



HEALTH-F4-2007-200754

www.gen2phen.org

D3.4 Scope and Range Requirements of Specialized Domain Models

WP3 – Standard data models and terminologies

**V6.0
Final**

Lead beneficiary: EMBL
Date: 10/08/2009
Nature: Report
Dissemination level: PU
(Public)


 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson	Version: v6.0	2/13


TABLE OF CONTENTS

DOCUMENT INFORMATION	3
DOCUMENT HISTORY	3
1. INTRODUCTION	4
2. DESCRIPTION OF WORK	5
3. BASIC LSDB BACKGROUND INFORMATION	6
4. LSDB DATA CONTENT STANDARDS	7
5. SECURITY CONSIDERATIONS	9
6. PHENOTYPE EXTENSIONS	11
7. FUTURE PLANS	12
7.1. DERIVATION AND SPECIFICATION OF EXCHANGE FORMAT (D3.7).....	12
7.2. WP4 – GENETICS G2P DATABASES	12
7.3. NEW TECHNOLOGIES	12
7.4. SECURITY.....	12
8. ABBREVIATIONS	13
REFERENCES	13

APPENDIX I - LSDB BACKGROUND INFORMATION

APPENDIX II - LSDB MINIMAL DATA CONTENT

APPENDIX III - RECORD OF DISCUSSION AND COMMENTS

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0

Document Information

Grant Agreement Number	HEALTH-F4-2007-200754	Acronym	GEN2PHEN
Full title	Genotype-To-Phenotype Databases: A Holistic Solution		
Project URL	http://www.gen2phen.org		
EU Project officer	Frederick Marcus (Frederick.Marcus@ec.europa.eu)		


Deliverable	Number	D3.4	Title	Scope and Range Requirements of Specialized Domain Models
Work package	Number	3	Title	WP3 – Standard data models and terminologies

Delivery date	Contractual	June 2009	Actual	August 2009
Status	Version 6.0		final <input checked="" type="checkbox"/>	
Nature	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

Authors (Partner)	Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson (EMBL)			
Responsible Author	Helen Parkinson		Email	parkinson@ebi.ac.uk
	Partner	EMBL-EBI	Phone	+44 (0)1223 494 672

Document History

Name	Date	Version	Description
Tomasz Adamusiak	15/6/2009	1	
Helen Parkinson	7/7/2009	2	
Tomasz Adamusiak	12/7/2009	3	
Ilkka Lappalainen	14/7/2009	4	
Helen Parkinson	15/7/2009	5	
Helen Parkinson	10/8/2009	6	Review

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson	Version: v6.0	4/13

Definitions

- Partners of the GEN2PHEN Consortium are referred to herein according to the following codes:

ULEIC – University of Leicester (UK) – Coordinator

EMBL – European Molecular Biology Laboratory (Germany) – Beneficiary

FIMIM – Fundació IMIM (Spain) – Beneficiary

LUMC – Leiden University Medical Center (Netherlands) – Beneficiary

INSERM – Institut National de la Santé et de la Recherche Médicale (France) – Beneficiary

KI – Karolinska Institutet (Sweden) – Beneficiary

FORTH – Foundation for Research and Tecnology Hellas (Greece) – Beneficiary

CEA – Commissariat à l’Energie Atomique (France) – Beneficiary

EMC – Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) – Beneficiary

UH.FGC – Helsingin Yliopisto (Finland) – Beneficiary

UAVR – Universidade de Aveiro (Portugal) – Beneficiary

UWC – University of the Western Cape (South Africa) – Beneficiary

CSIR – Council of Scientific and Industrial Research (India) – Beneficiary

SIB – Swiss Institute of Bioinformatics (Switzerland) – Beneficiary

UNIMAN – The University of Manchester (UK) – Beneficiary


BIOBASE – BioBase GmbH. (Germany) – Beneficiary

deCODE – Islensk Erfoagreining EH (Iceland) – Beneficiary

PHENO – Phenosystems S.A. (Belgium) – Beneficiary

BCP – Biocomputing Platforms Ltd. Oy (Finland) – Beneficiary

- Grant Agreement:** The agreement signed between the beneficiaries and the European Commission for the undertaking of the GEN2PHEN project (HEALTH-200754).
- Project:** The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- Work plan:** Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the GEN2PHEN project, as specified in Annex I to the Grant Agreement.
- Consortium:** The GEN2PHEN Consortium, conformed by the above-mentioned legal entities.
- Consortium agreement:** agreement concluded amongst GEN2PHEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties’ obligations to the Community and/or to one another arising from the Grant Agreement.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson	Version: v6.0	5/13

1. INTRODUCTION

Work package 3 ‘Standard data models and terminologies’ provides domain standards to develop GEN2PHEN specific architecture, facilitate data exchange and integrate data across existing and emerging resources. This work package is focused on providing standards to act as the foundation for much of the database development activities of other work packages.

The work package objectives include the rapid development of a standard data model(s) capable of representing the minimum agreed content standard (as determined by WP2) and a derived data exchange format. Data models developed in coordination with WP3 will have several uses in GEN2PHEN: data from pre-existing databases will be mapped to generate data in a derived data exchange format, thus offering a flexible solution for integrating and exchanging existing and new data. In this respect, data model development is a necessary prerequisite, initially separated from implementation details.

2. DESCRIPTION OF WORK

The GEN2PHEN deliverable ‘Scope and Range Requirements of Specialized Domain Models’ provides the focus for specific data model development in later phases of the grant to support future partner requirements. We have focussed on areas in which partners are actively involved and which are emerging in the community. Specifically in the areas of:


- Defining LSDB background information
- Establishment of data content standards in LSDB context
- Establishment of security model requirements
- Support for complex phenotype representation
- Identification of relevant new technologies

Documents describing minimal information standards were created by the LUMC Partner with HGVS community feedback. These will be summarized here and full versions are available in Appendix 1 and 2. These, especially the ‘optional’ parts of the recommendations, allowed us to expand the scope of the LSDB models.

During the second GEN2PHEN Modelling Workshop (Helsinki, January 19-22, 2009, hosted by UH.FGC) partners and invited experts discussed the LSDB minimal content provided by LUMC Partner and detailed proceedings are available in Appendix 3, as well as online from:

<http://askja.gene.le.ac.uk/drupal5/content/lsdb-minimal-requirements>

Security model requirements are drawn from the European Genome-phenome Archive (EGA) and community reaction to the Homer et al. paper[1] which has affected the way that the data are distributed within the genotype-to-phenotype field.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0 6/13

The first GEN2PHEN Phenotype Workshop (Geneva, May 7-8, 2009, hosted by SIB) produced the core phenotype model (reported in Deliverable 3.5) and subsequent work with the mouse, epidemiology, ontology and bio-banking communities has provided areas for extension which are described here.


3. Basic LSDB background information

The meta data describing a particular LSDB instance is of considerable importance for data exchange use cases. Specifically for diagnostic labs this information is required to establish trust in exchanged data. It is not enough to just share LSDB data content, as only a subset of available data is of interest to diagnostics labs. Some of the requirements include: up-to-dateness, sufficient cross-referencing to external resources, and well established IP rights that permit intended data usage. A data set without this accompanying information is practically useless for data consumers.

A list of obligatory, recommended and optional LSDB background information is provided in Appendix 1. It has also been summarized in table 1 below.

Table 1. Basic LSDB background information

ITEM	REQUIRED	EXAMPLE
Gene name	Yes	Calpain-3
Gene symbol	Yes	CAPN3
Chromosome Location	Yes	15q15.1-q21.1
Database location	Yes	http://www.DMD.nl
Curator(s)	Yes	Johan den Dunnen and Jacqui Beckmann
Date of creation	Yes	1997, January 10
Last update	Yes	2008, November 22
Version	Yes	CAPN3 081122
Reference sequence	Yes	LRG reference
Copyright & Disclaimer	Yes	<file>
Database policy	Yes	<file>
Link to registration	Recommended	
Link to variant submission	Recommended	
Link to query options	Recommended	
Collected data statistics	Recommended	number of unique DNA variants, number of individuals with variant(s) reported, number of variants reported
Sequence variant summary tables	Optional	
Links to other resources	Optional	Gene homepage

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0

ITEM	REQUIRED	EXAMPLE
		(http://www.DMD.nl/capn3_home.html) Entrez Gene, OMIM, HGMD

4. LSDB data content standards


During the second GEN2PHEN Modelling Workshop (Helsinki, January 19-22, 2009, hosted by UH.FGC) Partners and invited experts discussed the LSDB minimal content provided by LUMC Partner with feedback from HGVS community. Detailed proceedings are available in Appendix 3, as well as online from:

<http://askja.gene.le.ac.uk/drupal5/content/lsdb-minimal-requirements>


A list of obligatory, recommended and optional LSDB minimal data content is provided in Appendix 2. It has also been summarized in table 2 below. All variant descriptions have to follow the current HGVS recommendation and be verifiable by Mutalyzer software [2].

Table 2. Minimal LSDB data content

FIELD	REQUIRED	DESCRIPTION	EXAMPLE
Variant/Exon	Recommended	Number of exon where the variant was found	01 or 01e = exon number 1
Variant/DNA_genomic	Yes	Description of variant on genomic DNA level	g.456A>G
Variant/DNA_coding	Recommended	Description of variant on coding DNA level	c.123C>T
Variant/RNA	Yes	Description of variant on RNA level	r.123c>u
Variant/Protein	Yes	Description of variant on protein level	p. Pro123Arg
Variant/DBID	Yes	Facilitates direct links to record in LSDB	DMD_00123
Variant/Reference	Yes	Information on where the variant was described	Smith et al. 2008 (PMID012345), OMIM (60123.0001)
Variant/DNA_published	Recommended	Description of variant as published in Variant/Reference	521delT
Variant/Detection/Template	Yes	Template(s) used to detect the variant	DNA

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0 8/13

FIELD	REQUIRED	DESCRIPTION	EXAMPLE
Variant/Detection/Technique	Yes	Technique(s) used to detect the variant	DGGE, see 1.1 in Appendix 2 for a full list
Variant/DNA_remark	Recommended	Any worthwhile information not collected specifically in another Variant-field	
Variant/Frequency	Recommended	Frequency of the identified variant	23/400 variant found 23 times in 400 chromosomes analysed
Variant/Origin	Recommended	Inheritance of the variant	in vitro, see 1.2. in Appendix 2 for a full list
Variant/Restriction_site	Optional	Information on restriction enzyme recognition sites created or destroyed by the variant	BamHI-
Variant/Allele	Recommended	Information on the parent from which the variant was inherited	Parent #1
Variant/Pathogenicity	Recommended	Pathogenicity of the variant	No known pathogenicity
Patient/Patient_ID	Yes	Lab ID-code for the patient, used by submitters to unequivocally recognize individual patients	LGF20080099
Patient/Phenotype/Disease	Yes	Phenotype of the patient based on which DNA diagnosis was initiated (does NOT store the 'concluded' phenotype after a causative variant was detected)	DMD, OMIM term
Patient/Remarks	Recommended	Any additional information not collected specifically in another patient field, when not collected through individual columns, this field contains a description of phenotype	

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0 9/13

FIELD	REQUIRED	DESCRIPTION	EXAMPLE
Patient/Origin/Geographic	Recommended	Geographic origin of parents/family	AF = Afghanistan See 1.3 in Appendix 2 for a full list
Patient/Origin/Ethnic	Recommended	Ethnic origin of parents/family	Aboriginal
Variant/Origin/Ethnic	Optional	Similar to previous one, since parents may have different ethnic origin	Aboriginal
Patient/Gender	Recommended	Patient's gender	? = unknown, F = female, M = male
ID_submitterid	Yes	Submitter's id	001

LSDB data content standards were fully implemented and all the fields (required, recommended, optional) are already supported by the LSDB model reported previously in deliverable report 'D3.2 Development of High-Level Domain Model Version 1'.


There is some crossover with the GEN2PHEN Phenotype Model reported in deliverable report 'D3.5 High-Level Domain Model Version 2 with Sample Phenotype Focus'. Overlapping fields, which could potentially be implemented as a phenotype extension, are listed in table 3 below. Though a general recommendation is rather than to extend existing models, is to make them interoperable as the data may not all be in one place, available from a single resource.

Table 3. LSDB and Phenotype models overlap

LSDB Model	GEN2PHEN Phenotype Model
<i>Patient/Gender</i>	Attribute of <i>Individual</i> class
<i>Variant/Frequency</i>	Implemented as <i>InferredValue</i> for <i>Panel</i> or <i>Individual</i> classes
<i>Variant/Pathogenicity</i> <i>Patient/Phenotype/Disease</i>	<i>InferredValue</i> and <i>ObservableFeature</i>

5. Security considerations

There are legal and ethical security issues when dealing with human identifiable data. Legal issues on sharing human research data vary from country to country and are covered by country-specific data protection/privacy laws. Within the European Union these are governed by the Data Protection Directive (officially Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data), which is a European Union directive regulating the processing of personal data[3]. Under this directive personal data

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson	Version: v6.0	10/13


can only be processed when certain conditions of transparency, legitimate purpose and proportionality are met.

The Declaration of Helsinki, developed by the World Medical Association (WMA), is a paramount document for the ethical aspects of data sharing. It is a statement of ethical principles for medical research involving human subjects, including research on identifiable human material and data[4]. In particular two paragraphs, cited here in full, are of interest:

23. *Every precaution must be taken to protect the privacy of research subjects and the confidentiality of their personal information and to minimize the impact of the study on their physical, mental and social integrity.*
25. *For medical research using identifiable human material or data, physicians must normally seek consent for the collection, analysis, storage and/or reuse. There may be situations where consent would be impossible or impractical to obtain for such research or would pose a threat to the validity of the research. In such situations the research may be done only after consideration and approval of a research ethics committee.*

It has been a long tradition to release all research data as widely as possible to the public and among the research community. Individual level genetic data has also been distributed publicly, e.g. the data from HapMap (www.hapmap.org) or 1000 Genomes (www.1000genomes.org) projects. In each of these projects, participants are required to give freely an informed consent that allows further processing and distribution of their data. However, the consents acquired typically in a biomedical research projects have restrictions on the information management and dissemination. Furthermore, the paper by Homer et al.[1] showed that individuals can be located into a particular research cohort based on the publicly available aggregate data on the internet and acquired individual level genotypes, which may result in a violation of the informed consent. The paper changed the landscape of publicly available data from the biomedical research projects by moving the aggregate data under the authorized access and highlighting the importance of the applied security protocols for all data in such projects. Although one could say that stringent data privacy policies stifle biomedical research, these guidelines are in place to prevent abuse. For a more comprehensive view on the subject please see WP1 deliverable report ‘D1.3 – Report on General Ethical Issues in G2P Database Work’. Participants from Inserm participated in the ethics part of this activity (A. Pigeon, E. Rial-Sebbag and A. Cambon-Thomsen).

At the European Bioinformatics Institute all genetic and phenotypic data, consented for non-public distribution, is archived and distributed using the EGA security protocols. The EGA system includes a secure computing facility for data processing and a modularized archive model to provide high security and optimal performance for large data sets. As an example, the information about individuals and their phenotypes are stored in separate databases. The connections between these databases are made using abstract identifiers. The EGA is available at <http://www.ebi.ac.uk/ega>.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0

The EGA includes both authorized and public data for the archived data sets. As an example the published SNPs showing significant association for the studied phenotype are made publicly available at the EGA website, whereas the individual level data, and any aggregate data provided by the study authors, require authorized access. The EGA works closely with data access-granting organisation (DAO) that is normally formed from the same organisation that collated the original data. In all cases data access decisions are made by the appropriate DAO and not by the EGA. The EGA accepts only data that include a DAO-approved access plan and supports data access decisions that are based on ethics and respect the informed consent. The data access applications are made to the appropriate DAO and the data access agreement (DAA) is signed directly between the DAO and applicant. The DAA dictates how data can be stored, transferred and analyzed once downloaded from the EGA system. Therefore, the role of the EGA is to provide a secure way to distribute data that otherwise would not reach those researchers and clinicians authorised to view the data. The EGA experience dealing with secure information storage and dissemination has proved to be invaluable for the GEN2PHEN consortium.

6. Phenotype extensions


A workshop hosted by SIB partner was held in Geneva, May 7-8 2009. It was agreed that reporting of the phenotypes is inconsistent and should be improved. For example only some of the observation targets are annotated with *ultrasound of the liver was significant in one of the subjects*, but no information is given for other observation targets. Thus it is unclear whether they have also been tested. There are also a number of ethical ramifications which will be followed up in the ethics activity of WP1.

It was agreed that minimal information should be content specific, e.g. obligatory smoking status in reporting of hypertension, but published phenotypic information should at least contain the following information about observation targets:

- Age
- Gender
- Age of disease onset
- Use of an ontology (controlled vocabulary) term for signs and symptoms

Optionally therapy information should be included as well, although it was acknowledged that ontology coverage is coming short in this domain.

All those cases are supported by the GEN2PHEN Phenotype Model reported in deliverable report ‘D3.5 High-Level Domain Model Version 2 with Sample Phenotype Focus’. We are also beginning to evaluate if the model supports other phenotype use cases, e.g. DataSHaPER questions (www.p3gobservatory.org/datashaper/presentation.htm), use of ontologies to describe data, in particular Human Phenotype Ontology (HPO)[5] within GEN2PHEN, as well as prototype metabolic syndrome extensions of the Experimental Factor Ontology (EFO www.ebi.ac.uk/efo). Those efforts will be subsequently reported on.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0

7. FUTURE PLANS

7.1. Derivation and Specification of Exchange Format (D3.7)

The priorities for data formats in GEN2PHEN are the data exchange between locus specific databases and central repositories and HTP data. The modelling work to date has separated these domains to support immediate needs for data exchange.

Validation of LSDB data model commenced in 2009 by working with the existing LSDBs inside and outside the GEN2PHEN consortium, most of who have existing data formats. The LSDB minimal content reported here will be used to align and extend existing formats and to extend the existing models where necessary,

Validation of the MAGE-TAB OM is under way and progress is promising. We envisage that the Phenotypic descriptors, e.g. membership of a cohort through a shared phenotype, or trait will require an extension of MAGE-TAB, and the requirement to provide details of markers in context of HTP data will also require an extension.

Interoperability of models is key to GEN2PHEN and the requirements of the LSDB community will be tested using the Findis LSDB implementation generated with Molgenis by UH.FGC.

7.2. WP4 – Genetics G2P Databases


Data models provided by WP3 recommendations reported in this deliverable report will underpin this data gathering, import, curation, and data query capabilities implemented to support the developed LSDBs. Subsequently existing UMD and LOVD packages will adapt their databases and support software accordingly. These two major platforms will thus become fully interoperable and increasingly merge on the data-structure and data-exchange levels. A federation of LSDBs will thus be supported and increasingly unified. Data ownership/control will be solved with input from WP1.

7.3. New technologies

High-throughput sequencing, e.g. pyrosequencing, sequencing by ligation will dramatically change the available amount of human identifiable data. This will in turn bring new issues in data security, storage and modelling. WP3 work will continue upon providing components for flexible and future-proof database implementations.

7.4. Security

The EBI case study, provided here, is an example of how human identifiable data can be managed in the current climate where very little data can be made freely available. There is clearly work to be done evaluating other models and there will be ethical recommendations made on the basis of future work. Of key importance is the area of data exchange scenarios within GEN2PHEN and with external organisations for data and cases, where meta data only is

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Ilkka Lappalainen, Helen Parkinson		Version: v6.0


exchanged. An ethical working group was recently held with the GEN2PHEN Partners, where many of these issues were raised. This work will be continued by the INSERM Partner.

8. Abbreviations

EGA	European Genome-phenome Archive
HGVS	Human Genome Variation Society
LSDB	Locus Specific Database

REFERENCES

1. Homer, N., et al., *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays*. PLoS Genet, 2008. **4**(8): p. e1000167.
2. Wildeman, M., et al., *Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker*. Hum Mutat, 2008. **29**(1): p. 6-13.
3. *Data Protection in the European Union*. [cited 2009 July 12]; Available from: http://ec.europa.eu/justice_home/fsj/privacy/index_en.htm.
4. *WORLD MEDICAL ASSOCIATION DECLARATION OF HELSINKI Ethical Principles for Medical Research Involving Human Subjects*. [cited 2009 July 12]; Available from: <http://www.wma.net/e/policy/b3.htm>.
5. Robinson, P.N., et al., *The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease*. Am J Hum Genet, 2008. **83**(5): p. 610-5.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix I - LSDB background information		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	1/1

Appendix I - LSDB background information

1. LSDB background information to be supplied

Obligatory

ITEM	EXAMPLE
Gene name	Calpain-3
Gene symbol	CAPN3
Chromosome Location	15q15.1-q21.1
Database location	http://www.DMD.nl
Curator(s)	Johan den Dunnen and Jacqui Beckmann
Date of creation	1997, January 10
Last update	2008, November 22
Version	CAPN3 081122
Reference sequence	link to coding DNA reference sequence for describing sequence variants
Copyright & Disclaimer	<file>
Database policy	<file>

Recommended

Links to;
registration, variant submission, query options, etc.

Totals on data collected, e.g. number of unique DNA variants, number of individuals with variant(s) reported, number of variants reported, etc.


Optional

Sequence variant summary tables

Links to other resources

e.g. Gene homepage (http://www.DMD.nl/capn3_home.html)

Entrez Gene, OMIM, HGMD, etc.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final

Appendix II - LSDB minimal data content

1. LSDB data to be collected

The format chosen is =

{{ Name of field }}

field is obligatory, recommended, optional

purpose of field with short description

example entry

{{ Variant/Exon }}

recommended

contains redundant information (can be derived from {{ Variant/DNA }})

may cause confusion (for many genes opinions on exon number do not agree)

highly appreciated by clinical geneticists

purpose = stores the number of the exon where the variant was found

dependent on reference sequence field

example description

01 or 01e = exon number 1

02i = intron 2

03_07 = exons 3 to 7

{{ Variant/DNA_genomic }}

obligatory

essential information

purpose = stores a description of the variant at genomic DNA level

following HGVS recommendations

dependent on reference sequence field

example description

g.456A>G

{{ Variant/DNA_coding }}


recommended

basically redundant information (already covered in {{ Variant/DNA_genomic }}) but highly appreciated by users (represents the description they are most familiar with)

purpose = stores a description of the variant at coding DNA level

following HGVS recommendations

dependent on reference sequence field

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final

example description
c.123C>T

{{ Variant/RNA }}

obligatory

essential information

indicates whether RNA was analysed and thus quality of prediction

purpose = stores a description of the variant at RNA level

following HGVS recommendations

dependent on reference sequence field

example description

r.123c>u

r.? = RNA not analysed

{{ Variant/Protein }}

obligatory

redundant information

can be derived from {{ Variant/RNA }} or indirectly from

{{ Variant/DNA }}

usually predicted from {{ Variant/DNA }}

highly appreciated by clinical geneticists

purpose = stores a description of the variant at protein level

following HGVS recommendations

dependent on reference sequence field

example description

p. Pro123Arg

{{ Variant/DBID }}

obligatory

purpose = stores unique identifier for {{ Variant/DNA }}


facilitates direct links to record in LSDB

example description

DMD_00123

{{ Variant/Reference }}

obligatory

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final

purpose = stores information on where the variant was described (publication, poster congress) and/or who submitted the variant as well as links to other databases describing the variant (e.g. OMIM, dbSNP, etc.)

example description

Smith et al. 2008 (PMID012345)

Smith, Clinical Genetics, LUMC, Leiden, NL

dbSNP (rs012345)

OMIM (60123.0001)

{{ Variant/DNA_published }}

recommended

information from older publications not always available

purpose = stores description of the variant as published in {{ Variant/Reference }}

example description

521delT

Pro123Arg

{{ Variant/Detection/Template }}

obligatory

purpose = stores information on the template(s) used to detect the variant

example description

DNA, RNA, protein, ? (unknown)

{{ Variant/Detection/Technique }}

obligatory

purpose = stores information on the technique(s) used to detect the variant

example description

suggest to provide an example selection list (see 1.1)

DGGE = Denaturing-Gradient Gel-Electrophoresis

SEQ = SEQuencing

Southern = Southern blotting

...etc.

{{ Variant/DNA_remark }}


recommended

purpose = stores any worthwhile information not collected specifically in another

Variant-field

example description

free text

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final

{{ Variant/Frequency }}

recommended

purpose = stores information on the frequency the variant was found
usually used only for non-pathogenic variants

example description

23/400 = variant found 23 times in 400 chromosomes analysed

{{ Variant/Origin }}

recommended

purpose = stores information on the inheritance of the variant

example description

? (unknown), in vitro, familial, sporadic, de novo, ...

{{ Variant/Restriction_site }}

optional

redundant, can be derived from {{ Variant/DNA }}

highly appreciated by clinical geneticists

purpose = stores information on restriction enzyme recognition sites created or
destroyed by the variant

example description

BglII+ = creates a BglII site

BamHI- = destroys BamHI site

{{ Variant/Allele }}

recommended

essential in imprinted disorders

essential in recessive disorders (proves pathogenic variant inherited from both parents)

purpose = stores information on the parent from which the variant was inherited


example description

drop-down list

0 = Unknown, 1 = Parent #1, 2 = Parent #2, 10 = Paternal (inferred), 11 =

Paternal (confirmed), 20 = Maternal (inferred), 21 = Maternal
(confirmed)

{{ Variant/Pathogenicity }}

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	5/15

recommended
helpful for drawing conclusions on pathogenicity
purpose = stores information on the pathogenicity of the variant
example description
drop-down list
No known pathogenicity (-) , Probably no pathogenicity (-?), Unknown (?),
Probably pathogenic (+?), Pathogenic (+)
LOVD uses double descriptions = as published / acc. to curator(s) (e.g. + / ?)

{{ Patient/Patient_ID }}

obligatory

purpose = stores the lab ID-code for the patient, used by submitters to unequivocally recognize the individual patient

example description

LGF20080099

{{ Patient/Phenotype/Disease }}

obligatory

purpose = stores the phenotype of the patient based on which a DNA diagnosis was initiated (does NOT store the 'concluded' phenotype after a causative variant was detected)

example description

preferably using OMIM abbreviations
e.g. DMD, BMD, LGMD-2G, ...

{{ Patient/Remarks }}

recommended

purpose = stores any worthwhile information not collected specifically in another Patient field

when not collected through individual columns, this field contains a description of the phenotype

example description


free text

{{ Patient/Origin/Geographic }}

recommended

purpose = stores information on the geographic origin of the parents/family

example description

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	6/15

drop-down selection list (see 1.3, ISO 3166-1 alpha-2 International Organization for Standardization list)

AF = Afghanistan, AL = Albania, DZ = Algeria, AS = American Samoa, ...

Preference (see 1.3): Afghanistan=AF, ÅLAND ISLANDS=AX,

ALBANIA=AL, ALGERIA=DZ, etc.

NOTE - since parents may have different geographical origins the header {{ Variant/Origin/Geographic }} seems more appropriate

{{ Patient/Origin/Ethnic }}

recommended

purpose = stores information on the ethnic origin of the parents/family

example description

drop-down selection list (NOTE - a selection list is highly desired)

Aboriginal, ...

NOTE - since parents may have different ethnic origins the header {{ Variant/Origin/Ethnic }} seems more appropriate

{{ Patient/Gender }}

recommended

essential for X-linked or imprinted disorders

purpose = stores information on the gender of the patient

example description

drop-down selection list

? = unknown, F = female, M = male

{{ ID_submitterid_ }}

obligatory

purpose = stores information on the submitter

example description

001, 002, 003, ..

1.1. Selection list for {{ Variant/Detection/Technique }}

arrayCGH = array Comparative Genomic Hybridisation


arraySEQ = array resequencing

arrayCNV = array Copy Number Variation (SNP array, CNV array)

BEES = Base Excision Sequence Scanning

CMC = Chemical Mismatch Cleavage


CSCE = Conformation Sensitive Capillary Electrophoresis

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	7/15


ddF = dideoxy Fingerprinting
DGGE = Denaturing-Gradient Gel-Electrophoresis
DHPLC = Denaturing High-Performance Liquid Chromatography
DOVAM = Detection Of Virtually All Mutations (SSCA variant)
DSCA = Double-Strand DNA Conformation Analysis
EMC = Enzymatic Mismatch Cleavage
HD = HeteroDuplex analysis
IHC = Immuno-Histo-Chemistry
mPCR = multiplex PCR
MAPH = Multiplex Amplifiable Probe Hybridisation
MCA = high-resolution Melting Curve Analysis (hrMCA)
MLPA = Multiplex Ligation-dependent Probe Amplification
Northern = Northern blotting
PAGE = Poly-Acrylamide Gel-Electrophoresis
PCR = Polymerase Chain Reaction
PCRDig = PCR + restriction enzyme digestion
PFGE = Pulsed-Field Gel-Electrophoresis (+Southern)
PTT = Protein Truncation Test
RT-PCR = Reverse Transcription and PCR
SEQ = SEQuencing
Southern = Southern blotting
SSCA = Single-Strand DNA Conformation polymorphism Analysis (SSCP)
SSCAf = fluorescent SSCA (SSCP)
TaqMan = TaqMan assay
Western = Western Blotting

1.2. Selection list for {{ Variant/Origin }}

? (unknown)
in vitro
familial
sporadic
sporadic, consanguineous parents
sporadic, non-consanguineous parents
sporadic, consanguinity parents?
de novo
de novo, somatic mosaicism
de novo, germline mosaicism
de novo, germline and somatic mosaicism
de novo, in patient
de novo, in patient (maternal allele)
de novo, in patient (paternal allele)

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	8/15


de novo, in mother
de novo, in mother (grandmaternal allele)
de novo, in mother (grandpaternal allele)
de novo, in father
de novo, in father (grandmaternal allele)
de novo, in father (grandpaternal allele)

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	9/15


1.3. Selection list for {{Patient/Origin/Geographic }}

This list states the country names (official short names in English) in alphabetical order as given in ISO 3166-1 and the corresponding ISO 3166-1-alpha-2 code elements. The list is updated whenever a change to the official code list in ISO 3166-1 is effected by the ISO 3166/MA. It lists 240 official short names and code elements. One line of text contains one entry. A country name and its code element are separated by a semicolon (=).


AFGHANISTAN = AF
ÅLAND ISLANDS = AX
ALBANIA = AL
ALGERIA = DZ
AMERICAN SAMOA = AS
ANDORRA = AD
ANGOLA = AO
ANGUILLA = AI
ANTARCTICA = AQ
ANTIGUA AND BARBUDA = AG
ARGENTINA = AR
ARMENIA = AM
ARUBA = AW
AUSTRALIA = AU
AUSTRIA = AT
AZERBAIJAN = AZ
BAHAMAS = BS
BAHRAIN = BH
BANGLADESH = BD
BARBADOS = BB
BELARUS = BY
BELGIUM = BE
BELIZE = BZ
BENIN = BJ
BERMUDA = BM
BHUTAN = BT
BOLIVIA = BO
BOSNIA AND HERZEGOVINA = BA
BOTSWANA = BW
BOUVET ISLAND = BV
BRAZIL = BR
BRITISH INDIAN OCEAN TERRITORY = IO
BRUNEI DARUSSALAM = BN

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	10/15


BULGARIA = BG
BURKINA FASO = BF
BURUNDI = BI
CAMBODIA = KH
CAMEROON = CM
CANADA = CA
CAPE VERDE = CV
CAYMAN ISLANDS = KY
CENTRAL AFRICAN REPUBLIC = CF
CHAD = TD
CHILE = CL
CHINA = CN
CHRISTMAS ISLAND = CX
COCOS (KEELING) ISLANDS = CC
COLOMBIA = CO
COMOROS = KM
CONGO = CG
CONGO, THE DEMOCRATIC REPUBLIC OF THE = CD
COOK ISLANDS = CK
COSTA RICA = CR
CÔTE D'IVOIRE = CI
CROATIA = HR
CUBA = CU
CYPRUS = CY
CZECH REPUBLIC = CZ
DENMARK = DK
DJIBOUTI = DJ
DOMINICA = DM
DOMINICAN REPUBLIC = DO
ECUADOR = EC
EGYPT = EG
EL SALVADOR = SV
EQUATORIAL GUINEA = GQ
ERITREA = ER
ESTONIA = EE
ETHIOPIA = ET
FALKLAND ISLANDS (MALVINAS) = FK
FAROE ISLANDS = FO
FIJI = FJ
FINLAND = FI
FRANCE = FR
FRENCH GUIANA = GF

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final


FRENCH POLYNESIA = PF
FRENCH SOUTHERN TERRITORIES = TF
GABON = GA
GAMBIA = GM
GEORGIA = GE
GERMANY = DE
GHANA = GH
GIBRALTAR = GI
GREECE = GR
GREENLAND = GL
GRENADA = GD
GUADELOUPE = GP
GUAM = GU
GUATEMALA = GT
GUERNSEY = GG
GUINEA = GN
GUINEA-BISSAU = GW
GUYANA = GY
HAITI = HT
HEARD ISLAND AND MCDONALD ISLANDS = HM
HOLY SEE (VATICAN CITY STATE) = VA
HONDURAS = HN
HONG KONG = HK
HUNGARY = HU
ICELAND = IS
INDIA = IN
INDONESIA = ID
IRAN, ISLAMIC REPUBLIC OF = IR
IRAQ = IQ
IRELAND = IE
ISLE OF MAN = IM
ISRAEL = IL
ITALY = IT
JAMAICA = JM
JAPAN = JP
JERSEY = JE
JORDAN = JO
KAZAKHSTAN = KZ
KENYA = KE
KIRIBATI = KI
KOREA, DEMOCRATIC PEOPLE'S REPUBLIC OF = KP
KOREA, REPUBLIC OF = KR

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final


KUWAIT = KW
KYRGYZSTAN = KG
LAO PEOPLE'S DEMOCRATIC REPUBLIC = LA
LATVIA = LV
LEBANON = LB
LESOTHO = LS
LIBERIA = LR
LIBYAN ARAB JAMAHIRIYA = LY
LIECHTENSTEIN = LI
LITHUANIA = LT
LUXEMBOURG = LU
MACAO = MO
MACEDONIA, THE FORMER YUGOSLAV REPUBLIC OF = MK
MADAGASCAR = MG
MALAWI = MW
MALAYSIA = MY
MALDIVES = MV
MALI = ML
MALTA = MT
MARSHALL ISLANDS = MH
MARTINIQUE = MQ
MAURITANIA = MR
MAURITIUS = MU
MAYOTTE = YT
MEXICO = MX
MICRONESIA, FEDERATED STATES OF = FM
MOLDOVA, REPUBLIC OF = MD
MONACO = MC
MONGOLIA = MN
MONTENEGRO = ME
MONTSERRAT = MS
MOROCCO = MA
MOZAMBIQUE = MZ
MYANMAR = MM
NAMIBIA = NA
NAURU = NR
NEPAL = NP
NETHERLANDS = NL
NETHERLANDS ANTILLES = AN
NEW CALEDONIA = NC
NEW ZEALAND = NZ
NICARAGUA = NI

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	13/15


NIGER = NE
NIGERIA = NG
NIUE = NU
NORFOLK ISLAND = NF
NORTHERN MARIANA ISLANDS = MP
NORWAY = NO
OMAN = OM
PAKISTAN = PK
PALAU = PW
PALESTINIAN TERRITORY, OCCUPIED = PS
PANAMA = PA
PAPUA NEW GUINEA = PG
PARAGUAY = PY
PERU = PE
PHILIPPINES = PH
PITCAIRN = PN
POLAND = PL
PORTUGAL = PT
PUERTO RICO = PR
QATAR = QA
REUNION = RE
ROMANIA = RO
RUSSIAN FEDERATION = RU
RWANDA = RW
SAINT BARTHÉLEMY = BL
SAINT HELENA = SH
SAINT KITTS AND NEVIS = KN
SAINT LUCIA = LC
SAINT MARTIN = MF
SAINT PIERRE AND MIQUELON = PM
SAINT VINCENT AND THE GRENADINES = VC
SAMOA = WS
SAN MARINO = SM
SAO TOME AND PRINCIPE = ST
SAUDI ARABIA = SA
SENEGAL = SN
SERBIA = RS
SEYCHELLES = SC
SIERRA LEONE = SL
SINGAPORE = SG
SLOVAKIA = SK
SLOVENIA = SI

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Johan T. den Dunnen		Version: v1.0 Final

SOLOMON ISLANDS = SB
SOMALIA = SO
SOUTH AFRICA = ZA
SOUTH GEORGIA AND THE SOUTH SANDWICH ISLANDS = GS
SPAIN = ES
SRI LANKA = LK
SUDAN = SD
SURINAME = SR
SVALBARD AND JAN MAYEN = SJ
SWAZILAND = SZ
SWEDEN = SE
SWITZERLAND = CH
SYRIAN ARAB REPUBLIC = SY
TAIWAN, PROVINCE OF CHINA = TW
TAJKISTAN = TJ
TANZANIA, UNITED REPUBLIC OF = TZ
THAILAND = TH
TIMOR-LESTE = TL
TOGO = TG
TOKELAU = TK
TONGA = TO
TRINIDAD AND TOBAGO = TT
TUNISIA = TN
TURKEY = TR
TURKMENISTAN = TM
TURKS AND CAICOS ISLANDS = TC
TUVALU = TV
UGANDA = UG
UKRAINE = UA
UNITED ARAB EMIRATES = AE
UNITED KINGDOM = GB
UNITED STATES = US
UNITED STATES MINOR OUTLYING ISLANDS = UM
URUGUAY = UY
UZBEKISTAN = UZ
VANUATU = VU
VENEZUELA = VE
VIET NAM = VN
VIRGIN ISLANDS, BRITISH = VG
VIRGIN ISLANDS, U.S. = VI
WALLIS AND FUTUNA = WF
WESTERN SAHARA = EH

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix II - LSDB minimal data content		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Johan T. den Dunnen	Version: v1.0 Final	15/15

YEMEN = YE
 ZAMBIA = ZM
 ZIMBABWE = ZW

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Helen Parkinson	Version: v1.0 Final	1/8

Appendix III - Record of discussion and comments

Discussion with the GEN2PHEN Partners on the LSDB minimal content requirements

Record of discussion and comments.

Common themes. A common way of representing unknown, no data, not assayed etc with some definitions etc would be useful throughout this document and also for data exchange purposes.

For a list of participants see the workshop report at http://askja.gene.le.ac.uk/drupal5/Modelling_Workshop_2_Report

Variant/Exon (number) - recommended.

What to do when there are multiple transcripts? Must be in the context of the LRG, or other reference sequence.

Exon/introns are both relevant - there is no good word for this. Is a region of DNA and the role - exon or intron. Or 'splice transcript component' has been used. This is redundant info, people use to sort the data (e.g. in LOVD useful), sort on exon name is convenient. Re: LRGs - lab staff want this as this is how people think.

Mauno: This is redundant and can get this programmatically RNA/protein (Christophe thinks that this is usually predictive for proteins which are functionally assayed rather than sequenced)

Ivo: depends whether this is predicted or experimental info

Mauno: suggest that we add some info in the quality and source of the data - assayed/prediction.

Ivo: HGVS includes this, use parenthesis.

Christophe: exon not needed for data exchange, is useful for diagnostic use, for data exchange we can regenerate it.

Ray: 0-1 or 0-1e, second is redundant. Default is that if no 'i' included we assume this is exonic.


Andy: The 'i' is a problem as it changes the field from an integer to a string - in LOVD is text anyway. If this becomes std then this is fine. We have an integer, we don't add intron/exon as it is clear from the cDNA nomenclature.

Christophe: if you report the mutation at the cDNA, if genomic there is no transcript info - not useful. Sentence could be added to reflect this.

Jan: why not store genomic position?

Andy: genomic seqs change too much and are hard to work with as the numbers are too large.

Ray: if you have cDNA coords for splice site mutations, HGVS nomenclature is clear that this is donor/acceptor site not clear in genomic coordinates.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Helen Parkinson	Version: v1.0 Final	2/8

Christophe: Still use locally in UMD this info - but here we are to discuss exchange of data. Think about the minimum.

Mauno: different levels are important cDNA, genomic and protein, we use all of them.

Variant/DNA genomic - obligatory

Agreed.

Mummi: would be nice to do this from a genomic view rather than a reference sequence

Ray: as genome assemblies are revised these change. LRG will provide stability, and maps to an assembly is handled outside this.

Fiona: The other way is to do as in dbSNP with the flanking region, always would need to be remapped.

Variant/DNA_coding - recommended.

Andy: should we have obligatory DNA_genomic OR DNA_coding, not obligatory genomic and recommended DNA coding.

Helen: depends on the reference sequence.

Ray: I use genomic reference seq and present in terms of cDNA coordinates.

Fiona: is that the norm?

Ivo: in LOVD we host, there are no genes specifying genomic.

Christophe: same for UMD. But we can translate these using LRG. Again it's about exchange so redundant.

Fiona: we can translate after exchange if that will make the data flow easier.

Christophe: LRG was designed for genomic reference sequences, we can translate.

Ray: intended to move to LRG ref sequences. But still also have cDNA coordinates in the context of viewing.

Fiona: will contain the cDNA and genomic seq so is the link between the two.

Ray: avoid some of both, be consistent.

Helen: what will contain the cDNA and genomic seq so is the link between the two.

Christophe: we will just add another field, we will keep the cDNA.

Variant RNA - obligatory

Mauno: issue where evidence should be provided. Predicted/assayed.


Tomasz: how often filled?

heikki: redundancy can help catch typos. Before validate the protein need the RNA variant, based on the reference sequence.

Chrsit: can report multiple splice variants, needs these.

Ray: B globin, is an AA substitution and a splice site. Get some globin with AA subs and other additional frame shift transcripts. One too many needed here.

Ray: is null allowed - not analysed. Needs to be clarified what null means. Ask Johan to be clearer all the way through the document.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Helen Parkinson	Version: v1.0 Final	3/8

Variant/protein - obligatory

Andy: how tell difference between observed and predicted proteins.

Ivo: most people don't test on protein, in LOVD usually without proof and then people don't use brackets. Can check detection technique and see how assayed.

Andy: some genes we have this and is predicted.

Ray: 3 letter AA code is useful. Ask Johan if the three letter code is mandatory, thinks is more useful, or if single letter code is also allowed. Are both equally valid?

Andy: we need to be clear what 'not analysed' is across all fields - i.e. difference between no value and a? Needs a validation code as does RNA predicted/assayed.

Variant DBID - Obligatory

We think database source plus the dbid is needed and also version if specified at source database.

Mike: does that mean that each instance of the same variant has the same id.

Ivo: In LOVD it does, will get multiple patients.

Christophe: when we talk about this was to get a link back to the LSDB, can put anything you want. If you add this field will get a link to the patients with this mutation. Not designed as a unique id to a record. used as a link to get to patients with this record.

Variant/reference - obligatory

Pubmed id or DOI or dbSNP, OMIM? Need to be clear that there may be several references and the more the better

Heikki: this is merging two things, a database reference and a citation. dbSNP etc is a cross ref to a variant. Suggest that this is called bibliographic id and we split this from a database reference. May be many of both.

Rasko: Also seen that pubmed etc are treated as citations - these are just xrefs.

Mummi: you still need to resolve this so need a service.

suggest that we have:

variant/database

variant/bibref

and that use both recommended

Christophe: do we need dbSNP when we can get this manually?

Helen: What to do when a dbSNP doesn't map to ensembl?


Fiona: we have failed maps as well so can map to that.

Fiona: do we need both?

Ray: dbSNP maps to a specific substitution, a pubmed id maps to many. They are different things. *Helen:* if you were modelling this would split according to cardinalities citations are different from the dbxrefs, get multiple citations as well per variant.

Juha: do we need to acknowledge the first paper?

Mauno: need first and subsequent.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Helen Parkinson	Version: v1.0 Final	4/8

Ray: sometimes cite two papers for a single mutation. and first publication is simply a list - need to consider when modelling.

Helen: could reference OMIM for that.

Ray: if they are in OMIM.

Mauno: does OMIM ever publish OMIM that's not in a paper? Is a secondary reference.

Heikki: doesn't contain anything not published.

Ray: if this is the only place where we can add external references this is variant and patient reference. Could be split elsewhere.

Variant/DNA published - recommended

➔ optional Better name for this 'legacy description'

Legacy numbering systems are in the LRG specification, therefore not essential for exchange, suggest optional.

Mauno: not useful. No indication what was the reference for the paper.

Ray: counter argument: when people started numbering aa in collagens decided to number these at 1, at first glycine of triple helical region. Ignored upstream stuff. We now number from initiation M of translated product. Literature are described old way, added a custom field for the legacy numbering system. Without that confusion on a reader who looks at a paper.

Christophe: same experience with CFTR - should we reproduce this, or should we go forward to the correct system. I moved people to the new nomenclature don't want to go back is a concern.

Ray: feel same way. Want to use the new system, at the OI conference, people were not happy.

Andy: CFTR is a good example, clinicians use old labels, we need this. BIC doesn't use HGVS nomenclature, would want to label maps to old system

Ray: suggest change this to optional.

Fiona: people are using old systems still? YES would want to label maps to old system *Ray:* suggest change this to optional.

Fiona: In LRG we can add other naming schemes.

Morris: This can be a multiple reference.

Andy: this is not a clear name, not clear what it's there for - other or legacy is better.

Fiona: other naming is used in LRG.


Christophe: some people are using the wrong reference sequence and made an error in the case where you know it's wrong in the legacy should this be propagated?

Variant/Detection/Template - obligatory

Ray: template and method can be missing common in human mutation papers.

WE NEED A COMMON WAY OF DEFINING UNKNOWN across this domain.

Ray: we have not defined in OI database. Is always a value in that case.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Helen Parkinson	Version: v1.0 Final	5/8

Ivo: this is important, if RNA was analysed, provides pathogenicity importance, opinion on this has additional evidence

Christophe: then you will have this in the naming system for RNA nomenclature. People want to know if the whole gene or a subset of the gene was tested. This is different info. RNA is not often sequenced. Useful info may be how much of the sequence has been read. p53 - common exon, then bias in the info. This is redundant with nomenclature.

Ray: If describe at the cDNA level doesn't say what was analyzed.

Christophe: if you seq RNA and cDNA report both nomenclature so is redundant.

NO RESOLUTION REACHED

Variant/Detection/Technique obligatory

Ray: also need to say that the technique was not recorded.

Andy: with an LSDB it is implied in the data that only the relevant gene or even a part of gene is sequenced. When we transfer to large datasets we need to show what part of the genome was sequenced.

Andy: although multiple techniques may be used for a sample we do not give the history, we just describe the technology used to actually characterise the variant as it is reported.

Andy: have a list of techniques which can be used for this (can compare these to Johan's list).

Cross project connectivity - techniques could be added to the Protocol database from Ulf Landegran.

Variant/DNA_remark → remark or comment - recommended

Mike: is there an example of this, what is this.

Heikki: any comment so DNA can be removed from the name. Typically is always about variant, not patients. e.g. family of three analysed and all were identical.

Ray: mutation incorrectly described in x citation.

Variant/Frequency - recommended

Add what population is relevant - local DB content or external source e.g. HapMap

Mummi: rather than a min info in free text, report a relative frequency.


Helen: do you need to say what the population is? Needs to be clear if this is local or global.

Andy: is there a standard way to represent this data in dbSNP? Should align with dbSNP.

Variant/Origin - recommended

Remove sporadic as a separate term.

Purpose needs clarification - is this inheritance or the origin of the mutation. How did it arise vs how inherited. Also many more options in the list in the appendix, maternal and paternal. This is patient variant info, not just variant. Needs to be clearer. List is redundant with allele, these two are not cleanly separated. Is the disease and the variant assumed in this case, needs to be clearer.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Helen Parkinson		Version: v1.0 Final

Ray: Are sporadic/de novo different? *Helen:* is this including the parent? *Ivo:* is in the allele field. *Ray:* is inherited/not inherited, get germ line mosaicism as well. also needs to be covered. Also in cancer causing genes these are somatic.

Mike: clear from survey that people want segregation info, not sure how useful this is if doesn't also say that they have the disease.

NO RESOLUTION, see comments

Variant/Retriction_site - optional

Typically a list of sites, clarify definition.

Variant allele - recommended

Some redundancy with Variant/origin needs to be clarified. Suggestion that this can be optional. Need a way to link two variants with two parents. Labelling confusing. Suggests is a property of variant and is a property of the variable and allele.

Heikki: most common disorders are not imprinted so this may be optional.

Andy: if have two variants need to link two variants to two parents. This text doesn't do that. need a way to link two records together.

NO RESOLUTION

Variant/pathogenicity - recommended

Andy: is it likely that we will all agree on the same pathogenicity values?

Heikki: unlikely.

Ray: this is not the phenotype is about the variant, not linked to a patient.

Mauno: add evidence - e.g. list of GO codes that may work in this space.

Patient/Patient ID - obligatory

Purpose needs clarifying what about cases where there are multiple samples taken, does this need another field. Is this an internal id? external id? Make a recommendation that the language is changed to include anonymised or non identifiable ids NOT a lab id. Use these consistently when reporting.


Ray: where there is an id in a citation I record that id. Where there is no id I don't add that. I add an anon id to people who are submitting data to the database, map to anon no. They keep the mapping.

Christophe: same in UMD is anonymised id. Not a lab ID.

Mauno: maybe that is not available, change to recommended.

Ray: there are cases where there is no patient id.

Heikki: people ignore the samples per patient issue.

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies		Security: PU
	Authors: Tomasz Adamusiak, Helen Parkinson		Version: v1.0 Final

Patient/Phenotype/Disease - recommended

Problem that the initial phenotype and e.g. updated phenotypes - e.g. when a patient matures and gets onset of disease.

Need a way to update this in light of diagnosis.

Helen: presented as singular, is probably multiple.

Andy: reason the patient is referred may be a diagnosis which may change - Phenotype is a list of observable facts.

Helen: description using OMIM is probably the worst thing that can be used. Make a general statement about a recognised vocab.

Christophe: we do not store the original info - people can misdescribe the patient. If I put DMD this is wrong. We collect only validated information.

Ray: Need a way then to describe the primary phenotype and a concluded one.

Patient/remarks - recommended

unclear - change text to anything about the patient, phenotype suggests redundancy with the above statement. Needs work.

No problem with the idea of a comment.

Patient/origin/geographic recommended

We need to know what origin means - does it mean the place they were born, town, country.

Need clarity.

Christophe: if you add too much info you may be able to trace the info if rare enough.

Patient/Origin/Ethnic - recommended

Needs a vocab clearly.

MH: looking at the NCI meta data repository, there is a difference between race and ethnicity.

E.g. Hispanic, and race can get white, Asian. Hispanic has many origins. caDSR.

Andy: clinical geneticists came up with 'additional relevant ancestry'. Not record unless relevant to the disease and can be very specific. Not record unless relevant to the disease and can be very specific. Ashkenazi Jewish for e.g. is relevant, in other cases is not. Change to include this.


Heikki: use it when you have to, use an existing vocab.

Christophe: this is sensitive data, use it when you are allowed to, when you have consent for this. May reflect on recommended.

Patient/Gender - recommended

ISO std for this that includes the genetic disorders. Need to add unspecified.

Andy adds - standard is as follows: 'The most common internationally recognized standard for gender is ISO 5218. It is limited to values corresponding to Male (1), Female (2), Not Known (0) and Not Specified (9). Not Known means no information has been given, not specified means that the gender is non-specific (i.e. cannot be determined as male or female for whatever reason).

 HEALTH-200754	D3.4 Scope and Range Requirements of Specialized Domain Models. Appendix III. Record of discussion and comments		
	WP3 – Standard data models and terminologies	Security: PU	
	Authors: Tomasz Adamusiak, Helen Parkinson	Version: v1.0 Final	8/8

ID_submitterid - obligatory /recommended

Why is this obligatory? Does this add anything if exchanged, surely this can be tracked other ways. Many people curate from the literature.

Issues with credit, and also sharing contact details - maybe change to recommended as people may choose not to have this shared. If only an id is needed and not contact info then this adds little for data exchange.

Mummi: contributor concept issues of contributions. Microattribution get dozens of names per variant. Is this the minimum?

Additional Item; which regions were assayed when detecting a mutation, sequences, exons?

Andy and Christophe want to know what was sequenced/assayed e.g. genome, part gene, a single exon an an extra field in the proposed list.