Correspondence

The Human Gene Mutation Database: 2008 update

Peter D Stenson, Matthew Mort, Edward V Ball, Katy Howells, Andrew D Phillips, Nick ST Thomas and David N Cooper

Address: Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.

Correspondence: David N Cooper. Email: cooperdn@cardiff.ac.uk

Published: 22 January 2009

Genome Medicine 2009, **1**:13 (doi:10.1186/gm13)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/1/13

Abstract

The Human Gene Mutation Database (HGMD®) is a comprehensive core collection of germline mutations in nuclear genes that underlie or are associated with human inherited disease. Here, we summarize the history of the database and its current resources. By December 2008, the database contained over 85,000 different lesions detected in 3,253 different genes, with new entries currently accumulating at a rate exceeding 9,000 per annum. Although originally established for the scientific study of mutational mechanisms in human genes, HGMD has since acquired a much broader utility for researchers, physicians, clinicians and genetic counselors as well as for companies specializing in biopharmaceuticals, bioinformatics and personalized genomics. HGMD was first made publicly available in April 1996, and a collaboration was initiated in 2006 between HGMD and BIOBASE GmbH. This cooperative agreement covers the exclusive worldwide marketing of the most up-to-date (subscription) version of HGMD, HGMD Professional, to academic, clinical and commercial users.

The Human Gene Mutation Database (HGMD®) [1] records the first report of a disease-causing mutation or disease-associated/functional polymorphism and provides these data in a readily accessible form to all interested parties, whether they are from an academic, a clinical or a commercial background. HGMD has become the *de facto* central disease-associated mutation database available to the scientific community. The data comprise single base-pair substitutions in coding, regulatory and splicing-relevant regions of human nuclear genes, micro-deletions and micro-insertions, combined insertions/deletions (indels), repeat expansions, gross lesions (deletions, insertions and duplications) and complex rearrangements (including inversions). These categories of mutation data are summarized in Table 1.

Mutation and polymorphism data are obtained by means of a combination of manual and computerized search procedures. Thus, online library screening, the PubMed database and publicly available locus-specific mutation databases (LSDBs) are all used to optimize data acquisition. Each mutation or disease-associated/functional polymorphism is entered into HGMD only once under its earliest literature citation. Silent mutations within the coding region that do not alter the encoded amino acid are not recorded unless there is clear evidence of altered splicing and/or a direct disease association. Mutations that have not been adequately or unambiguously described in the corresponding literature report are also excluded unless full details can subsequently be obtained from the authors. Disease-associated/functional polymorphisms (see below) are excluded if the published data are deemed to be of insufficient quality (either because of the description provided or because of a tenuous/non-significant association with a clinical or laboratory phenotype). HGMD does not include somatic lesions or mitochondrial genome mutations. These are well covered by COSMIC [2] and MITOMAP [3], respectively.

Mutation data are viewable on a gene-wise basis and access to the subcategorized mutation data is available via hypertext link from each gene page. Additional links to

Table 1

Summary of mutation data present in HGMD (as of 1 December 2008)

	Number of entries	
Mutation type	HGMD public version	HGMD professional (subscription version)
Single base-pair substitutions		
Missense or nonsense	35,545	48,343
Splicing	5,803	8,219
Regulatory	817	1,400
Other lesions		
Small (≤20 bp) deletions	10,035	13,628
Small (≤20 bp) insertions	4,014	5,567
Small (≤20 bp) indels	909	1,244
Large (>20 bp) deletions	3,536	5,158
Large (>20 bp) insertions and duplications	559	1,003
Complex rearrangements (including inversions)	453	736
Repeat variations	151	260
Total	61,822	85,558

complementary data sources are provided here (Genome Database (GDB) [4], Online Mendelian Inheritance in Man (OMIM) [5], HUGO Gene Nomenclature Committee (HGNC) [6], Entrez Gene [7], GeneCards [8], GenAtlas [9], GeneClinics [10], UniGene [11], SwissProt [12] and the Human Protein Reference Database [13]). HGMD also provides annotated cDNA sequences for over 95% of the genes present in the database. The annual number of new entries being logged in HGMD has been steadily increasing over the last few years (Table 2) and now stands at over 9,000 mutations per annum. The number of new genes being entered into HGMD has also followed this upward trend, with 441 new genes being introduced during 2007 alone.

Disease-associated/functional polymorphisms

In addition to disease-causing mutations, HGMD seeks to include polymorphic DNA sequence variants that are either disease-associated and of direct functional significance, or of clear functional significance even though an associated clinical phenotype has yet to be identified. At present, these polymorphic variants comprise about 5% of HGMD data and approximately 55% of these are 'disease-associated'. The remainder represent variants that, despite manifesting no demonstrable disease association, have nevertheless been shown to significantly alter the expression of a gene or the structure/function of the gene product. Although functional polymorphisms with no known disease association do not have any immediate clinical relevance, these data are potentially very valuable in terms of understanding inter-individual differences in disease susceptibility.

Although the vast majority of polymorphic variants in HGMD are single-nucleotide polymorphisms (SNPs), a small number are of the insertion/deletion type. The polymorphic variants logged in HGMD are generally located in either the gene promoter or coding regions. However, it should be noted that SNPs occurring outside of these regions may nevertheless still have consequences for gene expression, splicing or transcription-factor binding. Polymorphic variants affecting individual drug response [14], patient survival times after diagnosis and responses to surgical intervention are not generally included in HGMD. Studies that simply report SNPs [15] in association with disease (and hence are likely to represent merely a linkage disequilibrium effect), but with no additional evidence of direct functional involvement of the variants in question, are also not included. Reports of haplotypes associated with an increased risk of disease are not included unless there is some indication as to precisely which variant(s) within the haplotype is/are responsible for the disease association or functional effect.

In some instances, the above criteria may be only partially satisfied, such that the HGMD curators remain unconvinced as to the clinical phenotypic relevance of the reported polymorphic variant. In such cases, the polymorphism may nevertheless still be included (i) as a result of supporting

Table 2

Number of new records entered into HGMD by year of entry

Year	Number of new mutation entries	Number of new gene entries
2001	7,451	189
2002	5,849	197
2003	5,989	214
2004	5,657	257
2005	7,649	241
2006	9,901	287
2007	9,371	441
2008 (to 1 December)	9,006	353

information becoming available subsequent to the publication of the original report, or (ii) because the associated gene/disease state was deemed to be of sufficient importance for it to warrant further study. Such variants are generally ascribed the descriptor 'association with?' to indicate that some degree of uncertainty is involved. The difficulty inherent in making decisions regarding the inclusion or exclusion of variants that have potential disease associations highlights the need for a methodical and methodologically uniform approach to assessing such reports as they appear in the literature [16].

Several other databases [17-19] have attempted to collate known polymorphism-disease associations but have met with only partial success owing to an over-reliance on computerized search procedures and automated data collection. This methodology tends to result in the creation of a database that comprises either verbatim and/or often inconsistent records of the disease-associated variants, or merely a list of PubMed citations rather than the actual variants in question. Polymorphism-disease association data curated in this way are also likely to comprise markers that occur in linkage disequilibrium with the presumed disease-associated/functional variants rather than being of functional significance themselves. We on the HGMD team believe that a manually curated database provides a rather better solution. Indeed, HGMD is currently the only database that focuses specifically on the collation of functional/disease-associated polymorphic variants to the exclusion of linkage markers.

A current limitation with regard to recording disease-associated polymorphic variants of functional significance within HGMD is the inclusion of only a single literature reference for each variant. A large proportion of those papers reporting a novel association between a disease and a polymorphic variant do not include functional data on that variant. HGMD will in the future address this by implementing a dual referencing system for polymorphisms: reference 1 will correspond to the first report demonstrating a functional effect (or disease-association) that meets the HGMD inclusion criteria, whereas reference 2 will (where appropriate) provide evidence of the first disease-association (or functional effect) of the polymorphism.

HGMD funding

Since HGMD does not receive any public funding to support its upkeep, it has been necessary to develop a sustainable model for both the current and future funding of the database. The ideal model (in the opinion of the curators) would be a mixture of income from both public and private sources. This, in principle, would allow HGMD to provide free database access to academic/non-profit users alongside a subscription-based distribution for commercial users marketed by a commercial company. With this eventual aim in mind, the HGMD curators opted to market their data in collaboration with BIOBASE GmbH [20]. As part of the commercial agreement, Cardiff University, as HGMD's host institution, agreed to provide BIOBASE with a period of exclusive access to newly added mutational information. This period extends to 2.5 years from the date of initial inclusion. BIOBASE provides HGMD (in the form of HGMD Professional; see below) as a stand-alone product as part of its database subscription package. The publicly available version of HGMD will, however, continue to be made available as a free service to registered users from academic/non-profit institutions via the Cardiff website [1]. By insisting that commercial entities pay for access to the latest HGMD data and software tools, while still providing a less up-to-date version free of charge to registered users from academic/non-profit institutions, the HGMD curators believe that they can continue to allow free access to the bulk of their mutation data, at the same time as generating sufficient income to support the maintenance of HGMD from its commercial distribution.

HGMD users and usage

HGMD has recently introduced a user registration scheme, which is free for users from academic/non-profit organizations. Prior registration is required to access and use HGMD. After completing the registration form, users are sent a password by email, which they can use to log on to the public HGMD website. Since the inception of the system in April 2006, over 23,000 user registrations have been recorded and HGMD is continuing to accrue about 800 new registrations every month. We have registered users from over 150 different countries (Table 3), providing an indication of how widely HGMD is used by the academic community worldwide. Each month, an average of 14,000 queries for genes are received (with an equal number accessing HGMD genes via an external link) from almost 6,000 users, with a total of over 160,000 pages served.

Table 3

Number of user registrations by geographic origin (as of 1 December 2008)

Country	Number of registrations
United States	5,647
China (including Hong Kong)	2,053
United Kingdom	1,949
Italy	1,340
India	1,072
Germany	891
Spain	857
Japan	786
France	746
Australia and New Zealand	778
Canada	646
Rest of Europe	3,129
Rest of Asia (including Middle East)	2,640
Central and South America	920
Africa	306
Others	11
Total	23,771

Users of the public site may not download HGMD data in their entirety without permission. This is, however, generally granted if the data are to be used exclusively for non-commercial collaborative research purposes. Collaborators who wish to access HGMD data in full are required to sign a confidentiality agreement. Recent successful collaborations include the projects to sequence the genomes of *Macaca mulatta* [21] and *Rattus norvegicus* [22] and a study into gains of glycosylation as a cause of inherited disease [23]. HGMD data have also been used by researchers in several other studies, including the newly reported sequencing of diploid genomes from individual humans [24,25] and recent mutation [26] and evolutionary [27] studies.

HGMD Professional

HGMD Professional [28] has been developed to serve as the subscription version of HGMD, and is available to both commercial and academic customers through a license from BIOBASE, our commercial partner. The Professional version allows access to up-to-date mutation data, with an updated product release every 3 months. This version is therefore essential for checking the novelty and/or pathogenicity of mutations newly found by researchers, clinicians and diagnosticians.

HGMD Professional contains many features not available in the free public version [29]. More powerful search tools in the form of an expanded search engine with full text Boolean searching are provided. Users can use these tools to perform additional searches for chromosomal locations, Entrez Gene ID numbers [7], HGNC database ID numbers [6], common gene symbol aliases [7], codon numbers, HGMD accession numbers and literature references. Enhanced gene and mutation viewing is also enabled, allowing improved navigation between different genes and mutation types.

Additional information is also provided on a mutation-specific basis. This includes curatorial comments pertaining to particular mutations (for example if the mutation data required correction in relation to the data presented in the original publication), and comparative biochemical information (including change in residue polarity, pH, weight, hydrophobicity [30,31], secondary structure propensity [32] and Grantham difference [33]) for the amino acid substitutions. These data are intended to assist with the assessment of the likely pathogenicity of each missense mutation. Pairwise alignments (21 amino acids long) are also provided for the majority of missense mutations in HGMD, using orthologous protein sequences obtained from the Entrez protein database [34]. In this context, HGMD Professional currently contains protein sequences from *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Canis lupus familiaris*, *Felis catus*, *Sus scrofa*, *Ovis aries*, *Bos taurus*, *Takifugu rubripes*, *Pan troglodytes* and *Macaca mulatta*. Amino acid sequences will continue to be added as they appear in the Entrez protein database. This feature allows HGMD users to ascertain the evolutionary conservation of a given amino acid residue and its surrounding sequence in relation to each missense mutation without having to load mutated sequences from HGMD into an external database.

Advanced search tools

The Advanced Search is essentially a suite of software tools, available as part of HGMD Professional, which are designed to enhance mutation searching, viewing and retrieval. Two of the main types of mutation in HGMD (single-nucleotide substitutions and microlesions) can be interrogated with this toolset. The datasets of each mutation type can be combined (for example, micro-deletions, microinsertions and indels) to enable more powerful searching across comparable types of mutation. When using the Advanced Search, users can tailor their queries with more specific criteria, including amino-acid exchange; nucleotide substitution; the size of a micro-deletion, microinsertion or indel; composition; motifs (both those created and those abolished by the mutation); dbSNP number; and keywords in the article title or abstract. Any mutation results returned by the Advanced Search can be downloaded in a tab-delimited format, ready to import into a different application.

Part of the Advanced Search tool includes a dynamic mutation viewer, which depicts coding-region mutations superimposed on the cDNA sequence of a gene. The wild-type cDNA sequence is represented in black whereas the mutated nucleotides are shown in different colors according to the type of mutation. Displays of each mutation type can be switched on or off using the appropriate buttons.

Future directions for HGMD

We have recently incorporated a fully comprehensive functional/disease-associated polymorphism dataset into HGMD to complement the existing disease-causing mutation data. The provision of additional orthologous protein sequences for alignments, fully annotated genomic sequences for all HGMD genes, and genomic coordinates for as many mutations as possible are also seen as high priorities.

The provision of further supplementary information, including additional clinical phenotypes observed with a given mutation, multiple additional references and gene and disease ontologies and *in vitro* characterization data will be added to HGMD once resources permit.

We believe that the development of HGMD Professional together with the suite of Advanced Search tools provides the user with a unique resource that can be used not only to acquire evidence to support the pathological authenticity and/or novelty of detected lesions, but also to obtain an overview of the mutational spectra for specific genes. It is hoped that these new developments will not only help to secure the future of HGMD, but will also enhance the facilities currently used for the long term storage and provision of mutation data to the scientific community.

Abbreviations

DbSNP, The Entrez Single Nucleotide Polymorphism database; GDB, Genome Database; HGMD, Human Gene Mutation Database; HGNC, HUGO Gene Nomenclature Committee; HUGO, Human Genome Organisation; LSDB, Locus-specific Mutation Database; OMIM, Online Mendelian Inheritance in Man; SNP, single-nucleotide polymorphism.

Competing interests

The authors wish to declare that HGMD is in receipt of funding from BIOBASE GmbH [2], their commercial partner. HGMD Professional, a version of the database that includes a suite of advanced analytical/search tools, is only available via subscription through BIOBASE. All financial support received from BIOBASE is used exclusively for the upkeep and maintenance of HGMD.

Authors' contributions

All authors are substantially involved in the upkeep and maintenance of HGMD. PDS and MM are primarily responsible for HGMD Professional and the Advanced Search toolset. All authors made a contribution to the text of this article and read and approved the final manuscript.

References

1. **The Human Gene Mutation Database** [http://www.hgmd.org]
2. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR: **COSMIC 2005.** *Br J Cancer* 2006, **94:**318-322.
3. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC: **An enhanced MITOMAP with a global mtDNA mutational phylogeny.** *Nucleic Acids Res* 2007, **35:**D823-D828.
4. Cuticchia AJ: **Future vision of the GDB human genome database.** *Hum Mutat* 2000, **15:**62-67.
5. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM.** *Am J Hum Genet* 2007, **80:**588-604.
6. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E: **The HGNC Database in 2008: a resource for the human genome.** *Nucleic Acids Res* 2008, **36:**D445-D448.
7. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2007, **35:**D26-D31.
8. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D: **GeneCards 2002: towards a complete, object-oriented, human gene compendium.** *Bioinformatics* 2002, **18:**1542-1543.
9. Frézal J: **Genatlas database, genes and development defects.** *C R Acad Sci III* 1998, **321:**805-817.
10. 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: **GeneTests-GeneClinics: Genetic testing information for a growing audience.** *Hum Mutat* 2002, **19:**501-509.
11. Pontius JU, Wagner L, Schuler GD: **UniGene: a unified view of the transcriptome.** In *The NCBI Handbook*. Bethesda: National Center for Biotechnology Information; 2003.
12. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase.** *Methods Mol Biol* 2007, **406:**89-112.
13. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, *et al.*: **Human protein reference database—2006 update.** *Nucleic Acids Res* 2006, **34:**D411-D414.
14. Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, Gong L, Owen R, Gong M, Gor W, Liu F, Truong C, Whaley R, Woon M, Zhou T, Altman RB, Klein TE: **The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge.** *Nucleic Acids Res* 2008, **36:**D913-D918.
15. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.
16. Cooper DN, Nussbaum RL, Krawczak M: **Proposed guidelines for papers describing DNA polymorphism-disease associations.** *Hum Genet* 2002, **110:**207-208.
17. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36:**431-432.
18. Rhee H, Lee JS: PADB: **Published association database.** *BMC Bioinformatics* 2007, **8:**348.
19. **HuGE navigator** [http://www.hugenavigator.net/]
20. **BIOBASE website** [http://www.biobase-international.com]
21. Rhesus Macaque Genome Sequencing and Analysis Consortium: **Evolutionary and biomedical insights from the rhesus macaque genome.** *Science* 2007, **316:**222-234.
22. Rat Genome Sequencing Consortium: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428:**493-521.
23. Vogt G, Chapgier A, Yang K, Chuzhanova N, Feinberg J, Fieschi C, Boisson-Dupuis S, Alcais A, Filipe-Santos O, Bustamante J, de Beaucoudrey L, Al-Mohsen I, Al-Hajjar S, Al-Ghonaium A, Adimi P, Mirsaeidi M, Khalilzadeh S, Rosenzweig S, de la Calle Martin O, Bauer

TR, Puck JM, Ochs HD, Furthner D, Engelhorn C, Belohradsky B, Mansouri D, Holland SM, Schreiber RD, Abel L, Cooper DN, *et al.*: **Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations.** *Nat Genet* 2005, **37:**692-700.
24. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.
25. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5:**e254.
26. Kryukov GV, Pennacchio LA, Sunyaev SR**: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.** *Am J Hum Genet* 2007, **80:**727-739.
27. Subramanian S, Kumar S: **Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome.** *BMC Genomics* 2006, **7:**306.
28. **HGMD Professional** [http://www.biobase-international.com/pages/index.php?id=hgmddatabase]
29. **The benefits of HGMD Professional** [http://www.biobase-international.com/pages/index.php?id=411]
30. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105-132.
31. Hopp TP, Woods KR: **Prediction of protein antigenic determinants from amino acid sequences.** *Proc Natl Acad Sci USA* 1981, **78:**3824-3828.
32. Chou PY, Fasman GD: **Prediction of protein conformation.** *Biochemistry* 1974, **13:**222-245.
33. Grantham R: **Amino acid difference formula to help explain protein evolution.** *Science* 1974, **185:**862-864.
34. **The Entrez Protein Database** [http://www.ncbi.nlm.nih.gov/sites/entrez?db=Protein]