



HEALTH-F4-2007-200754

www.GEN2PHEN.org

D 6.2 Successful initial integration of at least one LSDB and one Genomics database into Ensembl

WP6 – Integration and Data Access Technologies

**V3.0
Final**

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




 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

TABLE OF CONTENTS

DOCUMENT INFORMATION	4
DOCUMENT HISTORY	4
DEFINITIONS	5
1. INTRODUCTION	6
1.1. OVERALL GOALS OF DATA INTEGRATION	6
1.2. ENSEMBL AS A CENTRAL INTEGRATION RESOURCE	7
1.2.1. <i>Founding the LRG standard</i>	7
1.2.2. <i>Additional support for and current limitations on data integration</i>	8
1.3. EXISTING GENETICS DATABASE RESOURCES WITH A FOCUS ON GENOTYPE-TO-PHENOTYPE DATA 9	
1.3.1. <i>Locus specific databases (LSDBs)</i>	9
1.3.2. <i>Diagnostic Laboratory databases</i>	9
1.4. EXISTING GENOMICS DATABASES RESOURCES WITH A FOCUS ON GENOTYPE-TO-PHENOTYPE DATA	10
1.4.1. <i>Individual-level databases</i>	10
1.4.2. <i>Summary-Level databases</i>	11
2. ENSEMBL EXTENSIONS AND DEVELOPMENT	12
2.1. VARIATION ANNOTATION.....	12
2.1.1. <i>Requirement of common identifier space</i>	12
2.1.2. <i>Extension of Ensembl variation database schema</i>	12
3. INTEGRATION OF A GENOMICS GENOTYPE-TO-PHENOTYPE DATABASE INTO ENSEMBL	15
3.1. EUROPEAN GENOME-PHENOME ARCHIVE	15
3.2. NHGRI MANUALLY CURATED CATALOG	15
3.3. INTEGRATION WITH THE ENSEMBL WEBSITE	16
3.4. SEARCHING FOR PHENOTYPE DATA WITH ENSEMBL	18
4. INTEGRATION OF A GENETICS GENOTYPE-TO-PHENOTYPE DATABASE INTO ENSEMBL	20
4.1. INTRODUCTION	20
4.2. LRG IMPLEMENTATION WITHIN THE ENSEMBL CORE DATABASE	20
4.2.1. <i>LRG storage in Ensembl core database tables</i>	20
4.2.2. <i>Ensembl API developments to support LRG integration</i>	21
4.3. LRG DATAFLOW INTO ENSEMBL.....	21
4.4. VARIANT DATA AND ANNOTATIONS FROM LSDB DATABASES.....	22

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

5. INTERACTIONS WITH NCBI	22
5.1. INTRODUCTION	22
5.2. PLAN OF INTERACTION.....	23
5.3. FUTURE PLANS AND BENEFITS.....	24
ANNEXES	24
REFERENCES.....	29

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – 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 2	Title	Successful Initial Integration
Work package	Number 6	Title	Domain Analysis and Community Relations

Delivery date	Contractual	June 18	Actual	August 2009
Status	Version 3.0		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)	Paul Flicek and Fiona Cunningham (EMBL)		
Responsible Author	Paul Flicek		Email flicek@ebi.ac.uk
	Partner	EMBL	Phone +44 (0)1223 492581

Document History

Name	Date	Version	Description
P. Flicek and F. Cunningham	28/07/2009	1.0	Initial Draft
P. Flicek and F. Cunningham	28/07/2009	1.1	Internal corrections
P. Flicek and F. Cunningham	01/08/2009	2.0	Corrected with comments from R. Hastings
P. Flicek and F. Cunningham	08/08/2009	3.0	Incorporation of comments from A. Devereau

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

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 Technology 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	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

1. INTRODUCTION

1.1. Overall goals of data integration

One of the primary conceptual goals of the GEN2PHEN project is the integration of a number of diverse types of genotype-to-phenotype data resources in such a way as to maximise their utility and discoverability by researchers both within and beyond the community of those currently interested in and conducting research to increase our knowledge about the connections between genotype and phenotype. This is illustrated in the diagram of the three legged stool (Figure 1) that is one of the symbols of the GEN2PHEN project: effective methods for data integration will consist first of the development of the tools and techniques to access data that is contained in the existing resources at a central location (Figure 1: “Getting data in”). A key aspect of this is to create the middleware required to harmonise the way each individual resource thinks about its data internally, so that a resource created to integrate the data in the individual resources will be able to fully capture the details of the data in the external resources (Figure 1: “Data storage and infrastructure”). Although once collected into an integrated framework, the methods for accessing and using these integrated data sets is almost limitless, we will take our first steps in data presentation through existing resources which are already extensively used by the community (Figure 1: “Getting data out”).

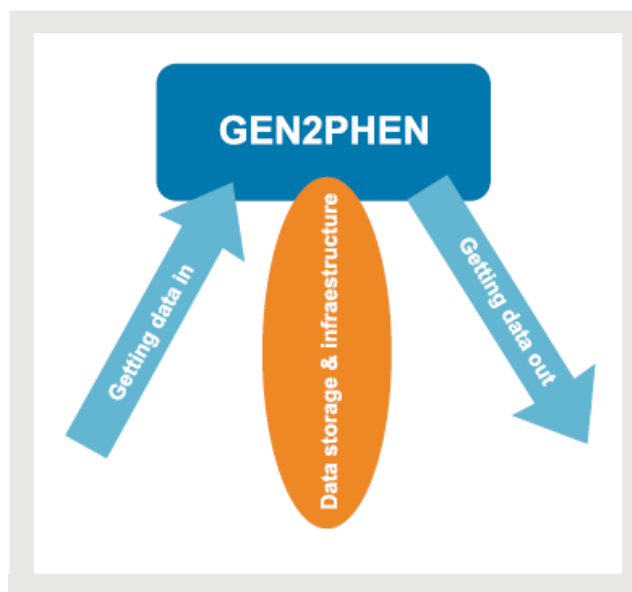



Figure 1: Major efforts of the GEN2PHEN project: (1) getting data in; (2) data storage and infrastructure; (3) getting data out

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

1.2. Ensembl as a central integration resource


Fundamentally, all of the information about the genotype-to-phenotype relationship can eventually be anchored to the genome sequence, although this process may be more complicated in some cases than others. A natural way to view the genome sequence is through a linear presentation on which important features (such as genes) and other features of interest are presented simultaneously with additional genome annotations. Such additional annotations comprise locations of variation including single nucleotide polymorphisms (SNPs), small insertion and deletion variations or larger copy number or structural variations (CNV/SV); regions of specific evolutionary conservation; experimentally known regions of the transcriptional activity or transcription factor binding; as well as regions predicted to be functional through computational analyses that focus on statistically significant patterns in the DNA sequence, patterns in multiple or pairwise alignments or any other information. In this context genotype-to-phenotype information can be considered one of these additional annotations and it makes scientific sense to attempt to integrate this new information into well established resources where most of the remaining annotations are already stably and corrected presented in an integrated manner.

Ensembl is one of the world's major sources of genome information and it serves as a means to create, collect, visualise and make available important genomic data sets to facilitate data analysis, integration and understanding. There are a number of specific challenges to integrating genotype-to-phenotype information into Ensembl and many of these will be addressed throughout the life of the GEN2PHEN project. In this report we will detail our initial efforts to integrate specific datasets that represent both genetics genotype-to-phenotype databases and genomics genotype-to-phenotype databases.

1.2.1. Founding the LRG standard

Successful data integration is most often built on data standards. In deliverable 3.3 we report extensively on the development and implementation of the LRG standard reference sequences, which are critical to the successful integration of data from genetics genotype-to-phenotype databases. For completeness, this development is summarised here.

The Locus Reference Genomic (LRG) standard reference sequences (Deliverable 3.3) are based on extensive discussions with locus specific database curators, diagnostic groups, genomics groups and central archive groups. These discussions crystallised during 2008, with a specific workshop occurring in Hinxton, phone conferences occurring during the year and another in-person meeting associated with American Society of Human Genetics (ASHG) annual meeting in November. Throughout the process, the LRG standard has been built from an international perspective that involved at every step the appropriate colleagues at the NCBI including those responsible for the RefSeq annotations [1]. We have registered the domain www.lrg-sequence.org and are using it as the central portal to the project.


 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

The LRG standard is critical for defining the genomic reporting of variants because it addresses a fundamental tension that arises when the same sequence is used for both the reporting of variant information and the biological interpretation of the variant. A key development of this standard has been separating the need for stable reporting of variants from the need for an up-to-date representation of biology in a region. Historically the predominant variant reporting method was based around cDNA coordinates. However, the use of this coordinate system has limitations due to irregular updates in the cDNA coordinates used for reporting and the desire for the addition of considerably richer sources of biological information (such as gene regulatory information) that are normally associated with the full genome sequence coordinates rather than the cDNA coordinates. By separating these two concepts the LRG provides an unchanging reporting standard that can be predictably coupled to richer information within Ensembl. This coupling, which is at the heart of the integration of LSDB data into Ensembl requires a considerable amount of engineering. Fortunately for the LSDB community, the integration of their data into Ensembl provides immediate and no-cost benefits including the ability to integrate the rich data resources from Ensembl into the LSDB databases. As already mentioned, additional details of this engineering are described in Deliverable 3.3

1.2.2. Additional support for and current limitations on data integration

The LRG standard is a necessary first step to Locus Specific Database (LSDB) data exchange and integration into Ensembl (although the LRG will be useful for more than just these two activities). Additional work to enable LSDB data exchange, done on behalf of the Human Genome Variation Society (HGVS), included recommendations for the data exchange between LSDBs and central repositories [2]. These recommendations for the data types to be shared have been implemented by the Leiden Open Variation Database (LOVD) in form of a "data sharing export tool" available in the most recent releases of LOVD.

There are significant limitations for data integration into public resources such as Ensembl. Most limitations are based on the potential for inadvertent identification of individual data from the participations in the corresponding research studies that led to the creation of the information to be integrated. In some cases, the individuals concerned have provided the necessary consent for all of their personally identifiable information to be publicly released, but this is very much the exception and not the rule. In the past, summary or aggregate level data was commonly released publicly, but a recent publication described a method to address the set membership problem for individuals participating in a genome-wide associated studies [3]. These results led to additional access controls for some of the summary GWAS data that had previously been fully integrated into Ensembl (and was originally planned to be a significant part of this deliverable report). All of this summary level data was removed from Ensembl and, beyond this brief comment, will not be covered in this report.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

Needless to say, these new results have significantly affected our plans for data search and display in this domain. Currently the policy in this area is still being developed, and we have participated in these policy discussions at a number of levels and are developing methods to enable data access, methods for search and methods for integration in light of the expected data access controls that will remain in place.

1.3. Existing genetics database resources with a focus on genotype-to-phenotype data


Database resources containing genotype-to-phenotype information are described in this section and the relationship between them and Ensembl is illustrated in Figure 2.

1.3.1. Locus specific databases (LSDBs)

Locus specific databases are focused on one or several genes. These resources are most often maintained by individual investigators with an interest in the diseases or syndromes which result from or are associated with the variants in the gene that is the focus of the database. In other cases, the curator may have a specific long term interest in the gene itself. There is little centralisation or coordination regarding the decision that a researcher makes to set up and maintain an LSDB, although several hundred LSDBs are known [4]. Those that do exist use a variety of implementations from sophisticated database applications designed to support and maintain LSDBs to simple text files and spreadsheets. Of the dedicated LSDB applications, two of the most popular are the Universal Mutation Database (UMD) [5] and the Leiden Open Variation Database (LOVD) [6]. These resources are important applications, but do not provide the entire informatics infrastructure to integrate their data with central resources such as Ensembl.

1.3.2. Diagnostic Laboratory databases

Data from diagnostic laboratory databases (e.g. that which is generated from the screening of patients with known or suspected disease risk) is rarely made available to outside resources for integration into more public databases such as Ensembl. In part this is because of the requirements to ensure that such data is effectively and completely anonymised before the DNA sequence can be released. Additionally, for the case of diagnostic laboratories any solution must achieve data submission and integration with minimal effort due to the demands of the clinical setting. However, there are a number of technical hurdles to this release as well including software to support the data submission to central resources and the requirements for standard reference sequences that can be used by the diagnostic laboratory and understood by the central resources. This technical hurdle is meant to be addressed by the LRG standard reference sequence (see Deliverable D3.3). However, other hurdles are still to be overcome and this report does not describe the integration of any clinical diagnostic data.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

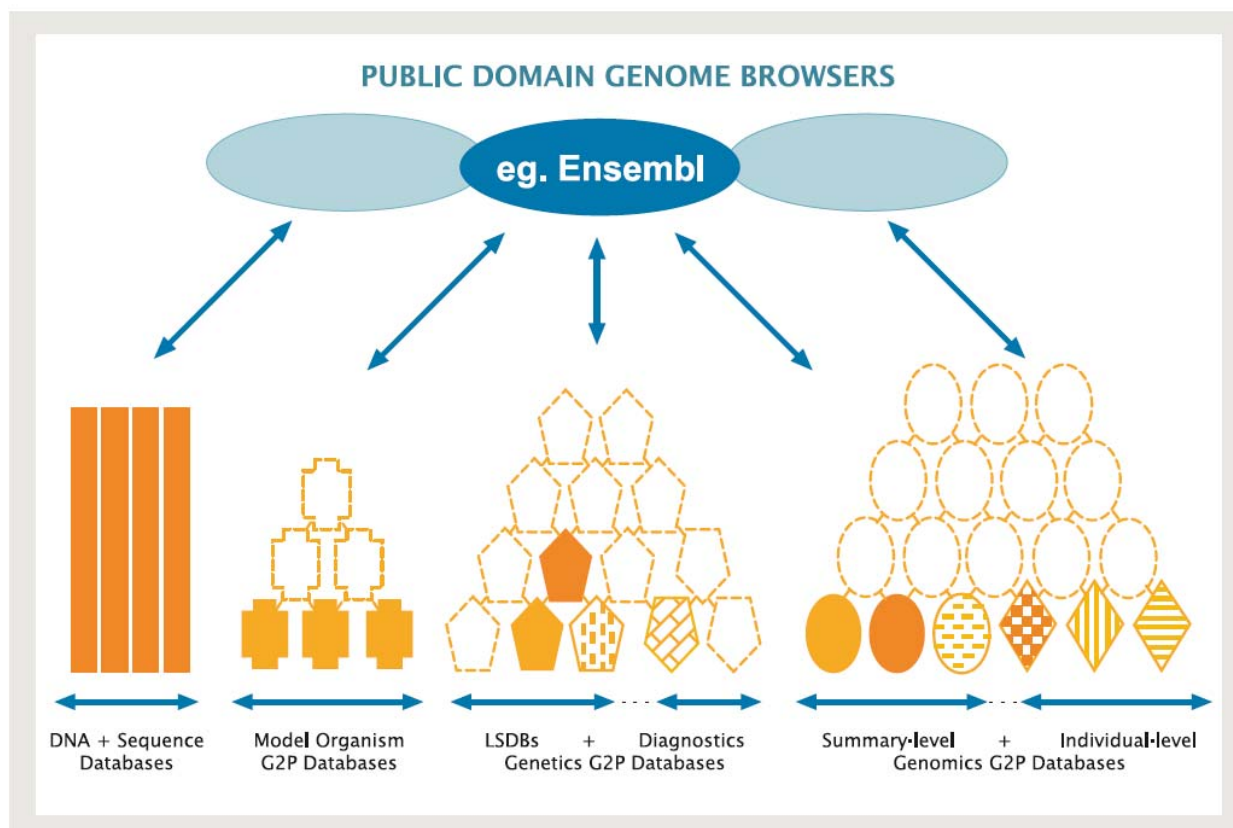



Figure 2: Idealised relationship between Ensembl and both genetics and genomics genotype-to-phenotype data resources

1.4. Existing genomics databases resources with a focus on genotype-to-phenotype data

1.4.1. Individual-level databases

Genomics genotype-to-phenotype databases contain individual-level data, which may include raw or processed data formats representing array-based dense genotype information, whole or partial genome sequences, extensive phenotype data or any other data source that might be reasonably considered “potentially identifiable” to the individual research participant. As noted in section 1.2.2 the information considered potentially identifiable under this standard is still an area of evolving policy and science.

The fact that this data is stored in databases and accessible to researchers that apply for access and meet certain minimum standards for that access is a consequence of the need for researchers to replicate the results found in other large scale studies as well as the considerable benefit of data release. Several major projects are working in this area and addressing these issues. The

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

GEN2PHEN project has involvement in or connections to almost all of these projects, including the European Genome-phenome Archive (EGA) at EBI (<http://www.ebi.ac.uk/ega/>) and the P3G project (<http://www.p3gconsortium.org>). We are also working closely with international partners and peer databases such as the NCBI's Database of Genotypes and Phenotypes (dbGaP) [7].


1.4.2. Summary-Level databases

Summary-level databases build naturally on the individual-level databases and, to a first approximation, represent the results of analysis that has used the individual-level data stored in the databases described in section 1.4.1 as input. There are a number of different kinds of summary-level analysis that are appropriate to consider in this context. Simple aggregate statistics such as allele frequencies for the populations or cohort collected are the most basic of these summary statistics although they are now restricted from public distribution as noted in section 1.2.2. Other summary-level information is the result of more comprehensive analysis and may include the genome-wide significance (in the form of p-values) of each of the assayed variants for a specific disease in a case-control study. In an ideal world, summary-level information would be publicly available and widely displayed in resources such as Ensembl.

A number of summary-level databases currently exist and some of these are linked to the individual-level databases described above. These include public data resources from the EGA linked to the individual data collections in the EGA as well as information from publications that give accession numbers to resources such as dbGAP and EGA. This linkage is natural because the individual-level databases tend to accession their data objects on the level of a study and the ability to release some summary-level information (such as the SNP most associated with a given disease) has already been provided in the form of the published results for a given study. The Wellcome Trust Case Control Consortium (WTCCC) is one example of a study that published the most significantly associated SNPs [8]. In this specific case all of the individual-level and summary-level data is stored in the EGA and it is trivial to provide from the EGA the locations of SNPs reported in the study. There are currently hundreds of genome regions that have been associated with common diseases using genome-wide association studies (see below).

Beyond the summary-level databases directly linked to individual-level databases, there are a number of resources that exist. Some of these already have limited availability in Ensembl through the DAS protocol [9] including the OMIM database (Online Mendelian Inheritance in Man) referenced to <http://www.ncbi.nlm.nih.gov/omim/>.

More recently an effort has been made by a group at the National Human Genome Research Institute (NHGRI) of the US National Institutes of Health to collect and curate the results of published genome-wide association studies [10]. This resource is especially topical considering the explosion of replicated results from genome-wide association studies and is quickly becoming the definitive source for staying abreast of this field.

	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

2. Ensembl extensions and development

2.1. Variation annotation

2.1.1. Requirement of common identifier space


The ability to annotate variation objects is required for the integration of data held in other resources. Simply put, annotation is the structured storage of information which describes characteristics of the object in question. In this case, we are interested in storing structured information about variation objects for which information exists in other resources. The first step in this process is the identification of the common identifier that will be used to ensure that both the external resource and Ensembl are referring to the same data object. For the case of variations there are two obvious identifiers that might be used to ensure that the same object is being referred to. The first is the location of the variation in genome coordinates and the second is the RefSNP identifier (rsID) assigned by the dbSNP build process. There are advantages to each option: genomic location can be readily determined as long as the genome assembly version is known and is useful to describe newly discovered genetic variation before such variations have been submitted to and processed by dbSNP, while the use of an rsID is well understood in the variation community though not every variant has an rsID assigned.

We have chosen to use the rsID as the primary identifier, although we confirm the genomic location as an additional validation check when it is possible to do so. Moreover, we are working closely with dbSNP to ensure that new rsID can be assigned quickly (see section 5 for more details).

2.1.2. Extension of Ensembl variation database schema

The integration into Ensembl of external genotype-to-phenotype data resources, whether genetics genotype-to-phenotype resources as described in section 1.3 or genomics genotype-to-phenotype databases as described in section 1.4, requires modifications to the underlying database schema and application programming interface (API). Specifically the Ensembl variation database had to be extended to support the storage and retrieval of phenotype data through the creation of a phenotype table and variation annotation table show in Figure 3 below. Both database tables required the development of API code compatible with the Ensembl-style API conventions [11]. These conventions specify a series of method types for data storage, retrieval and other operations required for database production, analysis and display on the Ensembl web site.

As well as setting up an import pipeline, we have added an additional two tables into our variation database for "phenotype" and "variation_annotation".

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

A (phenotype table)

Field	Type	Null	Key	Default	Extra
phenotype_id	Int(10) unsigned	NO	PRI		auto_increment
name	varchar(50)	YES	UNI		(null)
description	varchar(255)	YES	UNI		(null)

B (variation_annotation table)

Field	Type	Null	Key	Default	Extra
variation_annotation_id	int(10) unsigned	NO	PRI		auto_increment
variation_id	int(10) unsigned	NO	MUL		(null)
phenotype_id	int(10) unsigned	NO	MUL		(null)
source_id	int(10) unsigned	NO	MUL		(null)
study	varchar(30)	YES	(null)		(null)
study_type	set('GWAS')	YES	(null)		(null)
local_stable_id	varchar(255)	YES	(null)		(null)
associated_gene	varchar(255)	YES	(null)		(null)
associated_variant_risk_allele	varchar(255)	YES	(null)		(null)
variation_names	varchar(255)	YES	(null)		(null)
risk_allele_freq_in_controls	varchar(30)	YES	(null)		(null)
p_value	varchar(20)	YES	(null)		(null)

Figure 3: The full table structure for the phenotype table (A) and variation_annotation table (B) in the Ensembl variation database.

The extensions of the Ensembl variation database schema can be clearly seen in the context of the other data that is stored within the database from the schema extract that is shown in Figure 4. The extract is not a complete representation of the data stored in the Ensembl variation database, but instead represents the most important of the 33 tables that make up the Ensembl variation databases as of Ensembl release 55. Connections to the Ensembl core data base which stores all of the DNA sequence and the gene annotations are also noted where appropriate. Note that not all connections to other Ensembl databases are shown in the diagram.



HEALTH-200754

D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl

WP6: Integration and Data Access Technologies

Security: PU

Author(s): Fiona Cunningham and Paul Flicek (EMBL)

Version: v3.0 – Final

14/30

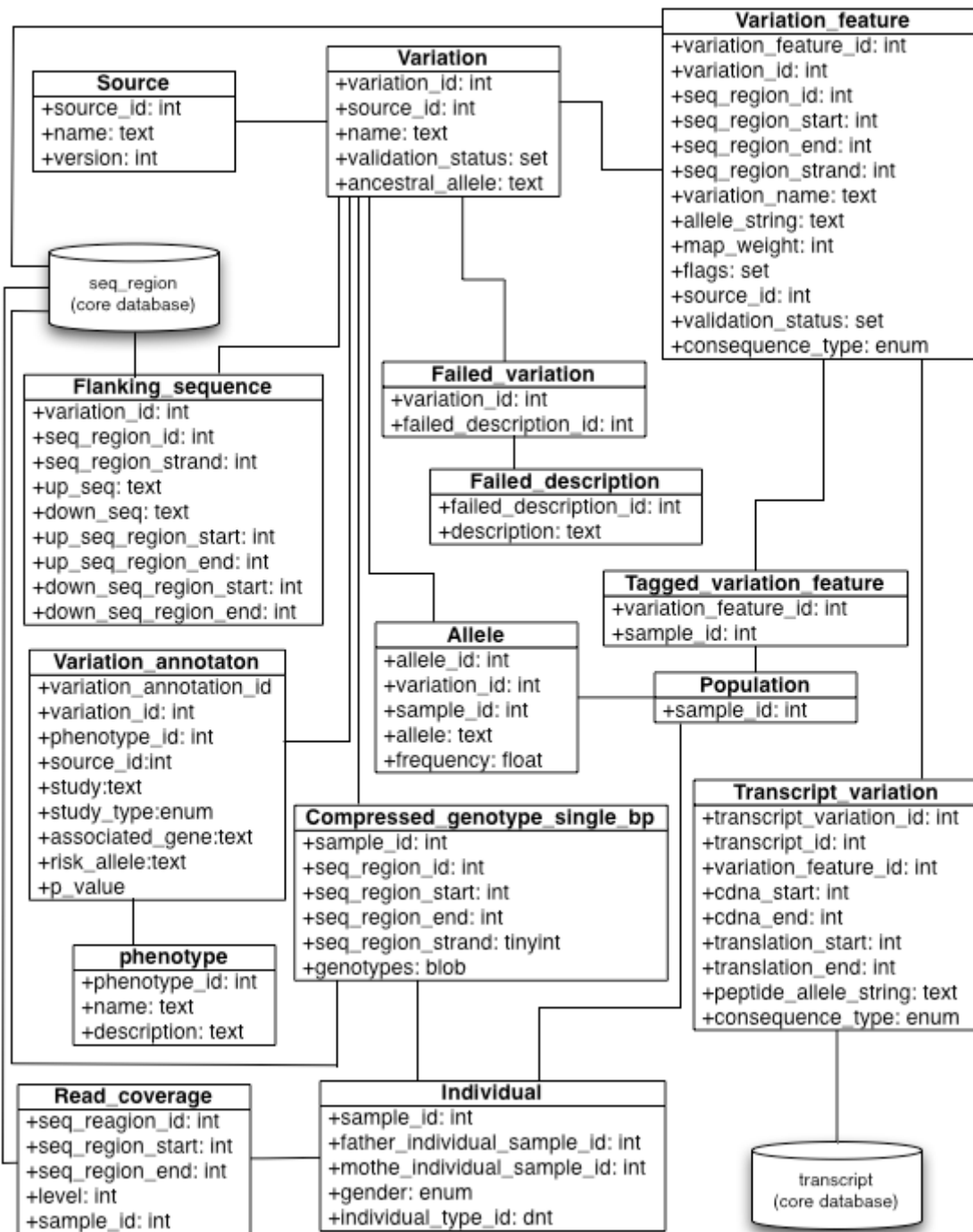



Figure 4: Extract of the Ensembl variation database schema showing new development for the storage and integration of genotype-to-phenotype data.

	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

3. Integration of a genomics genotype-to-phenotype database into Ensembl

We have taken advantage of close ties with the EGA (<http://www.ebi.ac.uk/ega/>) to set up a pipeline to integrate their publicly available data into Ensembl. Similarly, we have developed a collaboration with the National Human Genome Reference Institute (NHGRI) to incorporate their "Catalog of Published Genome-Wide Association Studies" (<http://www.genome.gov/GWAstudies>). As of Ensembl release 55 (July 2009) we have incorporated data from these two resources totalling 134 phenotypes from 1120 phenotype annotations.

We include a direct link from the web page to the original submitter's data entry under the "Source" column thereby acknowledging their contribution directly. In addition we provide a link to the publication via PubMed.


3.1. European Genome-phenome Archive

The European Genome-phenome Archive (EGA) database at the European Bioinformatics Institute (EBI) is designed to provide a permanent archive for all types of personally identifiable genetic data including genotypes, genome sequence and associated phenotype data. The EGA contains both data collected from individuals whose consent agreements limit data release to specific research uses or bona fide researchers and specific data approved for unlimited public release. For access to genomic and phenotypic data from samples important to biomedical research the EGA seeks to lead in implementation and play an important role in policy creation. There are currently data representing approximately 50,000 individuals and 10 studies within the EGA.

3.2. NHGRI manually curated catalog

The Office of Population Genomics at the National Human Genome Research Institute (NHGRI) of the US National Institutes of Health has developed a curated database of published significant genomic regions identified from genome-wide association studies [10]. This database content is shown graphically on the karyotype image downloaded from <http://www.genome.gov/GWAstudies> and shown in Figure 5. More regions are being added each month as additional GWAS studies are published and the data is being updated for every release of Ensembl.

The NHGRI catalog contains data about the associated SNP (listed by rsID), the disease, and the publication of reference for this association. All of this data has been integrated into Ensembl (see section 3.3, Figure 6 and Figure 7).

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

Published Genome-Wide Associations through March 2009
398 published GWA at $p \leq 5 \times 10^{-8}$

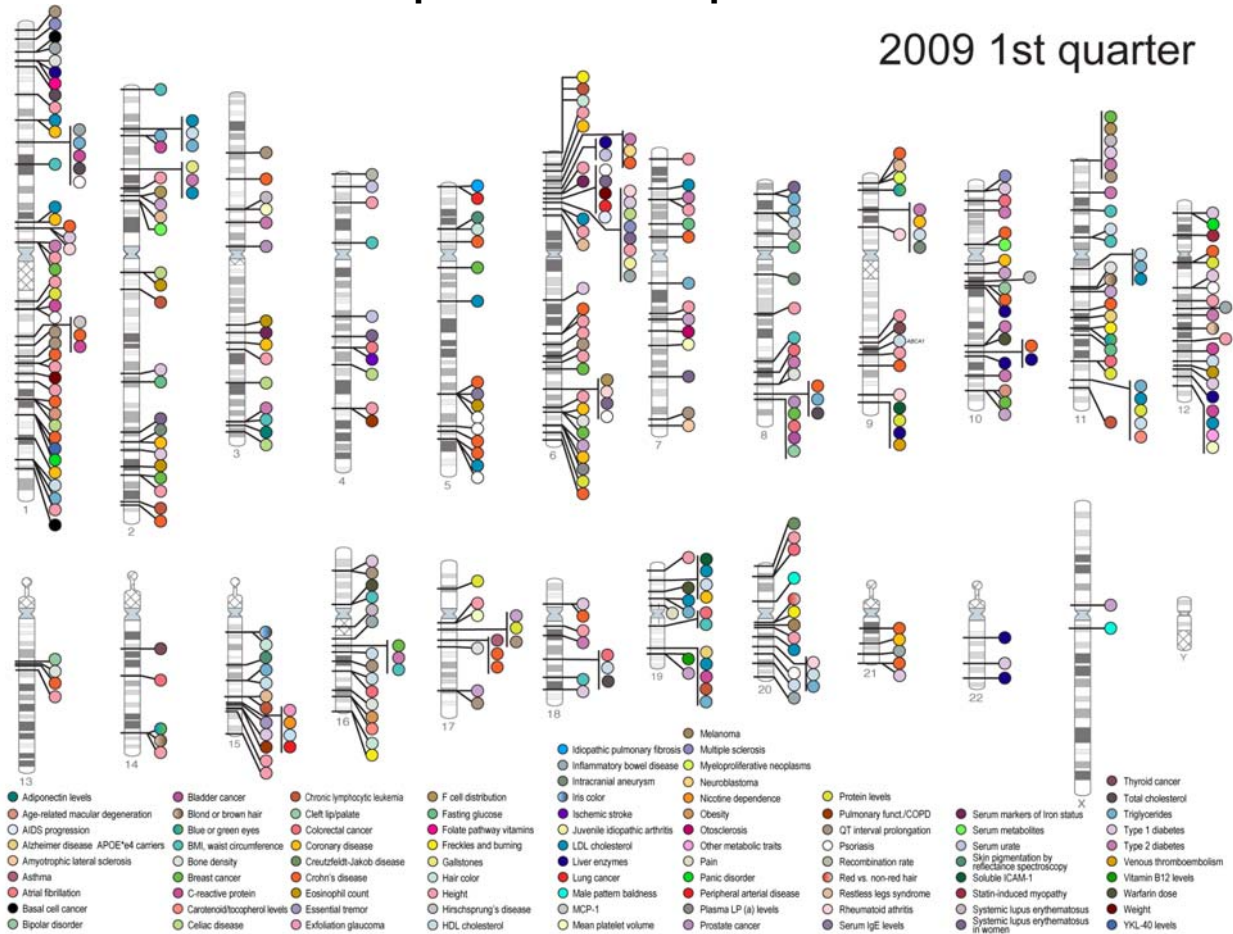



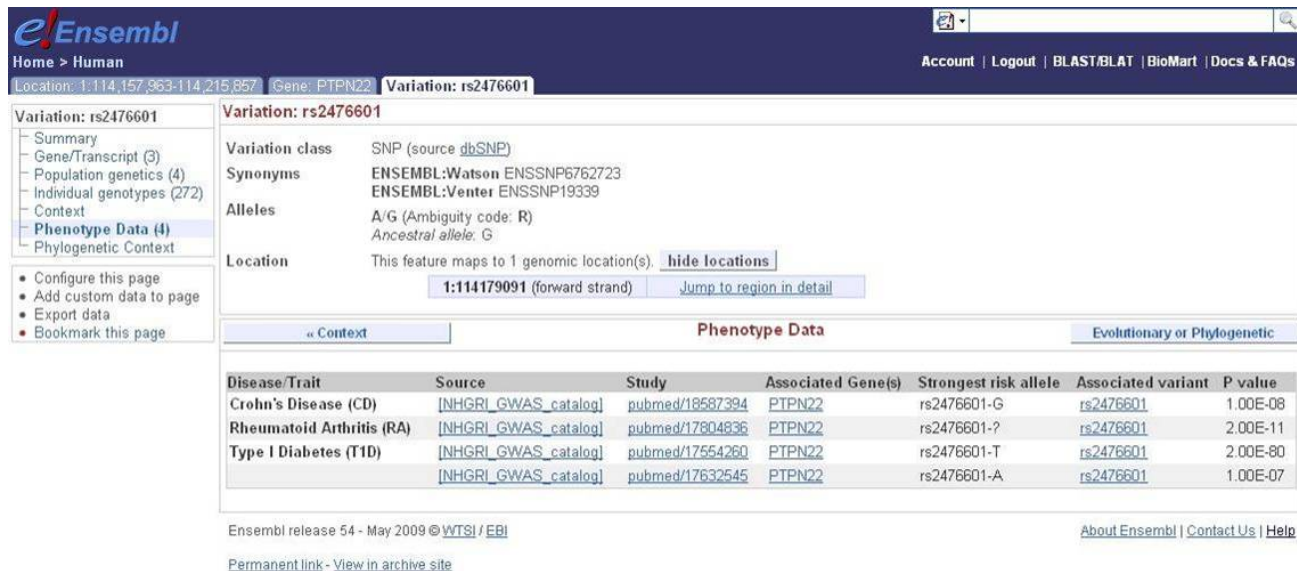
Figure 5: A karyotype view of the regions identified by genome-wide association studies and contained in the NHGRI curated database. A partial list of the diseases covered is listed by colour at the bottom of the figure.

3.3. Integration with the Ensembl website

To display the data we have added a new panel tab to the "Variation" data pages on the Ensembl website. This tab collects all the phenotype related data by SNP and groups the data across sources to enable easy comparison. The data stored under the phenotype tab may come from any of the genetics or genomics genotype-to-phenotype data resources described above. For each case the data source of the phenotype information will be prominently displayed (see Figure 6).

For detailed instructions about using variation data in Ensembl including information about the SNP displayed in Figure 6, please see Appendix I.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final



Ensembl
Home > Human
Location: 1:114,157,963-114,215,857 | Gene: PTPN22 | Variation: rs2476601

Variation: rs2476601


Variation class: SNP (source dbSNP)
Synonyms: ENSEMBL:Watson ENSNP6762723, ENSEMBL:Venter ENSNP19339
Alleles: A/G (Ambiguity code: R), Ancestral allele: G
Location: This feature maps to 1 genomic location(s). [hide locations](#)
1:114179091 (forward strand) [Jump to region in detail](#)

« Context **Phenotype Data** Evolutionary or Phylogenetic

Disease/Trait	Source	Study	Associated Gene(s)	Strongest risk allele	Associated variant	P value
Crohn's Disease (CD)	[NHGRI_GWAS_catalog]	pubmed/18587394	PTPN22	rs2476601-G	rs2476601	1.00E-08
Rheumatoid Arthritis (RA)	[NHGRI_GWAS_catalog]	pubmed/17804836	PTPN22	rs2476601-?	rs2476601	2.00E-11
Type I Diabetes (T1D)	[NHGRI_GWAS_catalog]	pubmed/17554260	PTPN22	rs2476601-T	rs2476601	2.00E-80
	[NHGRI_GWAS_catalog]	pubmed/17632545	PTPN22	rs2476601-A	rs2476601	1.00E-07

Ensembl release 54 - May 2009 © WTSI / EBI [About Ensembl](#) | [Contact Us](#) | [Help](#)
[Permanent link](#) - [View in archive site](#)

Figure 6: An example of rs2476601 from Ensembl release 54 showing that this SNP in the PTPN22 gene is associated with several diseases including Crohn's Disease, Rheumatoid Arthritis, and Type I Diabetes. It is interesting to note that the risk allele for each of these diseases is different in this position in the genome. The reference publications for each of these associations are linked to the appropriate record in the PubMed database. Additional links allow the user to immediately display the genomic region around the associated SNP in Ensembl which is shown in Figure 7 and immediately demonstrates the value of data integration within the Ensembl system.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

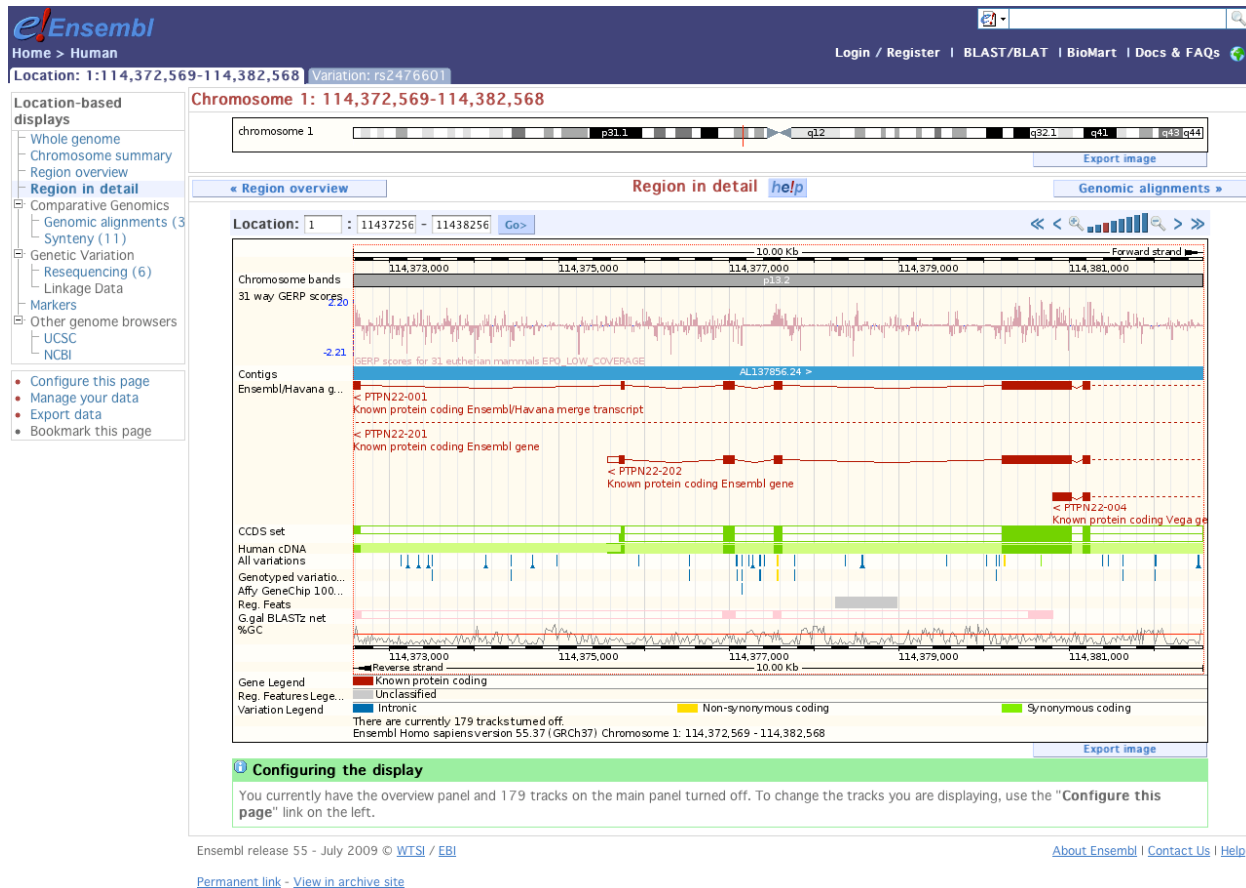

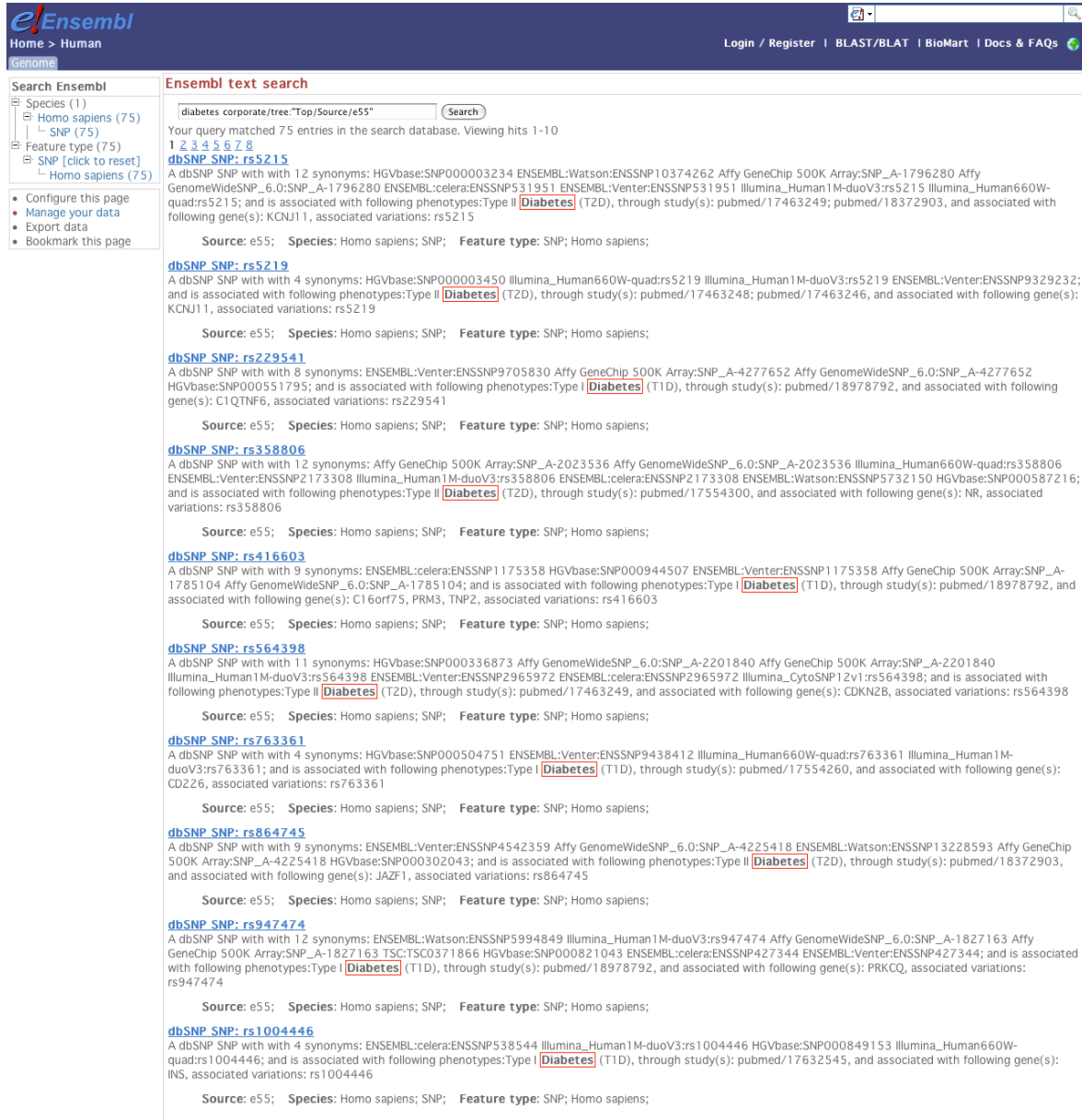


Figure 7: The genomic region immediately surrounding the disease associated SNP shown in Figure 5. Several benefits of data integration within Ensembl are immediately apparent from this figure. For example, the associated SNP, which is in a coding exon on the PTPN22 gene is at the centre of the figure in the variations track and is coloured yellow to indicate that it is a non-synonymous SNP within a protein coding gene. Additionally, we see the annotation of a putative regulatory region (in grey) very close to the SNP. We can also simultaneously view the evolutionary conservation as displayed by the GERP score calculated from the 31-way multi-species alignment (at the top of the image). Finally, we can see the deep evolutionary conservation of this region by displaying the human-chicken pairwise alignment (in pink at the bottom of the image), which shows that this region of the genome is likely to be under strong evolutionary constraint throughout vertebrate evolution.

3.4. Searching for phenotype data with Ensembl

It is also possible to search for the phenotype terms using the general Ensembl search as shown in Figure 8. The results of these searches include the associated SNPs for a given disease including the example shown in Figure 6. This is an entirely new feature introduced in Ensembl release 55.


 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final



The screenshot shows the Ensembl genome browser interface. At the top, there is a search bar with the query "diabetes corporate/tree:Top/Source/e55" and a search button. Below the search bar, it indicates that 75 entries were found. The results are listed as follows:

- dbSNP SNP: rs5215**
A dbSNP SNP with 12 synonyms: HGvbase:SNP000003234 ENSEMBL:Watson:ENSSNP10374262 Affy GeneChip 500K Array:SNP_A-1796280 Affy GenomeWideSNP_6.0:SNP_A-1796280 ENSEMBL:celera:ENSSNP531951 ENSEMBL:Venter:ENSSNP531951 Illumina_Human1M-duoV3:rs5215 Illumina_Human660W-quadr:rs5215; and is associated with following phenotypes: Type II **Diabetes** (T2D), through study(s): pubmed/17463249; pubmed/18372903, and associated with following gene(s): KCNJ11, associated variations: rs5215
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs5219**
A dbSNP SNP with 4 synonyms: HGvbase:SNP000003450 Illumina_Human660W-quadr:rs5219 Illumina_Human1M-duoV3:rs5219 ENSEMBL:Venter:ENSSNP9329232; and is associated with following phenotypes: Type II **Diabetes** (T2D), through study(s): pubmed/17463248; pubmed/17463246, and associated with following gene(s): KCNJ11, associated variations: rs5219
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs229541**
A dbSNP SNP with 8 synonyms: ENSEMBL:Venter:ENSSNP9705830 Affy GeneChip 500K Array:SNP_A-4277652 Affy GenomeWideSNP_6.0:SNP_A-4277652 HGvbase:SNP000551795; and is associated with following phenotypes: Type I **Diabetes** (T1D), through study(s): pubmed/18978792, and associated with following gene(s): C1QTNF6, associated variations: rs229541
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs358806**
A dbSNP SNP with 12 synonyms: Affy GeneChip 500K Array:SNP_A-2023536 Affy GenomeWideSNP_6.0:SNP_A-2023536 Illumina_Human660W-quadr:rs358806 ENSEMBL:Venter:ENSSNP2173308 Illumina_Human1M-duoV3:rs358806 ENSEMBL:celera:ENSSNP2173308 ENSEMBL:Watson:ENSSNP2173308 ENSEMBL:Venter:ENSSNP5732150 HGvbase:SNP000587216; and is associated with following phenotypes: Type II **Diabetes** (T2D), through study(s): pubmed/17554300, and associated with following gene(s): NR, associated variations: rs358806
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs416603**
A dbSNP SNP with 9 synonyms: ENSEMBL:celera:ENSSNP1175358 HGvbase:SNP000944507 ENSEMBL:Venter:ENSSNP1175358 Affy GeneChip 500K Array:SNP_A-1785104 Affy GenomeWideSNP_6.0:SNP_A-1785104; and is associated with following phenotypes: Type I **Diabetes** (T1D), through study(s): pubmed/18978792, and associated with following gene(s): C16orf75, PRM3, TNP2, associated variations: rs416603
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs564398**
A dbSNP SNP with 11 synonyms: HGvbase:SNP000336873 Affy GenomeWideSNP_6.0:SNP_A-2201840 Affy GeneChip 500K Array:SNP_A-2201840 Illumina_Human1M-duoV3:rs564398 ENSEMBL:Venter:ENSSNP2965972 ENSEMBL:celera:ENSSNP2965972 Illumina_CytoSNP12v1:rs564398; and is associated with following phenotypes: Type II **Diabetes** (T2D), through study(s): pubmed/17463249, and associated with following gene(s): CDKN2B, associated variations: rs564398
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs763361**
A dbSNP SNP with 4 synonyms: HGvbase:SNP000504751 ENSEMBL:Venter:ENSSNP9438412 Illumina_Human660W-quadr:rs763361 Illumina_Human1M-duoV3:rs763361; and is associated with following phenotypes: Type I **Diabetes** (T1D), through study(s): pubmed/17554260, and associated with following gene(s): CD226, associated variations: rs763361
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs864745**
A dbSNP SNP with 9 synonyms: ENSEMBL:Venter:ENSSNP4542359 Affy GenomeWideSNP_6.0:SNP_A-4225418 ENSEMBL:Watson:ENSSNP13228593 Affy GeneChip 500K Array:SNP_A-4225418 HGvbase:SNP000302043; and is associated with following phenotypes: Type II **Diabetes** (T2D), through study(s): pubmed/18372903, and associated with following gene(s): JAZF1, associated variations: rs864745
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs947474**
A dbSNP SNP with 12 synonyms: ENSEMBL:Watson:ENSSNP5994849 Illumina_Human1M-duoV3:rs947474 Affy GenomeWideSNP_6.0:SNP_A-1827163 Affy GeneChip 500K Array:SNP_A-1827163 TSC0371866 HGvbase:SNP000821043 ENSEMBL:celera:ENSSNP427344 ENSEMBL:Venter:ENSSNP427344; and is associated with following phenotypes: Type I **Diabetes** (T1D), through study(s): pubmed/18978792, and associated with following gene(s): PRKCO, associated variations: rs947474
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;
- dbSNP SNP: rs1004446**
A dbSNP SNP with 4 synonyms: ENSEMBL:celera:ENSSNP538544 Illumina_Human1M-duoV3:rs1004446 HGvbase:SNP000849153 Illumina_Human660W-quadr:rs1004446; and is associated with following phenotypes: Type I **Diabetes** (T1D), through study(s): pubmed/17632545, and associated with following gene(s): INS, associated variations: rs1004446
Source: e5S; Species: Homo sapiens; SNP; Feature type: SNP; Homo sapiens;

Figure 8: Results for searching for diabetes in Ensembl

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

4. Integration of a genetics genotype-to-phenotype database into Ensembl

4.1. Introduction

We will use the LRG standard reference sequences described in Deliverable 3.3 as the primary mechanism for incorporation of LSDB data into Ensembl. The LRG standard is uniquely appropriate for this purpose and, because we are also intimately involved with the implementation of the LRG project, both the LRG project and Ensembl will benefit from the sharing of knowledge and computer code between these efforts. We described above some of the development made to the Ensembl databases to support the integration of genotype-to-phenotype data. In this section we will provide details of the implementation of LRG standard in the Ensembl databases using the example of LRG_3 which corresponds to the COL3A1 LSDB that is currently available at https://eds.gene.le.ac.uk/home.php?select_db=COL3A1.

The LRG is a general format that will allow the integration of any LSDB that meets the LRG standard directly into Ensembl. Our initial integration, which is described in detail below uses the COL3A1 LSDB as an example, but could have chosen any LSDB that can be described with the LRG standard.

4.2. LRG implementation within the Ensembl core database

The LRG sequences are integrated into Ensembl through inclusion in the Ensembl core database, which is designed to store sequence and assembly information. The LRG itself is stored as a set of mappings between the reference chromosome and the LRG. They have a coordinate system called "LRG". Deletions are handled by the mappings in the assembly table. Insertions are handled using the seq_edit feature with the Ensembl core database (see below for an example).


4.2.1. LRG storage in Ensembl core database tables

An example here is LRG_3, as it is stored in the Ensembl core database:

```
select * from seq_region where seq_region_id in(1966011,226033);
```

seq_region_id	name	coord_system_id	length
226033	17	17	78774742
1966011	LRG3	102	22876

Segments of the LRG_3 sequence that match the reference assembly stored in the assembly table.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

```
select * from assembly where cmp_seq_region_id = 1966011;
+-----+-----+-----+-----+-----+-----+-----+
asm_seq_region_id | cmp_seq_region_id | asm_start | asm_end | cmp_start | cmp_end | ori |
+-----+-----+-----+-----+-----+-----+-----+
                | 1966011           | 45623706 | 45638067 |          940 | 15301 | -1 |
                | 1966011           | 45623449 | 45623705 |        15304 | 15560 | -1 |
                | 1966011           | 45616138 | 45623448 |        15566 | 22876 | -1 |
                | 1966011           | 45638068 | 45638239 |          763 |    934 | -1 |
                | 1966011           | 45638240 | 45638999 |           1 |    760 | -1 |
+-----+-----+-----+-----+-----+-----+-----+
```

From the above data we can see that LRG_3 is largely identical to Chromosome 17.

The additional sequence that is not on Chromosome 17 is stored as seq_edits in the seq_region_attr table i.e.


```
select * from seq_region_attr where attrib_type_id = 145 and seq_region_id = 1966011;
+-----+-----+-----+-----+-----+
| seq_region_id | attrib_type_id | value |
+-----+-----+-----+-----+-----+
| 1966011       | 145            | 15302 15303 AA |
| 1966011       | 145            | 15561 15565 AAAAA |
| 1966011       | 145            | 761 762 AA |
| 1966011       | 145            | 935 939 ATGAT |
+-----+-----+-----+-----+-----+
```

4.2.2. Ensembl API developments to support LRG integration

The Ensembl API contains methods for creation, retrieval and manipulation of the data objects stored with the Ensembl databases. A key feature of the API is that it provides a consistent interface to the data that allows the Ensembl web site and other resources to interact with the Ensembl data objects even if Ensembl changes the underlying database structure. A new module LRGslice.pm has been created which will act exactly the same within the Ensembl software system as any other DNA sequence (Ensembl refers to an arbitrary stretch of DNA sequence such as a chromosome or a haplotype as a “slice”), but will do all the internal mappings of features from the chromosome onto the LRG. The average user does not need to know about the LRGslice as this will be created automatically and hence no changes are expected to the user interface. In this way there will be seamless access of the data without a requirement to understand the internal storage model.

4.3. LRG dataflow into Ensembl

We recognise that the LRG creation process will be independent from the Ensembl release cycle and we have therefore developed a pipeline process to ensure that new LRG records are incorporated into Ensembl as soon as possible after their creation.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

In short, the Ensembl core team will download the latest batch of LRGs from the LRG FTP site. These will be stored in the core database as "seq regions" with a coordinate system called "LRG". We have created a special wrapper call to be able to retrieve features on an LRG slice.

4.4. Variant data and annotations from LSDB databases

The LRG standard provides a stable sequence coordinate system for reporting variants, but the LRG itself does not contain the variants or the annotation of these variants. These variants are integrated into Ensembl using the same mechanism as used for the genomics genotype-to-phenotype databases described in section 3. This involves the storage of the variant data within the Ensembl variation database and the interaction of the Ensembl variation database and the Ensembl core database. The major difference from the point of view of Ensembl is that the genomics genotype-to-phenotype data resources described in section 3 use the reference genome sequence assembly for their coordinate system; for the integration of genetics genotype-to-phenotype data resources such as LSDBs and clinical databases we use the LRG coordinate system as described in this section.


5. Interactions with NCBI

5.1. Introduction

As described in Deliverable 3.3, LRG reference sequences provide a coordinate system which can be used by diagnostic and locus specific databases in a stable manner going forward. This stability provides a key aspect in the transfer of information about specific variants worldwide.

However, this standard does not have any provision for the storage of variants in a worldwide, coordinated manner, with correct linkage to source database which is one of the purposes of this deliverable. As described above, we propose initially to provide a system for the transmission and storage of presence of variants stored in LSDBs and diagnostic databases. We would then, in a second phase to be developed in the future and built using the information reported here, like to create a graphical representation providing a concise interpretation of variant alleles with respect to phenotype based on the proposed HGVS reporting standard. In both cases, clear links to originating databases would be present for more detailed information extraction.


A key requirement for the success of the above is an effective working relationship with our colleagues from NCBI and dbSNP. This is required as LRGs will also be created at NCBI and data from LSDBs (and other sources) may be provided directly to NCBI.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

5.2. Plan of interaction

In consultation with colleagues at NCBI and dbSNP, we propose the following six point plan of interaction regarding the processing of genetics genotype-to-phenotype data.

1. Variants provided as positions on LRG reference coordinates will be sent, in a simple tab delimited format from the LSDB/Diagnostic database to dbSNP. Submitted data include a “local” identifier for the variant and a URL link. If the LSDB has already associated some variants to rsIDs then they can reported. If desired, it would be appropriate for an LSDB to create an entire dump of a locus, tracking the load from a previous release by virtue of the rsIDs. We believe that this system is already available at dbSNP through the Variation Batch Submission: <http://www.ncbi.nlm.nih.gov/projects/SNP/tranSNP/VarBatchSub.cgi>
2. dbSNP will provide a fast-turnaround (they have promised 2 weeks) assessment of this file, providing a similar tab delimited format which will then include a column of rsIDs that they have assigned. When a variant has the same position and same type as an existing rsID, this will be assigned the existing rsID. When it does not, dbSNP will create a new rsIDs. For LSDB submitted rsID cases, dbSNP will accept this if they agree, but flag it if they disagree on this assignment.
3. NCBI can optionally reject a submission if they believe there is systematic set of errors present in the file, and would report this promptly to the submitting LSDB.
4. On receipt of the rsIDs, there will be a two week period in which the LSDB can check the accuracy of the mapping. If there is no response, it will be assumed that this mapping is correct.
5. After this the LSDB submission will be accepted and dbSNP will regularly create aggregations of all SNPs present through this procedure outside of the normal dbSNP release cycle, though this may not be present on the main NCBI displays until the next dbSNP build. (Scheduling the integration of this data into NCBI main resources is an NCBI process.)
6. Ensembl will take these aggregations and show all these positions within two weeks of retrieval as a DAS source present as a one-click track on both genome and gene orientated displays. Ensembl will also integrate this information in the next available release cycle, with a worst case delay of approximately four months.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

5.3. Future plans and benefits


In phase 2, the format will be extended to take a limited set putative associations, based on HGVS reporting standards, (ie, the seven state flags proposed), and a simple flat text file of phenotype information. Ensembl and dbSNP will describe this as a “phenotype summary” and highlight the originating source for more details.

These proposals put the onus on providing clustering and tracking to dbSNP, with a relatively quick turnaround outside their normal cycle. The benefits of this are considerable especially considering the use of rsIDs as the coordinating ID system for SNPs. We think it is better for there to be one system worldwide than to create complex mappings between two competing systems. However, we know that this clustering in a fast turn around may be a more complex feature to implement and maintain than it appears, and we intend to evaluate its implementation over time and make changes as necessary.

ANNEXES

Appendix I: Information about making the most of Ensembl variation information.

More details can be found at <http://www.ensembl.org/info/website/tutorials/index.html>.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

Ensembl Walk-Through
9 May 2009

VARIATIONS WORKED EXAMPLE

In this worked example we will explore information with regard to the PTPN22 (Tyrosine-protein phosphatase non-receptor type 22) gene with a focus on variation. Note that this worked example only covers a small amount of all the information available in Ensembl!

Note: This worked example is based on Ensembl version 54 (March 2009). After in future a new version has gone live, version 54 will still be available through <http://May2009.archive.ensembl>.

☞ Go to <http://www.ensembl.org>.

(1) Searching for the human PTPN22 gene:

- ☞ Enter 'human PTPN22 gene' in the text box under 'Search Ensembl'.
- ☞ Click [Go].
- ☞ Click on 'Ensembl protein_coding Gene: ENSG00000134242 (HGNC (curated): PTPN22)' on the page with search results.

(2) Genomic sequence of the PTPN22 gene:

- ☞ Click on 'Sequence' in the side menu.


This page, as well as many others, can be customised using the 'Configure this page' link in the side menu.

To add variations to the display:

- ☞ Click on 'Configure this page' in the side menu.
- ☞ Select 'Show variations: Yes and show links'.
- ☞ Click [SAVE and close].

(3) Spliced sequence of the PTPN22-001 transcript:

- ☞ Click on 'ENST00000359785' in the list of transcripts.
- ☞ Click on 'Sequence - cDNA' in the side menu.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

(4) Unspliced sequence of the PTPN22-001 transcript:

☞ Click on 'Sequence - Exons' in the side menu.

To show the full intronic sequence and 1000 basepairs of flanking sequence:

☞ Click on 'Configure this page' in the side menu.

☞ Enter '1000' in the text box behind 'Flanking sequence at either end of transcript:'.

☞ Check the box behind 'Show full intronic sequence'.

☞ Click [SAVE and close].

(5) Cross references to other databases for the PTPN22-001 transcript:

☞ Click on 'External References - General identifiers' in the side menu.

(6) Genomic region around the PTPN22 gene:

☞ Click on the 'Location: 1:114,157,963-114,215,857' tab.

☞ Zoom out one step in the third panel using the zoom tool in the upper right hand corner of the panel.

To add variations and DECIPHER and DGV tracks to the display:

☞ Click on 'Configure this page' in the side menu.

☞ Enter 'variation' in the text box behind 'Search display' in the pop-up screen.

☞ Select 'All variations - Normal'.

☞ Enter 'decipher' in the text box behind 'Search display'.

☞ Select 'DECIPHER - Normal'.


☞ Enter 'dgv' in the text box behind 'Search display'.

☞ Select 'DGV loci - Normal'.

☞ Click [SAVE and close].

DECIPHER and DGV are DAS (Distributed Annotation System) sources; these data are not stored in the Ensembl database, but elsewhere. Ensembl is in this case only used as a means to display these data.

Clicking on a feature on this page will give a pop-up menu with information about the feature. Often the pop-up menu will also contain one or more links to pages with more detailed information.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

To get more information and go to the original DECIPHER entry for a DECIPHER feature:

- ☞ Click on the DECIPHER feature.
- ☞ Click on the link in the pop-up menu that links to the DECIPHER website.
- ☞ Go back to Ensembl by using the back button of the internet browser.

(7) Variations in the PTPN22 gene:

- ☞ Click on the 'Gene: PTPN22' tab.
- ☞ Click on 'Genetic Variation - Variation Table' in the side menu.

By default all variations that are located in the exons are shown as well as those that are located in the introns and flanking sequence and are within 100 bp from an exon.

To show all variations in exons, introns and within 5000 bp up- and downstream of the 5' and 3' end of the transcripts:

- ☞ Click on 'Configure this page' in the side menu.
- ☞ Select 'Context: Full Introns'.
- ☞ Click [SAVE and close].

To show only non-synonymous variations:

- ☞ Click on 'Configure this page' in the side menu.
- ☞ Deselect all options under 'Select Variation Type' except 'Non-synonymous'.
- ☞ Click [SAVE and close].

- ☞ Click on 'Genetic Variation - Variation image' in the side menu


(8) Variations in the PTPN22-001 transcript in different individuals:

- ☞ Click on the 'Transcript: PTPN22-001' tab.
- ☞ Click on 'Genetic Variation - Population comparison' in the side menu.

By default only data for Watson and Venter are shown.

To add other individuals:

- ☞ Click on 'Configure this page' in the side menu.
- ☞ Select all options under 'Select Individuals'.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

☞ Click [SAVE and close].

☞ Click on 'Genetic Variation - Comparison image' in the side menu.

(9) Variations in the PTPN22-001 protein:

☞ Click on 'Protein Information - Variations' in the side menu.

(10) Detailed information on variation rs2476601:


☞ Click on 'SNP ID rs2476601' in the list of variations.

☞ Click on 'Gene/Transcript' in the side menu.

☞ Click on 'Population genetics' in the side menu.

☞ Click on 'Individual genotypes' in the side menu.

☞ Click on 'Phenotype data' in the side menu.

 HEALTH-200754	D 6.2 Successful initial integration of at least one LSDB and one Genomics Database into Ensembl		
	WP6: Integration and Data Access Technologies		Security: PU
	Author(s): Fiona Cunningham and Paul Flicek (EMBL)		Version: v3.0 – Final

References

1. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**:D61-D65.
2. den Dunnen JT, Sijmons RH, Andersen PS, Vihinen M, Beckmann JS, Rossetti S, Talbot CC, Hardison RC, Povey S, Cotton RG: **Sharing data between LSDBs and central repositories.** *Hum Mutat* 2009, **30**:493-495.
3. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.** *PLoS Genet* 2008, **4**:e1000167.
4. Horaitis O, Talbot CC, Phommarinh M, Phillips KM, Cotton RG: **A database of locus-specific databases.** *Nat Genet* 2007, **39**:425.
5. Béroud C, Hamroun D, Collod-Béroud G, Boileau C, Soussi T, Claustres M: **UMD (Universal Mutation Database): 2005 update.** *Hum Mutat* 2005, **26**:184-191.
6. Fokkema IF, den Dunnen JT, Taschner PE: **LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach.** *Hum Mutat* 2005, **26**:63-68.
7. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat Genet* 2007, **39**:1181-1186.
8. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**:661-678.
9. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2**:7.
10. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106**:9362-9367.
11. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E: **The Ensembl core software libraries.** *Genome Res* 2004, **14**:929-933.