



Extension of Partial Gene Transcripts by Iterative Mapping of RNA-Seq Raw Reads

Kumar Saurabh Singh , Bartłomiej J. Troczka, Katherine Beadle, Linda M. Field, T. G. Emyr Davies , Martin S. Williamson, Ralf Nauen, and Chris Bass

Abstract—Many non-model organisms lack reference genomes and the sequencing and *de novo* assembly of an organisms transcriptome is an affordable means by which to characterize the coding component of its genome. Despite the advances that have made this possible, assembling a transcriptome without a known reference usually results in a collection of full-length and partial gene transcripts. The downstream analysis of genes represented as partial transcripts then often requires further experimental work in the laboratory in order to obtain full-length sequences. We have explored whether partial transcripts, encoding genes of interest present in *de novo* assembled transcriptomes of a model and non-model insect species, could be further extended by iterative mapping against the raw transcriptome sequencing reads. Partial sequences encoding cytochrome P450s and carboxyl/cholinesterase were used in this analysis, because they are large multigene families and exhibit significant variation in expression. We present an effective method to improve the contiguity of partial transcripts *in silico* that, in the absence of a reference genome, may be a quick and cost-effective alternative to their extension by laboratory experimentation. Our approach resulted in the successful extension of incompletely assembled transcripts, often to full length. We experimentally validated these results *in silico* and using real-time PCR and sequencing.

Index Terms—RNA sequencing, RNASeq, transcript extension, detoxification genes, *de novo* transcriptome assembly, iterative read mapping

1 INTRODUCTION

THE genomics revolution has led to a dramatic rise in the number of organisms with a sequenced genome. Despite these efforts many non-model species continue to lack this genetic resource. As an alternative, recent advances in next generation sequencing, specifically RNAseq [1], has made it possible to rapidly and affordably characterise the coding element of an organisms genome. The sequencing of an organisms transcriptome has therefore been particularly harnessed, especially by the community working on non-model species. Assembling a transcriptome without a known reference is computationally difficult and requires the use of a *de novo* assembler, such as Trinity [2], SOAPdenovo-trans [3], Velvet-Oases [4], Trans-ABYSS [5] and T-IDBA [6]. These *de novo* transcriptome assemblers have adopted algorithms commonly used for genome assembly. However, there are significant differences in sequence data obtained by RNAseq and DNAseq. For example, whilst genomic sequence coverage is generally uniform across the genome, transcriptome coverage is highly heterogeneous and this excludes the use of coverage information in resolving repeated motifs [7].

- K.S. Singh, K. Beadle, and C. Bass are with the College of Life and Environmental Sciences, University of Exeter Penryn Campus, Penryn, Cornwall TR10 9FE, United Kingdom. E-mail: {k.saurabh-singh, k.beadle, c.bass}@exeter.ac.uk.
- B.J. Troczka, L.M. Field, T.G.E. Davies, and M.S. Williamson are with the Department of Biological Chemistry and Crop Protection, Rothamsted Research Harpenden, Harpenden, Hertfordshire AL5 2JQ, United Kingdom. E-mail: {bartek.troczka, lin.field, emyr.davies, martin.williamson}@rothamsted.ac.uk.
- R. Nauen is with the Bayer AG, Crop Science Division, Monheim am Rhein 40789, Germany. E-mail: ralf.nauen@bayer.com.

Manuscript received 19 Mar. 2017; revised 25 June 2018; accepted 1 Aug. 2018. Date of publication 13 Aug. 2018; date of current version 31 May 2019.
(Corresponding author: Kumar Saurabh Singh.)

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2865309

All *de novo* assemblers are based on constructing, simplifying, and resolving *de bruijn* graphs [8], to extract likely isoforms in the form of transcripts. Among others, the Trinity *de novo* assembler has made considerable improvements in its algorithms used to assemble RNAseq data [9]. Instead of constructing *de-bruijn* graphs directly from reads, Trinity first creates a linear extended contig based on a fixed k-mer value of 25 base pairs (bp) [10]. It then groups overlapping contigs and constructs a *de-bruijn* graph for each component/sub-component. Sequencing reads are then mapped to these graphs and potential isoforms are extracted using parallel computing. Paired-end and mate-pair data can be utilized to scaffold contigs resulting in improved contiguity of the transcriptome. Despite these advances, all *de novo* assemblers, including Trinity, suffer from false positives, incompleteness and redundancy, which combine to result in a transcriptome comprising full length and partial transcripts [11]. Partial transcripts can result from sequencing errors, heterogeneity in expression, allelic variation or tandem repeats. Such factors present problems for *de novo* assemblers in finding optimal sets of linear paths constructed from the *de-bruijn* graphs. Methods have been developed to improve the contiguity of sequence assemblies. TransPS [12], scaffolds the pre-assembled transcriptome using a reference species to improve transcriptome coverage and reduce contig redundancy. Other methods like transcriptome optimization, though helpful in increasing total gene coverage and minimizing chimeras, do not specifically address the issues related to partial transcript assembly. IMAGE [13] was developed with the goal of improving genome assemblies by targeting local Illumina reads from gapped regions. TRAM [14] applies a *tblastn* [15] targeted search of Illumina sequencing datasets to recover protein-coding reads of interest and then locally assembles the subset of reads to get a longer sequence or contig. Unlike TRAM, aTRAM [16] uses a MapReduce approach to speed up the query search against the raw short reads. PRICE [17] was designed to address the challenge of assembling any nucleic acid species genome from metagenomic datasets and uses paired-read information to iteratively increase the size of existing contigs. Most contig extension methods are genome based and built on the assumption of uniform coverage and are therefore less suited to handling transcriptome data. Moreover, to our knowledge, there is currently no computational method available which specifically addresses the issue of improving or extending candidate mis-assembled or partial transcripts, derived from *de novo* transcriptome assemblies. The alternative to improving the contiguity of partial transcripts *in silico* is to extend them experimentally in the laboratory. A commonly used technique for this is Rapid Amplification of cDNA Ends (RACE). This approach can be used to obtain the full length sequence of an RNA transcript from a partial known sequence, but it is relatively costly and time consuming, so typically can only be carried out for just a small number of candidate partial gene sequences of interest. We recently sequenced the transcriptome of a non-model insect species, *Osmia bicornis*, and *de novo* assembled the transcriptome using Trinity. Our research focuses on characterising genes encoding detoxification enzymes in this species, specifically cytochrome P450s (P450s) and carboxyl/cholinesterases (CCEs). Transcripts belonging to these supergene families were present in our transcriptome assembly as both full length and partial gene sequences. These genes offer a good test for extension pipelines, because they are in large families (most insects have 36 - > 150 unique P450 genes) and different members can exhibit significant differences in their

expression. The latter means we would expect to observe marked differences in the transcriptome sequence coverage obtained for different genes. The aim of this study was to explore the possibility of extending *de novo* assembled incomplete transcripts up to full length genes using detoxification genes derived from transcriptome assemblies of *O. bicornis* (as a non-model species) and *D. melanogaster* (as a model species). We first carried out a comparative analysis using the most recent published methods for optimization of transcriptome assemblies to systematically evaluate their extension capabilities. We subsequently developed and tested an iterative read mapping strategy and used this to extend partial transcripts of our candidate genes of interest. Extended transcripts were validated *in silico* by BLAST against the NCBI nr database. Finally, several of the extended transcripts of *O. bicornis* were validated experimentally using RT-PCR, cloning and sequencing.

2 MATERIALS AND METHODS

2.1 Insects

Osmia bicornis (Red Mason Bee) cocoons were purchased from Dr. Schubert Plant Breeding (Landsberg, Germany) and stored at 4°C in complete darkness until emergence. Bees were hatched in a controlled environment (25°C; 50 percent relative humidity; L16:D8) and fed on 50 percent sucrose solution supplemented with pollen (Sussex Wholefoods, Sussex).

2.2 Extension Using Simulated Reads

The extension methodology was first checked using simulated reads. Reads were generated using the neat-genreads (<https://github.com/zstephens/neat-genreads>) tool, based on 3 reference P450 genes from *D. melanogaster* genome, namely, CYP12C1 (NM_001300180.1), CYP4G1 (NM_080292.4) and CYP9B2 (NM_078922.4). The performance of the extension method was checked on two types of simulated data. One with constant read coverage and second with variable read coverage. For the generation of simulated read datasets a default error model was used. Details on the simulations can be found here: https://github.com/kumarsaurabh20/transcripts_extension/tree/master/simulations

2.3 Transcriptome Assemblies

First transcriptomic datasets were collected from the FlyAtlas 2 project (<https://academic.oup.com/nar/article/46/D1/D809/4563305>). The project examined the expression of the genes of *D. melanogaster* in a range of tissues from adults and larvae. Data was downloaded from <https://www.ebi.ac.uk/ena/data/view/PRJEB22205> with run accessions: ERR2098815, ERR2098816, ERR2098817, ERR2098818, ERR2098819 and ERR2098820. A second transcriptomic dataset was generated from RNA sequencing of *O. bicornis*, across two lanes of an Illumina HiSeq 2,500 flow cell (100 bp paired end reads). Analysis of all RNA sequencing raw data was performed using custom pipelines. Briefly, the quality of raw data was checked using FastQC (version 0.11.5) [18] program (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and processed using Trimmomatic (version 0.32) [19] to remove Illumina specific adapters and low quality reads. Transcriptome assembly was done using the Trinity Software Suite and the assembled transcriptome was annotated with functional descriptions and gene ontology terms using Blast2GO (version 3.0.11) [20]. All transcripts, complete and incomplete, were extracted from the transcriptome data and those encoding P450s curated in Geneious (Biomatters, New Zealand).

2.4 Testing of Existing Methods

We evaluated the performance of 3 tools, TransPS (version 1.1.0), aTRAM (version 1.04) and PRICE (version 1.2). All the existing methods were tested using their default settings. Trans-PS was run

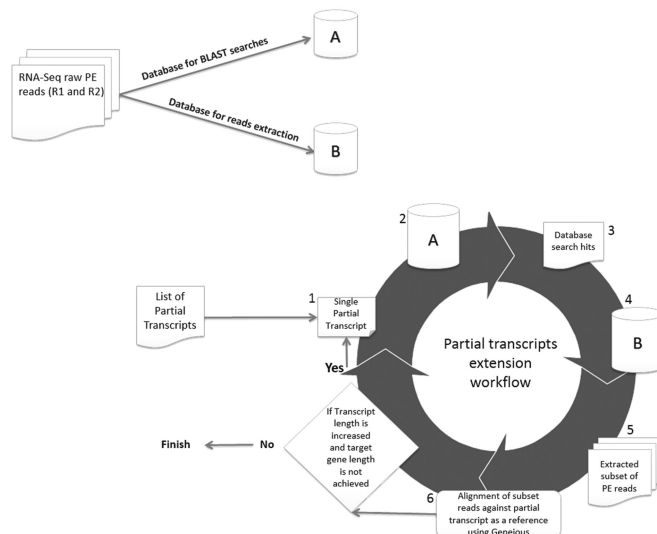


Fig. 1. Transcript extension pipeline (i) RNA-Seq raw reads formatted database. Database A is used for sequence searching and database B for mate pair retrieval. (ii) Schematic illustrating all the steps involved in the extension pipeline.

using the *O. bicornis* transcriptome and the command was called specifying a single fasta file with 92,484 transcripts. For aTRAM, raw RNA-seq reads were used to create an initial database. The program was run using 18 *O. bicornis* partial transcripts. aTRAM was initially run using the default number of iterations i.e., 5, followed by up to 10 iterations per run. Additionally, the aTRAM run was flagged as *complete* which automatically quits the program when a complete homolog is recovered. Similar to aTRAM, 18 partial transcripts were used as input seed contigs in PRICE with the following specifications: *-nc* 30/16, *-dbmax* 72, *-mol* 30, *-tol* 20, *-mpi* 80, *-target* 90 0. We tested PRICE using different numbers of cycles (*-nc*) starting from 16 to 30 and specified the target *-target* mode to limit the final contigs to extensions of input seeds, rather than keeping all of the assembled contigs.

2.5 Method Development

The developed workflow is based on iterative searching of raw reads against a query sequence. The query sequences are candidate partial transcripts that require extension. The extension workflow utilizes the BLAST algorithm for sequence database creation and sequence searching with modified parameters for highly optimized sequence searching. As an alternative to BLAST, nhmmer (version 3.1b2) [21] and LASTZ [22] algorithms were also tested. The workflow creates two raw reads database, referred here as A and B (Fig. 1), for subsequent use in downstream analysis. Database A is a BLAST database whereas B represents processed raw reads in FASTQ format. Once the databases are formatted, each query sequence is searched against database A using BLAST/nhmmer (Fig. 1 ii). The BLAST/nhmmer output file is further processed to a list file which is used to extract a subset of reads related to the query sequence. In the final step, the reads subset along with the query sequence are imported to Geneious software. Geneious in-built read aligner maps the subset of paired-end reads against the query sequence (Fig. 2). An extended sequence, obtained after the Geneious alignment, is again used as a query sequence to database A and B. This process is repeated multiple times until no more extension is achieved. A helper program to generate database A and B is available here: https://github.com/kumarsaurabh20/transcripts_extension

2.6 Experimental Validation

2.6.1 RNA and cDNA Synthesis

Total RNA was extracted from a single 24h-old adult female *O. bicornis* (which had been flash frozen in liquid nitrogen) using ISOLATE

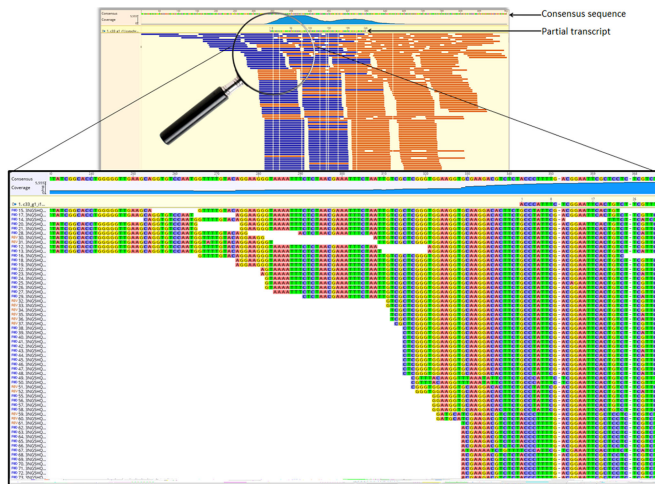


Fig. 2. Visual layout of how paired end reads are aligned against the partial transcript (CYP-2) and how the consensus sequence is constructed. Blue and orange coloured reads shows two mates of a paired end data. A coverage plot, in blue, displays the total read depth on the partial transcript.

II RNA Mini Kit (Bioline, UK) following the manufacturers protocol. The quality and quantity of eluted RNA was determined by visualisation on a 1 percent TAE agarose gel electrophoresis and Nanodrop 1,000 (Thermo Fisher Scientific, USA) measurements. 4g of total RNA was used to synthesise cDNA in 20l reactions containing SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, USA) and Oligo dT (15) primers (Promega, USA) following the manufacturers instructions.

2.6.2 PCR Reactions

PCR primers were designed using Geneious (version 8.1.3) software (Table 1) and synthesised by Sigma Aldrich (UK). 25L PCR reactions contained DreamTaq Green DNA Polymerase x2 mastermix (Thermo Fisher Scientific, USA), 1l of cDNA and 10 pmol of forward and reverse primers. Cycling conditions were 95°C for 2 min followed by 35 cycles of 95°C for 20s, 50°C-60°C for 30 s and 72°C for 2-6 min, with a final extension step of 72°C for 5 min. PCR products were visualised by agarose gel electrophoresis and bands of expected size were cut from the gel and the DNA either purified via QIAquick PCR purification kit (Qiagen, Germany) or if artefact bands were present via QIAquick gel extraction kit (Qiagen, Germany) following manufacturers protocols. PCR products were ligated into the positive selection cloning vector pJET1.2 according to the CloneJET PCR cloning kit guide (Thermo Fisher, Scientific, USA) and transformed into XL-1 blue competent cells (Agilent, USA). Plasmids were purified using the GeneJET plasmid mini prep kit following the manufacturers protocol. All sequencing was done by the Eurofins DNA value read service (Eurofins Scientific group, Belgium) and sequence analysis used Geneious (version 8.1.3) software.

3 RESULTS AND DISCUSSION

A prerequisite for the functional analysis of candidate genes of interest is, in many cases, knowledge of the complete open reading frames (ORFs) of that gene. In this study, we were specifically interested in identifying a method that could be used to extend *in silico* partial transcripts of candidate genes identified in *de novo* assembled transcriptomes using raw RNA-seq data. Any such method would offer a cheaper and quicker alternative to their extension using RACE or similar techniques. We initially tested existing software that had been developed for complete transcriptome improvements by scaffolding, performing micro-assemblies (*de novo* assemblies on subset of data) or by whole transcriptome optimization, as improving individual incomplete transcripts is an unexplored area of transcriptomic research.

TABLE 1
Primer Sequence Information for Successfully Validated Genes

Gene name	Primer name	Primer sequence
CYP-3	CYP-3-F1	GTAACATATTGAAACTGCA
	CYP-3-F2	TACAGAGGTTGAAAAAGGA
	CYP-3-R1	ACCATCTAGTATACAAGAGATATC
	CYP-3-R2	GAATTGTCTCATTTTGTCTTCT
CYP-4	CYP-4a-F1	GAACAGCAACTTCAGATT
	CYP-4a-F1.1	AACTACACTTTTGATGGCAC
	CYP-4a-F2	GACGTGATTGGTAGTTGT
	CYP-4a-R1	CTGAATACTGTTAAGCATTAC
	CYP-4a-R1.1	GTAAATATGGGGAGAGC
	CYP-4a-R2	TCATTCTCTGCCTGTATTT
	CYP-4a-R2.2	GGCTCAACTTGCCTCAAACCTCTCC
CYP-6	CYP-6-F1	TGCTACTCATCTTTAGTGTATAA
	CYP-6-F2	ATCTGTGGCTCCTTATTG
	CYP-6-R1	ATCATAGCTGTACAAAGATACT
	CYP-6-R2	CGTCTATAAAATTTGTACACTCGT
CYP-12	CYP-12-F1	GCAACCATATAATGTTTTCAAT
	CYP-12-R1	CATATTTGAAAAAGGTATCTCC
CYP-28	CYP-28b-F1	AGTGTTTTAGTCAACTCAATCA
	CYP-28b-R1	AGTGAAAAATTATTAGGTGTACAA
CYP-51	CYP-51b-F1	CGATGGTTGGAGTATTGG
	CYP-51b-R1	TTAAGCTTTTGTCAATTTTAAATAT
CYP-52	CYP-52-F1	GTCTTAGAATGCAACGTCTG
	CYP-52-F2	TACAAAATACGAAACAGGCA
	CYP-52-R1	GATTTCTATAGTTTGTCACTTGG
	CYP-52-R2	GAAGAATCGCCAGACTTG
CCE-8	CCE-8 F	CCTGTATCGAGTAGAAATACCACAAC
	CCE-8 R	TCGTCTACTGTACGATTATGCATCACC
CCE-11	CCE-11 F	GAAGTTAAACAGTCGTTCAACATG
	CCE-11 R	ACAGTATTATACAGAGCTTTTTGGCC

3.1 Extension of Gene Fragments Using Simulated Reads

The three reference genes (Section 2.2) used to generate simulated reads were of different length. Two types of read sets were generated, one with consistent coverage and another with variable coverage in order to check if variation in coverage impedes the extension of partial fragments. 30-50 bp fragments were extracted from full length genes and the current workflow was applied separately with the two read sets. Although, full length genes were recovered in all three cases, more iterations were needed to extend the partial fragments with reads with variable coverage. Extended details on simulation study can be found here: https://github.com/kumarsaurabh20/transcripts_extension/tree/master/simulations

3.2 Transcriptome Assemblies

The *D. melanogaster* and *O. bicornis* transcriptome assembly resulted in 93,099 and 92,484 total transcripts with contig N50 of 3223 and 4,002 bp respectively. The *Drosophila* transcriptome resulted in 222 transcripts with P450 annotations. Out of these, 24 transcripts were found to be partial and of less than 1,000 bp in lengths. In the case of *O. bicornis*, 24 full length and 18 partial transcripts encoding P450s were identified. In addition, 15 full length transcripts and 8 partial transcripts encoding CCEs were identified.

3.3 Testing of Existing Methods for Transcript Extension

We evaluated three existing programs; TransPS, PRICE and aTRAM for partial transcript extension using *O. bicornis* dataset. It is important to emphasize that all three programs were initially developed for somewhat different applications (see Introduction and Methods). TransPS is a transcriptome postscaffolding tool and requires whole transcriptomes, i.e., a complete *de novo* assembled transcript set, as input data. Out of 92,484 Trinity reported transcripts, Trans-PS selected 5,091 *O. bicornis* transcripts as one-to-one matches with the *Megachile rotundata* reference proteome with

TABLE 2
List of All Transcripts Used in the Extension Workflow with the Names, Original Trinity Assembled Length, and Final Extended Length

Gene family	Gene category	Partial transcript ID	Names	Original length	Extended length	Validation		
						In-Silico	Selected	Validated
CYPs	CYP-2	c33_g1_i1	CYP6AS124	246	1,742	YES	YES	NO
	CYP-3 ^{\$}	c3135_g1_i1	CYP6AS134	226	2,350	YES	YES	YES
	CYP-4	c7807_g1_i1	CYP6AS133	312	2,889	YES	YES	YES
	CYP-4	c7807_g2_i1	CYP6AS133	649	3,081	YES	YES	NO
	CYP-6*	c9730_g1_i1*	CYP343A1	1,274	1,644	YES	YES	YES
	CYP-12	c19755_g1_i1	CYP6AS127	970	2,763	YES	YES	YES
	CYP-28*	c24062_g1_i1*	CYP6AS121	254	2,049	YES	YES	YES
	CYP-34	c25730_g2_i1	CYP302A1	2,608	2,891	YES	YES	YES
	CYP-37	c25900_g6_i1	CYP343A1	942	1,736	YES	YES	NO
	CYP-37	c25900_g6_i2	CYP343A1	1,169	2,316	YES	YES	NO
	CYP-51	c27345_g1_i1	CYP6AS121	1,320	2,415	YES	YES	YES
	CYP-51	c27345_g1_i2	CYP6AS121	913	2,202	YES	YES	NO
	CYP-51	c27345_g1_i3	CYP6AS121	1,180	2,604	YES	YES	NO
	CYP-52	c35166_g1_i1	CYP334A1	222	1,874	YES	YES	YES
	CYP-53	c38361_g1_i1	NA	229	911	NO	NA	NA
	CYP-54	c30208_g1_i1	NA	237	2,353	NO	NA	NA
CCEs	CCE-5	c23930_g1_i5	NA	982	2,617	YES	YES	NO
	CCE-8	c24685_g1_i1	NA	664	1,974	YES	YES	YES
	CCE-8	c24685_g1_i2	NA	217	1,957	YES	NA	NA
	CCE-8	c24685_g1_i3	NA	397	2,115	NO	NA	NA
	CCE-8	c24685_g1_i4	NA	630	2,061	YES	YES	YES
	CCE-8	c24685_g1_i5	NA	1,997	1,997	YES	NA	NA
	CCE-8	c24685_g1_i6	NA	1,167	1,729	YES	NA	NA

A detailed overview of all evaluated transcripts is provided in Supplementary File 3, available online. *transcripts were extended using bait mapping approach. In silico validation involves confirmation of gene family using BLASTX and confirmation of the contiguity of ORFs. \$ See PCR validation of extended transcripts for an alternative CYP-3 transcript. All NA transcripts either shares $\geq 90\%$ homology with other isoforms under a gene category or could not be in silico validated, so not selected for experimental validation.

remaining transcripts unused and categorized as redundant transcripts. Based on BLASTX coordinates, TransPS further selected 473, out of 5,091 transcripts, for scaffolding. All scaffolded transcripts were then interrogated for those encoding P450s and CCEs. Only one P450, CYP450 6a23 was extended (to full gene length). In our list of partial P450 sequences, CYP450 6a23 equates to partial transcript CYP-51. Unlike Trans-PS, where full transcriptome was used, we then tested 16 partial P450 transcripts using aTRAM and PRICE for extension. The MapReduce algorithm implemented in aTRAM enables fast searching of paired-end data. It divides the read search into many smaller parallel executable jobs which speeds up computational time. Despite aTRAMs faster run time, we managed to assemble only 1 partial transcript, CYP-51 to full length. In this case the transcript length increased from 1,320 to 1,878 bp. A further 6 partial transcripts, though flagged as complete, were not extended to full length coding sequences. The aTRAM results suggest that the gene complexity impedes the extension process. Furthermore, the choice of assembler has a big impact on transcriptome extension. Of the existing extension methods that we tested, PRICE performed the best. Although, PRICE was originally designed to assemble complex metagenomic data, the core algorithm also shows potential for transcript extension, perhaps because metagenomic data also suffers from variable coverage. Using PRICES target flag, it is possible to extend a single partial sequence. Despite its longer run time, PRICE managed to fully extend 8 partial transcripts. One possible explanation for its success is that each assembly cycle includes a mapping step whereby reads are aligned to the target sequence followed by paired-end assembly using target sequence as a reference. This results in a longer extended sequence which includes the original target sequence in the middle with extension possible in both directions. Our analysis revealed that an exclusive k-mer based *de novo* assembly approach, as adopted by Trinity and aTRAM, is not sufficient to completely assemble a complex gene, unless k-mers are highly optimized for a particular set of data. Fine mapping of raw reads against a target sequence helps in resolving sequence abnormalities

and mis-assemblies. Despite the encouraging results, PRICE failed to extend 50 percent of the partial P450 transcripts. Further investigation revealed two reasons for this: 1) Low coverage of some of the transcripts and 2) PRICE failed to resolve some of the ambiguities in *de bruijn* graphs. Using PRICE for extension does not give improvements in sequence length if these graph based ambiguities are not resolved and collapsed within the same target sequence. All result files, including transcriptome assemblies, can be accessed from: <https://zenodo.org/record/1297276>

3.4 Development of a Novel Method for Transcript Extension

Independent local alignment of each sequencing read to a reference sequence is a logical approach to read mapping and works well when the target sequence is highly similar to the reference sequence. However, problems are encountered when there is significant divergence between the reference and target sequence, particularly if this is caused by insertions and deletions. Importantly, local alignment algorithms often tend to truncate the aligned region to improve the overall level of sequence identity and this truncation (also known as soft or hard clipping), as implemented by all popular short read aligners, restricts the extension of transcripts based on paired-end read information. Although truncation of short reads effectively generates short stretches of good quality alignments, it is also likely, in the case of extension, that the gaps required by the correct alignments are too costly, especially at the ends of reads, and algorithms then preferentially clip the matching regions instead of accepting the cost of the gaps (INDELS). After assessment of existing methods for partial sequence extension we developed our own method. In this method, the targeted reads mapping approach was utilized, but instead of *de novo* assembly, the targeted subset of reads were aligned with the partial transcript. Paired-end information of raw reads can facilitate transcript extension and iterative searching and mapping of these paired-end reads allowed us to completely extend partial transcripts. A key feature of the approach developed in this study is that it exploits the progressive global alignment algorithm

implemented in Geneious. We tested the latest versions of several alternative short read aligners including, BWA [23], Bowtie [24], SSAHA2 [25], SMALT (<http://www.sanger.ac.uk/resources/software/smalt/>) and TopHat [26] with our data but the Geneious inbuilt aligner outperformed other short read aligners. Furthermore, Geneious does not truncate the individual reads, so the consensus sequence produced after alignment is usually longer than the reference and the untruncated flanking regions provide some level of extension over the target sequence. Aligning the paired-end raw reads completely extends partial transcripts in 4-5 iterations (the number of iterations depends on the length of the partial transcript and complexity of the complete gene of interest). We first applied this methodology on cytochrome P450 partial transcripts obtained from the *D. melanogaster* transcriptome. 12 out of 24 partial transcripts were extended with 8 fully extended to full length coding sequences (Table S1, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2018.2865309>). For the remaining partial transcripts, the method could not find any read which could be used as a seed to extend the transcripts in either direction. We next used the *O. bicornis* transcriptome to investigate the performance of our method on a non-model insect species. Using this dataset, we managed to extend 15 out of 16 partial transcripts encoding P450s and 8 out of 9 CCEs to their complete gene lengths (Table 2). In the case of two of the partial P450 transcripts, CYP-6 and CYP-28, low coverage meant raw reads were not obtained from database A in the numbers necessary to extend using the primary pipeline. We therefore adopted a second approach based on bait mapping whereby, based on the BLAST result of partial transcripts, we used the complete coding sequence of the orthologous gene of a related bee species, in this case *M. rotundata*, extracting paired-end raw reads specific to this gene then mapping back to the partial transcript. This proved to be an effective alternative extension method and extended partial transcripts from 1,274 and 254 bp to 1,644 and 2,049 bp respectively. NCBI-BLAST verification confirmed their identity as P450 transcripts and they were subsequently validated experimentally (See PCR validation of extended transcripts section). The obvious limitation of this approach is the requirement for a related gene sequence from another species. Furthermore, the likelihood of success is heavily dependent on the level of sequence similarity between the reference species and the related species. Nevertheless, our study clearly demonstrates that where a good ortholog is available for the gene of interest it can be used to extend the partial transcript to full length.

3.5 PCR Validation of Extended Transcripts

PCR was used to assess the success rate of the *in silico* extension workflow, with 19 potentially unique sequences from the *O. bicornis* extended transcript set (3 CCEs and 16 P450s) selected for validation (Table 2). For the CCEs, two out of the three were successfully amplified from the extended sequences, and named as CCE-8 and CCE-11 (Table 2). Direct sequencing and alignment of both PCR fragments revealed that CCE-8 and CCE-11 are in fact the same gene. Interestingly, the original sequences for these two fragments were 664 bp (CCE-8) and 630 bp (CCE-11) long and mapped to the 5 and 3 end of the amplified fragments. For the P450s positive amplification was achieved for 8 (from 16 total *in silico* extended) transcripts with full length amplification for CYP-3, -12, -28, -34, -51 and -52 and partial amplification for two additional transcripts, CYP-4 and CYP-6 (Table 2). Where Trinity detects different transcripts of the same gene (for example as a result of alternative splicing) it outputs these as different isoforms. In our case, two isoforms were reported for CYP-4 and CYP-28 (a and b) and three isoforms for CYP-51 (51a, b and c). For two of the

P450s, the partially amplified CYP-6 and completely amplified CYP-28, a different mapping strategy, called the bait mapping strategy, was used for extension (see Methods). Both transcripts were subsequently validated by PCR and sequencing, with the sequence returned for CYP28 matching isoform b. PCR and sequencing of CYP-51 resulted in the successful validation of version 51b, which is therefore predicted to be the predominantly expressed version of this P450. In the case of CYP-3, direct sequencing of the PCR product showed that the sequence matched the *in silico* extension, but a number of double peaks were observed throughout the sequence, suggesting that two closely related genes, named as CYP-3a and CYP-3b, were amplified. Cloning of the PCR fragment followed by sequencing of four of the positive clones and BLAST searching against the known database of existing P450s in *O. bicornis* confirmed that this was indeed the case, with both of the P450s being unique, albeit highly similar, genes. Direct sequencing of CYP-28b validated the *in silico* extension and also revealed that this P450 and CYP-51b confirmed that they are in fact the same P450 transcript, with 99.9 percent identity to another previously identified *O. bicornis* gene (CYP-29). Amplification of CYP-12 produced a single band of the expected size and direct sequencing of both the PCR product and sequencing of cloned products showed a correct match with the *in silico* extension. Blast results confirmed this to be a novel *O. bicornis* P450 sequence. A similar result was found for CYP-52, however, a 45 bp deletion was observed in some clones, likely the result of alternative splicing. Novel sequence information was also obtained for CYP-4 and CYP-6 although neither extension allowed amplification of the full length ORF. A short fragment of approximately 500 bp was amplified for CYP-4 matching the *in silico* extension and for CYP-6, sequencing revealed a fragment of a CYP ORF interrupted by an intron. No amplification of the expected bands was possible for CYP-2, both variants of CYP-37 and CYP-4. However, for the later a short section of approximately 500 bp was amplified from the middle of the ORF matching the CYP-4 extension. It is most likely that the extended sequences of these 3 genes were a result of fusion of multiple individual P450s and they do not exist in *O. bicornis* in the predicted form. In summary PCR validation of the *in silico* extended transcripts revealed that 8 out of the initial 19 fully extended (Table 2) cases were successful in producing full length ORFs (CCE-8 and 11, CYP-3, 12, 28, 51, 52). Additionally, CYP-6 and CYP-4 successfully produced novel, unique sequences matching the expected enzyme group, although not full length ORFs. These data confirm the value of *in silico* extension of partial transcripts identified in *de novo* assembled transcriptomes of non-model organisms. Although we would suggest that all extended sequences extracted by the pipelines are verified by PCR and sequencing prior to further functional characterisation. This is both easier and more cost-effective than using RACE, a technique which often fails to extend partial transcripts. All the experimentally validated extended transcripts were submitted in NCBI and can be accessible using the accessions: MH500604-MH500655

4 CONCLUSION

In this study, we have presented a novel approach to extend incompletely assembled transcripts generated by next-generation *de novo* assemblers from paired read RNA-Seq datasets. The method is based on an iterative read mapping strategy. By combining read database searches, and targeted reference based assemblies, this methodology is capable of significantly extending partial transcripts, in many cases up to full length. Moreover, this approach can also be used for improving the assembly and variant correction of candidate genes. When RNAseq data is available for a non-model organism we believe that this workflow is a viable alternative to experimental extension of partial transcripts.

ACKNOWLEDGMENTS

The authors would like to thank Bee-Toxicogenomics (Rothamsted Research and Bayer AG Crop Sciences Division) group members for their useful comments and suggestions. KSS and CB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n°646625). KB received funding from the Biotechnology and Biological Sciences Research Council (BBSRC, award number 15076182). The work at Rothamsted forms part of the Smart Crop Protection (SCP) strategic programme (BBS/OS/CP/000001) funded through the Biotechnology and Biological Sciences Research Council's Industrial Strategy Challenge Fund.

REFERENCES

- [1] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. M. Jones, and M. A. Marra, "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing," *BioTechnol.*, vol. 45, no. 1, pp. 81–94, 2008.
- [2] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnol.*, vol. 29, no. 7, pp. 644–652, 2011.
- [3] W. Xie, Q. Wu, S. Wang, X. Jiao, L. Guo, X. Zhou, and Y. Zhang, "Transcriptome analysis of host-associated differentiation in *Bemisia tabaci* (Hemiptera: Aleyrodidae)," *Frontiers Physiology*, vol. 5, no. Nov., pp. 1–7, 2014.
- [4] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, "Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels," *Bioinf.*, vol. 28, no. 8, pp. 1086–1092, 2012.
- [5] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol, "De novo assembly and analysis of RNA-seq data," *Nature Methods*, vol. 7, no. 11, pp. 909–912, 2010.
- [6] Y. Peng, H. C. M. Leung, S. M. Yiu, X.-G. Zhu, M.-Z. Lv, and F. Y. L. Chin, "IDBA-Tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels," *Bioinf.*, vol. 29, no. 13, pp. i326–i334, Jul. 2013.
- [7] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, and A. Regev, "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis," *Nature Protocols*, vol. 8, no. 8, pp. 1494–1512, Jul. 2013.
- [8] P. E. C. Compeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nature Biotechnol.*, vol. 29, no. 11, pp. 987–991, 2011.
- [9] P. Jain, N. M. Krishnan, and B. Panda, "Augmenting transcriptome assembly by combining de novo and genome-guided tools," *PeerJ*, vol. 1, 2013, Art. no. e133.
- [10] Y. Yang and S. A. Smith, "Optimizing de novo assembly of short-read RNA-seq data for phylogenomics," *BMC Genomics*, vol. 14, no. 1, 2013, Art. no. 328.
- [11] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Rev. Genetics*, vol. 12, no. 10, pp. 671–682, 2011.
- [12] M. Liu, Z. N. Adelman, K. M. Myles, and L. Zhang, "TransPS: A transcriptome post scaffolding method for assembling high quality contigs," *Comput. Biol. J.*, vol. 2014, pp. 1–4, 2014.
- [13] I. J. Tsai, T. D. Otto, and M. Berriman, "Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps," *Genome Biol.*, vol. 11, no. 4, 2010, Art. no. R41.
- [14] K. P. Johnson, K. K. O. Walden, and H. M. Robertson, "Next-generation phylogenomics using a Target Restricted Assembly Method," *Mol. Phylogenetics Evolution*, vol. 66, no. 1, pp. 417–422, 2013.
- [15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, 2013, pp. 403–410.
- [16] J. M. Allen, D. I. Huang, Q. C. Cronk, and K. P. Johnson, "aTRAM - automated target restricted assembly method: A fast method for assembling loci across divergent taxa from next-generation sequencing data," *BMC Bioinf.*, vol. 16, no. 1, 2015, Art. no. 98.
- [17] J. G. Ruby, P. Bellare, and J. L. Derisi, "PRICE: Software for the targeted assembly of components of (Meta) genomic sequence data," *G3: Genes Genomes Genetics*, vol. 3, no. 5, pp. 865–80, 2013.
- [18] R. M. Leggett, R. H. Ramirez-Gonzalez, B. J. Clavijo, D. Waite, and R. P. Davey, "Sequencing quality assessment tools to enable data-driven informatics for high throughput Genomics," *Frontiers Genetics*, vol. 4, no. Dec., pp. 1–5, 2013.

- [19] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinf.*, vol. 30, no. 15, pp. 2114–2120, 2014.
- [20] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: A universal tool for annotation, visualization and analysis in functional Genomics research," *Bioinf.*, vol. 21, no. 18, pp. 3674–3676, 2005.
- [21] T. J. Wheeler and S. R. Eddy, "Nhmmer: DNA homology search with profile HMMs," *Bioinf.*, vol. 29, no. 19, pp. 2487–2489, 2013.
- [22] R. S. Harris, "Improved pairwise alignment of genomic DNA," PhD thesis, Pennsylvania State University, Dec. 2007.
- [23] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *arXiv:1303.3997v1*, 2013.
- [24] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, p. 25–24, 2011.
- [25] Z. Ning, A. J. Cox, J. C. Mullikin, Z. Ning, A. J. Cox, and J. C. Mullikin, "SSAHA: A fast search method for large DNA databases," *Genome Res.*, vol. 11, pp. 1725–1729, 2001.
- [26] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: Discovering splice junctions with RNA-Seq," *Bioinf.*, vol. 25, no. 9, pp. 1105–1111, 2009.

▷ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.