The human gene mutation database

David N. Cooper*, Edward V. Ball and Michael Krawczak

Institute of Medical Genetics, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, UK

Received August 13, 1997; Revised and Accepted September 26, 1997

ABSTRACT

The Human Gene Mutation Database (HGMD) represents a comprehensive core collection of data on published germline mutations in nuclear genes underlying human inherited disease. By September 1997, the database contained nearly 12 000 different lesions in a total of 636 different genes, with new entries currently accumulating at a rate of over 2000 per annum. Although originally established for the scientific study of mutational mechanisms in human genes, HGMD has acquired a much broader utility to researchers, physicians and genetic counsellors so that it was made publicly available at http://uwcm.ac.uk/uwcm/mg/hgmd0.html in April 1996. 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) and Online Mendelian Inheritance in Man (OMIM), Baltimore, USA. Hypertext links have also been established to Medline abstracts through Entrez, and to a collection of 458 reference cDNA sequences also used for data checking. Being both comprehensive and fully integrated into the existing bioinformatics structures relevant to human genetics, HGMD has established itself as the central core database of inherited human gene mutations.

INTRODUCTION

The Human Gene Mutation Database (HGMD), maintained at the Institute of Medical Genetics in Cardiff, represents a comprehensive core collection of data on germline mutations underlying human inherited disease. Thus, 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 and ‘indels’, plus a number of complex rearrangements not covered by the above categories. Somatic gene mutations and mitochondrial genome mutations are not included.

The curators of HGMD have adopted a policy of entering each mutation only once in order to avoid confusion between recurrent and identical-by-descent lesions. Reliable discrimination between these two alternatives would require information available only for a very small proportion of known lesions. Therefore, although data on the regional, ethnic and haplotype context of mutations would be extremely useful in terms of epidemiological and population genetics research, any unselective accumulation of literature reports would have resulted in an inflation of references with little immediate scientific use.

Although originally established for the scientific study of mutational mechanisms in human genes (1), HGMD has acquired a much broader utility in that it provides information of practical importance to researchers in human molecular genetics, physicians interested in a particular inherited condition in a given patient or family, and genetic counsellors. In view of its potential usefulness, the curators of HGMD made the database publicly available (2) through the WorldWideWeb in April 1996.

DATA COVERAGE AND STRUCTURE

By September 1997, HGMD contained >11 900 different lesions in a total of 636 different genes (Table 1). Entries are accumulating at a rate of >2000 per annum (Fig. 1). Coverage is limited to original published reports although some data are taken from ‘Mutation Updates’ or review articles. Mutations reported only in abstract form are not generally included. Data acquisition for HGMD has been accomplished by a combination of manual and computerised search procedures, scanning in excess of 250 journals on a weekly/monthly basis.

Table 1. Number of HGMD entries by mutation type (September 1997)

Mutation type	No. of entries
Single base-pair substitutions, missense/nonsense	7282
Single base-pair substitutions, splicing	1052
Single base-pair substitutions, regulatory	102
Small deletions (≤20 bp)	1857
Small insertions (≤20 bp)	653
Small indels (≤20 bp)	82
Repeat expansions	15
Gross deletions (>20 bp)	736
Gross insertions and duplications (>20 bp)	122
Complex rearrangements including inversions	71
Total	11972

*To whom correspondence should be addressed. Tel: +44 1222 744062; Fax: +44 1222 747603; Email: cooperdn@cardiff.ac.uk

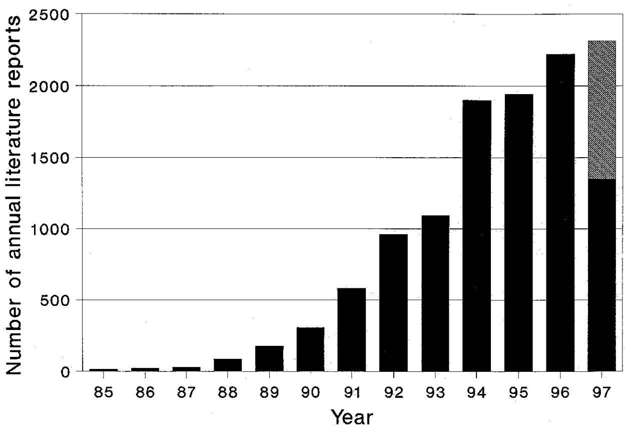

Figure 1. Number of different germline mutations underlying human genetic disease annually reported in the literature. The shaded section of the 1997 bar represents an extrapolation from data logged in HGMD by August 1997.

All HGMD entries comprise a reference to the first literature report of a mutation, the associated disease state as specified in that report, the gene name, symbol (as used by GDB) and chromosomal location. In cases where a gene symbol has not yet been made available owing to the recency of the cloning report, a provisional symbol has been adopted which is denoted by lower-case letters. 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 lies in either the first or third position in the triplet. While substitutions causing regulatory abnormalities are logged in with 8 nt flanking the site of mutation on both sides, no flanking sequence has been included yet for substitutions leading to aberrant splicing. Micro-deletions and micro-insertions (of ≤20 bp) are presented in terms of the deleted/inserted bases in lower case plus (in upper case) 10 bp DNA sequence flanking both ends of the lesion. Either the codon number or, in cases where a lesion extends outwith 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). Codon numbering may in some cases display inconsistencies owing to the common use of different numbering systems for the same protein. For the majority of genes, however, residue numbering has been standardized with respect to a generally accepted numbering system employing the appropriate reference cDNA sequence. For gross deletions, gross insertions and complex rearrangements, information regarding the nature and location of a lesion is logged in narrative form because of the extremely variable quality of the original data reported.

DATA ACCESS

HGMD is accessible on the basis of every gene being allocated one web page per mutation type, if data of that type are present. Since HGMD is partly dependent upon industrial funding and involves considerable editorial work over and above mere literature screening (e.g. to ensure the consistency of nucleotide sequence information, amino acid residue numbering and gene symbol usage), unsolved copyright problems have so far precluded HGMD from being downloadable in its entirety. However, once the closer cooperation with publically funded bioinformatics institutions currently envisaged has been put in place, unrestricted access to the database will become possible. During its first 17 months on the Internet, HGMD has been accessed >70 000 times.

Meaningful integration of the data with phenotypic, structural and mapping information on human genes has been accomplished through bi-directional links between HGMD and both the Genome Database (GDB) and Online Mendelian Inheritance in Man (OMIM), Baltimore, USA. In addition, hypertext links have been established from HGMD references to Medline abstracts through Entrez. Hypertext links have also been set up to ‘reference cDNA sequences’ (458 to date) which are used for data checking. The links to GDB and OMIM have enforced the standardisation of disease and gene nomenclature in HGMD. Thus HGMD can be searched either by HUGO-approved gene symbols, GDB accession numbers, or OMIM-compatible disease or gene names. For genes for which Locus-Specific Mutation Databases are available on the Internet, these databases (currently ~40) can be accessed either from the corresponding gene-specific HGMD pages or via the Locus-Specific Mutation Database page (3).

Table 2. Number of HGMD entries per gene by mutation type (September 1997)

Mutation type	Number of entries per gene								
	1	2	3	4–5	6–10	11–25	26–50	51–100	>100
Single bp substitutions									
Missense/nonsense	126	86	38	53	96	94	41	25	6
Splicing	100	58	30	17	31	20	3	2	0
Regulatory	19	4	3	3	4	2	0	0	0
Other lesions									
Small deletions (≤20 bp)	112	52	31	45	40	24	12	4	1
Small insertions (≤20 bp)	94	44	14	18	26	11	2	0	0
Small indels (≤20 bp)	45	12	1	1	1	0	0	0	0
Repeat expansions	15	0	0	0	0	0	0	0	0

CONCLUSIONS AND OUTLOOK

Being both comprehensive and fully integrated into the existing bioinformatics structures relevant to human genetics, HGMD has established itself as the central core database of inherited human gene mutations. In order to improve the accuracy, efficiency and rapidity of mutation publication, however, direct submission of mutation data to a central resource capable of (and responsible for) checking the novelty and consistency of data is both necessary and desirable. Although some Locus-Specific Databases have included mutations not published anywhere in the literature, even the close integration of these facilities will be inadequate to the task of meeting the demands likely to be made upon a central data repository. Table 2 illustrates that a substantial proportion of published mutation data are derived from genes in which only a handful of lesions have so far been characterised. In such cases the establishment of a Locus-Specific Database is not warranted. Indeed, such a resource is currently accessible via the Internet for only 58/628 (9%) of genes also referred to in HGMD. Although mutation data associated with these genes should comprise 48% mutations in HGMD (assuming the Locus-Specific Databases to be sufficiently comprehensive), the obvious lack of general coverage stresses the point that comprehensive collection of mutation data can only be performed in generalised fashion. To this end, HGMD has instituted a collaboration with Springer-Verlag GmbH, Heidelberg, to make online submission and electronic publication of human gene mutation data possible (4). These data will be published regularly by Springer's journal *Human Genetics* in both electronic and printed form. Once published, the data will be transmitted to Cardiff and deposited in HGMD. It is hoped that other journals may eventually follow suit.

ACKNOWLEDGEMENTS

The authors wish to thank SmithKline Beecham, Pfizer and the Deutsche Forschungsgemeinschaft for their financial support and Iain Fenton for computer assistance.

REFERENCES

1 Cooper,D.N. and Krawczak,M. (1993) *Human Gene Mutation*. BIOS, Oxford.
2 http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html
3 http://www.uwcm.ac.uk/uwcm/mg/oth_mut.html
4 http://link.springer.de/journals/humangen/mutation/