Serveur Académique Lausannois SERVAL serval.unil.ch

Author Manuscript
Faculty of Biology and Medicine Publication

This paper has been peer-reviewed but dos not include the final publisher proof-corrections or journal pagination.

Published in final edited form as:

serval
serveur académique lausannois

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Sharing Data between LSDBs and Central Repositories

Johan T. den Dunnen[1,*], **Rolf H. Sijmons**[2], **Paal S. Andersen**[3], **Mauno Vihinen**[4,5], **Jacques S. Beckmann**[6], **Sandro Rossetti**[7], **C. Conover Talbot Jr**[8], **Ross C. Hardison**[9], **Sue Povey**[10], and **Richard G.H. Cotton**[11,12]

[1]Leiden University Medical Center, Albinusdreef 2, Leiden, The Netherlands [2]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands [3]National Center for Antimicrobials and Infection Control, Statens Serum Institute, Copenhagen, Denmark [4]Insitute of Medical Technology, FI-33014 University of Tampere, Finland [5]Tampere University Hospital, FI-33520 Tampere, Finland [6]Service and Department of Medical Genetics, CHUV-UNIL, Lausanne, Switzerland [7]Nephrology Research, Mayo Clinic College of Medicine, Rochester Minnesota [8]The Johns Hopkins University School of Medicine, Baltimore, Maryland [9]Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania [10]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom [11]Genomic Disorders, VIC 3053, Australia [12]Research Centre, VIC 3053, Australia [13]Department of Medicine, The University of Melbourne, Parkville, VIC 3010, Australia

Abstract

Several Locus-Specific DataBases (LSDBs) have recently been approached by larger, more general data repositories (including NCBI and UCSC) with the request to share the DNA variant data they have collected. Within the Human Genome Variation Society (HGVS) a document was generated summarizing the issues related to these requests. The document has been circulated in the HGVS/LSDB community and was discussed extensively. Here we summarize these discussions and present the concluded recommendations for LSDB data sharing with central repositories.

Keywords

Introduction

With full genome sequencing at hand, we need to develop much better tools to store and share existing and new data on sequence variation and its (potential) phenotypic consequences [Cotton et al., 2007a]. Thus, there is an increasing pressure to share the data

*Correspondence to: Johan T. den Dunnen, LUMC, Albinusdreef 2, P.O. Box 9600, Zone S04-030, 2300 RC Leiden, The Netherlands. ddunnen@lumc.nl.

that curators of LSDBs have gathered [Cotton et al., 2008]. LSDB curators agree that data integration is more important than ever and information in LSDBs should be widely and easily available. Overall, they are supportive, but they also see data sharing potentially as a great risk and a threat. Their fear is that years of cautious quality-checked data collection and curation work will be given away for free, where it might be mixed with other lower quality data, and thus eventually lose much of its substance and raison d'être. Assuming appropriate conditions will be met to ensure adequate quality checks in these central repositories, additional concerns of LSDB curators need to be addressed, such as "What will be their recognition or their reward?" Currently, third parties sometimes take data without asking for permission, without giving proper acknowledgement or any reward, and sometimes the data are even used for profit. Furthermore, the LSDBs may envisage that the data will be shared once and then "repackaged" by a central repository. Who will need the LSDB after that? Who will warrant the data quality and provenance? Who owns the data?

For most, it is clear that there will be a role for everybody in this field. Currently, LSDBs provide the most accurate information resources with data quality and richness, especially regarding phenotype information, being much better than that available through other sources. Furthermore, compared to general data repositories, LSDBs mostly focus on a specific gene or condition and often develop specific additional tools.

A realistic threat exists that if LSDBs remain reluctant to share data, interested parties will find other ways to collect this information and bypass the LSDBs. In addition, LSDB curators often lack time and resources to devote to the maintenance and development of their database [Cotton et al., 2007b]. Consequently, they consider it the role of the data integrators to develop tools that accurately collect the information from LSDBs and make it more widely available. Overall, the central repositories have better resources and are much better qualified to perform this work (e.g., [Giardine et al., 2007]). The LSDB community can see the central collection, integration and redistribution of data as a positive issue; it will generate much more traffic and citations to their databases and thus provide more recognition. An intrinsic benefit for the LSDBs is that data quality and consistency will be checked when integrated with other resources. Also, redistribution ensures a more permanent record of the data. Finally, integration will facilitate collection and analysis of larger datasets, which adds reliability to the analysis of the effects of variants and to the evaluation of trends.

What Information Can We Expect to be Shared?

Minimal LSDB Information That Should Be Shared

1. Web address of the LSDB.

2. Contact details of the database curator(s).

3. Gene name, including the HGNC approved gene symbol, HGNC ID number, EntrezGene ID, MIM number (for genes and diseases).

4. DNA reference sequence used (GenBank or Ensembl accession + version number, preferably derived from the LRG-project as soon as these are available).

5. Description of published sequence variants at the DNA level using HGVS recommendations (see below) and, when available, the description as in the original report, dbSNP ID, and/or MIM number of variant.

6. LSDB specific identifier(s) to link directly to the specific variant(s) in the LSDB.

Further LSDB Information That Could Be Shared

Besides the basic and minimal information mentioned above, LSDBs contain a wealth of additional information, especially regarding the phenotype of patients carrying specific variants. What additional information could an LSDB consider to share with a central repository?

1. **Reference/PubMed ID:** sharing the PubMed ID in relation to the variants is desirable but not without consequences. In many cases an LSDB description of the variant has been changed from that published to follow HGVS recommendations and/or errors have been corrected (often after contacting the authors). Consequently, those following the publication link through a central repository might be unable to view the change in that publication. Such confusion can be prevented when the PubMed link is obtained through the LSDB, assuming the LSDB mentions when and how changes were made in relation to the original publication. When a central repository clearly indicates that such discrepancies may exist, an LSDB could share this information.

2. **Number of independent observations of a DNA variant:** The question is, what number should be exchanged here? (1) the total number of individual patients identified—which tells something about frequency in the population, but some variants will be specific for certain ethnic groups only; or (2) the total number of unrelated families identified. LSDB opinions currently favor option (3) as a more useful number, even though in certain cases one knows that multiple reports of a variant, in fact, represent a single founder mutation. The most favored format to share this information is through choice from a set of categories, for example, found once, 2–10 times, 11–99 times, >100 times. This also reduces the effect of patients that, without the LSDB curator's knowledge, have been reported several times in different publications.

3. **Change at protein level:** the change at protein level is in nearly 100% of cases a prediction based on the change found at DNA level. When desired, anybody can make this prediction. The consensus was that without experimental proof, an LSDB should not share this prediction. Experimental proof might exist when RNA or protein has been analyzed or when functional studies have been performed. If so, this information should be stored at the LSDBs (not usually done currently). LSDBs can share this information, but only when the central repository uses a discriminative display indicating the different levels of experimental proof.

4. **Change at RNA level:** at present, RNA is usually not analyzed, and any effect of a sequence variant at RNA level thus goes unnoticed. The increasing number of variants with unexpected consequences at the RNA level, mostly influencing splicing, that are reported underscores the danger of making a prediction based on

DNA sequence only. The considerations previously discussed under item 3 thus also apply to RNA. LSDBs should store this information when available and can then share it with a central repository.

5. **Associated pathogenicity:** although probably the most desired data field for sharing, this was generally considered the most problematic field. The definition of "pathogenic" itself varies depending on the specific clinical or biochemical criteria used [Easton et al., 2007; Plon et al., 2008]. For example, for the seemingly simple discrimination between Duchenne and Becker muscular dystrophies (DMD/BMD), submitters might use protein data, clinical data (e.g., the age ambulation is lost), or the genetic variant found. Opinions on disease symptoms may differ among clinical centres and countries, as does the naming of disease and symptoms. In rare diseases, small numbers make interpretation of variants rather risky. Large clinical heterogeneity (symptomatic and asymptomatic people) often can be observed even within a single family, attesting the influence of modifiers or incomplete penetrance. Furthermore, a given variant is generally tested on its own and only in relation to a specific disease; "no known pathogenicity" thus relates only to this specific phenotype and neglects any influence of the rest of the genome. In diseases such as Bardet-Biedl syndrome [Tayeh et al., 2008] or Retinitis Pigmentosa [Daiger et al., 2007], pathogenicity is dependent on the combined presence of variants in different genes that each alone may be nonpathogenic. Similarly, for multigenic diseases we can expect that frequent variants (single neucloetide polymorphisms [SNPs] or copy number variations [CNVs]), which most of the time behave as nonpathogenic, in certain cases change to pathogenic. To prevent easy misinterpretation, it will also be important to label pathogenic variants as either "dominant" or "recessive," that is, sufficient on its own to cause disease or only in combination with a variant on the other allele. From an LSDB, this distinction should be obvious, from a central repository it might not. One option is to share data on pathogenicity for clear-cut cases only [Plon et al., 2008], that is, those variants that have been shown many times to be associated with pathogenicity and those that have no known association with disease. All others could be classified as "unknown." Using more categories seems attractive, for example, "probably pathogenic" and "probably not pathogenic," but their use on a central site is probably confusing for nonexperts. Furthermore, it would distract from the fact that to draw such a conclusion the source, that is, the LSDB, should be consulted. Another problem is inconsistency among reports. For example, within an LSDB one may find five reports listing a specific variant as "pathogenic," three as "pathogenicity unknown" and another five as "no known pathogenicity." What should be reported in such cases? Thus, when the criteria used to discriminate these categories are clearly stated, LSDBs can share pathogenicity information labelled as either "no known pathogenicity," "unknown," or "pathogenic." The same holds true for a central repository, it should clearly state the criteria on which they made the distinction between the classes used. Preferably both parties use the same criteria, and it is advised to include strong disclaimers explaining the associated uncertainties.

6. **Unpublished/variants submitted directly to LSDB:** before this information can be shared, LSDB curators should make it clear to their submitters that they might share the data collected, how they will do this, and specify what information will be shared. For existing unpublished records, LSDB curators would have to ask permission from the original submitter before data could be shared. LSDBs should have a public "database policy" paragraph explaining their principles in this respect. When LSDBs receive confidential variants (i.e., full records submitted but not yet released, e.g., while awaiting publication), these should not be shared. It has been suggested that LSDBs should follow a policy where data can be kept confidential for only a limited time, for example, 12 months, after which the data become public automatically.

What Needs to be Done?

Central Repositories

1. List collaborating curators and funders of the LSDB on the central site.

2. Make it clear from where the displayed DNA variant data have been obtained (LSDB address, logo) and explain how the data are displayed (i.e., the categories and criteria used, e.g., regarding consequences at protein level and in relation to pathogenicity).

3. Provide for every variant a direct link to that variant in the respective LSDB. Share the central repository ID files (e.g., assigned dbSNP numbers) to allow the LSDB to link to the central repository.

4. Decide on a policy to follow when there is more then one LSDB for a gene. Will links be provided to all LSDBs containing a specific variant or will one LSDB be selected? What will be done when LSDBs have conflicting data?

5. Upon request, assist LSDB curators with fundraising by writing letters of support stating that the requesting LSDB actively shares its data and the value this has had for the central repository. Similarly, where possible, the central repository will keep track of data usage and give LSDB curators access to these figures.

6. Establish standard systems and semantics for LSDB data sharing (including data format, sites for data uploads, etc.), enabling LSDB curators to easily share their data and software developers and integrators to build tools for regular automated data exchange.

7. Help LSDBs to obtain data they would like to receive, e.g. a list of all variants known to the central repository. Such a list can be very helpful for a curator to initiate a new gene variant database.

LSDBs

LSDBs will need to draft a "database policy" clearly explaining what they expect of the data submitted (e.g., that the submitter obtained informed consent) and what the LSDB may do with the data collected. This database policy should be clearly displayed at the LSDB Web site and should indicate whether data will be shared, which data, with whom, how, under

what conditions, and whether they will contact their submitters to discuss changes in LSDB policy. In addition, LSDBs should realize that when data are shared they immediately will have a new task, that is, regularly updating this shared information. LSDB software developers should note this demand and accordingly build in tools for data sharing.

Conclusion

Ideally, all sequence variant data, including variants causing disease, should be available on central browsers for safekeeping and integration with other data on the genome such as that generated by the ENCODE project [ENCODE Project Consortium 2007]. Such sharing has begun at UCSC [Giardine et al., 2007] (e.g., FKRP at http://globin.bx.psu.edu/phencode/pui.html) and NCBI (Povey and Maglott, unpublished; e.g., TSC2 at http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=7249). The concerns with data sharing are expressed in this paper, and a guide to the future is presented. Any comments should be expressed to the corresponding author or in a letter to the Editor.

References

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. Human Genome Variation Society. J Med Genet. 2007b; 44:e72. [PubMed: 17400791]

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. Recommendations for locus-specific databases and their curation. Hum Mutat. 2008; 29:2–5. [PubMed: 18157828]

Daiger SP, Bowne SJ, Sullivan LS. Perspective on genes and mutations causing retinitis pigmentosa. Arch Ophthalmol. 2007; 125:151–188. [PubMed: 17296890]

Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ, Goldgar DE. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am J Hum Genet. 2007; 81:873–883. [PubMed: 17924331]

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh M, Patnaik SK, Blumenfeld O, Gottlieb B, Vihinen M, Väliaho J, Kent J, Miller W, Hardison RC. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat. 2007; 28:554–562. [PubMed: 17326095]

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV. IARC Unclassified Genetic Variants Working Group. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat. 2008; 29:1282–1291. [PubMed: 18951446]

Tayeh MK, Yen HJ, Beck JS, Searby CC, Westfall TA, Griesbach H, Sheffield VC, Slusarski DC. Genetic interaction between Bardet-Biedl syndrome genes and implications for limb patterning. Hum Mol Genet. 2008; 17:1956–1967. [PubMed: 18381349]