HVP/DA/001-02/EN

2014-05-08

WG05

GDSDBAC

DRAFT FOR APPROVAL

GENE/DISEASE SPECIFIC VARIANT DATABASE QUALITY PARAMETERS

Authors

Mauno Vihinen, Lund University, Lund, Sweden,
Christophe Beroud, INSERM UMR_S910, Marseille, France,
Andrew Devereau, National Genetics Reference Laboratories, Manchester, UK,
John Hancock, University of Cambridge, Cambridge, UK,
Peter Taschner, Leiden University Medical Center, Leiden, The Netherlands

Editor

Timothy D. Smith, Human Variome Project International Ltd, Australia

Published by:
Human Variome Project International Limited
Level 5, 234 Queensberry Street, The University of Melbourne VIC 3010, AUSTRALIA

Published 2014-05-08. Printed in Australia.

PDF: **ISBN <<PDF_ISBN>>**
Print: **ISBN <<PRINT_ISBN>>**

Contents

1

Foreword

The Human Variome Project is an international consortium of researchers, policy makers and healthcare professionals committed to the free and open collection, curation, interpretation and sharing of genomic knowledge.

The Human Variome Project Consortium envisions a world where the availability of and access to genetic variation information is not an impediment to diagnosis and treatment; where the burden of genetic disease on the human population is significantly decreased; and where the sharing of genetic variation information is standard clinical practice.

To facilitate worldwide and interoperable sharing of genomic knowledge, the Human Variome Project Consortium produces Standards and Guidelines. HVP Standards are those systems, procedures and technologies that the Human Variome Project Consortium has determined shall be used by the community. These carry more weight than the less prescriptive HVP Guidelines, which cover those systems, procedures and technologies that the Human Variome Project Consortium has determined would be beneficial for the community to adopt.

HVP Standards and Guidelines are central to supporting the work of the Human Variome Project Consortium and cover a wide range of fields and disciplines, from ethics to nomenclature, data transfer protocols to collection protocols for clinical data. They can be thought of as both technical manuals and scientific documents, and while the impact of HVP Standards and Guidelines differ, they are both generated in a similar fashion.

HVP Standards and Guidelines make the collection, curation and sharing of information more efficient and reliable by establishing consistent protocols that can be universally understood. They facilitate interconnection of and interoperability between different systems.

HVP Standards and Guidelines represent a consensus of the Human Variome Project Consortium, each member of which has had the opportunity to participate in the development and review of each standard and guideline. In addition, as every effort is made to include all interests in the activity, HVP Standards and Guidelines can be considered to be representative of all interests concerned within the scope of each Standard or Guideline.

The Human Variome Project defines consensus as significant agreement between all affected parties covered by the scope of the standard or guideline. Consensus requires that all views and objections be considered, and that a concerted effort be made toward their resolution.

More information on the Human Variome Project is available at the Project's website (http://www.humanvariomeproject.org/). Procedures for the development of HVP Standards and Guidelines can be found in *PD06-2011: Standards Development Process*, available at http://short.variome.org/PD06-2011.

This Document

This document has been prepared the HVP Working Group: WG05: Variant Database Quality Assessment. The Gene/Disease Database Advisory Council acted as Sponsoring Council.

An Exposure Draft of this Document was released for comment to the Human Variome Project Consortium on 2014-02-13. The consultation period ended 2014-04-14.

A Draft for Approval was submitted to the Gene/Disease Specific Database Advisory Council 2014-05-08.

Important Notice

HVP Standards and Guidelines are not intended to replace or substitute for any applicable legislation or regulation in any jurisdiction, or any institutional policy or funding agreement that a genetic variation information resource is operating under. Implementers of HVP Standards and Guidelines are responsible for determining and complying with all appropriate ethical and cultural protection practices and all applicable laws, regulations, policies and agreements.

Introduction

Thousands of variation databases have been developed during the last decades. They have widely varying contents, resources, as well as quality. Quality is becoming ever more important as the use of these databases by the community increases and users start to combine and compare information in the different resources.

Computer science and IT communities have discussed and developed criteria for database platform/system quality, however, there are no widely accepted systematic criteria and evaluation systems. Each domain has developed its own. Closest to this field comes BioDBcore, which has developed core attributes of biological databases. However, it is not a quality scheme.

The criteria described in computer science and information technology quality papers are more suitable to assess database platforms (e.g. DMuDB, LOVD, MUTbase, UMD etc.) than for variant database content. For some platforms, there might be little difference due to limited customization and flexibility. Even in the same installation using the latest version of LOVD, one gene variant database might be of much better quality than another depending on the curator(s).

1 Scope

This document details the quality assessment parameters that might be used to evaluate the quality of genetic variation databases and stimulate database curators to make improvements where they are needed.

How these quality assessment criteria might be implemented in a quality accreditation scheme, including the criteria that should be used to assess compliance with each parameter is outside the scope of this document.

2 Quality Assessment Principles

The guiding principle used throughout this document is simplicity while maintaining a good overview of the quality of the database and its content. The goal is to capture the major issues pertinent to quality balanced with a simple system for database users to comprehend and database administrators and curators to apply.

3 Quality Assessment Criteria

Quality evaluation for genetic variation databases should consider four specific areas. The basic principle is to have a simple system, which still can provide a good overview of the quality. There could be overall ranking in the style of hotels with stars, which could then be broken down to individual scores. The main evaluation areas suggested are:

a) Data quality
b) Technical quality
c) Accessibility
d) Timeliness

These four quality evaluation criteria can be divided into more specific items:

3.1 Data quality

a) Database scope and purpose
b) Contents (patient data, if so, how much, fulfillment of minimal requirements[1])
c) Completeness of data
d) Coverage of variations
e) Accuracy, error rate
f) Consistency (language, spelling, reference sequences) - type of curation
g) Integration to other resources
h) Use of standards (gene names, reference sequences, variation nomenclature, data models, ontologies …)
i) Date stamps
j) Authority, curatorial team competence
k) Contact details
l) Use of references
m) Data collection, sources
n) Definition of pathogenicity used
o) What kind of data (e.g. NGS data, SNVs)
p) Range of numerical values (min to max)
q) Correct units
r) No data lost when inputted
s) Consents, privacy
t) Ethical issues
u) Public/nonpublic data

3.2 Technical quality

a) Database management system, suitability
b) Speed of access
c) Quality control measures implemented
d) Use of automatic steps (e.g. HGVS names with Mutalyzer)
e) How corrections made
f) Reliability (uptime)
g) Version history
h) Functional links
i) Use on different browsers
j) Data security (operating system, backups, firewall)

3.3 Accessibility

a) Design
b) Readability
c) User friendly, logical interface
d) Navigation
e) Language, correctness
f) Ease of use
g) Consistency on the site

[1] HVP Standards and Guidelines for minimum requirements for molecular and clinical data are currently under development.

h) Interactivity
i) Available freely, is registration required
j) How to contact
k) Help, support, tutorials
l) Documentation (purpose, scope, motivation, copyright, disclaimer, database policy, data items, annotation guidelines)
m) Searchability (search engine/possibilities)
n) Downloadability of data/search results
o) Format(s) of output (for further analyses)
p) Graphics (use, clarity etc)
q) Contacts to community, support groups etc
r) Other modes of access, esp. web services access

3.4 Timeliness

a) Update frequency - including whether it is currently being maintained
b) Currency of updates (how new data included)
c) Versioning policy (are old versions available?)

4 Assessing Quality

Although consideration of how a database quality assessment program might be implemented is outside the scope of this document, the following issues may be of use when such a program is developed.

Technical information required to perform a quality audit of a database might be provided by the database administrator. This kind of self-assessment or would save time for the quality evaluators. For popular platforms (e.g. LOVD) scripts could be developed to automatically collect a number of facts.

5 Relevant references

Celli, J, R Dalgleish, M Vihinen, PE Taschner, and JT den Dunnen. "Curating gene variant databases (LSDBs): Toward a universal standard." *Human Mutation* 33 (2012): 291-297.

Cotton, RGH, et al. "Recommendations for locus-specific databases and their curation." *Human Mutation* 29 (2008): 2-5.

den Dunnen, JT, et al. "Sharing data between LSDBs and central repositories." *Human Mutation* 30 (2009): 493-495.

Gaudet, P, et al. "Towards BioDBcore: a community-defined information specification for biological databases." *Nucleic Acids Research* 39 (2011): D7-10.

Smedley, D, et al. "Finding and sharing: new approaches to registries of databases and services for the biomedical sciences." *Database (Oxford)*, 2010: doi:10.1093/database/baq1014.

Vihinen, M, JT den Dunnen, R Dalgleish, and RGH Cotton. "Guidelines for establishing locus specific databases." *Human Mutation* 33 (2012): 298-305.