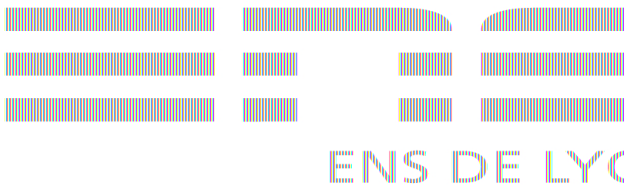
Finding Motifs in *Paramecium*: A Long and Hard Quest

Lab Meeting - June 19th 2014

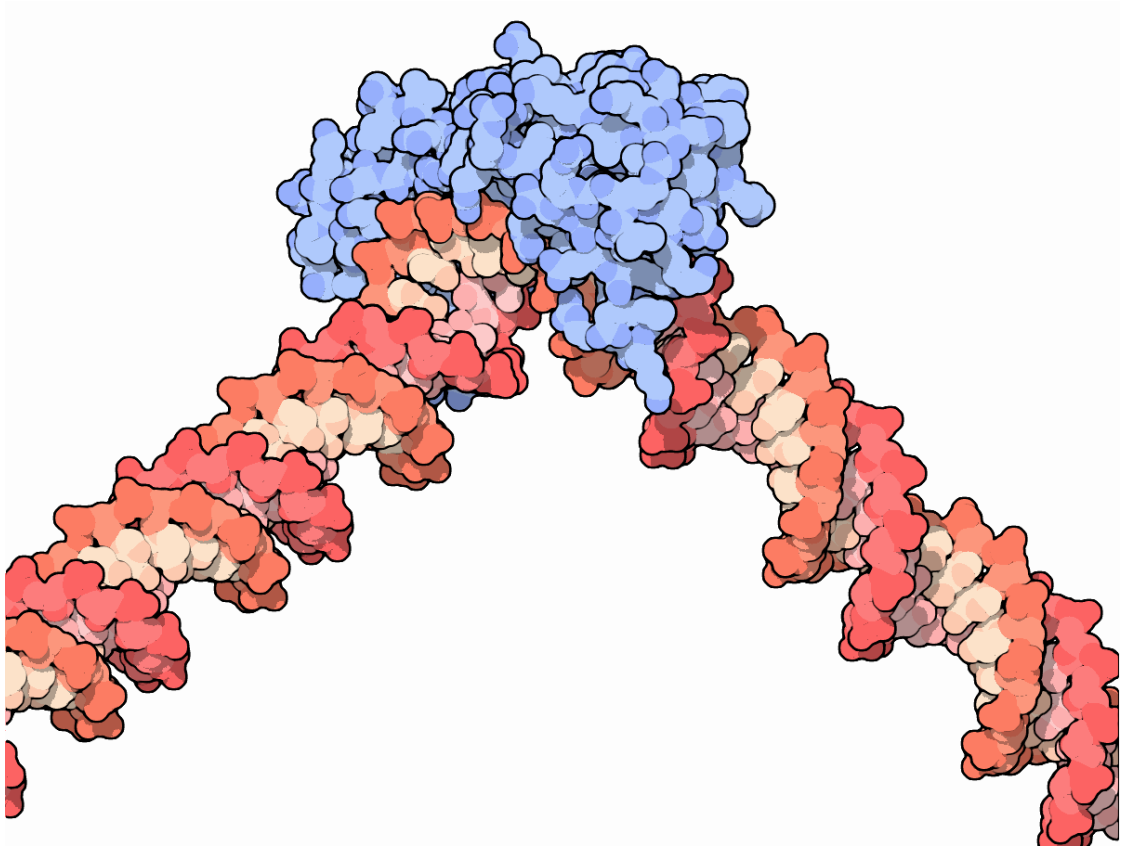
Matthias Grenié, Jean-François Goût, Michael Lynch



INDIANA UNIVERSITY
BLOOMINGTON

What have I been doing?

Gene Expression Regulation



Core Biological Process

Mediated by Transcription Factors

Transcription Factor Binding Sites

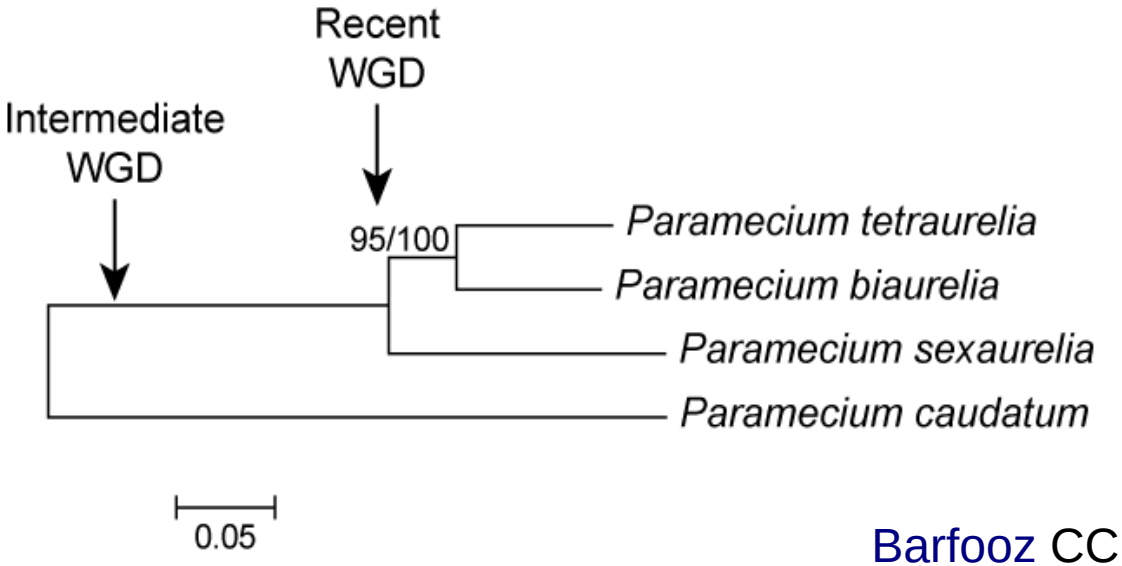
Diversity of regulation largely unknown

Major determinant of gene retention/loss after Whole-Genome Duplication

Model: *Paramecium*

Compact Genome
(intergenic regions ~250bp)

Closely related
species complex

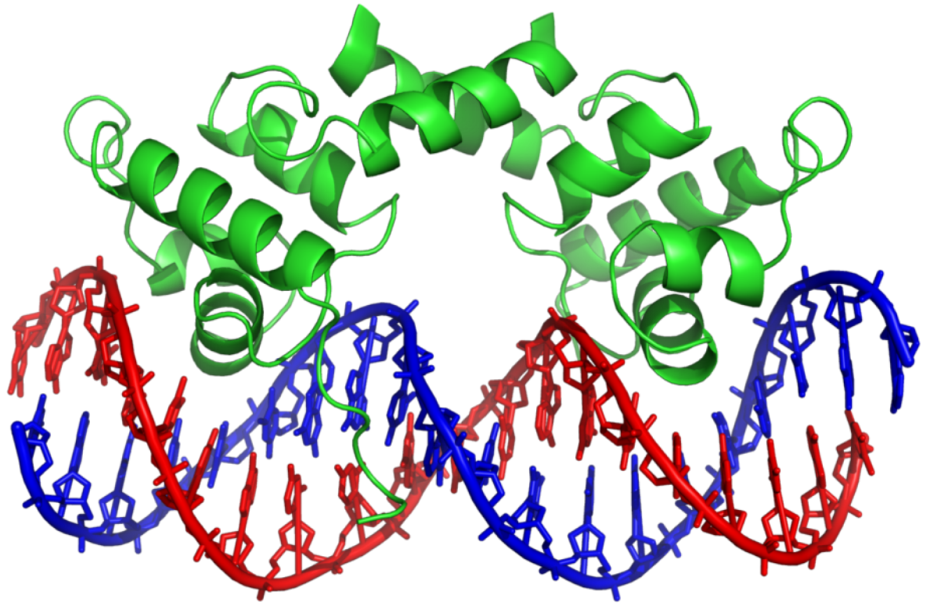


Experienced two WGDs

Transcription Factor Binding Sites

Small sequences (6-15nt long)

Degenerate



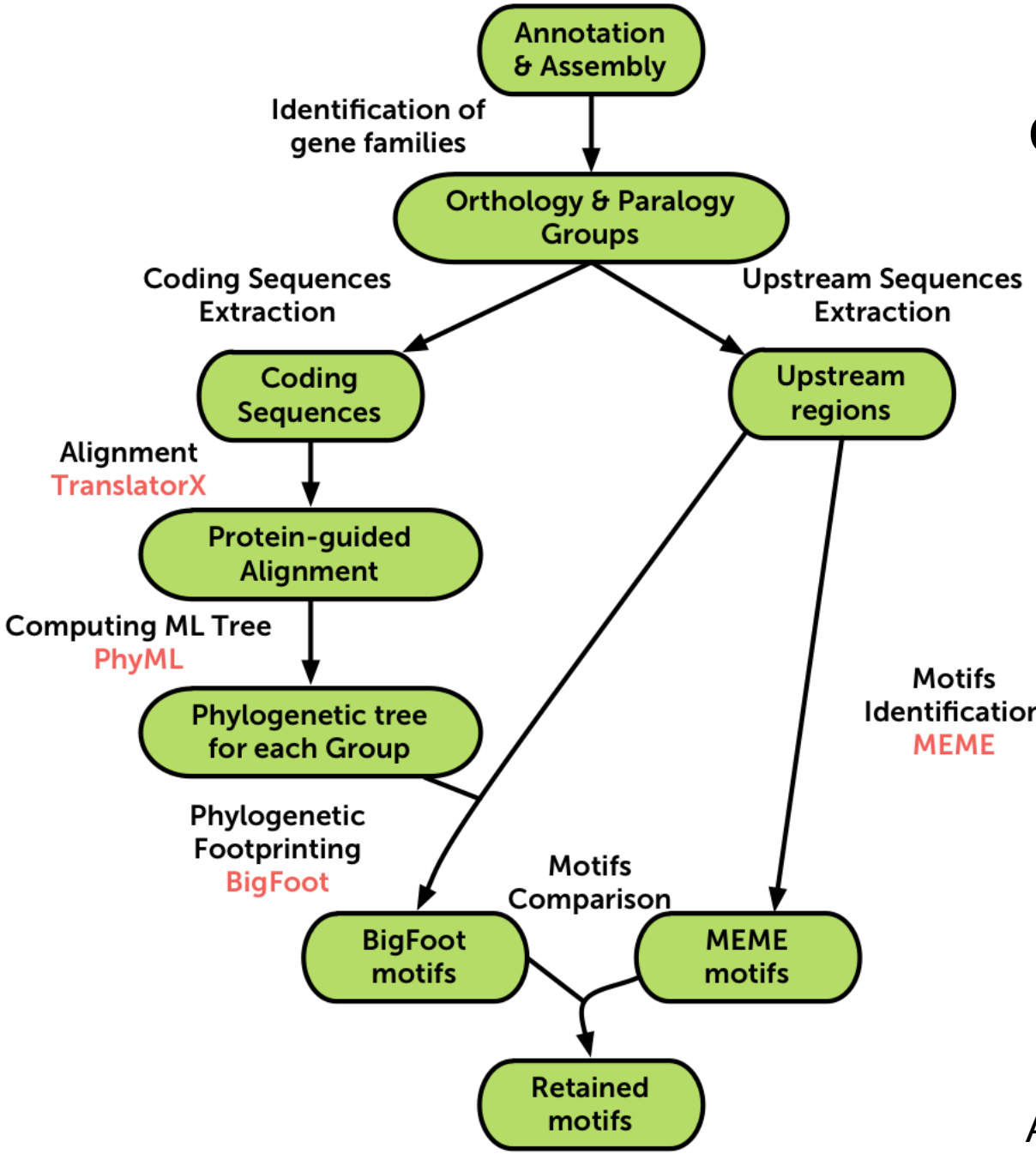
Conserved

Sometimes palindromic

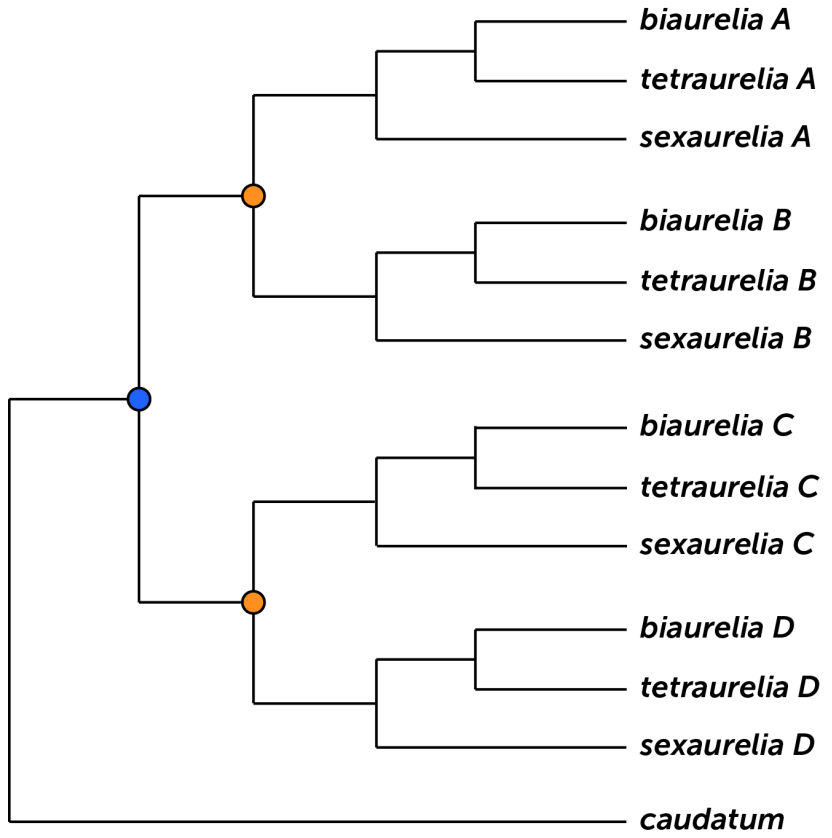
***e.g.* CACGTG**

Can we build an automated pipeline to identify Transcription Factor Binding Sites in the *Paramecium aurelia* complex?

Pipeline Overview



Orthology & Paralogy family



Adapted from McGrath et al. 2014 (right)

Classical Motif Finding

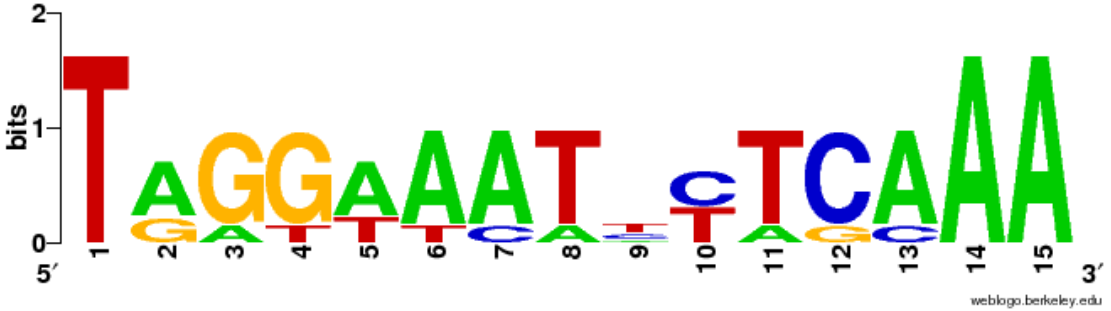
***De novo* motif finding, (around since the 1990s)**

Several tools available

Various methods (statistical, probabilistic, etc.)

Motif and Background sequences

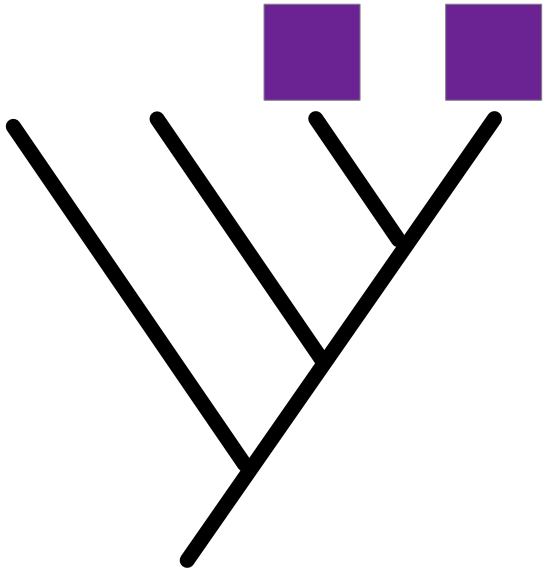
TG**G**GTTATCCTCC**AA**
TA**G**GAAATTTTCA**AA**
TA**G**GAAATTTTCA**AA**
TA**G**GAAATCCAGA**AA**
TG**A**TTAAAACTCA**AA**



Crooks et al. 2004, Bailey et al. 2006, Sahu et al. 2012

Phylogenetic Footprinting

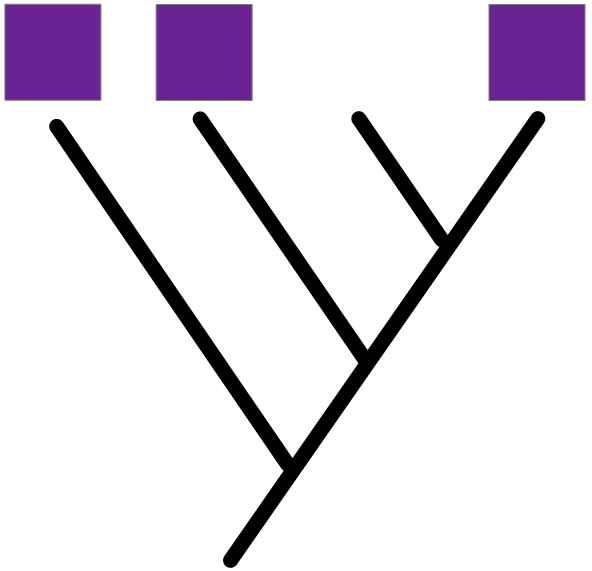
Compare ortholog sequences and use their phylogenetic relationships to identify motifs



Motif shared between close species

Motif rating

<



Motif shared between distant species

Motif rating

BigFoot, a phylogenetic footprinting software

Statistical alignment program

Model slowly and quickly evolving sequences

Takes the phylogeny into account to determine the boundary between regions

Does not output motifs directly → scores

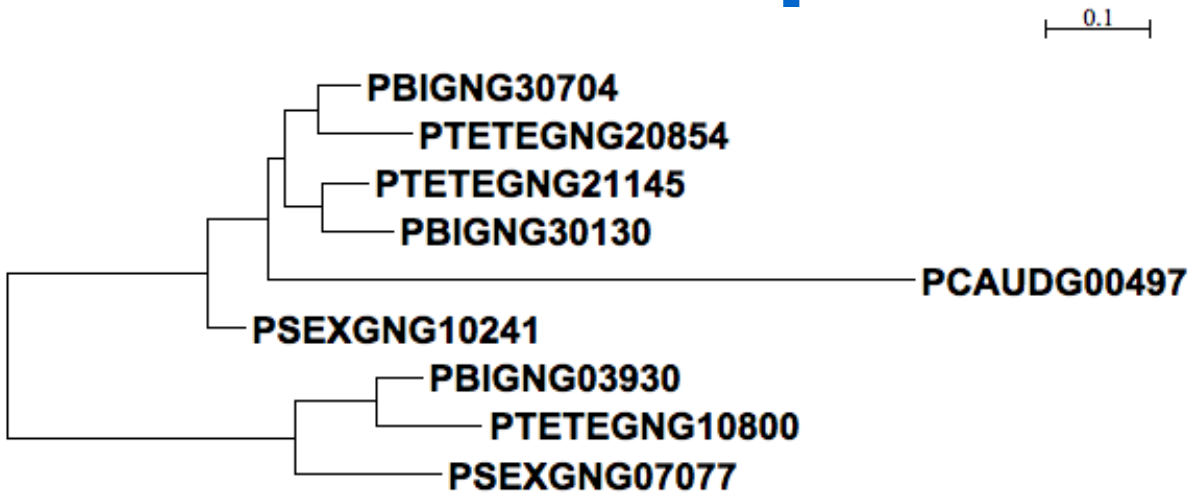
Alignment Score

Higher = better alignment

Phylogenetic Score

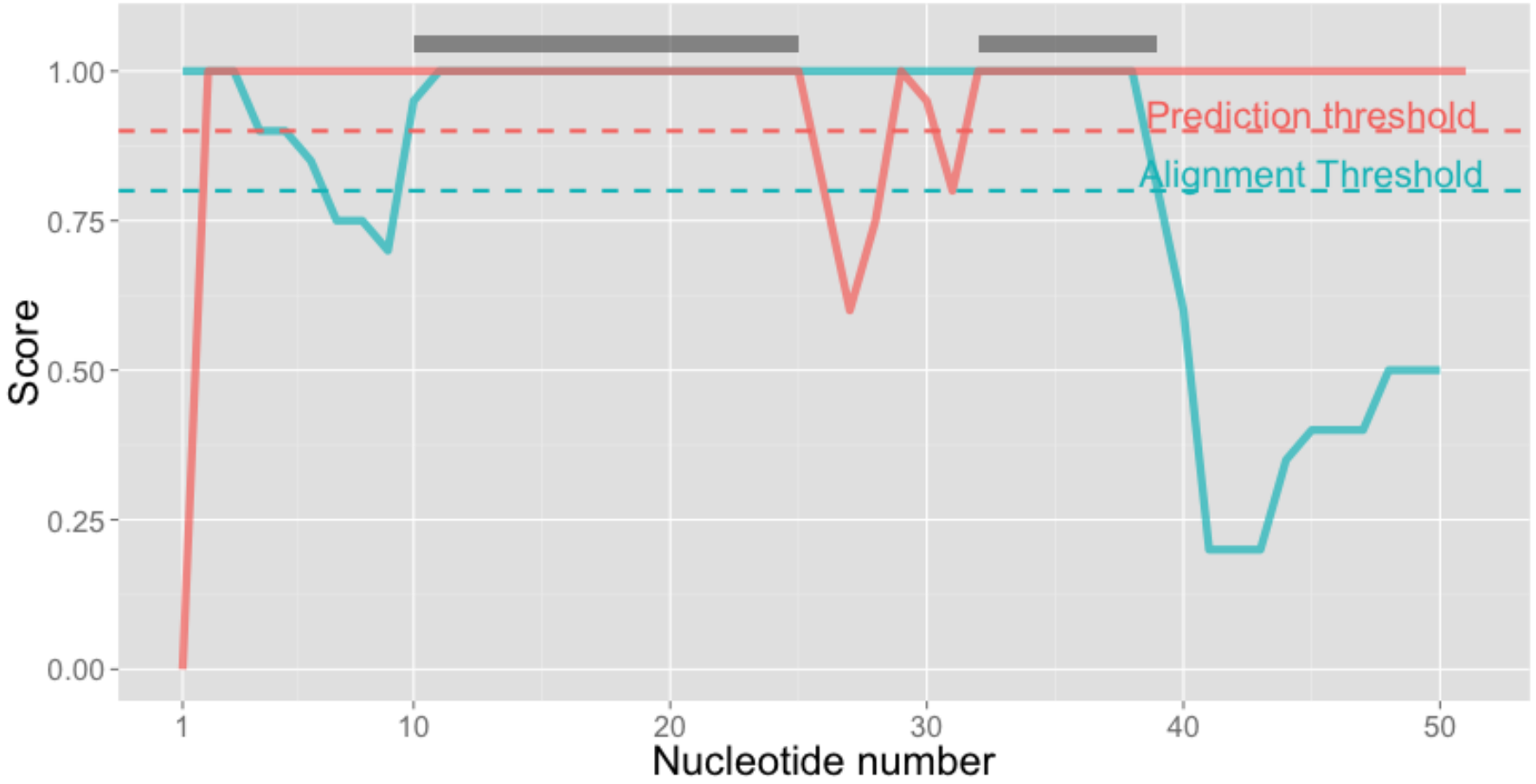
Higher = better conservation

Example of a family



Sliding window of 8nt

6 out of 8 correct positions



First Results

Extracted sequences from 15nt to 250nt

5781 families → 5008 included (min 4 genes, no overlap)

1614 motifs candidates found by BigFoot

6904 unique motifs found with MEME

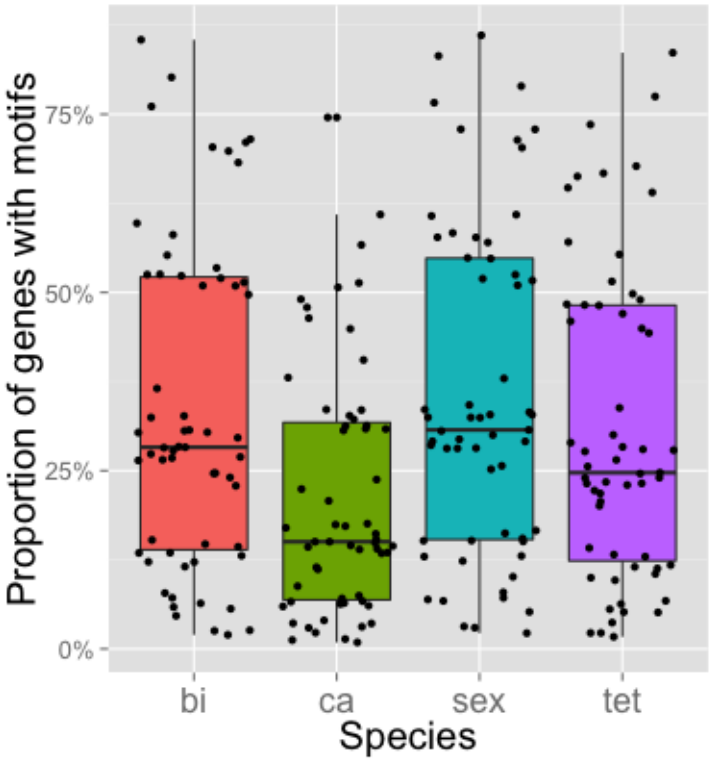
787 unique motifs after comparison with MEME

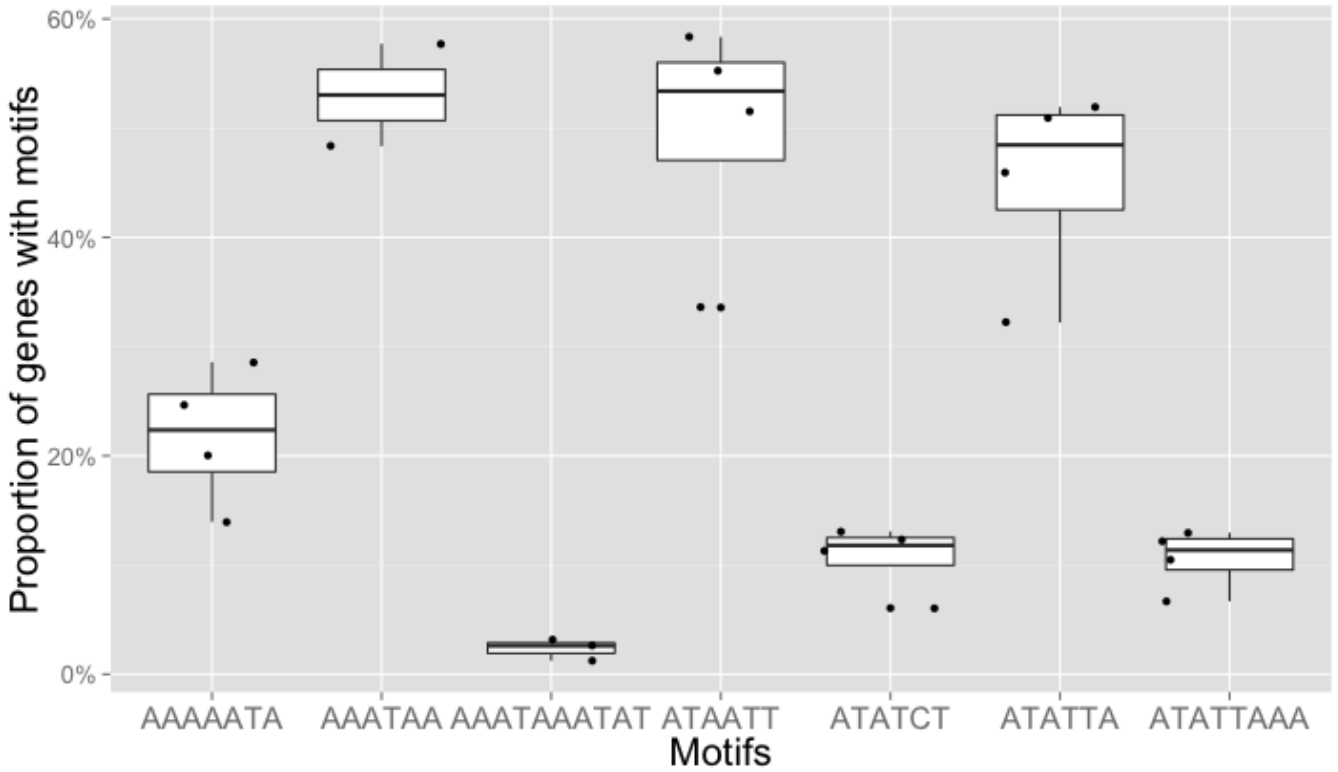
59 unique motifs common in all 4 species

More Results!

Took motifs found in all 4 species

Number of genes having motif / Total number of genes





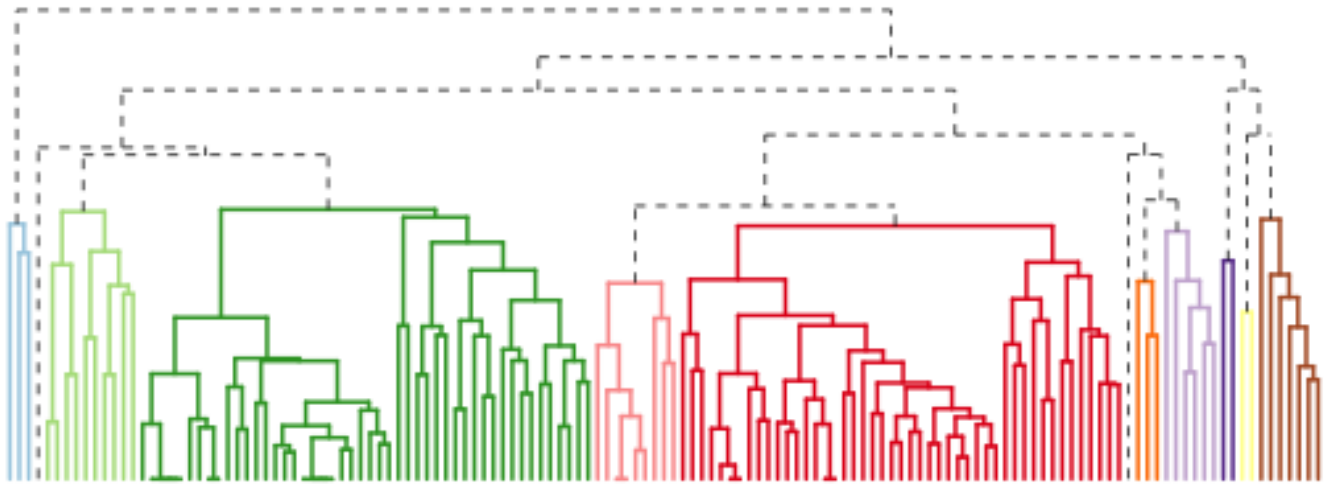
Perspectives/Caveats

Motif clustering (decreases redundancy)

Run pipeline with various thresholds

Motif extraction method for + and - strand

Look orthology family motifs conservation



New Learnings?

GitHub

Version-Control

Online

Open Source

Modular

Synchronize & Share easily

Code history

```
| * | commit 66f3276356e955302ac20333c34d6df1a3128180
| | | Author: Rekyt <matthias.grenie@ens-lyon.fr>
| | | Date:   Wed Jun 18 17:53:09 2014 -0400
| | |
| | |     Reversed all motifs for analysis
| | |
| * | commit 5982d26cc56cbe8e3133910aa70ae181bebdb43d
| | | Author: Rekyt <matthias.grenie@ens-lyon.fr>
| | | Date:   Wed Jun 18 17:52:54 2014 -0400
| | |
| | |     Simple list of motifs found in MEMEmotifsjune1714.txt
| | |
* | | commit e706ef01dce18e445f2cac338e75106ff3ca6268
| | | Author: Rekyt <matthias.grenie@ens-lyon.fr>
| | | Date:   Wed Jun 18 18:34:54 2014 -0400
| | |
```

Acknowledgements

I would like to thank:

Michael Lynch for having me in the lab even with all the French forms to fill...

Marie Sémon (without whom I wouldn't be here)

Jean-François Goût, for his constant and kind support

All of the Lynch lab team for the welcome and the discussions

Long live to cake day!

Questions?