BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 8 2013, pages 1004–1010
doi:10.1093/bioinformatics/btt092

Sequence analysis

Advance Access publication March 1, 2013

EBARDenovo: highly accurate *de novo* assembly of RNA-Seq with efficient chimera-detection

Hsueh-Ting Chu[1,2], William W. L. Hsiao[3,4], Jen-Chih Chen[5], Tze-Jung Yeh[5], Mong-Hsun Tsai[5,6], Han Lin[7], Yen-Wenn Liu[8], Sheng-An Lee[9], Chaur-Chin Chen[10], Theresa T. H. Tsao[6,7,*] and Cheng-Yan Kao[6,7,*]

[1]Department of Biomedical informatics, [2]Department of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan, [3]BCCDC Public Health Microbiology & Reference Laboratory, Vancouver, British Columbia V5Z 4R4, Canada, [4]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V5Z 4R4, Canada, [5]Institute of Biotechnology, [6]Graduate Institute of Biomedical Electronics and Bioinformatics, [7]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan, [8]National Research Institute of Chinese Medicine, No. 155-1, Sec. 2, Li Nung Street Peitou, Taipei 11221, Taiwan, [9]Department of Information Management, Kainan University, Taoyuan, 33857, Taiwan and [10]Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan

Associate Editor: Michael Brudno

ABSTRACT

Motivation: High-accuracy *de novo* assembly of the short sequencing reads from RNA-Seq technology is very challenging. We introduce a *de novo* assembly algorithm, EBARDenovo, which stands for Extension, Bridging And Repeat-sensing Denovo. This algorithm uses an efficient chimera-detection function to abrogate the effect of aberrant chimeric reads in RNA-Seq data.

Results: EBARDenovo resolves the complications of RNA-Seq assembly arising from sequencing errors, repetitive sequences and aberrant chimeric amplicons. In a series of assembly experiments, our algorithm is the most accurate among the examined programs, including de Bruijn graph assemblers, Trinity and Oases.

Availability and implementation: EBARDenovo is available at http://ebardenovo.sourceforge.net/. This software package (with patent pending) is free of charge for academic use only.

Contact: cykao@csie.ntu.edu.tw, htchu@asia.edu.tw or postergrey@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 9, 2012; revised on January 18, 2013; accepted on February 18, 2013

1 INTRODUCTION

RNA-Seq technology has revolutionized transcriptomic studies in recent years. In comparison with microarrays, this high-throughput sequencing technology allows the detection of novel transcripts. It also has a wider dynamic range in detecting transcript abundance than tiling arrays (Agarwal *et al.*, 2010; Kampa *et al.*, 2004), serial analysis of gene expression (Burke *et al.*, 1999; Velculescu *et al.*, 1995) and cap analysis of gene expression (Shiraki *et al.*, 2003). RNA-Seq is especially important for studying organisms without reference genomes. With reference genomes, accurate reads can be identified and mapped to the reference for accurate and sensitive quantification of even lowly expressed transcripts. However, for organisms without reference genomes, the short reads from RNA-Seq have to be assembled into contigs representing the transcripts before the expression level can be determined from the read coverage. In this case, sequencing errors may cause mis-assembly, create artificial contigs and consequently produce inaccurate interpretation of transcript abundance (Garcia *et al.*, 2012).

De novo assembly of short reads faces computational challenges from classical sequencing problems such as sequence repeats, homologous genes and artificial chimeric reads (Fig. 1) (Kircher *et al.*, 2011). Most *de novo* RNA-Seq assemblers, including Trans-Abyss (Robertson *et al.*, 2010), Oases (Schulz *et al.*, 2012) and Trinity (Grabherr *et al.*, 2011), are de Bruijn graph–based methods. In de Bruijn graphs, reads are either represented as *k*-mer nodes or as *k*-mer edges, and Eulerian paths between end nodes are examined for assembling sequences (Pevzner *et al.*, 2001). Both sequence repeats and artificial chimeric reads will cause erroneous *k*-mer nodes. Most of de Bruijn genome assemblers, such as Velvet (Zerbino and Birney, 2008) and ALLPATHS (Butler *et al.*, 2008), filter out low-frequency *k*-mer nodes to improve accuracy, as low-frequency *k*-mer nodes are more likely to be sequencing artifacts. However, *de novo* assembly of RNA-Seq data is more difficult because the abundance of RNA transcripts varies significantly (Grabherr *et al.*, 2011). Although de Bruijn–based algorithms have been successfully used to construct many transcriptomes using RNA-Seq data, it has always been difficult to keep the assembly both accurate and sensitive. Low-frequency *k*-mer nodes are important for discovering low-abundance transcripts. For sensitive detection of novel transcripts, RNA-Seq assemblers usually do not eliminate low-frequency *k*-mer nodes. Unfortunately, retaining low-frequency *k*-mer nodes would reduce assembly accuracy.

In this study, we introduce an efficient algorithm for RNA-Seq assembly, with special emphasize on detecting chimeric reads and

*To whom correspondence should be addressed.

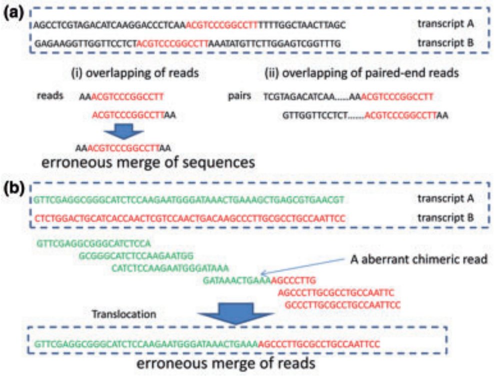
(a)
AGCCTCGTAGACATCAAGGACCCTCAAACGTCCCGGCCTTTTTTGGCTAACTTAGC
transcript A
GAGAAGGTTGGTTCCTCTACGTCCCGGCCTTAAATATGTTCTTGGAGTCGGTTTG
transcript B
(i) overlapping of reads
(ii) overlapping of paired-end reads
reads
AAACGTCCCGGCCTT
ACGTCCCGGCCTTAA
pairs
TCGTAGACATCAA.....AAACGTCCCGGCCTT
GTTGGTTCCTCT.......ACGTCCCGGCCTTAA
AAACGTCCCGGCCTTAA
erroneous merge of sequences
(b)
GTTCGAGGCGGGCATCTCCAAGAATGGGATAAACTGAAAGCTGAGCGTGAACGT
transcript A
CTCTGGACTGCATCACCAACTCGTCCAACTGACAAGCCCTTGCGCCTGCCAATTCC
transcript B
GTTCGAGGCGGGCATCTCCA
GCGGGCATCTCCAAGAATGG
CATCTCCAAGAATGGGATAAA
A aberrant chimeric read
GATAAACTGAAAAGCCCTTG
AGCCCTTGCGCCTGCCAATTC
GCCCTTGCGCCTGCCAATTCC
Translocation
GTTCGAGGCGGGCATCTCCAAGAATGGGATAAACTGAAAAGCCCTTGCGCCTGCCAATTCC
erroneous merge of reads

Fig. 1. The two major sources of errors in RNA-Seq *de novo* assembly. (**a**) The presence of common sub-sequences (repetitive sequences, homologous sequences) among transcripts causes reads to merge by mistake and produce mis-assembly (i). By using paired-end reads, the opportunity for mis-assembly is reduced. (**b**) Aberrant chimeric amplicons composed of disjointed regions of one or more transcripts can cause erroneous assembly of reads into artificial contigs

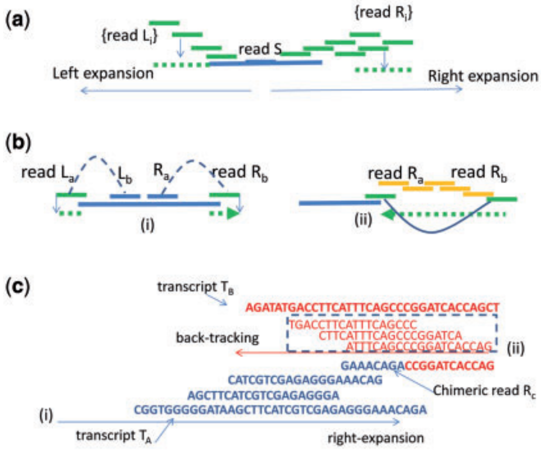

Fig. 2. The key principles in EBARDenovo for achieving high accuracy. (**a**) The outline of EBARDenovo algorithm. Candidate reads are iteratively added at both ends with three operations: extension, bridging and repeat sensing. (**b**) The two conditions for the alignment of an adjacent pair of reads (R_a and R_b) or (L_a and L_b). (i) the paired mate (L_b or R_a) is inside the current contig. (ii) The paired mate (R_b) is found in neighboring (usually overlapping) contigs. (**c**) Dynamic chimera detection. The aligned reads from the backtracking process are compared with the consensus sequence (T_A) to detect potential chimeric junctions

assembly errors. We showed that our new *de novo* assembly algorithm, designated EBARDenovo, has enhanced accuracy, while it still maintains high sensitivity in determining transcript abundance.

2 METHODS

2.1 Key principles of EBARDenovo

The new algorithm, EBARDenovo, which stands for Extension, Bridging And Repeat-sensing Denovo, applies a bi-directional expansion method using paired-end RNA-Seq data to guide the transcriptome assembly (Fig. 2a).

In this algorithm, the construction of a contig begins with a seed read. Searches are done to find proper neighboring reads at both ends of the seed. Two reads that come from the same paired-end sequencing reads are defined as read mates of each other. To avoid potentially ambiguous assembly created by repeat sequences, a proper extension must satisfy one of two conditions (Fig. 2b): i) a read mate can be found inside the contig that is currently being constructed; or ii) a read mate can be found in a neighboring (usually overlapping) contig. To ensure both conditions are satisfied for a contig extension, a chimera-detection test is performed on the paired candidates (Fig. 2c). Assuming there is a chimeric read R_c composed of two fragments from transcripts T_A and T_B, the correct assembly should always have higher read coverage than the contig containing the chimeric reads. In this case, the backtracking from the other end of chimeric read R_c will find a set of correct reads from transcript T_B, and the alignment of these reads is compared with the nucleotide bases of the current contig to determine whether the program can correctly expand the contig or cause a translocation of the two segments.

2.1 EBARDenovo: the proposed algorithm and key operations

EBARDenovo algorithm consists of three key operations: extension, bridging and repeat sensing. These operations are applied to both the right-end and the left-end expansions such that the selected sub-optimal reads are added in these operations.

Before starting the three assembly operations, EBARDenovo defines the first few base pairs (the default is 15 bp) of each read as its indexing keys. The reads are also sorted according to their abundance, and the most abundant read is selected as the first seed for the extension, followed by the second most abundant reads and so on.

To start the extension operation of assembly, the 2nd to the nth bp (the default is from 2nd to 16th bp corresponding to a 15-mer) will be used as the query key to search against the indexing keys of all reads. If no matching key is found, a query key will be generated from the 3rd to the 17th bp and so on. If a matching key is found, the reads containing this key will be aligned to the seed read at the error detection area (Supplementary Fig. S4). Error detection area is the overlapping region between the seed and the matching read, but excluding the key index region. At this step, all perfectly matched reads will be merged to extend the seed to a longer contig. If no perfect candidate is available, the reads with fewest mismatches will be used. However, >10% mismatch in the error detection area is not tolerated.

In the bridging operation, EBARDenovo uses paired-end information to constrain the search space. By using 'foreseeable reads' to confirm correct assembly, reads with as small as 15 bp overlaps can be constructed into reliable contigs. A read is 'foreseeable' if one of its paired-end sequenced mates is already in the contig (Fig. 3a). If a foreseeable read is already aligned in the contig, the extension of an assembled contig is continued (Fig. 3a). The extension at this stage is only one-way, either left to right or right to left. This is because bi-directional extension of a chimeric seed read would probably generate a chimeric contig, and there is insufficient information to detect chimeric reads at this stage. In the case that the foreseeable read is not

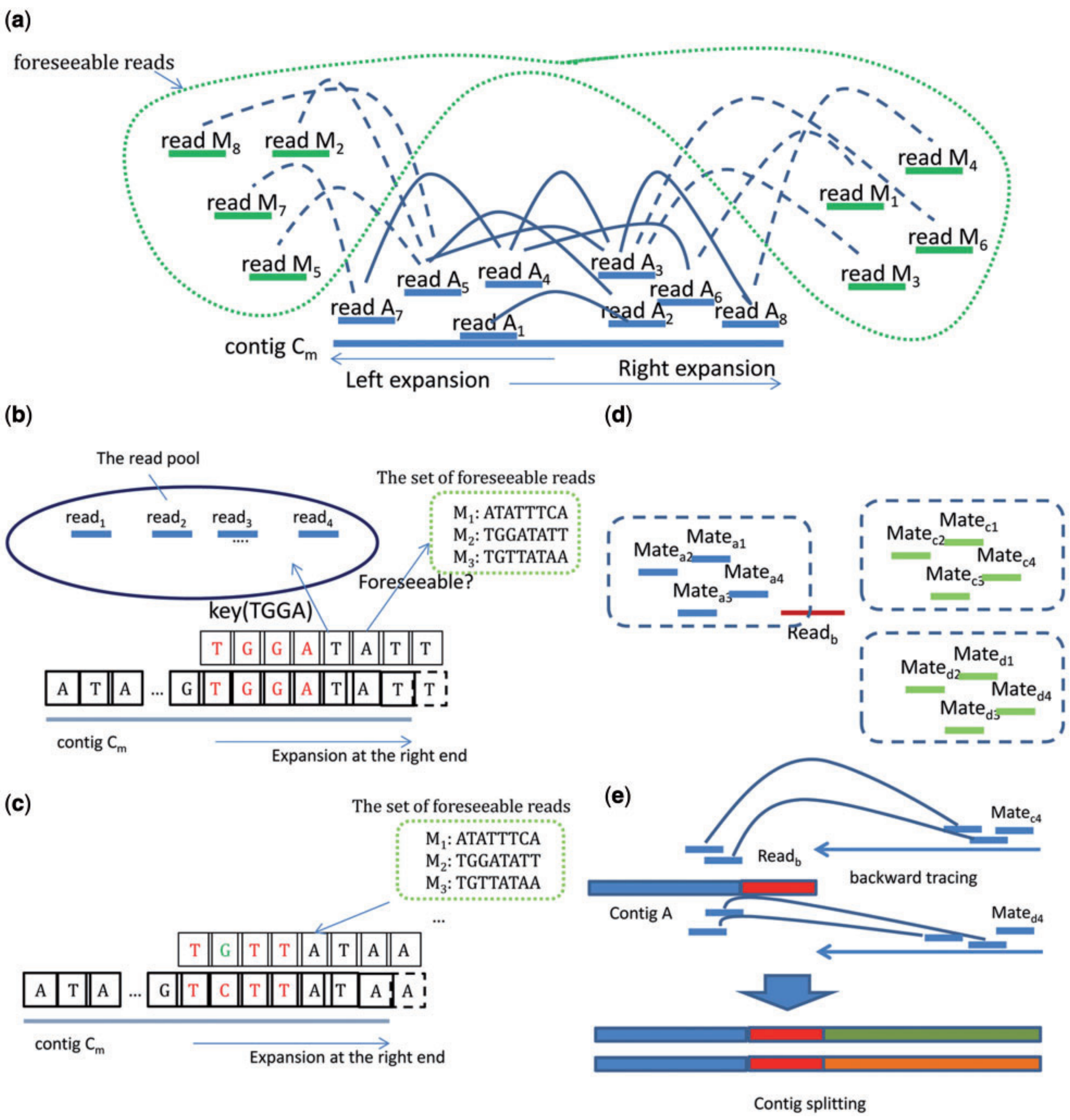

Fig. 3. The key steps of EBARDenovo. (**a**) 'Read foreseeing' tracks the paired mates of the aligned reads, which have not been aligned in the contig. These unaligned reads can be merged into the contig in upcoming operations of extension, bridging or repeat sensing. (**b**) In the extension operation, candidate reads are selected from the read pool according to a short indexing key. (**c**) In the bridging operation, foreseeable reads may match the expanding end of a contig. (**d**) In the repeat-sensing operation, if there are possible expansions for a contig, it is backtracked along each expanding possibility. The contig will be split if there are more than two possible expansions. (**e**) In the extension operation, when a non-foreseeable read is met, EBARDenovo attempts to align reads from the mate of the non-foreseeable read in the opposite direction of the extension

within the temporary contig, the foreseeable read will be used as seed read and extended backward to see whether the extension can reach the temporary contig (Fig. 2b). The insert sizes of paired-end RNA-Seq can be very different even within the same dataset. To avoid the assembly quality being affected by the varying insert sizes, the insert sizes are not evaluated in EBARDenovo.

The repeat-sensing operation also uses foreseeable reads to identify repetitive sequences naturally occurring in transcripts and resolve the issue of multiple expansions (Fig. 3b). A contig is split when multiple expansions occur, and the extension operation is then executed if there are no other foreseeable reads to indicate possible bridging (Fig. 3c). If the extension operation fails owing to sequencing errors, the algorithm will go back to the bridging operation (Fig. 3d). The extension and bridging operations are alternated until the contigs are no longer extendable. At the end, all the aligned reads are paired. The entire procedure of EBARDenovo is illustrated in Figure 4.

2.3 The verification of assembly accuracy using GMAP

The accuracy of assemblers was verified using GMAP (Wu and Watanabe, 2005) program. GMAP aligns the contigs with a reference genome to check the correctness of contigs. The *Arabidopsis thaliana* genome was downloaded from the TAIR database (http://www.arabidopsis.org/). The *Mus musculus* (mouse) and *Homo sapiens* (human) genomes were downloaded from the UCSC Genome database (http://hgdownload.cse.ucsc.edu). GMAP classifies mapping results into four classes: unique mapping (U), multiple mapping (M), translocated mapping (T) and no mapping (N). The no-mapping results were seen as contamination of the RNA samples. The error rates of assembly results were estimated by the ratio of translocated mappings T/(U + M + T). The RPKM values for genes were computed with the software CLC Genomics Workbench 5.1 by mapping the sequencing reads to the transcriptome databases. RPKM is defined as (Mortazavi *et al.*, 2008)

$$\frac{\text{the total number of reads}}{\text{the number of mapped reads} \times \text{gene length} \times 10^{9}}$$

2.4 Output of EBARDenovo for transcriptome analysis

To easily examine the assembly results, detect potential single nucleotide polymorphisms (SNPs) and calculate expression levels of assembled contigs, EBARDenovo offers four types of output information for each reconstructed sequence, including (i) the sequences, (ii) the aligned reads in each contig, (iii) the alignment of used pairs and (iv) the plot of sequence coverage (Supplementary Fig. S1). EBARDenovo also outputs the general information of contigs, the SNPs and the locations of small overlaps among reads in the same contig. The users can visualize the variants of the transcripts. In the past, users relied on supplemental mapping tools, e.g. Bowtie (Langmead *et al.*, 2009), to map reads to a contig and to estimate the abundance, whereas EBARDenovo alone can produce sufficient information for advanced analyses such as the identification of RNA editing sites (Brennicke *et al.*, 1999) and gene fusion candidates (Vega and Medeiros, 2003). We expect that EBARDenovo can enhance the experience of RNA-Seq analysis.

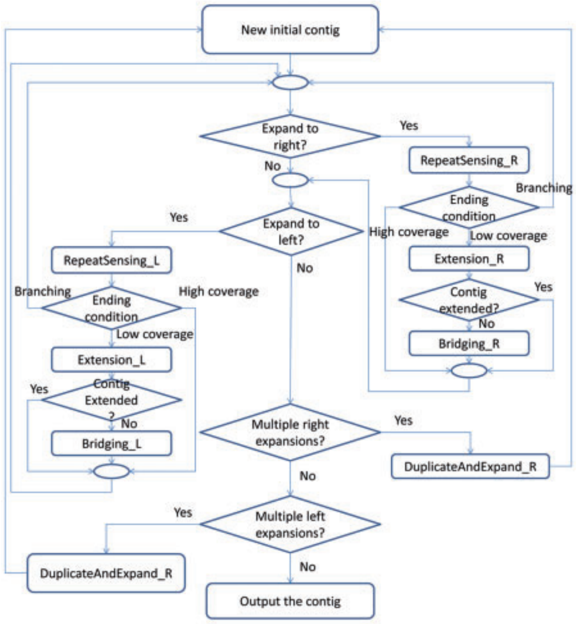

Fig. 4. The flowchart of the EBARDenovo algorithm

2.5 Software implementation

EBARDenovo is implemented in C# with the .NET Framework. The executable program can be run on 64 bit operating systems such as Windows, Linux and Mac OS. On Linux or Mac OS, Mono framework (http://www.mono-project.com/) needs to be installed to run the program. EBARDenovo program and the user manual can be downloaded from http://ebardenovo.sourceforge.net/. It is also available as Supplementary Software.

Table 1. List of test RNA-Seq data sets

Organism	Experiment	Assembly runs	Read length	Spots	Bases
A.Thaliana	Test data	ntubiotec[a]	90 bp	6.7 M	1.2 G
	SRX112186	SRR391052[b]	76 bp	27 M	4.1 G
Mouse	SRX118647	SRR404355	50 bp	13 M	1.3 G
	SRX064476	SRR212430	72 bp	21 M	3.0 G
Human	SRX135562	SRR453391	76 bp	16 M	2.4 G
	SRX087128	SRR324684	100 bp	18 M	3.6 G

[a] The test data of *A.thaliana* is provided by the Institute of Biotechnology, NTU.
[b] The other five datasets were available from the ENA (http://www.ebi.ac.uk/ena/).

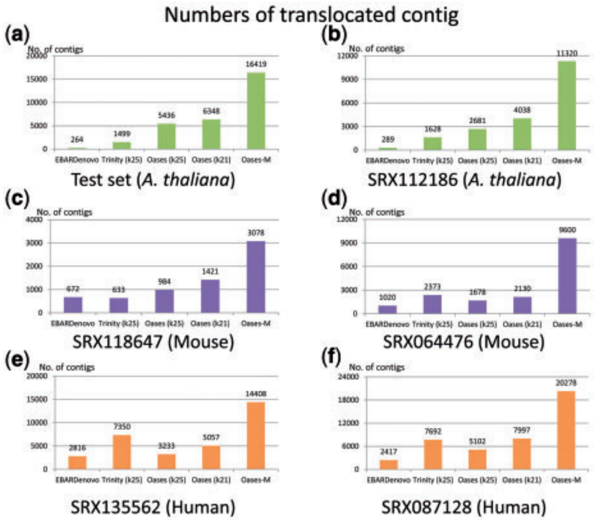

Fig. 5. Comparisons of GMAP tests for the assemblies of different RNA-Seq datasets by EBARDenovo, Trinity and Oases

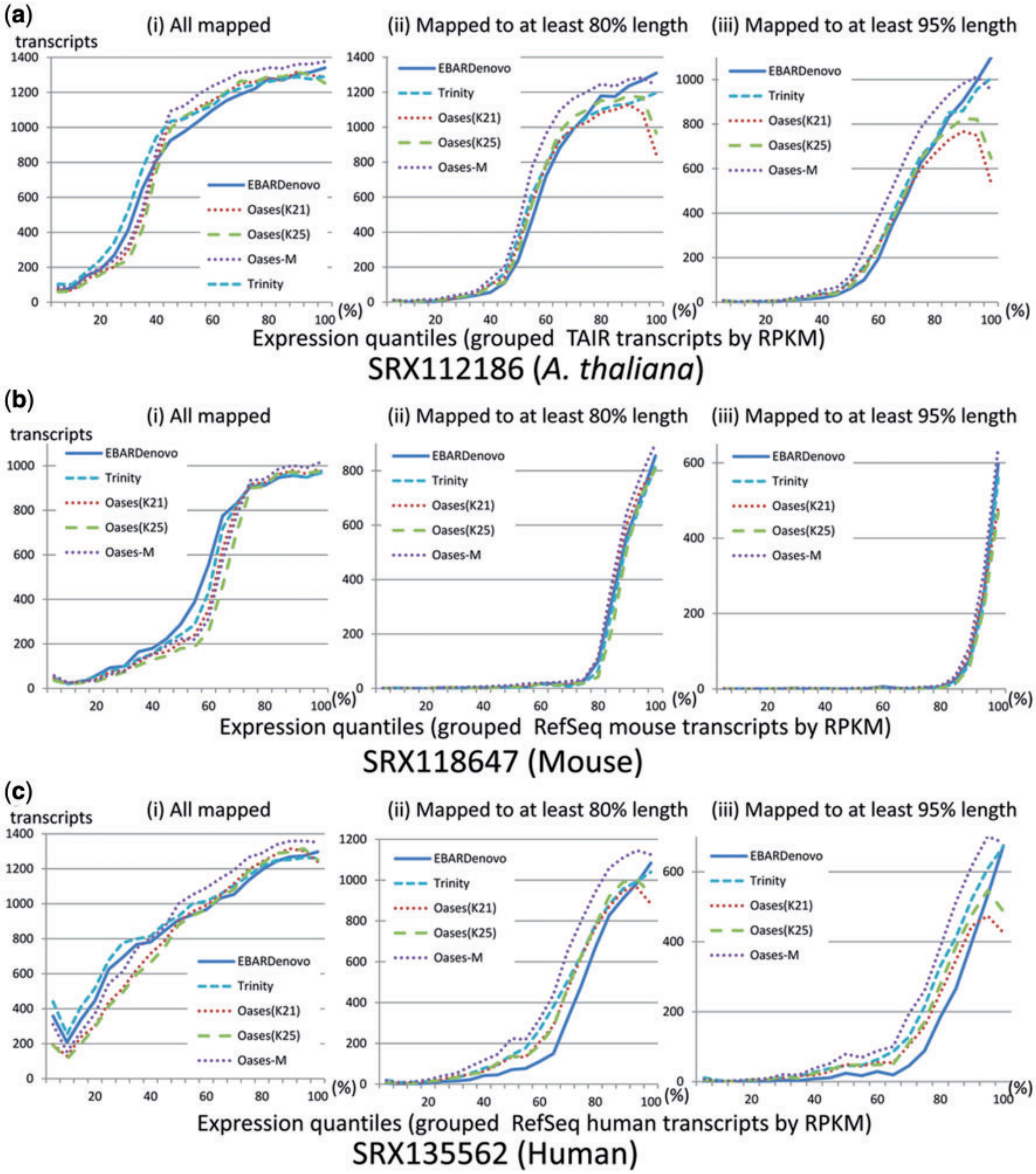

Fig. 6. Efficiency of different assemblers, including EBARDenovo, Trinity and Oases

3 RESULTS AND DISCUSSION

3.1 RNA-Seq datasets and tested *de novo* assembly algorithms

We tested our algorithm against two de Bruijn graph–based assemblers, Trinity and Oases, using six RNA-Seq datasets (listed in Table 1) from three model organisms: *A.thaliana*, *M.musculus* (mouse) and *H.sapiens* (human). We used four sizes of *k*-mers (21, 25, 29 and 33 bp) to run the Oases assembly. The assemblies with *k*-mers of 29 and 33 bp were used to produce the merging results known as Oases-M assemblies. Trinity only provides a fixed size of *k*-mer (25 bp). By default, EBARDenovo directly searches for candidate reads by an indexing key (15 bp). The detail assembly results are listed in Supplementary Table S1.

The assembly results were verified with the transcriptome and genome databases TAIR (Rhee *et al.*, 2003) (for *A.thaliana*) and NCBI Reference Sequences (RefSeq) (for mouse and human) using the mapping tool on CLC Genomics Workbench and GMAP (Wu and Watanabe, 2005). We showed the comparison of mapped transcripts by different algorithms in Supplementary Figure S2. Moreover, we checked the intersection of mapped transcripts between different assemblies in Supplementary

Figure S3. It is obvious that many transcripts can be successfully reconstructed by only one assembler in each dataset.

3.2 EBARDenovo produced most accurate assemblies for RNA-Seq datasets

To calculate accuracy of different assembly results, we mapped assembled contigs to their reference genomes using GMAP programs to detect translocation caused by potential mis-assembly. If a contig is mapped to different gene models in a chromosome or even different chromosomes, it is counted as an artificial translocation event. If a correct assembly of a natural gene fusion occurs, it will also be counted as a translocation event in the GMAP test. However, we assumed such natural fusion genes to be rare. We also recorded events of a whole contig mapped to multiple locations of the reference genome (multiple mapping) but did not count them as translocation errors because such events were detected even when reference transcripts were used in our GMAP test. It is obvious that EBARDenovo generated the least number of translocations among three programs in all the six experiments (Fig. 5). In the *A.thaliana* datasets, the number of erroneous contigs by EBARDenovo is less than one-fifth of those by Trinity. In the mouse and human datasets, Trinity produces two to three times more erroneous contigs than EBARDenovo. The assembly results by Oases with a smaller *k*-mer (21 bp) produced more translocation than with a bigger *k*-mer (25 bp). The merging assembly by Oases will accumulate errors from the assemblies with different *k*-mers. The detail results of GMAP test are listed in Supplementary Table S2.

The high accuracy of the proposed algorithm implies more stringent trimming of inaccurate contigs, which results in the reduction of contig length. For this reason, we checked the correctness of contigs at comparable lengths. We divided the contigs by their lengths into three groups (600∼799, 800∼999 and 1000∼1199 bp). We evaluated the accuracy of each contig group for the three assemblers using the GMAP. Supplementary Table S3 shows the results. In all the contig groups for all six datasets, EBARDenovo always produces the most accurate results.

3.3 The comparison of assembly coverage at different expression levels by different assemblers

To evaluate the performance of different assemblers on the reconstruction of full-length transcripts, EBARDenovo and the other two assemblers were compared using three RNA-Seq datasets: SRX112186 (*A.thaliana*), SRX118647 (mouse) and SRX135562 (human). The expression quantiles were calculated using the RNA-Seq Analysis tool in the CLC package. The completion of transcript reconstruction was estimated by comparing contigs with reference cDNA sequences from TAIR (Rhee *et al.*, 2003) and RefSeq (Pruitt *et al.*, 2012). Transcripts with higher coverage generally can be reconstructed much closer to full length, by merging contigs generated from different *k*-mers, Oases-M, therefore, produced a more completed reconstruction than other tools (Fig. 6), but paid a high price in accuracy (Fig. 5). Artificial translocations generated by Oases-M were much greater than those generated by the programs using only a fixed *k*-mer. In addition, performance of Oases clearly dropped for highly expressed transcripts. In contrast, EBARDenovo gave the best performance on transcripts with high expression.

Table 2. Performance comparison between RNA-Seq assemblers[a]

Algorithm	Maximal memory[b]	Time
EBARDenvo (key index = 15)[c]	3.3 G	39 min
Trinity (*k*-mer = 25)	5.1 G	264 min
Oases (*k*-mer = 21)	4.4 G	18 min
Oases (*k*-mer = 25)	3.9 G	16 min

[a]The results were from the assemblies of the *A.thaliana* test data (in Table 1).

[b]The maximal memory of running Trinity was the memory usage of the GraphFromFasta process at the stage of Chrysalis. The default Jellyfish process of the Inchworm stage usually consumes more memory (10∼100 G) for good performance.

[c]All the programs were run on VMware Virtual machines. The host machine is an i7-3820 PC with 32G RAM. Trinity and Oases were run on Ubuntu 10.04, and EBARDenvo was run on Windows 7. Both virtual machines are equipped with same memory space (24 G) and same CPU cores (2:2).

4 CONCLUSIONS

EBARDenovo is designed for accurate assembly of paired-end RNA-Seq data. It consumes less memory space than Trinity and Oases, whereas its speed is between that of Trinity and Oases (Table 2). More importantly, in our experiments, the contigs produced by EBARDenovo have much higher accuracy than those by both Trinity and Oases. The detection of aberrant chimeric reads is the key factor for high accuracy of EBARDenovo.

Accurately generating full-length transcripts and low-expression transcripts are important goals for the assembly of RNA-Seq data. There are chimeric transcripts created by natural gene fusion or sequence recombination. Methods such as EricScript have been developed to identify these naturally occurring chimeric sequences (Benelli *et al.*, 2012). Nonetheless, there are also aberrant chimeric reads, which are artifacts from sequencing processes. Chimeric read detection methods, such as ChimeraScan (Iyer *et al.*, 2011) and UCHIME (Edgar *et al.*, 2011), require reference genomes and are therefore not applicable in *de novo* assembly. The de Bruijn graph assemblers, such as Velvet (Zerbino and Birney, 2008), do not use active chimera-detection functions, but instead, remove low-coverage reads (occurrence ≤ 2) as potential aberrant chimeras. This rule makes the Velvet algorithm less suitable for RNA-Seq data, as the abundance of transcripts vary greatly. The other de Bruijn graph assemblers, such Oases and Trinity, relax the rule of eliminating low-coverage contigs to allow the assembly of rare reads (occurrence = 1) (Jiang *et al.*, 2011). However, the preservation of potential aberrant chimeric reads may produce artificial translocations of transcripts in the final assembly.

In the past years, de Bruijn graph methods had showed its advantage for full-length assemblies, but chimeric *k*-mer nodes limited the capability of modern de Bruijn graph algorithms to generate accurate low-coverage assemblies. Moreover, using shorter *k*-mer in de Bruijn graph assembler algorithms produces more mis-assemblies (Schulz *et al.*, 2012), whereas EBARDenovo can maintain high accuracy of assembly even with a short key length (Supplementary Fig. S5). In conclusion, the EBARDenovo algorithm is a promising new approach to identify transcripts from RNA-Seq analysis.

ACKNOWLEDGEMENTS

H.T.C. and C.Y.K. devised the method and wrote the software; J.C.C., T.J.Y. and M.H.T. designed the experiments; H.L., M.H.T. and S.A.L. tested the assembly algorithms; H.T.C., W.L.H., T.H.T., J.C.C., T.J.Y., Y.W.L. and C.C.C. jointly wrote the manuscript; C.Y.K. planned the project.

Conflict of Interest: none declared.

REFERENCES

Agarwal,A. *et al*. (2010) Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, **11**, 383.

Benelli,M. *et al*. (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics*, **28**, 3232–3239.

Brennicke,A. *et al*. (1999) RNA editing. *FEMS Microbiol. Rev*., **23**, 297–316.

Burke,J. *et al*. (1999) d2_cluster: a validated method for clustering EST and full-length cDNAsequences. *Genome Res*., **9**, 1135–1142.

Butler,J. *et al*. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res*., **18**, 810–820.

Edgar,R.C. *et al*. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 2194–2200.

Garcia,T.I. *et al*. (2012) RNA-Seq reveals complex genetic response to deepwater horizon oil release in Fundulus grandis. *BMC Genomics*, **13**, 474.

Grabherr,M.G. *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*., **29**, 644–652.

Iyer,M.K. *et al*. (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 2093–2094.

Jiang,L. *et al*. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res*., **21**, 1543–1551.

Kampa,D. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*., **14**, 331–342.

Kircher,M. *et al*. (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, **12**, 382–395.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*., **10**, R25.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Pevzner,P. *et al*. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.

Pruitt,K.D. *et al*. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*., **40**, D130–D135.

Rhee,S.Y. *et al*. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.

Robertson,G. *et al*. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.

Schulz,M.H. *et al*. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Shiraki,T. *et al*. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.

Vega,F. and Medeiros,L.J. (2003) Chromosomal translocations involved in non-Hodgkin lymphomas. *Arch. Pathol. Lab. Med*., **127**, 1148–1160.

Velculescu,V.E. *et al*. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Zerbino,D. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*., **18**, 821–829.