LSDBs and how they have evolved

Raymond Dalgleish, Department of Genetics, University of Leicester, Leicester, United Kingdom

Keywords: locus specific databases, LSDBs, reference sequences, sequence variants, software, data-quality assurance, variant nomenclature

Contact:

Raymond Dalgleish

Department of Genetics,

University of Leicester

University Road

Leicester

LE1 7RH

United Kingdom

Tel: +44 116 252 3425

Email: raymond.dalgleish@le.ac.uk

Abstract:

Locus specific databases (LSDBs) make a key contribution to our understanding of heritable and acquired human disorders, disease susceptibility, and adverse drug reactions. As data have accumulated in LSDBs, a greater reliance on their use has arisen in clinical practice. Even though LSDBs have existed in recognizable form for only a quarter of a century, their origin lies in the manual cataloging of data that began around 50 years ago. Analysis and recording of sequence variation in the globin genes, and the proteins which they encode, can confidently be said to be the foundation for what we now refer to as LSDBs. Their growth over the years has primarily been underpinned by software developments and the advent of the World Wide Web. However, it is also important to recognize the evolution of reporting standards and reference sequences, without which accurate and consistent reporting of sequence variants would be impossible. Nowadays, LSDBs exist for many human protein-coding genes and the focus of efforts has moved towards minor tidying up of the variant reporting nomenclature and processes for assuring the completeness, correctness and consistency of the data. The next twenty five years will doubtless witness further developments in the evolution of LSDBs.

Introduction:

The term locus specific database (LSDB) conjures the mental image of a computer-based database which provides public remote access to DNA variant data, usually by way of a web browser. More correctly, LSDBs should perhaps be referred to as genotype-phenotype databases as that is what they have become as more data have been accumulated that shed light on the relationship between underlying gene mutation events and the resulting disease manifestations. However, the first databases did not record DNA sequence variation. Arguably, LSDBs existed prior to the advent of DNA sequencing and, when they first emerged, they were repositories of protein sequence variation. With the advent of gene cloning and sequencing in the 1970s, and the rapid advances in sequence analysis enabled by the development of the polymerase chain reaction in the 1980s, the accumulation and cataloguing of gene variant data has become an essential task which underpins health care for both inherited and acquired conditions. The purpose of this review is to provide a historical perspective with respect to the evolution of LSDBs, rather than to provide a comprehensive review of the current state of genotype-phenotype databases in the next-generation sequencing and omics era. That subject has been thoroughly reviewed recently by others [Brookes and Robinson, 2015; Johnston and Biesecker, 2013].

The hemoglobinopathies: a model example for LSDBs

Before reviewing the individual developments, such as variant nomenclature and reference sequences, which have underpinned the evolution of LSDBs, it is worth considering the example of the timeline for the protein and gene system which arguably serves as the best model for all subsequent LSDBs: the hemoglobinopathies. The first well characterized protein system with respect to sequence variation was hemoglobin, the tetrameric heme-

containing molecule, comprising two alpha-like and two beta-like protein chains which transports oxygen in red blood cells. The first recognized hemoglobinopathy was sickle cell anemia which was described in 1910 by James Herrick. Its mode of inheritance was subsequently shown to be recessive by Taliaferro and Huck [1923], and its basis in terms of a protein charge difference was eventually elucidated by Pauling et al. [1949]: it was in this latter study that the term "molecular disease" was first coined. By the mid-1950s, Vernon Ingram had demonstrated the molecular basis of sickle cell disease by peptide mapping and protein sequencing, revealing a single amino acid substitution (glutamic acid to valine) at the sixth position in the mature adult beta globin chain [Ingram, 1956; Ingram, 1957]. By the mid-1960s, the genetics of the thalassemias were broadly understood though several gaps remained, especially with respect to the linkage arrangement of the globin structural loci [Rucknagel, 1964]. Several more globin variants were characterized in the following years and by 1967 there were around 30 known alpha-chain variants, around 50 known beta-chain variants and "other" globin variants numbered around 15 [Livingstone, 1967]. Some variants were characterized simultaneously in several laboratories leading to the same variant being known by two or more names. For example, the beta-chain glycine to aspartic acid substitution at position 16 was variously designated as hemoglobins J Baltimore, J Ireland, J Trinidad and N New Haven, with J Baltimore now being the conventional designation.

A key contemporary repository of globin variant data was the series of twelve editions of Victor McKusick's *Mendelian Inheritance in Man* which was first published in 1966. The successor web site, Online Mendelian Inheritance in Man (OMIM, http://www.omim.org/) continues to comprehensively catalog variant and phenotype information for the globin

genes, as well as for all other human genes and phenotypes [Amberger et al., 2015]. By the mid-1970s some 200 variants were known and were published in textbooks by Lehmann and Kynoch [1976] and by others the following year [Bunn et al., 1977a; Bunn et al., 1977b]. Although these books provided authoritative accounts of variants in each of the known globin chains which hemoglobins comprise, the presentation of variant data in textbooks tended to be narrative, rather than tabular. The honorable exception to this was the compilation by Lehmann and Kynoch [1976] which made an early attempt at a tabular presentational format. Over the years, additional variants were discovered and the need to record them was addressed by a series of reference works, sponsored by The Sickle Cell Anemia Foundation, and compiled by Titus Huisman and colleagues who systematically catalogued all known hemoglobin and thalassemia variants in printed tabular form [Huisman et al., 1997; Huisman et al., 1996; Huisman et al., 1998]. Although useful, these books did not provide a means of disseminating the data widely. The solution to that lay initially in careful reproduction of the data from the books in the form of web pages [Hardison et al., 1998]. However, the data could only be viewed by browsing, or by simple text searches. Subsequently, the data were transferred to an Oracle relational database, HbVar, [Hardison et al., 2002] hosted on the Globin Gene Server (http://globin.bx.psu.edu/). The server has been updated for more than a decade [Giardine et al., 2014; Giardine et al., 2007; Patrinos et al., 2004] and is now a knowledgebase hosting structural and expression data, in addition to sequence variants. A key contribution of the HbVar project to the field of variant databasing has been the submission of HbVar data to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). This has ensured that the location of globin gene variants is maintained in an archival database which can provide links back to the originating LSDB.

Although HbVar is the longest established globin gene database, it has recently been joined by the IthaGenes database (http://www.ithanet.eu/db/ithagenes) [Kountouris et al., 2014] which claims to provide an improved user interface relative to HbVar. The establishment of HbVar and IthaGenes as reliable sources of variant data for the human globin genes has depended on many parallel developments in the field of biological informatics. These developments and the policy decisions underpinning some of them are described in the following sections.

Describing sequence variants

Until the early 1990s, variants in proteins and in genes were not described systematically using an unambiguous nomenclature. It was not uncommon for variants to be known by a nickname based on the discoverer, the patient in which it was first found, or by the geographic location of its discovery. The various deficiencies of descriptions based on amino acids changes were described in 1993 by Ernest Beutler who recommended instead that variants be defined either in terms of a genomic DNA or a cDNA sequence change, with the latter providing the best compromise between the minor deficiencies of both. Notably however, Beutler refrained from actually making specific proposals for a precise nomenclature to describe DNA sequence base changes and this task was subsequently left to others [Beaudet and Tsui, 1993]. Although these proposals began to establish some order to variant descriptions, certain aspects remained very much at the discretion of those with a research interest in specific genes or proteins. Although they recommended that amino acids be numbered from the initiator codon, they acknowledged that this was problematic unless based upon cDNA sequence data for mRNAs with a single open reading frame. Many proteins are synthesized with a leader peptide and several undergo post-translational

cleavage to produce the mature protein. Even for proteins that do not undergo post-translational cleavage, the initiating methionine is commonly hydrolyzed, resulting in the mature protein being one amino acid shorter than the primary translation product [Giglione et al., 2004]. The consequence for beta globin is that the amino acid which is substituted by valine in sickle cell disease is the glutamic acid at position 7 of the primary translation product, rather than position 6 which is referred to in the commonly used description of this variant: “Glu6Val”. Pragmatically the authors recommended that where well established, but non-standard, amino acid numbering systems already existed “*The rule should be to utilize whatever is well established and conventional for the gene product in question.*” The introduction of a uniform numbering system for amino acids would have to wait.

The equally lengthy evolution of robust and sensible numbering schemes for DNA and RNA sequences was no less tortuous than that for proteins and amino acids. The HGVS nomenclature recommendations [den Dunnen and Antonarakis, 2000; Taschner and den Dunnen, 2011] continue to evolve and current developments aim to ensure that the nomenclature is capable of accurately describing large-scale cytogenetic genetic changes, including those characterized by DNA sequencing [ISCN, 2013]. Proposed changes to the nomenclature are reviewed by a Sequence Variant Description Working Group (SVD-WG) operating under the auspices of the Human Genome Variation Society (HGVS), the Human Variome Project (HVP), and the Human Genome Organization (HUGO) [den Dunnen et al., 2016].

Over the years, there has been a concerted effort to eliminate the use of “common” or trivial names for sequence variants. However, the use of traditional non-standard

nomenclature has persisted for many genes and diseases. A review of mutation nomenclature for the *CFTR* gene recommended in 2011 that variant reports should "*…include a description of the identified sequence variants in both HGVS and traditional nomenclature…*" [Berwouts et al., 2011]. In spite of the recommendations, reports still appear in the literature with variants described using only traditional nomenclature [Graeber et al., 2015; Pesci et al., 2015]. Database curators need to remain vigilant to the use of historical or legacy variant names in publications, and to simply accept that journal editors and reviewers remain largely immune to pleas from organizations, such as the Human Variome Project, that HGVS nomenclature be used as the primary means of describing variants. However, there is greater hope for the elimination of non-HGVS variant descriptions from clinical laboratory reports. Most clinical genetics testing laboratories are subject to external quality assessment (EQA) and many EQA providers exist across the globe. The results of a recent assessment of genetic reports from laboratories participating in four European EQA schemes indicate steady improvement in compliance with respect to standardized nomenclatures [Tack et al., 2016]. The key issue with respect to continued improvement is that EQA providers have the ability to rescind accreditation to laboratories who repeatedly flout their guidance concerning reporting standards.

Two decades ago, a proposal was made that all variants be assigned a unique identifier [Beutler et al., 1996]. At the time, the proposal was intended to mitigate problems associated with the use of nicknames, or trivial names, for variants, and the use of OMIM identifiers was the then favored scheme. This suggestion was never fully embraced but several variants are nowadays referenced by their identifiers in the NCBI dbSNP and ClinVar databases. Beyond that, no standard or mandatory system of variant identifiers has ever

been established. However, there remains a case for universal identifiers for variant alleles, and perhaps even for haplotypes for which disease-causation has been established. It ought to be possible, through cooperation with global resources such as NCBI and Global Alliance for Genomics & Health (GA4GH: https://genomicsandhealth.org/) to establish robust identifiers (as distinct from HGVS variant descriptions) for variant alleles that are not tied to specific genome builds or reference sequences.

Reference sequences

The concept of a reference sequence for the reporting of DNA and protein variants is nowadays taken for granted even though it was not formally proposed as a desirable or necessary component of reporting until 1996 [Beutler et al., 1996]. In the mid-1990s, comprehensive and complete sequence data existed for relatively few human genes and the quality of the data was not always as reliable as modern-day standards demand. In the early days of gene cloning, the only medium through which sequence data were shared was by publication in journals. Data were often limited and incomplete because the primary driver for their publication was simply the need to support assertions about the identity of cloned DNA sequences, rather than to provide extensive sequence information. There was certainly no established practice until well into the 1980s that published sequence data should also be submitted to public databases. Initially, submissions went to the Los Alamos Sequence Database (nowadays the GenBank database) and the EMBL Nucleotide Sequence Data Library (nowadays the European Nucleotide Archive), which were created in 1979 and 1980 respectively. The DNA Data Bank of Japan was created in 1986 and all three databases, which share sequence data with one another, are members of the International Nucleotide Sequence Database Collaboration (INSDC).

The cloning and partial sequencing of human globin mRNAs was first reported in 1974 [Marotta et al., 1974] with several additional related reports following in subsequent years. In some instances, individual published sequences did find their way eventually into individual database records but the fate for the majority was that they were aggregated much later into genomic DNA database records compiled from several sources. For example, the GenBank record J00179 (Human beta globin region on chromosome 11), which was compiled in 1993, comprises data taken from 96 individual publications.

The need at that time for this piecemeal assembly of gene sequences was relatively common especially for large genes. For some, only incomplete transcript and genes sequences existed initially, in spite of advances in DNA sequencing, and this presented an obstacle for the consistent reporting of sequence variants. However, databases had inconsistent policies with respect to submission of aggregated data. While GenBank appeared to embrace the practical imperative to build reference sequence contigs from overlapping incomplete sequences, the EMBL Nucleotide Sequence Data Library appeared to be not so keen. The need arose to have full-length reference sequences for the transcripts of the *COL1A1* and *COL1A2* genes for the reporting of variants giving rise to the heritable bone disorder osteogenesis imperfecta [Dalgleish, 1997]. Initially, EMBL announced that their policy was to accept only novel primary data and that derivative, or aggregated data, were not acceptable for submission. Fortunately they eventually relented and the accessions Z74615 and Z74616 were assigned respectively to the transcripts of the two genes.

By the 1990s many leading journals had adopted a policy requiring authors to submit their sequence data into databases as a condition of acceptance of manuscripts for publication. However, databases still relied at this time on the scanning of journals and of patent applications for sequence data [Benson et al., 1993; Emmert et al., 1994]. In 1993, 15% of new entries in GenBank were a consequence of journal scanning and the capture of protein and DNA sequence data from 1960 to 1993, previously unavailable electronically, was facilitated by collaboration between the European Patent Office and the European Bioinformatics Institute.

A side effect of journals' requirements to submit sequence data to databases was that the databases effectively became repositories for disease-causing sequence variants, in the form of variant DNA sequence entries, sometimes in parallel with the recording of these same variants in LSDBs. Analysis of four families with dominant forms of beta thalassemia [Thein et al., 1990] revealed two different sequence variants in exon three of the *HBB* gene that resulted either in extended or truncated beta-globin chains. DNA sequences from these patients are available from GenBank as records M34058.1, M34059.1 and AH001475.1. Thankfully, the unnecessary inflation of non-redundant DNA databases with data of this type is a practice that is no longer continued.

The next major advance with respect to improved reference sequences was the establishment by NCBI of the RefSeq Reference Sequence Database about 15 years ago [O'Leary et al., 2016]. RefSeq curates reference-standard sequences for chromosomes, transcripts (including non-protein-coding transcripts) and proteins based on careful review of all public sequence data, including RNA-Seq data, and in collaboration with other groups

via the LRG project [Dalgleish et al., 2010; MacArthur et al., 2014] and the Consensus CDS (CCDS) project [Harte et al., 2012]. It is through the efforts of the CCDS that reliable evidence exists for the many alternative gene transcripts for which RefSeq reference sequences have been created. In addition, RefSeqGene reference-standard sequences have been created for nearly 5600 gene-specific genome regions which are annotated with a sub-set of RefSeq transcripts and proteins. A current focus of RefSeq is the creation of RefSeqGene sequences for all genes for which clinical tests are commonly provided [O'Leary et al., 2016]. Consequently, RefSeq and RefSeqGene sequences provide the most comprehensive and reliable basis for the reporting of sequence variants.

However, two acknowledged features of RefSeq sequence records present minor obstacles for accurate and consistent reporting of variants. The first of these is that sequence-record accession numbers are versioned when any change to the reference sequence is made. For example, the full accession for the human beta globin mRNA reference sequence is NM_000518.4 with the “4” indicating version 4 of the primary sequence. This current version is exactly the same length as version 3 (dated 25 April 2001) but differs at nucleotide position 59 with a C at that position in version 3 and a T in version 4. This changes the codon for the third amino acid of the primary translation product from CAC to CAT, illustrating the need to explicitly specify sequence version numbers when reporting sequence variants. Unfortunately, authors of papers or clinical-testing reports may not be diligent in reporting the version number of the sequence used to report a sequence variant, with the potential for misreporting of variants which might be catastrophic in a clinical diagnostic setting. An associated issue is that particular bases in RefSeqGene and RefSeq transcript records might correspond with one another, but differ from the equivalent base in the RefSeq genome

record. This can arise when a particular variant allele is requested, through stakeholder consultation, in a RefSeq transcript because it represents the standard allele identified by alignment of public cDNA sequences.

The second limitation of RefSeq reference sequences is the need on the part of the user to assemble a complete set of sequence records for the gene in question, its transcripts and the proteins encoded by them. For the *TP53* gene, that comprises one record for the gene, eight for its transcripts and thirteen for the proteins making twenty-two in total. This problem, and that of sequence versioning, was solved by the introduction of Locus Reference Genomic (LRG) sequence records [Dalgleish et al., 2010; MacArthur et al., 2014] which are maintained through collaboration between EMBL-EBI and the RefSeqGene group at NCBI (http://www.lrg-sequence.org/). LRG records comprise the relevant genomic DNA, transcript and protein sequences and each is permanently stable with no versioning. This ensures that sequence variants reported with respect to LRGs remain authoritative even if the underlying genome sequence is ever revised. LRG sequences are recommended in the HGVS nomenclature guidelines for the reporting of variants and LRGs now exist for more than 650 genes of clinical interest. Even though this number falls far short of the nearly 5600 RefSeqGene sequences, the number of LRGs is expected to increase in response to requests from the research and diagnostic community for their creation. A further 218 LRGs are currently being compiled.

If RefSeq gene and transcript reference sequences are cited correctly by accession number and version, or LRGs are used instead, no issues should arise with respect to accurate variant reporting. However, the shift to variant detection through whole-genome or exome

sequencing brings the complication of variants now being reported primarily in the context of genome coordinates. Thankfully, robust tools, such as the NCBI Genome Remapping Service (http://www.ncbi.nlm.nih.gov/genome/tools/remap), provide convenient transformation of genome-build coordinates to RefSeqGene and LRG positions if such records exist for the genome region in question.

In spite of the availability of high-quality reference sequences for variant reporting, historic/legacy and "non-standard" sequences are sometimes used. This has the potential to complicate the validation of variant descriptions, and it is not confined to historic published accounts from the era before RefSeq and the advent of LRGs. For example, a relatively recently published account of *TNXB* gene variants uses the Ensembl transcript ENST00000375244 as the reference sequence [Gbadegesin et al., 2013]. The variant descriptions are reported correctly, but the use of a rare transcript as the reference sequence necessitates adjustment of some variant descriptions into the context of the RefSeq transcript NM_019105.6 for inclusion into the *TNXB* LSDB. Notably, there is no RefSeq transcript which corresponds with ENST00000375244, and this highlights the need for database curators to remain fully familiar with the full gamut of reference sequences.

Software solutions

Database software

The first major practical improvement upon the provision of human hemoglobin variant data in purely printed form was the reproduction of these same data in the form of web pages in the late-1990s on a site named the Globin Gene Server [Hardison et al., 1998]. Although the data provide by the server were comprehensive, they simply comprised a set

of hyperlinked web pages and a rudimentary query interface allowing the user to search on keywords. Interestingly, this forerunner of HbVar was not the first attempt to computerize the storage and retrieval of human hemoglobin variant data. Nearly ten years earlier, a FORTRAN 77 computer system, VARIANT, was written for the storage and retrieval of human hemoglobin variants [Macchiato and Tramontano, 1990]. However, in keeping with the era, VARIANT only had a command-line interface (CLI) requiring the user to type in the various storage and retrieval commands, rather than through the use of a web-page interface of the type which would not become commonplace for quite some time. VARIANT appears to have been written in a way that could have allowed it to be used for variants of other proteins, but its focus on amino acids variants limited its usefulness and there is no record in the literature of it having been used to record amino acid variants in any other protein. The other limitation, of course, was that VARIANT was a single-user system which provided no mechanism to share data publicly.

Subsequent development of LSDBs took one of two approaches. In some cases, bespoke web sites were created which were carefully tailored to present protein-, gene- or disease-specific data that might not be generally applicable to other genes or diseases. Others took the approach that a generic database system could provide for the need to report sequence variants for diverse genes and diseases. The former approach has resulted in several well respected LSDBs that might, in some instances, be considered nowadays to be knowledge-bases. These include, for example, the Clinical and Functional Translation of CFTR (CFTR2) site (http://www.cftr2.org/) for cystic fibrosis [Castellani, 2013], HbVAR site for human hemoglobin variants and thalassemias (http://globin.bx.psu.edu/hbvar/) [Patrinos et al., 2004], the Blood Group Antigen Gene Mutation Database

(http://www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.cgi?cmd=bgmut/home) [Patnaik et al., 2012], and the Alzheimer Disease & Frontotemporal Dementia Mutation Database (http://www.molgen.ua.ac.be/ADMutations/) [Cruts et al., 2012].

The generic database approach has been supported over the years by a series of "LSDB-in-a-box" solutions designed to perform the task of database creation and have included MUTbase [Riikonen and Vihinen, 1999], MuStaR [Brown and McKie, 2000], UMD [Béroud et al., 2005],and LOVD [Fokkema et al., 2011]. Of these, UMD arguably provides the richest set of curation and variant display tools. However, UMD has not proved to be a particularly popular solution to providing online LSDBs (http://www.umd.be/) and the software appears to be no longer available to download and install locally. It is LOVD (http://www.lovd.nl/) which is by far the most extensively adopted LSDB-in-a-box solution with the third iteration of the open-source software being released in December 2012. Although LOVD 3.0 is primarily a tool for gene-centric collection and display of DNA variants, it extends this idea by also providing storage for patient-centric data and NGS data, even of variants that lie outside of genes. A particularly attractive feature of LOVD is that its creators have established a database for most human protein-coding genes on their own servers (http://databases.lovd.nl/whole_genome/genes) and have invited interested parties to assume responsibility for maintaining databases for one or more genes of interest. This relieves prospective data curators of the tasks of providing servers to host the database and perform software maintenance tasks themselves, leaving them free to concentrate on the data. However, many of the genes in the LOVD "whole-genome" database have neither content nor curators. This issue could be mitigated in part if data could be migrated easily from existing curated LOVD2 databases to the "whole-genome" database. Even if there was

the will on the part of curators to migrate their existing data, no fully automated process exists to facilitate the transfer. At least in part, this has resulted in variant data for more than two thousand genes still being hosted using LOVD 2.0. A further issue is that the LOVD 3.0 "whole-genome" database is probably incomplete in terms of gene coverage. The HUGO Gene Nomenclature Committee (HUGO: http://www.genenames.org/) currently recognizes 19,001 protein-coding genes and 6,029 genes for non-coding RNAs. This makes a total of 25,030 which is considerably greater than the 22,002 genes in the "whole-genome" database. At present, disease-associated variants in non-coding RNA genes tend to be recorded mostly in databases dedicated to particular classes of RNA types [Bhattacharya and Cui, 2016; Zhang and Lupski, 2015; Zhao et al., 2016], rather than in gene-specific LSDBs of the type commonly established for protein-coding genes.

Finally, the MOLGENIS software framework [Swertz et al., 2010] offers an intermediate solution between fully bespoke and LSDB-in-a-box solutions by providing a method for the rapid development of LSDBs, among its other capabilities. An excellent example of this approach is the International Dystrophic Epidermolysis Bullosa Patient Registry which holds variant data for the *COL7A1* gene [van den Akker et al., 2011].

Variant nomenclature validation

The HGVS nomenclature standard for variant descriptions has evolved enormously over the years and has consequently become complex. This has resulted in frequent errors in variant descriptions in the literature and in clinical reports. Robust curation of LSDBs demands that variant reporting be accurate and that incorrect variant descriptions reports be fixed. This is especially important in the context of clinical reports where misreporting might result in

disastrous outcomes for patients [Ogino et al., 2007; Richards et al., 2015; Vihinen, 2015]. Mutalyzer (https://mutalyzer.nl/) [den Dunnen et al., 2016; Wildeman et al., 2008] is a program suite which supports the checking of sequence variants according to the HGVS nomenclature. More recently, the alternative "hgvs" validation package (https://bitbucket.org/biocommons/hgvs/) [Hart et al., 2015] has become available and has some useful features not implemented in Mutalyzer. These include pre-computed alignments of gene transcripts to the current genome assembly (https://bitbucket.org/biocommons/uta) which allows for rapid and comprehensive lift-over of variants from one gene transcript to another. This feature allows curators to quickly verify that a sequence variant described in the context of one particular transcript reference sequence is valid in the context of the reference sequence normally used for variant reporting in a given LSDB. To make hgvs more generally useful, Variant Validator (https://variantvalidator.org/) has been implemented to provide a web interface to the more useful validation functions provided by hgvs.

Variant effect prediction

The ability to detect putative disease-causing variants has advanced enormously with successive developments in massively parallel sequencing. Variants which survive the various filtering steps then need to be assessed with respect to their potential to be deleterious. This is important with respect to establishing a plausible disease-causation link in the first instance, but well-curated LSDBs might also want to verify the original pathogenicity assertions.

Predictive tools for the analysis of missense variants and RNA splicing are especially commonly used, and the resulting analyses are widely reported in the literature. However there is frequently a lack of clarity with the respect to the tools used, even if these have been previously published. The potential for misleading analyses of newly detected variants being propagated into LSDBs can be mitigated by the application of a simple set of criteria [Vihinen, 2013] with the key point perhaps being "*Before using prediction results, understand the principle of the method, its use, limitations, and applications.*" A comprehensive review of the available variant prediction tools is beyond the scope of this review and readers are instead directed to recent reviews [Niroula and Vihinen, 2016; Ritchie and Flicek, 2014] which highlight the functions provided by available tools and provide guidance on the choice of optimal solutions.

Discussion

The case for the creation and maintenance of LSDBs has been made by many commentators over the years. The earliest overt promotion of LSDBs is probably that made in the HUGO newsletter Genome Digest in the mid-1990s [Cotton et al., 1996]. By the start of this century, the case for LSDBs was well accepted and the focus of database advocates had begun to move towards ensuring that LSDBs complied with minimum standards with respect to the data held and how these data were presented [Claustres et al., 2002; Cotton and Horaitis, 2002; Horaitis and Cotton, 2004]. Subsequently there have been additional surveys and analyses of LSDBs, each attempting to assess the completeness, correctness and consistency of their provision of data, mostly in a general fashion [George et al., 2008a; Mitropoulou et al., 2010], but also with respect to specific diseases or genes [Savige et al., 2015; Soussi, 2014; Vail et al., 2015]. A particularly difficult issue for LSDBs has been the

ethical concerns associated with the inclusion of patient data that might have the potential to identify particular patients through the inclusion of even minimal data (e.g. a variant description and a phenotype) in an LSDB. That is perhaps an extreme viewpoint, but patient privacy has been a frequent subject of discussion and debate [Cotton et al., 2005; Povey et al., 2010; Vihinen et al., 2012] and guidance is provided in the literature for those contemplating the setting of a new LSDB or taking on the governance of one that already exists. The task of setting up a new LSDB or adopting one that has not recently been maintained can be daunting, but extensive advice is available for prospective curators [Celli et al., 2012; Vihinen et al., 2012].

Becoming a variant database curator is not an activity to be taken lightly, and due consideration needs to be given to the amount of time that the prospective curator can reasonably spend on the task. In part, the size of the task is determined by the number of genes for which variants will be collated and entered into an LSDB. For some heritable disorders, the number genes known to harbor disease-causing variants might initially be small, but will probably swell as understanding of disease etiology improves. An additional determinant of task size is the prevalence of the disorder, as this is likely to be a major factor in determining the effort required: rare disorders normally result in relatively few variants reports, either in the literature or directly from diagnostic laboratories. Whatever the size of the task, it is the quality of the data aggregation and evaluation that are critical if the resulting database is to be of value to users, and there is no substitute for expert curation. Some might argue that the evaluation of data from the literature can be reduced to a few simple rules and, in most cases, there might be an element of truth in that. However, it is through expert knowledge of disease etiology and accumulated wisdom that

deficiencies in published reports become apparent [Cotton and Scriver, 1998; Dalgleish, 2011]. Individuals and organizations have attempted to better define and codify the process of database creation and curation through reviews and analyses of locus-specific [Cotton et al., 2007; Mitropoulou et al., 2010] and general variant databases, but sometimes not without controversy [George et al., 2008a; George et al., 2008b; Stenson et al., 2008].

Compared with analyses of the data that may be found in LSDBs, the fundamental issue of how data elements are identified and evaluated for inclusion is the subject of much less attention in the literature. In most instances, newly published variants can be easily identified through the use of key-word alerting services such as those provided by PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Web of Science (http://webofscience.com/). Both services will provide alerts to publications in all languages, but there might be a tendency on the part of some curators to ignore foreign-language publications because of difficulty of access to reprints, and the problems of translation. Chinese-language publications are becoming notably more frequent in the field of osteogenesis imperfecta genetics and cannot be ignored, especially in instances where a publication might report variant data for as many as 200 cases [Zhao et al., 2015].

Although the case in favor of traditional LSDBs appears to be well made [Samuels and Rouleau, 2011], others have not necessarily disagreed, but have argued persuasively that true progress in the cataloguing of genetic variants will only come through international initiatives supported by substantial funding [Auerbach et al., 2011; Ayme et al., 2011]. An attendant issue is whether variant data should be aggregated into central databases or continue to be stored in individual LSDBs. Databases such as ClinVar

(http://www.ncbi.nlm.nih.gov/clinvar/) [Landrum et al., 2016], funded by the US Government, and the Human Gene Mutation Database (HGMD: http://www.hgmd.cf.ac.uk/) [Stenson et al., 2014], funded by commercial subscription, represent the current state of the art in centralized variant databases, however each has its limitations. Neither is ever likely to be entirely up-to-date with new variant data when compared with the best curated LSDBs. Several variant descriptions in ClinVar are not compliant with the HGVS nomenclature and the data are often incomplete. At the time of writing, there are 890 unique sequence variants recorded in the Osteogenesis Imperfecta Variant Database [Dalgleish, 1997; Dalgleish, 1998] for the *COL1A1* gene (version COL1A1 151214), but only 132 in ClinVar (version 20160202) for the same gene. It is necessary to understand, however, that ClinVar is entirely submitter-driven and is not equipped to review the literature and to build content via curation. Content will only accumulate when authors of papers and LSDB curators can be encouraged to submit their data. For HGMD, access to the most up-to-date variant data is by subscription only, even for academic use, and the freely accessible data are always three years out of date by design. In spite of these limitations, both ClinVar and HGMD are valuable resources, providing useful data that often complement those available from primary LSDBs. In the case of ClinVar, there is a clear opportunity to improve its content by fostering partnerships with LSDB curators, especially those of large resources such as InSiGHT (http://insight-group.org/variants/database/), RETTBase (http://mecp2.chw.edu.au/), BIC (http://research.nhgri.nih.gov/bic/) and ENIGMA (http://enigmaconsortium.org/).

The apparently unsolvable issue for variant databases is the seemingly constant duplication of data in multiple and diverse database locations. A search of the Locus Specific Database

list at Leiden University Medical Center (http://grenada.lumc.nl/LSDB_list/lsdbs) reveals, at the time of writing, that there are seven known databases for the *HBB* globin gene and six for the *COL1A1* collagen gene. This proliferation of databases is specifically discouraged [Celli et al., 2012; Vihinen et al., 2012] but the key underlying issue of reward and recognition needs to be properly addressed if there is to be any halt. Most LSDB curation is carried out without direct funding and is performed as a non-paid activity in addition to curators' primary employment responsibilities. Consequently, they lack the time to curate the multiple duplicate databases and feel that they will receive no reward or recognition for donating their curated data to large integrated databases such as ClinVar or by providing unrewarded curation services to help aggregate the existing data.

Occasionally, existing variant databases fail to meet the practical needs with respect to some diseases. Heritable disorders can be categorized as monogenic or polygenic, with the latter frequently involving environmental contributions to their causation. The simplest category of polygenic disease, and easiest to understand at the gene level, comprises those with digenic inheritance [Schäffer, 2013], which were first recognized in 1994. However, existing variant databases do not provide the ability to easily retrieve detailed data regarding digenic combinations of variants and so the Digenic Diseases Database (DIDA: http://dida.ibsquare.be/) [Gazzo et al., 2016] has recently been created to address the previously unmet needs of researchers in the field.

Accounts of the days when some LSDBs comprised word-processed documents, which were printed and posted to recipients, prior to the widespread availability of the internet must, nowadays, seem like apocryphal tales from a long-distant era. However, that was only

twenty five years ago: around the time that Human Mutation was established as a journal. The next twenty five years will doubtless bear witness to some additional dramatic developments.

Acknowledgements

I am grateful to Prof Sir David Weatherall for helpful discussions and guidance about the early days of globin and thalassemia genetics. I am also grateful to Dick Cotton for his unstinting support and especially for his boundless energy in promoting the growth of LSDBs.

Conflict of Interest

The author has no conflict of interest to declare.

References

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res 43:D789-D798.

Auerbach AD, Burn J, Cassiman JJ, Claustres M, Cotton RGH, Cutting G, den Dunnen JT, El-Ruby M, Vargas AF, Greenblatt MS, Macrae F, Matsubara Y et al. 2011. Mutation (variation) databases and registries: a rationale for coordination of efforts. Nat Rev Genet 12:2.

Ayme S, Terry SF, Groft S. 2011. Response to 'Mutation (variation) databases and registries: a rationale for coordination of efforts': an IRDiRC perspective. Nat Rev Genet 12:1.

Beaudet AL, Tsui LC. 1993. A suggested nomenclature for designating mutations. Hum Mutat 2:245-248.

Benson D, Lipman DJ, Ostell J. 1993. GenBank. Nucleic Acids Res 21:2963-2965.

Béroud C, Hamroun D, Collod-Béroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26:184-191.

Berwouts S, Morris MA, Girodon E, Schwarz M, Stuhrmann M, Dequeker E. 2011. Mutation nomenclature in practice: findings and recommendations from the cystic fibrosis external quality assessment scheme. Hum Mutat 32:1197-1203.

Beutler E. 1993. The designation of mutations. Am J Hum Genet 53:783-785.

Beutler E, McKusick VA, Motulsky AG, Scriver CR, Hutchinson F. 1996. Mutation nomenclature: nicknames, systematic names, and unique identifiers. Hum Mutat 8:203-206.

Bhattacharya A, Cui Y. 2016. SomamiR 2.0: a database of cancer somatic mutations altering microRNA-ceRNA interactions. Nucleic Acids Res 44:D1005-D1010.

Brookes AJ, Robinson PN. 2015. Human genotype–phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16:702-715.

Brown AF, McKie MA. 2000. MuStaR and other software for locus-specific mutation databases. Hum Mutat 15:76-85.

Bunn HF, Forget BG, Ranney HM. 1977a. Human Hemoglobins. Philadelphia: WB Saunders Company. 432 p.

Bunn HF, G. FB, Ranney HM. 1977b. Hemoglobinopathies. Philadelphia: WB Saunders Company. 308 p.

Castellani C. 2013. CFTR2: How will it help care? Paediatr Respir Rev 14 Suppl 1:2-5.

Celli J, Dalgleish R, Vihinen M, Taschner PEM, den Dunnen JT. 2012. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 33:291-297.

Claustres M, Horaitis O, Vanevski M, Cotton RGH. 2002. Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases. Genome Res 12:680-688.

Cotton RGH, Horaitis O. 2002. The HUGO Mutation Database Initiative. Pharmacogenomics J 2:16-19.

Cotton RGH, Phillips K, Horaitis O. 2007. A survey of locus-specific database curation. Human Genome Variation Society. J Med Genet 44:e72.

Cotton RGH, Sallée C, Knoppers BM. 2005. Locus-specific databases: from ethical principles to practice. Hum Mutat 26:489-493.

Cotton RGH, Scriver CR. 1998. Proof of "disease causing" mutation. Hum Mutat 12:1-3.

Cotton RGH, Scriver CR, McKusick VA. 1996. Locus-specific mutation databases: a resource. Genome Digest January:6-10.

Cruts M, Theuns J, Van Broeckhoven C. 2012. Locus-specific mutation databases for neurodegenerative brain diseases. Hum Mutat 33:1340-1344.

Dalgleish R. 1997. The human type I collagen mutation database. Nucleic Acids Res 25:181-187.

Dalgleish R. 1998. The human collagen mutation database 1998. Nucleic Acids Res 26:253-255.

Dalgleish R. 2011. Boning up on mutations: assessing the significance of candidate disease-causing DNA sequence variation. Genet Mol Res 10:1518-1521.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Béroud C, Dobson G et al. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2:24.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. Hum Mutat 15:7-12.

den Dunnen JT, Dalgleish R, Maglott D, Hart R, Greenblatt M, McGowan-Jordan J, Roux A-F, Smith T, Antonarakis SE, Taschner PEM. 2016. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat doi: 10.1002/humu.22981.

Emmert DB, Stoehr PJ, Stoesser G, Cameron GN. 1994. The European Bioinformatics Institute (EBI) databases. Nucleic Acids Res 22:3445-3449.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: The next generation in gene variant databases. Hum Mutat 32:557-563.

Gazzo AM, Daneels D, Cilia E, Bonduelle M, Abramowicz M, Van Dooren S, Smits G, Lenaerts T. 2016. DIDA: A curated and annotated digenic diseases database. Nucleic Acids Res 44:D900-D907.

Gbadegesin RA, Brophy PD, Adeyemo A, Hall G, Gupta IR, Hains D, Bartkowiak B, Rabinovich CE, Chandrasekharappa S, Homstad A, Westreich K, Wu G et al. 2013. *TNXB* mutations can cause vesicoureteral reflux. J Am Soc Nephrol 24:1313-1322.

George RA, Smith TD, Callaghan S, Hardman L, Pierides C, Horaitis O, Wouters MA, Cotton RGH. 2008a. General mutation databases: analysis and review. J Med Genet 45:65-70.

George RA, Smith TD, Callaghan S, Hardman L, Pierides C, Horaitis O, Wouters MA, Cotton RGH. 2008b. Response to Stenson et al on the review of general mutation databases. J Med Genet 45:319-320.

Giardine B, Borg J, Viennas E, Pavlidis C, Moradkhani K, Joly P, Bartsakoulia M, Riemer C, Miller W, Tzimas G, Wajcman H, Hardison RC et al. 2014. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. Nucleic Acids Res 42:D1063-D1069.

Giardine B, van Baal S, Kaimakis P, Riemer C, Miller W, Samara M, Kollia P, Anagnou NP, Chui DH, Wajcman H, Hardison RC, Patrinos GP. 2007. HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. Hum Mutat 28:206.

Giglione C, Boularot A, Meinnel T. 2004. Protein N-terminal methionine excision. Cell Mol Life Sci 61:1455-1474.

Graeber SY, Hug MJ, Sommerburg O, Hirtz S, Hentschel J, Heinzmann A, Dopfer C, Schulz A, Mainz JG, Tümmler B, Mall MA. 2015. Intestinal current measurements detect activation of mutant CFTR in patients with cystic fibrosis with the G551D mutation treated with Ivacaftor. Am J Respir Crit Care Med 192:1252-1255.

Hardison RC, Chui DHK, Giardine B, Riemer C, Patrinos GP, Anagnou N, Miller W, Wajcman H. 2002. *HbVar*: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum Mutat 19:225-233.

Hardison RC, Chui DHK, Riemer CR, Miller W, Carver MFH, Molchanova TP, Efremov GD, Huisman THJ. 1998. Access to a syllabus of human hemoglobin variants (1996) via the World Wide Web. Hemoglobin 22:113-127.

Hart RK, Rico R, Hare E, Garcia J, Westbrook J, Fusaro VA. 2015. A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. Bioinformatics 31:268-270.

Harte RA, Farrell CM, Loveland JE, Suner M-M, Wilming L, Aken B, Barrell D, Frankish A, Wallin C, Searle S, Diekhans M, Harrow J et al. 2012. Tracking and coordinating an international curation effort for the CCDS Project. Database (Oxford) 2012:bas008.

Herrick JB. 1910. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. Arch Intern Med 6:517-521.

Horaitis O, Cotton RGH. 2004. The challenge of documenting mutation across the genome: the Human Genome Variation Society approach. Hum Mutat 23:447-452.
Huisman THJ, Carver MFH, Baysal E. 1997. A Syllabus of Thalassemia Mutations. Augusta, GA: Sickle Cell Anemia Foundation. 309 p.
Huisman THJ, Carver MFH, Efremov GD. 1996. A Syllabus of Human Hemoglobin Variants. Augusta, GA: The Sickle Cell Anemia Foundation.
Huisman THJ, Carver MFH, Efremov GD. 1998. A syllabus of Human Hemoglobin Variants.2nd ed. Augusta GA: The Sickle Cell Anemia Foundation.
Ingram VM. 1956. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 178:792-794.
Ingram VM. 1957. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180:326-328.
ISCN. 2013. An international system for human cytogenetic nomenclature. Basel: S. Karger.
Johnston JJ, Biesecker LG. 2013. Databases of genomic variation and phenotypes: existing resources and future needs. Hum Mol Genet 22:R27-R31.
Kountouris P, Lederer CW, Fanis P, Feleki X, Old J, Kleanthous M. 2014. IthaGenes: an interactive database for haemoglobin variations and epidemiology. PloS one 9:e103020.
Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862-868.
Lehmann H, Kynoch PAM. 1976. Human Haemoglobin Variants and Their Characteristics. Amsterdam: North-Holland Publishing Company. 241 p.
Livingstone FB. 1967. Abnormal Hemoglobins in Human Populations. Chicago: Aldine Publishing Company. 476 p.
MacArthur JAL, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, Larsson P, Flicek P, Dalgleish R, Maglott DR, Cunningham F. 2014. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res 42:D873-D878.
Macchiato MF, Tramontano A. 1990. VARIANT: a store and retrieval system for human haemoglobin variants. Comput Methods Programs Biomed 31:113-114.
Marotta CA, Forget BG, Weissman SM, Verma IM, McCaffrey RP, Baltimore D. 1974. Nucleotide sequences of human globin messenger RNA. Proc Natl Acad Sci USA 71:2300-2304.
Mitropoulou C, Webb AJ, Mitropoulos K, Brookes AJ, Patrinos GP. 2010. Locus-specific database domain and data content analysis: evolution and content maturation toward clinical use. Hum Mutat 31:1109-1116.
Niroula A, Vihinen M. 2016. Variation interpretation predictors: principles, types, performance and choice. Hum Mutat,IN THIS SPECIAL ISSUE.
O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733-D745.
Ogino S, Gulley ML, den Dunnen JT, Wilson RB. 2007. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. J Mol Diagn 9:1-6.
Patnaik SK, Helmberg W, Blumenfeld OO. 2012. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. Nucleic Acids Res 40:D1023-1029.
Patrinos GP, Giardine B, Riemer C, Miller W, Chui DH, Anagnou NP, Wajcman H, Hardison RC. 2004. Improvements in the *HbVar* database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res 32:D537-D541.
Pauling L, Itano HA, Singer SJ, Wells IC. 1949. Sickle cell anemia, a molecular disease. Science 110:543-548.
Pesci E, Bettinetti L, Fanti P, Galietta LJV, La Rosa S, Magnoni L, Pedemonte N, Sardone GL, Maccari L. 2015. Novel hits in the correction of ΔF508-cystic fibrosis transmembrane conductance

regulator (CFTR) protein: synthesis, pharmacological, and ADME evaluation of tetrahydropyrido[4,3-d]pyrimidines for the potential treatment of cystic fibrosis. J Med Chem.

Povey S, Al Aqeel AI, Cambon-Thomsen A, Dalgleish R, den Dunnen JT, Firth HV, Greenblatt MS, Barash CI, Parker M, Patrinos GP, Savige J, Sobrido M-J et al. 2010. Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs). Hum Mutat 31:1179-1184.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. 2015. 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. Genet Med 17:405-424.

Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15:852-859.

Ritchie GRS, Flicek P. 2014. Computational approaches to interpreting genomic sequence variation. Genome Med 6:87.

Rucknagel DL. 1964. Current concepts of the genetics of thalassemia. Ann N Y Acad Sci 119:436-449.

Samuels ME, Rouleau GA. 2011. The case for locus-specific databases. Nat Rev Genet 12:378-379.

Savige J, Dalgleish R, Cotton RGH, den Dunnen JT, Macrae F, Povey S. 2015. The Human Variome Project: ensuring the quality of DNA variant databases in inherited renal disease. Pediatr Nephrol 30:1893-1901.

Schäffer AA. 2013. Digenic inheritance in medical genetics. J Med Genet 50:641-652.

Soussi T. 2014. Locus-specific databases in cancer: what future in a post-genomic era? The TP53 LSDB paradigm. Hum Mutat 35:643-653.

Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet 45:124-126.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133:1-9.

Swertz MA, Dijkstra M, Adamusiak T, van der Velde JK, Kanterakis A, Roos ET, Lops J, Thorisson GA, Arends D, Byelas G, Muilu J, Brookes AJ et al. 2010. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. BMC Bioinformatics 11 Suppl 12:S12.

Tack V, Deans ZC, Wolstenhome N, Patton S, Dequeker EMC. 2016. What's in a name? A co-ordinated approach towards a uniform nomenclature to improve patient reports and databases. Hum Mutat doi: 10.1002/humu.22975.

Taliaferro WH, Huck JG. 1923. The inheritance of sickle-cell anaemia in man. Genetics 8:594-598.

Taschner PEM, den Dunnen JT. 2011. Describing structural changes by extending HGVS sequence variation nomenclature. Hum Mutat 32:507-511.

Thein SL, Hesketh C, Taylor P, Temperley IJ, Hutchinson RM, Old JM, Wood WG, Clegg JB, Weatherall DJ. 1990. Molecular basis for dominantly inherited inclusion body beta-thalassemia. Proc Natl Acad Sci USA 87:3924-3928.

Vail PJ, Morris B, van Kan A, Burdett BC, Moyes K, Theisen A, Kerr ID, Wenstrup RJ, Eggington JM. 2015. Comparison of locus-specific databases for BRCA1 and BRCA2 variants reveals disparity in variant classification within and among databases. J Community Genet 6:351-359.

van den Akker PC, Jonkman MF, Rengaw T, Bruckner-Tuderman L, Has C, Bauer JW, Klausegger A, Zambruno G, Castiglia D, Mellerio JE, McGrath JA, van Essen AJ et al. 2011. The international dystrophic epidermolysis bullosa patient registry: an online database of dystrophic epidermolysis bullosa patients and their COL7A1 mutations. Hum Mutat 32:1100-1107.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34:275-282.

Vihinen M. 2015. Muddled genetic terms miss and mess the message. Trends Genet 31:423-425.

Vihinen M, den Dunnen JT, Dalgleish R, Cotton RGH. 2012. Guidelines for establishing locus specific databases. Hum Mutat 33:298-305.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6-13.

Zhang F, Lupski JR. 2015. Non-coding genetic variants in human disease. Hum Mol Genet 24:R102-R110.

Zhao X, Xiao J, Wang H, Ren X, Gao J, Wu Y, Lu C, Zhang X. 2015. Spectrum of COL1A1/2 mutations and gene diagnosis in Chinese patients with osteogenesis imperfecta. Zhonghua Yi Xue Za Zhi 95:3484-3489.

Zhao Y, Yuan J, Chen R. 2016. NONCODEv4: Annotation of noncoding RNAs with emphasis on long noncoding RNAs. Methods Mol Biol 1402:243-254.