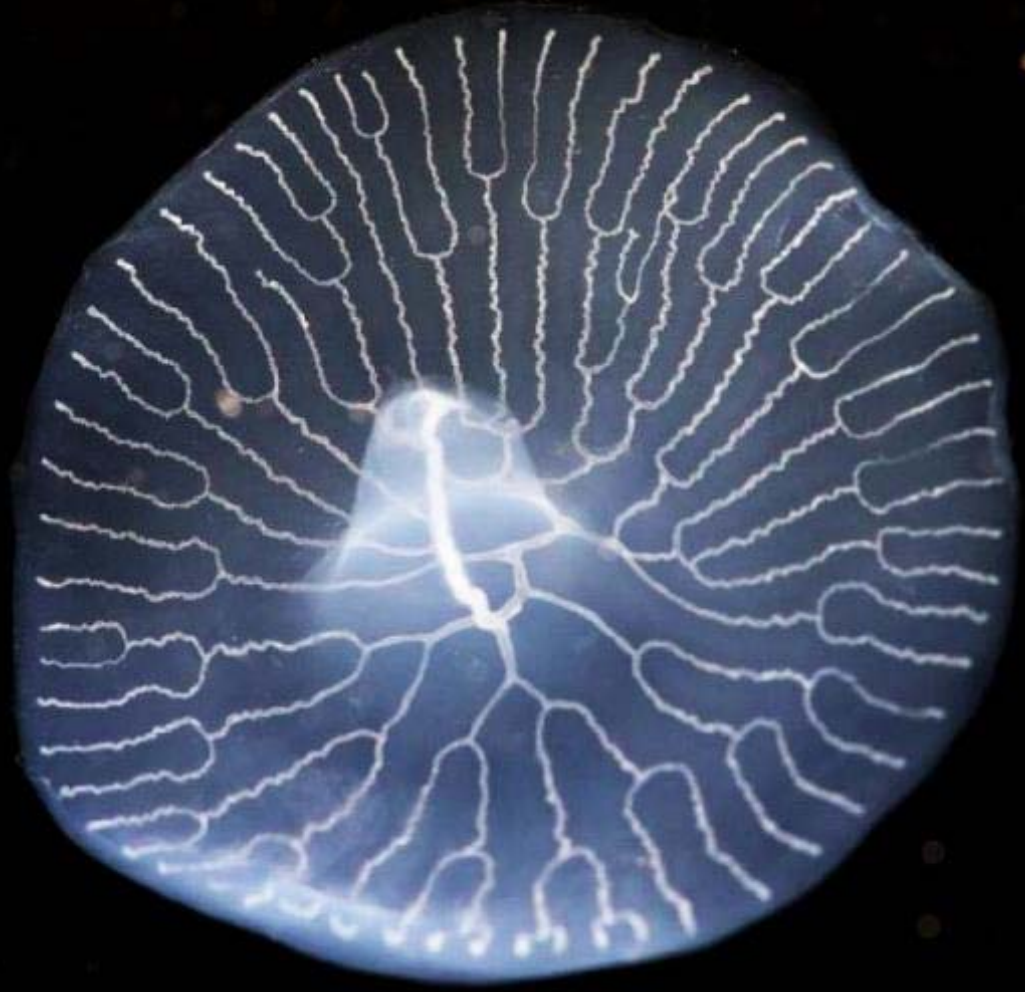
Evolutionary Genomics



Antonis Rokas
Department of Biological Sciences, Vanderbilt University
http://www.rokaslab.org
@RokasLab

Lecture Outline

- **Introduction to evolutionary genomics**

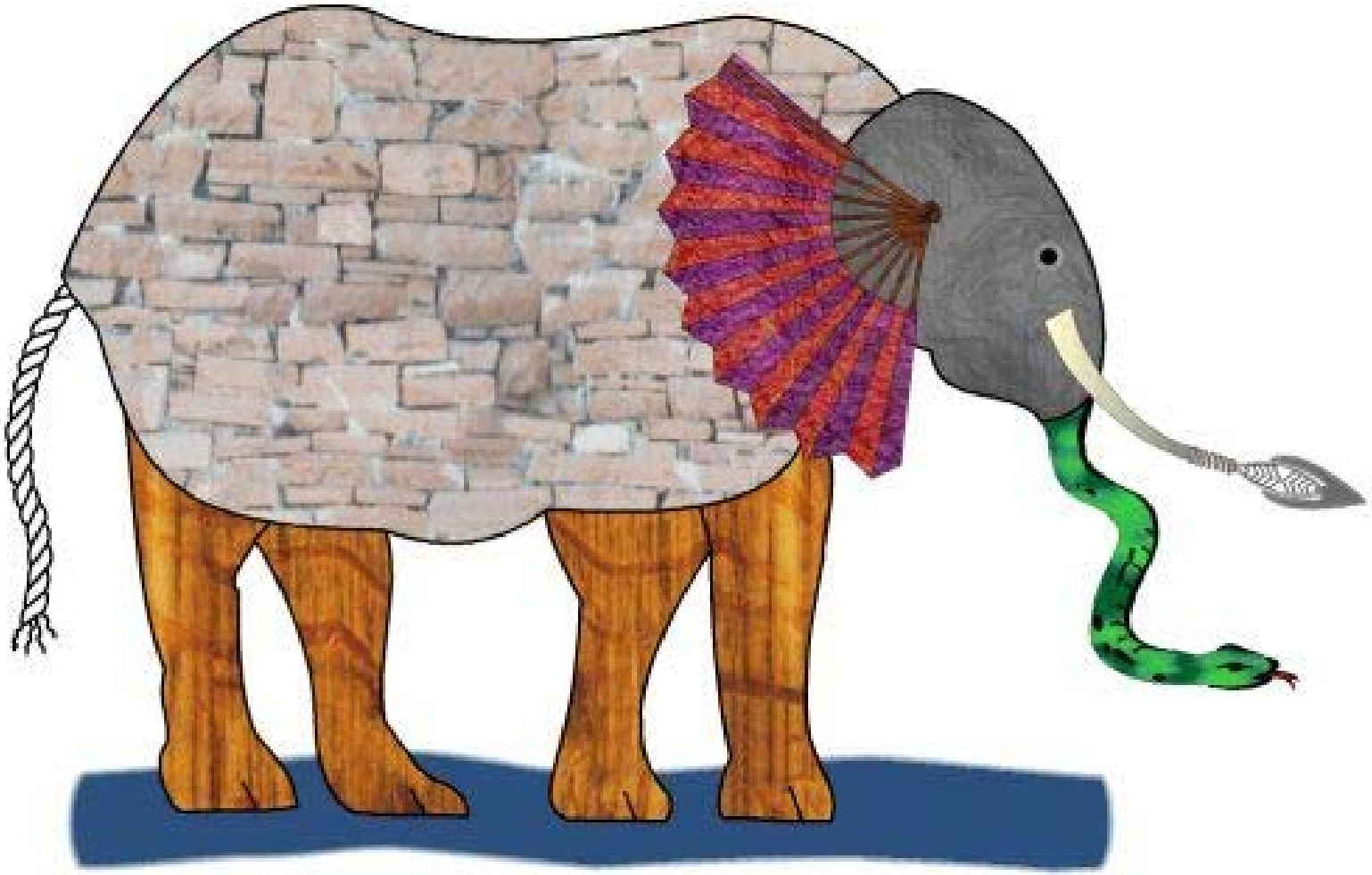
- **Phylogenomics, act 1**

---------- Coffee Break ----------

- **Phylogenomics, act 2**

- **Using genomes to understand lineage diversification**

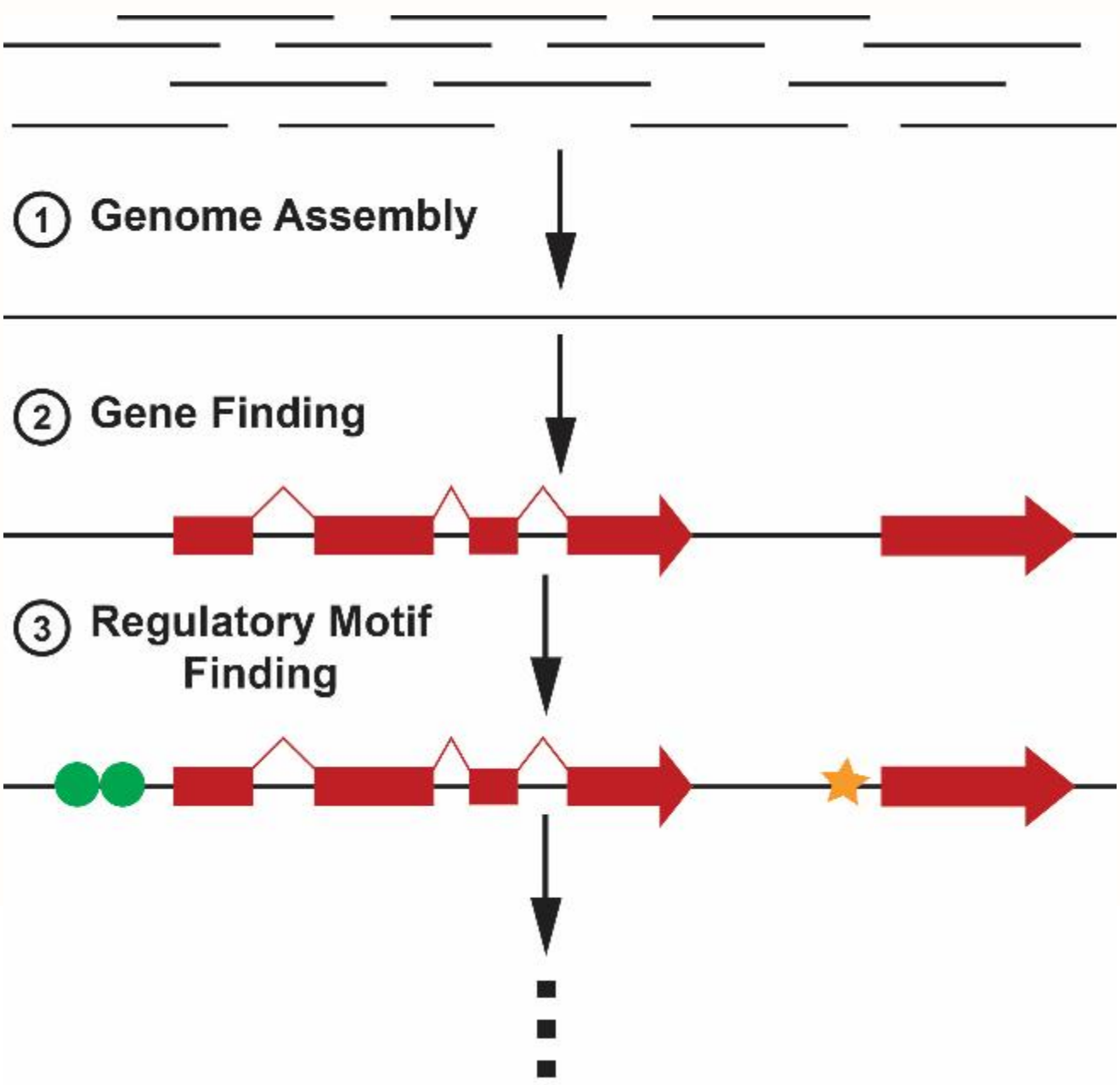
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

Understanding the Genome Requires Tools

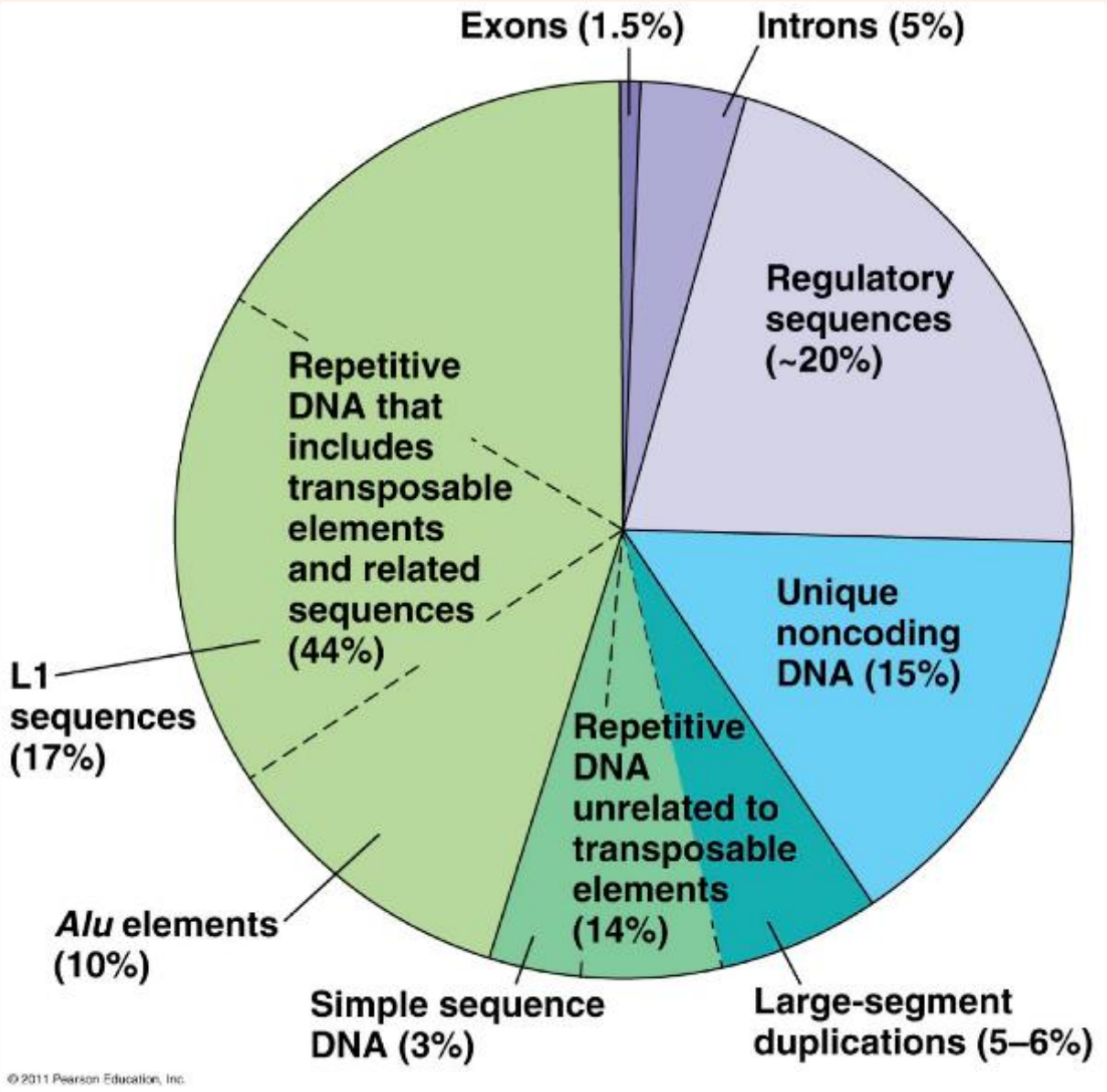


What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

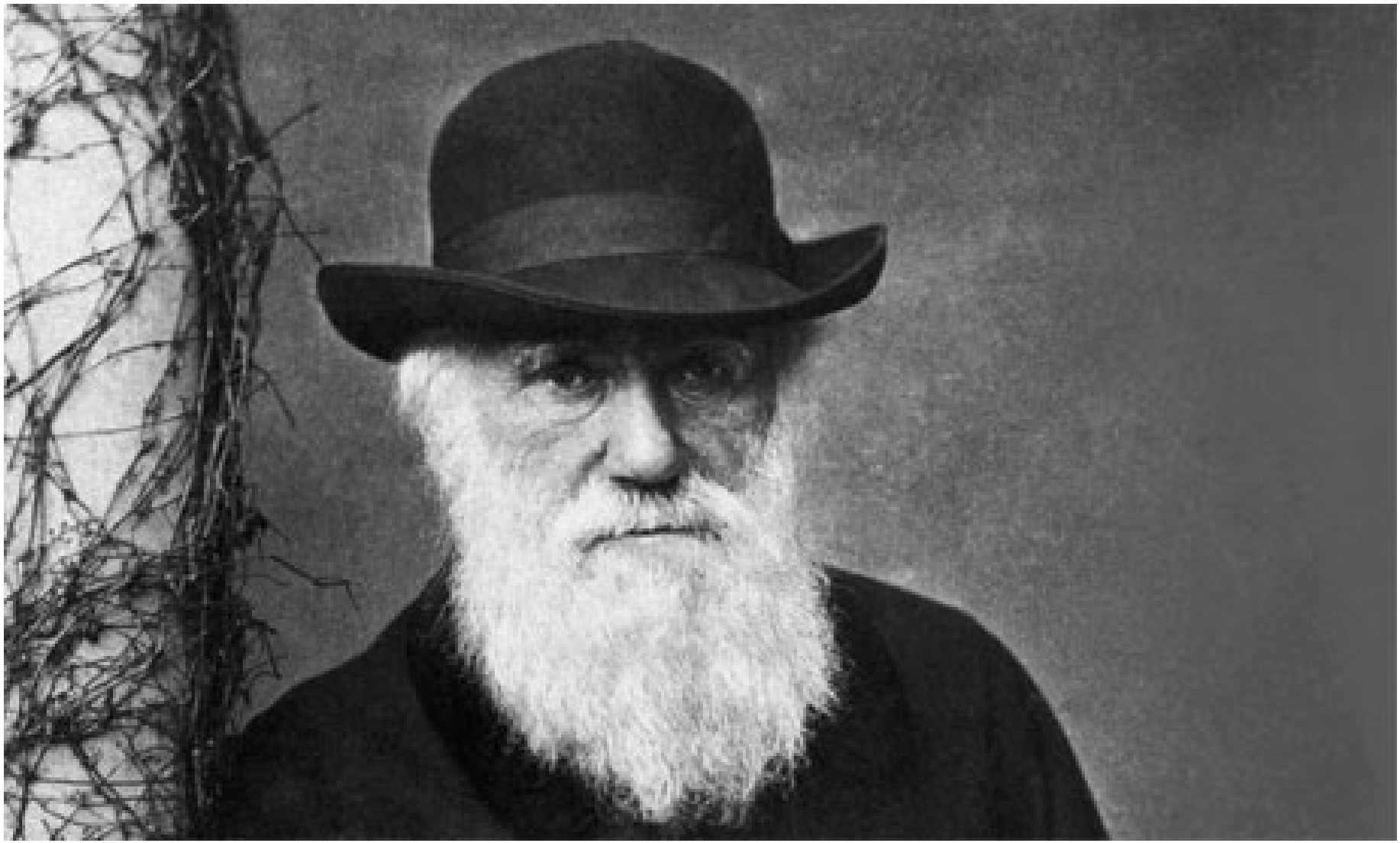
Transposon **Protein Binding Site** **Exon** **Intron**

Organization of the Human Genome

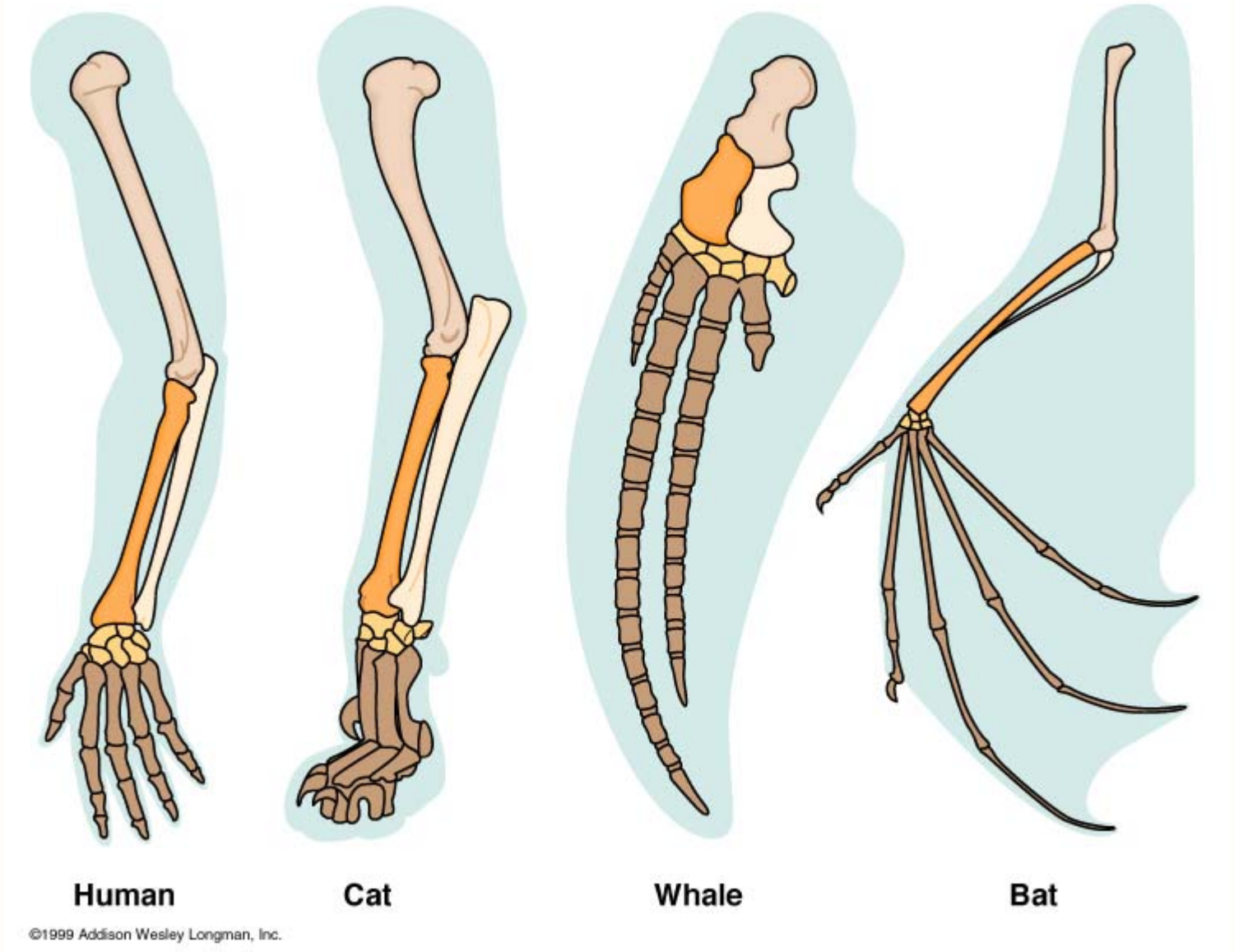


http://todd.jackman.villanova.edu/HumanGenome.jpg

Understanding the Genome Requires a Theory



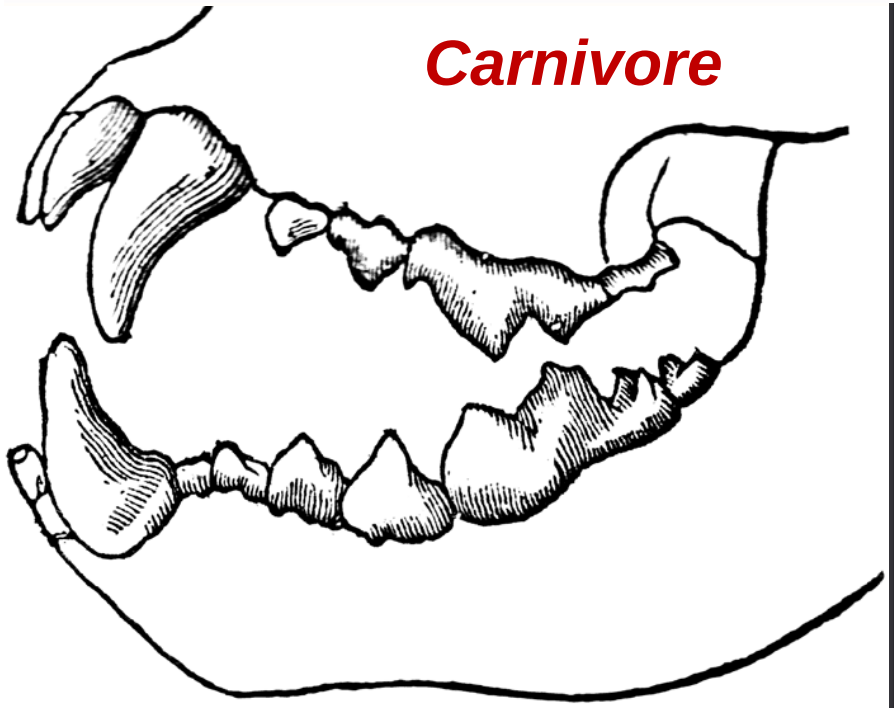
Similarity in Anatomy Suggests Common Origins



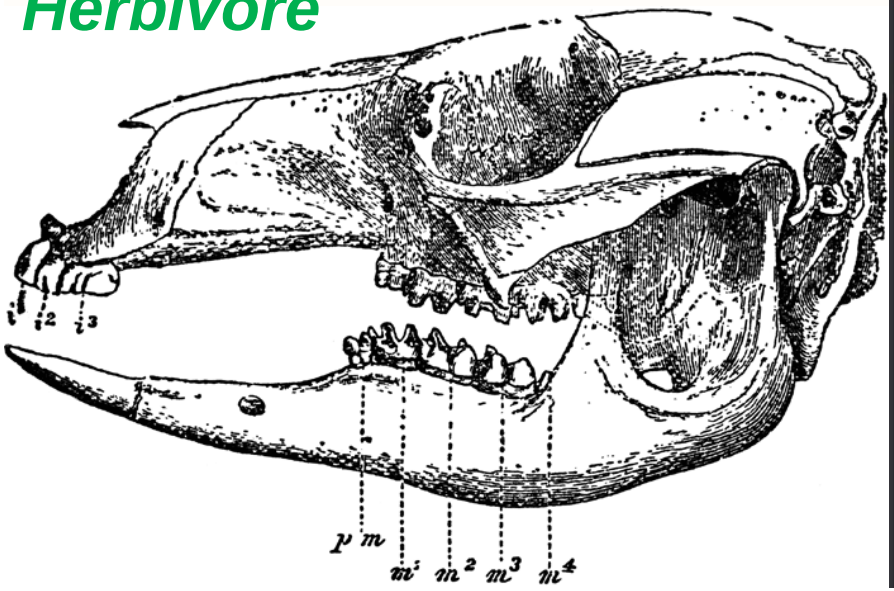
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg

Differences in Anatomy Suggest Adaptations

Carnivore



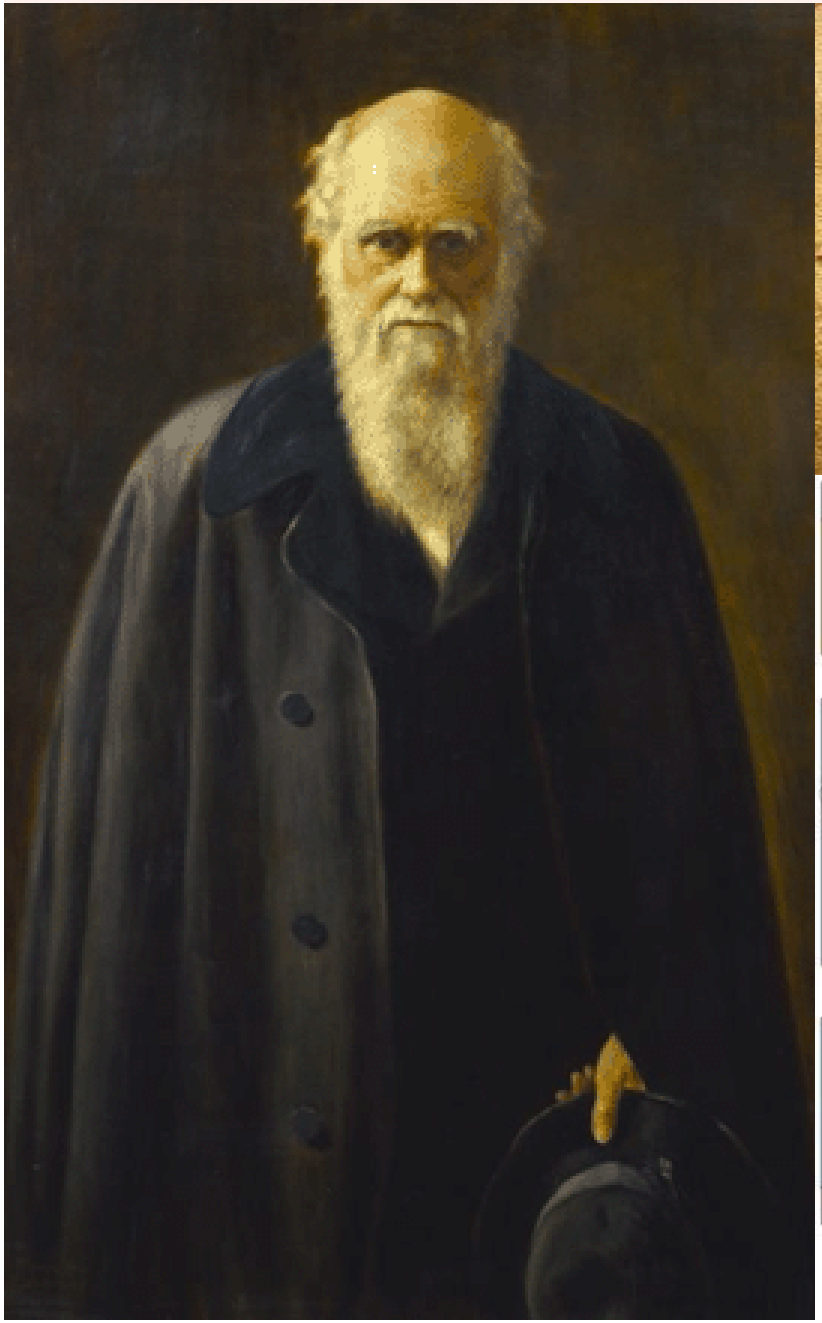
Herbivore

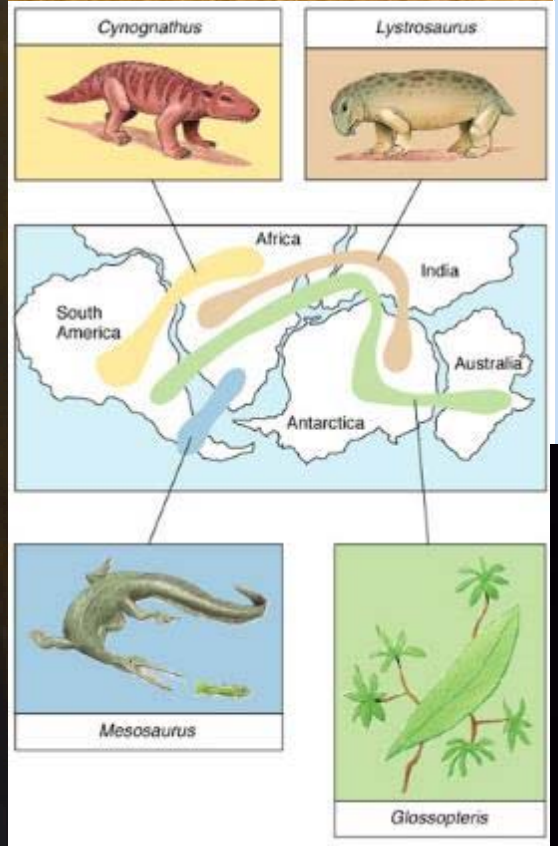


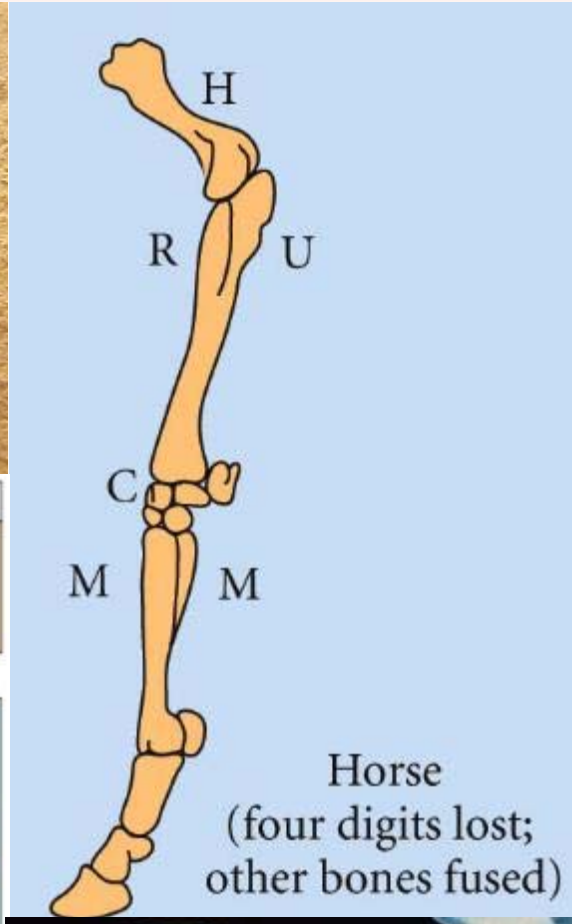
Incisivosaurus

Carnivore or Herbivore?

Darwin's Data

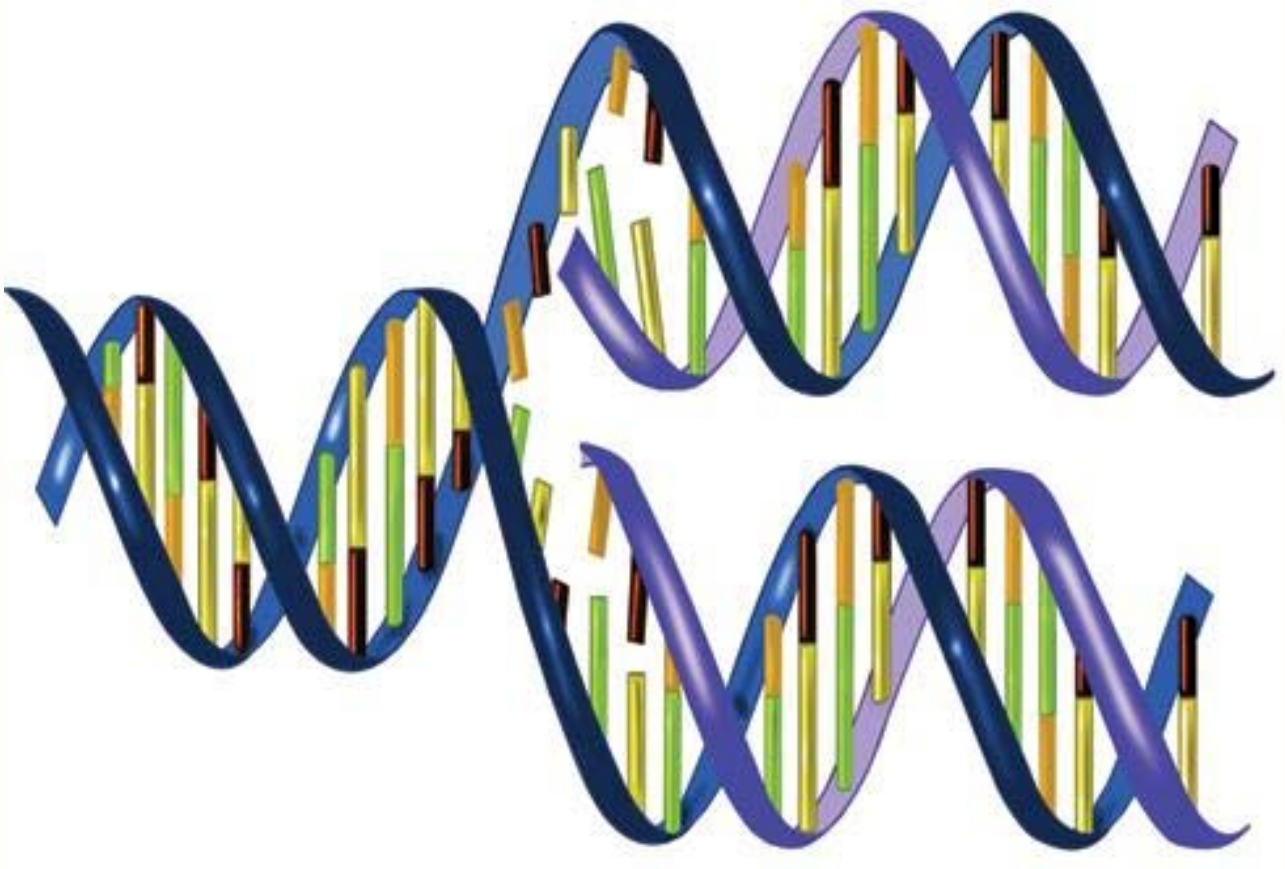






The DNA Record

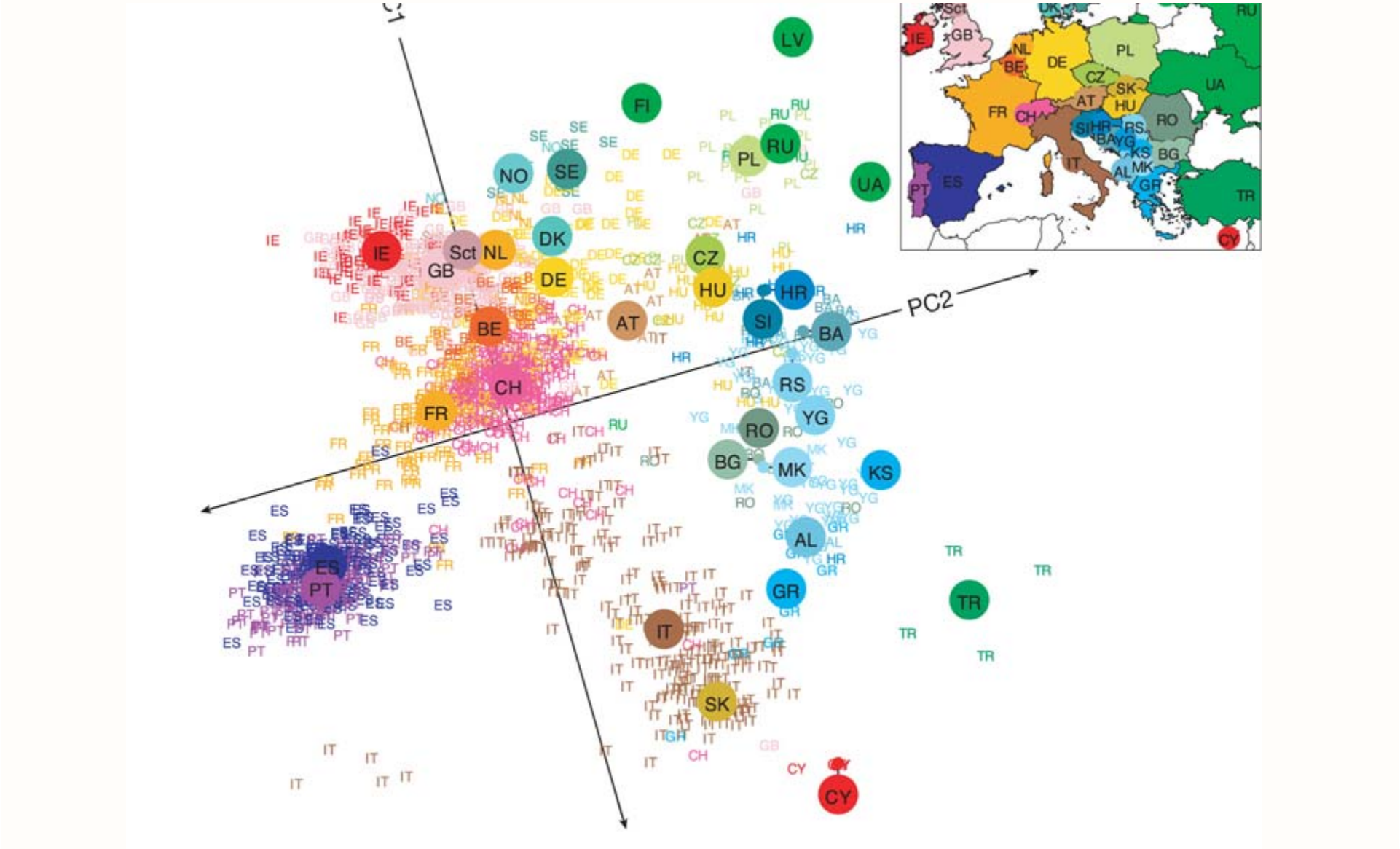
The DNA record contains important clues about organisms' biological past, and their history of change and adaptation



Similarities in the DNA record suggest common origin

Differences in the DNA record (might) suggest adaptations

Human Genes Mirror Geography



Novembre et al. (2008) Nature

Recent Positive Selection in Human Populations

in the Asian Population, involved in hair follicle development

The twenty-two strongest candidates for natural selection

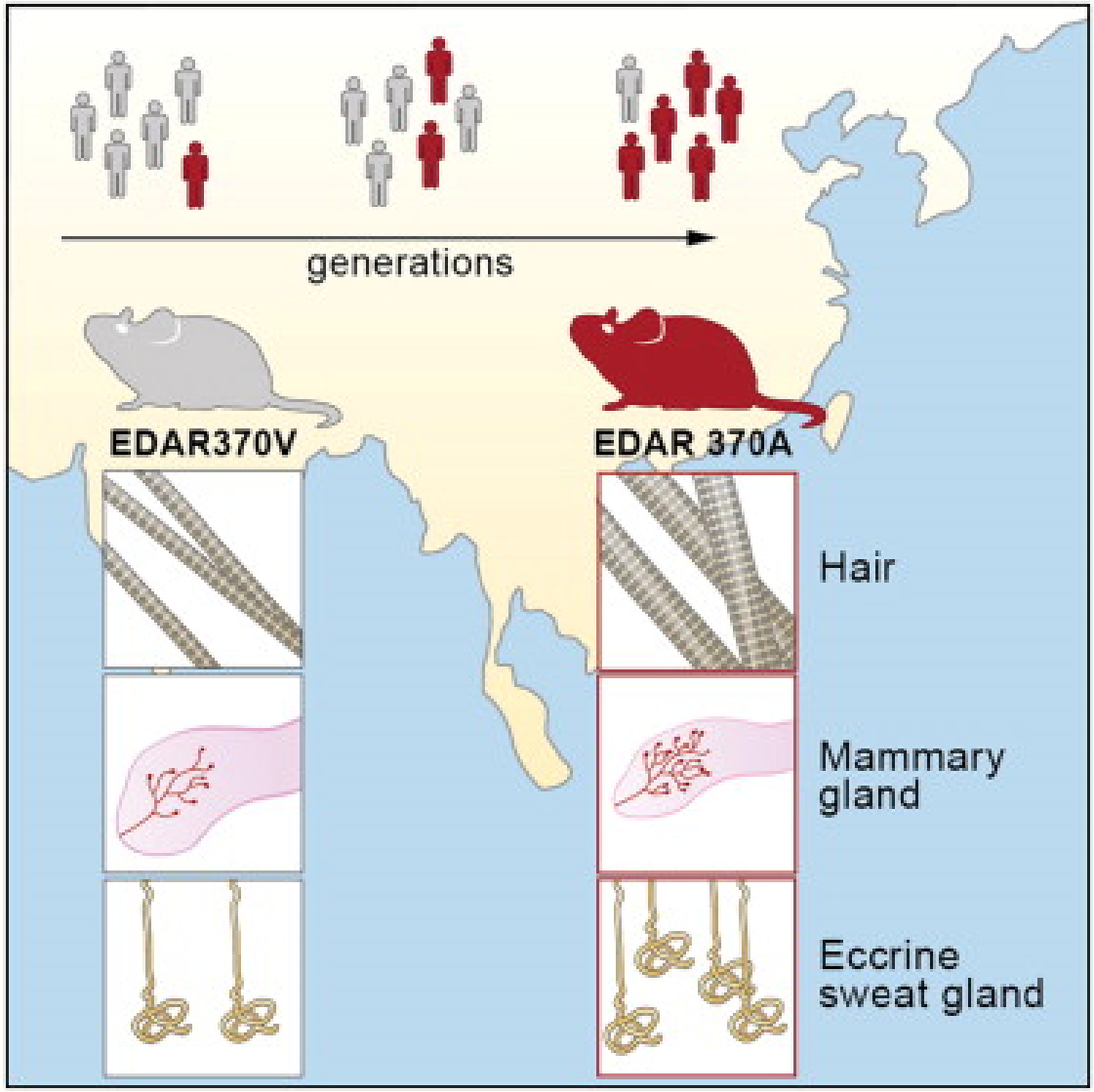
Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHST5, ADAT1, KARS*
chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
	Total SNPs		16.74	9,166	2,898	480	41	

In the European population, involved in skin pigmentation

In the West African population, related to Lassa virus infection

Sabeti et al. (2007) Nature

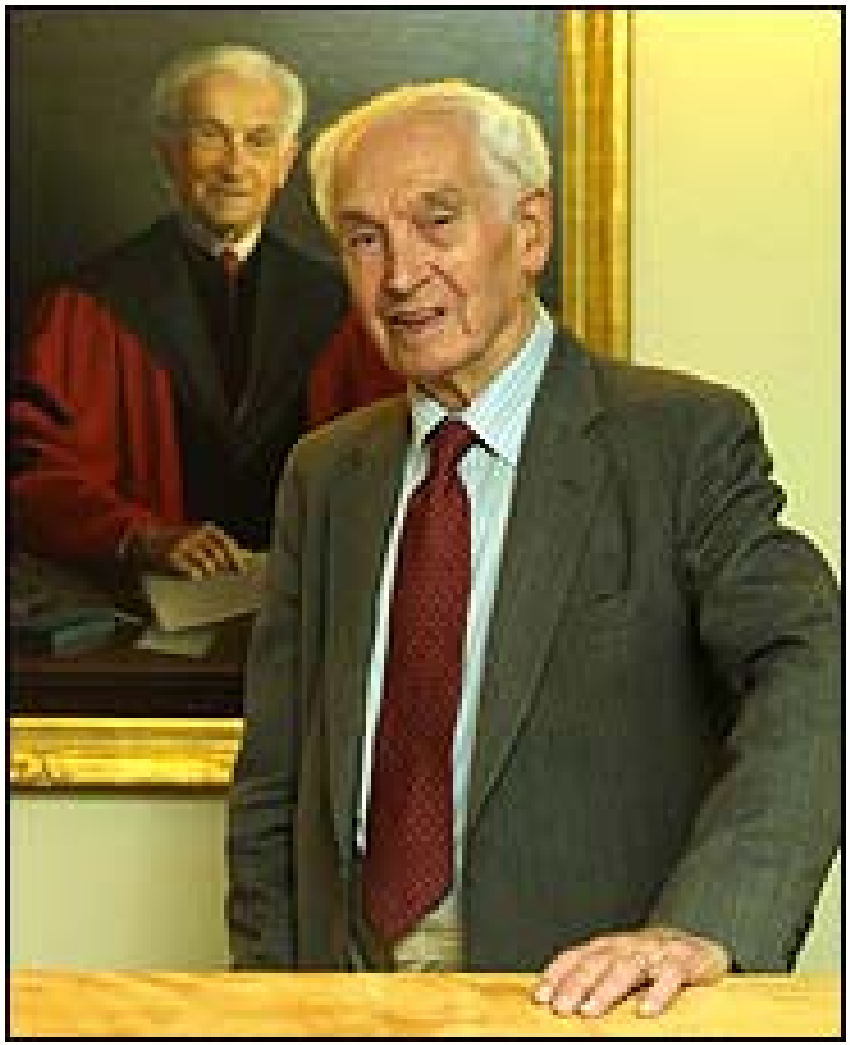
Phenotypic Effects of Recent Positive Selection



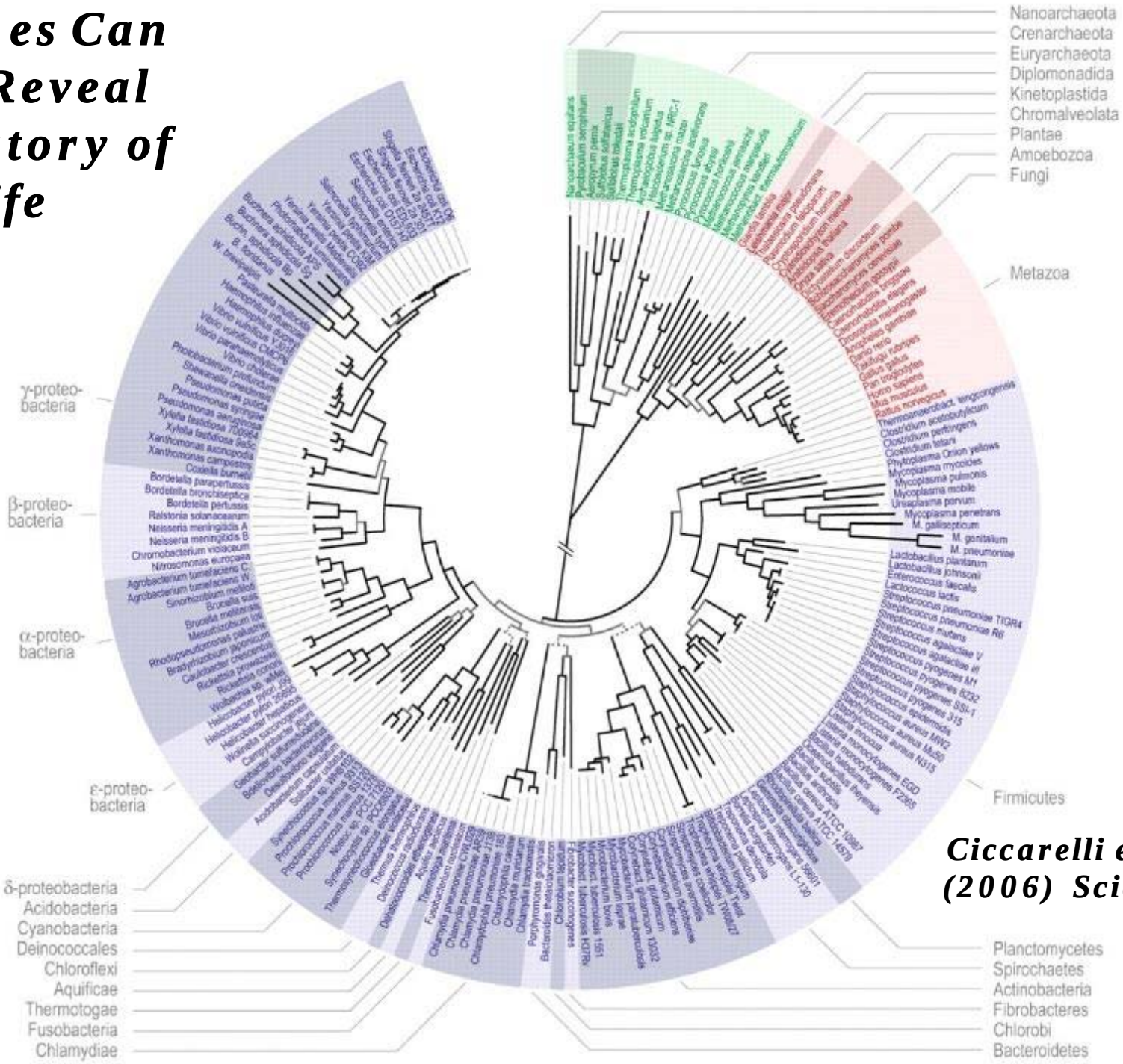
Kamberov et al. (2013) Cell

“…the search for homologous genes is quite futile except in very close relatives”

Ernst Mayr, 1963

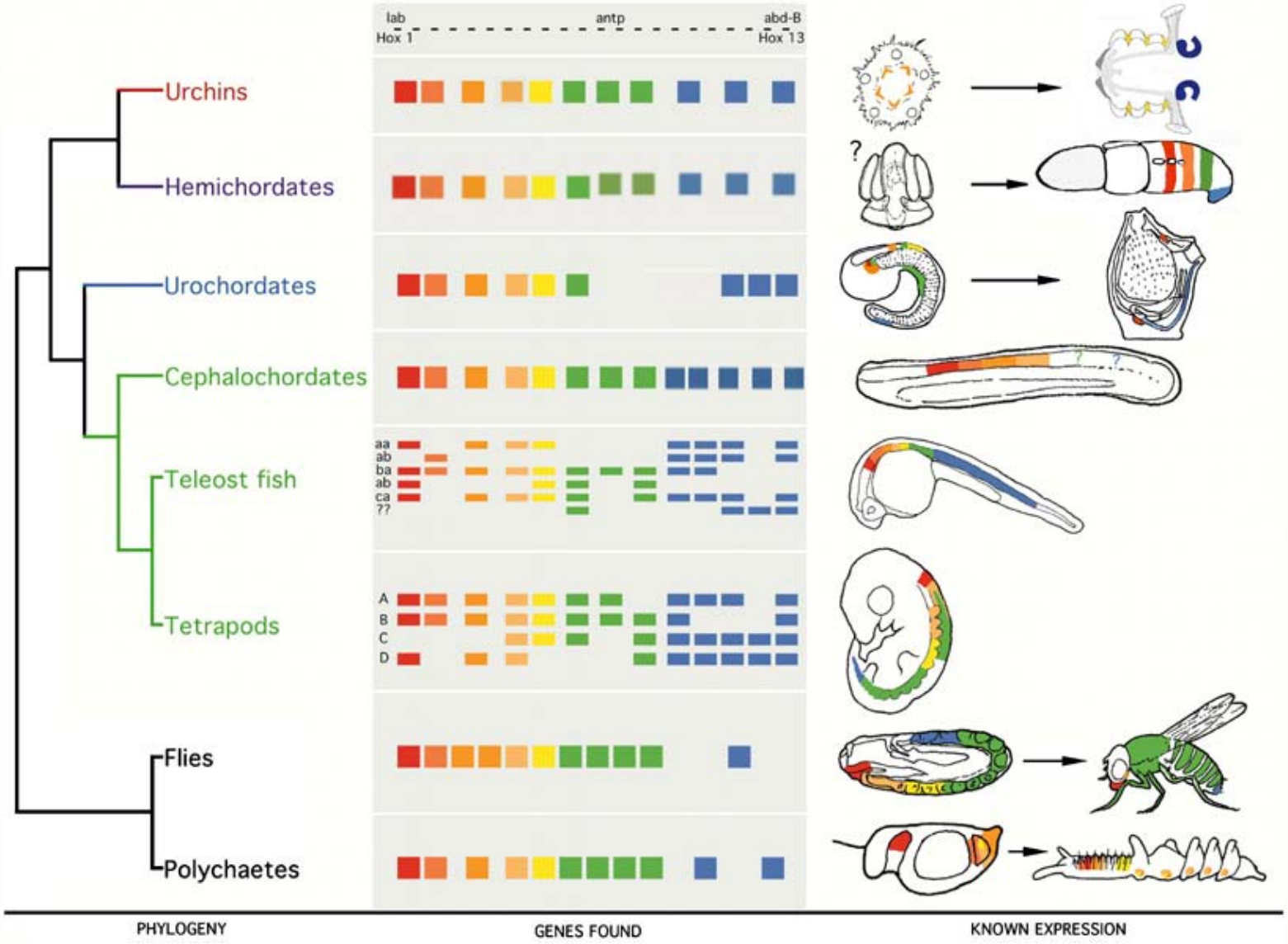


Genomes Can Help Reveal the History of Life



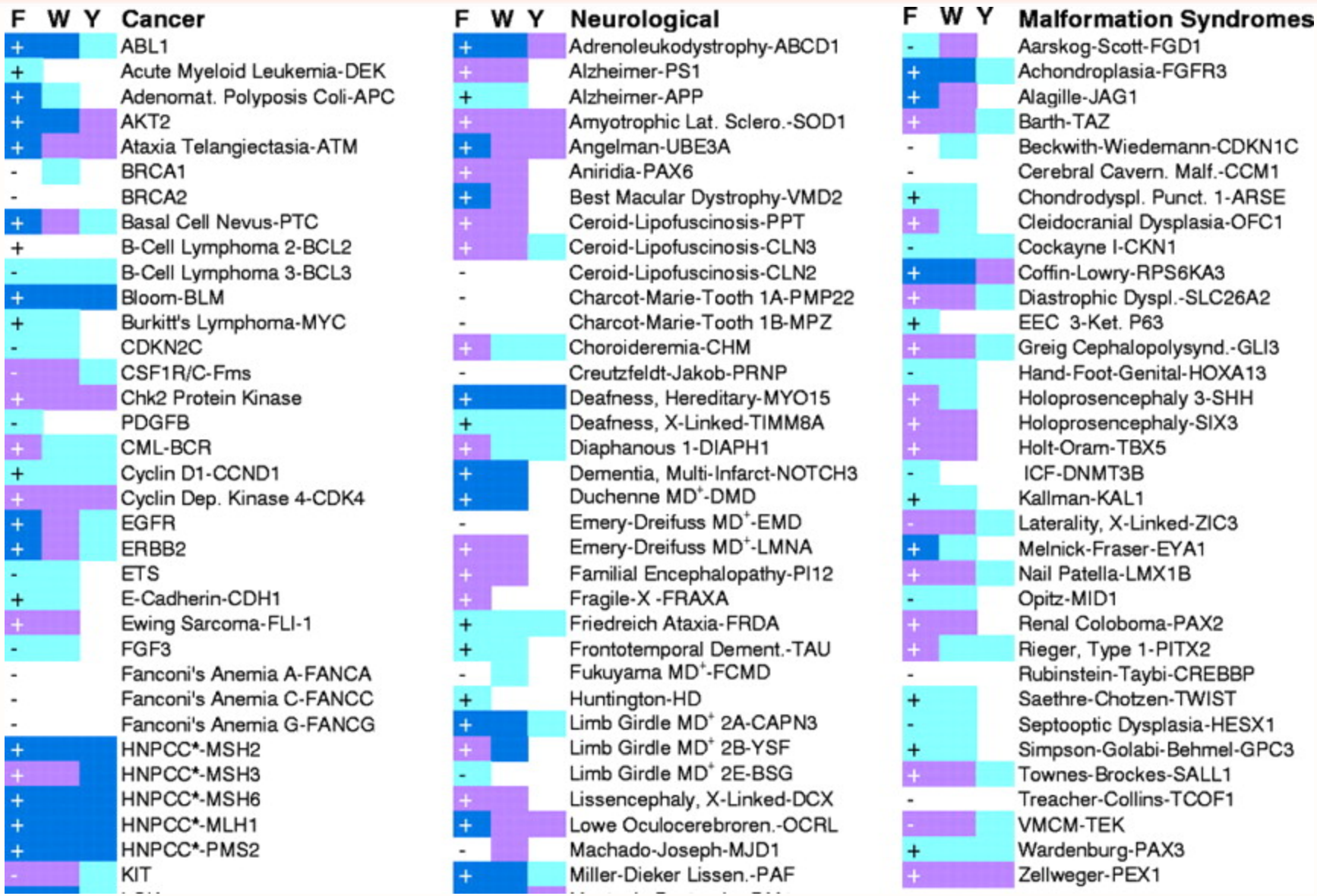
Ciccarelli et al. (2006) Science

Animal Bodies are Built from the Same Genetic Toolkit



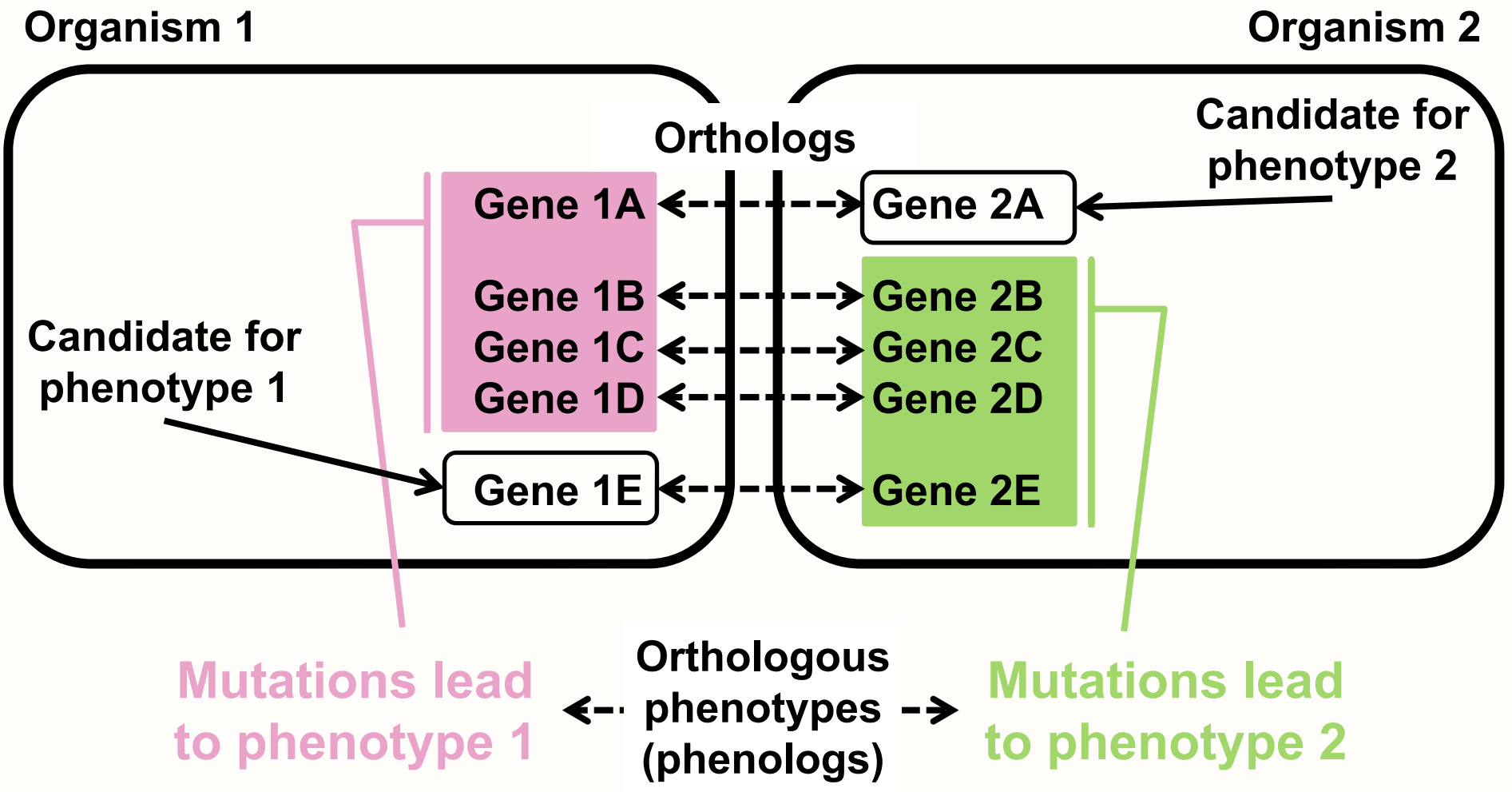
Swalla (2006) Heredity

What Makes Us Sick Is the Stuff of Life



Human disease-associated genes shared with flies (F), worms (W), and Yeast (Y); from Rubin et al. (2000) Science

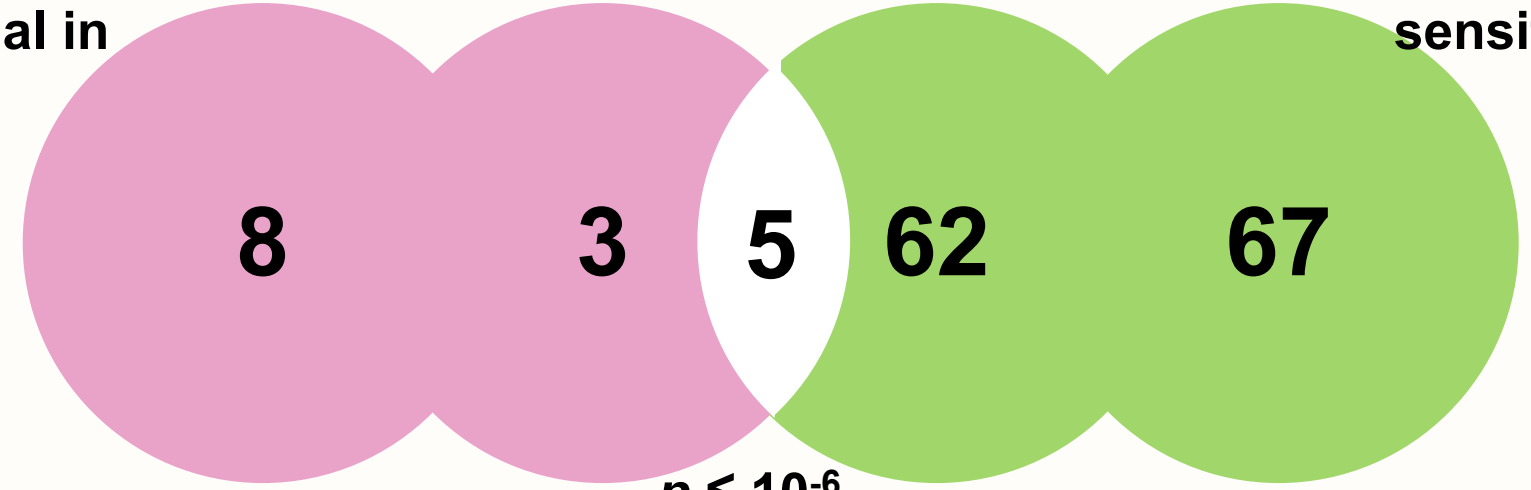
Evolution-Informed Analyses Have Great Predictive Power



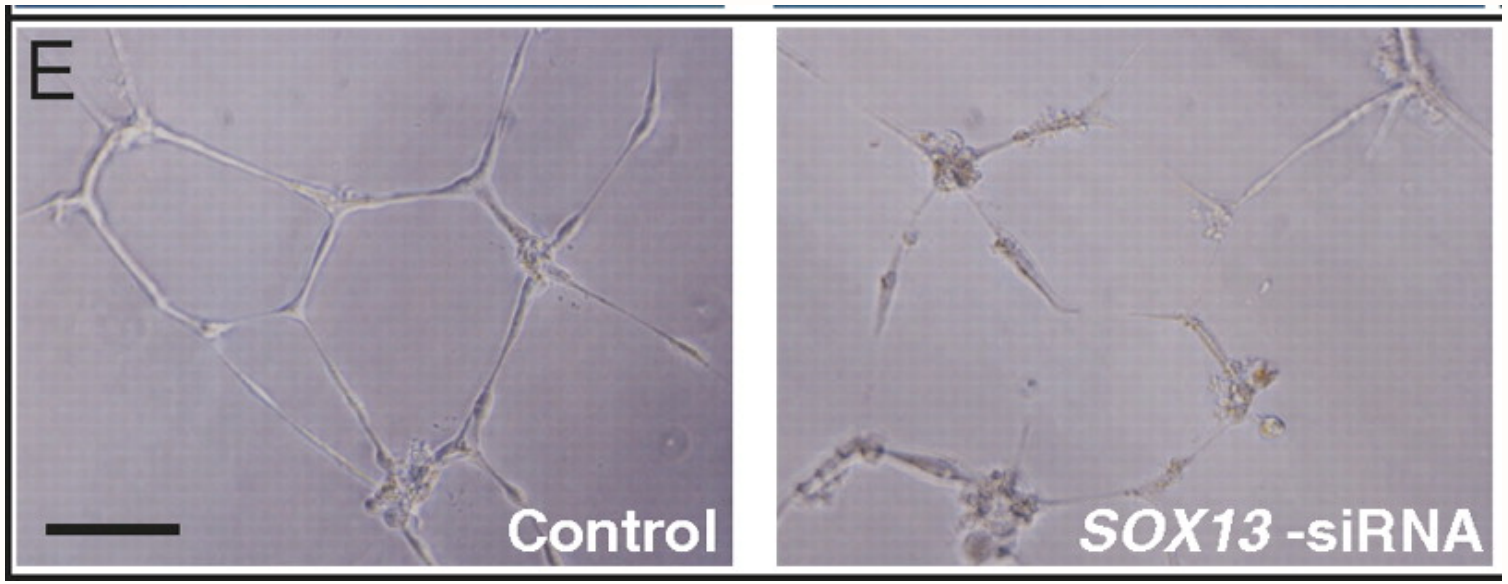
A Yeast Model for Angiogenesis

Angiogenesis abnormal in mice

Lovastatin sensitive in yeast



$p \leq 10^{-6}$



Genomics Used to Be "Big Science"…

wellcome trust sanger institute

BCM
Baylor College of Medicine

BROAD INSTITUTE

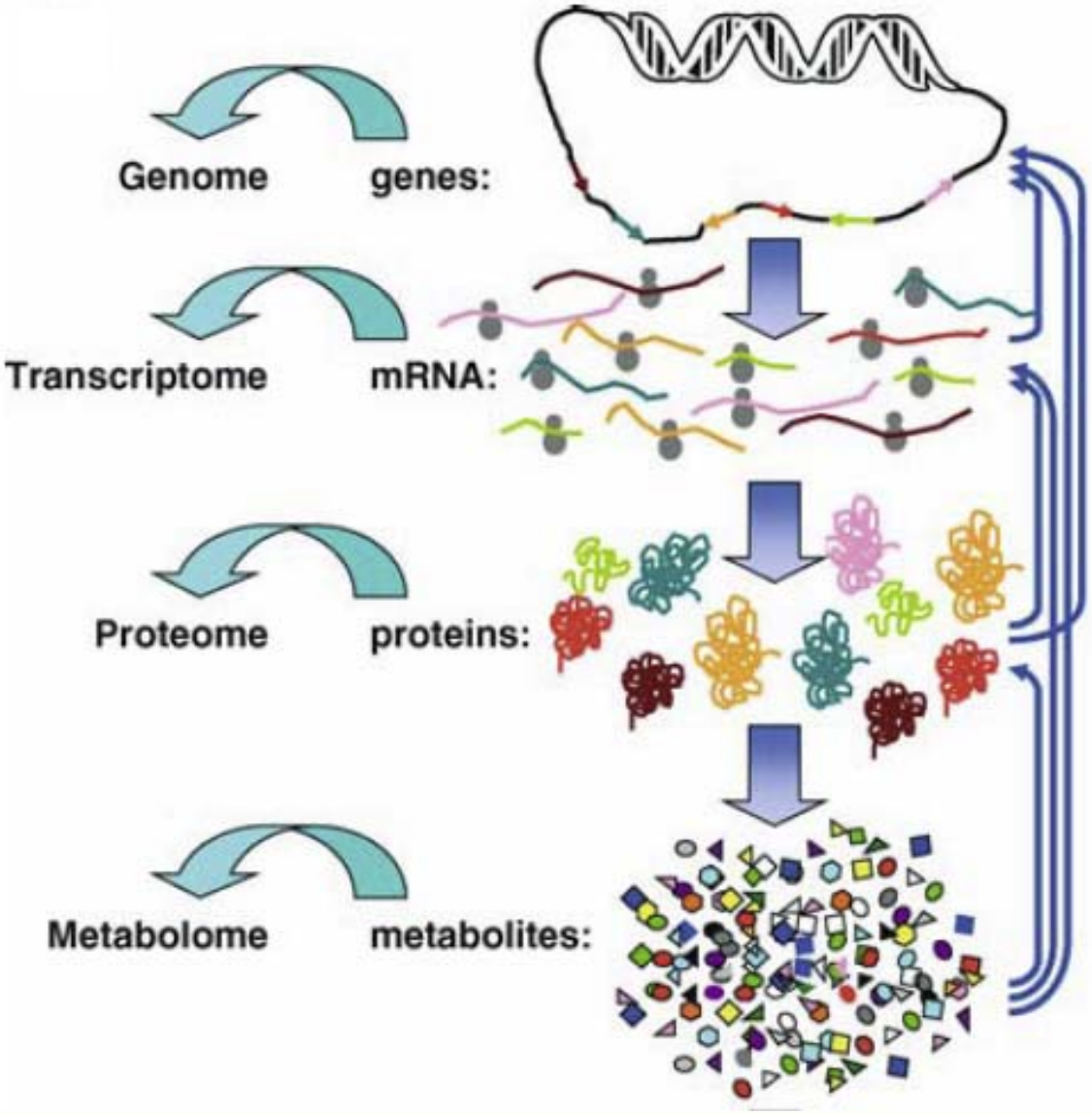
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

J. Craig Venter INSTITUTE

THE Genome CENTER
AT WASHINGTON UNIVERSITY

... But is now Accessible to Every Lab

The Age of High Throughput Technologies



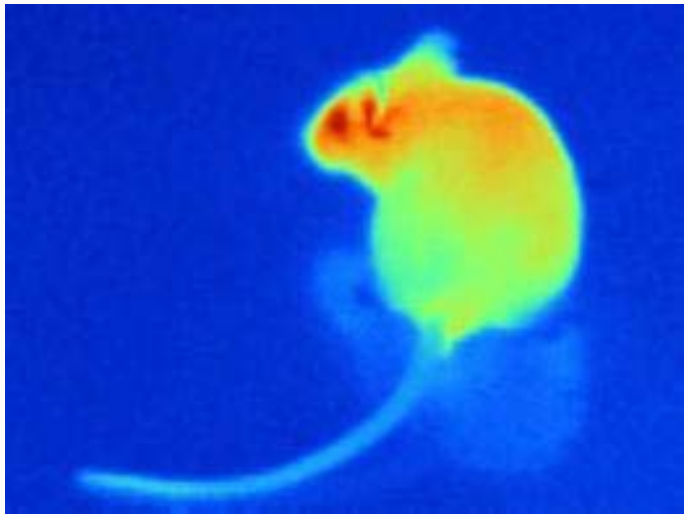
Novel Ways to Probe Gene Function in Any Organism

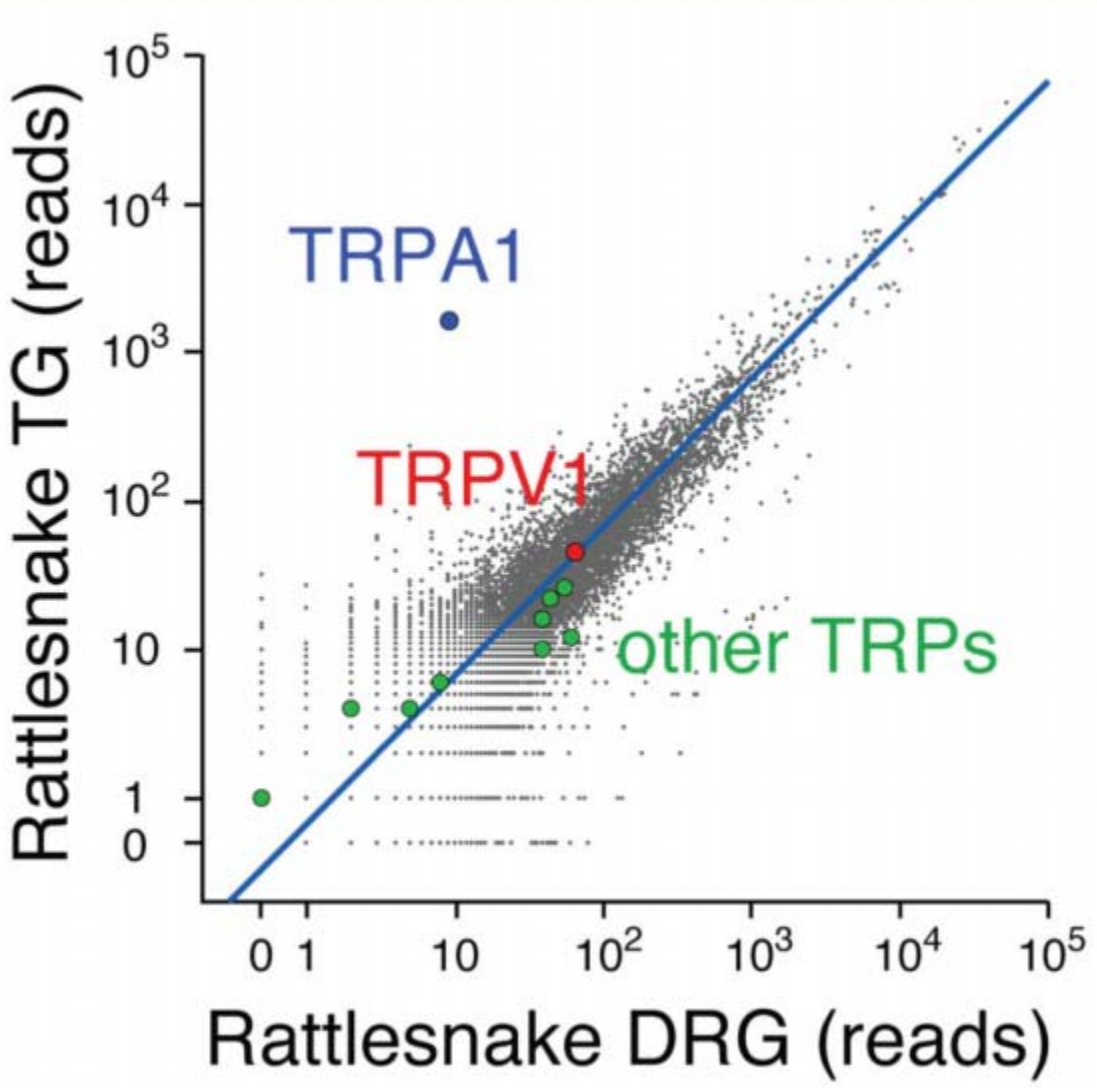
RNAi **TALENs / ZFNs and other nucleases / CRISPRs**

The Genomes of Non-Model Organisms are the New Frontiers

Rokas & Abbot (2009) Trends Ecol. Evol.

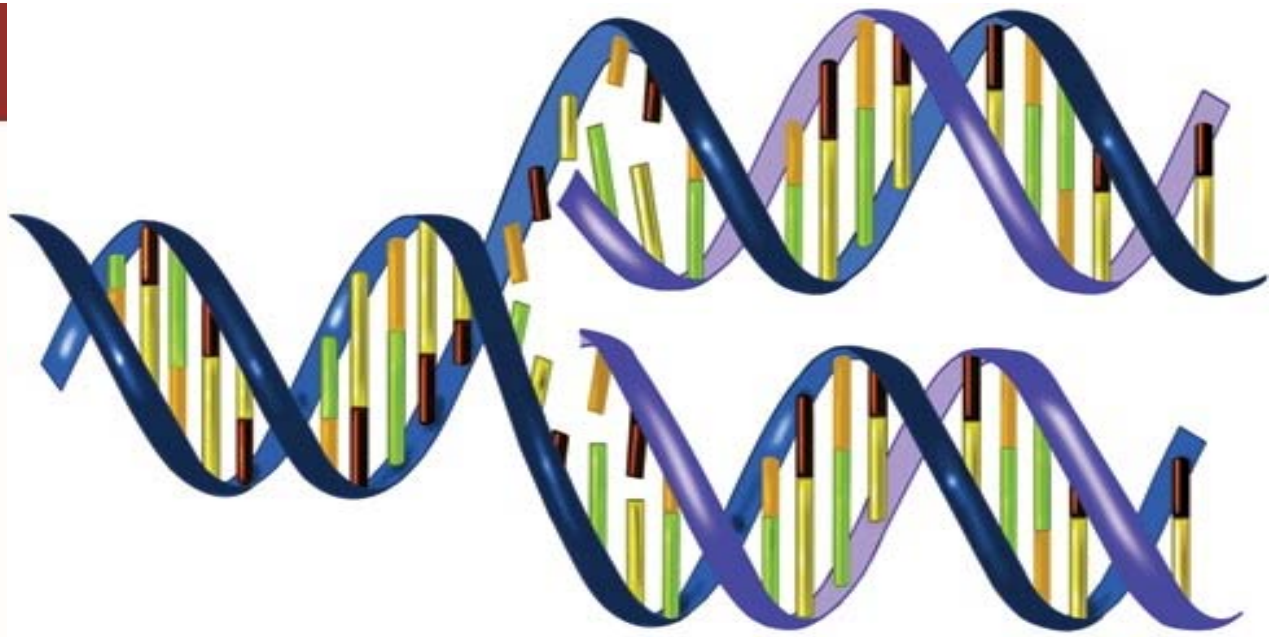
Snake Infrared Vision





Gracheva et al. (2010) Nature

The DNA Record



“The genome is, it's a fossil record; the genome is a landscape; the genome is a whole geography of distributions. [...] The genome is a storybook that's been edited for a couple of billion years, and you could take it to bed, like *A Thousand and One Arabian Nights*, and read a different story, in the genome, every night.”

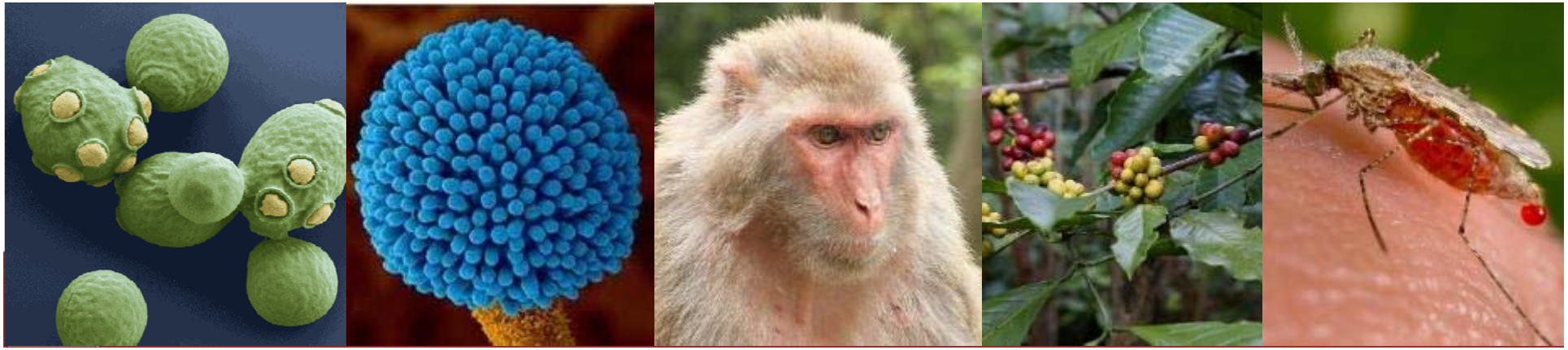
Eric Lander

The Rokas Lab

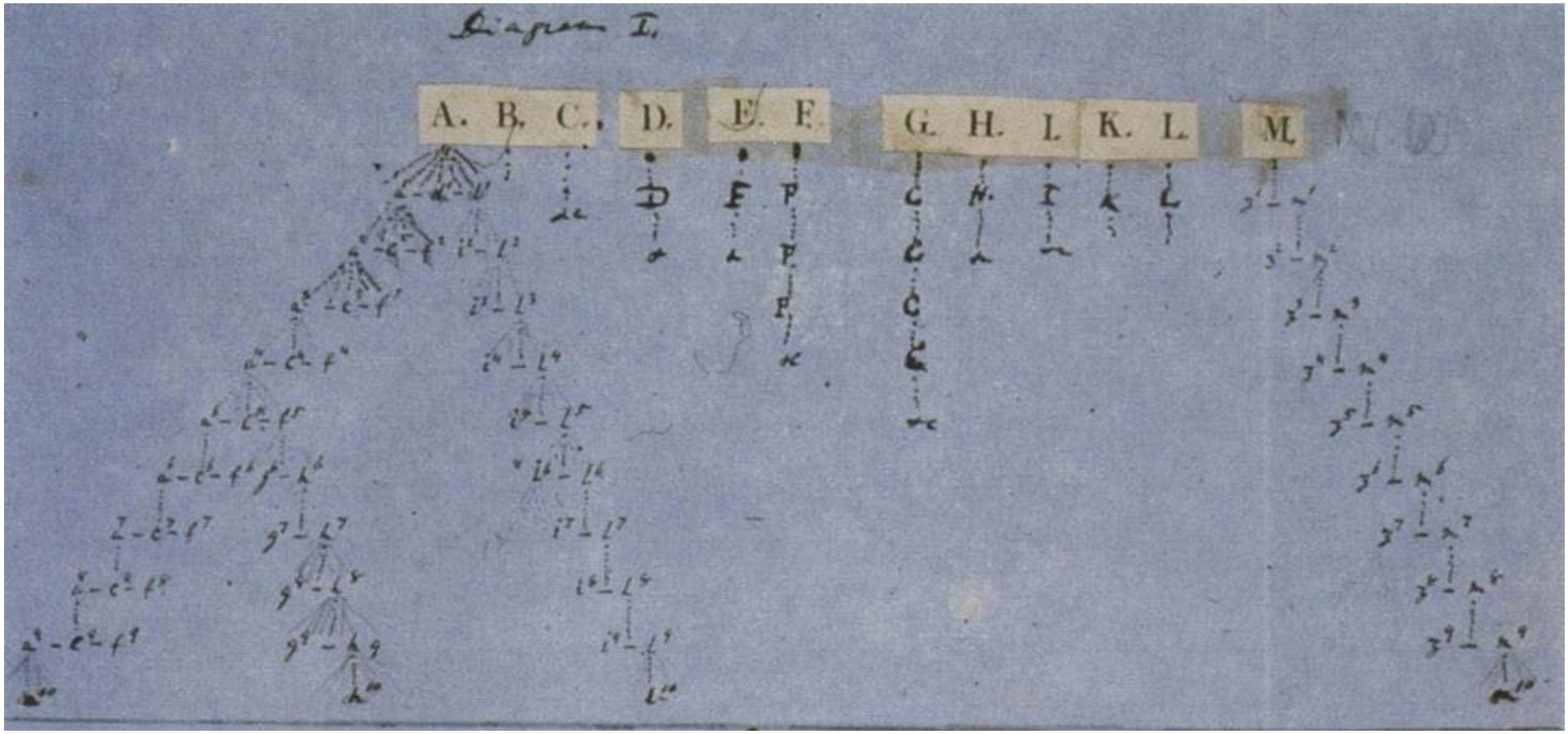
We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

Rokas Lab Research Themes

- **The evolution of fungal metabolism**
- **Phylogenomics of ancient divergences**
- **The evolution of mammalian pregnancy**



Darwin's Tree



Darwin's hand-made proof of the famous diagram from his *Origin of Species*

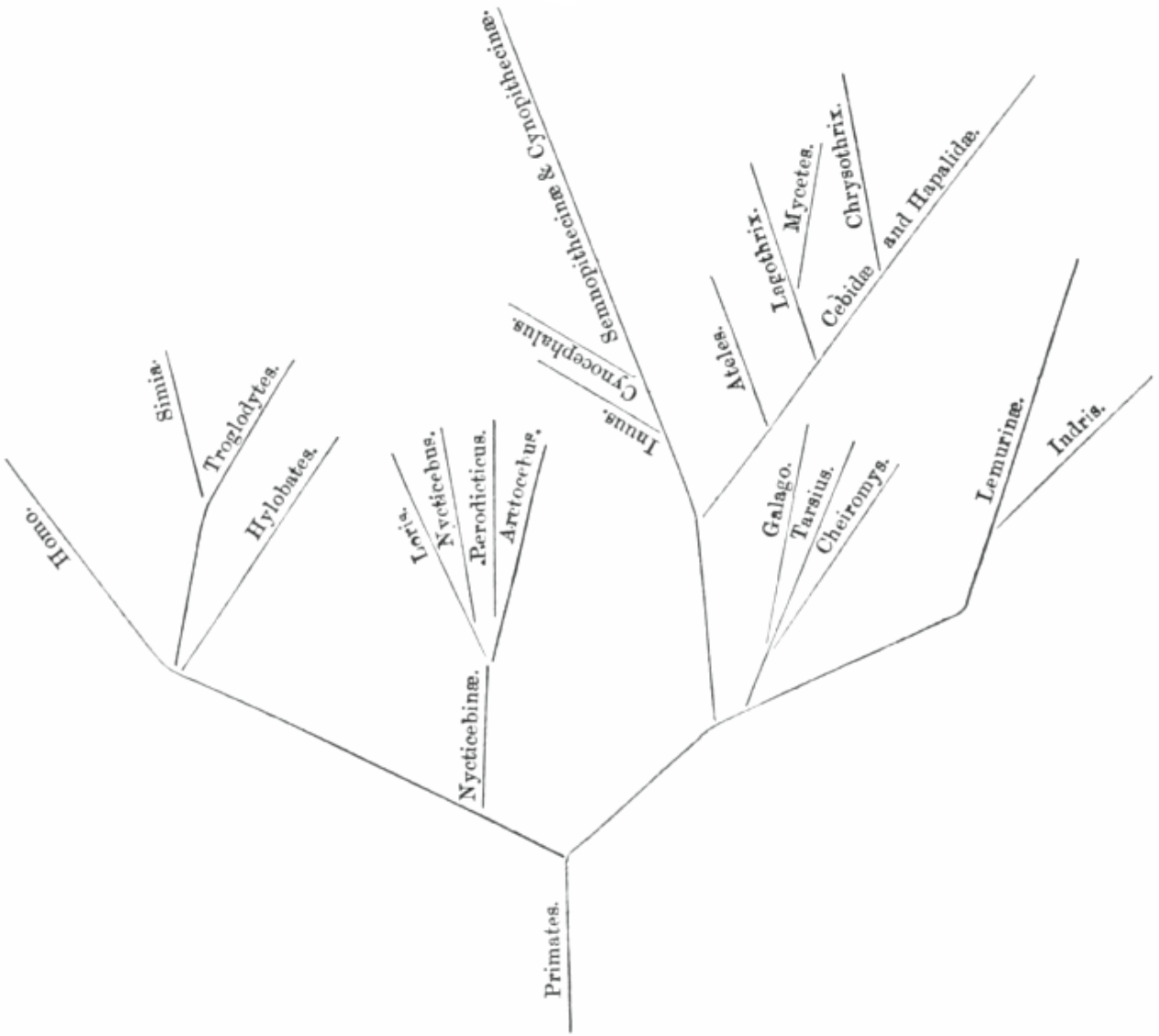
and instinct as the summing up of many contrivances, each useful to the possessor, nearly in the same way as when we look at any great mechanical invention as the summing up of the labour, the experience, the reason, and even the blunders of numerous workmen; when we thus view each organic being, how far more interesting, I speak from experience, will the study of natural history become!

A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised by man will be a far more important and interesting subject for study than one more species added to the infinitude of already recorded species. Our classifications will come to be, as far as they can be so made, genealogies; and will then truly give what may be called the plan of creation. The rules for classifying will no doubt become simpler when we have a definite object in view. We possess no pedigrees or armorial bearings; and we have to discover and trace the many diverging lines of descent in our natural genealogies, by characters of any kind which have long been inherited. Rudimentary organs will speak infallibly with respect to the nature of long-lost structures. Species and groups of species, which are called aberrant, and which may fancifully be called living fossils, will aid us in forming a picture of the ancient forms of life. Embryology will reveal to us the structure, in some degree obscured, of the prototypes of each great class.

When we can feel assured that all the individuals of the same species, and all the closely allied species of most genera, have within a not very remote period de-

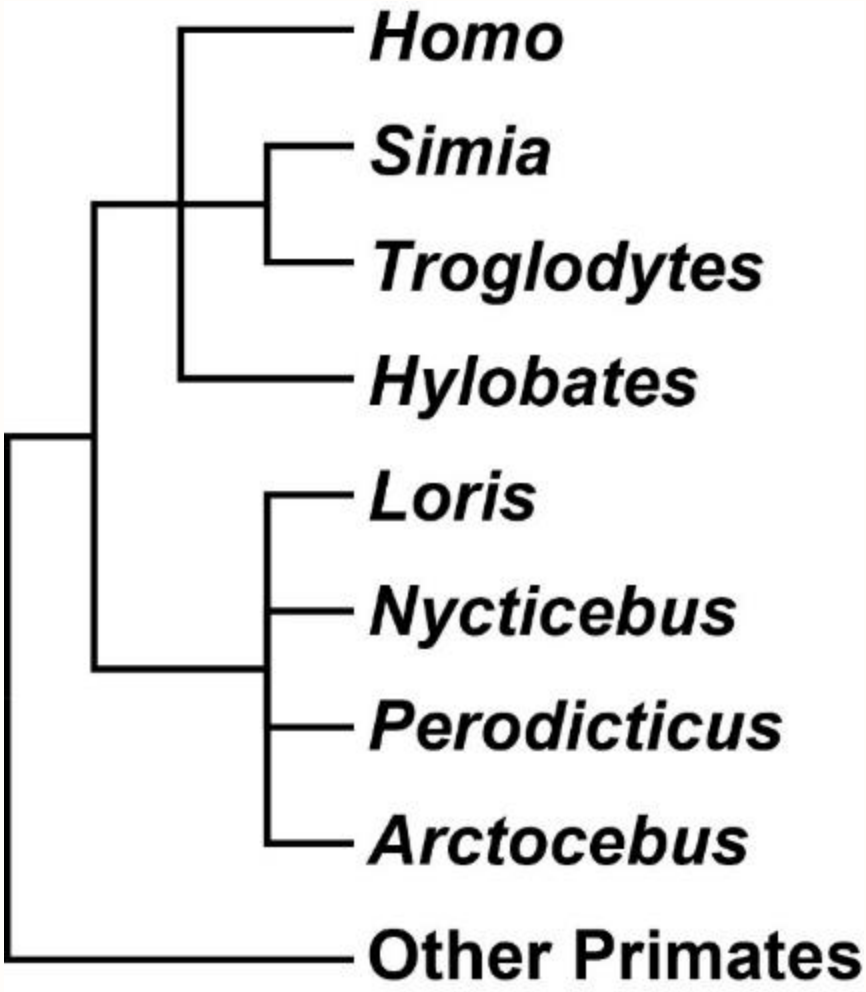
The First Published Phylogeny

St. George Jackson Mivart

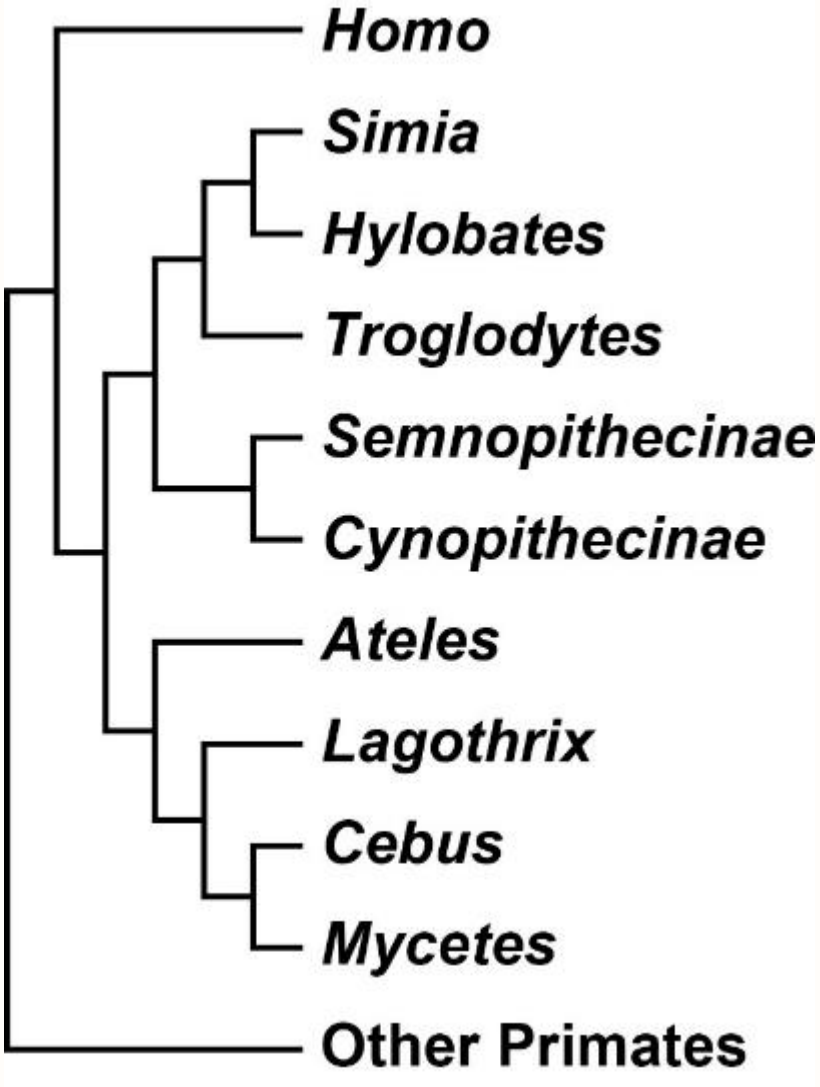


Discordance Between Trees There from the Beginning

1865: SPINAL COLUMN



1867: LIMBS

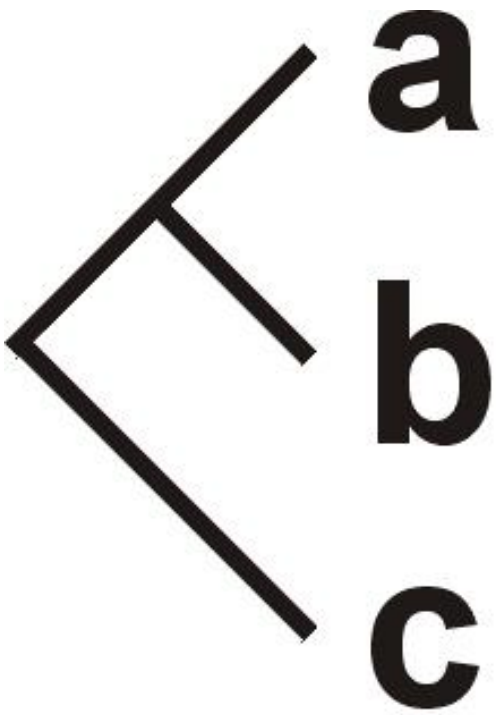


http://phylonetworks.blogspot.se/2012/09/the-first-network-from-conflicting.html

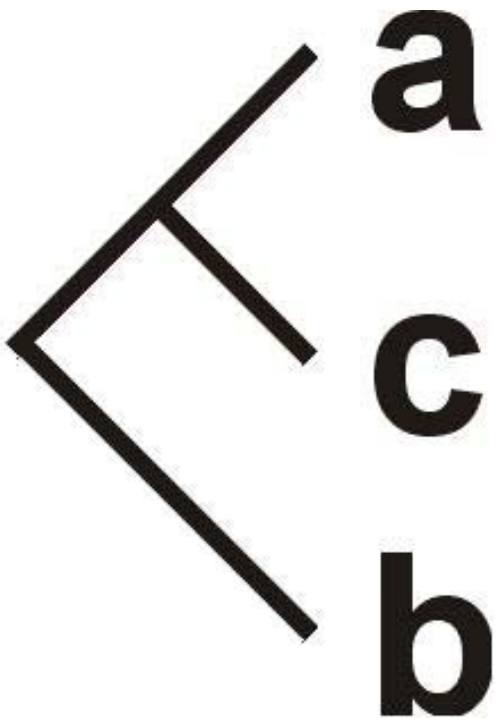
In some M.S. [… I say] that on genealogical principles alone, & considering whole organisation man probably diverged from the Catarhine stem a little below the branch of the anthropo:apes […]. I have then added in my M.S. that this is your opinion […]. Is this your opinion?

I have really expressed no opinion as to Man's origin nor am I prepared to do so at this moment. The [1865] diagram […] expresses what I believe to be the degree of resemblance as regards the spinal column *only*. The [1867] diagram expresses what I believe to be the degree of resemblance as regards the appendicular skeleton *only*

The Problem of Incongruence



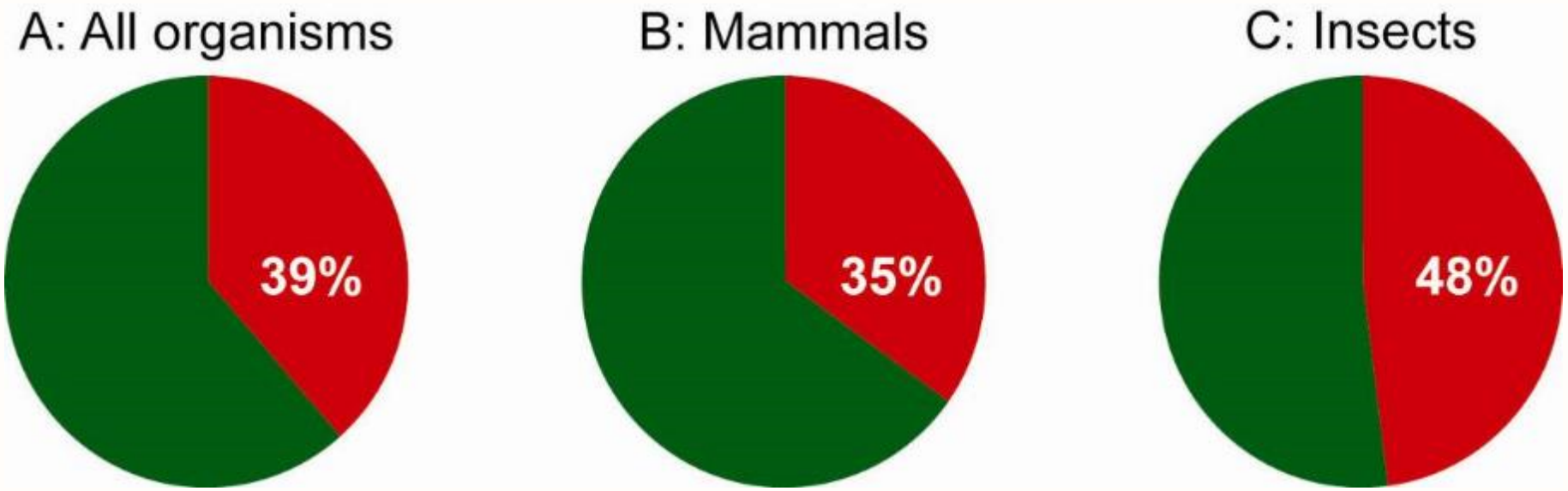
Gene X



Gene Y

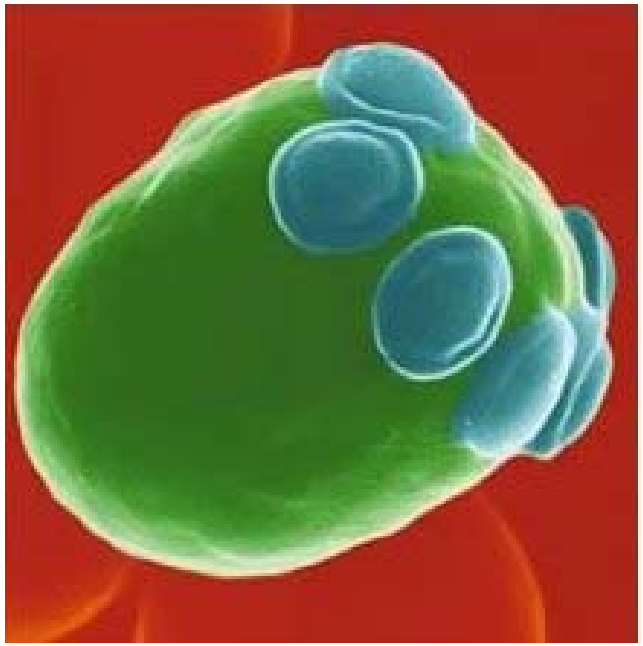
Species phylogeny?

Incongruence is Pervasive in the Phylogenetics Literature



Rokas & Chatzimanolis (2008) in Phylogenomics (W. J. Murphy, Ed.)

A Systematic Evaluation of Single Gene Phylogenies



S. cerevisiae

S. bayanus

S. paradoxus

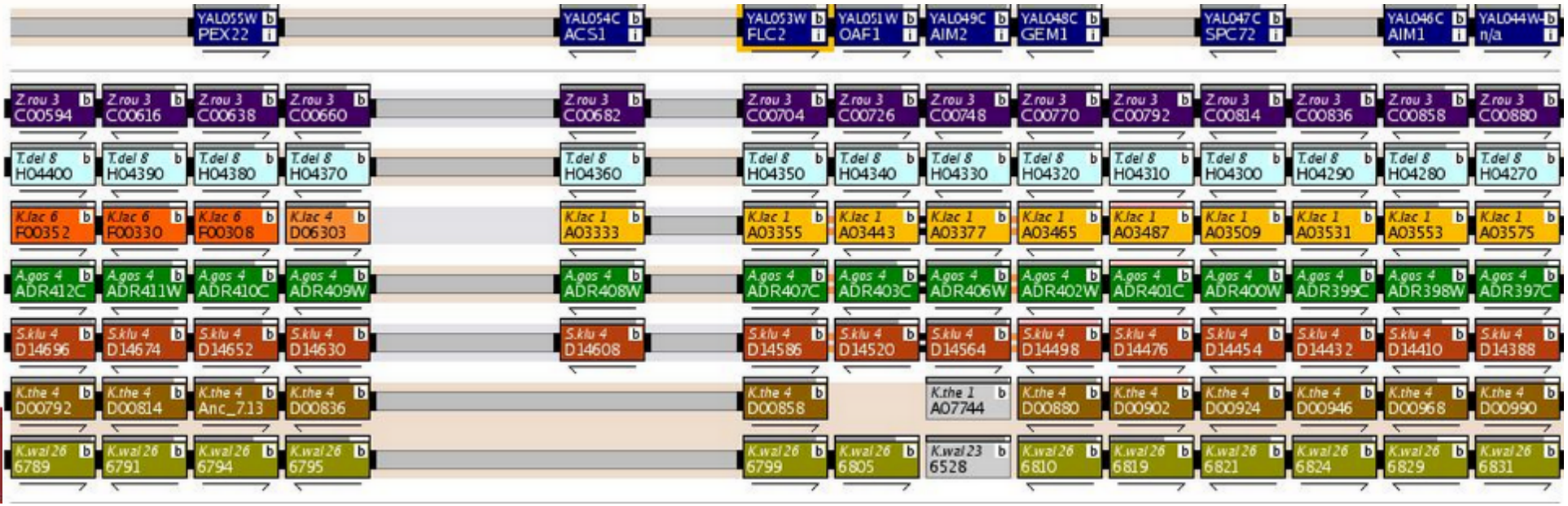
S. castellii

S. mikatae

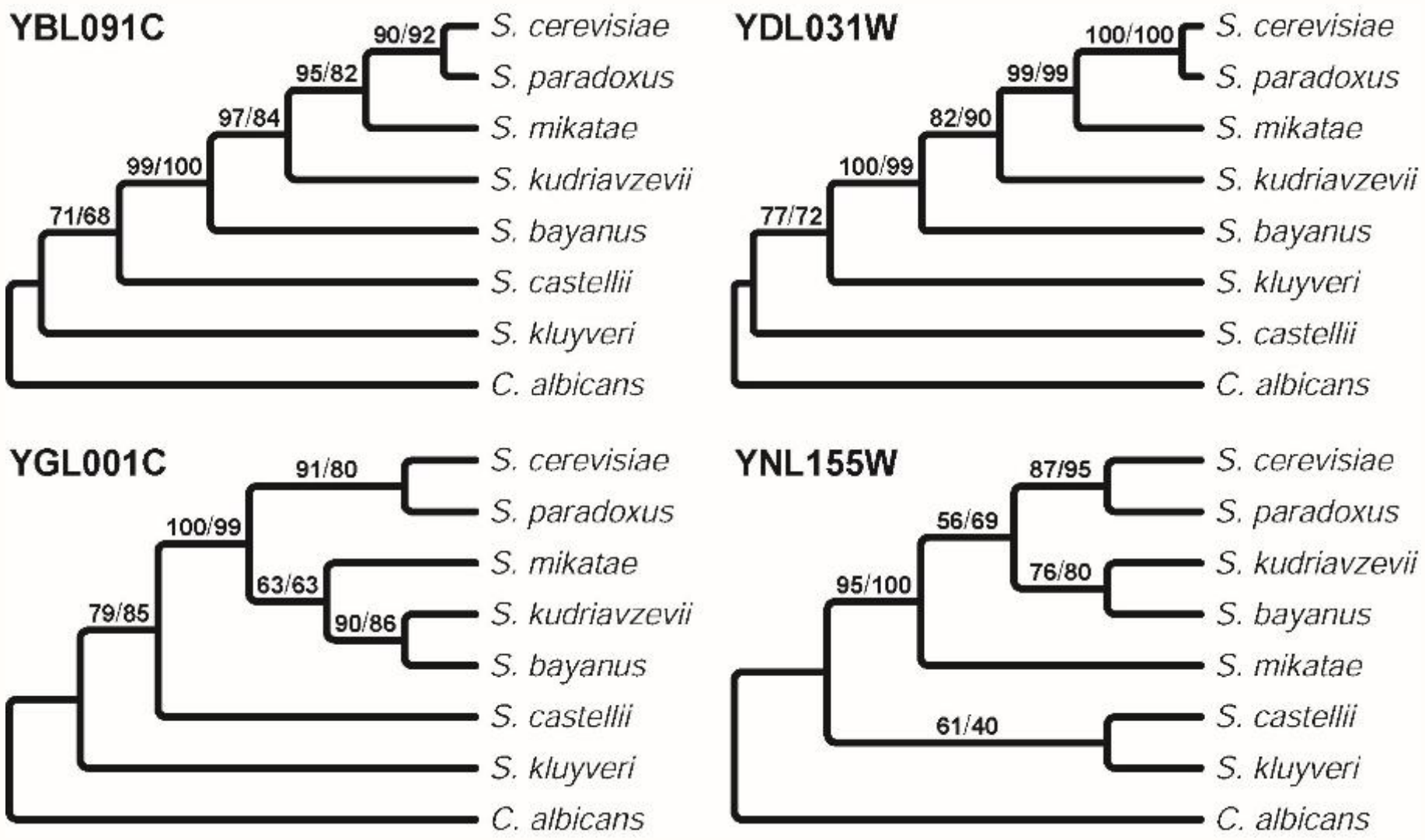
S. kluyveri

S. kudriavzevii

Candida glabrata



Incongruence at the Single Gene Level

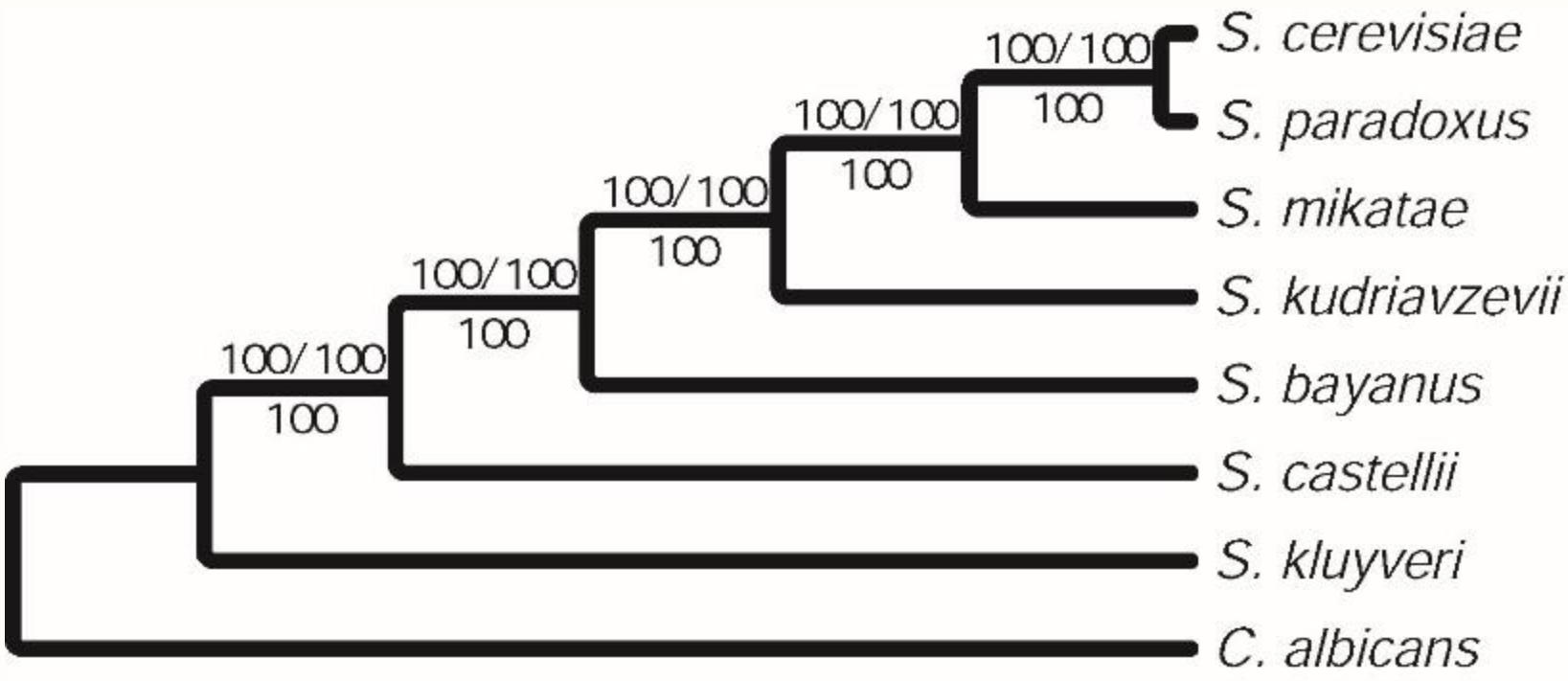


ML / MP

Anonymous Reviewer for Nature (2003)

Rokas et al. (2003) Nature

Concatenation of 106 Genes Yields a Single Yeast Phylogeny



ML / MP on nt
MP on aa

Rokas et al. (2003) Nature

The Phylogenomics Era – "Resolving" the Tree of Life

Syst. Biol. 61(1):150–164, 2012

DOI:10.1093/sysbio/syr089
Advance Access publication on September 7, 2011

LETT
LETT

Phylogenomic Analysis Resolves the Interordinal Relationships and Rapid Diversification of the Laurasiatherian Mammals

XUMING ZHOU, SHIXIA XU, JUNXIAO XU, BINGYAO CHEN, KAIYA ZHOU, AND GUANG YANG*

*Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China; *Correspondence to be sent to: Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China; E-mail: gyang@njnu.edu.cn.*

Resolving the evolutionary relationships of molluscs with phylogenomic tools

Stephen A. Smith[1,2], Nerida G. Wilson[3,4], Freya
Gonzalo Giribet[5] & Casey W. Dunn[1]

nature

LETTERS

Syst. Biol. 57(6):920–938, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802570791

Toward Resolving t
Tree: The Phylogen
of Jakobids and Cercozoans

Resolving Arthropod Phylogeny: Exploring Phylogenetic Signal within 41 kb of Protein-Coding Nuclear Gene Sequence

JEROME C. REGIER,[1] JEFFREY W. SHULTZ,[2] AUSTEN R. D. GANLEY,[3,6] APRIL HUSSEY,[1] DIANE SHI,[1] BERNARD BALL,[3] ANDREAS ZWICK,[1] JASON E. STAJICH,[3,7] MICHAEL P. CUMMINGS,[4] JOEL W. MARTIN,[5] AND CLIFFORD W. CUNNINGHAM[3]

Yeast

An

OPEN ACCESS Free

Toward Resolving Priors

Prion-Like Proteins in the Fungal Kingdom

Edgar M. Medina · Gary W. Jones · David A. Fitzpatrick

Towards

Renae C. Pratt, Gillian C. Gibb,* Mary Morgan-Richards,* Matthew J. Phillips,† Michael D. Hendy,* and David Penny**

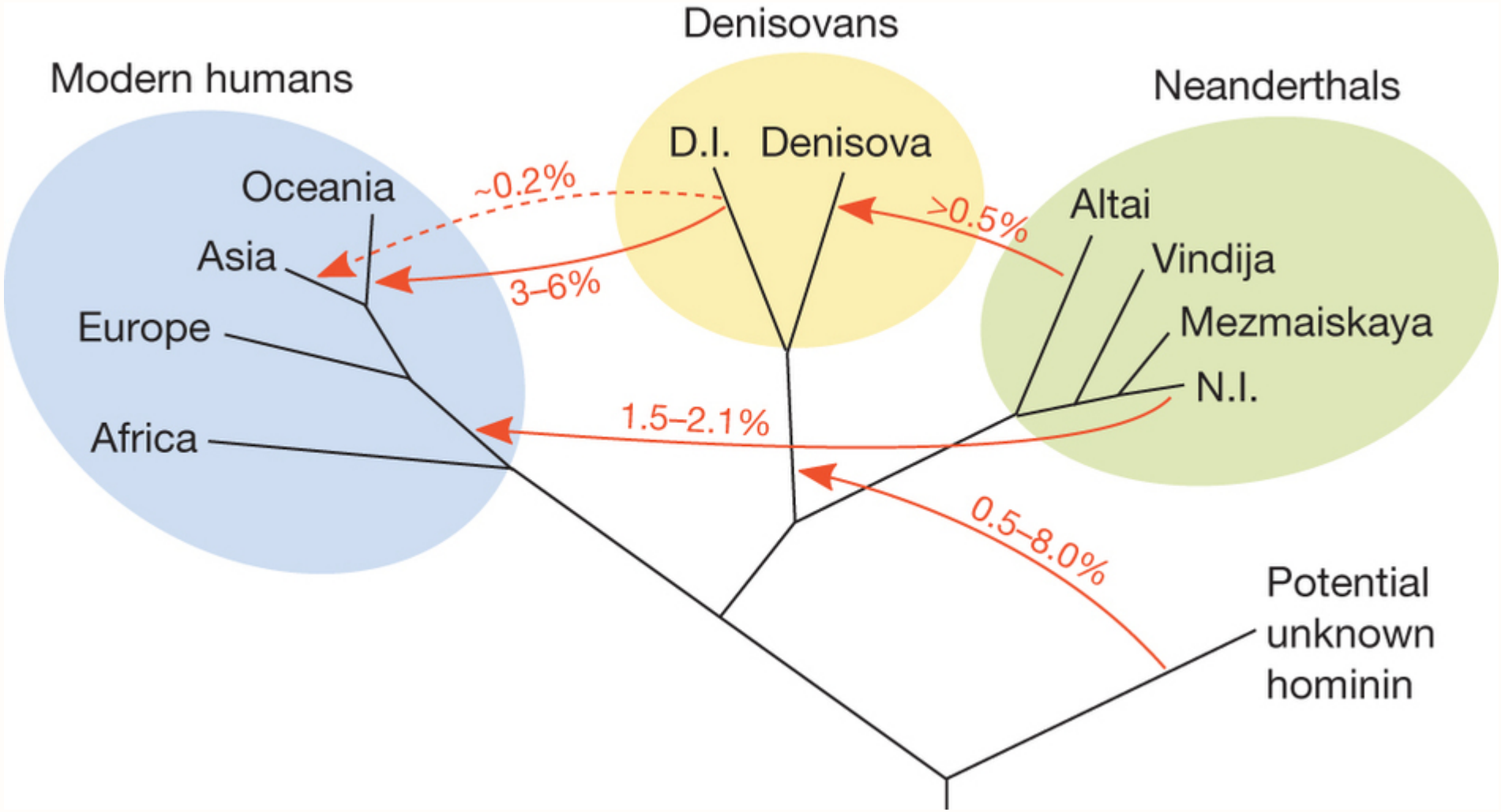
*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and †Centre for Macroevolution and Macroecology, School of Botany and Zoology, Australian National University, Canberra ACT, Australia

Samuli Lehtonen

Department of Biology, U

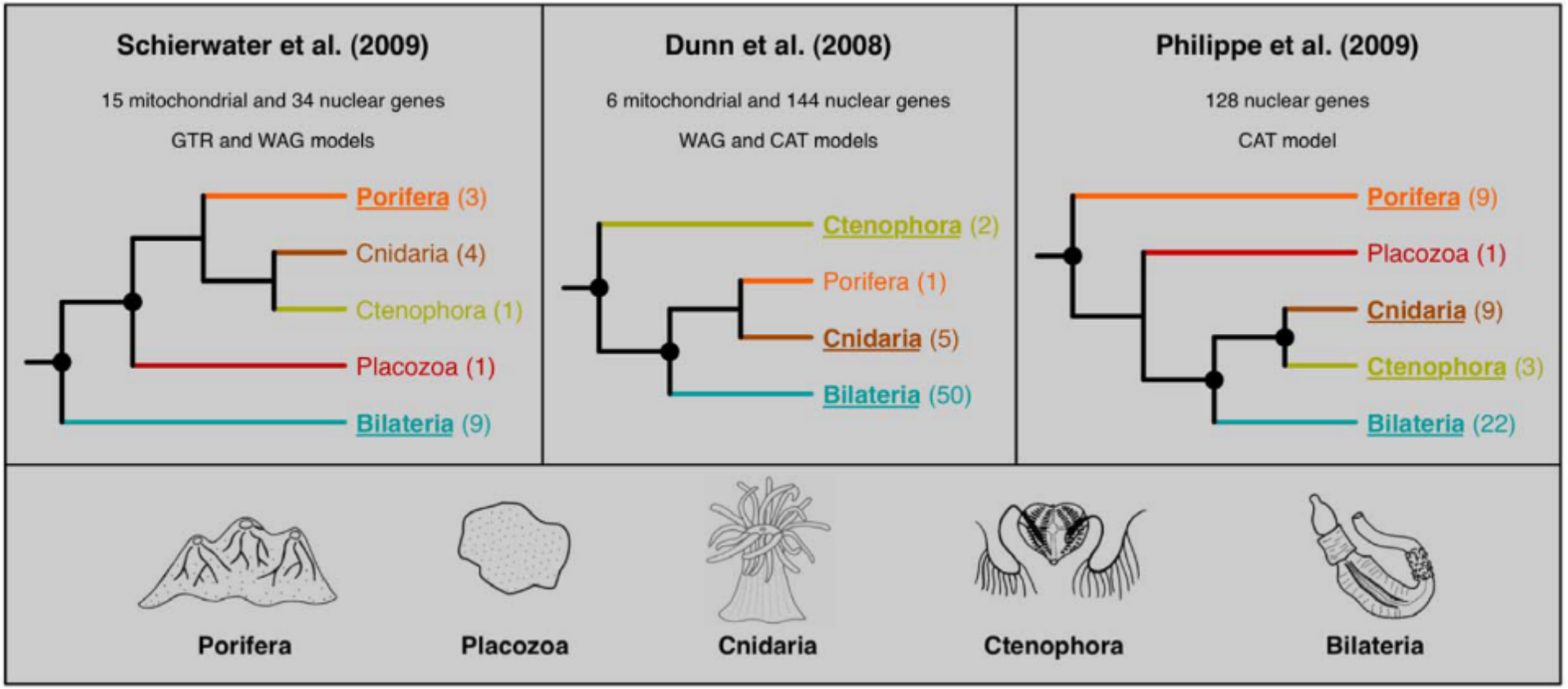
Have we eliminated incongruence?

At "Shallow" Depths, True History is Easier Seen & Quantified



Prufer et al. (2014) Nature

Incongruence in Deep Time is More Challenging



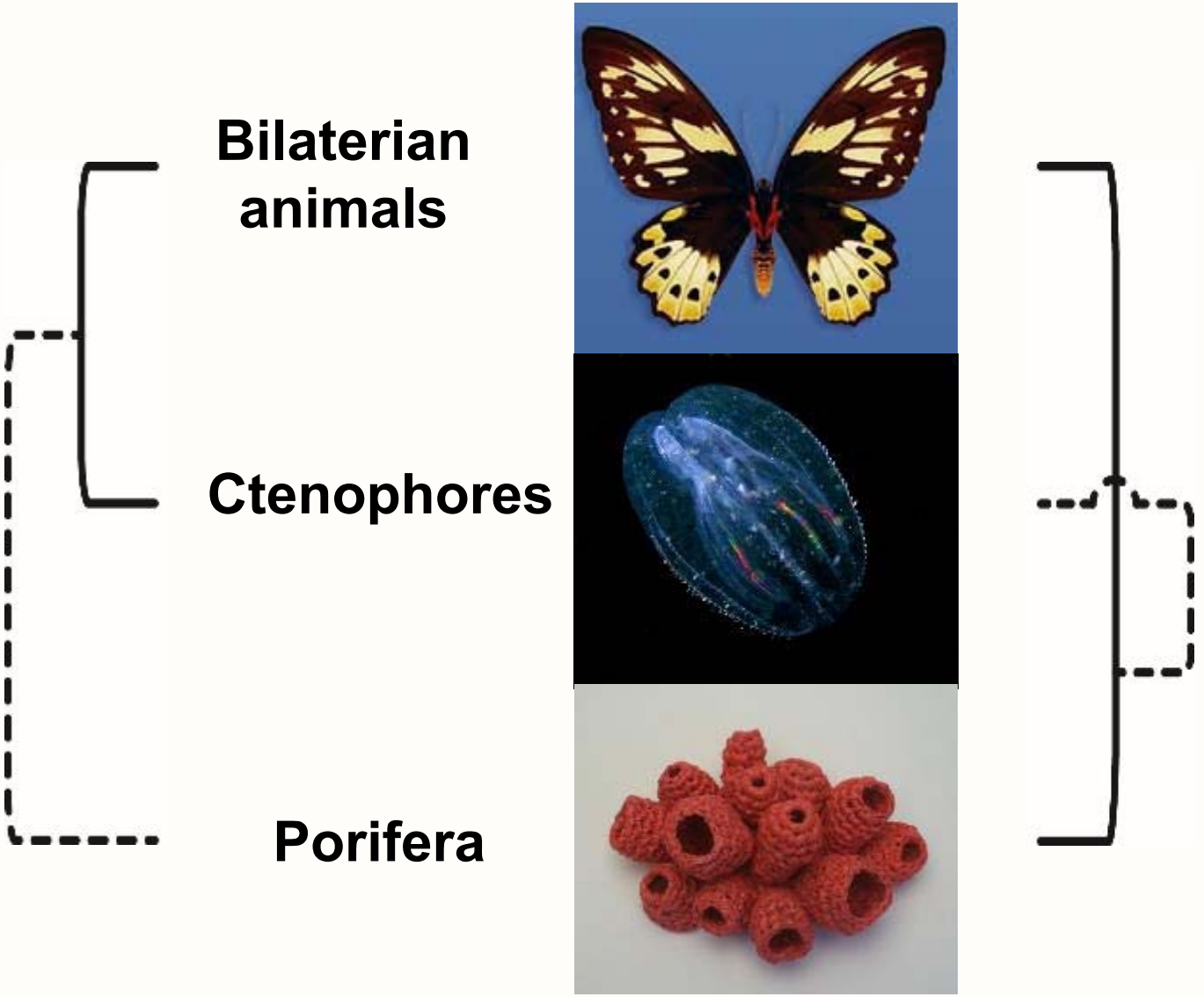
Philippe et al. (2011) PLoS Biol.

Incongruence in Deep Time is More Challenging

Kocot et al. (2011) Nature

Smith et al. (2011) Nature

Incongruence in Deep Time is More Challenging



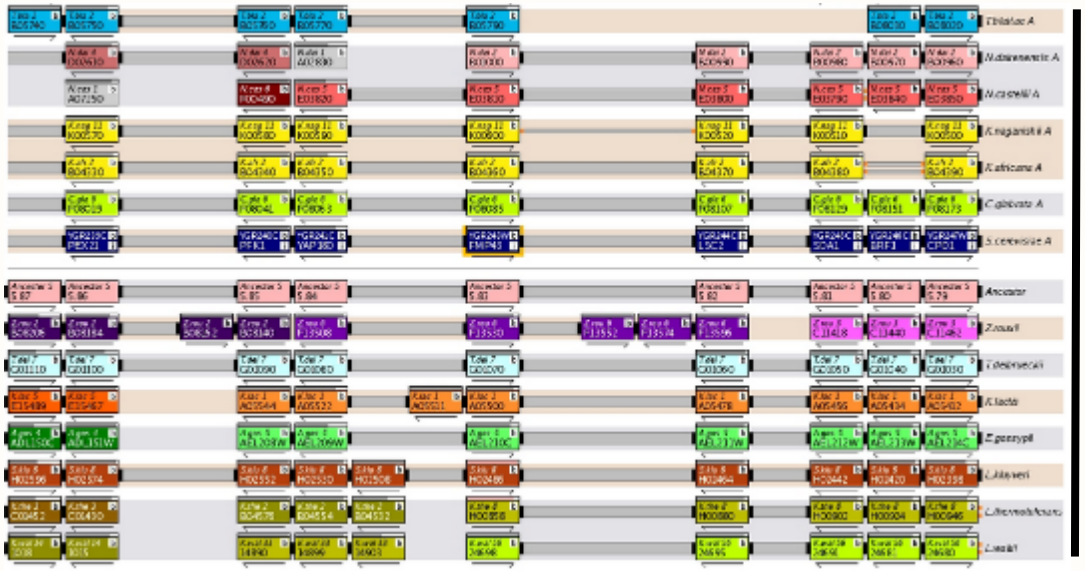
Pisani et al. (2015) PNAS

Chang et al. (2015) PNAS

Why the disconnect?

An Expanded Yeast Data Matrix

Yeast Gene Order Browser (YGOB)



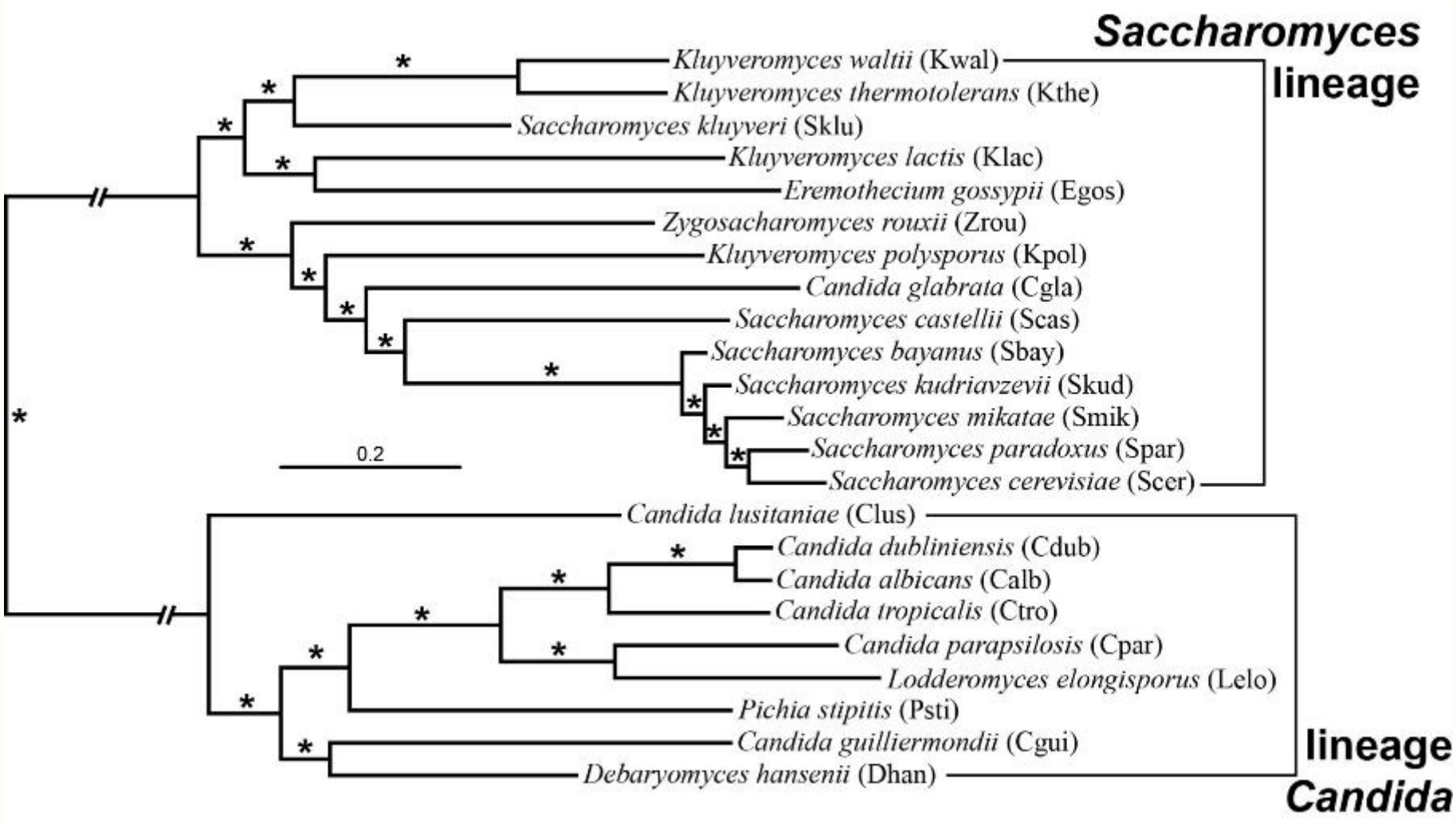
Candida Gene Order Browser (CGOB)

***Saccharomyces* lineage**

***Candida* lineage**

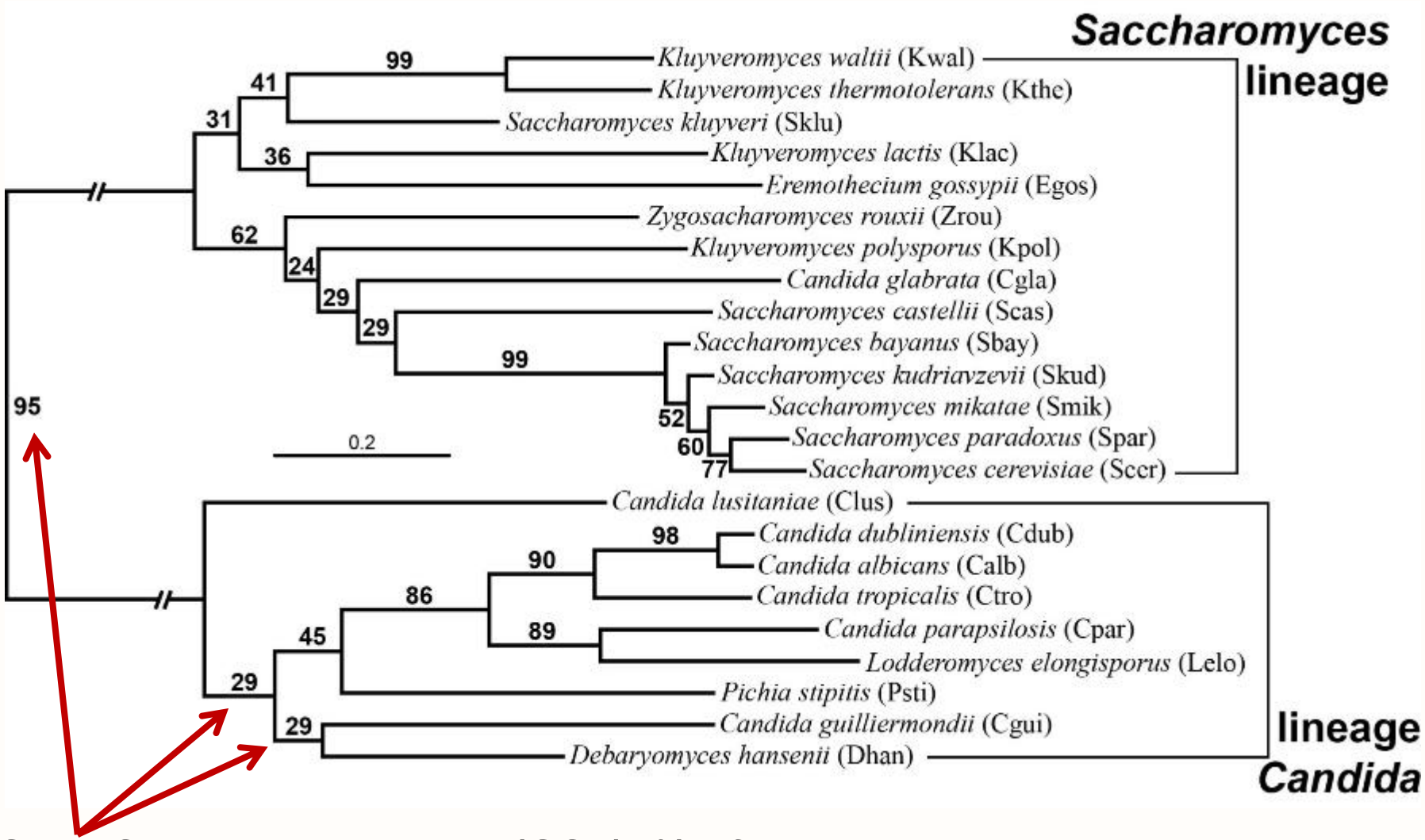
1,070 genes
23 taxa
no missing data

Concatenation Yields an Absolutely Supported Phylogeny



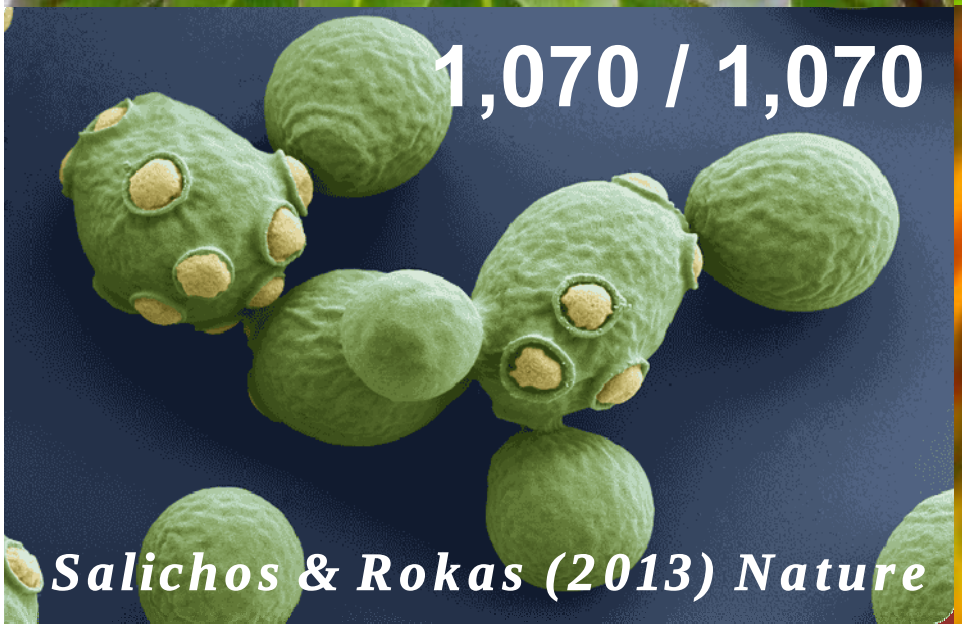
Salichos & Rokas (2013) Nature

The Yeast Phylogeny Inferred by Majority-Rule Consensus



Gene Support Frequency (GSF): % of single gene trees supporting a given internode

Gene Trees are Incongruent in Most Datasets

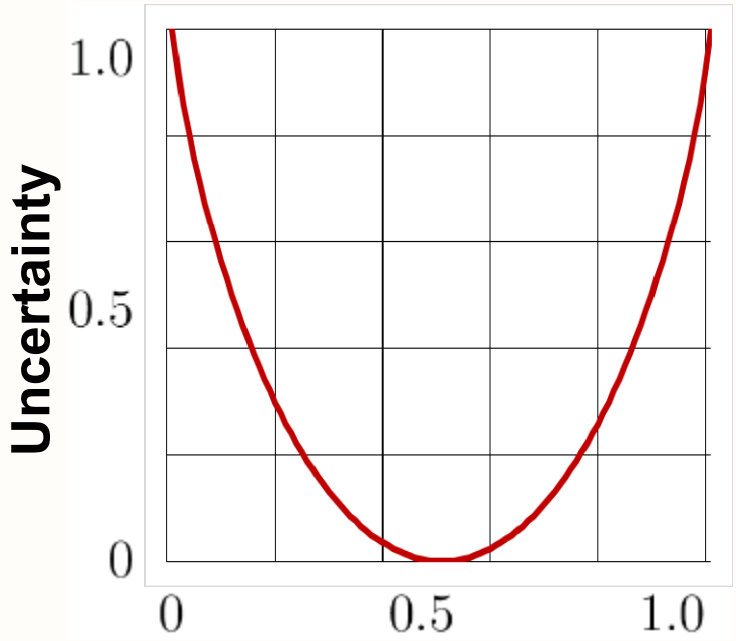


Quantifying Incongruence

Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting internode in the same set of trees

Tree Certainty (TC): the sum of IC across all internodes

IC and TC are implemented in the latest versions of RAxML

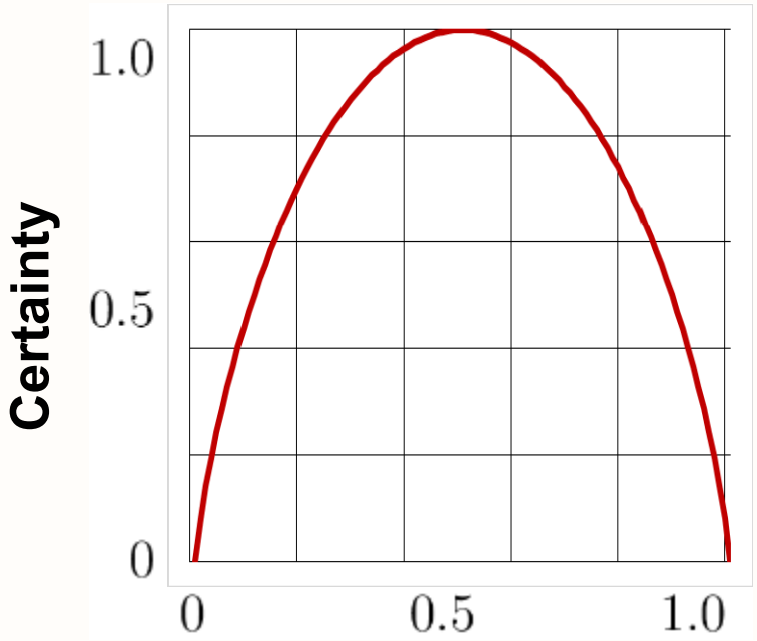


Quantifying Incongruence

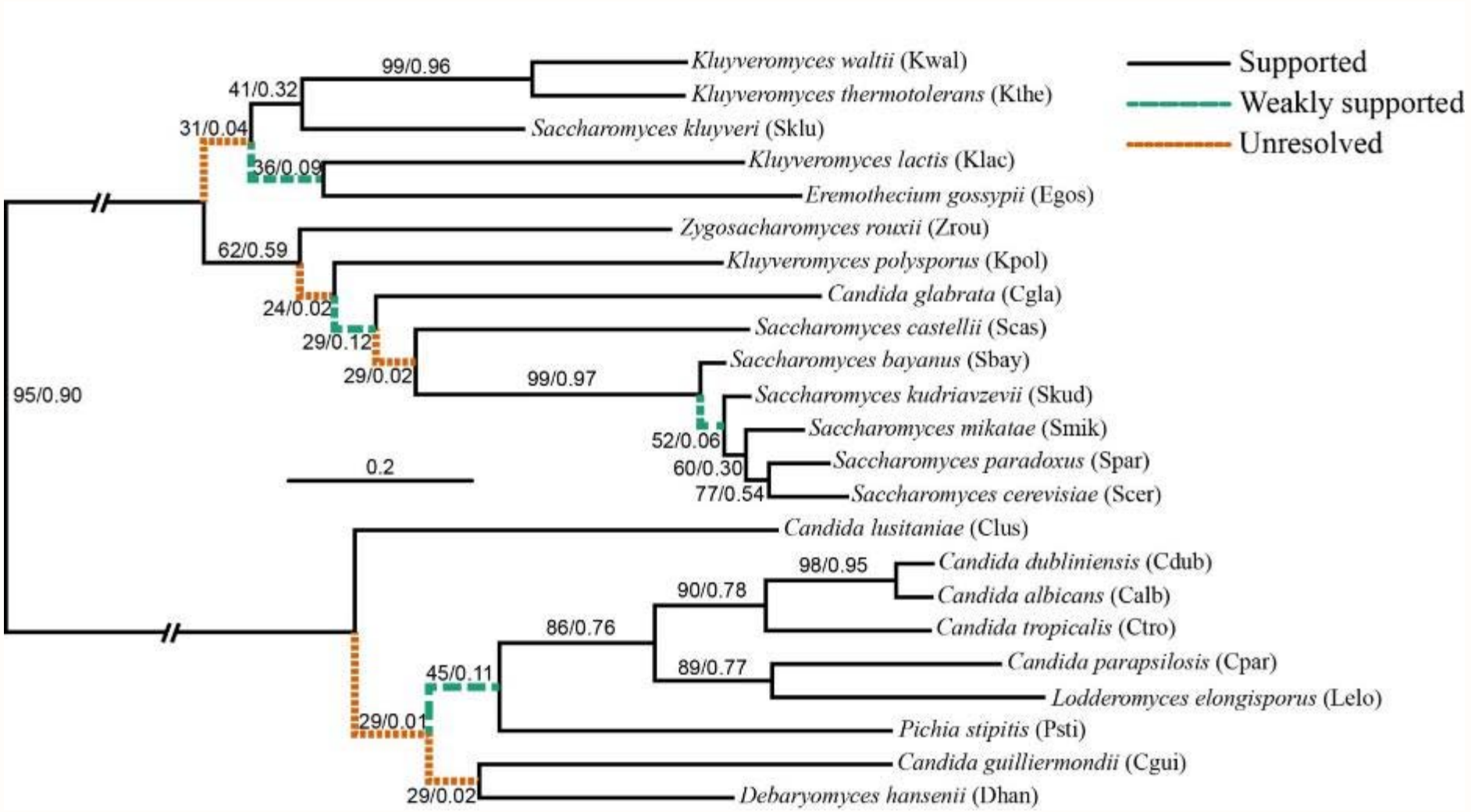
Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting internode in the same set of trees

Tree Certainty (TC): the sum of IC across all internodes

IC and TC are implemented in the latest versions of RAxML



Some Internodes are Poorly Supported at the Gene Level



Gene Support Frequency / Internode Certainty

Similar Results in Other Lineages

Vertebrates
(1,086 genes, 18 taxa)

Animals
(225 genes, 21 taxa)

Mosquitoes
(2,007 genes, 20 taxa)

Salichos & Rokas (2013) Nature; Wang et al. (2015) Genome Biol. Evol.

Incongruence in Phylogenomic Datasets

These debates concern internodes that are poorly supported by individual gene trees

Coffee Break

What is the phylogenetic signal in branches of the tree of life that are challenging to resolve?

Definitions of Phylogenetic Signal

A measure of the statistical dependence among species' trait values due to their phylogenetic relationships / the tendency of related species to resemble each other more than species drawn at random from the same tree

Revell et al. (2008) Syst. Biol.
Münkemüller et al. (2012) Methods Ecol. Evol.

The amount of support for a particular topology, e.g., the relative number of resolved internodes in a consensus tree

Sanderson (2008) Science

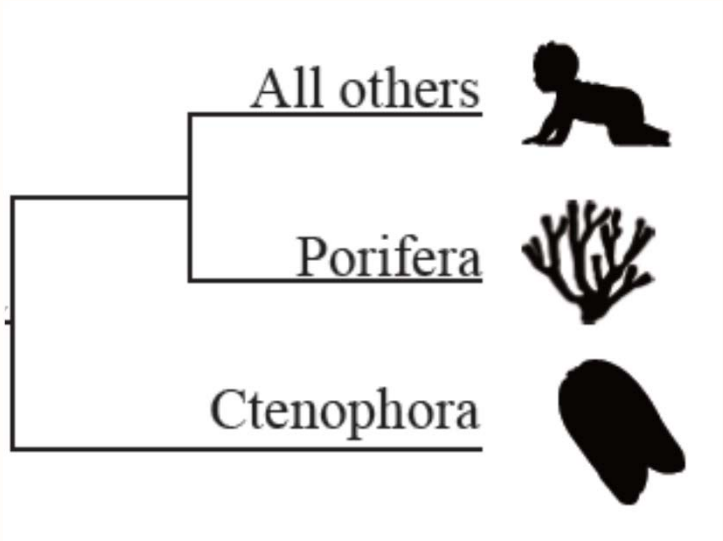
A measure of the substitutions occurring along a given branch of the evolutionary tree. In parsimony methods, the signal is encoded in shared derived characters. In probabilistic methods, the amount of phylogenetic signal actually extracted from a given dataset depends on the model and is expected to increase with the fit of the model to the data

Philippe et al. (2011) PLoS Biol.
Townsend et al. (2012) Syst. Biol.

Our Definition

Maximum Likelihood tree

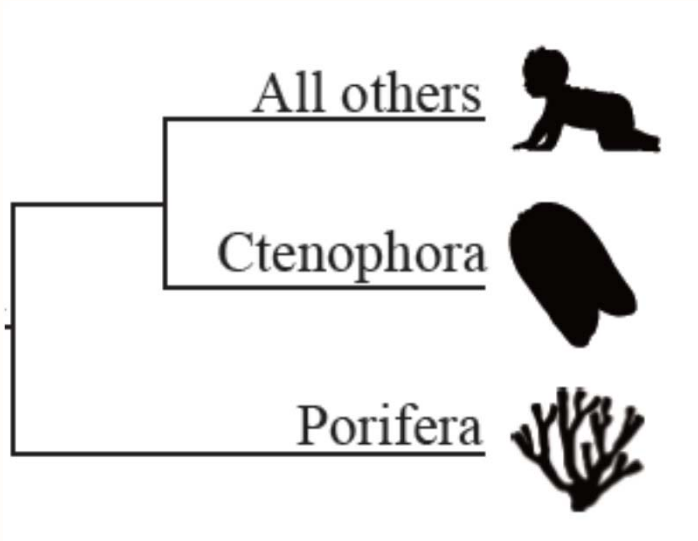
(T1)



$$\ln(T_1|X_i) = -100$$

Conflicting tree

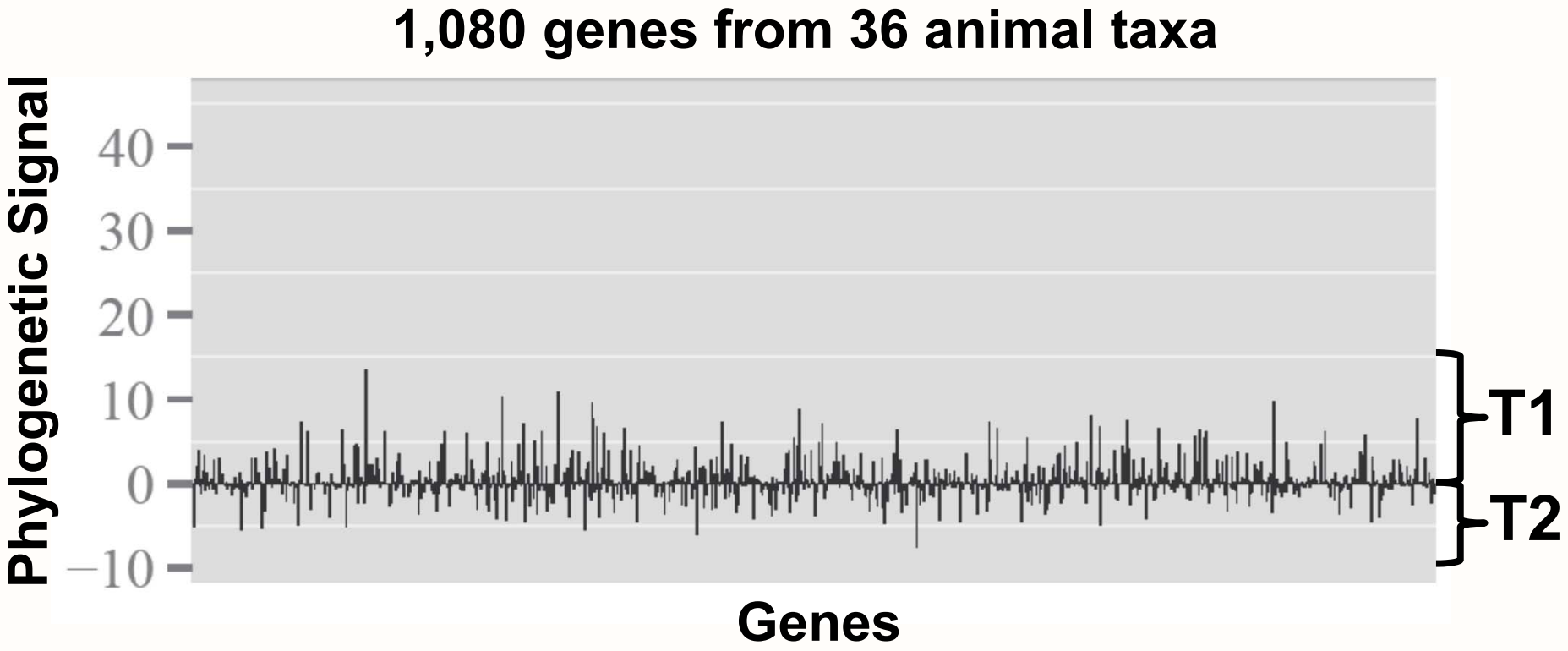
(T2)



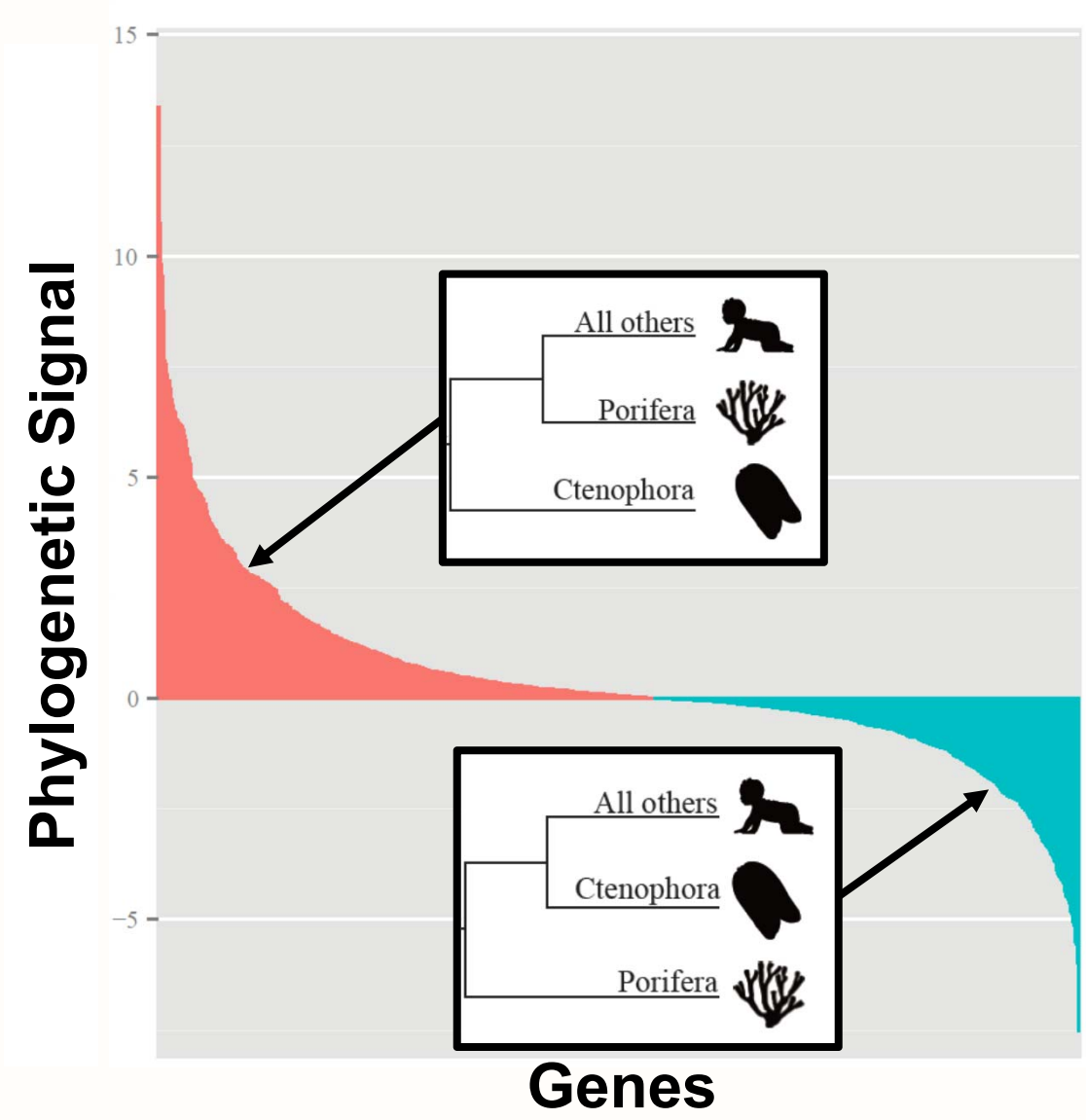
$$\ln(T_2|X_i) = -150$$

$$\textit{Phylogenetic Signal} = -(\ln(T_1|X_i) - \ln(T_2|X_i))$$

Signal of the Genes in a Phylogenomic Data Matrix

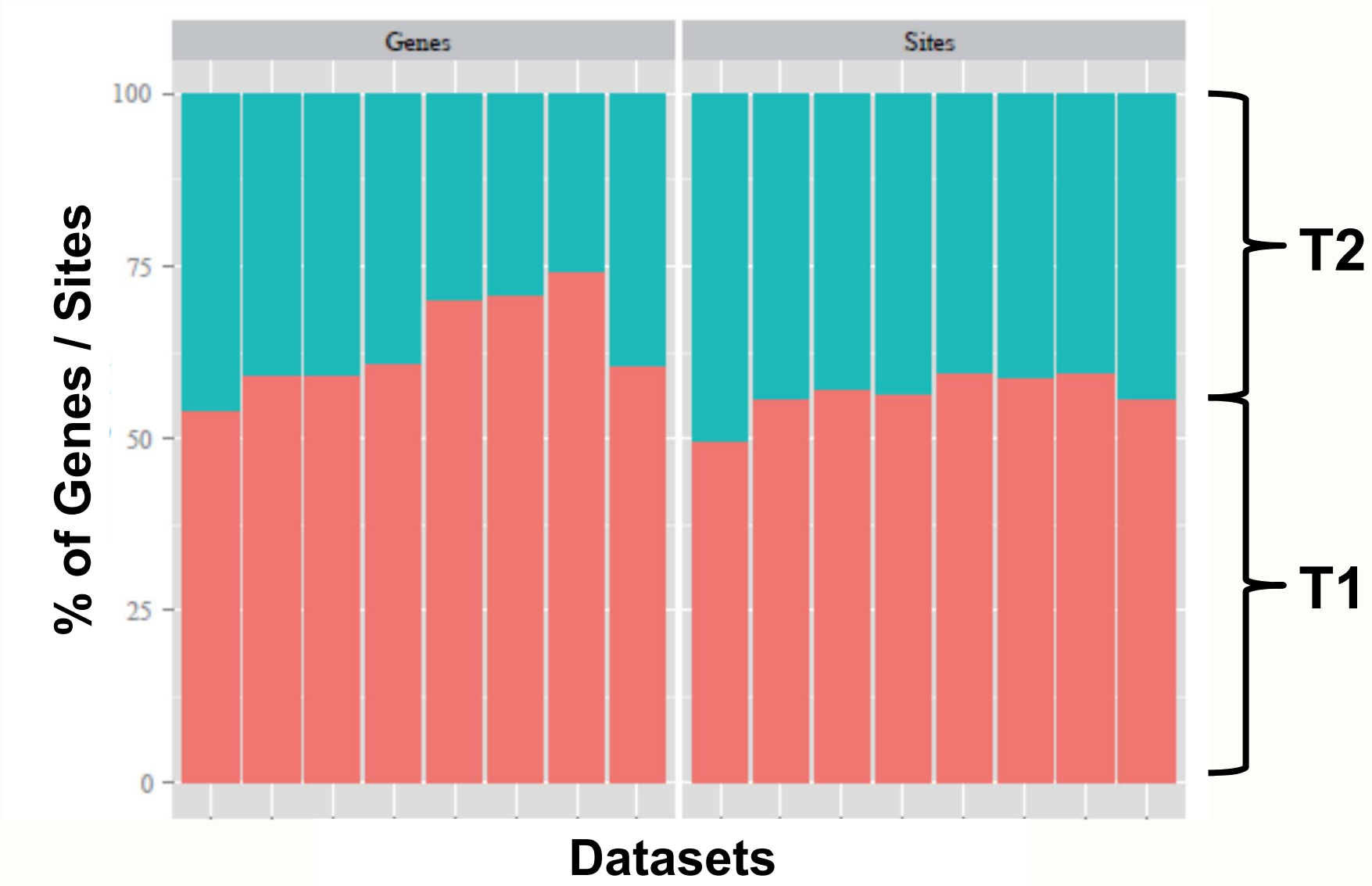


Signal of the Genes in a Phylogenomic Data Matrix



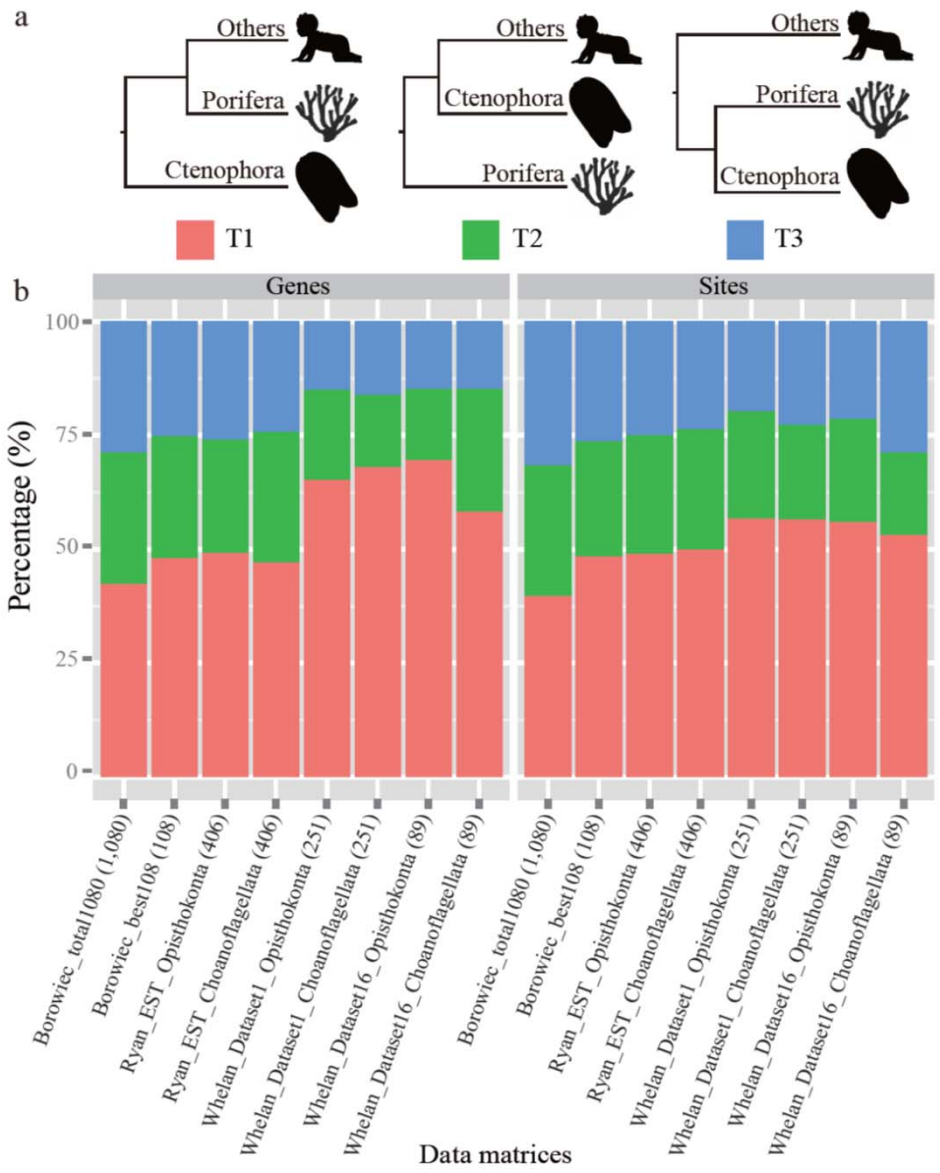
Shen et al. (2017) Nature Ecol. Evol.

Summarizing Phylogenetic Signal Across Genes and Sites



Shen et al. (2017) Nature Ecol. Evol.

Summarizing the Signal Across All 3 Possible Topologies



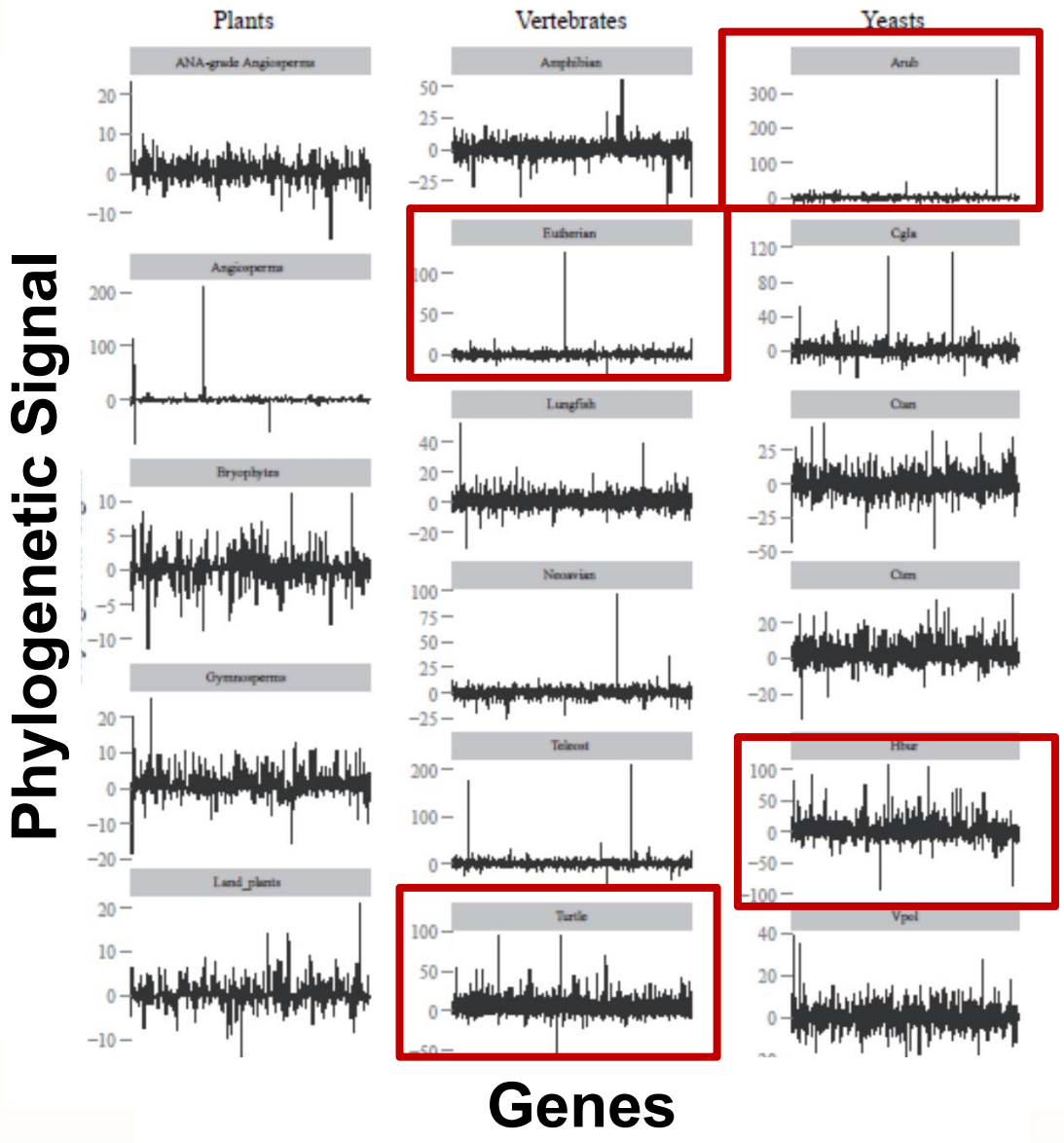
Shen et al. (2017) Nature Ecol. Evol.

Testing Several Contentious Branches of the Tree of Life

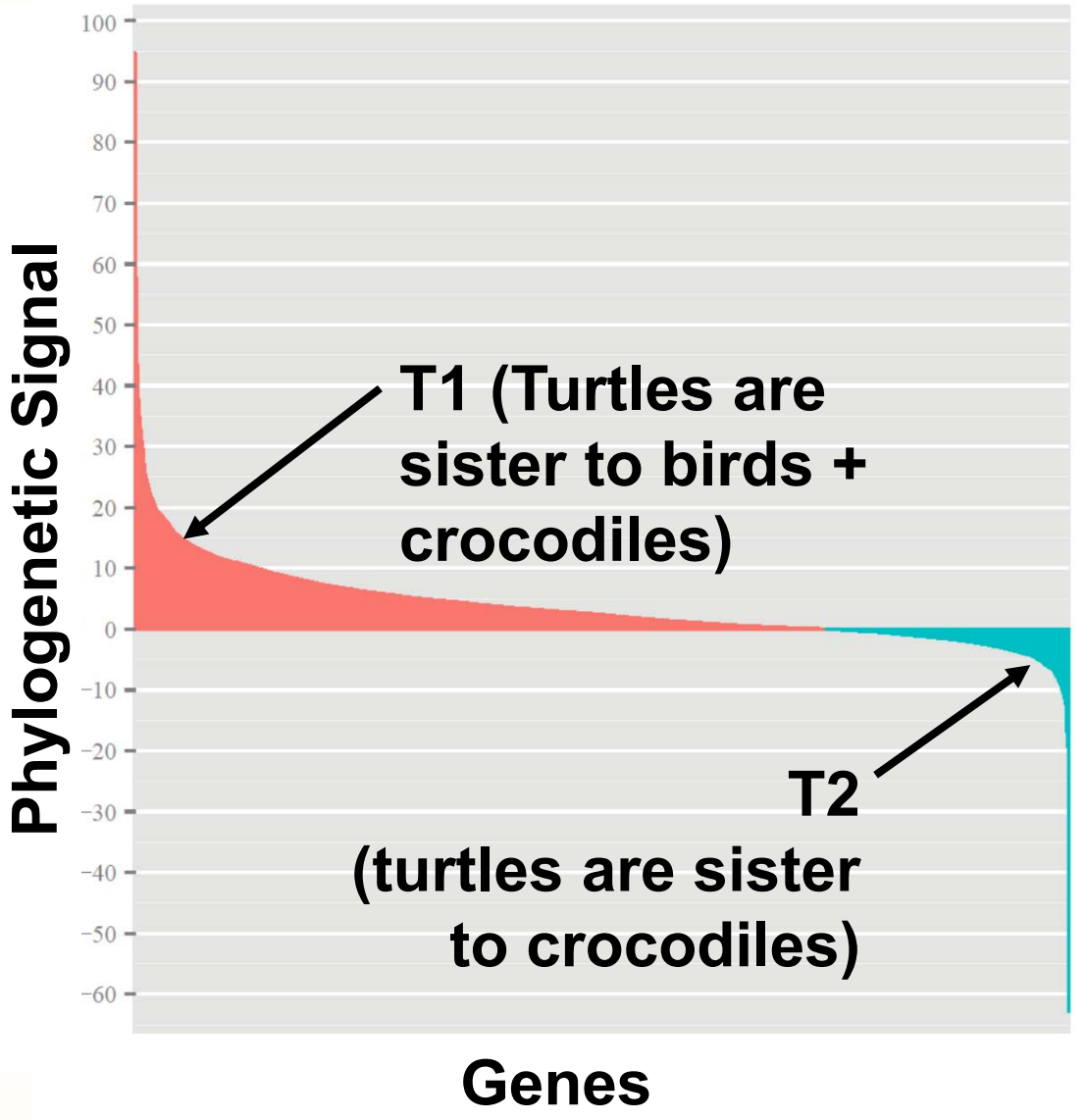
Clade	ML Tree (T1)	Conflicting Tree (T2)
Plants	*Amborella* as sister to all other flowering plants	*Amborella* + *Nuphar* as sister to all other flowering plants
	Magnoliids as sister to Eudicots + Chloranthales	Eudicots as sister to Magnoliids + Chloranthales
	Hornworts as sister to all other land plants, followed by a mosses + liverworts clade	Hornworts as sister to a mosses + liverworts clade
	Gnetales as sister to the Pinaceae, nested within the Coniferales	Gnetales as sister to the Coniferales
	Zygnematophyceae as sister to all land plants	Charales as sister to all land plants
Vertebrates	Gymnophiona as sister to all other amphibians	Anura as sister to all other amphibians
	Atlantogenata (Afrotheria + Xenarthra) as sister to all other placental mammals	Afrotheria as sister to all other placental mammals
	Lungfishes as sister to all tetrapods	Lungfishes + coelacanths as sister to all tetrapods
	Pigeons as sister to all other Neoaves	Falcons as sister to all other Neoaves
	Elopomorpha + Osteoglossomorpha as sister to all other teleosts	Osteoglossomorpha alone as sister to all other teleosts
	Turtles as sister to archosaurs (birds + crocodiles)	Turtles as sister to crocodiles
Yeasts	Ascoideaceae as sister to Phaffomycetaceae + Saccharomycetaceae	Ascoideaceae as sister to a clade comprising Pichiaceae, Debaryomycetaceae, Phaffomycetaceae, and Saccharomycetaceae
	Candida glabrata rather than *Naumovozyma castellii* as sister to Saccharomyces sensu stricto yeasts	*Naumovozyma castellii* rather than *Candida glabrata* sister to Saccharomyces sensu stricto yeasts
	Hyphopichia burtonii as sister to *Candida auris* + *Metschnikowia bicuspidata*	*Hyphopichia burtonii* as sister to *Debaryomyces hansenii*
	Zygosaccharomyces rouxii as sister to all other yeasts with occurring whole-genome duplication event	*Vanderwaltozyma polyspora* as sister to all other yeast with occurring whole-genome duplication event
	Meyerozyma guilliermondii as sister to *Debaryomyces hansenii*	*Meyerozyma guilliermondii* as sister to *Hyphopichia burtonii* + *Candida auris*
	Candida tanzawaensis as sister to *Pichia stipiti* + *Candida maltosa*	*Pichia stipiti* as sister to *Candida tanzawaensis* + *Candida maltosa*

Shen et al. (2017) Nature Ecol. Evol.

Phylogenetic Signal in Contentious Branches

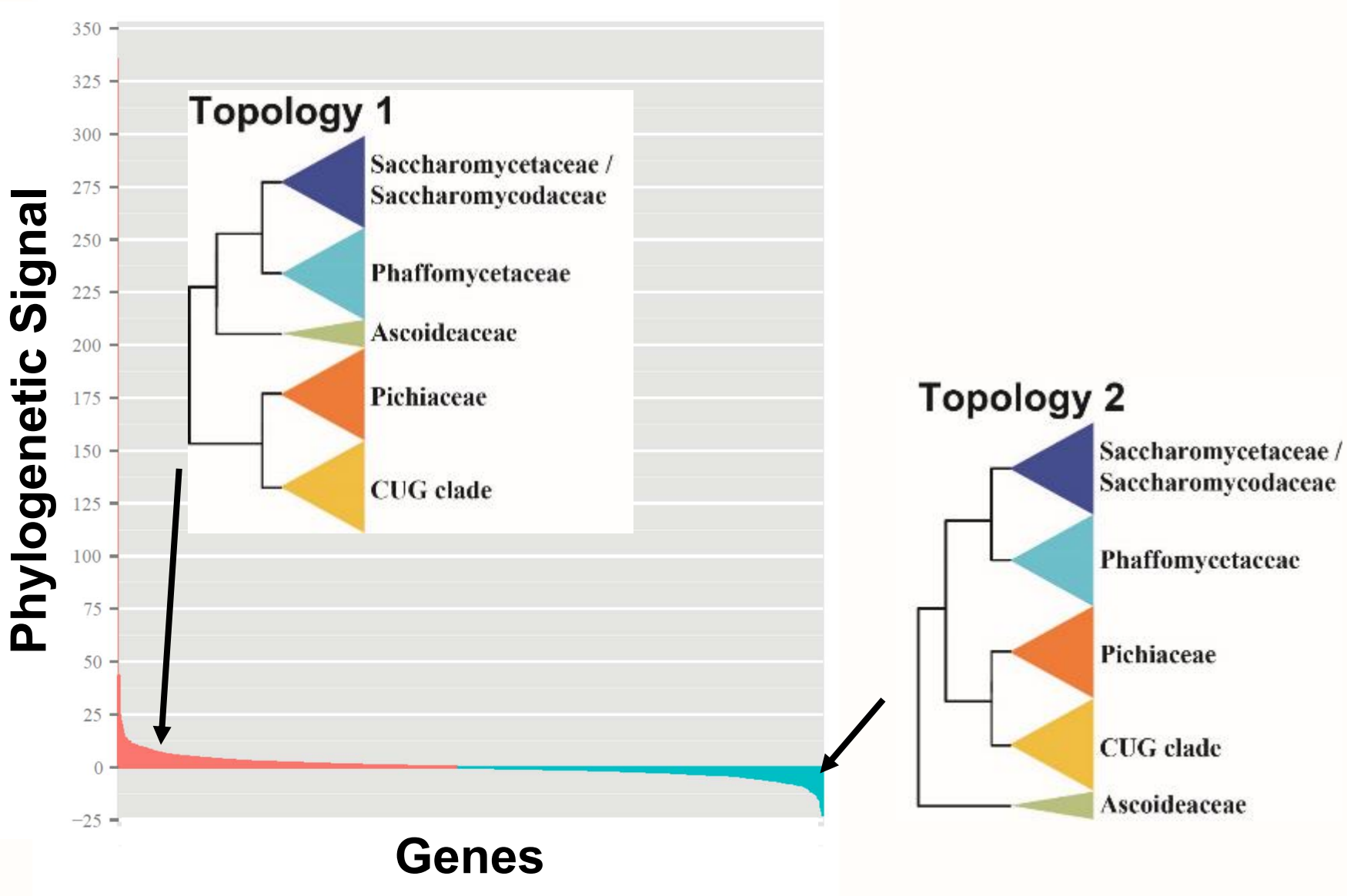


The Signal in Some Branches is Very Strong…

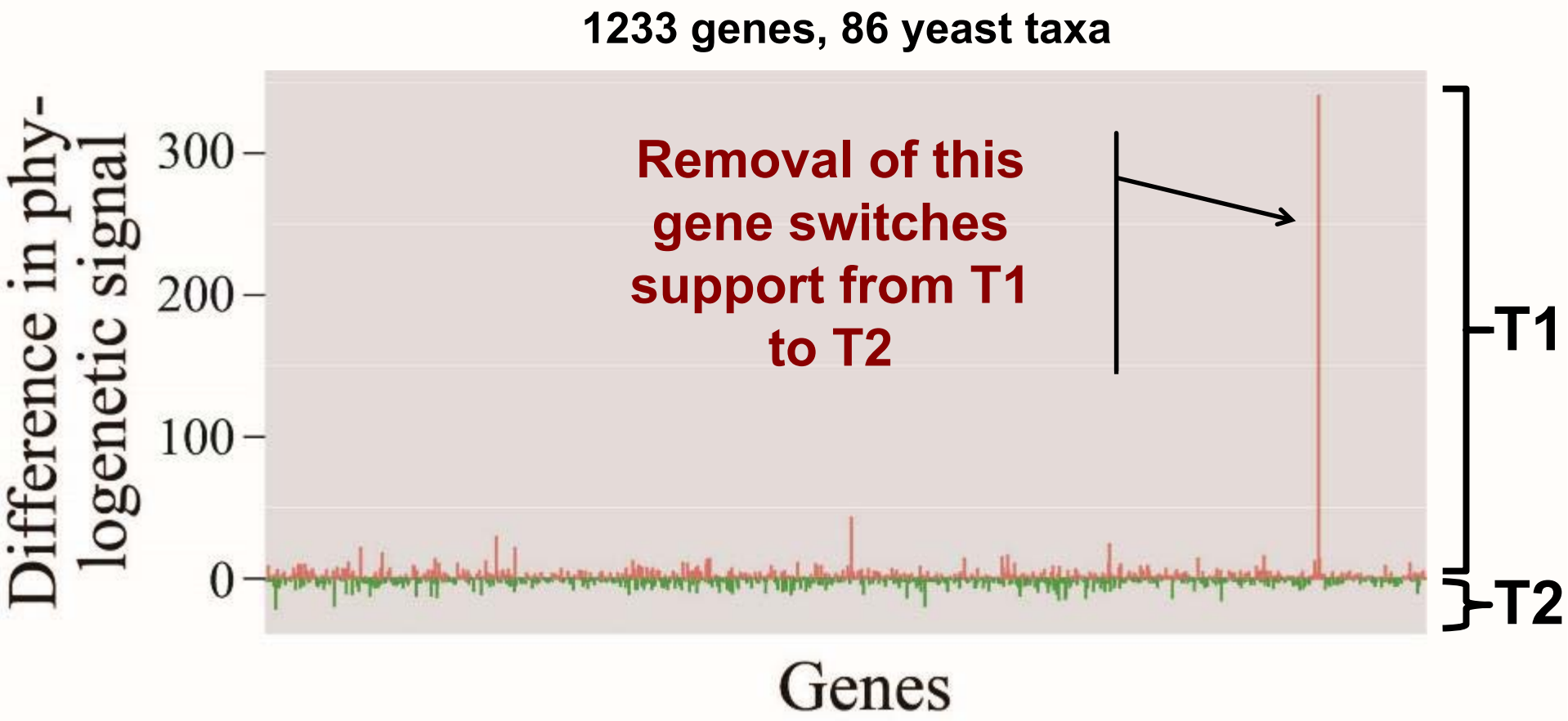


Shen et al. (2017) Nature Ecol. Evol.

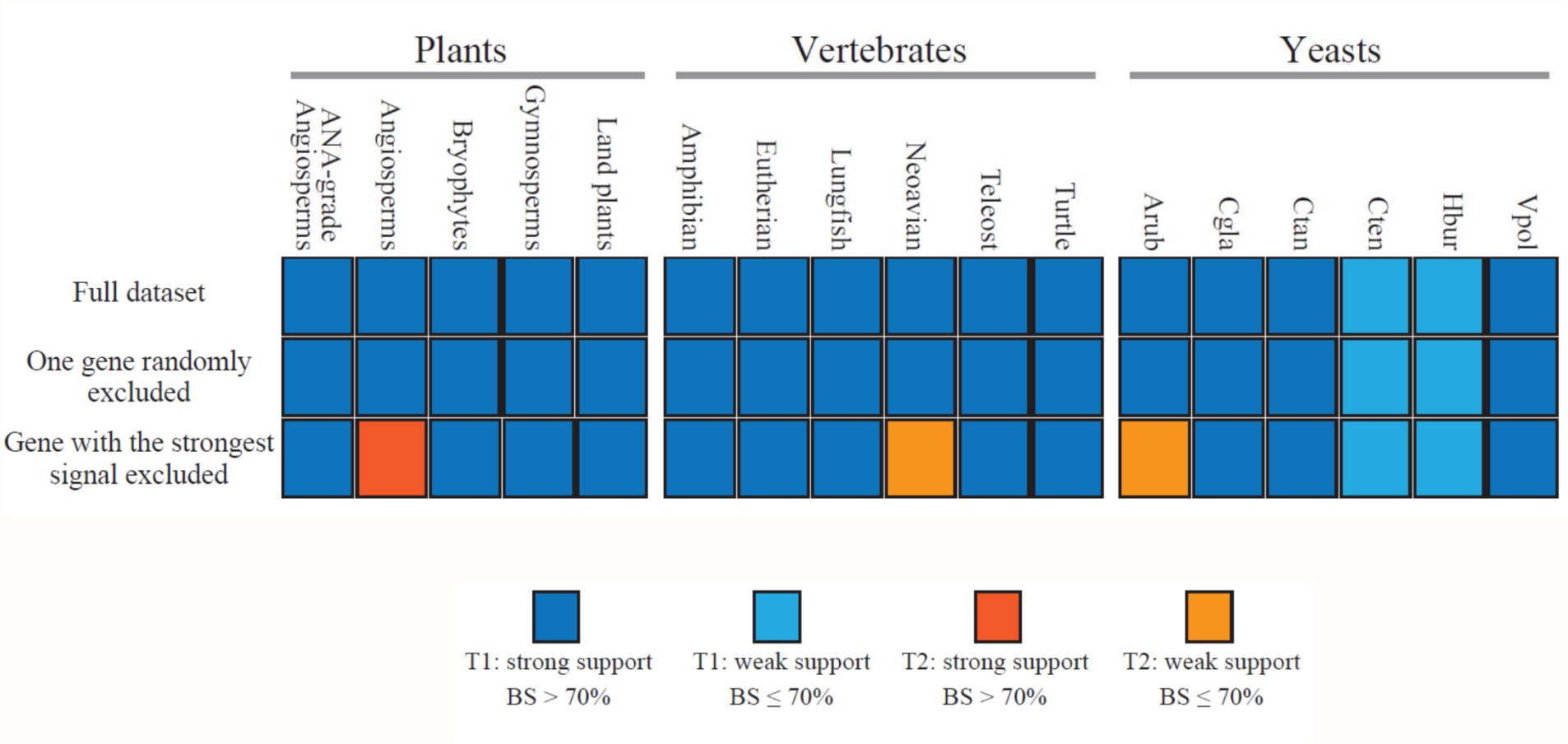
...But in Others It Stems from One or Two Genes



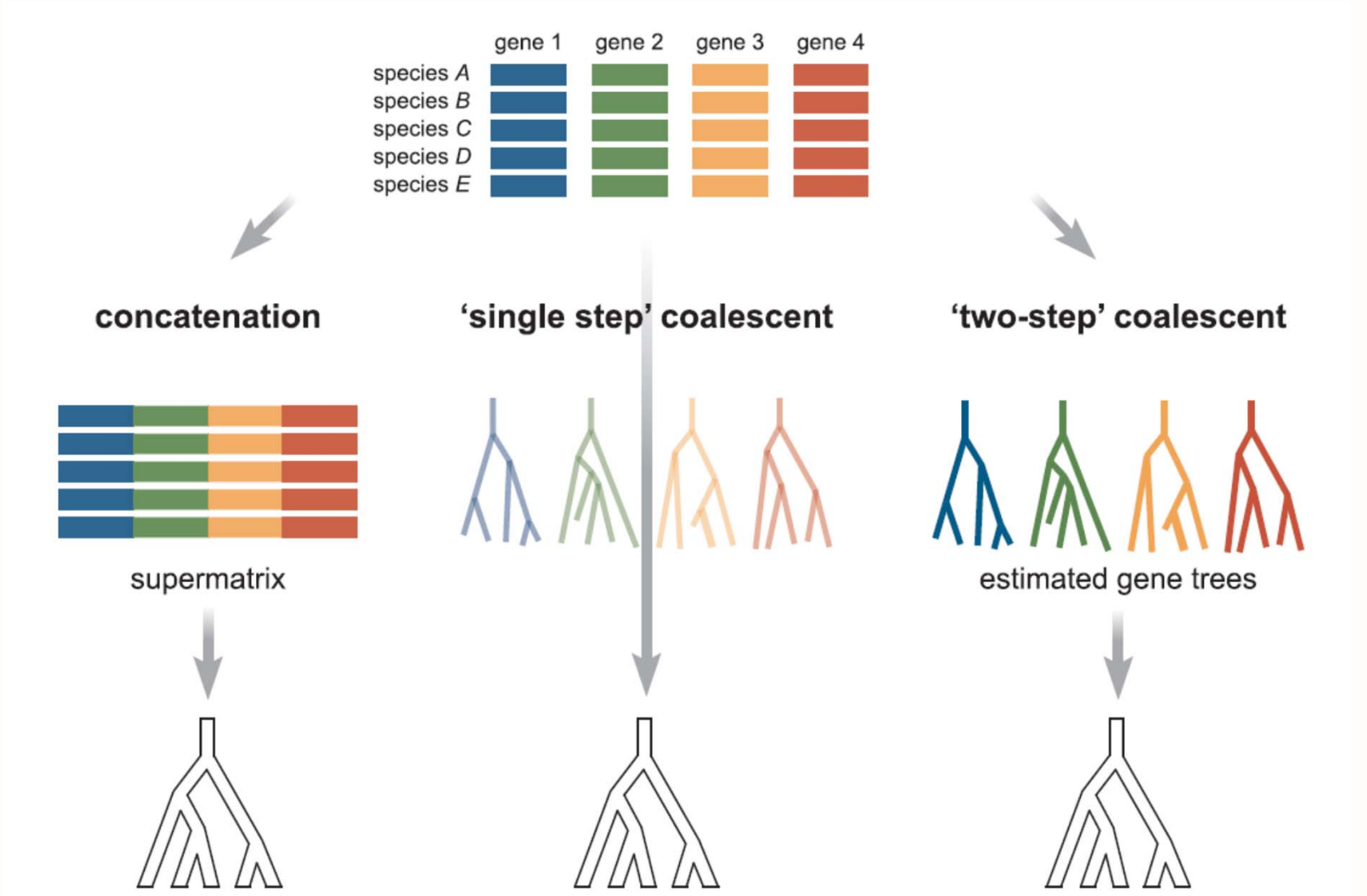
Phylogenetic Signal per Gene for the Two Hypotheses



Removing One Gene Alters the Topology

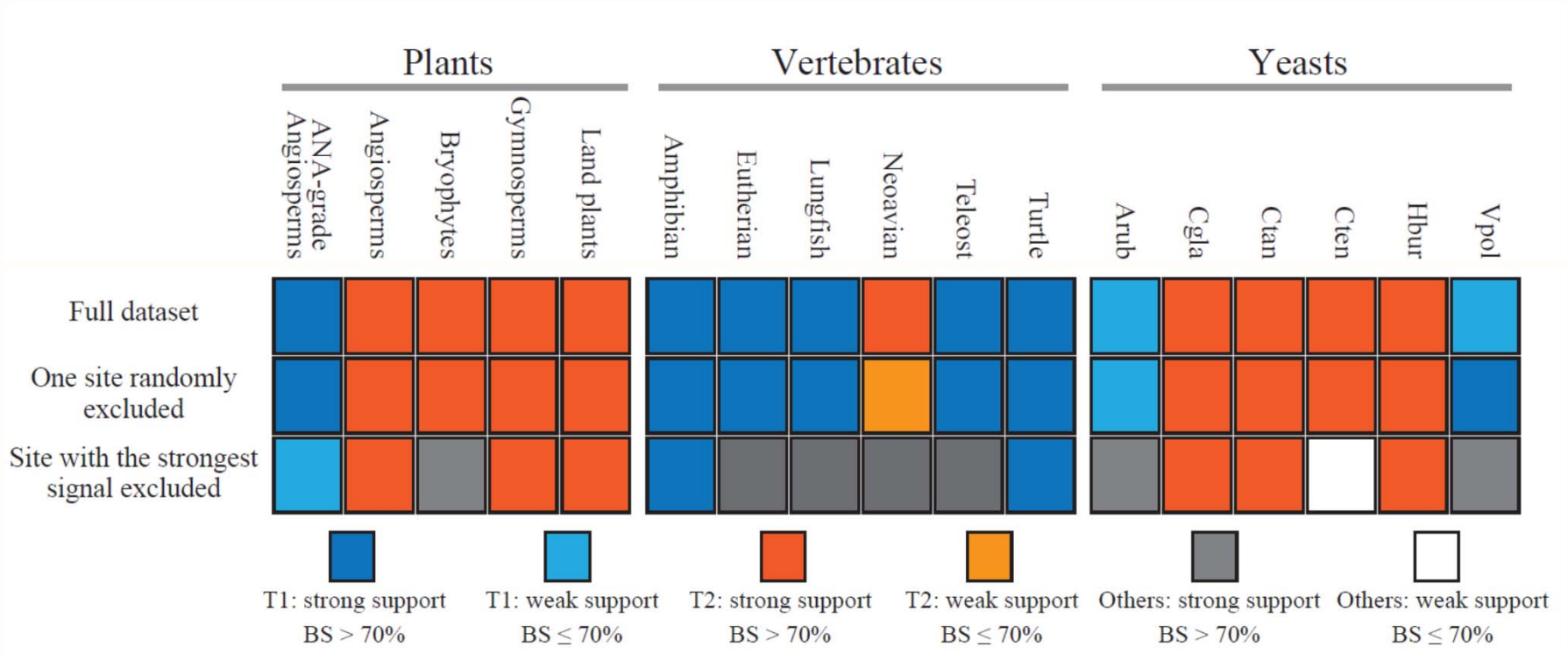


Methods for Phylogenomic Inference

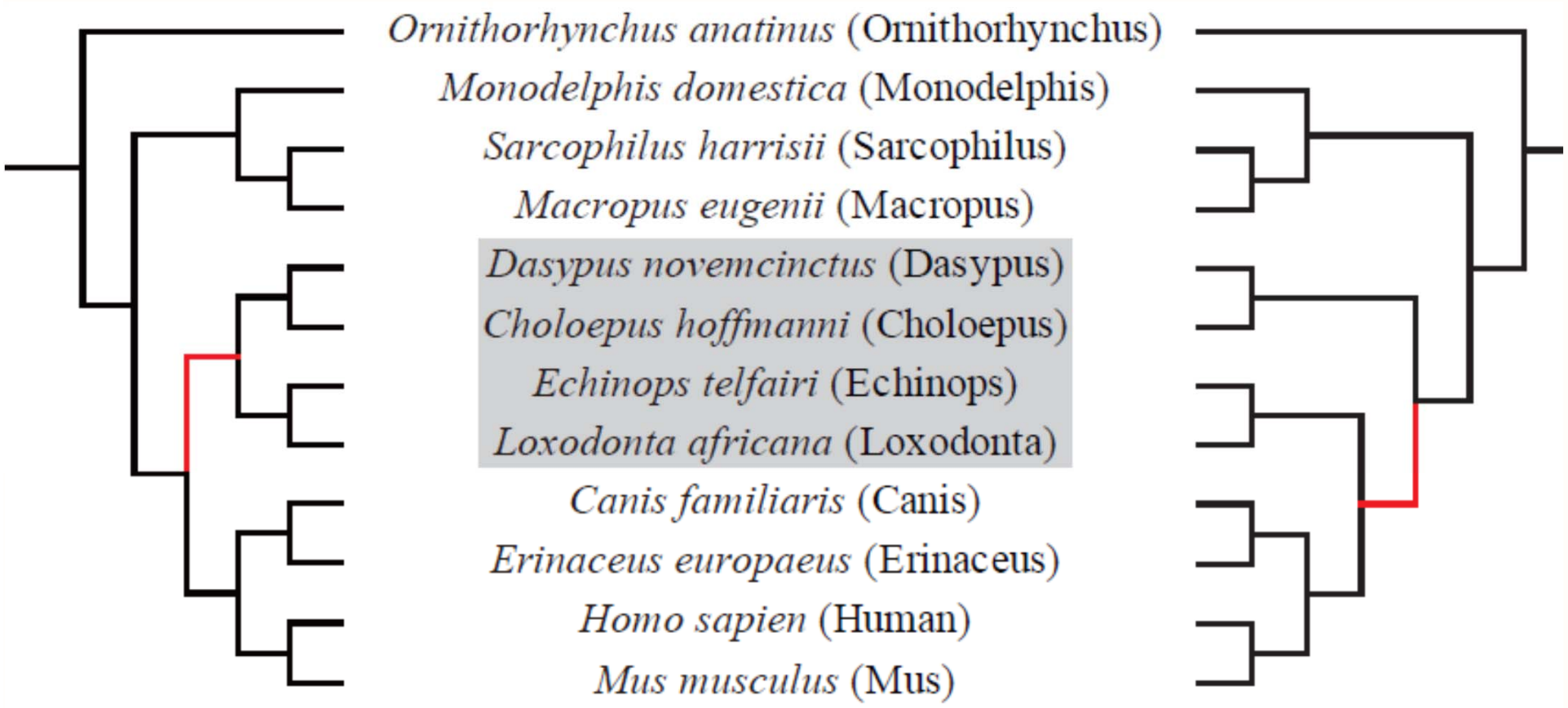


Liu et al. (2015) Ann. N. Y. Acad. Sci.

What Happens if we Remove One Site from Every Gene?



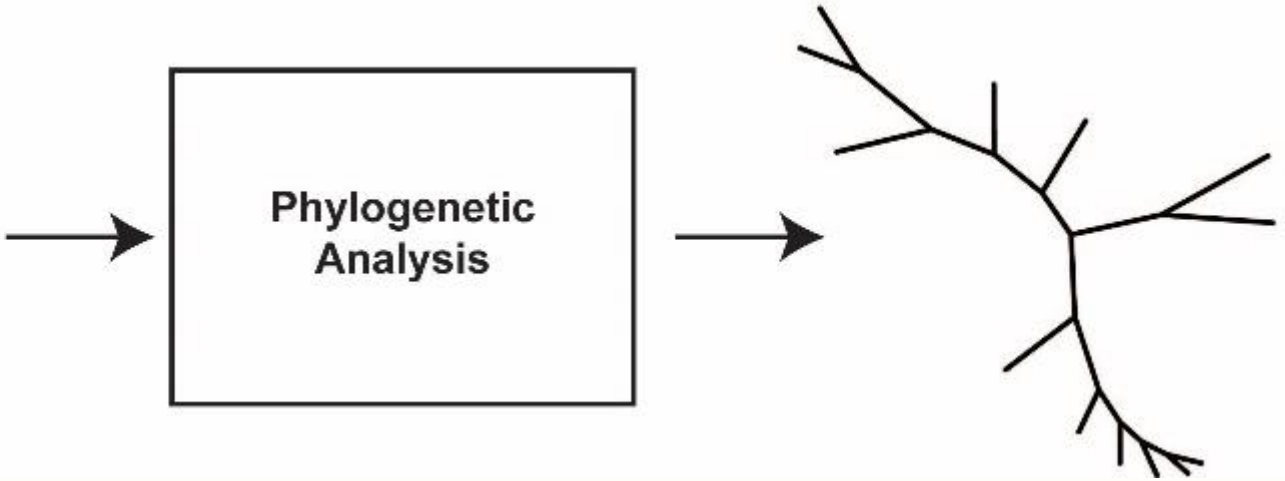
Removing 1 Site Alters the Topology



Shen et al. (2017) Nature Ecol. Evol.

What's Going On?

```
taxon_1    ACCCGATAGACAA
taxon_2    .G.G.........
taxon_3    .......CT....
taxon_4    ....A.......C
taxon_5    T.A..........
taxon_7    .......TT....
taxon_8    ..G....TT....
taxon_9    .......G.....
taxon_10   T............
taxon_11   T............
taxon_12   ..GG......T..
taxon_13   ..GG...C..T..
```

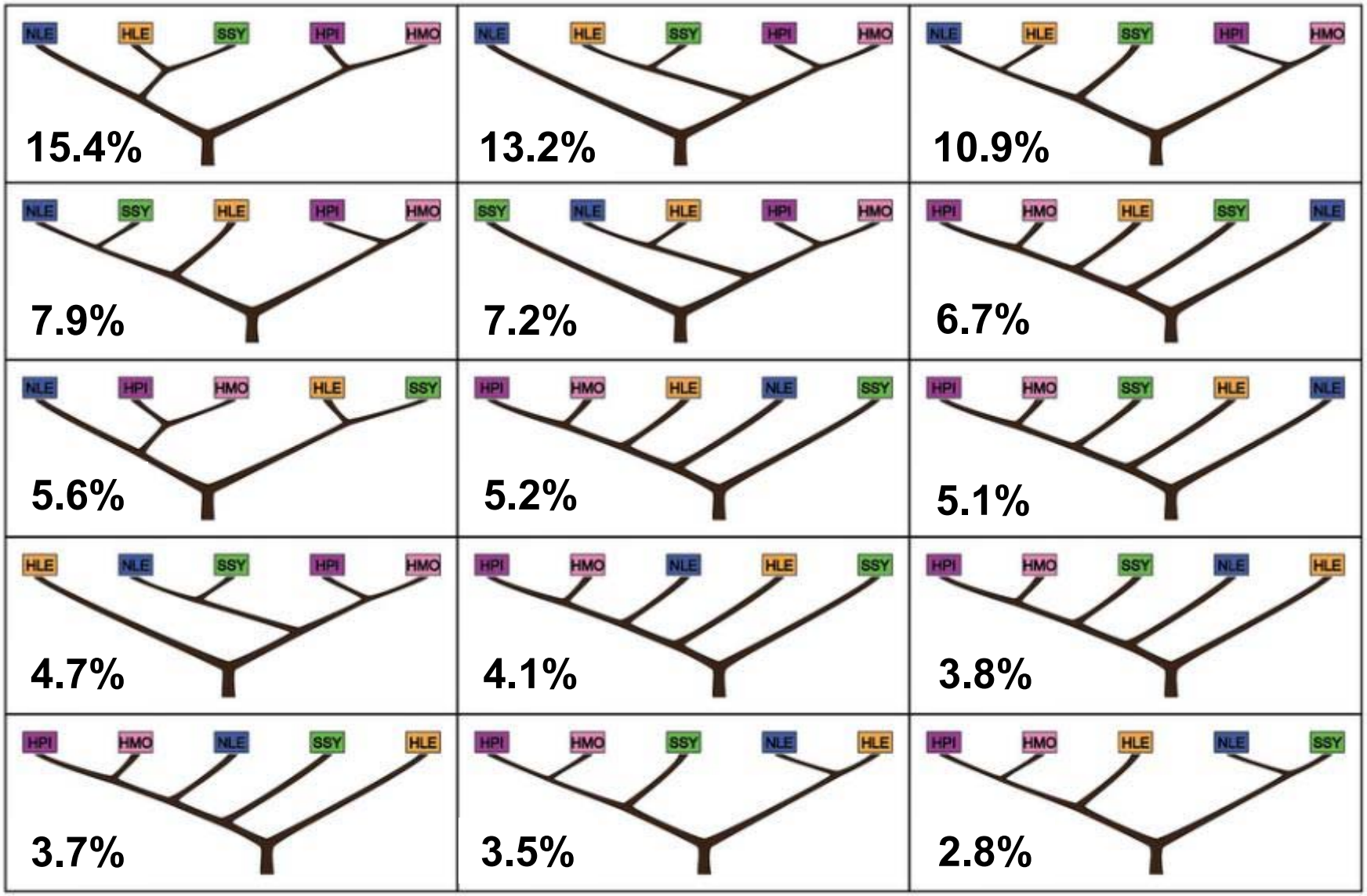


Parts of the tree of life are more likely to resemble bushes rather than trees – why do we expect that every ancestral branch will give rise to two, and only two, descendant branches?

The Phylogeny of Primate Genera

Nomascus leucogenys	*Hoolock leuconedys*	*Symphalangus syndactylus*	*Hylobates pileatus*	*Hylobates moloch*
				
NLE	HLE	SSY	HPI	HMO

Which is the Phylogeny of These 4 Genera?



Carbone et al. (2014) Nature

Genomfart?

- Parts of the tree of life are more likely to resemble a bush rather than a tree – do we expect that we can confidently infer every branch and twig?
- Bootstrap-based measures not useful in large data sets – methods evaluating conflict are preferable
- Methods evaluating conflict among data subsets (e.g., among genes or sites) are preferable
- Explicitly identify internodes that, despite the use of genome-scale data sets, robust study designs and powerful algorithms, are poorly supported

Lecture Outline

- Introduction to evolutionary genomics

- Phylogenomics

---------- Coffee Break ----------

- Phylogenomics

- Using genomes to understand lineage diversification

The Making of Biodiversity Across the Yeast Subphylum

- Sequence the genomes of all ~1,000+ known yeast species in the Saccharomycotina subphylum
- Construct their definitive phylogeny
- Revise their taxonomy
- Examine the impact of metabolism on yeast diversification

The Making of Biodiversity across the Yeast Subphylum

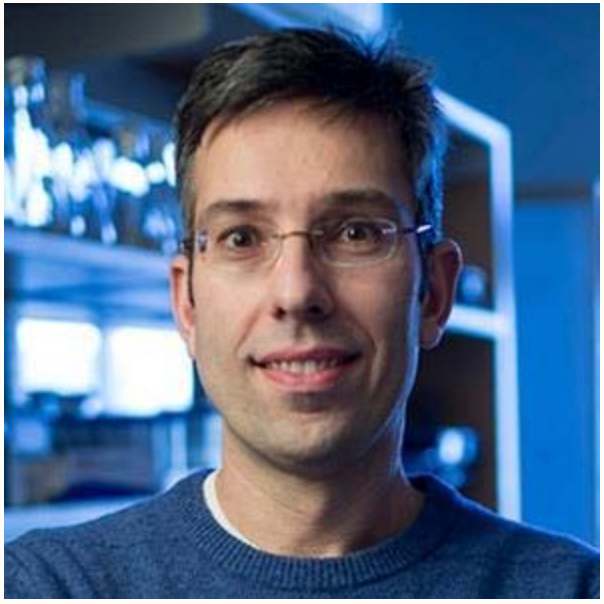
THE UNIVERSITY of WISCONSIN MADISON

USDA

VANDERBILT UNIVERSITY

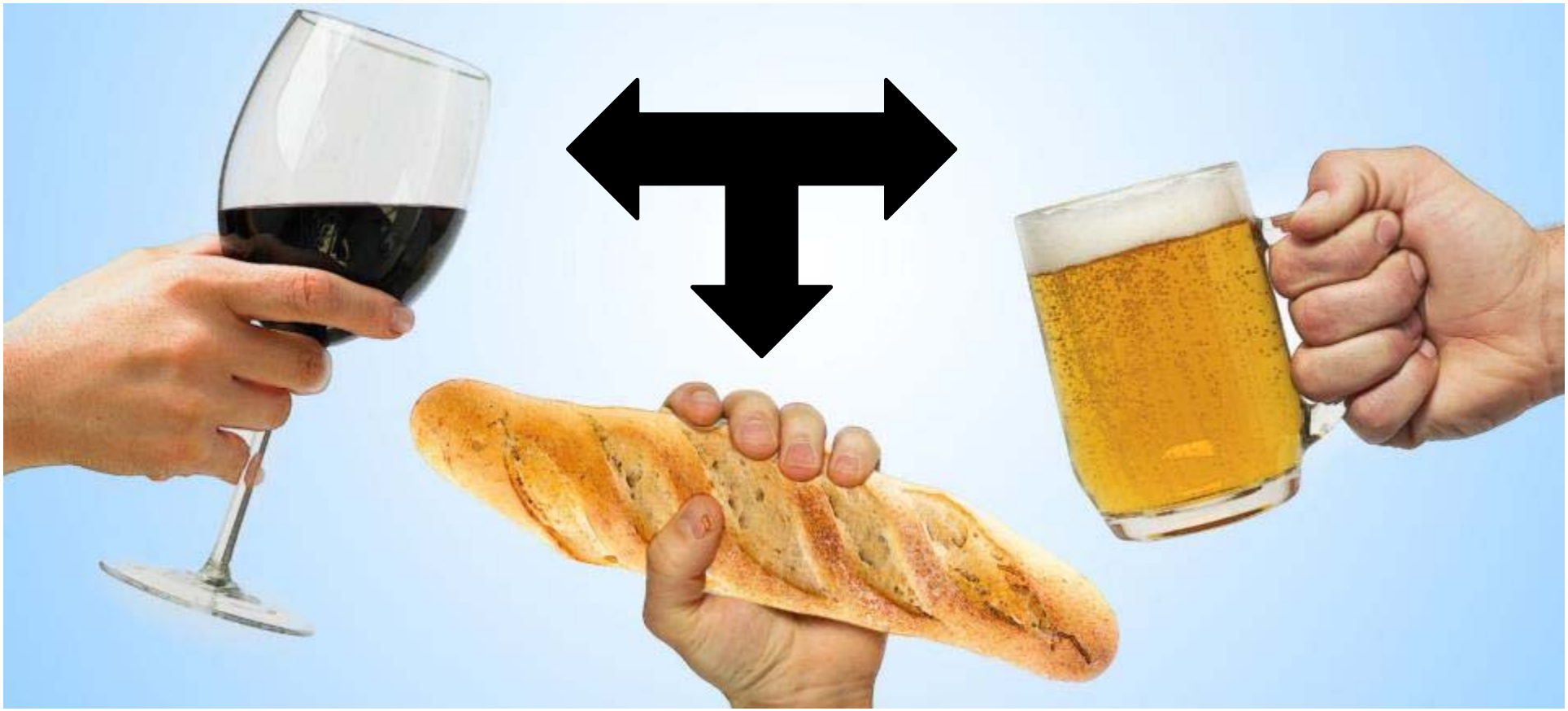
Hittinger lab

Kurtzman lab



Rokas lab

S. cerevisiae, *Cornerstone of Wine, Baking, and Brewing Industries*



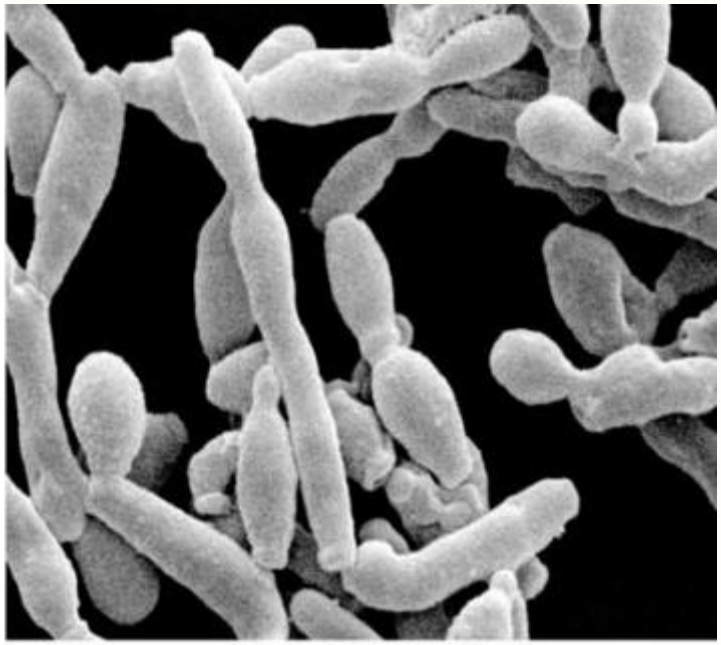
Several Other Genera Critical to the Food Industry

Aside from bread, beer, and wine, yeasts are critical for the production of kefir, soy sauce, sourdough, lambic beers, kimchi, dietary supplements, probiotics, and some cheeses

Genera involved: *Saccharomyces*, *Kluyveromyces*, *Zygosaccharomyces*, *Candida*, *Kazachstania*, *Pichia*, and *Dekkera* (*Brettanomyces*)

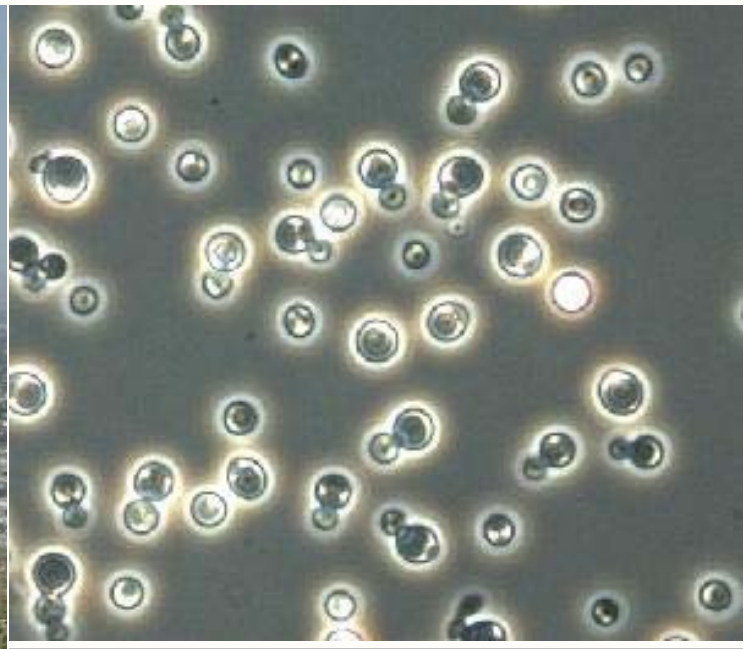
The Metabolisms of the 1,000+ Species Vary Widely

Xylose fermenters
(*Scheffersomyces (Pichia) stipitis*)

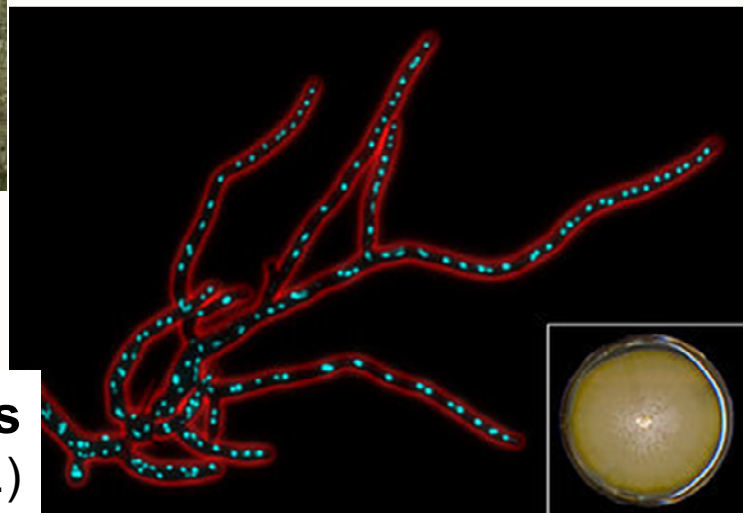


Cactophilic yeasts

Oil producers
(*Lipomyces*)



Animal pathogens
(*Candida*)



Plant pathogens
(*Eremothecium* sp.)

developing pipelines for high-throughput data analyses

Developing Pipelines for Genome Assembly

iWGS: *in silico* Whole Genome Sequencer & Analyzer

INPUT: Experimental Design + Reference Genome (Optional)

↓

(Optional) Step 1: Simulation of Illumina / PacBio Data

↓

Step 2: Quality Control (Quality / Adaptor Trimming; Error Correction)

↓

Step 3: Assembly (Illumina-only (10), PacBio-only (3), Hybrid (4))

↓

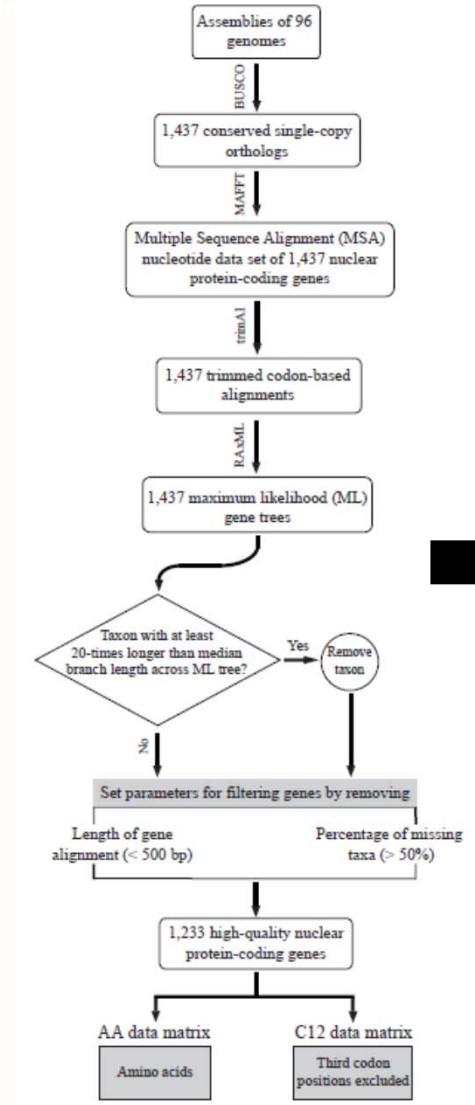
Step 4: QUAST Evaluation (Standalone or vs the Reference Genome**)**

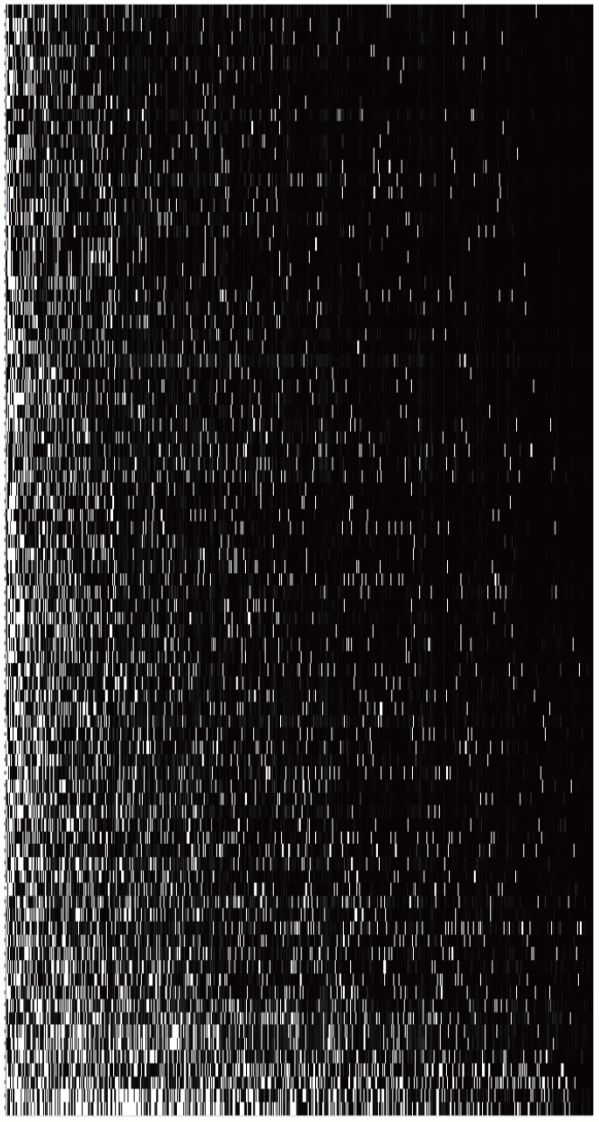
↓

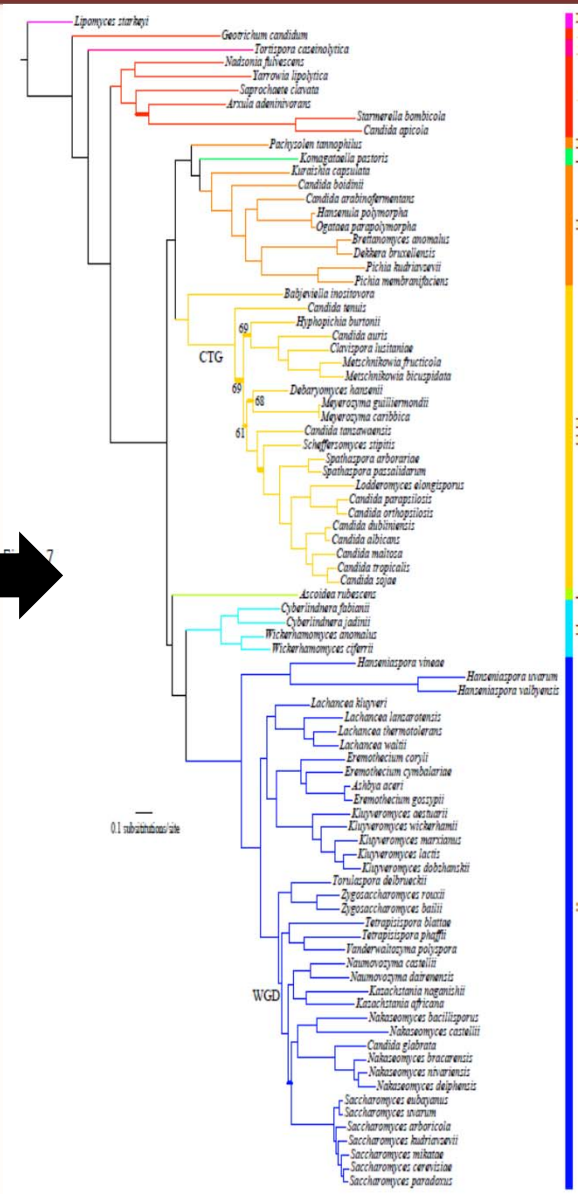
OUTPUT: Evaluation Report, Ranking of Experimental Designs Tested

Zhou et al. (2016) G3

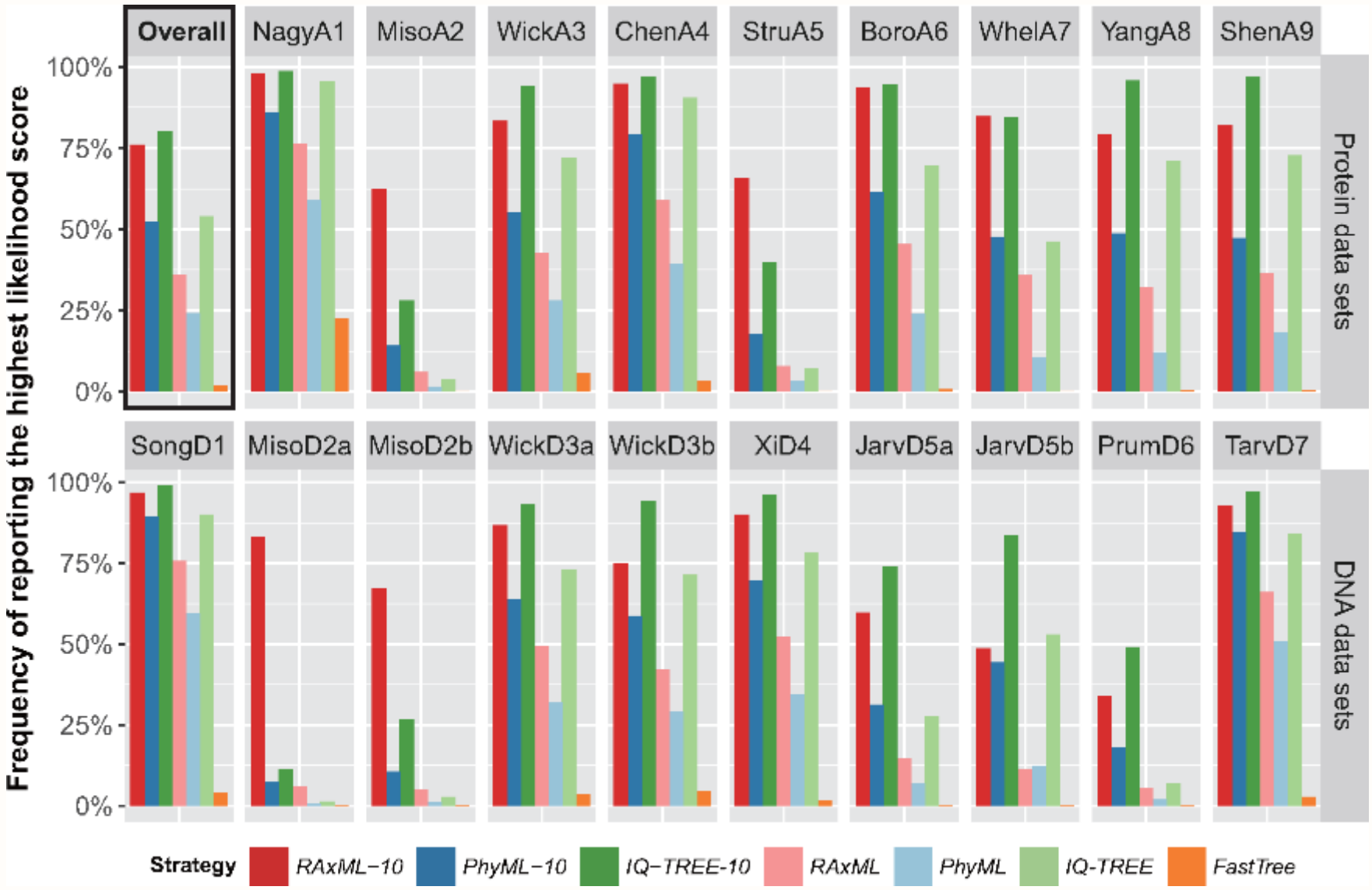
Developing Pipelines for Phylogenomic Inference







Assessing Speed and Accuracy of Phylogenomic Software



Acknowledgments

Hittinger lab – UW

Jacek Kominek

Dana Opulente

Amanda Hulfachor

David Peris

Kurtzman lab – USDA

Jeremy Devirgilio

Rokas lab – Vandy

Xing-Xing Shen

Jen Wisecaver

Xiaofan Zhou

Jacob Steenwyk

NSF

http://y100plus.org








http://www.rokaslab.org/