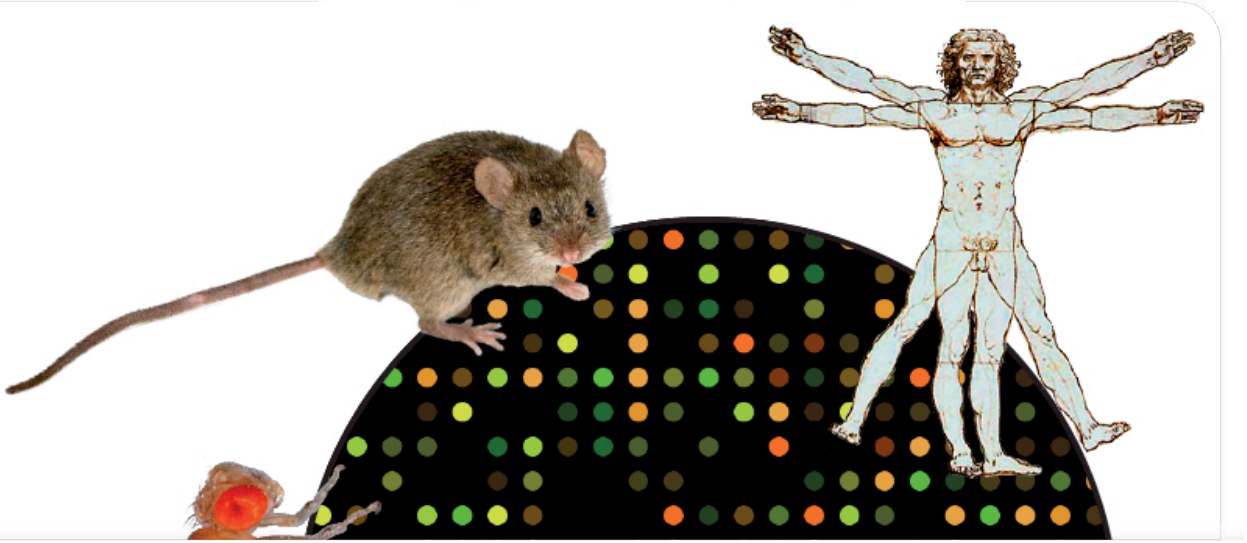
UNIL | Université de Lausanne

SIB



Frédéric Bastian, Marc Robinson-Rechavi

Bgee:
Integrating ontology and homology for the study of gene expression evolution

ASIH 2009

| le savoir vivant |

http://bgee.unil.ch/

Bgee: data**B**ase for **G**ene **E**xpression **E**volution

perform high throughput analyses of gene expression patterns.

http://bgee.unil.ch/

Bgee: data**B**ase for **G**ene **E**xpression **E**volution

Bgee
BGEE | Gene Expression Evolution

perform high throughput analyses of gene expression patterns.

Requirements:

1- integration of heterogeneous expression data

2- comparison criteria between anatomies, developmental stages, and genes

Integrating heterogeneous data

Integrating heterogeneous data poses several problems:

- comparing results of different types of techniques
- comparing results between experiments

Information we want to collect:

- where and when a gene is expressed
- with which confidence

Integrating heterogeneous data

Mapping expression data to ontologies:

- data annotations: need for manual curation
- data granularity: need for ontologies
- experimental factors: limitation to “normal” conditions

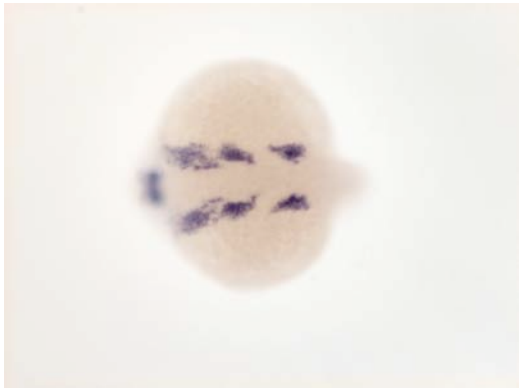
Integrating heterogeneous data

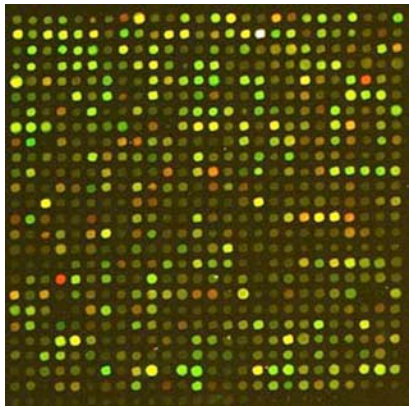
Expression confidence:

Bgee currently includes EST, Affymetrix, and *in situ* hybridization data

For each data: assign a level of confidence (low, high)

Require dedicated statistical tests





Integrating heterogeneous data

Expression confidence: experiments based on tag counting

High confidence: number of tags significantly ≠ 0

⇒ Gene expressed with **95%** confidence: **7 tags** mapped to this gene (Audic and Claverie, Genome Res., 1997)

⇒ Low confidence: 1 to 6 tags

⇒ High confidence: ≥ 7 tags

Integrating heterogeneous data

Expression confidence: Affymetrix data

All probes mapping to the same transcript
= a probeset

Gene significantly expressed:
probeset signal ≠ background signal

Integrating heterogeneous data

Expression confidence: Affymetrix data

Critical point: estimation of the background signal

- Use of the gcRMA algorithm to normalize the signal
- Use a subset of weakly expressed probesets for estimating the background

(Schuster *et al*, Genome Biology, 2007)

Wilcoxon test: probeset signal vs background signal

Low confidence: 1% < p-value ≤ 5%

High confidence: p-value ≤ 1% & consistency of all probes

http://bgee.unil.ch/

Bgee: data**B**ase for **G**ene **E**xpression **E**volution

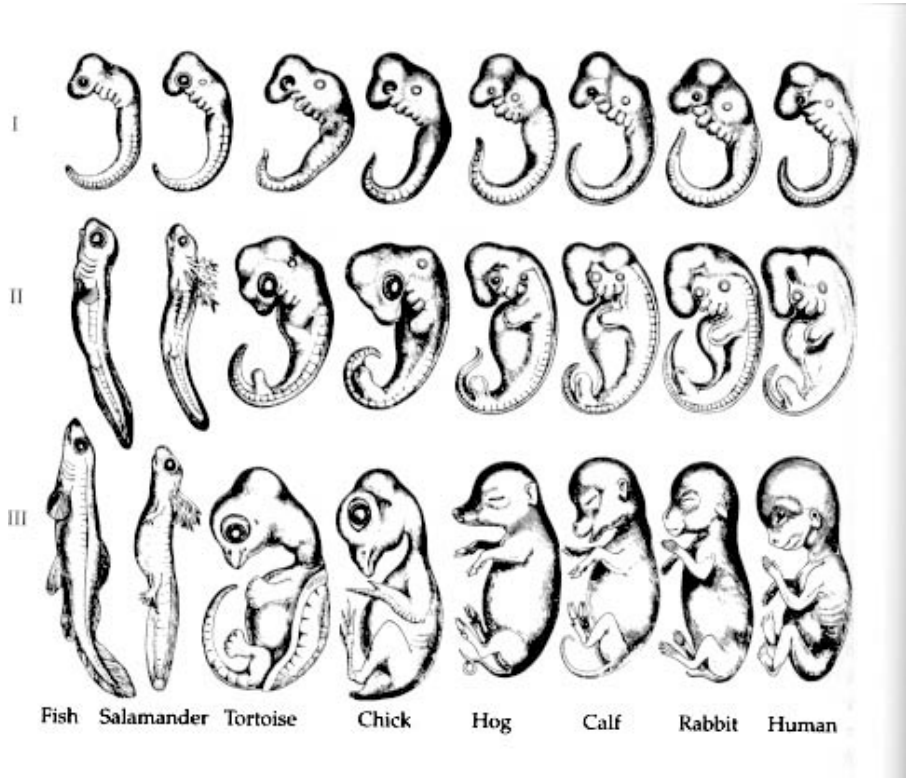
Bgee
BGEE | Gene Expression Evolution

perform high throughput analyses of gene expression patterns.

Requirements:

1- integration of heterogeneous expression data

2- comparison criteria between **anatomies**, **developmental stages**, and **genes**



http://bgee.unil.ch/

Bgee: data**B**ase for **G**ene **E**xpression **E**volution

Bgee
BGEE | Gene Expression Evolution

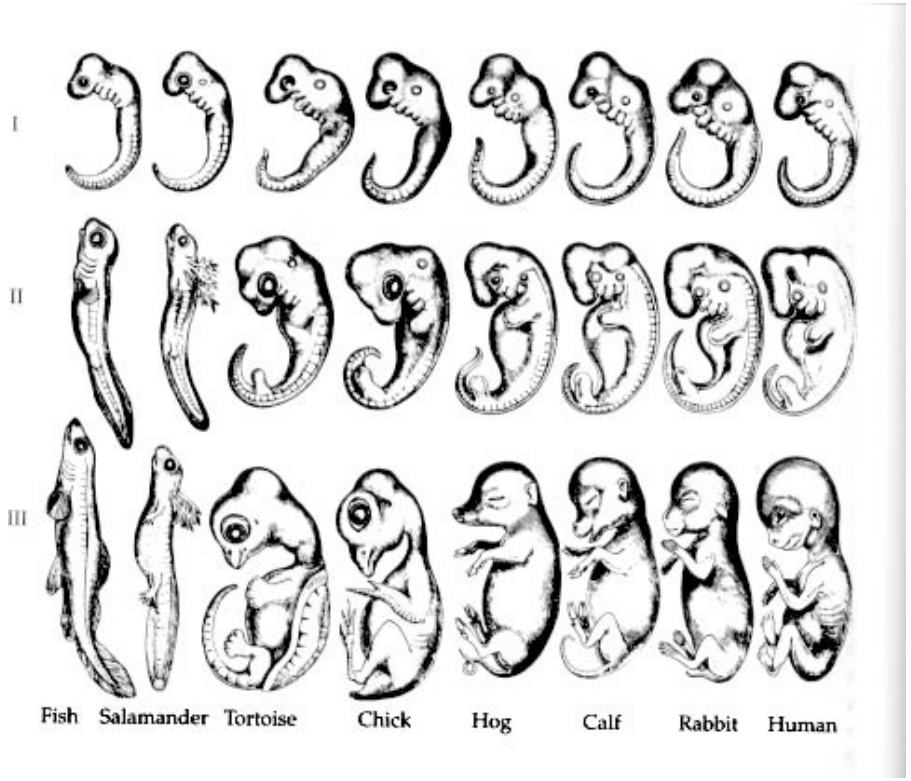
perform high throughput analyses of gene expression patterns.

Requirements:

1- integration of heterogeneous expression data

2- comparison criteria between **anatomies**, **developmental stages**, and **genes**

Orthologs from Ensembl

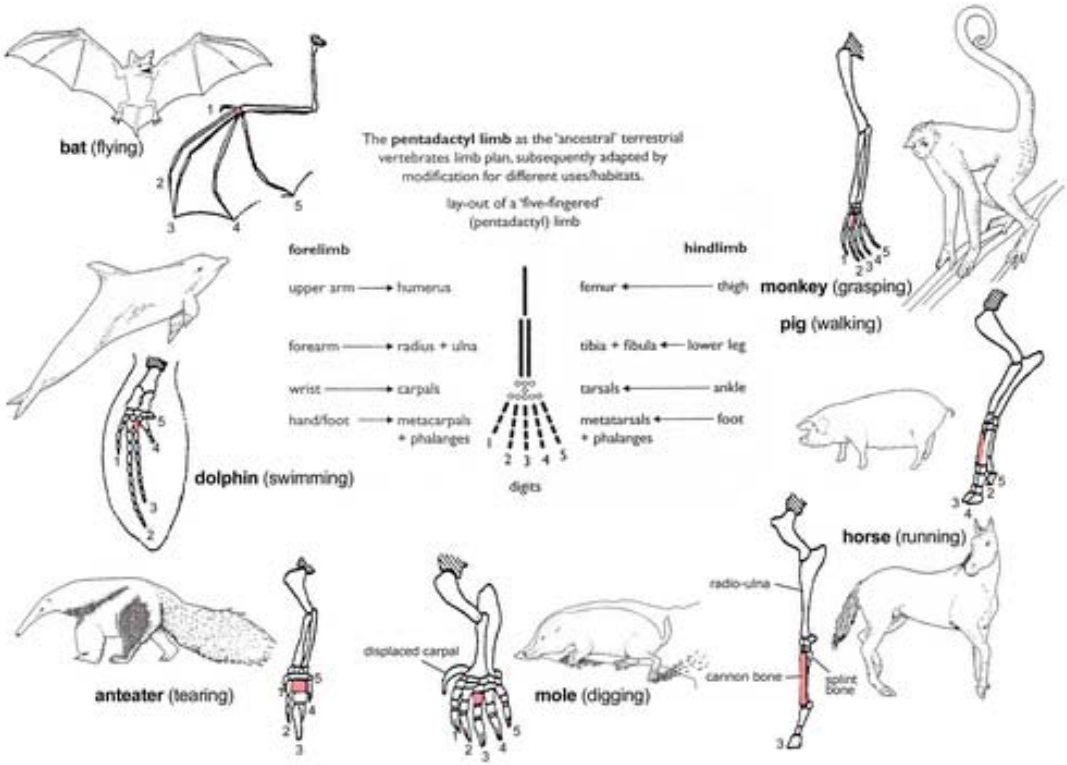


http://bgee.unil.ch/

Homology relationships between anatomies

Evolutionary approach => appropriate comparison criterion: **homology**

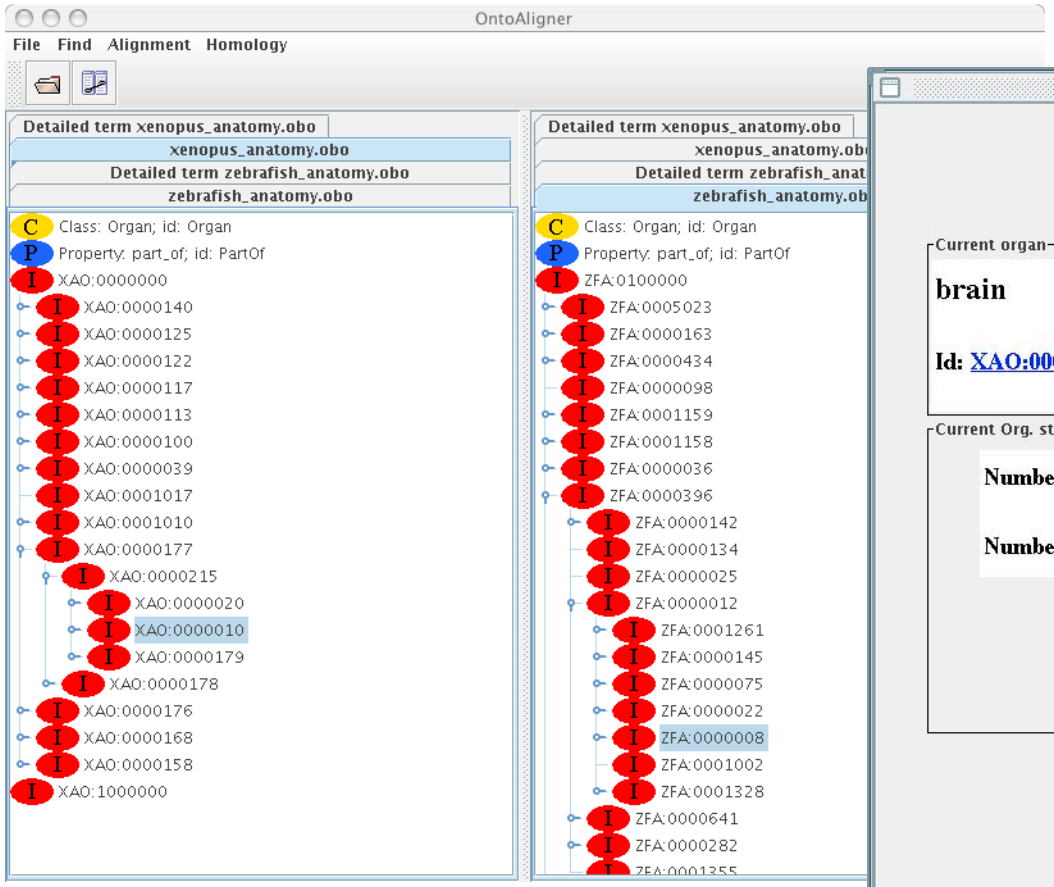
Homology: two anatomical structures within different organisms which originated from a structure of their common ancestral organism.

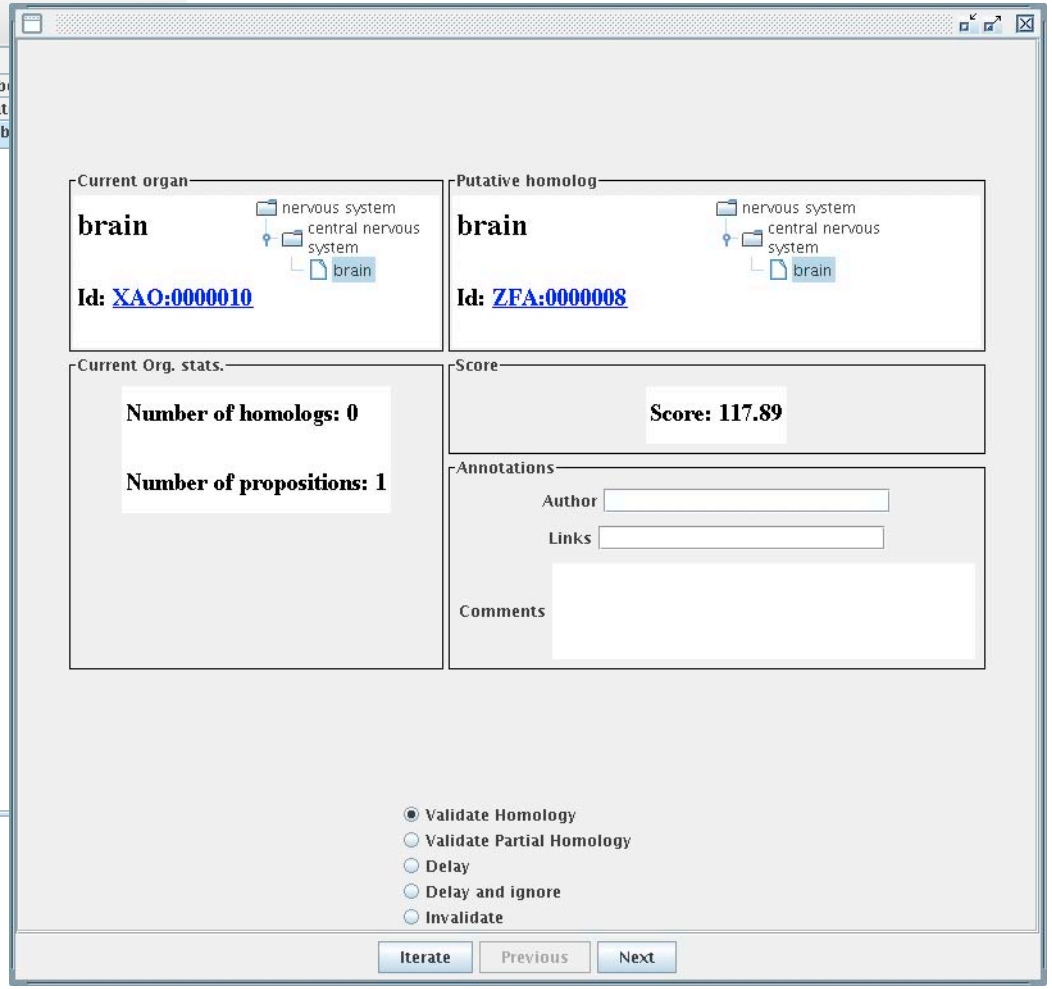


http://bgee.unil.ch/

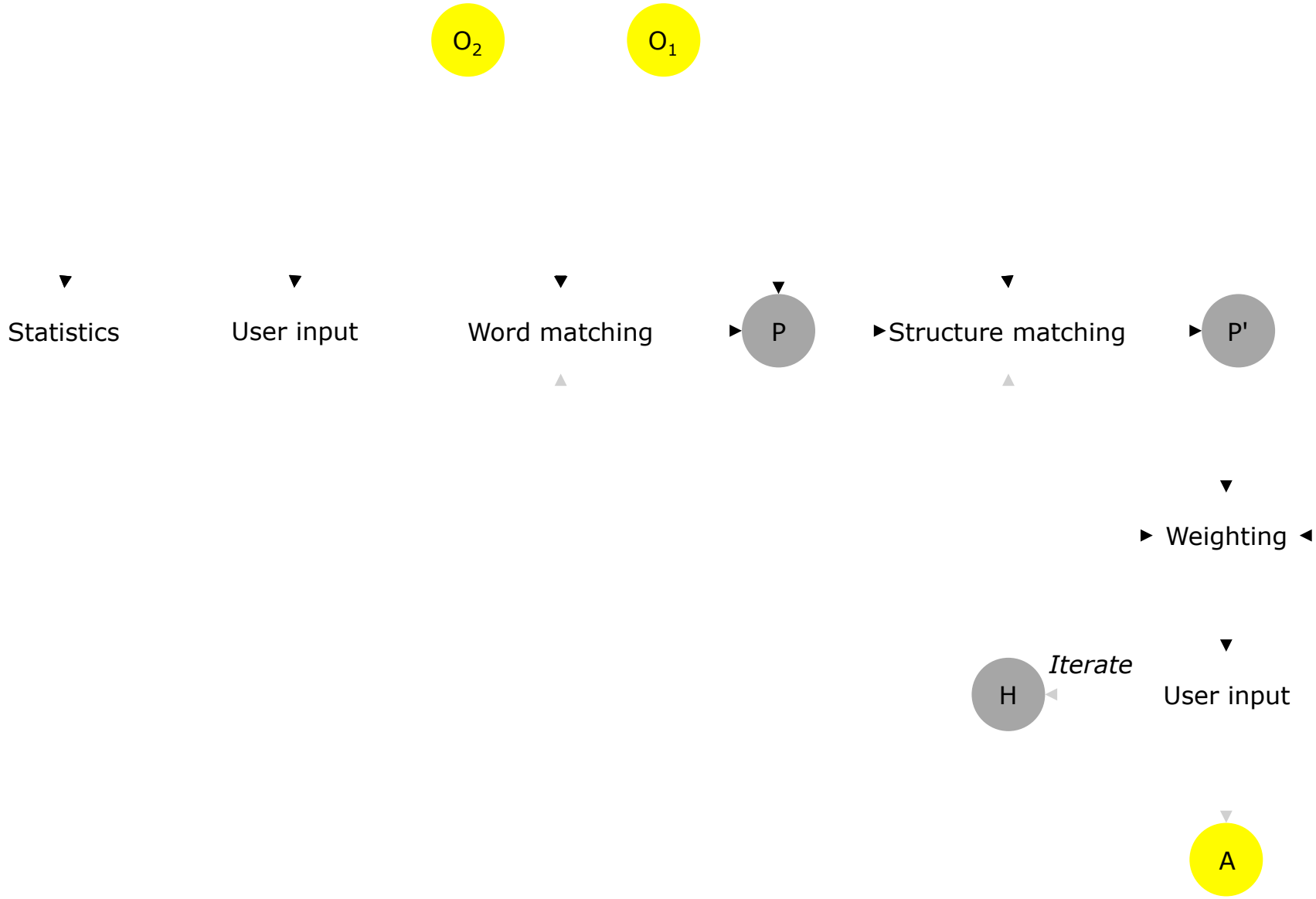
Homolonto

-Software to generate homology relationships
-Pairwise alignments of species-specific anatomical ontologies

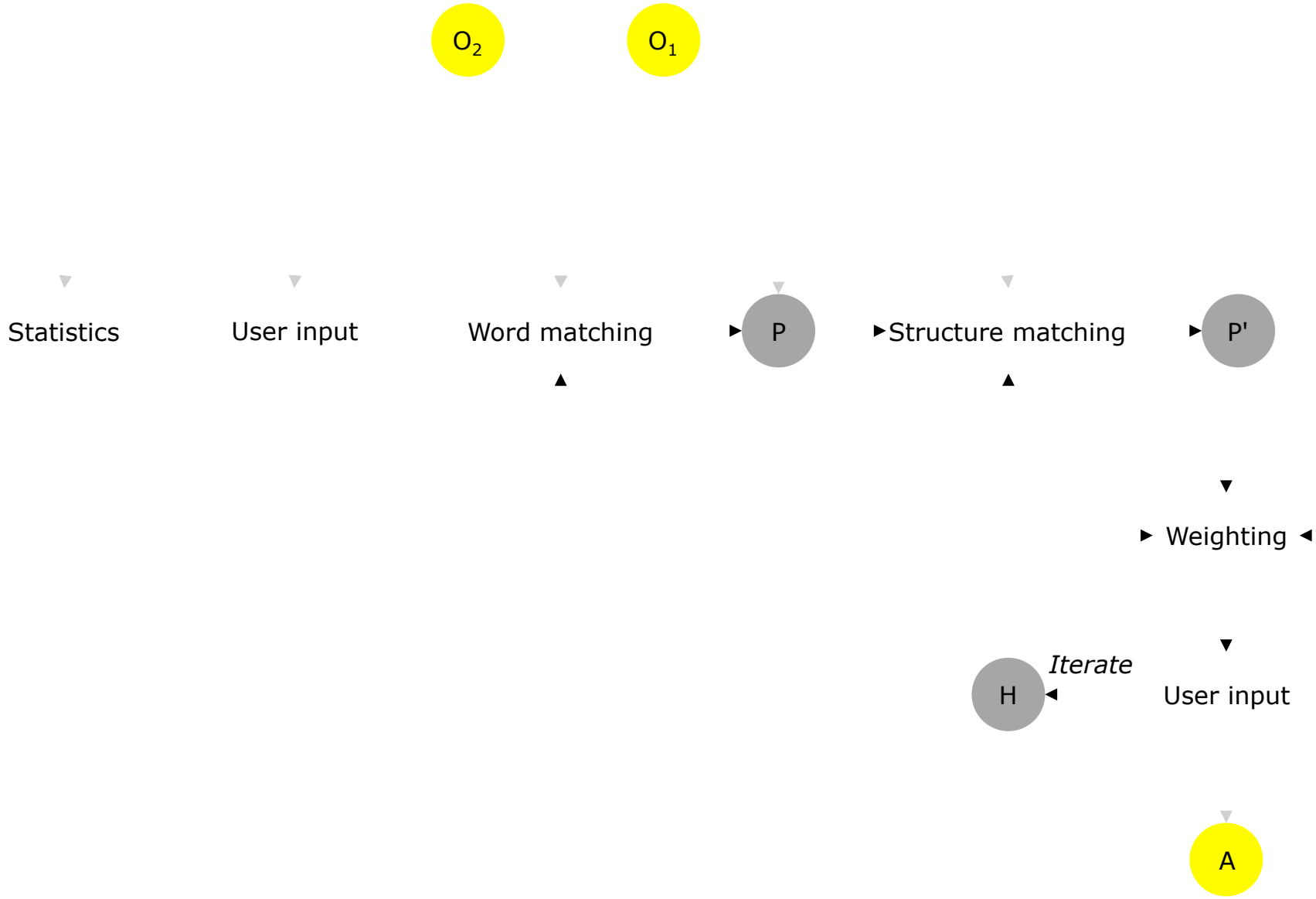




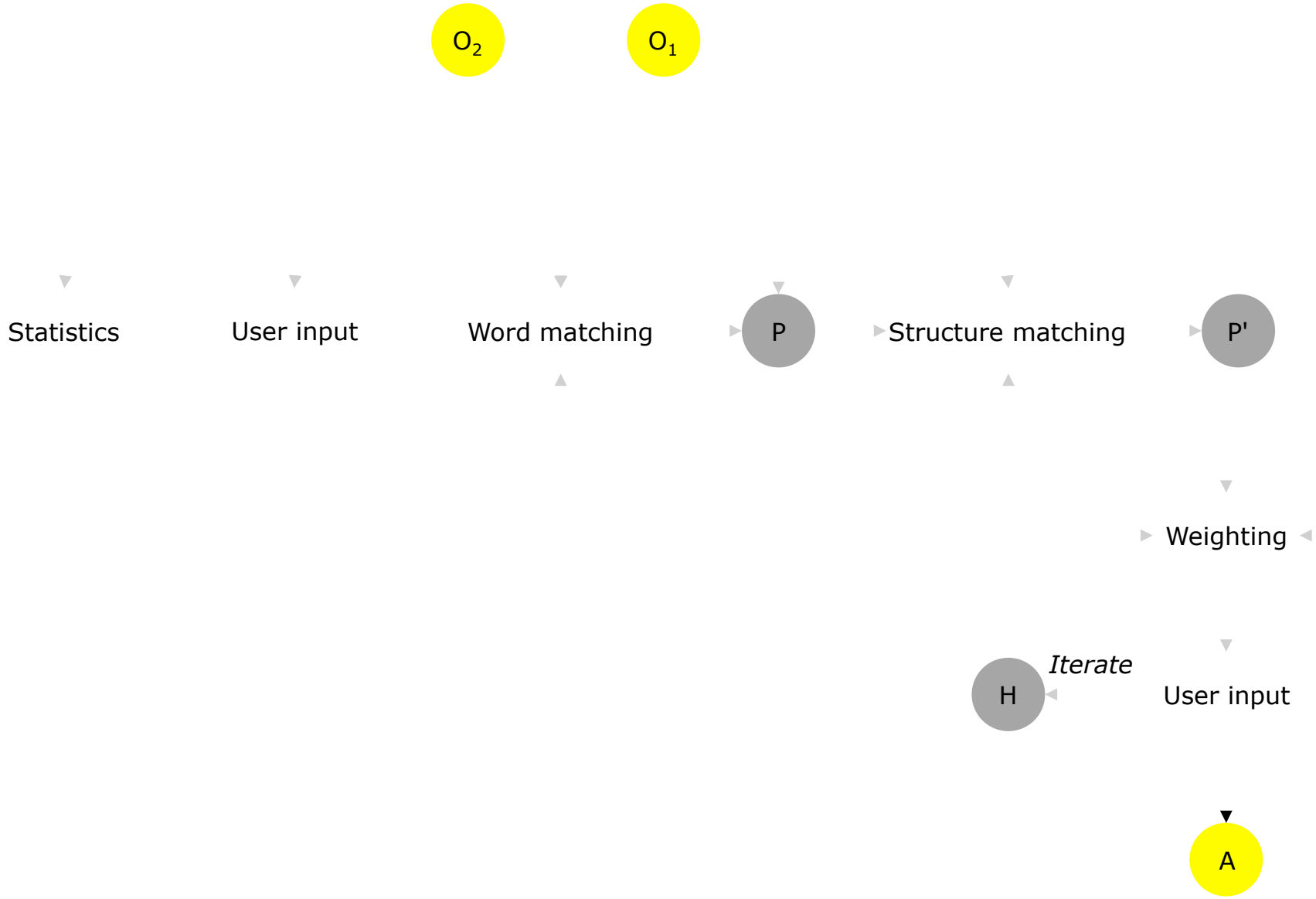
Homolonto


O_2
O_1
Statistics
User input
Word matching
P
Structure matching
P'
Weighting
H
Iterate
User input
A

Homolonto


O_2
O_1
Statistics
User input
Word matching
P
Structure matching
P'
Weighting
H
Iterate
User input
A

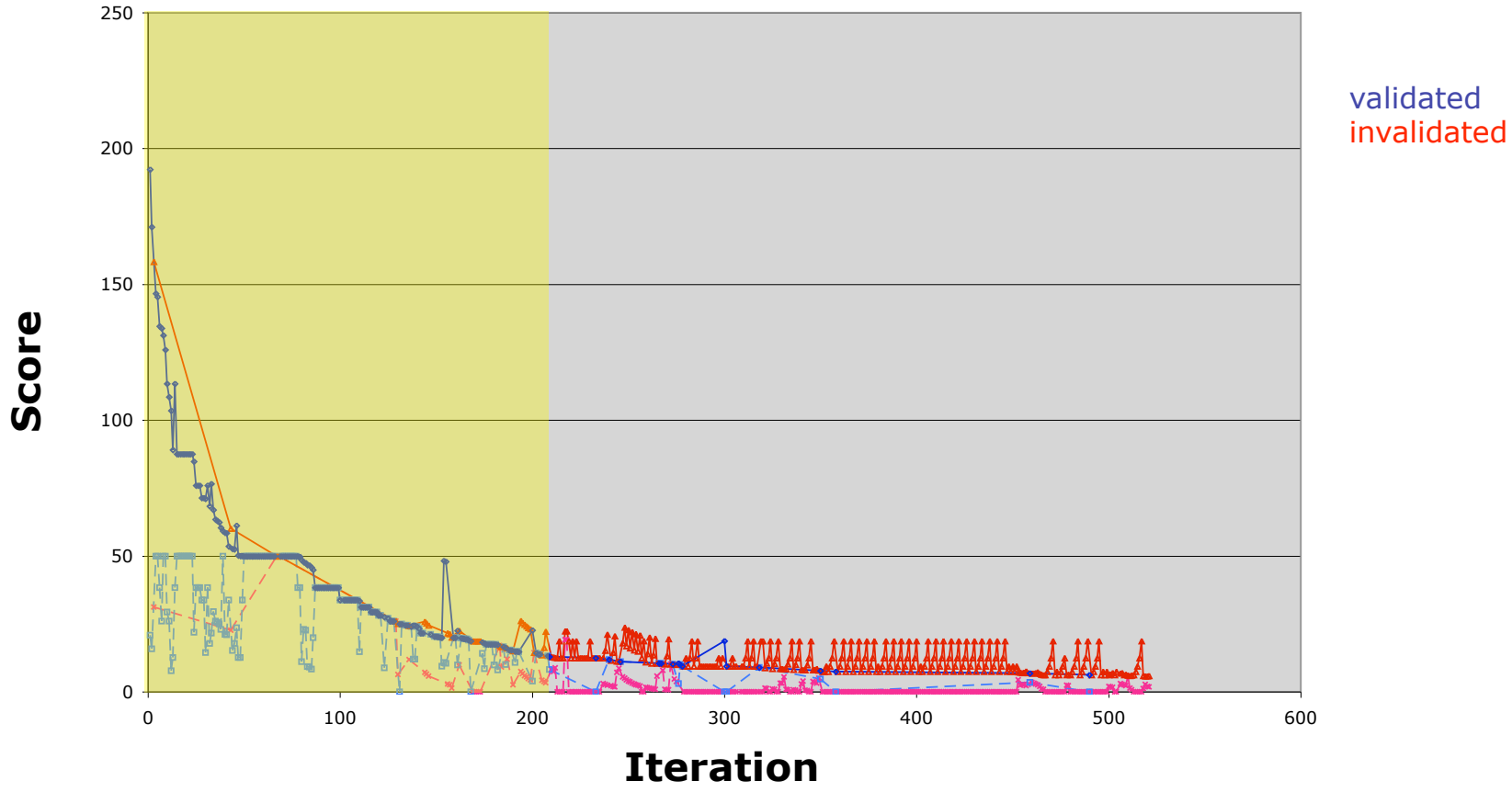
Homolonto


O_2
O_1
Statistics
User input
Word matching
P
Structure matching
P'
Weighting
H
Iterate
User input
A

http://bgee.unil.ch/

Test case: Xenopus-zebrafish ontologies

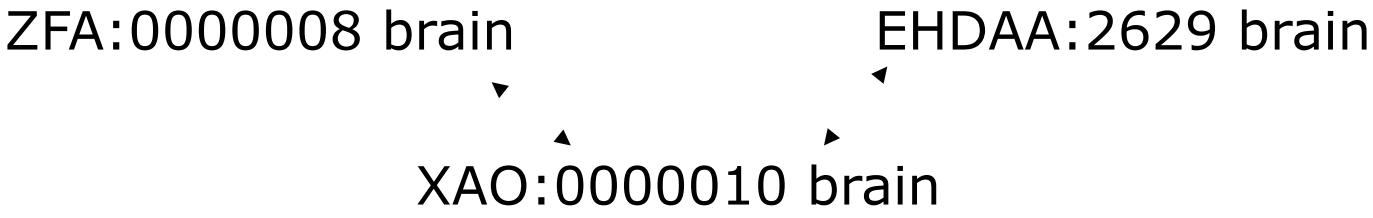
213 first pairs: 80% validated - contains 91% of homologs validated



http://bgee.unil.ch/

Generating a multi-species ontology

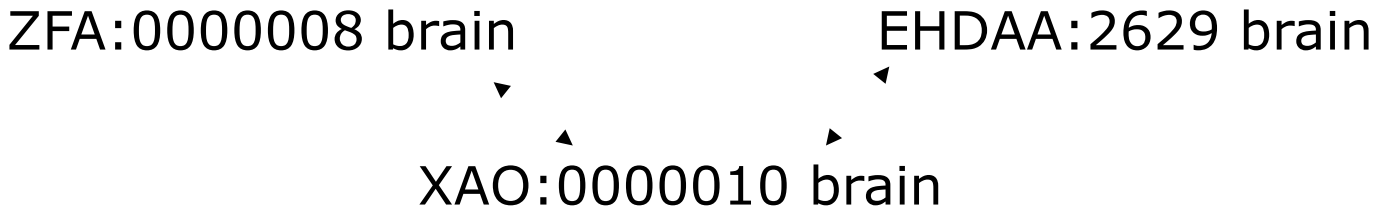
-Homolonto: generates pairwise relationships between ontologies



http://bgee.unil.ch/

Generating a multi-species ontology

-Homolonto: generates pairwise relationships between ontologies



-Merging pairwise alignments: generates groups of homologs

HOG:0000157 brain

ZFA:0000008

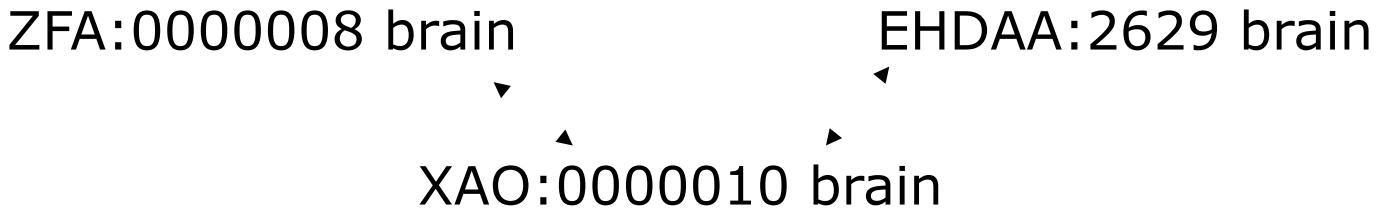
XAO:0000010

EHDAA:2629

http://bgee.unil.ch/

Generating a multi-species ontology

-Homolonto: generates pairwise relationships between ontologies



-Merging pairwise alignments: generates groups of homologs

HOG:0000157 brain

ZFA:0000008

XAO:0000010

EHDAA:2629

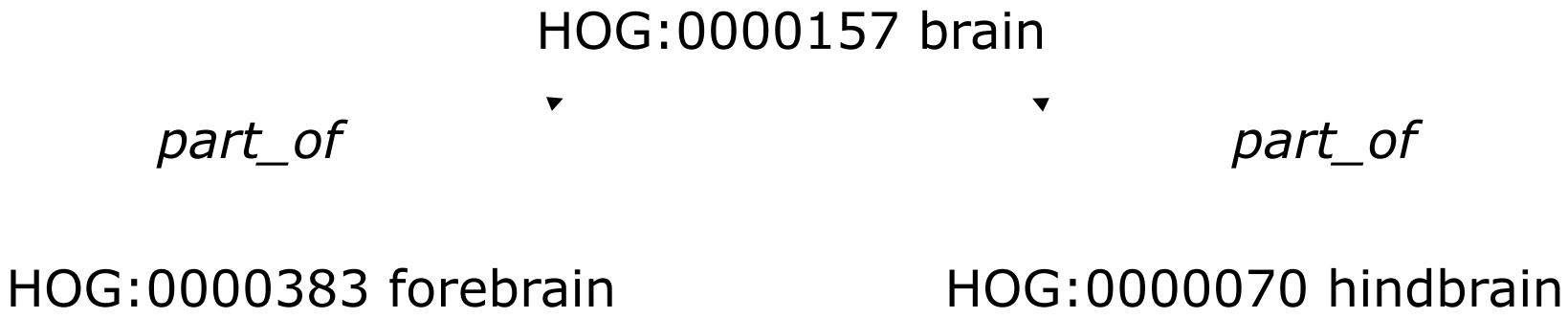
=>List of Homologous Organs Groups (HOGs)

http://bgee.unil.ch/

Generating a multi-species ontology

-HOGs need to be structured as an ontology to allow reasoning

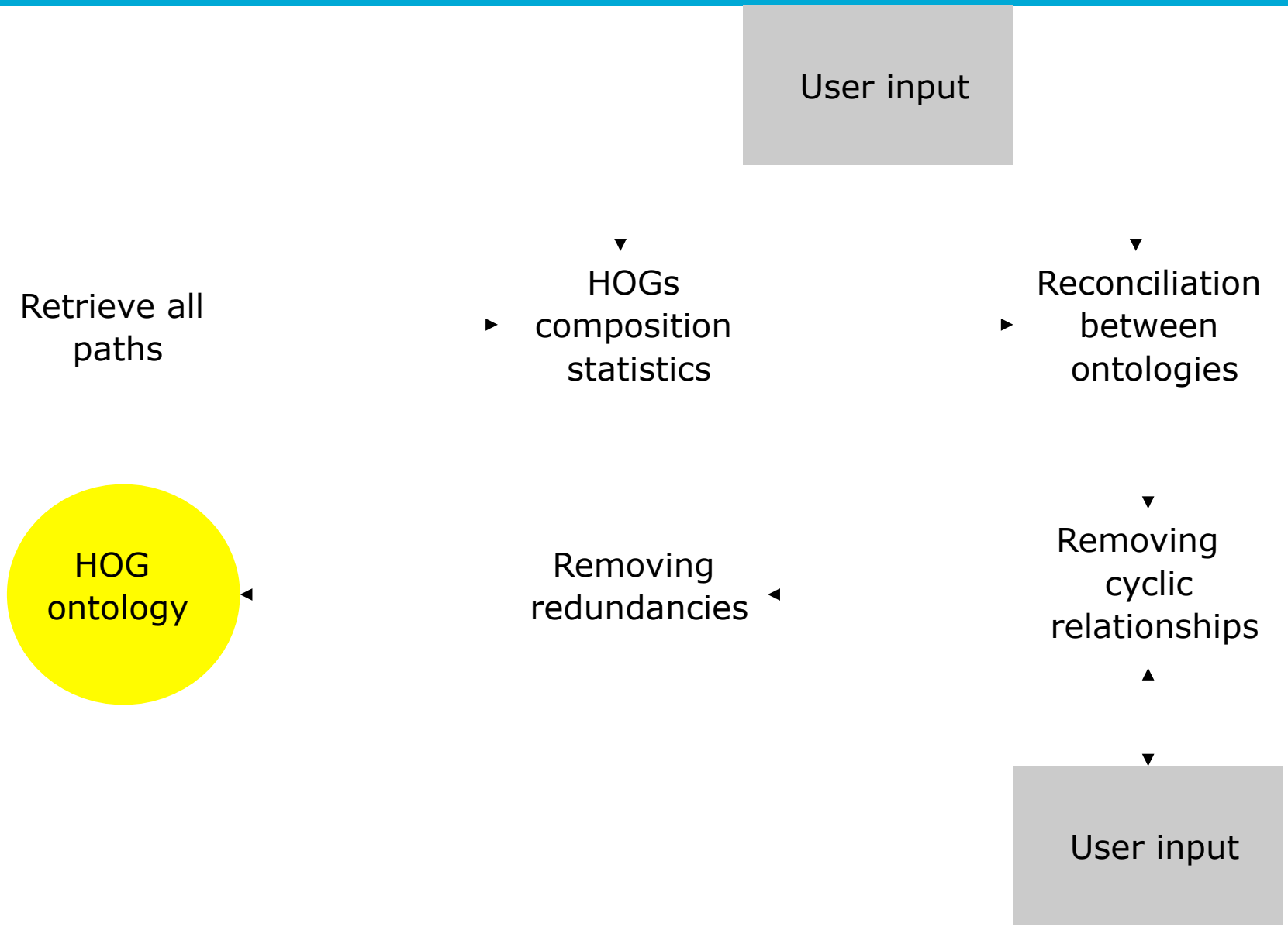
-At a minimum, relationships amongst them have to be designed



=>Algorithm to infer relationships between HOGs

http://bgee.unil.ch/

Inferring relationships between HOGs



http://bgee.unil.ch/

HOGs in Bgee

Use of Homolonto, followed by a curation process:

-4 species: human, mouse, zebrafish, Xenopus

-6 ontologies: ZFA, EHDAA, EV, EMAPA, MA, XAO

HOG ontology in OBO:

-1241 HOGs, 311 with description, 400 with synonyms

-1595 relations, 367 *part_of,* 12 *is_a*, 1216 *broader_than*

External Mapping file:

-involving 5314 anatomical structures

-all manually reviewed, with evidence codes and references

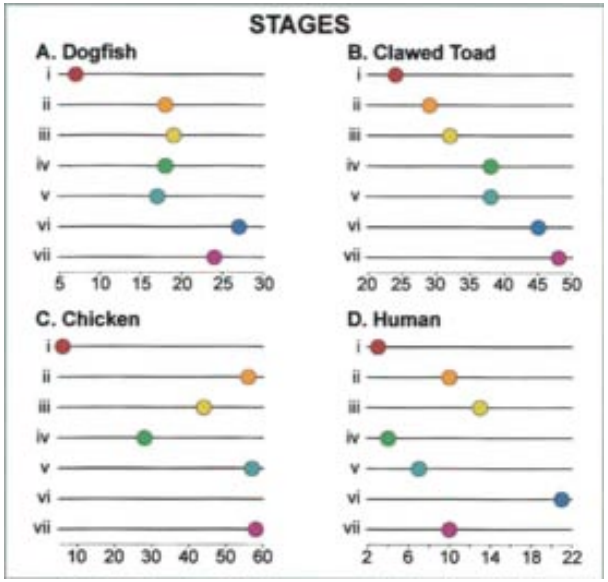
remarks and corrections are welcome!

Mapping of the developmental ontologies

To compare expression patterns, comparisons:

- between homologous organs
- at equivalent stage of development

Heterochrony: impossible to identify homologous stages



(i) First somite
(ii) Nasal placode
(iii) Optic cup
(iv) Heart loop
(v) Thyroid depression
(vi) Spleen anlage
(vii) Forelimb buds

(Jeffery *et al.*, Evol Dev., 2002)

Mapping of the developmental ontologies

Solution:

identify key events of development, common to all vertebrates

=> Ontology of “metastages”

```
All metastages
        Embryo
                Zygote
                Cleavage
                Blastula
                Gastrula
                Neurula
                Organogenesis
        Post-embryonic development
        Adult
```

http://bgee.unil.ch/

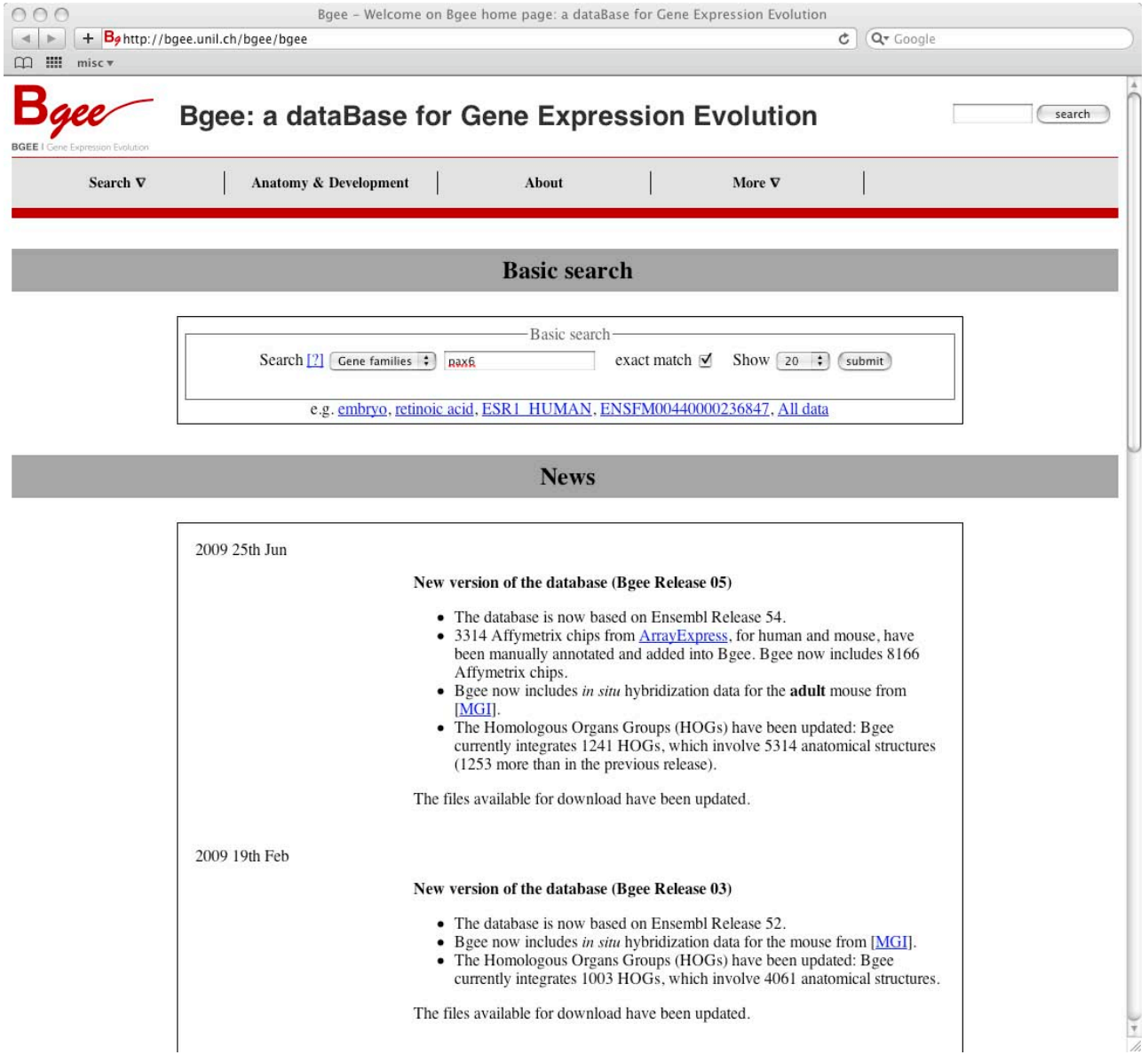
Bgee database

Database statistics

	Zebrafish	Mouse	Human	Xenopus
Ensembl Genes	24,233	31,804	37,435	19,017
Genes with expression data	15,820 65%	20,074 63%	19,354	3,721
Developmental stages	57	40	44	87
Developmental stages with expression data	57 100%	40 100%	19	18
Anatomical structures	2,173	6,408	2,724	645
Anatomical structures with expression data	1,109 51%	2,338 36%	200	52
EST libraries	101	686	2,396	66
Affymetrix chips	52	4,116	3,998	0
In situ figures/evidences	26,709	37,969	0	0

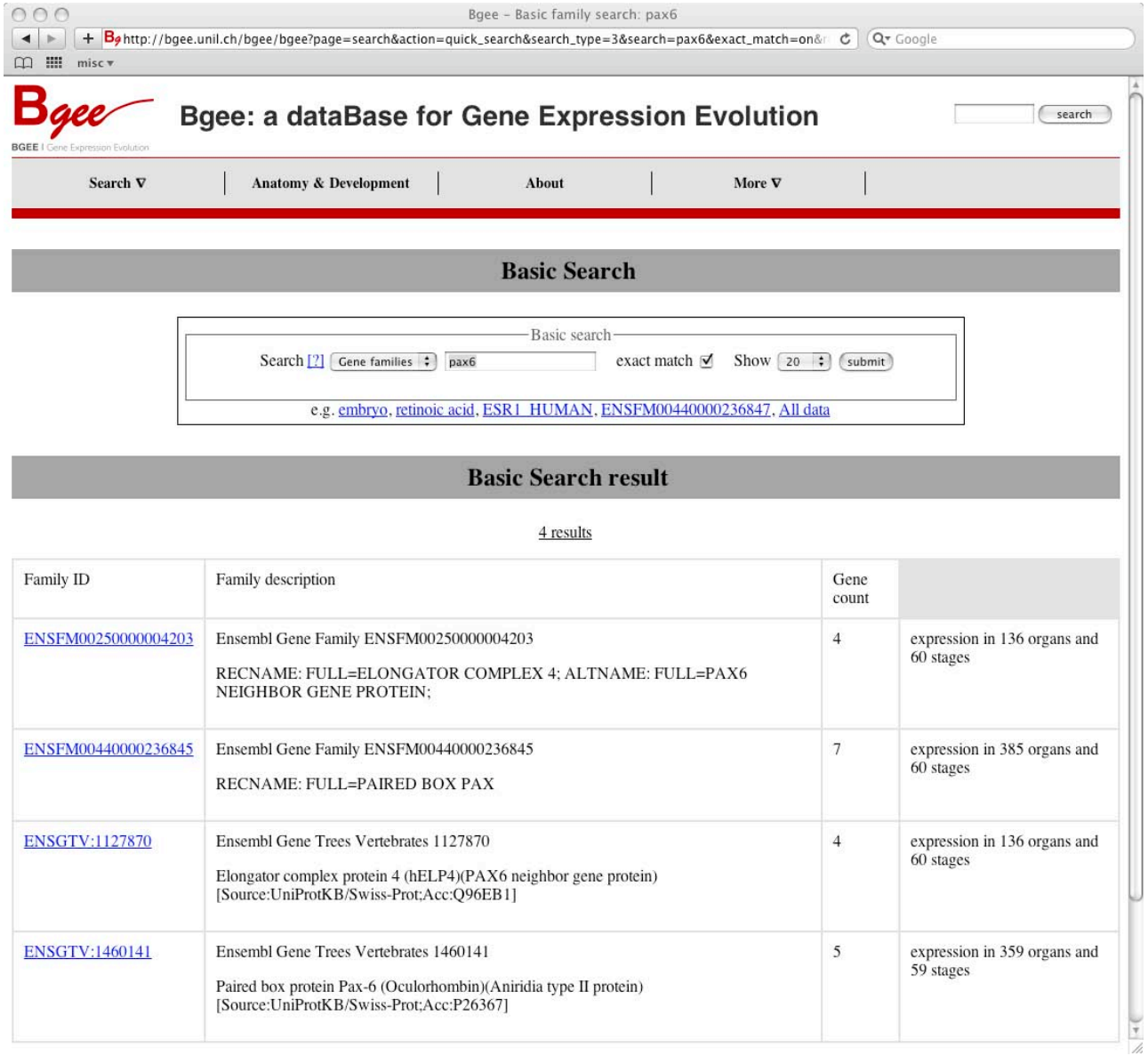
http://bgee.unil.ch/

Example: Pax-6 and the eye



http://bgee.unil.ch/

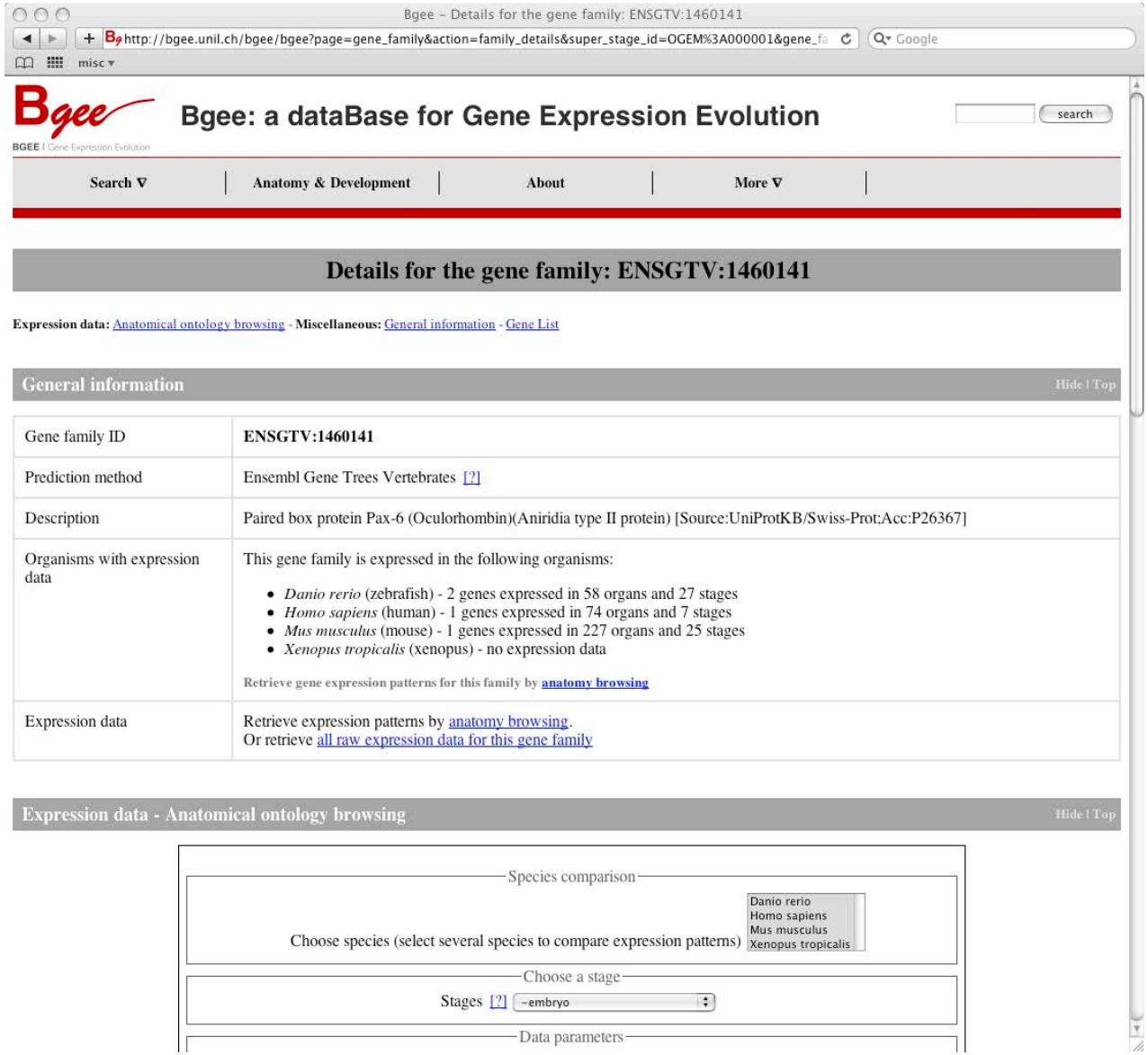
Example: Pax-6 and the eye



Family ID	Family description	Gene count	
ENSFM00250000004203	Ensembl Gene Family ENSFM00250000004203 RECNAME: FULL=ELONGATOR COMPLEX 4; ALTNAME: FULL=PAX6 NEIGHBOR GENE PROTEIN;	4	expression in 136 organs and 60 stages
ENSFM00440000236845	Ensembl Gene Family ENSFM00440000236845 RECNAME: FULL=PAIRED BOX PAX	7	expression in 385 organs and 60 stages
ENSGTV:1127870	Ensembl Gene Trees Vertebrates 1127870 Elongator complex protein 4 (hELP4)(PAX6 neighbor gene protein) [Source:UniProtKB/Swiss-Prot;Acc:Q96EB1]	4	expression in 136 organs and 60 stages
ENSGTV:1460141	Ensembl Gene Trees Vertebrates 1460141 Paired box protein Pax-6 (Oculorhombin)(Aniridia type II protein) [Source:UniProtKB/Swiss-Prot;Acc:P26367]	5	expression in 359 organs and 59 stages

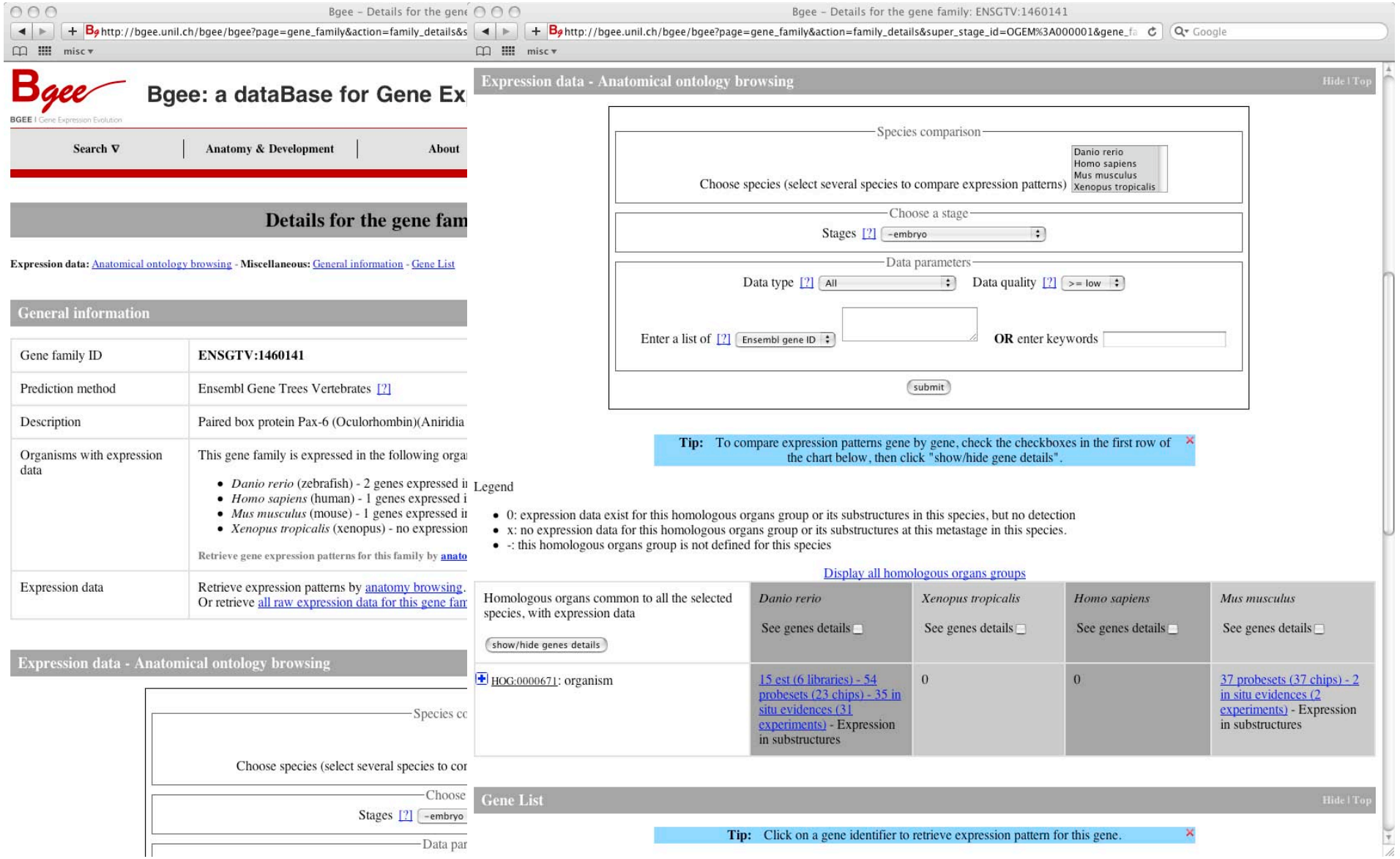
http://bgee.unil.ch/

Example: Pax-6 and the eye



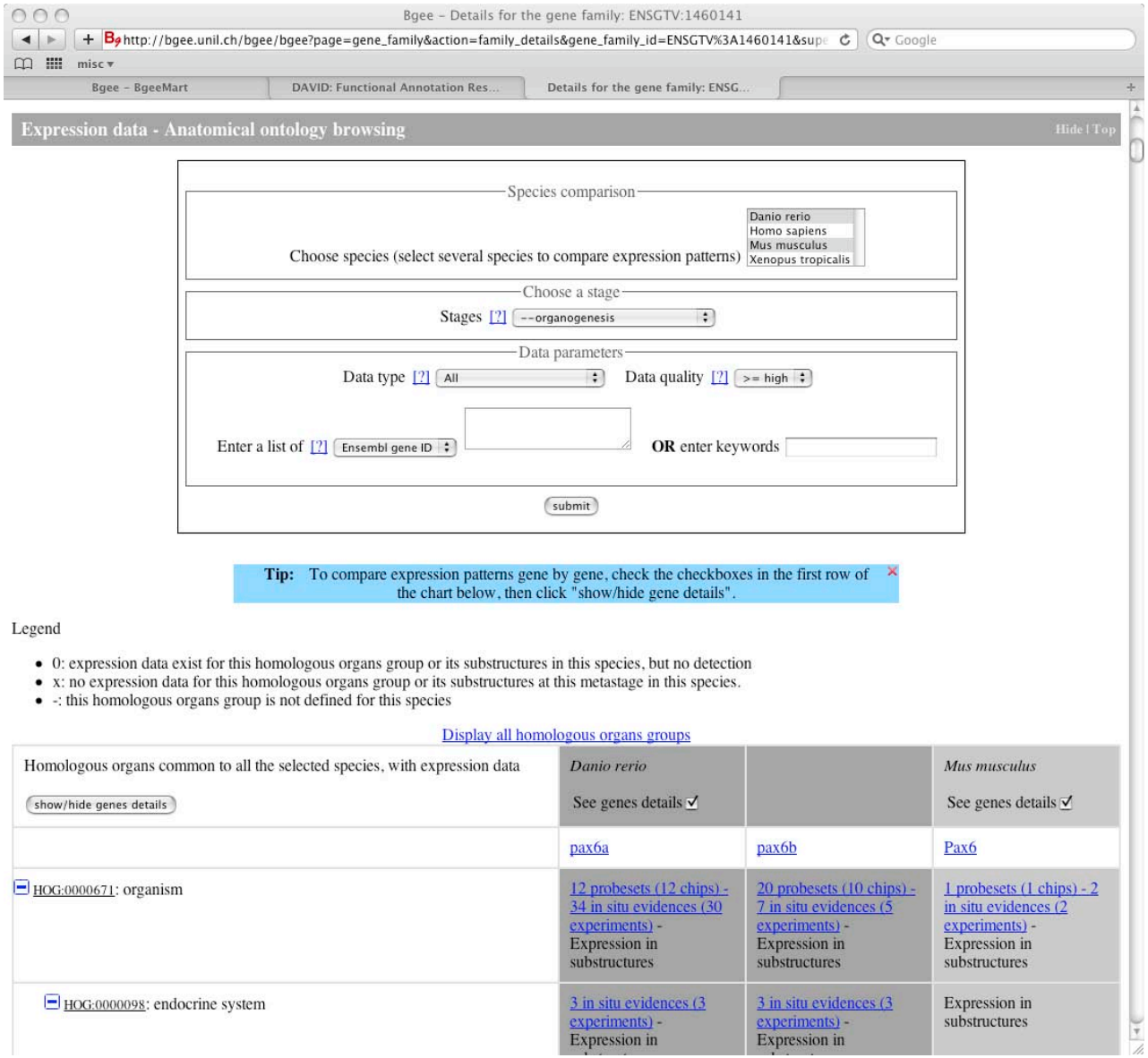
http://bgee.unil.ch/

Example: Pax-6 and the eye



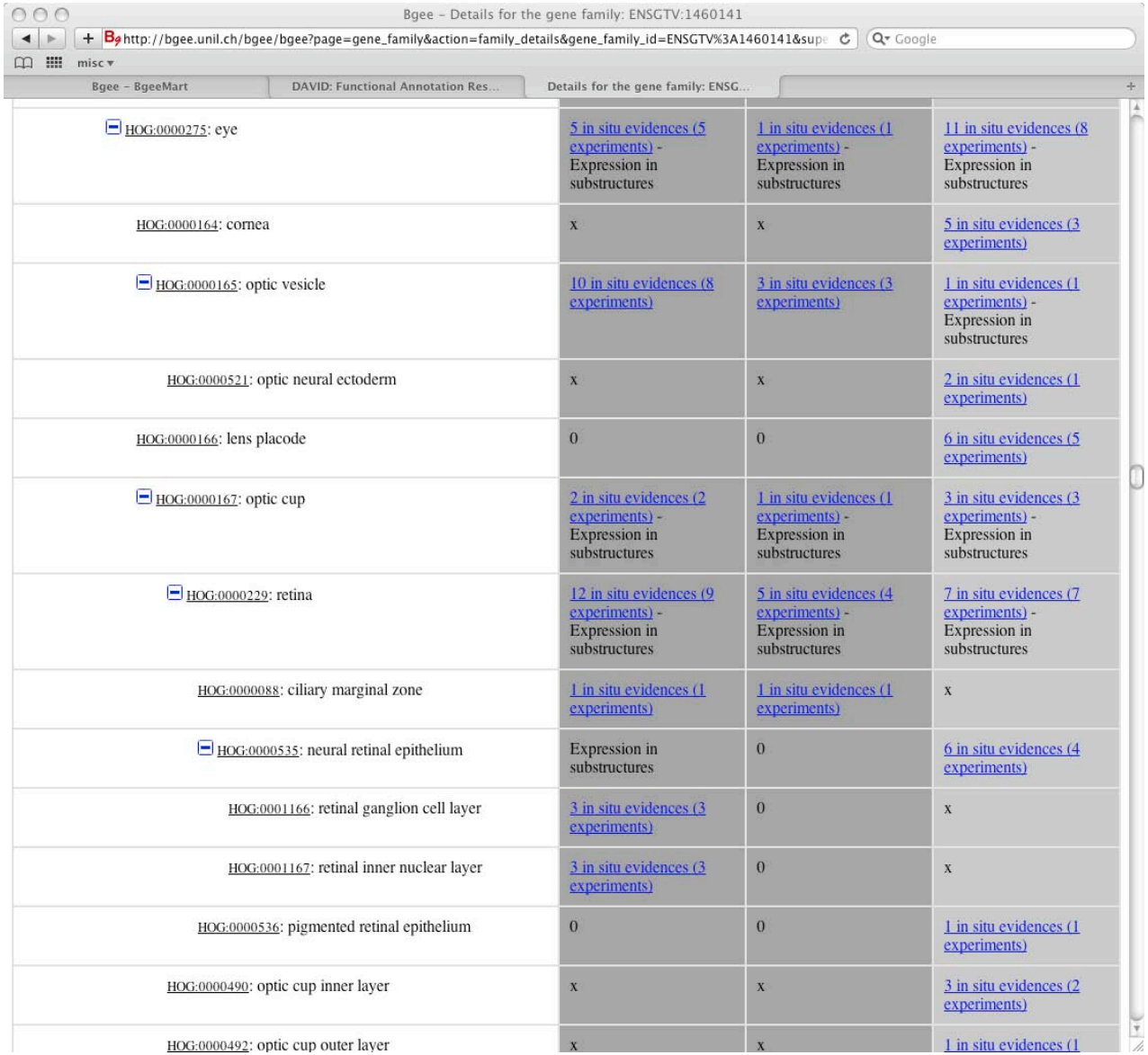
http://bgee.unil.ch/

Example: Pax-6 and the eye



http://bgee.unil.ch/

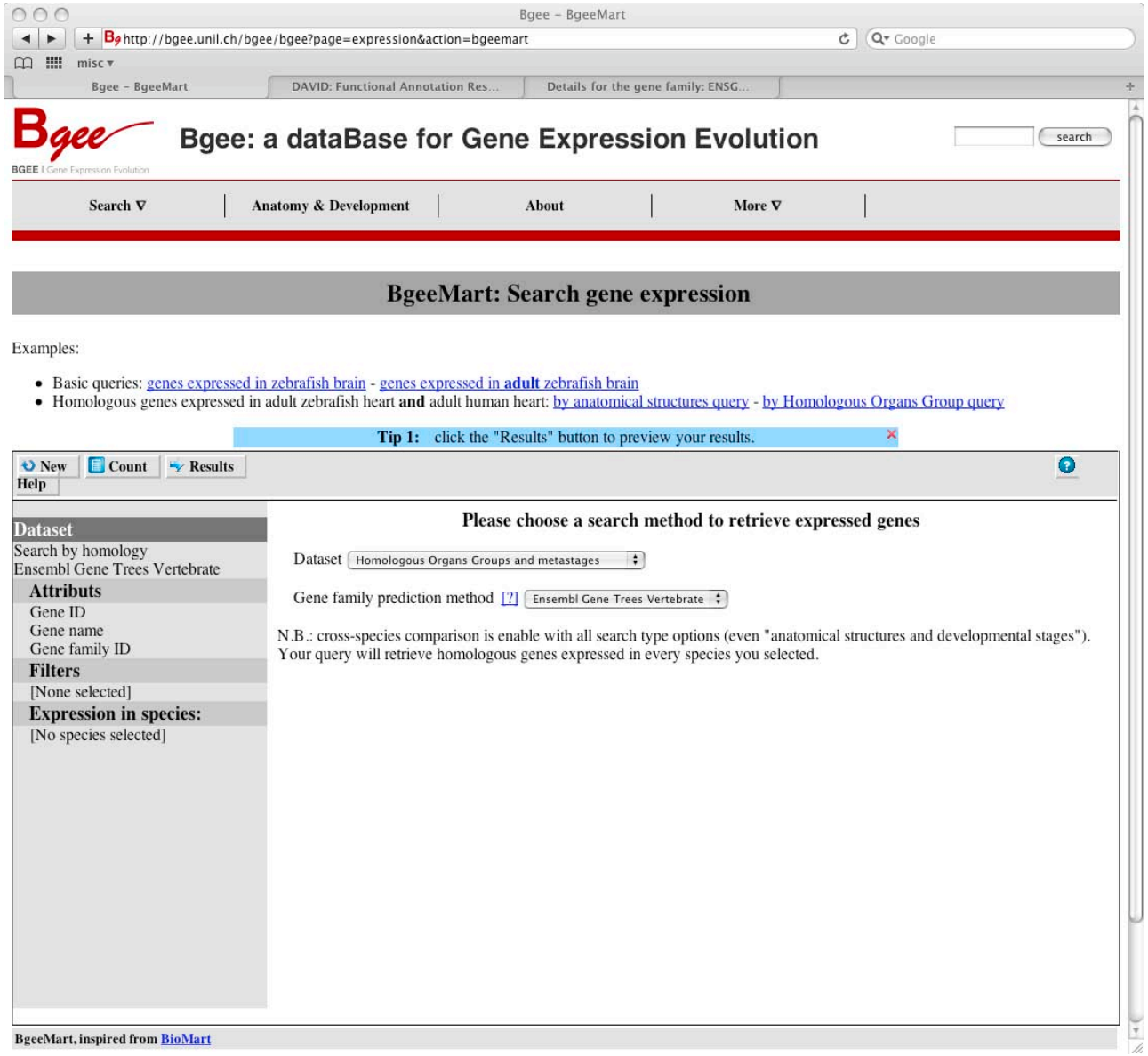
Example: Pax-6 and the eye



HOG:0000275: eye	5 in situ evidences (5 experiments) - Expression in substructures	1 in situ evidences (1 experiments) - Expression in substructures	11 in situ evidences (8 experiments) - Expression in substructures
HOG:0000164: cornea	x	x	5 in situ evidences (3 experiments)
HOG:0000165: optic vesicle	10 in situ evidences (8 experiments)	3 in situ evidences (3 experiments)	1 in situ evidences (1 experiments) - Expression in substructures
HOG:0000521: optic neural ectoderm	x	x	2 in situ evidences (1 experiments)
HOG:0000166: lens placode	0	0	6 in situ evidences (5 experiments)
HOG:0000167: optic cup	2 in situ evidences (2 experiments) - Expression in substructures	1 in situ evidences (1 experiments) - Expression in substructures	3 in situ evidences (3 experiments) - Expression in substructures
HOG:0000229: retina	12 in situ evidences (9 experiments) - Expression in substructures	5 in situ evidences (4 experiments) - Expression in substructures	7 in situ evidences (7 experiments) - Expression in substructures
HOG:0000088: ciliary marginal zone	1 in situ evidences (1 experiments)	1 in situ evidences (1 experiments)	x
HOG:0000535: neural retinal epithelium	Expression in substructures	0	6 in situ evidences (4 experiments)
HOG:0001166: retinal ganglion cell layer	3 in situ evidences (3 experiments)	0	x
HOG:0001167: retinal inner nuclear layer	3 in situ evidences (3 experiments)	0	x
HOG:0000536: pigmented retinal epithelium	0	0	1 in situ evidences (1 experiments)
HOG:0000490: optic cup inner layer	x	x	3 in situ evidences (2 experiments)
HOG:0000492: optic cup outer layer	x	x	1 in situ evidences (1

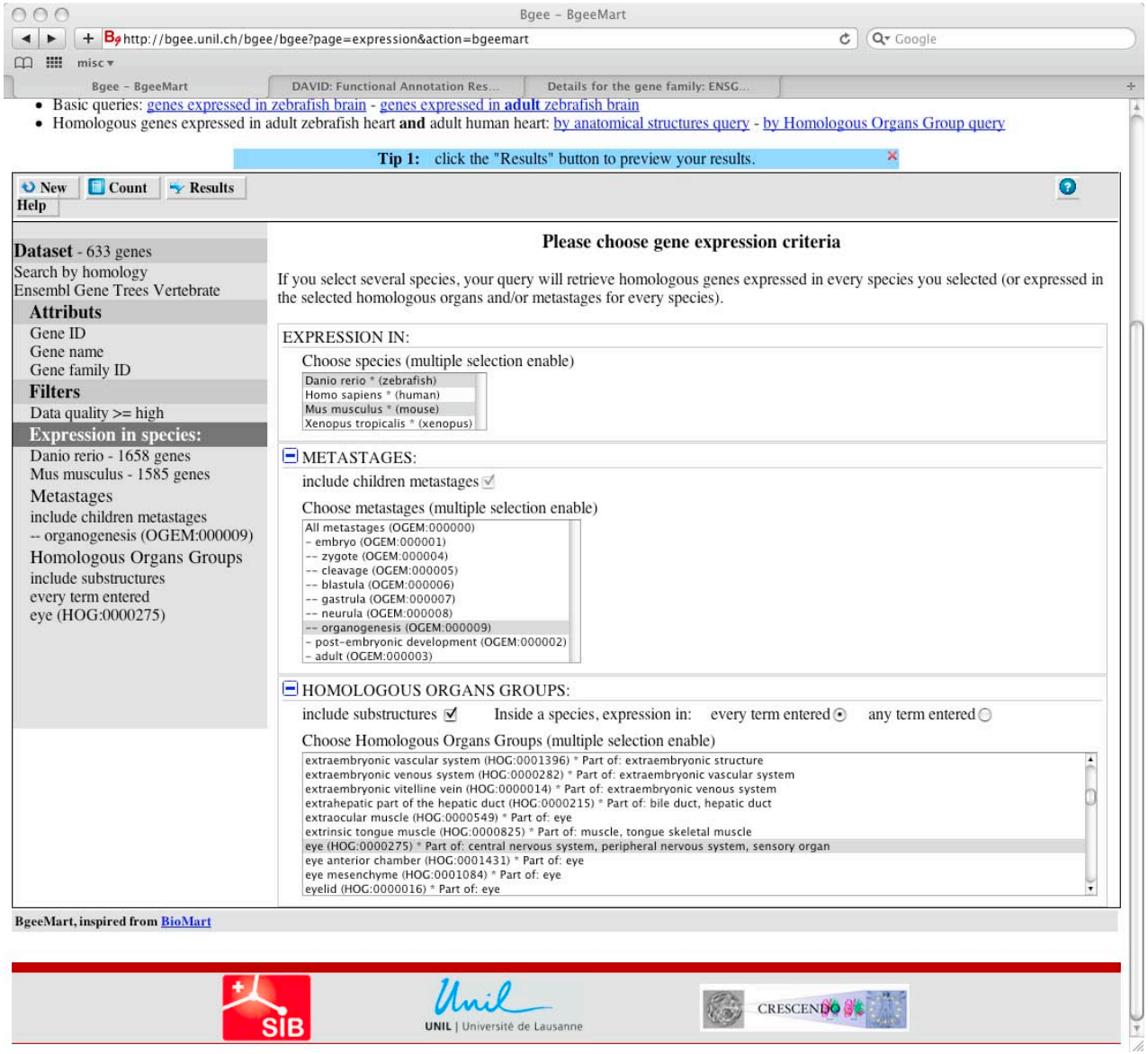
http://bgee.unil.ch/

Example: Pax-6 and the eye



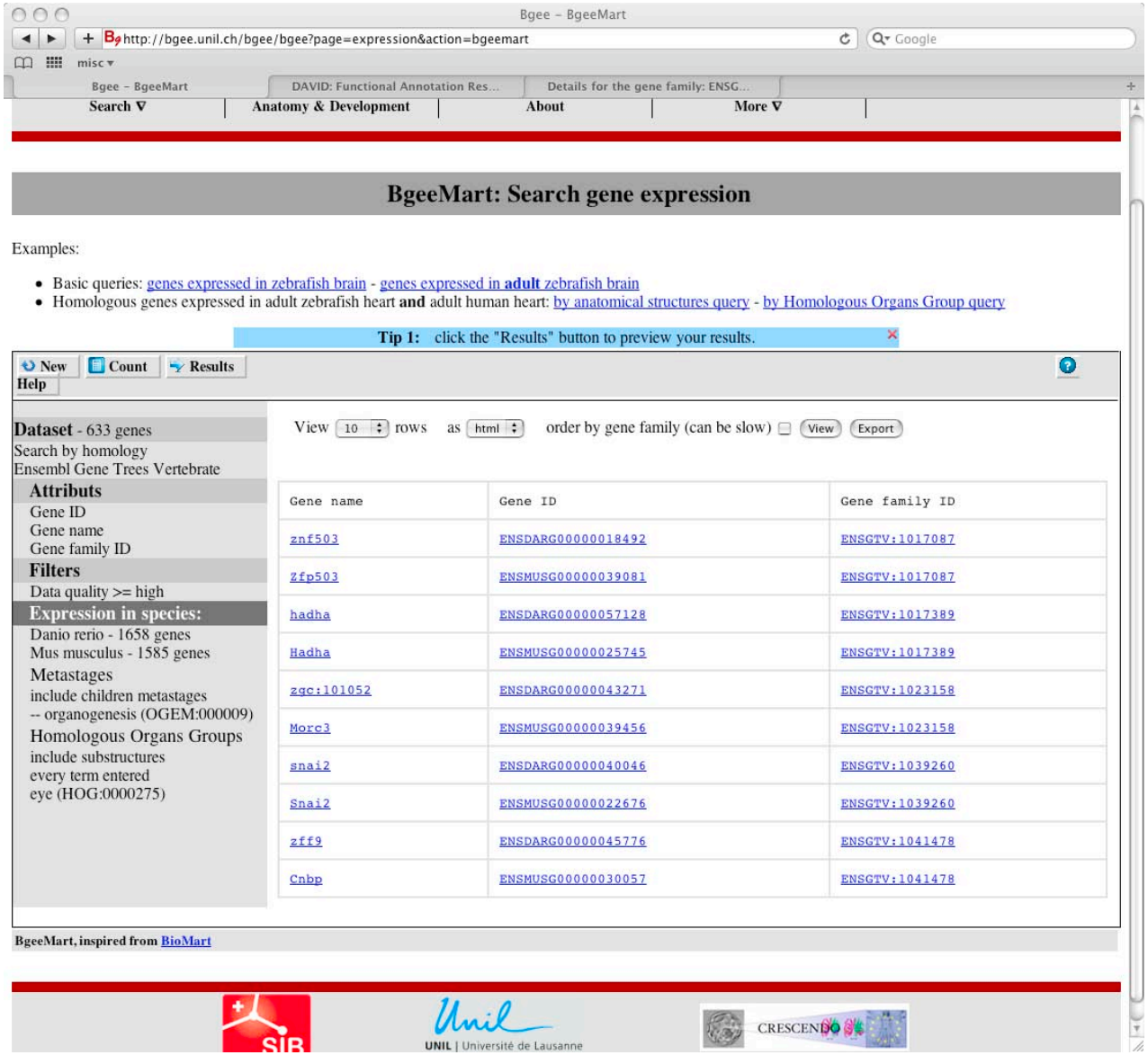
http://bgee.unil.ch/

Example: Pax-6 and the eye



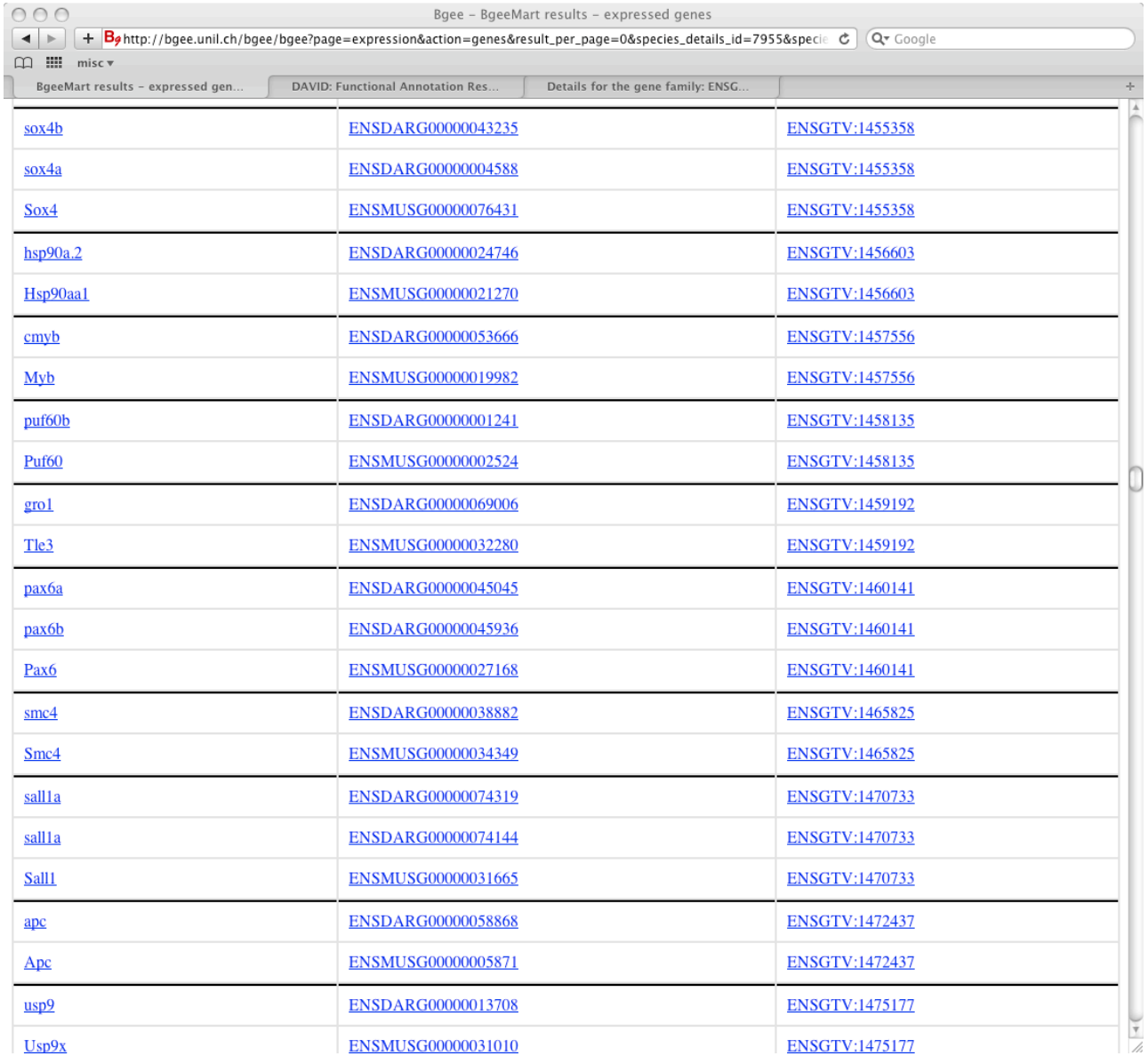
http://bgee.unil.ch/

Example: Pax-6 and the eye



http://bgee.unil.ch/

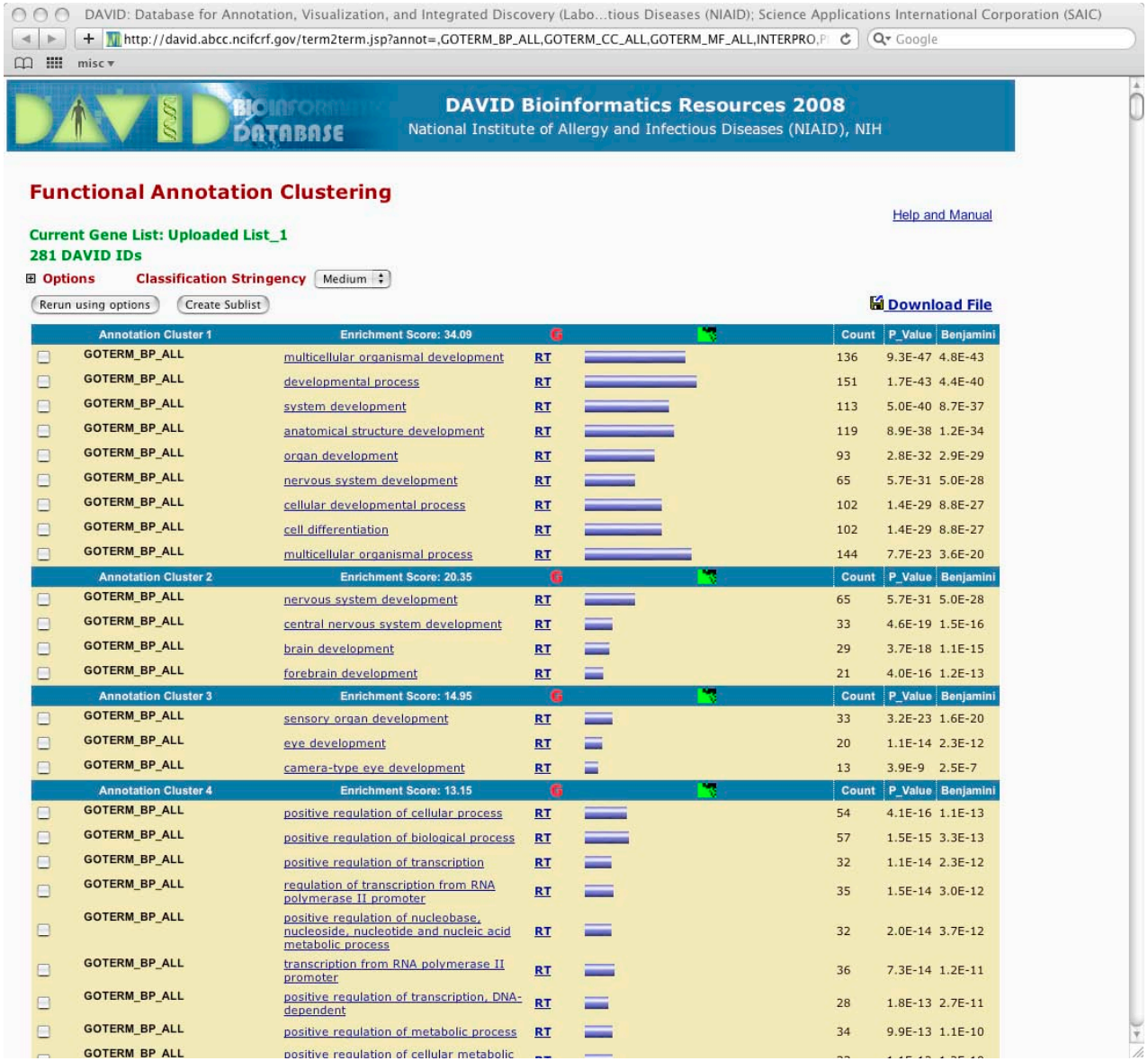
Example: Pax-6 and the eye



sox4b	ENSDARG00000043235	ENSGTV:1455358
sox4a	ENSDARG00000004588	ENSGTV:1455358
Sox4	ENSMUSG00000076431	ENSGTV:1455358
hsp90a.2	ENSDARG00000024746	ENSGTV:1456603
Hsp90aa1	ENSMUSG00000021270	ENSGTV:1456603
cmyb	ENSDARG00000053666	ENSGTV:1457556
Myb	ENSMUSG00000019982	ENSGTV:1457556
puf60b	ENSDARG00000001241	ENSGTV:1458135
Puf60	ENSMUSG00000002524	ENSGTV:1458135
gro1	ENSDARG00000069006	ENSGTV:1459192
Tle3	ENSMUSG00000032280	ENSGTV:1459192
pax6a	ENSDARG00000045045	ENSGTV:1460141
pax6b	ENSDARG00000045936	ENSGTV:1460141
Pax6	ENSMUSG00000027168	ENSGTV:1460141
smc4	ENSDARG00000038882	ENSGTV:1465825
Smc4	ENSMUSG00000034349	ENSGTV:1465825
sall1a	ENSDARG00000074319	ENSGTV:1470733
sall1a	ENSDARG00000074144	ENSGTV:1470733
Sall1	ENSMUSG00000031665	ENSGTV:1470733
apc	ENSDARG00000058868	ENSGTV:1472437
Apc	ENSMUSG00000005871	ENSGTV:1472437
usp9	ENSDARG00000013708	ENSGTV:1475177
Usp9x	ENSMUSG00000031010	ENSGTV:1475177

http://bgee.unil.ch/

Example: Pax-6 and the eye



DAVID: Database for Annotation, Visualization, and Integrated Discovery (Labo...tious Diseases (NIAID); Science Applications International Corporation (SAIC)

http://david.abcc.ncifcrf.gov/term2term.jsp?annot=,GOTERM_BP_ALL,GOTERM_CC_ALL,GOTERM_MF_ALL,INTERPRO,P

DAVID Bioinformatics Resources 2008
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manual

Current Gene List: Uploaded List_1
281 DAVID IDs

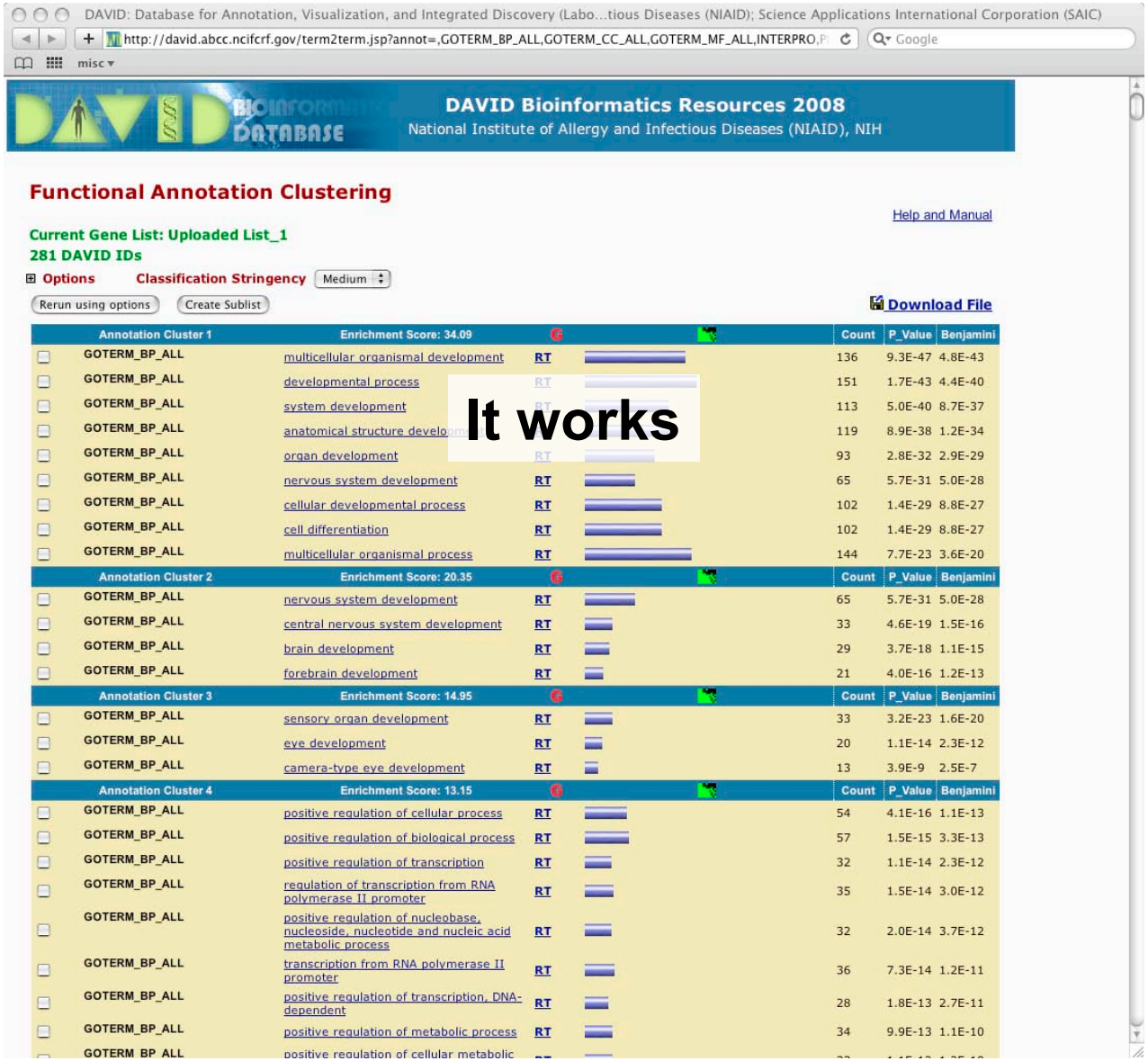
Options Classification Stringency: Medium

Rerun using options | Create Sublist | Download File

					Count	P_Value	Benjamini
Annotation Cluster 1	Enrichment Score: 34.09						
GOTERM_BP_ALL	multicellular organismal development	RT			136	9.3E-47	4.8E-43
GOTERM_BP_ALL	developmental process	RT			151	1.7E-43	4.4E-40
GOTERM_BP_ALL	system development	RT			113	5.0E-40	8.7E-37
GOTERM_BP_ALL	anatomical structure development	RT			119	8.9E-38	1.2E-34
GOTERM_BP_ALL	organ development	RT			93	2.8E-32	2.9E-29
GOTERM_BP_ALL	nervous system development	RT			65	5.7E-31	5.0E-28
GOTERM_BP_ALL	cellular developmental process	RT			102	1.4E-29	8.8E-27
GOTERM_BP_ALL	cell differentiation	RT			102	1.4E-29	8.8E-27
GOTERM_BP_ALL	multicellular organismal process	RT			144	7.7E-23	3.6E-20
Annotation Cluster 2	Enrichment Score: 20.35				Count	P_Value	Benjamini
GOTERM_BP_ALL	nervous system development	RT			65	5.7E-31	5.0E-28
GOTERM_BP_ALL	central nervous system development	RT			33	4.6E-19	1.5E-16
GOTERM_BP_ALL	brain development	RT			29	3.7E-18	1.1E-15
GOTERM_BP_ALL	forebrain development	RT			21	4.0E-16	1.2E-13
Annotation Cluster 3	Enrichment Score: 14.95				Count	P_Value	Benjamini
GOTERM_BP_ALL	sensory organ development	RT			33	3.2E-23	1.6E-20
GOTERM_BP_ALL	eye development	RT			20	1.1E-14	2.3E-12
GOTERM_BP_ALL	camera-type eye development	RT			13	3.9E-9	2.5E-7
Annotation Cluster 4	Enrichment Score: 13.15				Count	P_Value	Benjamini
GOTERM_BP_ALL	positive regulation of cellular process	RT			54	4.1E-16	1.1E-13
GOTERM_BP_ALL	positive regulation of biological process	RT			57	1.5E-15	3.3E-13
GOTERM_BP_ALL	positive regulation of transcription	RT			32	1.1E-14	2.3E-12
GOTERM_BP_ALL	regulation of transcription from RNA polymerase II promoter	RT			35	1.5E-14	3.0E-12
GOTERM_BP_ALL	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	RT			32	2.0E-14	3.7E-12
GOTERM_BP_ALL	transcription from RNA polymerase II promoter	RT			36	7.3E-14	1.2E-11
GOTERM_BP_ALL	positive regulation of transcription, DNA-dependent	RT			28	1.8E-13	2.7E-11
GOTERM_BP_ALL	positive regulation of metabolic process	RT			34	9.9E-13	1.1E-10
GOTERM_BP_ALL	positive regulation of cellular metabolic	RT					

http://bgee.unil.ch/

Example: Pax-6 and the eye

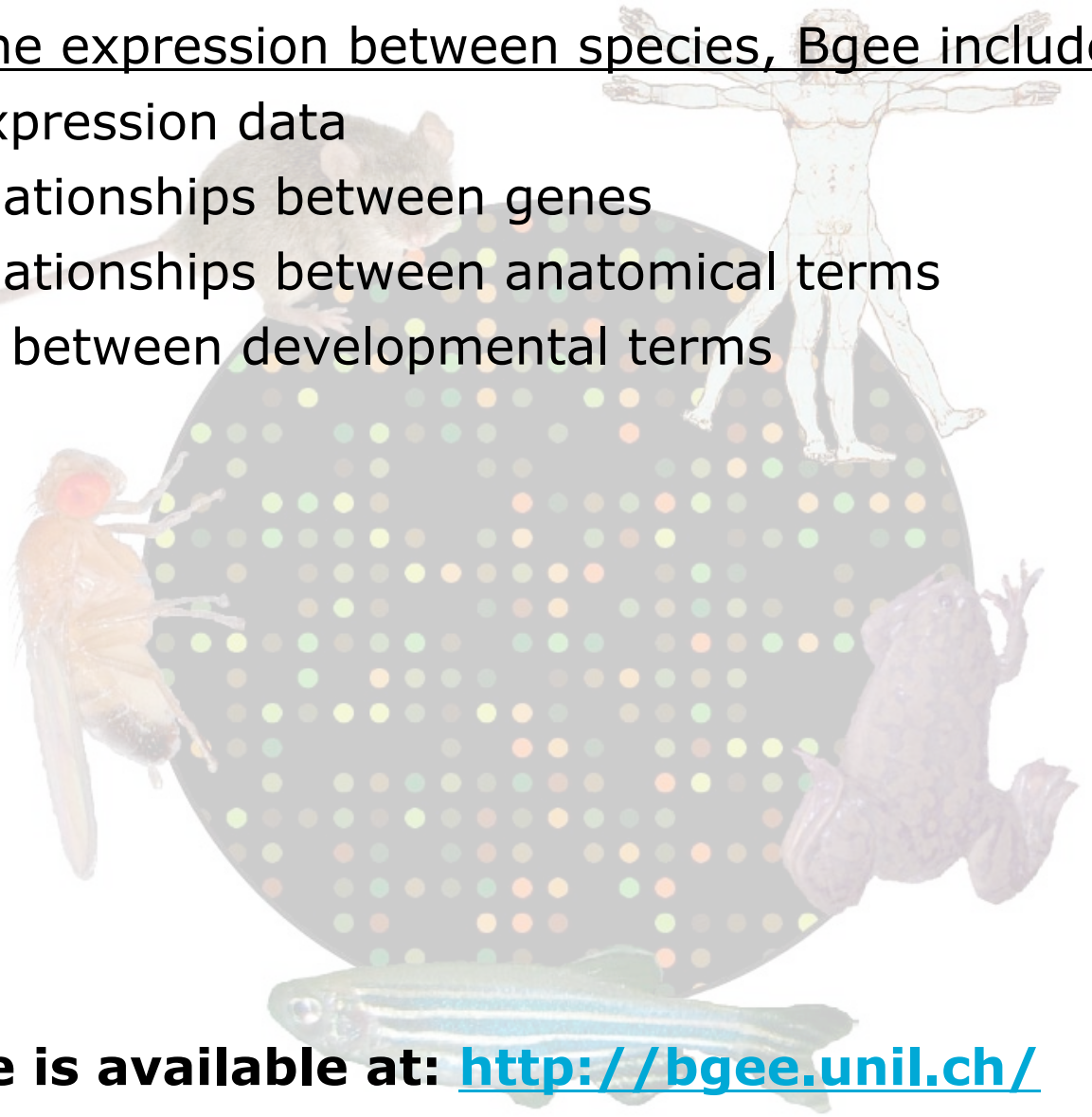


Annotation Cluster 1	Enrichment Score: 34.09			Count	P_Value	Benjamini
GOTERM_BP_ALL	multicellular organismal development	RT		136	9.3E-47	4.8E-43
GOTERM_BP_ALL	developmental process	RT		151	1.7E-43	4.4E-40
GOTERM_BP_ALL	system development	RT		113	5.0E-40	8.7E-37
GOTERM_BP_ALL	anatomical structure develo	RT		119	8.9E-38	1.2E-34
GOTERM_BP_ALL	organ development	RT		93	2.8E-32	2.9E-29
GOTERM_BP_ALL	nervous system development	RT		65	5.7E-31	5.0E-28
GOTERM_BP_ALL	cellular developmental process	RT		102	1.4E-29	8.8E-27
GOTERM_BP_ALL	cell differentiation	RT		102	1.4E-29	8.8E-27
GOTERM_BP_ALL	multicellular organismal process	RT		144	7.7E-23	3.6E-20
Annotation Cluster 2	Enrichment Score: 20.35			Count	P_Value	Benjamini
GOTERM_BP_ALL	nervous system development	RT		65	5.7E-31	5.0E-28
GOTERM_BP_ALL	central nervous system development	RT		33	4.6E-19	1.5E-16
GOTERM_BP_ALL	brain development	RT		29	3.7E-18	1.1E-15
GOTERM_BP_ALL	forebrain development	RT		21	4.0E-16	1.2E-13
Annotation Cluster 3	Enrichment Score: 14.95			Count	P_Value	Benjamini
GOTERM_BP_ALL	sensory organ development	RT		33	3.2E-23	1.6E-20
GOTERM_BP_ALL	eye development	RT		20	1.1E-14	2.3E-12
GOTERM_BP_ALL	camera-type eye development	RT		13	3.9E-9	2.5E-7
Annotation Cluster 4	Enrichment Score: 13.15			Count	P_Value	Benjamini
GOTERM_BP_ALL	positive regulation of cellular process	RT		54	4.1E-16	1.1E-13
GOTERM_BP_ALL	positive regulation of biological process	RT		57	1.5E-15	3.3E-13
GOTERM_BP_ALL	positive regulation of transcription	RT		32	1.1E-14	2.3E-12
GOTERM_BP_ALL	regulation of transcription from RNA polymerase II promoter	RT		35	1.5E-14	3.0E-12
GOTERM_BP_ALL	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	RT		32	2.0E-14	3.7E-12
GOTERM_BP_ALL	transcription from RNA polymerase II promoter	RT		36	7.3E-14	1.2E-11
GOTERM_BP_ALL	positive regulation of transcription, DNA-dependent	RT		28	1.8E-13	2.7E-11
GOTERM_BP_ALL	positive regulation of metabolic process	RT		34	9.9E-13	1.1E-10
GOTERM_BP_ALL	positive regulation of cellular metabolic	RT				

Conclusion & Perspectives

To compare gene expression between species, Bgee includes:

- integrated expression data
- homology relationships between genes
- homology relationships between anatomical terms
- relationships between developmental terms



Bgee is available at: http://bgee.unil.ch/

Conclusion & Perspectives

To compare gene expression between species, Bgee includes:

- integrated expression data
- homology relationships between genes
- homology relationships between anatomical terms
- relationships between developmental terms

In progress:

- miRNAs
- analogy (homoplasy) and deep homology relationships
  - classification of relevant types of relationships
  - management of pairwise (non transitive) relationships
  - integration of Drosophila
- differential gene expression
- programmatic access
  - DAS ready, SOAP in progress, EMBOSS access in progress

Acknowledgements

ex-members:

Gilles Parmentier
Frederic Ricci

UNIL | Université de Lausanne

FNSNF Swiss National Science Foundation

SIB

CRESCENDO

Bgee is available at: http://bgee.unil.ch/