BMC Research Notes

RESEARCH ARTICLE Open Access

Towards the integration of mouse databases - definition and implementation of solutions to two use-cases in mouse functional genomics

Michael Gruenberger[1], Rudi Alberts[2], Damian Smedley[3], Morris Swertz[4], Paul Schofield[1], The CASIMIR consortium[5], Klaus Schughart[2*]

Abstract

Background: The integration of information present in many disparate biological databases represents a major challenge in biomedical research. To define the problems and needs, and to explore strategies for database integration in mouse functional genomics, we consulted the biologist user community and implemented solutions to two user-defined use-cases.

Results: We organised workshops, meetings and used a questionnaire to identify the needs of biologist database users in mouse functional genomics. As a result, two use-cases were developed that can be used to drive future designs or extensions of mouse databases. Here, we present the use-cases and describe some initial computational solutions for them. The application for the gene-centric use-case, "MUSIG-Gen" starts from a list of gene names and collects a wide range of data types from several distributed databases in a "shopping cart"-like manner. The iterative user-driven approach is a response to strongly articulated requests from users, especially those without computational biology backgrounds. The application for the phenotype-centric use-case, "MUSIG-Phen", is based on a similar concept and starting from phenotype descriptions retrieves information for associated genes.

Conclusion: The use-cases created, and their prototype software implementations should help to better define biologists' needs for database integration and may serve as a starting point for future bioinformatics solutions aimed at end-user biologists.

Background

At present, we are just beginning to appreciate the complexity of genotype-phenotype association in humans, but more detailed and comprehensive analyses in basic research are urgently needed. Although studies in humans are important, they are limited because of the size of cohorts, strong but often unknown environmental influences, poor and inconsistently coded diagnosis, and lack of repeatability. Therefore, animal models are absolutely essential to complement human studies; they allow the investigation of underlying biological mechanisms in well-controlled experimental systems.

In particular, the mouse is an ideal model system for studying genetic factors that contribute to diseases because genetic reference populations (GRPs) with a large number of allelic variants in many genes, combinations thereof, and many knock-out mouse lines with deletions in single genes are available [1]. Research on mouse model systems has generated valuable discoveries for our understanding of the biological mechanisms of the normal function of the immune system as well as immune abnormalities, cardiovascular diseases, cancer, and infectious diseases [2].

Consequently, funding agencies around the world have supported an increasing number of functional genomics projects focused on the use of the laboratory mouse as a model for human disease. The results obtained have been collected in various databases. However, in most cases, these databases represent single project outputs and are maintained at different sites. Exceptions are, for example, the mouse genome database (MGD) database of MGI [3], the mouse phenome database (MPD) [4],

* Correspondence: kls@helmholtz-hzi.de
[2]Department of Infection Genetics, Helmholtz Centre for Infection Research & University of Veterinary Medicine Hannover, Inhoffenstr. 7, D-38124 Braunschweig, Germany

Europhenome [5] and the GeneNetwork database [6], which have collected information from many different sources. MGD is a database which has been optimized for researchers in the field of mouse functional genetics and genomics. It is constantly updated and manually curated and thus contains information of extremely high quality. Similarly, the GeneNetwork database contains phenotype and genotype information on mouse GRPs from the literature and directly entered source data, as well as tools to map quantitative trait loci. Both databases are extensively linked to other informatics resources.

However, there is a large volume of data in distributed databases that is not contained in MGI (Mouse Genome Informatics) or GeneNetwork and which are important for functional genomics studies (see the Mouse Resource Browser MRB [7]). Ad-hoc integration of these databases is very difficult. Many databases require a separate login procedure and need to be accessed using different methods (*e.g.* via a website, downloadable files or web services). Several resources do not adopt common standards *e.g.* using the same identifier for a given gene or protein [8]. In this case, a user may need to convert their gene identifiers to whatever the particular resource understands, e.g. MGI or Ensembl/mouse IDs, before starting a search.

As a first step towards new concepts for database integration, we have established a network of scientists from Europe, North America, Japan and Australia. The network is funded as a Coordination Action by the European Commission and called CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources) [9]. The Coordination Action is aimed at recommending standards to allow data sharing and integration between different projects.

Much can already be achieved using query tools that ease selection and joining of distributed data, such as BioMart [10], and/or workflow tools that support stepwise data retrieval, conversion and integration, such as Taverna [11] and Galaxy [12]. A prerequisite is that sources provide programmatic interfaces for queries or workflow tools that can be used to access or import the original data. However, such interfaces are often not available. This challenge was addressed by Smedley et al. who federated BioMart and MOLGENIS [13,14] in a Taverna workflow [15]. But these solutions are still too involved for many bench biologists to use directly for their research. Task-oriented user interfaces are needed on top of all these tools to more closely support biologists in their integrative analyses.

In order to gather the perspective of the end-users, the biologists, who will perform the actual data mining we designed use-cases together with biologists. Subsequently, two software implementations were developed on the basis of these use-cases to provide tools which could carry out the tasks requested by the users in the most practical format. Here, we describe two use-cases that arose as a result of our discussions with biologist-users during workshops, meetings and via a questionnaire. Furthermore, we demonstrate the first steps towards their implementation.

Methods and Results

Definition of the use-cases

During the first sessions with different user groups, some principle needs for data mining became apparent. These needs were further confirmed in subsequent meetings and demonstrations of development steps to biologist users. A user-friendly interface should not only query multiple databases but also allow for multiple search terms, allow iterative interactions, and contain a tool that allows storage of the results. Furthermore, most of the currently performed data mining in functional mouse genomics concerns genes, their functions and variants on one side; and phenotype descriptions on the other side. Based on these discussions, we designed two generic use-cases that should be suitable to a larger scientific community: a gene-centric and a phenotype-centric use-case.

Gene-centric use-case

The advent of high-throughput technologies in biology, such as gene expression microarrays, makes it now possible to identify, with the help of statistical and bioinformatics tools, large groups of candidate genes changing their expression levels in different experimental conditions. However, of the genes identified in this way, usually a few hundred, only a limited number of genes (in the order of 20-50) can feasibly be studied experimentally in the laboratory. Therefore, researchers prioritize the gene lists based on their own knowledge, literature, and additional information from many different web accessible databases, such as gene and protein descriptions, genetic diversity information, expression patterns in different tissues, *etc.* Since the searching of all these web databases by hand is very laborious and time-consuming, our user groups decided to describe a gene-centric use-case starting with an input of a limited number of gene names and aiming to facilitate easy and automatic collection of information about these genes from different sources. This process should be performed in an interactive fashion and allow storage and export of the results obtained.

An iterative user-driven strategy was developed based on the principles of an “online shop” (Fig. 1). Here, a customer can perform searches on the available data and collect them in a shopping cart. By performing additional searches for other data and by evaluating additional information on them, the customer can then decide to add or remove articles from his cart. Finally the collected articles are "exported" by executing an order.

Following the above principle, the integration of mouse databases via a gene-centric use-case should allow candidate gene symbols to be entered into a query form which then automatically collects basic information like synonyms, gene IDs, descriptions and genome locations for the entries (Fig. 1). Based on this information the user will then be able to refine the gene hit list by selecting the interesting genes and removing false hits. The final list will then be saved as a 'shopping cart' which can be revisited, modified, refined or extended. Finally, it should be possible to export the gene list in Excel-readable CSV format (Fig. 1).

A difficulty often encountered when performing analyses on genes, is that they have several synonyms and that in many scientific publications the systematic gene nomenclature is not followed (see [16]). Examples are *RANTES* (correct gene symbol *Ccl5*), *MIP1a* (*Ccl3*) and *IP-10* (*Cxcl10*). For other genes, it may be not known to the researcher that they represent members of large gene families, and one has to choose one or all to proceed with the analysis. Examples are *Hox*, *Fgf*, *Inhibin*, and interferon genes. Here, we consider as the "correct gene name" the name which is given by the international nomenclature committees: Mouse (International Committee on Standardized Genetic Nomenclature for Mice [17]), human (HUGO Gene Nomenclature Committee [18]), and rat (Rat Gene Nomenclature Committee [19]).

It is thus important that the use-case allows entering any gene name, synonyms, incomplete names, *etc*., but still makes sure that the correct genes will be found. For this, entries will be searched in a first step against the MGI database for disambiguation [20]. For each gene name multiple hits may appear and the user is then able to select the correct ones and add them to the cart.

In a second step, it is possible to collect additional information from different databases for the genes in the cart list. Examples of databases are MGI and ENSEMBL/mouse for information on gene structure and links to other resources; Eurexpress [21], SymAtlas [22] and ArrayExpress [23] for gene expression information; and INTACT [24] for gene interaction data. After retrieval of this information the user may refine his gene list in a given cart by searching for other genes or deleting genes in the current list.

The list of collected genes in a shopping cart can then be used to perform meta-analyses. For example, an analysis of GO-terms will allow finding out if certain GO-categories are over-represented in the particular gene list, indicating that the genes may belong to a specific pathway or biological process. Similarly, an analysis of expression patterns may reveal if there is a certain tissue in which the genes from the list are preferentially expressed.

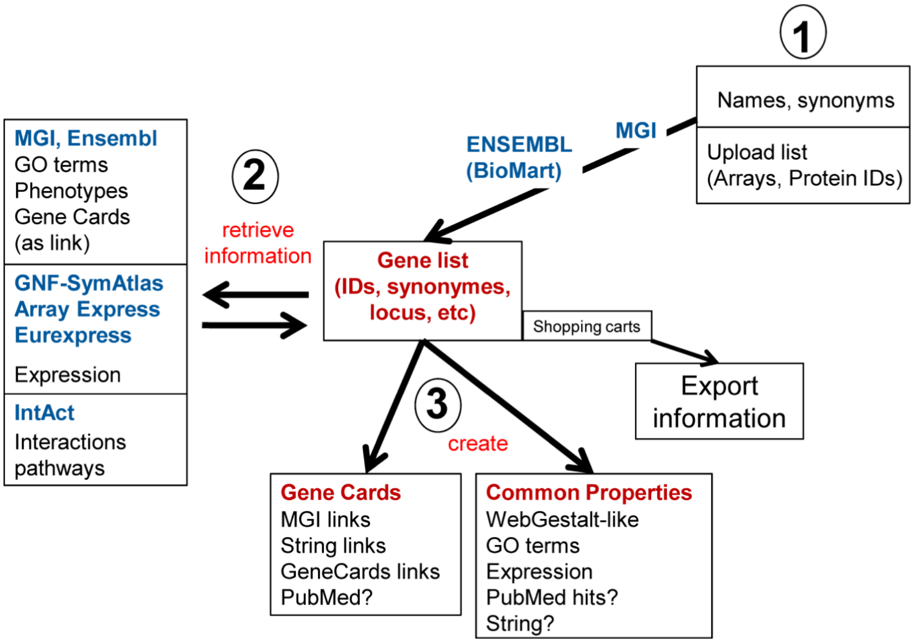


Figure 1 Schematic outline of gene-centric use case. The gene-centric use case should make it possible to enter gene names into a query form which then automatically collects basic information like synonyms, gene IDs, descriptions and genome locations for the entries. The user should then be able to select the interesting hits. The final list will then be saved as a 'shopping' cart which can be revisited and exported.

At present, only few of the currently existing databases offer some of the above-described functionalities, the most comprehensive one being MGI. And thus far, only BioMart represents an initiative which aims to allow the user to design queries on information from otherwise disparate databases. Also, BioMart allows refining searches and filtering out relevant information. However, Biomart is currently aimed at the advanced and trained user and is not yet designed for simple querying and collection of results in a shopping cart to which new genes and information can be added.

Phenotype-centric use-case

A second use-case was defined through the interaction with the user groups. It should allow researchers to begin their search with a phenotype description (Fig. 2). In this use-case, the scientist will search a phenotype ontology, obtain the closest hits and then decide which terms should be used in the following query. The use-case should also allow browsing of the phenotype ontology and the selection of terms of interest. The result of the searches for phenotype descriptions should then link to the associated genes.

At present, the most extensive and well structured phenotype ontology for the mouse is the Mammalian Phenotype (MP) ontology [25], accessible at MGI. MP is therefore used as a first standard which will allow querying MGI but also other databases that are using MP terms for phenotype descriptions, like EuroPhenome [26].

In the future, cross-referencing mouse MP terms with ontologies that describe diseases (such as the Disease Ontology - DO [27]) and phenotypes in humans (such as the Human Phenotype Ontology HPO [28] and Mouse Pathology Ontology MPATH [29]) should allow users to make cross-species searches by starting from phenotype descriptions. This will be particularly useful for human clinician researchers who are not familiar with mouse databases but who would like to know if there is a mouse model available for a given human disease.

The results from the phenotype-driven searches should then be linked to gene names associated with a given phenotype. These genes are presented as a list from which the user can choose the genes of interest and save them in a shopping cart. It is then possible to feed the genes into the gene-centric use-case and perform a more detailed data mining or meta-analysis.

The description and further development of the phenotype-driven use-case may represent a very useful concept for scientists and clinicians outside the mouse community. For example the Human Phenotype Ontology HPO is based on OMIM [30] and a search may be generated using HPO as a starting point to retrieve disease ID's from OMIM which can then be linked to gene symbols. The Drosophila phenotype ontology [31] developed by the Flybase group could be used to retrieve gene symbols and thereby gene function information from Flybase [32]. Or the *C. elegans* phenotype ontology [33] could be used to retrieve gene symbols from Wormbase [34]. Gene symbols retrieved from these databases could then be stored in a shopping cart.

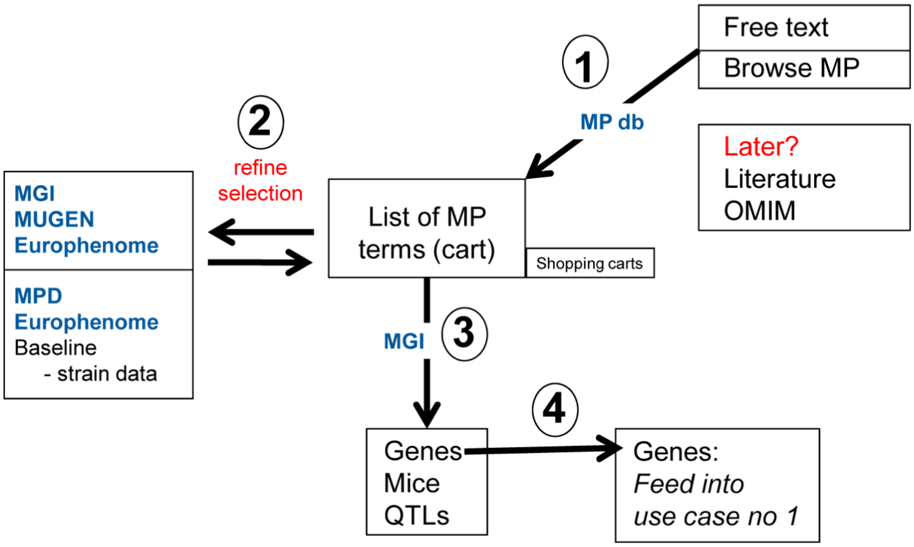


Figure 2 Schematic outline of phenotype-centric use case. The user will search or browse a phenotype ontology, obtain the closest hits and then decide which terms should be used for searches of phenotype descriptions. The latter should then link to genes associated with these phenotypes.

Implementation of the use-cases: MUSIG-Gen and MUSIG-Phen

Web services for database integration

A prerequisite for computer-supported data integration is programmatic access to select and retrieve data from distributed resources. As described by [15] there are several possible technical solutions to integrate data from different mouse informatics databases. The "CASIMIR strategy" is based on semantic standardization or wrapping of information transferred by web services. Currently the most popular implementations of web services use the SOAP/WSDL or the XML-REST protocols. The advantages of opening APIs and transferring information using XML schemas are discussed in [15].

For Europhenome and Mugen [35] SOAP/WSDL web services were available which could be used for MUSIG-Phen, and we set up a BioMart web service for part of the MGI data. Other databases such as the Ontology Lookup Service (OLS [36]) for ontology data and INTACT already had web services.

Users may want to integrate their local database or other databases. To demonstrate how this can be achieved, we generated web services for accessing GNF SymAtlas expression data. For this, we first saved the SymAtlas data locally. We then defined the Entrez Gene ID's as a common field which could be retrieved from the MGI Biomart and matched to the records in the local SymAtlas database. We then used MOLGENIS to create the relevant SOAP web services to retrieve the data from the local database, to subsequently load and display them in the shopping cart interface.

Implementation of MUSIG-Gen

After having defined the use-cases we wanted to provide users and developers with a first implementation which may then be tested and further revised in the future. Thus, certain parts of the use-case scheme outlined in Fig. 1 were implemented in the application MUSIG-Gen http://www.casimir.org.uk/usecase1/. In the following, we describe this tool from the perspective of the scientific user.

Fig. 3 displays the entry form of MUSIG-Gen where the user can type in gene names or synonyms (example: synonyms for chemokines). The result of the subsequent search query shows a list of hits from the MGI database which contain the query name (Fig. 4) and, in the default setting, additional information for each gene, like gene symbol, full gene name, all synonyms, and chromosomal location. This information allows the user to decide which one of the hits in the list corresponds to the gene of interest. As shown for the inputs "*RANTES*" and "*IP-10*", the correct gene names are displayed together with the search term and all other synonyms. If, for example, "*Fgf*" is used as query, all *Fgf* gene family members are displayed. The user may now decide which members to follow further. The genes selected in this process via the check box may then be saved in a shopping cart.

The gene list can subsequently be retrieved from the cart (Fig. 5) and additional information added, for example MGI IDs. These are hyper-linked to the corresponding entry at MGI so that the user has access to all MGI information on this particular gene with a single mouse click. Similarly, information on gene expression can be retrieved from the SymAtlas database. This query creates a new column for all genes on the list, displaying the SymAtlas IDs. The ID is again hyper-linked to SymAtlas and the corresponding data can be visualized with one mouse click (Fig. 6). Also, a search for information on Single Nucleotide Polymorphisms (SNPs) has been implemented. This function queries the Ensembl database and is currently set to display SNPs which result in non-synonymous coding changes in the open reading frame of the genes as well as the SNP Variation ID and a link to the Ensembl page with more details. (Fig. 6).

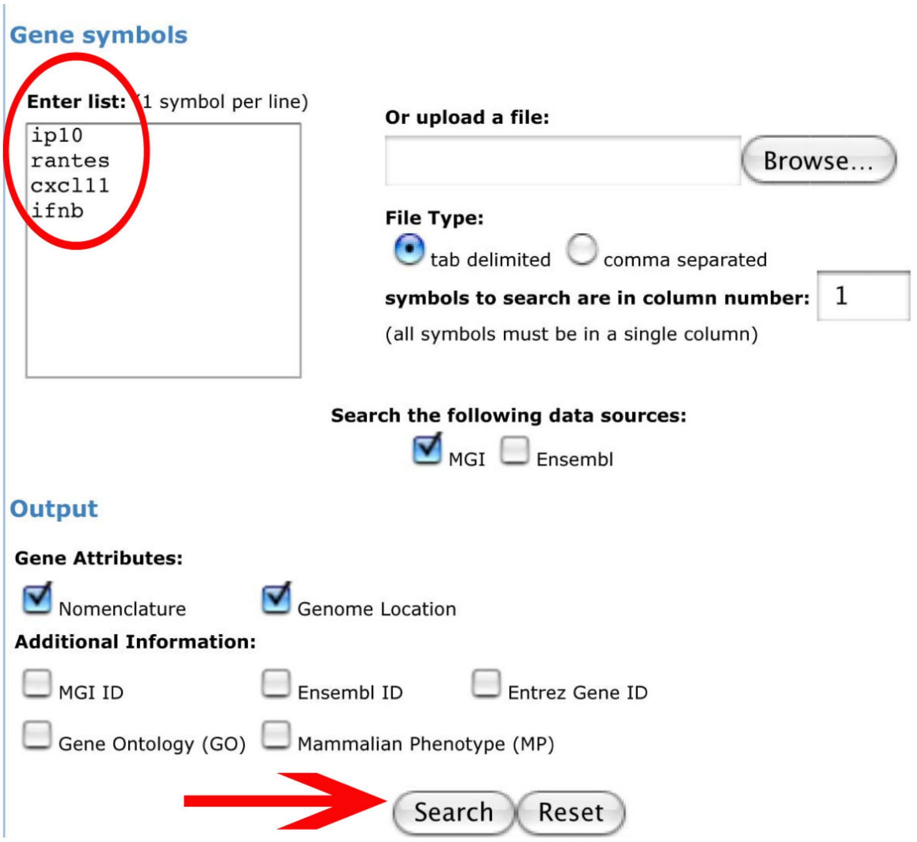
New genes can be easily added to an existing cart by calling up the entry form from within a cart and follow the same procedure as described above.

Because the genes listed in a cart contain a correct and unique identifier (MGI and/or Ensembl IDs) they can be directly used to query other databases. Such features and searches could be easily added to the existing MUSIG-Gen application. But even more important, it may now possible to perform an analysis on the entire group of genes in the cart. In the current version of the use-case, we implemented a GO term count as a proof-of-concept for the user interface. GO terms can be associated with all genes of the list using the 'load more data' feature and the representation of different GO-terms across the whole gene list be displayed (Fig. 7). These analyses may be extended to more sophisticated meta-analysis including also statistical evaluations in the future. Similarly, we added a tool to associate phenotype terms from the MP ontology and show their representation in the cart gene list.

As a final step, we added an export function to the shopping cart which allows the user to export his data in CSV format and then perform highly customized analysis locally.

Technical aspects of the implementation of MUSIG-Gen

The application layer of the shopping cart was developed in PHP. PHP proved to be a good choice for the development of the user interfaces, but did create some problems for the development of the web service client scripts because of a lack of multi-threading. The latter makes it impossible to retrieve data from different web services at the same time. The major problem is that some web pages access multiple services and depending



Gene symbols

Enter list: (1 symbol per line)

ip10
rantes
cxcl11
ifnb

Or upload a file:

File Type:

● tab delimited ○ comma separated

symbols to search are in column number: 1

(all symbols must be in a single column)

Search the following data sources:

☑ MGI □ Ensembl

Output

Gene Attributes:

☑ Nomenclature ☑ Genome Location

Additional Information:

□ MGI ID □ Ensembl ID □ Entrez Gene ID

□ Gene Ontology (GO) □ Mammalian Phenotype (MP)

Figure 3 Entry page for MUSIG-Gen. Here, the user can type in gene names or synonyms (circled entry box) and specify additional criteria under "output" which will be displayed on the hit list. Chemokine names and synonyms are displayed as an example. Clicking on the box "Search" (arrow) will start the search. A click on "Reset" will delete the gene names in the entry box and allow entering new names.

on the network speed and the kind of query some web services are slow to respond. This operation would thus stop the page from loading in the browser. We managed to mitigate this problem by creating an AJAX (Asynchronous JavaScript and XML) based loading system using the PHP PEAR AJAX [37] libraries. This system loads the main page first and then accesses each web service individually, thereby creating a more responsive system which lets the user interact with some data while the remainder of the data is still being retrieved.

The shopping cart system uses a Postgresql database to store user data. The data stored comprises the user's personal data (which is integrated into our web site management system to allow for a single login system) as well as the data retrieved from the different web services. The system imposes no limits as to how many data fields or data values a user can download and store in his shopping carts.

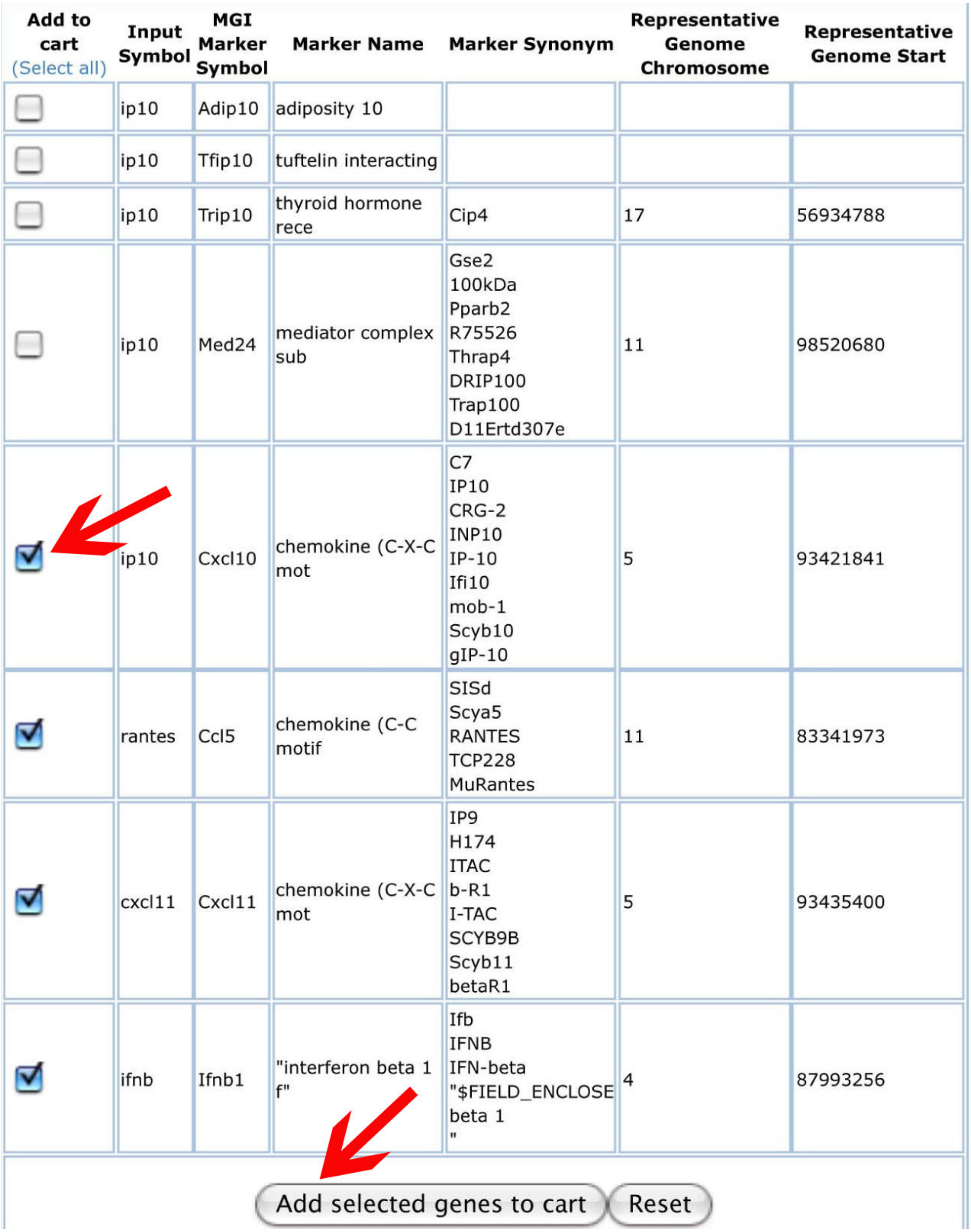
The application initially retrieves gene nomenclature and genome location data based on gene symbol: By default, nomenclature and genome location data is loaded from our MGI BioMart http://www.casimir.org.uk/biomart/martview/. Other data from the MGI BioMart can also be loaded, such as MGI, Ensembl, EntrezGene IDs as well as GO and MP ontologies. The Ensembl BioMart can also be queried at this stage for Uniprot IDs. Both BioMarts are accessed using the default BioMart XML-REST services. For this, we developed and used a generic BioMart XML-REST PHP client class which can be used to query any BioMarts.

Data may also be loaded from the Eurexpress BioMart or from the GNF and INTACT SOAP web services (using generic PHP SOAP libraries). There are also some fields which have the option of loading additional information, *e.g.* the GO and MP ID fields. The user can choose to load the ontology term names which are loaded from the OLS SOAP web service.

The source code and documentation for the MUSIG-Gen prototype may be downloaded form the following web server: http://www.casimir.org.uk/sourcecode/

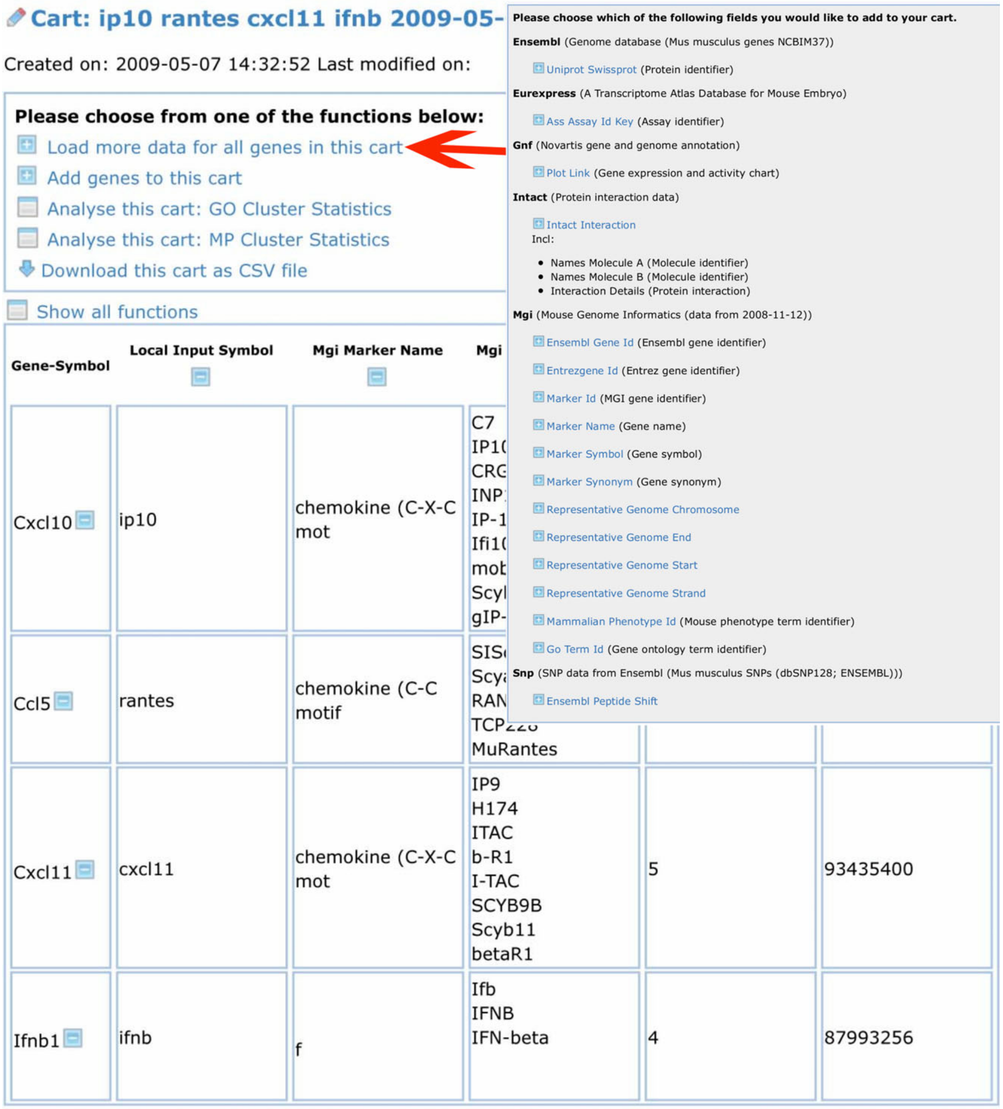
Implementation of MUSIG-Phen

Based on the scheme outlined in Fig. 7, certain parts of the phenotype-centric use-case were implemented in the application MUSIG-Phen http://www.casimir.org.uk/



Add to cart (Select all)	Input Symbol	MGI Marker Symbol	Marker Name	Marker Synonym	Representative Genome Chromosome	Representative Genome Start
□	ip10	Adip10	adiposity 10			
□	ip10	Tfip10	tuftelin interacting			
□	ip10	Trip10	thyroid hormone rece	Cip4	17	56934788
□	ip10	Med24	mediator complex sub	Gse2 100kDa Pparb2 R75526 Thrap4 DRIP100 Trap100 D11Ertd307e	11	98520680
☑	ip10	Cxcl10	chemokine (C-X-C mot	C7 IP10 CRG-2 INP10 IP-10 Ifi10 mob-1 Scyb10 gIP-10	5	93421841
☑	rantes	Ccl5	chemokine (C-C motif	SISd Scya5 RANTES TCP228 MuRantes	11	83341973
☑	cxcl11	Cxcl11	chemokine (C-X-C mot	IP9 H174 ITAC b-R1 I-TAC SCYB9B Scyb11 betaR1	5	93435400
☑	ifnb	Ifnb1	"interferon beta 1 f"	Ifb IFNB IFN-beta "$FIELD_ENCLOSE beta 1 "	4	87993256

Figure 4 First result page displaying list of all possible hits from MGI. The results of a search will be displayed as a list of hits from the MGI database which contain the query name and additional information for each gene. This information will allow the user select the hits which correspond to the gene of interest by clicking on the check box (arrow). A click on the button "Add selected genes to cart" will save the selected entries to a cart (arrow).



Cart: ip10 rantes cxcl11 ifnb 2009-05-

Created on: 2009-05-07 14:32:52 Last modified on:

Please choose from one of the functions below:

- Load more data for all genes in this cart
- Add genes to this cart
- Analyse this cart: GO Cluster Statistics
- Analyse this cart: MP Cluster Statistics
- Download this cart as CSV file

Show all functions

Gene-Symbol	Local Input Symbol	Mgi Marker Name	Mgi		
Cxcl10	ip10	chemokine (C-X-C mot	C7 IP1(CRG INP IP-1 Ifi1(mot Scyl gIP-		
Ccl5	rantes	chemokine (C-C motif	SIS(Scy RAN TCP228 MuRantes		
Cxcl11	cxcl11	chemokine (C-X-C mot	IP9 H174 ITAC b-R1 I-TAC SCYB9B Scyb11 betaR1	5	93435400
Ifnb1	ifnb	f	Ifb IFNB IFN-beta	4	87993256

Please choose which of the following fields you would like to add to your cart.

Ensembl (Genome database (Mus musculus genes NCBIM37))

- Uniprot Swissprot (Protein identifier)

Eurexpress (A Transcriptome Atlas Database for Mouse Embryo)

- Ass Assay Id Key (Assay identifier)

Gnf (Novartis gene and genome annotation)

- Plot Link (Gene expression and activity chart)

Intact (Protein interaction data)

- Intact Interaction
  Incl:
  - Names Molecule A (Molecule identifier)
  - Names Molecule B (Molecule identifier)
  - Interaction Details (Protein interaction)

Mgi (Mouse Genome Informatics (data from 2008-11-12))

- Ensembl Gene Id (Ensembl gene identifier)
- Entrezgene Id (Entrez gene identifier)
- Marker Id (MGI gene identifier)
- Marker Name (Gene name)
- Marker Symbol (Gene symbol)
- Marker Synonym (Gene synonym)
- Representative Genome Chromosome
- Representative Genome End
- Representative Genome Start
- Representative Genome Strand
- Mammalian Phenotype Id (Mouse phenotype term identifier)
- Go Term Id (Gene ontology term identifier)

Snp (SNP data from Ensembl (Mus musculus SNPs (dbSNP128; ENSEMBL)))

- Ensembl Peptide Shift

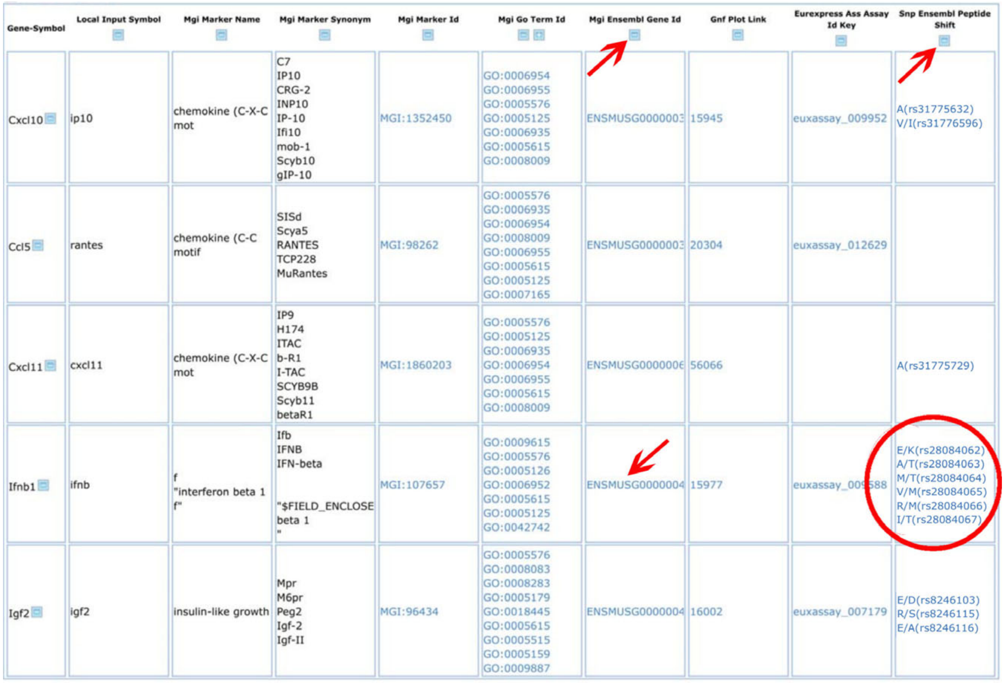
Figure 5 Retrieving list of genes from cart and adding more information. A gene list saved as a cart may be retrieved and additional information added by choosing the option "Load more data for all genes in this cart" (arrow). This will open a new window (insert) in which more information can be loaded from various databases.

usecase2/. The MUSIG-Phen prototype starts from a phenotype description, collects the genes associated with this phenotype in a cart and then performs all the analyses described above for MUSIG-Gen.

The starting point of MUSIG-Phen is a search page in which a free text entry will display a list of MP terms that most closely resemble the search term. The user may now choose the appropriate term, send a query to MGI and retrieve a list of genes that are associated with it. The list of genes can then be saved in a cart and further analyzed as described for MUSIG-Gen, *e.g.* add more information, perform meta-analysis, export lists. Alternatively, the user may start his query by browsing the hierarchical list of MP terms, select one and then retrieve the genes associated to the MP term (Fig. 8).

At this stage, the implementation is very similar to the services already provided by MGI. Thus, in addition to the current MGI search options, we implemented the possibility to query other external databases which contain phenotype descriptions based on MP terms. We demonstrated feasibility of this feature for searches of the Mugen and Europhenome databases.

At the present state, the MUSIG-Phen software was not designed for more sophisticated queries, because discussions with users revealed that further detailed queries very soon become highly specialized and



Gene-Symbol	Local Input Symbol	Mgi Marker Name	Mgi Marker Synonym	Mgi Marker Id	Mgi Go Term Id	Mgi Ensembl Gene Id	Gnf Plot Link	Eurexpress Ass Assay Id Key	Snp Ensembl Peptide Shift
Cxcl10	ip10	chemokine (C-X-C mot	C7 IP10 CRG-2 INP10 IP-10 Ifi10 mob-1 Scyb10 gIP-10	MGI:1352450	GO:0006954 GO:0006955 GO:0005576 GO:0005125 GO:0006935 GO:0005615 GO:0008009	ENSMUSG0000003	15945	euxassay_009952	A(rs31775632) V/I(rs31776596)
Ccl5	rantes	chemokine (C-C motif	SISd Scya5 RANTES TCP228 MuRantes	MGI:98262	GO:0005576 GO:0006935 GO:0006954 GO:0008009 GO:0006955 GO:0005615 GO:0005125 GO:0007165	ENSMUSG0000003	20304	euxassay_012629	
Cxcl11	cxcl11	chemokine (C-X-C mot	IP9 H174 ITAC b-R1 I-TAC SCYB9B Scyb11 betaR1	MGI:1860203	GO:0005576 GO:0005125 GO:0006935 GO:0006954 GO:0006955 GO:0005615 GO:0008009	ENSMUSG0000006	56066		A(rs31775729)
Ifnb1	ifnb	f "interferon beta 1 f"	Ifb IFNB IFN-beta "$FIELD_ENCLOSE beta 1 "	MGI:107657	GO:0009615 GO:0005576 GO:0005126 GO:0006952 GO:0005615 GO:0005125 GO:0042742	ENSMUSG0000004	15977	euxassay_009588	E/K(rs28084062) A/T(rs28084063) M/T(rs28084064) V/M(rs28084065) R/M(rs28084066) I/T(rs28084067)
Igf2	igf2	insulin-like growth	Mpr M6pr Peg2 Igf-2 Igf-II	MGI:96434	GO:0005576 GO:0008083 GO:0008283 GO:0005179 GO:0018445 GO:0005615 GO:0005515 GO:0005159 GO:0009887	ENSMUSG0000004	16002	euxassay_007179	E/D(rs8246103) R/S(rs8246115) E/A(rs8246116)

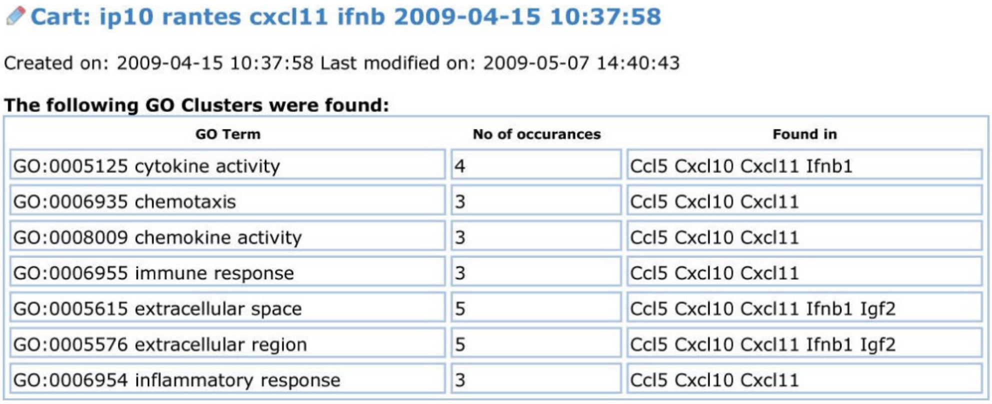
Figure 6 Extended result list. Result list displaying information from various databases. The MGI (column 5) and Ensembl IDs (arrows) are hyper-linked to the MGI and Ensembl databases, respectively. For the identification of non-synonymous SNPs, a query to the Ensembl database will create a new column (arrow column 10) in which predicted amino acid changes are displayed (circle).

complex for certain user subgroups. However, the present use-case implementation may already serve to query nascent databases (*e.g.* phenotype data from EUMODIC) and represents a very useful platform to test new developments which aim to connect mouse and human phenotype databases.

Technical aspects of the implementation of MUSIG-Phen

The implementation of the phenotype-centric use-case uses three SOAP/WSDL web services and our MGI BioMart web service: Initially the Mammalian Phenotype (MP) ontology is loaded from the OLS web service. The user-selected MP term is sent as query input to the MGI, EuroPhenome and MUGEN web services and matching gene symbols are returned. Gene symbols can then be selected and sent to the gene-centric use-case shopping cart.

Basic information about web services, such as type (for example BioMart or SOAP) and location URL is currently stored in a separate table. However, a larger web service catalogue such as BioMoby [38], Biocatalogue



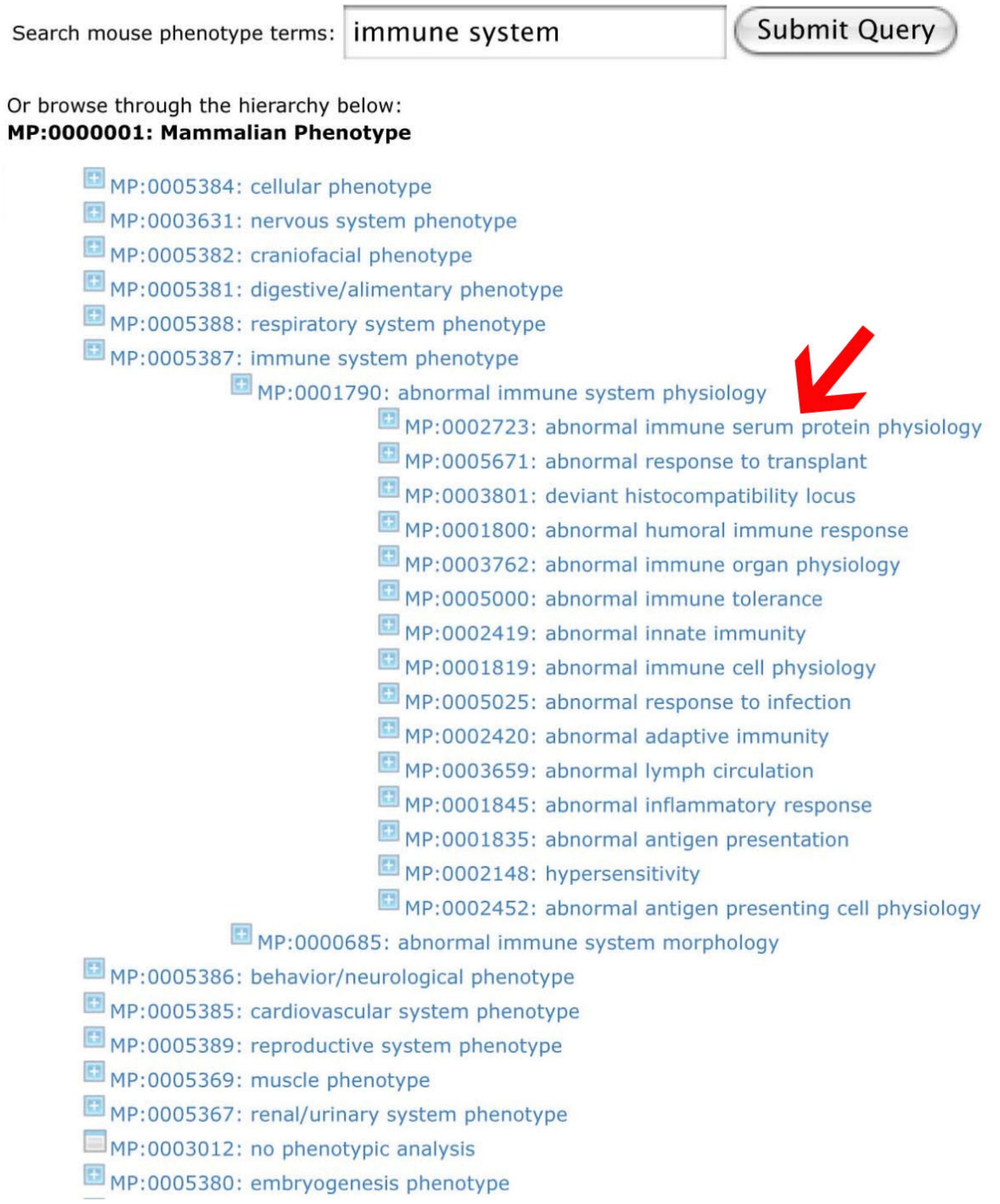
Cart: ip10 rantes cxcl11 ifnb 2009-04-15 10:37:58

Created on: 2009-04-15 10:37:58 Last modified on: 2009-05-07 14:40:43

The following GO Clusters were found:

GO Term	No of occurances	Found in
GO:0005125 cytokine activity	4	Ccl5 Cxcl10 Cxcl11 Ifnb1
GO:0006935 chemotaxis	3	Ccl5 Cxcl10 Cxcl11
GO:0008009 chemokine activity	3	Ccl5 Cxcl10 Cxcl11
GO:0006955 immune response	3	Ccl5 Cxcl10 Cxcl11
GO:0005615 extracellular space	5	Ccl5 Cxcl10 Cxcl11 Ifnb1 Igf2
GO:0005576 extracellular region	5	Ccl5 Cxcl10 Cxcl11 Ifnb1 Igf2
GO:0006954 inflammatory response	3	Ccl5 Cxcl10 Cxcl11

Figure 7 GO-term analysis. The representation of different GO-terms across the whole gene list in a cart can be displayed. For this option to be active, GO terms have to be loaded first with the "Load more data for all genes in this cart" option.



Search mouse phenotype terms: immune system

Or browse through the hierarchy below:

MP:0000001: Mammalian Phenotype

- MP:0005384: cellular phenotype
- MP:0003631: nervous system phenotype
- MP:0005382: craniofacial phenotype
- MP:0005381: digestive/alimentary phenotype
- MP:0005388: respiratory system phenotype
- MP:0005387: immune system phenotype
    - MP:0001790: abnormal immune system physiology
        - MP:0002723: abnormal immune serum protein physiology
        - MP:0005671: abnormal response to transplant
        - MP:0003801: deviant histocompatibility locus
        - MP:0001800: abnormal humoral immune response
        - MP:0003762: abnormal immune organ physiology
        - MP:0005000: abnormal immune tolerance
        - MP:0002419: abnormal innate immunity
        - MP:0001819: abnormal immune cell physiology
        - MP:0005025: abnormal response to infection
        - MP:0002420: abnormal adaptive immunity
        - MP:0003659: abnormal lymph circulation
        - MP:0001845: abnormal inflammatory response
        - MP:0001835: abnormal antigen presentation
        - MP:0002148: hypersensitivity
        - MP:0002452: abnormal antigen presenting cell physiology
    - MP:0000685: abnormal immune system morphology
- MP:0005386: behavior/neurological phenotype
- MP:0005385: cardiovascular system phenotype
- MP:0005389: reproductive system phenotype
- MP:0005369: muscle phenotype
- MP:0005367: renal/urinary system phenotype
- MP:0003012: no phenotypic analysis
- MP:0005380: embryogenesis phenotype

Figure 8 Browsing MP terms in MUSIG-Phen. On the MUSIG-Phen search site, the user may type in a phenotype description. Hitting the "Search" button will retrieve appropriate hits from the MP ontology. The different levels of the ontology may then be displayed by clicking in the "+" box. A click onto the MP term itself (arrow) will activate a query to the MGI database and retrieve all genes that are associated with it.

[39] or the mouse-centric MRB could easily be integrated and used to create a wider array of services. These services could also be linked to create a Taverna-like workflow tool which automatically matches IDs and fields from different services. The current limitation to this approach is the lack of standardization across databases and web services with respect to the use of ontologies and the naming of web service fields. For example a field for MGI gene IDs could be called mgi_id, gene_id, MGIGeneId *etc*. which would make automatic matching impossible. We therefore favor the idea to develop a web service field ontology which should be integrated into MRB or Biocatalogue to provide a look-up service for field names. Currently developments are ongoing within the Biocatalogue project to create a web service ontology to which web service developers annotate their fields which may provide a suitable solution to this problem.

The source code and documentation for the MUSIG-Phen prototype may be downloaded form the following web server: http://www.casimir.org.uk/sourcecode/

Discussion and Conclusion

The aim of generating the MUSIG-Gen and MUSIG-Phen applications was to provide a first set of solutions to user-defined use-cases and thereby generate a test environment for a fully distributed integration strategy. We also presented the applications to various user groups and collected their feed-back. All users appreciated the tools which were able to integrate data from several databases, and they especially liked the principle of the shopping cart. An additional, often mentioned suggestion was to link the genes in MUSIG-Gen to mouse mutants and phenotypes as well as gene expression information. We are planning to add these functionalities to future prototypes.

Our plan for a third use-case is to define the needs for an integration of mouse and human functional genomics databases. Here, we believe that the phenotype-centric use case may serve as a valuable basis to provide an entry point for clinical researchers. The concept would be to enter descriptions of human disease phenotypes as queries and to obtain mouse phenotype descriptions which relate to these terms. However, for such a query, it will first be necessary to relate the human phenotype descriptions with MP terms or with more detailed EQ-based phenotype descriptions.

Acknowledgements

We thank George Gkoutos for helpful discussions on the technical developments of the use cases and all the members of the international advisory board of the CASIMIR consortium http://www.casimir.org.uk/ for their helpful comments and support.

This work was supported by the Commission of the European Community, Framework Programme 6 contract no. LSHG-CT-2006-037811, NWO Rubicon, grant 825.09.008, CASIMIR and the HZI program Infection & Immunity.

Author details

[1]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, UK. [2]Department of Infection Genetics, Helmholtz Centre for Infection Research University of Veterinary Medicine Hannover, Inhoffenstr. 7, D-38124 Braunschweig, Germany. [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [4]Department of Genetics, University Medical Center Groningen Groningen Bioinformatics Centre, University of Groningen, P.O. Box 30001, 9700 RB Groningen, The Netherlands. [5]http://www.CASIMIR.org.uk/.

Authors' contributions

KLS conceived the study, organised the user workshops, developed the use-cases and wrote the manuscript. DS deployed Biomart for the various resources used for the use-case implementations. MS was involved in developing the use-cases and drafting the manuscript. RA developed the use-cases, set-up the Symatlas web service and drafted the manuscript. MG developed the prototypes, conducted the user demonstrations and wrote the manuscript. PNS coordinates the CASIMIR project and was involved in developing the use-cases and drafting the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 17 November 2009
Accepted: 22 January 2010 Published: 22 January 2010

References

1. Peters LL, Robledo RF, Bult CJ, Churchill GA, Paigen BJ, Svenson KL: **The mouse as a model for human biology: a resource guide for complex trait analysis.** *Nat Rev Genet* 2007, **8**:58-69.
2. Rosenthal N, Brown S: **The mouse ascending: perspectives for human-disease models.** *Nat Cell Biol* 2007, **9**:993-9.
3. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA: **The Mouse Genome Database (MGD): mouse biology and model systems.** *Nucl Acids Res* 2008, **36**:D724-8.
4. Bogue MA, Grubb SC, Maddatu TP, Bult CJ: **Mouse Phenome Database (MPD).** *Nucl Acids Res* 2007, **35**:D643-9.
5. Mallon AM, Blake A, Hancock JM: **EuroPhenome and EMPReSS: online mouse phenotyping resource.** *Nucl Acids Res* 2008, **36**:D715-8.
6. Wang J, Williams RW, Manly KF: **WebQTL: Web-based complex trait analysis.** *Neuroinformatics* 2003, **1**:299-308.
7. Zouberakis M, Chandras C, Hancock JM, Schofield PN, Aidinis V: **The Mouse Resource Browser (MRB) - A near-complete registry of mouse resources.** *BioInformatics and BioEngineering. BIBE 2008. 8th IEEE International Conference on (2008)* 2008, 1-5.
8. Hancock J, Chandras C, Zouberakis M, Aidinis V, Schofield PN: **Integrating information from EU-funded mouse functional genomics projects: a questionnaire-based analysis.** *BioInformatics and BioEngineering. BIBE 2008. 8th IEEE International Conference on (2008)* 2008, 1-5.
9. Hancock J, Schofield PN, Chandras C, Zouberakis M, Aidinis V, Smedley D, Rosenthal N, Schughart K: **CASIMIR: Coordination and Sustainability of International Mouse Informatics Resources.** *BioInformatics and BioEngineering. BIBE 2008. 8th IEEE International Conference on (2008)* 2008, 1-5.
10. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **Bio-Mart–biological queries made easy.** *BMC Genomics* 2009, **10**:22.
11. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services.** *Nucl Acids Res* 2006, , **34 Web Server:** W729-32.
12. **Galaxy.** http://galaxy.psu.edu/.
13. Swertz MA, De Brock EO, Van Hijum SA, De Jong A, Buist G, Baerends RJ, Kok J, Kuipers OP, Jansen RC: **Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases.** *Bioinformatics* 2004, **20**:2075-83.

14. Swertz MA, Jansen RC: **Beyond standardization: dynamic software infrastructures for systems biology.** *Nat Rev Genet* 2007, **8**:235-43.
15. Smedley D, Swertz MA, Wolstencroft K, Proctor G, Zouberakis M, Bard J, Hancock JM, Schofield P: **Solutions for data integration in functional genomics: a critical assessment and case study.** *Brief Bioinform* 2008, **9**:532-44.
16. Sundberg J, Schofield P: **A mouse by any other name.** *Journal of Investigative Dermatology* 2009, **129**:1599-1601.
17. **Guidelines for Nomenclature of Mouse and Rat Strains.** http://www.informatics.jax.org/mgihome/nomen/strains.shtml.
18. **HUGO Gene Nomenclature Committee.** http://www.genenames.org/.
19. **Rat Genome and Nomenclature Committee.** http://ratmap.gen.gu.se/RGNC/.
20. Eppig JT, Blake JA, Bult CJ, Richardson JE, Kadin JA, Ringwald M: **Mouse genome informatics (MGI) resources for pathology and toxicology.** *Toxicol Pathol* 2007, **35**:456-7.
21. **Eurexpress.** http://www.eurexpress.org/ee/.
22. **BioGPS.** http://biogps.gnf.org/?referer=symatlas#goto=welcome.
23. Parkinson , *et al*: **ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Res* 2009, , **37 Database:** D868-72.
24. Hermjakob H, *et al*: **IntAct - an open source molecular interaction database.** *Nucl Acids Res* 2004, **32**:D452-D455.
25. Smith CL, Goldsmith CA, Eppig JT: **The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.** *Genome Biol* 2005, **6**:R7.
26. **Eumodic.** http://www.eumodic.org/aboutus.html.
27. **Disease Ontology.** http://diseaseontology.sourceforge.net/.
28. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S: **The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.** *Am J Hum Genet* 2008, **83**:610-5.
29. Schofield , *et al*: **Pathbase: a database of mutant mouse pathology.** *Nucl Acids Res* 2004, D512-5.
30. **Online Mendelian Inheritance in Man, OMIM (TM).** McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) 2009http://www.ncbi.nlm.nih.gov/omim/.
31. **Fly phenotype ontology.** http://subversion.flymine.org/tags/flymine_release_2_1/flymine/model/phenotype/phenotype.ontology.
32. Tweedie S, *et al*: **FlyBase: enhancing Drosophila Gene Ontology annotations.** *Nucl Acids Res* 2009, **37**:D555-D559.
33. **C. elegans phenotype ontology.** http://www.obofoundry.org/cgi-bin/detail.cgi?id=worm_phenotype.
34. Tamberlyn Bieri, *et al*: **WormBase: new content and better access.** *Nucl Acids Res* 2007, **35**:D506-10.
35. Aidinis V, *et al*: **MUGEN mouse database; animal models of human immunological diseases.** *Nucl Acids Res* 2008, D1048-54.
36. Cote RG, Jones P, Apweiler R, Hermjakob H: **The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries.** *BMC Bioinformatics* 2006, **7**:97.
37. **HTML-AJAX.** http://pear.php.net/package/HTML_AJAX.
38. Wilkinson MD, Links M: **BioMOBY: an open source biological web services proposal.** *Brief Bioinform* 2002, 331-41.
39. Goble CA, Stevens RD, Hull D, Wolstencroft K, Lopez R: **Data Curation + Process Curation = Data Integration + Science.** *Brief Bioinform* 2008, **9**:506-517.

doi:10.1186/1756-0500-3-16
Cite this article as: Gruenberger *et al.*: Towards the integration of mouse databases - definition and implementation of solutions to two use-cases in mouse functional genomics. *BMC Research Notes* 2010 **3**:16.