Hum Genet (2014) 133:1–9
DOI 10.1007/s00439-013-1358-4

REVIEW PAPER

The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine

Peter D. Stenson · Matthew Mort · Edward V. Ball · Katy Shaw · Andrew D. Phillips · David N. Cooper

Received: 29 July 2013 / Accepted: 3 September 2013 / Published online: 28 September 2013

Abstract The Human Gene Mutation Database (HGMD®) is a comprehensive collection of germline mutations in nuclear genes that underlie, or are associated with, human inherited disease. By June 2013, the database contained over 141,000 different lesions detected in over 5,700 different genes, with new mutation entries currently accumulating at a rate exceeding 10,000 per annum. HGMD was originally established in 1996 for the scientific study of mutational mechanisms in human genes. However, it has since acquired a much broader utility as a central unified disease-oriented mutation repository utilized by human molecular geneticists, genome scientists, molecular biologists, clinicians and genetic counsellors as well as by those specializing in biopharmaceuticals, bioinformatics and personalized genomics. The public version of HGMD (http://www.hgmd.org) is freely available to registered users from academic institutions/non-profit organizations whilst the subscription version (HGMD Professional) is available to academic, clinical and commercial users under license via BIOBASE GmbH.

Introduction

The Human Gene Mutation Database (HGMD®) represents an attempt to collate all known gene lesions responsible for causing human inherited disease together with disease-associated/functional polymorphisms that have been published in the peer-reviewed literature. These data comprise single base-pair substitutions in coding, regulatory and splicing-relevant (both intronic and exonic) regions of human nuclear genes, as well as micro-deletions and micro-insertions, combined micro-insertions/micro-deletions (indels) of 20 bp or less, repeat variations, gross lesions (deletions, insertions and duplications of greater than 20 bp, up to and including a single characterized gene or group of contiguous genes that are directly involved in the aetiology of the disease/phenotype) and complex rearrangements (including inversions, translocations and complex indels). Mutation data are summarized in Table 1.

HGMD does not include either somatic or mitochondrial mutations, which are well covered by COSMIC (Forbes et al. 2011) and MitoMap (Ruiz-Pesini et al. 2007), respectively. HGMD also does not attempt to provide comprehensive coverage of pharmacological variants (except for those variants where evidence supporting a functional impairment has been provided); such variants are covered by PharmGKB (Thorn et al. 2010). Finally, HGMD is not a general genetic variation database; users interested in this type of variant should visit dbSNP (Sherry et al. 2001) or the Exome Variant Server (http://evs.gs.washington.edu/EVS/).

HGMD was originally established for the scientific study of mutational mechanisms in human genes causing inherited disease (Cooper et al. 2010), but has since acquired a much broader utility as a central unified repository for germ-line disease-related functional variation. It is now routinely accessed and utilized by next generation sequencing (NGS) project researchers, human molecular geneticists, molecular biologists, clinicians and genetic counsellors as well as by those specializing in biopharmaceuticals, bioinformatics and personalized genomics.

P. D. Stenson (✉) · M. Mort · E. V. Ball · K. Shaw · A. D. Phillips · D. N. Cooper (✉)
Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK
e-mail: StensonPD@Cardiff.ac.uk

D. N. Cooper
e-mail: cooperDN@cardiff.ac.uk

Table 1 Numbers of different mutations by mutation type present in HGMD Professional 2013.2 and the publicly available version of the database (June 28th 2013)

Mutation type	Total numbers of mutations		
	HGMD Professional	With chromosomal coordinates	Publicly available
Missense substitutions	62,368	61,845	44,933
Nonsense substitutions	15,781	15,574	11,306
Splicing substitutions	13,030	12,538	9,467
Regulatory substitutions	2,751	2,713	1,753
Micro-deletions $\leq$ 20 bp	21,681	21,134	15,796
Micro-insertions $\leq$20 bp	8,994	8,721	6,494
Micro-indels $\leq$20 bp	2,083	2,004	1,459
Gross deletions >20 bp	10,267	0	6,156
Gross insertions/ duplications >20 bp	2,376	0	1,253
Complex rearrangements	1,409	0	946
Repeat variations	421	0	305
Totals	141,161	124,529	99,868

Table 2 Differences between HGMD Professional and HGMD Public

	HGMD Professional	HGMD Public
Up-to-date mutation data	✔	✗
Curator comments	✔	✔
Quarterly updates^{+}	✔	✗
Gene-oriented search	✔	✔
Mutation-oriented search	✔	✗
Reference-oriented search	✔	✗
Batch search mode	✔	✗
Chromosomal coordinates	✔	✗
HGVS nomenclature	✔	✗
Additional literature references	✔	✗
Tracked variant history	✔	✗
dbSNP identifiers	✔	✗
Enhanced search options	✔	✗
Advanced search features	✔	✗
Disease ontology terms*	✔	✗
Data in VCF*	✔	✗
Downloadable version	✔	✗

$^{+}$ HGMD Public is updated on a 6-monthly basis

* Download customers only

HGMD is available in two versions: one public, one obtainable by subscription. The public version of HGMD (http://www.hgmd.org) is freely available to registered users from academic institutions/non-profit organizations. This version is, however, maintained in a basic form that is only updated twice per annum, is permanently 3 years out of date, and does not contain any of the additional annotations or extra features present in HGMD Professional (such as GRCh37/hg19 genomic chromosomal coordinates, HGVS nomenclature and additional literature references, see Table 2). The Professional version is available to both commercial and academic/non-profit users via subscription from BIOBASE GmbH (http://www.biobase-international.com).

Acquisition of mutation data

All HGMD mutation data are manually curated from the scientific literature. Identification of relevant literature reports is carried out via a combination of manual journal screening and automated procedures. The database currently contains mutation entries obtained from over 41,000 primary and 15,000 additional (supplementary) literature reports published in more than 1,950 different journals. Of >10,000 identified articles screened for mutation data during 2012, 35 % contained novel mutation data, 29 % contained additional useful information (e.g. in vitro functional data or further clinical or phenotypic information) and were therefore cited as additional references, whilst the remaining 36 % of articles contained no novel mutation data or supporting information to warrant their inclusion as either primary or supplementary references in HGMD. The number of articles screened by HGMD is increasing on a yearly basis; however, we impose no prior limit upon the number of articles we include as supplementary references for a given mutation.

For $\sim$4 % of all the missense/nonsense mutations reported in the literature during 2012, it was necessary for the HGMD Curators to contact the original authors to obtain correction and/or clarification of the nature or precise location of the mutations in question. However, only half of the mutations that required author contact were satisfactorily resolved by these means, leading to their inclusion in HGMD; the $\sim$2 % of unresolved missense/nonsense mutations will not be entered into HGMD unless or until the nature or precise location of the mutation(s) in question is determined to the satisfaction of the HGMD curators. Such data (currently 366 entries) are, however, retained indefinitely by HGMD as part of a "Bad Bank" of inadequately described mutations.

Classes of variant listed in HGMD

There are six different classes of variant present in HGMD (Figs. 1, 2). Disease-causing mutations (DM) are entered

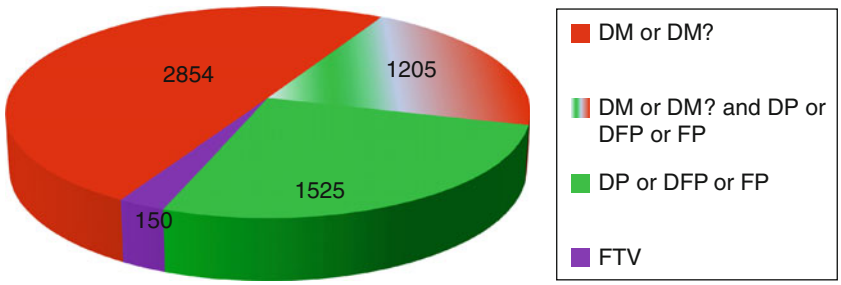

Fig. 1 5,734 genes are listed in HGMD professional 2013.2, subdivided here by variant class

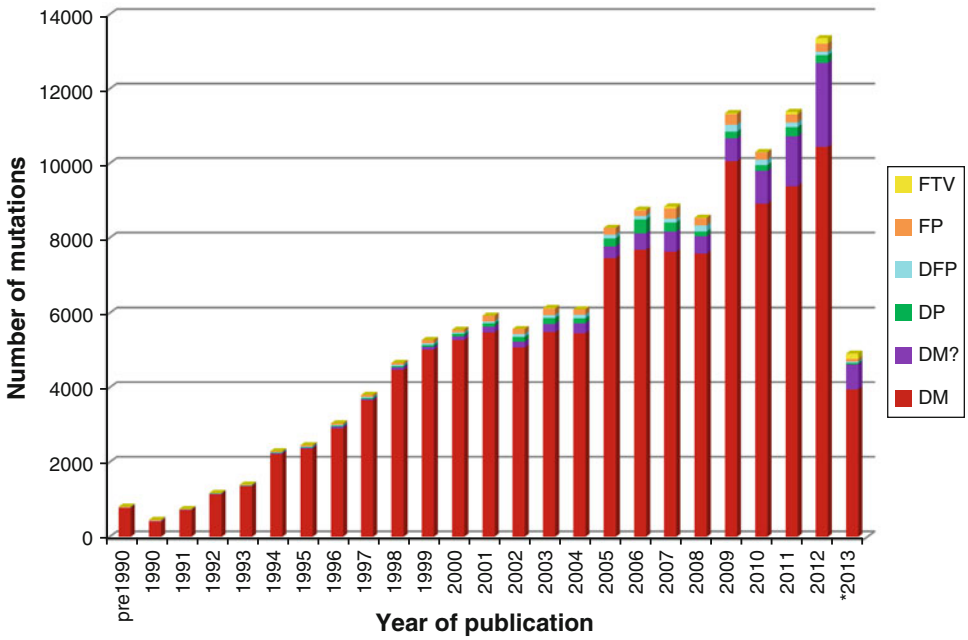

Fig. 2 HGMD annual mutation totals subdivided by variant class. *2013 figures to June 28th

into HGMD where the authors of the corresponding report(s) have demonstrated that the reported mutation(s) are involved in conferring the associated clinical phenotype upon the individuals concerned. The DM classification may, however, also appear with a question mark (DM?), denoting a probable/possible pathological mutation, reported to be pathogenic in the corresponding report, but where (1) the author has indicated that there may be some degree of uncertainty; (2) the HGMD curators believe greater interpretational caution is warranted; or (3) subsequent evidence has appeared in the literature which has called the putatively deleterious nature of the variant into question.

Disease-associated polymorphisms (DP) are entered into HGMD where there is evidence for a significant association with a disease/clinical phenotype along with additional evidence that the polymorphism is itself likely to be of functional relevance (e.g. as a consequence of genic/genomic location, evolutionary conservation, transcription factor binding potential, etc.), although there may be no direct evidence (e.g. from an expression study) of a functional effect. Functional polymorphisms (FP) are included in HGMD where the reporting authors have shown that the polymorphism in question exerts a direct functional effect (e.g. by means of an in vitro reporter gene assay or alternatively by protein structure, function or expression studies), but with no disease association reported as yet. Disease-associated polymorphisms with supporting functional evidence (DFP) must meet both of the above criteria in that the polymorphism should not only be reported to be significantly associated with disease but should also display evidence of being of direct functional relevance.

Copy number variations (CNVs) represent an important subset of potentially functional disease-associated variation. While HGMD does not wish to replicate the excellent curatorial work of other resources (e.g. the Database of Genomic Variants http://dgv.tcag.ca/dgv/app/home, DECIPHER http://decipher.sanger.ac.uk/ and Copy Number Variation in Disease http://202.97.205.78/CNVD/), we are nevertheless interested in including such variants (as gross deletions or duplications) if they meet certain criteria. Therefore, HGMD will include such variations if they have been shown to be both of functional significance and associated with disease, and involve a single characterized

gene that has itself been directly implicated in the disease association. Such variants would be entered under one of the above-mentioned polymorphism categories, depending upon the supporting evidence provided by the authors of the original reporting article.

In the opinion of the HGMD curators, the polymorphism data present in HGMD should be viewed with a considerable degree of caution owing to (1) the possibility that the observed disease association may be simply due to a linkage disequilibrium effect rather than a bona fide underlying functional mechanism and (2) the fact that in vitro studies are not invariably accurate indicators of in vivo functionality (Cirulli and Goldstein 2007; Dimas et al. 2009).

Finally, frameshift or truncating variants (FTV) are polymorphic or rare variants reported in the literature that are predicted to truncate or otherwise alter the length of the gene product (i.e. a stop-gain, stop-loss or frameshift variant) but with no disease association reported as yet. Most known FTVs have been identified during the course of large-scale genome/exome screening studies (involving either patient panels or apparently healthy individuals from the general population). They may be considered to represent either latent protein deficiencies or, potentially, heterozygous carrier states for recessive disorders. Coverage of FTVs is far from being comprehensive at this juncture, and it remains unclear what proportion will turn out to be clinically significant.

The HGMD curators have adopted a policy of continual reassessment of the curated content within the database. If and when additional and important new information pertaining to a specific mutation entry becomes available (e.g. questionable pathogenicity, confirmed pathogenicity, additional clinical or laboratory phenotypes, population frequency data, supporting functional studies, etc.), then the mutation entry may be revised or even re-categorized. Alternatively, a comment or additional reference may be added in order to communicate this new information to users. Where new information becomes available which suggests that a given disease-causing mutation (DM) is likely to be of questionable pathological relevance or possibly a neutral polymorphism (on the basis of additional case reports, genome/population screening studies, presence in dbSNP with reliable population frequency data, etc.), it may be flagged with a question mark (DM?) or even removed from the database entirely if it turns out to have been erroneously included ab initio. In a recent re-curation exercise, a total of 539 mutations were re-examined due to their presence in the 1000 Genomes Project dataset at a frequency of >1 %. Of the total re-examined, 33 mutations were removed from HGMD, 109 were re-categorized and 220 had additional comments or references added to further justify their inclusion in HGMD (Xue et al. 2012). One reason why some HGMD-listed mutations are often to be found among 1000 Genomes Project data is that many pathogenic lesions are found quite frequently in the population at large (Nishiguchi and Rivolta 2012; Andreasen et al. 2013; Lazarin et al. 2013; Cooper et al. 2013). In addition to internal curation, users of HGMD Professional may utilize a feedback function in order to inform the HGMD curators of relevant new or missing information, to request corrections or to ask for the reclassification or removal of a listed variant.

Most of the clinical phenotypes attributed to DMs in HGMD represent individually rare conditions that are generally regarded as monogenic diseases. However, it is important to note that HGMD also considers a few silent protein deficiencies or biochemical phenotypes (e.g. butyrylcholinesterase deficiency, reduced oxygen affinity haemoglobin, etc.) to be worthy of inclusion since they are potentially disease-relevant (even if they are relatively common in the general population); such variants may well be assigned to the DM category.

For individual mutations in HGMD, the provision of zygosity information (heterozygous, homozygous or compound heterozygous) has not been attempted. Reasons for this include (1) the fact that this information is not always unambiguously provided in the corresponding article; (2) the possibility that a given mutation may be pathogenic irrespective of the zygosity in which it is found; (3) the clinical consequences of zygosity may often be modified by other genetic variants either in *cis* or in *trans* and (4) the general phenomenon of variable or reduced penetrance which ensures that the genotype is not invariably predictive of the phenotype (Cooper et al. 2013). Thus, information pertaining to zygosity would not always be helpful or informative with regard to ascertaining or predicting the clinical phenotype, and indeed might even prove inaccurate or misleading.

HGMD users should not assume that just because a mutation is labelled "DM", that it automatically follows that the mutation is known or believed to be pathogenic in all individuals harbouring it (i.e. that the mutation exhibits 100 % penetrance). This is not invariably going to be the case and many "disease-causing mutations" will display reduced or variable penetrance for a variety of different reasons (reviewed by Cooper et al. 2013). Indeed, next generation sequencing programmes (such as the 1000 Genomes Project) are now identifying considerable numbers of "DM" mutations in apparently healthy individuals (MacArthur et al. 2012; Xue et al. 2012). Such lesions should not automatically be regarded as being clinically irrelevant because it is quite possible that they represent low-penetrance, mild or late onset, or more complex disease susceptibility alleles, as opposed to neutral variants (Cooper et al. 2013).

It has always been HGMD policy to enter a variant into the database even if its pathological relevance may be questionable (while indicating this fact wherever feasible to our users), rather than run the risk of inadvertently excluding a variant that may be directly (or indirectly) relevant to disease. We have taken several steps to highlight such equivocation in HGMD, viz. the recent introduction of the DM? variant class, a dbSNP 1000 Genomes frequency flag (to highlight HGMD variants that are also present in dbSNP, with allele frequency information included; see below) and the provision of additional literature citations where the pathogenicity of the variant may have been subsequently either questioned or confirmed. This latter point is particularly pertinent in the clinical setting, where a greater burden of proof may be required for use in diagnostic and predictive medicine, and when considering the return of incidental findings to patients after testing (Green et al. 2012, 2013; Ng et al. 2013; Gonsalves et al. 2013).

HGMD Professional

HGMD Professional has been developed to serve as the subscription version of HGMD, and is available to both commercial and academic customers under license from BIOBASE GmbH. HGMD Professional allows access to up-to-date mutation data with a quarterly release cycle; this version is therefore essential for checking the novelty of newly found mutations. HGMD Professional contains many features not available in the free public version (Table 2). More powerful search tools in the form of an expanded search engine with full text Boolean searching are provided. A batch search mode has recently been developed to allow users to search HGMD using gene (e.g. OMIM IDs) and variant (e.g. dbSNP IDs) oriented lists. Users can employ these tools to perform additional searches for gene-specific (e.g. chromosomal locations, gene names/aliases and gene ontology), mutation-specific (e.g. chromosomal coordinates, HGVS nomenclature, dbSNP ID) or citation-specific (e.g. first author, publication year, PubMed ID) information. The provision of chromosomal coordinates (hg19) for the vast majority of our nucleotide substitutions (98.7 % coverage) and other micro-lesions (97.3 % coverage) has made HGMD an invaluable tool for the large-scale analysis of NGS datasets such as the 1000 Genomes Project (1000 Genomes Project Consortium 2010, 2012). Additional information is also provided on a mutation-specific basis including curatorial comments pertaining to particular mutations (for example, if the mutation data presented required in-house correction in relation to the data presented in the original publication [5–10 % of entries], or if the clinical phenotype is associated with a more complex, i.e. a digenic or SNP in-*cis* inheritance pattern), additional reports comprising functional characterisation, further phenotypic information, comparative biochemical parameters, evolutionary conservation and SIFT (Sim et al. 2012) and MutPred (Li et al., 2009) predictions. These additional annotations are updated on a regular basis.

Recently, HGMD clinical phenotypes have been annotated against the Unified Medical Language System (UMLS) using a combination of manual curation and natural language processing. The UMLS is a comprehensive collection of biomedical concepts and the relationships between them (http://www.nlm.nih.gov/research/umls/). These UMLS mappings provide users with a more accurate and expanded phenotype search. Thus, searches using alternative disease names will return the same result-set, e.g. a search for "breast cancer" would yield identical results to a search for "malignant breast tumour". In addition, utilizing the UMLS allows for powerful semantic searching (e.g. searches for all mutations linked to blood disorders or all immune disorders).

Another new feature involves the highlighting of HGMD entries where the pathogenicity of the variant may have been cast into doubt by virtue of its allele frequency. HGMD Professional now displays a frequency flag when a listed variant is also found in dbSNP, and population frequency data from the 1000 Genomes Project are also provided. HGMD data have also recently been made available in Variant Call Format or VCF (Danecek et al. 2011), which will facilitate the comparison of HGMD with large NGS datasets. In addition to searching and viewing mutation data in a variety of ways, users of HGMD Professional may utilize a new feedback facility to submit corrections to the database curators or to request additional features.

HGMD Professional also contains an Advanced Search suite which has been designed to enhance mutation searching, viewing and retrieval. Two of the main types of mutation in HGMD (single-nucleotide substitutions and micro-lesions) can be interrogated with this toolset. Datasets for more than one mutation type may be combined (for example, micro-deletions, micro-insertions 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 functional profile (e.g. in vitro and in silico characterized transcription factor binding sites, post-translational modifications, microRNA binding sites, upstream ORFs, and catalytic residues, see Fig. 3); amino-acid change; nucleotide substitution; size and/or sequence composition of micro-deletions, micro-insertions or indels; pre- or user-defined sequence motifs (both those created and those abolished by the mutation); dbSNP number; keywords

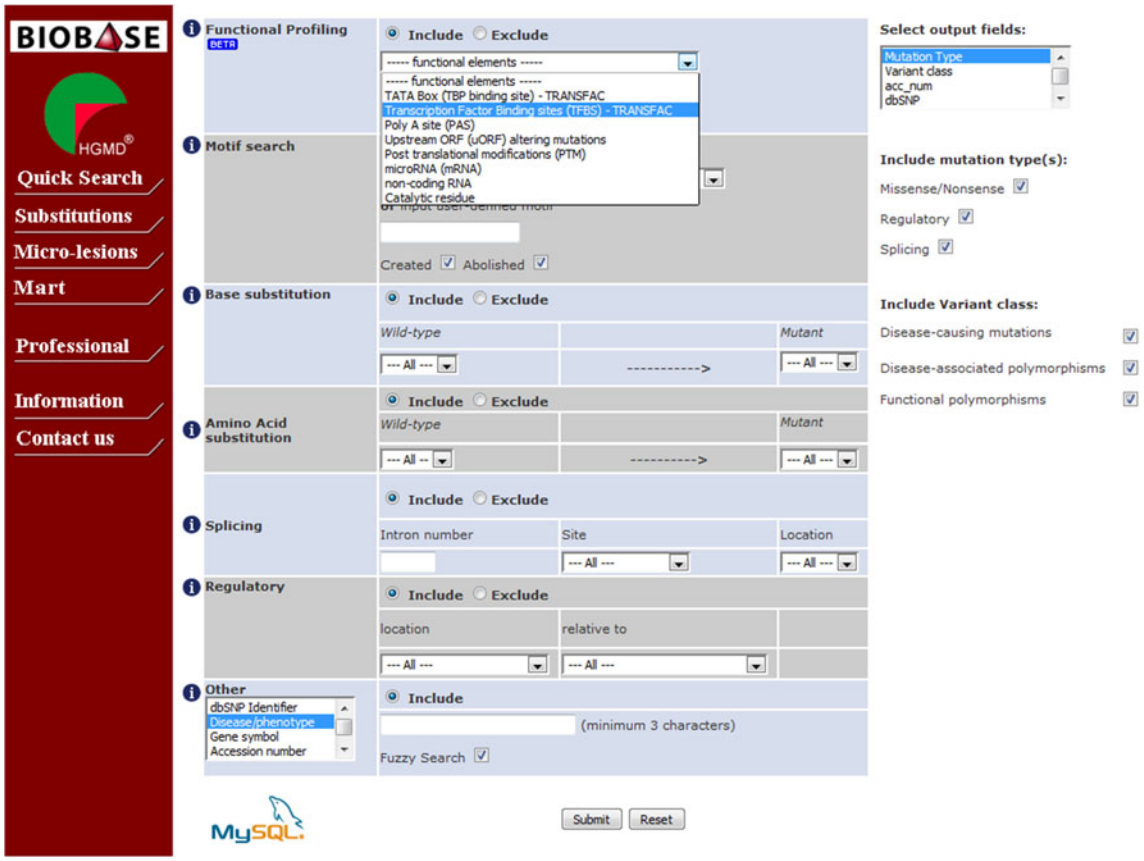

Fig. 3 Advanced nucleotide substitutions search in HGMD Professional

found in the article title or abstract. Results returned by the Advanced Search can be downloaded as tab-delimited text or a genome browser track, ready to be used in different applications. The Advanced Search also includes a batch mode called "Mutation Mart" to query HGMD via multiple identifiers including dbSNP, Entrez gene (http://www.ncbi.nlm.nih.gov/gene) and PubMed. HGMD Professional is available to subscribers either as an online only package or in downloadable form enabling users to incorporate HGMD data into their local variant analysis pipelines (http://www.biobase-international.com).

Other variant databases

Several other databases are available that attempt to record disease-causing or disease-associated (i.e. pathogenic) variation. These include the Online Mendelian Inheritance in Man, OMIM (http://www.omim.org/; Amberger et al. 2009), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), dbSNP (http://www.ncbi.nlm.nih.gov/SNP/; Sherry et al. 2001) and an assorted collection of locus-specific mutation databases (LSDBs) (http://www.hgvs.org/dblist/glsdb.html). OMIM does not provide statistics for allelic variants on its website; however, 22,901 germline OMIM variants appear to have been added to ClinVar, which itself contains a total of 25,375 pathogenic and probable pathogenic germline variants, while dbSNP contains 23,973 pathogenic or probable pathogenic germline variants (all databases were accessed July 10th 2013). Owing to the highly dispersed nature of the LSDBs and the potential for duplication between databases, accurate statistics with regard to like-for-like bona fide germline disease-causing (not merely neutral) variation is difficult to obtain. Since OMIM only records a limited number of variants per gene, and ClinVar is still in its infancy, HGMD is the only database of human pathological mutations that approaches comprehensive coverage of the peer-reviewed literature (Peterson et al. 2013). Since ClinVar and the LSDBs contain unpublished (non-peer reviewed) mutation data, the question has arisen as to whether HGMD should also include these data (Patrinos et al. 2012). However, several obstacles have been encountered by the LSDBs, including serious problems pertaining to data quality as well as issues of data provenance and consent. HGMD has therefore taken the decision not to include such data at this time.

How HGMD is utilized

Registered users of the public HGMD website currently number in excess of 60,000. Users may not download HGMD data in their entirety. However, mutation data may be made available at the discretion of the curators for non-commercial research purposes. Potential collaborators who wish to access HGMD data in full are required to sign a confidentiality agreement.

HGMD data have been used to perform an extensive series of meta-analyses on different types of gene mutation causing human inherited disease. These studies have helped to improve our understanding of mutational spectra and the molecular mechanisms underlying human inherited disease (Cooper et al. 2011). They have served to demonstrate not only that human gene mutation is an inherently non-random process but also that the nature, location and frequency of different types of mutation are shaped in large part by the local DNA sequence environment (Cooper et al. 2011). HGMD data have been used extensively in several international collaborative research projects including the 1000 Genomes Project (1000 Genomes Project 2010, 2012), where a surprising number of HGMD variants were found in apparently healthy individuals. They have also been used in the comparative analysis of several orthologous genomes including gorilla (Scally et al. 2012), cynomolgus and Chinese macaque (Yan et al. 2011), Rhesus macaque (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) and rat (Rat Genome Sequencing Project Consortium 2004), in which many apparently disease-causing mutations in human were found as wild type ('compensated mutations').

In a clinical setting, HGMD is widely utilized by many groups in ongoing NGS diagnostic (Johnston et al. 2012, Calvo et al. 2012, Bell et al. 2011) and human genome sequencing (Tong et al. 2010; Kim et al. 2009) programmes. HGMD has also been used by a number of different groups to aid the development of post-NGS variant interpretation algorithms including MutPred (Li et al. 2009), PROVEAN (Choi et al. 2012), CAROL (Lopes et al. 2012), CRAVAT (Douville et al. 2013), NEST (Carter et al. 2013) and FATHMM (Shihab et al. 2013). Finally, HGMD has been used as a resource for structural biologists in the reconstruction of protein interaction networks (Wang et al. 2012; Guo et al. 2013). A more complete list of articles which have utilized HGMD data or expertise in their production can be found on the HGMD website (http://www.hgmd.cf.ac.uk/docs/articles.html).

Data sharing

A limited HGMD data set, containing both chromosomal coordinates and HGMD identifiers, has been made available via academic data exchange programmes to the Gen2Phen project (Webb et al. 2011), the European Bioinformatics Institute (EBI)/Ensembl (Flicek et al. 2013) and the University of California, Santa Cruz (UCSC) (Meyer et al. 2013) and may be viewed in these projects' respective genome browsers. Data from HGMD Professional have additionally been made available to HGMD subscribers via Genome Trax™ (BIOBASE GmbH) and Alamut (Interactive Biosoftware), but are also accessible as part of the HGMD Professional stand-alone package (BIOBASE GmbH). Allowing free access to the bulk of the mutation data present in HGMD, while generating sufficient income from its commercial distribution to support its maintenance and expansion, represents a business model that should maximize the availability of HGMD at the same time as ensuring its long-term sustainability. Although we are necessarily obliged to be prudent with regard to data sharing with public data repositories, we have always taken the view that making as much data as possible publicly available is generally beneficial to both HGMD and its users worldwide.

Future plans

The provision of chromosomal coordinates for the vast majority of coding region micro-lesions in HGMD is now complete. Expanding this provision to include micro-lesions in non-coding regions and the gross and complex lesion dataset (where feasible) is a high priority, as is expanding the provision of genomic coordinates to include popularly utilized NGS formats such as General Feature Format (GFF) (http://www.sanger.ac.uk/resources/software/gff/) and BED format, to complement the recently added HGMD Variant Call Format (VCF) (Danecek et al. 2011). Mutations will also be mapped to the new genome build (GRCh38) in due course. A listing of removed variants will be implemented as time allows. Provision of genomic reference sequences based on the NCBI RefSeqGene project (Pruitt et al. 2009), links to available protein structures and homology models, and mapping HGMD phenotypes to the Human Phenotype Ontology (HPO) are also regarded as priorities.

In its current state of development, HGMD provides the user with a unique resource that can be utilized not only to obtain evidence to support the pathological authenticity and/or novelty of detected gene lesions and to acquire an overview of the mutational spectra for specific genes, but also as a knowledgebase for use in the bioinformatics and whole genome screening projects that underpin personalized genomics.

Conflict of interest The authors wish to declare an interest in so far as HGMD is financially supported by BIOBASE GmbH through a license agreement with Cardiff University.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online mendelian inheritance in man (OMIM). Nucleic Acids Res 37:D793–D796

Andreasen C, Refsgaard L, Nielsen JB, Sajadieh A, Winkel BG, Tfelt-Hansen J, Haunsø S, Holst AG, Svendsen JH, Olesen MS (2013) Mutations in genes encoding cardiac ion channels previously associated with sudden infant death syndrome (SIDS) are present with high frequency in new exome data. Can J Cardiol 29(9):1104–1109

Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3:65ra4

Calvo SE, Compton AG, Hershman SG, Lim SC, Lieber DS, Tucker EJ, Laskowski A, Garone C, Liu S, Jaffe DB, Christodoulou J, Fletcher JM, Bruno DL, Goldblatt J, Dimauro S, Thorburn DR, Mootha VK (2012) Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Sci Transl Med 4:118ra10

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 14(Suppl 3):S3

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012:e46688

Cirulli ET, Goldstein DB (2007) In vitro assays fail to predict in vivo effects of regulatory polymorphisms. Hum Mol Genet 16:1931–1939

Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum Mutat 31:631–655

Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM (2011) On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Hum Mutat 32:1075–1099

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132:1077–1130

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325:1246–1250

Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R (2013) CRAVAT: cancer-related analysis of VAriants Toolkit. Bioinformatics 29:647–648

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM (2013) ENSEMBL 2013. Nucleic Acids Res 41:D48–D55

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39(Database issue):D945–D950

Gonsalves SG, Ng D, Johnston JJ, Teer JK, NISC Comparative Sequencing Program, Stenson PD, Cooper DN, Mullikin JC, Biesecker LG (2013) Using exome data for opportunistic screening of malignant hyperthermia susceptibility. Anesthesiology 6(4):337–346

Green RC, Berg JS, Berry GT, Biesecker LG, Dimmock DP, Evans JP, Grody WW, Hegde MR, Kalia S, Korf BR, Krantz I, McGuire AL, Miller DT, Murray MF, Nussbaum RL, Plon SE, Rehm HL, Jacob HJ (2012) Exploring concordance and discordance for return of incidental findings from clinical sequencing. Genet Med 14:405–410

Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15:565–574

Guo Y, Wei X, Das J, Grimson A, Lipkin SM, Clark AG, Yu H (2013) Dissecting disease inheritance modes in a three-dimensional network challenges the "guilt-by-association" principle. Am J Hum Genet 93:78–89

Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, Mullikin JC, Biesecker LG (2012) Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. Am J Hum Genet 91:97–108

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS (2009) A highly annotated whole-genome sequence of a Korean individual. Nature 460:1011–1015

Lazarin GA, Haque IS, Nazareth S, Iori K, Patterson AS, Jacobson JL, Marshall JR, Seltzer WK, Patrizio P, Evans EA, Srinivasan BS (2013) An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals. Genet Med 15:178–186

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P (2009) Automated inference of

molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750

Lopes MC, Joyce C, Ritchie GR, John SL, Cunningham F, Asimit J, Zeggini E (2012) A combined functional annotation score for non-synonymous variants. Hum Hered 73:47–51

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ (2013) The UCSC Genome Browser database: extensions and updates. Nucleic Acids Res 41(1):D64–D69

Ng D, Johnston JJ, Teer JK, Singh LN, Peller LC, Wynter JS, Lewis KL, Cooper DN, Stenson PD, Mullikin JC, Biesecker LG (2013) Interpreting secondary cardiac disease variants in an exome cohort. Circ Cardiovasc Genet 6:337–346

Nishiguchi KM, Rivolta C (2012) Genes associated with retinitis pigmentosa and allied diseases are frequently mutated in the general population. PLoS ONE 7:e41902

Patrinos GP, Cooper DN, van Mulligen E, Gkantouna V, Tzimas G, Tatum Z, Schultes E, Roos M, Mons B (2012) Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. Hum Mutat 33:1503–1512

Peterson TA, Doughty E, Kann MG (2013) Towards precision medicine: advances in computational approaches for analysis of human variants. J Mol Biol. doi:10.1016/j.jmb.2013.08.008

Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI reference sequences: current status, policy and new initiatives. Nucleic Acids Res 37:D32–D36

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234

Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res 35:D823–D828

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 34:57–65

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40:W452–W457

Thorn CF, Klein TE, Altman RB (2010) Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics 11:501–505

Tong P, Prendergast JG, Lohan AJ, Farrington SM, Cronin S, Friel N, Bradley DG, Hardiman O, Evans A, Wilson JF, Loftus B (2010) Sequencing and analysis of an Irish human genome. Genome Biol 11:R91

Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 30:159–164

Webb AJ, Thorisson GA, Brookes AJ, GEN2PHEN Consortium (2011) An informatics project and online "Knowledge Centre" supporting modern genotype-to-phenotype research. Hum Mutat 32:543–550

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C, the 1000 Genomes Project Consortium (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91:1022–1032

Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, Du H, Chen J, Chen R, Zhang P, Huang Z, Thompson JR, Meng Y, Bai Y, Wang J, Zhuo M, Wang T, Huang Y, Wei L, Li J, Wang Z, Hu H, Yang P, Le L, Stenson PD, Li B, Liu X, Ball EV, An N, Huang Q, Zhang Y, Fan W, Zhang X, Li Y, Wang W, Katze MG, Su B, Nielsen R, Yang H, Wang J, Wang X, Wang J (2011) Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol 29:1019–1023