# FAQ 2015

# HGMD® - Frequently Asked Questions

## In HGMD®, there are many common SNPs listed as mutations. Are you going to correct these?

We are in the process of reviewing each of the variants listed in HGMD that have been found to occur at a higher frequency in normal populations than might be expected for a rare disease-causing variant. We are collaborating with the 1000 Genomes Consortium to achieve this. We are currently aware of about 700 variants assigned as Disease-Causing Mutations (DMs) in HGMD® 2012.2 that appear with an allele frequency of greater than 1% in the 1000 Genomes Project Data.
When a variant is observed in a normal population at a higher frequency than expected, it does not necessarily mean that the variant is not a disease-causing mutation. For example, variants may be common but give rise to a (recessive) disease only in those individuals where both alleles are affected e.g. CFTR dF508. Another mechanism might involve a potentially compensating variant (allelic or non-allelic) which could be present in much of the population, but disease will occur in the absence of the compensating variant. Alternatively, some variants may be compensated for by copy number variation. Even rare, disease-causing mutations typically do not exhibit 100% penetrance for the above (and other) reasons, although there are obvious exceptions (e.g. Huntington's disease). It is therefore not unreasonable that we should expect to find some disease-causing variants in healthy individuals. Indeed, we have estimated that human genomes (from normal apparently healthy individuals) typically contain ~100 genuine loss-of-function variants with ~20 genes completely inactivated (MacArthur et al., 2012). To come to a conclusion about the clinical relevance of a given mutation in a particular individual therefore requires the judgment of a medical professional, who can take such factors into account. Thus, mutational variants in HGMD® are likely to fall into a spectrum that ranges from spurious reports (especially in the older literature) where a variant may have been found simply in association with the disease while not being the actual causative variant, to cases where we cannot tell for sure, and cases where the variant, even though common, does indeed contribute to disease causation, and hence is correctly assigned as such.

To resolve this issue is likely to be a slow iterative process, because we have to review all the supporting evidence for each variant. After review, and if so required, we shall change the status of any incorrectly ascribed DM variant to disease-associated polymorphism (DP) or assign a question mark (DM?) or remove the mutation entry entirely.

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.

## How do I find the location of mutations in a specific variant for genes with multiple splicing variants?
## Is the codon numbering system for different mutations in a gene consistent; for example, is the numbering for different mutations all based on one splicing variant? If yes, where could I find the accession number of this mRNA splicing variant?

Codon numbering is consistent with the cDNA sequences provided (along with the NCBI accession numbers). HGMD® mutations are mapped to these sequences.

## Why doesn`t the lower case “acag” representation in TACTAC^414TTAGacagAGAAGCTGGG match the c.1245_1248delCAGA position for the deletion CD982750 in SMAD4?

Deletions and insertions in HGMD® are not necessarily represented at the most 3-prime (downstream) possible location of the sequence.
In the specific example for SMAD4, the mutation could be delACAG or delCAGA. It is not possible to tell which at the molecular sequence level.
HGVS nomenclature requires that the mutation is represented as delCAGA (most 3-prime nucleotides). That is the essential difference.

## How does the HGMD® database represent a simple, single-base insertion?

For all insertions, the start coordinate is one less than the end coordinate. Additionally, for insertions, wildBASE in the "Mutnomen" table is NULL and mutBASE represents the inserted bases located between the start and end coordinates. A risk allele of length one is an insertion of length one.

## Is there a way to get a complete SNP list for a gene, including mutations which are not disease-causing?

HGMD® records disease-causing mutations and disease-associated/functional polymorphisms only. Neutral polymorphism data are available in other databases (for example, dbSNP and HapMap). dbSNP data were integrated into HGMD® for missense/non-sense SNPs only.

## What is SIFT?

SIFT (Sorting Intolerant From Tolerant) is an algorithm which predicts whether an amino acid substitution (AAS) will affect protein function based on sequence homology and the physical properties of amino acids (NG et al, 2001). For disease-causing missense (DM) mutations in HGMD®, around 80% are predicted to be deleterious by SIFT (Mort et al, 2010).
References

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. (2001) *Genome Res* 11:863-874.
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. (2010) In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum Mutat* 31:335-346.

## How do I interpret the SIFT score?

An amino acid substitution (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.

## What is MutPred?

MutPred (Mutation Prediction) is an algorithm which predicts whether an amino acid substitution (AAS) will be disease-associated or neutral (Li et al. 2009). MutPred predicts the molecular cause of disease/deleterious AAS based upon the gain or loss of 14 different structural and functional properties, for example, loss of a phosphorylation site. This is the MutPred hypothesis.

References

Li B, Krishnan VG, Mort M, 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.

## How do I interpret the MutPred score?

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

## Is the first position in a chromosome counted as "0" or "1" for the purpose of determining genomic coordinates?

The first position in the chromosome is "1".

## Is the "risk allele" for a given site-disease association available in the HGMD® database?

In HGMD® you will find the wild-type allele and - of course - the mutated allele. This information is presented in the reports, where the base change (G-C) gives you the wild-type allele (G) first and the mutated allele (C) second (e.g. in the tables or the sequence). If you want to know if the actual chromosome combination behind this mutation is homozygous or heterozygous to understand whether the risk allele is CC or GG or one of them in combination with CG, then you will NOT find this info in the HGMD® database. You'll have to go and consult the cited article. Download users can find the allele associated with the HGMD® phenotype in the "Mutnomen" table (mutBASE column). The wildBASE column is the wild-type nucleotide sequence (NULL for insertions) and the mutBASE column is the mutated nucleotide sequence (NULL for deletions). When looking at the core data tables (e.g. mutation, prom), the phenotype allele should be the variant allele.

## Which strand are the sequences in the database taken from?

All sequence data (wild-type, mutated, and flanking bases) are given as they would appear on the strand which encodes the protein in question (i.e. the "coding" strand). For example, in the case of CM014827, the "Mutnomen" table lists wildBASE="T" and mutBASE="C" to indicate the T>C polymorphic change described in the referenced article. The strand encoding the STX1A gene is the minus strand on the assembly, so this substitution would be equivalent to an A>G change in the assembly sequence.

The wildBASE and mutBASE sequence are obtained from the core mutation tables while the flanking sequence is derived from the assembly. The flanking sequence is converted into the complementary sequence if the mutation was described on the non-coding strand in the reference, so as to correspond to the coding strand. Therefore all data in HGMD® are "coding strand" data.

## How many mutations in splice sites can I find in HGMD®?

15168 mutations with consequences for mRNA splicing are currently available in HGMD® release 2015.1.
To find updated statistics, please use the statistics page in HGMD® Professional.

## How do I find the location of mutations in a specific variant for genes with multiple splicing variants?

## Is the codon numbering system for different mutations in a gene consistent; for example, is the numbering for different mutations all based on one splicing variant? If yes, where could I find the accession number of this mRNA splicing variant?

Codon numbering is consistent with the cDNA sequences provided (along with the NCBI accession numbers). HGMD® mutations are mapped to these sequences.

## Why doesn`t the lower case “acag” representation in TACTAC^414TTAGacagAGAAGCTGGG match the c.1245_1248delCAGA position for the deletion CD982750 in SMAD4?

Deletions and insertions in HGMD are not necessarily represented at the most 3-prime (downstream) possible location of the sequence.
In the specific example for SMAD4, the mutation could be delACAG or delCAGA. It is not possible to tell which at the molecular sequence level.
HGVS nomenclature requires that the mutation is represented as delCAGA (most 3-prime nucleotides). That is the essential difference.

## How does the HGMD® database represent a simple, single-base insertion?

For all insertions, the start coordinate is one less than the end coordinate. Additionally, for insertions, wildBASE in the "Mutnomen" table is NULL and mutBASE represents the inserted bases located between the start and end coordinates. A risk allele of length one is an insertion of length one.

## Is there a way to get a complete SNP list for a gene, including mutations which are not disease-causing?

HGMD® records disease-causing mutations and disease-associated/functional polymorphisms only. Neutral polymorphism data are available in other databases (for example, dbSNP and HapMap). dbSNP data were integrated into HGMD® for missense/non-sense SNPs only.

## I have tried to open "Get map" under the Mutation viewer for ACTA2, but the window could not be opened and I had to re-login. What should I do?

Try installing "Java Runtime Environment version 6" locally on your computer.
One other known “Security” problem with the Mutation Viewer could be an issue of the newest Java version which introduced Internet Explorer like security settings. The settings can be changed in Windows 7 by opening “Programs -> All programs -> Java -> Configure Java”.
In the Security tab you need to add the site “https://portal.biobase-international.com” to the site list using the “Edit Site List” button.

## Does HGMD® provide information about genomic polymorphisms in genes? How does the HGMD® define a genomic alteration - as a mutation or as a polymorphism?

Only disease-associated/functional polymorphisms are included in HGMD®. 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 case no clinical phenotype is known to be associated with a polymorphic variant, but sufficient *in vitro* or *in vivo* expression/functional data have nevertheless been presented to indicate functional significance, then the variant will be included in HGMD®.
NCBI dbSNP numbers (where identified) are also included in the comment field. A polymorphism is a mutation found at a frequency of >1% in any population.

## Two new germinal mutations recently identified in our lab, are not reported in HGMD® database. How can they be added?

HGMD® records disease-causing mutations published in the literature. At the moment, the only way to get these mutations into HGMD® would be to publish them.

## How frequently is the HGMD® database updated?

HGMD® has quarterly releases.

## Is the reference identified by the PMID in a mutation record one of the studies that established the genetic association with the disease/phenotype? If multiple studies examined the variation, which study is given in the PMID field (for example, the first published, the most authoritative, etc.) record. Will all papers specified for a record be in agreement about the risk allele? In what database table/columns are these additional articles in stored?

The Allmut mutation entry should contain the first literature report. HGMD® also may provide where available additional references for a given entry in case additional information is reported in the reference.

Additional references are stored in the "Extrarefs" table. The "risk allele" will not always be the same between different literature reports (which will report different phenotypes and functional studies).

## Do you include variants from genome-wide association studies (GWAS) papers in HGMD®?

HGMD® does include SNPs from GWAS studies whenever there is evidence for a likely effect on function (which is in fact lacking for most GWAS studies). HGMD® 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). To be included, there must be a convincing association of the polymorphism with the phenotype. These polymorphisms are currently identified in the database by an addition to the phenotypic description. These additions are limited to association, association with and increased or lower risk, depending on how the polymorphism was reported.

## Which genome build is currently used for HGMD®?

Genome build 37 (hg19) is used for HGMD® since release 2012.1. Users still requiring coordinates based on hg18 should use LiftOver based at UCSC.

## How many mutation entries in HGMD® have dbSNP identifiers?

47125 mutation entries (about 25%) in HGMD® Professional 2015.1 have dbSNP identifiers. HGMD® entries that have been mapped to a corresponding entry in dbSNP display the FREQ symbol where the associated dbSNP entry contains population frequency data. 12518 mutations in HGMD® 2015.1 have FREQ=frequency
information. To find updated statistics, please use the predefined dbSNP identifier search available on the mutation search page.

## How many disease terms and phenotypes are reported in HGMD®?

HGMD® Professional contains currently about 14314 diseases/phenotypes (conditions).
The disease descriptions can be different variants of one disease as taken from the literature, and is therefore sometimes redundant.
The hgmd_phenbase provides disease terms from MeSH, ICD-10 and other sources mapped to HGMD® disease phenotypes to provide standard disease terms and unique identifiers. 13505 ICD10 phenotypes have been mapped to the 2012.2 HGMD® version and 2086 unique MeSH terms are annotated against HGMD® phenotypes.
Please use the statistics page for updated information.

## In the quick search part of the HGMD® advanced search, what does the ranking score relate to?

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).

## Can I find mutations associated with two (or more) phenotypes in HGMD®?

Yes, if a paper reports a variant to cause two genuinely different phenotypes (a very rare occurrence) this should be reflected in the disease field (e.g. CM013504 Stargardt disease and macular degeneration).
In case two (or more) papers describe the same lesion as responsible for different phenotypes we will record the earliest report as the "primary" mutation reference and add the subsequent reports as "Additional phenotype" secondary references. In each case, the phenotype is associated with the reference in which it is described. The details of the additional phenotypes (and other information from the secondary references) being provided in the expanded mutation record (that reached by clicking on the accession number button).
Currently (2014.2) 11552 mutations have additional disease/phenotype information from secondary references.

## How to perform a batch search in HGMD® Professional?

HGMD® online provides two options for batch searches. The first option at the main search “Professional” is designed to accept a list of up to 500 variant or gene identifiers. Identifiers accepted by the batch search include dbSNP, chromosomal coordinate and HGMD® accession for variants and HUGO Nomenclature Committee gene symbols and IDs, Entrez Gene IDs and OMIM IDs for genes, and Variant Call Format (VCF).
At the Advanced search tool we provide “MART” has an option for batch queries which can be saved as text files. A maximum of 50 identifiers of dbSNP, EntrezGene or PubMed IDs can be loaded and searched for.

## Why do numbers for sequences in HGMD® Professional differ from numbering in (primary) references?

Mutations presented in HGMD® may utilise an alternate (up-to-date) transcript numbering compared to that used in the original report (which may have been published many years ago). Please don`t expect a 13 year old manuscripts to continue to match modern transcript sequences in all cases.

## Why are there differences in the nucleotides and coordinates for deletions between the mutation description, the HGVS description, and the VCF file?

Consider for example the mutation CD961751 in ABCD1.

The deletion description:
GCTGCAG^TGGctcctcatcgccCTCCCTGCTA

The genomic coordinates and sequence (at the time of this writing, based on GRCh37.3)
ChrX:152991137-152991148

CTTTTGGCTGGCAGCTGCTGCAGTGGCTCC(TCATCGCCCTCC/-)CTGCTACCTTCGTCAACAGTGCCATCCGTT

The HGVS description
NM_000033.3: c.416_427delTCATCGCCCTCC

The VCF description
X 152991132 CD961751 GCTCCTCATCGCC G

If you look at them lined up against each other here is how they align:

CTTTTGGCTGGCAGCTGCTGCAGTGGCTCCTCATCGCCCTCCCTGCTACCTTCGTCAACAGTGCCATCCGTT Deletion

CTTTTGGCTGGCAGCTGCTGCAGTGGCTCCTCATCGCCCTCCCTGCTACCTTCGTCAACAGTGCCATCCGTT Genomic

CTTTTGGCTGGCAGCTGCTGCAGTGGCTCCTCATCGCCCTCCCTGCTACCTTCGTCAACAGTGCCATCCGTT HGVS

CTTTTGGCTGGCAGCTGCTGCAGTGGCTCCTCATCGCCCTCCCTGCTACCTTCGTCAACAGTGCCATCCGTT VCF

There are two possible positions for the deletion because of the flanking CTCC repeat (highlighted in red) that is part of the deletion. While the deleted sequences and their coordinates appear to look different, the remaining sequence after taking out either one looks exactly the same, namely

CTTTTGGCTGGCAGCTGCTGCAGTGGCTCCCTGCTACCTTCGTCAACAGTGCCATCCGTT.

Since the remaining sequence is the only thing that can be sequenced, one cannot know which of the sequences was actually deleted. However, VCF and HGVS do have conventions on which sequence to pick: VCF requires insertions/deletions to be placed as far 5-prime (upstream) as possible, whereas HGVS requires them to be as far 3-prime (downstream) as possible. Because of this difference in policy, we have to pick different sequences and their positions in these different formats.

We provide the original deletion as it has been described in the paper. We provide HGVS descriptions and genomic coordinates in the records of HGMD® following the HGVS conventions, i.e. the most 3-prime variant. We provide sequences and coordinates in VCF following the VCF convention, i.e. the most 5-prime variant.

## How do you decide which reference cDNA has to be linked?

As a rule, if there is more than one refseq for the gene in question, the curation team would try match the one used in paper that fits the data, or use the longest available transcript or the transcript used by the refseqgene project if the refseq is not specified or not clear.

## How should I reference HGMD® Professional in a scientific article?

The current version of HGMD® Professional is described at:
The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.
Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN
Hum. Genet. (2014) v133, p1-9