapped onto a single consensus sequence among the 1939 OMIM entries used for the analysis.

Moreover, 391 OMIM entries could not mapped onto any single sequence. We manually checked the data to check the reason for the inconsistency and found that the peptide sequence offset and reference sequence confliction are the major reasons. Other reasons include N-terminal methionine offset, mature protein offset and typing error.

DISCUSSION

Our results show that currently tracing sequences based on the amino or nucleotide acid variance information is tedious and sometimes prone to error. As we have shown above, there are several entries which have several hundred reported mutations. Actually, based on the report from HbVar (Giardine et al., 2007), the number is probably higher. At the same time, the reported mutation numbers may be even greater for some other genes, such as the LDL receptor gene (*LDLR*) (Cambien and Tiret, 2007).

Generally, biologists can link to a reference sequence to check the positions while using a central mutation database like the OMIM. In this paper, we automatically retrieve candidate sequences and mapping mutation data. This approach is commonly used by biologists. The results indicate biologists would have approximately a 20% possibility of obtaining *inconsistent* results by using a central or general mutation database for extracting sequences. Although a constant offset would be useful for finding the consistent reference sequence, such as a signal peptide or N-terminal methionine offset, this will not solve the inconsistency problem. The greater the mutation data is, the higher the rates of inconsistent results. This is mainly caused by the inconsistency of reference sequences used by the authors of the original literature.

Moreover, there is also inconsistency between sequence variances annotated by database and natural mutation published in literature. Currently, all of the sequence variances annotated in UniProtKB are based on the up-to-date sequence. However, our result indicates that 30% OMIM entry with protein point mutation could not be mapped onto the up-to-date sequence entry in UniProtKB. Considering that most of the mutagenesis data published in literature are included in UniProtKB annotated sequence variance, we can refer that there is also inconsistency between natural variant and mutagenesis types of mutations published in literature.

Although there are still other types of mutations not addressed in this study, such as insertion and deletion mutations in non-coding regions, the results for such data are predictable. Insertion data could not be checked using reference sequence consistency analysis. It will be helpful to use local flanking sequences to assign sequence features that will give us greater confidence in the alignment of sequences.

By using the methods mentioned in this article, in addition to other methods such as constant offset to the entire residue numbers and more sequences, about 80% of the entries in mutation databases like the OMIM could be mapped onto a consistent reference sequence.

Considering the expanding amount of mutation data in the genomic era, the earlier that a biologist can use a common reference sequence, the less inconsistency there will be. The capture of the reference sequence data from literature before or after publication might be helpful. Recently, several natural language tools have been developed to identify protein names and mutation terms from Medline abstract or full text articles, such as MutationFinder (Caporaso et al., 2007), MEMA (Rebholz-Schuhmann et al., 2004), mSTRAP (Kanagasabai et al., 2007) and MuteXt (Horn et al., 2004). Verifying the extracted mutation information based on corresponding position within the protein sequence will improve their performance. However, the inconsistency among different studies could not be solved by these approaches. Therefore, a consensus reference database like CCDS database (Ostell, 2009) is needed by genotype-phenotype studies, especially for the multicenter studies with sequencing techniques.

The results indicate biologists will have approximately a 20% possibility of obtaining inconsistent results when mapping the protein point mutation data from the OMIM database to a single reference sequence in public sequence database. This is mainly caused by sequence database changing or mature sequences used by the authors in the original literature. In order to solve this problem in the future, a consensus

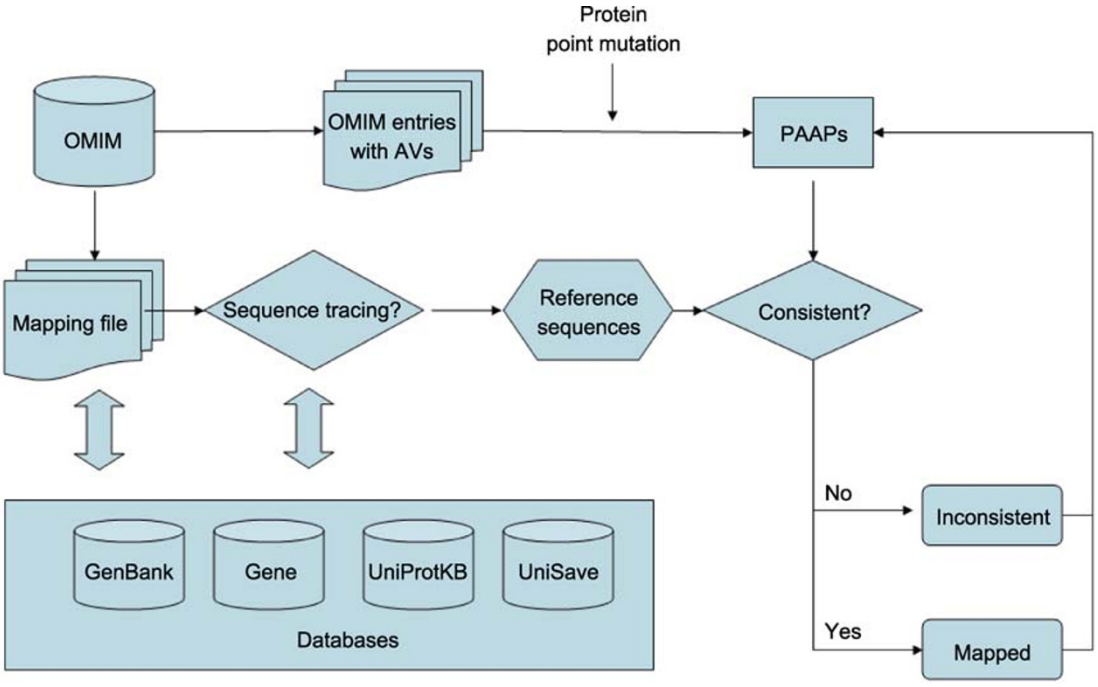

Figure 4. Schema for reference sequence consistency analysis.

reference database is needed.

MATERIALS AND METHODS

Study design

We accessed the OMIM database on 30 April 2009 by eUtils tool and deposited all OMIM entries that included an allelic variance record in a local SQLite database. Then, each allelic variance entry was parsed into extracted point mutation data, i.e. the position and amino acid pairs (PAAPs), with the regular expression in Perl syntax shown below:

/^(gly|ala|val|leu|ile|ser|thr|cys|met|asn|gln|asp|glu|lys|arg|phe|tyr|trp|his|pro)(\d+)(gly|ala|val|leu|ile|ser|thr|cys|met|asn|gln|asp|glu|lys|arg|phe|tyr|trp|his|pro|ter|del)$/i

We then explored two approaches for mapping the extracted point mutation data onto sequences. A flowchart of the two approaches appears as Fig. 4.

One method we used is called *RefSeq-link*. With the aid of the database mapping files, each OMIM ID was converted to sequence IDs in GenBank or Uniprot. Then, the database tools (eUtils or dbFetch) were employed to retrieve the up-to-date reference sequences from both databases. Then, all PAAPs under this entry were mapped onto the sequences. When all of the PAAPs could be mapped onto one sequence, the OMIM entry was placed into a *mapped* group; if not it was placed into an *inconsistent* group. This process is referred to as reference sequence consistency analysis (Fig. 4).

The other method we used is named *Sequence tracing*. Based on the sequences obtained by the method described in *RefSeq-link*, we tracked all of the versions of each entry in history by using Unisave for UniProtKB and dot version retrieve for GenBank entries. The group classification is the same as the *RefSeq-link* method.

Generally, biologists can link to reference sequences to check positions by using the provided links to up-to-date sequences while using a central mutation database like OMIM. In this study, we use the *reference-link* method to simulate this process and evaluate the suitability of the OMIM database for mutation sequence mapping. However, it is said that the reference sequence used by biologists may not be the final version. Therefore, we designed the sequence tracing method to automatically retrieve old versions of sequences for the reference sequence consistency analysis.

We applied the statistics package R to compare the results of Pearson's *Chi*-squared test with Yates' continuity correction.

ACKNOWLEDGEMENTS

This work was partly supported by grants from the Talents Developmental Fund of Shanghai in 2011 to ZF Li and the National High Technology Research and Development Program of China (863 Program) to XY Zhang (Grant Nos. 2007AA02Z332 and 2008AA02Z126).

ABBREVIATIONS

AV size, the number of allelic variance for each OMIM entry; CSF-1R, colony stimulating factor 1 receptor; HbVar, a database of human hemoglobin variants; HGMD, Human Gene Mutation Database; LDLR, LDL receptor gene; LSDB, locus-specific Database; OMIM, Online Mendelian Inheritance in Man; PAAP, the position and amino

acid pairs

REFERENCES

Alonso, G., Koegl, M., Mazurenko, N., and Courtneidge, S.A. (1995). Sequence requirements for binding of Src family tyrosine kinases to activated growth factor receptors. J Biol Chem 270, 9840–9848.

Cambien, F., and Tiret, L. (2007). Genetics of cardiovascular diseases: from single mutations to the whole genome. Circulation 116, 1714–1724.

Caporaso, J.G., Baumgartner, W.A. Jr, Randolph, D.A., Cohen, K.B., and Hunter, L. (2007). MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics 23, 1862–1865.

George, R.A., Smith, T.D., Callaghan, S., Hardman, L., Pierides, C., Horaitis, O., Wouters, M.A., and Cotton, R.G.H. (2008). General mutation databases: analysis and review. J Med Genet 45, 65–70.

Giardine, B., van Baal, S., Kaimakis, P., Riemer, C., Miller, W., Samara, M., Kollia, P., Anagnou, N.P., Chui, D.H.K., Wajcman, H., et al. (2007). HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. Hum Mutat 28, 206.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33, D514–D517.

Horaitis, O., Talbot, C.C. Jr, Phommarinh, M., Phillips, K.M., and Cotton, R.G.H. (2007). A database of locus-specific databases. Nat Genet 39, 425.

Horn, F., Lau, A.L., and Cohen, F.E. (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.

Kanagasabai, R., Choo, K.H., Ranganathan, S., and Baker, C.J.O. (2007). A workflow for mutation extraction and structure annotation. J Bioinform Comput Biol 5, 1319–1337.

Lee, A.W., and States, D.J. (2000). Both src-dependent and -independent mechanisms mediate phosphatidylinositol 3-kinase regulation of colony-stimulating factor 1-activated mitogen-activated protein kinases in myeloid progenitors. Mol Cell Biol 20, 6779–6798.

Leinonen, R., Nardone, F., Zhu, W., and Apweiler, R. (2006). UniSave: the UniProtKB sequence/annotation version database. Bioinformatics 22, 1284–1285.

Li, Zuofeng, Xingnan Liu, Jingran Wen, Ye Xu, Xin Zhao, Xuan Li, Lei Liu, and Xiaoyan Zhang. 2011. “DRUMS: A human disease related unique gene mutation search engine”. Human Mutation 32, E2259–E2265.

Ostell, J. (2009). Data Sharing: Standards for Bioinformatic Cross-Talk. Hum Mutat 30, vii–vii.

Rebholz-Schuhmann, D., Marcel, S., Albert, S., Tolle, R., Casari, G., and Kirsch, H. (2004). Automatic extraction of mutations from Medline and cross-validation with OMIM. Nucleic Acids Res 32, 135–142.

Tatusova, T.A., and Madden, T.L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174, 247–250.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. & Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. Nucl Acids Res 35(Database), D5–D12.

Xi, H., Park, J., Ding, G., Lee, Y.-H., and Li, Y. (2009). SysPIMP: the web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. Nucl Acids Res 37, D913–D920.

