QIAGEN HGMD® 2020

HGMD® Professional 2020.3 Global Documentation

Sample to Insight

QIAGEN

Contents

Introduction

Citing HGMD®

If you refer to HGMD® in any publication, please cite:

Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, Cooper DN (2020) The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* epub. PubMed 32596782.

Rationale

Human gene mutation is a highly specific process, and this specificity has important implications for the nature, prevalence and therefore diagnosis of genetic disease. Indeed, the recognition that certain DNA sequences are hypermutable has yielded clues as to the endogenous mutational mechanisms involved and provided insights into the intricacies of the processes of DNA replication and repair (Cooper and Krawczak 1993). In practical terms, a fuller understanding of the mutational process may prove important in molecular diagnostic medicine by contributing to improvements in the design and efficacy of mutation search procedures and strategies in different genetic disorders.

The Human Gene Mutation Database (HGMD®) represents an attempt to collate known (published) gene lesions responsible for human inherited disease. This database, whilst originally established for the study of mutational mechanisms in human genes (Cooper and Krawczak 1993), has now acquired a much broader utility in that it embodies an up-to-date and comprehensive reference source to the spectrum of inherited human gene lesions. Thus, HGMD® provides information of practical diagnostic importance to (i) researchers and diagnosticians in human molecular genetics, (ii) physicians interested in a particular inherited condition in a given patient or family, (iii) genetic counsellors, (iv) personal genomics and NGS researchers.

Data coverage

The Human Gene Mutation Database includes the first example of all mutations causing or associated with human inherited disease, plus disease-associated/functional polymorphisms reported in the literature. HGMD® may also include additional reports for certain mutations if these reports serve to enhance the original entry (e.g. functional studies).

These data comprise various types of mutation within the coding regions, splicing and regulatory regions of human nuclear genes. Somatic mutations and mutations in the mitochondrial genome are thus not included, although in the latter case, links to Mitomap are provided. Each mutation is entered only once in order to avoid confusion between recurrent and identical-by-descent lesions.

HGMD® does not usually include mutations lacking obvious phenotypic consequences although a few such variants have been included where they could conceivably have some clinical effect (e.g. albumins, butyrylcholinesterases). Many published mutation searches identify more than one genetic change in a single patient. In such cases, the relationship between a given lesion and the clinical phenotype is not always immediately clear, and the curators of HGMD® have had to rely exclusively upon the judgements of authors, peer reviewers and journal editors. The possibility of unintentional inclusion of some lesions with little or no pathological significance can therefore not be ruled out.

HGMD® includes disease-associated/functional polymorphisms. To be included, there must be a convincing association of the polymorphism with the disease/functional phenotype. More information on this can be found in the polymorphism inclusion criteria.

HGMD® gene and variant data are supplemented by various meta-data from different sources. Such data are included to enhance the core gene/variant report to assist users to interpret the data in the way that they desire. These data include *in silico* predictions from the dbNSFP3 database (https://sites.google.com/site/jpopgen/dbNSFP), population frequency data from gnomAD (http://gnomad.broadinstitute.org/), typical inheritance pattern expected for the predominant phenotype(s) in each gene (AD, AR, ADAR, XLD, XLR, and YL, only where such data may be unequivocally assigned), gene ontology (http://www.geneontology.org/) and the Unified Medical Language System disease ontology terms (https://www.nlm.nih.gov/research/umls/).

Evidence for pathological authenticity

Pathological mutations that dramatically disrupt the structure of a given gene are self-evidently very likely to be responsible for the associated clinical phenotype. However, for other categories of lesion, pathological mutations are often difficult to distinguish from polymorphisms with little or no clinical significance, particularly if their structural or functional consequences are subtle (Cotton and Scriver, 1998). Evidence for their authenticity in a pathological context therefore usually comes from one or more different lines of evidence:

- Absence in normal controls.
- Novel appearance and subsequent cosegregation of the lesion and disease phenotype through the family pedigree.
- Absence of any other lesion in the gene that could be responsible for the observed clinical phenotype.
- Previous independent occurrence in an unrelated patient.
- Non-conservative amino acid substitutions are more likely to disrupt protein function.
- Location in a protein region of known structural or functional importance.
- Location in an evolutionarily conserved nucleotide sequence and/or amino acid residue.
- In vitro demonstration of reduced gene expression/mRNA splicing/activity or stability of protein product consequent to mutation.
- Demonstration that the mutant protein has the same properties in vitro as its in vivo mutant counterpart.

- Reversal of the pathological phenotype in patient/cultured cells by gene replacement.

Despite the best efforts of the HGMD® curators, it may be assumed that some categories of gene lesion listed in HGMD® (e.g. missense mutations, regulatory mutations, splicing mutations) are likely to include entries that are not actually causative even though they have been reported as such. In some cases the evidence for pathogenicity may be dubious; such variants can be identified by the addition of a question mark (?) to the given disease/phenotype, which indicates that some degree of uncertainty is involved.

Data collection

Data are collected by the manual and computerised screening of journals and publicly available locus specific databases (LSDBs). Where possible, data are included from the original reports; entries are referenced to 'Mutation Updates' and review articles if the original publication is not available. Please note that ambiguously-described mutations are not included in the database until clarification has been obtained from the authors.

Curation policy

Disease-causing mutations are entered into HGMD® where the authors of the associated reference indicate that the alteration described confers the clinical phenotype specified upon the individuals concerned. Disease-associated polymorphisms (DPs) are entered into HGMD® where there is evidence for a significant association with a clinical or laboratory phenotype along with additional evidence that the polymorphism is itself of likely functional relevance (e.g. missense change, alters transcription factor/miRNA binding site etc.). Functional polymorphisms (FPs) are entered into HGMD® where the authors have demonstrated that the polymorphism in question exerts a direct functional effect (e.g. as evidenced by a luciferase reporter gene assay). Disease-associated polymorphisms with supporting functional evidence (DFPs) must meet both of the above criteria.

The HGMD® curators have adopted a policy of continual assessment of the curated content with respect to the mutation entries in the database. If and when additional important new information pertaining to a specific mutation entry becomes available (e.g. questionable pathogenicity, confirmed pathogenicity, additional phenotypes, population frequency, functional studies etc.), the mutation entry may be revised or recategorized. 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) entry is likely to be of questionable pathological relevance or even a neutral polymorphism (on the basis of additional case reports, genome/population screening studies etc.), it may be flagged with a question mark (DM?) or even retired from the database entirely if it turns out to have been erroneously included ab initio.

The majority of clinical phenotypes assigned to DMs in HGMD® represent rare conditions that most people would consider to be "diseases". However, it is important to note that HGMD® also considers a "silent" protein deficiency or biochemical phenotype (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.

HGMD® users should not assume that just because a mutation is labelled "DM", that it automatically follows that the HGMD® curators are certain that the mutation is disease causing in all individuals harbouring it (i.e. that this mutation is deemed to be fully penetrant). As geneticists, we know that this is not invariably going to be the case. Indeed, it is likely that next generation sequencing programmes (such as the 1000 Genomes Project) will identify considerable numbers of "DM" mutations in apparently healthy individuals (MacArthur et al. 2012). Such lesions should not be regarded automatically as being clinically irrelevant because it is quite possible that these mutations may be low-penetrance or late onset disease susceptibility alleles rather than neutral variants. 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 to our users wherever feasible), rather than run the risk of inadvertently excluding a variant that may be directly relevant to disease.

Mutation categories

Table 1. Summary of mutation categories in HGMD®

Types	Description	Genomic coordinates	HGVS	Web interface
Missense/nonsense	Single base-pair substitutions in coding regions	YES	YES	HGMD PRO and ADVANCED
Splicing	Single base-pair substitutions with consequences for mRNA splicing	YES	YES	HGMD PRO and ADVANCED
Regulatory	Single base-pair substitutions causing regulatory abnormalities	YES	NO	HGMD PRO and ADVANCED
Small deletions	Micro-deletions (20 bp or less)	YES	YES	HGMD PRO and ADVANCED
Small insertions	Micro-insertions (20 bp or less)	YES	YES	HGMD PRO and ADVANCED
Small indels	Micro-indels (20 bp or less)	YES	YES	HGMD PRO and ADVANCED
Gross deletions	Deletions over 20 bp	NO	NO	HGMD PRO
Gross insertions	Insertions over 20 bp	NO	NO	HGMD PRO
Complex rearrangements	Recorded in narrative format	NO	NO	HGMD PRO
Repeat variations	Recorded in narrative format	NO	NO	HGMD PRO

All HGMD® entries comprise a reference to the first literature report (primary reference) of a mutation, the associated disease state as specified in that report, and the gene name and official symbol (as recommended by the HUGO Gene Nomenclature Committee; HGNC). In cases where no official gene symbol exists, a provisional symbol has been assigned by the HGMD® curators, which is denoted by lower-case letters.

NCBI dbSNP numbers (where identified) may also be recorded. The inclusion of a dbSNP identifier for a HGMD® entry in no way implies that the entry in question is a polymorphism.

Missense/nonsense

Single base-pair substitutions in coding regions are presented in terms of the triplet change.

Splicing

Single base-pair substitutions with consequences for mRNA splicing are presented as the nucleotide change and the position relative to the donor or acceptor splice site of a specified intron. Positions given as positive integers refer to a 3' (downstream) location and negative integers refer to a 5' (upstream) location.

Regulatory

Substitutions causing regulatory abnormalities are logged with thirty nucleotides of upstream and downstream flanking sequence; the location of the mutation relative to the transcriptional initiation site, start codon or termination codon is given.

Small deletions

Micro-deletions (of 20 bp or less) are presented with the deleted bases in lower case plus, in upper case, 10 bases of sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character ("^"). Where deletions occur outside the coding region of the gene, other positional information is occasionally provided e.g. 5' UTR (5' untranslated region) or E6I6 (denotes exon 6/intron 6 boundary).

Small Insertions

Micro-insertions (of 20 bp or less) are shown with the inserted bases in lower case and 10 bases of upstream and downstream sequence (in upper case). The numbered codon is preceded in the sequence by the caret character ("^"). Where insertions extend outside the coding region, other positional information may be provided e.g. 3' UTR (3' untranslated region) or E12I12 (denotes exon 12/intron 12 boundary).

Small Indels

Micro-indels (of 20 bp or less) are presented with the deleted bases in lower case flanked by, in upper case, 10 bases of upstream and downstream sequence. The inserted nucleotides are shown in lower case. The numbered codon is preceded in the given sequence by the caret character ("^"). Where the lesion occurs outside the coding region of the gene, other positional information is occasionally provided e.g. c.252-32_-28 to indicate that the deleted bases occur at the -32_-28 positions relative to the exon starting at c.252.

Other mutation types

For gross deletions, gross insertions, repeat variations 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.

Sub-categorisation of data

Disease-associated polymorphism (DP)

A polymorphism reported to be in significant association with a disease/phenotype ($p<0.05$) that is assumed to be functional (e.g. as a consequence of location, evolutionary conservation, replication studies etc.), although there is as yet no direct evidence (e.g. from an expression study) of functional effect.

Disease-associated polymorphism with supporting functional evidence (DFP)

A polymorphism reported to be in significant association with disease ($p<0.05$) for which evidence of direct functional importance (e.g. as a consequence of altered expression, mRNA studies etc.) has been presented (either in the original report or in a subsequent study, which will be included as a secondary reference).

In vitro/laboratory or in vivo functional polymorphism (FP)

A polymorphism reported to affect the structure, function or expression of the gene (or gene product), but with no disease association reported as yet.

Disease causing mutation? (DM?)

Likely pathological mutation reported to be disease causing in the corresponding report, but where the author has indicated that there may be some degree of doubt, or subsequent evidence has come to light in the literature, calling the deleterious nature of the variant into question.

De novo mutations identified as part of a large-scale mutation screen for such variants in patients with disorders such as autism, schizophrenia and intellectual disability will be entered under the DM? category unless there is cogent evidence to support their inclusion as DMs. All likely disruptive sequence changes identified in cases (not controls, or unaffected siblings in parent-offspring groups) will be entered. Such variants will include single base substitutions causing missense, nonsense or canonical splice site changes as well as both small and large exonic frameshift deletions/insertions or other complex exonic rearrangements. Other variant types (e.g. synonymous substitutions) may be considered for inclusion if additional evidence supportive of pathogenicity is presented.

Disease causing mutation (DM)

Pathological mutation reported to be disease causing in the corresponding report (i.e. all other HGMD® data).

Retired entry (R)

An entry retired from HGMD® due to being found to have been erroneously included ab initio, or subject to correction in the literature resulting in the record becoming obsolete, merged or otherwise invalid.

HGMD computed rankscore

The ranking score is a single probability score between 0 and 1, with 1 being most likely disease-causing. The score is computed using a machine learning approach, and is based upon multiple lines of evidence, including HGMD literature support for pathogenicity, evolutionary conservation (100 way vertebrate alignment), variant allele frequency and in-silico pathogenicity prediction. Individually, scores may be interpreted as probabilities of pathogenicity (i.e. the higher the score the more likely the variant is disease-causing). Scores may also be utilised in aggregate to prioritize and rank multiple HGMD variants which have been found in the same sample (for example, one approach would be to select variants within the same variant class [e.g. DM], then sort by the rankscore in descending order). This feature is under ongoing development.

Inclusion Criteria for Disease-associated/Functional polymorphisms

Aim

HGMD® seeks to include DNA sequence variants that are either (i) disease-associated and of likely functional significance, or (ii) of clear functional significance even though no associated clinical phenotype may have been identified to date. In order to deal with published polymorphism reports describing potential disease associations in a methodical and uniform manner we have adopted the inclusion criteria set out below.

Background

At present, ~58% of the polymorphic variants recorded in HGMD® are 'disease-associated'. However, even in cases where no disease association has yet been demonstrated, functional polymorphisms that alter the expression of a gene or the structure/function of the gene product are potentially very important. Although a functional polymorphism with no disease association may not have any direct and/or immediate clinical relevance, these data are potentially very valuable in terms of understanding inter-individual differences in disease susceptibility. The vast majority of polymorphic variants in HGMD® are single nucleotide polymorphisms (SNPs) but a small number are of the insertion/deletion type. The polymorphic variants logged in HGMD® are generally located in either regulatory or coding regions of genes although it should be noted that SNPs occurring outside of these regions may nevertheless still have consequences for gene expression, splicing, transcription factor binding etc.

Definitions

The distinction between a disease-associated polymorphism and a pathological mutation is in practice often fairly arbitrary and is generally made in the context of the prevalence of the variant in the population as well as its penetrance (the frequency with which a specific genotype manifests itself as a given clinical phenotype). Variants with a minor allele frequency of >1% in the population being studied are, by convention, termed polymorphisms. These polymorphisms are identified in the database by the addition of 'association with' and 'association with?' to the clinical/laboratory phenotype description (the question mark indicates that the association is judged by the HGMD® curators to be somewhat tenuous).

Polymorphic variants logged in HGMD® usually fall into two discrete categories:

Disease-associated polymorphisms of presumed or proven functional significance (DP or DFP)

To be included as disease-associated, a statistically significant ($p<0.05$) association between the polymorphism and a clinical phenotype must have been reported. In addition, other information (e.g. in vitro or in vivo expression/functional data, replicated association studies, epidemiological studies, evolutionary conservation data etc.) should have been made available to support the contention that the polymorphism in question is itself of bona fide functional significance. Such a polymorphism (DP) could have consequences for gene expression, protein structure/function, gene splicing, etc. These supporting experimental data are required to ensure that non-causative variants (i.e. those merely in linkage disequilibrium with the actual causative variants) are not included. If the functional data required to support the inclusion of a disease-associated variant are contained in a subsequent article, the primary reference logged in HGMD® will be that in which the disease association was reported, with the paper containing the functional evidence being given as a secondary reference ('Functional characterisation').

Polymorphisms of functional significance with no reported disease association (FP)

If no clinical phenotype is known to be associated with a polymorphic variant, but sufficient in vitro or in vivo expression/functional data1 have nevertheless been presented to indicate functional significance, then the variant will be included in HGMD®. Typically, such data provide evidence for a direct effect on gene expression, protein structure and/or function, gene splicing etc. These variants can thus, in a very real sense, be considered as giving rise to a 'deficiency' (or occasionally a surfeit) of a given gene transcript or protein product. The phenotype recorded in HGMD® would give a brief description of the functional effect e.g. 'Reduced gene expression, association with'. If, at a later date, evidence becomes available to indicate that a disease/clinical phenotype is associated with the polymorphism, the 'functional' phenotype will be replaced by the disease/clinical phenotype, with the corresponding paper added as the primary reference. The earlier functional report will be included as a secondary reference and the entry will be re-categorised as "DFP". Polymorphic variants affecting individual drug responses, patient survival times after diagnosis, and responses to surgical intervention, are not included in HGMD®. Studies which simply report dbSNP numbers in association with disease (e.g. from large scale genome-wide association studies), with no additional evidence of direct functional involvement are also not included in HGMD®. Users interested in this particular category of variation should try other databases such as the Catalogue of Published Genome-Wide Association Studies (http://www.genome.gov/26525384/) or the Genetic Association Database (http://geneticassociationdb.nih.gov/).

One caveat to bear in mind is that in vitro studies are not invariably accurate indicators of the in vivo situation [see for example Cirulli and Goldstein (2007) & Dimas et al. (2009)].

In some instances, the above criteria may be only partially satisfied, such that the HGMD® curators remain unconvinced as to the functional/phenotypic relevance of the variant reported. In such cases, the polymorphism may be included as a result of (i) supporting information becoming available subsequent to publication of the original (first)

report, or (ii) because the associated gene/disease state was deemed to be of sufficient importance for the variant to warrant further study. 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.

Replication studies for disease-associated polymorphic variants

The replication of disease-association studies can be a source of additional information to satisfy the inclusion criteria. If a replication study serves to support a previously tenuous genotype-phenotype correlation, then the phenotype can be 'promoted' from 'association with?' to 'association with' and the replication study will be added as a secondary reference.

HGMD® wild type and mutant alleles may be reversed

In HGMD® the mutant allele is recorded as the allele responsible for the reported disease/phenotype. In some HGMD® records (typically polymorphism data, where both the wild type and mutant allele may be found at high frequency), the wild type as given by RefSeq may be the same as HGMD® mutant allele; this is so we preserve the relationship between reported disease/phenotype and mutant allele.

Other Categories of Variation

Copy number variations

Copy number variations (CNVs) are DNA segments >1 kb in length that present with variable numbers of copies in a given population. These variants are being reported in the literature with an ever increasing frequency. CNVs are potentially functionally significant and should therefore in principle be treated by HGMD® in a similar manner to any other polymorphism. However, human CNVs are already being collected by other databases such as the Database of Genomic Variants (http://projects.tcag.ca/variation/) and the Human Genome Structural Variation Project (http://humanparalogy.gs.washington.edu/structuralvariation/). CNVs that are disease-associated are also being collated in databases such as DECIPHER (http://www.sanger.ac.uk/PostGenomics/decipher/), the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (http://www.ecaruca.net) and the Chromosome Abnormality Database (http://www.ukcad.org.uk/cocoon/ukcad/). Whilst HGMD® does not wish to replicate the excellent curatorial work of other organisations, HGMD® is still interested in such variants if they are shown to be both of functional significance and associated with disease, and if they involve a single characterised gene that is itself clearly involved in the disease association.

Risk haplotypes

Reports of haplotypes associated with an increased risk of disease are not included in cases where there is no indication as to precisely which variant (or variants) within the haplotype is (are) responsible for the disease

association/functional effect. If, however, evidence is presented to support the contention that a single variant within the risk haplotype is causative and/or of functional significance to a degree which satisfies the inclusion criteria, then it would be included in HGMD®.

Web Interfaces (HGMD® PRO + HGMD® Advanced)

Have I got the right browser?

The HGMD® web pages should work correctly with any modern browser. These include the most popular browsers such as Microsoft Internet Explorer, Mozilla, FireFox and Opera. Text-only browsers such as Lynx can also operate HGMD® correctly although some pages will look odd.

We are working toward making our code compliant with HTML v5. If your browser is not capable of this, it will only mean that the presentation on your screen is not identical to ours. The results generated from search requests should not be affected.

Can I bookmark a particular page?

With web browsers it is possible to 'bookmark' individual pages from HGMD. However, we do not recommend this as not only are you likely to miss important notices placed on the HGMD® home page, but we reserve the right to change the internal structure of the database as required.

HGMD® PRO searches

There are five primary searches in HGMD® PRO: Gene, Mutation, Reference, Batch and Advanced. The Gene and Reference searches support Boolean searching, with the use of wildcards (*). UK and US spelling variations are accepted; for example, both "haemophilia" and "hemophilia" will return the F8 and F9 genes. Please note that alternate spelling functions only for whole words.

Boolean fulltext searching

+ operator indicates that the search term must be present in each result.

e.g. *+breast +cancer* returns results where both breast and cancer are present.

- operator indicates that the search term must not be present in any result.

e.g. *-breast +cancer* returns results that contain cancer but not breast.

* operator serves as the truncation (or wildcard). Unlike the other operators, it should be appended to the word to be affected.

e.g. *poly** returns results such as polyposis, polycystic and polypeptide.

" operator when used to enclose your search terms means that the search terms are treated literally (as they were typed).

e.g. *"hum mutat"* will return results containing the exact phrase "hum mutat".

Note1: If Boolean operators are not utilised, then multiple search terms are treated by a Boolean search as separate entities (e.g. a search for breast cancer without quotes [""] will return all results containing breast OR cancer.

Note2: HGMD® fulltext searching in MySQL has a minimum word length of 4 characters by default. Any search terms entered with less than 4 characters may not be recognised by the HGMD® fulltext search engine.

Non-boolean searching

This type of search works in a different way to the Boolean fulltext search. The fulltext index is not utilised, therefore there is no minimum word length in force (2 and 3 character searches will function). No operators are used, and multiple search terms will be treated as a literal phrase. Partial matches will also be returned by the search (except for the HGNC/OMIM/GDB/Entrez ID search, will return exact matches only).

Search for the word transporter.

- Returns results where the word transporter is present. Will also return partial matches such as transporters and cotransporter.

Search for the words prostate cancer.

- Returns results where the words prostate cancer occur as a literal phrase.

Search for the gene symbol APC.

- Returns results where APC occurs, including the APC gene itself, plus partial matches e.g. APCDD1 and PROC (where APC is an alias).

Note: This search does not utilise an index, and therefore may be somewhat slower than a Boolean fulltext search.

Gene search

Enter search term:

The genes present in HGMD® may be found by utilising seven different search options.

1. All fields search - Searches for your search terms in all fields listed below at once (2-7).
2. Gene symbol search - Searches HGMD® for the official HUGO Gene Nomenclature Committee gene symbol. Any gene symbol aliases that have been identified are also be included in this search.

Official symbol example: 'ABCC2'

Gene symbol alias example: 'MRP2'

3. Gene description search - Searches HGMD® for the official HUGO Gene Nomenclature Committee gene name. Any gene name aliases that have been identified are also be included in this search.

Official description example: 'ATP-binding cassette, sub-family C (CFTR/MRP), member 2 (CMOAT)'

Gene description alias example: 'Canalicular multispecific organic anion transporter'

4. Chromosomal location search - Searches HGMD® for the chromosomal location of HGMD® genes.

Example: '10q24'

5. HGNC/OMIM/GDB/Entrez ID - Searches HGMD® for the gene identifiers assigned by the HUGO Gene Nomenclature Committee database, Online Mendelian Inheritance in Man, the Genome Database (legacy only) and the Entrez Gene database.

HGNC Example: '5384'

OMIM Example: '601107'

GDB Example: '6089489'

Entrez Example: '1244'

6. Disease/phenotype search - Searches HGMD® for the disease/phenotype associated with reported mutations in HGMD® genes.

Example: 'Dubin-Johnson syndrome'

7. Gene ontology search - Searches for the ontology terms that have been assigned (by the Gene Ontology Consortium) to the genes present in HGMD®.

Example: 'organic anion transmembrane transporter activity', 'GO:0008514' or '0008514'

Detailed results will contain a list of gene information, diseases and external links/IDs associated with your search terms. Concise results are limited to gene information (symbol, description, chromosomal location) only.

To access the HGMD® record for a gene, click on the gene symbol listed. Some portions of the returned text may be highlighted in green. This indicates the part of the results that matches the search terms.

Secondary gene search

Symbol:

The secondary search allows users to go directly to a particular gene if the gene symbol is known. This search will only function with the correct HUGO Nomenclature Committee gene symbol.

Mutation search

There are five ways HGMD® may be searched for specific mutations.

1. Codon number search - Searches for mutations affecting a particular codon in the coding region. This search will return results from the missense/nonsense, small deletions, small insertions and small indels mutation categories. Results will contain a list of genes/diseases with links to both the gene page and specific mutations found during your search. For small deletions, small insertions and small indels, the codon number searched is that of the first affected codon, not the last whole codon as marked by "^" in the HGMD® entry (i.e. - in the entry CD010589 AAAS 156 TTGCGT^GTCTttGCATGGCACC, the first affected codon is 157). Please note that this search will not pick up mutations either beginning or wholly within an intron as there will not be a first affected codon to search for.

Example: '157'

NOTE: You can restrict your search to disease causing mutations or disease-associated/functional polymorphisms. The default is to search both types. Please note that for convenience, "frameshift or truncating variant" is included under both types.

2. Accession number search - Searches HGMD® for specific accession numbers (if known). Partial accession numbers will be accepted (with wildcards *). Results will contain a list of genes with links to both the gene page and specific accession number(s) found during your search.

Example: 'CM035497' or 'CM035*'

3. Search using official HGVS mutation nomenclature - Searches HGMD® for mapped mutations using the official nomenclature as described by den Dunnen and Antonarakis (2001). This search includes only the missense/nonsense mutations, small deletions, small insertions and small indels that have been mapped to the genome. Please note that for deletions, insertions and indels, HGVS nomenclature requires that the most 3' affected nucleotides are specified.

Example: '298C>T' or 'R100X'

Example: '20_21delTT' or '20_21del2' or '20_21del'

Example: '104_105insAA' or '104_105ins2' or '104_105ins' and '2195_2198dupAACA' or '2195_2198dup4' or '2195_2198dup'

Example: '385_386delAGinsGTT' or '385_386del2ins3' or '385_386delins'

4. Search using dbSNP identifier - Searches HGMD® for mutations that have a corresponding entry present in the NCBI dbSNP database. Partial 'rs' numbers will be accepted (with wildcards checked).

Example: 'rs1799963'

Example: 'rs1042522'

5. Search using chromosomal coordinates - Searches HGMD® for mutations using chromosomal coordinates. This search includes only then missense/nonsense mutations, small deletions, small insertions and small indels that have

been mapped to the genome (the February 2009 build; GRCh37/hg19). Users may enter a coordinate range, a single coordinate or an entire chromosome.

Example: 'chr1:2327114_2333800'

Example: 'chrX:77130681_77132008'

Example: 'chr7:1942983'

Example: 'chr2'

Phenotype search

There are two ways HGMD may be searched for phenotypes.

1. HGMD phenotype search - Searches the phenotypes recorded in HGMD directly. The phenotypes present in HGMD are generally recorded as they were initially reported in the corresponding literature article.
   Example: 'Cancer'

2. UMLS semantic search - Searches mapped UMLS ontology which includes OMIM, SNOMED CT and MeSH terms.
   Example: 'Cutis laxa'

Results will contain an alphabetical list of HGMD phenotypes and genes associated with your search terms.

Other phenotype search options

The ability to browse the phenotypes found in HGMD alphabetically (A-Z), by chromosome, or by mapped UMLS ontologies (currently OMIM, SNOMED CT, MeSH and HPO terms) is also available.

Reference search

There are six ways HGMD® may be searched for references.

1. All fields search - Searches for your search terms in all fields listed below (2-6).
2. First author search - Searches HGMD® for the first author (surname) of the mutation references found in HGMD®.

Example: 'Edwards'

3. PubMed journal search - Searches HGMD® for the PubMed journal title abbreviations associated with the mutation references found in HGMD®. Full journal names may also be used.
   Example: 'Am J Hum Genet' or 'American journal of human genetics'

4. PubMed ID search - Searches HGMD® for PubMed IDs associated with the mutation references found in HGMD®.

   Example: '9042910'

5. Publication year search - Searches HGMD® for the specific years in which the mutation references found in HGMD® were published.

Example: '1997'

6. HGMD® gene search - Searches HGMD® for the gene symbols associated with the mutation references found in HGMD®.

Example: 'TCOF1'

Results will contain a list of genes and references associated with your search terms. To access the HGMD® record for that gene, click on the gene symbol listed. You may also view the PubMed record (if available) associated with each search result.

Secondary reference search

Medline journal abbreviation:

The secondary search allows users to retrieve all data derived from a particular journal.

Batch search

The batch search is designed to accept a list of up to 500 variant or gene identifiers of various types. Identifiers accepted by the batch search include dbSNP, chromosomal coordinate, HGMD accession and VCF for variants and HUGO Nomenclature Committee gene symbols and IDs, Entrez Gene IDs and OMIM IDs for genes. Users should specify the required search option (via the radio button panel) and then either paste or type their list of identifiers into the search box (one per line), or upload their list via the upload box.

Users may also restrict their searches to the disease causing variant class (DM and DM?) only, or the disease-associated/functional polymorphic variant class (DFP, DP and FP) only. The default is to search both types at once. Users may further prioritise their results to return the most likely disease relevant variants first (based on literature evidence, in silico functional predictions and population frequency data).

Plain text format is required for uploaded files (one search term/identifier per line). Tab-delimited text should be used for VCF. Users should restrict their VCF files to CHROM POS ID REF ALT to keep the file size small, as there is a default upload limit of 2 MB.

1. dbSNP - Searches for variants with specified dbSNP 'rs' identifiers.

Example: 'rs6025'

2. Chromosomal coordinate - Searches for variants with specified chromosomal coordinates (GRCh37/hg19).

Example: 'chr1:169519049'

3. Variant Call Format (VCF) – Searches for variants with specified v4.0 compliant VCF. The CHROM, POS, ID, REF and ALT (tab-delimited) fields are relevant here.

Example: '1 2338019 ID1 CAG C . . .'

4. HGMD accession - Searches for variants with specified HGMD® accession numbers.

Example: 'CM940389'

5. PubMed ID – Searches by variants derived from specified PubMed identifiers.

Example: '24077912'

6. HUGO gene symbol - Searches for genes with specified HUGO Nomenclature Committee gene symbols.

Example: 'F5'

7. HUGO gene ID - Searches for genes with specified HUGO Nomenclature Committee gene identifiers.

Example: '3542'

8. Entrez Gene ID - Searches for genes with specified Entrez Gene identifiers.

Example: '2153'

9. OMIM ID - Searches for genes with specified OMIM gene identifiers.

Example: '612309'

Other search options

The ability to browse HGMD® genes by gene symbol (A-Z), chromosomal location (1-22, X and Y), or to view pre-queried HGMD® data (a random HGMD® gene entry, genes newly added for the current release, genes updated with new mutation data for the current release, genes by total number of mutations, or genes sorted by ontology term) is also available.

F.A.Q. search

The HGMD® Frequently Asked Questions (accessed through the Information menu) may be searched for keywords, or retrieved in its entirety. Boolean operators may be used (see above) and, as with HGMD® searching, there is a minimum word length in force, usually of 4 characters.

General remarks

Firstly, always make sure you are searching with the correct option selected. You cannot use a gene symbol search to find a disease. If you are getting error messages you need to alter your search strategy. If you are getting too many "gene not found" errors when searching for gene symbols, you should try searching alternate fields instead. For example, CD95 is the old symbol for FAS. It will not appear in a gene symbol search, but it will appear in a gene description or alias search. If you are not getting the results required when using a gene description or

disease/phenotype search, you can try to narrow your search. For example, entering "cancer" as a disease/phenotype search term will produce too many results. Narrowing it to something like "gastric cancer" may produce the desired results. Note also that there is a minimum word length of 4 characters for HGMD® searching, so entering any term with less than that will not produce any results.

Search types

The HGMD® Advanced Search has five different interfaces:

1. Nucleotide Substitutions PLUS search
2. Micro-lesions search
3. Quick search
4. MutationMart
5. Gene-based search

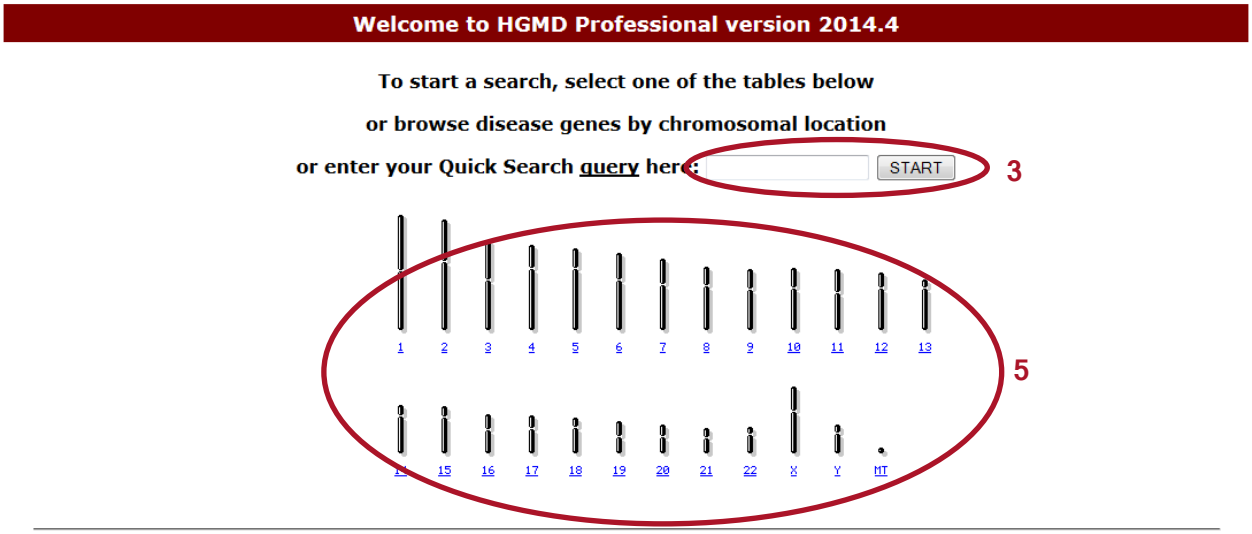

Figure 1. Access to HGMD® Advanced Search Interface

Nucleotide Substitutions PLUS search

Overview

The Nucleotide Substitutions PLUS search allows all HGMD® single-base substitutions (missense/nonsense, regulatory and splicing) to be interrogated via one interface.

The mutation types used by this search have been mapped onto genomic reference sequences (hg19) to allow flexible data-mining and motif-searching for transcription factors, conserved DNA domains or user-defined motifs.

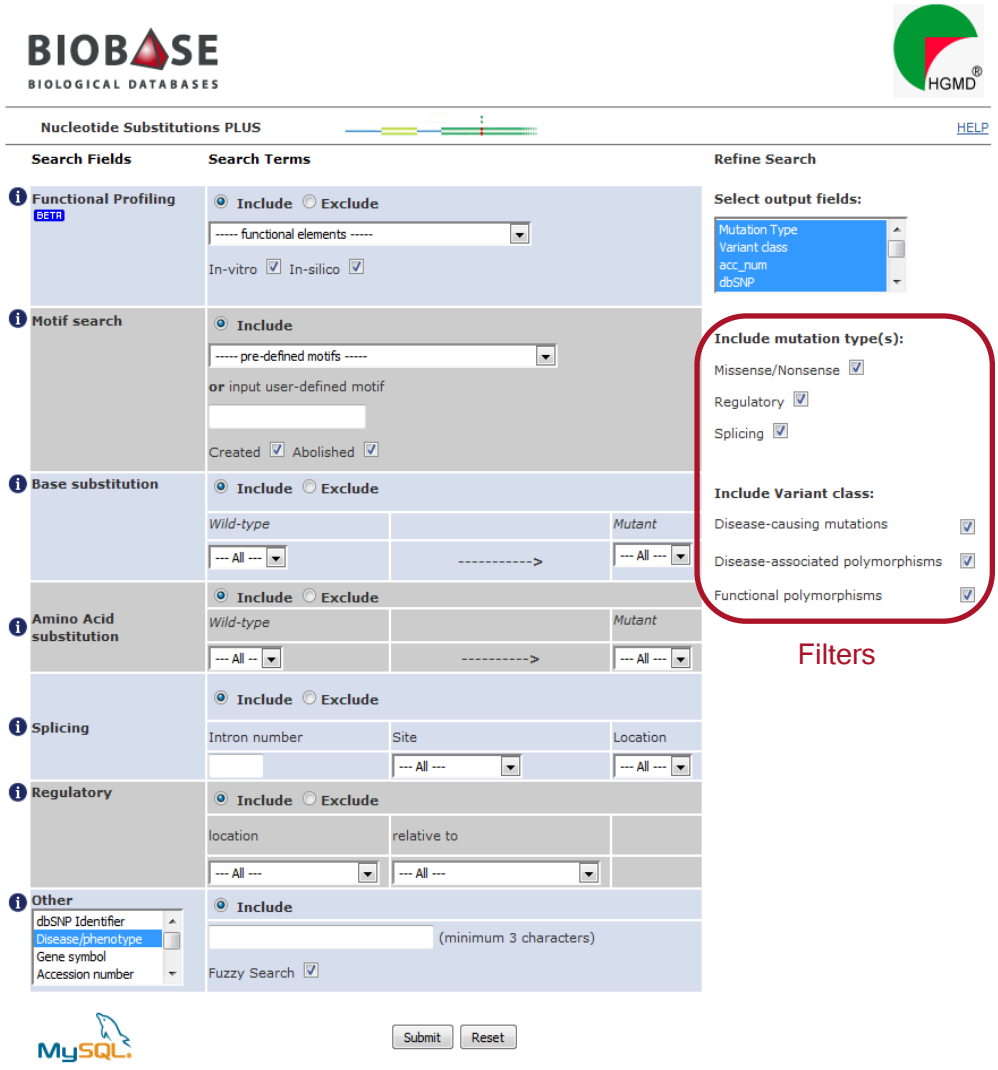

Figure 2. Nucleotide Substitutions PLUS Interface

How do I search?

Nucleotide Substitutions can be mined using one or more of the following categories:

- Motif Search
- Base Substitution
- Amino Acid Substitution
- Splicing
- Regulatory
- Other

Motif Search

Using the drop-down box predefined motifs can be selected or a user-defined motif can be typed into the text field to the side of the drop-down box.

The user-defined motif can be defined by using regular expressions. Some characteristics of extended regular expressions are:

'.' matches any single character.

A character class '[...]' matches any character within the brackets. For example, '[abc]' matches 'a', 'b', or 'c'. To name a range of characters, use a dash. '[a-z]' matches any letter, whereas '[0-9]' matches any digit.

Example 1: The regular expression [CA][CA]AGGTAGGTAA would match to the 5' splice site consensus sequence.

Example 2: The regular expression [GC]ATG would match to both GATG and CATG.

Example 3: The regular expression [AG][CT][AG] would match the sequence RYR.

Note - the motif search feature is not case sensitive and single letter codes are not supported.

If the check-box 'Created' is selected then all substitutions that create this motif will be returned.

If the check-box 'Abolished' is selected then all substitutions that disrupt this motif will be returned.

Base Substitution

Using the drop-down menus the wild type and/or the mutant nucleotides can be selected.

Amino Acid Substitution

Using the drop-down menus the wild type and/or the mutant amino acids can be selected.

Splicing

Mutations that have been shown to affect splicing can be searched for by:

- Entering an intron number into the ivs text box.
- The site of the substitution can be selected (either donor or acceptor splice site).
- The location relative to the splice site can also be selected.

Other

Select the field to search from the list. Enter the search term in the text field. If the 'Fuzzy Search' check box is selected then wildcards will be added to the beginning and end of the search term.

Combining of the "other" option with the categories listed above will narrow the results to e.g. to a certain disease/phenotype or gene.

Filters

Advanced Search Results can be filtered by mutation type:

- Missense/nonsense
- Regulatory
- Splicing

Advanced Search Results can also be filtered by variant class:

- Disease-causing mutations
- Disease-associated polymorphisms
- Functional polymorphisms

Output Fields:

1. Mutation type e.g. regulatory
2. Variant class e.g. DM
3. HGMD® id
4. dbSNP
5. Functional profiling result
6. Disease/Phenotype
7. Gene symbol
8. Entrez gene id
9. Link to PROTEOME locus report
10. HGVS nomenclature
11. Mutation description
12. Genomic coordinates (hg19 default)
13. MutPred
14. SIFT
15. Genomic sequence context
16. Primary Reference
17. cDNA Mutation display link

Functional Profiling of Mutation Data

HGMD® is in the process of annotating variants with both in vitro/in vivo and in silico data to help in the ascertainment of the molecular mechanism underlying the functional effect of a given mutation.

When this process has been completed, HGMD® aims to annotate mutation data for over 40 different types of functional site including exonic splice enhancers (ESE), post-translational modification sites and numerous transcription factor binding sites (TFBS) etc.

Please see the following paper for more information:

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P, Mooney SD: In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat. 31(3):335-346

Access to the functional profiling search tools is via the 'Nucleotide Substitutions Plus' Search.

An example functional profiling search:

Search for HGMD® regulatory variants disrupting transcription factor binding sites (TFBS) from TRANSFAC. Select 'TFBS' from functional profiling drop down menu (Figure 3).

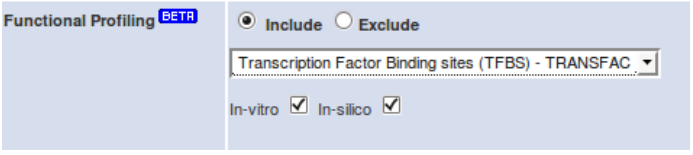

Figure 3. Drop down functional profiling menu from Nucleotide Substitutions PLUS

Functional sites disrupted by the mutation in question are shown in the functional Profile column (Figure 4, highlighted by red box).

Figure 4. Functional profiling results

Clicking on the relevant functional site shows the functional profiling mutation report (Figure 5) which displays both in vitro/in vivo and in silico evidence (where available) for any functional site disruption.

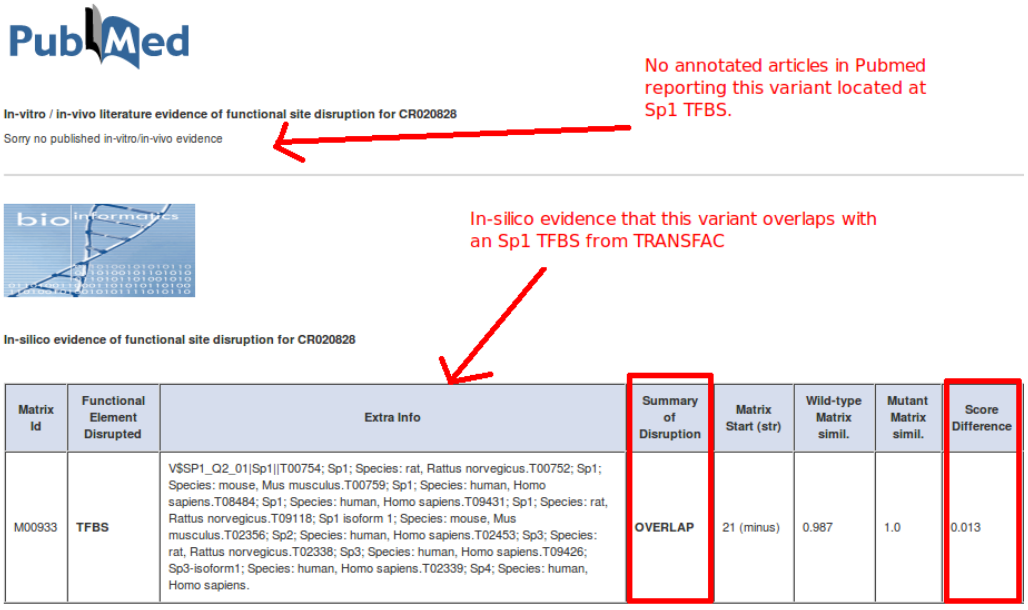

Figure 5. Functional profiling mutation report

Ability to download results (Figure 6):

1. Results downloaded as a tab delimited file
2. Genome browser track using hg19 coordinates

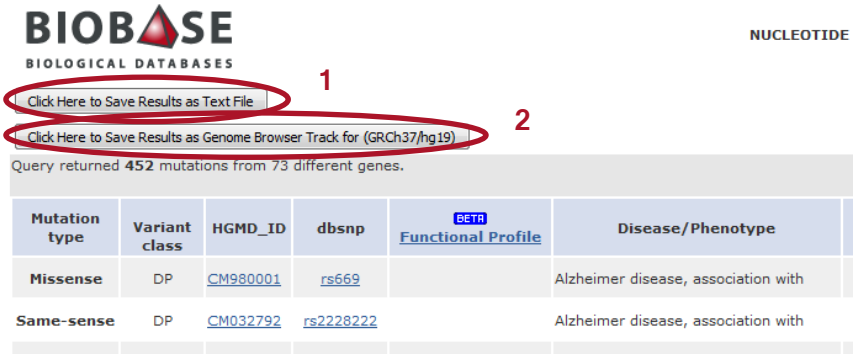

Figure 6. Downloading results from an Advanced Search query

Micro-lesions （Insertions, Indels and Deletions <21bp） search

This category comprises all small deletions, small insertions and small indels. This allows data-mining to be performed on a number of fields including deletion size, insertion size and user-defined motifs. Search results are available to download, using the same method as with the Nucleotide Substitutions Plus search.

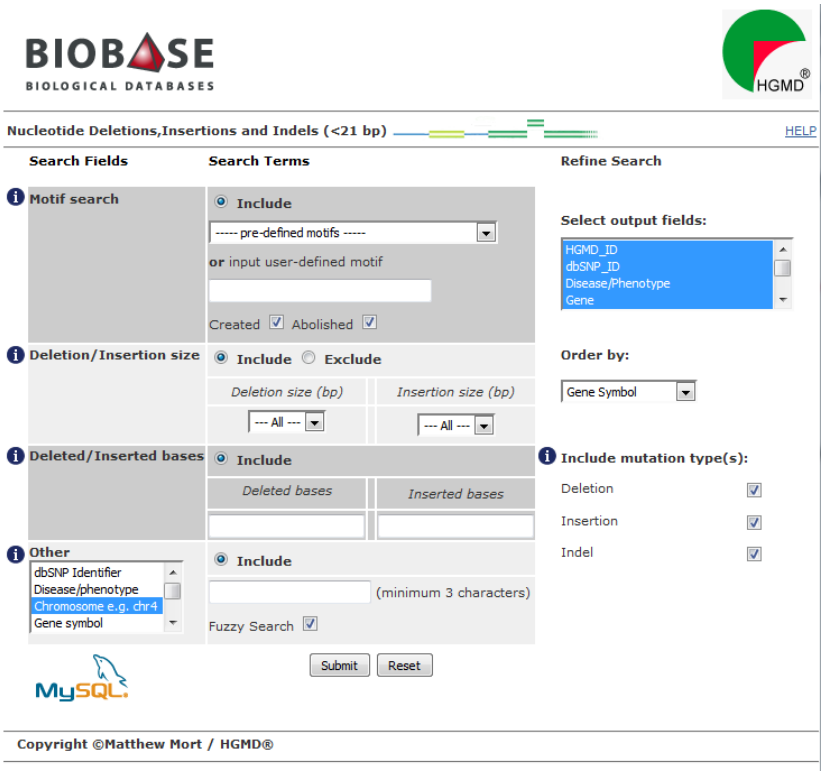

Figure 7. Micro-lesions search page

Output fields include:

Mutation type, HGMD® ID, dbSNP, Disease/Phenotype, Gene, HGVS, Genomic coordinates, Sequence context, codon, Deleted bases, Inserted bases, Nucleotide, Reference

Quick search (Search in both Nucleotide substitutions and Micro-lesions)

With the Quick search, selected fields in different tables (Nucleotide substitutions and Micro-lesions) can be searched at the same time. This search also queries the title of the original mutation report for key words.

This search option is especially useful to get a quick overview of the information available in different tables, e.g. for a certain disease.

The Quick search results are ranked by relevance and are assigned a Ranking rating. The 'quick search' ranking score relates to the number of matches found for the query keyword(s) across the gene symbol, disease term, title of mutation report, abstract of mutation report and dbSNP identifier fields. The higher the score the more relevant the mutation to the query keyword(s).

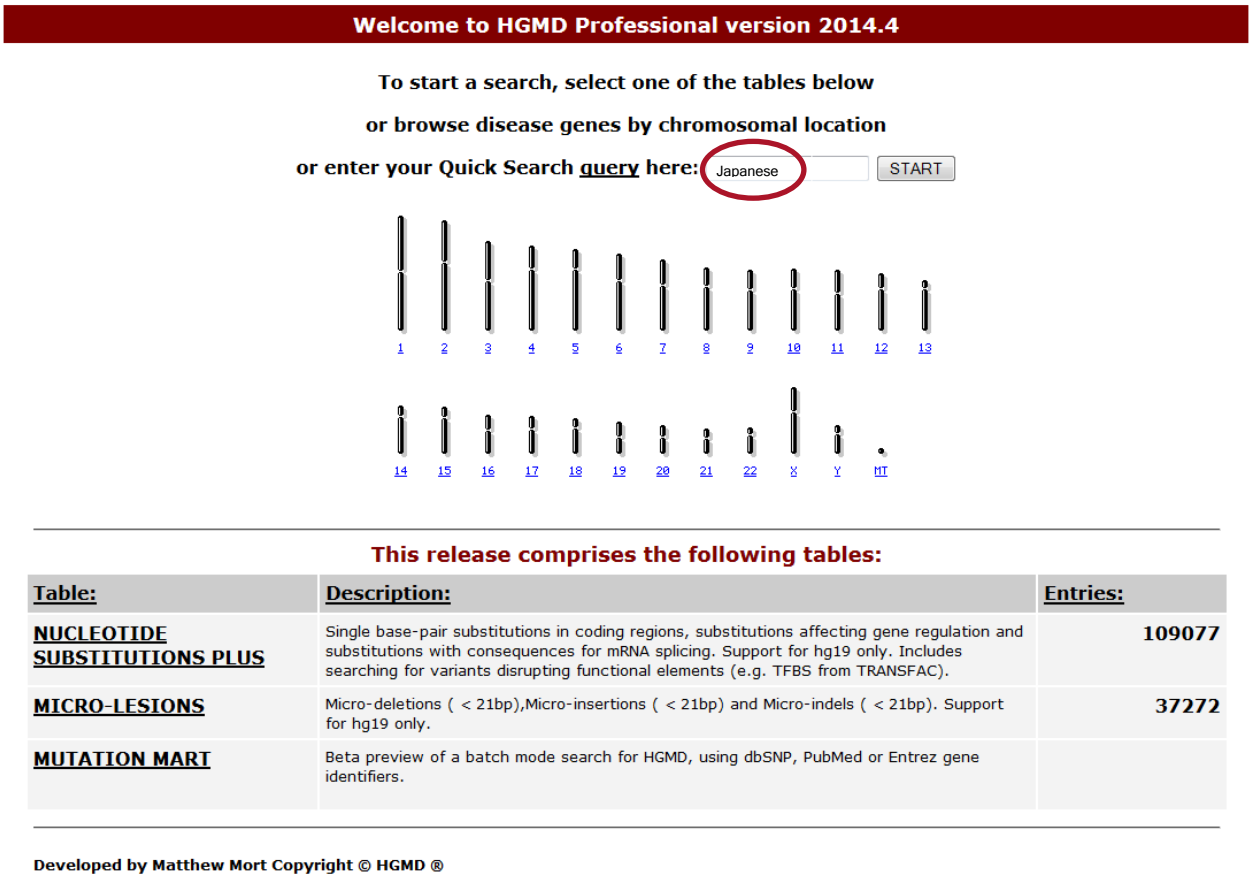

Figure 8. HGMD® Advanced Quick Search

Quick search examples

Example 1: Enter 'Japanese' as the search term (Figure 8).

This returns all mutations linked to a Mutation Report with 'Japanese' in the title and/or 'Japanese' in the Gene or Disease fields.

Example 2: Enter 'stroke candidate gene' as the search term.

This returns all mutations linked to a Mutation Report with 'stroke candidate gene' in the title and/or 'stroke candidate gene' in the Gene or Disease fields.

MutationMart: batch mode search

HGMD® can be queried in a batch mode using three different types of identifier:

1. dbSNP identifier e.g. rs1800072
2. PubMed identifier e.g. 20981092
3. Entrez gene identifier e.g. 1080

Up to 50 identifiers can be queried at once (place each identifier on a new line).

MutationMart Results can be downloaded to a tab delimited text file.

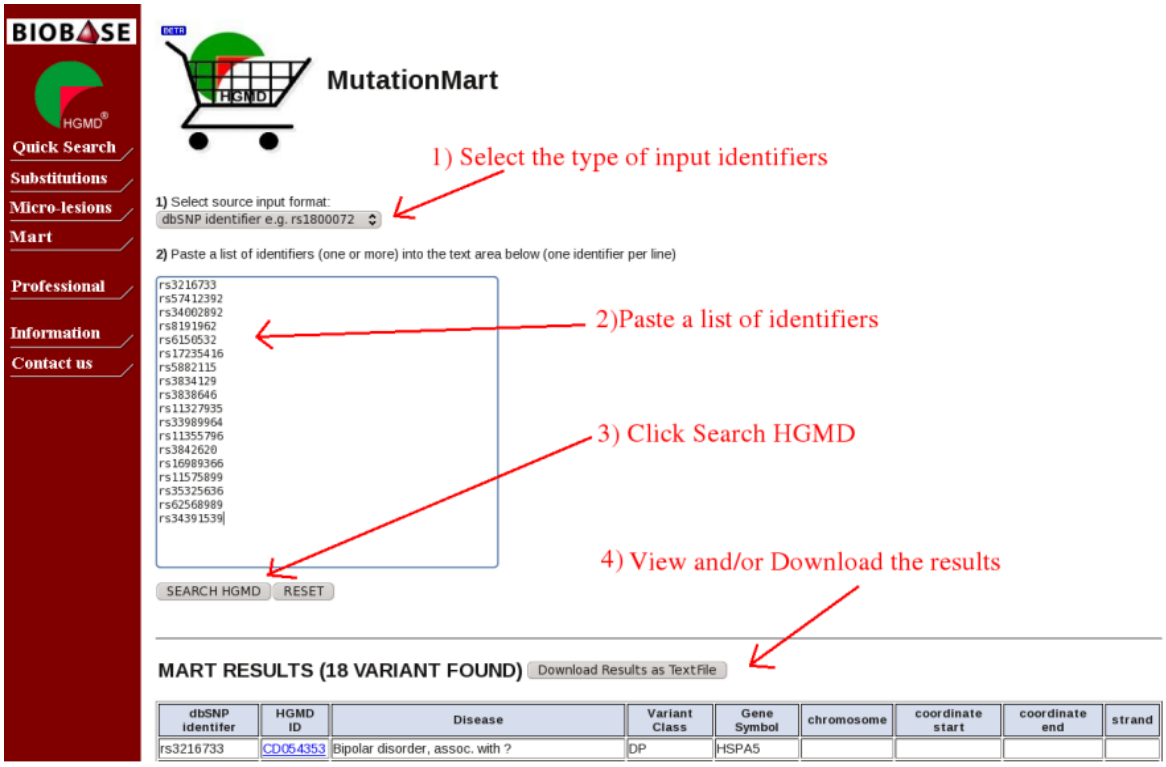

Figure 9. Description of the MutationMart

Gene-based search

Clicking on a chromosome identifier (1-22, X, Y, MT) will return a list of all the genes on that chromosome recorded in HGMD®. The list will include links to the cDNA mutation display for the genes in which substitutions, small deletions, small insertions and/or small indels in the coding region are recorded.

Genome-interpretation and NGS-based analysis using HGMD® Advanced

HGMD® Advanced provides a collection of utilities to assist in the analysis of NGS data:

1. Results from the Nucleotide Substitutions PLUS search, Micro-lesions search and Mutation Mart can be exported into custom genome browser tracks to allow importing into other software tools such as the UCSC genome browser and Genome Trax.
2. Functional profiling of mutation data to identify the underlying molecular mechanism including the mapping of HGMD® variants to TFBS from Transfac.
3. Providing annotations from SIFT (**S**orting **I**ntolerant **F**rom **T**olerant).
4. Providing annotations from MutPred.

Option 1: Exporting HGMD® data to other applications (e.g. Genome Trax or UCSC) by creating a Custom Genome Browser Track

Step 1: Creating your HGMD® Custom Genome Browser Track

- Use the Advanced Search to generate your search query.
- Click submit and navigate to results page
- Click download results as genome browser track （Figure 10**Error! Reference source not found.**）
- Save custom genome browser track to your computer （Hint: remember where you save it）

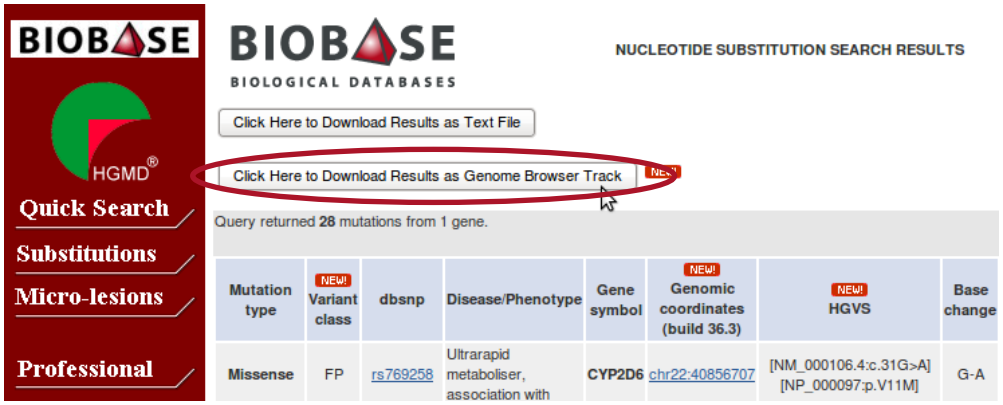

Figure 10. Download custom Genome Browser track

Step 2: Importing your HGMD® Custom Genome Browser Track into an external program （e.g. UCSC genome browser or Genome Trax）

- Navigate to UCSC genome browser gateway （http://genome.ucsc.edu/cgi-bin/hgGateway）
- Click manage/add custom tracks （Figure 11**Error! Reference source not found.**）

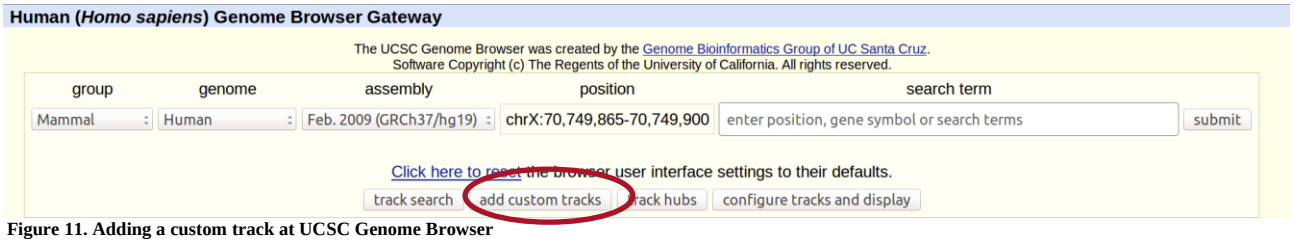

Figure 11. Adding a custom track at UCSC Genome Browser

Step 3：Finding the Custom Genome Browser track to upload.

- Click Browse （Figure 12） and upload previously saved browser track
- Click submit

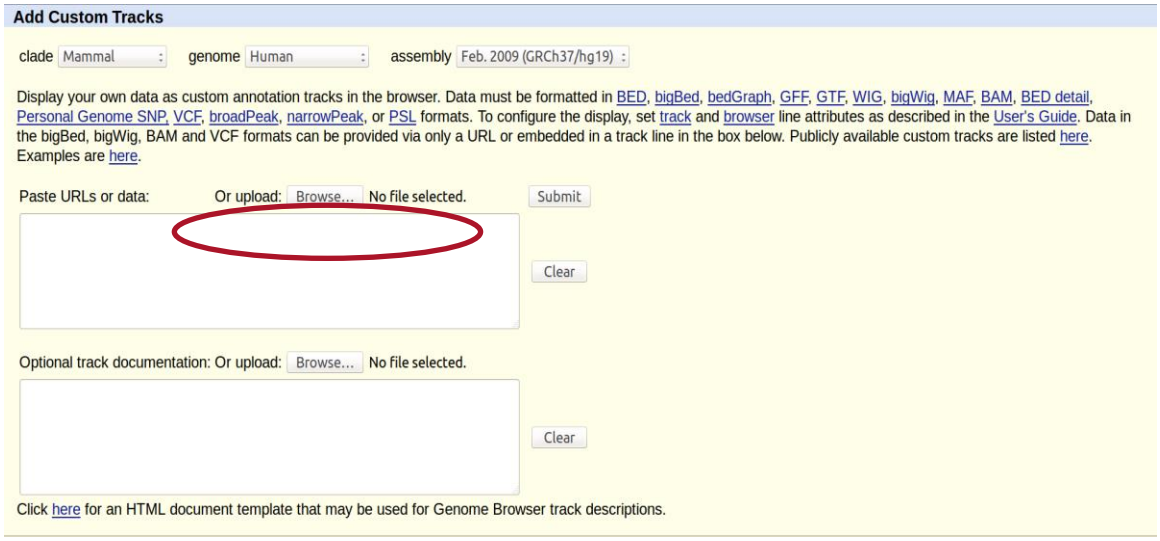

Figure 112. Uploading the custom Genome Browser track

Step 4

- Your track should now be imported.
- Click "go to genome browser" （Figure 13**Error! Reference source not found.**）.

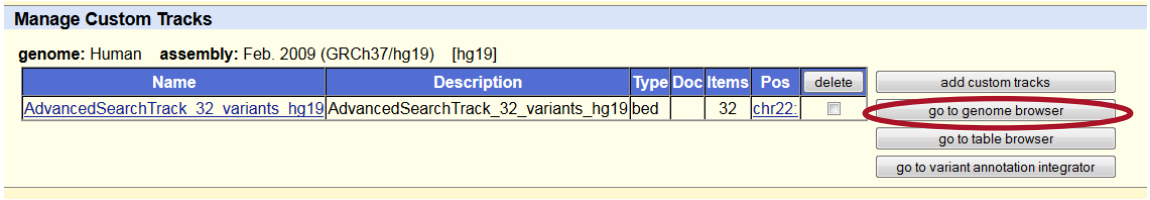

Figure 13. Imported track

Step 5: View your custom track in the genome browser (Figure 14**Error! Reference source not found.**).

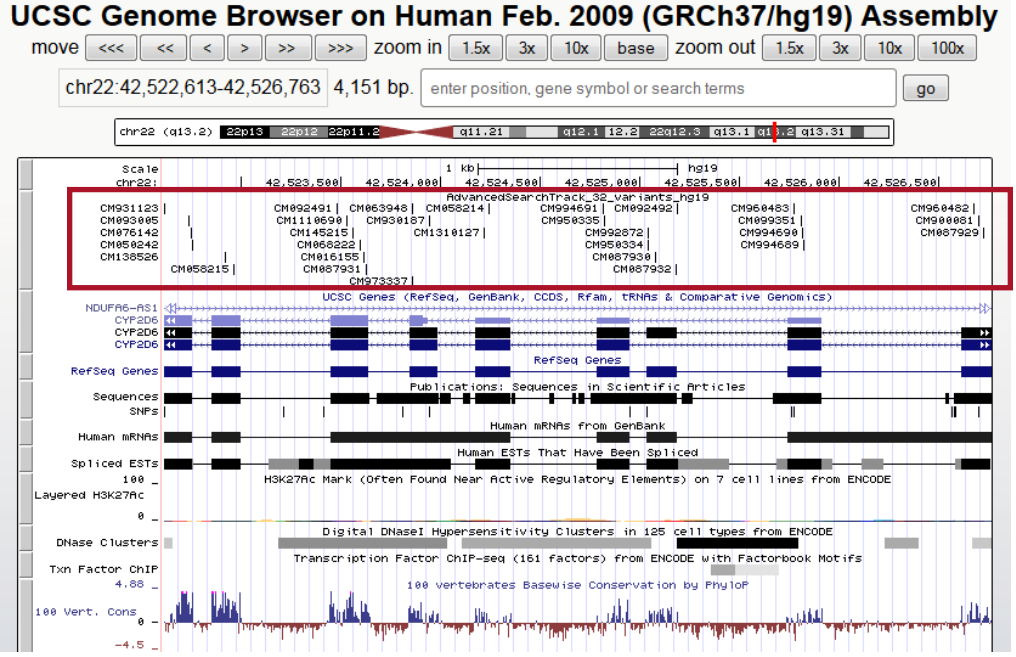

Figure 14. An Advanced Search result been viewed on the UCSC genome browser

Note: Large custom tracks may take a while to load.

HGMD® custom annotation tracks can be viewed only on the machine from which they were uploaded and are automatically deleted if unused for 48 hours. The HGMD® custom annotation tracks are for the sole use of HGMD® Professional subscribers.

Tools used to predict harmful missense mutations (SIFT and MutPred)

HGMD® missense mutations have been annotated with two different tools which predict the pathogenicity of missense mutations. The MutPred tool also makes predictions about the underlying molecular mechanism disrupted, e.g. loss of phosphorylation site.

SIFT (Sorting Intolerant From Tolerant)

SIFT predicts whether an amino acid substitution (AAS) affects protein function based on sequence homology and the physical properties of amino acids (NG and Henikoff 2001). For disease-causing (DM) missense mutations in HGMD® around 80% are predicted to be deleterious by SIFT. An AAS with a SIFT score of less than 0.05 is predicted to be deleterious, one with a score greater than or equal to 0.05 is predicted to be tolerated. For more information please refer to the SIFT website (http://sift.jcvi.org/www/SIFT_help.html).

MutPred

The MutPred Score is the probability (expressed as a figure between 0 and 1) that an AAS is deleterious/disease-associated. A missense mutation with a MutPred score > 0.5 could be considered as 'harmful', whereas a MutPred score > 0.75 should be considered a high confidence 'harmful' prediction.

The MutPred hypothesis refers to the underlying structural and functional properties that the missense mutation impacts upon. The accompanying P-value indicates the assigned probability that the specified structural or functional property has been impacted upon by the mutation (a P value <0.05 indicates a statistically significant probability). Around 20% of missense mutations in HGMD® have been assigned a MutPred hypothesis. For more information please refer to the MutPred website (http://mutpred.mutdb.org/about.html).

cDNA Mutation Display

The cDNA mutation display depicts coding region mutations superimposed on the cDNA sequence of the gene (therefore, splicing and other non-coding mutations are not displayed).

The wild-type cDNA sequence is displayed with the mutation class superimposed over the sequence, with hypertext links to the main HGMD mutation page for each variant (see Fig. 15). Mouse-hovering over the mutation will display the HGVS notation (e.g. c.115C>T p.Q39* below). Other options include the ability to display exon boundaries and optional nucleotide/protein sequence numbering. All features may be toggled on and off using the tick-box options.

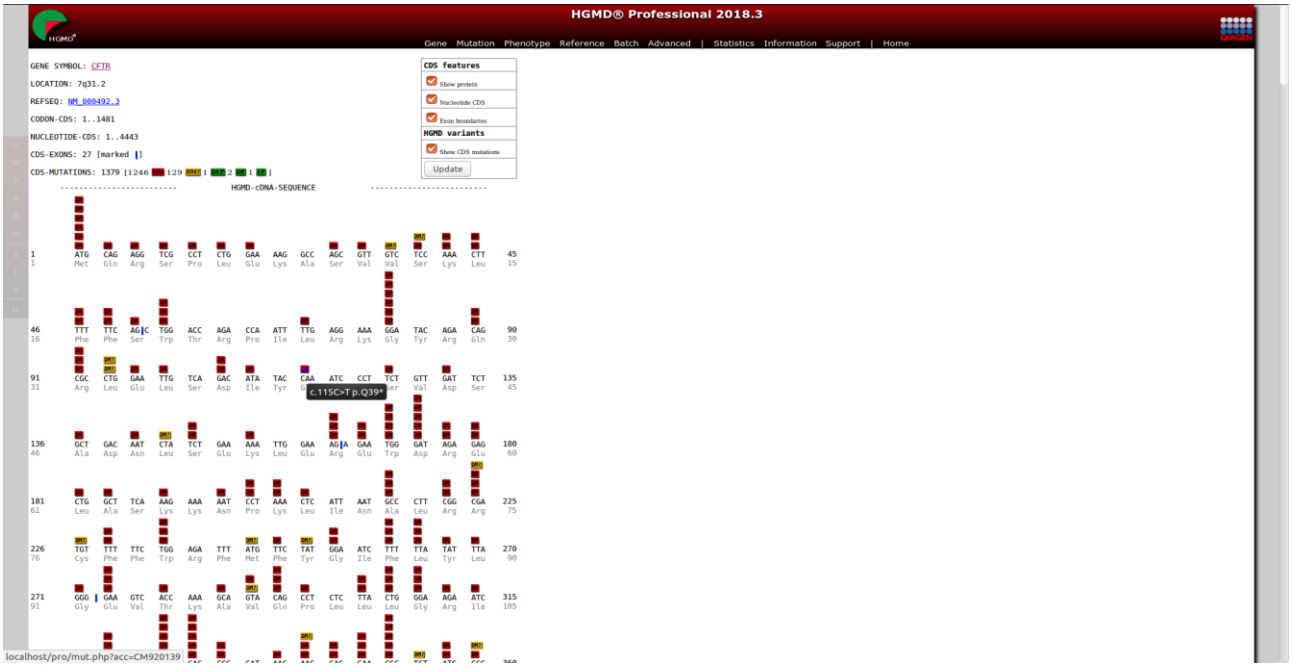

Figure 15. CFTR cDNA mutation display example with all options ticked.

Copyright Notice:

References

Stenson PD et al. (2014), The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133:1-9.

Cooper DN et al. (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.

MacArthur DG et al. (2012), A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823-828.

Dimas AS et al. (2009), Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246-1250.

Stenson PD et al. (2008), The Human Gene Mutation Database (HGMD®): 2008 Update. Genome Med 1(1):13.

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

Balakirev ES and Ayala FJ (2003), Pseudogenes: are they "junk" or functional DNA?. Ann Rev Genet 37: 123-51.

Ng PC and Henikoff S (2001), Predicting deleterious amino acid substitutions. Genome Res. 11:863-74.

den Dunnen JT and Antonarakis SE (2001), Nomenclature for the description of human sequence variations. Hum Genet 109: 121-24.

Cotton RG and Scriver CR (1998), Proof of "disease causing" mutation. Hum Mutat 12:1-3.

Krawczak M and Cooper DN (1995), Core database. Nature 374(6521): 402

Cooper DN and Krawczak M (1993), Human Gene Mutation. BIOS Scientific Publishers, Oxford.