DATABASES

Human Gene Mutation Database (HGMD®): 2003 Update

Peter D. Stenson,[1] Edward V. Ball,[1] Matthew Mort,[1] Andrew D. Phillips,[1] Jacqueline A. Shiel,[1] Nick S.T. Thomas,[1] Shaun Abeysinghe,[1] Michael Krawczak,[2] and David N. Cooper[1]*

[1]*Institute of Medical Genetics, University of Wales College of Medicine, Heath Park, Cardiff, UK;* [2]*Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität, Kiel, Germany*

Communicated by Richard G.H. Cotton

The Human Gene Mutation Database (HGMD) constitutes a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease (www.hgmd.org). Data catalogued includes: single base-pair substitutions in coding, regulatory and splicing-relevant regions; micro-deletions and micro-insertions; indels; triplet repeat expansions as well as gross deletions; insertions; duplications; and complex rearrangements. Each mutation is entered into HGMD only once in order to avoid confusion between recurrent and identical-by-descent lesions. By March 2003, the database contained in excess of 39,415 different lesions detected in 1,516 different nuclear genes, with new entries currently accumulating at a rate exceeding 5,000 per annum. Since its inception, HGMD has been expanded to include cDNA reference sequences for more than 87% of listed genes, splice junction sequences, disease-associated and functional polymorphisms, as well as links to data present in publicly available online locus-specific mutation databases. Although HGMD has recently entered into a licensing agreement with Celera Genomics (Rockville, MD), mutation data will continue to be made freely available via the Internet. Hum Mutat 21:577–581, 2003. © 2003 Wiley-Liss, Inc.

KEY WORDS: HGMD; mutation database; inherited disease; human; bioinformatics

DATABASES: www.hgmd.org (HGMD)

DATABASE STRUCTURE AND CONTENT

The Human Gene Mutation Database (HGMD, www.hgmd.org) represents the only comprehensive collation of germline mutations underlying human inherited disease [Krawczak and Cooper, 1997; Cooper et al., 1998; Krawczak et al., 2000]. HGMD comprises published single base-pair substitutions in coding, regulatory and splicing-relevant regions of human nuclear genes, as well as deletions, duplications, insertions, repeat expansions, combined micro-insertions/deletions (indels), and a number of different types of complex rearrangement (Table 1) that occur in association with inherited disease. HGMD does not however include somatic lesions or mitochondrial genome mutations. The latter are well covered by MITOMAP [www.mitomap.org; Kogelnik et al., 1998], to which HGMD provides links for mitochondrial genes harboring known pathological mutations.

Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base occurs in either the first or third position in the triplet. Substitutions causing gene regulatory abnormalities are logged with 30 nucleotides flanking the site of mutation on both sides. The location of the mutation relative to the transcriptional initiation site, initiator ATG, or polyadenylation site is given. Mutations affecting mRNA splicing are presented in brief with information specifying the relative position of the lesion with respect to a numbered intron donor or acceptor splice site. Positions logged as positive integers refer to a 3′ (downstream) location and negative integers refer to a 5′ (upstream) location. Micro-deletions of 20 bp or less are presented in terms of the deleted bases in lower case and in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^). In cases where any location parameter is listed as "?", either the location is unknown or a consistent nucleotide/codon

Received 3 October 2002; accepted revised manuscript 29 January 2003.

*Correspondence to: David N. Cooper, Institute of Medical Genetics, University of Wales College of Medicine, Heath Park, Cardiff, CF14 4XN, UK. E-mail: cooperdn@cardiff.ac.uk

DOI 10.1002/humu.10212
Published online in Wiley InterScience (www.interscience.wiley.com).

TABLE 1. Summary of Mutation Data in HGMD, March 2003

Mutation type	Number of entries
Single base-pair substitutions	
Missense/nonsense	22,682
Splicing	3,783
Regulatory	370
Other lesions	
Micro-deletions (≤20 bp)	6,587
Micro-insertions (≤20 bp)	2,573
Indels (≤20 bp)	377
Gross (>20 bp) deletions	2,172
Gross (>20 bp) insertions and duplications	348
Complex rearrangements (including inversions)	452
Repeat variations	71
Total	39,415

Numbers include data not yet made available through the public website.

numbering system is lacking. Where deletions extend outside the coding region of the gene in question, other positional information is provided, e.g., 5′-UTR (5′-untranslated region) or E6I6 (denotes exon 6/intron 6 boundary). It should be noted that codon numbering may display inconsistencies with the literature in some cases because different residue numbering systems are adopted for the same protein. For most genes where there is no risk of error or ambiguity, residue numbering has been standardized with respect to the generally accepted numbering system. For gross deletions and insertions, duplications, repeat variations, and complex rearrangements, information regarding the nature, location, and extent of the lesion is logged in narrative form owing to the extremely variable quality of the data reported.

Mutation data in HGMD are accessible on the basis of every gene being allocated one web page per mutation type if data of that type are present. Meaningful integration with phenotypic, structural, and mapping information has been accomplished through bi-directional links between HGMD and both the Genome Database (GDB) [www.gdb.org/; Cuticchia, 2000] and Online Mendelian Inheritance in Man (OMIM) [www.ncbi.nlm.nih.gov/Omim/; Hamosh et al., 2002]. Links to GenAtlas [www.dsi.univ-paris5.fr/genatlas/; Frézal, 1998], the HUGO Nomenclature Committee [www.gene.ucl.ac.uk/nomenclature/; Povey et al., 2001], GeneCards [http://bioinformatics.weizmann.ac.il/cards; Rebhan et al., 1998], GeneClinics [www.geneclinics.org/; Pagon et al., 2002], and LocusLink [www.ncbi.nlm.nih.gov/LocusLink/; Pruitt and Maglott, 2001] have also been established.

Each mutation is entered into HGMD only once (citing the first literature report) in order to avoid confusion between recurrent and identical-by-descent lesions. Silent mutations within the coding region that do not alter the encoded amino acid are not recorded unless there is good evidence for altered splicing and/or a disease association. Mutations that have not been adequately or unambiguously described in the original report are also excluded unless full details can subsequently be obtained from the authors. Such problems could be minimized if authors were to strictly follow published mutation nomenclature guidelines [den Dunnen and Antonarakis, 2001; http://archive.uwcm.ac.uk/uwcm/mg/docs/mut_nom.html; also see information at the Human Genome Variation Society website: www.hgvs.org/mutnomen/]. In addition to mutation data, HGMD provides supplementary information that can be used to assist in data interpretation, e.g., links to cDNA reference sequences (currently numbering 1,316), mutation maps for coding sequence mutations, and splice junction data (for 60 genes). Missense and nonsense mutations and micro-deletions are automatically checked for accuracy and consistency against the cDNA reference sequences.

Data are obtained by means of a combined electronic and manual search procedure. HGMD currently contains data derived from more than 500 different life-science and medical journals, with entries accumulating at a rate exceeding 5,000 per annum (Table 2). Data from five journals, ranked by number of published mutations listed in HGMD, account for just over 50% of HGMD entries (Fig. 1): *Human Mutation* (5,167), *The American Journal of Human Genetics* (5,001), *Human Molecular Genetics* (3,206), *Human Genetics* (2,277), and *Nature Genetics* (2,267). The top 100 journals account for more than 95% of the entries (Fig. 1).

A law of diminishing returns in terms of journal coverage is apparent. Thus, some 20% of the total number of mutations currently listed in HGMD are distributed between approximately 435 of the journals scanned. We may safely assume that this proportion will tend to increase in the future as mutation data

TABLE 2. Summary of Entries in HGMD byYear, March 2003

Year	Number of entries
up to 1985	130
1986	52
1987	77
1988	139
1989	255
1990	403
1991	683
1992	1,145
1993	1,340
1994	2,282
1995	2,345
1996	2,939
1997	3,574
1998	4,415
1999	4,894
2000	5,181
2001	5,277
2002	3,882[a]
2003	402[a]

[a] Figure likely to increase as new data are acquired.

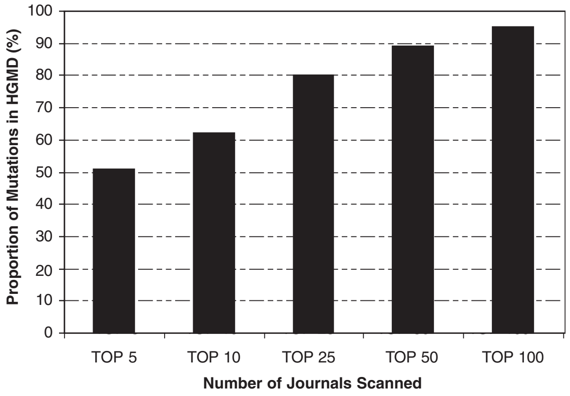

FIGURE 1. **Proportion of mutation entries in HGMD by the number of journals scanned.**

becomes published in an ever more diverse selection of journals covering specialized subject areas. This should ensure that the scanning of the more obscure journals continues to be a worthwhile proposition in terms of data acquisition. In recognition of this, we have recently begun to implement a broader search strategy than previously pursued. This involves an expanded computerized search of all articles listed in the PubMed literature database (www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=PubMed), in addition to the manual screening of a selection of core journals.

HGMD does not usually include mutations lacking obvious phenotypic consequences. However, some variants have been included on the basis that they significantly reduce the expression of a given gene or the functional activity of its protein product, even though these mutations may not have been shown yet to be of direct clinical relevance.

DISEASE-ASSOCIATED POLYMORPHISMS

Disease-associated polymorphisms currently comprise approximately 1.5% of mutation entries in HGMD. Most are single base-pair substitutions but a small number are of an insertion/deletion type. They generally occur within gene promoters or coding regions and can therefore serve to alter either the level of expression of the gene or the functional activity of the gene product in question. The distinction between a disease-associated polymorphism and a pathological mutation sensu stricto is inevitably fairly arbitrary to some extent.

Many reports of disease-associated polymorphisms that appear in the literature are of uncertain significance and may yet turn out to be unreliable as a result of inappropriate case-control matching, inadequate choice of statistical testing, inconsistent phenotype definition, etc. Therefore, the decision as to whether to include disease-associated polymorphisms in HGMD has involved an exercise of judgement. To be included, either a statistically significant ($p < 0.05$) association between the polymorphism and a clinical phenotype must have been observed, or in vitro or in vivo expression/functional data must have indicated that the polymorphism influences either the level of expression or the function of the protein product. In some instances, the above criteria have only been partially satisfied so that the HGMD curators remained unconvinced as to the phenotypic relevance of the variants reported. A decision to include the polymorphism may nevertheless have been made: either as a result of other supporting information that became available since publication of the first report; or because the associated gene/disease state was deemed to be of sufficient importance for it to warrant confirmation by further work. Such variants have been ascribed the descriptor "association with ?" (as opposed to "association with" without a question mark) to indicate that some degree of uncertainty is involved. The difficulty inherent in making decisions to include or exclude such potential disease associations highlights the need for a methodical and methodologically uniform approach to assessing such reports as they appear in the literature [Cooper et al., 2002].

COLLABORATION WITH CELERA GENOMICS

HGMD has undergone some significant changes over the last 2 years. In the summer of 2000, HGMD entered into a licensing agreement with Celera Genomics (Rockville, MD) (www.celera.com). As part of this agreement, the University of Wales College of Medicine in Cardiff agreed to provide Celera with a period of exclusive access to new information added to HGMD. This period currently extends to 1 year from the date of initial inclusion of mutation data. HGMD has therefore been made available along with the single nucleotide polymorphism (SNP) reference database as part of the Celera Discovery System™ (CDS) [www.celeradiscoverysystem.com; Kerlavage et al., 2002]. The Celera Human SNP Reference Database is a comprehensive database with over 4 million nonredundantly mapped variants distributed throughout the genome (on average, one variation every 700 bp). All variants are correlated with relevant genes, gene structure, and protein changes in order to aid the identification or validation of potential disease genes. It is a highly integrated database including SNPs discovered during the assembly of the genomes of five individuals of diverse ethnic backgrounds in addition to mutation and polymorphism data from HGMD, HGVbase, and dbSNP. An independently searchable version of HGMD has also been made available by Celera for its CDS subscribers. The data are presented in a very similar format to that already employed by HGMD. Celera is also evaluating making a stand-alone version of HGMD available for researchers who are not CDS subscribers. For further information, please refer to the Celera website.

The publicly available version of HGMD will continue to be maintained and made available free of charge, albeit with a time delay, via the Cardiff website. The public website also contains extra information such as a comprehensive listing of locus-specific mutation databases (http://archive.uwcm.ac.uk/uwcm/mg/docs/oth_mut.html), current guidelines for mutation nomenclature (http://archive.uwcm.ac.uk/uwcm/mg/docs/mut_nom.html, courtesy of den Dunnen and Antonarakis [2001]), and links to MITOMAP for mitochondrial genes with known pathological mutations.

ONLINE SUBMISSION OF MUTATION DATA

In collaboration with the journal *Human Genetics*, HGMD provides access to a free online submission system for human gene mutation data (http://link.springer.de/journals/humangen/mutation/form.htm). Mutation data submitted online via the aforementioned URL are assigned a temporary accession number before being checked for originality, accuracy, and uniformity. Once validated, *Human Genetics* formally accepts the submitted material for electronic publication and the accession number becomes permanent. After publication, the mutation data are transmitted to Cardiff for inclusion in HGMD.

Online submission is becoming an ever more important source of mutation data. *Human Mutation* used to publish Mutation and Polymorphism Reports (MPRs) online (http://interscience.wiley.com/jpages/1059-7794/mutnote1.html), although this was discontinued in October 2000. Recent developments within the HUGO Mutation Database Initiative (MDI) have led to the formation of the Human Genome Variation Society (HGVS; www.hgvs.org), with HGVbase (formerly HGbase) [http://hgvbase.cgb.ki.se/; Fredman et al., 2002], a central submission site for polymorphism data, acting as the proposed repository for variants submitted to HGVS. This implies that HGMD/*Human Genetics* currently represents the only central submission site available for human gene mutations of pathological significance. However, it is worth noting that many locus-specific mutation databases (LSDBs) also provide facilities for online submission of data, although only for genes in their own specific areas of interest.

The importance of online publication is likely to increase in the future. As yet, however, fewer than 5% of entries in HGMD have been originally brought to public attention through an online submission system (*Human Genetics*, *Human Mutation* MPRs, or an LSDB). Although the availability of a website accepting online submission of all mutation data should encourage electronic submission, for the foreseeable future there can be no doubt as to the continuing importance of the conventional biomedical literature as the main forum for publishing these data.

LOCUS-SPECIFIC MUTATION DATABASES

HGMD has recently established links to unpublished mutation data presented online by 100 publicly available locus-specific mutation databases (LSDBs), some of which contain data obtained via online submission. These links were established in order to provide HGMD users with ready access to these hitherto unpublished mutation data as well as to material published only in meeting abstracts or book chapters (not normally covered by HGMD). Although the 100 LSDBs covered constitute fewer than 40% of the total number of electronically available LSDBs [Claustres et al., 2002], the remaining 60% were not found to contain any mutation data omitted by HGMD. However, to put these data in their proper perspective, at present only approximately 1.5% of mutations contained in the public version of HGMD can be attributed exclusively to an LSDB.

Locus-specific mutation databases nevertheless have a vital role to play in the publication of, and subsequent access to, mutation data. These databases are curated by experts in the specific gene/disease covered by the corresponding LSDB. Therefore, they often include gene- or disease-specific information such as ethno-geographic origin, phenotypic sequelae, and population frequency data. However, LSDBs also have their shortcomings. A lack of uniform layout and content makes searching for mutations in multiple unrelated genes very difficult. An attempt to circumvent this problem has been implemented (www.ebi.ac.uk/mutations/), but has only been partially successful owing to the limited resources available as compared to the burgeoning number of available LSDBs.

Many LSDBs, once created, are not routinely updated and maintained and can therefore rapidly become obsolete. Inconsistent links to external sources of complementary data (e.g., OMIM and HGMD) are also prevalent. Assuming similar comprehensive coverage, only 18% of genes and 56% of mutations presently listed in HGMD would be covered by available LSDBs. Examples of currently maintained multi-gene LSDBs include the Inherited Peripheral Neuropathies Mutation Database (http://molgen-www.uia.ac.be/CMTMutations/), which contains mutations from 17 different genes and the Familial Hypertrophic Cardiomyopathy Mutation Database (www.angis.org.au/Databases/Heart/heartbreak.html) containing mutations from nine different genes.

Other sources of mutation data comparable to HGMD include OMIM [www.ncbi.nlm.nih.gov/Omim/] and SWISS-PROT [http://ca.expasy.org/sprot; Boeckmann et al., 2003]. OMIM currently contains examples of allelic variants identified in a total of 1,389 human nuclear genes, including some neutral polymorphisms and somatic mutations. By comparison, SWISS-PROT contains some 15,348 variants in a total of 1,888 different protein sequences (www.expasy.org/cgi-bin/lists?humpvar.txt), although only approximately 830 of these proteins exhibit disease-associated variants. The remainder appear to contain polymorphisms and rare neutral variants.

The lack of complete coverage from other complementary sources of information, combined with the rigorous quality control and uniform/user-friendly structure provided by HGMD, serves to ensure that this database will continue to represent an important tool for academic, clinical, and commercial users.

REFERENCES

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370.

Claustres M, Horaitis O, Vanevski M, Cotton RGH. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12:680–688.

Cooper DN, Ball EV, Krawczak M. 1998. The Human Gene Mutation Database. Nucleic Acids Res 26:285–287.

Cooper DN, Nussbaum RL, Krawczak M. 2002. Proposed guidelines for papers describing DNA polymorphism-disease associations. Hum Genet 110:207–208.

Cuticchia AJ. 2000. Future vision of the GDB human genome database. Hum Mutat 15:62–67.

den Dunnen JT, Antonarakis SE 2001. Nomenclature recommendations: nomenclature for the description of human sequence variations. Hum Genet 109:121–124.

Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. 2002. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30:387–391.

Frézal J. 1998. Genatlas database, genes and development defects. C R Acad Sci III 321:805–817.

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55.

Kerlavage A, Bonazzi V, di Tommaso M, Lawrence C, Li P, Mayberry F, Mural R, Nodell M, Yandell M, Zhang J, Thomas P. 2002. The Celera Discovery System. Nucleic Acids Res 30:129–136.

Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. 1998. MITOMAP: a human mitochondrial genome database – 1998 update. Nucleic Acids Res 26:112–115.

Krawczak M, Cooper DN. 1997. The Human Gene Mutation Database. Trends Genet 13:121–122.

Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN. 2000. The Human Gene Mutation Database – a biomedical information and research resource. Hum Mutat 15:45–51.

Pagon RA, Tarczy-Hornoch P, Baskin PK, Edwards JE, Covington ML, Espeseth M, Beahler C, Bird TD, Popovich B, Nesbitt C, Dolan C, Marymee K, Hanson NB, Neufeld-Kaiser W, Grohs GM, Kicklighter T, Abair C, Malmin A, Barclay M, Palepu RD. 2002. GeneTests-GeneClinics: genetic testing information for a growing audience. Hum Mutat 19:501–509.

Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. 2001. The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 109:678–680.

Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140.

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 14:656–664.