Comparative analysis of fungal gene structures reveals intron loss

Jason Stajich
Duke University

Fungal Comparative Genomics

- Understand genome evolution
- How does gene structure change over evolutionary time?
- Relative frequency and importance of these events

Existing Genome Annotation

New Annotation

Hemiascomycota

Euascomycota

Archiascomycota

Basidiomycota

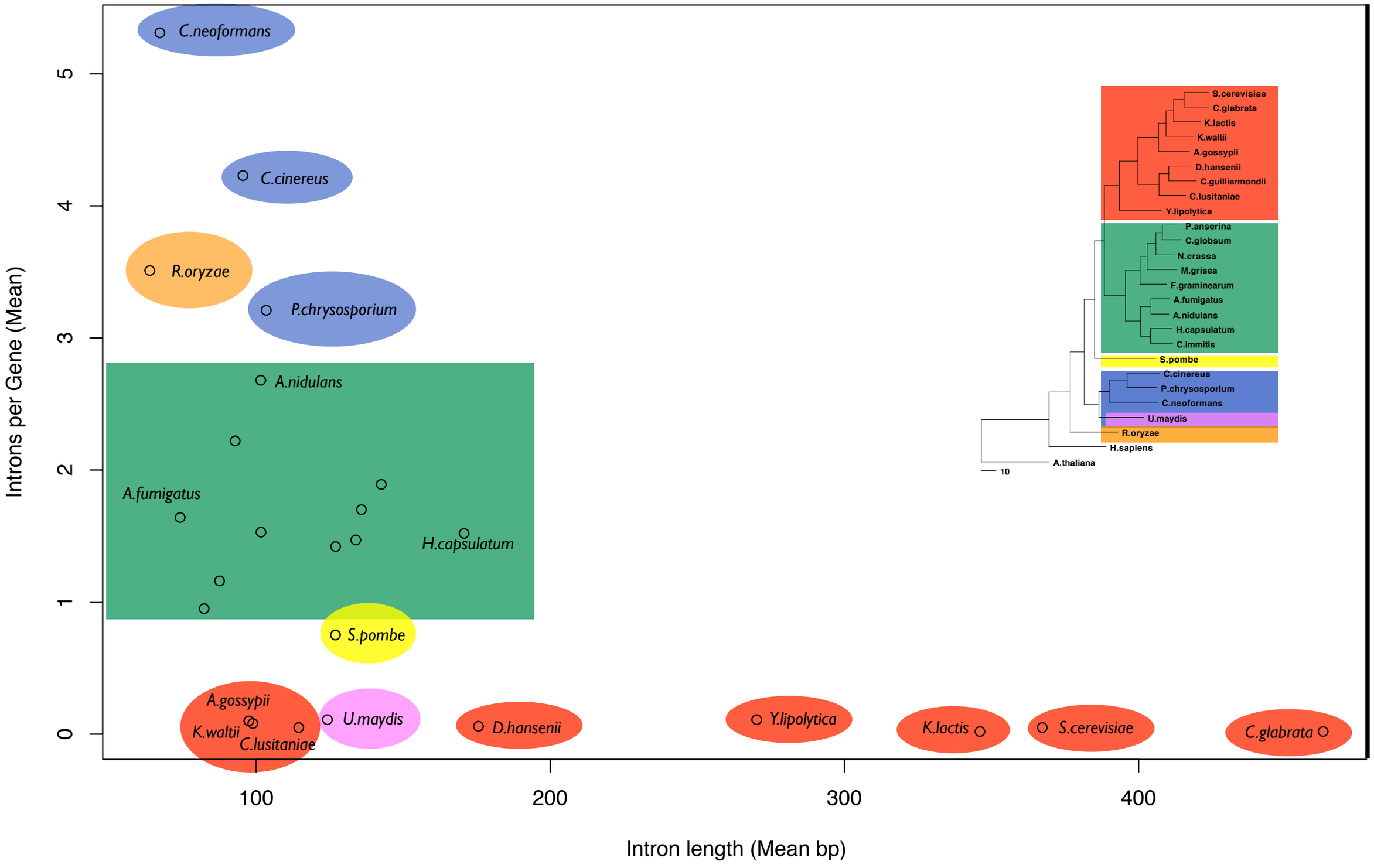
Zygomycota

S.cerevisiae
C.glabrata
K.lactis
K.waltii
A.gossypii
D.hansenii
C.guilliermondii
C.lusitaniae
Y.lipolytica
P.anserina
C.globsum
N.crassa
M.grisea
F.graminearum
A.fumigatus
A.nidulans
H.capsulatum
C.immitis
S.pombe
C.cinereus
P.chrysosporium
C.neoformans
U.maydis
R.oryzae
H.sapiens
A.thaliana

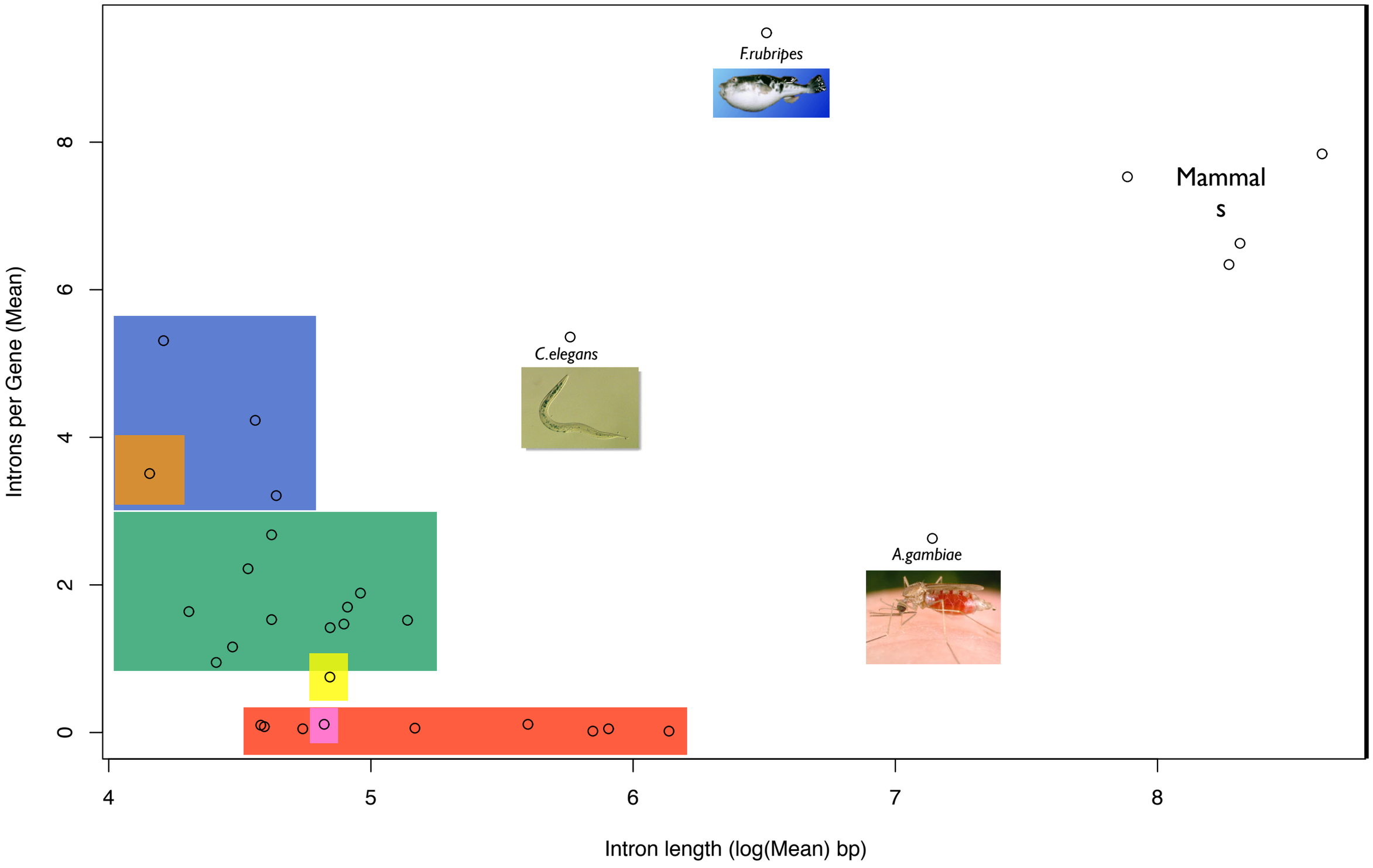
10

Protein ML tree based on 100 random orthologs

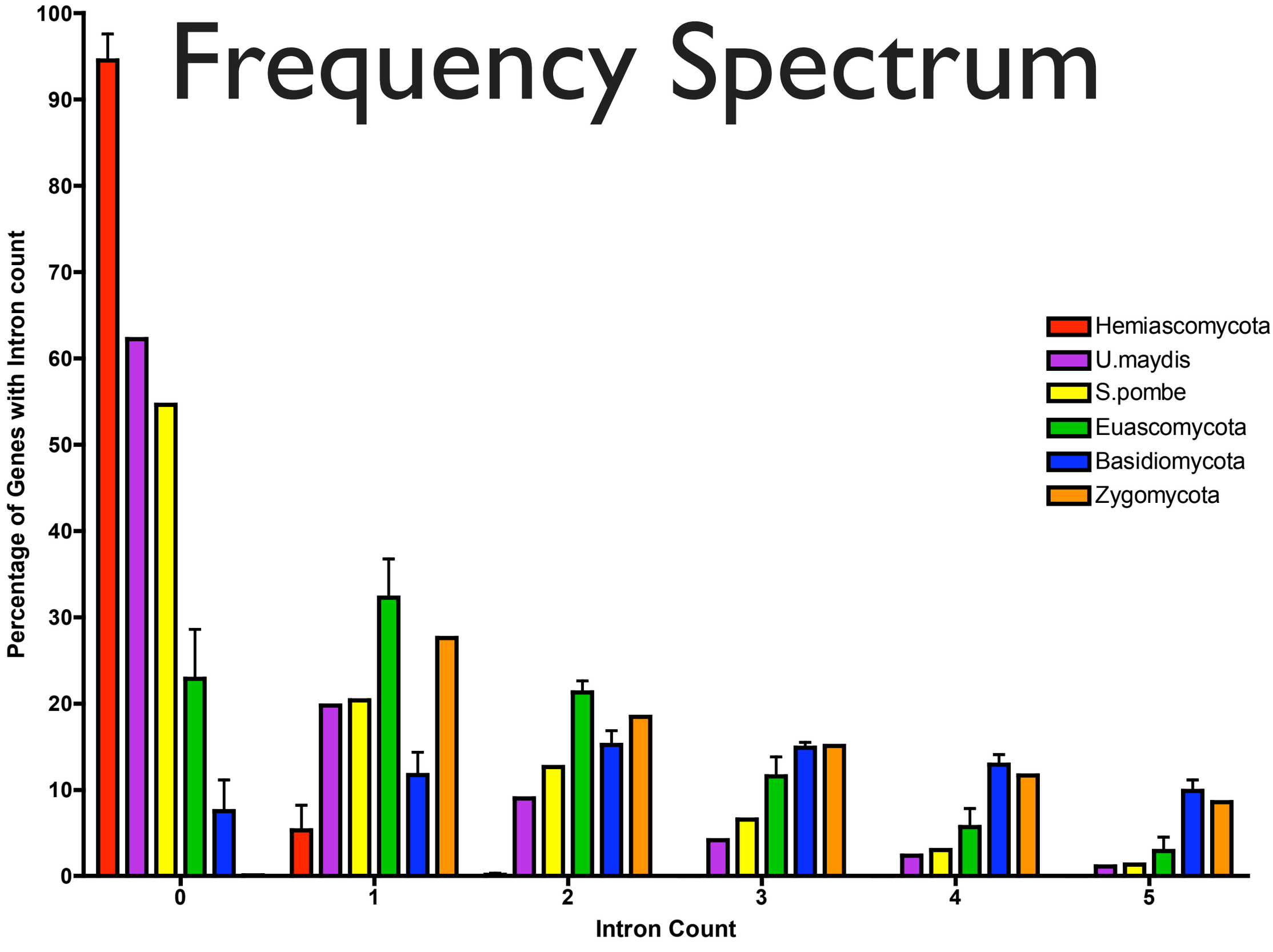
Fungal Introns per Gene versus Intron Size



Introns per Gene versus Intron Size



Introns per Gene Frequency Spectrum



Why have introns?

- Enable alternative splicing
- Nuclear export machinery may be coupled to splicing
- Nonsense Mediated Decay (NMD)

Evolution of gene structure

- How does exon-intron structure evolve?
- How the hemiascomycete pattern of intron poor genes arise? Ancestral or derived?
- Was the fungal ancestor intron rich or poor?
- What is mechanism of intron loss and gain?

Genome Annotation Pipeline

Rfam
tRNAscan

SNAP
Twinscan

ZFF to GFF3

GFF2 to GFF3

AUGUSTUS

GFF2

Genome

BLASTZ
BLASTN

Bio::SearchIO

Proteins

exonerate
protein2genome

GFF2 to GFF3

ESTs

exonerate
est2genome

Bio::DB::GFF

GFF to AA

predicted
proteins

FASTA
all-vs-all

Proteins

SearchIO

protein to
genome
coordinates

SearchIO

HMMER

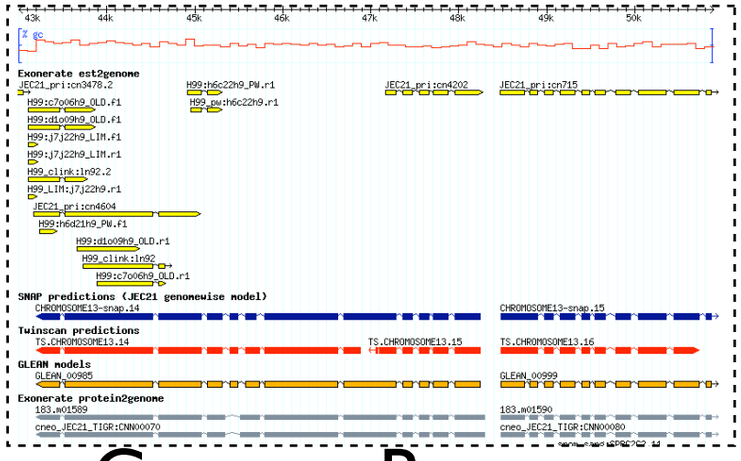
homologs
and
orthologs

TRIBE-MCL

Tools::GFF

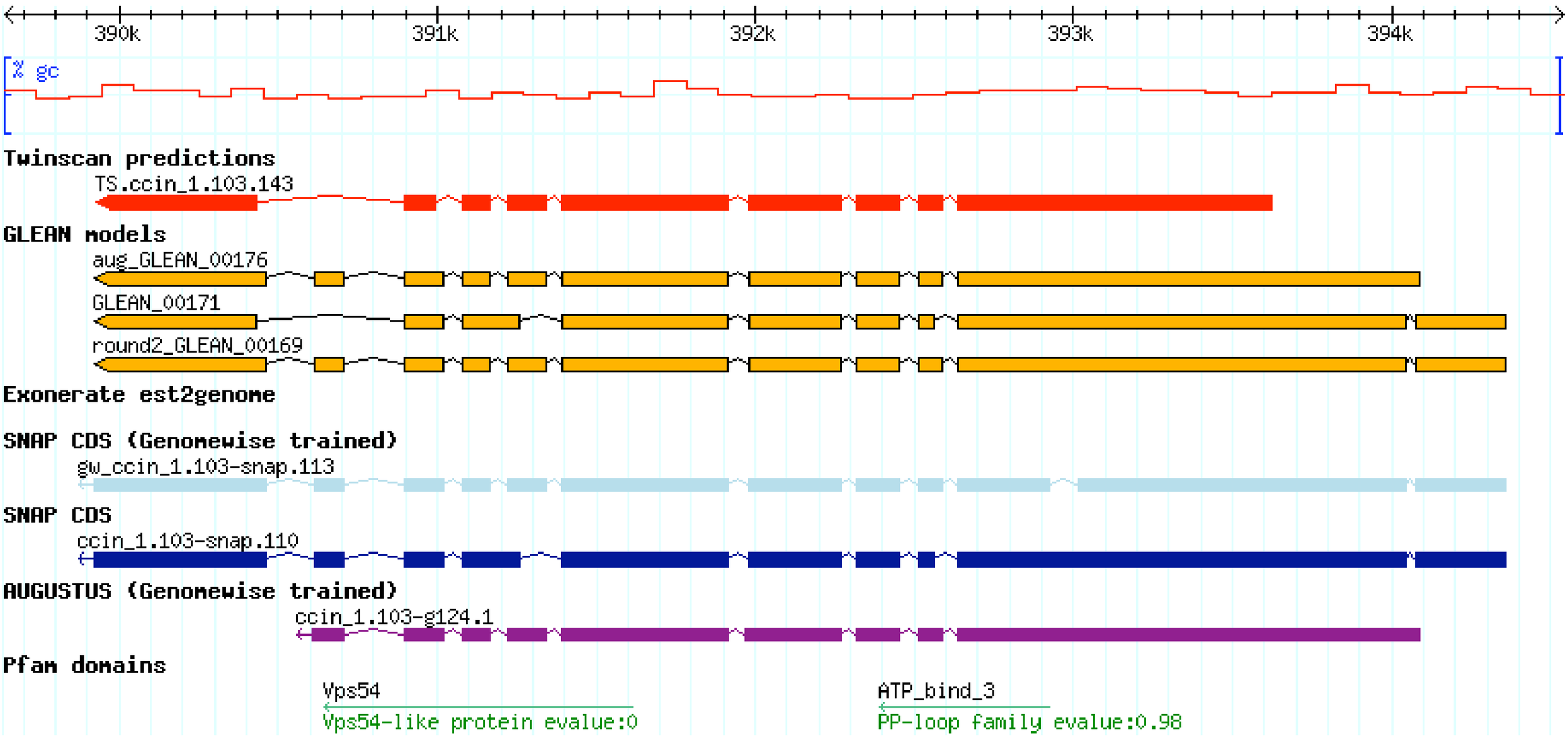
GLEAN
(combiner)

Gene
families



Genome Browser

Automated Annotation of Gene Models



Ortholog and Intron Evaluation Method

- Pairwise orthologs from best-reciprocal hits of FASTA all-vs-all search
- Multi-way orthologs tie together consistent cycles of pairwise orthologs
- Compute multiple sequence alignment (protein)
- Map intron position back into alignment
- Score shared position on species tree

Number of Shared Introns

Number of Orthologs

	scer	ylip	ncra	mgri	fgra	anid	spom	ccin	pchr	cneo	umay	rory	hsap	atha
scer	**6701**	51	31	30	32	36	23	10	10	27	EFB1 UBC13	21	8	13
ylip	3015	**6521**	147	118	147	136	91	50	48	92	MMS2 RPS18A	96	71	54
ncra	2493	3105	**10112**	5037	5902	3831	570	824	771	1103	RPS27B	996	791	441
mgri	2326	2945	5357	**11109**	6055	3910	560	858	806	1107	RPL14B RPL7A	999	783	447
fgra	2593	3256	5735	5787	**11640**	5085	692	1066	969	1373	ASC1 NOG2	1232	984	541
anid	2416	3037	4681	4672	5369	**9541**	698	1236	1103	1589	1134	1399	1008	566
spom	2516	2814	2651	2530	2766	2599	**4970**	580	602	938	1125	996	874	511
ccin	1779	2179	2495	2374	2633	2450	2025	**10119**	9168	5930	1255	2907	2413	3998
pchr	1827	2269	2551	2465	2754	2573	2120	4104	**12466**	5185	1255	2781	2191	1150
cneo	2192	2576	2788	2671	2951	2767	2407	2884	3072	**6652**	1367	3608	2991	1627
umay	1950	2378	2703	2594	2842	2667	2139	2623	2702	3079	**6522**	1239	1195	1139
rory	2349	2800	2724	2619	2930	2701	2673	2467	2685	2723	2488	**6468**	4089	2032
hsap	1791	2078	2061	2018	2203	1994	2132	2025	1838	2053	2803	1844	**33965**	3398
atha	1732	2011	1977	1965	2130	1983	2027	1688	1722	1962	2509	2509	2942	**29993**

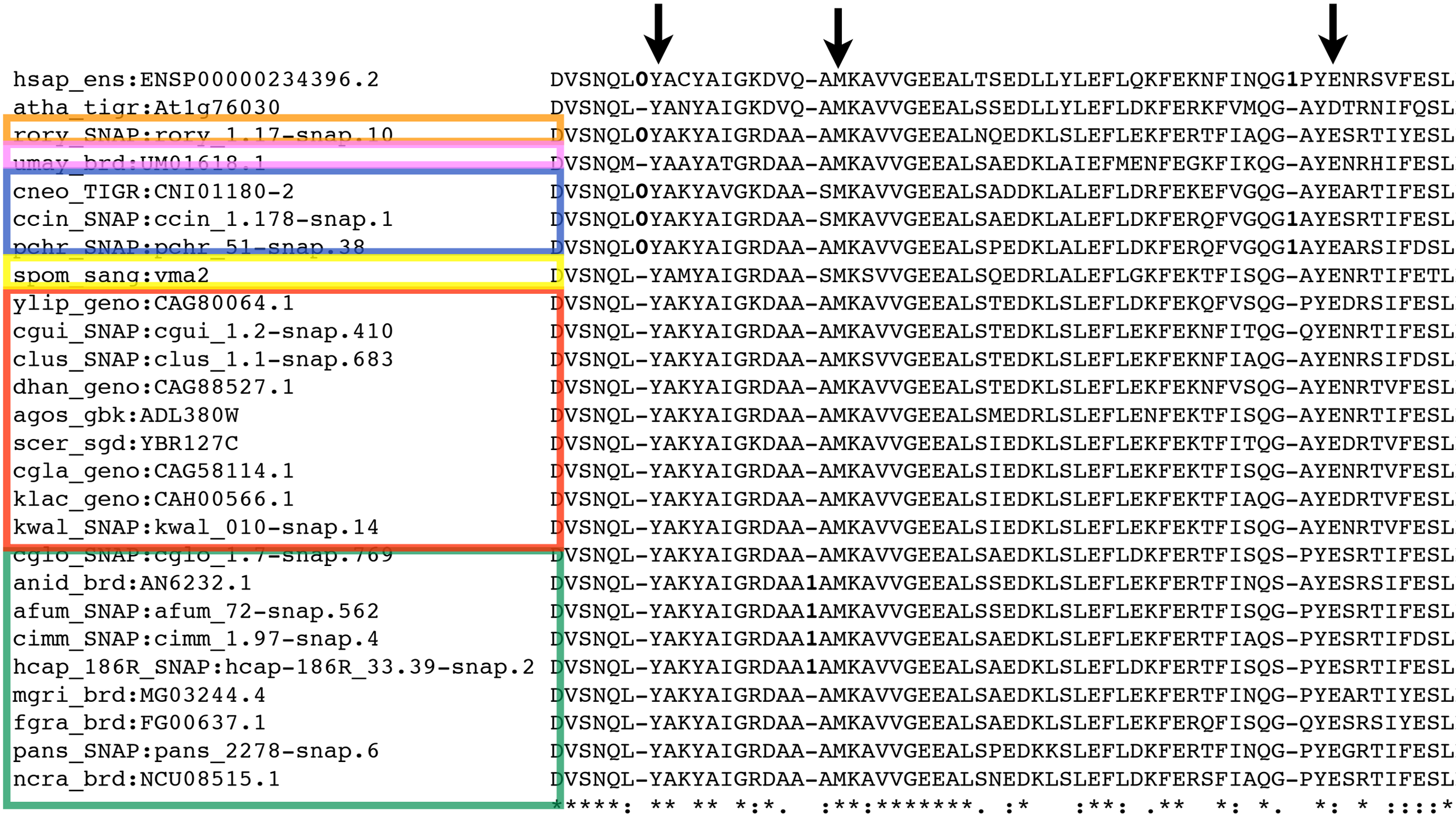
Pairwise orthologs summary

- Basal lineages in a clade tend to share more introns with outgroup (loss as an ongoing process)
- *Y.lipolytica* shares roughly 5x as many intron positions with species outside of Hemiascomycota as *S.cerevisiae* does.
- 20% more pairwise orthologs for *H.sapiens-R.oryzae* than *H.sapiens-A.thaliana*.
  - [Fungi-Metazoa more closely related]
- *C.cinereus* or *C.neoformans* have 2x as many introns shared with *H.sapiens* as with *A.thaliana* even with roughly comparable numbers of pairwise orthologs.

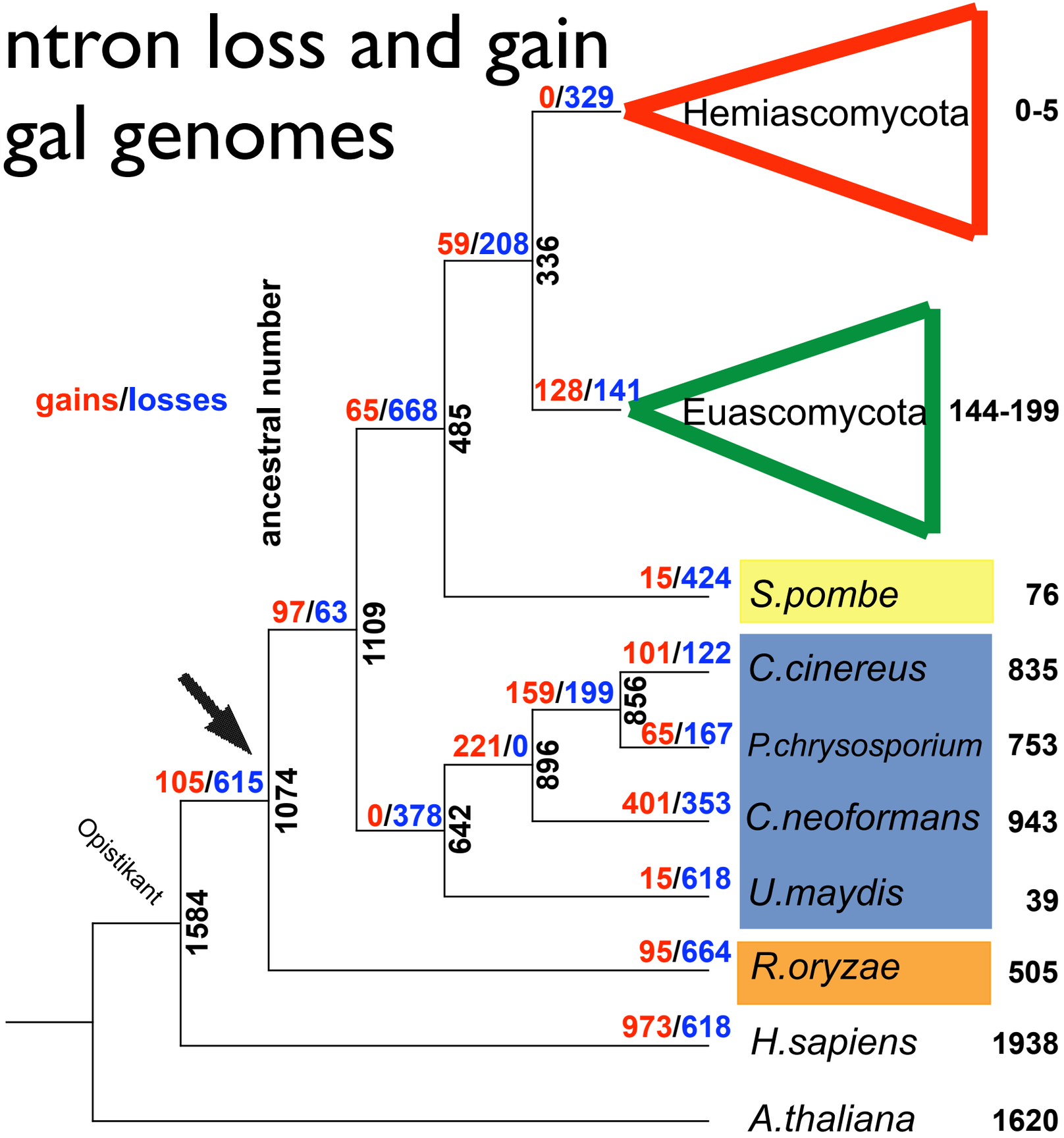
Multi-way orthologs & Introns

- 768 orthologous gene clusters across 26 species
- Filtering positions with no gaps in alignment, min 40% average similarity.
- 1311 shared intron positions can be considered
- Use ML method of Roy and Gilbert 2005 to infer ancestral states and rates of gain/ loss.

Example Alignment With Intron



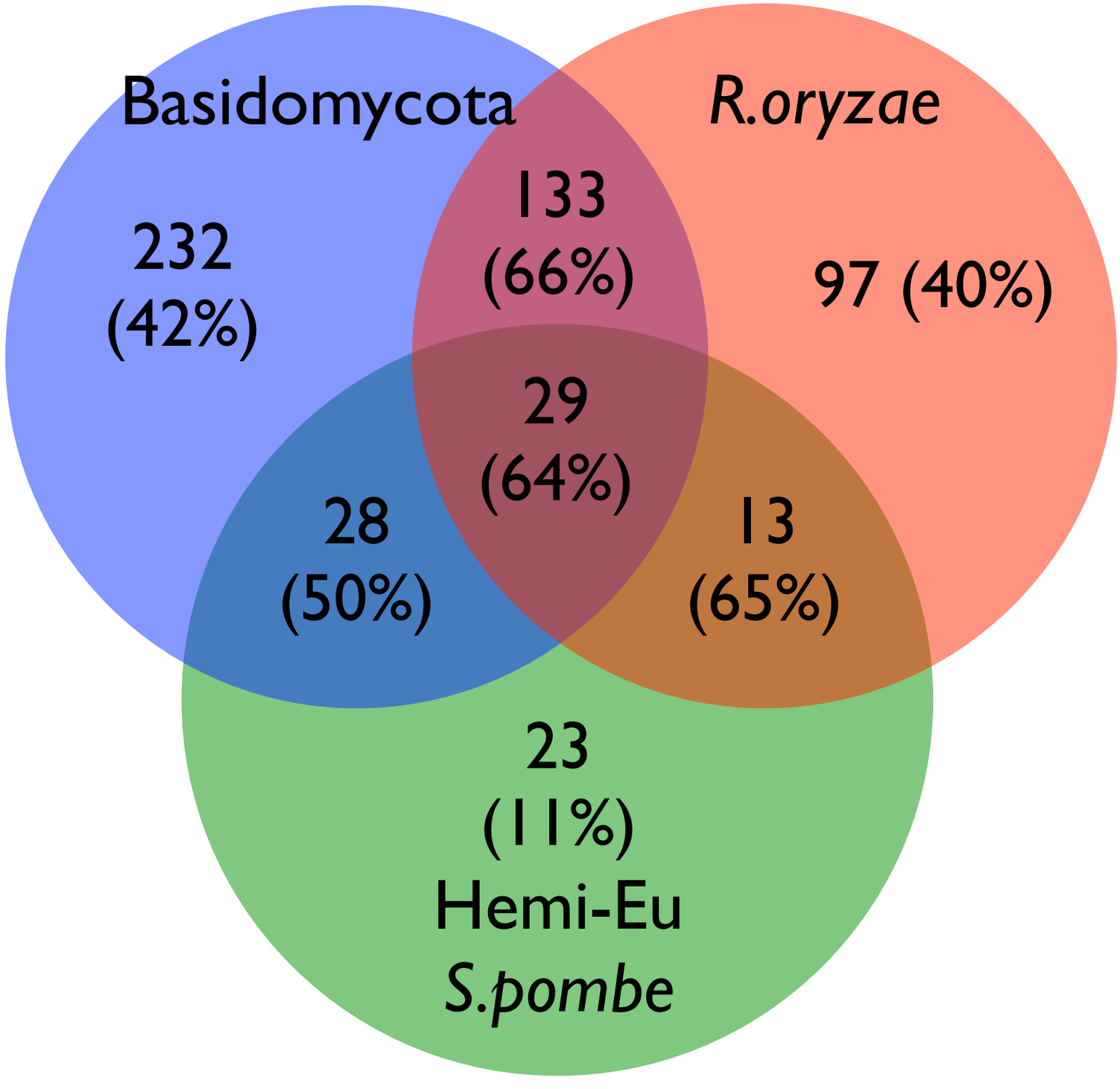
Patterns of intron loss and gain in fungal genomes



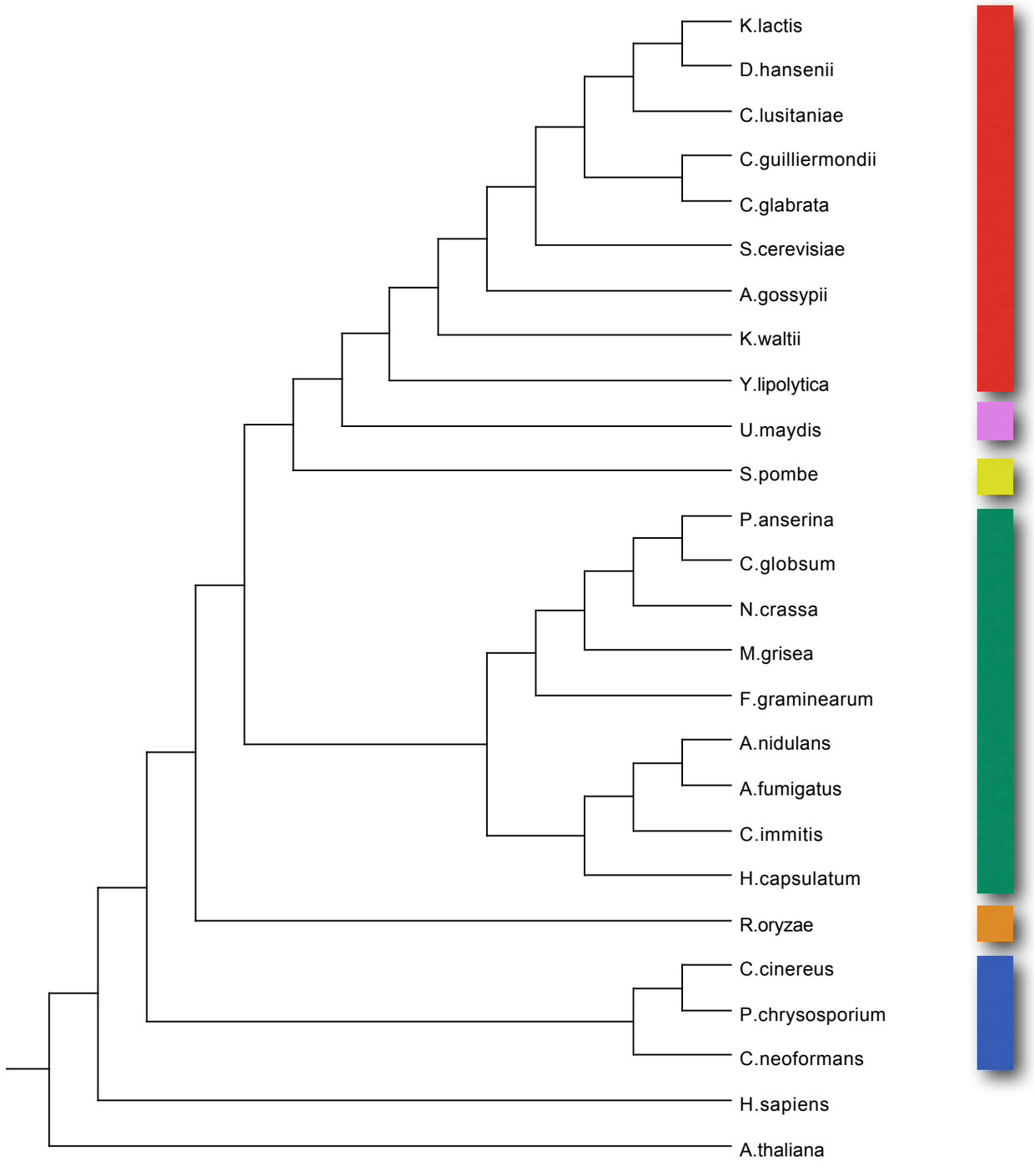
768 orthologous genes

Introns present
in extant species

Introns shared with Athaliana/Human outgroup



Parsimony Tree based on Intron Position



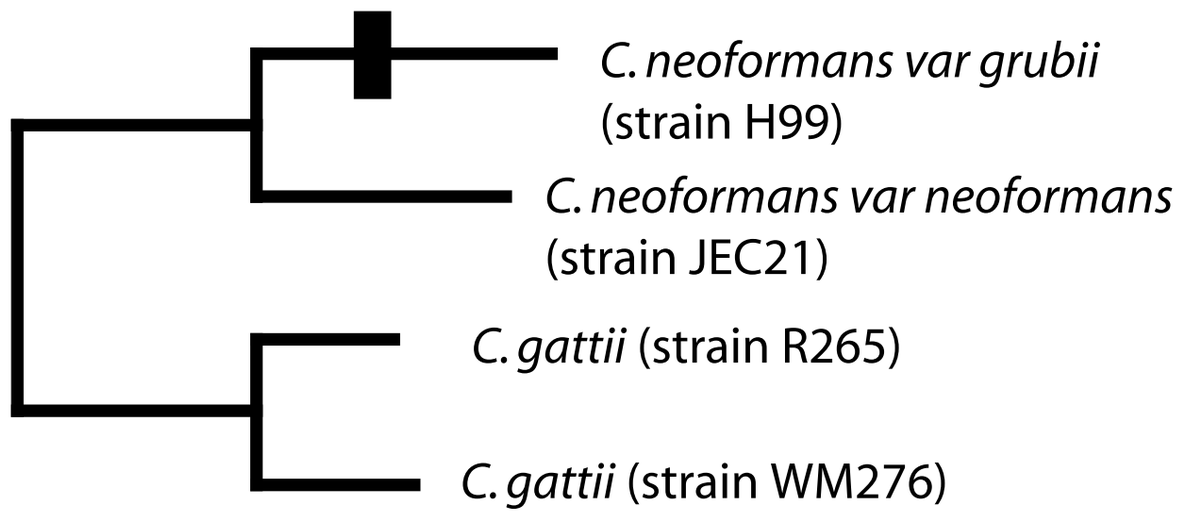
1327 characters
1 of 12 MP trees

Intron loss

- If loss predominates, how are introns lost?
- Fink (1987) model proposes mRNA integration into genome
- Boeke et al (1985) showed loss intron through RNA intermediate which is integrated into genome
- Large scale comparisons are too far awy to determine recent loss events

Searching for Intron loss in *C. neoformans*

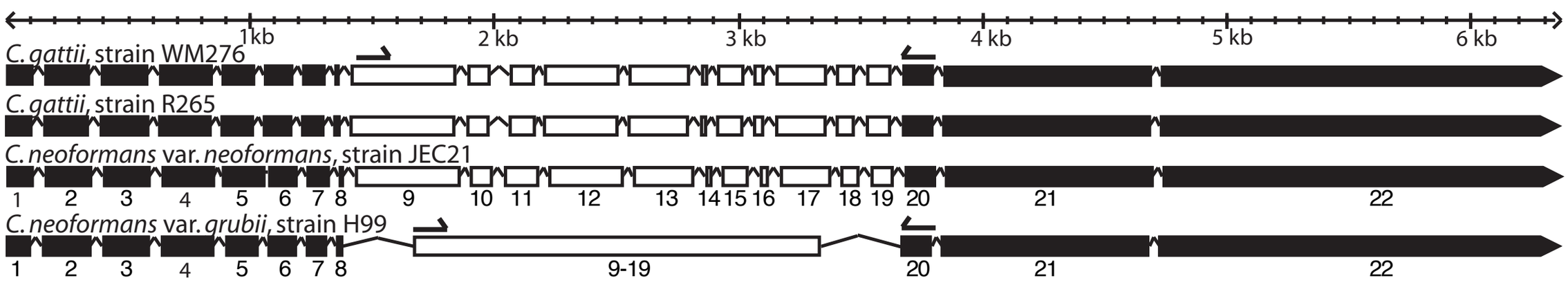
- Average of 5 introns per gene (Loftus et al, 2005)
- Average K_s between var. *grubii* and var. *gattii* 0.22 (roughly mouse-rat divergence), C.gattii vs C.neoformans K_s is 0.35
- 5133 4-way orthologs

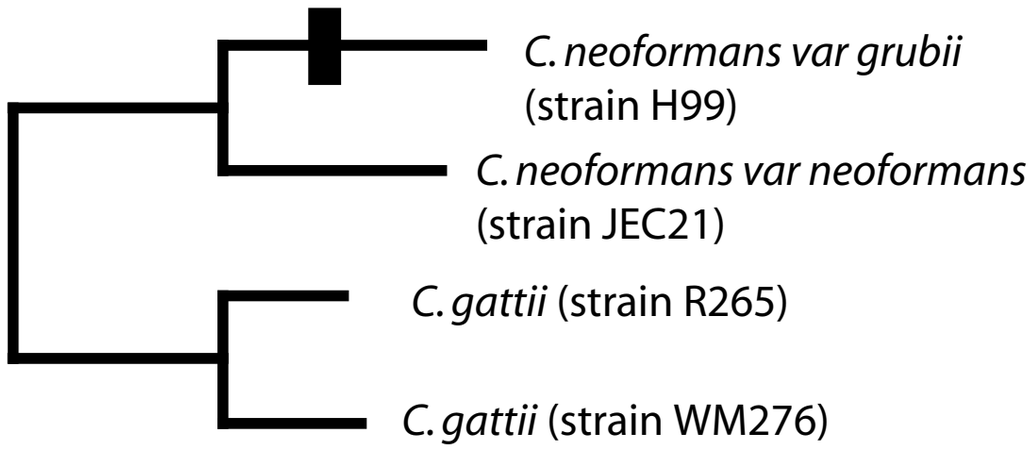


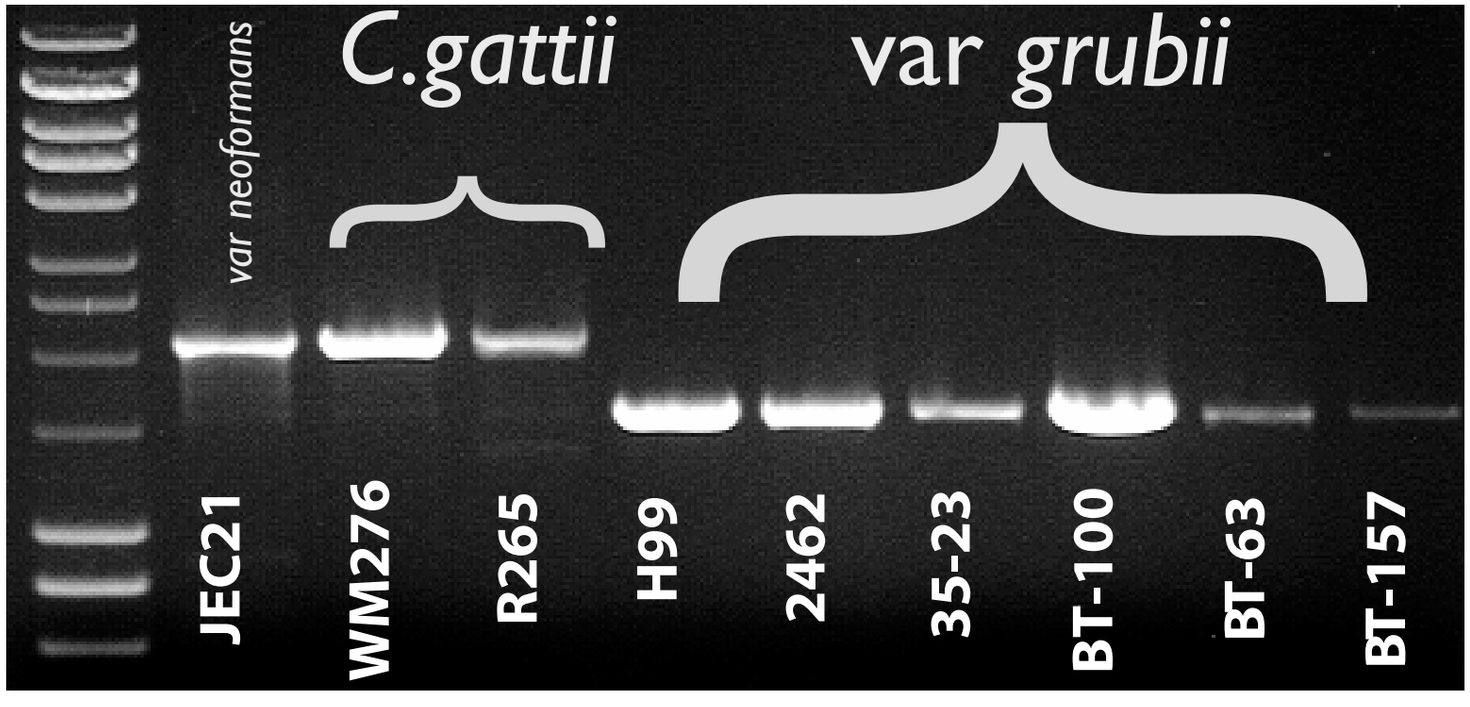
Few loss events observed

- 25 out of the 5133 loci had loss or gain events
- 2 had evidence 4 or more intron loss/gain events
- CNI01550, putative RNA helicase missing 10 introns in var. *grubii*.

10 introns lost in *C.neoformans* var. *grubii*







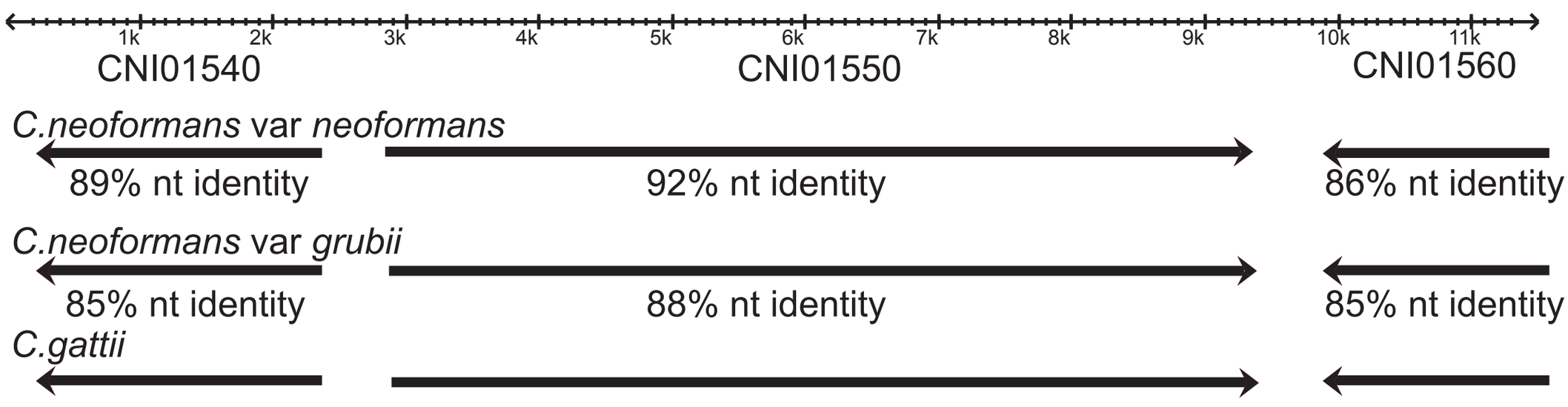
Perfect deletions of introns at the locus

```
B_R265       CGACAAGTACATAAAACTTTTTTTGTGCCTGGCGCAAAGACTTTCCATTGCTGACAGAAAACAGGTTGAATCAAGCCAATCTCT
B_WM276      AGACAAGTACATAAAACTTTTTTTGTGTCTGGTGCAAAGATTTTTCATTGCTGACAGAAAACAGGTTGAATCAAGCCAATCTCT
A_H99        AGACAA----------------------------------------------------GTTGAATCAAGCCAATCTCT
D_JEC21_CDS  AGACAA----------------------------------------------------GTTGAATCAAGCCAATCTCT
D_JEC21      AGACAAGTACATACTAGTCCTTGTG---CTATCCCAAAGACTTT-CATTGCTGACAGAAAACAGGTTGAATCAAGCCAATCTCT
              *****                                                      ********************
```

Intron 9

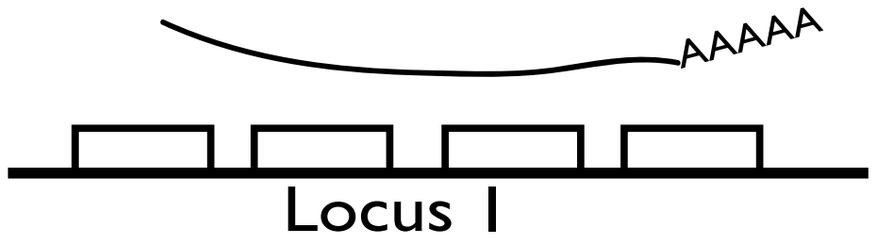
```
B_R265       CGCTGCCGAATTATGTCGATGTTGGAGATTTCTTGAGGTAAGCAACAGACTCGTAACAGCTTGTTCGGTCGCAA---ACCAGCT
B_WM276      CGCTGCCGAACTATGTCGATGTTGGAGATTTCTTGAGGTAAGCAACAGACTCGTAACAGCTTGTTCGGTCGCAA---ACCAGCT
A_H99        CCCTGCCGAATTATGTCGACGTTGGAGATTTCTTGAG-----------------------------------------------
D_JEC21_CDS  CCCTGCCGAATTATGTCGATGTTGGAGATTTCTTGAG-----------------------------------------------
D_JEC21      CCCTGCCGAATTATGTCGATGTTGGAGATTTCTTGAGGTACGTCGCAAACTCGTAACAGCTTGTTCGATCGCAAACCACCATCT
             * ******** ******** ******************
```

Intron 10

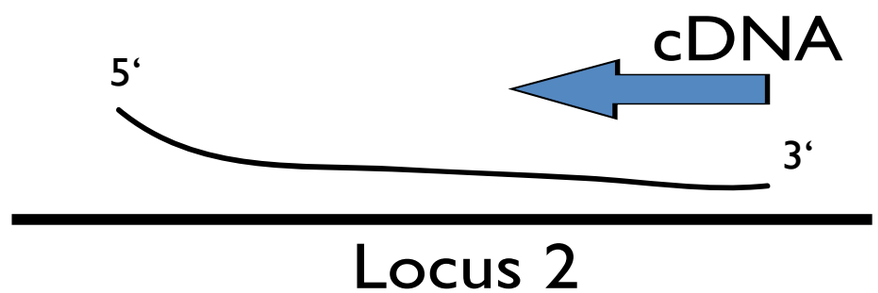


Model for intron loss

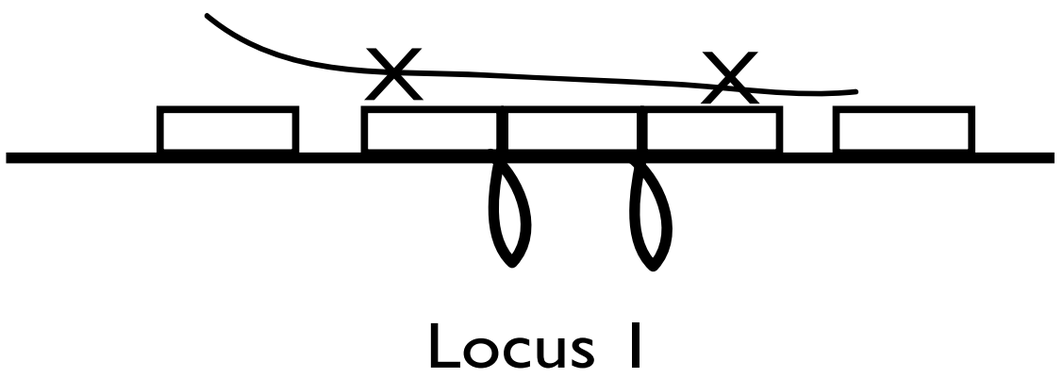
A transcription and splicing produce intronless transcript



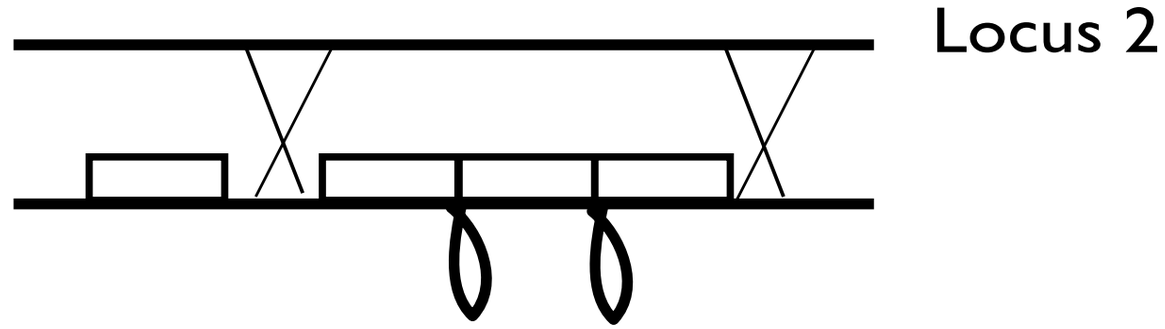
B1 cDNA integrates into genome



C homologous recombination of cDNA



B2 gene conversion from locus 2



Summary

- Fungal ancestor was intron rich
  - Estimate 2+ introns per gene (conservatively)
- Differential rates of loss and gain of introns
  - Loss was ongoing in Hemiascomycota, not all at once (more basal organisms have more introns)
- Homologous recombination at the locus can explain intron loss in some systems

Future work

- Identify unambiguous intron gain
- Evaluate gene structure change in paralogous gene families
- Evolutionary model for change in intron length

Acknowledgments

Fred Dietrich (Duke)
Scott Roy (Harvard)

Aaron Mackey (U Penn)
Ian Korf (UC Davis)
Mario Stanke (Göttingen)

TIGR, Genoluvres, Sanger Centre,
Stanford, DOE JGI,
Broad - Fungal Genome Initiative

Duke

Data and genome browsers available at
http://fungal.genome.duke.edu