A Comparative Genomic Investigation of Fungal Genome Evolution

Jason Stajich
Duke University
University Program in Genetics & Genomics

Evolutionary genomics

Evolutionary & Organismal Biology

Phylogeny
Population genetics and structure
Phenotype
Ecological adaptations

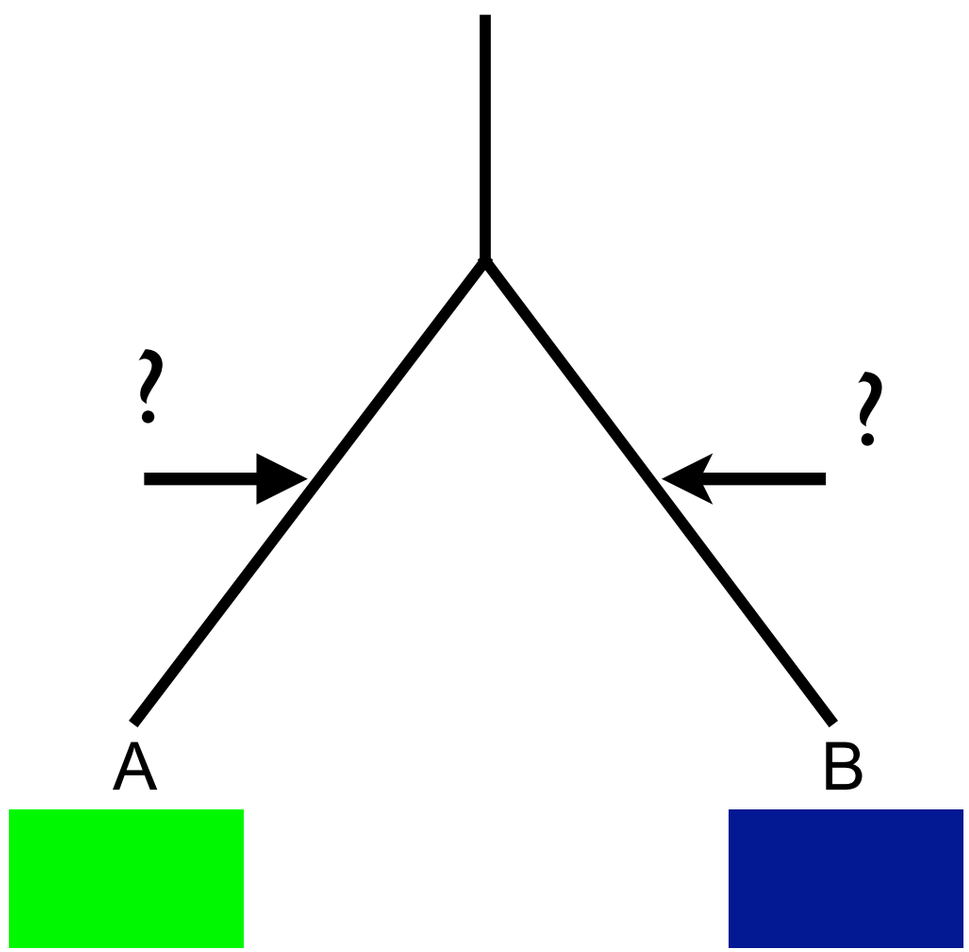
Comparative Genomics

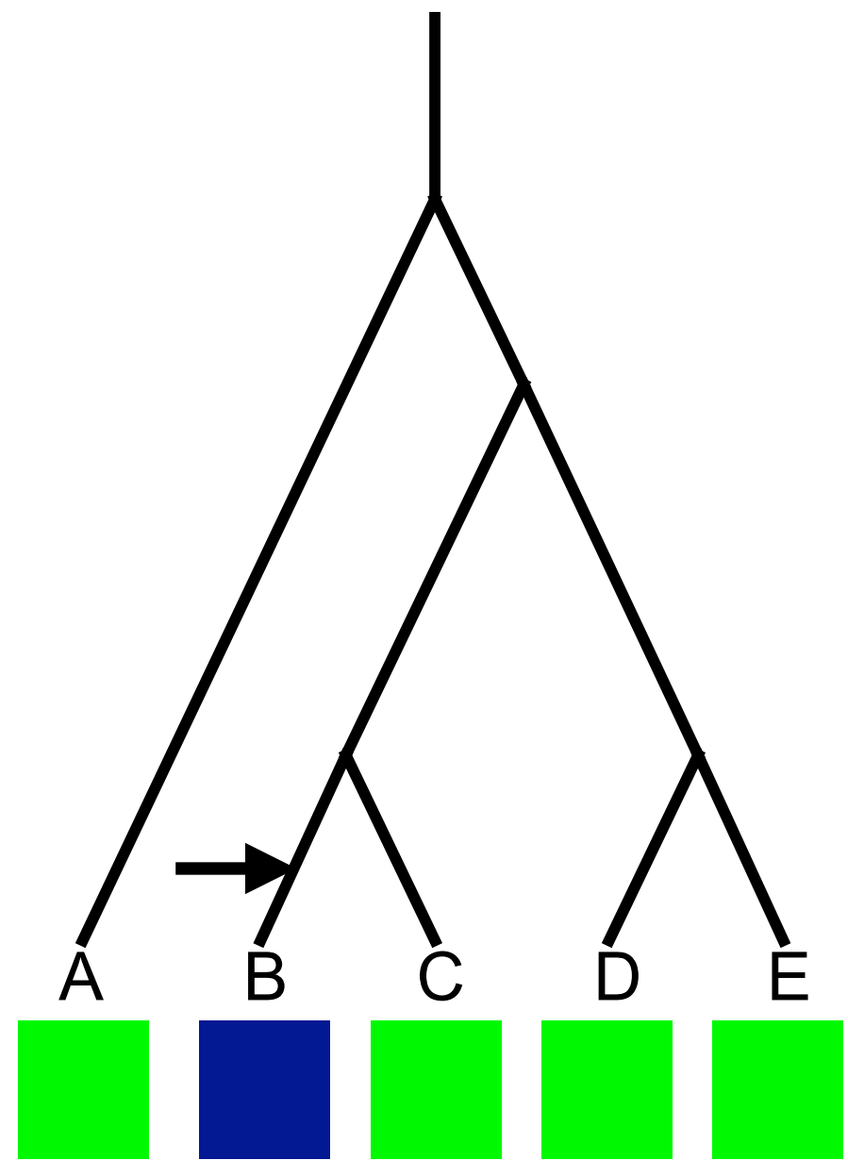
Molecular evolution
Gene order
Gene families
Gene and genome structure
Gene content
Conserved elements
Rates of molecular evolution
Gene function inference

Model Systems

Genetic tools
Gene function & expression
Regulatory networks
Pathways
Molecular & cellular biology
Disease models

Power of the comparative approach





Industrial uses of fungi

- Bread, beer, wine - *Saccharomyces cerevisiae*
- Sake and soy sauce - *Aspergillus oryzae*
- Dairy - *Penicillium roqueforti*, *Kluyveromyces lactis*
- Citric acid - *Aspergillus niger*
- Riboflavin - *Ashbya gossypii*
- Stonewashed jeans - *Trichoderma reesei*
- Penicillin antibiotic - *Penicillium notatum*
- Button Mushrooms - *Agaricus bisporus*

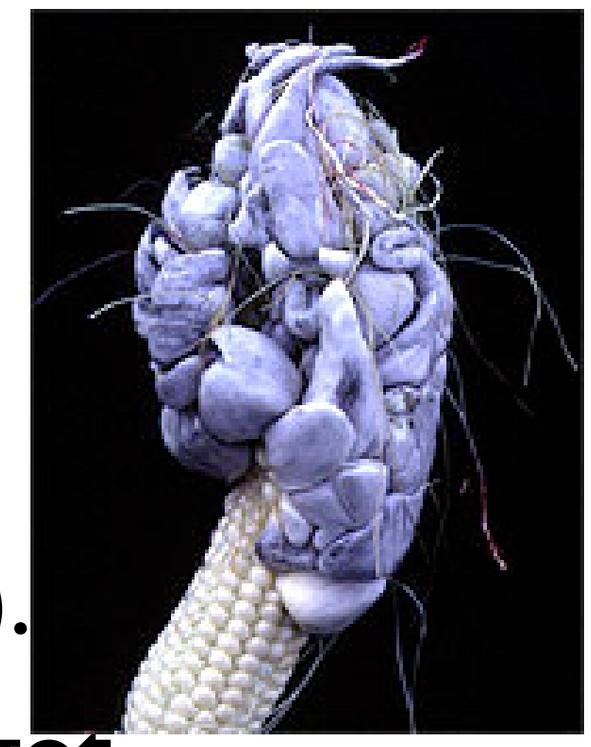
Agricultural impact of fungi

s of plant disease is caused by

www.gov.mb.ca

USDA

A.G. Bölker



posit mycotoxins - e.g. ergot

tal fungi provide nutrient

and nitrogen fixation

Impact of fungi on human health

- Mostly immunocompromised individuals are at risk of life-threatening infections
- Primary pathogens
  - *Histoplasma, Coccidioides, Cryptococcus gattii*
- Opportunistic pathogens
  - *Candida albicans, Aspergillus fumigatus, Cryptococcus neoformans, Rhizopus oryzae*

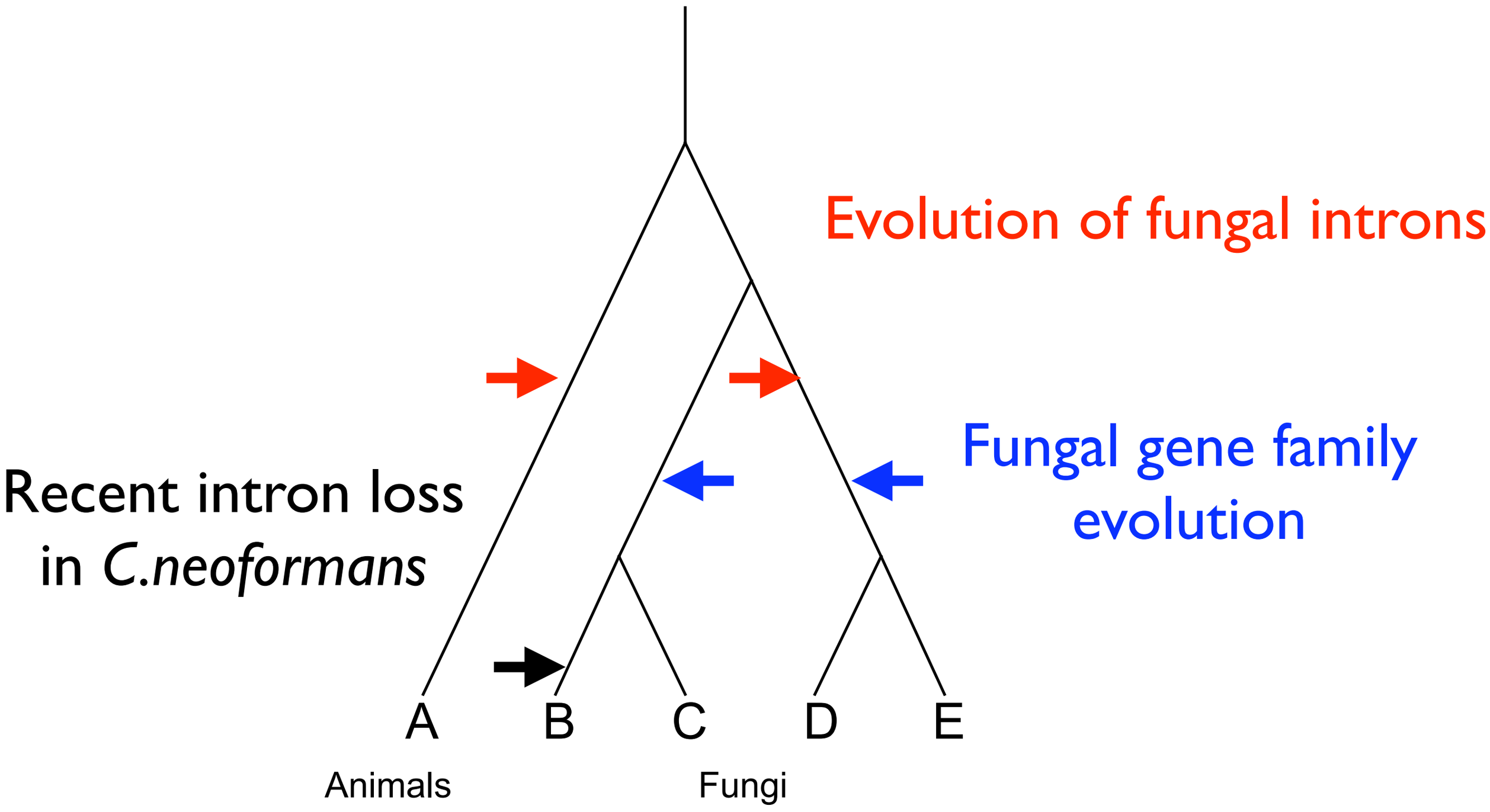
Fungi as genetic models

- Beadle and Tatum (1941) - one gene, one enzyme hypothesis in *Neurospora crassa*
- Cell cycle, cell model - *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*
- Molecular biology tools to investigate phenotype-genotype
- Evolutionary models

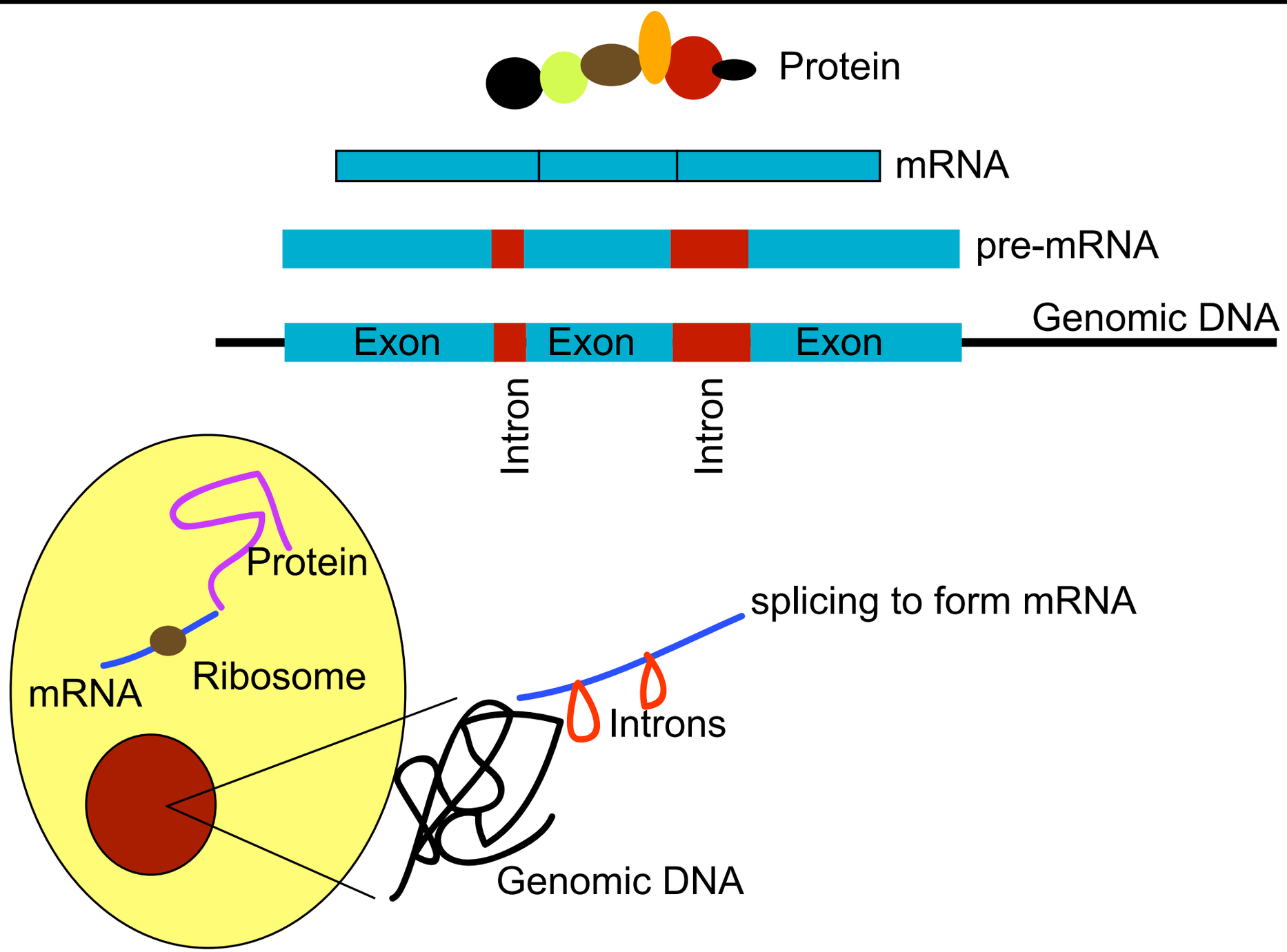
Fungal genomes

- Smaller than most vertebrate and plant genomes
  - *A. gossypii* 8.5 Mb; *S. cerevisiae* 12 Mb
  - *N. crassa* 40 Mb
  - Animals: 100 Mb worm; 3000 Mb Human
- Vary in protein coding gene content
  - 4,700 in *A. gossypii*; 5,800 in *S. cerevisiae*
  - 16,000 in *R. oryzae* or *S. nodorum*
  - 19,000 in Fruitfly; 25,000 in worm

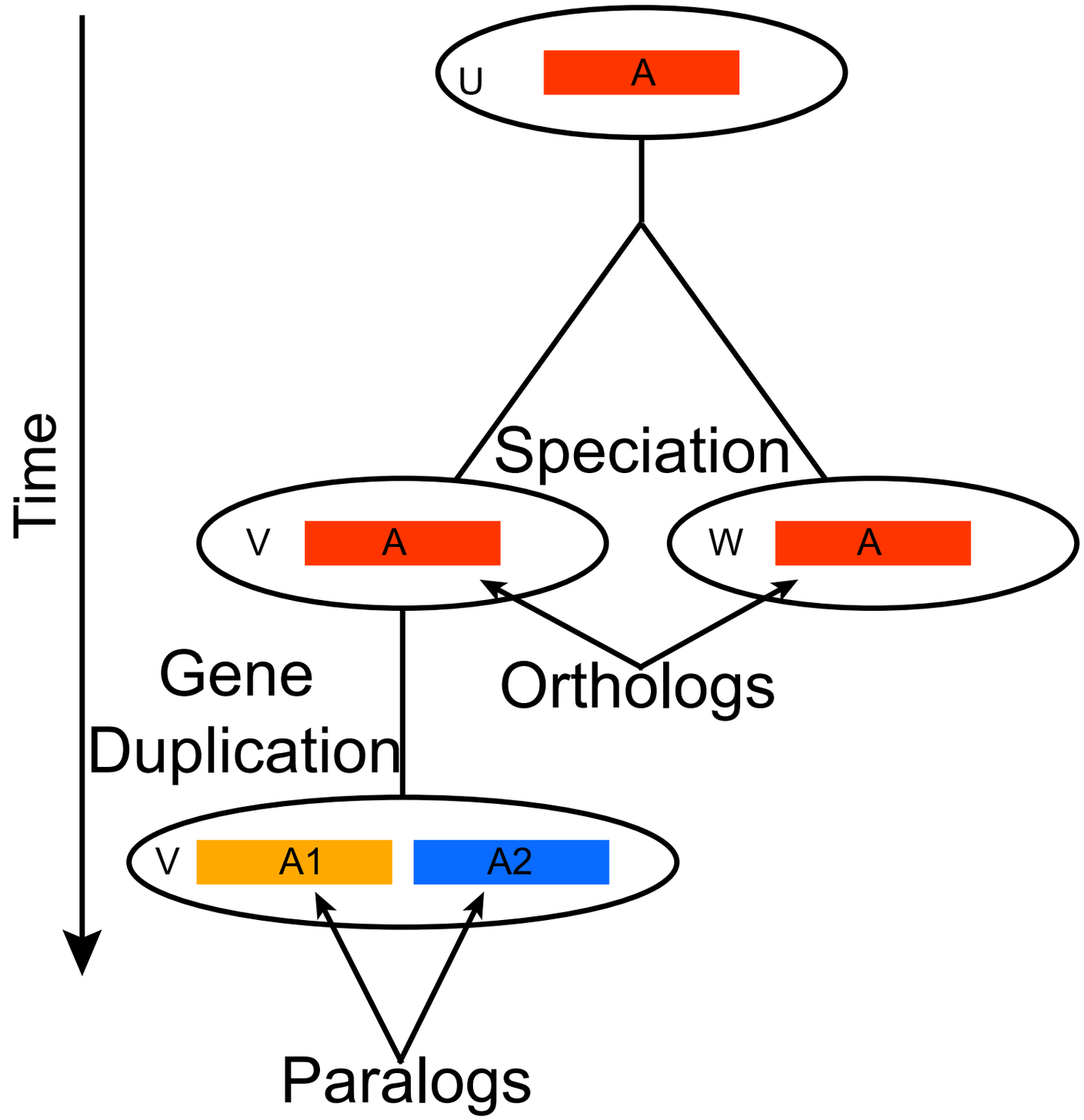
Fungal comparative genomics



Central dogma of eukaryotic biology

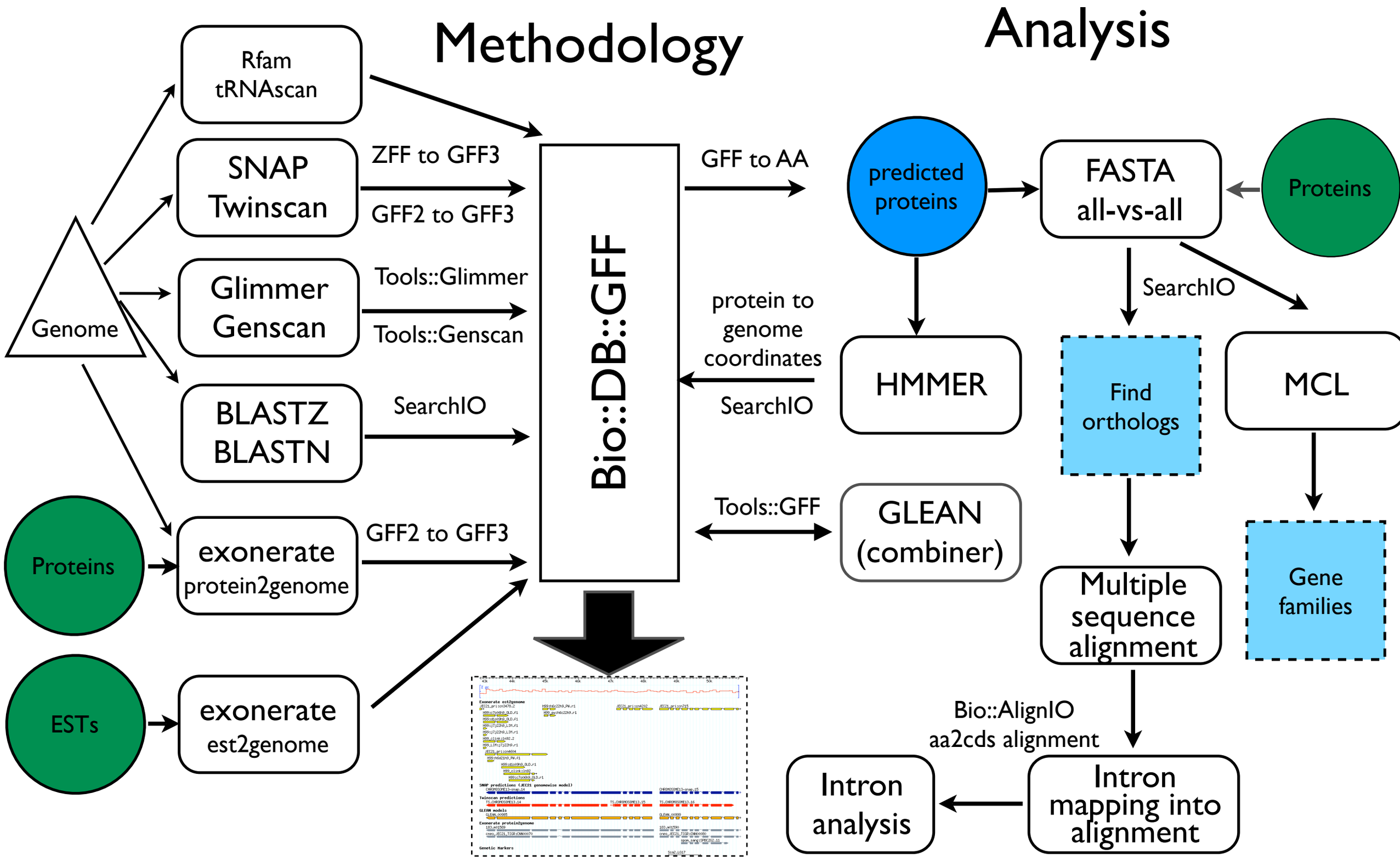


Orthologs and Paralogs



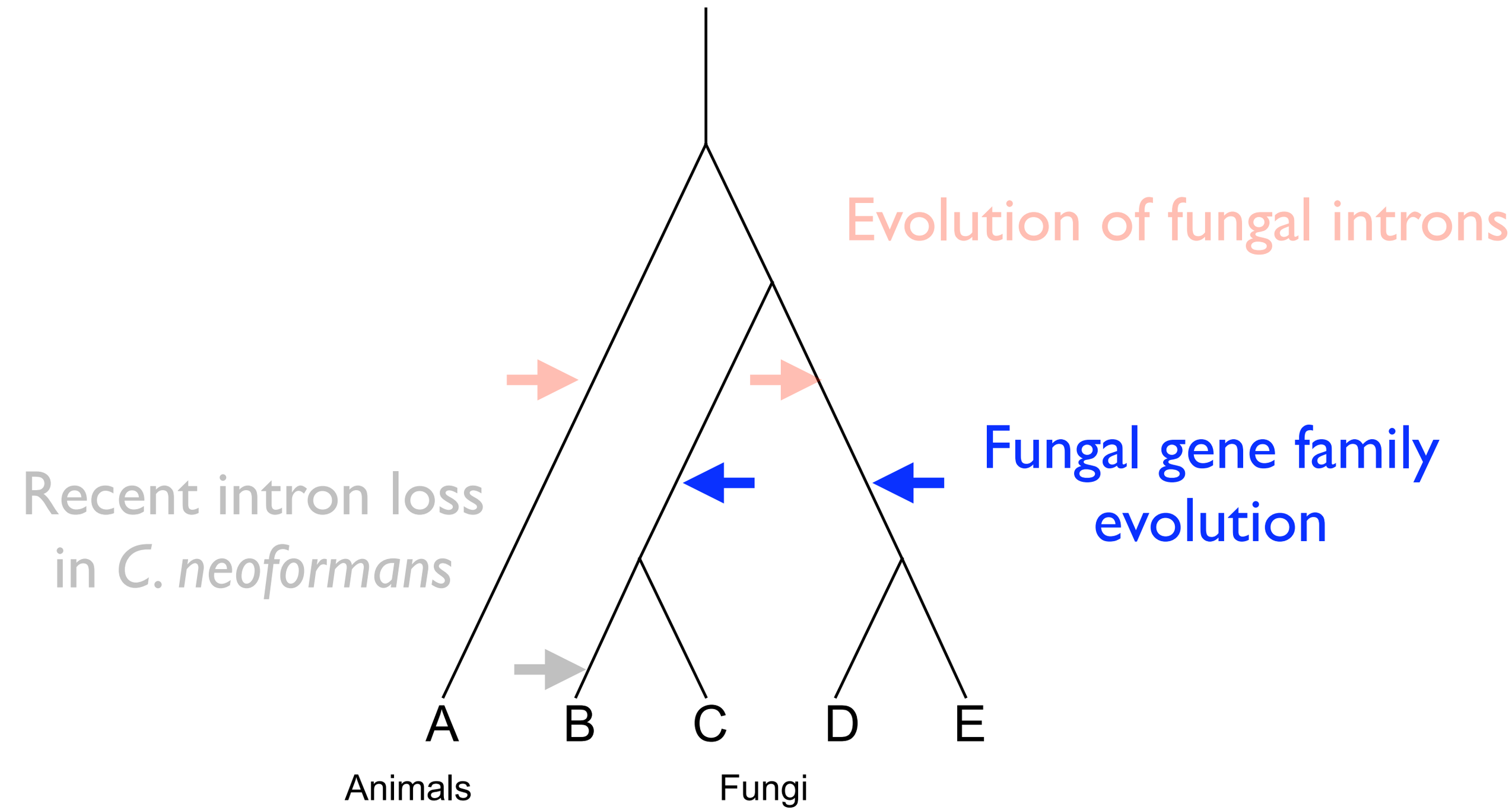
Genome annotation

- Available fungal genomes were only assembled genomic sequence.
- Need systematic and consistent gene predictions for genome comparisons
- Automated annotation pipeline for gene prediction.



http://fungal.genome.duke.edu

Fungal comparative genomics



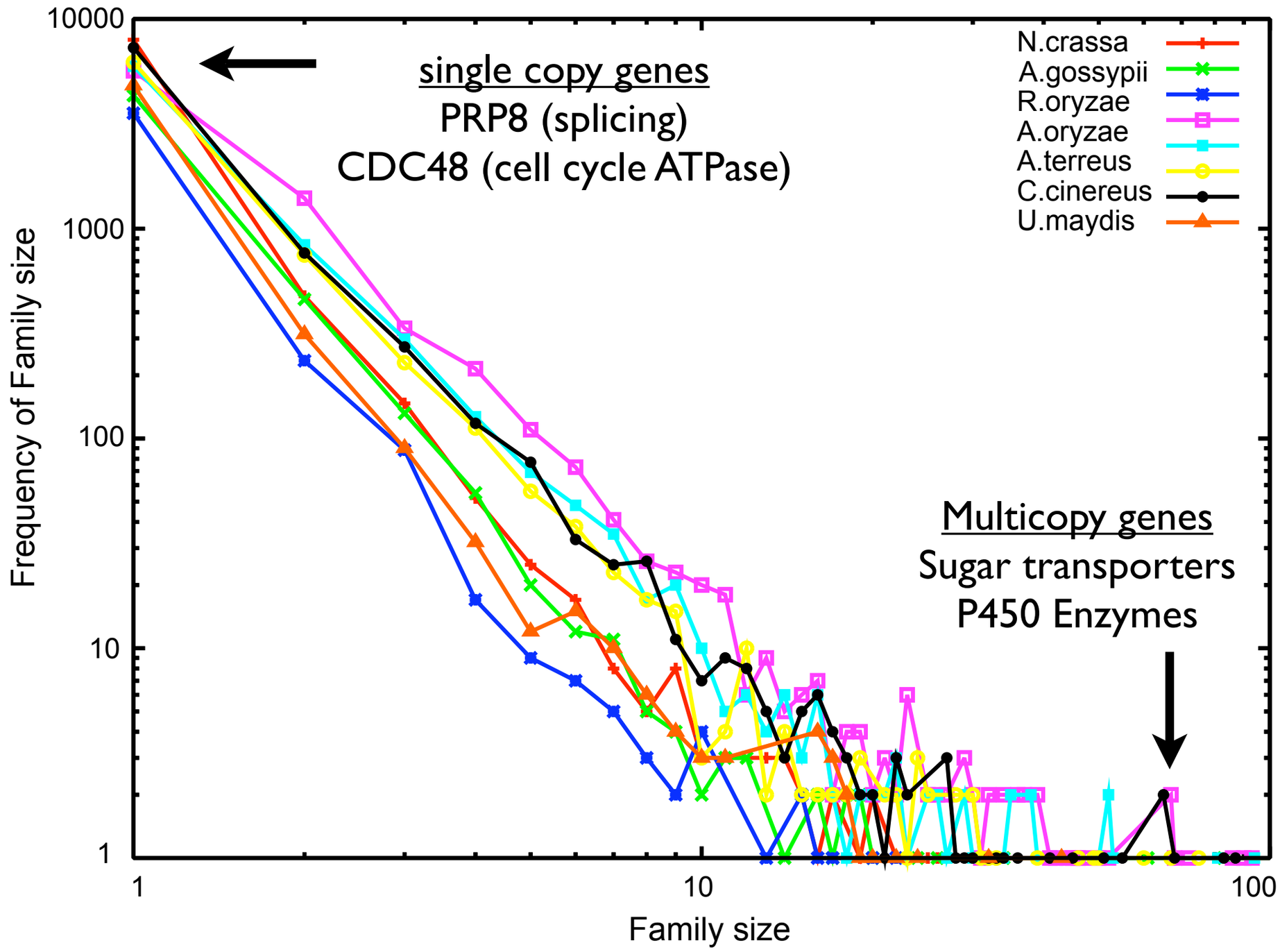
Gene family evolution

- Gene families are the crucible of new genes and thus new functions
- Signature of adaptive evolution often confounded in multi-gene families
- Can we identify families that are have unexpectedly large changes in size across a phylogeny?
  - Follow up these families with more focused studies

Identifying gene family expansions

- Previous work only considered pairwise
- *Ad hoc* comparison of gene family sizes
  - *C.elegans*-*C.briggsae* - GPCR family expansions (Stein et al, *PLOS Biology* 2004)
  - *A. gambiae*-*D. melanogaster* - Mosquito specific family expansions related to symbiotic bacteria (Holt et el, *Science* 2002).

Gene family sizes follow power law distribution

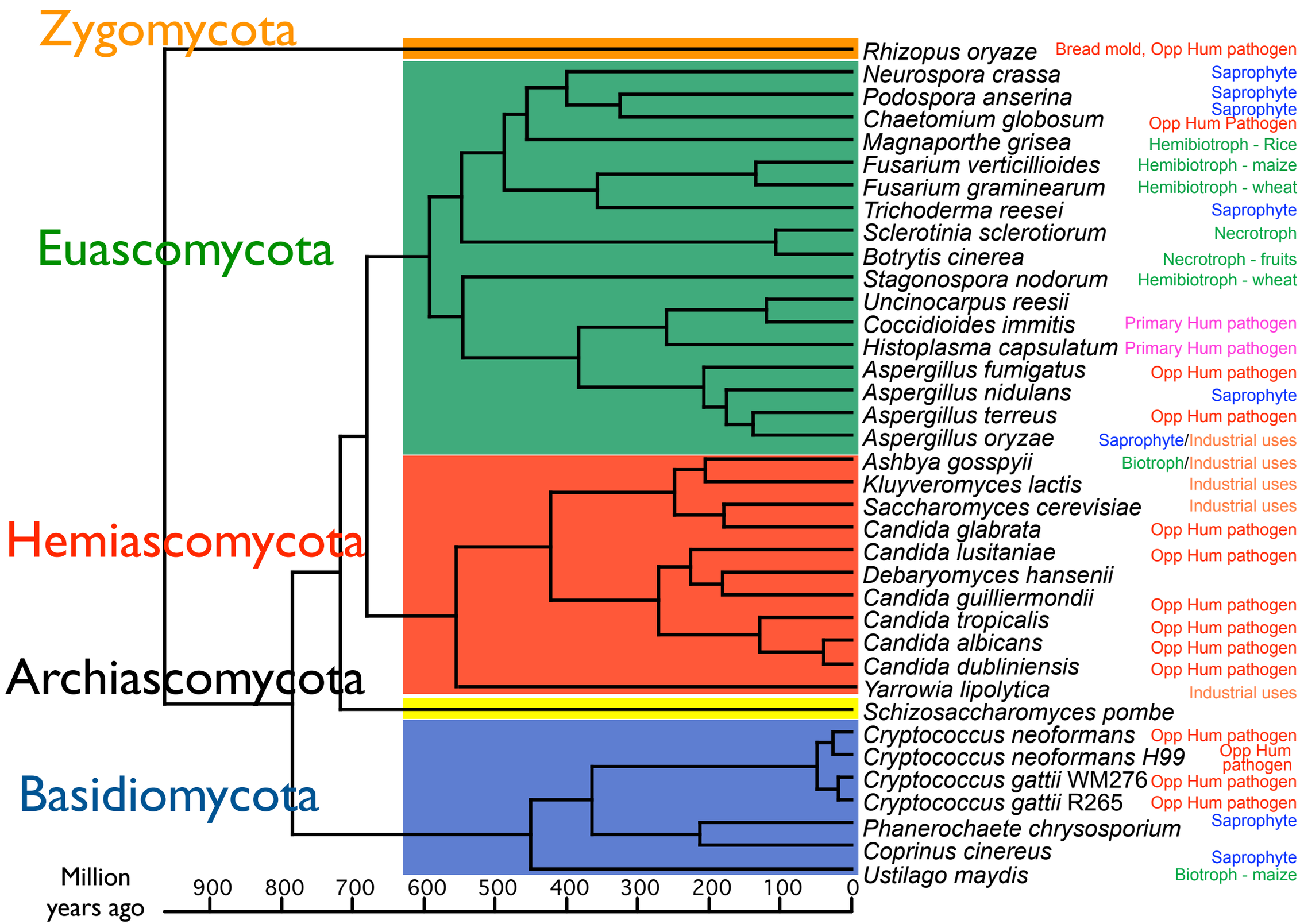


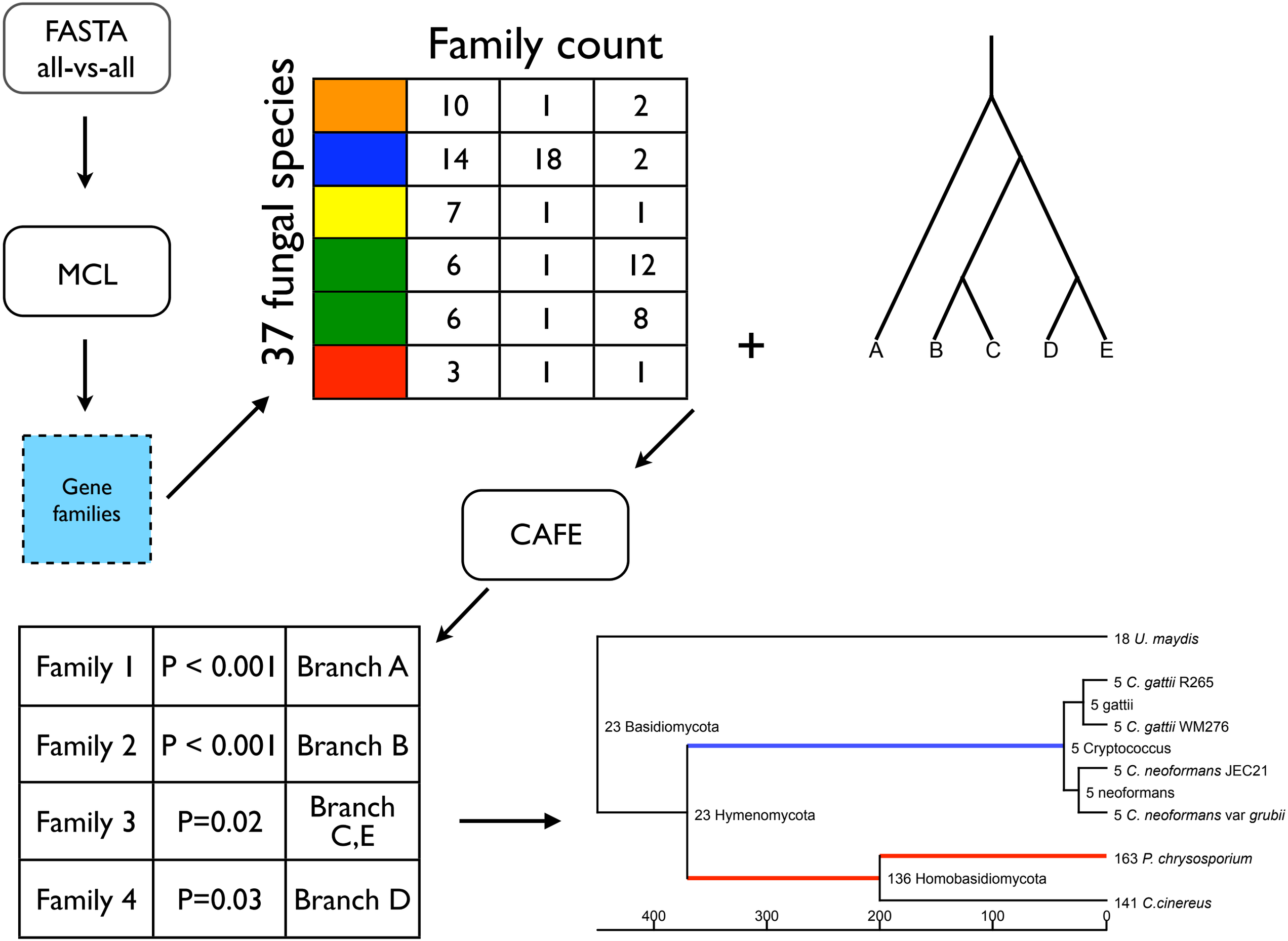
Phylogenetic evaluation of gene family size change

- Previous methods only used *ad hoc* statistics
- Explicit model for gene family size change according to a Birth-Death (BD) models
- Apply BD to family size along phylogeny using probabilistic graph models
- CAFE - Computational Analysis of gene Family Evolution

Hahn et al, *Genome Res* 2005
De Bie, et al *Bioinformatics* 2006
Demuth et al, *submitted*

Fully sequenced fungal genomes




FASTA
all-vs-all
MCL
Gene
families
Family count
37 fungal species
10 1 2
14 18 2
7 1 1
6 1 12
6 1 8
3 1 1
+
A B C D E
CAFE
Family 1 P < 0.001 Branch A
Family 2 P < 0.001 Branch B
Family 3 P=0.02 Branch C,E
Family 4 P=0.03 Branch D
18 U. maydis
5 C. gattii R265
5 gattii
5 C. gattii WM276
23 Basidiomycota
5 Cryptococcus
5 C. neoformans JEC21
5 neoformans
5 C. neoformans var grubii
23 Hymenomycota
163 P. chrysosporium
136 Homobasidiomycota
141 C.cinereus
400 300 200 100 0

Families with significant expansions

Transporters
Kinases
P450
Oxidation

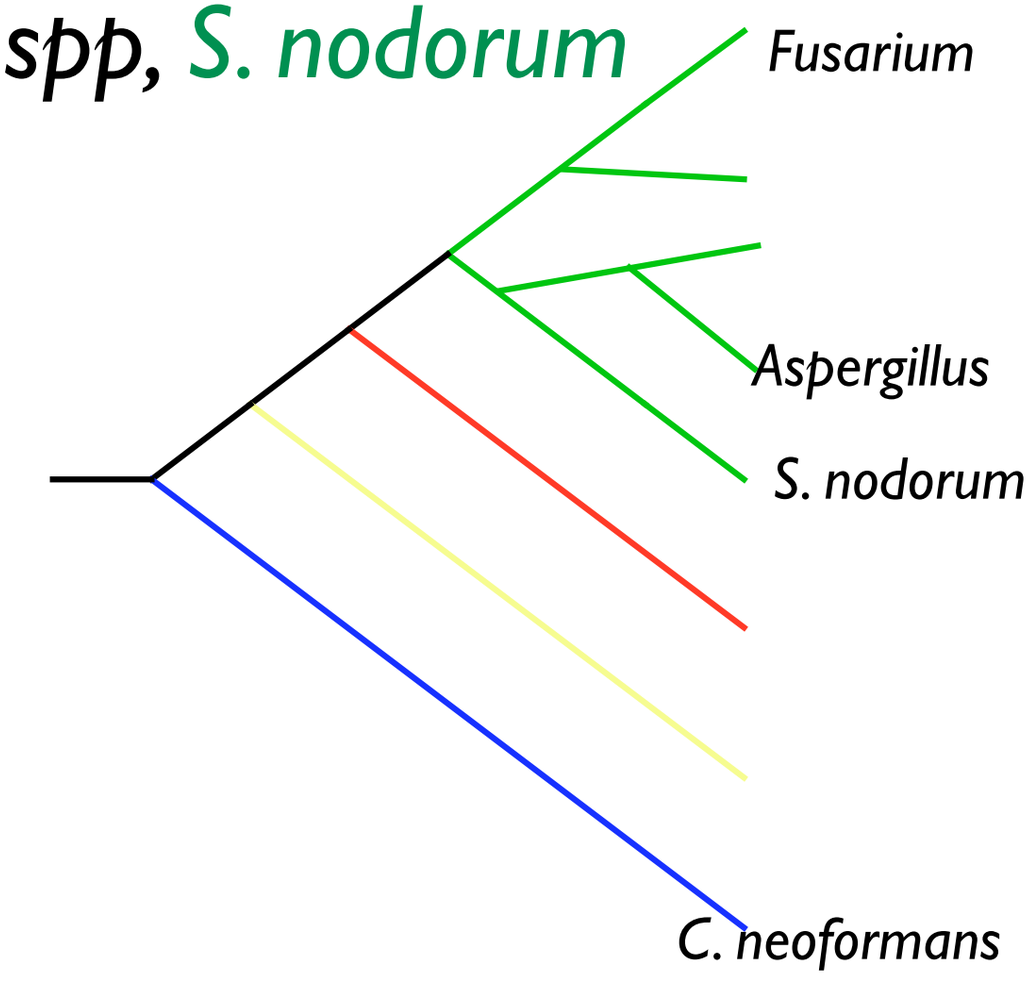
Vitamin & Cofactor transport	Methytransferase
Lactose & sugar transport	Cytochrome P450: CYP64
Amine transport	Cytochrome P450: CYP53,57A
Myo-instol, quinate, and glucose transport	Cytochrome P450
Oligopeptide transport	Kinase
ABC transporter	Subtilase family
MFS, drug pump, & sugar transport	NADH flavin oxidoreductase
Transport	Aldehyde dehydrogenase
Monocarboxylate & sugar transport	Aldo/kedo reductase
ABC transport	Multicopper oxidase
Amino acid permease	AMP-binding enzyme

Transporters

- Of 45 significant families, 22 were related to transport
- Vitamin and amino acid transport
- Sugar and sugar-like transporters
- Multidrug and efflux pumps
- ABC transporters (ATP Binding Cassette)

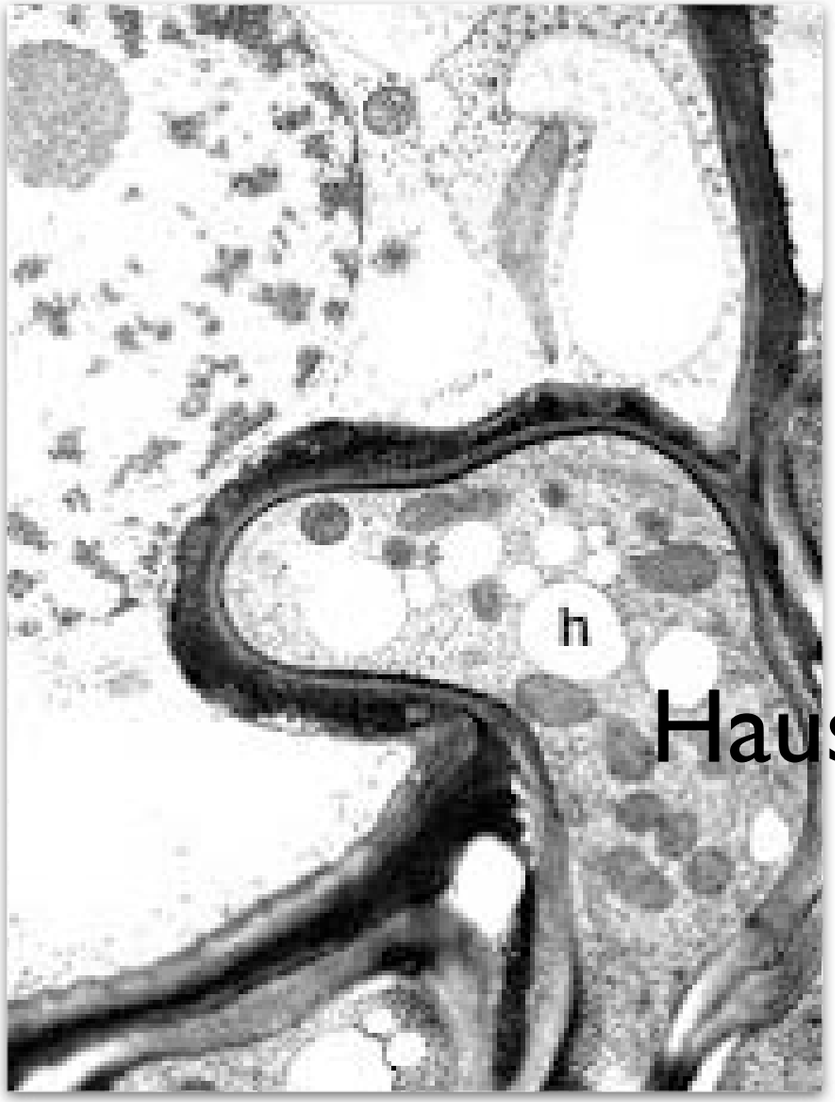
Branches with transporter expansions

- Sugar related, Drug pump, and Major Facilitator Superfamily
  - *Aspergillus spp*, *Fusarium spp*, *S. nodorum*
  - Euascomycota
- Vitamin transport
  - *C. neoformans*, *Fusarium*
  - *A. nidulans* (Biotin)



Sugar transporter use in phytopathogens

- Sugar transporters are used to extract nutrients from host
  - Haustorium: specialized structure for plant parasitism
  - Many sugar transporters highly and specifically expressed in haustoria

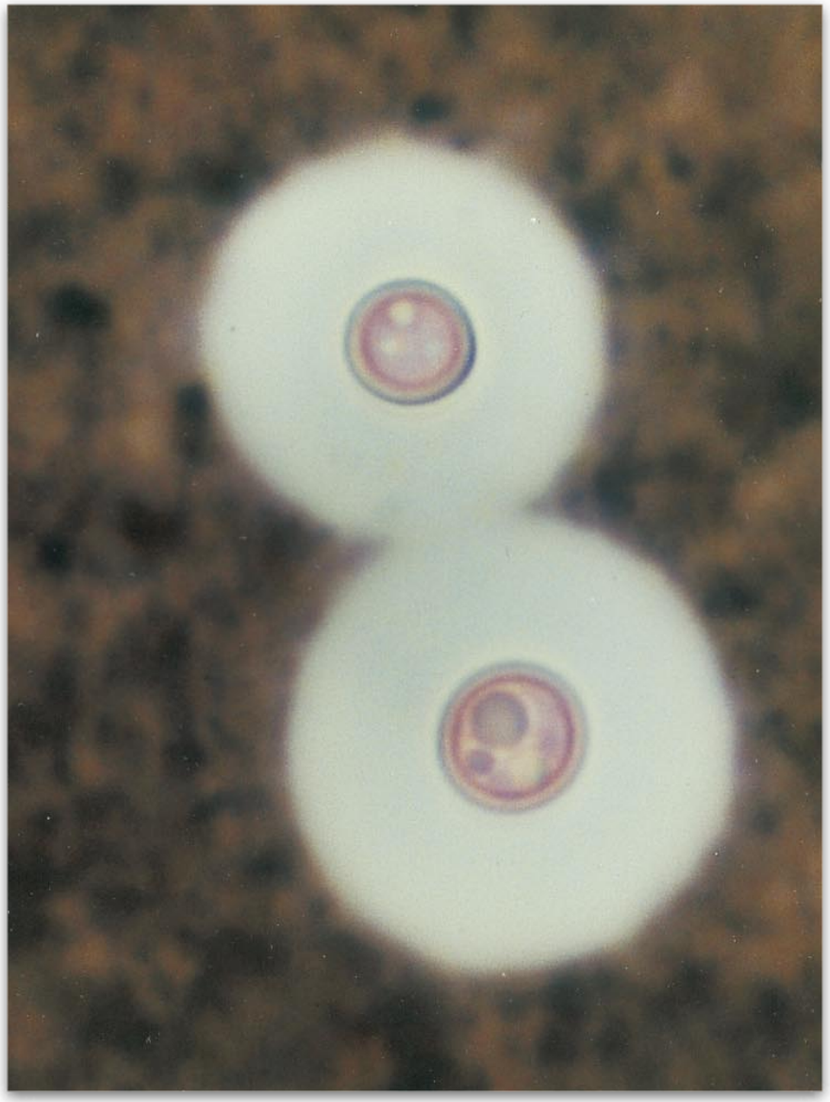


Haustorium

Robert Bauer http://tolweb.org/

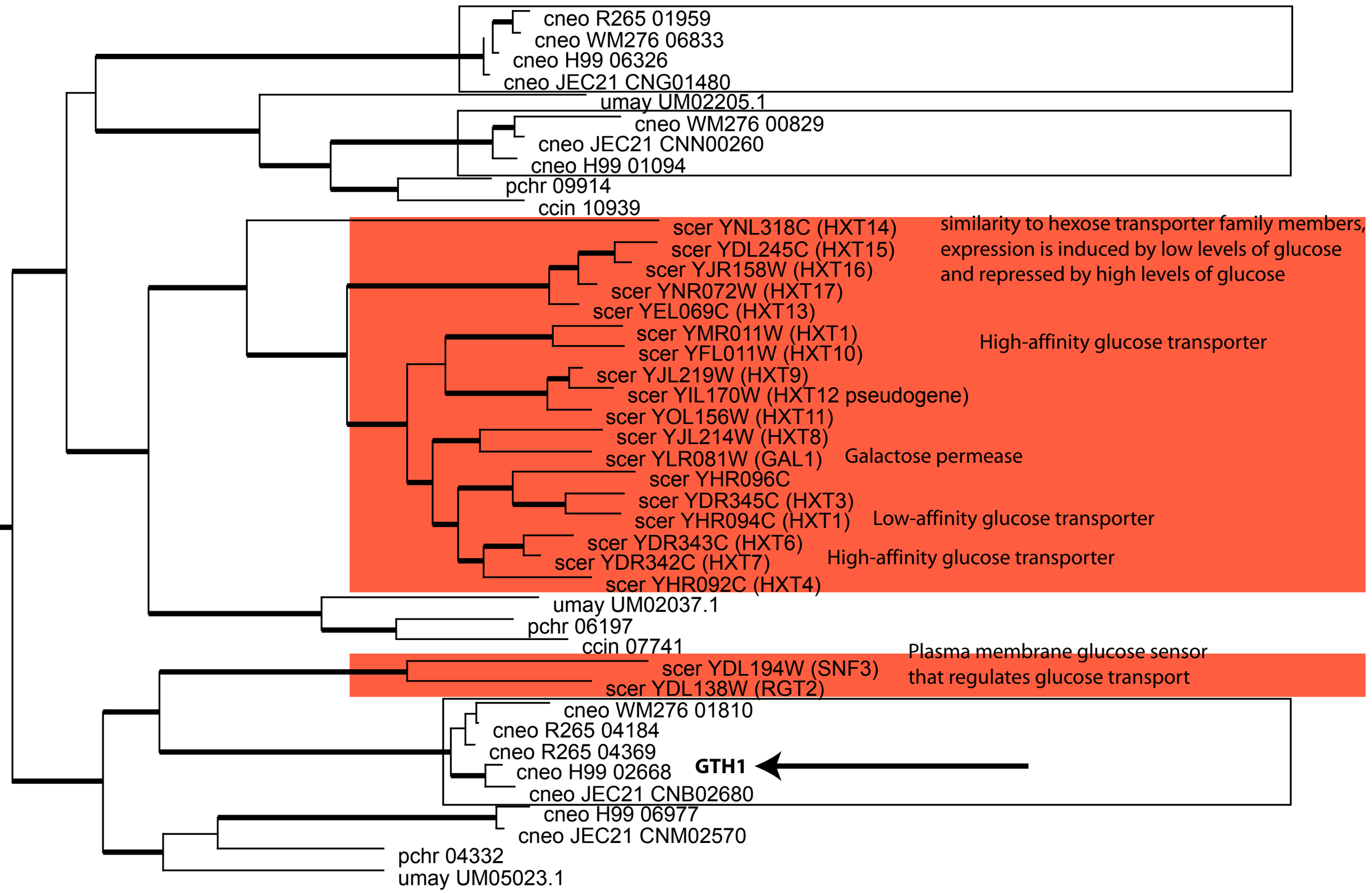
Cryptococcus sugar transporters

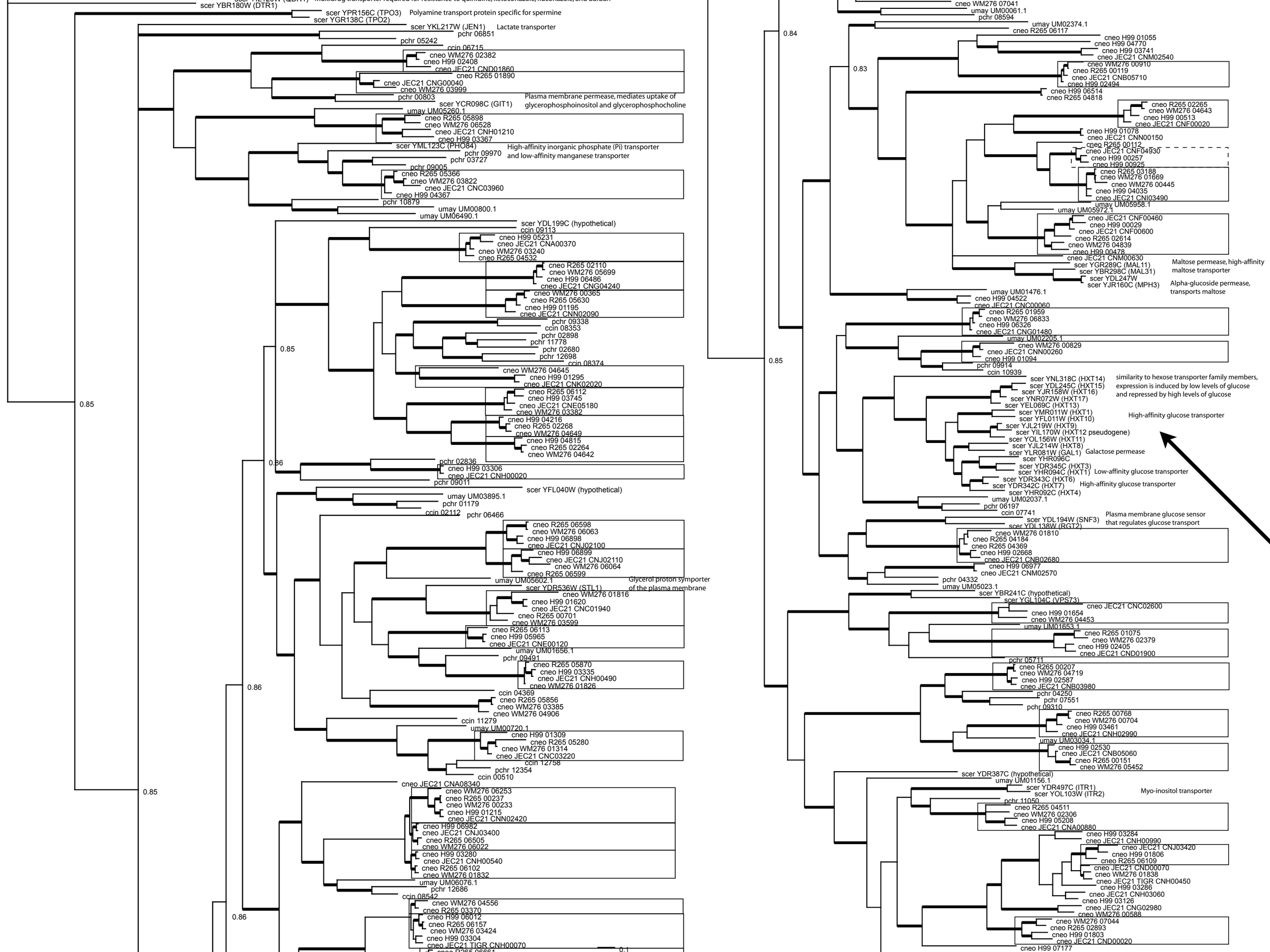
- 3x as many sugar transporters in *C. neoformans* (~50) than other basidiomycetes
- "sugar coated killer"
- Capsule is a mixture of glucose, xylose, and mannose.
- Transporters could be important in capsule synthesis



Zerpa et al, 1996

Analysis of sugar transporter sub-family


cneo R265 01959
cneo WM276 06833
cneo H99 06326
cneo JEC21 CNG01480
umay UM02205.1
cneo WM276 00829
cneo JEC21 CNN00260
cneo H99 01094
pchr 09914
ccin 10939
scer YNL318C (HXT14)
scer YDL245C (HXT15)
scer YJR158W (HXT16)
scer YNR072W (HXT17)
scer YEL069C (HXT13)
scer YMR011W (HXT1)
scer YFL011W (HXT10)
scer YJL219W (HXT9)
scer YIL170W (HXT12 pseudogene)
scer YOL156W (HXT11)
scer YJL214W (HXT8)
scer YLR081W (GAL1)
scer YHR096C
scer YDR345C (HXT3)
scer YHR094C (HXT1)
scer YDR343C (HXT6)
scer YDR342C (HXT7)
scer YHR092C (HXT4)
similarity to hexose transporter family members,
expression is induced by low levels of glucose
and repressed by high levels of glucose
High-affinity glucose transporter
Galactose permease
Low-affinity glucose transporter
High-affinity glucose transporter
umay UM02037.1
pchr 06197
ccin 07741
scer YDL194W (SNF3)
scer YDL138W (RGT2)
Plasma membrane glucose sensor
that regulates glucose transport
cneo WM276 01810
cneo R265 04184
cneo R265 04369
cneo H99 02668
GTH1
cneo JEC21 CNB02680
cneo H99 06977
cneo JEC21 CNM02570
pchr 04332
umay UM05023.1


scer YBR180W (DTR1)
scer YPR156C (TPO3)
scer YGR138C (TPO2)
Polyamine transport protein specific for spermine
scer YKL217W (JEN1)
Lactate transporter
pchr 06851
pchr 05242
ccin 06715
cneo WM276 02382
cneo H99 02408
cneo JEC21 CND01860
cneo R265 01890
cneo JEC21 CNG00040
cneo WM276 03999
pchr 00803
scer YCR098C (GIT1)
Plasma membrane permease, mediates uptake of glycerophosphoinositol and glycerophosphocholine
umay UM05260.1
cneo R265 05898
cneo WM276 06528
cneo JEC21 CNH01210
cneo H99 03367
scer YML123C (PHO84)
High-affinity inorganic phosphate (Pi) transporter and low-affinity manganese transporter
pchr 09970
pchr 03727
pchr 09005
cneo R265 05366
cneo WM276 03822
cneo JEC21 CNC03960
cneo H99 04367
pchr 10879
umay UM00800.1
umay UM06490.1
scer YDL199C (hypothetical)
ccin 09113
cneo H99 05231
cneo JEC21 CNA00370
cneo WM276 03240
cneo R265 04532
cneo R265 02110
cneo WM276 05699
cneo H99 06486
cneo JEC21 CNG04240
cneo WM276 00365
cneo R265 05630
cneo H99 01195
cneo JEC21 CNN02090
pchr 09338
ccin 08353
pchr 02898
pchr 11778
pchr 02680
pchr 12698
ccin 08374
cneo WM276 04645
cneo H99 01295
cneo JEC21 CNK02020
cneo R265 06112
cneo H99 03745
cneo JEC21 CNE05180
cneo WM276 03382
cneo H99 04216
cneo R265 02268
cneo WM276 04649
cneo H99 04815
cneo R265 02264
cneo WM276 04642
pchr 02836
cneo H99 03306
cneo JEC21 CNH00020
pchr 09011
scer YFL040W (hypothetical)
umay UM03895.1
pchr 01179
ccin 02112
pchr 06466
cneo R265 06598
cneo WM276 06063
cneo H99 06898
cneo JEC21 CNJ02100
cneo H99 06899
cneo JEC21 CNJ02110
cneo WM276 06064
cneo R265 06599
umay UM05602.1
Glycerol proton symporter of the plasma membrane
scer YDR536W (STL1)
cneo WM276 01816
cneo H99 01620
cneo JEC21 CNC01940
cneo R265 00701
cneo WM276 03599
cneo R265 06113
cneo H99 05965
cneo JEC21 CNE00120
umay UM01656.1
pchr 09491
cneo R265 05870
cneo H99 03335
cneo JEC21 CNH00490
cneo WM276 01826
ccin 04369
cneo R265 05856
cneo WM276 03385
cneo WM276 04906
ccin 11279
umay UM00720.1
cneo H99 01309
cneo R265 05280
cneo WM276 01314
cneo JEC21 CNC03220
ccin 12758
pchr 12354
ccin 00510
cneo JEC21 CNA08340
cneo WM276 06253
cneo R265 00237
cneo WM276 00233
cneo H99 01215
cneo JEC21 CNN02420
cneo H99 06982
cneo JEC21 CNJ03400
cneo R265 06505
cneo WM276 06022
cneo H99 03280
cneo JEC21 CNH00540
cneo R265 06102
cneo WM276 01832
umay UM06076.1
pchr 12686
ccin 08542
cneo WM276 04556
cneo R265 03370
cneo H99 06012
cneo R265 06157
cneo WM276 03424
cneo H99 03304
cneo JEC21 TIGR CNH00070
0.1
0.85
0.86
0.86
0.86
0.85
0.85
0.84
0.83
0.85
cneo WM276 07041
umay UM00061.1
pchr 08594
umay UM02374.1
cneo R265 06117
cneo H99 01055
cneo H99 04770
cneo H99 03741
cneo JEC21 CNM02540
cneo WM276 00910
cneo R265 00119
cneo JEC21 CNB05710
cneo H99 02494
cneo H99 06514
cneo R265 04818
cneo R265 02265
cneo WM276 04643
cneo H99 00513
cneo JEC21 CNF00020
cneo H99 01078
cneo JEC21 CNN00150
cneo R265 00112
cneo JEC21 CNF04930
cneo H99 00257
cneo H99 00925
cneo R265 03188
cneo WM276 01669
cneo WM276 00445
cneo H99 04035
cneo JEC21 CNI03490
umay UM05958.1
umay UM05972.1
cneo JEC21 CNF00460
cneo H99 00029
cneo JEC21 CNF00600
cneo R265 02614
cneo WM276 04839
cneo H99 00478
cneo JEC21 CNM00630
scer YGR289C (MAL11)
scer YBR298C (MAL31)
scer YDL247W
scer YJR160C (MPH3)
Maltose permease, high-affinity maltose transporter
Alpha-glucoside permease, transports maltose
umay UM01476.1
cneo H99 04522
cneo JEC21 CNC00060
cneo R265 01959
cneo WM276 06833
cneo H99 06326
cneo JEC21 CNG01480
umay UM02205.1
cneo WM276 00829
cneo JEC21 CNN00260
cneo H99 01094
pchr 09914
ccin 10939
scer YNL318C (HXT14)
scer YDL245C (HXT15)
scer YJR158W (HXT16)
scer YNR072W (HXT17)
scer YEL069C (HXT13)
similarity to hexose transporter family members, expression is induced by low levels of glucose and repressed by high levels of glucose
scer YMR011W (HXT1)
scer YFL011W (HXT10)
High-affinity glucose transporter
scer YJL219W (HXT9)
scer YIL170W (HXT12 pseudogene)
scer YOL156W (HXT11)
scer YJL214W (HXT8)
scer YLR081W (GAL1)
Galactose permease
scer YHR096C
scer YDR345C (HXT3)
scer YHR094C (HXT1)
Low-affinity glucose transporter
scer YDR343C (HXT6)
scer YDR342C (HXT7)
High-affinity glucose transporter
scer YHR092C (HXT4)
umay UM02037.1
pchr 06197
ccin 07741
scer YDL194W (SNF3)
scer YDL138W (RGT2)
Plasma membrane glucose sensor that regulates glucose transport
cneo WM276 01810
cneo R265 04184
cneo R265 04369
cneo H99 02668
cneo JEC21 CNB02680
cneo H99 06977
cneo JEC21 CNM02570
pchr 04332
umay UM05023.1
scer YBR241C (hypothetical)
scer YGL104C (VPS73)
cneo JEC21 CNC02600
cneo H99 01654
cneo WM276 04453
umay UM01653.1
cneo R265 01075
cneo WM276 02379
cneo H99 02405
cneo JEC21 CND01900
pchr 05711
cneo R265 00207
cneo WM276 04719
cneo H99 02587
cneo JEC21 CNB03980
pchr 04250
pchr 07551
pchr 09310
cneo R265 00768
cneo WM276 00704
cneo H99 03461
cneo JEC21 CNH02990
umay UM03034.1
cneo H99 02530
cneo JEC21 CNB05060
cneo R265 00151
cneo WM276 05452
scer YDR387C (hypothetical)
umay UM01156.1
scer YDR497C (ITR1)
scer YOL103W (ITR2)
Myo-inositol transporter
pchr 11050
cneo R265 04511
cneo WM276 02306
cneo H99 05208
cneo JEC21 CNA00880
cneo H99 03284
cneo JEC21 CNH00990
cneo JEC21 CNJ03420
cneo H99 01806
cneo R265 06109
cneo JEC21 CND00070
cneo WM276 01838
cneo JEC21 TIGR CNH00450
cneo H99 03286
cneo JEC21 CNH03060
cneo H99 03126
cneo JEC21 CNG02980
cneo WM276 00588
cneo WM276 07044
cneo R265 02893
cneo H99 01803
cneo JEC21 CND00020
cneo H99 07177

GTH1 is a Glucose Sensor

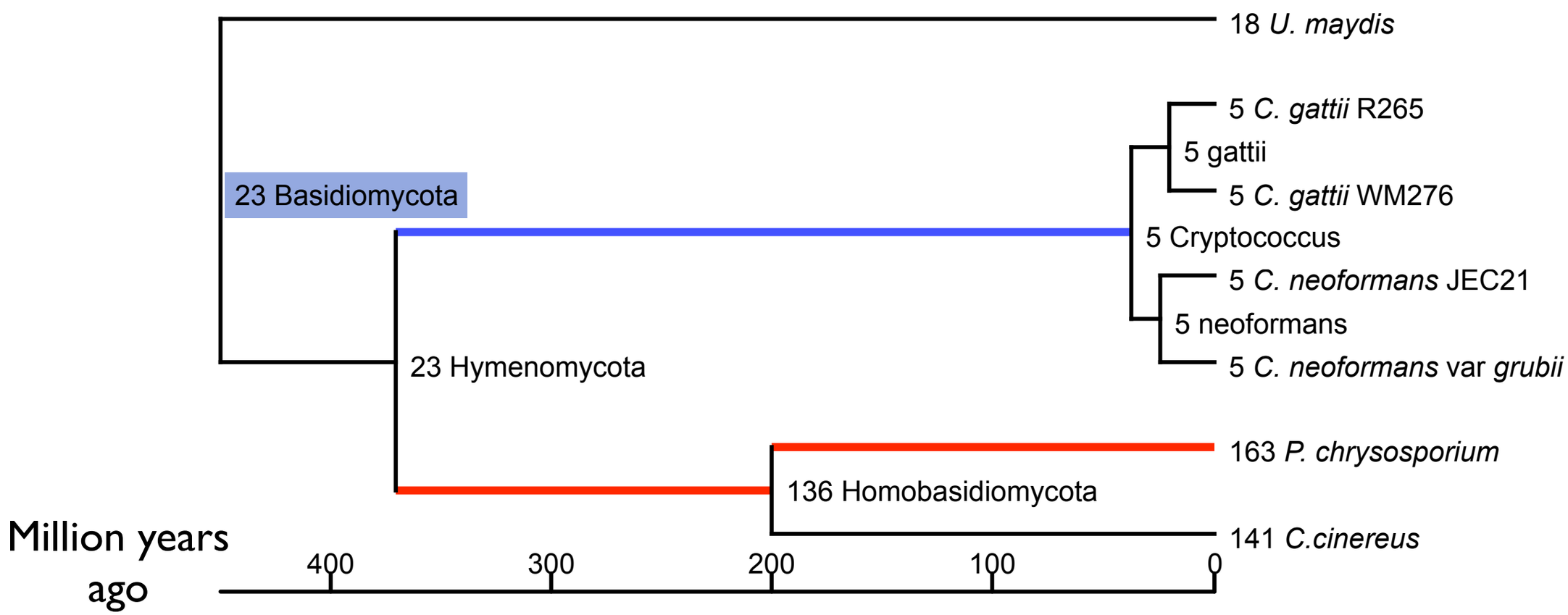
- Melanin production pathway is starvation induced
- GTH1 KO is hypermelanized (always glucose starved)
- GTH1 overexpression is hypomelanized

Alspaugh et al 1999

P450 CYP64

P450 enzymes involved in synthesis and cleavage of chemical bonds. Drug metabolism in animals.

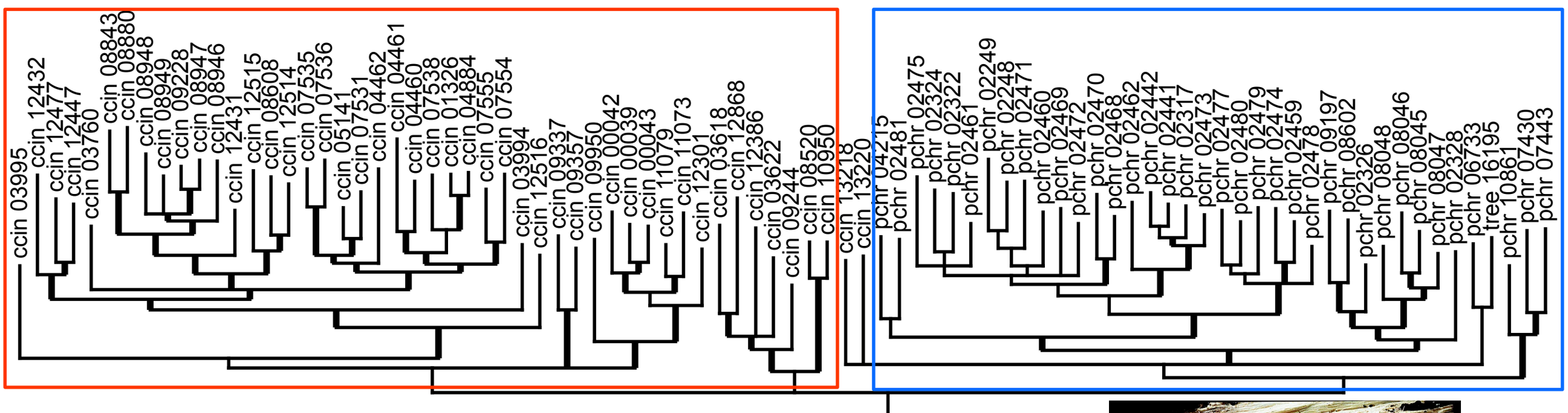
CYP64: Step in *Aspergillus* spp aflatoxin pathway
P. chrysosporium implicated in lignin and hydrocarbon degradation.



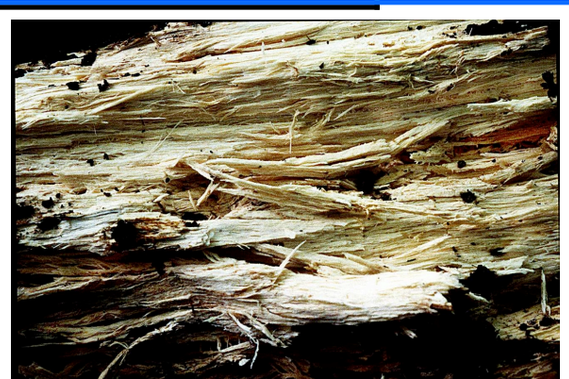
CYP64 was from independent duplication

C. cinereus expansion

P. chrysosporium expansion

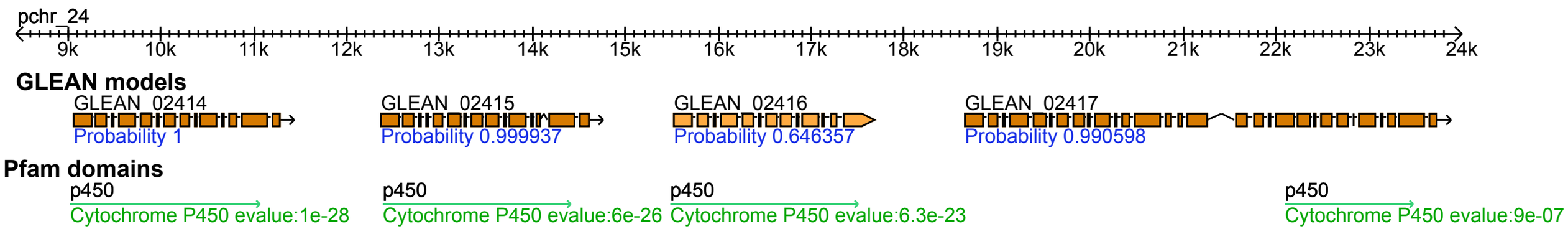


Mario Cervini



Tom Volk

Local duplications created CYP64 expansion



http://fungal.genome.duke.edu

Family size contractions

- *Histoplasma*, *Coccidioides* many families
- Hemiascomycetes - P450
- *C. neoformans* - P450
- *U. maydis* - Lactose transport

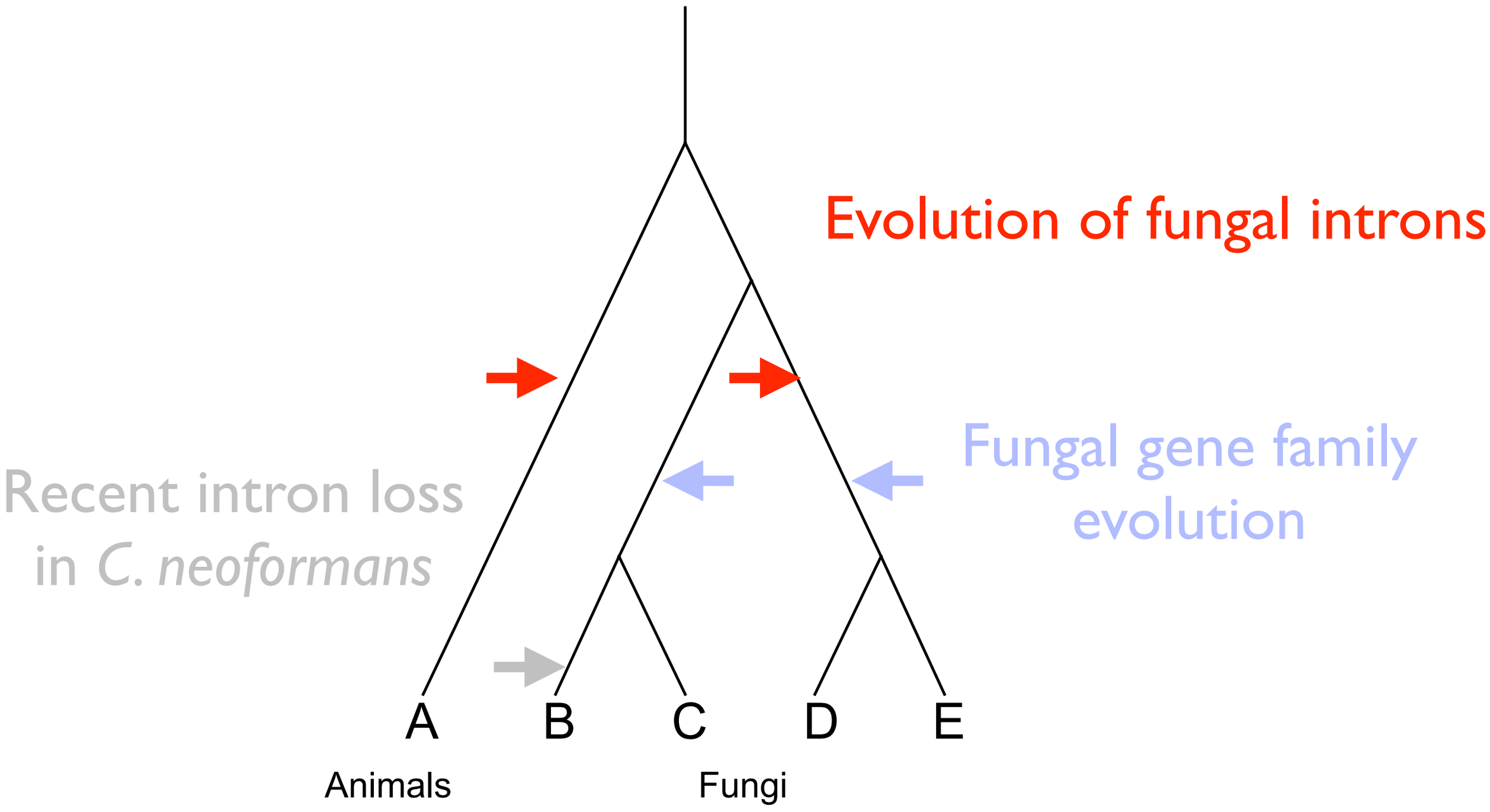
Basidiomycetes

fam	U.maydis	C.cinereus	P.chrysosporium	H99	JEC21	R265	WM276	Cryptococcus & GO Notes	Protein domains
0	1	420	68	140	4	14	70	transposon	Integrase core domain(57);
1	74	108	119	78	81	84	82	kinases. homobasidio	Protein kinase domain(306);
2	44	67	87	48	50	49	52	G-protein coupled receptor protein	WD domain, G-beta repeat(193)
3	42	113	53	27	4	23	106	transposon	Reverse transcriptase
4	18	141	163	5	5	5	5	metal ion binding; cation binding	Cytochrome P450 (39)
5	32	26	76	44	47	46	45	drug transport	Major Facilitator Superfamily(157);
6	17	36	44	47	55	40	39	"hexose transporter	dicarboxylate symporter
7	19	20	23	59	57	47	50	sugar transporter related	Sugar (and other) transporter(167);
8	30	33	64	32	33	31	35	"siderophore biosynthesis?	- dihydroxybenzoat
9	35	39	35	33	40	34	36	transmembrane receptor activity;	Ras family(130); ADP-ribosylation
10	32	44	45	29	27	31	30	localization hydrolase	ABC transporter (112); RecF/RecN/
11	5	125	61	11	11	11	13	hydrolyzing O-glycosyl	WSC domain (24); Glycosyl
18	4	2	5	46	43	22	30	oxidoreductase activity;RNA	Oxidoreductase family, NAD-
20	0	109	36	0	0	0	0	kinase activity; phosphotransferas	

Conclusions

- Sugar transporters are highly expanded in independent lineages
  - Saprophytic and phytopathogenic lifestyles
- P450 CYP64 independent expansions in Homobasidiomycetes
  - Lignin degradation and saprophytic lifestyles
- Family size contractions among lineages containing primary pathogens
  - Genome streamlining?

Fungal comparative genomics

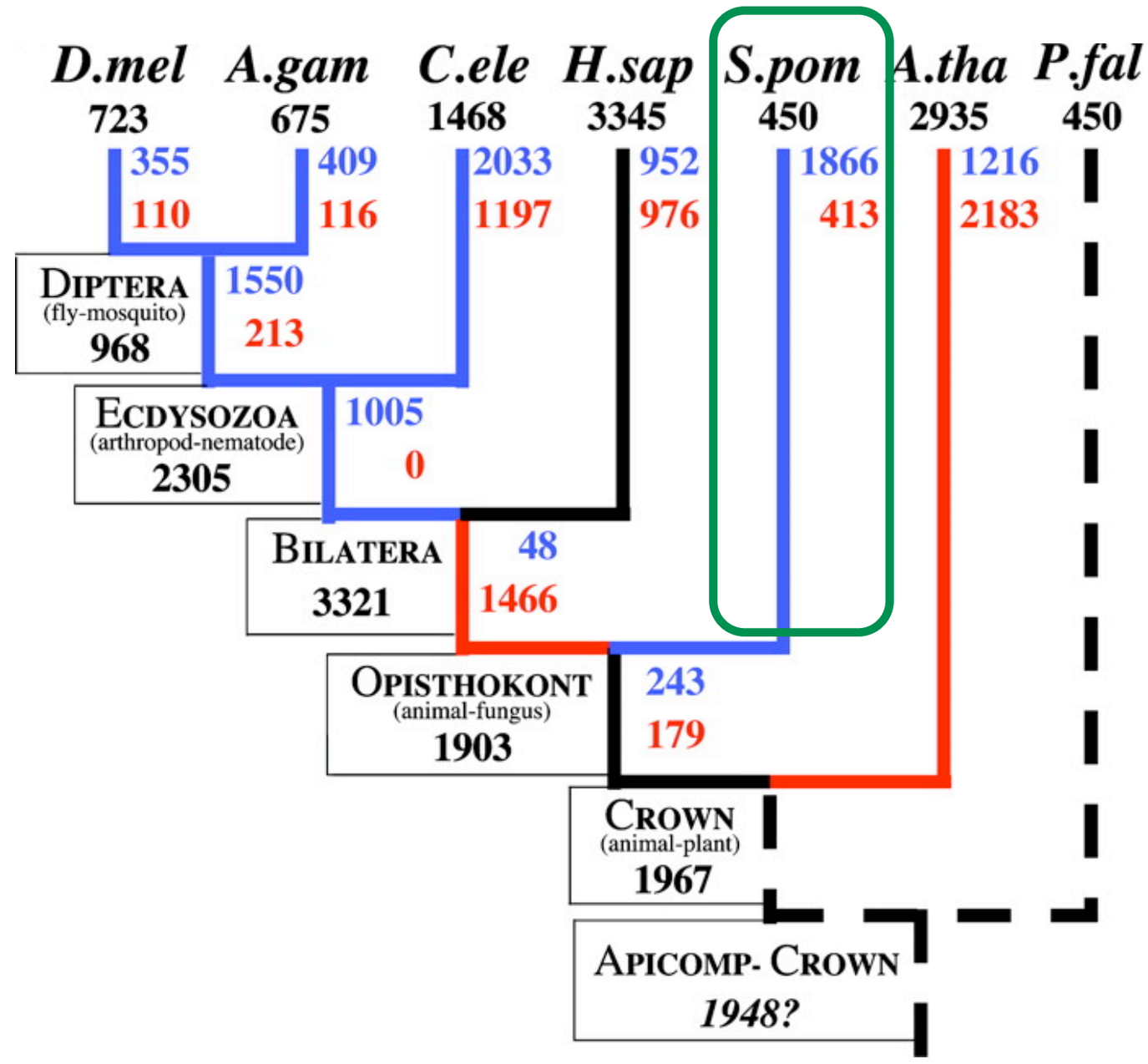


Evolution of gene structure

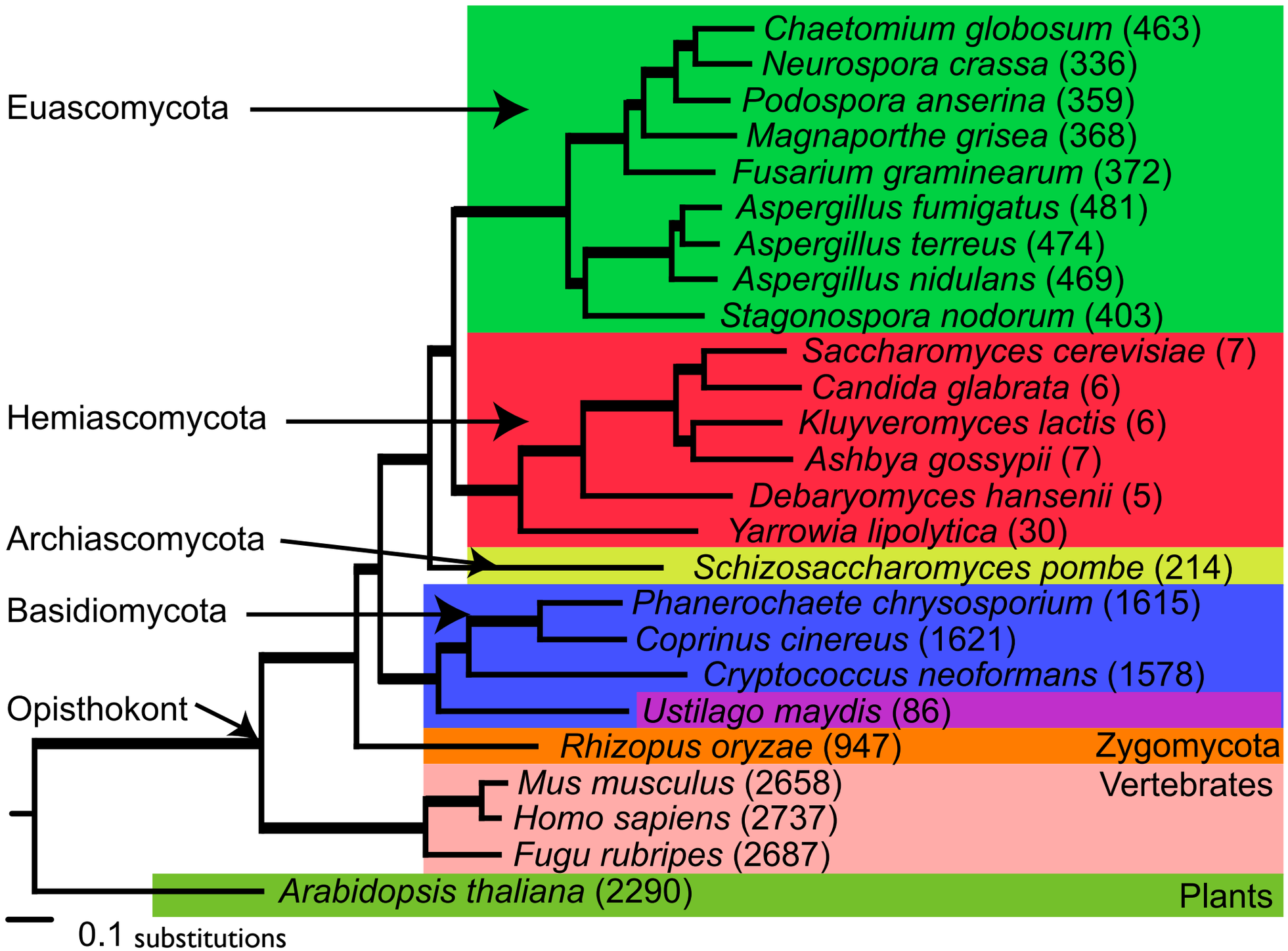
- Present day introns
  - Recent insertions?
    - Introns late hypothesis
  - Present in eukaryotic ancestor?
    - Introns early hypothesis / exon theory of genes
  - Mixture of two?

Previous work on intron evolution

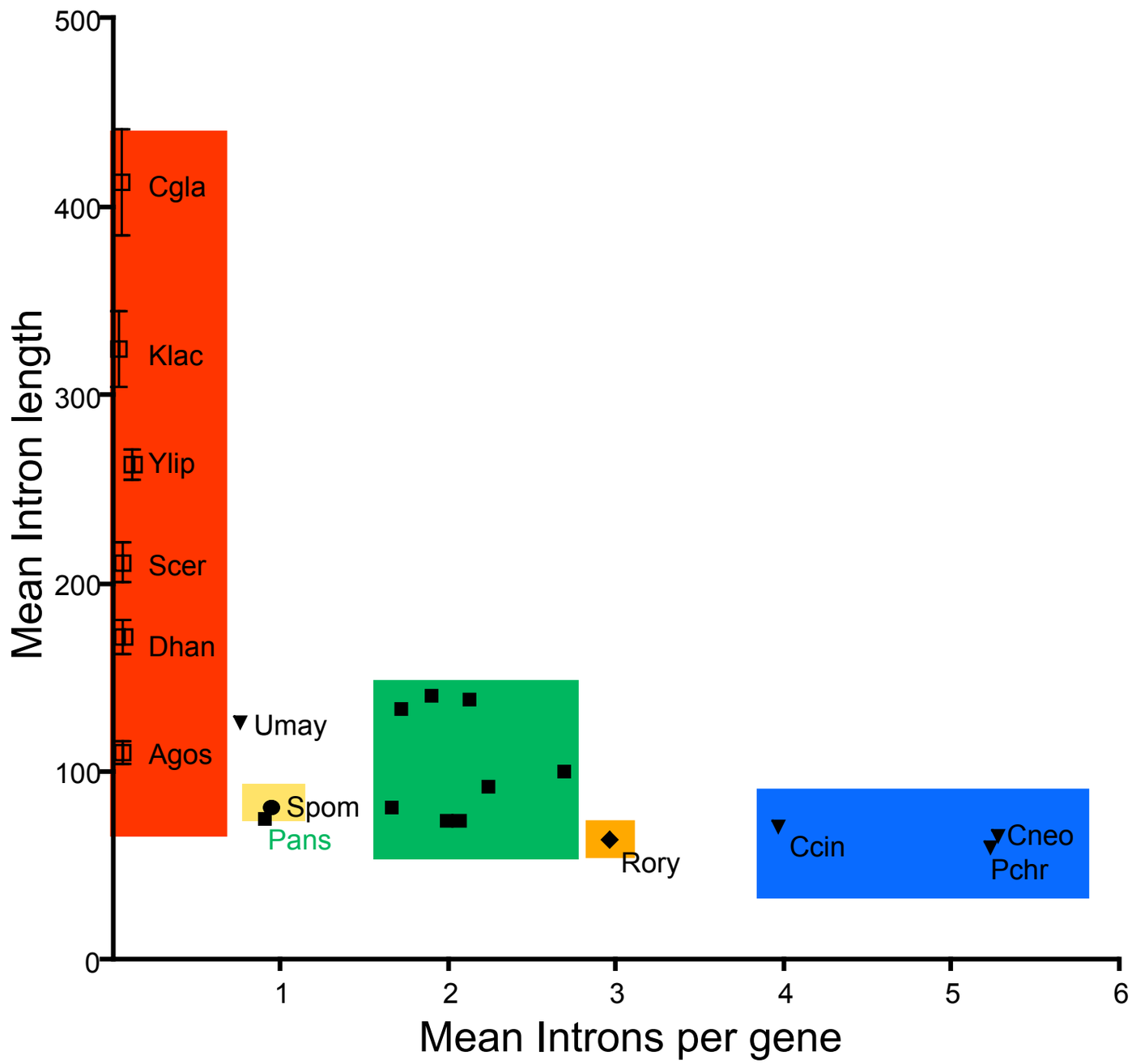
- Rogozin et al. 2003
  - 7 genomes
  - 684 genes, 7236 positions
  - Parsimony analysis
- Analysis methods
  - Roy and Gilbert. 2005
  - Csũrös. 2005
  - Nguyen et al. 2006



Calculating intron densities across a phylogeny



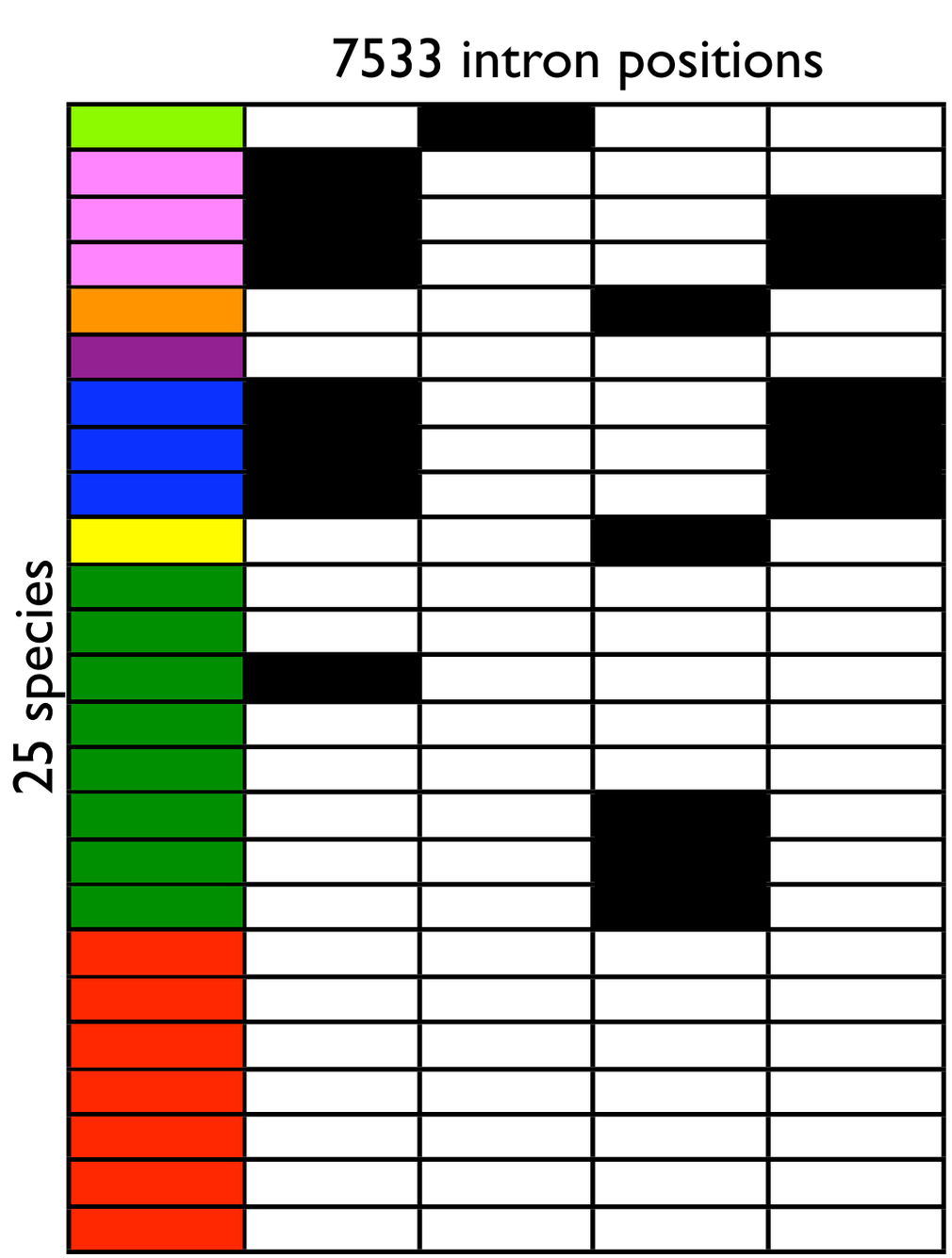
Intron frequency varies among the fungi

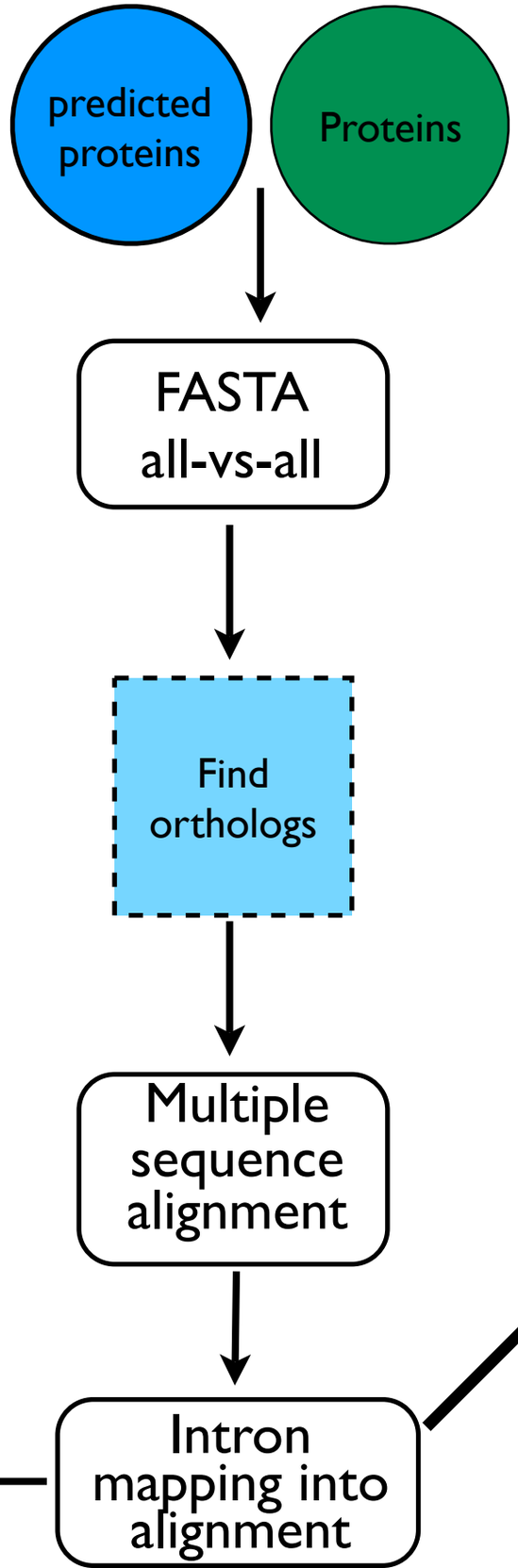


Analysis of whole genomes

- 25 entire genomes
  - 21 fungi, 3 vertebrates, 1 plant
- Largest dataset ever assembled for intron analysis
- 1160 orthologous genes
- 7533 intron positions
- 4.15 Mb conserved coding sequence per genome

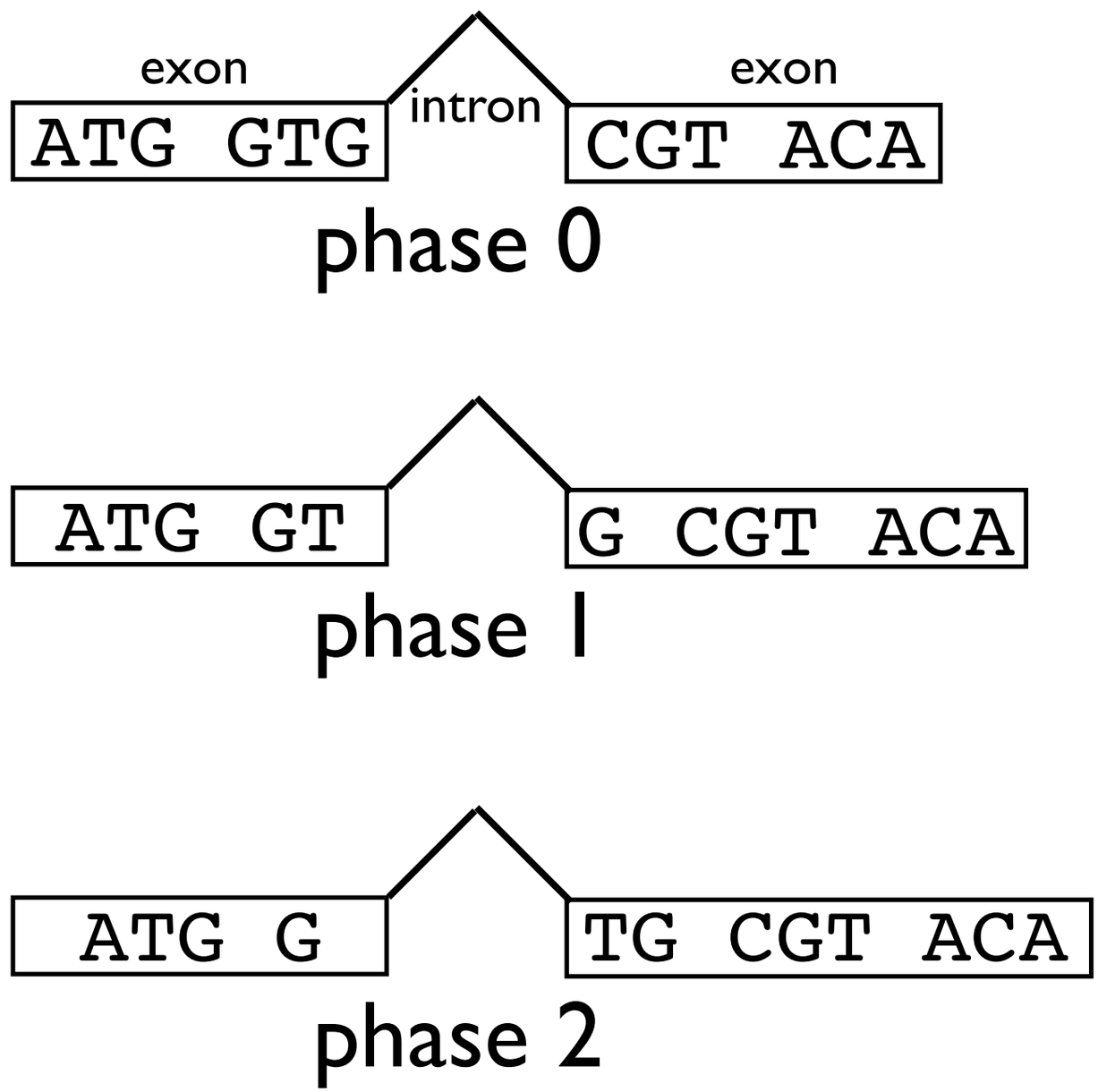
Analysis methods






```
A.thaliana      TWLPEQVA-FIQS1MG-NDKANS-YW
Fugu            QWTQEQVQ0SVQE-MG-NAKAKR-LY
Human           QWTQEQIQ0CMQE-MG-NGKANR-LY
Mouse           QWTQEQIQ0CMQE-MG-NGKANR-LY
R.oryzae        TWVQDQVE-NMVR-WG-NERANK2YW
U.maydis        TWTREQVD-RMKE-VG-NLKSNR-KY
C.neoformans    IWTPEQME0SIQK-WG-NKRANM-YW
C.cinereus      VWTPEQME0SIQK-WG-NRRANL-YW
P.chrysosporium VWTVEQMN0SIQK-WG-NKRANI-YW
S.pombe         SWTDEQTE-NMTR-WG-NERANL2YW
F.graminearum   SWTDEQLQ-SVLK-WG-NARANK-YW
N.crassa        AWTDEQLQ-SVLN-WG-NARANK-YW
C.globosum      SWTNEQVD0NMRK-VG-NVVSNK-LY
P.anserina      SWTDEQLQ-SVLN-WG-NARANK-YW
M.grisea        SWTDEQLQ-SVLS-WG-NARANK-YW
S.nodorum       TWTDEQLQ-SVLK-WG-NARANK-YW
A.nidulans      AWTDEQLQ-SVVR-WG-NARANK2YW
A.fumigatus     SWTDEQLQ-SVIK-WG-NARANK2YW
A.terreus       SWTDEQLQ-SVLR-WG-NARANK2YW
Y.lipolytica    AWTEEQLA-SMMK-WG-NTRCNM-FW
D.hansenii      AWTDEQVE-LMIK-WG-NEKCNI-YW
K.lactis        TWTEEHLE-AVLE-FGNNKKFNE-YY
A.gossypii      TWKEEHLQ-QVVR-FGNNQQANK-VF
C.glabrata      TWKEEHLK-ELVQ-MRNNVNANR-VY
S.cerevisiae    TWKEEHLV-KLIQ-FKNNLRANS-YY
                 *  ::     :       *    :
```

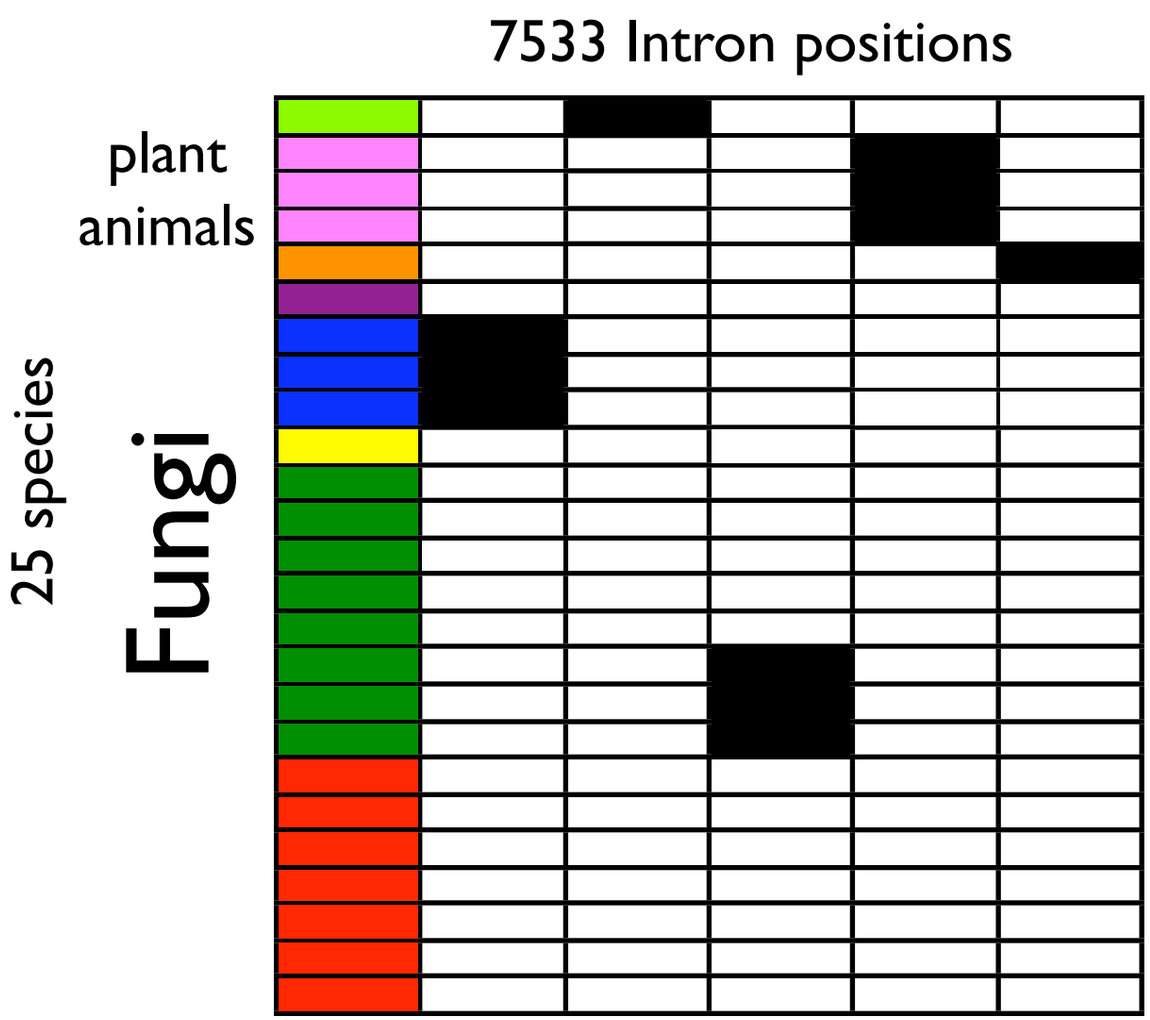
Intron phase

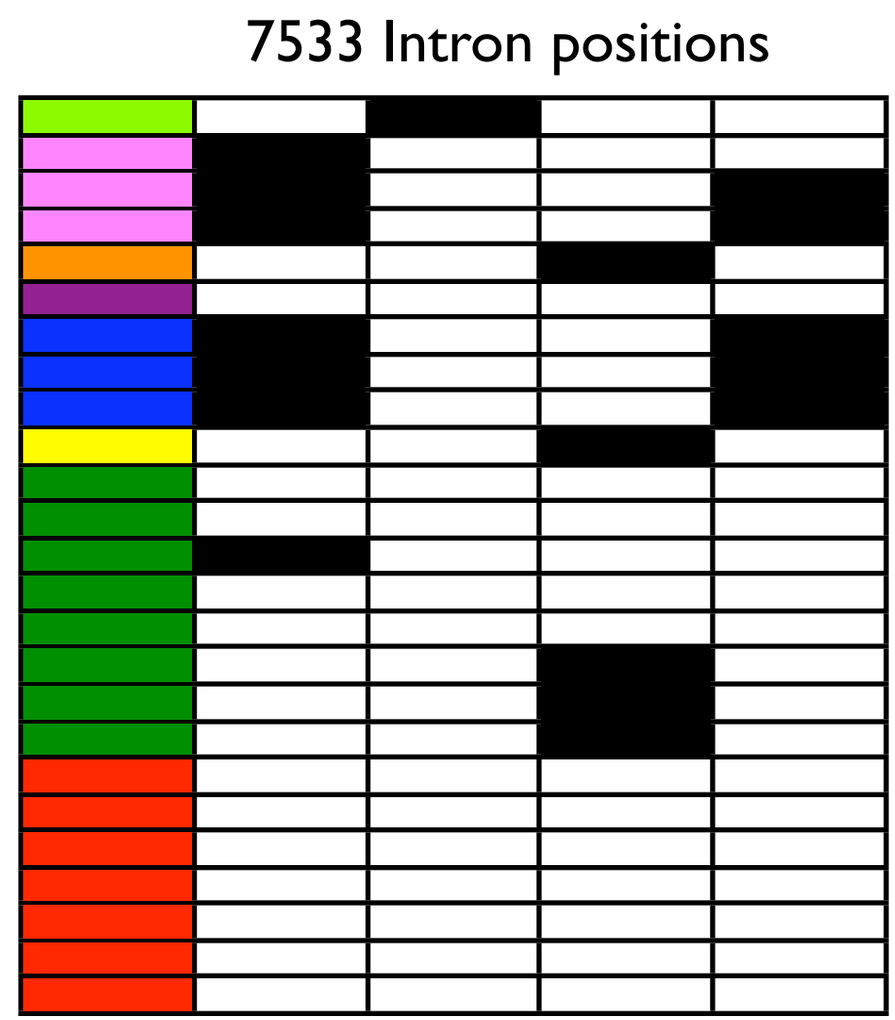


Conserved intron positions

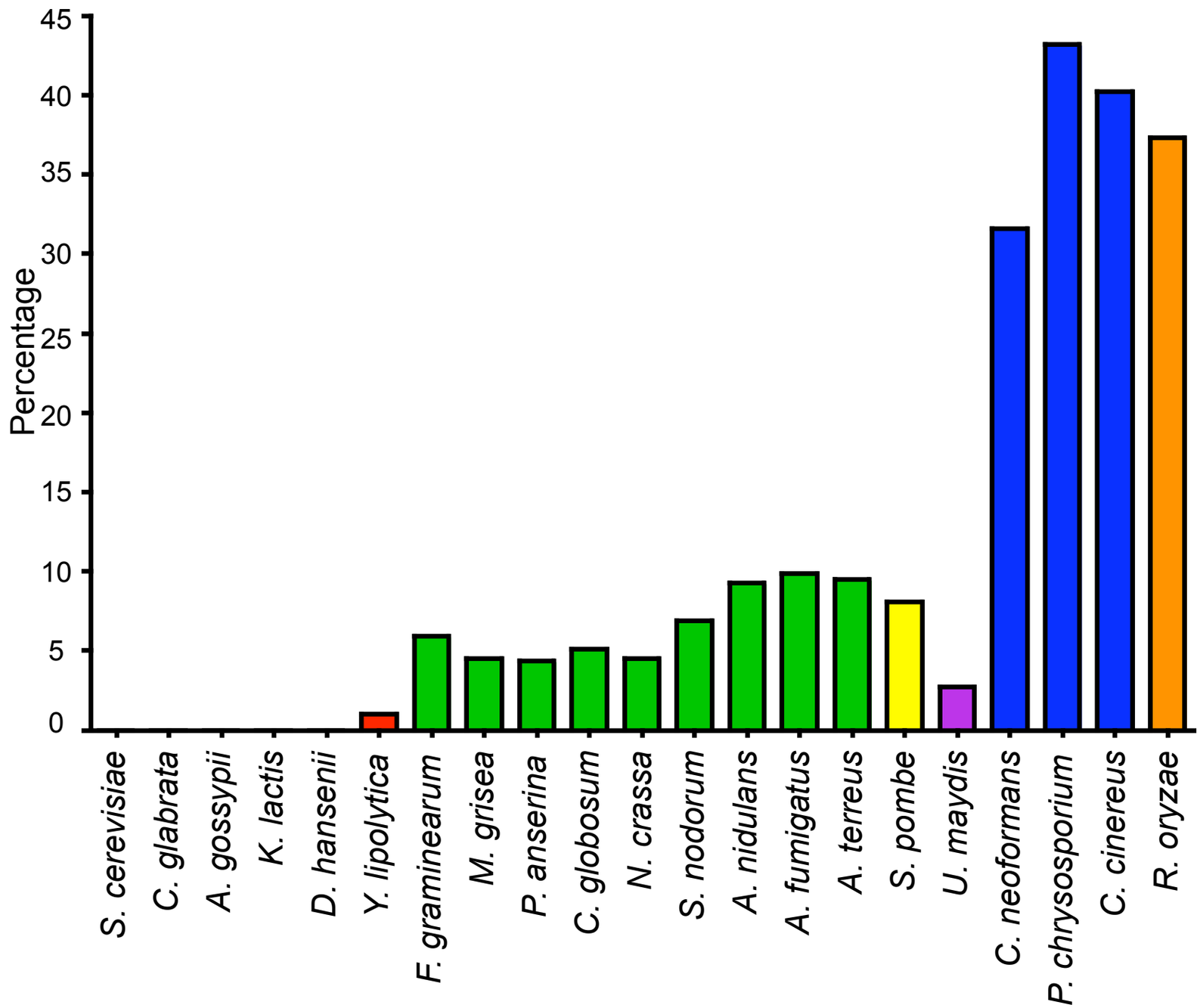
```
A.thaliana      TWLPEQVA-FIQS1MG-NDKANS-YWEA-----ELPP-----NYD---RV-GIENFIRAK2Y-------EEKRWV--
Fugu            QWTQEQVQ0SVQE-MG-NAKAKR-LYEA-----FLPK-----CFQRPETDQ-SAEIFIRDK-Y-------DKKKYMDK
Human           QWTQEQIQ0CMQE-MG-NGKANR-LYEA-----YLPE-----TFRRPQIDP2AVEGFIRDK-Y-------EKKKYMDR
Mouse           QWTQEQIQ0CMQE-MG-NGKANR-LYEA-----YLPE-----TFRRPQIDP2AVEGFIRDK-Y-------EKKKYMDR
R.oryzae        TWVQDQVE-NMVR-WG-NERANK2YWEA-----NL-------GDRKPS-ES-NMEMWIRAK-Y-------EQKRWA--
U.maydis        TWTREQVD-RMKE-VG-NLKSNR-KYNPDEMRNRPPT-----NMEESERDS-ELEKYIRRK-Y-------EFRRFV--
C.neoformans    IWTPEQME0SIQK-WG-NKRANM-YWER-----HLKA-----GHI-PS-DH2KIESFIRSK-Y-------ETRRWA--
C.cinereus      VWTPEQME0SIQK-WG-NRRANL-YWEA-----HLKP-----GHN-PP-EH2KMESFVRSK-Y-------ESRRWA--
P.chrysosporium VWTVEQMN0SIQK-WG-NKRANI-YWEA-----HLKA-----GHI-PP-DH2KMESFIRSK-Y-------ESKRWA--
S.pombe         SWTDEQTE-NMTR-WG-NERANL2YWEA-----KLAG-----GHV-PS-DS2KIATFIKTK-Y-------EFKKWV--
F.graminearum   SWTDEQLQ-SVLK-WG-NARANK-YWEA-----KLAA-----GHA-PS-EA-KIENFIRTK-Y-------ELKRWV--
N.crassa        AWTDEQLQ-SVLN-WG-NARANK-YWEA-----KLAQ-----GHV-PS-ES-KIENFIRTK-Y-------ELKRWV--
C.globosum      SWTNEQVD0NMRK-VG-NVVSNK-LYNPDN---KNPPVPIDADEA----DG-AMERFIRQK-YIARTLSIGKRRPGGD
P.anserina      SWTDEQLQ-SVLN-WG-NARANK-YWEA-----KLAP-----GHV-PS-EA-KIENFIRTK-Y-------ELKRWV--
M.grisea        SWTDEQLQ-SVLS-WG-NARANK-YWES-----KLAA-----GHA-PS-EA-KIENFIRTK-Y-------ELKRWV--
S.nodorum       TWTDEQLQ-SVLK-WG-NARANK-YWEA-----KLAP-----GHV-PS-EA-KIENFIRTK-Y-------ESKRWT--
A.nidulans      AWTDEQLQ-SVVR-WG-NARANK2YWEA-----KLAP-----GHV-PP-EA2KIENFIRTK-Y-------ESKRWV--
A.fumigatus     SWTDEQLQ-SVIK-WG-NARANK2YWEA-----KLAP-----GHV-PS-EA2KIENFIRTK-Y-------ESKRWV--
A.terreus       SWTDEQLQ-SVLR-WG-NARANK2YWEA-----KLAP-----GHV-PS-EA2KIENFIRTK-Y-------ESKRWV--
Y.lipolytica    AWTEEQLA-SMMK-WG-NTRCNM-FWEA-----KLPK-----GHV-PD-DN-KIENFIRTK-Y-------DMKKWA--
D.hansenii      AWTDEQVE-LMIK-WG-NEKCNI-YWES-----KLPD-----GYV-PD-QL-KIDNFIRTK-Y-------DLKKWV--
K.lactis        TWTEEHLE-AVLE-FGNNKKFNE-YYEN-----KLGG-----GTYVPD-QS-KIGQFIRTK-Y-------ELKKWV--
A.gossypii      TWKEEHLQ-QVVR-FGNNQQANK-VFEG-----RLGG-----GSYVPD-QS-KMGQFIKTK-Y-------EVRKWY--
C.glabrata      TWKEEHLK-ELVQ-MRNNVNANR-VYEA-----KLPDSSKFNGKSLGNDIN-LLQEFIRQK-Y-------ERKRWM--
S.cerevisiae    TWKEEHLV-KLIQ-FKNNLRANS-YYEATL-ADELKQ-----RKI-TD-TS-SLQNFIKNK-Y-------EYKKWI--
                 *  ::    :      *   :   ::                                ::: * *         ::
```

Patterns of conservation





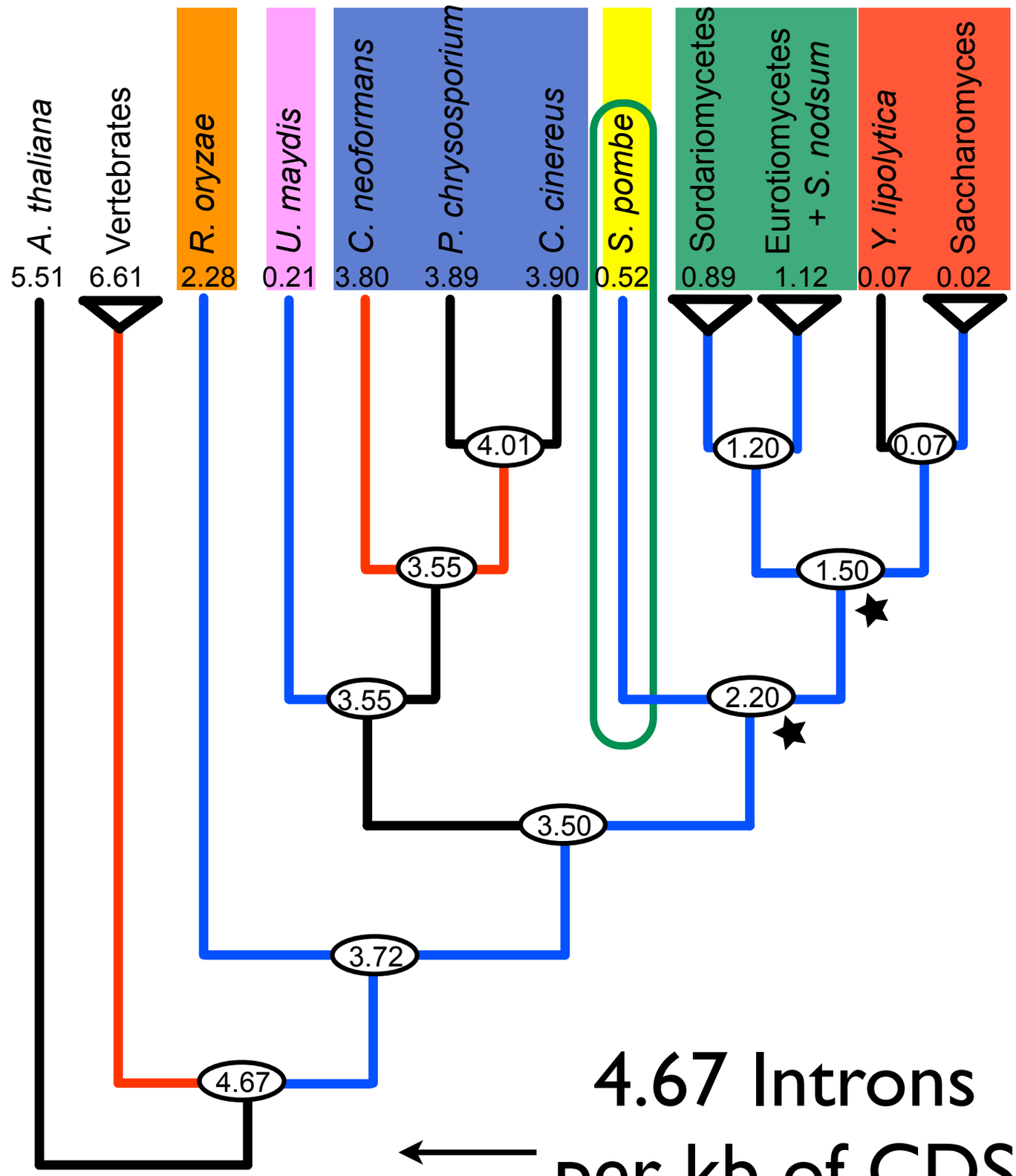
Intron positions shared with animals or plants



Intron position reconstruction

- 3 Methods
  - Roy and Gilbert. 2005
  - Csűrös. 2005
  - Nguyen et al. 2006
- Methods agree for all but 2 nodes in tree

Reconstruction of ancestral intron densities



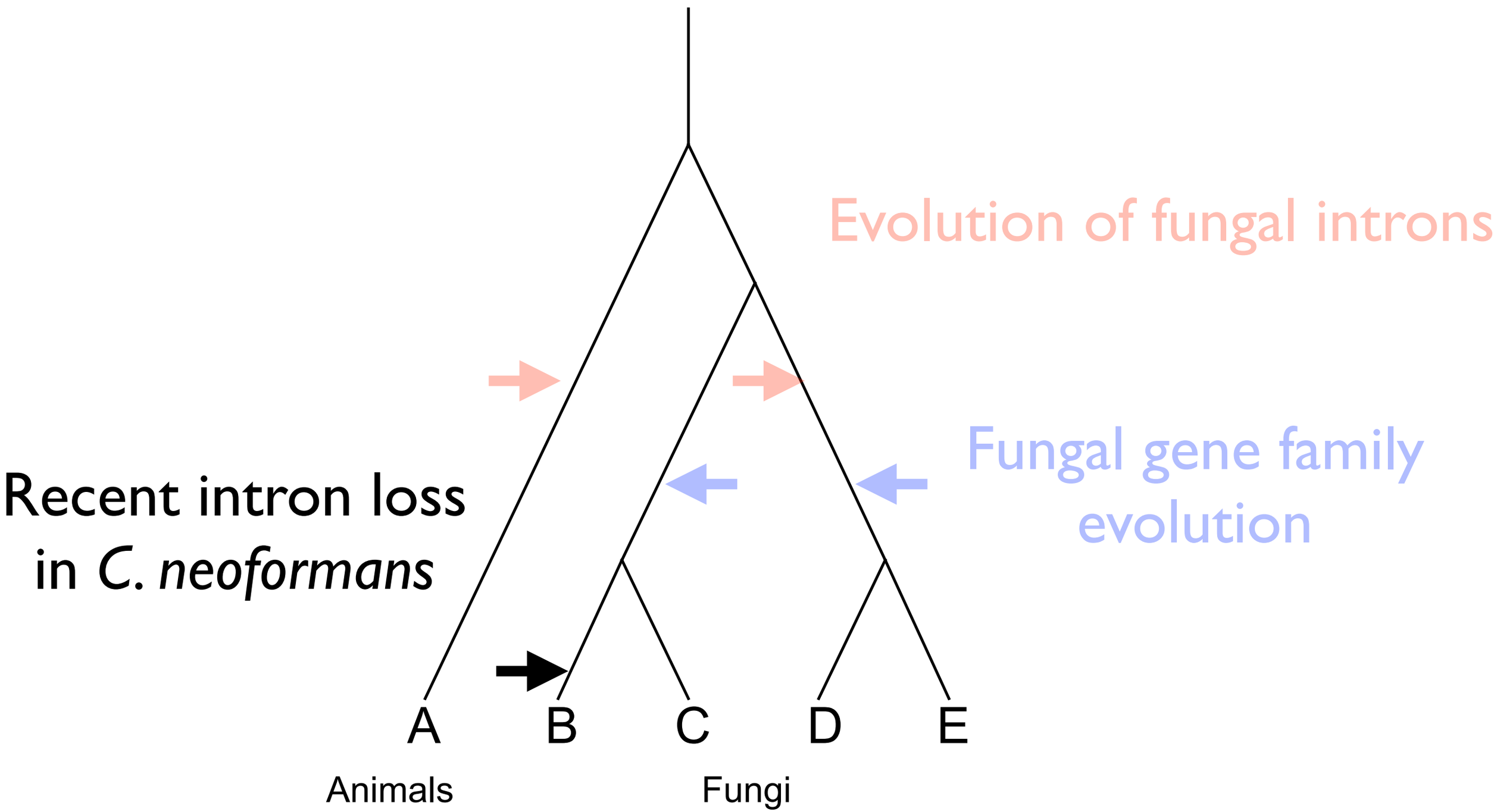
4.67 Introns per kb of CDS

Exon length of 214 bp

Conclusions

- Early eukaryotic crown genes were complex!
  - Ancestor had 70% of the introns in vertebrates - many more than previously reported
- Intron loss has dominated among the fungi
  - Hemiascomycota experienced loss
- No significant evidence for intron sliding or double insertions
- Sampling can bias interpretations - all fungi are not equal.

Fungal comparative genomics



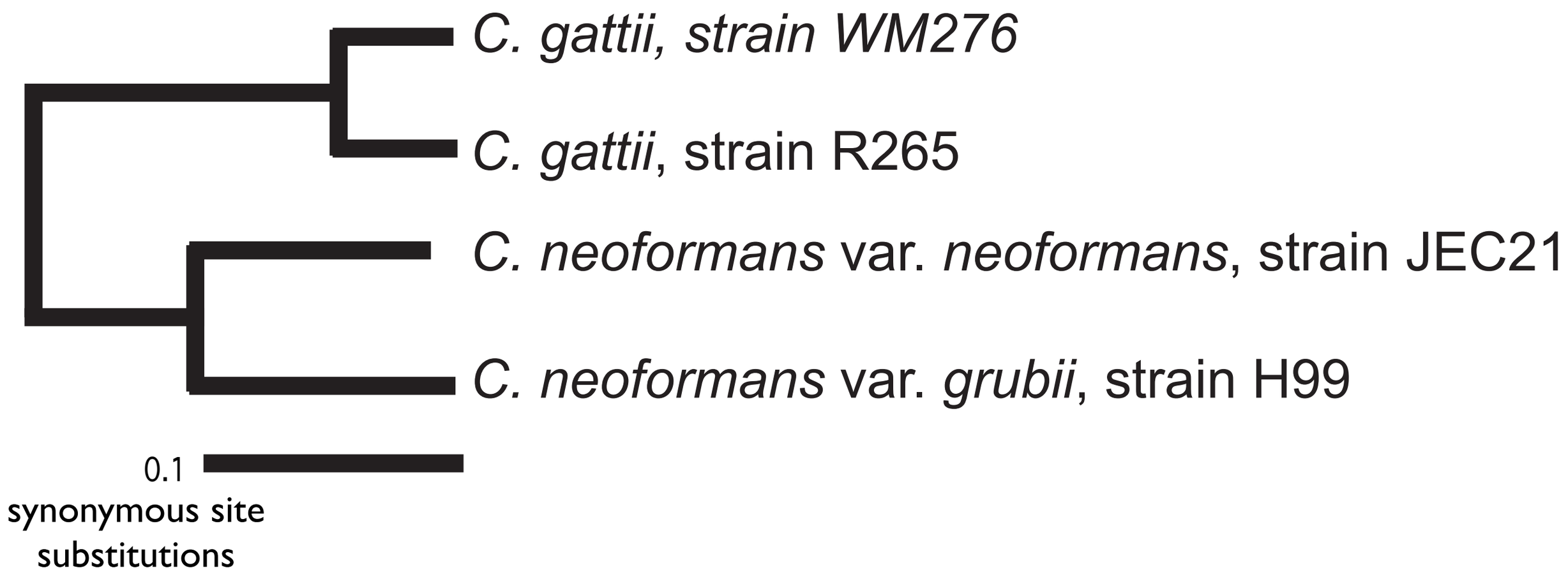
Mechanism of intron loss in fungi

- *S. cerevisiae* and Hemiascomycota have undergone intron loss.
- How are introns lost from the genome?
  - Are they lost independently?
  - Are they lost many at a time?
- What is the molecular mechanism of loss?

Models of intron loss

- All introns in *S. cerevisiae* are in 5' end of gene.
- G. Fink proposed transcripts recombine with genome 3' -> 5' explaining 5' retention bias.
- Most lost events in *S. cerevisiae* occurred too long ago to find evidence of mechanism with a comparative approach.

Sequenced *Cryptococcus* genomes

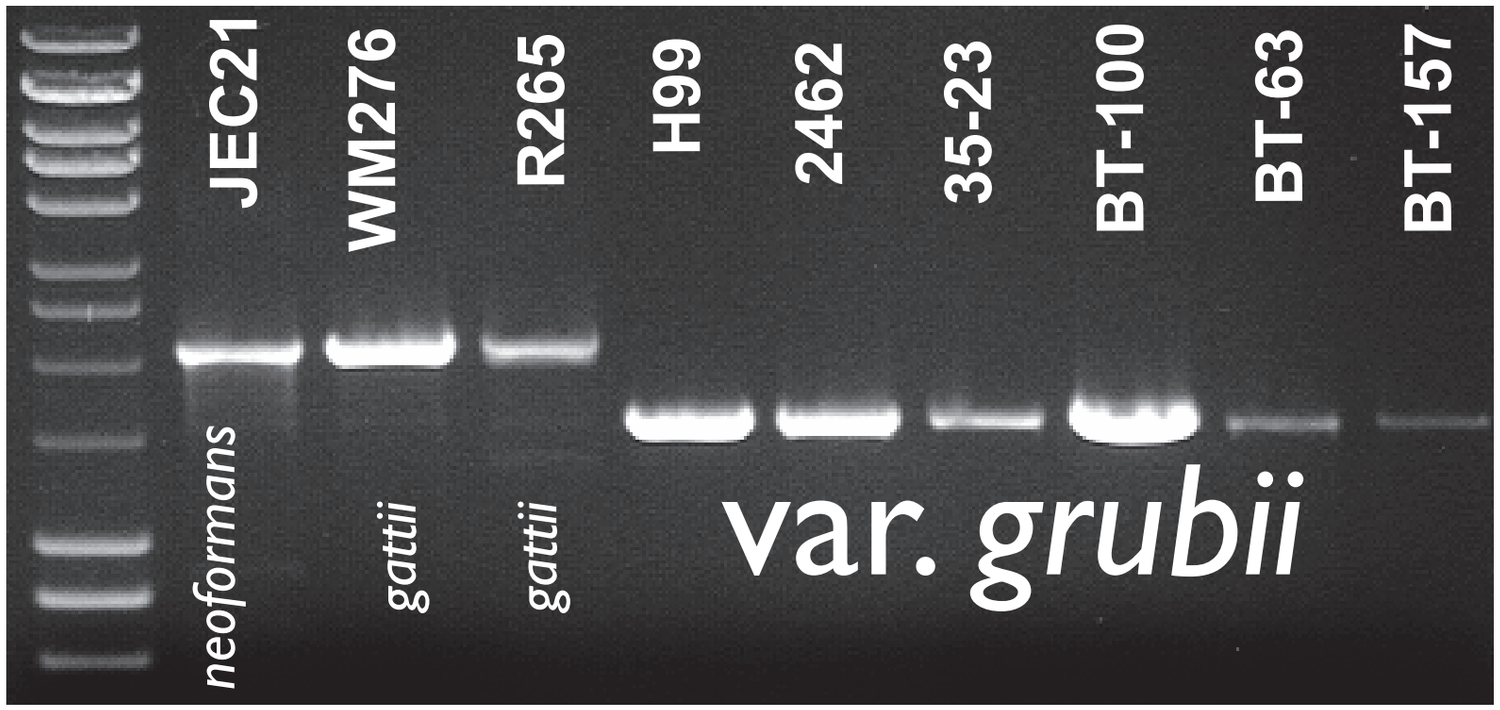


Screen for intron changes

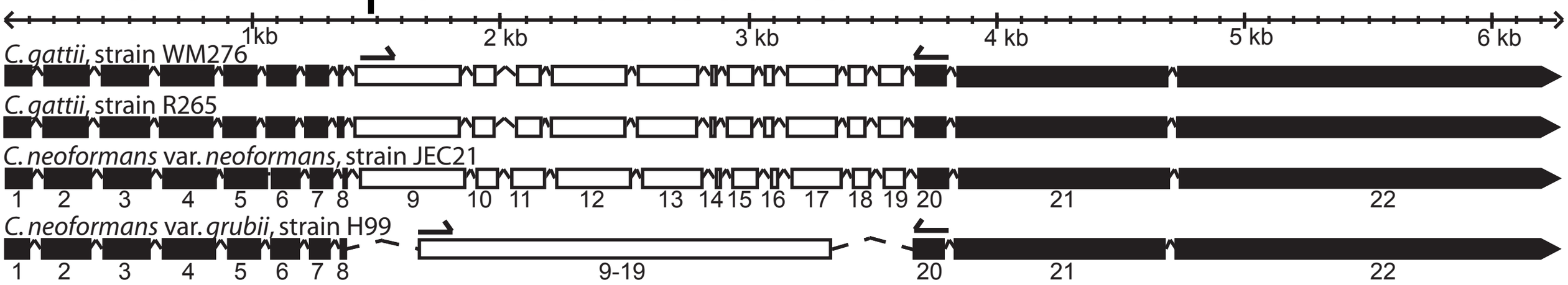
- Annotate 3 *Cryptococcus* genomes (var. *grubii* and 2 var. *gattii* genomes)
- Identify and align 4-way orthologous genes
  - 5298 orthologous genes (out of ~6500)
- Identify intron position changes

Stajich and Dietrich. Euk Cell *In press*.

Intron loss in var. *grubii*



CNI01550 - putative RNA helicase



Stajich and Dietrich. Euk Cell *In press.*

Intron loss was a precise excision

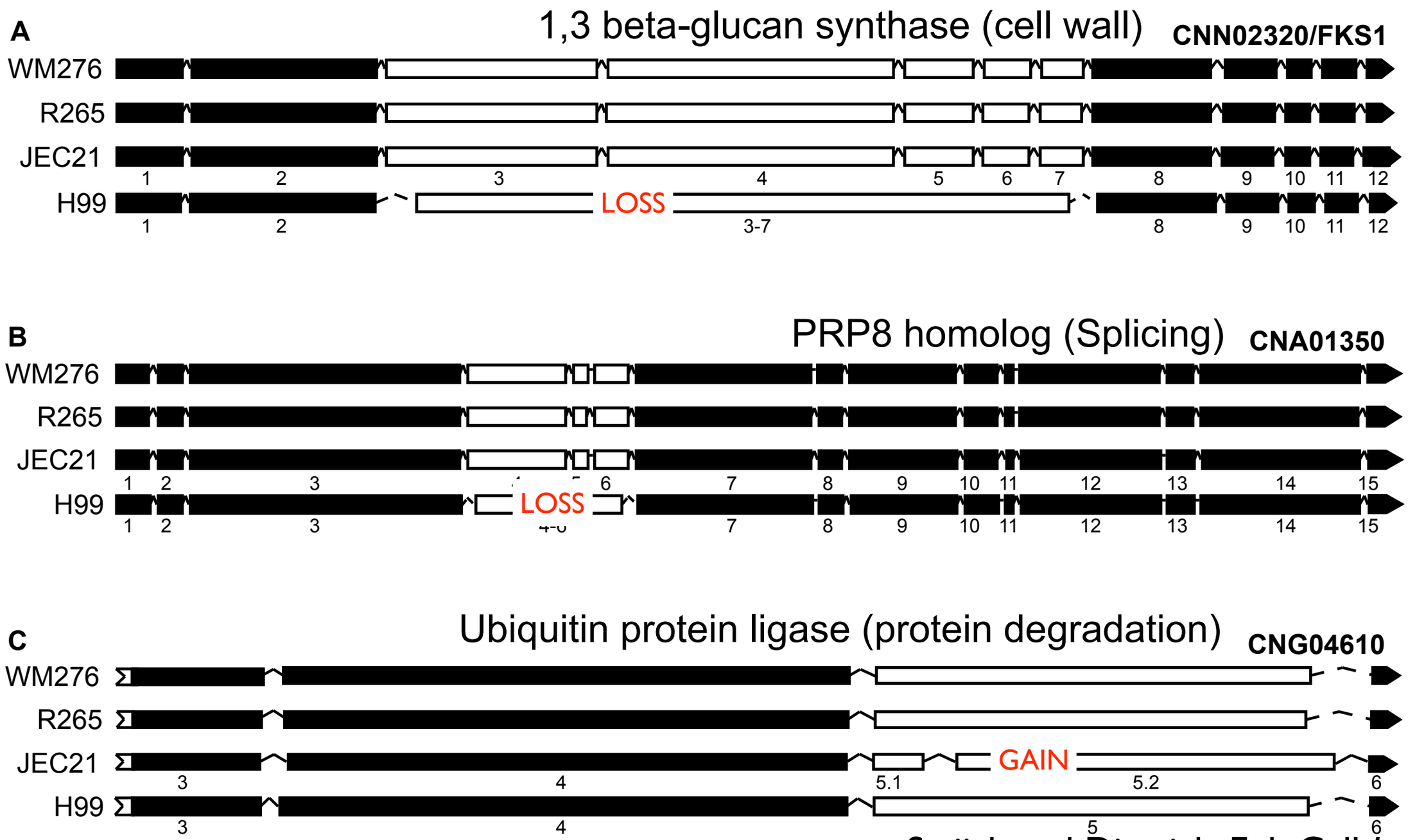
```
R265        CGACAAGTACATAAAACTTTTTTTGTGCCTGGCGCAAAGACTTTCCATTGCTGACAGAAAACAGGTTGAA
WM276       AGACAAGTACATAAAACTTTTTTTGTGTCTGGTGCAAAGATTTTTCATTGCTGACAGAAAACAGGTTGAA
H99         AGACAA-------------------------------------------------------------GTTGAA
JEC21_CDS   AGACAA-------------------------------------------------------------GTTGAA
JEC21       AGACAAGTACATACTAGTCCTTGTG---CTATCCCAAAGACTTT-CATTGCTGACAGAAAACAGGTTGAA
             *****                                                      ******

R265        CGCTGCCGAATTATGTCGATGTTGGAGATTTCTTGAGGTAAGCAACAGACTCGTAACAGCTTGTTCGGTC
WM276       CGCTGCCGAACTATGTCGATGTTGGAGATTTCTTGAGGTAAGCAACAGACTCGTAACAGCTTGTTCGGTC
H99         CCCTGCCGAATTATGTCGACGTTGGAGATTTCTTGAG-------------------------------
JEC21_CDS   CCCTGCCGAATTATGTCGATGTTGGAGATTTCTTGAG-------------------------------
JEC21       CCCTGCCGAATTATGTCGATGTTGGAGATTTCTTGAGGTACGTCGCAAACTCGTAACAGCTTGTTCGATC
            * ******** ******** ****************
```

Intron Missing

Stajich and Dietrich. Euk Cell *In press.*

Other examples of loss



Stajich and Dietrich. Euk Cell *In press.*

Conclusions

- Intron loss via homologous recombination with spliced transcript.
  - Multiple adjacent introns are lost.
  - Precise deletions of introns.
- Loss biased towards the middle of gene not 3'.

Overall conclusions

- Multiple genome sequences have helped resolve several outstanding questions in evolution introns
  - Origin of introns still mystery, but early eukaryotic genes were complex
  - Suggested intron function in splicing
- Gene family expansions can be important in identifying molecular basis for adaptation