Fungal Intron Evolution:

Why a small genome has many introns?

Kemin Zhou, Alan Kuo, Asaf Salamov, and Igor Grigoriev

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Introduction

Here we are trying to answer the question why one of the smallest genome *Sporobolomyces roseus* has one of the most introns of all fungal genomes in the context of fungal intron evolution. In this study we used a statistical comparative genomics approach toward intron number evolution among 16 fungal genomes.

Table1. Fungal genomes used in this study.

	Database	Species
Ascomycota	Aspni1	*Aspergillus niger*
	Mycfi1	*Mycosphaerella fijiensis*
	Mycgr1	*Mycosphaerella graminicola*
	Necha2	*Necria haematococca*
	Picst3	*Pichia stipitis*
	Trire2	*Trichoderma reesei*
	Trive1	*Trichoderma virens*
Basidiomycota	copci1	*Coprinus cinereus*
	cryneo1	*Cryptococcus neoformans*
	Lacbi1	*Laccaria bicolor*
	Phchr1	*Phanerochaete chrysosporium*
	Pospl1	*Postia placenta*
	Sporo1	*Sporobolomyces roseus*
	ustma1	*Ustilago maydis*
Zygomycota, Chytridiomycota	Phybl1	*Phycomyces blakesleeanus*
	Batde5	*Batrachochytrium dendrobatidis*

Exon number reduction half loss rule. *S. roseus* is an exception

Table 2. Intron evolution within genomes. Coding exon number between differently conserved genes were compared. The genes were divided in four conservation groups: all--all genomes (GCAS), between-- between different phyla (GCBP), phylum--within the same phylum (GCWP), and species-specific genes (SSG). The p-values for t-test are colored red if less than 10e-4, pink if less than10e-3, yellow if less than 10e-2, and green if less than 0.05.

dbname	all	p-val	between	p-val	phylum	p-val	species
Aspni1	3.76	8.61E-08	3.31	3.24E-09	3.02	0.002406	2.83
Batde5	5.90	0.000305	5.18	0.224571	4.86	8.79E-07	3.57
copci1	7.32	0.000322	6.61	8.94E-18	5.65	4.17E-34	4.38
cryneo1	7.29	0.000356	6.69	0.171129	6.37	1.32E-06	5.18
Lacbi1	7.89	2.61E-07	6.84	2.97E-10	6.11	1.43E-35	4.83
Mycfi1	2.49	0.5253	2.44	0.065689	2.52	5.96E-09	2.23
Mycgr1	2.48	0.087217	2.59	0.000774	2.75	0.00569	2.91
Necha2	3.33	0.001856	3.09	0.002464	2.97	0.00023	3.14
Phchr1	7.08	1.40E-05	6.32	2.41E-22	5.18	5.14E-12	4.34
Phybl1	6.18	0.002574	5.68	0.003718	6.54	2.89E-14	4.21
Picst3	1.44	0.425451	1.41	0.510146	1.44	0.098549	1.54
Pospl1	6.90	0.31744	6.69	0.983844	6.68	1.28E-06	5.92
Sporo1	7.21	0.677986	7.29	0.519538	7.48	0.405048	7.21
Trire2	3.31	3.97E-05	2.99	0.001305	2.85	0.046557	3.06
Trive1	3.35	5.51E-06	2.99	0.000228	2.84	0.045978	2.95
ustma1	1.67	0.846634	1.69	0.778802	1.67	0.003047	1.90

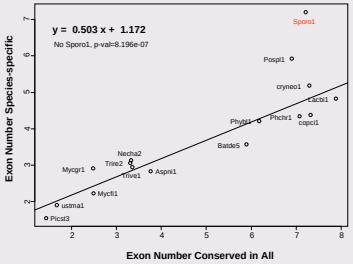

Figure 4. Half loss rule. Showing the linear relationship between the average number of exons in species specific genes (SSG) and that of genes conserved in all species (GCAS). Sporo1 is an exception although its inclusion still make the correlation statistically significant to p-value of 9.996e-06.

Most frequent and the shortest exon length and evidence of intron loss

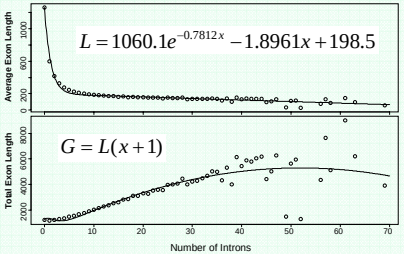

Figure 8. The shortest most frequent exon length. Top half, mean exon length as a function of number of introns. The equation set x to 60 to 70, the estimated exon length is 66-86 nt long. The bottom half is simply plots the total exon length against the intron numbers.

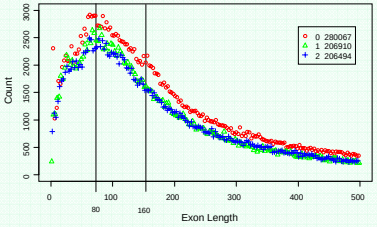

Figure 9. Exon length distribution. Exon length shorter than 500 nt from all 16 genomes are plotted with exon of different phases. Exon phase is defined as the remainder of the length of exon divided by 3. The total number of exons (all sizes) in different phases are shown in the legend. Phase 0 exon dominates.

No intron loss for *S. roseus* (Sporo1)

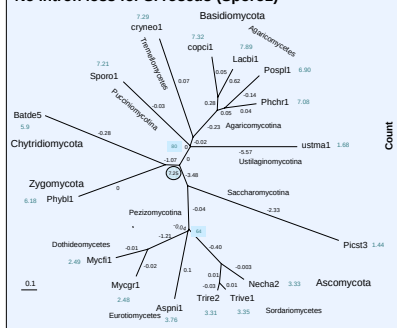

Figure 1. Whole genome phylogenetic tree and intron gain/loss estimates with Linear Least Square (LLS) method. The average number of coding exons from GCAS are labeled next to each database name (used for abbreviation of Species names). Bootstrap values are all 100% except for the two values shown in light blue boxes. Each value on the branch represents the estimated intron gain or loss. The major phyla and subphyla are labeled. The number in circle is the estimated number of coding exons of ancestor of fungi.

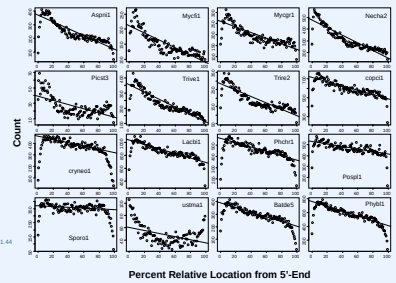

Figure 2. Intron relative location distribution. A regression line was drawn with data excluding the extreme values from both ends. The dip from both ends are due to edge effects. Data are grouped for every 1%.

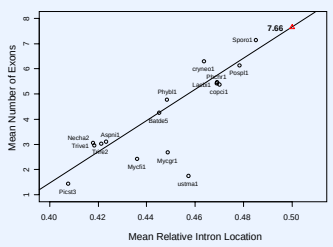

Figure 3. Estimating the number of exons in the ancestor of fungi with relative intron location.

Table 3. Conserved gene have shorter introns. Average of the log intron length (ALIL) were compared between conservation level all and species. Only Aspni1 showed no significant difference. The two genomes with the most intron loss ustma1 and Picst3 showed the opposite trend. Column diffexp is the natural exponential of the differences of ALIL (species – all).

dbname	All	Species	diffexp	P-value
Aspni1	4.277	4.279	0.2	0.88794301
Batde5	4.605	4.653	5.0	1.47E-05
copci1	4.130	4.208	5.0	4.93E-22
cryneo1	4.083	4.175	5.7	4.99E-30
Lacbi1	4.035	4.319	18.6	0
Mycfi1	4.288	4.776	45.8	1.82E-81
Mycgr1	4.283	4.878	58.9	7.27E-137
Necha2	4.180	4.235	3.7	9.06E-05
Phchr1	4.026	4.069	2.5	8.99E-08
Phybl1	4.584	4.653	7.0	2.87E-12
Picst3	4.428	4.251	-13.6	0.01634888
Pospl1	4.200	4.507	23.9	5.11E-129
Sporo1	4.412	4.529	10.2	7.64E-31
Trire2	4.460	4.527	6.0	0.00102148
Trive1	4.361	4.450	7.3	5.26E-08
ustma1	4.689	4.506	-18.2	0.0001481

Reversetranscriptase have divergent effects on exon number

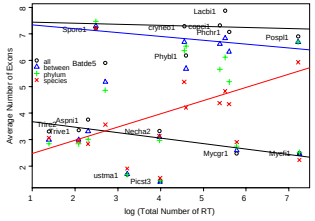

Figure 13. Average number of exons and amount of reverser transcriptase relationship. At species level, there is a positive correlation for 10 out of 16 genomes. At more conserved levels there is a negative correlation.

***S. roseus* has the least number of genes and the fourth smallest genome. Four yeast genomes are clustered at the lower end.**

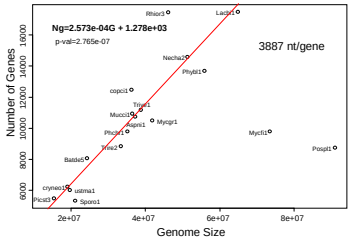

Figure 5. Constant carrying capacity of fungal genomes. Genome size was estimated by the total length of scaffolds belonging to each genome. The linear regression equation was derived from excluding Pospl1 and Mycfi1. Ng is the number of genes; G is the genome size. Pospl1 genome was highly polymorphic and assembled as diploid. Mycfi1 had large number of retroposons in the genome.

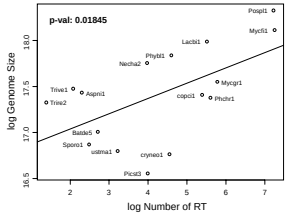

Figure 6. Correlation between genome size and number of reverse transcriptase. We got similar plot if using the total length of RT.

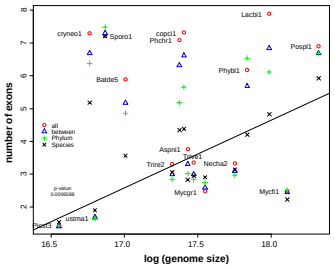

Figure 7. Larger genomes tend to have more exons per gene. We divided the genes into four groups as described in table 2. There is a linear relationship between the number of exons of SSG and log genome size. The p-value is 0.01588 when Sporo1 and cryneo1 are excluded, and 0.0006588 when Mycif1 is excluded in addition (line shown).

***S. roseus* Is an exception to number of exons and genome size correlation for SSG. It also has little RT footprints**

Summary and Discussion

Why S. roseus has the smallest genome

No intron loss detected by phylogenetic tree method, or very few by the relative intron location method. One of the smallest genomes with the least number of genes. Very few RT footprints. Exception to the rule of exon number reduction in less conserved genes.

Number of exons in ancestor genomes and gene birth big bang

Our phylogenetic tree gave us 7.25 which is an under estimation since intron loss do occur to even conserved genes such as those from the Ascomycota. There were 7.66 coding exons in the common ancestor of fungi. This method has limitation in ignoring any intron loss mechanism the leave the uniform distribution of intron undisturbed.

We found a linear relationship between the average exon length and the number of introns based on the 16 genomes: L=1060.1*exp(-0.8712) – 1.8961*x + 198.5 where x is intron number and L is average exon length. The exponential term provide evidence supporting intron loss. Another usage of the equation is to find the most dominant exon length of ancient ancestors. It seems that intron loss operates on gene level. If each gene have the same probability to have intron loss which assume to be a fixed number, then exons in longer genes will more likely to represent the ancestral state. From this study, 60-70 is the largest number of introns per gene, the equation tells us that the average length of this is 66-86 nt. If we independently look at the exon length distribution, we found that the peak of the exon length of all the fungal genomes is also at around 70-80 nt with a second shoulder at around 160 nt. Except for very short exons, fungal exons are dominated by phase 0 exons. This supports a hypothesis that there was an earlier gene birth "big bang" in the earlier evolution of eukaryotes. In this process, genes are predominantly generated by exon-shuffling of exons dominated by size of less than 100 nt with peak at 70-80 nt. Phase 0 exons dominates because they can be formed from intron loss between phase 1 and phase 2 exons. The average proteins are 400 aa accordingly the genes should have 15-16 exons. After this intron evolution was dominated by intron loss although intron gain still happen on an order of magnitude less through exon-shuffling or intron insertion into existing genes. The exon-shuffling has become less frequent largely due to selection pressure against protein function disruption. There could be a period of intron loss from 16 to 8 by mechanism other than RT.

Intron length variability

Intron lengths in fungi assume roughly log normal distribution so our analysis was carried out in log scale. The average introns from GCAS range from 56 to 109 nt, but those from SSG range from 58 to 131 nt. Our analysis clearly showed that there is an overall trend of less conserved genes tend to have longer introns in most genome, but shorter introns from P. stipitis and U. maydis the only two genomes where intron loss has reached the maximum extend and both have smaller genomes. Even with these small scale changes, overall the introns in fungi are really short. This may explain lack of correlation between intron size and genome size or amount of RT. It looks that the small size of intron has excluded transposable elements such as RT.

The carrying capacity of fungal genomes

There is a linear relationship between the number of genes and genome size for most fungi. Two out of the 18 genomes are exceptions to this rule. The genomic sequence of the first one *P. placenta* was assembled from diploid DNA. The other M. fijiensis is highly populated by RT. Nearly all fungal chromosomes are small and highly condensed. Most fungal genomes contain little repetitive DNA sequences. Fungi are the only major eukaryotic groups that are haploid. This linear equation that states for every 3887 nt there is a gene might be a summary for the above features. Furthermore there is a correlation between the number or total length of RT and genome size. It appears that RT and genes compete for the same genome space. It is reasonable to assume that larger genomes can afford having extra "useless" intron DNA and so there should be a correlation between the number of exons per gene and the genome size, but this is not so straight forward. There is only correlation, with exceptions of two Basidiomycota yeasts *S. roseus* and *C. neoformans* within the SSG.

Conclusion

We have proposed a new approach to look at the fungal intron evolution at genome scale. Out novel method to estimated the number of exons in the common ancestor of fungi could be applied to other genomes such animal or plants. Our proposal of a gene birth big bang based on solid fungal comparative genomic analysis could be a useful framework for further analysis. We also observed that reverse transcriptase can have divergent effect on the number of exon in the genome.

Acknowledgement

Statistical and mathematical assistance from Mingkun Li. Annotation efforts from Andrea Aerts, Bobby Otillar, Frank Korzeniewski, and Xueling Zhao. Computer support from IT: Jereme Brad and Ilya Malinov. Tireless support from GDS groups.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.