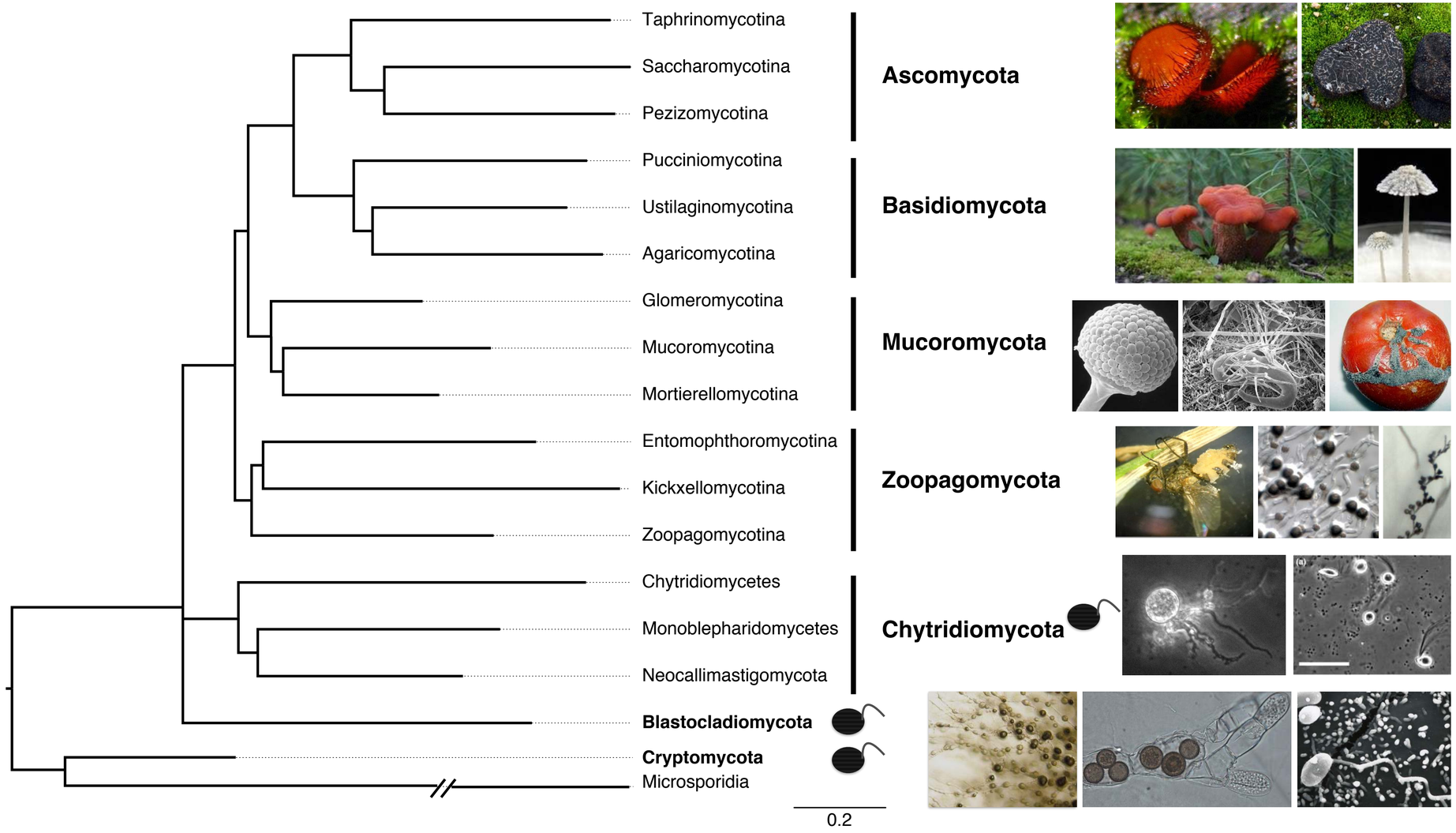
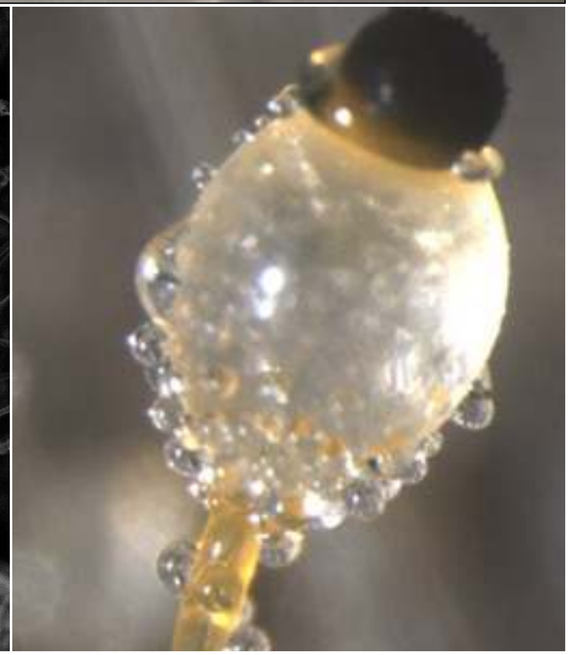
EVOLUTIONARY GENOMICS

Jason Stajich
Univ of California-Riverside

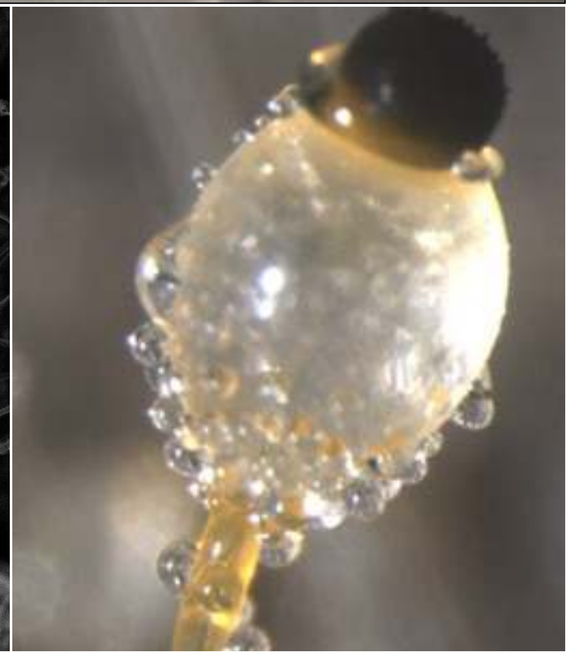
KINGDOM FUNGI





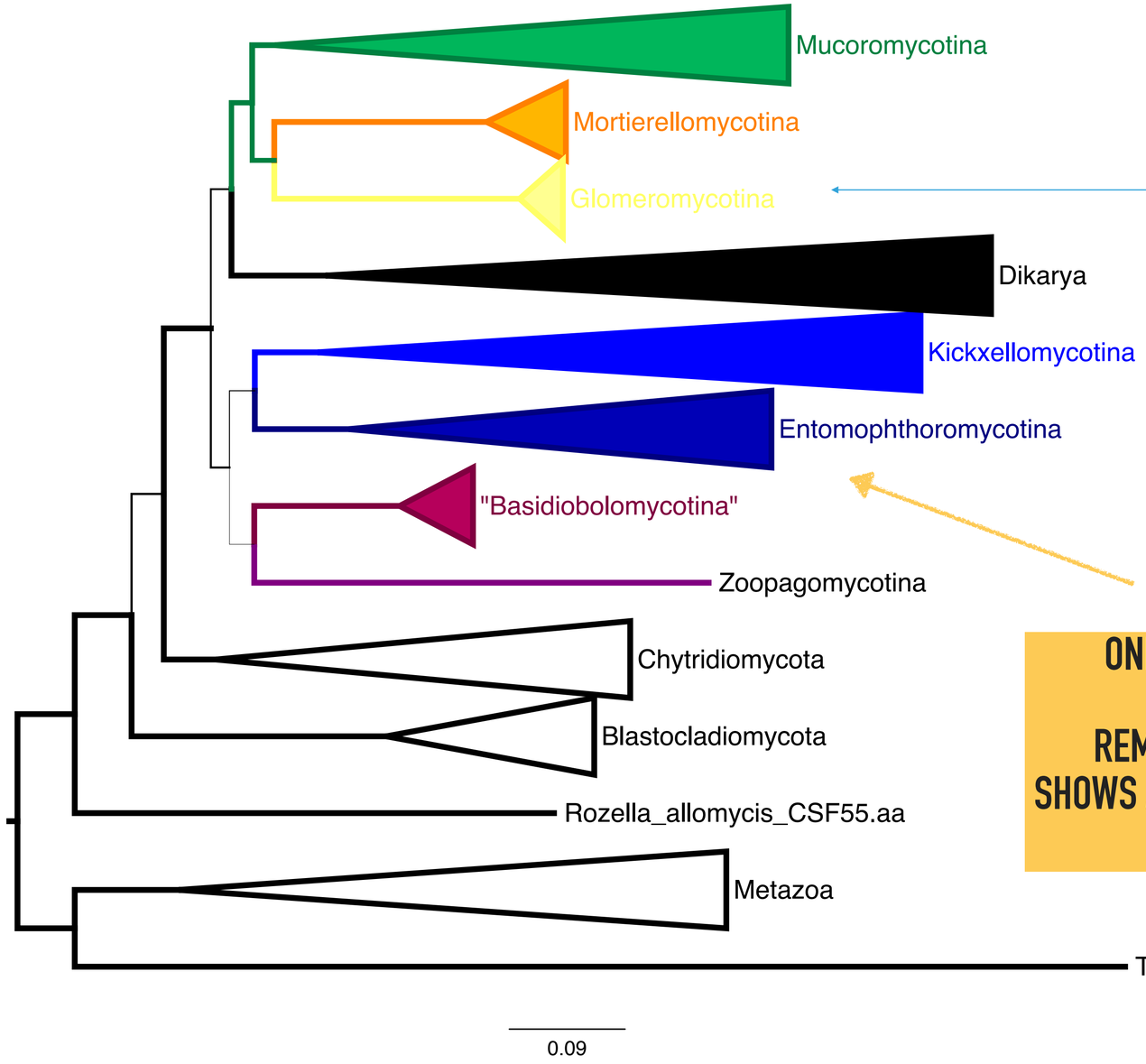
ZYGOMYCETE GENEALOGY OF LIFE

- Revisit phylogenetic relationships with whole genome data. Reference genomes and light coverage.
  - Extensive genome sampling of the zygomycete phyla
  - Estimating divergence time incorporating fossil data
- Examine evolution of sub-cellular characters on phylogeny (Spitzenkörper; Septa)



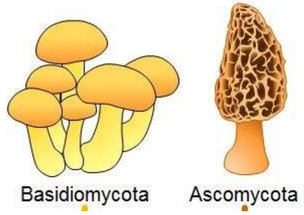
ZYGOMYCETE GENEALOGY OF LIFE

- Revisit phylogenetic relationships with whole genome data. Reference genomes and light coverage.
  - Extensive genome sampling of the zygomycete phyla
  - Estimating divergence time incorporating fossil data
- Examine evolution of sub-cellular characters on phylogeny (Spitzenkörper; Septa)



Mucoromycota

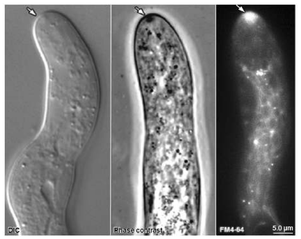
Glomeromycotina



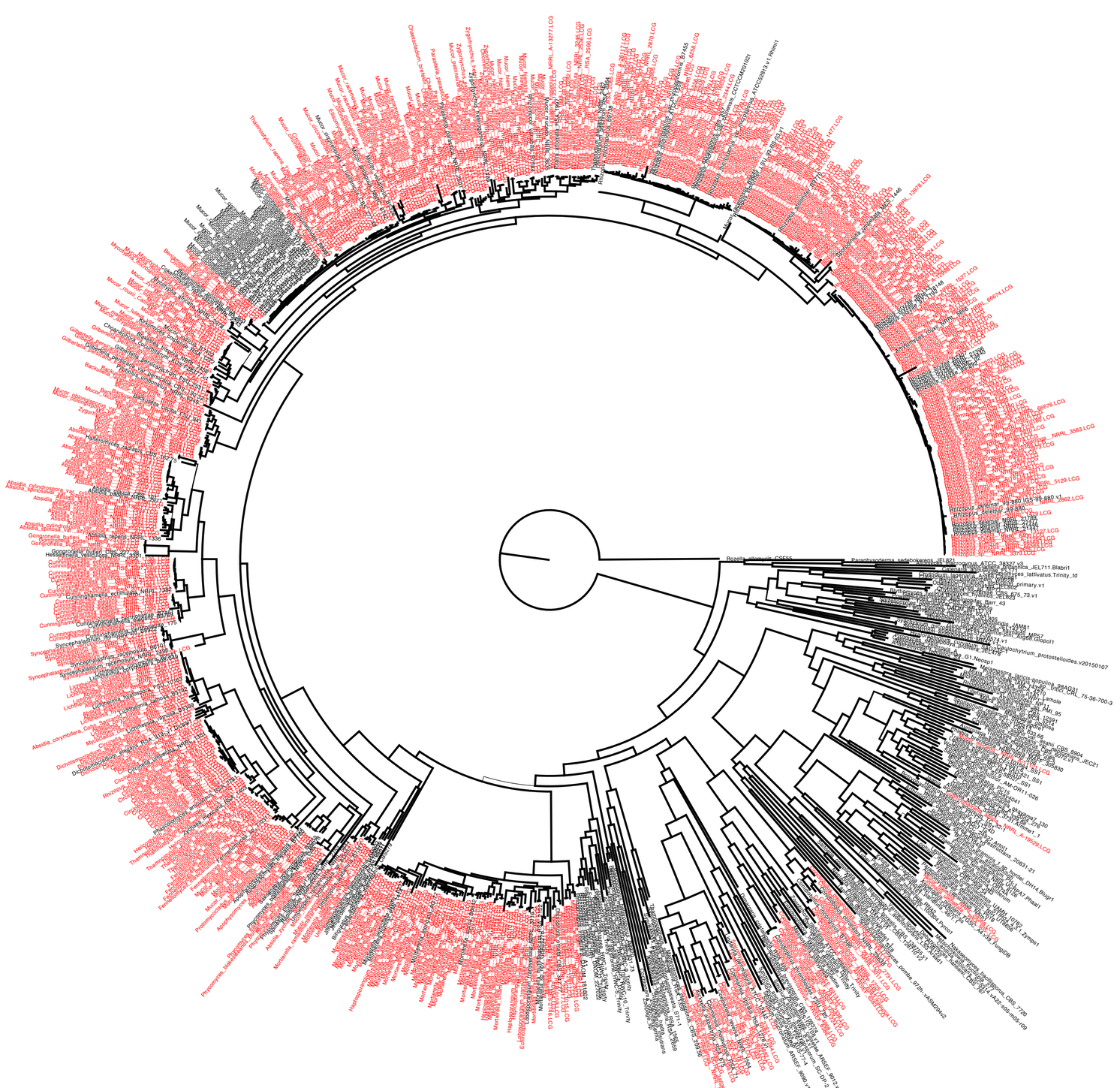
Zoopagomycota

ONE DIFFICULT TO PLACE LINEAGE IS BASIDIOBOLUS

REMAINS OPEN QUESTION AS TO WHERE IT BEST FITS. SHOWS AFFINITY TO THE ENTOMOPTHOROMYCOTINA BUT ALSO HAS OTHER ARRANGEMENTS

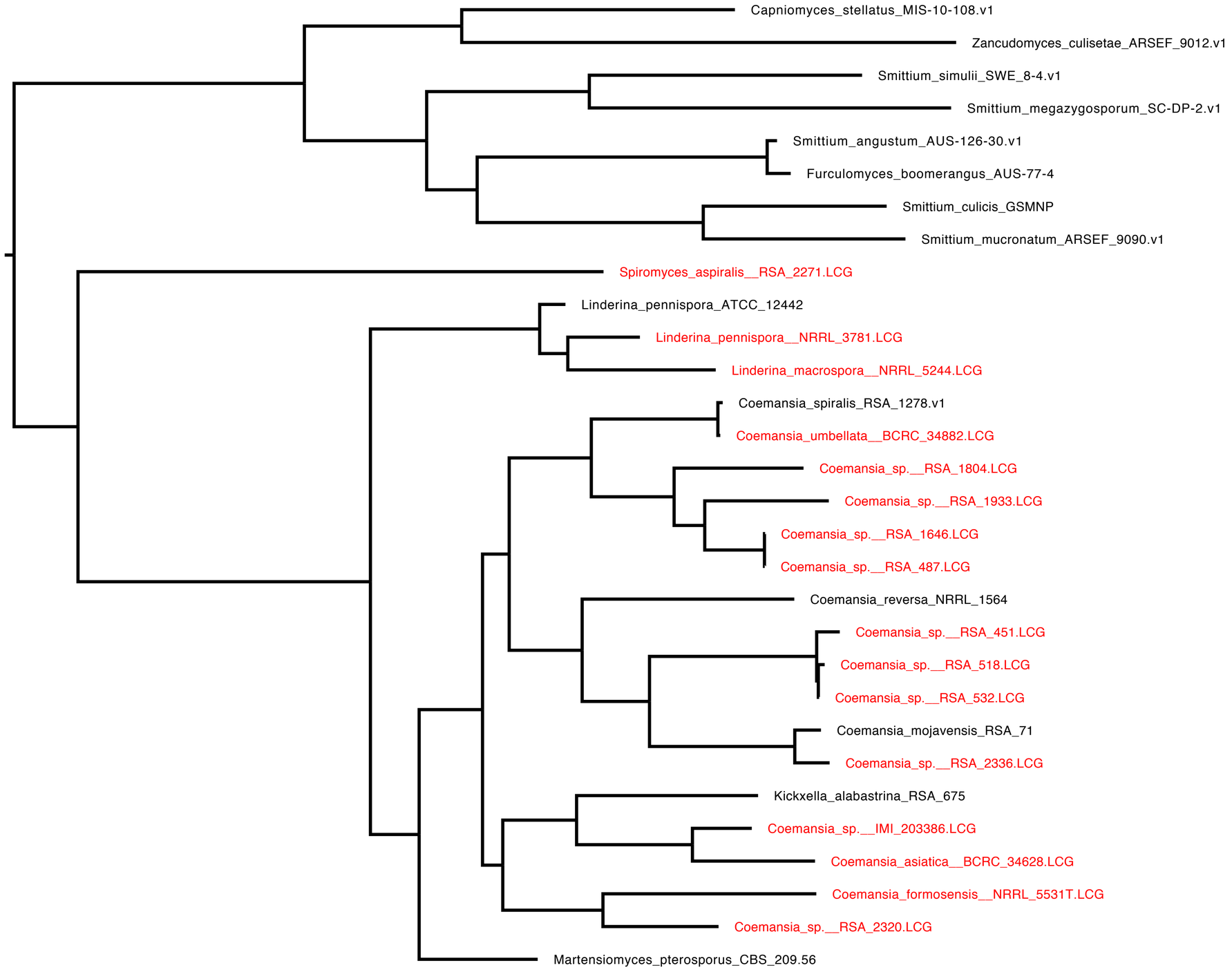


SPATAFORA ET AL, MYCOLOGIA 2016



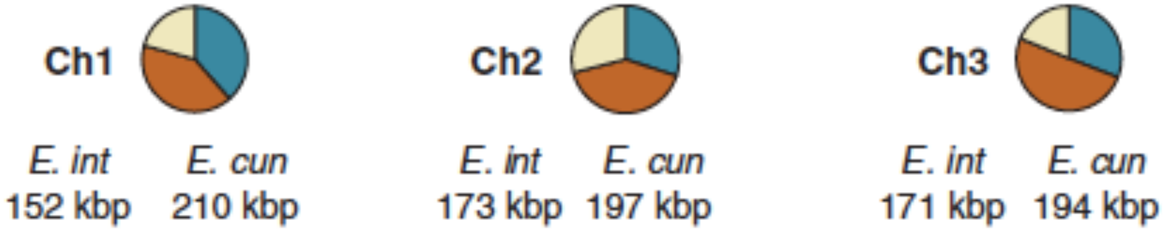
857 Taxa
520 “Light Coverage” Genomes

434 Protein coding Marker genes

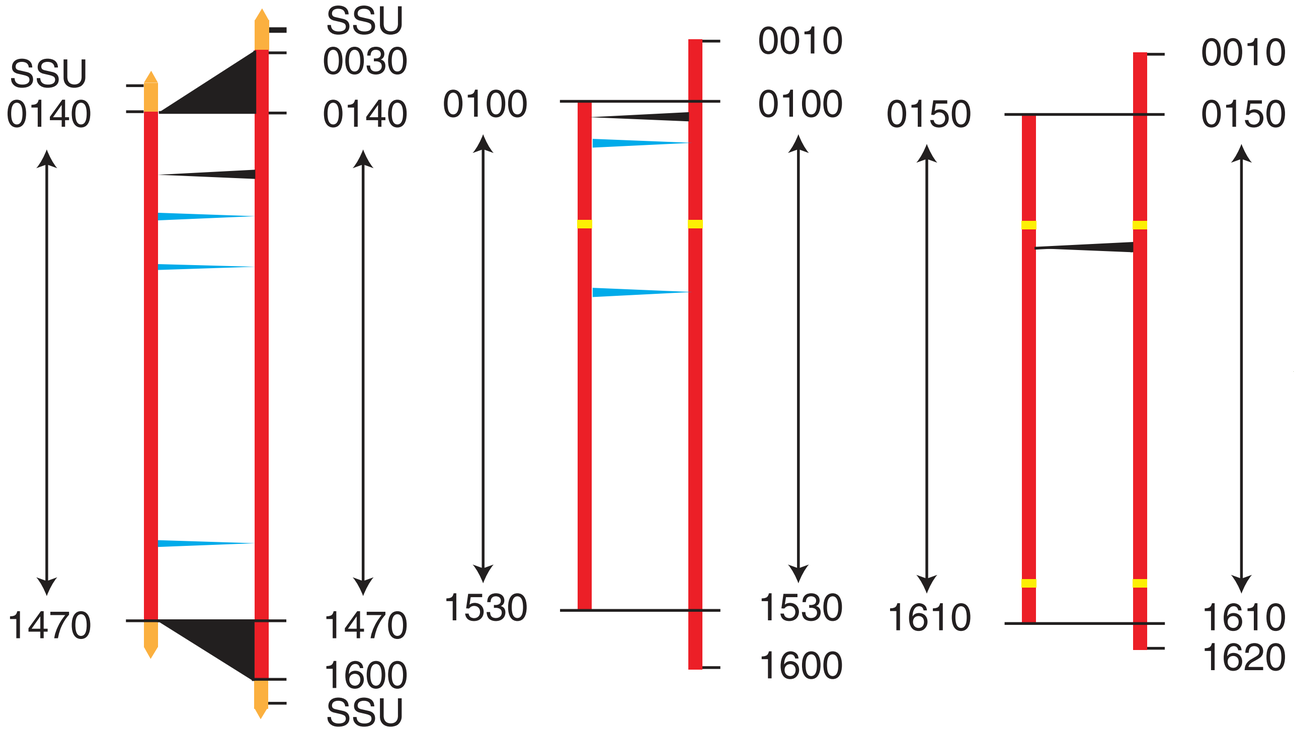

Capniomyces_stellatus_MIS-10-108.v1
Zancudomyces_culisetae_ARSEF_9012.v1
Smittium_simulii_SWE_8-4.v1
Smittium_megazygosporum_SC-DP-2.v1
Smittium_angustum_AUS-126-30.v1
Furculomyces_boomerangus_AUS-77-4
Smittium_culicis_GSMNP
Smittium_mucronatum_ARSEF_9090.v1
Spiromyces_aspiralis__RSA_2271.LCG
Linderina_pennispora_ATCC_12442
Linderina_pennispora__NRRL_3781.LCG
Linderina_macrospora__NRRL_5244.LCG
Coemansia_spiralis_RSA_1278.v1
Coemansia_umbellata__BCRC_34882.LCG
Coemansia_sp.__RSA_1804.LCG
Coemansia_sp.__RSA_1933.LCG
Coemansia_sp.__RSA_1646.LCG
Coemansia_sp.__RSA_487.LCG
Coemansia_reversa_NRRL_1564
Coemansia_sp.__RSA_451.LCG
Coemansia_sp.__RSA_518.LCG
Coemansia_sp.__RSA_532.LCG
Coemansia_mojavensis_RSA_71
Coemansia_sp.__RSA_2336.LCG
Kickxella_alabastrina_RSA_675
Coemansia_sp.__IMI_203386.LCG
Coemansia_asiatica__BCRC_34628.LCG
Coemansia_formosensis__NRRL_5531T.LCG
Coemansia_sp.__RSA_2320.LCG
Martensiomyces_pterosporus_CBS_209.56

GENOME EVOLUTION AND COMPARATIVE ANALYSES

Within Single Individual Genome	Genomes of multiple individuals, species
Summary Statistics	Comparative analysis
Genome & Chromosome Structure	Gene Families content
Total genome size	Orthologs & Paralogs
Organization and compactness	Shared vs Unique genes
Gene/Feature Content (domains)	Tests for horizontal transfer
	Gene Trees
Patterns of gene duplication	
	Tests for Selection



Blue, brown and beige colours represent the portion of proteins that are, respectively, shorter, identical or longer in *E. intestinalis* compared with *E. cuniculi* orthologues



METHODS

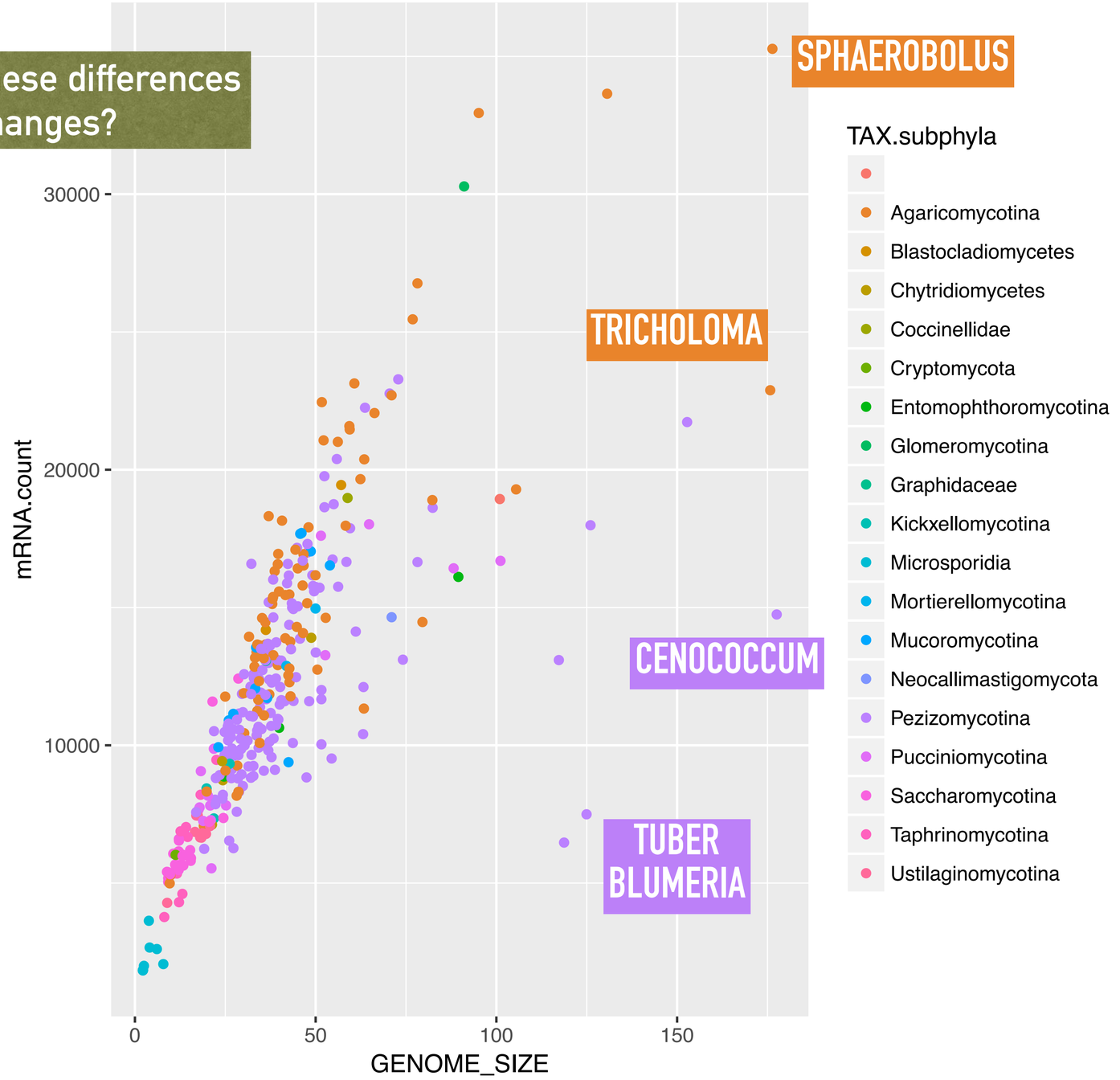
- Genome feature investigation
  - Basic gene or genome descriptions
  - Gene content and gene

The Microsporidian *E. intestinalis* has a more compact and reduced genome as compared to its sister species *E. cuniculi*

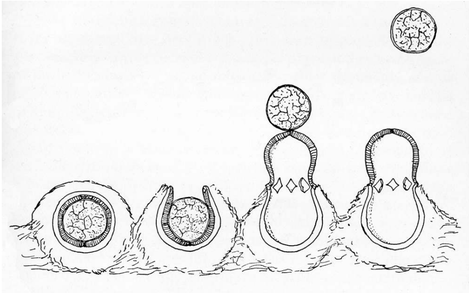
Corradi et al Nat Comm 2010

Fungal Genome size vs Gene number

What drives these differences and changes?



SPHAEROBOLUS

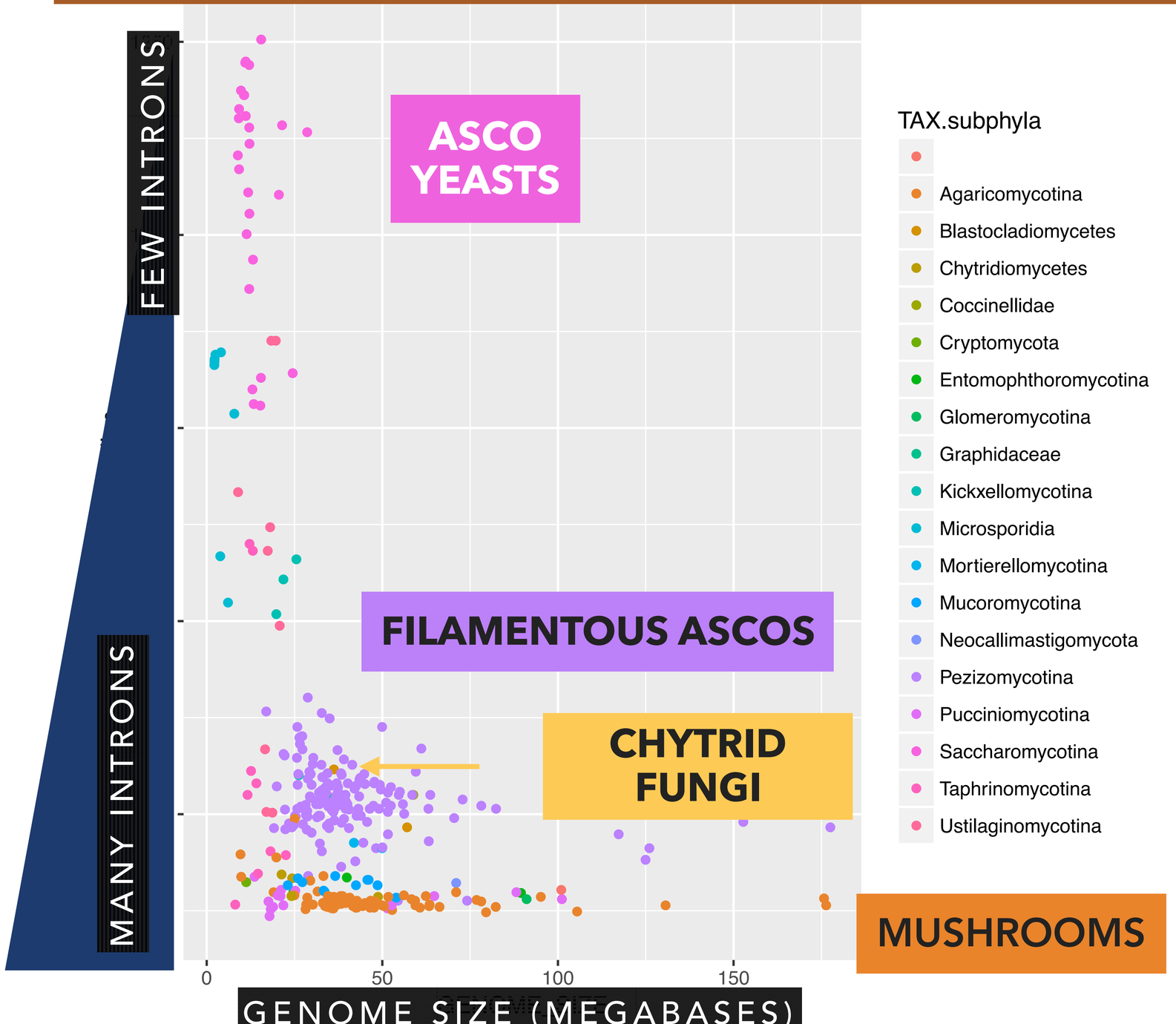


CENOCOCCUM

TUBER

MICROSPORIDIA

Genome size vs Intron density



TAX.subphyla

- Agaricomycotina
- Blastocladiomycetes
- Chytridiomycetes
- Coccinellidae
- Cryptomycota
- Entomophthoromycotina
- Glomeromycotina
- Graphidaceae
- Kickxellomycotina
- Microsporidia
- Mortierellomycotina
- Mucoromycotina
- Neocallimastigomycota
- Pezizomycotina
- Pucciniomycotina
- Saccharomycotina
- Taphrinomycotina
- Ustilaginomycotina

ANCESTRAL FUNGI HAD GENES RICH WITH INTRONS BASED ON OBSERVATIONS OF MANY SHARED INTRON POSITIONS AMONG ORTHOLOGOUS GENES FROM PHYLOGENETICALLY DIVERSE SPECIES

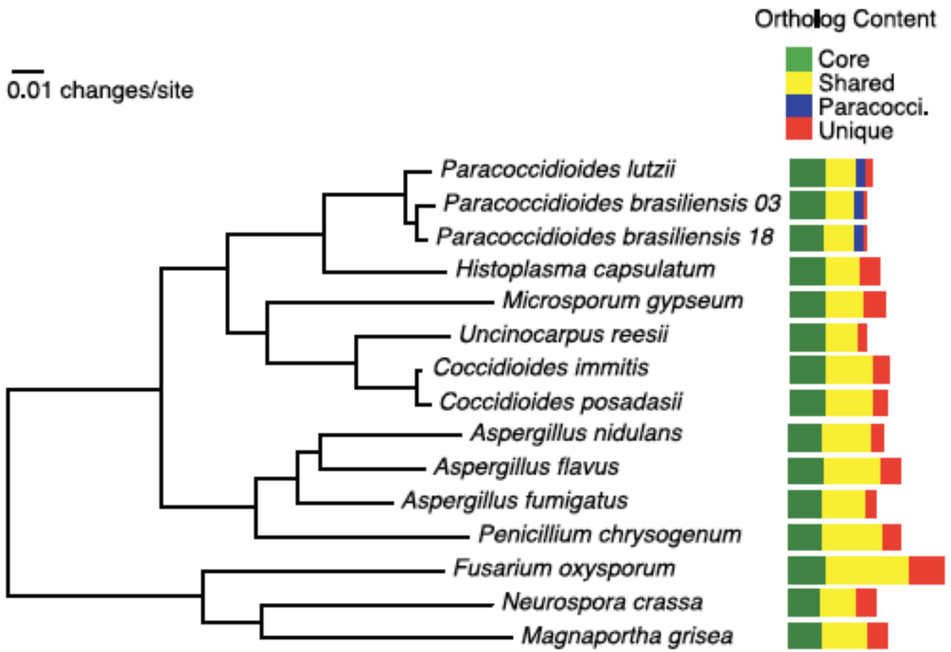
INDEPENDENT INTRON LOSS PRESSURE IN SEVERAL YEAST LINEAGES (SACCHAROMYCOTINA AND TAPHRINOMYCOTINA)

SUMMARY STATISTICS ABOUT GENOME CONTENT WITH SOME COMPARISONS

Table 1. Assembly and gene statistics.

	P. lutzii	*P. brasiliensis*	
		Pb03	**Pb18**
Coverage	8.0X	8.9X	9.8X
Assembly size (Mb)	32.9	29.1	30.0
Total contig length (Mb)	32.6	28.8	29.4
Scaffolds	111	65	57
Scaffold N50 (Mb)	1.02	1.97	2.13
Contigs	885	552	669
Contig N50 (kb)	84.3	114.9	109.7
Quality ≥Q40 (%)	98.9	98.9	98.9
GC (%)	42.8	44.5	44.4
Predicted protein-coding genes	9,132	7,875	8,741
Dubious genes	1,002	265	699
High-confidence genes	8,130	7,610	8,042
Mean gene length (nt)	1,814	1,833	1,802
Mean coding sequence length (nt)	1,330	1,433	1,346
Mean intron length (nt)	126	140	132
Mean intron number per gene	3.1	2.5	2.8
Mean exon number per gene	4.1	3.5	3.8
GC exonic (%)	49.8	50.4	50.1
GC intronic (%)	41.7	42.4	42.3
Mean intergenic length (nt)	1,799	1,848	1,628
tRNAs	118	103	103
Transmembrane proteins	1121	1057	1084
Secreted proteins	297	291	322
GPI-anchored proteins	61	63	62

doi:10.1371/journal.pgen.1002345.t001



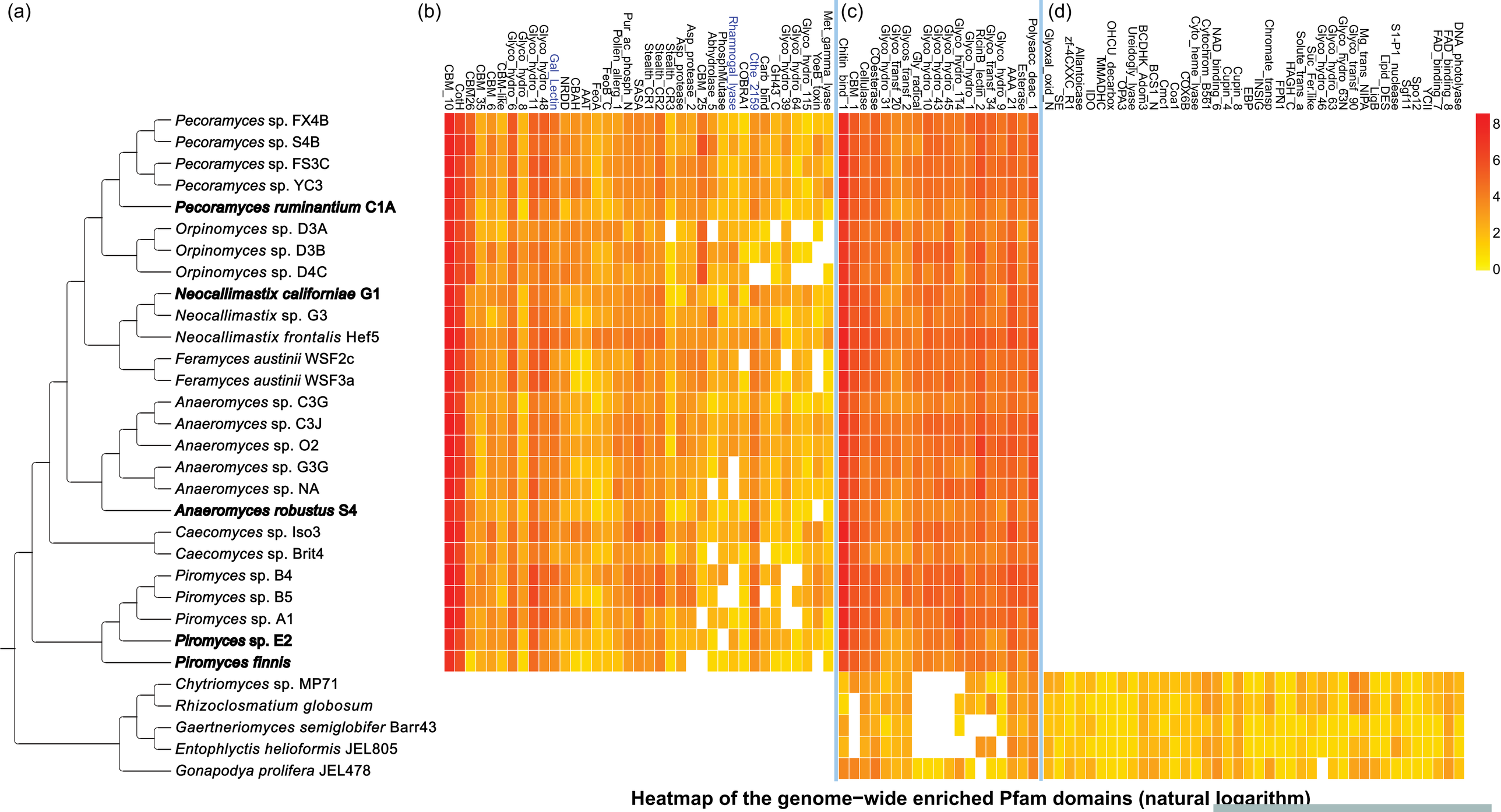
Carbohydrate Metabolism					Peptidase Activity					
CAZY					MEROPS					
GH	GT	CE	PL	CT	A	C	G	M	S	T
65	64	9	0	22	11	54	0	76	136	19
64	65	10	0	20	6	53	0	75	142	19
67	68	10	0	21	7	54	0	74	135	19
73	64	10	0	21	8	53	0	76	137	19
75	70	9	0	17	9	58	0	104	184	21
72	71	10	0	19	7	49	0	89	160	19
72	86	10	0	26	7	58	0	91	175	19
65	79	9	0	24	8	52	0	86	168	19
275	116	40	23	65	11	63	2	113	233	20
306	136	43	23	103	17	65	3	136	297	22
256	112	31	16	56	9	62	2	107	209	19
236	118	32	10	60	19	65	3	102	224	19
324	105	24	17	37	30	77	1	168	422	21
179	87	27	3	11	19	54	2	87	159	20
159	108	51	4	17	20	61	2	121	248	21
*	*	*	*		*	*	*	*	*	

Dejardins et al PLoS Gen

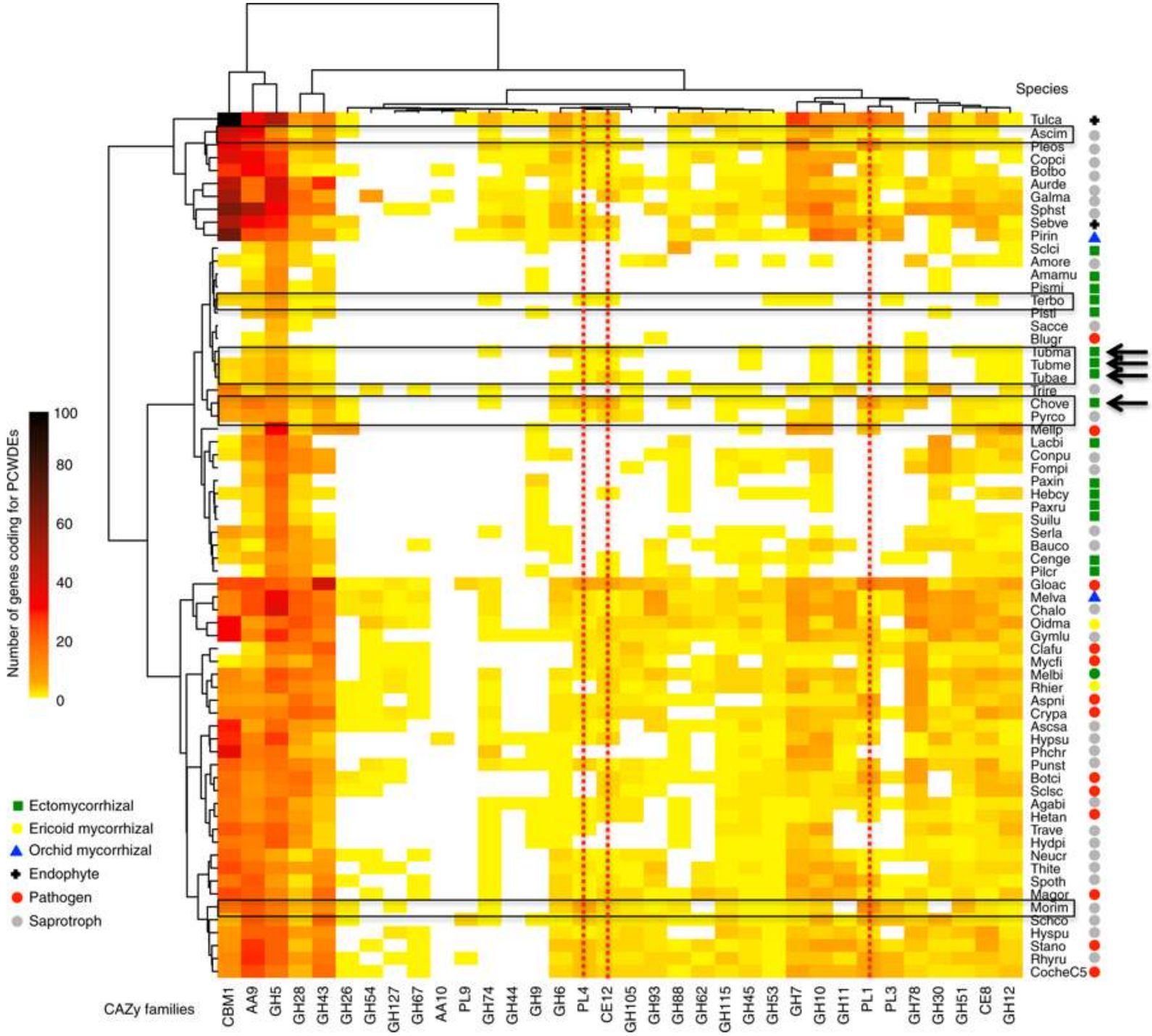
Summary / aggregated Counts by species

Phylum	Subphylum	Species	Ecology 1	Ecology 2	Oxidoreductases	CAZymes	CBM1
Basidiomycota	Agaricomycotina	*Amanita muscaria*	Biotroph	Ectomycorrhizal	29	28	0
		Hebeloma cylindrosporum			22	39	2
		Laccaria bicolor			34	47	1
		Paxillus involutus			23	39	0
		Paxillus rubicundulus			24	35	0
		Piloderma croceum			41	50	0
		Pisolithus microcarpus			17	29	0
		Pisolithus tinctorius			22	26	0
		Scleroderma citrinum			24	25	0
		Suillus luteus			29	40	0
		Tremella mesenterica		Mycoparasite	5	14	0
		Sebacina vermifera		Orchid symbiont	21	121	50
		Tulasnella calospora			9	178	110
		Piriformospora indica		Root endophyte	15	103	64
		Auricularia delicata	Saprotroph	White rot	60	140	42
		Botryobasidium botryosum			16	97	29
		Fomitiporia mediterranea			40	81	10
		Galerina marginata			77	128	52
		Heterobasidion annosum			35	65	17
		Hypholoma sublateritium			52	83	28
		Jaapia argillacea			17	102	24
		Phanerochaete chrysosporium			31	75	28
		Pleurotus ostreatus			43	111	31
		Plicaturopsis crispa			21	70	9
		Punctularia strigosozonata			46	93	26
		Schizophyllum commune			8	103	5
		Sphaerobolus stellatus			(284)	(195)	0
		Trametes versicolor			50	94	23
		Coniophora puteana		Brown rot	16	80	3
		Dacryopinax sp.			14	50	1
		Fomitopsis pinicola			16	65	0
		Gloeophyllum trabeum			13	61	1
		Hydnomerulius pinastri			21	86	16
		Serpula lacrymans			12	52	8
		Agaricus bisporus		Soil, litter or other Saprotroph	47	65	13
		Amanita thiersii			27	71	10
		Coprinopsis cinerea			41	96	44
		Gymnopus luxurians			66	118	32
	Ustilaginomycotina	*Ustilago maydis*	Biotroph	Plant pathogen	6	22	0
	Pucciniomycotina	*Melampsora larici-populina*			31	71	0
Ascomycota	Pezizomycotina	*Tuber melanosporum*	Biotroph	Ectomycorrhizal	5	19	3
		Oidiodendron maius		Ericoid symbiont	31	119	35
		Cryphonectria parasitica		Plant pathogen	18	102	12
		Stagonospora nodorum			16	104	10
		Aspergillus nidulans		Animal pathogen	9	83	6
		Trichoderma resei	Saprotroph	Soil, litter or other Saprotroph	10	39	14
	Saccharomycotina	*Pichia stipitis*			2	10	0
NA	Mucoromycotina	*Phycomyces blakesleeanus*			2	18	1
Chytridiomycota	NA	*Batrachochytrium dendrobatidis*	Biotroph	Animal pathogen	3	5	0

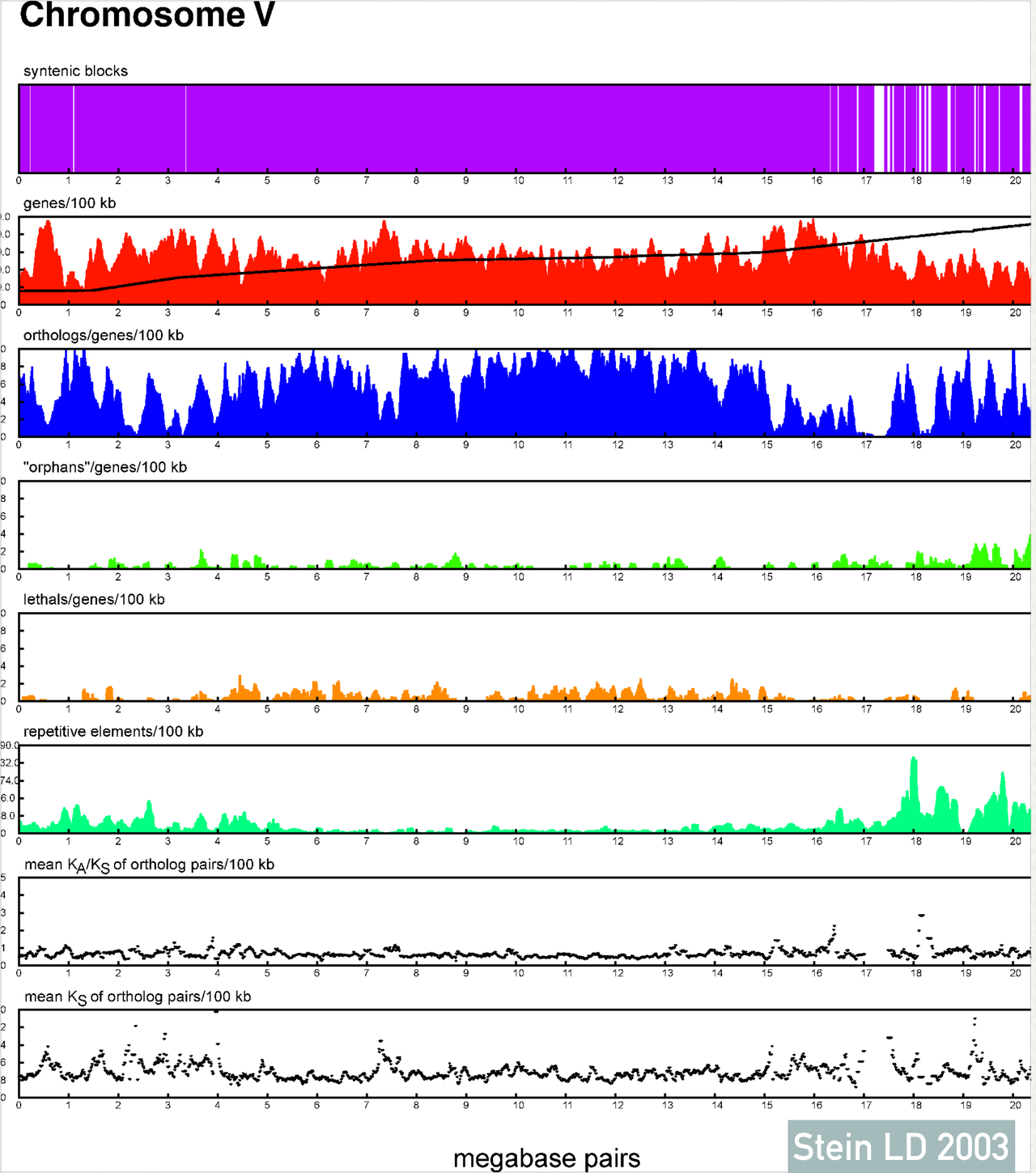
Kohler et al. Nat Gen. 2015


(a)
(b)
(c)
(d)
Pecoramyces sp. FX4B
Pecoramyces sp. S4B
Pecoramyces sp. FS3C
Pecoramyces sp. YC3
Pecoramyces ruminantium C1A
Orpinomyces sp. D3A
Orpinomyces sp. D3B
Orpinomyces sp. D4C
Neocallimastix californiae G1
Neocallimastix sp. G3
Neocallimastix frontalis Hef5
Feramyces austinii WSF2c
Feramyces austinii WSF3a
Anaeromyces sp. C3G
Anaeromyces sp. C3J
Anaeromyces sp. O2
Anaeromyces sp. G3G
Anaeromyces sp. NA
Anaeromyces robustus S4
Caecomyces sp. Iso3
Caecomyces sp. Brit4
Piromyces sp. B4
Piromyces sp. B5
Piromyces sp. A1
Piromyces sp. E2
Piromyces finnis
Chytriomyces sp. MP71
Rhizoclosmatium globosum
Gaertneriomyces semiglobifer Barr43
Entophlyctis helioformis JEL805
Gonapodya prolifera JEL478
CBM_10
CotH
CBM26
CBM_35
CBM_6
CBM-like
Glyco_hydro_8
Glyco_hydro_11
Glyco_hydro_48
Gal_Lectin
NRDD
CBAH
AAT
FeoA
FeoB_C
Pollen_allerg_1
Pur_ac_phosph_N
SASA
Stealth_CR1
Stealth_CR2
Stealth_CR3
Asp_protease_2
Asp_protease_5
CBM_25
Abhydrolase_5
PhosphMutase
RhamnogaL_lyase
Cthe_2159
COBRA1
Carb_bind
GH43_C
Glyco_hydro_39
Glyco_hydro_64
Glyco_hydro_115
YoeB_toxin
Met_gamma_lyase
Chitin_bind_1
CBM_1
Cellulase
COesterase
Glyco_transf_20
Glyco_transf_31
Glycos_transf_2
Gly_radical
Glyco_hydro_10
Glyco_hydro_43
Glyco_hydro_45
Glyco_hydro_114
Glyco_hydro_1
RicinB_lectin_2
Glyco_transf_34
Glyco_hydro_9
AAA_21
Esterase
Polysacc_deac_1
Glyoxal_oxid_N
SE
zf-4CXXC_R1
Allantoicase
IDO
MMADHC
OHCU_decarbox
OPA3
Ureidogly_lyase
BCDHK_Adom3
BCS1_N
Cmc1
Coa1
COX6B
Cyto_heme_lyase
Cytochrom_B561
NAD_binding_6
Cupin_4
Cupin_8
EBP
FPN1
HAGH_C
INSIG
Chromate_transp
Solute_trans_a
Suc_Fer-like
Glyco_hydro_46
Glyco_hydro_63
Glyco_hydro_63N
Glyco_transf_90
Mg_trans_NIPA
LigB
Lipid_DES
S1-P1_nuclease
Spo12
Sgf11
YCII
FAD_binding_7
FAD_binding_8
DNA_photolyase
8
6
4
2
0
Heatmap of the genome−wide enriched Pfam domains (natural logarithm)
Wang et al. Biorxiv 2018

Distribution of secreted plant cell wall degrading enzyme (PCWDE) across collection fungi to test how Tuber group interacts

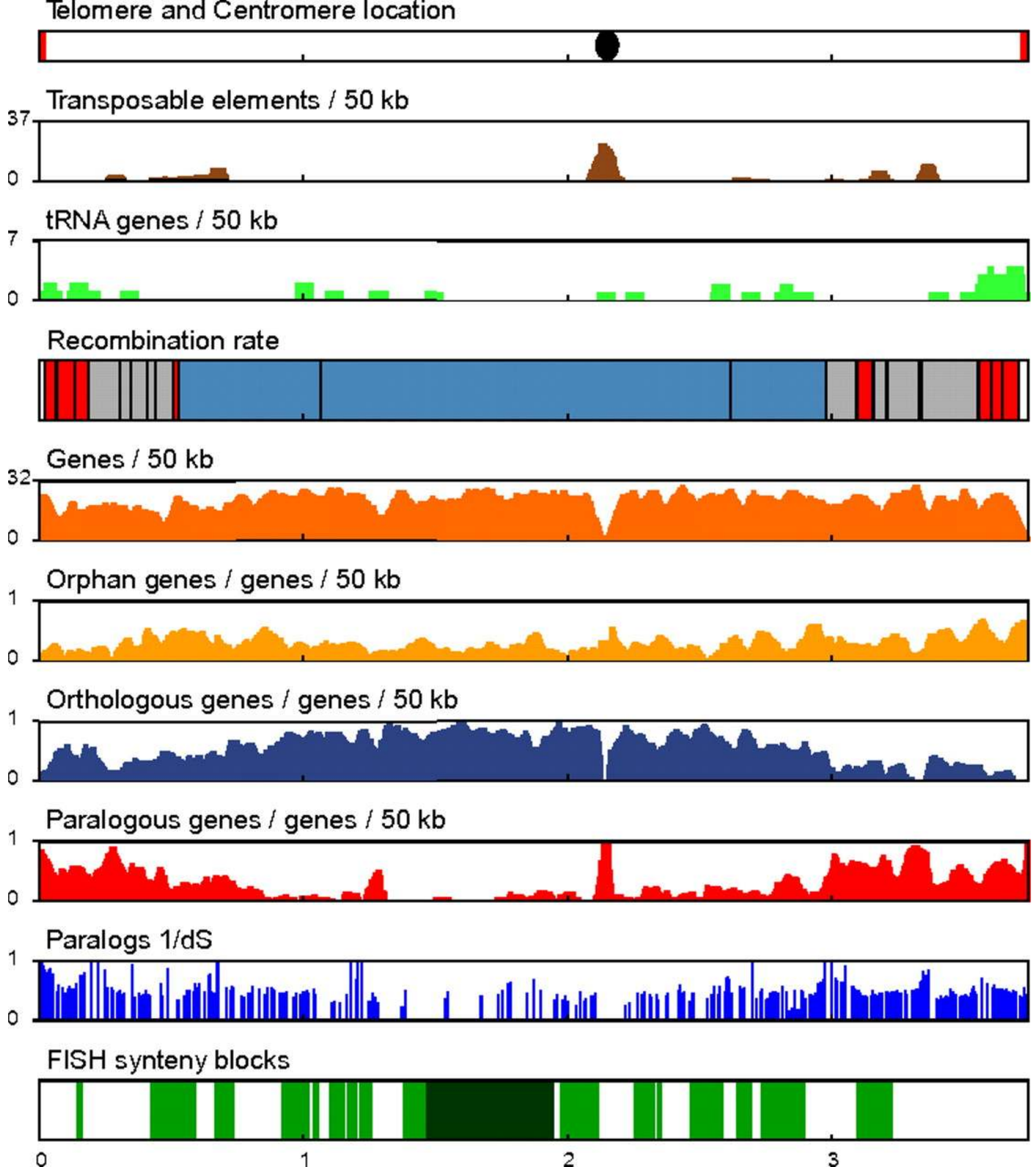


Black frames highlight Pezizomycetes taxa, whereas black arrows indicate Tuberaceae taxa and red dotted lines highlight PL1, PL4 and CE12 families. The ecology of each species is indicated at the right of the species abbreviations.



GENOME CONTENT WITH TRACKS

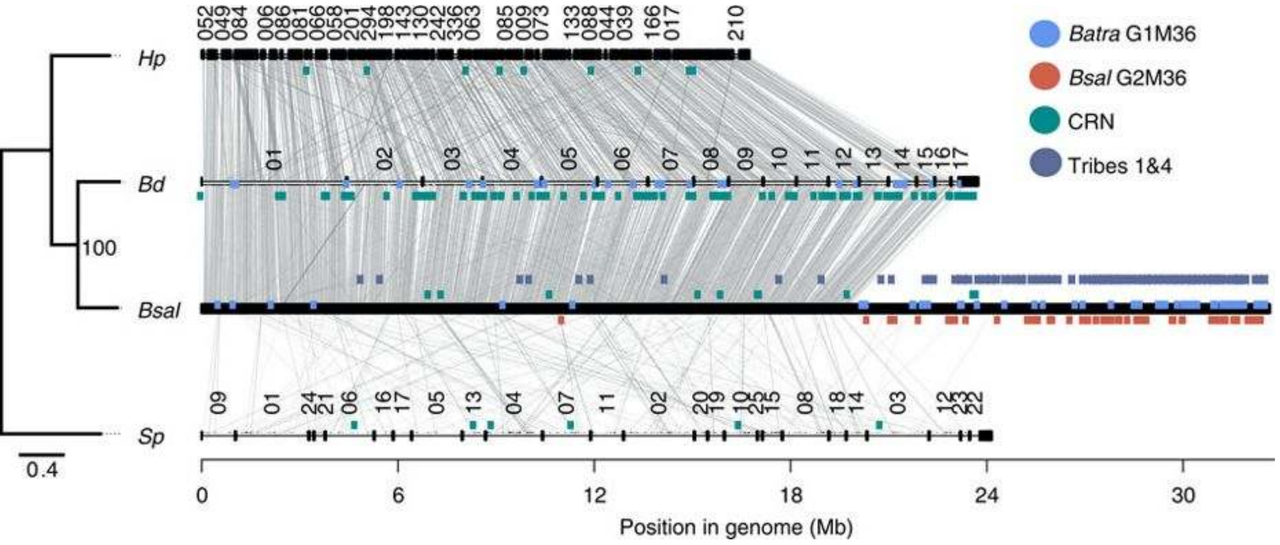
- C. briggsae gene content and features
- Show information about gene content, which genes are shared/unique
- Inverse relationship between Transposons and called genes.

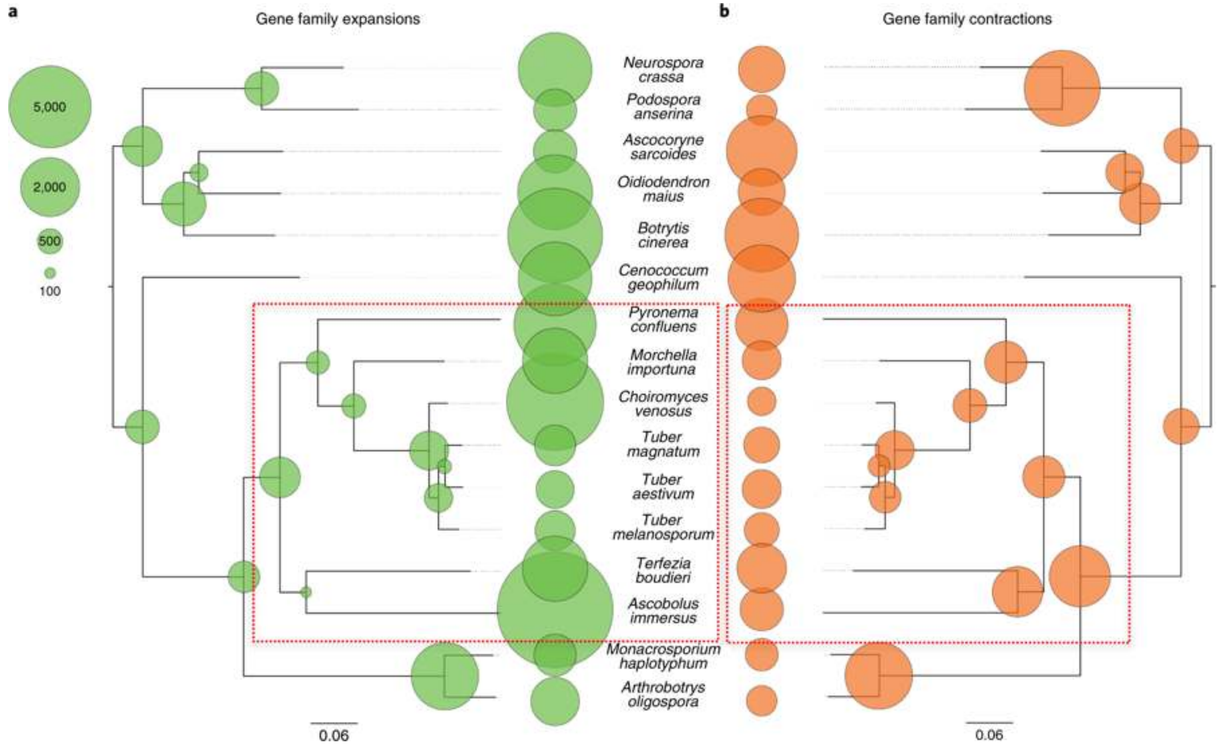


SIMILAR CHROMOSOME CONTENT PLOT

- Comparison of chromosome II for Mushroom *Coprinopsis*.
- Demonstrating other ways to compare content
- Notice the regions of high synteny and having orthologous genes are located in center
- While gene duplications (paralogs) are enriched at telomere

Stajich et al PNAS 2010





METHODS: COMPARATIVE GENOMICS

- Identifying Shared and Unique Genome content
- Synteny
- Whole genome alignment
- Gene family
  - Gene clustering - examine orthology and paralogy
- Functional content comparison

GENOME ALIGNMENT AND SYNTENY INFERENCE

Goals

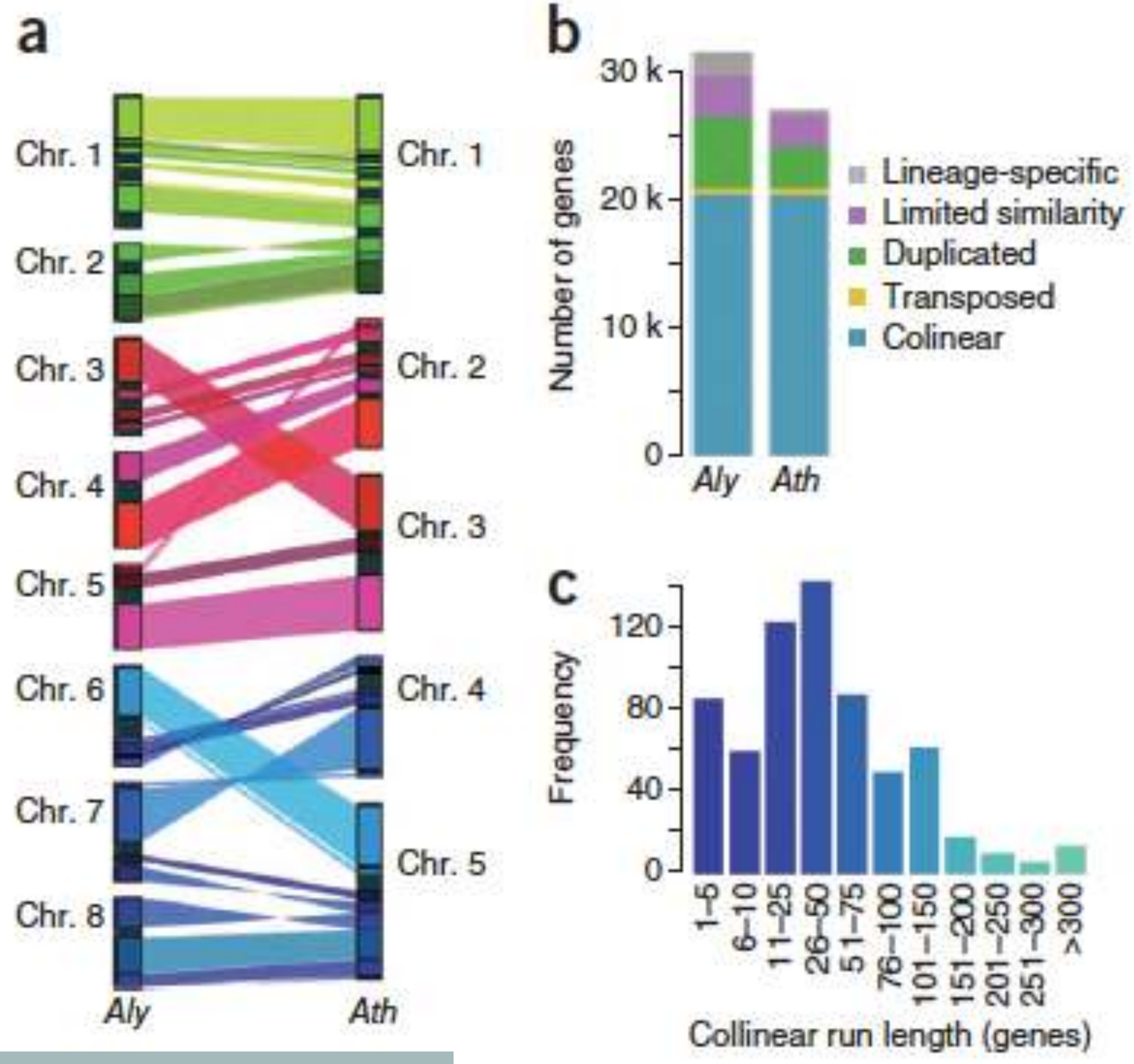
- Identify homologous segments between regions of genome

- Reconstruct evolutionary history for every nucleotide in the genome

Challenges: Genomes are not necessarily (rarely!) co-linear between species

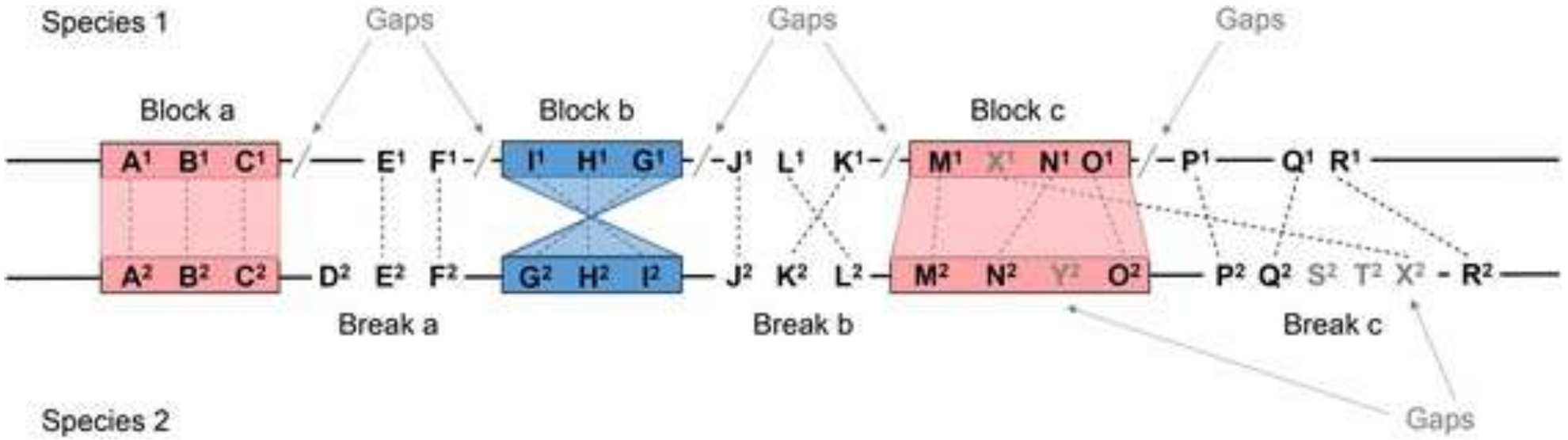
Identifying Insertion / Deletion / Rearrangements

Not feasible to just run a giant multiple alignment for all chromosomes



Hu et al Nat Genetics 2011

SYNTENY

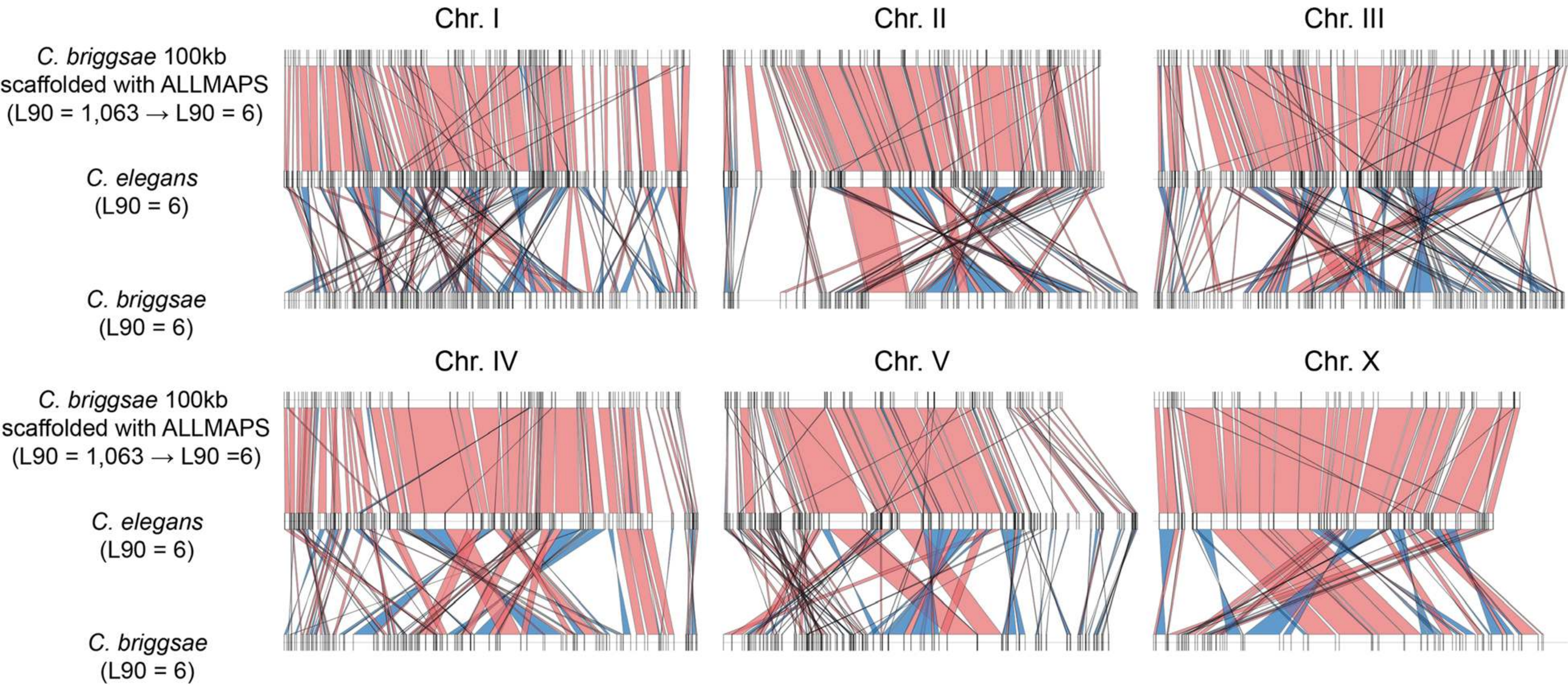


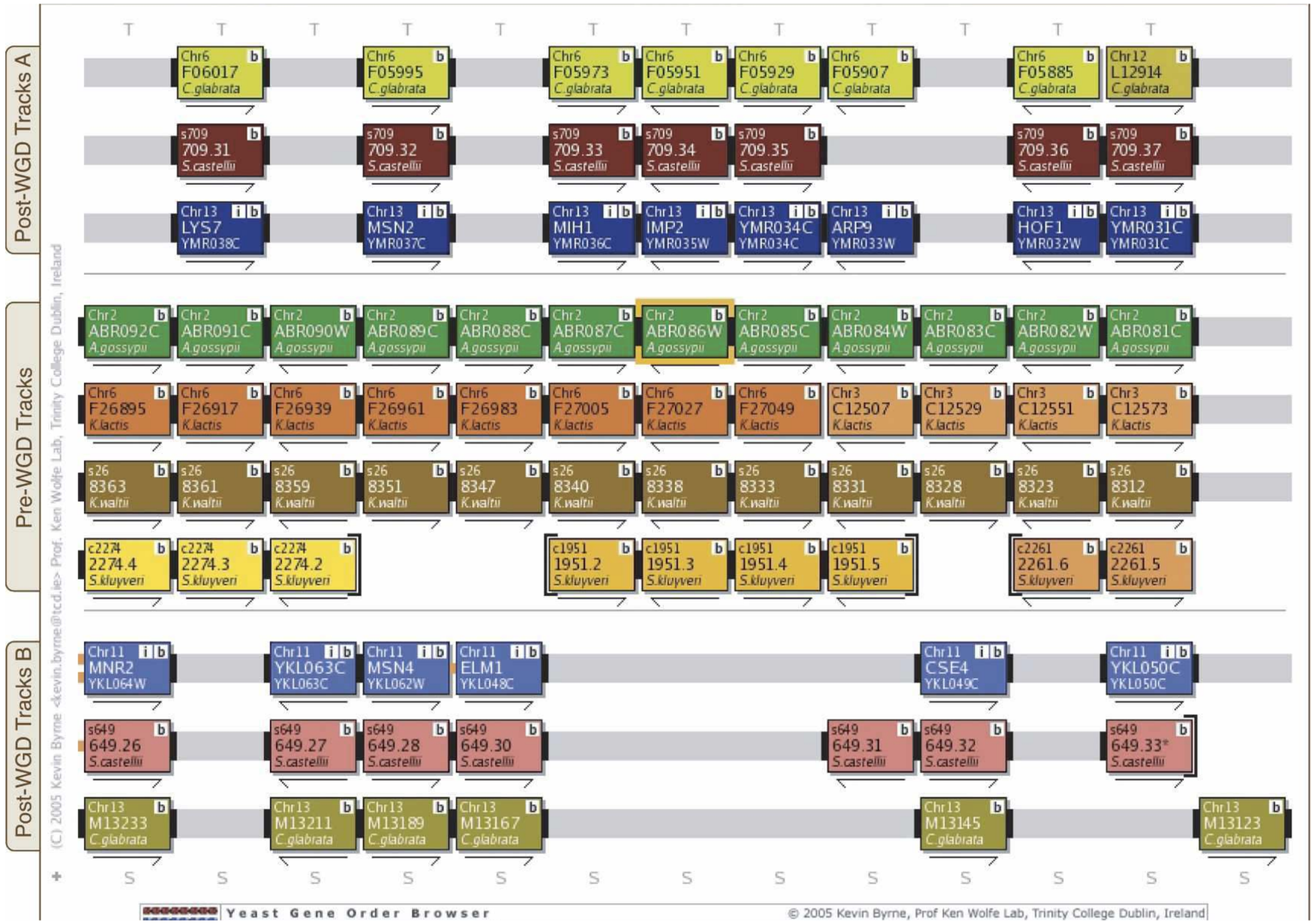
Synteny was originally defined as two more pairs of homologous genes on same chromosomal segment

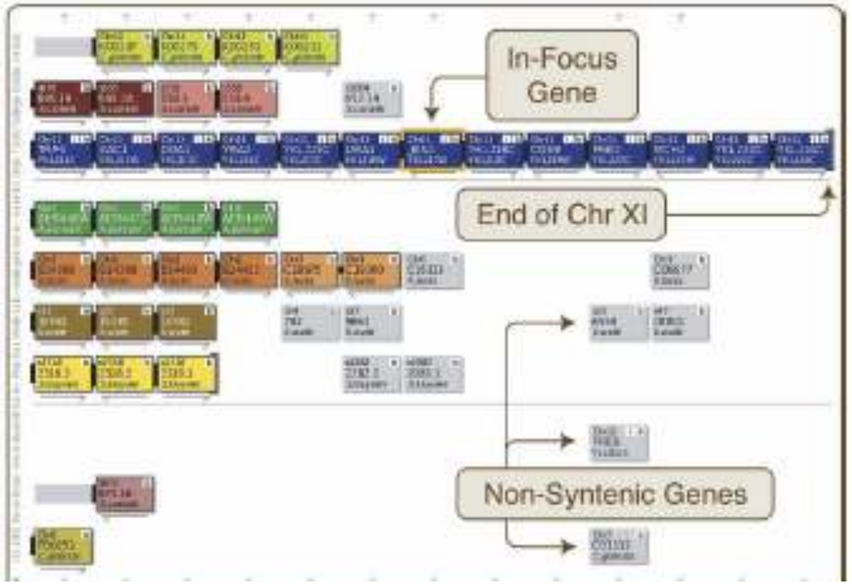
The stricter definition require collinearity of the genes and orientation

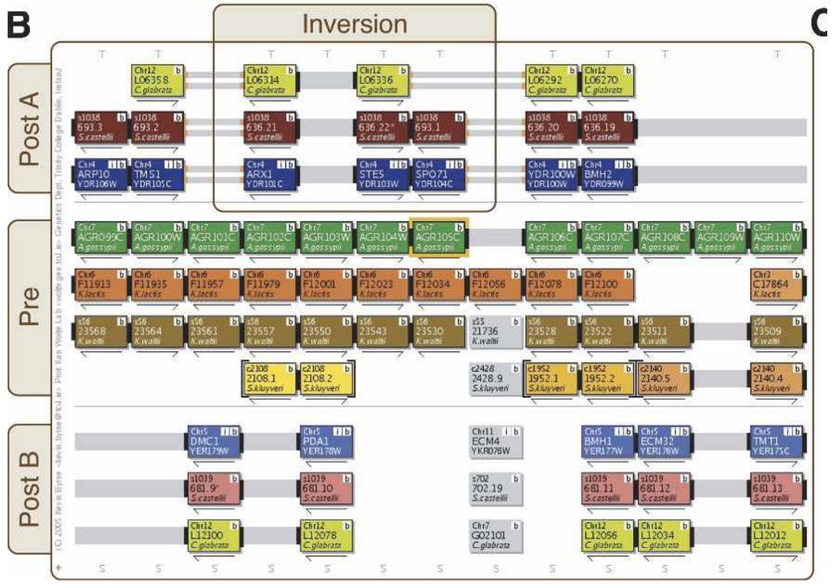
Requires detection of similar regions and chaining these into “runs” of genes

SYNTENY TO EXAMINE GENOME ORGANIZATION AND CONSTRAINT



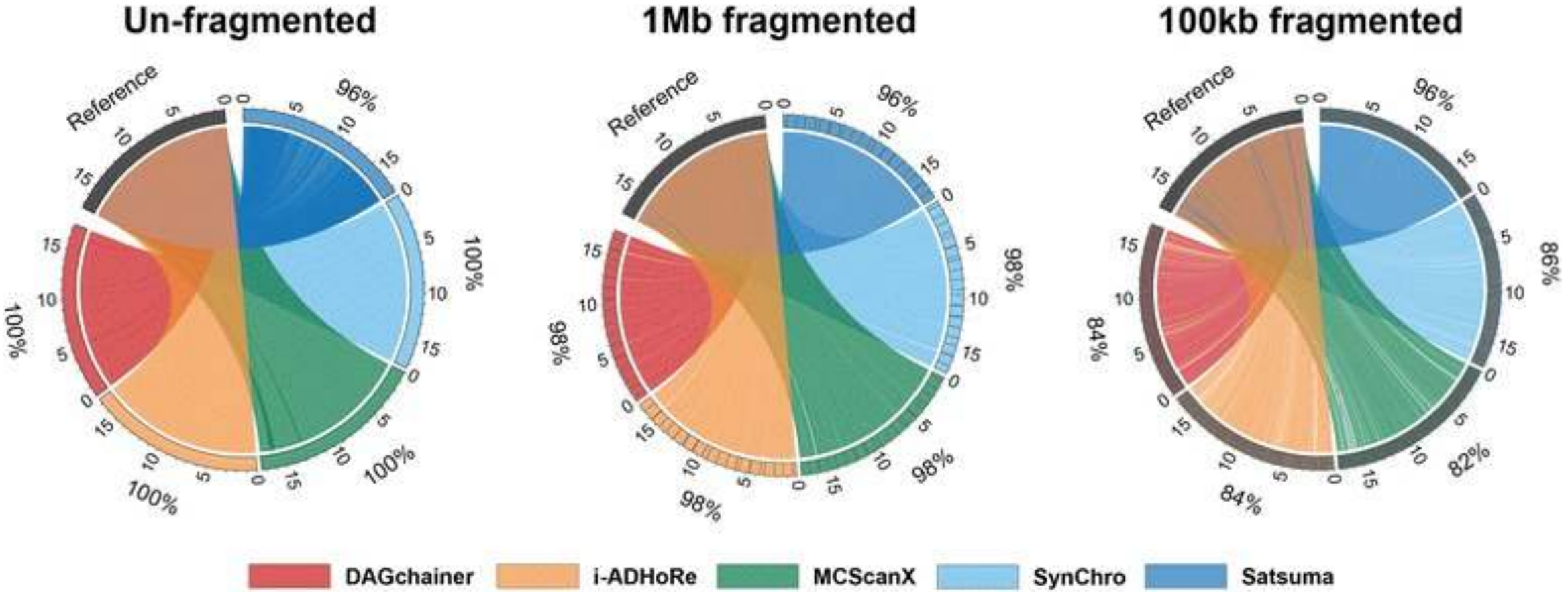






Yeast Gene Order Browser: Automated and curated synteny Gold standard in synteny and orthology comparisons in Saccharomycotina yeasts - http://ygob.ucd.ie/

RECOVERY OF SYNTENIC REGIONS BY DIFFERENT TOOLS ON SIMULATED FRAGMENT C.ELEGANS CHROMOSOME



TOOLS FOR SYNTENY COMPUTATION

Gene or gene anchor based

i-ADHoRe - http://bioinformatics.psb.ugent.be/beg/tools/i-adhore30 Multi-way species compare

"highly sensitive software tool to detect degenerated homology relations within and between different genomes."

DAGchainer- http://dagchainer.sourceforge.net/ Multi-way species compare

"computes chains of syntenic genes found within complete genome sequence"

Mercator- https://www.biostat.wisc.edu/~cdewey/mercator/ Multi-way species compare

constructs orthology maps from protein anchored maps followed by refinement from multiple alignment of DNA

TOOLS FOR SYNTENY COMPUTATION AND ALIGNMENT

Whole genome alignment based

Satsuma2 - https://github.com/bioinfologics/satsuma2 Pairwise

"FastFourierTransform cross-correlation based synteny aligner, to reliably align large and complex DNA sequences providing maximum sensitivity (to find all there is to find), specificity (to only find real homology) and speed"

MUMMER - https://github.com/mummer4/mummer (v4) and http://mummer.sourceforge.net/ (v3) Pairwise

"a system for rapidly aligning entire genomes, whether in complete or draft form"

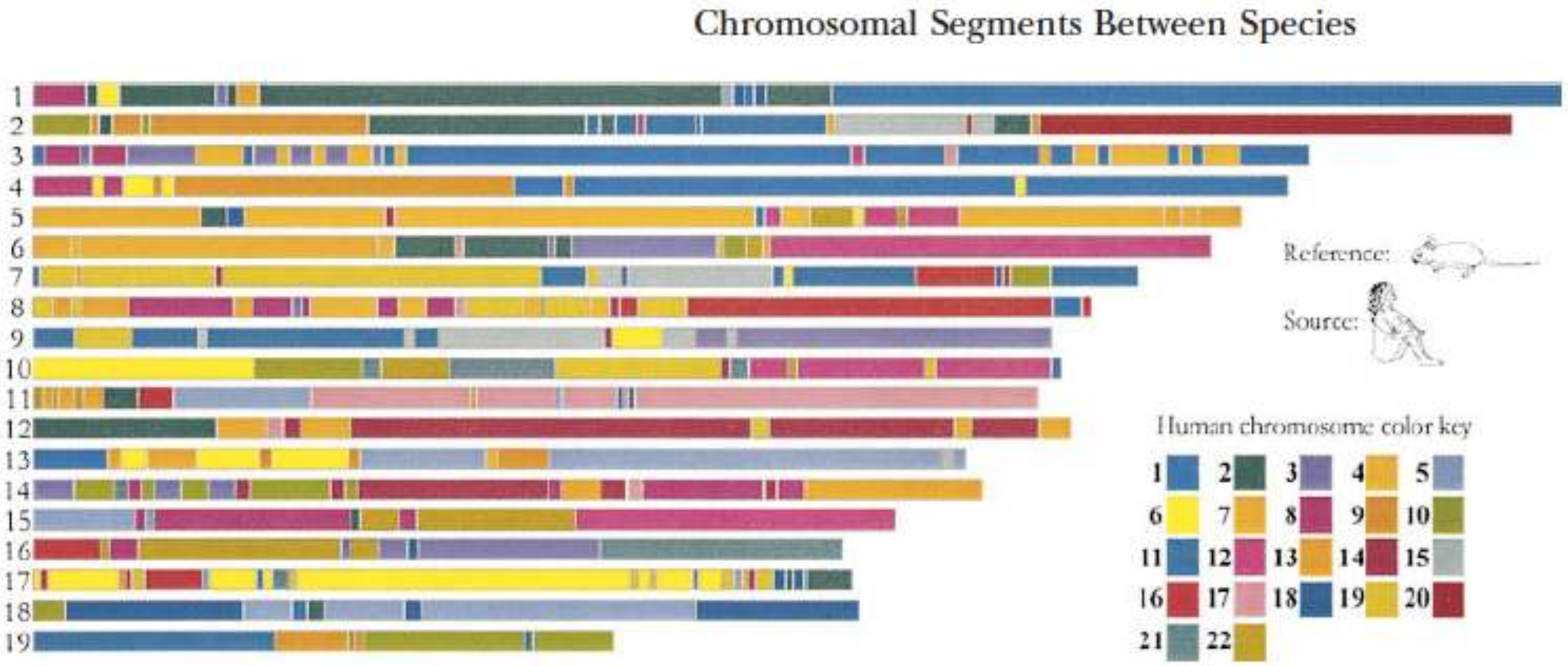
LAST - http://last.cbrc.jp/ Pairwise

"LAST finds similar regions between sequences, and aligns them."

Minimap2 - https://lh3.github.io/minimap2/ Pairwise

A versatile pairwise aligner for genomic and spliced nucleotide sequences

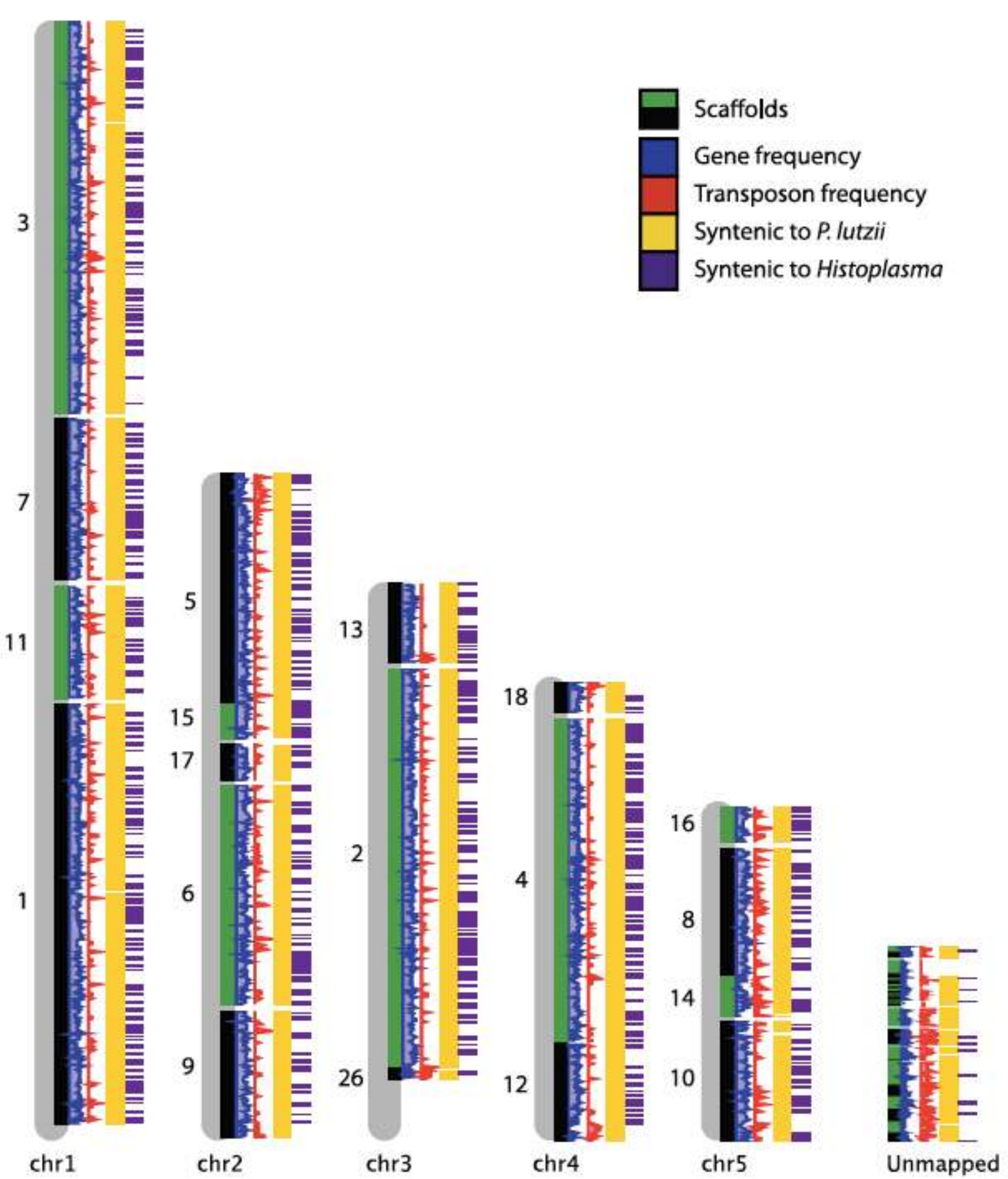
SYNTENY AND HOMOLOGY: MOUSE HUMAN SHARED GENE CONTENT & SYNTENY

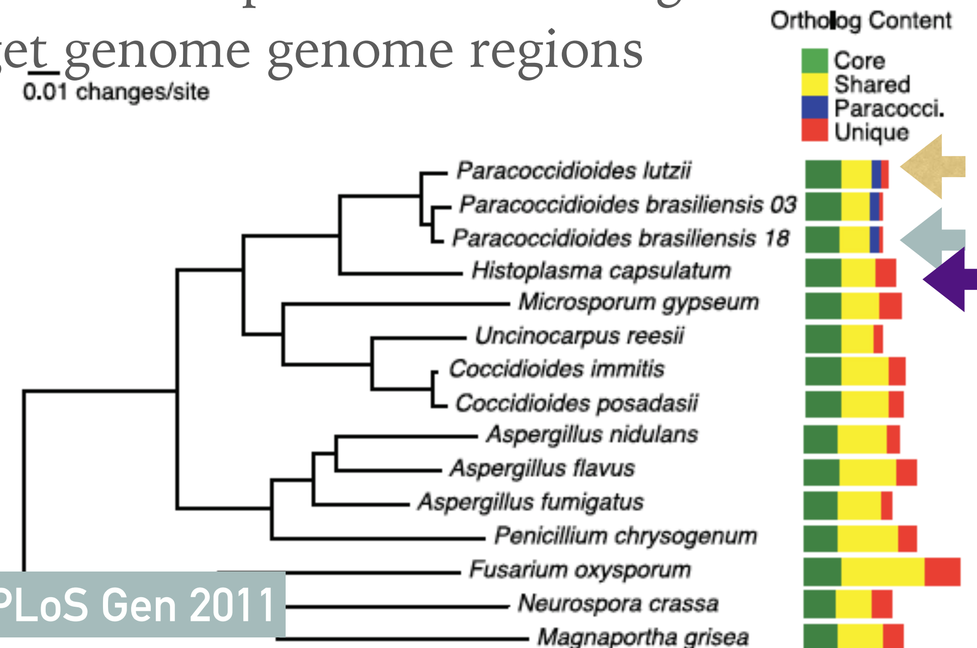


Kumar et al. Genetics 2001

SUMMARIZING SHARED GENOME CONTENT

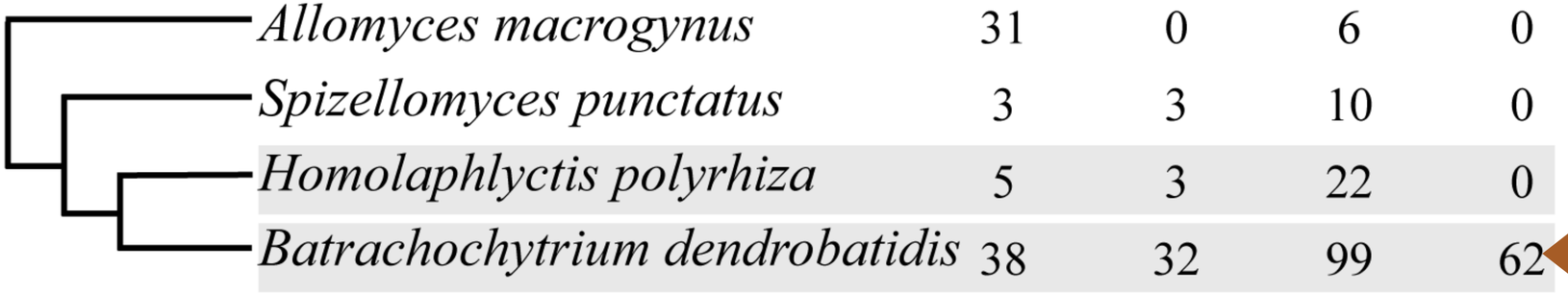
- 3 fungi are compared here: Pb18 supercontigs (grey) to the five chromosomes from the optical map (chr1-5).
- Summary statistics about gene and transposon density
- Synteny blocks expressed as coverage of the target genome genome regions





Dejardins et al PLoS Gen 2011

EXAMINATION OF AN EXPANDED GENE FAMILY WITH FURTHER FOLLOWUP OF GENE TREES



	M36	S41	Asp	CRN
Allomyces macrogynus	31	0	6	0
Spizellomyces punctatus	3	3	10	0
Homolaphlyctis polyrhiza	5	3	22	0
Batrachochytrium dendrobatidis	38	32	99	62

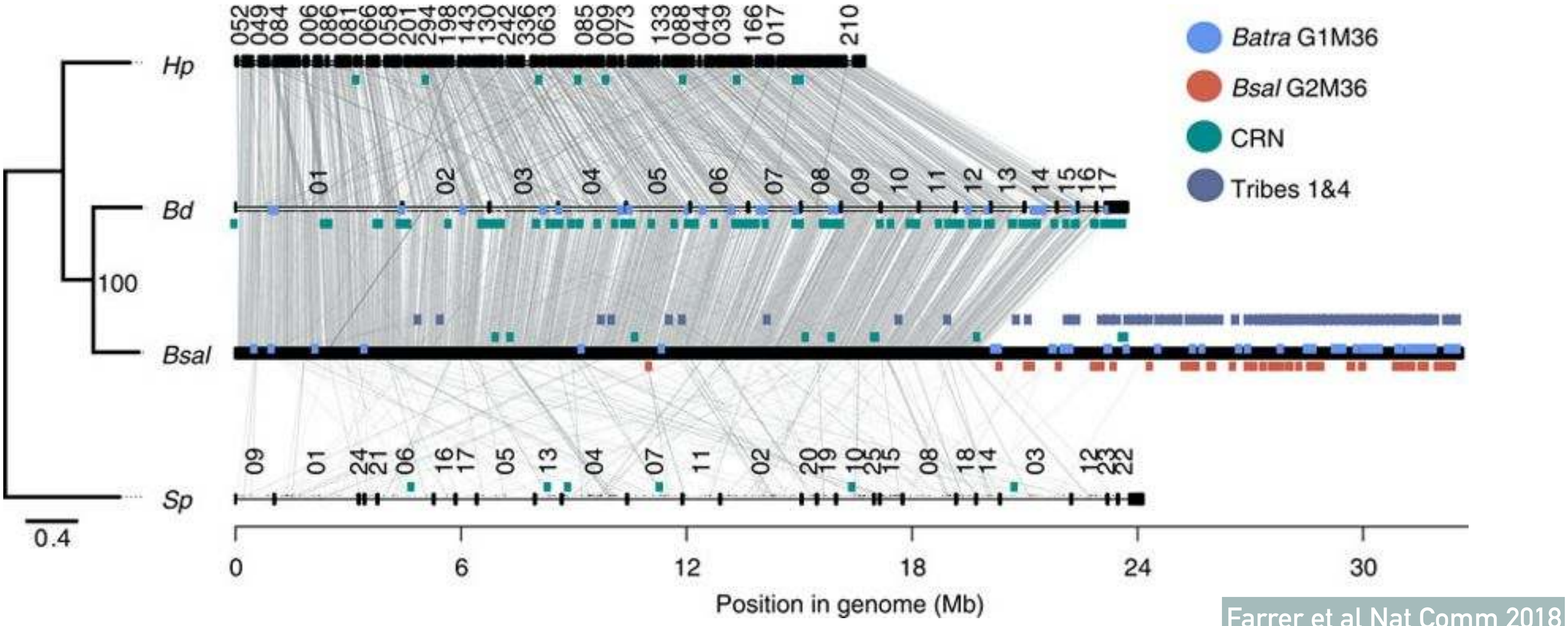
pathogen

Examination of all Pfam and metalloprotease domains among the available chytrids found only a few which were expanded in the pathogenic lineage

M36 Gene Tree

Joneson et al PLoS Pathogens 2011

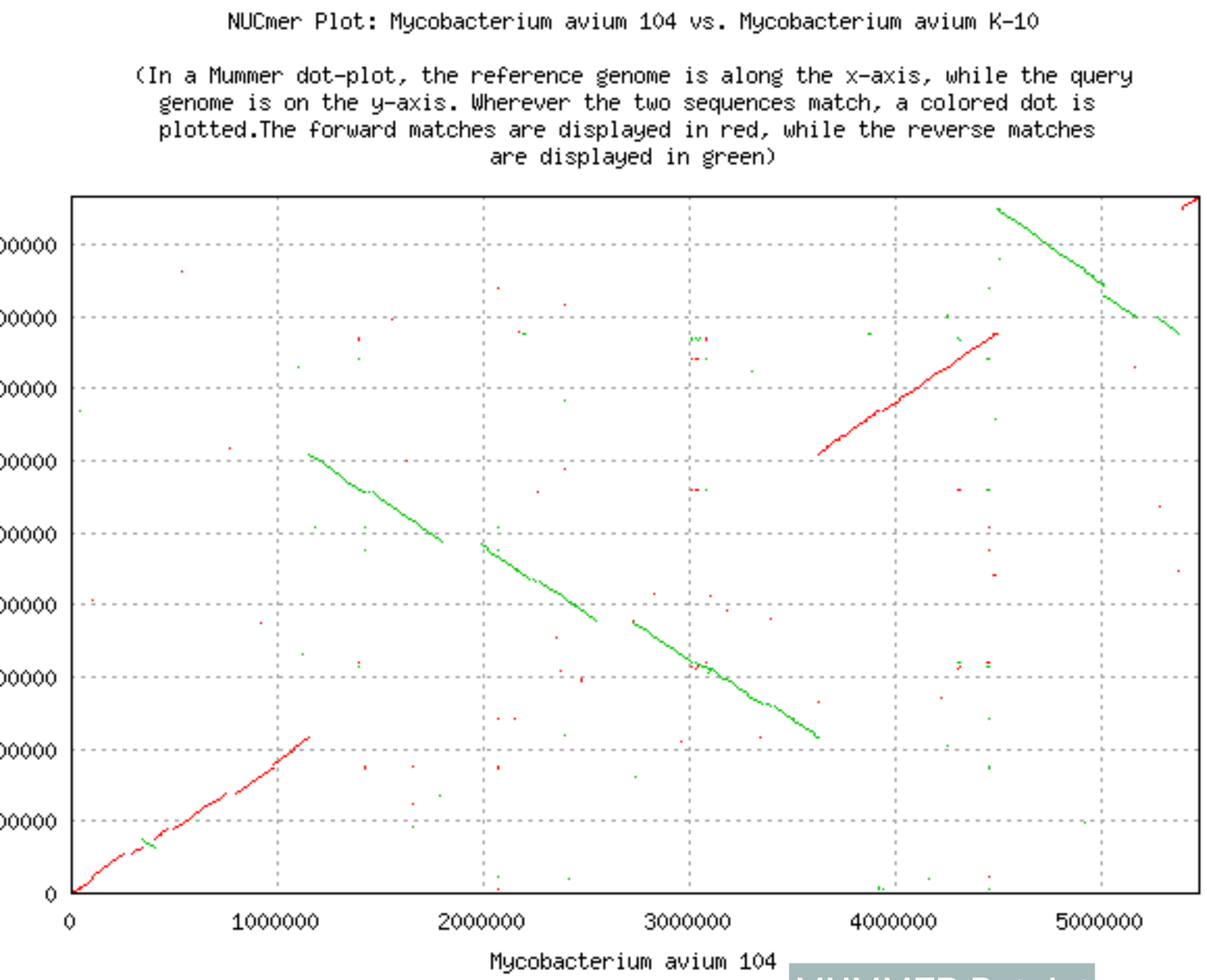
SYNTENY, NOVEL EXPANSION OF M36 FAMILY IN B. SALAMANDRIVORANS – AMPHIBIAN PATHOGEN



Farrer et al Nat Comm 2018

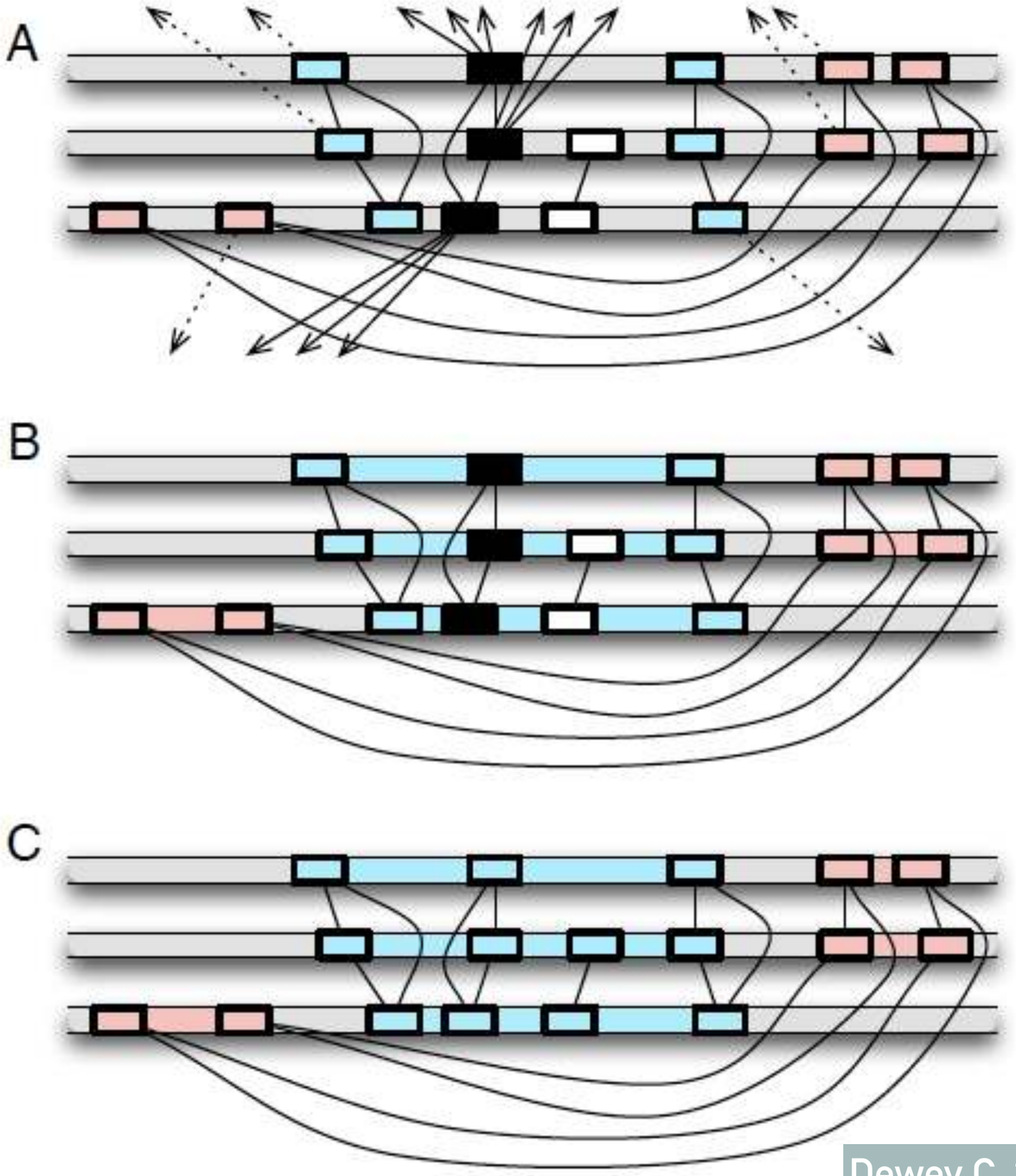
WHOLE GENOME ALIGNMENT

WORD MATCHING & ALIGNMENTS



MUMMER Dotplot

- Word-based, fixed or variable size matching
- **MUMMER**, **LAST**, LASTZ, BLASTZ, YASS
  - MUMMER: Fast lookup with suffix tables by building an index query and target sequences and apply fast suffix-table matching “ultra fast:
  - LAST finds initial matches based on their multiplicity, instead of using a fixed length (e.g. BLAST uses 11-mers). To find these variable-length matches, it also uses a suffix array
- Generally implemented as a pairwise analysis but can be hierarchically chained to support multiple taxa
- Annotation free - use sequence matches to define regions to anchor alignments and extend match region



Dewey C. 2006

ANCHORING AND ALIGNING – MERCATOR

- Use other features to draw connections between
- Exons - protein coding regions in particular - can be matched between species
- Extend blocks based on shared gene order
- Targeted sequence alignment in the regions that are considered matching based on these features
- Mercator (Dewey and Pachter)

WHOLE GENOME ALIGNMENT

Examine evolutionary history of every base in the genome

Questions that can be answered

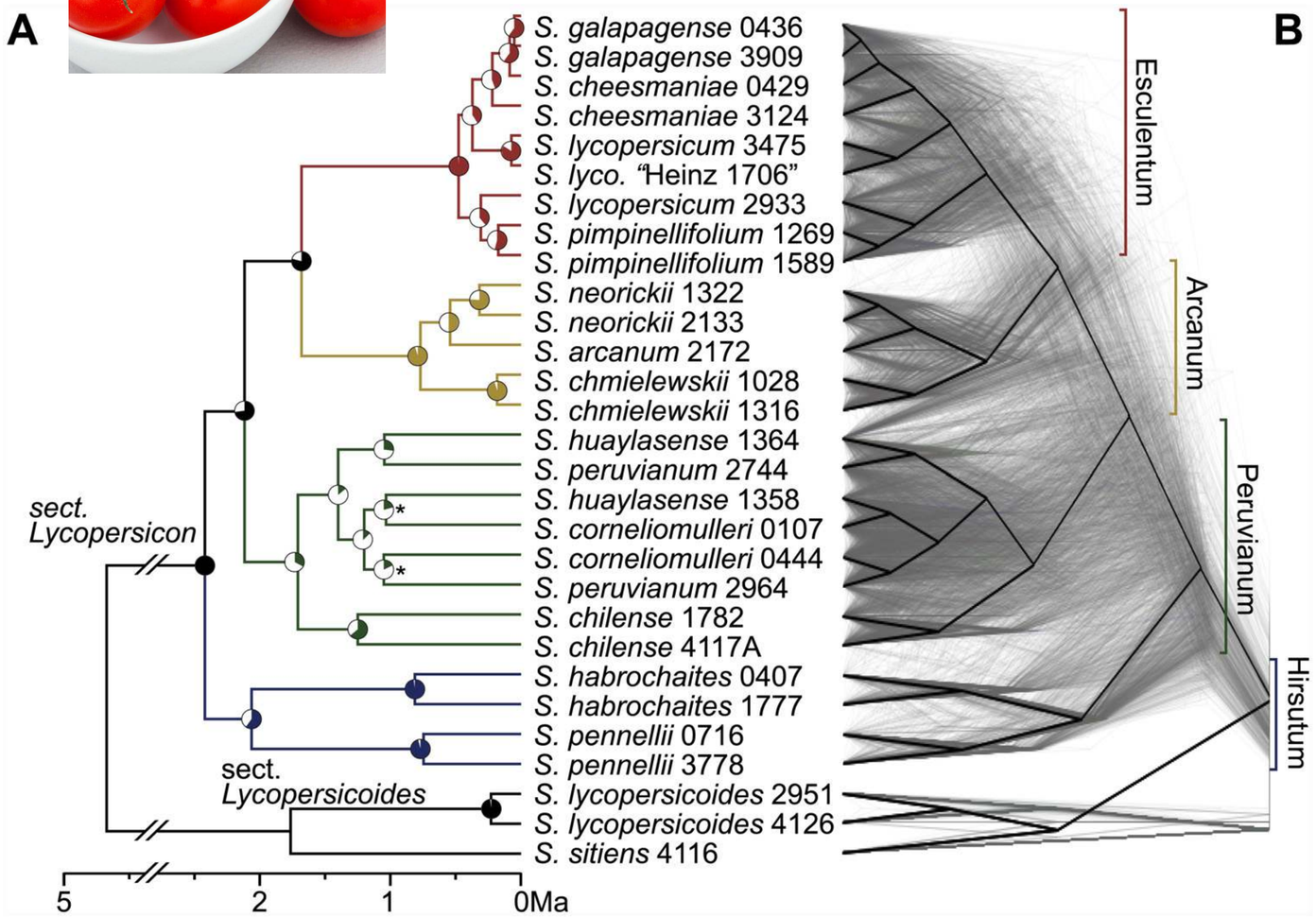
Rates of evolution across the genome - using a sliding windows or evaluating each gene.
Tests for directional selection

Individual gene trees from these alignments using subsampled windows

Identify unique regions and shared insertion/deletions

Testing for constraint / co-evolution

EXAMINING SPECIES TREE AND GENE/ WINDOW PHYLOGENY FROM WGA

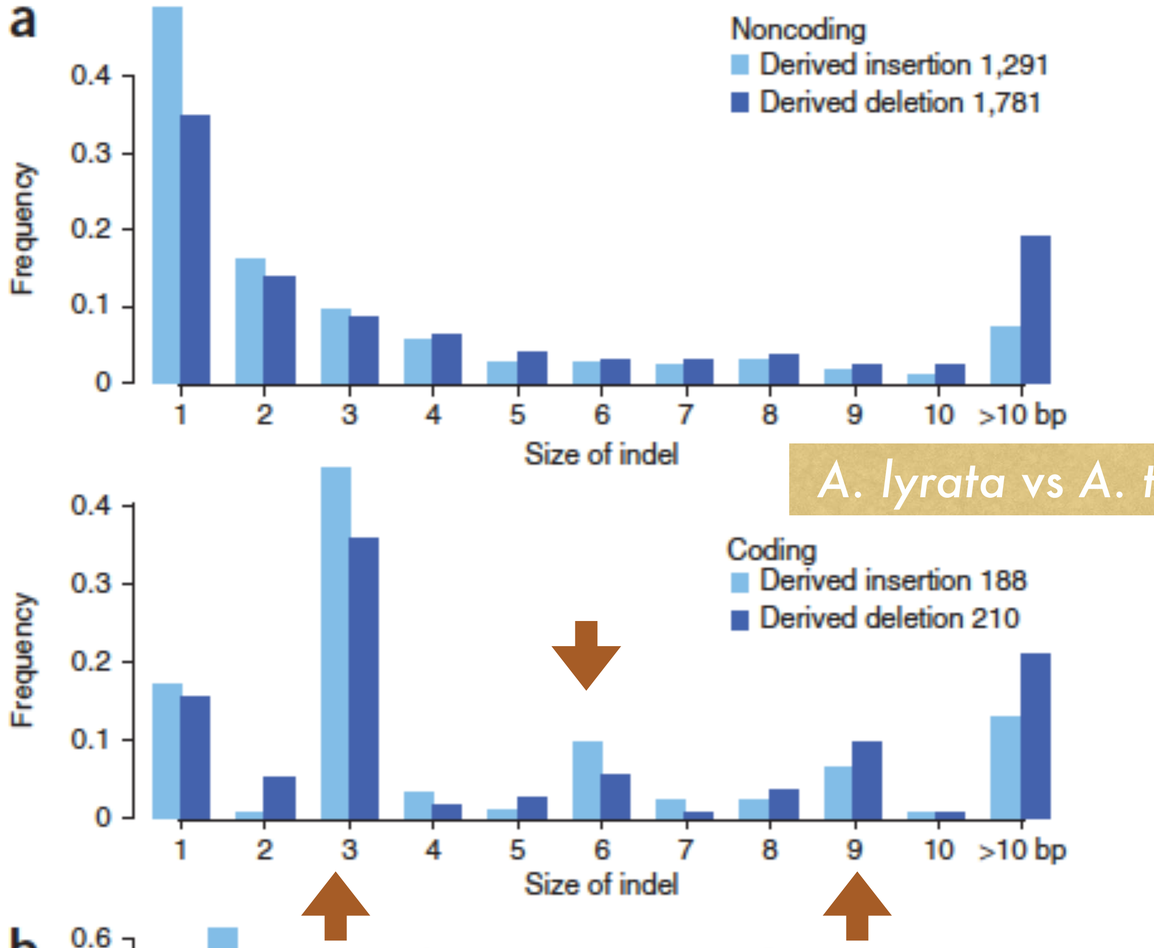


Solanum sect. *Lycopersicon*)

- Using a whole genome alignment (or in this case of transcriptome alignments) to examine consistency of phylogenetic signal across the genome.
- Individual gene trees can be inferred from sliding windows across the whole genome/transcriptome alignment
- Examine the consistency of these gene trees in a Cloudogram - drawing these as patterns impose on species tree.

Pease et al PLoS Biol 2016

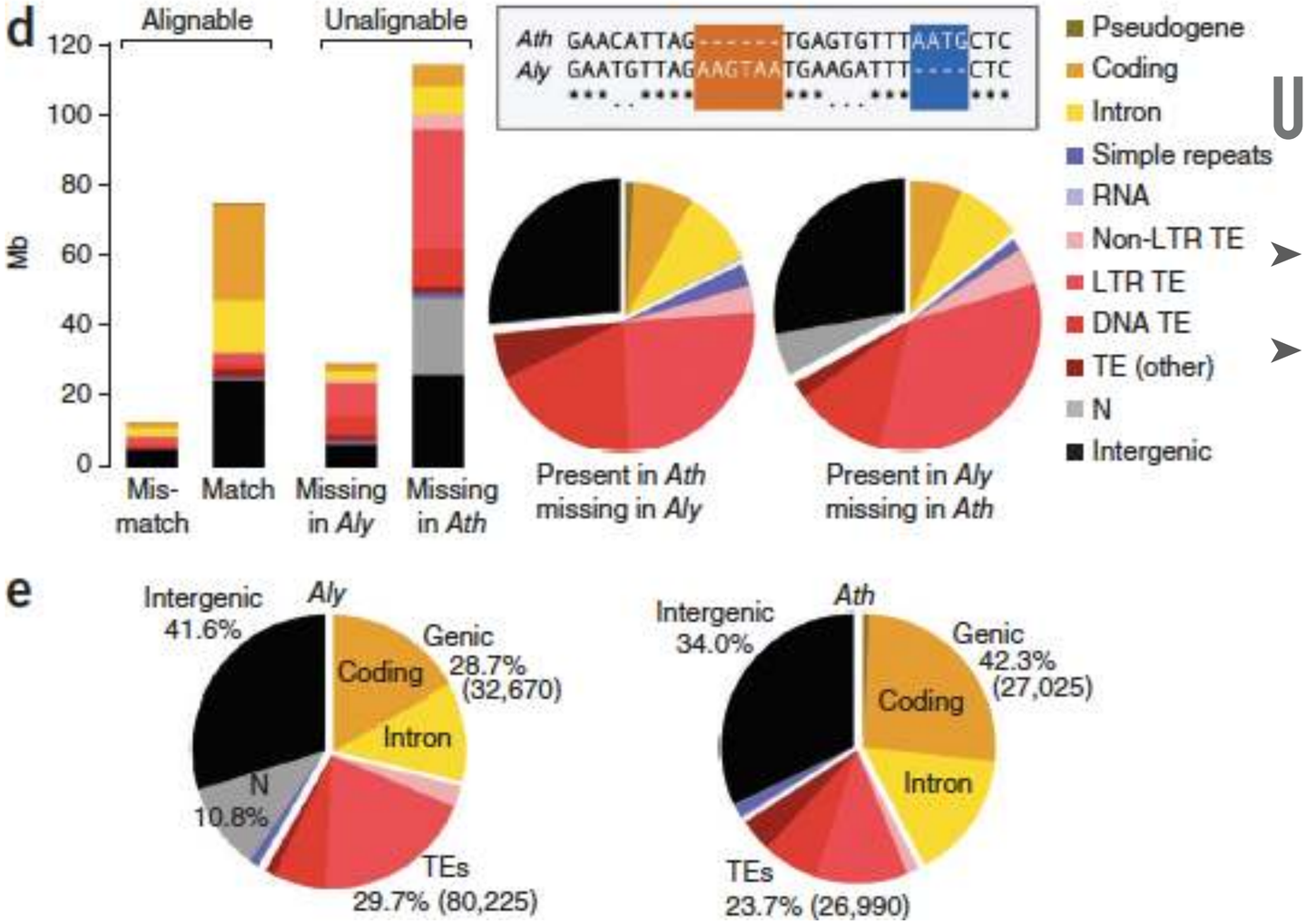
GENOME WIDE DIVERGENCE PATTERNS



A. lyrata vs *A. thaliana*

Hu et al Nat Genetics 2011

- Examination of insertion/deletion differences between two *Arabidopsis* (plant) genomes.
- Comparing the sizes of INDELs reveals constraint on the size of changes in coding regions
  - "Derived" status was inferred by using the 95 *A. thaliana* individuals and only considering a site that is fixed in the *A. thaliana* population. *A. lyrata* allele assumed to be ancestral allele when these are different.



Arabidopsis lyrata genome provides comparative context to how genome size can change

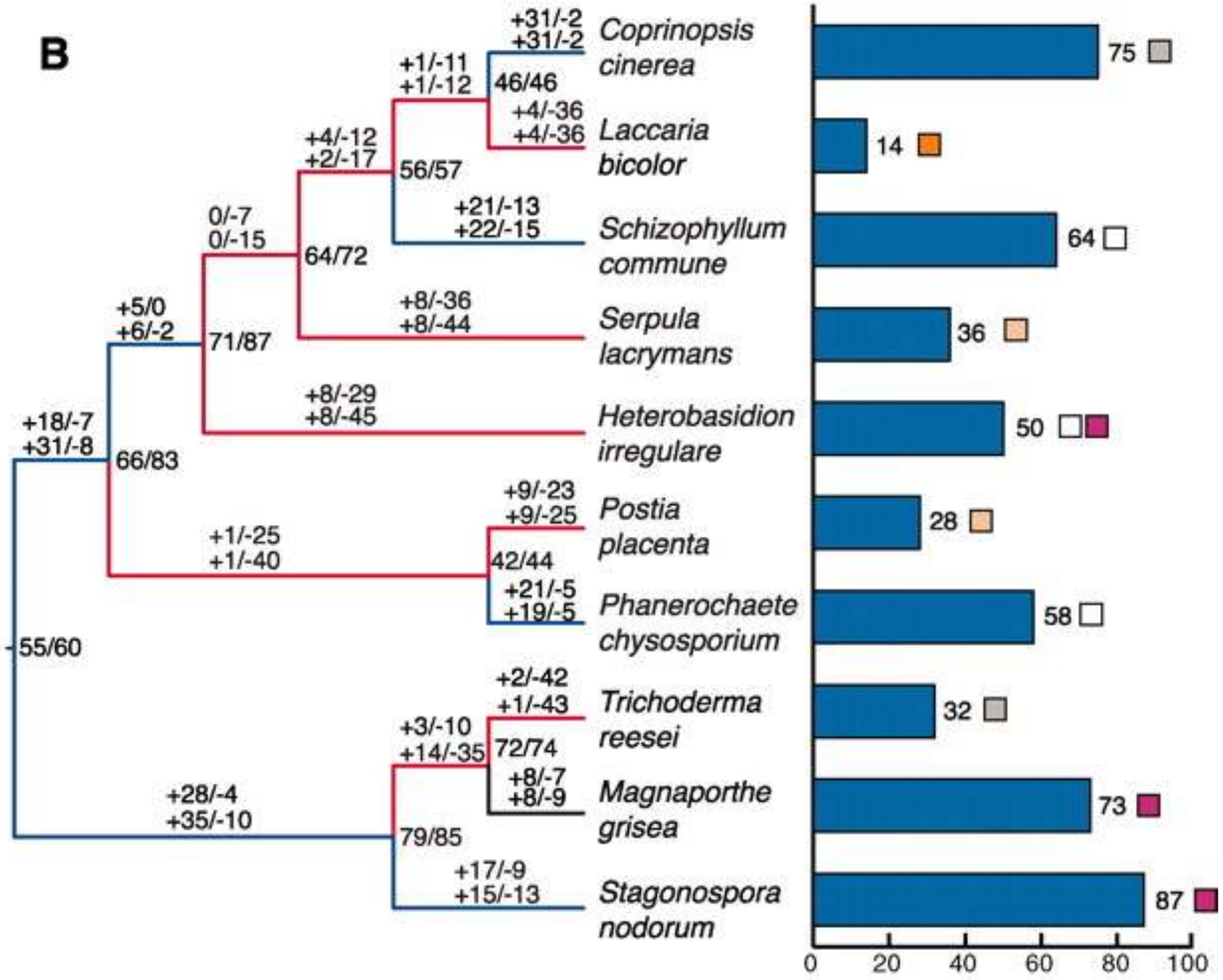
Hu et al Nat Genetics 2011

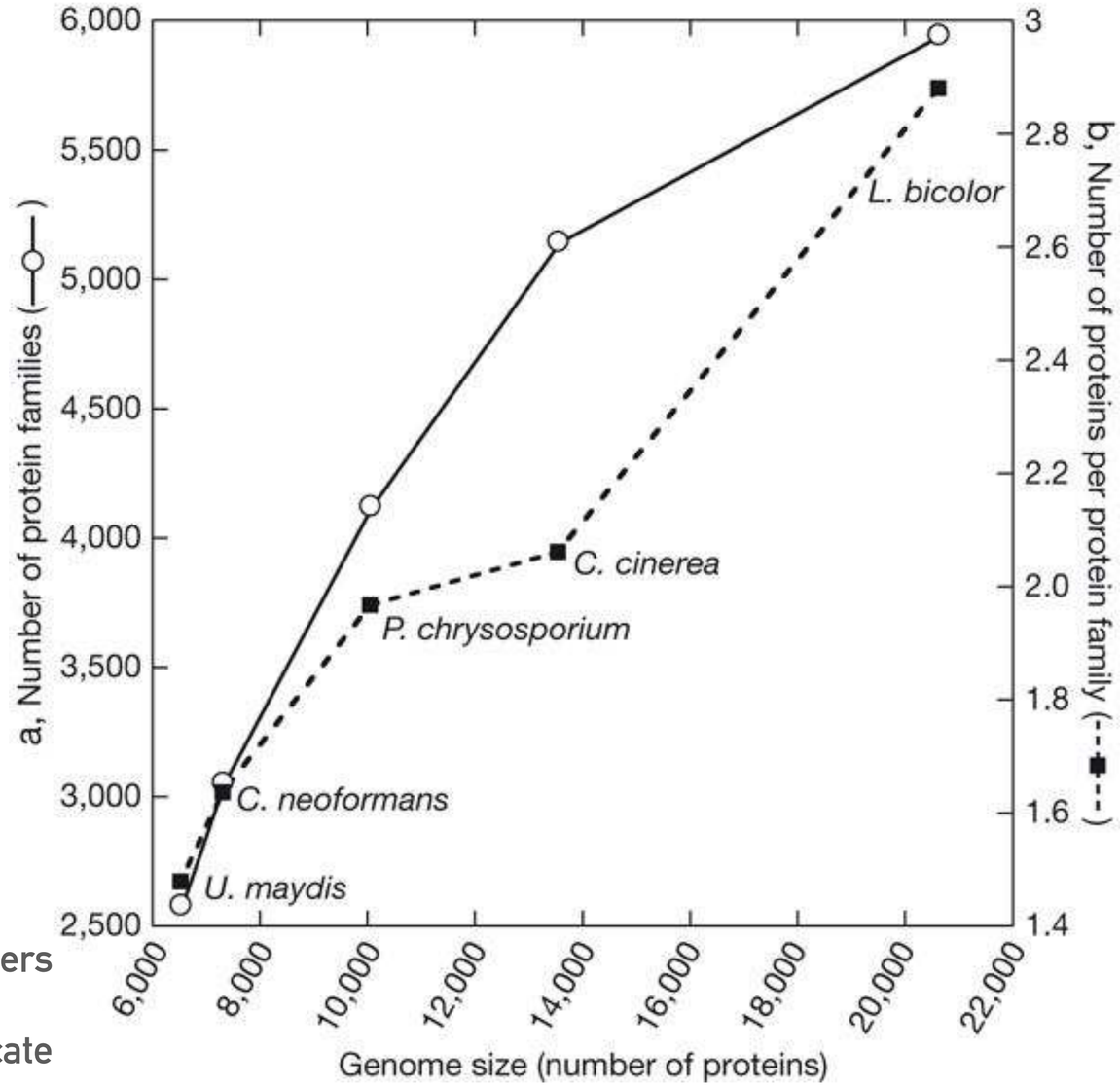
UNALIGNABLE REGIONS

- Finding unalienable regions
- Where did they come from?

"Unalignable sites can be considered as present in one species and absent in the other, as shown in the boxed sequence diagram; matches are indicated by asterisks, and mismatches by periods. The histogram on the left indicates the absolute number of unalignable sites, and the pie charts in the middle compare their relative distribution over different genomic features."

GENE FAMILY SIZE EVOLUTION



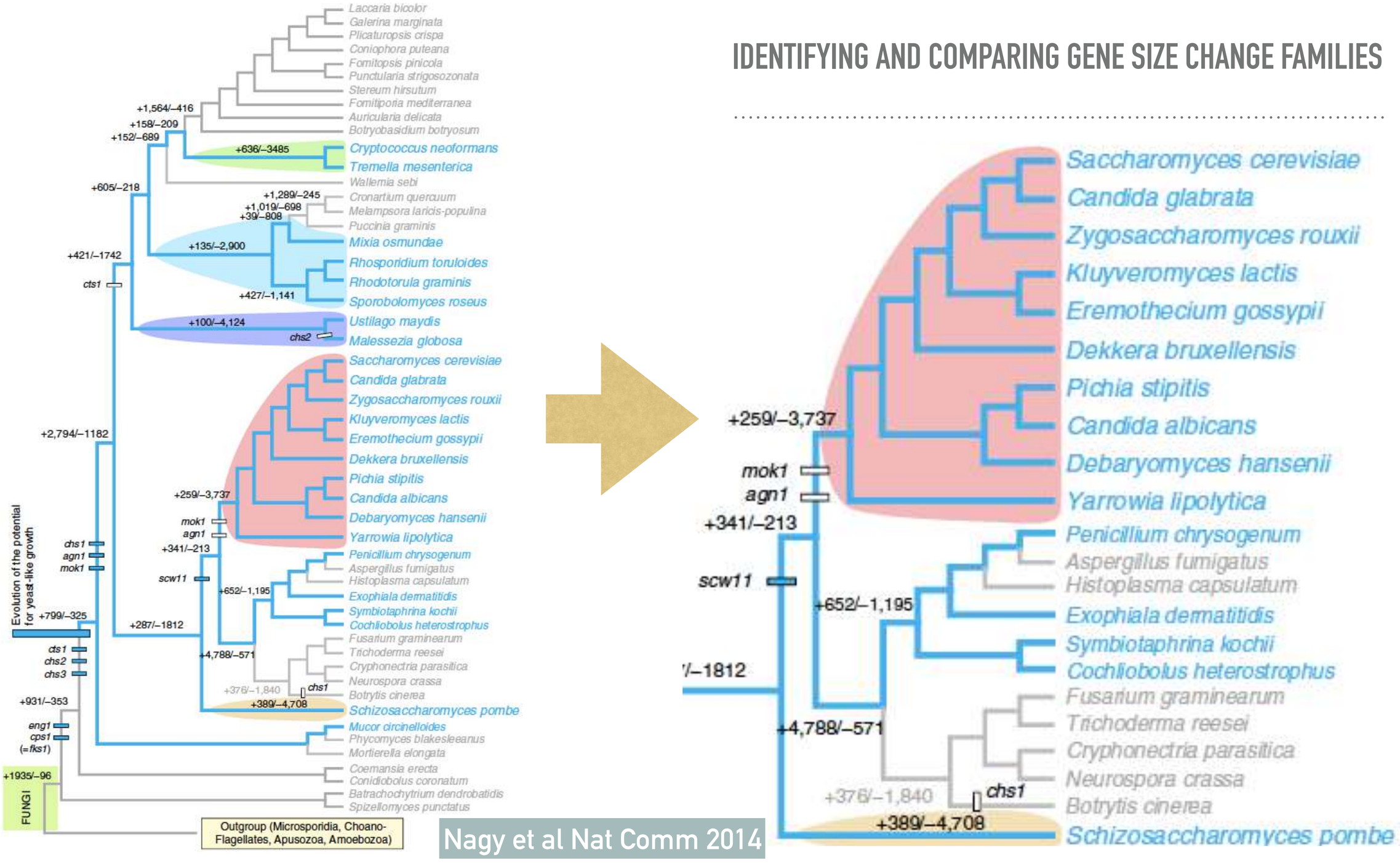


Gain / Loss Numbers at nodes and along branches indicate estimated copy numbers for ancestral species and ranges of gains and losses, respectively, estimated by using 90 and 75% bootstrap thresholds for gene trees in reconciliations. Bars indicate copy numbers in sampled genomes.

Eastwood et al Science 2011

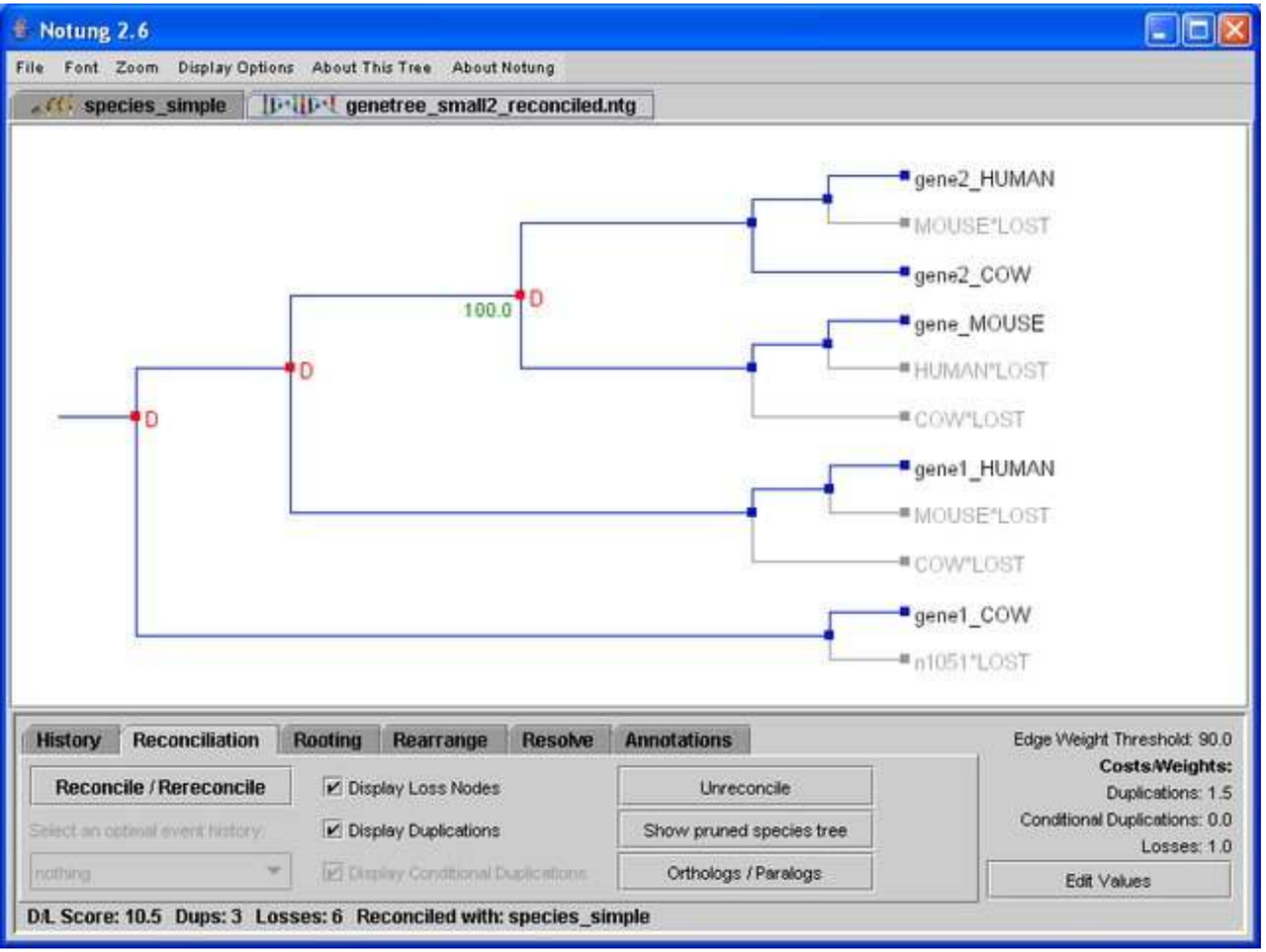
Martin et al Nature 2008

IDENTIFYING AND COMPARING GENE SIZE CHANGE FAMILIES



Nagy et al Nat Comm 2014

HOW TO INFER LOSSES OR GAINS?

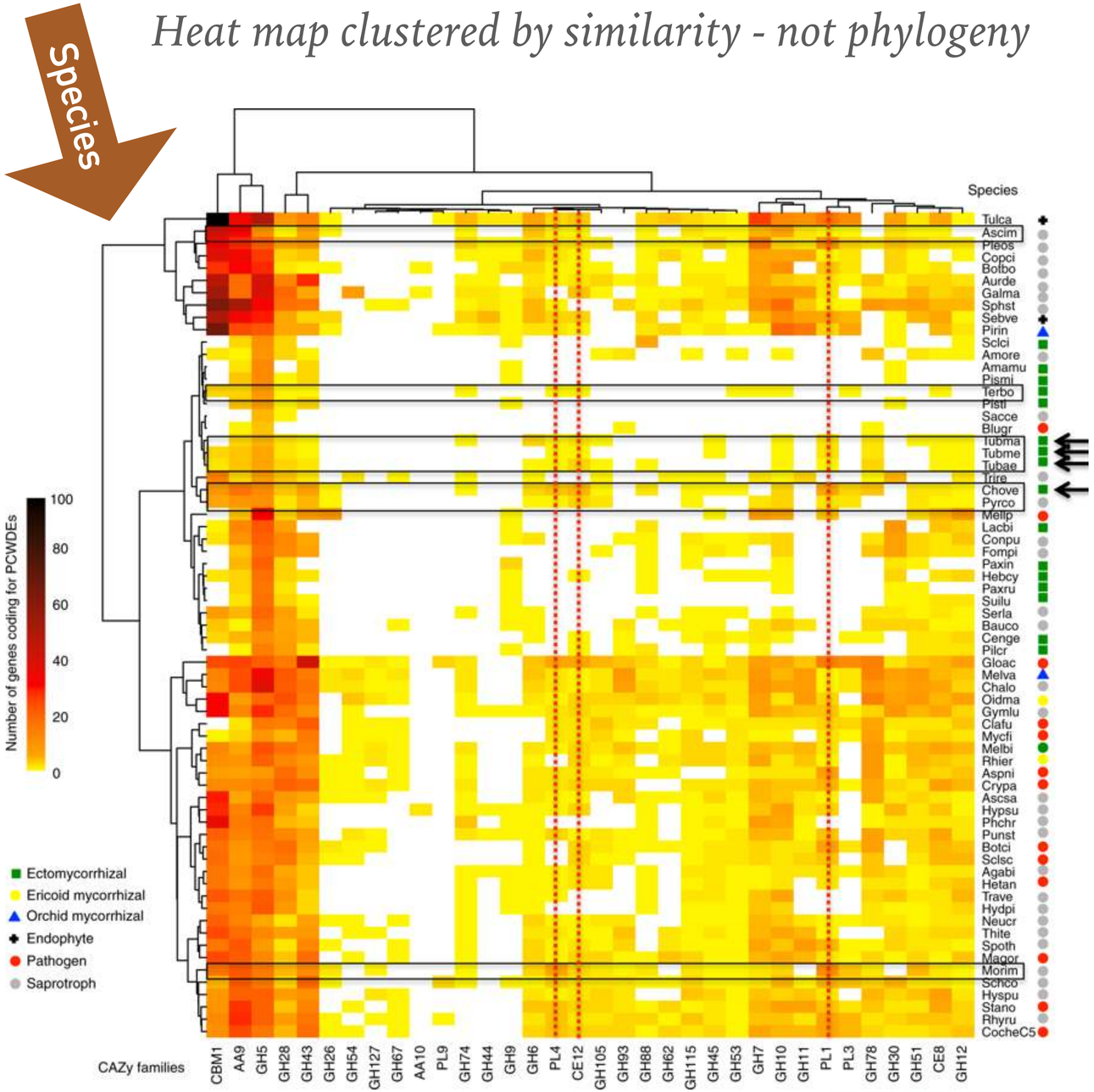


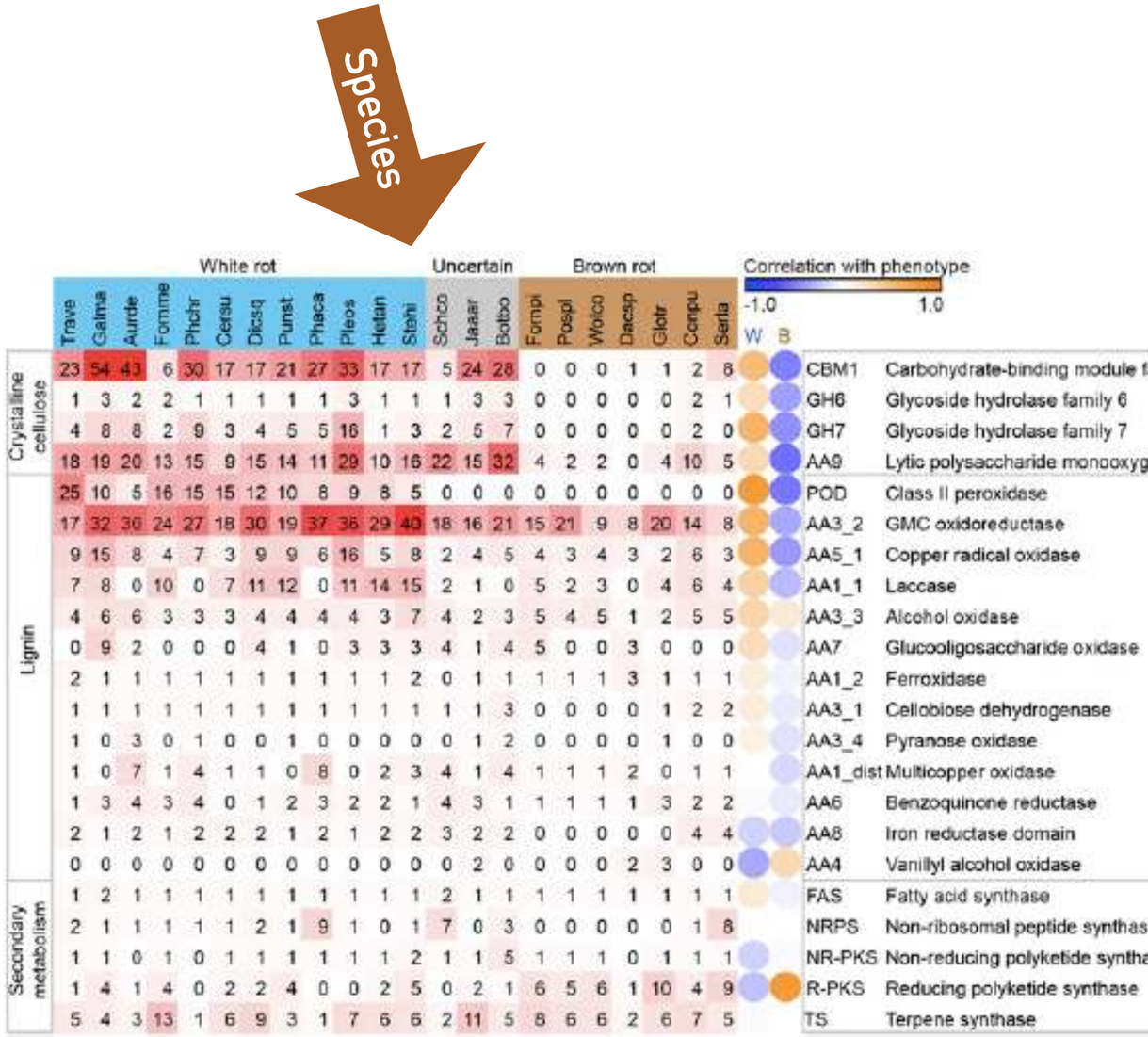
- Gene / Tree species tree reconciliation
- NOTUNG can classify nodes and lineages on the gene tree as gains or losses
- Requires a known species tree and then input gene trees, each is reconciled with the species tree to determine

COMPARING GENE CONTENT

Heat map clustered by similarity - not phylogeny

Heat map columns grouped by ecological status

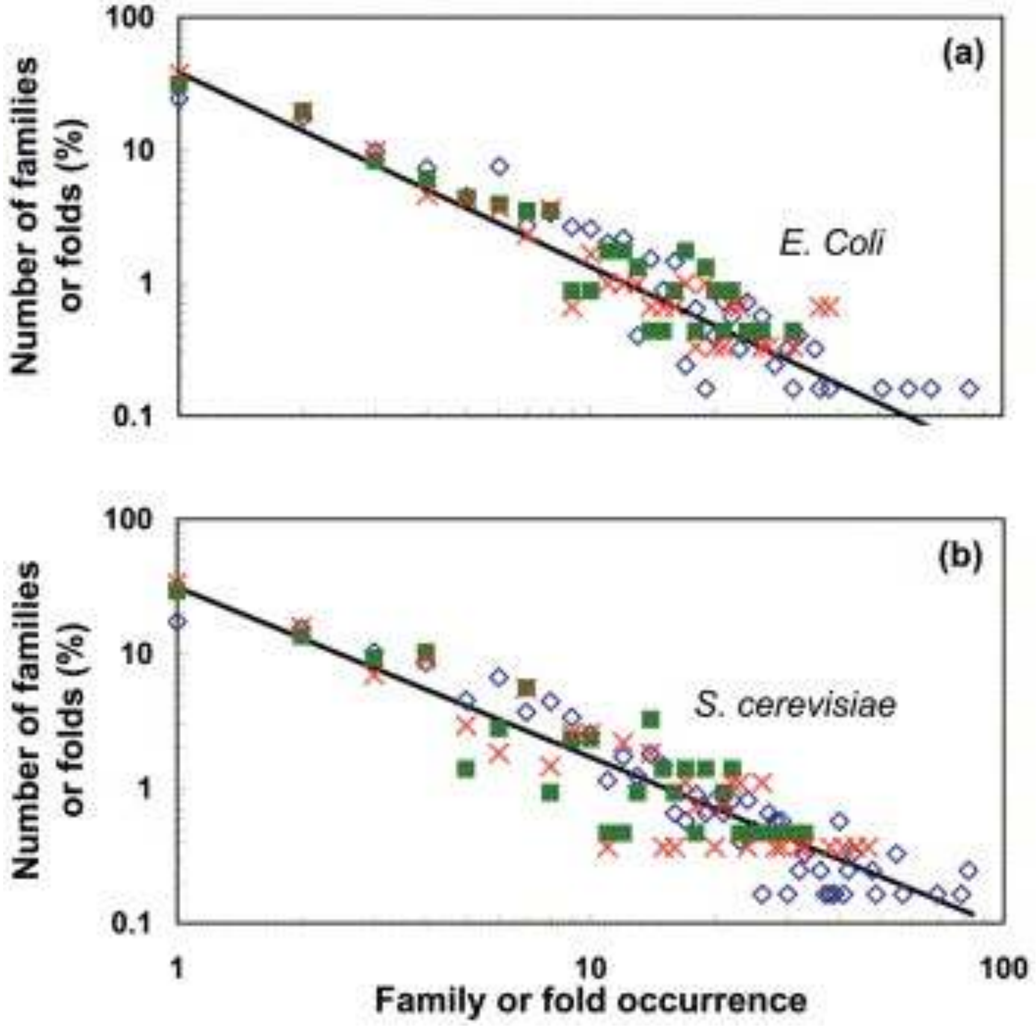




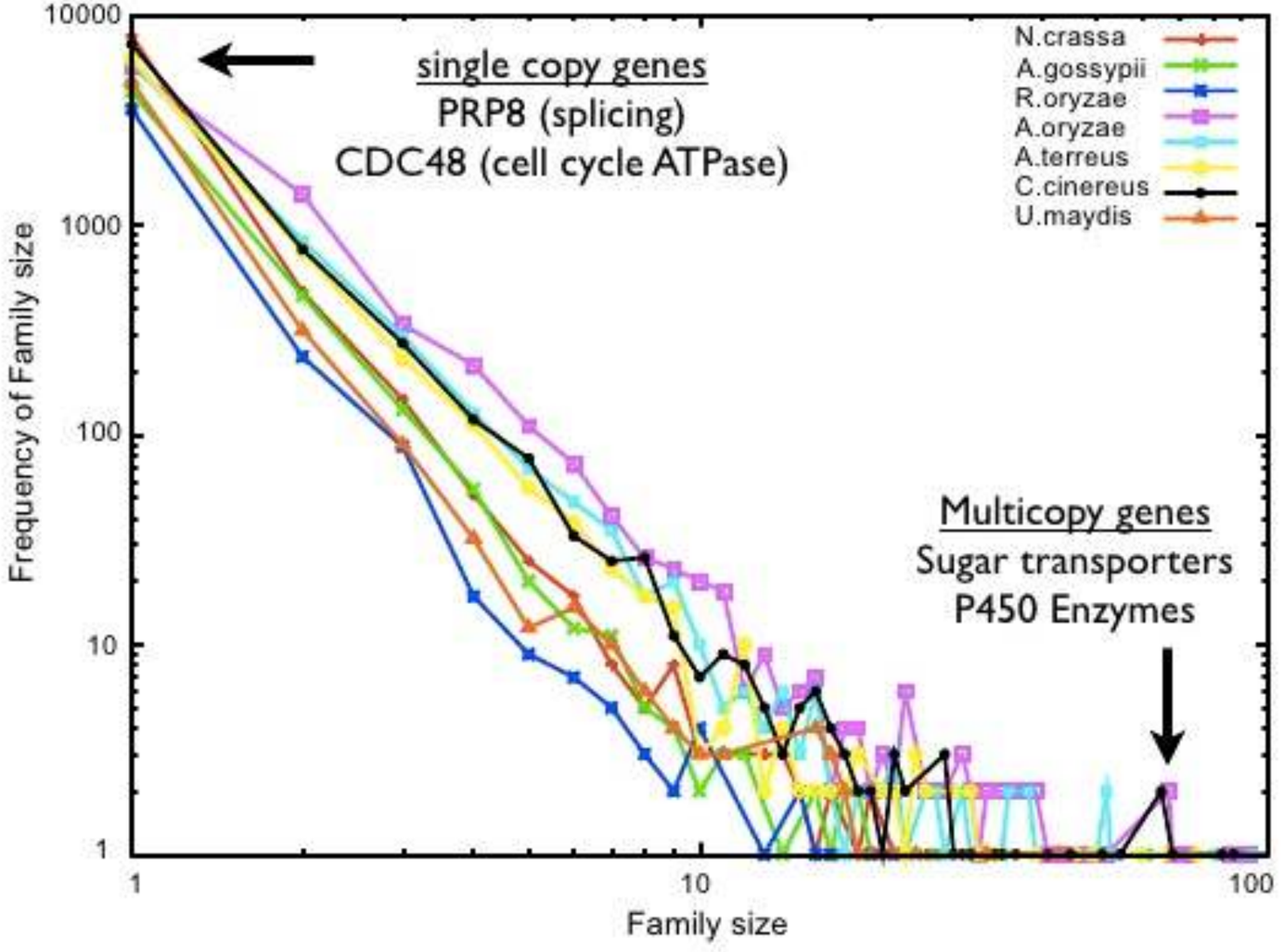
Murat et al Nat Ecol Evol 2018

Riley et al PNAS 2014

GENE FAMILY SIZES FOLLOW A POWER LAB AND CAN BE MODELED WITH A BIRTH-DEATH MODEL

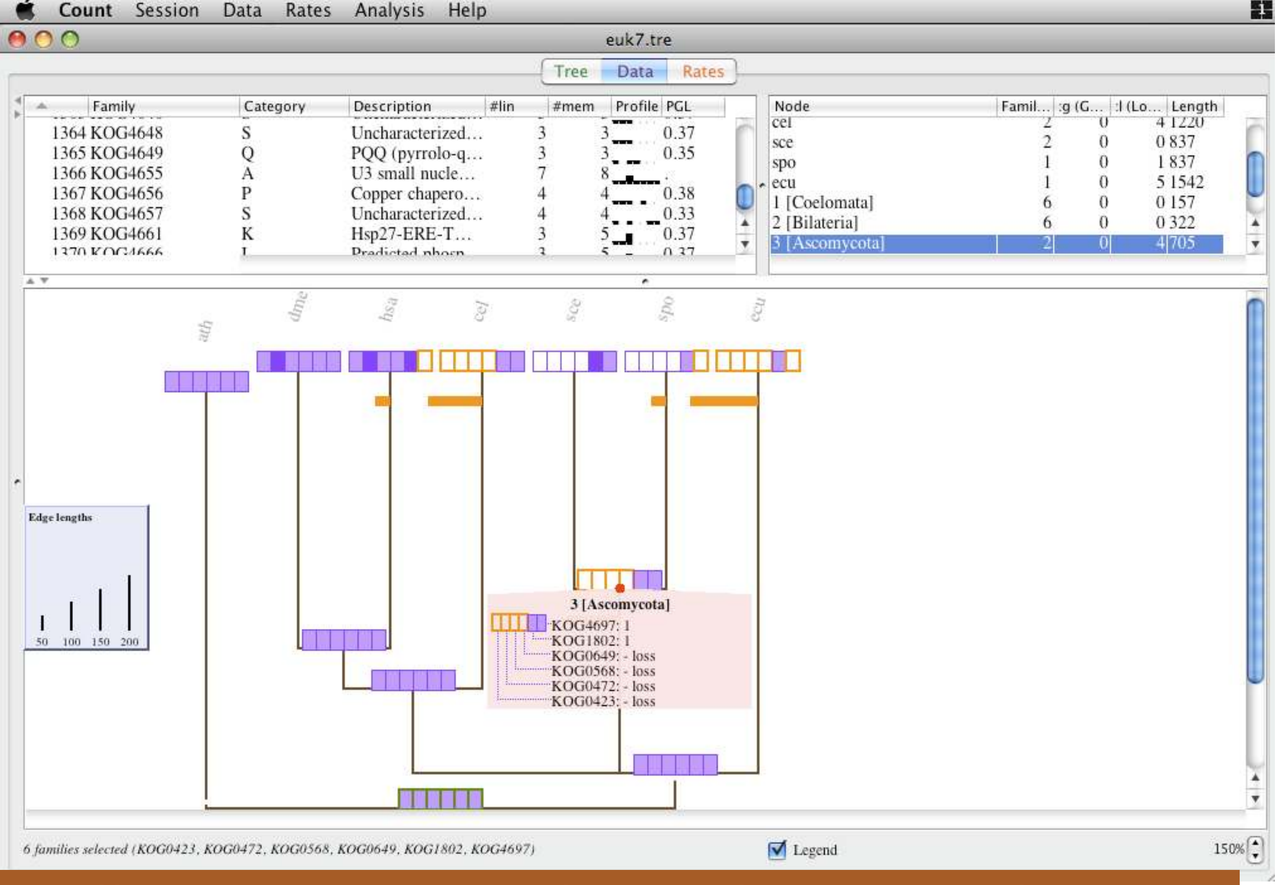


Qian et al JMB 2008



Stajich 2006

GENE FAMILY SIZE CHANGE USING COUNTS



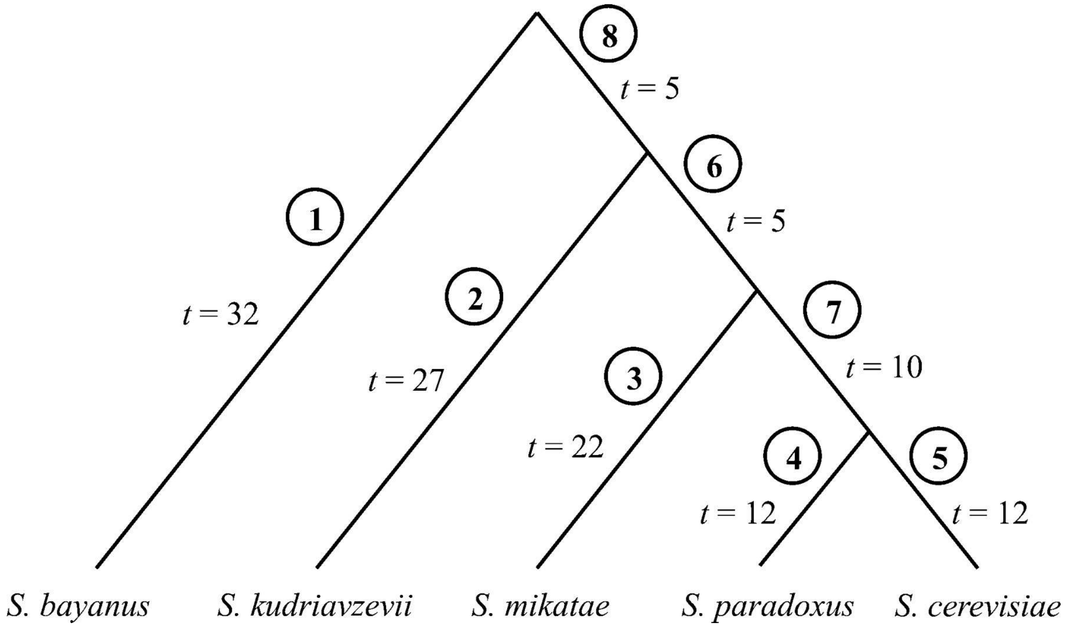
http://www.iro.umontreal.ca/~csuros/gene_content/count.html

- Using counts of copy number of genes in a gene family can examine the patterns of gains and losses to:
  - Find branches with excess of gains/ losses
  - Identify gene families with extreme changes

Species ->	A	B	C
OG1	1	3	1
OG2	10	2	2
OG3	5	5	2

- Count is a tool to evaluate homolog family sizes (phylogenetic profiles), or other numerical census-type characters along a phylogeny.

CAFE – COMPUTATION ANALYSIS OF (GENE) FAMILY EVOLUTION



Branch #	Expansions	No change	Contractions	Average expansion
1 (t = 32)	97	3181	239	-0.050
2 (t = 27)	383	3032	102	0.095
3 (t = 22)	509	2922	86	0.147
4 (t = 12)	96	3383	38	0.019
5 (t = 12)	44	3426	47	0.021
6 (t = 5)	3	3491	23	-0.005
7 (t = 10)	10	3313	194	-0.052
8 (t = 5)	2	3515	0	0.001

Family	Family sizes	Branch	Score
Flavodoxin	(2 (3 (**5** (1 1))))	3	0.11

Hahn et al Genome Res 2005

https://hahnlab.github.io/CAFE/

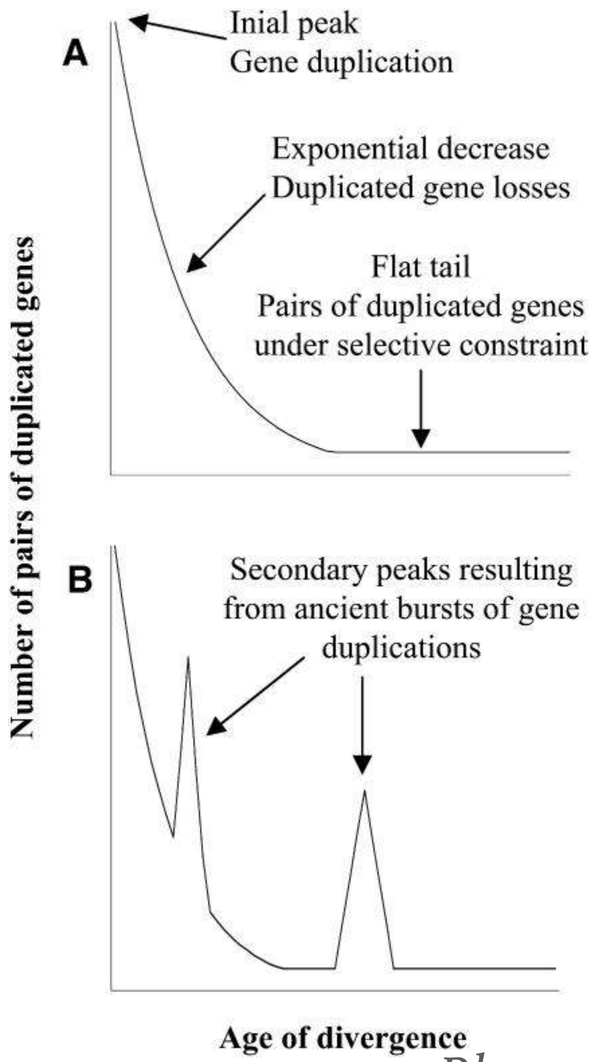
EVOLUTIONARY PATTERNS
GENOME DUPLICATION

```
GBNTN_000002-T1 GACCTGCTTGACTATGAAGAAGCCTCGCATGGTATTCTTGTTGAACCTCCTCCACCCAAT
GBNWO_006121-T1 GACCTGCTTGACTATGAAGAAGCCTCGCATGGTATTCTTGTTGAACCACCTTTACCCAAT
                ********************************************** ***  *******

GBNTN_000002-T1 ACACTGTTACTAGGATCTGAGAATGATGGTGGTACTTCTTGTTATATCGATTCCTTATTG
GBNWO_006121-T1 ACGCTGCTATTAGGATCTGAGAATGATGGTGGTACTTCTTGTTATATTGATTCCTTATTG
                ** *** ** ************************************ ************

GBNTN_000002-T1         EAKWNGQDEAGHWTQEDASELFLFITE--------TFDLPYLPFQIRLFHGANKDSDDDR
GBNWO_006121-T1         EAKWNGQDEAGHWTQEDASELFLFITETFDLPYLPTFDLPYLPFQIRLFHGANKDSDDDR
                        ************************        ************************

GBNTN_000002-T1         VMTDRTLTLSIPETNNAVKEIKLQDILVDYFYNSIITGVRRQVDSSHSHEEVTENVAASP
GBNWO_006121-T1         VMTDRTLTLSIPETNNSVKEIKLQDILVDYFYNSIITGVRRQVDSSHSHEEVTENVAASH
                        ***************:********************************************

GBNTN_000002-T1         TIQSMINSADEKPNSKRALIKKNDSQVAVTAWQVLELLPFYSPTNEQGISINTQVDALFP
GBNWO_006121-T1         TIQSMINNTDEKPSSKRALVKKNDGHVAVTAWQVLELLPFYSPTNEQGISINAQVDASFP
                        *******.:****.*****:****.:***************************:**** **
```

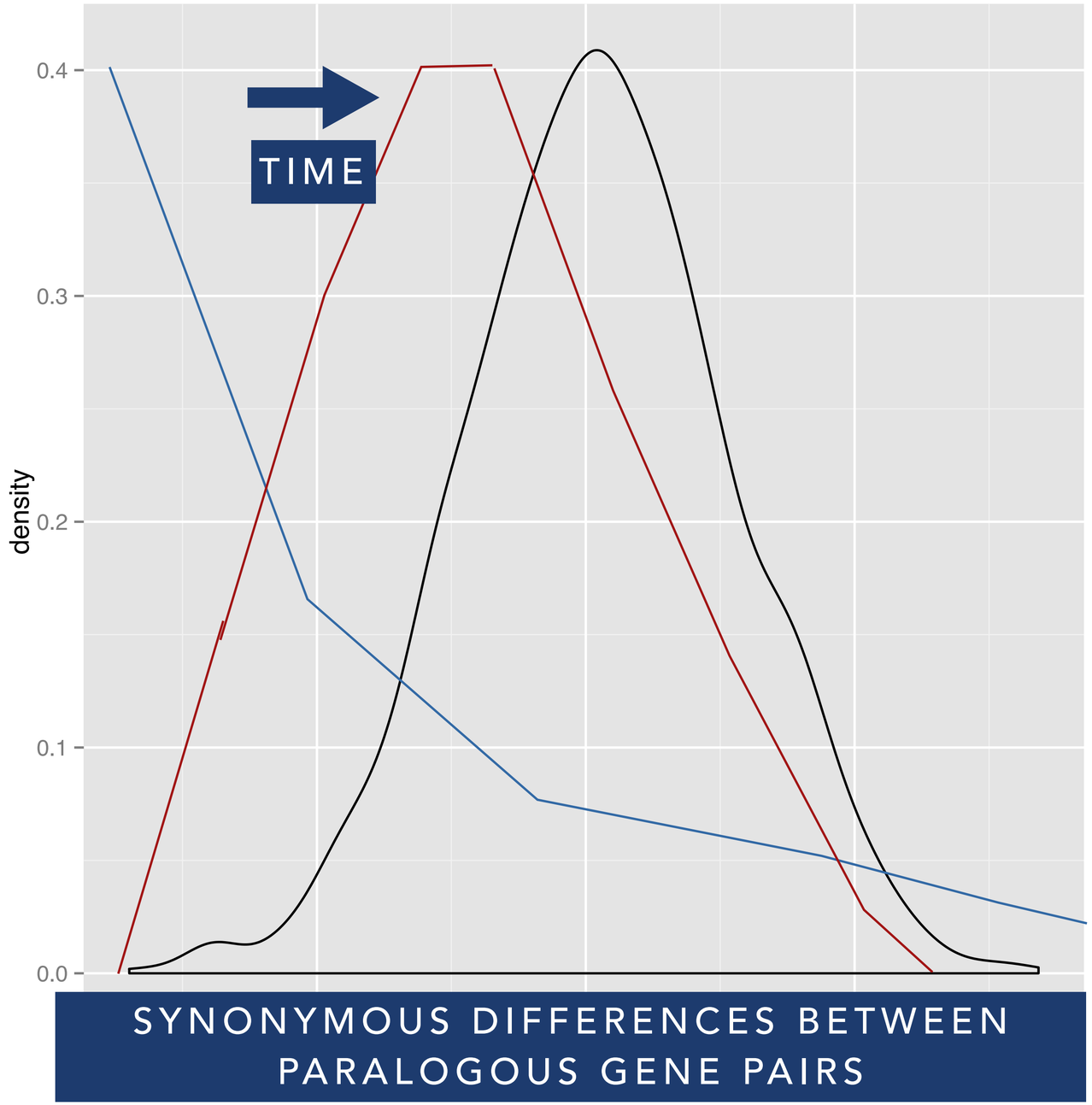
MOLECULAR EVOLUTION

- Phylogenetics is about inferring evolutionary history of group of sequences, characters, taxa, etc
- Molecular evolution is specifically focusing on changes in molecular sequence (protein, DNA/RNA). One aspect is examining rates of change of sequences.
- changes in sequence calculated under different substitution models and calculate the distance.
- Ks = synonymous site change
  dS = synonymous site rate of change
- Ka = non-synonymous site change
  dN = non-synonymous site rate of change

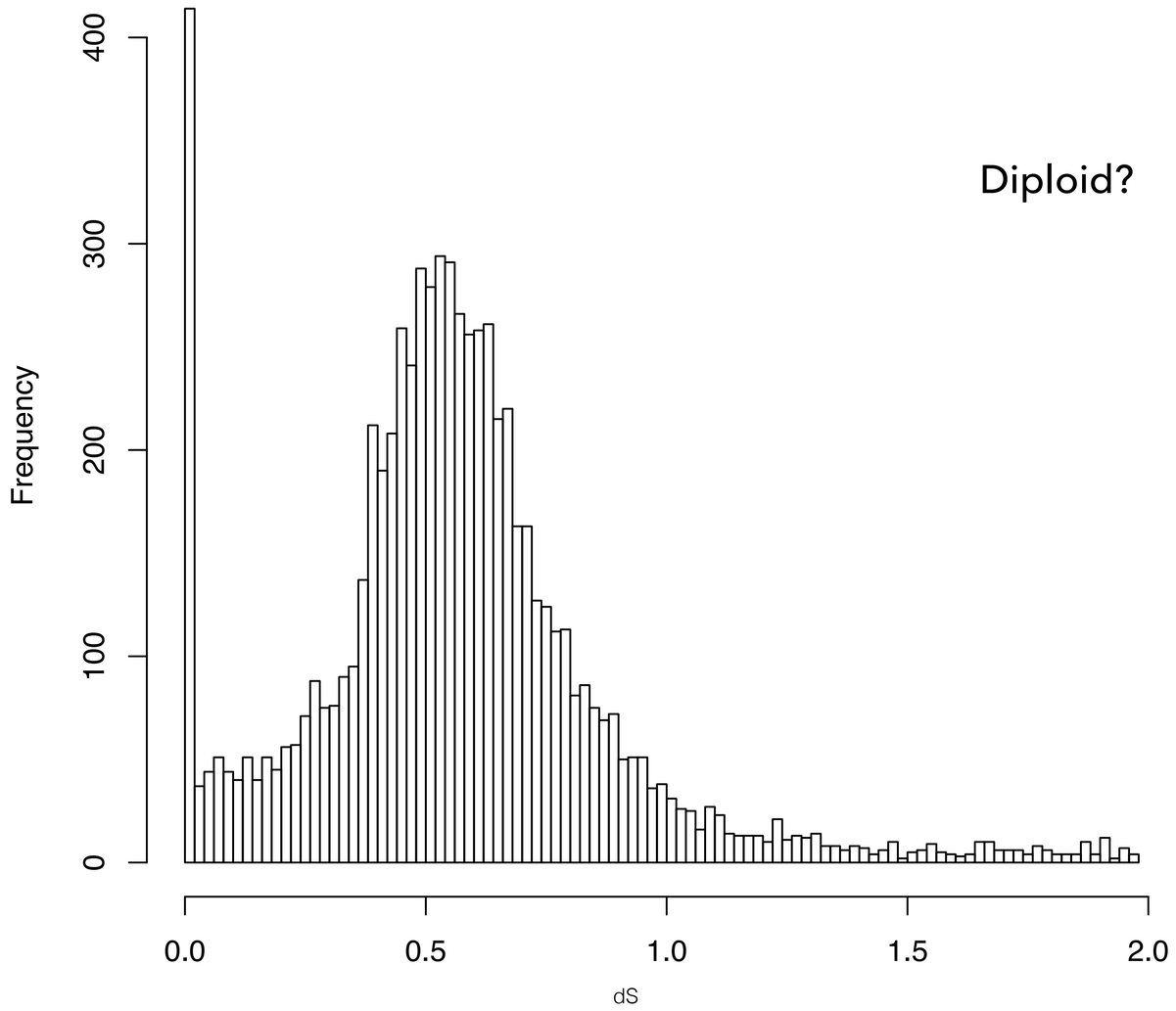
KS PLOTS AND DUPLICATION



Blanc and Wolfe Plant Cell 2004

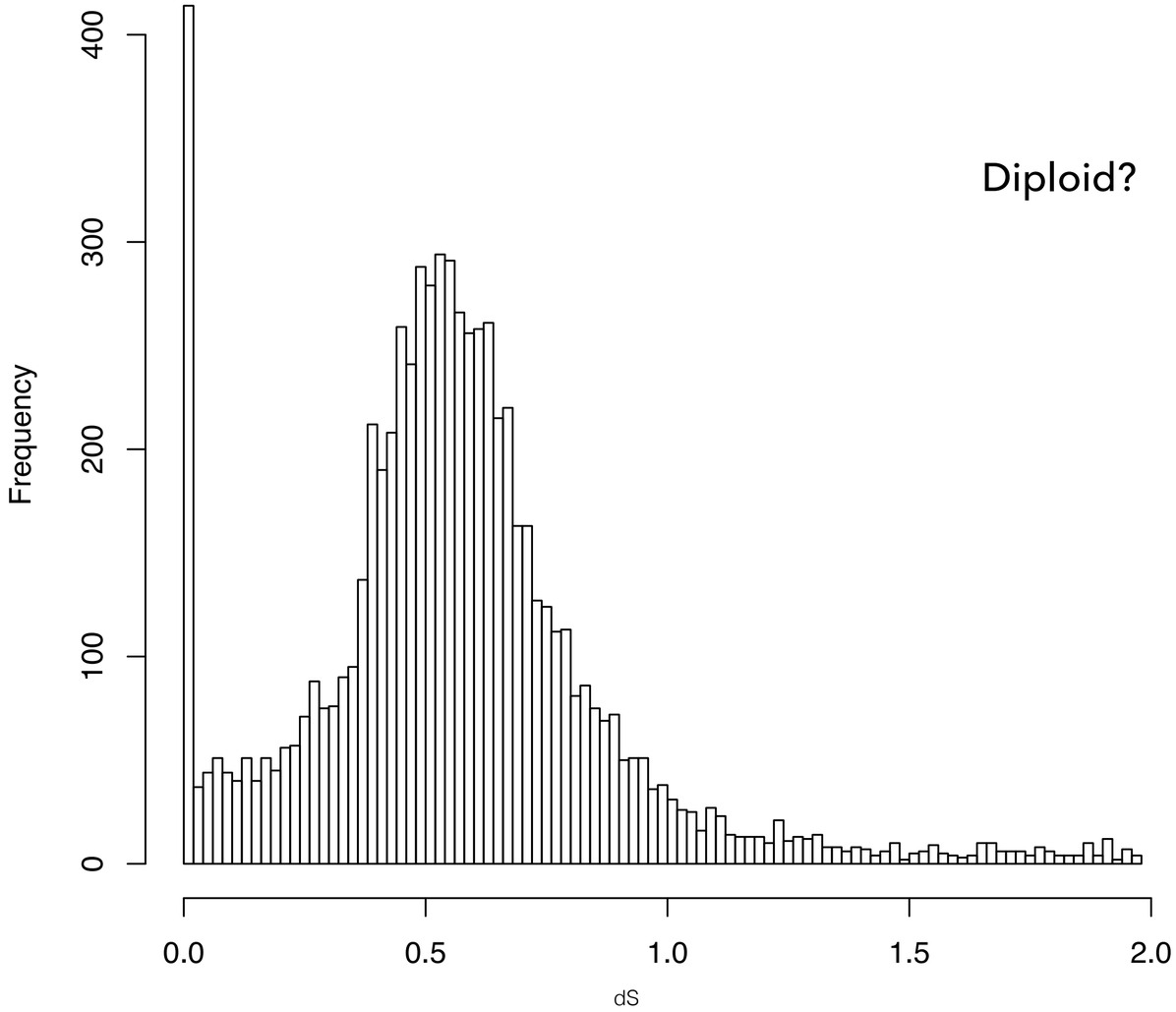


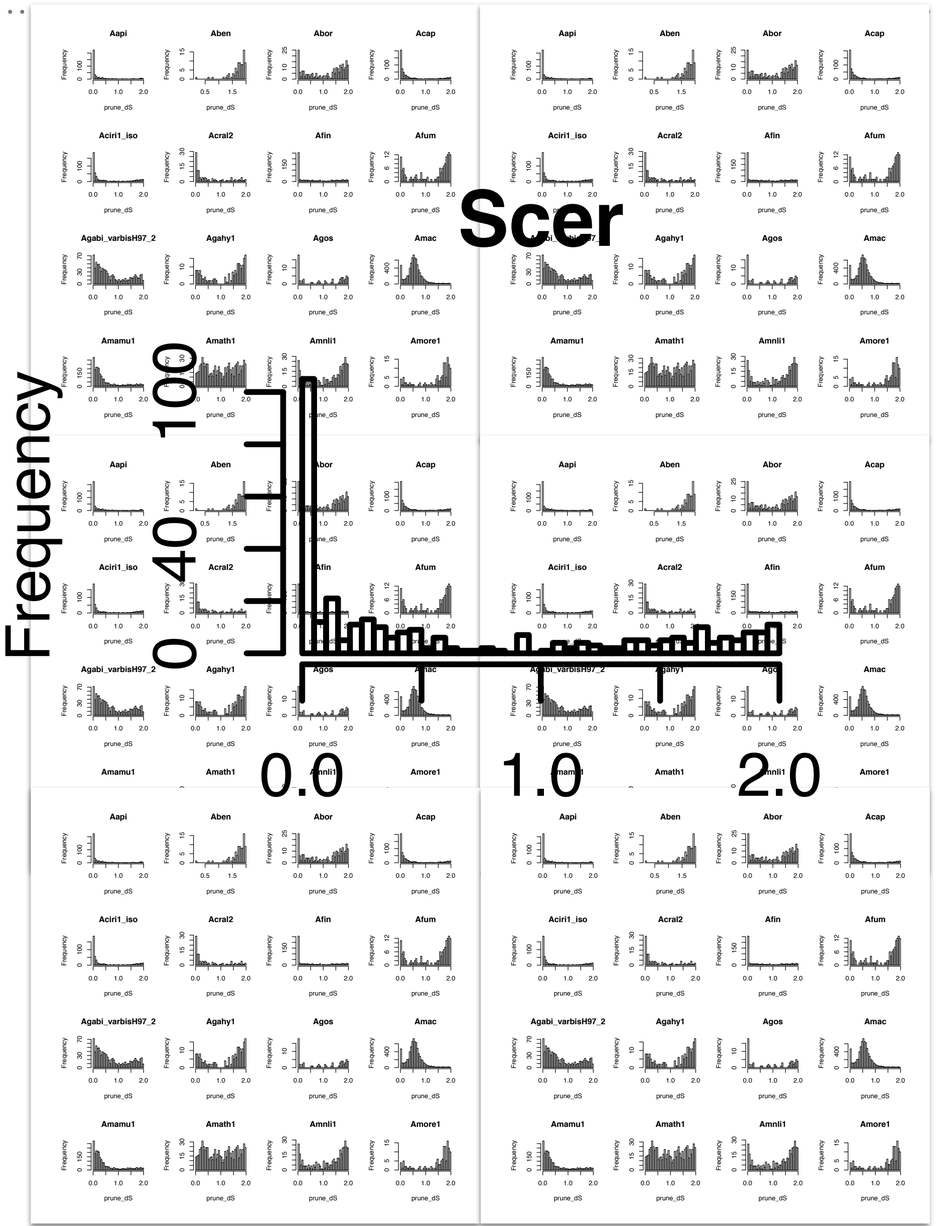
Allomyces macrogynus (Blastocladiomycota Chytrid)



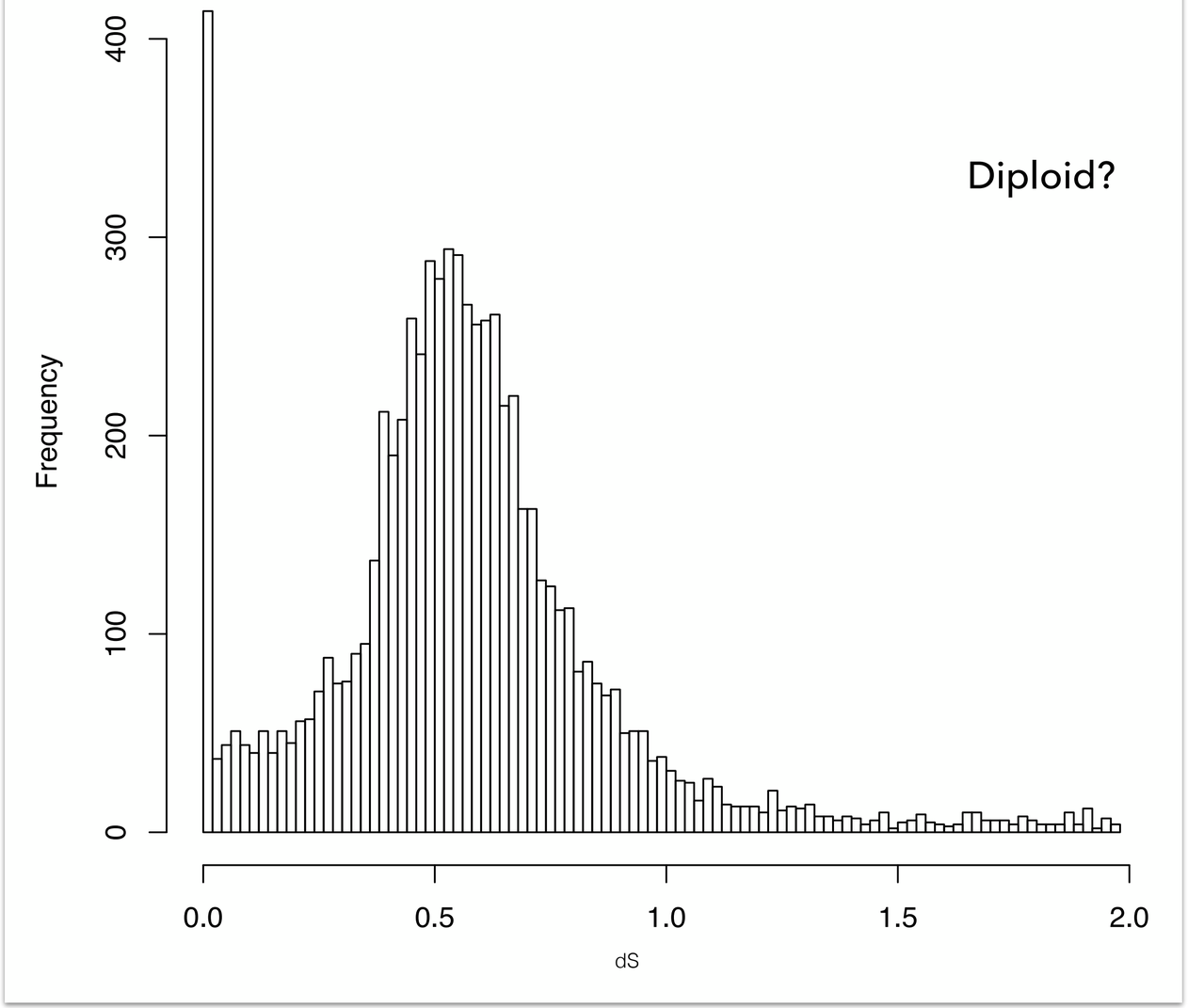
Allomyces macrogynus (Blastocladiomycota Chytrid)

Diploid?

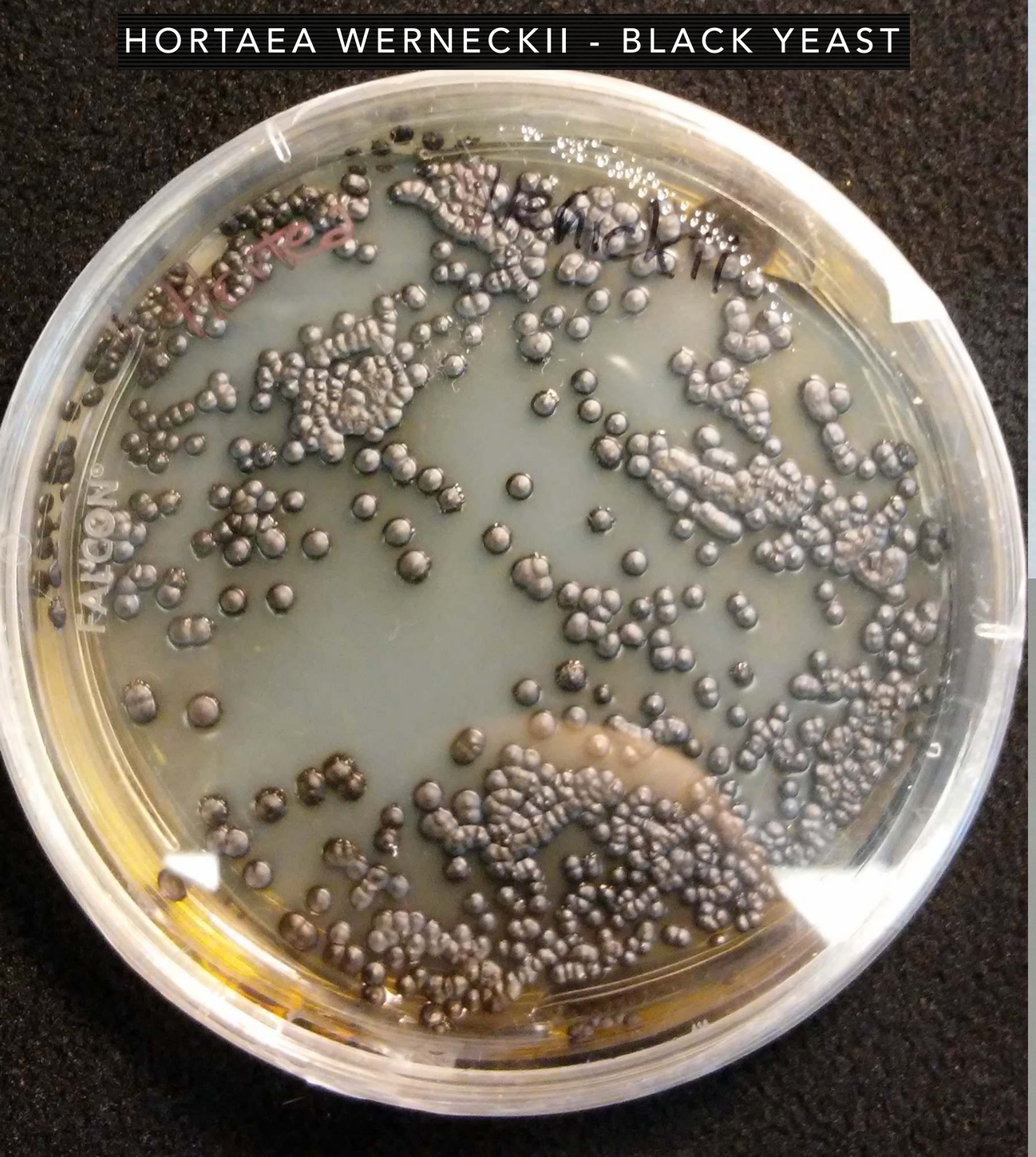


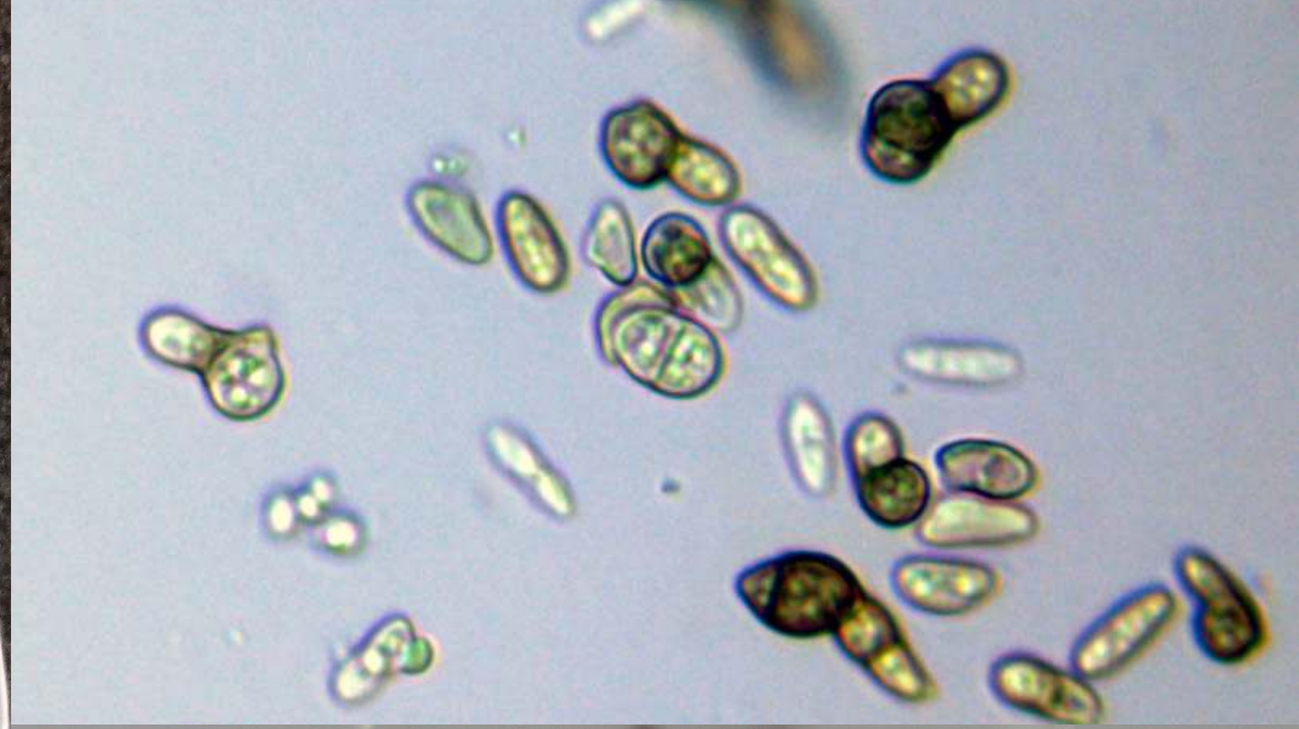


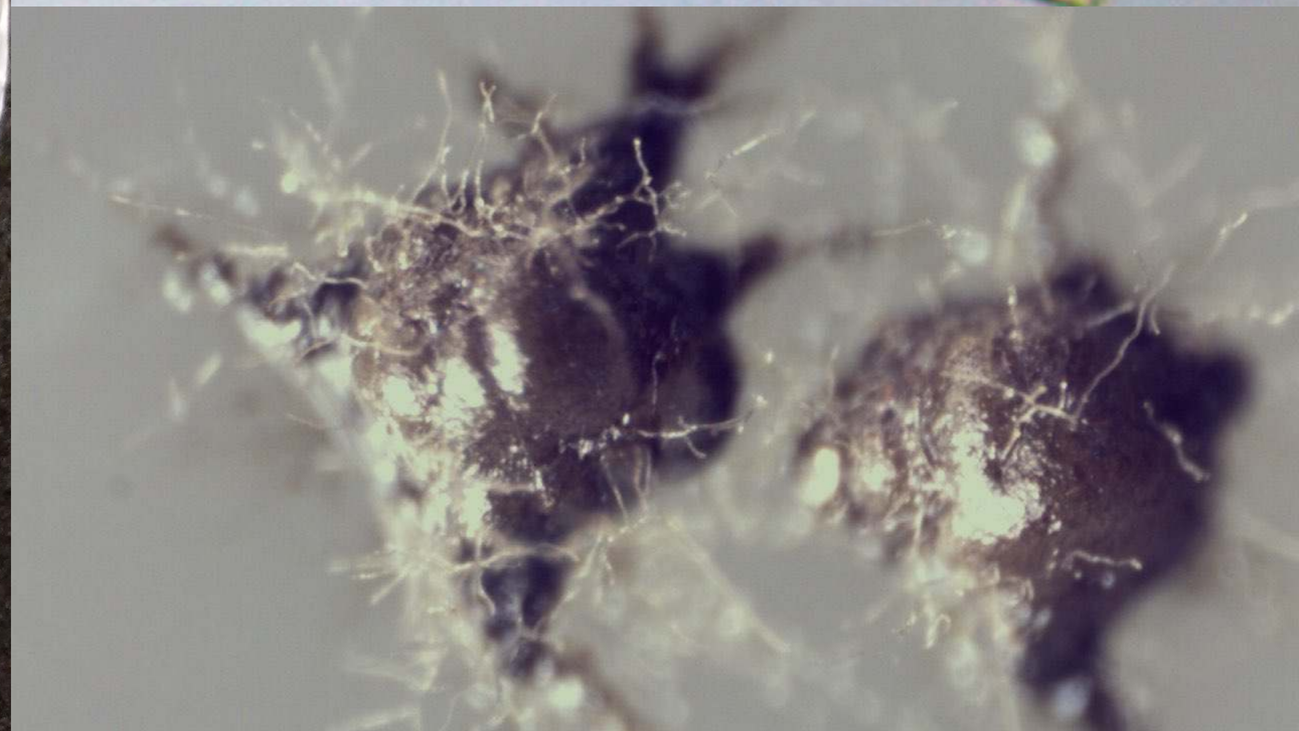
Allomyces macrogynus (Blastocladiomycota Chytrid)



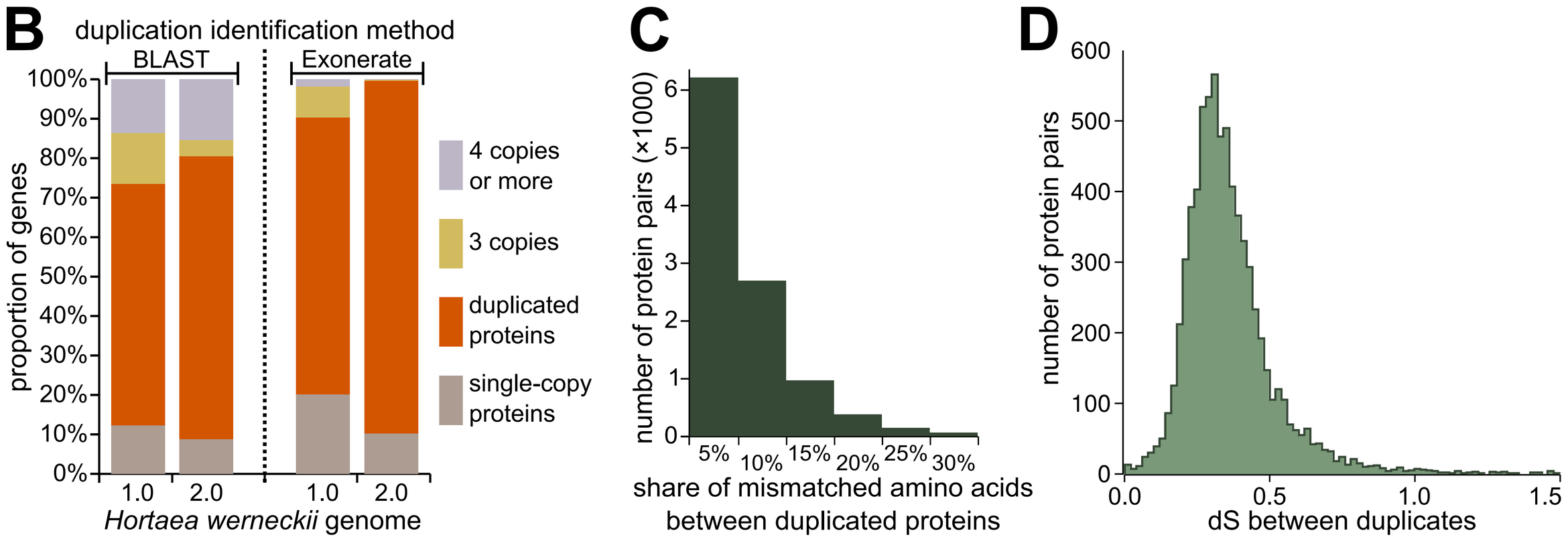
HORTAEA WERNECKII - BLACK YEAST







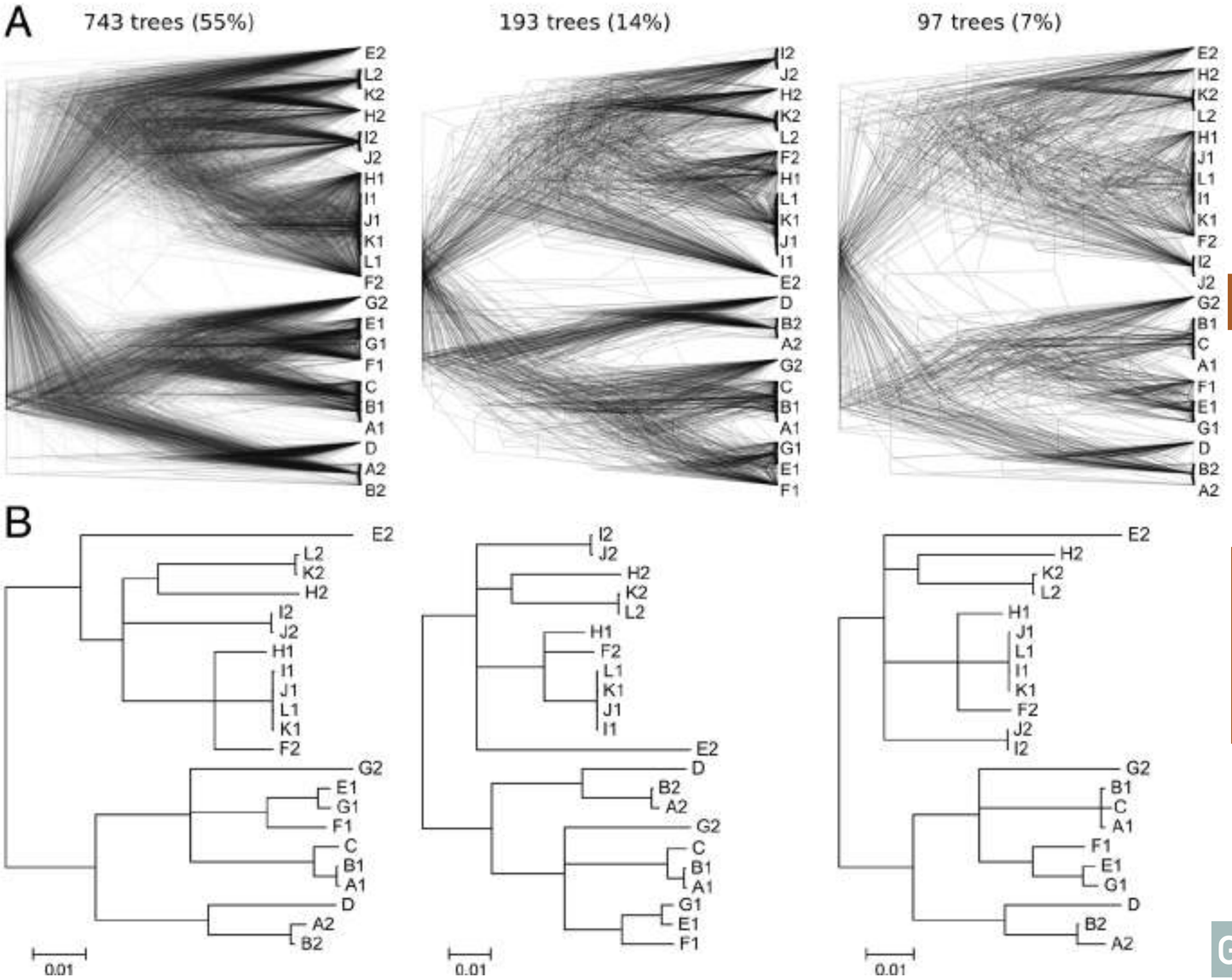
Recent fungal whole genome duplications- *Hortaea werneckii*



Cene Gostinčar, Metka Lenassi
Sunita Sinha, Stephane Flibotte, Corey Nislow

Sinha, Stajich et al. G3 2017

HORTAEA WERNECKII HISTORY IS MORE COMPLICATED THAN A WGD



Examinig 11 more strains

Multiple evolutionary histories among genomic loci

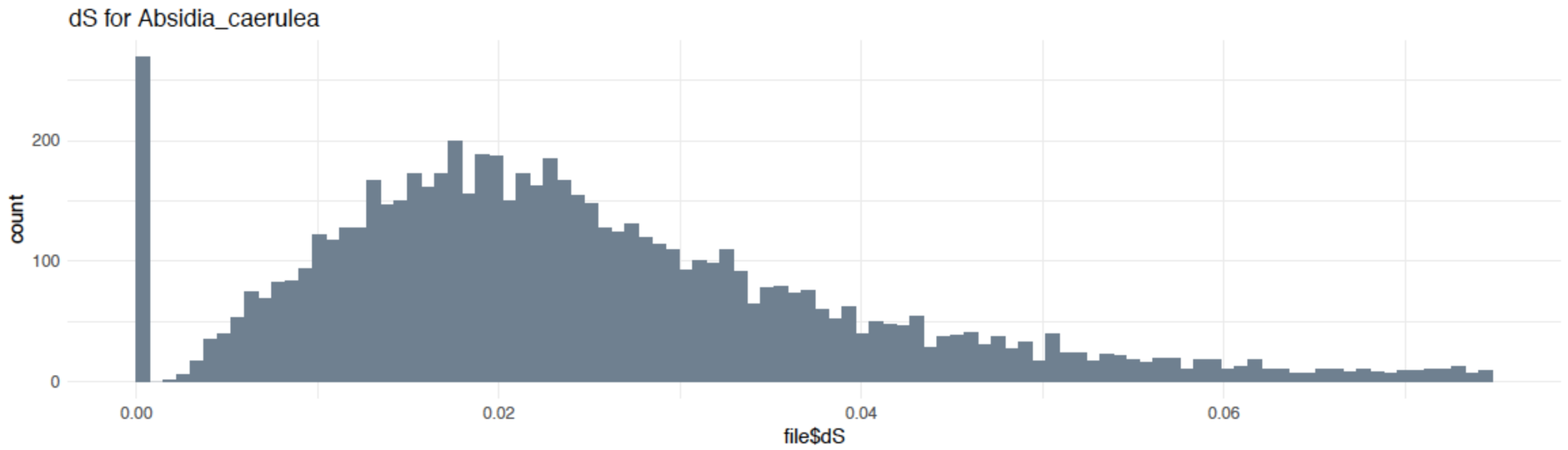
Evidence that some strains are haploid and some are diploid. First genome sequenced was diploid and where haplotypes had sufficient divergence they assembled into separate scaffolds

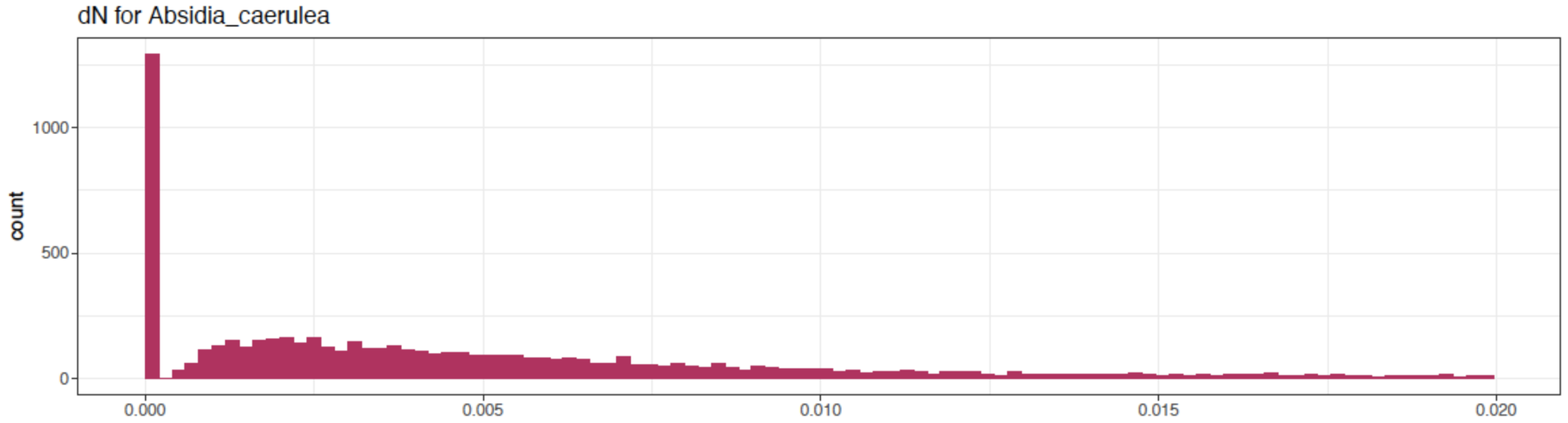
Gostinčar et al BMC Genomics 2018

HOW TO COMPUTE KS PLOTS AND TEST FOR DUPLICATIONS?

- All vs All clustering of proteins (or gene DNA sequence)
- Identify 1:1 paralogs, 3 or more members of a gene family are going to be hard to resolve so ignore these as the hypothesis for a WGD is it creates a lot of duplicate pairs
- Align proteins and project in to coding sequence alignment to get codon alignment
- calculate Ks (and Ka if you like) for all pairs.
- PAML, YN00 are suitable. I also wrote a fast tool based on YN00 that generates simple tabular output - https://github.com/hyphaltip/subopt-kaks/
  - yn00_cds_prealigned or yn00_cds_optimal (let it align coding sequences for you)

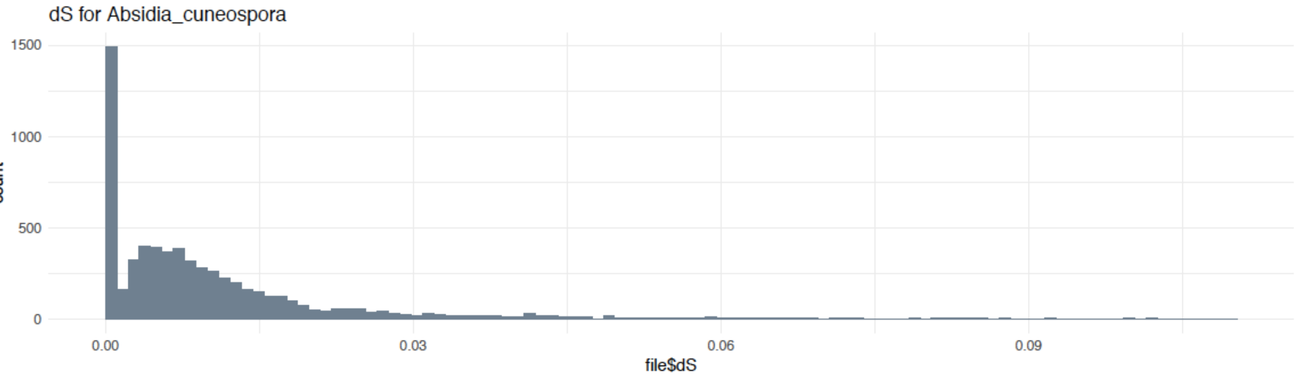
USEFUL INVESTIGATING BETWEEN SPECIES DIVERGENCES

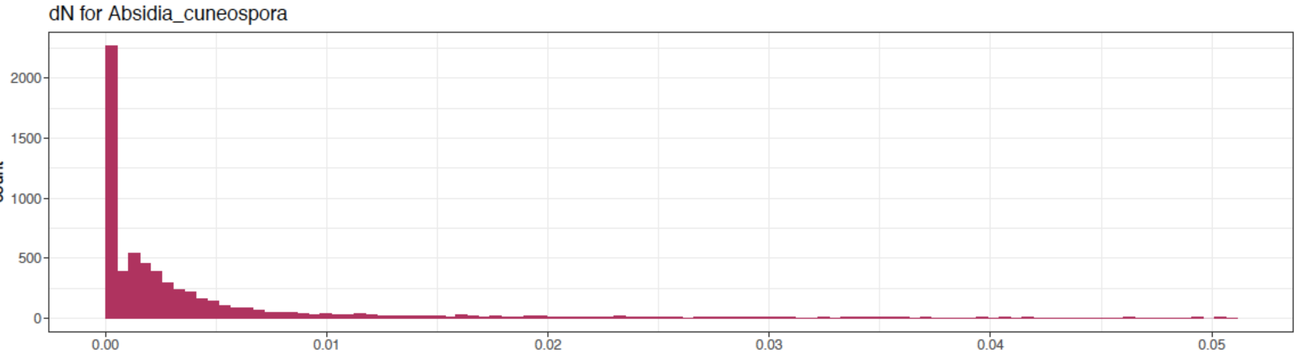




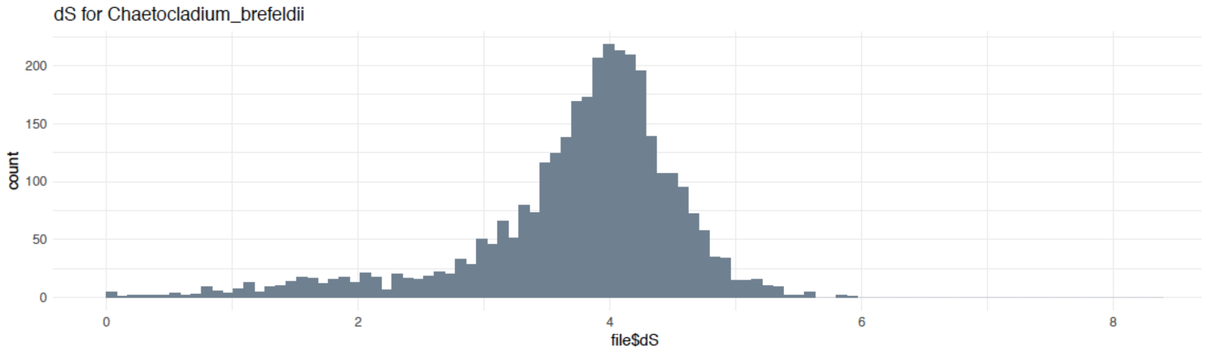
Comparing 2 strains

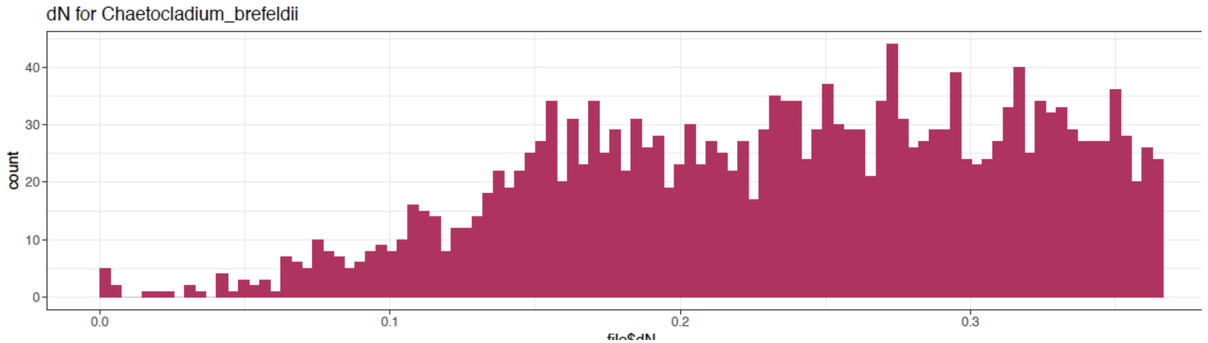
USEFUL INVESTIGATING BETWEEN SPECIES DIVERGENCES





Comparing 2 strains





Comparing 2 species

DATA SOURCES

- Primary sequence data
  - Genome assembly, annotation, raw sequencing reads - GenBank (NCBI,EMBL,DDBJ)
- Curated, organized, managed
  - Model organism databases (flybase.org, yeastgenome.org, wormbase.org)
  - Collected databases of genomes EnsEMBL, EuPathDB/FungiDB (ensembl.org, fungidb.org, eupathdb.org)
- Data providers and integrated tools
  - US Joint Genome Institute Mycocosm, Phytozome, IMG (Bacteria)
  - Seq center websites/databases (Genoscope, WUSTL)

DATA FORMATS

Fasta
>sequenceID
AGAGCATAT

GFF - Generic feature format
9 columns, tab delimited
chrom, source, type, start,end,score,strand,frame,Group

Genbank: Sequence + annotation

```
LOCUS       MSJE01000001.1        430834 bp    DNA     linear   PLN 03-JAN-2019
DEFINITION  Aspergillus terreus strain IMV 01167.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Aspergillus terreus
  ORGANISM  Aspergillus terreus
            Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
            Eurotiomycetes; Eurotiomycetidae; Eurotiales; Aspergillaceae;
            Aspergillus.
REFERENCE   1  (bases 1 to 430834)
  AUTHORS   Stajich,J.E.
  TITLE     Annotation of genomes of fungal isolates from surfaces in
            International Space Station
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 430834)
  AUTHORS   Stajich,J.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (03-JAN-2019) Plant Pathology and Microbiology,
            University of California-Riverside, 900 University Ave, Riverside,
            CA 92521, USA
COMMENT     'Annotated using funannotate v1.5.1'.
FEATURES             Location/Qualifiers
     source          1..430834
                     /organism="Aspergillus terreus"
                     /mol_type="genomic DNA"
                     /strain="IMV 01167"
                     /db_xref="taxon:33178"
     gene            complement(<14..1332)
                     /locus_tag="BS087_000011"
     mRNA            complement(join(<14..151,209..1006,1071..1332))
                     /product="hypothetical protein"
     CDS             complement(join(<14..151,209..1006,1071..1332))
```

GenBank format

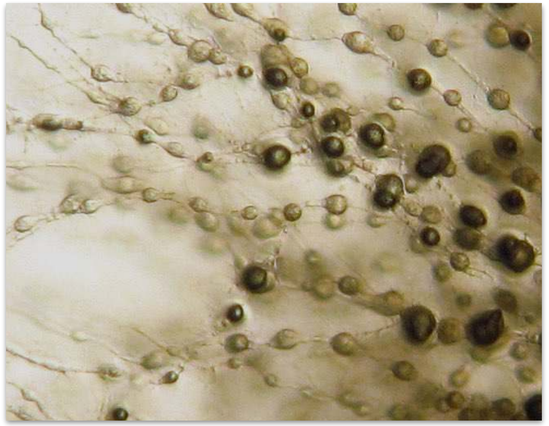
GFF FORMAT

```
MSJE01000001.1    GenBank  gene 14   1332 .    -    .    ID=BS087_000011;
MSJE01000001.1    GenBank  mRNA 14   1332 .    -    .    ID=BS087_000011-T1;Parent=BS087_000011;product=hypothetical protein;
MSJE01000001.1    GenBank  exon 1071 1332 .    -    .    ID=BS087_000011-T1.exon1;Parent=BS087_000011-T1;
MSJE01000001.1    GenBank  exon 209  1006 .    -    .    ID=BS087_000011-T1.exon2;Parent=BS087_000011-T1;
MSJE01000001.1    GenBank  exon 14   151  .    -    .    ID=BS087_000011-T1.exon3;Parent=BS087_000011-T1;
MSJE01000001.1    GenBank  CDS  1071 1332 .    -    0    ID=BS087_000011-T1.cds;Parent=BS087_000011-T1;
MSJE01000001.1    GenBank  CDS  209  1006 .    -    2    ID=BS087_000011-T1.cds;Parent=BS087_000011-T1;
MSJE01000001.1    GenBank  CDS  14   151  .    -    2    ID=BS087_000011-T1.cds;Parent=BS087_000011-T1;
```

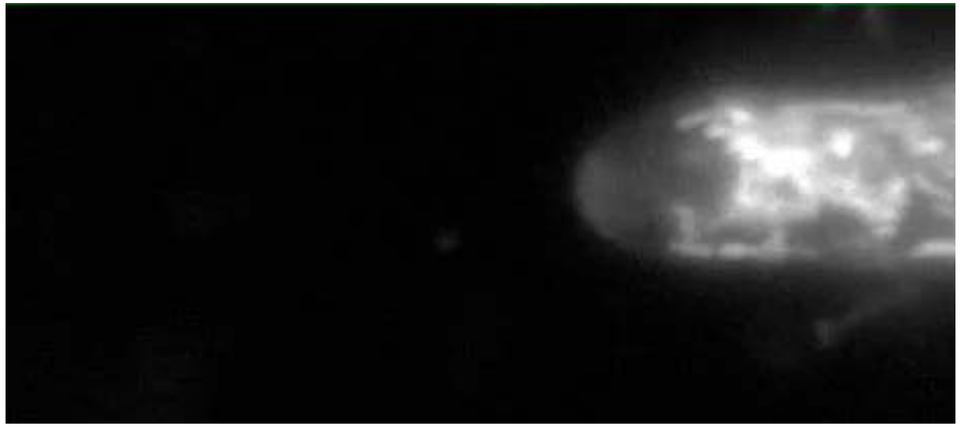
GENOME ANNOTATION

- Pipelines to run automatic prediction. Eukaryotic gene annotation harder than Bacteria or Archaea prediction
  - expressed RNA (RNASeq, ESTs)
  - Comparative data - alignment of proteins and transcripts from other closely related species
  - Ab initio gene predictors trained on the organism (Augustus, GeneMark.HMM, etc)
  - Data are combined into consensus gene models

  - Existing pipelines that I have familiarity with MAKER and Funannotate



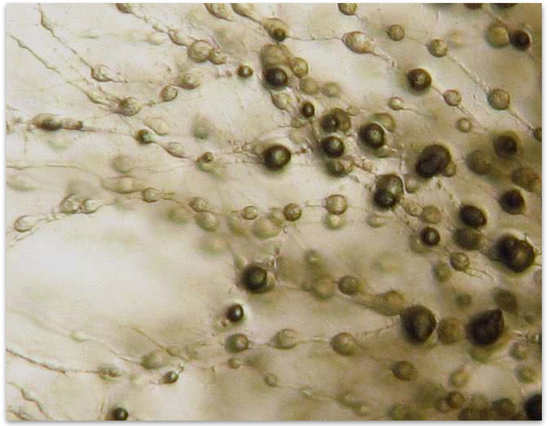
Catenaria (Blastocladiomycota)



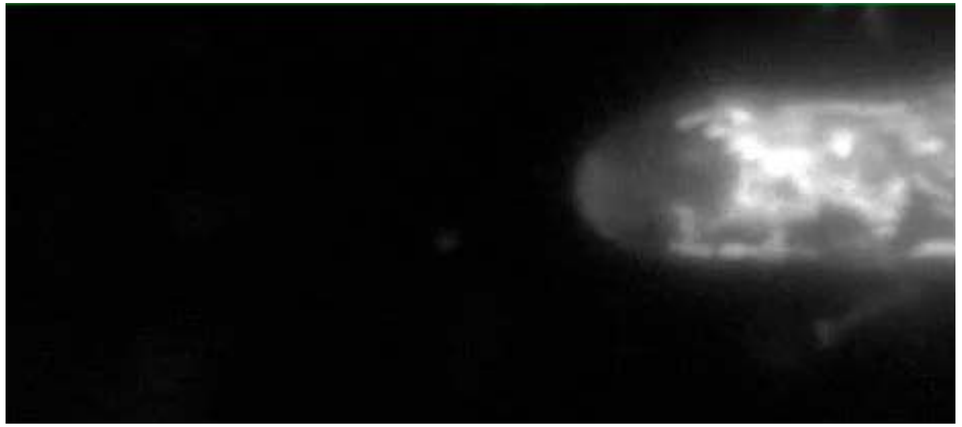
Rhizopus (Mucoromycota)

ANN NGUYEN, GREG JEDD
OUSMANE CISSÉ, MINOU NOWROUSIAN, DAVID HEWITT

EVOLUTION OF MULTICELLULARITY IN FUNGI



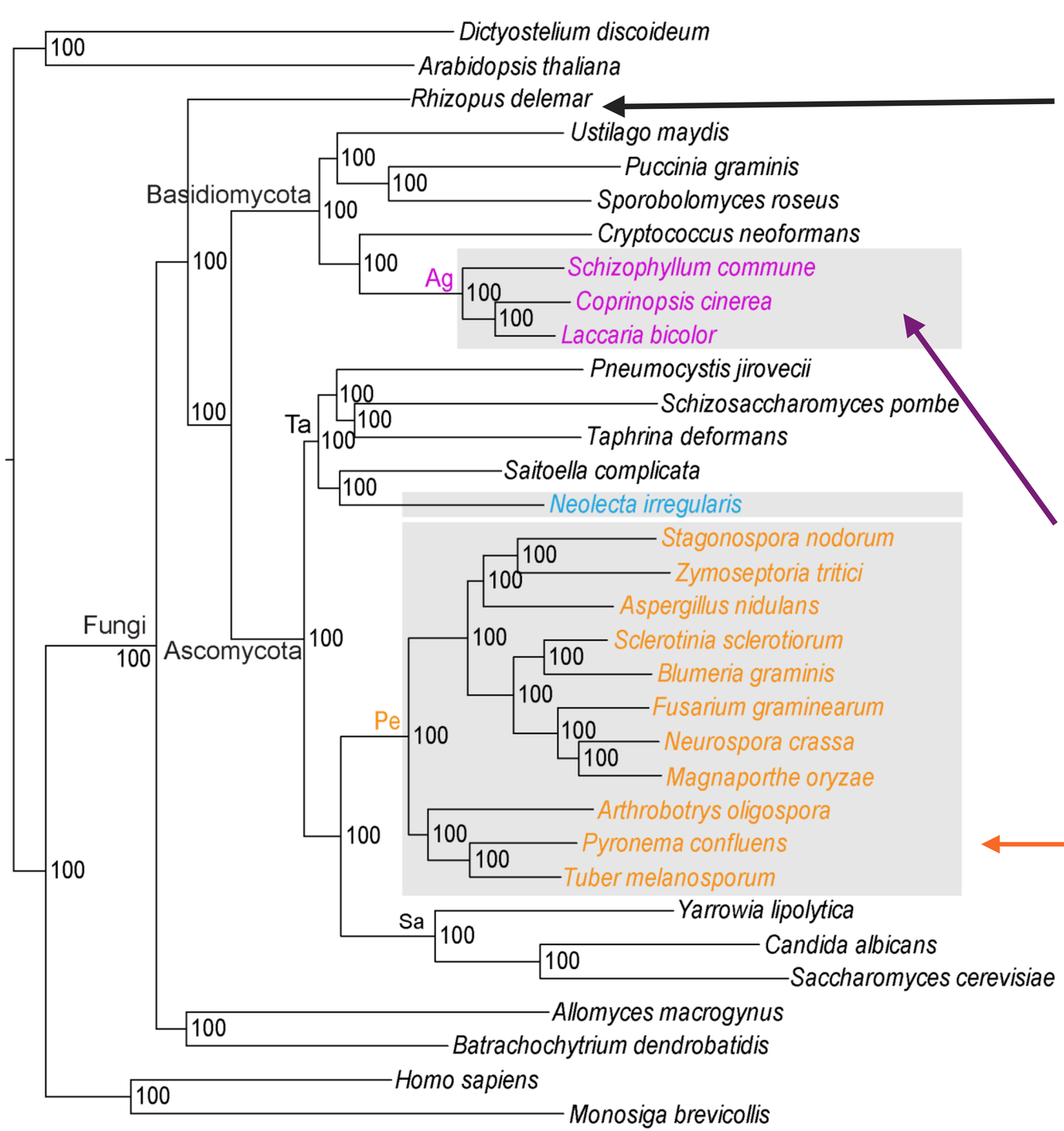
Catenaria (Blastocladiomycota)

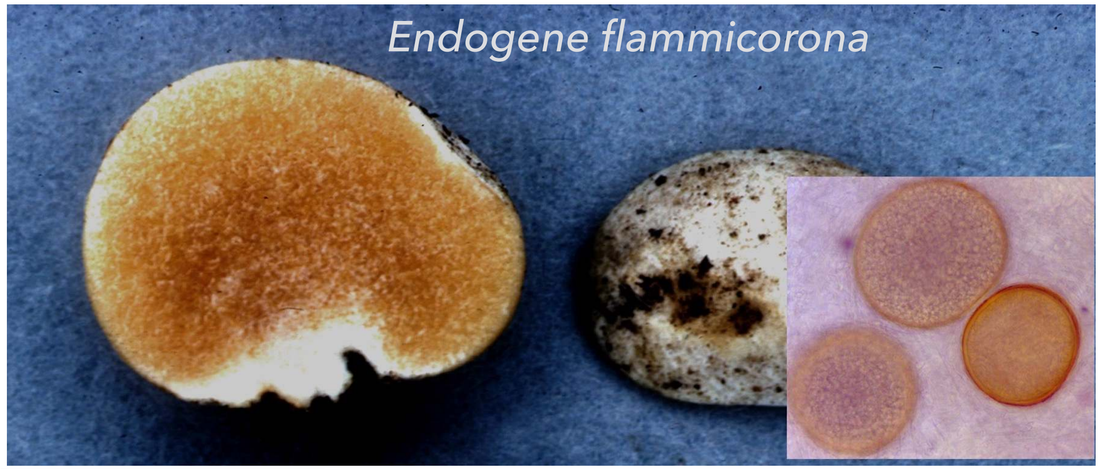


Rhizopus (Mucoromycota)

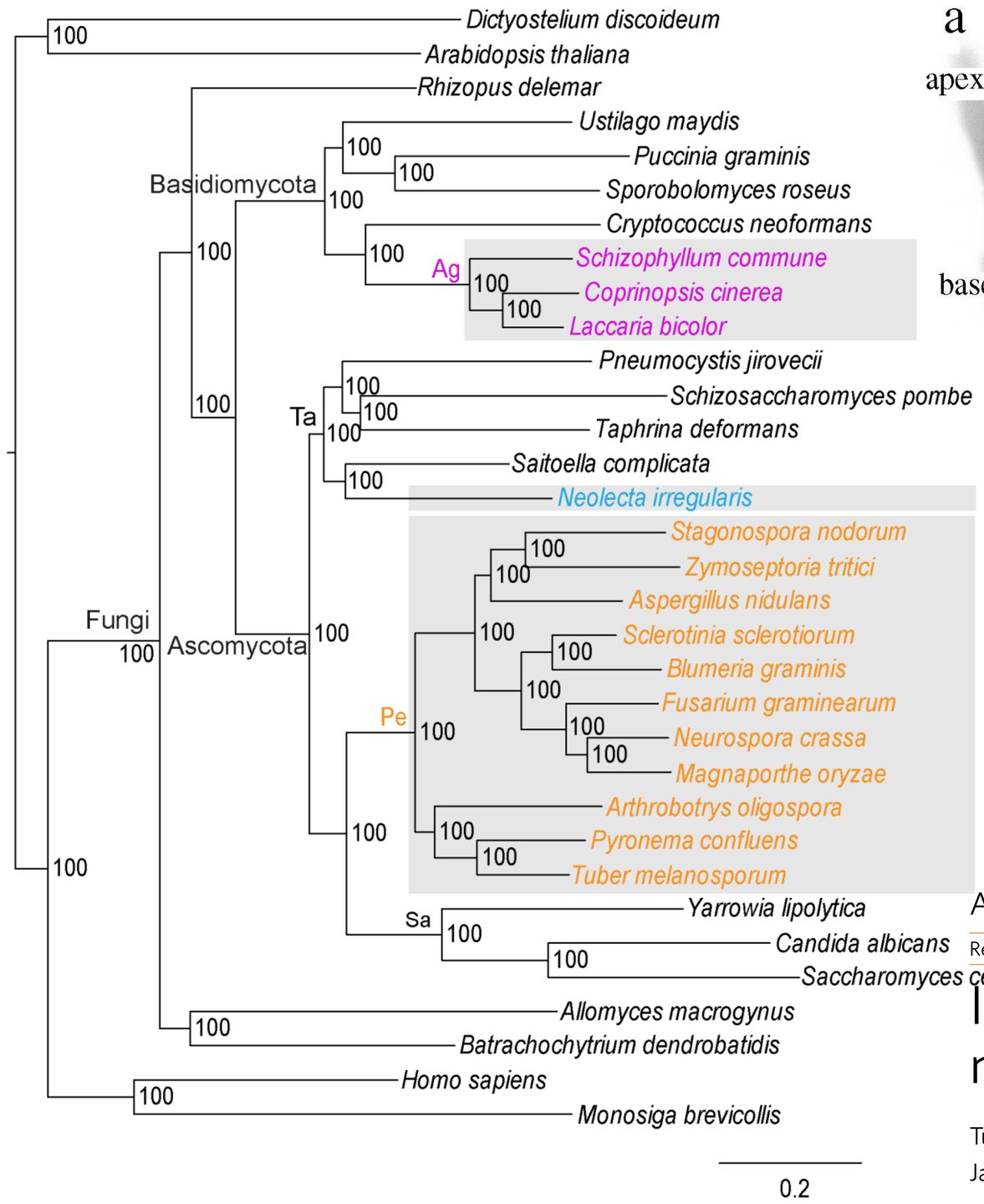
ANN NGUYEN, GREG JEDD
OUSMANE CISSÉ, MINOU NOWROUSIAN, DAVID HEWITT

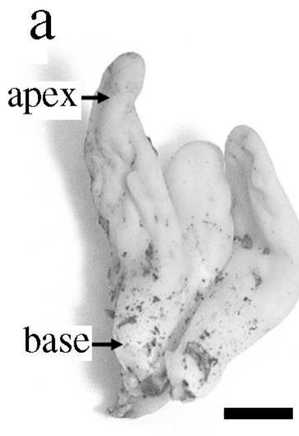
EVOLUTION OF MULTICELLULARITY IN FUNGI


Dictyostelium discoideum
Arabidopsis thaliana
Rhizopus delemar
Ustilago maydis
Puccinia graminis
Sporobolomyces roseus
Cryptococcus neoformans
Basidiomycota
Ag
Schizophyllum commune
Coprinopsis cinerea
Laccaria bicolor
Pneumocystis jirovecii
Schizosaccharomyces pombe
Ta
Taphrina deformans
Saitoella complicata
Neolecta irregularis
Stagonospora nodorum
Zymoseptoria tritici
Aspergillus nidulans
Fungi
Ascomycota
Sclerotinia sclerotiorum
Blumeria graminis
Fusarium graminearum
Pe
Neurospora crassa
Magnaporthe oryzae
Arthrobotrys oligospora
Pyronema confluens
Tuber melanosporum
Yarrowia lipolytica
Sa
Candida albicans
Saccharomyces cerevisiae
Allomyces macrogynus
Batrachochytrium dendrobatidis
Homo sapiens
Monosiga brevicollis


Endogene flammicorona


Tuber melanosporum





NEOLECTA HAS A MULTICELLULAR FRUITING BODY AND FORMS HYPHAE

ARTICLE

Received 22 Jul 2016 | Accepted 29 Dec 2016 | Published 8 Feb 2017

DOI: 10.1038/ncomms14444 OPEN

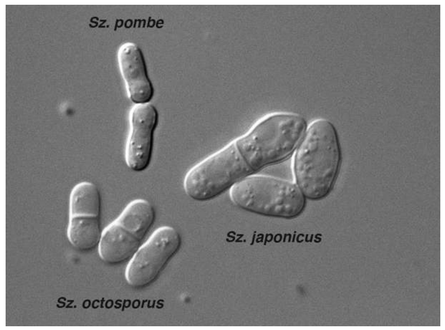
Innovation and constraint leading to complex multicellularity in the Ascomycota

Tu Anh Nguyen[1,*], Ousmane H. Cissé[2,*,†], Jie Yun Wong[1], Peng Zheng[1], David Hewitt[3], Minou Nowrousian[4], Jason E. Stajich[2] & Gregory Jedd[1]

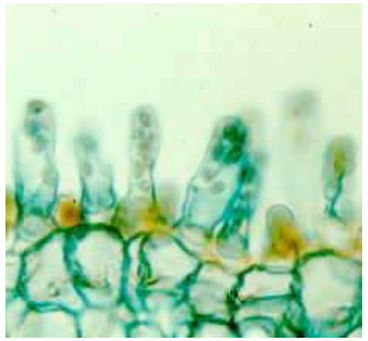
NEOLECTA AS MODEL FOR MULTICELLULAR EVOLUTION

Ta

NEOLECTA IRREGULARIS
* PLANT ROOT ASSOCIATED
* SO FAR UNCULTURABLE IN THE LAB
* FORMS SPOROCARPS - FRUITING BODIES

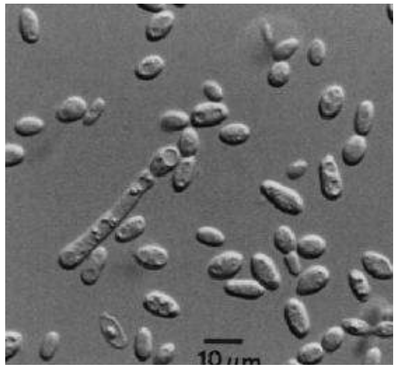


Schizosaccharomyces pombe



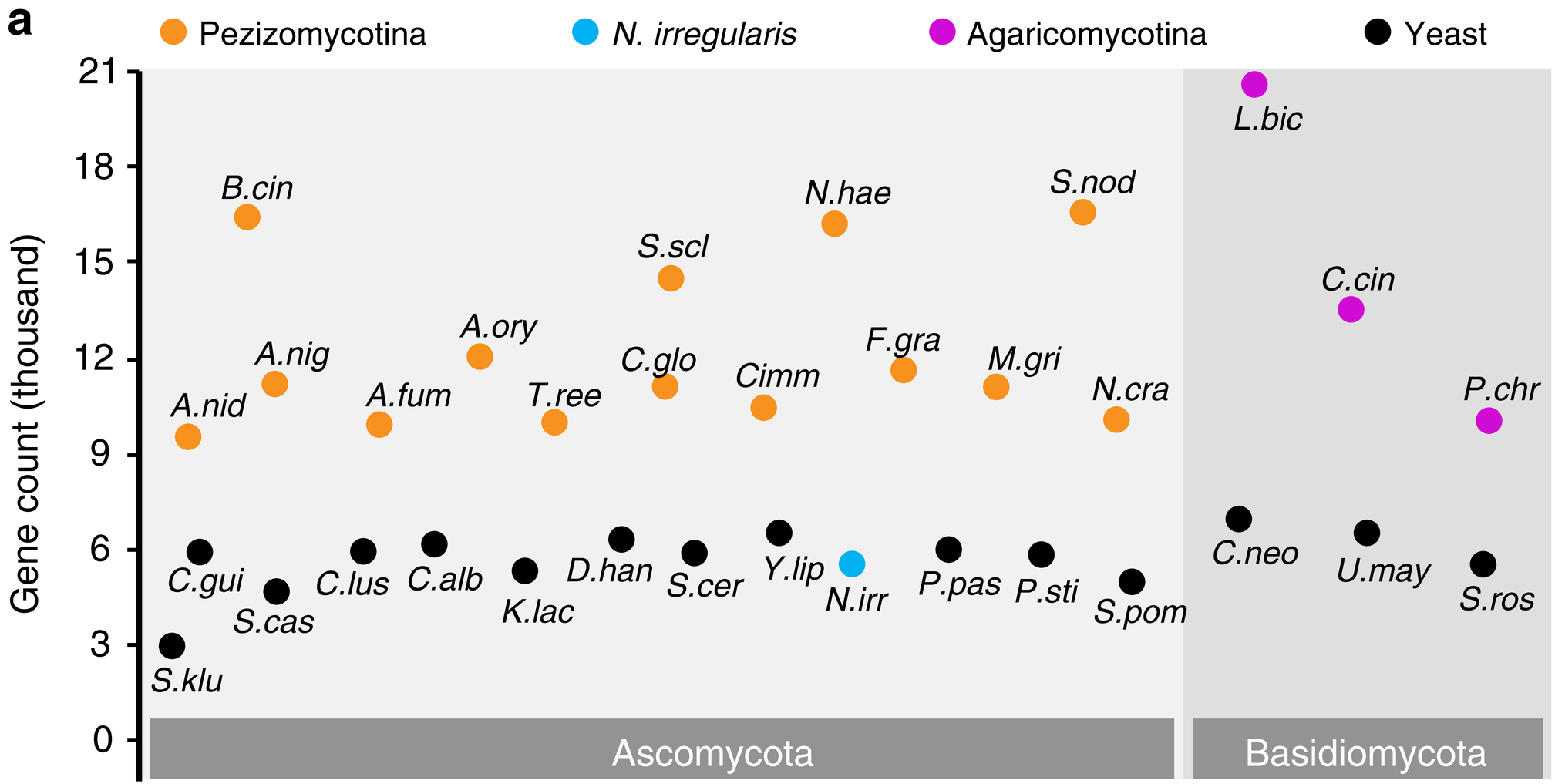
TAPHRINA DEFORMANS

NEOLECTA IRREGULARIS

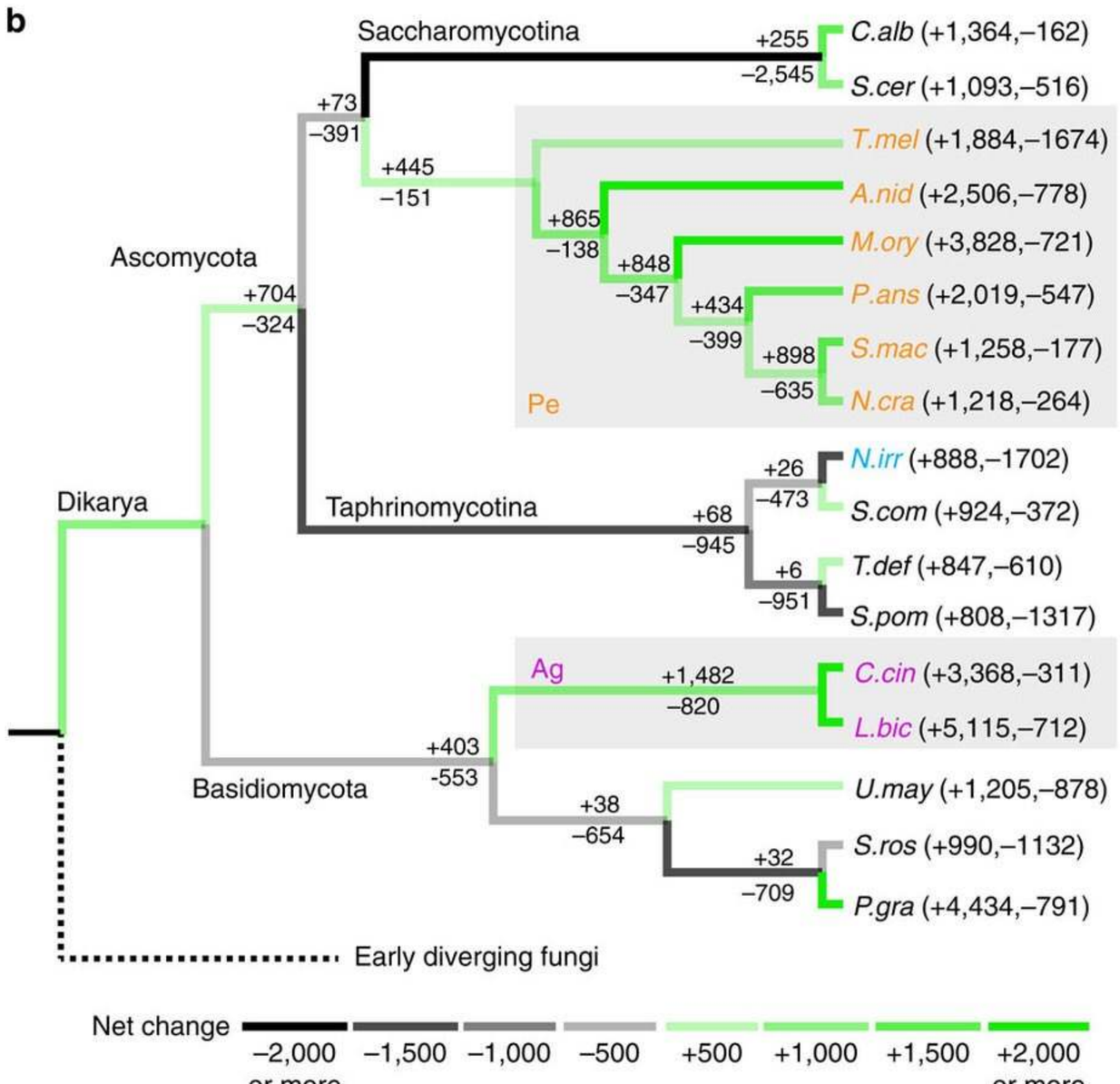


SAITOELLA COMPLICATA

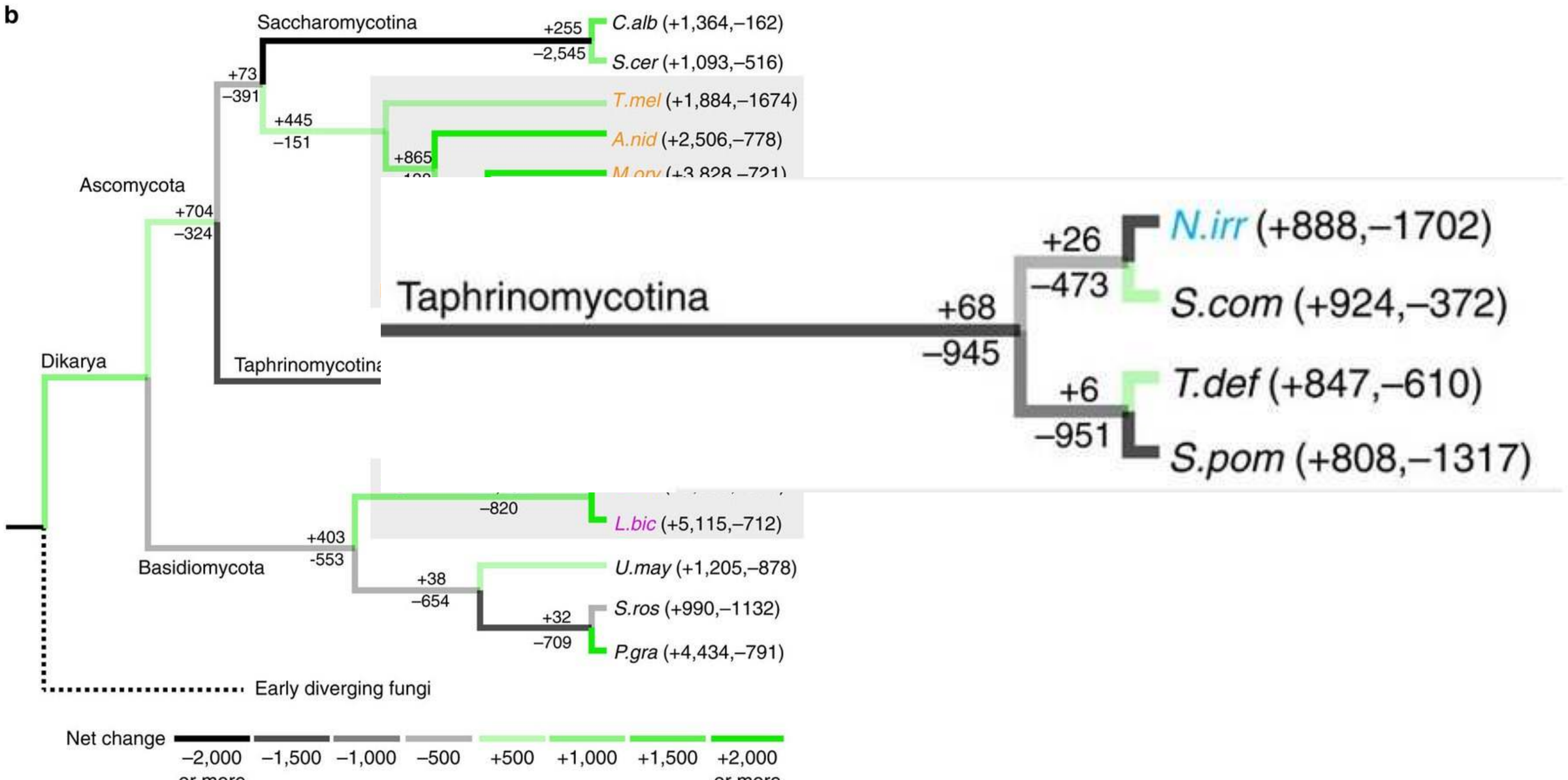
GENOME SIZE DOES NOT PREDICT COMPLEX MULTICELLULARITY



NEOLECTA LINEAGE DID NOT EXPERIENCE LARGE RECENT GAINS OF GENES



NEOLECTA LINEAGE DID NOT EXPERIENCE LARGE RECENT GAINS OF GENES

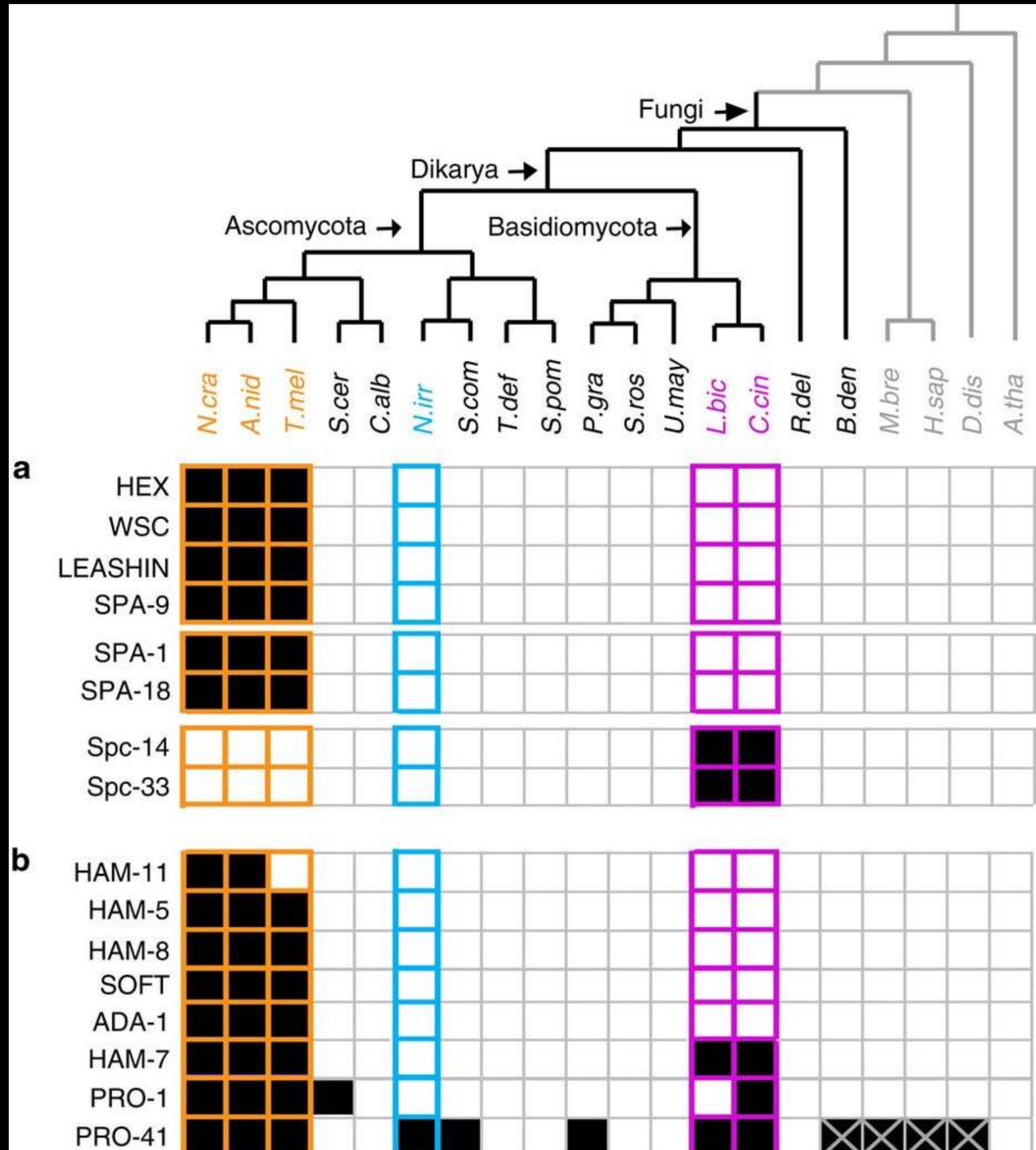


SEARCHING FOR COMPLEX MULTICELLULARITY (CM) SIGNATURES

- Are there genes that fungi that have CM share?
- These fungi make septated hyphae, with gating channels to separate cells.
  - When did these gating mechanisms evolve?
- Hyphal fusion and remodeling needed to make complex multicellular structures (e.g. fruiting bodies)

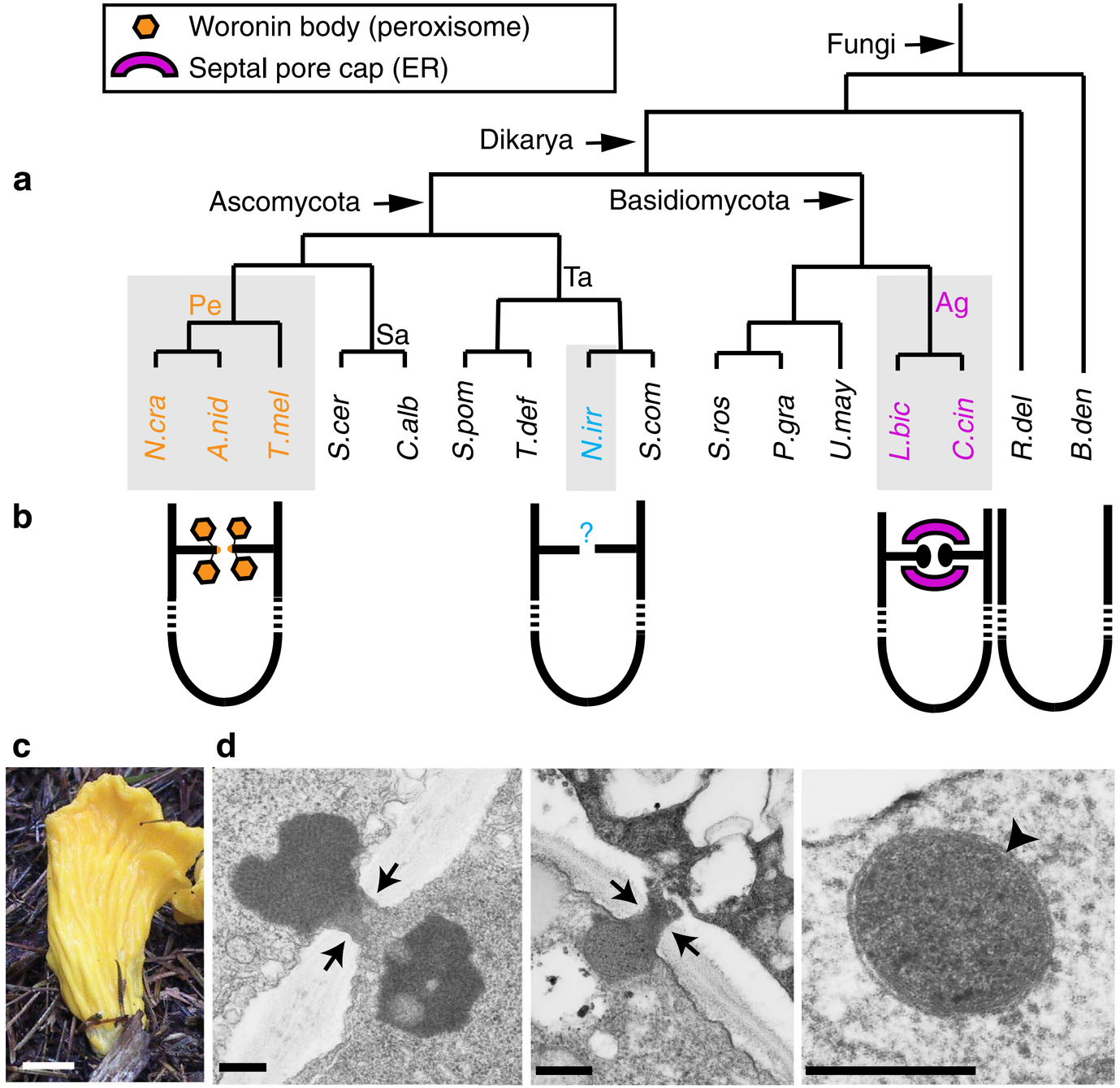
SEARCHING FOR CONSERVED GENES AMONG FUNGI WITH CM

- Search *N. crassa* proteins against collection of fungi
  - Identify those shared among CM fungi or lost in yeasts.
- 1,050 genes found in a CM associated
- 37% are absent and 47% highly divergent in yeasts
- over-represented functional categories: endomembrane transport and organization (transport routes, substrate transport, peroxisome) and aerobic respiration (electron transport and redox-related enzymes)

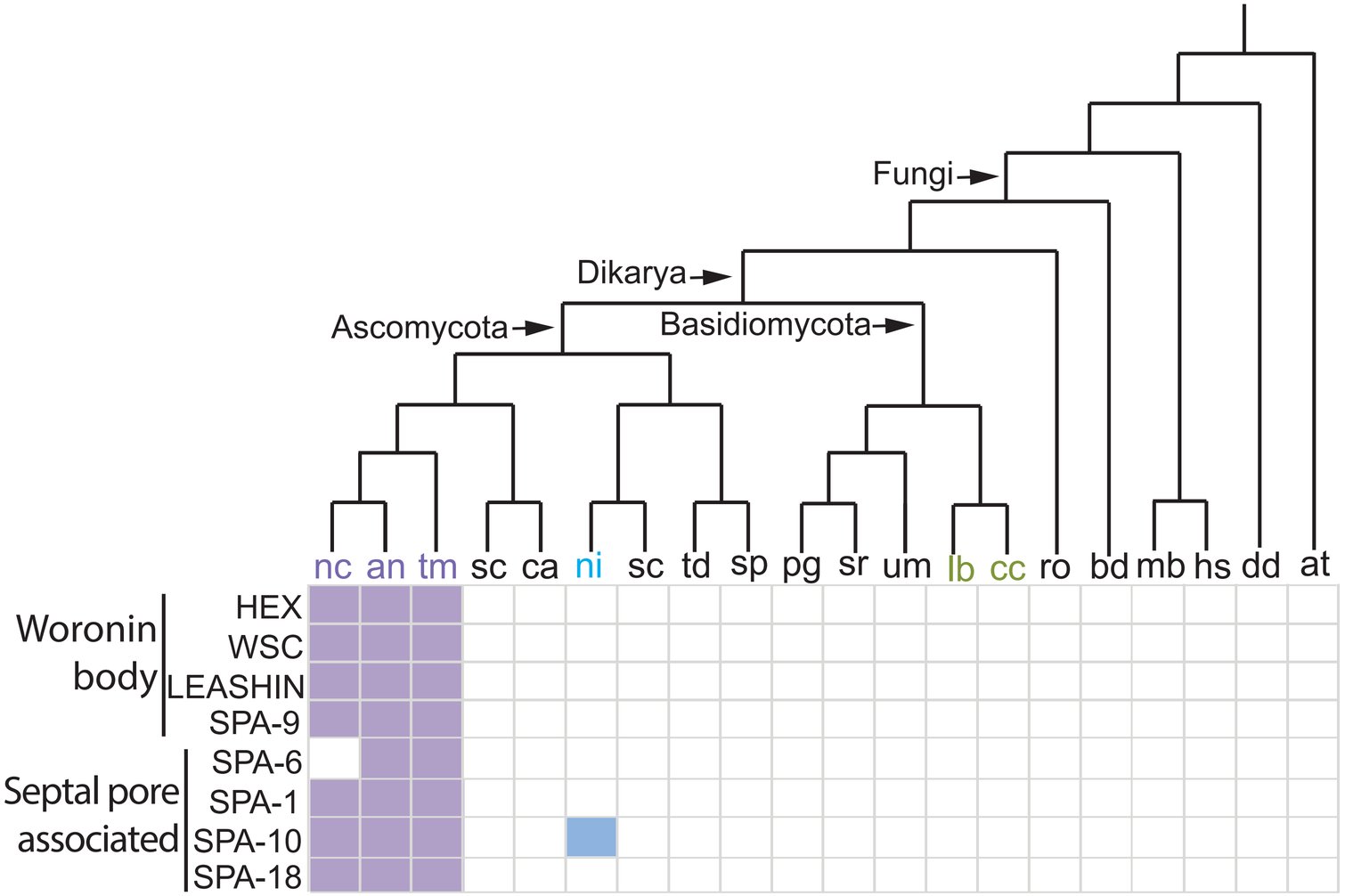


MECHANISMS FOR SEPTAL PLUGGING VARIES ACROSS THE FUNGI

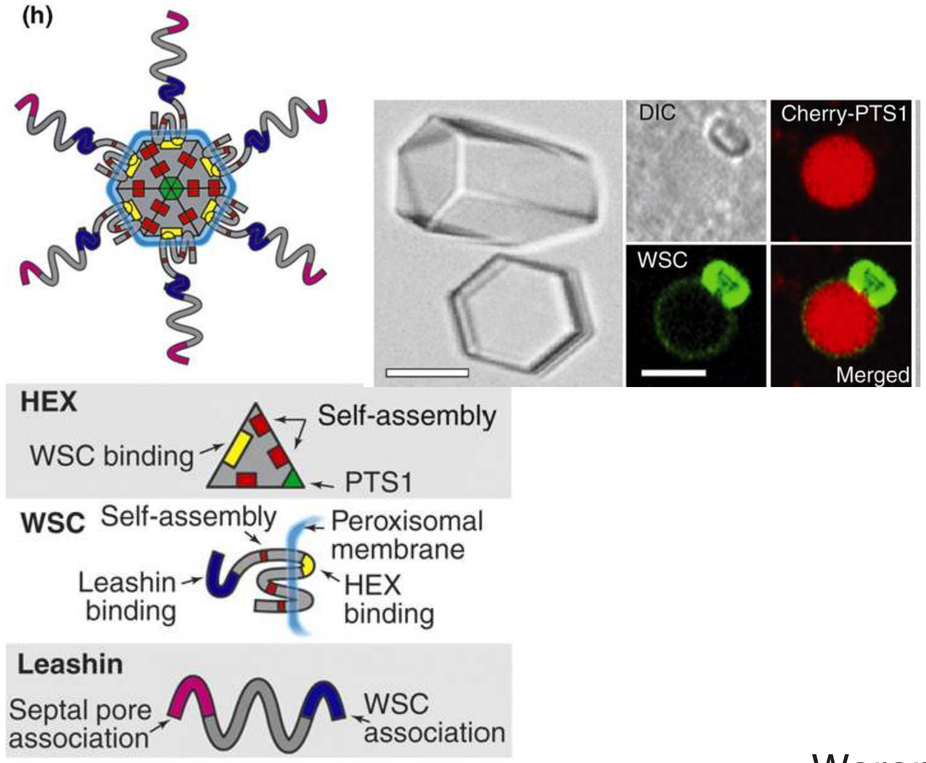
ARE MOLECULES SHARED THAT PERFORM THIS PLUGGING?



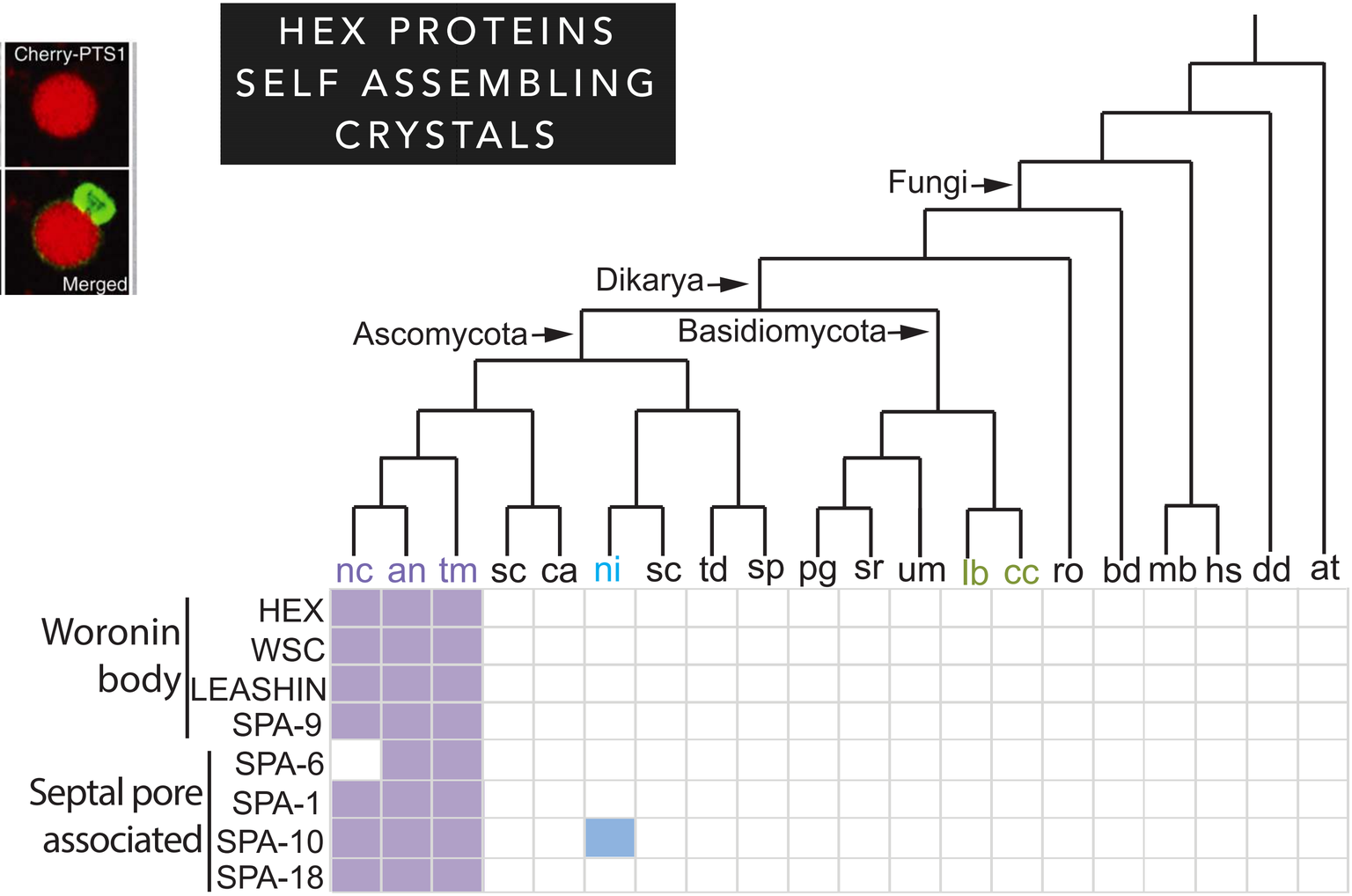
NO WORONIN BODYGENES IN NEOLECTA: RESTRICTED TO PEZIZOMYCOTINA



NO WORONIN BODYGENES IN NEOLECTA: RESTRICTED TO PEZIZOMYCOTINA

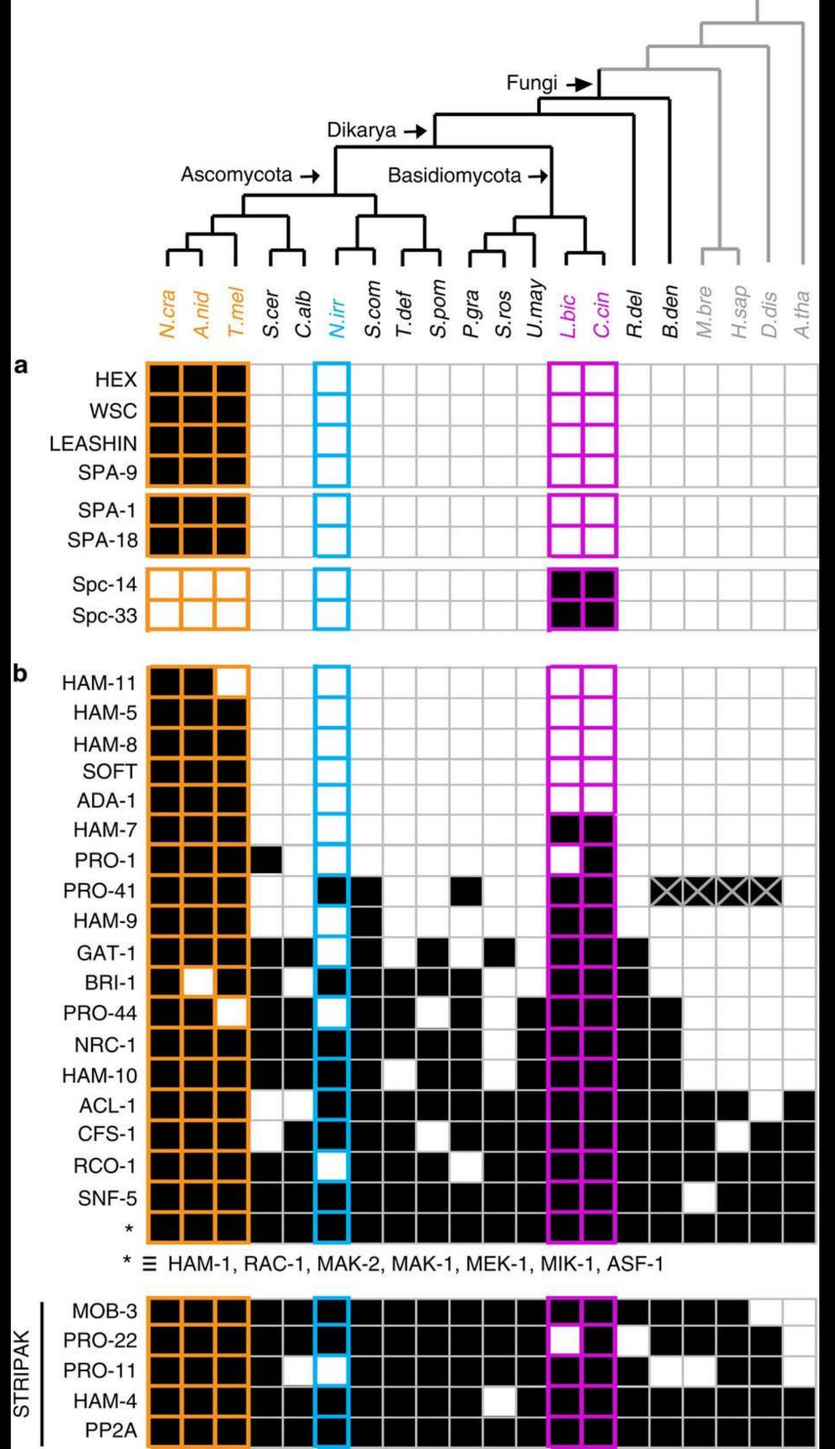


HEX PROTEINS SELF ASSEMBLING CRYSTALS



GENE SHARING PATTERNS ACROSS MULTICELLULAR FUNGI

- Septal pore gating and hyphal fusion proteins (a)
  - Pezizomycotina & Agaricomycotina specific proteins for pore gating are missing in Neolecta
- Some hyphal fusion proteins are missing Neolecta (b)
- Striatin-interacting phosphatases and kinases (STRIPAK) complex is ancient and conserved throughout. Important in development (c)

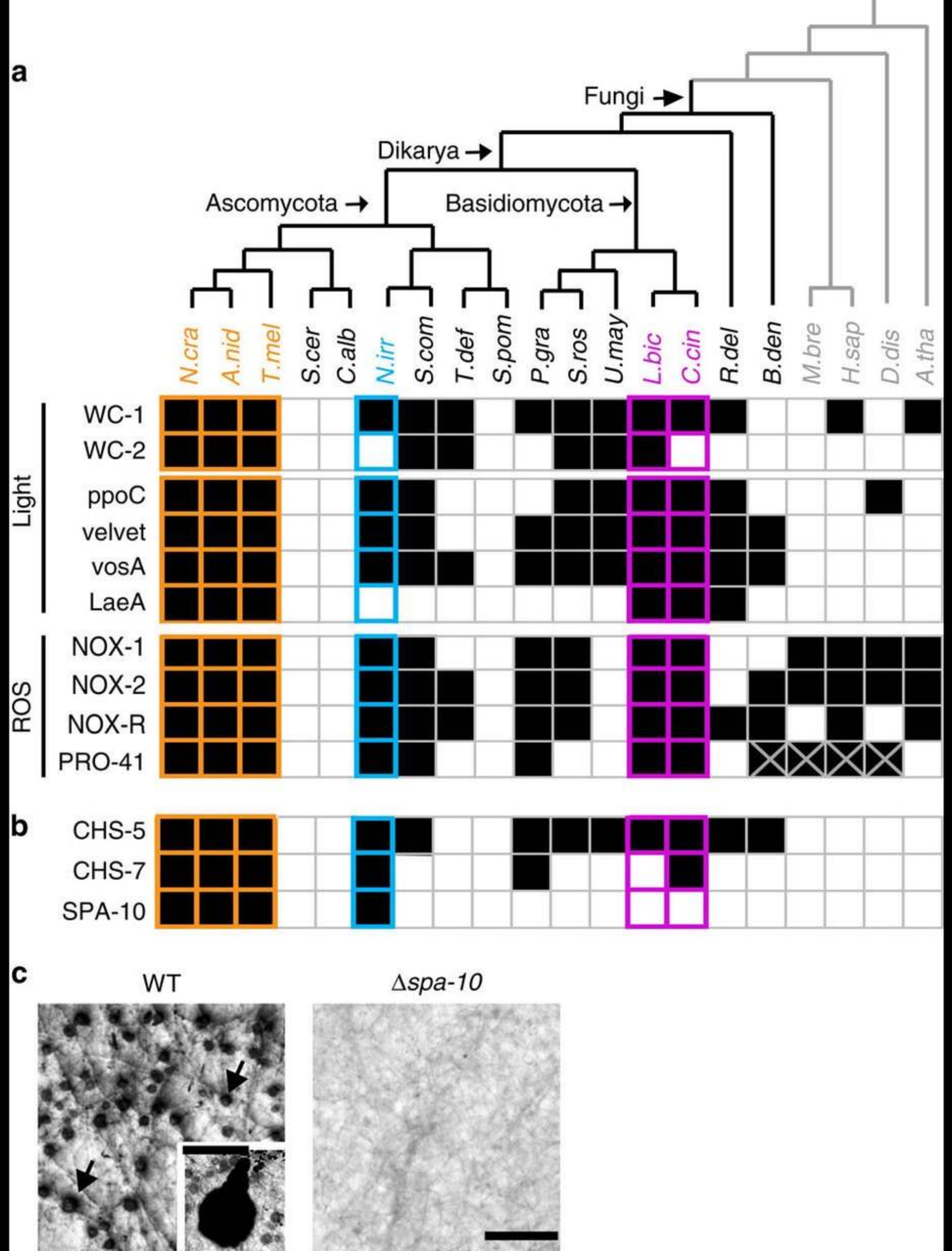


NGUYEN ET AL 2017

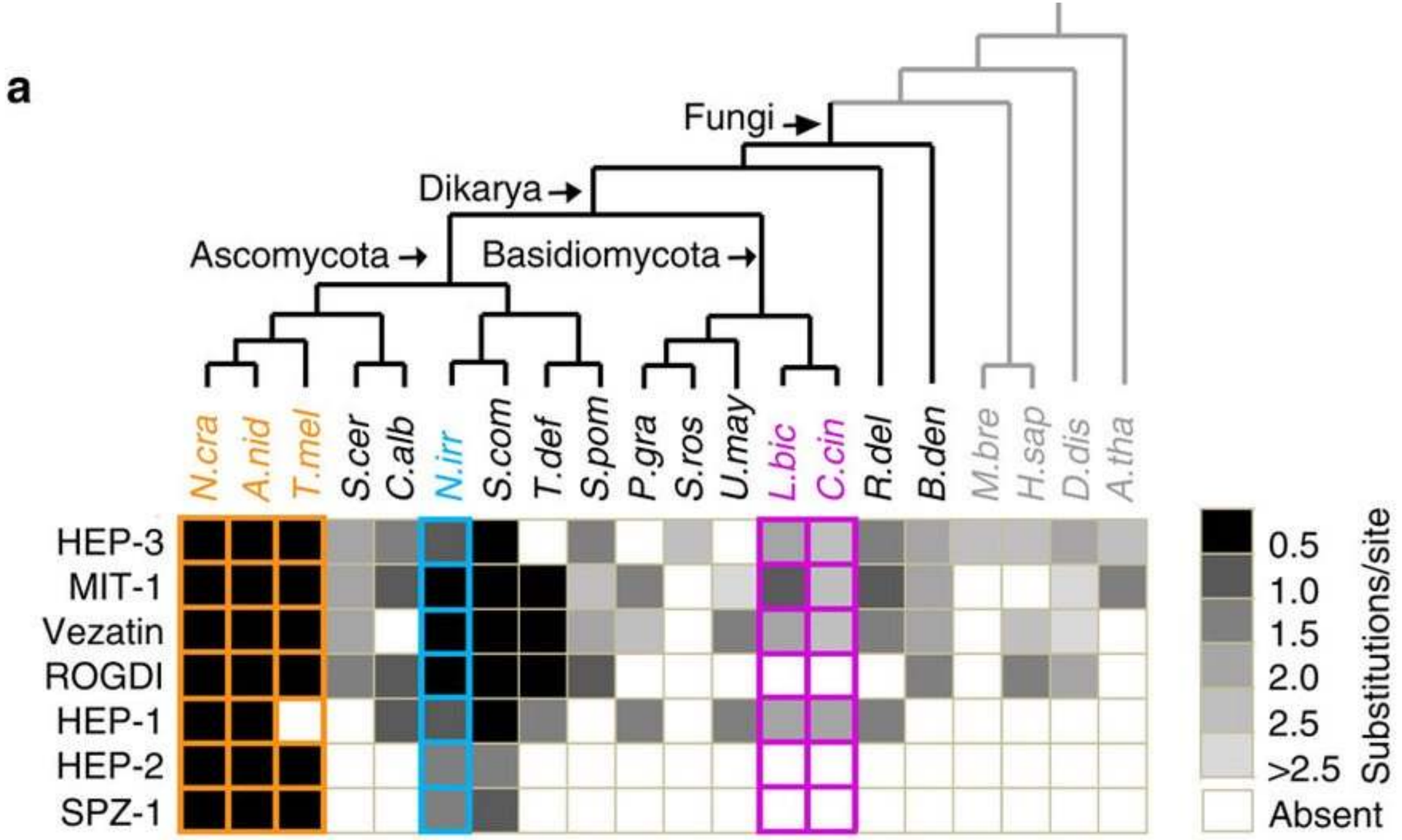
GENES SHARED AMONG SPECIES WITH COMPLEX MORPHOLOGY

- Proteins important for signaling and morphogenesis
- Reactive Oxygen Species Signaling
- Remodeling and construction

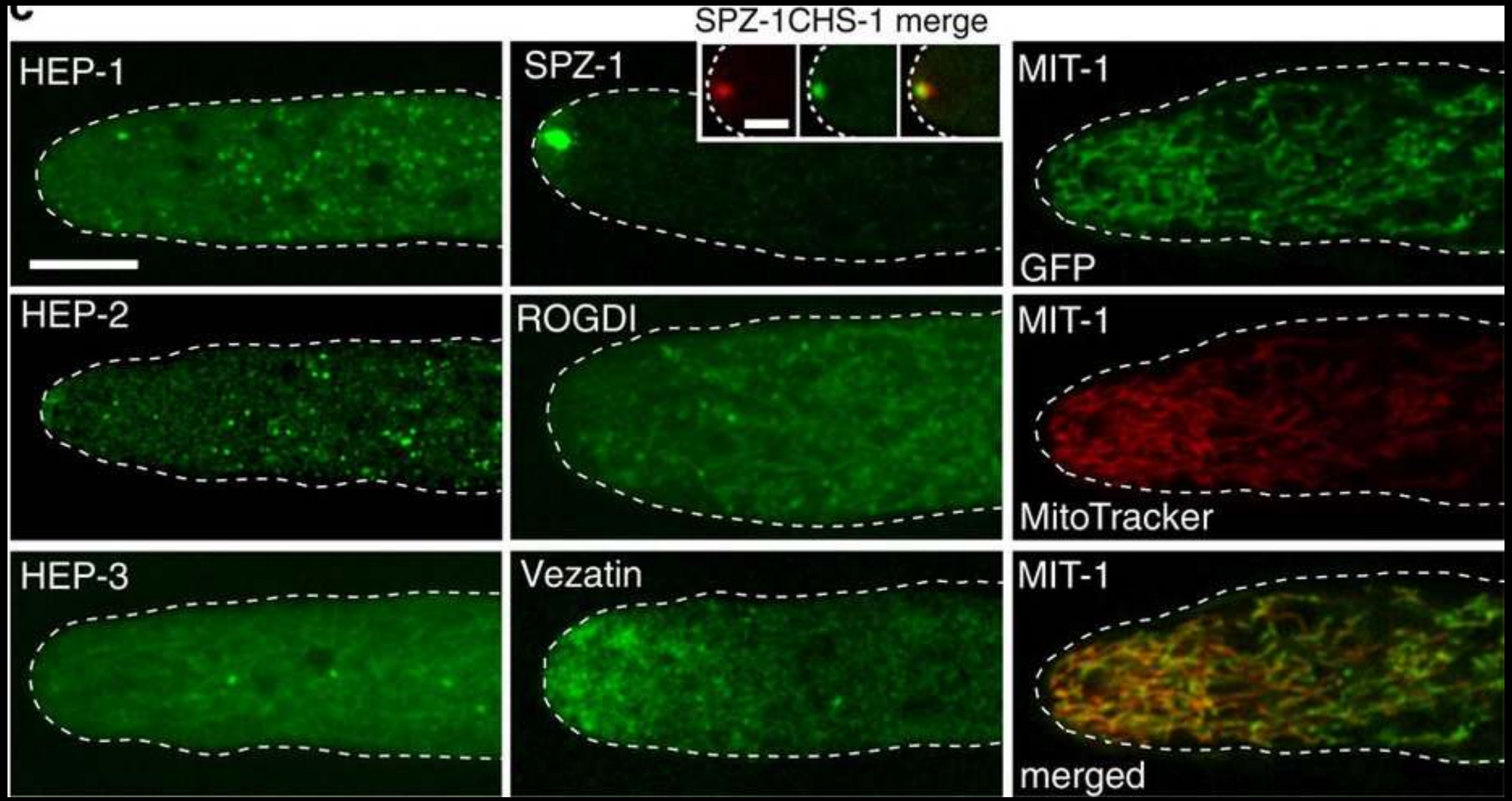
- Δ*spa-10* are female sterile



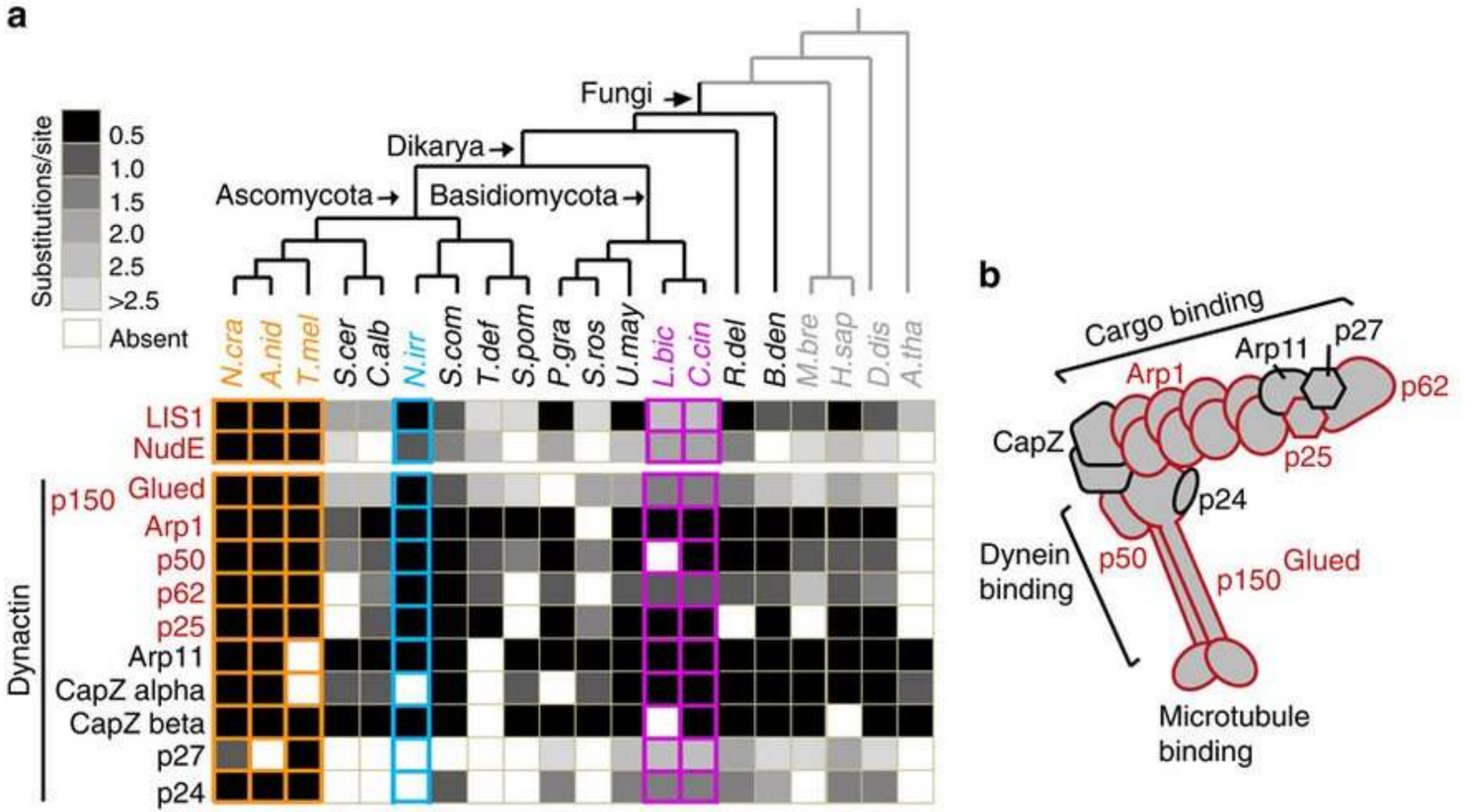
THERE ARE ALSO NOVEL PROTEINS IMPLICATED IN COMPLEX MULTICELLULARITY

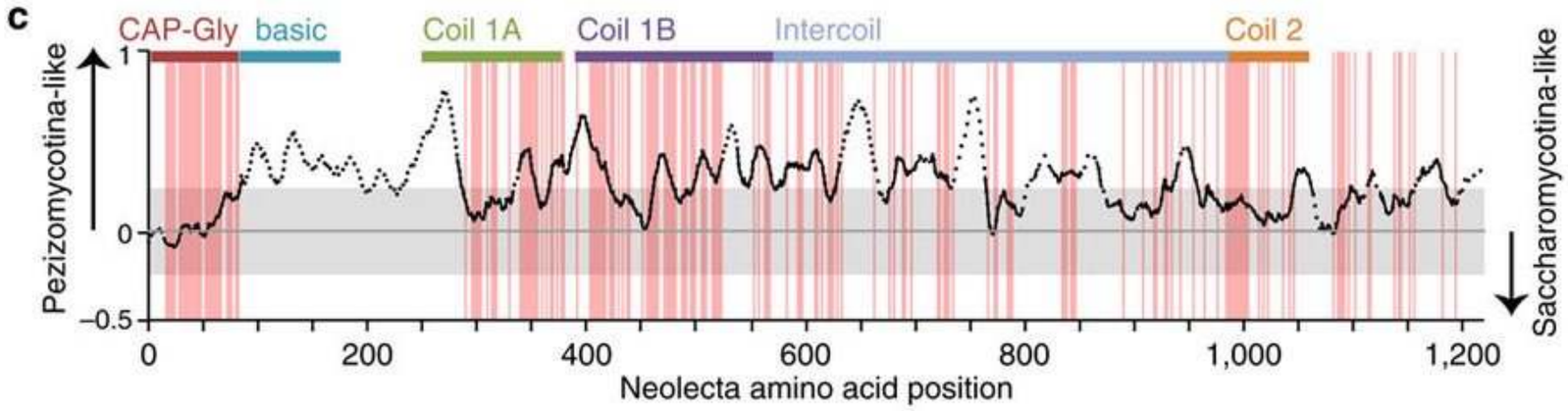


Novel proteins' localization
Enriched for transmembrane domains
MIT-1 is novel mitochondrial localized protein



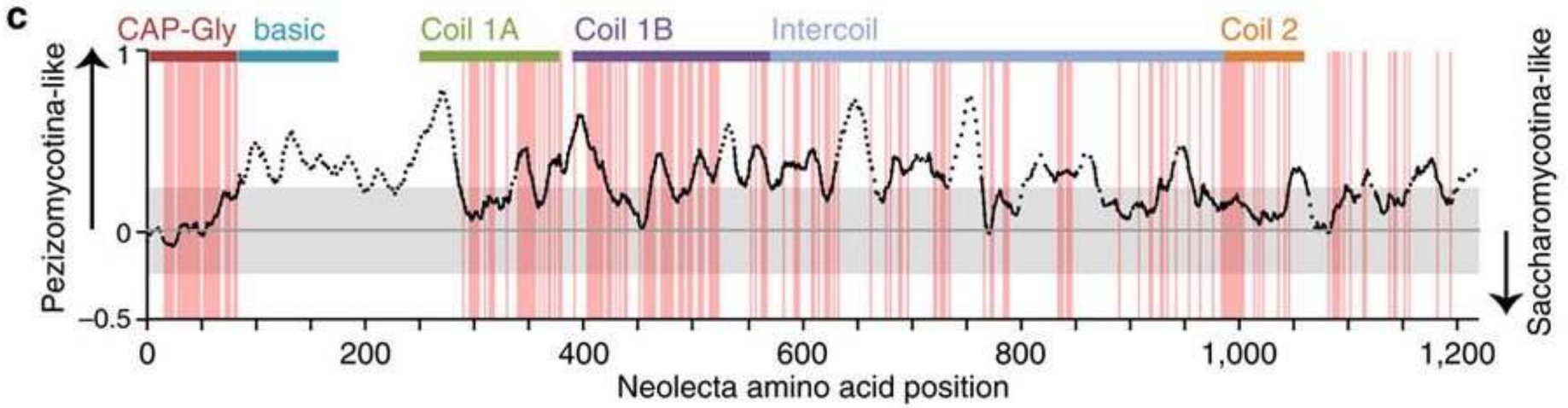
DYNEIN AND ITS REGULATORS HAVE A COMPLEX MULTICELLULARITY SIGNATURE OF CONSERVATION





P150-GLUED PROTEIN

REGIONS OF CONSERVATION IN P150^{GLUED} PROTEIN

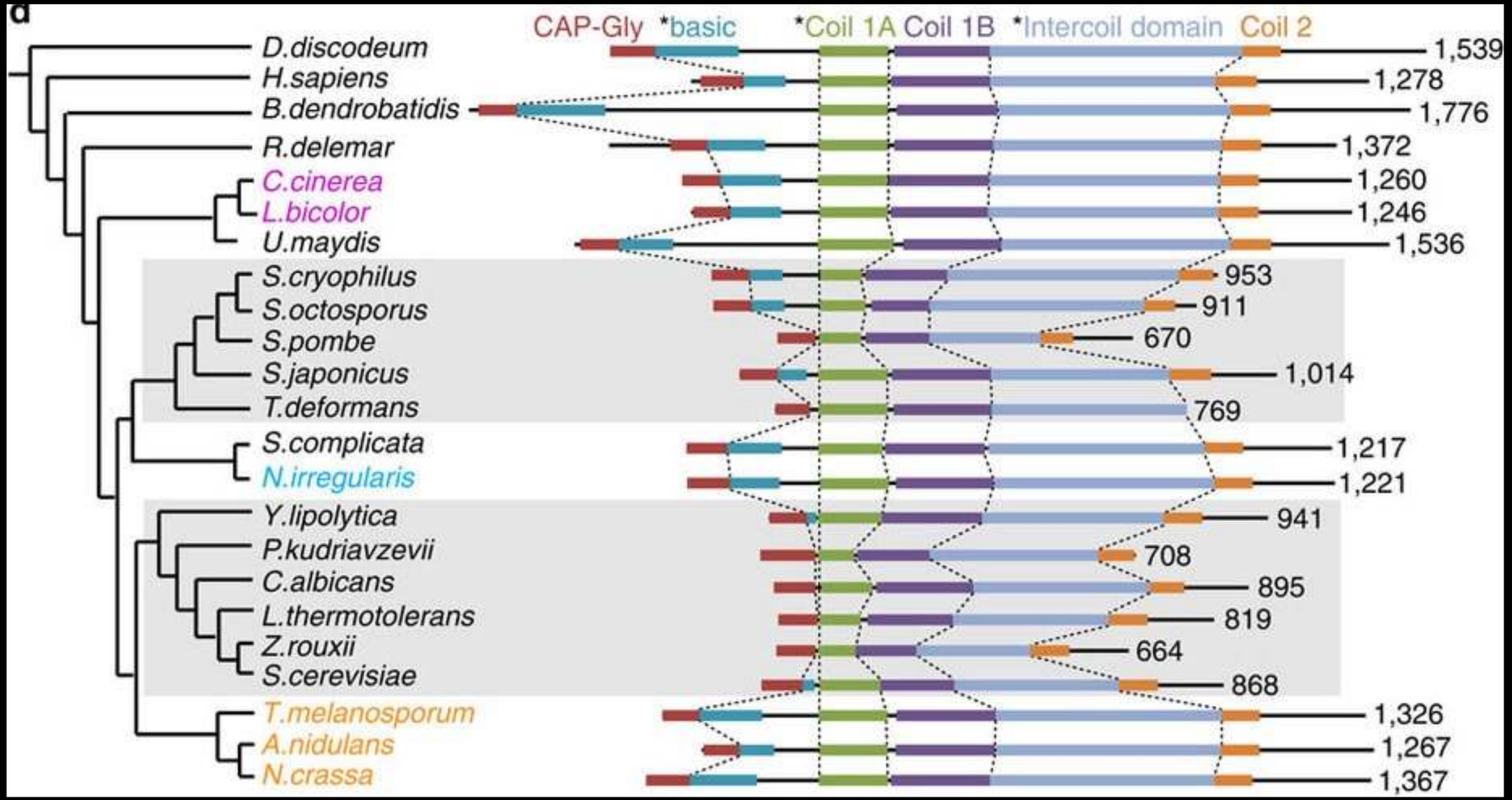


P150^{GLUED} PROTEIN

DOTTED LINES ARE MISSING IN YEAST - LOSING REGIONS RELATED TO MOTOR PROCESSIVITY

NEOLECTA P150^{GLUED} IS MORE SIMILAR TO PEZIZOMYCOTINA ORTHOLOGUES OVER THE ENTIRE LENGTH OF THE PROTEIN

INDEPENDENT DYNEIN PROTEIN CONTRACTION IN YEASTS



SUMMARY

- Patterns of gene loss in fungi from ancient genes: e.g. flagella, also independent losses dominate in formation of many yeast lineages
- Some aspects of complex multicellularity in fungi is likely not a dramatic gain of genes. Losses of Transcription factors and families in yeast lineages (e.g Laszlo Nagy et al Nat Com 2014.)
- Complexity may be result of repurposing or redirection of existing core genes.
- Ancestral fungus had gene more extensive than what we see in models yeasts
- Utility of comparing similarities of retained genes in species with complex morphology

FIN