Ancient nature of alternative splicing and functions of introns

Kemin Zhou[1], Asaf Salamov,[1] Alan Kuo[1,] Andrea Aerts[1] and Igor Grigoriev[1]

[1]DOE Joint Genome Institute

[1]*To whom correspondence may be addressed. E-mail:* kzhou@lbl.gov.

March 14, 2011

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Ancient nature of alternative splicing and functions of introns

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

kzhou@lbl.gov

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract

Using four genomes: *Chamydomonas reinhardtii*, *Agaricus bisporus*, *Aspergillus carbonarius*, and *Sporotricum thermophile* with EST coverage of 2.9x, 8.9x, 29.5x, and 46.3x respectively, we identified 11 alternative splicing (AS) types that were dominated by intron retention (RI; biased toward short introns) and found 15, 35, 52, and 63% AS of multiexon genes respectively. Genes with AS were more ancient, and number of AS correlated with number of exons, expression level, and maximum intron length of the gene. Introns with tendency to be retained had either stop codons or length of 3n+1 or 3n+2 presumably triggering nonsense-mediated mRNA decay (NMD), but introns retained in major isoforms (0.2-6% of all introns) were biased toward 3n length and stop codon free. Stopless introns were biased toward phase 0, but 3n introns favored phase 1 that introduced more flexible and hydrophilic amino acids on both ends of introns which would be less disruptive to protein structure. We proposed a model in which minor RI intron could evolve into major RI that could facilitate intron loss through exonization.

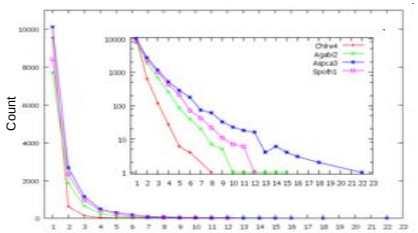
Characteristics of Input Data

Producer is the sequencing technology used to generate EST sequences. Count is the total number of EST used for the analysis. Average len. Is the average length of the EST. Min/max len. Is the minimum and maximum length of the EST input. Frac. Mapped is the fraction of EST that are mapped to the genomic DNA. Num. models is taken from the GeneCatalog or FilteredModels (If GeneCatalog does not exist) track of the JGI portal. GeneCatalog and FilteredModels are very similar to each others, with the former derived from the later. Exons/model is the number of exon per gene model taken from the GenaCatalog or FilteredModels track. Expr. Fraction is the average of total exon length divided by the total genomic length of the gene model derived from the COMBEST program. EST Coverage is total EST nucleotides divided by total non-gapped genomic length, then divided by expression fraction.

	Genome	Chlre4	Agabi2	Aspca3	Spoth1
EST	Producer	Sanger	454	454	Solexa+454
	Count	309,185	1,140,141	2,466,463	42,173,117
	Average len.	927.3	221.6	401.8	40.8
	min/max len.	15/5159	50/1479	47/961	26/1014
	Fraction Mapped	0.66	0.87	0.92	0.95
Genome	Size (mb)	112	30	36	39
	Gap fraction	0.075	0.007	0.056	0.003
	Num. models	16,696	10,443	11,624	8808
	Exons/model	7.37	5.99	3.47	3.02
	Expr. fraction	0.62	0.82	0.91	0.91
	GC Content	0.64	0.46	0.52	0.52
	EST Coverage	2.87x	8.94x	29.52x	46.30x

Results

1. Alternative Spliced Forms Distribution. There is an exponential decay of number of AS, as number of AS increases. There can be up to 22 AS in one gene.



Number of Models per Gene

2. Alternative Splicing Can Reach 63% and Antisense is Frequent

Num EST/Assembly is the number of EST per transcript assembly. Alt. of all/multiexon is the fraction of alternative spliced genes with respect to all genes and multiexon genes respectively. mRNA length from all genes.

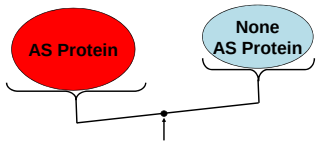
Genome	Chlre4	Agabi2	Aspca3	Spoth1
EST Coverage	2.87	8.94	29.52	46.30
Num EST/Assembly	22.1	134.0	252.2	5567.4
mRNA length	1054.1	1085.3	1739.0	1829.0
Alt. of all/multiexon	0.08/0.15	0.28/0.35	0.33/0.52	0.33/0.63
Antisense fraction	0.0675	0.1433	0.2375	0.2880

3. Number of exons (numexon), expression level (profmaxh), and maximum intron length (maxintronlen) determine number of AS. Linear regression analysis.

Genome	Chlre4		Agabi2		Aspca3		Spoth1	
Coefficients	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value
intercept	1.012e+00	<2e-16	1.043e+00	<2e-16	5.364e-01	<2e-16	5.599e-01	<2e-16
numexon	2.508e-02	<2e-16	9.289e-02	<2e-16	3.686e-01	<2e-16	3.629e-01	<2e-16
profmaxh	1.267e-03	<2e-16	1.038e-03	<2e-16	1.113e-03	<2e-16	2.180e-04	<2e-16
maxintronlen	7.174e-05	1.33e-4	1.154e-03	7.12e-09	1.553e-03	<2e-16	6.165e-04	3.12e-07
mRNAlen			-8.125e-05	2.05e-05			9.177e-05	3.36e-16
overall		<2.2e-16		<2.2e-16		<2.2e-16		<2.2e-16

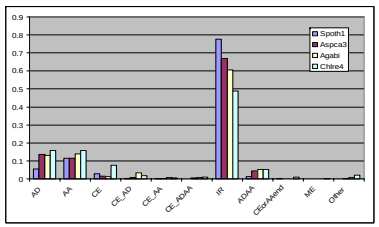
4. Alternatively Spliced Genes are More Ancient

Genome	AS	NoAS	P-value
Agabi2	9.81%	5.62%	1.60E-14
Aspca3	10.17%	7.15%	1.86E-10
Spoth1	10.23%	4.51%	<2.2e-16



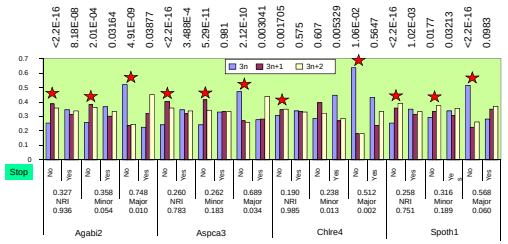
Blast search against nr - Fungi
% Hits All Three Kingdoms

5. Intron Retention Dominates the Eleven Types of AS



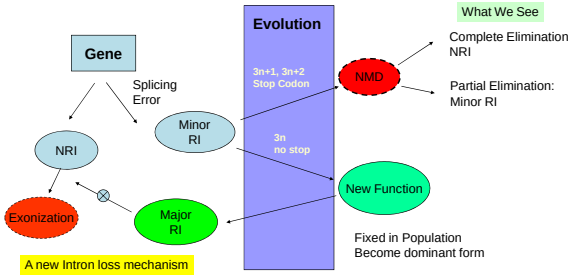
The types of alternative splicing are AD alternative donor, AA alternative acceptor, CE cassette exon, CE_AD cassette exon with alternative donor, CE_AA cassette exon with alternative acceptor, CE_ADAA cassette exon with both alternative donor and acceptor, IR intron retention, ADAA both alternative donor and alternative acceptor, CEorAAend cassette exon or alternative donor at end of the model, ME mutually exclusive exon, other none of the above types.

6. Minor or Non-Retained (NRI) Introns tend to have either stop codon, or 3n+1 or 3n+2 length, Major RI Introns favor 3n length and avoid stop codon



Combined effect of stop codon and RI type on intron 3n length bias. All introns are from coding regions. A consistent pattern of avoidance of 3n in stopless NRI and minor RI, and enrichment of 3n in stopless major RI. Star marks significant Chi-Square test p-values that are shown on the top. Numbers above RI type are fractions of stopless introns within each RI type, and below are fractions of each RI type. The insignificance of Chlre4 minor stopless introns is mainly attributed to the very low counts in this category.

Role of RI In Evolution: Minor ➔ Major → Loss



Observed NRI is a mixture of true NRI and minor RI after NMD clean up. Major RI can facilitate intron loss through exonization. Minor RI could evolve into major isoform. Observation of Minor and Major RI depends on environmental condition.

Conclusion

Abundant AS can be detected given enough EST coverage. Number of exons, expression level, and longest intron determine the extent of AS. AS is ancient. Intron retention is the most frequent AS in this study and can served as intermediates in evolution of new functions.

Acknowledgments

We appreciate the help from Frank Korzeniewski and Xiueling Zhao for supporting the genome annotation pipeline, Jasmyn Pangilinan and Erika Lindquist for running the Newbler EST assembler. We also thank Dr. Zhong Wang for insightful discussions.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231