Title: Recommendations for locus-specific databases and their curation.
Authors: Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehvaslaiho H, Liang P, Marsh S, Nebert DW, Povey S, Rossetti S, Scriver CR, Summar M, Tolan DR, Verma IC, Vihinen M, den Dunnen JT
Journal: Human mutation
Year: 2008 Jan
Volume: 29
Issue: 1
Pages: 2-5
DOI: 10.1002/humu.20650

Published in final edited form as:
Hum Mutat. 2008 January ; 29(1): 2–5. doi:10.1002/humu.20650.

Recommendations for Locus-Specific Databases and Their Curation

R.G.H. Cotton[1,*], **A.D. Auerbach**[2], **J.S. Beckmann**[3], **O.O. Blumenfeld**[4], **A.J. Brookes**[5], **A.F. Brown**[6], **P. Carrera**[7], **D.W. Cox**[8], **B. Gottlieb**[9], **M.S. Greenblatt**[10], **P. Hilbert**[11], **H. Lehvaslaiho**[12], **P. Liang**[13], **S. Marsh**[14], **D.W. Nebert**[15], **S. Povey**[16], **S. Rossetti**[17], **C.R. Scriver**[18], **M. Summar**[19], **D.R. Tolan**[20], **I.C. Verma**[21], **M. Vihinen**[22,23], and **J.T. den Dunnen**[24]

[1]Genomic Disorders Research Centre, St. Vincent's Hospital Melbourne, Australia [2]Laboratory of Human Genetics and Hematology, Rockefeller University, New York, New York [3]Service and Department of Medical Genetics, Centre Hospitalier Universitaire Vaudois and Faculty of Medicine, University of Lausanne, Lausanne, Switzerland [4]Department of Biochemistry, Albert Einstein College of Medicine, New York, New York [5]Department of Genetics, University of Leicester, Leicester, United Kingdom [6]MRC Human Genetics Unit, Edinburgh, United Kingdom [7]San Raffaele Scientific Institute and Laboraf, Laboratory of Molecular Biology, Milano, Italy [8]Department of Medical Genetics, University of Alberta, Edmonton, Alberta, Canada [9]Lady Davis Institute for Medical Research, Montreal, Quebec, Canada [10]University of Vermont College of Medicine, Burlington, Vermont [11]Institut de Pathologie et de Génétique (IPG)-Molecular Biology, Gosselies, Belgium [12]South African National Bioinformatics Institute (SANBI), University of the Western Cape, Cape Town, South Africa [13]Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, New York [14]Division of Oncology, Washington University, St. Louis, Missouri [15]Department of Environmental Health, University of Cincinnati Medical Center, Cincinnati, Ohio [16]Galton Laboratory, University College London, London, United Kingdom [17]Mayo Clinic College of Medicine, Rochester, Minnesota [18]Montreal Children's Hospital Research Institute, McGill University, Montreal, Quebec, Canada [19]Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee [20]Biology Department, Boston University, Boston, Massachusetts [21]Department of Genetic Medicine, Sir Ganga Ram Hospital, New Delhi, India [22]Institute of Medical Technology, University of Tampere, Tampere, Finland [23]Research Unit, Tampere University Hospital, Tampere, Finland [24]Leiden University Medical Center, Leiden, The Netherlands

Abstract

Expert curation and complete collection of mutations in genes that affect human health is essential for proper genetic healthcare and research. Expert curation is given by the curators of gene-specific mutation databases or locus-specific databases (LSDBs). While there are over 700 such databases, they vary in their content, completeness, time available for curation, and the expertise of the curator. Curation and LSDBs have been discussed, written about, and protocols have been provided for over 10 years, but there have been no formal recommendations for the ideal form of these entities. This work initiates a discussion on this topic to assist future efforts in human genetics. Further discussion is welcome.

*Correspondence to: Richard G.H. Cotton, PhD, DSc, Professor, Genomic Disorders Research Centre, 7th Floor Daly Wing, St. Vincent's Hospital Melbourne, 35 Victoria Parade, Fitzroy VIC 3065, Australia. E-mail: cotton@unimelb.edu.au.

Keywords

mutation; variome; database; curation; LSDB; bioinformatics

Introduction

There has been much discussion of gene databases or locus-specific databases (LSDBs) and their curation. This is happening because it is acknowledged that curation of mutations in genes is best done by experts in that gene or disease [Cotton, 2000], rather than by workers at central databases who collect information from the literature. This notion has been underscored recently by Gout et al. [2007], who, using their expertise as LSDB curators, documented 5% errors in the literature, and Murphy et al. [2004] found that 43% of CDKN2A mutations had at least one error in the report. Obviously, these errors cannot be tolerated since serious decisions, such as abortion and radical therapy, are being made on published data. Similarly, wasteful research and diagnostic strategies could be based on such faulty data.

There have been many publications on LSDBs, including a review of content and recommendations for essential data fields [Claustres et al., 2002], recommendations for setting up an LSDB [Horaitis and Cotton, 1999] (updated in 2003 and 2005), recommendations for submission forms [Horaitis and Cotton, 1999] (see also form by A.D. Auerbach et al. at www.hgvs.org/entry.html, and Table 1) and listings of LSDBs [Horaitis et al., 2007]. However, despite all this, and despite the existence of almost 700 databases that we refer to as LSDBs, there has been neither an agreed definition of the ideal LSDB nor a definition of how they should be curated. Such definitions are needed, because in the future we will rely increasingly on expertly curated information. An aim of the Human Variome Project (HVP) [Anonymous, 2007; Cotton et al., 2007a; Cotton and Kazazian, 2005] is to have such expertly curated information. The ideal vision is that each of the 20,488 protein coding human genes [Pennisi, 2007] will ultimately have curated LSDBs. This principle should probably be extended further to an as yet unknown number of short or regulatory non-protein coding genes (microRNAs and others) as well as structural elements or copy number variable regions of clinical relevance e.g. CNV disease database (see Table 1). An updated detailed protocol for establishing an LSDB is in preparation (J.T. den Dunnen and C. Beroud, personal communication).

LSDBS

Up until now, a working definition of an LSDB has been “a listing of sequence variants in a specific gene(s) causing a Mendelian disorder or a change in the phenotype, curated by an expert in that gene” (unpublished assumption). These listings have usually been initiated and driven by the interests of the curator, which might be research, clinical, or diagnostic in nature. Most LSDBs are now Internet-based and most are easily accessible by the general public, hence it is essential that they be as accurate as possible. This definition sets LSDBs apart from central mutation databases (CMDBs), the most prominent being the Human Gene Mutation Database (HGMD) [Stenson et al., 2003] (see Table 1) and Online Mendelian Inheritance in Man (OMIM) [Hamosh et al., 2005] (see Table 1), which collect variants in all genes, mainly from the literature. It should be noted that these databases do not collect all variants. OMIM generally collects the first variants described and later some with unique characteristics. HGMD collects only the first report of a variant, some associated phenotype but not its recurrence. In addition, there are some collections/databases that appear to be intermediate in status. These include MutationView [Minoshima et al., 2001] (see Table 1), Blood Group Antigens [Blumenfeld and Patnaik, 2004] (see Table 1), the Immunogenetics database [Piirila et al., 2006] (see Table 1), and the collection by Retina International (Table 1). The existence of these latter databases, curated by experts in the *field* of the collection, has prompted this work. These databases contain

variants involving approximately 350, 40, 115, and 77 genes, respectively. Finally, collections of LSDBs are emerging that use one software tool and run on a single server, but each is curated by an expert [Beroud et al., 2000; Fokkema et al., 2005]. The software for creating these LSDB collections reflect the defined characteristics obtained by the survey of Claustres et al. [2002] and is freely available (LOVD and UMD; see Table 1). The software used for the Immunogenetics database on a single server uses a third software [Riikonen and Vihinen, 1999] (see Table 1).

The Ideal LSDB

The review of Claustres et al. [2002] examined 100 representative LSDBs and found 80 characteristics that the curators had included. This list must be close to the ideal content and was therefore used as a basis for recommendations [Claustres et al., 2002; Horaitis and Cotton, 1999 (updated in 2003 and 2005)]. The variant submission form suggestion reflected this content [Horaitis and Cotton, 1999] (updated in 2003 and 2005) (see form by A.D. Auerbach et al., at www.hgvs.org/entry.html, and see Table 1), although priority was given to any information that was considered to be crucial. To be most useful to the widest possible community, the database needs to be complete, up-to-date, cover all aspects of the variants, provide quality assessment for each entry, and information about the disease and the resulting clinical phenotype (or at least links to such data). For recessive traits it is necessary that annotations make it clear in which allelic combinations the (pathogenic) variants have been identified as well as when only one variant could be detected. LSDBs also ought to be designed in a dynamic and extendable format, as new requirements may appear as we progress in the description of all their elements (e.g., for partially penetrant traits, one may consider entering additional information as needed). While there is no absolute requirement that every gene alteration have its pathogenicity proven at least by reconstituting the mutation in a standard reporting system, the ideal LSDB should have such a category at least available, particularly with the discovery of an increasing number of “functional SNPs,” etc.

As almost all LSDBs have been constructed to examine a specific phenotypic condition (i.e., disease), it is important that they be designed to be able to record cases in which a mutation is not reported after the gene is sequenced. This could then lead to the identification of novel genes causing these diseases if additional genes are suspected to exist. The ideal LSDB needs to be as flexible a tool as possible for studying disease phenotypes.

As the importance of LSDBs for diagnostic purposes increases, a quality assurance system will be required, including a periodic evaluation (similar to an accreditation process) to ensure no relaxation of these criteria. A positive review can then be used to award qualifying LSDBs a certification label.

Table 2 indicates logical broad recommendations for LSDBs to allow maximum usage and efficiency. Databases following these guidelines could be designated Human Genome Variation Society (HGVS) LSDBs, possibly with two stages, “initial” and “definite”; others could be referred to simply as LSDBs (or some other name) because they are more similar to listings in central variant databases, e.g., a specialized central database.

Curation

The original reason for collecting the details of variants in the first gene specific database (globin) was to facilitate structural and functional studies of the encoded protein [Huisman et al., 1997]. Later, the reason for collecting details on DNA variants was mostly to assist diagnosis and prognosis of patients with a Mendelian disorder. A recent meta-analysis allowed biological conclusions to be drawn [Marini et al., 2007], demonstrating another value of properly curated LSDBs.

LSDBs can also be used to examine phenotype–genotype correlations, provided that clinical data are collected, either directly or through adequate links. However, to successfully detect positive correlations, these data have to be collected in a carefully designed and structured way and for a large set of cases. Also, for phenotype collection, with association to specific genotypes, patient privacy issues will have to be addressed by Institutional Review Boards at the institutions of the curator and/or submitter. It is urgent that ethical guidelines be established so that phenotype data can be made available in public databases for genotype correlation to be studied.

Curation has ranged from a small number of non-gene experts extracting variants from literature at OMIM and HGMD, through a range of genes being curated by an expert in the field, to individuals or small consortia curating variants in single genes. In the latter case, the larger databases tend to be run by a team of curators, experts in specific subfields (e.g., PAHdb; see Table 1) [Scriver et al., 2003].

Gene-specific curation by experts is difficult to attain, mostly due to time constraints of the curator as well as the difficulty in finding funding for these activities. However, expert curation is needed as indicated by the recent description of errors in the literature [Gout et al., 2007]. In fact, the success of the LSDBs as an indispensable tool supporting DNA diagnosis puts an increasing demand on their quality and "up-to-dateness," which is in direct conflict with the work being carried out mostly by enthusiastic volunteers [Cotton et al., 2007b]. It is evident that the availability of a good LSDB has a significant positive and steadily increasing effect on evidence-based diagnostic decision making. In addition, it avoids clinicians and diagnosticians spending much time on collecting this information locally (and in multiplicate world-wide for everybody working on the same gene). In fact, immediately generating an LSDB for a gene without one is imperative. Since the software required is freely available (e.g., UMD and LOVD), assistance in setting-up the LSDB is offered, including free hosting (LOVD), there should be no limitations to establish these missing LSDBs.

Interestingly, however, although it is realized how useful the LSDBs are, the noncommercial diagnostic professionals using them turn out to have little time to submit their own findings. First, it is evident that this attitude is counterproductive for the completeness and quality of the resource they use and consider essential. Data on recurrence and frequency are not building up and new variants detected (either pathogenic, unclassified, or not pathogenic) are not registered. Second, the great virtue of the Internet, being a simple, easy, cheap, and very fast tool to exchange data, is not used. Especially for complicated cases in which the underlying pathogenicity can not be determined easily, the so called unclassified variants, the LSDB would be extremely useful. It would provide a perfect low-level possibility to share such findings and request colleagues to contact each other when additional information regarding the pathogenicity of the variant of interest is obtained. Finally, it should be noted that submission results in a free-of-charge error and quality check by the LSDB curator before the new variant is accepted and uploaded.

Ideal Curation

If funding was not a limiting factor, the characteristics of desirable curation are given in Table 3.

Future Prospects

Some LSDBs indicate that they have been constructed according to HUGO-Mutation Database Initiative/HGVS guidelines (e.g., Blood Group Antigens database) [Blumenfeld and Patnaik, 2004]. This is currently done on a voluntary basis, even though these guidelines have not yet been strictly defined. When these recommendations are finalized, they should become part of

an LSDB quality and approval system, which should benefit both LSDB curators and users. A discussion work to define formal guidelines for LSDBs is currently in preparation.

Funds for curation have been difficult to find because of the rareness of many diseases. However, as a group, these disorders are very important, especially when phenotypic consequences will be eventually linked to most or all of the 20,488 protein-coding genes. Minimal curation after database setup has been estimated to be about 1 day per week, largely depending on the LSDB size, e.g., TP53 (somatic) being at one extreme, and many genes having only a handful of variants at the other extreme [Cotton et al., 2007b]. This 1 day per week translates into about US $10,000 per year; therefore, for all 20,488 protein-coding genes, the cost is prohibitive. Thus, strategies are being developed to spread the load, e.g., by gene-specific sponsoring of LSDBs.

By having expertly curated LSDBs, which are kept up-to-date, both the user community and the patient will benefit. It should be realized that LSDBs create a considerable cost saving for DNA diagnosis; for each gene without a well-curated LSDB, each clinician performing DNA diagnosis for this gene is spending costly time trying to keep track of all variants reported world-wide in local private lists.

Acknowledgments

We thank Rania Horaitis for help in preparation of this manuscript.

Grant sponsor: National Health and Medical Research Council (NHMRC).

References

What is the Human Variome Project? Nat Genet 2007;39:423. [PubMed: 17392793]

Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. UMD (Universal Mutation Database): a generic software to build and analyze locus-specific databases. Hum Mutat 2000;15:86–94. [PubMed: 10612827]

Blumenfeld OO, Patnaik SK. Allelic genes of blood group antigens: a source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. Hum Mutat 2004;23:8–16. [PubMed: 14695527]

Claustres M, Horaitis O, Vanevski M, Cotton RG. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 2002;12:680–688. [PubMed: 11997335]

Cotton RG. Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. Hum Mutat 2000;15:4–6. [PubMed: 10612814]

Cotton RG, Kazazian HH Jr. Toward a human variome project. Hum Mutat 2005;26:499.

Cotton RG, 2006 Human Variome Project. Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A, Brown AF, Byers P, Cantu JM, Cassiman JJ, Claustres M, Concannon P, Cotton RG, den Dunnen JT, Flicek P, Gibbs R, Hall J, Hasler J, Katz M, Kwok PY, Laradi S, Lindblom A, Maglott D, Marsh S, Masimirembwa CM, Minoshima S, de Ramirez AM, Pagon R, Ramesar R, Ravine D, Richards S, Rimoin D, Ring HZ, Scriver CR, Sherry S, Shimizu N, Stein L, Tadmouri GO, Taylor G, Watson M. Recommendations of the 2006 Human Variome Project meeting. Nat Genet 2007a;39:433–436. [PubMed: 17392799]

Cotton RG, Phillips K, Horaitis O. A survey of locus-specific database curation. J Med Genet 2007b; 44:e72. [PubMed: 17400791]

Fokkema IF, den Dunnen JT, Taschner PE. LOVD: easy creation of a locus-specific sequence variation database using an “LSDB-in-a-box” approach. Hum Mutat 2005;26:63–68. [PubMed: 15977173]

Gout AM, ADPKD Gene Variant Consortium. Ravine D, Harris PC, Rossetti S, Peters D, Breuning M, Henske EP, Koizumi A, Inoue S, Shimizu Y, Thongnoppakhun W, Yenchitsomanus PT, Deltas C, Sandford R, Torra R, Turco AE, Jeffery S, Fontes M, Somlo S, Furu LM, Smulders YM, Mercier B,

Ferec C, Burtey S, Pei Y, Kalaydjieva L, Bogdanova N, McCluskey M, Geon LJ, Wouters CH, Reiterova J, Stekrová J, San Millan JL, Aguiari G, Del Senno L. Analysis of published PKD1 gene sequence variants. Nat Genet 2007;39:427–428. [PubMed: 17392796]

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33 (Database issue):D514–D517. [PubMed: 15608251]

Horaitis, O.; Cotton, RGH. Current protocols in human genetics, Unit 7.11. New York: Wiley-Liss; 1999. Human mutation databases; p. 7.11.1-7.11.11.Originally published in 1999 and updated in 2003 and 2005

Horaitis O, Talbot CC Jr, Phommarinh M, Phillips KM, Cotton RG. A database of locus-specific databases. Nat Genet 2007;39:425. [PubMed: 17392794]

Huisman, THJ.; Carver, MFH.; Efremov, GD. A syllabus of Human Hemoglobin Variants. The Sickle Cell Anemia Foundation; Augusta, GA: 1996. p. 420

Marini JC, Forlino A, Cabral WA, Barnes AM, San Antonio JD, Milgrom S, Hyland JC, Korkko J, Prockop DJ, De Paepe A, Coucke P, Symoens S, Glorieux FH, Roughley PJ, Lund AM, Kuurila-Svahn K, Hartikka H, Cohn DH, Krakow D, Mottes M, Schwarze U, Chen D, Yang K, Kuslich C, Troendle J, Dalgleish R, Byers PH. Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans. Hum Mutat 2007;28:209–221. [PubMed: 17078022]

Minoshima S, Mitsuyama S, Ohtsubo M, Kawamura T, Ito S, Shibamoto S, Ito F, Shimizu N. The KMDB/MutationView: a mutation database for human disease genes. Nucleic Acids Res 2001;29:327–328. [PubMed: 11125127]

Murphy JA, Barrantes-Reynolds R, Kocherlakota R, Bond JP, Greenblatt MS. The CDKN2A database: integrating allelic variants with evolution, structure, function, and disease association. Hum Mutat 2004;24:296–304. [PubMed: 15365986]

Pennisi E. Working the (gene count) numbers: finally, a firm answer? Science 2007;316:1113. [PubMed: 17525311]

Piirilä H, Väliaho J, Vihinen M. Immunodeficiency mutation databases (IDbases). Hum Mutat 2006;27:1200–1208. [PubMed: 17004234]

Riikonen P, Vihinen M. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 1999;15:852–859. [PubMed: 10705438]

Scriver CR, Hurtubise M, Konecki D, Phommarinh M, Prevost L, Erlandsen H, Stevens R, Waters PJ, Ryan S, McDonald D, Sarkissian C. PAHdb 2003: what a locus-specific knowledgebase can do. Hum Mutat 2003;21:333–344. [PubMed: 12655543]

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003;21:577–581. [PubMed: 12754702]

TABLE 1

List of Internet Sites

Blood Group Antigens	http://www.ncbi.nlm.nih.gov/gv/rbc/xslcgi.fcgi?cmd=bgmut/home
dbSNP	www.ncbi.nlm.nih.gov/projects/SNP
Decipher, Copy Number Variation Database	http://www.sanger.ac.uk/PostGenomics/decipher
HGMD	www.hgmd.cf.ac.uk
HGVS listing of LSDBs	www.hgvs.org/dblist/glsdb.html
HGVS recommendation for mutation nomenclature	www.hgvs.org/mutnomen
HGVS recommendations for submission forms	www.hgvs.org/entry.html
Human Variome Project	www.humanvariomeproject.org
ImmunoDeficiency Mutation Databases	http://bioinf.uta.fi/IDbases
LOVD software	www.lovd.nl
MutationView	http://mutview.dmb.med.keio.ac.jp
MUTbase	http://bioinf.uta.fi/MUTbase
OMIM	www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM
PAH Knowledgebase	www.pahdb.mcgill.ca
Retina International	www.retina-international.org/sci news/mutation.htm
UMD	www.umd.be

TABLE 2
The Ideal LSDB Characteristics

1	Use of standardized nomenclature.
2	Use of uniform or standardized LSDB software.
3	Use of standard data fields across LSDBs.
4	All important fields filled by expert curators.
5	HUGO-MDI/HGVS guidelines followed, and such stated.
6	Publication on, and registration in, the list of LSDBs at HGVS.
7	Use of the HGVS/HUGO MDI/form for incoming unpublished entries.
8	Curation as per guidelines.
9	Links to clinical, phenotype, and protein 3D structure databases.
10	Periodic quality review (e.g., every 2–4 years, depending on the importance of the database).
11	LSDBs publicly available.
12	Ethical guidelines need to be followed.
13	Funding and a permanent site identified for sustainability and for the permanent record.
14	Accreditation required.

TABLE 3
Ideal Curation of LSDBs

1	Initially, collection through a review of all mutations that have been published.
2	Collection not only from the literature, but also from a network of collaborators, diagnostic laboratories, other databases, etc.
3	LSDB maintained by a team of expert curators, having complementary expertise, e.g. clinical, metabolic, pathogenicity, protein structure, etc., and including the original discoverer(s) of the disease/gene association, if relevant.
4	Data collection aided by software tools (e.g., mutation checker) to ensure data is correct and up-to-date.
5	Promotion of a strategy to determine the pathogenicity of variants of the gene.