University of Dundee

Improved annotation with de novo transcriptome assembly in four social amoeba species

Singh, Reema; Lawal, Hajara M.; Schilde, Christina; Glöckner, Gernot; Barton, Geoffrey J.; Schaap, Pauline; Cole, Christian

RESEARCH ARTICLE

CrossMark

Improved annotation with *de novo* transcriptome assembly in four social amoeba species

Reema Singh[1,2], Hajara M. Lawal[2], Christina Schilde[2], Gernot Glöckner[3,4], Geoffrey J. Barton[1], Pauline Schaap[2] and Christian Cole[1*] (iD)

Abstract

Background: Annotation of gene models and transcripts is a fundamental step in genome sequencing projects. Often this is performed with automated prediction pipelines, which can miss complex and atypical genes or transcripts. RNA sequencing (RNA-seq) data can aid the annotation with empirical data. Here we present *de novo* transcriptome assemblies generated from RNA-seq data in four Dictyostelid species: *D. discoideum*, *P. pallidum*, *D. fasciculatum* and *D. lacteum*. The assemblies were incorporated with existing gene models to determine corrections and improvement on a whole-genome scale. This is the first time this has been performed in these eukaryotic species.

Results: An initial *de novo* transcriptome assembly was generated by Trinity for each species and then refined with Program to Assemble Spliced Alignments (PASA). The completeness and quality were assessed with the Benchmarking Universal Single-Copy Orthologs (BUSCO) and Transrate tools at each stage of the assemblies. The final datasets of 11,315-12,849 transcripts contained 5,610-7,712 updates and corrections to >50% of existing gene models including changes to hundreds or thousands of protein products. Putative novel genes are also identified and alternative splice isoforms were observed for the first time in *P. pallidum*, *D. lacteum* and *D. fasciculatum*.

Conclusions: In taking a whole transcriptome approach to genome annotation with empirical data we have been able to enrich the annotations of four existing genome sequencing projects. In doing so we have identified updates to the majority of the gene annotations across all four species under study and found putative novel genes and transcripts which could be worthy for follow-up. The new transcriptome data we present here will be a valuable resource for genome curators in the Dictyostelia and we propose this effective methodology for use in other genome annotation projects.

Keywords: Dictyostelia, Social amoeba, De novo, Transcriptome assembly, RNA-seq

Background

Whole genome sequencing projects are within the scope of single laboratories. The Genomes OnLine Database [1] reports (as of 13th May 2016) there are 76,606 sequenced organisms, of which 12,582 are eukaryotes. However, only 8,047 are reported as being complete. Annotation of gene models is a requirement for a complete genome [2]. There are several complementary strategies for achieving gene annotation in novel genomes including gene prediction [3, 4], expressed sequence tag (EST) libraries [5] and RNA sequencing (RNA-seq) data [6]. Gene prediction methods are limited in the complexity of the gene models they are able to produce; alternative splice sites are unpredictable and untranslated regions (UTRs) have subtle signals [7]. EST libraries, if available, are usually fragmented and incomplete. RNA-seq data is dependent on good alignments to the reference. *De novo* transcriptome assembly is equally able to fulfil this function, although it can be computationally challenging [8–10]. Transcriptome assembly methods can be either reference-guided or reference-free [11, 12]. Reference-guided methods have the advantage of simplifying the search space, but are dependent on the

* Correspondence: c.cole@dundee.ac.uk
[1]Computational Biology, School of Life Sciences, University of Dundee, Dow Street, Dundee, UK
Full list of author information is available at the end of the article

Singh *et al. BMC Genomics* (2017) 18:120

Page 2 of 17

relevance, quality and completeness of the reference. Reference-free methods do not have any dependencies, but need to deal with sequencing errors sufficiently well to avoid poor assemblies [11–13]. We present the application of a *de novo* transcriptome assembly to four eukaryotic species: *Dictyostelium discoideum*, *Polysphondylium pallidum*, *Dictyostelium fasciculatum* and *Dictyostelium lacteum*. The genome of *D. discoideum* was published in 2005, it is 34 Mb in size and has been assembled into six chromosomes, a mitochondrial chromosome, an extra-chromosomal palindrome encoding ribosomal RNA (rRNA) and three 'floating' chromosomes [14]. The genome was generated via dideoxy sequencing and contigs were ordered into chromosomes by HAPPY mapping [14, 15] and still contains 226 assembly gaps.

In contrast, the similar sized genomes of *P. pallidum*, *D. lacteum* and *D. fasciculatum* were sequenced more recently using both dideoxy and Roche 454 sequencing. Their assembly was assisted by a detailed fosmid map and primer walking, leading to only 33 to 54 gaps per genome, but are more fragmented with 41, 54 and 25 supercontigs, respectively [15, 16]. The *D. discoideum* genome has been extensively annotated via the Dictybase project [17], whereas the gene models for *P. pallidum*, *D. fasciculatum* and *D. lacteum*, available in the Social Amoebas Comparative Genome Browser [18], are primarily based on computational predictions.

The social amoeba *D. discoideum* is a widely-used model organism for studying problems in cell-, developmental and evolutionary biology due to their genetic tractability allowing elucidation of the molecular mechanisms that underpin localized cell movement, vesicle trafficking and cytoskeletal remodeling as well as multicellular development and sociality. The social amoebas form a single clade within the Amoebozoa supergroup and are divided into four major taxon groups according to molecular phylogeny based on SSU rRNA and α-tubulin sequences [15]. The four species under study here represent each of the four groups: *D. discoideum* (group 4), *P. pallidum* (group 2), *D. fasciculatum* (group 1) and *D. lacteum* (group 3). Genome annotations are not static and benefit from the application of additional evidence and new methodologies [7, 19]. Therefore we present, for the first time, substantially updated annotations based on a *de novo* transcriptome assembly for the *D. discoideum*, *P. pallidum*, *D. fasciculatum* and *D. lacteum* genomes.

Methods
Sample preparation
Sequencing data were obtained from four RNA-seq experiments. The *D. discoideum* data were obtained from an experiment comparing gene expression changes between wild-type cells and a diguanylate cyclase (dgcA) null mutant at 22 h of development [20]. The *P. pallidum* data were obtained at 10 h of development in an experiment comparing wild-type and null mutants in the transcription factor cudA (Du, Q. and Schaap, P. unpublished results). In this experiment *P. pallidum* cells were grown in HL5 axenic medium (Formedium, UK), starved for 10 h on non-nutrient agar, and harvested for total RNA extraction using the Qiagen RNAeasy kit. The data for *D. lacteum* and *D. fasciculatum* were obtained from developmental time series [21]. For these series cells were grown in association with *Escherichia coli* 281, washed free from bacteria, and plated on non-nutrient agar with 0.5% charcoal to improve synchronous development. Total RNA was isolated using the Qiagen RNAeasy kit at the following stages: growth, mound, first fingers, early-mid culmination, fruiting bodies. *D. lacteum* RNAs were also sampled at three time points intermediate to these stages.
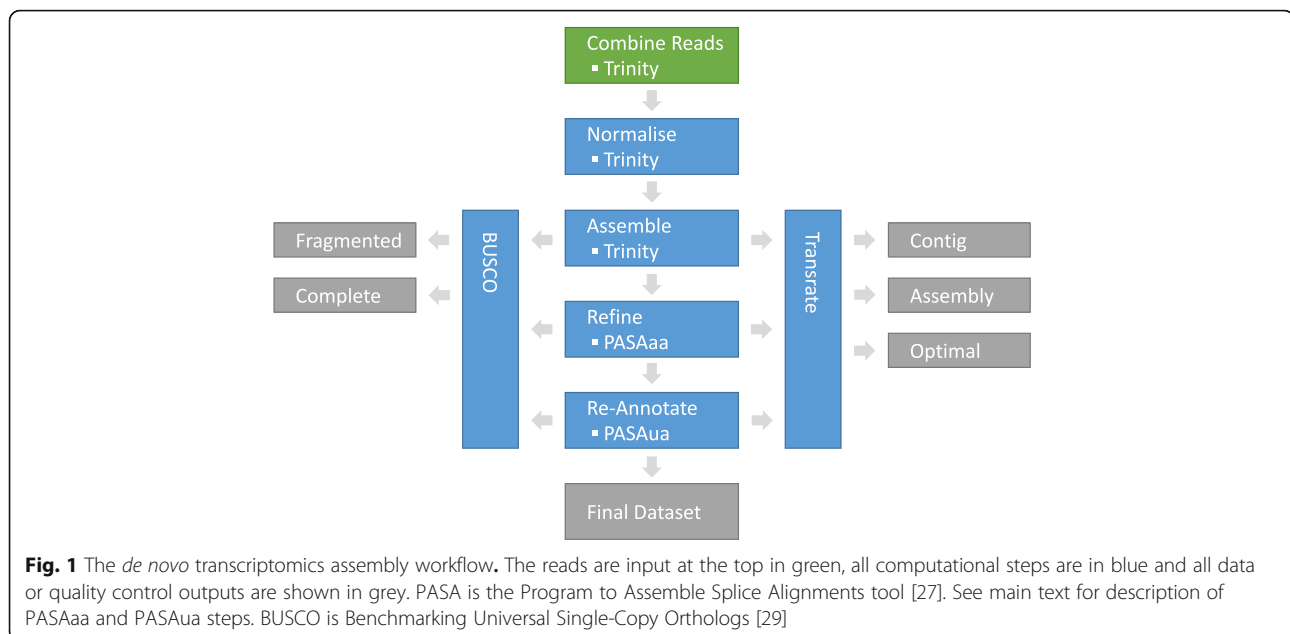
Illumina paired end sequencing
Total RNA was enriched for messenger RNA (mRNA) using poly-T oligos attached to magnetic beads and converted to a sequencing ready library with the TruSeq mRNA kit (Illumina), according to manufacturer's instructions and 100 basepairs (bp) paired-end sequenced using an Illumina HiSeq instrument. For the *D. discoideum* and *P. pallidum* samples, 1 μg of total RNA was used as starting material, with 4 ul of 1:100 dilution External RNA Controls Consortium (ERCC) ExFold RNA Spike-In Mixes (Life Technologies) added as internal controls for quantitation for the RNA-Seq experiment and sequenced at the Genomic Sequencing Unit, Dundee. In total there were 433 M, 413 M, 171 M and 319 M reads respectively for *D. discoideum*, *P. pallidum*, *D. fasciculatum* and *D. lacteum*.

Data processing and *de novo* transcriptomics assembly
The quality of the raw reads was checked with FastQC [22] and the reads were found to have high quality scores across their full length. No trimming of the data was performed, as aggressive trimming can negatively impact on the quality of assemblies [23]. All reads for each species were separately combined prior to *de novo* assembly. Being a more mature genome the *D. discoideum* data was used to verify the methodology, thereby giving a reference point for the other, less well characterised, species. Figure 1 shows a schematic of the overall workflow.

Trinity version 2013.11.10 [8] was used for *de novo* assembly, and normalisation of the read data was achieved with a kmer of 25 and aiming for 50x coverage of the genome. Following normalisation there remained 5.3 M, 8.3 M and 16.0 M read pairs in *D. discoideum*,

Singh *et al. BMC Genomics* (2017) 18:120

Page 3 of 17

Fig. 1 The *de novo* transcriptomics assembly workflow. The reads are input at the top in green, all computational steps are in blue and all data or quality control outputs are shown in grey. PASA is the Program to Assemble Splice Alignments tool [27]. See main text for description of PASAaa and PASAua steps. BUSCO is Benchmarking Universal Single-Copy Orthologs [29]

P. pallidum and *D. lacteum*, respectively. *D. fasciculatum* reads were not normalised as there were fewer than the recommended 300 million reads as per the Trinity manual. Trinity was run on the normalised reads using the –jacard-clip parameter and setting –k-min-cov to 4 in an attempt to reduce the number of fused transcripts in *P. pallidum* only. In the other species the parameter made little difference. For the initial transcript set of *D. discoideum* and *P. pallidum* assemblies, any transcripts with BLAT (BLAST-like alignment tool) v35x1 [24] hits to the ERCC spike-in sequences were removed from the *D. discoideum* and *P. pallidum* assemblies. *D. fasciculatum* and *D. lacteum* were not cultured axenically and thus the samples were contaminated by their bacterial food source. In order to remove the bacterial contamination *D. fasciculatum* and *D. lacteum*, transcripts were filtered with the TAGC (taxon-annotated GC-coverage) plot pipeline [25]. TAGC determines for each contiguous sequence (contig) the proportion of GC bases, their read coverage and best phylogenetic match. With this information it is possible to identify which transcripts are mostly likely to be contaminants and removed. In order to remove the contamination, first all the transcripts were aligned to the BLAST 'nt' database using BLAST megablast. Using the trinity assembled transcripts, the BAM file of the reads mapped back to the transcripts and the transcripts to species mapping, non-target related transcripts were removed. The contaminant transcripts were differentiated on the coverage vs GC plots (see Additional file 1: Figure S1).

The normalised set of reads were aligned with bowtie (0.11.3, with parameters applied as per Trinity script alignReads.pl) [26] to the whole transcript set and the total number of reads matching to each transcript were stored (see Additional file 1: Figure S2 for the read distributions for each dataset).

Transcript refinement

Program to Assemble Spliced Alignments (PASA) v2.0.0 [27] was used to refine the Trinity transcripts into more complete gene models including alternatively spliced isoforms. Initially developed for EST data, PASA has been updated to also work with *de novo* transcriptome data. Using the seqclean tool available with PASA, all the transcripts were screened and trimmed for low complexity regions, poly (A) tails and vector sequences. GMAP (Genome Mapping and Alignment Program) [28] and BLAT [24] were used to align the transcripts to their respective genomes. Trinity transcripts that failed to align to the already existing genome in both GMAP and BLAT were removed as 'failed'. Remaining 'good' transcripts at this stage are termed the PASAaa dataset. Next, PASA takes existing annotations and compares them to the PASAaa dataset. PASA uses a rule-based approach for determining which transcripts are consistent or not with the existing annotation and updates the annotation as appropriate: new genes, new transcript isoforms or modified transcripts. PASAua is the term used for the PASA assembled transcripts after updating with existing annotation.

Assembly quality check

At each stage, the transcript datasets were assessed with Benchmarking Universal Single-Copy Orthologs, BUSCO, [29] and Transrate v1.0.0 [13] These methods take complementary approaches in assessing completeness and/or accuracy. Transrate v1.0.0 uses the read data and

Singh *et al. BMC Genomics* (2017) 18:120

Page 4 of 17

optionally the reference sequence as input. BUSCO defines a set of 429 core eukaryotic genes. These genes are used as a proxy for minimum completeness based on the assumption that a eukaryotic genome or transcriptome assembly should encode a large proportion of the core set of genes. The BUSCO (v1.1b1) tool uses hidden Markov models (HMMs), defined for each of the core genes in the set, returning whether there are complete or partial matches within the *de novo* transcripts. When run in genome mode, BUSCO additionally uses Augustus [3] to generate a predicted gene set against which the HMMs are tested. Transrate calculates the completeness and accuracy by reporting contig score and assembly score. Contig score measures the quality of the individual contig, whereas assembly score measures the quality of whole assembly.

Orphan RNAs

The full set of Trinity transcripts constitutes the best approximation of the assembly of transcripts expressed in the RNA-seq sequencing data. The transcripts were aligned against the existing genome and coding DNA (cDNA) references (from Dictybase (*D. discoideum)* and SACGB [18], (*D. fasciculatum, D. lacteum* and *P. pallidum*)) using BLAT. Any transcripts not matching the existing references were searched against the NCBI 'nt' database with BLAST [30] and with PSI-BLAST against the NCBI 'nr' database for the longest predicted ORF in any remaining transcripts without a match to 'nr'. This exhaustive search allowed the categorisation of 'annotated' (transcript with match to known genome and/or cDNA), 'known' (match to related species), 'artefact' (match to non-related species (non-Dictyostelid)) and 'putative novel' (remainder) datasets.

PCR and subcloning

D. discoideum genomic DNA (gDNA) was extracted using the GenElute mammalian genetic DNA extraction kit (Sigma). Polymerase chain reaction (PCR) reactions were run for 30 cycles with 50 ng of gDNA and 1 μM of primers with 45 s annealing at 55 °C, 2 min extension at 70 °C and 30 s denaturation at 94 °C. The reaction mixtures were size-fractionated by electrophoresis, and prominent bands around the expected size were excised, purified using a DNA gel extraction kit (Qiagen) and subcloned into the PCR4-TOPO vector (Invitrogen). After transformation, DNA minipreps of clones with the expected insert size were sequenced from both ends.
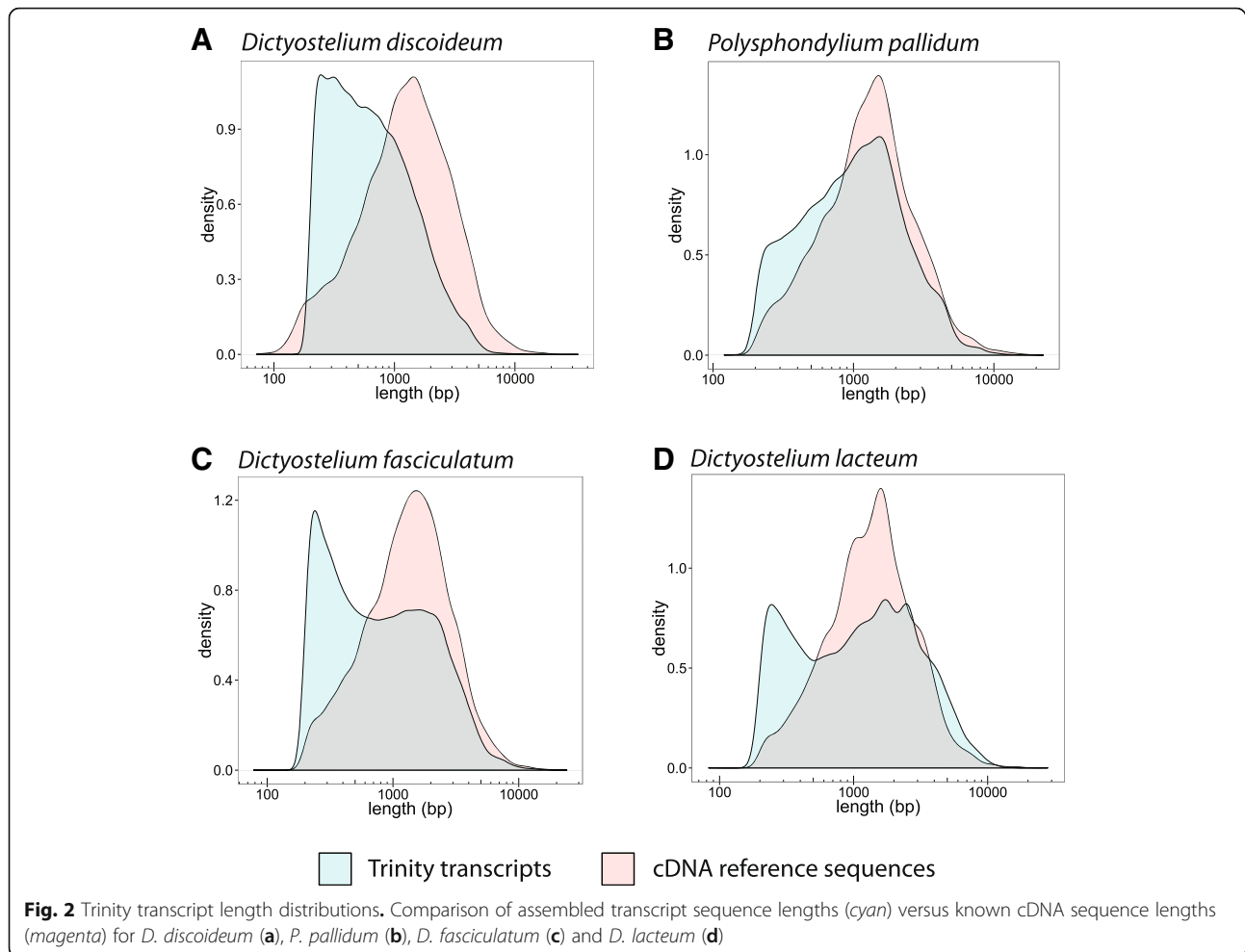
Results and discussion

De novo transcript assembly

Table 1 shows a summary of the Trinity output for the *D. discoideum, P. pallidum, D. lacteum* and *D. fasciculatum de novo* transcriptome assemblies. Overall, the raw assemblies are similar in terms of total transcripts, GC content, and contig N50 or E90N50 (N50 for the top 90% expressed transcripts). *D. discoideum* is slightly anomalous in N50, E90N50, mean length and transcripts ≥ 1,000 bp with all features being smaller than the other three assemblies. The mean length over all the annotated Dictybase coding sequences is 1,685 bp which is substantially larger than in the assembled transcripts (867 bp) suggesting that the *D. discoideum* transcripts are fragmented.

Figure 2 shows the distribution of transcript lengths for *D. discoideum, P. pallidum, D. fasciculatum, D. lacteum* (cyan) when compared to the available cDNA datasets (magenta). The *D. discoideum* cDNAs are manually curated, whereas the others are predicted. The transcript sets are enriched in short transcripts (<1000 bp) as

Table 1 Trinity assembly summary statistics

	D. discoideum	*P. pallidum*	*D. fasciculatum*[a]	*D. lacteum*[a]
Total raw read pairs	216,284,941	206,385,299	85,379,416	159,455,838
Normalised read pairs	5,278,467	8,273,023	NA	15,994,900
Assembled Statistics				
Total Trinity transcripts	31,259	35,631	46,779	38,508
Transcripts > = 1000 bp	8,818	17,988	18,977	20,637
GC content (%)	28.1	35.7	35.6	31.2
Maximum Transcript Length (bp)	21,679	17,271	19,435	15,026
Stats based on all transcript contigs				
Contig N50	1,268	1,871	2,107	2,759
Contig E90N50	1,423	2,069	2,424	3,341
Mean length (bp)	871	1,338	1,212	1,671
Total assembled bases	27,239,648	47,658,623	56,707,337	64,358,120

[a]Assembly statistics calculated after removal of bacterial transcripts (see Additional file 1)

Singh *et al. BMC Genomics* (2017) 18:120

Page 5 of 17

Fig. 2 Trinity transcript length distributions. Comparison of assembled transcript sequence lengths (*cyan*) versus known cDNA sequence lengths (*magenta*) for *D. discoideum* (**a**), *P. pallidum* (**b**), *D. fasciculatum* (**c**) and *D. lacteum* (**d**)

compared to their cDNAs with the effect being most marked in *D. discoideum*, *D. fasciculatum* and *D. lacteum* (Fig. 2a,c and d). The *P. pallidum* assembly is more similar to its cDNA reference dataset (Fig. 2b). Interestingly, the longest assembled transcript in *D. discoideum* (21,679 bp) was found to be approximately half of the mitochondrial chromosome. We speculate that as the mitochondrion is gene rich and highly expressed, Trinity was unable to resolve overlapping reads from adjacent genes thereby joining them all into one 'supercontig'.

The subsequent steps in the assembly were performed with PASA [27] which uses reference genome and transcript datasets to generate a refined and updated transcriptome assembly. The first stage takes the transcriptome assemblies, aligns them against the genome and clusters them into gene structures according to their genome alignments. Any transcripts, which do not align adequately to the genome are filtered out by PASA, under the assumption that they are misassemblies. This dataset will be referred to as 'PASA annotated assemblies', PASAaa. In unfinished and complex genomes, it is possible that there are missing gene loci in the genome reference. The missing loci may appear in a *de novo* transcriptome assembly and would be filtered out by PASA. The second stage uses the aggregated and filtered set of transcripts to refine the existing annotations for each of the species. At this stage, the gene models are updated with new or extended UTRs, new alternatively spliced isoforms are added, and introns are added or removed. New genes are identified, and existing genes are split or merged as required by the *de novo* assembly data. This dataset is referred to as 'PASA updated annotations', PASAua.

Table 2 shows the results of each stage of the assembly workflow from Trinity to each of the PASA steps and compared to the existing set of gene models from DictyBase (*D. discoideum*) or Augustus predictions (*P. pallidum*, *D. lacteum* and *D. fasciculatum*). It is clear that at each stage the assemblies become more similar to the existing gene models (Table 2). For example, in all the species the total number of transcripts was 3-4-fold larger in the Trinity data than in the existing annotations. Although *de novo* assembly

Singh *et al. BMC Genomics* (2017) 18:120

Page 6 of 17

Table 2 Comparison of transcript statistics at each stage of assembly

	Trinity	PASAaa	PASAua	Existing Models[a]
D. discoideum				
No. transcripts	31,259	19,920	11,523	12,283
Mean length	871	782	1,787	1,685
No. exons	392,850	32,389	26,905	27,293
Mean exons/transcript	12.78	1.65	2.35	2.41
P. pallidum				
No. transcripts	35,631	21,230	12,849	11,440
Mean length	1,338	1,231	1,822	1,635
No. exons	94,636	45,582	37,906	33,179
Mean exons/transcript	2.67	2.44	3.12	2.90
D. fasciculatum				
No. transcripts	46,779	21,574	12,714	11,879
Mean length	1,212	1,357	1,951	1,696
No. exons	88,425	43,406	39,021	36,789
Mean exons/transcript	2.25	2.35	3.21	3.10
D. lacteum				
No. transcripts	38,508	18,045	11,315	10,232
Mean length	1,671	1,871	1,964	1,712
No. exons	75,249	27,430	22,396	20,682
Mean exons/transcript	2.46	2.26	2.15	2.02

[a]Existing models retrieved from DictyBase for *D. discoideum* and derived from Augustus predictions for *P. pallidum*, *D. lacteum* and *D. fasciculatum*

has the potential to identify novel genes and transcripts, a 3-fold increase is unlikely. By the end of PASAua, the transcript counts were within 1,500 of the existing models, with *D. fasciculatum*, *D. lacteum* and *P. pallidum* having more genes than in their Augustus-predicted models, and *D. discoideum* having 760 fewer genes than in the DictyBase-curated models. This is to be expected as the gene prediction algorithms are unlikely to have found all transcripts, whereas the *D. discoideum* curated set will include genes expressed under certain conditions only (e.g. developmental time points) that were not part of the experiment included here. Mean transcript lengths increased through the workflow. In particular, for *D. discoideum*, the mean Trinity transcript length was 871 bp and the final PASAua length was 1,787 bp indicating that the high fragmentation observable by an excess of short transcripts (Fig. 2) has been reduced. Similarly, the total number of identified exons was reduced from the initial Trinity dataset.

Overall, the initial Trinity assemblies have been refined from a fragmentary and redundant dataset to a more full-length and less redundant set of transcripts, which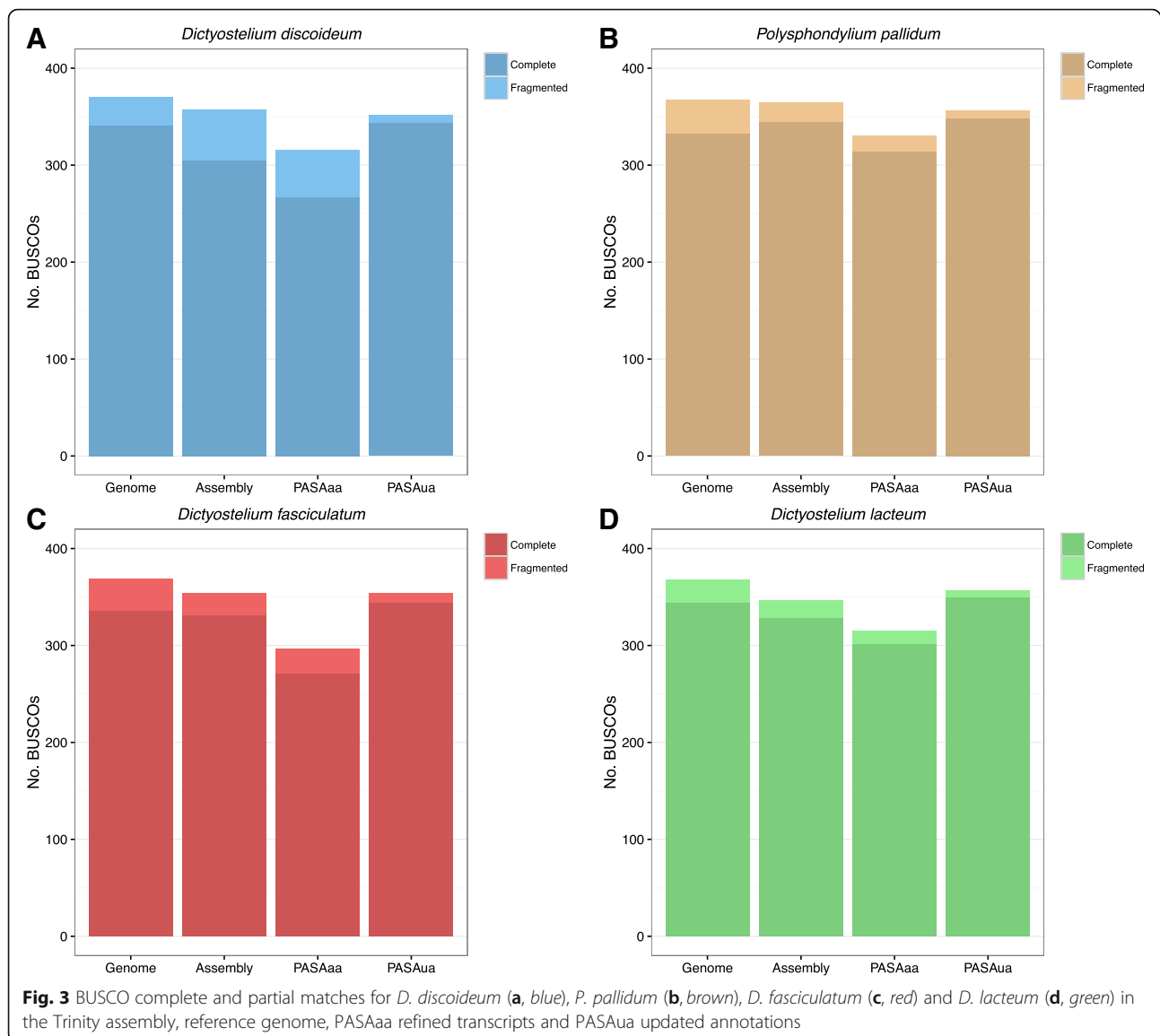 are more similar to the existing reference datasets in terms of total transcript counts, mean length, number of exons and exons per transcript (Table 2).

Quality assessment

Benchmarking Universal Single-Copy Orthologs (BUSCO) and Transrate are tools which allow the assessment of completeness and accuracy of transcriptome assemblies. A set of 429 core eukaryotic genes (CEGs) was defined by BUSCO, for the purpose of assessing completeness in eukaryotic genomes [29]. CEGs are conserved across taxa and the majority should be present in the majority of eukaryotic species. A large fraction of missing BUSCO genes could be indicative of an incomplete assembly. Figure 3 shows the comparison of complete and partial BUSCO matches in all four species for the genome reference, Trinity assembly, PASAaa refined transcripts and PASAua updated annotations. In the ideal situation all the BUSCOs would be detected in an assembly, however high sequence divergence or absence in the species will give a lower maximum detection level. The whole genome BUSCO score represents the upper limit for any of the assemblies. All except the PASAaa datasets have >80% complete or fragmented BUSCOs and are close to the whole genome count, suggesting the transcriptome assemblies are nearly complete. It is noticeable, that the number of identified BUSCOs is consistently lower in the PASAaa data for all four species (Fig. 3). This drop is due to the strict PASA filtering during transcript assembly. PASAaa only retains transcripts, which align to the reference with 95% identity and 90% length coverage. Manual checking of the BUSCOs that are identified in the Trinity data, but not in PASAaa reveals that they all are labelled as failed alignments. This suggests that either BUSCO is overly permissive in defining the orthologues or that PASAaa is overly aggressive in filtering transcripts. PASAua appears to 'rescue' this behaviour, presumably by including good annotations for genes that are poorly assembled in the Trinity data.
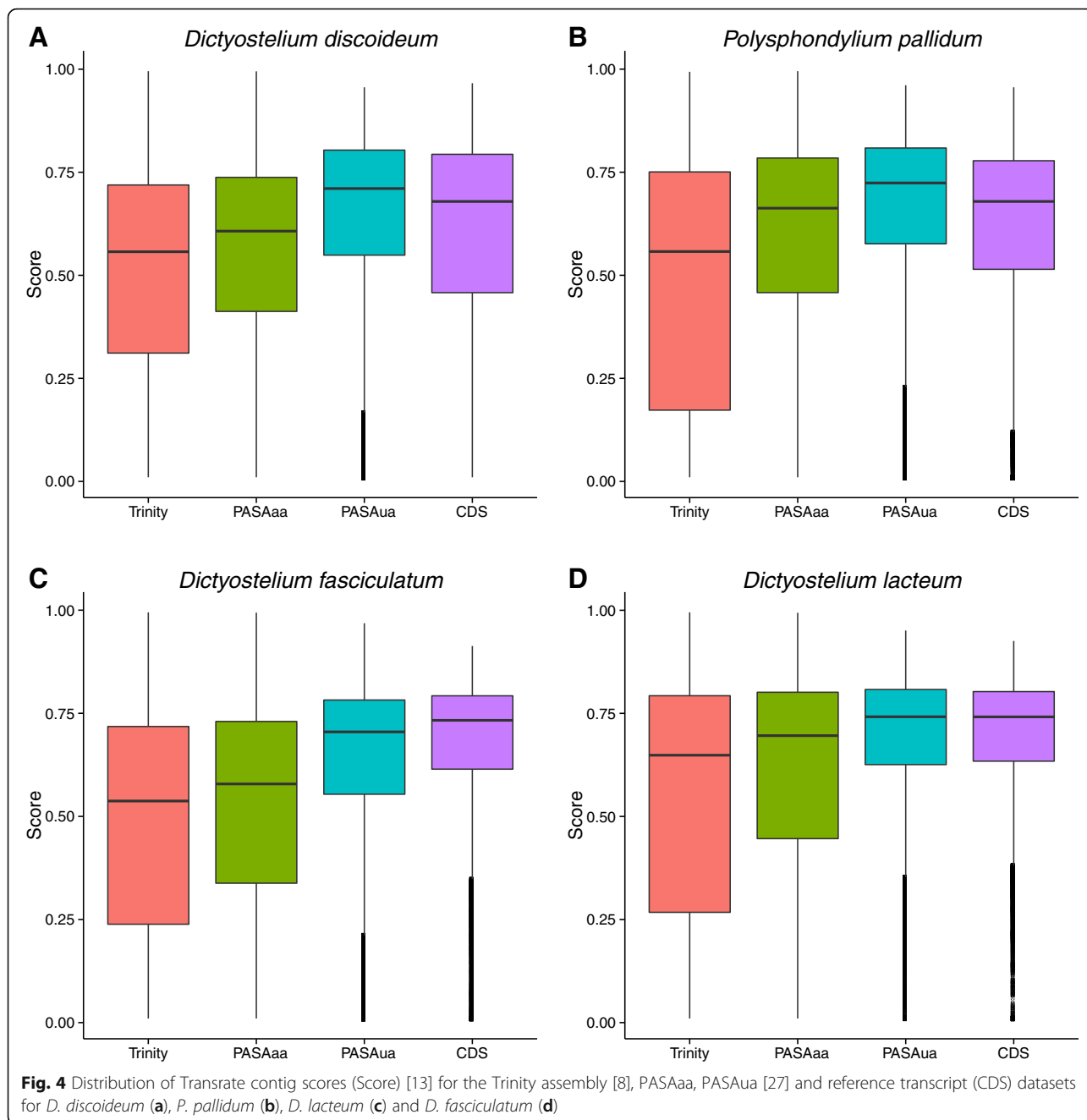
Transrate assesses transcript quality by calculating several contig-level metrics based on the input RNA-seq data, and measures how well the read data support the contigs. Contigs are scored individually and then combined into an overall assembly score which ranges from 0 to 1. An optimal score is also reported, which predicts the best potential assembly score achievable by removing the worst scoring contigs in the dataset. An assembly score of 0.22 and optimised score of 0.35 were found to be better than 50% of 155 published *de novo* transcriptome assemblies [13]. A high Transrate score with a small improvement in the optimal score indicates a good *de novo* assembly, which is unlikely to be improved without further data or information.

Figure 4 compares the distribution of Transrate contig scores from the Trinity assembly, PASAaa refinement,

Singh *et al. BMC Genomics* (2017) 18:120

Page 7 of 17

Fig. 3 BUSCO complete and partial matches for *D. discoideum* (**a**, *blue*), *P. pallidum* (**b**, *brown*), *D. fasciculatum* (**c**, *red*) and *D. lacteum* (**d**, *green*) in the Trinity assembly, reference genome, PASAaa refined transcripts and PASAua updated annotations

PASAua update and reference transcript/coding sequence (CDS) datasets for each of the four species. In contrast to the BUSCO data, the PASAaa data shows an improvement in Transrate contig scores when compared to the raw Trinity output meaning that the PASAaa transcripts are more consistent with the data, confirming that perhaps BUSCO is too permissive when assigning orthologues rather than PASAaa being too aggressive with its filtering. Notably the reference sequence datasets ('CDS' Fig. 4) for *D. discoideum* and *P. pallidum*, show a lower median score than the PASAua data, indicating that PASAua is working well in combining the data with the existing annotations. There is little difference in *D. lacteum*. In *D. fasciculatum* the CDS data shows the best Transrate score of any of the assemblies.
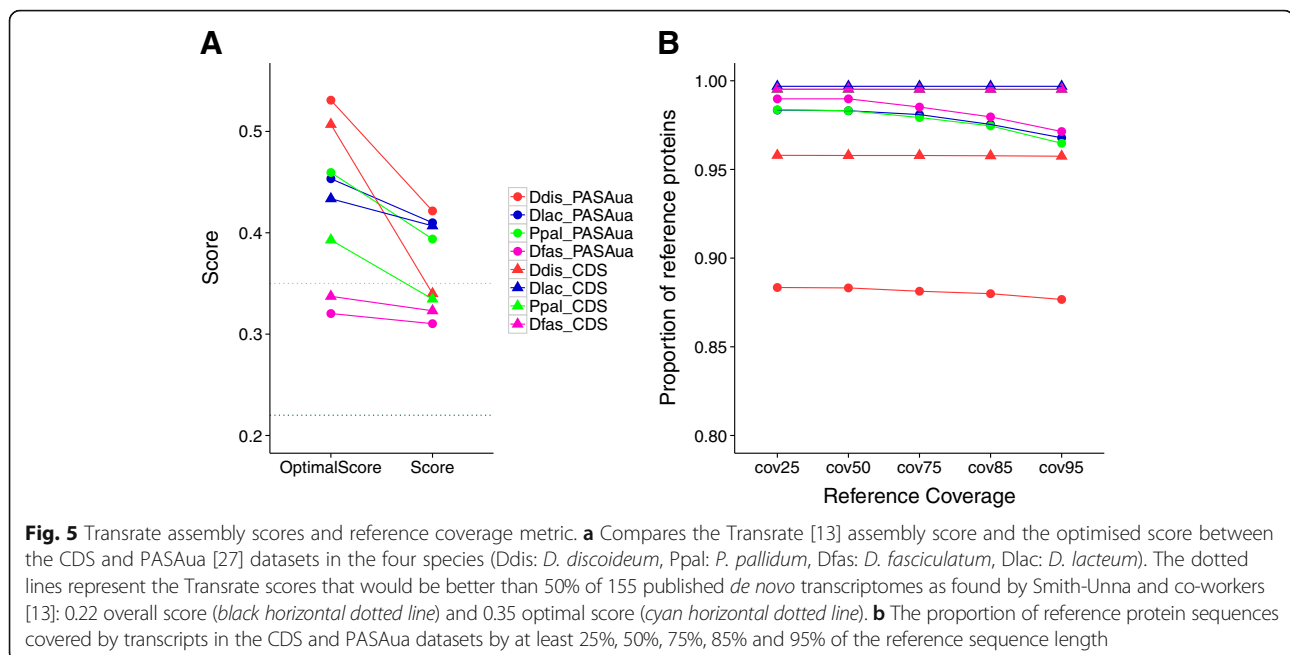
Figure 5 compares the Transrate assembly scores and optimal scores between PASAua and the annotated CDS over the four species. The assembly scores range from 0.31 (*D. fasciculatum*) to 0.42 (*D. discoideum*) and the optimal scores range from 0.32 (*D. fasciculatum*) to 0.53 (*D. discoideum*). It is clear that PASAua has better Transrate scores (Fig. 5a filled circles) than the annotated CDS (Fig. 5a filled triangles), except for *D. fasciculatum*, with all the PASAua assemblies scoring better than 50% of published transcriptome assemblies (Fig. 5a dotted black line). The optimal scores for PASAua are also all better than 50% of published transcriptome assembly data (Fig. 5a dotted cyan line), with the exception of *D. fasciculatum*. In *D. fasciculatum* the difference between the assembly (0.31) and optimal PASAua scores

Singh *et al. BMC Genomics* (2017) 18:120

Page 8 of 17

Fig. 4 Distribution of Transrate contig scores (Score) [13] for the Trinity assembly [8], PASAaa, PASAua [27] and reference transcript (CDS) datasets for *D. discoideum* (**a**), *P. pallidum* (**b**), *D. lacteum* (**c**) and *D. fasciculatum* (**d**)

(0.32) is small (Fig. 5a green filled circles), suggesting that there is little improvement to the assembly possible given the read data for this species. Using the optimal score, Transrate defines a set of 'good' contigs which best fit the data. The proportion of PASAua good contigs ranges from 79.9% (*D. discoideum*) to 97.2% (*D. fasciculatum*) which, for all species, is a higher proportion than the annotated CDS (Additional file 1: Table S3).

Transrate additionally has a reference-based measure, which aligns the transcripts to the reference protein sequences and the results are shown in Fig. 5b. The y-axis

in Fig. 5b shows the proportion of reference protein sequences covered with transcript sequences at several thresholds (25, 50, 75, 85 and 95%) of the reference. Clearly the majority of the datasets recapitulate the reference annotations completely - only when looking at the raw numbers can a tiny (<0.01) difference between the CDS and PASAua assemblies be observed. The exceptions are the *D. discoideum* datasets where the CDS shows a 0.96 proportional coverage of the reference and the PASAua dataset is worse at 0.88 proportional coverage. It is unclear why the *D. discoideum* set would be

Singh *et al. BMC Genomics* (2017) 18:120

Page 9 of 17

Fig. 5 Transrate assembly scores and reference coverage metric. **a** Compares the Transrate [13] assembly score and the optimised score between the CDS and PASAua [27] datasets in the four species (Ddis: *D. discoideum*, Ppal: *P. pallidum*, Dfas: *D. fasciculatum*, Dlac: *D. lacteum*). The dotted lines represent the Transrate scores that would be better than 50% of 155 published *de novo* transcriptomes as found by Smith-Unna and co-workers [13]: 0.22 overall score (*black horizontal dotted line*) and 0.35 optimal score (*cyan horizontal dotted line*). **b** The proportion of reference protein sequences covered by transcripts in the CDS and PASAua datasets by at least 25%, 50%, 75%, 85% and 95% of the reference sequence length

different to the other datasets with this metric given its comparable behaviour using other metrics. It is possible that by focussing on the known annotations only, there are features in the predominantly manually curated reference in *D. discoideum*, which are not captured in the assembly; for example developmental specific genes not observed under the conditions here presented.

Interpretation

What does an RNA-seq-based *de novo* assembly achieve when there is an already existing annotation either manually curated or generated via prediction? Is it worth it?

Table 3 details the results following PASA refinement of the existing gene models. Despite being a manually curated genome, the *D. discoideum* gene models where extensively modified by PASA with 7,182 being updated. Most of the updates in *D. discoideum* (6,750, 94%) are the result of UTR additions at 5′ and 3′ ends of genes, which were mostly missing in the existing models. The assemblies in the other species have a similar number of updates, but UTR-only updates to transcripts are a smaller fraction of the total. 187 new alternatively spliced transcripts, in 170 genes, were identified in *D. discoideum* (Table 3). There are currently 70 alternatively spliced

transcripts, in 34 genes, annotated in Dictybase so this new data represents a 2.7-fold increase in the number alternatively splice transcripts and a 5-fold increase in genes. This number in *D. discoideum* could be an underestimate as the *D. fasciculatum*, *D. lacteum* and *P. pallidum* assemblies all have ~1000 alternate splice isoforms.

Figure 6 provides examples, in each of the four species, of changes to the transcript models determined by PASA that are well supported by all the data. Each panel highlights a different type of change to the reference model. Gene DDB_G0295823 has a single transcript (DDB0266642 Fig. 6a) with two exons and a single intron. The RNA-seq data (brown), Trinity assembly (purple) and PASAaa refinement (red) identifies extensions to the model, adding 5′ and 3′ UTRs to the annotation (green, narrow bars). The Trinity transcript (purple) is on the opposite strand to the reference transcript (black) and is corrected by PASA (red & green). The example in *P. pallidum* (Fig. 6b) shows three new alternatively spliced products of the gene (Fig. 6b green bars 1, 2, 3 labels). The three new models have the same coding region, but differ in their 5′-UTRs: two with differently sized introns and one without an intron. The new models also include a longer second coding exon (Fig. 6b arrow),

Table 3 Summary data following PASA transcript refinement and re-annotation

	D. discoideum	*P. pallidum*	*D. fasciculatum*	*D. lacteum*
Gene model updated	7,182	7,712	6,664	5,610
New alternate splice isoforms	187	1,321	842	1,088
Novel genes	44	175	19	21
Update results in modified protein	554	2,252	5,393	4,741

Singh *et al. BMC Genomics* (2017) 18:120

Page 10 of 17
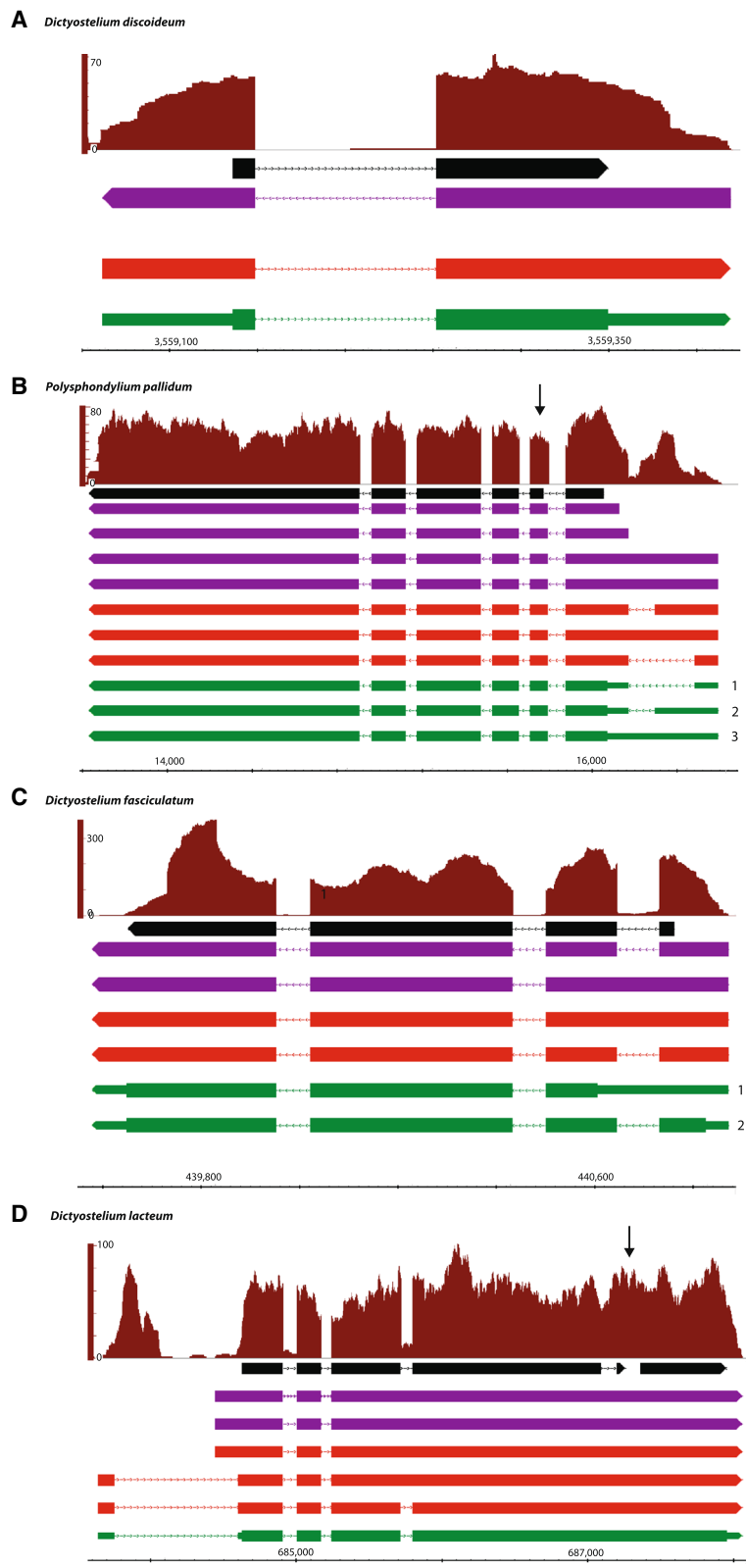
Fig. 6 (See legend on next page.)

Singh et al. BMC Genomics (2017) 18:120
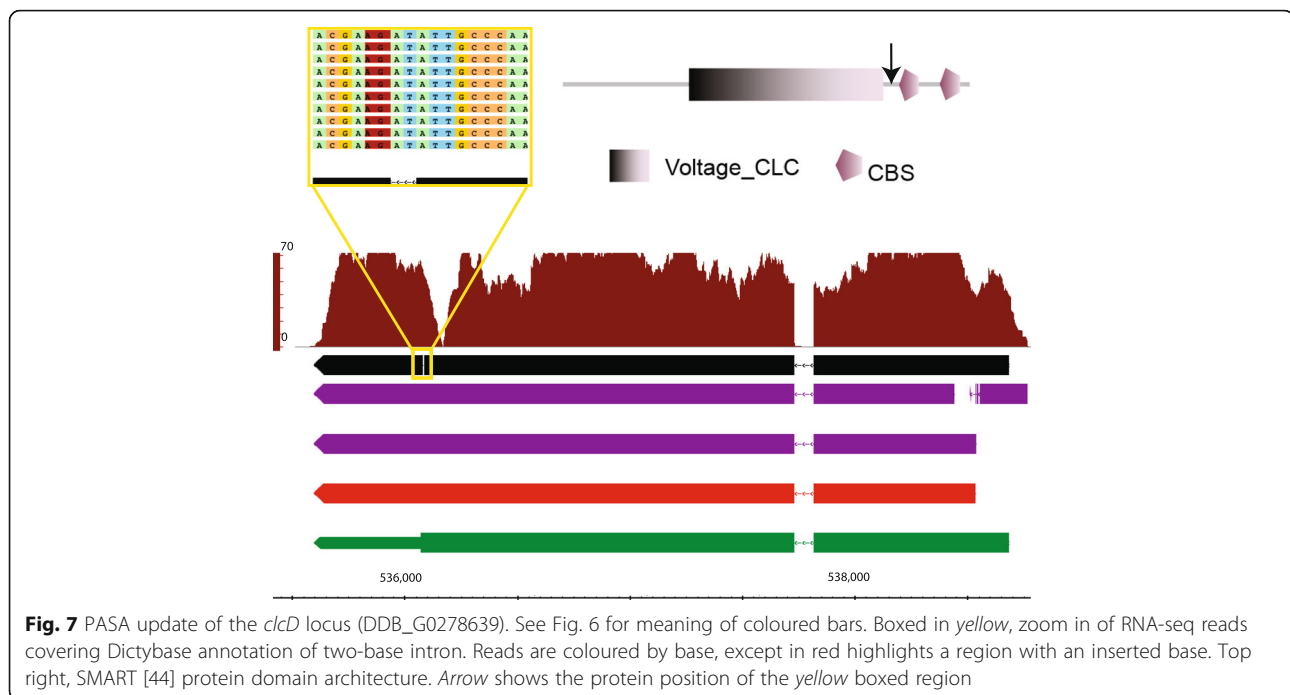
Page 11 of 17

which increased the sequence of the protein product by 9 amino acids. Figure 6c shows an example, in *D. fasciculatum*, where an alternatively spliced transcript alters the protein product. The alternatively spliced isoform (Fig. 6c, labelled 1) removes the first intron and extends the 5′-UTR when compared to the updated gene model (labelled 2). The CDS is shortened by 45 amino acids with the use of alternate start site, but the rest of the protein is identical. In the RNA-seq data it appears that this new alternative transcript is not the dominantly expressed isoform in the context of the whole organism. The final example is the merging of two *D. lacteum* genes into one (Fig. 6d). The black bars show two distinct genes (DLA_11596 and DLA_04629), but the RNA-seq data (brown) and the Trinity assembly (purple bars) show uninterrupted expression across the intergenic region between the two genes (arrow). The PASA refinement and re-annotation (red and green bars) encapsulate the expression as a contiguous region with the coding region being in-frame over the two existing gene models. The annotation for the upstream DLA_11596 gene in SACGB [17] gives its best bi-directional hit in Uniprot/TrEMBL as *gxcN* in *D. discoideum* (DDB0232429, Q55 0V3_DICDI). *gxcN* codes for a 1,094 amino acid protein where DLA_11596 codes for a 762 amino acid protein and the pairwise alignment of DLA_11596 with DDB0232429 shows no overlap over the C-terminal 300 residues. The PASAua gene fusion of DLA_11596/DLA_04629 (Fig. 6d) codes for a longer, 1,029 protein which aligns across the full length of DDB0232429 in a pairwise alignment. We suggest that the existing gene model, DLA_11596, is a truncated form of a *D. discoideum gxcN* orthologue and that the fusion with the downstream DLA_04629 gene represents the more accurate gene model.

Given that *D. discoideum* has been extensively studied and the annotation curated by Dictybase, it is of note that our pipeline identified putative changes which altered the protein sequence of 554 genes (4.5% of total reference models) (Table 3). *D. discoideum* has been the focus of many functional studies including about 400 deletions in genes that are required for normal multicellular development [16]. Comparing the 554 *D. discoideum* genes with modified proteins to the developmentally essential genes, we found 16 genes (2.9%) that overlapped (see Additional file 2: Figure S3 for domain diagrams). Out of the 16, nine are either truncated or extended at the N- or C-terminal. In the remaining seven proteins, there is loss or gain of exons. Five proteins were updated

with additional exons: DDB_G0268920, DDB_G0269160, DDB_G0274577, DDB_G0275445 and DDB_G0277719, and two proteins have an exon deletion: DDB_G0271502 and DDB_G0278639.

Investigating these protein changes in more detail revealed some errors in the underlying genome sequence, which resulted in some unusual gene models. Figure 7 shows clcD (chloride channel protein, DDB_G0278639) as an example. In the domain architecture of clcD, there are two CBS (cystathionine beta-synthase) domains present at positions 827–876 and 929–977 in the transcript sequence. In the updated sequence the protein is truncated and these two domains have been removed. This is likely to be incorrect since all eukaryotic CLC proteins require the two C-terminal CBS domains to be functional [31]. How did this change occur in the *de novo* transcript assembly? In the existing annotation, there is an impossibly short two-base intron between the CLC domain and first CBS domain. Splicing requires a two-base donor and a two-base acceptor at either end of the splice site meaning at least four bases are required, not including any insert sequence. Careful investigation of the RNA-seq genome aligned reads reveals a single-base insertion immediately after the intron in 22 out of 23 reads overlapping the region (yellow inset, Fig. 7). The RNA-seq data turns the two-base intron into a three-base, in-frame codon inserting an isoleucine into the protein sequence and retaining the CBS domains. By implication there is a missing base in the genome reference, which interrupts the open reading frame with a premature stop upstream of the CBS domains (arrow, Fig. 7). PASA cannot deal with missing bases in the reference and erroneously truncates the, now out-of-frame, coding region four codons downstream of the missing base at a TGA stop codon. It also cannot create an impossible intron, which a human annotator presumably added in order to keep the transcript in-frame and retain the conserved CBS domains. PASA did make an error updating this gene, but it does not seem possible for it to have dealt with the missing base any other way.

Inspection of all the *D. discoideum* gene models identified 119 sites in 102 genes with introns shorter than 5 bp (see Additional file 3: Table S1). Of these genes, five have three tiny introns each. Four of them are either in poorly expressed genes or in poorly expressed regions within genes. One gene (DDB_G0279477), however, is well expressed across the full length. The gene contains two 3 bp introns and one 1 bp intron. The two 3 bp

Singh *et al. BMC Genomics* (2017) 18:120

Page 12 of 17

Fig. 7 PASA update of the *clcD* locus (DDB_G0278639). See Fig. 6 for meaning of coloured bars. Boxed in *yellow*, zoom in of RNA-seq reads covering Dictybase annotation of two-base intron. Reads are coloured by base, except in red highlights a region with an inserted base. Top right, SMART [44] protein domain architecture. *Arrow* shows the protein position of the *yellow* boxed region

introns contain a TAA sequence encoding a stop codon, but according to the RNA-seq data the codons should be TTA (Leu) with evidence from 56 and 33 reads in the two sites, respectively, 100% of which contain the TTA codon. The 1 bp intron region is covered by 38 reads and one would not expect to see introns in RNA-seq data, by definition, it does seem highly unlikely for a 1 bp intron to exist given our current knowledge of mRNA splicing: canonical GU-AG dinucleotides and a branch point >18 bp upstream from the 3′ splice site. For this gene, there are clear errors in the genome sequence, which have lead to the creation of an erroneous gene model to compensate for them. It is arguable that none of the 119 < 5 bp introns are genuine but are artificial constructs to fix problems with the gene models. We recommend that gene annotators revisit these genes and consider updating the models [7, 32] and the underlying genome using RNA-seq data as evidence [33, 34].

In addition to what we have shown here, it would be possible to use the RNA-seq data to directly improve the genome assembly of the four dictyostelid species mentioned herein. Xue et al. [35] have shown with their 'L_RNA_Scaffolder' tool that improved scaffolding of complex genomes such as human and zebrafish is possible with RNA-seq indicating the feasibility in more gene dense species.

The protein changes in *D. fasciculatum*, *D. lacteum* and *P. pallidum* number in the thousands (Table 3) highlighting that computational gene prediction is only a first step in annotating a genome. A reliable genome annotation requires evidence from many sources of information [19]. The types of protein changes seen in these three species range from inappropriately fused or split genes (see Fig. 6 bottom panel for an example) via insertions/deletions to changes in protein coding start/stop codons positions resulting in extended or truncated coding sequences. All the PASAua outputs are in the form of GFF files viewable within any genome browser. We have made an IGB Quickload server available for easy browsing of the data (http://www.compbio.dundee.ac.uk/Quickload/Dictyostelid_assemblies).

In the *D. discoideum*, *D. fasciculatum*, *D. lacteum* and *P. pallidum* datasets 44, 19, 21 and 175 novel putative genes were identified by PASA respectively (Table 3). These novel genes are in genomic loci with no current annotated gene model or where an existing model is substantially modified. The 44 *D. discoideum* novel genes, defined by 47 transcripts, were examined by eye in IGB [36] against all known *D. discoideum* reference datasets, including predicted gene models (see Additional file 4: Table S2). Of the 47 transcripts, 8 are novel alternate splice transcripts (Additional file 4: Table S2). Although 'novel' suggests there is no existing annotation at the locus of interest, if a gene update is sufficiently different from the reference gene model, PASA may consider that locus as a novel gene. In most of these cases the new transcript represents a corrected model for a previously computationally predicted gene. Many of the predicted gene models were annotated in Dictybase as pseudogenes and were originally ignored by PASAua, which only considers protein coding genes. Fragments of the pseudogenes do

Singh *et al. BMC Genomics* (2017) 18:120

Page 13 of 17

encode ORFs and PASA has reported them as being novel genes (Additional file 4: Table S2), but it is not possible to be sure whether the protein products are expressed in vivo with this data. Out of the 47, it appears only 6 are truly novel as they do not overlap any previously annotated transcripts: novel_model_13, novel_model_23, novel_model_30, novel_model_31, novel_model_38 and novel_model_39. All except novel_model_23 have a sequence match to existing genes, suggesting that they are paralogues. The longest novel unannotated model is 510 AA in length (novel_model_31) and appears to be a duplicate copy of the leucine rich repeat protein lrrA present on the chromosome 2.

Notwithstanding the large number of updates to the existing *D. discoideum* annotations it is clear from Table 3 that there are substantially more changes in the other three species. In particular, the numbers of modified protein sequences are 4, 9 and 10-fold larger in *P. pallidum* (2,252), *D. lacteum* (4,741) and *D. fasciculatum* (5,393), respectively. Similarly, there are 7, 5 and 6-fold more novel alternate splice isoforms in the three species, respectively. For *P. pallidum* (1,321), *D. lacteum* (1,088) and *D. fasciculatum* (842), the gene models were predicted with Augustus (G. Glöckner, *personal communication*) which, given the updates found with PASA, suggests that although the predicted gene models are in the correct locus, many are inconsistent with empirical RNA-seq evidence. With respect to novel genes annotated by PASA, it is notable that *D. fasciculatum* and *D. lacteum* have fewer than either *D. discoideum* or *P. pallidum*. It is unclear why this would be. Many genes were inspected by eye with IGB [36] and overall the annotations appear appropriate, but there are many occasions where human intervention would make further improvements.

Orphan RNAs

As mentioned above, PASA requires that transcripts align to the genome before it can consider them for further analysis. It makes sense to use the genome as a filter for valid transcripts, however this makes the assumption that the genome is complete. Any gaps in the genome that include genes will result in filtering out perfectly valid transcripts.

To determine whether this has happened here, we isolated the transcripts that did not align to the genome and used a process of elimination to identify those transcripts that could be genuine. Table 4 breaks down the number of orphan RNAs and whether they match non-dictyostelid genes ('artefact'), genes in other dictyostelids ('known') or neither ('novel'). *D. fasciculatum* and *D. lacteum* have far more 'novel' non-genome transcripts (6,559 and 6,465, respectively) than *D. discoideum* (69) or *P. pallidum* (26). This is likely due to the fact that these species, which were cultured on bacteria, contain chimeric misassemblies of bacterial and dictyostelid

Table 4 Annotation of Trinity transcripts

	D. discoideum	*P. pallidum*	*D. fasciculatum*	*D. lacteum*
Annotated	28,698	29,405	32,875	28,463
Aligned to reference				
Known	672	170	3,028	240
Novel	1,277	4,505	2,978	2,040
Artefact	98	376	288	134
Not aligned to reference				
Known	14	82	525	945
Novel	69	26	6,559	6,465
Artefact	431	81	208	366

transcripts. Despite this, they still have 525 and 945 'known' transcripts which have sequence matches to other Dictyostelids, higher than seen in *D. discoideum* (14) and *P. pallidum* (82). These transcripts are probably the best candidates for experimental assessment as genuinely non-genome transcripts.

We further investigated the 69 *D. discoideum* 'novel' transcripts with a more sensitive PSI-BLAST search on their longest ORFs and queried their cognate proteins for functional domains using SMART. Table 5 shows the 11 most interesting hits based on the sequence match, read count and ORF length. They are all well expressed and have ORF lengths consistent with functional proteins. Three novel transcripts (comp4660_c0_seq1, comp4660_c4_seq1 and comp5569_c2_seq1) show similar sequence matches to DDB_G0292950 via PSI-BLAST searching, in spite of very low sequence similarity between them. DDB_G0292950 codes for a hypothetical protein which is not conserved in other dictyostelids and is poorly expressed (RPKM <1) at all time points in dictyExpress [37]. The three transcripts match across different parts of DDB_G0292950 indicating that they are different parts of the same larger gene. All transcripts identified in Table 5 were selected for experimental validation via PCR amplification, 8/11 were confirmed.

The comp5787_c28_seq1 transcript has putative homologues in *D. fasciculatum*, *D. purpureum* and *P. pallidum* as shown in Fig. 8. Sequence conservation is high as well as conservation of the Importin-beta N-terminal domain (IBN_N) and HEAT-like repeat (HEAT_EZ) domain architecture although the *D. discoideum* sequences appears to have an additional HEAT repeat domain (Fig. 8).

Genomic cloning of orphan *Dictyostelium discoideum* mRNAs

The newly assembled transcripts that could not be mapped onto the genome are either contaminants or genuine mRNAs for which the genomic counterpart is in an assembly gap of the genome. To investigate the latter option, we used PCR to attempt to amplify the genes

Singh *et al. BMC Genomics* (2017) 18:120

Page 14 of 17

Table 5 Homology and functional information for novel transcripts in *D. discoideum*

Transcript ID	Length	Read Count	ORF Length	% Identity	Closest Sequence Match[a]	Domain
comp1545_c0_seq1	1,035	560	345 aa	41	DFA_11908[b]	XPGI
				29	DDB_G0284987[c]	HhH2
				29	DICPUDRAFT_31355[d]	Low complexity
comp4660_c0_seq1	592	41	198 aa	29	DDB_G0292950[c]	Low complexity
				27	DDB_G0293040[c]	
comp4660_c4_seq1	650	36	217 aa	28	DDB_G0292950[c]	Low complexity
comp5569_c2_seq1	1,891	692	631 aa	30	DDB_G0292950[c]	Low complexity
				29	DDB_G0293040[c]	
				25	DDB_G0292936[c]	
				25	DDB_G0292934[c]	
comp5787_c28_seq1	2,554	1,658	831 aa	72	DICPUDRAFT_51827[d]	IBN_N
				50	XP_004351151[b]	Pfam:HEAT_EZ Low complexity
				51	PPL_10069[e]	
comp5953_c11_seq1	882	425	294 aa	50	XP_004359777.1[b]	AAA_11, AAA_12
comp5953_c48_seq1	844	293	281 aa	63	DICPUDRAFT_84867[d]	AAA_11
				63	XP_004359777.1[b]	
				41	DFA_01811[b]	
comp6065_c2_seq1	556	209	169 aa	49	DICPUDRAFT_84867[d]	Pfam:DUF2439
				41	DFA_01810[b]	Low complexity
comp2066_c1_seq1	580	212	193 aa	82	DICPUDRAFT_160104[d]	Pfam:UAA
comp4678_c1_seq1	455	102	67 aa		No significant hit	No domain
comp470_c0_seq1	779	114	71 aa		No significant hit	Transmembrane region

[a]Matches to related species via PSI-BLAST
[b]*D. fasciculatum*
[c]*D. discoideum*
[d]*Dictyostelium purpureum*
[e]*P. pallidum*

from *D. discoideum* genomic DNA (gDNA). Oligo-nucleotide primers were designed to amplify regions of about 0.5 – 1.4 kb of 11 transcripts (Additional fie 1: Table S4). The amplified size can however be larger due to the presence of introns. For eight transcripts, corresponding gDNAs could be amplified, but for two genes two transcripts were part of the same gene (Additional file 2: Figure S3). The six genes in total were all protein coding genes. For three transcripts, comp470_c0_seq1, comp4678_c1_seq1 and comp2066_c1_seq1 no PCR products were obtained, but the first two transcripts contained multiple stop codons in all reading frames and are likely assembly errors. The amplified PCR products were sub-cloned and sequenced from both ends. Sequences were assembled and aligned with the transcript sequence. Apart from just a few mismatches, the transcript and gDNA sequences were identical (Additional fie 1: Table S5). Only one amplified fragment contained introns (Additional file Table 1: S5). Six out of seven of the protein coding orphan transcripts therefore had a counterpart in the genome. Overall,

deciphering the genomes of organisms is a key step in being able to probe their biology. With the advent of high-throughput sequencing technologies this has become a simpler problem to solve. Yet it is still not trivial to finish a genome assembly without any gaps [38]. The genome sequence on its own, however, imparts very little functional information and requires annotation of genes, transcripts and regulatory regions to be scientifically useful [7]. Many gene annotation methods are dependent on either homology to related species [30, 39] or via gene finding prediction algorithms [40, 41] or ideally both. However, the first method will miss all unusual or species-specific genes, while both methods fall short of accurately predicting intron-rich genes, genes with alternative or non-canonical splice sites or genes with very short exons. The ability to generate a whole transcriptome for a given species and use it to empirically annotate the genome has the power to confirm and correct any errors introduced with other methods. This has been achieved with expressed sequence tags (ESTs) in the past [42], but now can be performed with RNA-Seq short read data [32].
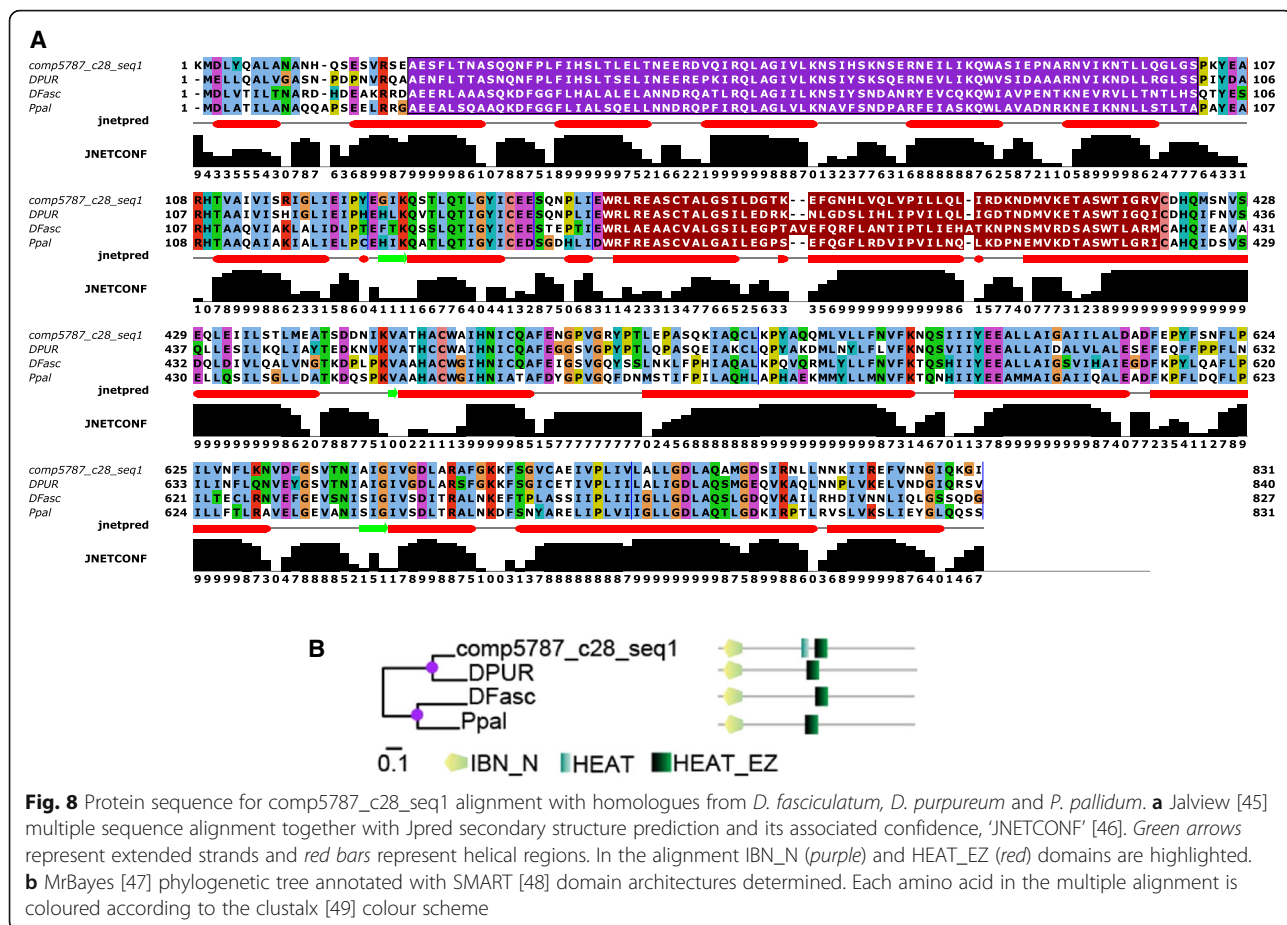
Singh *et al. BMC Genomics* (2017) 18:120

Page 15 of 17

Fig. 8 Protein sequence for comp5787_c28_seq1 alignment with homologues from *D. fasciculatum, D. purpureum* and *P. pallidum*. **a** Jalview [45] multiple sequence alignment together with Jpred secondary structure prediction and its associated confidence, 'JNETCONF' [46]. *Green arrows* represent extended strands and *red bars* represent helical regions. In the alignment IBN_N (*purple*) and HEAT_EZ (*red*) domains are highlighted. **b** MrBayes [47] phylogenetic tree annotated with SMART [48] domain architectures determined. Each amino acid in the multiple alignment is coloured according to the clustalx [49] colour scheme

This evidence-based methodology is non-trivial and is not perfect. There are examples where the data is not adequately represented in the final transcript set when interpreted by the human eye. In addition, PASA only defines protein-coding genes meaning that all non-coding RNAs (ncRNAs) will be ignored and will not be in the final annotation unless already identified in the reference. Identifying ncRNAs is difficult as they have no obvious products and well-defined sequence features [43]. This does not negate their importance or relevance to the Dictyostelia.

Conclusion

In this study, we present a *de novo* transcriptome assembly in four social amoeba species for the first time and with these data we have:

- Created a final set of of 11,523 (*D. discoideum*), 12,849 (*P. pallidum*), 12,714 (*D. fasciculatum*) and 11,315 (*D. lacteum*) transcripts.
- Substantially updated the existing transcript annotations by altering models for more than half of all the annotated transcripts.

- Identified changes to thousands of transcripts in the predicted gene models of *P. pallidum, D. lacteum* and *D. fasciculatum* many of which affect the protein coding sequence.
- Identified and validated six novel transcripts in *D. discoideum*.
- Putatively identified dozens to hundreds of novel genes in all four species.
- Identified errors in the genome sequence of at least two *D. discoideum* genes (*clcD* and DDB_G0279477). With the possibility of, at least, another 104 genes having sequence errors.
- Found hundreds of putatively alternatively spliced transcripts in all species, something which has not been identified before in *P. pallidum, D. lacteum* or *D. fasciculatum*.

By combining methodologies we now have a better and more complete description of the transcriptome for these four species. This is not an end-point, however, but a further step towards fully finished genomes. More data and more manual refinement will be required to improve the annotations further.

Singh *et al. BMC Genomics* (2017) 18:120

Page 16 of 17

Additional files

Additional file 1: Figure S1. TAGC plot for *D. fasciculatum* (A) and *D. lacteum* (B) before and after filtering. Each colour blob represents different taxa with unmatched transcripts are shaded in grey. The unannotated grey coloured transcripts after filtration set further filtered by high GC and low read coverage. This plot shows a major blob of transcripts that are annotated with the Dictyostelium fasciculatum species with high coverage and lower GC content. Other contaminations form E.coli, pseudomonas fluorescence and other species has also been highlighted with different colours. These contaminations clearly make different blobs with lower read coverage and high GC content. However, it's good to see that there are some other transcripts that showing matched to dictyostelium discoideum- that clearly reflect the presence of some novel unannotated transcripts in the new assembly. **Figure S2** A comparison of assembled transcripts read count. The boxplots represent the range between the 1st and 3rd quartiles of the data by the coloured boxes, the median is the horizontal bar and points shown beyond the whiskers are >95% of the data. **Table S3** Transrate good contigs. **Table S4** Olignucleotide sequences. **Table S5** Alignment with DNA sequence of PCR product. (DOCX 787 kb)

Additional file 2: Figure S3. Annotated domain diagrams for 16 *D. discoideum* developmentally relevant proteins which have had their protein sequence altered. (PDF 367 kb)

Additional file 3: Table S1. *D. discoideum* GFF file of genomic positions with introns < 5 bp. (TXT 9 kb)

Additional file 4: Table S2. List of PASAua novel genes in *D. discoideum*. (XLSX 13 kb)

Abbreviations

BLAT: BLAST-like alignment tool; bp: Basepairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; cDNA: Coding DNA; CDS: Coding sequence; contig: Contiguous sequence; ERCC: External RNA controls consortium; EST: Expressed sequence tag; gDNA: Genomic DNA; GMAP: Genome Mapping and Alignment Program; mRNA: Messenger RNA; ncRNA: Non-coding RNA; PASA: Program to Assemble Spliced Alignments; PCR: Polymerase chain reaction; RNA-seq: RNA sequencing; rRNA: Ribosomal RNA; UTR: Untranslated region

Availability of data and materials

The datasets supporting the conclusions of this article are available at figShare, http://dx.doi.org/10.6084/m9.figshare.3384364, the European Nucleotide Archive (project IDs: PRJEB12875, PRJEB12907, PRJEB12908, PRJEB12909) and as an Integrated Genome Browser [36] QuickLoad server at http://www.compbio.dundee.ac.uk/Quickload/Dictyostelid_assemblies.

Authors' contributions

RS and CC developed experimental design, performed analyses and wrote the manuscript. HML performed gene amplification by PCR. PS and GJB contributed to experimental design and manuscript preparation. CS and GG prepared and sequenced RNAs, respectively. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Computational Biology, School of Life Sciences, University of Dundee, Dow Street, Dundee, UK. [2]Cell and Development Biology, School of Life Sciences, University of Dundee, Dow Street, Dundee, UK. [3]Institute of Biochemistry I, Medical Faculty, University of Cologne, D-50931 Cologne, Germany. [4]Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, D-12587 Berlin, Germany.

References

1. Reddy TBK, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA, Kyrpides NC. The genomes OnLine database (GOLD) v. 5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2015;43(D1):D1099–106.
2. Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. Genome project standards in a new era of sequencing. Science. 2009;326(5950):236–7.
3. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33:W465–7.
4. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–9.
5. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks (vol 7, pg 562, 2012). Nat Protoc. 2014;9(10):2513.
7. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329–42.
8. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech. 2011;29(7):644–52.
9. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92.
10. Xie YL, Wu GX, Tang JB, Luo RB, Patterson J, Liu SL, Huang WH, He GZ, Gu SC, Li SK, et al. SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30(12):1660–6.
11. Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015;31:3938–45.
12. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Hubbard TJ, Guigo R, Harrow J, Bertone P, Consortium R. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10(12):1177–84.
13. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Res. 2016;26:1134–44.
14. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, et al. The genome of the social amoeba Dictyostelium discoideum. Nature. 2005;435(7038):43–57.
15. Heidel AJ, Lawal HM, Felder M, Schilde C, Helps NR, Tunggal B, Rivero F, John U, Schleicher M, Eichinger L, et al. Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. Genome Res. 2011;21(11):1882–91.
16. Glöckner G, Lawal HM, Felder M, Singh R, Singer G, Weijer CJ, Schaap P. The multicellularity genes of dictyostelid social amoebas. Nat Commun. 2016;7:12085.
17. Chisholm RL, Gaudet P, Just EM, Pilcher KE, Fey P, Merchant SN, Kibbe WA. dictyBase, the model organism database for Dictyostelium discoideum. Nucleic Acids Res. 2006;34:D423–7.
18. Felder M, Romualdi A, Petzold A, Platzer M, Sühnel J, Glöckner G. GenColors-based comparative genome databases for small eukaryotic genomes. Nucleic Acids Res. 2013;41(Database issue):D692–9.

Singh *et al. BMC Genomics* (2017) 18:120

Page 17 of 17

19. Schurch NJ, Cole C, Sherstnev A, Song J, Duc C, Storey KG, McLean WHI, Brown SJ, Simpson GG, Barton GJ. Improved annotation of 3′ untranslated regions and complex loci by combination of strand-specific direct RNA sequencing, RNA-Seq and ESTs. Plos One. 2014;9(4):e94270.

20. Chen ZH, Singh R, Cole C, Lawal HM, Schilde C, Febrer M, Barton GJ, Schaap P. Adenylate cyclase A acting on PKA mediates induction of stalk formation by cyclic diguanylate at the Dictyostelium organizer. Proc Natl Acad Sci USA. 2017; Early edition, 5 January 2017.

21. Schilde C, Lawal HM, Noegel AA, Eichinger L, Schaap P, Glöckner G. A set of genes conserved in sequence and expression traces back the establishment of multicellularity in social amoebae. BMC Genomics. 2016;17(1):871.

22. FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

23. Macmanes MD. On the optimal trimming of high-throughput mRNA sequence data. Front Genet. 2014;5:13.

24. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.

25. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet. 2013;4:237.

26. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

27. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–66.

28. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.

29. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

31. Jentsch TJ, Stein V, Weinreich F, Zdebik AA. Molecular structure and physiological function of chloride channels. Physiol Rev. 2002;82(2):503–68.

32. Hassan MA, Melo MB, Haas B, Jensen KDC, Saeij JPJ. De novo reconstruction of the toxoplasma gondii transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. BMC Genomics. 2012;13:696.

33. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-Seq data. Am J Hum Genet. 2013;93(4):641–51.

34. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, Corvin AP, Morris DW. Development of strategies for SNP detection in RNA-Seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. Plos One. 2013;8(3):e58815.

35. Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, Sun XW. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 2013;14(1):1–14.

36. Nicol JW, Helt GA, Blanchard SG, Raja A, Loraine AE. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics. 2009;25(20):2730–1.

37. Rot G, Parikh A, Curk T, Kuspa A, Shaulsky G, Zupan B. dictyExpress: a Dictyostelium discoideum gene expression database with an explorative data analysis web-based interface. BMC Bioinformatics. 2009;10:265.

38. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11(January 2010):31–46.

39. Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics. 2002;18(3):440–5.

40. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.

41. Rabiner LR. A tutorial on hidden Markov-models and selected applications in speech recognition. P Ieee. 1989;77(2):257–86.

42. Gissi C, Pesole G. Transcript mapping and genome annotation of ascidian mtDNA using EST data. Genome Res. 2003;13(9):2203–12.

43. Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP. Identification and properties of 1,119 candidate LincRNA loci in the drosophila melanogaster genome. Genome Biol Evol. 2012;4(4):427–42.

44. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 2015;43(D1):D257–60.

45. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.

46. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 2008;36:W197–201.

47. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754–5.

48. Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res. 1999;27(1):229–32.

49. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with clustal x. Trends Biochem Sci. 1998;23(10):403–5.