Bioinformatics for evolutionary developmental biology

Marc Robinson-Rechavi

Evolutionary Bioinformatics group, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland

1. INTRODUCTION

Evolutionary developmental biology ("Evo-Devo") has emerged as one of the most exciting areas of biology. It is providing for the first time some answers to long standing questions, such as the origin of novelty, sources of phenotypic variation, or the relationships between diverse animal body plans. And forcing us to reconsider apparently known answers, such as the relation between micro and macro evolution, or the definition of homology *(1-6)*. Diverse fields of research have contributed to this success, including of course developmental biology and evolutionary biology, but also classical zoology, paleontology, or molecular genetics *(7, 8)*. Most recently, genome projects from diverse have also brought important insight into the evolution of developmentally important genes *(9-12)*.

Thus Evo-Devo is by its very nature an interdisciplinary science; and so is bioinformatics. We are interested in this chapter in the interface between these two interdisciplinary fields. While some bioinformatic studies have implications for Evo-Devo, and some Evo-Devo studies (especially analyzing genome sequences) make use of bioinformatics, this interface has been rather neglected up to now *(*see also *13)*.

One field where Evo-Devo has been motor in posing questions that may be answered using bioinformatics is the study of whole genome duplication. Although suggestions of the importance of duplication in evolution have been recurrent *(14)*, the main evidence in support of this theory came in the 1990's from the study of Evo-Devo. Most notably the discovery of four Hox gene complexes in human and mouse, compared to only one Hox complex in many invertebrates, such as the fruit fly *(15)*. This was rapidly suggested to be consistent with an hypothesis of two rounds of genome duplication at the origin of vertebrates, suggested by Susumo Ohno in a book *(16)* which provided the classical framework for the study of genome duplication. Ohno suggested that mutation of existing genes cannot generate new functions without risking loss of the original function, whereas duplication creates redundancy of this original function, allowing one copy to diverge and adopt a new function. Thus the divergence of orthologs (homologs diverging after speciation) would be conservative, whereas the divergence of paralogs (homologs diverging after duplication) would allow for the evolution of novelty. This idea has become mainstream in comparative genomics *(*e.g. *17)*. Ohno also suggested that these duplications could be linked to the "complexity" of lineages such as vertebrates, an appealing idea in light of the duplications of Hox complexes (key regulators of animal development), but one which has proven difficult to test. Especially

that the position of mammals as the apex of such "complexity" was short lived, with the discovery of seven Hox complexes in the zebrafish *(18)*.

Genome duplication, although rare, as emerged as an important factor in genome evolution. Genome scale evidence first came, surprisingly, from the simple yeast *Saccharomyces cerevisiae*, with the discovery that the yeast genome was tiled by non overlapping duplicated blocks *(19)*. Further studies in yeast established comparative mapping on duplicated and non duplicated species as the best way to prove and date whole genome duplication *(20)*. In addition to yeast, evidence for ancient whole genome duplications has notably been found in *Arabidopsis thaliana (21, 22)*, cereals *(23)*, teleost fishes *(24, 25)*, and paramecium *(26)*. Finally, a combination of comparative mapping and phylogeny has provided support for Ohno's *(16)* hypothesis of two whole genome duplications at the origin of vertebrates *(27-29)*. More recent tetraploids are also known in various vertebrate lineages *(30)*.

We will first present results on the use of sequence analysis, notably phylogenetics, which shed some light on questions from Evo-Devo. In a second part, we will present ongoing research to model more complicated anatomical and developmental data, to provide a bioinformatic platform for Evo-Devo studies.

2. THE EASY PART: THE EVOLUTION OF GENE SEQUENCES

2.1. Rapid Overview of Bioinformatics involved

The basic task in relating sequence evolution to developmental biology is finding genes of interest, listing all their homologs, and determining their phylogenetic relationships. To study their evolution, we may also be interested in functional classification, and in evidence for selective pressure. For example, a change in selective pressure on some sites may be evidence for a change in function of the protein.

The first task can be considered in two manners: we may start with candidate genes, and search for their homologs, or we may start by determining homologs genome-wide, and use the result of this analysis to select genes of interest. In both cases, a key step is identifying homologs by sequence similarity. This is a topic abundantly treated elsewhere, but we need to note here that many genes of interest in Evo-Devo are characterized by short conserved domains which may be difficult to identify in standard scans using e.g. BlastP *(31)*. Transcription factors such as the Hox, bZIPs or bHLHs are thus best identified by the careful use of Hidden Markov Models *(*e.g. *32)*, and are often missed in large scale scans for

homologs. An interesting exception to this is the nuclear hormone receptor superfamily, whose members can be readily thanks to their ligand binding domain of approximately 200 amino acids *(33, 34)*.

The topic of genome duplication raises that of the distinction between orthologs and paralogs *(17, 35)*. The theoretically correct way to do this is through phylogenetic inference, although numerous alternative methods have been proposed. We will not treat these methods in detail here, but note that we use likelihood methods systematically *(*e.g. PhyML, *36)*. While the most common use of phylogenies in such studies will be to start with the target genes, and analyze their phylogeny, in some case we perform the reverse task. For this, we rely on existing databases of gene trees, such as TreeFam *(37)* or the databases of the PBIL *(38-40)*, combined with tree reconciliation tools *(41)*. The latter allow us to specify a topology and search for all gene trees (or sub-trees) which match it. Thus we can identify all genes which were retained in duplicate after whole genome duplication in fishes, but not duplicated in tetrapodes, specifying also species or lineages in which gene loss is allowed or forbidden.

2.2. The Importance of Duplication and Loss

2.2.1. Why don't flies have retinoic acid receptors?

Nuclear hormone receptors (or nuclear receptors, NRs) are transcription factors which are specific to Metazoa (animals). They include receptors of major hormone, such as steroids or thyroid hormone. NRs play important roles in many central biological processes, notably development *(*reviewed in *42)*. A typical example is the group of retinoic acid receptors, which includes in human RARα, RARβ and RARγ. RARs mediate the regulation of antero-posterior expression of Hox genes in vertebrate development by retinoic acid. Whereas the Hox are largely conserved between mammals and flies, no ortholog of RAR is found in the Drosophila genome. Neither for that matter are orthologs of classical steroid receptors (ERs, AR, PR, MR and GR), nor of thyroid hormone receptors (TRs). Moreover, orthologs of these genes are not found in nematode genomes either. The steroid receptors are even absent from the genome of *Ciona intestinalis*. In fact, while most of the 48 human nuclear receptors have known ligands (hormones or fatty acids), most of the 21 fly nuclear receptors are so-called "orphans", without a known ligand. These observations led to the suggestion that most liganded nuclear receptors were vertebrate innovations *(9, 43)*.

When a sufficient sampling of animal genomes became available, we took advantage of the conserved structure of nuclear receptors to search for all homologs, and performed a

global phylogenetic analysis of the superfamily *(44)*. By combining this gene tree with the known species phylogeny, we can date duplications, but also losses. On a rooted tree, events which are closer to the tips are more recent, and events which are closer to the root are more ancient. This does not provide an absolute dating (in years), but it does provide a relative dating. Thus if we start with human RARs (Fig. 1), we see that all speciations among vertebrates (e.g. tetrapode / fish) happened more recently that the duplications that gave rise to RARα, β and γ, but that these duplications occurred after the speciation between Ciona and vertebrates. Thus these duplications date to the origin of vertebrates. If we go back in time before the split between the Ciona and vertebrate RAR orthologs, we find not a speciation node, but an older gene duplication which gave rise to RARs and other nuclear receptors. When did this older duplication occur? The ROR/HR3 sub-tree includes not only speciations among chordates and vertebrate specific duplications, like RARs, but also a speciation node between chordates on the one hand, and insects and nematode on the other. In other words, the speciation between ecdysozoans and deuterostomes at the origin of bilaterian animals *(45, 46)*. In the tree, this speciation is clearly more recent that the duplication leading to RARs, RORs, and other nuclear receptors. Thus the order of events was the following (Fig. 1): duplications leading to proto-RAR, proto-ROR, proto-Rev-erb, and other NRs; then speciation between ecdysozoans and deuterostomes. Then, to explain the lack of any RAR ortholog in the sequenced ecdysozoan genomes, we must infer that this gene was lost in ecdysozoans. Thus, RAR does not appear as a vertebrate innovation, bur rather as an ecdysozoan loss. Of note, an RAR ortholog has also been identified in the sea urchin genome *(47)*, a deuterostome but not a chordate.

Figure 1

In the analysis of the whole superfamily, we find this pattern repeating itself: vertebrate innovations are invertebrate losses *(44)*. Symmetrically, insect specific genes (e.g. E78) are ancestral bilaterian genes lost in the chordate lineage. This pattern has been spectacularly confirmed for steroid receptors, with the cloning and characterization of an estrogen receptor ortholog from a mollusk *(48, 49)*. Indeed ancestral sequence reconstruction shows that the ancestor of bilaterian animals probably had a receptor activated by estrogen.

2.2.2. Why do humans have three retinoic acid receptors?

Flies may not have retinoic acid receptors, but humans have three. RARα, β and γ are paralogs, kept from the genome duplications at the origin of vertebrates. All three bind All Trans Retinoic Acid, and activate transcription of target genes. Yet they are not redundant. Not only have the three copies been kept over 400 MY of vertebrate evolution, but Knock-Out experiments show paralog-specific phenotypes, mostly affecting development *(50)*. Like other nuclear receptors, RARs have a DNA binding domain and a ligand binding domain. The latter is about 270 amino acids in RARs, and is composed of 12 α-helices. Of these, 25 amino acids in the hydrophobic ligand binding pocket make direct contact with the ligand. The binding pockets of the human RAR paralogs differ in three of these 25 positions. They also differ in their *in vitro* binding to different synthetic retinoids, and in the resulting transactivation of target genes.

To gain further insight into the evolution of these differences, we compared all chordate RARs, including vertebrates, amphioxus and tunicates *(51)*. The phylogeny confirms the dating of the duplications at the origin of vertebrates, with single copy orthologs in amphioxus and tunicates. The amino acid sequence of the ligand binding domain of the RAR immediately predating the duplications was predicted using Maximum Likelihood, and synthesized. All homologs and the predicted ancestral protein expectedly transactivate with all-trans retinoic acid, with similar EC50 values. On the other hand, the ancestral, amphioxus and tunicate RARs do not transactivate in the presence of retinoids specific of human RARα or γ. They do bind the RARβ specific retinoid, with strong transactivation in amphioxus, weaker for tunicate and predicted ancestral. Targeted mutations of the amphioxus RAR show that the three positions identified in human are indeed key to the differences in specificity. These results suggests that RARβ is closest to the ancestral function, with changes evolving by point mutations in the ligand binding pocket in RARα and γ. This is confirmed by limited proteolytic experiments, which show that all RARs do bind the β specific retinoid, even when it is not sufficient for transactivation. They do not bind the other specific retinoids. Interestingly, *in situ* hybridization shows that the expression pattern of amphioxus RAR is most similar to that of RARβ. Thus it appears that after whole genome duplications, one copy kept close to the ancestral function, both in terms of sequence and expression pattern, while the two others acquired derived characteristics *(51)*.

This study focused on differences between paralogs, due to duplication. But it is worth noting that we also found differences between orthologs. For example both zebrafish and

Xenopus RARγ transactivate with both the α and γ specific retinoids (as established in mammals), and indeed have one amino acid in helix H3 which is identical to mammalian RARα, not γ *(51)*. A more general consideration of nuclear receptors shows that functional differences between orthologs are not rare when distant organisms are compared. For example vertebrate Rev-erbs are orphan receptors involved in circadian cycle regulation *(52)*, but E75, their insect ortholog (Fig. 1), is a heme receptor involved in ecdysone regulation *(53)*. Thus function may change after duplication, but also between species *(*see also *54)*.

2.2.3. Biased gene loss after whole genome duplication

After whole genome duplication, duplicate copies of genes may evolve in different manners, gaining or losing functions *(*reviewed in *55)*. But the most common fate is certainly loss of one of the copies. Although this may not seem very exciting in itself, the contrast between which genes are lost, and which are kept in double, has emerged as one of the most important features of whole genome duplications. The rate of gene loss has been estimated at 88% in about 80 Myr since genome duplication in yeasts *(20)*, 70% in ≤86 Myr in Arabidopsis *(56)*, and 79% in about 61-67 Myr in cereals *(23)*. By comparing only genes which were mapped to chromosomes in Tetraodon fish and human, and whose evolutionary fate could be determined by phylogenetic analysis, we obtained a figure of 85% of gene loss after whole genome duplication in teleost fishes *(57)*, despite the greater age of the event. These similar figures are best explained if most loss occurs rapidly after duplication *(58)*, so that subsequent evolution does not change the figure significantly.

In an important study, Davis and Petrov *(59)* showed that slowly evolving genes are more likely to be found duplicated. The bias is similar in yeast and in nematode worm, and is maintained over evolutionary time, indicating that gene retention was also biased after the whole genome duplication in yeast. An important aspect of the work of Davis and Petrov *(59)* was to use estimates of selective pressure which are phylogenetically independent of the duplication. We conducted a similar study in fishes *(57)*, and found that these conclusions also applied to a more ancient genome duplication, in a vertebrate lineage. The selection pressure is measured by the ratio of the number of non synonymous substitutions per site (dN) to the ratio of synonymous substitutions per site (dS), between the human and mouse orthologs of genes which either lost one copy ("singletons") or did not after the fish whole genome duplication (Fig. 2). Using only human - mouse dN to measure the rate of evolution of the proteins encoded, we find that non-duplicated orthologs of gene pairs retained after

duplication evolve 30% slower. This is comparable to observations for nematode (25%) and yeast (50%) genes.

Figure 2

What is the relevance of these observations to Evo-Devo? First, if whole genome duplication has really played a key role in the establishment of the developmental diversity of fishes *(60, 61) (*but see *62)*, we need to understand its mechanisms as well as possible. Second, gene retention may also be biased relative to function, and enrichments in communication and developmental genes has been reported in insects, yeasts *(63)* and Arabidopsis *(64)*. In fishes, we found an excess of genes annotated with terms related to development and signaling functions *(57)*, which supports the putative link between genome duplication and developmental innovations.

Second, once the importance of biased gene retention after duplication is established, this provides a convenient measure of selective pressures on the genome. We have used this to test developmental constraints on genome evolution. Two models have been proposed for developmental constraints on morphological evolution. The first, dating back to pre-Darwinian observations by von Baer *(1, 65)*, is that there is a progressive divergence of morphological similarities between vertebrate embryos. More general characters would form in early development, which would be highly constrained, while species-specific characters would form in late development, which would be more open to innovation. An alternative models was proposed more recently *(66, 67)*: the "hourglass model" is based on the observation of large morphological diversity in very early development (e.g. blastula). This model assumes a constrained stage in middle development, around the vertebrate "pharyngula" stage. To evaluate the impact of the constraints postulated by both models on the genome, we investigated the pattern of expression during development according to gene retention after whole genome duplication. We expect genes which are kept in double to be highly expressed at developmental stages which are open to evolutionary innovation, not at stages which are highly constrained. These should be characterized by conservatism: no duplication, no loss of highly expressed genes (Fig. 3). We also expect a high cost of gene loss (e.g. lethal phenotype for gene Knock-Out) in more constrained stages.

Figure 3

By combining a zebrafish time series microarray experiment (E-TABM-33 accession from ArrayExpress) with phylogenetic definition of retention or loss after duplication in fishes (Fig. 2), we find that genes kept in duplicate are lowly expressed in early development, then increase regularly their expression to reach a maximum in late development (Roux and Robinson-Rechavi, unpublished). In principle, this could be the result of biased evolution after duplication (e.g. duplicate genes evolving lower expression in early development), as well as of biased retention and loss. To check this, we used EST data to determine expression in mouse development: the orthologs of genes kept in duplicate in fishes are lowly expressed in early mouse development. Moreover, the data fit a simple linear correlation, while excluding the parabola curve expected from the hourglass model (Fig. 3). The same results are obtained contrasting genes kept or lost after whole genome duplications at the origin of vertebrates. Finally, gene KO phenotypes are also consistent with decreasing constraints over development both in zebrafish and mouse.

These results together show that (i) timing of expression during development is a strong and conserved constraint on genome evolution; (ii) gene duplication is restricted when phenotype is constrained; and (iii) at the genomic level, the traditional "von Baer-like" model provides a much better fit than the hourglass model.

In conclusion of this section, gene duplication and loss must both be understood to clarify the relationship between genome evolution and developmental (thus morphological) evolution.

3. DEVELOPING BIOINFORMATIC TOOLS FOR EVO-DEVO

3.1. Defining homology for bioinformatics

Homology is one of the most fundamental concepts in biology. It has also generated abundant terminological and conceptual discussion *(*e.g. *68)*. It was originally defined based on similarity of anatomical structures, with emphasis on their relations, thus clarifying that a bird wing and a mammalian forelimb are homologous. The term later acquired a historical dimension: homologous structures are assumed to derive from a same ancestral structure, to which they owe their similarity, whereas analogous structures have converged to similarity from different ancestral starting points. This definition can create complex situations, since the bat wing and the bird wing are analogous as wings, having converged from an ancestral walking limb, but they are homologous as forelimbs, since they derive from the forelimb of an ancestral tetrapode. Moreover, different fields of research have developed different

operational definitions *(69)*: historical homology, defined by phylogenetic continuity; morphological homology, defined by structural similarity; or biological homology, defined by similar developmental constraints. Morphologists also define serial homology between organs repeated along the axis of a same organism, such as vertebrae *(*discussed in *70)*. Since the original evolutionary formulations of homology, two fields of research have hugely influenced our view of biological diversity and evolution: molecular evolution and Evo-Devo.

At the molecular level, homologs between species have been discovered whose origin predate the divergence of bacteria and eukaryotes, while inside each species, genomes include many families of homologous genes. To clarify this situation, three main types of molecular homology are distinguished: orthology, or divergence by speciation; paralogy, or divergence by sequence duplication; and xenology, or divergence after gene transfer between species *(35)*. Such molecular homology is probably the only type which has been well formalized in bioinformatics up to now.

Evo-Devo is dependent on definitions of homology both for anatomical structures and for genes, and has also brought important new information relative to homology. Some of this new information has raised new questions. This is best exemplified by the now classical case of animal eyes: if insect eyes and vertebrate eyes are organized in fundamentally different ways, they are probably analogs; but if they are determined during development by orthologous genes, are they not homologs? Such cases have led to the controversial proposal that the presence of key orthologous genes in development suffices to define homology *(*discussed in *71)*. An attempt to solve this issue is the new term "homocracy", which is defined by sharing the expression of the same patterning genes *(72)*. Homocratic structures may or may not be homologous; homologous structures are often homocratic, but this is not a logical necessity.

When defining homology in development, we must also take into account heterochrony, changes in the relative timing of development during evolution. For each organ homology should be defined at specific developmental stages, which differ between organs. For example, the heart develops later in primates relative to rodents, while the ear develops earlier, thus changing the timing of development of these organs relative to each other *(73)*. This makes it impossible to define homology between developmental stages as a whole in many cases, and greatly complicates automatic comparison of embryos between species. Moreover, developmental "stages" as defined in the literature are somewhat arbitrary divisions of a continuous process. The limits of these divisions may not be consistent between species, even without heterochrony.

Despite the importance of homology, little attention has been paid to careful implementation in bioinformatics. Notably, few ontologies contain any notion of homology. Ontologies are formal representations of a field of knowledge, including terms, definitions and relations between the terms (e.g. hydrolase is_a enzyme). They have become an important tool in bioinformatics, corresponding to the need to formalize many complex descriptions in biology. Some ontologies do provide homology relationships. PATIKA *(74)*, a pathway ontology, includes a "homology" relation; in practice it is used to manage paralogy inside gene families. Protein homology is defined as a "synonym" in the Molecule role ontology of the INOH Pathway Database. The most detailed implementation to our knowledge is in the Sequence Ontology *(75)*, which includes a "homologous_to" relation, which is the only child of "similar_to", and has three children, "orthologous_to", "paralogous_to" and "non_functional_homolog_to". The latter is an interesting formalization of the relation between a gene and a pseudogene. The child relation to "similar_to" shows that a morphological definition of homology was chosen, whereas the Cell Ontology uses a definition of historical homology *(76)*. But in the Cell Ontology the relation is not implemented explicitly. Instead, homology is the default for the same term in different species (e.g. "muscle_cell" in human and fly); otherwise, several lineage specific terms are created, as in "pigment_cell_(sensu_Vertebrata)" and "pigment_cell_(sensu_Nematoda_and_Protostoma)". A similar approach is used in the Plant Ontology *(77)*. In several ontologies, homology is not defined as a type of relation, but is discussed in the definitions. For example, good discussions of anatomical homology, including serial homology and analogy, appear in definitions of the ontologies of mosquito *(78)* or corn *(79)*.

3.2. Modeling homology relationships

To conduct Evo-Devo studies computationally, we need to define homology relationships between ontologies describing the anatomy and development of different species. Designing such relationships consists in finding correspondences (homology relationships) between the concepts (organs) of these ontologies. This problem is a special case of "schema matching", or "ontology alignment". Ontology alignment *(80)* is the process of determining correspondences between ontology concepts. Usually, this technique is used to find the common concepts present in two ontologies. In the case of anatomical ontologies, the concepts to align are not strictly common, but rather, related: a homology relationship is not an equivalence relationship. For this reason, classical ontology alignment approaches cannot

be applied here: these methods would be misled by the existence of elements of same names and related to the same concept, but not homologous (eye of insects and of vertebrates for instance), or homologous elements with different names (caudal fin and upper limb for instance). This is why in our modified ontology alignment technique (Parmentier and Robinson-Rechavi, unpublished) an expert has to manually validate the putative homologs.

Our process is a supervised one: at each step, some homology relationships are proposed to the expert, who may validate them or not. Computations are made based on these decisions, and new propositions are made to the expert.

The algorithm starts with a list of pairs, which have identical names. This is based on the assumption that two structures that have the same name are likely homologous. For example, "optic cup" of ZFIN (zebrafish) *(81)* and "optic cup" of EHDA (human) will be paired, but "optic cup" of ZFIN will not be initially paired with "optic nerve" of EHDA. The score of similarity between terms is up weighted by the proportion of common words, and down weighted by the frequency of these words (frequent words are less informative, e.g. "endoderm"). Moreover, scores are propagated between pairs which are neighbors in both ontologies. For example, the score of the "optic cup" pair is added to the score of the "eye" pair, as "optic cup" is part of "eye".

Each pair is proposed to the expert, in descending order of scores. The expert may validate or invalidate the hypothesis of homology, or delay decision. The expert may choose to evaluate any number of pairs before triggering an iteration, in which computations are performed. Computation creates or extends homology groups. The new homology information is propagated through the ontologies. The underlying idea is that if two concepts A and B are homologous, then one of the sub-concepts of A is probably homologous to one of the sub-concepts of B. Of note, validated homology contributes a significantly higher score than name similarity. Propagation is down weighted by the number of sub-concepts, to avoid generating many false positives (i.e. all the children of "whole body").

Evaluation of pairs, ordered by total score (base score + propagated score), and iteration, are repeated until the expert decides to terminate, or no more pairs are proposed.

Our method is implemented in Homolonto (Parmentier and Robinson-Rechavi, unpublished), a software that we have developed in Java. Compared to manual alignment of the ontologies, Homolonto reduces time considerably, with high sensitivity. Thus aligning the zebrafish (ZFIN; 2087 terms) and Xenopus (Xenbase; 480 terms) ontologies took one month by hand, but 2 days using Homolonto. The first 213 pairs proposed to the expert (i.e. one day of work) were valid at 80%, and contained 91% of all true positives.

3.3. Bgee, a database for gene expression evolution

To be useful for Evo-Devo and other comparative studies, the homology relationships determined using Homolonto are implemented in a database, Bgee. This "dataBase for Gene Expression Evolution" is being developed to facilitate comparisons of gene expression between animal species (http://bgee.unil.ch) *(82)*.

To enable large scale gene expression pattern comparison, Bgee must answer three conditions: (i) Precise description of the anatomy and developmental stages of each species, stored in a computer-understandable way. This is done using existing ontologies, such as ZFIN *(81)*. (ii) Comparison criteria between anatomies, developmental stages, and genes. For anatomy, this is done using Homolonto. For development, we have developed a small ontology of "metastages" which are common to all bilaterian animals, such as "blastula part_of embryo". For genes, we use homology predictions from other sources *(*i.e. Ensembl, *83)*. (iii) Integration of expression data in order to know in which anatomical features (spatial mapping) and which developmental stages (temporal mapping) genes are expressed. The relationships between these types of information are represented in a very simplified manner in Fig. 4.

Figure 4

Concerning developmental stages, we have seen that it is not possible to detail homologous stages in a similar way to organs. This is why we developed a simplified ontology of metastages. Despite the resulting loss of accuracy, it allows comparison of gene expression patterns taking into account developmental time.

Concerning expression data, we face two challenges: integrating heterogeneous data types *(84, 85)*; and transforming often quantitative data (level of expression) into the qualitative information which is standard in typical developmental studies ("expressed" or not). The reason to integrate heterogeneous expression data is that they complement each other in terms of coverage. For example EST libraries present typically an incomplete picture of the transcriptome, but they are available for many species, and allow good identification of closely related paralogues. Oligonucleotide microarrays are much more complete, but different experiments are difficult to compare, and non model species are not covered. Both ESTs and microarrays are usually annotated to coarse anatomical and developmental descriptions (e.g. "adult brain"), whereas *in situ* hybridizations can provide very detailed

accounts of gene expression. But *in situ* hybridization provides limited genome coverage *(*although see *86)*, is not applicable to humans, and can be more challenging to treat automatically than other transcriptome data types *(85, 87)*.

Briefly, the basic approach chosen in Bgee is to recode expression data for a gene in an organ and developmental stage as "not detected", "expressed with high confidence", or "expressed with low confidence". For experiments based on tag counting, such as ESTs, SAGE, or MPSS, we have considered a gene as expressed with a high confidence if the 95% confidence interval does not include zero *(88)*. For microarray data, a gene is considered expressed if the normalized signal is significantly above the background signal. In the future, we plan to add information on whether a gene is significantly more expressed in one condition than another (e.g. more expressed in muscle than brain), and integrate other types of data.

The database is developed with mySQL, and currently includes four vertebrate species. The website allows users to retrieve information on gene expression by querying the database for keywords or gene identifiers, or browsing anatomical or developmental ontologies. In addition to species specific or gene specific views, users may view all gene families expressed in homologous organs between chosen species, and the complete expression information of a gene family across species. All queries may be constrained by data type, data quality, and keywords or identifiers.

Bgee is a promising tool to enable Evo-Devo studies on a larger scale. We hope it will also be useful to put functional genomics studies in a comparative context, and provide a platform for integration of anatomical homology information into bioinformatics.

4. CONCLUSION

The intersection of bioinformatics and Evo-Devo is still relatively small, but holds a large potential for bringing the tools of high throughput biology to illuminate our understanding of some of the most fundamental questions in biology: the origin of novelty, the role of constraints, the importance of loss vs. gain, or the extent of conservation between distant taxa. In this chapter, we have discussed briefly two aspects of the integration of bioinformatics and Evo-Devo: sequence based analysis, and modeling homology relationships. A third approach should be mentioned in conclusion: gene regulatory networks. The characterization of gene regulatory networks controlling development is still a recent field *(90)*, and data has proven often costly to produce even in one species. It is to hoped that in the future sufficient

functional data will be available from a wider array of species. This should allow mechanistic yet large scale studies of the control of morphological diversity by the genome.

5. ACKNOWLEDGEMENTS

I thank Frédéric Bastian, Gilles Parmentier and Julien Roux for their help in preparing this chapter. Research in the laboratory of MRR is supported by the EU program Crescendo, the Swiss National Science Foundation, Etat de Vaud, the Swiss Institute of Bioinformatics, and the Decrypthon program.

6. FIGURE LEGENDS

Figure 1: Simplified phylogenetic tree of three groups of nuclear receptors
Maximum likelihood phylogeny (PhyML *(36)*, JTT model, four rate categories, gamma shape parameter and proportion of invariant estimated) of 78 nuclear receptors from the groups NR1B (RAR), NR1D (Rev-erb, E75) and NR1F (ROR, HR3). Branch length is proportional to substitutions per amino acid site. In blue, the relative timing of key speciation events. In red, the relative timing of duplication events.

Figure 2: Comparison of selection pressure on duplicated and singleton genes.
Schematic phylogenetic classification of genes according to duplication and loss. Tn = *Tetraodon nigroviridis*; Tr = *Takifugu rubripes*; Hs = *Homo sapiens*; Mm = *Mus musculus*. dN = number of non synonymous substitutions per site; dS = number of synonymous substitutions per site. The arrow represents the unpaired t-test between dN/dS values.

Figure 3: Schematic predictions of two models of developmental constraints on evolution.

Figure 4: Schematic representation of the relationships between types of information in the Bgee database.
Expression data is central to relating genes to anatomical and developmental terms in each species.

7. REFERENCES

1- Gould SJ (1977) *Ontogeny and phylogeny*. The Belknap Press of Harvard University Press. Cambridge, Mass.

2- Gould SJ (2002) *The Structure of Evolutionary Theory*. Belknap Press. Boston.

3- Carroll S (2005) *Endless Forms Most Beautiful: The New Science of Evo Devo and The Making of the Animal Kingdom*. W. W. Norton & Company. New York.

4- Sanetra M, Begemann G, Becker MB, Meyer A (2005) Conservation and co-option in developmental programmes: the importance of homology relationships. *Front Zool* **2**: 15.

5- Theissen G (2005) Birth, life and death of developmental control genes: New challenges for the homology concept. *Theory Biosci* **124**: 199-212.

6- Brakefield PM (2006) Evo-devo and constraints on selection. *Trends Ecol Evol* **21**: 362-368.

7- Raff RA (2007) Written in stone: fossils, genes and evo-devo. *Nat Rev Genet* **8**: 911-920.

8- Breuker CJ, Debat V, Klingenberg CP (2006) Functional evo-devo. *Trends Ecol Evol* **21**: 488-492.

9- Dehal P, Satou Y, Campbell RK, et al. (2002) The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins. *Science* **298**: 2157-2167.

10- Sea Urchin Genome Sequencing C, Sodergren E, Weinstock GM, et al. (2006) The Genome of the Sea Urchin Strongylocentrotus purpuratus. *Science* **314**: 941-952.

11- King N, Westbrook MJ, Young SL, et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451**: 783-788.

12- Canestro C, Yokoi H, Postlethwait JH (2007) Evolutionary developmental biology and genomics. *Nat Rev Genet* **8**: 932-942.

13- Mabee PM (2006) Integrating Evolution and Development: The Need for Bioinformatics in Evo-Devo. *BioScience* **56**: 301-309.

14- Taylor JS, Raes J (2004) Duplication and Divergence: The Evolution of New Genes and Old Ideas. *Ann Rev Genet* **38**: 615-643.

15- Holland PW, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. *Development* **supplement**: 125-133.

16- Ohno S (1970) *Evolution by gene duplication*. Springer-Verlag. Heidelberg.

17- Koonin EV (2005) Orthology, paralogs and evolutionary genomics. *Ann Rev Genet* **39**: 309-338.

18- Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**: 1711-4.

19- Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708-13.

20- Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**: 617-24.

21- Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: genome duplication in *Arabidopsis* and syntheny between soybean and *Arabidopsis*. *Proc Natl Acad Sci USA* **97**: 4168-4173.

22- Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **99**: 13627-13632.

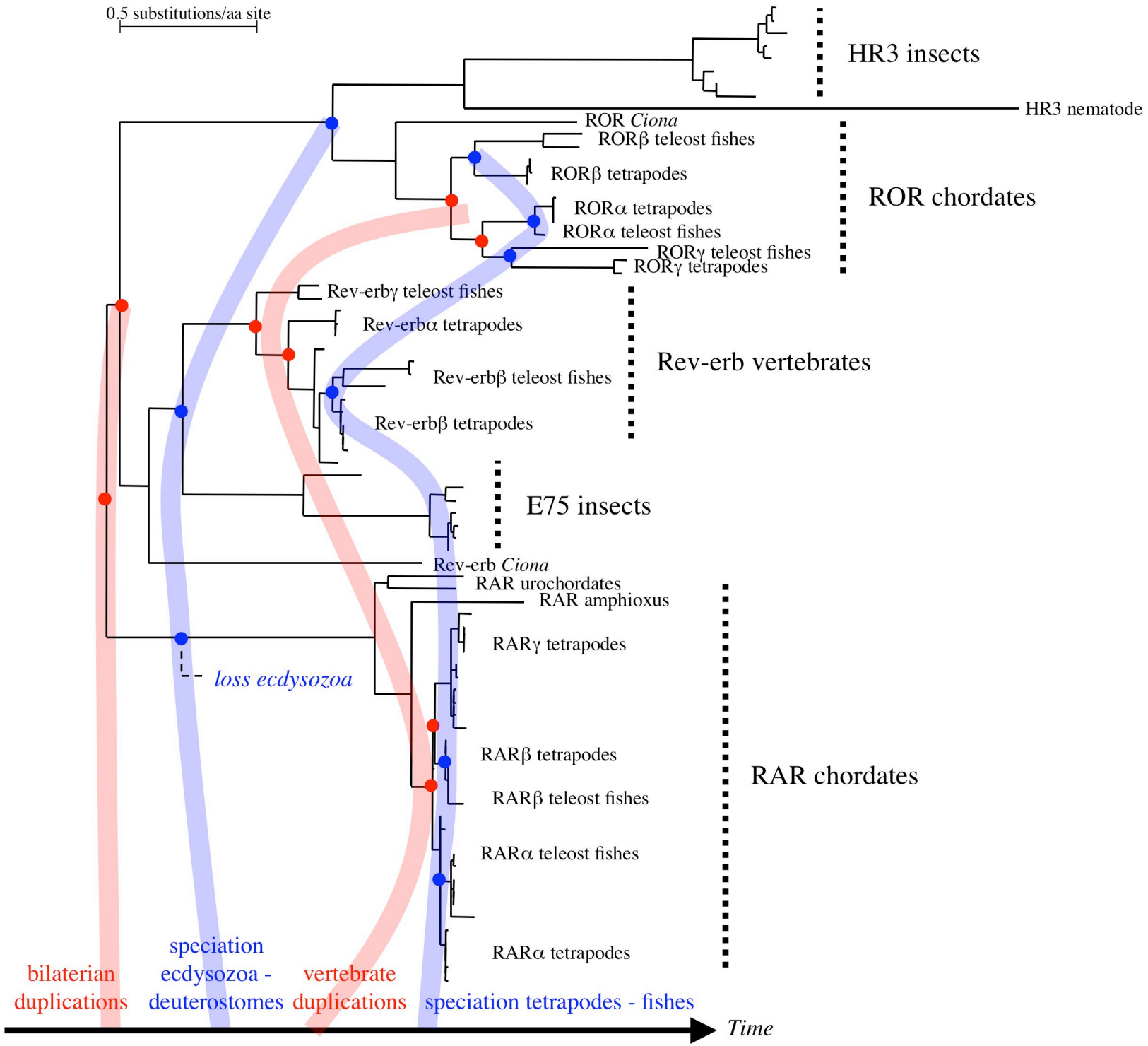
23- Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* **101**: 9903-9908.

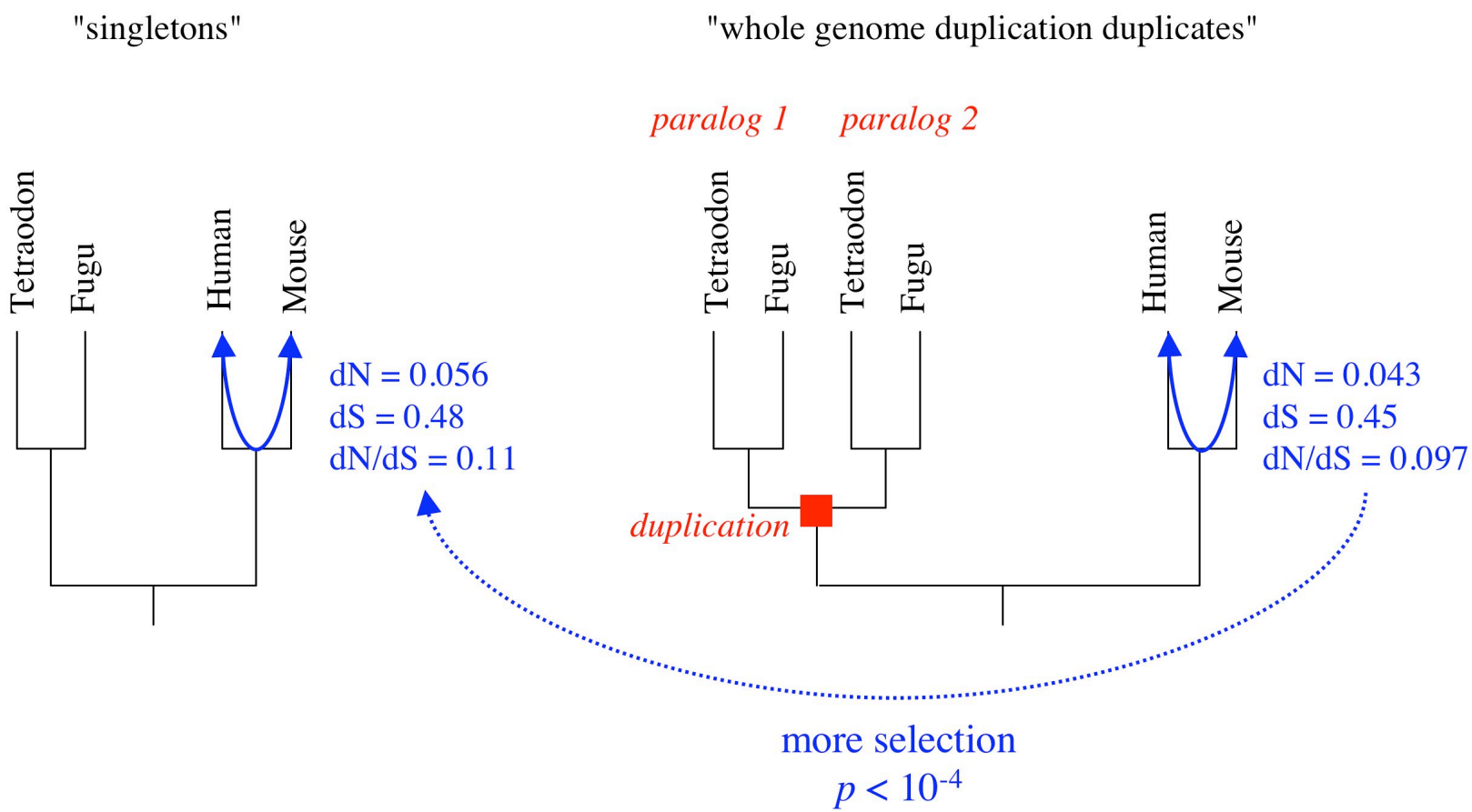
24- Jaillon O, Aury J-M, Brunet F, et al. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**: 946-957.
25- Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* **15**: 1307-14.
26- Aury JM, Jaillon O, Duret L, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**: 171-8.
27- Dehal P, Boore JL (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol* **3**: e314.
28- Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**: 1254-1265.
29- Putnam NH, Hellsten U, Yu JS, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype *Nature* **in press**.
30- Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Ann Rev Genet* **34**: 401-437.
31- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.
32- Amoutzias GD, Veron A, Weiner AJ, Robinson-Rechavi M, Bornberg-Bauer E, Oliver SG, Robertson DL (2007) One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization, and DNA-Binding Site Specificity. *Mol Biol Evol* **827-835**: 827-835.
33- Escriva Garcia H, Laudet V, Robinson-Rechavi M (2003) Nuclear receptors are markers of animal genome evolution. *J Struct Funct Genomics* **3**: 177-84.
34- Robinson-Rechavi M, Laudet V (2003) Bioinformatics of Nuclear Receptors, In: Russell DW, Mangelsdorf DJ (eds). *Methods in Enzymology*, pp. 93-118 Academic Press.
35- Fitch WM (2000) Homology: a personal view on some of the problems. *Trends Genet* **16**: 227-231.
36- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.
37- Li H, Coghlan A, Ruan J, Coin LJ, Heriche J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK-S, Zheng W, Dehal P, Wang J, Durbin R (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucl Acids Res* **34**: D572-580.
38- Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* **22**: 2360-5.
39- Perrière G, Combet C, Penel S, Blanchet C, Thioulouse J, Geourjon C, Grassot J, Charavay C, Gouy M, Duret L, Deléage G (2003) Integrated databanks access and sequence/structure analysis services at the PBIL. *Nucleic Acids Res* **31**: 3393-3399.
40- Perriere G, Duret L, Gouy M (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res* **10**: 379-85.
41- Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21**: 2596-603.
42- Laudet V, Gronemeyer H (2002) *The nuclear receptors factsbook*. Academic Press. London.
43- Escriva H, Delaunay F, Laudet V (2000) Ligand binding and nuclear receptor evolution. *Bioessays* **22**: 717-727.

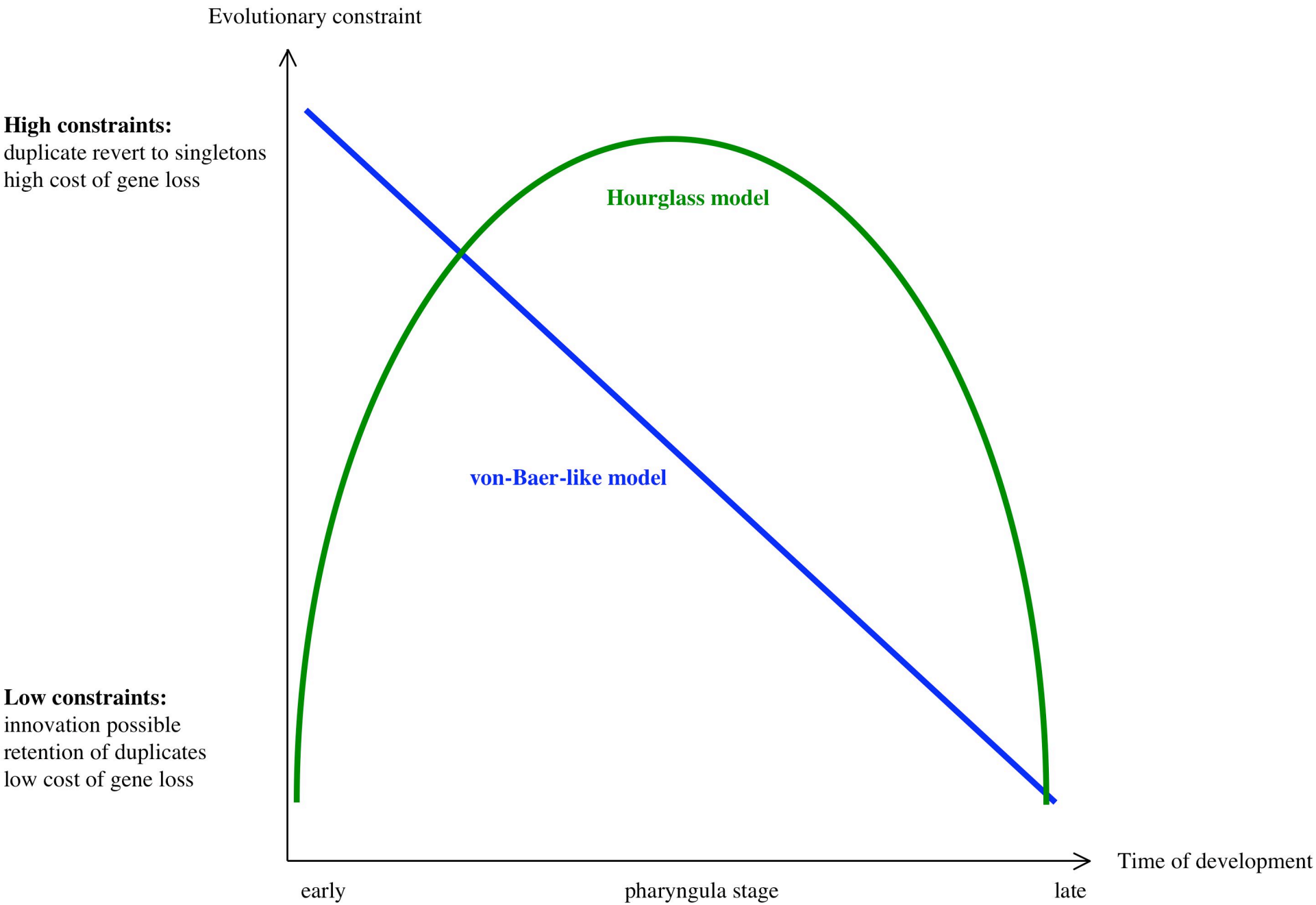
44- Bertrand S, Brunet FG, Escriva H, Parmentier G, Laudet V, Robinson-Rechavi M (2004) Evolutionary Genomics of Nuclear Receptors: From Twenty-Five Ancestral Genes to Derived Endocrine Systems. *Mol Biol Evol* **21**: 1923-1937.
45- Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R (2000) The new animal phylogeny: Reliability and implications. *Proc Natl Acad Sci U S A* **97**: 4453-4456.
46- Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*.
47- Marletaz F, Holland LZ, Laudet V, Schubert M (2006) Retinoic acid signaling and the evolution of chordates. *Int J Biol Sci* **2**: 38-47.
48- Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* **301**: 1714-7.
49- Keay J, Bridgham JT, Thornton JW (2006) The Octopus vulgaris estrogen receptor is a constitutive transcriptional activator: evolutionary and functional implications. *Endocrinology*: en.2006-0363.
50- Mark M, Ghyselinck NB, Chambon P (2006) FUNCTION OF RETINOID NUCLEAR RECEPTORS: Lessons from Genetic and Pharmacological Dissections of the Retinoic Acid Signaling Pathway During Mouse Embryogenesis. *Ann Rev Pharmacol Toxicol* **46**: 451-480.
51- Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V (2006) Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genetics* **2**: e102.
52- Duez H, Staels B (2008) Rev-erb[alpha] gives a time cue to metabolism. *FEBS Letters* **582**: 19-25.
53- Reinking J, Lam MMS, Pardee K, Sampson HM, Liu S, Yang P, Williams S, White W, Lajoie G, Edwards A, Krause HM (2005) The Drosophila Nuclear Receptor E75 Contains Heme and Is Gas Responsive. *Cell* **122**: 195-207.
54- Markov G, Lecointre G, Demeneix B, Laudet V (2008) The "street light syndrome", or how protein taxonomy can bias experimental manipulations. *Bioessays* **30**: 349-57.
55- Semon M, Wolfe KH (2007) Consequences of genome duplication. *Curr Opin Genet Dev* **17**: 505-12.
56- Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433-438.
57- Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* **23**: 1808-16.
58- Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 341-345.
59- Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biology* **2**: e55.
60- Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* **27**: 937-45.
61- Volff JN (2005) Genome evolution and biodiversity in teleost fish. *Heredity* **94**: 280-94.
62- Donoghue PCJ, Purnell MA (2005) Genome duplication, extinction and vertebrate evolution. *Trends Ecol Evol* **20**: 312-319.

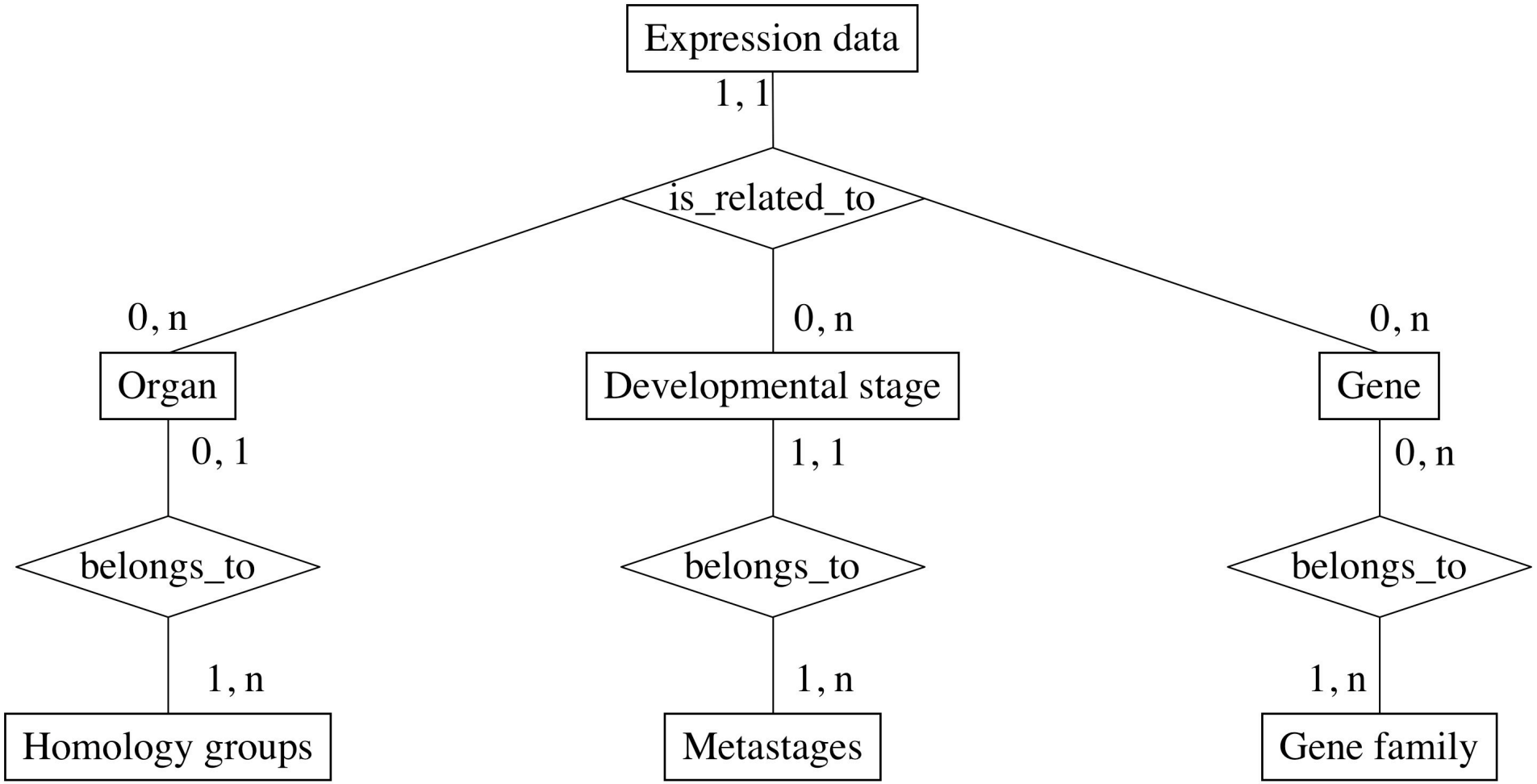
63- Jordan IK, Wolf YI, Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol* **4**: 22.
64- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* **102**: 5454-9.
65- von Baer KE (1828) *Ueber Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion*. Bornträger. Königsberg.
66- Duboule D (1994) Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*: 135-42.
67- Raff RA (1996) *The shape of life : genes, development, and the evolution of animal form*. University of Chicago Press. Chicago; London.
68- Hall B (1999) *Homology: The Hierarchical Basis of Comparative Biology*. John Wiley & Sons.
69- Abouheif E (1997) Developmental genetics and homology: a hierarchical approach. *Trends Ecol Evol* **12**: 405-408.
70- McKitrick MC (1994) On homology and the ontological relationship of parts. *Syst Biol* **43**: 1-10.
71- Wray GA, Abouheif E (1998) When is homology not homology? *Curr Opin Genet Dev* **8**: 675-680.
72- Nielsen C, Martinez P (2003) Patterns of gene expression: homology or homocracy? *Development Genes and Evolution* **213**: 149-154.
73- Jeffery J, Bininda-Emonds O, Coates M, Richardson M (2005) A New Technique for Identifying Sequence Heterochrony. *Syst Biol* **54**: 230-240.
74- Demir E, Babur O, Dogrusoz U, Gursoy A, Ayaz A, Gulesir G, Nisanci G, Cetin-Atalay R (2004) An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics* **20**: 349-356.
75- Eilbeck K, Lewis SE (2004) Sequence Ontology annotation guide. *Comp Funct Genom* **5**: 642-647.
76- Bard J, Rhee S, Ashburner M (2005) An ontology for cell types. *Genome Biol* **6**: R21.
77- Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F (2005) Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages. *Comp Funct Genom* **6**: 388-397.
78- Topalis P, Koutsos A, Dialynas E, Kiamos C, Hope LK, Strode C, Hemingway J, Louis C (2005) AnoBase: a genetic and biological database of anophelines. *Insect Mol Biol* **14**: 591-597.
79- Vincent PLD, Coe JEH, Polacco ML (2003) Zea mays ontology - a database of international terms. *Trends Plant Sci* **8**: 517-520.
80- Shvaiko P, Euzenat J (2007) *Ontology Matching*. Springer Verlag. Berlin Heidelberg.
81- Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe DG, Mani P, Ramachandran S, Schaper K, Segerdell E, Song P, Sprunger B, Taylor S, Van Slyke CE, Westerfield M (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res* **34**: D581-5.
82- Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M (2008) *Bgee: Integrating and comparing heterogeneous transcriptome data among species*. in *DILS, International Workshop on Data Integration in the Life Sciences 2008*. Evry, France: Springer LNBI series.
83- Hubbard TJ, Aken BL, Beal K, et al. (2007) Ensembl 2007. *Nucleic Acids Res* **35**: D610-7.

84- Kuo WP, Liu F, Trimarchi J, et al. (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotech* **24**: 832-840.
85- Lee CK, Sunkin SM, Kuan C, Thompson CL, Pathak S, Ng L, Lau C, Fischer S, Mortrud M, Slaughterbeck C, Jones A, Lein E, Hawrylycz M (2008) Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome Biol* **9**: R23.
86- Thisse B, Heyer V, Lux A, Alunni V, Degrave A, Seiliez I, Kirchner J, Parkhill JP, Thisse C (2004) Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. *Methods Cell Biol* **77**: 505-19.
87- Ye J, Chen J, Li Q, Kumar S (2006) Classification of Drosophila embryonic developmental stage range based on gene expression pattern images. *Comput Syst Bioinformatics Conf*: 293-8.
88- Audic S, Claverie J-M (1997) The Significance of Digital Gene Expression Profiles. *Genome Res* **7**: 986-995.
89- Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* **99**: 909-917.
90- Davidson EH, Erwin DH (2006) Gene Regulatory Networks and the Evolution of Animal Body Plans. *Science* **311**: 796-800.


0.5 substitutions/aa site
HR3 insects
HR3 nematode
ROR Ciona
RORβ teleost fishes
RORβ tetrapodes
RORα tetrapodes
RORα teleost fishes
RORγ teleost fishes
RORγ tetrapodes
ROR chordates
Rev-erbγ teleost fishes
Rev-erbα tetrapodes
Rev-erbβ teleost fishes
Rev-erbβ tetrapodes
Rev-erb vertebrates
E75 insects
Rev-erb Ciona
RAR urochordates
RAR amphioxus
RARγ tetrapodes
loss ecdysozoa
RARβ tetrapodes
RARβ teleost fishes
RARα teleost fishes
RARα tetrapodes
RAR chordates
bilaterian duplications
speciation ecdysozoa - deuterostomes
vertebrate duplications
speciation tetrapodes - fishes
Time


"singletons"
"whole genome duplication duplicates"
paralog 1
paralog 2
Tetraodon
Fugu
Human
Mouse
Tetraodon
Fugu
Tetraodon
Fugu
Human
Mouse
dN = 0.056
dS = 0.48
dN/dS = 0.11
duplication
dN = 0.043
dS = 0.45
dN/dS = 0.097
more selection
$p < 10^{-4}$


Evolutionary constraint
High constraints:
duplicate revert to singletons
high cost of gene loss
Hourglass model
von-Baer-like model
Low constraints:
innovation possible
retention of duplicates
low cost of gene loss
Time of development
early
pharyngula stage
late



Comparisons between species:
Homologous genes expressed in homologous organs at equivalent stages