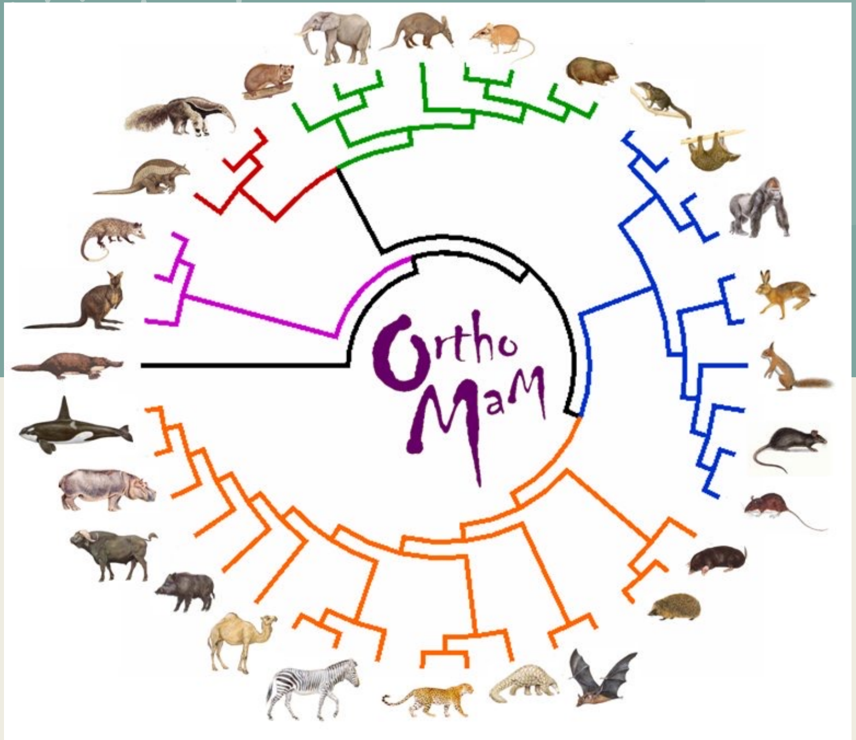
Comparative Genomics

PhyloGenomics Lab



Finding genes involved in a trait using comparative genomics

Prof. Nathan Clark

University of Pittsburgh

About me...

B.S. in Organic Chemistry

- University of Texas at Austin

PhD in Genome Sciences

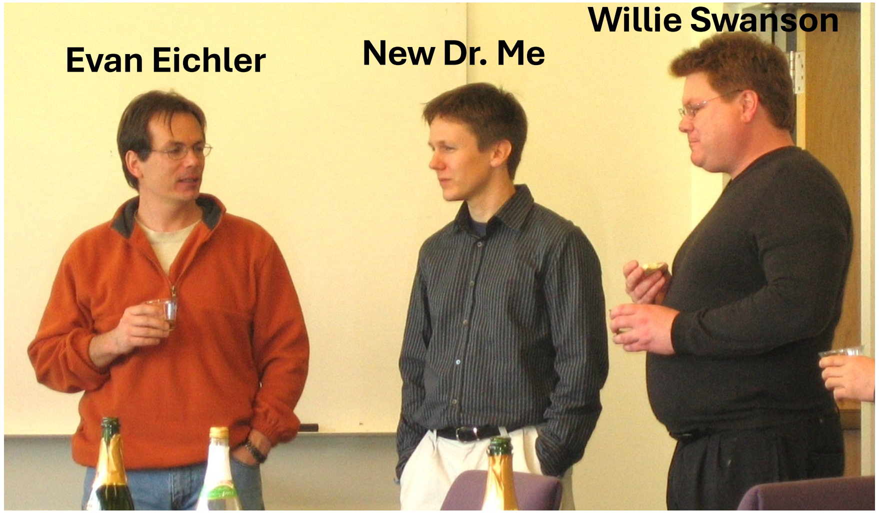
- University of Washington – Seattle
- Willie Swanson's lab

Postdoc

- Cornell with Chip Aquadro

Professorship

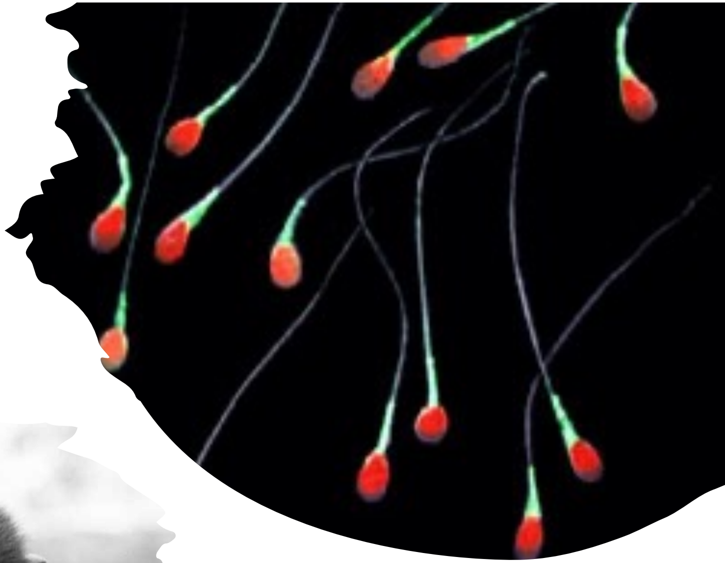
- University of Pittsburgh

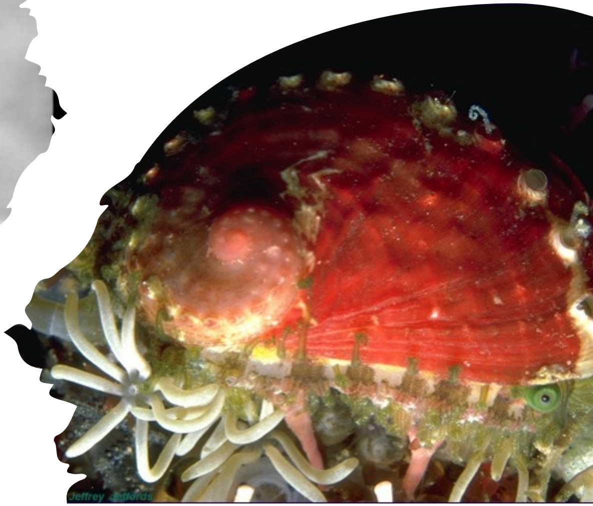


Early studies in the adaptive evolution of reproductive proteins

- Cooperation Conflict and Competition within and between the sexes
- Abalone, butterflies, primates

- This was all in early genomic era so often Sanger sequencing to compare to a reference species
- Then, more and more species genomes led to growth of comparative genomics…

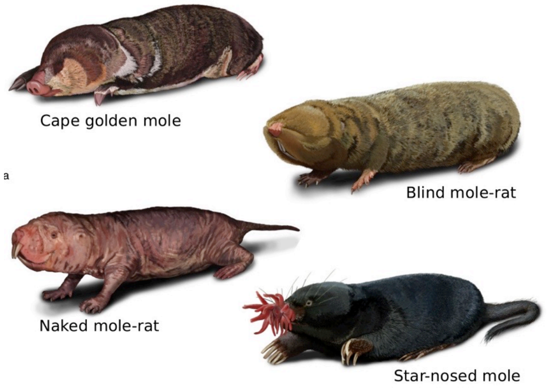




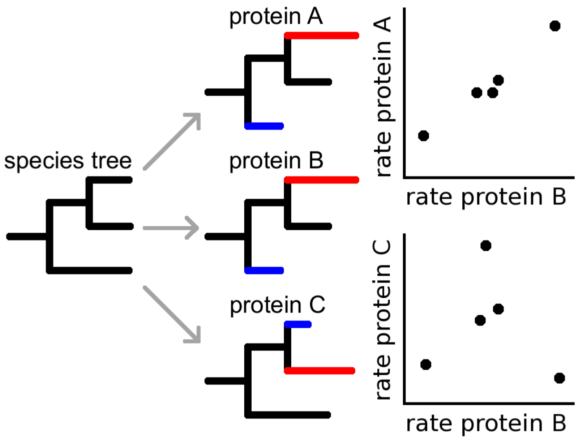
Jeffrey Jeffords

Known for phylogenetic comparative methods and applications

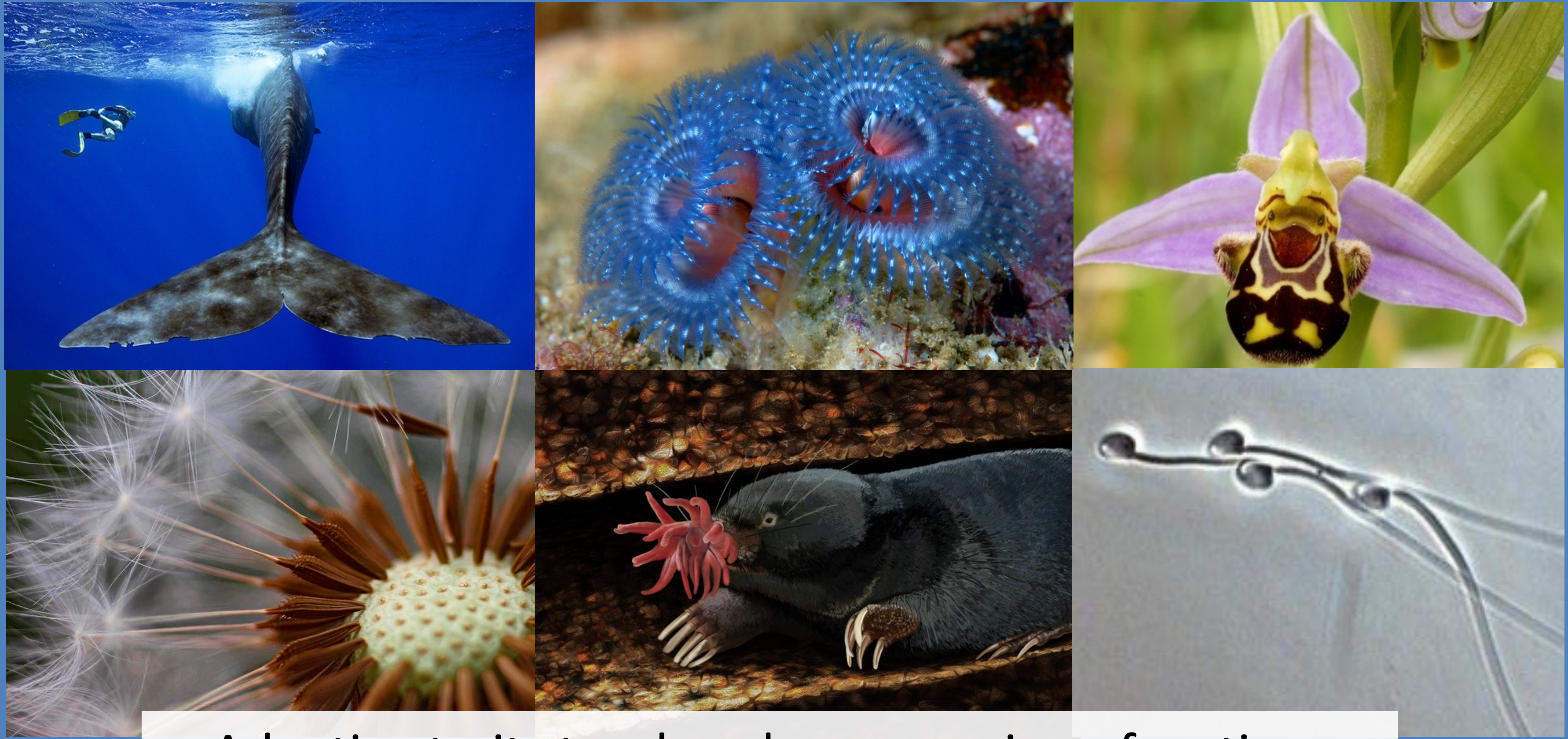
- Linking phenotypes to genotypes through comparative species analysis
  - RERconverge
  - BUSTED-PH



- Connecting genes into functional networks with evolutionary patterns
  - Evolutionary Rate Covariation (ERC)



Evolutionary Adaptations



- Adaptive traits teach us how organisms function.
- How do we assign genetic regions to phenotypes?

Outline

Today

- Understanding phylogenic trees
- Measuring molecular divergence
- Inferring phylogenies
- Natural Selection
- Methods to infer adaptation
- Lab structure

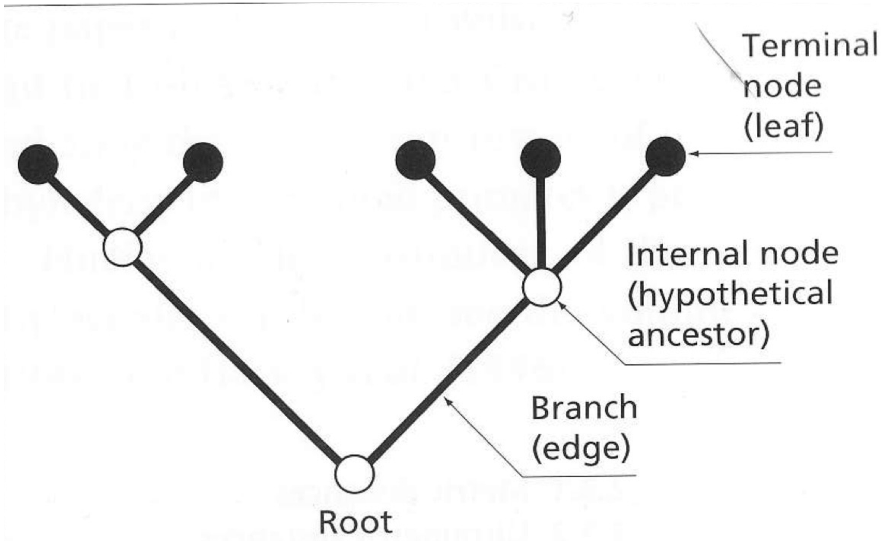
On Friday

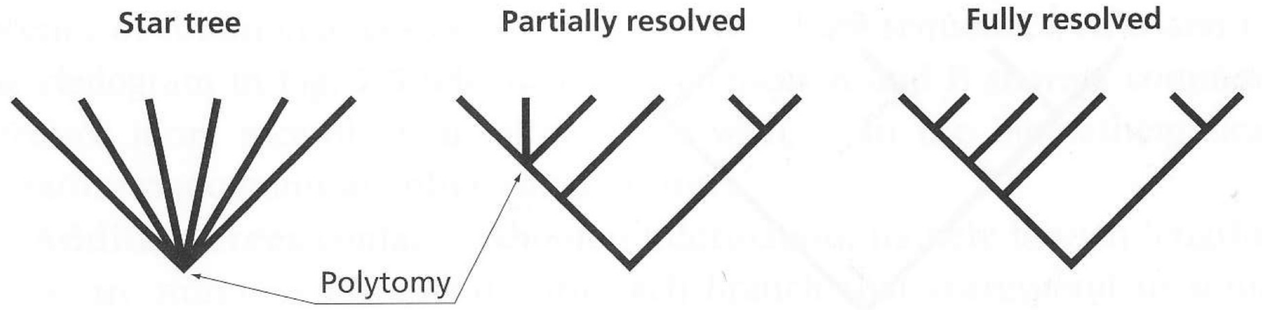
- – How to begin comparative genomics studies
- – Finding and making multiple sequence alignments of orthologs
- – Many many more comparative genomics approaches

Phylogenetic Trees

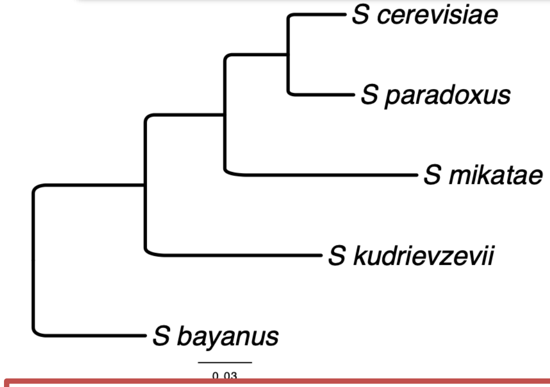
- Phylogenetics is a powerful set of methods to describe change between species or genes.
  - Relationships between species or genes are inferred.
  - Specific hypotheses about origin and function can be tested using phylogenetics.

Tree Terminology

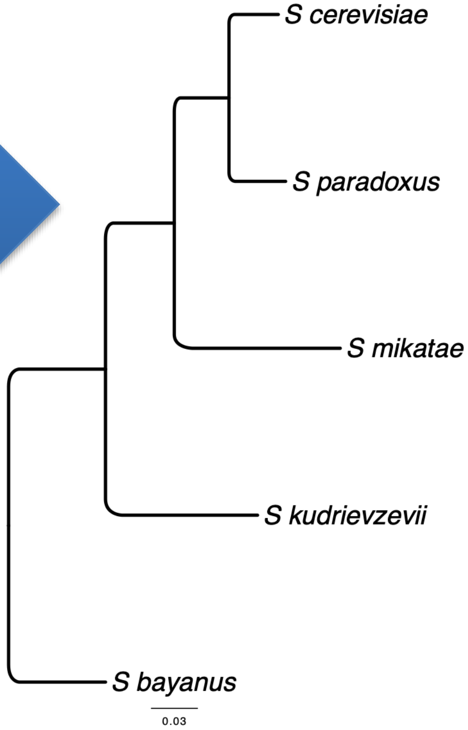




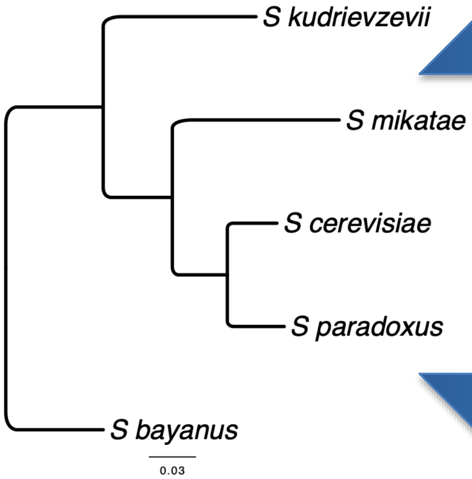
Rectangular trees



Vertical spacing is meaningless.

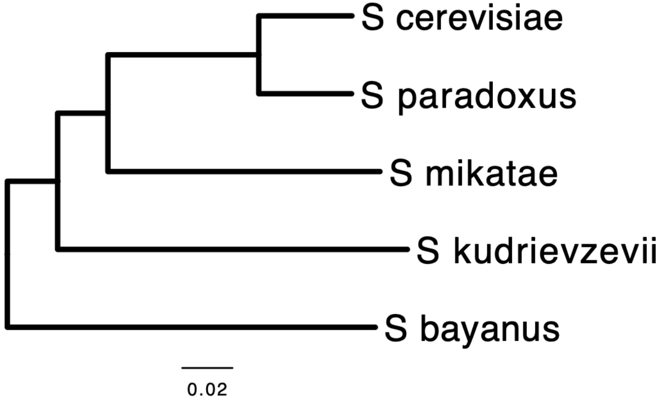


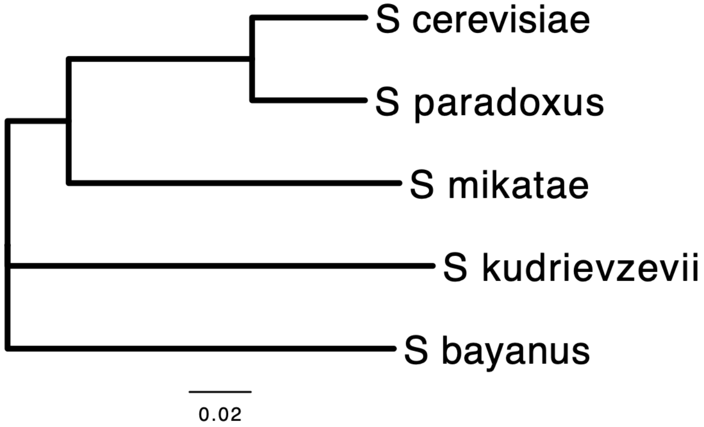
Only horizontal spacing indicates time/genetic distance.



Vertical arrangement of leaves has no effect.

Rooted *versus* Unrooted





Rooted

N species

2N – 2 branches

N – 1 internal nodes

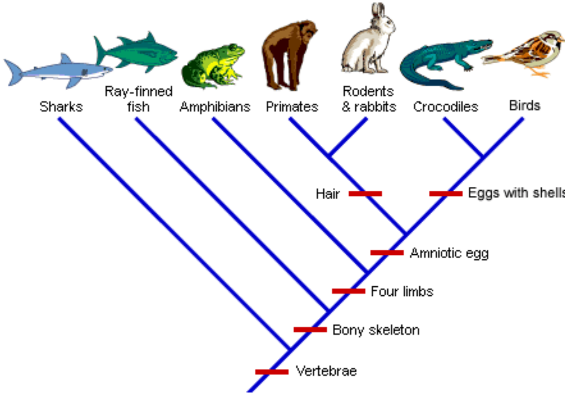
Unrooted

N species

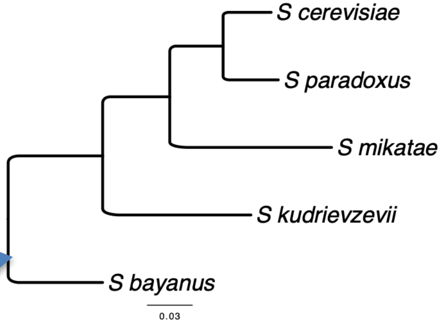
2N – 3 branches

N – 2 internal nodes

Cladogram – only intended to show connectivity, branch lengths do not reflect amount of divergence



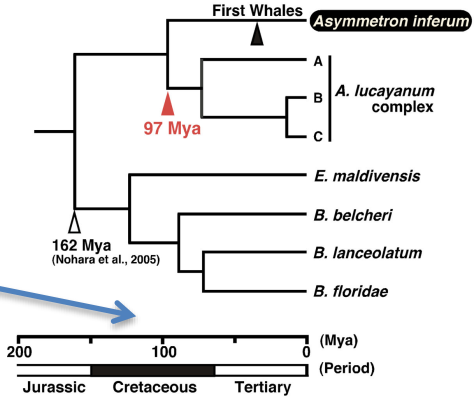
Phylogram – both connectivity and branch lengths are represented. Branch length could be genetic distance



or

Time (ultrametric tree)

Tips are always aligned.



Writing trees in Newick format

Parentheses surround clades

Commas separate taxa

:numbers are branch lengths

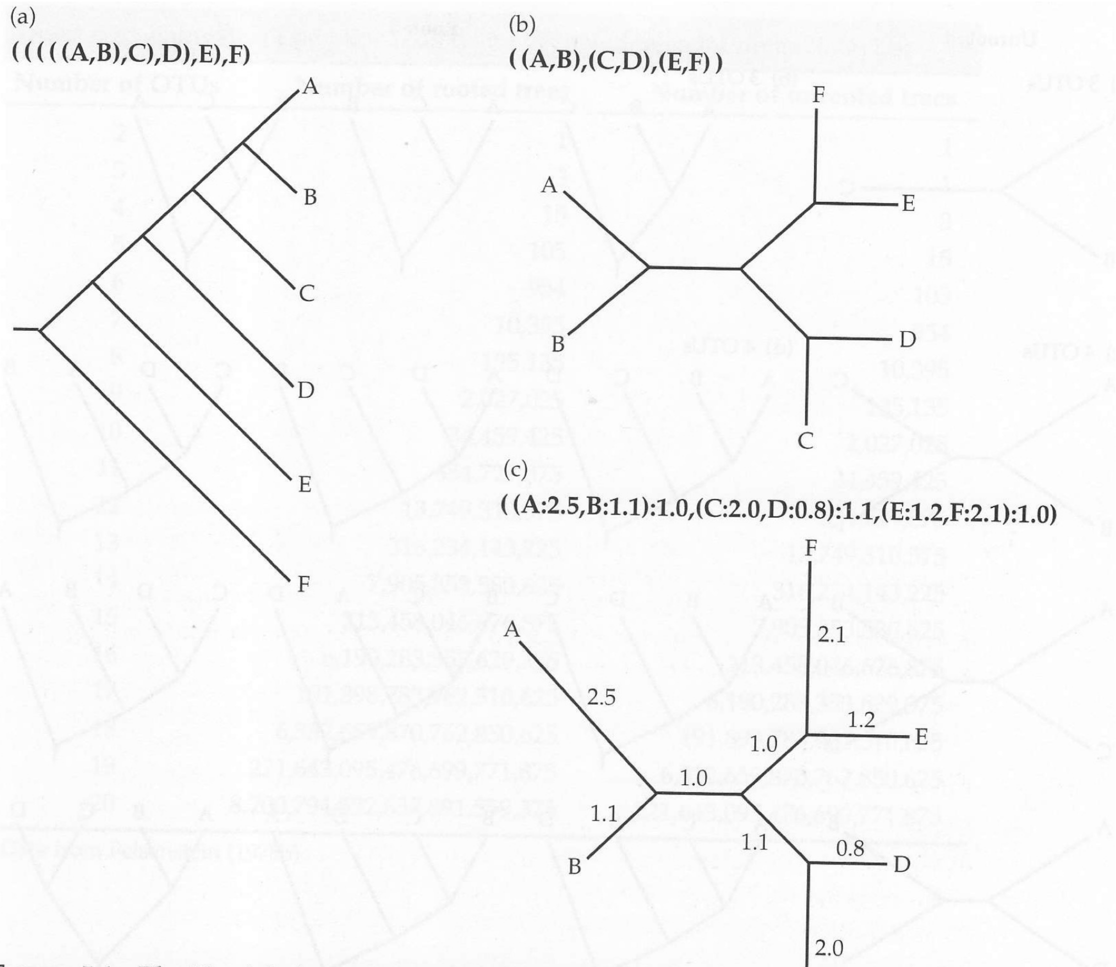
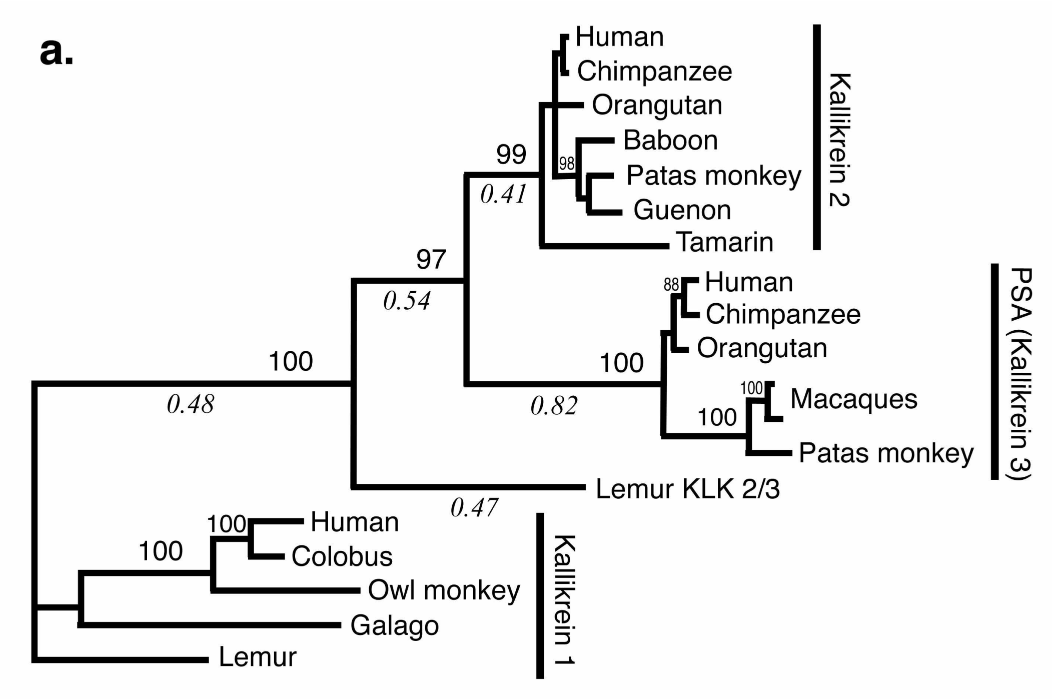


FIGURE 5.4 The Newick format representation for rooted

Orthology and Paralogy



Orthologous genes are those that diverged due to a **speciation event**.

- Orthologous groups will tend to “recreate” the species tree.

Paralogs are those that diverge due to **gene duplication**.

Measuring Molecular Divergence

Genetic distance = metric of sequence evolution.

Evolutionary rate = metric of change per unit time.

Why measure divergence?

Rate reveals functional clues via conservation and adaptation.

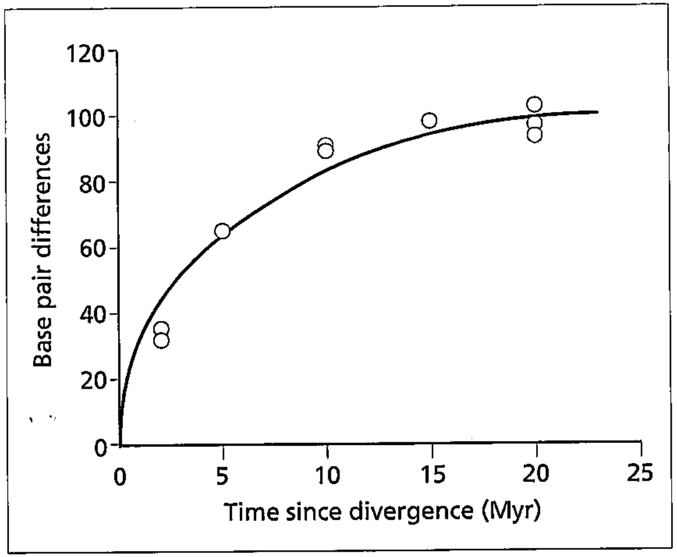
Rate is crucial for phylogenetic inference

- determining species relationships
- dating evolutionary events with the
  "molecular clock"

Percent divergence (%) is the simplest metric of divergence

It’s intuitive and good for broad audiences.
But the **major problem** is that it underestimates divergence after only a moderate amount of change.

Fig. 5.11 Number of nucleotide substitutions between pairs of bovid mammal mitochondrial sequences (684 basepairs from the *COII* gene) against estimated time of divergence. Notice that the observed number of substitutions is not linear with time but curvilinear. Data from Janecek *et al.* (1996).



← Saturation

The bending of this curve

<u>Saturation</u> is when divergence reaches a level that observed differences will be lower than the actual number of changes.

Saturation bias occurs because some types of substitutions over branches are “invisible” when simply studying extant sequences.

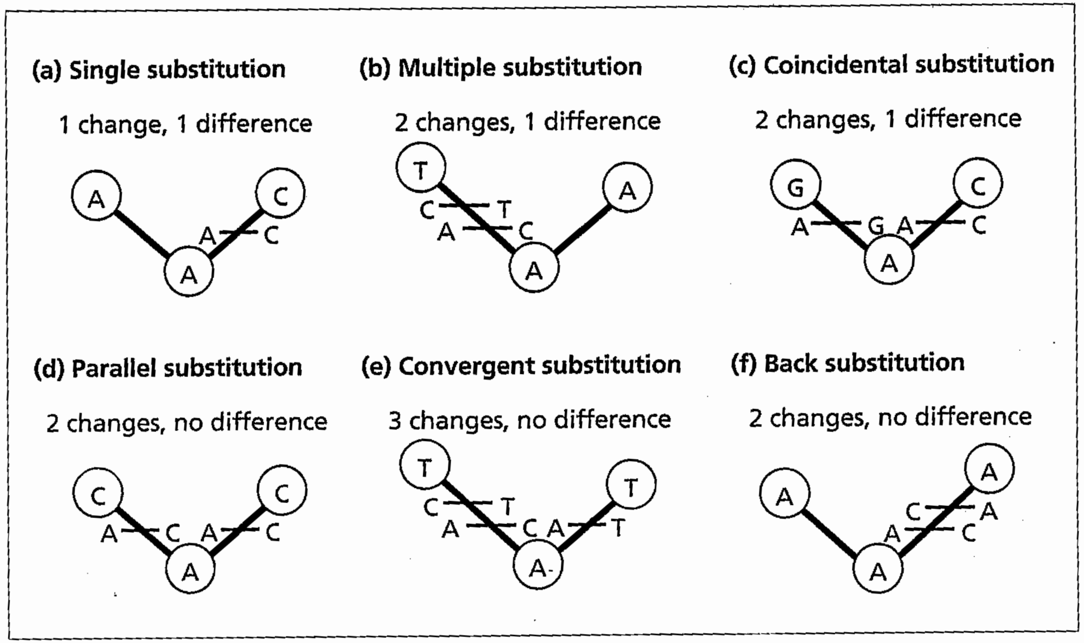


Fig. 5.9 Six kinds of nucleotide substitution. In each case the ancestral nucleotide was A. In all except the case of a single substitution, the number of substitutions that actually occurred is greater than would be counted if we just compared the two descendant sequences. In the lower three cases the nucleotides are identical in both descendant sequences, but this similarity has not been directly inherited from the ancestral sequence. Such similarity is termed ‘homoplasious’.

Saturation

Saturation causes a systematic underestimating bias when measuring molecular divergence.

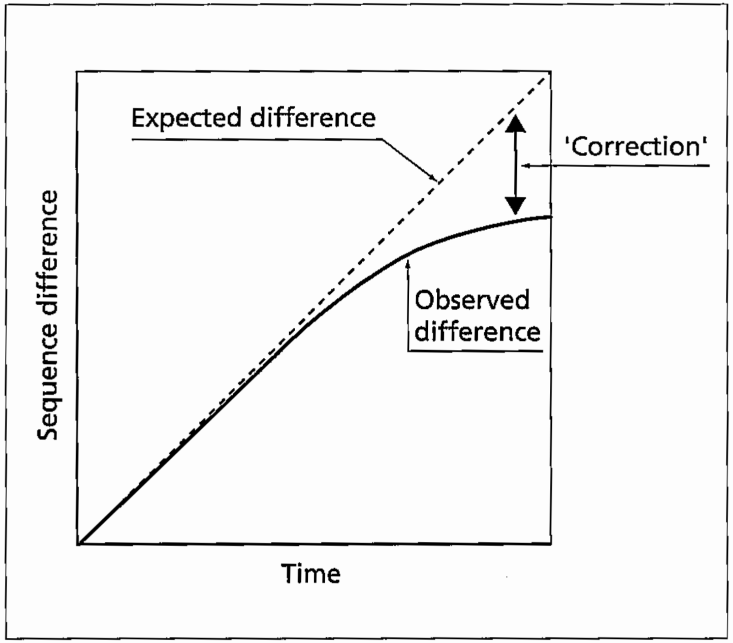


Fig. 5.12 The need to correct observed sequence differences. The extent of observed differences between two sequences is not linear with time (as we would expect if the rate of molecular evolution is approximately constant) but curvilinear due to multiple hits. The goal of distance correction methods is to recover the amount of evolutionary change that the multiple hits have overprinted and to 'correct' the distances for unobserved hits. In effect, the methods seek to 'straighten out' the line representing observed differences.

Nucleotide Substitution Models

- To make functional inferences we typically don’t model insertions and deletions, large or small, because of the difficulty in assigning homology.

- We model nucleotide substitutions under the assumptions:
  - They occur as single, independent events
  - They don’t affect substitutions at other sites

- More general models account for natural differences in rates between nucleotides

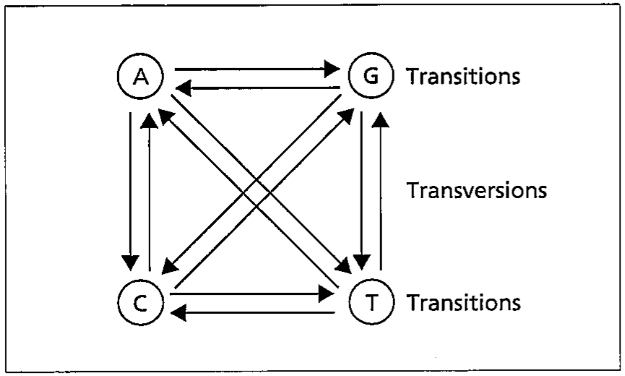
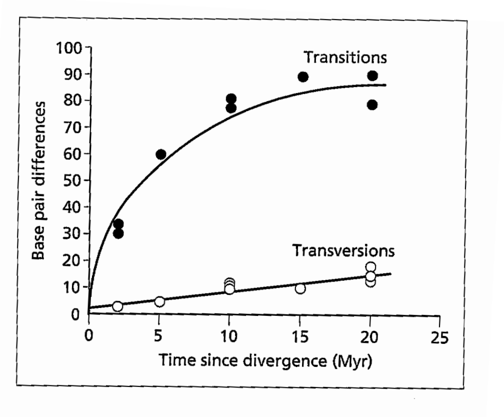
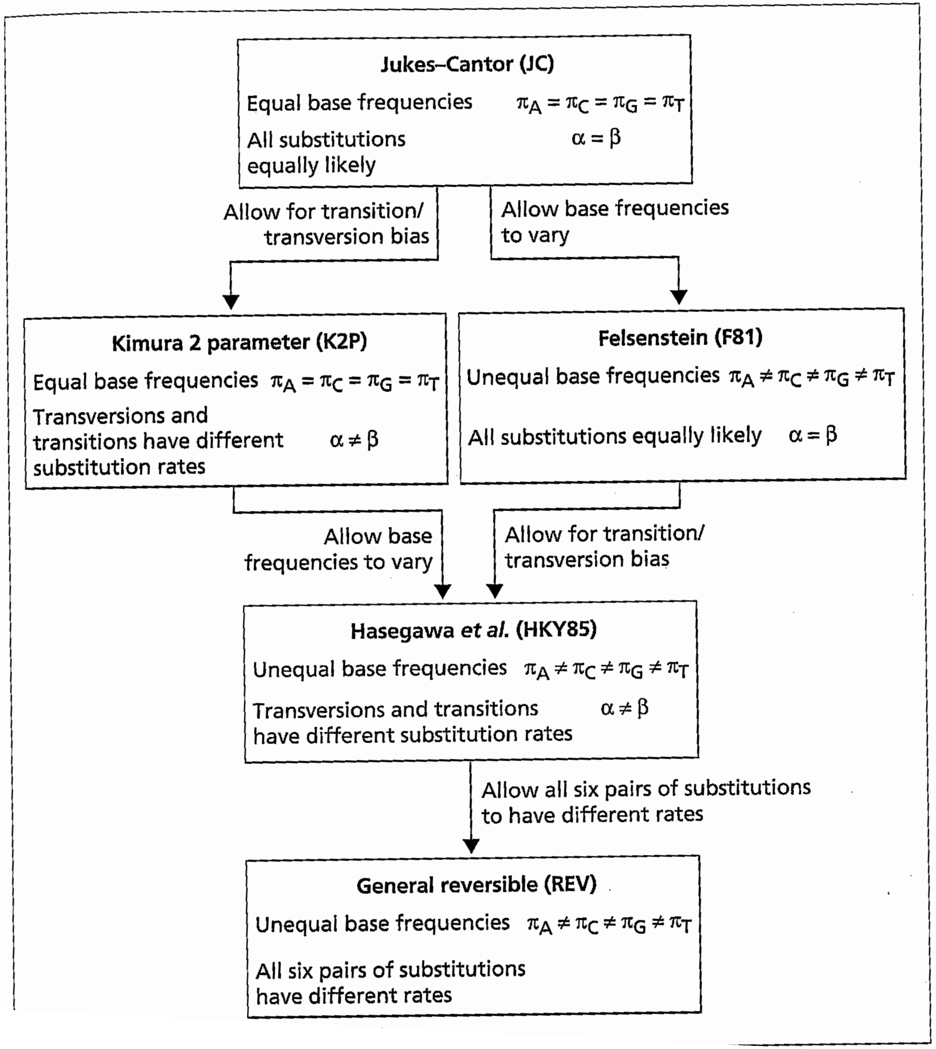


Fig. 5.10 The possible substitutions among the four nucleotides.



Overview of Models



Jukes–Cantor (JC)

Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$

All substitutions equally likely $\alpha = \beta$

Allow for transition/ transversion bias

Allow base frequencies to vary

Kimura 2 parameter (K2P)

Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$

Transversions and transitions have different substitution rates $\alpha \neq \beta$

Felsenstein (F81)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

All substitutions equally likely $\alpha = \beta$

Allow base frequencies to vary

Allow for transition/ transversion bias

Hasegawa *et al.* (HKY85)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

Transversions and transitions have different substitution rates $\alpha \neq \beta$

Allow all six pairs of substitutions to have different rates

General reversible (REV)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

All six pairs of substitutions have different rates

Few parameters
+ Fast
- Less realistic

Parameter-rich
+ more realistic
- slow
- could overfit

Overview of Models

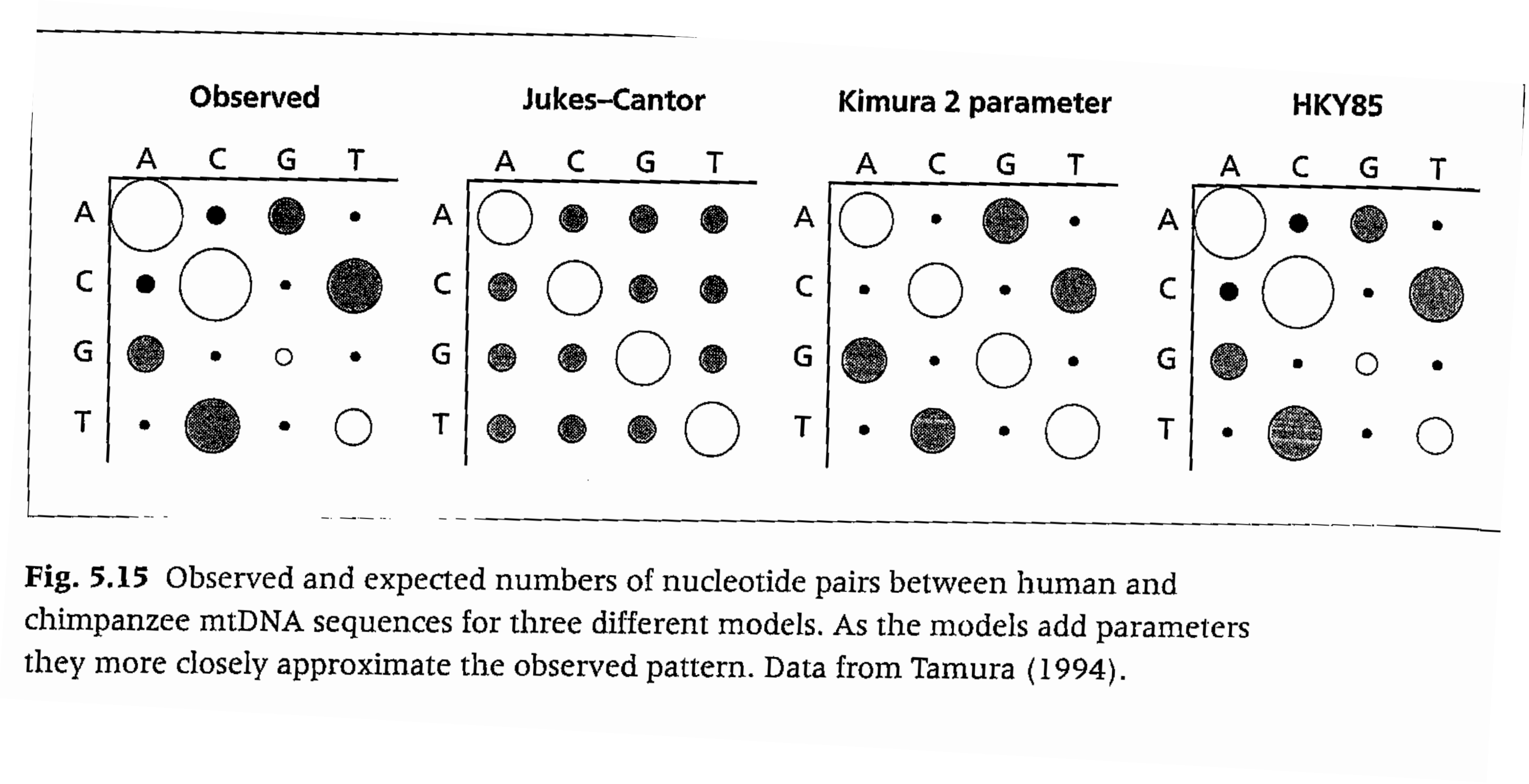


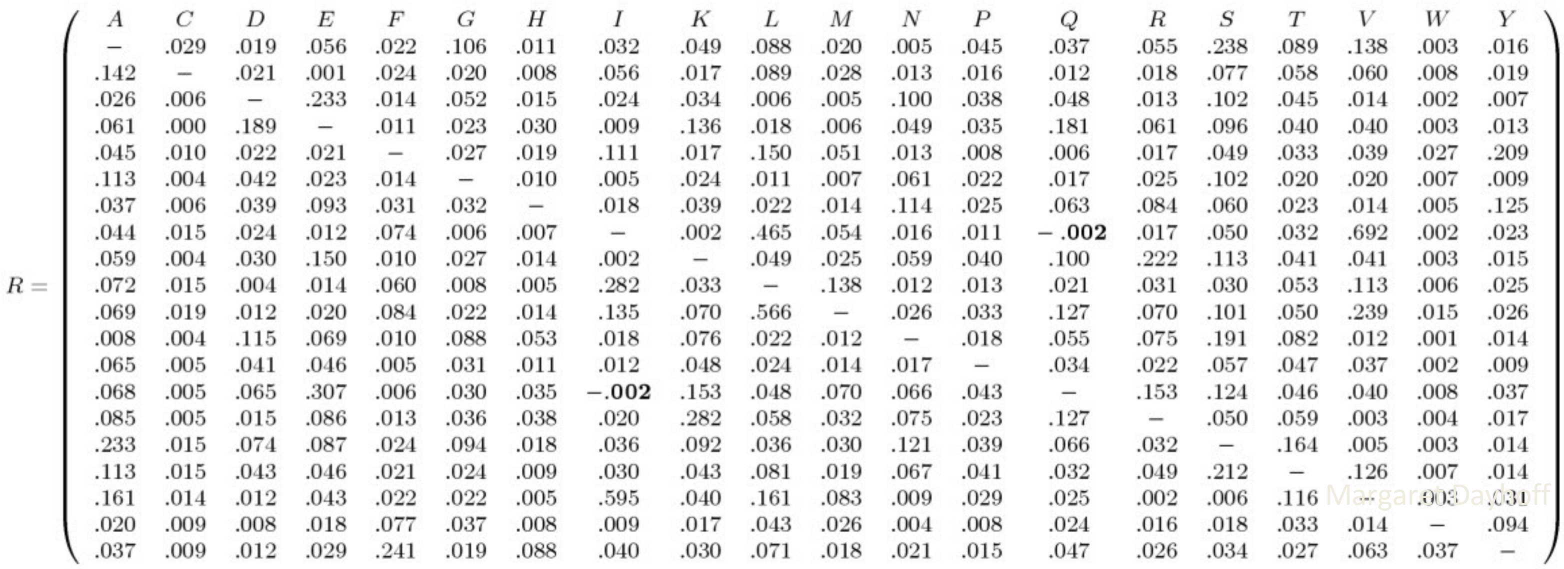
Fig. 5.15 Observed and expected numbers of nucleotide pairs between human and chimpanzee mtDNA sequences for three different models. As the models add parameters they more closely approximate the observed pattern. Data from Tamura (1994).

Protein evolution models

- Protein evolutionary rates range about 3 orders of magnitude.
- Proteins are easier to align because of homologous amino acid positions.
  - This allows study of longer timescales over which nucleotides have been long saturated.
- However, the natural rates of change between which amino acids vary greatly
  - The codon table governs exchanges
  - Physical and chemical differences are important and acted upon by natural selection. (Some are more similar than others.)
    - Leucine<-->Valine != Arginine<-->Lysine

Amino Acid Substitution Models

Rate Matrices model the expected substitution rates for proteins

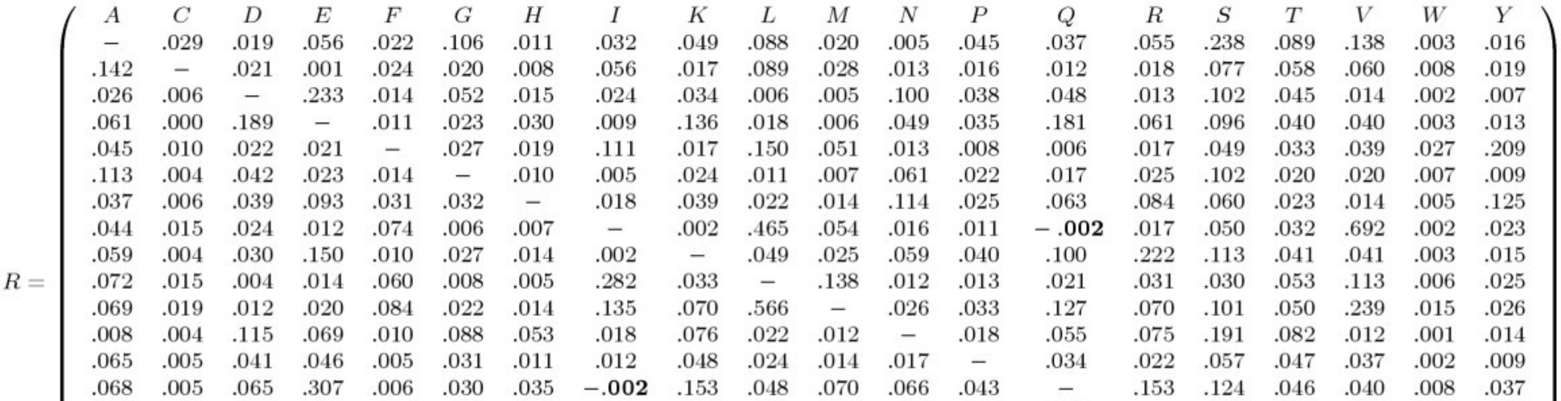


$$R = \left(\begin{array}{cccccccccccccccccccc}
A & C & D & E & F & G & H & I & K & L & M & N & P & Q & R & S & T & V & W & Y \\
- & .029 & .019 & .056 & .022 & .106 & .011 & .032 & .049 & .088 & .020 & .005 & .045 & .037 & .055 & .238 & .089 & .138 & .003 & .016 \\
.142 & - & .021 & .001 & .024 & .020 & .008 & .056 & .017 & .089 & .028 & .013 & .016 & .012 & .018 & .077 & .058 & .060 & .008 & .019 \\
.026 & .006 & - & .233 & .014 & .052 & .015 & .024 & .034 & .006 & .005 & .100 & .038 & .048 & .013 & .102 & .045 & .014 & .002 & .007 \\
.061 & .000 & .189 & - & .011 & .023 & .030 & .009 & .136 & .018 & .006 & .049 & .035 & .181 & .061 & .096 & .040 & .040 & .003 & .013 \\
.045 & .010 & .022 & .021 & - & .027 & .019 & .111 & .017 & .150 & .051 & .013 & .008 & .006 & .017 & .049 & .033 & .039 & .027 & .209 \\
.113 & .004 & .042 & .023 & .014 & - & .010 & .005 & .024 & .011 & .007 & .061 & .022 & .017 & .025 & .102 & .020 & .020 & .007 & .009 \\
.037 & .006 & .039 & .093 & .031 & .032 & - & .018 & .039 & .022 & .014 & .114 & .025 & .063 & .084 & .060 & .023 & .014 & .005 & .125 \\
.044 & .015 & .024 & .012 & .074 & .006 & .007 & - & .002 & .465 & .054 & .016 & .011 & -\mathbf{.002} & .017 & .050 & .032 & .692 & .002 & .023 \\
.059 & .004 & .030 & .150 & .010 & .027 & .014 & .002 & - & .049 & .025 & .059 & .040 & .100 & .222 & .113 & .041 & .041 & .003 & .015 \\
.072 & .015 & .004 & .014 & .060 & .008 & .005 & .282 & .033 & - & .138 & .012 & .013 & .021 & .031 & .030 & .053 & .113 & .006 & .025 \\
.069 & .019 & .012 & .020 & .084 & .022 & .014 & .135 & .070 & .566 & - & .026 & .033 & .127 & .070 & .101 & .050 & .239 & .015 & .026 \\
.008 & .004 & .115 & .069 & .010 & .088 & .053 & .018 & .076 & .022 & .012 & - & .018 & .055 & .075 & .191 & .082 & .012 & .001 & .014 \\
.065 & .005 & .041 & .046 & .005 & .031 & .011 & .012 & .048 & .024 & .014 & .017 & - & .034 & .022 & .057 & .047 & .037 & .002 & .009 \\
.068 & .005 & .065 & .307 & .006 & .030 & .035 & -\mathbf{.002} & .153 & .048 & .070 & .066 & .043 & - & .153 & .124 & .046 & .040 & .008 & .037 \\
.085 & .005 & .015 & .086 & .013 & .036 & .038 & .020 & .282 & .058 & .032 & .075 & .023 & .127 & - & .050 & .059 & .003 & .004 & .017 \\
.233 & .015 & .074 & .087 & .024 & .094 & .018 & .036 & .092 & .036 & .030 & .121 & .039 & .066 & .032 & - & .164 & .005 & .003 & .014 \\
.113 & .015 & .043 & .046 & .021 & .024 & .009 & .030 & .043 & .081 & .019 & .067 & .041 & .032 & .049 & .212 & - & .126 & .007 & .014 \\
.161 & .014 & .012 & .043 & .022 & .022 & .005 & .595 & .040 & .161 & .083 & .009 & .029 & .025 & .002 & .006 & .116 & - & .003 & .031 \\
.020 & .009 & .008 & .018 & .077 & .037 & .008 & .009 & .017 & .043 & .026 & .004 & .008 & .024 & .016 & .018 & .033 & .014 & - & .094 \\
.037 & .009 & .012 & .029 & .241 & .019 & .088 & .040 & .030 & .071 & .018 & .021 & .015 & .047 & .026 & .034 & .027 & .063 & .037 & -
\end{array}\right)$$

Margaret Dayhoff

Amino Acid Substitution Models

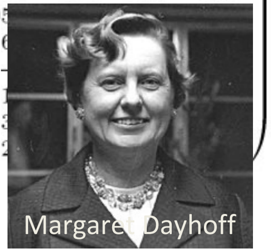
Rate Matrices model the expected substitution rates for proteins



$$R = \left(\begin{array}{cccccccccccccccccccc}
A & C & D & E & F & G & H & I & K & L & M & N & P & Q & R & S & T & V & W & Y \\
- & .029 & .019 & .056 & .022 & .106 & .011 & .032 & .049 & .088 & .020 & .005 & .045 & .037 & .055 & .238 & .089 & .138 & .003 & .016 \\
.142 & - & .021 & .001 & .024 & .020 & .008 & .056 & .017 & .089 & .028 & .013 & .016 & .012 & .018 & .077 & .058 & .060 & .008 & .019 \\
.026 & .006 & - & .233 & .014 & .052 & .015 & .024 & .034 & .006 & .005 & .100 & .038 & .048 & .013 & .102 & .045 & .014 & .002 & .007 \\
.061 & .000 & .189 & - & .011 & .023 & .030 & .009 & .136 & .018 & .006 & .049 & .035 & .181 & .061 & .096 & .040 & .040 & .003 & .013 \\
.045 & .010 & .022 & .021 & - & .027 & .019 & .111 & .017 & .150 & .051 & .013 & .008 & .006 & .017 & .049 & .033 & .039 & .027 & .209 \\
.113 & .004 & .042 & .023 & .014 & - & .010 & .005 & .024 & .011 & .007 & .061 & .022 & .017 & .025 & .102 & .020 & .020 & .007 & .009 \\
.037 & .006 & .039 & .093 & .031 & .032 & - & .018 & .039 & .022 & .014 & .114 & .025 & .063 & .084 & .060 & .023 & .014 & .005 & .125 \\
.044 & .015 & .024 & .012 & .074 & .006 & .007 & - & .002 & .465 & .054 & .016 & .011 & -\mathbf{.002} & .017 & .050 & .032 & .692 & .002 & .023 \\
.059 & .004 & .030 & .150 & .010 & .027 & .014 & .002 & - & .049 & .025 & .059 & .040 & .100 & .222 & .113 & .041 & .041 & .003 & .015 \\
.072 & .015 & .004 & .014 & .060 & .008 & .005 & .282 & .033 & - & .138 & .012 & .013 & .021 & .031 & .030 & .053 & .113 & .006 & .025 \\
.069 & .019 & .012 & .020 & .084 & .022 & .014 & .135 & .070 & .566 & - & .026 & .033 & .127 & .070 & .101 & .050 & .239 & .015 & .026 \\
.008 & .004 & .115 & .069 & .010 & .088 & .053 & .018 & .076 & .022 & .012 & - & .018 & .055 & .075 & .191 & .082 & .012 & .001 & .014 \\
.065 & .005 & .041 & .046 & .005 & .031 & .011 & .012 & .048 & .024 & .014 & .017 & - & .034 & .022 & .057 & .047 & .037 & .002 & .009 \\
.068 & .005 & .065 & .307 & .006 & .030 & .035 & -\mathbf{.002} & .153 & .048 & .070 & .066 & .043 & - & .153 & .124 & .046 & .040 & .008 & .037 \\
.085 & .005 & .015 & .086 & .013 & .036 & .038 & .020 & .282 & .058 & .032 & .075 & .023 & .127 & - & .050 & & & &
\end{array}\right)$$

PAM – Accepted Point Mutation

Schwarz R & Dayhoff M (1979) Matrices for detecting distant relationships. In Dayhoff M, editor, Atlas of protein sequences, pages 353 - 58. National Biomedical Research Foundation.



Margaret Dayhoff

BLOSUM

Henikoff and Henikoff (1992) Amino acid substitution matrices from protein blocks. PNAS 89: 10915-19.

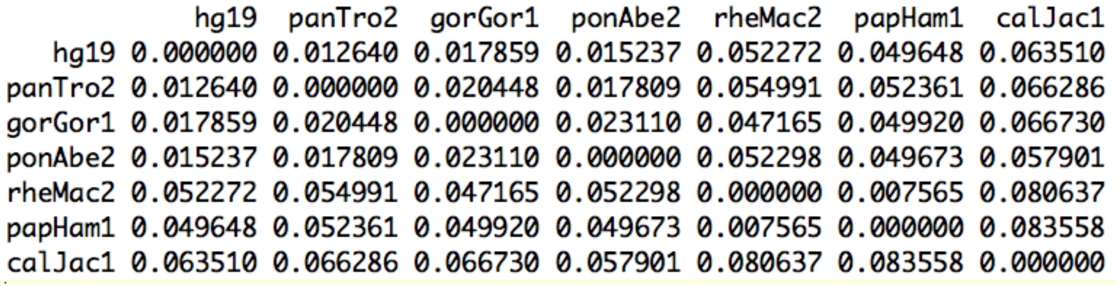
Jorja and Steve Henikoff

Inference of Phylogenies

Phylogeny is inferred through distance data or discrete characters

Distance – all pairwise distances are computed from characters.

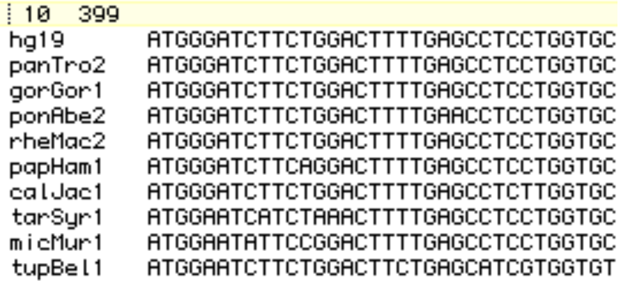
- Neighbor-Joining is widely used distance algorithm
- FAST-ME 2 is newer and faster



```
            hg19   panTro2  gorGor1  ponAbe2  rheMac2  papHam1  calJac1
   hg19 0.000000 0.012640 0.017859 0.015237 0.052272 0.049648 0.063510
panTro2 0.012640 0.000000 0.020448 0.017809 0.054991 0.052361 0.066286
gorGor1 0.017859 0.020448 0.000000 0.023110 0.047165 0.049920 0.066730
ponAbe2 0.015237 0.017809 0.023110 0.000000 0.052298 0.049673 0.057901
rheMac2 0.052272 0.054991 0.047165 0.052298 0.000000 0.007565 0.080637
papHam1 0.049648 0.052361 0.049920 0.049673 0.007565 0.000000 0.083558
calJac1 0.063510 0.066286 0.066730 0.057901 0.080637 0.083558 0.000000
```

Discrete character methods use sequences directly during inference.

- Maximum parsimony, **Maximum likelihood,** and **Bayesian methods**



```
10  399
hg19        ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
panTro2     ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
gorGor1     ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
ponAbe2     ATGGGATCTTCTGGACTTTTGAACCTCCTGGTGC
rheMac2     ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
papHam1     ATGGGATCTTCAGGACTTTTGAGCCTCCTGGTGC
calJac1     ATGGGATCTTCTGGACTTTTGAGCCTCTTGGTGC
tarSyr1     ATGGAATCATCTAAACTTTTGAGCCTCCTGGTGC
micMur1     ATGGAATATTCCGGACTTTTGAGCCTCCTGGTGC
tupBel1     ATGGAATCTTCTGGACTTCTGAGCATCGTGGTGT
```

Distance Method: Neighbor-Joining

Genetic distances are first calculated between all pairs of sequences in the alignment.

Distance methods are **very fast** and often a good choice, especially when there are a large number of sequences (> ~100).

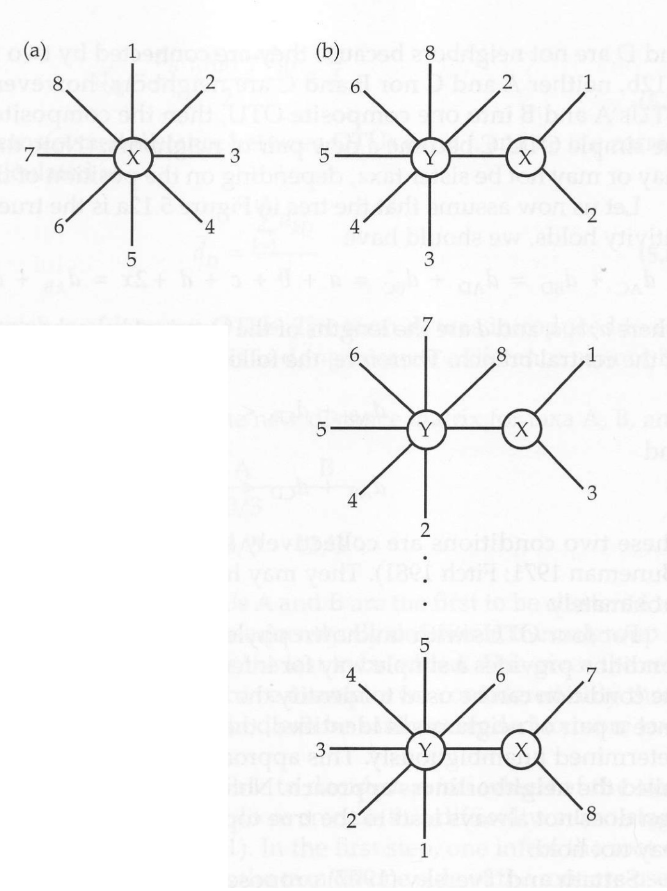
Neighbor-Joining: Saitou and **Nei**. (1987) *MBE*

Clustering method that quickly generates a single tree reproducibly.

Begin with all taxa in a star-tree – one node polytomy.

$$S_0 = 1/(N-1) * \Sigma_{(i<j)}D_{ij}$$

Sequentially add nodes to pairs of taxa that minimize the total tree length.



Statistical models in phylogenetics

Likelihood and Bayesian methods discern between evolutionary models using probability.

- Probability provides a criterion to select the best tree topology and parameter estimates.
- Can test hypotheses using parameters such as rate of evolution (K, d_N/d_S, divergence times…)

Evolutionary models include:

- Substitution-related parameters
  - Branch-specific substitution rates
  - Transition/transversion rate ratio
  - Equilibrium frequencies for nucleotides/amino acids
  - Rate classes, which allow sites to have different rates
- Tree topology
  - The pattern of connections between sequences
  - Topology is a strange parameter space, and is harder to search than a continuous numerical parameter such as a rate or the transition/transversion ratio.

These methods are the **most accurate**, but are computationally expensive and **slower**.

Maximizing likelihood numerically

Complicated models must be maximized through a guided trial-and-error, “hill climbing” algorithm.

1. Set initial parameter values and tree.
2. Calculate likelihood.
3. Propose new parameter value or tree.
4. Calculate likelihood.
5. Decide whether to accept the new value.
6. Repeat steps 3-5 until changes no longer improve likelihood.

Because local maxima can trap this algorithm below the best model, must try multiple initial parameter values.

This is computationally expensive but allows us to determine the maximum likelihood model in many cases.

If the algorithm converges on the same parameter estimates after starting from several different initial values, it is a sign that the algorithm is converging on the maximum likelihood model.

Examples of likelihoods for different ape tree topologies.

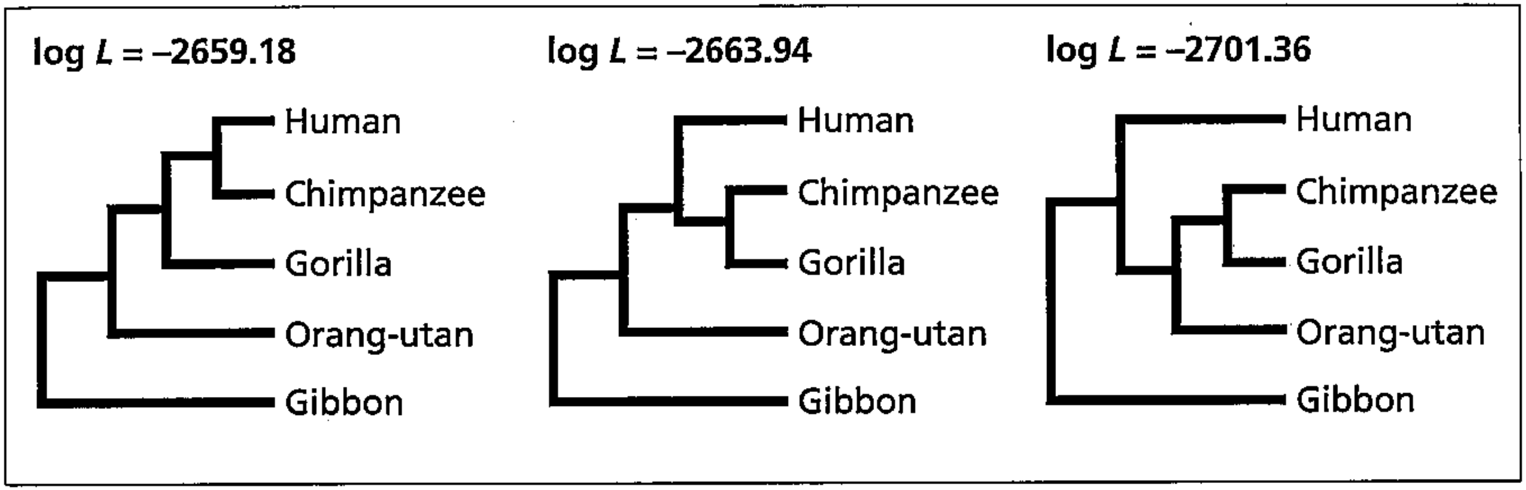


Fig. 6.19 Three different hypotheses of relationship among the hominoids and the likelihoods that each tree has given rise to the observed data.

- Do these represent all possible topologies?
- Which is most likely?

Phylogenetic software

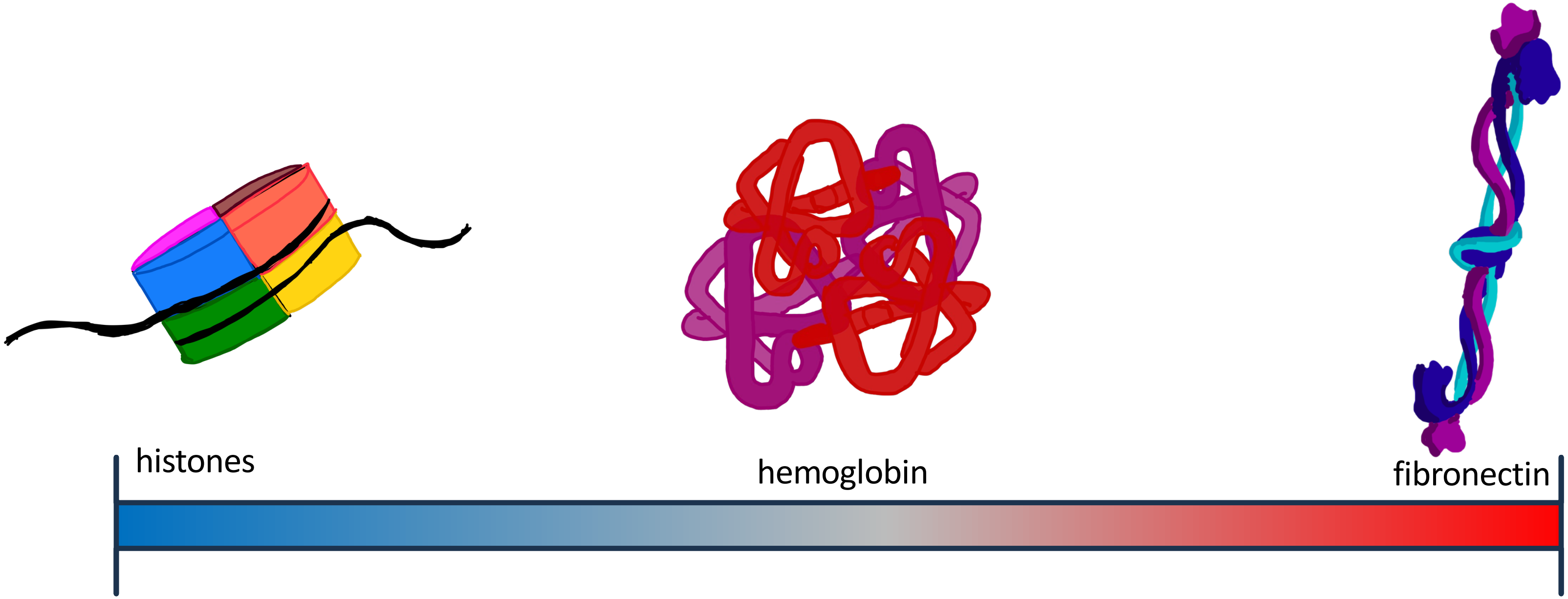
- Distance
  - Command-line: FastME 2.0 (Fast Minimum Evolution)
    - http://www.atgc-montpellier.fr/fastme/
  - Graphic user interface: Seaview https://doua.prabi.fr/software/seaview
- Maximum Likelihood
  - PhyML http://www.atgc-montpellier.fr/phyml/
  - RAxML-NG https://github.com/amkozlov/raxml-ng
  - IQ-TREE 2 https://github.com/iqtree/iqtree2
  - Phangorn package in R (nice in R but slower) https://klausvigo.github.io/phangorn/
- Bayesian
  - BEAST 2 / StarBeast3 https://www.beast2.org/
  - RevBayes https://revbayes.github.io/

Forms of Natural Selection

- Positive selection
  - Beneficial alleles increase in frequency
  - Over phylogenetic timescales, leads to **increased divergence** of nucleotide and/or protein sequences.
  - "Diversifying selection"

- Negative selection
  - Deleterious alleles decrease in frequency
  - On phylogenetic scales, leads to **conservation** of sequences and overall slower divergence.
  - "Purifying selection"

Genes/proteins evolve at different average rates



histones

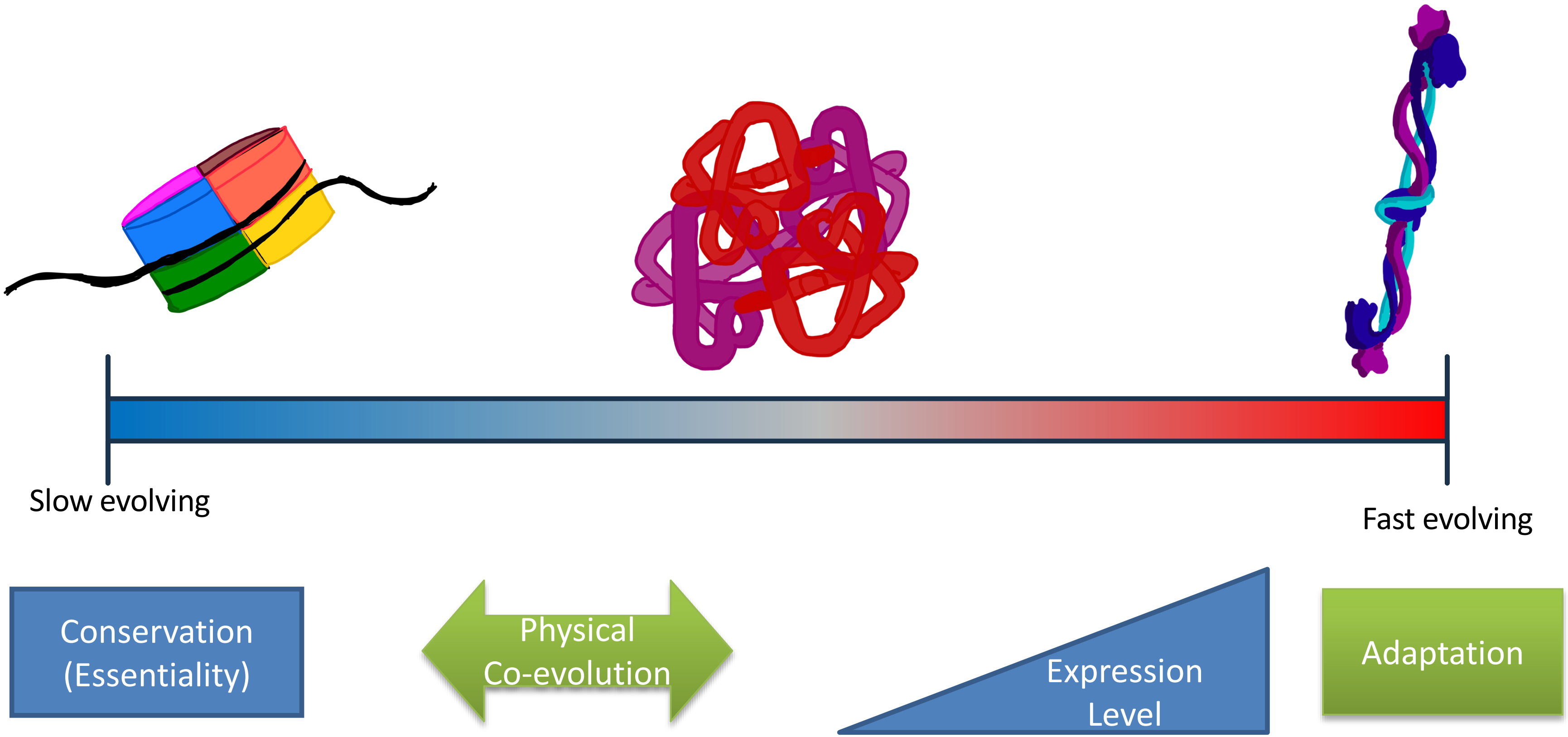
hemoglobin

fibronectin

Slow evolving

Fast evolving

Which forces control evolutionary rate?



Slow evolving

Fast evolving

Conservation (Essentiality)

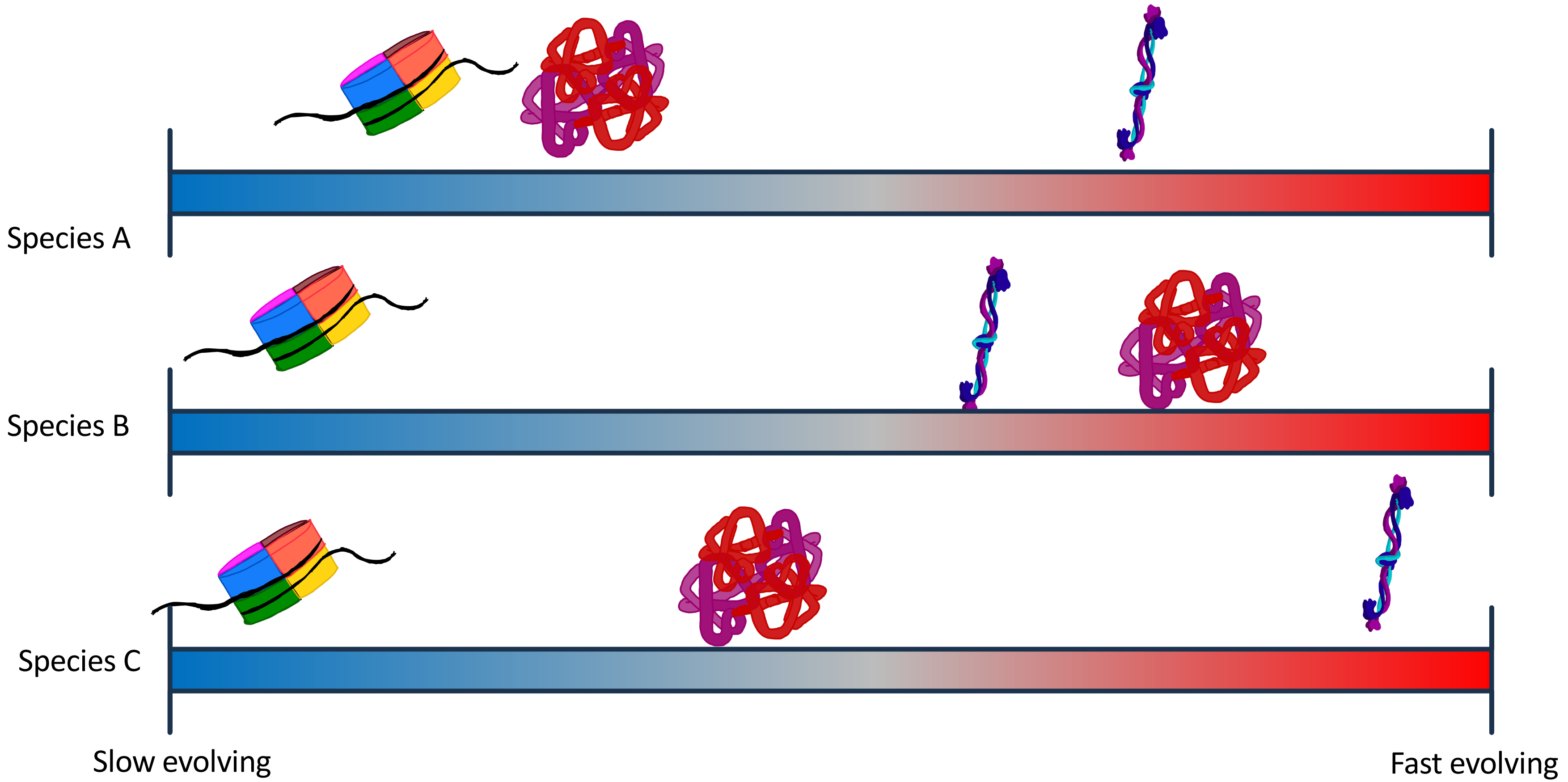
Physical Co-evolution

Expression Level

Adaptation

- Rates vary over time due to fluctuation in these forces

And… the evolutionary rate of any specific gene/protein changes over time (between species)



Comparative Approaches to Map Adaptive Phenotypes to Genotypes

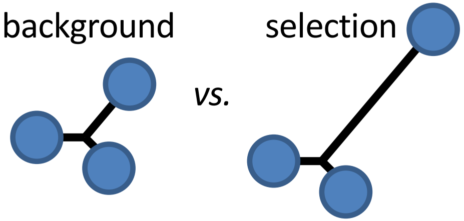
Genetic crosses between populations or species.

- Genetic mapping
- Peromyscus, blind cavefish

Linnen et al *Science* 2009

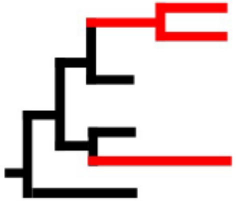
Positively selected regions leading to trait

- Population Genetics (F_{ST})
- Phylogenetics (d_N/d_S)
- High altitude populations and HIF1a



PhyloG2P – phylogenetic genotype to phenotype

Coincident evolution of genetic change and a **convergent** trait (i.e., repeatedly evolved trait)



Rate acceleration correlated with convergent trait

PhyloG2P

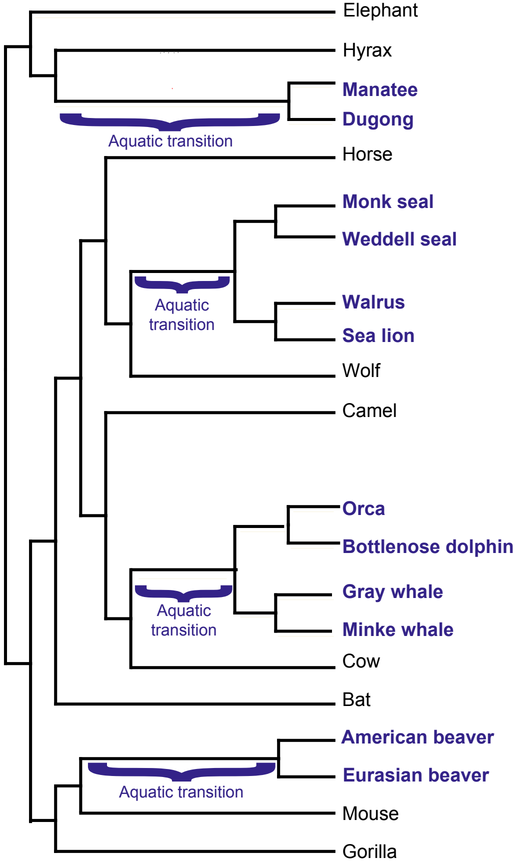
Phylogenetic Genotype to Phenotype

Smith et al. *TREE* 2020

Which genetic events are associated with a trait?

Genetic events include:

- Rate changes
- Positive selection
- Gene loss/ pseudogenization
- Gene family expansion/contraction



PhyloG2P

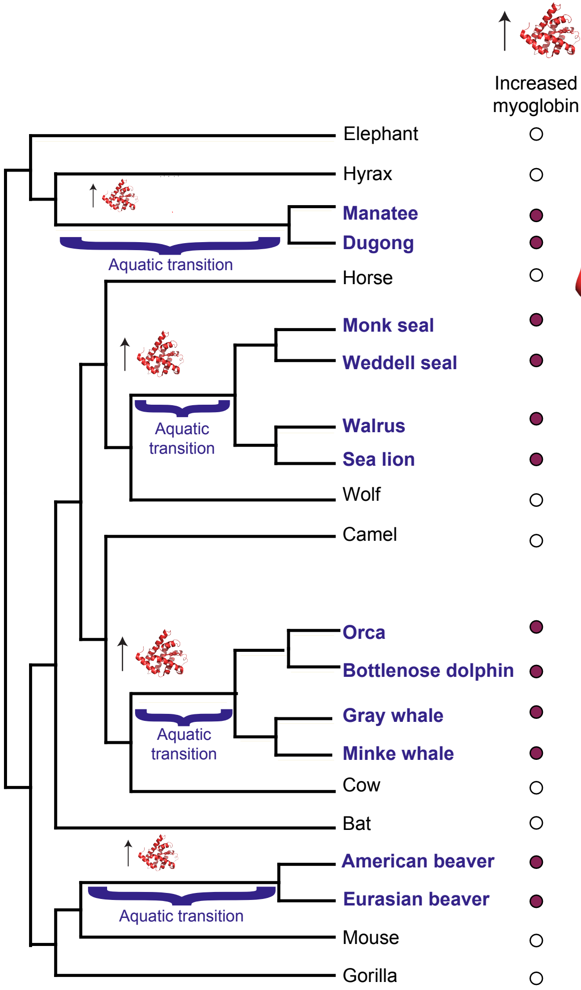
Phylogenetic Genotype to Phenotype

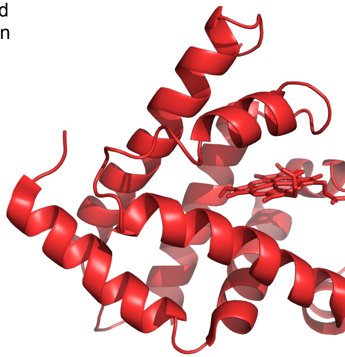
Which genetic events are associated with a trait?

Genetic events include:

- Rate changes
- Positive selection
- Gene loss/ pseudogenization
- Gene family expansion/contraction

Novel traits produce **transient genetic signals**.





Increased Myoglobin

PhyloG2P

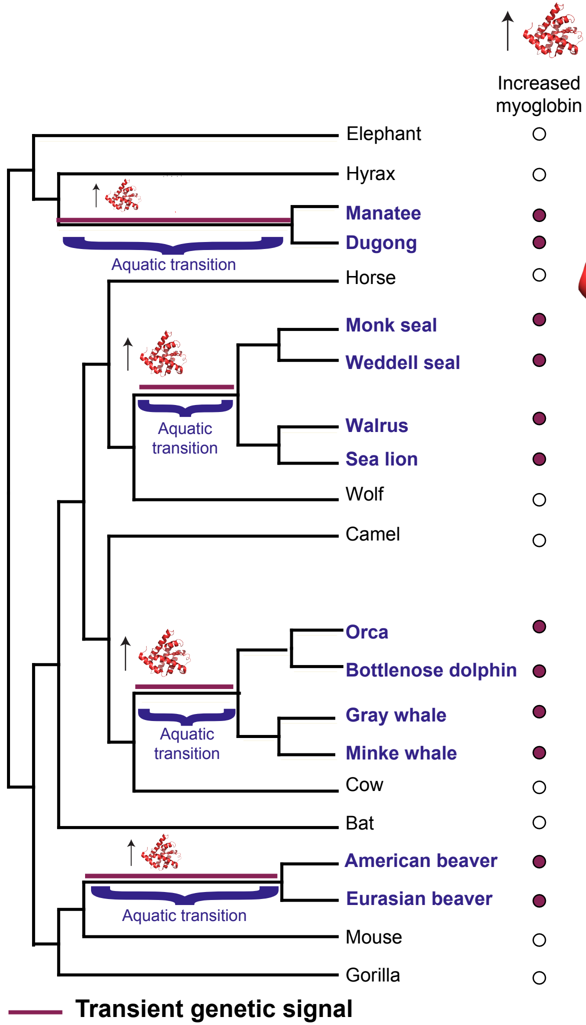
Phylogenetic Genotype to Phenotype

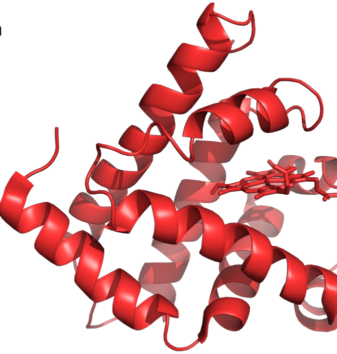
Which genetic events are associated with a trait?

Genetic events include:

- Rate changes
- Positive selection
- Gene loss/ pseudogenization
- Gene family expansion/contraction

Novel traits produce **transient genetic signals**.





Increased Myoglobin

PhyloG2P

Phylogenetic Genotype to Phenotype

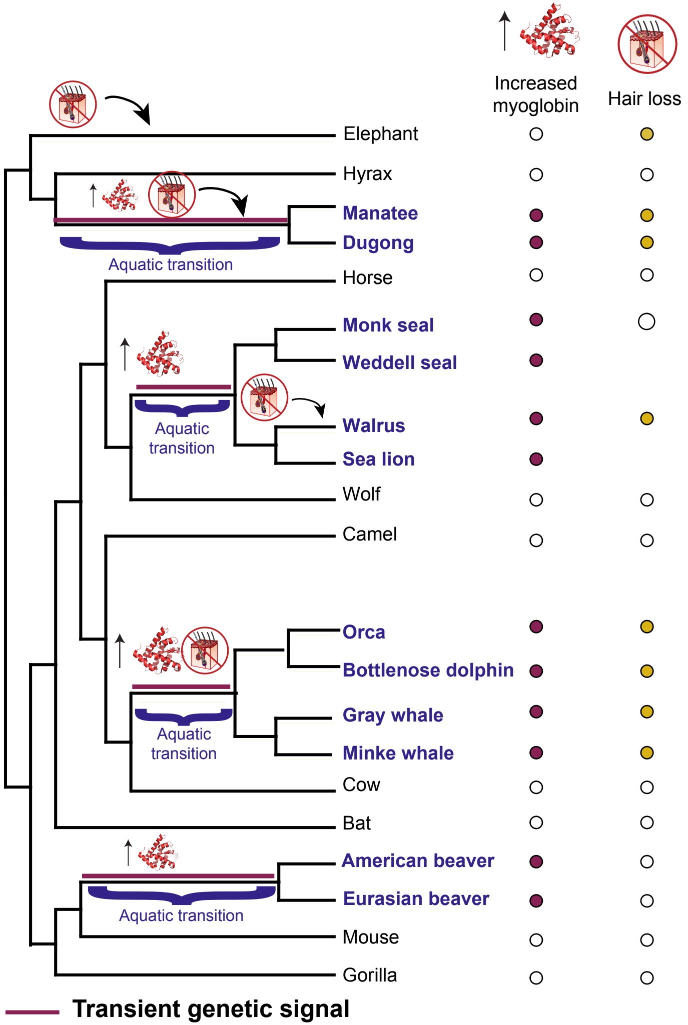
Which genetic events are associated with a trait?

Genetic events include:

- Rate changes
- Positive selection
- Gene loss/ pseudogenization
- Gene family expansion/contraction

Novel traits produce **transient genetic signals**.

Trait loss leads to **persistent genetic signals**.



Transient genetic signal

Indicates putative branches on which pictured phenotypes first appeared

PhyloG2P

Phylogenetic Genotype to Phenotype

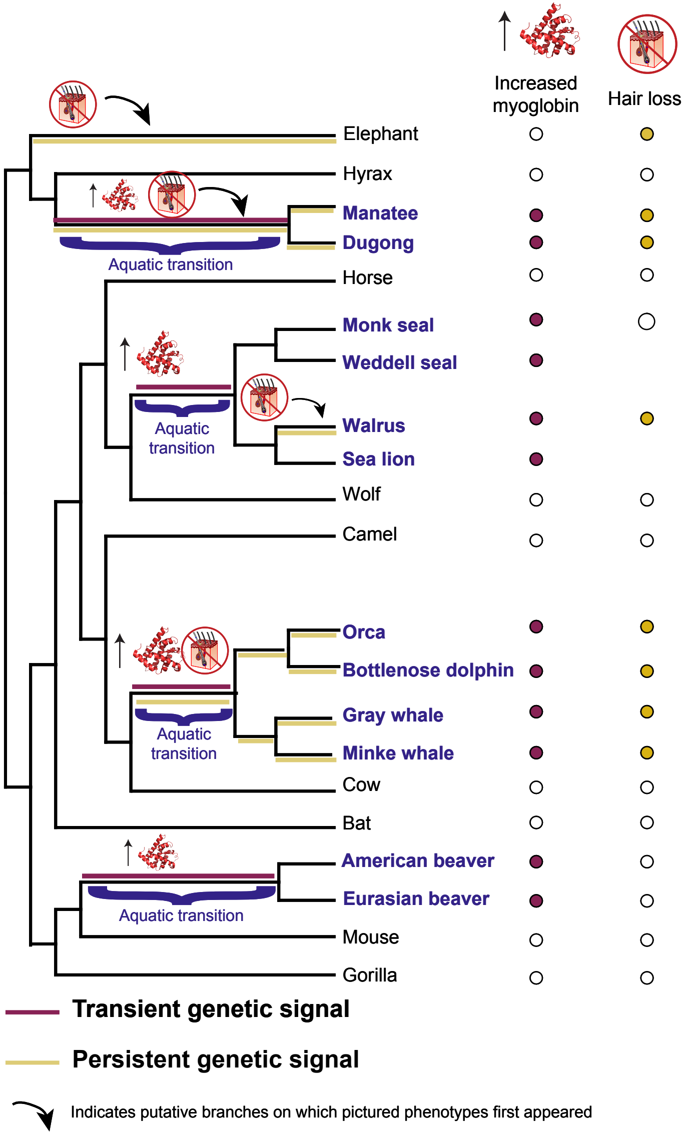
Which genetic events are associated with a trait?

Genetic events include:

- Rate changes
- Positive selection
- Gene loss/ pseudogenization
- Gene family expansion/contraction

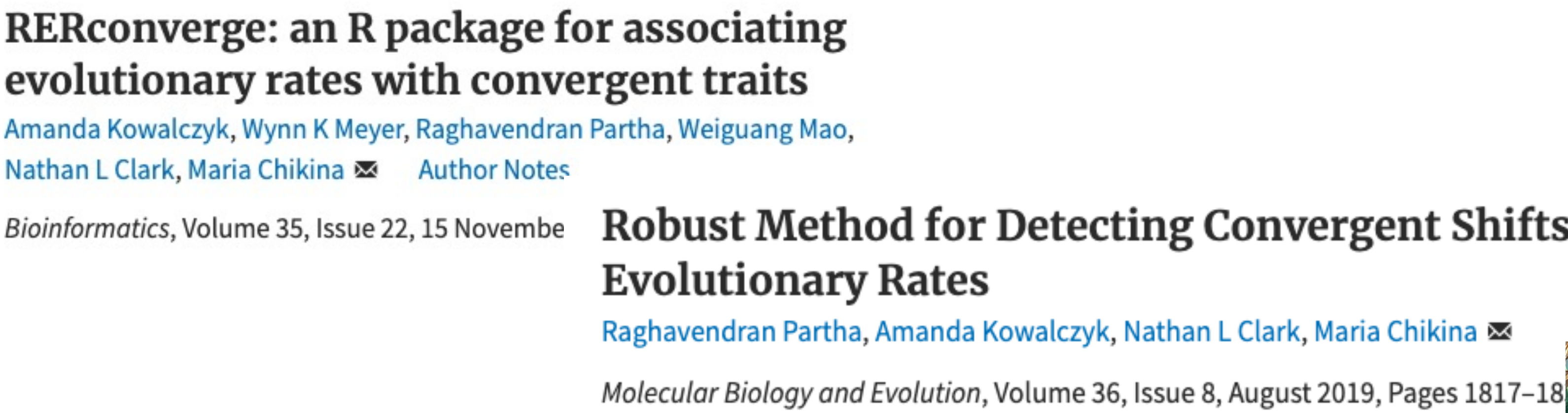
Novel traits produce **transient genetic signals**.

Trait loss leads to **persistent genetic signals**.



RERconverge package, available on GitHub

https://github.com/nclark-lab/RERconverge



RERconverge: an R package for associating evolutionary rates with convergent traits

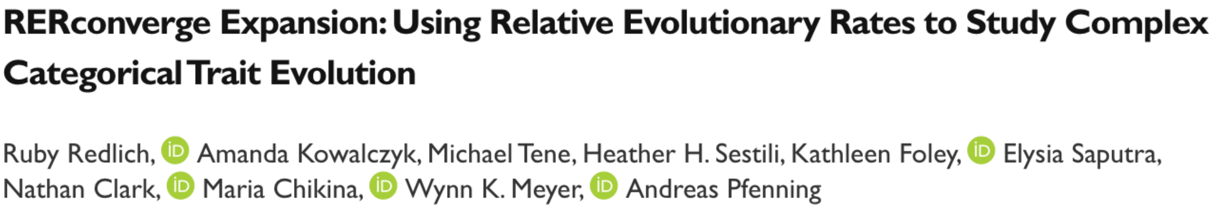
Amanda Kowalczyk, Wynn K Meyer, Raghavendran Partha, Weiguang Mao, Nathan L Clark, Maria Chikina ✉ Author Notes

Bioinformatics, Volume 35, Issue 22, 15 Novembe

Robust Method for Detecting Convergent Shifts in Evolutionary Rates

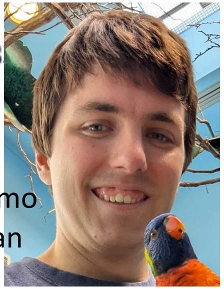
Raghavendran Partha, Amanda Kowalczyk, Nathan L Clark, Maria Chikina ✉

Molecular Biology and Evolution, Volume 36, Issue 8, August 2019, Pages 1817–18



RERconverge Expansion: Using Relative Evolutionary Rates to Study Complex Categorical Trait Evolution

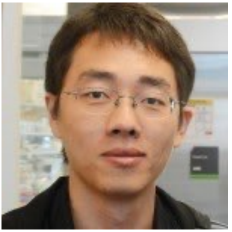
Ruby Redlich, Amanda Kowalczyk, Michael Tene, Heather H. Sestili, Kathleen Foley, Elysia Saputra, Nathan Clark, Maria Chikina, Wynn K. Meyer, Andreas Pfenning



Guillermo Hoffman Meyer

Maria Chikina

Amanda Kowalczyk



Wayne Mao

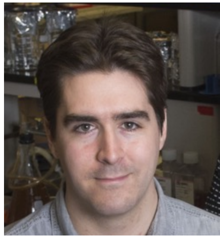
Wynn Meyer

Raghav Partha

Joe Robinson

Elysia Saputra

Ruby Redlich



Andreas Pfenning

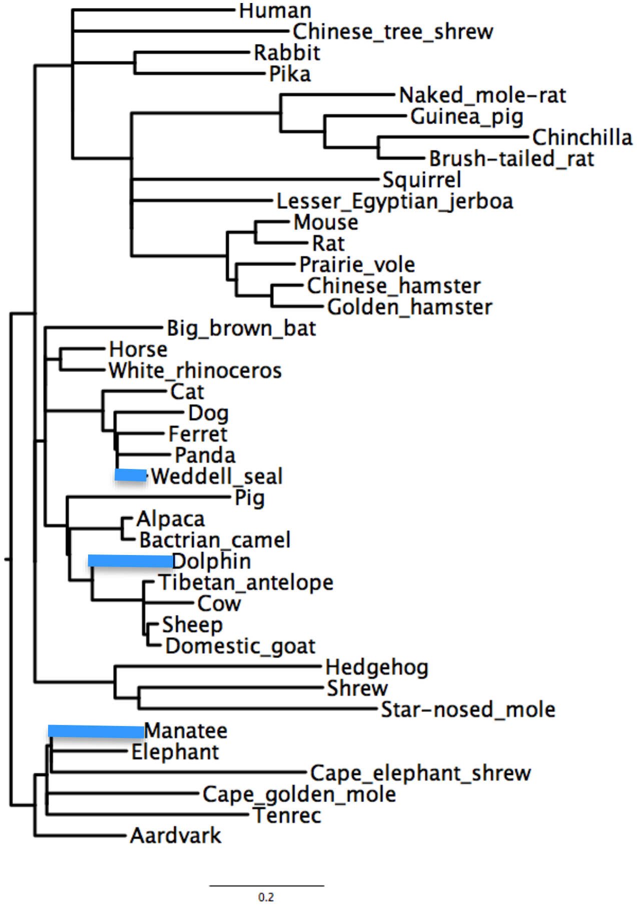
RERCONVERGE is widely used in comparative genomics

The perplexing figure behind a crucial virus database p. 332
Regulatory reforms to advance psychedelic therapies p. 347
A compact galaxy in the early Universe p. 416
Science
$15
28 APRIL 2023
SPECIAL ISSUE
science.org
AAAS
CRYPTOPROCTA FEROX
ZOONOMIA
Diverse genomes reveal mammalian secrets p. 356
DAUBENTONIA MADAGASCARIENSIS
CHOLOEPUS HOFFMANNI
PHATAGINUS TRICUSPIS
ECHINOPS TELFAIRI

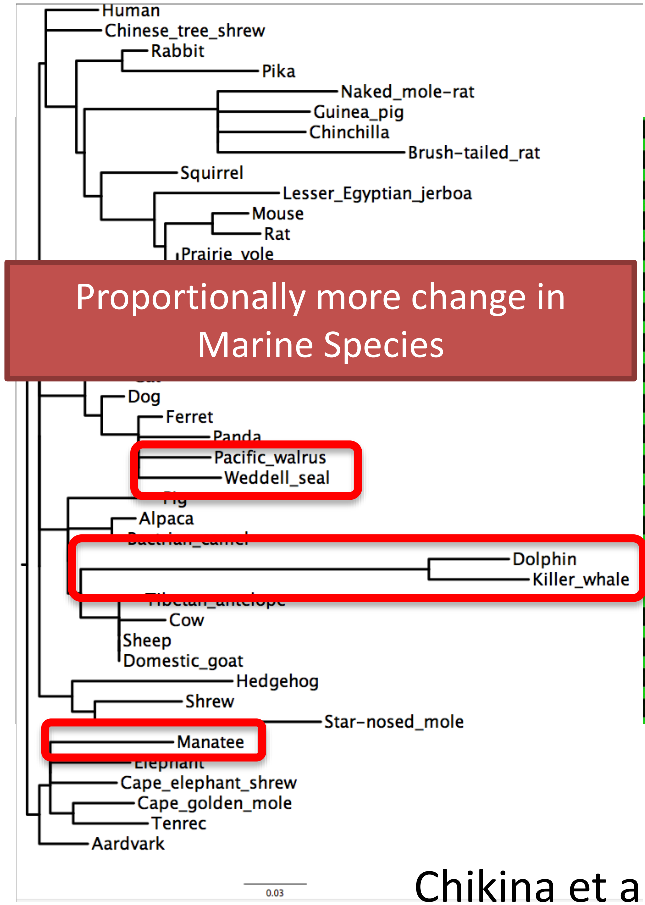
The need for research on solar geoengineering pp. 795, 812, & 815
How the Webb space telescope will revolutionize astronomy p. 806
Chang'e-5 samples show late lunar volcanism p. 887
Science
$15
12 NOVEMBER 2021
science.org
AAAS
EXTREME LIFE SPAN
The genetics of aging in rockfishes pp. 824 & 842

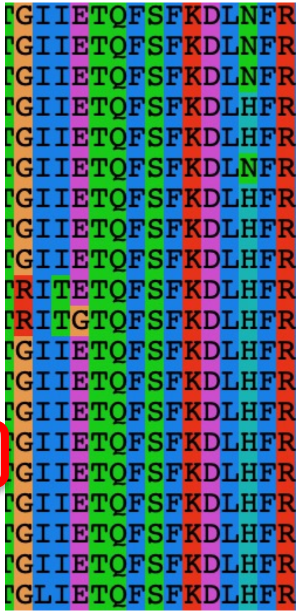
Gene evolutionary rates vary across species

Average Evolutionary Rate



Gustatory G-protein (GNAT3)



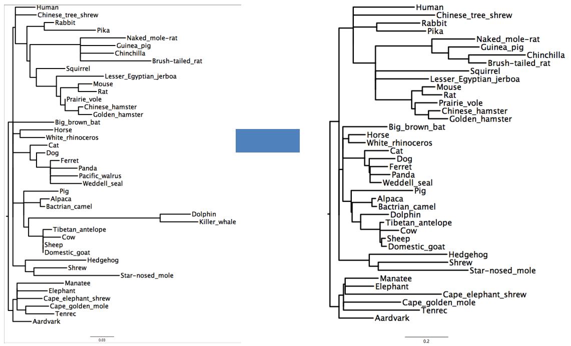


Proportionally more change in Marine Species

Chikina et al. *MBE* 2016

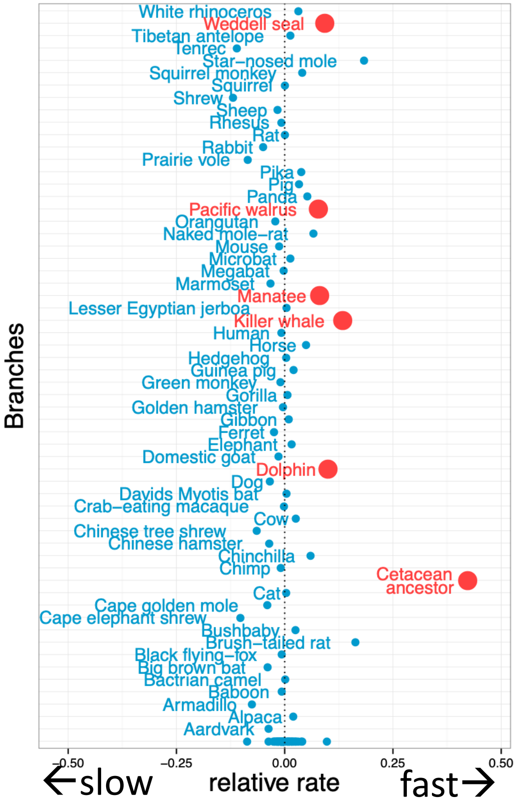
Relative evolutionary rates (RER)
a method to study convergence

“Subtract” **average tree** from **gene tree**



Actually, we take residuals from a weighted regression

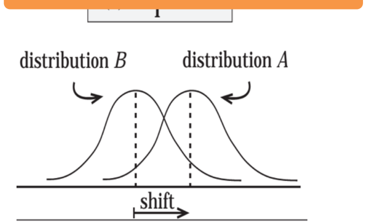
Relative Evolutionary Rates (RERs) of *GNAT3*



Hypothesis test:
RER acceleration in marine *vs* terrestrial species

P = 0.00008

hence convergent marine acceleration

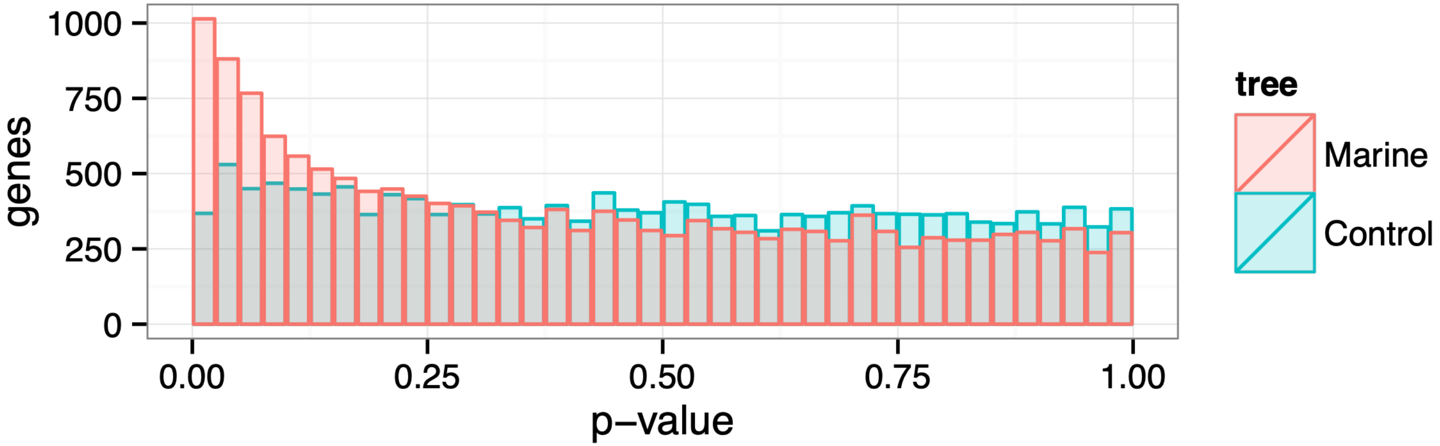


Chikina et al. *MBE* 2016

Marine-accelerated genes are in statistical excess

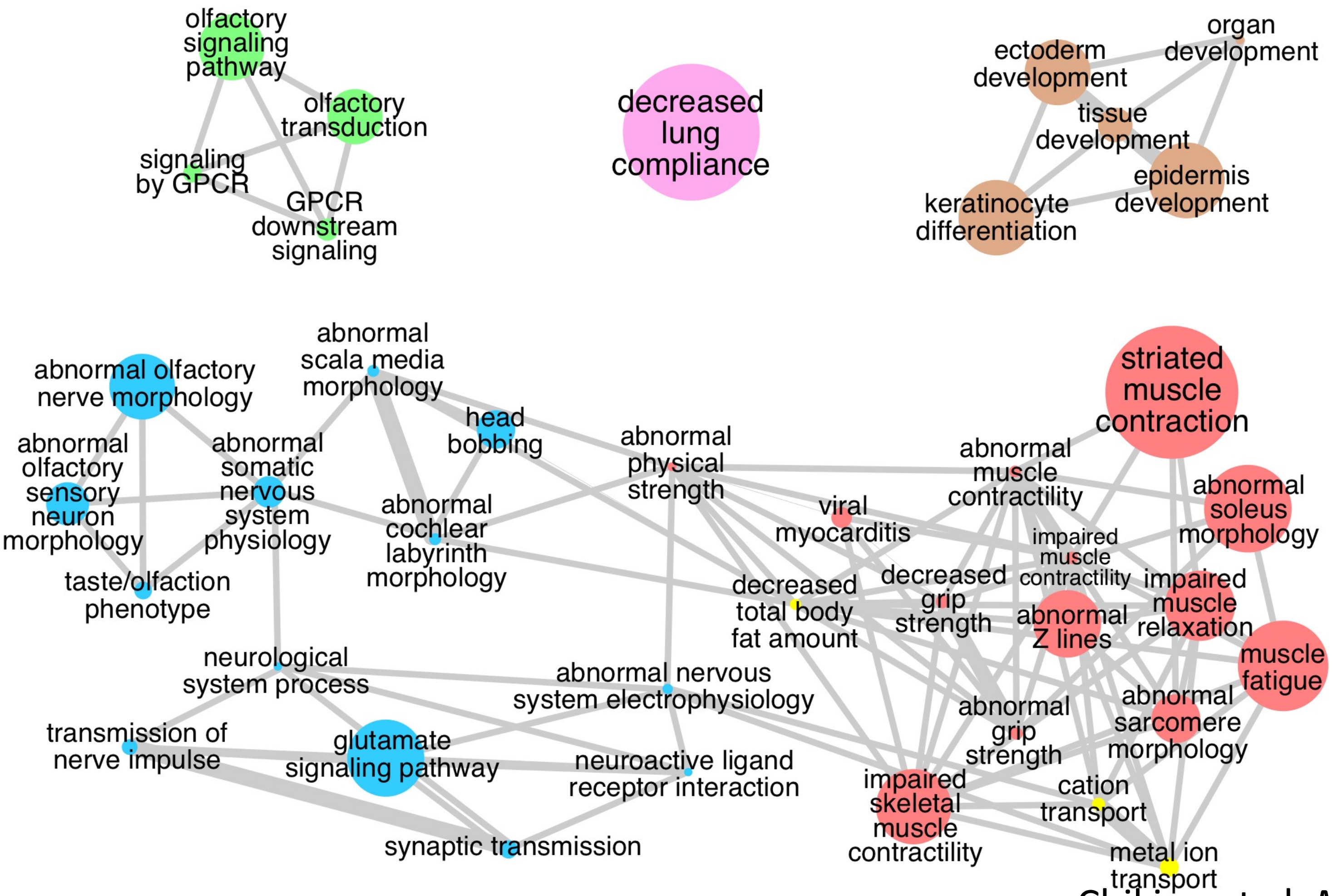
Table 1 | Top marine-accelerated genes

Gene	P-value	Description	Function	Evolutionary Mode
GNAT3	0.00008	G protein subunit in bitter, sweet, and umami taste transduction	Taste	Relaxed
OR6B1	0.00014	olfactory receptor	Olfaction	Relaxed
CALHM1	0.00014	ion channel required for sweet, bitter and umami tastes	Taste	Relaxed
SSTR4	0.00016	somatostatin receptor 4		Relaxed
COL9A2	0.00023	collagen, type IX, alpha 2; hyaline joint cartilage protein	Cartilage	Pos. selection
FGF11	0.00033	fibroblast growth factor. Nervous system development		Relaxed
HMGCS2	0.00042	catalyzes ketogenesis, which provides lipid-derived energy	Lipid metabolism	Relaxed
CLSTN2	0.00046	calsyntenin 2	Nervous system	Relaxed
PERP	0.00049	component of intercellular desmosome junctions in epithelia	Skin	Pos. selection
S100A5	0.00059	S100 calcium binding protein A5		Relaxed
PIK3R5	0.00153	phosphoinositide-3-kinase, regulatory subunit 5		Relaxed
RNF222	0.00154	ring finger protein 222		Relaxed
OR10Z1	0.00164	olfactory receptor, family 10, subfamily Z, member 1	Olfaction	Relaxed
TGM3	0.00166	transglutaminase 3, cell envelope formation in the epidermis and hair follicle	Skin, Hair	Pos. selection
SFTPB	0.00168	Lung-specific surfactant protein B	Lung	Relaxed



Chikina et al. *MBE* 2016

Functions enriched among marine-accelerated genes



olfactory signaling pathway
olfactory transduction
signaling by GPCR
GPCR downstream signaling
decreased lung compliance
ectoderm development
organ development
tissue development
epidermis development
keratinocyte differentiation
abnormal olfactory nerve morphology
abnormal scala media morphology
head bobbing
abnormal olfactory sensory neuron morphology
abnormal somatic nervous system physiology
abnormal cochlear labyrinth morphology
taste/olfaction phenotype
abnormal physical strength
viral myocarditis
abnormal muscle contractility
striated muscle contraction
abnormal soleus morphology
impaired muscle contractility
decreased total body fat amount
decreased grip strength
abnormal Z lines
impaired muscle relaxation
muscle fatigue
neurological system process
abnormal nervous system electrophysiology
transmission of nerve impulse
glutamate signaling pathway
neuroactive ligand receptor interaction
synaptic transmission
abnormal grip strength
abnormal sarcomere morphology
impaired skeletal muscle contractility
cation transport
metal ion transport

Chikina et al. *MBE* 2016

Lab Structure

Convergent evolution provides a powerful natural experiment for identifying the genetic basis of complex traits.

- Goal: Use a PhyloG2P method to study a convergent trait
- Learning Objectives
  - Understand and view multiple sequence alignments
  - Infer trees
  - Use a relative rates to study a convergent trait
  - Find functional enrichments among gene sets

Lab Structure

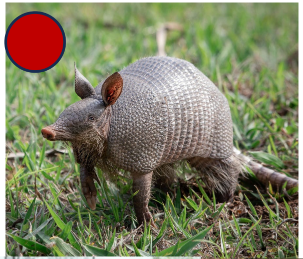
Activity: Use RERconverge to find genes whose rates associate with the loss of teeth and/or tooth enamel.

Steps

1. Make gene trees for sample genes in 108 mammal species
2. Read in large number of orthologous gene trees
3. Calculate Relative Evolutionary Rates (RERs)
4. Encode a convergent trait in a phylogenetic tree
5. Correlate RERs with the convergent trait
6. Study top correlated genes in with faster RERs and slower RERs in the convergent species
7. Calculate enrichments of functional annotations in top genes

There are mammals that have convergently lost their enamel and/or teeth

Xenarthra

● 

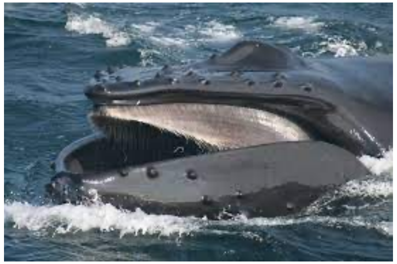
■

●

■ Pangolins

● Aardvarks

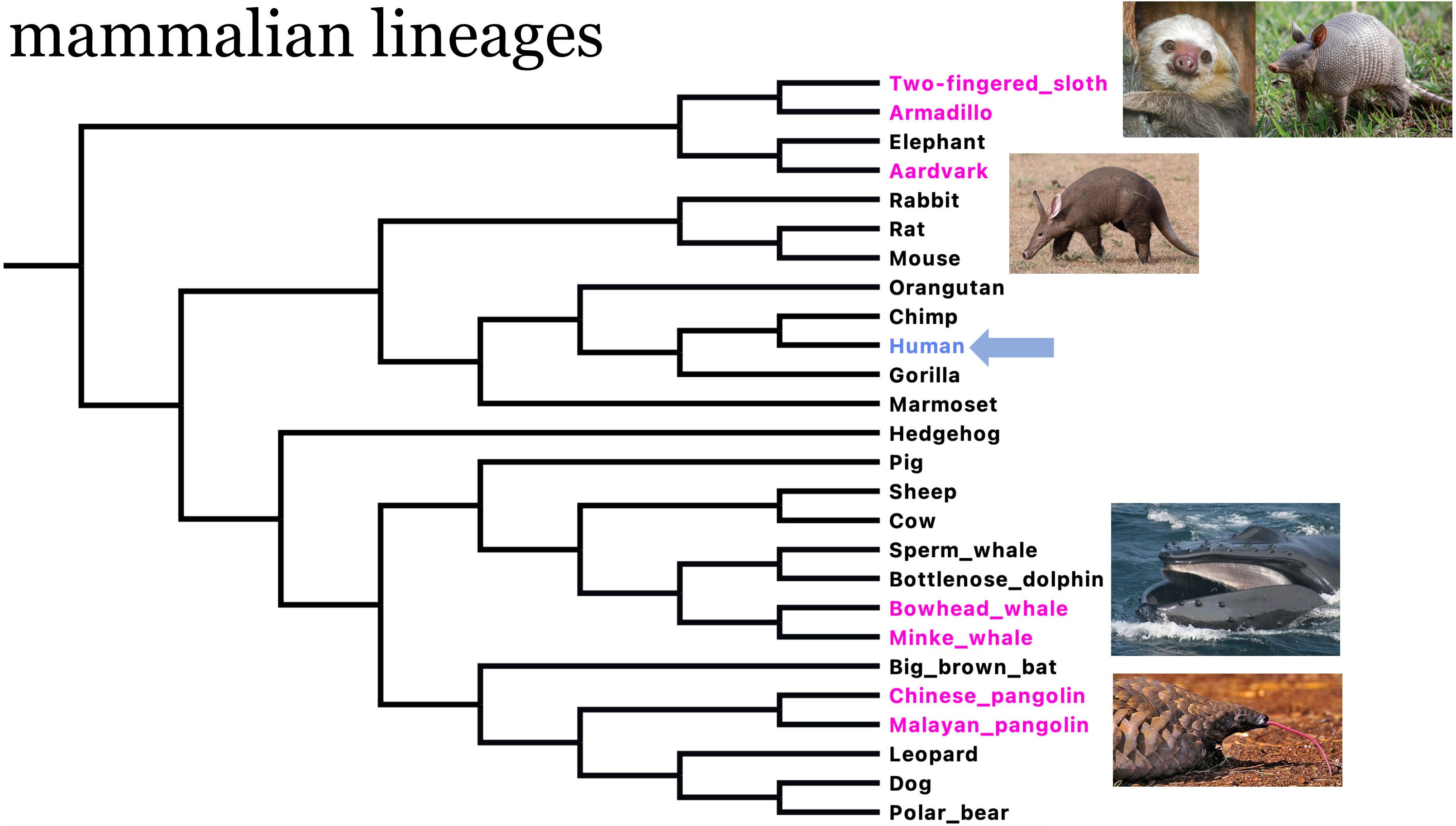
■ Baleen whales



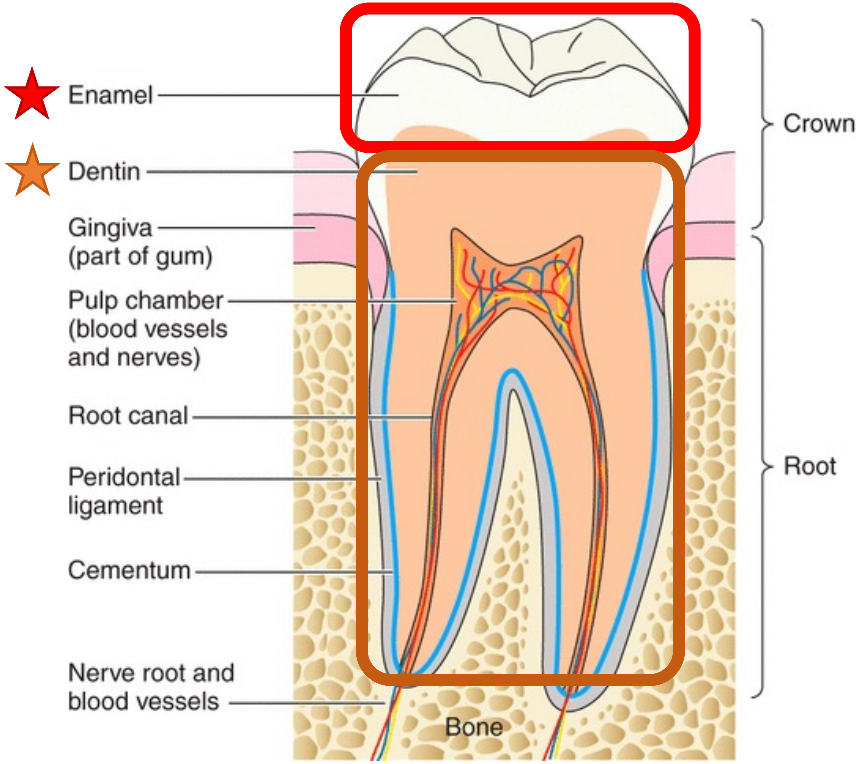
● Enamelless only

■ Enamelless and toothless

Tooth/enamel loss has occurred in several mammalian lineages



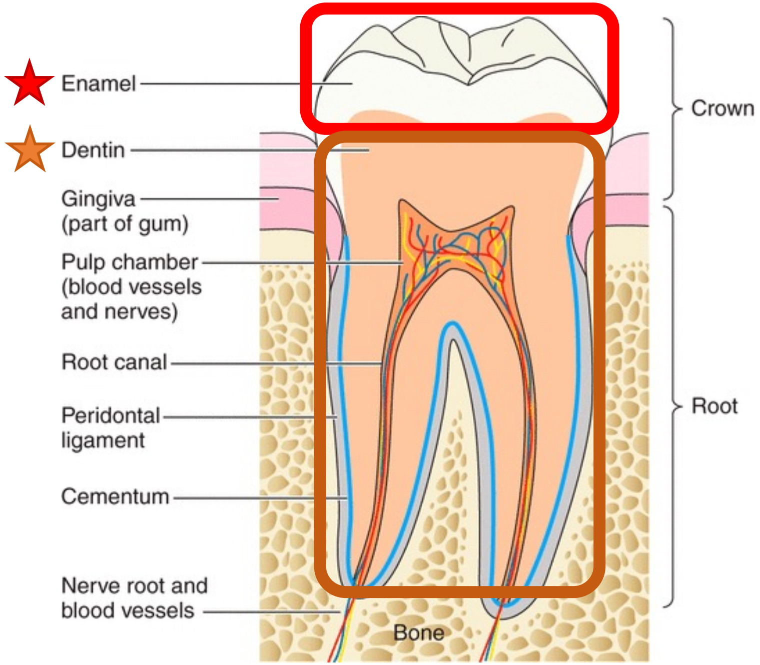
Teeth have two main parts: the crown and root



LaCruz, et al. 2017

There are several genes that contribute to tooth development… or tooth loss

ACP4, AMBN, AMEL, AMTN, ODAPH, DSPP, ENAM, KLK4, MMP20, ODAM, DMP1, MEPE



Lab Structure

Activity: Use RERconverge to find genes whose rates associate with the loss of teeth and/or tooth enamel.

Steps

1. Make gene trees for sample genes in 108 mammal species
2. Read in large number of orthologous gene trees
3. Calculate Relative Evolutionary Rates (RERs)
4. Encode a convergent trait in a phylogenetic tree
5. Correlate RERs with the convergent trait
6. Study top correlated genes in with faster RERs and slower RERs in the convergent species
7. Calculate enrichments of functional annotations in top genes

Terminology

- Relative evolutionary rate (RER) – rate of change for 1 gene over 1 branch, normalized by the expected genome-wide and tree-wide rates
- Trait tree – a phylogenetic tree encoding which branches saw trait change, and their relative weighting
- Rho – estimate of correlation coefficient, sign indicates direction (+ or -) of relationship, range -1 to 1
- signedLogP / stat [RERconverge output] – the $\log_{10}$ of the P-value, carrying the sign of Rho.
  - Top accelerated genes have extreme high values.
  - Top decelerated genes have low values.
- Stat [enrichment analysis] – statistic of the strength of enrichment of an annotation in a gene ranking