RESEARCH ARTICLE

YeATSAM analysis of the walnut and chickpea transcriptome reveals key genes undetected by current annotation tools

[version 1; peer review: 2 approved, 1 not approved]

Sandeep Chakraborty, Pedro J. Martínez-García, Abhaya M. Dandekar

Department of Plant Sciences, University of California, Davis, USA

Abstract

Background: The transcriptome, a treasure trove of gene space information, remains severely under-used by current genome annotation methods. Methods: Here, we present an annotation method in the YeATS suite (YeATSAM), based on information encoded by the transcriptome, that demonstrates artifacts of the assembler, which must be addressed to achieve proper annotation. Results and Discussion: YeATSAM was applied to the transcriptome obtained from twenty walnut tissues and compared to MAKER-P annotation of the recently published walnut genome sequence (WGS). MAKER-P and YeATSAM both failed to annotate several hundred proteins found by the other. Although many of these unannotated proteins have repetitive sequences (possibly transposable elements), other crucial proteins were excluded by each method. An egg cell-secreted protein and a homer protein were undetected by YeATSAM, although these did not produce any transcripts. Importantly, MAKER-P failed to classify key photosynthesis-related proteins, which we show emanated from Trinity assembly artifacts potentially not handled by MAKER-P. Also, no proteins from the large berberine bridge enzyme (BBE) family were annotated by MAKER-P. BBE is implicated in biosynthesis of several alkaloids metabolites, like anti-microbial berberine. As further validation, YeATSAM identified ~1000 genes that are not annotated in the NCBI database by Gnomon. YeATSAM used a RNA-seq derived chickpea (Cicer arietinum L.) transcriptome assembled using Newbler v2.3.

Conclusions: Since the current version of YeATSAM does not have an ab initio module, we suggest a combined annotation scheme using both MAKER-P and YeATSAM to comprehensively and accurately annotate the WGS.

Keywords

RNA-seq, transcriptome, MAKER-P, genome annotation, berberine bridge enzyme, Trinity, walnut genome sequence
Introduction

The genome of a particular organism is static in all cells, unlike the dynamic transcriptome, which is the transcription of the gene space into RNA molecules in a fashion responsive to a variety of factors, such as developmental stage, tissue, and external stimuli. RNA-seq, a high-throughput RNA sequencing method, has radically transformed the identification of transcripts and quantification of transcriptional levels (Flintoft, 2008; Wang et al., 2009). It is supported by a diverse set of computational methods for analyzing the resulting data (Chakraborty et al., 2015; Chang et al., 2015; Chu et al., 2013; Fu et al., 2012; Grabherr et al., 2011; Loth et al., 2012; Mbandi et al., 2015; Schulz et al., 2012; Simpson et al., 2009; Trapnell et al., 2009; Trapnell et al., 2012; Wang et al., 2010; Zerbino & Birney, 2008).

Rapid advances in genome sequencing technologies have generated sequences for a deluge of organisms and species. The task of annotating these sequences has been addressed by several flows. These pipelines are categorized in http://omictools.com/genome-annotation-category and http://genometools.org/ and reviewed in (Yandell & Ence, 2012). Here, we focus specifically on MAKER-P (Campbell et al., 2014; Holt & Yandell, 2011; Law et al., 2015; Neale et al., 2014), which was used to annotate the recently published walnut genome sequence (WGS) (Martínez-García et al., 2016).

In the current study, the YeATS suite (Chakraborty et al., 2015) was enhanced to include genome annotation capabilities using RNA-seq-derived transcriptomes (YeATS annotation module - YeATSAM). First, the Trinity-assembled transcriptome obtained from twenty different tissues was compared to the WGS, excluding transcripts emanating from extraneous sources. This step incidentally revealed both biodiversity and plant-microbe interactions in walnut tree(s) from Davis, California (Chakraborty et al., 2016a). The WGS-derived transcripts were split into three open reading frames (ORFs), which were subjected to BLAST analysis using a plant proteome database obtained from the Ensembl database (Kersey et al., 2016). Transcripts can contain more than one significant ORF and must be handled differently depending on whether they map to the same or a different protein. The resulting analysis provided the WGS annotation.

Both MAKER-P and YeATSAM failed to annotate several hundred proteins annotated by the other. Many of the proteins had repetitive sequences or domains that, although difficult to detect, do not represent critical proteins during annotation. An egg cell-secreted protein (Sprunck et al., 2012), a copper chaperone (Shim et al., 2012), and a clavata3/ESR-Related protein (Kinoshita et al., 2007) were among the proteins not detected through the YeATSAM flow. Some proteins undetected in the MAKER-P flow are more significant in the context of a plant genome: several photosynthesis-related proteins encoded by the chloroplast (Nelson & Yocom, 2006) and the large family of FAD-binding berberine bridge enzymes (BBE) involved in biosynthesis of antimicrobial benzophenanthridines (Cheney, 1963; Winkler et al., 2008). We posited possible reasons for such exclusions and recommend incorporating both flows for comprehensive enumeration of genes in the WGS.

As further validation, YeATSAM was applied to chickpea (Cicer arrietinum L.), an important pulse crop with many nutritional and health benefits (Jukanti et al., 2012). The RNA-seq-derived transcriptome of chickpea has also been sequenced (Garg et al., 2011) and was processed through the YeATSAM pipeline to identify ~1000 proteins that are encoded by these transcripts, but are not annotated in the NCBI database, most of which were annotated using Gnomon (Souvorov et al., 2010).

Methods

The input to YeATSAM is a set of post-assembly transcripts (TRS) and the walnut genome sequence (WGS) (Figure 1). Transcripts that do not align to the WGS were removed (Chakraborty et al., 2016a). A BLAST database of protein peptides (plantpep.fasta: 1M sequences) using ~30 organisms (list.plants) from the Ensembl genome was created (Kersey et al., 2016). The three longest open reading frames (ORF), obtained using the ‘getorf’ utility in the EMBOSS suite (Rice et al., 2000), for each transcript in (TRS) underwent BLAST analysis (Camacho et al., 2013) to the ‘plantpep.fasta’. For cutoff E-value=1E-8, depending on the number of matches, the transcripts were clustered as:

1. None - either a previously unknown gene or non-coding RNA.
2. One - unique ORF.
3. Multiple ORFs matching to the same gene - merge the ORFs if the Evalue of the combined ORF is significantly lower.
4. Multiple ORFs matching to different genes - duplicate the transcripts, associating each transcript with a different ORF.

In vitro methods

Fifteen samples of walnut tissue were gathered from Chandler trees growing in the Stuke block at UC Davis between April and October 2008. Four additional samples were taken from Chandler plant material from the same orchard maintained in tissue culture. Several grams of leaf and root tissue from each plant were frozen in liquid nitrogen and then transferred to a -80 C freezer. RNA was isolated from each sample using the hot borate method (Wilkins & Smart, 1996) followed by purification and DNase treatment using an RNA/DNA Mini Kit (Qiagen, Valencia, CA) per the manufacturer’s protocol. High-quality RNA was confirmed by running an aliquot of each sample on an Experion Automated Electrophoresis System (Bio-Rad Laboratories, Hercules, CA). The cDNA libraries were constructed following the Illumina mRNA-sequencing sample preparation protocol (Illumina Inc., San Diego, CA). Final elution was performed with 16μL RNase-free water. The quality of each library was determined using a BioRad Experion (BioRad, Hercules, CA). Each library was run as an independent lane on a Genome Analyzer II (Illumina, San Diego, CA) to generate 85bp paired-end sequences from each cDNA library. Over a billion reads were obtained. Prior to assembly, all reads underwent quality control for paired-end reads and trimming using Sickle v1.33 (Joshi & Fass, 2011). The minimum read length was 45bp with a minimum Sanger quality score of 35. The quality-controlled reads were de novo assembled
First, transcripts from extraneous organisms are pruned. Next, the three longest open reading frames (ORFs) from each transcript undergo BLAST analysis to a database of plant peptides. Depending on the number of significant matches, the transcripts are clustered as: (a) None - either a previously unknown gene, or non-coding RNA. (b) One - Unique ORF (c) Multiple ORFs matching to the same gene - merge the ORFs if the Evalue of the combined ORF is significantly lower. (d) Multiple ORFs matching to different genes - duplicate the transcripts, associating each with a different ORF. Subsequently, the ORFs are merged based on overlapping amino acid sequences and exact substrings are removed.

Figure 1. YeATSAM flow. First, transcripts from extraneous organisms are pruned. Next, the three longest open reading frames (ORFs) from each transcript undergo BLAST analysis to a database of plant peptides. Depending on the number of significant matches, the transcripts are clustered as: (a) None - either a previously unknown gene, or non-coding RNA. (b) One - Unique ORF (c) Multiple ORFs matching to the same gene - merge the ORFs if the Evalue of the combined ORF is significantly lower. (d) Multiple ORFs matching to different genes - duplicate the transcripts, associating each with a different ORF. Subsequently, the ORFs are merged based on overlapping amino acid sequences and exact substrings are removed.

with Trinity v2.0.6 (Grabherr et al., 2011). Standard parameters were used and the minimum contig length was 300bp. Individual assemblies for each library and a combined assembly of all tissues were performed.

The walnut genome sequence has been released to the public domain (http://ucanr.edu/sites/wgig/). The Illumina (Genome Analyzer II) for all 20 tissues can be accessed at http://www.ncbi.nlm.nih.gov/sra/PRJNA232394.

The transcriptome of Cicer arietinum (transHybrid.fasta, ICC4958; Desi chickpea) was obtained from http://www.nipgr.res.in/ctdb.htm (Garg et al., 2011). The dataset ‘represents optimized de novo hybrid assembly of 454 and short-read sequence data.’ About two million 454 reads were assembled using Newbler v2.3 followed by hybrid assembly with 53409 transcripts generated by optimized short-read data assembly using TGICL, as reported previously (Garg et al., 2011). The set of annotated proteins from chickpea was obtained from the NCBI database (chickpea.pep.fasta, N=34198).

PHYML v3.0 was used to generate phylogenetic trees from alignments (Guindon et al., 2005). Multiple sequence alignment was done using ClustalW (Larkin et al., 2007) and figures were generated using the ENDscript server 2.0 (Robert & Gouet, 2014). The source code written in Perl is provided as Dataset 1 (YeATSAM.tgz). A README is provided inside the top-level directory for installation and running the programs.
Results and discussion

The input to YeATSAM was ~111K Trinity-assembled transcripts (Combined TrinityFull.fasta) (Figure 1). Each transcript was aligned to the WGS (wgs.5d_scafSeq200+.trimmed) using BLAST (Camacho et al., 2013). Transcripts that did not align to the WGS (cutoff BLAST bitscore=75) were excluded (Chakraborty et al., 2016a). Those transcripts that aligned to the WGS (list.transcriptome-clean: 106K) were split into the three longest open reading frames (ORF) (list.transcriptome.clean.ORFS: 320K).

A BLAST database of protein peptides (plantpep.fasta:1M sequences) using ~30 organisms (list.plants) from the Ensembl genome was created (Kersey et al., 2016). The availability of proteomes from related organisms accelerates the annotation. The BLAST results of list.transcriptome.clean.ORFS: 320K on ‘plantpep.fasta’ was processed using a cutoff: bitscore=60, Evalue~<1E-10.

Merging ORFs: broken transcripts

There are two instances in which ORFs can be merged to create a longer amino acid sequence. The first scenario occurs when a particular transcript has multiple ORFs that match to the same protein with high significance, indicating that a sequencing or assembly error has broken a contiguous ORF (Chakraborty et al., 2015). In total, 5% of the present transcripts (5,000 of 106,000) had two or more ORFs matching with high significance to the same protein, exactly mirroring the 5% error rates seen in transcripts restricted to the transcriptome from the tissue at the heartwood/sapwood transition zone in black walnut (Chakraborty et al., 2015). While most of these transcripts have repetitive elements, there were other non-repetitive sequences with this particular problem. C20727_G1_I1 is one example: it has two ORFS, ORF_15 and ORF_36, that match a DNA repair metallo-β-lactamase family protein (Accession number: XP007043420.1) with Evalues=9E-70 and 6E-96, respectively (Figure 2a). The two ORFs were merged (inserting the sequence ‘ZZZ’, although the length of the missing fragment is not known) since the Evaluate of the combined ORF reduces to 2E-175 and the merged sequence was chosen as representative for the transcript. ORFs are not merged when the combined ORF did not significantly decrease the Evaluate and the longer ORF was selected to represent the transcript.

The other scenario occurs when the assembler fails to merge two transcripts into a single one. In this instance, two ORFs emanating from different transcripts with contiguous sequences were merged. While the merging of two ORFs was described previously (Chakraborty et al., 2015), we introduced an additional filter to select mergeable ORFs based on whether the E-value obtained by merging the two ORFs is significantly reduced. For example, transcripts C53209_G8_I1 and C53209_G6_I1 both map to the scaffold SUPER472 and their corresponding ORFs can be merged based on the sequence string ‘PNRSSLP’ (Figure 2b). The individual ORFs and the combined ORFs align to an autophagy-related protein (TAIR ID: AT3G49590.2) with Evaluues=2e-106, 8e-63, and 1e-180, respectively. The increased significance of the combined ORF, in addition to other checks, like ensuring that mapping is to the same scaffold, adds further support to the fact that these transcripts should have been contiguous in the final assembled transcriptome.

Transcripts with multiple ORFs

About 3% of transcripts have ORFs that map to different proteins. Some transcripts should not have been merged. C1089_G1_I1 is an interesting example: a 4574 nt transcript that maps to the chloroplast and encodes two genes. One is highly variable and the other is conserved. The two ORFS, ORF_64 (fwd: 1117-2631) and ORF_108 (fwd: 3195 - 4271), map to maturase K (TAIR ID: ATCG00040.1) and photosystem II reaction center protein (TAIR ID: ATCG00020.1) with very high significance. Maturase K is a good candidate for barcoding angiosperms because it has highly variable coding sequences (Yu et al., 2011), while the photosystem II reaction center protein is completely conserved (100% similarity with Arabidopsis). Another example is C19241_G1_I1 (4702 nt), split into ORF_68 (fwd: 176-3487) and ORF_115 (reverse: 4509-4096) encoding a damaged DNA binding protein (TAIR ID: AT4G05420.1) and photosystem I subunit K (TAIR ID: AT1G30380.1) with high significance, respectively. These transcripts are split in the YeATSAM flow, resulting in one ORF per transcript. Subsequently, this artifact of the Trinity assembly led to several unannotated proteins in the MAKER-P flow.

Identifying genes not detected by either YeATSAM or MAKER-P

We compared the annotations of walnut by MAKER-P (walnut. wgs.5d.all.maker.proteins.fasta) and YeATSAM (DB.ORFBEST60). MAKER-P and YeATSAM each failed to annotate several proteins identified by the other (MAKER-P=~4000; YeATSAM=700). Although most of these unannotated proteins have repetitive sequences (transposable elements), some vital, non-repetitive proteins were excluded by each method. For example, an egg cell-secreted protein (‘WALNUT 00001389-RA’) (Sprunck et al., 2012), a Clavata3/esr-related gene (‘WALNUT 00023705-RA’) (Kinoshita et al., 2007) and a copper chaperone (‘WALNUT 00006344-RA’) (Shin et al., 2012) were not annotated in the YeATSAM flow. These genes do not have transcripts in the twenty tissues analyzed in the current study and are most likely pseudogenes.

Proteins unannotated by MAKER-P

MAKER-P fails to annotate many key photosystem-related proteins (Table 1). The transcript C59245_G1_I1 has ORF_43 (fwd: 176-1714) and ORF_70 (fwd: 2212-2496) mapping to photosystem II reaction center protein B (PSBB) and photosystem II reaction center protein H (PSBH), respectively. While MAKER-P does annotate PSBB, it failed to detect PSBH. These proteins map to transcripts encoding two significant ORFs (>1E-10), indicating that failure to handle this might have excluded these proteins. Also, these proteins are encoded by the chloroplast. However, this limitation of MAKER-P is not confined to transcripts emanating from the chloroplast. For example, C48031_G3_I1 encodes a leucine-rich repeat transmembrane protein kinase (AT5G48940.1) and a metallo-β-lactamase family protein (TAIR ID: AT4G33540.1) and is mapped to scaffold ‘SUPER374’. MAKER-P failed to annotate the β-lactamase family protein.

Furthermore, MAKER-P failed to annotate any FAD-binding berberine bridge enzymes (BBE) in the WGS (Kutchan & Dittrich, 1995). These enigmatic enzymes are implicated in the transformation of (S)-reticuline to (S)-scoulerine during
Open reading frames (ORF) that can be merged. (a) ORFs from the same transcript: C20727_G1_I1 has two ORFs (ORF 15 and ORF 36) matching to a DNA repair metallo-β-lactamase family protein (Accession number: XP007043420.1) with high significance. We merged the two ORFs (inserting ‘ZZZ’) since the Evalue of the combined ORF is significantly reduced. (b) ORFs from different transcripts: We merged ORFs from two different transcripts (C53209_G8_I1 and C53209_G6_I1), since both transcripts map to the same scaffold (SUPER472) can be overlapped based on the sequence string ‘PNRSSLP’, and the merged ORF has a significantly reduced Evalue.
Table 1. Key photosystem-related proteins in the chloroplast not annotated by MAKER-P and YeATSAM. These transcripts have multiple open reading frames (ORFs) mapping to different proteins with high significance. For example, C59245_G1_I1 has another ORF (43) which maps to photosystem II reaction center protein B (PSBB). MAKER-P annotates PSBB, but not PSBH. These transcripts all emanate from the chloroplast, although not all genes that MAKER-P failed to annotate were from the chloroplast. Genes predicted by MAKER-P that are not identified by YeATSAM are listed with their homology to corresponding genes in the TAIR database.

<table>
<thead>
<tr>
<th>TRS</th>
<th>ORF</th>
<th>Len</th>
<th>TAIR</th>
<th>Description</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C52274_G4_I1_B</td>
<td>189</td>
<td>231</td>
<td>ATCG00720.1</td>
<td>PETB photosynthetic electron transfer B</td>
<td>4.00e-155</td>
</tr>
<tr>
<td>C52274_G4_I1_C</td>
<td>231</td>
<td>177</td>
<td>ATCG00730.1</td>
<td>PETD photosynthetic electron transfer D</td>
<td>1.00e-108</td>
</tr>
<tr>
<td>C53854_G1_I1_A</td>
<td>45</td>
<td>98</td>
<td>ATCG00070.1</td>
<td>PSBK photosystem II reaction center protein K precursor</td>
<td>1.00e-27</td>
</tr>
<tr>
<td>C53854_G1_I1_B</td>
<td>62</td>
<td>62</td>
<td>ATCG00080.1</td>
<td>PSBI photosystem II reaction center protein I</td>
<td>3.00e-20</td>
</tr>
<tr>
<td>C54343_G2_I1_A</td>
<td>8</td>
<td>91</td>
<td>ATCG00580.1</td>
<td>PSBE photosystem II reaction center protein E</td>
<td>4.00e-54</td>
</tr>
<tr>
<td>C59245_G1_I1_B</td>
<td>70</td>
<td>95</td>
<td>ATCG00710.1</td>
<td>PSBH photosystem II reaction center protein H</td>
<td>4.00e-43</td>
</tr>
<tr>
<td>WALNUT_00014004-RA</td>
<td>-</td>
<td>1117</td>
<td>AT5G16850.1</td>
<td>TERT Telomerase reverse transcriptase</td>
<td>0.0</td>
</tr>
<tr>
<td>WALNUT_00018632-RA</td>
<td>-</td>
<td>295</td>
<td>ATMG00560.1</td>
<td>RPL2 Nucleic acid-binding, OB-fold-like protein</td>
<td>9e-152</td>
</tr>
<tr>
<td>WALNUT_00019747-RA</td>
<td>-</td>
<td>326</td>
<td>AT1G24040.1</td>
<td>Acyl-CoA N-acetyltransferases (NAT) superfamily protein</td>
<td>5e-121</td>
</tr>
<tr>
<td>WALNUT_00031866-RA</td>
<td>-</td>
<td>311</td>
<td>AT5G07810.1</td>
<td>SNF2 domain-containing protein/helicase domain-containing</td>
<td>9e-115</td>
</tr>
<tr>
<td>WALNUT_00020600-RA</td>
<td>-</td>
<td>155</td>
<td>ATG01240.1</td>
<td>RPS7.2 ribosomal protein S7 chrC:140704-141171</td>
<td>1e-108</td>
</tr>
<tr>
<td>WALNUT_00016414-RA</td>
<td>-</td>
<td>231</td>
<td>AT5G14850.1</td>
<td>alpha/beta-Hydrolases superfamily protein chr5:16756698-16757791</td>
<td>6e-96</td>
</tr>
<tr>
<td>WALNUT_00027509-RA</td>
<td>-</td>
<td>289</td>
<td>AT2G43190.3</td>
<td>ribonuclease P family protein chr2:17956220-17957833</td>
<td>2e-94</td>
</tr>
<tr>
<td>WALNUT_00022174-RA</td>
<td>-</td>
<td>389</td>
<td>AT2G07707.1</td>
<td>Plant mitochondrial ATPase, F0 complex, subunit</td>
<td>5e-86</td>
</tr>
<tr>
<td>WALNUT_00018616-RA</td>
<td>-</td>
<td>124</td>
<td>ATCG00890.1</td>
<td>NDH6.1 NADH-Ubiquinone/plastoquinone (complex I)</td>
<td>1e-79</td>
</tr>
<tr>
<td>WALNUT_00003702-RA</td>
<td>-</td>
<td>924</td>
<td>AT5G14990.1</td>
<td>BEST Arabidopsis thaliana protein match is: myosin</td>
<td>2e-79</td>
</tr>
</tbody>
</table>

benzophenanthridine alkaloid biosynthesis in plants (Winkler et al., 2006). This pathway is over-expressed upon osmotic stress and pathogen attack (Attila et al., 2008; González-Candelas et al., 2010), provides resistance in lettuce, sunflower and transgenic tobacco by generating anti-microbial compounds (Custers et al., 2004), and has unknown functions at specific developmental stages in Arabidopsis (Irshad et al., 2008; Pagnussat et al., 2005). Moreover, it is expressed in floral nectar (Nectarin V, NBBE) (Carter & Thornburg, 2004) and roots of tobacco (Kajikawa et al., 2011), and in xylem sap of cabbage (Ligat et al., 2011) and grapevine (Chakraborty et al., 2016b). NBBE was constitutively expressed in the Phytophthora infestans-resistant potato genotype SW93-1015 (Ali et al., 2012). Benzophenanthridines are antimicrobial; the California poppy (Eschscholzia californica) is used as a traditional medicine (Cheney, 1963; Oldham et al., 2010). Oral administration of the alkaloid berberine isolated from a Chinese herb lowered cholesterol in 32 hypercholesterolemic patients over three months (Kong et al., 2004). Berberine has also been shown to possess antidiabetic properties (Lee et al., 2006).

The number of BBE genes in different plant species varies significantly from one in moss (Physcomitrella patens) to 64 in western poplar (Populus trichocarpa) (Daniel et al., 2015). A. thaliana has 27 TAIR IDs assigned to BBE enzymes, with two splice variants (Supplementary Table 1) (Daniel et al., 2015). Based on the current transcriptome, there are four full length BBE genes (JrBBE1 to 4) that map to different scaffolds in the WGS, in addition to other fragmented transcripts (Table 2 and Table 3). JrBBE1 (C54052_G1_I1) maps to the scaffold JCF7180001213852 and encodes a 564 aa long ORF, which has significant matches to Uniprot:P30986. The closest match of Uniprot:P30986 (with a low significance of 1E-07) to the MAKER-P annotation is ‘WALNUT_00019599-RA’, a 476 aa long cytokinin dehydrogenase. The sequence alignment of JrBBE genes to Uniprot (P30986) is shown (Figure 3a).

As with the walnut transcriptome, the chickpea transcriptome (transHybrid.fasta: n=34760) (Garg et al., 2011) was split into three ORFs, each of which was BLAST-ed to the subset of plant proteins created from the Ensembl database. Subsequently, the ORFs with significant homology to this database (n=29263) were BLAST-ed to the set of annotated chickpea proteins in the NCBI database (n=34198). Most of these annotations were done using Gnomon (Souverov et al., 2010) (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA190909), which analyzed ~35000 transcripts. There are ~1500 proteins identified by YeATSAM that are absent in the NCBI database (Evalue cutoff 1E-10). Some of these proteins and their corresponding genes in the TAIR database are shown (Table 4). TC00902 is an interesting example with two merged genes: a hydrogen ion-transporting ATP synthase (TAIR ID: ATMG00640.1) and a cytochrome C biogenesis (TAIR ID: ATMG00900.1). While Gnomon identified the cytochrome C
Table 2. FAD-binding berberine bridge enzymes (BBE) are undetected in MAKER-P. These oxidases are involved in the benzophenanthridine alkaloid biosynthesis in plants. *Arabidopsis* has 27 loci for this family (and a splice variant) (Table 3). Here, there are four full length berberine bridge enzyme (BBE) genes (named JrBBE1-4) identified using the transcriptome. Some of the proteins are truncated (like C54286_G1_I1), which might be an artifact of the Trinity assembler. Thus, this is not a complete enumeration of the JrBBE genes.

<table>
<thead>
<tr>
<th>Id</th>
<th>Transcript</th>
<th>Length</th>
<th>Scaffold</th>
<th>ORF</th>
<th>TAIR Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>JrBBE1</td>
<td>C54052_G1_I1</td>
<td>564</td>
<td>JCF7180001213852</td>
<td>34</td>
<td>AT1G26420.1</td>
</tr>
<tr>
<td>JrBBE2</td>
<td>C53871_G1_I1</td>
<td>564</td>
<td>JCF7180001217410</td>
<td>28</td>
<td>AT1G30700.1</td>
</tr>
<tr>
<td>JrBBE3</td>
<td>C55152_G1_I1</td>
<td>552</td>
<td>JCF7180001222284:2429142-2890931</td>
<td>37</td>
<td>AT4G20820.1</td>
</tr>
<tr>
<td>JrBBE4</td>
<td>C7952_G1_I1</td>
<td>559</td>
<td>JCF7180001218369</td>
<td>110</td>
<td>AT2G34790.1</td>
</tr>
<tr>
<td></td>
<td>C54286_G2_I1</td>
<td>307</td>
<td>JCF7180001217076</td>
<td>35</td>
<td>AT1G11770.1</td>
</tr>
<tr>
<td></td>
<td>C54286_G1_I1</td>
<td>128</td>
<td>JCF7180001217076</td>
<td>7</td>
<td>AT4G20830.1</td>
</tr>
<tr>
<td></td>
<td>C12765_G1_I1</td>
<td>114</td>
<td>JCF7180001218369</td>
<td>8</td>
<td>AT4G20840.1</td>
</tr>
<tr>
<td></td>
<td>C51815_G1_I4</td>
<td>168</td>
<td>JCF7180001218369</td>
<td>29</td>
<td>AT4G20860.1</td>
</tr>
</tbody>
</table>

Table 3. Expression counts (normalized) of transcripts from the FAD-binding berberine bridge enzyme (BBE) family. The genes have tissue-specific expression - JrBBE3 is highly expressed in the roots and transition zone. The tissue abbreviations are from Chakraborty et al., 2016a.

<table>
<thead>
<tr>
<th>Id</th>
<th>Transcript</th>
<th>CE</th>
<th>CI</th>
<th>CK</th>
<th>EM</th>
<th>FL</th>
<th>HC</th>
<th>HL</th>
<th>HP</th>
<th>HU</th>
<th>IF</th>
<th>LE</th>
<th>LM</th>
<th>LY</th>
<th>PK</th>
<th>PL</th>
<th>PT</th>
<th>RT</th>
<th>SE</th>
<th>TZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>JrBBE1</td>
<td>C54052_G1_I1</td>
<td>44</td>
<td>4</td>
<td>136</td>
<td>197</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JrBBE2</td>
<td>C53871_G1_I1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>79</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JrBBE3</td>
<td>C55152_G1_I1</td>
<td>43</td>
<td>34</td>
<td>25</td>
<td>62</td>
<td>1</td>
<td>2</td>
<td>35</td>
<td>1040</td>
<td>346</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JrBBE4</td>
<td>C7952_G1_I1</td>
<td>32</td>
<td>85</td>
<td>8</td>
<td>55</td>
<td>11</td>
<td>711</td>
<td>15</td>
<td>8</td>
<td>241</td>
<td>137</td>
<td>37</td>
<td>123</td>
<td>315</td>
<td>420</td>
<td>160</td>
<td>217</td>
<td>5</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C54286_G2_I1</td>
<td>33</td>
<td>33</td>
<td>19</td>
<td>23</td>
<td>23</td>
<td>19</td>
<td>5</td>
<td>22</td>
<td>9</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

biogenesis protein (Genbank: XP_004500083.1), it failed to identify the ATP synthase. Unlike MAKER-P, Gnomon generates transcripts through predictive algorithms and does not take the transcriptome as an input. Notwithstanding, these chickpea genes remain unannotated despite the presence of a straightforward method to detect them from available transcripts.

**Future work**

Among the ~700 genes not detected by YeATSAM, there are ~500 genes with no matches in the complete ‘nr’ database. Of these, ~300 have no transcripts (SetA), while the remaining ~200 have matches among the transcripts (SetB). Considering the sensitivity of RNA-seq and the wide coverage of twenty tissues, it is a definite possibility that SetA are pseudogenes. Future work in YeATSAM will focus on methods to distinguish these two classes of genes.

**Conclusions**

The availability of a RNA-seq-derived transcriptome from a newly sequenced organism like walnut, for which there are related annotated genomes (Arabidopsis, Vitis, etc), immensely simplifies annotation of the genome and influences the choice of annotation software. Here, we introduce a new annotation method in the YeATS suite (YeATS Annotation Module - YeATSAM), which was used to annotate the newly-sequenced walnut genome using a simple workstation. The key differentiating factor in YeATSAM is the splitting of the assembled transcriptome into multiple ORFs (Chakraborty et al., 2015). Transcripts often have more than one significant ORF that must be handled differently depending on whether they map to the same or different proteins. We show that YeATSAM failed to annotate ~700 genes identified by MAKER-P, while identifying ~4000 genes missed by MAKER-P. While most of these genes have repetitive stretches, both methods missed vital genes identified by the other. Since many of the additional genes identified by MAKER-P have no known transcripts, we posit that these were identified using *ab initio* methods. In the absence of such an *ab initio* module in YeATSAM, we propose a combined method using both MAKER-P and YeATSAM to annotate the WGS. YeATSAM was also applied to the chickpea transcriptome and identified ~1000 proteins that are not annotated in the NCBI database. This transcriptome was assembled using Newbler v2.3 (Garg et al., 2011) and most of the 34198 chickpea proteins in the NCBI database were annotated using Gnomon, the standard annotation tool (http://www.ncbi.nlm.nih.govgenome/guide/gnomon.shtml).
Figure 3. Multiple sequence alignment of BBE from walnut and other organisms. (a) The JrBBE sequences were aligned to berberine bridge enzyme (BBE) genes from *Eschscholzia californica* (EcBBE; California poppy), *Arabidopsis thaliana* (AtBBE15) and *Nicotiana tabacum* (Nectarin V). Secondary structure information from the structure PDBid:3D2D (*E. californica*) was used to annotate the sequences. The signal peptides are different in these proteins, suggesting different localization of these proteins in walnut. (b) Phylogenetic tree generated from the multiple sequence alignment.
Table 4. Selected genes in chickpea that are not annotated in the NCBI database. Most of the NCBI genes were predicted using Gnomon. YeATSAM used the publicly available transcriptome from chickpea to identify these genes. The corresponding genes from the TAIR database are shown. Several transcripts (like TC20962) encode multiple genes, while others (like TC01181) have only one significant ORF. TRid, transcription id; TAIRid: Arabidopsis thaliana id.

<table>
<thead>
<tr>
<th>TRid</th>
<th>TAIRid</th>
<th>Description</th>
<th>Evalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC20962 A</td>
<td>ATMG00070.1</td>
<td>NAD9 NADH dehydrogenase subunit 9 chrM:23663-24235</td>
<td>3e-116</td>
</tr>
<tr>
<td>TC20962 B</td>
<td>AT2G07687.1</td>
<td>Cytochrome c oxidase, subunit III chr2:3311854-3312651</td>
<td>3e-107</td>
</tr>
<tr>
<td>TC20962 C</td>
<td>AT2G07674.1</td>
<td>Unknown conserved protein chr2:3269151-3269906</td>
<td>6e-41</td>
</tr>
<tr>
<td>TC01181</td>
<td>ATMG01360.1</td>
<td>COX1 cytochrome oxidase chrM:349830-351413</td>
<td>0.0</td>
</tr>
<tr>
<td>TC11063</td>
<td>AT3G30841.1</td>
<td>Cofactor-independent phosphoglycerate mutase chr3:12591595-12593401</td>
<td>0.0</td>
</tr>
<tr>
<td>TC06038</td>
<td>ATMG00090.1</td>
<td>Structural constituent of ribosome;protein binding chrM:25482-28733</td>
<td>3e-124</td>
</tr>
<tr>
<td>TC13206</td>
<td>AT3G13440.1</td>
<td>S-adenosyl-L-methionine-dependent methyltransferases superfamily</td>
<td>1e-118</td>
</tr>
<tr>
<td>TC07586</td>
<td>AT2G07725.1</td>
<td>Ribosomal L5P family protein chr2:3448402-3448959</td>
<td>2e-113</td>
</tr>
<tr>
<td>TC19047</td>
<td>ATMG00570.1</td>
<td>Sec-independent periplasmic protein translocase</td>
<td>8e-107</td>
</tr>
<tr>
<td>TC09002 B</td>
<td>ATMG00640.1</td>
<td>Hydrogen ion transporting ATP synthases, rotational</td>
<td>3e-104</td>
</tr>
<tr>
<td>TC15163</td>
<td>AT4G28360.1</td>
<td>Ribosomal protein L22p/L17e family protein chr4:14029294-14030926</td>
<td>1e-100</td>
</tr>
<tr>
<td>TC13677</td>
<td>AT5G05210.1</td>
<td>Surfeit locus protein 6 chr5:1548198-1549534</td>
<td>9e-91</td>
</tr>
<tr>
<td>TC13780 A</td>
<td>AT2G07707.1</td>
<td>Plant mitochondrial ATPase, F0 complex, subunit 8 protein</td>
<td>2e-90</td>
</tr>
<tr>
<td>TC18786</td>
<td>AT1G73440.1</td>
<td>Calmodulin-related chr1:27611418-27612182</td>
<td>5e-45</td>
</tr>
</tbody>
</table>

Software availability
Latest source code: [https://github.com/sanchak/YeATSAM](https://github.com/sanchak/YeATSAM)


License: GNU General Public License

**Author contributions**
AMD and SC were involved in the study design. SC developed the software and designed pipeline that enabled the annotations, PJM-G was involved in the validation with the walnut genome sequence. SC wrote the first draft and the rest of the authors were involved in subsequent editing and modifications.

**Competing interests**
No competing interests were disclosed.

**Grant information**
AMD wishes to acknowledge grant support from the California Walnut Board.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Acknowledgements**
The authors wish to thank the California Walnut Board for their generous support. We appreciate the efforts of Mary Lou Mendum in editing this manuscript.

**Supplementary material**
Supplementary Table 1. FAD-binding (Berberine family) reticuline oxidases in *Arabidopsis*. These proteins are involved in the benzophenanthridine alkaloid biosynthesis in plants. Benzophenanthridine is an anti-microbial metabolite. There are 27 loci in chromosomes 1,2,4 and 5, with one splice variant (AT4G20830) marked with an asterisk.

Click here to access the data.


The article presents an annotation method, YeATSAM, that leverages the information contained in RNA-Seq derived transcriptomes. The method was compared with two other annotation methods using two organisms: MAKER-P (a RNA evidence based and \textit{ab initio} hybrid method) with walnut, and NCBI Gnomon (a homology based and \textit{ab initio} hybrid method) with chickpea. Although YeATSAM and MAKER-P identified same genes, there were also genes that were identified by only one of them (about 4,000 by YeATSAM and about 700 by MAKER-P), as well as genes that both methods failed to identify. Similarly, YeATSAM identified about 1,000 genes that Gnomon failed to identify.

The article is well written, the analysis is technically sound, the tables and figures present the results well, and the conclusions are supported by the data. Nonetheless, I would suggest the following changes:

- Address discrepancies in the numbers reported, e.g., 20 tissues (in abstract, introduction, results, and future work) v. 19 tissues (in methods: 15 samples + four additional samples); 700 (in results) v. \textasciitilde700 (in future work, and conclusions); \textasciitilde1,500 chickpea proteins (in results) v. \textasciitilde1,000 (in conclusions). Instead of approximate values report actual values.

- As the tool is designed to be used with other organisms (besides walnut and chickpea), make the method and workflow (Figure 1) independent of any organism (e.g., the input to YeATSAM is the genome sequence rather than the walnut genome).

- Figure 3 (b) can be resized without losing its readability.

\textit{Competing Interests:} No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
This work focuses on a current major challenge in improving genome and transcriptome automated annotation. It also deals with difficulties derived from imperfect de novo assemblies, such as transcripts representing fused and split genes. The increasing affordability to generate sequencing data enhances the demand for more powerful annotation predicting tools and pipelines, although exact annotations will still remain for wet-lab experimentation. This paper compares the YeATSAM tool to previously annotated genomes, in which existing de novo assemblies are used as generated and analyzed with blast, interproscan or similar tools for homology-based annotation. Even though the paper indicates novelty of the method, there are critical points that need modification.

The method reported here - YeATSAM - is not clearly different from the work already reported in a previous paper\(^1\). This method reported here (identify 3 longest ORFs, then blast to known proteins, then merge or split if needed) looks identical to previously published in F1000 Research\(^1\) - for instance, Figure 1 in the previously published paper is an almost identical replica of the Figure 1 in this paper. The current work does appropriately cite this previous paper. However, if there is a novel algorithm to describe here, it needs to be clearly delineated from this previous work. Otherwise, it should just be cited.

This previous publication also compares the annotation of the walnut genome by YeATS and Maker-P. The previous paper and this paper both profile walnut transcripts where ORFs were merged and transcripts that match multiple proteins; this paper does use different transcripts to demonstrate the methodology and results. To emphasize the novelty of the present paper, the authors should clarify exactly what this paper offers in addition to the previous paper. In this regard, the paper does go a bit further than the previous one by detailing genes that were unannotated by MAKER-P but found via this method; those genes were not reported previously. If the algorithm has not changed from the previous work, a new focus for this paper is needed, possibly reporting these novel genes such as the BBEs.

The addition of the chickpea genome annotation is barely described - a single short results paragraph. The author also has an existing F1000 research article describing the use of YeATS on chickpea transcripts and describing the detection of missed genes and describing multiple ORFs mapping to different proteins and fragmented ORFs of the same protein\(^2\). How does this report differ from that one? That one is not cited in this report.

Data reproducibility and accessibility - the new annotations are not made available for either walnut or chickpea (unless they are the same as the ones provided already in Chakraborty et al. 2015\(^1\)). It would be very difficult to replicate this experiment. No parameters or commands are provided to determine how
PHYML, ClustalW or ENDscript server were utilized. I confirmed that YeATSAM.zip (listed as YeATSAM.tgz in manuscript) with README is available for download and the links to data are functional. I was unable to install YeATSAM; the installation and usage instructions are very vague.

Specifics:

The joined results of MAKER-P and YeATSAM look promising for improving genome annotations. However, a figure or table describing the total number of genes predicted by each software and the overlap would be very helpful to visualize the results. The report commonly has words like “several” or “many” and the usage of “~” in front of numbers. Numbers should be reported exactly where they are important to the method and results. Examples:

- “A BLAST database of protein peptides (plantpep.fasta: 1M sequences) using ~30 organisms (list.plants)” - also list.plants does not link to anything.
- “About 3% of transcripts have ORFs that map to different proteins”
- “MAKER-P and YeATSAM each failed to annotate several proteins identified by the other (MAKER-P=4000; YeATSAM=700)”
- “Among the ~700 genes not detected by YeATSAM, there are ~500 genes with no matches in the complete ‘nr’ database. Of these, ~300 have no transcripts (SetA), while the remaining ~200 have matches among the transcripts (SetB).”

Based on the content of the manuscript, the introduction focuses adequately on the explanation of the problematic of annotating newly assembled genomes and transcriptomes. However, a deeper introduction to the software utilized may be relevant for a better understanding of their choice and also of their basic mechanics. In relation to the results commented on the introduction, the relevance of some of the selected genes is not clear. Specifically, the relevance of the three “critical” proteins not detected by YeATSAM, which are not transcribed and are thus considered pseudogenes, is confusing.

In relation to the generation of de novo assemblies, the authors are suggested to provide detail on how the assemblies were combined, considering that the simple addition of libraries would lead to high redundancy. For the walnut genome, were the MAKER-P and YeATSAM packages using the same set of RNASeq reads? This would be an important point to emphasize - a true comparison of the two methods would preferably use the same starting point. The original walnut paper reports using 19 libraries (Martinez-Garcia et al. 2016); this paper reports 20 libraries.

In the results and discussion section, the manipulation of ORFs is an interesting concept, although the difference to the described methodology in Chakraborty et al. 2015 is not clear. The use of the term ORF is confusing here since it appears that the merged sequences are the encoded peptides, while ORFs are nucleotide sequences. Moreover, it seems likely the ORFs from the same gene might match different proteins because they are being compared to 30 different organisms. The ORFs could match to the orthologs of the gene in question from different organisms. (i.e. they have different matches to database entries, both orthologs, but they are legitimately from the same gene). In this case, merging is the best avenue, but the software would actually split the transcript apart. Was this seen in some transcripts? Also, when referring to significance with similar proteins, values should be provided.

The authors mention that many genes unannotated by MAKER-P have repetitive stretches. What types of repetitive stretches? There is no methodology given for this analysis? This needs to be
described/explained. In regards to the sentence “Although most of these unannotated proteins have repetitive sequences (transposable elements)” - does that mean the unannotated proteins originate from within transposable elements, or transposable elements have inserted into the gene itself?

The authors do not address the overall differences in proteins detected by each annotation program - is a pattern that may explain these? Pseudogenes are mentioned twice, but this idea is not fully explained. While 20 tissues will capture many genes, it is probably not exhaustive – is there any additional evidence these “genes” are actually pseudogenes such as premature stop codons or frameshift mutations?

Minor:

This sentence needs improved clarity:
“The BLAST results of list.transcriptome.clean.ORFS: 320K on ‘plantpep.fasta’ was processed using a cutoff: bitscore=60, Evalue=1E-10”

Heading “Transcripts with multiple ORFs” - the section above also deals with transcripts with multiple ORFs. This heading could be clarified.

Some revision on the writing would improve readability. Abbreviations are recommended to be properly specified at first use in the manuscript and always in figures. Also, numbers and units should be spaced. In relation to the language, the authors are advised to review the use of scientific English, as well as verb tense consistency.

In Table 1, the main line indicates proteins not annotated by either program while the last line indicates listing of genes predicted by MAKER-P. These two sentences in the same caption lead to confusion. In addition, sizing and description of other figures might be improved.

References


Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.
Carmen Cuadrado  
Department of Food Technology, SGIT - INIA (Deputy Directorate General of Research and Technology-National Institute for Agricultural and Food Research and Technology), Madrid, Spain  
Africa Sanchiz  
Departamento de Tecnología de Alimentos, SGIT_INIA, Madrid, Spain

In this paper the authors investigate a new annotation method in the YeATS suite (YeATS Annotation Module - YeATSAM), which was used to annotate the newly-sequenced walnut genome using a simple workstation. In YeATSAM the assembled transcriptome is splitting into multiple ORFs. They show that YeATSAM failed to annotate ~700 genes identified by MAKER-P, while identifying ~4000 genes missed by MAKER-P. While most of these genes have repetitive stretches, both methods missed important genes identified by the other. Since many of the additional genes identified by MAKER-P have no known transcripts, the authors suggest that these were identified using ab initio methods. In the absence of such an ab initio module in YeATSAM, they propose a combined method using both MAKER-P and YeATSAM to annotate the WGS. This work is very interesting because the results probe the adequacy of this new annotation method. In general, the presentation is clear and the conclusions are adjusted to the results obtained. The figures and tables are also clear. Some comments are listed below:

- In the abstract, please change the order in “Results and Conclusions” part, from lines 17 to 21. Consider to mention first “YeATSAM used a […] chickpea transcriptome assembled using Newbler v2.3” and then that “1000 genes were identified, which were not previously annotated by Gnomon annotation tool”.

- Fourth and fifth paragraphs of Introduction could be changed to the discussion and in the introduction leave some short sentences about this.

- En fifth line of Methods section correct “seqeunces”.

- Please consider to explain further section “future work”.

**Competing Interests:** No competing interests were disclosed.

*We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.*
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com