Phylogenetic footprinting, a pipeline for motif identification in *Paramecium*

Matthias Grenié[1], Jean-François Goût[2], Michael Lynch[2]
1 Départment de Biologie, École Normale Supérieure de Lyon, Lyon, France
2 Biology Department, Indiana University, IN, United States of America

Abstract

Whole-Genome Duplication (WGD) are widespread among the tree of life, in plants, animals or even Ciliates. Duplicate genes have different rate of retention after a WGD, some of them are fully conserved while other are lost. *Paramecium* is a great system to study WGD, especially because of the numerous genomes available. In addition, the *aurelia* complex is a cryptic species complex, whom species experienced two rounds of WGDs. Recently, retention rate was correlated to duplicate retention in *Paramecium*, the higher expressed a gene was, the greater chances of retention he had. Thus, trying to understand the links between cis-regulatory sequences and gene expression would let us eventually understand better gene retention. Here we used predefined orthology and paralogy gene families (McGrath et al., 2014) to identify conserved motifs in upstream regions of genes in *Paramecium caudatum*, *P. biaurelia*, *P. sexaurelia* and *P. biaurelia*. We developed a phylogenetic approach with BigFoot (Satija et al., 2009), complemented by analyses with MEME (Bailey et al., 2006). We identified more than 117 motifs conserved among all species. We found that 10 of them were present in the 4 species at various rates. *P. caudatum* seemed to have a smaller of ratio of gene containing conserved motifs. We still have to improve our analysis pipeline to identify better candidates and suppress redundancy in found motifs.

Introduction

Since Ohno first hypothesized the influence of Whole-Genome Duplications (WGD) (Ohno, 1970), scientists kept showing that a broad number of organisms experienced WGD: *Saccharomyces cerevisiae* (Kellis et al., 2004), Angiosperms (Arrigo and Barker, 2012), Vertebrates (Dehal and Boore, 2005), Salmonids (Alexandrou et al., 2013), and many others. WGD are evolutionary event when the genome of a given individual is duplicated, meaning that the whole genome is in two copies, duplicated pairs of genes are called *paralogs*. WGD can also occur after two closely related species hybridize to avoid hybrid incompatibility issues. WGDs may be involved in many evolutionary radiations (Alexandrou et al., 2013) as they provide the raw material to explore new evolutionary landscapes.

Since Ohno numerous models have been developed to explain the retention rate we observe between duplicate genes (Chain et al. (2011), reviewed in Innan and Kondrashov (2010)). Gene Balance Hypothesis is for example a well explored hypothesis in the literature, according to this hypothesis dosage-sensitive gene are more retained because of stoichiometry issues, it would thus explain the over retention of transcription factors and multi-complexes proteins.

To understand the consequences of WGDs we have been studying various *Paramecium* species (Beisson et al., 2010). *Paramecium* are a Ciliates group (See Figure 1). Aury *et al.* showed that at least three round of WGDs occurred in the *Paramecium* genus (Aury et al., 2006), two of which occurred in the *aurelia* complex (see Figure 2). This cryptic species complex of 13 reproductively isolated species is a great model to study the fate of duplicate genes (Catania et al. 2009, McGrath et al. 2014).

We know that WGD retained duplicates are biased in functional category (Edger and Pires, 2009). But recently, several studies unraveled the link between gene expression in gene retention after WGD in *Paramecium* (Gout et al. 2010, Arnaiz et al. 2010, Chain et al. 2011). Genes that are highly expressed are more likely to be retained than genes with a low expression. The COSTEX model proposed by the authors states that gene evolution is more constrained as gene expression is high. It explains well why

highly expressed genes are more retained and why they evolve more slowly thant other genes.

Having established this link between gene retention and expression level of the gene, it is normal to try to get better understanding of *Paramecium* regulatory sequences. The compactness of *Paramecium* genomes makes the study of regulatory elements easier than in other eukaryotic species (McGrath et al., 2014). *Paramecium* genomes have on average 300nt long inter-genic regions close to the size of *Saccharomyces cerevisiae* ones (Chen et al. 2011 and Hahn and Young 2011). Some short inter-genic regions are known, in *S. cerevisiae*, to regulate genes directly downstream of them. Transcription regulator is thus far easier to study as for most gene its *cis*-regulatory sequence is directly upstream of it, as a rough approximation, and not hundreds of thousand bases away as it can be the case in mammalian genomes. Because *Paramecium* also have very short inter-genic regions, it is reasonable to assume that their promoters share the same mechanisms.

Various strategies to study regulatory elements evolution have been developed (Wittkopp and Kalay, 2012); the idea being that regulatory elements can be conserved along evolution, some tools can detect these more conserved regions and thus output putative motifs (D'haeseleer, 2006). Phylogenetic footprinting is one of the approaches developed when trying to detect motifs among several species (Zhang and Gerstein, 2003), the idea being that regulatory motifs tend to be conserved by purifying selection (Nelson and Wardle, 2013), while non-functional elements will accumulate mutation along evolution. The motifs are weighted by the phylogenetical relationships of the given sequences, if a motif is found in two closely related species it will have less weight than if it is detected between two distant species.

In this paper we develop a phylogenetic footprinting workflow to study the *cis*-regulatory elements among three *aurelia* species: *P. biaurelia*, *P. sexaurelia*, *P. tetraurelia* and a more basal species *P. caudatum*. The three *aurelia* experienced two rounds of WGD compared to *P. caudatum*, thus when studying orthology and paralogy groups between genes, they can contain up to 13 different genes: 1 from *P. caudatum* and 4 from each *aurelia* species. Using those groups we use our workflow to identify conserved motifs in the upstream regions in each group.

M. G. designed the pipeline and ran the analysis, he used previously acquired annotation, assembly and expression data. J.-F. G. supervised the project while M.L. advised.

Methods

We set up a pipeline using Biopython (Cock et al., 2009) to make our analysis (Fig. 3), the code is available at `https://github.com/LynchLabGroup/para`

Genomes and Annotation

We used annotation and sequence from our previous analyses for *P. caudatum* & the species from the *aurelia* complex. (see McGrath et al. 2014)

Gene families

Looking at the phylogenetic tree of the *aurelia* species, two WGDs occured at the root and affected three of our species. We have established some gene families from WGD2 using comparison, each family contains a set of orthologous genes between the four *aurelia* species studied, and eventually, the paralogous gene found in each species ; at maximum the families contain 13 different genes. Those families were established previously in our team. For details in the method see (McGrath et al. 2014 & McGrath *et al.* in press)

Upstream sequences extraction

For each gene, we extracted 250nt upstream of the start codon. If the previous gene was less than 250nt away, we reduced the extracted region so that it includes only inter-genic region. *Paramecium* genomes have very small inter-genic regions, on average 110bp for *caudatum* and around 300bp for the *aurelia* (McGrath et al., 2014). If the region was less than 15nt long, we removed the gene from the family. Sequences were extracted so that the first nucleotide of each one was the nearest from the start codon.

Coding Sequences extraction and alignment

Phylogenetic footprinting requires a phylogenetic tree to weigh the evolutionary signal of given motifs (Zhang and Gerstein, 2003). A motif conserved between two close species will be more likely to be conserved by chance than because of purifying selection than a motif conserved in two distant species. Because we are focusing on the conservation of upstream sequences we chose not to use them to avoid circularity. Instead,corresponding coding sequences (CDS) were extracted and used to model phylogenetic trees for each family. We preferred to have a gene tree over species tree, to avoid eventual inconsistencies because of gene conversion (ref. needed). See Challenges section for explanations on the use of gene tree over species tree.

CDSs in each family were aligned using TranslatorX (Abascal et al., 2010) a protein-guided alignment software. The Maximum Likelihood (ML) tree was then computed using PhyML (Guindon et al., 2010) the HKY85 model.

Phylogenetic footprinting

We used a phylogenetic footprinting software BigFoot (Satija et al., 2009) to detect highly conserved motifs in upstream sequences. We used 10000 burn-in cycles and 20000 cycles with a sampling rate of 1000 for the Hidden Monte-Carlo Markov Chain (HMCMC) process. The HMCMC is a stochastic that would run freely during the burn-in cycles, so that it can refine itself, then for a certain number of cycle we conserve its parameters at a given sampling rate. BigFoot aligns the given sequences with gaps and tries to identify conserved and non-conserved regions ; it models the evolution of those regions along the phylogenetic tree assuming conserved regions evolve more slowly than non-conserved ones. At the end of the analysis BigFoot outputs an alignment of sequences used to identify slow and fast evolving regions as well as, for each nucleotide in the alignment, the posterior probability of the alignment, higher values show higher confidence in the alignment, and the phylogenetic footprinting result, higher values indicating higher posterior probability of purifying selection.

Using a phylogenetic footprinting program means we have to use a phylogenetic tree and depending on the phylogenetic tree we are using, the evolutionary signal used in the footprinting is not identical.

The species tree (see Figure 2) gives us the relation between all considered, accounting for the various splits between species along with WGDs. The problem is that, not all gene families follow this tree. Because of the high similarity between duplicates, gene conversion occurs, the paralogs may recombine in a "copy-paste" way, we do not know what is the tree in each family and have to compute it for each of them.

Transcription factor binding sites are known to be generally conserved but degenerate on certain positions. For example, Whitfield et al. 2012 showed that this motif was conserved in human promoters with several highly variable positions (see Figure 4). Thus, to seek biologically relevant motifs, we had to take into account that in the middle of motifs, the phylogenetic score could drop on several positions, before rising again.

BigFoot does not output directly identified motifs, instead it produces two files with an alignment of the sequences and associated phylogenetic and alignment scores, as explained above.Using these scores we detected motifs of at least 6 nucleotides long, alignment score over 0.8 and phylogenetic score over 0.9.

Because of the known biological nature of Transcription Factor Binding Sites, we allowed for a "gap" in motifs of 2 nucleotide, so that the scores could drop under the thresholds in those gaps (see Figure 5). (Still need to add scores distribution)

To answer this problem we use a sliding window method of 8 nucleotides in our analysis: for each family, we looked at the scores of 8 nucleotides at the time and slide along the sequences. If the window contained at least 6 bases with scores above our thresholds, we would retain this motif. Then, from this particular region we would try to extend the sequence by adding adjacent nucleotides with good scores.

Comparison with MEME

To check our predictions and assess the conservation of found motifs, we compared motifs prediction with those of MEME (Bailey et al., 2006), a widely used *ab initio* motif finding tools (D'haeseleer, 2006). It searches for statistically significant motifs, with a gap-less, local, multi-alignment system.

MEME was shown to have a very high False Positive Rate of the discovery (Zia and Moses, 2012). That is why many studies combine multiple motif detection tools (Liseron-Monfils et al., 2013). In our case, MEME is of particular interest as it obtains motifs using a totally different method from BigFoot.

For each family we identified overlapping motifs between MEME and BigFoot. We computed an overlapping index as follows:

$$0 \leq \frac{\text{nucleotides in common}}{\text{size of the smallest motif}} \leq 1 \tag{1}$$

if this index was over 0.9 we would then considered the motif as relevant. Thus at the end we would only retain motifs that were pretty conserved among the different species.

Motif classification and data analysis

All the analyses were produced using R, scatter plots and graphs were produced using the R package `ggplot2`.

For each motif found by the pipeline, we extracted all genes having this motif in each species, we then compared the expression of these genes to those without the motifs using a Wilcoxon non-parametric test. We considered as significant all the tests with a p-value less than 0.001, if the mean of the expression of gene with the motif was higher than the expression of those without, we classified the motifs as increasing the expression, otherwise, we classified it as decreasing the expression.

To test whether motifs were equally conserved in all *Paramecium* species, we used generalized linear models (GLM) included in the package `stats`. The GLM was parametrized with the binomial family, logit link, and with the number of genes containing the motif and number of those that did not contain it as response variables. The number of genes containing the motif is thus considered as the number of "successes" out of the number of trials (total number of genes) in the binomial model. We then used a Wald type II analysis (ANOVA) using the package `car` to determine whether the proportion of genes containing the motifs differed among species. We then repeated these tests on each identified motifs. Thus, if the ANOVA was significant, it would show that the ratios between species were different in a statistically significant way.

Results

From our orthologs and paralogs gene familes we had 5751 families, only 5008 gene families had at least 4 genes. We ran our pipeline on these 5008 families, to detect conserved motifs among upstream sequences of *Paramecium* genomes. The pipeline was set to detect motifs, in upstream regions ranging between 15nt and 250nt (see Methods), of at least 6 nucleotides with an eventual gap in conservation (see Figure 4) for more biological relevance.

Using BigFoot we identified 811 different motifs in 608 families. For each of these, we extracted the name of all genes containing the motif, in each species. Only 10 motifs were present in the four species. We then computed the ratio of genes containing the motif over the total number of genes (results are summed up in Table 1).

One might note that all retained motifs seem to be modified TATA box, also, the small number of match may be explained by too stringent conditions. Still, if we look at the proportions of matching genes in each species, we can see different patterns (see Figure 6). First, looking by species (Figure 6A), *P. caudatum* seems to have a lower ratio than the other species. Indeed, we proceeded to a GLM as explained in Methods section, and found a very significant difference between the proportions (p-value $<$ 2e-16). Two hypotheses can explain this pattern: either *P. caudatum* lost the gene with these motifs, or the three *aurelia* species retained more genes with them.

Looking at all the different motifs (Figure 6C), we can see that between two and four motifs have a lower of genes containing them: `TAAATCT` has a mean ratio around 5% of the genes, `TTAATATT` around 12% and `TTAAATT` & `TTAATTA` around 30%, while other motifs have mean proportion around 40%. This may be due to different functions done by motifs, some motifs may be very specific to a given category of genes, while others are more widely used.

Indeed Figure 6B underlines the differences between motifs, to compare the proportions between species, we computed a GLM for each motifs, then used an ANOVA on this GLM. All the tests were highly significant (pvalue <2e-16) showing that in each motifs, the ratios between species are significantly different. Because we observed overall lower ratios in *P. caudatum*, we redid the analysis without it and all the comparisons were still significant. Thus, motif do not seem to be spread in the same way in each species.

We also tried to classify each motif as increasing or decreasing the overall expression of genes. We compared the expression distribution of genes with and without a motif using a Wilcoxon test, if it was significant, than we would classify the motif as "increasing the expression" or "decreasing the expression" according on the difference between means (results in Table 1). All the found motifs seem to be associated with an increase of expression for the genes possessing them in at least a species and 9 out of 10 at least in two species. `TTAAATT` and `TTAATTA` are associated with high expression levels in, respectively, 3 and 4 species, they may be good candidates to study the evolutionary fate of genes.

For the 117 unique found motifs, we also correlated the average expression levels with the distance from the start codon. For the 10 retained motifs we found a very significant (pvalue $<$ 5e-5 for all motifs) positive correlation with distances (data not shown). The farer the motif from the start codon the higher the expression of the genes containing it would be.

Future Directions

We identified motifs using an automated pipeline (Figure 3), comparing motifs identified using phylogenetic footprinting and statistical enrichment methods. We were able to dectect X motifs, associated with specific expression values and distance. Still, we plan to improve this pipeline in several ways.

After identifying motifs, we scan the entire *Paramecium* genomes to find associated genes, however we only search for perfect matches. Thus, we do not take into account degenerate motifs. Thus we can miss genes that have slightly different motifs. Recently Nelson and Wardle showed that looking exact matches during motif identification may miss most of the signal. Indeed we have increase evidence that real motifs are most of the time degenerate and that Transcription Factor binding mostly depends on the general DNA context.

One might want to group the motifs by similarity. We generally think about motifs as families, having specific sequences with some variation. For the moment, in our pipeline, we do not cluster motifs using "supposed" family, we consider each exact motifs as unique. However, several clustering methods exist to measure the diversity among motifs. The small sizes of motifs (about 6nt to 15nt) make these methods

complicated to develop. Several methods have been implemented, but still, are not adapted for motifs and do not output families of motifs.

Another way to think about the problem can be to cluster motifs according to features one can measure on them. For example, one can compute several statistical indexes on motifs to compare. Thus we do not need to implement a new algorithm to cluster motifs as other classical methods such as principal components analysis, hierarchal clustering or k-means clustering can be applied. Still, these methods rely on the indexes you choose to measure, the motifs you are studying have to have well-defined clusters using those variables. We tried to cluster the studied motifs with this hierarchal clustering method (data not shown), however, the chosen variables were not enough to classify motifs correctly.

Several other new methods have been proposed to cluster motifs as they are, considering the sequences and the variability along each nucleotide. Some of them rely on position-specific weight matrices (PSM) which are matrices of the abundances of each base at each position. In a future improvement of the pipeline we cool try to implement such methods to reduce the redundancy of our data.

It has been proven that most motif identification tools have a high false positive discovery rate, thus, we need ways to confirm the relevance of found motifs. One way used by Liseron-Monfils et al. 2013 is to increase the number of tools used in the pipeline, in their analysis they used MEME, BioProspector and Weeder and filtered out significant results and then only combined them. They have selected in each set, significant motifs using p-values and then only compare them. Liseron et al. showed that by combining these tools they were able to increase the sensitivity by 22% over the best standalone tool. Here we used MEME and BigFoot but we could include other tools in our workflow.

Acknowledgments

I would to thank my advisor Jean-François Goût for his patient and constant support, Michael Lynch for having me in his lab. More broadly, I would like to thank the whole Lynch lab team, for great scientific and non-scientific discussions.

Figures and Tables

Motif	Ratio bi.	Ratio ca.	Ratio sex.	Ratio tet.	Type bi.	Type ca.	Type sex.	Type tet.
AAAAAT	44.74 (17555)	25.60 (4739)	51.15 (17872)	38.61 (15833)	*NS*	*NS*	*NS*	**H**
TAAATCT	5.97 (2341)	2.54 (471)	6.25 (2184)	5.18 (2124)	**H**	**H**	*NS*	*NS*
TAAATT	52.36 (20549)	31.23 (5780)	56.81 (19849)	49.60 (20340)	**H**	*NS*	*NS*	**H**
TATTTA	52.00 (20406)	31.26 (5786)	54.76 (19131)	47.02 (19283)	**H**	*NS*	*NS*	**H**
TTAAAT	51.47 (20199)	30.36 (5619)	55.55 (19407)	48.11 (19728)	**H**	**H**	*NS*	*NS*
TTAAATT	27.34 (10727)	14.86 (2750)	30.57 (10680)	25.59 (10493)	**H**	*NS*	**H**	**H**
TTAATATT	13.89 (5452)	7.68 (1421)	14.18 (4956)	12.26 (5026)	**H**	*NS*	*NS*	*NS*
TTAATT	58.09 (22797)	40.53 (7501)	60.91 (21282)	55.34 (22693)	**H**	*NS*	*NS*	*NS*
TTAATTA	32.62 (12802)	23.29 (4311)	34.27 (11975)	30.30 (12427)	**H**	**H**	**H**	**H**
TTTATT	54.12 (21238)	32.14 (5949)	58.26 (20357)	49.61 (20344)	**H**	*NS*	*NS*	**H**

Table 1. Common motifs in all species. bi: *P. biaurelia*, ca: *P. caudatum*, sex: *P. sexaurelia*, tet: *P. tetraurelia*. Ratio: Proportion of genes containing the motif over all genes in percentage. The number of gene matching the given motif and species is in parentheses. Type: is the motif associated with a higher expression level? H: associated with higher expression, NS: non-significant.

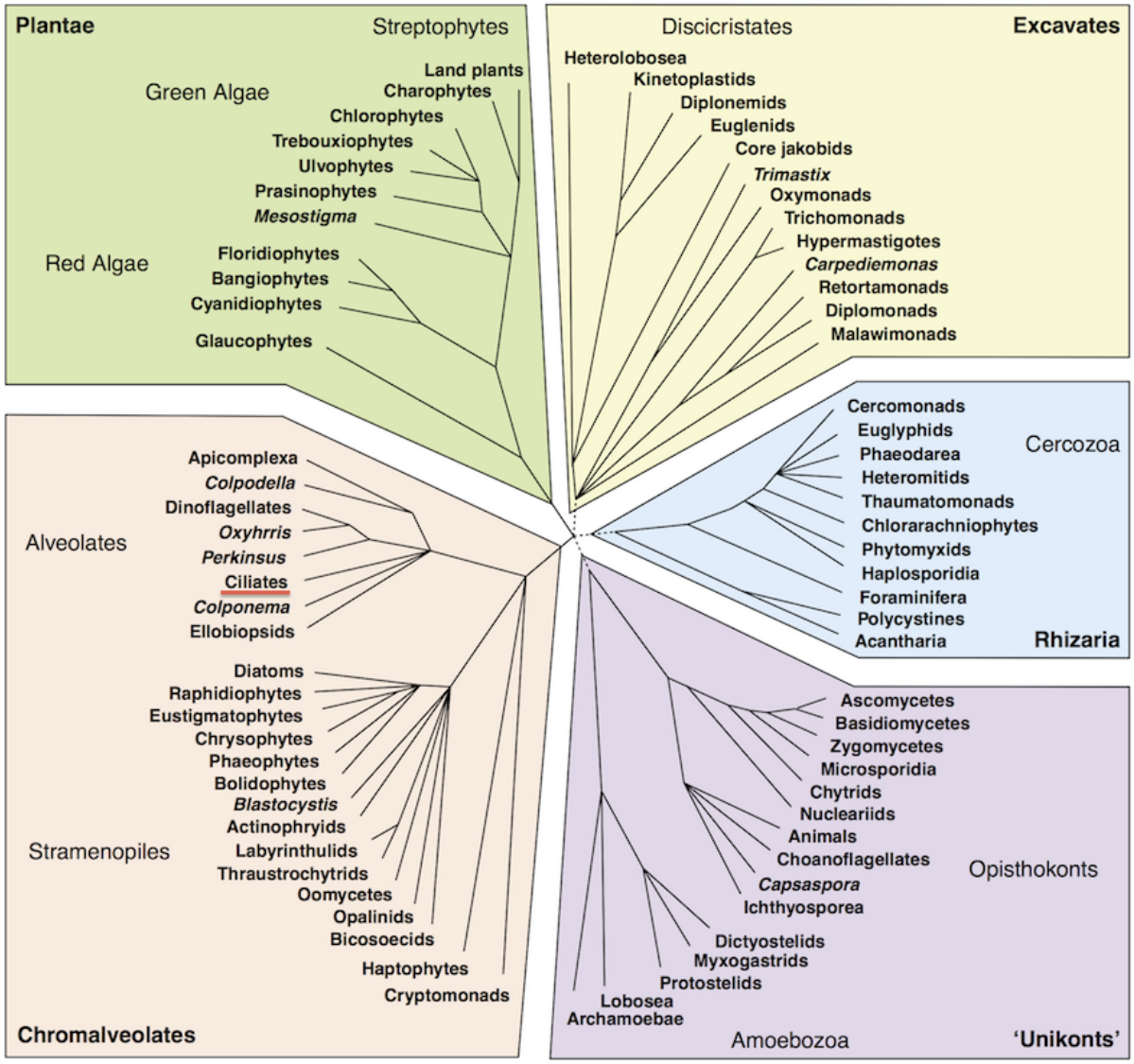

Figure 1. Phylogenetic tree of eukaryotes phylas. The *Ciliates* group, underlined in red in the figure, inside the Alveolates among the Chromoalveloates, it contains the *Paramecium* genus. While the Animals (Metazoa) belong to the Opisthokonts group. From (Keeling et al., 2005)

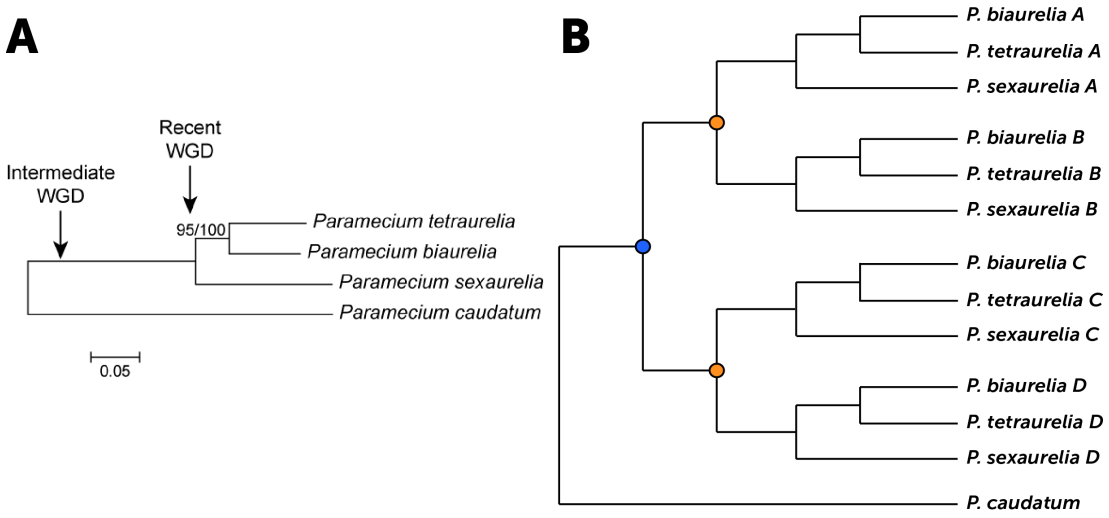

Figure 2. Phylogenetic trees of *Paramecium* and according gene tree. A: *Paramecium* species tree with WGD positions, *caudatum* is the more basal species and we used it here as an outgroup, while the *aurelia* cluster tightly together (Adapted from McGrath et al. 2014); B: An example of a family gene tree. With a single *caudatum* gene and four genes for each *aurelia* species, this is the maximum gene family size possible. In some families, certain genes of particular lineages may have been lost, the average size is 6.5, meaning that most of families actually lost some of the duplicate genes. *Legend:* blue circle, intermediary WGD; orange circle, more recent WGD; as dated by (Aury et al., 2006). (Adapted from McGrath et al. 2014).

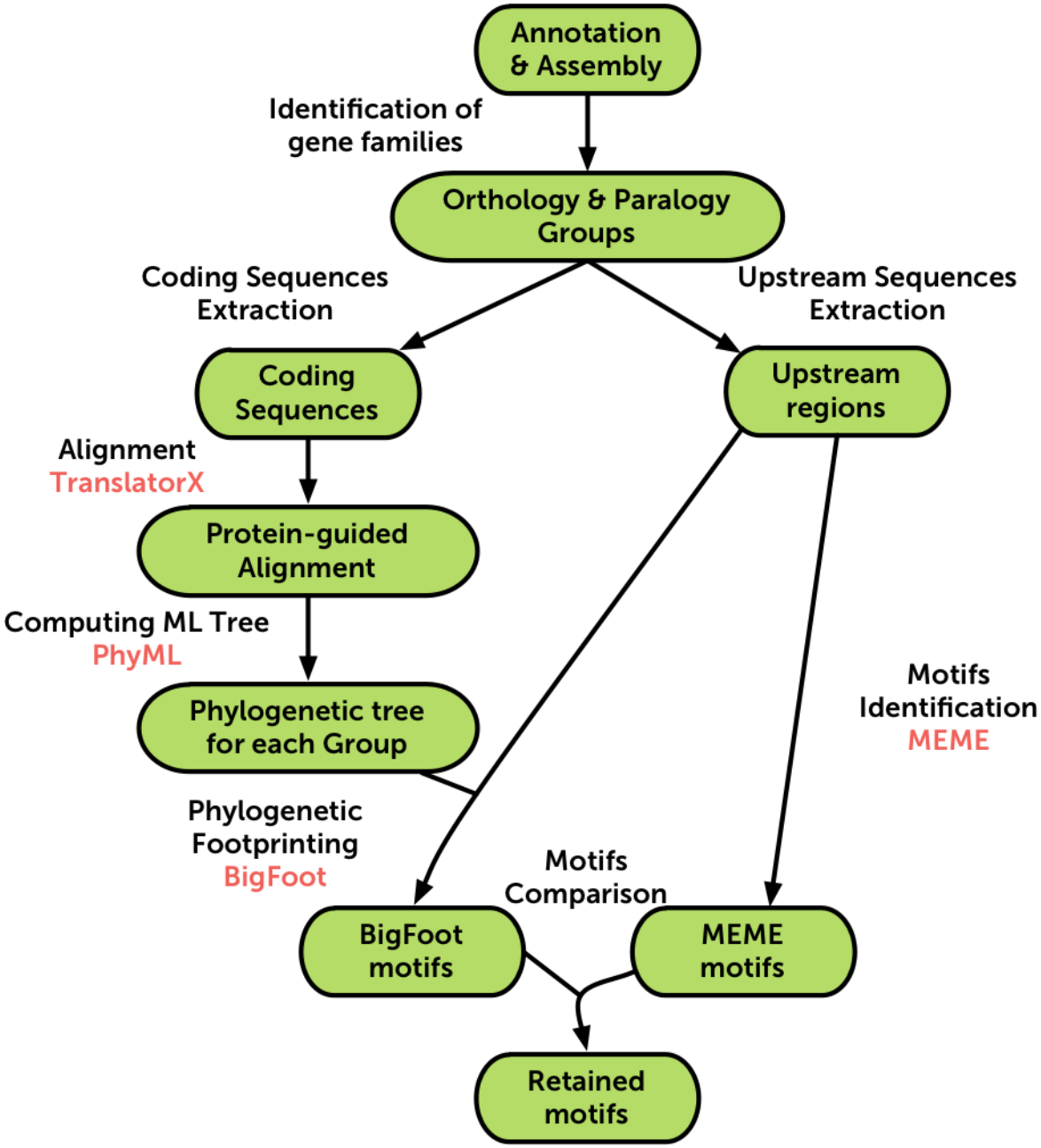

Figure 3. Flow chart of the whole pipeline. From the genome assembly and annotation of the four species, we used orthology and paralogy groups from (McGrath et al., 2014). For each of these families we extracted the coding sequences (CDS) as well as upstream regions from the start codon. We built a maximum likelihood phylogenetic tree using PhyML on pre-aligned CDS with TranslatorX (Abascal et al., 2010). On the one hand we computed the first fifth motifs of size of at least 4 nt using MEME on upstream regions (Bailey et al., 2006), while on the other hand using both the tree and the upstream sequences were used to detect motifs with a phylogenetic footprinting approach using BigFoot (Satija et al., 2009). We then retained only conserved motifs between MEME and BigFoot.

Figure 4. Degenerate Motif Example. This is a binding site from Whitfield et al. 2012, in human promoters. Some positions are clearly conserved, like position 5 with T, while others are highly variable (positions 1 and 4). Transcription Factor Binding Sites are very conserved on some positions while highly variable on others.

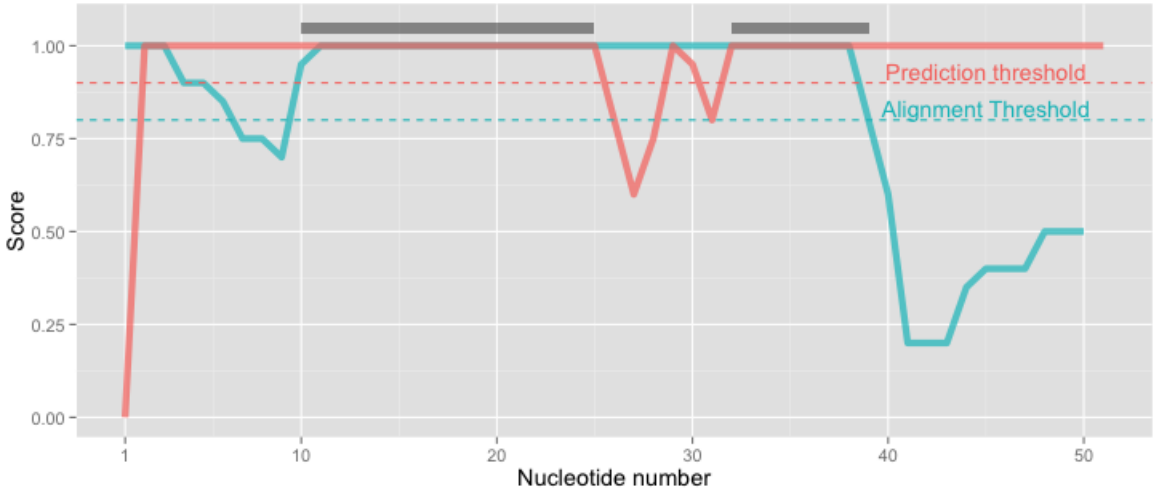

Figure 5. Motif detection from BigFoot example. BigFoot proceeds to phylogenetic prediction using an alignment technique, refer to (Satija et al., 2009) for more informations. For each position in the alignment, BigFoot assigns an alignment confidence score and phylogenetic footprinting,*i.e.* motif prediction, confidence score. In our analysis we used a detection technique to allow gaps in both the prediction and the alignment score. Curves: Red solid: Prediction score, Red dashed: Prediction threshold used, Blue solid: Alignment score, Blue dashed: Alignment threshold. Gray boxes: potential motifs detected in the sequence.

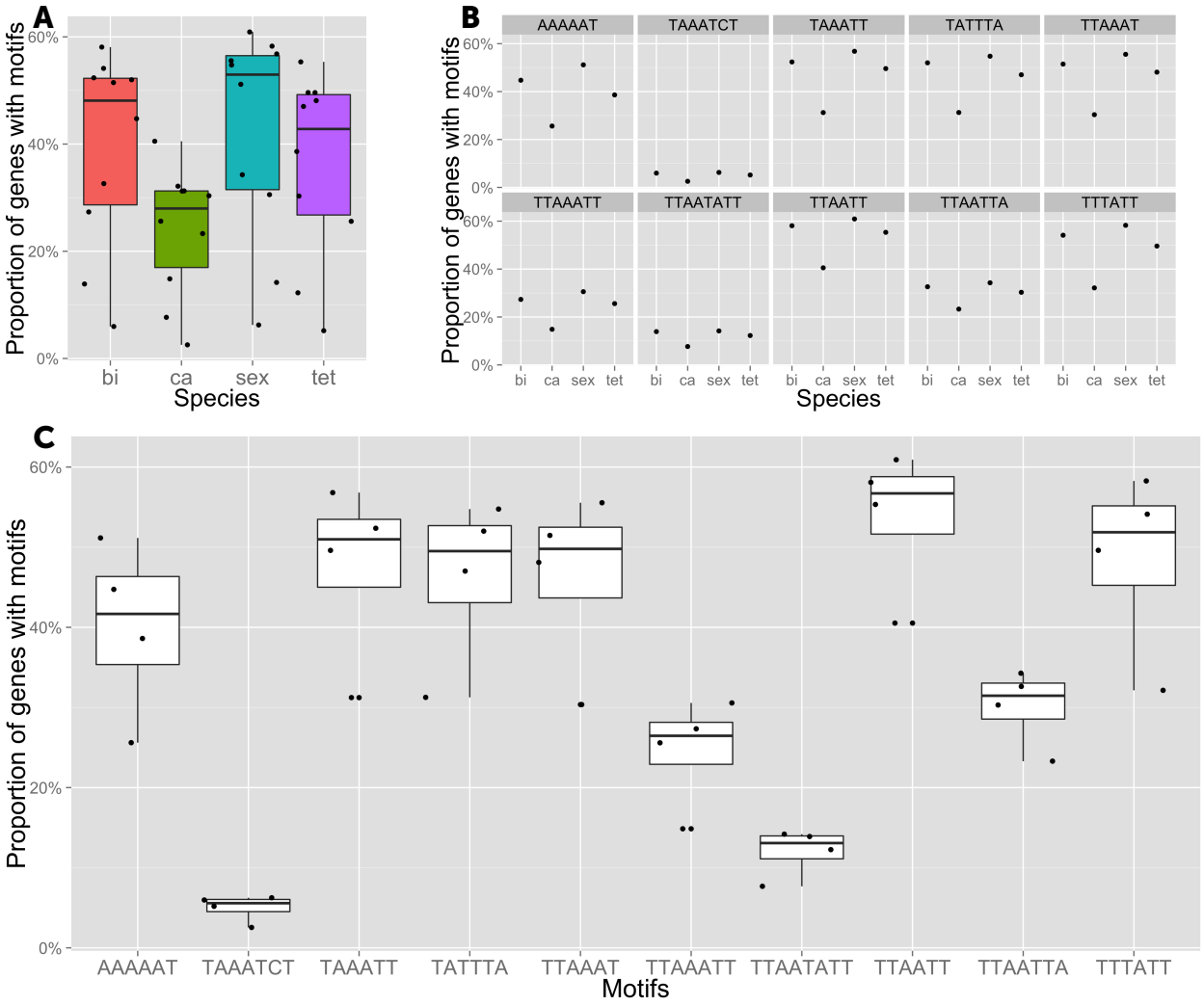

Figure 6. Ratio of genes with detected common motifs in all the species. Legend: A, Proportion of genes containing the motif by species ; B, separated by motifs and by species ; C, grouped by motif. bi: *P. biaurelia*; ca: *P. caudatum*; sex: *P. sexaurelia*; tet: *P.tetraurelia*. *P. caudatum* seems, for all motifs, to have less genes containing them then the other species. While two motifs seems importantly lower than others: TAAATCT and TTAATATT, but TTAATATT and TTAATTA seem also lower than other motifs. For more details on the ratios, see Table 1

References

F. Abascal, R. Zardoya, and M. J. Telford. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research*, 38(suppl 2):W7–W13, July 2010. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq291. URL http://nar.oxfordjournals.org/content/38/suppl_2/W7. PMID: 20435676.

M. A. Alexandrou, B. A. Swartz, N. J. Matzke, and T. H. Oakley. Genome duplication and multiple evolutionary origins of complex migratory behavior in salmonidae. *Molecular phylogenetics and evolution*, 69(3):514–523, Dec. 2013. ISSN 1095-9513. doi: 10.1016/j.ympev.2013.07.026. PMID: 23933489.

O. Arnaiz, J.-F. Gout, M. Betermier, K. Bouhouche, J. Cohen, L. Duret, A. Kapusta, E. Meyer, and L. Sperling. Gene expression in a paleopolyploid: a transcriptome resource for the ciliate paramecium tetraurelia. *BMC Genomics*, 11:547, Oct. 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-547. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3091696/. PMID: 20932287 PMCID: PMC3091696.

N. Arrigo and M. S. Barker. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology*, 15(2):140–146, 2012. ISSN 1369-5266. doi: 10.1016/j.pbi.2012.03.010. URL http://www.sciencedirect.com/science/article/pii/S1369526612000453.

J.-M. Aury, O. Jaillon, L. Duret, B. Noel, C. Jubin, B. M. Porcel, B. Ségurens, V. Daubin, V. Anthouard, N. Aiach, O. Arnaiz, A. Billaut, J. Beisson, I. Blanc, K. Bouhouche, F. Câmara, S. Duharcourt, R. Guigo, D. Gogendeau, M. Katinka, A.-M. Keller, R. Kissmehl, C. Klotz, F. Koll, A. Le Mouël, G. Lepère, S. Malinsky, M. Nowacki, J. K. Nowak, H. Plattner, J. Poulain, F. Ruiz, V. Serrano, M. Zagulski, P. Dessen, M. Bétermier, J. Weissenbach, C. Scarpelli, V. Schächter, L. Sperling, E. Meyer, J. Cohen, and P. Wincker. Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, 444(7116):171–178, Nov. 2006. ISSN 0028-0836. doi: 10.1038/nature05230. URL http://www.nature.com/nature/journal/v444/n7116/abs/nature05230.html.

T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(suppl 2):W369–W373, July 2006. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkl198. URL http://nar.oxfordjournals.org/content/34/suppl_2/W369. PMID: 16845028.

J. Beisson, M. Bétermier, M.-H. Bré, J. Cohen, S. Duharcourt, L. Duret, C. Kung, S. Malinsky, E. Meyer, J. R. Preer, and L. Sperling. Paramecium tetraurelia: The renaissance of an early unicellular model. *Cold Spring Harbor Protocols*, 2010(1):pdb.emo140, Jan. 2010. ISSN 1940-3402, 1559-6095. doi: 10.1101/pdb.emo140. URL http://cshprotocols.cshlp.org/content/2010/1/pdb.emo140. PMID: 20150105.

F. Catania, F. Wurmser, A. A. Potekhin, E. Przyboś, and M. Lynch. Genetic diversity in the paramecium aurelia species complex. *Molecular Biology and Evolution*, 26(2):421–431, Feb. 2009. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msn266. URL http://mbe.oxfordjournals.org/content/26/2/421. PMID: 19023087.

F. J. Chain, J. Dushoff, and B. J. Evans. The odds of duplicate gene persistence after polyploidization. *BMC genomics*, 12(1):599, 2011. URL http://www.biomedcentral.com/1471-2164/12/599/.

W.-H. Chen, W. Wei, and M. J. Lercher. Minimal regulatory spaces in yeast genomes. *BMC Genomics*, 12(1):320, June 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-320. URL http://www.biomedcentral.com/1471-2164/12/320/abstract. PMID: 21679449.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. d. Hoon. Biopython: freely available python tools for computational molecular biology and. *Bioinformatics*, 25(11):1422–1423, June 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp163. URL http://bioinformatics.oxfordjournals.org/content/25/11/1422. PMID: 19304878.

P. Dehal and J. L. Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10):e314, Sept. 2005. doi: 10.1371/journal.pbio.0030314. URL http://dx.doi.org/10.1371/journal.pbio.0030314.

P. D'haeseleer. How does DNA sequence motif discovery work? *Nature Biotechnology*, 24(8):959–961, 2006. ISSN 1087-0156. doi: 10.1038/nbt0806-959. URL http://www.nature.com.ezproxy.lib.indiana.edu/nbt/journal/v24/n8/full/nbt0806-959.html.

P. P. Edger and J. C. Pires. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, 17(5):699–717, July 2009. ISSN 0967-3849, 1573-6849. doi: 10.1007/s10577-009-9055-9. URL http://link.springer.com/article/10.1007/s10577-009-9055-9.

J.-F. Gout, D. Kahn, L. Duret, and Paramecium Post-Genomics Consortium. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet*, 6(5):e1000944, 2010. doi: 10.1371/journal.pgen.1000944. URL http://dx.doi.org/10.1371/journal.pgen.1000944.

S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–321, May 2010. ISSN 1076-836X. doi: 10.1093/sysbio/syq010. PMID: 20525638.

S. Hahn and E. T. Young. Transcriptional regulation in saccharomyces cerevisiae: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, 189(3):705–736, Nov. 2011. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.111.127019. URL http://www.genetics.org/content/189/3/705. PMID: 22084422.

H. Innan and F. Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, 2010. ISSN 14710056. doi: 10.1038/nrg2689. URL http://ezproxy.lib.indiana.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=47586173&site=ehost-live&scope=site.

P. J. Keeling, G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray. The tree of eukaryotes. *Trends in Ecology & Evolution*, 20(12):670–676, 2005. ISSN 0169-5347. doi: 10.1016/j.tree.2005.09.005. URL http://www.sciencedirect.com/science/article/pii/S0169534705003046.

M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, 2004. ISSN 0028-0836. doi: 10.1038/nature02424. URL http://www.nature.com.ezproxy.lib.indiana.edu/nature/journal/v428/n6983/full/nature02424.html.

C. Liseron-Monfils, T. Lewis, D. Ashlock, P. D. McNicholas, F. Fauteux, M. Strömvik, and M. N. Raizada. Promzea: a pipeline f or discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the maize development atlas. *BMC plant biology*, 13(1):42, 2013. URL http://www.biomedcentral.com/content/pdf/1471-2229-13-42.pdf.

C. L. McGrath, J.-F. Gout, T. G. Doak, A. Yanagi, and M. Lynch. Insights into three whole-genome duplications gleaned from the paramecium caudatum genome sequence. *Genetics*, page genetics.114.163287, May 2014. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.114.163287. URL http://www.genetics.org/content/early/2014/05/19/genetics.114.163287. PMID: 24840360.

A. C. Nelson and F. C. Wardle. Conserved non-coding elements and cis regulation: actions speak louder than words. *Development*, 140(7):1385–1395, Apr. 2013. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.084459. URL http://dev.biologists.org/content/140/7/1385. PMID: 23482485.

S. Ohno. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, 99(1):120–130, Jan. 1970. ISSN 0002-8487, 1548-8659. doi: 10.1577/1548-8659(1970)99⟨120:TEDIGS⟩2.0.CO;2. URL http://www.tandfonline.com/doi/abs/10.1577/1548-8659%281970%2999%3C120%3ATEDIGS%3E2.0.CO%3B2.

R. Satija, Á. Novák, I. Miklós, R. Lyngsø, and J. Hein. BigFoot: bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evolutionary Biology*, 9(1):217, Aug. 2009. ISSN 1471-2148. doi: 10.1186/1471-2148-9-217. URL http://www.biomedcentral.com/1471-2148/9/217/abstract. PMID: 19715598.

T. W. Whitfield, J. Wang, P. J. Collins, E. C. Partridge, S. F. Aldred, N. D. Trinklein, R. M. Myers, and Z. Weng. Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13(9):R50, Sept. 2012. ISSN 1465-6906. doi: 10.1186/gb-2012-13-9-r50. URL http://genomebiology.com/2012/13/9/R50/abstract. PMID: 22951020.

P. J. Wittkopp and G. Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69, Jan. 2012. ISSN 1471-0056. doi: 10.1038/nrg3095. URL http://www.nature.com/nrg/journal/v13/n1/full/nrg3095.html.

Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *Journal of Biology*, 2(2):11, June 2003. ISSN 1475-4924. doi: 10.1186/1475-4924-2-11. URL http://jbiol.com/content/2/2/11/abstract. PMID: 12814519.

A. Zia and A. M. Moses. Towards a theoretical understanding of false positives in DNA motif finding. *BMC Bioinformatics*, 13(1):151, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-151. URL http://www.biomedcentral.com.ezproxy.lib.indiana.edu/1471-2105/13/151.