Database, Vol. 2011, Article ID bar029, doi:10.1093/database/bar029

Original Article

The BioMart interface to the eMouseAtlas gene expression database EMAGE

Peter Stevenson*, Lorna Richardson, Shanmugasundaram Venkataraman, Yiya Yang and Richard Baldock

The Institute of Genetics and Molecular Medicine, MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, UK

***Corresponding author:** Tel: +44 (0)131 332 2471; Fax: +44 (0)131 467 8456; Email: peter.stevenson@hgu.mrc.ac.uk

Submitted 1 April 2011; Revised 2 June 2011; Accepted 13 June 2011

Here, we describe the BioMart interface to the eMouseAtlas gene expression database EMAGE. EMAGE is a spatiotemporal database of *in situ* gene expression patterns in the developing mouse embryo. BioMart provides a generic web query interface and programmable access using web services. The BioMart interface extends access to EMAGE via a powerful method of structuring complex queries and one with which users may already be familiar with from other BioMart implementations. The interface is structured into several data sets providing the user with comprehensive query access to the EMAGE data. The federated nature of BioMart allows scope for integration and cross querying of EMAGE with other similar BioMarts.

Database URL: http://biomart.emouseatlas.org

Project description

EMAGE (http://www.emouseatlas.org/emage) is a freely available, curated, online database of *in situ* gene expression patterns in the developing mouse embryo (1–3). Gene expression domains extracted from raw images are integrated spatially into a set of standard 3D virtual mouse embryo models at different stages of development. In addition, an anatomy ontology (4) is used to describe sites of expression using text annotations. The EMAGE web site provides specialized interfaces that allow users to search for gene expression patterns. Uniquely, EMAGE searches can be made using spatially defined regions, as well as standard text based querying. Search results concentrate on providing access to images of original experiments and mapped expression patterns. However, the database also contains a considerable amount of information describing the experiments that were carried out to produce the gene expression patterns, as well as the data provenance. Much of this ancillary data, although visible on the detailed submission pages, was not directly searchable in the EMAGE web site before the implementation of the BioMart (5) interface (http://biomart.emouseatlas.org). BioMart allows a greater range of possible queries that can include experiment details and other ancillary data. It also allows the possibility of integration with other BioMarts. However, BioMart cannot currently provide spatial searches or ready access to image data. The EMAGE web site and the BioMart interface therefore should be seen as complementary methods of querying EMAGE data content.

Data content

The EMAGE data is divided into three BioMart databases.

EMAGE browse repository BioMart database

This database provides a summary view of the curated EMAGE submissions as well as additional assays that are not suitable for spatial mapping and have therefore not been included as full EMAGE submissions. It has a single data set. The *Repository Browse* data set has a limited number of filters and attributes providing an

Database, Vol. 2011, Article ID bar029, doi:10.1093/database/bar029

overview of the EMAGE data. It is the equivalent of the 'browse' and 'quick search' options on the EMAGE web site.

EMAP anatomy ontology BioMart database

This database provides a view of the standard anatomy ontology of mouse development (EMAP) as used in the EMAGE database. The anatomy ontology is constructed as a directed acyclic graph (DAG), with a parent and child relationship, the relationship being 'part-of', for example 'elbow' is part-of 'arm'. The ontology is defined for each Theiler Stage (6) of development and for all stages combined (the abstract ontology). In the BioMart interface these are differentiated by the descriptions 'timed stages' and 'abstract'. Each node in the DAG is given a component name and a unique ID. The database contains two data sets.

Anatomy components (component details) data set. This data set provides details of the components of the ontology. The component name, stage and ID form the BioMart filters for this data set as well as being default attributes. Further attributes include the DAG relationship information plus additional ontological data.

This data set can be linked in a query to the EMAGE submission data set and to the EMAGE text expression data set. However, a query for gene expression using this data set will find results only for that single anatomy component of the ontology. To search for expression in a component and its children, the Anatomy Components (including subcomponents) data set should be used.

Anatomy components (including subcomponents). This data set consists of the anatomy ontology components at each timed stage along with each of the child subcomponents that are 'part-of' that particular component. The filters for this data set are the component name, ID and Theiler stage. The attributes available are name, ID and component path of the parent and child term, plus Theiler stage. The 'component path' is the series of component names from the highest level parent through its descendants to the component term.

This data set can be linked in a query to the EMAGE submission data set and to the EMAGE text expression data set.

EMAGE gene expression BioMart database

This database comprises the main source of EMAGE gene expression and ancillary data. The database contains four data sets. Of these, the primary data set is the EMAGE submission data set. The other three data sets, as well as the anatomy ontology data set, can be linked in a query to allow comprehensive search capabilities where this is required. Limitations in the translation of the database schema from native EMAGE to BioMart have resulted in this somewhat fragmented data model. Nonetheless, access to all the EMAGE data can be achieved either in a single or linked BioMart query.

EMAGE submission data set. This data set comprises the bulk of data in EMAGE. An EMAGE submission generally will be one specimen stained for the expression of one gene or protein at one point in embryonic development. Associated with each submission there may be details of the experiment, annotated sites of expression, the names and details of the data submitter or source, relevant links to other data, acknowledgements to data providers and/or references, details of the images, associated gene synonyms and GO terms. Many search attributes are available and have been grouped broadly in the categories just described. The filters for this data set are grouped by gene, submission, experiment, and links. The EMAGE submission data set can be linked in a single query with the EMAGE experiment data set or the EMAGE spatial expression annotation data set if required.

EMAGE experiment data set. This data set consists of data pertinent to the *in situ* experiment that produced the gene expression image used in a particular EMAGE submission. This covers preparation and description of the specimen used the targeted gene, detection reagent type, probe sequences and cell line data. Common filters and attributes are grouped together for convenience and certain attributes are set as default for results. The EMAGE experiment data set can be linked in a single query with the EMAGE submission data set and the EMAGE text expression annotation data set if required.

EMAGE spatial expression data set. This data set consists of data pertinent to spatial gene expression data recorded in EMAGE. A relatively limited set of filters are available for this data set and the attributes grouped as 'Mapped position attributes' are probably only useful in the context of the full EMAGE database where the spatial information can be used for querying and visualization. Nonetheless, an indication of spatial gene expression can be found using the filters provided. The most useful of the expression filters is probably 'strength'. EMAGE gene expression strength is indicated with values of strong, moderate, weak, possible or not detected. The default attribute 'EMAGE ID' is returned in the results table as a URL link to the submission page where full details and visualizations of the mapped expression patterns can be seen. The EMAGE spatial expression data set can be linked in a single query with the EMAGE submission data set if required.

EMAGE text expression annotation data set. This data set comprises the EMAGE gene expression data that has been annotated using text descriptions from the anatomy ontology. The filters for this data set, which are grouped as expression text annotation filters, include ontology ID along with expression strength and pattern. Additional filters are also available, under the group headings of gene, submission and experiment filters. Attributes are grouped in the same way as the filters. This data set can be linked in a single query with the EMAGE experiment data set if required. This allows additional experiment attributes to be found that are not available from the text expression data set. This data set can also be linked in a query with the anatomy ontology data set in order to obtain more information about an ontology term such as its name, path and stage range.

Query examples

A number of example queries are given in this section. These examples demonstrate the use of the different data sets described earlier, and show how these can be used to provide powerful and flexible searches of the EMAGE data. Screenshots of each query are available as online Supplementary Data.

(1) EMAGE browse repository database.

Query #1. 'Find a summary of gene expression for Fgf family genes at Theiler Stage 18'

Data set	Filters	Attributes
Repository Browse	Gene/Protein: fgf% Theiler Stage: 18	Resource ID Gene/Protein Detection Reagent Theiler Stage Stage Given Assay Specimen Type Mutant Allele URL

Querying the Repository Browse data set provides an overview of gene expression data available in the EMAGE repository as illustrated by Query 1. The query has incorporated the wild-card symbol (%) in order to find all gene symbols starting with 'fgf'. The attribute 'URL' provides a link to either the full submission description in EMAGE for fully curated data, or to the original image source in an external resource if the submission is not fully curated in EMAGE.

(2) EMAP anatomy ontology database.

Query #2. 'Find text annotated gene expression in the limb for the Wnt and Hox gene families'.

Data set	Filters	Attributes
Anatomy Components (including subcomponents)	Anatomy component name: limb	Parent component name Parent component path Child component name Child component path
EMAGE text expression annotation data set	MGI gene symbol: Wnt%, Hox%	Strength (text annotation) Pattern (text annotation) MGI gene symbol Emage ID Theiler stage

The Anatomy Components (including subcomponents) can be used to find gene expression annotation in EMAGE when used as a linked database query as illustrated by Query 2. The query will return results for the component chosen as a filter and all of its subcomponents. Here, the query term 'limb' also retrieves expression in, for example, 'forelimb bud'. The 'Emage ID' attribute appears in the results as a link to the full submission description in EMAGE (as it does in all cases where the EMAGE ID is chosen as an attribute).

(3) EMAGE gene expression database.

Query #3. 'Find the EMAGE submissions and details of probes used to detect the Fgf gene family at Theiler stages 16–20 where the specimen type is section data (or unknown) and the assay quality is the highest value'.

Data set	Filters	Attributes
EMAGE submission data set	MGI gene symbol: fgf% Theiler stage: 16–20 Assay quality: 3 Specimen type: section, unknown	MGI gene symbol Emage ID Theiler stage Staining procedure Embedding reagent Clearing method Fixation Specimen type Assay quality
EMAGE experiment data set		Detection reagent identifier Detection reagent sequence type Detection reagent notes ISH probe generated from ISH probe chemistry ISH probe strand ISH probe label

Querying the EMAGE gene expression BioMart database allows potentially complex sets of query conditions and output parameters to be defined. This provides a powerful tool to filter views of the gene expression data, its experimental details, and other associated data as illustrated by Query 3. The query is constructed to filter the EMAGE submission data set for gene, stage, assay quality and

Database, Vol. 2011, Article ID bar029, doi:10.1093/database/bar029

specimen type. The query requires to be linked with the EMAGE experiment data set in order to show attributes of the experiment and detection reagents.

Query #4. 'For the detection reagent with the identifier MGI:1334951 (a specific Fgf8 riboprobe) find the EMAGE submissions and the publication details of the experiment along with any further published references to this experiment'.

Data set	Filters	Attributes
EMAGE experiment data set	Detection reagent identifier: MGI:1334951	MGI gene symbol Emage ID Theiler stage Detection reagent identifier Authors
EMAGE submission data set		Accession (reference) Authors Publication year Title Publication name Publication issue Publication volume Pages

For a particular *in situ* hybridization probe of interest a user can find all of the instances of gene expression recorded in EMAGE for that probe, along with references to the experiment in the published literature as illustrated by Query 4. The query is filtered in the EMAGE experiment data set using the detection reagent identifier. The query is linked with the EMAGE submission data set in order to show the various attributes of the publication details.

Discussion and future direction

The EMAGE database was conceived primarily to provide a unique method of collating, curating, querying and analysing spatiotemporal patterns in the mouse embryo including gene expression. In order for the gene expression data to be provided in context, additional metadata pertaining to the *in situ* experiments is also included in the database. This data is curated rigorously by EMAGE editorial staff to provide a high quality data set. The primary focus of the EMAGE website is to provide tools to search for gene expression, to display this data to users in a visual form as far as possible, and to enable spatial query and analysis. The implementation of BioMart in EMAGE now allows an additional text based interface that addresses a need for focussed searches across all of the text based information held in EMAGE.

The data mining aspect of BioMart is an important factor in providing additional search functionality to EMAGE. However, BioMart also provides a standardized web interface, an application programming interface (API) and a RESTful web service. These interfaces are being increasingly adopted by a variety of well known biological databases and are described further by Haider *et al.* (7). This means that biological researchers and programmers may already be familiar with the BioMart interface and may prefer to access a database such as EMAGE using tools and techniques that they have already used with other BioMart implementations.

Another important aspect of BioMart is that it allows federated queries between databases (5). The ability to integrate data in this way will be the focus of future development of the EMAGE BioMart. A number of existing and proposed gene expression databases share a high degree of commonality with the EMAGE database. These include Eurexpress (8) a transcriptome-wide mouse gene expression study that shares the same anatomy ontology. Eurexpress currently has its own BioMart implementation but it is planned to absorb this data into EMAGE, in which case a unified Biomart will be provided. GUDMAP (9), a genito–urinary molecular anatomy project also has a common anatomy ontology and shares database administration and development with EMAGE. It is planned to provide a BioMart implementation for GUDMAP in the future. It may be possible to integrate this data into a single BioMart instance with EMAGE but more likely would be to link these in a federated manner. The eChickAtlas project (http://www.echickatlas.org/) is a chicken embryo gene expression database that is currently under development. This database will use an ontology derived from, and mapped to, the same source as the EMAP mouse ontology. This provides scope for interesting cross species queries to be carried out and could be implemented using the federated nature of BioMart.

Integration of databases in this way should provide researchers concerned with gene expression data access to a large volume of information in an easily accessible and unified manner.

Finally, extending BioMart to spatially mapped data opens up the possibility of specifically spatial queries and filters. These could be implemented via simple 'punch-out' type capabilities to the EMAGE graphical interfaces or via stored queries held in the database but will require extension of the underlying BioMart model. If this were possible then it would provide a more significant level of integration through to spatially organized data.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors thank Damian Smedley at the EBI and Bernard Haggerty here in the MRC HGU for helping the

development of the EMAGE BioMart and both the OICR and EBI for linking EMAGE on their portal.

Funding

UK Medical Research Council (MRC; as part of the MouseAtlas project at the MRC Human Genetics Unit). Funding for open access charge: Medical Research Council.

Conflict of interest. None declared.

References

1. Richardson,L., Venkataraman,S., Stevenson,P. *et al*. (2010) EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res.*, **38 (**Suppl 1), D703–D709.
2. Venkataraman,S., Stevenson,P., Yang,Y. *et al*. (2008) EMAGE: Edinburgh mouse atlas of gene expression: 2008 update. *Nucleic Acids Res.*, **36 (**Suppl 1), D860–D865.
3. Christiansen,J.H., Yang,Y., Venkataraman,S. *et al*. (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res.*, **34 (**Suppl 1), D637–D641.
4. Bard,J.B.L., Kaufman,M.H., Dubreuil,C. *et al*. (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech. Dev.*, **74**, 111–120.
5. Smedley,D., Haider,S., Ballester,B. *et al*. (2009) BioMart – biological queries made easy. *BMC Genomics*, **10**, 22.
6. Theiler,K. (1989) *The House Mouse - Atlas of Embryonic Development*. Springer-Verlag, New York.
7. Haider,S., Ballester,B., Smedley,D. *et al*. (2009) BioMart Central Portal—unified access to biological data. *Nucleic Acids Res.*, **37 (**Suppl 1), W23–W27.
8. Diez-Roux,G., Banfi,S., Sultan,M. *et al*. (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.*, **9**, e1000582.
9. McMahon,A.P., Aronow,B.J., Davidson,D.R. *et al*. (2008) GUDMAP: the genitourinary developmental molecular anatomy project. *J. Am. Soc. Nephrol.*, **19**, 667–671.