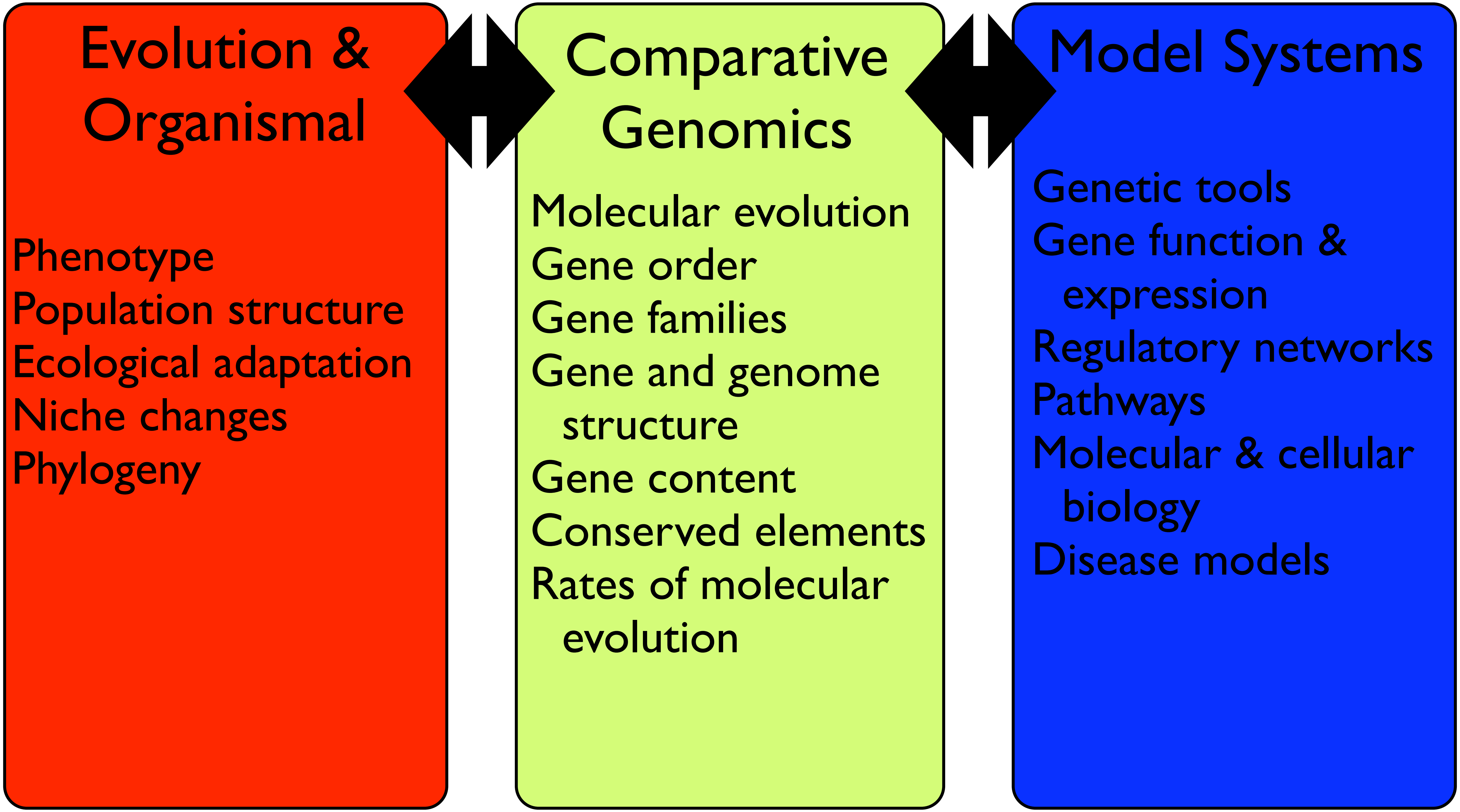
A Comparative Genomic Investigation of Fungal Genome Evolution

Jason Stajich
Duke University
University Program in Genetics & Genomics

Evolutionary genomics



Evolution & Organismal

Phenotype
Population structure
Ecological adaptation
Niche changes
Phylogeny

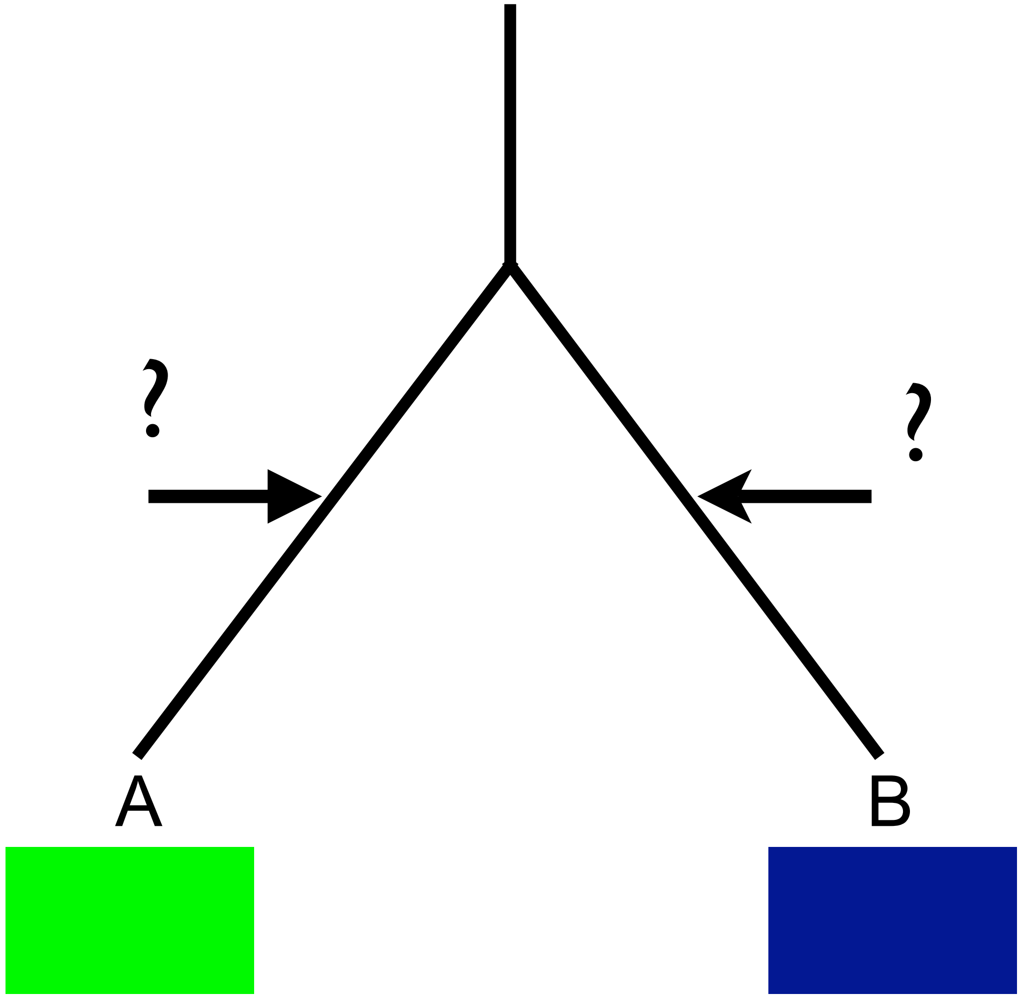
Comparative Genomics

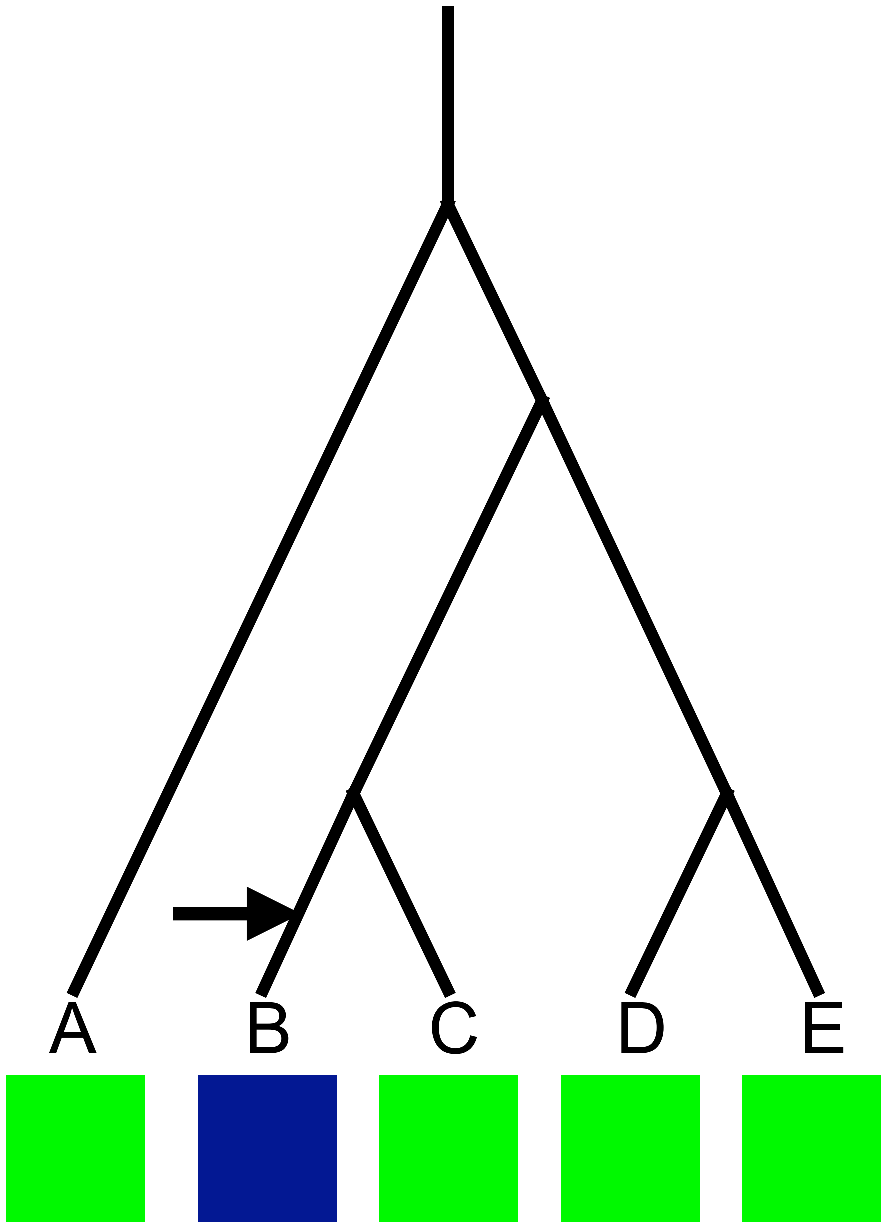
Molecular evolution
Gene order
Gene families
Gene and genome structure
Gene content
Conserved elements
Rates of molecular evolution

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*

Agricultural impact of fungi

www.gov.mb.ca

s of plant disease is caused by

USDA

A.G. Bölker

posit mycotoxins - e.g. ergot

zal fungi provide nutrient

and nitrogen fixation

Impact of fungi on human health

- Biggest risk for immunocompromised individuals
- Primary pathogens
  - *Histoplasma, Coccidioides, Cryptococcus gattii*
- Opportunistic pathogens
  - *Candida albicans, Aspergillus fumigatus, Cryptococcus neoformans, Rhizopus oryzae*

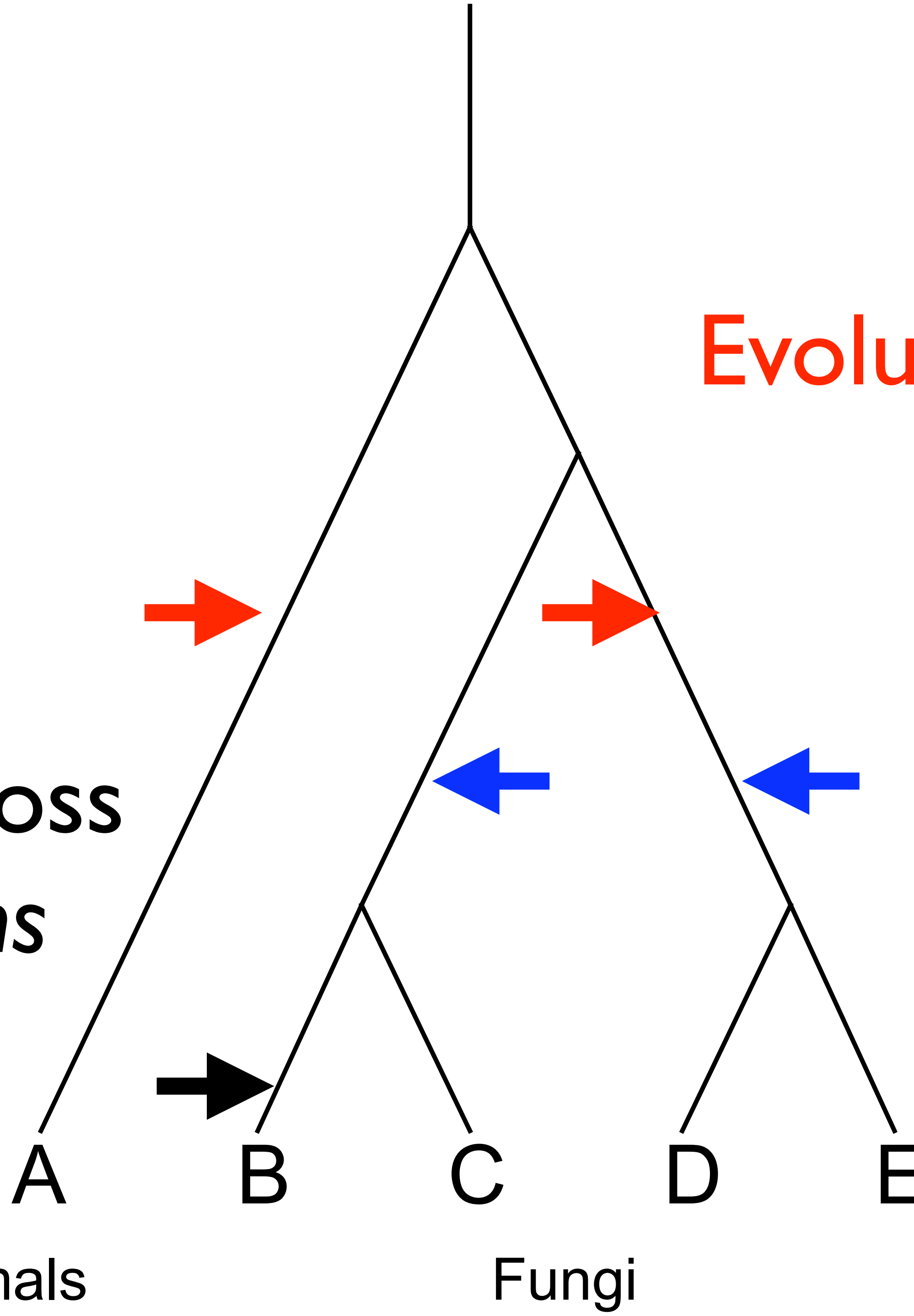
Fungi as genetic models

- Beadle and Tatum (1941) - one gene, one enzyme hypothesis in *Neurospora crassa*
- Cell cycle, cell model - *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*
- Straightforward 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
  - 4700 in *A. gossypii*, 5800 in *S. cerevisiae*
  - 16,000 in *R. oryzae* or *S. nodorum*
  - 19,000 in Fruitfly, 25,000 in worm

Fungal comparative genomics

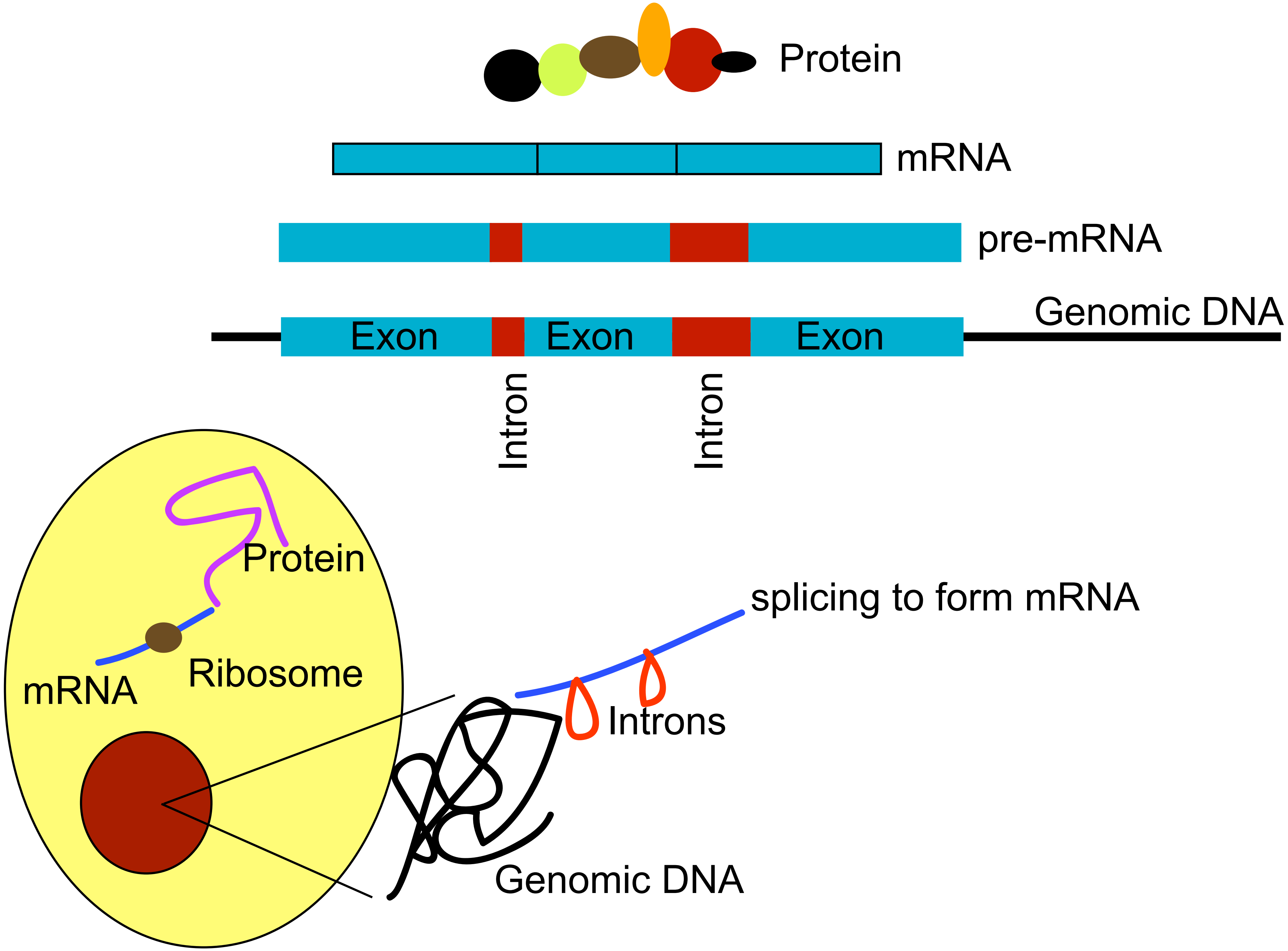


Evolution of fungal introns

Fungal gene family evolution

Recent Intron Loss in *C.neoformans*

Central dogma of eukaryotic biology



Protein

mRNA

pre-mRNA

Genomic DNA

Exon

Exon

Exon

Intron

Intron

Protein

Ribosome

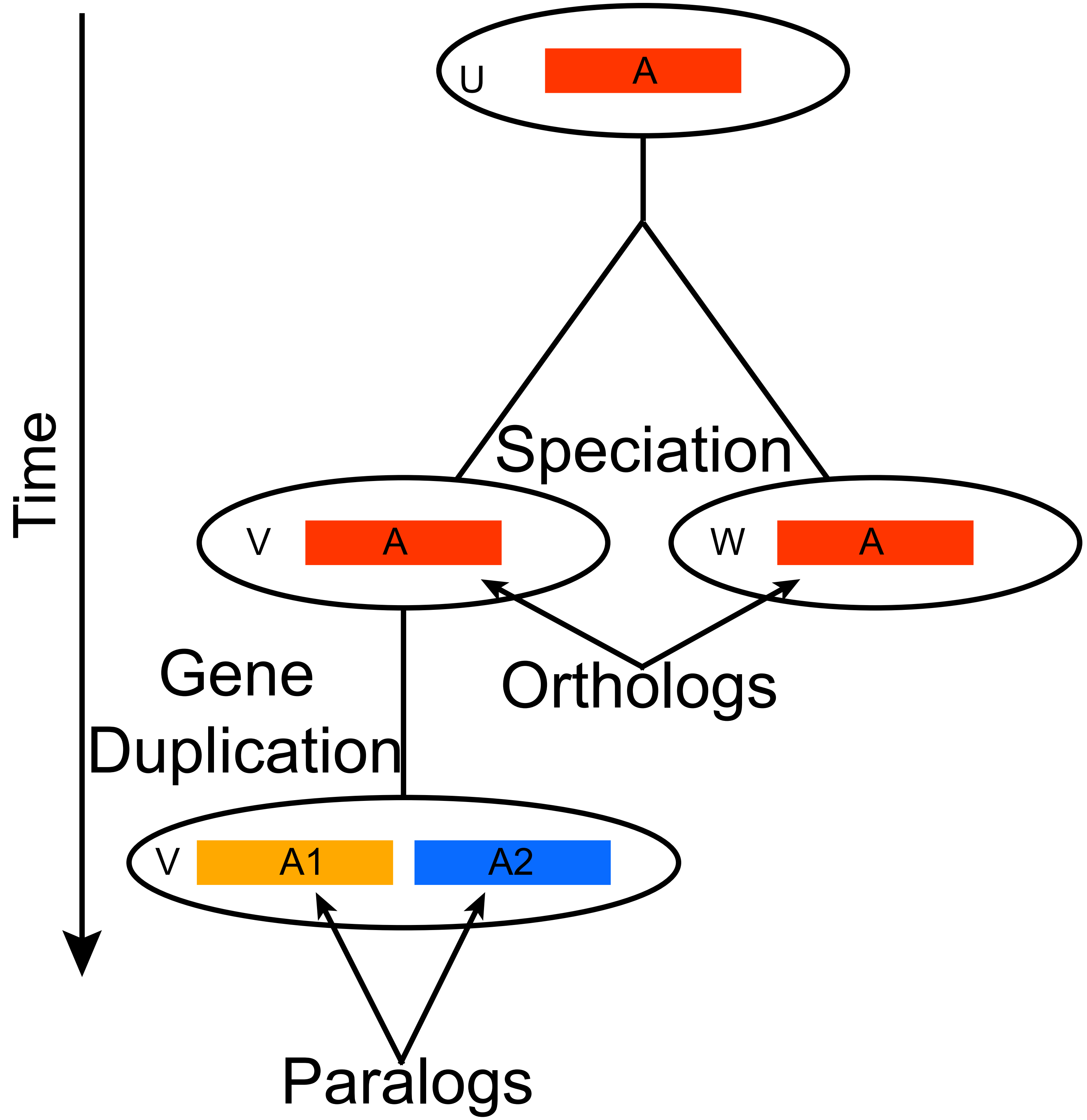
mRNA

splicing to form mRNA

Introns

Genomic DNA

Orthologs and Paralogs

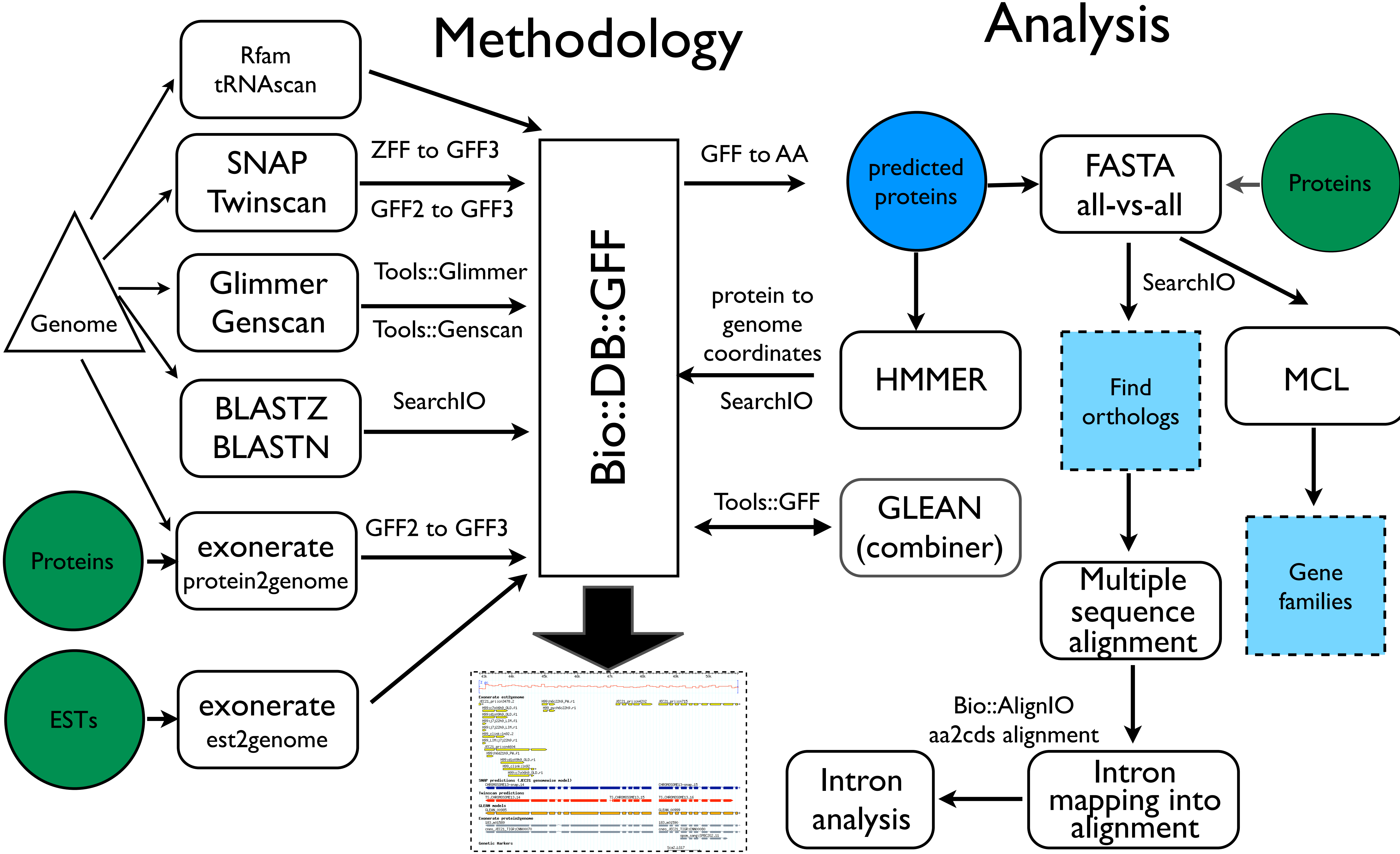


Genome annotation

- Many of the fungal genomes were only assembled genomic sequence.
- Automated annotation pipeline was built to generate to get systematic gene prediction.
- Several gene prediction programs were trained and results were combined to produce composite gene calls

Methodology

Analysis



Genome

Rfam
tRNAscan

SNAP
Twinscan

ZFF to GFF3

GFF2 to GFF3

Glimmer
Genscan

Tools::Glimmer

Tools::Genscan

BLASTZ
BLASTN

SearchIO

Proteins

exonerate
protein2genome

GFF2 to GFF3

ESTs

exonerate
est2genome

Bio::DB::GFF

GFF to AA

protein to
genome
coordinates

SearchIO

Tools::GFF

GLEAN
(combiner)

predicted
proteins

HMMER

FASTA
all-vs-all

Proteins

SearchIO

Find
orthologs

MCL

Gene
families

Multiple
sequence
alignment

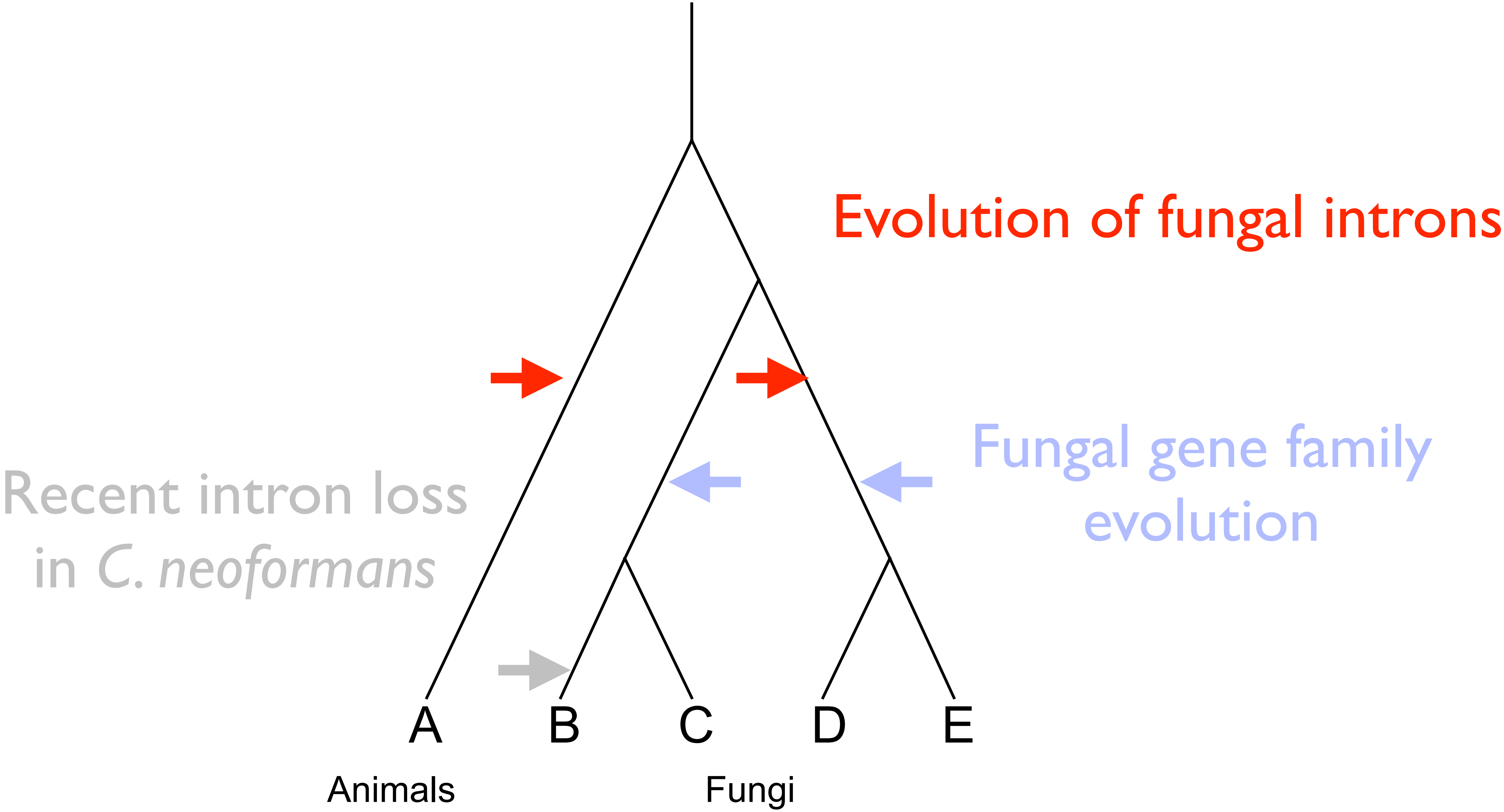
Bio::AlignIO
aa2cds alignment

Intron
mapping into
alignment

Intron
analysis

http://fungal.genome.duke.edu

Fungal comparative genomics

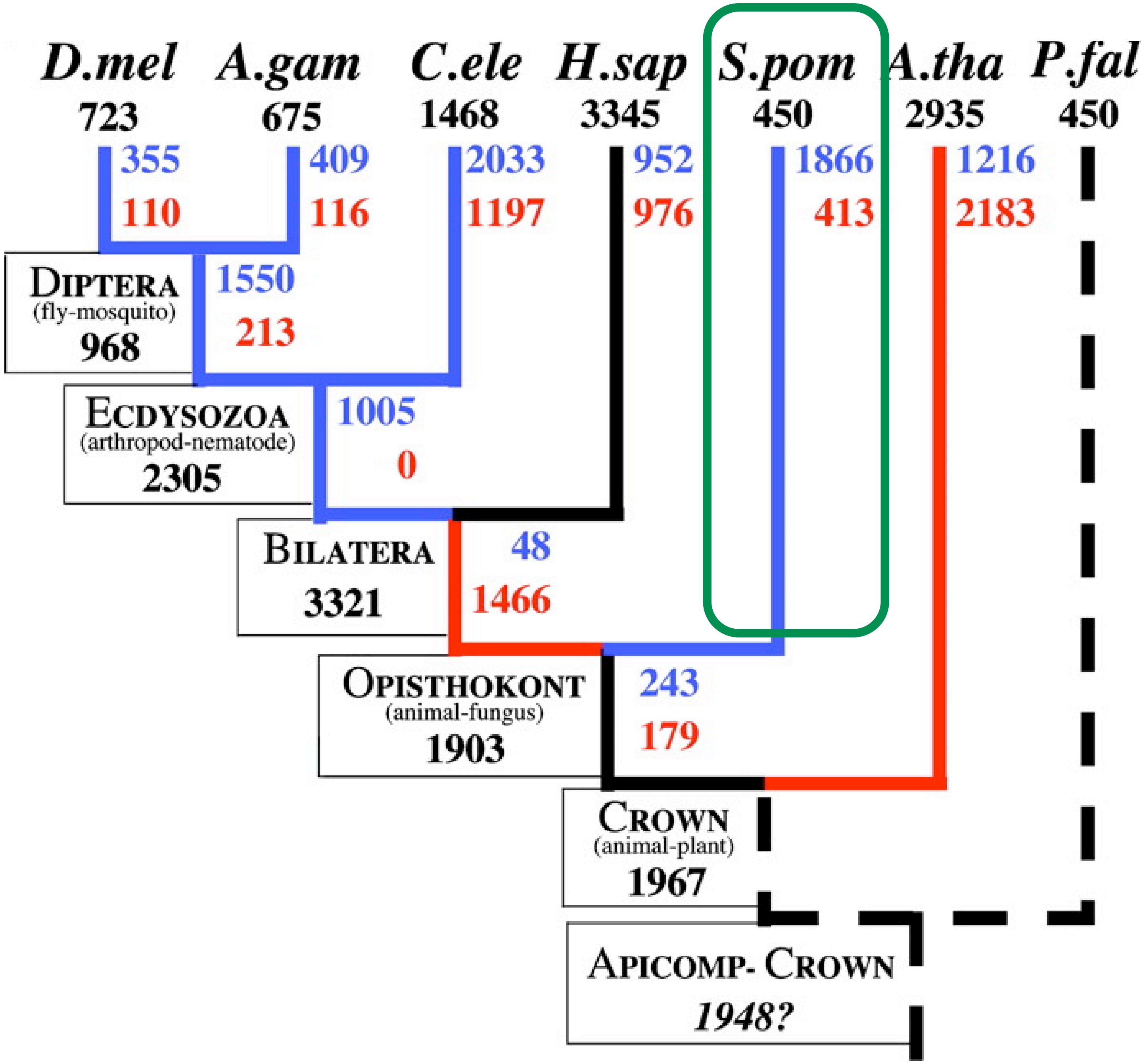

Evolution of fungal introns
Fungal gene family evolution
Recent intron loss in *C. neoformans*
A
B
C
D
E
Animals
Fungi

Evolution of gene structure

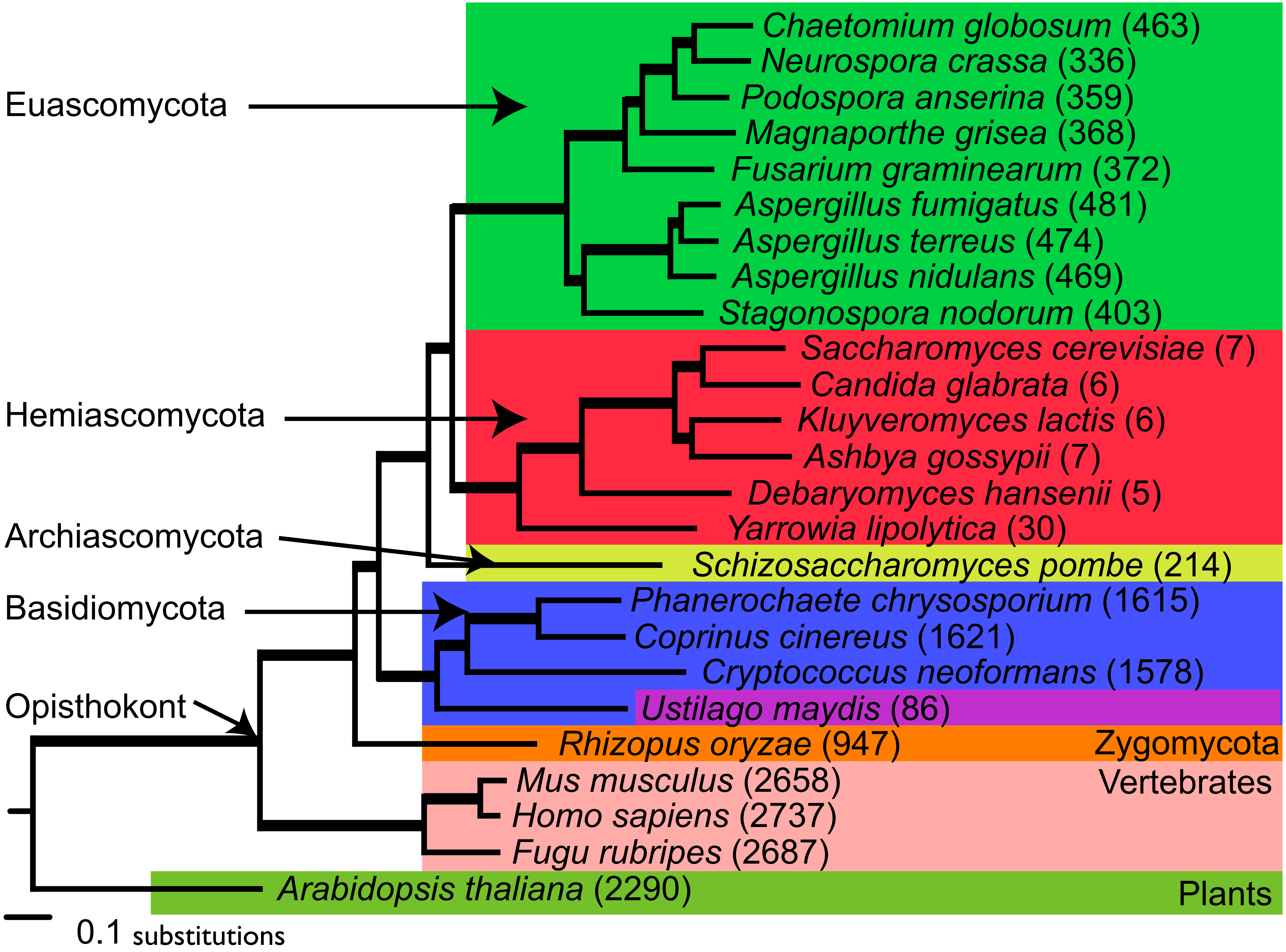
- Present day introns
  - Recent insertions?
    - Introns late hypothesis
  - Formed in eukaryotic ancestor?
    - Introns early hypothesis / exon theory of genes
  - Mixture of two?

Previous work on intron evolution

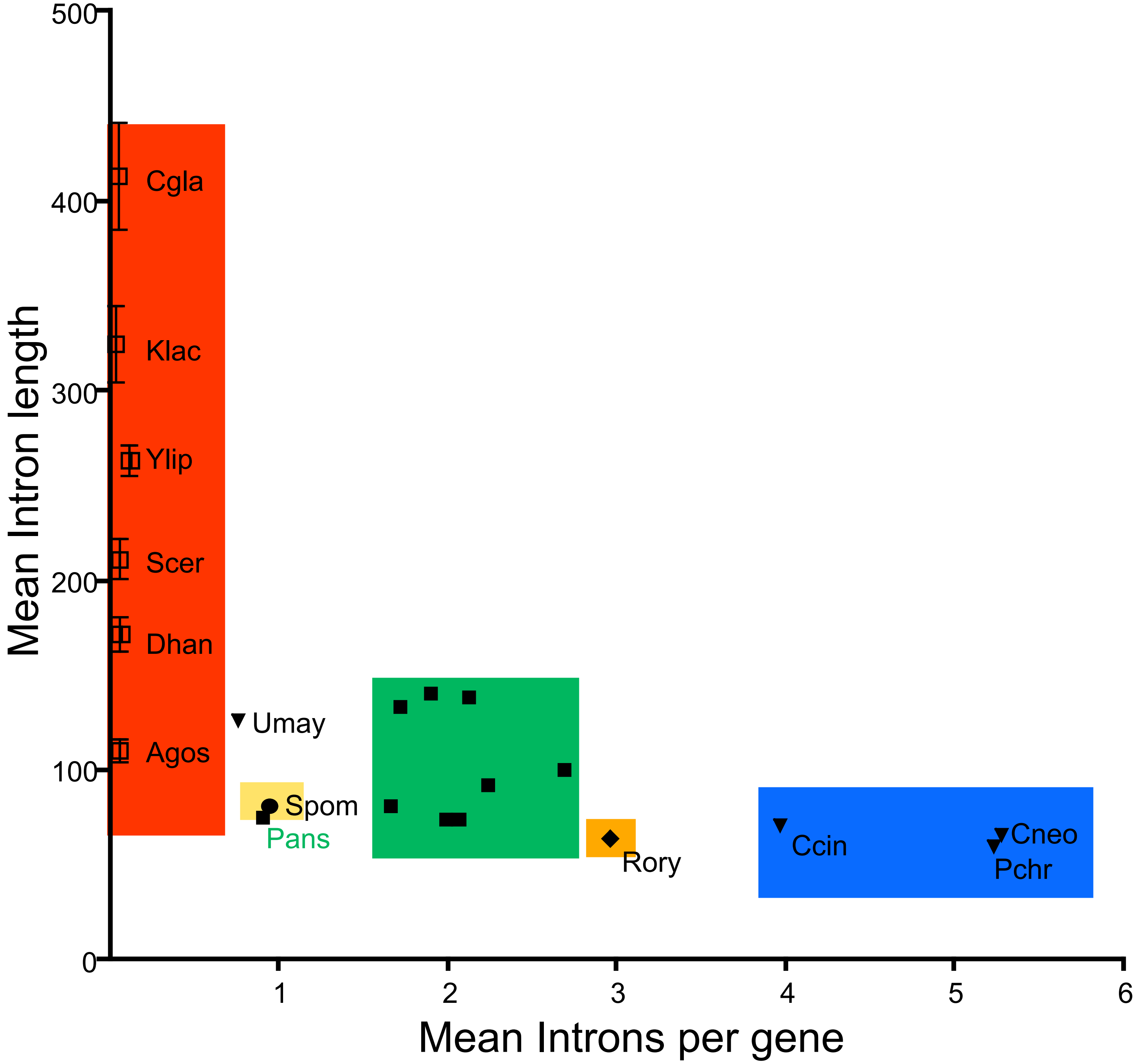
- Rogozin et al. 2003
  - 7 genomes
  - 684 genes, 7236 positions
- Other methods
  - Roy and Gilbert. 2005
  - Csũrös. 2005
  - Nguyen et al. 2006



Calculating intron densities across a phylogeny


Euascomycota
Hemiascomycota
Archiascomycota
Basidiomycota
Opisthokont
Chaetomium globosum (463)
Neurospora crassa (336)
Podospora anserina (359)
Magnaporthe grisea (368)
Fusarium graminearum (372)
Aspergillus fumigatus (481)
Aspergillus terreus (474)
Aspergillus nidulans (469)
Stagonospora nodorum (403)
Saccharomyces cerevisiae (7)
Candida glabrata (6)
Kluyveromyces lactis (6)
Ashbya gossypii (7)
Debaryomyces hansenii (5)
Yarrowia lipolytica (30)
Schizosaccharomyces pombe (214)
Phanerochaete chrysosporium (1615)
Coprinus cinereus (1621)
Cryptococcus neoformans (1578)
Ustilago maydis (86)
Rhizopus oryzae (947)
Zygomycota
Mus musculus (2658)
Vertebrates
Homo sapiens (2737)
Fugu rubripes (2687)
Arabidopsis thaliana (2290)
Plants
0.1 substitutions

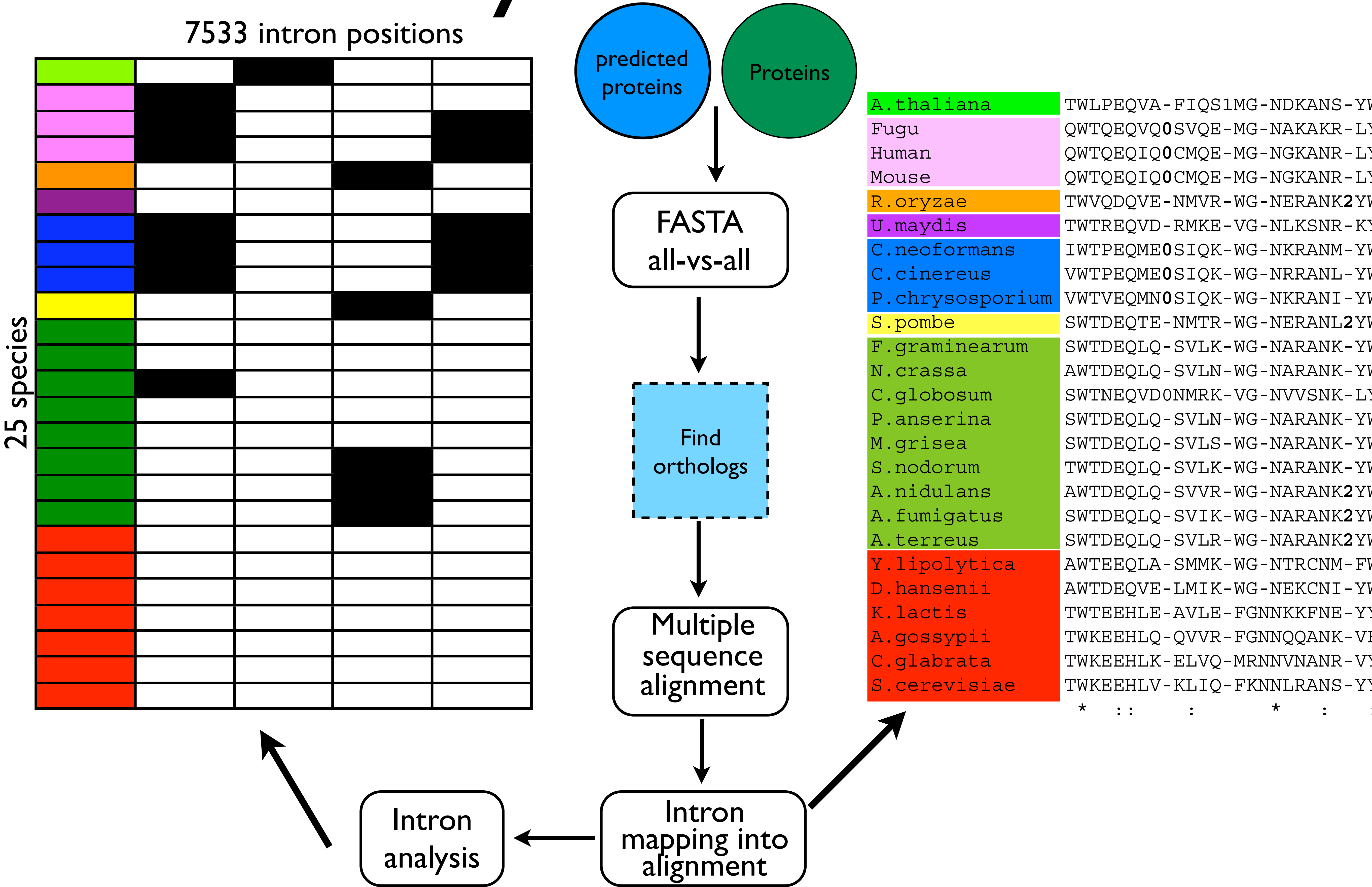
Intron frequency varies among the fungi



Analysis of whole genomes

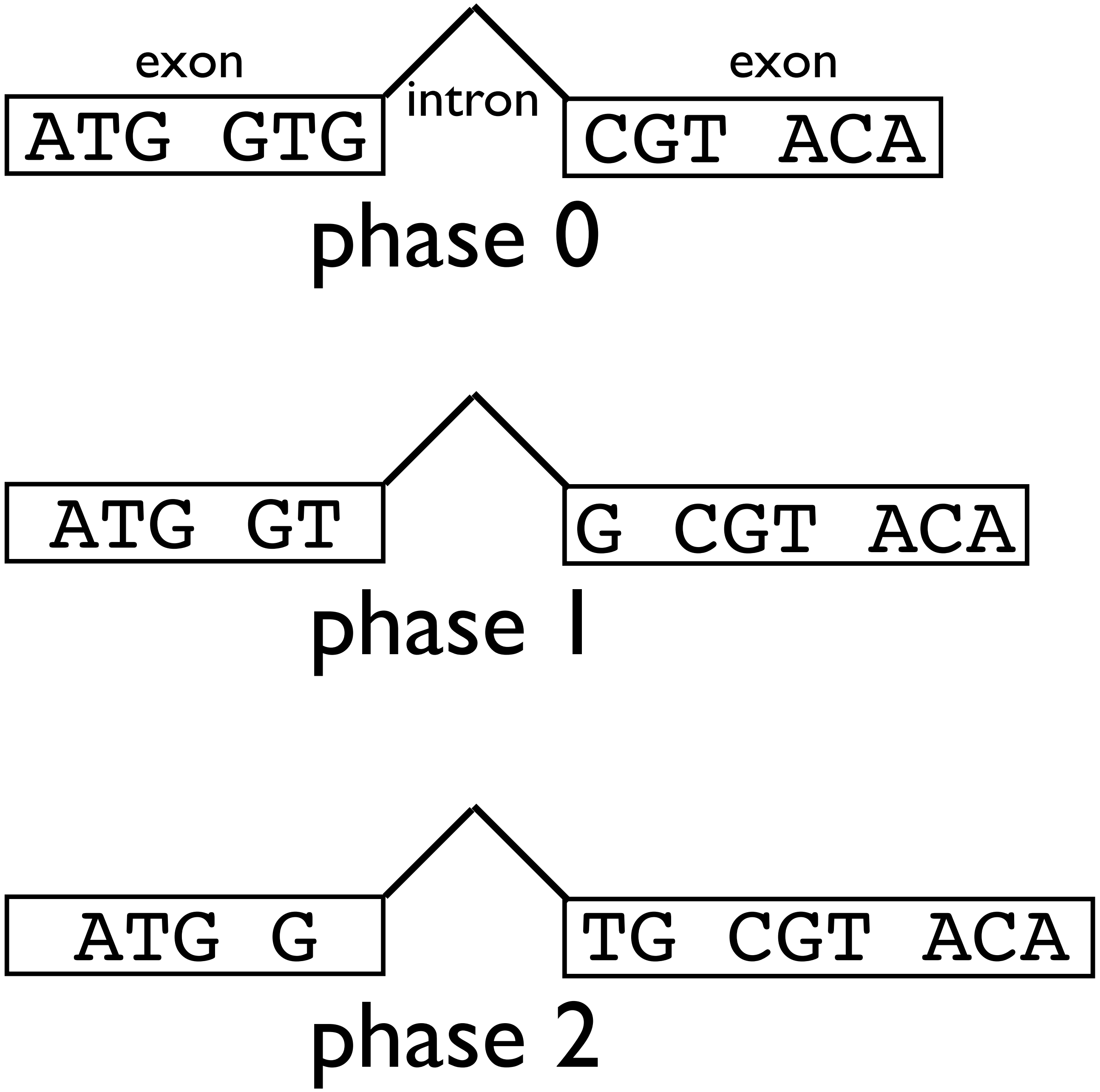
- 25 entire genomes
  - 21 fungi, 3 vertebrates, 1 plant
- Largest dataset ever assembled for intron analysis
- 1160 orthologous genes
- 7533 intron positions
- 4.15 Mb coding sequence (CDS) per genome

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-V
C.glabrata        TWKEEHLK-ELVQ-MRNNVNANR-VY
S.cerevisiae      TWKEEHLV-KLIQ-FKNNLRANS-YY
                   *   ::     :       *   :
```

Intron phase



exon

intron

exon

ATG GTG

CGT ACA

phase 0

ATG GT

G CGT ACA

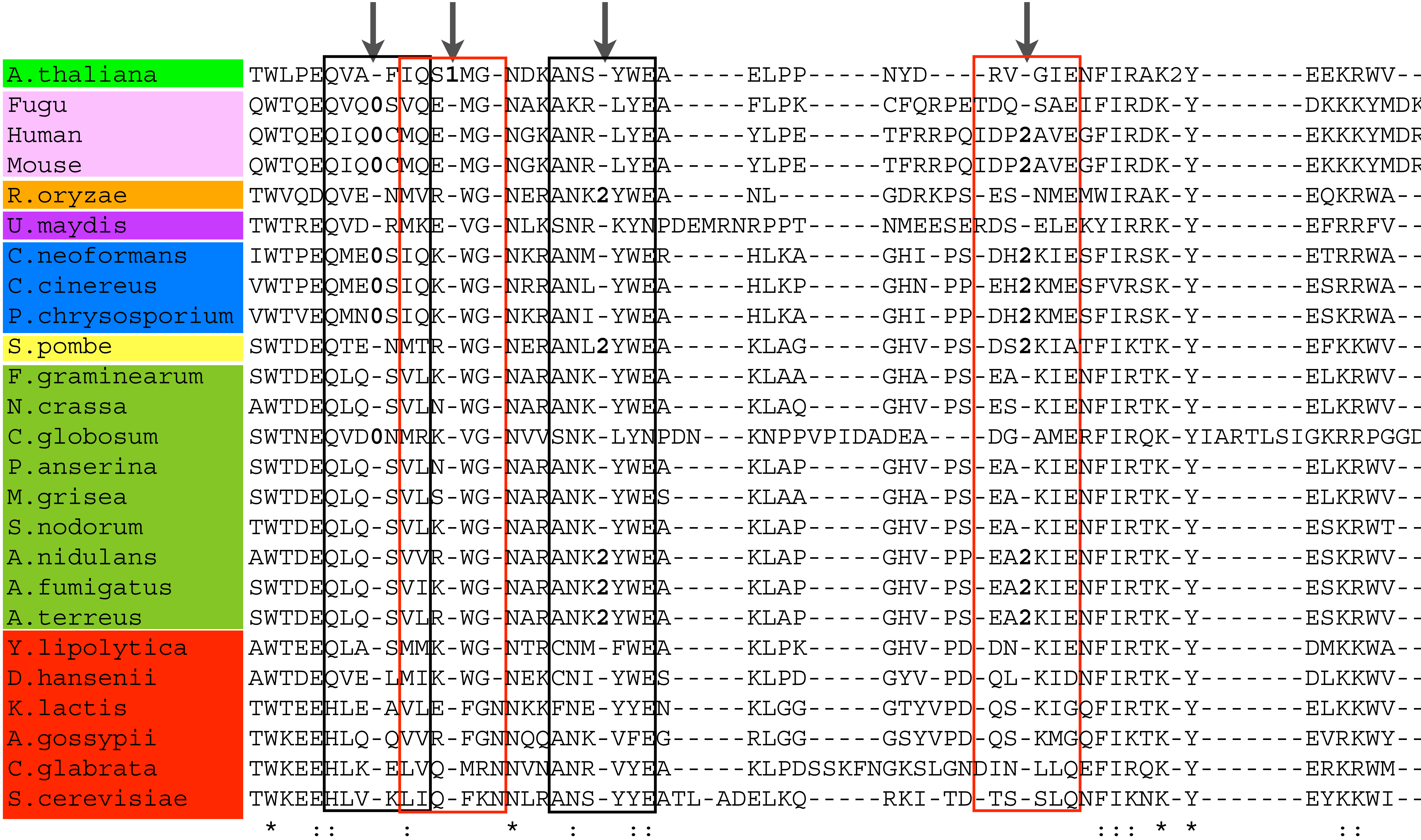
phase 1

ATG G

TG CGT ACA

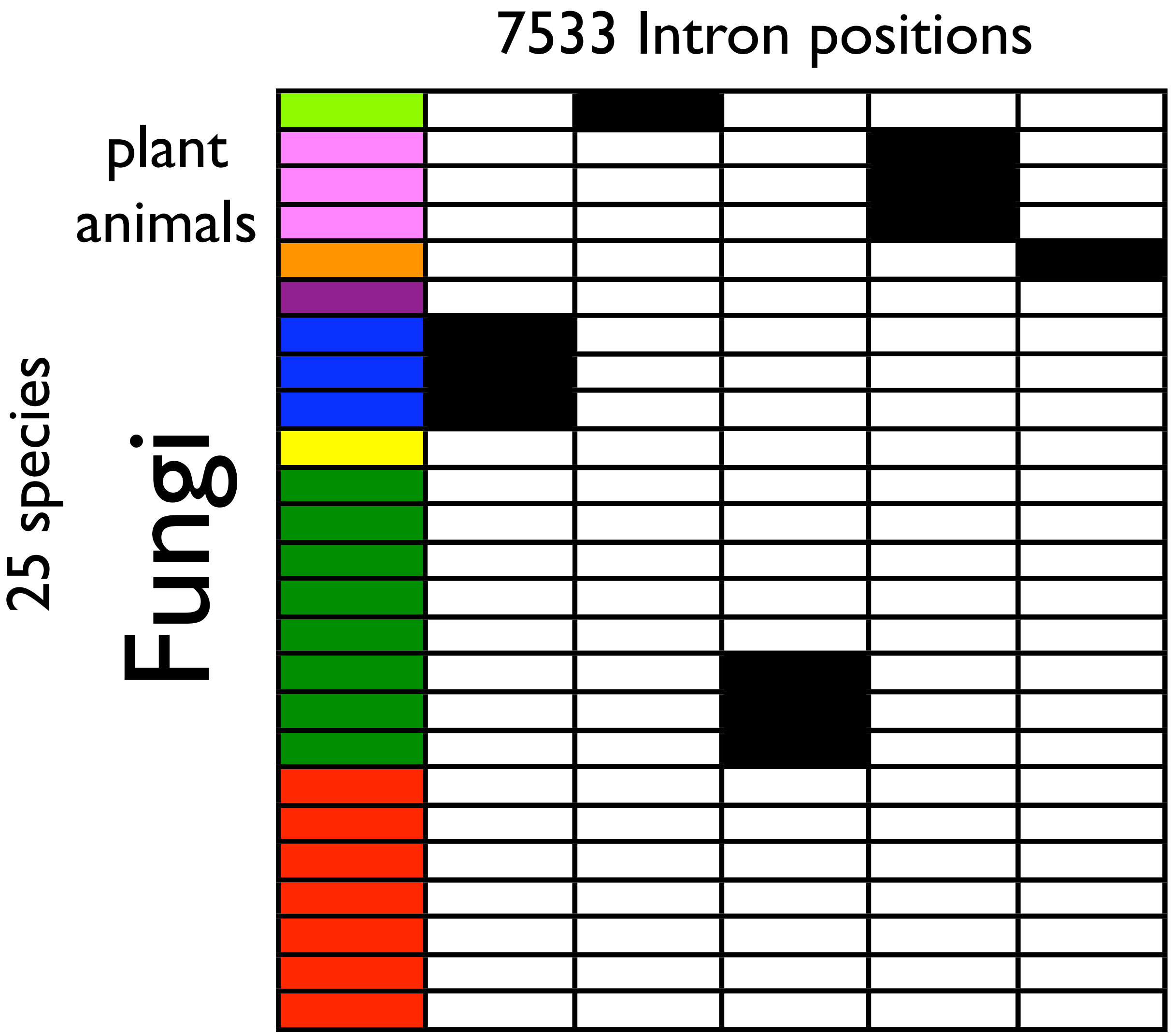
phase 2

Conserved intron positions

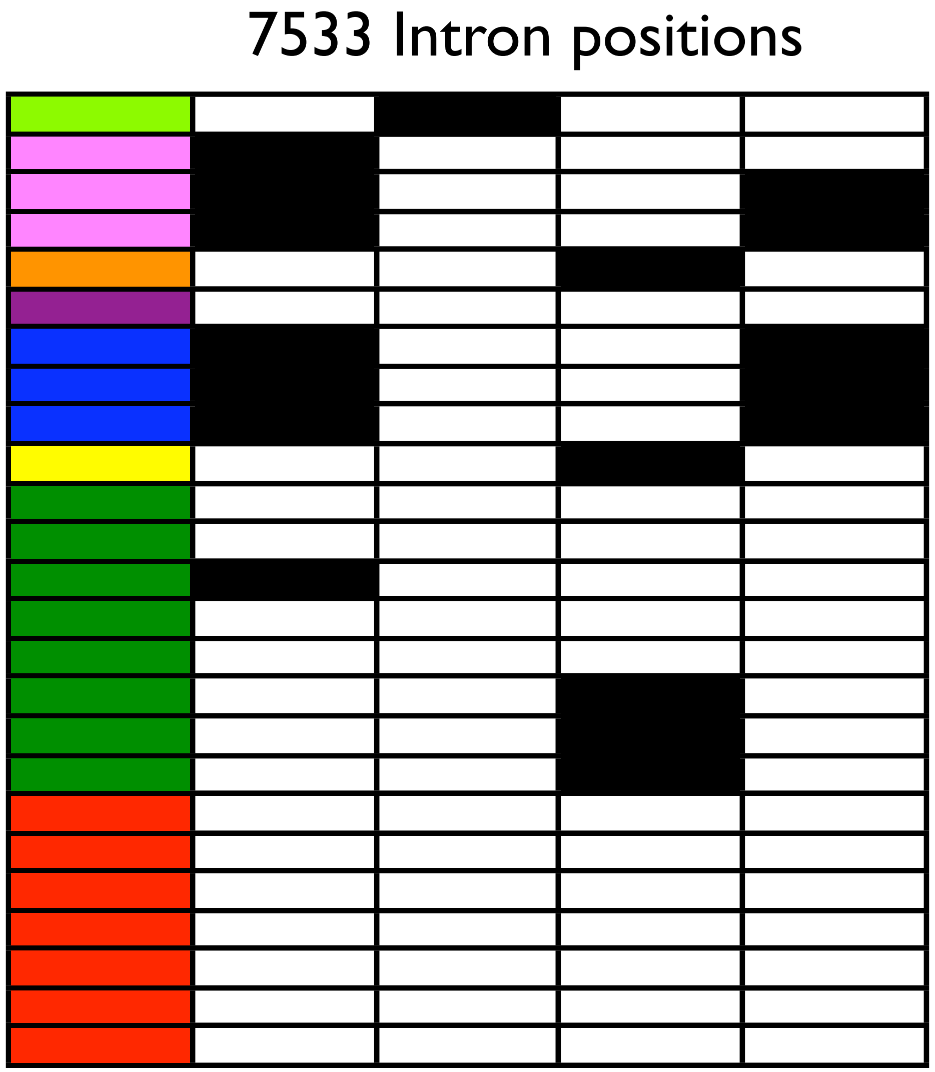


```
A.thaliana       TWLPEQVA-FIQS1MG-NDKANS-YWEA-----ELPP-----NYD----RV-GIENFIRAK2Y-------EEKRWV--
Fugu             QWTQEQVQ0SVQE-MG-NAKAKR-LYEA-----FLPK-----CFQRPETDQ-SAEIFIRDK-Y-------DKKKYMDK
Human            QWTQEQIQ0CMQE-MG-NGKANR-LYEA-----YLPE-----TFRRPQIDP2AVEGFIRDK-Y-------EKKKYMDR
Mouse            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

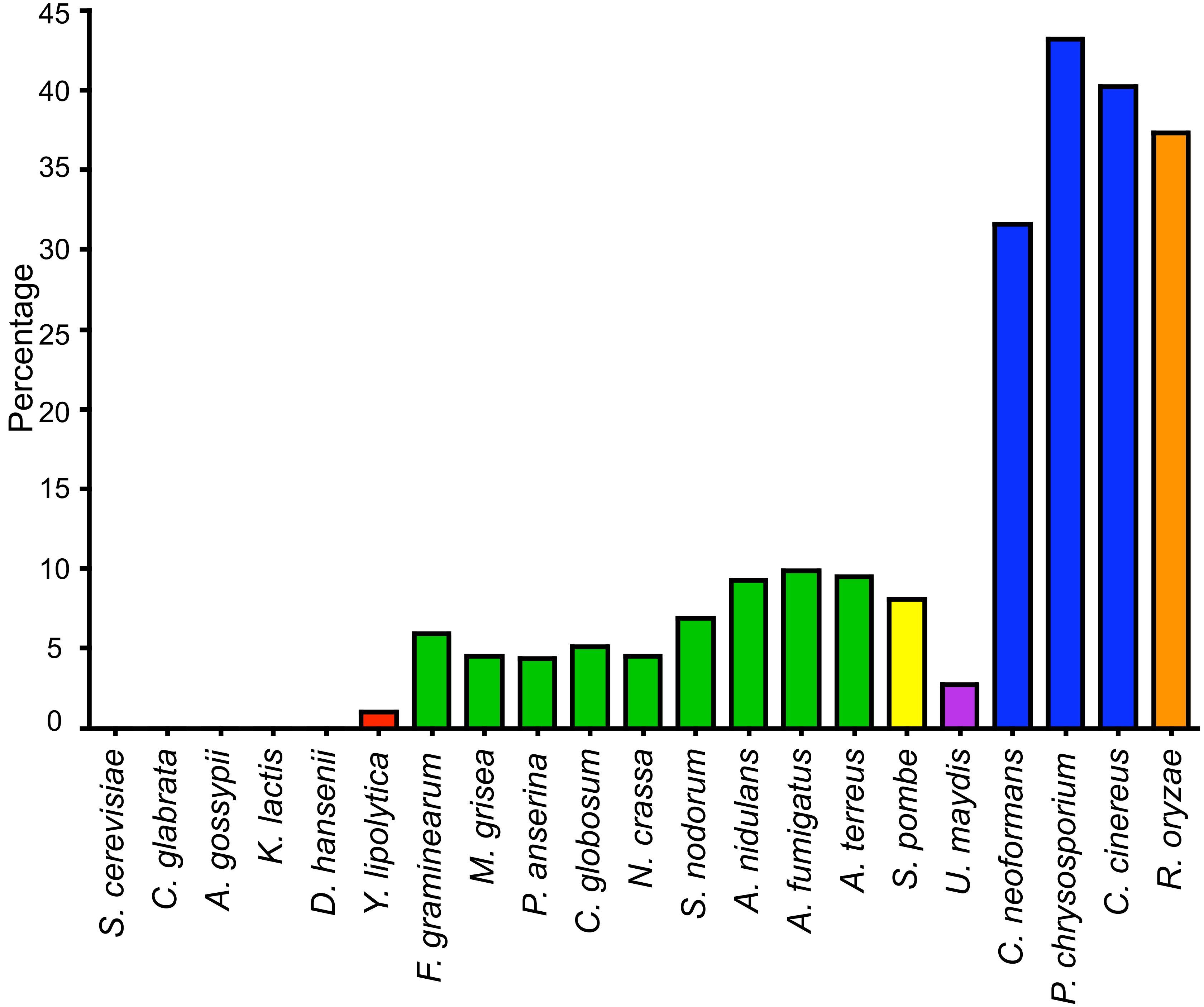


Introns late



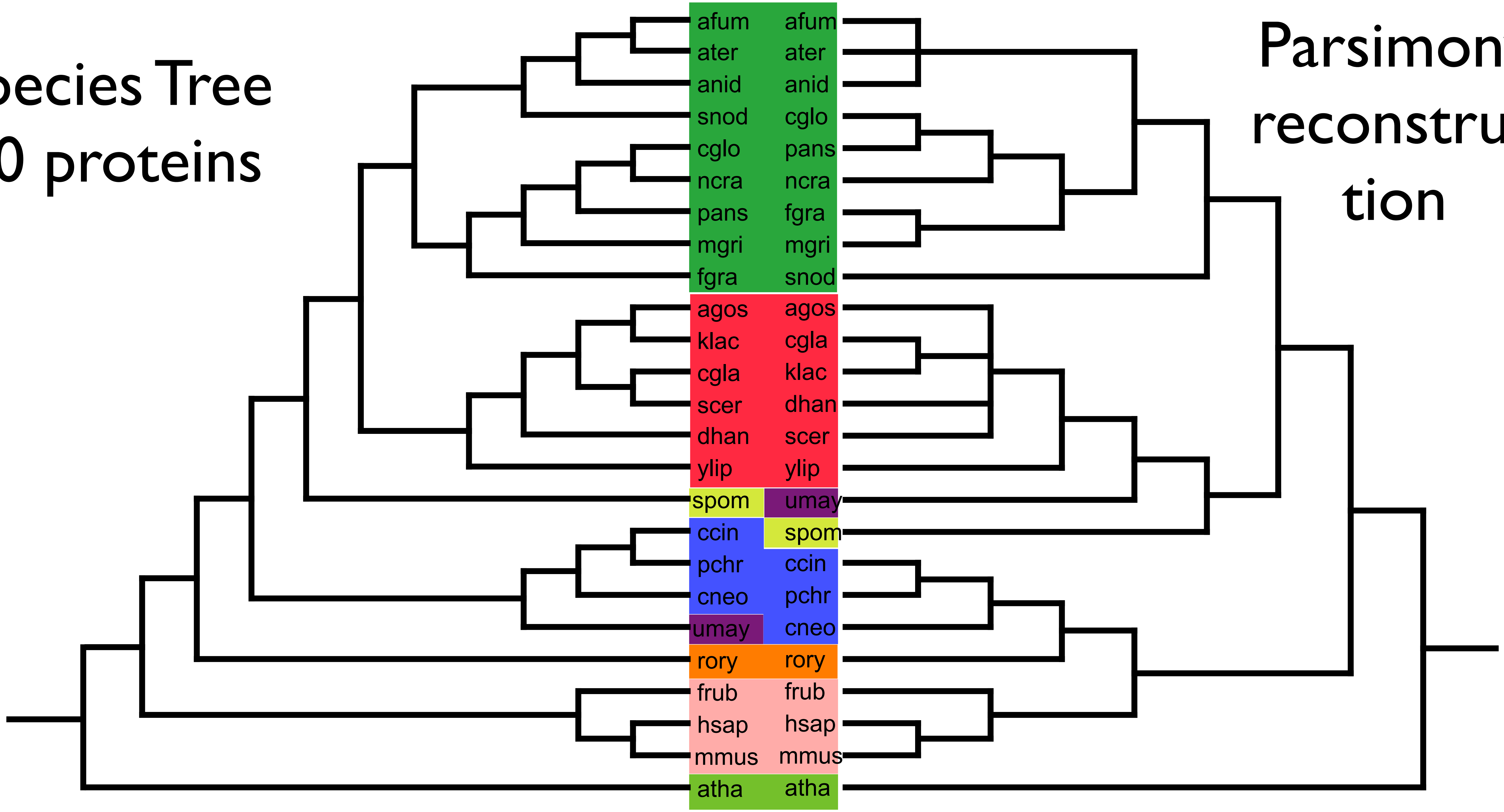
Introns early(ier)

Intron positions shared with animals or plants



Phylogenetic signal in intron positions

Species Tree
30 proteins

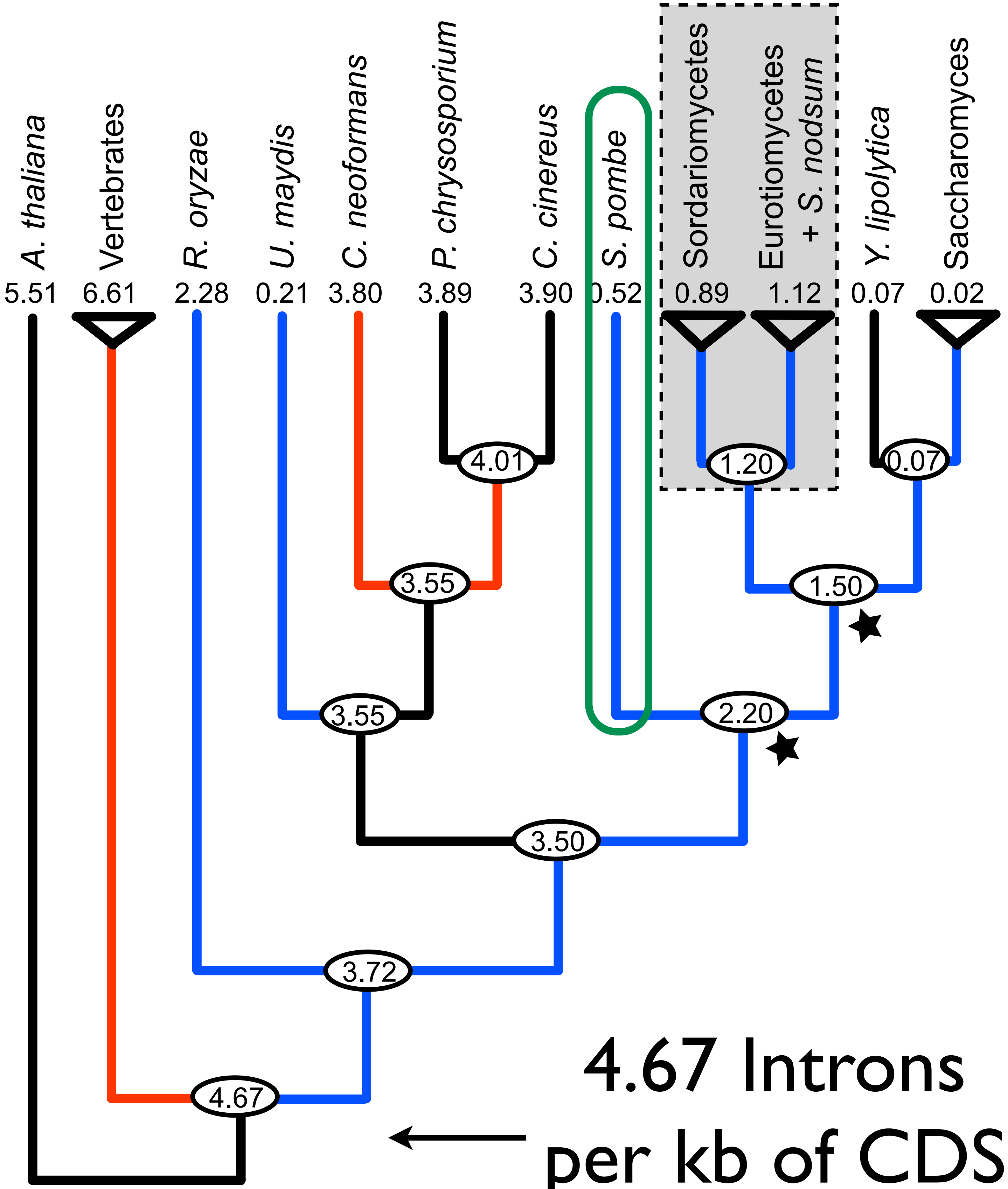


Parsimony
reconstruc
tion

Intron position reconstruction

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

Reconstruction of ancestral intron densities



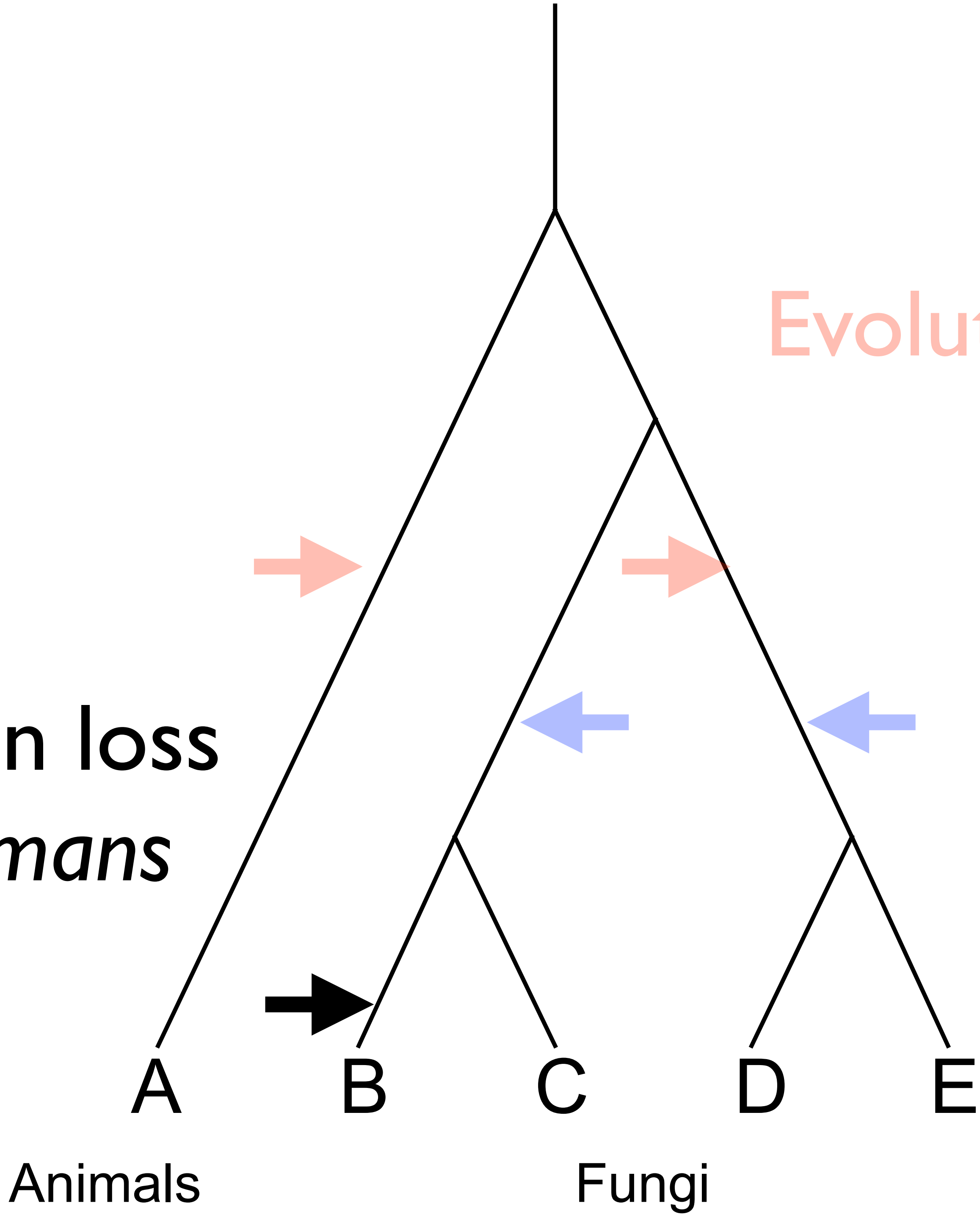
4.67 Introns per kb of CDS

Exon length of 214 bp

Conclusions

- Early eukaryotic crown genes were complex!
  - Ancestor had 70% of the introns in vertebrates
  - More introns than previously reported
- Intron loss has dominated among the fungi
  - Hemiascomycota experienced loss
- Sampling can bias interpretations - all fungi are not equal.

Fungal comparative genomics



Evolution of fungal introns

Fungal gene family evolution

Recent intron loss in *C. neoformans*

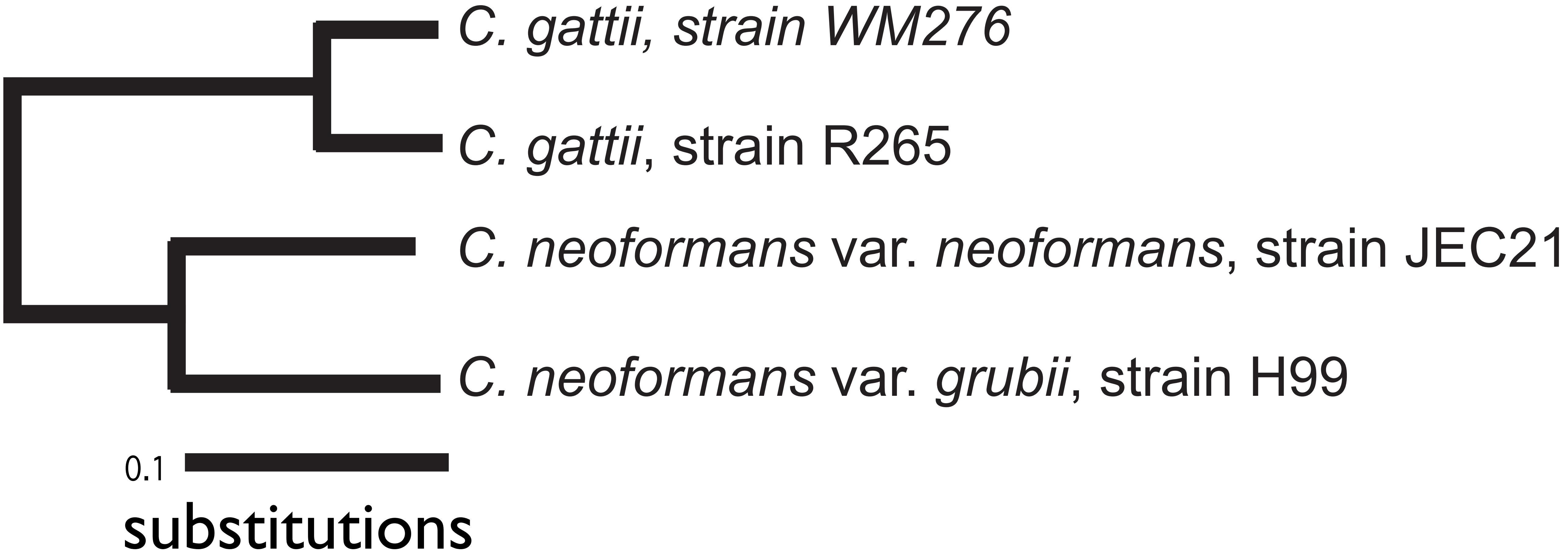
Mechanism of intron loss

- *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?
- 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.
- In *S. cerevisiae* most intron loss events occurred too long ago so little evidence supporting any mechanism

Sequenced *Cryptococcus* genomes



C. gattii, strain WM276

C. gattii, strain R265

C. neoformans var. *neoformans*, strain JEC21

C. neoformans var. *grubii*, strain H99

0.1

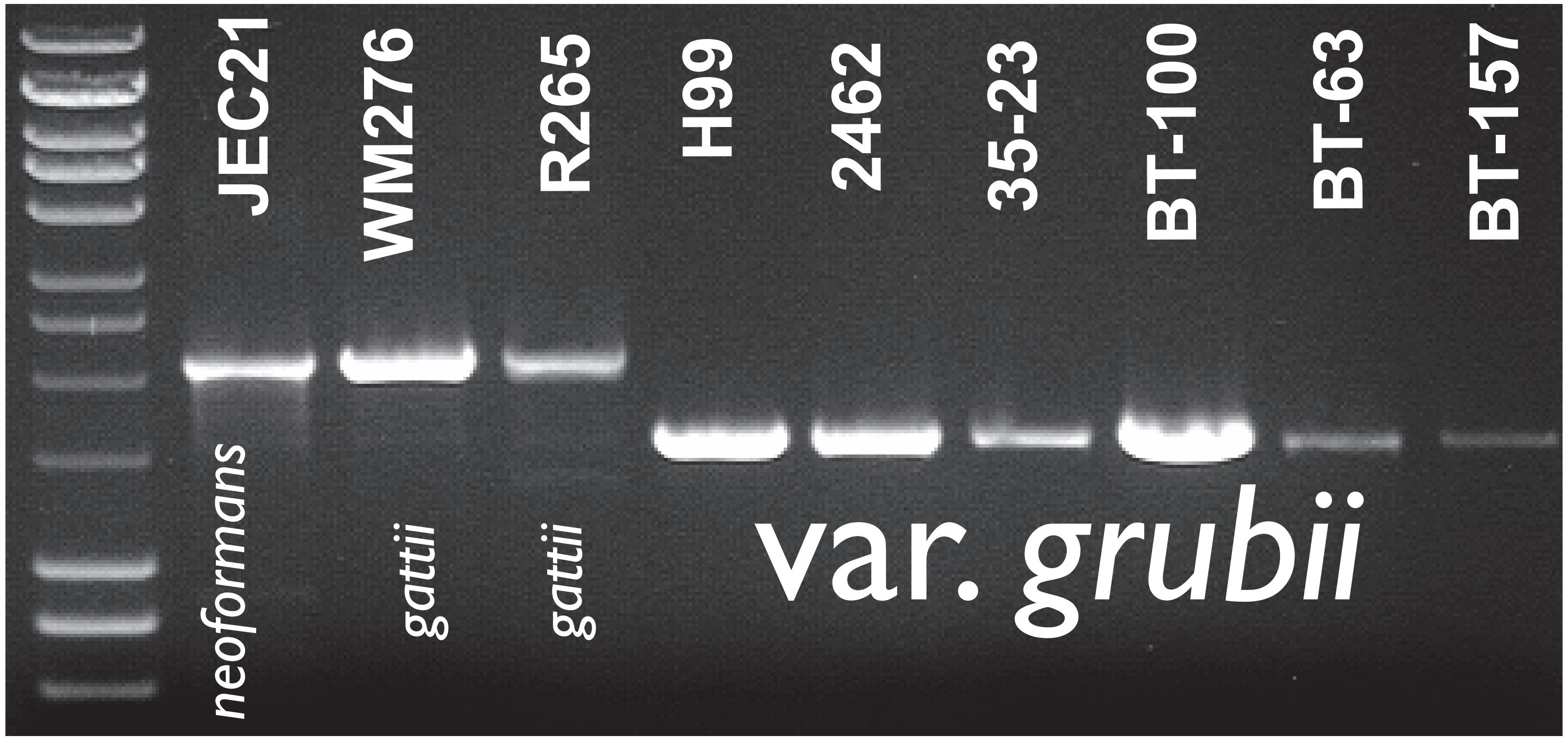
substitutions

Screen for intron changes

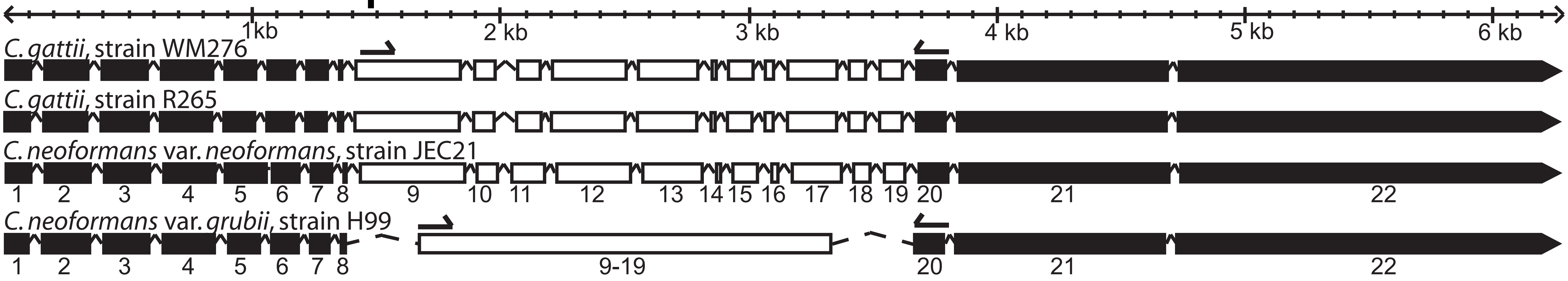
- Annotate 3 *Cryptococcus* genomes (var. *grubii* and 2 var. *gattii* genomes)
- Identify and align 4-way orthologous genes
  - 5298 orthologous genes (out of ~6500)
- Identify intron position changes

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

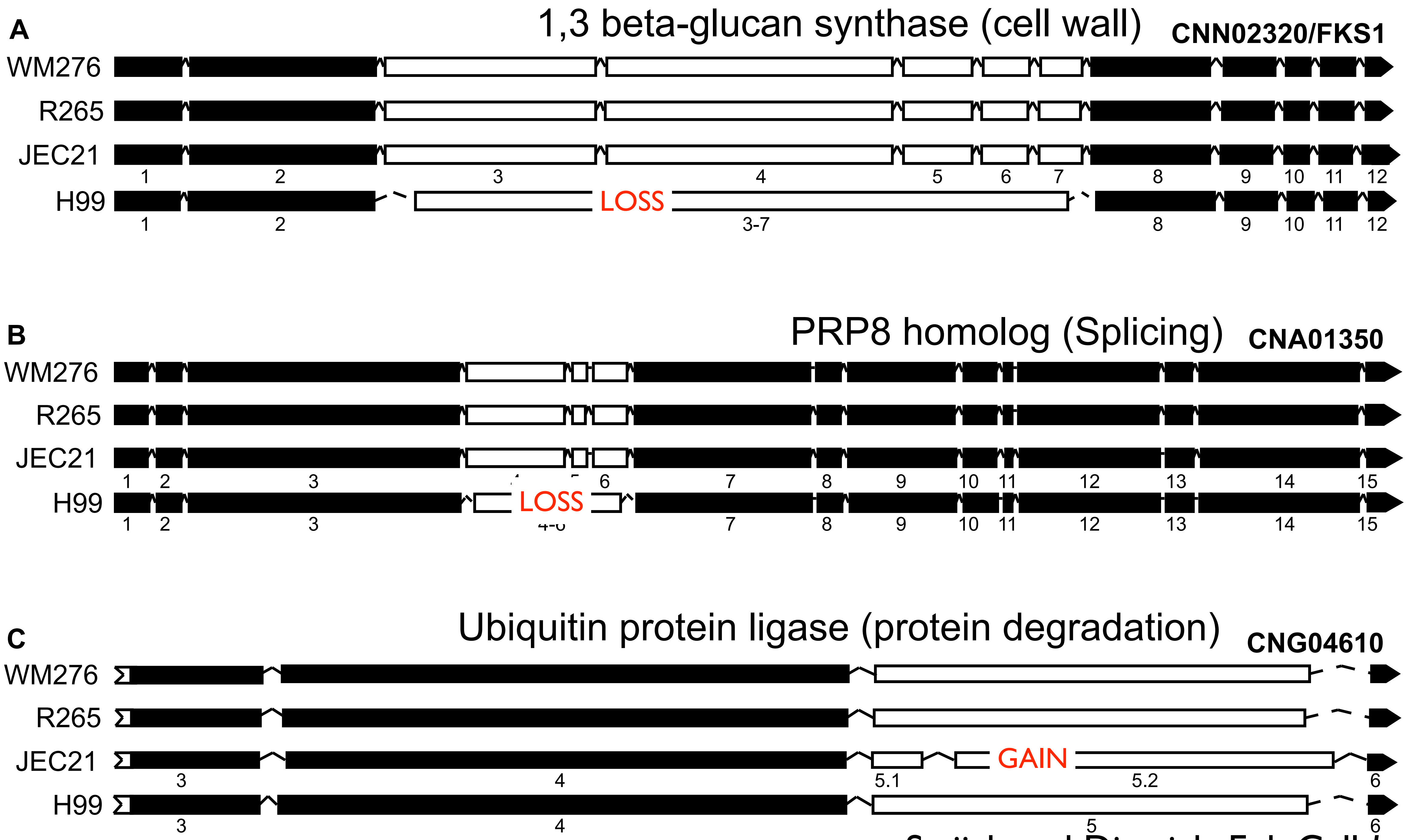
```
R265       CGACAAGTACATAAAACTTTTTTTGTGCCTGGCGCAAAGACTTTCCATTGCTGACAGAAAACAGGTTGAA
WM276      AGACAAGTACATAAAACTTTTTTTGTGTCTGCTGCAAAGATTTTTCATTGCTGACAGAAAACAGGTTGAA
H99        AGACAA----------------------------------------------------------GTTGAA
JEC21_CDS  AGACAA----------------------------------------------------------GTTGAA
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

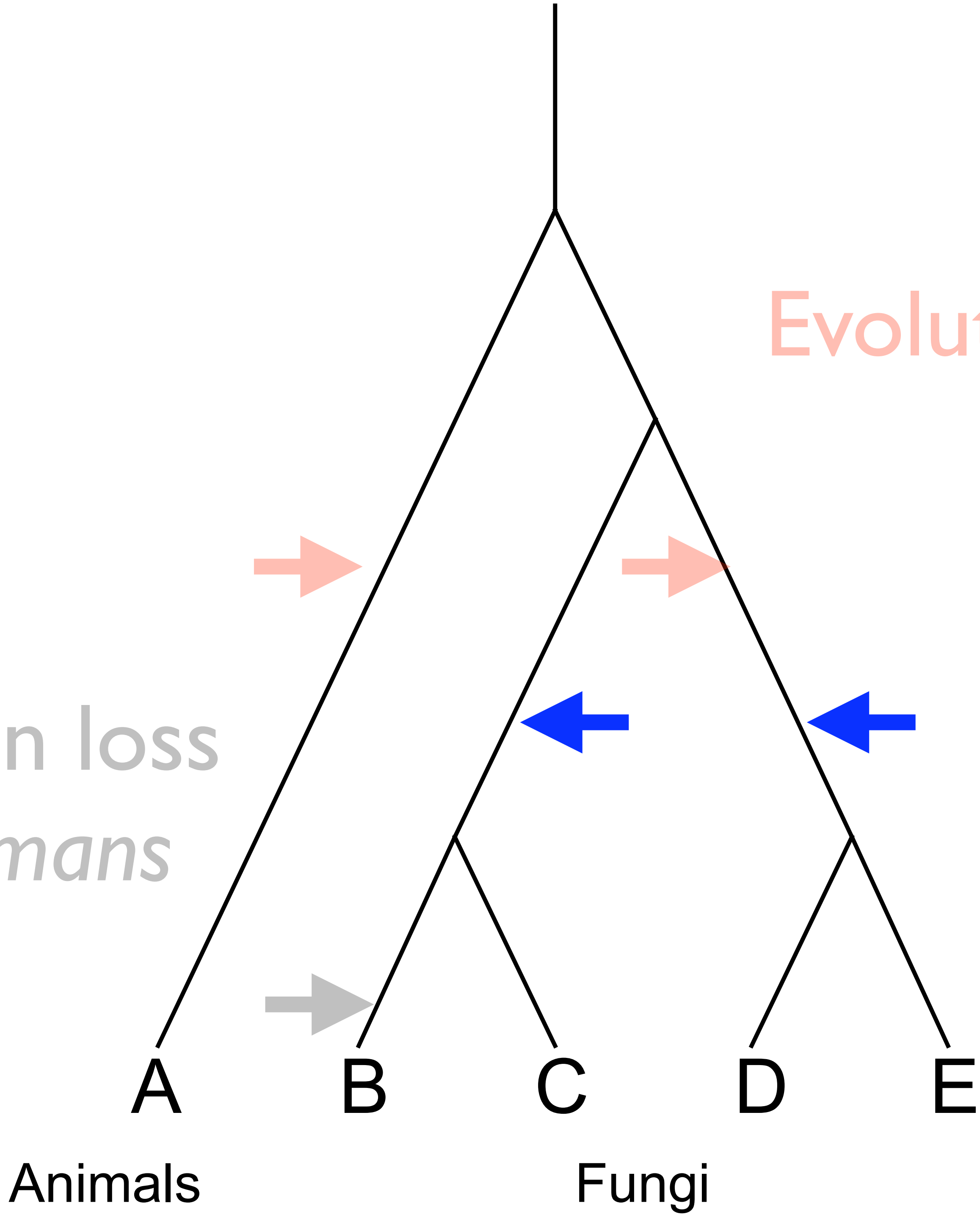


Stajich and Dietrich. Euk Cell *In press*.

Conclusions

- Intron loss via homologous recombination with spliced transcript
    - Large losses are all adjacent introns
    - Precise deletion
- Loss biased towards the middle of gene not 3’

Fungal comparative genomics



Evolution of fungal introns

Fungal gene family evolution

Recent intron loss in *C. neoformans*

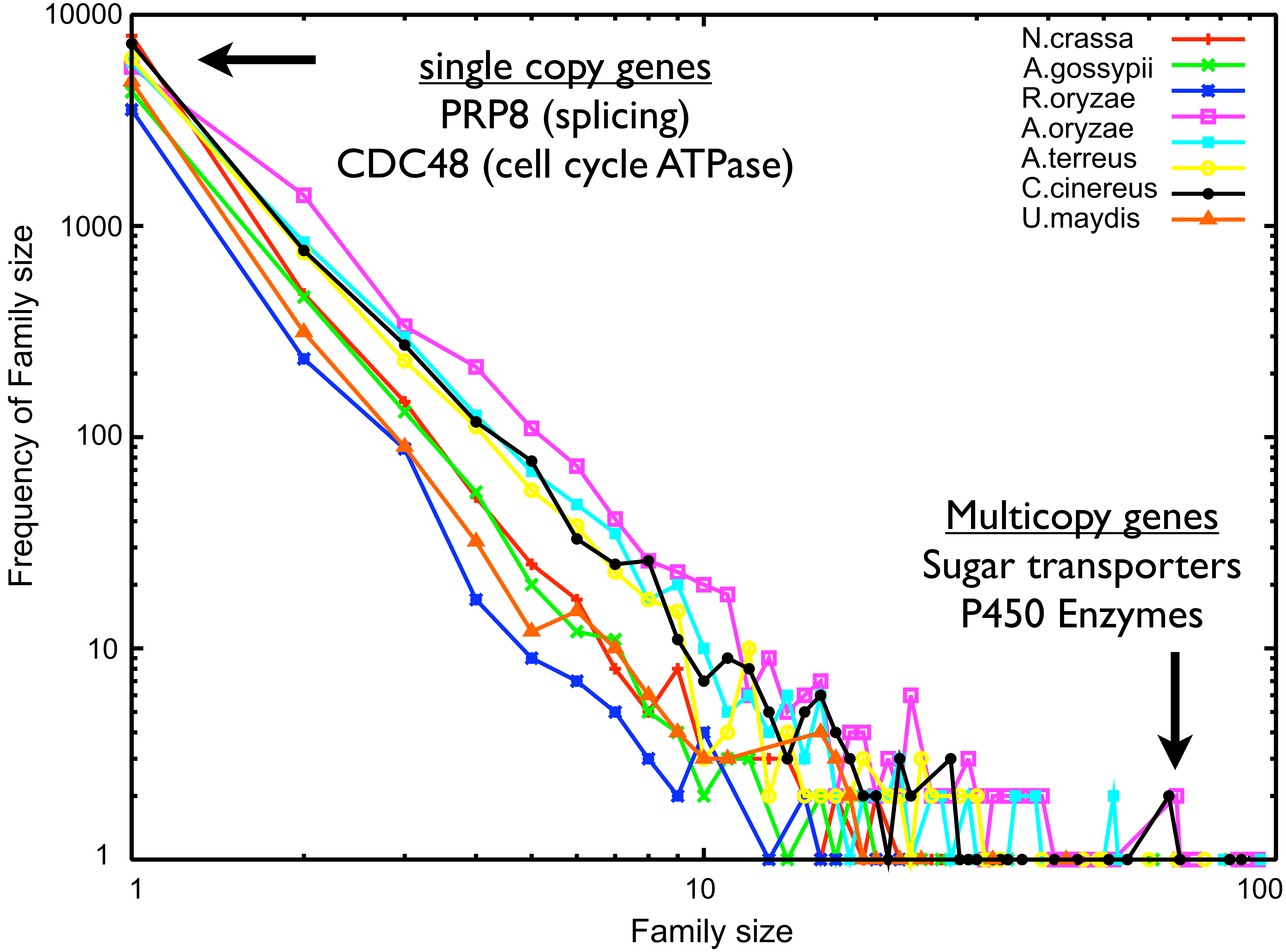
Gene family evolution

- 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 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



single copy genes

PRP8 (splicing)

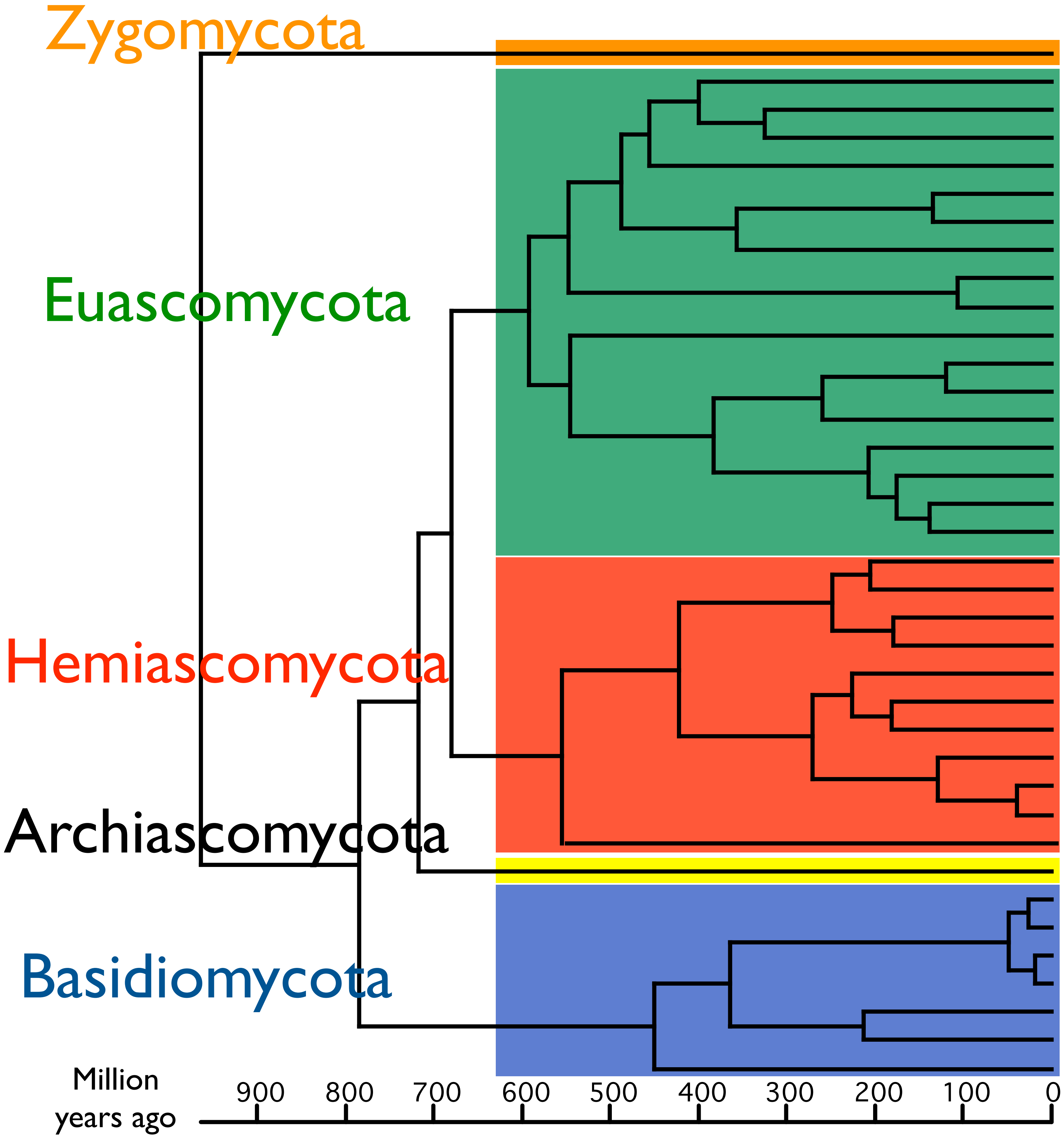
CDC48 (cell cycle ATPase)

Multicopy genes

Sugar transporters

P450 Enzymes

Fully sequenced fungal genomes

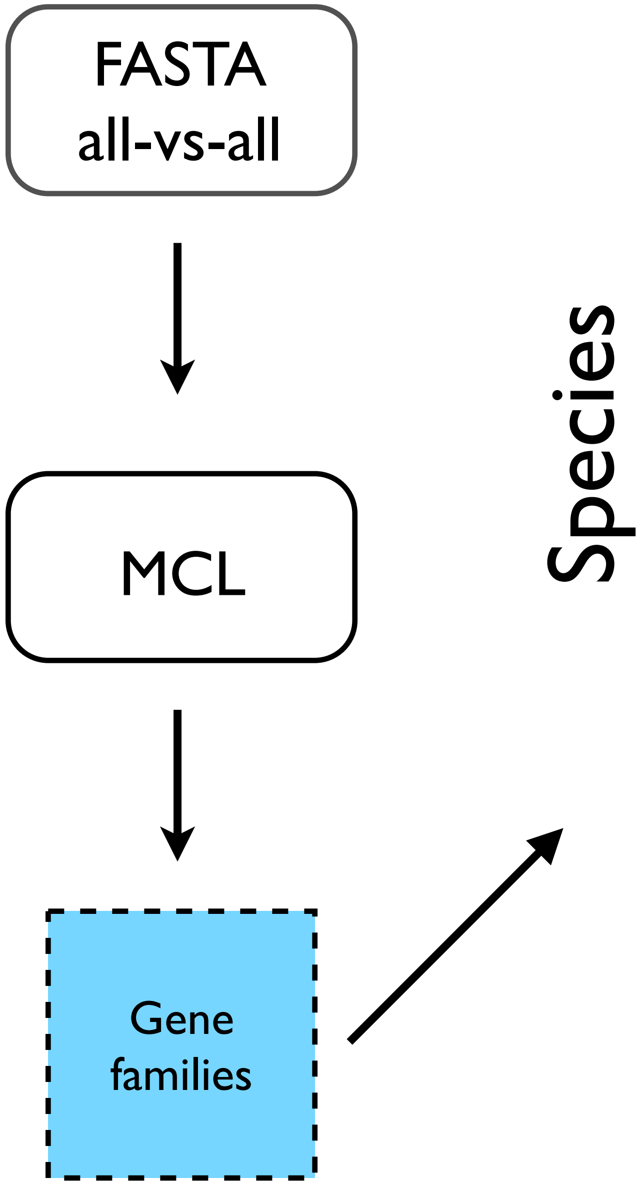


Rhizopus oryaze	Bread mold, Opp Hum pathogen
Neurospora crassa	Saprophyte
Podospora anserina	Saprophyte
Chaetomium globosum	Saprophyte Opp Hum Pathogen
Magnaporthe grisea	Hemibiotroph - Rice
Fusarium verticillioides	Hemibiotroph - maize
Fusarium graminearum	Hemibiotroph - wheat
Trichoderma reesei	Saprophyte
Sclerotinia sclerotiorum	Necrotroph
Botrytis cinerea	Necrotroph - fruits
Stagonospora nodorum	Hemibiotroph - wheat
Uncinocarpus reesii	
Coccidioides immitis	Primary Hum pathogen
Histoplasma capsulatum	Primary Hum pathogen
Aspergillus fumigatus	Opp Hum pathogen
Aspergillus nidulans	Saprophyte
Aspergillus terreus	Opp Hum pathogen
Aspergillus oryzae	Saprophyte/Industrial uses
Ashbya gosspyii	Biotroph/Industrial uses
Kluyveromyces lactis	Industrial uses
Saccharomyces cerevisiae	Industrial uses
Candida glabrata	Opp Hum pathogen
Candida lusitaniae	Opp Hum pathogen
Debaryomyces hansenii	
Candida guilliermondii	Opp Hum pathogen
Candida tropicalis	Opp Hum pathogen
Candida albicans	Opp Hum pathogen
Candida dubliniensis	Opp Hum pathogen
Yarrowia lipolytica	Industrial uses
Schizosaccharomyces pombe	
Cryptococcus neoformans	Opp Hum pathogen
Cryptoccocus neoformans H99	Opp Hum pathogen
Cryptococcus gattii WM276	Opp Hum pathogen
Cryptococcus gattii R265	Opp Hum pathogen
Phanerochaete chrysosporium	Saprophyte
Coprinus cinereus	Saprophyte
Ustilago maydis	Biotroph - maize

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 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*



FASTA all-vs-all

MCL

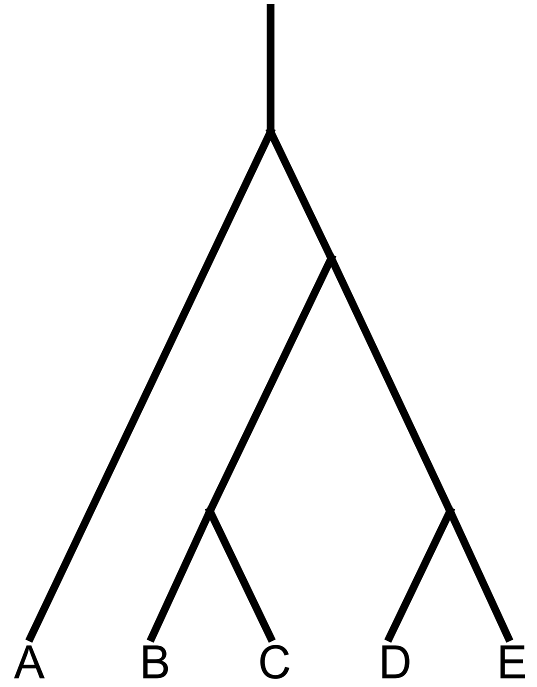
Gene families

Family count

Species

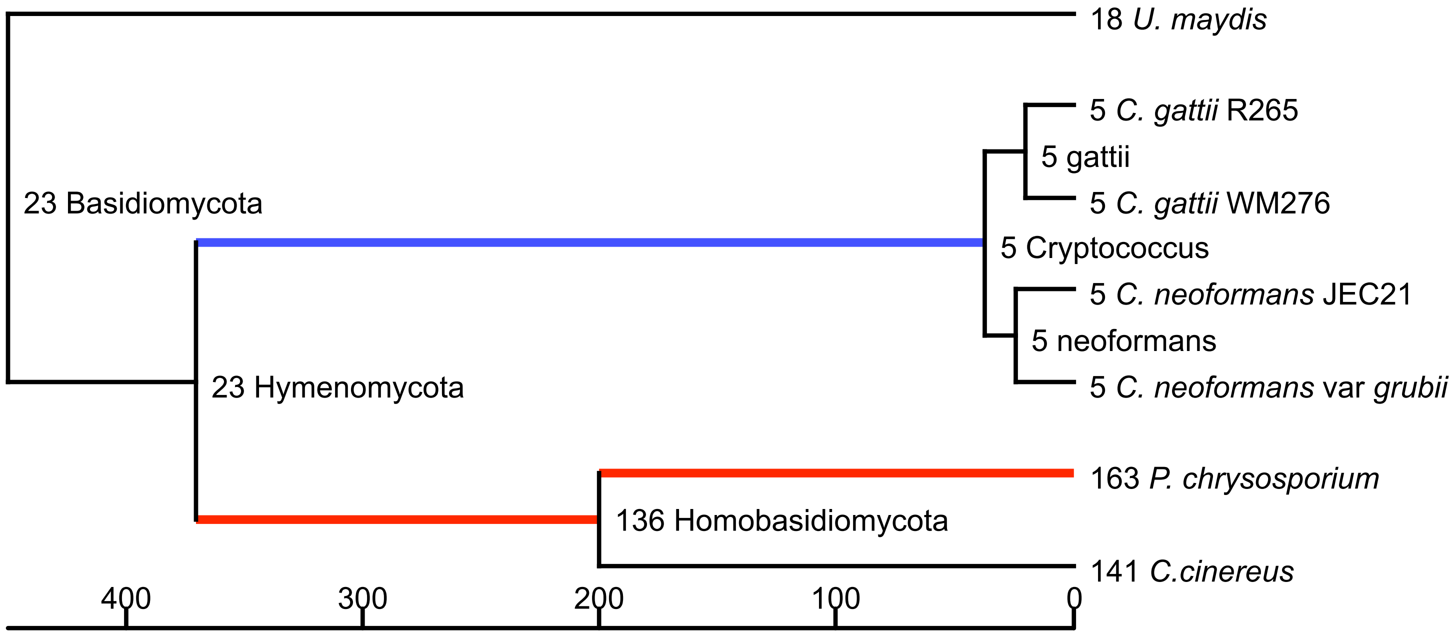
🟧	10	1	2
🟦	14	18	2
🟨	7	1	1
🟩	6	1	12
🟩	6	1	8
🟥	3	1	1

+



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



Families with significant expansions

Transporters
Kinases
P450
Oxidation

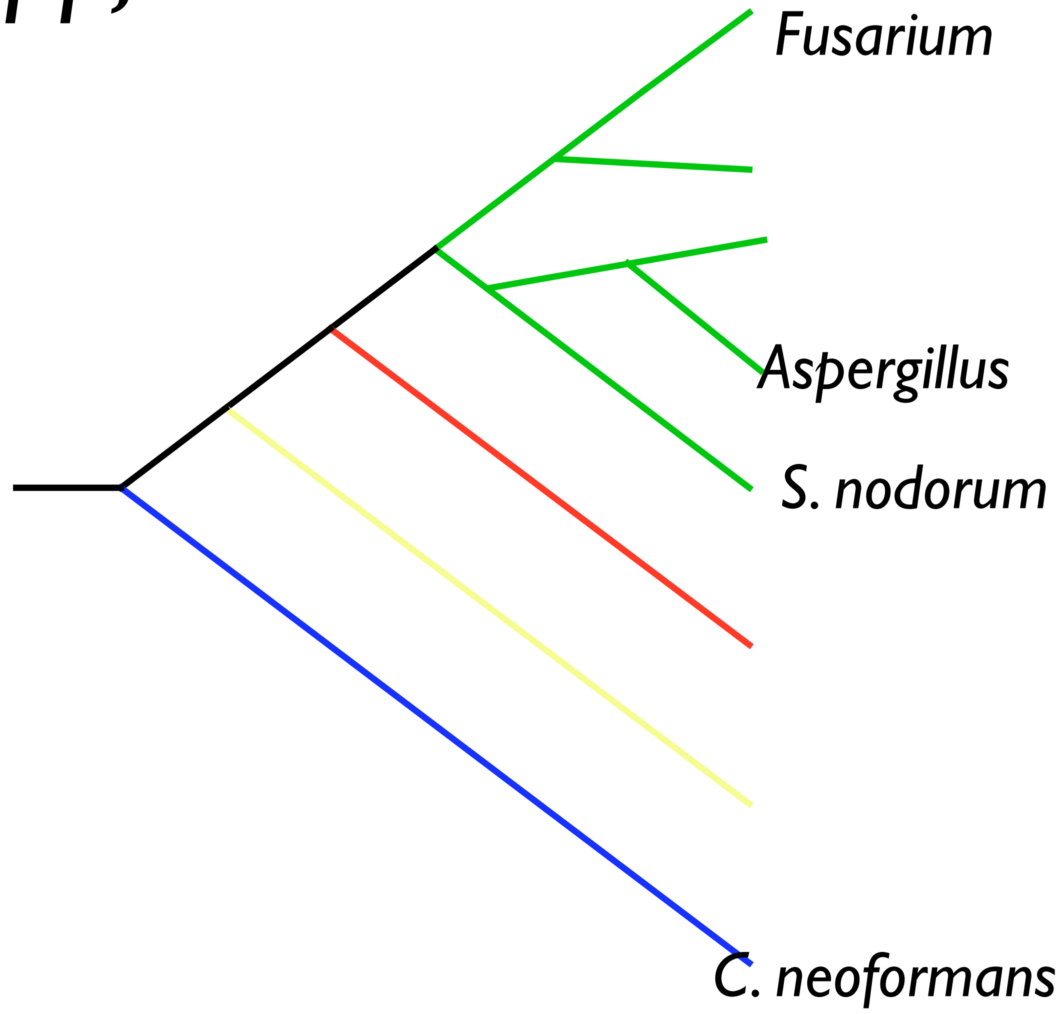
Vitamin & Cofactor transport	Methytransferase
Lactose & sugar transport	Cytochrome P450: CYP64
Amine transport	Cytochrome P450: CYP53,57A
Myo-instol, quinate, and glucose transport	Cytochrome P450
Oligopeptide transport	Kinase
ABC transporter	Subtilase family
MFS, drug pump, & sugar transport	NADH flavin 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)

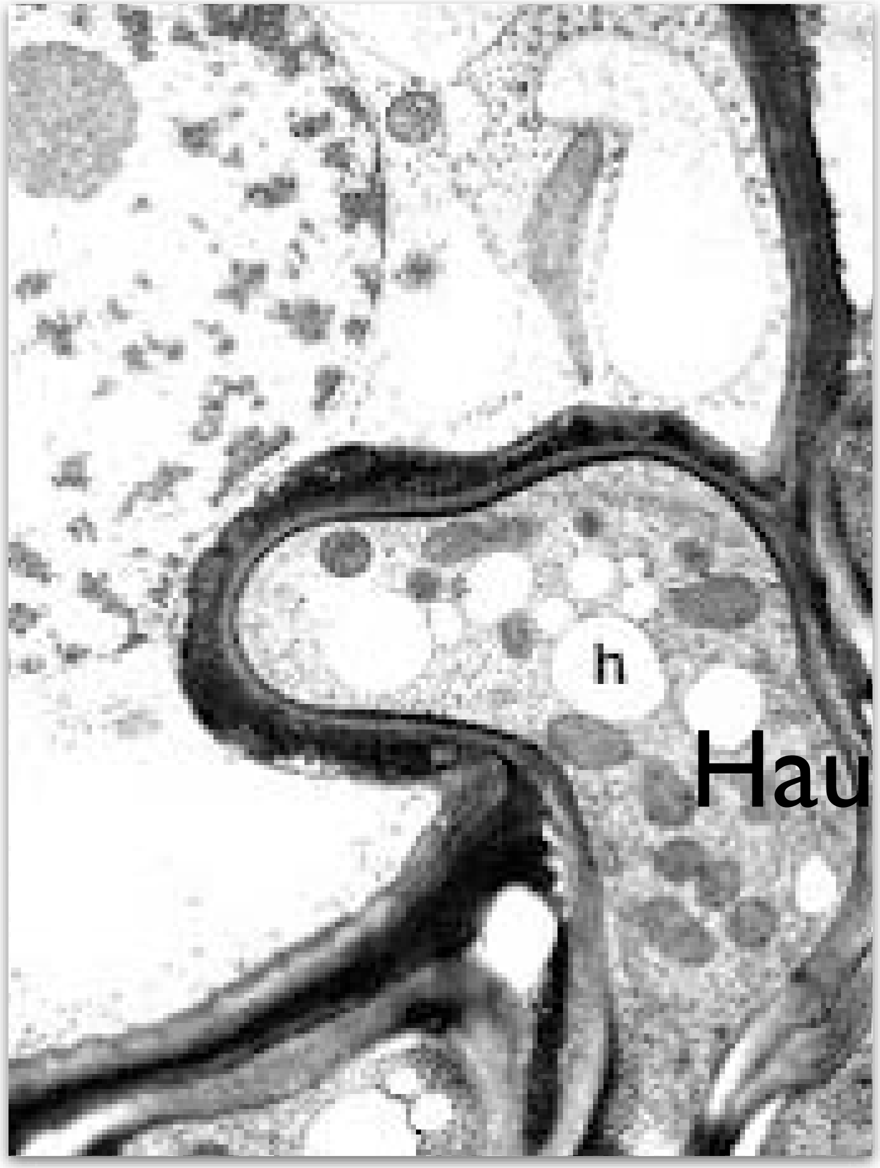
Branches with transporter expansions

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



What do phytopathogens use transporters for?

- 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

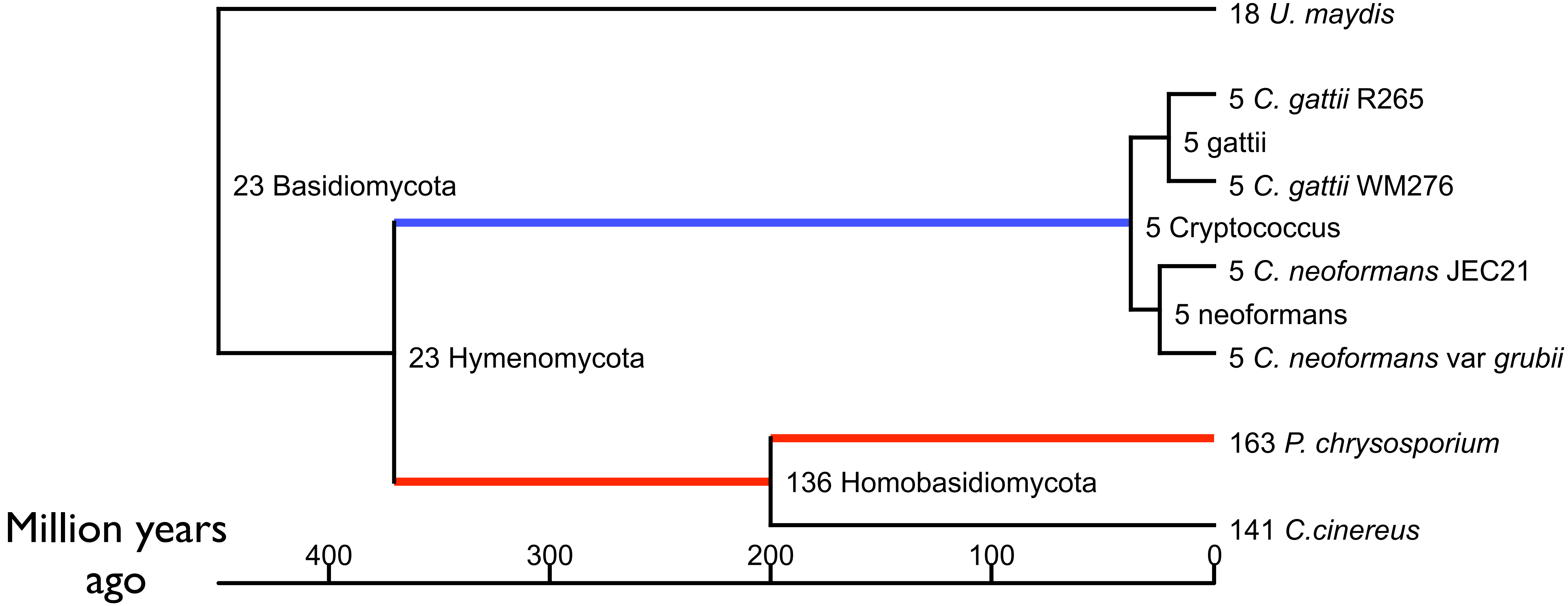
- 3x as many sugar transporters in *C. neoformans* than other basidiomycetes
- "sugar coated killer"
- Capsule is mixture of glucose, xylose, and mannose.

Zerpa et al, 1996

P450 CYP64

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

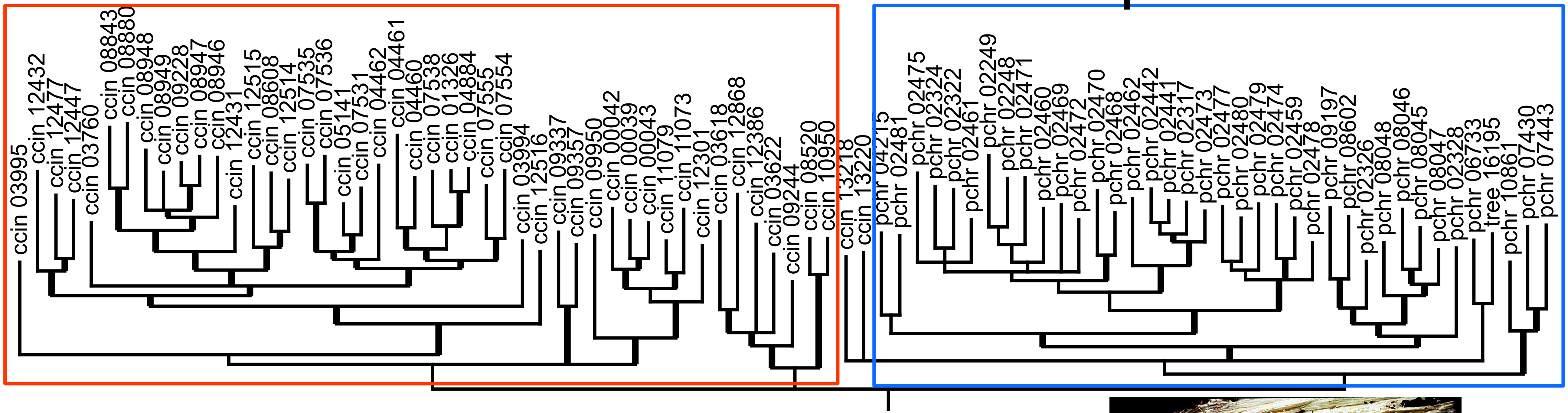
CYP64: Step in *Aspergillus* spp aflatoxin pathway
P. chrysosporium implicated in lignin and hydrocarbon degradation.



CYP64 was from independent duplication

C. cinereus expansion

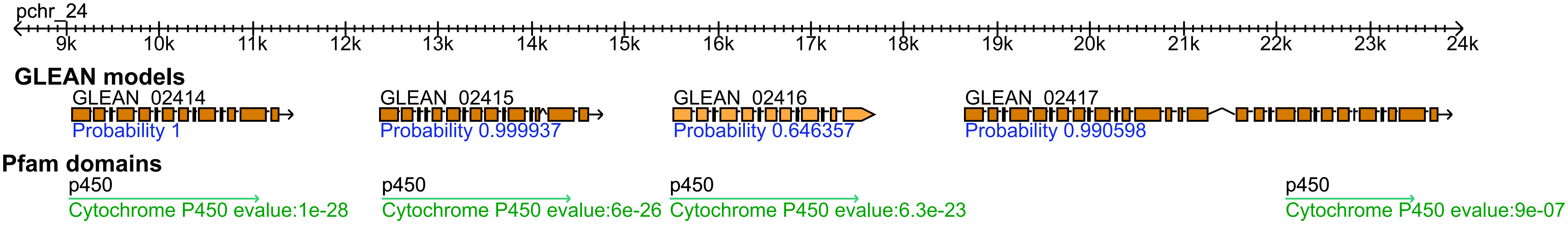
P. chrysosporium expansion



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

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?

Overall conclusions

- Multiple genome sequences have helped resolve several outstanding questions in evolution introns
- Gene family expansions can be important in identifying molecular basis for adaptation

Future directions

- UC Berkeley with John Taylor
- Adaptation and speciation in fungi
- Focus on pathogenic fungus *Coccidioides*
    - Signatures of adaptation among genomes of 12 sequenced strains

Acknowledgments

Fred Dietrich

Greg Wray
Lincoln Stein
John McCusker
Alex Hartemink

Joseph Heitman
Marcy Uyenoyama
Rytas Vilglyas
Tim James
Pat Pukkilla

Matthew Hahn
DUMRU
John Perfect
Andy Alspaugh
Tom Mitchell

Ewan Birney

<u>Sequencing centers</u>
Broad Institute
Duke University
Joint Genome Institute
Génolevures
Stanford University
(NIH and NSF)

UPGG NSF

CSEM

Robert Cramer
James Fraser
Steven Giles
Alex Idnurm
Scott Roy

Dave Des Marais
Heath O'Brien
Matt Rockman
MDG
<u>Dietrich Lab</u>
Andria Allen
Stephanie Diezmann
Charles Hall
Shan Huang
Philippe Lüdi
Laura Kavanaugh

Sandra Reynolds
Mark DeLong