Large scale analyses of positive selection using codon models

Romain A. Studer[1,2] and Marc Robinson-Rechavi[1,2,3]
[1] Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland;
[2] Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;
[3] Corresponding author.
[3] E-mail marc.robinson-rechavi@unil.ch; fax 41-21-6924165.

Abstract Positive selection is the mechanism of adaptation to the environment, as well as the main source of novelty in evolution, and thus it is of great interest to find its trace in genomes. During the last decade, different evolutionary models have been developed to detect positive selection at the gene level, based on divergence between species. Most recently, these models have been applied to large scale comparisons of genomes. We present in this chapter some strengths and limitations of such genomic scans for positive selection, and discuss the main recent large-scale studies, as well as relevant databases. We particularly discuss our recent results concerning the impact of genome duplication in vertebrate evolution, and our related database Selectome.

1 Introduction.

1.1 Positive selection as a mechanism of adaptation

A fundamental concept in evolution is selective pressure from the environment. Mutations in genomes occur at random. These mutations could have an impact in the phenotype by modifying e.g. biochemical function, or the expression of genes affected. Most mutations have a negative effect on the fitness (they are deleterious), and thus almost all genes are under purifying selection to preserve their function. Some mutations have no effect on fitness, and may be fixed under the neutral process of drift. Others will have a beneficial effect and they will be kept in the genome by the process of positive selection (also called Darwinian, adaptive, or directional selection). This positive selection is the mechanism of adaptation to the environment, and thus it is of great interest to find its trace in genomes. During the last decade, many molecular analyses have found different categories of genes preferentially affected by adaptive selection (Yang 2006).

1.2 Functional categories of genes

The most representative categories of genes under positive selection are involved in arm-race adaptation. Virus and bacteria are rapidly mutating due to the absence of control in replication and to the large population size, respectively. These mutations could affect various genes like HIV proteins (env (Nielsen and Yang 1998), gag, pol) or the wsp protein in the outer membrane in Wolbachia bacteria (Jiggins, Hurst et al. 2002). These genes are on perpetual adaptation against drugs and immune systems. Similarly, other organisms have to continually optimize their defenses in order to counter these attacks. The Major Histocompatibility Complex (MHC) classes I and II are evolving to recognize any kind of external peptides and are subject to positive selection in both classes (Hughes and Nei 1988; Hughes and Nei 1989; Hughes, Hughes et al. 1994). Many other immunity genes are also under positive selection, like Glycophorin A (Baum, Ward et al. 2002), CD4 glycoprotein (Zhang, Weinstock et al. 2008) or TRIMα (Sawyer, Wu et al. 2005).

Others main categories are genes involved in sexual reproduction, like the sperm lysin in abalone (Lee, Ota et al. 1995) or the protamine P1 in primates (Rooney and Zhang 1999). Genes of perception are also found to be under positive selection (ie Olfactory Receptor OR5I1 (Moreno-Estrada, Casals et al. 2008)). Cases of genes involved in digestion have also been reported, notably the Lysosyme in primates (Messier and Stewart 1997). This could be explained by a change in diet.

1.3 The case of duplicated genes.

Another group of genes, which are frequently reported to evolve under positive selection, are not related to a particular functional category: duplicated genes. After a duplication event, a gene will be present twice in the genome, implying a cost of redundancy, which may be balanced by a selective gain in function. Different theoretical models have been elaborated to predict the fate of these duplicated genes (see review (Zhang 2003)). The main fate is simply loss of one copy, by **nonfunctionalization**. Assuming that duplicates are deleterious (in term of stochiometry or cost of expression), most additional copies will be rapidly erased from the genome, and become pseudogenes. It has been estimating that around 60-80% of copies are lost after a whole genome duplication (Brunet, Crollius et al. 2006; Semon and Wolfe 2007). But what is the fate of genes which stay in two copies in the genome? The first predicted fate, of special interest for studies of positive selection, is **neofunctionalization (NF)** (Ohno 1970; Force, Lynch et al. 1999). In its simplest version, one copy will keep the ancestral function, and the other will rapidly acquire mutations during a period a relaxation of selection. These mutations at strategic positions will promote a new function. Function is a sometimes ambiguous term, and can be defined notably as the biochemical function of the protein, its interaction partners, or the spatio-temporal expression of the gene. This model invokes positive selection to fix advantageous mutations (Kondrashov and Kondrashov 2006; Shiu, Byrnes et al. 2006). An alternative to neofunctionalization is **subfunctionalization (SF)**. It assumes that the ancestral gene has several functions. After the duplication event, each copy will loss one or more functional parts and these genes will be complementary. For example, the developmental gene *Xhox3* of the Xenopus frog is expressed in the tail, the analia–genitalia and the nervous system. In the Zebrafish, there are two genes homologous to *Xhox3:* the gene *evx1*, which is expressed in the analia–genitalia and the nervous system, and its duplicate *eve1,* which is expressed in the tail and during gastrulation (Avaron, Thaeron-Antono et al. 2003). The best known version of this model is the Duplication-Degeneration-Complementation (DDC) model (Force, Lynch et al. 1999). Importantly, it does not involve positive selection, *in contrario* to the neofunctionalization model. The **sub-neo-functionalization (SNF)** model (He and Zhang 2005) is a mix between the two others models, in which duplicated genes may have a rapid subfunctionalization step followed by a longer neofunctionalization step. Other models have also been proposed but we will not describe them here (Conant and Wolfe 2008). We just would like to mention that it is usually difficult to attribute the evolutionary pattern of a pair of genes exactly to one model.

We can now use whole genome sequences to test hypotheses related to the occurrence of positive selection, such as what makes the differences between chimpanzee and us, or the level of implication of positive selection in the preservation of duplicate genes. We will present in this chapter different analyses of positive selection across evolution of genomes. The emphasis will be on problems of scaling codon based models; excellent discussion of methods for single gene studies, and/or polymorphism based methods, can be found elsewhere (Eyre-Walker 2006; Anisimova and Liberles 2007).

2 Which codon model for which problem?

The selective pressure, which occurred during the divergence between two homologous genes, can be measured at the nucleotide level by computing the *dN/dS* ratio (ω). *dN* is defined as the number of non-synonymous mutations per non-synonymous site, and indicates in first approximation the substitutions which are generated by mutation and fixed by a combination of drift and selection (on the function of the protein). *dS* is the number of synonymous mutations per synonymous site, and indicates in first approximation the substitutions which are generated by mutation and fixed by drift alone. If a gene is under purifying selection for the function of the encoded protein, this ω ratio is expected to be lower than 1, since most amino acid changes will be rejected; this is the case for most genes. If a gene is evolving without selective constraints (neutral evolution), the ω ratio is expected to be equal to 1, with no impact of either synonymous nor amino acid changes on fitness. Finally, if there is positive selection to change the structure or function of the encoded protein, it is possible that ω be higher than 1: amino acid changes are selected for their new role in the protein and thus kept in the genome more frequently than expected under mutation + drift. These expectations mean that ω can be used to estimate the direction and intensity of selection on protein coding genes.

The most popular package for such an estimation of evolutionary pressure is PAML (Phylogenetic Analysis by Maximum Likelihood) (Yang 1997; Yang 2007). For an overview of the other methods, we refer to Anisimova and Liberles (2007). In the following, we present the main codon models of the PAML package.

2.1 *Pair-wise estimate of dN/dS*

The pair-wise measure consists in simply estimating the *dN* and *dS* values between two genes. This method is interesting for genes that are closely related, when no additional information is available. This is typically the case when comparing orthologs from two closely related genomes. But it can be problematic when measuring more divergent genes. Notably this is a risk of saturation in the estimate of *dS*, when a synonymous site has multiple substitutions. We discuss this in section 3.3.

2.2 *Branch models*

The "branch models" (Yang 1998) were the first codon models to be implemented in PAML. They take advantage of a multiple sequence alignment to estimate *dN* and *dS* on specific branches using a mixture of Markov process and maximum likelihood inference. Depending on the model used, it can assign different *dN/dS* values to all branches, or to a few categories of branches. These models are useful to detect genes that have undergone strong positive selection, which has modified deeply and rapidly the amino acid sequences. The basic model is the Model 0, or "one-ratio" model. It assumes only one *dN/dS* for all branches of the tree, but the *dN* and *dS* could vary for each branch. At the opposite, the “free-ratio” model estimates one *dN/dS* for each branch. It is in principle useful when we have no a-priori hypothesis, but the result should be taken with caution, due to wide number of free parameters. Between these two extremities, the “two-or-more ratios” models can be used to test an a-priori hypothesis, by specifying the branch or branches that are thought to be under positive selection. We will obtain one *dN/dS* for the branch(es) of interest, and another for all other branches. We can then construct a likelihood ratio test (LRT) by contrasting this model against a simpler model, such as the one-ratio model. This test is more precise than the free ratio model. But one potential disadvantage in all the "branch models" is that *dN/dS* values are averaged over all positions in the alignment.

2.3 *Site models*

It is obvious that not all sites in a protein are equivalent, and it may not be a reasonable assumption to expect positive selection to act on all the protein sequence. Thus the second type of codon models are the "site models" (Yang, Nielsen et al. 2000). These models are quite useful to detect specific amino acids that are continuously under positive selection in a gene family. This is expected to be the case in arm-races such as experience by HIV proteins (Yang, Nielsen et al. 2000)or the MHC (Yang and Swanson 2002), or in sexual conflict, as identified in the mollusk Abalone sperm lysin (Yang, Swanson et al. 2000). The classic usage of these models is to perform an LRT contrasting the positive selection model M2a, which assigns sites into three different *dN/dS* classes ($\omega_0<1$, $\omega_1=1$ and $\omega_2\geq1$), to a nearly-neutral model M1a that assigns sites into only two classes ($\omega_0<1$, $\omega_1=1$). Other LRTs can be constructed using model M3 (3 classes of sites under no constraints (ω_0, ω_1, ω_2)) vs. M0 (1 class of site with ω_0), or model M8 (10 classes + 1 class $\omega\geq1$) vs. model M7 (10 classes, no positive selection allowed). These models need at least 6 sequences to be reasonably powerful (see Sect. 3). Moreover, they will only detect positive selection that acts over long periods of time (relative to the sequence sampling). In practice, this makes the corresponding tests very conservative, and adapted only to detecting the most extreme examples of positive selection, such as HIV proteins.

2.4 *Branch-site models*

The most recent class of models are “branch-site” models (Yang and Nielsen 2002; Zhang, Nielsen et al. 2005). They present an interesting mix of the two previous models. These models allow an estimation of the proportion of sites under positive selection (if any) during a specific evolutionary time (determined by a specific phylogenetic branch). This model is intuitively appealing, as positive selection could be expected to affect only some sites, during a limited time of functional change (Fig. 1). It is also consistent with theoretical expectations from realistic models of molecular evolution, such as the model of episodic selection (Gillespie 1991). If only a few substitutions occurred during the change of function, a simple branch model will fail to detect them, because the *dN/dS* per branch is averaged over all positions, and most of these remain under purifying selection. Thus branch-site models are best suited to identifying mutations that fine tune proteins, as in the case of adaptation of plant photosynthesis with the optimization of the RubisCO enzyme (Christin, Salamin et al. 2008).

In branch-site models, a branch of interest must be defined as the foreground, while all the other branches are defined as background. Positive selection is excluded on the background branches, while it may be allowed on the foreground. The branch-site model A will estimate three different *dN/dS* ratios ($\omega_{0,}$ ω_1 and ω_2), and assign sites into four different classes: class K0 sites are under purifying selection ($0\leq\omega_0\leq1$) on all branches (foreground and background); class K1 sites are under neutral evolution ($\omega_1=1$) on all branches; class K2a sites may be under positive selection ($\omega_2\geq1$) on the foreground branch, but under purifying selection ($0\leq\omega_0\leq1$) on background branches; and finally class K2b sites may be under positive selection ($\omega_2\geq1$) on the foreground branch, but under neutral evolution ($\omega_1=1$) on background branches.

There are two different likelihood ratio-tests (LRT) to infer the significance of this model. The original version compared the positive selection model A against the nearly-neutral site model M1a (Yang and Nielsen 2002). In this case, relaxation of purifying selection on the foreground branch could be wrongly interpreted as significant positive selection (Zhang 2004). The improved version (Zhang, Nielsen et al. 2005) contrasts the branch-site model A with positive selection ($\omega_2\geq1$) against a constrained branch-site model A where sites can be only under purifying or neutral evolution (ω_2 fixed to 1). Thus the LRT is significant only if positive selection on the foreground branch is a better explanation of the data than a possible relaxation of purifying selection on the foreground branch. When the test is significant, a Bayes Emipircal Bayes (BEB) prediction (Yang, Wong et al. 2005) can identify sites under positive selection, according to their posterior probability (PP) (Fig. 1). Of note, positive selection may be supported by the LRT even in cases where the BEB has insufficient power to predict specific sites.

3 Issues in deep and large-scale analysis

3.1 *Sampling*

Different problems could appear when analyzing large datasets of genes using codon models, but the main one is probably sequence sampling. Simulations suggest a minimum number of six sequences in the alignment to have enough power and accuracy (Anisimova, Bielawski et al. 2001; Anisimova, Bielawski et al. 2002; Anisimova and Yang 2007). This is in itself a major issue for genomic studies involving too few species (e.g. two or three primates). Moreover, sequences that are too close will not contain enough information for reliable estimation of *dN* and *dS*, while sequences which are too divergent may be difficult to align (Sect. 3.2) or have issues of saturation of *dS* (Sect. 3.3). Good sampling can help resolve the latter problem (Sect. 3.3).

3.2 *Alignment quality.*

The quality of the multiple sequences alignment is critical, as in many comparative analyses (e.g. phylogeny, molecular modeling). If the alignment is of poor quality, the final result will also be of poor quality ("Garbage In, Garbage Out") (Landan and Graur 2007; Wong, Suchard et al. 2008). In most phylogenetic software, any column with at least a gap will be removed from the alignment, mainly because interpretation of gaps in evolutionary context is still poorly understood. But the residues immediately surrounding gaps are often difficult to align. The GBLOCK method (Castresana 2000) has been developed to extract the best parts of a multiple sequence alignment, based on gap patterns, and can be used automatically in large scale analyses.

3.3 *Saturation of dS*

When sequence divergence increases, so does the probability that each synonymous site have undergone multiple substitutions. This leads to the classical issue in molecular evolution of saturation, whence it can become difficult or impossible to estimate the number of substitutions. Pairwise analysis of sequence is very sensitive to this issue. But, like in phylogeny, an appropriate sampling scheme can "break" long branches, and improve the estimation of *dS* (and *dN*) values. For example in a Putative RNA methylase family, the pairwise method of PAML will estimate a dS of 0.43 between the Human gene ENSG00000066651 and its Mouse ortholog ENSMUSG00000019792, which is below saturation. But with the longer divergence between this Human gene and its Zebrafish ortholog ENSDARG00000040033, the pairwise dS is estimated at 13, clearly saturated. Using the corresponding Ensembl gene tree adds 13 orthologs to these two genes. Then the one-ratio branch model estimates a maximum *dS* value of 1.5 for any one branch, and a total of 3.4 if we sum the *dS* values of all branches separating the two original genes (Fig. 2). While distance based estimates of *dS* may saturate relatively early ($dS \geq 1$), simulations indicate that maximum likelihood methods are more robust against saturation problems (Anisimova, Bielawski et al. 2001).

3.4 *False Discovery Rate*

A large-scale scan for positive selection will encounter the same problem as any other large scan, namely test repetition. Both testing different branches for one gene tree, and testing many different genes, may lead to false positive results. A simulation study suggested that the q-value (Storey and Tibshirani 2003) provides a good compromise between power and specificity in the case of multiple testing in one gene tree (Anisimova and Yang 2007). This method evaluates the proportion of false positives according to the global distribution of all p-values (Storey and Tibshirani 2003). Since the q-value is also a method of choice for genomic scans (indeed was developed for genomic data), it should also be appropriate for the second issue, of testing many genes. Indeed, our simulations found that q-value correction over p-values from multiple branches and multiple genes was both powerful and specific (Studer, Penel et al. 2008).

4 Large scale studies

We present here some selected whole genome scans for positive selection. More examples are reviewed elsewhere (Biswas and Akey 2006; Eyre-Walker 2006).

4.1 General scans for positive selection

One of the first large scale scans for positive selection was performed by Endo et al (1996). They scanned the DDBJ/EMBL/Genbank database and classified more than 24000 sequences into 3595 groups of homologous genes. They used pairwise computations of *dN/dS* to estimate the level of positive selection and they found 17 (0.48%) groups susceptible to be under positive selection, of which 9 are proteins from parasites or viruses. They also used a window analysis to find regions under selection in each gene. This study was made before the rise of more sophisticated codon model. It should be noted that problems of false positives have recently been reported for the sliding window approach (Schmid and Yang 2008).

A more recent study searched for genes evolving under positive selection in *Escherichia coli* (Petersen, Bollback et al. 2007). The authors used site models (Nielsen and Yang 1998) to scan 3757 genes from strain K12 with at least two other orthologs in other strains. They found positive selection is present in eight gene categories, based on the EcoCyc Database, especially in genes encoding cell surface proteins. High incidence of positive selection in E. coli is consistent with expectations from large population size.

4.2 From human - chimpanzee comparisons to a study of vertebrates

In recent years, several scans of positive selection have been performed with a special focus on the human and other primate lineages. Most recently, these approaches have been generalized to mammalian and vertebrate gene trees.

The simplest approach is to perform a pairwise *dN/dS* analysis between all orthologs of two genomes. Thus Nielsen et al. (2005) scanned more than 13,000 genes between Human and Chimpanzee. The aim was to identify genes that experimented positive selection in either or both lineages; 733 genes were reported, but only 35 have a p-value under 5%. They found over-representation of immune-defense-related genes and sensory perception, as well as genes on the X chromosome; and under-representation of genes expressed in the brain.

Tests for positive selection can be made directional by including a third genome, but usually at the cost of analyzing less genes. Clark et al. (2003) searched for positive selection between Human and Chimpanzee, using the Mouse as outgroup. They analyzed 7645 ortholog groups of Human-Chimpanzee-Mouse. They used two models: the branch model (Yang 1998) and the original version of the branch-site model (Yang and Nielsen 2002). They checked for a number of potential biases, such as GC content, repeat density, local recombination and segmental duplication. The scan identified positive selection on the human branch for 1547 genes, and on the chimpanzee branch for 1534 genes. They found an overrepresentation of genes present in OMIM (Hamosh, Scott et al. 2005) , as well as an overrepresentation of olfactory genes in human. Interestingly, some genes involved in speech were found under positive selection in human. More surprisingly, they found genes involved in the metabolism of amino acids. This has been interpreted as linked to a change in diet, since humans but not chimpanzees eat meat. Finally, Clark et al. (2003) propose that, as suggested by King and Wilson (1975), most difference between Chimpanzees and Humans may be due to regulatory changes. Jorgensen et al. (2005) used the three different types of codon model (Branch, site and branch-site (original version)) to infer positive selection in 1120 trios of human-mouse-pig genes. They found only 1 gene with the branch model, but 3.0% of genes in the human lineage show positive selection with the branch-site model, relatively to 2.0% for pig and 2.2% for mouse. However, Jorgensen et al. (2005) are cautious due to the small taxon sampling in their study, and suggest using it only as a first step for further analyses. The most recent gene trio study scanned 13 888 genes using the Macaque as an outgroup to identify branches leading to Human or Chimpanzee lineages (Bakewell, Shi et al. 2007). A notable advance over previous studies is the use of the improved branch-site model (Zhang, Nielsen et al. 2005). They found more genes with positive selection in chimpanzee evolution than in human evolution, and indeed after correcting for multiple testing identified only 2 genes in the human lineage (59 in the chimpanzee lineage). Even so, it has been suggested that the small sequence sampling used induced false-positives (Suzuki 2008).

The recent availability of multiple genome sequences has allowed for studies which combine a large number of genes with a phylogenetic framework. Arbiza et al. (2006) thus investigated more than 13,000 genes with orthologs in Human, Chimpanzee, Mouse, Rat and Dog, to estimate positive selection, relaxation, or evolutionary rate acceleration in the human lineage. They used Relative-Rate Tests (Robinson-Rechavi and Huchon 2000) and Branch-site models (original and improved versions) to estimate divergence between human and chimpanzee genes. They found positive selection associated with Gene Ontology terms such as sensory perception, GPCR signaling pathway, immune response, DNA/RNA metabolism and transcription. An investigation in the mammalian lineage used the power of six different species to analyze approximately 16,500 human genes, with at least two orthologs in another mammalian (Kosiol, Vinar et al. 2008). All duplicated genes were excluded. All the six species are represented in 42% of gene families. They used the site model to look over the entire tree, and the branch-site model (modification of the improved version) to infer lineage specific positive selection. Again, they found over-representation in processes of immunity/defense and sensory perception. They found also several pathways containing large numbers of gene under positive selection, such as the complement immunity system and the FAS/p53 apoptotic pathway. Some differences were found between primates and rodents: perception is over-represented in primates, and immunity in rodents. The analysis of microarray expression data indicated that genes under positive selection are less expressed and more tissue-specific than other genes. A more focused study investigated the molecular cause of species differences in disease (Vamathevan, Hasan et al. 2008). In this study, they analysed 3079 orthologs genes of the same five mammalian genomes as Arbiza et al. (2006), with a special focus on human and primate diseases. They first looked for positive selection using the free-ratio model in order to have an overview of the evolutionary rate. They found an ω between 0.14 to 0.20, depending the branch tested. Using the branch-site model, they found 511 genes under positive selection, with an excess in the chimpanzee branch (162 in chimp. vs 52 in human), as seen in Bakewell et al. (2007). They confirmed overrepresentation in nucleic acid metabolism, neuronal activities and immunity/defense. Of special interest, genes under positive selection are more likely to be present in OMIM. Finally genes under positive selection appear to interact more often with other positively selected genes than expected by chance. Finally, we have investigated positive selection using the improved branch-site test in 884 gene trees, including at least four mammals, Chicken, Xenopus, and five fishes (Studer, Penel et al. 2008). We found evidence for positive selection, after correcting for multiple testing, in a surprising 77% of gene trees. This appears to be due to increased power of the test with more species and greater divergence between sequences. We discuss this study more in detail in the next section.

4.3 *What is the effect of genome duplication on the incidence of positive selection?*

As discussed in Section 1, positive selection has been suggested to be important in the evolution of duplicated genes. To test this, we studied positive selection in vertebrate gene trees, which were constrained to include only specific duplication patterns. During vertebrate evolution, three different rounds of whole genome duplication occurred: two rounds before the split tetrapodes-Teleost fishes (known as "2R" (Putnam, Butts et al. 2008)), and a third before the diversification of Teleost fishes ("3R" or "FSGD" (Jaillon, Aury et al. 2004)). The availability of many vertebrate sequenced genomes, especially the five Teleost fishes provides us with enough power to test the hypothesis of positive selection as a retention mode of duplicated genes. We found that positive selection is pervasive along the vertebrate tree, but is surprisingly independent of the evolutionary event, speciation or duplication (Studer, Penel et al. 2008).

To perform a large-scale analysis with constraints on gene phylogeny, we used the database HomolEns version 3 (http://pbil.univ-lyon1.fr/databases/homolens.html), based on Ensembl release 41 (Oct 2006). HomolEns is built on the same model as Hovergen (Duret, Mouchiroud et al. 1994) or Hobacgen (Perriere, Duret et al. 2000). All predicted peptides are compared to each other with BLASTP2 (BLOSSUM62 as substitution matrix and e-value cut-off = 10-4), followed by a transitive clustering of genes (cluster size varies from 1 to more than 1000). For each family, a multiple sequences alignment is built with MUSCLE (Edgar 2004) and a maximum-likelihood phylogenetic tree is estimated, on conserved blocks of the alignments selected with GBLOCKS, with PhyML (substitution model = JTT, estimated proportion of invariable sites, 4 categories, estimated gamma, initial tree with BIONJ) (Guindon and Gascuel 2003). For each family, tree reconciliation is performed using a taxonomic reference tree (Dufayard, Duret et al. 2005; Kuzniar, van Ham et al. 2008). Each node is a defined either as a speciation event (separating orthologs) or a duplication event (separating paralogs). Importantly, the query tool FamFetch includes the TreePattern editor (Dufayard, Duret et al. 2005), which allows searching for families according to specific tree topologies. In our study, we used three different topologies (Fig. 3), which all share strong constraints on species sampling and phylogenetic relations among vertebrates: (a) gene trees where no retention of duplicates is allowed, after the fish specific genome duplication or otherwise ("singleton") and (b) and (c) gene trees where duplicated genes after the FSGD are retained in all fishes, forbidding or not other duplications in the tree. The genes with retention after FSGD provide the dataset to test the impact of duplication, and the singleton genes provide a control in the absence of duplication. In addition, we performed a search for genes retained in duplicate after 2R, to control for the eventual impact of this event on evolution of the vertebrate genes. It should be noted that our stringent selection procedure enriched our dataset in basic cell processes, which are usually under-represented in reports of positive selection.

For all our analyses, we used CodeML 3.15 from the PAML package (Yang 1997). We first tried the **"branch model"** to estimate *dN/dS* among branches. As we said before, this model is interesting to detect branches under strong positive selection, but suffers from two major problems in regards of our dataset: the number of parameters to estimate (one *dN/dS* per branch) and the long evolutionary time. Taken together, these problems generated important convergence problems. We next tried the **"site model"** to estimate the selective pressure acting on specific sites. However we found no significant results. Although this model can be useful on very fast evolving genes such as MHC or HIV proteins, it fails to detect weaker or transient positive selection on genes, which mostly evolve under strong purifying selection. Thus neither of these models was used in the final analysis. Finally, we implemented the improved **"branch-site"** model (Zhang, Nielsen et al. 2005) in a bioinformatic pipeline for large gene trees. Of note, the original branch-site model was used in one of the rare studies of ancient positive selection, on 2R duplication in the Troponin C gene family (Bielawski and Yang 2003). We corrected for test repetition using the q-value method (Storey and Tibshirani 2003), with q = 10%. A specificity of results from the improved branch-site test is the bimodal distribution of p-values, due to many cases in which the best fit of the alternative model is identical to the null model, and thus p = 1. Because of this, we use the *bootstrap* option to evaluate π_0 in the R package QVALUE.

The first striking result is that more than one third of FSGD duplicated genes (36%) have experienced positive selection shortly after duplication. Although this seems a high figure, it should be noted that (i) positive selection only concerns on average 3.0% of sites, and (ii) we lack an expectation of the amount of positive selection on such genes, with this method and sampling. To solve the second point, we focused on others vertebrate branches. Surprisingly, in the singleton dataset, we found larger proportions of genes having experienced positive selection (Fig. 4). As reported above, we found that on our total dataset (884 families), 77% of gene families have at least one branch with significant results. It must be noted again that most positive selection concerns a few sites, depending on the branch tested: only between 0.9% to 4.7% of sites are affected by positive selection on average. It is probable that this low percentage, over one branch, would be not detected by a simpler evolutionary model.

We checked further the influence of whole genome duplication. Using Wilcoxon tests on different model parameters, no significant differences in the incidence of positive selection are found comparing either Fish duplicate branches against others in the same gene tree, FSGD topology genes against singleton genes, or families with vertebrate 2R detected against no 2R detected.

As our dataset is based on specific patterns, the multiple alignments are generally good, due the tree constraints we imposed. To remove any doubt, we also performed additional analyses with GBLOCK and with MAFFT (Katoh, Kuma et al. 2005). In all cases, the results were consistent, with small differences between alignment methods.

As said previously, a common criticism in analyses involving estimation *dN* and *dS* is the possibility of *dS* saturation. To ensure this is not the case in our study, we performed simulations using the same global parameters as the real data for each gene: number of sequences, sequence length, tree topology, branch lengths, *dN/dS* ration ω, transition/transversion ratio κ, and codon usage. This procedure guaranties that the simulated dataset has the same distribution of parameters as the real data set, including potential confounding factors such as codon usage bias or long branches. These results and those of Anisimova et al. (Anisimova, Bielawski et al. 2002; Anisimova and Yang 2007) showed that the maximum likelihood estimate is robust to *dS* saturation, even for large divergence as shown by our simulations with doubled branch lengths. This is probably due to the use of more sequences, which "break" the long branches of the gene tree.

Finally, the conclusion of our study is threefold: (i) positive selection affects diverse phylogenetic branches and diverse genes categories during vertebrate evolution; (ii) it concerns only a small proportion of sites (1%-5%); and (iii) whole genome duplication had no detectable incidence on the prevalence of this positive selection.

5 Selectome, a database of branch-site positive selection

Most of these different methods of positive selection are time consuming, and can be difficult to use, thus it is of interest to have precomputed results available in a database. One of the first such database is TAED (The Adaptive Evolution Database), which contains gene sequences, multiple alignment and trees from “higher plants” and chordates (Roth, Betts et al. 2005). For each branch of the tree, a *dN/dS* estimation is performed in a pairwise manner, using ancestral sequences reconstruction. Another database is the HumanPAML browser, which presents results of positive selection specific to human genes (Nickel, Tefft et al. 2008). The computation of *dN/dS* is performed in ~14 000 human genes with mammalian orthologs. Several models from PAML are used, including branch models, site models, and branch-site models (original and improved). When a branch has to be specified for the model, results are computed only for the human branch. In our study of vertebrate genomes (Studer, Penel et al. 2008), we developed a fully automatic procedure to detect positive selection in any branch of phylogenetic trees, using the branch-site test. The main core consists in formatting the multiple alignment sequences of nucleotides, preparing the tree file by specifying automatically each branch of interest (coded with a #1), launching CodeML and finally retrieving the data and statistically analyzing them. In that study, we focused only in some specific patterns of evolution. But it should be of interest to have an exhaustive catalog of branches under positive selection, using the same methodology. Thus, we then developed Selectome (Proux, Studer et al. 2008). We used an improved version of the previous bioinformatic pipeline to scan all vertebrates genes families in the TreeFam database (Ruan, Li et al. 2008). Release 1 focused only on manually curated part (TreeFam-A). The database, built in MySQL, can be explored through a web interface at http://bioinfo.unil.ch/selectome. The user can perform searches using different criteria, such as gene name, gene ID, keywords, or by selecting specific branches of the phylogeny, with or without duplication. The result is a display of relevant sub-families, with annotated phylogenetic trees. Duplicated nodes are colored in red and branches under positive selection are in green boxes, with the associated *p-value* indicated. The user can also view the protein multiple sequences alignment with the Jalview applet (Clamp, Cuff et al. 2004). For each significant branch, bars identify sites detected by BEB methods (Yang, Wong et al. 2005), color coded according to level of posterior value. We think this could be very useful not only in evolutionary biology, but also in molecular biology studies to identify e.g. candidates for site-directed mutagenesis.

6 Conclusion

In these studies presented here, different model have been used to infer evolutionary rate. An important point is that these models do not always detect the same pattern of positive selection. Pairwise estimation or branch models are useful to detect positive selection between two closely related genes, or in a specific lineage in well-sampled family. Site models are more specific to detect amino acids that make rapid adaptation to external factors. And the branch-site models can be used to detect ancient or specific adaptive mutational events. Another point to note in conclusion is that there is of necessity a trade-off between the number of genes analyzed, and phylogenetic sampling. Thus pairwise comparisons of closely related genomes will always include more genes than studies spanning many species. Obviously, the information from both types of studies are useful, and complete each other.

We have seen through different analyses that various classes of genes could under positive selection. The genes for arm-race, for sexuality, for sensorial perception are the most dominant, but positive selection can also be detected in housekeeping genes. Duplicated genes somewhat surprisingly do not appear to experience more positive selection than other genes. It will be of special interest in the future to be able to combine increasingly complex and realistic models of codon evolution (e.g. Mayrose, Doron-Faigenboim et al. 2007; Yang and Nielsen 2008) with large amounts of data and good species sampling.

7 References

Anisimova, M., J. P. Bielawski, et al. (2001). "Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution." Mol Biol Evol **18**(8): 1585-92.

Anisimova, M., J. P. Bielawski, et al. (2002). "Accuracy and power of bayes prediction of amino acid sites under positive selection." Mol Biol Evol **19**(6): 950-8.

Anisimova, M. and D. A. Liberles (2007). "The quest for natural selection in the age of comparative genomics." Heredity **99**(6): 567-79.

Anisimova, M. and Z. Yang (2007). "Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites." Mol Biol Evol **24**(5): 1219-28.

Arbiza, L., J. Dopazo, et al. (2006). "Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome." PLoS Comput Biol **2**(4): e38.

Avaron, F., C. Thaeron-Antono, et al. (2003). "Comparison of even-skipped related gene expression pattern in vertebrates shows an association between expression domain loss and modification of selective constraints on sequences." Evol Dev **5**(2): 145-56.

Bakewell, M. A., P. Shi, et al. (2007). "More genes underwent positive selection in chimpanzee evolution than in human evolution." Proc Natl Acad Sci U S A **104**(18): 7489-94.

Baum, J., R. H. Ward, et al. (2002). "Natural selection on the erythrocyte surface." Mol Biol Evol **19**(3): 223-9.

Bielawski, J. P. and Z. Yang (2003). "Maximum likelihood methods for detecting adaptive evolution after gene duplication." J Struct Funct Genomics **3**(1-4): 201-12.

Biswas, S. and J. M. Akey (2006). "Genomic insights into positive selection." Trends Genet **22**(8): 437-46.

Brunet, F. G., H. R. Crollius, et al. (2006). "Gene loss and evolutionary rates following whole-genome duplication in teleost fishes." Mol Biol Evol **23**(9): 1808-16.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Mol Biol Evol **17**(4): 540-52.

Christin, P. A., N. Salamin, et al. (2008). "Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis." Mol Biol Evol **25**(11): 2361-8.

Clamp, M., J. Cuff, et al. (2004). "The Jalview Java alignment editor." Bioinformatics **20**(3): 426-7.

Clark, A. G., S. Glanowski, et al. (2003). "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios." Science **302**(5652): 1960-3.

Conant, G. C. and K. H. Wolfe (2008). "Turning a hobby into a job: How duplicated genes find new functions." Nat Rev Genet **9**(12): 938-950.

Dufayard, J. F., L. Duret, et al. (2005). "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases." Bioinformatics **21**(11): 2596-603.

Duret, L., D. Mouchiroud, et al. (1994). "HOVERGEN: a database of homologous vertebrate genes." Nucleic Acids Res **22**(12): 2360-5.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-7.

Endo, T., K. Ikeo, et al. (1996). "Large-scale search for genes on which positive selection may operate." Mol Biol Evol **13**(5): 685-90.

Eyre-Walker, A. (2006). "The genomic rate of adaptive evolution." Trends Ecol Evol **21**(10): 569-75.

Force, A., M. Lynch, et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-45.

Gillespie, J. H. (1991). The Causes of Molecular Evolution, Oxford University Press, USA.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol **52**(5): 696-704.

Hamosh, A., A. F. Scott, et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucl. Acids Res. **33**(suppl_1): D514-517.

He, X. and J. Zhang (2005). "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution." Genetics **169**(2): 1157-64.

Hughes, A. L., M. K. Hughes, et al. (1994). "Natural selection at the class II major histocompatibility complex loci of mammals." Philos Trans R Soc Lond B Biol Sci **346**(1317): 359-66; discussion 366-7.

Hughes, A. L. and M. Nei (1988). "Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection." Nature **335**(6186): 167-70.

Hughes, A. L. and M. Nei (1989). "Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection." Proc Natl Acad Sci U S A **86**(3): 958-62.

Jaillon, O., J. M. Aury, et al. (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-57.

Jiggins, F. M., G. D. Hurst, et al. (2002). "Host-symbiont conflicts: positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae." Mol Biol Evol **19**(8): 1341-9.

Jorgensen, F. G., A. Hobolth, et al. (2005). "Comparative analysis of protein coding sequences from human, mouse and the domesticated pig." BMC Biol **3**: 2.

Katoh, K., K. Kuma, et al. (2005). "MAFFT version 5: improvement in accuracy of multiple sequence alignment." Nucleic Acids Res **33**(2): 511-8.

King, M. C. and A. C. Wilson (1975). "Evolution at two levels in humans and chimpanzees." Science **188**(4184): 107-16.

Kondrashov, F. A. and A. S. Kondrashov (2006). "Role of selection in fixation of gene duplications." J Theor Biol **239**(2): 141-51.

Kosiol, C., T. Vinar, et al. (2008). "Patterns of positive selection in six Mammalian genomes." PLoS Genet **4**(8): e1000144.

Kuzniar, A., R. C. van Ham, et al. (2008). "The quest for orthologs: finding the corresponding gene across genomes." Trends Genet **24**(11): 539-51.

Landan, G. and D. Graur (2007). "Heads or tails: a simple reliability check for multiple sequence alignments." Mol Biol Evol **24**(6): 1380-3.

Lee, Y. H., T. Ota, et al. (1995). "Positive selection is a general phenomenon in the evolution of abalone sperm lysin." Mol Biol Evol **12**(2): 231-8.

Mayrose, I., A. Doron-Faigenboim, et al. (2007). "Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates." Bioinformatics **23**(13): i319-27.

Messier, W. and C. B. Stewart (1997). "Episodic adaptive evolution of primate lysozymes." Nature **385**(6612): 151-4.

Moreno-Estrada, A., F. Casals, et al. (2008). "Signatures of selection in the human olfactory receptor OR5I1 gene." Mol Biol Evol **25**(1): 144-54.

Nickel, G. C., D. Tefft, et al. (2008). "Human PAML browser: a database of positive selection on human genes using phylogenetic methods." Nucleic Acids Res **36**(Database issue): D800-8.

Nielsen, R., C. Bustamante, et al. (2005). "A scan for positively selected genes in the genomes of humans and chimpanzees." PLoS Biol **3**(6): e170.

Nielsen, R. and Z. Yang (1998). "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene." Genetics **148**(3): 929-36.

Ohno, S. (1970). Evolution by Gene Duplication.

Perriere, G., L. Duret, et al. (2000). "HOBACGEN: database system for comparative genomics in bacteria." Genome Res **10**(3): 379-85.

Petersen, L., J. P. Bollback, et al. (2007). "Genes under positive selection in Escherichia coli." Genome Res **17**(9): 1336-43.

Proux, E., R. A. Studer, et al. (2008). "Selectome: a database of positive selection." Nucleic Acids Res **37**: D404-D407.

Putnam, N. H., T. Butts, et al. (2008). "The amphioxus genome and the evolution of the chordate karyotype." Nature **453**(7198): 1064-71.

Robinson-Rechavi, M. and D. Huchon (2000). "RRTree: Relative-Rate Tests between groups of sequences on a phylogenetic tree." Bioinformatics **16**(3): 296-297.

Rooney, A. P. and J. Zhang (1999). "Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection?" Mol Biol Evol **16**(5): 706-10.

Roth, C., M. J. Betts, et al. (2005). "The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics." Nucleic Acids Res **33**(Database issue): D495-7.

Ruan, J., H. Li, et al. (2008). "TreeFam: 2008 Update." Nucleic Acids Res **36**(Database issue): D735-40.

Sawyer, S. L., L. I. Wu, et al. (2005). "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain." Proc Natl Acad Sci U S A **102**(8): 2832-7.

Schmid, K. and Z. Yang (2008). "The Trouble with Sliding Windows and the Selective Pressure in BRCA1." PLoS ONE **3**(11): e3746.

Semon, M. and K. H. Wolfe (2007). "Consequences of genome duplication." Curr Opin Genet Dev **17**(6): 505-12.

Shiu, S. H., J. K. Byrnes, et al. (2006). "Role of positive selection in the retention of duplicate genes in mammalian genomes." Proc Natl Acad Sci U S A **103**(7): 2232-6.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A **100**(16): 9440-5.

Studer, R. A., S. Penel, et al. (2008). "Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes." Genome Res **18**(9): 1393-402.

Suzuki, Y. (2008). "False-positive results obtained from the branch-site test of positive selection." Genes Genet Syst **83**(4): 331-8.

Vamathevan, J. J., S. Hasan, et al. (2008). "The role of positive selection in determining the molecular cause of species differences in disease." BMC Evol Biol **8**: 273.

Wong, K. M., M. A. Suchard, et al. (2008). "Alignment uncertainty and genomic analysis." Science **319**(5862): 473-6.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-6.

Yang, Z. (1998). "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution." Mol Biol Evol **15**(5): 568-73.

Yang, Z. (2006). Computational Molecular Evolution, Oxford University Press, USA.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-91.

Yang, Z. and R. Nielsen (2002). "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages." Mol Biol Evol **19**(6): 908-17.

Yang, Z. and R. Nielsen (2008). "Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage." Mol Biol Evol **25**(3): 568-79.

Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-49.

Yang, Z. and W. J. Swanson (2002). "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes." Mol Biol Evol **19**(1): 49-57.

Yang, Z., W. J. Swanson, et al. (2000). "Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites." Mol Biol Evol **17**(10): 1446-55.

Yang, Z., W. S. Wong, et al. (2005). "Bayes empirical bayes inference of amino acid sites under positive selection." Mol Biol Evol **22**(4): 1107-18.

Zhang, J. (2004). "Frequent false detection of positive selection by the likelihood method with branch-site models." Mol Biol Evol **21**(7): 1332-9.

Zhang, J., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Mol Biol Evol **22**(12): 2472-9.

Zhang, J. Z. (2003). "Evolution by gene duplication: an update." Trends in Ecology & Evolution **18**(6): 292-298.

Zhang, Z. D., G. Weinstock, et al. (2008). "Rapid evolution by positive Darwinian selection in T-cell antigen CD4 in primates." J Mol Evol **66**(5): 446-56.

8 Figures

8.1 *Fig. 1*

The gene family "Protein kinase C and casein kinase substrate in neurons protein (SwisProt:PACN1_HUMAN)" (code HBG059468 in Homolens release 3) presents positive selection at three different evolutionary times, as revealed by the branch-site model of PAML (Yang and Nielsen 2002; Zhang, Nielsen et al. 2005). We highlighted three different sites: a) site 220 was selected for a Threonine in the Mammalia branch, b) site 238 for a Histidine in the longest branch in Fishes and c) in site 263 for a Serine in Vertebrates against a Glutamine in Fishes. These sites are above a cut-off of 95% in the Bayes Empirical Bayes analysis from CodeML (Yang, Wong et al. 2005).

8.2 *Fig. 2*

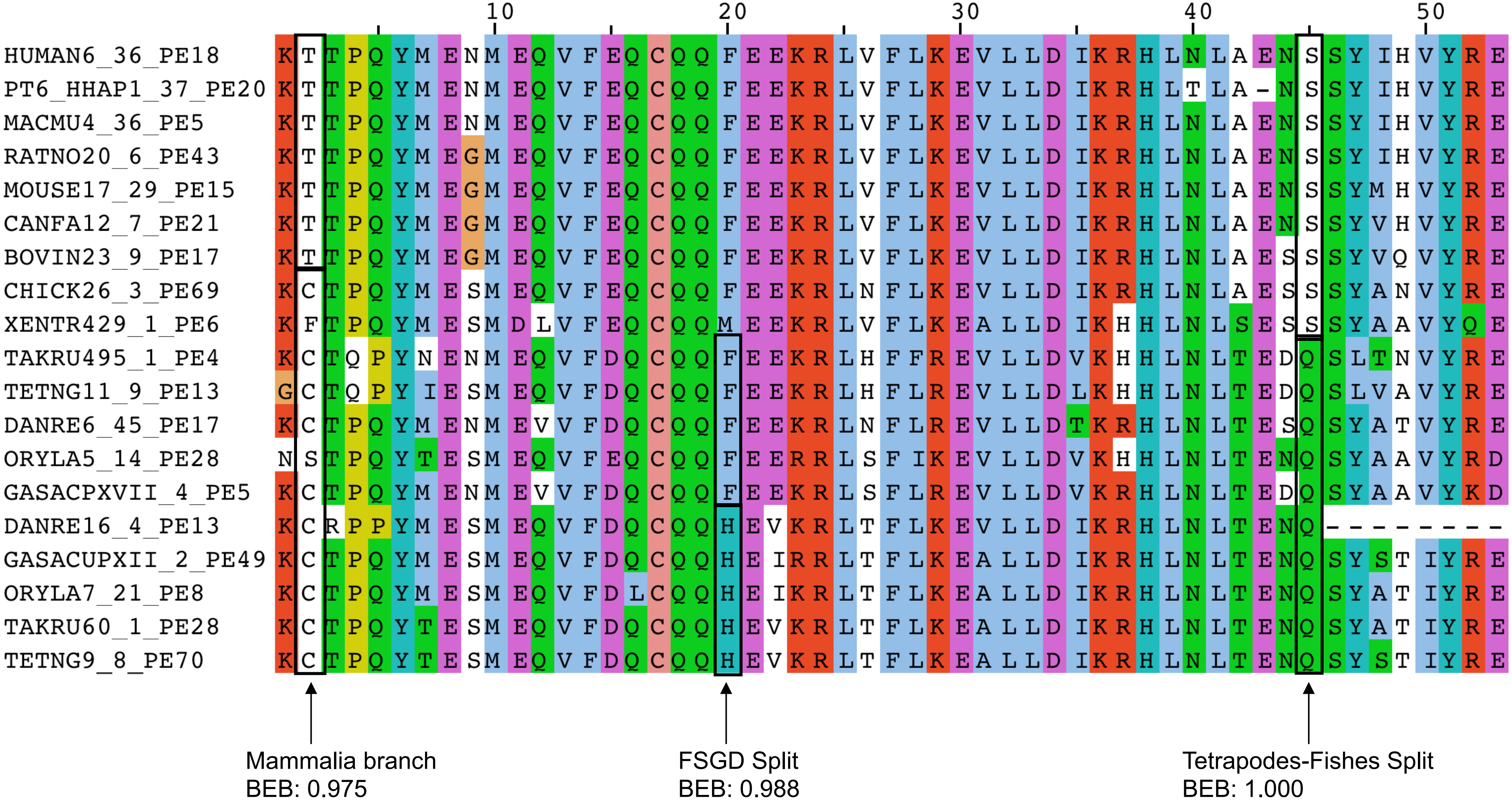
Phylogenetic tree of the putative RNA methylase family. The tree comes from Homolens release 3 (code: HBG000007). The pairwise analysis of Human and Zebrafish genes results in a *dS* of 13.37, which is saturated. Adding all others species and computing under the "one-ratio" model of PAML, the codon method takes advantage of breaking the branches and results in more precise *dS* estimation, with a maximum of 1.45 and a sum of 3.43 for all branches separated Zebrafish and Human branches.

8.3 *Fig. 3*

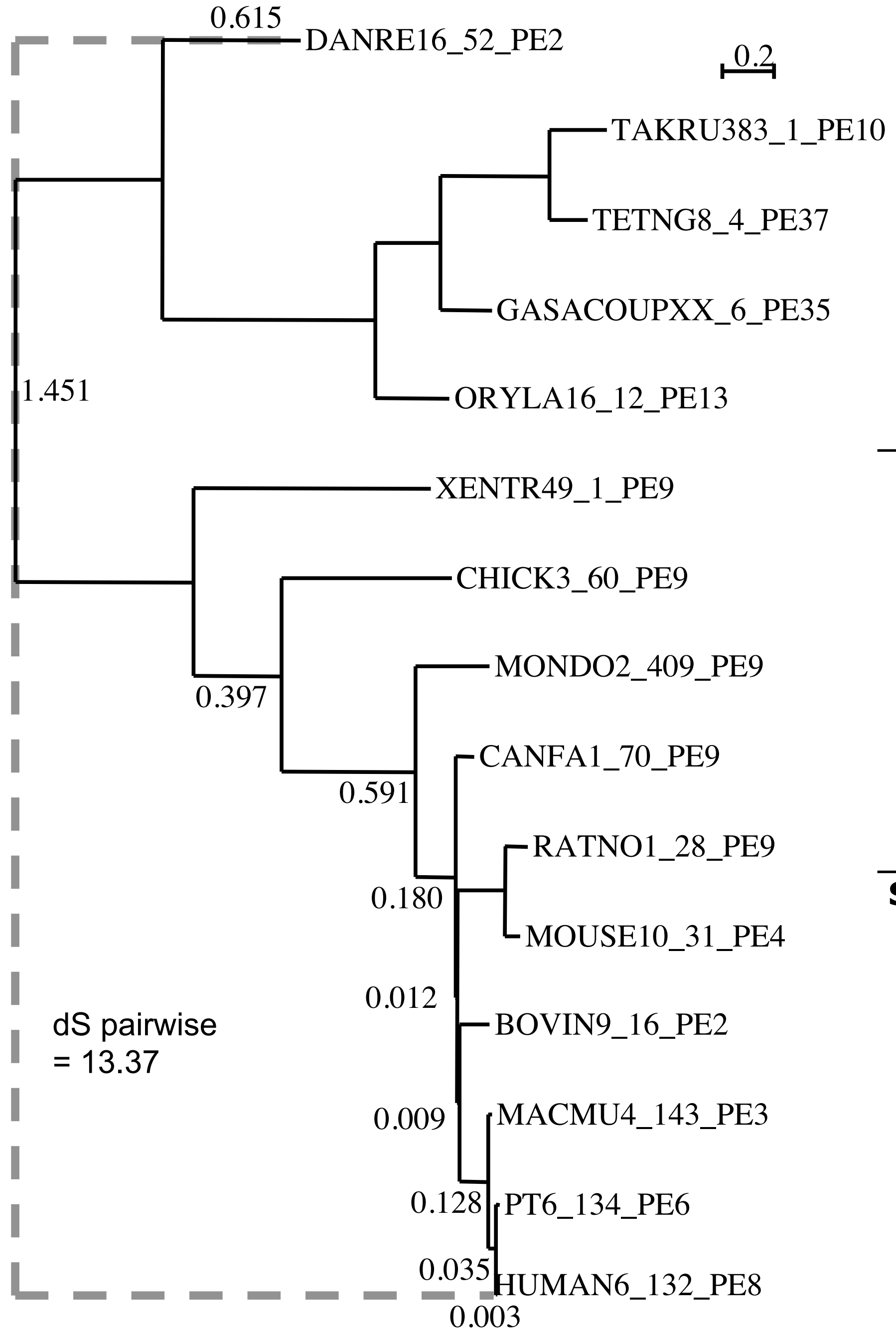
The different tree patterns used in our study. Diamonds represent speciation nodes and boxes represent duplication nodes. Clear branches have no constraint whereas duplication is forbidden on dark grey branches. (a) Singleton topology: no duplication allowed, the gene tree must follow the expected species tree. (b) Conservation of paralogs from the fish specific duplication enforced in all five fishes; no other duplication allowed, and the gene tree must follow the expected species tree. (c) Same as B, but other duplications are allowed, and there might be slight differences between the gene tree and the species tree.

8.4 *Fig. 4*

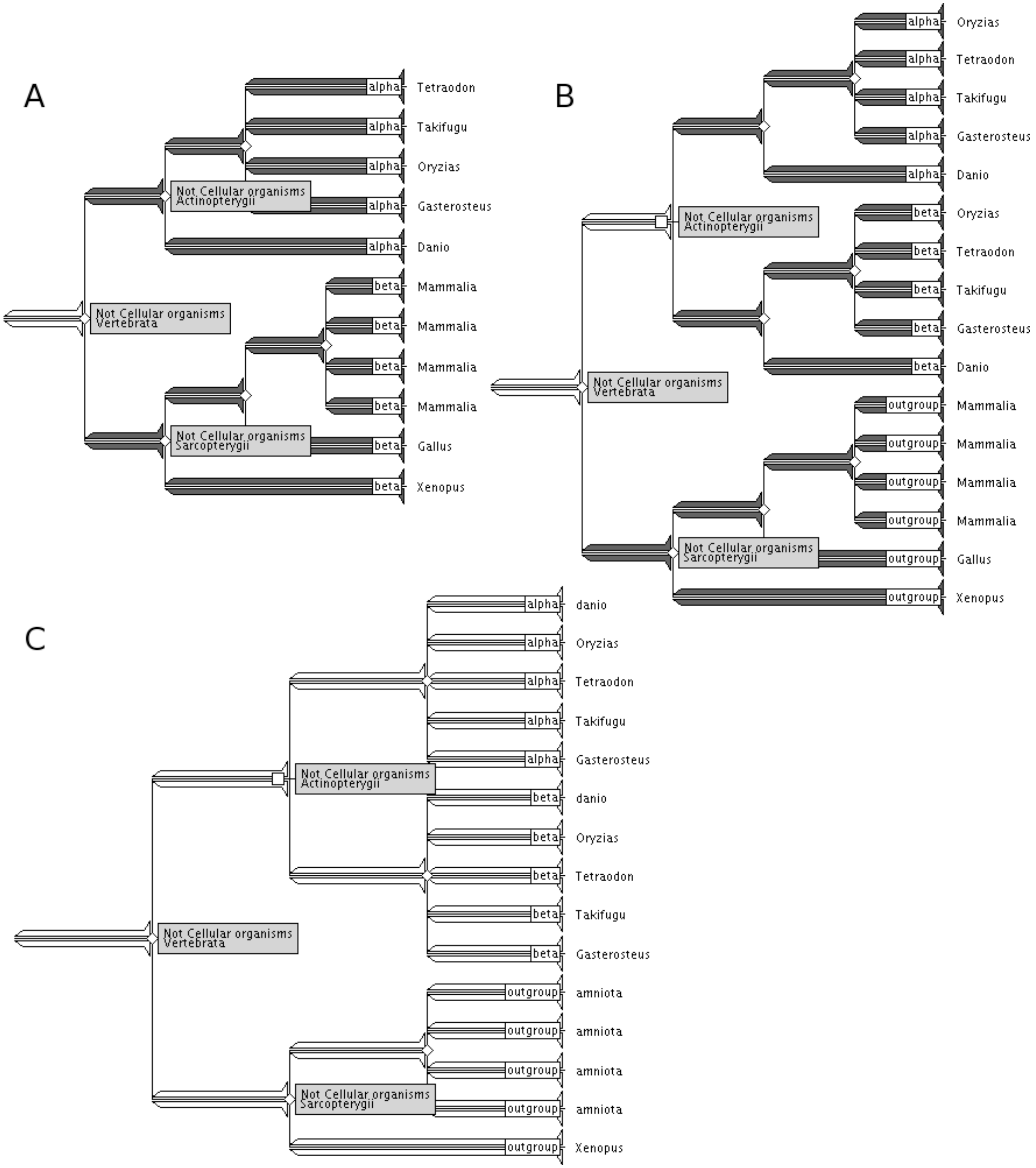
The taxonomic tree of species used in our analysis of positive selection (Studer, Penel et al. 2008). The numbers in brackets after lineage names are the number of species in taxa represented by more than one species. The doubled branches are the split between duplicates from the Fish Specific Genome Duplication. The speciation branches analyzed are in bold. The values on branches represent the proportion of cases where positive selection has been detected on a branch, with the mean proportion of sites involved in brackets.

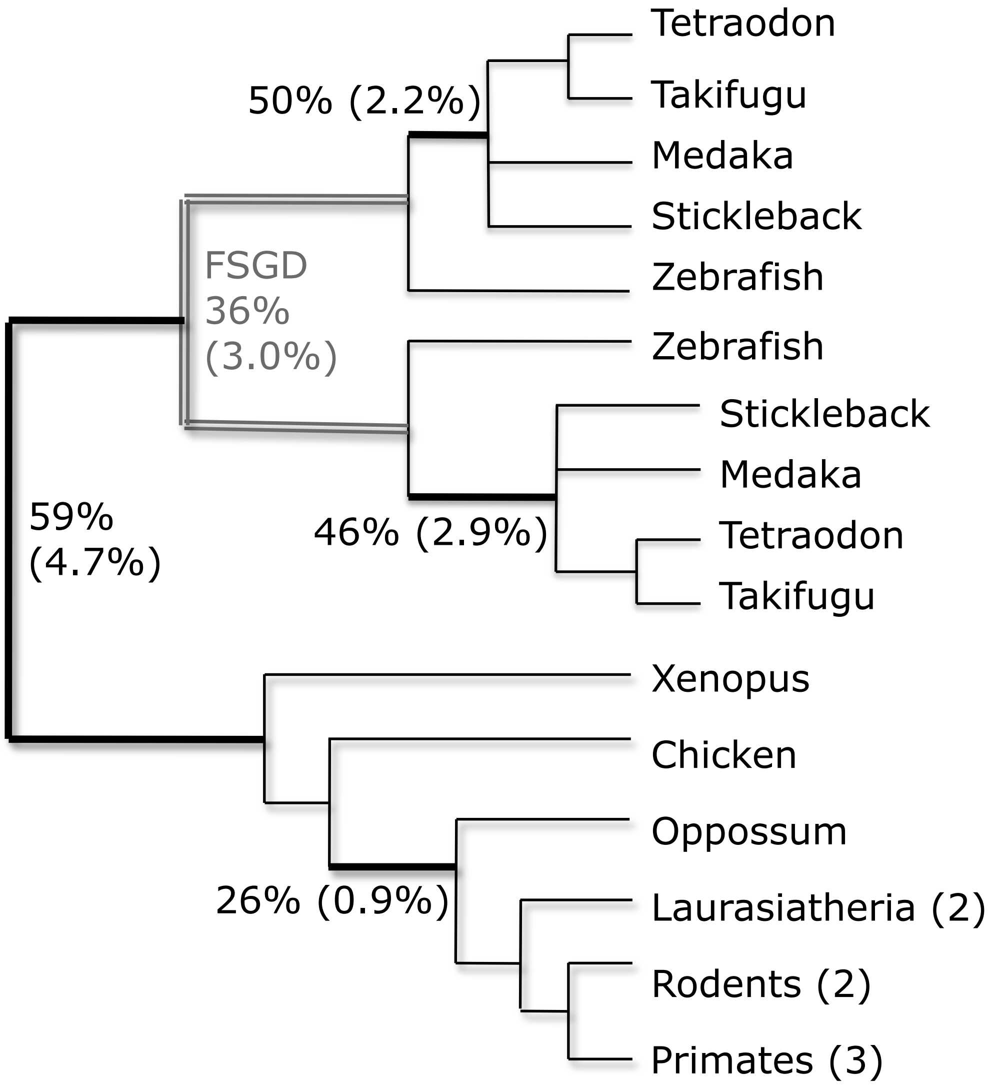


```
HUMAN6_36_PE18       KTTPQYMENMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLNLAENSSYIHVYRE
PT6_HHAP1_37_PE20    KTTPQYMENMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLTLA-NSSYIHVYRE
MACMU4_36_PE5        KTTPQYMENMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLNLAENSSYIHVYRE
RATNO20_6_PE43       KTTPQYMEGMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLNLAENSSYIHVYRE
MOUSE17_29_PE15      KTTPQYMEGMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLNLAENSSYMHVYRE
CANFA12_7_PE21       KTTPQYMEGMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLNLAENSSYVHVYRE
BOVIN23_9_PE17       KTTPQYMEGMEQVFEQCQQFEEKRLVFLKEVLLDIKRHLNLAESSSYVQVYRE
CHICK26_3_PE69       KCTPQYMESMEQVFEQCQQFEEKRLNFLKEVLLDIKRHLNLAESSSYANVYRE
XENTR429_1_PE6       KFTPQYMESMDLVFEQCQQMEEKRLVFLKEALLDIKHHLNLSESSSYAAVYQE
TAKRU495_1_PE4       KCTQPYNENMEQVFDQCQQFEEKRLHFFREVLLDVKHHLNLTEDQSLTNVYRE
TETNG11_9_PE13       GCTQPYIESMEQVFDQCQQFEEKRLHFLREVLLDLKHHLNLTEDQSLVAVYRE
DANRE6_45_PE17       KCTPQYMENMEVVFDQCQQFEEKRLNFLREVLLDTKRHLNLTESQSYATVYRE
ORYLA5_14_PE28       NSTPQYTESMEQVFEQCQQFEERRLSFIKEVLLDVKHHLNLTENQSYAAVYRD
GASACPXVII_4_PE5     KCTPQYMENMEVVFDQCQQFEEKRLSFLREVLLDVKRHLNLTEDQSYAAVYKD
DANRE16_4_PE13       KCRPPYMESMEQVFDQCQQHEVKRLTFLKEVLLDIKRHLNLTENQ--------
GASACUPXII_2_PE49    KCTPQYMESMEQVFDQCQQHEIRRLTFLKEALLDIKRHLNLTENQSYSTIYRE
ORYLA7_21_PE8        KCTPQYMESMEQVFDLCQQHEIKRLTFLKEALLDIKRHLNLTENQSYATIYRE
TAKRU60_1_PE28       KCTPQYTESMEQVFDQCQQHEVKRLTFLKEALLDIKRHLNLTENQSYATIYRE
TETNG9_8_PE70        KCTPQYTESMEQVFDQCQQHEVKRLTFLKEALLDIKRHLNLTENQSYSTIYRE
```



	branch	dS
Zebrafish	25..3	0.6157
	16..25	1.4515
	16..17	0.3979
	17..18	0.5915
	18..19	0.1808
	19..20	0.0124
	20..21	0.0091
	21..22	0.1285
	22..23	0.0350
Human	23..9	0.0037
Sum of dS		**3.4261**


A
Not Cellular organisms Vertebrata
Not Cellular organisms Actinopterygii
Not Cellular organisms Sarcopterygii
alpha Tetraodon
alpha Takifugu
alpha Oryzias
alpha Gasterosteus
alpha Danio
beta Mammalia
beta Mammalia
beta Mammalia
beta Mammalia
beta Gallus
beta Xenopus
B
Not Cellular organisms Vertebrata
Not Cellular organisms Actinopterygii
Not Cellular organisms Sarcopterygii
alpha Oryzias
alpha Tetraodon
alpha Takifugu
alpha Gasterosteus
alpha Danio
beta Oryzias
beta Tetraodon
beta Takifugu
beta Gasterosteus
beta Danio
outgroup Mammalia
outgroup Mammalia
outgroup Mammalia
outgroup Mammalia
outgroup Gallus
outgroup Xenopus
C
Not Cellular organisms Vertebrata
Not Cellular organisms Actinopterygii
Not Cellular organisms Sarcopterygii
alpha danio
alpha Oryzias
alpha Tetraodon
alpha Takifugu
alpha Gasterosteus
beta danio
beta Oryzias
beta Tetraodon
beta Takifugu
beta Gasterosteus
outgroup amniota
outgroup amniota
outgroup amniota
outgroup amniota
outgroup Xenopus


Tetraodon
Takifugu
Medaka
Stickleback
Zebrafish
50% (2.2%)
FSGD
36%
(3.0%)
Zebrafish
Stickleback
Medaka
Tetraodon
Takifugu
46% (2.9%)
59%
(4.7%)
Xenopus
Chicken
Oppossum
Laurasiatheria (2)
Rodents (2)
Primates (3)
26% (0.9%)