



HEALTH-F4-2007-200754

<http://www.gen2phen.org>

D2.3 Technical State-of-the-art Document for G2P Databases

WP2 – Domain analysis and community relations

**V1.3
Final**

Lead beneficiary: EMC
Date: 14/08/2009
Nature: Report
Dissemination level: PU
(Public)


 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

TABLE OF CONTENTS


DOCUMENT INFORMATION	3
DOCUMENT HISTORY	3
DEFINITIONS	4
1. INTRODUCTION	5
2. APPROACH OVERVIEW	5
2.1. LSDBs/DIAGNOSTIC DATABASES	5
2.1.1. <i>Approach</i>	6
2.1.2. <i>Results</i>	6
2.1.3. <i>Conclusions</i>	8
2.2. WHOLE-GENOME GENOTYPE-PHENOTYPE DATABASES	9
2.2.1. <i>Whole genome genotype phenotype databases</i>	9
2.2.2. <i>Discussion and conclusions</i>	12
2.3. SPECIFICATION FOR LOCUS REFERENCE GENOMIC SEQUENCES	13
2.3.1. <i>Rationale</i>	13
2.3.2. <i>Implementation</i>	14
2.3.3. <i>Outline of LRG production process</i>	14
2.4. BIORESOURCE IMPACT FACTOR (BRIF)	15
2.4.1. <i>Introduction</i>	15
2.4.2. <i>Objectives</i>	16
2.4.3. <i>Work performed in the course of GEN2PHEN</i>	16
2.4.4. <i>Conclusions</i>	20
3. REFERENCES	21

APPENDIX I - LSDBs/Diagnostic databases

APPENDIX II - Central G2P Databases, State of the Art Catalog

APPENDIX III- The concept of Bio-Resource Impact Factor

APPENDIX IV- LSDB domain analysis

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		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	D2.3	Title	Technical State-of-the-art Document for G2P databases
Work package	Number	2	Title	WP2 – Domain analysis and community relations

Delivery date	Contractual	June 2009	Actual	August 2009
Status	Version 1.3		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)	Christina Mitropoulou (EMC) Anne Cambon-Thomsen (INSERM), Frank Schacherer (BIOBASE), Raymond Dalgleish (ULEIC), and George P. Patrinos (EMC)			
Responsible Author	George P. Patrinos		Email	g.patrinos@erasmusmc.nl
	Partner	EMC	Phone	+30-6958.008355

Document History

Name	Date	Version	Description
C. Mitropoulou	25.07.2009	1	Draft
G. P. Patrinos	31.07.2008	1.1	Edit
A. J. Brookes	11.08.2009	1.2	Edit
G. P. Patrinos	14.08.09	1.3	Final

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – 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	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

1. INTRODUCTION

One of main objectives of Work Package 2 (WP2) is to analyze, on a technical level, each sub-type of the G2P databases of interest to the GEN2PHEN project, namely LSDBs, Diagnostic databases, and Genomics databases containing either individual or summary level datasets, to understand their key technical features. An emphasis in this work was placed upon parameters of standardization – particularly data models, data exchange formats, and ontologies/nomenclatures. Actual data models in current use have been fully documented. A comparative analysis has considered these features, and also taken note of how they are used in conjunction with specific data curation criteria. These documented data model summaries is expected to provide important supporting materials for the data model development work of WP3.

A sub-focus on ontologies assessed and documented the state-of-the-art for relevant projects (see Deliverable 2.2). That guidance is expected to be used later on in specific implementation activities in WP4 and WP5. Also, ethics was specifically addressed related to LSDBs and diagnostic databases.

The analysis work covered by this activity will spanned human and model organism databases, as these must all, ultimately, be properly integrated. To that end, the analysis also considered empirically-determined pros and cons of various current integration strategies. This broad technology assessment exercise was executed primarily by teams within the Consortium, though external advice was sought as needed.

2. APPROACH OVERVIEW


Our technical domain analysis involved the following areas in the G2P field:

- (a) LSDBs and diagnostic databases (activity 2.1),
- (b) Whole-genome genomics databases (activity 2.2),
- (c) Specifications for locus-reference genomic sequences (activity 2.3), and
- (d) Ethics for LSDBs and diagnostic databases (activity 2.4).

This document summarizes the main findings and conclusions from our technical domain analysis efforts in the fields listed above.

2.1. LSDBs/Diagnostic databases

Mutation databases of human genes are assuming an increasing importance in all areas of health care. In addition, more and more experts in the mutations and diseases of particular genes are curating published and unpublished mutations in locus-specific databases (LSDB). LSDBs provide an invaluable tool for analyzing gene expression and phenotype in both normal and disease conditions, as the curators are closely in touch with molecular biologists very

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

experienced with the analysis of a specific gene and its anomalies. This system generally promotes submission of data and maintains an accurate and up-to-date data source. We performed a thorough domain analysis of 727 existing LSDBs and analyzed the structure and content of each of the LSDBs currently available through the Internet.


2.1.1. Approach

We examined websites containing LSDBs of mutations of individual genes available through the HGVS website (<http://www.hgvs.org>). This site has an extensive list of 727 LSDBs with their URLs that are routinely updated as LSDB curators submit new databases to the list. There may be a few databases available and not included on the list that we have not been able to locate yet, but it is impossible to know for sure. The HGVS list may be viewed on the HGVS website. LSDBs were examined for the presence or absence of 58 content criteria describing the general presentation, locus-specific information, database structure and data collection, fields in the mutation table, structure, content, or overall ease-of use of the database (Appendices 1 and 2). These criteria were evaluated using binary scoring (0: Absence/No; 1: Presence/Yes) and have been selected on the basis of objectivity. Criteria which are based on a more or less subjective judgment by the evaluator (e.g. ease of use, LSDB design, *etc*) were excluded from our study.

2.1.2. Results

Of 727 LSDBs available to our knowledge, 59 were redundant. Also, 33 did not qualify as LSDBs, 21 LSDBs were not available to the general public and 9 were no longer available. This left 604 LSDBs for different nuclear genes available for study. By way of comparison, the Human Gene Mutation Database (HGMD) (Stenson *et al.* 2009) lists 2,426 genes in its public version (and 3,461 in its Professional 2009.2 Release) that contain at least one mutation. The redundancy found among LSDBs indicated that mutations were reported by two databases in the case of 49 genes, 9 databases in the case of three genes, and 4 different databases reported the gene TP53.

Almost 60% of the LSDBs had a home page that provided a clear explanation of content and aim of database and a minimum set of cross-references (active links and pointers) for the user to access additional information. Important links included HGMD (22.2%), OMIM (Online Mendelian Inheritance in Man) for clinically related information (59.5%), other LSDBs (0.6%) and PubMed literature database for access to published references, GenBank/EMBL for detailed DNA sequence information, HGVS nomenclature website, and other useful links (84.1%). 66.1% of LSDBs advised users that information in the database is copyrighted intellectual property, such that they should cite the database in the appropriate manner when using data, while 50.9% had a disclaimer notice. In 38.5% of the cases studied, the database description was published in a peer-reviewed scientific journal. Finally, only 0.6% of the LSDBs were written in a language other than English, namely French in all cases. LSDBs were composed of mutation entries, such


 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases	
	WP2 – Domain analysis and community relations	Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos	Version: v1.3 – Final 7/21

that each entry usually corresponds to a mutation in a single patient and is added (generally, but not always) after curator inspection. Almost all databases were a compilation of data derived from both published literature (99.6%) and submissions directly to the LSDB contributed by researchers throughout the world (96.7%). Direct online submission by contacting curators directly was available in 56.6% of the cases.

Contrary to previous observations, very few LSDBs were structured as flat files containing a number of fields for each entry (1.6%). The majority of LSDBs were relational [SQL (Structured Query Language)-based; 68.6%]. However, a substantial number (29.2%) are still displayed through Hypertext Markup Language (HTML). This number, though is significantly lower, compared to the previous Claustres and coworkers (2002) study¹. A complete table listing all mutations was available in 78.4% of LSDBs, from which only a small portion included downloadable formats (19%) that were sometimes difficult or impossible to download. Some databases showed mutation maps or mutation visualization tools (22.5%) depicting the location of mutations throughout the gene (or even the protein) sequence, and a few added graphical displays, including dynamic graphing tools. Because there is no standard yet, there are still a substantial amount of LSDBs are custom-based, resulting in data content heterogeneity. It is rather encouraging though that 40.4% of LSDBs are based on a Database Management System namely LOVD (141 LSDBs), MUTbase (116 LSDBs) and UMD (10 LSDBs) that contributes to data uniformity. A large number of LSDBs are updated frequently. In particular, 513 out of 727 LSDBs were updated recently (198 LSDBs in 2009, 188 LSDBs in 2008 and 127 LSDBs in 2007), accounting for 70.3% of the total LSDBs. However, almost 30% of the LSDBs were outdated, being updated in 2006 or earlier.

A number of databases contained much information in addition to the list of mutations, making the registries valuable for physicians and scientists from many fields. About 77.1% of the LSDBs included the reference sequence, a vital piece of information that was, unfortunately, not available in the rest. Also, 68.1% included information on the gene, and 54.6% of the chromosome location. However, only 35.9% provided information on the disease and 4% had information on protein function. Proper mutation nomenclature was followed in 56.6% of LSDBs. Finally, only 34.4% of LSDBs displayed links to patient associations or to websites with clinical content.

Information on the gene of interest, protein function, protein structure, and protein sequence alignment could be found in some LSDBs. A few of the LSDBs qualify as “knowledgebases” because they combine scientific and diagnostic data on mutations with associated information useful for clinicians or students (e.g., population distribution of alleles, haplotype associations) and information for patients and their families (e.g., treatment, diagnosis, dedicated organizations, or parent associations). Some LSDBs aimed to facilitate the detection and characterization of mutations by providing technical support in the form of primer sequences and mutation detection protocols. Seventy nine (79)% of LSDBs included a complete reference list and in 90.4% of LSDBs references were directly linked to PubMed literature database. Summary phenotypic descriptions (i.e., pathogenicity: YES/NO, etc) were available in 53.7% of LSDBs whereas detailed phenotypic descriptions were available only in 12.6% of LSDBs analyzed. Only 2.1% of LSDBs were cross-linked with other LSDBs. In addition to the mutation listing, many databases provided data fields for associated information, such as mutation

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final


detection methodology (25.6%), mutation frequency (2.7%), ethnic group/geographic origin of patients (45.9%), restriction enzyme change with mutation (37.6%).

A search engine in 58.2% of LSDBs, mostly relational databases or those utilizing a DBMS, provided the opportunity to interrogate the database for specific information contained in a number of fields, such as mutation name (41%), gene region (33.8%), Codon number (32.3%), author name (44.3%), phenotype (44.7%), ethnic group (28.1%), geographic location (27.4%). In 47.8% of LSDBs, other fields for querying were also available, such as cancer type and classification, DNA source, protein domain, etc. Finally, a number of URLs that we accessed contained databases that did not qualify as LSDBs. In particular, 33 databases did not qualify as LSDBs, e.g. were MS Excel-based with no querying capacity. Also, 21 were password protected and registration for membership was requested to ensure that individuals using the database agree to a set of guidelines covering data submission, confidentiality, appropriate data use, and acknowledgment. These LSDBs were not publicly available and, hence, were not included in the analysis. Finally, 9 LSDBs were no longer available.

2.1.3. Conclusions

The main observations from our technical domain analysis of the LSDB field are:

- (a) A vital characteristic that would enhance the rate of new database creation would be the availability of “off-the-shelf” tailor-made software. Software needs to be able to allow collection, correction, and review and be able to store the data, both published and unpublished. This software needs two main functionalities: to allow operation by official curators, the software either remote or in a central facility, and to allow permanent storage of data. Approximately 40% of the LSDBs analyzed are based on a database management system, namely LOVD, UMD or MUTbase. Therefore, contrary to previous observations, there is less data content heterogeneity than previously reported¹.
- (b) Also, a large number of “LSDBs” are still displayed in HTML and lack basic querying tools, although the number is less than previously measured.
- (c) Our data also showed that data visualization tools are still underdeveloped in the existing and there is much that needs to be done regarding means for visualizing data entries in LSDBs and diagnostic databases.
- (d) Finally, mutation frequency data are under-represented in LSDBs (4%), compared to ethnic group data (45%). This underlines the need for developing separate modules recording ethnic-specific mutation frequency data in LSDBs or separate databases that serve that need, such as National/ethnic mutation databases that currently exist².

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

2.2 Whole-genome genotype-phenotype databases

This part of our technical domain analysis aims to highlight differences in approach, comprehensiveness, and focus of the major ones. Understanding the existing resources can facilitate the creation of a federated overall resource, as well as point out gaps in the current data.


There is a large number of resources that cover specific areas, such as individual diseases or genes. Taken together, these also represent a general cross-section of available knowledge, but since the knowledge in them is fragmented and their number is large, covering each of them in detail is beyond the scope of this text. It must be noted however, that some systems have been used repeatedly to create locus specific databases, and the installed base of these tools might be seen as one distributed data source. Excluded from review are tools that aggregate or display data from various databases, but do not collect and provide unique original data, like for example DiseaseCards (<http://www.diseasecard.org/>). It is acknowledged that there is value in this kind of aggregation being built on top of the original resources, but this is not the focus of this report.

To identify the resources in this list, input from domain experts was combined with the results of an online search. All information on these resources was taken from the publicly available material on their respective web sites. A detailed overview of the genomic databases analyzed is provided in the Appendix 3.

2.2.1. Whole genome genotype phenotype databases

Online Mendelian Inheritance in Man (OMIM; www.ncbi.nlm.nih.gov/sites/entrez?db=omim): OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources. OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. OMIM focuses primarily on inherited, or heritable, genetic diseases.

Human Gene Mutation database (HGMD; www.hgmd.org): HGMD records all germ-line disease-causing mutations and disease-associated/functional polymorphisms reported in the literature, and provides these data in a readily accessible format to all interested parties, whether they are from an academic, clinical or commercial background. HGMD now constitutes, de facto, the central disease-associated mutation database available to the scientific community. The data comprise single base-pair substitutions in coding (e.g. missense and nonsense), regulatory and splicing-relevant regions of human nuclear genes, micro-deletions and micro-insertions, indels, repeat expansions, as well as gross gene lesions (deletions, insertions and duplications) and complex gene rearrangements. HGMD is freely available to registered academic/non-profit users. Mutation data are currently made available on this public website 2½ years after initial inclusion in the database. An up-to-date subscription version is available to both commercial and

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases	
	WP2 – Domain analysis and community relations	Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos	Version: v1.3 – Final


academic customers via license. Various types of mutation within the coding regions, splicing and regulatory regions of human nuclear genes causing inherited disease. Somatic mutations and mutations in the mitochondrial genome are thus not included.

The Pharmacogenomics Knowledgebase (PharmGKB; www.pharmgkb.org): The PharmGKB database is a central repository for genetic, genomic, molecular and cellular phenotype data and clinical information about people who have participated in pharmacogenomics research studies. Stated mission is to develop, implement, and disseminate a public genotype-phenotype resource focused on pharmacogenetics and pharmacogenomics and serve a broad community including geneticists, molecular biologists, pharmacologists, physicians, policy makers and the lay public. The data includes, but is not limited to, clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathways, metabolic and transporter domains. Manual curation is done from research literature, submission of primary data and integration of data from a range of other sites like UniProt, HapMap etc. Database use is free for research use only.

Human Genome Variation database of Genotype-to-Phenotype (HGVBbaseG2P; www.hgvbaseg2p.org/index): HGVBbaseG2P provides a centralized compilation of summary level findings from genetic association studies, both large and small. The creators actively gather datasets from public domain projects, and encourage direct data submission. Submissions, Import of third party data (curators actively gather large datasets, such as Whole Genome Association Study findings, from public domain projects). HGVBbaseG2P aims to provide an extensive, centralized compilation of summary level findings from human genetic association studies.

Catalogue Of Somatic Mutations In Cancer (COSMIC; www.sanger.ac.uk/genetics/CGP/cosmic): COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers. The database contains information on publications, samples and mutations. It includes samples which have been found to be negative for mutations during screening therefore enabling frequency data to be calculated for mutations in different genes in different cancer types. In order to provide a consistent view of the data a histology and tissue ontology has been created and all mutations are mapped to a single version of each gene. The mutation data and associated information is extracted from the primary literature. This resource focuses on somatic mutations.

DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER; <http://decipher.sanger.ac.uk>): This database is devoted to store submicroscopic chromosomal imbalance that collects clinical information about chromosomal microdeletions/duplications/insertions, translocations and inversions and displays this information on the human genome map. Direct data submissions by a network of academic centres of Clinical Genetics. The focus is on medical care and genetic advise for patients in addition to supporting research.


 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

Single Nucleotide Polymorphism Database (dbSNP; www.ncbi.nlm.nih.gov/projects/SNP): The dbSNP database is meant to serve as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms. dbSNP takes the looser 'variation' definition for SNPs, so there is no requirement or assumption about minimum allele frequency. Very few dbSNP records have phenotype and disease descriptions. Originally, the great majority of data in dbSNP was collected and defined as variations simply using sets of co-aligned genomic or DNA sequences. Because this process typically had little to no focus on disease condition, only about 250 records in dbSNP were successfully associated with phenotype-causing variations or a clinical outcome in OMIM. Starting in the Spring of 2008, however, dbSNP began accepting submissions of Clinical/LSDB variations as well as annotations to existing variations (including phenotype). As of 10 May 2009, there are a total of 2220 records in dbSNP that were submitted as Clinical/LSDB variations or have OMIM links.

UniProt Knowledgebase (UniProtKB; www.uniprot.org): UniProtKB is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data. Phenotype related genotype changes are not the focus of UniProtKB, but are curated as part of the overall sequence-related annotation.

Database of Genotypes and Phenotypes (dbGAP; <http://www.ncbi.nlm.nih.gov/gap>): The database of Genotypes and Phenotypes was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Open-access data can be browsed online or downloaded from dbGaP without prior permission or authorization.

European Genome-phenome Archive (EGA; www.ebi.ac.uk/ega/page.php): The European Genome-phenome Archive (EGA) is designed to be a repository for all types of genotype experiments, including case control, population, and family studies. We will include SNP and CNV genotypes from array based methods and genotyping done with re-sequencing methods. This data may be either publicly available or limited access, depending on the design of the study. Like dbGAP, EGA was developed to archive and distribute the results of studies concerning the interaction of genotypes and phenotypes. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The content of these depositories is provided to

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

researchers only after successful application to appropriate Data Access Committees (DAC), and all data transfer involves encoded data files. This precaution is designed to prevent mischievous use of G2P datasets, such as the identification of research subjects whose identity and participation in a study needs to be kept secret. This data access mechanism also necessarily limits possibilities for broad analysis, integration, and sharing of the archived data.


Indian Genome Variation database (IGVdb; www.igvdb.res.in): The Indian Genome Variation database aims to provide data on validated SNPs and repeats, both novel and reported, along with gene duplications, in over a thousand genes, in 15,000 individuals drawn from Indian subpopulations. The database is freely available for all academic use around the world. Discoveries arising out of the IGV project will be IPR-protected and will be licensed for commercial exploitation.

Disease specific databases using a standard platform: Such resources include Alzgene (www.alzgene.org), PDGene (www.pdgene.org/) and SZGene (www.szgene.org/)

2.2.2. Discussion and conclusions

The number and diversity of these genomic databases reinforces the importance of the subject. At the same time, it highlights the limitations of the different solutions. Two related dimensions in which databases differ are the level of granularity and verification of the data. At one end of the spectrum are user submitted experimental results from large scale experiments, which may not have undergone independent verification or peer review, and where the data represent inputs for research on phenotypical association rather than validated findings (e.g. dbSNP). At the other end are manually crafted databases of peer reviewed research results on mutations, which have undergone editing in addition to the review process and where the phenotype to genotype relationship has been validated in study (e.g. HGMD, OMIM). Both approaches are valuable but lead to structurally very different results. Other dimensions in which the resources differ are focus on certain kinds of mutation (e.g. germ line or somatic, SNPs or deletions/insertions), how much of the supporting experimental data supporting the conclusion is stored, detail in description of the phenotype, focus on certain areas of research like diseases, populations or even individual genes, or focus on certain applications like pharmacogenomics, medical counselling, and bioinformatics infrastructure.

To conclude, the many different purposes and approaches expressed in the databases reviewed make it clear that if there is a new kind of data generated, it will soon spawn the creation of databases. The challenge seems to be less one of finding missing pieces in the data puzzle, and more one of fitting together a large number of pieces that are cut to have extensive overlap. It appears improbable to the author that one system to integrate all of the content could be devised, and questionable whether such a system would be desirable, even if it could, if only because of the resulting size and complexity. However, a number of concepts like genes and sequence positions are shared between many databases and could act as a common basis to bring

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

together data that is not present in all the databases, like occurrence frequencies, disease information or experimental detail, in a kind of mix and match, plug-in architecture.

2.3. Specification for Locus Reference Genomic Sequences


2.3.1. Rationale

The primary goal is to create a universally acceptable standard for a new specification for human genomic DNA Reference Sequences that addresses the primary shortcomings of existing systems, namely RefSeq, such as the lack of universally agreed genomic Reference Sequences for certain genes, the inconsistent and incomplete (and sometimes outdated and inappropriate) annotation of the existing ‘ad hoc’ reference sequences, the lack of long-term stability of reference sequences and, most importantly, the confusion in end-users’ minds concerning ‘versioning’ of DNA sequences. A solution would be the concept of a Locus Reference Genomic (LRG) sequence, which builds on the initial ideas from NCBI for RefSeqGene. Specifically, LRG is a system for providing a genomic DNA sequence representation of a single gene that is idealised, has a permanent ID (with no versioning), and core content that never changes (*i.e.* nucleotide sequence, transcripts, exons, start & stop codon positions). In other words, an LRG sequence and its core annotation are not meant to represent current biology knowledge, but to provide a standard for reporting variation in a stable coordinate system. The combination of the LRG plus the updatable-annotation layer will be used to support the biological interpretation of variants.

A Locus Reference Genomic (LRG) sequences is a permanently stable identifier for the combination of a genomic sequence plus its core ‘locked’ annotation. The sequence content and coordinates are defined by the sequence presented in its transcriptional orientation, accompanied by sufficient 5’ and 3’ flanking sequence for unique placement in the genome. LRGs will be written in the DNA alphabet and will encompass a single gene. In other words, overlapping genes encoded on the other strand, and hence transcribed in the opposite direction, will require their own separate LRGs. A gene might give rise to one or more transcripts. In the context of an LRG, the term “transcript” means a fully processed functional RNA that is either coding (*i.e.*, an mRNA) or is non-coding (*e.g.*, tRNA or long & short ncRNAs). To avoid any confusion, “transcript” is not synonymous with a primary transcription product (*e.g.*, hnRNA). Different transcripts from a gene might share exons, or regions of exons, in common.

For each transcript “t” (numbered, sequentially, with an Arabic numeral (*e.g.*, t1, t2, t3, *etc.*)) the following information will be annotated: (a) The sequence coordinates comprising the transcript, (b) For coding transcripts, amino acids will be numbered sequentially from the start to the stop codon, (c) The conceptual translation protein sequence, (d) Creation date (e) Molecule type (f) URL for LRG home page (*i.e.* <http://www.lrg-sequence.org>).

In the fixed layer, the gene will be defined by placement of standard transcripts. The exons so placed will not be assigned explicit identifiers (labels or names), but will be defined by their coordinates on the LRG, the cDNA, and the coding region. The updatable layer, however,

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

can be used to represent additional information about the same exons, or additional exons, thus allowing for both systematic and legacy numbering according to the needs of the stakeholders.

In summary, the fixed annotation layer contains just the DNA sequence, the coordinates and sequence of the major transcripts, the coordinates of the exons that comprise these transcripts and the sequence of each conceptual translated protein. No other information concerning the biology of the gene is recorded in the fixed annotation layer.

On the other hand, the updatable layer, which will be updated when necessary, provides coordinates to map the LRG onto the current genome build, onto legacy reference sequences and to support legacy exon- and amino-acid numbering systems. The sequence of an LRG is intended to be stable and not change whenever human genome assemblies are revised and updated. In addition, the sequence of an LRG may be a composite of several natural alleles.

2.3.2. Implementation


Ideally, there will be joint EBI/HGNC/NCBI implementation of LRGs to ensure that gene IDs and symbols are correctly assigned. Journal editors and LSDBs curators will be informed of LRGs with the intention of raising awareness of the advantages of using LRGs as reference sequences. It is hoped that journals might mandate the use of LRGs in the description of gene variants.

LRGs will be numbered sequentially with Arabic numerals with no leading zeros to pad the number to a fixed length (*i.e.* LRG_1, LRG_2, LRG_3, *etc.*). There will be no versioning of LRGs. Additional transcripts will be named by sequential integers (*i.e.* LRG_{n}t2, LRG_{n}t3, *etc.*). HGVS nomenclature will be use o report conventions

2.3.3. Outline of LRG production process

First of all, an initial LRG proposal should be prepared with the assistance with data stakeholders. Subsequently, all stakeholders should be provided with information about the proposed LRG and requested to review the proposed sequence and associated annotations. Ultimately, LRG formats should be generated for download, such as *LRG* (LRG DNA sequence and core (stable) annotations) and *LRGplus* (LRG DNA and protein sequences, and core and updateable annotations). The formats can be either in XML, based on the INSDSeq XML standard by INSDC or other useful formats, *e.g.* Genbank, EMBL, FASTA, gff, *etc.* that might be generated, if necessary, by transformation using XSL style sheets.

Also, tools will be required to visualise LRGs and their relationship to previous reference sequences (*e.g.* RefSeq and RefSeqGene entries) and to other related LRGs, otherwise the community acceptance of the LRG standard might be limited. Decisions about tool features ought to be informed by community requests and the primary goal should be ease of use. Such

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dagleish, and George P. Patrinos		Version: v1.3 – Final

tools could be based on NCBI Genome Workbench, Ensembl or the NGRl variant browser, but a tool dedicated to LRGs and variation reporting would be desirable. Functionality should include: (a) Selecting the appropriate LRG by HGNC gene symbol, (b) Displaying of alignment of previous reference to current (c) Alignment of DNA and protein sequences, (d) Ability to edit sequence in a “what if” fashion and view the consequences in terms of alterations to translation and/or splicing using HGVS-compliant nomenclature (e) Other tools, such as Mutalyzer, Mutation Checker, will need to be adapted to parse LRGs.

2.4. Bioresource Impact Factor (BRIF)

2.4.1. Introduction


The term Bio-Resource can refer to many different things, such as a biobank, a database, a set of bioinformatics software tools, a collection of well characterized animal strains etc. It is characterized by having various purposes and by involving various actors and institutions of the scientific and biomedical community, in academic, associative or industry contexts. To optimize the construction of a Bio-Resource, it needs to be made accessible in a clear manner, for the retrieving of data and/or samples. Addressing this by dealing with technical considerations is necessary, but this not sufficient for the development of a useful Bio-Resource. For example, there will often be resistance to populating a public database with data that a clinician has generated, so that the data can be accessed widely. Of course, some such issues are amenable to technical solutions, such as compatibility of data formats, but progress also needs to be made on ethical aspects such as the protection of rights to privacy and confidentiality of individuals whose data and/or samples are part of the Bio-Resource.

A particularly important consideration in the setting up and using a Bio-Resource that has value for the scientific and biomedical community, is ensuring a good level of recognition of the work it represents and the effort that is integral to sharing bio-resources. But such an arrangement is far from being typical practice. For this reason, to assist the concerned community, work has been done on the question of how to achieve incentivisation: to promote both the construction and the sharing of Bio-Resources.

As part of this work, a central concept that has been introduced by Cambon-Thomsen in 2003, and further developed in 2004, is that of a Bio-Resource Impact factor (BRIF)¹. There is presently no standardised and easy way to assess the importance of a Bio-Resource, nor to relate it to the scientific impact of the discoveries enabled by its use. To address this, the idea was formulated of constructing a quantitative parameter, modelled on the publication ‘Impact Factor’. Such a BRIF would make it possible to document; 1) the quantitative use of a Bio-

¹ CAMBON-THOMSEN A. Assessing the impact of biobanks. Nature Genetics (Correspondence), 2003, 34, (1) : 25-26

CAMBON-THOMSEN A. The social and ethical issues of post-genomic human biobanks *Nature Reviews Genetics*, 2004 5: 866-873

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

Resource, 2) the quality and the importance of research results involving it, and 3) the scientific and management recognition of those who set up and made available a valid Bio-Resource and their institution. The system would operate in a way similar to the citation impact factor that has been long established and used, along with other parameters, to assess journals and publications, to evaluate individual researchers or research team activities, etc. BRIF could be used much more rationally than “reputation” in the evaluation of the impact of researcher or institution Bio-Resource activities over time. Also, if such a factor was taken into account in assessing professional results, it would increase the quality and sharing of Bio-Resources. This is the desire of donors, and data providers. This idea has been discussed and referred to in the literature (ref Cambon-Thomsen 2003, cited 17 times and Cambon-Thomsen 2004, cited 35 times in web of science and more in Google scholar), but no practical implementation has yet occurred.

2.4.2. Objectives

Activities in GEN2PHEN in the context of Deliverable D3.2 involve; 1) disseminating the concept of BRIF, 2) exploring its practical aspects, 3) consulting with the community about it, 4) mapping it in the context of incentivisation of data sharing in modern biosciences, and 5) deciding upon further steps in order to perform a pilot implementation that will be developed towards Deliverable D9.3 and Deliverable D9.5.


2.4.3. Work performed in the course of GEN2PHEN

The 5 listed steps were accomplished through the following activities, over the first 18 months of the project. People involved in Inserm Toulouse were A. Cambon-Thomsen, A Pigeon, PA Gourraud, E Rial-Sebbag, A Mahalatchimy, M Thomsen, D Chartier. At GAM and other meetings, other Partners were also involved (especially Anthony Brookes and Gudmundur A. Thorisson).

The actions were: Debate, presentations at various meetings; questionnaires within GEN2PHEN, publications, preparation of a working group and of a web forum in the website of the project. The concept was presented in various settings, further discussed, and a plan of work has been developed.

1) Dissemination of the concept:

Within the GEN2PHEN Consortium: the concept of BRIF was discussed informally and also in an organized way at the GAM2; GAM3 and GAM4 meetings. In addition, a questionnaire that considered incentives, the required recognition, and tools suitable to achieve this, was introduced. This “Ethics questionnaire” has been circulated throughout the Consortium, and garnered 25 responses (see Deliverable 1.3).

	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

External to the GEN2PHEN Consortium: the concept was presented and discussed in various international and national meetings (listed in Appendix 4) during 2008 and the first half of 2009. This activity also built on previous discussions in various forums.

2) Exploration of BRIF practical aspects

First, to construct such an index or impact factor, we need to set up is a standardized way to refer to the Bio-Resources used in articles and publications.

- One task will be to set up a section and format for the citation; and this should involve Journal editors;
- Another task will involve defining an unambiguous system for the designation of bio-Resources (uniquely numbering each Bio-Resource); this aspect has been a focus of this activity, addressed by discussing the logic and format of such a numbering in various forums (see list of meetings in Appendix 4). A main question is who, or which institution(s), will be responsible for attributing the number and what would be its structure. A proposal has been put forward by Kauffmann & Cambon-Thomsen, 2008², that entails a hybrid between the ISBN type numbers and the numbering of clinical trials (attributed by WHO). Discussions have occurred in connection with other DOI systems (digital object identification).


Secondly, the parameters of the index itself have to be worked out and balanced (i.e., citation only, citation as unique bio-Resource, or combined with others such as unique author or multiple authors or collective authors in publications), and a decision made on whether the impact factor of the Journal where the article citing the Bio-Resource should be taken into account (cascade index), etc. The relation between BRIF and other initiatives like micro-attribution have been discussed, especially at the San Feliu Human Variome Planning meeting (May 2008), the European Society of Human genetics Conference and the P3G meeting in Barcelona, May-June 2008, and the Identifying Researchers on the Biomedical Web (IRBW2009), Workshop, Toronto, May 13-14, 2009.

Thirdly the advantages/disadvantages and pitfalls of using such an index need to be carefully considered.

Finally it is deemed necessary that work is done on a “prototype”, based on some case studies and on a dedicated working group, and this shall be developed in relation with the BBMRI prototype (BBMRI: Biobank and biomolecular research infrastructure, part of the ESFRI EU initiative in FP7).

3) Consultation of the community about BRIF

² KAUFFMANN F., CAMBON-THOMSEN A. **Tracing biological collections: between books and clinical trials.** *JAMA*, 2008;299(19): 2316-2318

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final


This was logically included in the general framework of incentives, as BRIF can only be one of the tools for this purpose.

Questions about incentives were posed in the context of the GEN2PHEN ethics questionnaire (see Deliverable 1.3) that received 25 responses. In addition, discussions at the various meetings confirmed that the issues of data sharing, tools to facilitate it, and systems of academic recognition, are major preoccupations of the community. Answers to the questionnaire showed that the matter of individual recognition was the main preoccupation, rather than institutional recognition. But this must be balanced against the fact that people who answered the questionnaire were individuals, not representatives of institutions. Discussion at the Data Release Workshop in Toronto, May 2009, Canada that involved funders, showed that institutions and funders view the institutional recognition aspect as very important. Thus the syntax and pilot studies that are envisaged have a high degree of feasibility.

4) Mapping of BRIF in the context of incentivisation to data sharing in modern biosciences

This activity is, of course, related to the others above, and it has thereby already been partly documented. Some aspects, however, are particularly relevant and important to underline.

- a) discussion with Editors especially at the “Human Variome Project planning meeting, San Feliu de Guixols May 25 – 29, 2008”, the ESHG Conference in Barcelona June 1-3 2009, the IRBW 2009, May 2009, Toronto Canada.
- b) discussion with funders and scientists at the Data release workshop in Toronto, Canada, May 2009
- c) discussion of policies for sharing data and for incentivization, in national and international projects in the context of the following initiatives:
 - biobank and biomolecular European research infrastructure (BBMRI): meetings in Hinxton, February 2008, BBMRI Steering Committee, Munich, February 2009, The biobank Conference: Harmonising Biobank Research: Maximising Value – Maximising Use 25-27 March 2009
 - the International Cancer Genomics consortium (ICGC), and the National Cancer Institute (INCa) in France, where cooperation and sharing of Bio Resources are at the heart of a task force
 - Ga2len FP6 Network of Excellence, the Final Conference of the PHGEN (Public Health Genomics Network), a DG SANCO funded EU project, Istanbul, November 2008
 - During the consultation period on the OECD guidelines on Human biobanks and genetic research databases (HBGRD), discussion on sharing policies occurred and this dimension is clearly the object of several recommendations (principles and guidelines) of this important document that should be made available by OECD in the fall of 2009. A. Cambon-Thomsen as part of the participants in this OECD

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

working group could reflect on the Community views as developed in GEN2PHEN activities.

It was highly apparent in these discussions that practical tools to facilitate exchanges were needed, besides the rewarding system, of which BRIF is an integral part. For example, a web based platform that would allow easy access and validation of information on legal requirements for exchanging samples and data (Ga2len contribution). Such a tool has been started and recently launched (www.hSERN.eu), and this was presented in 2 meetings (Oxford Conference on Governing biobanks in June 2008, with a chapter being in press in the book issued from the conference proceedings³, and as an invited conference in the Autumn P3G meeting 2008 by E Rial-Sebbag). It seems that facilitating Bio-Resource sharing, and making operational a rewarding system, are complementary activities.

- 5) to decide upon further steps in order to then perform a pilot study towards implementation that will be developed towards D9.3 and D9.5


The above activities make it possible to draw up the following plans for work to be started in Autumn 2009:

- A working group will be set up including scientists, editors, experts in ISI impact factors, lawyers, ethicists, bioinformaticians and experts in unique identifiers. Two meetings in addition to telephone and electronic communications will be held. This group should work out a set of parameters for the BRIF and prepare a report on the parameters, advantages and pitfalls and present several options and a SWOT analyses.
- A web forum will be open on the topic on the Knowledge Center, whose main conclusions will be taken into account by the working group.

Three pilot studies are planned;

- A study on some of the Bio-Resources (biobanks) that have volunteered as part of the prototype of BBMRI (CEPH Biobank, France; - EGP, The Estonian Genome Project of University of Tartu, Estonia; - HUNT Biobank, Norway; KORA-gen, Augsburg, Germany; THL, The Institute for Health and Welfare, Finland; UK Biobank, UK; UMCG, University Hospital Groningen, Netherlands)


³ RIAL-SEBBAG E, MAHALATCHIMY A, CHARTIER D, CAMBON-THOMSEN A. A new proposal to help researchers in the respect of legal requirements for the exchange of biological material: human sample exchange regulation navigator (hSERN) in *Governing Biobanks -What are the Challenges?* Jane Kaye and Mark Stranger Eds, Ashgate, 2009, (in press)

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dagleish, and George P. Patrinos		Version: v1.3 – Final

- A national group of laboratories having bio-Resources in the cancer domain in France (as part of INCA task force)
- A database within GEN2PHEN still to be defined, such as an LSDB.


2.4.4. Conclusions

The completed work on BRIF has been performed in order to prepare the ground for pilot studies and substantiation of the parameters of BRIF. Major steps have been taken to map this tool among the general picture of data sharing European and international debate, to re-consider this in the light of other tools and rewarding systems such as micro-attribution, to promote technical information on exchanges and sharing processes, to make progress towards the identification aspect of bio-Resources, to sensitize the community to this concept, and to work out the relations between sharing policies, rewarding systems and ethical dimensions. Coordination activities with other initiatives, both European and international⁴, have been a main part of the completed work. Plans are now clear for the next steps and ways to proceed.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	WP2 – Domain analysis and community relations		Security: PU
	Author(s): Christina Mitropoulou, Anne Cambon-Thomsen, Frank Schacherer, Raymond Dalgleish, and George P. Patrinos		Version: v1.3 – Final

References

1. Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 12:680–688.
2. Patrinos GP. 2006. National and Ethnic mutation databases: Recording populations' genography. *Hum Mutat* 27:879-887.
3. Cotton RG, Sallée C, Knoppers BM. 2005. Locus-specific databases: from ethical principles to practice. *Hum Mutat* 26:489-493.
4. Kaput J, Cotton RG, Hardman L, Watson M, Al Aqeel AI, Al-Aama JY, Al-Mulla F, Alonso S, Aretz S, Auerbach AD, Bapat B, Bernstein IT, Bhak J, Bleoo SL, Blöcker H, Brenner SE, Burn J, Bustamante M, Calzone R, Cambon-Thomsen A, Cargill M, Carrera P, Cavedon L, Cho YS, Chung YJ, Claustres M, Cutting G, Dalgleish R, den Dunnen JT, Díaz C, Dobrowolski S, Dos Santos MR, Ekong R, Flanagan SB, Flicek P, Furukawa Y, Genuardi M, Ghang H, Golubenko MV, Greenblatt MS, Hamosh A, Hancock JM, Hardison R, Harrison TM, Hoffmann R, Horaitis R, Howard HJ, Barash CI, Izagirre N, Jung J, Kojima T, Laradi S, Lee YS, Lee JY, Gil-da-Silva-Lopes VL, Macrae FA, Maglott D, Marafie MJ, Marsh SG, Matsubara Y, Messiaen LM, Möslein G, Netea MG, Norton ML, Oefner PJ, Oetting WS, O'Leary JC, de Ramirez AM, Paalman MH, Parboosingh J, Patrinos GP, Perozzi G, Phillips IR, Povey S, Prasad S, Qi M, Quin DJ, Ramesar RS, Richards CS, Savige J, Scheible DG, Scott RJ, Seminara D, Shephard EA, Sijmons RH, Smith TD, Sobrido MJ, Tanaka T, Tavtigian SV, Taylor GR, Teague J, Töpel T, Ullman-Cullere M, Utsunomiya J, van Kranen HJ, Vihinen M, Webb E, Weber TK, Yeager M, Yeom YI, Yim SH, Yoo HS. 2009. Planning the Human Variome Project. The Spain report. *Hum Mutat* 30:496-510.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

APPENDIX I

LSDBs/Diagnostic databases

Mutation databases of human genes are assuming an increasing importance in all areas of health care. In addition, more and more experts in the mutations and diseases of particular genes are curating published and unpublished mutations in locus-specific databases (LSDB). We analyzed 727 databases and their content for the presence or absence of 58 content criteria. No criterion studied gave unanimous agreement in every database. Our study provided a strong case for uniformity of data to make LSDB content maximally useful.


Introduction

Both completion of the human genome sequencing project and new methods for mutation detection, will lead to a tremendous increase of mutation identification in a growing number of genes. Consequently, the task of reporting and analyzing germ-line or somatic DNA variation will be a major challenge for the future of biological and medical science. Mutation databases are repositories in which allelic variations are described and assigned within a specific gene. Currently, three types of genetic databases are available: general (core) databases, locus-specific databases (LSDBs) and National/Ethnic mutation databases (NEMDBs).

LSDBs provide an invaluable tool for analyzing gene expression and phenotype in both normal and disease conditions, as the curators are closely in touch with molecular biologists very experienced with the analysis of a specific gene and its anomalies. This system generally promotes submission of data and maintains an accurate and up-to-date data source. We performed a thorough domain analysis of 727 existing LSDBs and analyzed the structure and content of each of the LSDBs currently available through the Internet.

Methods


We examined websites containing LSDBs of mutations of individual genes available through the HGVS website (<http://www.hgvs.org>). This site has an extensive list of 727 LSDBs with their URLs that are routinely updated as LSDB curators submit new databases to the list. There may be a few databases available and not included on the list that we have not been able to locate yet, but it is impossible to know for sure. The HGVS list may be viewed on the HGVS website.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations	Security: PU	
	Authors: Christina Mitropoulou and George P. Patrinos	Version: v1.4 –Final	2/12

LSDBs were examined for the presence or absence of 58 content criteria describing the general presentation, locus-specific information, database structure and data collection, fields in the mutation table, structure, content, or overall ease-of use of the database.

Table 1.

A. GENERAL PRESENTATION	24. Flat-file database
1. Explanation of content and aim	25. Relational database
2. Useful links	26. Password required for data access
3. Links to OMIM	27. Year of last update
4. Links to HGMD	28. Mutation visualization tool
5. Links to other LSDBs	29. Use of specific software
6. Database description published	D. FIELDS IN MUTATION TABLE
7. Copyright	30. Complete reference list
8. Counter	31. Summary phenotypic description
9. Disclaimer	32. Detailed phenotypic description
10. Language other than English	33. Links to references
B. LOCUS-SPECIFIC INFORMATION	34. Cross-reference with other databases
11. Reference sequence available	35. Restriction enzyme change
12. Information about disease	36. Information of ethnic group
13. List of patient associations	37. Mutation/allele frequency
14. Chromosome location	38. Detection method
15. Information on gene	E. DATABASE QUERYING
16. Information on protein function	39. Querying tool(s) available
17. HGVS nomenclature followed	40. Querying field: Mutation name
C. DATABASE STRUCTURE, DATA COLLECTION	41. Querying field: Gene region
18. Data collection via literature	42. Querying field: Codon number

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

19. Data submission by contacting the curators	43. Querying field: Author name
20. Direct online data submission	44. Querying field: Phenotype
21. Summary table listing all mutations	45. Querying field: Ethnic group
22. Downloadable mutation table	46. Querying field: Geographic location
23. Database in HTML format	47. Other fields for querying

These criteria (summarized in Table 1) were evaluated using binary scoring (0: Absence/No; 1: Presence/Yes) and have been selected on the basis of objectivity. Criteria which are based on a more or less subjective judgment by the evaluator (e.g. ease of use, LSDB design, *etc*) were excluded from our study.


Results

Of 727 LSDBs available to our knowledge, 59 were redundant. Also, 33 did not qualify as LSDBs, 21 LSDBs were not available to the general public and 9 were no longer available. This left 604 LSDBs for different nuclear genes available for study. By way of comparison, the Human Gene Mutation Database (HGMD) (Stenson *et al.* 2009) lists 2,426 genes in its public version (and 3,461 in its Professional 2009.2 Release) that contain at least one mutation. The redundancy found among LSDBs indicated that mutations were reported by two databases in the case of 49 genes, 9 databases in the case of three genes, and 4 different databases reported the gene TP53.

Criteria Examined

General Presentation of LSDBs

Almost 60% of the LSDBs had a home page that provided a clear explanation of content and aim of database and a minimum set of cross-references (active links and pointers) for the user to access additional information (Fig. 1). Important links included HGMD (22,2%), OMIM (Online Mendelian Inheritance in Man) for clinically related information (59,5%), other LSDBs (0,6%) and PubMed literature database for access to published references, GenBank/EMBL for detailed DNA sequence information, HGVS nomenclature website, and other useful links (84,1%). 66,1% of LSDBs advised users that information in the database is copyrighted intellectual property, such that they should cite the database in the appropriate manner when using data, while 50,9% had a disclaimer notice. In 38,5% of the cases studied, the database

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

description was published in a peer-reviewed scientific journal. Finally, only 0.6% of the LSDBs were written in a language other than English, namely French in all cases.

Data Collection and Submission (Data Source)

LSDBs were composed of mutation entries, such that each entry usually corresponds to a mutation in a single patient and is added (generally, but not always) after curator inspection. Almost all databases were a compilation of data derived from both published literature (99.6%) and submissions directly to the LSDB contributed by researchers throughout the world (96.7%). Direct online submission by contacting curators directly was available in 56.6% of the cases (Fig. 2).

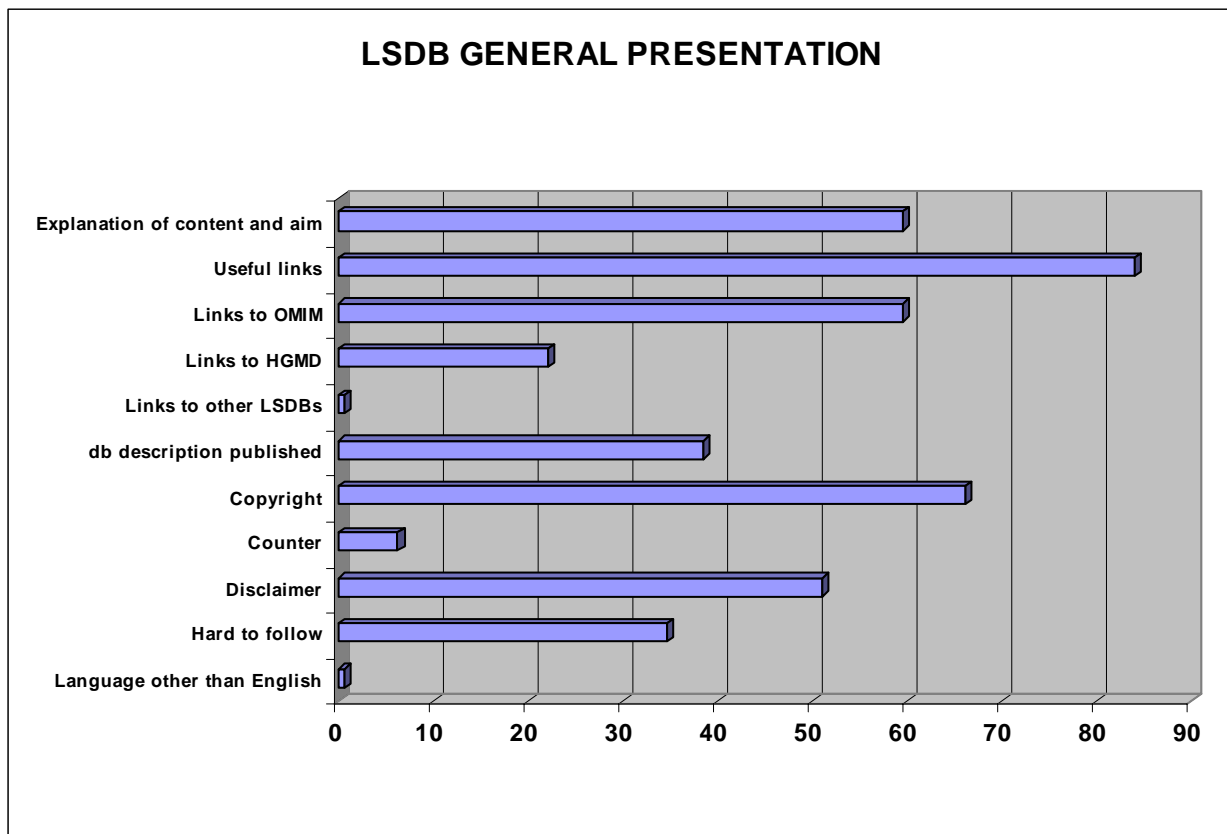




Fig.1. LSDB general presentation

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

Contrary to previous observations, very few LSDBs were structured as flat files containing a number of fields for each entry (1.6%). The majority of LSDBs were relational [SQL (Structured Query Language)-based; 68.6%]. However, a substantial number (29.2%) are still displayed through Hypertext Markup Language (HTML). This number, though is significantly lower, compared to the previous Claustres and coworkers (2002) study¹. A complete table listing all mutations was available in 78.4% of LSDBs, from which only a small portion included downloadable formats (19%) that were sometimes difficult or impossible to download.

Some databases showed mutation maps or mutation visualization tools (22.5%) depicting the location of mutations throughout the gene (or even the protein) sequence, and a few added graphical displays, including dynamic graphing tools. Because there is no standard yet, there are still a substantial amount of LSDBs are custom-based, resulting in data content heterogeneity. It is rather encouraging though that 40.4% of LSDBs are based on a Database Management System namely LOVD (141 LSDBs), MUTbase (116 LSDBs) and UMD (10 LSDBs) that contributes to data uniformity (Fig. 3).

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

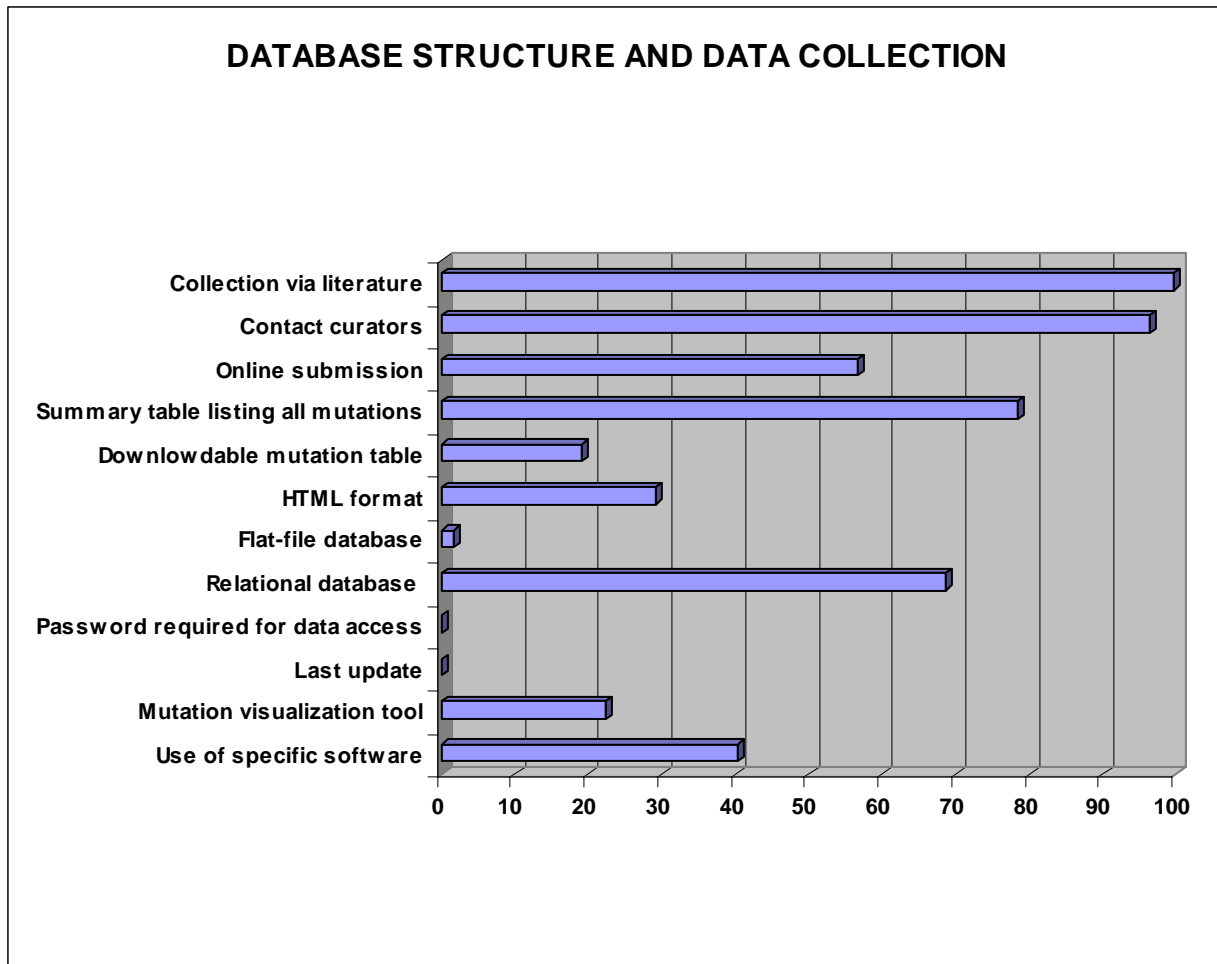



Fig. 2. Database structure and data collection

A large number of LSDBs are updated frequently. In particular, 513 out of 727 LSDBs were updated recently (198 LSDBs in 2009, 188 LSDBs in 2008 and 127 LSDBs in 2007), accounting for 70.3% of the total LSDBs. However, almost 30% of the LSDBs were outdated, being updated in 2006 or earlier (Fig. 4).

Locus-specific information

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

A number of databases contained much information in addition to the list of mutations, making the registries valuable for physicians and scientists from many fields. About 77.1% of the LSDBs included the reference sequence, a vital piece of information that was, unfortunately, not available in the rest.

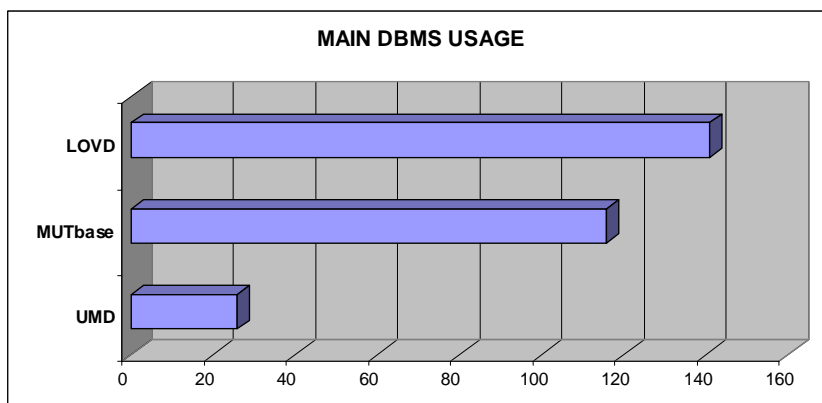


Fig. 3. Main DBMS usage

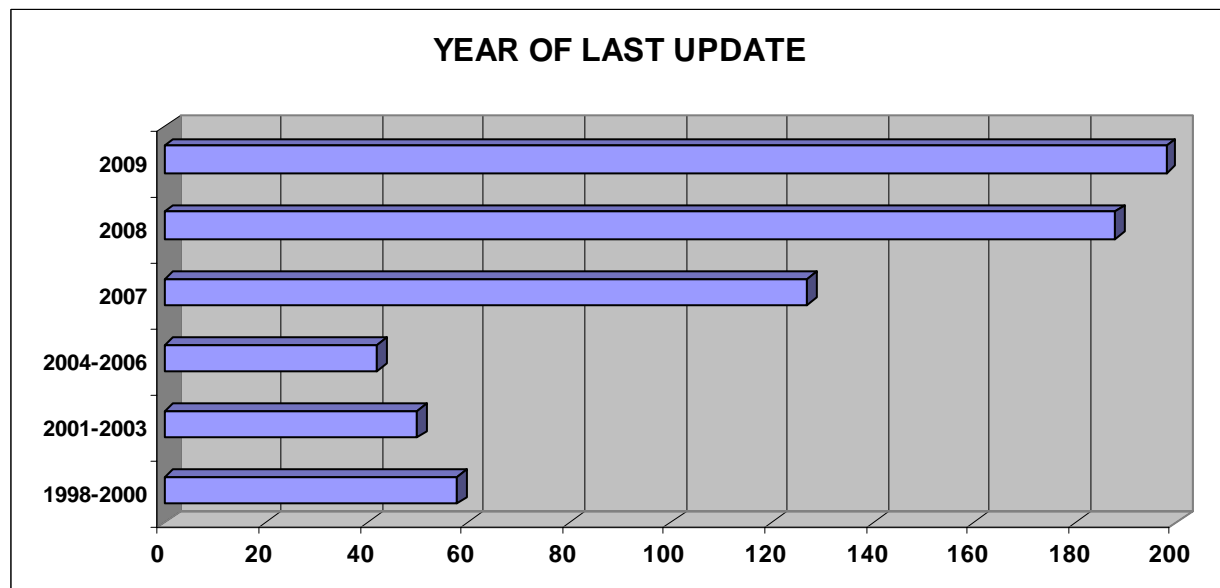



Fig. 4. Year of the last LSDB update

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

Also, 68.1% included information on the gene, and 54.6% of the chromosome location. However, only 35.9% provided information on the disease and 4% had information on protein function. Proper mutation nomenclature was followed in 56.6% of LSDBs. Finally, only 34.4% of LSDBs displayed links to patient associations or to websites with clinical content. (Fig. 5).

Fields in the mutation table

Information on the gene of interest, protein function, protein structure, and protein sequence alignment could be found in some LSDBs. A few of the LSDBs qualify as “knowledgebases” because they combine scientific and diagnostic data on mutations with associated information useful for clinicians or students (e.g., population distribution of alleles, haplotype associations) and information for patients and their families (e.g., treatment, diagnosis, dedicated organizations, or parent associations). Some LSDBs aimed to facilitate the detection and characterization of mutations by providing technical support in the form of primer sequences and mutation detection protocols.

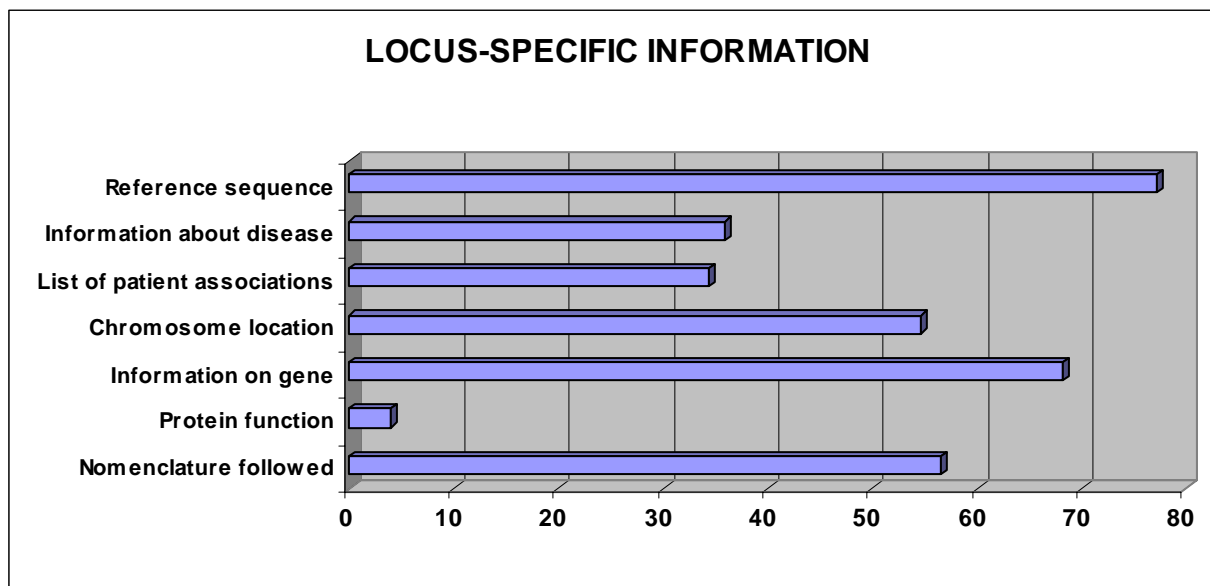



Fig. 5. Locus-specific information

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

Seventy nine (79)% of LSDBs included a complete reference list and in 90.4% of LSDBs references were directly linked to PubMed literature database. Summary phenotypic descriptions (i.e., pathogenicity: YES/NO, etc) were available in 53.7% of LSDBs whereas detailed phenotypic descriptions were available only in 12.6% of LSDBs analyzed. Only 2.1% of LSDBs were cross-linked with other LSDBs.

In addition to the mutation listing, many databases provided data fields for associated information, such as mutation detection methodology (25.6%), mutation frequency (2.7%), ethnic group/geographic origin of patients (45.9%), restriction enzyme change with mutation (37.6%). These data are summarized in Fig. 6.

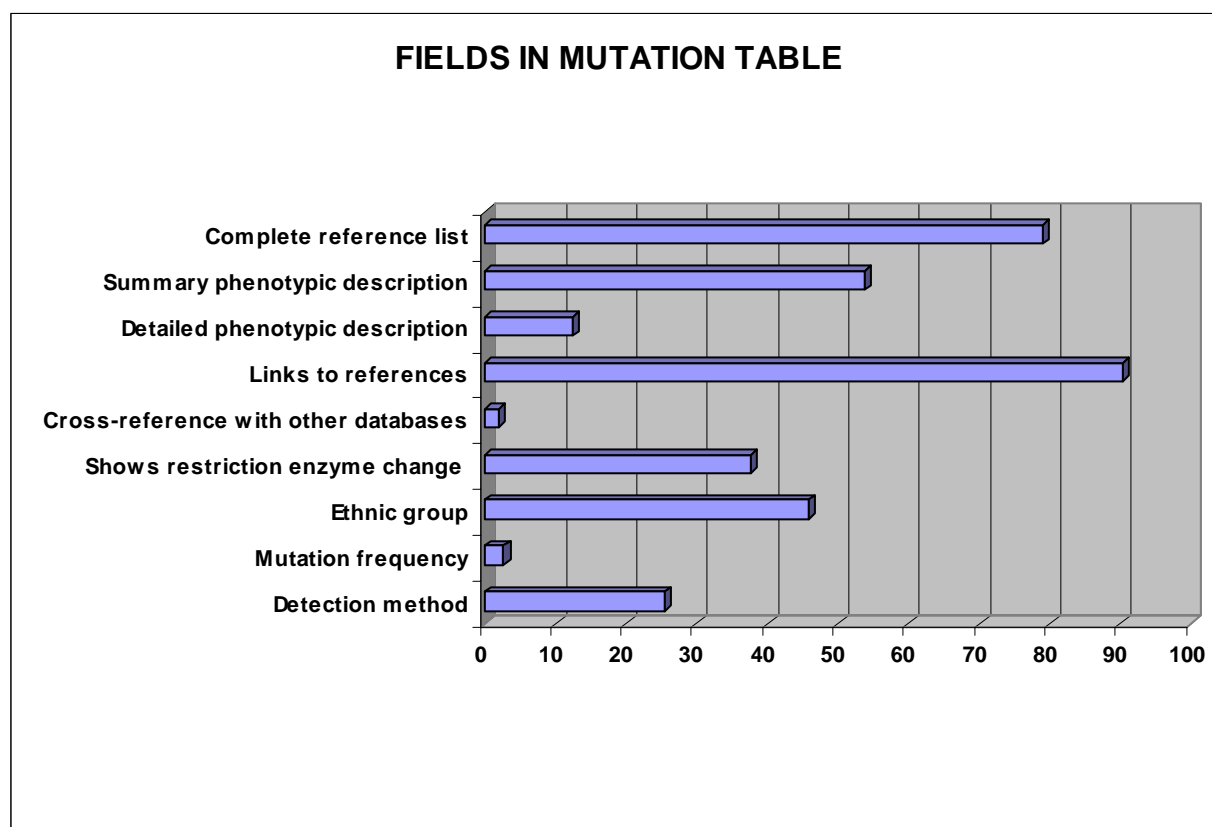



Fig. 6. Fields in mutation table

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

A search engine in 58.2% of LSDBs, mostly relational databases or those utilizing a DBMS, provided the opportunity to interrogate the database for specific information contained in a number of fields, such as mutation name (41%), gene region (33.8%), Codon number (32.3%), author name (44.3%), phenotype (44.7%), ethnic group (28.1%), geographic location (27.4%; Fig. 7). In 47.8% of LSDBs, other fields for querying were also available, such as cancer type and classification, DNA source, protein domain, etc.

Finally, a number of URLs that we accessed contained databases that did not qualify as LSDBs. In particular, 33 databases did not qualify as LSDBs, e.g. were MS Excel-based with no querying capacity. Also, 21 were password protected and registration for membership was requested to ensure that individuals using the database agree to a set of guidelines covering data submission, confidentiality, appropriate data use, and acknowledgment. These LSDBs were not publicly available and, hence, were not included in the analysis. Finally, 9 LSDBs were no longer available (Fig. 8).

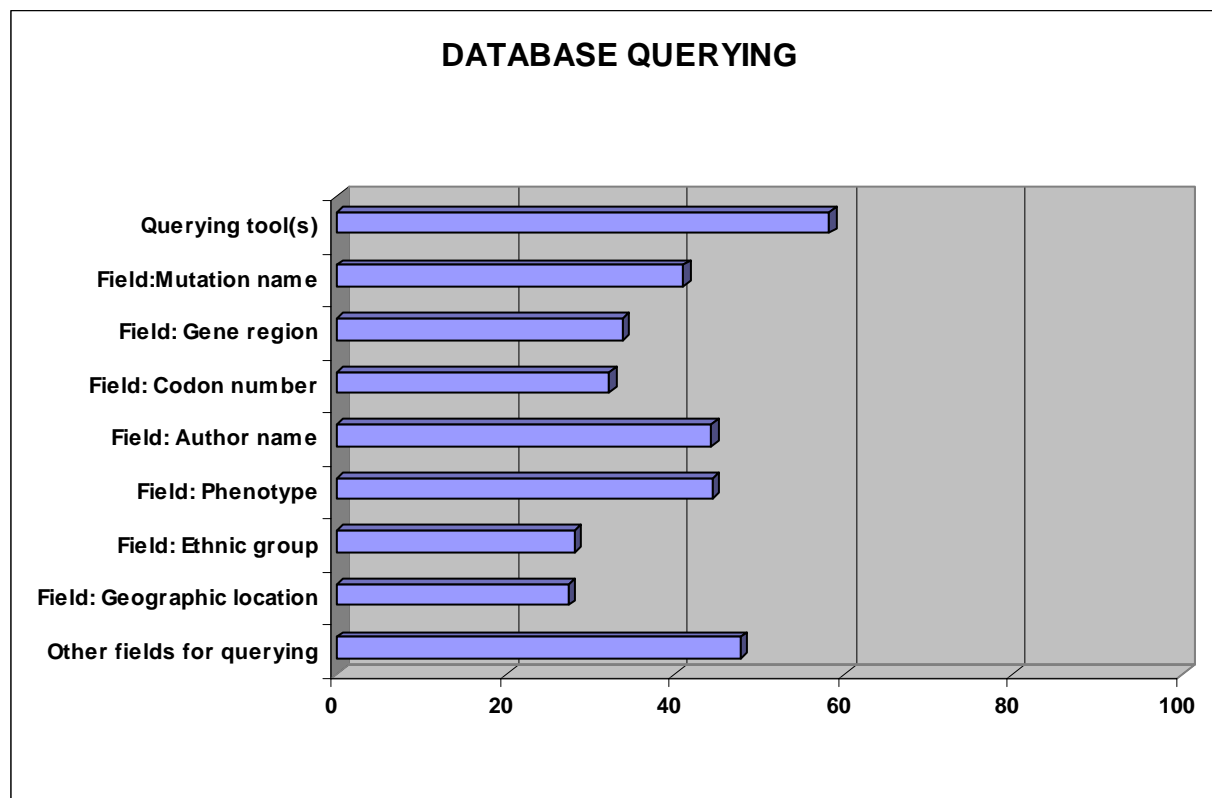



Fig. 7. Database querying

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

Conclusions

The main observations from our technical domain analysis of the LSDB field are:

- (a) A vital characteristic that would enhance the rate of new database creation would be the availability of “off-the-shelf” tailor-made software. Software needs to be able to allow collection, correction, and review and be able to store the data, both published and unpublished. This software needs two main functionalities: to allow operation by official curators, the software either remote or in a central facility, and to allow permanent storage of data. Approximately 40% of the LSDBs analyzed are based on a database management system, namely LOVD, UMD or MUTbase. Therefore, contrary to previous observations, there is less data content heterogeneity than previously reported¹.
- (b) Also, a large number of “LSDBs” are still displayed in HTML and lack basic querying tools, although the number is less than previously measured.

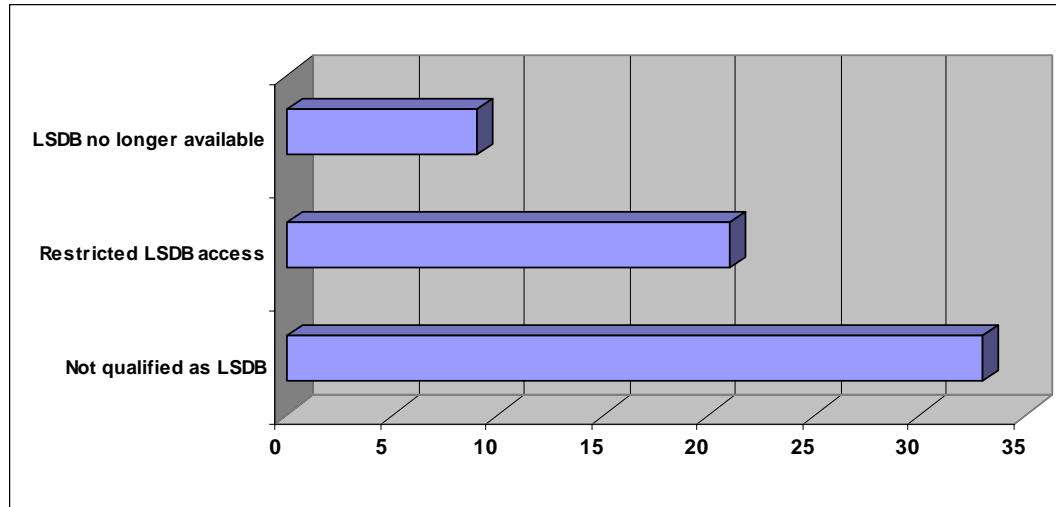



Fig. 8. Databases that do not qualify as LSDBs


- (c) Our data also showed that data visualization tools are still underdeveloped in the existing and there is much that needs to be done regarding means for visualizing data entries in LSDBs and diagnostic databases.
- (d) Finally, mutation frequency data are under-represented in LSDBs (4%), compared to ethnic group data (45%). This underlines the need for developing separate modules

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix I - LSDBs/Diagnostic databases		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Christina Mitropoulou and George P. Patrinos		Version: v1.4 –Final

recording ethnic-specific mutation frequency data in LSDBs or separate databases that serve that need, such as National/ethnic mutation databases that currently exist².

References

1. Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 12:680–688.
2. Patrinos GP. 2006. National and Ethnic mutation databases: Recording populations' genography. *Hum Mutat* 27:879-887.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

APPENDIX II

Central G2P Databases, State of the Art Catalog

This summary aims to document the existing genotype-phenotype databases (not central browsers) that cover the whole genome in their focus, summarizing what each project is trying and managing to achieve.

1. Whole-genome genotype-phenotype databases

1.1.1. Introduction


This report aims to give an overview on the currently available genotype to phenotype databases for *Homo sapiens* that cover the whole genome. It aims to highlight differences in approach, comprehensiveness, and focus of the major ones. The concern is for unique sources of data, not on interfaces or aggregation tools that can be built on top of this data. Understanding the universe of existing resources can facilitate the creation of a federated overall resource, as well as point out gaps in the current data.

Because the focus is on whole genome, general-purpose databases, those are covered in more detail. There is a large number of resources that cover specific areas, such as individual diseases or genes. Taken together, these also represent a general cross-section of available knowledge, but since the knowledge in them is fragmented and their number is large, covering each of them in detail is beyond the scope of this text. It must be noted however, that some systems have been used repeatedly to create locus specific databases, and the installed base of these tools might be seen as one distributed data source.

Excluded from review are tools that aggregate or display data from various databases, but do not collect and provide unique original data, like for example DiseaseCards (<http://www.diseasecard.org/>). It is acknowledged that there is value in this kind of aggregation being built on top of the original resources, but this is not the focus of this report.

1.1.2. Process

To identify the resources in this list, input from domain experts was combined with the results of an online search. All information on these resources was taken from the publicly available material on their respective web sites.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

1.1.3. List of data sources

1. Whole genome genotype phenotype databases

Short name **OMIM**

Full name **Online Mendelian Inheritance in Man**

URL <http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim>

Last observed 19 May 2009

Description OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources. OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine.


Data source Manual Curation from research literature

Availability Free

Pubmed Nucleic Acids Res. 2009 Jan;37(Database issue):D5-15. PMID: 18940862

Contents

	Autosomal	X-Linked	Y-Linked	Mito	Total
Gene with known sequence	12098	581	48	37	12764
Gene with known sequence and phenotype	349	25	0	0	374
Phenotype description, molecular basis known	2285	207	2	26	2520

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Mendelian phenotype or locus, molecular basis unknown	1596	141	5	0	1742
Other, mainly phenotypes with suspected Mendelian basis	1900	139	2	0	2041
Total	18228	1093	57	63	19441

Last update May 19, 2009

Comments OMIM focuses primarily on inherited, or heritable, genetic diseases.
/ Focus

Short name **HGMD**

Full name **Human Gene Mutation Database**


URL <http://www.hgmd.org/>

Last observed 19 May 2009

Description HGMD records all germ-line disease-causing mutations and disease-associated/functional polymorphisms reported in the literature, and provides these data in a readily accessible format to all interested parties, whether they are from an academic, clinical or commercial background. HGMD now constitutes, de facto, the central disease-associated mutation database available to the scientific community. The data comprise single base-pair substitutions in coding (e.g. missense and nonsense), regulatory and splicing-relevant regions of human nuclear genes, micro-deletions and micro-insertions, indels, repeat expansions, as well as gross gene lesions (deletions, insertions and duplications) and complex gene rearrangements.

Data source Manual Curation from research literature

Availability HGMD is freely available to registered academic/non-profit users. Mutation data are currently made available on this public website 2½ years after initial inclusion in the database. An up-to-date subscription version is available to both commercial and academic customers via

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final


license.

Pubmed Stenson et al. (2009) The Human Gene Mutation Database (HGMD®):
2008 Update. Genome Med. 1(1):13. PMID: 19348700


Contents	Public entries:	Total entries:
Mutation totals (as of 2009-05-19)	64970	88317
Gene symbol	2388	3337
cDNA sequence	2341	3078
Missense/nonsense	37423	49806
Splicing	6144	8548
Regulatory	858	1459
Small deletions	10545	14063
Small insertions	4229	5751
Small indels	944	1295
Gross deletions	3611	5303
Gross insertions	583	1053
Complex rearrangements	479	772
Repeat variations	154	267

Last update March, 2009


Comments /
Focus Various types of mutation within the coding regions, splicing and
regulatory regions of human nuclear genes causing inherited disease.
Somatic mutations and mutations in the mitochondrial genome are
thus not included.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations	Security: PU	
Authors: Frank Schacherer	Version: v1.4 –Final	5/16	


Short name	PharmGKB
Full name	Pharmacogenomics Knowledge Base
URL	http://www.pharmgkb.org/
Last observed	19 May 2009
Description	The PharmGKB database is a central repository for genetic, genomic, molecular and cellular phenotype data and clinical information about people who have participated in pharmacogenomics research studies. The data includes, but is not limited to, clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathways, metabolic and transporter domains.
Data source	Manual Curation from research literature, submission of primary data and integration of data from a range of other sites like UniProt, HapMap etc.
Availability	Free for research use only
Pubmed	T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart and R.B. Altman, "Integrating Genotype and Phenotype Information: An Overview of the PharmGKB Project" The Pharmacogenomics Journal (2001) 1, 167-170. PMID: 11908751
Contents	Annotations documenting the relationship of over 500 drugs, 450 diseases and 600 variant genes (Jan 2008). 1,575 annotated SNPs relating to human genetic variants that confer differential responsiveness to drugs and genome-wide association study data.
Last update	6 May 2009
Comments / Focus	Stated mission is to develop, implement, and disseminate a public genotype-phenotype resource focused on pharmacogenetics and pharmacogenomics and serve a broad community including geneticists, molecular biologists, pharmacologists, physicians, policy makers and the lay public.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Short name	HGVbaseG2P
Full name	Human Genome Variation database of Genotype-to-Phenotype information
URL	http://www.hgvbaseg2p.org/index
Last observed	02 June 2009
Description	HGVbaseG2P provides a centralized compilation of summary level findings from genetic association studies, both large and small. The creators actively gather datasets from public domain projects, and encourage direct data submission.
Data source	Submissions, Import of third party data (curators actively gather large datasets, such as Whole Genome Association Study findings, from public domain projects).
Availability	Free (restrictions based on original data sources apply)
Pubmed	G.A.Thorisson, O.Lancaster, R.C.Free, R.K.Hastings, P.Sarmah, D.Dash, S.K.Brahmachari, A.J.Brookes HGVbaseG2P: a central genetic association database. Nucleic Acids Research, (2009) 37:D797-802. PMID: 18948288
Contents	Data from 119 studies on various diseases and disorders, as well as core content from all marker (rsID) entries in dbSNP.
Last update	19 May 2009
Comments / Focus	HGVbaseG2P aims to provide an extensive, centralized compilation of summary level findings from human genetic association studies.
Short name	COSMIC
Full name	Catalogue Of Somatic Mutations In Cancer
URL	http://www.sanger.ac.uk/genetics/CGP/cosmic/

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Last observed	02 June 2009
Description	COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers. The database contains information on publications, samples and mutations. It includes samples which have been found to be negative for mutations during screening therefore enabling frequency data to be calculated for mutations in different genes in different cancer types. In order to provide a consistent view of the data a histology and tissue ontology has been created and all mutations are mapped to a single version of each gene.
Data source	The mutation data and associated information is extracted from the primary literature.
Availability	Free, with attribution requirement
Pubmed	The Catalogue of Somatic Mutations in Cancer (COSMIC). Forbes et al. Curr Protoc Hum Genet. 2008 Apr;Chapter 10:Unit 10.11. PMID: 18428421
Contents	Experiments 1111579 Tumours 339481 Mutations 78933 References 7386 Genes 4775 Fusions 2424 Structural Variants 40
Last update	28 May 2009
Comments / Focus	Focus is on somatic mutations.
Short name	DECIPHER

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Full name **DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources**

URL <http://decipher.sanger.ac.uk/>

Last observed 02 June 2009

Description A database of submicroscopic chromosomal imbalance that collects clinical information about chromosomal microdeletions/duplications/insertions, translocations and inversions and displays this information on the human genome map.

Data source Direct data submissions by a network of academic centres of Clinical Genetics.

Availability Free, with attribution requirement. Some data not available to the general public to protect patient privacy.

Pubmed Recommend citing the website, which contains a list of publications that made use of DECIPHER

Contents

All Patients	2644
Consented	1140
Syndromes	58
Array Types	42
Studies	147


Last update 18 Dec 2008

Comments / Focus Focus is on medical care and genetic advise for patients in addition to supporting research.


Short name **dbSNP**

Full name **Single Nucleotide Polymorphism Database**


URL <http://www.ncbi.nlm.nih.gov/projects/SNP/>

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Last observed	19 May 2009
Description	The dbSNP database is meant to serve as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms. dbSNP takes the looser 'variation' definition for SNPs, so there is no requirement or assumption about minimum allele frequency.
Data source	Submissions
Availability	Free
Pubmed	Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11. PMID: 11125122
Contents	Very few dbSNP records have phenotype and disease descriptions. Originally, the great majority of data in dbSNP was collected and defined as variations simply using sets of co-aligned genomic or DNA sequences. Because this process typically had little to no focus on disease condition, only about 250 records in dbSNP were successfully associated with phenotype-causing variations or a clinical outcome in OMIM. Starting in the Spring of 2008, however, dbSNP began accepting submissions of Clinical/LSDB variations as well as annotations to existing variations (including phenotype). As of 10 May 2009, there are a total of 2220 records in dbSNP that were submitted as Clinical/LSDB variations or have OMIM links.
Last update	30 April 2009
Comments / Focus	Focus is on nucleotide variation, phenotype association is only a side activity.
Short name	UniProtKB
Full name	UniProt Knowledgebase
URL	http://www.uniprot.org
Last observed	04 Jun 2009


 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Description	UniProtKB is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.
Data source	Manual Curation (SwissProt/reviewed part) and computational knowledge transfer (TrEMBL/unreviewed part)
Availability	Free
Pubmed	The Universal Protein Resource (UniProt) Nucleic Acids Res. 36:D190-D195(2008).
Contents	The reviewed section of UniProtKB in human returned 1775 protein results for annotation type mutagen. (There are also 14302 for annotation type natural variant, many from dbSNP, which do not generally have phenotypes associated).
Last update	26 May 2009
Comments / Focus	Phenotype related genotype changes are not the focus of UniProtKB, but are curated as part of the overall sequence-related annotation.
Short name	dbGAP
Full name	Database of Genotypes and Phenotypes
URL	http://www.ncbi.nlm.nih.gov/gap
Last observed	15 June 2009
Description	The database of Genotypes and Phenotypes was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

assays, as well as association between genotype and non-clinical traits.

Data source	By submission.
Availability	dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Open-access data can be browsed online or downloaded from dbGaP without prior permission or authorization.
Pubmed	The NCBI dbGaP database of genotypes and phenotypes. Mailman MD, et al. Nat Genet. 2007 Oct;39(10):1181-6. PMID: 17898773
Contents	19047 Variables, 2828 Analyses, 914 Documents, and 378 Datasets in 156 Studies.
Last update	June 8 th , 2009
Comments / Focus	Focus is on studies.
Short name	EGA
Full name	European Genome-phenome Archive
URL	http://www.ebi.ac.uk/ega/page.php
Last observed	15 June 2009
Description	The European Genome-phenome Archive (EGA) is designed to be a repository for all types of genotype experiments, including case control, population, and family studies. We will include SNP and CNV genotypes from array based methods and genotyping done with re-sequencing methods..
Data source	By submission.
Availability	This data may be either publicly available or limited access, depending on the design of the study.
Pubmed	The NCBI dbGaP database of genotypes and phenotypes. Mailman

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

MD, et al. Nat Genet. 2007 Oct;39(10):1181-6. PMID: 17898773

Contents No summary statistics available.

Last update June 8th, 2009

Comments / Focus Focus is on studies.

DbGAP and EGA were developed to archive and distribute the results of studies concerning the interaction of genotypes and phenotypes. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The content of these depositories is provided to researchers only after successful application to appropriate Data Access Committees (DAC), and all data transfer involves encoded data files. This precaution is designed to prevent mischievous use of G2P datasets, such as the identification of research subjects whose identity and participation in a study needs to be kept secret. This data access mechanism also necessarily limits possibilities for broad analysis, integration, and sharing of the archived data.

Short name **IGVdb**

Full name **Indian Genome Variation database**


URL <http://www.igvdb.res.in/>

Last observed 15 June 2009

Description The Indian Genome Variation database aims to provide data on validated SNPs and repeats, both novel and reported, along with gene duplications, in over a thousand genes, in 15,000 individuals drawn from Indian subpopulations.

Data source Samples collected using the Sequenom massarray system.


Availability Freely available for all academic use around the world. Discoveries arising out of the IGV project will be IPR protected and will be licensed for commercial exploitation.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Pubmed	Indian Genome Variation Consortium 2008 Genetic landscape of the people of India: a canvas for disease gene exploration. J. Genet. 87, 3–20. PMID: 18560169
Contents	75 genes on the discovery panel of 43 samples.
Last update	May 2009
Comments / Focus	Database is still under development.

2. Locus specific databases using a standard platform

Short name	LOVD
Full name	Leiden Open Variation Database
URL	http://www.lovd.nl
Last observed	03 June 2009
Description	LOVD is designed to provide a flexible, freely available tool for gene-centered collection and display of DNA variations.
Data source	Direct data submissions by maintainers of the individual installations.
Availability	The software is free. Sharing of data contained depends on the decision of the by maintainers of the individual installations.
Pubmed	Fokkema IFAC, Den Dunnen JT and Taschner PEM (2005). <i>LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-Box" approach</i> . Hum Mutat. 2005 Aug;26(2):63-8.
Contents	96,307 variants 42,439 patients

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

752 genes

41 installations

All numbers based on maintainers that share these data with the LOVD central site. There are many more non-public, local installs.

Last update 27 Apr 2009

Comments /
Focus

Short name **MUTbase**

Full name **MUTBase**

URL <http://www.lovd.nl>

Last observed 03 June 2009

Description MUTbase program suite provides an easy, interactive and quality controlled submission of information to mutation databases.

Data source Direct data submissions by maintainers of the individual installations.

Availability The software available through license.


Pubmed Riikonen, P. and Vihinen, M. (1999) Bioinformatics, 15, 852-859

Contents 122 public installations, 3 are under construction
These databases contain altogether data for 5359 patients. There are also 27 immunodeficiency mutation databases.

Last update 27 Apr 2009

Comments /
Focus Man of the public installations are hosted by the creators of MUTbase, the immunodeficiency databases are maintained by other groups.

LSDBs available on UMD platform (see <http://www.umd.be/>)

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases		
	Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations	Security: PU	
Authors: Frank Schacherer	Version: v1.4 –Final	15/16	

3. Disease specific databases using a standard platform

Alzgene (see <http://www.alzgene.org>)

PDGene (see <http://www.pdgene.org/>)

SZGene (see <http://www.szgene.org/>)

4. Further resources

BIOBASE Knowledge Library (mammalian): up-to-date protein or disease information that has been expertly curated from the published literature. Phenotypes are associated with genes and proteins. <http://www.biobase-international.com/index.php?id=469>

PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics


Metalife AG <http://www.phenomicDB.de>

HGV human genome variation database. A list of URLs for the ~650 internet-accessible LSDBs provided on the Human Genome Variation Society website. <http://www.hgvs.org/dblist/glsdb.html>

1.1.4. Discussion

The number and diversity of these databases reinforces the importance of the subject. At the same time, it highlights the limitations of the different solutions.

Two related dimensions in which databases differ is the level of granularity and verification of the data. At one end of the spectrum are user submitted experimental results from large scale experiments, which may not have undergone independent verification or peer review, and where the data represent inputs for research on phenotypical association rather than validated findings (e.g. dbSNP). At the other end are manually crafted databases of peer reviewed research results on mutations, which have undergone editing in addition to the review process and where the phenotype to genotype relationship has been validated in study (e.g. HGMD, OMIM). Both approaches are valuable but lead to structurally very different results.


 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix II - Central G2P Databases, State of the Art Catalog		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Frank Schacherer		Version: v1.4 –Final

Other dimensions in which the resources differ are focus on certain kinds of mutation (e.g. germ line or somatic, SNPs or deletions/insertions), how much of the supporting experimental data supporting the conclusion is stored, detail in description of the phenotype, focus on certain areas of research like diseases, populations or even individual genes, or focus on certain applications like. pharmacogenomics, medical counseling, and bioinformatics infrastructure.

1.1.5. Conclusions

The many different purposes and approaches expressed in the databases reviewed make it clear that if there is a new kind of data generated, it will soon spawn the creation of databases. The challenge seems to be less one of finding missing pieces in the data puzzle, and more one of fitting together a large number of pieces that are cut to have extensive overlap.

It appears improbable to the author that one system to integrate all of the content could be devised, and questionable whether such a system would be desirable, even if it could, if only because of the resulting size and complexity. However, a number of concepts like genes and sequence positions are shared between many databases and could act as a common basis to bring together data that is not present in all the databases, like occurrence frequencies, disease information or experimental detail, in a kind of mix and match, plug-in architecture.

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix III - The concept of BRIF (Bio-Resource Impact Factor)		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Anne Cambon-Thomsen		Version: v1.4 –Final

APPENDIX III

The concept of Bio-Resource Impact Factor

Meetings where the BRIF concept and the rewarding systems were considered and discussed, as part of GEN2PHEN activities

A. 2008

1. *BBMRI kick off meeting* Hinxton, UK; 10-12 Feb 2008

Presentation of various aspects of the project for the infrastructure, incentive sbeing one of the aspect (Participants : A. Cambon-Thomsen, E. Rial-Sebbag)

2. *Translating ELSI: Ethical, legal and social implications of genomics. Conference NGHRI*, Cleveland, 1-3 May, 2008.

Poster: Ethics and policies for sharing data from biobanks E Rial-Sebbag, F Kauffmann, A Cambon-Thomsen


3. *Human genome variome planning meeting*, San Feliu de Guixols May 25 – 29, 2008

Invited conference by A Cambon-Thomsen on “Sharing samples and data: ethics embedded in the tools, the tricks and the incentives” authored by Cambon–Thomsen A, Rial–Sebbag E, Ducournau.P, Gourraud PA, Milanovic F, Kauffmann F

Participation in publication Kaput et al. **Planning the Human Variome Project: The Spain Report** *Human Mutation*, Hum Mutat. 2009 Jan 20;30(4):496-510

4. *P3G and PHOEBE steering committee meeting*, Barcelona, May 30 – 31 2008 (Presentation of the article in JAMA, 2008 on numbering Bioresources, se footnote 2)

5. *ESHG Conference 2009*, June 1-3 2008, Barcelona; specific conversation with Myles Axton, Editor of Nature Genetics on incentives and microattribution (informal discussion).

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix III - The concept of BRIF (Bio-Resource Impact Factor)		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Anne Cambon-Thomsen		Version: v1.4 –Final

6. International Conference: Governing biobanks: what are the challenges? Oxford, June 24–26, 2008

Presentation by E Rial-Sebbag : hSERN: human sample exchange regulation navigator. A new proposal to help researchers in the respect of legal requirements for the exchange of human biological material

Invited contribution on Disciplinary and regional challenges in the context of multidisciplinary research and globalization. A CAMBON–THOMSEN

(Chapter in press, see footnote 3 in the main deliverable)

7. ESF Research Conference in Biomedicine “Biobanks: Introduction and next steps”: Sant Feliu de Guixols (Costa Brava), Spain 1 – 5 November 2008

Invited conference: Promoting the use of biobank resources: the challenges of exchanges and openness policies par A. CAMBON–THOMSEN

8. Post Genome Respiratory Epidemiology II: An Interdisciplinary Challenge”: November 6–8, 2008 Cernay, France


Invited Conference and animation of a round table discussion on Ethical issues in genetic research by CAMBON – THOMSEN A. The sharing policies were part of the discussion with a group of young scholars. (Collaboration with the Ga2len EU project)

9. Autumn annual meeting of thePublic Population Project in Genomics Consortium (P3G), November 10-11, 2008, Philadelphie, USA

Invited Conference by E Rial-Sebbag: A New Proposal To Help Researchers In The Respect Of Legal Requirements For The Exchange Of Human Biological Material (hSERN)

10. International Cancer Genome Consortium: First ICGC Scientific Workshop November 15–17, 2008 – Bethesda, USA.

Discussion by A. Cambon-Thomsen as member of the working group on Ethics and data access

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix III - The concept of BRIF (Bio-Resource Impact Factor)		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Anne Cambon-Thomsen		Version: v1.4 –Final

11. Networking meeting for EU-funded Biobanking projects; 20-21 November 2008; European commission, DG research, directorate for health, Brussels

12. Public Health Genomics European Network PHGEN Final Conference; Thursday 27th November 2008; Istanbul Turkey.

Invited conference on “Policy Development for Biobanking” PA Gourraud on behalf of A Cambon-Thomsen

B. 2009

1. Symposium INCa « Is genomic revolution changing our approaches to cancer and treatment? at the 20th ICACT (International Congress on Advanced Cancer Therapy) congress, Paris 04/02/09.

Invited Conference by A. Cambon-Thomsen on “Is the ethics landscape also changing? “ where sharing data aspects were part of the discussion.


2. Cancéropole Grand Sud, 12 March 2009 : Journée biobanques. Montpellier, France. A. Cambon-Thomsen (Discussions on sharing policies)

3. International Conference organized by PHOEBE, P3G and BBMRI: Harmonising Biobank Research: Maximising Value – Maximising Use; 25-27 March 2009, Brussels

Discussions on identification of bioresources and of individual contributors (Invited conference by F Kauffmann)

4. The biology of genomes 2009; – Cold Spring Harbor, USA, May 5 - 8, 2009:

Invited contribution by A. Cambon-Thomsen to the ELSI session and discussion on “Do we face genomics challenges of ethical issues or ethical challenges of genomics issues?” (sharing policies being one aspect developed and debated)

 HEALTH-200754	D2.3 – Technical State-of-the-art Document for G2P Databases Appendix III - The concept of BRIF (Bio-Resource Impact Factor)		
	WP2 – Domain analysis and community relations		Security: PU
	Authors: Anne Cambon-Thomsen		Version: v1.4 –Final

5. Data release workshop; May 12 & 13, 2009, Toronto, Canada organised by Funding agencies (Genome Canada, Wellcome Trust, European Commission and others).

A. Cambon-Thomsen, A Brookes among the delegation from Europe; a working group on incentives was organized and the rewarding systems and BRIF concepts discussed. A report is currently being circulated and will be published.

6. Identifying Researchers on the Biomedical Web (IRBW2009), Workshop; May 13-14 May, 2009, Toronto, Canada

Invited presentation by A. Cambon-Thomsen on “Collective authorship and microattribution”.

Discussions on the various ways of numbering and identifying not only individuals but also Bio-Resources

7. ESHG 2009 Conference; Vienna, May 23-26, 2009

Poster and oral presentation in an educational session (by A. Cambon-Thomsen)

Poster: Ethical issues and subsequent governance in the GEN2PHEN project, au A. Cambon Thomsen^{*^1} , A. Pigeon^{^1} , E. Rial-Sebbag^{^1}

, P. A. Gourraud^{^1} , M. Thomsen^{^1} , & GEN2PHEN consortium where the answers also on the incentive part were analysed and presented.

8. EBI – EMBL genotype to phenotype workshop; Wellcome Trust Genome Campus, Hinxton, UK, June 10, 2009

Invited Conference on “Ethical issues in genotype to phenotype domain”, by A. Cambon-Thomsen (sharing policies being one aspect developed and debated)

9. International Cancer Genome Consortium:

Second ICGC Scientific Workshop, Hinxton 23-24 June 2009 (E Rial-Sebbag, on behalf of A Cambon-Thomsen, member of the working group on Ethics and data access).