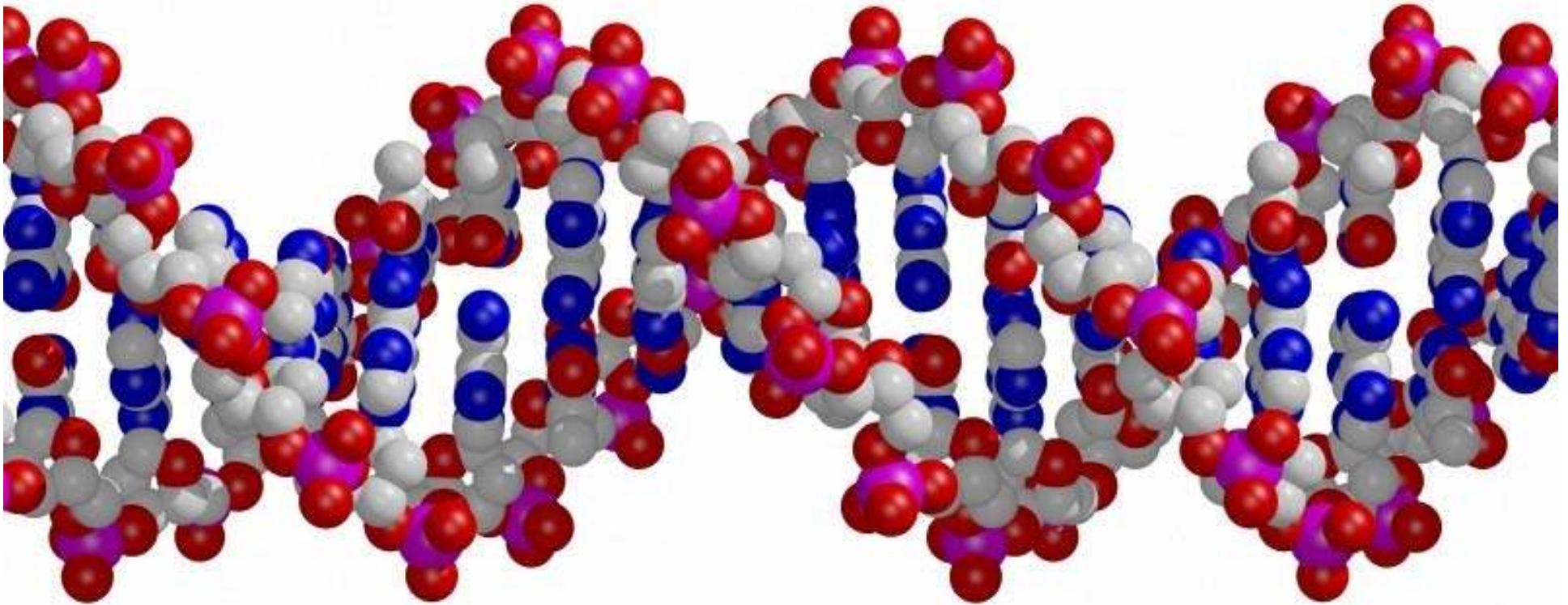


Evolutionary Genomics



Antonis Rokas
Department of Biological Sciences
Vanderbilt University



<http://as.vanderbilt.edu/rokaslab>

Lecture Outline

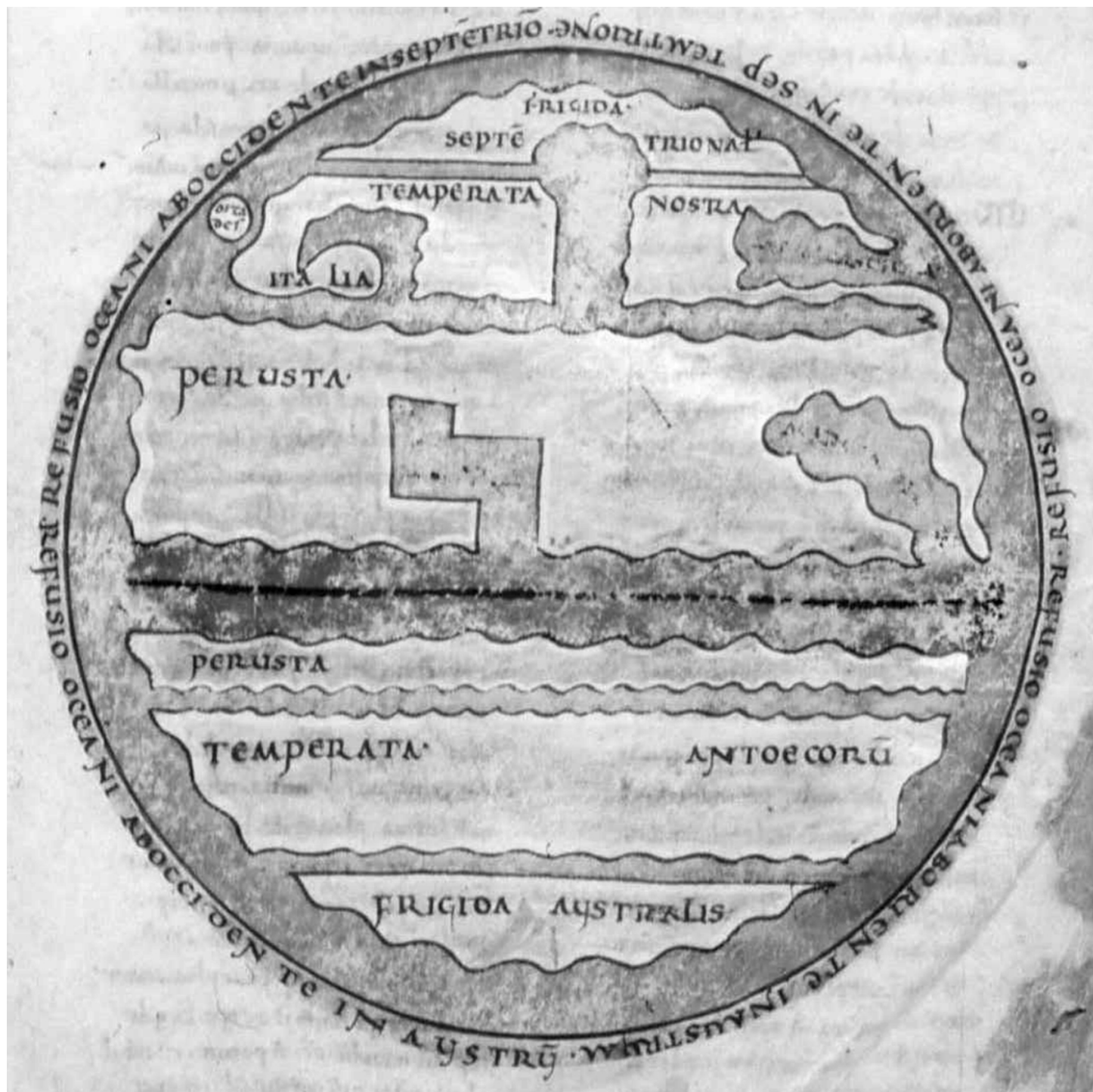
❖ **Introduction to Genome Evolution**

❖ **Comparative Functional Genomics**

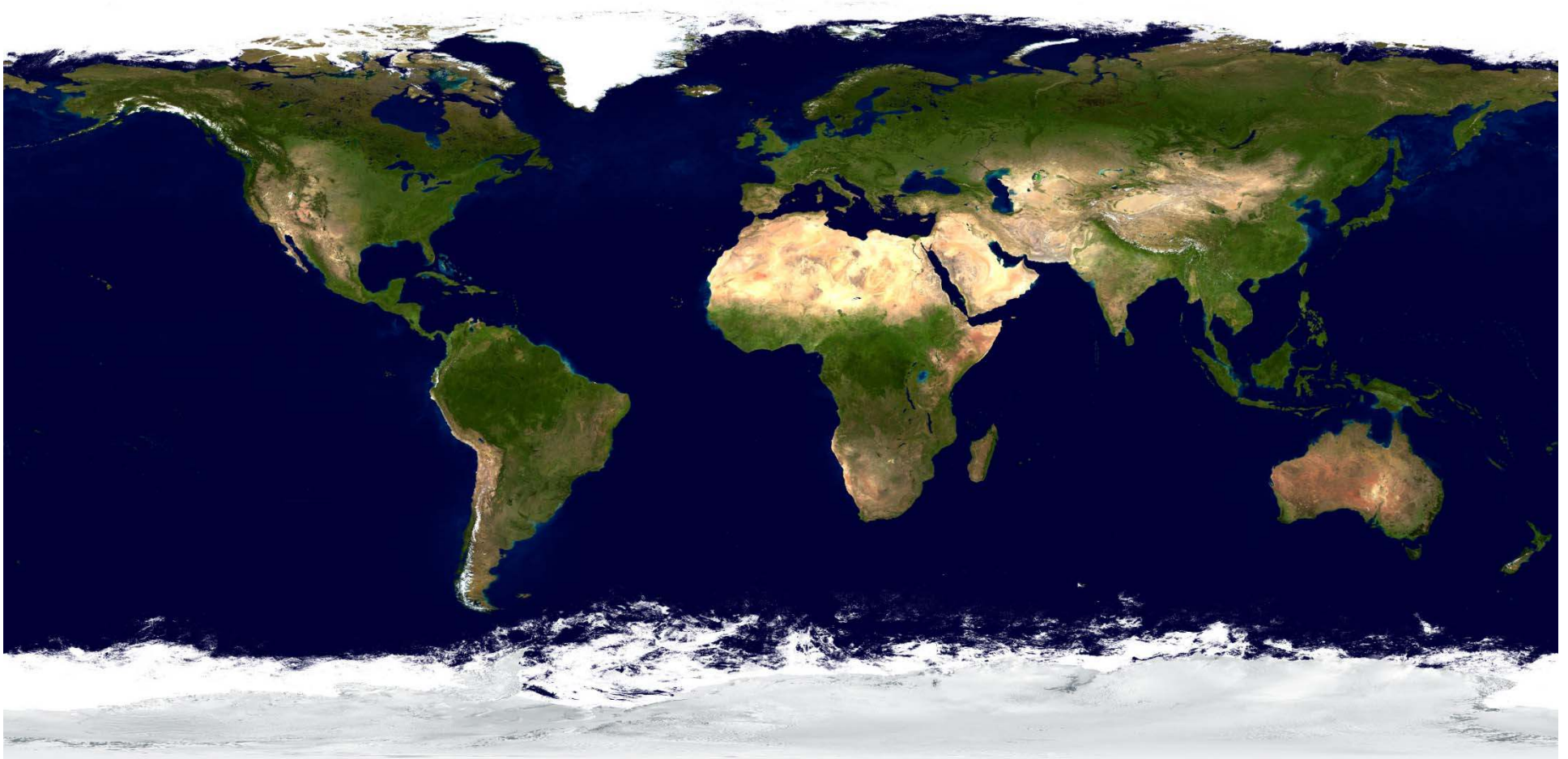
----- **Coffee Break** -----

❖ **Population Genomics**

❖ **Phylogenomics**



The World's Map



http://www.oera.net/How2/PlanetExs/EarthMap_2500x1250.jpg

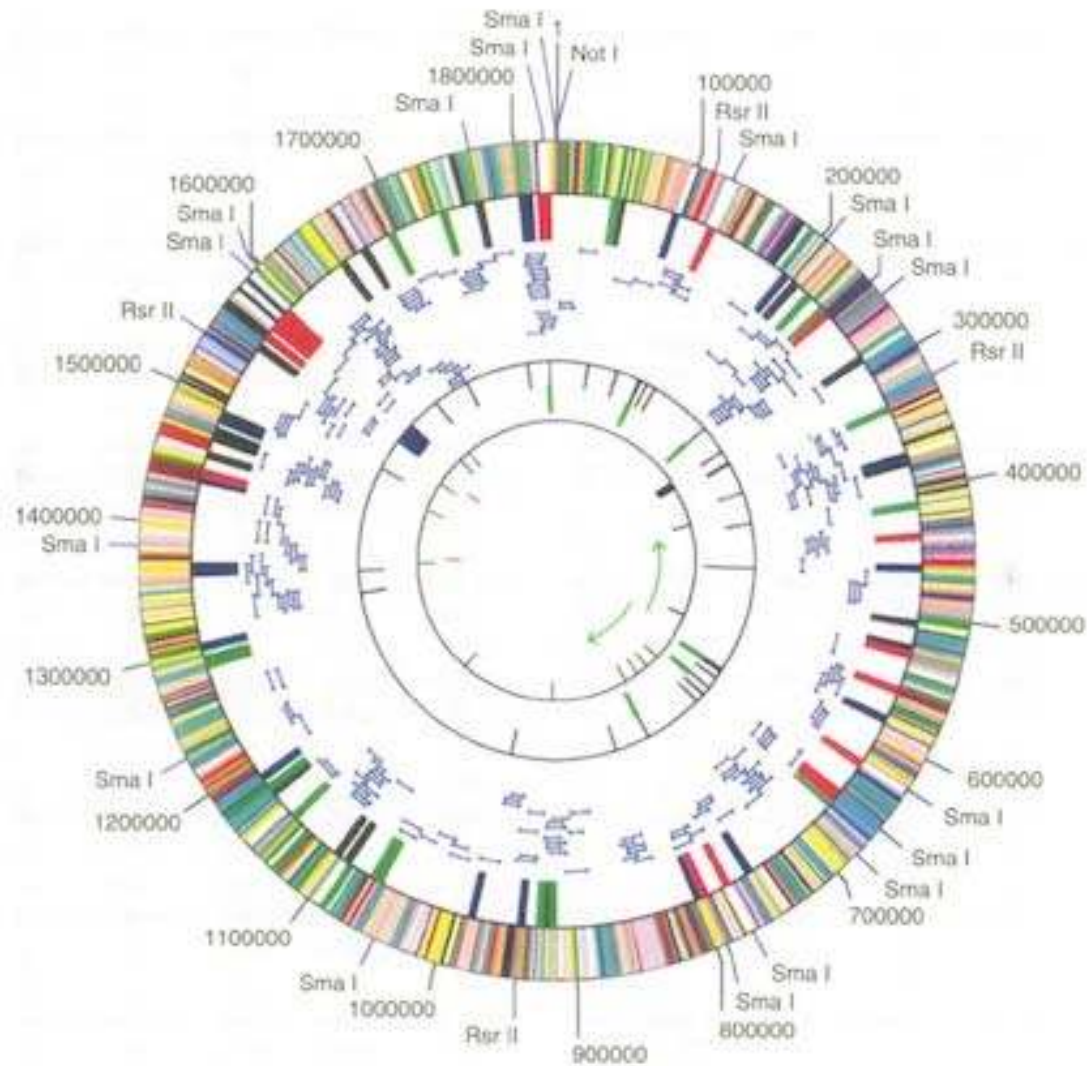


The Periodic Table

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period																		
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	* 71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	** 103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Uuq	115 Uup	116 Uuh	117 Uus	118 Uuo
*Lanthanoids			* 57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb		
**Actinoids			** 89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No		



1995 – 1996: The First Sequenced Genomes

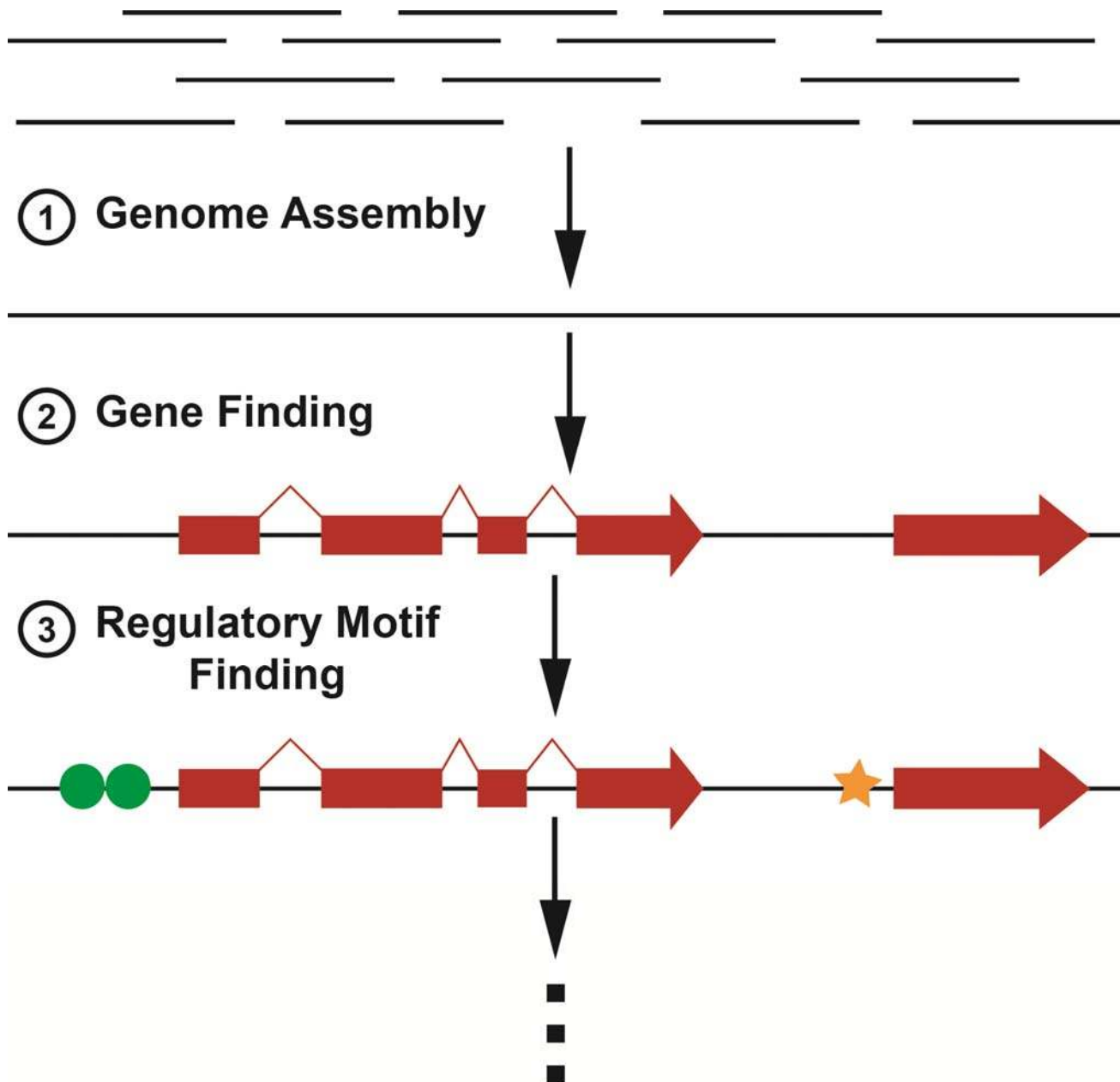


What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCACTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTGAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCCTCCAGCTCCCTAGCACTGGCAGGGTTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCCTCACAGGTAAGTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGTTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACAGCCGCATCCCCTACAAAATCTACCCCAAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTGTGAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAATAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAACCCCTTGGGCCCAGCGGAACAAGGAAA
AGGGGATACAGTGCGGATTTCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT



Understanding the Genome Requires Tools



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTTCGGGCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGTTCTCTCACTACTCAGTTCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGCGCAGGGAATAGGTCAGCCATGGTGGAAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGCGGAGATAAACTTACCGCAGTGACACAGATCAAAAGATGTCTTTGACATAAGCAAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTAAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTTATCTGCTCTCCAAGTCT
GCGATATTGCTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGTAAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTTCGTAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCTACAAAATCTACCCCAAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTGAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCAGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCTACATTCATATGGGCCAGCGATTGGAACCCTTCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

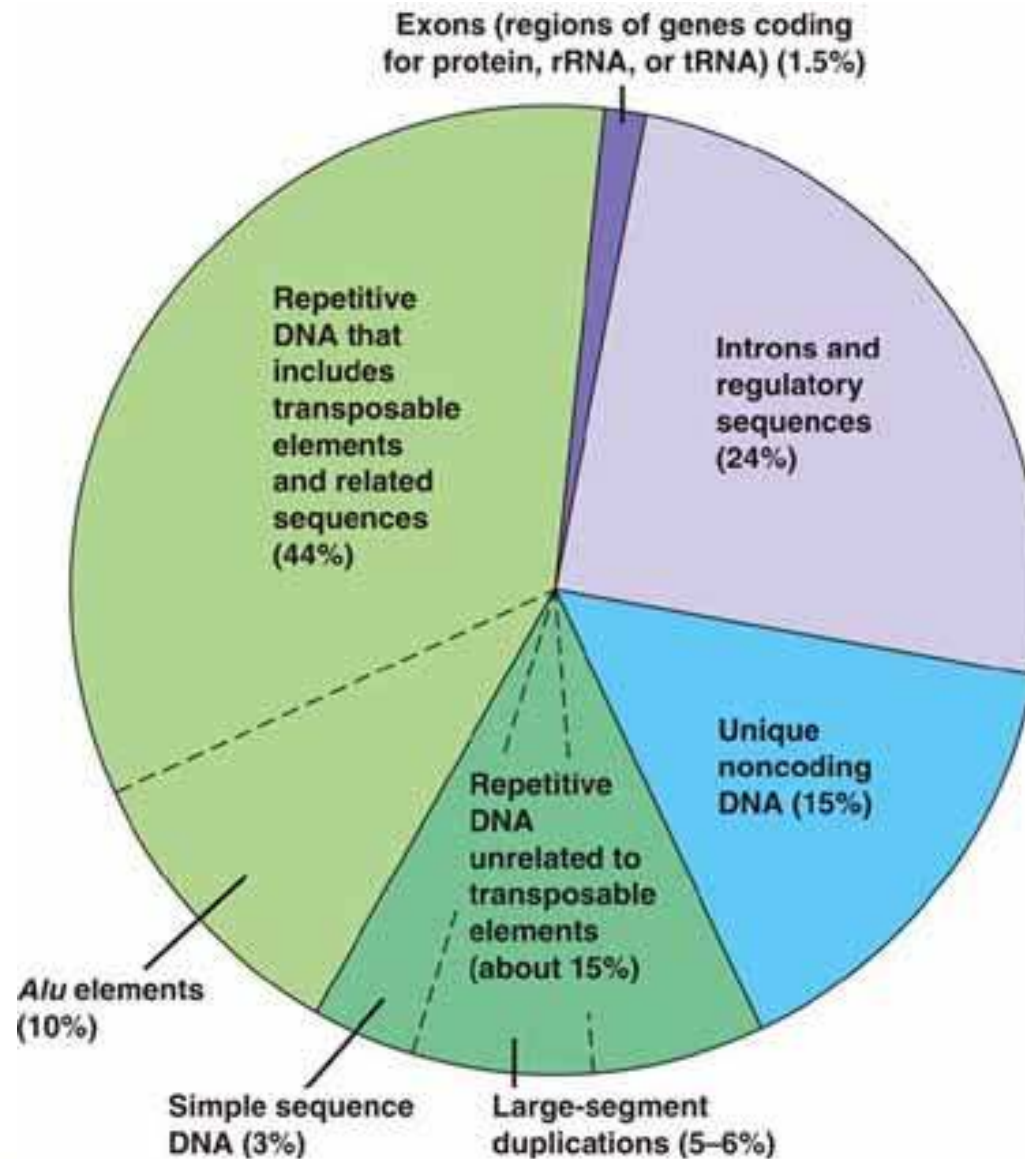
Non-coding DNA

Regulatory Motifs

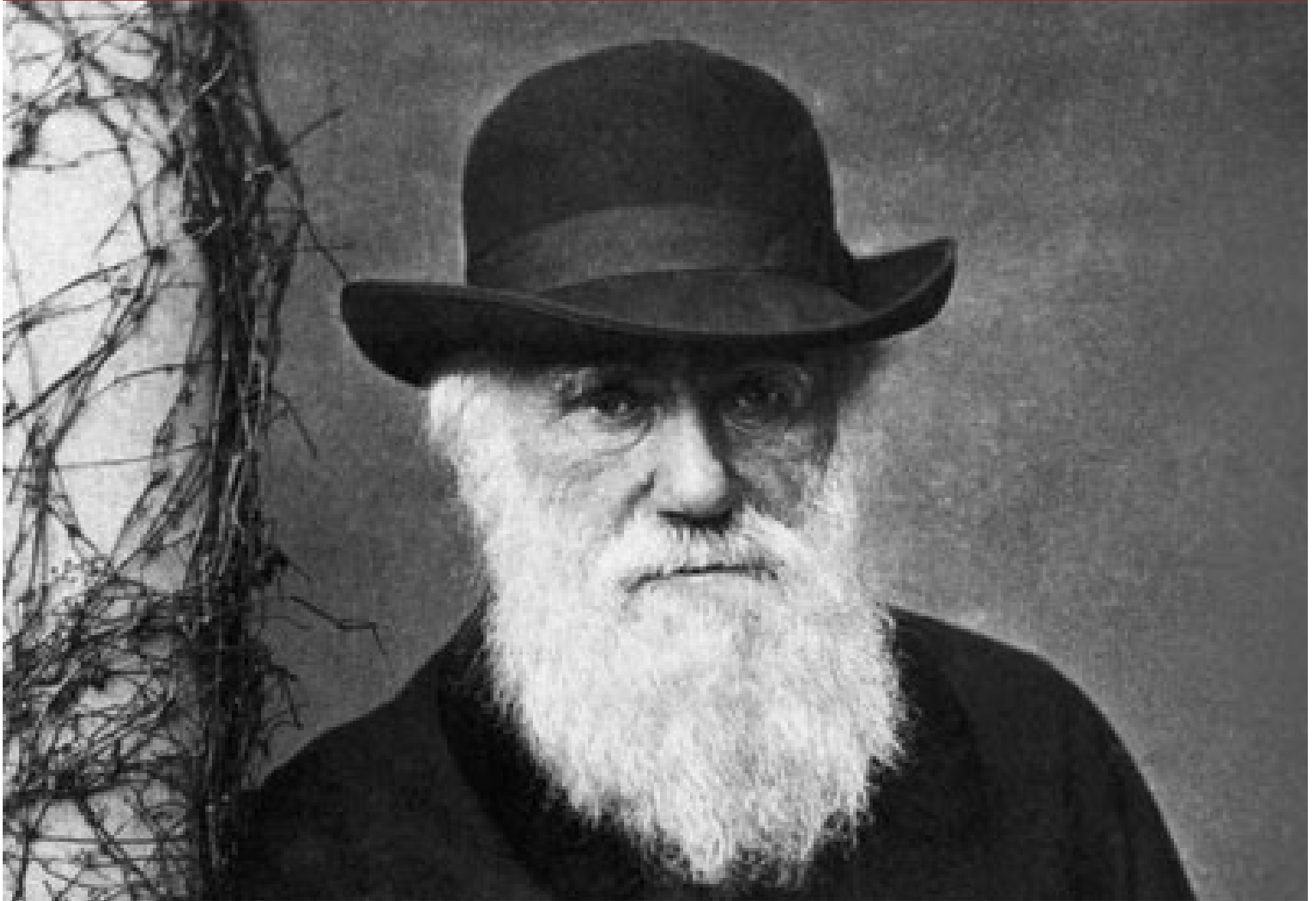
Exons

Introns

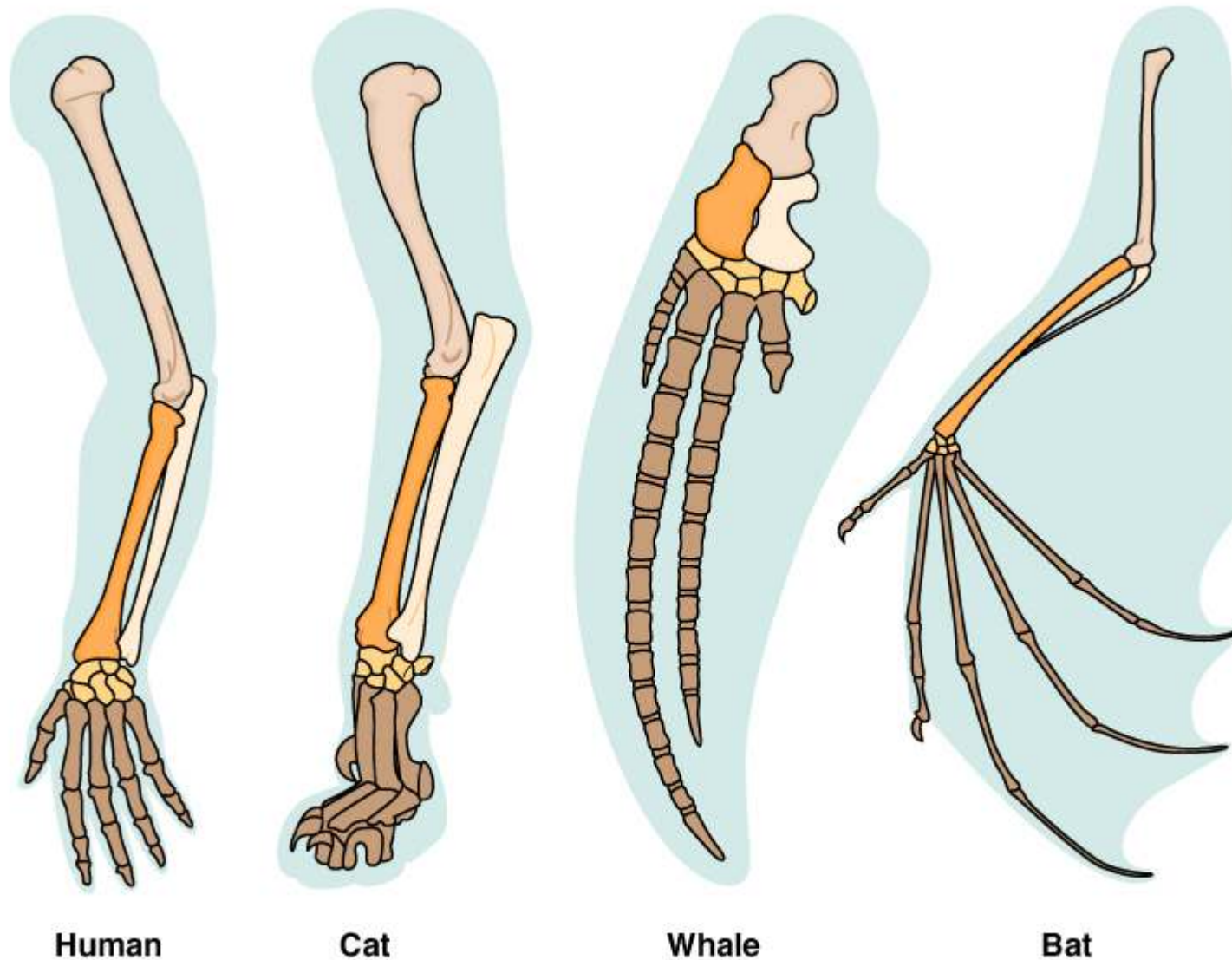
Organization of the Human Genome



Understanding the Genome Requires a Theory



Similarity in Anatomy Suggests Common Origins



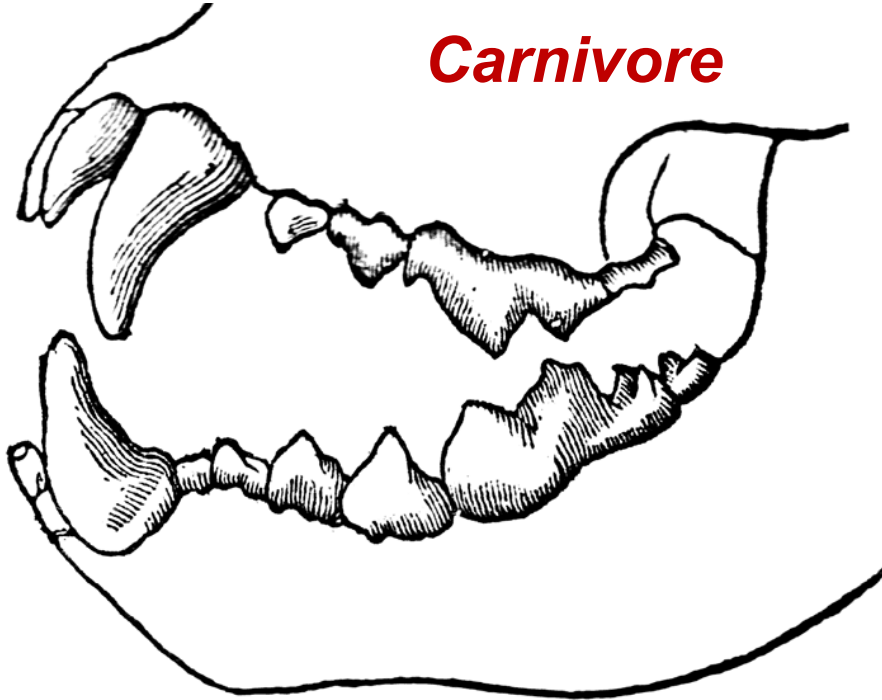
©1999 Addison Wesley Longman, Inc.



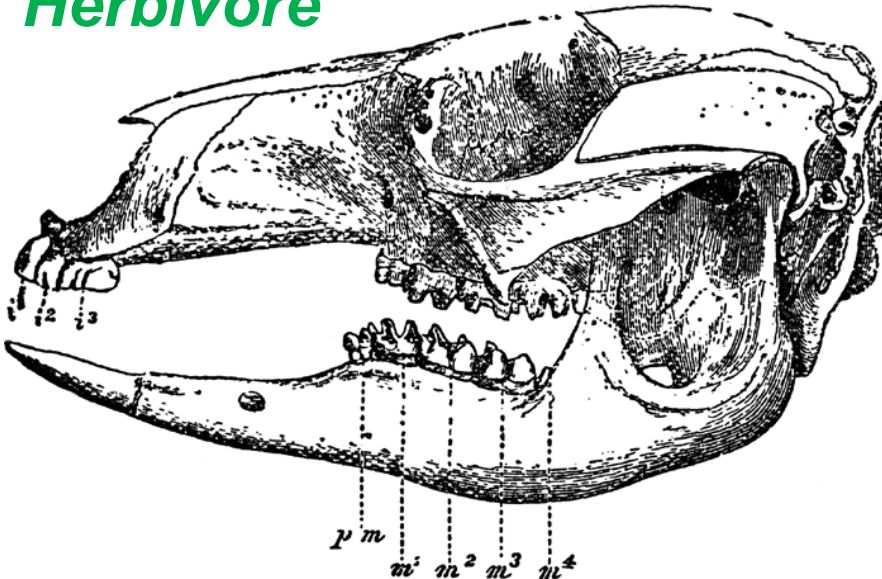
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg

Differences in Anatomy Suggest Adaptations

Carnivore



Herbivore

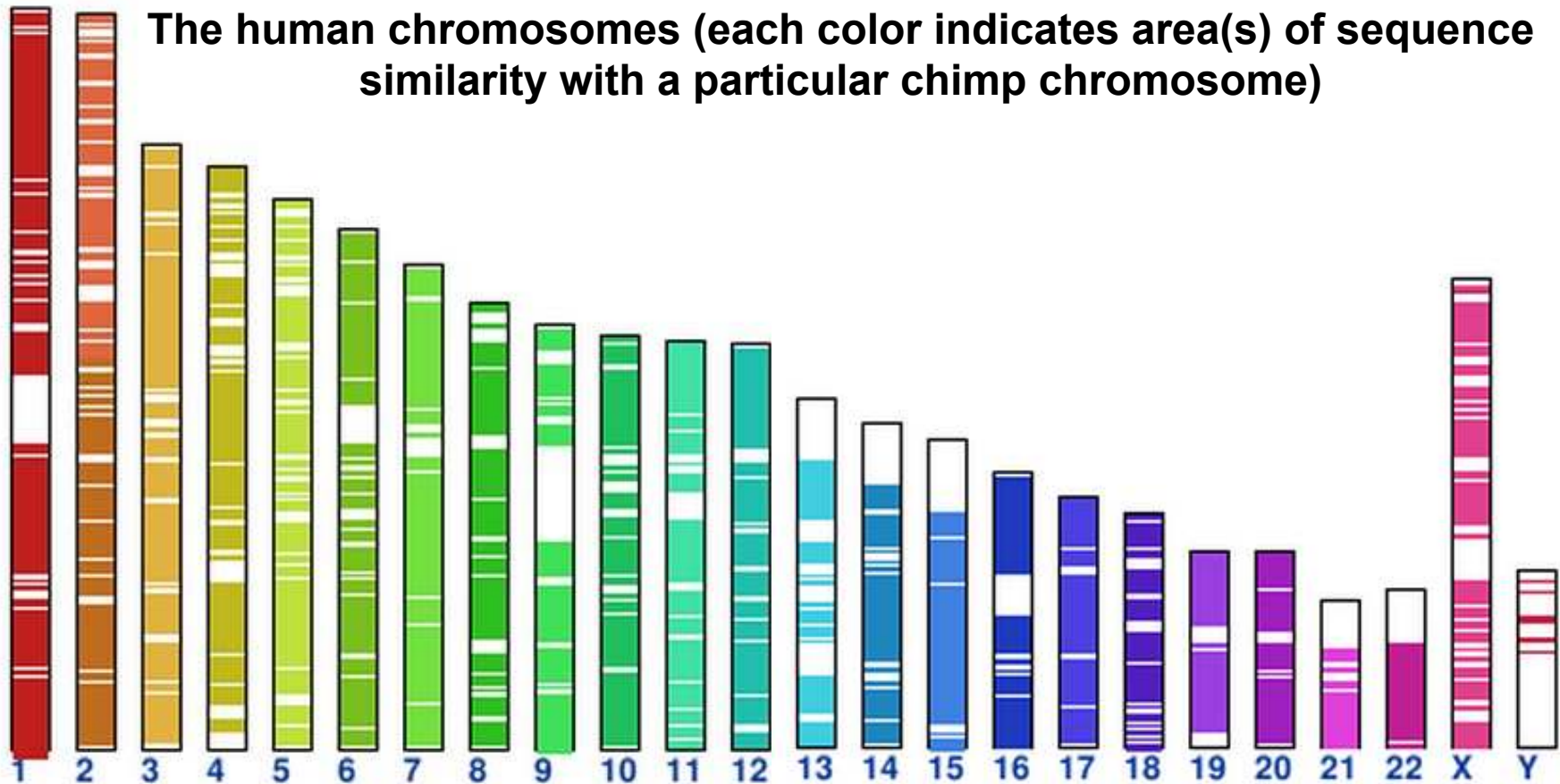


Incisivosaurus

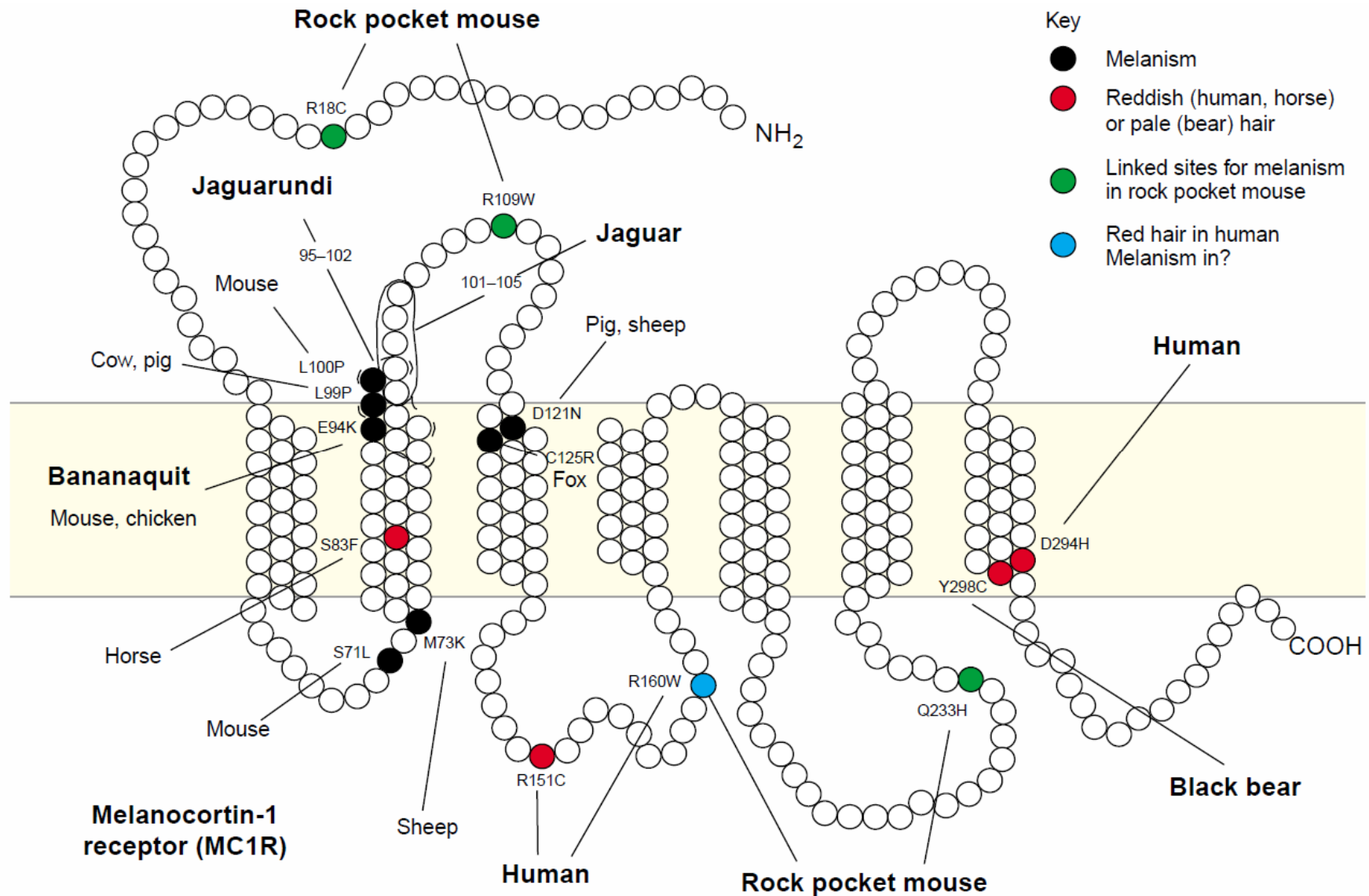


Carnivore or Herbivore?

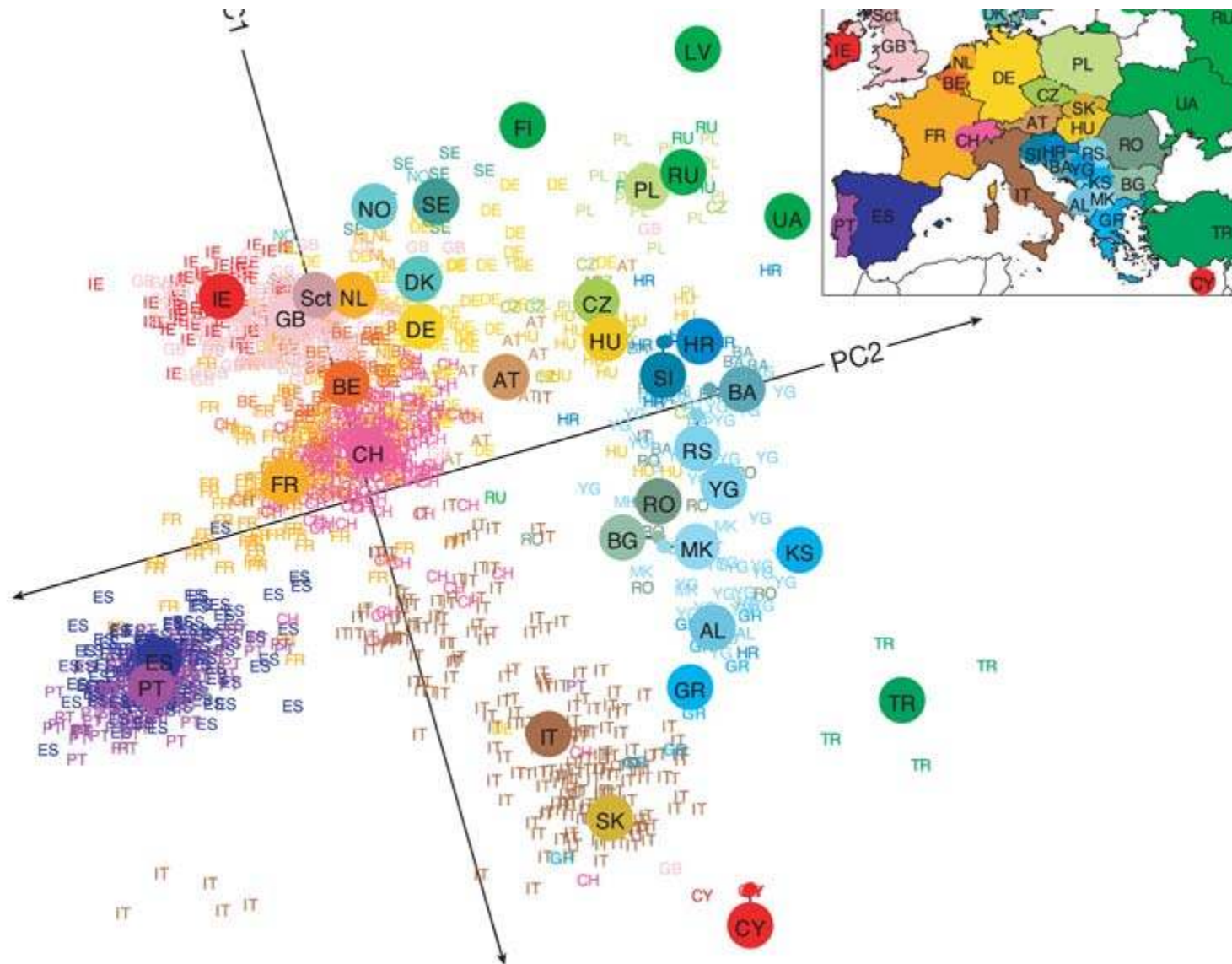
Similarity in the DNA Record Suggests Common Origins



Differences in the 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	
chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	LARGE
Total SNPs			16.74	9,166	2,898	480	41	

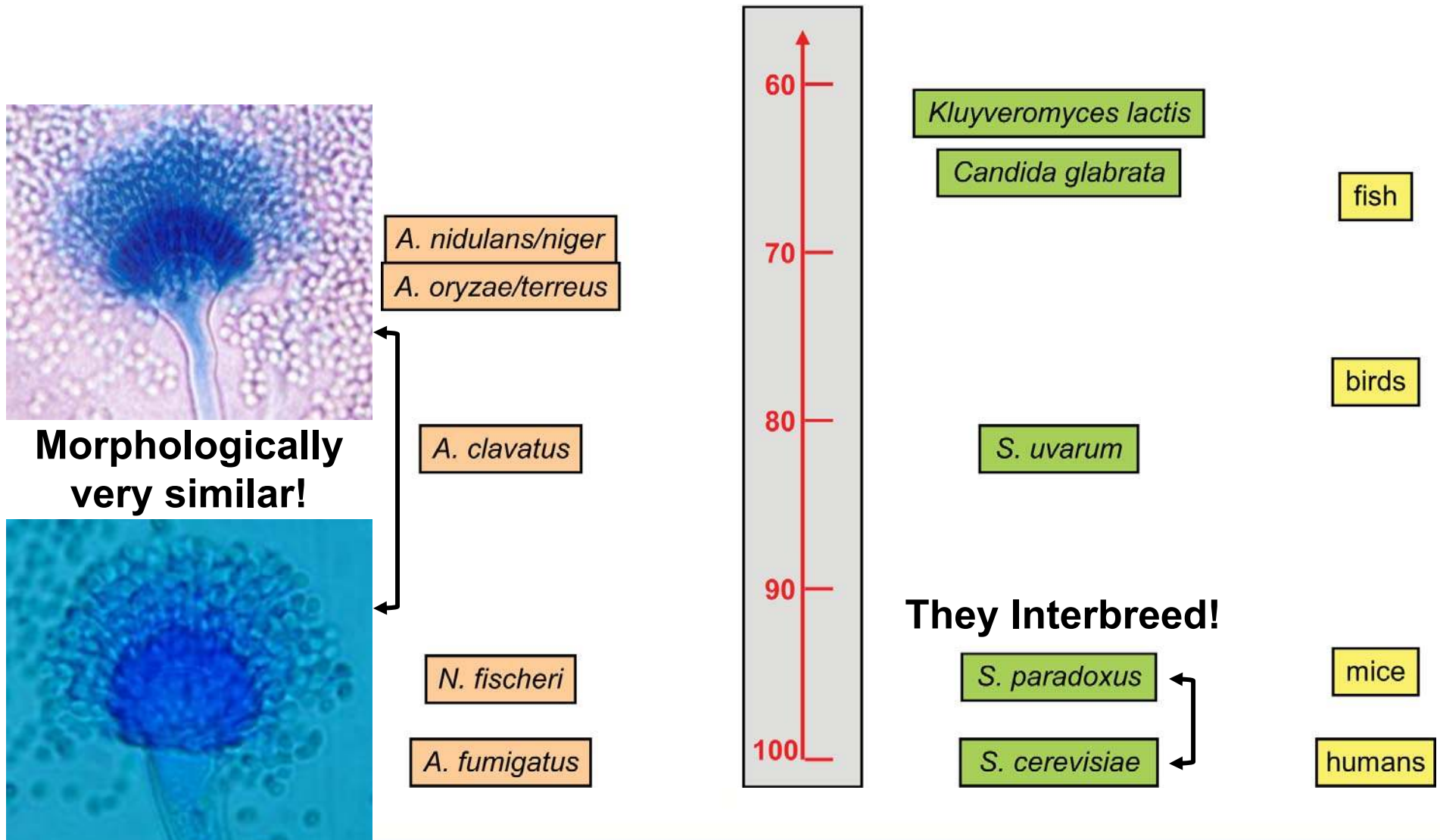
In the European population,
involved in skin pigmentation

In the West African population,
related to Lassa virus infection



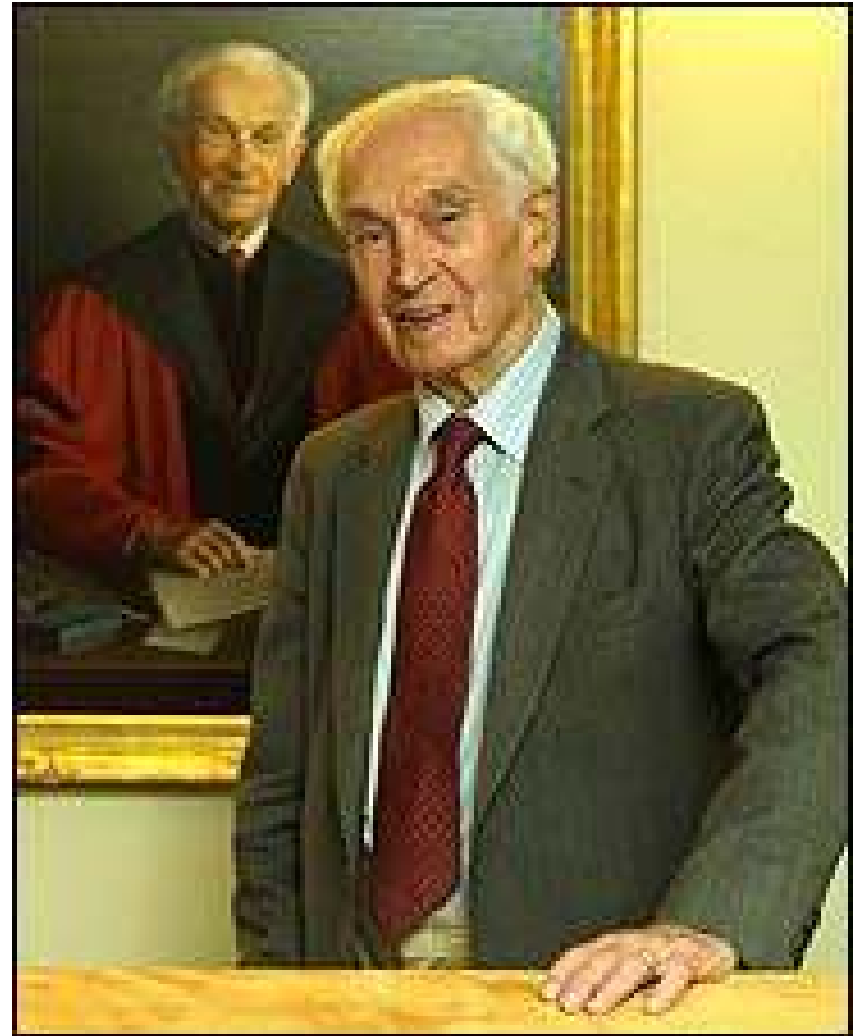
Genomes Provide a Common Yardstick for Comparison

Average proteome sequence similarity



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

Ernst Mayr, 1963



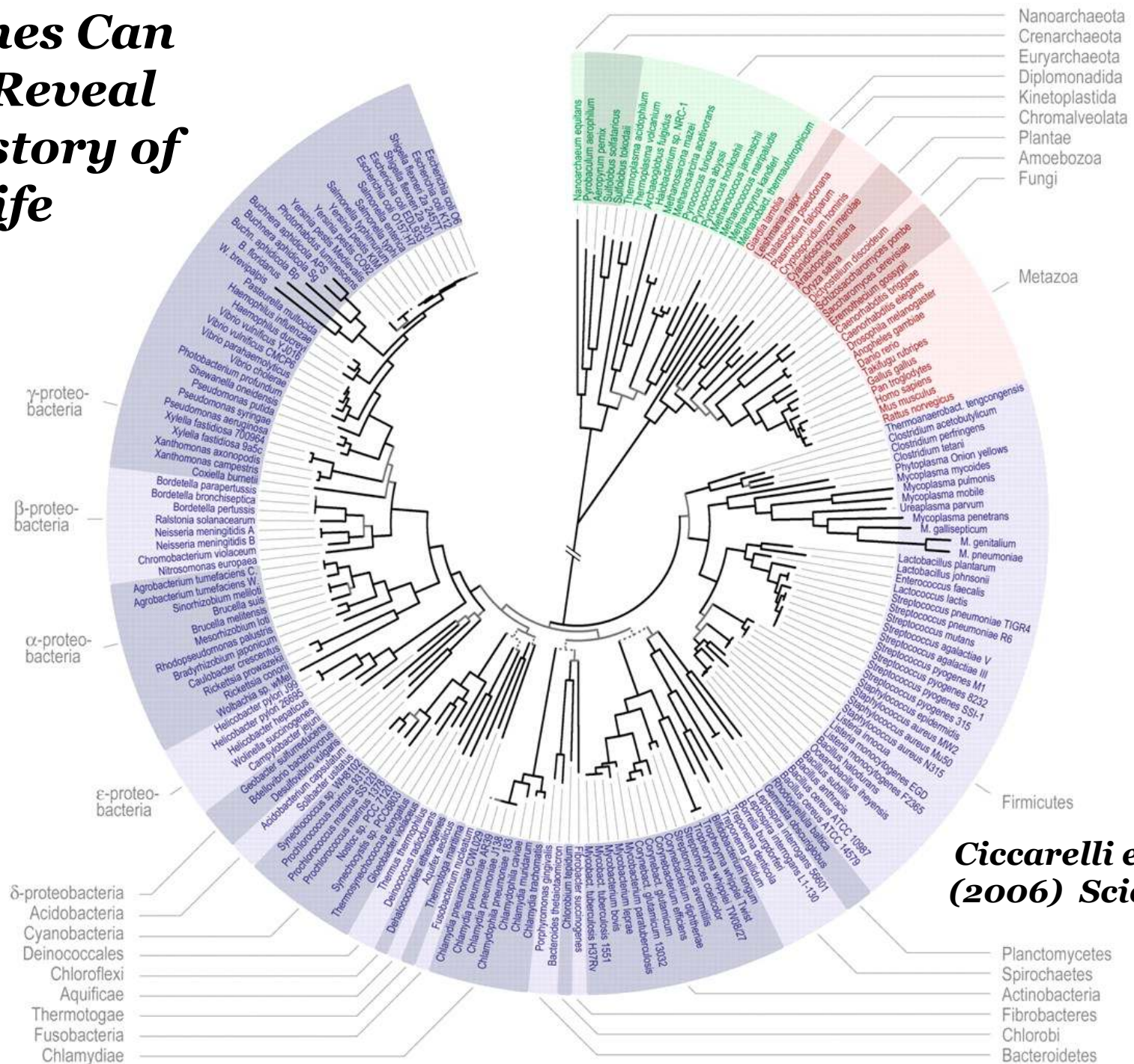
What Makes Us Sick Is the Stuff of Life

F	W	Y	Cancer	F	W	Y	Neurological	F	W	Y	Malformation Syndromes
+			ABL1	+			Adrenoleukodystrophy-ABCD1	-			Aarskog-Scott-FGD1
+			Acute Myeloid Leukemia-DEK	+			Alzheimer-PS1	+			Achondroplasia-FGFR3
+			Adenomat. Polyposis Coli-APC	+			Alzheimer-APP	+			Alagille-JAG1
+			AKT2	+			Amyotrophic Lat. Sclero.-SOD1	+			Barth-TAZ
+			Ataxia Telangiectasia-ATM	+			Angelman-UBE3A	-			Beckwith-Wiedemann-CDKN1C
-			BRCA1	+			Aniridia-PAX6	-			Cerebral Cavern. Malf.-CCM1
-			BRCA2	+			Best Macular Dystrophy-VMD2	+			Chondrodyspl. Punct. 1-ARSE
+			Basal Cell Nevus-PTC	+			Ceroid-Lipofuscinosis-PPT	+			Cleidocranial Dysplasia-OFC1
+			B-Cell Lymphoma 2-BCL2	+			Ceroid-Lipofuscinosis-CLN3	-			Cockayne I-CKN1
-			B-Cell Lymphoma 3-BCL3	-			Ceroid-Lipofuscinosis-CLN2	+			Coffin-Lowry-RPS6KA3
+			Bloom-BLM	-			Charcot-Marie-Tooth 1A-PMP22	+			Diastrophic Dyspl.-SLC26A2
+			Burkitt's Lymphoma-MYC	-			Charcot-Marie-Tooth 1B-MPZ	+			EEC 3-Ket. P63
-			CDKN2C	+			Choroideremia-CHM	+			Greig Cephalopolysynd.-GLI3
-			CSF1R/C-Fms	-			Creutzfeldt-Jakob-PRNP	-			Hand-Foot-Genital-HOXA13
+			Chk2 Protein Kinase	+			Deafness, Hereditary-MYO15	+			Holoprosencephaly 3-SHH
-			PDGFB	+			Deafness, X-Linked-TIMM8A	+			Holoprosencephaly-SIX3
+			CML-BCR	+			Diaphanous 1-DIAPH1	+			Holt-Oram-TBX5
+			Cyclin D1-CCND1	+			Dementia, Multi-Infarct-NOTCH3	-			ICF-DNMT3B
+			Cyclin Dep. Kinase 4-CDK4	+			Duchenne MD ⁺ -DMD	+			Kallman-KAL1
+			EGFR	-			Emery-Dreifuss MD ⁺ -EMD	-			Laterality, X-Linked-ZIC3
+			ERBB2	+			Emery-Dreifuss MD ⁺ -LMNA	+			Melnick-Fraser-EYA1
-			ETS	+			Familial Encephalopathy-PI12	+			Nail Patella-LMX1B
+			E-Cadherin-CDH1	+			Fragile-X-FRAXA	-			Opitz-MID1
+			Ewing Sarcoma-FLI-1	+			Friedreich Ataxia-FRDA	+			Renal Coloboma-PAX2
-			FGF3	+			Frontotemporal Dement.-TAU	+			Rieger, Type 1-PITX2
-			Fanconi's Anemia A-FANCA	-			Fukuyama MD ⁺ -FCMD	-			Rubinstein-Taybi-CREBBP
-			Fanconi's Anemia C-FANCC	+			Huntington-HD	+			Saethre-Chotzen-TWIST
-			Fanconi's Anemia G-FANCG	+			Limb Girdle MD ⁺ 2A-CAPN3	-			Septo-optic Dysplasia-HESX1
+			HNPCC*-MSH2	+			Limb Girdle MD ⁺ 2B-YSF	+			Simpson-Golabi-Behmel-GPC3
+			HNPCC*-MSH3	-			Limb Girdle MD ⁺ 2E-BSG	+			Townes-Brockes-SALL1
+			HNPCC*-MSH6	+			Lissencephaly, X-Linked-DCX	-			Treacher-Collins-TCOF1
+			HNPCC*-MLH1	+			Lowe Oculocerebroren.-OCRL	+			VMCM-TEK
+			HNPCC*-PMS2	-			Machado-Joseph-MJD1	+			Wardenburg-PAX3
-			KIT	+			Miller-Dieker Lissen.-PAF	+			Zellweger-PEX1

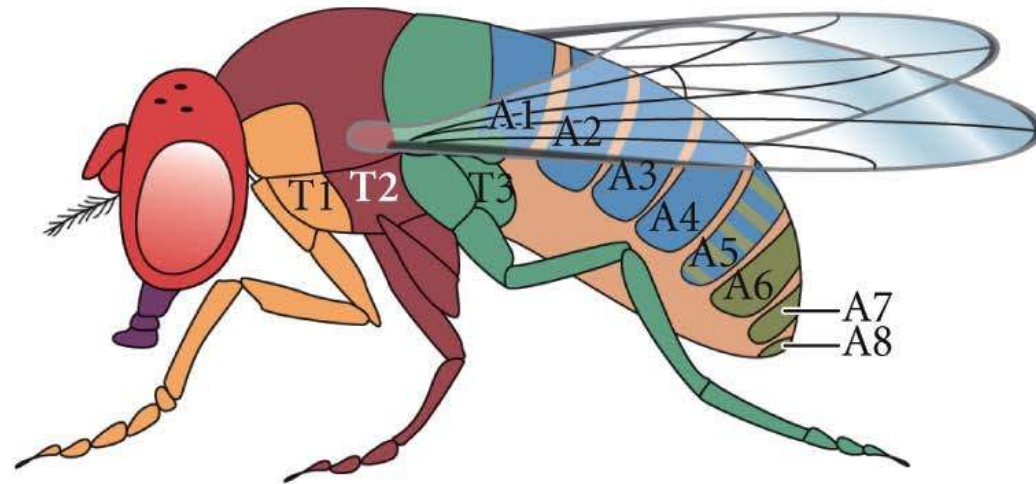


*Human disease-associated genes shared with flies (F), worms (W), and Yeast (Y);
from Rubin et al. (2000) Science*

Genomes Can Help Reveal the History of Life



The Drosophila Body- Building Genes



Antennapedia complex

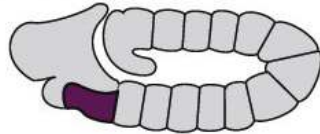
bithorax complex

lab *Pb* *Dfd* *Scr* *Antp*

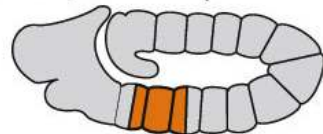
Ubx *abdA* *AbdB*



labial (*lab*)



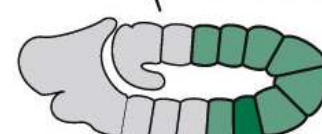
Deformed (*Dfd*)



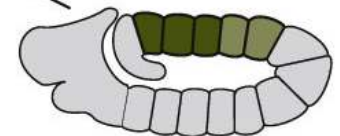
Sex combs reduced (*Scr*)



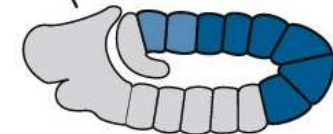
Antennapedia (*Antp*)



Ultrabithorax (*Ubx*)

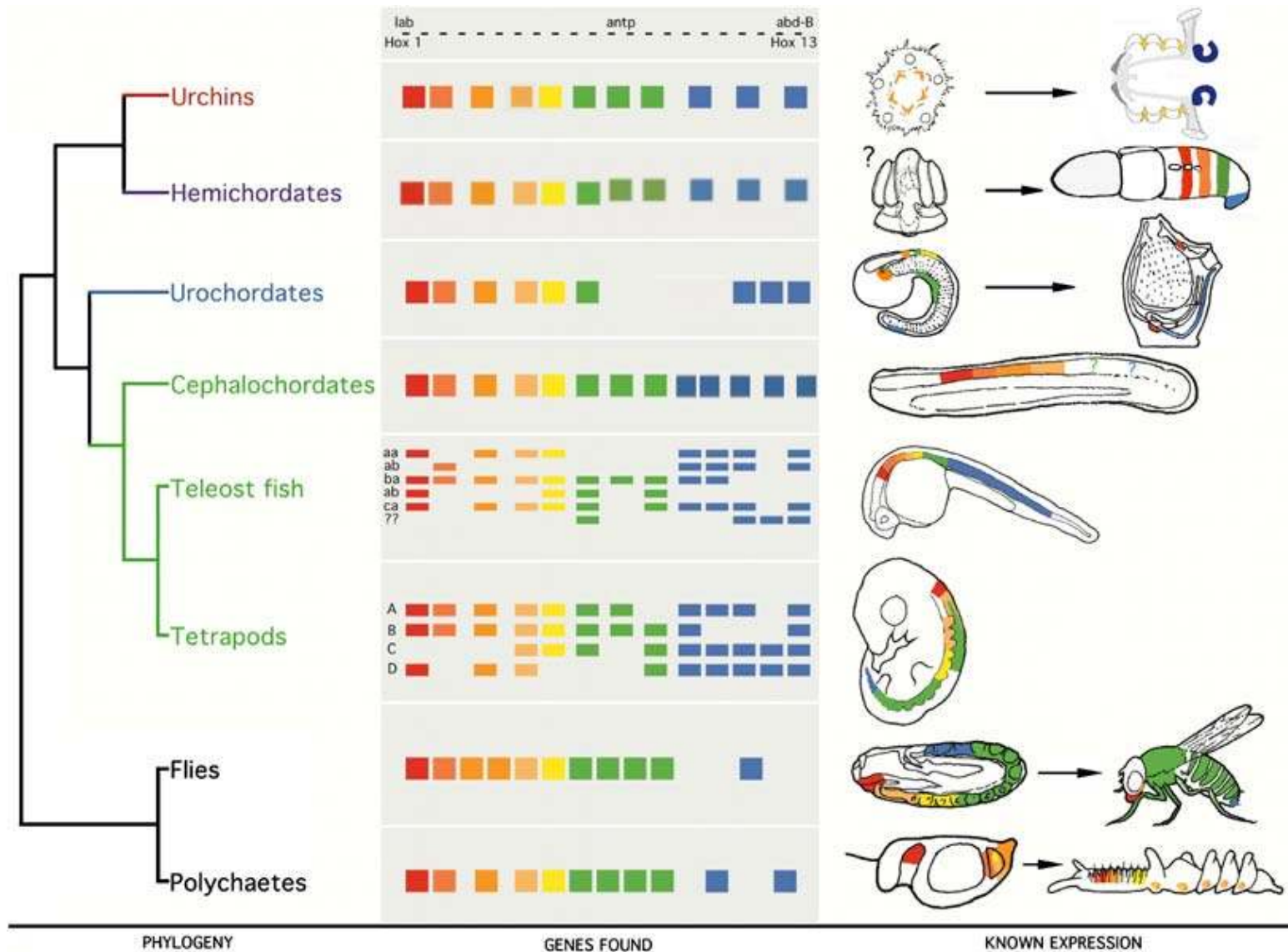


Abdominal B (*AbdB*)

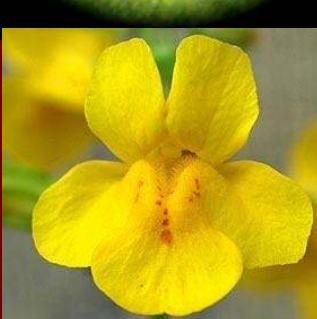
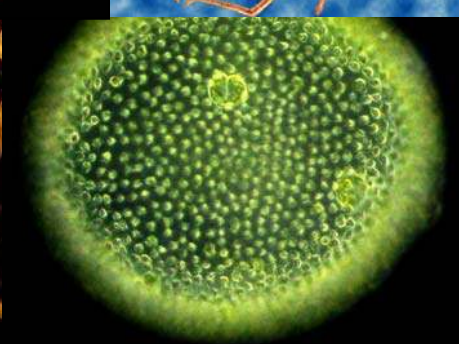
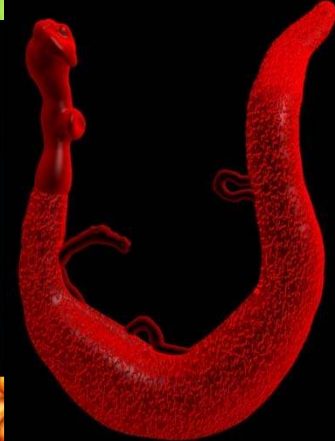
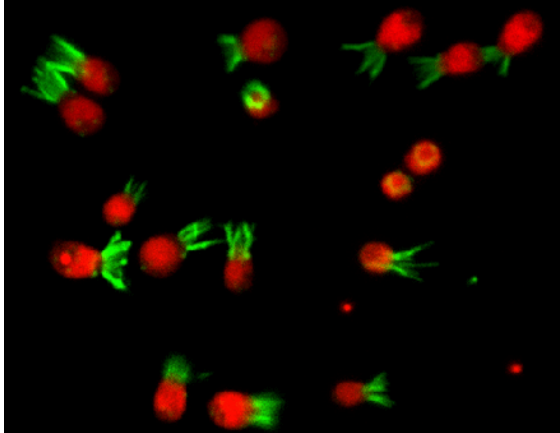
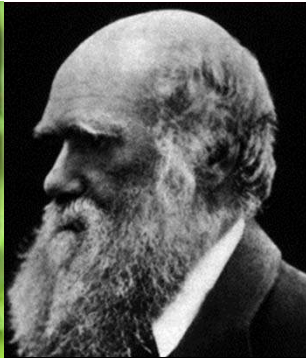


abdominal A (*abdA*)

Animal Bodies are Built from the Same Genetic Toolkit

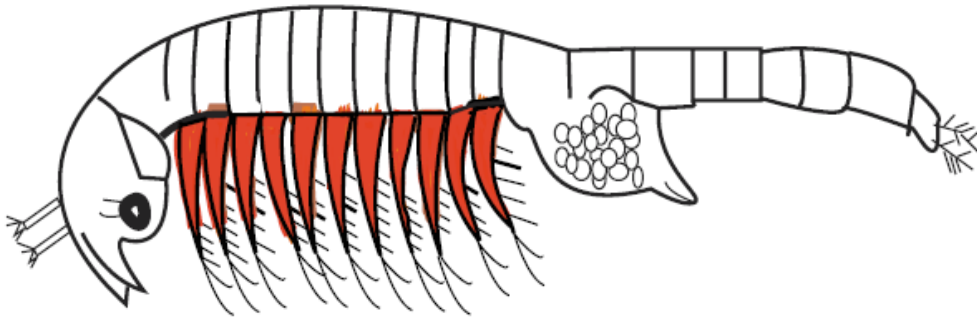


Swalla (2006) Heredity

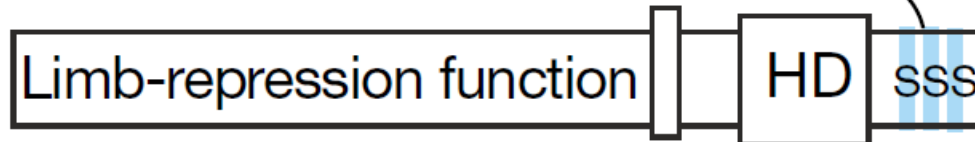
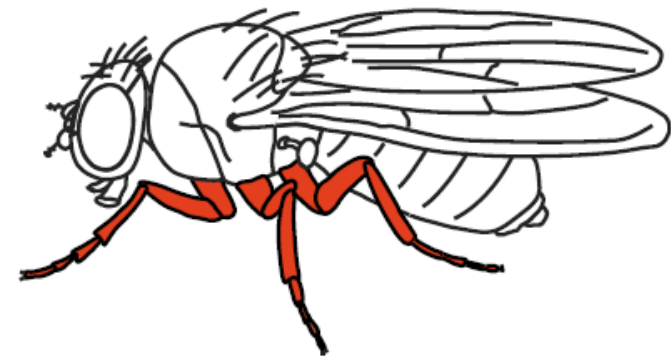


Altering the Coding Sequence of Genes

Brachiopod crustaceans



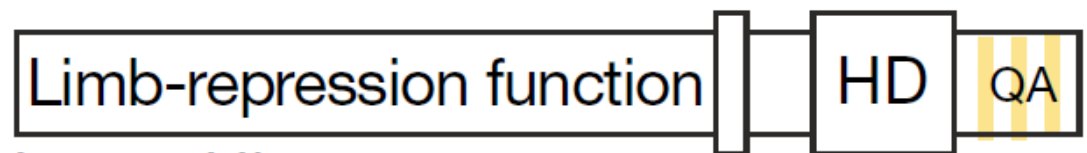
Insects



Branchiopod crustacean Ubx

❖ Loss of Ser phosphorylation sites

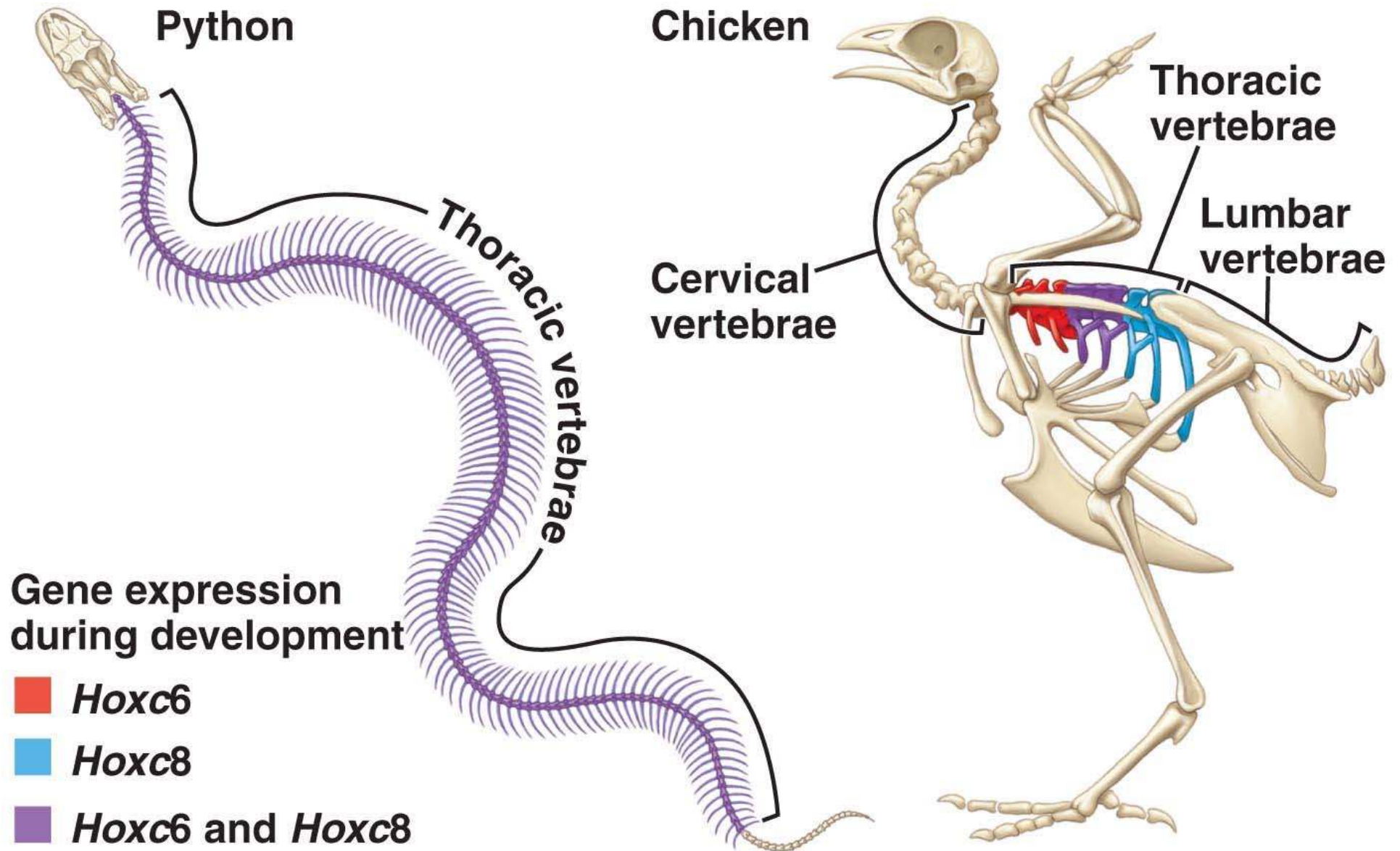
❖ Expansion of QA repeat



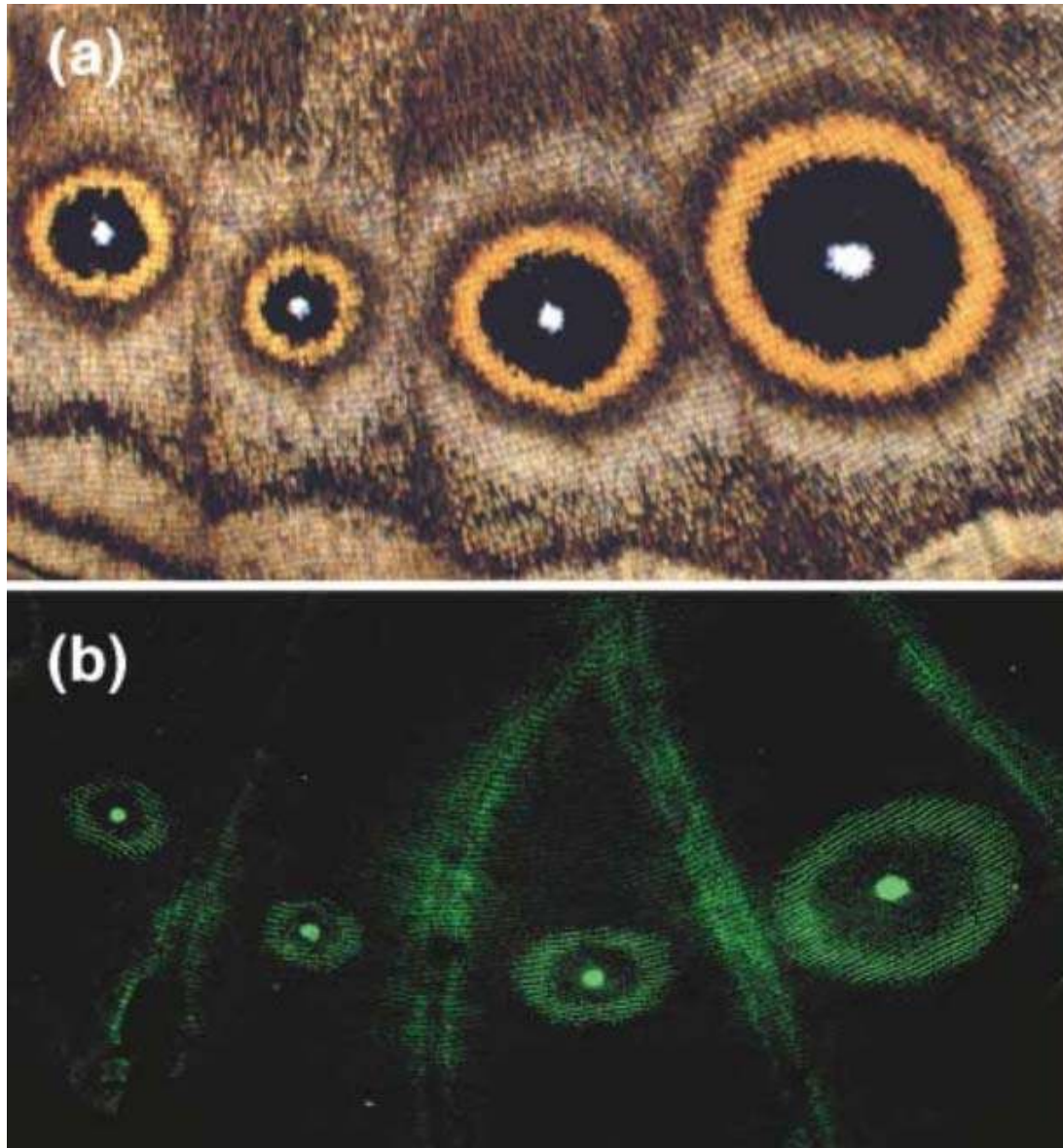
Insect Ubx



Altering the Expression of Genes



Teaching Genes New Tricks (Gene Co-Option)



Brunetti et al. (2001) Curr. Biol.

High-Throughput DNA Sequencing Technologies

454 / Roche

450 bp 1.5 Gbp / day



Illumina

150 bp 35 Gbp / day



Helicos

55 bp 4.5 Gb / day



“Traditional” Sanger / Capillary Sequencing

650 bp

2 Mbp / day



SOLiD ABI

75 bp 22 Gbp / day



PacBio

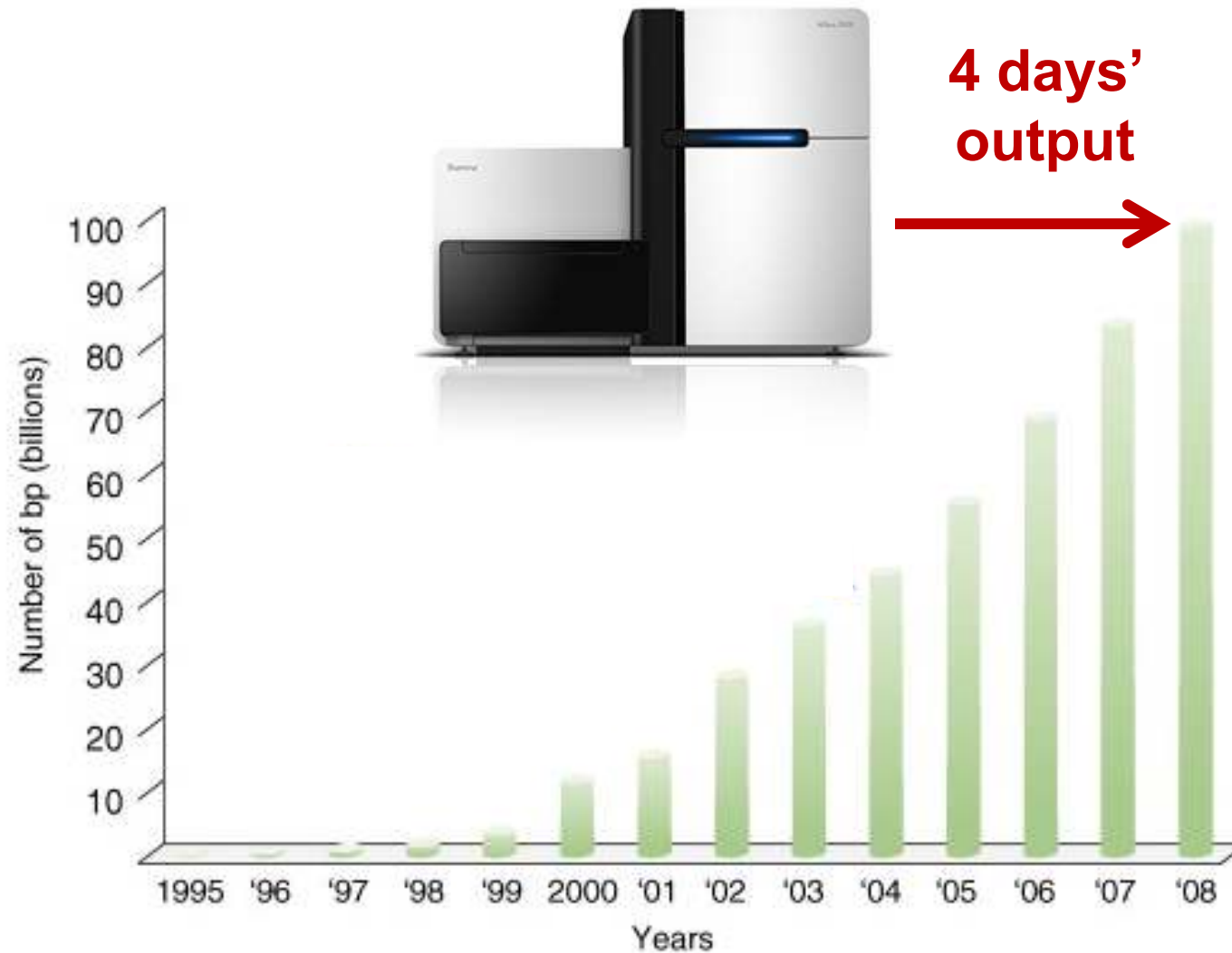
1000 bp 70 Gbp / day



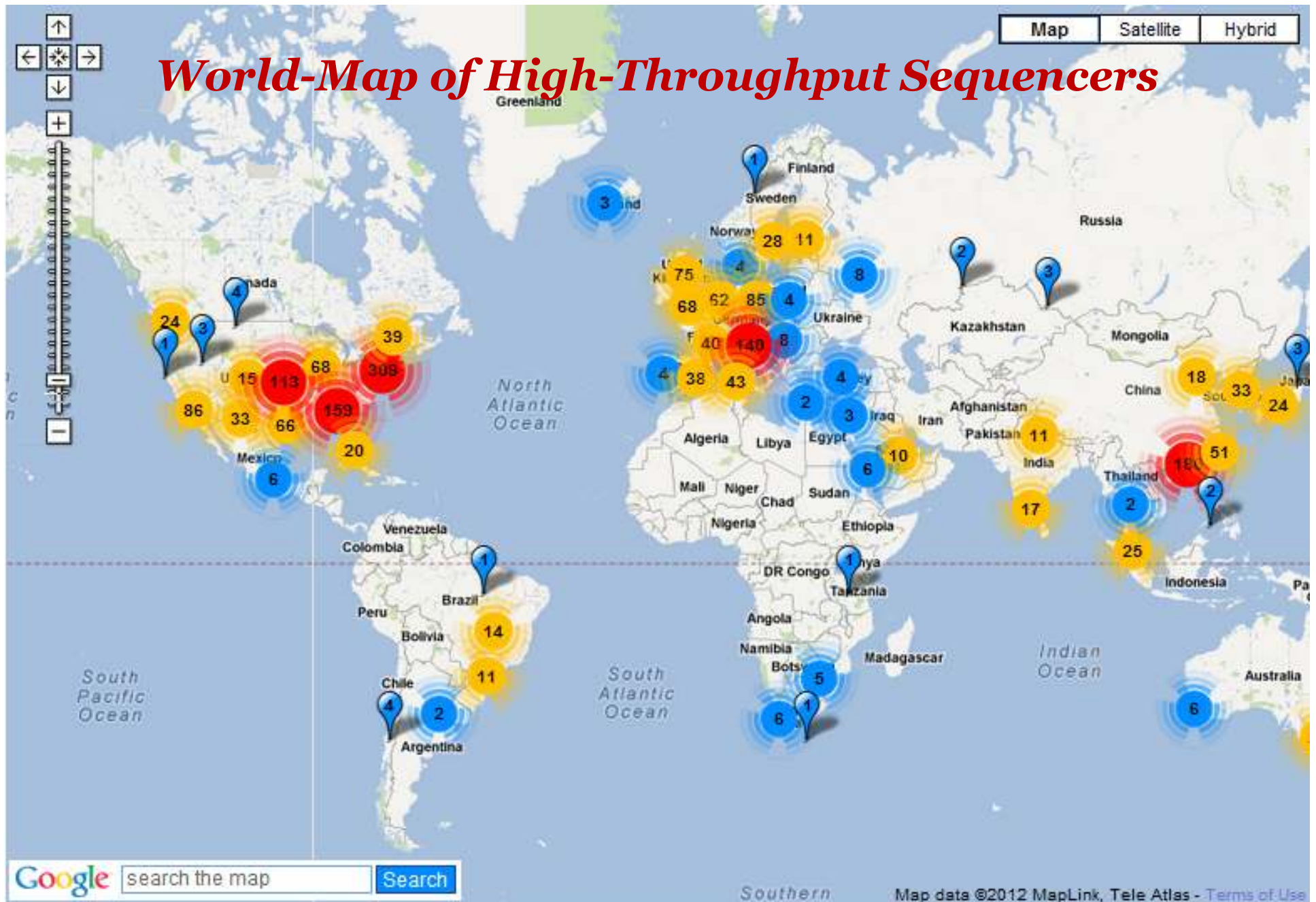
Ion PGM

100 bp 120 Gbp / day

Why Is High-Throughput DNA Sequencing So Exciting?

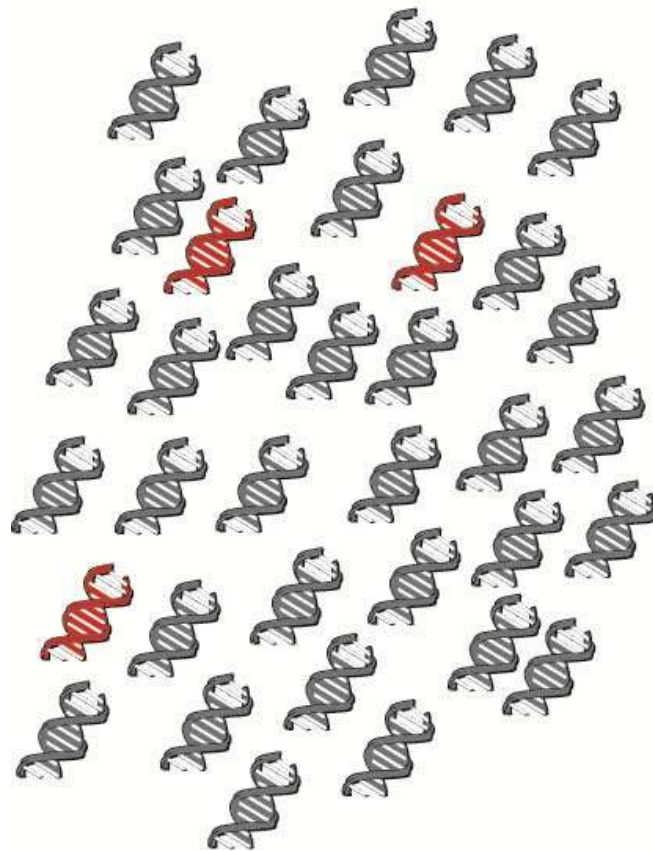


Gilad et al. (2009) Trends Genet.



<http://pathogenomics.bham.ac.uk/hts/>

Next-Gen Sequencing is Qualitative and Quantitative



NGSTs

Each DNA template
is sequenced directly

Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript

Capillary Sequencing

All DNA templates are sequenced
together to create a single
consensus sequence

Grey transcript



Lecture Outline

❖ **Introduction to Genome Evolution**

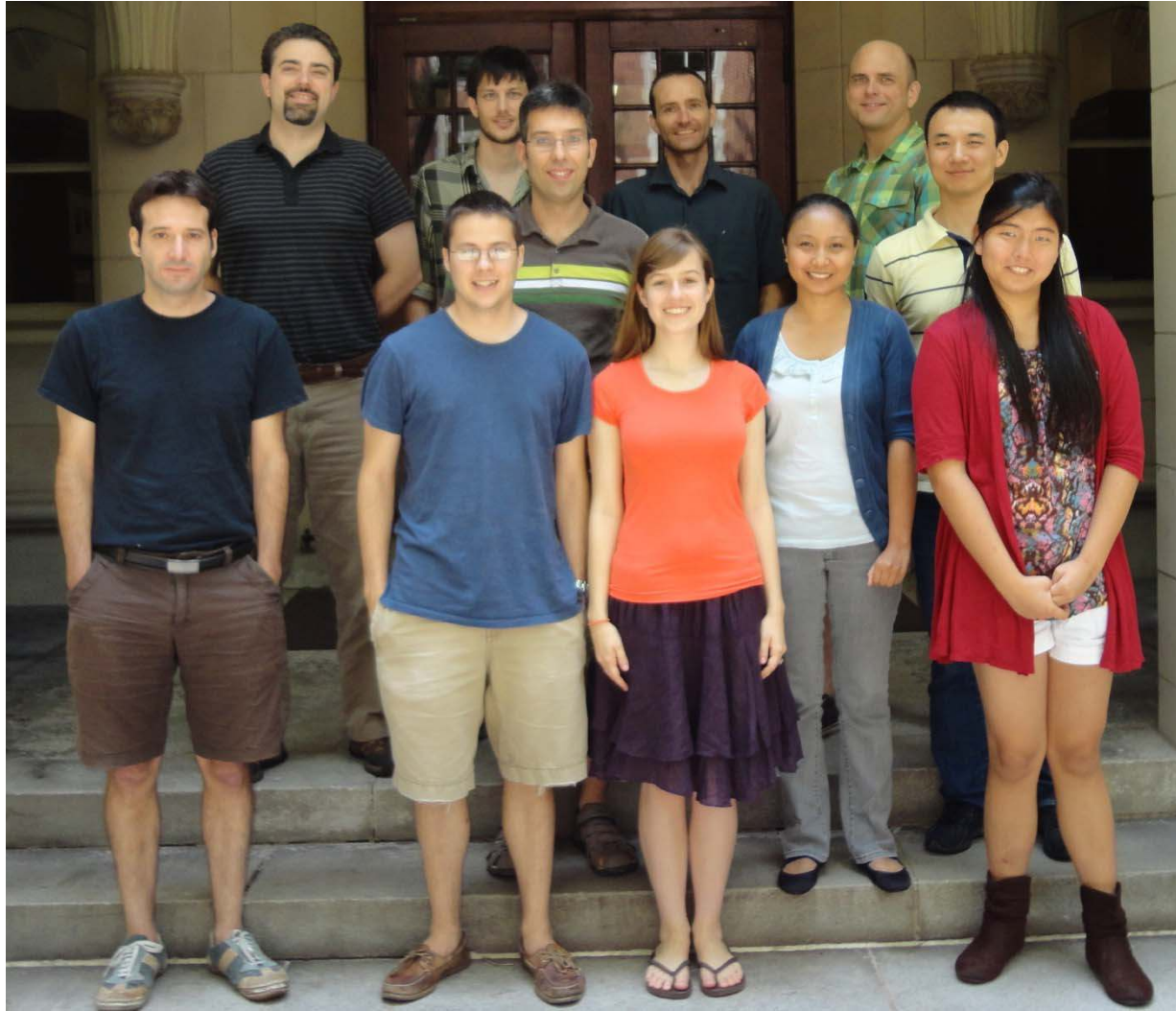
❖ **Comparative Functional Genomics**

----- **Coffee Break** -----

❖ **Population Genomics**

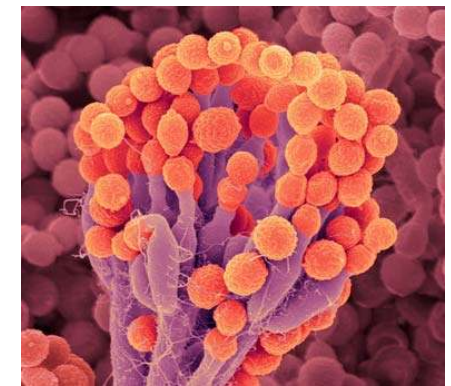
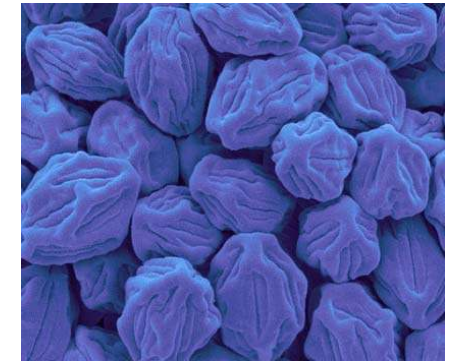
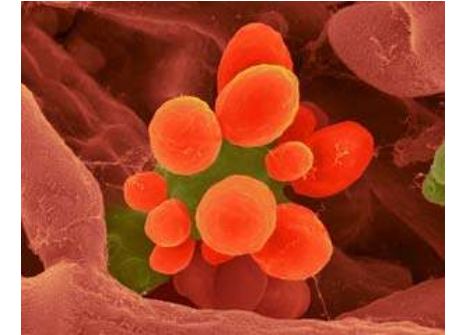
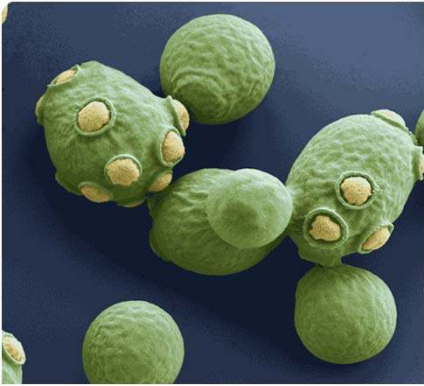
❖ **Phylogenomics**

The Rokas Lab

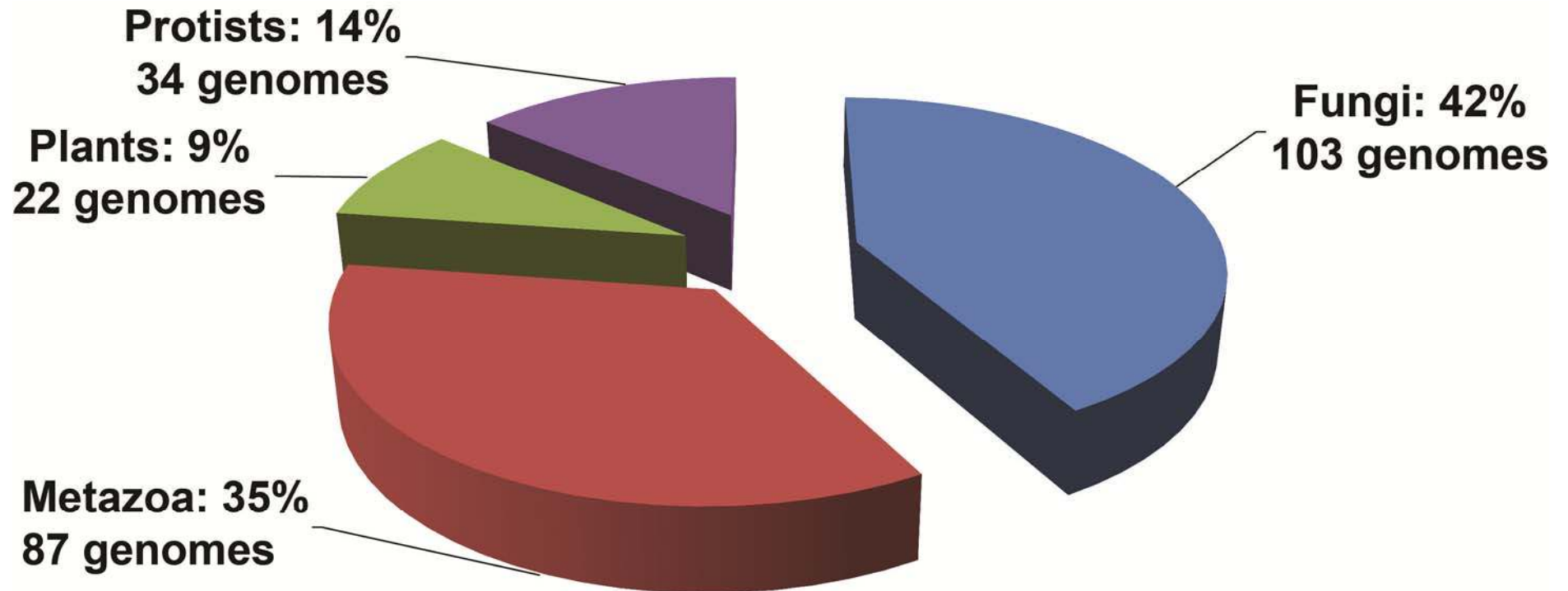


We study the DNA record to gain insight into evolution's patterns and processes using computational and experimental approaches in a variety of lineages

Fungi: A Model for the Study of the DNA Record



Fungi: the Most Sequenced Eukaryotic Lineage



fungi: 1000genomes

<http://www.jgi.doe.gov/sequencing/cspseqplans2012.html>



Data from GOLD 3.0 (www.genomesonline.org), March 2011

What Questions Should We Ask?

“Does the genome sequence *per se* now enable us to decode the language of life? The answer, of course, is an emphatic “no”.

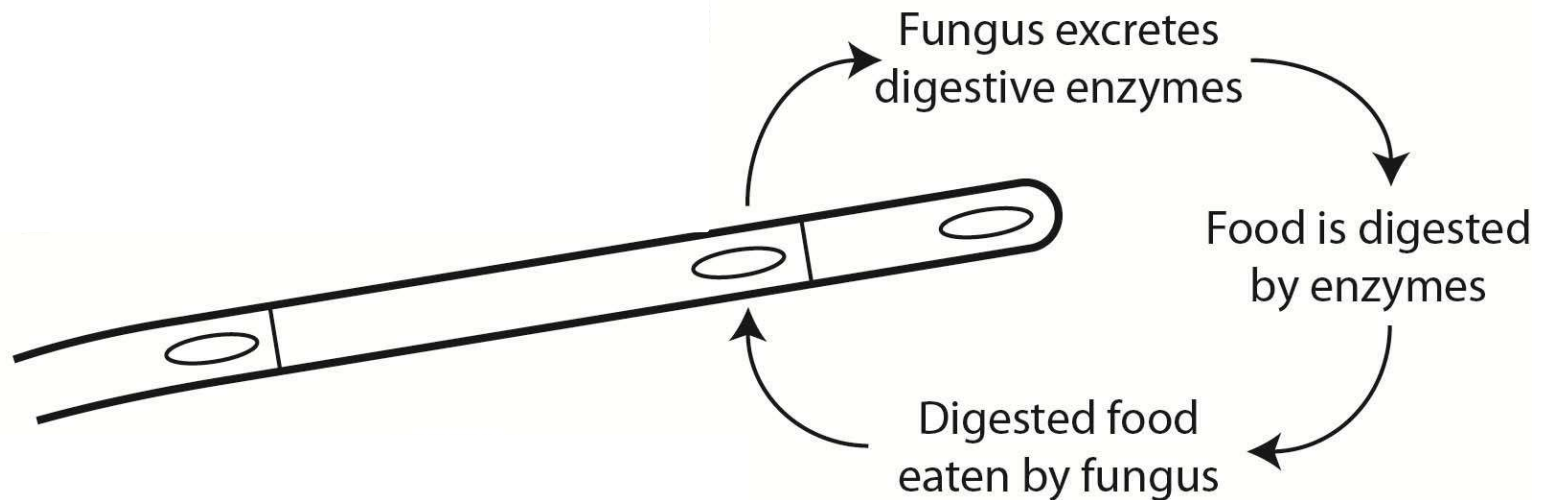
[...] Thanks to genomics and reverse genetics, we now know that the inactivation of most genes yields either no obvious phenotype or early death, neither of which is immediately informative of the gene's function.

In any case, even a gene whose mutant phenotype is, say, a change in nose shape, cannot be called “a gene for the nose”. The genome does not say “make nose”, it says “make serpentine receptors”—it speaks biochemistry, not phenotype.”





Fungi Eat by Absorption

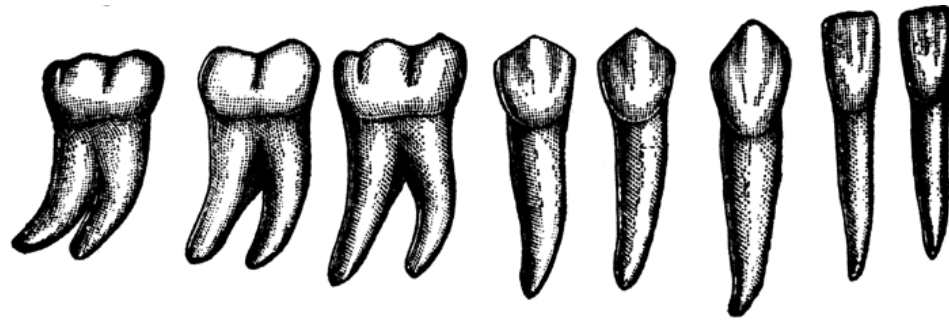


❖ **Fungi have evolved powerful “weapons” to defend their food**

❖ **Different fungi have specialized to “eating” different foods**

Metabolism is Fundamental to the Fungal Lifestyle

- ❖ The genes involved in fungal primary metabolism are their “teeth”



- ❖ The genes involved in fungal secondary metabolism are their “horns”, “spines”, and “claws”



Beetle horns



Osprey claws



Echidna spines



Fungi “Eat” Almost Everything, Grow on Any Surface



Wood, leaves, nails, leather, cloth, manure, animal carcasses, live hosts, ink, syrup, paint, glue, hair

“I put maybe a shot of whiskey in a liter of agar and filled the petri plates with it. That made [the fungus] grow a hell of a lot faster”



*Dr. James Scott, Mycologist
Wired Magazine, 06/11 issue*

What Fungi Cannot Eat!



Diner fries

McDonald's fries

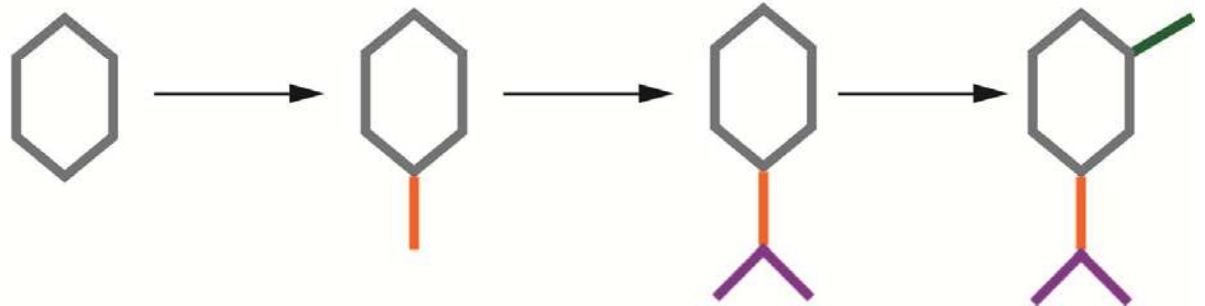
Fungal Metabolic Genes Are Often Physically Clustered



Enzymes



Metabolic pathway

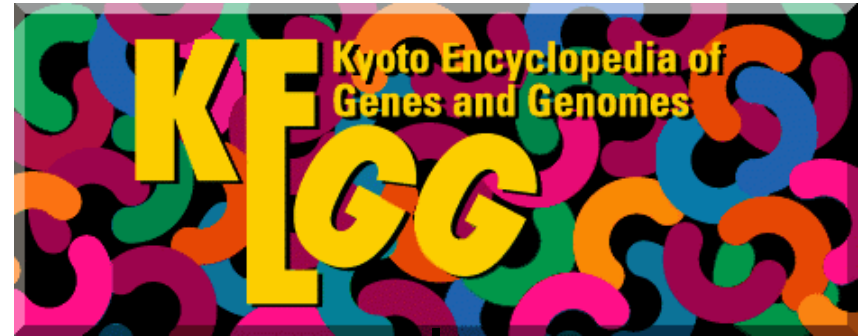


HOW? WHY?

The CRAP (Cluster Reconstruction And Phylogeny) Pipeline



Jason Slot



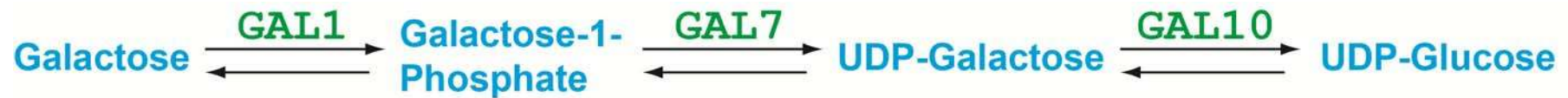
Retrieve genes from 173 pathways from
35 KEGG-annotated genomes (23 fungi)

Search for presence of pathway
homologs in ~130 genomes (105 fungi)
by BLAST and OrthoMCL

Infer phylogenetic
relationships of
pathway homologs

Search for
physical linkage
(clustering) of
pathway homologs

Variation in GAL Gene Clustering Across Fungi



Saccharomyces (Ascomycota: Saccharomycetes)



Aspergillus (Ascomycota: Eurotiomycetes)



Cryptococcus (Basidiomycota: Tremellomycetes)



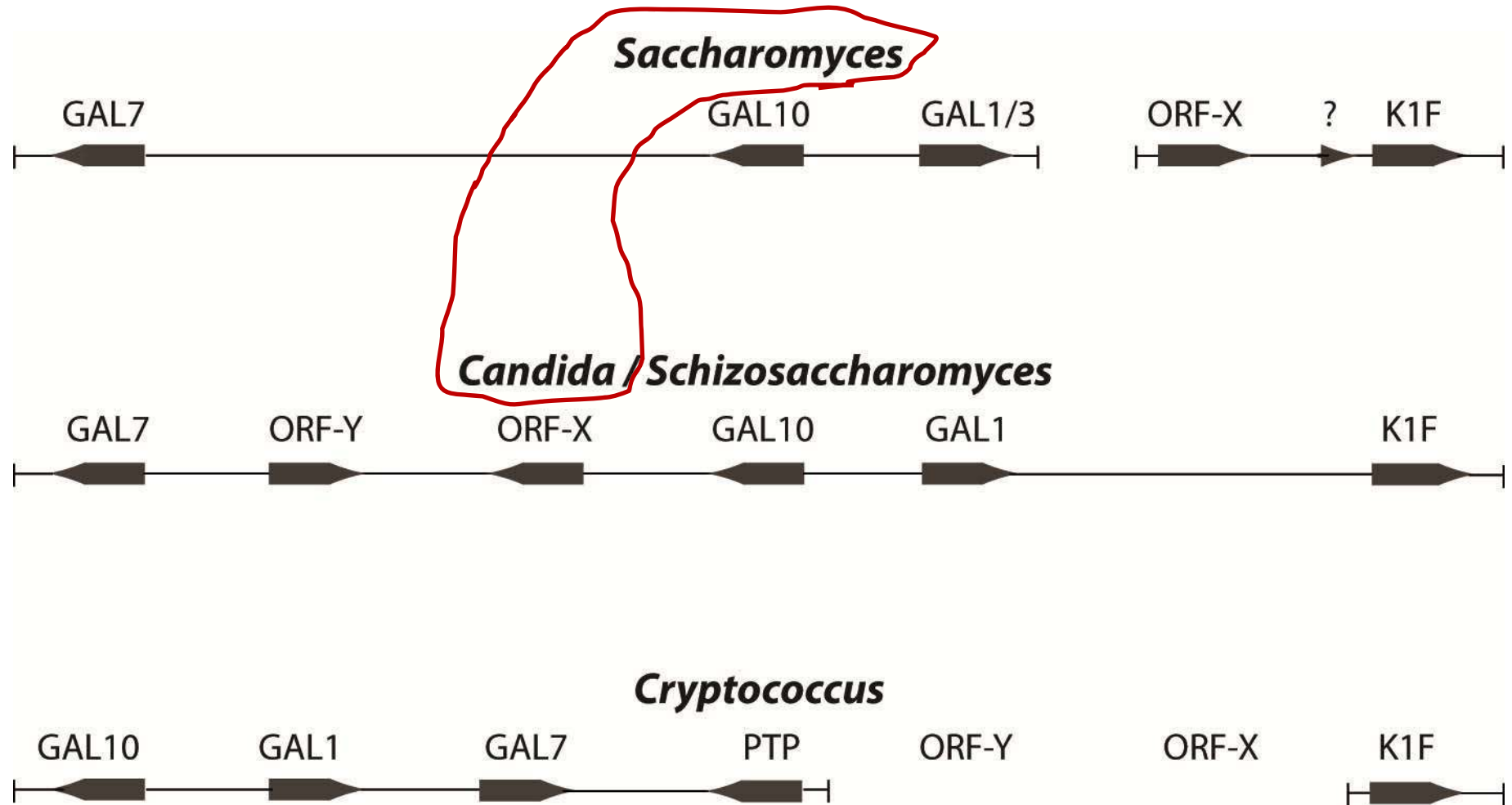
Ustilago (Basidiomycota: Ustilaginomycetes)



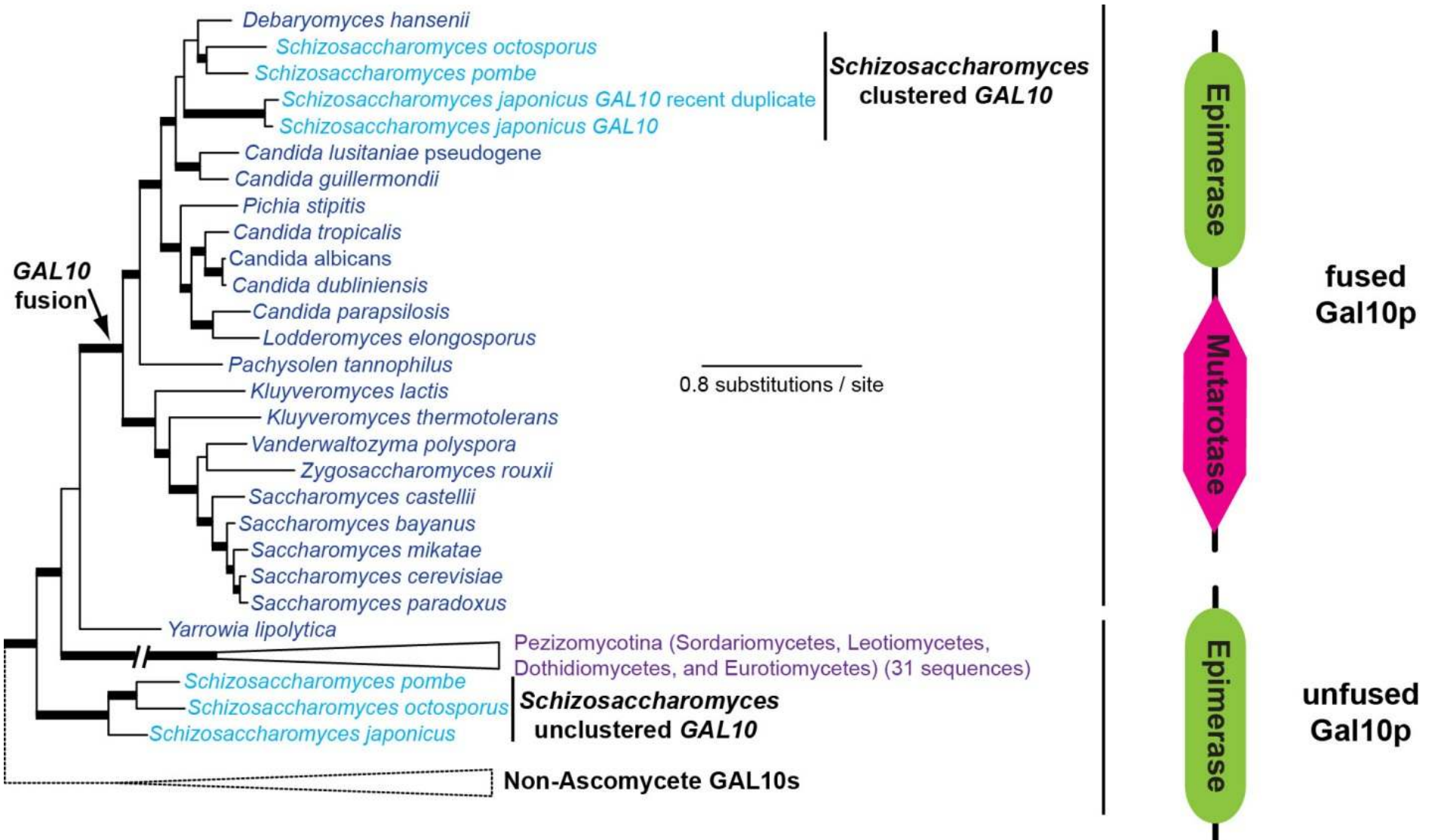
Batrachochytrium (Chytridiomycota: Chytridiomycetes)



Three Distinct Types of Fungal GAL Gene Clusters



Schizosaccharomyces Gained Its GAL Cluster by HGT from Candida



Even Very Large Gene Clusters are Subject to HGT

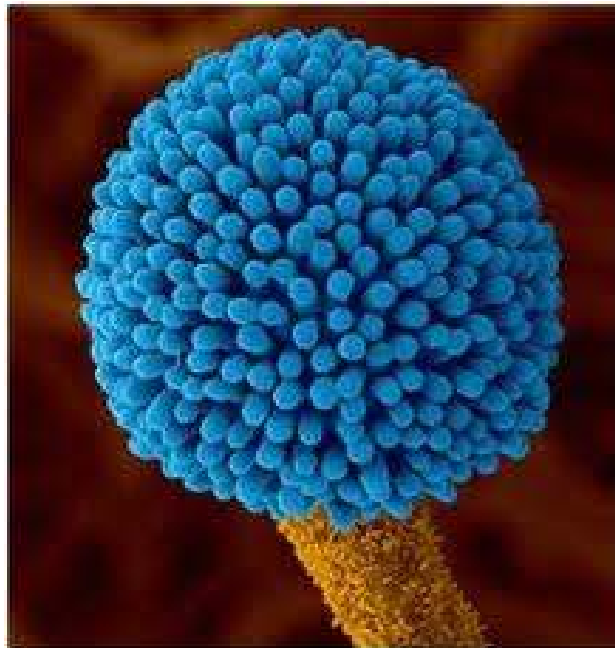


The Sterigmatocystin Gene Cluster

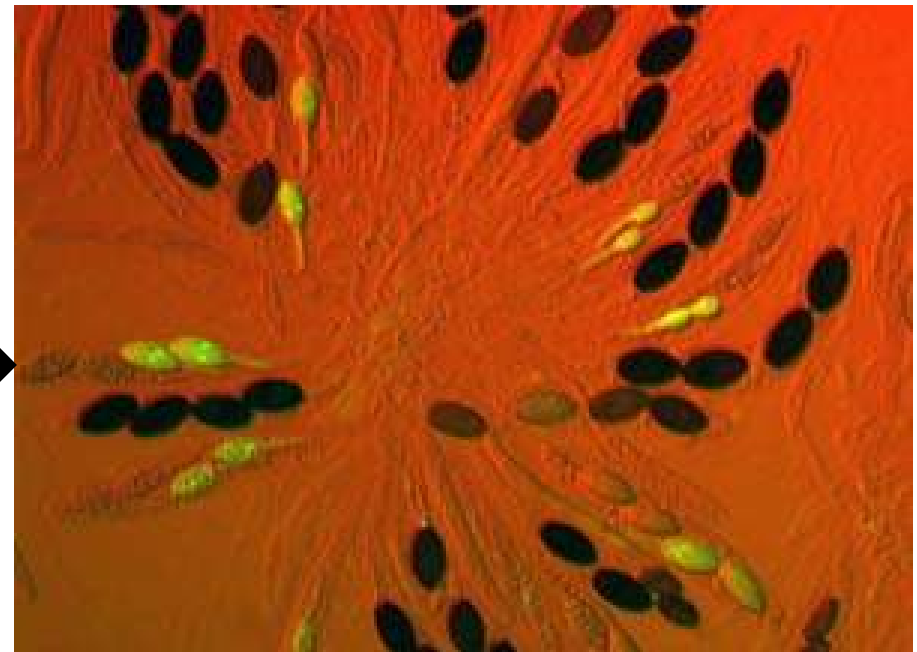


In the US, agricultural loss due to aflatoxin contamination is ~ \$1B/yr

Structural genes
Regulators



HGT



***Aspergillus* (Eurotiomycetes) *Podospora* (Sordariomycetes)**



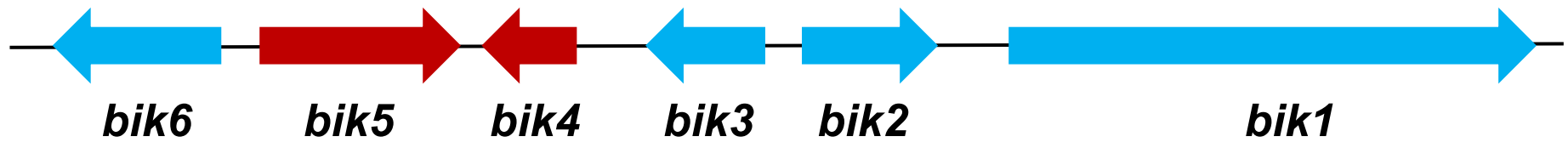
Slot & Rokas (2011) Curr. Biol.

Larvicidal activity of metabolites from the endophytic *Podospora* sp. against the malaria vector *Anopheles gambiae*

Josphat C. Matasyoh • Birger Dittrich •
Anja Schueffler • Hartmut Laatsch

Abstract In a screening for natural products with mosquito larvicidal activities, the endophytic fungus *Podospora* sp. isolated from the plant *Laggera alata* (Asteraceae) was conspicuous. Two xanthones, sterigmatocystin (1) and secosterigmatocystin (2), and an anthraquinone derivative (3) 13-hydroxyversicolorin B were isolated after fermentation on M₂ medium. These compounds were characterised using spectroscopic and X-ray analysis and examined against third instar larvae of *Anopheles gambiae*. The results demonstrated that compound 1 was the most potent one with LC₅₀ and LC₉₀ values of 13.3 and 73.5 ppm, respectively. Over 95% mortality was observed at a concentration 100 ppm after 24 h. These results compared

Does HGT of Metabolic Gene Clusters Contribute Anything Else?



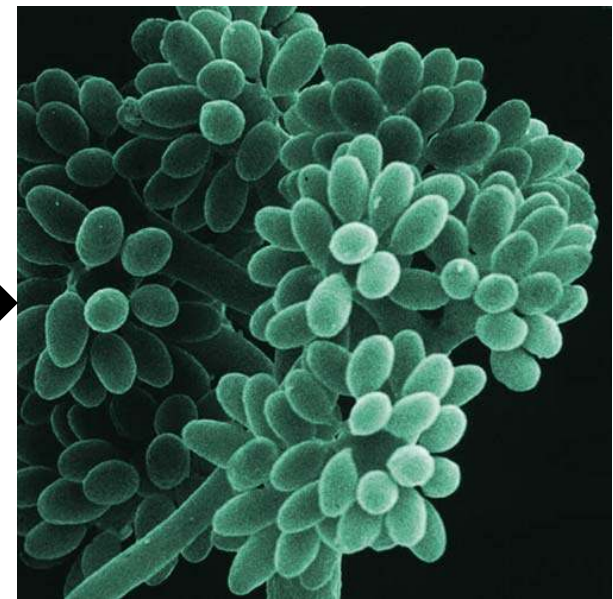
The Bikaverin Gene Cluster

 Structural genes
 Regulators



Fusarium (Sordariomycetes)

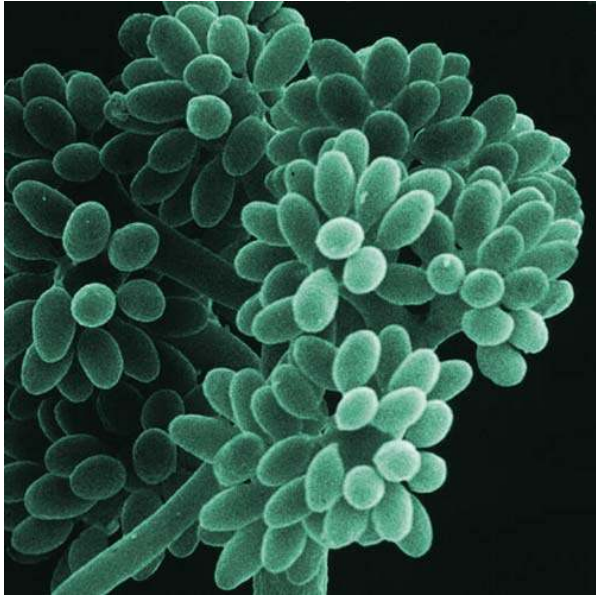
HGT

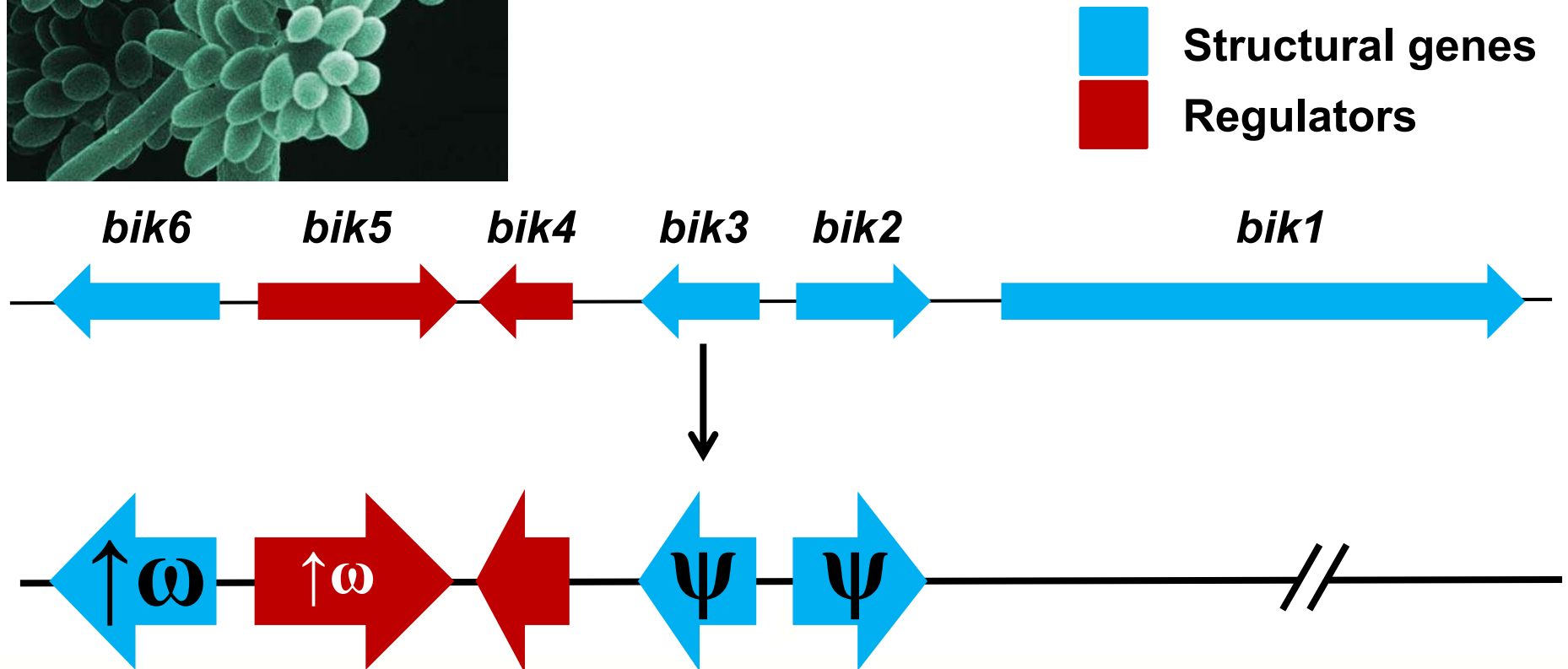
Botrytis (Leotiomycetes)



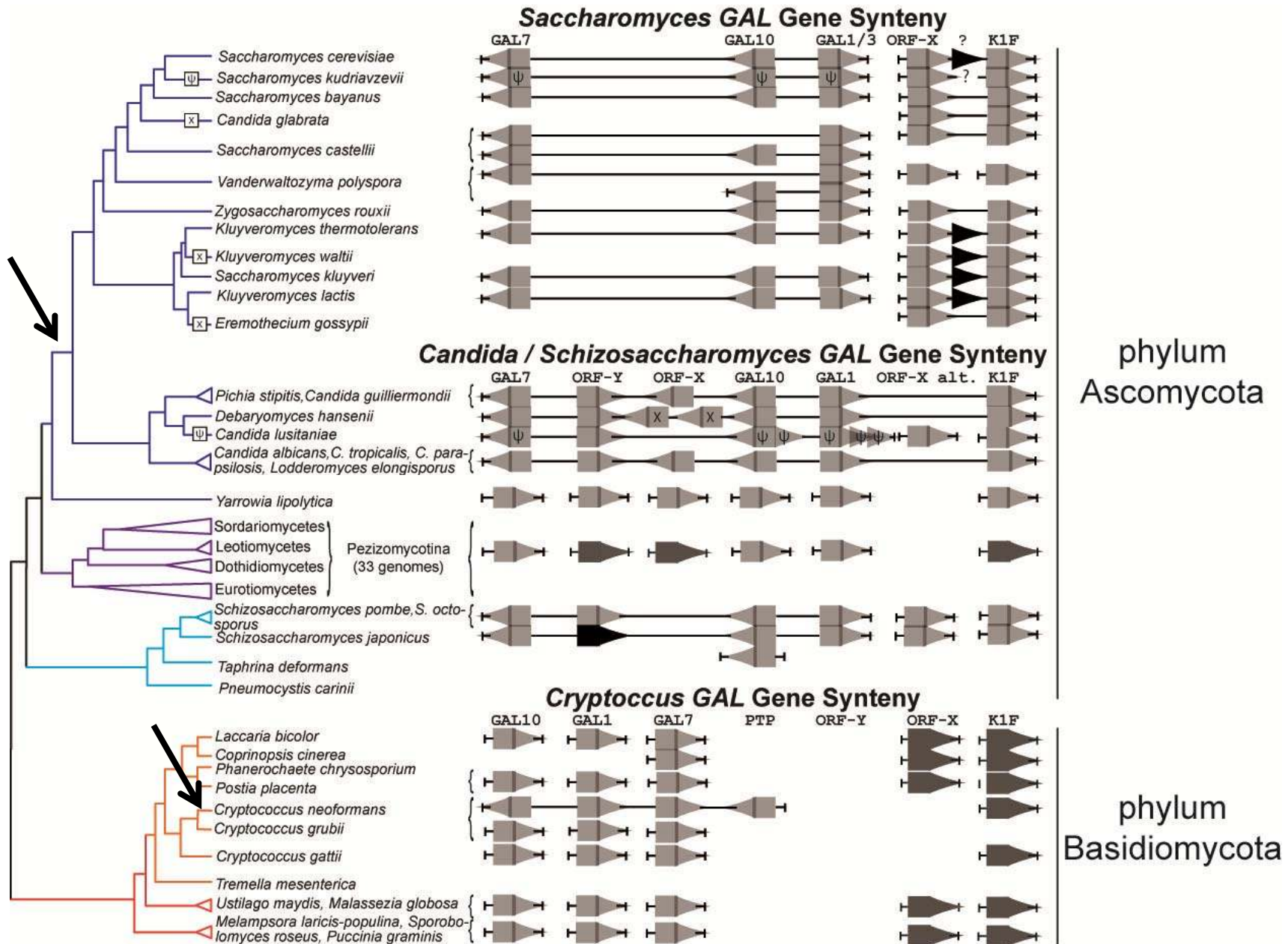
HGT of Metabolic Clusters Also Contributes Novel Regulation



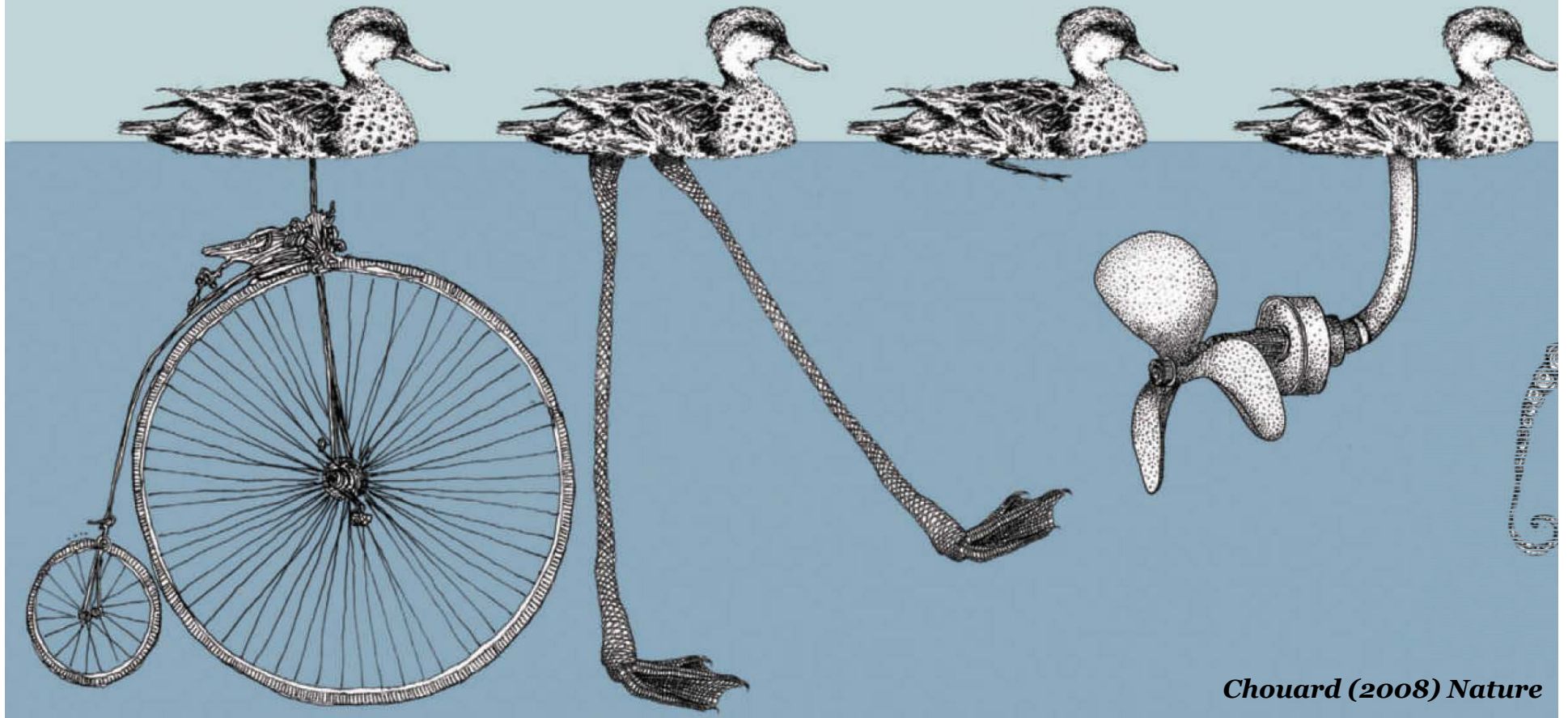
***Botrytis* (Leotiomyces)**



The *Saccharomyces* & *Cryptococcus* Clusters Evolved Independently



Independently Evolved (Analogous) Traits



Chouard (2008) Nature

Saccharomyces / Candida lineage

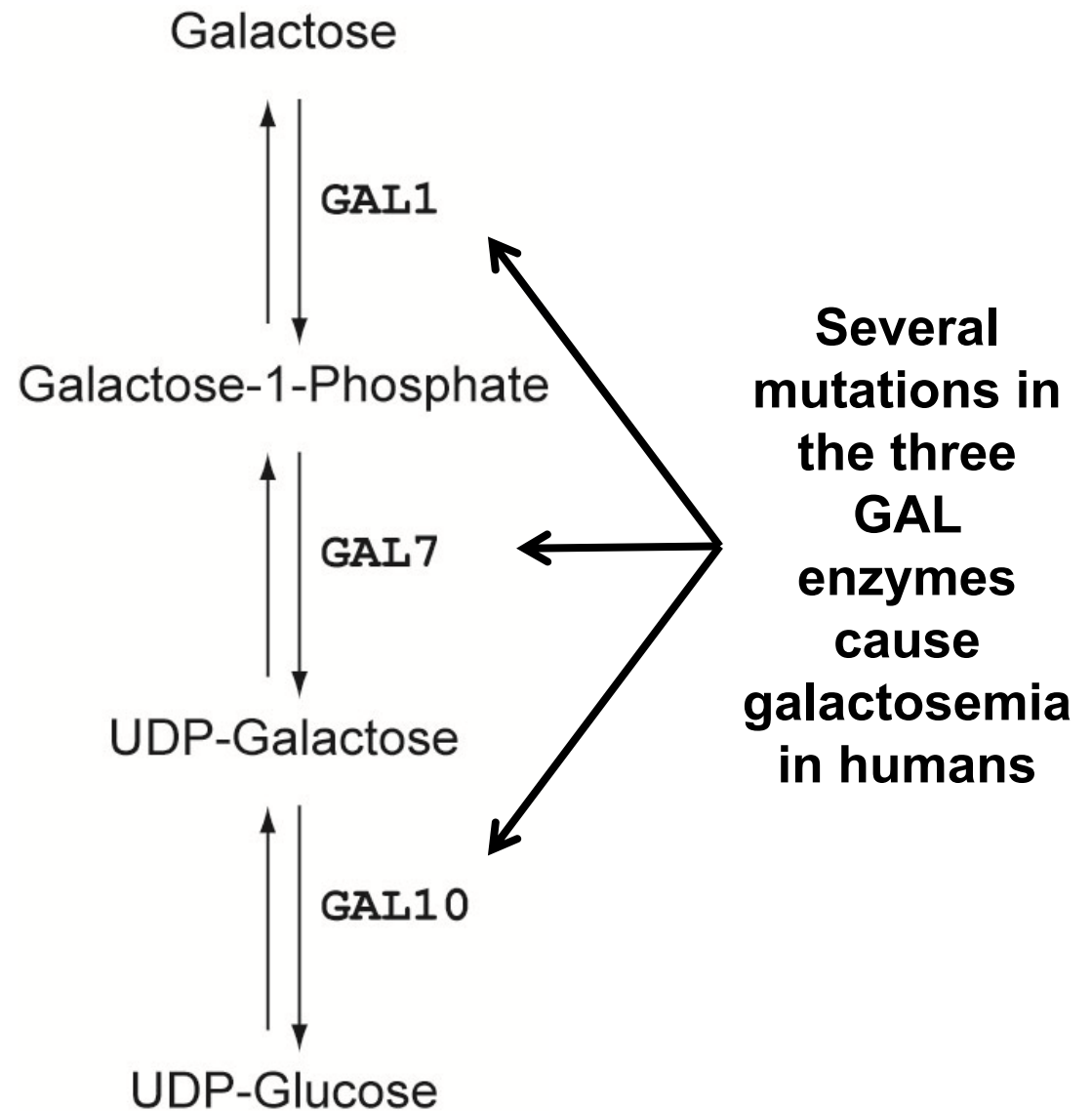
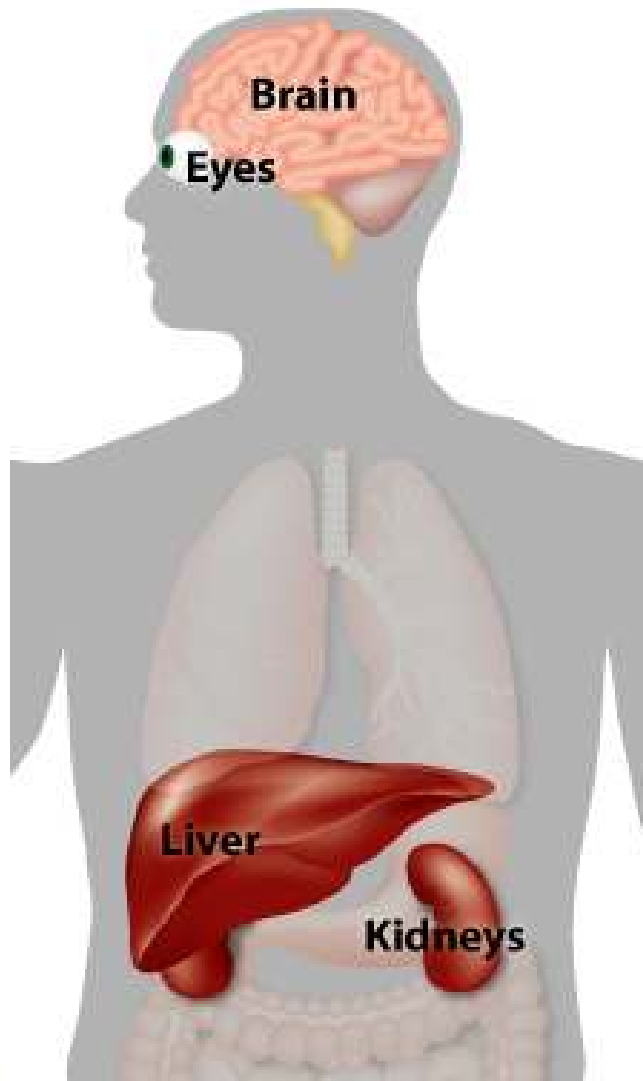


Cryptococcus lineage

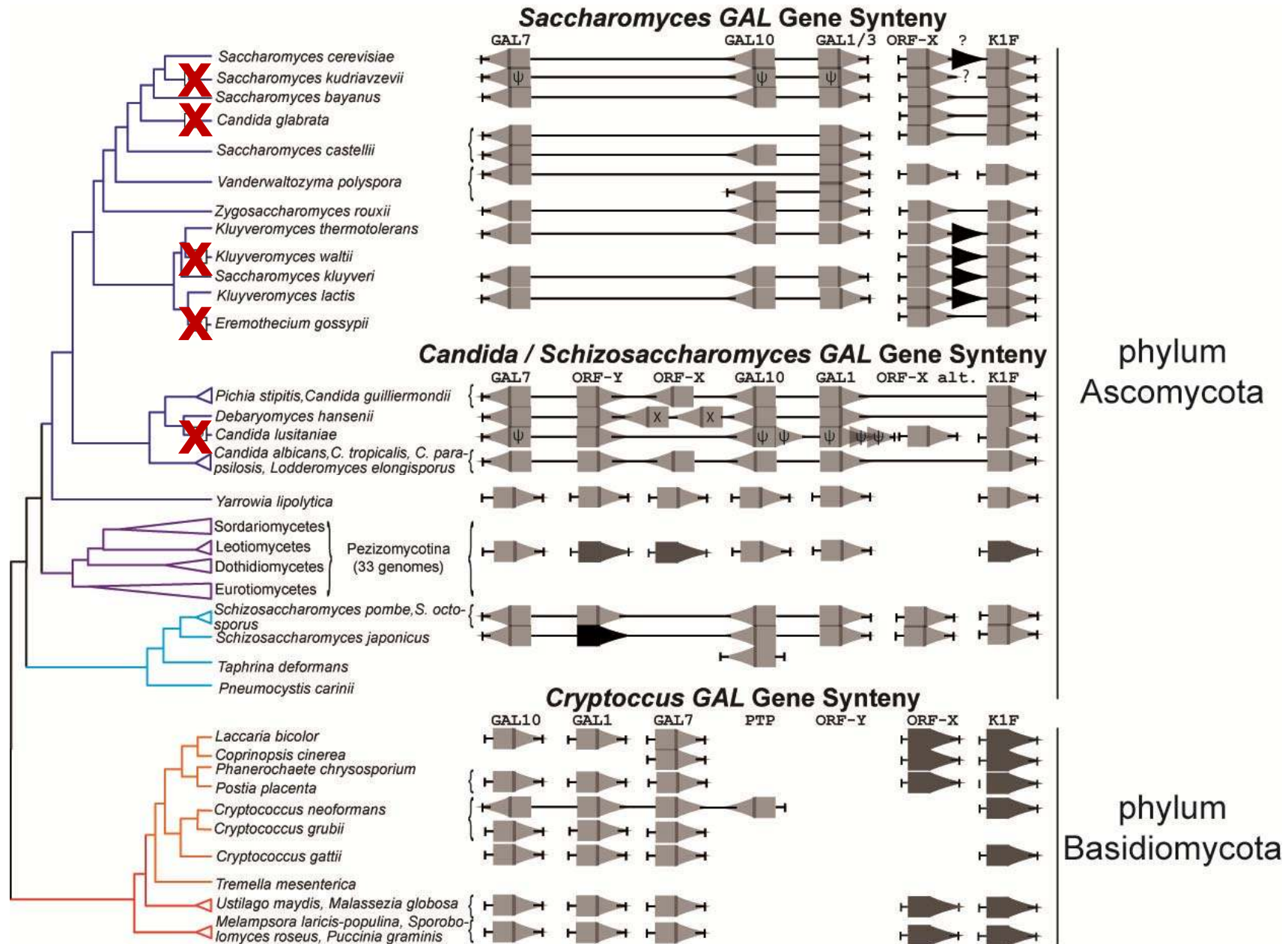
Slot & Rokas (2010) PNAS

Why Cluster?

Organs Affected by Galactosemia



What are the Evolutionary Effects of Pathway Clustering?



Clustered Pathways Are Easier Lost

Two potentially functional GAL states:

clustered – C

unclustered – U

One nonfunctional GAL state:

nonfunctional – N

Rate Restriction	LnL	2 Δ L	DF	p-value	Mean Transition rate
none	-28.82				qCN = 30.9 qUN = 1.0 qNC = 77.8 qNU = 2.2
qNC = qNU	-29.64	1.62	1	0.25	qNC = qNU = 12.9
qCN = qUN	-32.95	8.24	2	0.02	qCN = qUN = 2.4

Species with Clusters Have Higher Rates of GAL Pathway Loss



Analyses done using Mark Pagel's BayesTraits program

Lecture Outline

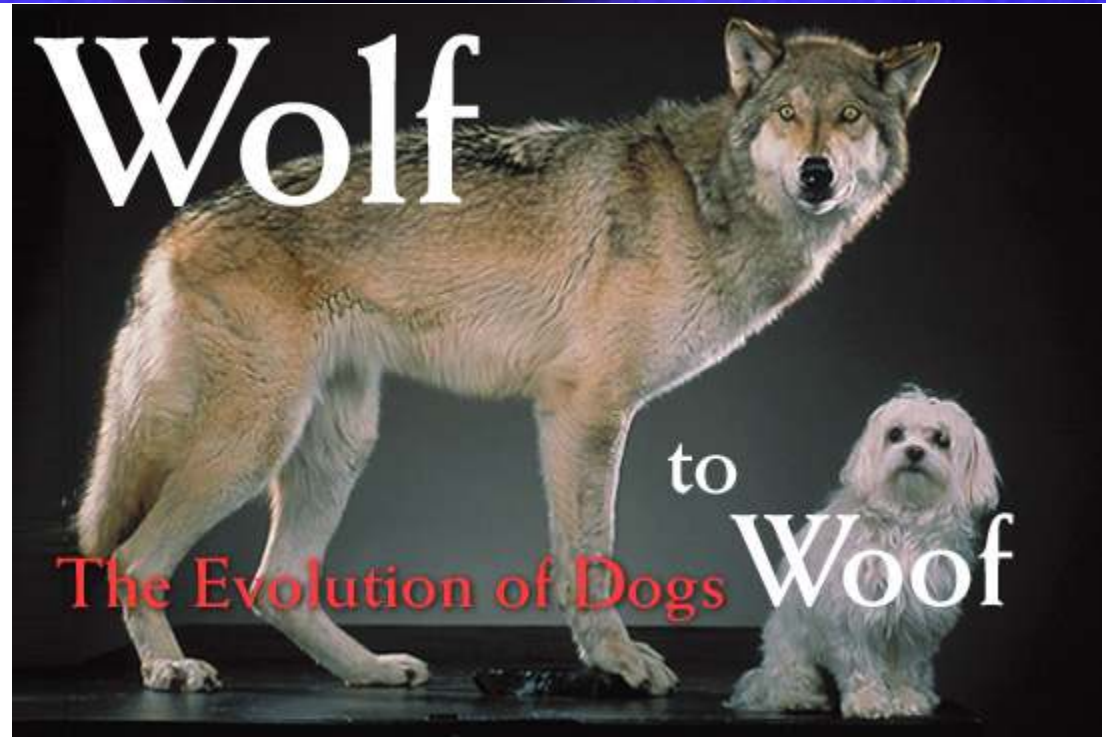
❖ **Introduction to Genome Evolution**

❖ **Comparative Functional Genomics**

----- **Coffee Break** -----

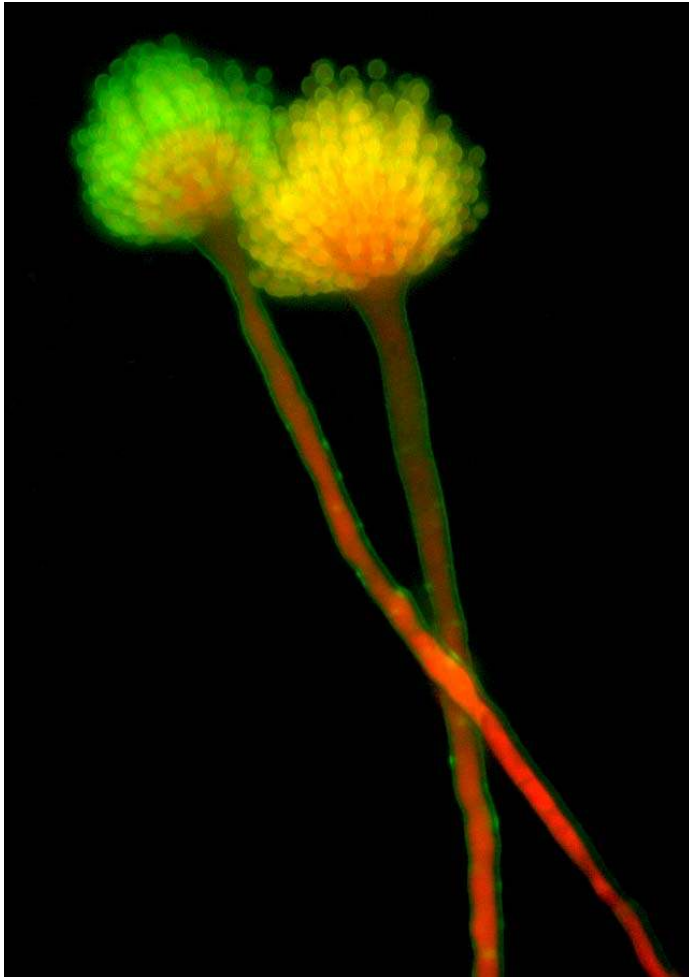
❖ **Population Genomics**

❖ **Phylogenomics**





***Aspergillus oryzae*: Cornerstone of Several Japanese Tasty Liquids**



Archaeological evidence suggests that mixed fermented alcoholic beverage of rice, honey and fruit was made in China as early as 7 – 9 millennia ago

***Aspergillus oryzae*, a filamentous fungus, is involved in the production of sake (rice wine), *miso* (soy bean paste), *su* (vinegar) and *shoyu* (soy sauce)**



How Sake is Made

1. Polishing



2. Washing



3. Steaming



4. Koji making



5. Yeast starter



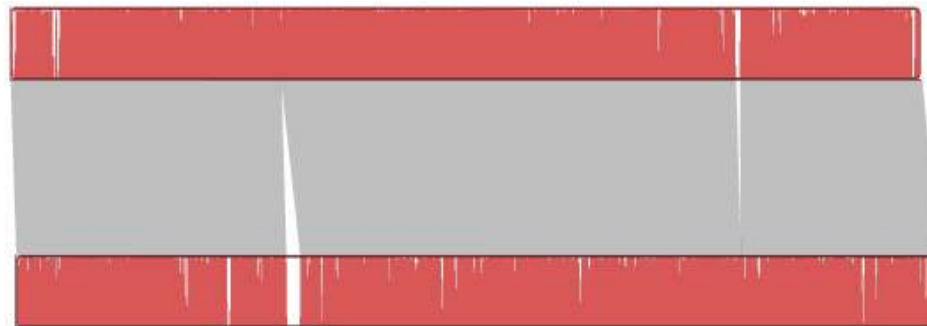
6. Pressing



http://www.sakejapan.com.au/whats_nihonshu.html#what

The *A. oryzae* and *A. flavus* Genomes are Nearly Identical

A. oryzae



A. flavus

6.5 Mb Chromosome 1

- ❖ 8 Chromosomes
- ❖ 37 Mb
- ❖ 12,000 genes
- ❖ 99.5% nt identity

A. oryzae

- ❖ Sake, soy sauce, miso
- ❖ Non-aflatoxin producer
- ❖ USDA GRAS species

A. flavus

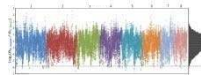
- ❖ Agricultural pest
- ❖ Aflatoxin producer
- ❖ ~\$1 billion annually



The Road to Domestication



Evolutionary Relationship?



Positively Selected Genes?



Functional Differences?

A. flavus



DOMESTICATION RD

ONE WAY

A. oryzae



Sequencing 16 Genomes

A. oryzae

SRRC 302	Sake
RIB 331	Miso
RIB 333	Miso
RIB 537	Sake
RIB 632	Sake
RIB 642	Sake
RIB 949	Soy Sauce
RIB 40	Sake, Reference Strain

A. flavus

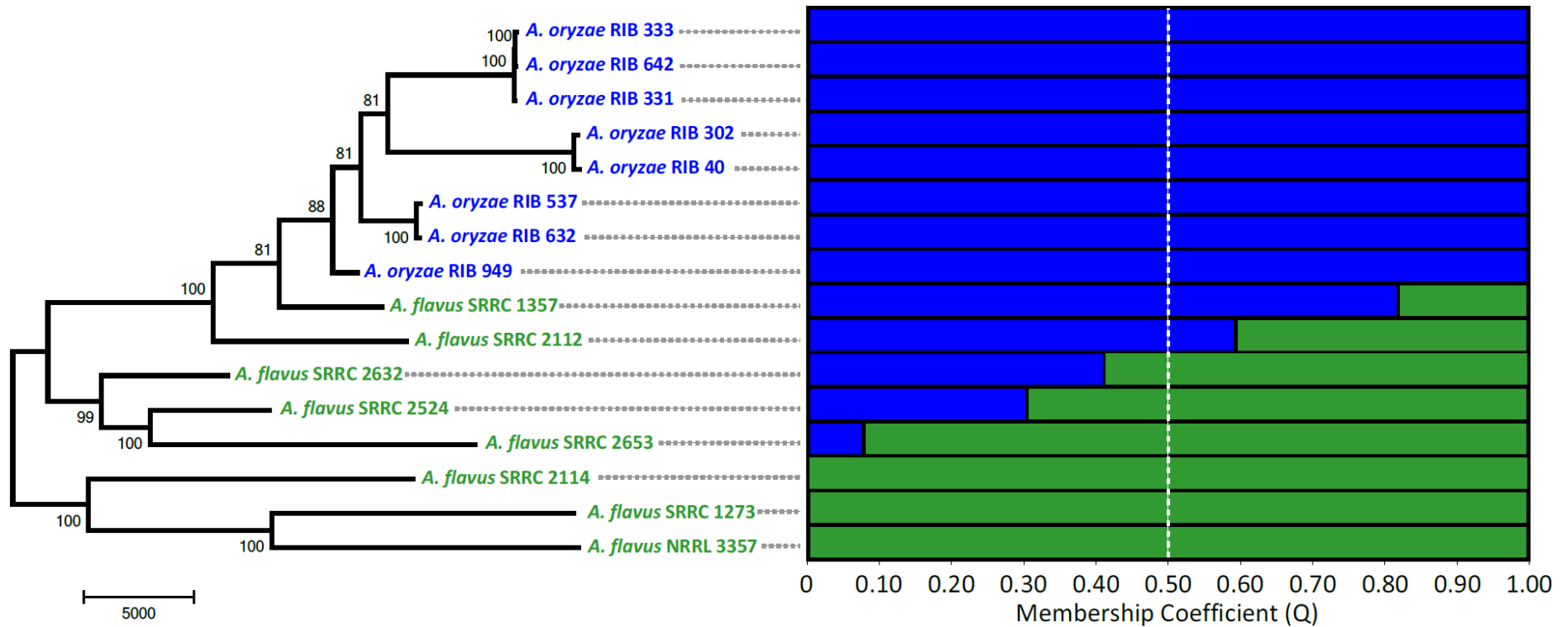
SRRC 1273	Soil, Arizona
SRRC 1357	Dried bacon, Croatia
SRRC 2112	Hazelnut, Turkey
SRRC 2114	Wheat, USA
SRRC 2524	Dead termites, China
SRRC 2632	Blood, Chicago, Illinois
SRRC 2653	Corneal ulcer, Miami, Florida
NRRL 3357	Peanut, Reference strain, USA

❖ Illumina Sequencing

❖ 12-30 million 80bp reads

❖ > 20x coverage

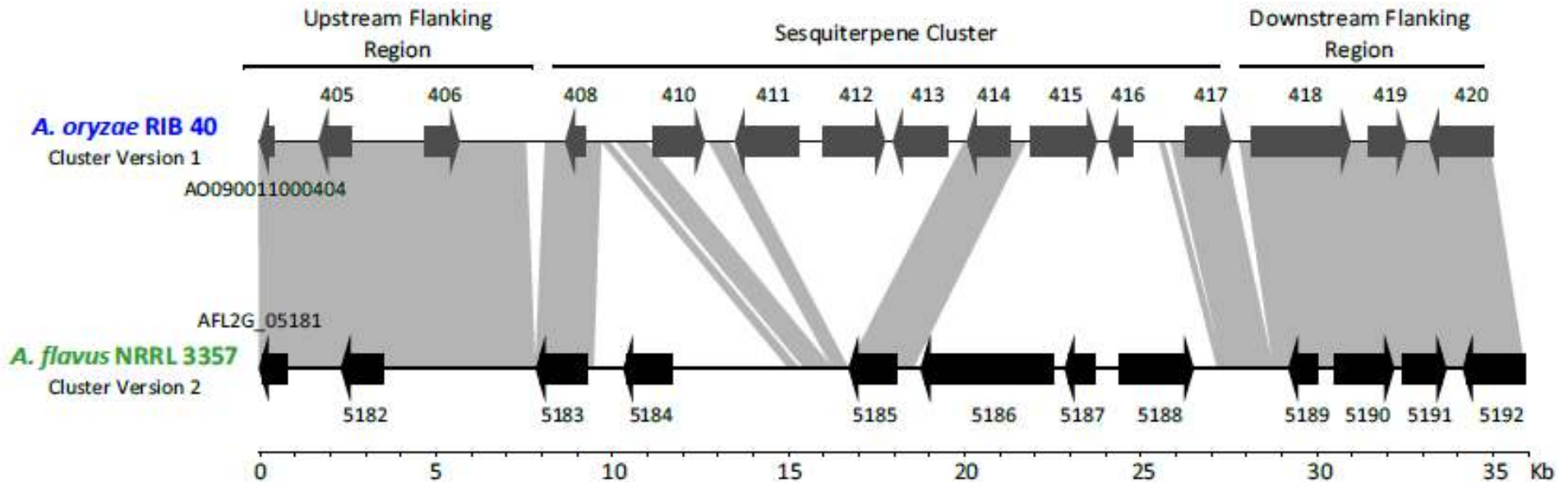
Aspergillus oryzae Isolates are Genetically Distinct



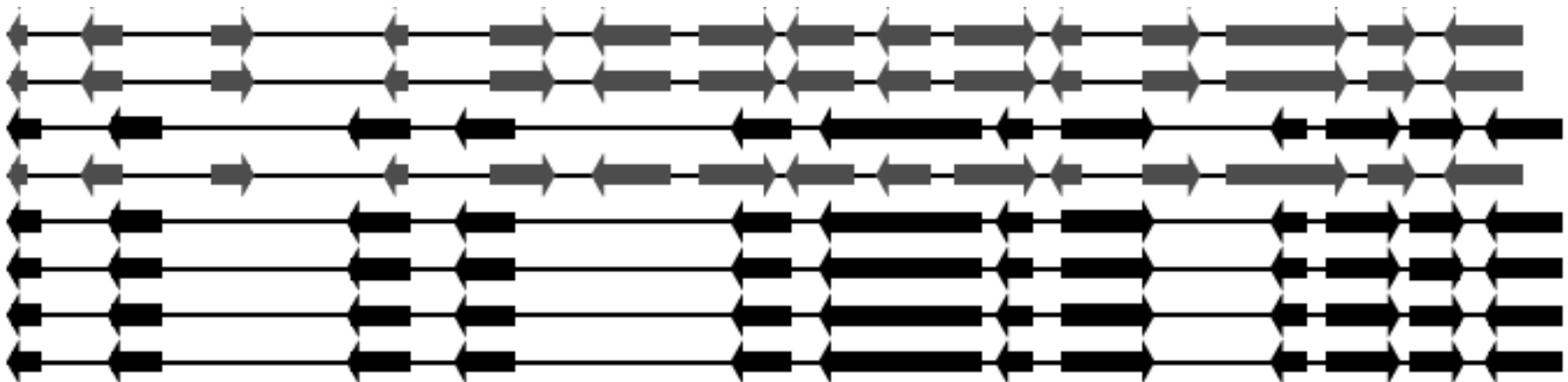
100,084 SNPs



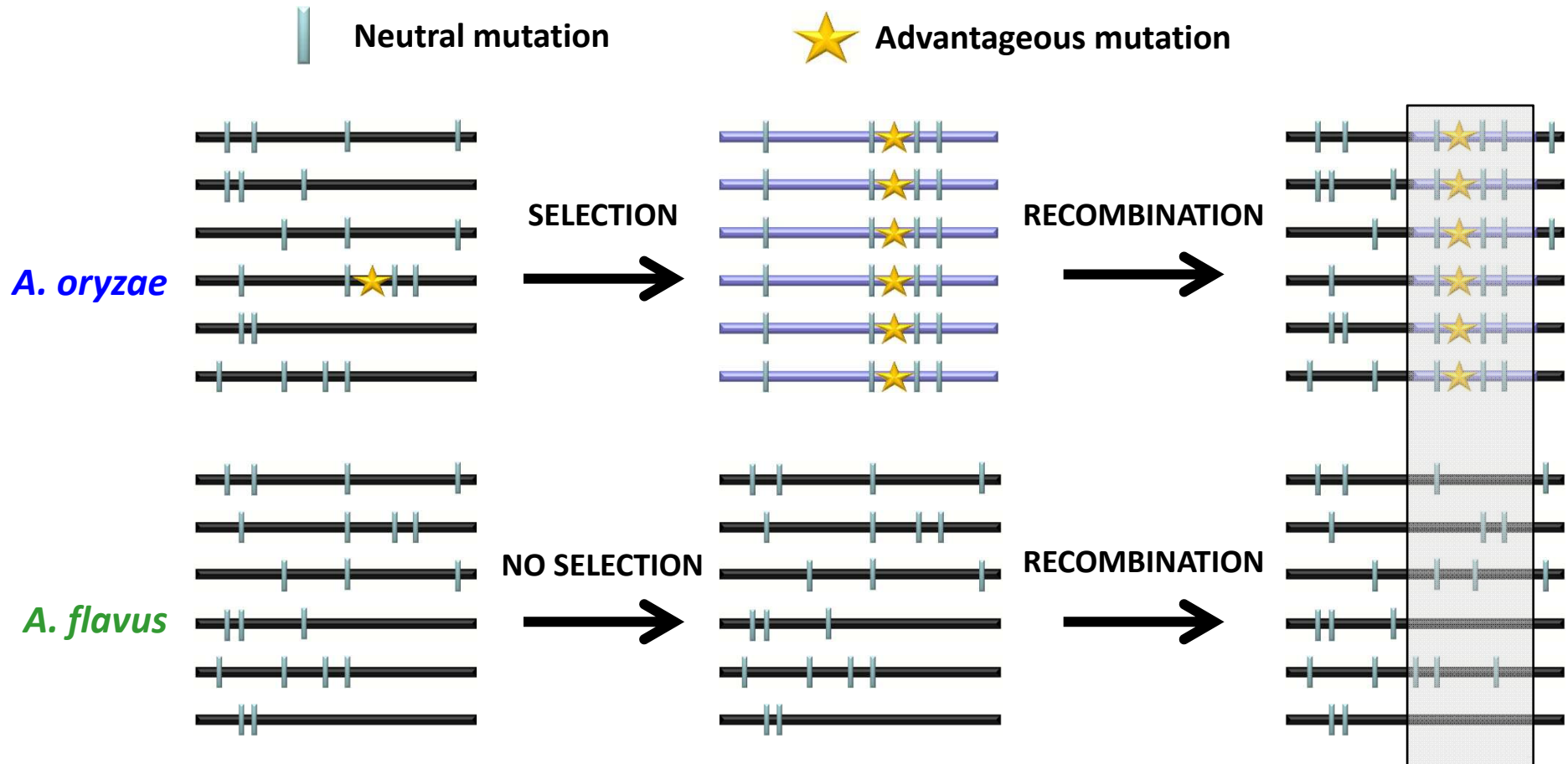
Non-homologous Alleles



A. flavus



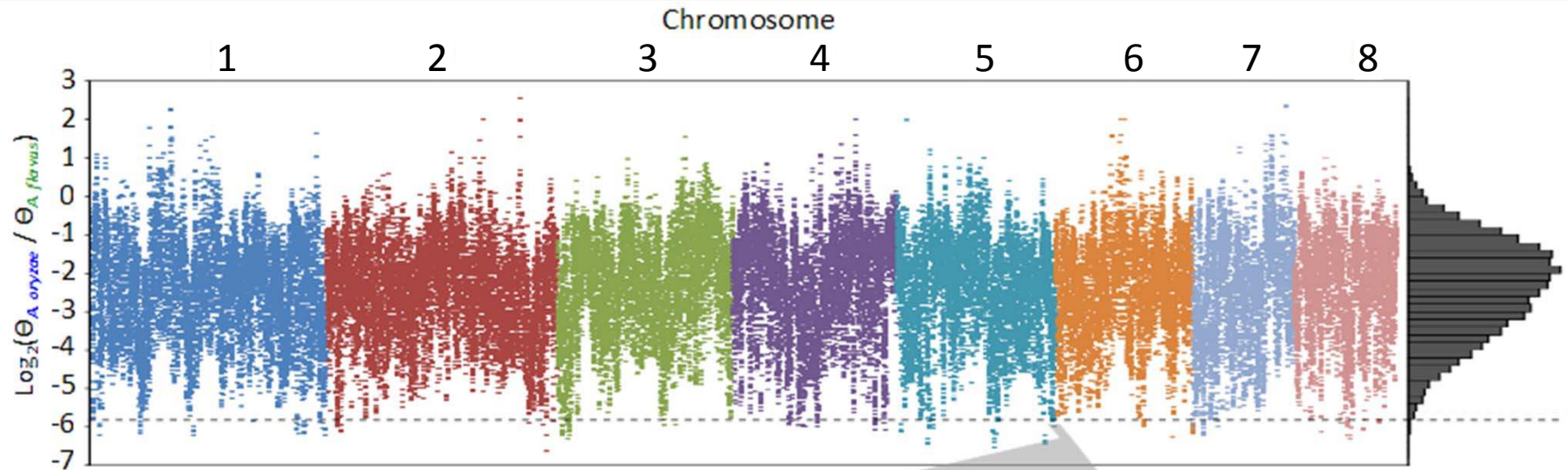
Detecting Recent Positive Selection in *A. oryzae*



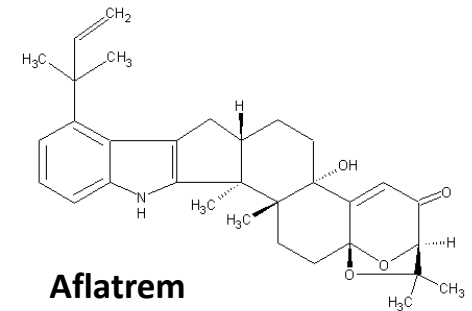
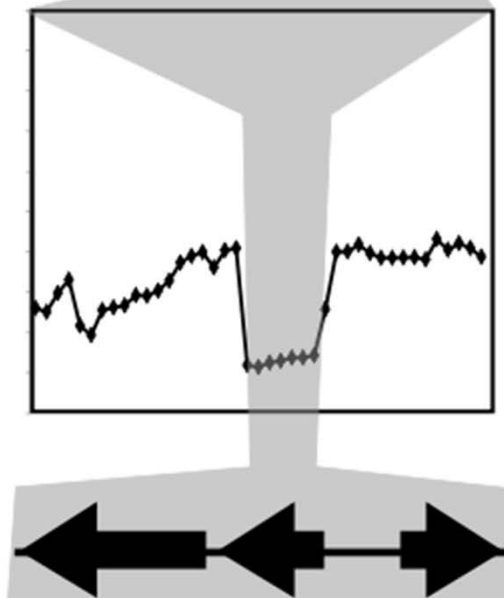
- ❖ Relative nucleotide diversity ($\Theta_{A.oryzae} / \Theta_{A.flavus}$)
- ❖ 5kb sliding windows, 500bp steps; >65,000 windows
- ❖ Overlapping regions of windows in the lowest 0.25%



Recent Positive Selection in the *A. oryzae* Genome



65,894 total windows
↓
164 candidate windows
↓
61 Putative Selective
Sweep Regions (PSSRs)
↓
148 Genes



**Genes in Sweep Regions
are Enriched in Secondary
Metabolism ($P < 0.0006$)**



The Complex Flavor of Sake

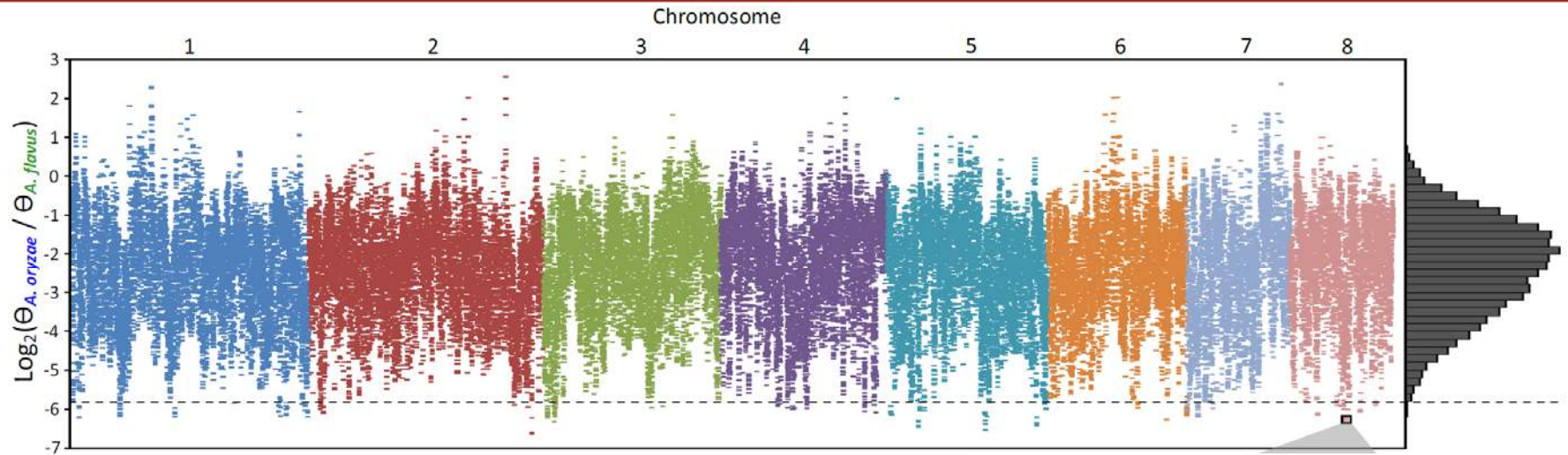


- ❖ Fragrance
- ❖ Impact
- ❖ Sweet/Dry
- ❖ Acidity
- ❖ Presence
- ❖ Complexity
- ❖ Earthiness
- ❖ Tail

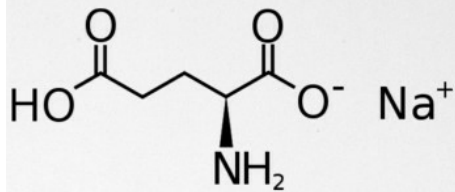
Have “flavor” genes been under selection?



Differences in a Flavor Associated Gene



MONOSODIUM GLUTAMATE



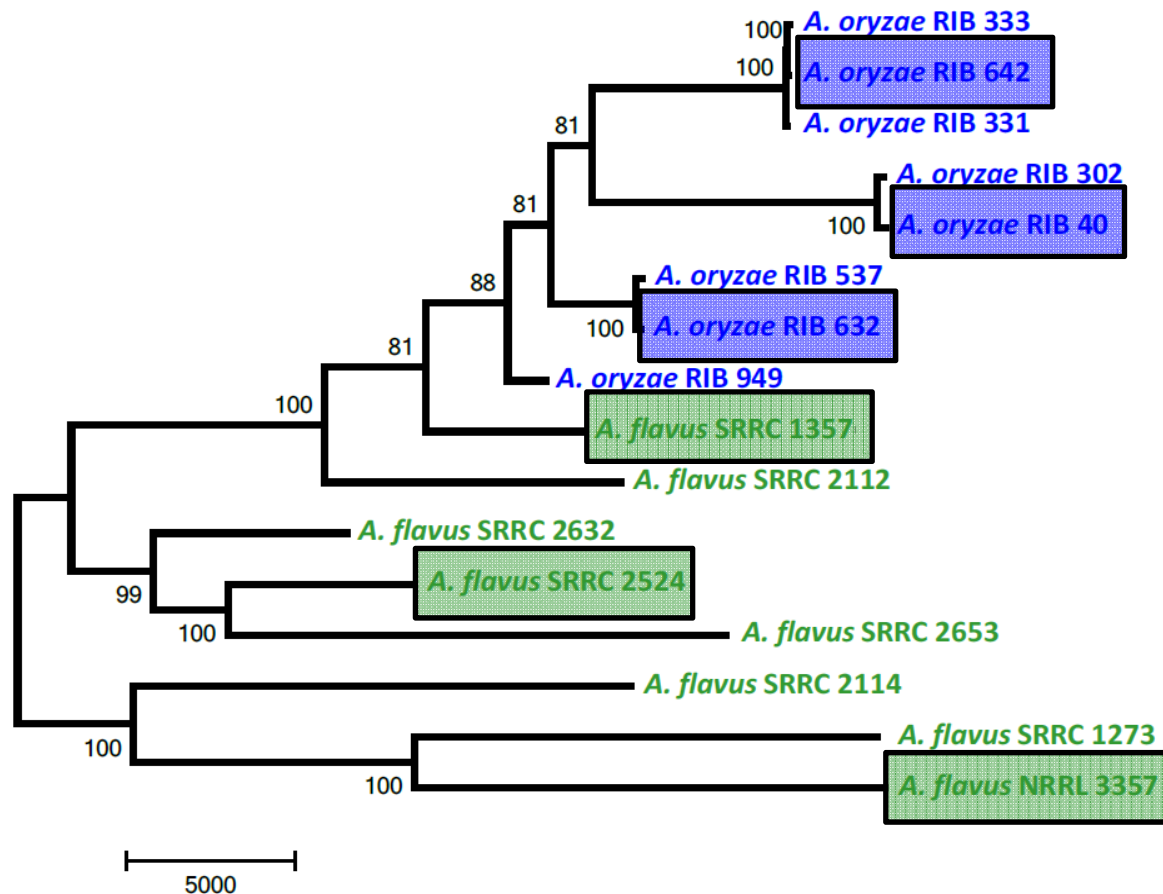
SODIUM SALT OF GLUTAMIC ACID



Glutamic Acid



Detecting Differences in Transcript and Protein Abundance

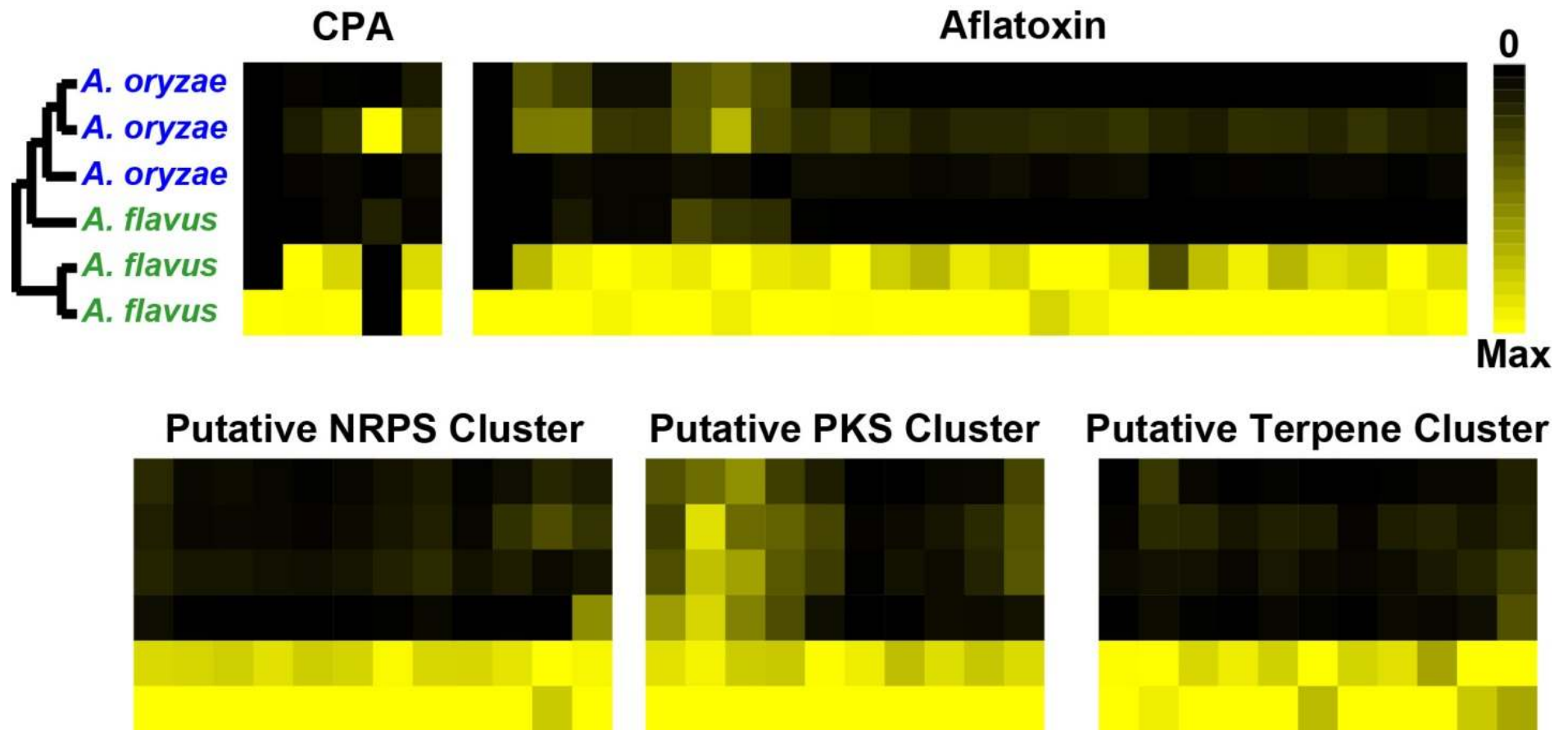


- ❖ Gene Expression (RNA-Seq)
- ❖ Protein Abundance (MudPIT)

- ❖ Sake strains
- ❖ 30° C
- ❖ 24 hours

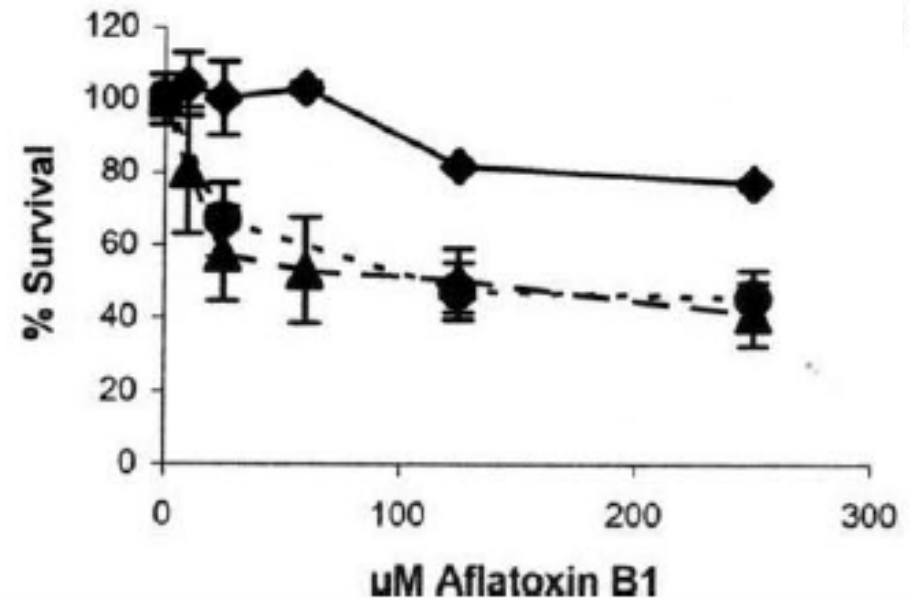
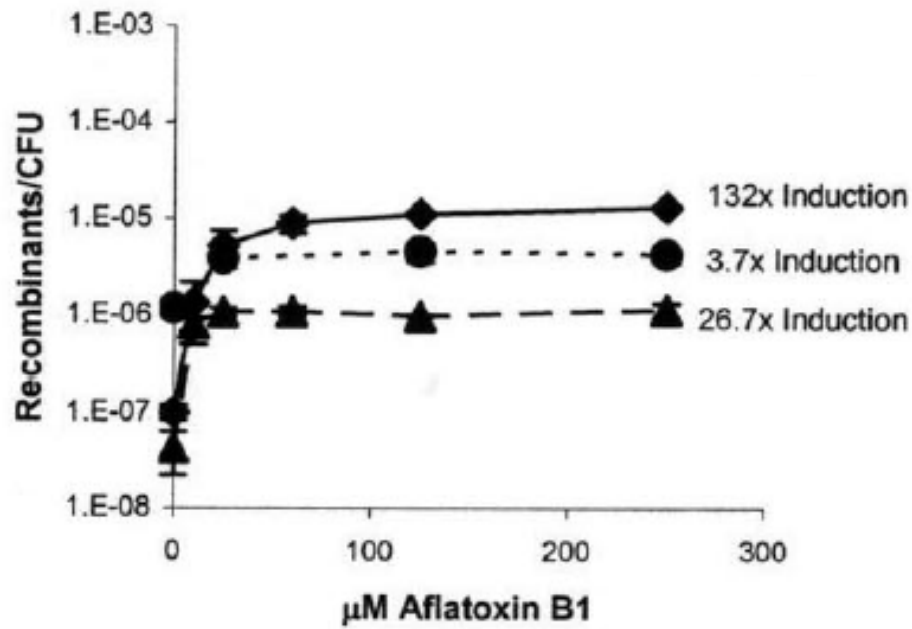


Down-Regulation of Secondary Metabolism in *A. oryzae*



Why is A. oryzae Atoxic?

Aflatoxin is genotoxic to *S. cerevisiae*



The atoxicity of *A. oryzae* might have been driven by its impact on yeast survival and, as a consequence, fermentation for making sake



A. oryzae Breaks Down Starch into Sugar

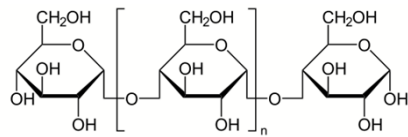
Rice Preparation



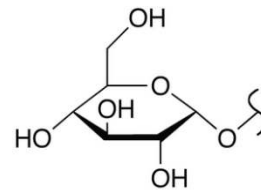
Saccharification



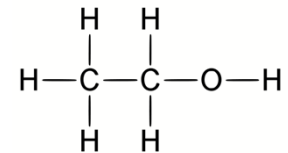
Fermentation



A. oryzae



Yeast



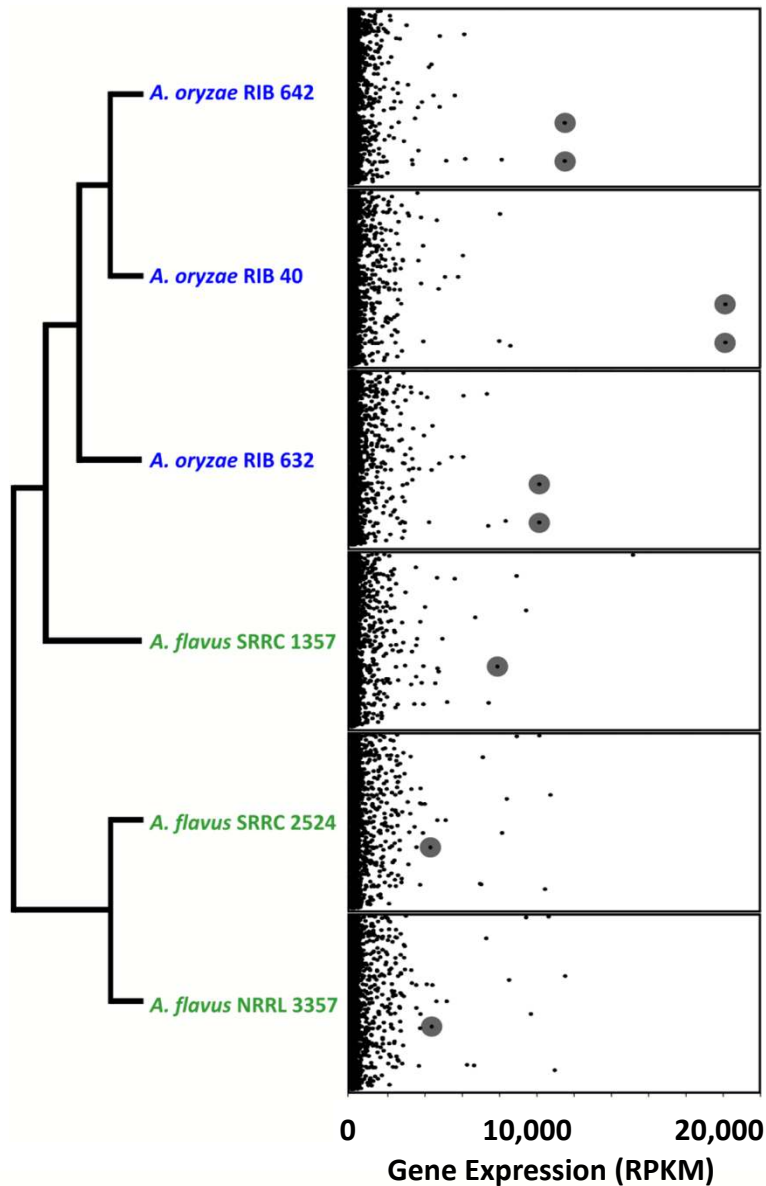
Starch

Sugar

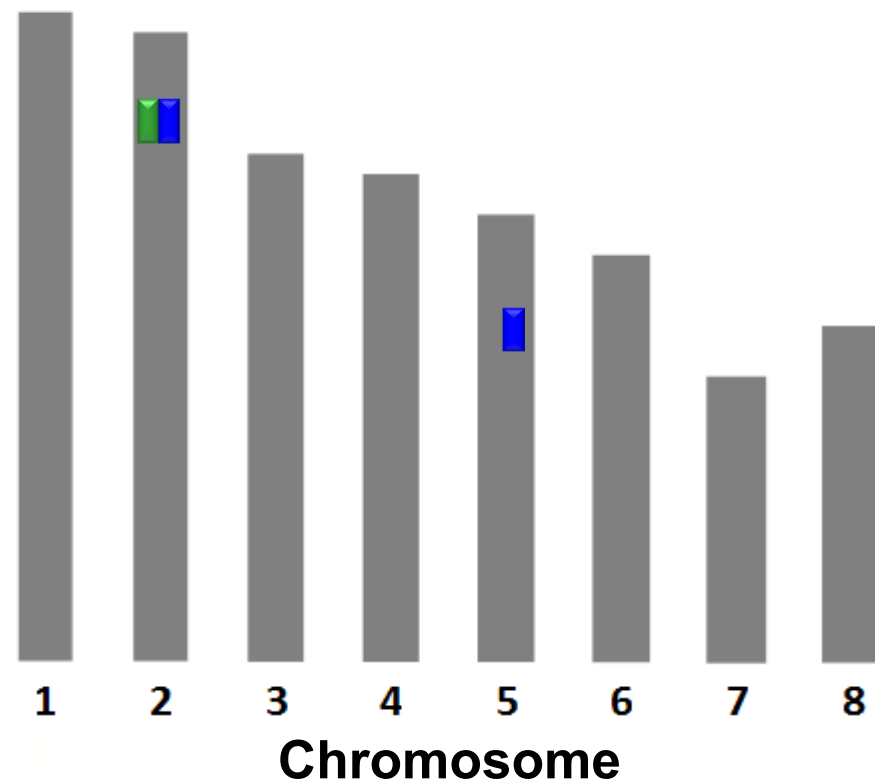
Alcohol



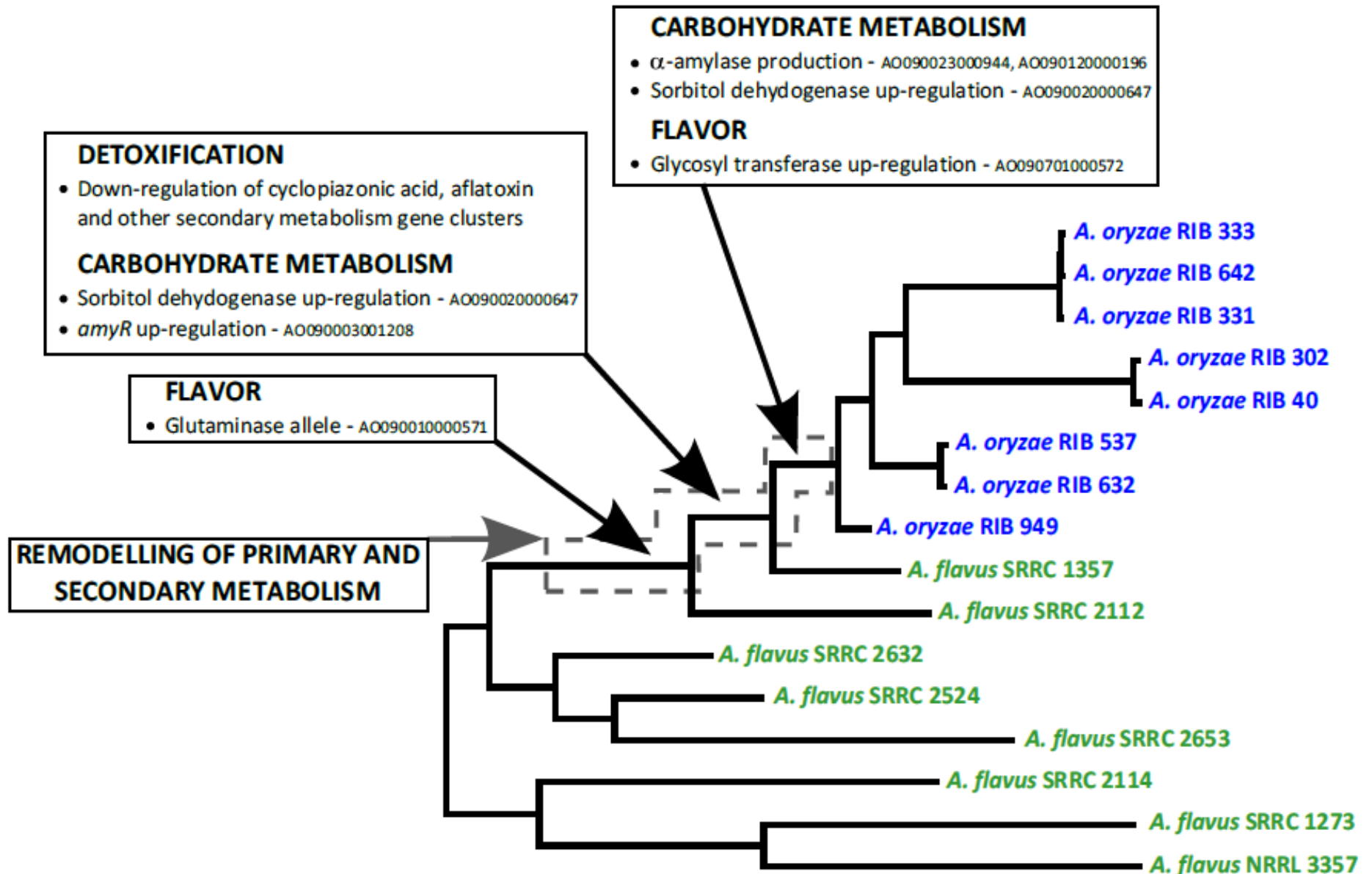
α -amylase is Highly Abundant



- ❖ Gene Expression $\uparrow P = 1e-300$
- ❖ Protein Enrichment $\uparrow P = 8e-63$
- ❖ Carbohydrate Metabolism $\uparrow P = 6e-5$



DOMESTICATION RD

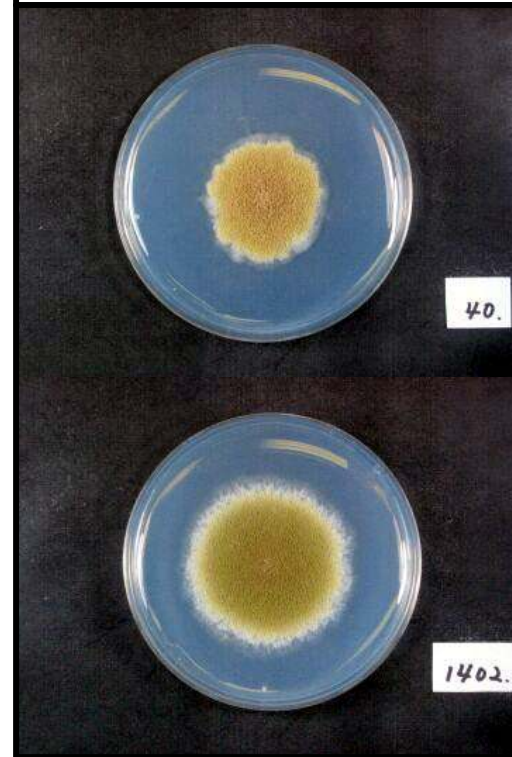


Contrasting Domestication Patterns Across Kingdoms

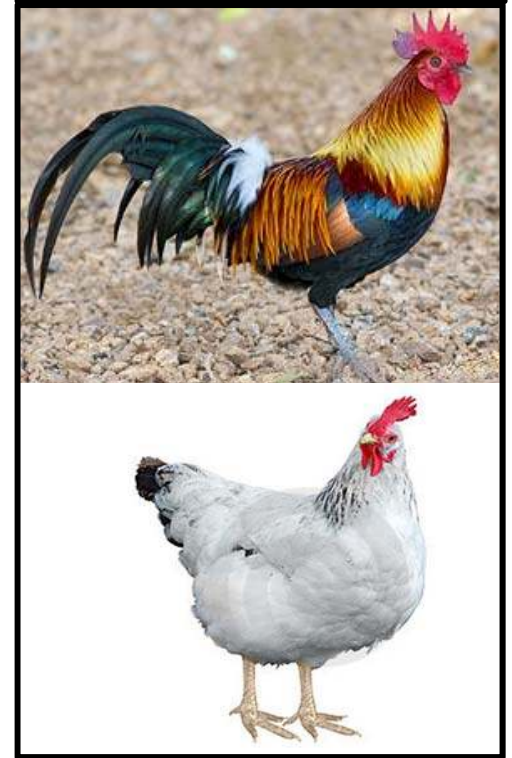
Plants



Microbes



Animals



Lecture Outline

❖ **Introduction to Genome Evolution**

❖ **Comparative Functional Genomics**

----- **Coffee Break** -----

❖ **Population Genomics**

❖ **Phylogenomics**

The Dawkins Delusion

“... there is, after all, one true tree of life [...]. It exists. It is in principle knowable. We don’t know it all yet. By 2050 we should – or if we do not, we shall have been defeated only at the terminal twigs, by the sheer number of species.”

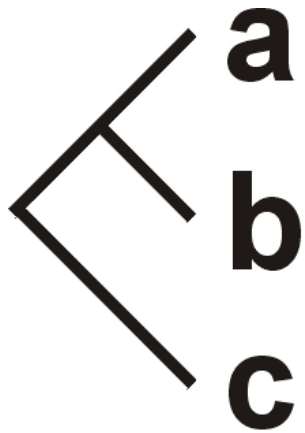
Richard Dawkins



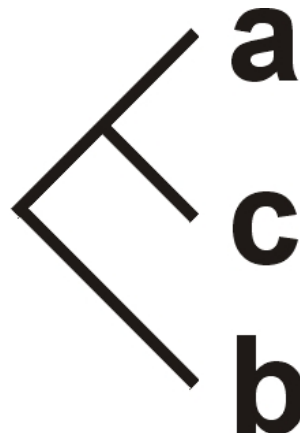
Dawkins (2003), A Devil’s Chaplain

The Problem of Incongruence

Gene X

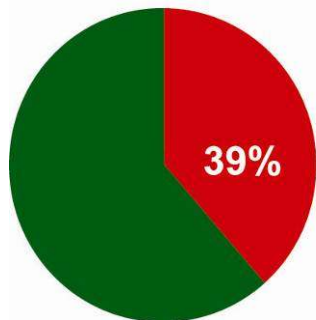


Gene Y

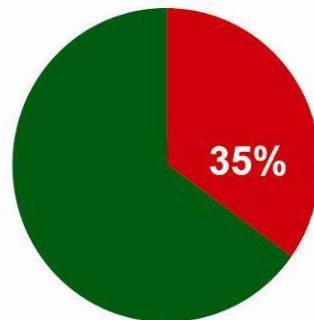


Species tree?

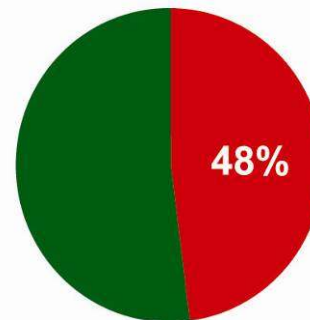
A: All organisms



B: Mammals



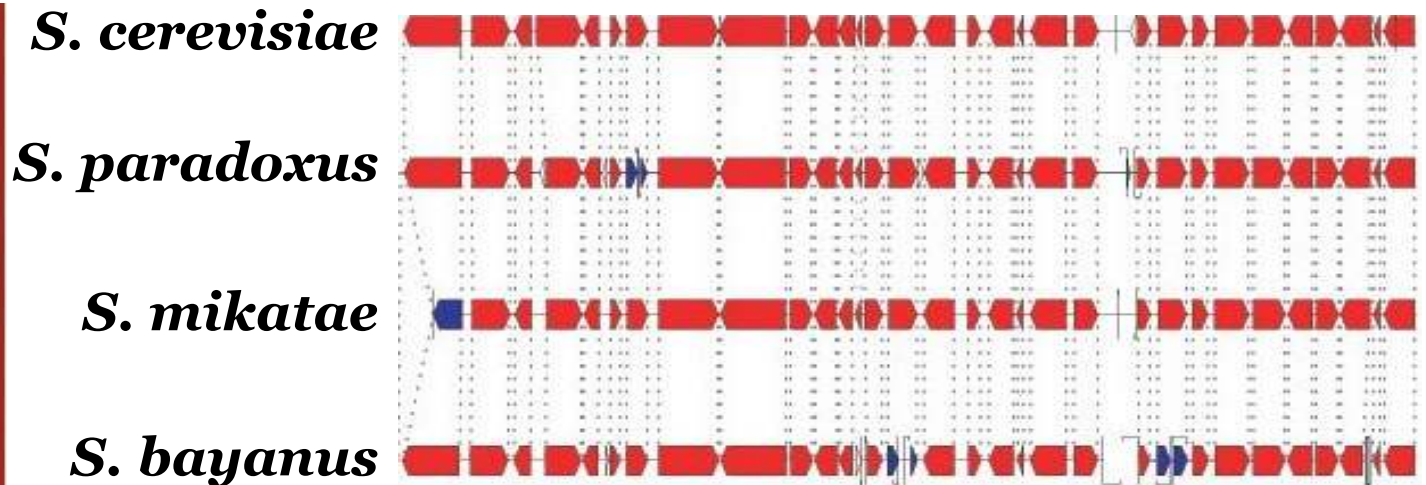
C: Insects



Incongruence is
pervasive in the
phylogenetics
literature



A Systematic Evaluation of Single Gene Phylogenies

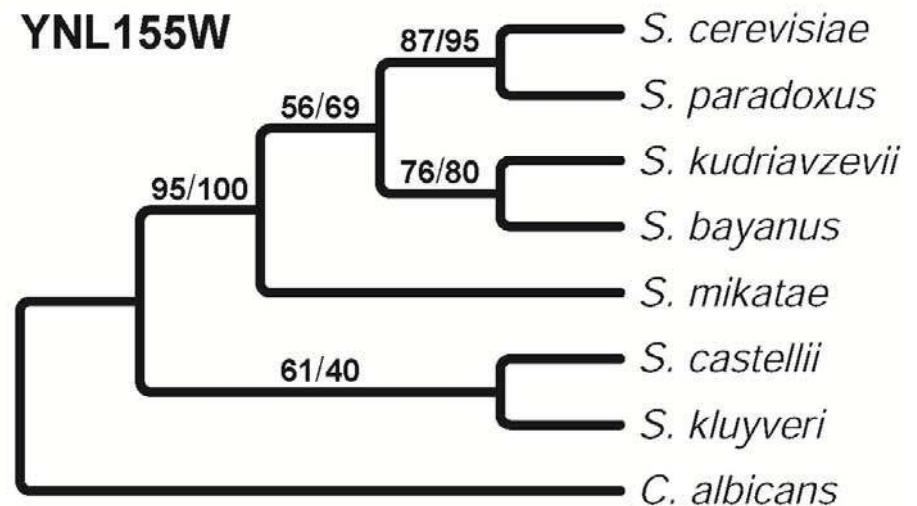
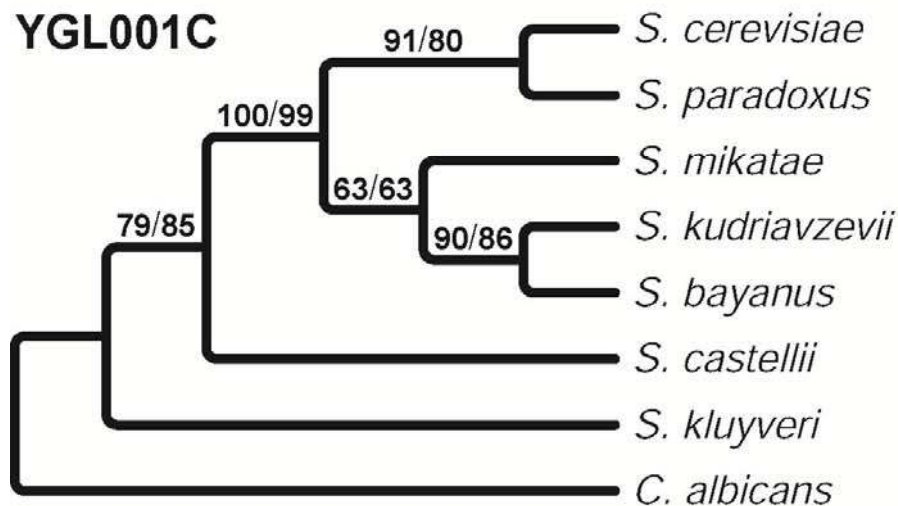
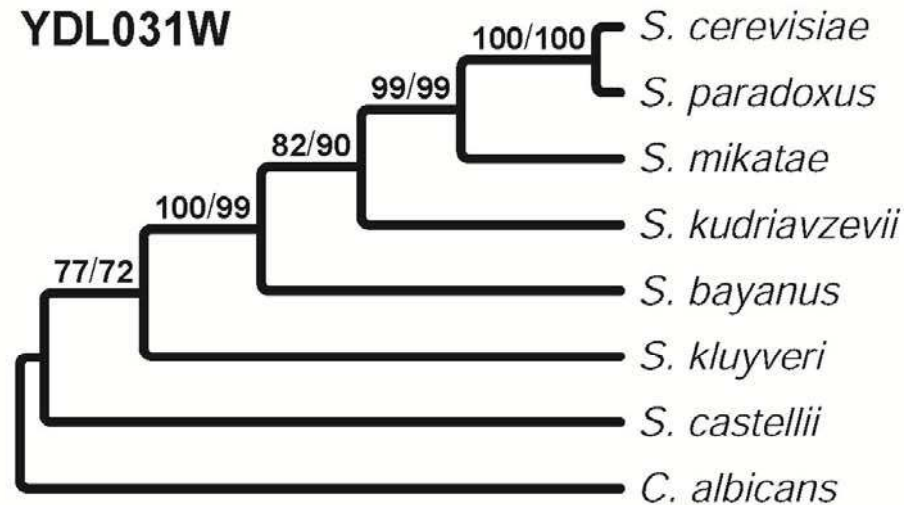
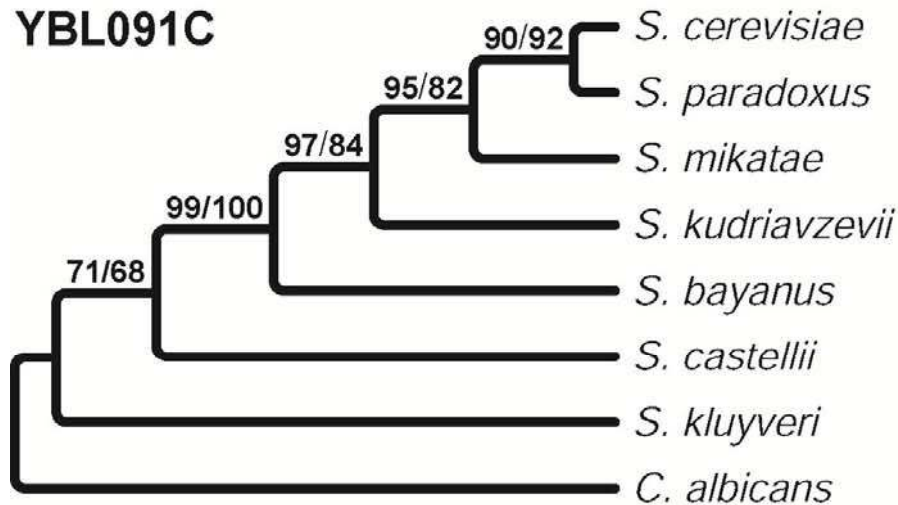


Dataset: 106 genes on all 16 chromosomes totaling 127kb corresponding roughly to 1% of the genomic sequence, 2% of genes

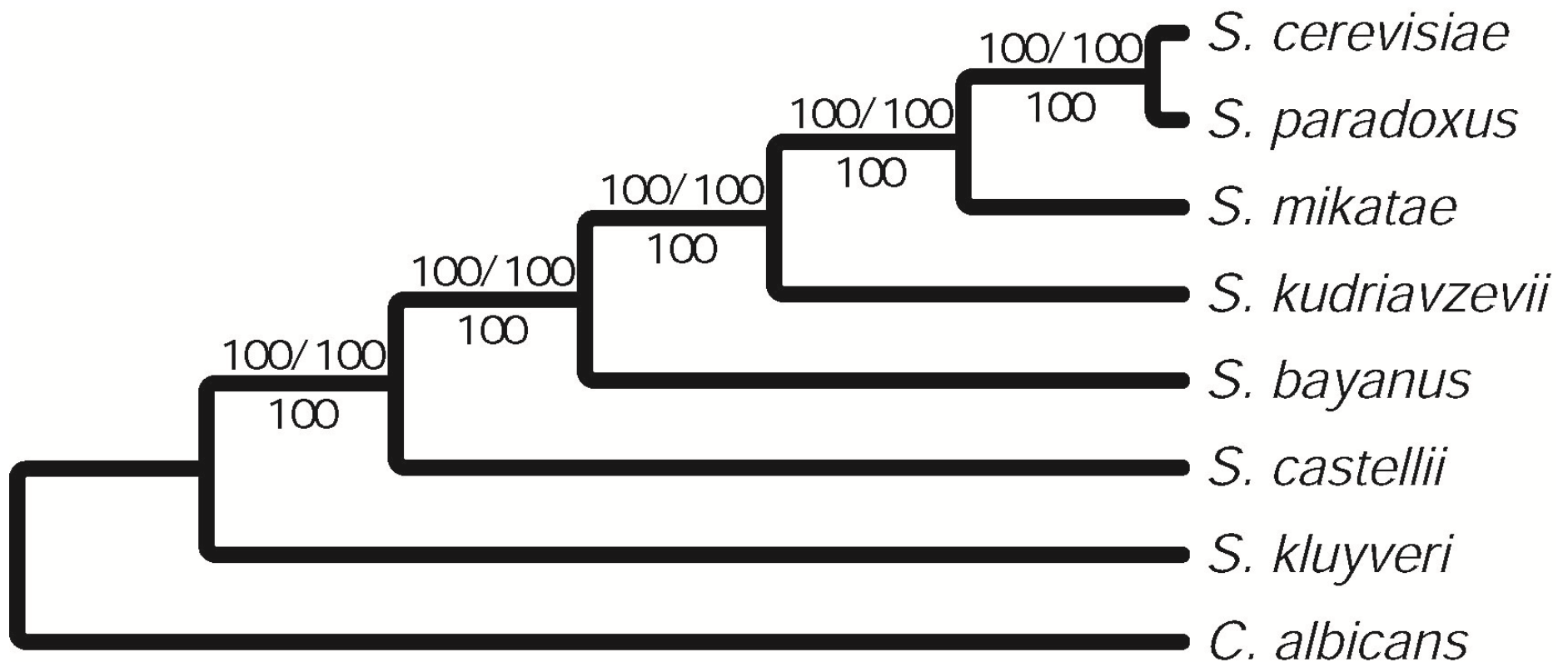
Analyses: Maximum Likelihood (ML) & Maximum Parsimony (MP) on nt data sets and MP on amino acid data sets



Incongruence at the Single Gene Level



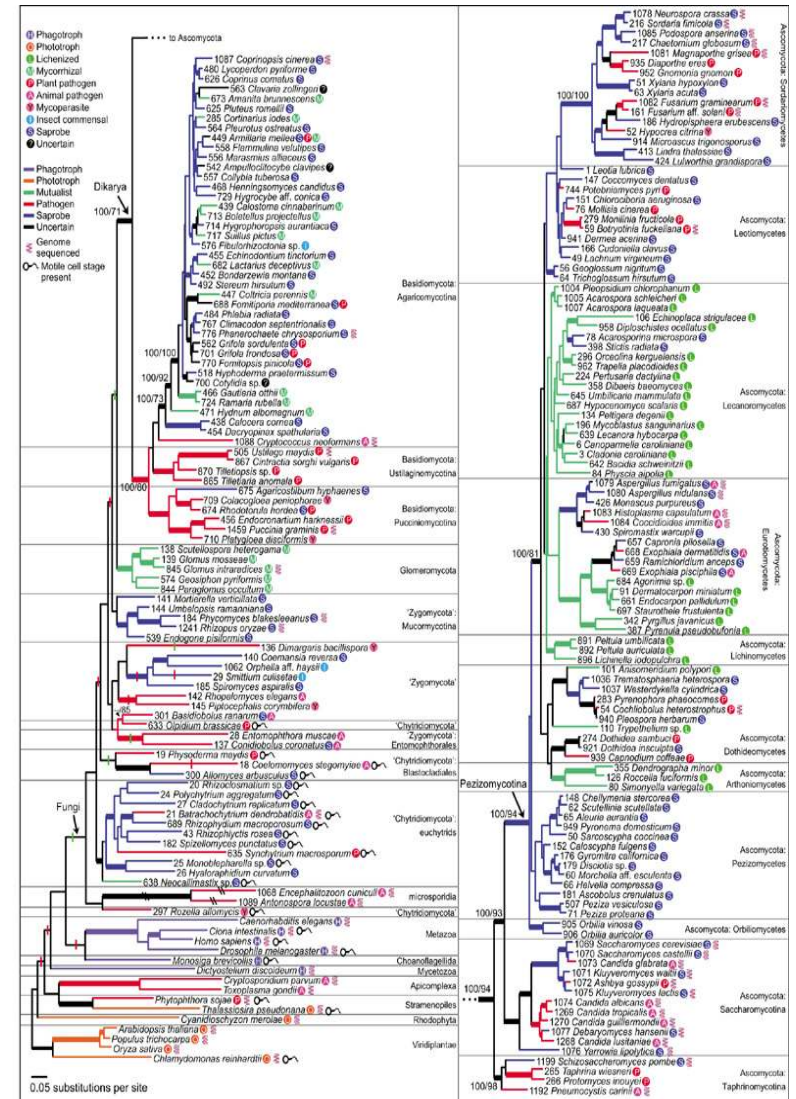
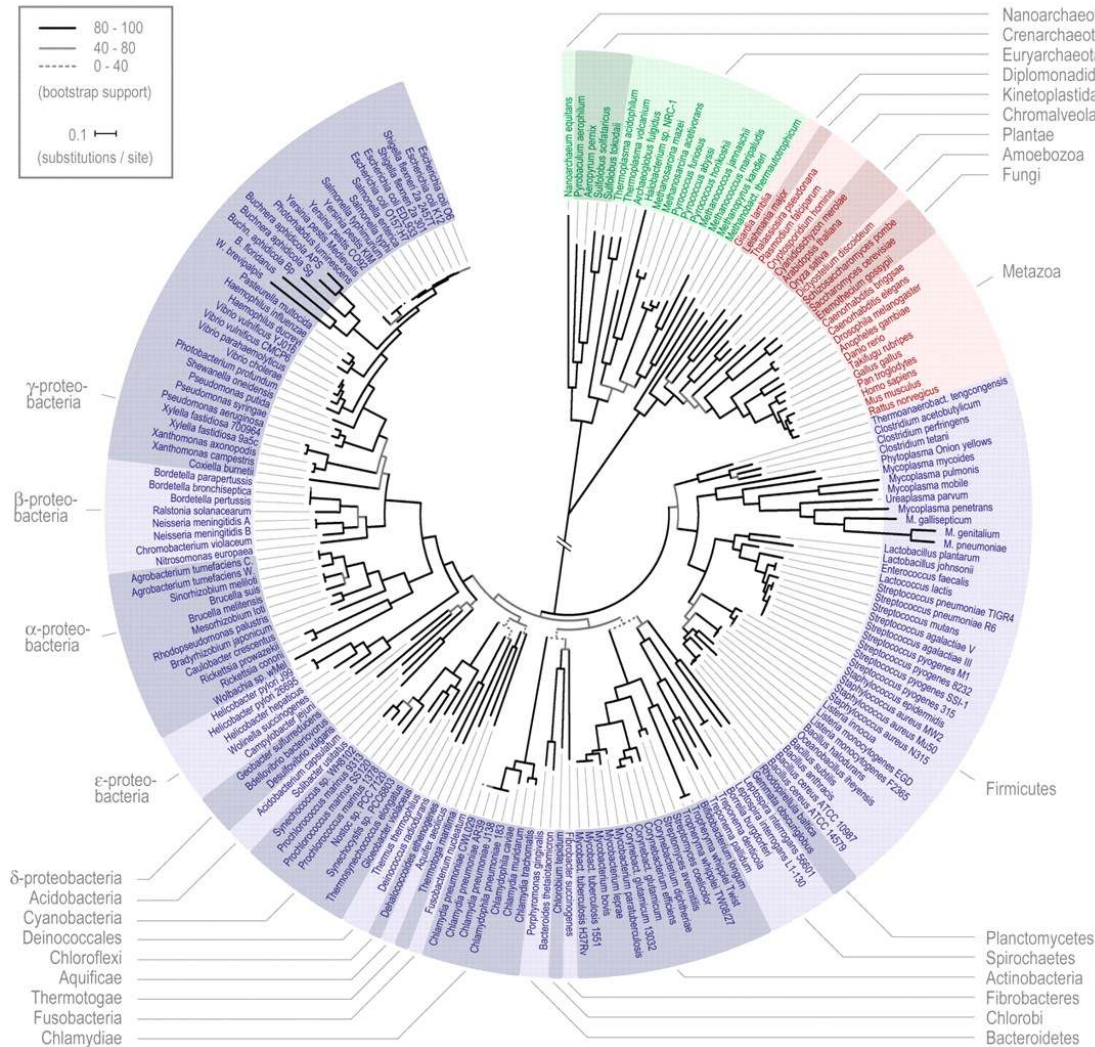
Concatenation of 106 Genes Yields a Single Yeast Phylogeny



Clade support values = $\frac{\text{ML / MP on nucleotide alignments}}{\text{MP on amino acid alignments}}$



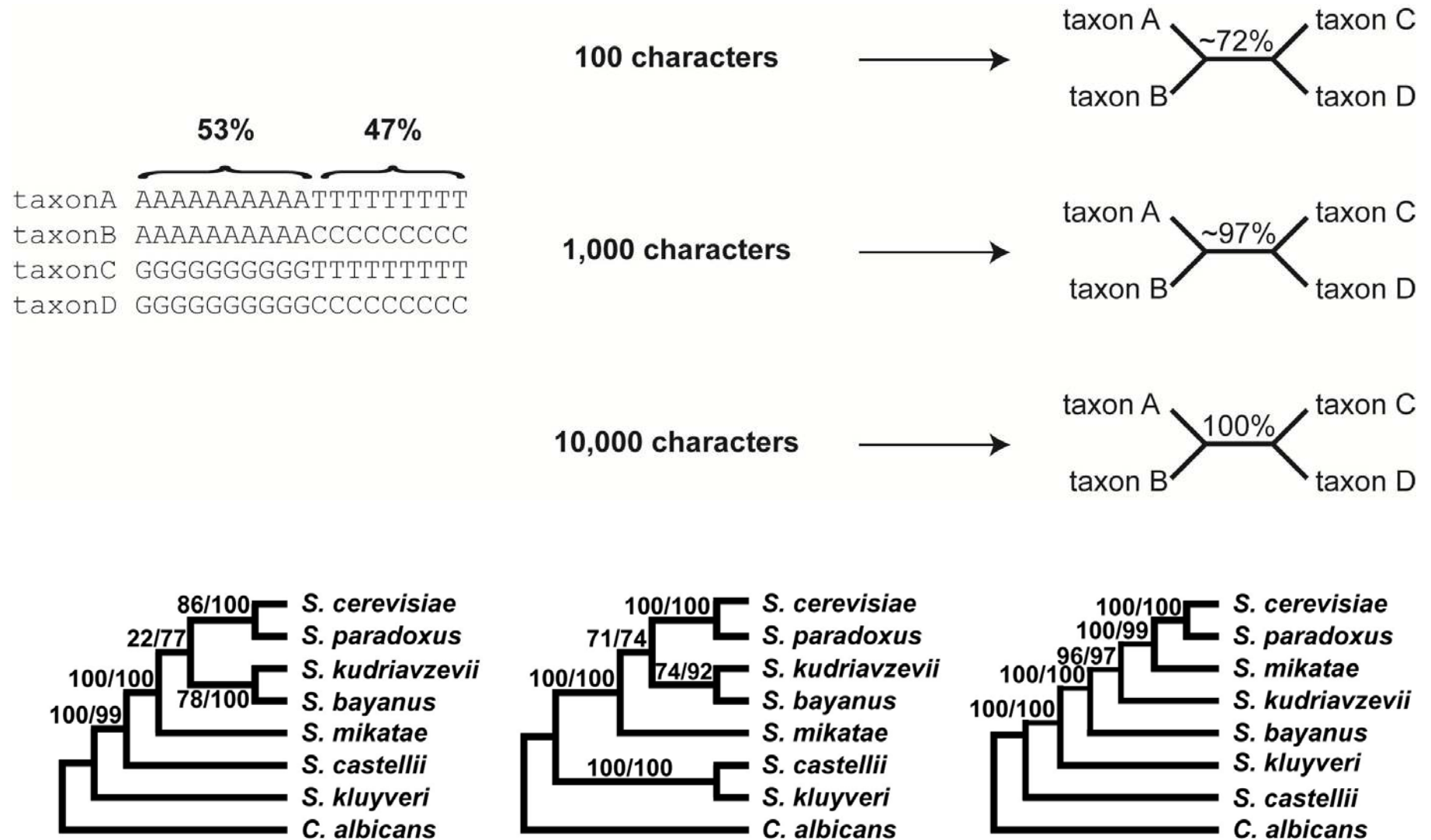
Reconstructing the Tree of Life



Ciccarelli et al. (2006) Science

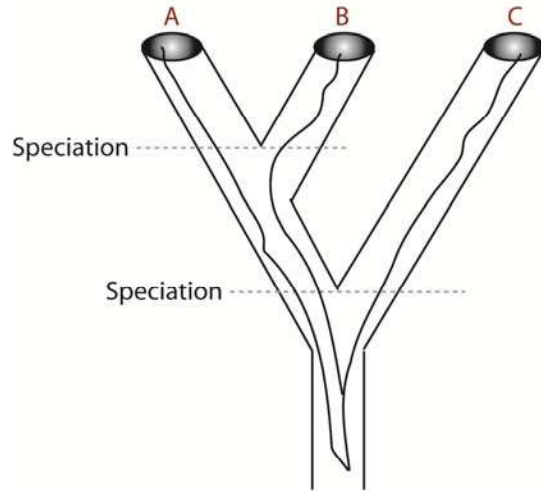
James et al. (2006) Nature

Methods of Assessing Robustness are Subject to Systematic Error

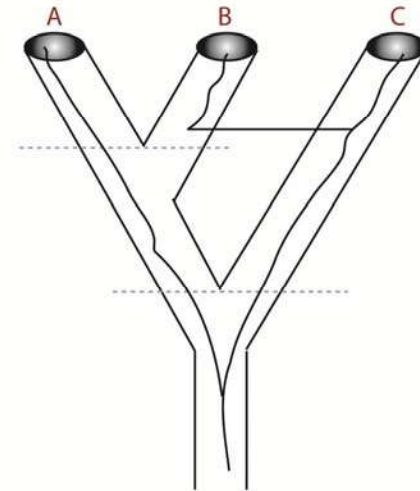


Gene Trees Can Differ from Species Trees

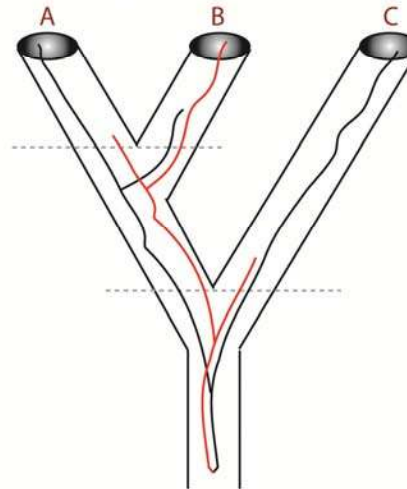
Lineage Sorting



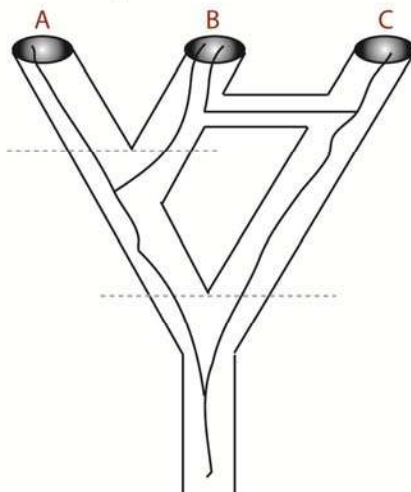
Horizontal Gene Transfer



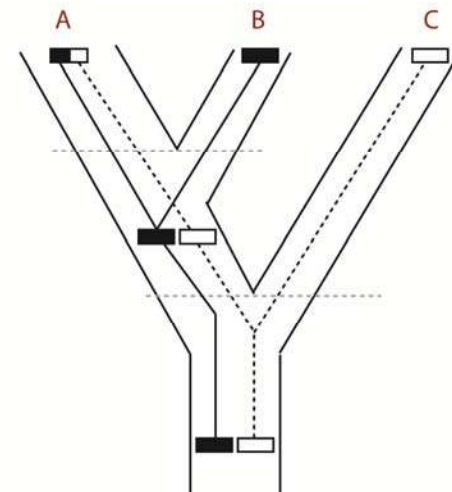
Gene Duplication and Loss



Hybridization

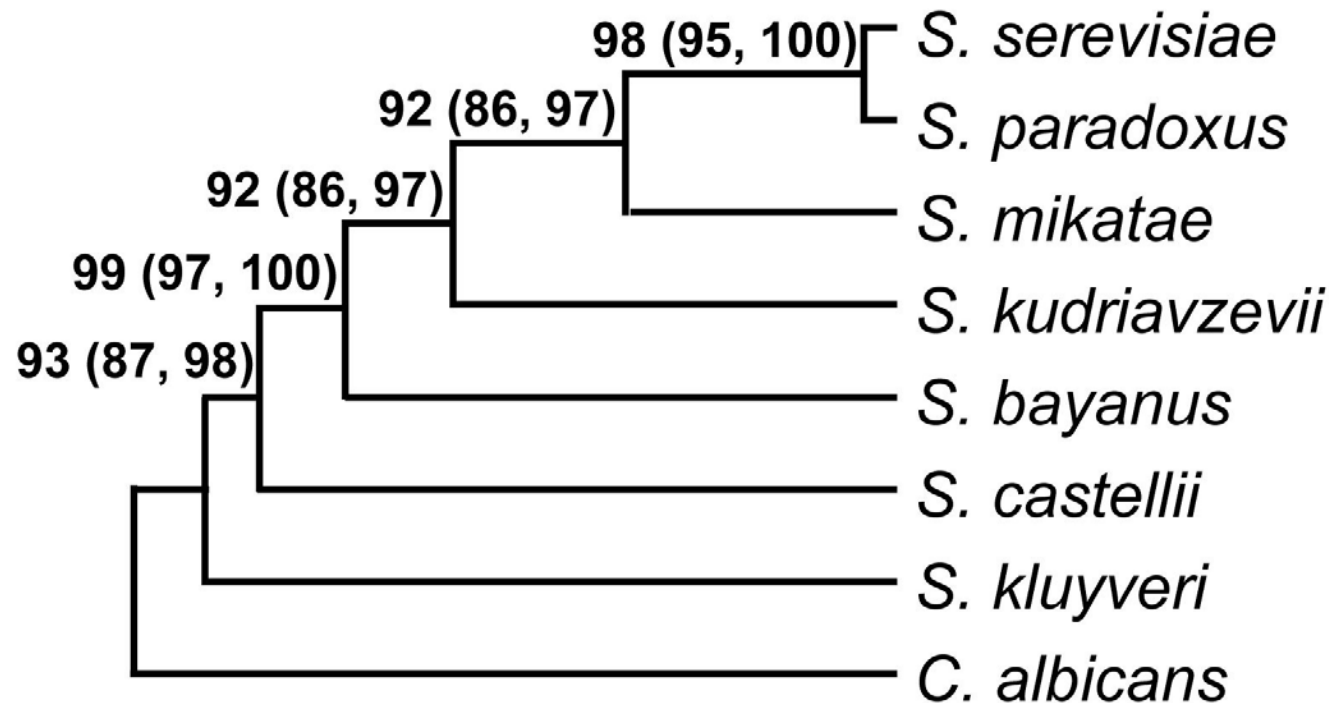


Recombination



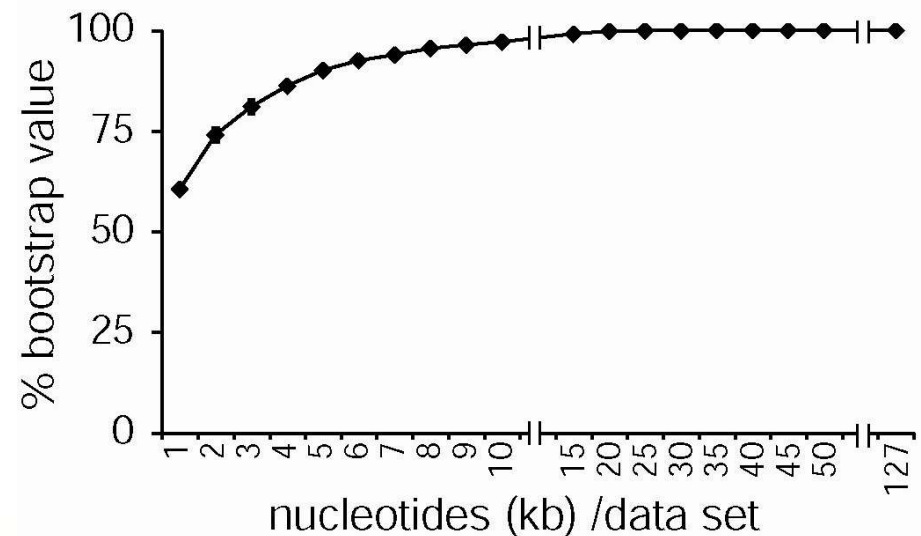
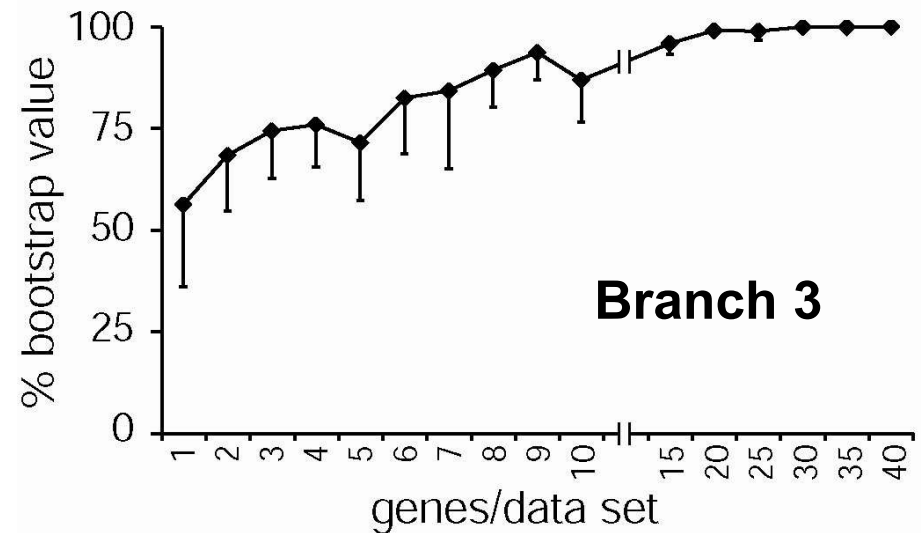
Inferring the Species Tree from Individual Gene Histories

Concordance Factor: The proportion of the genome for which a clade is true



How Much Data is Sufficient to Recover the Species Tree?

✓ Phylogenies based on single / few-genes are vulnerable to *sampling artifacts*





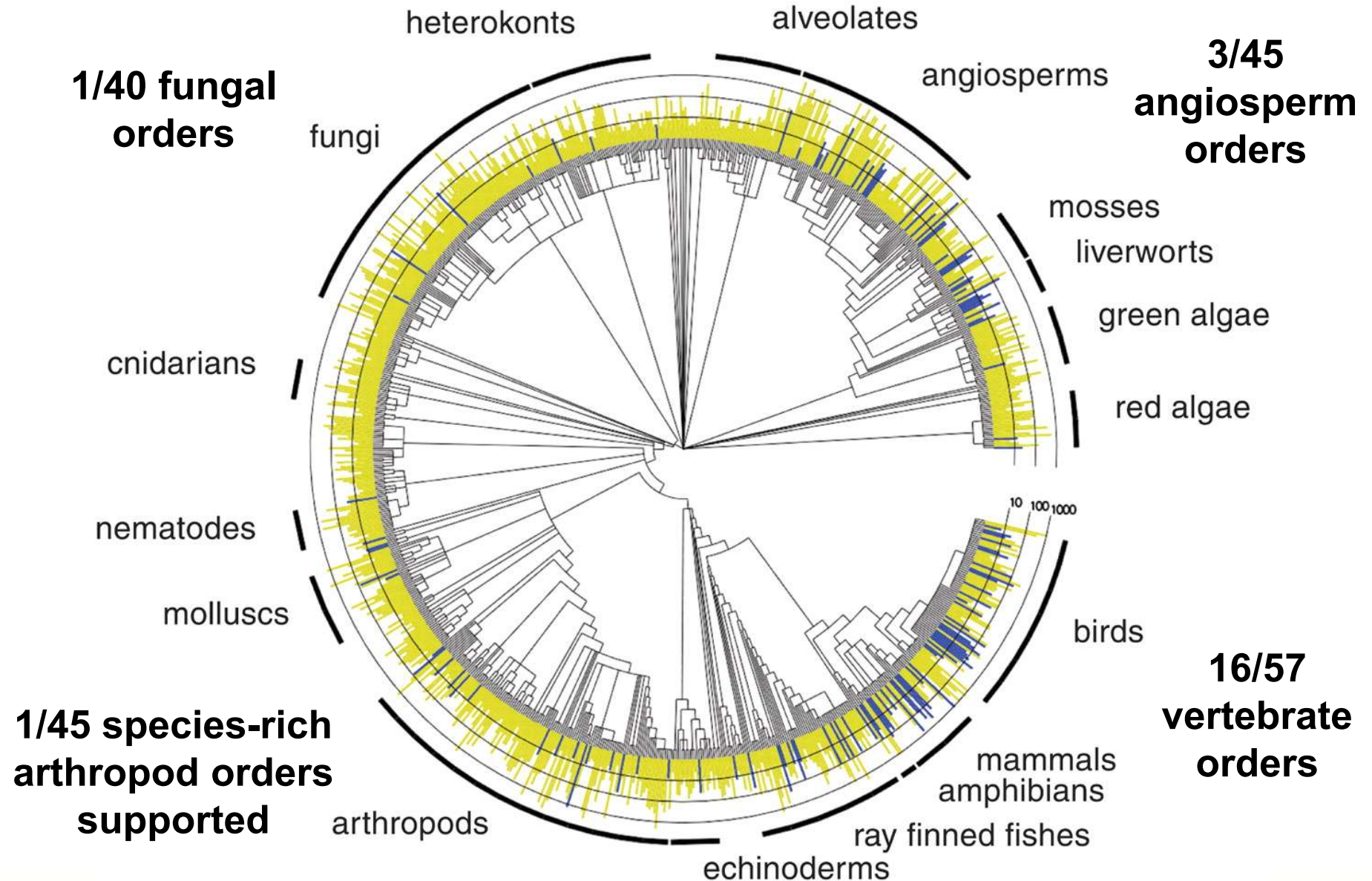
Estimating the Taxonomic Breadth of the Tree of Life

2002: Cracraft
“guess-timates”
that 0.4% of all
known species
have been
included in at least
one published
phylogeny

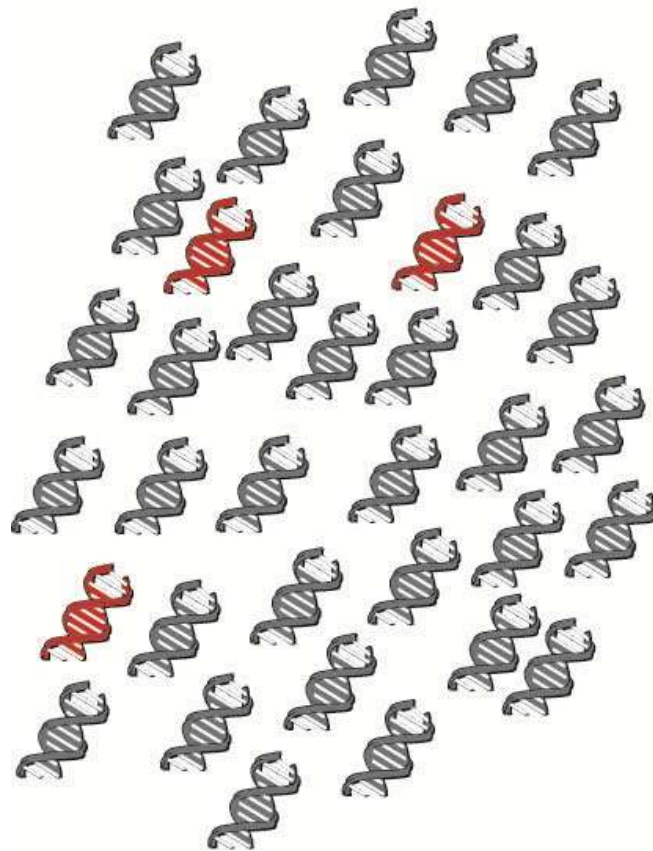
2008: Sanderson
reports that
molecular
sequence data
have been sampled
from 10% of all
known species



The Genomic Depth of the Tree of Life



Next-Gen Sequencing is Qualitative and Quantitative



NGSTs

Each DNA template
is sequenced directly

Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript

Capillary Sequencing

All DNA templates are sequenced
together to create a single
consensus sequence

Grey transcript



Transcriptome Sequencing to Increase Genomic Depth

- 1. Smaller than genome (in *Anopheles gambiae* the transcriptome makes up 7% of the genome)**
- 2. Fewer repetitive and transposable elements**
- 3. Unequal representation (up to 6-7 orders of magnitude)**

Enriched for housekeeping and energy genes (they tend to be conserved)
- 4. The overwhelming majority of sequence evolution models have been developed for and tested in coding sequences**

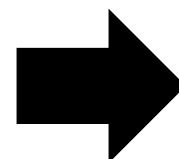


Can we Use RNA-Seq to Increase Genomic Depth?

Species	Stock No.	Collection Location
<i>Anopheles albimanus</i> (<i>Nyssorhynchus</i>)	MRA-126	El Salvador
<i>Anopheles arabiensis</i> <i>Cellia</i>)	MRA-339	Zimbabwe
<i>Anopheles dirus</i> (<i>Cellia</i>)	MRA-700	Thailand
<i>Anopheles farauti</i> (<i>Cellia</i>)	MRA-489	Papua New Guinea
<i>Anopheles freeborni</i> (<i>Anopheles</i>)	MRA-130	USA
<i>Anopheles gambiae</i> (<i>Cellia</i>)	MRA-765	Liberia
<i>Anopheles quadriannulatus</i> (<i>Cellia</i>)	MRA-761	South Africa
<i>Anopheles quadrimaculatus</i> (<i>Anopheles</i>)	MRA-139	USA
<i>Anopheles stephensi</i> (<i>Cellia</i>)	MRA-128	India
<i>Aedes aegypti</i> (<i>Stegomyia</i>)	MRA-735	West Africa

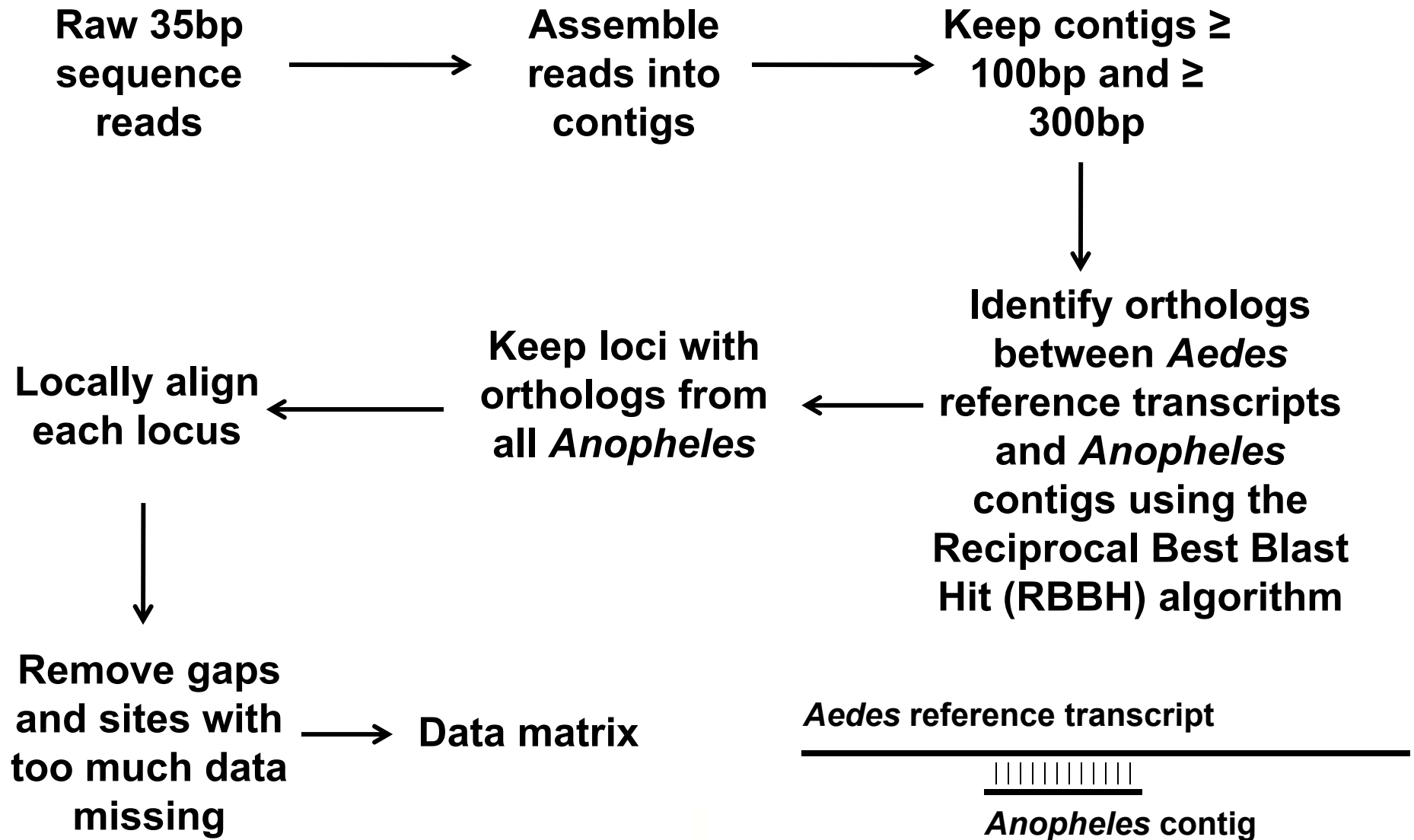


**Illumina
RNA-Seq**



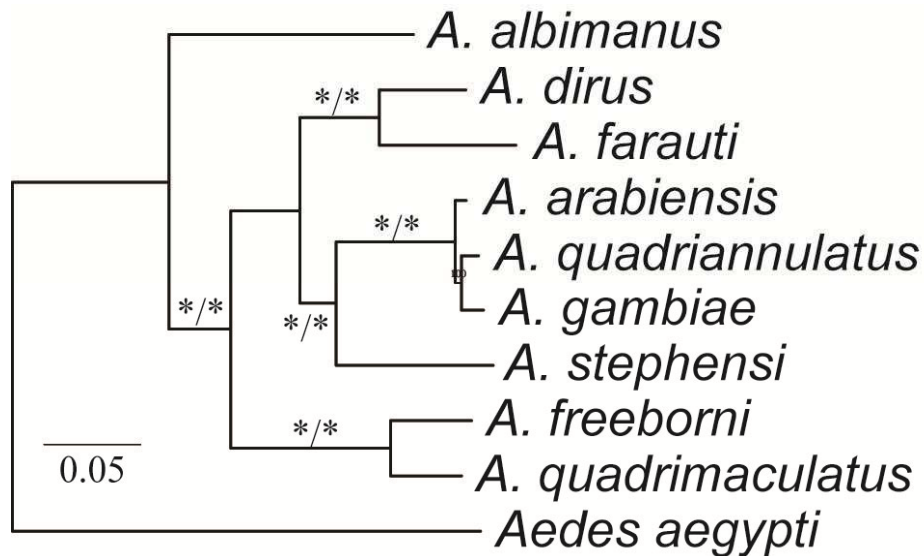
**~150,000,000 reads
5,250,000,000 bp**

Data Matrix Construction: The “Singlecontig” Strategy



Robust Phylogenetic Inference from RNA-Seq Data

Using ≥ 100 bp contigs

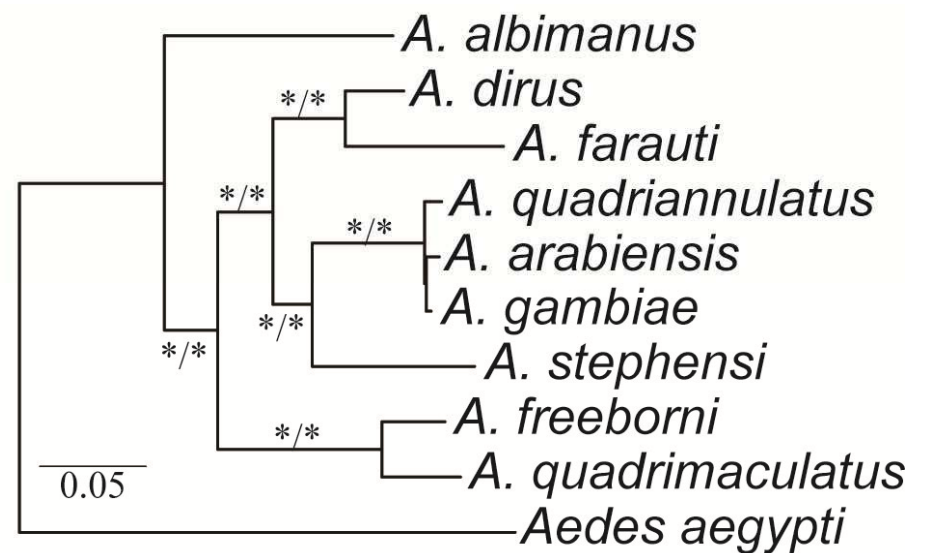


Loci = 553

Aln Length = ~390 Kb

% Missing data = 51

Using ≥ 300 bp contigs



Loci = 69

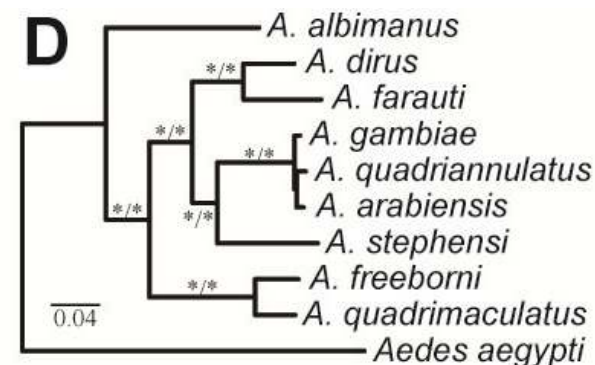
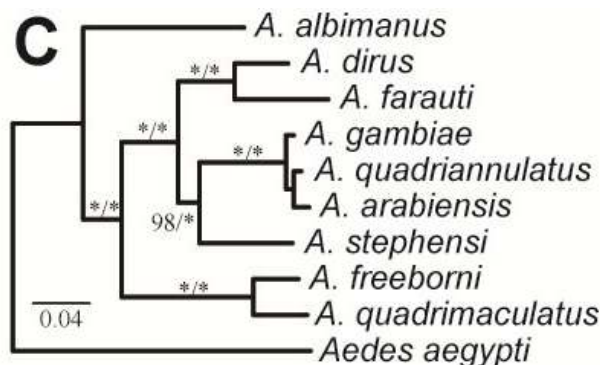
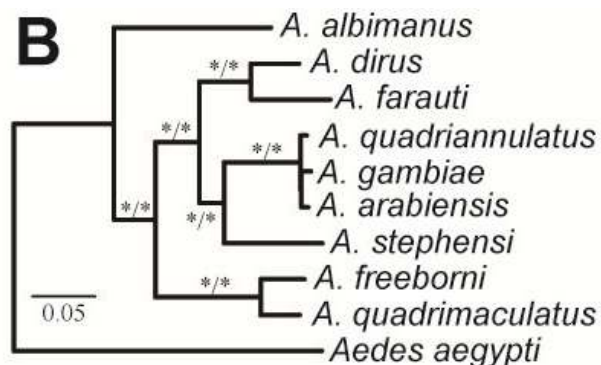
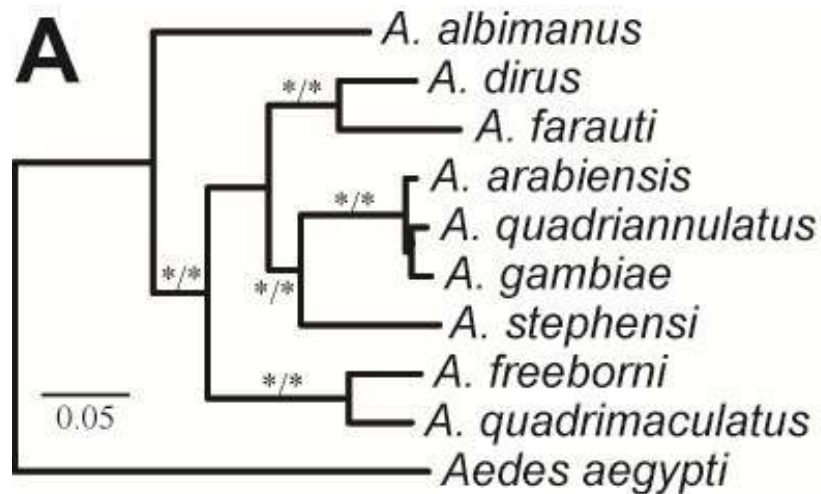
Aln Length = ~73 Kb

% Missing data = 44



Accurate Phylogenetic Inference From our Data

553 loci
Aln L: ~390 Kb
Missing data: 51%



Exclude erroneous loci
491 loci
Aln L: ~329 Kb
Missing data: 50%

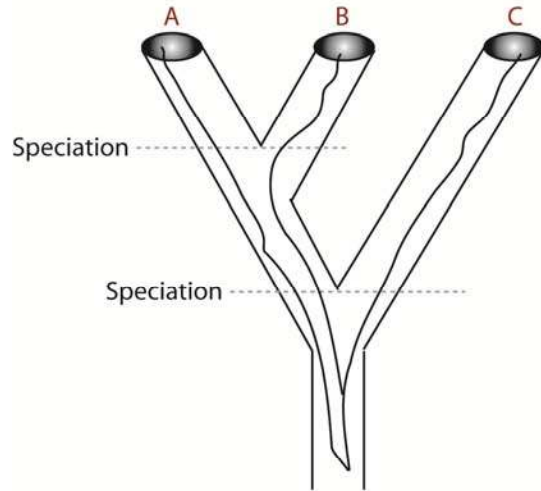
Use only sites without
data missing
Aln L: ~15 Kb
Missing data: 0%

Use *A. gambiae* as ref.
634 loci
Aln L: ~472 Kb
Missing data: 50%

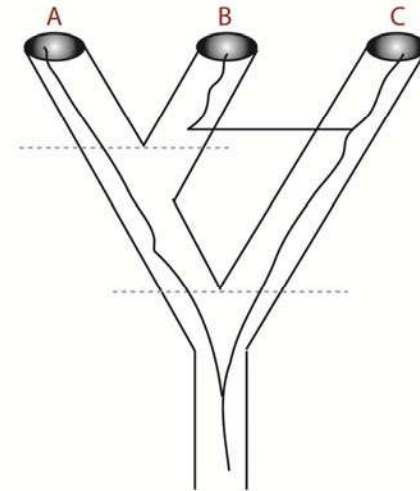


Gene Trees Can Differ from Species Trees

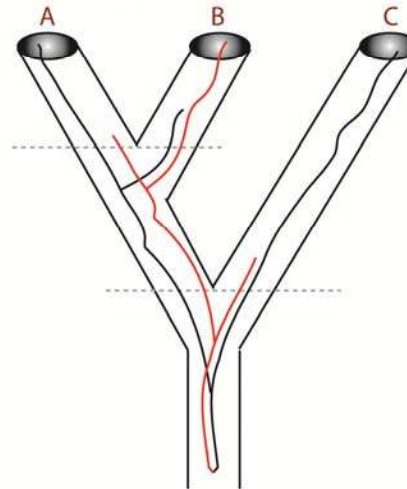
Lineage Sorting



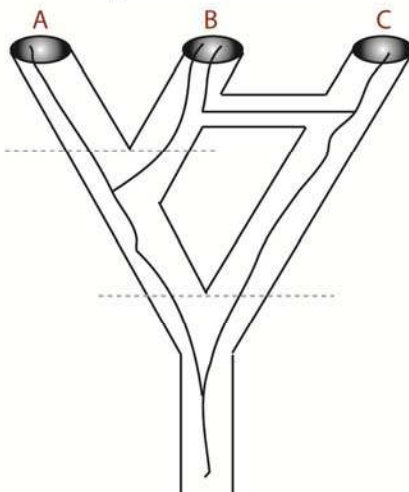
Horizontal Gene Transfer



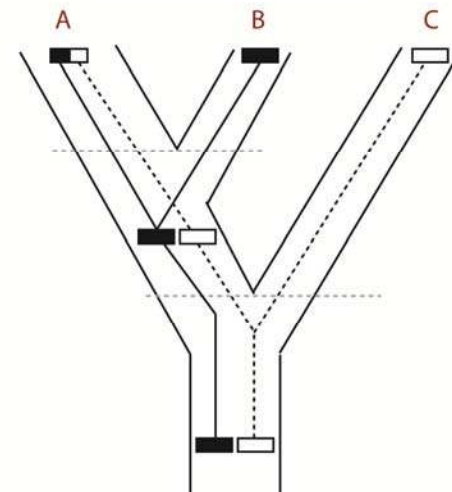
Gene Duplication and Loss



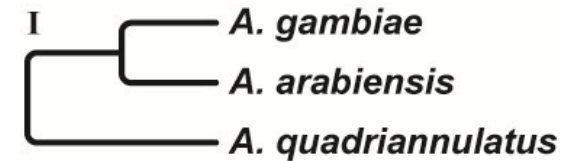
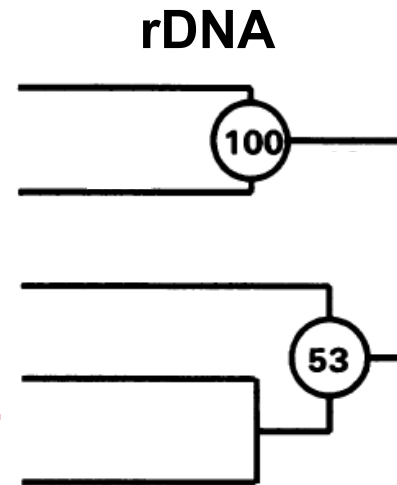
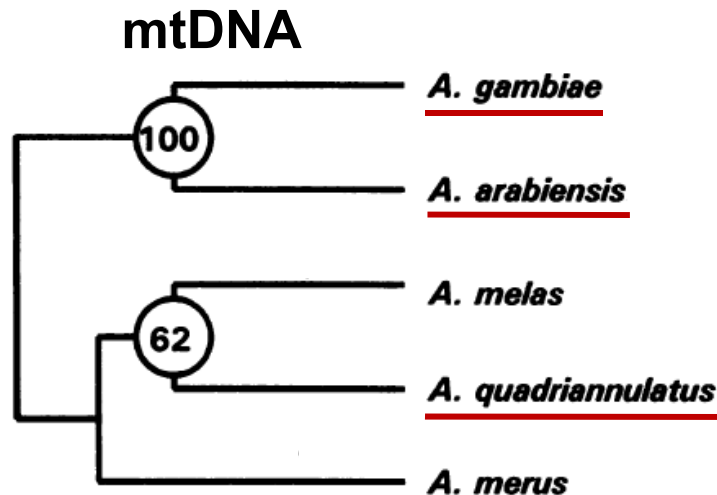
Hybridization



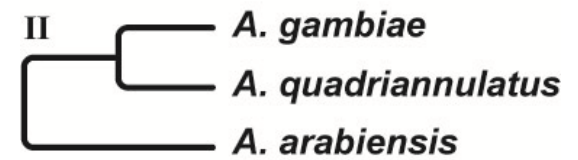
Recombination



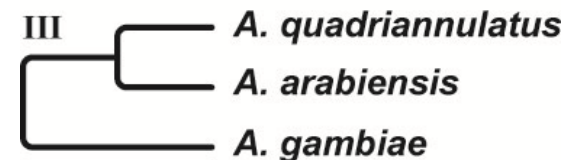
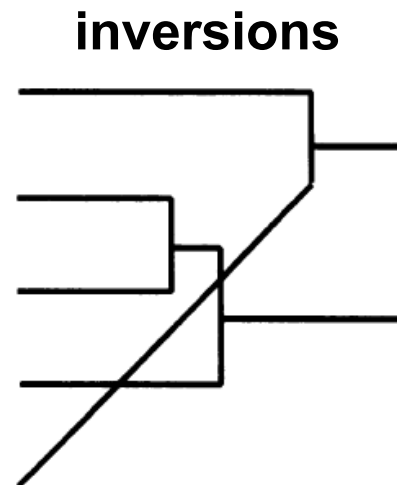
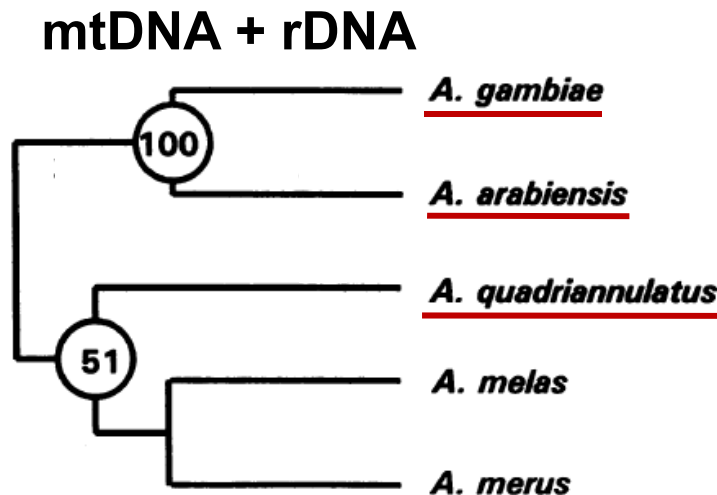
Our Data Matrices Can Detect Population-Level Events



10/237



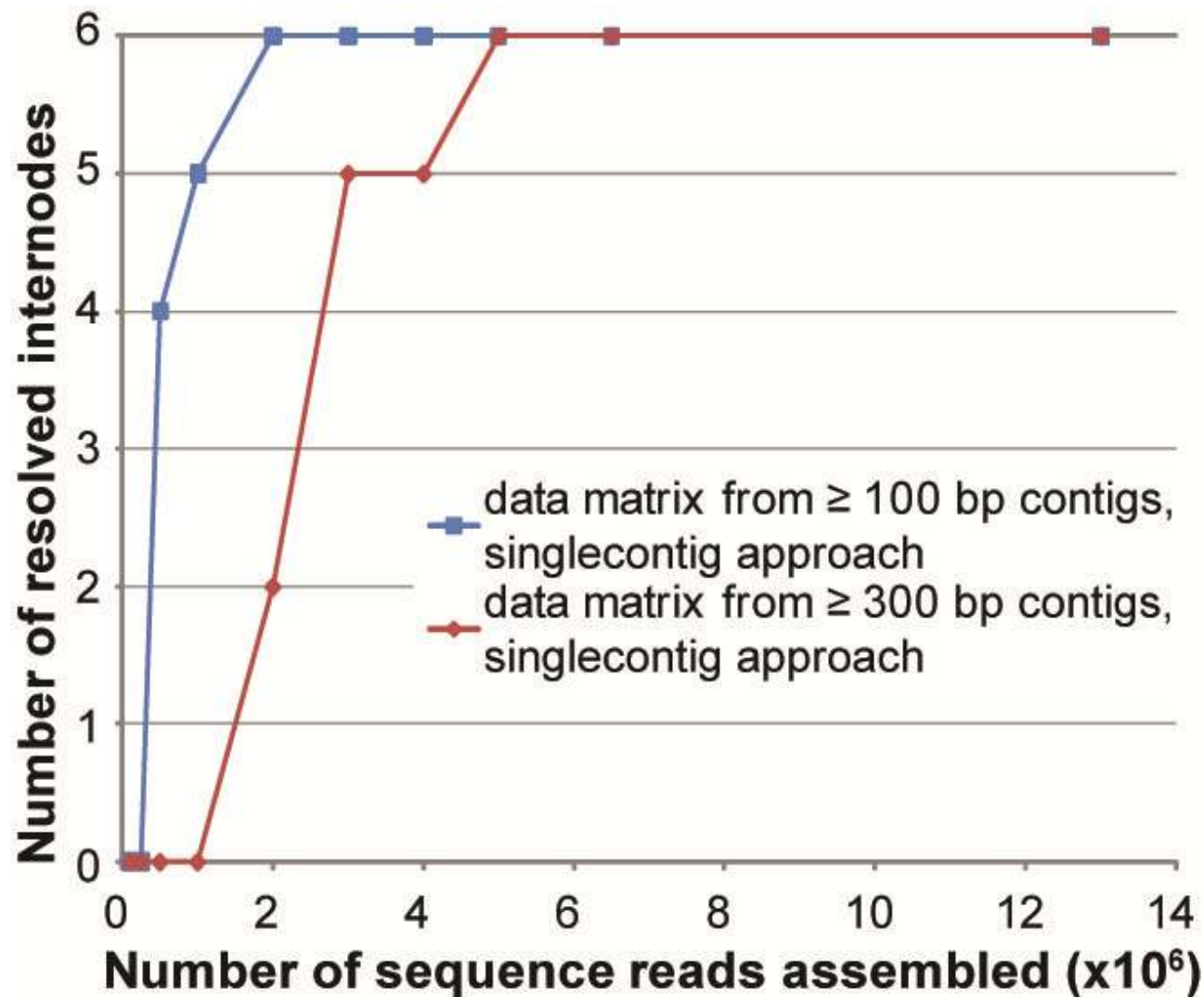
12/237



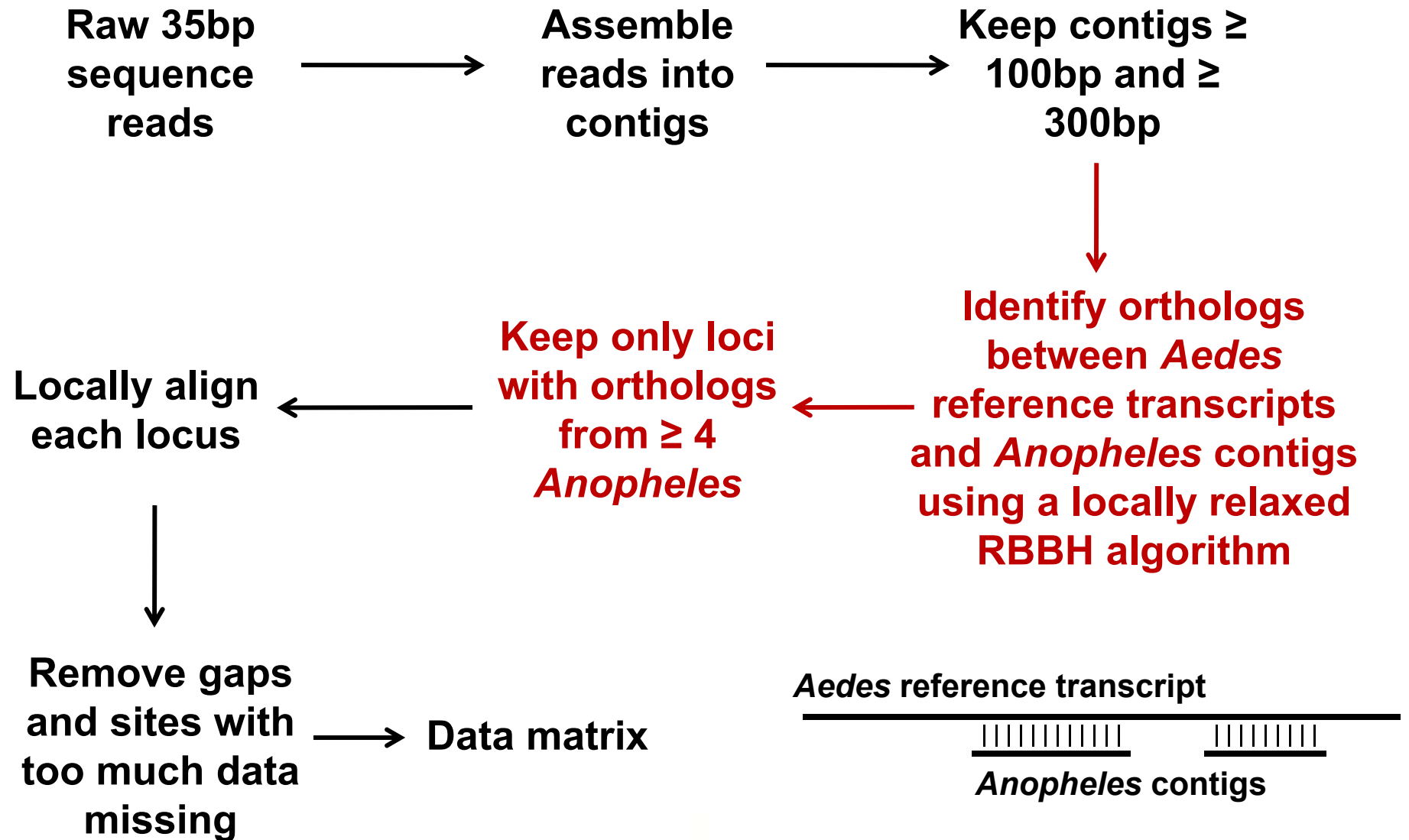
10/237



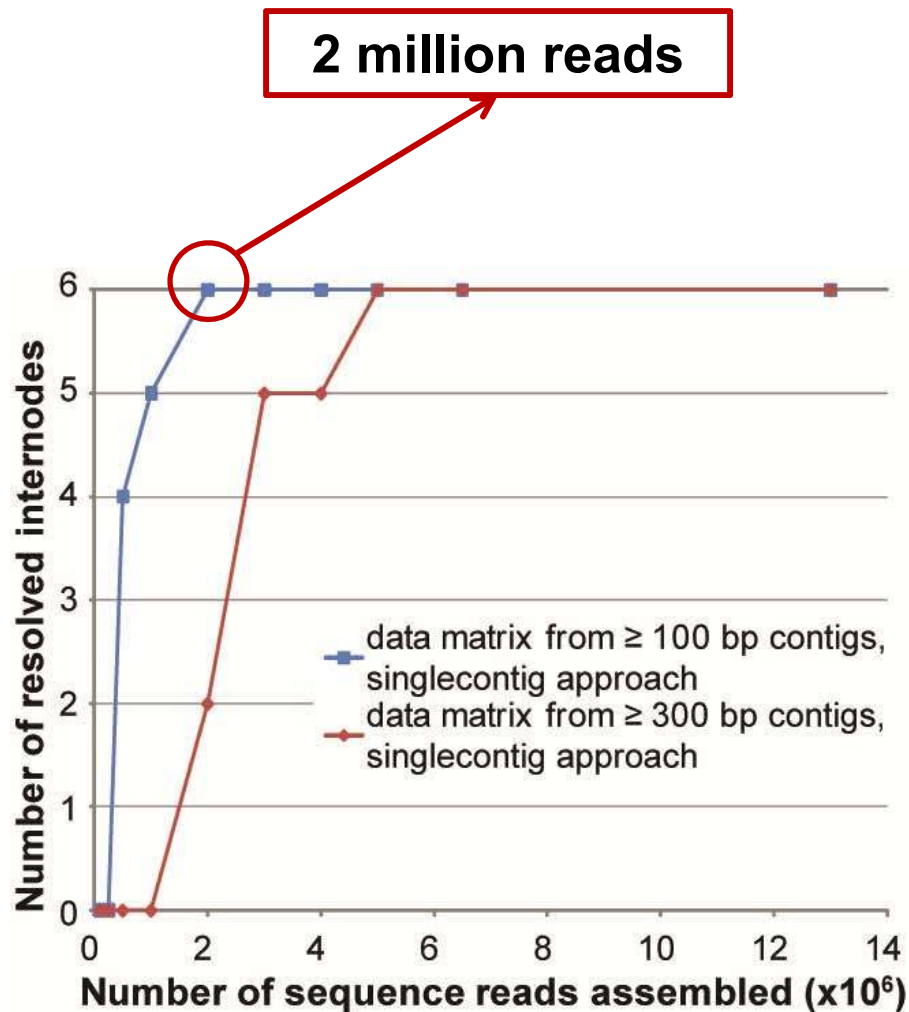
Robust Phylogenetic Inference From Few Sequence Reads



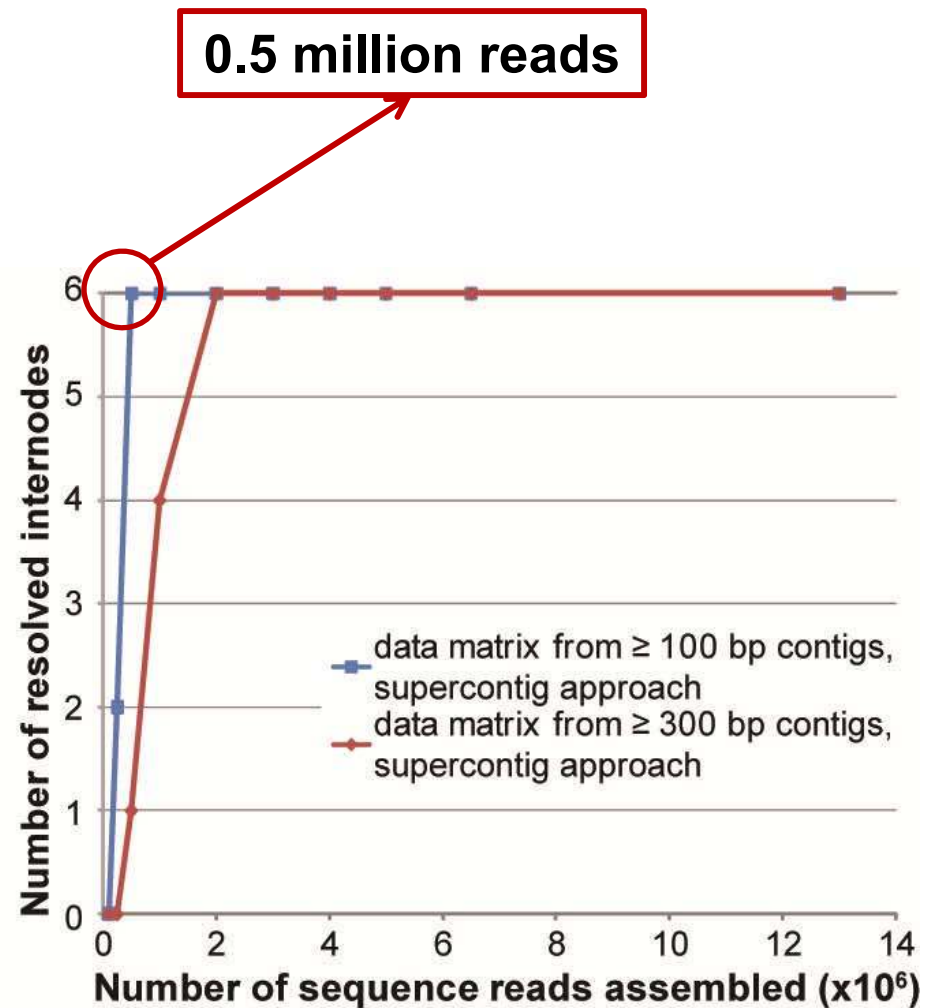
Experimental Design: The “Supercontig” Strategy



Robust Phylogenetic Inference From Very Few Reads



553 loci, AInL: ~390Kb, %miss: 51



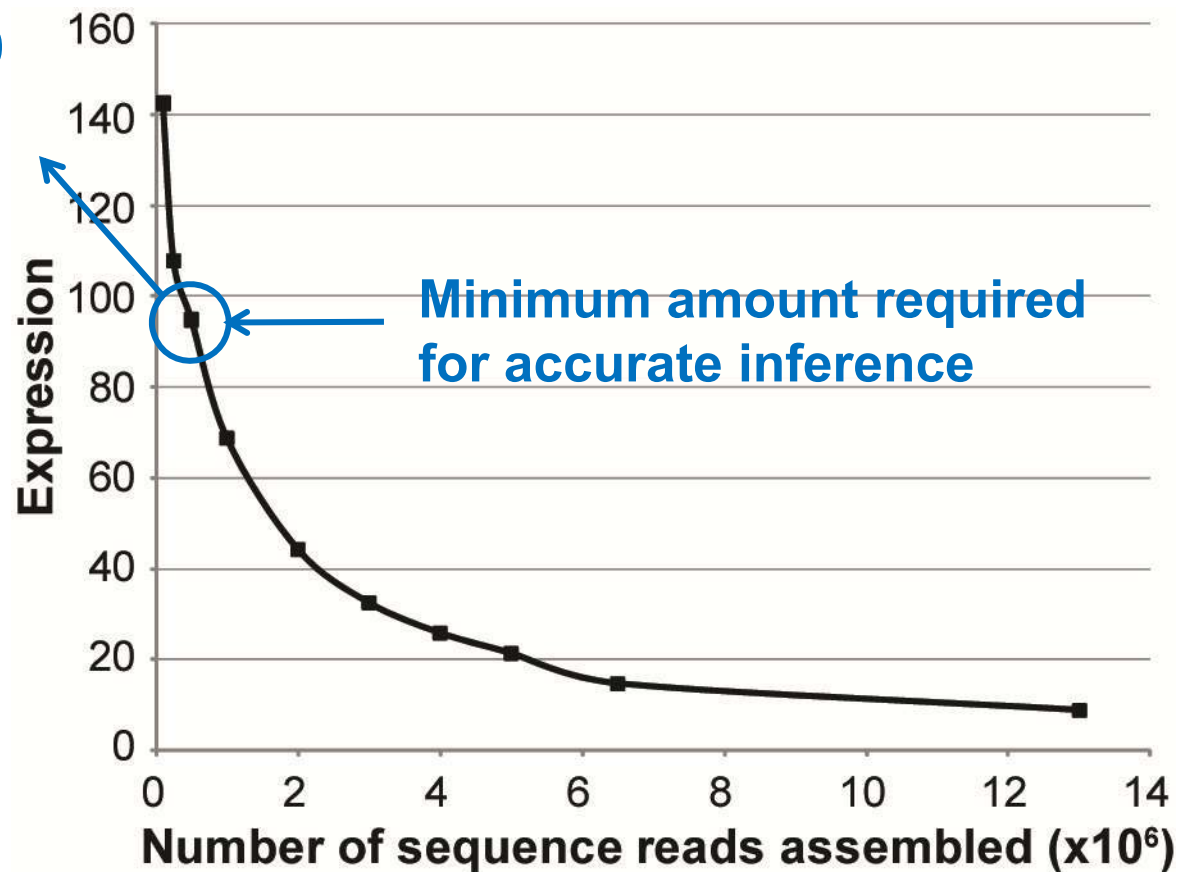
2,661 loci, AInL: ~971Kb, %miss: 62



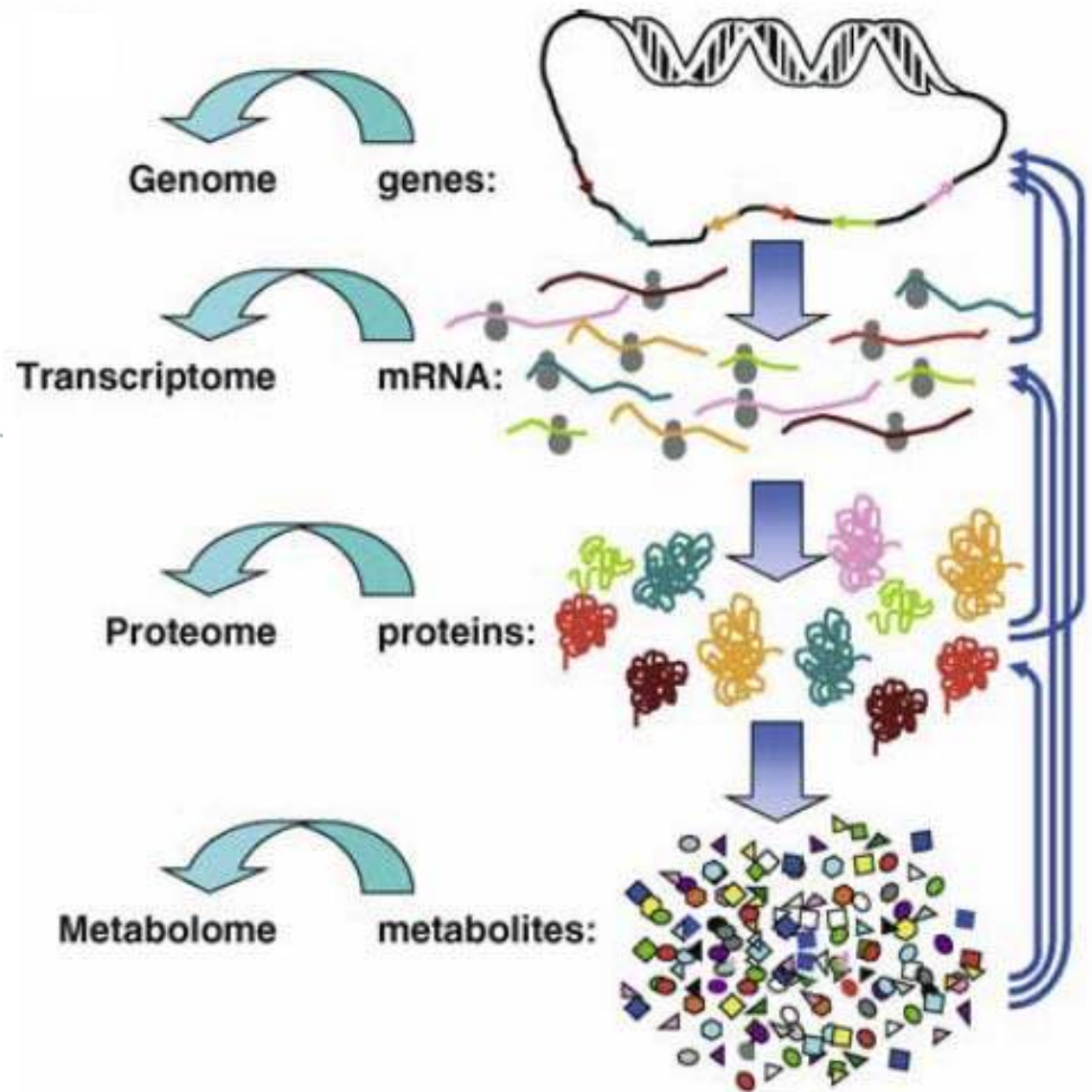
Our Sequences are from Highly-Expressed Transcripts

2008 cost: ~\$50

2013 cost: ~\$5

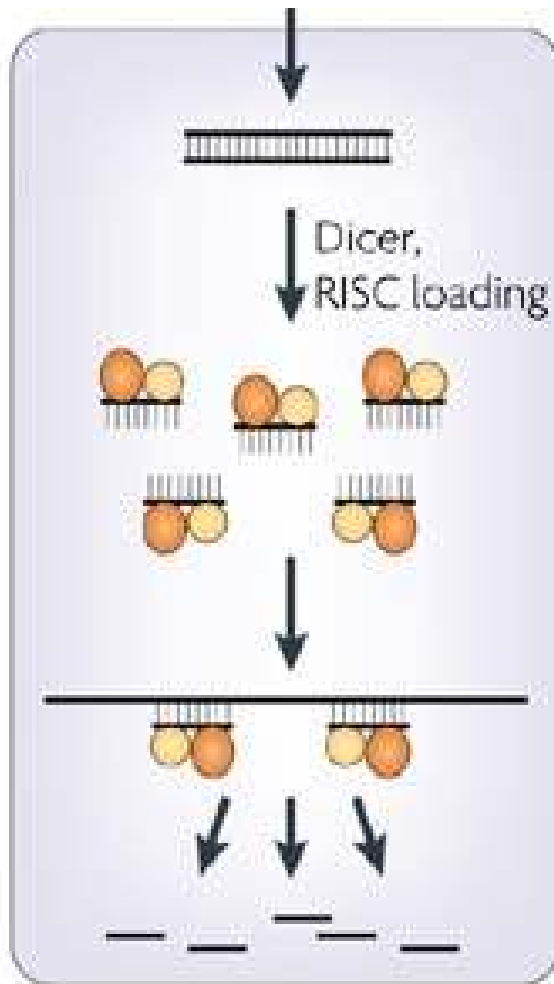


The Age of High Throughput Technologies

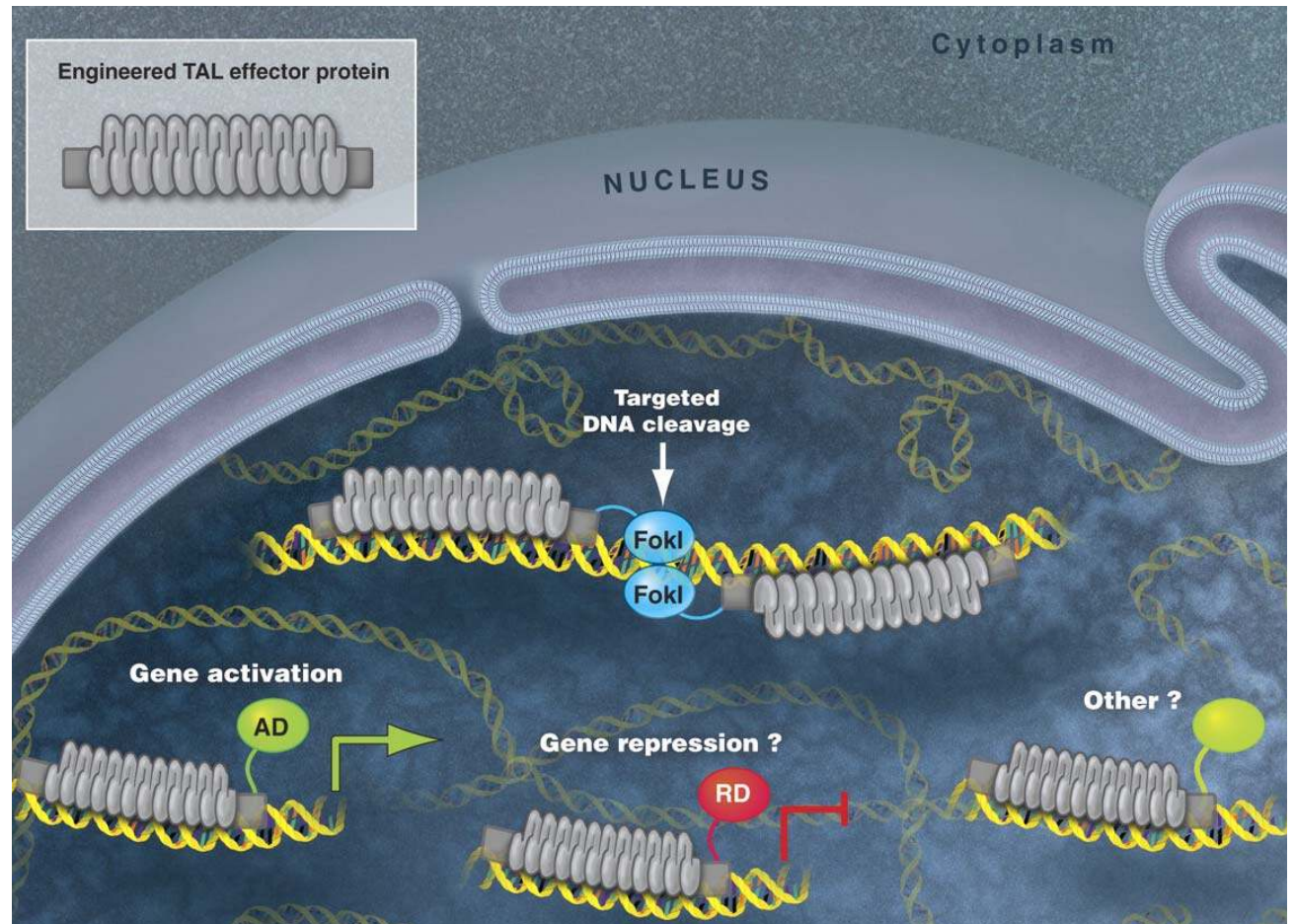


Novel Technologies for Probing Gene Function

RNAi



TAL Effectors



Boutros & Ahringer (2008) Nat. Rev. Genet.;
Bogdanove & Voytas (2011) Science

The Genomes of Non-Model Organisms are the New Frontiers



Rokas & Abbot (2009) Trends Ecol. Evol.

Acknowledgements



Jason Slot



John Gibbons



Chris Hittinger



National Science Foundation
WHERE DISCOVERIES BEGIN

SEARLE SCHOLARS PROGRAM



<http://as.vanderbilt.edu/rokaslab>