Intro to PAML/codeML for estimating dN/dS

Catherine Armbruster
Postdoc in Jen Bomberger's Lab
Pitt Dept of Micro & Molecular Genetics (MMG)

Microbial Genomics Workshop
8 Jan 2021

@catarmbruster

Many slides adapted from:

Arkadiy Garber (ASU)

github.com/Arkadiy-Garber

@TheIronArk

Bioinformatics Virtual Coordination Network

biovcnet.github.io

PAML (Phylogenetic Analysis by Maximum Likelihood)

A program package by Ziheng Yang

(Demonstration by Joseph Bielawski)

abacus.gene.ucl.ac.uk/software/paml.html

Other popular tools for dN/dS:

HyPhy
Hypothesis Testing using Phylogenies

hyphy.org & galaxy.hyphy.org

MrBayes

nbisweden.github.io/MrBayes/index.html

paup.phylosolutions.com/

Outline

- Example of how I have used PAML in my research
- Background on dN/dS and PAML
- Hands-on demonstration of a simple PAML/codeML run in Jupyter notebook

P. aeruginosa evolves in the airways of people with cystic fibrosis (CF) over their lifetime

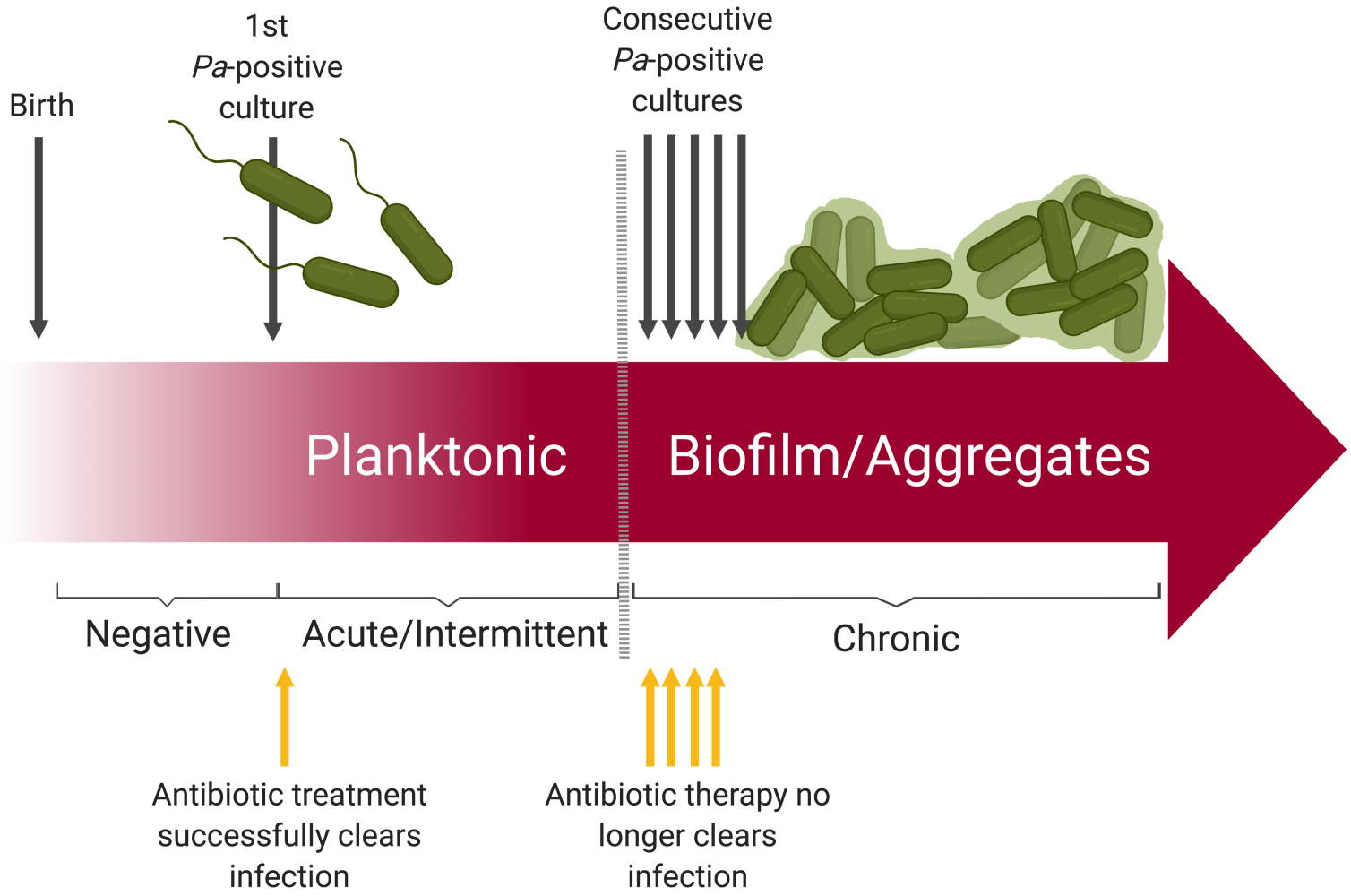
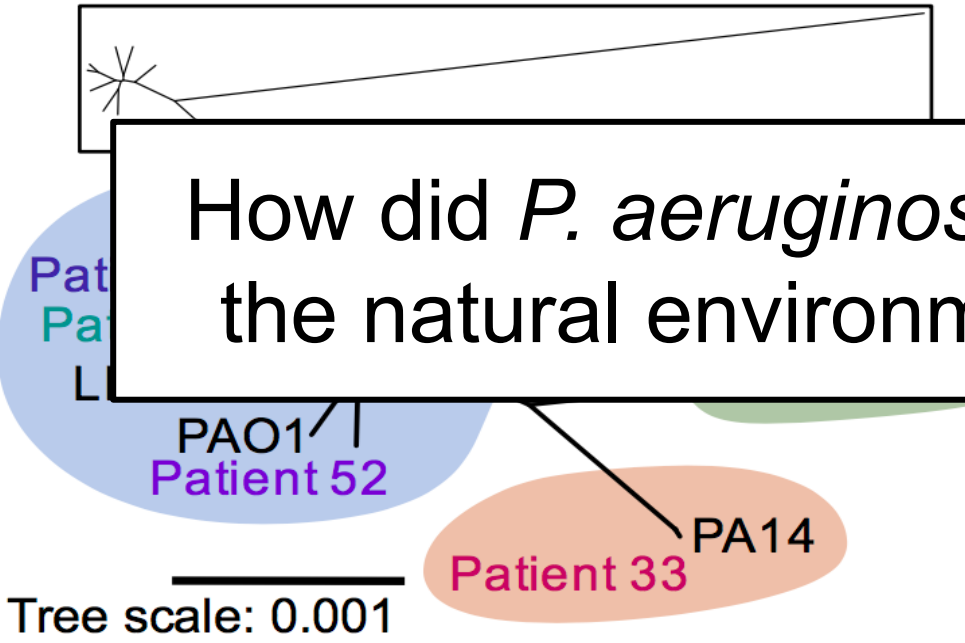
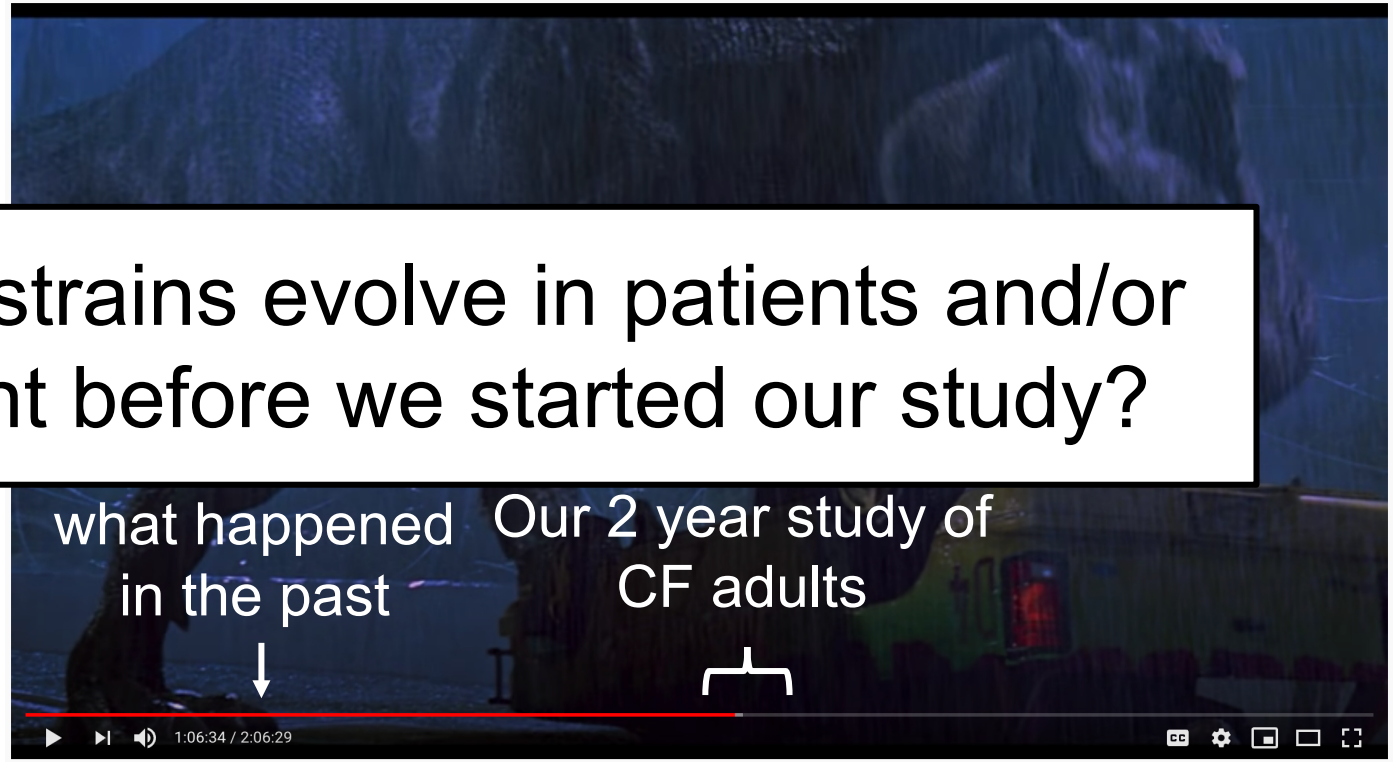


Figure: Allison Welp

CF adults' *P. aeruginosa* isolates were already highly host-adapted before we started our 2-year study



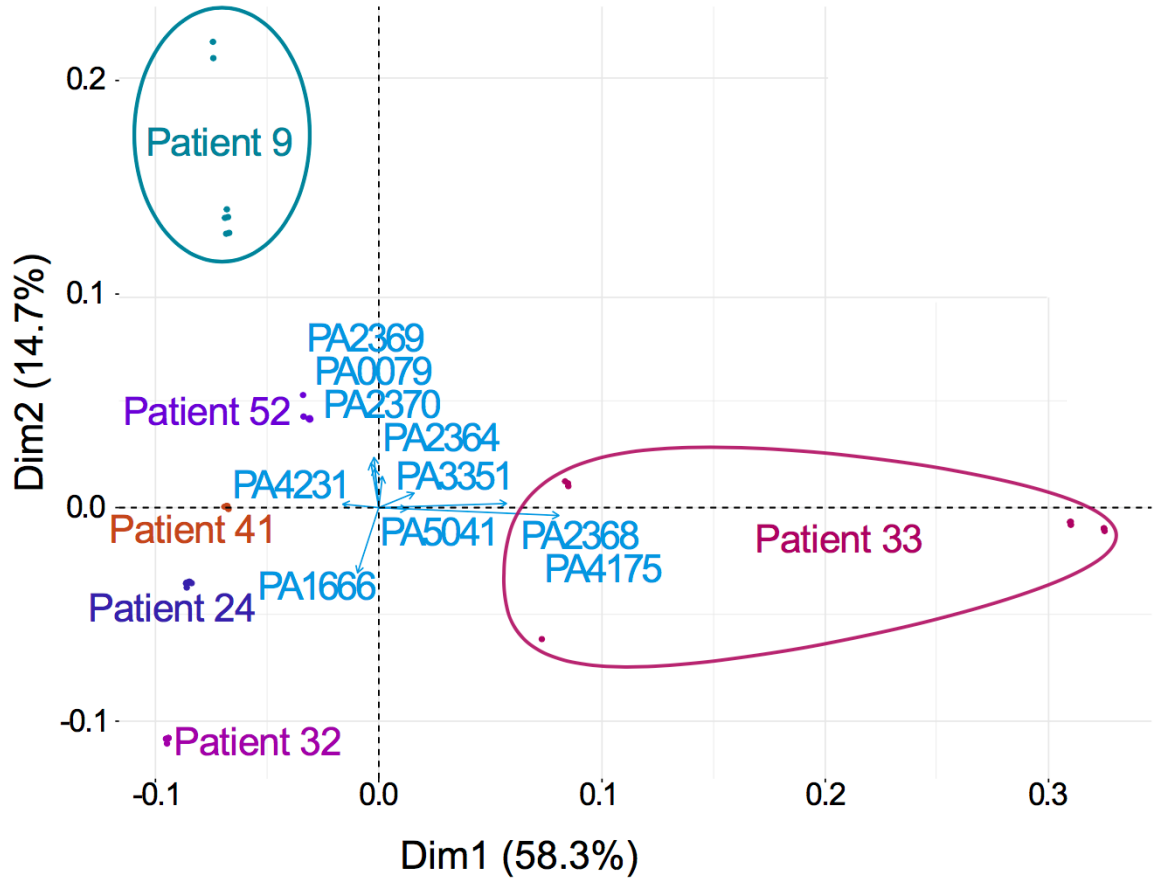
How did *P. aeruginosa* strains evolve in patients and/or the natural environment before we started our study?



Jurassic Park

Armbruster, *et al*. bioRxiv 2020.

We used codeML to estimate dN/dS for all genes in all patients' *P. aeruginosa* lineages



- PCA plot of individual isolates based on dN/dS values of >200 virulence-associated genes (>400,000 data points)
- Vectors = top 10 genes that differed between strains
  - Type VI secretion, motility, nutrient acquisition

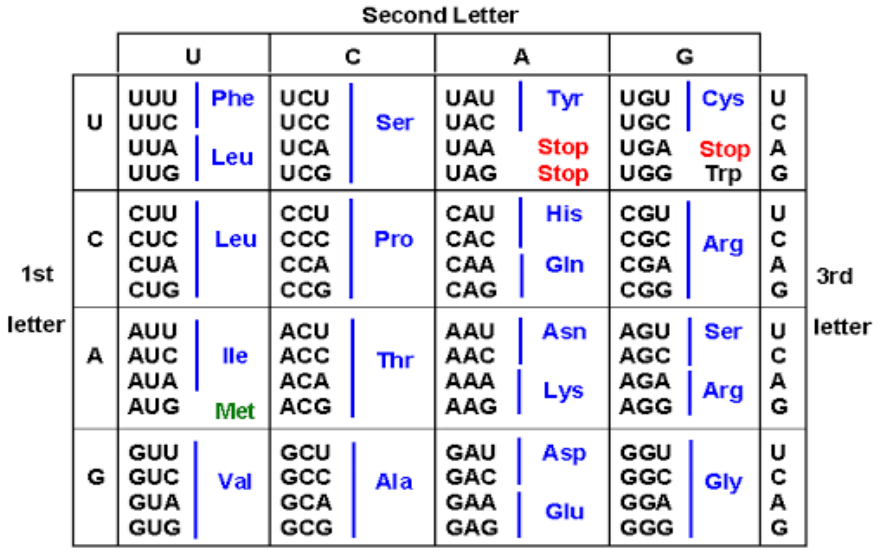
Armbruster, *et al*. bioRxiv 2020.

Synonymous vs. Nonsynonymous mutations

- **Synonymous mutation**: change in the nucleotide sequence, but no change in the encoded peptide.
  - Provides a metric of neutral evolution of the sequence

- **Nonsynonymous mutation**: change in the nucleotide sequence that also results in a change in amino acid sequence.
  - Provides a metric for the conservation of protein sequence, in the context of the neutral changes estimated by dS (the rate of synonymous substitutions)

The genetic code is degenerate (redundant)



1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA UCG Ser	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU His CAC CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU Asn AAC AAA AAG Lys	AGU Ser AGC AGA AGG Arg	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU Asp GAC GAA GAG Glu	GGU GGC GGA GGG Gly	U C A G

<u>dN/dS (aka ω)</u>

dS = synonymous mutations per possible synonymous sites

dN = nonsynonymous mutations per possible nonsynonymous sites

dN/dS → infer conservation or divergence of a particular peptide sequence given the amount of neutral (synonymous) genetic drive that has occurred

Figure: *http://biology.kenyon.edu/courses/biol114/Chap05/Chapter05.html*

Estimating selection using dN/dS

dN/dS = 1 → = neutral selection of entire sequence

High dN/dS (> 1) = high rate of nonsynonymous substitutions, given a certain rate of synonymous substitutions = **positive selection and adaptive evolution**

Low dN/dS (< 1) = low rate of nonsynonymous substitutions, given a certain rate of synonymouns substitutions = **purifying selection**

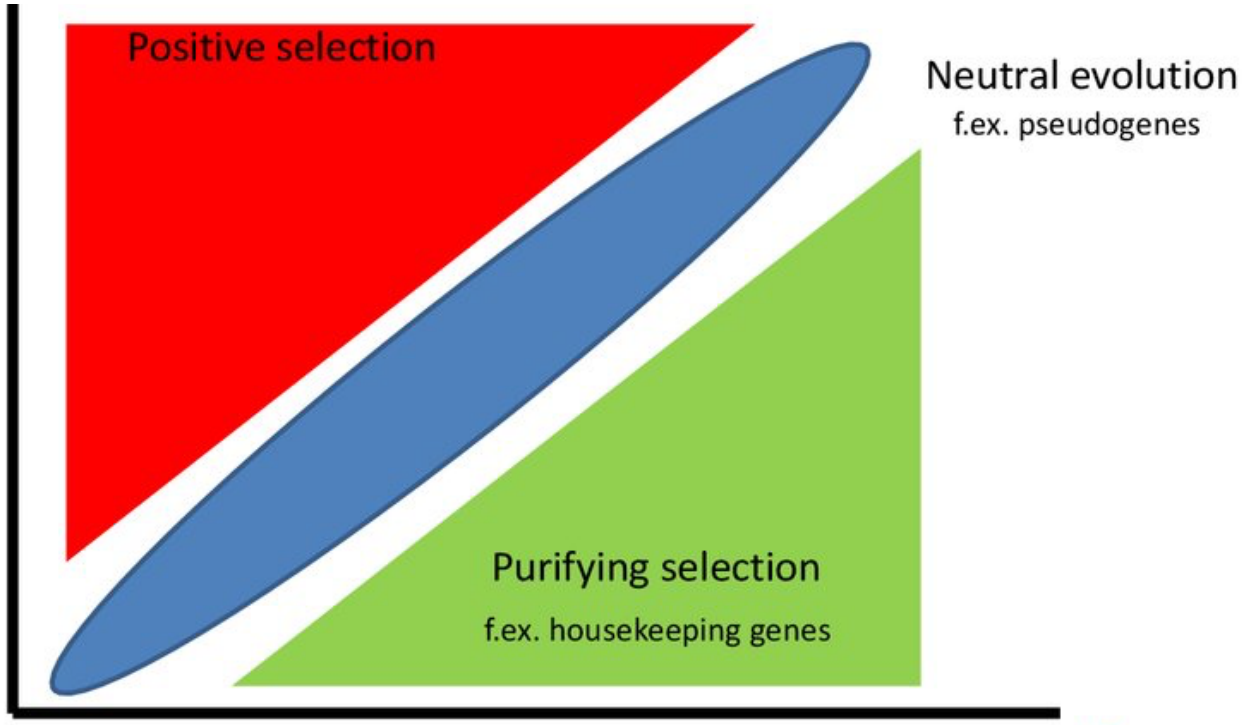
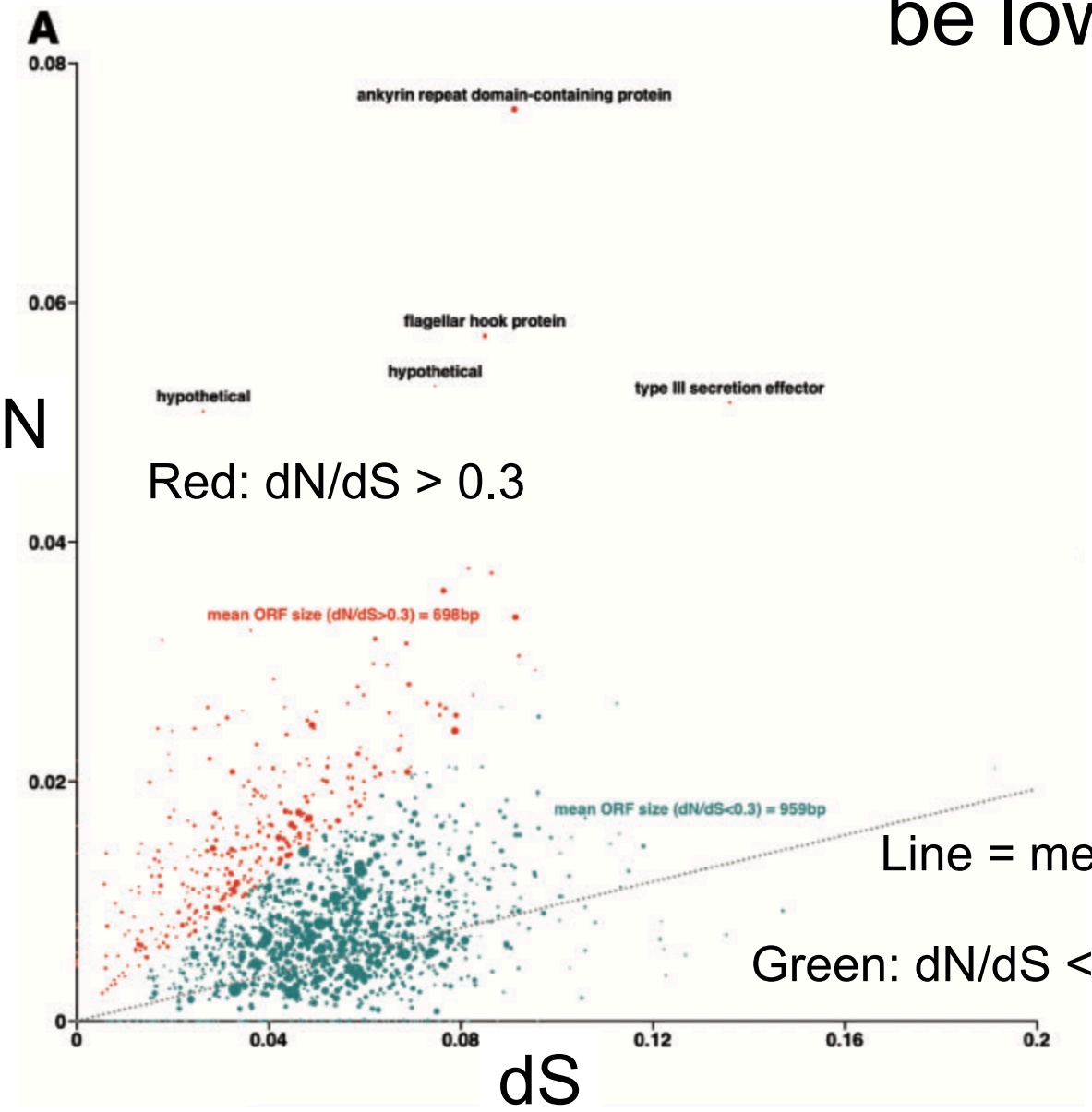


Figure: Max Planck Institute IMPRS seminar summer term 2020

In reality, dN/dS values for whole genes tend to be lower



Red: dN/dS > 0.3

Line = mean dN/dS

Green: dN/dS < 0.3

Usually:

High dN/dS (> 0.2/0.3) = neutral or positive selection

Low dN/dS (< 0.2) = purifying selection

Oakeson et al., Genome Biol. Evol. 2014.

Limitations of dN/dS

- Assumes synonymous mutations are silent/not selectable
- Not appropriate for non-coding sequences (rRNA, tRNA, tmRNA, other types of small regulatory RNAs)
- Mutational saturation can lead to back-mutations that underestimate divergence (recommended dS < 1)
- Near-identical sequences with only a few SNPs can lead to over-estimation of dN/dS → (recommended dS > 0.01)

OPEN ACCESS Freely available online

PLoS GENETICS

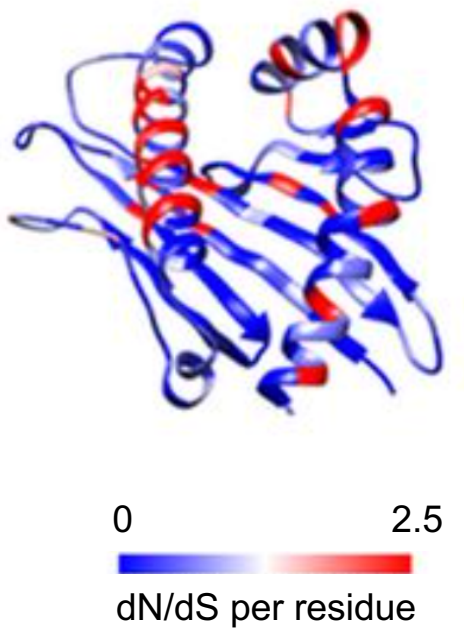
The Population Genetics of dN/dS

Sergey Kryazhimskiy[1], Joshua B. Plotkin[1,2]*

1 Biology Department, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **2** Program in Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Limitations of dN/dS

- Assumes synonymous mutations are silent/not selectable
- Not appropriate for non-coding sequences (rRNA, tRNA, tmRNA, other types of small regulatory RNAs)
- Mutational saturation can lead to back-mutations that underestimate divergence (recommended dS < 1)
- Near-identical sequences with only a few SNPs can lead to over-estimation of dN/dS → (recommended dS > 0.01)
- If estimating one dN/dS for a gene, the signal of one position or domains under high positive selection can be lost, if the rest of the sequence is conserved
  - Site models address this



Aswad *et al.*, J of Virol. 2018.

Intro to PAML (Phylogenetic Analysis by Maximum Likelihood)

PAML and dependencies can be downloaded and installed in a conda environment:

```
mmg-bomberg-16:~ caa78$ conda create -n PAML pal2nal muscle paml --yes
```

More PAML resources:

PAML website & download link:
http://abacus.gene.ucl.ac.uk/software/PAML.html

PAML FAQ:
http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf

PAML discussion/bug reports:
http://www.rannala.org/phpBB2/

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

Z. Yang. 1997. Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. Mol Biol Evol. 15(5):568-73.

What can PAML do?

- Estimate synonymous and nonsynonymous rates
  - Detect positive selection in protein-coding DNA sequences with likelihood ratio tests
- Compare different evolutionary models or assumptions
- Compare and test phylogenetic trees
- Estimate divergence times with different clock models
- Ancestral sequence reconstruction (DNA, codon, or AAs)
- Monte Carlo simulation of nucleotide, codon, or AA sequence datasets

Programs in the PAML package

baseml	for bases
basemlg	continuous gamma for bases
codeml	aaml for amino acids & codonml for codons
evolver	simulation, tree distances
yn00	d_N and d_S by Yang & Nielsen (2000)
chi2	chi square table
pamp	parsimony (Yang and Kumar 1996)
mcmctree	Bayesian MCMC divergence time estiamtion, under soft bounds (Yang & Rannala 2006)

PAML website

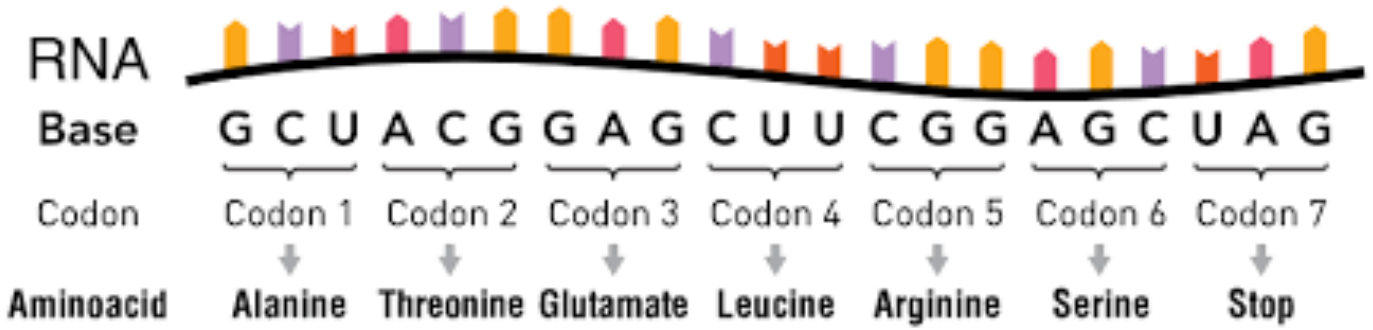
PAML takes 3 files as inputs:

1) Sequence data file (codon alignment)

2) Tree file (optional)

3) Control file (*.ctl)

1) Codon alignment = nucleotide sequence arranged in triplets (as codons)



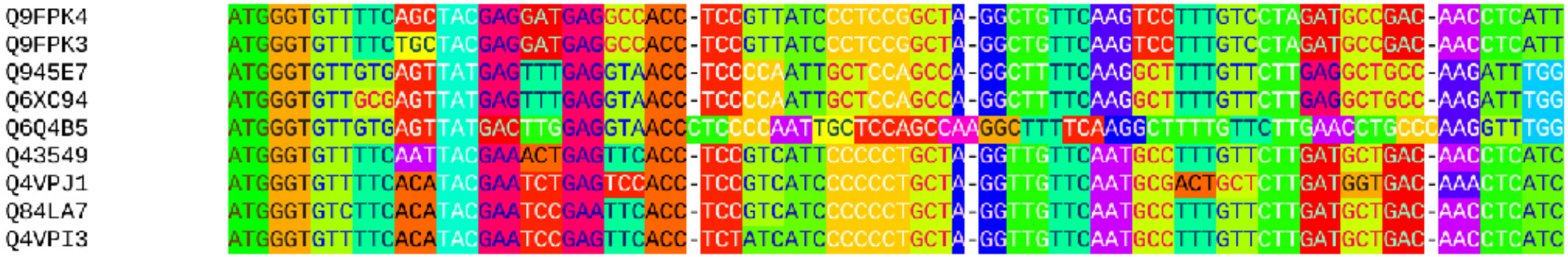
pal2nal.pl is a tool that makes a codon alignment

Inputs to pal2nal:

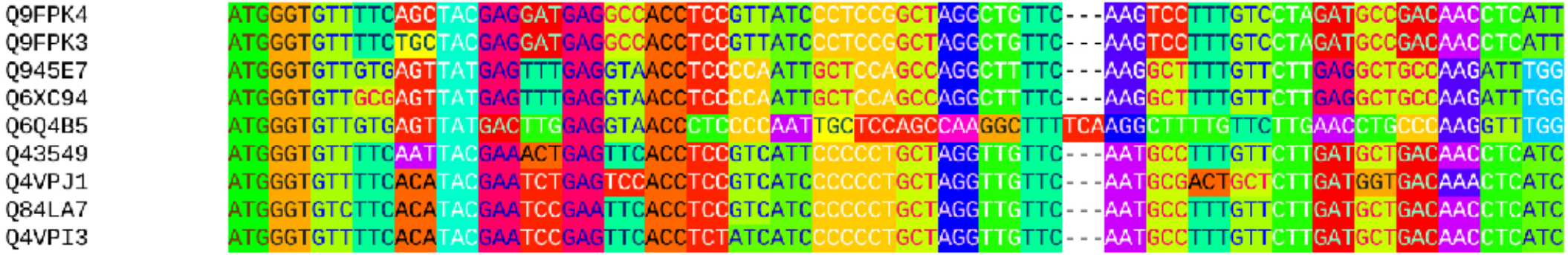
- protein alignment (e.g. MUSCLE aligned .faa)
- associated nucleotide sequences with same headers (after the “>”)

Example of a codon alignment improved by first aligning amino acids

Alignment only using nucleotide sequences (.ffn)

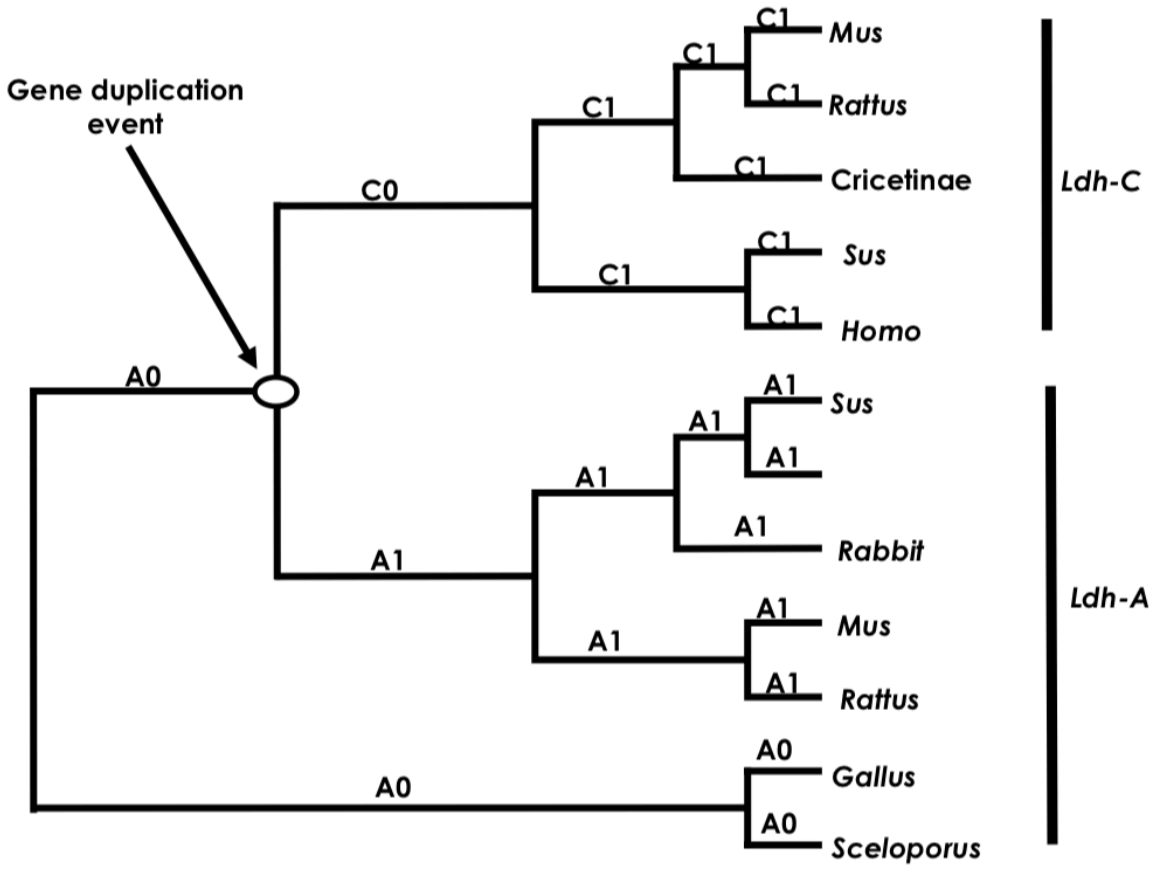


Aligned amino acids first (.faa), then used that to arrange .ffn into codons for alignment



DNATagger Sherrer and Basso, Gen & Molec Res 2008

2) Phylogenetic tree – optional input as parenthetic (Newick) format



E.g. For use with branch and branch-site models.

Does the strength and direction of selection acting on Ldh-C differ from Ldh-A?

Figure: PAML website

3) **codeml.ctl** file – contains locations of input and output files, as well as parameters

codeml-2.ctl

```
     seqfile = zinc_transporter_ZitB.codonalign.fna * sequence data filename
     treefile = gtotree_out3.tree
     outfile = mlc           * main result file name

       noisy = 0  * 0,1,2,3,9: how much rubbish on the screen
     verbose = 1  * 0: concise; 1: detailed, 2: too much
     runmode = -2  * 0: user tree;  1: semi-automatic;  2: automatic
                 * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

     seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
   CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

*       ndata = 10
       clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
      aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
  aaRatefile = dat/jones.dat  * only used for aa seqs with model=empirical(_F)
                 * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

       model = 2
                 * models for codons:
                     * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                 * models for AAs or codon-translated AAs:
                     * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                     * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

     NSsites = 0  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                 * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                 * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                 * 13:3normal>0

       icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
       Mgene = 0
                 * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
                 * AA: 0:rates, 1:separate

    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 2  * initial or fixed kappa
    fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
        omega = .4 * initial or fixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 8  * # of categories in dG of NSsites models

        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 1  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 1  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Table 2. Parameters in the site models

Model	NSsites	#p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$
M3 (discrete)	3	5	p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$

E.q. Run with NS sites = 1 2 to detect positive selection on a gene

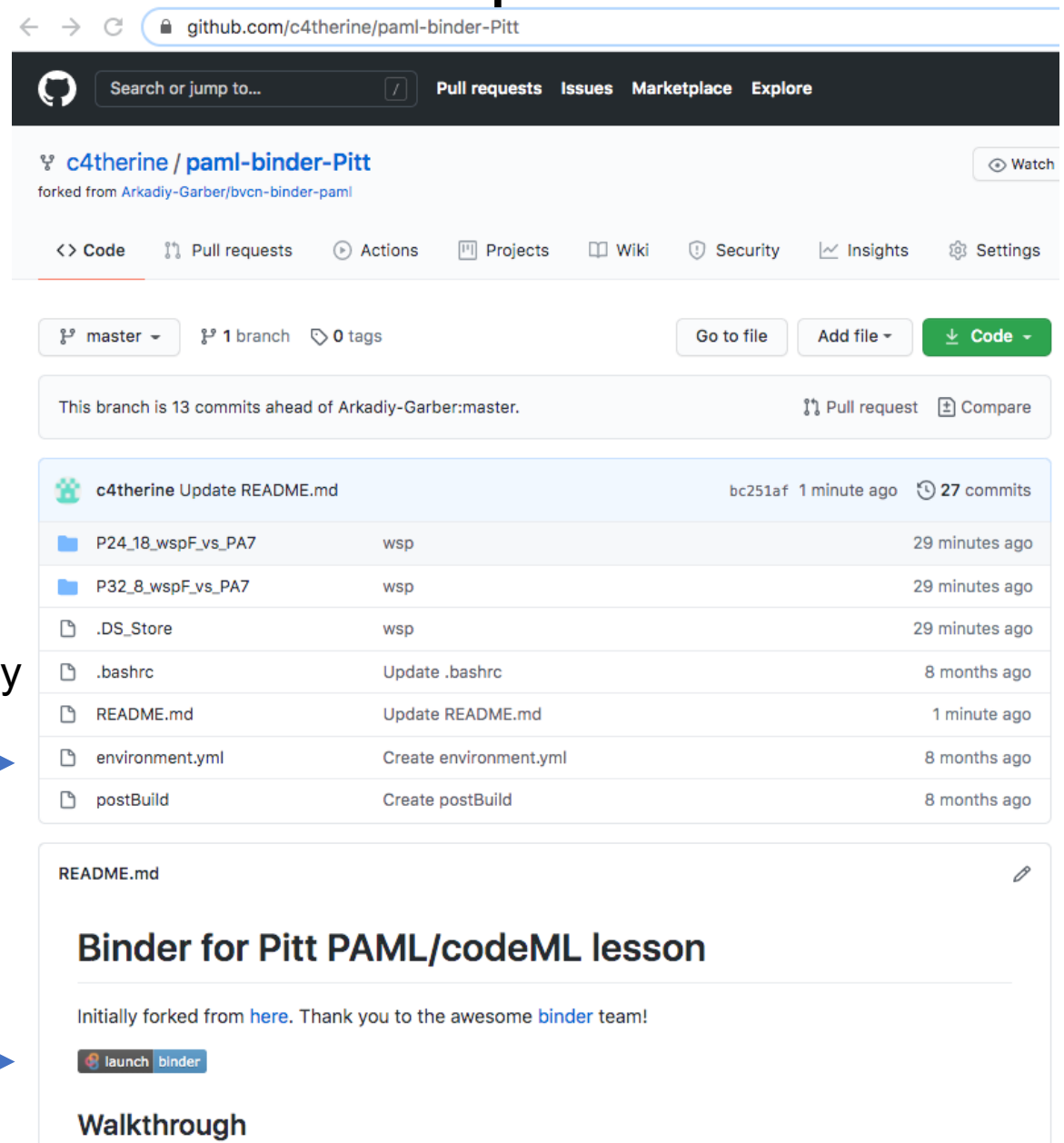
- Compare likelihood of value from the neutral model (M1a) to M2a with likelihood ratio test

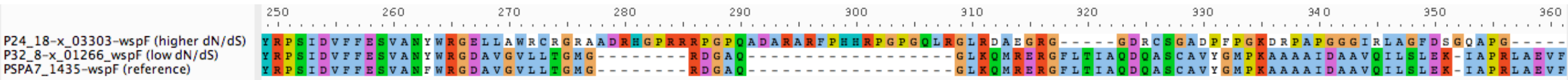
Jupyter binder tutorial – Estimate dN/dS by pairwise comparisons

Github url

Dependencies (already loaded in binder)

Link to Jupyter binder




250
260
270
280
290
300
310
320
330
340
350
360
P24_18-x_03303-wspF (higher dN/dS)
P32_8-x_01266_wspF (low dN/dS)
PSPA7_1435-wspF (reference)