IDBA-MT: De novo Assembler for Metatranscriptomic Data generated from Next-Generating Sequencing Technology

Henry C.M. Leung*
Department of Computer Science,
The University of Hong Kong,
Pokfulam Road, Hong Kong
Email: cmleung2@cs.hku.hk
Telephone: 852-2857-8269

S.M. Yiu
Department of Computer Science,
The University of Hong Kong,
Pokfulam Road, Hong Kong
Email: smyiu@cs.hku.hk
Telephone: 852-2857-8242

John Parkinson
Biochemistry & Molecular and Medical Genetics
University of Toronto, 27 King's College Circle, Toronto, Ontario, Canada M5S 1A1
Email: john.parkinson@utoronto.ca
Telephone: 416-813-5746

Francis Y.L. Chin
Department of Computer Science,
The University of Hong Kong,
Pokfulam Road, Hong Kong
Email: chin@cs.hku.hk
Telephone: 852-2859-2178

* Corresponding authors

ABSTRACT

Motivation: 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, i.e. incorrect contigs formed by merging multiple mRNA sequences. To our best knowledge, there is no assembler designed for metatranscriptomic data.
Results: In this paper, we introduced an assembler called IDBA-MT which is designed for assembling reads from metatranscriptomic data. Compared with other assemblers, IDBA-MT produces much fewer chimeric contigs (reduce by 50% or more) when compared with existing assemblers like Oases, IDBA-UD and Trinity.

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 (IBD) (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.

Although metagenomic analysis provide 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 (Booijink et al, 2010; Poretsky et al., 2005; Tartar et al., 2009) 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, say bacteria. By growing the host, multiple copies of the implanted cDNA can be obtained for analysis. However, constructing cDNA clone libraries is labour 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 technology (NGS) (Bosch and Grody, 2008; Fullwood et al., 2009; Morozova and Marra, 2008; Petterssonet 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; Poretsky et al., 2010; Urich et al., 2008) 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, e.g. 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 (Peng et al., 2012; Simpson, et al., 2009; Zerbino and Birney, 2008), 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, i.e. 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 microbe with low abundance ratio is much smaller that the number of reads sampled from high expressed mRNA from microbe with 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 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. Thus, shorter contigs or incorrect contigs are produced. 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 (Peng et al., 2012) work on genomic data can solve part of the problem by considering local coverage of each contig and produce longer contigs. However, it produces many incorrect contigs among complicate repeat regions as the local coverage of repeat regions are much different from the other part.

- *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, say 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 geneBank (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 de Bruijn graph (Peng et al., 2011a; Peng et al., 2011b; Peng et al., 2012; Simpson et al., 2009; Zerbino and Birney, 2008) 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.

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 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 exists 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*, which 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.

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 higher coverage 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, i.e. 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 de Bruijn graph with 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 different 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 applies a lower thresholds for tips and paired-end reads. In order to detect wrong merging junctions alone a 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 problem as a read sampled from repeat regions of an mRNA may align cross over a junction and affects the number of reads covering the junction. Figure 3 gives an example of wrongly aligned reads across a junction. The junction should be between position 8 and 9. However, as substitution error is allowed, there are three reads start with "CAACT" aligned across the junction.

In order to detect the junction more accurate, all contigs are mapped back to the de Bruijn graph with $k = k_{min}$, i.e. each contig is represented by a path in the de Bruijn graph. Based on the assumption that k-mers in the repeat regions are sampled more than k-mers in unique regions, boundary of repeat regions (potential junctions) can be detected if the multiplicity of two adjacent vertices (k-mers) is larger than one standards 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 position 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 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 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 positions by random. If this probability is low (say < 5%), we could assume there is a repeat region nearby with high confident.

For simplicity, we assume 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 position i, i.e. 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

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

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

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

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

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

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

2.4 *Decomposing chimeric contig based on repeat region*

Consider a contig with exactly two junctions which 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. On contig is the concatenation of region A and B and the other is the catenation of region 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 tip (branches with relative low coverage) may become part of the simple path again. The broken contig can be extended further along to these tips.

3 *Experimental result*

We evaluated the assembler Oases (Schulz et al., 2012), Trinity (Grabherr et al., 2011), IDBA-UD (Peng et al., 2012) and IDBA-MT on a real dataset from 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 due to errors, such as removing tips, merging bubbles, etc. 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, i.e. 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.

For each dataset, we randomly picked length-75 bp paired-end reads from the sequences with 1% sequencing error. The mean insert distance of the paired-end reads are 200 bp with standard derivation 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 1000x and the rest 100 sequences having sequencing depth 3x. The experimental results are shown in Table 2.

Since the abundance ratios of 20 mRNA sequences are much higher than the other sequences, Both Oases and IDBA-UD can assemble the reads sampled from high abundance mRNAs quite well. However, for those mRNA with low abundance ratios, only IDBA-UD can assemble a small portion of them. Thus, the coverages of the contigs produced by both algorithms are lower 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 with similar abundances that introduce branches in the de Bruijn graph and prevent Trinity from merging contigs. Thus, Trinity has low error rate for the dataset mentioned in Section 3.3 (Tables 3).

IDBA-MT can determine wrong merging junctions in the contigs and break them down into correct 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 also 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 3x. Compared with the dataset in Section 3.3.2, 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 to these assemblers. It has the highest coverage (61.85%) and the second lowest error rate (5.50%) among all assemblers. Oasis 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 much 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* value 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 break the three long correct contigs into shorter contigs of length 500 bp to 700bp, these shorter contigs are correct and can be annotated. It may not worth to produce longer contigs with some of them are incorrect chimeric contigs.

3.4 *Simulated Data with mixture of abundance ratios*

In 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 high to low abundance ratio. Thus, we generate a more realistic dataset with the abundance ratios of mRNAs vary from 1000x to 3x following the power law (number of mRNAs with a certain abundance is directly proportional to the negative of abundance ratio).

Compared with the datasets in 3.3.2 and 3.3.3, 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 3.3.3.

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 lower coverage (31.00%) than IDBA-MT (58.12%). IDBA-UD produced longer contigs in average. However, similar as Section 3.3, IDBA-UD produces longer contigs with more error. The longest and second longest sequences produced by IDBA-UD are incorrect and the length of the rest contigs do not differs a lot.

3.5 *Real data on mouse gut*

Xiong et al. (Xiong et al., 2012) isolated RNAs from the lumen of the cecum and colon of 4 mice at 12 weeks old. 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 merges all paired-end reads in the sample and compare the assembly results from Oases, Trinity, IDBA-UD and IDBA-MT (Table 5).

Since there is 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 protein. 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 cannot be aligned to known protein sequences may not be false positive as there are many unknown proteins.

Figure 5 shows an example of a chimeric contig produced by IDBA-UD. The whole contig 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. Although IDBA-UD produces 22 longer contigs than IDBA-MT (IDBA-MT resolve the potential repeat regions in these contigs), alignment with known protein sequences shown that in all cases, when a contigs produced by IDBA-UD broken into 2 or more contigs by IDBA-MT, at least one of short contigs can be aligned to the known protein sequences better (in term of both the proportion of alignment length and alignment score). It may reflect that IDBA-UD produce more chimeric contigs than IDBA-MT.

4 *Conclusions and future works*

Next-generation sequencing 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 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 very high sequencing depth.

5 *Acknowledgment*

This research is partially supported by Juvenile Diabetes Research Foundation #17-2011-520, RGC HKU 7111/12E and HKU 719709E. Thanks Jayne Danska and Janet Markle for providing the real dataset on mouse gut used in this analysis.

6 *Author Disclosure Statement*

No competing financial interests exist.

7 *References*

Benson,D., Karsch-Mizrachi, I., Lipman, D., et al. 2000. GenBank. *Nucleic Acids Research* 28(1), 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(16), 5533–5540.

Bosch, J.T., 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 biology*. 5(3), e82.

Frias-Lopez, J. Shi, Y., Tyson, G. et al. 2008. Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci.* 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. Kechris, K. 2009. Conserved Amino Acid Sequence Features in the α Subunits of MoFe, VFe, and FeFe Nitrogenases. *PLoS ONE*. 4(7), 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(7), 644–652.

Huang, X., Wang, J., Aluru, S. et al. 2003. PCAP: AWhole-Genome Assembly Program. *Genome Res*. 13, 2164–2170.

Kent, J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res*. 12(4), 656–664.

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

Khachatryan, Z., Ktsoyan, Z., Manukyan, G. et al. 2008. Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS One*. 3(8), e3064.

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

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

Mullikin, J., Ning, Z. 2003. The Phusion Assembler. *Genome Res*. 13, 81–90.

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(13), 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(11), 1420–1428.

Pettersson, E., Lundeberg, J., 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(7285), 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(8), 1086–1092.

Simpson, J., Durbin, R. 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*. 26(12), 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(6), 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. *Biotech. for 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(6), 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(4), e36009.

Zerbino, D., Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18(5), 821–829.

Repeat length	Repeat patterns	# genes contain repeats	# genes contain repeats at the two ends (100 bp)
30 bp	6,856,156	51,504 (53.27% genes)	49,100 (51.77% genes)
40 bp	4,253,013	42,655 (43.96% genes)	38,922 (41.05% genes)
50 bp	3,677,168	38,682 (39.99% genes)	32,996 (34.80% genes)
60 bp	3,226,316	36,307 (37.49% genes)	29,040 (30.62% genes)
70 bp	2,840,548	34,662 (35.79% genes)	26,559 (28.01% genes)
80 bp	2,495,914	31,791 (32.96% genes)	23,053 (24.31% genes)
90 bp	2,215,876	26,341 (27.37% genes)	16,219 (17.10% genes)
100 bp	2,004,400	23,582 (24.53% genes)	13,108 (13.82% genes)

Table 1. Number of Repeat pattern in known gene sequences.

Software	Coverage	Maximum Length	Average Length	# of wrong contig (length)	# of correct contig (length)	Error Rate
Oases	5.29%	542 bp	191 bp	11 (1,357 bp)	161 (30,798 bp)	4.22%
Trinity	3.67%	1,117 bp	351 bp	14 (15,696 bp)	64 (22,601 bp)	40.98%
IDBA-UD	27.68%	1,172 bp	342 bp	5 (2702 bp)	154 (52,879 bp)	4.86%
IDBA-MT	39.63%	1,675 bp	462 bp	3 (1,358 bp)	227 (104,979 bp)	1.28%

Table 2. Experimental Result on simulated data with Extreme abundance ratios.

Software	Coverage	Maximum Length	Average Length	# of wrong contig (length)	# of correct contig (length)	Error Rate
Oases	23.29%	598 bp	175 bp	34 (4,169 bp)	436 (76,394 bp)	5.17%
Trinity	16.00%	713 bp	300 bp	84 (41,898 bp)	348 (105,074 bp)	28.51%
IDBA-UD	57.49%	1,857 bp	364 bp	37 (23,600 bp)	624 (227,795 bp)	9.39%
IDBA-MT	61.85%	704 bp	280 bp	18 (6,664 bp)	408 (114,476 bp)	5.50%

Table 3. Experimental Result on simulated data with equal abundance ratios.

Software	Coverage	Maximum Length	Average Length	# of wrong contig (length)	# of correct contig (length)	Error Rate
Oases	31.00%	676 bp	194 bp	63 (8,471 bp)	1009 (196,162 bp)	4.14%
Trinity	15.10%	1,270 bp	319 bp	106 (75,713 bp)	310 (99,603 bp)	43.18%
IDBA-UD	54.12%	1,279 bp	419 bp	43 (24,241 bp)	489 (205,150 bp)	10.57%
IDBA-MT	58.20%	1,511 bp	317 bp	36 (15,084 bp)	847 (268,949 bp)	5.31%

Table 4. Experimental Result on simulated data with Mixture of abundance ratios.

Software	Maximum Length	Average Length	# of contig	Total Length	# of contig aligned to known proteins (length)
Oases	693 bp	127 bp	99,611	12,655,199 bp	489 (84,044 bp)
Trinity	15,857 bp	500 bp	19,721	9,862,469 bp	7,188 (2,994,588 bp)
IDBA-UD	10,741 bp	490 bp	18,951	9,287,101 bp	9,510 (4,178,162 bp)
IDBA-MT	8,863 bp	490 bp	18,972	9,301,484 bp	9,515 (4,181,949 bp)

Table 5. Experimental Result on real mouse gut data.