



HEALTH-F4-2007-200754

www.gen2phen.org

D4.2 Graphical Software for the Presentation of LSDB Data

WP4 – Genetics G2P Databases

**V1.3
Final**

Lead beneficiary: LUMC
Date: 11/02/2010
Nature: Prototype
Dissemination level: PU



 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

TABLE OF CONTENTS

DOCUMENT INFORMATION	3
DOCUMENT HISTORY	3
DEFINITIONS	4
1. INTRODUCTION.....	5
2. DESCRIPTION OF WORK	5
3. FUTURE WORK.....	5
ANNEX	9
REFERENCES.....	15

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

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	4.2	Title	Graphical Software for the Presentation of LSDB Data
Work package	Number	4	Title	Genetics G2P Databases

Delivery date	Contractual	Month 24	Actual	11/02/2010
Status	Version 1.3		final <input checked="" type="checkbox"/>	
Nature	Report <input type="checkbox"/> Prototype <input checked="" type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

Authors (Partner)	P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)			
Responsible Author	Johan T. den Dunnen	Email	J.T.den_Dunnen@lumc.nl	
	Partner	LUMC	Phone	+31-71-5269501

Document History

Name	Date	Version	Description
P. Taschner (LUMC), I. Fokkema (LUMC), J. den Dunnen (LUMC)	13/11/09	1.0	Draft
C. Bérout, G. Collod-Bérout	04/12/09	1.1	Draft
P. Taschner	14/12/09	1.2	Draft
P. Taschner	11/02/10	1.3	Final draft

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases	Security: PU	
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)	Version: v1.3–Final	4/16

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.

	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

1. INTRODUCTION

Locus-specific databases (LSDBs) store information about one or more specific genes. They have been developed by curators with different interests, leading to patient-centered, sequence variation-centered, disease-centered, and protein-centered databases. Clinicians and researchers can search these databases to assess the potential disease-causing effects of sequence variants in specific genes or to see which diseases or phenotypes are associated with specific variants in different genes. Most LSDB data are presented in tabular format, but many data are generally easier to interpret when presented as graphical displays. In addition, graphical displays may help to avoid errors when entering variant data and errors when searching for variants.

To facilitate collection and expert curation of gene sequence variants in these databases, GEN2PHEN partners INSERM and LUMC have developed their ‘LSDB-in-a-box’ software: the Universal Mutation Database (UMD - <http://www.umd.be/>, 1) and the Leiden Open-source Variation Database (LOVD- <http://www.lovd.nl>, 2). UMD and LOVD can store DNA sequence variation information data generated by different clinicians and researchers for one or more specific genes in combination with phenotypic and disease-related information.

Since Work package 4 “Genetics G2P Databases” is focussed on creating locus-specific database solutions, this also includes the development of graphical software to display LSDB data.


2. GRAPHICAL DISPLAY OF LSDB DATA

In this deliverable, we describe the approaches and software components used by UMD and LOVD to create graphical displays of the LSDB contents. Detailed information about the graphical tools can be found in the user manuals for UMD (version December 2009) and LOVD (version 2.0-23), which are available from the UMD and LOVD websites. Two complementary approaches are used: creating graphical displays within the database application and providing data to generate displays using third party tools.

2.1 Graphical displays created within ‘LSDB-in-a-box’ applications

LOVD:

After setup, LOVD has a sequence variation-centered focus with a limited set of fields to store the minimal LSDB contents as described previously (3). Curators are able to extend this set by creating custom columns, which may contain a large variety of data. This variable content makes it difficult to use an automatic approach for graphical display of custom data. Therefore, the standard graphical tools included in LOVD are limited mostly to display data present in all LOVD installations. These are mainly used to display variant statistics for a (sub)set of the database contents (Annex I, Fig. 1). For more integrative views of variant data, LOVD supports data exchange with central repositories and genome browsers (including NCBI, UCSC, Ensembl,

	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final


see below). The development of a data exchange format by WP3 will also allow LOVD data display by third parties (e.g. Phenosystems (4), Alamut (5)).

UMD:

Since its creation, the UMD system has been designed to help users to extract knowledge from crude data and therefore includes many sophisticated algorithms with graphical display options. They address a wide range of topics ranging from molecular to genotype-phenotype analysis or even to therapeutic hypothesis validation. In order to facilitate the interpretation of these data, graphical displays are used in combination with tables. Because these tools are frequently unique, we developed an easy-to-use import interface to rapidly integrate data from any source within the UMD system. This allows all curators to use these algorithms easily no matter which system they had used to set-up their LSDB (LOVD, MutDB (6, <http://mutdb.org/>) or any other system). This procedure is detailed in the UMD manual (version December 2009, pages 29-42).

The graphical displays can be used to:

- Present general characteristics of the gene (UMD manual pages 53-54):
 - o Coding sequence
 - o Exon phasing (cf. annex 2)
 - o Genomic organization
 - o Constitutive splice sites characteristics
- Analyze molecular mechanisms (UMD manual pages 63-64):
 - o Involved in small deletions
 - o Involved in small insertions/duplications
- Analyze the various reported mutations (UMD manual pages 68-107):
 - o Detailed mutational events
 - o Frequency of mutations
 - o Distribution of mutations (small or large rearrangements)
 - o Mutation maps (cf. annex 3)
 - o Geographic distribution (cf. annex 4)
 - o Binary or multiple comparisons
 - o Sequence analysis for splicing auxiliary motifs identification
 - o Hydrophobicity impact of missense mutations
 - o 3D graph
 - o Haplotype and phylogeny
 - o Modules analysis
 - o Branch point maps
 - o Splice site maps
 - o ESE maps
- Analyze phenotypic (mostly clinical) data
 - o Expressivity analysis
 - o Onset of symptoms
 - o Diagnosis delay
 - o Survival rate

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final


- Weight graph
- Graph per symptom
- Graph per age
- 3D graph with two symptoms (cf. annex 5)
- Validation of therapeutic hypothesis
 - Nonsense mutations and read-through
 - Stop codons map
 - Exon skipping

With the development of the data exchange format, UMD is exporting data that can also be displayed by genome browsers or third parties (e.g. Phenosystems (4), Alamut (5)).

2.2 Graphical displays of LSDB data created by third party tools


Many genome browsers (Ensembl, UCSC, NCBI Genome Workbench) support display of LSDB data and other genome related information as custom tracks, which are generated on demand (7, 8). To accomplish this, the LSDB curator has to convert the LSDB data into the standard General Feature Format (GFF, <http://www.sanger.ac.uk/Software/formats/GFF>) or in genome browser-specific formats (e.g. the BED format specified by the UCSC, see <http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html>). Clicking the appropriate link on the LSDB website will submit the selected data to the genome browser. The LSDB variant data are then shown as a separate track in a genome browser window. This functionality has been included in LOVD (See Annex I, Fig. 2). LSDB curators can also set up Distributed Annotation System (DAS) servers to display their data in Ensembl (See Deliverable 3.8 DAS Implementation for GEN2PHEN Data and <http://www.ensembl.org/info/docs/das/index.html> for more information).

Another possibility is to use a well-described data exchange format as developed within WP3 (See Deliverable 3.7 Derivation and Specification of Exchange Format). The exchange format contains all the information necessary to support automatic exchange of data between different databases and allows visualization by third party tools. In combination with a web service, automatic import of LSDB data for display can then be initiated by genome browsers and other visualization tools (Universal Browser, NCBI Sequence Viewer, etc.)(7-10). One important issue is that LSDB curators have to allow data exchange. To limit concerns that the availability of LSDB data in genome browsers will render individual LSDB web sites obsolete, the HGVS recently published recommendations for LSDB data sharing with central repositories (11). In general, this will result in the display of sequence-related information with a link to the original LSDB. To demonstrate the use of minimal LSDB data described by the HGVS, variants in the LOVD FKRP database have been visualized in the UCSC genome browser as part of the PhenCode project (<http://globin.bx.psu.edu/phencode/pui.html>, 12) (See Annex I, Fig. 3).

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

3. FUTURE WORK

New graphical tools are expected to be added as part of the extension of the UMD and LOVD ‘LSDB-in-a-box’ software to fit the general Gen2Phen data model developed by WP3 (D3.2). The rapid development of new technology is expected to lead to a growing demand for new data displays from clinicians and researchers. Many patient and control samples are in increasing numbers analysed using high throughput omics approaches. Due to a lack of standards for emerging technology, omics data differ in format and nature depending on the platforms used. Therefore, the incorporation of these data and their display are difficult to accomplish using “LSDB in a box” software. The most practical solution to generate integrated views of high-dimensional omics data and sequence variant data is to provide LSDB data for use in integrative tools developed by third parties, such as the Integrative Genomics Viewer (13). The data exchange format developed by WP3 (D3.7) is expected to enhance the integration of LSDB data in integrative tools.

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3-Final 9/16

ANNEX I

DNA variants

variant	number	location				percentages
		5'start	coding	intron	3'stop	
substitutions	18645	118	3464	14868	195	
deletions	1889	3	477	1198	211	
duplications	1553	6	156	1111	280	
insertions	457	2	19	431	5	
insertion/deletions	97	0	61	36	0	
inversions	1	0	0	1	0	
2 variants in 1 allele	202	-	-	-	-	
complex	4362	-	-	-	-	
unknown	66	-	-	-	-	
totals	27272	129	4177	17645	691	

RNA variants

variant	number	percentages
substitutions	588	
deletions	176	
duplications	26	
insertions	25	
insertion/deletions	9	
2 variants in 1 allele	16	
complex	112	
unknown	22830	
no effect	157	
no RNA produced	3333	
total	27272	

Variants not observed: inversions, splice variants


Protein variants

variant	number	percentages
substitutions	confirmed: 136 predicted: 1497	
deletions	confirmed: 23 predicted: 30	
duplications	predicted: 1	
insertions	confirmed: 1	
insertion/deletions	confirmed: 1 predicted: 2	
2 variants in 1 allele	confirmed: 30	
frame shifts	confirmed: 211 predicted: 476	
no protein variants	confirmed: 3333	
nonsense	confirmed: 255 predicted: 953	
translation initiation variant	predicted: 1	
silent	confirmed: 369	
complex	19476	
unknown	477	
total	27272	

Variants not observed: nonstop variants

Legend:  confirmed  predicted

Fig. 1. Variants statistics of the LOVD DMD gene database

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

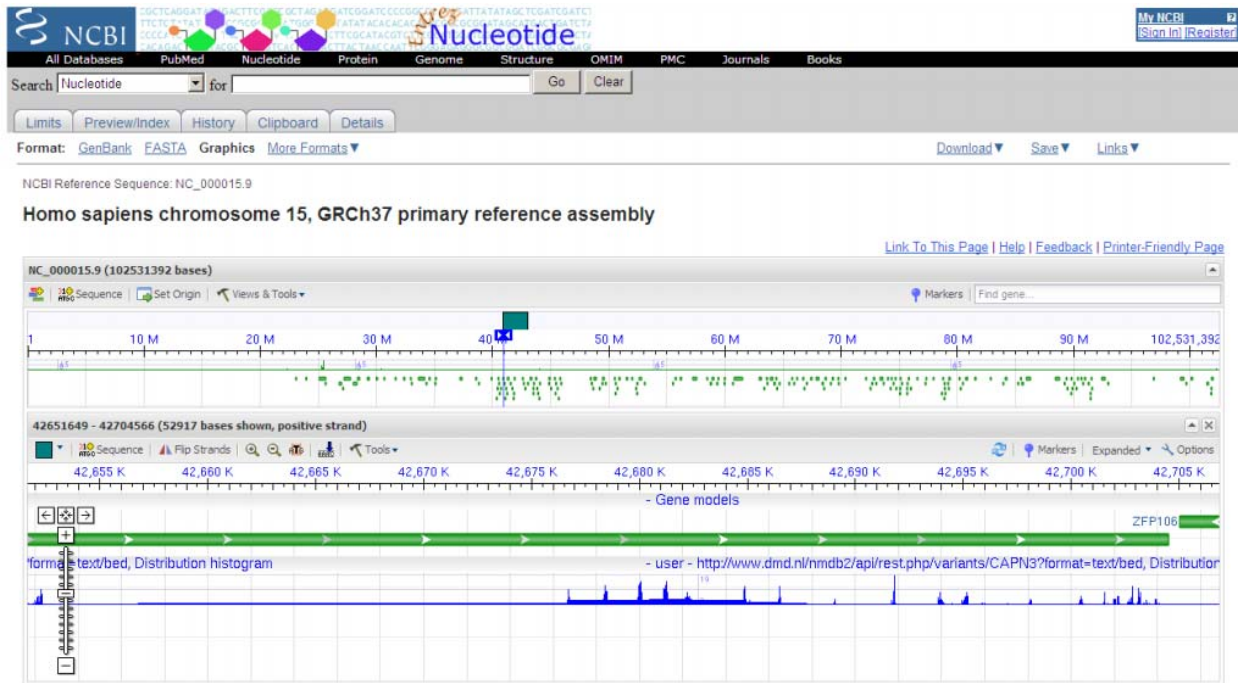



Fig. 2. Distribution of CAPN3 variants displayed using the NCBI Sequence Viewer.

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3-Final 11/16

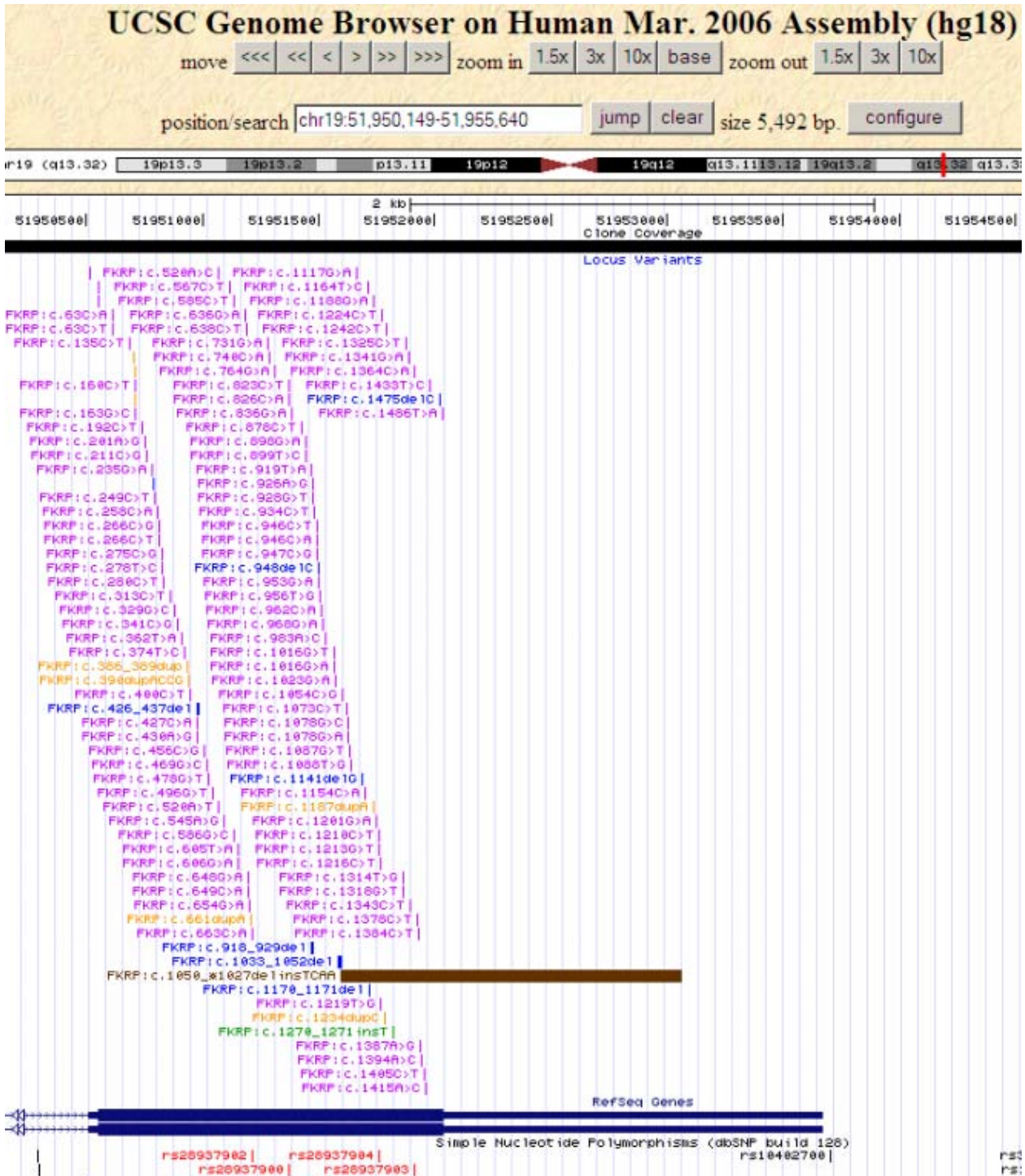

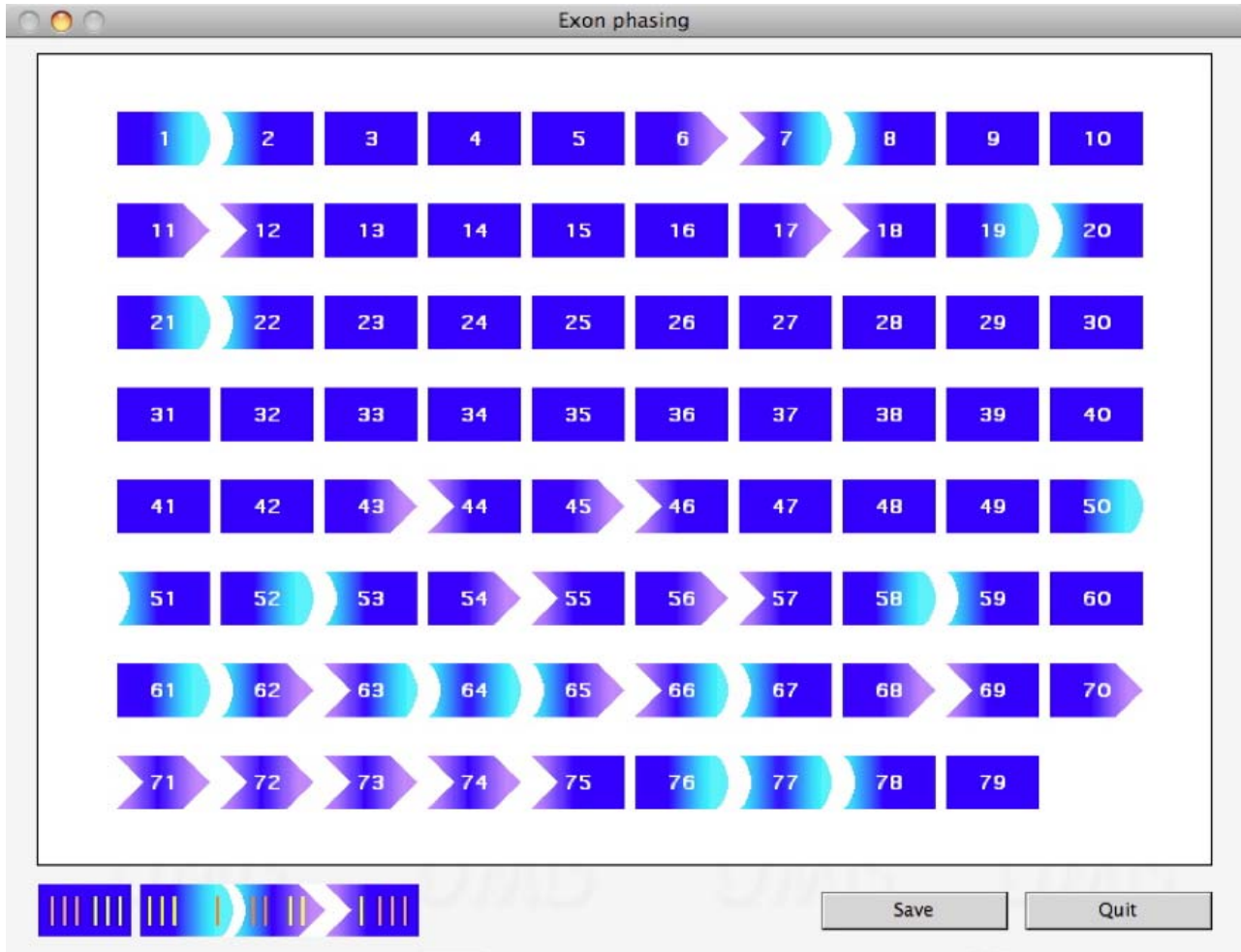



Fig. 3. FKRP variants displayed using the Phencode version of the UCSC Genome Browser.

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3-Final

ANNEX 2

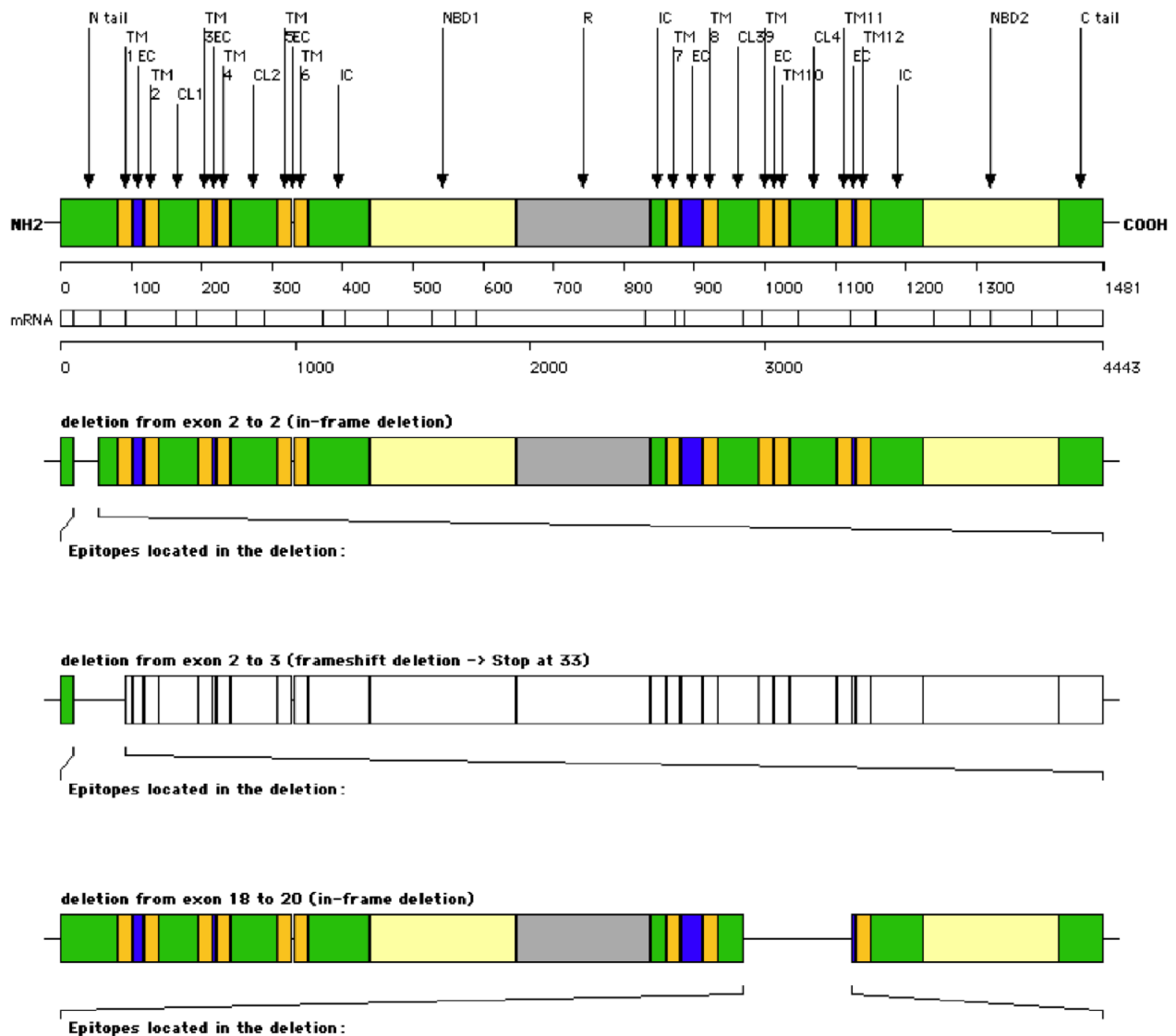
Example of the Exon phasing graphical display of the DMD gene extracted from the UMD-DMD database (<http://www.umd.be/DMD/>):



 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3-Final

ANNEX 3

Example of the theoretical consequences of large deletions on the reading frame of the *DMD* gene. Note that antibody epitopes localized within the deleted region are also displayed. These data were extracted from the UMD-DMD database (<http://www.umd.be/DMD/>):





HEALTH-200754

D4.2 Graphical Software for the Presentation of LSDB Data

WP4 - Genetics G2P databases

Security: PU

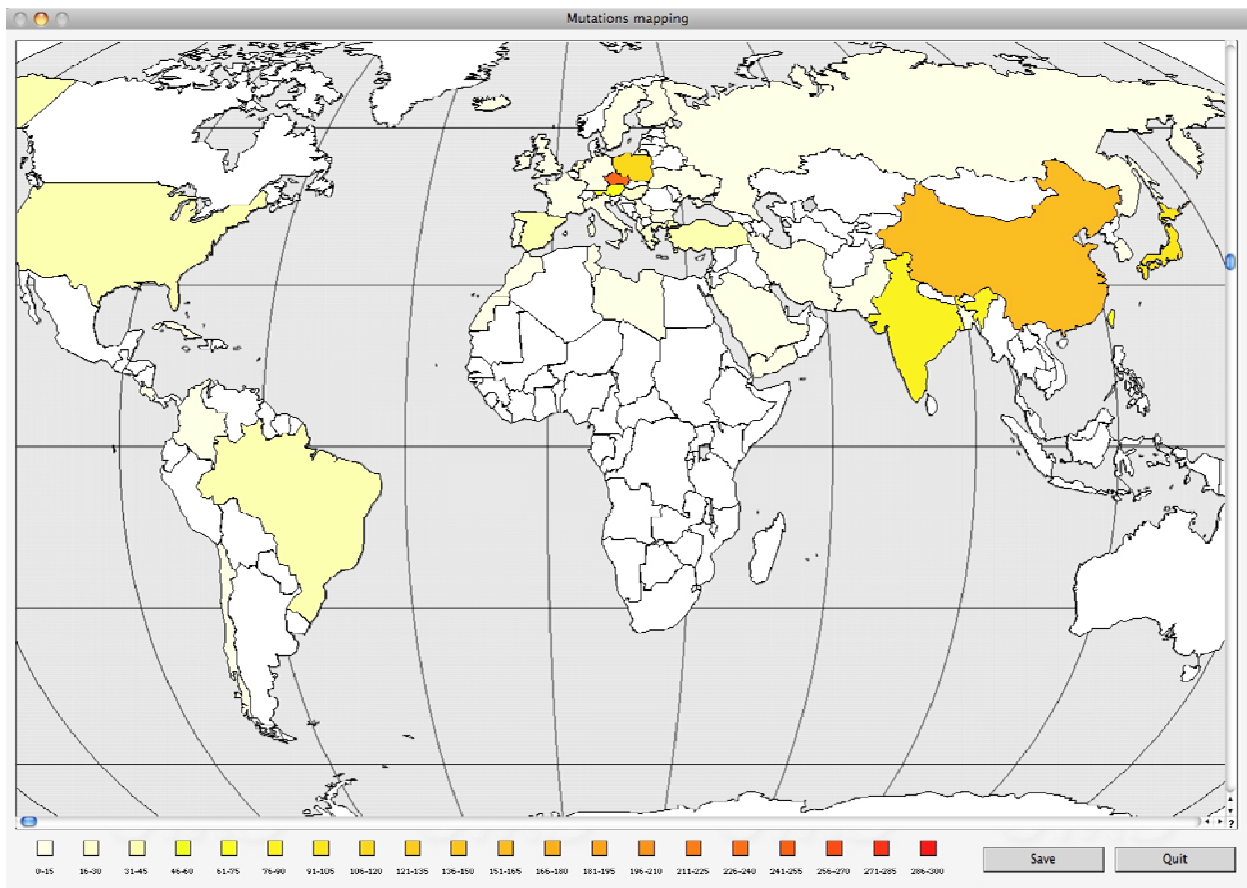
Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)


Version: v1.3–Final

14/16

ANNEX 4


Example of the distribution of missense mutations of the *ATP7B* gene. These data were extracted from the UMD-ATP7B database (<http://www.umd.be/ATP7B/>):



 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

References

1. Beroud C, et al. 2005. UMD (universal mutation database). Hum Mutat 26:184-191.
2. Fokkema IF, den Dunnen JT, & Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-Box" approach. Hum Mutat 26:63-68.
3. 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. 2008. Recommendations for locus-specific databases and their curation. Hum Mutat 29:2-5.
4. Phenosystems (<http://www.phenosystems.com>)
5. Alamut (<http://www.interactive-biosoftware.com/alamut.html>)
6. Mooney SD, Altman RB. 2003. MutDB: annotating human variation with functionally relevant data. Bioinformatics 19:1858-1860.
7. T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Gräf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle and P. Flicek. 2009. Ensembl 2009. Nucleic Acids Res 37 (Database issue):D690-D697. -1006.
8. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996.
9. Universal Browser (<http://www.ngrl.org.uk/Manchester/universalbrowser.html>)
10. NCBI Sequence Viewer (<http://www.ncbi.nlm.nih.gov/projects/sviewer/>)
11. den Dunnen JT, Sijmons RH, Andersen PS, Vihinen M, Beckmann JS, Rossetti S, Talbot CC Jr, Hardison RC, Povey S, Cotton RG. 2009. Sharing data between LSDBs and central repositories. Hum Mutat 30:493-495.

 HEALTH-200754	D4.2 Graphical Software for the Presentation of LSDB Data		
	WP4 - Genetics G2P databases		Security: PU
	Author(s): P. Taschner (LUMC), I. Fokkema (LUMC), C. Beroud (INSERM), J. den Dunnen (LUMC)		Version: v1.3–Final

12. 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, Va`liaho J, Kent J, Miller W, Hardison RC. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28:554–562.

13. Integrative Genomics Viewer (<http://www.broadinstitute.org/igv>)