JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 7, 2013

Pp. 540–550
DOI: 10.1089/cmb.2013.0042

Late Submission for RECOMB/DREAM 2012 Special Issue

IDBA-MT: *De Novo* Assembler for Metatranscriptomic Data Generated from Next-Generation Sequencing Technology

HENRY C.M. LEUNG[1], SIU-MING YIU[1], JOHN PARKINSON[2], and FRANCIS Y.L. CHIN[1]

ABSTRACT

High-throughput next-generation sequencing technology provides a great opportunity for analyzing metatranscriptomic data. However, the reads produced by these technologies are short and an assembling step is required to combine the short reads into longer contigs. As there are many repeat patterns in mRNAs from different genomes and the abundance ratio of mRNAs in a sample varies a lot, existing assemblers for genomic data, transcriptomic data, and metagenomic data do not work on metatranscriptomic data and produce chimeric contigs, that is, incorrect contigs formed by merging multiple mRNA sequences. To our best knowledge, there is no assembler designed for metatranscriptomic data. In this article, we introduce an assembler called IDBA-MT, which is designed for assembling reads from metatranscriptomic data. IDBA-MT produces much fewer chimeric contigs (reduce by 50% or more) when compared with existing assemblers such as Oases, IDBA-UD, and Trinity.

Key words: algorithms, alignment, computational molecular biology, dynamic programming, genomic rearrangements, metagenomics, next generation sequencing.

1. INTRODUCTION

STUDYING THE INTERACTION of different microbes in an environmental sample is important for understanding the microbial world and its effect on the host. For example, the diversity of microbes in human gut was found to be related to common diseases such as inflammatory bowel disease (Poretsky et al., 2005) and gastrointestinal disturbance (Khachatryan et al., 2008). Studying each microbe in a sample separately cannot provide much insight on how microbes interact with each other and, more important, most microbes cannot be cultured and separated in laboratories. Thus, directly studying the collective genomes sampled from a natural microbial community, called *metagenomics*, has become a standard way of studying the interaction among different microbes in a community.

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, People's Republic of China.
[2]Biochemistry & Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada.

This article was presented at the 5th Annual RECOMB Conference on Regulatory and Systems Genomics, with DREAM Challenges, held on November 12–15, 2012, in Redwood City, CA. It was the conference organizers' intention to include this as part of a special RECOMB/DREAM supplement published in the May 2013 issue of the *Journal of Computational Biology*, but it did not appear due to an oversight on the organizers' part. We thank the Journal for publishing it at this time.

Although metagenomic analysis provides some insights into what kinds of microbes exist in a sample and their relative abundance ratios, it is difficult to understand how the microbes work together, especially how they respond to different environmental changes. This question can be answered properly by sequencing the mRNAs existing in the sample, as they are highly related with the proteins produced by the microbes in the sample.

Traditional metatranscriptomic studies have been based on microarrays or cDNA (Poretsky et al., 2005; Tartar et al., 2009; Booijink et al., 2010) clone libraries. Microarrays (Parro et al., 2007) can detect the existence of some selected mRNA sequences in a sample, and their relative abundance can be roughly estimated by the signal of microarray. However, since the microarrays are designed for known mRNA sequences, those RNAs in the sample without reference sequences cannot be detected. Moreover, some signals of microarray may be noisy and the abundance of mRNA cannot be accurately estimated. cDNA clone libraries randomly select some mRNAs and convert them into cDNAs. Each cDNA will then be implanted in the genome of some host, for example, bacteria. By growing the host, multiple copies of the implanted cDNA can be obtained for analysis. However, constructing cDNA clone libraries is labor intensive, and the relative abundance of mRNAs will be biased, for example, if the protein encoded by the mRNA is toxic with respect to the host.

With the help of high-throughput next-generation sequencing (NGS) technology (Leininger et al., 2006; Bosch and Grody, 2008; Morozova and Marra, 2008; Fullwood et al., 2009; Pettersson et al., 2009), biologists have overcome the limitation of the above methods by sequencing the mRNA sequences directly and collectively in a sample. Several metatranscriptomic studies using pyrosequencing technology (with read length about 400 bp) have been performed (Frias-Lopez et al., 2008; Gilbert et al., 2008; Urich et al., 2008; Poretsky et al., 2010) and have achieved promising results for soil samples (Urich et al., 2008) and marine samples (Frias-Lopez et al., 2008; Gilbert et al., 2008). However, as the throughput of reads from pyrosequencing technology is low when compared with other NGS technology, for example, Illumina (at least 100 times lower), sequencing mRNAs in a sample by NGS technology becomes a trend and this introduces a new computational problem.

The current Illumina sequencing technology can produce reads of length around 100 bp. Since the read length is short, reads cannot be aligned to known protein sequences or be annotated easily. Thus, an additional assembling step is required to combine reads to construct a longer sequence, called a *contig*, with length at least 300 bp for better alignment and annotation. Note that the sampled reads might still not be able to align to known gene sequences because either the mutation rate of microbes is high or the gene sequences of many microbes are still unknown (Eisen, 2007). However, a longer sequence of mRNA will make the analysis easier.

To our best knowledge, there is no existing assembler specially designed for assembling reads from metatranscriptomic data. Existing assemblers for genomic data (Huang et al., 2003; Mullikin and Ning, 2003; Zerbino and Birney, 2008; Simpson et al., 2009; Peng et al., 2012), transcriptomic data (Peng et al., 2011b; Schulz et al., 2012), and metagenomic data (Peng et al., 2011a) cannot be applied on the metatranscriptomic data because of the following reasons:

- *Uneven sequencing depth.* In genomic data, reads are sampled uniformly along a single genome. Thus, the number of reads sampled from each position is similar; that is, the sequencing depth at each position is about the same. For transcriptomic data and metagenomic data, reads are sampled from a mixture of mRNAs and genomes, respectively. Since the expression levels of genes vary and the abundances of species in a sample are different, the sequencing depth of different mRNAs (transcriptomic data) and genome (metagenomic data) can vary a lot (can be over 100 times different). However, this problem becomes more serious in metatranscriptomic data. Since the abundance ratios of different kinds of microbes in the sample are different [over 1,000 times difference (Qin et al., 2010)] and the expression levels of different mRNA from the same kind of microbes also vary, the number of reads sampled from low-expressed mRNA from a microbe with a low abundance ratio is much smaller than the number of reads sampled from high-expressed mRNA from the microbe with a high abundance ratio (can be 100,000 times difference). Alignment of reads from mouse gut metatranscriptomic data (Xiong et al., 2012) to known mRNAs shows that the differences in the abundance can be over 20,000 times among mRNAs in the top 20 most abundant families in the sample. The abundances of these low-abundance families cannot be estimated accurately as the numbers of reads aligned to these families are small. As a result, error reads sampled from high-abundance mRNAs

Table 1. Number of Repeat Patterns in Known Gene Sequences

Repeat length, bp	*Repeat patterns*	*No. of genes containing repeats (% genes)*	*No. of genes containing repeats at the two ends, 100 bp (% genes)*
30	6,856,156	51,504 (53.27)	49,100 (51.77)
40	4,253,013	42,655 (43.96)	38,922 (41.05)
50	3,677,168	38,682 (39.99)	32,996 (34.80)
60	3,226,316	36,307 (37.49)	29,040 (30.62)
70	2,840,548	34,662 (35.79)	26,559 (28.01)
80	2,495,914	31,791 (32.96)	23,053 (24.31)
90	2,215,876	26,341 (27.37)	16,219 (17.10)
100	2,004,400	23,582 (24.53)	13,108 (13.82)

appear much more [70 times more (Xiong et al., 2012)] than the correct reads sampled from low-abundance mRNAs. Existing assemblers usually determine error reads based on their sampling rates based on an assumption that correct reads appear more than erroneous reads in the sample. MetaIDBA (Peng et al., 2011a) designed for metagenomic data can resolve some of these problems based on the idea that there are not many similar patterns between two different genomes. However, MetaIDBA cannot work on metatranscriptomic data as the sequences of mRNAs contain many similar patterns as mentioned below. IDBA-UD [18] works on genomic data and can solve part of the problem by considering local coverage of each contig and produce longer contigs. However, it produces many incorrect contigs among complicated repeat regions as the local coverages of repeat regions are much different from other parts.

- *Repeat patterns occurring in different mRNAs.* Repeat patterns in the genomes and mRNAs usually introduce ambiguity, which leads to very short contigs for existing assemblers. Compared with genomic and metagenomic data, there are much more repeat patterns in metatranscriptomic data. Proteins with similar functionality, for example, proteins in the same family, usually have similar substructures and conserved patterns in the amino acid sequences (Glazer and Kechris, 2009). Thus, the mRNAs encoding proteins with similar functionality usually contain similar patterns even if they are from genomes of different microbes. As microbes in the same environment usually produce some proteins with similar functionality with respect to the environment, the number of repeat patterns in metatranscriptomic data is high. For example, Table 1 shows the number of repeats for different lengths in the bacteria gene sequences from the GenBank (Benson et al., 2000) of known sources (different versions of the same genes from the same bacteria have been removed before the analysis). We can see that 24.53% of genes contain at least one repeat region with length at least 100 bp. Thus, these repeats cannot be resolved using length-100 reads. Because of the existence of repeat patterns, existing assemblers based on the de Bruijn graph (Zerbino and Birney, 2008; Simpson et al., 2009; Peng et al., 2011a; Peng et al., 2011b; Peng et al., 2012) or string graph (Simpson and Durbin, 2010) construct a graph with many branches, which leads to short contigs. More seriously, when the repeat pattern occurs at the beginning of one mRNA and at the end of another mRNA (13.82% of genes contain a length-100 repeat near the end), existing assemblers may merge these two mRNAs incorrectly as a single chimeric contig. Figure 1 shows an example of a chimeric contig for mouse gut transcriptomic data (Xiong et al., 2012). These kinds of chimeric contigs produced by existing assemblers will affect the annotation of mRNAs or be considered as mRNAs from fusion genes.

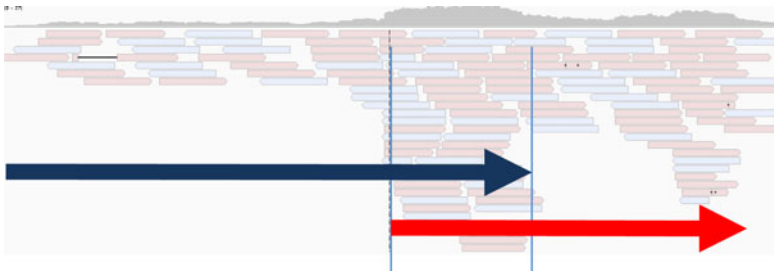

FIG. 1. Chimeric contig produced by IDBA-UD. The blue arrow and red arrow represent part of two different mRNAs sharing common patterns merged incorrectly by IDBA-UD. The short arrows in pale colors represent sampled reads aligned to the contig.

In order to solve the above problems, we introduce the algorithm IDBA-MT for assembling metatranscriptomic data. IDBA-MT applies IDBA-UD local coverage idea (Step 1) which can produce longer contigs for data with uneven sequencing depth. Although some chimeric contigs are produced also, IDBA-MT can determine chimeric contigs (Step 2) and resolve merged mRNAs using the *k*-mer multiplicity (local support) at each vertex and paired-end information (Step 3).

IDBA-MT constructs contigs in a similar fashion as IDBA-UD (Peng et al., 2012) using the de Bruijn graph with multiple *k*. However, IDBA-MT will not consider all simple paths in the de Bruijn graph as a single mRNA. With the assumption that reads are sampled uniformly from each mRNA, the vertex supports along the contigs should be similar, even though each might differ a lot from each other because of the different abundance of each mRNA (we shall discuss later if this assumption is invalid). Thus, the mis-assembled chimeric contigs be identified from the abrupt change of vertex support along the contig. A sudden increase (decrease) of support *junction* signals the start of a repeat region. If the insert distance of the paired-end reads is longer than the repeat region, IDBA-MT might be able to identify the repeat region and decompose the chimeric contig into two separate contigs with a common similar region by means of the paired-end information (Step 4). Note that if the reads are not sampled evenly (the assumption is invalid), there may exist false-positive junctions (due to change of support) along the contig. These false-positive junctions can be easily identified if there exist two paired-end reads aligned to the contig with different end reads covering the junction. The uneven sequencing depth problem can also be alleviated. Some correct short dead-end paths with low support, called *tips*, that were removed in the error correction step can be recovered after the problem of chimeric contigs is solved (Step 5). The resolved contigs can then be further extended by reads that were considered as tips before. A workflow of IDBA-MT is shown in Figure 2.

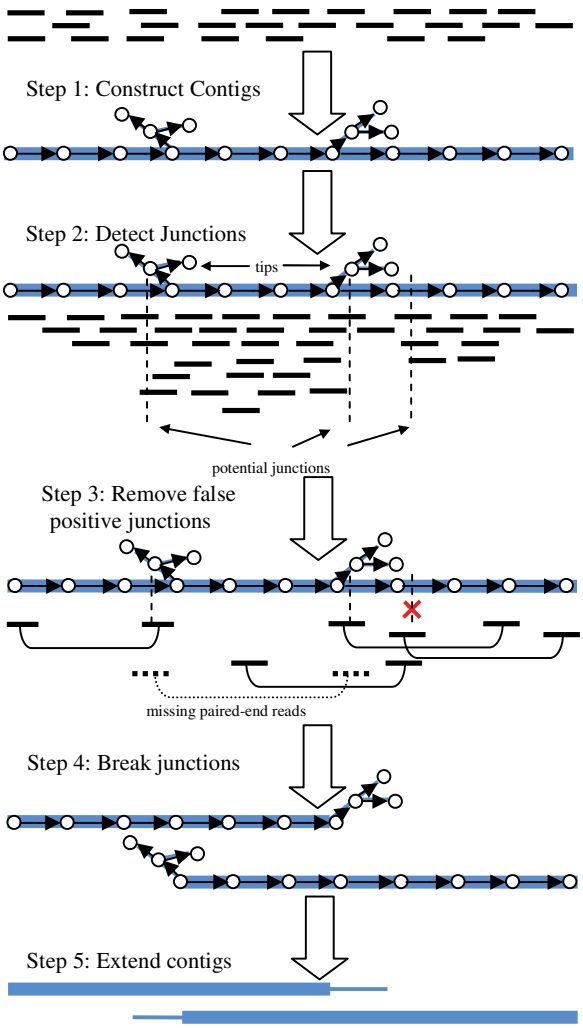

FIG. 2. Flowchart of algorithm IDBA-MT.

We have tested the performance of IDBA-MT on the contigs produced by Oases (Schulz et al., 2012), Trinity (Grabherr et al., 2011), and IDBA-UD (Peng et al., 2012) based on simulated data. IDBA-MT can reduce the error rate (due to chimeric contigs) from 4.22%, 40.98%, 4.86% to 1.28% (in term of total length) when compared with Oases, Trinity, and IDBA-UD, respectively, and with a higher coverage of 61.85% compared with 23.29%, 16.00%, and 57.49% for Oases, Trinity, and IDBA-UD, respectively. For real biological data, IDBA-MT can also produce more correct contigs (in term of length) that can be aligned to known protein sequences.

2. METHODOLOGY

IDBA-MT is a program for assembling metatranscriptomic data. It applies a similar approach as IDBA-UD (Peng et al., 2012) for constructing a set of contigs using multiple k values (from $k_{\min}$ to $k_{\max}$) and local assembling technique. These contigs will then be mapped to a path in the de Bruijn graph with $k_{\min}$. Note that this path might not be simple and contains many branches. For each contig, a set of potential wrong merging junctions will be detected based on the support of vertices in the de Bruijn graph. False-positive junctions will be determined using paired-end information. Repeat regions are then determined based on the positions of junctions, and the contigs will be broken down into shorter contigs. These shorter contigs will be further extended into some dead-end branches (tips) based on some low-coverage reads treated as erroneous before. In this section, we will describe these four steps in details.

2.1. Construct potential contigs using IDBA-UD

Similar to IDBA-UD (Peng et al., 2012), IDBA-MT starts with a de Bruijn graph using a small k value; that is, each vertex represents a length-k substring (k-*mer*) in the read, and there is an edge from vertex u to v if the length-$(k-1)$ suffix of u is the same as the length-$(k-1)$ prefix of v and the k-mers u and v appear consecutively in a read. Short dead-end branches (tips) are removed, similar paths are merged (merging bubbles), and some repeats are solved by local assembling technique. Short contigs are constructed from each simple path. These short contigs and all input reads will then be used to construct a de Bruijn graph with a larger k value for constructing longer contigs.

However, since the lengths of mRNAs are much shorter than a chromosome, instead of using a large threshold ($2k$) for removing tips, IDBA-MT uses a smaller threshold (k) for removing tips. Note that since IDBA-UD removes tips at each iteration, although the difference between the two thresholds are small, the accumulate effect is large after many iterations (e.g., k value ranges from $k_{\min}=20$ to $k_{\max}=100$). IDBA-MT also applies a lower threshold when using paired-end information (only one paired-end read is needed as support for connecting two contigs). Using a low threshold may lead to longer contig with more error (chimeric contigs) in normal situation. However, since IDBA-MT will break the chimeric contig in the next few steps, IDBA-MT prefers to produce longer contig in the first step even many of them may be chimeric contigs.

2.2. Detect potential wrong merging junctions

Although IDBA-UD produces fewer chimeric contigs than other existing assemblers, it still produces some chimeric contigs especially IDBA-MT and applies a lower threshold for tips and paired-end reads. In order to detect wrong merging junctions alone, an erroneous contig (a junction represent the starting or ending position of a repeat region), input reads should be aligned to the contigs and the number of reads aligned at each position can be determined. Junctions can be determined when the number of aligned reads at two adjacent positions differ a lot (higher than some threshold). However, as there are errors in reads, substitution errors should be allowed when aligning reads to contigs. These substitution errors introduced a problem, as a read sampled from repeat regions of an mRNA may align across a junction and affect the number of reads covering the junction. Figure 3 gives an example of wrongly aligned reads across a junction. The junction should be between positions 8 and 9. However, as substitution error is allowed, there are three reads that start with ''CAACT'' aligned across the junction.

In order to detect the junction more accurately, all contigs are mapped back to the de Bruijn graph with $k=k_{\min}$; that is, each contig is represented by a path in the de Bruijn graph. On the basis of the assumption that k-mers in the repeat regions are sampled more than k-mers in the unique regions, the boundary of

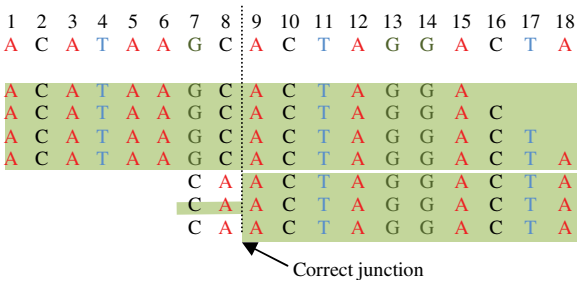

FIG. 3. Example of wrongly aligned reads across a junction.

repeat regions (potential junctions) can be detected if the multiplicity of two adjacent vertices (k-mers) is larger than one standard deviation of coverage of the contig. For example, in Figure 3, the multiplicity of the 5-mer ''CACTA'' start at position 8 is 3, while the multiplicity of the 5-mer ''ACTAG'' start at position 9 is 7; there is a potential junction between positions 8 and 9 of the contigs (between the 5-mer ''CACTA'' and ''ACTAG'' in the de Bruijn graph).

2.3. *Determine repeat regions*

Every paired-end read should be sampled from the same mRNA. Consider the paired-end reads sampled along a true-positive contig. For any position, there should be both first-end and second-end reads aligned unless the position is at the ends of the contigs. Here, the definition of first end and second end are based on the position of reads aligned to the contigs instead of the sequencing order. If there is a position in the middle of a contig such that only first (second) ends of paired-end reads can be aligned, there should be a repeat region before (after) that position. Figure 4 shows an example of a junction that only second-end reads aligned. Thus, the starting and ending positions of the repeat region can be determined near the junctions.

When the sequencing depth of an mRNA is low, some positions may have only first- or second-end reads aligned. These positions with a low sampling rate may mislead IDBA-MT when determining repeat regions. Thus, we should calculate the probability that only one end of paired-end reads aligned at a particular position by random. If this probability is low (say $<5\%$), we could assume that there is a repeat region nearby with high confident.

For simplicity, we assume that the read length is l and the insert distance is exactly d without error. The calculation can be extended easily when the read length varies (because of sequencing quality) and the insert distance following normal distribution with mean distance d and standard deviation σ.

Given a position i (start at 0) on a contig with q paired-end reads aligned at the position i; that is, the starting position of the aligned length-l read is in the region $[i-l+1, i]$. If the second end of a paired-end read aligned at i, the possible position of the first aligned position is $[i-d+1, i-d+l]$. Thus, the number of possible aligned positions for the first end is

$$r_{1\text{st}} = \begin{cases} 0 & \text{i-d } +1<0 \\ l & \text{i-d } +1 \geq 1-1 \\ i-d+l+1 & \text{otherwise} \end{cases}$$

Similarly, if the first end of a paired-end read aligned at the position i, the possible position of the second aligned position is $[I+d-l, I+d-1]$. The number of possible aligned positions for the second end is

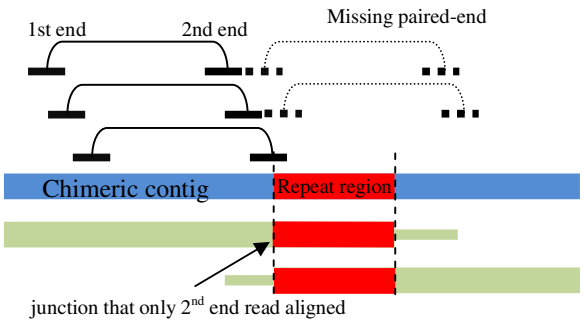

FIG. 4. Example of removing a false-possible junction.

$$r_{2nd} = \begin{cases} 0 & \text{i} + \text{d} - \text{l} \geq \text{n} \\ l & \text{i} + \text{d} - \text{l} \leq \text{n} - 1 \\ n - (i + d - l) & \text{otherwise} \end{cases}$$

The probability that a read aligned at the position i be the first end and second end are $r_{1st}/(r_{1st} + r_{2nd})$ and $r_{2nd}/(r_{1st} + r_{2nd})$, respectively. The probability that all q paired-end reads aligned at the position i have the same end aligned at the position i is

$$p_{rand} = \left(\frac{r_{1st}}{r_{1st} + r_{2nd}}\right)^q + \left(\frac{r_{2nd}}{r_{1st} + r_{2nd}}\right)^q$$

By considering $p_{rand} < 5\%$, we can calculate the minimum support q for determining the repeat region near the position i.

2.4. Decomposing chimeric contig based on repeat regions

Consider a contig with exactly two junctions that divide the contig into three regions A, B, and C, where A is the region before the first junction, B is the region between the two junctions, and C is the region after the second junction. B is a repeat region and the contig should be broken down into two shorter contigs. One contig is the concatenation of regions A and B, and the other is the catenation of regions B and C. Thus, the corresponding path in the de Bruijn graph should be divided into two paths with two copies of paths corresponding to region B. The end of the region A path should be connected to the start of the region B path. The end of the region B path should be connected to the start of the region C path (Figure 2). If there are multiple repeat regions in the contig, the above step can be performed for each repeat region.

2.5. Extending contig

After dividing the de Bruijn graph, some of the removed tips (branches with a relatively low coverage) may become part of the simple path again. The broken contig can be extended further along these tips.

3. EXPERIMENTAL RESULT

We evaluated assemblers Oases (Schulz et al., 2012), Trinity (Grabherr et al., 2011), IDBA-UD (Peng et al., 2012), and IDBA-MT on a real dataset from the mouse gut (Xiong et al., 2012) and three simulated datasets generated from known bacteria gene sequences obtained from GenBank (Benson et al., 2000). Oases, Trinity, and IDBA-UD are designed for assembling transcriptomic data based on constructing a de Bruijn graph. A simple path in the graph with no branches (vertex with in-degree or out-degree larger than one) represents a contig. All these algorithms employ different procedures for removing branches because of errors, such as removing tips and merging bubbles. However, as the de Bruijn graph for metatranscriptomic data contains many branches, these procedures may not be able to remove some of the erroneous branches or may remove some correct branches and then lead to chimeric contigs. Trinity has a butterfly procedure for handing alternative splicing, but this butterfly procedure in Trinity might not be helpful for our bacteria datasets as bacteria seldom have alternative splicing.

3.1. Simulated data

We downloaded all bacteria gene sequences from the GenBank (Benson et al., 2000) with known sources. However, since the same genes from the same species (same or different strains) may be recorded several times, these duplicated sequences are removed and only one version is kept. As a result, we obtained 94,827 gene sequences. Among these sequences, many of them share similar regions to different extents; that is, two different genes from the same species or different species share some similar patterns. Among them, there are 658 sequences share at least half of the sequences with other gene sequences. We generate the simulated dataset based on these gene (mRNA) sequences. We target for these sequences because the problem of generating chimeric contigs is more serious for reads sampled from these sequences, and the problem of generating chimeric contigs by different assemblers can be evaluated directly.

Table 2. Experimental Results on Simulated Data with Extreme Abundance Ratios

Software	Coverage, %	Maximum length, bp	Average length, bp	No. of wrong contig (length, bp)	No. of correct contig (length, bp)	Error rate, %
Oases	5.29	542	191	11 (1,357)	161 (30,798)	4.22
Trinity	3.67	1,117	351	14 (15,696)	64 (22,601)	40.98
IDBA-UD	27.68	1,172	342	5 (2,702)	154 (52,879)	4.86
IDBA-MT	39.63	1,675	462	3 (1,358)	227 (104,979)	1.28

For each dataset, we randomly picked length-75 bp paired-end reads from the sequences with sequencing error rate of 1%. The mean insert distance of the paired-end reads are 200 bp with a standard deviation of 10 bp. We generate three datasets with different numbers of randomly picked sequences and abundance ratios.

The assemblers are tested on these sampled reads. We determine if a contig is correctly assembled by aligning it to the mRNA sequences using blat (Kent, 2002) with at least 95% of the regions of a contig aligned to the mRNA sequence. Otherwise, it is considered as wrong. Those positions of mRNA sequences aligned by a correct contig are considered as covered by the contig and the coverage of a set of mRNA sequences is the percentage of positions covered by the set of correct contigs. Oases and IDBA-UD use paired-end reads for merging contigs to form scaffolds. However, as their procedures for forming scaffolds are not designed for metatranscriptomic data, the error rates of the scaffolds are several times higher than those of the contigs. Thus, we only compare the contigs' performance produced by the assemblers.

3.2. Simulated data with extreme abundance ratios

A total of 120 mRNA sequences are randomly picked and reads are sampled from the sequences with 20 sequences having sequencing depth 1000× and the rest 100 sequences having sequencing depth 3×. The experimental results are shown in Table 2.

Since the abundance ratios of 20 mRNA sequences are much higher than the rest sequences, Both Oases and IDBA-UD can assemble the reads sampled from high-abundance mRNAs quite well. However, for those mRNAs with low abundance ratios, only IDBA-UB can assemble a small portion of them. Thus, the coverages of the contigs produced by both algorithms are less than 30%. Trinity does not perform well in this dataset because it merges several contigs into longer chimeric contigs. This problem becomes less serious when there are more mRNAs in the sample that introduce branches in the de Bruijn graph and prevent Trinity from merging contigs (Tables 2 and 4).

IDBA-MT can determine wrong merging junctions in the contigs and break them down into shorter contigs (long enough for protein annotation). Thus, it has the highest coverage (39.63%) and the lowest error rate (1.28%). The average length of the contigs is 462 bp, which is longer than the contigs produced by other assemblers.

3.3. Simulated data with similar abundance ratios

Reads are sampled from all 658 gene sequences having long repeats with sequencing depth 3×. Compared with the simulated data with extreme abundance ratios, the coverages of contigs produced by all assemblers increase (Table 3). It is because the sequencing depths of genes are similar and the assemblers will not treat reads sampled from low-abundance mRNAs as erroneous.

Although all assemblers perform well in this ideal situation, IDBA-MT still produces the most correct contigs when compared with these assemblers. It has the highest coverage (61.85%) and the second lowest

Table 3. Experimental Results on Simulated Data with Equal Abundance Ratios

Software	Coverage, %	Maximum length, bp	Average length, bp	No. of wrong contig (length, bp)	No. of correct contig (length, bp)	Error rate, %
Oases	23.29	598	175	34 (4,169)	436 (76,394)	5.17
Trinity	16.00	713	300	84 (41,898)	348 (105,074)	28.51
IDBA-UD	57.49	1,857	364	37 (23,600)	624 (227,795)	9.39
IDBA-MT	61.85	704	280	18 (6,664)	408 (114,476)	5.50

Table 4. Experimental Results on Simulated Data with Mixture of Abundance Ratios

Software	*Coverage, %*	*Maximum length, bp*	*Average length, bp*	*No. of wrong contig (length, bp)*	*No. of correct contig (length, bp)*	*Error rate, %*
Oases	31.00	676	194	63 (8,471)	1,009 (196,162)	4.14
Trinity	15.10	1,270	319	106 (75,713)	310 (99,603)	43.18
IDBA-UD	54.12	1,279	419	43 (24,241)	489 (205,150)	10.57
IDBA-MT	58.20	1,511	317	36 (15,084)	847 (268,949)	5.31

error rate (5.50%) among all assemblers. Oases has the lowest error rate (5.17%), but it produces much shorter contigs (average length = 175 bp) than IDBA-MT (average length = 280 bp) and has a lower coverage (23.29%) than IDBA-UD (61.85%). The lengths of the contigs produced by Oases are too short for annotation. On the other hand, IDBA-UD produces the longest contig in this dataset and the average length of their contigs is the highest among all assemblers. It is because it applies multiple k values for filling in gaps and extends a contig even the support is low. However, long error contigs are produced too. Among the contigs longer than 1,000 bp produced by IDBA-UD, 40% of them are incorrect. Moreover, since contigs with length longer than 300 bp can be annotated well, although IDBA-MT breaks the three long contigs into shorter contigs of length 500 bp to 700 bp, these shorter contigs are correct and can be annotated. It may not be worth to produce longer contigs as some of them are incorrect chimeric contigs.

3.4. Simulated data with mixture of abundance ratios

In a real situation, the abundance ratios of different mRNAs are not in two extremes or almost the same. Instead, the abundance ratios of mRNAs vary continuously from a high to a low abundance ratio. Thus, we generate a more realistic dataset with the abundance ratios of mRNAs range from 1000× to 3× following the power law (number of mRNAs with a certain abundance is directly proportional to the negative of abundance ratio).

Compared with the other simulated datasets, the contigs produced by IDBA-UD and Trinity contain more errors as the reads sampled from low-abundance mRNAs act as noise for the high-abundance mRNAs. Moreover, this noise cannot be distinguished from the reads sampled from high-abundance mRNAs because there is a mixture of mRNAs with different abundances. As a result, IDBA-UD, Trinity, and IDBA-MT have lower coverages than the dataset in III.C.

For these datasets, IDBA-MT still has the highest coverage (58.20%) and the second lowest error rate (5.31%) among all assemblers (Table 4). Again, Oases achieves the lowest error rate (4.14%) but produces shorter contigs (average length = 194 bp) than IDBA-MT (average length = 317 bp) and has a lower coverage (31.00%) than IDBA-MT (58.12%). IDBA-UD produced longer contigs on average. However, similar as simulated data with similar abundance ratios, IDBA-UD produces longer contigs with more error. The longest and second longest contigs produced by IDBA-UD are incorrect and the length of the rest of the contigs do not differ a lot.

3.5. Real data on the mouse gut

Xiong et al. (2012) isolated RNAs from the lumen of the cecum and colon of four 12-week-old mice. Paired-end reads were generated using Illumina sequencing technology. The read length is about 75 bp and the insert distance is about 300 bp. Since the number of paired-end reads generated for each sample is small, we merge all paired-end reads in the sample and compare the assembly results from Oases, Trinity, IDBA-UD, and IDBA-MT (Table 5).

Table 5. Experimental Results on Real Mouse Gut Data

Software	*Maximum length, bp*	*Average length, bp*	*No. of contig*	*Total length, bp*	*No. of contig aligned to known proteins (length, bp)*
Oases	693	127	99,611	12,655,199	489 (84,044)
Trinity	15,857	500	19,721	9,862,469	7,188 (2,994,588)
IDBA-UD	10,741	490	18,951	9,287,101	9,510 (4,178,162)
IDBA-MT	8,863	490	18,972	9,301,484	9,515 (4,181,949)

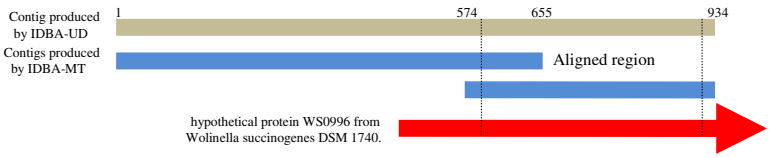

FIG. 5. Example of chimeric contigs resolved by IDBA-MT.

Since there are no reference mRNAs for verification, we align the contigs to known protein sequences using Blastx with default parameters. A contig is considered ''correct'' if at least 90% of the contig sequence can be aligned to a single known protein. Since Oasis produces very short contigs, many of them cannot be aligned to known proteins. Thus, the number of correct contigs is small. Trinity can construct longer contigs than Oasis. More contigs can be aligned to known proteins. IDBA-UD and IDBA-MT, which both apply local support and paired-end information and have similar performance, can assemble longer and more correct contigs than Oases and Trinity. Note that contigs that cannot be aligned to known protein sequences may not be false positive, as there are many unknown proteins.

Figure 5 shows an example of chimeric contigs produced by IDBA-UD. The whole contigs cannot be aligned to any known protein sequences. However, the last 300 bp of the sequences can be aligned to a hypothetical protein predicted from the genome of Wolinella succinogenes DSM 1740. IDBA-MT can detect this chimeric contig and decompose it into two shorter contigs with the second one aligned to this hypothetical protein.

4. CONCLUSIONS AND FUTURE WORKS

NGS technology provides a great opportunity for analyzing metatranscriptomic data. However, to our best knowledge, no assembler works well on metatranscriptomic data. Existing assemblers for genomic data, transcriptome data, and metagenomic data produce many chimeric contigs when work with metatranscriptomic data. We have introduced a software tool called IDBA-MT, which can assemble metatranscriptomic datasets with a much lower error rate than existing assemblers.

However, when the number of repeat regions in the mRNAs increases or the abundances of most of the mRNAs are very low, all assemblers, including IDBA-MT, do not perform well. Our next target is to improve this assembly tool for datasets with a very high sequencing depth.

ACKNOWLEDGMENTS

This research was partially supported by Juvenile Diabetes Research Foundation #17-2011-520, RGC HKU 7111/12E, and HKU 719709E. We thank Jayne Danska and Janet Markle in the Hospital for Sick Children, Faculty of Medicine, University of Toronto, for providing the data used in this analysis.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Benson, D., Karsch-Mizrachi, I., Lipman, D., et al. 2000. GenBank. *Nucleic Acids Res.* 28, 15–18.

Booijink, C., Boekhorst, J., Zoetendal, E., et al. 2010. Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. *Appl. Environ. Microbiol.* 76, 5533–5540.

Bosch, J.T., and Grody, W. 2008. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J. Mol. Diagn*. 10, 484–492.

Eisen, J. 2007. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol*. 5, e82.

Frias-Lopez, J., Shi, Y., Tyson, G., et al. 2008. Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA* 105, 3805–3810.

Fullwood, M., Wei, C., Liu E., et al. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*. 19, 521–532.

Gilbert, J., Field, D., Huang, Y., et al. 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* 3, e3042.

Glazer, A., and Kechris, K. 2009. Conserved amino acid sequence features in the α subunits of MoFe, VFe, and FeFe nitrogenases. *PLoS ONE* 4, e6136.

Grabherr, M., Haas, B., Yassour, M., et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol*. 29, 644–652.

Huang, X., Wang, J., Aluru, S., et al. 2003. PCAP: a whole-genome assembly program. *Genome Res*. 13, 2164–2170.

Kent, J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*. 12, 656–664.

Khachatryan, Z., Ktsoyan, Z., Manukyan, G., et al. 2008. Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS ONE* 3, e3064.

Leininger, S., Urich, T., Schloter, M., et al. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442, 806–809.

Morozova, O., and Marra, M. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92, 255–264.

Mullikin, J., and Ning, Z. 2003. The phusion assembler. *Genome Res*. 13, 81–90.

Parro, V., Moreno-Paz, M., and Gonzalez-Toril, E. 2007. Analysis of environmental transcriptomes by DNA microarrays. *Env. Microbiol*. 9, 453–464.

Peng, Y., Leung, H., Yiu, S., et al. 2011a. T-IDBA: a *de novo* iterative de Bruijn graph assembler for transcriptome. *In Proceedings of RECOMB*, 337–338.

Peng, Y., Leung, H., Yiu, S., et al. 2011b. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27, i94–i101.

Peng, Y., Leung, H., Yiu, S., et al. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428.

Pettersson, E., Lundeberg, J., and Ahmadian, A. 2009. Generations of sequencing technologies. *Genomics* 93, 105–111.

Poretsky, R., Bano, N., Buchan, A., et al. 2005. Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* 71, 4121–4126.

Poretsky, R., Sun, S., Mou, X., et al. 2010. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ. Microbiol.* 12, 616–627.

Qin, J., Li, R., Raes, J., et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65.

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

Simpson, J., and Durbin, R. 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26, i367–i373.

Simpson, J., Wong, K., Jackman, S., et al. 2009. Assembly by short sequences—a *de novo*, parallel, paired-end sequence assembler. *Genome Res.* 19, 1117–1123.

Tartar, A., Wheeler, M., Zhou, X., et al. 2009. Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. *Biotechnol. Biofuels*. 2, 25.

Urich, T., Lanzen, A., Qi, J., et al. 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* 3, e2527.

Xiong, X., Frank, D., Robertson, C., et al. 2012. Generation and analysis of a mouse intestinal metatranscriptome through Illumina based RNA-sequencing. *PLoS ONE* 7, e36009.

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

Address correspondence to:
Dr. Henry C.M. Leung
Department of Computer Science
The University of Hong Kong
LG101, CYC Building, Pokfulam Road
Hong Kong
People's Republic of China

E-mail: cmleung2@cs.hku.hk