



HEALTH-F4-2007-200754

www.gen2phen.org

D 5.2 First report on Tools for Data Collection for Genomics G2P Databases

WP5 – Genomics G2P Databases

**V1.1
Final**

Lead beneficiary: ULEIC
Date: 10/02/2010
Nature: Report
Dissemination level: PU (Public)



 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

TABLE OF CONTENTS

DOCUMENT INFORMATION	3
DOCUMENT HISTORY	3
DEFINITIONS	4
1. EXECUTIVE SUMMARY	5
2. INTRODUCTION	6
2.1. GENETIC ASSOCIATION STUDIES.....	6
2.2. GENOME WIDE ASSOCIATION STUDIES (GWAS)	6
2.3. G2P DATA SOURCES.....	7
2.4. GWAS DATA AVAILABILITY	7
3. DATA COLLECTION	7
3.1. TYPES OF DATA	8
3.1.1. <i>Marker database</i>	8
3.1.2. <i>Study database</i>	9
3.1.3. <i>European Genome-phenome Archive (EGA)</i>	10
3.1.4. <i>dbGaP</i>	10
3.1.5. <i>Broad Institute Diabetes Genetics Initiative</i>	11
3.1.6. <i>1958 Birth Cohort (B58C)</i>	11
3.1.7. <i>NHGRI GWAS catalog</i>	11
3.1.8. <i>Open Access Database of Genome-wide Association Results</i>	13
3.1.9. <i>The Human Gene Mutation Database (HGMD®)</i>	13
3.1.10. <i>Other sources</i>	13
3.1.11. <i>Direct submissions</i>	14
4. SOFTWARE FOR G2P DATA COLLECTION AND VALIDATION	14
4.1. SUBMISSION PROCESS AND SUBMISSION TOOLS	14
4.2. GETTING DATA IN	15
4.2.1. <i>Marker import pipeline (dbSNP-lite)</i>	16
4.2.2. <i>Study metadata import pipeline</i>	16
4.2.3. <i>Frequency and association data pipeline</i>	17
4.2.4. <i>Study Feedback and Communication</i>	20
4.2.5. <i>Software for G2P federation</i>	21
ANNEXES	23
REFERENCES	26

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 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 5.2	Title	First Report on Tools for Data Collection for Genomics G2P Databases
Work package	Number 5	Title	Genomics G2P Databases

Delivery date	Contractual Month 24	Actual	10/02/2010
Status	Version 1.1		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)	University of Leicester (ULEIC)		
Responsible Author	R.K Hastings (rkh)	Email	rkh7@le.ac.uk
	Partner ULEIC	Phone	

Document History

Name	Date	Version	Description
Robert Hastings	21/01/2010	1.0	First draft
Robert Hastings	10/02/2010	1.1	Final draft

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

Definitions

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

ULEIC – University of Leicester (UK) – Coordinator

EMBL – European Molecular Biology Laboratory (Germany) – Beneficiary

FIMIM – Fundació IMIM (Spain) – Beneficiary

LUMC – Leiden University Medical Center (Netherlands) – Beneficiary

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

KI – Karolinska Institutet (Sweden) – Beneficiary

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

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

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

UH.FGC – Helsingin Yliopisto (Finland) – Beneficiary

UAVR – Universidade de Aveiro (Portugal) – Beneficiary

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

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

SIB – Swiss Institute of Bioinformatics (Switzerland) – Beneficiary

UNIMAN – The University of Manchester (UK) – Beneficiary


BIOBASE – BioBase GmbH. (Germany) – Beneficiary

deCODE – Islensk Erfoagreining EH (Iceland) – Beneficiary

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

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

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

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final


1. EXECUTIVE SUMMARY

Current solutions for the reporting and archiving of genetic association data are far from optimal. A key part of the GEN2PHEN project was providing genomics database software to fill this role, in the form of the Human Genome Variation Genotype to Phenotype database (HGVbaseG2P). This report summarises the steps taken in this project to develop tools for Genotype-to-Phenotype (G2P) data collection from a number of different sources.

We have collected G2P data in the form of Genome Wide Association Studies (GWAS) from 8 different sources. These include GWAS from: Wellcome Trust Case Control Consortium (WTCCC); database of Genotypes and Phenotypes (dbGaP), Broad institute, NHGRI GWAS catalog, 1958 Birth Cohort (B58C); and the Open Access Database of Genome-wide Association Results (OADGAR). In addition to these resources, literature mining by the HGVbaseG2P team provided additional GWAS data to complement the publically available data.

Two main areas have been developed to deal with the complex data that is associated with GWAS studies, the meta-data import and the association data-import. A basic submission tool has been developed to encourage submissions and will be further developed to make the experience of submitting a study easier.

In the future HGVbaseG2P will be developed into a tool to allow researchers to deposit summary-level data into their own copy of HGVbaseG2P and publish to the wider scientific community results that they consider safe to share (*i.e.* the results that are published in scientific journals and supplementary information).

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

2. INTRODUCTION


2.1. Genetic association studies

Genetic association studies provide a means to explore the genetic basis of complex traits, such as disease and drug response. Recent improvements in genotyping technologies and in sample bio-banking have dramatically increased the scale and the accuracy of the data being produced (Kingsmore, Lindquist et al. 2008). The reporting of results from such studies is, however, far from optimal; they are typically disseminated in diverse and disconnected databases, journals and meetings. Negative studies are all too often not reported at all (Shields 2000). Consequently, there is no convenient way to gather together, compare and contrast findings from comprehensive subsets of related studies. This presents a major problem for the field, since association studies produce both positive and negative signals that may be real or false, and which can only be resolved by comparing independently generated data sets.

2.2. Genome wide association studies (GWAS)

A particular type of study, known as Genome Wide Association Studies (GWAS), further compounds the situation described above. GWA studies have, within the last few years, emerged as a powerful tool to assay thousands of the most common genetic variants, typically Single Nucleotide Polymorphisms (SNPs) and more recently Copy Number Variation (CNVs) and relate these variants to disease phenotypes or traits (Pearson, Manolio 2008).

However for the majority of diseases these variants only explain a small proportion of individual differences in disease predisposition, and further replication studies, follow up studies, and fine mapping studies are needed to ensure results are not false-positive.

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

2.3. G2P data sources


With large volumes of data being generated by G2P studies, resources are needed for the collection and storage of this data. Public archival databases of genetic association data such as the database of Genotypes and Phenotypes (dbGaP) ((Mailman, Feolo et al. 2007), <http://www.ncbi.nlm.nih.gov/gap>) and the European Genome-phenome Archive (EGA) (<http://www.ebi.ac.uk/ega/page.php>) provide an obvious potential solution of this problem. There are also some small disease-specific initiatives (Allen, Bagade et al. 2008, Bertram, McQueen et al. 2007, Hulbert, Smink et al. 2007), along with resources such as the NHRI GWAS catalog (Hindorff, Sethupathy et al. 2009) and the Open Access Database of Genome-wide Association Results (Johnson, O'Donnell 2009) providing summary level resources of published GWAS studies. However, none of these resources bring together a globally comprehensive list of GWAS studies, while enabling direct submission of smaller studies (both positive and negative) by research groups.

2.4. GWAS data availability

Data collection of summary-level GWAS (typically allele/genotype frequencies and association results) was until August 2008, available from major data repositories such as the National Institutes of Health (NIH) dbGaP database, the Wellcome Trust Case Control Consortium (WTCCC) website (<http://www.wtccc.org.uk>) and the EGA. However since a publication which suggested allele frequency or genotype counts from summary-level data do not mask an individuals identity within GWAS studies (Homer et al. 2008), this type of data has been restricted for use by approved users only and are not freely available for distribution, due to concerns about protecting the identities of participants in studies.

3. DATA COLLECTION

Building tools for G2P data collection was a key software deliverable in the GEN2PHEN

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

project. The result of this deliverable so far is the data collection and import of G2P data, more specifically Genome Wide Association Studies (GWAS) from different sources into the Human Genome Variation Genotype to Phenotype database (HGvbaseG2P) (Thorisson, Lancaster et al. 2009) (<http://www.hgvbaseg2p.org>). The data collection methods and tools developed during this initial data collection are discussed below in the context of HGvbaseG2P but are designed to be flexible enough that they can be adapted for use by other research groups. Ultimately the whole system (HGvbaseG2P platform and validation tools) will be available for researchers to set up their own G2P resource.


3.1. Types of data

The data content of HGvbaseG2P can be broadly split into two types,

- Genetic polymorphic variations (*i.e.* Single Nucleotide polymorphisms (SNPs)), which composes the Marker side of the database.
- Genetic association studies (typically Genome Wide Association Studies (GWAS)) composing the Study side of the database.

3.1.1. Marker database

The reference information layer of simple, sequence-level variation as archived in dbSNP is dynamic and changes constantly over time, as new data are gathered and existing records are altered. G2P database resources that rely on dbSNP data as a reference therefore need to deal with these changes in the source database. In the case of HGvbaseG2P, such change tracking is important in order to maintain a consistent link between association study data and a basal data layer of genetic marker data, such that references to primary marker information in the former can be properly updated to match changes in the latter. However, the marker revision information provided by dbSNP is not adequate for this task.

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

To address this need for more detailed change tracking, the "dbSNP-lite" processing pipeline was created (see WP Deliverable 'D7.1 dbSNP-lite Established' for a more detailed report on this tool).


Currently all markers from dbSNP build 129 are imported into the database using the dbSNP-lite tool (see section 4.2.1 and Work Package (WP) deliverable 'D7.1 dbSNP-lite Established' for more information). While the SNP layer is a stripped down version of dbSNP we have added additional information allowing us to track the status of markers from different dbSNP builds. This layer of markers helps to validate markers used in GWAS studies and prevent erroneous data being imported (such as incorrect alleles), tracks the status of each marker in dbSNP and provides information on whether a marker has been deleted, merged, or replaced.

In the future we plan to be able to import other types of genetic variation data, such as Copy Number Variation (CNV) to allow the inclusion of CNV association studies into HGVbaseG2P.

3.1.2. *Study database*

The study database of HGVbaseG2P consists of all the elements need to represent an association study. In HGVbaseG2P we can display different types of association studies, but typically we collect case and control and quantitative trait loci (QTL) studies. As of December 2009 we have 366 studies available.

A Study in HGVbaseG2P is similar in scope to a journal article, comprising information relevant to a given research question or set of related questions. A study may contain data and analysis results from one or more experiments, one or more Sample Panels of test subjects, and one or more Phenotypes. Sample Panels may be characterised in terms of various Phenotypes, and they also may be combined and/or split into Assayed Panels. The Assayed Panels are used as the basis for reporting allele/genotype frequencies (in 'Genotype Experiments') and/or genetic association findings (in 'Analysis Experiments'). Environmental factors are handled as part of the Sample

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

Panel and Assayed Panel data structures. This type of data is referred to from this point as the study-meta data (*i.e.* the data about the data).

To assist with the importing of experiment data and to ensure consistent and unambiguous representation of alleles and genotypes we have developed a new nomenclature system. For more information and examples please see http://www.hgvbaseg2p.org/docs/hgvbaseg2p_nomenclature_system.pdf. The system caters not only for simple sequence alleles and traditional presence/absence genotypes, but also copy-number variants and somatic variants, as well as quantitative and ratio classes of genotypes. It also offers a robust way to represent long alleles.


Apart from HGVbaseG2P there are a number of GWAS databases or GWAS data providers. Below is a brief summary of the resources we collect data from for the Study part of HGVbaseG2P and the types of data they provide.

3.1.3. *European Genome-phenome Archive (EGA)*

The European Genome-phenome Archive (EGA) (<http://www.ebi.ac.uk/ega/>) resource is designed to be a repository for all types of genotype experiments, including case control, population, and family studies. Currently this resource has data from 11 projects that may be either publicly available or limited access depending on the request of the authors. In HGVbaseG2P we currently have the Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control Consortium 2007) (originally from the WTCCC site). We are currently displaying no frequency or association data from the EGA in HGVbaseG2P.

3.1.4. *dbGaP*

The database of Genotypes and Phenotypes (dbGaP) is an archive that distributes the results of studies that have investigated the interaction of genotype and phenotype. This resource currently has 61 summary-level GWAS studies of which HGVbaseG2P initially had 4 GWAS before

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)	Version: v1.1 Final	11/28

summary-level association data sharing was halted. We are not showing any frequency or association data at the request of the authors of these studies due to issues raised in the Homer et al. 2008 paper and only the study-meta data is available.

3.1.5. *Broad Institute Diabetes Genetics Initiative*


The Diabetes Genetics Initiative (DGI) is a collaboration of the Broad Institute of MIT and Harvard, Lund University, and Novartis Institutes for BioMedical Research. (<http://www.broadinstitute.org/science/projects/diabetes-genetics-initiative/diabetes-genetics-initiative>). The DGI had made a Type 2 Diabetes study (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena et al. 2007) available which was imported into HGVbaseG2P. HGVbaseG2P is currently showing the study-meta data, however no frequency or association data are available at the request of the authors of these studies due to the Homer et al. 2008 publication.

3.1.6. *1958 Birth Cohort (B58C)*

This resource is a web presentation (<http://www.b58cgene.sgu.ac.uk/>) of genetic data derived from the British 1958 birth cohort DNA collection, a national research resource created with funding from the Wellcome Trust and the Medical Research Council (MRC). We have an agreement to show all p-values for each of the Quantitative Trait Loci (QTL) shown in Table 1 and we plan to release this data in Spring 2010.

3.1.7. *NHGRI GWAS catalog*


The National Human Genome Research Institute (NHGRI) has created a summary-level online catalog of SNP-trait associations from published genome-wide association studies (<http://www.genome.gov/26525384>) known as the NHGRI GWAS catalog (Hindorff, Sethupathy et al. 2009). The researchers running this site manually mine the list of GWAS studies contained on the HuGE Navigator site (<http://hugenavigator.net/>) via PUBMED. As of January 2010 the GWAS catalog has 472 studies of which over 300 are present in HGVbaseG2P.

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

We are currently working on the next update due in spring 2010 to capture the remaining studies in the NHGRI GWAS catalog that are not present in HGVbaseG2P at this time.

Phenotype	Description
log ₁₀ IgE (kU/L)	Log (base 10) serum total immunoglobulin E concentration at age 44-45, adjusted for sex, date & time of collection, postal delay, laboratory batch and nurse (which also adjusts for area of residence)
systolic BP (mm Hg)	Average of 3 systolic blood pressure measures at age 44-45, adjusted for sex, sphygmomanometer, date & time of test, air temperature and nurse (which also adjusts for area of residence)
adult BMI (kg/m ²)	Body mass index at age 44-45 adjusted for sex, floor surface and and nurse (which also adjusts for area of residence)
adult height (cm)	Height measured at age 44-45, adjusted for sex and nurse (which also adjusts for area of residence)
FEV1 (litres)	Highest forced expiratory volume in 1 second at age 44-45, adjusted for sex, height, month of test, recent chest infection and nurse (which also adjusts for spirometer and area of residence)
4 kHz threshold (dB)	Hearing threshold at 4 kHz in the better ear at age 44-45, adjusted for sex, audiometer and nurse (which also adjusts for area of residence)
cholesterol (mmol/L)	Serum total cholesterol at age 44-45, adjusted for sex, date & time of collection, postal delay and nurse (which also adjusts for area of residence)
log ₁₀ HbA1c (%)	Log (base 10) glycosylated haemoglobin at age 44-45, adjusted for sex, date & time of collection, postal delay and nurse (which also adjusts for area of residence)
log ₁₀ fibrinogen (g/L)	Log (base 10) plasma fibrinogen at age 44-45, adjusted for sex, date & time of collection, postal delay, laboratory batch and nurse (which also adjusts for area of residence)
birth weight (kg)	Birth weight, adjusted for sex, gestational age and nurse (which also adjusts for area of residence)

Table 1: List of studies imported into HGVbaseG2P from B58C

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

3.1.8. *Open Access Database of Genome-wide Association Results*

The **Open Access Database of Genome-wide Association Results** (OADGAR, DOI: **10.1186/1471-2350-10-6**) is a summary-level, open access, centralized database of significant published GWAS results. This resource provides G2P data and annotation in the form of Excel files (collected from journal publications, and study author websites) for 118 GWAS studies published up to March 1, 2008. We have currently incorporated 111 studies in HGVbaseG2P, and we will plan to update when this resource updates in the future.


3.1.9. *The Human Gene Mutation Database (HGMD®)*

The Human Gene Mutation Database (HGMD®) (<http://www.hgmd.cf.ac.uk/>) represents an attempt to collate known (published) gene lesions responsible for human inherited disease. HGVbaseG2P has arranged a collaboration with HGMD® to provide a track of HGMD® data on the HGVbaseG2P browser. This now allows users to observe where these mutations are in the human genome while investigating GWAS data. The information HGVbaseG2P provides shows the HGMD® identifier, location of the mutation, gene name and provides a direct link to the gene page at HGMD®.

3.1.10. *Other sources*

We currently have 7 GWAS studies imported into HGVbaseG2P that we have mined from PUBMED searches of GWAS journal publications. We have taken to this approach to find GWAS studies that the NHGRI GWAS catalog did not have in their resource or did not have any data for.

We have also taken to approaching researchers directly to ask for their data and we are currently awaiting responses for these requests.

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

3.1.11. Direct submissions

HGVbaseG2P actively encourages G2P data submissions and has provided a way for researchers to submit their data to us in an easy and understandable way. Currently researchers are encouraged to submit their data into a simple excel sheet, split into six sections, study, panels, phenotypes, genotype experiments, association experiments and experiment data. This can be found on the HGVbaseG2P website at (http://www.hgvbaseg2p.org/docs/Submission_Data_Template.xls). A supporting document is available for descriptions on each of the six sections detailing what is essential for a submission to HGVbaseG2P (http://www.hgvbaseg2p.org/docs/Submission_Guidance_Notes.pdf). Submissions can be emailed directly to submissions@hgvbaseg2p.org.


As of Jan 2010 we have had one direct submission to HGVbaseG2P that is currently being processed and will be available on the HGVbaseG2P website in the next study database release scheduled for Spring 2010.

4. SOFTWARE FOR G2P DATA COLLECTION AND VALIDATION

This section describes the tools that have been developed for data submission, collection, and processing.

4.1. Submission process and submission tools

Summary-level GWAS data can be submitted to HGVbaseG2P via the submission template described in section 3.1.11. All data submitted to HGVbaseG2P remains the property of the data generators and/or submitters, and all records will be presented in the database with links and acknowledgements leading back to the original data source. Any users who might wish to obtain non-aggregated data are instructed to make suitable requests to the relevant submitter and their

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

data access authorities.

Submissions can be submitted with embargo dates or conditions attached. We still immediately process such datasets to ensure each submission is complete and useable, but we do not release the submitted data to the public until instructed to do so by the submitters.


When submitting genetic association data and/or allele/genotype frequency data to HGVbaseG2P, we require that the utilised Markers are all present in a major public marker/variation database (*e.g.*, dbSNP). If this is not the case, we assist in depositing the Markers into a suitable database.

To submit genetic association and/or allele/genotype frequency data, submitters are required to adhere to the information as specified in the http://www.hgvbaseg2p.org/docs/Submission_Guidance_Notes.pdf and paste it into the http://www.hgvbaseg2p.org/docs/Submission_Data_Template.xls form for submission. Each submission will equate to one Study in HGVbaseG2P, but each Study (*i.e.*, each submission) can include one or more Experiments.

Currently we are devising a software tool that will guide submitters through the process of gathering and checking their data in a ‘wizard style’ before submitting it. The tool will organise submission content into a compatible format and gather related information from sites across the Internet (*e.g.*, journal citation details, Marker Ids, and Allele specifications). The tool will also check for any inconsistencies in the total submission making it simpler for users to assemble and check their submissions before being imported into HGVbaseG2P. This tool will also have an added benefit in that submitters will be able to reuse components (*e.g.*, assay details, clinical materials, and phenotype descriptions) from previous submissions.

4.2. Getting data in

The data collected from publically available resources and direct submissions is imported into

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

HGVbaseG2P using specially written import pipelines, using the programming language Perl. These pipelines deal with marker data, study meta-data and frequency/association data. Each of the approaches for importing the types of data is described in more detail below.

4.2.1. Marker import pipeline (dbSNP-lite)

This pipeline is a collection of custom built Perl scripts that extract core data fields from dbSNP and synchronises them with a local copy of marker data in HGVbaseG2P (extracted from a previous dbSNP release, *i.e.* the Marker database see section and WP 7.1 ‘dbSNP-lite Established’). A key function of the tool is to compare marker entries in a new dbSNP release with existing marker entries in HGVbaseG2P and to highlight a key set of changes in the former.


The primary output of dbSNP-lite is a simplified or “lite” version of dbSNP marker data in a standard format, enhanced with revision information for each marker describing the changes such as marker and allele strand-flips, mergers and deletions if any were found.

Without this basal layer of marker information in HGVbaseG2P it would make validating and importing GWAS studies extremely difficult.

The dbSNP-lite output may have broader utility for others who wish to utilise dbSNP data in a similar way, and the tool can be extended to extract additional data elements from dbSNP if required. Capabilities for handling other reference sources of variation data, such as the Database of Genomic Variants (DGV: <http://projects.tcag.ca/variation/>) and dbVar (<http://www.ncbi.nlm.nih.gov/projects/dbvar/>) will be added in the near future as need arises.

4.2.2. Study metadata import pipeline

The meta-data import pipeline in HGVbaseG2P allows study-meta data (*i.e.* things like the number of participants in a study, the experiment details, author details and detailed phenotype information) to be imported for each study. The meta-data pipeline also performs a validation

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

check on any supplied PUBMED identifiers (available when a study has a publication) and uses this to gather more information, such as authors, abstract. The PUBMED identifier is also a quick and easy way of automatically detecting if a study has already been imported into HGVbaseG2P. The study-meta system is flexible enough to have very detailed manual curation (mined from a publication or from the author of the study) inputted by user or can be generated automatically from the import file, provided the correct fields are available. The study-meta data result files are extensible markup language (XML) files that can be further curated by the HGVbaseG2P team before being imported into the HGVbaseG2P database.


For single studies a specifically formatted XML file is manually populated, while in the case of composite data sources (*i.e.* files containing data from multiple studies), it is easier to generate multiple XML files automatically for each study with additional minor human annotation and curation when needed. The automatically generated XML files are generated using custom built Perl scripts.

In some studies within HGVbaseG2P we show just the study meta-data at the request of some data providers (WTCCC, dbGaP and Broad) as the frequency/association data is not considered to be safe by them to release to the general scientific community. This is due to the publication by Homer et al. 2008 describing the identification of individuals from summary-level data.

4.2.3. Frequency and association data pipeline

Once the study meta-data exists within HGVbaseG2P the frequency and association data can be imported. This pipeline consists of a number of Perl modules housing code for validation and error checking along with code to calculate missing data if needed (such as genotype and allele frequencies). A user of the pipeline describes all the rules for validation and calculation based on the fields in the G2P data file in a template configuration file.

This makes the frequency/association import pipeline a useful and flexible system for G2P data


 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

import, particularly as the customizable template-file based solution can deal with any type of G2P data format, allowing the rapid import of data from a number of different sources.

A typical frequency and association import follows these steps. First the pipeline checks to see if the study identifier described in the template configuration file exists in the study database (else there is no study or experiments for the data to be imported into). Any compound fields are separated (*e.g.* genotype frequencies (11/12/22)) and all fields to be imported are put into appropriate variables for import. Next if the genotype platform is an Affymetrix one and the dbSNP identifier is not available (in some instances just the Affymetrix probe identifier is present in the association file) a pre-import step converts this identifier to a dbSNP identifier obtained from a lookup file provided by Affymetrix. This is an optional plugin to the import pipeline that is only utilised when the dbSNP identifiers are not present and the genotype platform is an Affymetrix one. This lookup feature will be extended to incorporate other local identifier lookups from other genotyping platforms if needed, Additional plugins can also be created to deal with other non-typical imports making the frequency and association a flexible import tool.


Next each SNP from the association dataset is checked against the information for the corresponding SNP in the Marker database reporting back if the SNP has been replaced by another SNP identifier, merged with another SNP identifier, deleted from dbSNP or remained unchanged. If the SNP does not exist with the Marker database, that SNP and all associated association data are not imported to the study.

Once the SNP identifier has been checked that it can be imported in HGVbaseG2P, validation checks on the supplied frequency and association data are performed. The supplied alleles are checked against the alleles for the corresponding SNP in the marker database to make sure they match and they are not strand-flipped. Furthermore, if the genotype and allele frequencies are not included in the G2P data file they are calculated from the genotype and allele counts and

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

imported into the database along with any association data.

Additionally for markers on the X chromosome, we check the heterozygosity rates for X chromosome markers (against a dataset known only to contain female participants) to check if there were many males and females used in the case and controls, if the hemizygous genotype class are not included in the association data. We do this because X chromosome markers in males cannot give rise to homozygous or heterozygous genotypes, as they are hemizygous for X chromosome sequence. Therefore it is not statistically valid to bundle hemizygous (male) and homozygous (female) genotypes together as it will lead to false association (Simpsons paradox) due to the distortions in gender percentages between cases and controls. If the validation check proves that both male and female X chromosome genotype classes were bundled together we exclude them from import and contact the study authors for additional clarification and information.

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

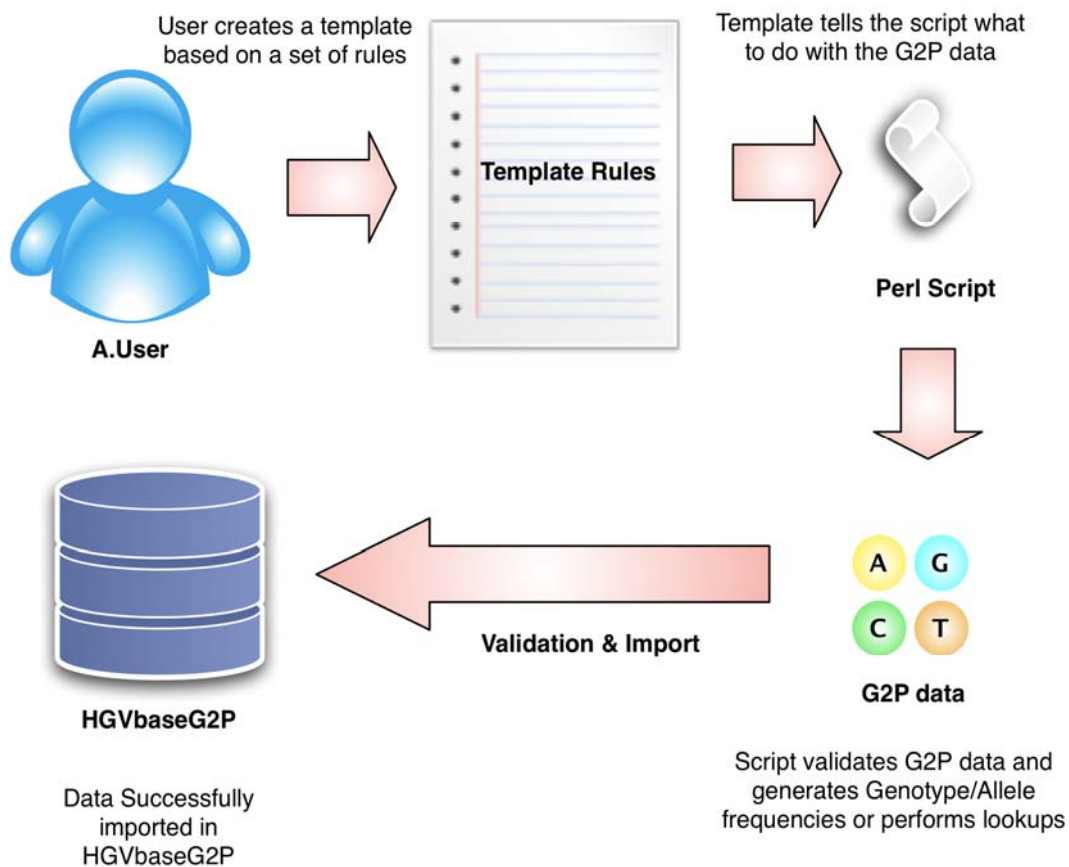



Figure 1: Basic steps of data import pipeline

4.2.4. Study Feedback and Communication

Markers in studies that do not pass the HGvbaseG2P frequency and association validation process are logged in an error file with an appropriate error code that describes the type of error detected. We contact the authors of the study directly via email with our concerns, or if the data was downloaded from dbGaP, WTCCC, Broad or NHGRI GWAS catalog we contact the appropriate resource. The email we send will explain any errors or inconsistencies we have found in their data and we will ask them if they also agree with what we have discovered and await their response or feedback. If they agree the data has errors they have two options, 1) if the errors are created from the way they have generated their association data they can re-run the analysis and we will then import the corrected data 2) If the data cannot be fixed (*i.e.* the error

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final


stems from the experimental stage) or they do not fix the data then we exclude the erroneous data from HGVBbaseG2P and detail any excluded markers in an error log made available with the data downloads on the HGVBbaseG2P site (this is currently switched off due to Homer et al 2008). In some instances where the researcher needs to visualise either the marker data or the study-meta data to see if HGVBbaseG2P has represented it correctly, we provide secure access to a test server so they can view their study and approve its content before release on the live HGVBbaseG2P site. Currently for each study we detail the number of markers successfully imported out of the total number genotyped to provide users of the site with an indication of a Study's completeness, but this will be extended to include information on the number of markers that did not pass the validation process so users of the site can get an indication of study quality.

4.2.5. Software for G2P federation

The aim is to make HGVBbaseG2P available 'in-a-box', initially as a 'virtual' system (hosted by ULEIC but controlled by the researcher) but later as a separate 'federated' solution that can be downloaded and setup on a researcher's own server. The federated versions of HGVBbaseG2P will also include all the validation and import tools described above that are needed to import studies to create a fully functional G2P resource.

Using these G2P tools any researcher can set up their own G2P study resource based on HGVBbaseG2P platform and administer access to data within. Options will be included to control the level of data access collaborators, or the general scientific community has to particular studies. Also, rather than include a large marker database, the in-a-box system will be able to link to the marker database on the central site (see Figure 2).

Options will also be provided for studies within the user's system to be made available on the central HGVBbaseG2P site. This will allow them to use the visualization tools to compare/contrast their study with the hundreds available in the full HGVBbaseG2P database. A curation tool will

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

also form part of the ‘in-a-box’ system allowing researchers to curate data to a high specification and ultimately provide them with a way to submit and publish data on their local system to the main HGVbaseG2P site.

This system will allow the more efficient data collection of G2P data, as researchers will be able import their own data easily saving time and effort while having the potential to share their data with a wider scientific audience if desired.

‘Public and Private’ HGVbaseG2P

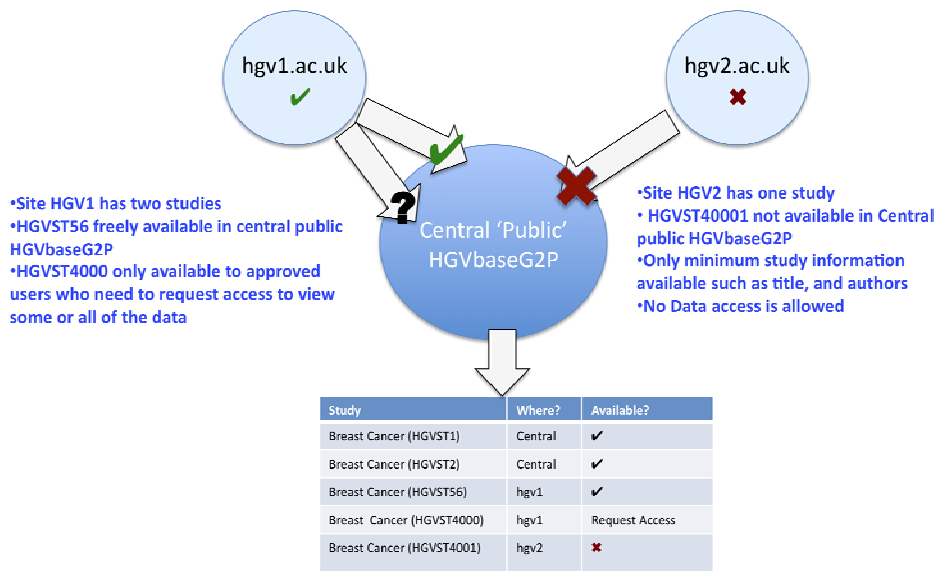



Figure 2: Federated HGVbaseG2P enabling researchers to collect and visualise data G2P data

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

ANNEXES

Annex I - ABBREVIATIONS

Abbreviations

HGVbaseG2P

Full name

Human Genome Variation Genotype to Phenotype database

G2P

Genotype-to-Phenotype

GWAS

Genome Wide Association Studies

WTCCC

Wellcome Trust Case Control Consortium

dbGaP

database of Genotypes and Phenotypes

NHGRI GWAS catalog

National Human Genome Research Institute: - A Catalog of Published Genome-Wide Association Studies

B58C

1958 Birth Cohort

OADGAR

Open Access Database of Genome-wide Association Results

dbSNP

Database of Single Nucleotide polymorphisms

SNP

Single Nucleotide Polymorphism

CNV

Copy Number Variation

EGA

European Genome-phenome Archive

NIH

National Institute of Health

NCBI


National Center for Biotechnology Information

DGI

Diabetes Genetics Initiative

DGV


The Database of Genomic Variants

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

dbVar Database of Genomic Structural Variation
MRC Medical Research Council
Quantitative Trait Loci QTL
HGMD The Human Gene Mutation Database

Annex II - WEB RESOURCES

Resource	URL
HGVbaseG2P	http://www.hgvbaseg2p.org/ Last accessed 18/01/2010
WTCCC	http://www.wtccc.org.uk/ Last accessed 18/01/2010
dbGaP	http://www.ncbi.nlm.nih.gov/gap Last accessed 18/01/2010
NHGRI GWAS catalog	http://www.genome.gov/26525384 Last accessed 18/01/2010
B58C	http://www.b58cgene.sgul.ac.uk/ Last accessed 18/01/2010
OADGAR	DOI: http://dx.doi.org/10.1186/1471-2350-10-6 Last accessed 18/01/2010
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/ Last accessed 18/01/2010
EGA	http://www.ebi.ac.uk/ega/ Last accessed 18/01/2010
DGI	https://www.broadinstitute.org/science/projects/diabetes-genetics-initiative/diabetes-genetics-initiative

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

Last accessed 18/01/2010

DGV

<http://projects.tcag.ca/variation/>

Last accessed 18/01/2010

dbVar

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

Last accessed 18/01/2010

HuGE Navigator


<http://hugenavigator.net/>

Last accessed 18/01/2010

HGMD®

<http://www.hgmd.cf.ac.uk/>

Last accessed 18/01/2010

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

References


KINGSMORE, S.F., LINDQUIST, I.E., MUDGE, J., GESSLER, D.D. and BEAVIS, W.D., 2008. Genome-wide association studies: progress and potential for drug discovery and development. *Nature reviews.Drug discovery*, **7**(3), 221-230.

SHIELDS, P.G., 2000. Publication bias is a scientific problem with adverse ethical outcomes: the case for a section for null results. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, **9**(8), 771-772.

ALLEN, N.C., BAGADE, S., MCQUEEN, M.B., IOANNIDIS, J.P., KAVVOURA, F.K., KHOURY, M.J., TANZI, R.E. and BERTRAM, L., 2008. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature genetics*, **40**(7), 827-834.

BERTRAM, L., MCQUEEN, M.B., MULLIN, K., BLACKER, D. and TANZI, R.E., 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature genetics*, **39**(1), 17-23.

DIABETES GENETICS INITIATIVE OF BROAD INSTITUTE OF HARVARD AND MIT, LUND UNIVERSITY, AND NOVARTIS INSTITUTES OF BIOMEDICAL RESEARCH, SAXENA, R., VOIGHT, B.F., LYSENKO, V., BURTT, N.P., DE BAKKER, P.I., CHEN, H., ROIX, J.J., KATHIRESAN, S., HIRSCHHORN, J.N., DALY, M.J., HUGHES, T.E., GROOP, L., ALTSHULER, D., ALMGREN, P., FLOREZ, J.C., MEYER, J., ARDLIE, K., BENGTSSON BOSTROM, K., ISOMAA, B., LETTRE, G., LINDBLAD, U., LYON, H.N., MELANDER, O., NEWTON-CHEH, C., NILSSON, P., ORHO-MELANDER, M., RASTAM, L., SPELIOTES, E.K., TASKINEN, M.R., TUOMI, T., GUIDUCCI, C., BERGLUND, A., CARLSON, J.,

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final


GIANNINY, L., HACKETT, R., HALL, L., HOLMKVIST, J., LAURILA, E., SJOGREN, M., STERNER, M., SURTI, A., SVENSSON, M., SVENSSON, M., TEWHEY, R., BLUMENSTIEL, B., PARKIN, M., DEFELICE, M., BARRY, R., BRODEUR, W., CAMARATA, J., CHIA, N., FAVA, M., GIBBONS, J., HANDSAKER, B., HEALY, C., NGUYEN, K., GATES, C., SOUGNEZ, C., GAGE, D., NIZZARI, M., GABRIEL, S.B., CHIRN, G.W., MA, Q., PARIKH, H., RICHARDSON, D., RICKE, D. and PURCELL, S., 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y.)*, **316**(5829), 1331-1336.

HINDORFF, L.A., SETHUPATHY, P., JUNKINS, H.A., RAMOS, E.M., MEHTA, J.P., COLLINS, F.S. and MANOLIO, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(23), 9362-9367.

HOMER, N., SZELINGER, S., REDMAN, M., DUGGAN, D., TEMBE, W., MUEHLING, J., PEARSON, J.V., STEPHAN, D.A., NELSON, S.F. and CRAIG, D.W., 2008b. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, **4**(8), e1000167.

HULBERT, E.M., SMINK, L.J., ADLEM, E.C., ALLEN, J.E., BURDICK, D.B., BURREN, O.S., CASSEN, V.M., CAVNOR, C.C., DOLMAN, G.E., FLAMEZ, D., FRIERY, K.F., HEALY, B.C., KILLCOYNE, S.A., KUTLU, B., SCHUILENBURG, H., WALKER, N.M., MYCHALECKYJ, J., EIZIRIK, D.L., WICKER, L.S., TODD, J.A. and GOODMAN, N., 2007. T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic acids research*, **35**(Database issue), D742-6.

JOHNSON, A.D. and O'DONNELL, C.J., 2009. An open access database of genome-wide association results. *BMC medical genetics*, **10**, 6.

 HEALTH-200754	D 5.2 First Report on Tools for Data Collection for Genomics G2P Databases		
	WP5: Genomics G2P Databases		Security: PU
	Author(s): Robert Hastings (ULEIC)		Version: v1.1 Final

MAILMAN, M.D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L., POPOVA, N., PRETEL, S., ZIYABARI, L., LEE, M., SHAO, Y., WANG, Z.Y., SIROTKIN, K., WARD, M., KHOLODOV, M., ZBICZ, K., BECK, J., KIMELMAN, M., SHEVELEV, S., PREUSS, D., YASCHENKO, E., GRAEFF, A., OSTELL, J. and SHERRY, S.T., 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*, **39**(10), 1181-1186.

PEARSON, T.A. and MANOLIO, T.A., 2008. How to interpret a genome-wide association study. *JAMA : the journal of the American Medical Association*, **299**(11), 1335-1344.

THORISSON, G.A., LANCASTER, O., FREE, R.C., HASTINGS, R.K., SARMAH, P., DASH, D., BRAHMACHARI, S.K. and BROOKES, A.J., 2009. HGVbaseG2P: a central genetic association database. *Nucleic acids research*, **37**(Database issue), D797-802.

WELLCOME TRUST CASE CONTROL CONSORTIUM, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661-678.