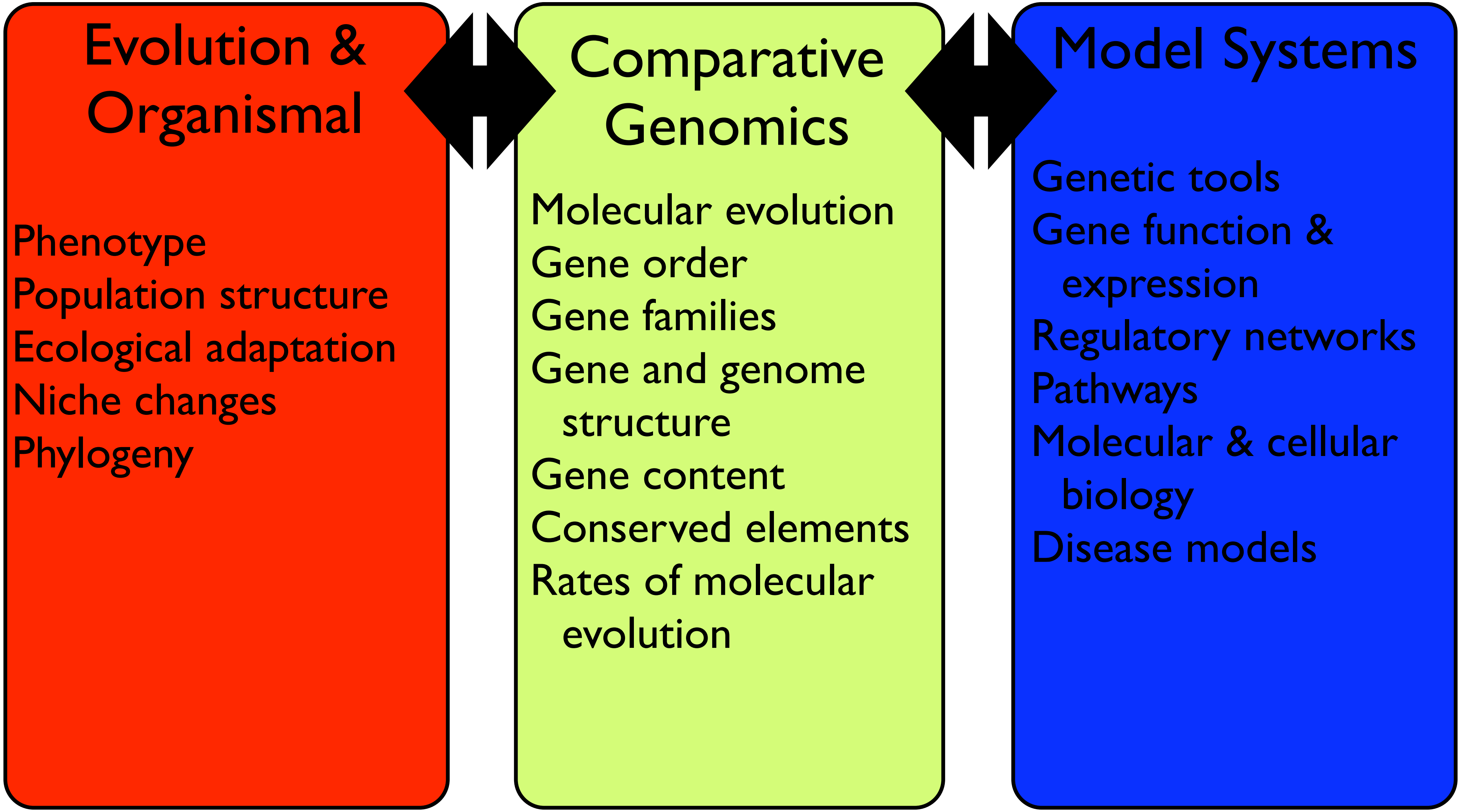
Intron evolution in Fungi

Jason Stajich
Taylor Lab

Evolutionary genomics



Evolution & Organismal

- Phenotype
- Population structure
- Ecological adaptation
- Niche changes
- Phylogeny

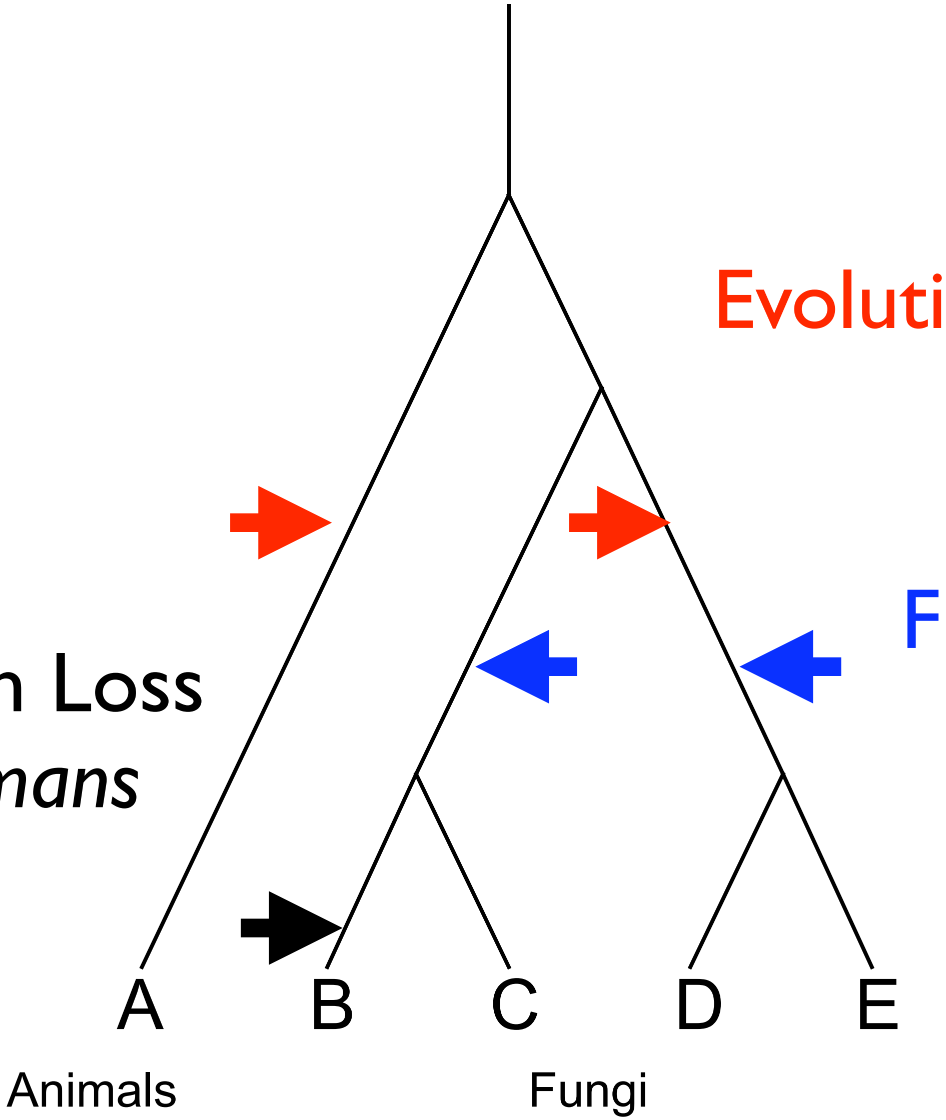
Comparative Genomics

- 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

Fungal comparative genomics

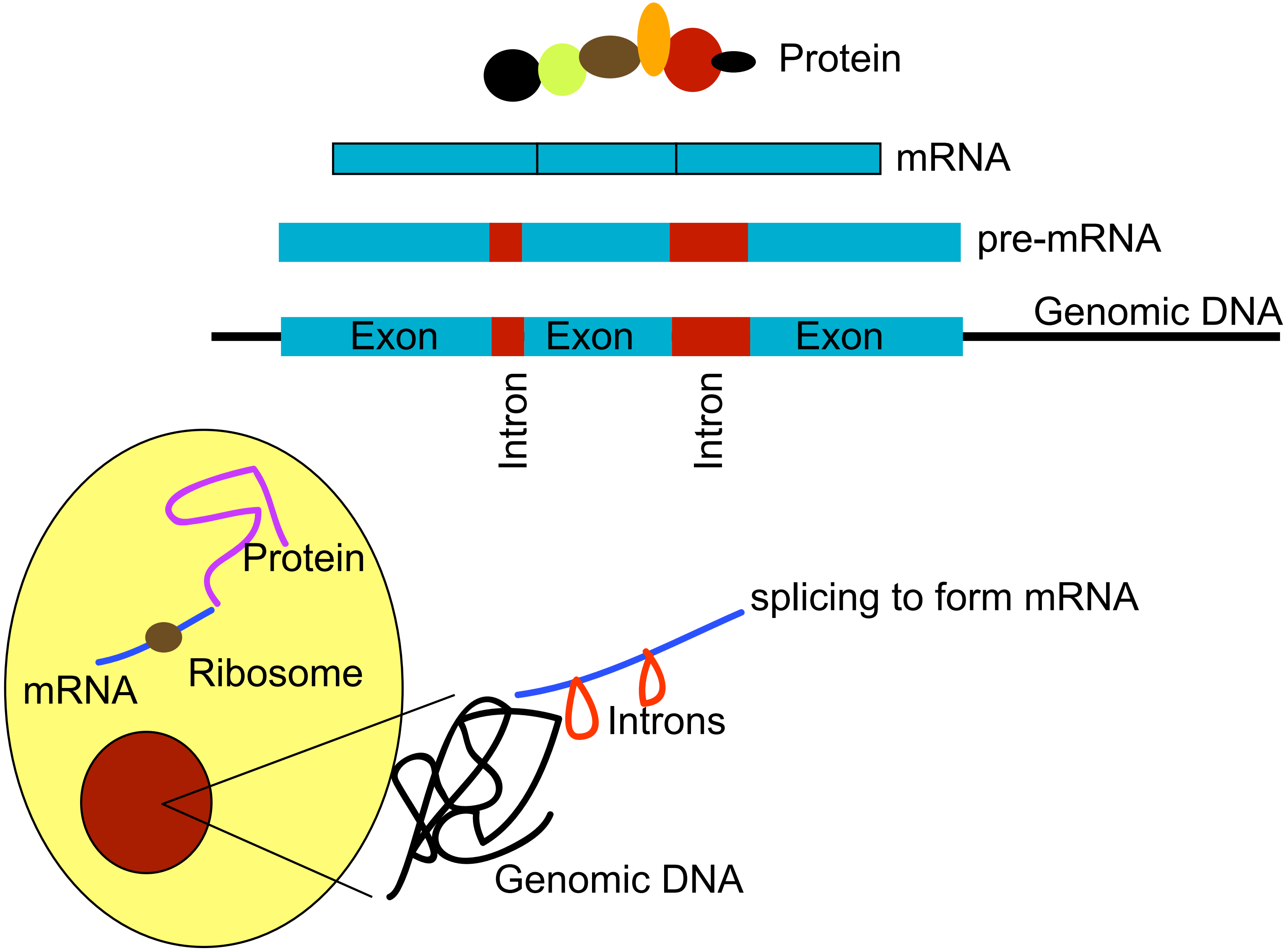


Evolution of fungal introns

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

mRNA

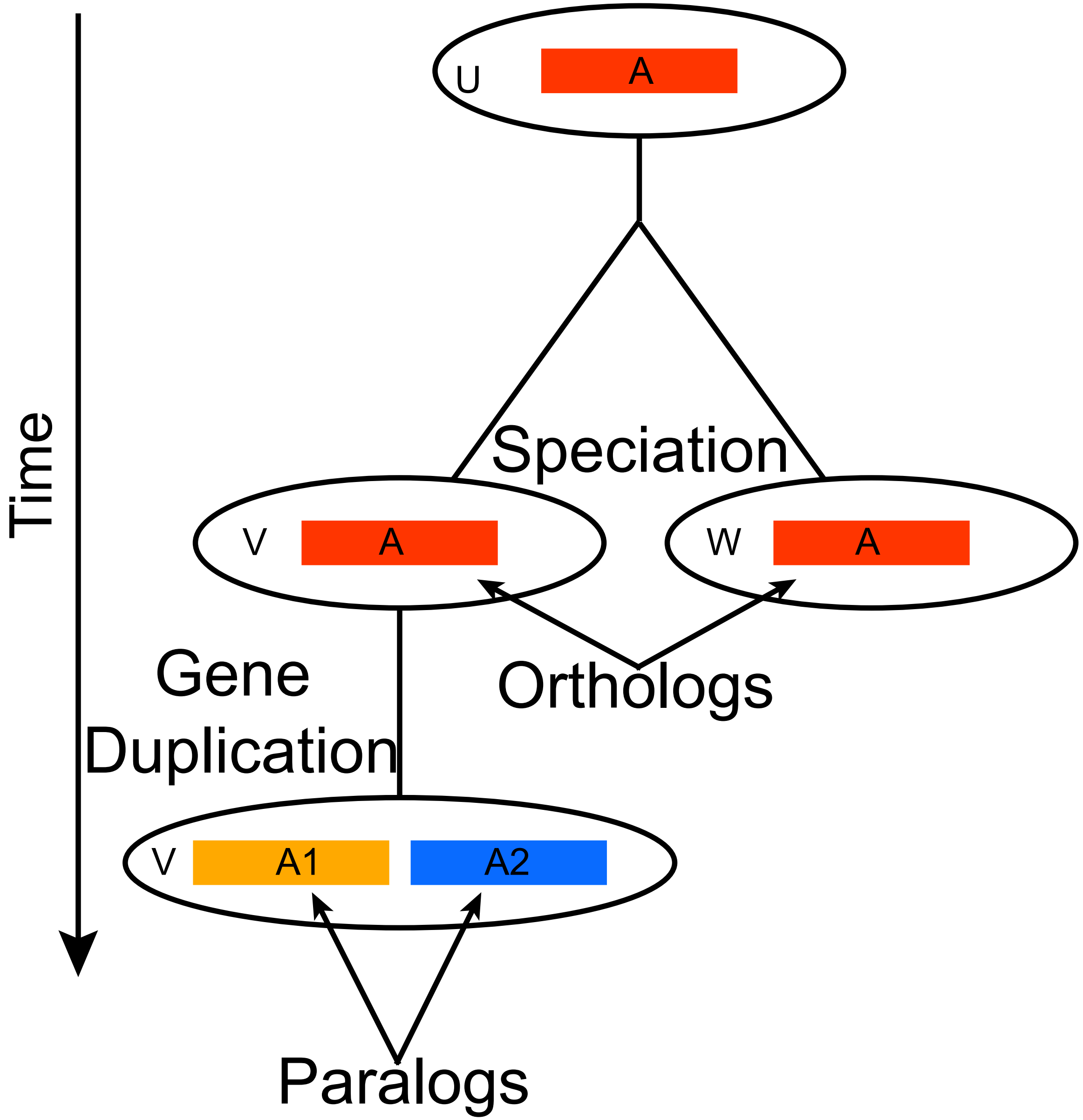
Ribosome

splicing to form mRNA

Introns

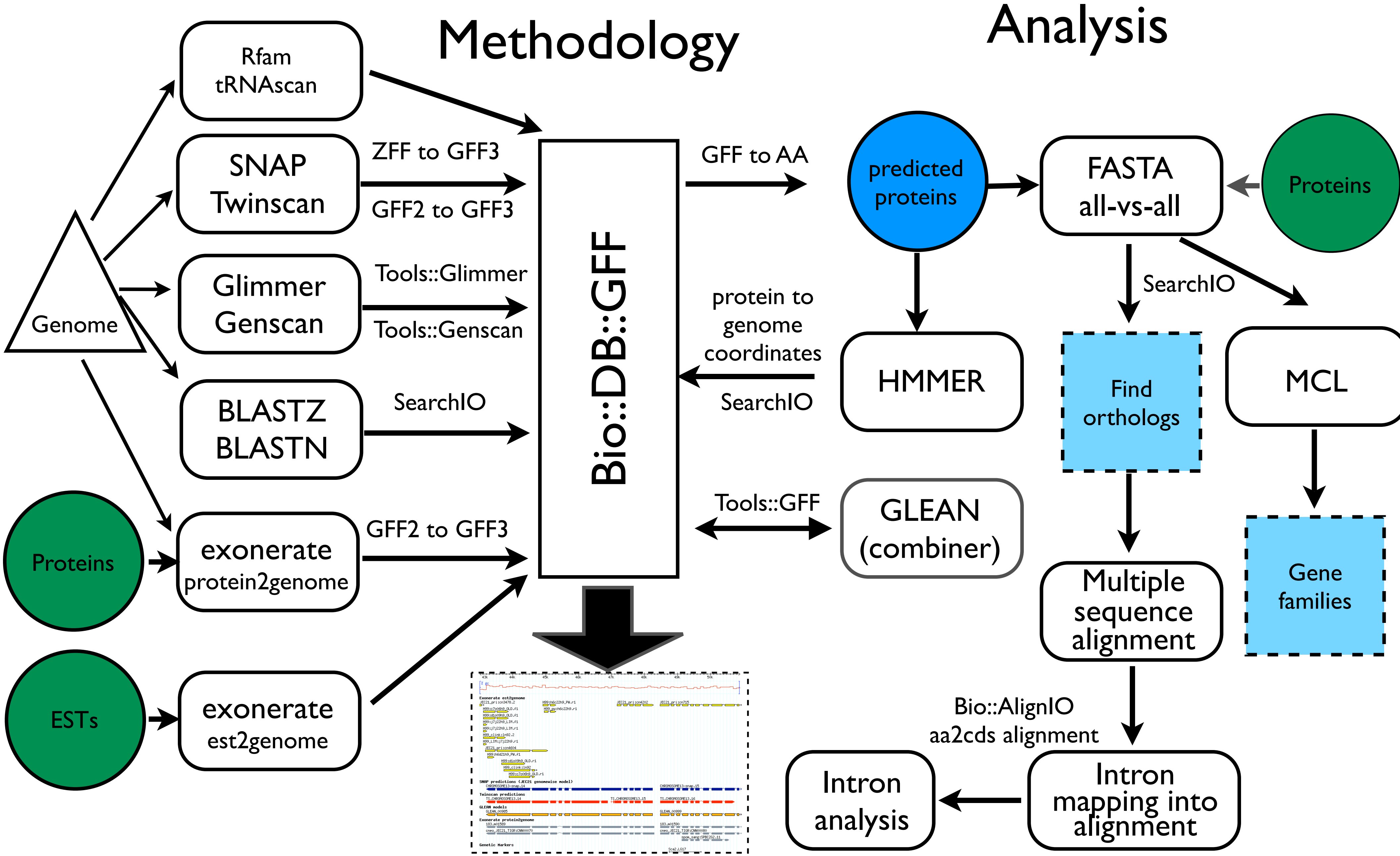
Genomic DNA

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

predicted proteins

FASTA
all-vs-all

Proteins

protein to
genome
coordinates

SearchIO

HMMER

SearchIO

Find
orthologs

MCL

Tools::GFF

GLEAN
(combiner)

Multiple
sequence
alignment

Gene
families

Bio::AlignIO
aa2cds alignment

Intron
analysis

Intron
mapping into
alignment

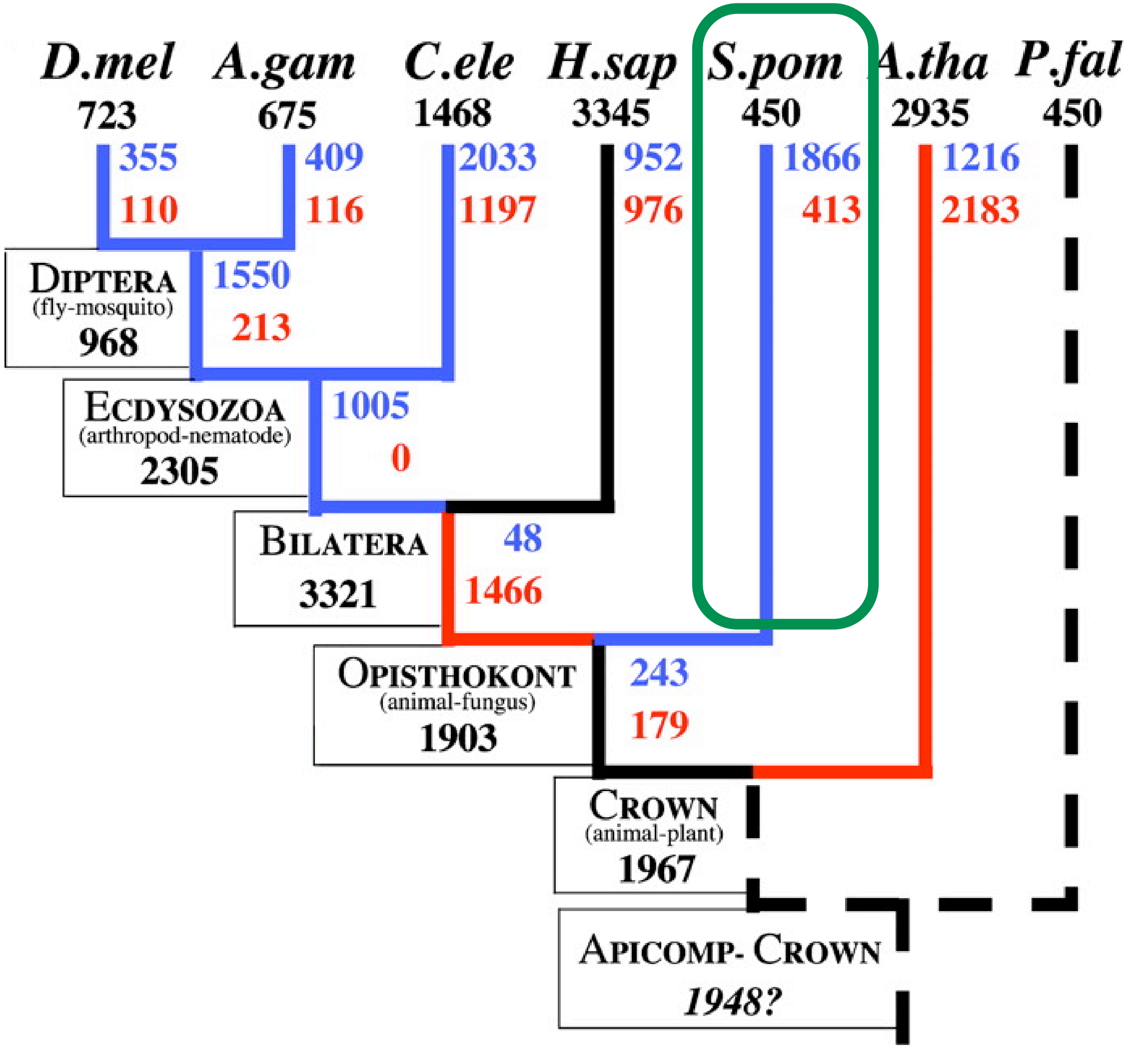
http://fungal.genome.duke.edu

Evolution of gene structure

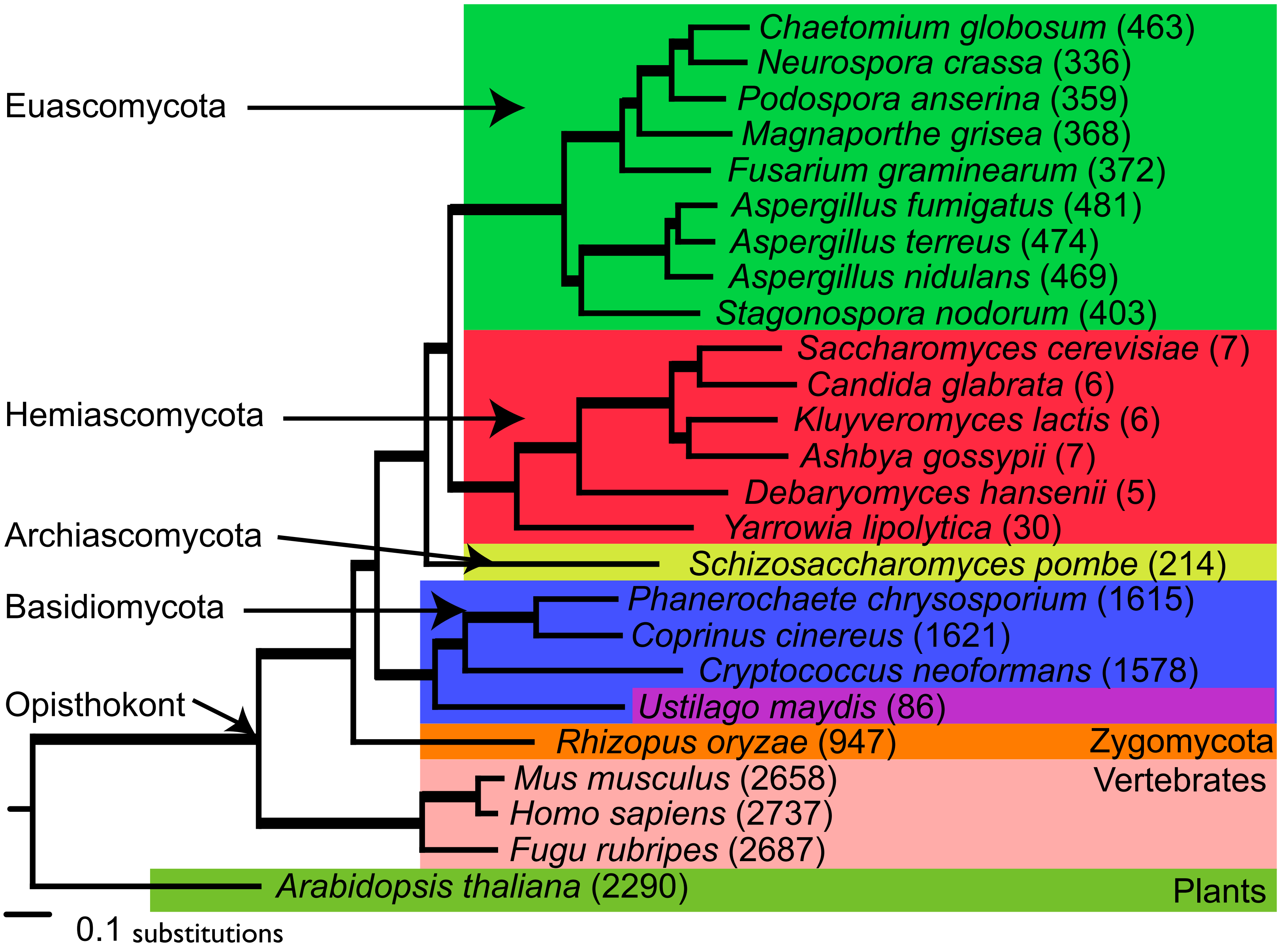
- Present day introns
  - Recent insertions?
    - Introns late hypothesis
  - Formed in eukaryotic ancestor?
    - Introns early hypothesis / exon theory of genes
  - Mixture of two?

Previous work on intron evolution

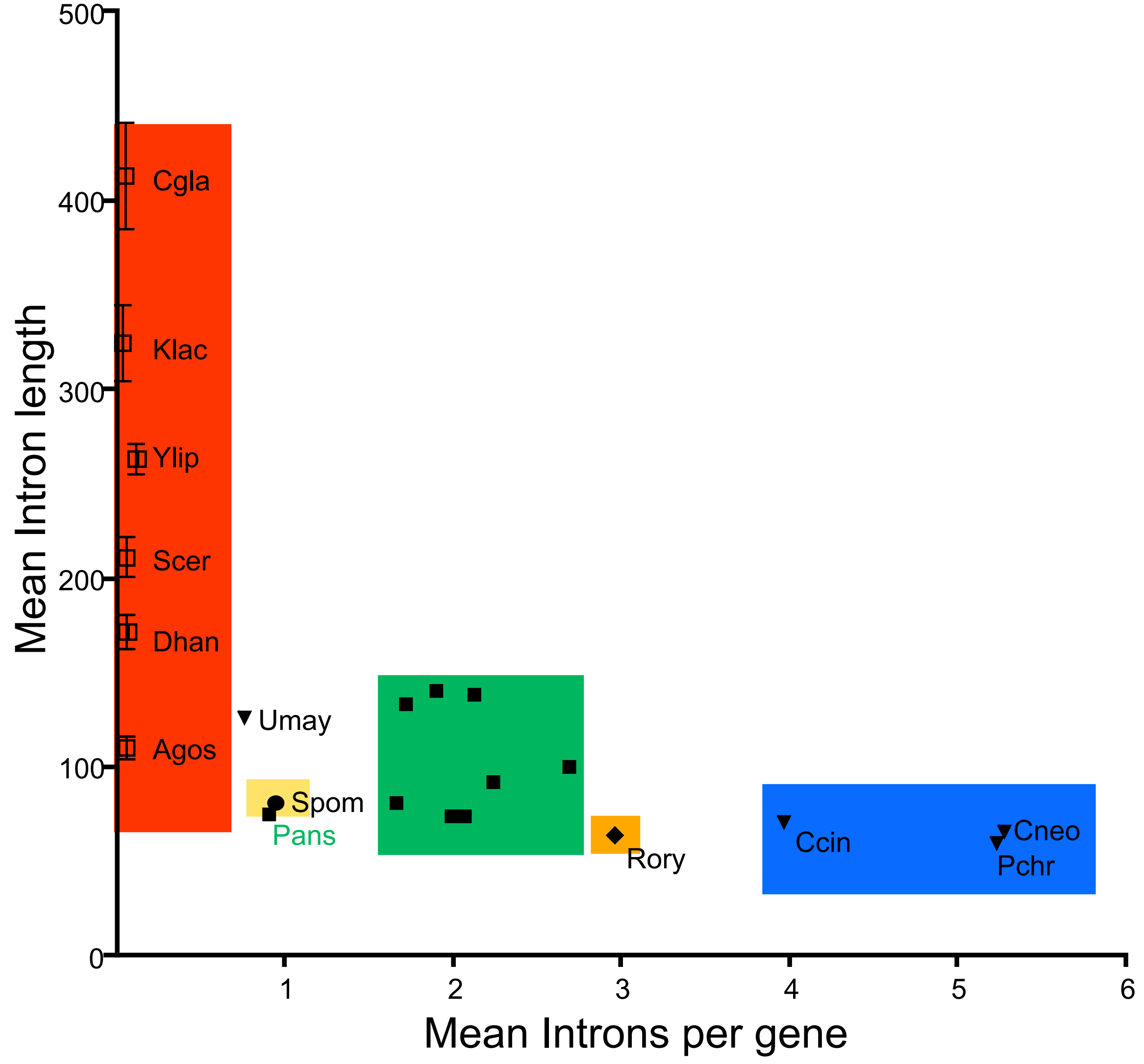
- Rogozin et al. 2003
  - 7 genomes
  - 684 genes, 7236 positions
- Other methods
  - Roy and Gilbert. 2005
  - Csũrös. 2005
  - Nguyen et al. 2006



Calculating intron densities across a phylogeny



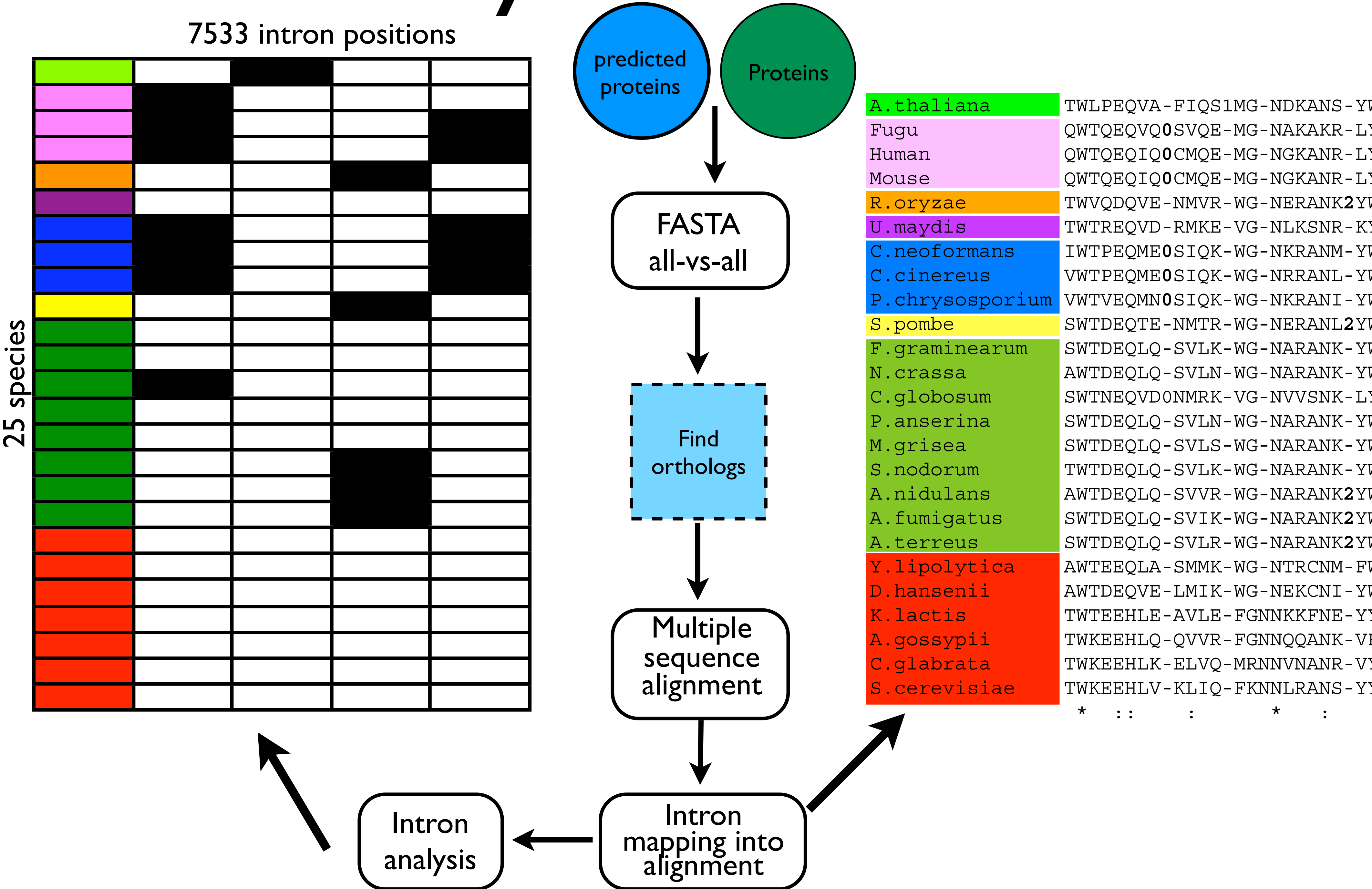
Intron frequency varies among the fungi



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



7533 intron positions

25 species

predicted proteins

Proteins

FASTA all-vs-all

Find orthologs

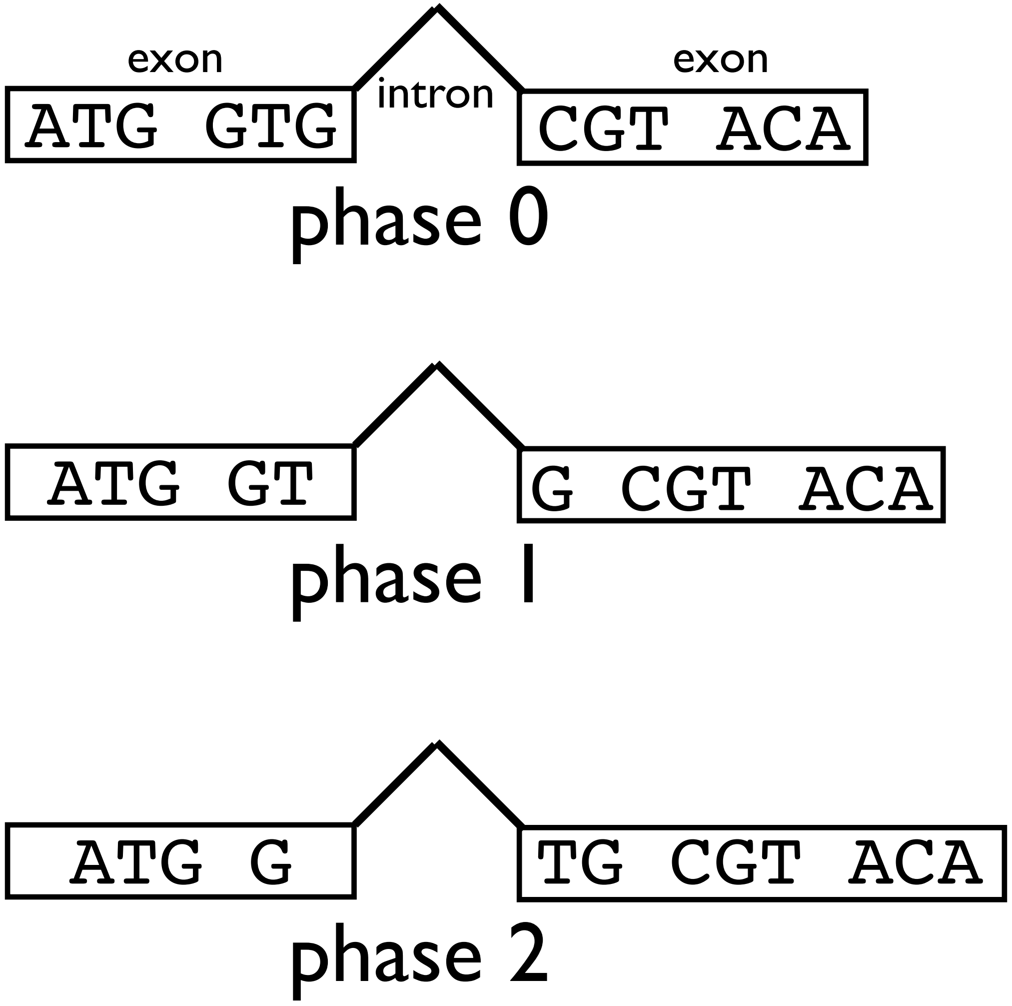
Multiple sequence alignment

Intron mapping into alignment

Intron analysis

```
A.thaliana      TWLPEQVA-FIQS1MG-NDKANS-Y
Fugu            QWTQEQVQ0SVQE-MG-NAKAKR-L
Human           QWTQEQIQ0CMQE-MG-NGKANR-L
Mouse           QWTQEQIQ0CMQE-MG-NGKANR-L
R.oryzae        TWVQDQVE-NMVR-WG-NERANK2Y
U.maydis        TWTREQVD-RMKE-VG-NLKSNR-K
C.neoformans    IWTPEQME0SIQK-WG-NKRANM-Y
C.cinereus      VWTPEQME0SIQK-WG-NRRANL-Y
P.chrysosporium VWTVEQMN0SIQK-WG-NKRANI-Y
S.pombe         SWTDEQTE-NMTR-WG-NERANL2Y
F.graminearum   SWTDEQLQ-SVLK-WG-NARANK-Y
N.crassa        AWTDEQLQ-SVLN-WG-NARANK-Y
C.globosum      SWTNEQVD0NMRK-VG-NVVSNK-L
P.anserina      SWTDEQLQ-SVLN-WG-NARANK-Y
M.grisea        SWTDEQLQ-SVLS-WG-NARANK-Y
S.nodorum       TWTDEQLQ-SVLK-WG-NARANK-Y
A.nidulans      AWTDEQLQ-SVVR-WG-NARANK2Y
A.fumigatus     SWTDEQLQ-SVIK-WG-NARANK2Y
A.terreus       SWTDEQLQ-SVLR-WG-NARANK2Y
Y.lipolytica    AWTEEQLA-SMMK-WG-NTRCNM-F
D.hansenii      AWTDEQVE-LMIK-WG-NEKCNI-Y
K.lactis        TWTEEHLE-AVLE-FGNNKKFNE-Y
A.gossypii      TWKEEHLQ-QVVR-FGNNQQANK-V
C.glabrata      TWKEEHLK-ELVQ-MRNNVNANR-V
S.cerevisiae    TWKEEHLV-KLIQ-FKNNLRANS-Y
                 *  ::    :      *   :
```

Intron phase



exon

ATG GTG

intron

exon

CGT ACA

phase 0

ATG GT

G CGT ACA

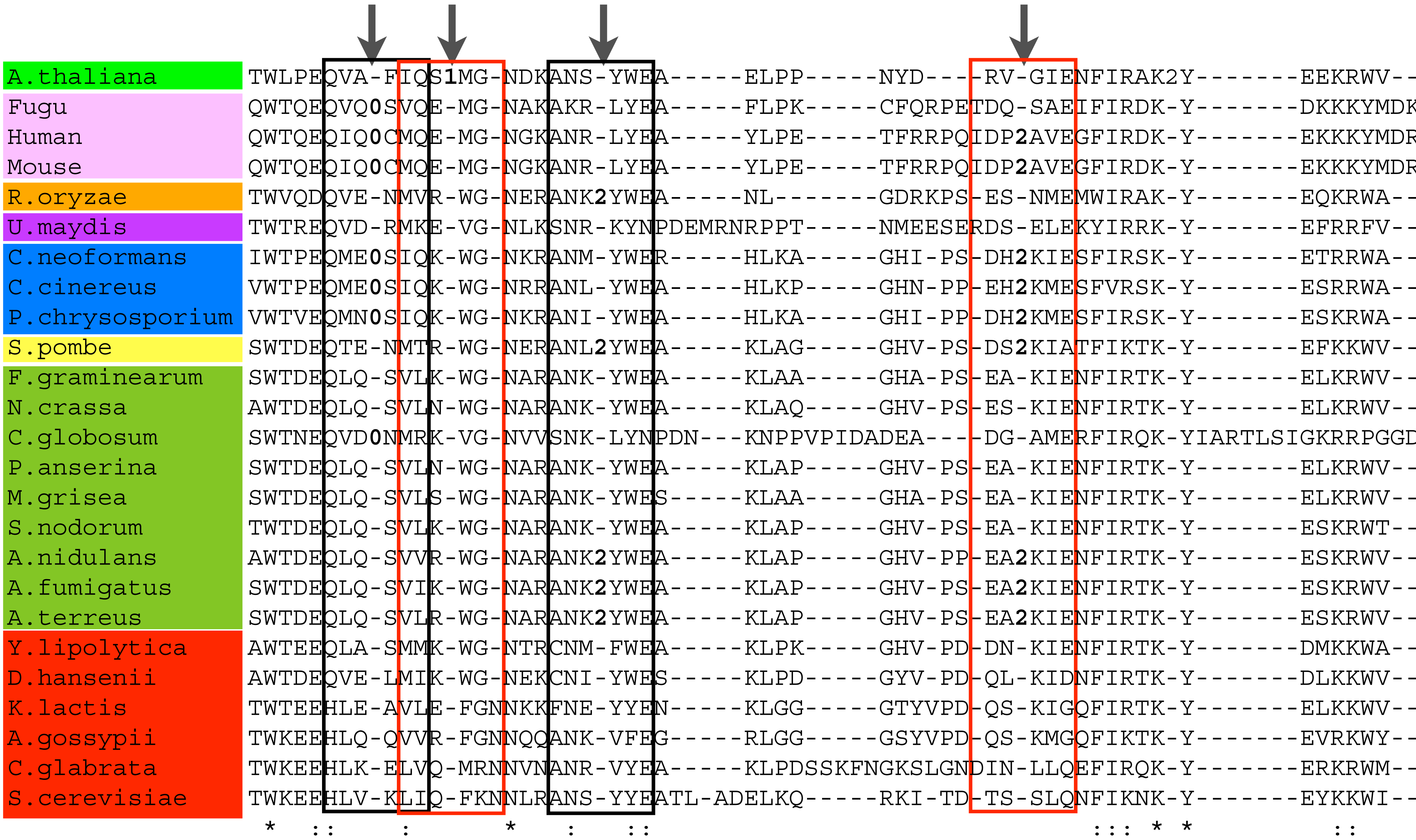
phase 1

ATG G

TG CGT ACA

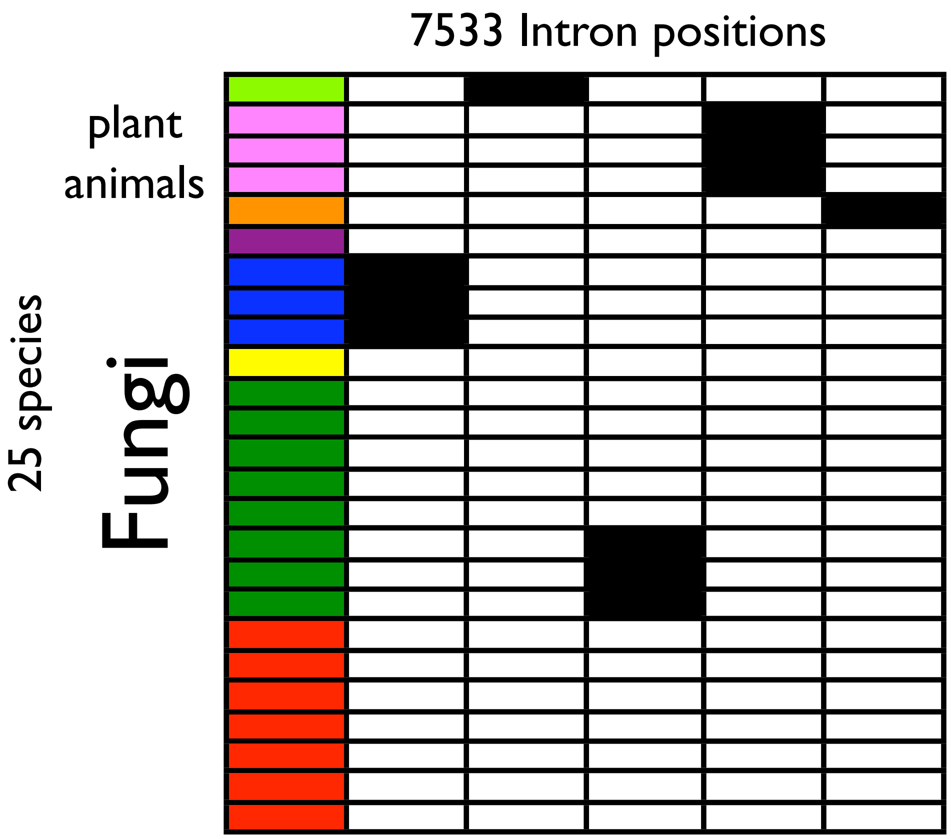
phase 2

Conserved intron positions

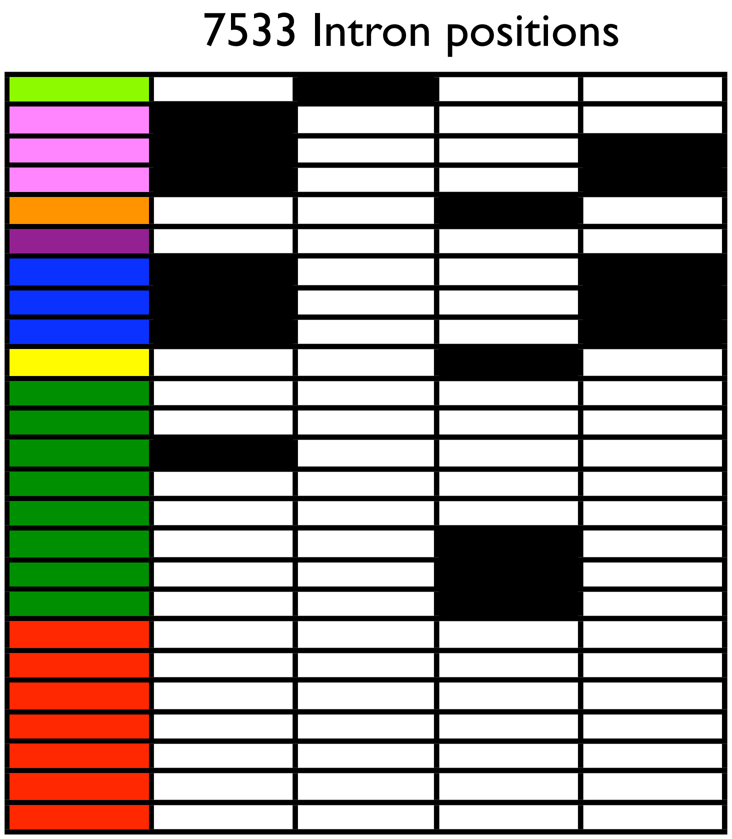


```
A.thaliana       TWLPEQVA-FIQS1MG-NDKANS-YWEA-----ELPP-----NYD----RV-GIENFIRAK2Y-------EEKRWV--
Fugu             QWTQEQVQ0SVQE-MG-NAKAKR-LYEA-----FLPK-----CFQRPETDQ-SAEIFIRDK-Y-------DKKKYMDK
Human            QWTQEQIQ0CMQE-MG-NGKANR-LYEA-----YLPE-----TFRRPQIDP2AVEGFIRDK-Y-------EKKKYMDR
Mouse   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

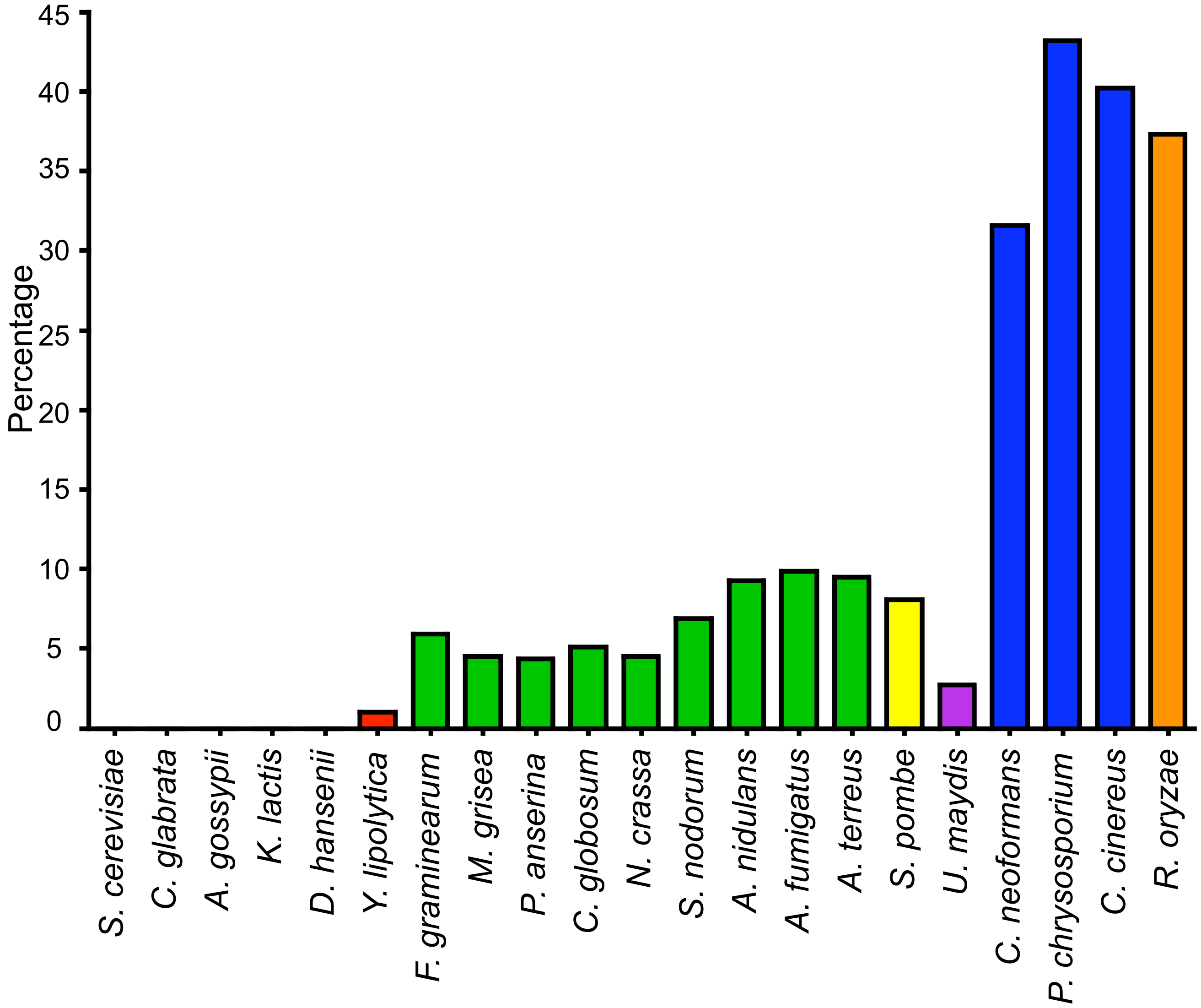


Introns late



Introns early(ier)

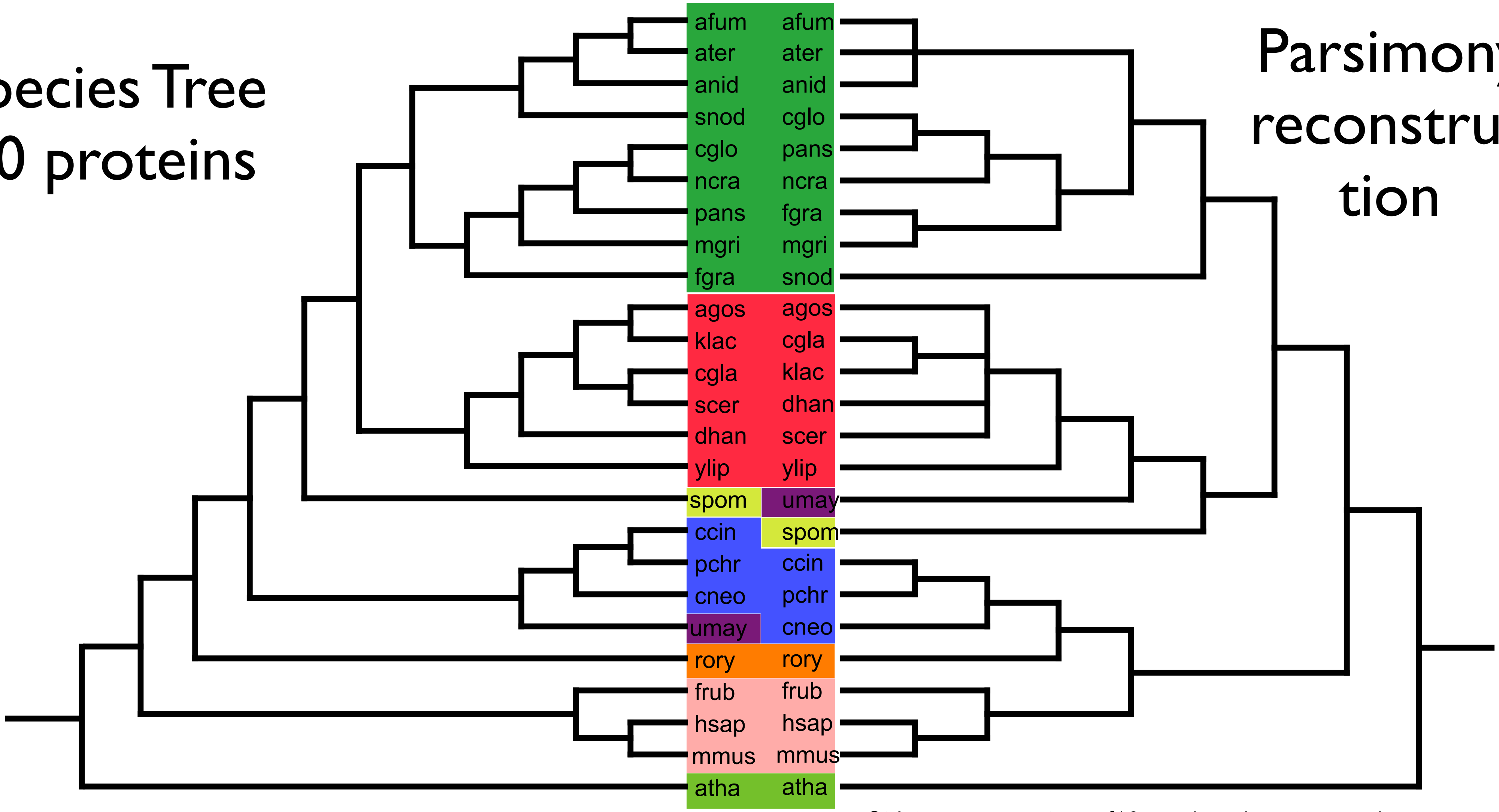
Intron positions shared with animals or plants



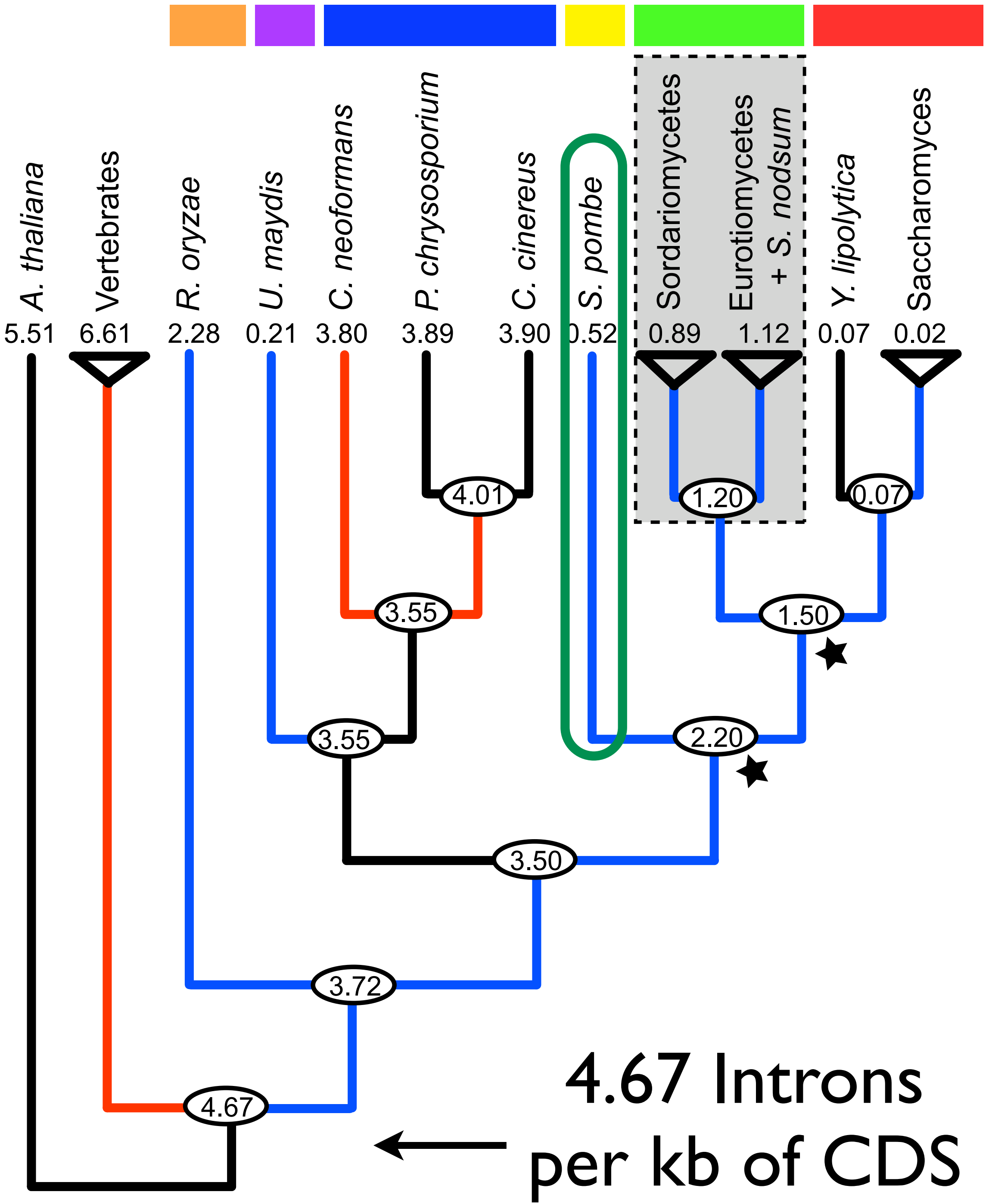
Phylogenetic signal in intron positions

Species Tree
30 proteins

Parsimony
reconstruc
tion



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.

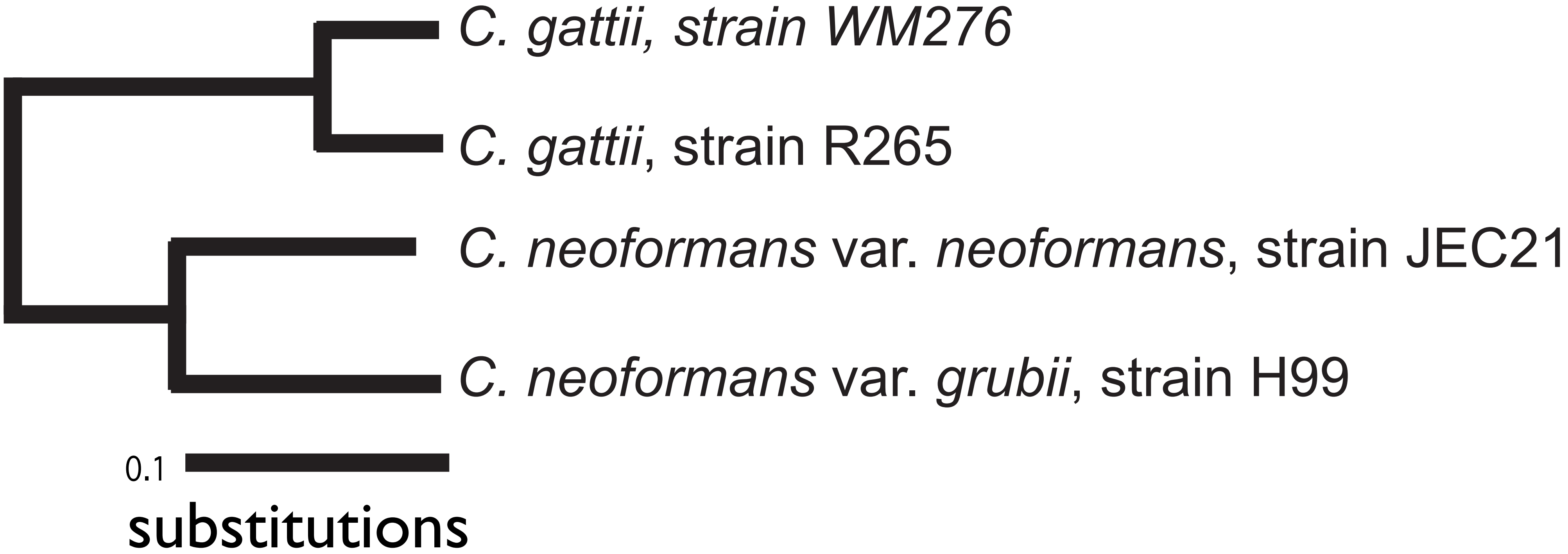
Mechanism of intron loss

- *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

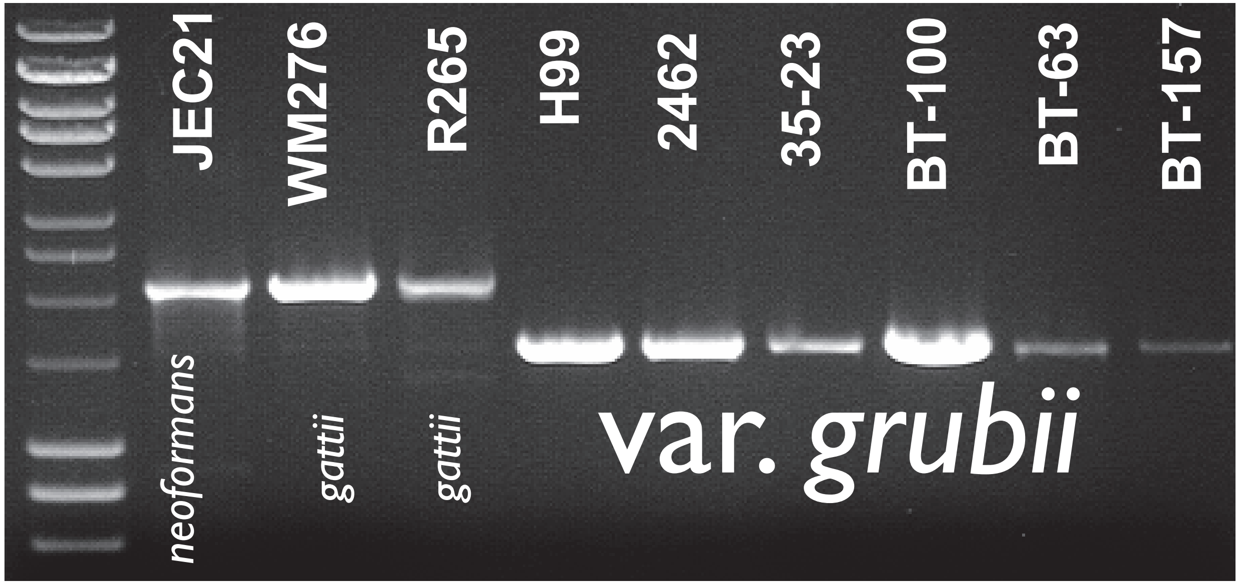
substitutions

Screen for intron changes

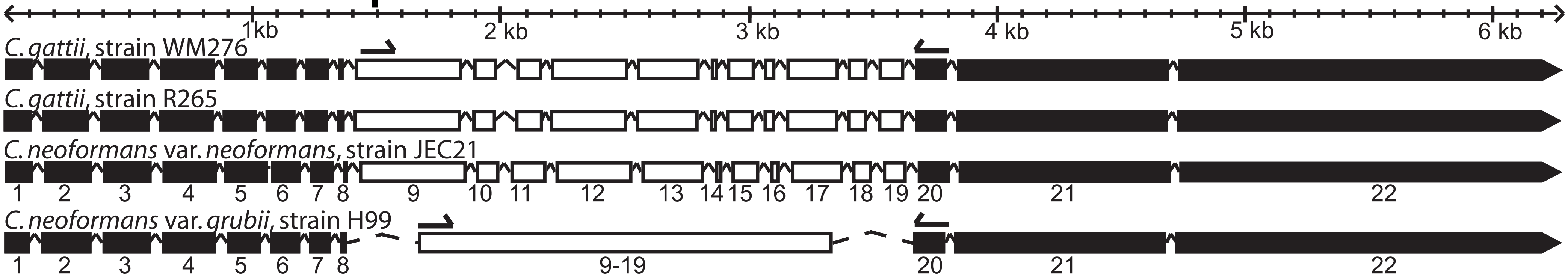
- Annotate 3 *Cryptococcus* genomes (var. *grubii* and 2 var. *gattii* genomes)
- Identify and align 4-way orthologous genes
  - 5298 orthologous genes (out of ~6500)
- Identify intron position changes

Stajich and Dietrich. Euk Cell 2006.

Intron loss in var. *grubii*



CNI01550 - putative RNA helicase



Stajich and Dietrich. Euk Cell 2006.

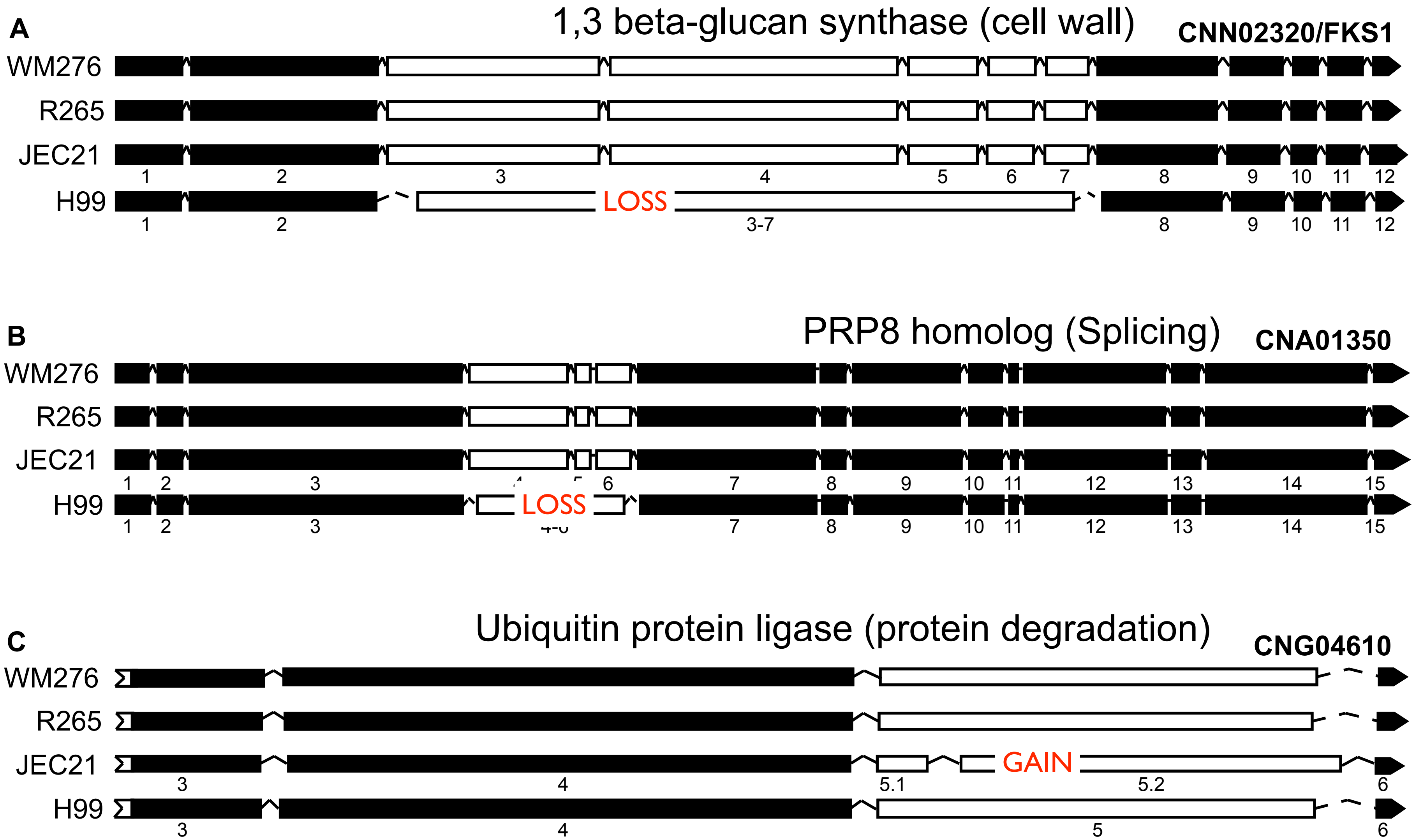
Intron loss was a precise excision

```
R265       CGACAAGTACATAAAACTTTTTTTGTGCCTGGCGCAAAGACTTTCCATTGCTGACAGAAAACAGGTTGAA
WM276      AGACAAGTACATAAAACTTTTTTTGTCTCTCCTCCAAACATTTTTCATTGCTGACAGAAAACAGGTTGAA
H99        AGACAA-------------------  Intron Missing  ------------------GTTGAA
JEC21_CDS  AGACAA-------------------                  ------------------GTTGAA
JEC21      AGACAAGTACATACTAGTCCTTGTG---CTATCCCAAAGACTTT-CATTGCTGACAGAAAACAGGTTGAA
            *****                                                           *****

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

Stajich and Dietrich. Euk Cell 2006.

Other examples of loss



Stajich and Dietrich. Euk Cell 2006.

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’

Acknowledgements

Fred Dietrich
Scott Roy

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