Evolutionary Adaptation & Comparative Genomics


Cape golden mole


Blind mole-rat


Naked mole-rat


Star-nosed mole

Comparative Genomics
Workshop on Genomics 2026
evomics
Český Krumlov, Czechia

Nathan Clark
University of Pittsburgh
https://nclarklab.org/
nclark.bsky.social

https://github.com/nclark-lab/PhyloGenomicsLab/tree/main

Outline: Evolutionary Adaptation and Comparative Genomics

- Goals of comparative genomics
- Preparing to compare genomes
- Phylogenetic Genotype to Phenotype Approaches
  - Gene loss / Pseudogenization
  - Gene family expansion and contraction
  - Rates-based inference
- Inferring positive selection in proteins: d_N/d_S methods
  - Branch-specific positive selection
- Inferring gene networks with Evolutionary Rate Covariation (ERC)
- Analysis of gene regulatory regions (enhancers and promoters)
- Developing Frontiers

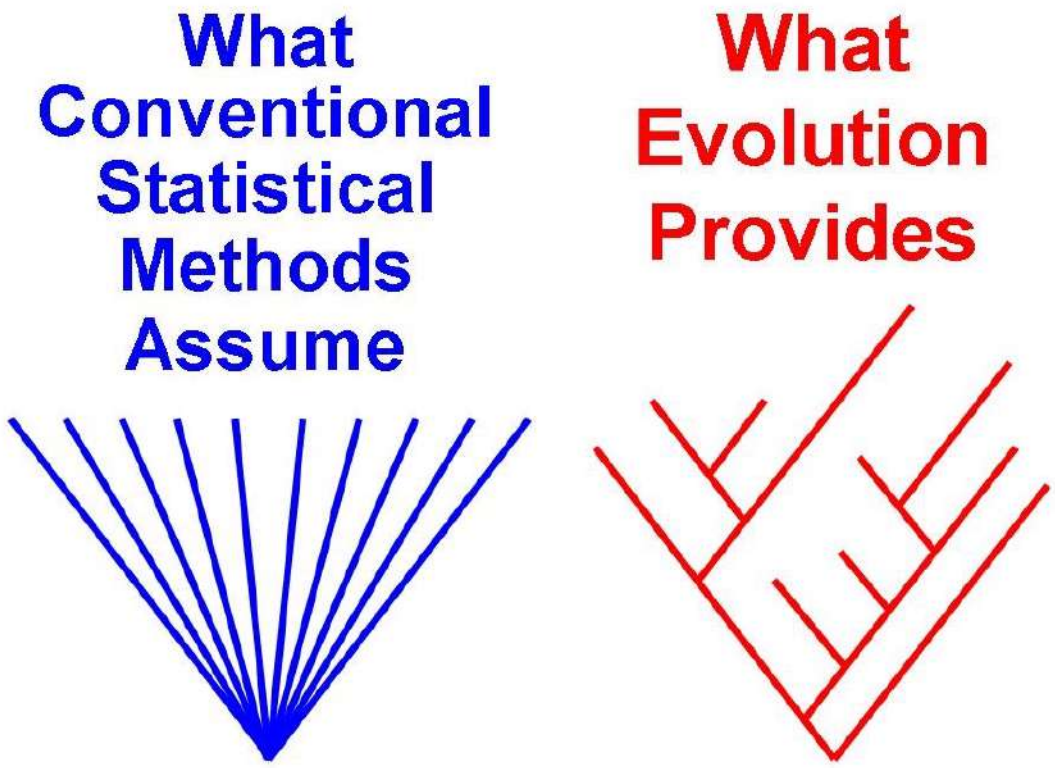
Comparative studies **must** be done using phylogeny-based methods

- Species samples are not independent. They are correlated in proportion to their relation to other species.

- Most classical statistical methods assume independence b/w samples, so applying them to comparative traits creates false conclusions.

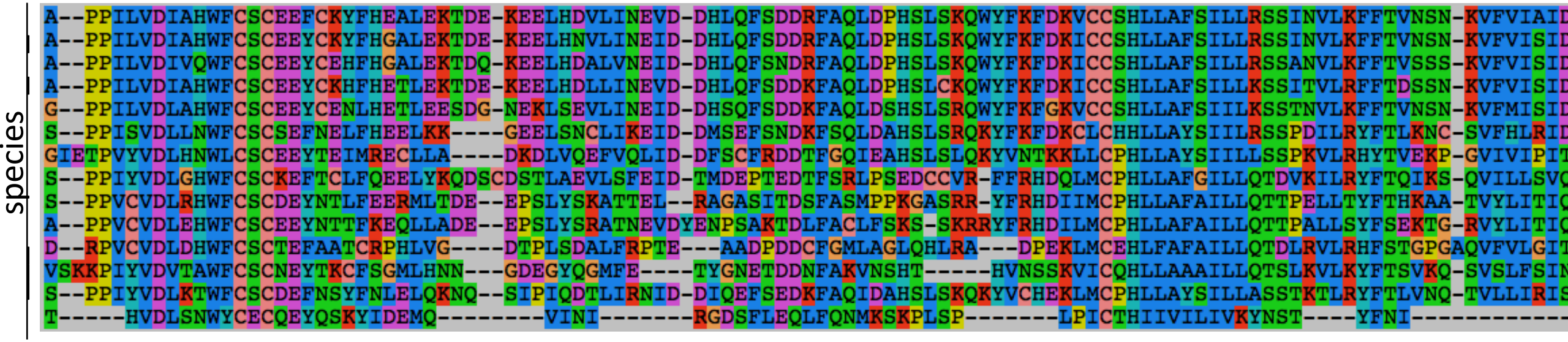
Required reading:

- Phylogenies and the Comparative Method (Felsenstein. Am. Nat. 1985)



Write down 5 observations about this alignment.

Each row is a different species.



Goals of Comparative Genomics

- Decipher the nature of genetic changes.
  - The end process of sorting through Mutation, Selection, & Drift
- Identify important functional regions (conservation)
  - phyloP, phastCons regions, conserved amino acid positions
- Infer species/gene relationships
  - phylogenetics from Wednesday and Thursday
- Identify and determine the genetic basis of Adaptive changes
- Annotate the high-level functions of genetic elements

Where to get orthologous gene alignments

1. Download MSAs from projects such as:
   - Zoonomia https://zoonomiaproject.org/
   - Vertebrate Genomes Project https://vertebrategenomesproject.org/
   - Hiller lab https://genome.senckenberg.de/download/TOGA/
   - ENSEMBL, etc...

2. Extract genes from a whole-genome alignment of many species in MAF or HAL format.
   - MAF is an alignment based on the coordinates of a reference species
     - Use **RPHAST** R package to extract sub-regions of the alignment using “sub.msa()”
   - HAL is reference-free format, produced by *ProgressiveCactus* program.

Where to get orthologous gene alignments

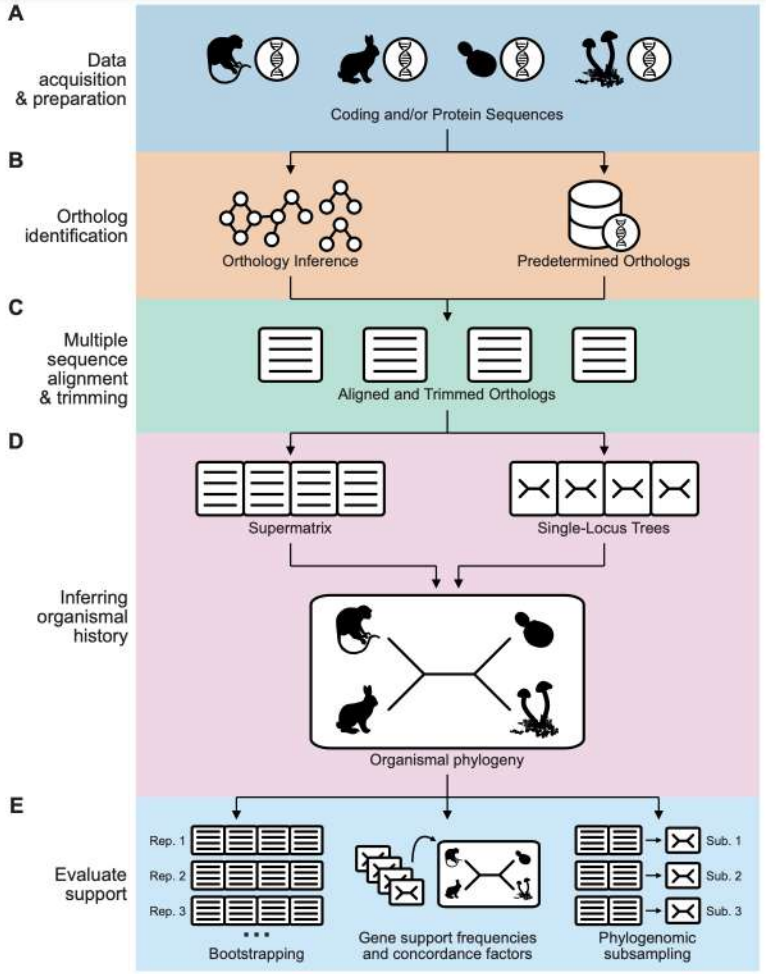
3. Download Clark lab gene trees and alignments
   - https://github.com/nclark-lab/ComparativeData/wiki
   - Available for 330 Yeast species and 420 mammals
   - Proteins and putative regulatory regions (conserved non-coding regions)

4. Group annotated genes from your species' genome annotations.
   - Collect gene models from all species in study
   - *Potentially just use the BUSCO genes already determined during gene annotation*
   - Identify orthologous gene groups using OrthoFinder followed by OrthoSnap

Jacob Steenwyk tools

Advance the broader mission to democratize rigorous bioinformatic research

Facilitating phylogenomic workflows and beyond



BioKIT

OrthoHMM

orthofisher

OrthoSNAP

ClipKIT

PhyKIT

RemarKIT

treehouse

ggpubfigs

Bjornson et al. (2024), Molecular Evolution and Phylogenetics

Engineering software for 'omic inquiry

Ortholog identification

OrthoHMM

Steenwyk et al. (2024),
bioRxiv

OrthoSNAP

Steenwyk et al. (2022),
PLOS Biology

orthofisher

Steenwyk & Rokas (2021),
G3 Genes|Genomes|Genetics

Phylogenomics

ClipKIT

Steenwyk et al. (2020),
PLOS Biology

PhyKIT

Steenwyk et al. (2021),
Bioinformatics

ORTHO FLOW

Turnbull & Steenwyk et al. (2023),
bioRxiv

Genomics

BioKIT

Steenwyk et al. (2022),
GENETICS

LVBRS

Le and Steenwyk et al. (2022),
bioRxiv

RemarKIT

Steenwyk et al. (in prep.)

Other

ggpubfigs

Steenwyk & Rokas (2021),
Micro. Resource Announcements

treehouse

Steenwyk & Rokas (2019),
BMC Research Notes

solu

Moilanen et al. (2025)
BMC Bioinformatics

450,000+ downloads

Align the Orthologous gene sets

- MUSCLE
  - Simple, fast, and effective for DNA and Protein
- Prank 
  - Phylogeny-aware alignment produces highly biologically accurate alignments
  - DNA, Codons, Protein
  - Is slower but reasonable after providing a guide species tree
- MACSE
  - Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons

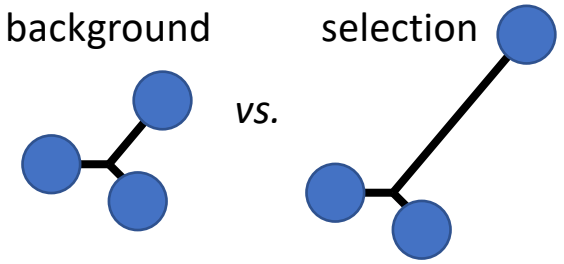
Comparative Approaches to Map Adaptive Phenotypes to Genotypes

Genetic crosses between populations or species.

- Genetic mapping
  - Peromyscus coat color Mc1r (Linnen et al *Science* 2009)
  - blind cavefish

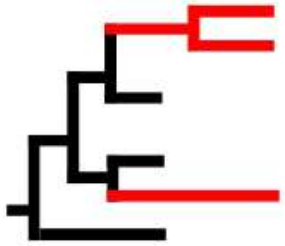
Linnen et al *Science* 2009

- Differential expression (DE) analysis
  - Which genes are changing between species?
  - Should be analyzed with phylogeny-based tools
    - Such as phylogenetic generalized least squares regression (PGLS)



Positively selected regions leading to trait

- Population Genetics (F_{ST} , Selective sweeps…)
  - High altitude populations and HIF1a
  - (Simonson et al *Science* 2010; Beale et al PNAS 2010)
- Phylogenetics (d_N/d_S)



Rate acceleration correlated with convergent trait

PhyloG2P – phylogenetic genotype to phenotype

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

Relating genotype to phenotype - **Population Genetics Approach**

- Advantages
  - Large samples sizes -- power to detect small effect sizes at **stringent cutoffs**
  - Easy-to-correct confounders so that we can establish causality
- Disadvantages
  - Limited genotype diversity
  - Limited phenotype diversity
    - No flying people!

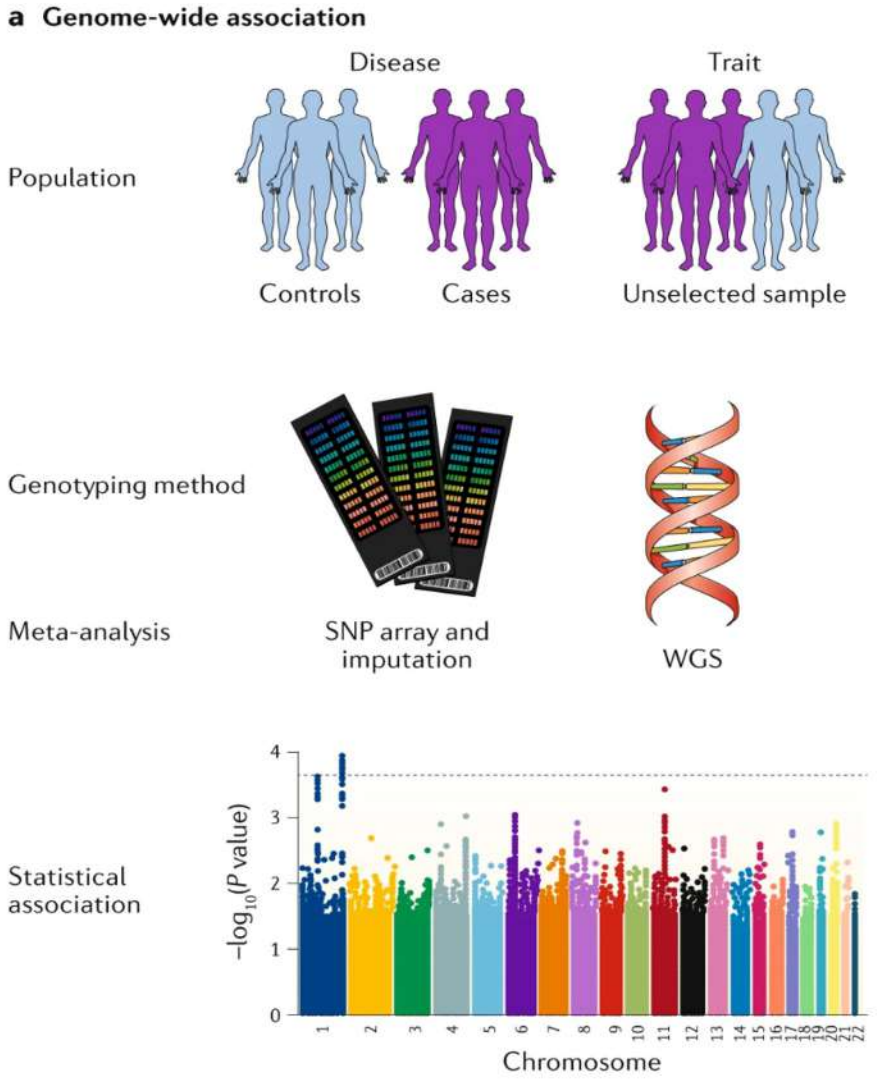


Image Credit: Benefits and limitations of genome-wide association studies

Relating genotype to phenotype - **Comparative Genomics Approach**

- Disadvantages
  - Fundamentally limited samples sizes
  - Impossible-to-correct-for confounders
- Advantages
  - Large genotype diversity (maybe a disadvantage)
  - **Large phenotype diversity**
    - **Bats can fly!**

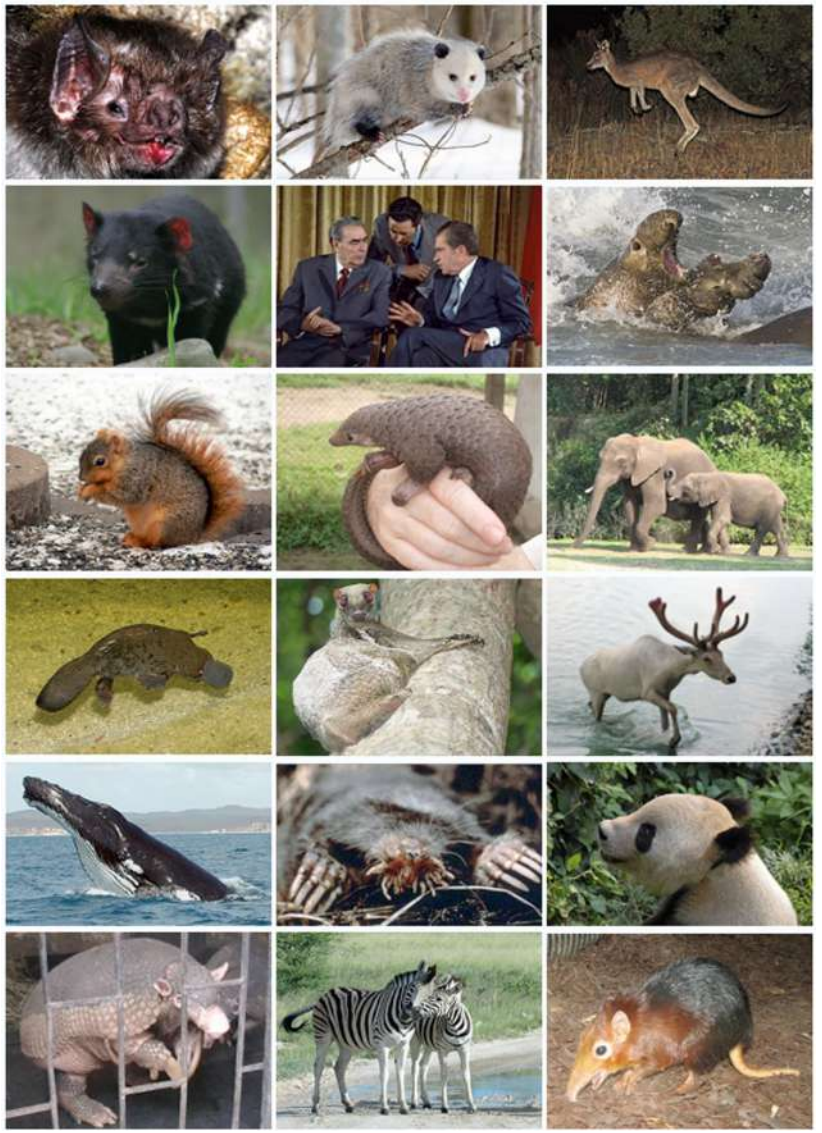
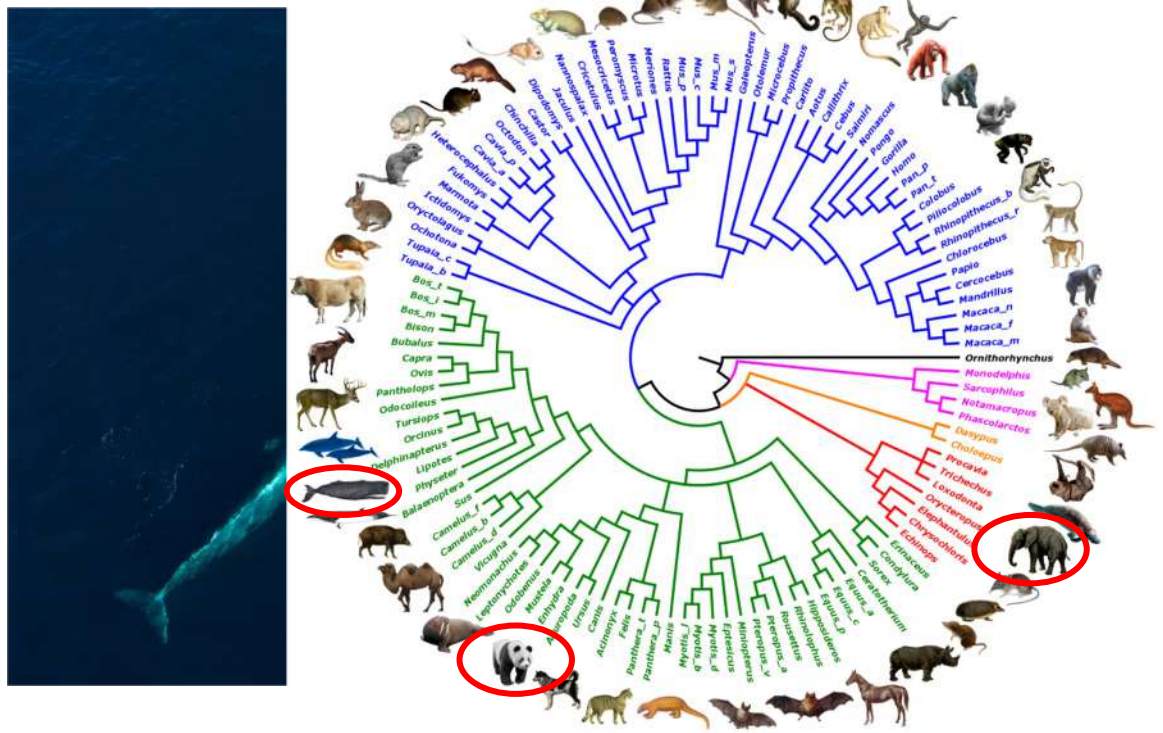


Image Credit: Wikipedia Mammals

Large phenotype diversity example

Dramatic body size diversity



Shrew: 2g

Whale: 10^8g

Much like in the GWAS case we need more more than one species with a the phenotype of interest -- **convergent traits**

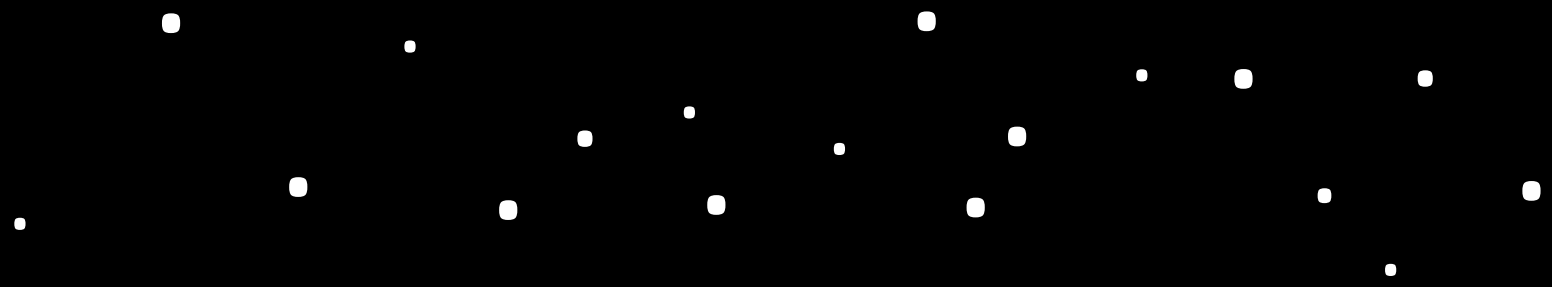
Convergent evolution provides unique opportunities to study repeated adaptation and genetic change

We aim to:

1. Identify genetic changes underlying novel, responsive traits.
2. Discover unexpected phenotypic and genetic changes.

from genes

Protein-coding genes Regulatory sequences RNA genes



from genes to high level function

PhyloG2P

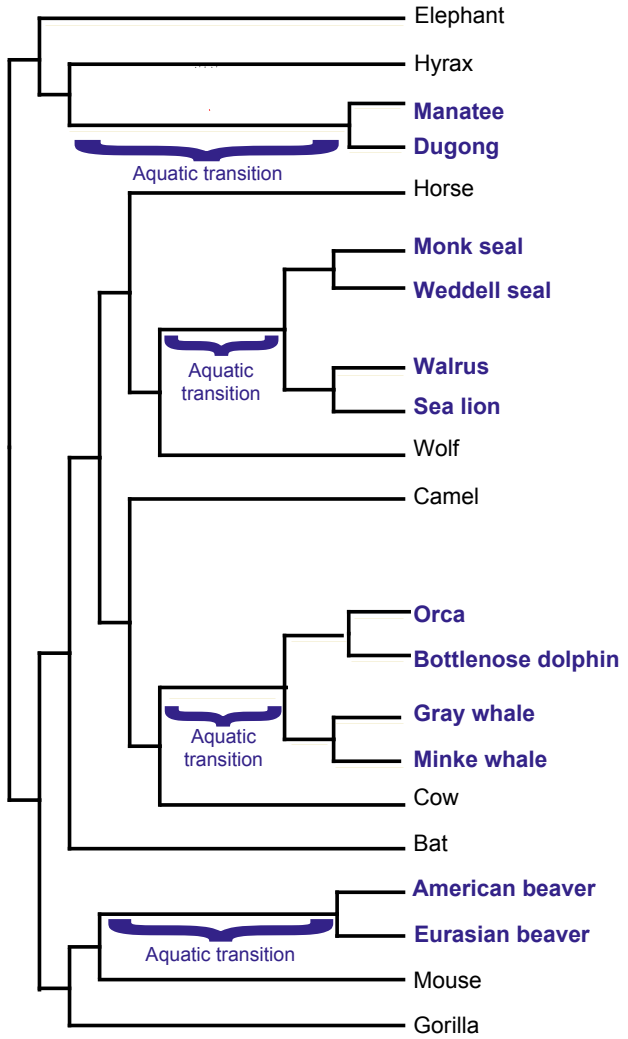
Phylogenetic Genotype to Phenotype

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



Clark et al. *Ann Rev Genet* 2025

PhyloG2P

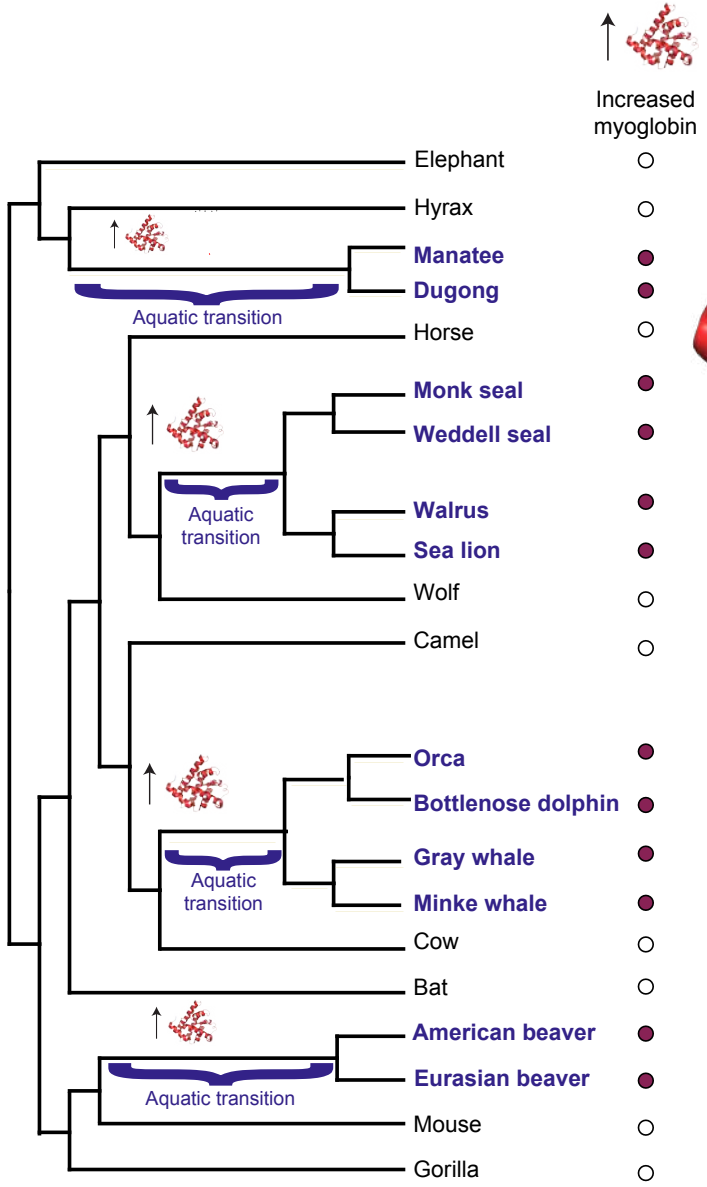
Phylogenetic Genotype to Phenotype

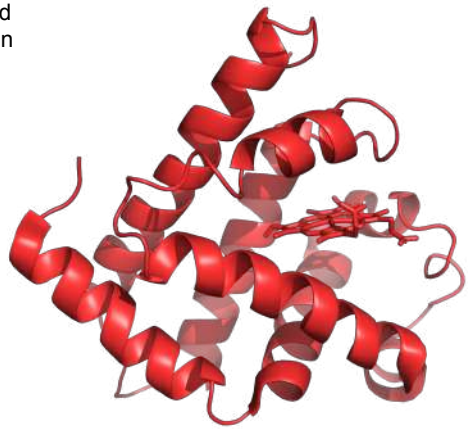
Which genetic events are associated with a trait?

Genetic events include:

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

Novel traits produce **transient genetic signals**.





Increased Myoglobin

Clark et al. *Ann Rev Genet* 2025

PhyloG2P

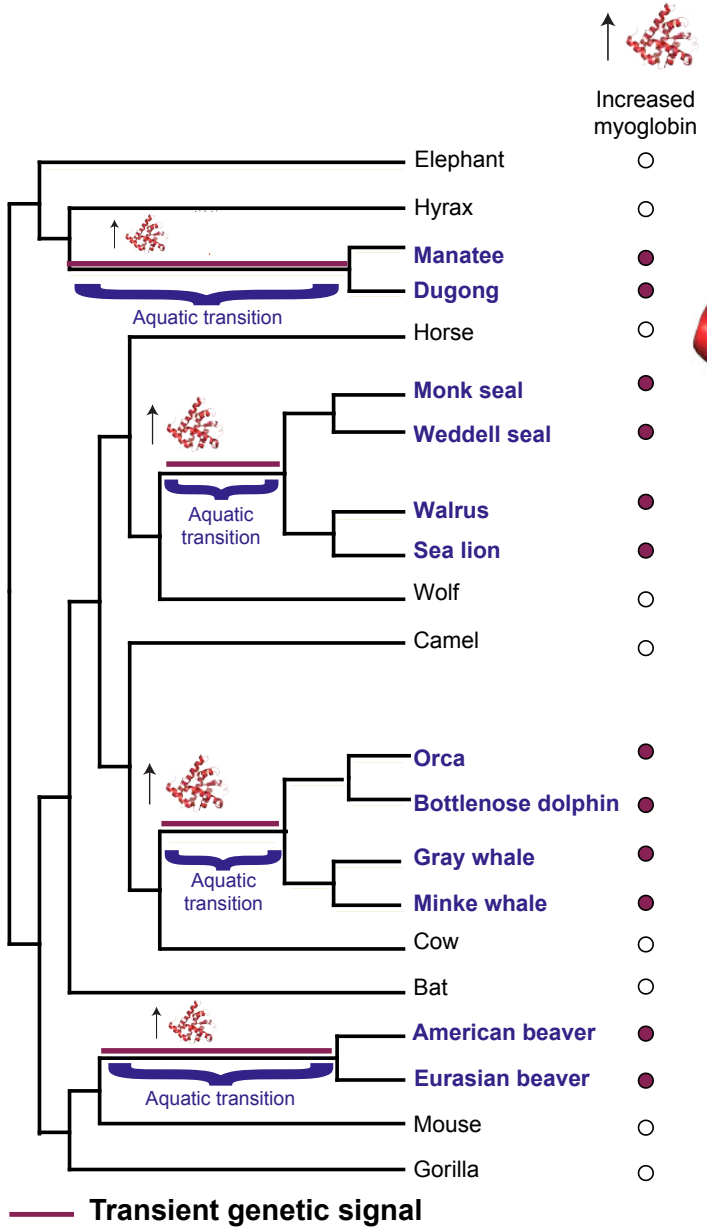
Phylogenetic Genotype to Phenotype

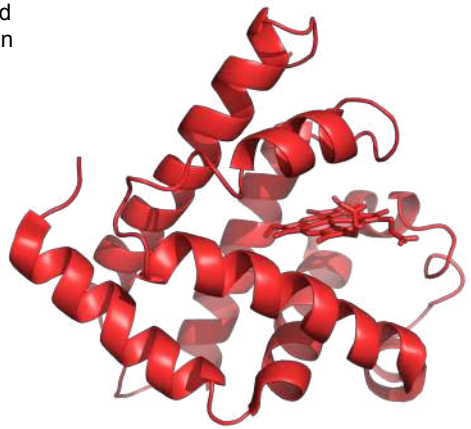
Which genetic events are associated with a trait?

Genetic events include:

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

Novel traits produce **transient genetic signals**.





Increased Myoglobin

Clark et al. *Ann Rev Genet* 2025

PhyloG2P

Phylogenetic Genotype to Phenotype

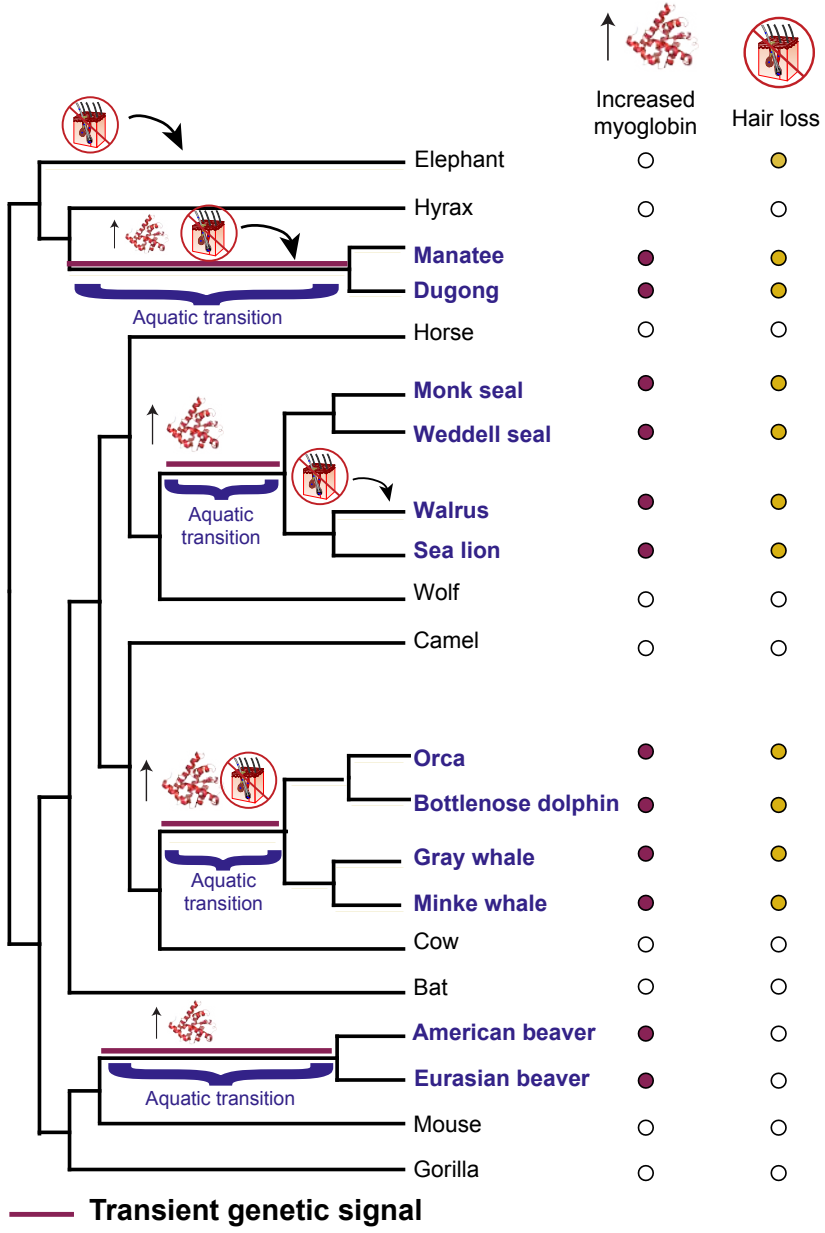
Which genetic events are associated with a trait?

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**.



Clark et al. *Ann Rev Genet* 2025

Indicates putative branches on which pictured phenotypes first appeared

PhyloG2P

Phylogenetic Genotype to Phenotype

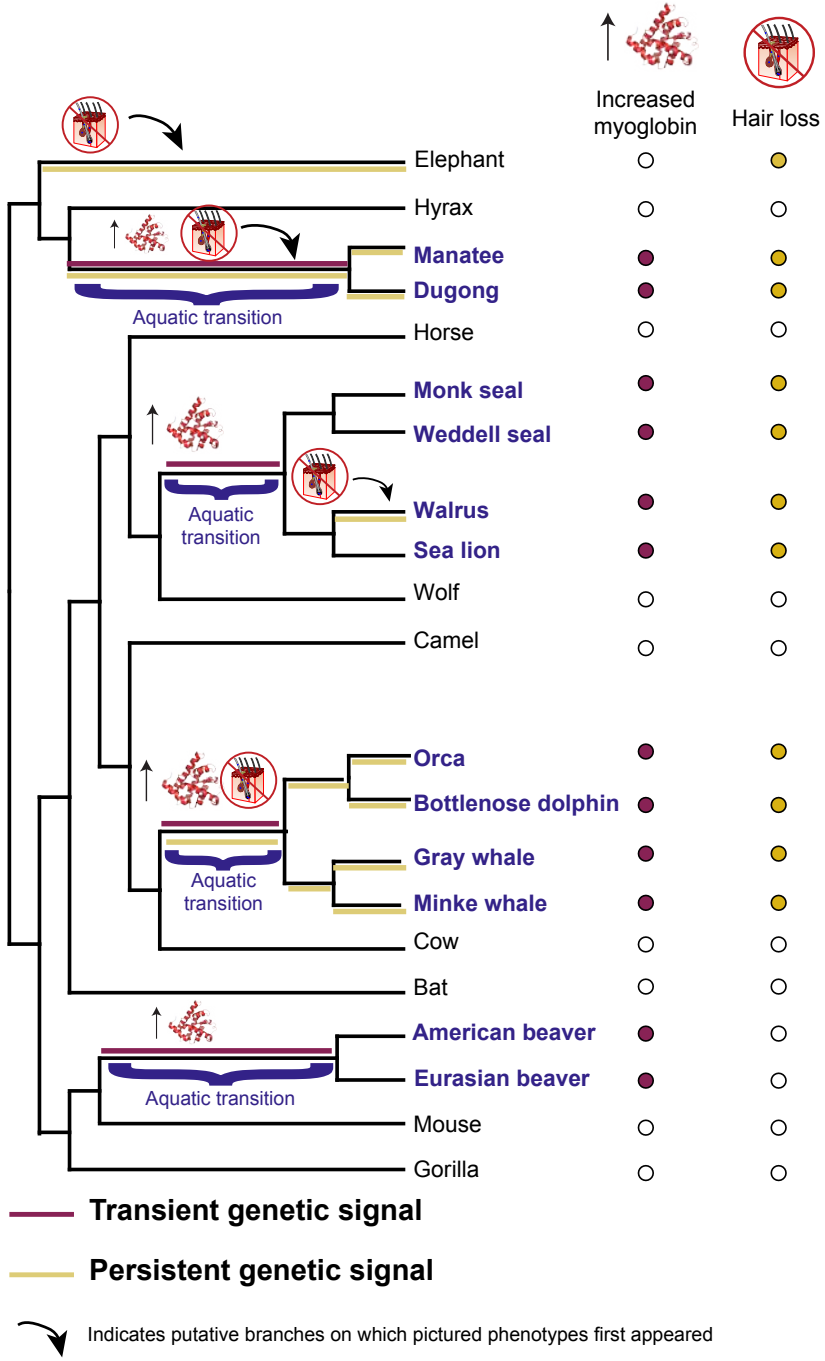
Which genetic events are associated with a trait?

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**.



Clark et al. *Ann Rev Genet* 2025

PhyloG2P

Phylogenetic Genotype to Phenotype

Which genetic events are associated with a trait?

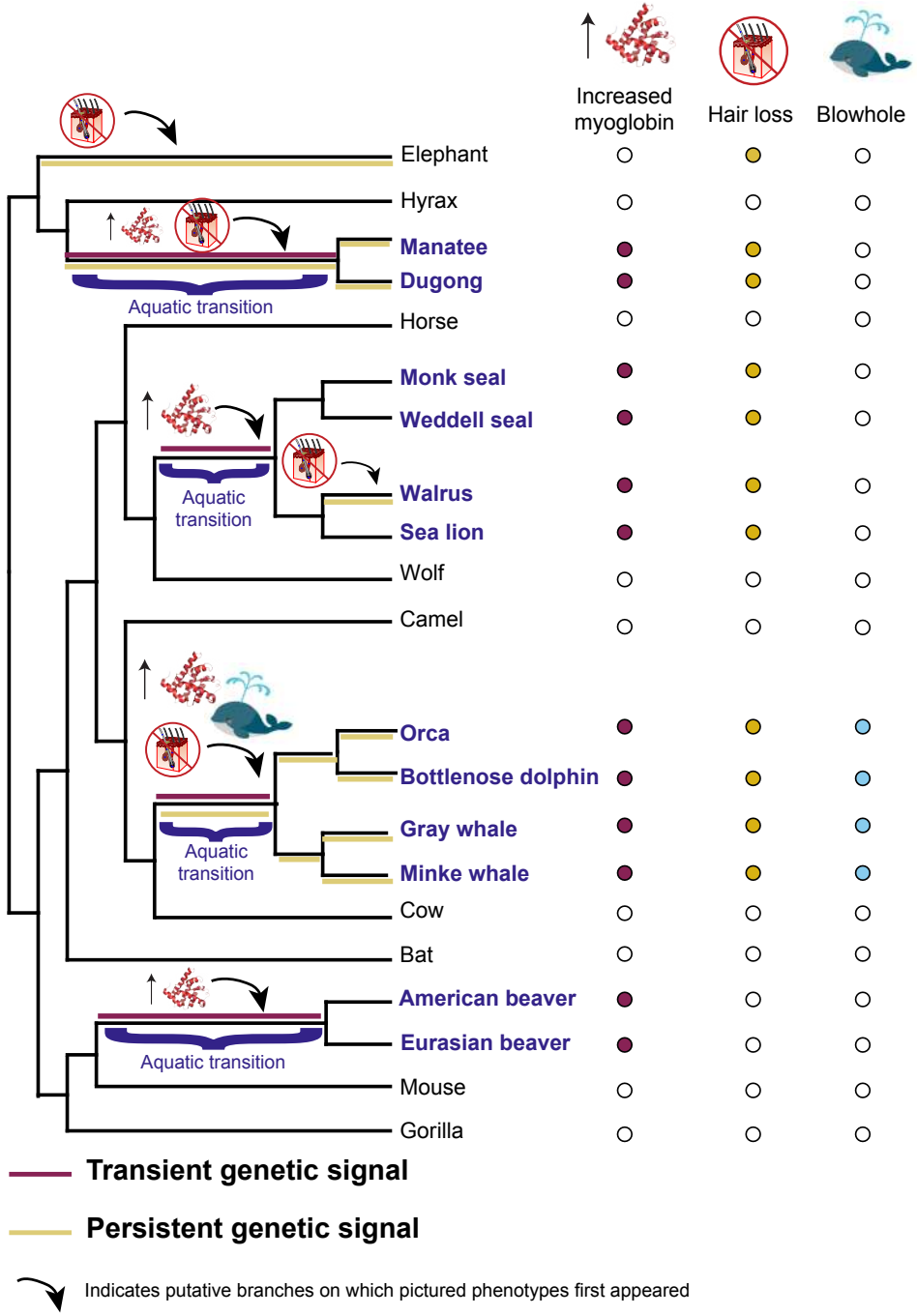
Genetic events include:

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

Novel traits produce **transient genetic signals**.

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

Idiosyncratic changes are not captured.



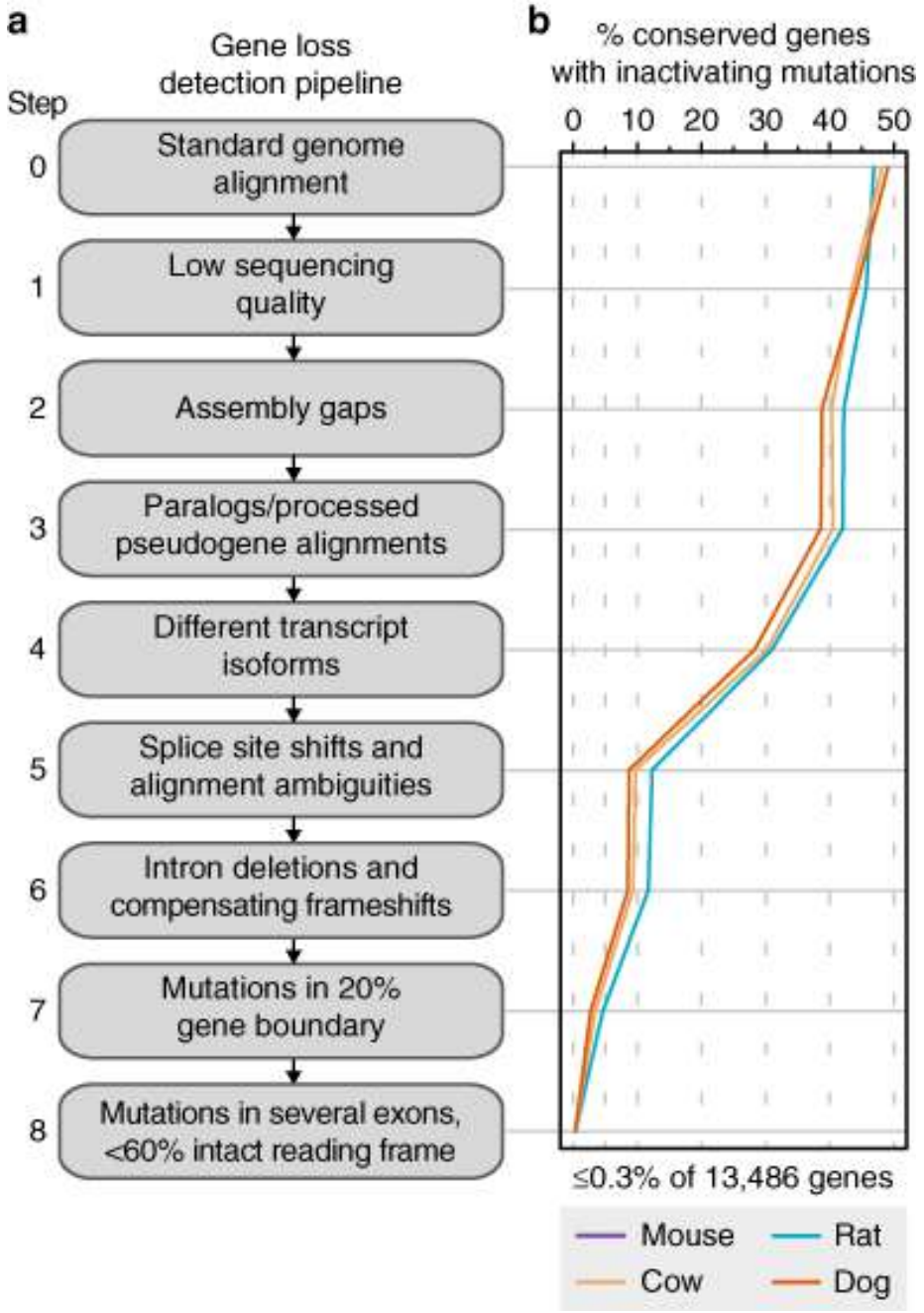
Clark et al. *Ann Rev Genet* 2025

Gene Loss / Pseudogenization

Scanning genomics for deactivated (pseudogenized) gene copies

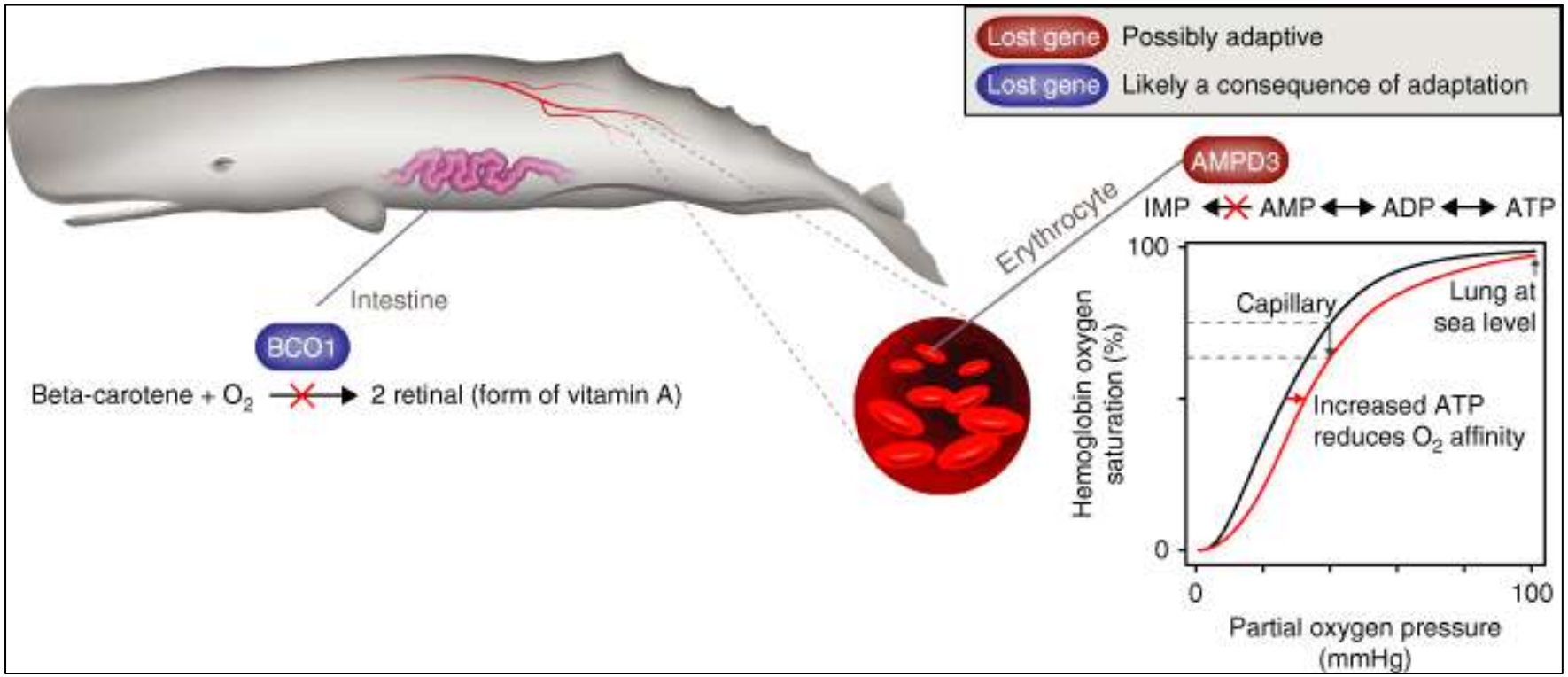
The different steps address the many challenges related to assembly and alignment issues.

Michael Hiller lab's **GENE LOSS DETECTION PIPELINE** or **TOGA2** identify function and non-functional gene copies in collections of closely related species.



Sharma *et al*. Nature Comm. 2018

Diving and dietary adaptations in sperm whales

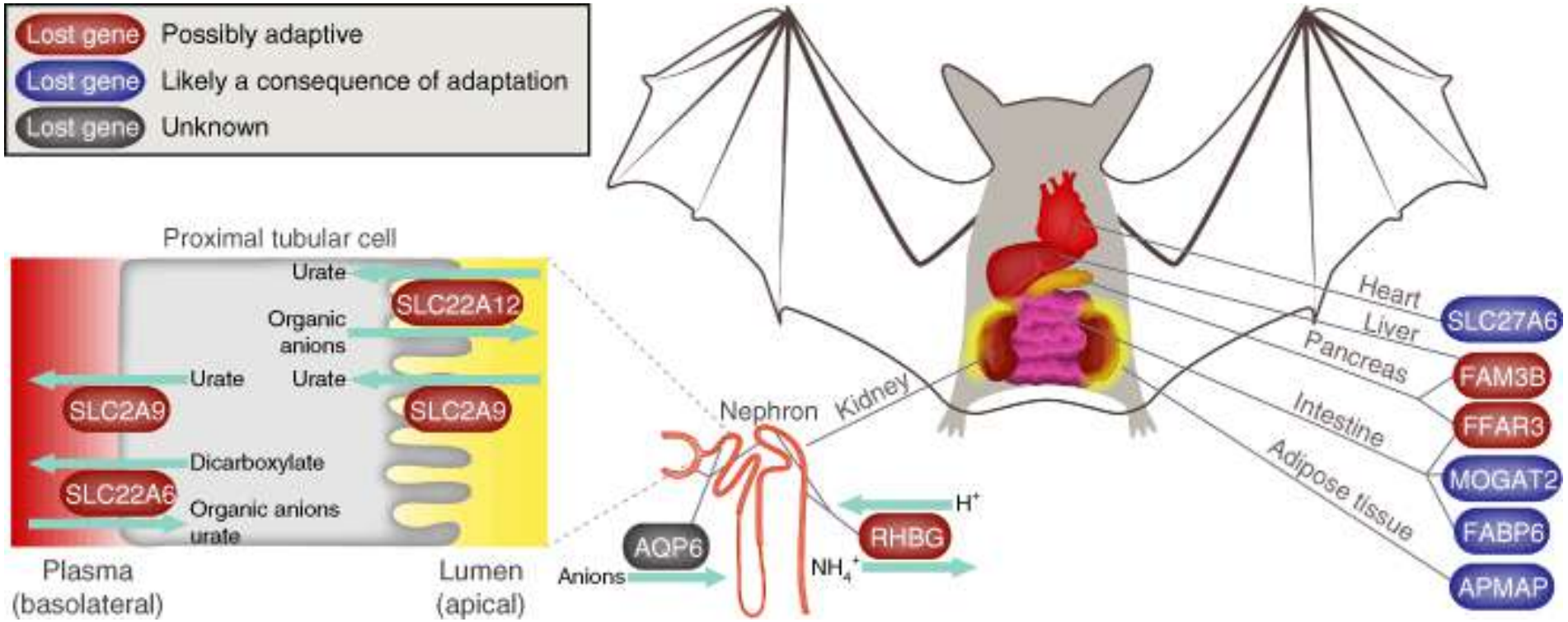


Loss of *AMPD3* is likely an adaptation to the extreme diving ability of sperm whales. *AMPD3* loss increases the level of ATP (an allosteric hemoglobin effector), which facilitates O_2 release.

The loss of the vitamin A synthesizing enzyme ***BCO1*** is likely due to relaxed selection from sperm whale's specialized diet that mainly consists of vitamin A-rich but beta-carotene poor squid.

Sharma *et al*. Nature Comm. 2018

Renal and metabolic adaptations in frugivorous bats



A number of **renal transporter genes** (left side) that are specifically lost in fruit bats reduce urine osmolality. Thus, these gene losses likely contribute to the ability of fruit bats to efficiently excrete excess dietary water.

Losses of **metabolic genes** (right side) are likely adaptive by improving the processing of the sugar-rich fruit juice. In contrast, **gene losses shown in blue** are probably a consequence of adapting to the frugivorous diet.

Sharma *et al*. Nature Comm. 2018

Convergent Evolution in Marine Mammals

Cetaceans

Sirens

Pinnipeds

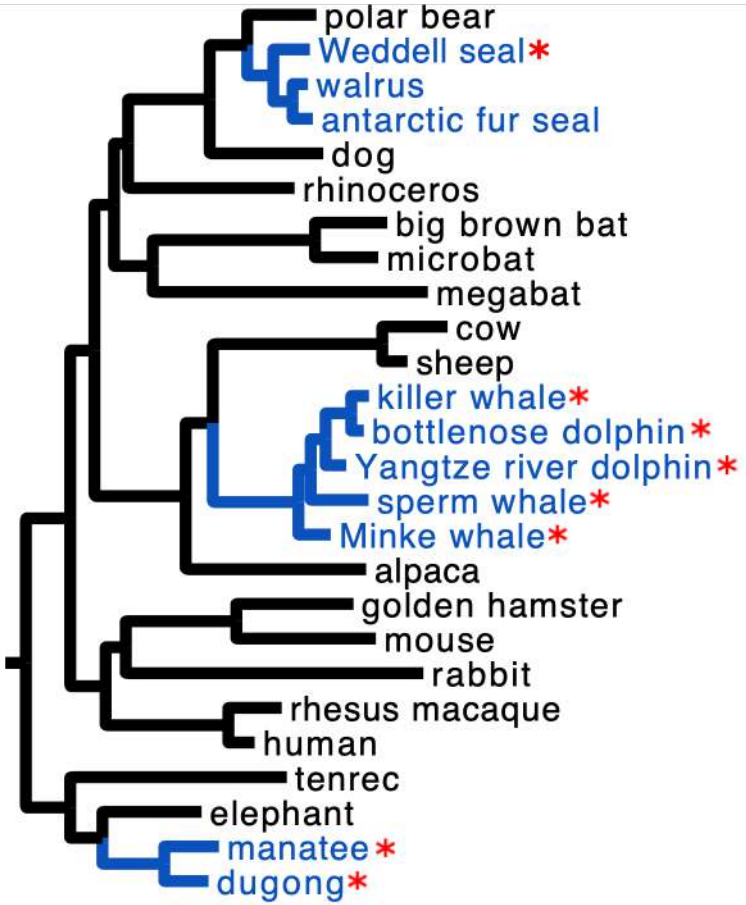
Convergent changes:

- streamlined morphology
- thick, compact epidermis to provide a low friction and impermeable barrier
- sensory systems are adapted for underwater
- respiratory physiology is tuned for diving under great pressure and hypoxia

At the molecular level, convergent evolution has been more difficult to identify, until now…

Identifying genes whose ***repeated loss*** correlates with aquatic environment

*gene contains lesions
(early stop codons, frameshifts)



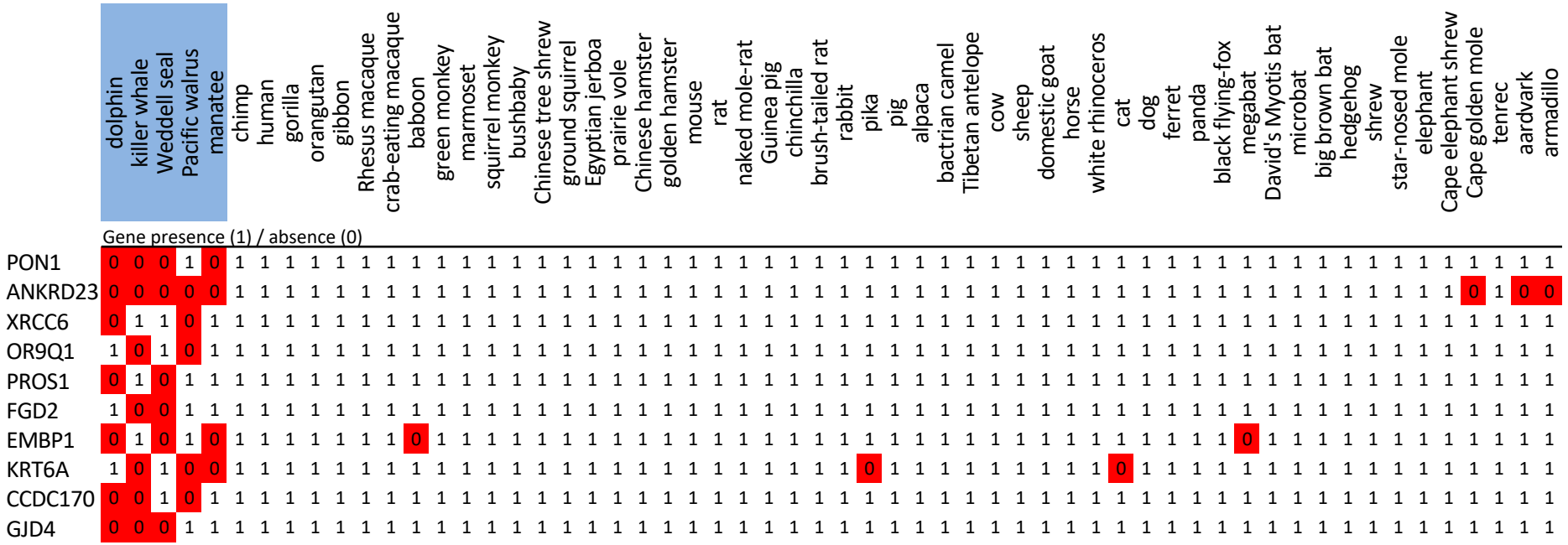
Gene	Loss rate (independent)	Marine loss rate (dependent)	Terrestrial loss rate (dependent)	LRT statistic	Empirical P-value	FDR	Description
PON1	0.672	49.7	0	22.24	3.08×10^{-6}	0.0154	paraoxonase 1
OR10Z1	1.15	100	0.467	19.99	7.25×10^{-6}	0.0201	olfactory receptor
OR8D4	1.25	100	0.510	19.21	1.60×10^{-5}	0.0201	olfactory receptor
TAS2R1	1.32	100	0.535	19.20	1.60×10^{-5}	0.0201	taste receptor
OR1F2P	2.03	100	1.18	15.86	5.40×10^{-5}	0.0831	olfactory receptor
GSTM1	1.48	100	0.762	15.82	3.90×10^{-5}	0.0831	glutathione S-transferase mu 1
OR6K2	2.02	100	1.22	15.79	4.50×10^{-5}	0.0831	olfactory receptor
OR51D1	1.13	49.3	0.466	15.59	8.60×10^{-5}	0.0831	olfactory receptor
TAAR5	1.17	48.2	0.484	15.16	9.90×10^{-5}	0.0936	trace amine associated receptor 5
OR4C13	1.77	100	0.915	14.88	7.00×10^{-5}	0.0972	olfactory receptor

Claire Kronk

Wynn Meyer

Jerrica Jamison

Using BayesTraits to infer coincidence of evolutionary events



	Independent model		Dependent model				***via BayesTraits by Mark Pagel***
Gene	Log likelihood	gene loss rate	Log likelihood	marine loss rate	terrestrial loss rate	P-value	Description
PON1	-26.7294	2.31	-16.3059	100.00	0.00	4.97E-06	paraoxonase 1
ANKRD23	-33.0866	4.90	-25.7259	100.00	2.52	1.25E-04	ankyrin repeat domain 23 (skeletal muscle)
XRCC6	-25.3000	1.51	-18.0866	59.75	0.00	1.46E-04	X-ray repair cross-complementing 6 (Ku70)
OR9Q1	-25.4322	1.51	-18.2294	59.37	0.00	1.47E-04	olfactory receptor
PROS1	-25.4304	1.51	-18.2478	58.62	0.00	1.51E-04	protein S (alpha)
FGD2	-25.5627	1.51	-18.3903	58.26	0.00	1.52E-04	FYVE, RhoGEF and PH domain containing 2
EMBP1	-38.7641	3.81	-31.6236	100.00	1.58	1.57E-04	embigin pseudogene 1
KRT6A	-33.7456	3.89	-26.6269	100.00	1.61	1.61E-04	keratin 6A (epithelial, type II cytokeratin)
CCDC170	-23.3382	1.53	-16.3374	77.81	0.00	1.83E-04	coiled-coil domain containing 170

BayesTraits is a likelihood and Bayesian program by Mark Pagel that implements phylogenetic models of trait change.

It estimates rates of change, including in the context of another variable.

In this study, the models were used to test the hypothesis that there is a higher rate of trait loss in the aquatic condition compared to terrestrial.

Has Pon1 been inactivated in manatees and dugongs?

Sample: 7 manatees from Florida

3 dugongs from Australia

Jerrica Jamison

Wynn Meyer

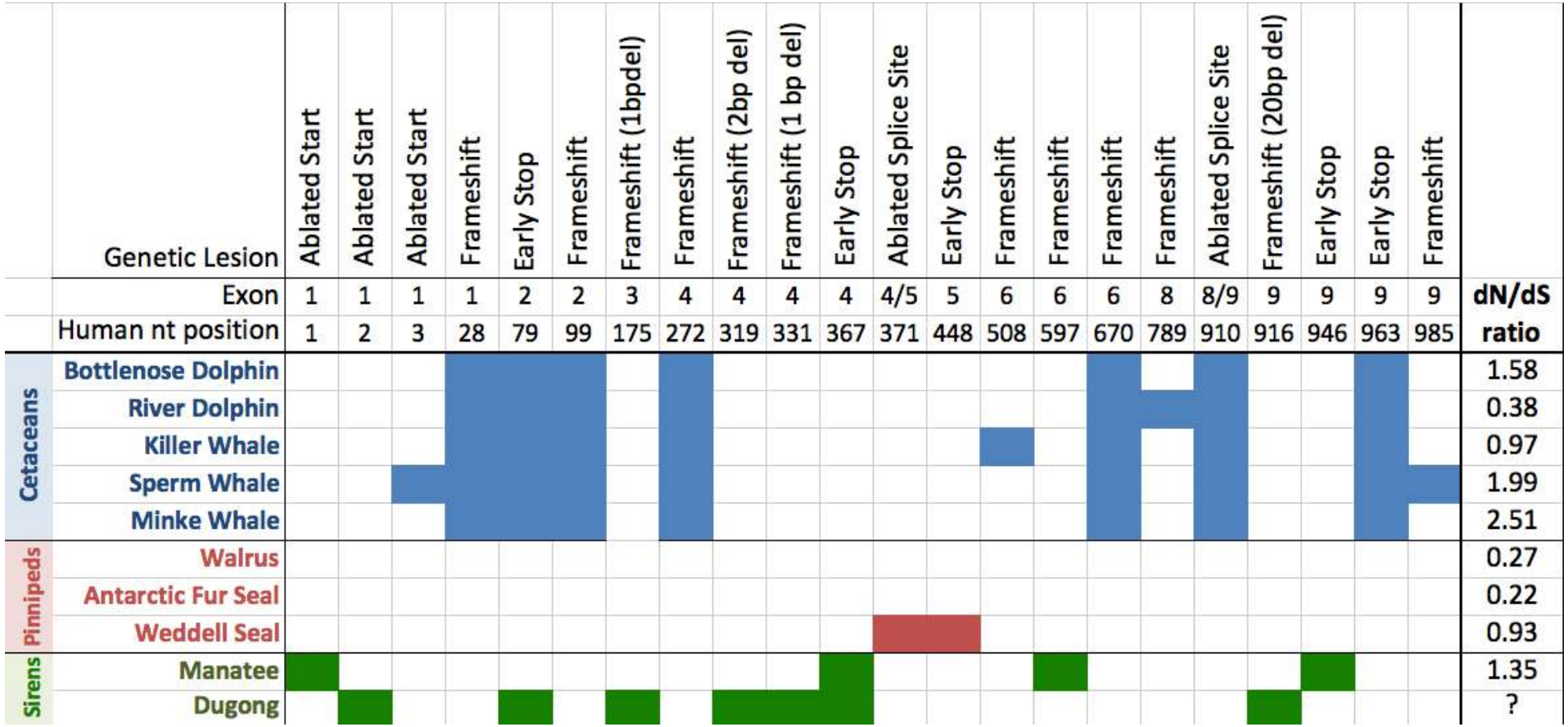
PITTSBURGH ZOO & PPG AQUARIUM

Joe Gaspard

Bob Bonde
USGS Gainsville

Janet Lanyon
University of Queensland

Summary of *Pon1* Genetic Lesions in Marine Mammals



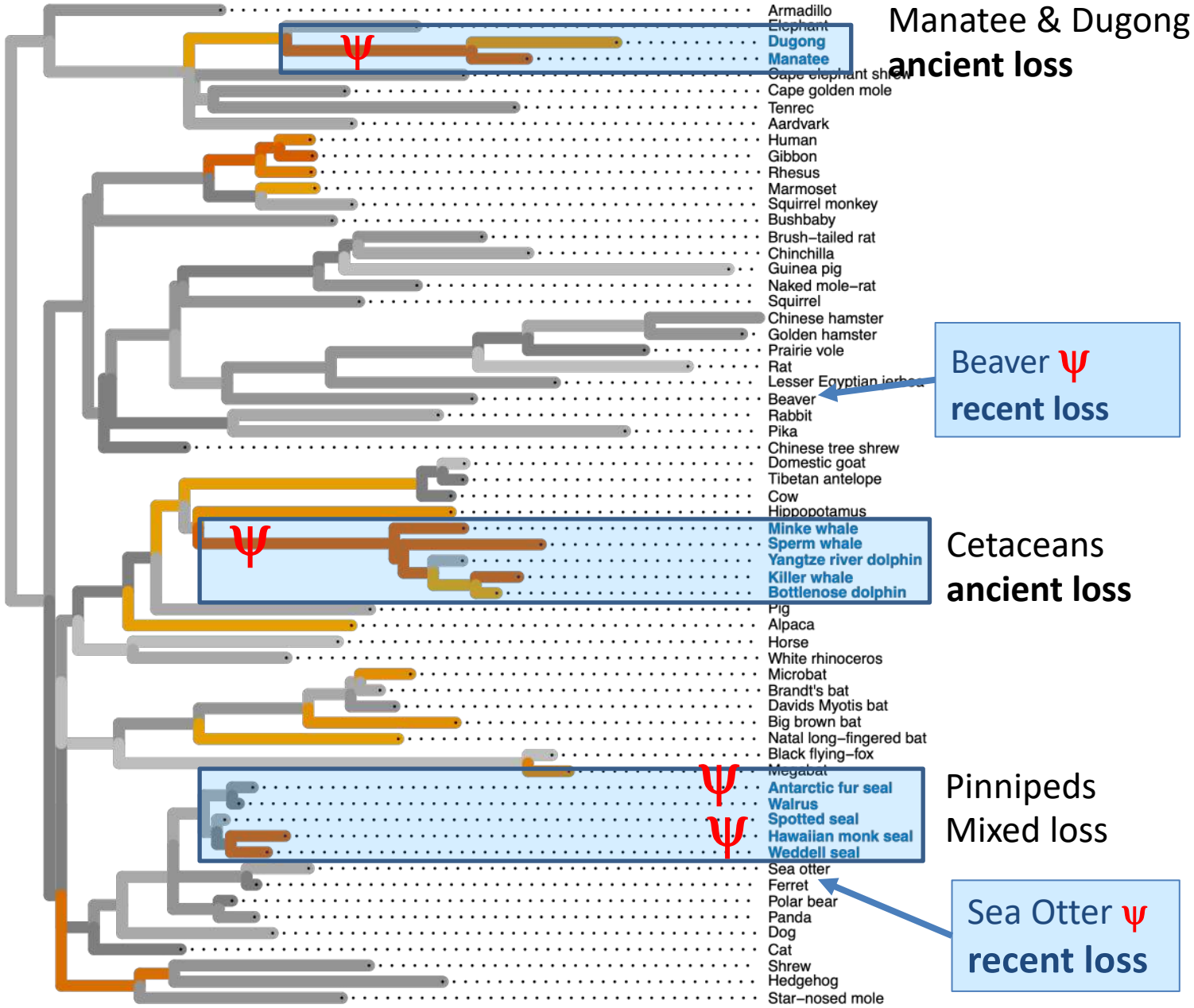
PON1 gene loss across mammals

Evidences of loss:

- DNA gene sequence
- mRNA levels in liver
- activity in blood plasma

Examined in >120 species

ψ = pseudogenization



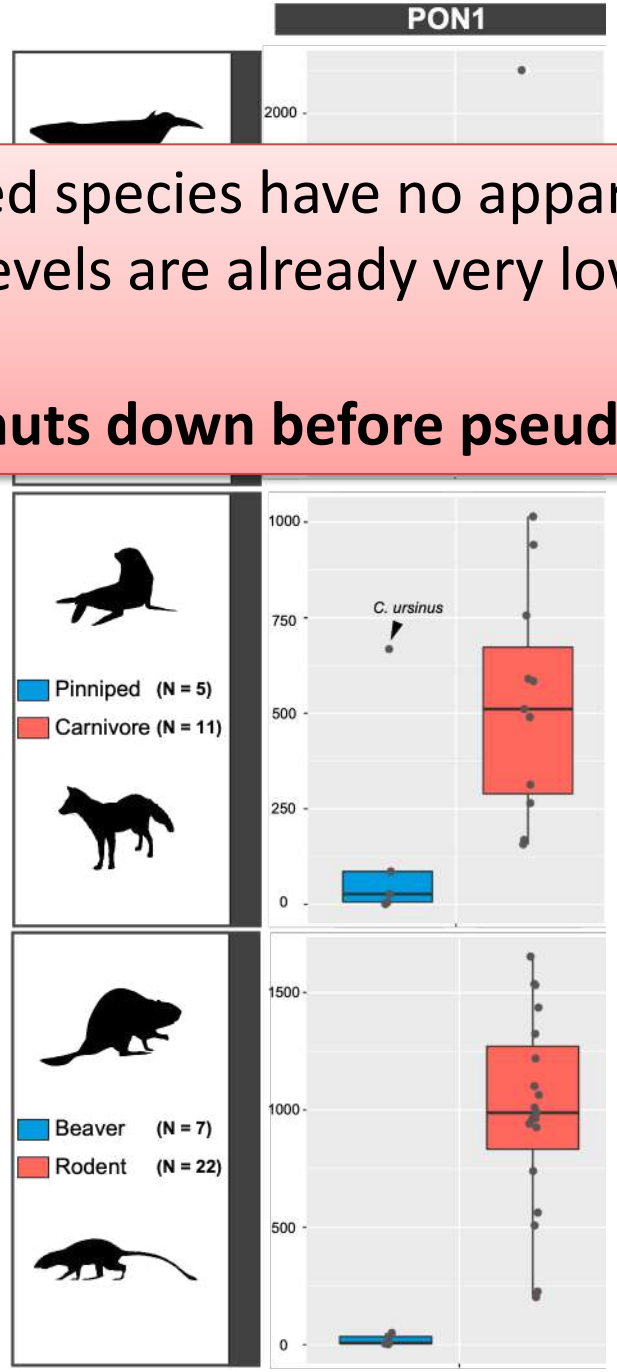
Pon1 mRNA is very low in aquati species

Many Pinniped species have no apparent genetic lesions, but their mRNA levels are already very low!

Expression shuts down before pseudogenization!

Pon3 is also low in pinnipeds and beavers

TPM (transcripts per million) in liver



Allie Graham

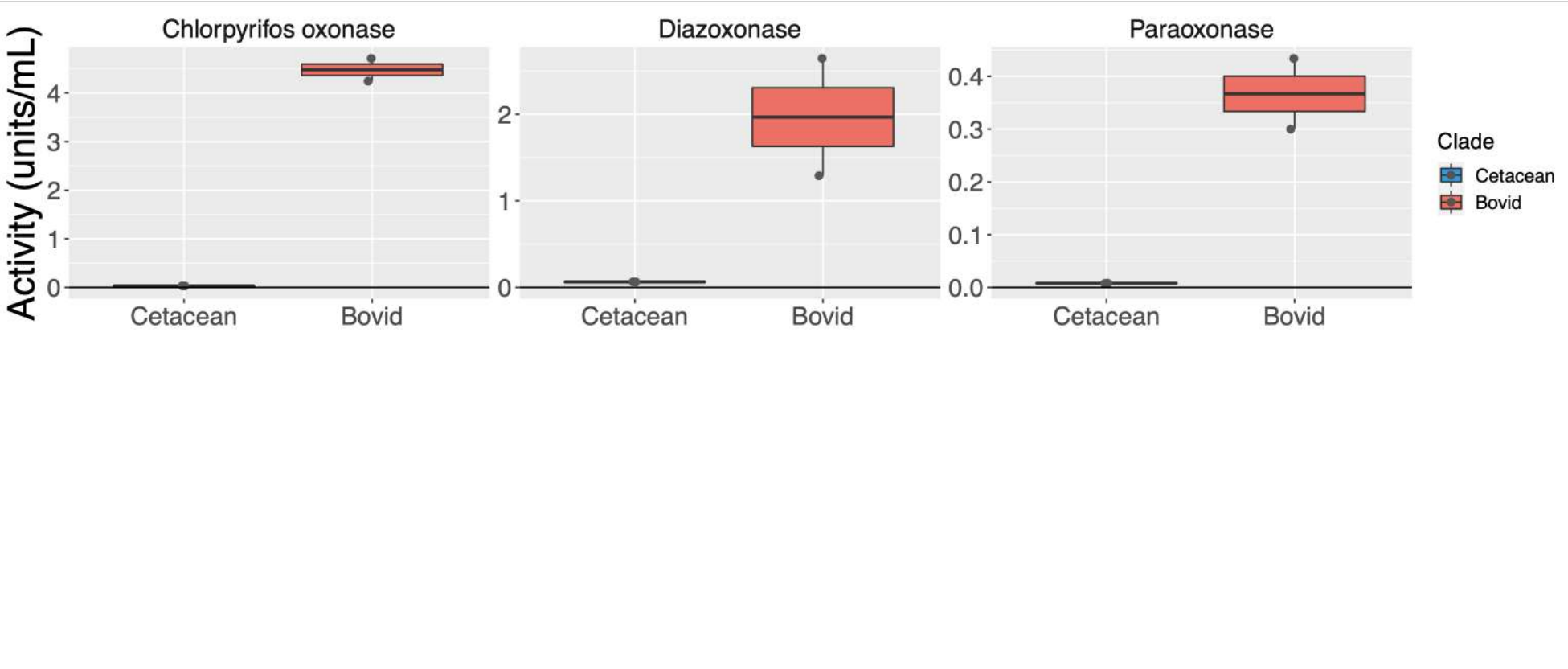
Biochemical activities in blood plasma support loss

Clem Furlong

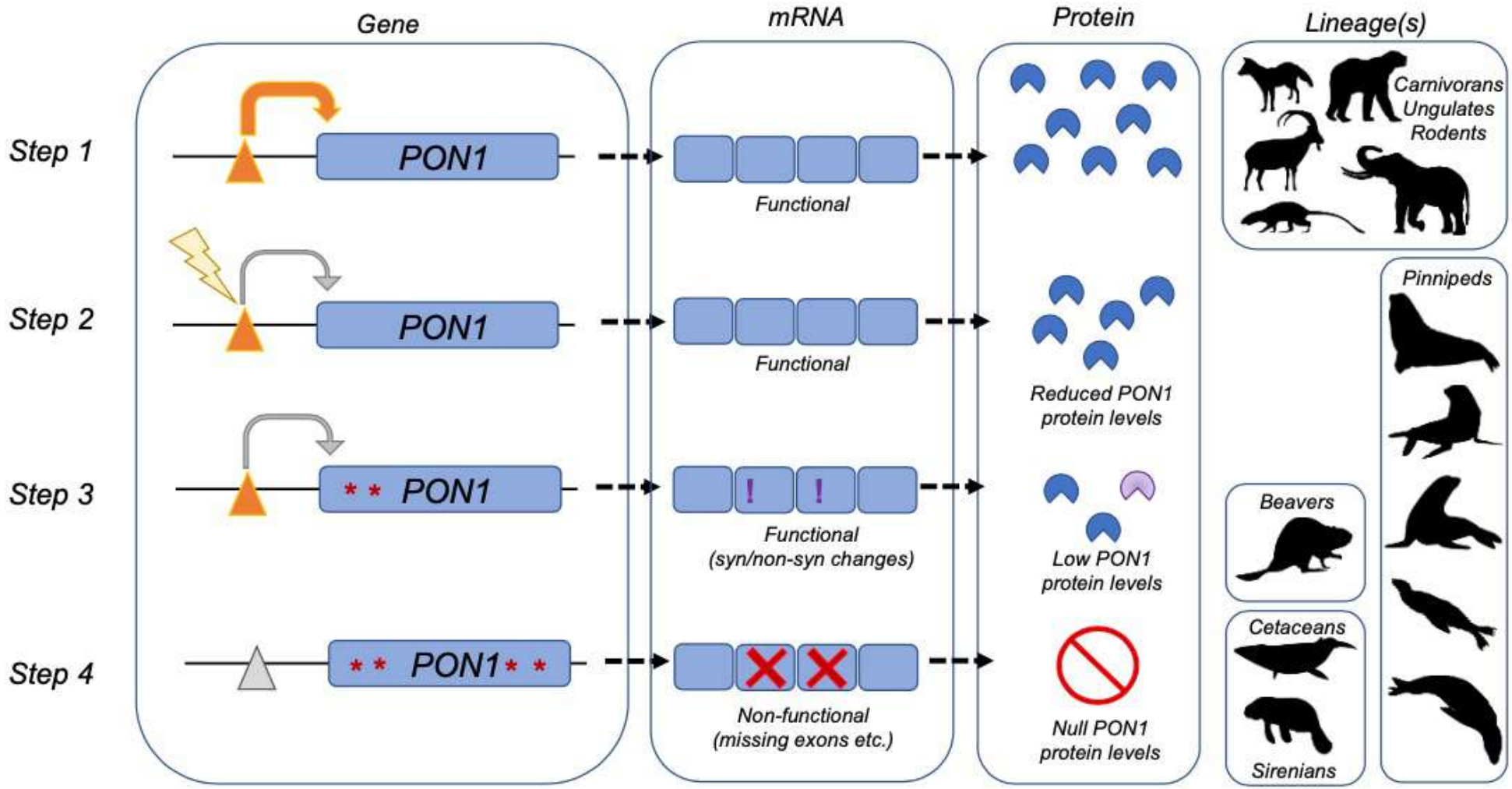
Rebecca Richter

Judit Marsillach

University of Washington

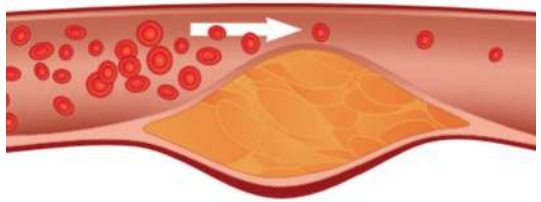


Model for Pon1 inactivation

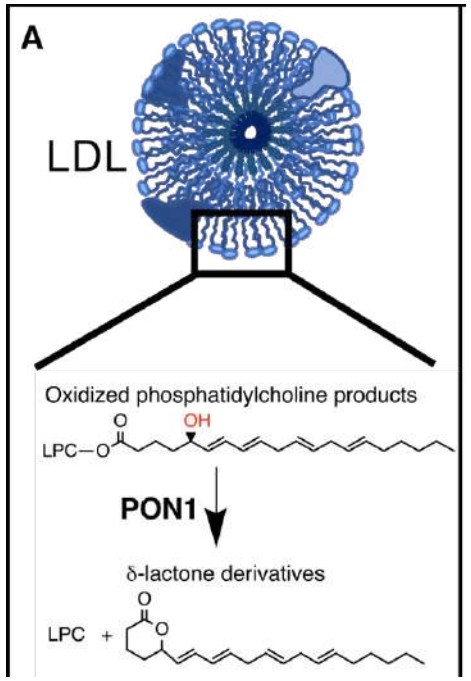


What does *Paraoxonase 1* (*Pon1*) do?

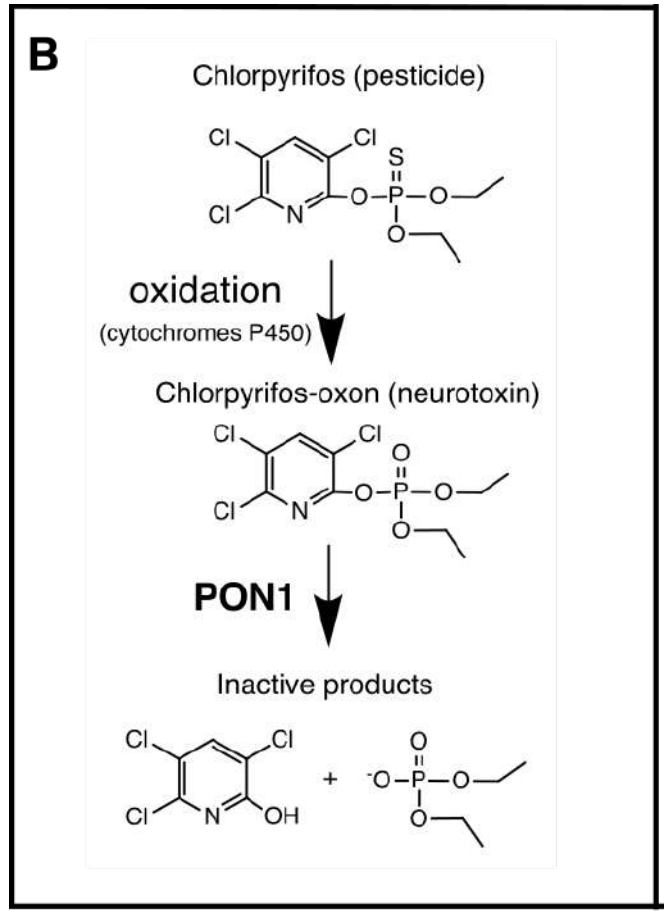
PON1 alleles associated with risk of atherosclerosis.



Thought to mitigate damage to oxidized lipids in bloodstream



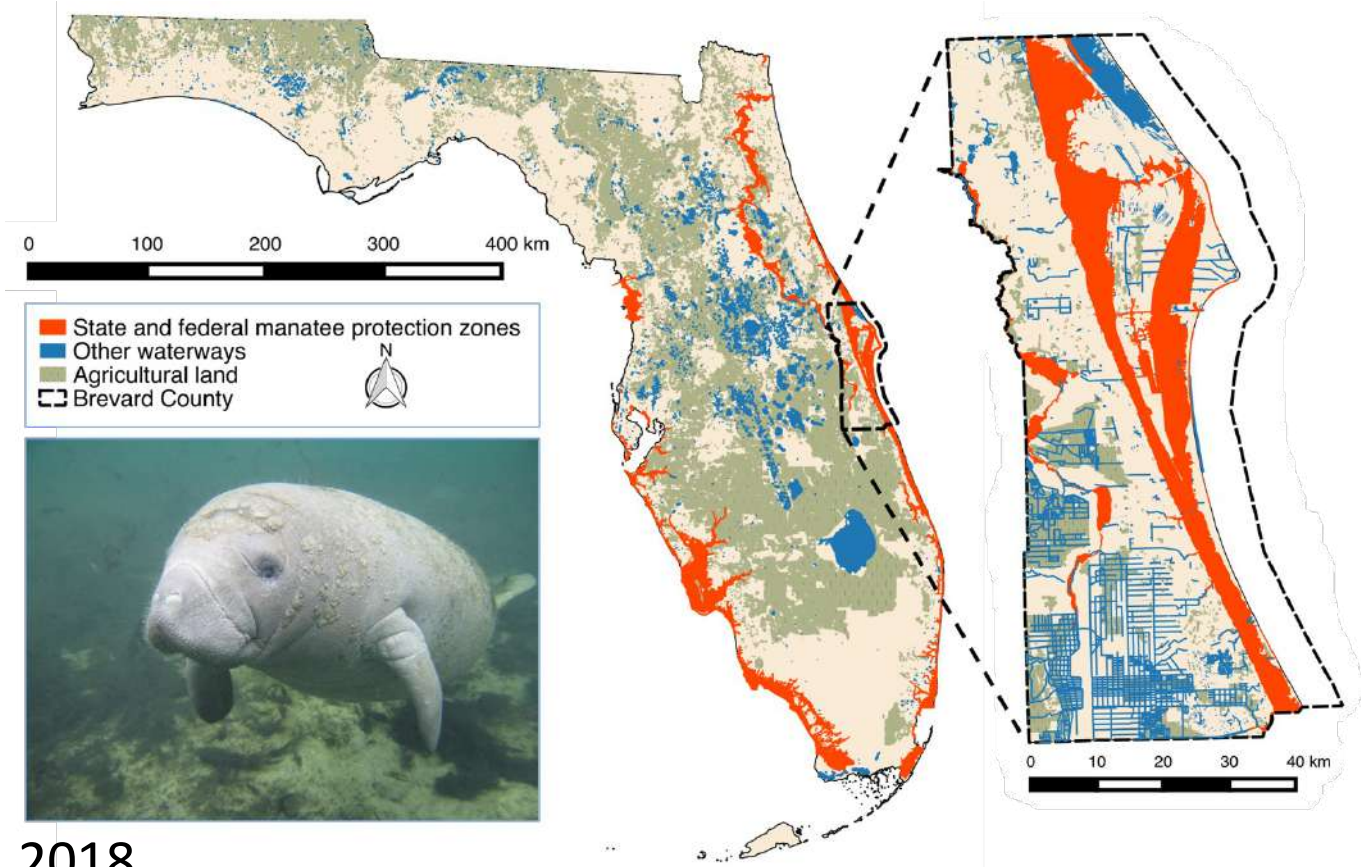
Breaks down organophosphate pesticides



What are the potential implications for loss of Pon1 function in aquatic mammals?

Manatees are potentially at risk from agricultural run-off.

- Acute exposure is neurotoxic and kills pon1-/- mice
- Chronic exposure through bioloading



Meyer et al. *Science.* 2018

Photo: Robert K Bonde, USGS | Maps: Stacy E Woods, Johns Hopkins University

EVOLUTION

Ancient convergent losses of *Paraoxonase 1* yield potential risks for modern marine mammals

Wynn K. Meyer[1], Jerrica Jamison[2], Rebecca Richter[3], Stacy E. Woods[4]*, Raghavendran Partha[1], Amanda Kowalczyk[1], Charles Kronk[2], Maria Chikina[1], Robert K. Bonde[5], Daniel E. Crocker[6], Joseph Gaspard[7], Janet M. Lanyon[8], Judit Marsillach[3], Clement E. Furlong[3,9], Nathan L. Clark[1,10]†

Pon1 loss was likely Adaptive

Pon1 could have been lost because it is no longer important

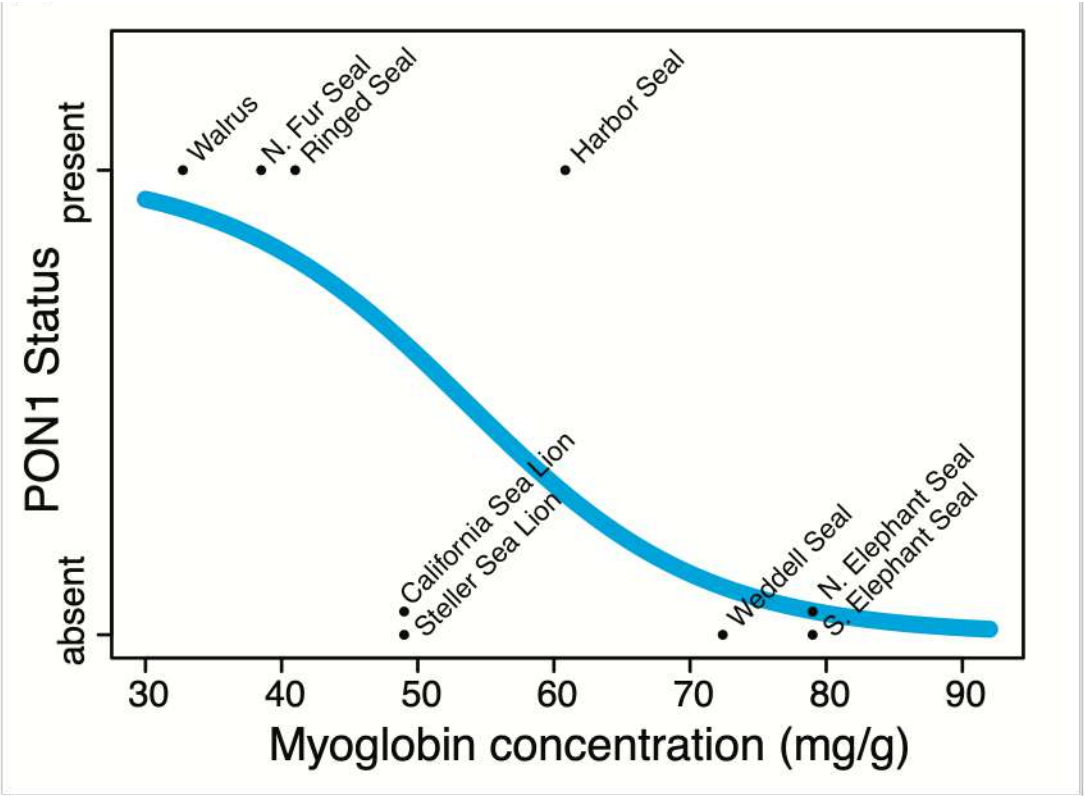
However, we hypothesize its loss was adaptive because

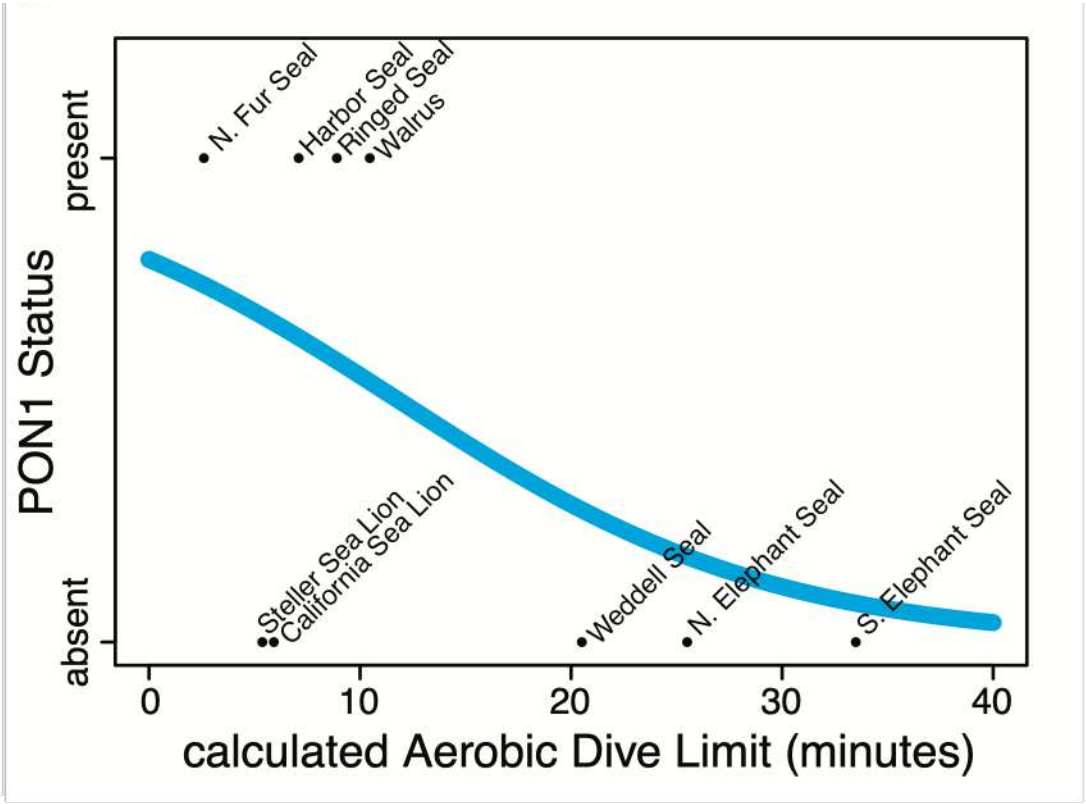
- It is the most consistently lost in aquatic mammals. More so than genes lacking constraint, such as olfactory receptors.
- It is lost very quickly, first by loss of expression, then pseudogenization.

We believe it was lost as an adaptation to diving.

The best predictors of *Pon1* loss are diving traits

We test hypotheses in pinnipeds because they are a large set of species with active and inactive *Pon1*

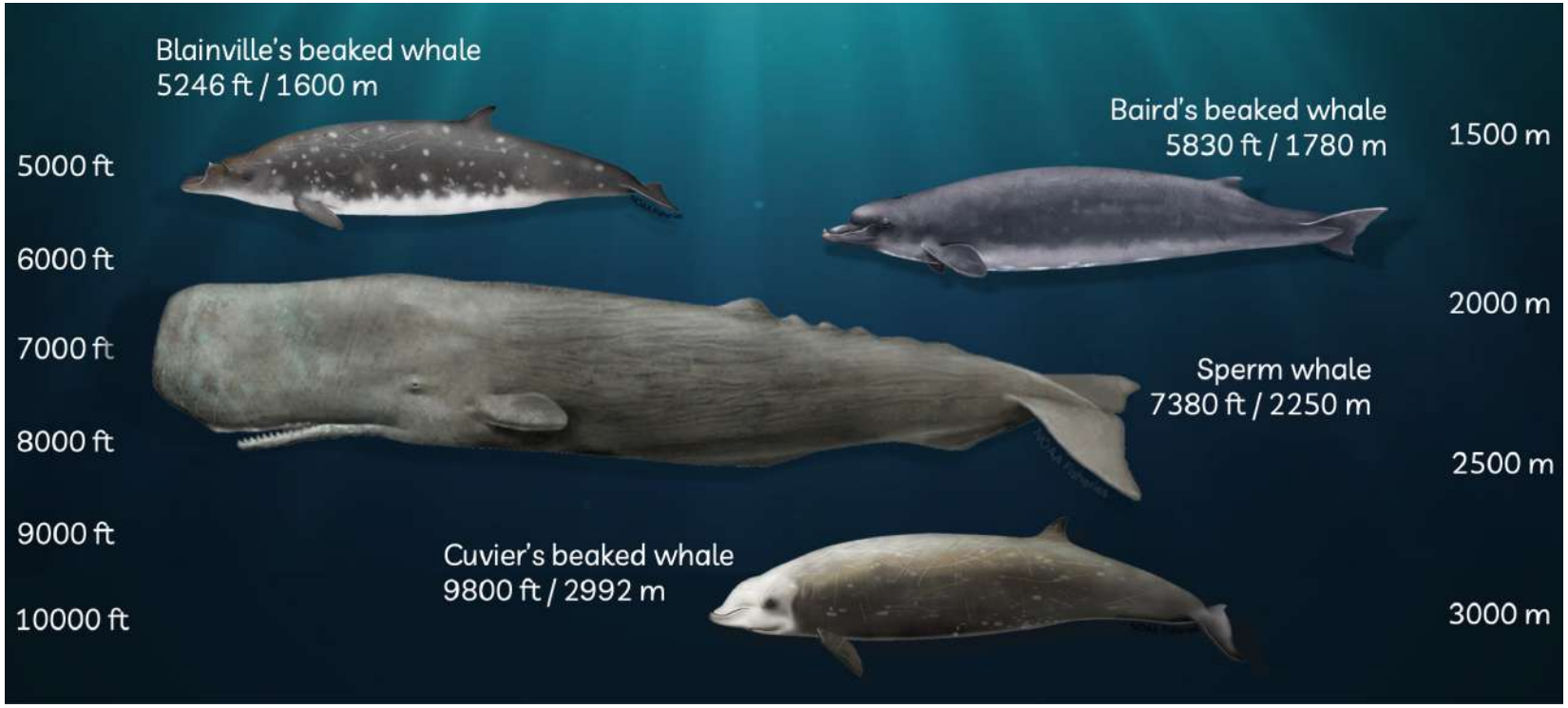




Phylogenetic logistic regression P=0.05

Diving-related ischemia creates oxidative stress

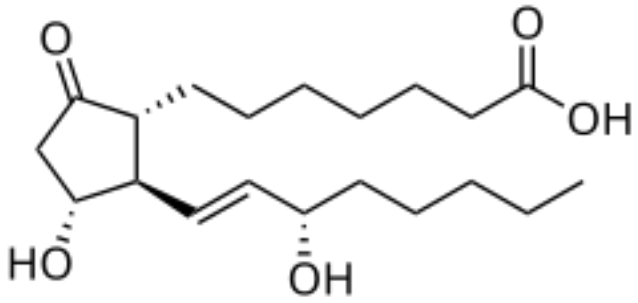
- Oxygen depletion and rapid reperfusion triggers inflammation that leads to production of oxidative species
- Many diving species have constitutively upregulated enzymes protecting against oxidation, such as Superoxide dismutase and Catalase.



Pon1 Hypothesis

Shutting down Pon1 dampens the inflammatory response to acute hypoxia/ischemia

- Pon1 is capable of producing certain eicosanoids (signaling lipids)
- Eicosanoids modulate inflammation



Working with Pittsburgh Vascular Institute

- Stephen Chan's lab has Pon1 knock out mouse
- Hypoxic chambers and ischemia tests
- Vascular inflammatory markers

PITTSBURGH HEART, LUNG, AND BLOOD VMI VASCULAR MEDICINE INSTITUTE
VITALANT • HCWP • UNIVERSITY OF PITTSBURGH • UPMC

Gene family expansion and contraction

Convergent genetic patterns in high-altitude mammals

Which genes are pseudogenized?

- 17 high-altitude specialist species
- 120 low-altitude
- Identify genes preferentially lost at altitude
- Used BayesTraits (Pagel)

Allie Graham

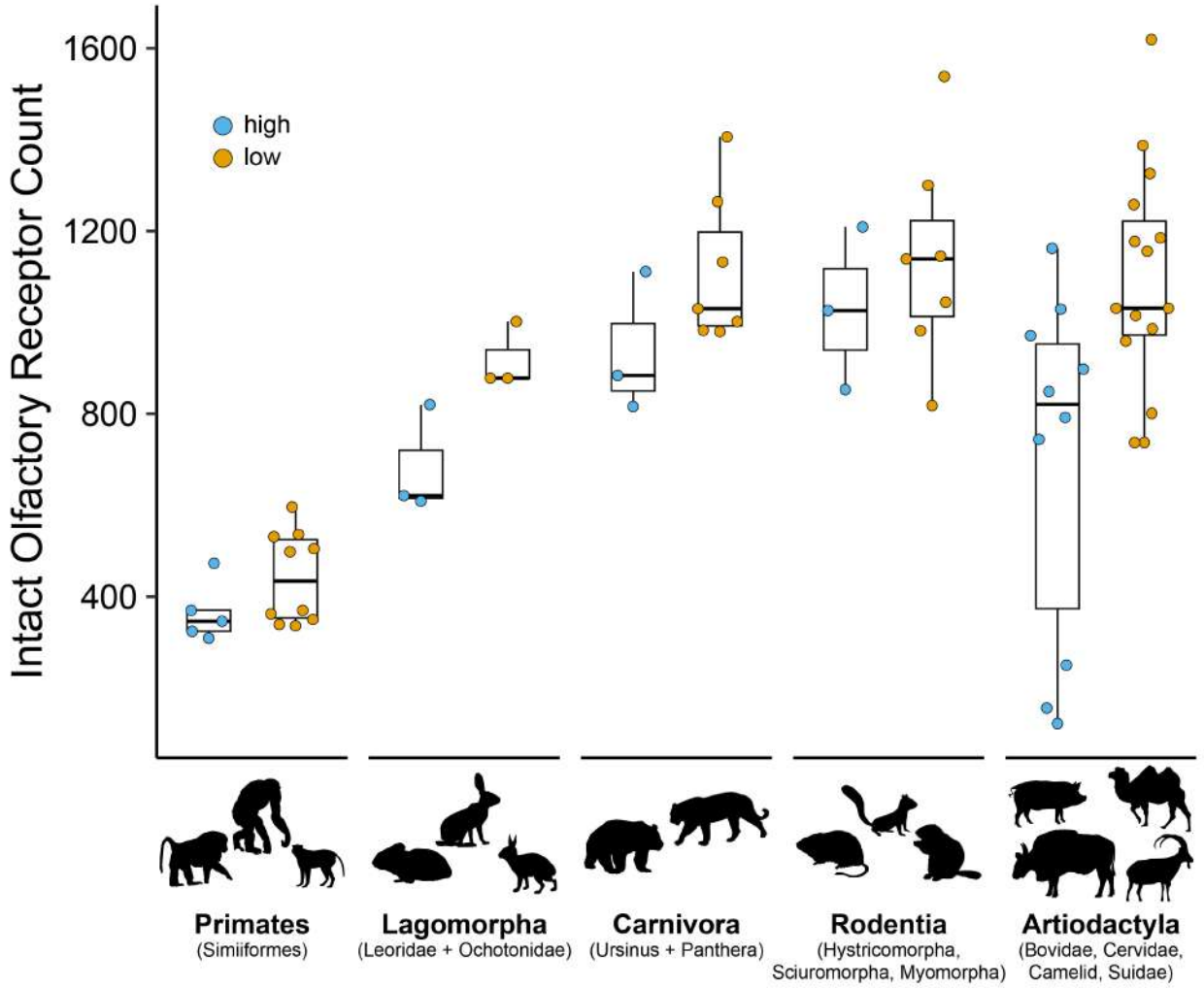
Convergent reduction of olfactory genes in mammalian species at high altitude

New pseudogenes in high altitude mammals were highly enriched for ORs.

Sensory Perception of Smell

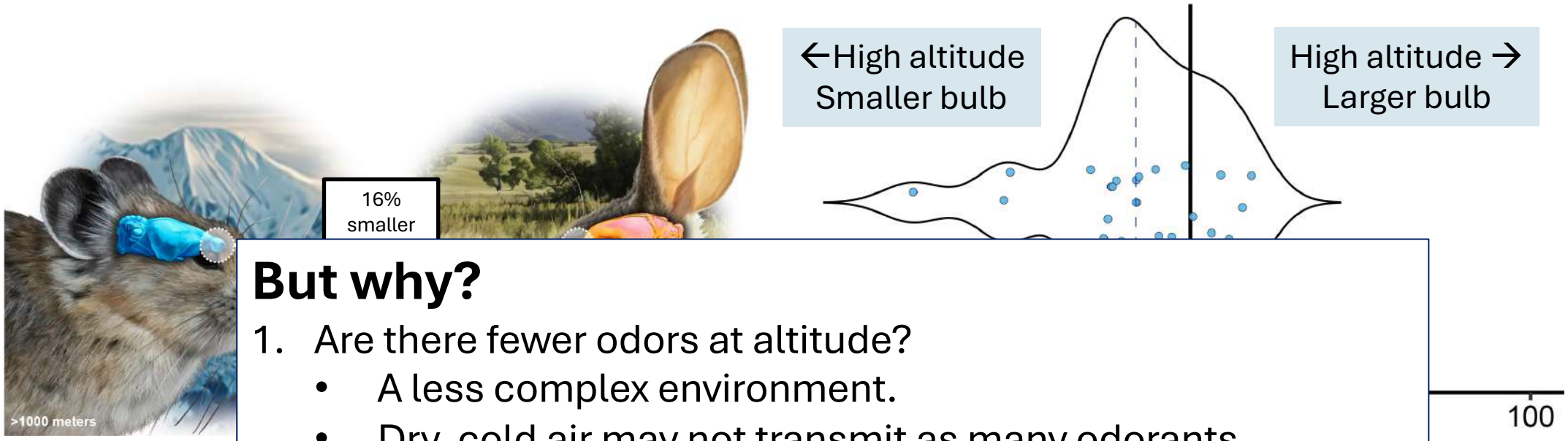
FDR = 7.7×10^{-36}

~23% reduction in OR genes



Olfactory bulb size is reduced in species at altitude

Brain endocasts from 21 Low- and High-altitude “sister”



Wilcox Paired,

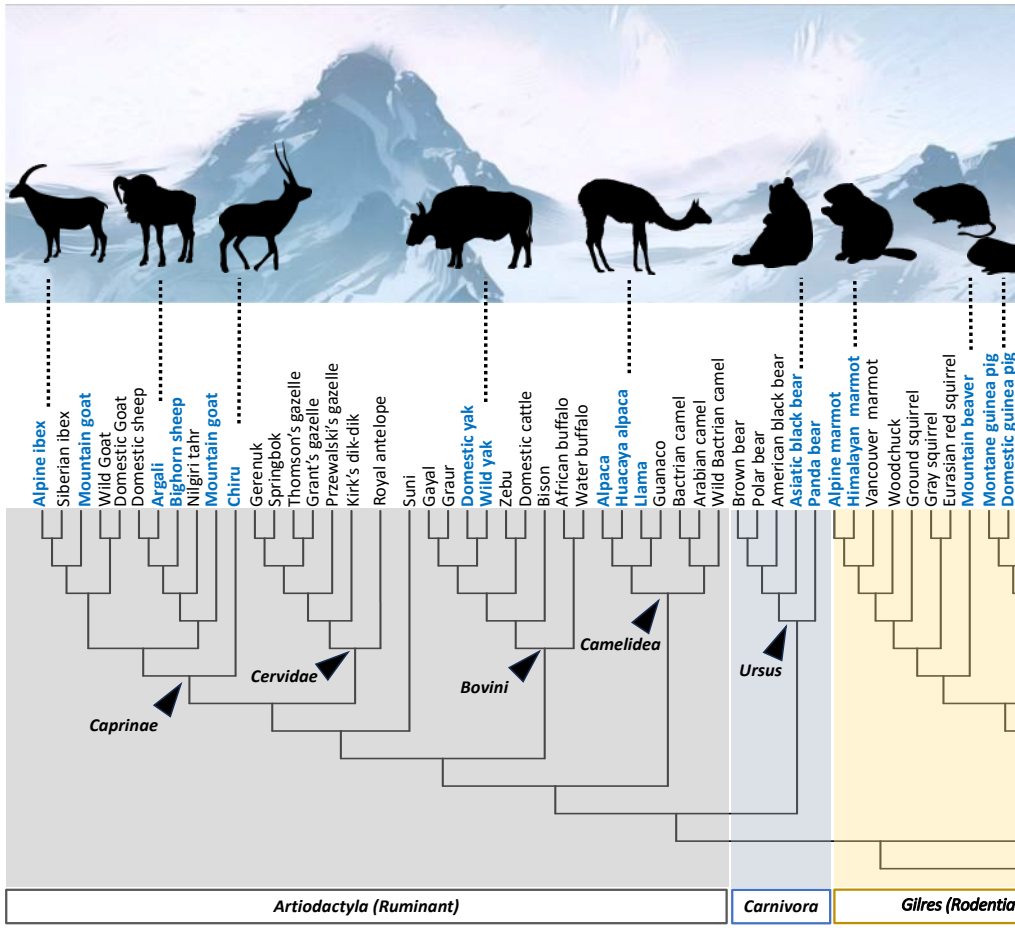
But why?

1. Are there fewer odors at altitude?
   - A less complex environment.
   - Dry, cold air may not transmit as many odorants.

2. Do olfactory anatomy and physiology work poorly at altitude?

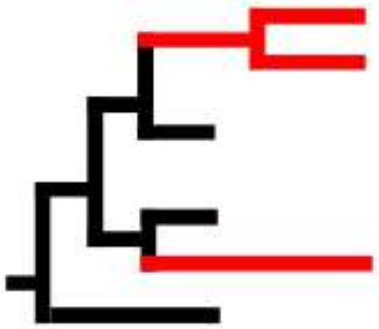
Gene Family Methods

- Scanning annotated genes from all species for Olfactory Receptor (OR) homology.
  - Hidden Markov Model (HMM) for OR family
  - Used **HMMer** by Sean Eddy
- Tested for association between high altitude and OR copy number using Phylogenetic Generalized Least Squares (PGLS) models
  - Used R package “phytools”
  - Revell and Harmon’s excellent book
    - http://www.phytools.org/Rbook/
- CAFE - tool for analyzing gene family evolution
  - https://github.com/hahnlab/CAFE



Rates-based PhyloG2P

Asking if a gene's branch-specific evolutionary rates correlate with the evolution of a trait on those branches



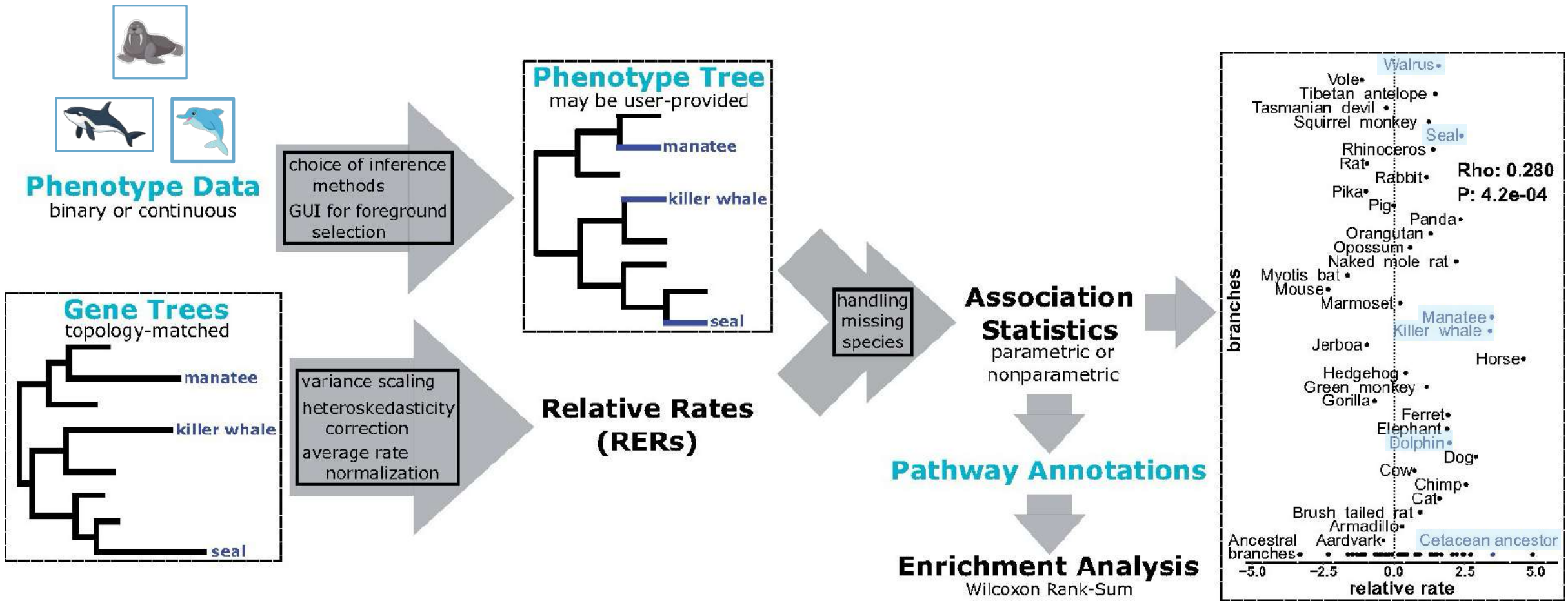
Rate acceleration correlated with convergent trait

<u>Computational Methods</u>

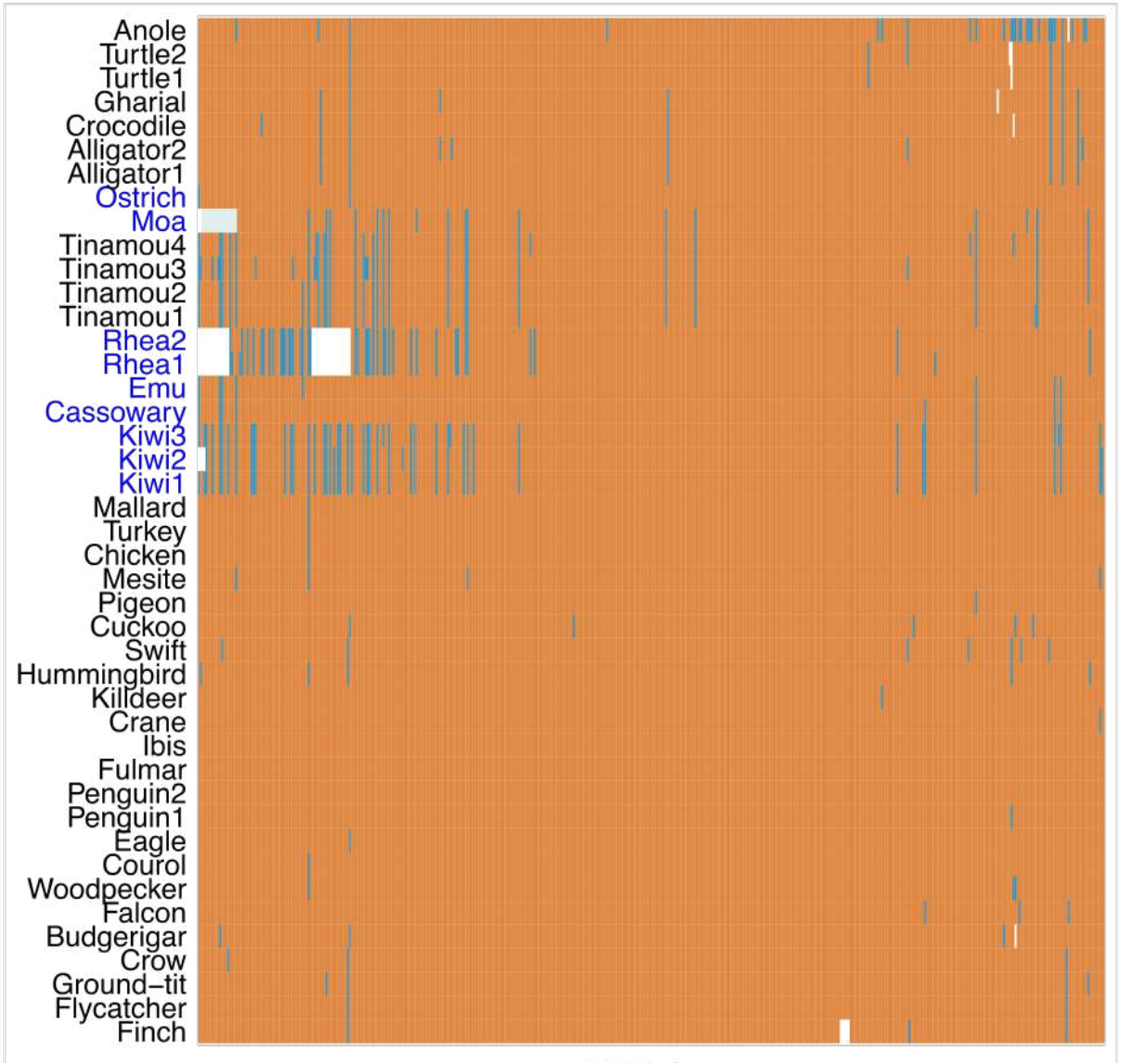
- Coevol
- ForwardGenomics
- RERconverge
- PhyloACC

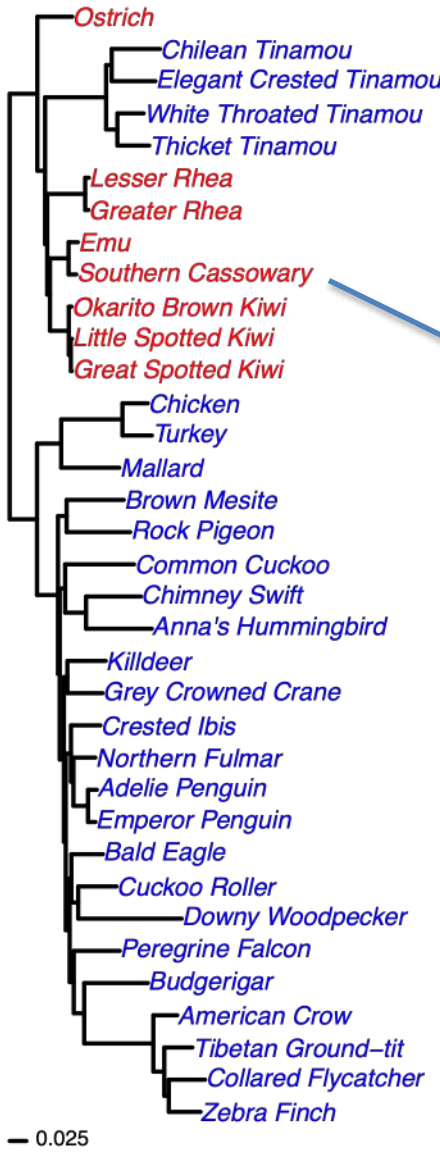
RERConverge Background

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

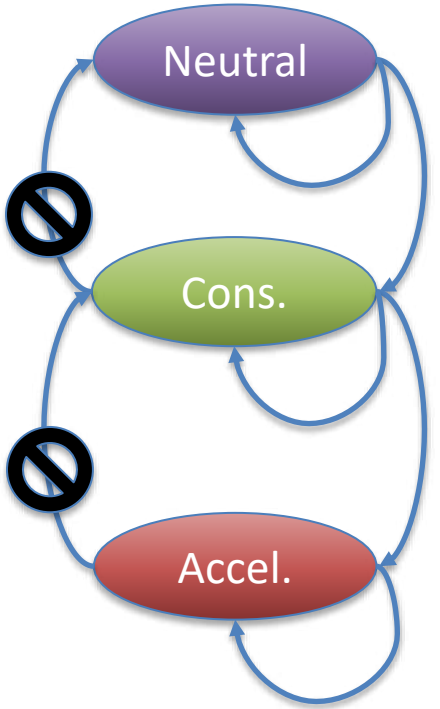


phyloAcc: detecting convergent acceleration in CNEEs

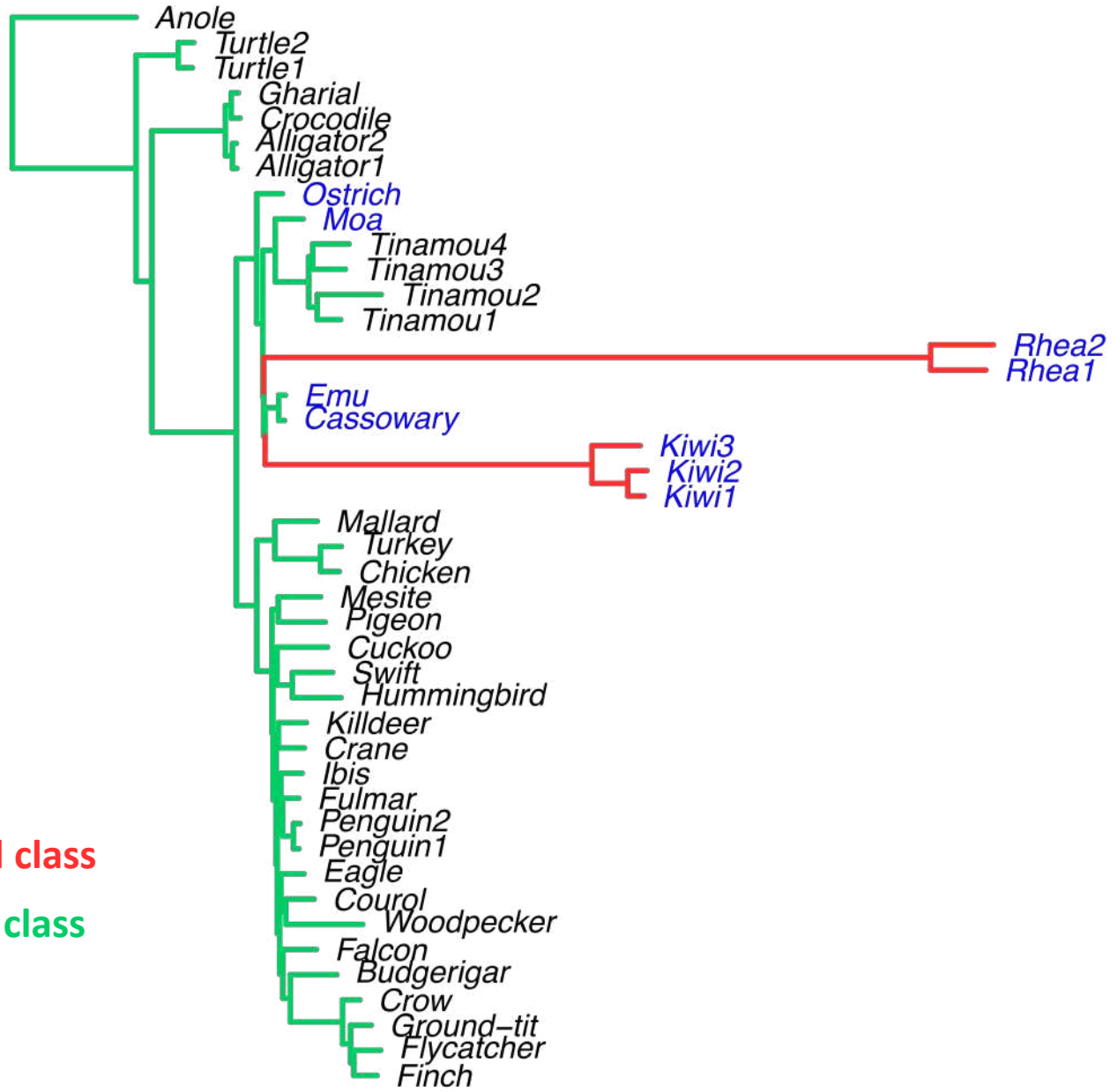




Hidden states for each branch



phyloAcc: detecting convergent acceleration in CNEEs



Accelerated class

Conserved class

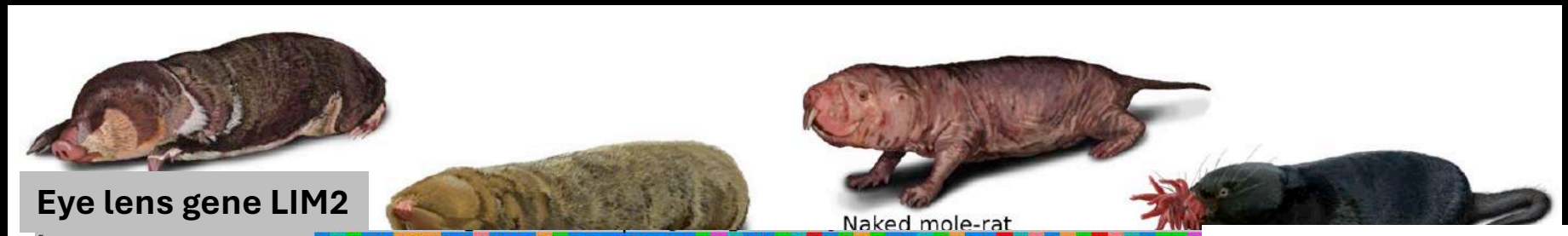
Zhirui Hu

Scott Edwards

Gregg Thomas

Tim Sackton

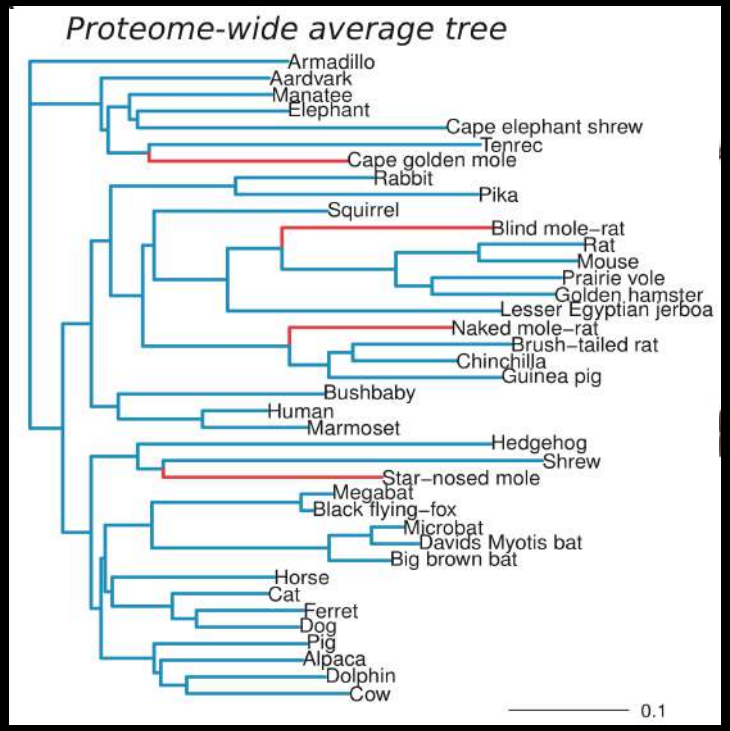
Convergent evolution in the subterranean niche

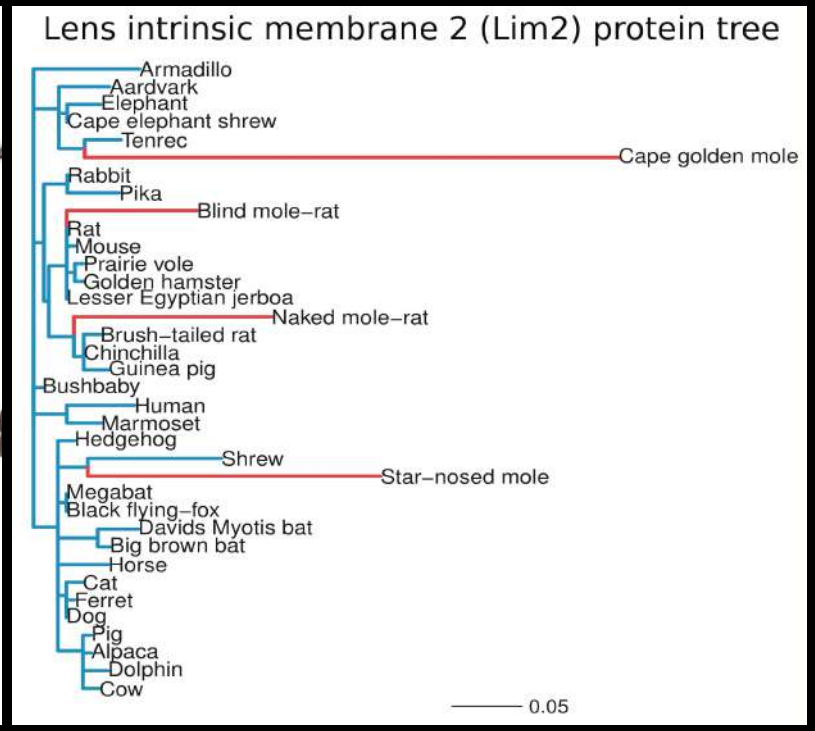


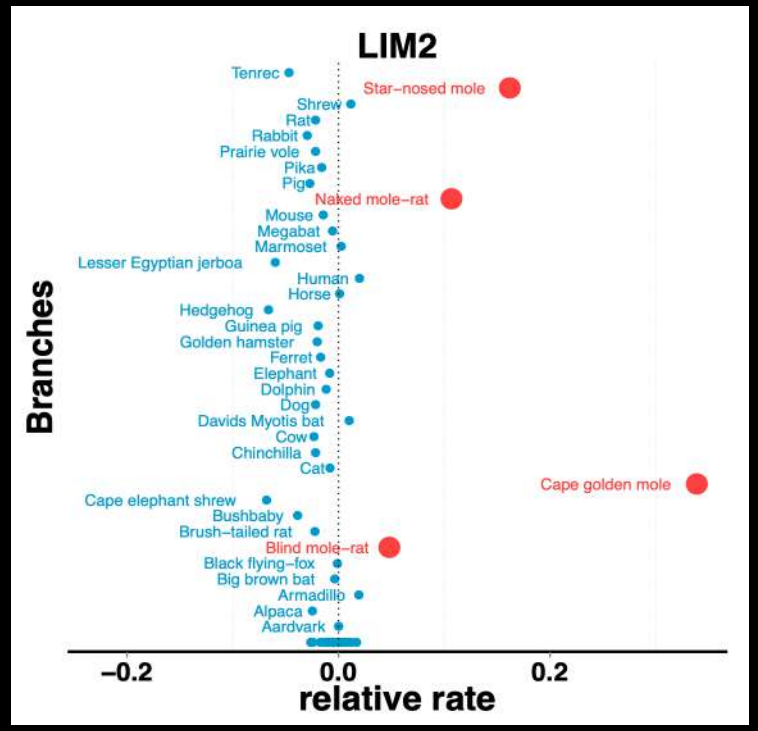
Eye lens gene LIM2

```
human              MYSFMGGGLFCAWVGTILLVVAMATDHWMQYRLSGSFAHQGLWRYCLGNKCYLQTD
rabbit             MYSFMGGGLFCAWVGTILLVVATATDHWMQYRLSGSFAHQGLWRYCLGSKCFLQTE
lion               MYSFMGGGLFCAWVGTILLVVATATDHWMQYRLSGSFAHQGLWRYCLGNKCYLQTE
Cape golden mole   RYSFISGSLFCVWVG-------TATDHWMQFRLLGFFAHRRLWQCYRGNKCYLLIE
star-nosed mole    MQGLGGGGLLCACAGAVLLVGATATDHWMQYRLARSFAHQGLWRYCLGSRCFLQTE
```

RER > 0, evolving faster
RER = 0, evolving at expected rate
RER < 0, evolving slower





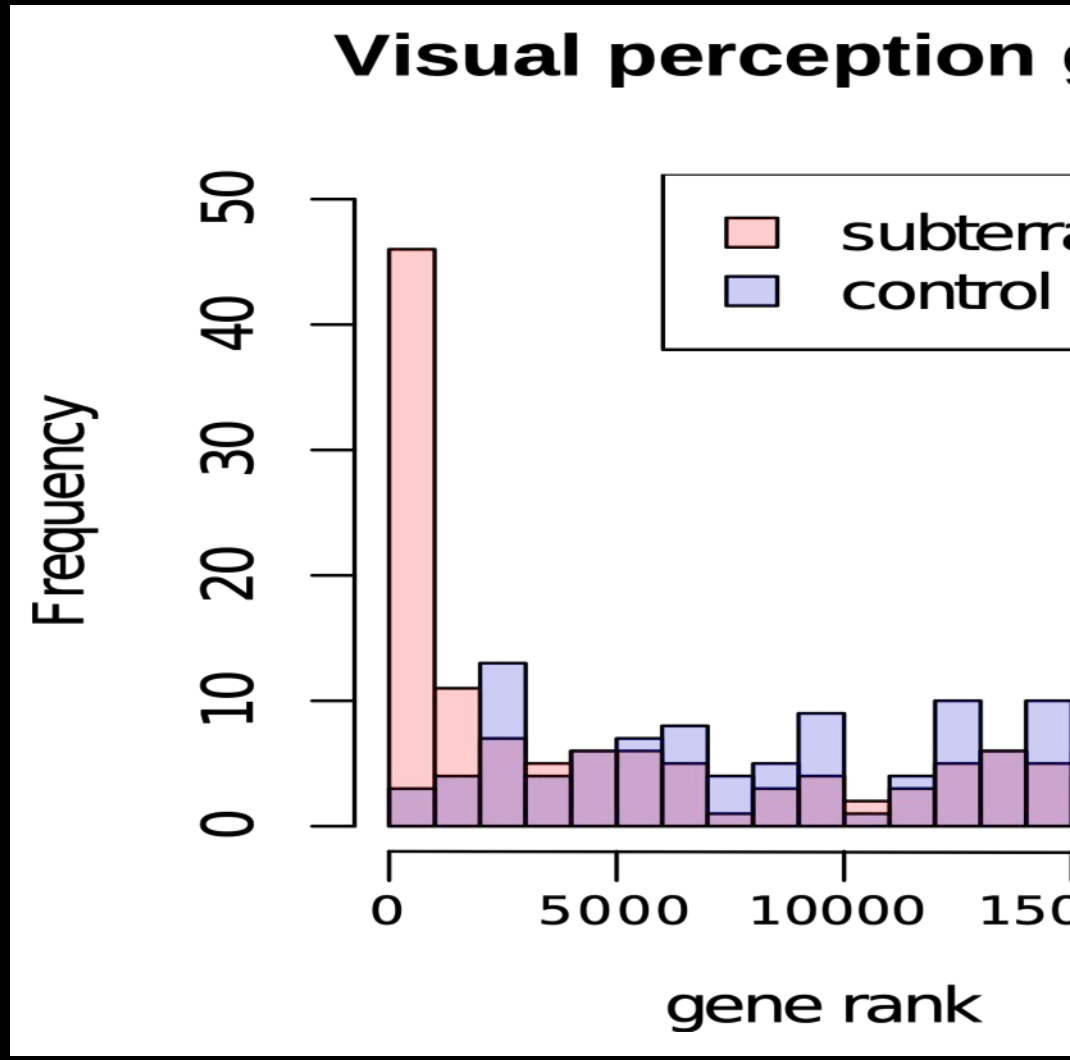


Red = foreground species Blue = background species

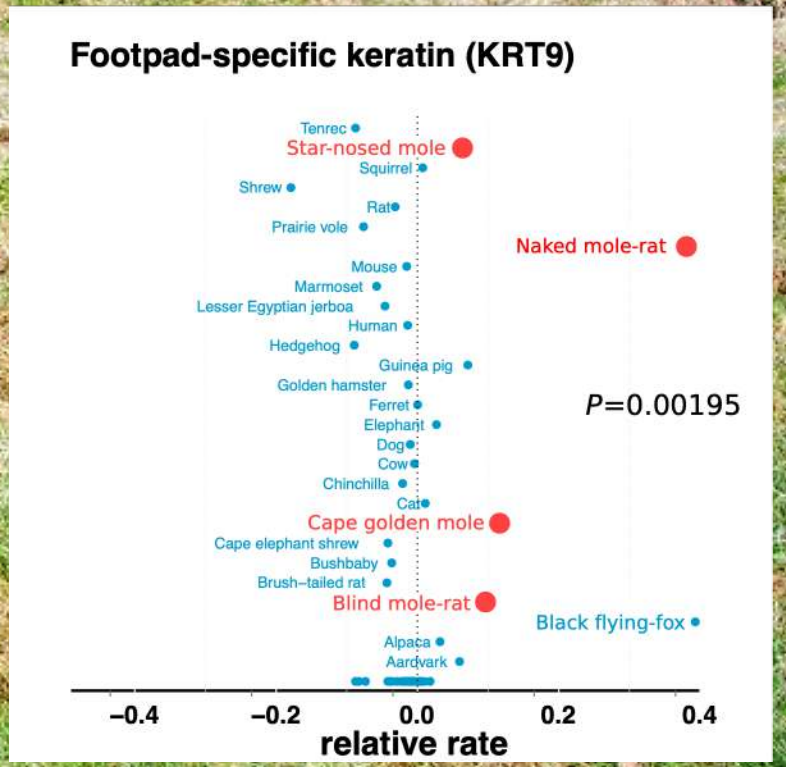
Partha et al. *eLife* 2017

Mole-acceleration readily identifies ocular genes

	rank_slp	gene_list	Rho	N	P	p.adj
CRYBA1	1	yes	0.5057838	304	3.777225e-21	7.379187e-17
CRYAA2	2	no	0.4804796	250	7.614028e-16	6.577785e-12
CRYAA	3	yes	0.4769891	252	1.010102e-15	6.577785e-12
LIM2	4	yes	0.4510504	280	1.942252e-15	9.485957e-12
MIP	5	yes	0.3908291	340	7.456527e-14	2.913414e-10
PDE6B	6	no	0.3219138	496	2.015540e-13	6.036333e-10
CRYBB2	7	yes	0.4321436	263	2.162896e-13	6.036333e-10
GNAT2	8	yes	0.3606896	360	1.679258e-12	4.100748e-09
GRM6	9	yes	0.2994954	491	1.238885e-11	2.689206e-08
CRYBB1	10	yes	0.3213292	420	1.526337e-11	2.981852e-08
CRYGC	11	yes	0.3037436	396	6.754765e-10	1.189845e-06
CRYBA2	12	yes	0.3355903	320	7.308630e-10	1.189845e-06
CRYGS	13	no	0.3262378	337	8.497429e-10	1.276968e-06
CRYGD	14	yes	0.2957452	398	1.780587e-09	2.484682e-06
BFSP2	15	yes	0.2686057	476	2.606225e-09	3.394348e-06
GJA8	16	yes	0.2603423	490	4.934390e-09	6.024891e-06
LENEP	17	yes	0.3152818	328	5.290935e-09	6.080218e-06
SLC1A7	18	yes	0.2600910	472	9.772791e-09	1.060674e-05
ABCA4	19	no	0.2397980	547	1.357747e-08	1.396050e-05
CRYGA	20	yes	0.2835281	380	1.859315e-08	1.816179e-05
CRYBB3	21	yes	0.2948070	340	3.032531e-08	2.821120e-05
RS1	22	yes	0.2894511	350	3.502801e-08	3.110487e-05



Tunneling adaptations?
Skin-related proteins are accelerated in moles



Keratin 9 loss in humans and mice leads to thickening of the palms and footpads, **palmoplantar keratoderma**.

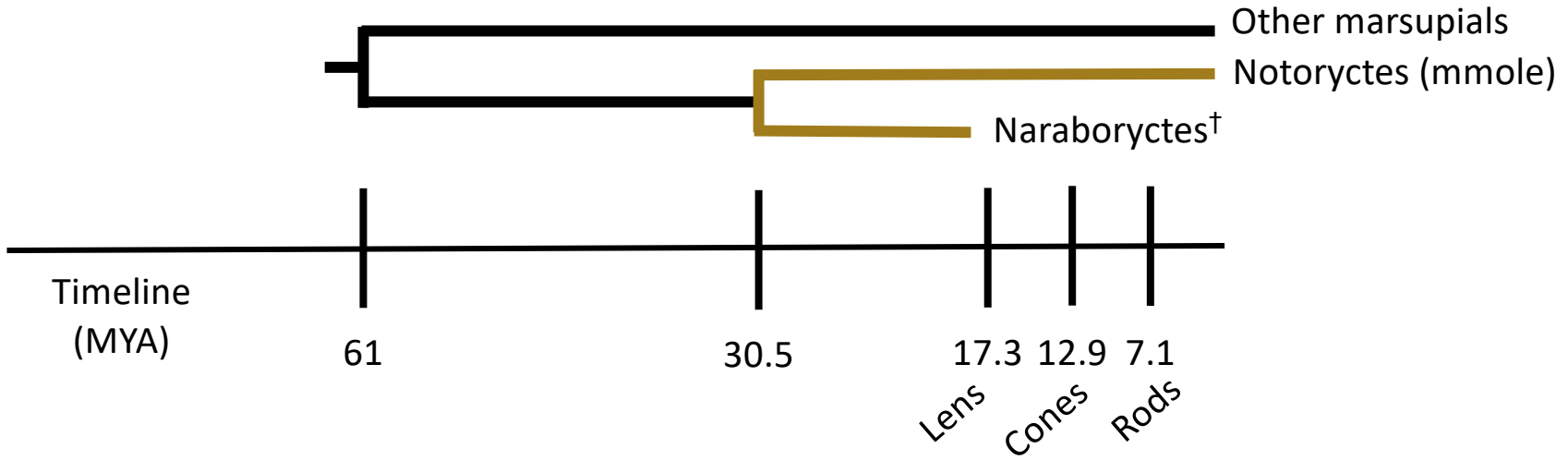
Keratin 9 has stop codons in 3 subterranean lineages.

Marsupial Mole (*Notoryctes typhlops*) genome project

OMG
OZ MAMMALS GENOMICS

Eye components were released from constraint millions of years apart.

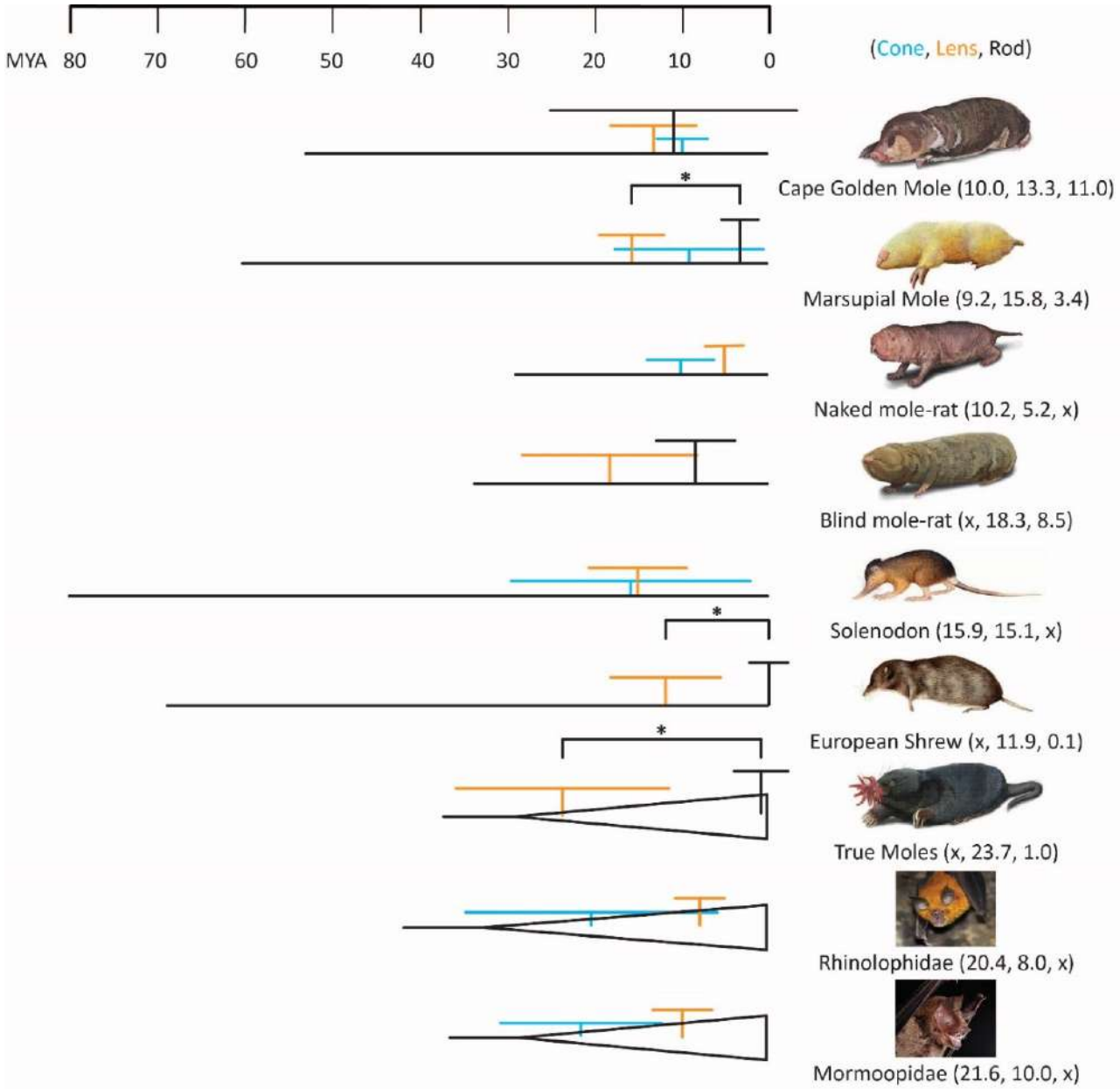
Cell Type	Total # Genes	Average Mmole d_N/d_S	Average marsupial d_N/d_S	Difference	Release from constraint (MYA)
Lens Fiber	15	0.345	0.086	0.259	17.3
Cone Photoreceptor	6	0.307	0.120	0.187	12.9
Rod Photoreceptor	15	0.248	0.149	0.099	7.1



Sarah Lucas

Poorly sighted species lost constraint ocular abilities variably

Sarah Lucas



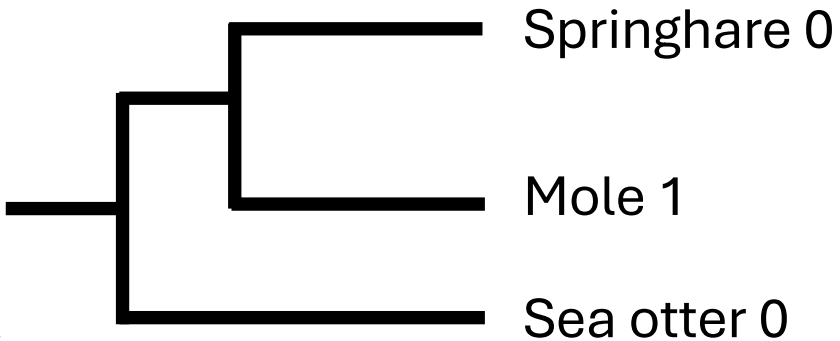
Can we improve power with better trait descriptions?

Binary vs continuous trait analysis

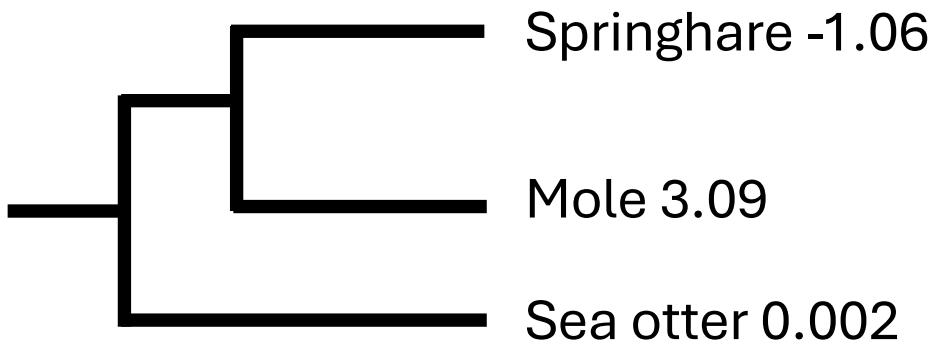
Binary

0 = not poor eyesight

1 = poor eyesight

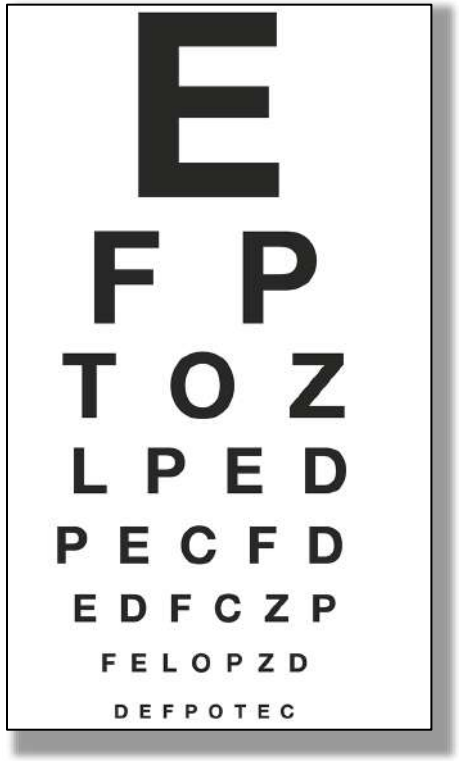


Continuous

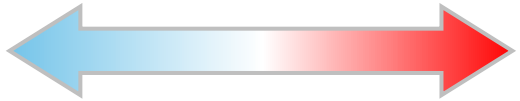


Courtney Charlesworth

Binary vs continuous trait analysis



Continuous



Courtney Charlesworth

Trait Scores are a continuous measure reflecting constraint on a process

Ocular Trait Scores

Calculate mean rate of the trait gene list for every species

Ocular Trait gene list

65 eye-specific genes

	Eastern mole	Sea otter	Springhare
CRYBB3	10.31	-0.52	-2.86
GNAT2	2.47	0.81	-0.53
GRM6	2.39	0.08	-0.62
RHO	2.09	-0.48	-0.95

Sarah Lucas

RER > 0 = faster rate of evolution
RER = 0 expected rate
RER < 0 = slower rate of evolution

Now, several mammalian lineages with regressive eyes score highly with continuous trait scores

Indus river dolphin

Mole rat

Mole rats

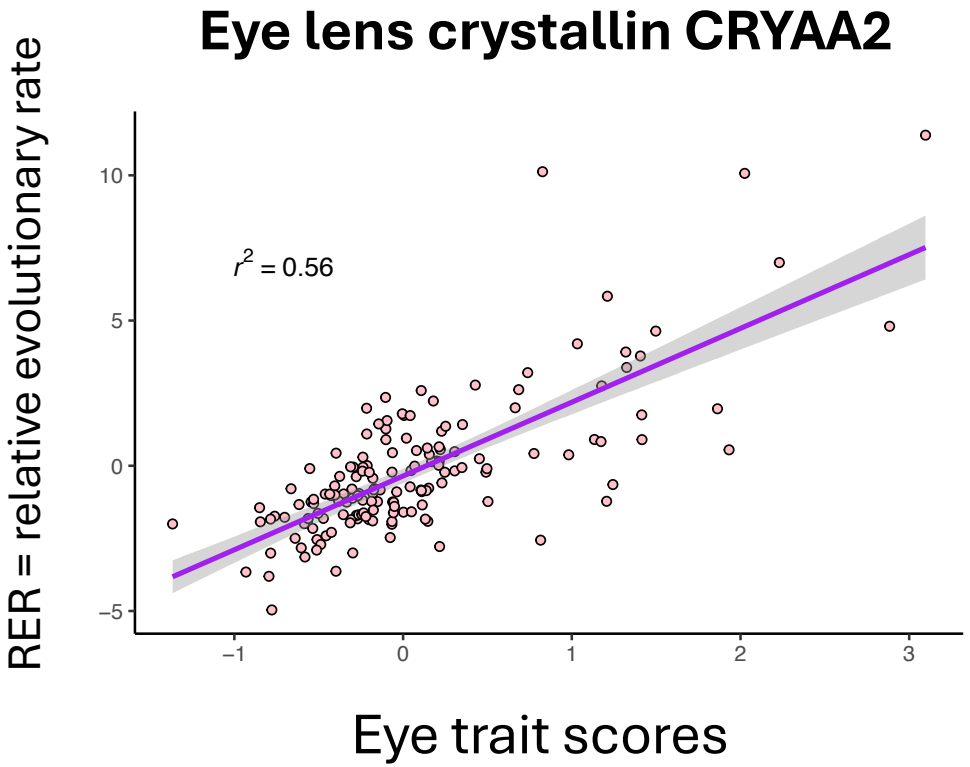
Eye trait score

Armadillo,
Tenrec

Bats

Moles,
Solenodon

Known eye genes correlate with the trait scores, which is, of course, expected

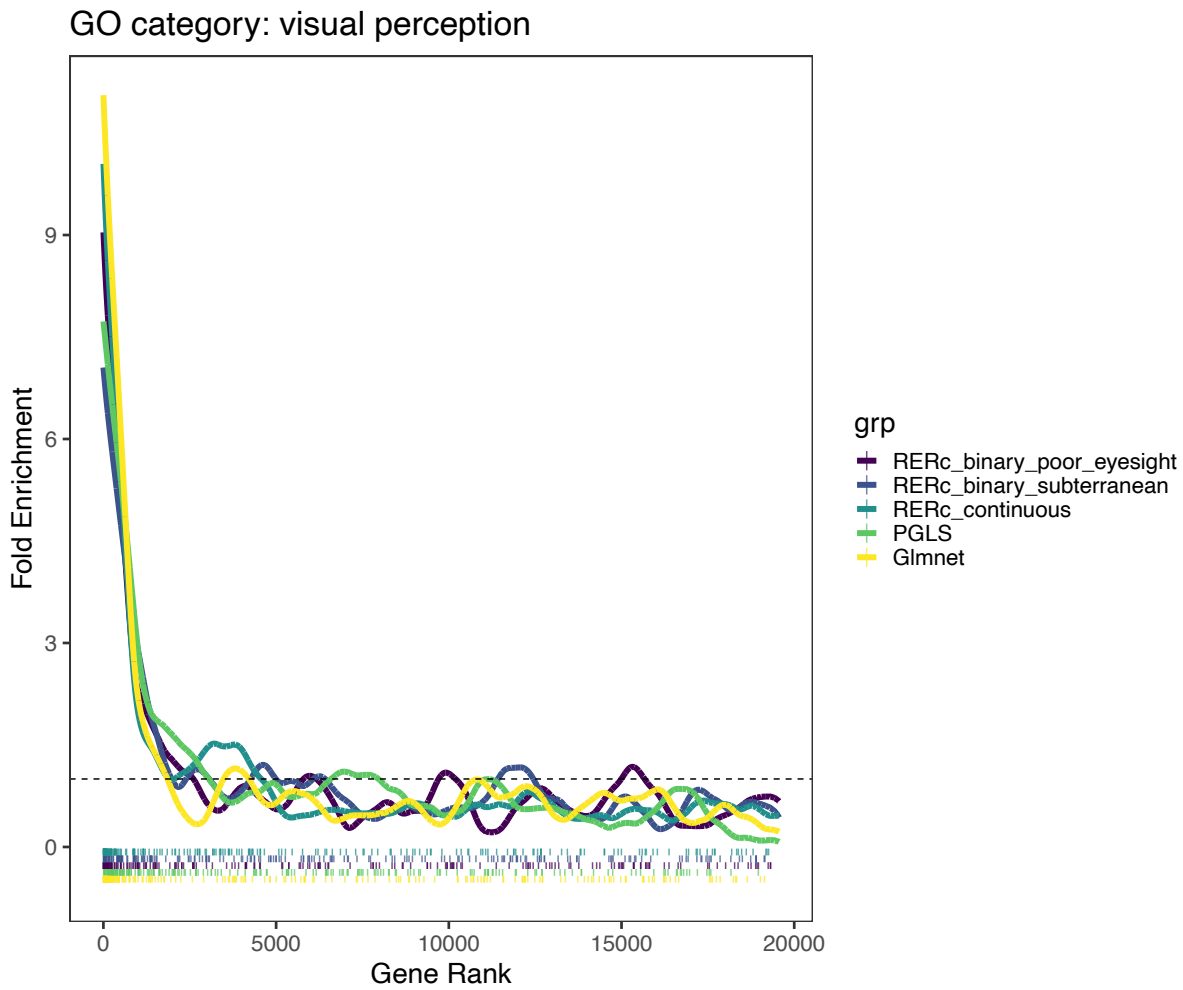


Infer new ocular genes and regulatory regions→

Each method enriches for the Visual Perception GO term

Trait Score-based methods are large improvement.

Method	Enrichment in top 100 genes
PGLS on TraitScores	51x
RERcon binary on TraitScores	35x
GLMNET – training gene set	30x
RERcon continuous on TraitScores	29x
RERc binary “Moles”	19x

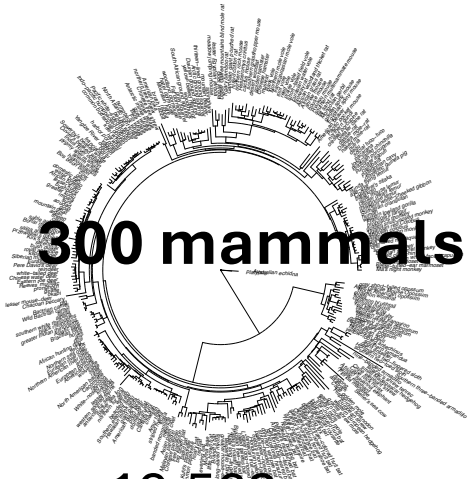


←Accelerated in poorly sighted species

Inference of novel ocular genes and validation via the Optokinetic response in zebrafish

gene	full_name	function	eye_expr.	expr._other_tissues	zebrafish_ortholog
		apoptosis	lens	many (broad expr.)	yes, 1
		GABA receptor	retina	none (eye-specific)	yes, 2
		likely cell adhesion	none	mostly brain-specific	yes, 1
		unknown, paralog LACTB = lipid metabolism	retina	none (eye-specific)	yes, 2
		unknown	retina	none (eye-specific)	no

Analysis



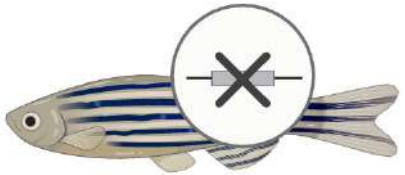
300 mammals

- 19,562 genes
- 1M+ conserved noncoding elements

Candidates

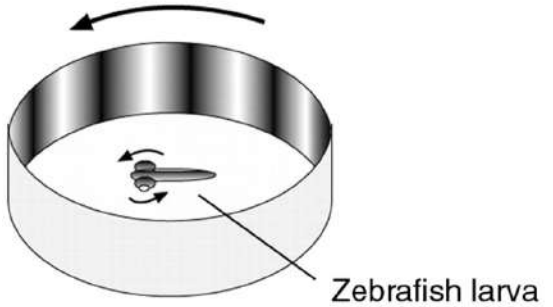
Done

Knockout in zebrafish



In progress

Assess phenotype

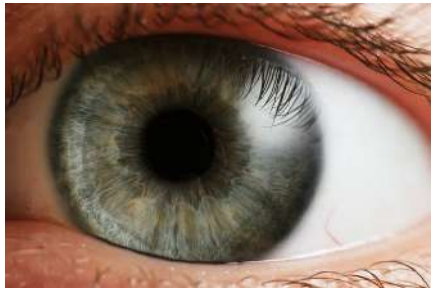


Biorender; Roeser and Baier 2003

Can mole-accelerated regulatory regions guide us to new regions involved in congenital eye disease?

Causal mutations not found in ~50% congenital abnormality cases.

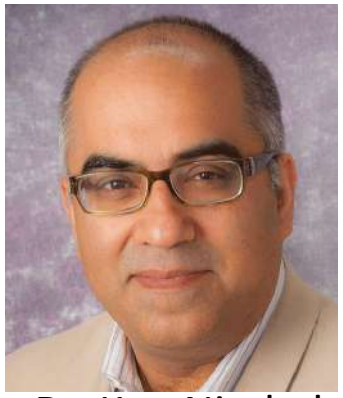
- Diagnostic panels only examine protein-coding regions.
- Could **regulatory regions** hold their causal mutations?



We sequence patient and parent genomes for difficult cases.

We constructed a GATK-best practices variant calling pipeline for SNPs and indels

Children's Hospital of Pittsburgh | of UPMC



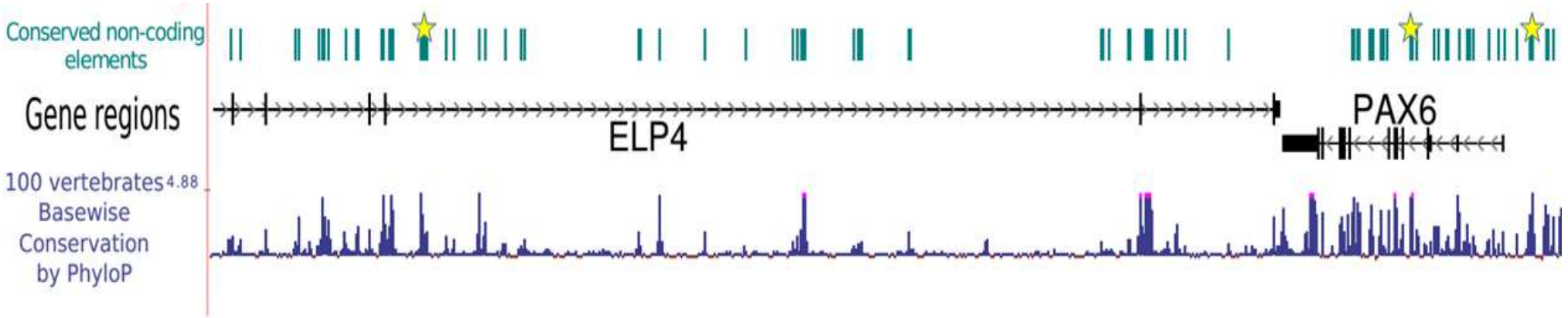
Dr. Ken Nischal

First 2 patients yielded unrecognized *de novo* mutations in:

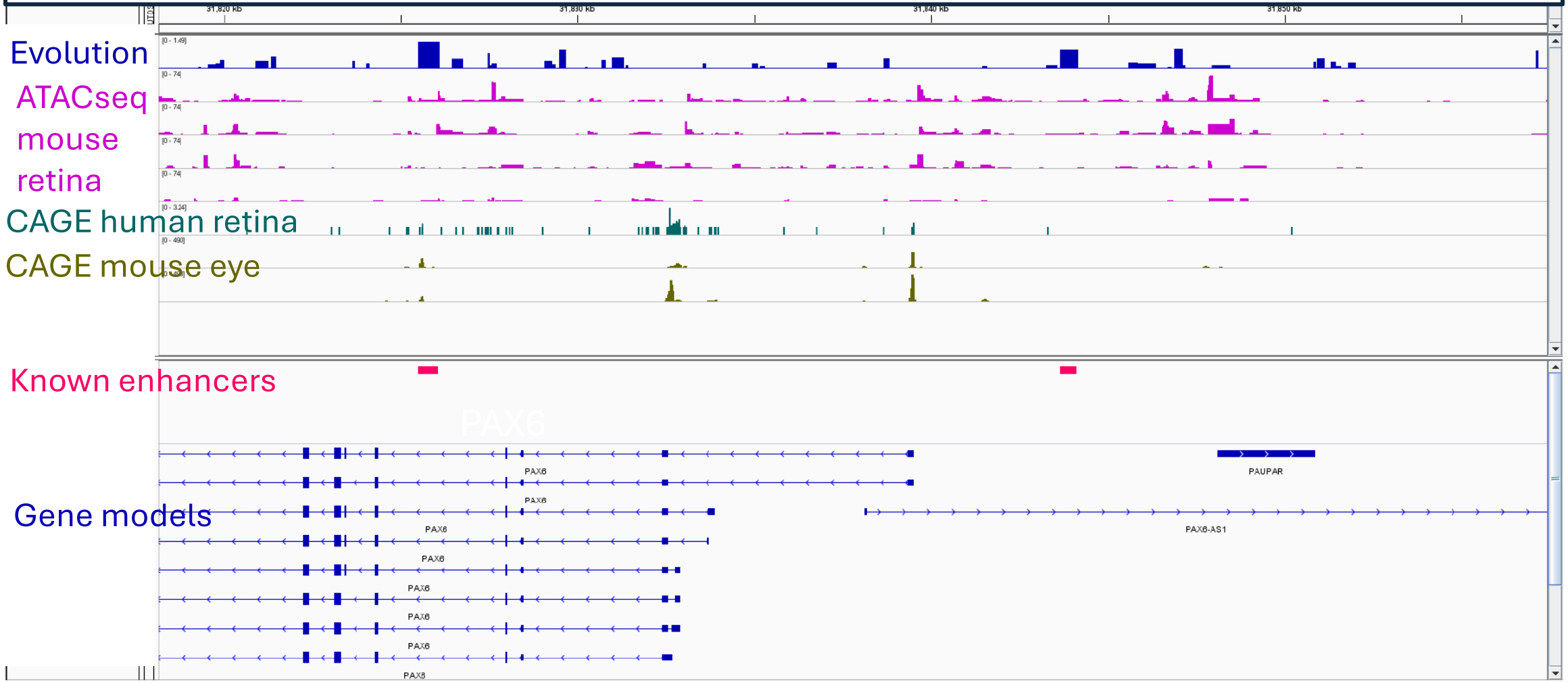
gap junction A8 (***GJA8***) for congenital aphakia (absent lens)

crystallin beta A1 (***CRYBA1***) for bilateral cataracts

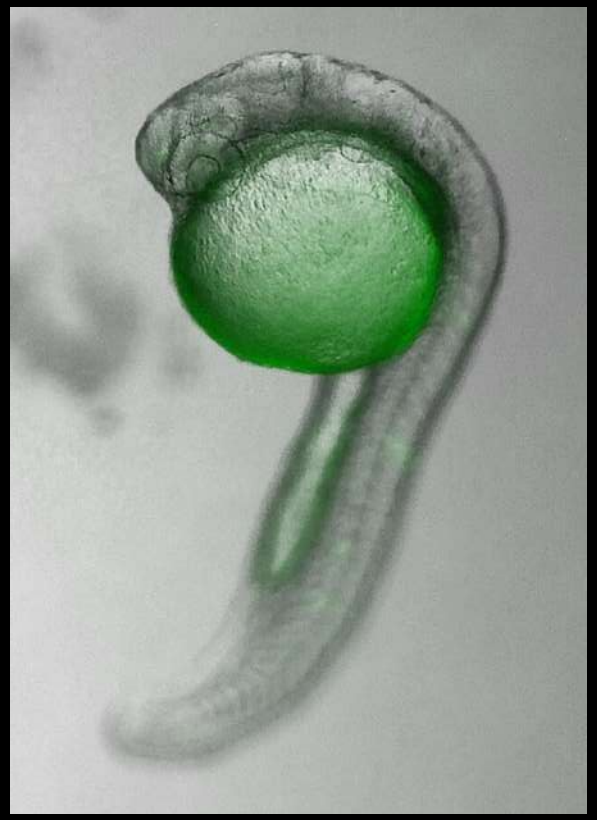
Can we predict eye-specific regulatory elements across the genome?



Convergent Evolution locates known enhancers missed by experimental methods



Do high-scoring regions drive vision-related expression in embryonic zebrafish?



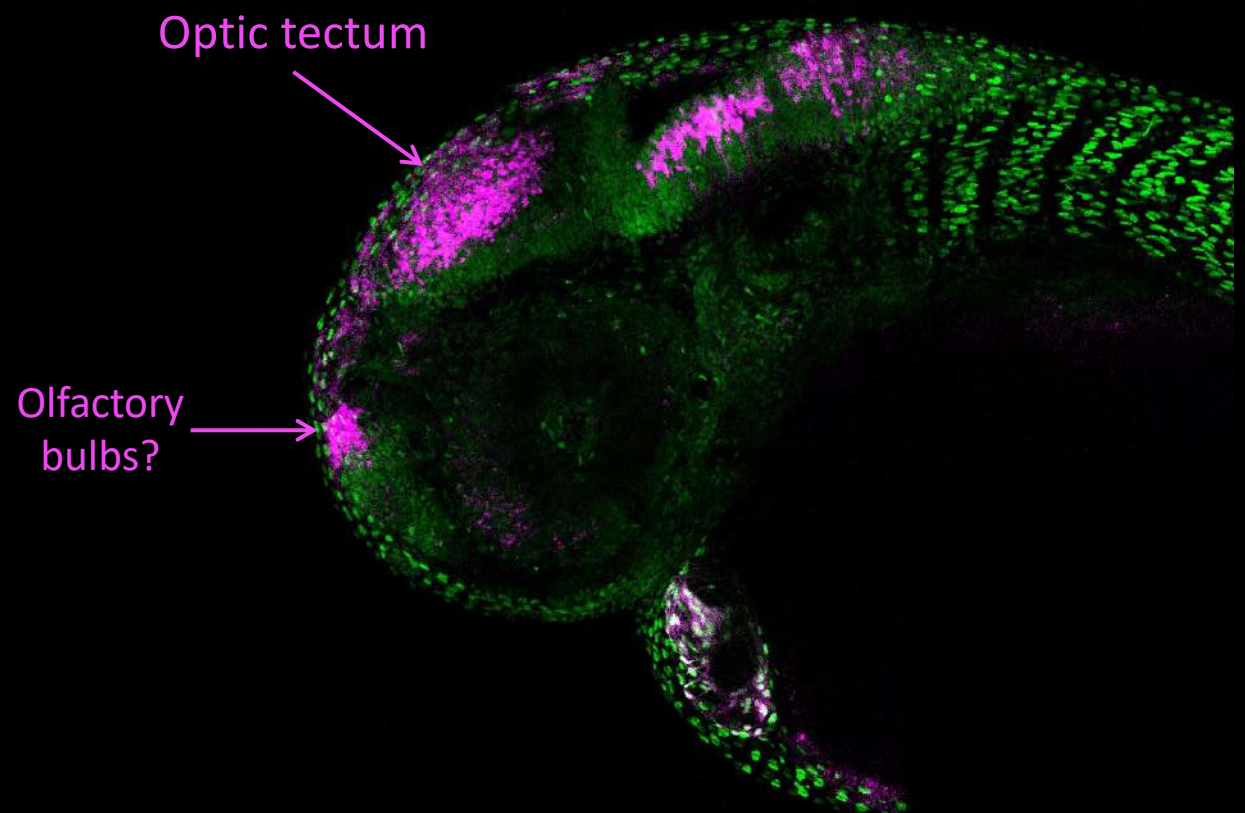
Jason Presnell

Jaret Lieberth

Conserved noncoding element 6078 enhaces expression in sensory regions of the brain

Tg(H2A-mCherry)

Tg(cne6078;E1b:EGFP)



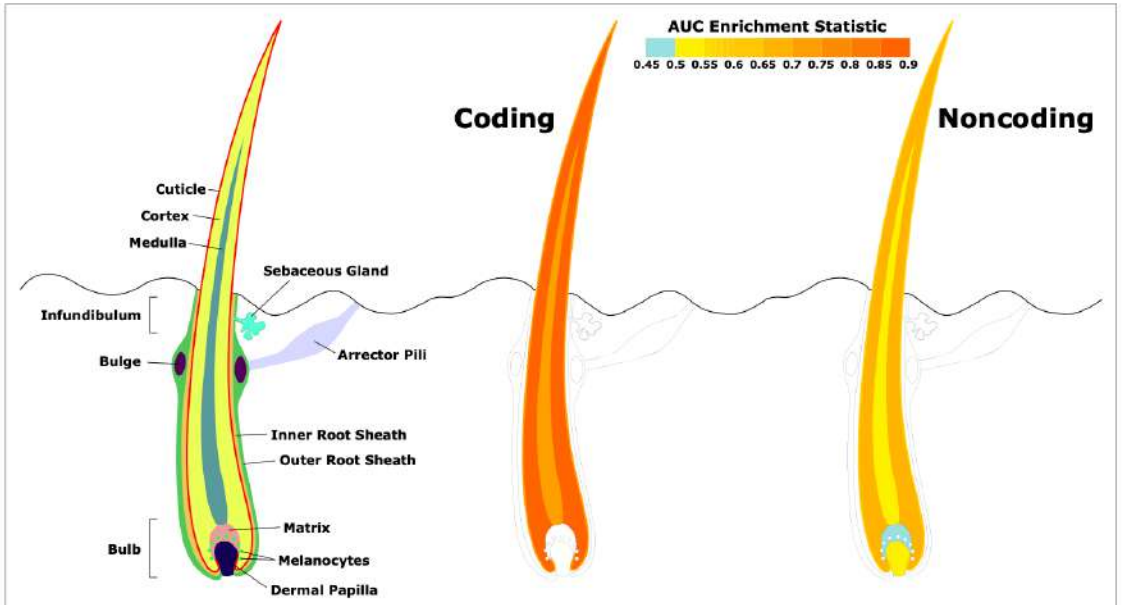
Convergent Evolution of Hairlessness

Amanda Kowalczyk

Protein-coding and Regulatory regions of hair structural genes were accelerated due to relaxed constraint

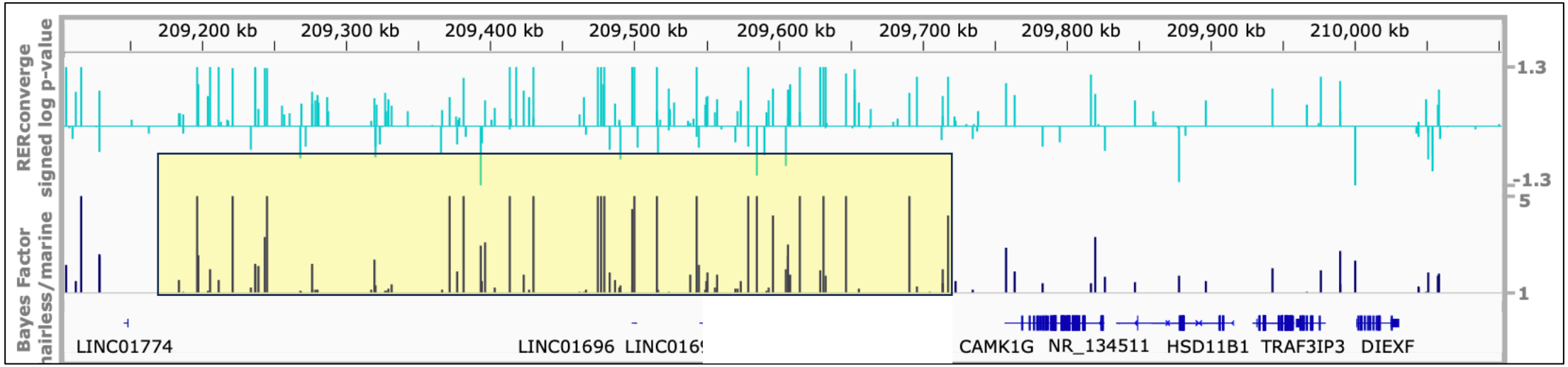
Scanned:

- 19,149 protein-coding genes
- 343,598 conserved non-coding elements (CNEs)



	Coding				Noncoding			
	p-value	adjusted p-value	AUC	Number of Genes	p-value	adjusted p-value	AUC	Number of CNEs
Cortex	0.00042	0.0050	0.86	8	4.20e-7	2.94e-6	0.65	93
Cuticle	0.010	0.035	0.72	11	1.58e-7	2.21e-6	0.59	287
Medulla	0.0059	0.028	0.72	13	5.80e-3	2.71e-2	0.55	263
Inner Root Sheath	0.15	0.41	0.60	19	0.54	0.66	0.51	626
Bulge	0.34	0.52	0.57	15	0.25	0.50	0.48	397
Outer Root Sheath	0.30	0.52	0.56	28	0.55	0.66	0.49	695
Matrix	0.71	0.90	0.55	4	0.078	0.18	0.46	191
Arrector Pili	0.30	0.52	0.54	45	0.39	0.66	0.51	2421
Sebaceous Gland	0.21	0.48	0.54	94	0.75	0.75	0.50	2709
Bulb	0.96	0.96	0.50	12	0.57	0.66	0.49	470
Infundibulum	0.87	0.94	0.48	7	0.73	0.75	0.51	218
Dermal Papilla	0.88	0.94	0.48	5	0.049	0.16	0.55	139
Melanocytes	0.52	0.72	0.46	19	0.48	0.66	0.49	663

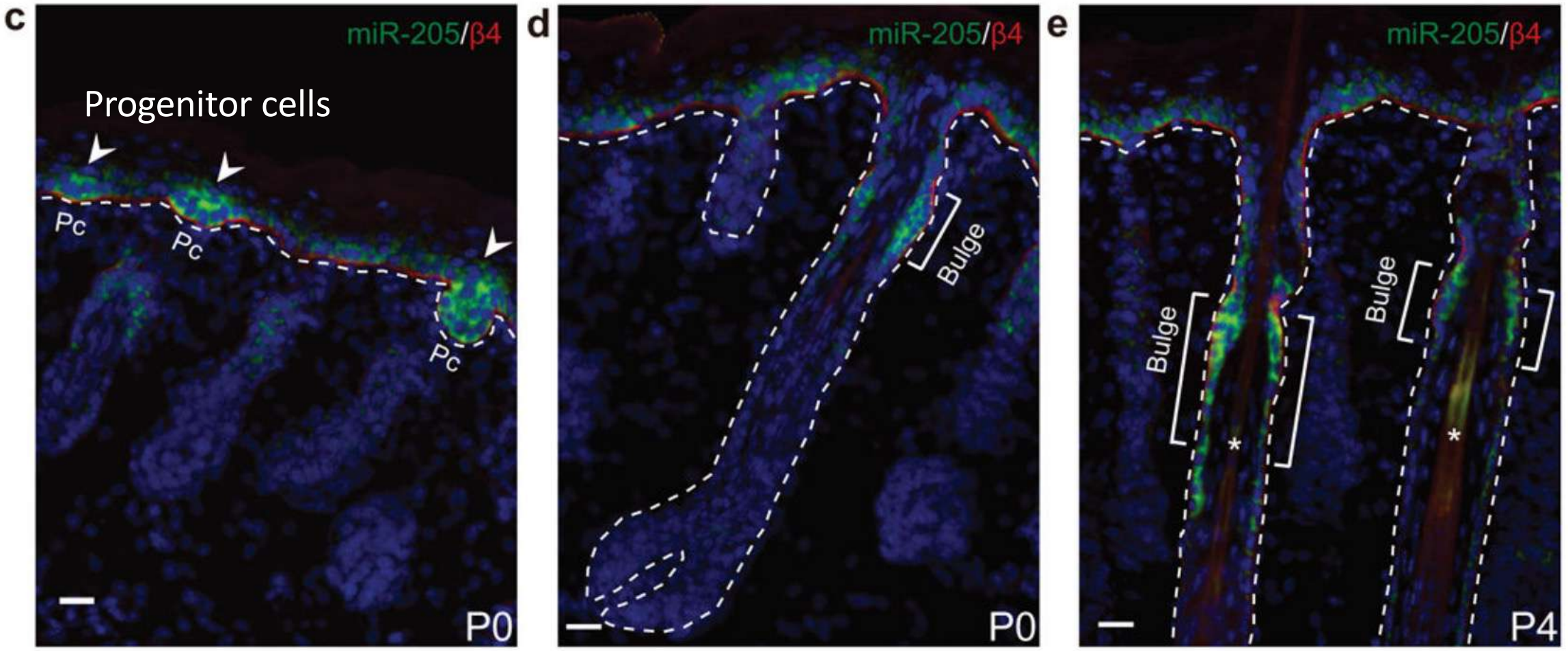
Most extreme cluster of hairless accelerated non-coding regions



MicroRNA MIR205

FDR q-value=1.4x10^{-4}

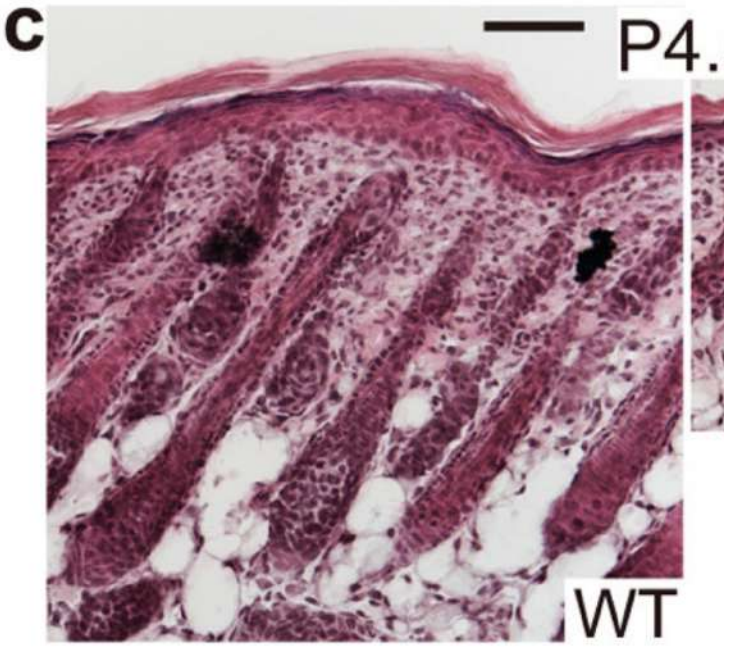
miR-205 is most highly expressed in the Hair Follicle Stem Cells located in the bulge of mature hair follicles



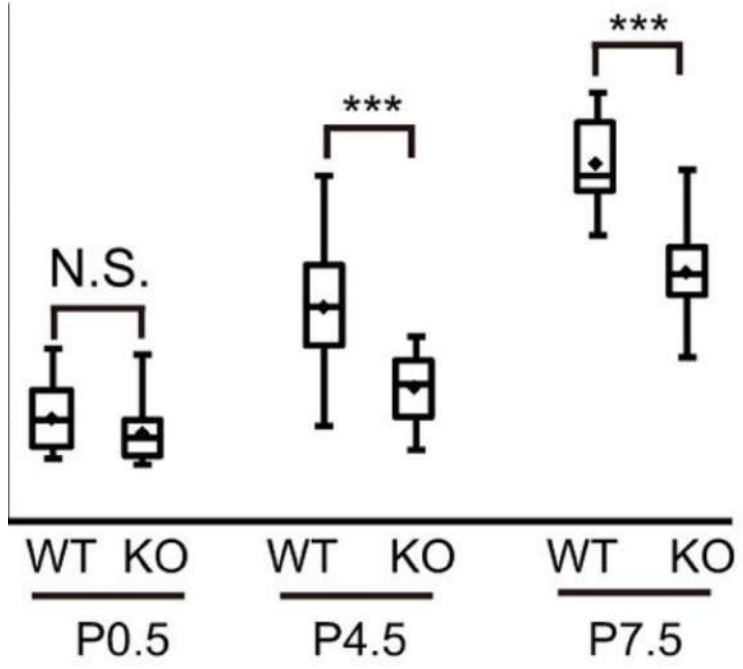
(Wang et al. *Nat. Cell. Biol.* 2013)

Abl
pro

Five additional uncharacterized microRNAs are in clusters of hair-associated elements



	Statistic	p-adj
MIR205	0.238	8.6e-10
MIR1305	0.162	1.1e-6
MIR924	0.178	6.8e-6
MIR124-1	0.218	2.4e-5
MIR346	0.283	2.4e-5
MIR759	0.152	8.3e-5
MIR3167	0.150	2.1e-4
MIR4255	0.189	4.4e-4



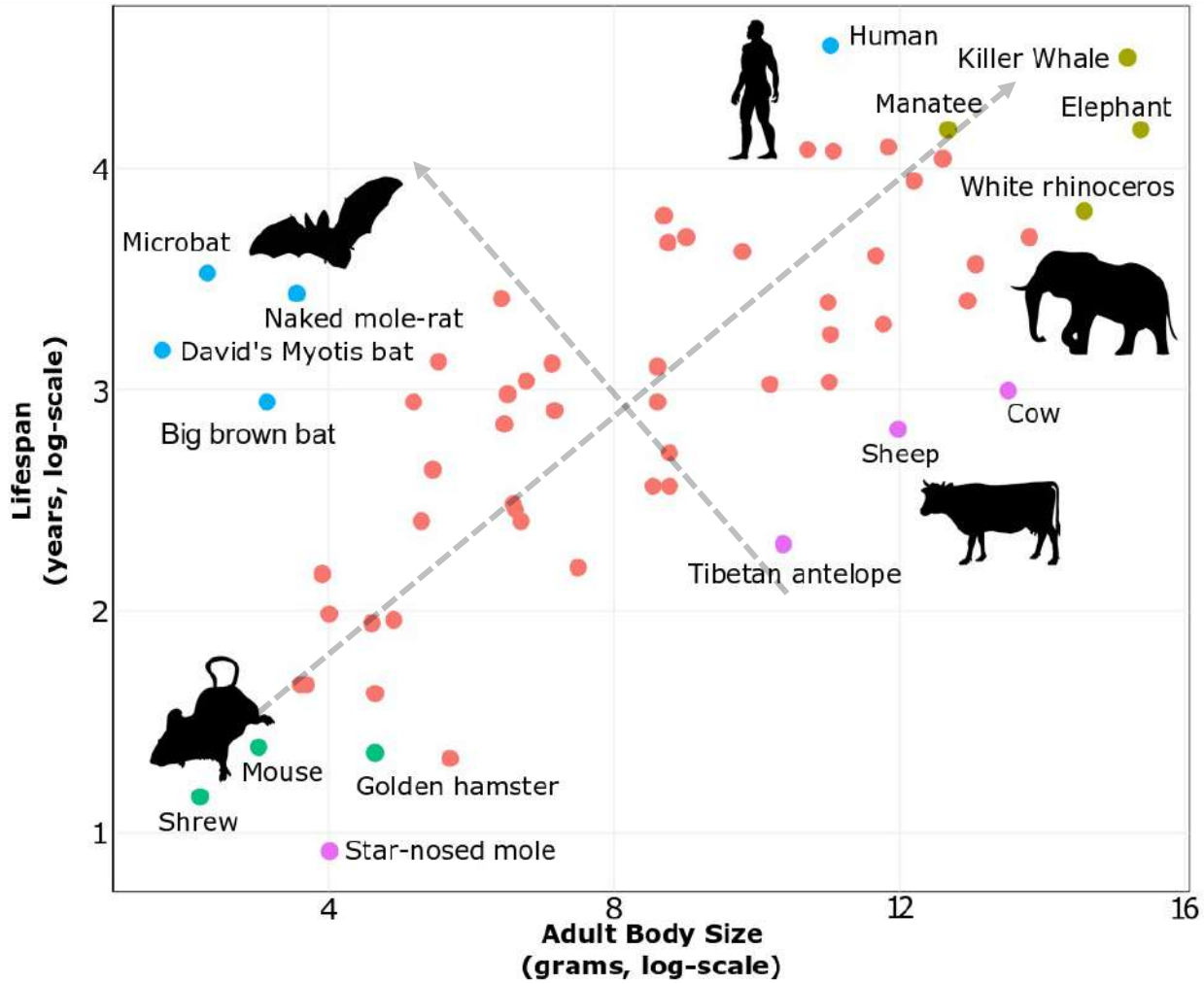
MicroRNA-205 Controls the PI3K Pathway. (War

m Cells by Modulating

Lab Achievement Unlocked

Which genes enable long lifespan?

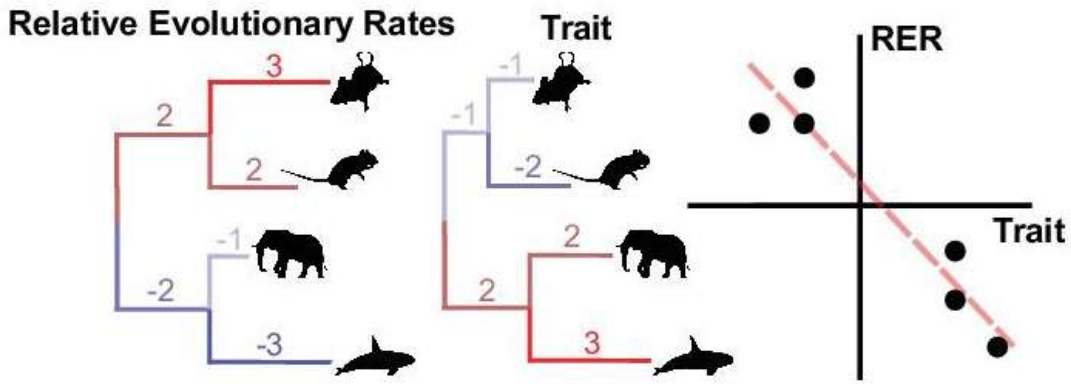
Extracting principal components of longevity and body size



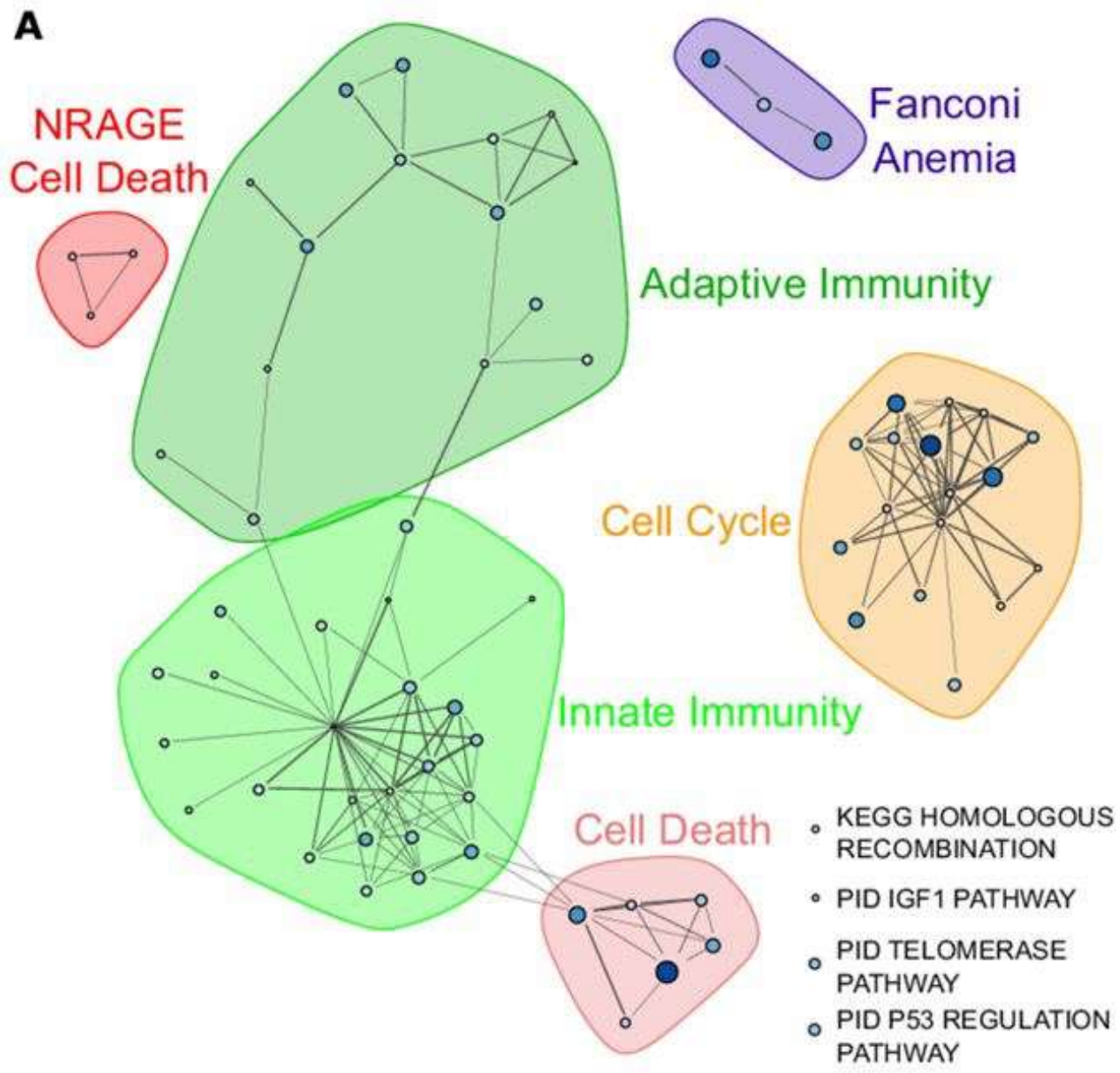
PC1 – long-lived, large-bodied
PC2 – independently long-lived

- PC1 high
- PC1 low
- PC2 high
- PC2 low

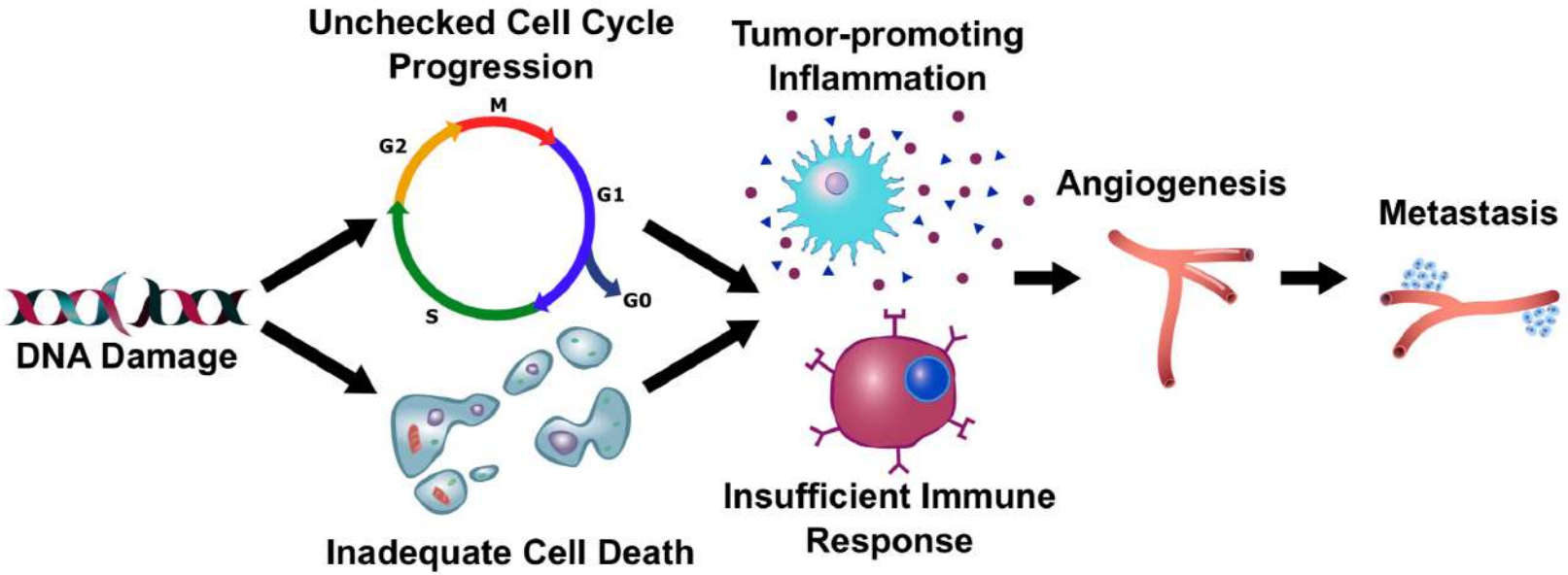
Kowalczyk et al. *eLife* 2020

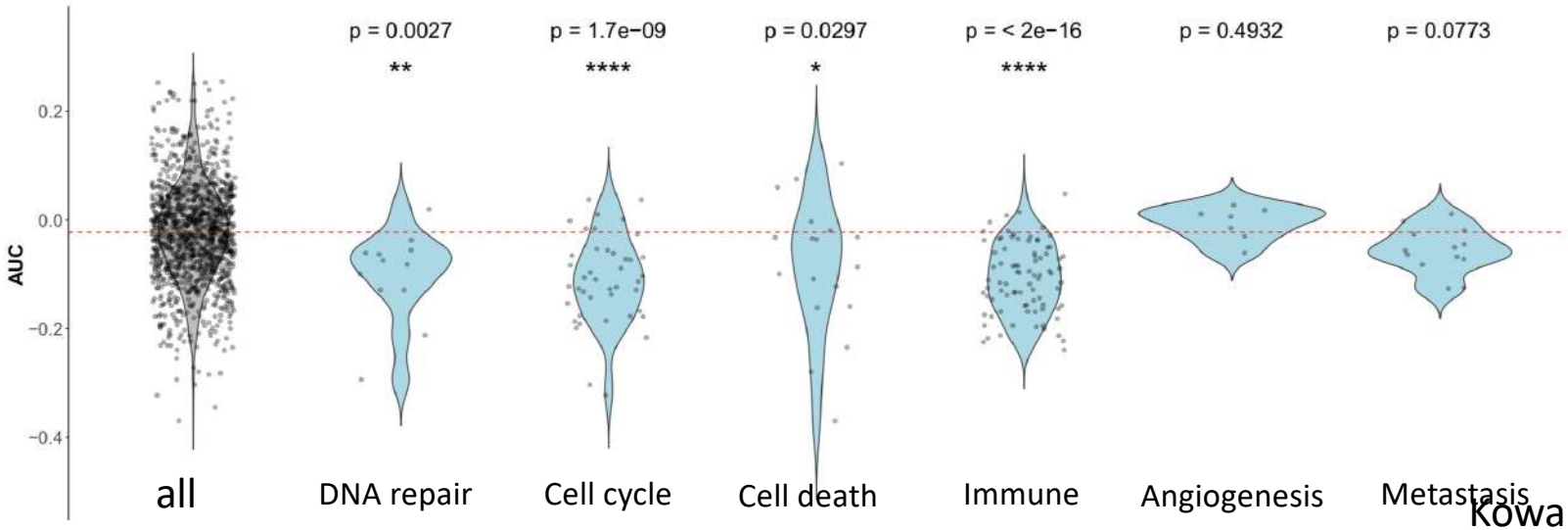


Highly Constrained (decelerated) genes in Long-lived species



Long-lived species are under intense pressure to prevent **cancer**



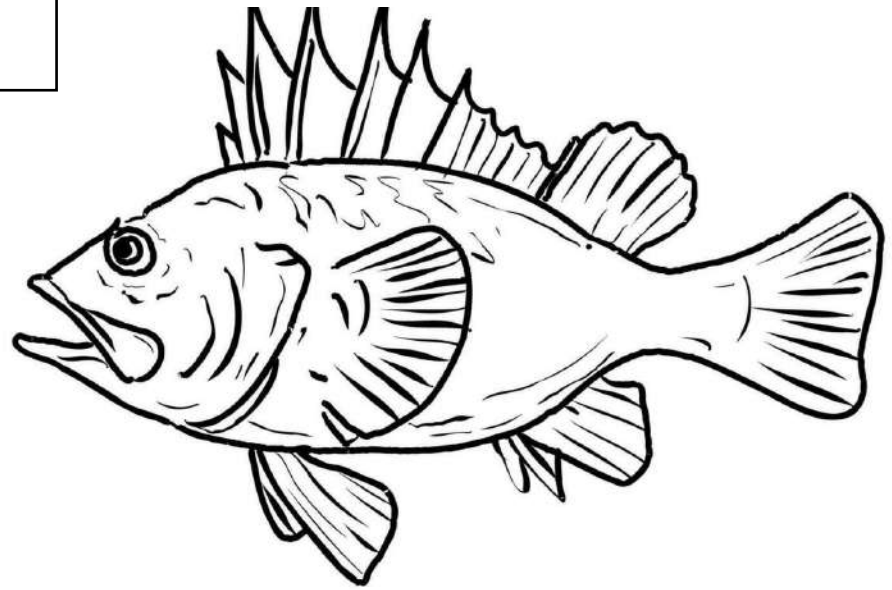


Kowalczyk et al. *eLife* 2020

Lifespan studies in separate taxonomic groups agree on importance of specific functions

- insulin signaling
- telomere maintenance
- DNA repair
- immunity

Pacific Rockfishes



Kowalczyk et al. *eLife* 2020

Kolora et al. *Science* 2021

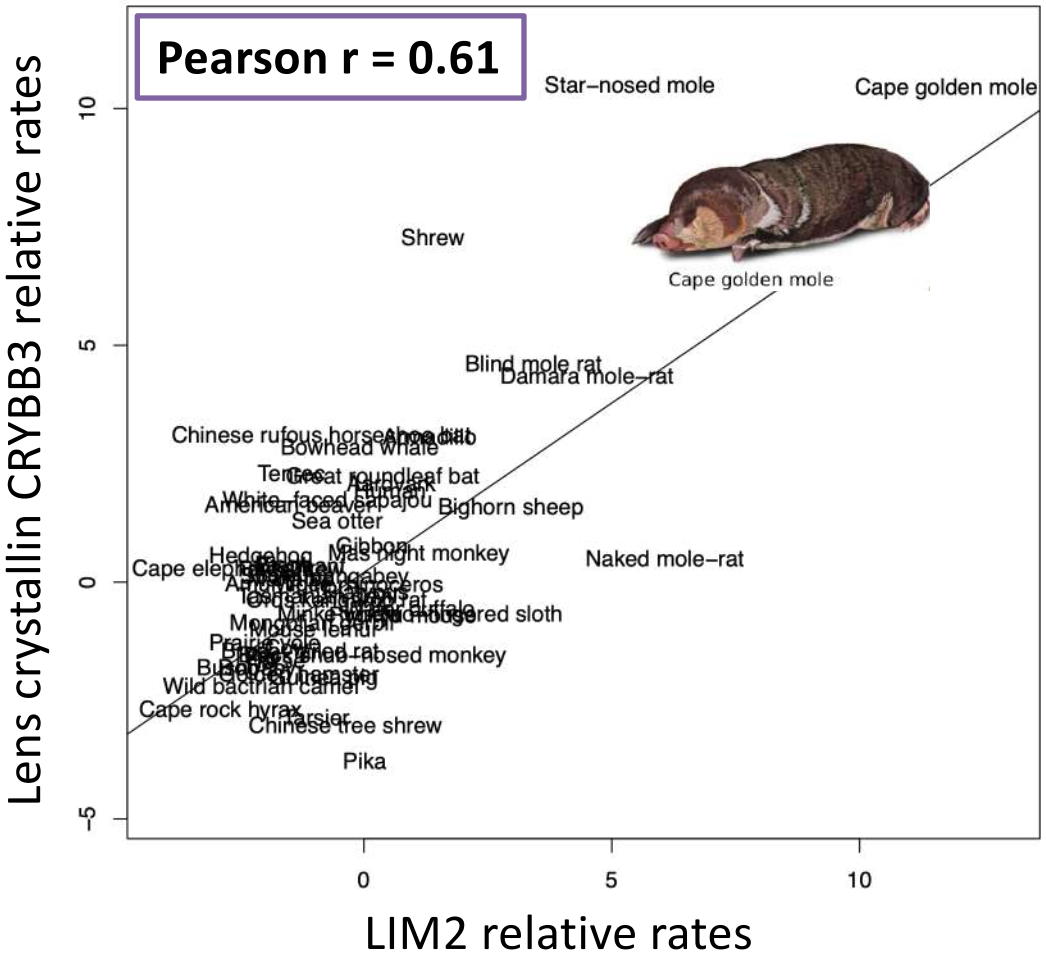
Convergent changes across traits

Trait	Discarded or relaxed constraint	Altered (positive selection)	Constraint increased
Aquatic	Taste and Smell Muscle proteins Nervous system	Skin Cell junction proteins Lung surfactant proteins	N/A
Subterranean	Ocular proteins & regulatory regions Footpad keratin	N/A	N/A
Hairless	Hair structural proteins Hair and skin regulatory regions	N/A	N/A
Long lifespan	N/A	N/A	Cell cycle DNA repair Telomerase IGF1 pathway
High altitude	Olfactory receptors Fetal hemoglobin	Nitric oxide signaling Hemoglobin	N/A

Evolutionary Rate Covariation
ERC

Predicting functional networks of genes using only orthologous sequences

LIM2's rates covary with other lens proteins



Evolutionary Rate Covariation (ERC)

expressed as correlation coefficient [-1,1]

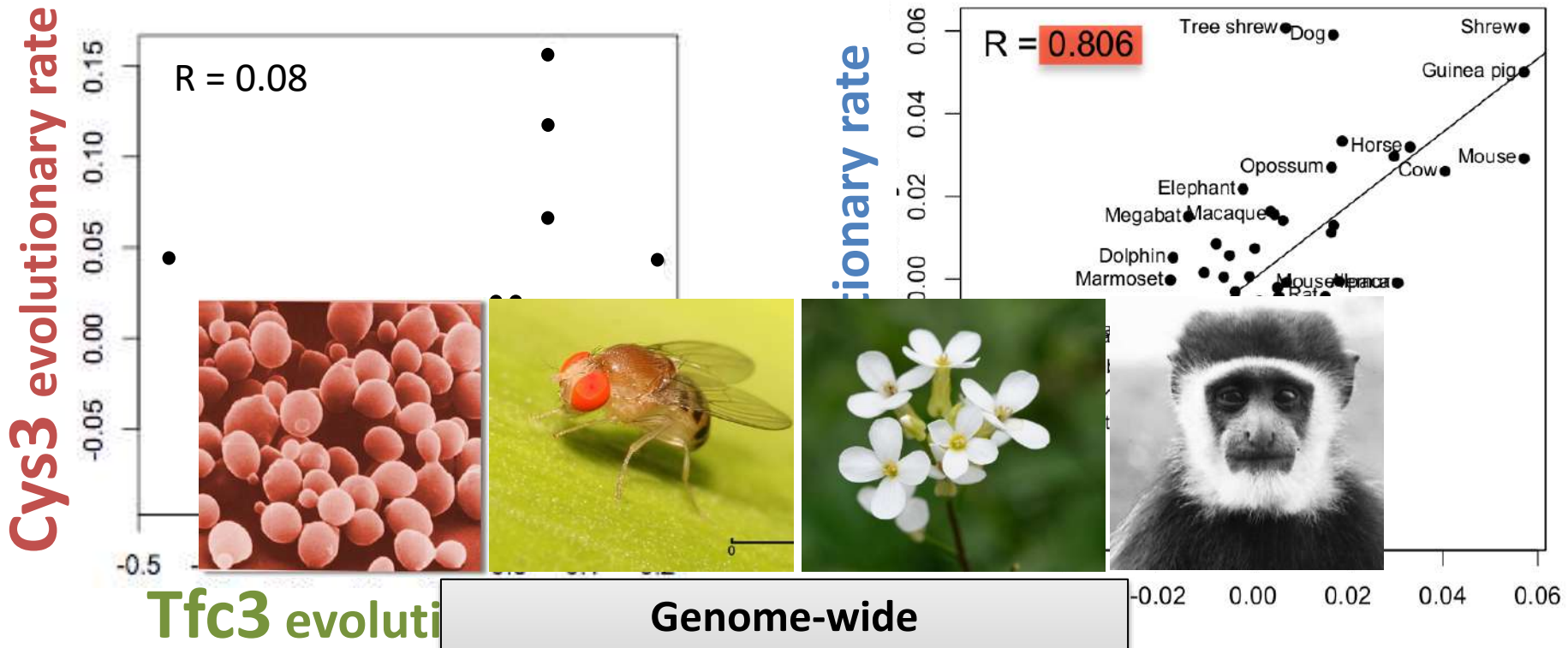
Covariation reflects co-function

Selective pressures affect entire functions.
Hence, co-functional genes have covarying rates.

Evolutionary Rate Covariation

Genes with shared pressures have rates that covary

We don't need to know the environmental pressure to infer functional relationships!



Genome-wide

Thousands of genes

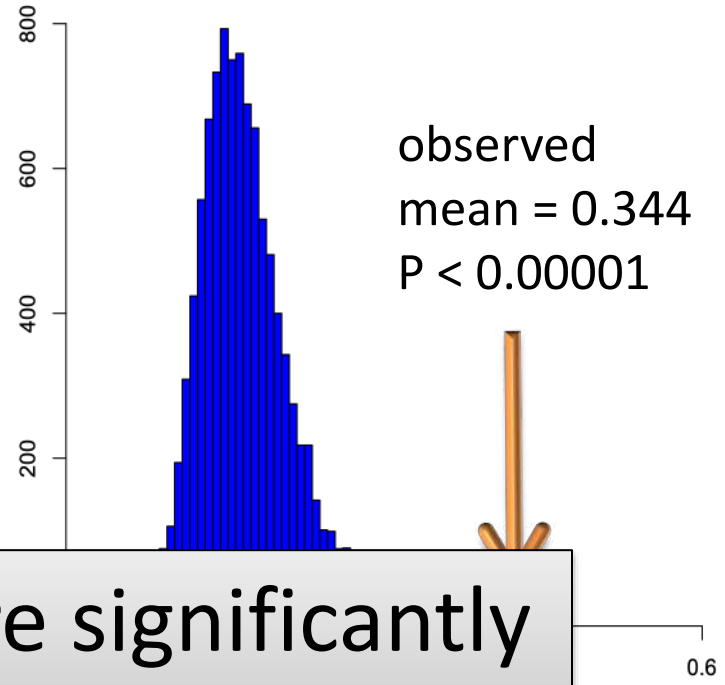
Millions of gene-gene comparisons

Members of protein complexes and pathways exhibit ERC with each other

Complement Cascade Pathway

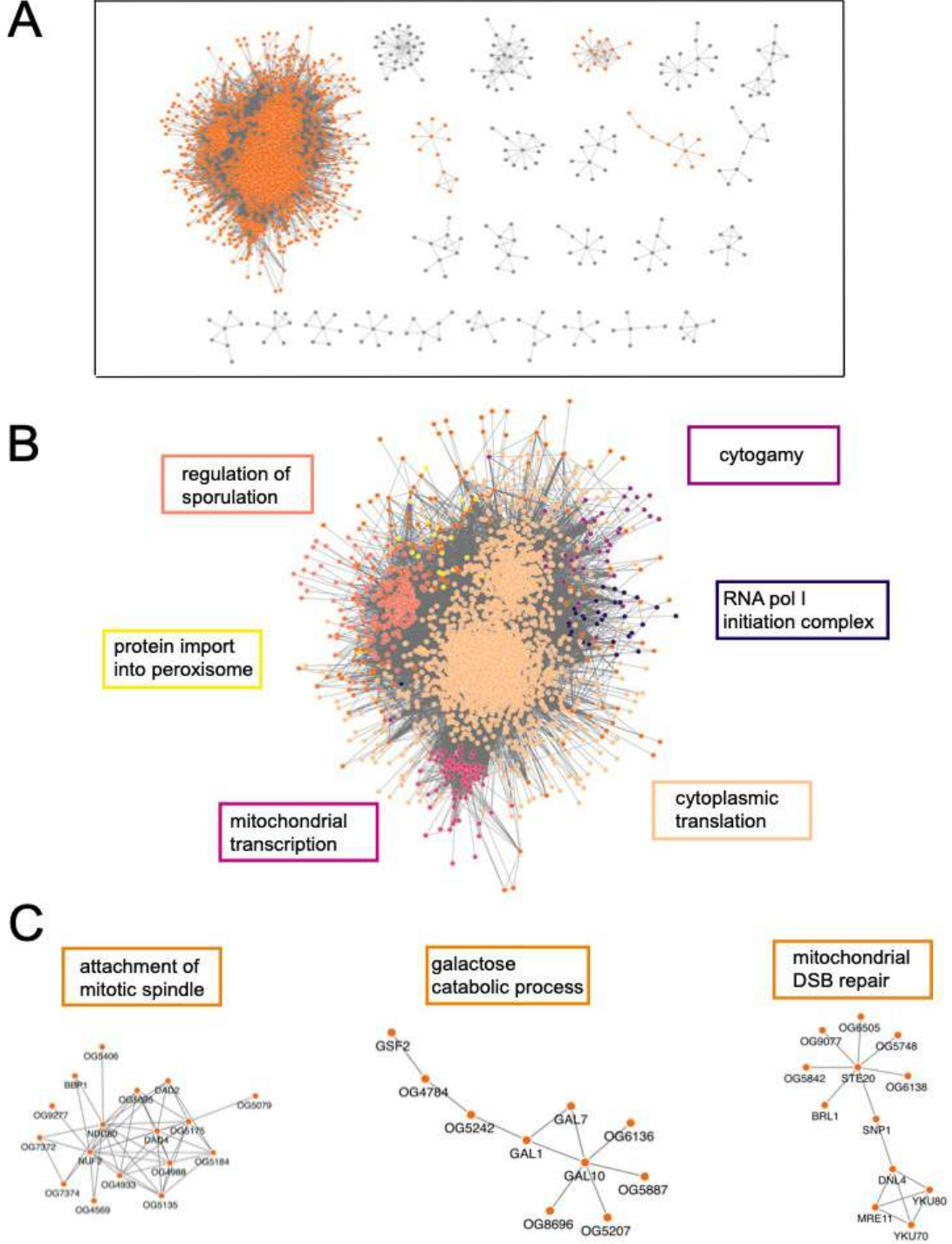
	C1QA	C1QB	C1QC	C1R	C1S	C3	C5	C7	C8A	C8B	C9	CD8A	CD96	CFD	CFH	CFI	SERPING1
C1QA	ND	0.50	0.56	0.54	0.30	0.24	0.34	0.28	0.27	0.35	0.31	0.18	0.32	0.09	0.26	0.30	0.41
C1QB	0.50	ND	0.55	0.60	0.62	0.24	0.33	0.18	0.53	0.49	0.22	0.37	0.16	-0.24	0.52	0.49	0.54
C1QC	0.56	0.55	ND	0.51	0.64	0.31	0.50	0.40	0.48	0.50	0.35	0.21	0.23	-0.29	ND	0.50	0.63
C1R	0.54	0.60	0.51	ND	0.54	0.26	0.25	0.17	0.40	0.45	0.24	0.26	0.40	-0.26	0.20	0.46	0.41
C1S	0.30	0.62	0.64	0.54	ND	0.44	0.45	0.44	0.35	0.41	0.55	0.56	0.25	0.04	0.58	0.81	0.56
C3	0.24	0.24	0.31	0.26	0.44	ND	0.68	0.18	0.12	0.21	0.24	0.05	-0.15	0.24	0.66	0.53	0.27
C5	0.34	0.33	0.50	0.25	0.45	0.68	ND	0.46	0.34	0.29	0.44	0.02	-0.07	0.48	0.66	0.58	0.35
C7	0.28	0.18	0.40	0.17	0.44	0.18	0.46	ND	0.46	0.40	0.58	0.15	0.14	-0.03	0.35	0.50	0.40
C8A	0.27	0.53	0.48	0.40	0.35	0.12	0.34	0.46	ND	0.79	0.34	-0.02	0.45	-0.26	0.32	0.39	0.21
C8B	0.35	0.49	0.50	0.45	0.41	0.21	0.29	0.40	0.79	ND	0.46	0.02	0.57	-0.12	0.34	0.41	0.21
C9	0.31	0.22	0.35	0.24	0.55	0.24	0.44	0.58	0.34	0.46	ND	0.36	ND	0.44	0.65	0.51	0.57
CD8A	0.18	0.37	0.21	0.26	0.56	0.05	0.02	0.15	-0.02	0.02	0.36	ND	0.55	ND	0.27	0.35	0.34
CD96	0.32	0.16	0.23	0.40	0.25	-0.15	-0.07	0.14	0.45	0.57	ND	0.55	ND	ND	ND	0.14	0.22
CFD	0.09	-0.24	-0.29	-0.26	0.04	0.24	0.48	-0.03	-0.26	-0.12	0.44	ND	ND	ND	ND	0.15	-0.04
CFH	0.26	0.52	ND	0.20	0.58	0.66	0.66	0.35	0.32	0.34	0.65	0.27	ND	ND	ND	0.62	0.39
SERPIN																	

Distribution of means of random draws

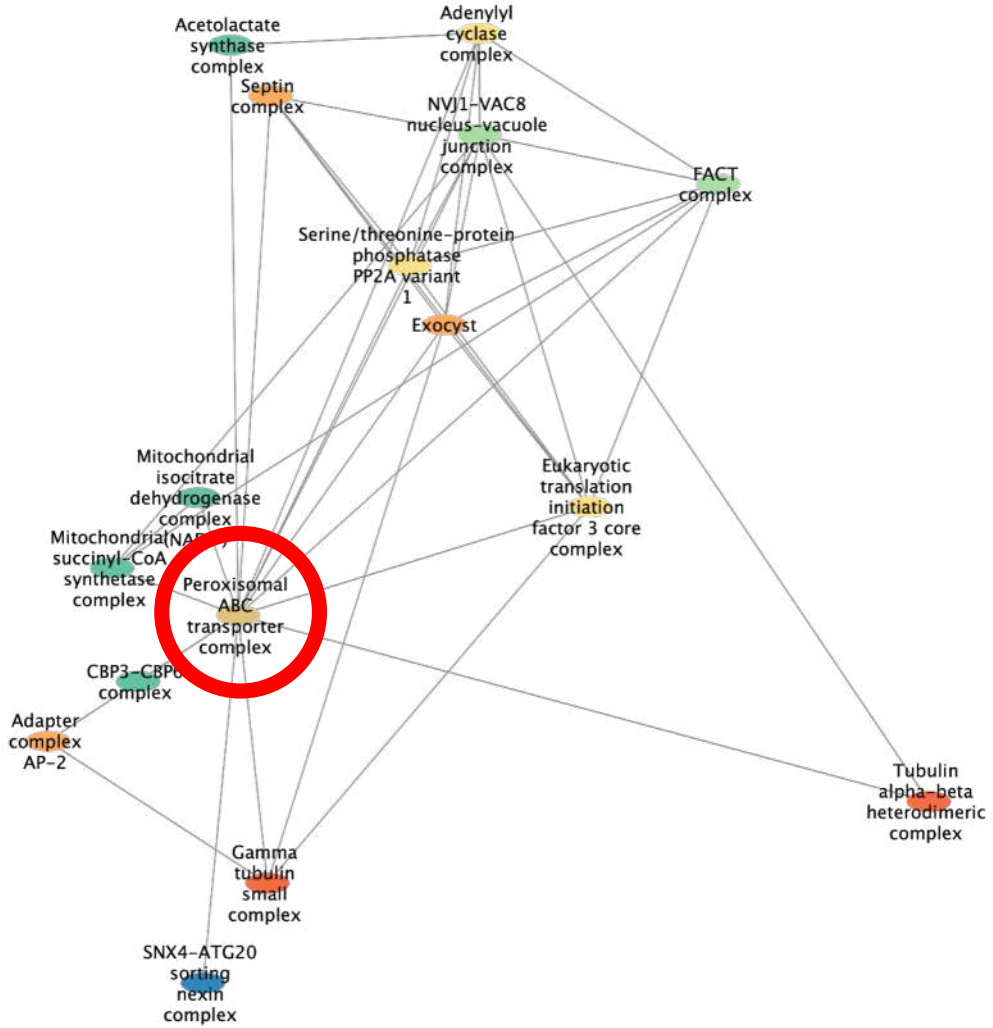


Most protein complexes are significantly correlated

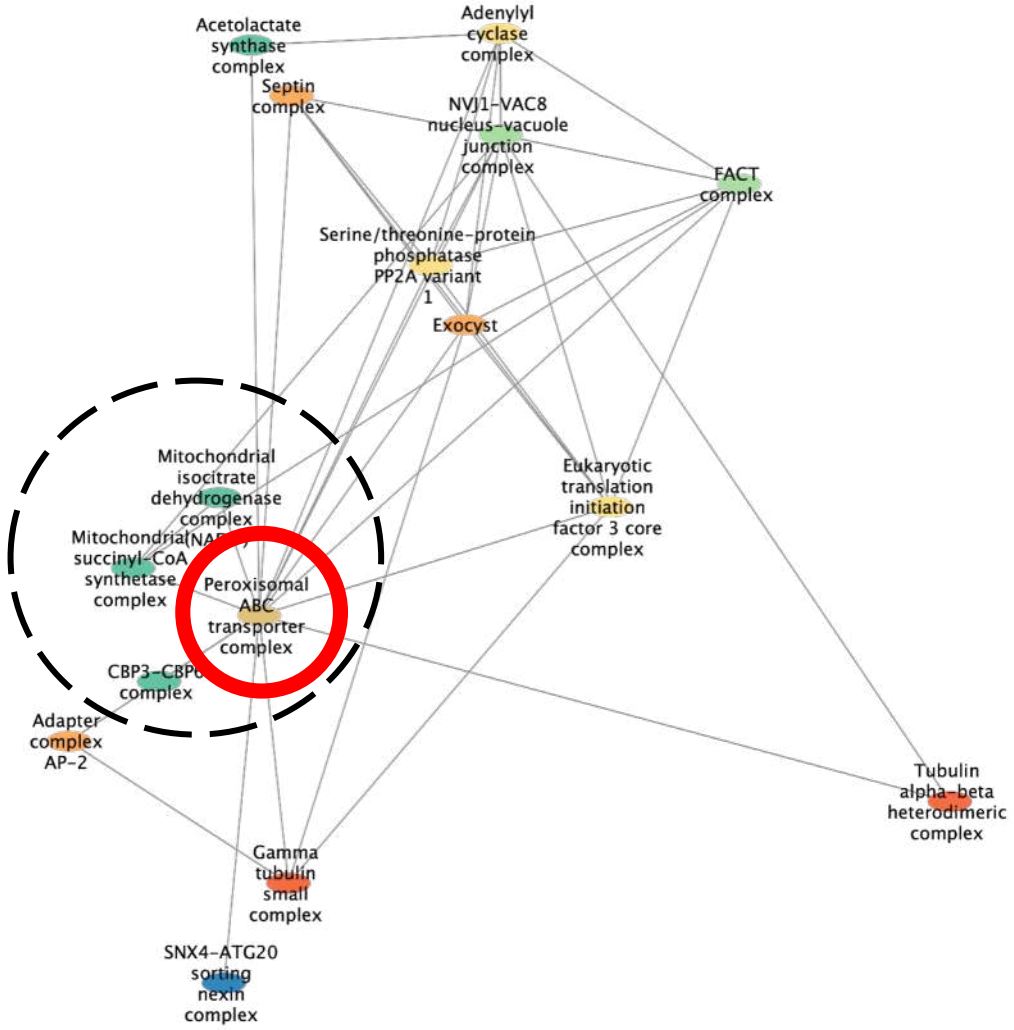
ERC-based clustering constructs functionally related clusters of proteins



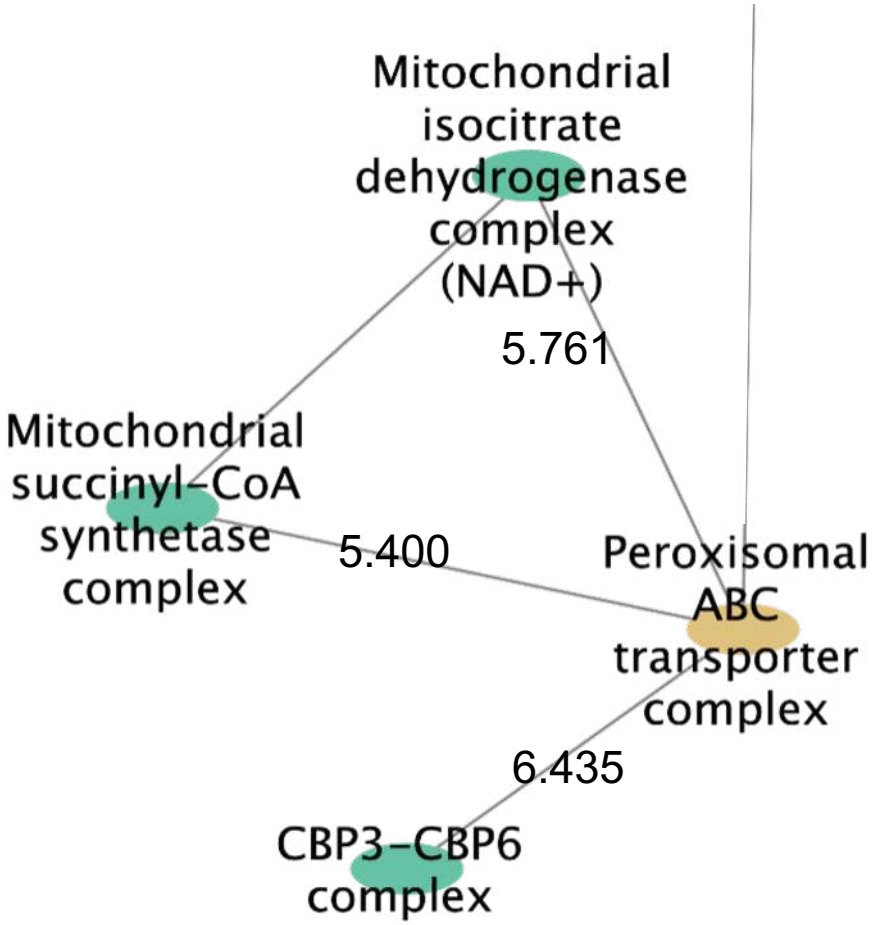
Edges capture functional associations

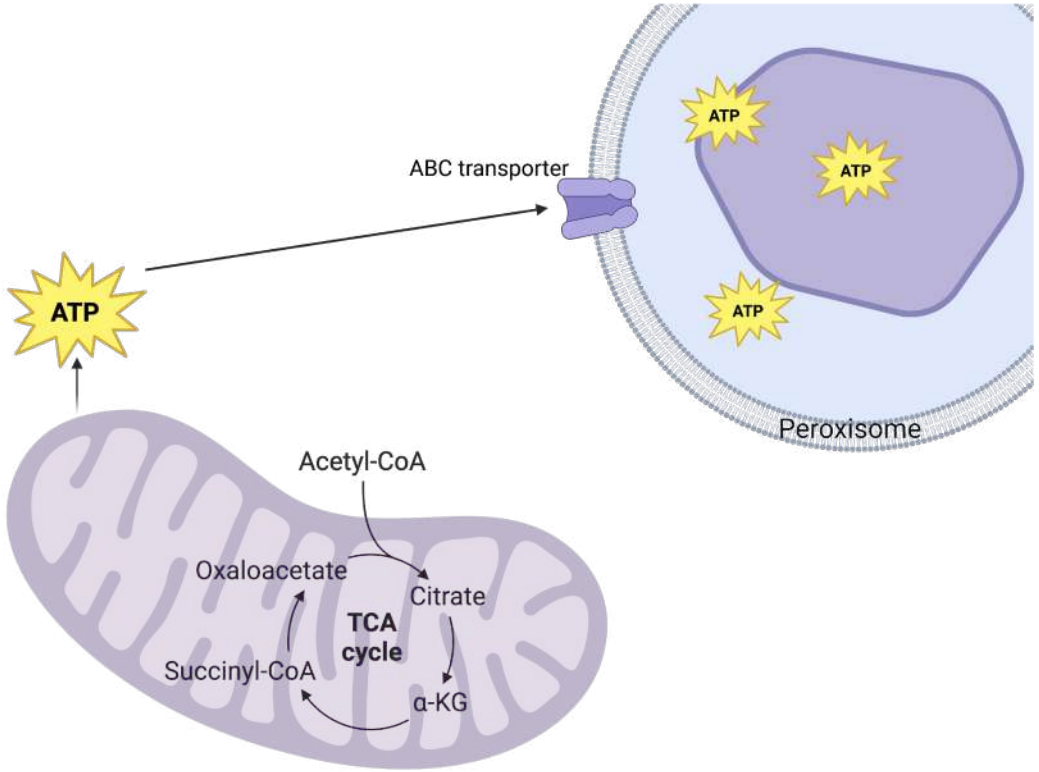


Edges capture functional associations



Edges capture functional associations



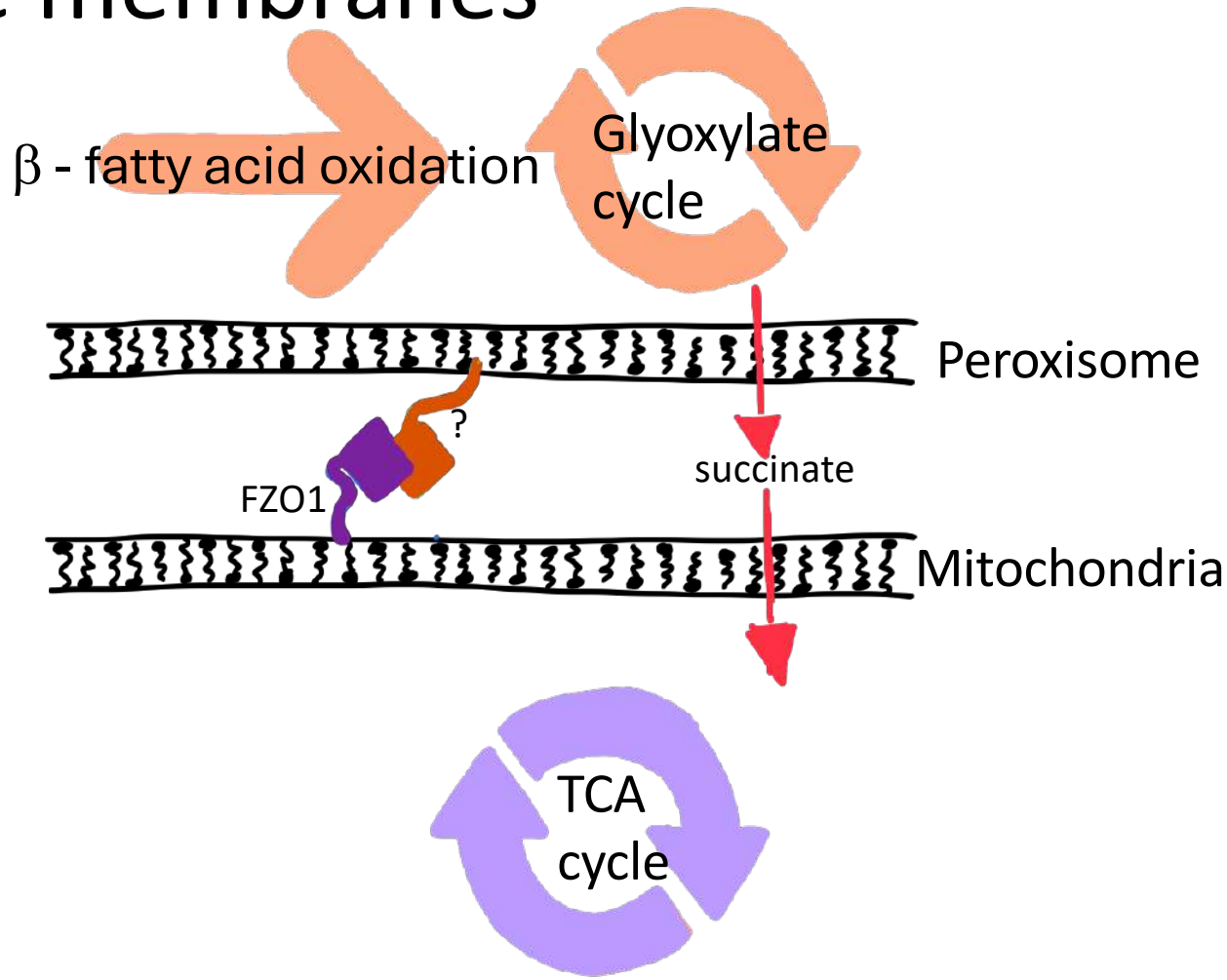


van Roermund, C. W., IJlst, L., Linka, N., Wanders, R. J., & Waterham, H. R. (2022). Peroxisomal ATP Uptake Is Provided by Two Adenine Nucleotide Transporters and the ABCD Transporters. *Frontiers in Cell and Developmental Biology*, *9*. https://doi.org/10.3389/fcell.2021.788921

ERC is elevated for functions shared across organelle membranes

	FAA1	FAA2	FAT1	POX1	FOX2	ECI1
FAA1						
FAA2	3.07					
FAT1	7.71	3.54				
POX1	4.87	NA	5.21			
FOX2	9.78	5.34	8.27	9.18		
ECI1	7.26	5.39	7.50	NA	11.80	
POT1	NA	4.33	6.01	7.99	16.65	10.23

Average ftERC = 8.652
permutation p-value < 0.001



Physical amino acid coevolution is a minor contributor to ERC

	AME1_1	AME1_2	AME1_3	AME1_4	OKP1_1	OKP1_2	OKP1_3	CTF19_1
MCM21_1	2.71	2.74	2.80	2.00	1.71	1.59	1.97	0.75
MCM21_2	2.01	1.83	2.19	1.88	2.04	3.06	1.91	2.42
AME1_1					3.82	1.20	3.06	0.07
AME1_2					4.18	4.85	4.10	3.06
AME1_3					5.02	5.56	5.05	2.38
AME1_4					4.73	5.10	4.80	4.40
OKP1_1								2.26
OKP1_2								2.46
OKP1_3								1.58

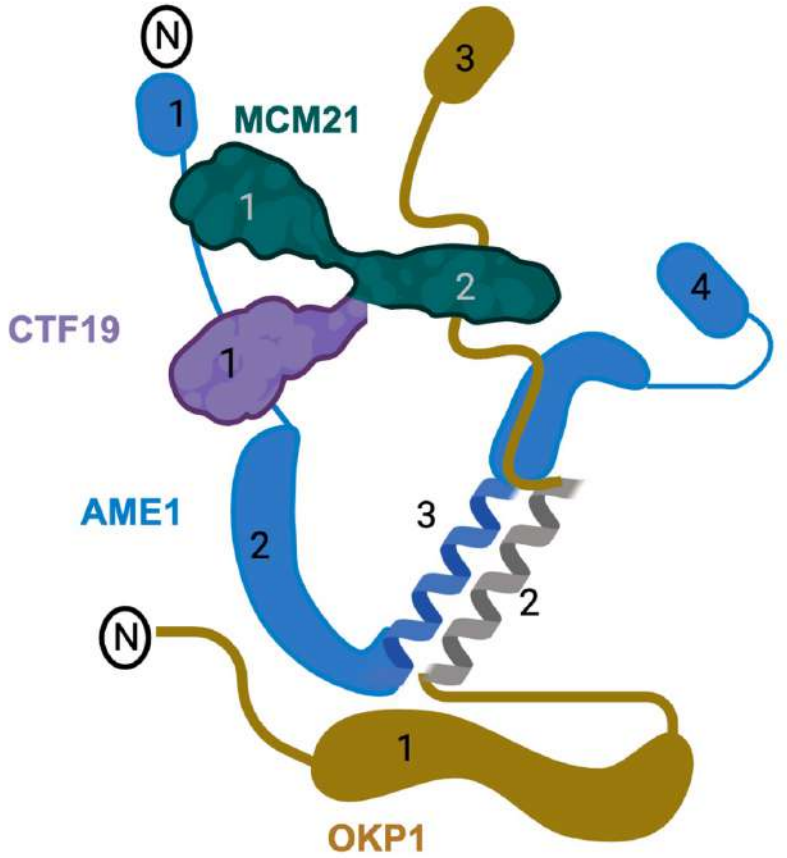


Figure 3. Recreation of ***Figure 1B*** with the COMA complex. The table shows the evolutionary rate covariation (ERC) values for each domain pair as labeled in the cartoon on the right. The domain pairs with physical interactions are highlighted in green.

Little Chikina Clark. *eLife*. 2024

Physical amino acid coevolution is a minor contributor to ERC

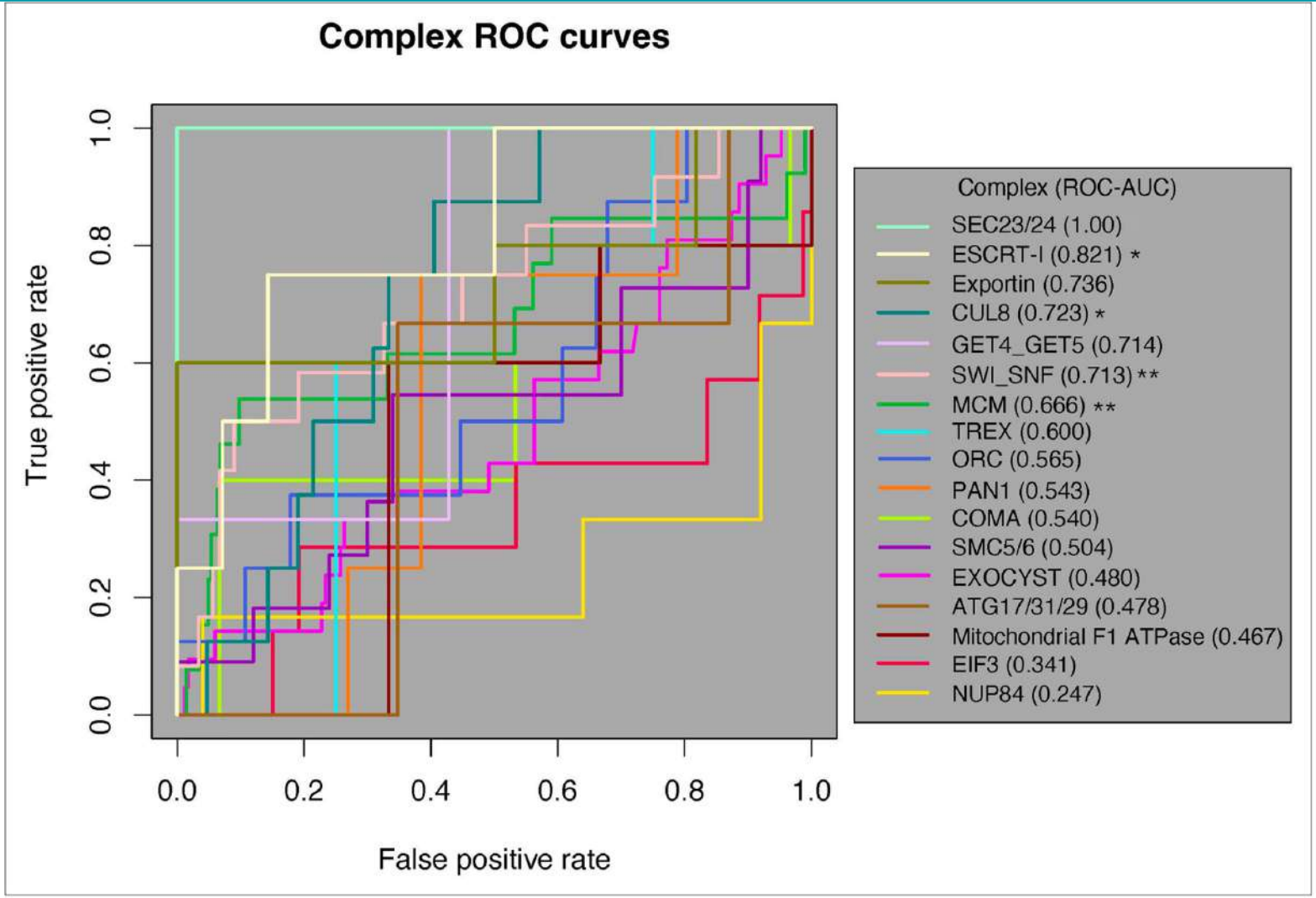


Figure 4. Receiver-operating characteristic (ROC) curve analysis of all 17 protein complexes. Of the 17 complexes, 12 have an ROC-AUC > 0.5. The SEC23/24 complex (bright green) has the highest ROC-AUC at 1, and the NUP84 complex (marigold) has the lowest AUC of 0.247. One-tailed Mann–Whitney *U* test, *p<0.05, **p<0.01. AUC, area under the curve.

Little Chikina Clark. *eLife*. 2024

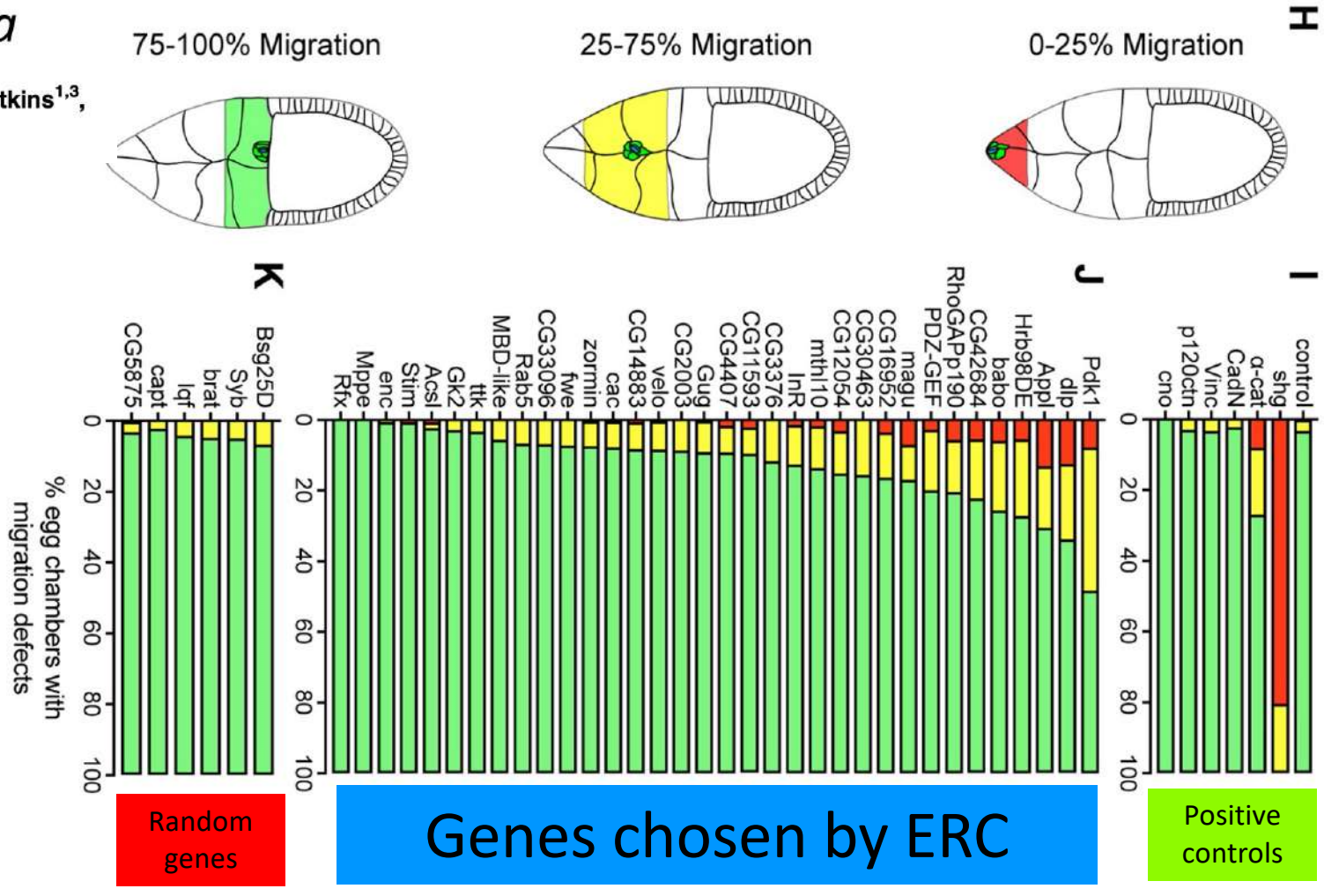
ERC is a result of fluctuating selective pressures

- Co-functional genes experience the same pressures
- Those pressures fluctuate over time, causing changes in their rates of sequence evolution
- The result is that co-functional proteins have evolutionary rates that change in correlation

ERC uncovers relationships missed by genetic and biochemical assays.

Evolutionary rate covariation analysis of E-cadherin identifies Raskol as a regulator of cell adhesion and actin dynamics in *Drosophila*

Qanber Raza[1], Jae Young Choi[2], Yang Li[1], Roisin M. O'Dowd[1], Simon C. Watkins[1,3], Maria Chikina[4], Yang Hong[1], Nathan L. Clark[4], Adam V. Kwiatkowski[1]*



ERC uncovers relationships missed by genetic and biochemical assays.

- ERC Identifies 6 New Members of a Protein Network Required for *Drosophila* Post-mating Reponses. (*PLoS Genetics*. 2014)
- Amino Acid Transporter Jhl-21 Coevolves with Glutamate Receptors and Impacts NMJ Physiology in *Drosophila* Larvae. (*Sci.Reports*. 2016)
- ERC identifies new signalin
  Genetics. 2019)
- Evolution-based screening
  DNA repair genes. (*PNAS*. 2
- Evolutionary rate covariati
  transporter. (*Biochem J*. 20
- Identifying Trafficking Regu
  O'Donnell – University of P
- Correlated evolutionary ra
  family (Choi, Clark, Purugg
- A *Drosophila* screen identi
  deficiency. (*eLife*. 2020)
- MAIA, Fc receptor-like 3, s
  fertilization. (Svobodova et
- Multiple 9-1-1 complexes promote homolog synapsis, DSB repair, and ATR signaling during mammalian meiosis. (*bioRxiv*. 2021)
- Experimental exchange of paralogous domains in the MLH family provides evidence of sub-functionalization after gene duplication. (*G3*. 2021)

PLOS GENETICS

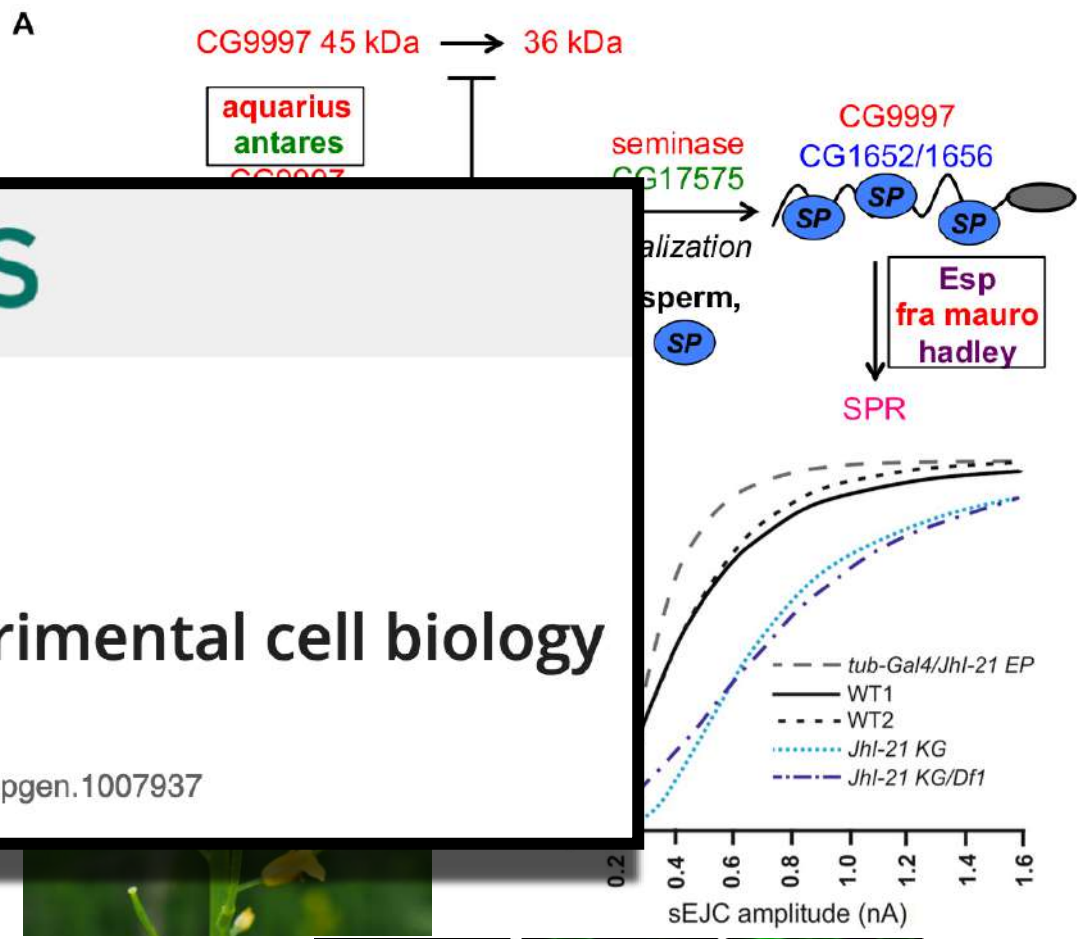
OPEN ACCESS

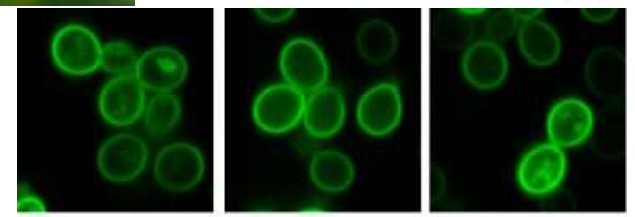
PERSPECTIVE

Evolution as a guide for experimental cell biology

Jeffrey Colgren, Scott A. Nichols

Published: February 14, 2019 • https://doi.org/10.1371/journal.pgen.1007937

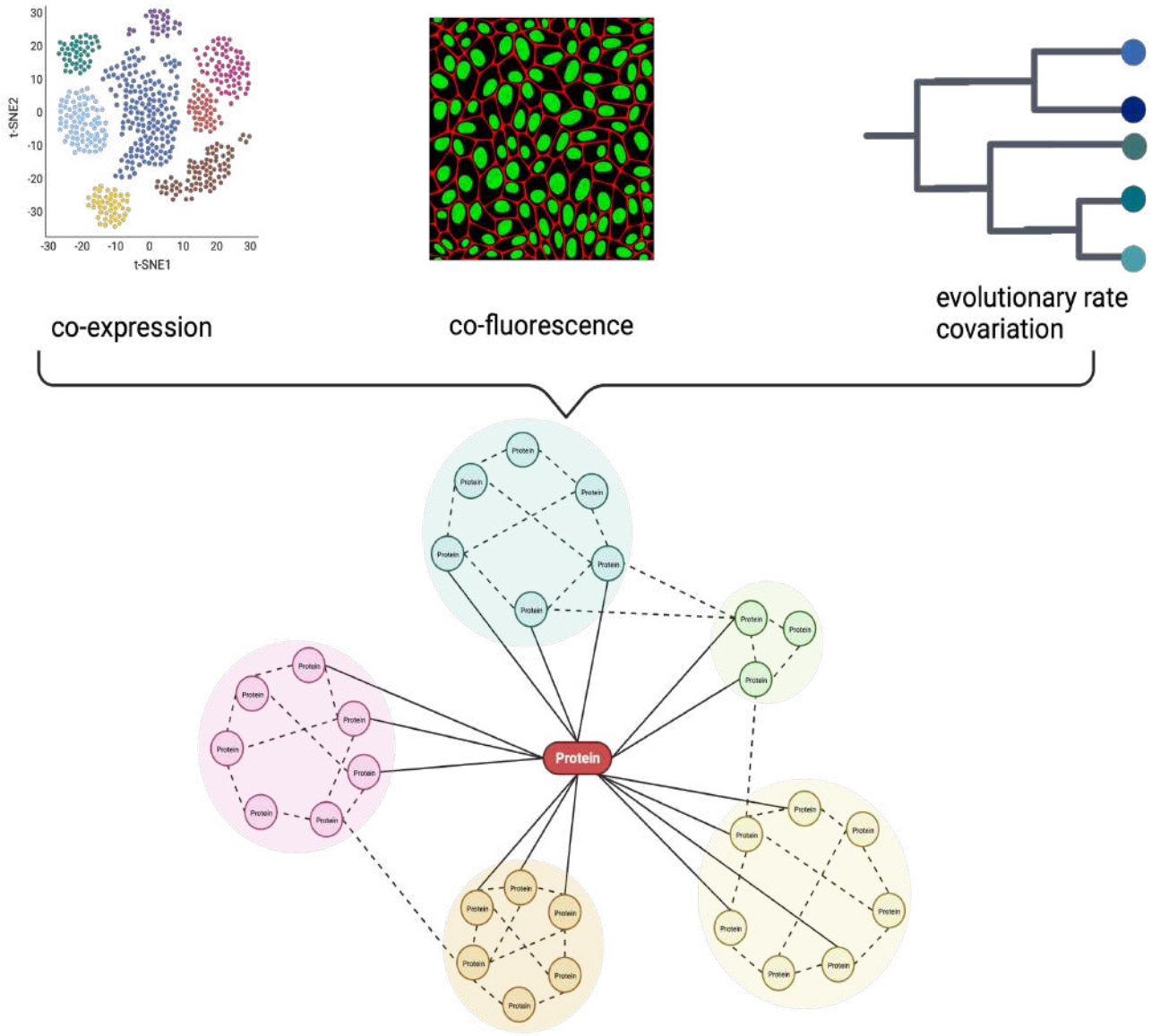




Clement Chow lab - University of Utah
Ellen Leffler lab - University of Utah

Steenwyk *et al. Sci Advances* 2022.
Tao *et al. Nature Aging.* 2024

ERC as a complementary approach to infer gene functional networks



Larger datasets should provide enormous power

343 yeast species !

Rokas & Hittinger

(Shen et al. *Cell* 2018)

Highly contiguous assemblies of 101 drosophilid genomes

(Kim et al. *eLife* 2021)

zoonomia

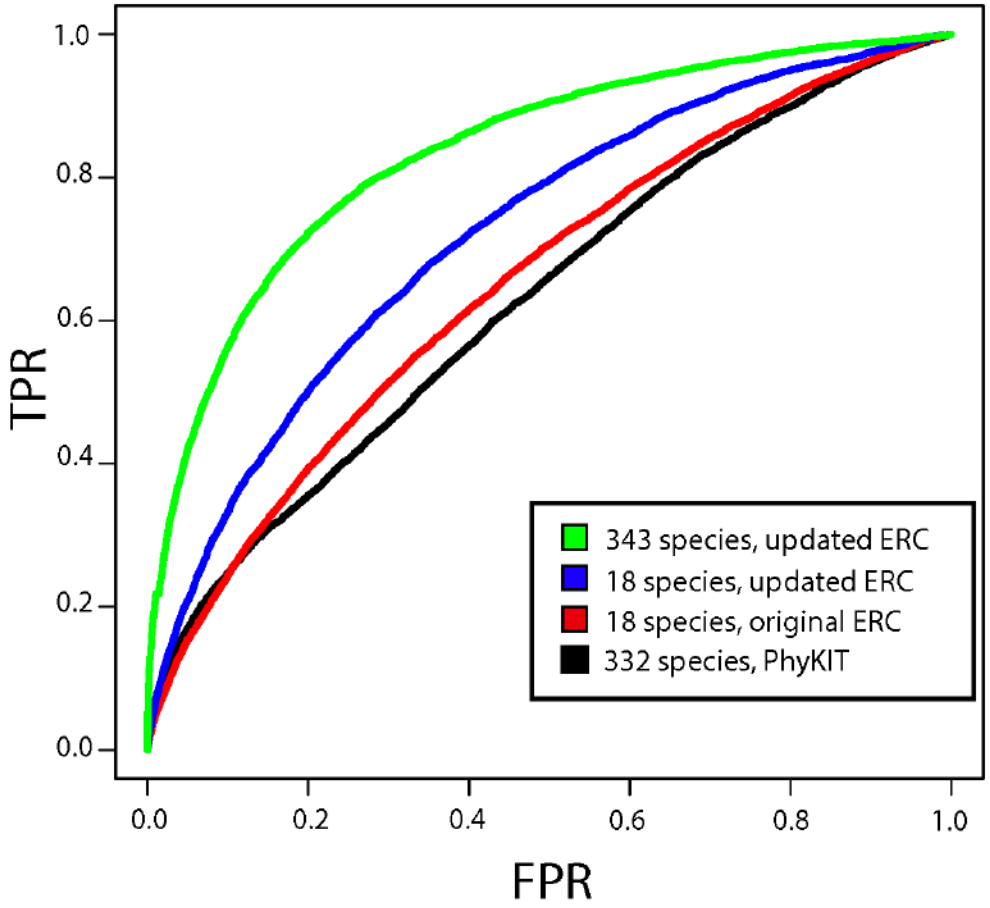
428 mammals

ERC 2.0 method advances

- Much faster
- Handles hundreds or thousands of species
- More powerful

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

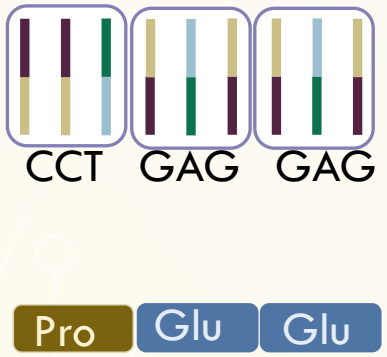
Jordan Little



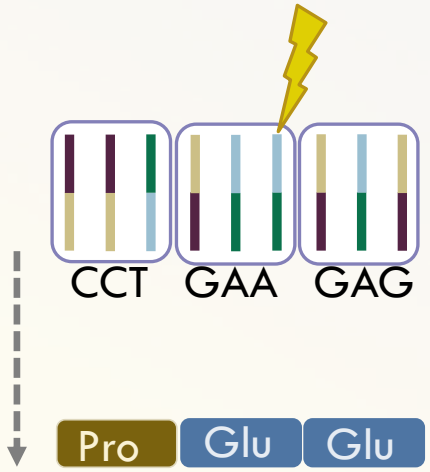
Little et al. *Genome Research* 2025

Inferring past positive selection for amino acid divergence using d_N/d_S

Initial genotype



Genotype after a single nucleotide change

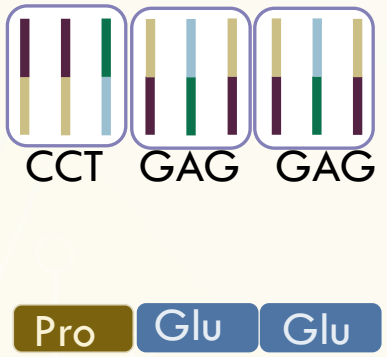


No change in amino acid encoded

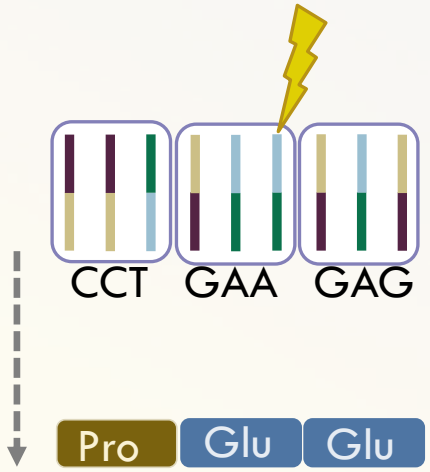
Direct connection to phenotype

Protein from: Al-Haggar et al., Eur J Hum Genet 2012

Initial genotype



Genotype after a single nucleotide change



No change in amino acid encoded

Direct connection to phenotype

Protein from: Al-Haggar et al., Eur J Hum Genet 2012

dS

Initial genotype

CCT GAG GAG

Pro Glu Glu

CCT GAG GAG

Pro Glu Glu

Genotype after a single nucleotide change

CCT GTG GAG

Pro Val Glu

Change in amino acid encoded

CCT GAA GAG

Pro Glu Glu

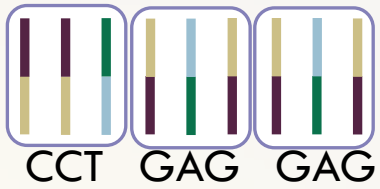
No change in amino acid encoded

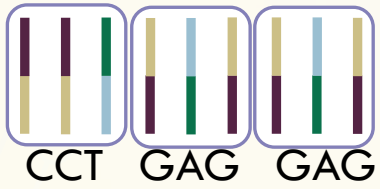
Direct connection to phenotype

dS

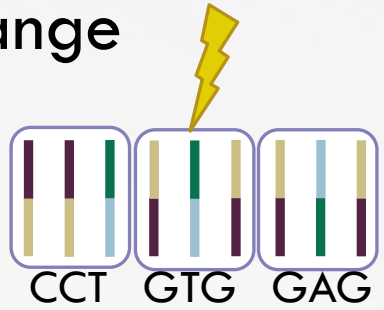
Protein from: Al-Haggar et al., Eur J Hum Genet 2012

Initial genotype

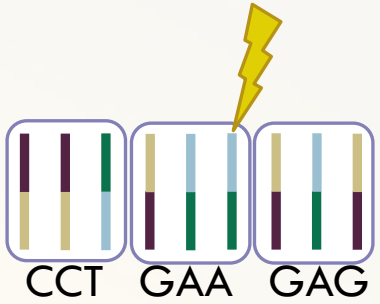




Genotype after a single nucleotide change



Change in amino acid encoded



No change in amino acid encoded

Direct connection to phenotype

dN

dS

Protein from: Al-Haggar et al., Eur J Hum Genet 2012

SELECTIVE PRESSURES

Negative selection (purifying selection)	$\frac{dN}{dS} < 1$	e.g. Highly conserved genes – DNA polymerase
Neutral evolution	$\frac{dN}{dS} = 1$	e.g. pseudogenes
Positive selection (adaptive evolution)	$\frac{dN}{dS} > 1$	e.g. Adaptively evolving proteins – immune genes

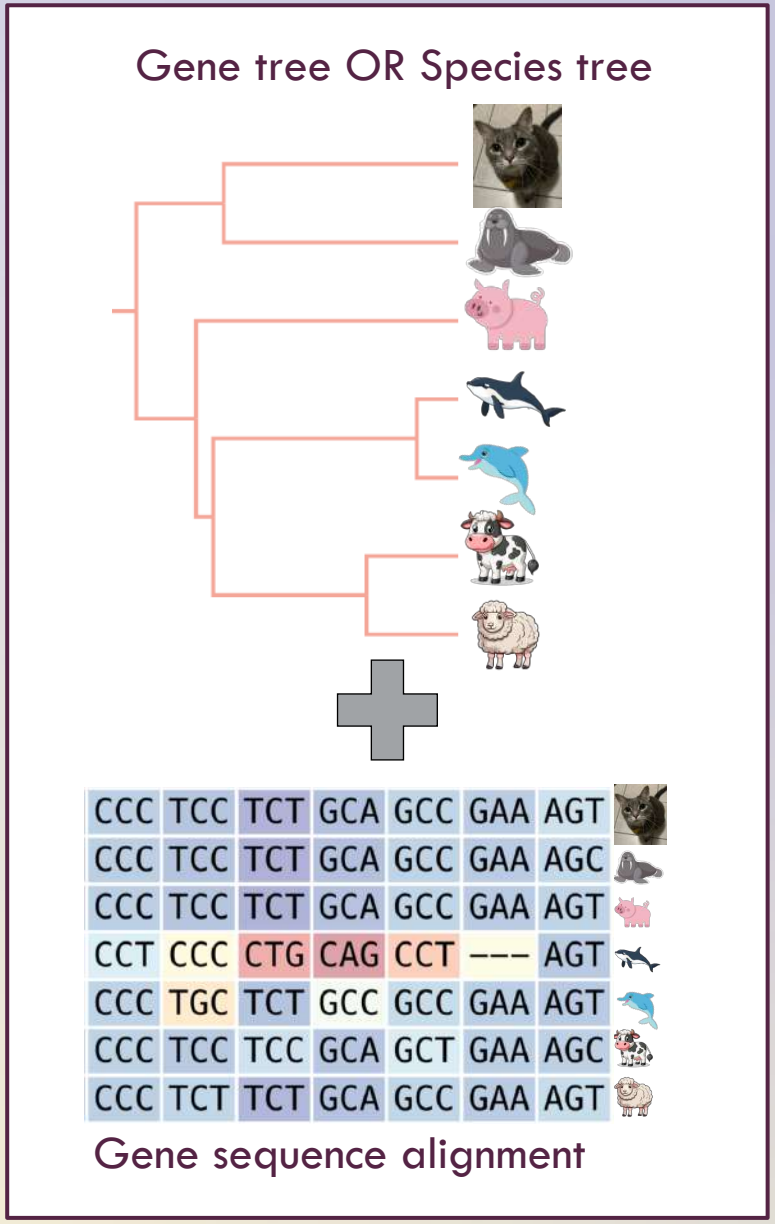
BUSTED methods: estimate dN/dS as ω

The d_N/d_S ratio (ω) measures the selective pressure

d_N = # nonsynonymous substitutions/# nonsynonymous sites
d_S = # synonymous substitutions/# synonymous sites

CODON MODELS

- We use codon models to estimate ω (dN/dS) values
- Finite state continuous time Markov process
- LRT to test for positive selection
- <u>To use:</u>
  - Coding sequence alignment (codon)
  - Tree



dN/dS methods identify protein-coding genes potentially involved in adaptive evolution.

Common pressures leading to positive selection include:

- Pathogens and immunity
- New environments
- Toxins
- Sexual selection

Recommended methods:

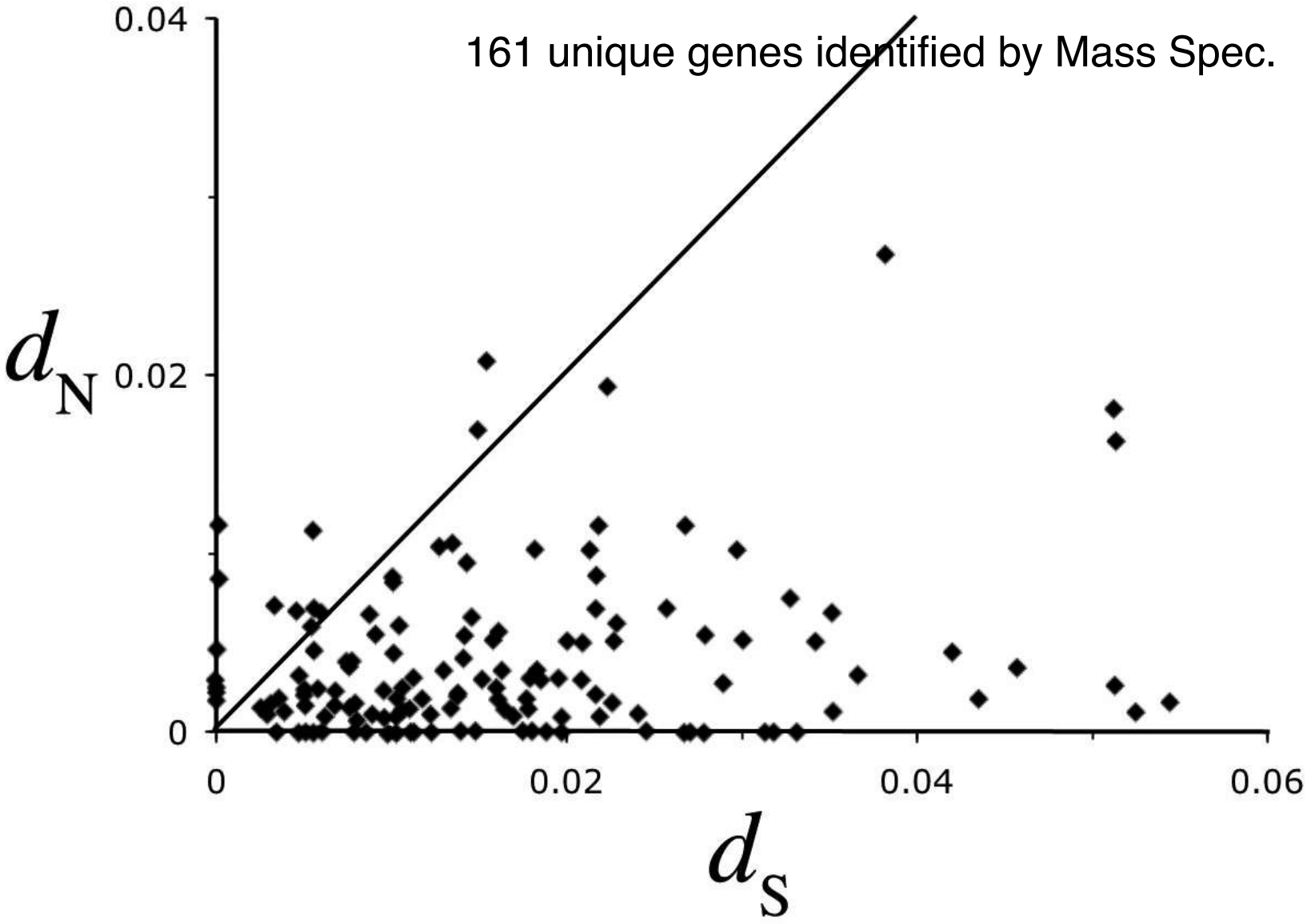
- HyPhy package: specifically BUSTED and BUSTED-E
- PAML: phylogenetic analysis by maximum likelihood

Which primate reproductive proteins participate in sexual selection?

Nine candidates ~11 primate species.

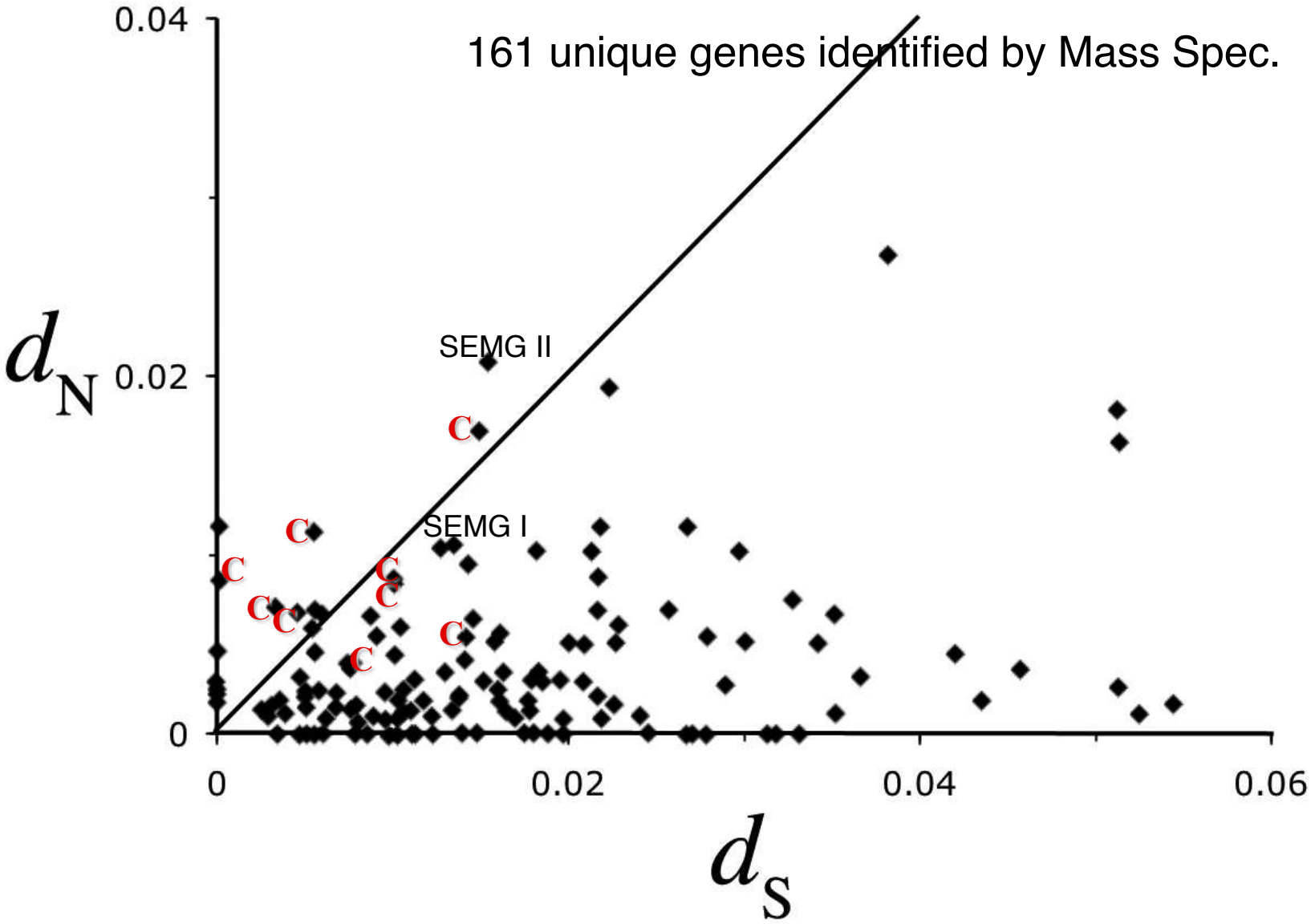
Evolutionary Screen

Human Seminal Genes vs. Chimpanzee



Evolutionary Screen

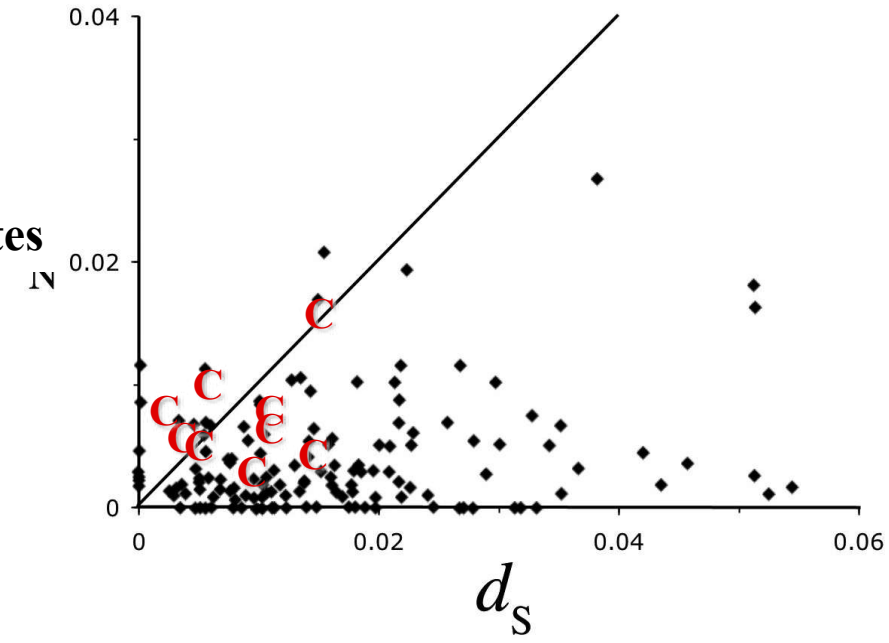
Human Seminal Genes vs. Chimpanzee



Significant Positive Selection in 8 Candidates

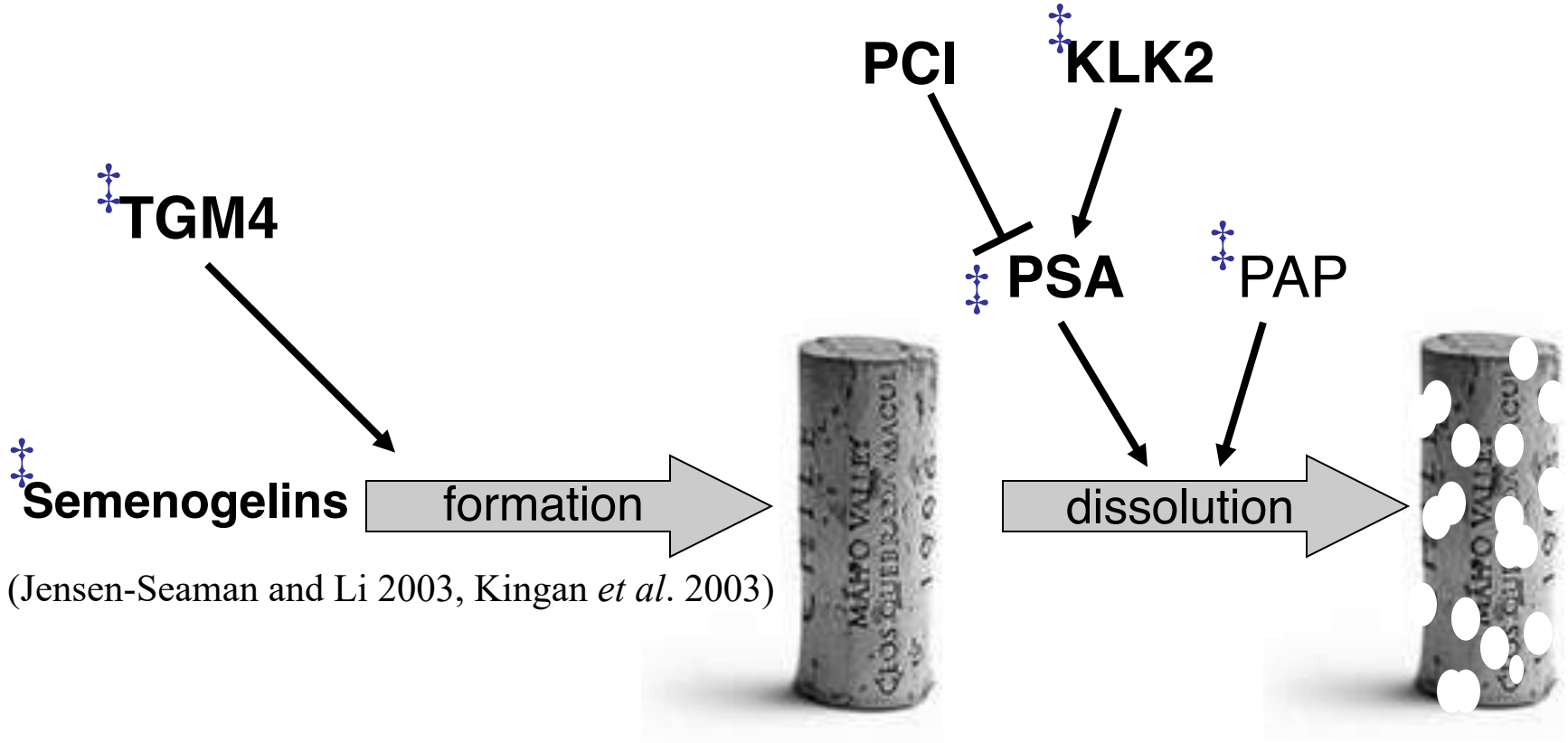
Divergent Analysis

Gene	Human-chimpanzee d_N/d_S	d_N/d_S and %codons	P-value: all species
TGM4 prostate-specific transglutaminase 4	**2.14** (10/2)	**13.9** (0.3%)	0.0018
KLK2 kallikrein 2, prostate-specific	**1.42** (3/1)	**2.11** (24%)	0.0046
ACPP prostatic acid phosphatase	**0.51** (3/3)	**6.47** (4.6%)	0.0072
PSA prostate-specific anitigen	**0.38** (3/3)	**2.94** (6.7%)	**branch-sites**
DBI acyl-coA-binding protein	infinity (2/0)	**2.91** (30%)	0.015
VAT1 synaptic vesicle VAT-1 homolog	**2.05** (7/1)	N/A	NS
PIP prolactin-induced protein	**1.13** (5/2)	**7.56** (25%)	0.000054
TMPRSS2 transmembrane serine protease 2	**0.87** (9/4)	**4.44** (5.4%)	0.00019
MSMB beta-microsemino protein	**0.84** (2/1)	**2.90** (42%)	0.000032



Primate Copulatory Plug Pathway

‡ Significant positive selection in primates.

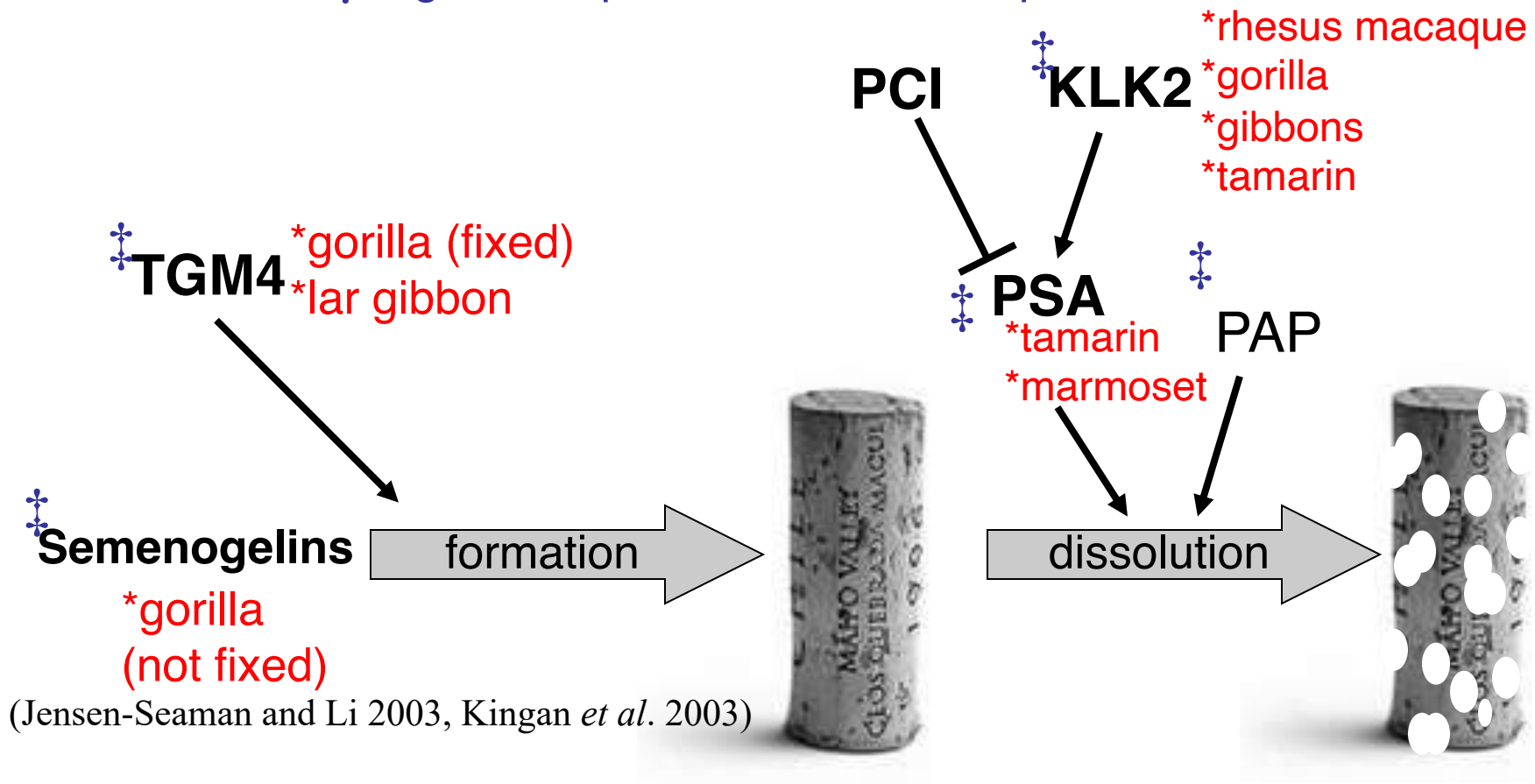


(Jensen-Seaman and Li 2003, Kingan *et al*. 2003)

Primate Copulatory Plug Pathway

* Species showing loss of function at this gene.

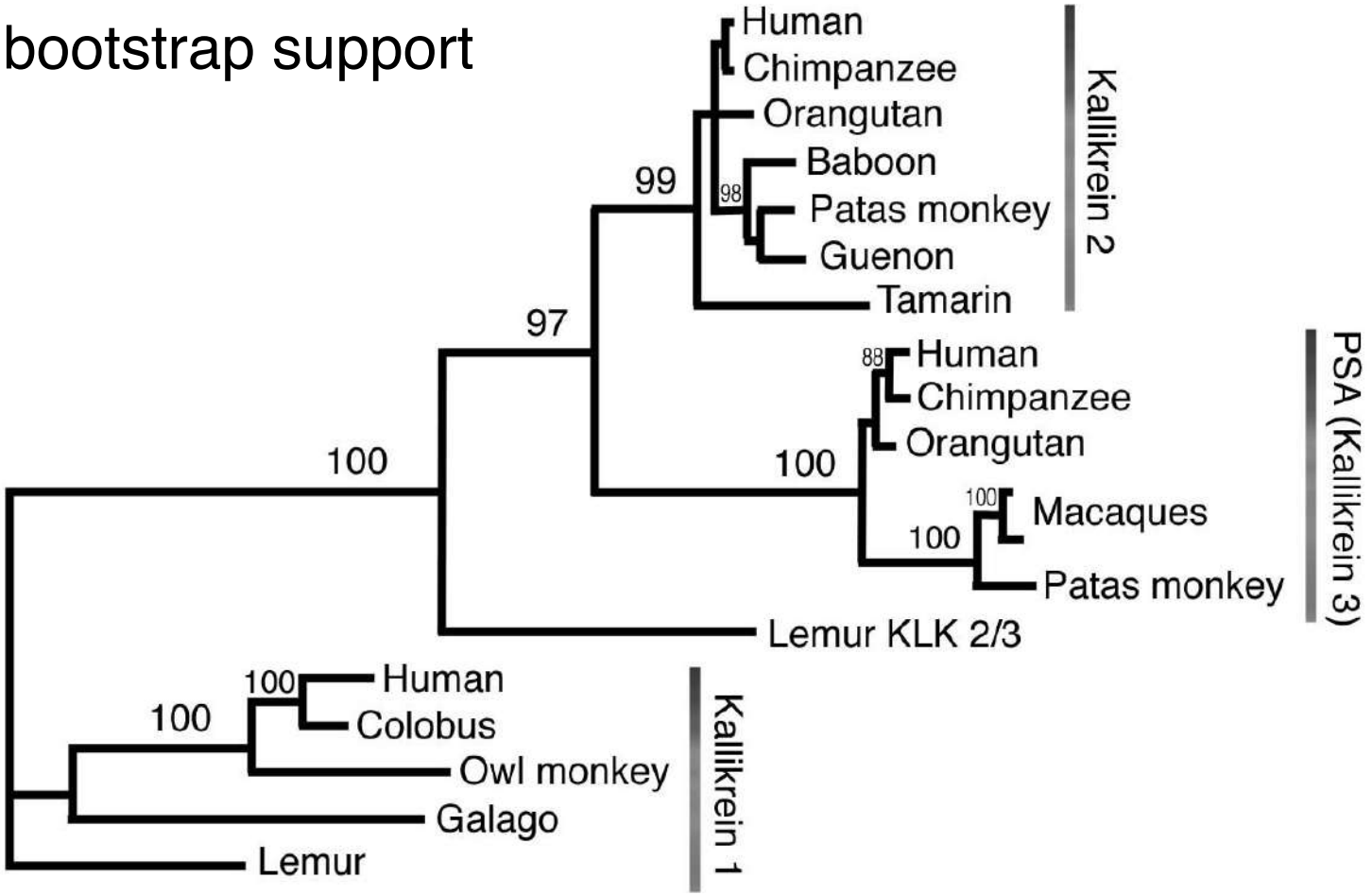
‡ Significant positive selection in primates.



(Jensen-Seaman and Li 2003, Kingan *et al*. 2003)

Simian duplication

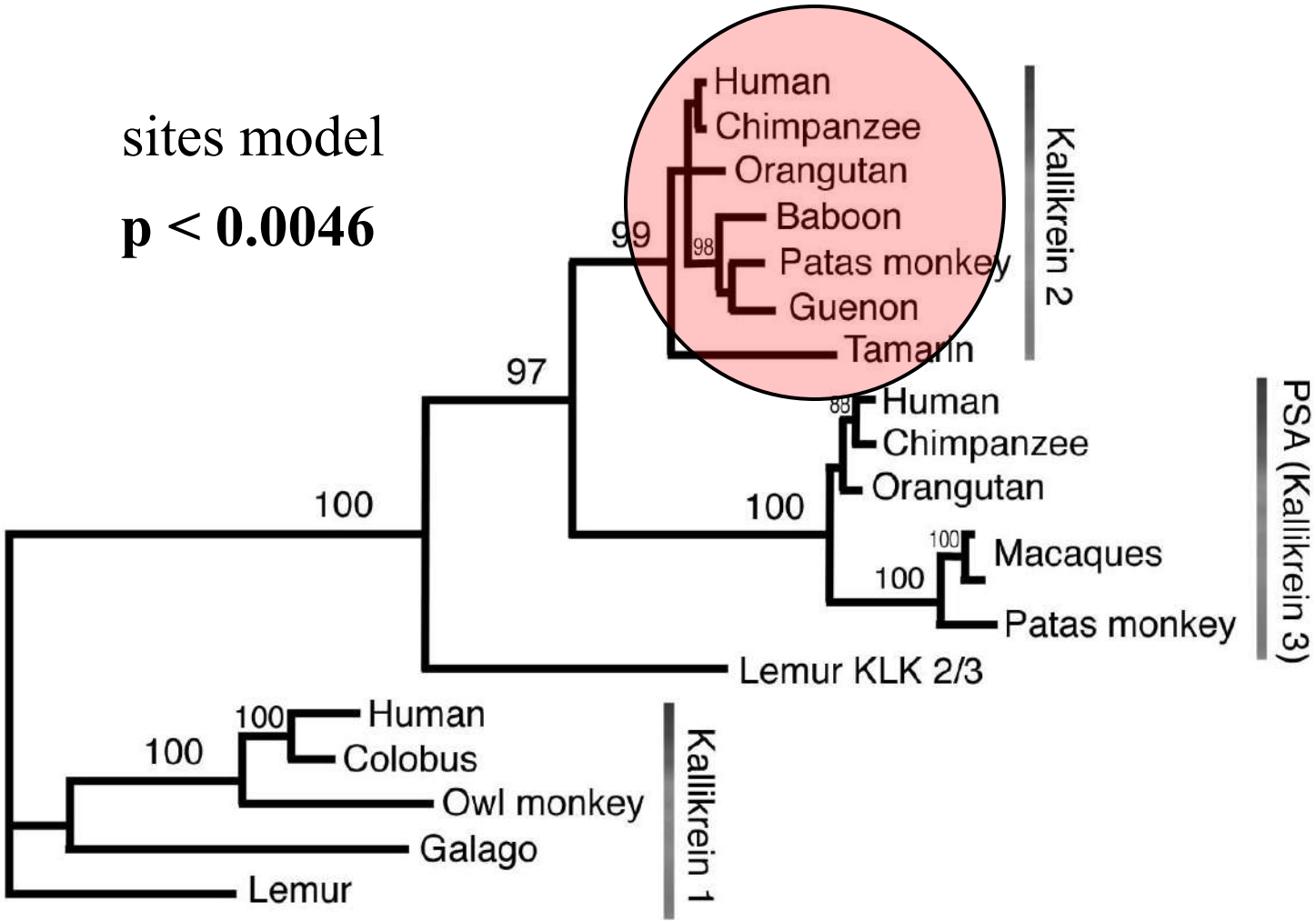
bootstrap support



Simian duplication

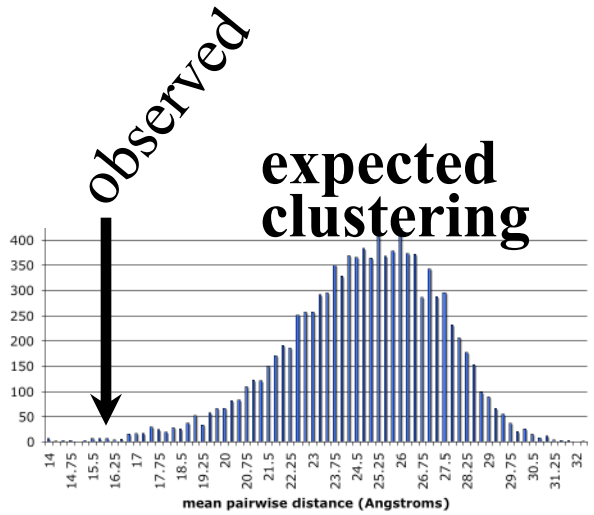
sites model

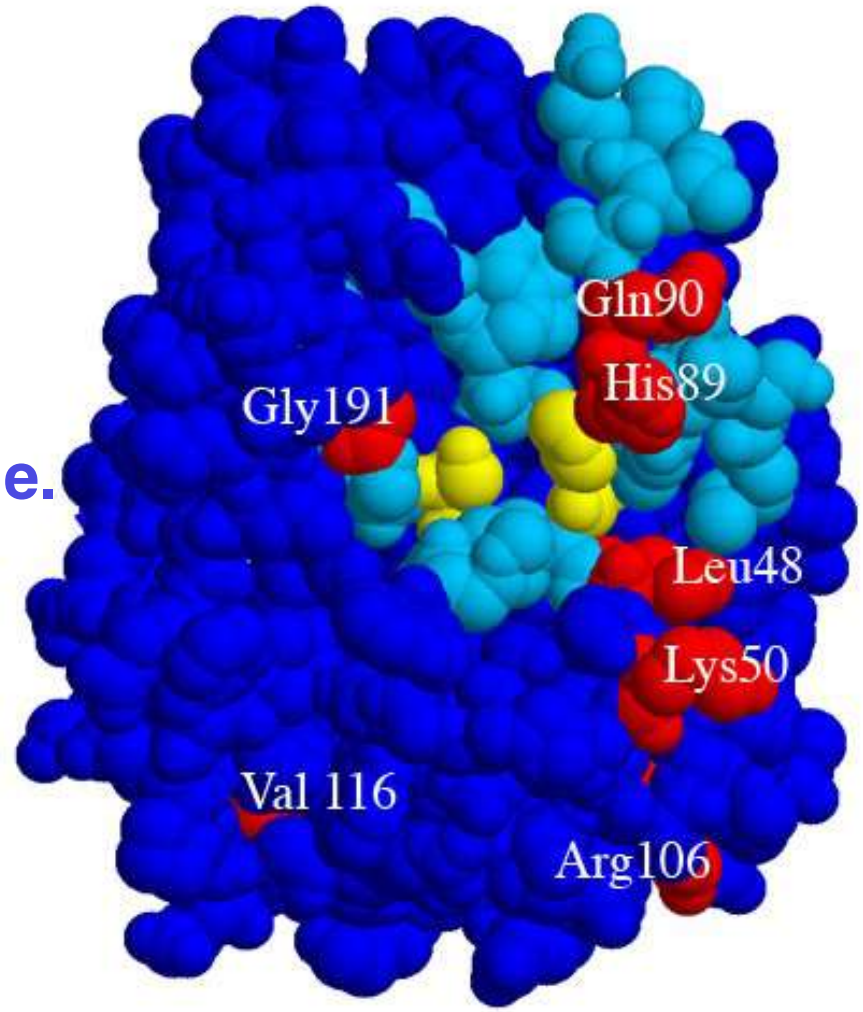
p < 0.0046



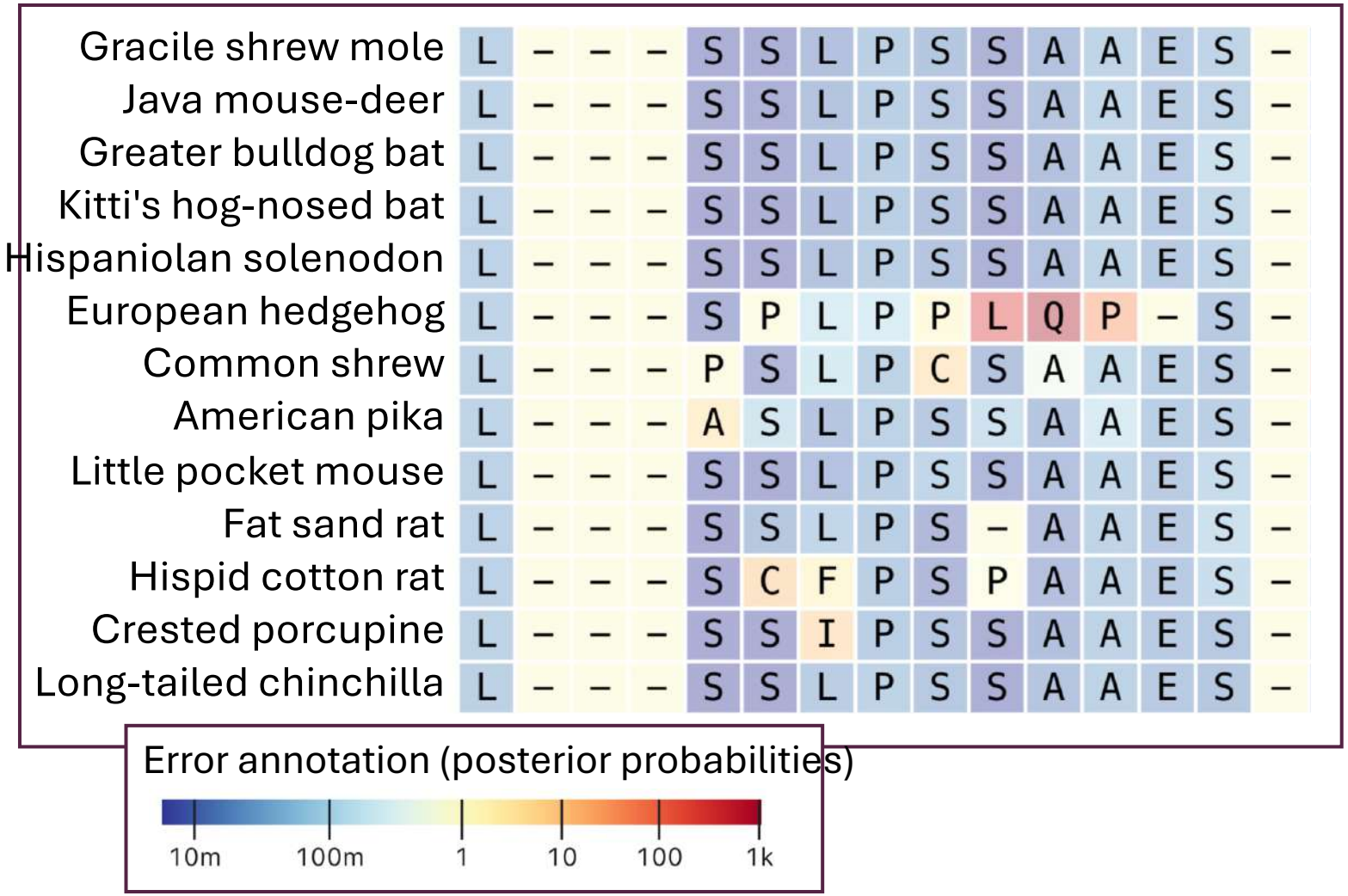
Positively selected sites are non-randomly distributed on the surface of Kallikrein2.

Selected sites in red.
Active sites in yellow.
Substrate binding in light blue.



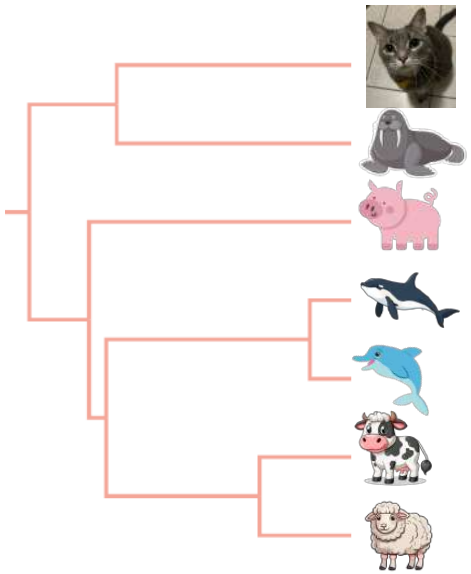


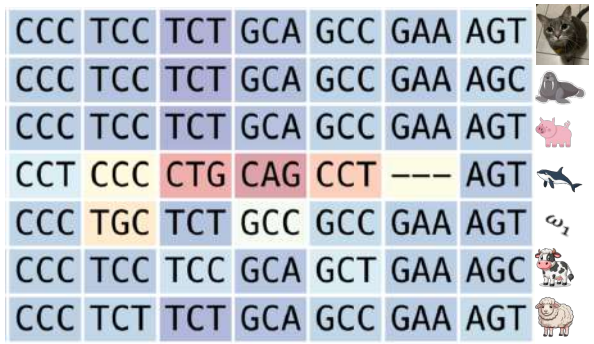
We noticed an alignment problem that was leading to false inferences of positive selection



BUSTED-E: mitigating sequencing and alignment errors in multi-species orthogroups

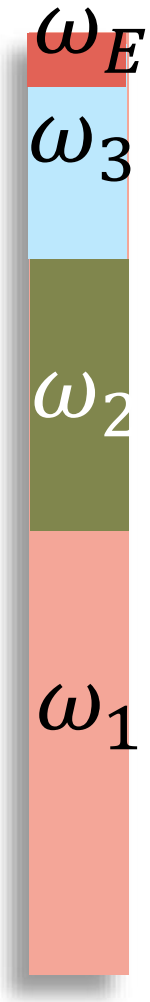
Gene tree OR Species tree





Gene sequence alignment

Estimate ω values



LRT for positive selection

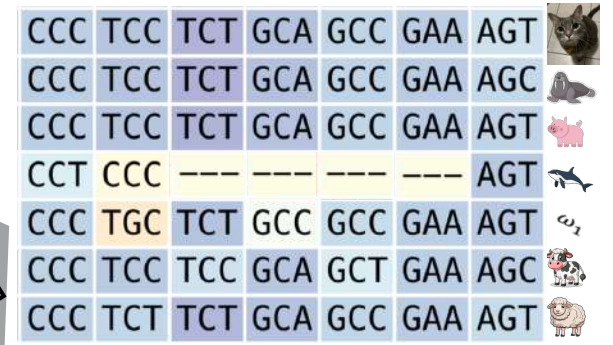
p-values from test for positive selection

Pathway Annotations

Enrichment Analysis

Wilcoxon Rank-Sum

Masked alignment



Masked gene sequence alignment

Avery Selberg, Sergei Pond, Maria Chikina

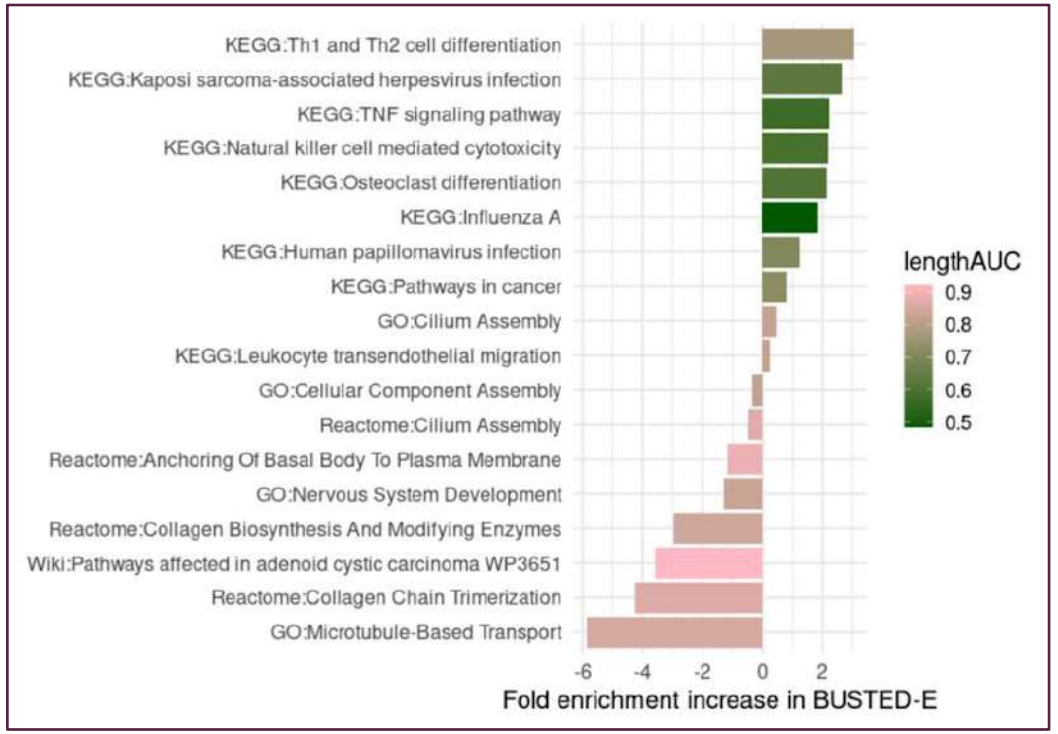
BUSTED-E
Tested with mammalian and bird dataset

BUSTED set of positively selected genes

- 4,704 (42% of total genes)
- No clear pattern, random signal

BUSTED-E set of positively selected genes

- 894 (7.9% of total genes)
- Enrichment of immune genes

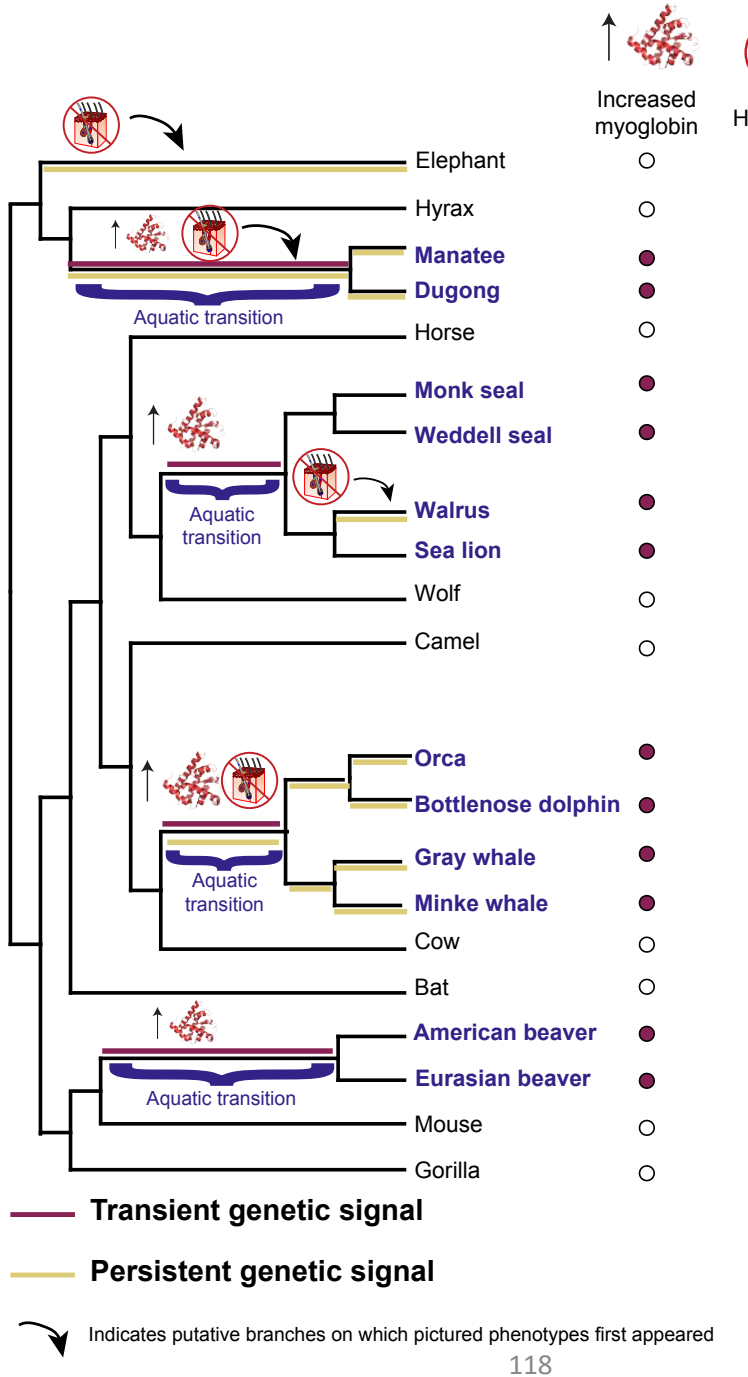


Trait transitions are challenging

It is challenging to identify genetic changes associated with **trait gain**.

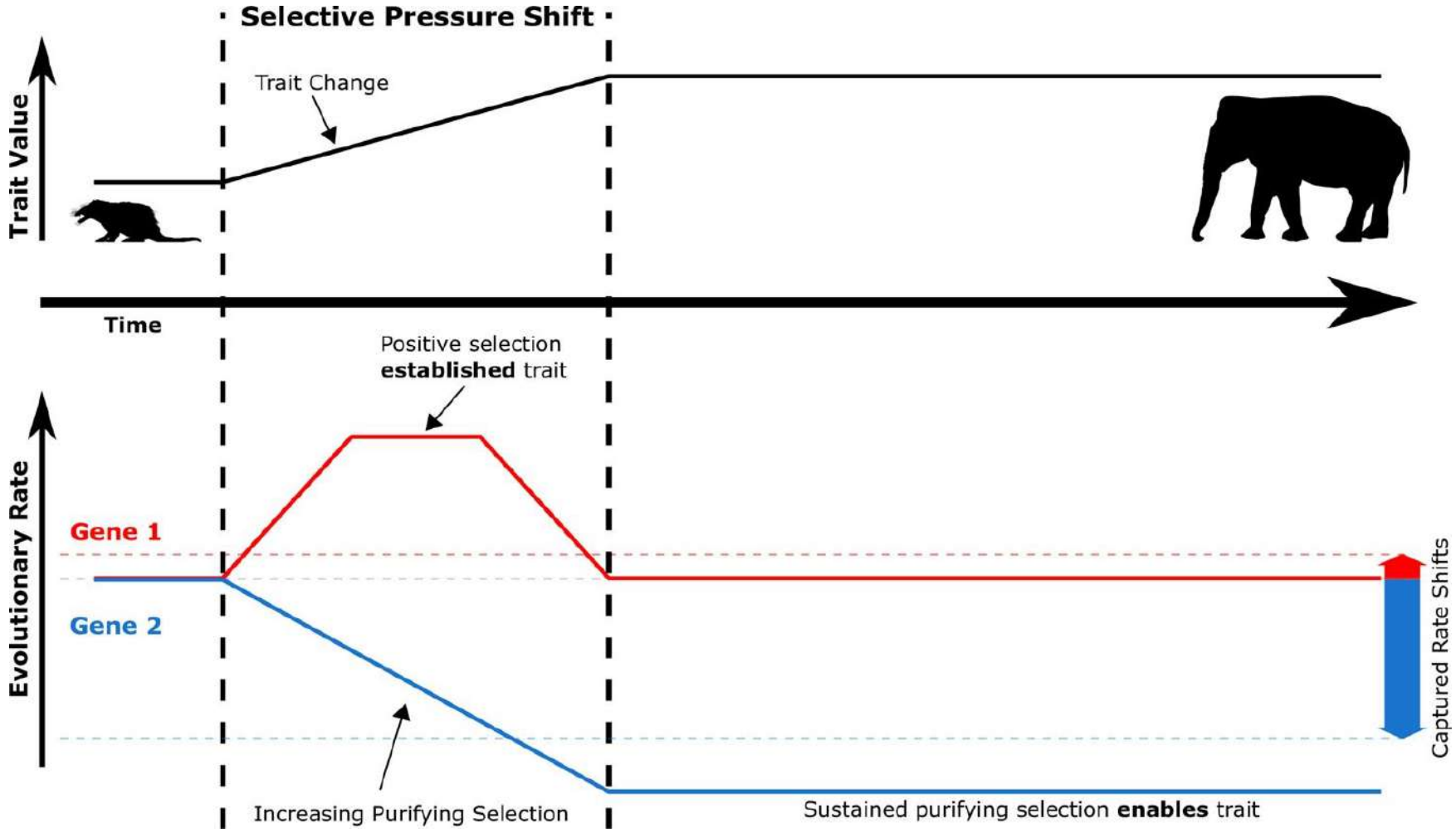
One strategy is scan for positively selection associated with transitional branches.

However, current branch-site tests in PAML and HYPHY do not guarantee positive selection was specific to the foreground branches.



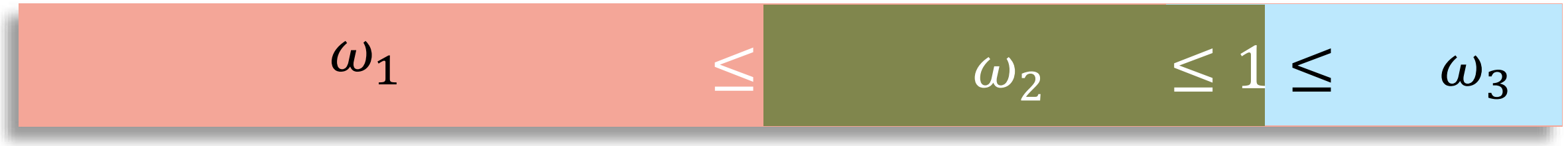
Branch-site Models for d_N/d_S

To determine genes contributing to bursts of adaptive change we need to detect positive selection on specific branches on which the trait transitioned/changed.



BUSTED branch site methods

BUSTED ω classification:

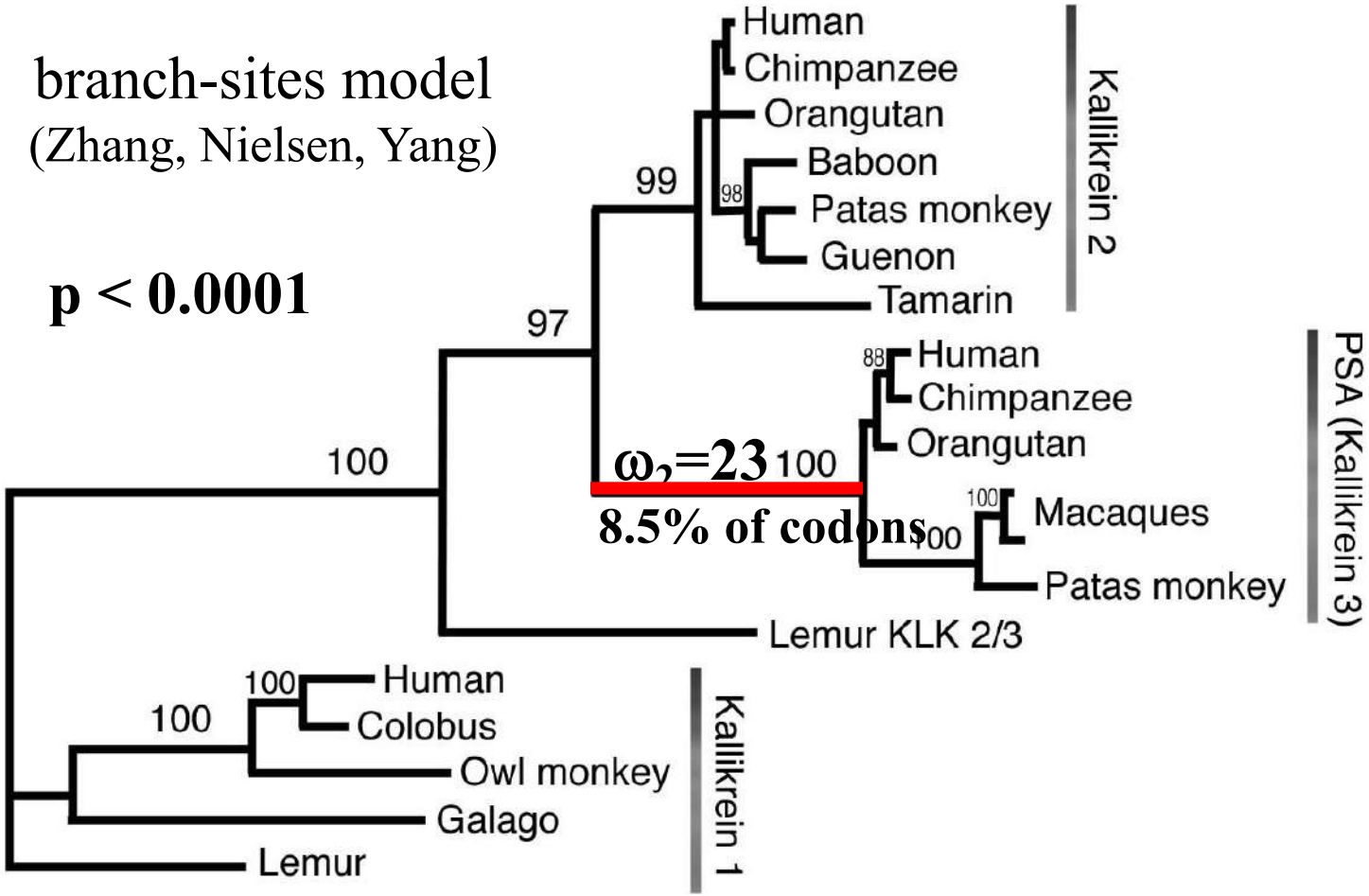


ω varies across both sites and branches (across the entire tree)

Simian duplication

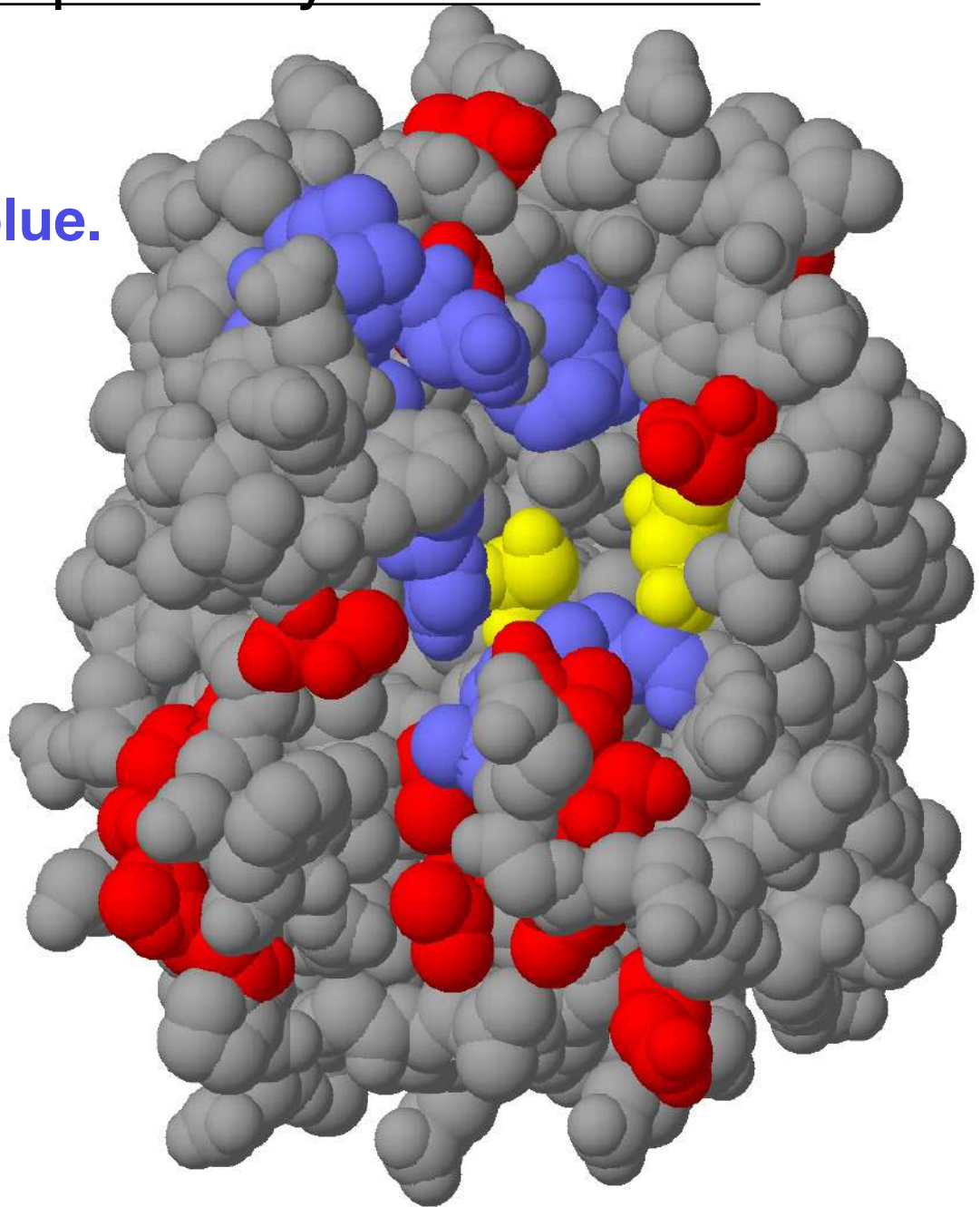
branch-sites model
(Zhang, Nielsen, Yang)

p < 0.0001

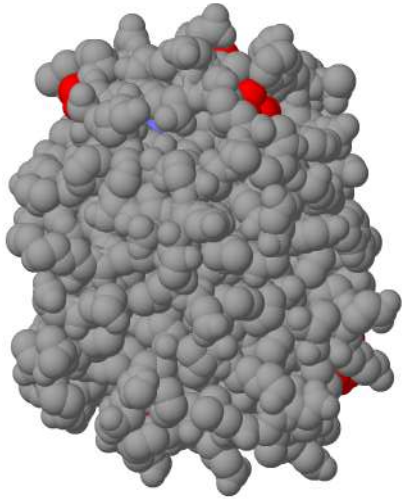


Post-duplication *PSA* positively selected sites

Selected sites in red.
Active sites in yellow.
Substrate binding in light blue.



Rear View



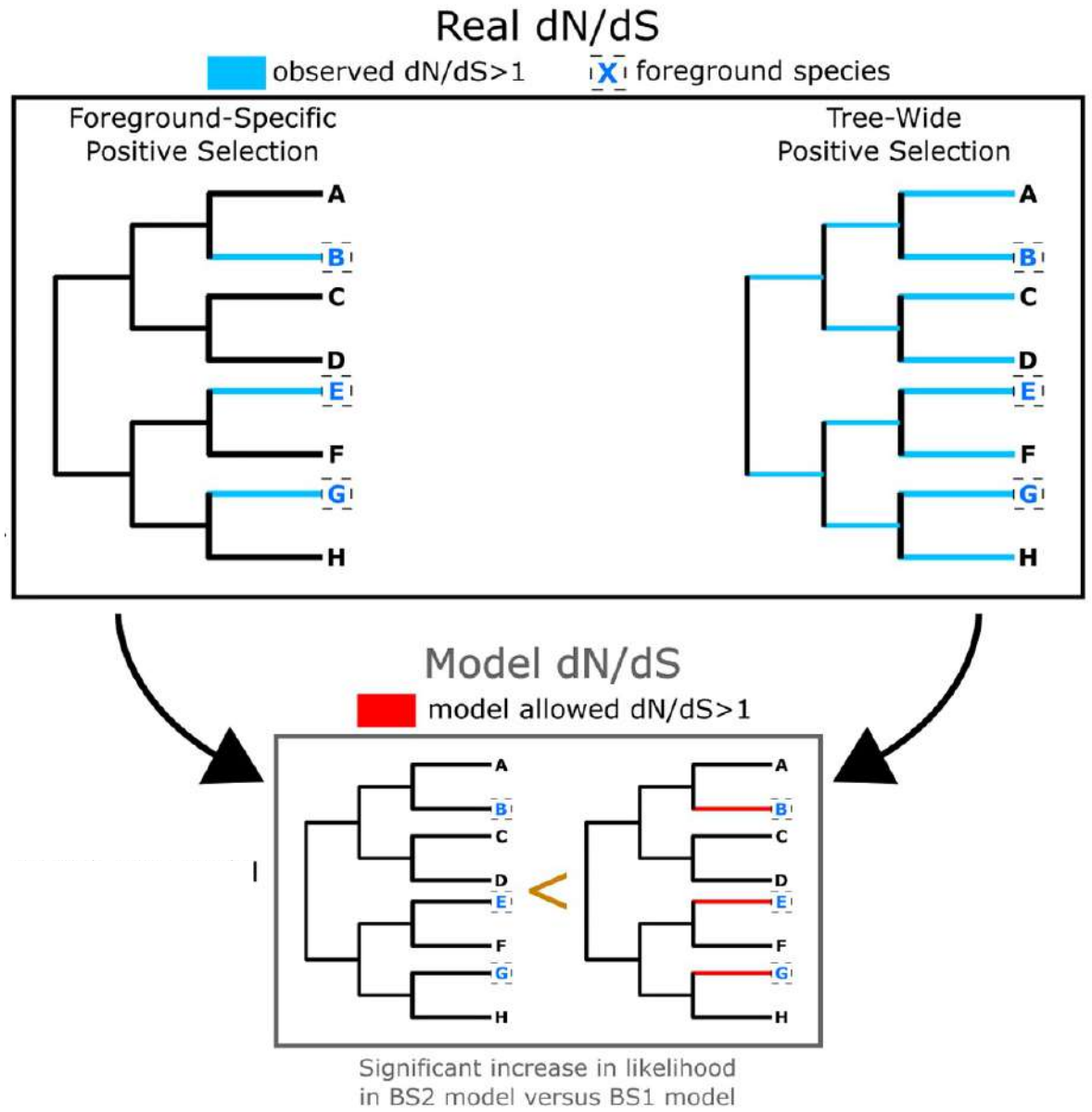
A cautionary tale on proper use of branch-site models to detect convergent positive selection

Amanda Kowalczyk, Maria Chikina, Nathan L Clark

Traditional branch-site models in PAML and BUSTED did not guarantee that positive selection was **specific** to foreground branches.

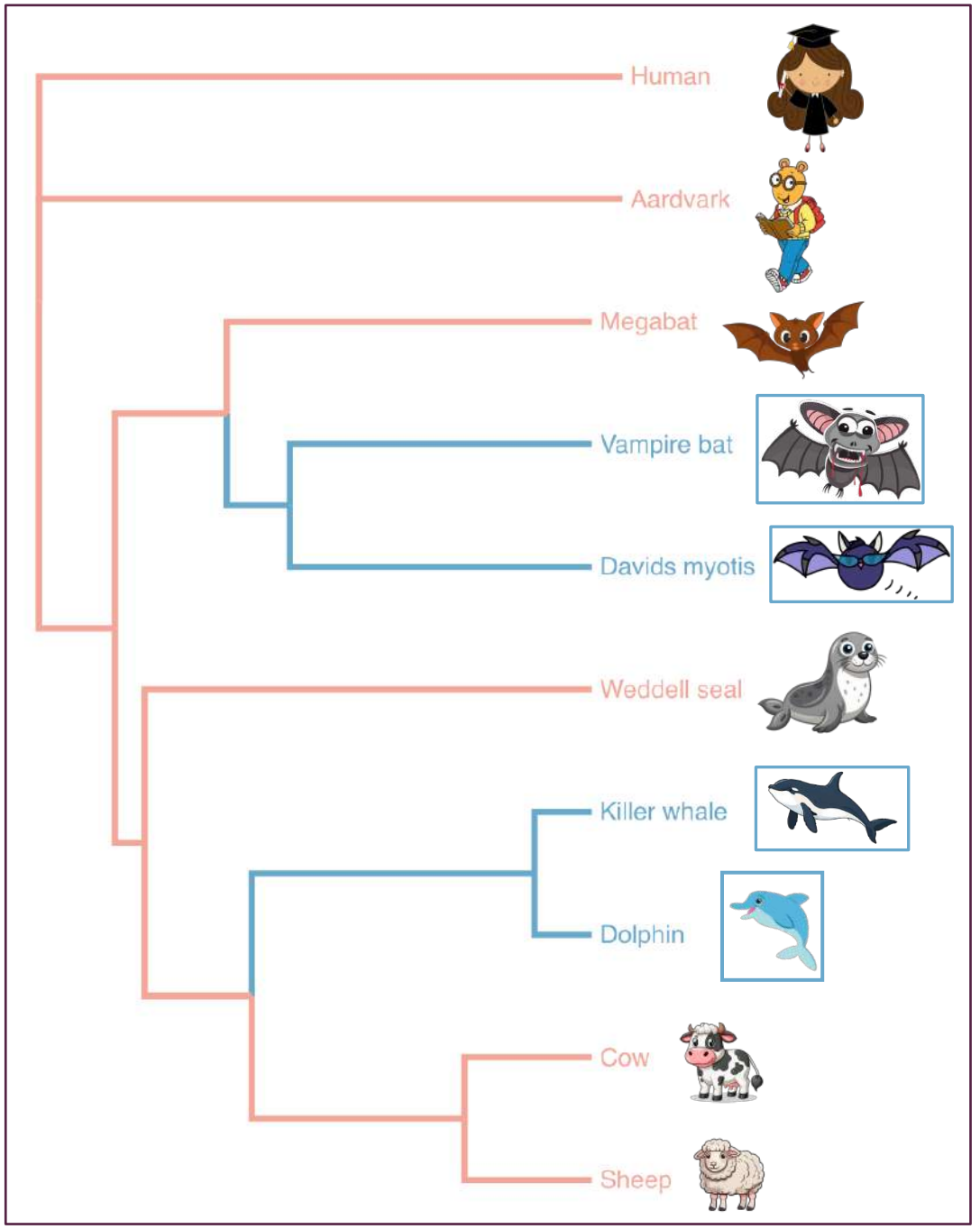
This led to the publication of many false conclusions of positively selected genes for a specific trait or species!

Sergei Pond and our groups developed a solution called BUSTED-PH (Phenotype)



BUSTED-PH

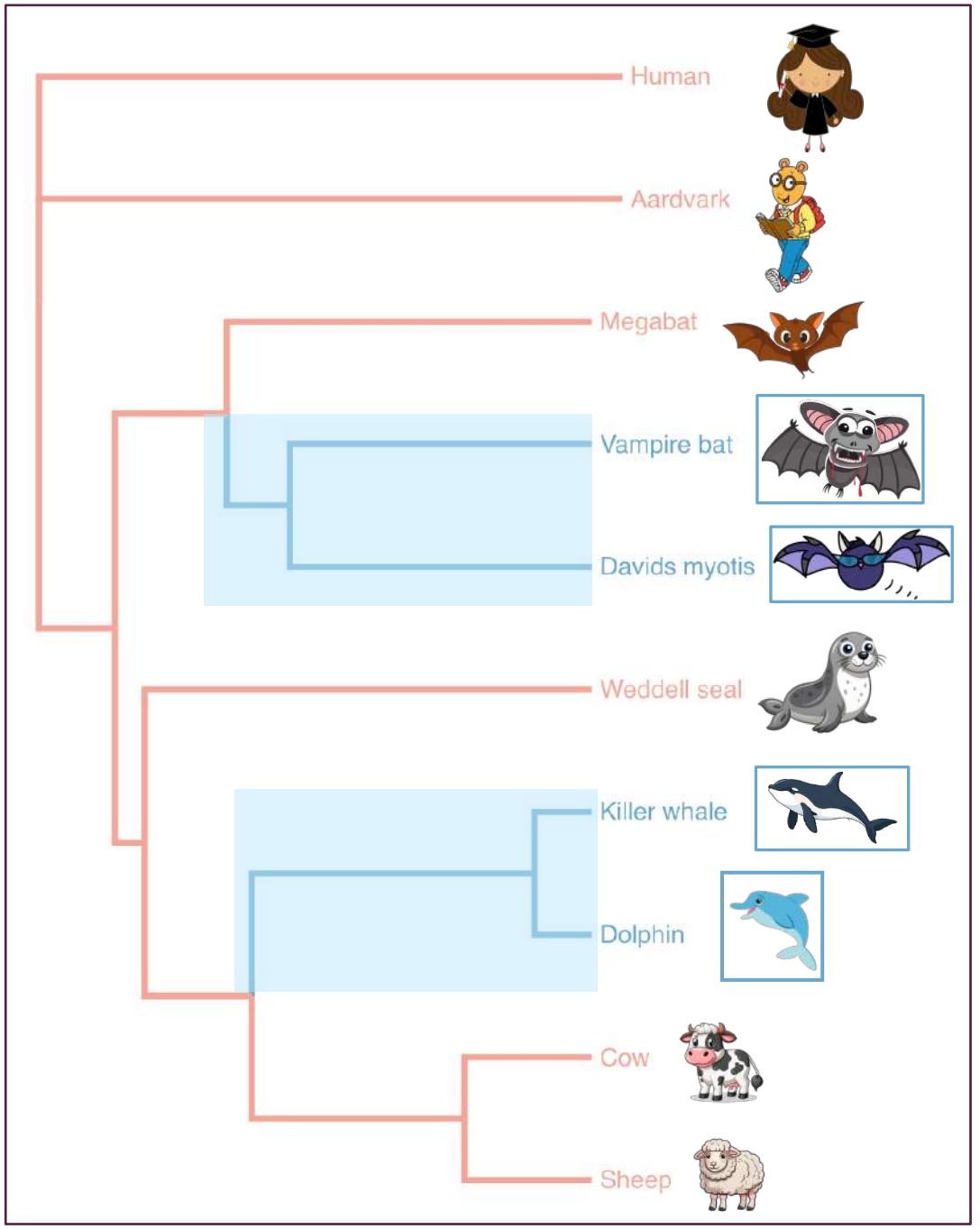
Tests for trait-associated positive selection



BUSTED-PH

Tests for trait-associated positive selection

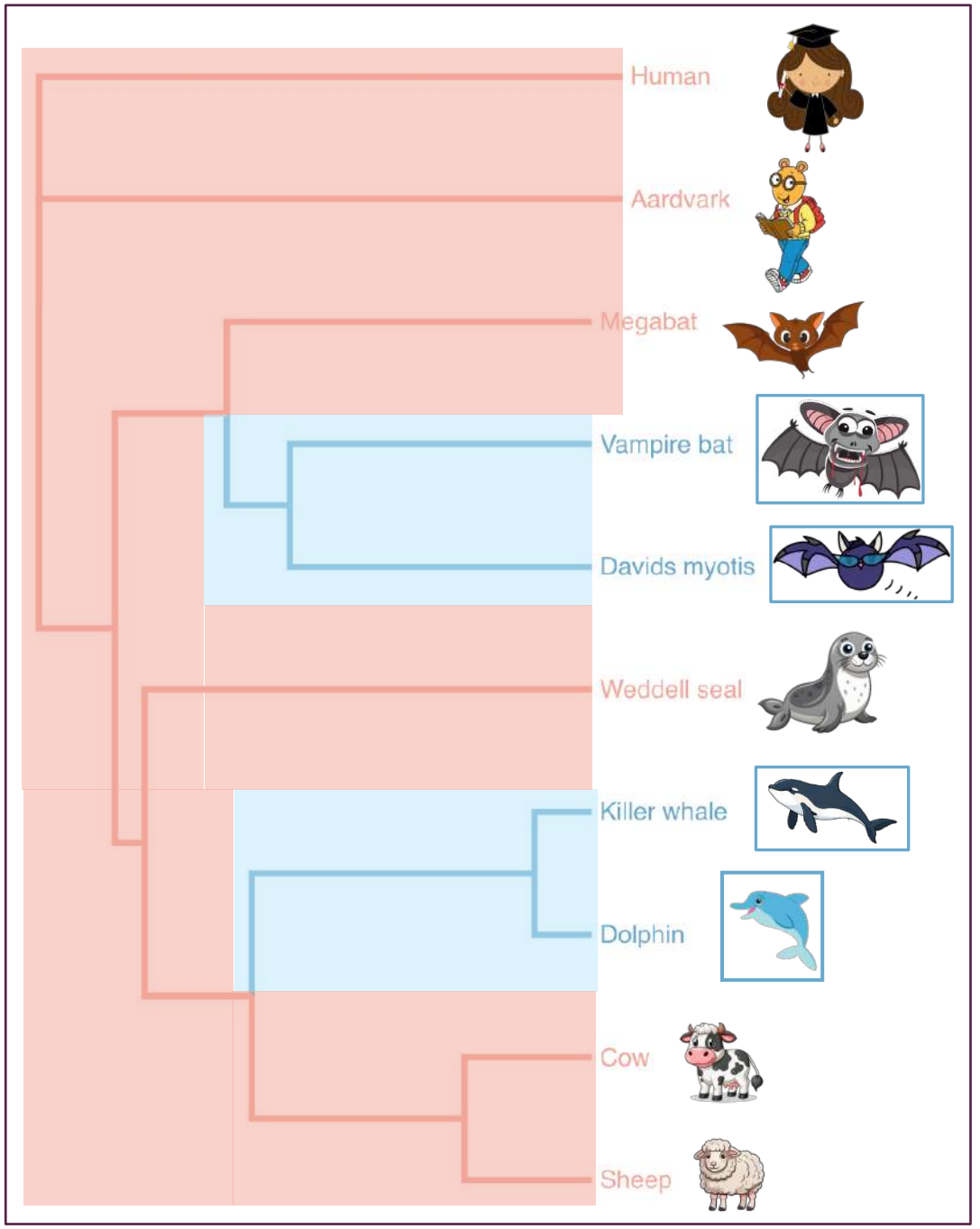
1. Test for evidence of positive selection in foreground branches



BUSTED-PH

Tests for trait-associated positive selection

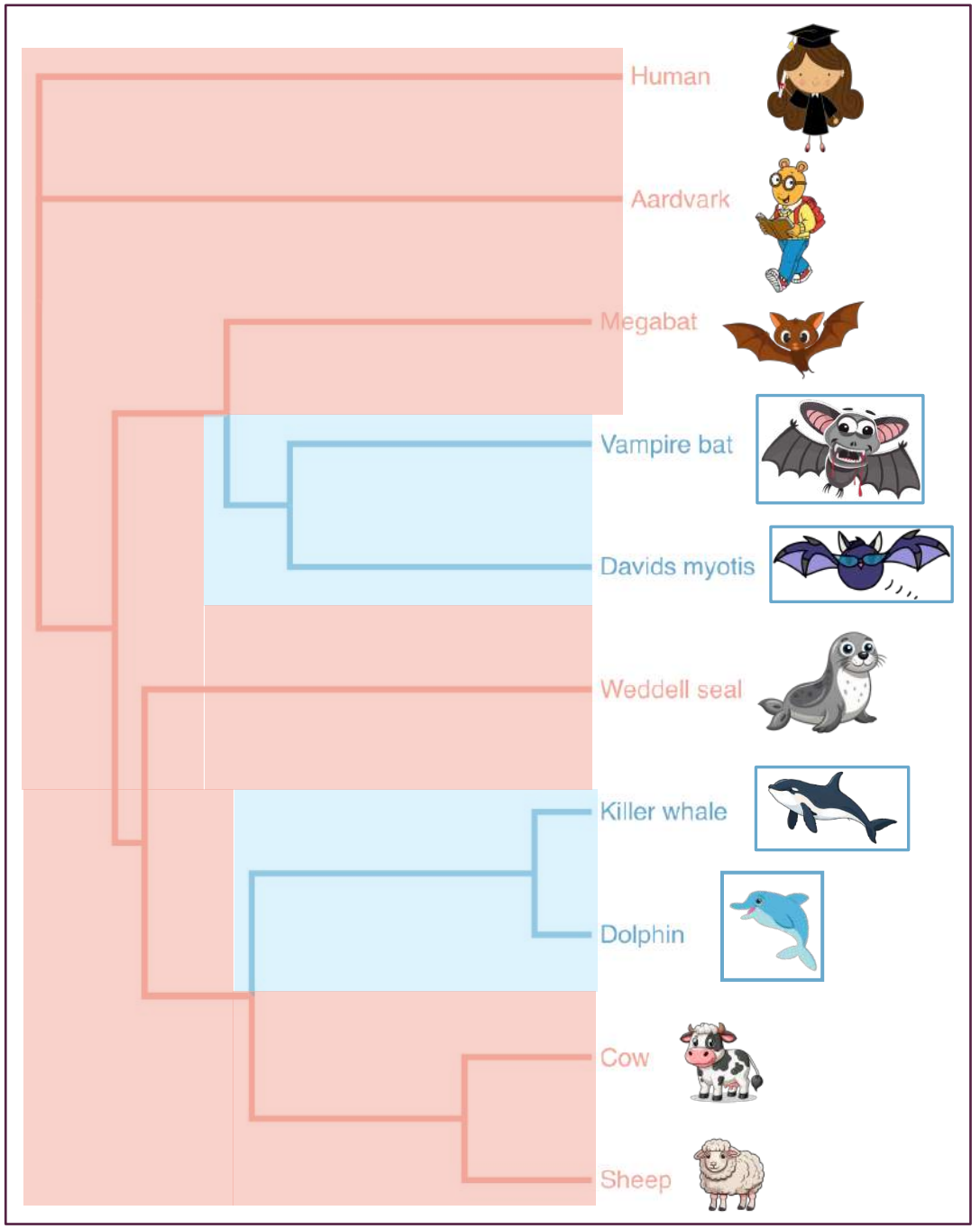
1. Test for evidence of positive selection in foreground branches
2. Test for evidence of positive selection in background branches



BUSTED-PH

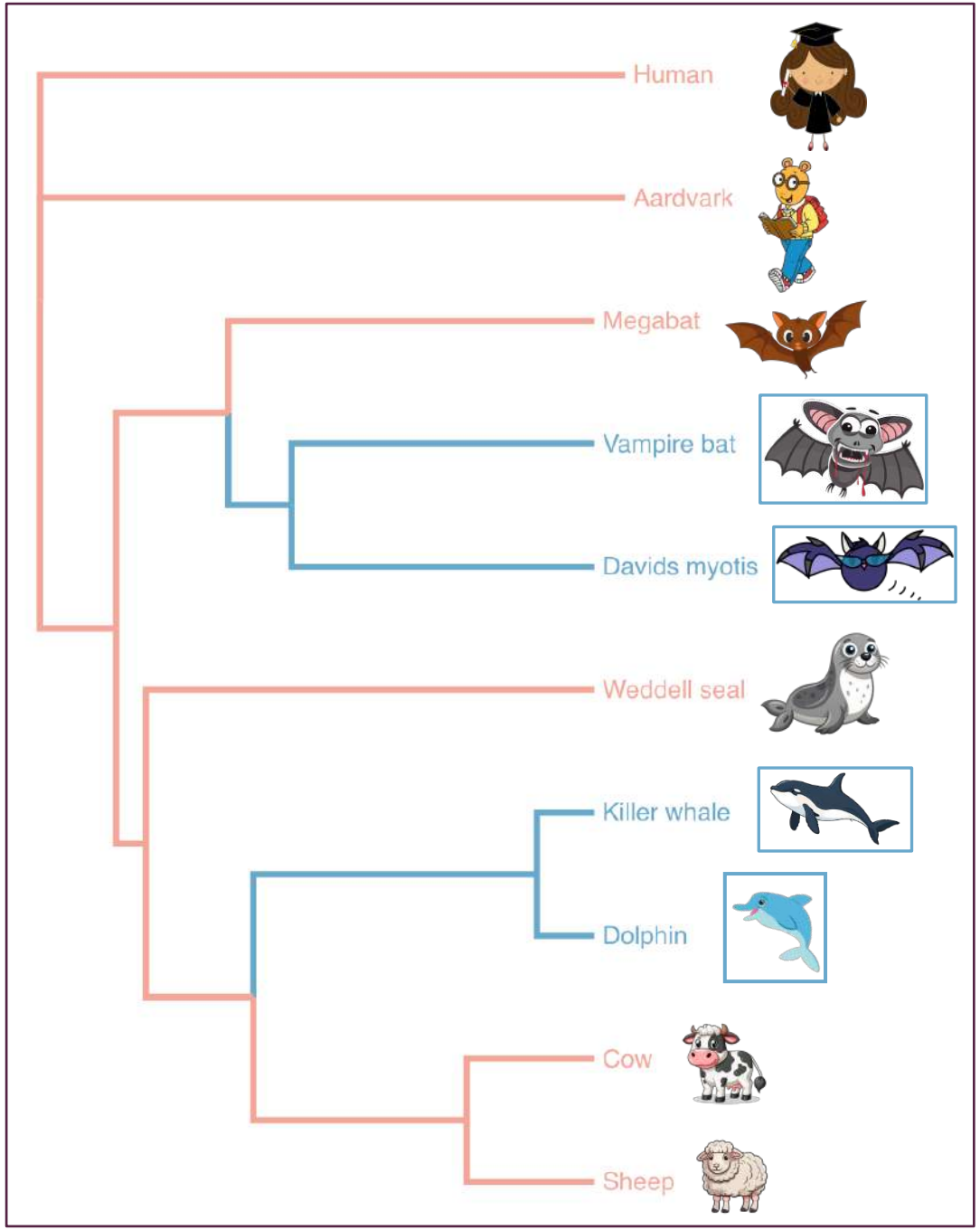
Tests for trait-associated positive selection

1. Test for evidence of positive selection in foreground branches
2. Test for evidence of positive selection in background branches
3. Test for statistically significant difference between foreground and background selective pressures



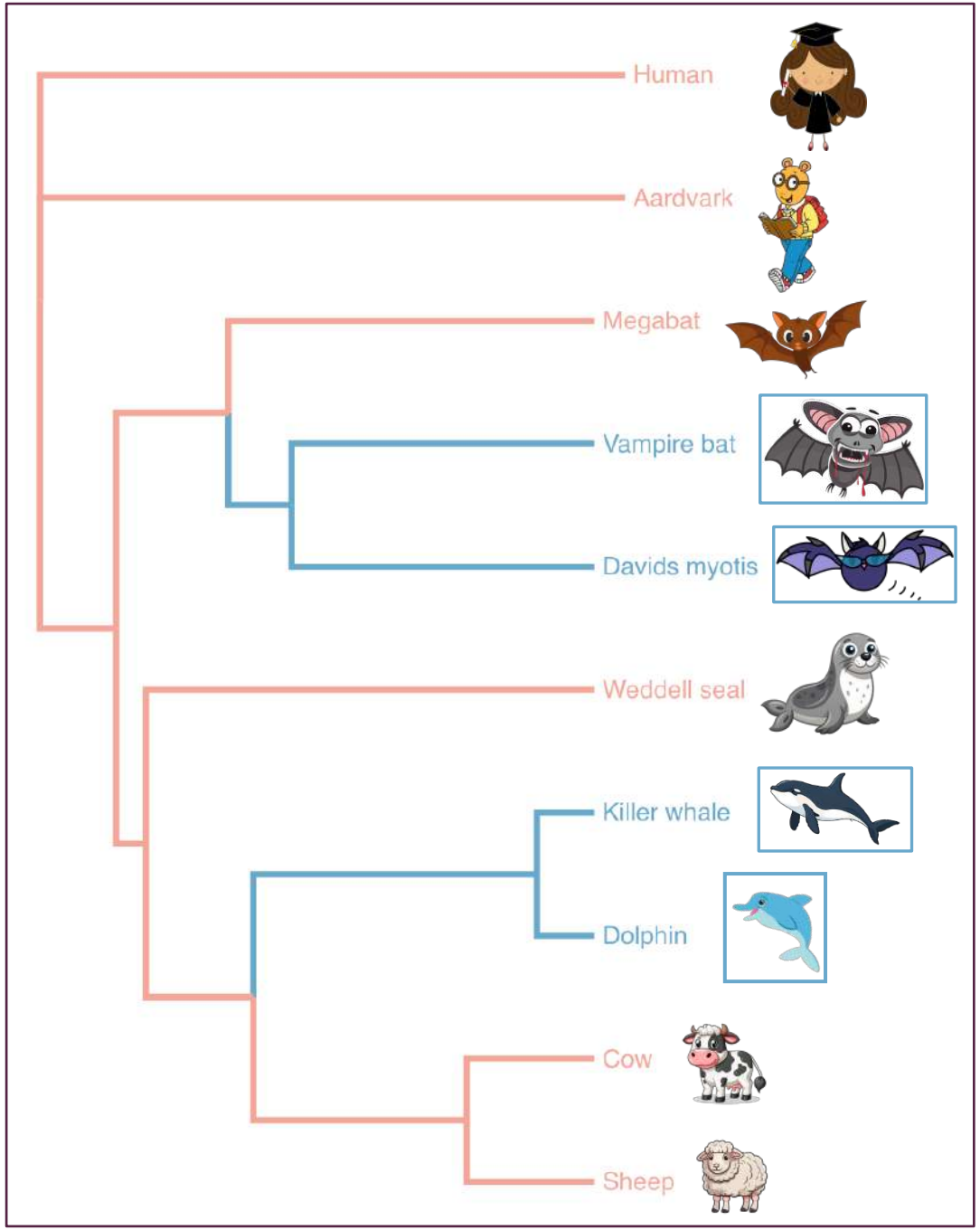
Detecting trait-associated positive selection

- Test BUSTED-PH genome-wide for echolocation-associated selection
- Analyze 18,940 orthologous genes from Nathan Clark, Nikolai Hecker and Michael Hiller
- Build on Allard et al. (2025), which focused on directional selection



Echolocation BUSTED-PH results:

240 trait-associated positively selected genes



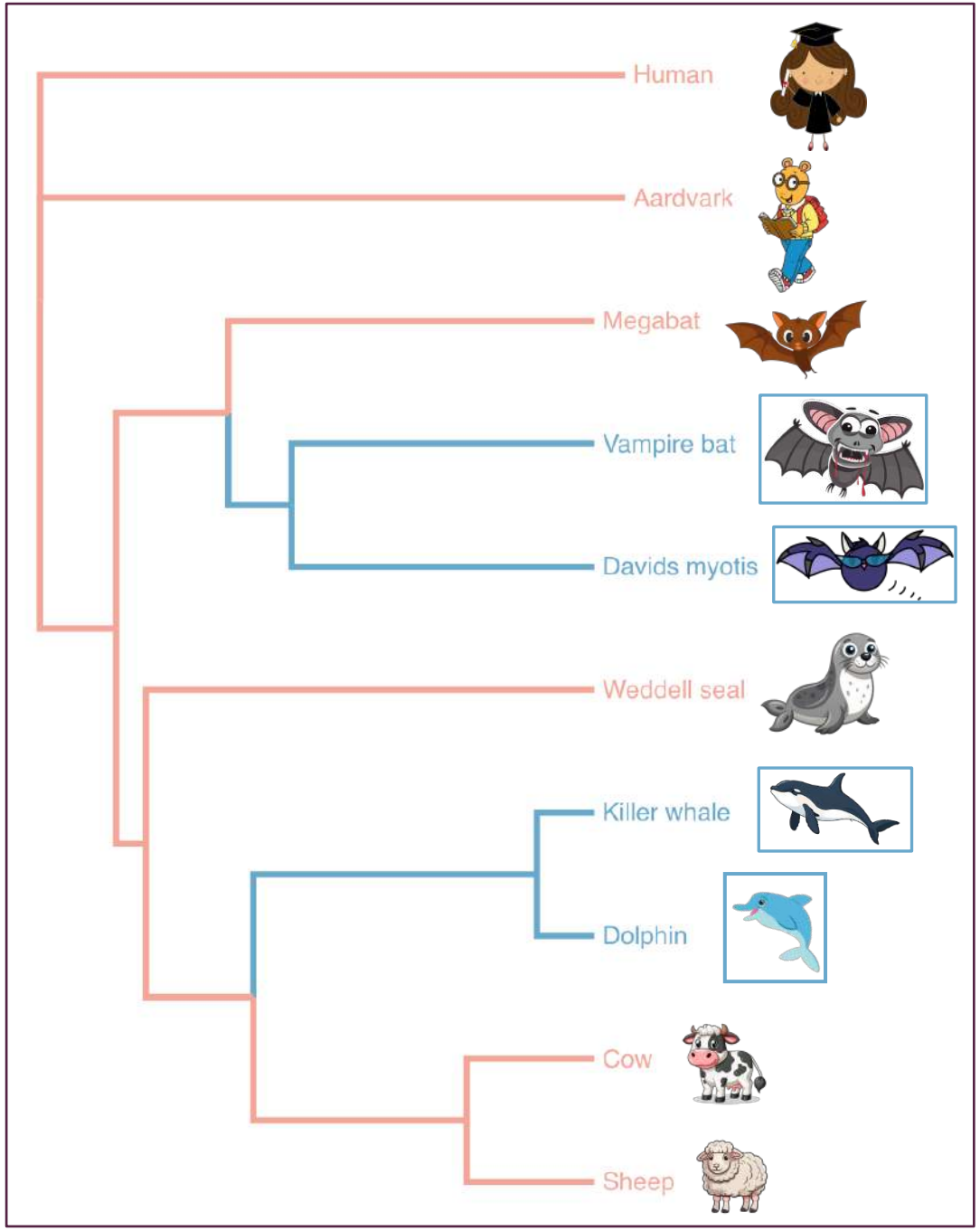
Echolocation BUSTED-PH results

Enriched for Gene Ontologies related to

- Sensory perception

Highly expressed in tissues related to

- Outer hair cells (inner ear)



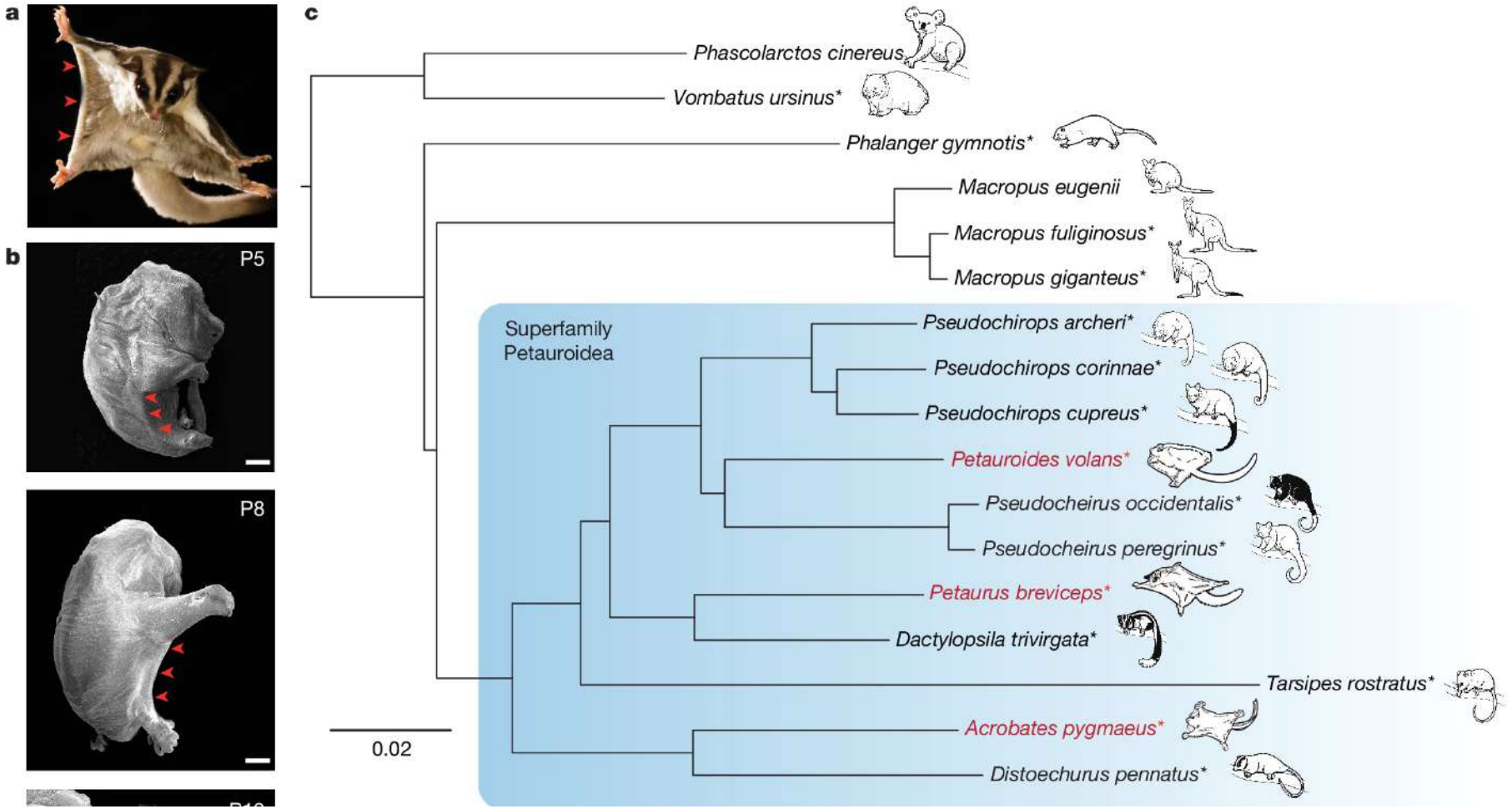
Studying Regulatory Region using PhyloG2P

- Rates-based methods have already been applied, with some success
  - phyloP, Forward Genomics, REforge, RERconverge, PhyloACC

  - When applied genome-wide to conserved non-coding elements, high-scoring elements are enriched near genes relevant to the trait, or in tissue-relevant open chromatin (ATAC-seq) regions.
  - However, the enrichment is weak and is rarely informative enough to justify substantial experimental validation.

Emx2 underlies the development and evolution of marsupial gliding membranes

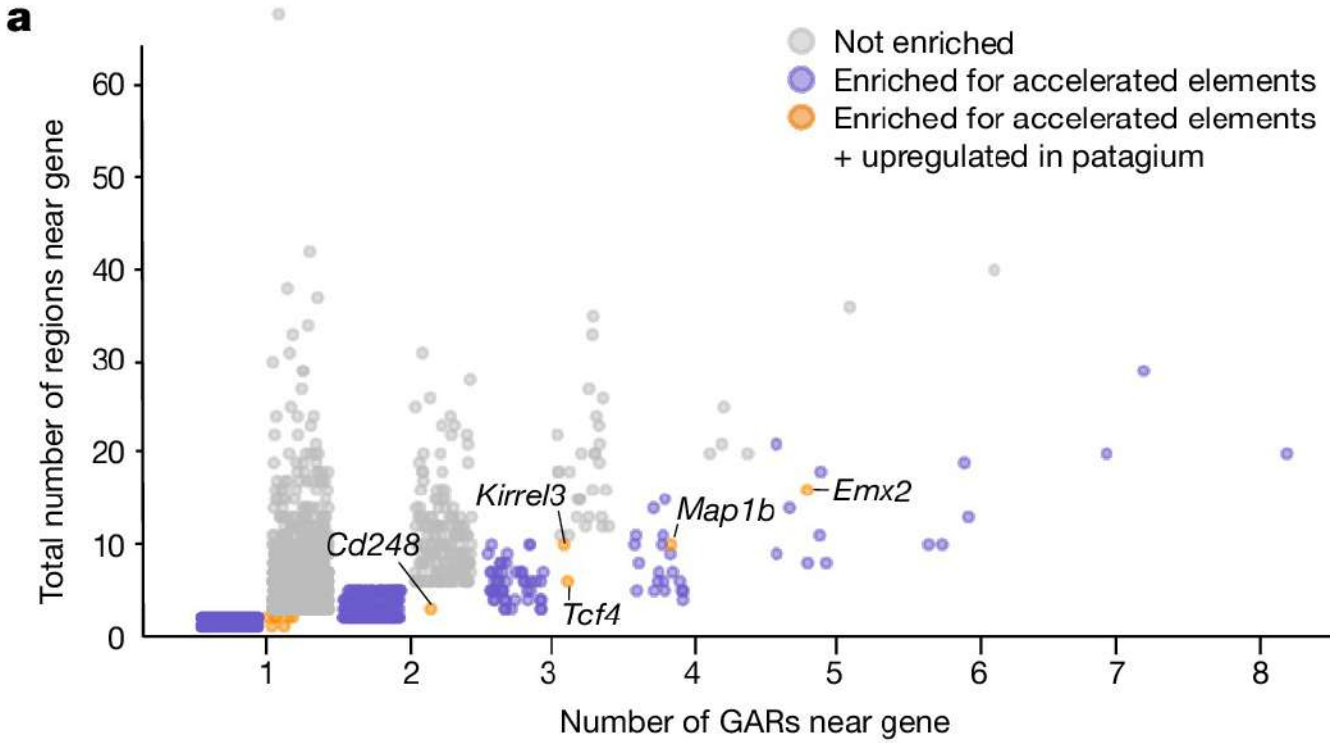
Moreno … Mallarino. *Nature* 2024

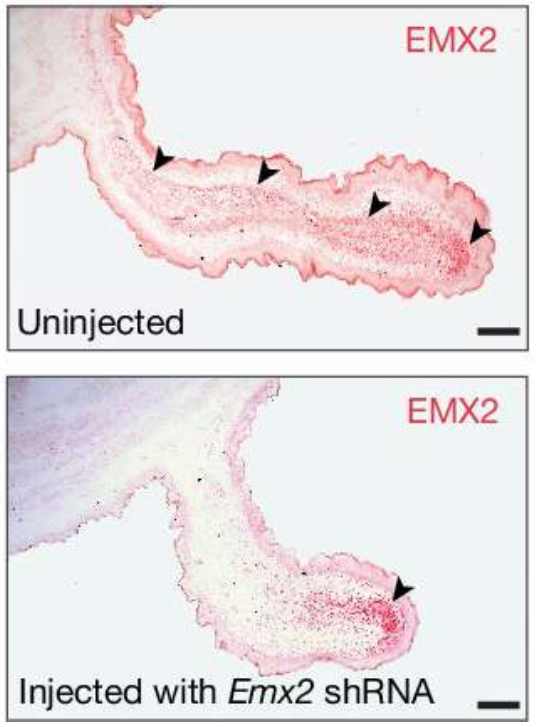


Convergent evolution of gliding membranes

Changes in enhancers near *Emx2* identified contributing to gliding membrane development

- Identified open chromatin (ATAC-seq) in gliding membrane primordial tissues.
- Used **phyloP** to identify putative enhancers accelerated in gliding species.
- Knocking down Emx2 in membrane regions led to smaller membranes!

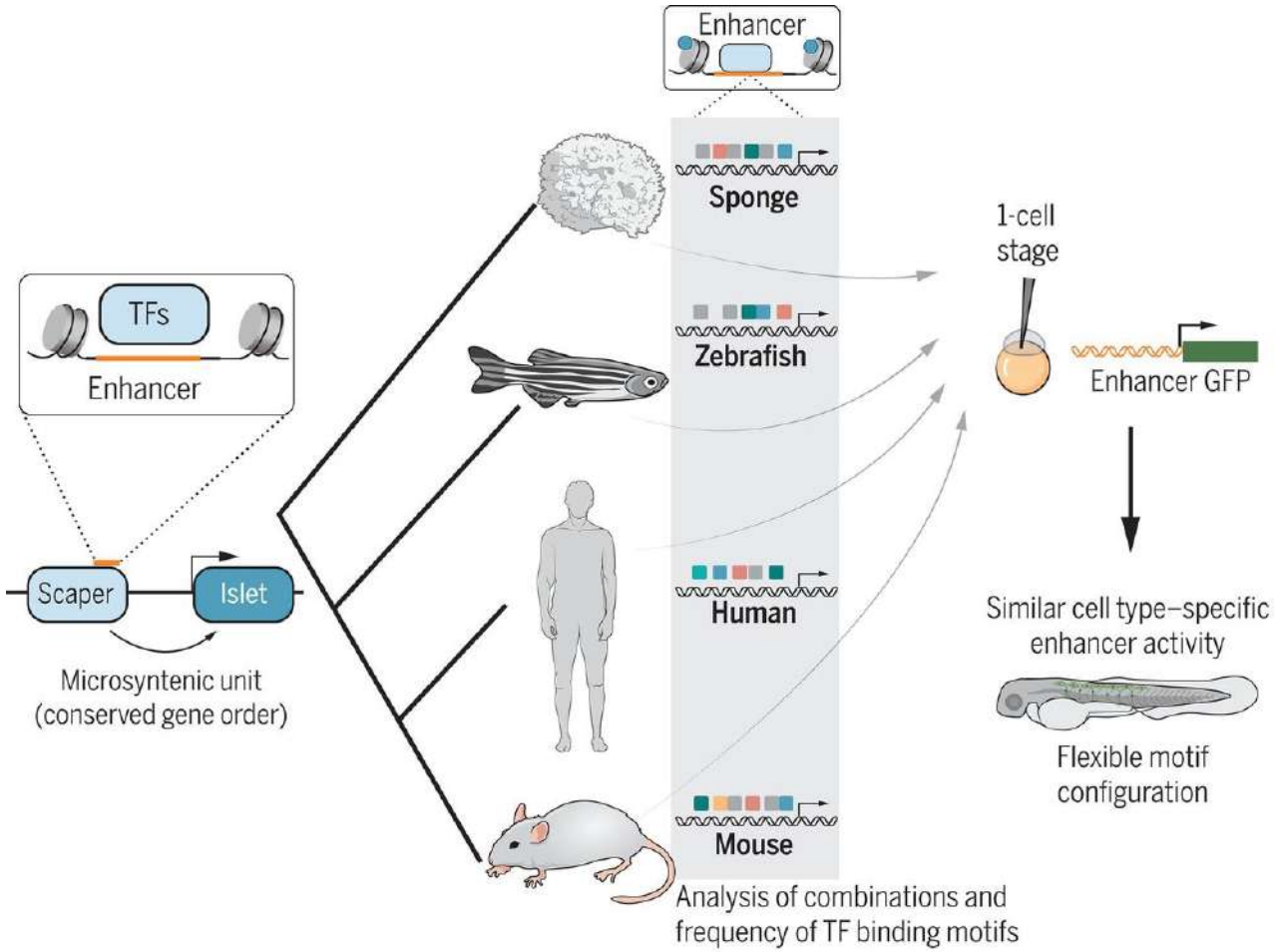




The Challenge with Regulatory Regions is that **Transcription Factor Sites Turn Over**

- Alignment methods are perhaps inadequate.

- Alignment-free methods:
  - Study transcription factor binding sites or motifs, themselves
  - Use machine learning models of chromatin states (*e.g.*, Sei) and gene expression (*e.g.*, Borzoi)



Wong et al. *Science* 2020

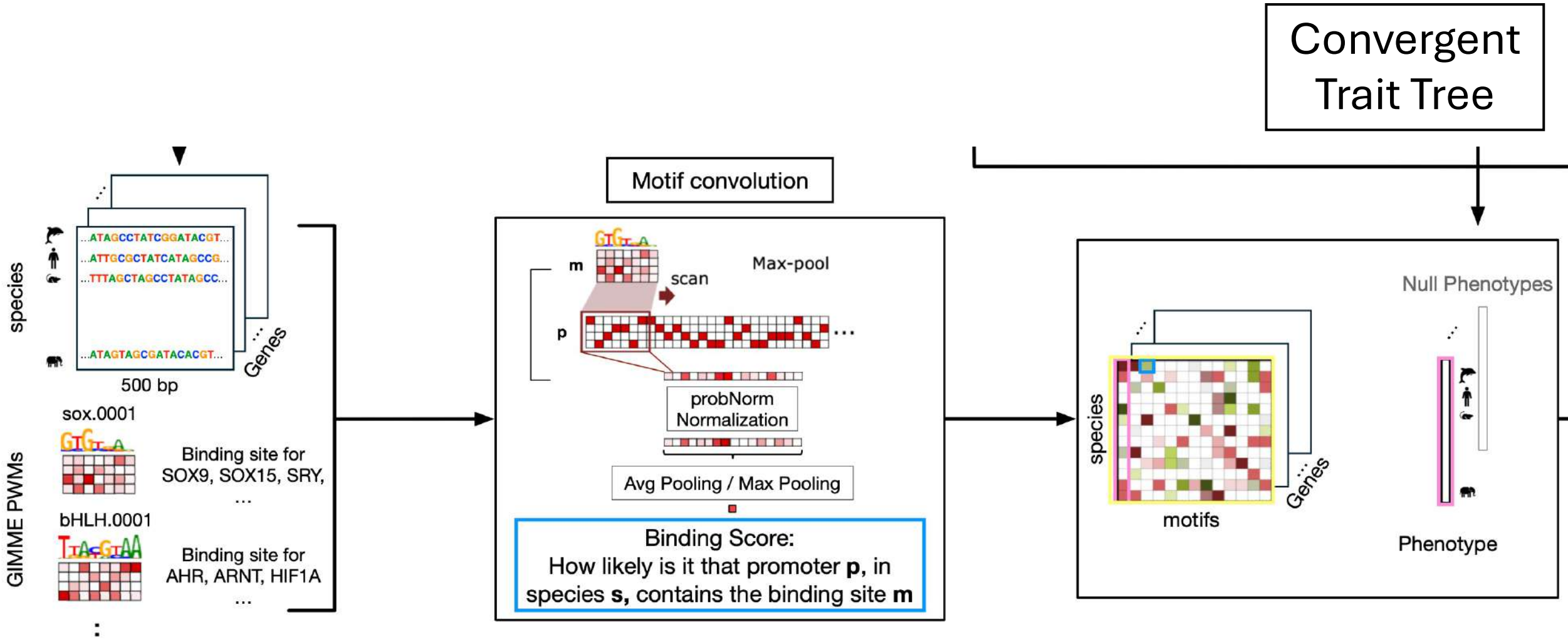
AFconverge –
alignment-free convergent trait analysis of regulatory regions

- Count and score known Transcription Factor binding motifs in orthologous promoter and enhancer elements

- Evaluate changes in each TF, finding those that correlate with your trait

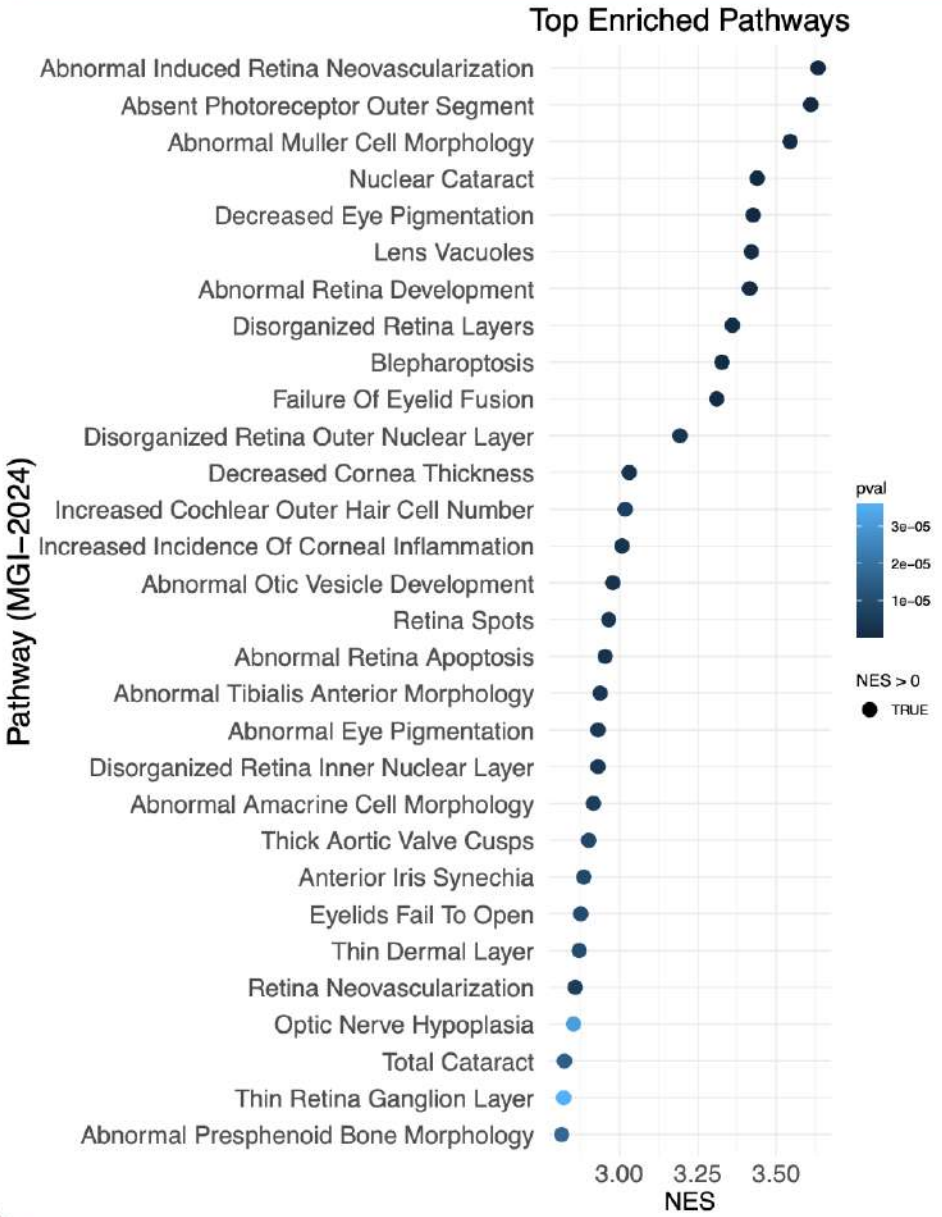
Rezwan Hosseini and Maria Chikina

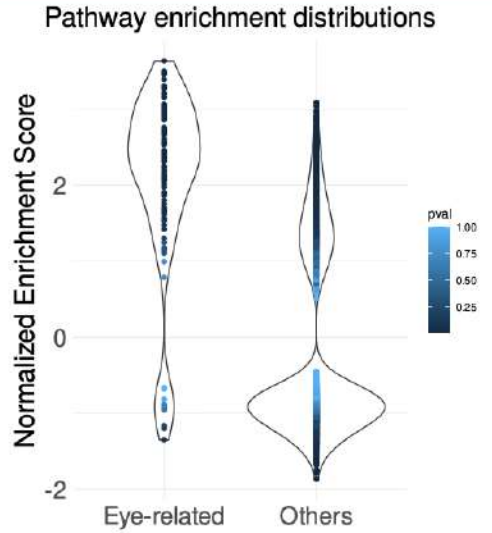
AFconverge – alignment-free convergent trait analysis of regulatory regions



AFconverge

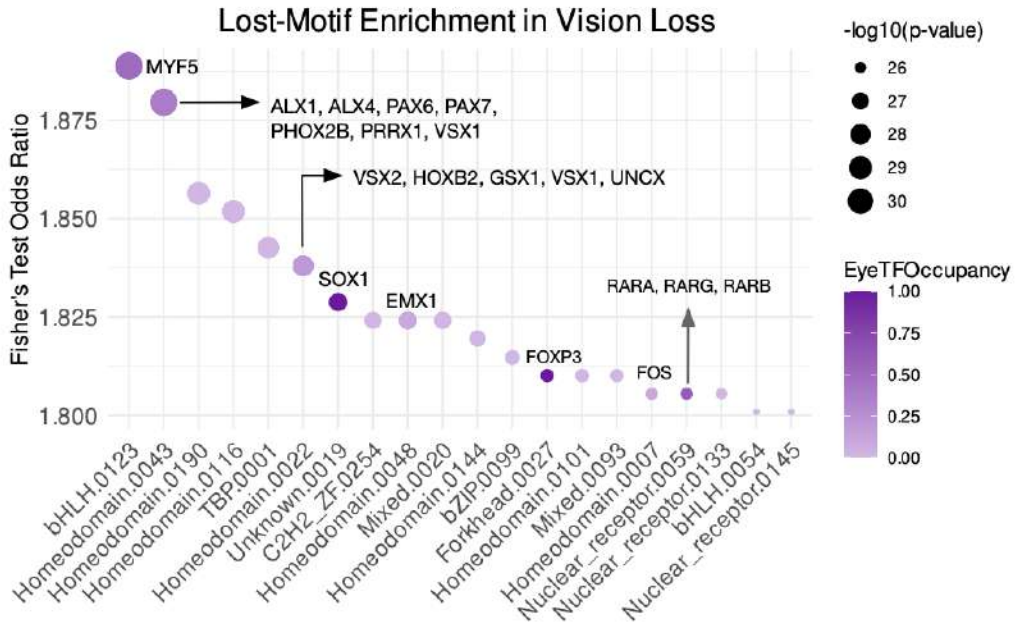
Convergent Vision Loss





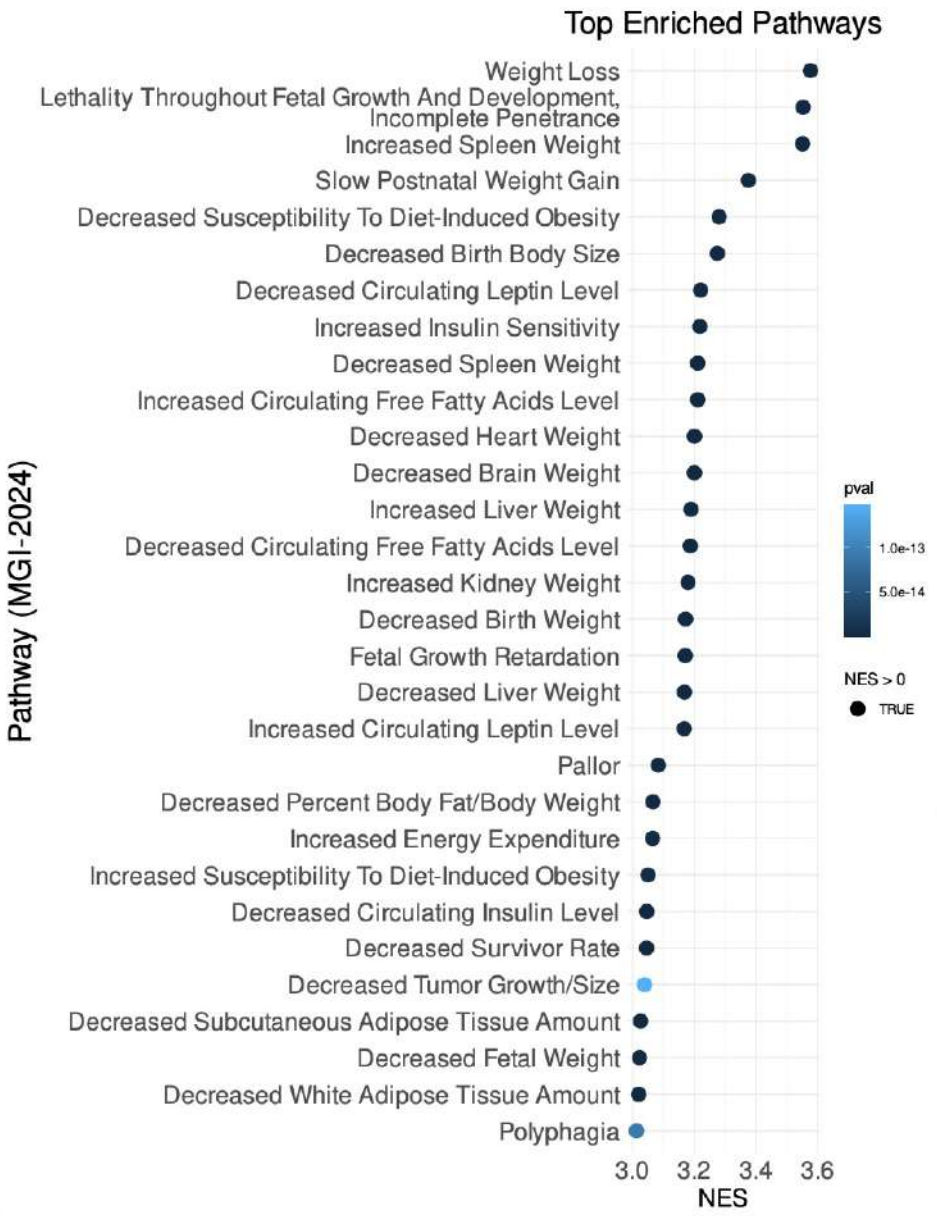
keywords to separate Eye related vs others:

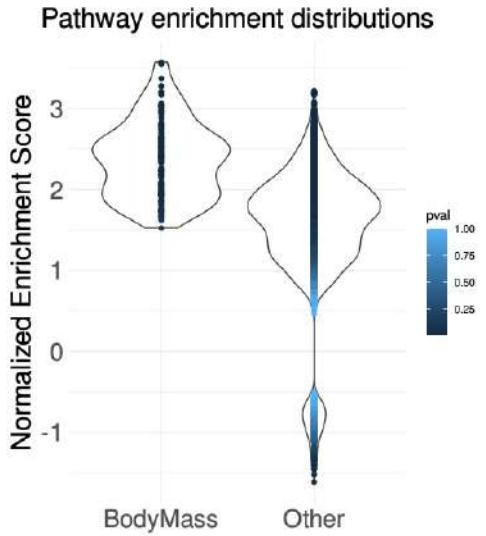
"Eye", "retina", "rod", "cone","lens", "optic", "pupil", "photoreceptor", …



AFconverge

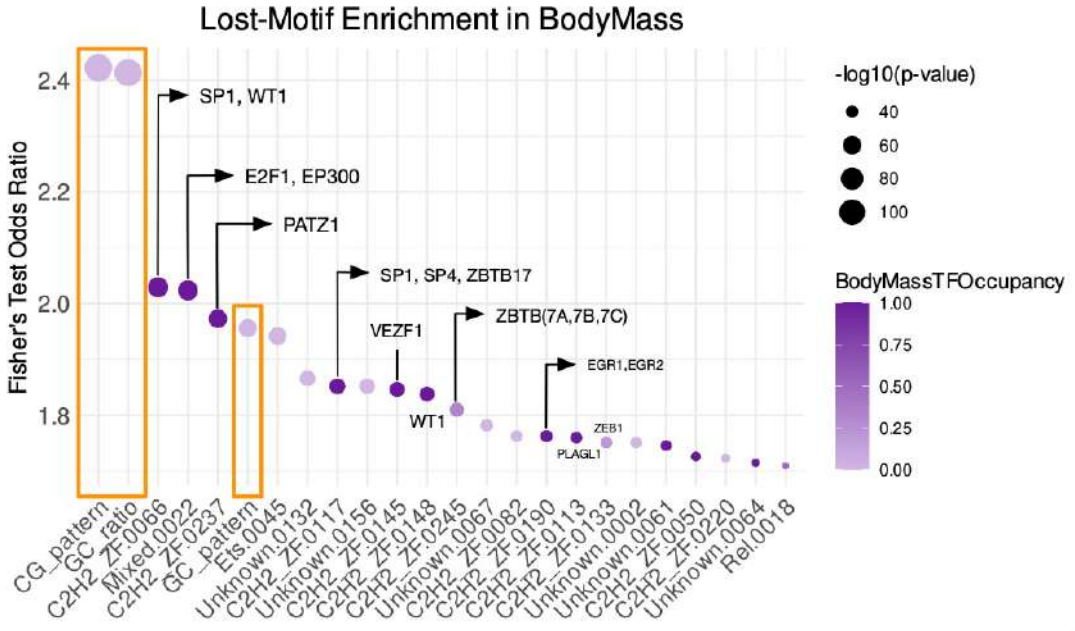
Body Mass





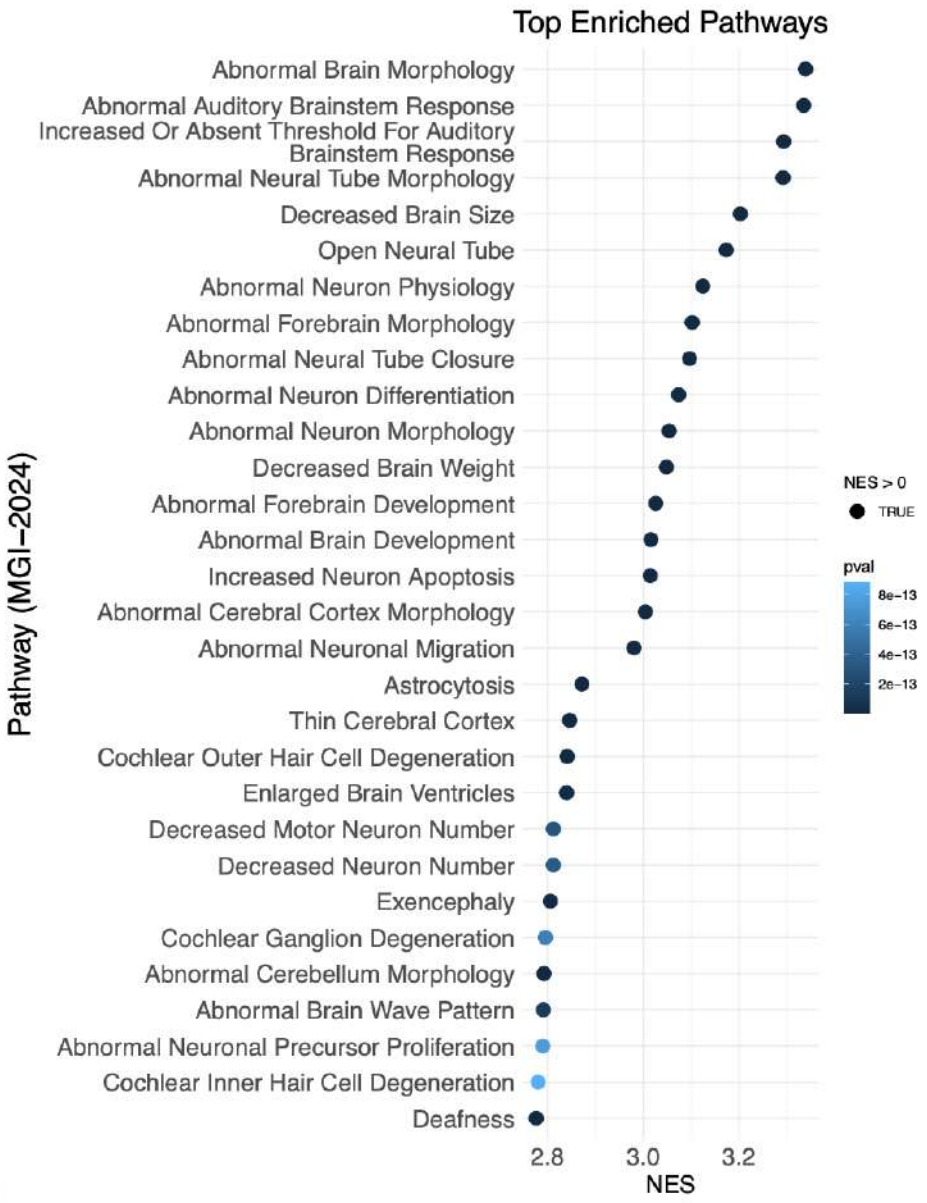
keywords to separate BodyMass related vs others:

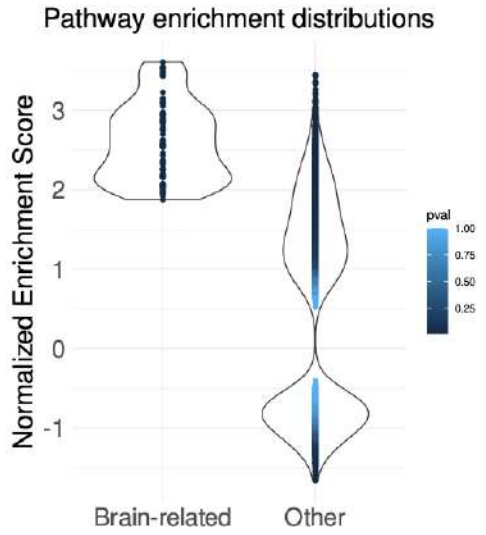
"body mass", "body size", "weight", "growth", "obesity", "fat", "adipose"



AFconverge

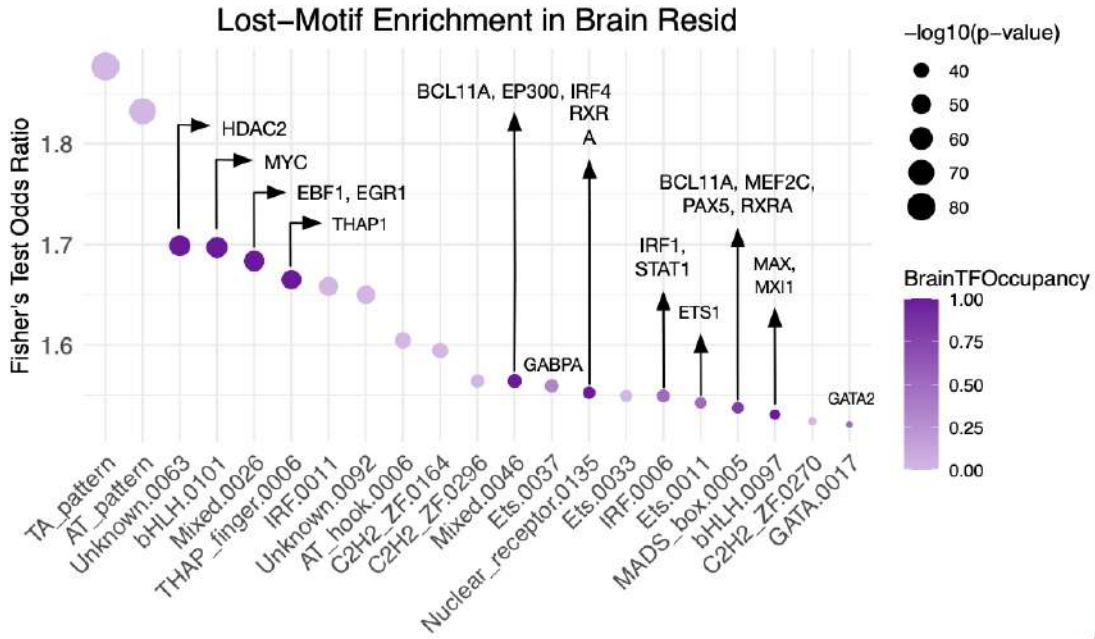
Relative Brain Mass



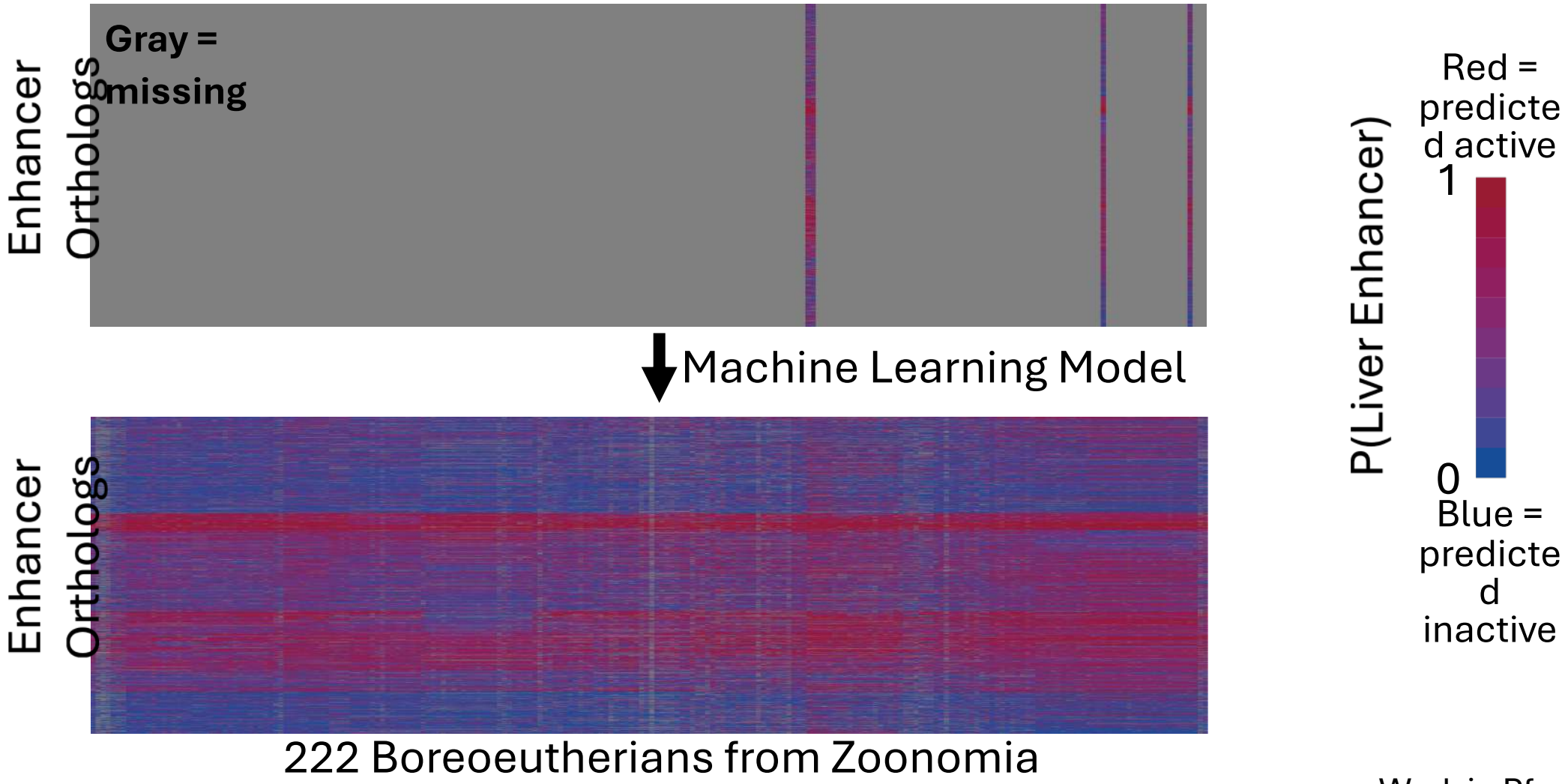


keywords to separate BodyMass related vs others:

"brain", "neural", "neuron"

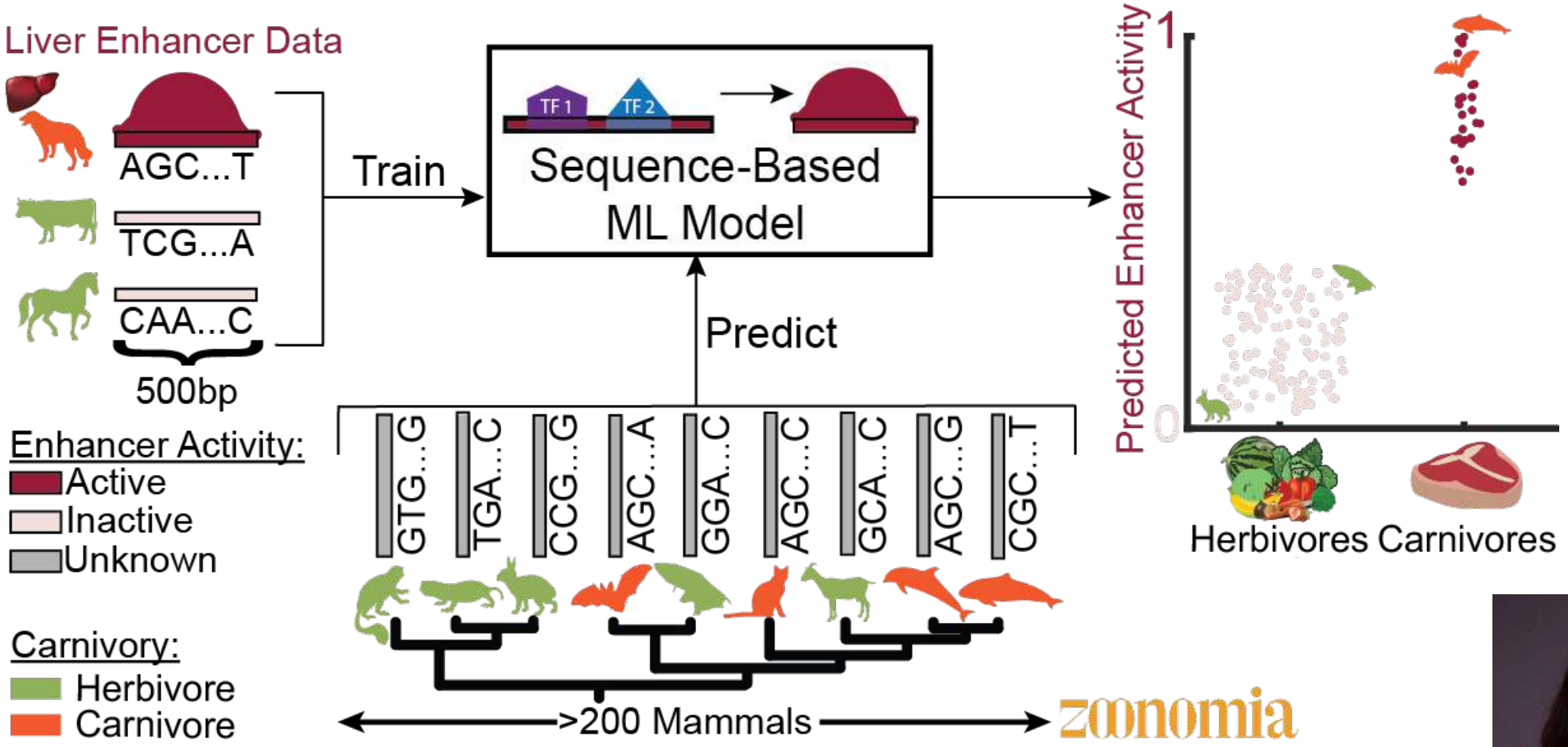


Machine learning models can use DNA sequence to predict enhancer activity across species



Work in Pfenning

The Tissue Aware Conservation Inference Toolkit (TACIT) can identify enhancers involved in convergent trait evolution



Irene Kaplow

Sequence-to-Function genome models

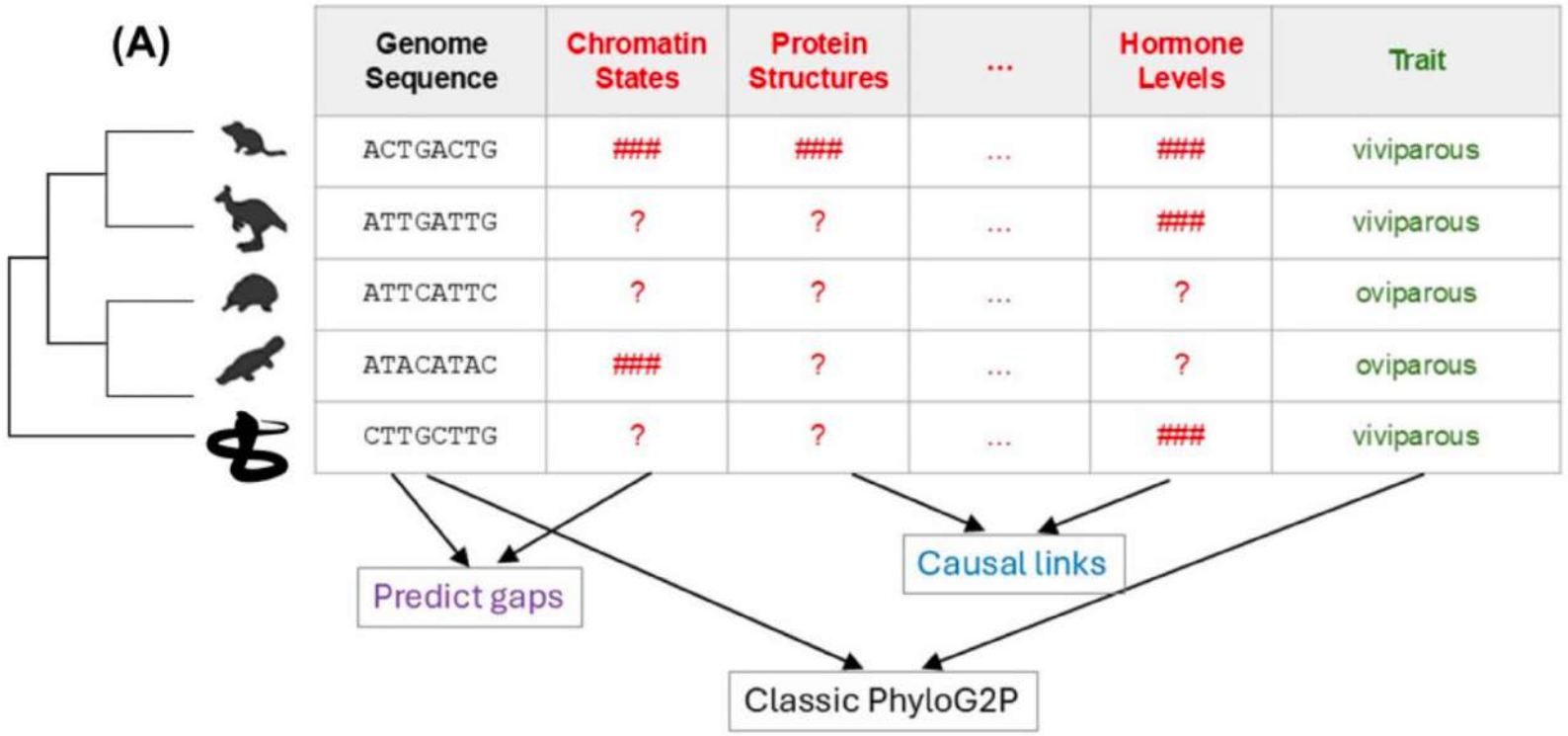
- LLMs were trained on thousands of chromatin and gene expression data from mice and humans
- When given a ~500kbp regions, they predict expression levels of genes in all tissues on which they were trained.
  - AlphaGenome
  - Borzoi

- We are using them in other species with promising results.
- Our goal is to bypass regulatory regions and just study predicted changes in tissue-specific expression levels across hundreds of species!

Integrating Intermediate Traits in Phylogenetic Genotype-to-Phenotype Studies

Nathan L. Clark*, Chris Todd Hittinger†, Hongmei Li-Byarlay‡, Antonis Rokas§, Timothy B. Sackton¶ and Robert L. Unckless||,⊥

Making inferences any of these levels will facilitate comparative studies and **provide mechanism**.
Missing species data (?) can be imputed or predicted using machine learning, "AI".



Intermediate phenotypes could also include:
Gene expression, Protein interactions, Biochemical activities, Metabolism, Interaction networks

What are the Genomics Frontiers that you will advance?

- Machine learning, LLMs, “AI”
  - But how exactly?
- Custom models of individual disease.
  - Personalized organioids
- CRISPR-based gene therapy
  - CAR-T therapy
- Ultra precise pharmaceutical prescriptions based on genome sequence and/or enzymatic function
- Determine functional effects of all possible variants
  - MPRAs
  - deep mutational scanning
- Meta-sequencing as monitoring in real time, constantly!
  - What are the DNA sequences that pass through us, our habitats, our communities, our planet?

The team that makes it happen

The Clark Lab

nclarklab.org

@nclark.bsky.social

Jered Stratton
Emily Kopania
Dwon Jordana
Courtney Charlesworth
Justine Denby
Shahd Elsayed
Sneha Chaudhuri
Nethan Chauhuri
Emma Mahoney

Lab alumni in this work:
Alex Preble
Nico Schwartz
Jordan Little
Guillermo Hoffmann Meyer
Amanda Kowalczyk
Raghav Partha
Wynn Meyer
Allie Graham
Jason Presnell
Sarah Lucas

Maria Chikina
University of Pittsburgh

Sergei Pond
Temple University

National Human Genome Research Institute

NIH National Eye Institute
Research Today...Vision Tomorrow

Methods cited

- BayesTraits https://isu-molphyl.github.io/EEOB563/computer_labs/lab8/BayesTraits.html
  - https://www.evolution.reading.ac.uk/BayesTraitsV5.0.3/BayesTraitsV5.0.3.html
- BUSTED-E
- BUSTED-PH https://github.com/veg/hyphy-readmes/tree/master/busted-ph
- CAFE https://github.com/hahnlab/CAFE
- ERC https://github.com/nclark-lab/erc
- GENE LOSS / TOGA https://tbg.senckenberg.de/hillerlab/tools-and-data/
- HMMer (motif searching) https://www.ebi.ac.uk/Tools/hmmer/home
- MACSE https://www.agap-ge2pop.org/macse/
- MUSCLE 5 https://github.com/rcedgar/muscle
- OrthoFinder https://github.com/davidemms/OrthoFinder
- OrthoSnap and other Steenwyk software
  - https://jlsteenwyk.com/orthosnap/other_software/index.html
- PAML https://github.com/abacus-gene/paml
- PGLS with PhyTools (R library) http://www.phytools.org/Rbook/
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC10773453/
- phyloP, PHASTCONS with RPHAST
  - https://github.com/CshlSiepelLab/RPHAST
- PRANK https://ariloytynoja.github.io/prank-msa/
  - https://github.com/ariloytynoja/prank-msa/blob/master/src/prank.1.pod
- RERconverge https://github.com/nclark-lab/RERconverge