YeATSAM analysis of the chloroplast genome of walnut reveals several putative un-annotated genes and mis-annotation of the trans-spliced rps12 gene in other organisms

Sandeep Chakraborty,
R - 44/ 1, Celia Engineers, T. T. C Industrial Area, Rabale, Navi Mumbai, 400701, India.

Abstract

An open reading frame (ORF) is genomic sequence that can be translated into amino acids, and does not contain any stop codon. Previously, YeATSAM analyzed ORFs from the RNA-seq derived transcriptome of walnut, and revealed several genes that were not annotated by widely-used methods. Here, a similar ORF-based method is applied to the chloroplast genome from walnut (Accid:KT963008). This revealed, in addition to the ∼84 protein coding genes, ∼100 additional putative protein coding genes with homology to RefSeq proteins. Some of these genes have corresponding transcripts in the previously derived transcriptome from twenty different tissues, establishing these as bona fide genes. Other genes have introns, and need to be manually annotated. Importantly, this analysis revealed the mis-annotation of the rps12 gene in several organisms which have used an automated annotation flow. This gene has three exons - exon1 is ∼28kbp away from exon2 and exon3 - and is assembled by trans-splicing. Automated annotation tools are more likely to select an ORF closer to exon2 to complete a possible protein, and are unlikely to properly annotate trans-spliced genes. A database of trans-spliced genes would greatly benefit annotations. Thus, the current work continues previous work establishing the proper identification of ORFs as a simple and important step in many applications, and the requirement of validation of annotations.

Introduction

Common walnut (*Juglans regia* L.) derives its economic importance for its wood and the high nutritional value of the nuts [1]. The walnut genome sequence was recently obtained from the cultivar Chandler [2]. Also, the complete chloroplast sequence (GenBank: KT963008) was annotated with CpGAVAS [3], revealing 86 protein-coding genes [4].

Previously, the YeATS [5] suite was developed to address assembly artifacts in RNA-seq derived transcriptomes [6], and was applied to identify metagenomes [7, 8]. The possibility of mis-annotation of genes (transcript from heat shock protein of the fungi *Cladosporium cladosporioides* has been erroneously annotated as a saffron gene) was demonstrated, emphasizing the need to remove contamination from transcriptomes as an important first step [8]. YeATS was extended (YeATSAM) to demonstrate several important omissions by other tools [9]. Finally, it was proposed that the merged transcripts might bias expression counts [10].

Here, the walnut chloroplast sequence is analyzed to identify additional protein coding genes that are not annotated. In addition to finding ∼100 putative genes, some of them occuring in the transcriptome obtained earlier [2], a major finding in this study was the identification of the mis-annotation of the rsp12 gene in several organisms. rsp12 is a trans-spliced gene having three exons - exon1 is located ∼28kbp from exon2 and exon3 [11, 12]. The mis-annotation is a direct result of the complexity faced by *de novo* annotation methods in detecting such genes.

Methods

The 'getorf' program from the EMBOSS suite [13] was used to obtain the ORFs. As described previously, a BLAST database of protein peptides using ∼30 organisms from the Ensembl genome [14] was created [9,10]. This is done to reduce computational times. All ORFs $>$ 60 aa were BLAST'ed to this database [15]. Final verification is done on the 'nr' database choosing RefSeq proteins. The transcriptome from twenty different tissues is at http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/Reju/transcriptome/Trinity_Assembly [2, 5, 7]. Multiple sequence alignment was done using MAFFT (v7.123b) [16], and figures generated using the ENDscript 2.0 server [17]. PHYML (v3.0) was used to generate phylogenetic trees from alignments [18]. All protein structures were rendered by PyMOL(TM) Molecular Graphics System, Version 1.7.0.0. (http://www.pymol.org/). Results reported here are obtained using a simple workstation in a few hours.

Results

The walnut chloroplast sequence analysis revealed ∼10000 ORFs (FILE:genome.chloroplast.orf.fa), with 530 ORFs having $>$ 60 aa ($\text{ORF}_{530}^{Chloroplast}$, FILE:genome.chloroplast.orf.fa.filterlen.60.fa). There are currently 56036 annotated genes for *J. Regia* in the NCBI database ($\text{NCBI}_{56036}^{JRegia}$, FILE:JRegia.nr.fa). As a corroboration step, $\text{ORF}_{530}^{Chloroplast}$ were BLAST'ed [15] to $\text{NCBI}_{56036}^{JRegia}$, and revealed 84 (and not 86) protein coding genes (FILE:annotated.ncbi.txt). The two missing genes might be due to the 60 aa cutoff applied here.

Next, $\text{ORF}_{530}^{Chloroplast}$ were BLAST'ed to the local BLAST database from ∼ 30 organisms (see Methods), and revealed 118 protein coding fragments (FILE:not.annotated.local.ncbi.txt), which have not been annotated. These 118 ORFs were then BLAST'ed to the RefSeq proteins in the 'nr' database, resulting in a total of 99 entries (FILE:not.annotated.ncbi.txt). An example is ORF KT963008.1_2812 (125 aa) homologous to the RefSeq protein (*Pharus latifolius*, Accid:YP_008080537.1, Evalue=1E-67) (Fig 1a). This ORF is 100% identical to transcript C43892_G2_I1 identified in previous work [2], and is thus a bona fide gene (Table 3). Some of these ORFs are fragments, and might need manual curation before annotation. For example, three ORFs (KT963008.1_4166, KT963008.1_4182 and KT963008.1_4200) can be merged to create a protein homologous to the RefSeq gene from *Medicago truncatula* (Accid:XP_013455718.1, Evalue=4E-45) (Fig 1b). This gene is "encoded by transcript MTR_4g451295", and is a bona fide gene at least in *M. truncatula*. Another strategy to exclude pseudo-genes compared the predicted genes to the transcriptome derived from twenty different tissues [2], confirming the presence of at least five real genes (Table 4).

An important finding in the current study is the mis-annotation of the ribosomal rps12 gene in several organisms (Table 1, Fig 2). The rps12 gene has three exons - exon2 and exon3 are proximal, while exon1 (38 aa) is distant, and the complete protein is assembled through trans-splicing [11] (Table 2). A pentatricopeptide repeat protein facilitates the trans-splicing of rps12 [19] (Fig 3). It is not surprising that automated annotation tools will pick up the closest ORF with a start codon prior to exon2 to complete a theoretically possible protein sequence.

Discussion

ORF based annotation is trivially simple. Yet, as the current manuscript shows, several protein coding genes remain un-annotated in the walnut chloroplast. Comparison to transcriptomes enables further corroboration of non-pseudo genes. On the other hand, *de novo* identification of trans-spliced genes through automated methods is almost impossible. Mis-annotation of the rps12 gene by automated methods need to be purged before they become mainstream. Also, a database of trans-spliced genes would greatly facilitate their proper annotation. Future work in the YeATS suite will address the automated stitching of spliced genes, identified through fragmented ORFs.

References

1. Aradhya MK, Potter D, Gao F, Simon CJ (2007) Molecular phylogeny of juglans (juglandaceae): a biogeographic perspective. Tree Genetics & Genomes 3: 363–378.

2. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, et al. (2016) The walnut (juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of nonstructural polyphenols. The Plant Journal .

3. Liu C, Shi L, Zhu Y, Chen H, Zhang J, et al. (2012) Cpgavas, an integrated web server for the annotation, visualization, analysis, and genbank submission of completely sequenced chloroplast genome sequences. BMC genomics 13: 715.

4. Hu Y, Woeste KE, Dang M, Zhou T, Feng X, et al. (2016) The complete chloroplast genome of common walnut (juglans regia). Mitochondrial DNA Part B : 1–2.

5. Chakraborty S, Britton M, Wegrzyn J, Butterfield T, Martinez-Garcia PJ, et al. (2015). YeATS-a tool suite for analyzing RNA-seq derived transcriptome identifies a highly transcribed putative extensin in heartwood/sapwood transition zone in black walnut.

6. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.

7. Chakraborty S, Britton M, Martínez-García P, Dandekar AM (2016) Deep RNA-seq profile reveals biodiversity, plant–microbe interactions and a large family of NBS-LRR resistance genes in walnut (*juglans regia*) tissues. AMB Express 6: 1.

8. Chakraborty S (2016) Transcriptome from saffron (crocus sativus) plants in jammu and kashmir reveals abundant soybean mosaic virus transcripts and several putative pathogen bacterial and fungal genera. bioRxiv : 079186.

9. Chakraborty S, Martinez-Garcia PJ, Dandekar A (2016). YeATSAM analysis of the walnut and chickpea transcriptome reveals key genes undetected by current annotation tools [version 1; referees: 1 approved, 1 not approved].

10. Chakraborty S (2016) Rna-seq assembler artifacts can bias expression counts and differential expression analysis - case study on the chickpea transcriptome emphasizes importance of freely accessible data for reproducibility [version 2; referees: 2 not approved]. F1000Research 5.

11. Koller B, Fromm H, Galun E, Edelman M (1987) Evidence for in vivo trans splicing of pre-mrnas in tobacco chloroplasts. Cell 48: 111–119.

12. Sharma MR, Wilson DN, Datta PP, Barat C, Schluenzen F, et al. (2007) Cryo-em study of the spinach chloroplast ribosome reveals the structural and functional roles of plastid-specific ribosomal proteins. Proceedings of the National Academy of Sciences 104: 19315–19320.

13. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.

14. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, et al. (2016) Ensembl genomes 2016: more genomes, more complexity. Nucleic acids research 44: D574–D580.

15. Camacho C, Madden T, Ma N, Tao T, Agarwala R, et al. (2013) BLAST Command Line Applications User Manual .

16. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772–780.

17. Robert X, Gouet P (2014) Deciphering key features in protein structures with the new endscript server. Nucleic acids research 42: W320–W324.

18. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33: W557–559.

19. Schmitz-Linneweber C, Williams-Carrier RE, Williams-Voelker PM, Kroeger TS, Vichas A, et al. (2006) A pentatricopeptide repeat protein facilitates the trans-splicing of the maize chloroplast rps12 pre-mrna. The Plant Cell 18: 2650–2663.

20. Souvorov A, Kapustin Y, Kiryutin B, Chetvernin V, Tatusova T, et al. (2010) Gnomon–ncbi eukaryotic gene prediction tool. National Center for Biotechnology Information : 1–24.

21. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, et al. (2003) Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research 31: 5654–5666.

Table 1: **Enumerating mis-annotated rps12 genes in the NCBI RefSeq database:** The bona fide rps12 gene is 124 aa long, and is identified correctly through conceptual translation. rps12 is a trans-spliced gene with three exons - exon1 is ∼28kbp away from exon2 and exon3. Automated methods identify a ORF in the proximity of exon2 with a start codon, leading to a mis-annotated gene. Gnomon is the method for NCBI eukaryotic genome annotation pipeline [20]. JCVI Eukaryotic Genome Annotation Pipeline uses the EVidenceModeler (EVM) [21]. Accid: Accession id.

Accid	Organism	Length	Method	Date
ALO71482.1	*Juglans regia*	124	conceptual translation	20-MAY-2016
YP_007375022.1	*Quercus rubra*	124	conceptual translation	18-JAN-2013
YP_008963715.1	*Liquidambar formosana*	124	conceptual translation	20-DEC-2013
XP_015867265.1	*Ziziphus jujuba*	143	Gnomon	24-MAR-2016
XP_015898012.1	*Ziziphus jujuba*	143	Gnomon	24-MAR-2016
XP_016724328.1	*Gossypium hirsutum*	135	Gnomon	18-MAY-2016
XP_012575527.1	*Cicer arietinum*	212	Gnomon	08-JUN-2015
XP_013443673.1	*Medicago truncatula*	159	EVM, PASA	25-AUG-2015

Table 2: **Mis-annotation of rps12 gene**: The rps12 protein has three exons - exon1 is ∼ 28kbp away from exon2 and exon3, and is assembled into the mature protein by trans-splicing [19]. There are other theoretically possible protein sequences that can be assembled by replacing exon1 with a different ORF (KT963008.1_6221 and KT963008.1_6242 in the current case) proximal to exox2.

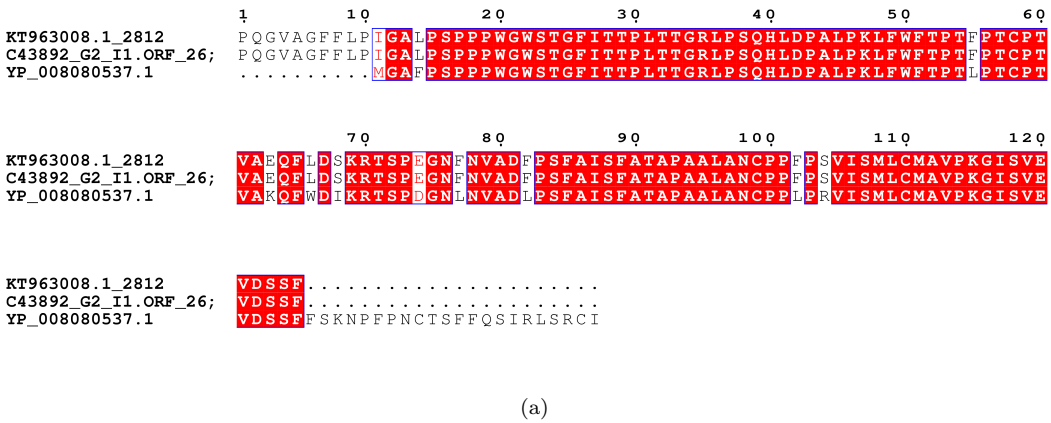
	ORF	ORF span		δ Protein span
Red oak	KT963008.1_7092	[75367 to 75224]		1 to 38
Quercus rubra	KT963008.1_6275	[104156 to 103881]	28932	39 to 116
124 aa	KT963008.1_6291	[103445 to 103326]	436	116 to 124
Chickpea	KT963008.1_6221	[106133 to 105825]		1 to 91
Cicer arietinum	KT963008.1_6242	[105141 to 105031]	684	109 to 123
212 aa	KT963008.1_6275	[104156 to 103881]	875	124 to 212

Table 3: **Expression counts of transcript C43892_G2_I1 which is identical to the ORF KT963008.1_2812**: This gene is not annotated, and is homologous to a RefSeq protein (*Pharus latifolius*, Accid:YP_008080537.1, Evalue=1E-67) These are raw counts, and are not normalized. The abbreviations are provided in FILE:20LIBS.pdf.

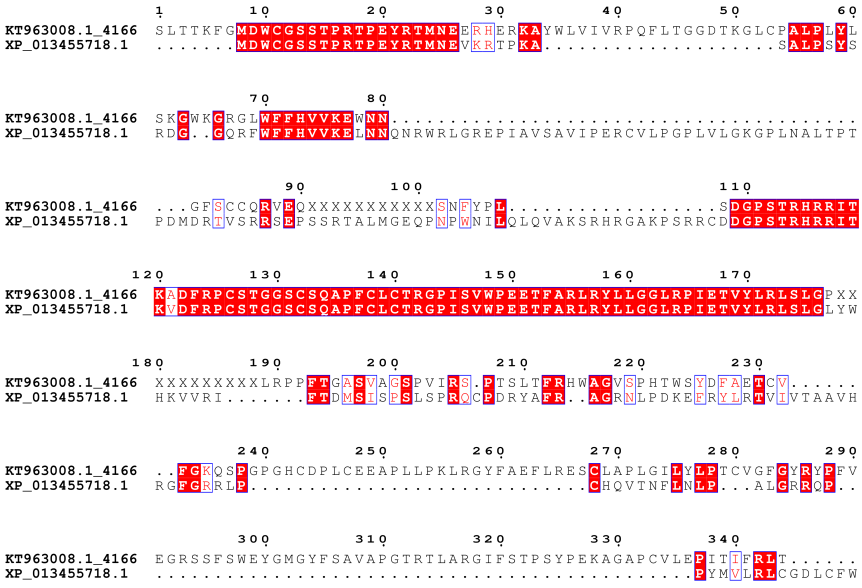
Transcript	CE	CI	CK	EM	FL	HC	HL	HP	HU	IF	LE	LM	LY	PK	PL	PT	RT	SE	TZ	VB
C43892_G2_I1	122	402	327	133	685	171	1086	265	359	292	2293	742	1859	226	707	291	505	118	540	564

Table 4: **Comparing predicted ORFs in the walnut chloroplast to the transcriptome:** It can be seen there are at least five bona-fide genes. TRS: transcript id from the RNA-seq transcriptome derived previously [2].

	ORF	RefSeq id	Description	TRS
1	KT963008.1_1240	ERN11400.1	hypothetical protein *A. trichopoda*	C41199_G1_I1
2	KT963008.1_4328	OIW21822.1	hypothetical protein *L. angustifolius*	C54203_G1_I1
	KT963008.1_6215	OIW21822.1	hypothetical protein *L. angustifolius*	C54203_G1_I1
3	KT963008.1_2812	YP_008080537.1	orf137 (chloroplast) *P. latifolius*	C43892_G2_I1
	KT963008.1_4790	YP_008080537.1	orf137 (chloroplast) *P. latifolius*	C43892_G2_I1
4	KT963008.1_2840	NP_817236.1	ORF58d *P. koraiensis*	C43892_G2_I2
	KT963008.1_4818	NP_817236.1	ORF58d *P. koraiensis*	C43892_G2_I2
5	KT963008.1_7519	DAA50039.1	TPA: hypothetical protein *Z. mays*	C62693_G1_I1



(a)



(b)

Figure 1: **Chloroplast genes that are not annotated in the current NCBI database:** (**a**) ORF KT963008.1_2812 (125 aa) is homologous to a RefSeq protein (*Pharus latifolius*, Accid:YP_008080537.1, Evalue=1E-67). Furthermore, this ORF is 100% identical to transcript C43892_G2_I1 from a previously derived transcriptome, establishing it as a bona fide gene. (**b**) Manually merging three ORFs to obtain an previously un-annotated gene: The three ORFs are KT963008.1_4166, KT963008.1_4182 and KT963008.1_4200, which have been merged manually by the insertion of a random number of "X". This is compared to a RefSeq gene from *Medicago truncatula* (Accid:XP_013455718.1), "encoded by transcript MTR_4g451295" (Evalue=4E-45). Thus, this is not a pseudo-gene, at least in *M. truncatula*.

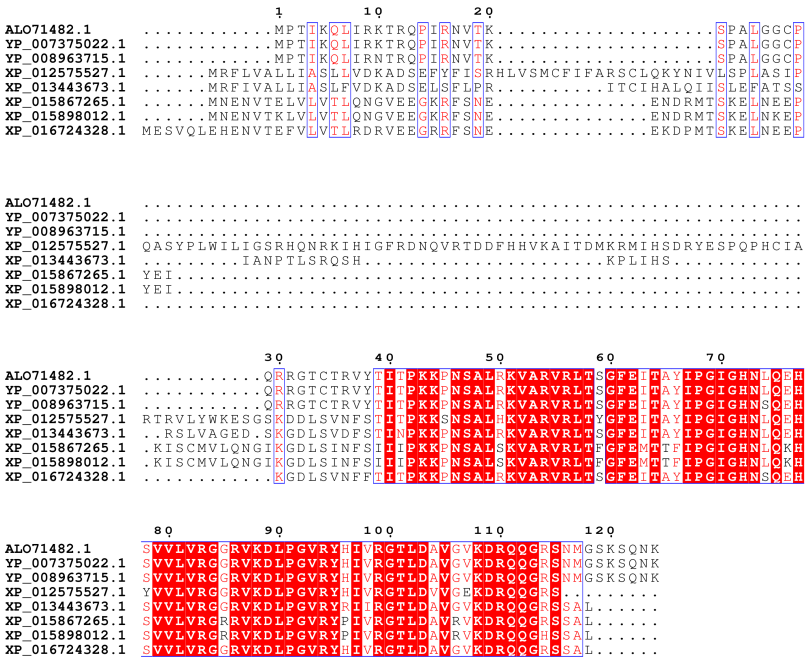


Figure 2: **Multiple sequence alignment of rps12 genes, both bona fide and mis-annotated**: rps12 is a trans-spliced gene with three exons - exon1 is ∼28kbp away from exon2 and exon3. ALO71482.1 , YP_007375022.1 and YP_008963715.1 are the bona fide rsp12 genes (124 aa long) with exon1 (38 aa long) ending in the sequence "RGTCTRVY". Mis-annotated genes have chosen ORFs prior to exon2 that lead to a theoretically possible gene.

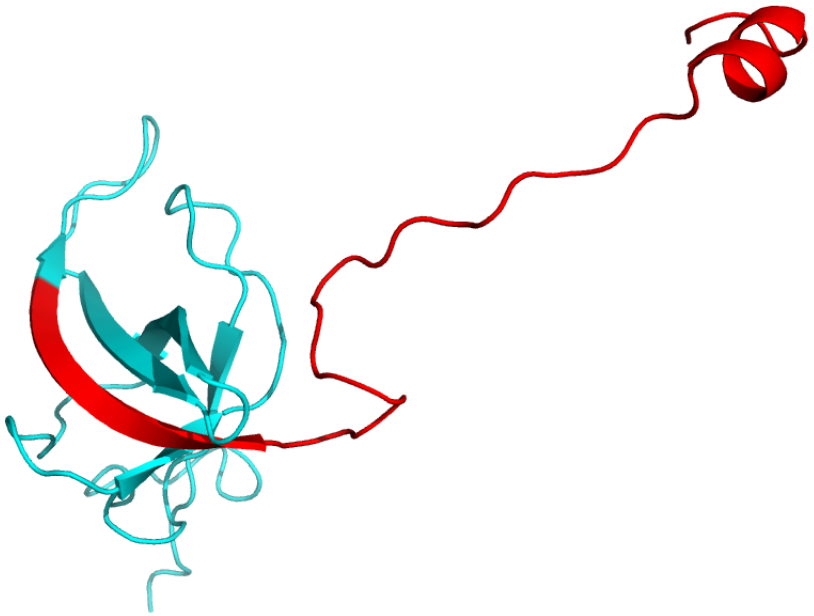


Figure 3: **Structure of rps12 from spinach chloroplast 30s subunit (PDBid:3BBN [12], chain L):** rps12 is a trans-spliced gene with three exons - exon1 is ∼28kbp away from exon2 and exon3. Exon1 is 38 aa long, and is marked in red.