Title

High EST Coverage Revealed Abundant Alternatively Spliced Transcripts

Permalink

https://escholarship.org/uc/item/66c8x2kp

Authors

Zhou, Kemin
Salamov, Asaf
Kuo, Alan
et al.

Publication Date

2010-03-24

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

High EST Coverage Revealed Abundant Alternatively Spliced Transcripts

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

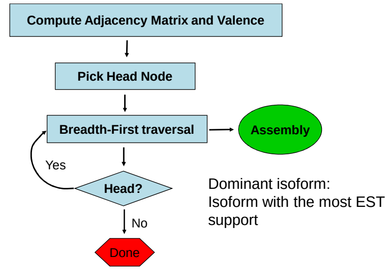
kzhou@lbl.gov

Abstract

Gene modeling has always been a challenge for computational biologists, but it becomes trivial when informed by expressed sequence tags (ESTs). New sequencing technologies such as 454 and Solexa can generate huge number of ESTs, but algorithms used in our production pipeline such as Newbler and PASA are inadequate in generating quality gene models from EST sequences. We developed a new algorithm COMBEST to generate partial or complete gene models from EST and genomic sequences. When applied to three genomes - *Chamydomonas reinhardtii*, *Agaricus bisporus*, and *Aspergillus carbonarius*- with coverage of 1.7 (2.7x), 22.5 (6.1x), and 51.9 (24.3x) ESTs per kb genomic sequence, we found different fractions of genes with alternative spliced forms of 6%, 16%, and 29% for three genomes respectively. These numbers are 11%, 25%, and 49% respectively if normalized to multi exon genes. The fraction of alternatively spliced genes is an inherent feature of a particular genome and the living condition of the organism; however, deep EST coverage is essential to reveal alternative splicing to the fullest extent. Since our algorithm also calculates the relative expression level for each splicing isoform, the results from COMBEST can be a useful resource for studying intron splicing and evolution in addition to being a tool for gene modeling in the high-throughput sequencing era. One of the interesting results from our analysis is that minor alternative forms with much shorter protein sequences occur at much lower frequencies as compared to the dominant isoform.

Input Data and Methods

EST Assemble Algorithm: COMBEST


Compute Adjacency Matrix and Valence

Pick Head Node

Breadth-First traversal

Assembly

Yes

Head?

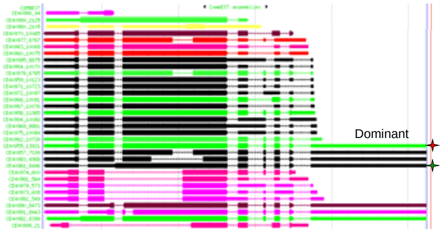
No

Done

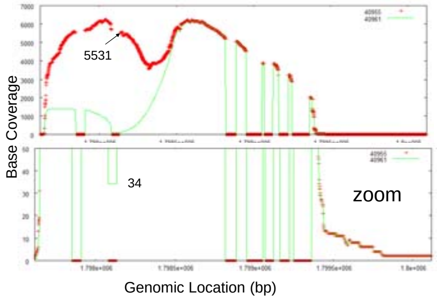
Dominant isoform:
Isoform with the most EST support

Example Assembly Result (GAPDH)

Model color: red relative partial; green: genuine; purple: with non-canonical intron; yellow: unusual model; brown: relative partial with non-canonical intron; black: non-of the above



Base coverage profile and relative expression level



Input genome and phylogeny

Genome	Organism	Phylogeny
Chlre4	Chlamydomonas reinharditii	Green Alga
Agabi2	Agaricus bisporus H97	Basidiomycota
Aspca3	Aspergillus cabonarius	Ascomycota

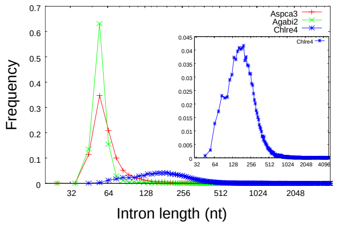
Input Data Summary

	Genome	Chlre4	Agabi2	Aspca3
EST	Count	309,185	1,140,141	2,466,463
	Average Len.	927.3	221.6	401.8
	Source	Sanger	454	454
	Fraction Mapped	0.604	0.597	0.764
Genomic	Genome Size (mb)	112	30	36
	Gap fraction	0.075	0.007	0.056
	Num. models	16,696	10,443	11,624
	Exons/model	7.37	5.99	3.47
	Coding fraction	0.62	0.82	0.91
	GC Content	0.64	0.46	0.52
	EST Coverage	2.68x	6.10x	24.28x

Results

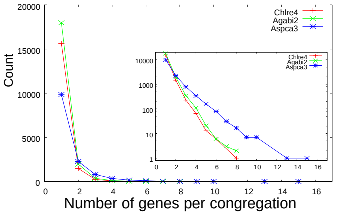
1. Intron length distribution and canonical splice site percentage

Genome	Chlre4	Agabi2	Aspca3
Canonical	98.98%	99.08%	97.41%



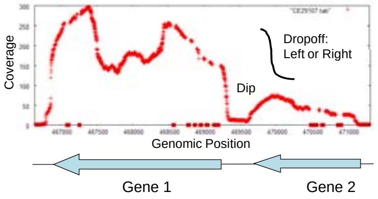
Assemblies with non-canonical introns are excluded from this analysis.

2. Degree of transcription overlap of neighboring genes



Congregation is a cluster of overlapping EST/genome alignments in genomic space.

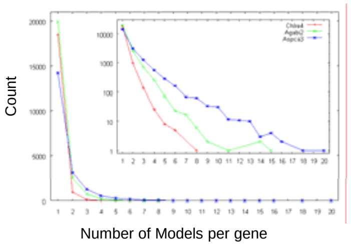
Dips or Drop-offs usually connects neighboring genes in base coverage profile



Percentage of Overlapping

Genome	Chlre4	Agabi2	Aspca3
Congreg. > 1 gene	10%	12%	27%
Gene Congregated	20%	23%	50%

3. Alternative Spliced Forms Distribution



4. Fraction of Alternative Splicing and Antisense

Alt. of all/multiexon: Fraction of genes with alternative spliced forms in all genes or in multi-exon genes. Full-length: is a subset of all models containing start and stop codons.

Genome	Chlre4	Agabi2	Aspca3
Num EST per Assembly	10.9	54.4	190.8
mRNA length	812.5	708.0	1525.1
Alt. of all/multiexon	0.06/0.11	0.16/0.25	0.29/0.49
Full-length Alt. of all/multiexon	0.08/0.16	0.34/0.37	0.39/0.56
Antisense fraction of all genes	0.0644	0.1213	0.2461

5. Linear regression of number of splice forms vs. number of exons (numexon), expression level (profmaxh), length of longest intron (maxintronlen), and log mRNA length.

Genome	Chlre4		Agabi2		Aspca3	
Factors	coefficient	p-value	coefficient	p-value	coefficient	p-value
Intercept	1.101609	<2e-16	1.626	<2e-16	0.5767	<2e-16
numexon	0.027572	<2e-16	0.06856	<2e-16	0.3266	<2e-16
profmaxh	0.001123	<2e-16	0.000839	<2e-16	0.001446	<2e-16
maxintronlen			0.00107	3.40e-06	0.002291	<2e-16
log(mRNAlen)			-0.0827	0.00281		
overall		<2.2e-16		<2.2e-16		<2.2e-16

⇨ The amount of alternative splicing in all three genomes correlates with number of exons and expression levels. In the two fungal genomes, length of longest intron also contribute to alternative splicing. There is a weak nagative correlation in Agagi2 with log mRNA length.

6. Introns boost transcription levels

T-test results for the expression level of genes as measured as the maximum height of base coverage profile (profmaxh). All of the p-values are less than 2.2e-16.

Genome	Chlre4		Agabi2		Aspca3	
Exon Structure	Single	Multi	Single	Multi	Single	Multi
All Genes	7.1	15.3	5.5	31.3	31.1	93.9
Full-Length	7.1	21.8	10.6	51.8	41.2	103.1

7. Top 10 Isoforms from Aspca3.

#Isoform	Peplen	Blast Hit Definition
20	337	glyceraldehyde-3-phosphate dehydrogenase
18	331	malate dehydrogenase, NAD-dependent
16	193	zinc knuckle domain protein
16	226	60S ribosomal protein L13
15	522	extracellular alpha-amylase
15	137	60S ribosomal protein L35a
15	395	conserved hypothetical protein
15	25	NOHIT
14	107	60S ribosomal protein L30
14	179	nucleosome binding protein

Conclusion

1. **COMBEST is a useful tool for studying gene expression and intron splicing given large number of ESTs and reasonably assembled genomes**
2. **The higher the EST coverage, the more alternative splicing, antisense, and transcription overlap are detected**
3. **Transcipt tends to run into neighboring genes (25% in Aspca3), but the frequency of this happening is low as characterized by Dips and Drop-offs**
4. **As much as 50% of multi-exon genes have alternative splicing in Fungi**
5. **Number of alternatively spliced isoforms correlates with number of exons and expression levels as well as length of longest introns in genomes with short average introns**
6. **Ancient genes tends to have more alternative spliced isoforms**
7. **The expression level of genes with introns is significantly higher than those without introns**
8. **Manual examination of alternatively spliced genes showed that minor isoforms that produces much shorter proteins than the dominant isoform usually occur at very low frequencies.**

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 insightfull discussions.

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.