Knowledge multiplies when shared — when calling things by their right name: improving the validation and exchange of genetic data in research and diagnostics

Fokkema, I.F.A.C.

CHAPTER 3

Data sharing and gene variant databases

Ivo F.A.C. Fokkema[1], Johan T. den Dunnen[1,2]

1 - Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands.
2 - Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands.

Adapted from: *Data sharing and gene variant databases.*
Chapter in: *Clinical DNA variant interpretation: Theory and Practice.* (2021) Elsevier.
DOI: 10.1016/C2019-0-01138-7.

3.1 Abstract

DNA diagnostics heavily relies on shared information on genes, variants, and phenotypes. Without this data sharing, DNA diagnostics would be severely hampered. Data sharing is also, by far, the cheapest way to classify variants. A large fraction of the many Variants of Unknown Significance (VUSes), i.e., variants for which there is too little information to classify them reliably, will likely immediately be solved when data is shared more efficiently. In the ideal situation, diagnostic laboratories would automatically share their data with a public international repository.

Besides their role in diagnostics, databases direct research, give insights into a gene's function, guide rationally designed treatments, and promote interaction between scientists, between scientists and patients, and between patients. There are hundreds of databases containing information on genes, variants, and phenotypes. We focus here on those containing information aiding the clinical classification of a variant. We discriminate two major types: general and focused databases. Of the general databases, we briefly describe OMIM, dbSNP/EVA, and gnomAD, the information they contain, and their use in diagnostics. The focused databases HGMD, LOVD, and ClinVar are described in more detail.

3.2 Introduction

DNA diagnostics heavily relies on shared information on genes, variants, and phenotypes. Without this data sharing, DNA diagnostics would be severely hampered. When we do not share our findings, we do not offer optimal care to the patients and their families. To get to reliable, evidence-based variant classification, every observation counts, strengthening the evidence we have (disease-associated or not) as well as ultimately allowing accurate risk estimates. Given these evident basic requirements, it is astonishing to see that for a long time, sharing was rather the exception than the standard and that stable financial support for many databases is still largely lacking. Data sharing is, by far, the cheapest way to classify variants. A large fraction of the so-called Variants of Unknown Significance (VUSes), i.e., variants for which there is too little information to classify them reliably, will likely immediately be solved when data is shared more efficiently and more broadly.

The daily routine in a clinical diagnostics laboratory is, in general, as follows. A sample comes in and gets sequenced, the variants found are stored in a private database, and external repositories are queried to determine the population frequency of each variant and all known information about the associated phenotypic consequences (clinical variant classification). When necessary, and where possible, additional analyses are performed, e.g., analyzing RNA to check for possible consequences on RNA processing (e.g., RNA splicing), an immunohistochemical analysis to study protein amount and localization, and *in silico* analyses to predict possible variant consequences. Finally, a conclusion is drawn, and a diagnostic report is written.

Unfortunately, the step missing from the daily routine mentioned above is the automatic sharing with a public international repository. Occasionally, interesting cases that highlight new genotype-phenotype relationships or new disease pathways are published in scientific journals. However, large-scale data sharing to online public databases is rarely achieved. This has serious consequences. First, most importantly, the data available for accurate variant classification are far from complete; in the end, probably less than 10% of all observations are published. Many variants have, therefore, to be classified as a VUS.[1] A Variant of Unknown Significance is, therefore, often a Variant of Unsufficient Sharing. Furthermore, even when variants are published in the scientific literature, this does not mean the variants can easily be found there. Most databases do not have the resources to check the literature and upload all published data. Variants published in journals are also known to often not adhere to the proper standards for describing them, leading to misinterpretations at worst or inability to use the data at best (see also Chapter 6). Because of these factors, a growing number of journals have started to require database submissions before accepting manuscripts for publication.

The main aim of variant databases is to catalog all variants encountered, collect all available information, and display and share the data collected. Depending on the focus of the database, the information collected will differ. Some will focus on (monogenic) gene-disease links, others on variant (population) frequencies, functional predictions, or associated polygenic phenotypes. The variants collected are stored and displayed using a standard description, primarily based on the HGVS recommendations to describe variants in DNA, RNA, and protein sequences[2] (**`hgvs-nomenclature.org`**). Although, in general, they will store variant descriptions based on several coordinate systems, depending on their focus, the main display is based on either chromosomal genomic coordinates, a gene, or a protein. All major databases link to each other.

This chapter focuses on the use of DNA variant databases in DNA diagnostics. Besides the diagnostic application, databases have a much broader use. First, they support research in several ways. The variants, the variant type, their location in the gene/protein, and the consequences reported give important insights into the gene's function, the cellular processes they are involved in, and how this affects an organism. Variants for which too little information is available to classify them reliably are obvious candidates to be tested in functional assays, in vitro gene-splicing experiments, or animal model systems. Known variant effects can be used to train and benchmark predictive computational tools. Further research is directed toward the aspects that are not understood, i.e., variants with unexpected or unexplained effects and genetic modifiers that influence disease outcomes. When the correlation between genotype and phenotype is understood, this can be used to design a rational approach towards a possible treatment. Possible treatments may target specific variant types, e.g., exon-skipping or stop-codon read-through, where the database can be used to guide a design based on the observed frequencies.[3] Validating treatment success is guided by collected data on disease progression, pointing out important outcome measures, and comparing these between treated and untreated individuals. Another database function is to promote interaction between scientists, between scientists and patients, and between patients. Scientists find information on open questions that need further research and colleagues who may have useful biological samples to support such studies. Using the databases, scientists find leads to get into contact with reporting labs for collaboration and biological samples and to patients for further research or to participate in clinical trials. Patients use the databases to see what information is available regarding their condition and to find patients with similar or identical genotypes ("a patient like me"). Ultimately, when a treatment becomes available, the databases are a source to find patients who may benefit from the treatment.

3.3 General databases

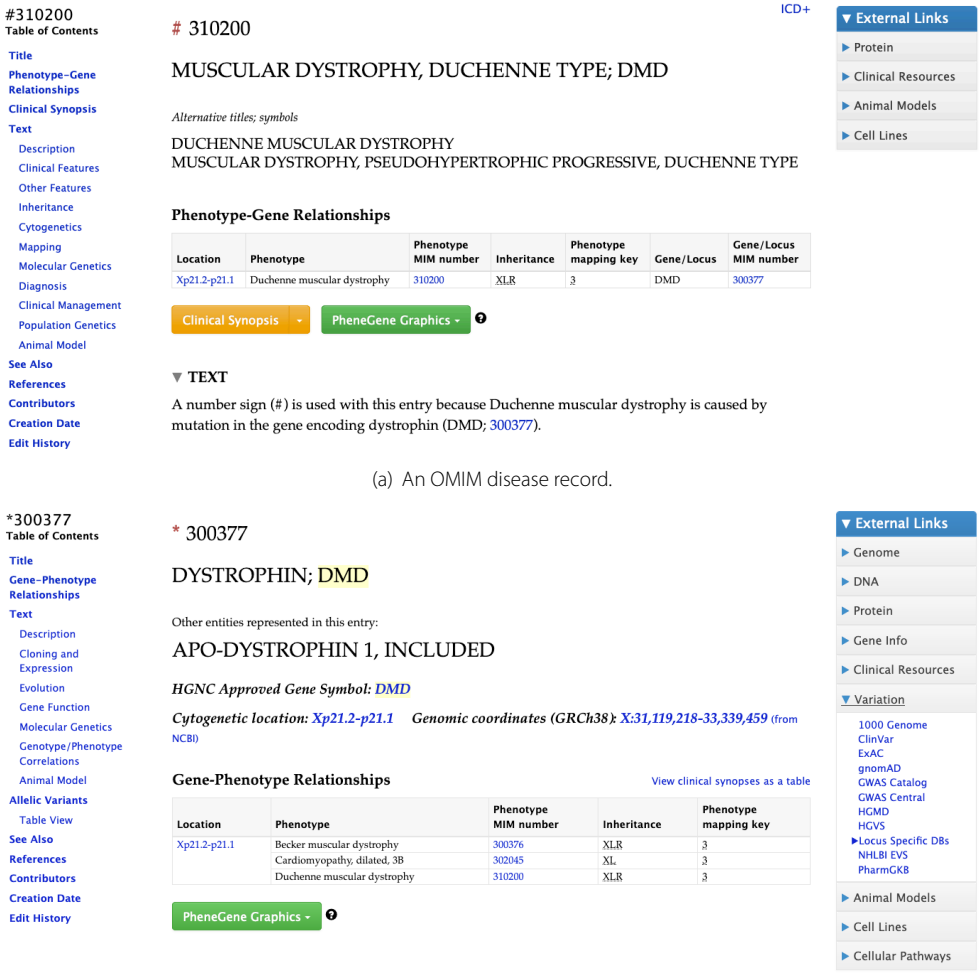
There are two main types of variant databases: general and focused databases. The general databases are "a mile wide and an inch deep", while the focused databases are "an inch wide and a mile deep". The general databases try to catalog all variants published and direct to focused databases for further details. The focused gene or disease variant databases try to collect all information from all available cases, published and unpublished. Discussing all available databases in this chapter will be impossible, so we will focus on those used most frequently in clinical diagnostics. The genome browsers, e.g., UCSC (`genome.ucsc.edu`) and Ensembl (`ensembl.org`), serve an important intermediate function by offering tracks in their display linking to many available datasets.

3.3.1 OMIM

The OMIM database (Online Mendelian Inheritance in Man, `omim.org`) focuses on the relationship between genes and phenotypes.[4] Although the focus is on genetic disorders, all genetic phenotypes are covered. OMIM is a freely available authoritative compendium which is updated daily. The database has two types of records: phenotype records (Figure 3.1A) and gene records (Figure 3.1B). The records link to each other, but not all phenotypes have been linked to a gene, nor have all genes been linked to a phenotype.

Phenotype records include a short historical perspective and detail the clinical features, the modes of inheritance observed, disease diagnosis, and how the disease was mapped and linked to a specific gene or genomic region. Gene records give details on the gene, the encoded protein, its (suggested) function, and the existence of animal models. Details are provided when a gene has been linked to one or more phenotypes, and a range of phenotype-associated variants are listed. The number of variants listed differs greatly; the focus is on the first gene-phenotype reports. In addition, a series of variants may be described, representing both typical and exceptional cases, together giving a comprehensive overview of the observed genotype-phenotype relations. The reported variants are linked to dbSNP, ClinVar, and gnomAD. Links to other sources are provided through an "External Links" menu.

OMIM does not yet use HPO terms (Human Phenotype Ontology)[5] when describing phenotypes, but for frequently observed features, these can be obtained from, e.g., the HPO website (`hpo.jax.org`). Clinical labs mainly use OMIM identifiers and disease names when reporting individual phenotypes.



(a) An OMIM disease record.

(b) An OMIM gene record.

Figure 3.1: **The OMIM database.** (a) OMIM disease record for Duchenne Muscular Dystrophy (#310200). The record gives summary data on the disease and, when known, the inheritance pattern(s) observed, the genomic location, and the gene(s) involved. Every record has a specific format containing the topics discussed on the left and links to other resources on the right. (b) OMIM gene record for the dystrophin (*DMD*) gene (*300377). The record gives summary data on the gene and its location and, when known, the phenotypes in which it is known to be involved. Every record has a specific format containing the topics discussed on the left and links to other resources on the right (the Variation menu has been opened).

3.3.2 GWAS

While OMIM focuses on monogenic disorders, it does store information on all genetic phenotypes, including polygenic traits revealed by genome-wide association studies (GWAS). Specific GWAS databases, e.g., the GWAS catalog (**ebi.ac.uk/gwas**) and GWAS Central (**mart.gwascentral.org**), focus on such studies in more detail and contain large amounts of information on variants which, through large population studies, have been linked to a specific phenotype. While these databases are rarely used in clinical genetics, GWAS signals mapped to a genomic region may add additional evidence to establish new gene-phenotype links.

3.3.3 dbSNP and EVA

dbSNP (**ncbi.nlm.nih.gov/snp**) and EVA (European Variation Archive, **ebi.ac.uk/eva**) are databases that try to catalog all small human DNA variants, including single-nucleotide variants, microsatellites, and small insertions and deletions. To every variant, an identifier is assigned, an "rs" number (e.g., rs104894790), and details are given regarding its genomic location and possible transcripts covering the variant position. In addition, information is provided about the populations in which the variant has been identified and its frequency. Both databases link to a range of other sites containing additional information.

3.3.4 gnomAD

While in essence, the gnomAD database[6] (**gnomad.broadinstitute.org**, Figure 3.2) and its predecessor ExAC collect similar information, they are different from dbSNP and EVA that collect data from many different sources. gnomAD is unique since it is based on data from 730,947 exomes and 76,215 genomes (**gnomad.broadinstitute.org/stats**, visited 2024-06-23) from unrelated individuals from different populations, analyzed by one group using a standard analysis pipeline and calling variants using the same thresholds. Given the high quality of the data and the large number of samples analyzed, the gnomAD database can be used to study several aspects of gene function. Using a mutation rate model, gnomAD classified all protein-coding genes from mutation tolerant to mutation intolerant. Over 3000 genes, more than half of which have currently not been linked to a human disease phenotype, were shown to be largely devoid of variants predicted to be protein-truncating. In other genes, high frequencies of variants predicted to be truncating were found, even in homozygous cases, making it highly unlikely that these genes are associated with human disease phenotypes. Most labs use databases like gnomAD, dbSNP, and EVA as indispensable sources to find variant frequencies, where a high frequency reduces the chance that a variant has serious consequences. For variants in genes that have not yet been linked to a disease

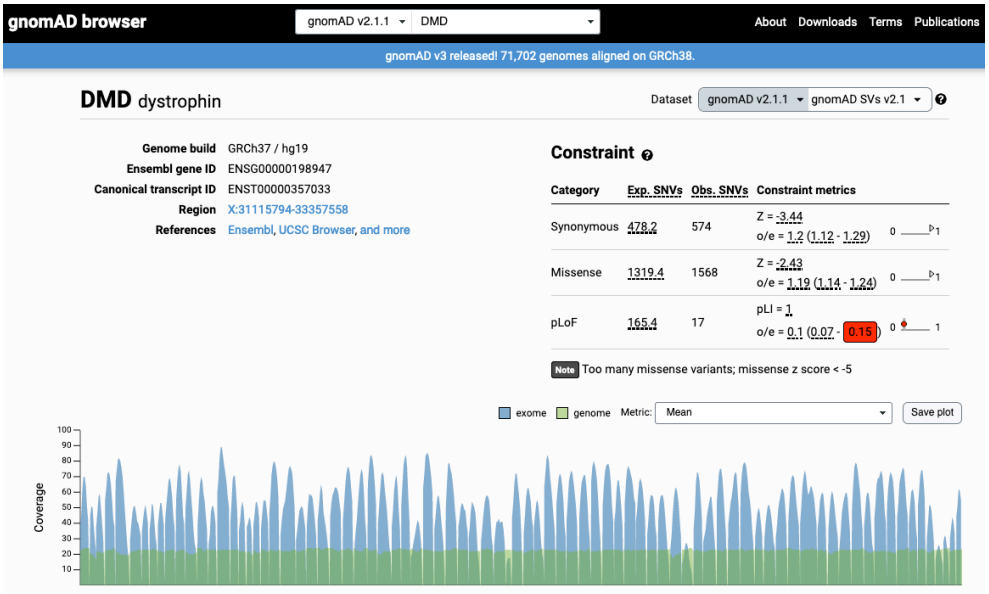


Figure 3.2: **The gnomAD database.** The record for the *DMD* gene, its exon/intron structure, and the mean coverage obtained in exome and genome sequencing studies are shown. The panel on the right ("Constraint") shows the number of variants in the *DMD* gene as observed and expected for three categories: synonymous, missense, and pLoF (predicted loss-of-function). The number of pLoF variants observed is significantly below the number expected, indicating that the gene has an essential function.

phenotype, the overall mutation rate available from gnomAD gives an indication of how likely it is that the gene has a critical function.

3.4 Focused databases: Gene variant databases

Historically, focused variant databases were started by experts working in, or associated with, DNA diagnostics. These so-called Locus-Specific Databases (LSDBs) focus mainly on a gene or a disease caused by variants in several genes. The databases differed widely, using different platforms/database software, collecting details on variable phenotypic features, and displaying variant data in various formats.

Initially, there was little collaboration, and often, several databases were started for the same major disease genes, e.g., *TP53*, *BRCA1* and *BRCA2*, the colon cancer genes, etc. Stimulated by the HUGO Mutation Database initiative,[7] LSDBs joined forces, developed shared international standards,[8] and gradually merged into larger consortia. The development of freely available gene variant database software packages like LOVD[9] (`LOVD.nl`, see also Chapter 4) and UMD[10] (`umd.be`) boosted standardization and collaboration. Ultimately, the LOVD version

3 software, facilitating the collection and display of genome-wide data, persuaded many independent database installations to join efforts and merge into a central repository, the "Global Variome shared LOVD". The shared LOVD is now, by far, the largest collaborative effort to share data on genes, variants, and phenotypes. The database operates under the auspices of Global Variome, a UK charity linked to the Human Variome Project (HVP), a non-governmental organization (NGO) recognized by UNESCO (United Nations Educational, Scientific and Cultural Organization). Different LOVD installations still exist, but most public instances share basic information with the central LOVD server. The information shared is used to automatically update the LSDB gene variant database list (lsdb.variome.org) and to offer a centralized variant query service, redirecting positive hits to the LOVD installation containing the data (`LOVD.nl/3.0/search`).

Currently, besides the shared LOVD, there are two other major human gene variant database initiatives covering all human genes, HGMD[11] and ClinVar,[12] each with its own focus. All three are discussed in detail below.

3.4.1 HGMD

HGMD (the Human Gene Mutation Database, `www.hgmd.cf.ac.uk`) focuses on variants causative or associated with inherited human disease as well as disease-associated/functional variants. Unlike other databases, HGMD covers only published data and reports variant classifications as they are published, listing only the first report unless an additional report extends the original entry, e.g., based on functional studies. Also, only generic phenotype information is stored (disease name). The data stored covers all variant types within the coding regions, splicing, and regulatory regions of human nuclear genes. Somatic variants and variants in the mitochondrial genome are not included. Variants that do not alter the encoded amino acid sequence are not recorded unless they have been shown to affect mRNA splicing or gene expression or have been reported as associated with disease. Unless they have some clinical relevance, variants lacking obvious phenotypic consequences are not collected. This does limit HGMD's clinical relevance, as it does not store variants known not to be associated with disease (likely benign and benign). Also, when other sites, e.g., a genome browser, display HGMD data, it will only give positional information and not display the variant. This is because while the shared LOVD and ClinVar are freely accessible public archives, access to HGMD is restricted, with recent data requiring a paid subscription.

3.4.2 ClinVar

ClinVar (`ncbi.nlm.nih.gov/clinvar/`) is an NIH-funded public repository reporting the relationships between variants and an individual's health status, with supporting evidence to facilitate access to and communication about the relationships and the history of that

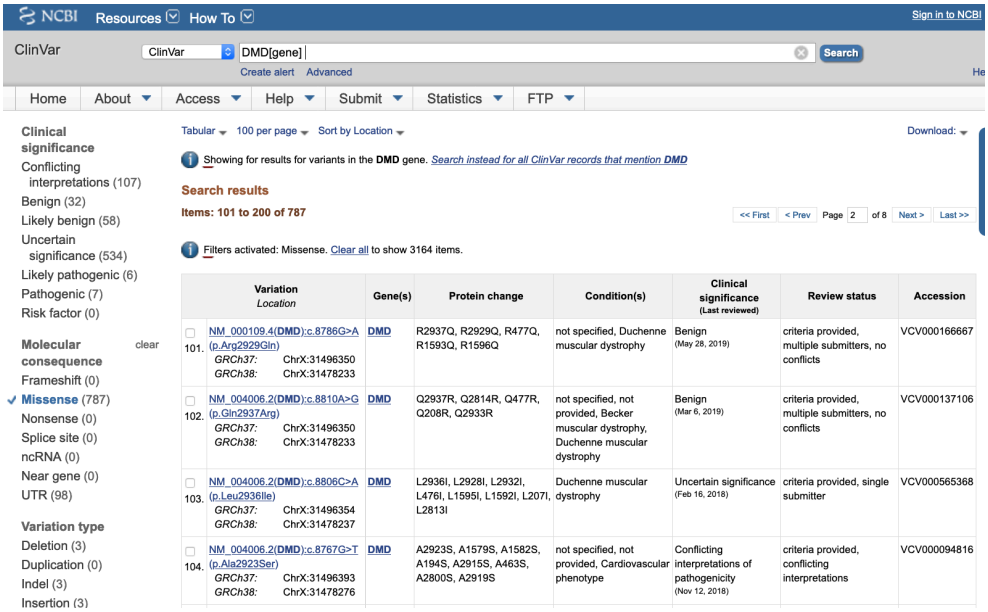


Figure 3.3: **The ClinVar database.** The database was queried for variants in the *DMD* gene (top middle) and then filtered for only missense variants (selected left under "Molecular Consequences"). Four variants are shown. The "Variation Location" column gives a variant description based on a coding DNA reference sequence (c.) and the location of the variant relative to two different genome builds, GRCh37 and GRCh38. The "c." variant description is shown based on different reference transcripts (here, `NM_000109.4` and `NM_004006.2`). The "Protein change" column gives predicted protein variant descriptions using the one-letter amino acid code and based on a range of different transcripts (up to eight are listed). Note that the HGVS Nomenclature recommends descriptions with the format `p.(Gln2937Arg)`, instead. Subsequent columns "Condition(s)", "Clinical significance", and "Review status" show summary data for all variants submitted more than once. For variant `c.8767G>T`, conflicting interpretations were shared; opening the record lists all submissions and interpretations.

interpretation. Unlike HGMD, ClinVar does not have the capacity to review published literature; therefore, it fully depends on submissions from external sources. ClinVar partners with the ClinGen project, providing data for evaluation and archiving the results of interpretation by recognized expert panels and providers of practice guidelines. ClinVar archives and versions submissions; when submitters update their records, the previous version is retained. When submitters register, they need to provide details regarding their institute, the diagnostics methods used, and the process used to evaluate and ultimately classify variants. Based on these data, ClinVar rates submitting labs. A large fraction of the data available from ClinVar comes from the major DNA diagnostics labs in the United States, sharing their variant data and classification.

The standard entry point into ClinVar is by using a gene symbol. The main display obtained

is a listing of all variants linked to that gene (e.g., `ncbi.nlm.nih.gov/clinvar/?term=DMD`), ordered based on chromosomal position (so from 3' to 5' for genes on the minus strand, see Figure 3.3). Zooming in on specific subsets of the data is achieved using a menu offering categories like "clinical significance", "molecular consequence", "variation type", "variant length", "allele origin", "review status", etc., each with a set of predefined choices (e.g., for "molecular consequence": "frameshift", "missense", "nonsense", "splice site", "ncRNA", "UTR", and "near gene"). Another option is to use the "Search builder" at the top of the screen, which can, for example, be used to query for a specific variant. Variant descriptions given mostly follow HGVS Nomenclature and show both a transcript-based (c.) and genomic chromosomal description (g.) — the latter for genome builds GRCh37 and GRCh38.

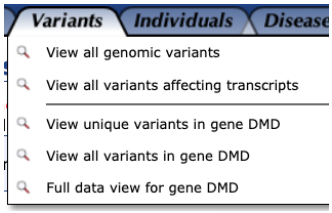
When the record for a specific variant is selected (e.g., `NM_004006.2:c.10108C>T` for *DMD*) the new variant display (`ncbi.nlm.nih.gov/clinvar/variation/11213/`) shows all independent reports of that variant, split over submitted interpretations (11 for `c.10108C>T`, visited 2024-06-23) and mentions in the literature (18 for `c.10108C>T`, linked to PubMed). The confidence level and accuracy of variant classifications depend largely on the number of observations and supporting evidence available. All data, both consensus and conflicting, are displayed. Each variant record includes a star-rated "Review status" summarizing the current status of the variant's classification, for `c.10108C>T`, a two-star rating meaning "criteria provided, multiple submitters, no conflicts". Submitted variant entries provide an "Evidence details" link showing the submitter's comments regarding the clinical significance of the variant.

3.4.3 Global Variome shared LOVD

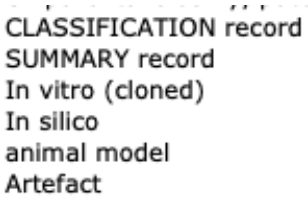
While ClinVar and the "Global Variome shared LOVD" (`LOVD.nl/shared`) have largely overlapping goals, the origin of the latter is quite different, being an unfunded community-driven initiative. The database design of the shared LOVD, as defined by the LOVD software, follows international standards.[7,8] LOVD-powered databases collect four basic types of data: the **individual** investigated, the **phenotype** observed, the **screening** (analysis) performed, and the **variant(s)** detected, including their clinical classification. While ClinVar focuses only on variants and their classification, LOVD prefers full case-level submissions containing data on all four categories listed above. LOVD also accepts simple variant-level data, e.g., for consortia reporting a variant and its overall interpretation or for functional data from studying the effect of a variant in a model system. The main goal of LOVD is to offer access to all available data to support the clinical evaluation of the possible consequences of a variant on the health of the individual carrying the variant.

The historic roots of the initiative are evident from the displays shown, which include a

(a) The LOVD page header.



(b) A drop-down menu under a tab.



(c) Various variant record types.

Figure 3.4: **The Global Variome shared LOVD.** Standard LOVD display after selecting the *DMD* gene. (a) The database offers displays based on "Genes", "Transcripts", "Variants", "Individuals", "Diseases" (phenotypes), and "Screenings". Clicking a tab will display the information selected. (b) Per tab, specific menu options become available through a mouse-over. The Variant tab offers a display of all genomic variants, all variants affecting a transcript, or variants affecting a specific gene (here *DMD*). (c) The database offers a range of special record types to label specific records. These labels can also be used to select for specific records (see text).

limited number of custom options. These options, allowing database curators to personalize their database and acknowledge their home institution, funding agencies, and others, were essential to accomplish collaboration from the diverse LSDB community. While in the shared LOVD, most curators use the preset standard data fields, some use the "custom columns" option available for the phenotype and variant displays. These columns contain data considered essential by the expert curator and are displayed in addition to the standard columns. Also, individuals and screenings can have custom columns, but these will be system-wide and can not be set by curators.

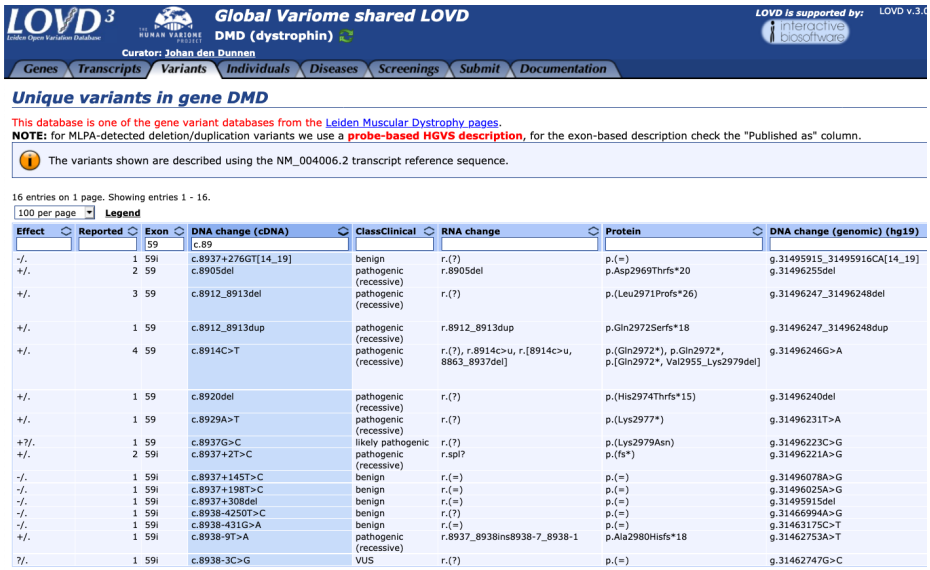
The standard entry point into the database is by using a gene symbol (e.g., `LOVD.nl/DMD`). The main display obtained is a gene variant homepage giving important details on the database. Shown are the name of the responsible curator, specific database details, links to other relevant sources, and a range of data display options, including links to view the data using the main genome browsers. The gene homepage also links to general data summaries ("Graphs") and specific tools like the "Reading Frame Checker" predicting the consequences on the protein level of multi-exon deletions or duplications.

The gene homepage display shows a series of tabs ("Genes", "Transcripts", "Variants", "Individuals", "Diseases", and "Screenings") linking to the data collected (Figure 3.4A). Per tab, a mouse-over menu shows options for more specific subsets of the data (Figure 3.4B). Selecting the "Variants" tab lists all variants, ordered based on their position in the active transcript (from 5' to 3'). Variant descriptions given follow HGVS Nomenclature and offer a choice for a transcript based (c.) and genomic chromosomal description (g.), the latter for genome builds GRCh37 and GRCh38 (Figure 3.5A). Zooming in on specific subsets of the data is achieved using the query boxes at the top of each data column. Complex queries are supported by the option to query per column and using Boolean operators for AND (space), OR ("|"), and NOT ("!"). An overview of all variants in the database, displayed on one of the major genome browsers, can be obtained from the gene homepage (Figure 3.5B). A local genome browser view around a variant selected can be obtained from the variant record.

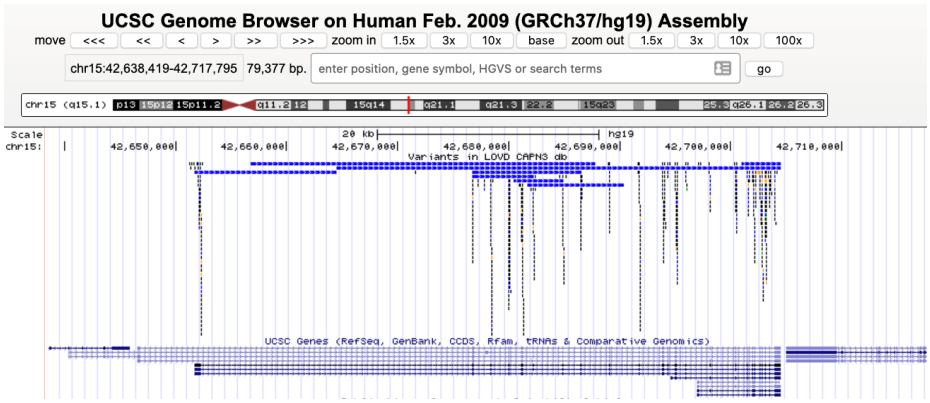
Compared to ClinVar, LOVD has some unique features. Given the structure of the database and the data collected, it is rather easy to get from a variant to an individual and the associated phenotype. In cases where a gene has an active curator, often a lot of detailed phenotypic data will have been collected and displayed. For a specific gene, several different transcripts can be linked to show variant listings (e.g., `LOVD.nl/CDKN2A`). In an LOVD database there is a mandatory "RNA" column showing whether the consequences of the variant on RNA level have been investigated and, when yes, how they affected the transcript, described using HGVS recommendations. When a variant affects RNA processing, the consequences on protein level are indicated in the "Protein" column. Other databases, including ClinVar, have no RNA field and go directly from DNA to (predicted) protein consequences, neglecting and not reporting available experimental data. Moreover, ClinVar does not offer their submitters a separate protein field for their data; instead, they only show predicted protein changes, presented as confirmed changes, as ClinVar chooses not to follow the HGVS Nomenclature guidelines on separating predictions from confirmed observations.

Unlike other databases, LOVD shows the combination of variants identified in an individual, in one gene or in different genes. LOVD therefore allows to see in which combinations variants have been found in recessive or di-genetically inherited diseases and whether this has consequences for disease severity. For variants in the *CYP* genes (see `LOVD.nl/CYP2D6`), this allows the detailed display of all variants on a specific allele (e.g., *CYP2D6*56A* `databases.lovd.nl/shared/individuals/00074484`) as well as listing the two alleles in one individual (see `databases.lovd.nl/shared/individuals/00046493`).

Besides standard variant records linked to an individual and phenotype, the shared LOVD contains a set of specifically labeled records (Figure 3.4C). Each variant can have one "summary record", when available, showing the opinion of the curator(s) or an international expert panel



(a) A variant query in the *DMD* gene.



(b) The UCSC genome browser showing LOVD data.

Figure 3.5: **Different ways of viewing variants from the Global Variome shared LOVD.** (a) The database was queried for variants in the *DMD* gene and next for all unique variants (selected from "Variants" tab). Using the "Exon" column query box (top of the column), variants in exon 59 were selected, and using the "DNA change" (cDNA) column, variants starting with "c.89" were shown. A range of variants have been retrieved, each column showing summary data. The "Effect" column shows the reported consequence(s) for the variant on gene function, the "Reported column" the number of independent variant submissions, and the "ClassClinical" column the clinical classification(s) submitted. Variant descriptions given are shown at DNA, RNA, and protein level, based on one transcript reference sequence (NM_004006.2, listed above the table) and two genome builds (only hg19 shown), all following HGVS recommendations. The "RNA change" column shows whether RNA was analyzed and, when yes, what was observed (for variants `c.8914C>T` and `c.8938-9T>A`, splicing was affected). (b) The UCSC genome browser showing the location of all *CAPN3* variants in the shared LOVD. Variants are concentrated around the exons. Large deletions are shown as blue lines, and duplications as orange lines.

regarding the classification of the variant (e.g., ENIGMA for *BRCA1/BRCA2* and InSiGHT for the colon cancer genes). "Classification records" are used when labs are not able to share individual and phenotype data but are willing to share their classification of the variant, an option pioneered by the Dutch diagnostic labs[13] (see also Chapter 8). "In vitro (cloned)" records are used to show data resulting from assays testing the consequences of the variant on the function of the gene/protein. "In silico" records show consequences as predicted by bioinformatic tools, "animal model" records data available from other organisms, and "artifact" records are used to warn for false-positive variant calls. Since these records are labeled specifically, users can either zoom in specifically on these records or exclude one or more types from their display.

Detailed phenotype queries can be performed using the "Diseases" tab. Selecting a specific phenotype opens a display showing all individuals linked to that phenotype. Selecting subsequently "Phenotype entries for this disease" opens a page facilitating a detailed comparison of all individuals linked to that phenotype. Linking variants and phenotypes can be achieved using the "Full data view for gene" link available in the drop-down menu under the "Variants" tab.

LOVD databases contain a range of APIs, facilitating the exchange of information with the central LOVD server, responding to variant queries as well as to submit data. The submission API is very powerful, allowing submitters to directly link their hospital LIMS system to automatically submit their data. Other databases (including ClinVar) rarely have this advanced option, for submissions demanding active human interference with e-mail steps and specific file formats. The submission API, initially developed for two German labs, is used by a growing number of submitters.

Access to LOVD databases is supported by a range of short URLs. An HGNC-approved gene symbol can be used to go to the list of known databases for a gene (`DMD.variome.org`) or the database for that gene in the shared LOVD installation (`LOVD.nl/DMD`). From the shared LOVD, based on the two-letter country code, a URL like `mx.LOVD.org` will retrieve a list of all variants linked to individuals from that country (in this example, Mexico) and to a list of all variants shared by submitters from that country. A link based on the database ID (`LOVD.nl/DMD_000007` for DMD_000007) immediately displays all records in the shared LOVD for that variant. Data linked to specific publications, when referenced in the submitted data, can be retrieved using their PubMed ID or DOI (`databases.lovd.nl/shared/references/PMID:23900271`, `databases.lovd.nl/shared/references/DOI:10.1038/ejhg.2013.169`).

Variant queries across all LOVD installations are offered through a central API and the LOVD website (`LOVD.nl/3.0/search`), facilitating queries using a specific genomic position as well

as using a (short) range of positions. As mentioned, LOVDs facilitate access using an API and they participate in the GA4GH Beacon project (**beacon-network.org**). When the major databases only contain a few observations of a variant, or do not contain any variant reports, the beacon project can be very helpful to quickly check whether data may be available from other less frequently used resources.

3.4.4 Other databases

While their primary focus is on small variants, both the shared LOVD and ClinVar databases include large rearrangements and genomic structural variation (multi-gene deletions and duplications, translocations, deletion-insertions, transposition, etc.) as well. There are, however, other databases that focus on structural variation specifically, e.g., NCBI's database of human genomic Structural Variation (dbVar, **ncbi.nlm.nih.gov/dbvar**) and EBI's Database of Genomic Variants archive (DGVa, **ebi.ac.uk/dgva**), currently transitioning its data to the European Variation Archive (EVA). The Database of Genomic Variants (DGV, **dgv.tcag.ca**) includes selected high-quality datasets from dbVar and DGVa, further curated for accuracy and validity. The DECIPHER database (**decipher.sanger.ac.uk**) was initiated to store, analyze, and share plausibly pathogenic structural variants from well-phenotyped patients suffering from genetic disorders. While it initially contained data from large structural variants only, when whole-exome sequencing became available, it also started to include small variant data. The focus of DECIPHER is not to catalog variants but to offer a platform to establish new gene–disease links by providing tools for variant analysis and the identification of other patients exhibiting similar genotype–phenotype characteristics.

Another important source of genetic variation is the Catalogue Of Somatic Mutations In Cancer (COSMIC, **cancer.sanger.ac.uk/cosmic**). COSMIC is the world's largest resource containing somatic variants and their impact on human cancer and can, for example, be used to determine cancer-specific mutation profiles. COSMIC is manually curated and updated four times each year. Since it is often difficult to discriminate between somatic and germline variants, COSMIC will also contain many germline variants.

Just as for DNA, hundreds of databases focus on proteins and their many different features only.[14] Most informative in relation to clinical diagnosis are those focusing on protein structure and collecting information on the consequences of variants on protein level. We would like to mention two specifically: the Protein Data Bank (PDB, **rcsb.org**), containing information about the 3D shapes of proteins, and UniProt (**uniprot.org**), containing sequences and annotations for over 120 million proteins across all branches of life. These databases facilitate the analysis of evolutionary conservation of a specific protein, highlighting evolutionary conserved and non-conserved residues and the conservation of specific functional protein

domains, e.g., an ATPase or DNA-binding domain, and their conservation across different proteins.

3.5 Final considerations

Most gene variant databases do not have the capacity to review published literature and manually add the data reported in relevant publications. Consequently, getting data published in scientific literature does not guarantee that these data will be incorporated into the major central repositories and thereby become available for automated API-based queries from exome and genome sequencing annotation pipelines. A growing number of journals have, therefore, started to demand database submission as an inherent step of the process to review and accept manuscripts for publication. Although database submission is obviously interesting to authors as their data will be easier to find, they mostly consider it just as additional work. However, a lot is gained when data are first submitted to a database.[15] The author will get a free consistency check, improving overall data quality. In addition, database submission results in the ability to link to the database for tabular overviews (e.g., the shared LOVD DOI link mentioned above), obviating the need to add supplemental data to the submitted manuscript.

Before submitting data, authors should first check the options available, read the manual or follow explanatory presentations/videos offered, and browse the database to get an idea of the data collected and the format used to store them. It is also wise to compare the options offered. While most databases have the option to submit data individually through web forms, others may offer a batch upload using a standard format or assistance when data have been stored in other formats. The shared LOVD even supports automated submission by linking to a hospital LIMS system.

It is important to mention that while databases check data quality before it is uploaded, they only display the information submitted by users. They do, in general, not rate this information or, for example, classify variants. Therefore, it is incorrect to state that according to LOVD or ClinVar, the variant is classified as, e.g., pathogenic (class 5). Classifying a variant is the investigator's responsibility, and the databases aim to support this classification by displaying relevant data. This may mean "conflicting data" is shown for a specific variant, i.e., different variant classifications submitted by more than one submitter. Conflicting classifications often derive from a lack of initial information, i.e., when the variant was reported for the first time (older submissions), it could not be accurately classified. Over time, more information will be collected, and a more specific classification will become possible (toward benign or pathogenic). Unfortunately, submitters rarely come back to update older submissions. In other cases, conflicting classifications show that current opinions did not yet converge to a

consensus classification.

An excellent starting point for performing variant analysis is a genome browser. Using the browser, the gene, transcript, and protein involved, as well as its genomic location, evolutionary conservation, etc., can be visualized. In addition, when the proper tracks have been activated, one can immediately see what variants and variant types have been identified before and where more detailed information can be found. Informative tracks to activate include dbSNP/EVA, gnomAD (ExAC, EVS), LOVD, ClinVar, COSMIC, HGMD, DGV, DECIPHER, OMIM, UniProt, GWAS, and variants in papers. When a variant is present in any of these resources, the browser will provide a direct link to this information.

A problem that cannot be solved easily is that there are way too many databases.[16] While the intention behind starting a new database in general is good, in the end, it just worsens the problem. Instead of starting a new database, the way forward is to find one that comes close but lacks features or data that are considered essential and then join efforts and add what is missing. Unfortunately, such collaborations are not supported by funding agencies, where it is often easier to get support to build a new database than to extend and sustain an existing one.

The database curator performs an important task. In general, a curator is an unpaid, voluntary expert willing to spend some free time on the database. A concise database curator who actively collects information and contacts colleagues, convincing them to share unpublished information, will experience that the efforts invested are well appreciated. The curator will receive many compliments and invitations to present the database work, be considered a world-leading expert on the gene/disease, get many opportunities for collaborative research, and will have the option to publish database updates. Given the number of genes in the human genome and the number of gene variant databases that still lack an active curator, anybody working in the field of clinical diagnostics should feel obliged to become involved and volunteer for at least one gene!

3.6 References

[1] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. *Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. Genetics in Medicine 2015; 17 (5) 405–424.

[2] Johan T. den Dunnen, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, and Peter

E.M. Taschner. *HGVS Recommendations for the Description of Sequence Variants: 2016 Update*. Human Mutation 2016; 37 (6) 564–569.

[3] Annemieke Aartsma-Rus, Ivo F.A.C. Fokkema, Jan Verschuuren, Ieke Ginjaar, Judith van Deutekom, Gert Jan van Ommen, and Johan T. den Dunnen. *Theoretic applicability of antisense-mediated exon skipping for Duchenne muscular dystrophy mutations*. Human Mutation 2009; 30 (3) 293–299.

[4] Joanna S. Amberger, Carol A. Bocchini, Francois Schiettecatte, Alan F. Scott, and Ada Hamosh. *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders*. Nucleic Acids Research 2015; 43 (Database issue) D789–D798.

[5] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O.B. Jacobsen, Daniel Danis, Jean Philippe Gourdine, Michael Gargano, Nomi L. Harris, Nicolas Matentzoglu, Julie A. McMurry, David Osumi-Sutherland, Valentina Cipriani, James P. Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C. Jansen, Ahmed Muaz, Willie H. Chang, Jenna Bergerson, Stanley J.F. Laulederkind, et al. *Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources*. Nucleic Acids Research 2019; 47 (D1) D1018–D1027.

[6] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, et al. *Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes*. bioRxiv 2019; 531210.

[7] R.G.H. Cotton, V. McKusick, and C.R. Scriver. *The HUGO Mutation Database Initiative*. Science 1998; 279 (5347) 10–15.

[8] Mauno Vihinen, Johan T. den Dunnen, Raymond Dalgleish, and Richard G.H. Cotton. *Guidelines for establishing locus-specific databases*. Human Mutation 2012; 33 (2) 298–305.

[9] Ivo F.A.C. Fokkema, Peter E.M. Taschner, Gerard C.P. Schaafsma, J. Celli, Jeroen F.J. Laros, and Johan T. den Dunnen. *LOVD v.2.0: the next generation in gene variant databases*. Human Mutation 2011; 32 (5) 557–563.

[10] Christophe Béroud, Dalil Hamroun, Gwenaëlle Collod-Béroud, Catherine Boileau, Thierry Soussi, and Mireille Claustres. *UMD (Universal Mutation Database): 2005 update*. Human Mutation 2005; 26 (3) 184–191.

[11] Peter D. Stenson, Matthew Mort, Edward V. Ball, Katy Evans, Matthew Hayden, Sally Heywood, Michelle Hussain, Andrew D. Phillips, and David N. Cooper. *The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies*. Human Genetics 2017; 136 (6) 665–677.

[12] Melissa J. Landrum and Brandi L. Kattman. *ClinVar at five years: Delivering on the promise*. Human Mutation 2018; 39 (11) 1623–1630.

[13] Ivo F.A.C. Fokkema, Kasper J. van der Velde, Mariska K. Slofstra, Claudia A.L. Ruivenkamp, Maartje J. Vogel, Rolph Pfundt, Marinus J. Blok, Ronald H. Lekanne Deprez, Quinten Waisfisz, Kristin M. Abbott, Richard J. Sinke, Rubayte Rahman, Isaäc J. Nijman, Bart de Koning, Gert Thijs, Nienke Wieskamp, Ruben J.G. Moritz, Bart Charbon, Jasper J. Saris, Johan T. den Dunnen, Jeroen F.J. Laros, Morris A. Swertz, and Marielle E. van Gijn. *Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data*. Human Mutation 2019; 40 (12) 2230–2238.

[14] Dong Xu. *Protein databases on the internet*. Current Protocols in Protein Science 2012; 1 (SUPPL.70).

[15] Johan T. den Dunnen. *Efficient variant data preparation for Human Mutation manuscripts: Variants and phenotypes*. Human Mutation 2019; 40 (8) 1009.

[16] Johan T. den Dunnen. *Yet another database?* Human Mutation 2018; 39 (6) 755.