Human Genetics (2020) 139:1197–1207
https://doi.org/10.1007/s00439-020-02199-3

REVIEW

The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting

Peter D. Stenson[1] · Matthew Mort[1] · Edward V. Ball[1] · Molly Chapman[1] · Katy Evans[1] · Luisa Azevedo[1,2] · Matthew Hayden[1] · Sally Heywood[1] · David S. Millar[1] · Andrew D. Phillips[1] · David N. Cooper[1]

Received: 18 May 2020 / Accepted: 19 June 2020 / Published online: 28 June 2020

Abstract
The Human Gene Mutation Database (HGMD®) constitutes a comprehensive collection of published germline mutations in nuclear genes that are thought to underlie, or are closely associated with human inherited disease. At the time of writing (June 2020), the database contains in excess of 289,000 different gene lesions identified in over 11,100 genes manually curated from 72,987 articles published in over 3100 peer-reviewed journals. There are primarily two main groups of users who utilise HGMD on a regular basis; research scientists and clinical diagnosticians. This review aims to highlight how to make the most out of HGMD data in each setting.

Introduction

The Human Gene Mutation Database (HGMD®) available via http://www.hgmd.org represents an attempt to systematically collate all known gene lesions underlying human inherited disease that have been published in the peer-reviewed literature. Mutation data catalogued by HGMD (summarized by mutation type) are listed in Table 1.

HGMD was originally established in 1996 with the goal of facilitating the scientific study of mutational mechanisms in human genes underlying inherited disease (Cooper et al. 2010; Stenson et al. 2017). However, over the last 20 years, it has acquired a much broader utility as it has become the central unified repository for disease-related genetic variation in the germ-line.

Brief history of the resource

The first Public version of HGMD containing ~ 10,000 variants in around 600 genes was made freely available from Cardiff via http://www.hgmd.org in April 1996. From that point, the database expanded swiftly to become the de facto central database for mutations causing human inherited disease. HGMD has been supported over the years by commercial partnerships with various industry leading biomedical research companies. Through a partnership with Celera Genomics from 2000 to 2005, HGMD data were made available as part of the Celera Discovery System. The years 2006–2015 saw the creation and continued development of HGMD Professional, a stand-alone web application, made available under license from BIOBASE GmbH. In 2015, QIAGEN Bioinformatics acquired BIOBASE, and our commercial partnership continued with HGMD data being made available via HGMD Professional (including data download) plus integration into Ingenuity Variant Analysis and Qiagen Clinical Insight. The latest version of HGMD (2020.2) contains 289,346 different mutations in 11,076 genes (Fig. 1).

Sources of mutation data

HGMD screens the peer-reviewed biomedical literature on an ongoing basis, and currently contains data derived from over 72,000 manuscripts published in more than 3100 different journals. Relevant articles are identified via manual inspection of a core selection of journals, supplemented by the use of online computerised procedures utilising Google Scholar, publishers' websites and PubMed, to survey the wider literature. Articles identified as potential sources of mutation data are assessed by a team of experienced

✉ Peter D. Stenson
stensonPD@cardiff.ac.uk

[1] Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

[2] i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4200-135 Porto, Portugal

Table 1 Numbers and types of different variants and genes present in HGMD Professional release 2020.2 and the publicly available version of the database (as of June 7th 2020)

Mutation type	Number of Mutations in HGMD Professional 2020.2 (disease-associated/functional polymorphism sub-total)	Number of Mutations (publicly available via http://www.hgmd.org)
Missense substitutions	136,383 (6435)	85,225
Nonsense substitutions	31,407 (392)	20,779
Splicing substitutions (intronic and exonic)	24,976 (735)	17,183
Regulatory (5′ and 3′ and intergenic)	4723 (3006)	3544
Small deletions (≤ 20 bp)	41,749 (369)	28,155
Small insertions/duplications (≤ 20 bp)	17,760 (212)	11,745
Small indels (≤ 20 bp)	3813 (70)	2679
Gross deletions (> 20 bp)	20,448 (170)	14,186
Gross insertions/duplications (> 20 bp)	5219 (98)	3445
Complex rearrangements	2299 (138)	1747
Repeat variations	569 (331)	498
All HGMD data	289,346 (11,954)	189,186[a]
HGVS nomenclature provided[b]	263,452 (10,923)	0
Genomic coordinates/Variant Call Format (VCF) provided[c]	263,160 (10,845)	168,473[d]
Genes (subdivided by variant class)	**Number of Genes in HGMD Professional 2020.2**	**Number of Genes (publicly available via http://www.hgmd.org)**
Number of genes (with DM and/or DM? entries only)	7141	4218
Number of genes (with either DP, FP or DFP only)	1198	904
Number of genes (with a mixture of DM and/or DM? plus DP, FP and/or DFP)	2737	2522
Number of disease genes (containing at least one DM or DM? entry)	9878	6740
Total number of genes in HGMD[e]	11,076	7644

DM disease-causing mutation, *DM?* Likely disease-causing, but with questionable pathogenicity, *DP* disease-associated polymorphism, *DFP* disease-associated polymorphism with supporting functional evidence, *FP* in vitro/laboratory or in vivo functional polymorphism

[a]Mutations available via the HGMD Public Website (http://www.hgmd.org)

[b]As described in den Dunnen et al. (2016)

[c]As described by Danecek et al. (2011)

[d]The Ensembl HGMD_PUBLIC release (https://www.ensembl.org/) contains hg19/hg38 genomic coordinates and HGMD accession numbers only

[e]Total excludes mitochondrial genes (searchable but no variant data) and retired records

curators (with an average of more than 12 years experience in curation). Discrepancies in variant reporting that require additional scrutiny are identified in approximately 20% of articles. Some 25% of these can be resolved by utilising other information reported in the manuscript or by referring to supplementary material (chromosomal coordinate, sequence chromatogram etc.). However, approximately 75% of these ambiguities necessitate direct contact with the authors. Author responses that are sufficient to allow us to include the mutation data in question are received for ~ 55% of queries; however, the reported variants from the other 45% (comprising ~ 7% of all papers screened) remain unresolved.

One other challenge we have encountered is that an increasing number of journals do not appear to be systematically indexed by Medline, at least not immediately upon publication (i.e. the NLM catalogue states that the journal is not currently indexed for MEDLINE, although individually submitted abstracts may still be present in PubMed); with 729 mutation entries, the journal *Front Genet* is the most highly represented of these, followed by *Neurol Genet* (488 entries) and *Mol Syndromol* (333 entries). There are a total of 172 journals listed in HGMD that the NLM catalogue lists as not currently indexed by Medline/PubMed. This number represents approximately 5% of all the journals currently cited by HGMD. A summary of the top 20 journals cited

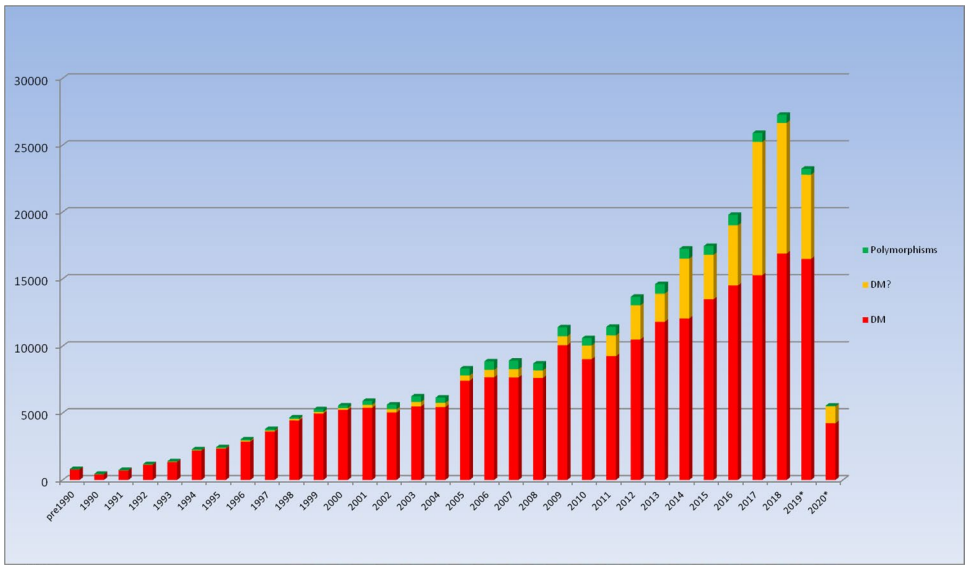

Fig. 1 Mutation totals by year of publication subdivided by variant class. *Figures for 2019 and 2020 not yet complete. *DM* disease-causing mutation, *DM?* Likely disease-causing, but with questionable pathogenicity

in HGMD (by number of mutation entries listed) is shown in Fig. 2.

Utilisation

There are many different ways in which HGMD data may be utilised in both a research and a clinical setting, all of which are dependent upon on the version of HGMD available to the user. For checking known genotype–phenotype relationships (i.e. relatively small numbers of variants found in specific genes involved in a known disease), the Professional (https://www.qiagenbioinformatics.com/products/human-gene-mutation-database/) online-only interface will most likely suffice. Users may search using the gene symbol, disease name utilising the Universal Medical Language System or UMLS (https://www.nlm.nih.gov/research/umls/), literature reference, HGVS description or genomic coordinate. More recently, nucleotide-level annotations for multiple non-canonical mRNA transcripts have been added to HGMD Professional, in recognition of the fact that clinically relevant variants are more likely to impact those exons that are present in multiple transcripts (Subramanian 2018).

Academic or non-profit users without a subscription may utilise the public version of HGMD (http://www.hgmd.org). However, this version is provided in a basic form that is searchable only by gene symbol or disease name, is only updated twice annually, is maintained permanently at least 3 years out of date, and does not contain any of the additional annotations or extra features present in HGMD Professional (e.g. dbSNP ID, chromosomal coordinates, HGVS nomenclature, Variant Call Format (VCF), population frequency data, additional literature reports, advanced search features, evolutionary conservation data and functional predictions).

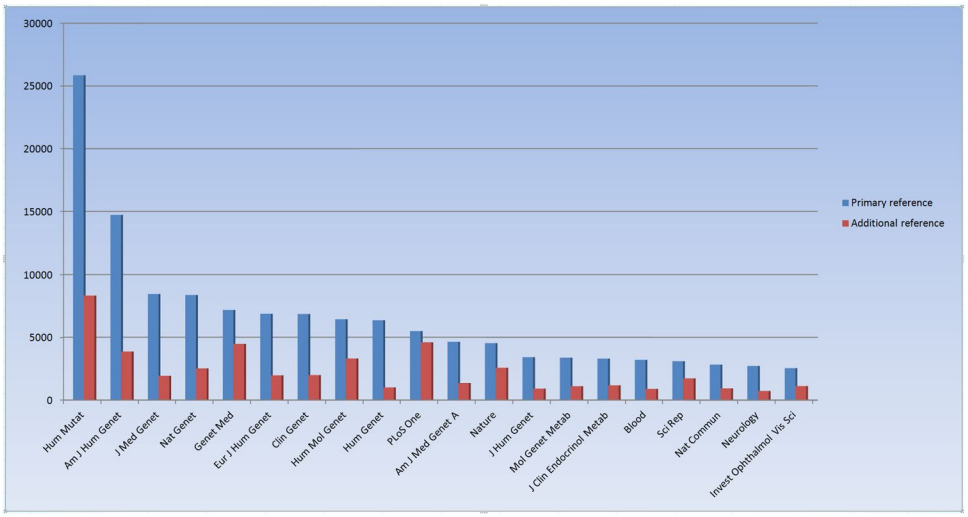

Fig. 2 Top 20 journals by number of mutation entries (HGMD Professional release 2020.2 June 7th 2020) in relation to both primary and additional (secondary) references

Increasingly in clinical practice (as has happened in research), the use of next-generation sequencing (NGS) technology has greatly expanded in recent years. HGMD has adapted to these changes by providing mutation data in standardized formats for ease of computational analysis. The most useful of these in a high-throughput context is Variant Call Format or VCF (Danecek et al. 2011), which mimics the format of the data that will typically be produced by a bioinformatician after processing the output from NGS. This format is available as part of the licensed HGMD Professional download version. There is also a publicly available dataset containing coordinates (but not precise nucleotide changes) released via Ensembl (https://www.ensembl.org). Users looking for a potential clinical diagnosis may wish to utilise the HGMD online batch search mode (utilising dbSNP identifier, hg19/hg38 coordinate, hg19/hg38 VCF, HGMD accession number, PubMed ID, HUGO gene symbol, HUGO gene ID, Entrez gene ID or OMIM ID). Results are limited to the first 500 found; however, these may be prioritised based on likely causation or population frequency based on gnomAD (http://gnomad.broadinstitute.org/) data. Results may also be restricted to a particular UMLS disease concept (see Fig. 3). In this way, Next Generation Sequencing results may be directly compared to HGMD data, and any relevant variants that have been previously implicated in disease causation are returned. Filtering a typical exome will generally yield a list of approximately 400 damaging variants, comprising up to 8 highly damaging DMs (Xue et al. 2012) and up to 15 potential risk alleles (Tabor et al. 2014). To make sense of these data, the user must prioritize their results in line with their own particular needs and priorities (Table 2).

An exome screen will most likely return a combination of the different HGMD variant classes listed in Table 2. Results may be further prioritized within a variant class by (i) utilizing the population frequency data from dbNSFP3, a database of functional predictions and annotations based on genomic location (Liu et al. 2016), present in the HGMD download and (ii) making use of the HGMD computed ranking score. This ranking score is a single relative probability score between 0 and 1, with 1 being most likely disease-causing; it is currently available only for coding region nucleotide substitutions (both missense and nonsense). The score is computed by HGMD using a supervised machine learning approach known as Random Forest (Breiman 2001), and is based upon multiple lines of evidence, including HGMD literature support for pathogenicity (placed on a scale of 1–10, with 1 being the lowest score and 10 being the highest), evolutionary conservation (100-way vertebrate alignment), variant allele frequency and in silico pathogenicity prediction including CADD (Rentzsch et al. 2019), PolyPhen2 (Adzhubei et al. 2010) and MutPred (Li et al. 2009). HGMD data are used to train the model with disease-causing mutations (DM) forming the positive class and possible/probable disease-causing mutations (DM?) making up the negative class. Individually, ranking scores may be interpreted as relative probabilities of pathogenicity (i.e. the higher the score, the more likely the variant is to be disease-causing).

Search terms	Gene symbol	HGMD mutation	HGVS	hg38 coordinate	Variant class	dbSNP identifier	gnomAD	HGMD accession	Clinvar
6 51752011 CM112375 G T	PKHD1 AR	Tyr2343Term	NM_138694.3:c.7029C>A	chr6:51887213	DM	rs202223718		CM112375	
1 215821963 CM080617 G C	USH2A AD/AR	Ser4830Term	NM_206933.2:c.14489C>G	chr1:215648621	DM	rs184351619	0.000258264	CM080617	
18 44109190 CM1211193 G A	LOXHD1 AR	Arg1494Term	NM_144612.6:c.4480C>T	chr18:46529227	DM	rs201587138	0.000905387	CM1211193	178396 Pathogenic/Likely_pathogenic
2 26418053 CM940884 C G	HADHA AR	Glu510Gln	NM_000182.4:c.1528G>C	chr2:26195184	DM	rs137852769	0.00181019	CM940884	100085 Pathogenic
1 94473287 CM970017 G A	ABCA4 AD/AR	Leu1970Phe	NM_000350.2:c.5908C>T	chr1:94007731	DM	rs28938473	0.00303755	CM970017	7892 Conflicting_interpretations_of_pathogenicity
14 23890217 CM087590 C A	MYH7 AD	Asp1096Tyr	NM_000257.3:c.3286G>T	chr14:23421008	DM?	rs45478699	0.0000645703	CM087590	42953 Uncertain_significance
16 89842176 CM112454 C G	FANCA AR	Cys625Ser	NM_000135.3:c.1874G>C	chr16:89775768	DM?	rs139235751	0.00264755	CM112454	265136 Conflicting_interpretations_of_pathogenicity
1 31211894 CM166387 G A	LAPTM5	Pro135Ser	NM_006762.2:c.403C>T	chr1:30739047	DM?	rs34101571	0.00862347	CM166387	
17 46022065 CM144885 G A	PNPO AR	Arg116Gln	NM_018129.3:c.347G>A	chr17:47944699	DM?	rs17679445	0.0529054 HI	CM144885	129981 Benign/Likely_benign
7 117170947 CS930762 T C	CFTR AR	IVS3 as T-C -6	NM_000492.3:c.274-6T>C	chr7:117530893	DM?	rs371315549	0.000419788	CS930762	237855 Conflicting_interpretations_of_pathogenicity
19 49206674 CM950483 G A	FUT2	Trp154Term	NM_000511.5:c.461G>A	chr19:48703417	DFP	rs601338	0.422822	CM950483	12945 Benign,_association,_protective
1 169519049 CM940389 T C	F5	Arg534Gln	NM_000130.4:c.1601G>A	chr1:169549811	DFP	rs6025	0.016952	CM940389	642 Pathogenic,_risk_factor
1 94476467 CM015091 T A	ABCA4	Asn1868Ile	NM_000350.2:c.5603A>T	chr1:94010911	DFP	rs1801466	0.0388723	CM015091	99390 Conflicting_interpretations_of_pathogenicity
1 98348885 CM970421 G A	DPYD	Cys29Arg	NM_000110.3:c.85T>C	chr1:97883329	DFP	rs1801265	0.280106	CM970421	435 Pathogenic
2 38298203 CM004465 C G	CYP1B1	Leu432Val	NM_000104.3:c.1294C>G	chr2:38071060	DFP	rs1056836	0.497509	CM004465	456637 Benign
4 100268190 CM050165 A C	ADH1C	Gly78Term	NM_000669.4:c.232G>T	chr4:99347033	DP	rs283413	0.006395	CM050165	18181 risk_factor
16 89261482 CM067985 C A	CDH15	Tyr788Term	NM_004933.2:c.2364C>A	chr16:89195074	DP	rs2270416	0.0583398	CM067985	128650 Likely_benign
14 24470138 CM1610830 C T	DHRS4L2	Arg159Term	NM_198083.3:c.475C>T	chr14:24000929	DP	rs1811890	0.0667897	CM1610830	
1 152323132 CM1311867 G T	FLG2	Ser2377Term	NM_001014342.2:c.7130C>A	chr1:152350656	DP	rs12568784	0.19868	CM1311867	
11 48286231 CM074974 T A	OR4X1	Tyr273Term	NM_001004726.1:c.819T>A	chr11:48264679	DP	rs10838851	0.700679	CM074974	
6 132859609 CM034133 T A	TAAR9	Lys61Term	NM_175057.3:c.181A>T	chr6:132538470	FP	rs2842899	0.236676	CM034133	
19 58003580 CS115477 A G	ZNF419	IVS4 ds G-A +1	NM_024691.3:c.298+1G>A	chr19:57492212	FP	rs2074071	0.324509	CS115477	
19 41350664 CM057913 A T	CYP2A6	Tyr392Phe	NM_000762.5:c.1175A>T	chr19:40844759	FP	rs1809810	0.01161	CM057913	
19 15789140 CM043943 A G	CYP4F12	Val90Ile	NM_023944.3:c.268G>A	chr19:15678330	FP	rs609290	0.093566	CM043943	
22 42522613 CM931123 G C	CYP2D6	Ser486Thr	NM_000106.5:c.1457G>C	chr22:42126611	FP	rs1135840	0.58043	CM931123	242701

Fig. 3 Example, online batch result set from HGMD Professional 2020.2

Table 2 HGMD variant classes

HGMD variant class	Relevance	Clinical diagnostic setting	NGS research setting
DM—Disease-causing mutation	Literature indicates causal (or likely causal) link with disease	Most important. These data should be prioritized	Depending on the user's remit, these should be looked at first
DM?—Likely disease-causing, but with additional uncertainty	As for DM, but the authors, curators or other literature evidence indicate that further caution is warranted	If no DM variants are found, these should be looked for next, or they should be ranked lower priority if there are DM results	These data may also be of interest, depending upon requirements (e.g. gene ontology or disease concept stratification)
DP—Disease-associated polymorphism	Significant statistical association with a clinical phenotype. Likely to be functionally relevant	Likely to be irrelevant in a clinical diagnostic setting	These should be included if personal disease risk is being assessed
DFP—DP with supporting functional evidence	As for DP, but definitive functional evidence exists (e.g. via an in vitro luciferase assay)	Potentially important in terms of calculating disease risk (e.g. venous thrombosis risk and Factor 5 Leiden). Other relevant disease risk or drug response variants are also present in this class	If the aim were to look at personal disease risk, or for disease modifiers, then these should be included
FP—functional polymorphism with no reported disease association	Functional effect has been demonstrated, but no disease association has been reported as yet	Likely to be irrelevant in a clinical diagnostic setting, although drug response variants may be present	Interesting from a research perspective as potential risk modifier variants
R—retired from HGMD	Record has been retired and is no longer considered to be phenotypically relevant	Potentially relevant for the purpose of variant exclusion	Potentially relevant if the researcher is interested in variant re-annotation etc

Ranking scores may also be utilised in aggregate to prioritize and rank multiple HGMD variants that have been found in the same sample. A representative example result set (utilizing VCF search terms) from the HGMD Professional batch search showing the first five results from each mutation class present in a normal exome from an apparently healthy individual and sub-ranked by the HGMD ranking score, is provided in Fig. 3.

The top five results shown in Fig. 3 are all DM entries present in HGMD. The five DMs on the list are known to cause autosomal recessive disorders (*USH2A*, *PKHD1, LOXHD1*, *HADHA* and *ABCA4*), and the healthy individual concerned is, therefore, an asymptomatic carrier of these variants. However, *USH2A* and *ABCA4* have also been implicated in late-onset dominant phenotypes, and so may be of long-term clinical interest. The next five results in our illustrative list are from the DM? class and should therefore be treated with an additional degree of caution. However, upon closer inspection, some of these entries could prove to be of clinical relevance in the longer term. The *MYH7* variant is an autosomal dominantly inherited potential cardiomyopathy risk factor, whereas the *PNPO* variant may also be clinically relevant in relation to the phenotype of Pyridoxamine 5′-phosphate oxidase deficiency (although this phenotype can be highly variable). The remaining DM? entries listed in Fig. 3 display either recessive or more complex inheritance, or else give rise to biochemical phenotypes that are of relatively minor clinical concern, which is often the case for this class of variant.

The next three classes of variant (DFP, DP, FP) all occur at polymorphic frequencies, and therefore may confer increased disease risk, and/or may give rise to an alteration in the function of the gene/gene products involved. They are not, however, generally expected to be of immediate clinical concern. That said, HGMD would nevertheless recommend that DFPs, in particular the alleles with a low population frequency for the minor allele (Kido et al. 2018), should be treated as "honorary DMs" for the purposes of returning results, particularly in light of the problems that are often encountered when attempting to classify low-penetrance/hypomorphic alleles, or those with combinatory effects (Wang and Chiang 2019). Notable examples of clinically significant DFPs present in HGMD include the *DPYD* allele p.Cys29Arg (global MAF 0.28) which may be relevant for 5-fluorouracil toxicity and *F5* p.Arg534Gln (F5 Leiden – global MAF 0.02), both of which are listed in Fig. 3. Other selected examples of clinically important DFPs include *PROS1* p.Ser501Pro (Protein S Heerlen polymorphism – global MAF 0.002), *CD36* p.Tyr325* (global MAF 0.03), *AMPD1* p.Gln45* (global MAF 0.10), *CES1* p.Gly143Glu (global MAF 0.01), *FCN3* c.349delC (global MAF 0.02), *ABCA4* p.Asn1868Ile (global MAF 0.04). Some of the

returned alleles in the DFP, DP and FP classes may also be relevant to drug metabolism or as potential modifiers of the clinical phenotype [for example the *SCN5A* p.His558Arg (global MAF 0.25) or *CYP2C19* c.681G > A *2 allele (global MAF 0.17)].

Population frequency data are often employed to screen out potentially benign alleles (Whiffin et al. 2017). Indeed, the HGMD curators have periodically utilised this method to re-annotate or remove questionable variants from HGMD. This practice should, however, in the opinion of the HGMD curators, be utilised with great caution, as it may reduce or even prevent the return of potentially clinically relevant alleles for certain later-onset diseases (Zernant et al. 2017; Wang and Chiang 2019). Despite these concerns, filtering by population frequency remains the best first method to de-prioritize low risk alleles from result sets. HGMD has therefore included population frequency data from 1000 Genomes (1000 Genomes Project Consortium et al. 2015), ExAC (Lek et al. 2016) and gnomAD (http://gnomad.broadinstitute.org/) to facilitate this process. FINDbase may also be used for specific populations (Kounelis et al. 2020). As a precaution, and owing to the fact that some disease-causing alleles occur at relatively high frequencies in certain populations, users may wish to consider "positive filtering", and add HGMD-flagged alleles back into their result set if it is felt that they have been inadvertently excluded [e.g. *HFE* p.Cys282Tyr (MAF 0.038), *LPA* c.4289 + 1G > A (MAF 0.029), *HBB* p.Glu7Val (MAF 0.012) or *G6PD* p.Val68Met (MAF 0.033)].

Further filtering

The simplest way to filter variants (apart from by population frequency) is using in silico pathogenicity prediction methods. HGMD contains many of the predictions provided by dbNSFP3 (Liu et al. 2016), which may be utilised for this purpose. HGMD also contains terms present in the Gene Ontology database (The Gene Ontology Consortium 2017). This information (e.g. "mismatch repair", "ATP binding" etc.) may be utilised if the user is aware of which ontology terms may be linked to their phenotype(s) or gene(s) of interest. Mapped UMLS disease concepts may be utilised to stratify results according to a broad set of disease-related terms (such as "Blood Disorder", "Cancer" etc.). Where the phenotype is known, this approach is very efficient at identifying and returning HGMD alleles most relevant to the phenotype of interest, a method that is increasingly being recognized as important (Amin and Wilde 2018). An example of an exome prioritization/filtering workflow is given in Fig. 4.

De novo mutations

Human germline de novo mutations are both a driver of evolution and an important cause of genetic disease (Goldmann et al. 2019; Veltman and Brunner 2012; Acuna-Hidalgo et al. 2016). Indeed, whole-genome studies have suggested that de novo mutations may be responsible for a considerable proportion of congenital or early-onset neurodevelopmental disorders, including autism spectrum disorder, epilepsy and intellectual disability/developmental delay (Neale et al.

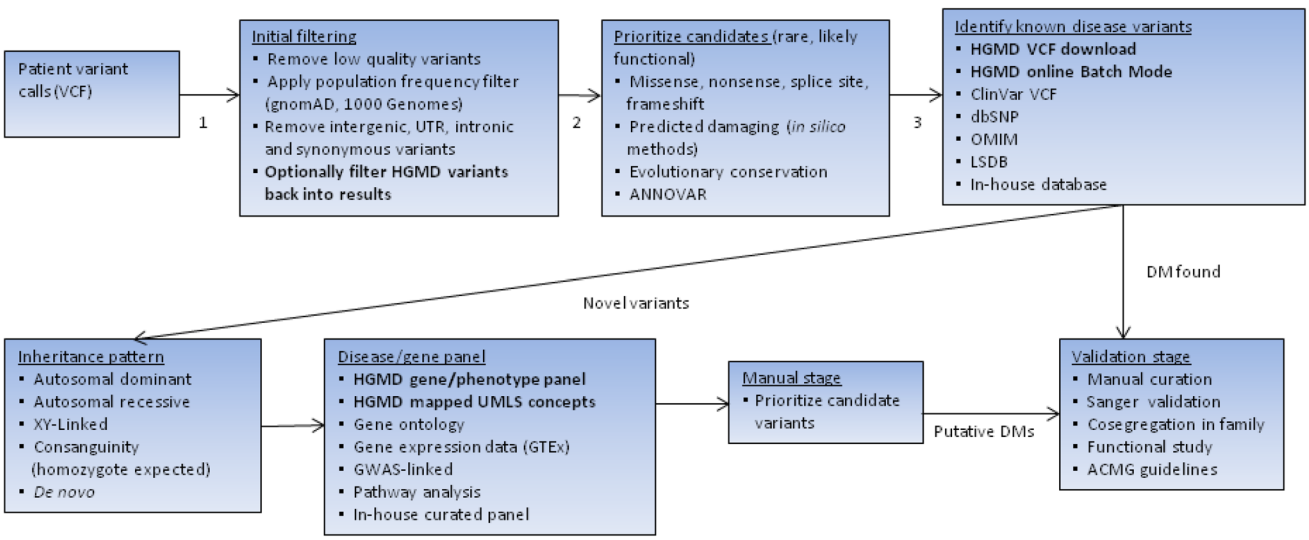

Fig. 4 Example of an NGS/diagnostic workflow

2012; Iossifov et al. 2014; Hamdan et al. 2017). Although such disorders often display a complex multifactorial aetiology (Guo et al. 2018), it is thought that autism spectrum disorder in particular has a large risk of recurrence in families (Breuss et al. 2020). Although these studies are still at an early stage, they generally show a measurable effect on disease risk, especially for exonic loss-of-function (LOF) de novo variation (Takata 2019). HGMD has, therefore, taken the decision to include these mutations, owing to the increased likelihood of their being involved in de novo (non-familial) phenotypes.

De novo mutations identified as part of large-scale mutation screening programs in patients with developmental disorders are entered into HGMD under the DM? variant class unless there is convincing additional evidence to support their inclusion as DMs. All likely disruptive sequence changes identified in cases (but not controls, or unaffected siblings in parent–offspring groups) are entered. Such variants 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 rearrangements. Other variant types (e.g. synonymous substitutions) may be considered for inclusion if additional evidence supportive of pathogenicity has been presented. This collection of de novo variants should prove useful to those undertaking large-scale screening programs in terms of checking for the known or suspected involvement of a particular gene or specific mutation (or mutation type) in a given neurodevelopmental disorder. Additional references will be added in the case of those mutations found recurrently in the literature, thereby adding to the weight of evidence supporting the involvement of a specific mutation or gene in a given disorder. There are now approximately 15,000 de novo mutations logged in HGMD; 2500 of these have at least one additional reference and 33 of them have been reclassified as DMs.

Other resources

There are a relatively small number of other sources (both publicly funded and commercial) of mutation data analogous to HGMD that are available to the scientific community. These include ClinVar (Landrum et al. 2020) (public), CentoGene (https://www.centogene.com) (commercial), LOVD (Fokkema et al. 2011) (software is public), COSMIC (Tate et al. 2019) (public and commercial license), DECIPHER (Bragin et al. 2014) (public), dbSNP (https://www.ncbi.nlm.nih.gov/snp) (public) and OMIM (Amberger et al. 2019) (public and commercial license). Like-for-like comparisons between these resources are very difficult, as obtaining the data can be problematic due to licensing requirements (CentoGene), or being distributed over many installations with potentially different usage/permission terms (LOVD). Although HGMD links to COSMIC, the data present in the latter resource are somatic in nature, and therefore, not directly comparable to HGMD. DECIPHER records data that are complementary to HGMD (i.e. large chromosomal rearrangements involving multiple genes which HGMD does not systematically catalogue). A basic comparison between the number of variants annotated by HGMD, ClinVar and OMIM is presented in Fig. 5.

Data for this comparison were limited to HGMD DM or DM? entries with genomic coordinates (March 2020 release 2020.1 VCF file) versus those ClinVar variants labelled with Pathogenic, Likely_pathogenic or Pathogenic/Likely_pathogenic assertions from the ClinVar VCF file (downloaded 2020-02-03). OMIM data were identified via the OMIM allelic variant identifier present in the ClinVar download. It can be seen that HGMD captures all OMIM pathogenic allelic variants, plus almost 60% of those present in ClinVar. In contrast, ClinVar captures only 23% of HGMD variants, with the majority (77%) remaining specific to HGMD. The remaining 43,342 ClinVar-only entries appear to comprise unpublished variants submitted by diagnostic laboratories (which complement HGMD's attempts to provide comprehensive cover of the peer-reviewed literature). There is in addition a small overlap of ClinVar variants of uncertain significance/VUS (17,166 variants—not shown in Fig. 5) between HGMD and ClinVar. The bulk of ClinVar VUS (204,337 entries) are however not present in HGMD. The reason for this lies with HGMD editorial policy. Our curators will include a published VUS (as a DM?) if it is deemed plausible that the variant is the cause of the patient's phenotype (e.g. VUS present in a known disease gene where it is rare in the general population, labelled VUS by the authors and is considered to be a reasonable potentially causative finding). ClinVar, however, appears to include practically all submitted VUS, even when the affected status of the individual or family is unknown, leading to much larger numbers of this type of variant being present in their dataset.

Variant reclassification

HGMD will reclassify or retire a variant if published evidence comes to light (e.g. via functional, case-level or mass exome variant frequency studies) that supports reclassification. Variants may be reclassified from DM? to DM if the new evidence increases support for the potential pathogenicity of the variant in question. The reclassification can of course go the other way (DM to DM?) where new data emerge that argue against variant pathogenicity. Disease-associated polymorphisms (DP) may be reclassified to DFP if new evidence supporting a functional effect is published. A variant may also be retired (R) if found to have been included erroneously ab initio, or

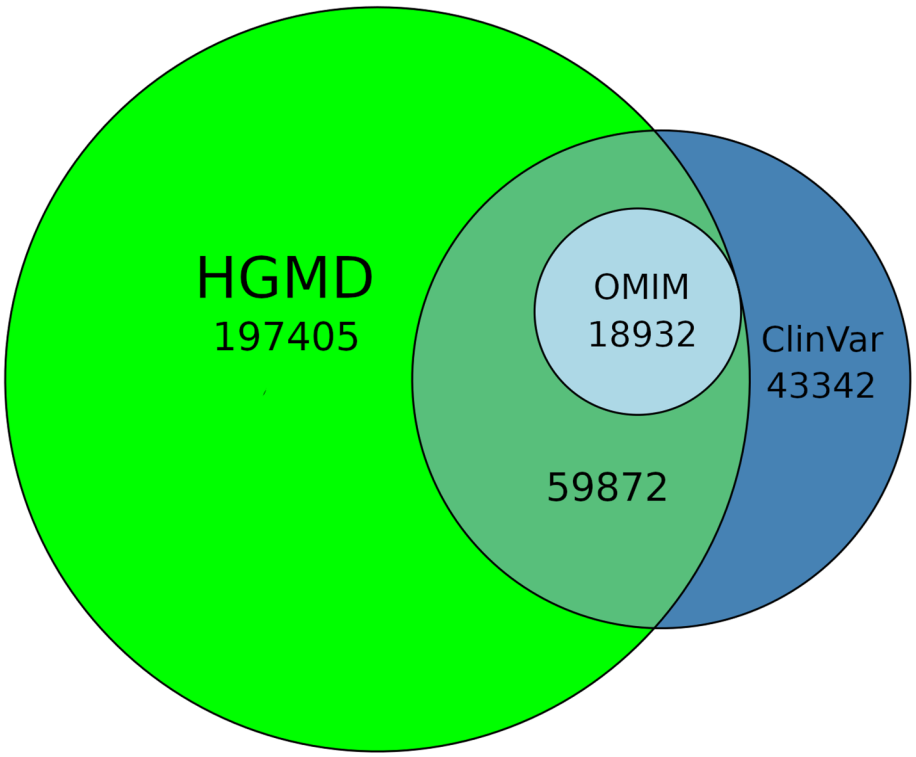

Fig. 5 HGMD vs ClinVar vs OMIM comparison (as of March 2020)

has subsequently been shown beyond reasonable doubt to be benign (either by virtue of its apparent population frequency, or literature reclassification or retraction). ClinVar variant reclassification rates (where a submitting laboratory updates a variant classification) are broadly similar to those of HGMD. The ClinVar reclassification rate has been reported to be 0.79% (Harrison and Rehm 2019), whereas the equivalent rate for HGMD data was 1.12% over the same time period (all data entered into HGMD between January 2016 and July 2019). However, if all HGMD entries are included (irrespective of the date when they were first entered), then the HGMD reclassification rate rises to 2.06%. This is to be expected, as HGMD has pursued a policy of continuous curation, re-annotating older data wherever necessary, whereas ClinVar appears to rely almost exclusively on the original submitter updating their submission (hence the lower rate of reclassification). Recent literature suggests that if ClinVar alleles are independently re-interpreted, then a large number of reclassifications become necessary (Xiang et al. 2020). Our view is that reclassification is to be expected with any well maintained mutation database, which should always be considered to be "work in progress".

Automated mutation retrieval

Researchers at HGMD have been involved for several years in attempts to automatically extract mutation data from the literature. We recently contributed towards the Automatic Variant Evidence Database (AVADA), a novel machine learning tool that uses natural language processing to automatically identify pathogenic genetic variant evidence in full-text primary literature (Birgmeier et al. 2020). AVADA automatically retrieved 58% of the likely disease-causing variants deposited in HGMD. Automatic retrieval and verification of novel likely disease-causing variants involved in inherited disease (i.e. those not already verified via manual curation) is, however, much more challenging. Our own internal assessment has demonstrated that > 90% of novel (i.e. unknown to HGMD) computationally derived automated literature mutation "hits" are in fact (from HGMD's point of view) false positives comprising a mixture of several different types; (i) somatic mutations (e.g. cancer driver mutations or mutations conferring cancer therapy resistance), i(i) non-human mutations (i.e. from mouse, or another model organism), iii) artificially engineered mutations (e.g. mutagenesis experiments looking at catalytic or other active sites of a protein), (iv) so-called "A.I. artifacts" (i.e. supposed variant matches a specified text mining pattern, but is not a genuine mutation), (v) incorrectly or inadequately described human mutations (e.g. protein description not matching nucleotide description, requiring manual verification, but entered verbatim by the algorithm), (vi) secondary mRNA or other post-transcriptional sequelae (e.g. skipped exons or intron inclusions), (vii) coincidentally co-located mutations matching positions in two or more different genes, but only being genuine for one of them, and finally (viii)

benign variants (polymorphisms or rare variants only found in healthy controls). The tracking down of such false leads involved a great deal of HGMD editorial/curation time, but did not lead to a corresponding increase in identified novel disease-relevant mutations. Owing to these limitations, we have opted to implement a strictly controlled form of automated additional reference retrieval, using pre-existing and well-described human-curated HGMD variants.

Automated additional reference retrieval

To minimise the possibility of accruing the aforementioned false positives, HGMD has taken the decision to strictly limit the application of automated mutation retrieval/identification methods to our previously catalogued human mutation data set, identifying only the additional literature mentions of these already verified mutations. The computerised method employed will examine the full-text of identified articles (HTML or PDF including any supplementary material). Any literature reference in which a mention of a previously described mutation is found will be recorded and entered into HGMD as an additional reference for that mutation. These additional auto-curated references will be clearly marked as non-human curated when presented to HGMD users.

Future plans

HGMD plans to include the data from the GTEx project (GTEx Consortium 2013) to allow filtering on tissue expression of particular genes. This, in combination with the Gene Ontology and UMLS, should allow even more efficient filtering (e.g. combining "central nervous system disorder" from the UMLS with "brain expressed" genes from GTEx). We also have plans to expand our provision of in silico variant predictions and to include computed ACMG classifications, based on the ACMG 2.0 rules recently published (Kalia et al. 2017). Roll-out of our automated additional reference retrieval system is also a priority.

Conclusion

In conclusion, HGMD contains an expansive set of tools that may be utilised by users in the fields of clinical diagnostics, personalised genomics and NGS/bioinformatics research to search and prioritise results derived from its comprehensive mutation data set. The onus is, however, on the clinician or researcher to use these tools and data sensibly and appropriately to obtain results that are suitable for their own use cases.

Acknowledgements We would like to acknowledge previous financial support provided by Celera Genomics and Biobase GmbH, as well as the ongoing support and collaboration from Qiagen Inc.

Compliance with ethical standards

Conflict of interest The authors wish to declare an interest insofar as HGMD is financially supported by Qiagen Inc. through a License agreement with Cardiff University.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68–74

Acuna-Hidalgo R, Veltman JA, Hoischen (2016) A New insights into the generation and role of *de novo* mutations in health and disease. Genome Biol 17:241

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249

Amberger JS, Bocchini CA, Scott AF, Hamosh A (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 47:D1038–D1043

Amin AS, Wilde AAM (2018) The phenotype is equally important in promoting variants from benign to pathogenic as well as in demoting variants from pathogenic to benign. Heart Rhythm 15:562–563

Birgmeier J, Deisseroth CA, Hayward LE, Galhardo LMT, Tierno AP, Jagadeesh KA, Stenson PD, Cooper DN, Bernstein JA, Haeussler M, Bejerano G (2020) AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature. Genet Med 22:362–370

Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ (2014) DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res 42:D993–D1000

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breuss MW, Antaki D, George RD, Kleiber M, James KN, Ball LL, Hong O, Mitra I, Yang X, Wirth SA, Gu J, Garcia CAB, Gujral M, Brandler WM, Musaev D, Nguyen A, McEvoy-Venneri J, Knox R, Sticca E, Botello MCC, Uribe Fenner J, Pérez MC, Arranz M, Moffitt AB, Wang Z, Hervás A, Devinsky O, Gymrek M, Sebat J, Gleeson JG (2020) Autism risk in offspring can be assessed through quantification of male sperm mosaicism. Nat Med 26:143–150

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

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

den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE (2016) HGVS recommendations for the description of sequence variants: 2016 Update. Hum Mutat 37:564–569

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT (2011) LOVD vol 2.0: the next generation in gene variant databases. Hum Mutat 32:557–563

Goldmann JM, Veltman JA, Gilissen C (2019) De novo mutations reflect development and aging of the human germline. Trends Genet 35:828–839

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585

Guo H, Wang T, Wu H, Long M, Coe BP, Li H, Xun G, Ou J, Chen B, Duan G, Bai T, Zhao N, Shen Y, Li Y, Wang Y, Zhang Y, Baker C, Liu Y, Pang N, Huang L, Han L, Jia X, Liu C, Ni H, Yang X, Xia L, Chen J, Shen L, Li Y, Zhao R, Zhao W, Peng J, Pan Q, Long Z, Su W, Tan J, Du X, Ke X, Yao M, Hu Z, Zou X, Zhao J, Bernier RA, Eichler EE, Xia K (2018) Inherited and multiple *de novo* mutations in autism/developmental delay risk genes suggest a multifactorial model. Mol Autism 13(9):64

Hamdan FF, Myers CT, Cossette P, Lemay P, Spiegelman D, Laporte AD, Nassif C, Diallo O, Monlong J, Cadieux-Dion M, Dobrzeniecka S, Meloche C, Retterer K, Cho MT, Rosenfeld JA, Bi W, Massicotte C, Miguet M, Brunga L, Regan BM, Mo K, Tam C, Schneider A, Hollingsworth G, FitzPatrick DR, Donaldson A, Canham N, Blair E, Kerr B, Fry AE, Thomas RH, Shelagh J, Hurst JA, Brittain H, Blyth M, Lebel RR, Gerkes EH, Davis-Keppen L, Stein Q, Chung WK, Dorison SJ, Benke PJ, Fassi E, Corsten-Janssen N, Kamsteeg EJ, Mau-Them FT, Bruel AL, Verloes A, Õunap K, Wojcik MH, Albert DVF, Venkateswaran S, Ware T, Jones D, Liu YC, Mohammad SS, Bizargity P, Bacino CA, Leuzzi V, Martinelli S, Dallapiccola B, Tartaglia M, Blumkin L, Wierenga KJ, Purcarin G, O'Byrne JJ, Stockler S, Lehman A, Keren B, Nougues MC, Mignot C, Auvin S, Nava C, Hiatt SM, Bebin M, Shao Y, Scaglia F, Lalani SR, Frye RE, Jarjour IT, Jacques S, Boucher RM, Riou E, Srour M, Carmant L, Lortie A, Major P, Diadori P, Dubeau F, D'Anjou G, Bourque G, Berkovic SF, Sadleir LG, Campeau PM, Kibar Z, Lafrenière RG, Girard SL, Mercimek-Mahmutoglu S, Boelman C, Rouleau GA, Scheffer IE, Mefford HC, Andrade DM, Rossignol E, Minassian BA, Michaud JL, Deciphering Developmental Disorders Study (2017) High rate of recurrent *de novo* mutations in developmental and epileptic encephalopathies. Am J Hum Genet 101:664–685

Harrison SM, Rehm HL (2019) Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. Genome Med. 11:72

Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paeper B, Nickerson DA, Dea J, Dong S, Gonzalez LE, Mandell JD, Mane SM, Murtha MT, Sullivan CA, Walker MF, Waqar Z, Wei L, Willsey AJ, Yamrom B, Lee YH, Grabowska E, Dalkic E, Wang Z, Marks S, Andrews P, Leotta A, Kendall J, Hakker I, Rosenbaum J, Ma B, Rodgers L, Troge J, Narzisi G, Yoon S, Schatz MC, Ye K, McCombie WR, Shendure J, Eichler EE, State MW, Wigler M (2014) The contribution of *de novo* coding mutations to autism spectrum disorder. Nature 515:216–221

Kalia SS, Adelman K, Bale S, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, Miller DT (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med 19:249–255

Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, Chen R, Sirota M, Kodama K, Hadley D, Butte AJ (2018) Are minor alleles more likely to be risk alleles? BMC Med Genom 11:3

Kounelis F, Kanterakis A, Kanavos A, Pandi MT, Kordou Z, Manusama O, Vonitsanos G, Katsila T, Tsermpini EE, Lauschke VM, Koromina M, van der Spek PJ, Patrinos GP (2020) Documentation of clinically relevant genomic biomarker allele frequencies in the next-generation FINDbase worldwide database. Hum Mutat 41:1112–1122

Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C, Lyoshin V, Maddipatla Z, Maiti R, Mitchell J, O'Leary N, Riley GR, Shi W, Zhou G, Schneider V, Maglott D, Holmes JB, Kattman BL (2020) ClinVar: improvements to accessing data. Nucleic Acids Res 48:D835–D844

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291

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

Liu X, Wu C, Li C, Boerwinkle E (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat 37:235–241

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH Jr, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ (2012) Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. Nature 485:242–245

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47:D886–D894

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136:665–677

Subramanian S (2018) Abundance of clinical variants in exons included in multiple transcripts. Hum Genom 12:33

Tabor HK, Auer PL, Jamal SM, Chong JX, Yu JH, Gordon AS, Graubert TA, O'Donnell CJ, Rich SS, Nickerson DA, Bamshad MJ, NHLBI Exome Sequencing Project (2014) Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. Am J Hum Genet 95:183–193

Takata A (2019) Estimating contribution of rare non-coding variants to neuropsychiatric disorders. Psychiatry Clin Neurosci 73:2–10

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA (2019) COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 47:D941–D947

The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 45:D331–D338

Veltman JA, Brunner HG (2012) *De novo* mutations in human genetic disease. Nat Rev Genet 13:565–575

Wang NK, Chiang JPW (2019) Increasing evidence of combinatory variant effects calls for revised classification of low-penetrance alleles. Genet Med 21:1280–1282

Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, Barton PJR, Funke B, Cook SA, MacArthur D, Ware JS (2017) Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med 19:1151–1158

Xiang J, Yang J, Chen L, Chen Q, Yang H, Sun C, Zhou Q, Peng Z (2020) Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. Sci Rep 10:331

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C, 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

Zernant J, Lee W, Collison FT, Fishman GA, Sergeev YV, Schuerch K, Tsang Sparrow JR, Tsang SH, Allikmets R (2017) Frequent hypomorphic alleles account for a significant fraction of ABCA4 disease and distinguish it from age-related macular degeneration. J Med Genet 54:404–412