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 Cotton[2,3]

► Additional tables are published online only at http://jmg.bmj.com/content/vol45/issue2

[1] Structural and Computational Biology Program, Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales, Australia; [2] Genomic Disorders Research Centre, St Vincent's Hospital Melbourne, Fitzroy, Victoria, Australia; [3] Department of Medicine, The University of Melbourne, Melbourne, Victoria, Australia; [4] Victorian Bioinformatics Consortium, Monash University, Melbourne, Victoria, Australia

Correspondence to: Dr Richard George, Structural and Computational Biology Program, Victor Chang Cardiac Research Institute, 384 Victoria Street, Darlinghurst, NSW 2010, Australia; r.george@victorchang.edu.au

S Callaghan is deceased

Received 29 June 2007
Revised 5 September 2007
Accepted 9 September 2007
Published Online First 24 September 2007

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)[4–6] and the Human Gene Mutation Database (HGMD),[7] 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,[9] 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.[10] These databases are a rich source of information on the gene itself and its mutations, and have been referred to as knowledge bases.[11–13] A recent survey indicated some 678 genes have databases[14] that are curated by experts in each gene. In a 2002 review Claustres *et al* compared 80 data fields occurring over a sample of 100 representative databases.[15] 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 al*[15] (see supplementary table 1, available online). Four other databases were included as controls or comparisons. These included the SNP databases: HGVbase[16] and dbSNP;[17] and two LSDBs: *PAH*[18] 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,

Box 1: Definitions (Cotton and Scriver 1998)[8]

- **Mutation:** Base changes shown to cause single gene, or Mendelian, disorders. This usage has been general in clinical practice.
- **Polymorphism:** Base changes having no clinical effect. This usage has also been general in clinical practice.
- **Sequence variant:** A recently preferred option for any base change pathogenic at one extreme and with no effect at the other. Its nature is specified as either causing or not causing functional or pathological consequences.
- **Single nucleotide polymorphism (SNP):** Initially coined to refer to single base changes of little or no functional consequence, which were sought by researchers for use as aids in gene mapping, common disease and pharmacogenomic studies. The term certainly did not include mutations that cause single gene disorders. It appears that usage has strayed from that initial definition both from single nucleotide changes to all sequence variants, and whether or not they cause single gene disorders.
- **General mutation database (GMDB):** Sometimes referred to as central mutation databases, GMDBs strive to collect mutations in all genes and curate them centrally—for example, HGMD and OMIM.
- **Locus specific database (LSDB):** Databases of mutations, and relevant polymorphisms, in a single gene that are curated by an expert or experts in that gene. Sometimes the curator may manage several genes they are researching or diagnosing.
- **Conduit mutation database (CMDB):** Displays mutations by linking to LSDBs or GMDBs—for example, HOWDY.

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.[19] 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 GMDBs, 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

Database	Address	Broad features	Currency of information (as at 27 October 2006)
HGMD	http://www.hgmd.org	Collects all mutations in all genes from public sources	Quarterly release
OMIM	http://www.ncbi.nlm.nih.gov/entrez/query. fcgi?db = OMIM	First and interesting mutations with references. Initially collected mutations	Updated daily
GDB	http://www.gdb.org	Initially collected mutations	11 August 2005
GeneCards	http://www.genecards.org	Integrates fragments of information from specialised databases	9 July 2006
Mutation Discovery	http://www.mutationdiscovery.com	Initiated to collect DHPLC data	25 August 2003
EBI Mutation Database	http://www.ebi.ac.uk	Now defunct	NA
Howdy	http://howdy.jst.go.jp	A system to retrieve human genome information from different public resources	Updated daily
HGVbase	http://hgvbase.cgb.ki.se	Phenotype/genotype database	23 July 2003
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP	Catalogue of SNPs	18 May 2006
PAH	http://www.pahdb.mcgill.ca	Locus specific database for *PAH*	3 August 2006
Fanconi Anemia	http://www.rockefeller.edu/fanconi/mutate	Locus specific database for genes associated with Fanconi Anemia	?

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 under 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), where a mutation is a single base change within an exon, intron or regulatory region of the gene. In comparison, HGMD (commercial release version 6.3) 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.

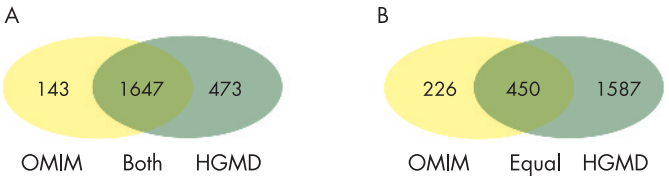

Figure 1 Analysis of genes with mutations annotated in Human Gene Mutation Database (HGMD) and Online Mendelian Inheritance in Man (OMIM) as of October 2006, where a mutation is a single base pair change within an exon, intron or regulatory region of a gene. (A) Venn diagram comparing the number of unique genes in OMIM and HGMD. (B) Venn diagram comparing the number of unique genes with more annotated mutations in one database than the other.

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.[7] 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

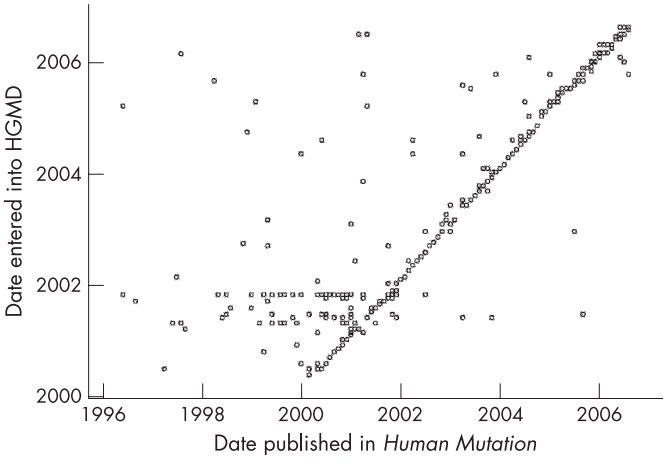

Figure 2 The date of publication versus the date of entry into Human Gene Mutation Database (HGMD) for mutations published in *Human Mutation* since May 2000. A mutation is a single base pair change within an exon, intron or regulatory region of the gene.

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,[20 21] 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.[22 23] The reference used by the *PAH* LSDB[23] is an initial report of the allele, but the database should consider referencing a second paper by the same authors (Okano *et al*[24]) 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 GMDBs 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.[25]

Data presentation for OMIM has traditionally been textual using the Human Genome Variation Society nomenclature.[25–27] The *PAH* LSDB also uses this nomenclature.

Table 2 Number of mutations in general mutation databases (GMDBs) compared with the number in specific locus specific databases (LSDBs)

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

NF, not found.
*Consulted LSDBs: *PAH* (http://www.pahdb.mcgill.ca/), *PKD2* (http://pkdb.mayo.edu), *AR* (http://androgendb.mcgill.ca) and *CFTR* (http://www.genet.sickkids.on.ca/cftr).

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*,[15] 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/prototype" 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 GMDBs 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 30 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

A

No.	Reference ID	Systematic name	Mutation name	Other name	Region	Mutation type	Length	CpG	Comments	Date
1	29	c.473G>A	p.R158Q	CGG/CAG	E5	Missense	1	Y		Not available

B

Accession number	Codon change	Amino acid change	Codon number	Phenotype	Reference
CM890093	CGG-CAG	Arg-Gln	158	Phenylketonuria	Dwomiczak (1989) Hum Genet 84, 95

C

0.0010 Phenylketonuria (PAH, ARG158GLN)

In 7 out of 94 PKU alleles, Dwomiczak et al. (1989) identified a G-to-A transition in nucleotide 695 in exon 5. Twenty-four per cent of the PKU alleles were in a background of haplotype 4; all 7 of the G-to-A transitions were on the haplotype 4 background.The base substitution predicted an arg158-to-gln change.

Figure 3 Comparison of data presentation techniques across (A) the phenylalanine hydroxylase (PAH) locus specific database, (B) Human Gene Mutation Database (HGMD), and (C) Online Mendelian Inheritance in Man (OMIM).

in 2003. GDB, initiated as part of the Human Genome Project, has not been updated since 2005. Although they are not current, GDB, EBI and Mutation Discovery may still have either useful software or other information. Only GeneCards, Howdy, HGMD and OMIM seem to be current and have regular, if sometimes infrequent, updates. Of the four current databases, only HGMD and OMIM have useful numbers of mutations. GeneCards and Howdy seem to be mainly conduit databases—that is, they rely on links to the other collections of mutations. Anecdotally, it is well known that LSDBs usually contain more mutations than the two main GMDBs, HGMD and OMIM, and this has been confirmed.

Notable are the small numbers of SNPs in dbSNP that have links to OMIM, indicating that they are variations causing Mendelian disease. It should be kept in mind that dbSNP and the other mutation databases contain very different data. dbSNP aims to capture common allelic variants in the human population.[17] The mutation databases, on the other hand, contain mutation data for Mendelian disease in specific families that may be rare. dbSNP is often used by researchers to determine whether a sequencing variant is in fact a rare mutation in their families or a harmless common variant found in the general population. Currently, OMIM has very high standards as to whether a disease associated SNP is included into the database. This might explain disease–gene associations in HGMD that are absent in OMIM. It is likely that as rarer allelic variants are catalogued by dbSNP and the genetic basis of complex disease emerges, the overlap between the two types of databases will increase.

Several observations can be made with regard to the quality of the data. Firstly, some data are missing from both of the major GMDBs. It was found that there were 226 genes in OMIM that had more single base mutations than in HGMD. This was surprising because HGMD aims to collect all mutations whereas OMIM addresses only the first and most interesting ones. Also, there were 143 genes in OMIM with mutations that were not present in HGMD—for example, *COL9A1* and *PTCH2*. These observations have no clear explanation and tend to reduce confidence because one would expect the two databases to be consistent with each other. Our study of HGMD found omissions of mutations, and in some cases, the gene itself, reported in *Human Mutation*, suggesting incomplete collection of data was at least part of the problem.

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[28] 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.

One major issue that distinguishes HGMD from the LSDBs 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,[29] 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. Some 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[30] and the Human Variome Project[31] 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 Smibert Vacation Studentship (TS) supported this work. The authors would also like to thank Conover Talbot for his critical comments.

Competing interests: None

REFERENCES

1. **Baglioni C.** The fusion of two peptide chains in hemoglobin lepore and its interpretation as a genetic deletion. *Proc Nat Acad Sci* 1962;**48**:1880–6.
2. **McKusick VA.** *Mendelian inheritance in man. A catalogue of human genes and genetic disorders*. Baltimore, Maryland: Johns Hopkins University Press, 1998.
3. **Cooper DN,** Krawczak M. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum Genet* 1990;**85**:55–74.
4. **Hamosh A,** Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl Acids Res* 2002;**30**:52–55.
5. **Hamosh A,** Scott AF, Amberger J, Bocchini AC, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl Acid Res* 2005;**33**:514–17.
6. **Hamosh A,** Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* 2000;**15**:57–61.
7. **Stenson PD,** Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 2003;**21**:577–81.
8. **Cotton, R, Scriver C**. Proof of disease causing mutation. *Hum Mutat* 1998;**12**:1–3.
9. **Cotton, RG, Horaitis O**. The HUGO mutation database initiative. Human genome organization. *Pharmacogenomics* 2002;**2**:16–19.
10. **Huisman TH,** Carver MF, Efremov GD. *A syllabus of human haemoglobin variants*. Augusta, Georgia: The Sickle Cell Anaemia Foundation, 1996.
11. **Scriver CR,** Cotton RG, Antonarakis S, McKusick VA. Mutation database initiative to go forward under HUGO. *Genome Digest* 1997;**4**:12–15.
12. **Scriver CR,** Nowacki PM, Lehväslaiho H. Guidelines and recommendations for content, structure and deployment of mutation databases. *Hum Mutat* 1999;**13**:344–50.
13. **Scriver CR,** Nowacki PM, Lehväslaiho H. Guidelines and recommendations for content, structure, and deployment of mutation databases: II. Journey in progress. *Hum Mutat* 2000;**15**:13–15.
14. **Horaitis O,** Talbot CC Jr, Phommarinh M, Phillips KM, Cotton RG. A database of locus-specific databases. *Nat Genet* 2007;**39**:425
15. **Claustres M,** Horaitis O, Vanevski M, Cotton RGH. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 2002;**12**:680–8.
16. **Fredman D,** Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucl Acids Res* 2002;**30**:387–91.
17. **Sherry ST,** Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucl Acids Res* 2001;**29**:308–11.
18. **Scriver CR,** Hurtubise M, Konecki D, Phommarinh M, Prevost L, Erlandsen H, Stevens R, Waters PJ, Ryan S, McDonald D, Sarkissian C. PAHdb 2003: what a locus-specific knowledgebase can do. *Hum Mutat* 2003;**21**:333–44.
19. **White JA,** McAlpine PJ, Antonarakis S, Cann H, Eppig JP, *et al*. HUGO Nomenclature Committee. *Genomics* 1997;**45**:468–71.
20. **Hennermann JB,** Vetter B, Wolf C, Windt E, Bührdel P, Seidel J, Mönch E, Kulozik AE. Phenylketonuria and hyperphenylalaninemia in eastern Germany: a characteristic molecular profile and 15 novel mutations. *Hum Mutat* 2000;**15**:254–60.
21. **Deltas CC.** Mutations of the human polycystic kidney disease 2 (PKD2) gene. *Hum Mutat* 2001;**18**:13–24.
22. **Dworniczak B,** Aulehla-Scholz C, Horst J. Phenylketonuria: detection of a frequent haplotype 4 allele mutation. *Hum Genet* 1989;**84**:95–6.
23. **Okano Y,** Wang T, Eisensmith RC, Güttler F, Woo SL. Recurrent mutation in the human phenylalanine hydroxylase gene. *Am J Hum Genet* 1990;**46**:919–24.
24. **Okano Y,** Wang T, Eisensmith RC, Güttler F, Woo SL. Missense mutations associated with RFLP haplotypes 1 and 4 of the human phenylalanine hydroxylase gene. *Am J Hum Genet* 1990;**46**:18–25.
25. **Antonarakis SE,** The Nomenclature Working Group. Recommendations for a nomenclature system for human gene mutations. *Hum Mutat* 1998;**11**:1–3.
26. **den Dunnen JT,** Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 2000;**15**:7–12.
27. **den Dunnen JT,** Paalman MH. Standardizing mutation nomenclature: why bother? *Hum Mutat* 2003;**22**:181–2.
28. **Berman H,** Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 2003;**10**:980.
29. **Smyth DJ,** Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, *et al*. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature Genet* 2006;38:617–19.
30. **Cotton RG,** Phillips K, Horaitis O. A survey of locus-specific database curation. Human Genome Variation Society. *J Med Genet* 2007;**44**:e72.
31. **Cotton RG,** Kazazian HH Jr. Toward a human variome project. *Hum Mutat* 2005;**26**:499.

Supplementary Table 3 – Summary of mutations published in *Human Mutation* sampled in HGMD at three time points

		May 2004			May 2005			January 2007			January 2007 (commercial)		
	Issue	1	2	3	1	2	3	1	2	3	1	2	3
12/2006	27 (12)	-	-	-	-	-	-	-	-	-	-	-	-
	27 (11)	-	-	-	-	-	-		▼			▼	✓
	27 (10)	-	-	-	-	-	-		○		✓	○	✓
	27 (9)	-	-	-	-	-	-		○		✓	○	○
	27 (8)	-	-	-	-	-	-				✓	✓	✓
	27 (7)	-	-	-	-	-	-				✓		✓
	27 (6)	-	-	-	-	-	-		○		✓	○	✓
	27 (5)	-	-	-	-	-	-	-	-	-	-	-	-
	27 (4)	-	-	-	-	-	-				✓	✓	✓
	27 (3)	-	-	-	-	-	-				✓	✓	✓
	27 (2)	-	-	-	-	-	-			▼	✓	✓	▼
1/2006	27 (1)	-	-	-	-	-	-	○	○		○	○	✓
	26 (6)	-	-	-	-	-	-				✓	✓	✓
	26 (5)	-	-	-	-	-	-				✓		✓
	26 (4)	-	-	-	-	-	-		▼		✓	▼	✓
	26 (3)	-	-	-	-	-	-		✓		✓	✓	✓
	26 (2)	-	-	-	-	-	-		•		✓	•	✓
	26 (1)	-	-	-	-	-	-			•	✓	✓	•
	25 (6)	-	-	-	-	-	-	•	✓	✓	•	✓	✓
	25 (5) S	-	-	-	-	-	-				✓	✓	✓
	25 (4)	-	-	-	▼			-	-	-	-	-	-
	25 (3)	-	-	-				-	-	-	-	-	-
	25 (2)	-	-	-				-	-	-	-	-	-
1/2005	25 (1)	-	-	-	▼			-	-	-	-	-	-

	24 (6)	-	-	-				-	-	-	-	-	-
	24 (5)	-	-	-			▼	-	-	-	-	-	-
	24 (4)	-	-	-				-	-	-	-	-	-
	24 (3)	-	-	-				-	-	-	-	-	-
	24 (2)	-	-	-				-	-	-	-	-	-
	24 (1)	-	-	-				-	-	-	-	-	-
	23 (6)	-	-	-	▼		•	-	-	-	-	-	-
	23 (5) S	-	-	-				-	-	-	-	-	-
	23 (4)	-	-	-				-	-	-	-	-	-
	23 (3)	-	-	-	✓	✓		-	-	-	-	-	-
	23 (2)	-	-	-	•	✓		-	-	-	-	-	-
1/2004	23 (1)				✓	✓		-	-	-	-	-	-
	22 (6)				✓	✓		-	-	-	-	-	-
	22 (5)				✓	✓	✓	-	-	-	-	-	-
	22 (4)					✓	✓	-	-	-	-	-	-
	22 (3)		-	-	✓	-	-	-	-	-	-	-	-
	22 (2)				✓	✓	✓	-	-	-	-	-	-
	22 (1)	-	-	-	✓			-	-	-	-	-	-
	21 (6)			-	✓	✓	-	-	-	-	-	-	-
	21 (5)				✓	✓	✓	-	-	-	-	-	-
	21 (4)				✓		✓	-	-	-	-	-	-
	21 (3)	-	-	-	-	-	-	-	-	-	-	-	-
	21 (2)		✓	✓		✓	✓	-	-	-	-	-	-
1/2003	21 (1)	✓	-	-	✓	-	-	-	-	-	-	-	-
	20 (6)	✓		○	✓	✓	○	-	-	-	-	-	-
	20 (5)	○	✓	✓	○	✓	✓	-	-	-	-	-	-
	20 (4)	○	○		○	○		-	-	-	-	-	-
	20 (3)	✓	✓	○	✓	✓	○	-	-	-	-	-	-
	20 (2)	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-
	20 (1)	-	-	-	✓	✓	○	-	-	-	-	-	-

	19 (6)	-	-	-	✓	✓	-	-	-	-	-	-	-
	19 (5)	-	-	-	✓		▼	-	-	-	-	-	-
	19 (4)	-	-	-	✓	✓	✓	-	-	-	-	-	-
	19 (3)	-	-	-	✓	✓		-	-	-	-	-	-
	19 (2)	-	-	-	•	✓	✓	-	-	-	-	-	-
1/2002	19 (1)	-	-	-	✓	✓	✓	-	-	-	-	-	-

-	Not surveyed
○	Unable to determine due to inconsistency between nomenclature
▼	Gene not present in HGMD
•	Variant present but cited source different from *Human Mutation*
✓	Variant present
	Variant not present
S	Survey date