Title

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

Author Summary

Introduction

Since Ohno first hypothesized the influence of Whole-Genome Duplications (WGD) [Ohno, 1970], scientists kept showing that a broad number of organisms experienced several rounds of WGD: *Saccharomyces cerevisiae* [**?**], 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. A recent study on the horshoe crab genome underlined that WGDs may be a more common phenomenon than stated. They showed that at least a WGD occurred in this conserved lineage, thus WGDs may not be evolutionary drivers.

Still, it seems that WGDs are widespread along the tree of life and since Ohno numerous models have been developed to explain the retention rate we observe between duplicate genes (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 Fig. 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 Fig. 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]).

Recently, several studies unraveled the link between gene expression in gene retention after WGD in *Paramecium* (Gout et al. [2010], Arnaiz et al. [2010]). 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 250nt long inter-genic regions close to those of *Saccharomyces cerevisiae* [Chen et al., 2011]. These 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 each gene its *cis*-regulatory sequence is directly upstream of it 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 the 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. 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 than weighted by the phylogenetical relationships of the given sequences, *id est*, 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 specific 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.

Methods

We set up a pipeline to make our analyse (Fig. 3), the code is available at.

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. 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 have less importance 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. 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 Fig. 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 successive round of WGDs there are several fates possible for pair of duplicate after the first WGD. Some of these genes may cluster together in the same leaves, if the pairs diverge between species ; another possible outcome is the subfunctionnalization of each gene before the second WGD, meaning before the speciation of the *aurelia* complex, thus genes from different species would cluster together in a phylogenetic tree ; or even a combination of the previous outcome and gene conversion, leading paralogs to recombine in a copy-paste way, changing dramatically the gene tree.

Transcription factor binding sites are known to be generally conserved but degenerate on certain positions. For example, ... showed that this motif was conserved ....NN... with two highly variable positions (denoted by "N", meaning "A", "T", "C" or "G" using IUPAC notation). 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 Fig. 4).

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. 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 (Ref. needed). That is why many studies combine multiple motif detection tools. In our case, MEME is of particular interest as it obtains motifs using a totally different method from BigFoot.

To assess the relevance of motifs found from BigFoot's outputs, we compared them to a well-known motif finding program: MEME. 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.

Motif classification and data analysis

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

Results

From our orthologs and paralogs gene familes we had 5751 families, only 5008 gene families had at least 4 genes. Among these 5008 families BigFoot found 811 motifs in 608 families. From these 811 motifs 117 matches with motifs found by MEME.

Perspectives

After identifying motifs should relate presence/absence of motifs to duplicated genes fate.

Motifs detection should take phylogeny into account for comparative analysis. Not the same value.

Need to improve pipeline for degenerate motifs, for the moment, extract only exact motif in the genome. (Ref. needed)

Need to measure diversity among detected motifs -¿ clustering tools and suppress rendundancy

Implement other motif finding tools to validate results.

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.

References

References

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

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

Olivier Arnaiz, Jean-Francois Gout, Mireille Betermier, Khaled Bouhouche, Jean Cohen, Laurent Duret, Aurelie Kapusta, Eric Meyer, and Linda Sperling. Gene expression in a paleopolyploid: a transcriptome resource for the ciliate paramecium tetraurelia. *BMC Genomics*, 11:547, October 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.

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

Jean-Marc Aury, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, Vincent Daubin, Véronique Anthouard, Nathalie Aiach, Olivier Arnaiz, Alain Billaut, Janine Beisson, Isabelle Blanc, Khaled Bouhouche, Francisco Câmara, Sandra Duharcourt, Roderic Guigo, Delphine Gogendeau, Michael Katinka, Anne-Marie Keller, Roland Kissmehl, Catherine Klotz, France Koll, Anne Le Mouël, Gersende Lepère, Sophie Malinsky, Mariusz Nowacki, Jacek K. Nowak, Helmut Plattner, Julie Poulain, Françoise Ruiz, Vincent Serrano, Marek Zagulski, Philippe Dessen, Mireille Bétermier, Jean Weissenbach, Claude Scarpelli, Vincent Schächter, Linda Sperling, Eric Meyer, Jean Cohen, and Patrick Wincker. Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, 444(7116):171–178, November 2006. ISSN 0028-0836. doi: 10.1038/nature05230. URL http://www.nature.com/nature/journal/v444/n7116/abs/nature05230.html.

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

Janine Beisson, Mireille Bétermier, Marie-Hélène Bré, Jean Cohen, Sandra Duharcourt, Laurent Duret, Ching Kung, Sophie Malinsky, Eric Meyer, John R. Preer, and Linda Sperling. Paramecium tetraurelia: The renaissance of an early unicellular model. *Cold Spring Harbor Protocols*, 2010(1):pdb.emo140, January 2010. ISSN 1940-3402, 1559-6095. doi: 10.1101/pdb.emo140. URL http://cshprotocols.cshlp.org/content/2010/1/pdb.emo140. PMID: 20150105.

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

Wei-Hua Chen, Wu Wei, and Martin 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.

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

Jean-François Gout, Daniel Kahn, Laurent 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, May 2010. doi: 10.1371/journal.pgen.1000944. URL http://dx.doi.org/10.1371/journal.pgen.1000944.

Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier 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.

Hideki Innan and Fyodor Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, February 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.

Casey L. McGrath, Jean-Francois Gout, Thomas G. Doak, Akira Yanagi, and Michael 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.

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

Susumu 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, January 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.

Rahul Satija, Ádám Novák, István Miklós, Rune Lyngsø, and Jotun Hein. BigFoot: bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evolutionary Biology*, 9(1):217, August 2009. ISSN 1471-2148. doi: 10.1186/1471-2148-9-217. URL http://www.biomedcentral.com/1471-2148/9/217/abstract. PMID: 19715598.

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

Zhaolei Zhang and Mark 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.

Figure Legends

Tables

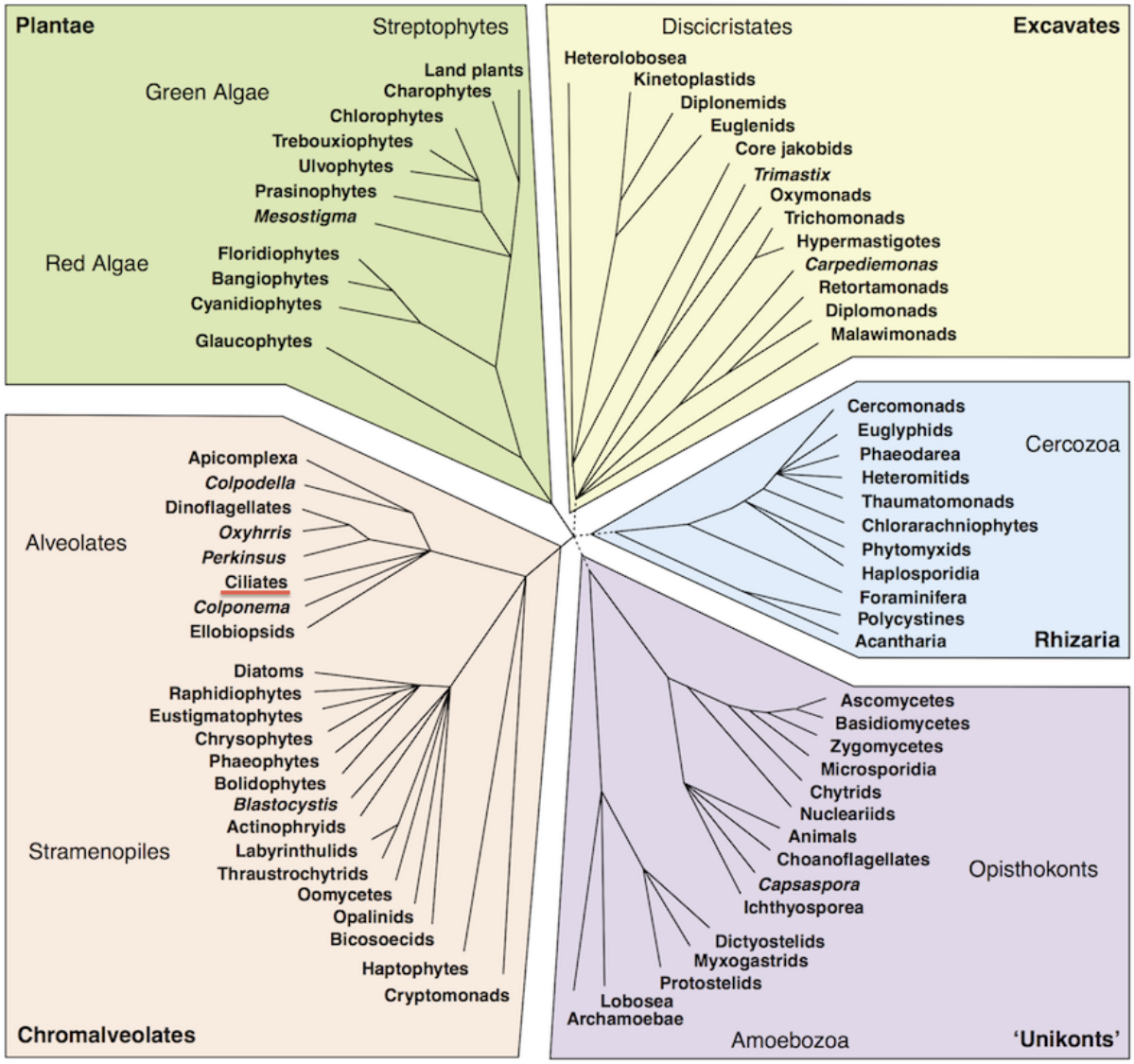

Figure 1. Phylogenetic tree of eukaryotes phylas. The *Ciliates* group, underlined in red in the figure, inside the Alveolates among the Chromoalveloates, it contains the *Paramecium* genus. While the Animals (Metazoa) belong to the Opisthokonts group. From (Ref. needed)

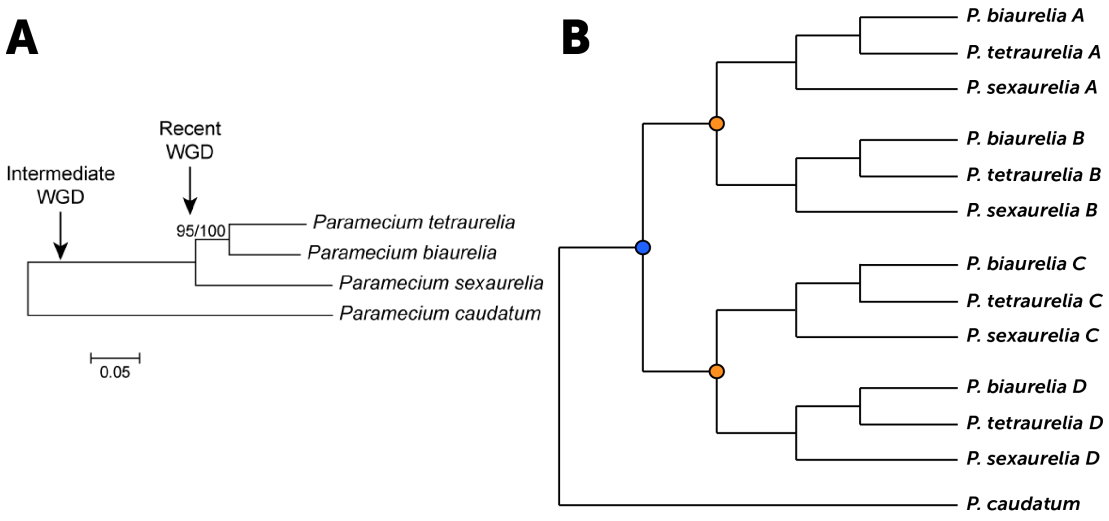

Figure 2. Phylogenetic trees of *Paramecium* and according gene tree. A: *Paramecium* species tree with WGD positions (Adapted from Ref. needed); B: An example of a family gene tree. A single *caudatum* gene and four genes for each *aurelia* species. *Legend:* blue circle, intermediary WGD; orange circle, more recent WGD; as dated by (Ref. needed). (Adapted from Ref. needed)

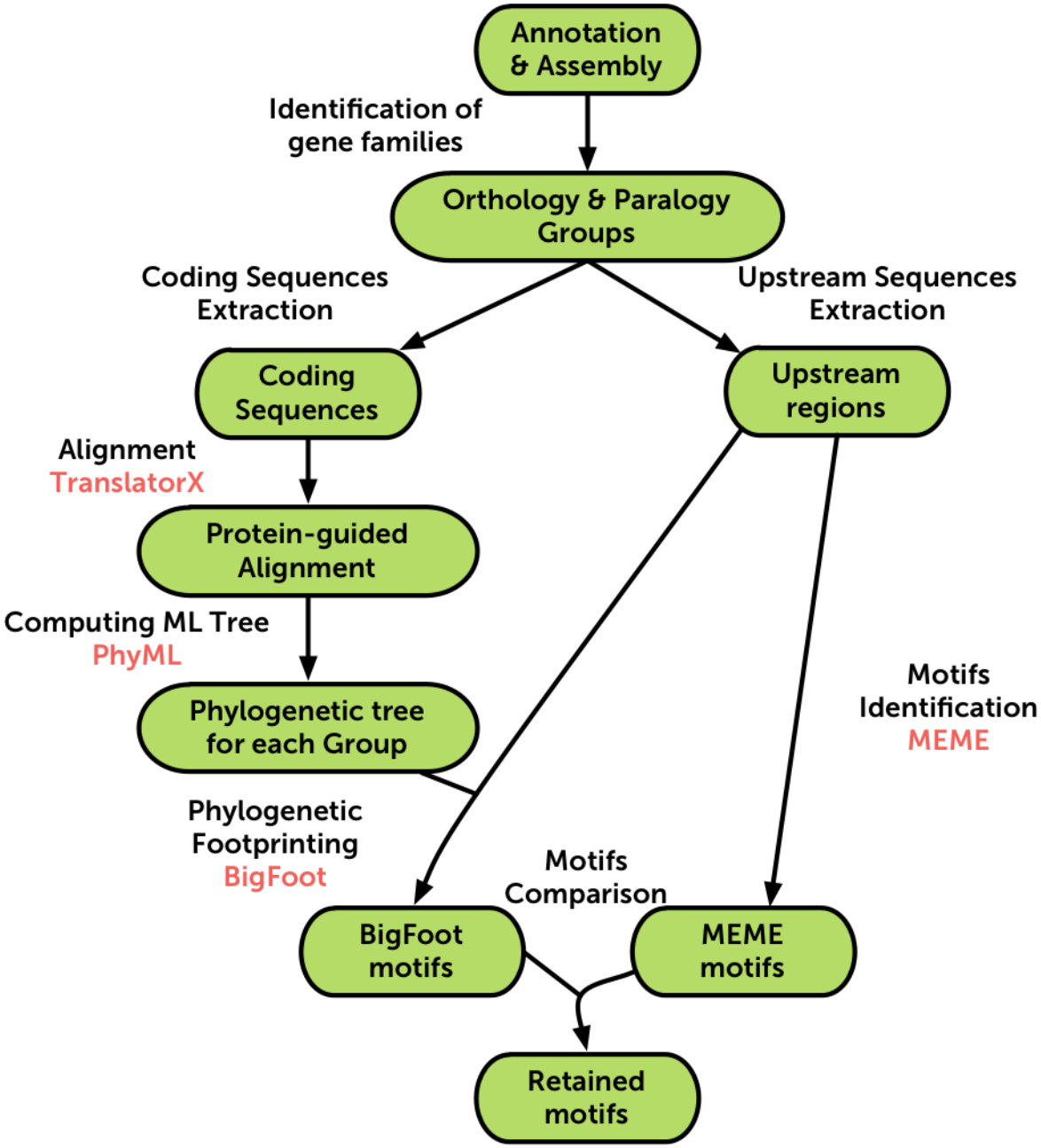

Figure 3. Flow chart of the whole pipeline. From the genome assembly and annotation of the four species, we used orthology and paralogy groups from (Ref. needed). For each of these families we extracted the coding sequences (CDS) as well as upstream regions from the start codon. We built a maximum likelihood phylogenetic tree using PhyML on pre-aligned CDS with TranslatorX. On the one hand we computed the first fifth motifs of size of at least 4 nt using MEME on upstream regions, 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. We then retained only conserved motifs between MEME and BigFoot.

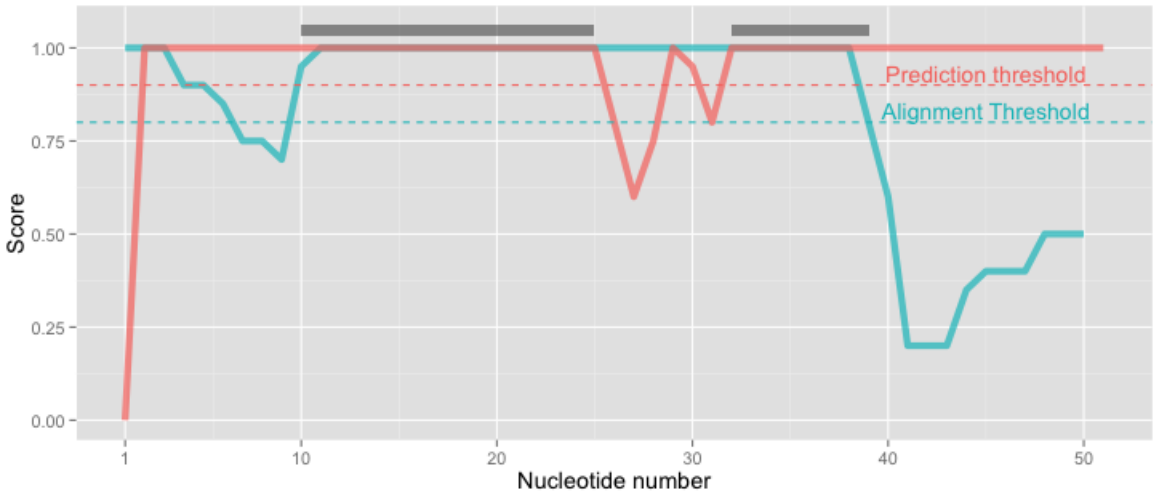

Figure 4. Motif detection from BigFoot example. BigFoot proceeds to phylogenetic prediction using an alignment technique, refer to (Ref. needed) for more informations. For each position in the alignment, BigFoot assigns an alignment confidence score and phylogenetic footprinting, *i.e.* motif prediction, 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.