General mutation databases: analysis and review

R A George,1 T D Smith,2,3 S Callaghan,4 L Hardman,2 C Pierides,2,3 O Horaitis,2 M A Wouters,1 R G H Cotton2,3

ABSTRACT

Databases of mutations causing Mendelian disease play a crucial role in research, diagnostic and genetic health care and can play a role in life and death decisions. These databases are thus heavily used, but only gene or locus specific databases have been previously reviewed for completeness, accuracy, currency and utility. We have performed a review of the various general mutation databases that derive their data from the published literature and locus specific databases. Only two—the Human Gene Mutation Database (HGMD) and Online Mendelian Inheritance in Man (OMIM)—had useful numbers of mutations. Comparison of a number of characteristics of these databases indicated substantial inconsistencies between the two databases that included absent genes and missing mutations. This situation strengthens the case for gene specific curation of mutations and the need for an overall plan for collection, curation, storage and release of mutation data.

The collection of lists of mutations causing single gene disorders began when the definition of globin gene mutations at the protein level became possible.1 It was then that Victor McKusick began collecting a compendium of inherited syndromes under the title Mendelian Inheritance in Man (MIM).2 After the advent of DNA sequencing, the rate of discovery of mutations accelerated and MIM began adding mutations to the compendium as they were characterised. Around the same time, David Cooper began specifically collecting genetic mutations to analyse their nature and frequency in order to find the most common sequence changes in humans.3 Databases of this type are referred to as general or central mutation databases (see Box 1 for definitions) and today these two are available online as Online Mendelian Inheritance in Man (OMIM)4,5 and the Human Gene Mutation Database (HGMD),6 respectively. These and a number of newer general mutation databases reviewed in this study are summarised in table 1.

Although official usage statistics are not readily accessible on the websites, the information stored in these databases is extremely important and widely used. These databases are particularly useful for those dealing in mutations found in patients as the first questions asked are, “Is this sequence variation pathogenic?” and, “Has it been found before?” If others have seen such a change and have reported investigations showing it to be pathogenic, this makes the work of the enquiring laboratory or clinicians easier and more cost effective. Among many other uses,7 those studying the function of genes and their products are using the experiments of nature to define essential base or amino acid changes. Besides defining the actual mutations and indicating their source, general databases list other properties of the mutation and the patient concerned. Further, other features such as mutation maps and links, etc, may be included.

In contrast to general mutation databases, databases of mutations in individual genes (Locus Specific Databases, LSDBs) have been developed, beginning with the globin mutations.8 These databases are a rich source of information on the gene itself and its mutations, and have been referred to as knowledge bases.9–11 A recent survey indicated some 678 genes have databases12 that are curated by experts in each gene. In a 2002 review Claustres et al compared 30 data fields occurring over a sample of 100 representative databases.12

This review was based in part on earlier recommendations of ideal database content which produced an entry form for submitting variants to such a database (http://www.hgvs.org/entry.html).12–13

Because of the need for a rapid and complete access to mutation data we have reviewed a number of GMDBs to assess their content, currency and completeness.

METHODS

Comparison of all GMDBs in relation to a number of features

Seven GMDBs containing or providing access to mutations were listed (table 1) and compared according to the relevant criteria used by Claustres et al12 (see supplementary table 1, available online). Four other databases were included as controls or comparisons. These included the SNP databases: HGVbase16 and dbSNP;17 and two LSDBs: PAH18 and Fanconi Anaemia. Characteristics were assessed by one operator in October 2006.

Comparison of the main GMDBs containing mutations: OMIM and HGMD

All genes and their variations listed in OMIM were taken from the omim.txt and mim2gene downloadable files (ftp.ncbi.nih.gov; downloaded on 10 October 2006). By text-mining the omim.txt file we were able to group each variation into “insertion”; “deletion”; “mutation” (including those occurring in exons, introns and regulatory regions); and “other” categories. Programs written for text mining were developed in Perl5 and will be made available upon request.

We compared the variations in the “mutation” category extracted from OMIM to those in the HGMD commercial release (version 6.3, 30 September 2006). Mutation data held in HGMD were collected by querying the Missense/nonsense,
Splicing and Regulatory tables only in the HGMD database.

The nomenclature for many genes is often derived from individual researchers, resulting in many aliases for a single gene. To ensure that we compared like with like, genes in both OMIM and HGMD were converted to their Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) approved gene name. This was achieved by first converting each alias gene name to their unique ENTREZ identifier, by querying the NCBI text files: gene_info and gene_history. ENTREZ identifiers were then converted to their corresponding HGNC gene name, again using the NCBI files. For genes that have not yet been allocated a HGNC name, the ENTREZ gene name was retained.

Currency of data in HGMD in relation to mutations in the literature

Issues of Human Mutation, starting with those that coincide with the most recent HGMD public release, were reviewed to determine the currency of information in HGMD. Three different novel mutations reported in each issue (see supplementary tables 1 and 2, available online) were randomly selected and then searched for in HGMD. This process was undertaken on three separate occasions for the public version of HGMD: May 2004, May 2005 and January 2007. In January 2007 the same mutations were also assessed in the commercial release of HGMD (version 6.4, 15 December 2006).

Search for specific mutations in HOWDY, GeneCards, EBI, Mutation Discovery and GDB

Two specific mutations were taken from the literature and searched for in the databases other than OMIM, HGMD and their LSDBs. These were a PAH mutation p.R158Q\(^{20}\) and PKD2 p.Q405X.\(^{21}\) The presence or absence of each mutation was recorded.

Number of mutations in GMDBs compared with the number in LSDBs

Four LSDBs of varying size were randomly selected and the number of mutations and polymorphisms found in each were recorded. These were then compared against mutation entries found in GMDs, including the HGMD public release.

RESULTS

Comparison of all GMDBs in relation to a number of features

Of the general databases checked against our 68 criteria (see supplementary table 1, available online) those that display mutations are: GDB, HGMD, OMIM and Mutation Discovery (although some polymorphisms are included as well in these databases). Those displaying both mutations and SNPs are: EBI, GeneCards and HOWDY, and those displaying SNPs alone, which we studied for comparative purposes, are HGVbase and dbSNP (although dbSNP does include a limited number of mutations).

Table 1 General and locus specific databases studied

<table>
<thead>
<tr>
<th>Database</th>
<th>Address</th>
<th>Broad features</th>
<th>Currency of information (as at 27 October 2006)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGMD</td>
<td><a href="http://www.hgmd.org">http://www.hgmd.org</a></td>
<td>Collects all mutations in all genes from public sources</td>
<td>Quarterly release</td>
</tr>
<tr>
<td>GDB</td>
<td><a href="http://www.gdb.org">http://www.gdb.org</a></td>
<td>Initially collected mutations</td>
<td>11 August 2005</td>
</tr>
<tr>
<td>GeneCards</td>
<td><a href="http://www.genecards.org">http://www.genecards.org</a></td>
<td>Integrates fragments of information from specialised databases</td>
<td>9 July 2006</td>
</tr>
<tr>
<td>Mutation Discovery</td>
<td><a href="http://www.mutationdiscovery.com">http://www.mutationdiscovery.com</a></td>
<td>Initiated to collect DHPLC data</td>
<td>25 August 2003</td>
</tr>
<tr>
<td>EBI Mutation Database</td>
<td><a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a></td>
<td>Now defunct</td>
<td>NA</td>
</tr>
<tr>
<td>Howdy</td>
<td><a href="http://howdy.jst.go.jp">http://howdy.jst.go.jp</a></td>
<td>A system to retrieve human genome information from different public resources</td>
<td>Updated daily</td>
</tr>
<tr>
<td>HGVbase</td>
<td><a href="http://hgvbase.cgb.ki.se">http://hgvbase.cgb.ki.se</a></td>
<td>Phenotype/genotype database</td>
<td>23 July 2003</td>
</tr>
<tr>
<td>PAH</td>
<td><a href="http://www.pahdb.mcgill.ca">http://www.pahdb.mcgill.ca</a></td>
<td>Locus specific database for PAH</td>
<td>3 August 2006</td>
</tr>
<tr>
<td>Fanconi Anemia</td>
<td><a href="http://www.rockefeller.edu/fanconi/mutate">http://www.rockefeller.edu/fanconi/mutate</a></td>
<td>Locus specific database for genes associated with Fanconi Anemia</td>
<td>?</td>
</tr>
</tbody>
</table>
The studied databases can be divided into three categories:

1. Those that collect and curate data—OMIM, HGMD and GDB
2. Those merely acting as a conduit—EBI, GeneCards and HOWDY
3. Those that both collect and act as a conduit—Mutation Discovery

Databases displaying mutations that are clearly currently active are GeneCards, HOWDY, OMIM and HGMD. Both OMIM and HGMD are regularly updated, but it is not clear how often GeneCards and Howdy are updated.

All active databases displaying mutations, except GeneCards, HOWDY and OMIM, allow submission by users, but no indication is given as to which mutations have been submitted or altered what criteria or review structure such submissions are accepted. Some sort of history akin to what is available for entries in Wikipedia (www.wikipedia.org) would be desirable.

It is difficult to compare the 68 criteria across the seven databases, suffice to say the quality is patchy with no one database being judged perfect with a correct/expected entry for each criterion. Further criteria were assessed and compared between HGMD and OMIM because they are both highly used and both collect original data. Because HOWDY is a conduit database only, OMIM and HGMD received the most critical appraisal.

Comparison of the main GMDBs containing mutations: OMIM and HGMD

OMIM contains over 12 000 genes with disease association; however, HGMD contains only those genes that have variation data. Here we compare only those genes in OMIM and HGMD that have mutation data.

The number of mutations listed for each particular gene was the focus of a detailed study. OMIM (as at 10 October 2006) had mutation data for 1790 genes (11 392 mutations), whereas mutations were a single base change within an exon, intron or regulatory region of the gene. In comparison, HGMD (commercial release version 6.5) had mutation data for 2120 genes (43 627 mutations). Figure 1 details the comparison of entries in OMIM and HGMD.

Because OMIM only collects the first and notable mutations, HGMD should always show more mutations per gene in every gene. However, this is not always the case, with 226 genes in OMIM having more mutations per gene than the same gene in HGMD. As expected, 1587 genes in HGMD show more mutations per gene than the same genes in OMIM. It is interesting that there are 143 genes with mutations in OMIM that are not present in HGMD; examples include COL9A1 and PTCH2 (COL9A1 was added to HGMD after the survey). The origin of these discrepancies is not clear and indicates one of the problems for users relying on one or both systems.

Other inconsistencies include missing genes or the complete absence of mutations for a gene. At the time of analysis 17 genes with mutations were present in HGMD but the same genes were absent in OMIM; examples include CYP4F12 and SRPX2. However, SRPX2 can now be found in OMIM, entered in March 2007. At the mutational level, 457 genes that have mutations listed in HGMD do not have mutations recorded in OMIM, even though the gene can be found in OMIM; examples include ACAD8 and PAX1.

Some of the annotated genes in HGMD have variation data other than single base mutations. Consistently, where a gene in HGMD only has variation data that are not simple base mutations, no single base mutations were noted in OMIM.

Currency of data in HGMD in relation to mutations in the literature

Mutations published in Human Mutation make up the major data source for the HGMD database. At three times, we sampled up to three novel mutations in successive earlier issues of Human Mutation, starting from the issue that was published on the survey date, and looked to see if they were present in the public release of the HGMD database (see supplementary table 2 and its summary in supplementary table 3, available online).

At each time point that this survey was conducted, it was found that a number of mutations reported in the issues of Human Mutation were missing. For example, two mutations in the 2002 issues of Human Mutation 19(5) and 19(3) did not appear in HGMD in the May 2005 survey. It is also interesting to note that, in some cases, not only was the mutation missing from the HGMD database, but the gene itself was also missing. This would seem to indicate an incomplete collection of data from the published literature.

As HGMD relies on commercial funding to finance its activities, it has adopted a strategy of charging for access to the most up-to-date version of its data. For this reason we also conducted a search for mutations in the HGMD commercial release, the results of which are shown in supplementary table 3 (available online). Clearly, in January 2007, the commercial version was 1–2 months behind currency and the public version around 18 months behind, up from the 14 months in the...
Review

previous surveys. The patchiness of genes and mutations covered is also evident in the commercial release, so the quality of the public data is just as good as the commercial release, if somewhat dated.

Figure 2 illustrates the time taken for a mutation to be entered into HGMD after publication. Although mutations are entered continually, this information is only available to users in bursts upon a new database release. Mutations that appear in the database before their publication in Human Mutation most likely represent mutations that have been republished. For example, a new publication was entered for the mutations recorded in the gene ATP7B after the LSDB for this gene updated its source publication.

It is worth noting that it was not always clear if a particular mutation found in Human Mutation had been included in the HGMD collection due to name changes in moving from the published data to HGMD. For example, mutations may have been reported in Human Mutation with a name based on the genomic sequence but, due to the naming conventions employed by HGMD, the identifier was converted to a cDNA name when included in the database. In other cases, the mutations were present in the HGMD database, but a journal other than Human Mutation was cited as the source, possibly indicating publication by multiple groups.

Search for specific mutations in HOWDY, GeneCards, EBI, Mutation Discovery and GDB

We searched for the specific mutations PAH p.R158Q and PKD2 p.Q405X, published in Human Mutation in 2000 and 2001, respectively, in the various mutation databases. At the time, none of the five databases showed the p.Q405X mutation. HGMD, OMIM and PAH LSDB do have an entry for p.R158Q but cite earlier research articles. The reference used by the PAH LSDB is an initial report of the allele, but the database should consider referencing a second paper by the same authors (Okano et al) that discusses the precise mutation.

Number of mutations in GMDBs compared with the number in LSDBs

As expected, the LSDBs have more mutations than HGMD, and HGMD has more mutations than OMIM. The other four GMDs are clearly further behind and dbSNP only links a small number of its listed variations to OMIM (table 2).

Data presentation for mutations in LSDBs, HGMD and OMIM

Across the range of GMDBs, the method employed to display the mutation and associated data differs greatly (fig 3). At the two extremes of this scale are OMIM and HGMD. OMIM lists variants and their associated information in textual form, allowing phenotypic effects and discovery information to be clearly described and easily read. However, the layout prohibits efficient computational data mining. HGMD, on the other hand, relies exclusively on tabular layouts, presenting its data in a series of columns. Extra information is obtained through the listed references, either to published articles or LSDBs. HGMD is held in a MySQL database (licence required), which makes the data easily accessible through simple database queries.

The most notable disadvantage of HGMD is the lack of adherence to the Human Genome Variation Society (HGVS) recommended mutation nomenclature (http://www.hgvs.org/mutnomen/), which perhaps explains some of the problems in the survey and the extra work required to find particular mutations in HGMD. Under the HGVS nomenclature it is possible to name any variant using three frames of reference for positional number: genomic DNA position, cDNA position and amino acid position. HGMD only provides the codon number to position exonic mutations and the IVS numbering system for intronic mutations.

Data presentation for OMIM has traditionally been textual using the Human Genome Variation Society nomenclature. The PAH LSDB also uses this nomenclature.

DISCUSSION

Databases that summarise information and contain or point to original or other sources are an essential part of today’s data driven world. Currency and accuracy is particularly crucial in the area of human health as there can be life or death issues involved. Because of the important role that databases of mutations play in genetic healthcare or research, central databases containing mutations were reviewed for content. As “controls” two primarily SNP databases and two LSDBs were also reviewed (see supplementary table 1, available online).

To assess utility against an “ideal database” we chose some of the relevant characteristics from another review, Claustres et al, and applied these to the 11 databases. The survey was difficult and perhaps qualitative because only partial replies could be given in some cases. Nevertheless, some valuable conclusions can be drawn to provide a perspective for those dealing with mutations causing single gene disorders.

All 11 databases were given a score by counting the number of “yes” entries. This clearly can only give rise to a qualitative figure for comparison because certain characteristics should be weighted—for example, nomenclature or public availability—and some boxes contained partial agreement with the question. Nevertheless, it is interesting that among all the mutation displaying databases reviewed, a “gold standard/prototyph” database—the phenylalanine hydroxylase (PAH) LSDB—comes out on top, as perhaps expected, and interestingly GeneCards came out second. However, if mutation content had been scored and weighted, GeneCards would have scored much lower. Notionally the ideal general database should apply all the desirable characteristics of an LSDB to all genes with reported variants.

With regard to content, for sequence variants that cause disorders, the most widely used GMDs are HGMD and OMIM. All other general databases have insignificant numbers of mutations in “all genes”. The mutation section of EBI, which employed the SRS system to harvest mutations from LSDBs, contained some 50 genes, but this section seems to have disappeared between October 2006 and March 2007. Mutation Discovery was initiated by Transgenomic Inc as a service and a portal for those using dHPLC instruments, but was last updated.

Table 2

<table>
<thead>
<tr>
<th>Database</th>
<th>PAH</th>
<th>PKD2</th>
<th>AR</th>
<th>CFTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSDB*</td>
<td>528</td>
<td>308</td>
<td>514</td>
<td>1529</td>
</tr>
<tr>
<td>OMIM</td>
<td>59</td>
<td>5</td>
<td>50</td>
<td>136</td>
</tr>
<tr>
<td>HGMD</td>
<td>444, 44</td>
<td>289</td>
<td>1208</td>
<td></td>
</tr>
<tr>
<td>HOWDY</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>NF</td>
</tr>
<tr>
<td>GeneCards</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mutation Discovery</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>GDB</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>dbSNP (linked to OMIM)</td>
<td>350 (5)</td>
<td>273 (0)</td>
<td>402 (0)</td>
<td>515 (3)</td>
</tr>
</tbody>
</table>

NF, not found.

The currency of the available HGMD commercial release is well ahead of the currency of the public release, with the data appearing 1–2 months and around 18 months post-publication, respectively. This is not a problem if users can afford to access the commercial release. However, public benefit must be balanced with commercial need. The Protein Data Bank allows researchers to quarantine commercially valuable structures for up to 12 months before they are released to the public, and recently the National Institute of Health announced that scientists contributing data from genome-wide association studies would be granted data exclusivity for up to 12 months. Given the fast pace of discovery it would be desirable for the public version of HGMD to be no further than 12 months behind the commercial release. This situation was recently ameliorated by the introduction of a discounted academic licence fee for HGMD, allowing access to the most recent version.

In conclusion, this survey has indicated that while the two major GMDBs is that it collects only a single instance of any mutation, which is usually the first published account. Any subsequent published account of the same mutation is not recorded. However, different individuals harbouring exactly the same mutation might not have identical phenotypes. This is especially true of dominant mutations and mutations causing variable expression.

The archiving of mutation data is a complex and laborious task but one that has enormous social impact and benefit. The visionary efforts to set up these databases and the labour of maintaining them should be applauded. However, efforts to improve them must continue. This has been difficult from several perspectives. The exponential growth of data, particularly in the light of recent association studies, has been a problem. With this growth of knowledge our understanding of the relationship between genotype and phenotype is gradually changing, bringing new challenges for database design.

In conclusion, this survey has indicated that while the two major GMDBs are clearly useful and well used, they have some characteristics that are less than ideal. These are: mutation nomenclature usage; failure to assimilate all available data; and lack of important characteristics usually confined to LSDBs. Of these adverse characteristics derive from historical and financial realities and are readily explained, but some, such as missing genes and mutations, are not. It is unrealistic to expect curators of large GMDBs to be experts on all genes. Indeed, a resolution of a meeting in 1994 of some of the world’s prominent geneticists concluded that mutations in genes were best curated by experts in each gene.11 We need to ensure that efforts to curate individual genes are utilised rather than the
GMDBs trying to collect the data de novo. Ideally a GMDB would transfer data from existing online LSDBs. This is already occurring to some extent. However, gene curation is time consuming and expensive and the Human Variome Project aims to address these problems. Data in publications and databases needs to be correct as life and death decisions rest on them, and this can best be ensured by curators.

Acknowledgements: The NHMRC (RC), Ronald Geoffrey Arnott Foundation (RG), Victorian Bioinformatics Consortium (SC) and Helen Smbert Vacation Studentship (TS) supported this work. The authors would also like to thank Conover Talbot for his critical comments.

Competing interests: None

REFERENCES