Published online 4 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D847–D853***
doi:10.1093/nar/gkm797

4D*Xpress*: a database for cross-species expression pattern comparisons

Yannick Haudry[1], Hugo Berube[2], Ivica Letunic[1], Paul-Daniel Weeber[1], Julien Gagneur[1], Charles Girardot[1], Misha Kapushesky[2], Detlev Arendt[1], Peer Bork[1], Alvis Brazma[2], Eileen E. M. Furlong[1], Joachim Wittbrodt[1] and Thorsten Henrich[1,]*

[1]European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany and [2]European Bioinformatics Institute, EMBL-EBI Wellcome Trust Genome Campus Hinxton, Cambridge, CB10 1SD, UK

Received August 14, 2007; Revised September 14, 2007; Accepted September 17, 2007

ABSTRACT

In the major animal model species like mouse, fish or fly, detailed spatial information on gene expression over time can be acquired through whole mount *in situ* hybridization experiments. In these species, expression patterns of many genes have been studied and data has been integrated into dedicated model organism databases like ZFIN for zebrafish, MEPD for medaka, BDGP for *Drosophila* or GXD for mouse. However, a central repository that allows users to query and compare gene expression patterns across different species has not yet been established. Therefore, we have integrated expression patterns for zebrafish, *Drosophila*, medaka and mouse into a central public repository called 4D*Xpress* (expression database in four dimensions). Users can query anatomy ontology-based expression annotations across species and quickly jump from one gene to the orthologues in other species. Genes are linked to public microarray data in ArrayExpress. We have mapped developmental stages between the species to be able to compare developmental time phases. We store the largest collection of gene expression patterns available to date in an individual resource, reflecting 16 505 annotated genes. 4D*Xpress* will be an invaluable tool for developmental as well as for computational biologists interested in gene regulation and evolution. 4D*Xpress* is available at http://ani.embl.de/4DXpress .

INTRODUCTION

Precise spatio-temporal gene expression is crucial during the development of an organism. Combinations of transcription factors give distinct identities to embryonic structures, tissues and cell types and trigger complex developmental processes like embryonic patterning, morphogenesis and differentiation. To know the exact time and location of gene transcripts is essential when studying the functions of genes involved in developmental processes as well as for trying to decipher the code of *cis*-regulatory modules. Therefore expression localization data has been gathered by the dedicated model species databases like ZFIN for zebrafish (1), BDGP (2) and FlyBase (3) for *Drosophila*, MEPD (4) for medaka, Aniseed for ciona, XDB3 for *Xenopus* or GXD (5) and EMAGE (6) for mouse. A central platform, which allows users to compare gene expression in different species, however, has not yet been established. Such a resource would be invaluable not only to complement lacking expression information in one species by annotations done in other species, but also to study the evolutionary origin of embryonic structures.

Here we provide a platform for a cross-species expression pattern resource. 4D*Xpress* (expression database in four dimension) stores images, which lets biologists see and judge expression patterns together with an organized annotation. It allows users to query the data and makes data accessible to computational analysis. Our vision is that in a few years time the exact localization of each single transcript will be known for the major model species. We hope that our resource will help to store them in an organized way, to compare different species expression patterns and to provide tools to analyse this data.

DATA INTEGRATION

Data integration is a major challenge of the project. Besides the differences of the model organism themselves, databases provide gene expression data in different formats (flat files, sql-dumps, direct database access) and annotation has been done differently (screens, literature, curators).

*To whom correspondence should be addressed. Tel: +49 6221 387 516; Fax: +49 6221 387 166; Email: henrich@embl.de

Table 1. Content of 4D*Xpress*. Annotation status of gene expression patterns at present time

	Source	Genes	Stages	Stages per gene	Anatomy terms	Anatomy terms per gene	Anatomy terms per stage	Distinct anatomy terms
Drosophila	bdgp	5951	21 048	3.54	29 867	5.02	1.42	288
Medaka	mepd	882	27 46	3.11	5047	5.72	1.84	338
Zebrafish	zfin	5779	102 671	17.77	178 851	30.95	1.74	694
Mouse	mgi	3893	127 99	3.29	17 291	4.44	1.35	1661
		16 505	139 264	8.44	231 056	14.00	1.66	2981

Expression data

So far we have integrated expression data for zebrafish (1), *Drosophila* (2), medaka (4) and mouse (5). Table 1 gives an overview on the expression pattern annotations that have been integrated for 4D*Xpress*. The best-annotated model species at the moment are *Drosophila* and zebrafish with almost 6000 annotated genes each. Mouse follows with 3893 annotated genes; some annotations were done using a 3D virtual embryo (6).

Also expression data has been gathered differently. For medaka and *Drosophila* the major annotation results from a screen. Expression has been analysed at distinct time points and cover between 3 and 4 stages per gene on average (Table 1, stages per gene), whereas zebrafish expression patterns are additionally annotated from literature by a team of database curators. Annotation is done for continuous developmental stages.

Anatomy ontologies are often very rich, however only a limited fraction of the terms is actually used for expression annotation (Table 1, distinct annotations). Again, ZFIN uses a rich vocabulary with almost 700 distinct terms. The values for mouse and medaka need to be treated with care, as the ontologies used for annotation here are the cross product of anatomy and stage ontologies and therefore overestimates vocabulary richness.

Our database schema can store all information required by the MISFISHIE standard (minimum information specification for *in situ* hybridization and immunohistochemistry experiments) (7). This will allow us to efficiently adapt other model species as well as developing a data exchange format to keep up to date with other resources.

Cross-species relationships

One of the major goals of our project is to be able to compare gene expression patterns between the different model species. For doing so, relationships need to be established between genes (orthology), between time windows (developmental stages) and most challenging between anatomical structures (homologue/analogue).

Orthology mapping. EnsEMBL compara (8) provides a reliable source of sequence homology relationships, which was computed using a tree-based approach. We have chosen to use this and update regularly upon new EnsEMBL releases. We assigned each gene to a cluster of orthologues using the EnsEMBL notification: one2one-, one2many- and many2many-orthology relationships. Through the web interface (described below), these clusters are visualized as a network and homology relationships are used to sort the gene list retrieved from a query as well as for allowing quick links from one gene to the orthologues in other species.

Developmental stage mapping. It is very difficult to identify corresponding developmental stages in two species, even when comparing two closely related fish species like medaka and zebrafish. For instance in medaka, the head and brain develop faster, whereas the tail and somites develop slower than in zebrafish. So a matching zebrafish stage regarding the number of somites (which is a very popular staging feature) would correspond to an earlier stage than a matching zebrafish stage based on head features.

However there are key events in development, which allow researchers to define a list of eight stages that is described in all developmental biology text books and is common to all bilaterian animals: zygote, cleavage, blastula, gastrula, neurula, organogenesis, juvenile and adult. By mapping each of the species stages onto one of the bilaterian stages the link between species stages can be done and combinatorial explosion can be prevented. A new species will only need to be mapped to the common stages (Figure 1, top right) and not against all stages of all other species (Figure 1, top left).

Obviously temporal resolution is lost when mapping a list of 40 developmental stages onto a list of only eight common stages, but the eight stages seem to be the largest set shared by all bilaterian species and they represent the key events in the development of an organism. The original species-specific stage annotation is not replaced by the stage mapping terms to keep high temporal resolution. However, the stage mapping establishes temporal relationships that can be used for cross-species queries.

Anatomy mapping. The anatomy mapping will be an ongoing process the same as it is also an ongoing debate in the scientific community about which structures can be defined as being homologous. We have not yet carried out a complete anatomy mapping, but we have set up the resources and tools for doing so. Evidence from different analyses will need to be integrated for approaching this problem. One can use lexical, anatomy structure and co-expression cues to establish relationships between the anatomical terms. The first two cues can be used by just comparing the anatomy ontologies available for the model species (9). For the inclusion of co-expression we are currently examining conserved network patterns in species-specific co-expression networks via orthology

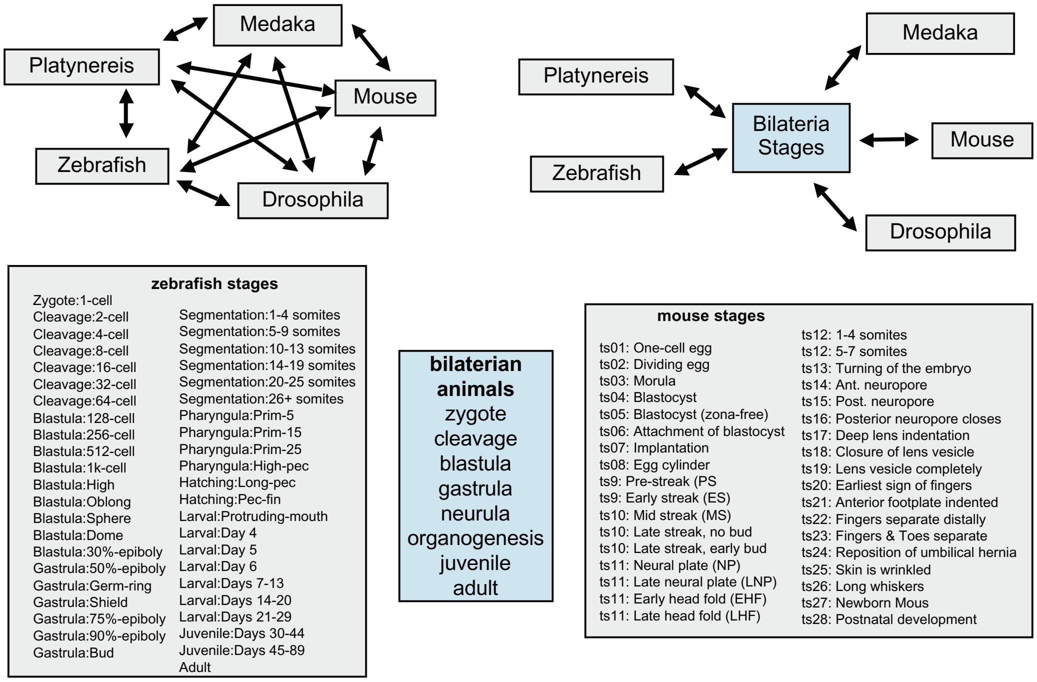


Figure 1. Mapping of developmental species was done via a list of stages common to all bilaterian animals.

relationships. The user can exploit lexical cues already, using the term-based expression search (described below).

The common anatomy reference ontology (CARO) is being developed to facilitate interoperability between existing anatomy ontologies for different species. It aims to provide a template for building new anatomy ontologies. We think CARO could serve as a template to build an anatomy ontology shared by all bilaterians. Similar to the stage mapping we then want to map species-specific anatomy terms onto this common ontology.

4D ArrayExpress data warehouse

Expression data acquired through *in situ* hybridization, antibody or transgenic expression can be complemented through microarray data. The first methods provide high-resolution data in both space and time, which microarray data cannot provide; microarray experiments however can quickly give a quantitative overview on the overall expression of all genes in a genome. Especially useful are time series that provide insight in expression changes during development. That is why we have set up a complementary project at ArrayExpress (10), which stores corresponding microarray data. The project is called 4D ArrayExpress data warehouse (4DDW) and is accessible at: http://www.ebi.ac.uk/microarray-as/4DDW_EMBL/. The 4DDW will be described in detail elsewhere.

So far we have established 4737 reciprocal links for mouse, *Drosophila* and zebrafish. When querying microarray data at the 4DDW users can quickly go to 4D*Xpress* and vice versa. The close linkage of these two resources allows researchers for example to quickly examine the gene expression patterns of a list of genes that cluster together in a microarray experiment.

Expression similarity

Expression patterns within a species can easily be compared when representing the expression annotation as a binary vector (1 for expressed, 0 for not expressed). Different methods to calculate the similarity between these vectors can be applied.

We have chosen the Jaccard coefficient as a similarity measure for a start, which is simple to calculate and has been used in the first BDGP release (2) for the same purpose.

The Jaccard distance has been calculated between the expression vectors of gene pairs. The expression binary vector was compiled considering stage and anatomy. If a gene is expressed (has positive annotation) at a given stage in a given anatomical structure the vector value is set to true, otherwise to false.

The Jaccard similarity coefficient is defined as the size of the intersection divided by the size of the union of the sample vectors:

$$\text{Jaccard similarity coefficient}: J(\mathrm{A}, \mathrm{B}) = |\mathrm{A} \cap \mathrm{B}|/|\mathrm{A} \cup \mathrm{B}|$$

$$\text{Jaccard distance}: J_{\delta}(\mathrm{A,B}) = 1 - \mathrm{J(A,B)}$$

The Jaccard distance is supposed to estimate how different expression patterns are. However this value depends on the extent and quality of the expression annotation. Thus, in the cases where annotations are incomplete or have been done inconsistently, this measure might be misleading. Also, this method treats all anatomical structures equally. Relations defined in the anatomy ontology are not taken into account. In future we will provide additional similarity measures e.g. the semantic similarity, which accounts for that.

Still, the Jaccard distance provides a quick and easy way for identifying similarly annotated genes. The values are stored in the database and helps users to find genes within the species with similar expression patterns. This measure can also be used to cluster genes with similar gene expression pattern annotations as shown for *Drosophila* (2). We use these similarity relationships to generate co-expression networks and plan to search for conserved network patterns across species using orthology relationships.

CROSS-SPECIES OVERLAP

Model species differ; they differ in morphology and function; they differ in accessibility by molecular methods; they differ in the genomic and computational resources and in the size of the scientific community working on them. This is reflected in the number of genes annotated for each model species throughout development (Figure 2). Whereas mouse has a large scientific community behind, it is not producing similar amounts of offspring as egg laying fish and *Drosophila*. Embryos are developing internally resulting in a smaller number of annotated genes; however, with high quality annotation. Zebrafish and *Drosophila* are the most complete data sets. There are huge differences in how they were compiled. Whereas the *Drosophila* data was only acquired in a single screen with only a few annotators, the zebrafish data was collected from several large-scale screens and annotated expression patterns from the literature.

When comparing expression data, it is important to examine the data overlap between the species. For example: How many genes are annotated at corresponding developmental stages? And: How many orthologues have expression annotation?

In Figure 2, we have marked corresponding developmental stages with the same colour (stage mapping as described above).

Zebrafish is spanning most developmental stages and largest temporal overlap exists with *Drosophila* (from cleavage till organogenesis). Neurula and organogenesis stages are the best-annotated stages in all four species and therefore most promising to be compared to each other.

Besides temporal overlap we need orthology overlap to be able to compare expression patterns across species. The overlap we have between annotated genes in the different species combinations is shown in Table 2.

The numbers in Table 2 are getting particularly important when doing global computational analyses. When focusing on two species comparisons, zebrafish and *Drosophila* annotations will yield the largest overlap of 964 annotated orthologous groups, when going for three species mouse should be taken into account.

WEB INTERFACE

4D*Xpress* is a JAVA-based application with a web-based front-end powered by the servlet container TOMCAT and data are stored in a PostgreSQL relational database. The web application is based on a model-view-controller (MVC) architecture using the Struts Framework, and enhanced with applets, JavaScript and AJAX (Asynchronous JavaScript and XML) technologies to build a powerful, interactive, user-friendly interface. 4D*Xpress* is available at http://ani.embl.de/4DXpress.

It is also possible to link to 4D*Xpress* gene entry pages from other projects using the following link http://ani.embl.de/4DXpress/reg/all/search/bquery.do?id = with a gene identifier as the ID value that can be either an EnsEMBL ID, gene symbol or primary identifier from other public resources (e.g. FlyBase IDs, MGI IDs or ZFIN IDs).

Query genes

Genes can be searched either by a range of external identifiers, symbols, names or by their expression pattern annotation. Using the ontology-based form, by selecting a species, the corresponding stage and anatomy ontologies are loaded and information is provided on how many genes are annotated with the listed terms. The term-based form allows more complex queries and cross-species queries can be preformed by selecting 'Bilateria'. Then, a list of search terms can be entered manually or guided by auto-completion of terms, which have been used for annotation. The fact that corresponding structures often have similar names in the different species allows meaningful cross-species queries using this tool.

Upon sending the query a gene list is returned, which provides the user with a summary overview. By default this list is ordered by orthologous groups, which facilitates the comparison of orthologous genes in the different species.

When picking an individual gene entry the full information on that gene is displayed: external identifiers, gene description, expression pattern annotation using stage and anatomy ontologies, images of stained embryos and orthology relationships (visualized as a network). From the gene view a list of orthologues can be selected and their expression annotation and images can be compared to each other on a single page. A cropped screenshot comparing medaka Six3 and its *Drosophila* orthologue Optix is shown in Figure 3.

Also, a list of similarly expressed genes within the same species is provided, which was calculated using the Jaccard coefficient (see above). Users can select some of them and compare them like shown in Figure 3.

Ontology browser

Ontologies are becoming widely used to annotate units of information by providing controlled vocabularies and structured knowledge. Therefore, anatomy ontologies are useful to enforce standard terminology for gene expression annotation as well as for making this information accessible to computational analysis, but at the same time database usage becomes more complex to non-expert users. We provide a tree-based tool to help users to browse ontologies that were used for expression pattern annotation. It allows users to query terms and expand or collapse individual nodes.

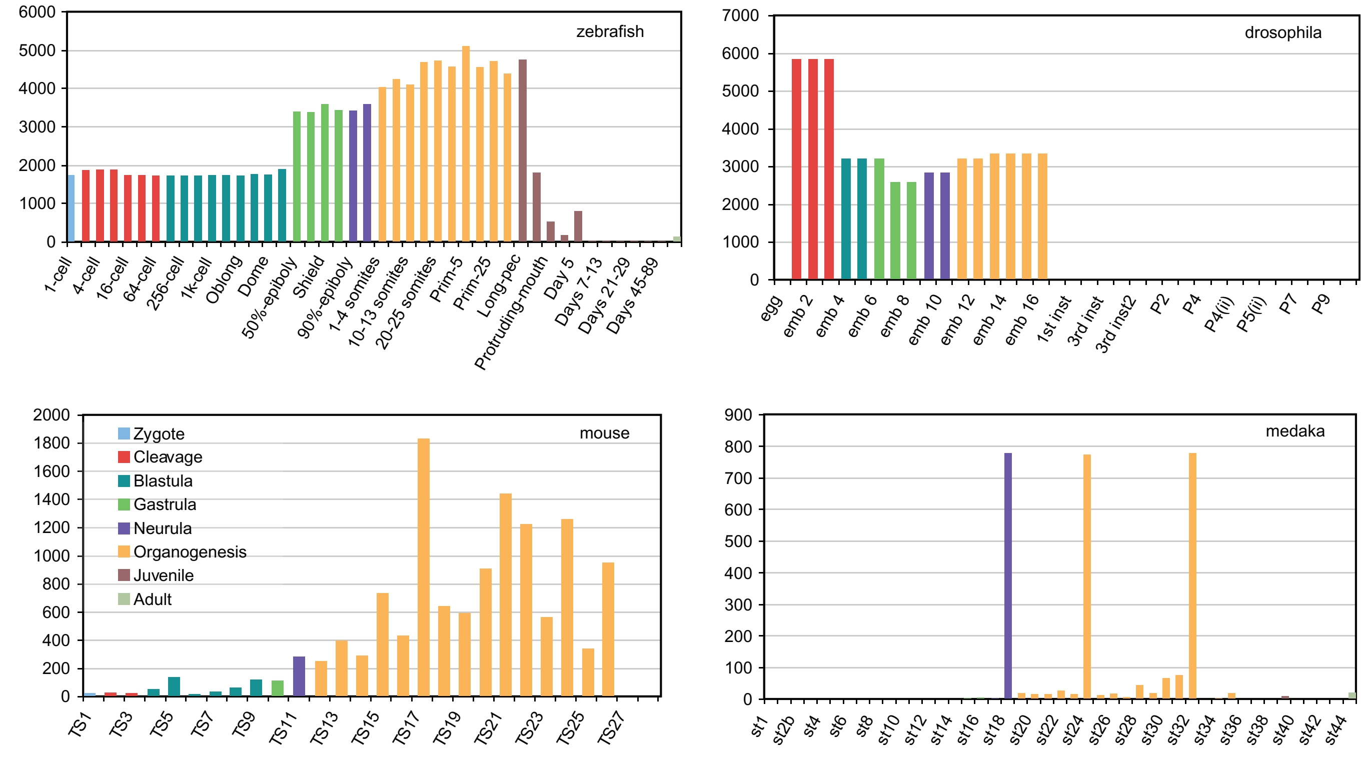


Figure 2. Comparison of the number of genes annotated at species-specific developmental stages in zebrafish, *Drosophila*, mouse and medaka. Corresponding developmental stages have the same colour. Colour legend for all panels is shown in the mouse panel.

Table 2. Number of orthologous groups with genes that are annotated in more than one species

Number of species	COGs
Two species	
Zebrafish, *Drosophila*	964
Mouse, Zebrafish	913
Mouse, *Drosophila*	764
Drosophila, Medaka	341
Zebrafish, Medaka	329
Mouse, Medaka	260
Three species	
Mouse, Zebrafish, *Drosophila*	336
Zebrafish, *Drosophila*, Medaka	156
Mouse, Zebrafish, Medaka	135
Medaka, Mouse, *Drosophila*	124
Four species	
Mouse, Zebrafish, *Drosophila*, Medaka	68

Developmental stage ontologies can be browsed by species and external links provide more information on stage definitions. Species-specific stage ontologies were mapped onto a common stage list (Figure 1) and thereby temporal relationships were established, which can be accessed via web interface.

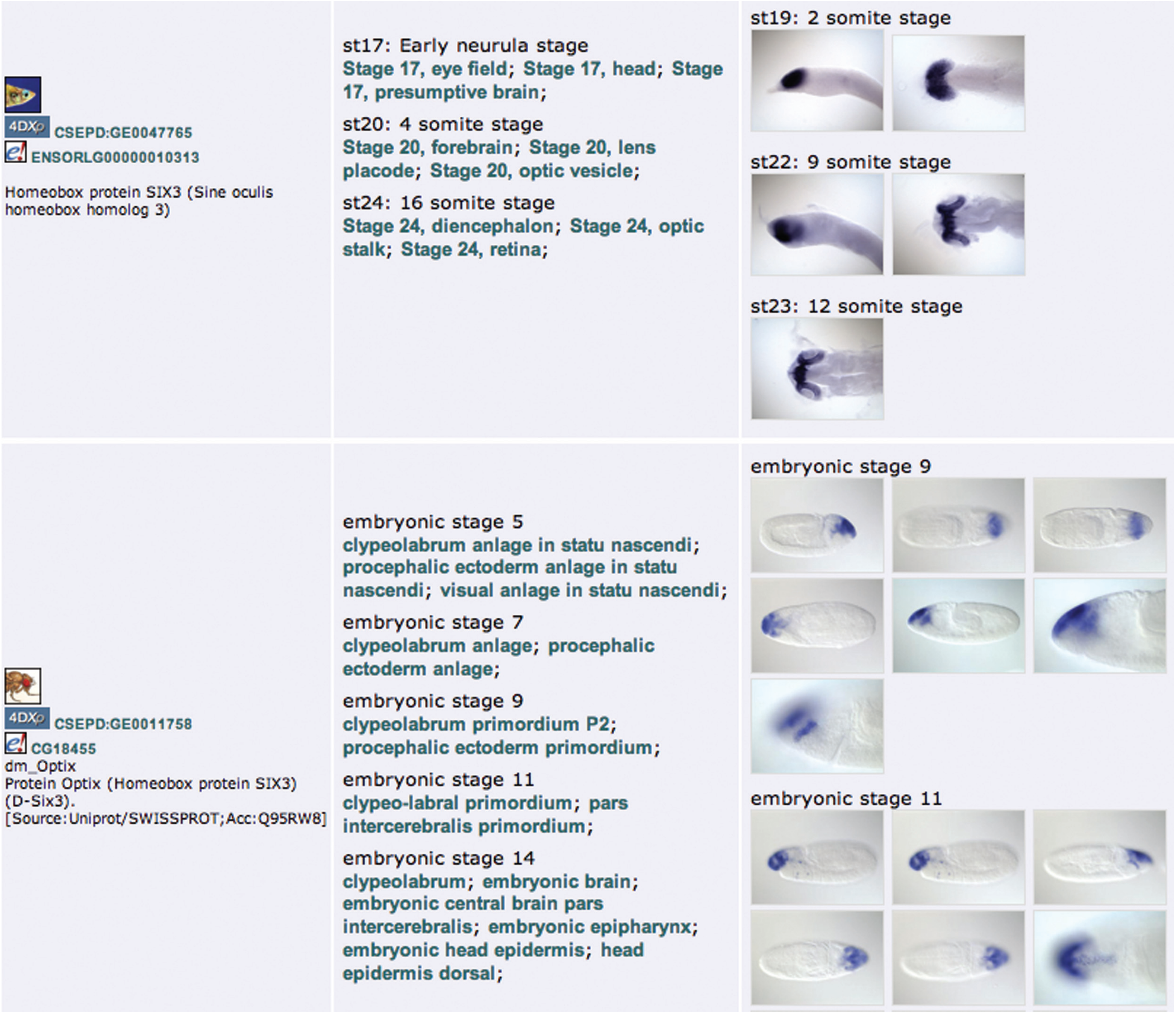
Annotation tool

Our annotation tool allows users to annotate gene expression patterns resulting from any of the three types of experiments: whole mount *in situ* hybridization, transgenic reporter gene expression or antibody staining. The same tool can be used for all supported species (for now: zebrafish, mouse, medaka, *Drosophila*, platynereis). Species-specific ontologies for developmental stages and anatomies can be loaded and users can customize a list of favorite terms to be used.

CONCLUSIONS AND FUTURE DIRECTIONS

We have integrated expression data on 16 505 genes in the four important developmental model species: mouse, zebrafish, *Drosophila* and medaka. We developed a stable database schema and a powerful web interface to access this data. With the interface cross-species queries can be done, facilitated by a stage mapping. An expression similarity measure is implemented to find genes with similar expression patterns and links to all original data sources are provided.

With these tools in place we aim to integrate more species, which are available in the public domain like *Xenopus laevis* with 17.000 images, Ciona and *Caenorhabditis elegans*. We have set up the infrastructure to analyse and compare the data. We will analyse the features of *in situ* co-expression networks and compare them between the species and to co-expression networks derived from microarray data. We will use conserved network patterns to assess mapping of anatomical structures.



CSEPD:GE0047765
ENSORLG00000010313
Homeobox protein SIX3 (Sine oculis homeobox homolog 3)

st17: Early neurula stage
Stage 17, eye field; Stage 17, head; Stage 17, presumptive brain;

st20: 4 somite stage
Stage 20, forebrain; Stage 20, lens placode; Stage 20, optic vesicle;

st24: 16 somite stage
Stage 24, diencephalon; Stage 24, optic stalk; Stage 24, retina;

st19: 2 somite stage

st22: 9 somite stage

st23: 12 somite stage

CSEPD:GE0011758
CG18455
dm_Optix
Protein Optix (Homeobox protein SIX3) (D-Six3).
[Source:Uniprot/SWISSPROT;Acc:Q95RW8]

embryonic stage 5
clypeolabrum anlage in statu nascendi; procephalic ectoderm anlage in statu nascendi; visual anlage in statu nascendi;

embryonic stage 7
clypeolabrum anlage; procephalic ectoderm anlage;

embryonic stage 9
clypeolabrum primordium P2; procephalic ectoderm primordium;

embryonic stage 11
clypeo-labral primordium; pars intercerebralis primordium;

embryonic stage 14
clypeolabrum; embryonic brain; embryonic central brain pars intercerebralis; embryonic epipharynx; embryonic head epidermis; head epidermis dorsal;

embryonic stage 9

embryonic stage 11

Figure 3. The comparative view shows expression patterns of a list of selected genes (medaka Six3 and its *Drosophila* orthologue Optix). Expression annotation and images can be easily compared between the genes on a single page.

ACKNOWLEDGEMENTS

This work was carried out in the Centre for Computational Biology at EMBL. We are grateful to the model organism database crews for providing us their expression pattern data. In particular we thank Martin Ringwald and Susan McClatchy to give us direct access to the MGI database; thanks to Monte Westerfield and Judy Sprague for helping us with the ZFIN data reports; we thank Pavel Tomancak for helping us to understand the BDGP mySQL schema.

We thank Mirana Ramialison for extending the medaka annotation and Francois Spitz for mouse *in situ* images. We are grateful to the MISFISHIE team for initiating an expression pattern data exchange format. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P. *et al.* (2006) The zebrafish information network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
2. Tomancak,P., Berman,B.P., Beaton,A., Weiszmann,R., Kwan,E., Hartenstein,V., Celniker,S.E. and Rubin,G.M. (2007) Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.*, **8**, R145.
3. Grumbling,G. and Strelets,V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res.*, **34**, D484–D488.
4. Henrich,T., Ramialison,M., Wittbrodt,B., Assouline,B., Bourrat,F., Berger,A., Himmelbauer,H., Sasaki,T., Shimizu,N. *et al.* (2005) MEPD: a resource for medaka gene expression patterns. *Bioinformatics*, **21**, 3195–3197.

5. Smith,C.M., Finger,J.H., Hayamizu,T.F., McCright,I.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Ringwald,M. (2007) The mouse gene expression database (GXD): 2007 update. *Nucleic Acids Res.*, **35**, D618–D623.
6. Christiansen,J.H., Yang,Y., Venkataraman,S., Richardson,L., Stevenson,P., Burton,N., Baldock,R.A. and Davidson,D.R. (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res.*, **34**, D637–D641.
7. Deutsch,E.W., Ball,C.A., Bova,G.S., Brazma,A., Bumgarner,R.E., Campbell,D., Causton,H.C., Christiansen,J., Davidson,D. *et al.* (2006) Development of the minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE). *Omics*, **10**, 205–208.
8. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
9. Zhang,S. and Bodenreider,O. (2003) Aligning representations of anatomy using lexical and structural methods. *AMIA Annu. Symp. Proc.*, **00**, 753–757.
10. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P. *et al.* (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.