



HEALTH-F4-2007-200754

[www.gen2phen.org](http://www.gen2phen.org)

# **D5.1 Summary Document for Genomics Database V-1 Software**

**WP5 – Genomics G2P Databases**


**V1.01  
Final**

Lead beneficiary: ULEIC

Date: 03/08/2009


Nature: Report

Dissemination level: PU  
(Public)


 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security: PU</b>
	<b>Author(s): R. Free (ULEIC)</b>		<b>Version: v1.01 – Final</b>

## TABLE OF CONTENTS

<b>DOCUMENT INFORMATION</b> .....	<b>4</b>
<b>DOCUMENT HISTORY</b> .....	<b>4</b>
<b>DEFINITIONS</b> .....	<b>5</b>
<b>1. EXECUTIVE SUMMARY</b> .....	<b>6</b>
<b>2. INTRODUCTION</b> .....	<b>7</b>
2.1. GENETIC ASSOCIATION STUDIES.....	7
2.2. CURRENT SOLUTIONS.....	7
2.3. HGVBASEG2P V1 .....	7
<b>3. SYSTEM DESIGN</b> .....	<b>8</b>
<b>4. DATABASE DESIGN AND CONTENT</b> .....	<b>9</b>
4.1. MARKER DATABASE .....	10
4.2. STUDY DATABASE .....	10
4.2.1. <i>Studies</i> .....	10
4.2.2. <i>Samples</i> .....	11
4.2.3. <i>Phenotypes</i> .....	11
4.2.4. <i>Experiments</i> .....	11
<b>5. SOFTWARE COMPONENTS</b> .....	<b>11</b>
5.1. THE HGVBASEG2P APPLICATION PROGRAM INTERFACE (API).....	11
5.1.1. <i>Database/Schema Libraries</i> .....	11
5.1.2. <i>Web Library</i> .....	12
5.1.3. <i>Browser Library</i> .....	12
5.1.4. <i>Validation/DataImport Library</i> .....	12
5.2. DBSNP-LITE.....	12
5.3. PERL SCRIPTS .....	12
5.4. TESTING.....	12
<b>6. GETTING DATA IN</b> .....	<b>13</b>
6.1. MARKER IMPORT PIPELINE (DBSNP-LITE).....	13
6.2. STUDY METADATA .....	13
6.3. FREQUENCY AND ASSOCIATION DATA PIPELINE.....	13
6.4. PROCESSING.....	13
<b>7. GETTING DATA OUT</b> .....	<b>14</b>
7.1. THE WEB APPLICATION .....	14
7.2. HGVMART.....	14
7.3. BULK DATA EXPORT.....	15

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security: PU</b>
	<b>Author(s): R. Free (ULEIC)</b>		<b>Version: v1.01 – Final</b> 3/23

7.4. DAS.....	15
<b>8. USING HGVBASEG2P.....</b>	<b>15</b>
8.1. THE HOME PAGE.....	15
8.2. SEARCHING HGVBASEG2P .....	15
8.3. REPORTS .....	16
8.4. THE HGVBASEG2P BROWSER.....	17
8.4.1. <i>Study Information</i> .....	17
8.4.2. <i>Genome View</i> .....	17
8.4.3. <i>Region View</i> .....	17
8.4.4. <i>Marker Info</i> .....	19
8.5. HGVMART.....	19
8.5.1. <i>MartView</i> .....	19
8.5.2. <i>BioMart API and Web Services</i> .....	19
8.6. BULK DATA EXPORT.....	19
8.7. DAS.....	19
<b>ANNEXES .....</b>	<b>22</b>
ANNEX I – CORE DATABASE DATA MODELS .....	22
<b>REFERENCES.....</b>	<b>23</b>

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

## Document Information

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


<b>Deliverable</b>	<b>Number</b>	5.1	<b>Title</b>	Summary Document for Genomics Database V-1 Software
<b>Work package</b>	<b>Number</b>	5	<b>Title</b>	Genomics G2P Databases

<b>Delivery date</b>	<b>Contractual</b>	Month 18	<b>Actual</b>	03/08/2009
<b>Status</b>	Version 1.01		final <input checked="" type="checkbox"/>	
<b>Nature</b>	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
<b>Dissemination Level</b>	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

<b>Authors (Partner)</b>	Robert Free (ULEIC)			
<b>Responsible Author</b>	Robert Free		<b>Email</b>	rcf8@le.ac.uk
	<b>Partner</b>	ULEIC	<b>Phone</b>	0116 2237273

## Document History

Name	Date	Version	Description
R. Free	10/07/09	1.00	Initial draft
R. Free	03/08/09	1.01	Amended draft

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – 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	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – 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 (HGVBbaseG2P).

HGVBbaseG2P is a modular system based on a combination of custom-built components and third-party libraries and tools. There is also the provision to deactivate data sensitive components, in response to the Homer *et al.* study (Homer et al. 2008).

HGVBbaseG2P comprises relational databases containing core data used by the web application and secondary databases derived from these that provide data for the Browser and HGVMart. The core databases consist of a Marker database (providing a basal-layer of markers), which is linked to study-based genotyping and association data contained in a Study database. This latter database also contains study metadata including information about samples and phenotypes.

The software components of HGVBbaseG2P revolve around a Perl application programming interface (API), which: i) provides access to HGVBbaseG2P databases; ii) contains the framework for the web application and browser, and iii) provides logic for marker, data and metadata import routines.


To import data into HGVBbaseG2P it is necessary to have it in a compatible format. Our metadata import system uses an intermediate database compatible extensible markup language (XML) format. While, our frequency/association import pipeline uses a more flexible and customisable text-file based solution to cater for the various data formats.

The dynamic, integrated web front-end for HGVBbaseG2P is provided by a combination of third-party and custom built components, on both the server and client. These components were optimised to work together effectively as required.

The web application's home page provides access to all major functions including reports related to studies, panels, phenotypes and markers; the advanced search; and faceted browsing, from which they can be added to the HGVBbaseG2P Browser.

The Browser provides advanced visualisation of multiple association result sets, allowing them to be compared and contrasted at a genome-wide summary and a region specific level. This latter view contains multiple resolutions allowing the user to get to the individual marker level.

HGVBbaseG2P contains several data mining components including HGVMart (which can be accessed through the web application, and programmatically), bulk data downloads for individual experiments and a distributed annotation server (DAS).

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – 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 biobanking have dramatically increased the scale and the accuracy of the data being produced (Kingsmore 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. The situation is further compounded by the recent emergence of large genome-wide association studies (GWAS) data sets, which involve the study of many thousands of subjects. These enormous studies bring extra complications with respect to data handling, data sensitivity and statistical interpretation.


### 2.2. CURRENT SOLUTIONS

Public archival databases of genetic association data such as the database of Genotypes and Phenotypes (dbGaP) (Mailman et al.) and the European Genome-phenome Archive (EGA) (<http://www.ebi.ac.uk/ega/page.php>) provide an obvious potential solution to this problem and there are also some small disease-specific initiatives (N. C. Allen et al. 2008; Bertram et al. 2007; Hulbert et al. 2007). However, none of these bring together a globally comprehensive list of GWAS studies, while enabling direct submission of smaller studies. Providing genomics database software to fill this role was a key software deliverable in the GEN2PHEN project. The result of this work is the Human Genome Variation Genotype to Phenotype database (HGVBbaseG2P) (<http://hgvbbaseg2p.org>).

### 2.3. HGVBbaseG2P V1

HGVBbaseG2P V1 represents the world's first central database for summary-level (*i.e.* not person-specific) genetic association data concerning any and all traits. It has been populated with an initial series of GWAS data sets acquired from other public depositories (e.g. dbGaP, EGA); and subsets of results from multiple association studies. To enable these data sets to be effectively integrated and optimally used, HGVBbaseG2P provides extensive web-based tools for result set browsing, visualization and mining.

This document contains a summary of the system design and a brief guide to using the web application in practice. The summary describes database content and structure; the software used and developed; and approaches made available to get data in and out of the system. The guide includes illustrated examples of the web page reports; the visualisation tools available to researchers; and the data mining tools.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security: PU</b>
	<b>Author(s): R. Free (ULEIC)</b>		<b>Version: v1.01 – Final</b>

### 3. SYSTEM DESIGN

HGVbaseG2P V1 was designed as a modular system, which could easily be maintained, extended and tested (Figure 1). It is based on a combination of custom-built components and appropriate third-party libraries/tools (Table 1).

In response to data anonymity issues raised by a recent study (Homer et al. 2008), we can deactivate components that provide access to ‘sensitive’ data (e.g. p-values and genotype frequencies for a large number of markers) at the request of the data providers. This includes access to individual data sets within the Browser component and to the associated DAS tracks. Currently, the whole HGVmart must be deactivated, but in the future we intend to include fine grained controls for individual data sets.

HGVbaseG2P contains components that deal with three main themes: i) getting data in; ii) processing the data; iii) getting data out. Good database design was of critical importance in order to achieve synergy between these three themes.

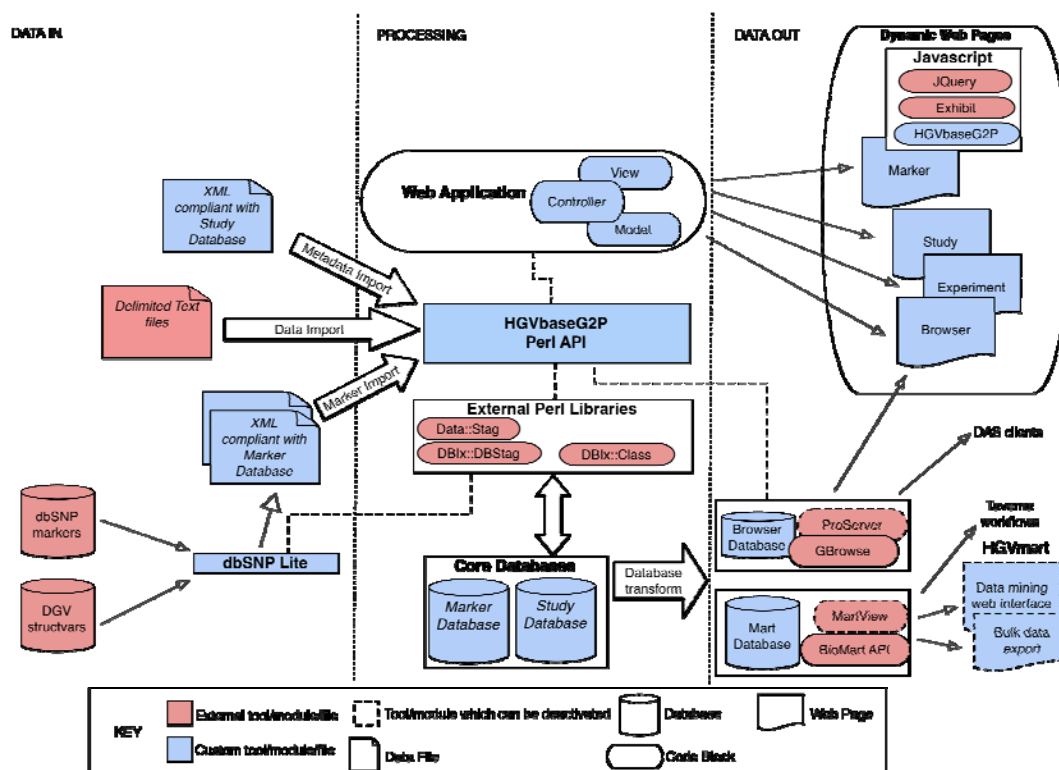



Figure 1: Summary of the design of HGVbaseG2P V1

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final


**Table 1: Third-party libraries and tools used in HGVbaseG2P**

Library/Tool	Source	Use
Catalyst 5.8	<a href="http://www.catalystframework.org/">http://www.catalystframework.org/</a>	Basis of the web-application and frontend
DBIx::Class 0.08102	<a href="http://search.cpan.org/dist/DBIx-Class/lib/DBIx/Class.pm">http://search.cpan.org/dist/DBIx-Class/lib/DBIx/Class.pm</a>	Object Relational Mapping for primary databases via HGVbaseG2P API
DBIx::DBStag 0.09	<a href="http://search.cpan.org/~cmungall/DBIx-DBStag/">http://search.cpan.org/~cmungall/DBIx-DBStag/</a>	Used by dbSNP-lite to provide primary database access via a structured data format.
Xapian 1.0.6	<a href="http://xapian.org/">http://xapian.org/</a>	Boolean searches of data in the core databases.
Template Toolkit 2.19	<a href="http://template-toolkit.org/">http://template-toolkit.org/</a>	Templating of hyper-text markup language (HTML) with complex logic
GBrowse 1.69 <sup>(Stein 2002)</sup>	<a href="http://gmod.org/">http://gmod.org/</a>	Basis of the Browser component
GBrowse_karyotype		
BioMart 0.6 <sup>(Kasprzyk 2003)</sup>	<a href="http://www.biomart.org/">http://www.biomart.org/</a>	Basis of the HGVmart component
JQuery 1.2.6	<a href="http://jquery.com/">http://jquery.com/</a>	Dynamic web pages and cross web-browser compatibility
JQuery UI 1.5.3	<a href="http://jqueryui.com/">http://jqueryui.com/</a>	
Exhibit <sup>(Huynh et al. 2007)</sup>	<a href="http://www.simile-widgets.org/exhibit/">http://www.simile-widgets.org/exhibit/</a>	Faceted searching of core data
MySQL 5.0.32	<a href="http://www.mysql.com/">http://www.mysql.com/</a>	Database system for back-end
Apache 2.2	<a href="http://httpd.apache.org/">http://httpd.apache.org/</a>	Web application server
ProServer 2.0 <sup>(Finn et al. 2007)</sup>	<a href="http://www.sanger.ac.uk/Software/analysis/proserver/">http://www.sanger.ac.uk/Software/analysis/proserver/</a>	DAS server for core and Browser data

#### 4. DATABASE DESIGN AND CONTENT

There are several databases that are required by HGVbaseG2P (Table 2).

The primary databases are complete relational databases containing all core data (Study and Marker), while the secondary databases are generated from the primary databases. These derived databases provide denormalised forms of data specific to the Browser and HGVmart components of the system.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

**Table 2: Databases used in HGVBbaseG2P and their roles in the system**

Database	Use	Role
Study	Core system	Study metadata, and genotyping and association data related to those studies.
Marker	Core system	Marker data, including a summary of changes to the original source databases and possible alleles and genotypes.
Browser	Browser	Denormalised data from the Study and Marker databases, and pre-calculated data based on counts of significant markers. The Browser uses this data to provide a rapid response to user requests.
Features	Browser	Gene feature file (GFF) compatible database containing basic genomic features
Mart	HGVmart	Denormalised form of Study and Marker databases in BioMart compatible format

#### 4.1. Marker Database

The Marker database (see Annex I for a diagram) comprises core information on all the markers present in NCBI's database of Single Nucleotide Polymorphisms (dbSNP) database (Mailman et al.).

Changes to marker information in the source database are tracked and updated in HGVBbaseG2P via the 'dbSNP-lite' marker import pipeline (see below). The Marker database also stores the location(s) of the markers and the flanks. Frequency and association data in the Study database are linked to the basal-layer of markers in the Marker database.


#### 4.2. Study Database

The Study database (see Annex I for a diagram) closely follows that of the new Phenotype and Genotype Experiment (PaGE) object model standard (<http://www.pageom.org/>).

The main association data layer of HGVBbaseG2P comprises four principal components that emulate the main concepts used in standard literature reporting of genetic association studies, namely 'Study', 'Sample', 'Phenotype' and 'Experiment' entities.

##### 4.2.1. Studies

A 'Study' wraps the three other main data entities that make up a single submission ('Sample', 'Phenotype' and 'Experiment'). Each 'Study' contains summary information, plus various details relating to the study design.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

#### 4.2.2. Samples

HGVbaseG2P uses ‘Sample Panels’ to represent a named collection of individuals that are present in a ‘Study’—such as disease cases, or matched controls. Typically, individuals in a ‘Sample Panel’ are affected by one or more similar disease phenotypes, or have some other key metric in common (*e.g.* age, gender, ethnicity). Data generated by performing genotyping experiments are reported in terms of an ‘Assayed Panel’. This is a group of test subjects derived by splitting and/or merging one or more sample panels to create new collections, on the basis of some explicit criteria such as severity or subclass of disease or some environmental criteria.

#### 4.2.3. Phenotypes

Phenotypes are stored in a very flexible but straightforward data structure. Whereas other databases typically use unstructured free-text descriptions to hold phenotype information, HGVbaseG2P splits phenotype information into three sub-components: (i) the ‘Phenotype Property’ (the character or trait investigated); (ii) the ‘Phenotype Method’ (how the trait was measured); and (iii) the ‘Phenotype Value’ (the value obtained by measuring the trait). Identical ordinal or nominal values in groups of individuals are thereby easily represented, as are categories of disease affection status. For quantitative traits in patient groups, statistical values that describe a distribution (*e.g.* median, standard deviation, maximum, minimum) are stored as a series of ‘Phenotype Values’. The same data model allows phenotype thresholds to be specified and used as criteria for ‘Assayed Panel’ selection.

#### 4.2.4. Experiments

A Study can contain one or more ‘Experiments’. These are packages within studies that contain one or both of the following elements:

- A genotyping data set pertaining to a single phenotype examined using a specific set of ‘Assayed Panels’ (*e.g.* Case and Control)
- One or more Result Sets containing association data and obtained using distinct statistical tests (*e.g.* allelic trend test, a genotypic test *etc.*)


## 5. SOFTWARE COMPONENTS

### 5.1. The HGVbaseG2P Application Program Interface (API)

In addition to establishing a large depository of summary-level association data, one of the goals of HGVbaseG2P is to provide a powerful toolkit to allow the extensive database content to be explored, so that new knowledge may be created. Most of the logic behind these tools is contained in separate libraries within a custom-built Perl API.

#### 5.1.1. Database/Schema Libraries

The Database modules provide controlled programmatic access to the primary databases, by wrapping common tasks performed by the Schema modules. Conversely, the Schema modules

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

are object relational mapping (ORM) modules, which provide complete access to all of the tables in the primary databases.

### **5.1.2. Web Library**

This library contains Catalyst-based modules that deal with page serving and session management. The HTML templates themselves are contained in a separate directory. Any complex logic (e.g. Browser) is separated into its own module or library.

### **5.1.3. Browser Library**

This library contains modules that generate data in a GBrowse compatible form. This is achieved by retrieving data from the Browser database, and then either passing it on and/or processing this data by extrapolating it into ‘bins’ of a particular size, based on its location on the chromosome.

### **5.1.4. Validation/DataImport Library**

This library provides modules that deal with the import and validation of study metadata, and frequency/association data.

## **5.2. dbSNP-lite**

This is a separate Perl system, which deals with the logic behind marker and allele content alterations, mergers and deletions. It accepts XML files downloaded from dbSNP, and then validates/processes them (taking into account the current data in HGVbaseG2P). The result is an XML file that can be loaded into the primary databases using DBStag.


## **5.3. Perl scripts**

These wrap the logic in the HGVbaseG2P modules and libraries and allow the user to do a number of things including:

- generate secondary databases from primary databases
- import data from markers, studies, frequencies and associations
- update pre-calculated data
- startup a test web-server
- generate search indices for Xpian

## **5.4. Testing**

The HGVbaseG2P system was tested using a combination of regression testing and manual testing. Regression testing is used to check complex logic via Perl test scripts and modules. Systematic manual testing was used to ensure the web application and its user interface work as expected.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

## 6. GETTING DATA IN

Data is imported into HGVbaseG2P using specially written import pipelines. These deal with marker data, study metadata and frequency/association data. Each of the approaches for achieving this is described in more detail below.

We are adding data from available GWAS studies (and from smaller data sets collating many studies) but in many cases we are unable to access data due to the previously mentioned Homer *et al.* study. At the request of some data providers we have also restricted access to data already present in the system.

### 6.1. Marker import pipeline (dbSNP-lite)

This import pipeline appropriately handles marker and allele stand-flips, mergers and deletions in a way that ensures the integrity of any existing connections to frequency or association record items. The pipeline generates XML in a format that can be loaded into HGVbaseG2P using DBStag.

### 6.2. Study Metadata


Metadata is currently loaded into HGVbaseG2P using DBStag. For single studies containing more details, a specifically formatted XML file is manually populated. In the case of composite data sources (ie. files containing data from multiple studies), it is easier to generate multiple XML files automatically for each study. In both cases, minor curation of the data is undertaken.

### 6.3. Frequency and Association Data Pipeline

The data import pipeline is a flexible system devised to deal with the many different file formats used in association studies. Data elements from text files are processed/validated in batches, and then stored in the database. The pipeline employs simple text-file templates containing the minimal information required to tie data to the Study metadata; and the arrangement of data fields in the input files. For each marker, identifiers (generally rs IDs) are validated against expected alleles (strand-flipped if required). Then where required, the system calculates values for missing data (e.g. genotype frequencies if only numbers are supplied). Finally correctly validated data is imported into the Study database.

### 6.4. Processing

The processing stage of HGVbaseG2P V1 involves the execution of data import pipelines to import and update marker, study metadata and frequency/association data files. Once the data in the primary databases is stable on the development server (ie. a data freeze), the databases are rebuilt on the live server alongside the previous data freeze. During this time the primary databases are also transformed into denormalised secondary databases, used by the Browser and HGVmart components. The process is completed once all databases have been built. At this point the HGVbaseG2P configuration is changed so that the system makes use of these new databases. The data can then be examined using the various outputs available to the user. Attempts are made to reduce downtime by simply switching the system to use the new databases.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

## 7. GETTING DATA OUT

Several methods are used to retrieve data from HGVbaseG2P. The web application provides pages containing comprehensive reports of study, genotype, association and marker data; access to tools for visual investigation (Browser); data mining (HGVMart); and help and information about the project. Data can also be retrieved directly using the BioMart API and associated web-services, and embedded as DAS tracks in Ensembl or other genome browsers. Archived bulk data for individual studies can also be downloaded from the web site.

### 7.1. The Web Application

HGVbaseG2P runs on a Linux based server running MySQL and Apache with Modperl (see Table 1 for versions). It uses a combination of client and server based approaches to provide a dynamic, integrated web-application for users.

On the server side, the Catalyst web application framework is used. This system is much easier to maintain and extend than other alternatives and can be scaled up into a complex web application.

Access to the primary databases (Study and Marker) is provided via ORM, and complex searches enabled by the Xapian library. All HTML pages are generated at run-time using the template-based TemplateToolkit system.

Client-side interactive elements are provided through custom written Javascript modules which make use of third-party libraries (Exhibit and JQuery/JQuery UI).


A modified version of GBrowse, was integrated into the web-application and provides interactive views for the HGVbaseG2P Browser component. Rather than retrieve data from primary databases directly, secondary databases were generated with simple data structure to improve the speed of the Browser component significantly.

To tie the disparate Catalyst, GBrowse and BioMart systems together, it was necessary to use a variety of clever client and server techniques including: use of asynchronous Javascript and XML (AJAX); parameter passing; auto resized IFrames; and showing/hiding web page elements. Changes were also made to some parts of GBrowse to allow direct retrieval of data, which further increased Browser performance.

### 7.2. HGVMart

The HGVMart data-mining tool was built using the BioMart system. HGVMart uses a BioMart compatible, transformed version of the primary databases, which were generated using bespoke scripts.

HGVMart can be accessed via the web application (through the MartView interface), or externally through web-services and the BioMart API.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

### 7.3. Bulk data export

Bulk-data downloads are available containing all data for single Experiments. These were generated from the HGVMart database using the BioMart API.

### 7.4. DAS

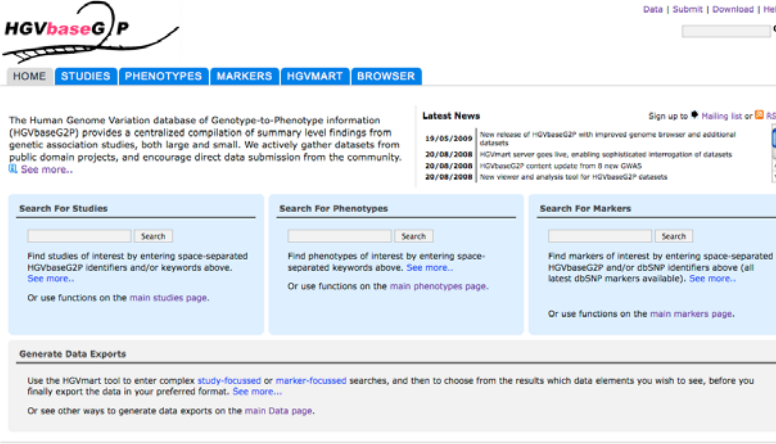
DAS support is provided through the external ProServer system<sup>(Finn et al. 2007)</sup>. This uses the Browser database to provide DAS tracks for each individual Result Set in HGVbaseG2P. These tracks are currently provided in the form of individual marker p-values.

## 8. USING HGVBASEG2P

### 8.1. The home page

The home page (<http://hgvbasesg2p.org/index>) for the web application (Figure 2) provides access to all useful pages.


The main tabs provide direct access to i) browsable lists for the studies and phenotypes; ii) a single/multiple marker search; iii) the HGVMart mining tool and iv) an advanced HGVbaseG2P browser. Links at the top and bottom provide access to general information pages (including a disclaimer, about and help pages).



**Figure 2: The HGVbaseG2P home page.**

### 8.2. Searching HGVbaseG2P

The top search box searches all domains for keywords (using OR logic), while keyword searches within specific domains can be carried out using the three distinct search boxes. Searches across all domains result in multiple tabs containing separate results for each of the three domains (studies, phenotypes and markers).

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

HGVbaseG2P V1 also provides a faceted search for studies and phenotypes, which provide comparable results to the standard search (Figure 3).

» Browse all studies  
The following genotype-to-phenotype Studies are currently available in the database. [What is a Study?](#) [Show Studies added to Browser](#)

No Studies added to Browser  
No Result Sets added to Browser  
[Go to Browser](#) [Remove all](#)

**Study List Filters**  
Use the filters below to refine the Studies displayed in the list on the right.

**Search Studies by keyword(s)**  
cancer  
e.g. disease name, year, genotyping platform, HGVbaseG2P Study identifier

**Select Study design(s)**  
22 GWAS

**22 Studies filtered from 119 originally (Reset All Filters)**

Study ID	Study Name	Study Design	Authors	Date	Add
HGVST101	Colorectal cancer	GWAS	Tomlinson <i>IP et al.</i>	2009-05-06	+
HGVST103	Prostate cancer	GWAS	Gudmundsson <i>J et al.</i>	2009-05-06	+
HGVST110	Colorectal cancer	GWAS	Tenesa <i>A et al.</i>	2009-05-06	+
HGVST114	Breast cancer	GWAS	Kibriya <i>MG et al.</i>	2009-05-06	+
HGVST118	Lung cancer	GWAS	Wang <i>Y et al.</i>	2009-05-06	+
HGVST126	Lung cancer	GWAS	Hung <i>RJ et al.</i>	2009-05-06	+
HGVST129	Breast cancer	GWAS	Gold <i>B et al.</i>	2009-05-06	+
HGVST130	Lung cancer	GWAS	Amos <i>CI et al.</i>	2009-05-06	+
HGVST136	Prostate cancer	GWAS	Eeles <i>RA et al.</i>	2009-05-06	+
HGVST32	Cancer, breast cancer and prostate cancer	GWAS	Murabito <i>JM et al.</i>	2009-05-06	+
HGVST42	Colorectal cancer	GWAS	Broderick <i>P et al.</i>	2009-05-06	+
HGVST57	Breast cancer	GWAS	Stacey <i>SN et al.</i>	2009-05-06	+
HGVST62	Breast cancer	GWAS	Easton <i>DF et al.</i>	2009-05-06	+
HGVST63	Colorectal cancer	GWAS	Zanke <i>BW et al.</i>	2009-05-06	+
HGVST70	Prostate cancer	GWAS	Duggan <i>D et al.</i>	2009-05-06	+
HGVST72	Lung cancer	GWAS	Spinola <i>M et al.</i>	2009-05-06	+
HGVST73	Colorectal cancer	GWAS	Tomlinson <i>I et al.</i>	2009-05-06	+
HGVST74	Lung cancer, smokers with versus smokers without	GWAS	Spinola <i>M et al.</i>	2009-05-06	+
HGVST77	Prostate cancer	GWAS	Gudmundsson <i>J et al.</i>	2009-05-06	+
HGVST81	Prostate cancer, Type II Diabetes Mellitus	GWAS	Gudmundsson <i>J et al.</i>	2009-05-06	+
HGVST93	Prostate cancer (aggressive)	GWAS	Thomas <i>G et al.</i>	2009-05-06	+
HGVST98	Urinary bladder cancer	GWAS	Klemeny <i>LA et al.</i>	2009-05-06	+


Figure 3: Faceted browsing of Studies using ‘cancer’ as a keyword.

### 8.3. Reports

HGVbaseG2P provides a number of tabular reports that summarise studies and markers, and their associated data (Table 3).

Table 3: Web-based reports available in HGVbaseG2P

Report Name	Content	Where?
Study report	Displays a Study summary including associated panels and phenotypes, and Experiments. This report also allows the user to add Result Sets to display in the Browser.	Links from browseable study list
Marker report	Displays a summary of the selected marker. But also provides dynamic views which allow users to show studies containing a particular marker at different significance levels. There are also views of allele/genotype frequencies.	Links from marker search results or if a search presents a single results goes straight to the report.
Panel reports	Reports for individual Sample Panels and Assayed Panels.	Links from ‘Panels’ tab in study report
Phenotype reports	Reports for individual Phenotype Methods and Values.	Link from ‘Phenotypes’ tab in study report

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

## 8.4. The HGVbaseG2P Browser

For advanced visualisation of the data sets, a Browser is provided. This allows the user to view a graphical representation of the p-value data to compare and contrast multiple (up to 8) association result sets.

The Browser allows researchers to:

- i) Find interesting p-value signals in a genome-wide summary of the data and then drill down into the details.
- ii) Go straight to a region of interest by entering co-ordinates, a gene name or a marker identifier into the Search box of the region view.

Studies of interest are added or removed from the browser using icons in the faceted lists and in the study reports. Users are presented with a dialog containing the result sets available in the chosen study, which they can select as appropriate.

To describe how the Browser works in practice, GWA studies comparing controls with ‘Crohn’s disease’ patients are used as an example.

### 8.4.1. Study Information

This tab displays the Studies that have been added, along with a key to the colours used in the graphical displays. It is also possible to view information about Experiments, and add/remove Result Sets from the Browser.

### 8.4.2. Genome View


This tab (Figure 4) shows a genome-wide summary of association results based on the number of p-values above a significance threshold, in 3-Mb bin for each selected Result Set. Options are provided for the user to: i) set the chromosomes displayed; ii) change the orientation and size of chromosomes; iii) switch the scale type of the stacked plot; and iv) display marker coverage for the selected result sets.

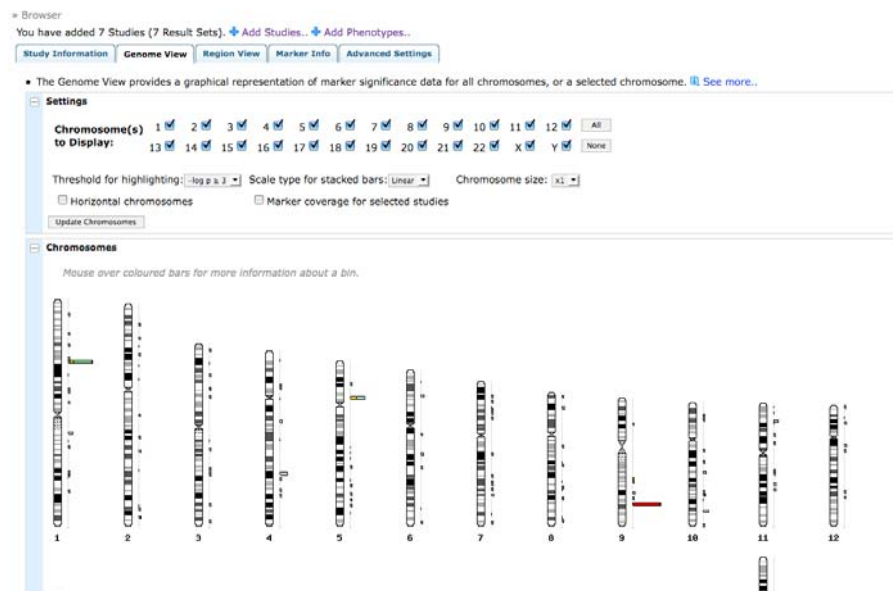
A popup is displayed if the user moves the mouse pointer over a bin of interest (Figure 5). This popup contains detailed information about the number of markers above the threshold, in each result set, for the highlighted 3-Mb bin.

Clicking on a bin switches the display to the Region View tab, which shows the region of interest in more detail.

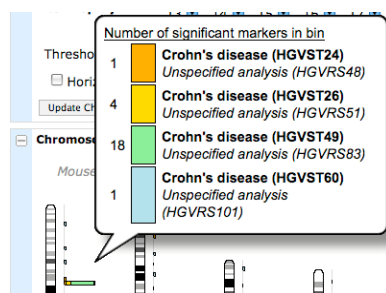
### 8.4.3. Region View

This view provides low and high-resolution views of data (depending on the size of the region being viewed). Low-resolution views are similar to the displays present in the ‘Genome View’, except the significant marker counts are calculated for 1-Mb windows. In addition a gene density plot is shown at this level of detail.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final



**Figure 4: The Browser page with Genome View showing Crohn's disease association studies.**




**Figure 5: Demonstration of a 'popup' showing the number of significant markers in a single 3Mb.**

The user can 'zoom in' to a region by using provided controls or clicking on a bin of interest.

The high-resolution view (Figure 6) shows the following information for the selected Result Sets:

1. The marker coverage
2. A line trace of the maximum  $-\log$  p-values (within 15-Kb bins)
3. A stacked plot of the number of markers above the defined significance (within 15Kb bins)
4. All individual markers above the significance threshold
5. Markers, which are significant in multiple result sets.

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

More details on the markers present in 4 and 5 can be obtained by clicking on any of interest to show the Marker Info tab.

#### **8.4.4. Marker Info**

The ‘Marker’ Info tab (Figure 7) shows general information about a selected marker and the significance data for a single result set or multiple result sets (including frequency and genotyping data). For anonymity reasons, sensitive data (*e.g.* odds ratios, frequency data) can be disabled for particular studies. This would generally be done upon request from the data producer.

## **8.5. HGVMart**

### **8.5.1. MartView**

The MartView interface is accessible through the ‘HGVMart’ tab. There are two mart data sets provided: G2P STUDY and MARKER. Once a data set is selected, the user can use filters to query the data as required, and attributes to select the data fields that are obtained.

Clicking the ‘Results’ button returns a user-selectable formatted version of the data (Figure 8). HGVMart has also been extended to provide compacted versions of frequency data (*e.g.* so all frequencies for a single ‘frequency cluster’ appear as a single row).

### **8.5.2. BioMart API and Web Services**

The BioMart Perl API can be used directly with HGVMart if the correct parameters are used in a BioMart registry file.

Another alternative for programmatic access is to use the web-service capabilities of HGVMart and the ‘query’ parameter to pass an XML query to the MartService system.

## **8.6. Bulk data export**

Bulk-data downloads for individual experiments are available from the study summary reports and from the ‘Download’ link. The data is provided as archived text files containing frequency and association data dumped from HGVMart.

## **8.7. DAS**

HGVbaseG2P data can be embedded into Ensembl and other DAS supporting genome browsers by using our DAS track URLs.

» Browser

You have added 4 Studies (4 Result Sets) [+ Add Studies..](#) [+ Add Phenotypes..](#)

[Study Information](#)
[Genome View](#)
[Region View](#)
[Marker Info](#)
[Advanced Settings](#)

- The Region View provides a graphical representation of one or more Result Sets' marker significance data for a genomic region or gene. [See more..](#)


**Settings**

Threshold for highlighting:  Scale type for stacked bars:

Showing 327 kbp from Chr1, positions 67,333,000 to 67,660,000



**Figure 6: The Browser page with Region View showing Crohn's disease association studies.**

 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

» Browser

You have added 4 Studies (4 Result Sets) [+ Add Studies..](#) [+ Add Phenotypes..](#)

[Study Information](#) | [Genome View](#) | [Region View](#) | [Marker Info](#) | [Advanced Settings](#)

• Showing information for HGVM3917208 in 3 Studies across 3 Result Sets

<b>HGVbaseG2P Marker ID</b>	HGVM3917208
<b>Source Database Info</b>	External marker accession <a href="#">rs11209026</a> , imported from <a href="#">dbSNP</a>
<b>Allele Information</b>	AGATCATTC(A):(G)AACTGGGTAG
<b>Genomic location</b>	Chr1: 65,995,725 - 65,995,725

Study Name and Analysis	Phenotype	P-value	Frequency Data	Odds Ratio		
				0.1	1	10
<b>Crohn's disease (HGVST26)</b> <i>Unspecified analysis (HGVR551)</i>	Crohn's disease	2.2e-18	Not supplied	Not supplied		
<b>Crohn's disease (HGVST49)</b> <i>Unspecified analysis (HGVR583)</i>	Crohn's disease	6.62e-19	Not supplied	Not supplied		
<b>Crohn's disease (HGVST60)</b> <i>Unspecified analysis (HGVR5101)</i>	Crohn's disease	2.17e-07	Not supplied	Not supplied		

**Figure 7: The Browser page with Marker Info showing a significant marker in three Crohn's disease association studies.**

» HGVMart (Study dataset selected)

[New](#) | [Count](#) | [Results](#)

[XML](#) | [Perl](#) | [Help](#) | [User Details](#) | [Logout](#)


Dataset: G2P STUDY  
Export all results to:  |   
 Unique results only   
Email notification to:

View:  rows as   Unique results only

HGVbase Study ID	HGVbase Marker ID	Allele Set	Chromosome	Marker Start	Marker Stop	Number Genotyped Samples
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	686
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	686
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	686
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	485
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	485
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	485
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	1102
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	1102
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	1102
HGVST1	HGVM1937142	(C)(T)	10	130394896	130394896	686

**Figure 8: HGVMart query of the 'G2P STUDY' dataset showing association data from the 'CGEMS Prostate Cancer' study.**



 HEALTH-200754	<b>D5.1 Summary Document for Genomics Database V-1 Software</b>		
	<b>WP5: Genomics G2P Databases</b>		<b>Security:</b> PU
	<b>Author(s):</b> R. Free (ULEIC)		<b>Version:</b> v1.01 – Final

## References

- Allen, N.C. et al., 2008. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genetics*, 40(7), 827-834.
- Bertram, L. et al., 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, 39(1), 17-23.
- Finn, R.D. et al., 2007. ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, 23(12), 1568-1570.
- Homer, N. et al., 2008. 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. et al., 2007. T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Research*, 35(Database), D742-D746.
- Huynh, D.F., Karger, D.R. & Miller, R.C., 2007. Exhibit. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*. Banff, Alberta, Canada, p. 737. Available at: <http://portal.acm.org/citation.cfm?doid=1242572.1242672>.
- Kasprzyk, A., 2003. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Research*, 14(1), 160-169.
- Kingsmore, S.F. et al., 2008. Genome-wide association studies: progress and potential for drug discovery and development. *Nature Reviews Drug Discovery*, 7(3), 221-230.
- Mailman, M.D. et al., The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10). Available at: <http://www.nature.com/doifinder/10.1038/ng1007-1181>.
- 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.
- Stein, L.D., 2002. The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research*, 12(10), 1599-1610.