Title	T-IDBA: A de novo Iterative de Bruijn Graph Assembler for Transcriptome
Author(s)	Peng, Y; Leung, HCM; Yiu, SM; Chin, FYL
Citation	The 15th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2011), Vancouver, BC., Canada, 28-31 March 2011. In Lecture Notes in Computer Science, 2011, v. 6577, p. 337-338
Issued Date	2011
URL	http://hdl.handle.net/10722/152000
Rights	The original publication is available at www.springerlink.com

T-IDBA: A de novo Iterative de Bruijn Graph Assembler for Transcriptome

Yu Peng, Henry C.M. Leung, S.M. Yiu, Francis Y.L. Chin

Department of Computer Science, The University of Hong Kong
Pokfulam Road, Hong Kong
{ypeng,cmleung2,smyiu,chin}@cs.hku.hk

Abstract. RNA sequencing based on next-generation sequencing technology is useful for analyzing transcriptomes, discovering novel genes and studying exon/intron structures. Similar to genome assembly, de novo transcriptome assembly does not rely on a reference genome and additional annotated information. Most, if not all, existing de novo transcriptome assemblers rely heavily on de novo genome assembly techniques without fully utilizing the properties of transcriptomes and may result in short contigs because of the splicing nature (shared exons) of the genes and the repeats that exist in different genes.
In this paper, we analyze the properties of the mammalian transcriptome and propose an algorithm to reconstruct expressed isoforms without a reference genome. We extend the iterative de Bruijn graph approach (IDBA) by using pair-end information to solve the problem of long repeats in different genes and the problem of branching in the same gene due to alternative splicing. The graph will be decomposed into small components, each of which corresponds to a few, if not single, genes. The most possible isoforms with sufficient support from the pair-end reads will be found heuristically by depth-first search. In practice, our de novo transcriptome assembler, T-IDBA, outperforms Abyss (one of the newest de novo transcriptome assembler) substantially in terms of sensitivity and precision for both simulated and real data. The experimental results also match with our theoretical analysis of the performance of T-IDBA, which guarantees most isoforms can be reconstructed as long as their coverage exceeds a certain threshold.
Availability: T-IDBA is available at http://www.cs.hku.hk/~alse/idba/

Keywords: de novo transcriptome assembly, de Bruijn graph, mate-pair reads, pari-end reads, mRNAs, exons

1 Introduction

RNA sequencing (RNA-Seq) is a recently developed technique to sequence cDNAs (complementary DNAs) generated from RNAs using the next-generation sequencing technologies (e.g. Illmina Genome Analyzer and Applied Biosystems SOLID). RNA-seq is becoming more important in the analysis of transcriptomes and has been used successfully in identifying novel genes, refining 5' and 3' ends of genes, studying gene functions [1] and locating exon/intron boundaries [2, 3]. By aligning the reads obtained from RNA-seq to a reference genome, [4] developed a method to discover a complete transcriptome for yeast. Furthermore, RNA-seq has been used to determine the expression levels of transcripts [5]. [6] studied the complexity of the transcriptome reconstruction problem (i.e. the reconstruction of all expressed isoforms and their expression levels) and showed that theoretically short reads (both single-end and pair-end) cannot guarantee a unique solution even if information of genes, exon

boundaries and isoforms are all known and given, IsoInfer [7] provides a practical solution by formulating the transcriptome reconstruction problem as a convex quadratic problem; by determining their abundance ratios (i.e. expression levels) based on the annotated information of the reference genome such as exon-intron boundaries and TSS-PAS (transcript start sites and polyadenylation sites) pair information; and heuristically searching for the best possible isoforms and expression levels. Cufflink [8] and Scripture [9] are two other recently published methods using gene and exon-intron boundary information generated by TopHat [3] to reconstruct isoforms. Both approaches build a graph in which exons are the nodes and two exons are connected if there are reads that connect them. Cufflink [8] assigns weights to the edges of the graph and models the isoform reconstruction problem as a minimum path cover problem while Scripture [9] creates a statistical model to identify significant segments as isoforms.

Similar to the genome assembly problem, the de novo transcriptome assembly problem (the problem of reconstructing isoforms without a reference genome and annotated information) is also very important. Transcriptome assembly methods that rely on a reference genome and additional annotated information may suffer from missing and erroneous information of some genes or exons in the database and also cannot detect structural variations in the sample. Moreover, the quality of these methods depends heavily on the accuracy of the alignment tools [3]. As RNA-seq technology becomes more mature, there will be an increasing need to reconstruct unknown mRNAs in the sample without any reference genome information. However, there has been little progress on the de novo transcriptome assembly problem. Most, if not all, existing approaches apply de novo genome assembly techniques (i.e. de Bruijn graph [10-12], string graph [13]) directly to solve the de novo transcriptome assembly problem (e.g. [14, 15]) without fully utilizing the properties of transcriptomes. The performance of these approaches, in particular for the reconstruction of isoforms for the same gene, is not satisfactory. There are other approaches (e.g. [16]) that construct isoforms based on ESTs (Expressed Sequence Tag); for example, [16] used the de Bruijn graph to construct a splicing graph for ESTs. These approaches are usually not scalable and not applicable to massive short reads.

Seemingly, transcriptome assembly is an easier problem than genome assembly for eukaryotics, such as mouse and human), as there are at most 40~50 thousands of transcripts, of length at most a few thousands of nucleotides, while the chromosomes are much longer (up to hundreds of millions). The following issues make the de novo transcriptome assembly problem different from the genome assembly problem. (1) Due to the splicing nature of the genes (for eukaryotes[1]), the same exon may be used in many different isoforms. This implies that, in both the de Bruijn graph and the string graph, there exist many branches in the subgraph that corresponds to a particular gene. The algorithms designed for de novo genome assembly problem usually stop extending the contigs at branches. In order to perform well in de novo transcriptome assembly, one has to make a decision as to which edge to traverse at these branches; otherwise, the reported contigs will be short and long contigs corresponding to isoforms cannot be constructed. (2) Ideally, each subgraph corresponding to a particular gene can be isolated as a connected component. However, due to repeats, subgraphs corresponding to different genes may merge together and it is difficult to identify correct paths that correspond to isoforms of a gene in the graph. This also represents a major difference

[1] Without splicing, the problem becomes a lot easier. In this paper, we focus on the transcriptome assembly problem for eukaryotes such as mouse and human for which splicing occurs in the majority of the genes.

between solving the de novo transcriptome assembly problem and the problem with a reference genome, because with a reference genome, we can focus on one gene at a time and reconstruct all its isoforms based on the alignment. (3) Isoforms may have different expression levels and it is difficult to identify low-expressed isoforms as the majority of the reads may come from those with relatively higher expression levels. Note that the uneven expression levels are quite different from uneven coverage in genome assembly because the reads from these 'weak' isoforms can have their exons (k-mers) well covered by other isoforms.

In this paper, we tackle the de novo transcriptome assembly problem. To resolve problems arising from repeats in different genes (such as merged subgraphs of different genes and more branches by shared exons), we analyzed the properties of mammalian transcriptomes and observed that not too many genes (less than 1.4%) contain repeat patterns of length greater than 90 bp. This implies that if we can construct a de Bruijn graph using substrings of length 90 in the reads, subgraphs that correspond to different genes are more likely to be isolated. However, the current next-generation sequence technology may not produce such long reads and, even if the technology is available, constructing such a de Bruijn graph directly using such long substrings may suffer from the gap problem. To resolve this problem (Section 2.1), we first build an accumulated de Bruijn graph [17] based on single-end reads up to say 50bp (for reads of length 75bp) which can resolve the gap and repeats problem up to 50bp, and then, extend to 90bp based on pair-end reads. The graph will decompose into many connected components, most of which contain only a single or a few mRNAs. Finally, the branching problem introduced by the shared exons is resolved by a heuristic path finding algorithm (also based on pair-end information of reads) to generate all possible isoforms and the most possible will be output according to heuristic depth-first search (Section 2.2).

We still cannot reconstruct the isoforms with low expression levels (i.e. mRNAs with low coverage of reads). However, based on theoretical analysis (Section 2.3), we can guarantee that most mRNAs can be reconstructed by our T-IDBA software as long as their coverage exceeds a certain threshold. We have implemented T-IDBA and evaluated its performance on both simulated and real data which match well with our theoretical analysis. The results show that T-IDBA outperforms other de novo transcriptome assembly approaches substantially.

2 Method

Different from genome assembly whose input reads are sampled from a species genome, the input reads of mammalian transcriptome assembly are sampled from the (expressed) mRNAs of a mammal. As the total length of genes is much shorter than the genome, at first glance, mammalian transcriptome assembly problem seems easier than the genome assembly problem. However, because of alternative splicing, some long patterns representing exons may occur in multiple mRNAs (isoforms) from the same gene. Thus, the de Bruijn graph (i.e. the graph with each vertex representing a k-mer and an edge from u to v if u and v adjacently occur in a read) has more branches when constructed for the mammalian transcriptome assembly problem than those for the genome assembly problem. Therefore, traditional de novo genome assemblers [11] would not work well for the mammalian transcriptome assembly problem as they would usually stop at branches resulting in very short contigs which represent only part of the exons instead of the whole isoform.

≥ k	30	40	50	60	70	80	90	100	110	120
# of genes	5384	3528	2005	996	620	448	367	294	247	233

Table 1. The number of genes in mouse containing a repeated pattern with length ≥ k as some other genes.

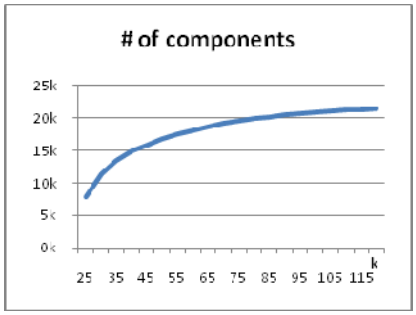

Figure 1. The number of components in de Bruijn graph with different k value.

Besides, many de novo genomic assemblers [10, 11] remove both ends of a contig to increase the accuracy. As the mRNAs are usually relatively short, e.g. 500~5000nt, the lengths of contigs decrease significantly.

In order to reconstruct the isoforms of different genes, we developed T-IDBA which first divides the de Bruijn graph into many connected components, most of which represent isoforms from a single gene. Then T-IDBA determines each isoform from each component using pair-end information.

2.1 Constructing Connected Components

We observe that, although the repeat patterns in the whole genome can be very long, the number of genes that contain the same long repeated patterns is actually quite few. Table 1 shows the number of genes of mouse[2] having repeated patterns of length at least 30. In particular, there are only 367 genes, out of 16,491 genes, containing repeated patterns of length at least 90bp. If we can construct a de Brujin graph with large k, most connected components should contain isoforms from single genes.

We have also built de Bruijn graphs of the reference mRNA of the mouse (UCSC: mm9, NCBI build 37) for different values of k. Figure 1 shows the number of connected components. The number of connected components increases when k increases and there are 20,457 connected components for the de Bruijn graph with $k = 90$. As there are 46,104 mRNAs in the mouse database, each component contains on average 2 mRNAs. Table 2 shows the distribution of the numbers of mRNAs in components. About 91% of mRNAs are in components containing no more than 10 mRNAs, with the majority containing only one mRNA.

[2] Data obtained from EMBL-EBI (http://www.ebi.ac.uk/astd/main.html).

# of mRNAs	1	2	3	4	5	6	7	8	9	10	>10
# of components	10113	4684	2483	1380	748	434	249	117	80	48	121

Table 2. The distribution of mRNAs in components

k	30	40	50	60	70	80	90	100	110	120
# of mRNAs	5339	3336	1922	1234	819	588	472	369	321	262

Table 3. The number of mRNAs with length-*k* repeats.

Algorithm 1. T-IDBA algorithm

1. Apply IDBA on input reads from $k = k_{min}$ to k_{mod} to get a de Bruijn graph *G*.
2. Align pair-end reads to *G* and find connections *C*(*x*, *y*) between nodes with support of at least α pair-end reads. *C* = { *(x,y)* | there are at least α pairs of reads connecting *x* and *y*} (The default value of α is 5.)
3. For each connection *C*(*x*, *y*) if there exists a unique path *p* connecting *x* and *y* which is consistent with the insert distance, then treat this path as a long read for the next step. *P* = {*p* | *p* is the unique path in *G* connecting *(x, y)* in *C*}
4. Apply IDBA on input reads and *P* with $k = k_{mod}$ to k_{max} to get a de Bruijn graph *G'*.
5. For each component in *G'*, find paths with highest support from pair-end reads.

The current next-generation sequencing technology usually only produces reads of length about 75[3]. Thus, it is impossible to construct a de Bruijn graph with $k = 90$ from single-end reads directly. Even if the reads were long enough, there would be a lot of gaps in the graph when constructed directly. In order to solve this problem, at the first step, T-IDBA applies the IDBA algorithm [17] to construct a de Bruijn graph with $k = k_{mod} < l$, where *l* is the length of the input read. At the second step, T-IDBA aligns pair-end reads (by exact match) to the graph for confident connection between nodes. Each connection of nodes is validated by finding a unique path in the graph connecting the pair of nodes with length matching the insert distance of the pair-end reads (within specified error). Note that, the connection between two nodes will be discarded if the number of paths between them is zero or more than one. Although this problem is NP-hard, since there are only a few loops in the graph when $k_{mod} = 50$, the unique path can be found in practice. Table 3 shows the number of mRNAs with length-*k* repeats. As we can see, less than 5% (1,922 out of 46,104) of mRNAs contain loops (length-*k* repeats) in the graph. All the unique paths for validation are recorded for resolving branches and treated as extra long reads for IDBA to construct de Bruijn graph with $k_{max} \geq k \geq l$.

Note that IDBA needs to be tuned specifically for transcriptome assembly. Tips removal in de Bruijn graph is performed using very short length to avoid removing too many *k*-mers, since transcripts are usually very short. In addition, a bigger threshold *m* is used for filtering those incorrect *k*-mers due to sequencing errors. Since this may filter out some low-coverage *k*-mers, it is unlikely that we can reconstructed mRNAs with low

[3] Although some genome centers can produce longer reads, the majority of them are still working with reads of length 75 or shorter.

converage (i.e. low expression level) and only those well-expressed mRNAs are considered.

2.2 Discovering Isoforms in Connected Components

For each connected component in the de Bruijn graph with $k = k_{max}$, T-IDBA discovers those paths starting from a vertex with zero in-degree to a vertex with zero out-degree with the highest support from pair-end reads. A path is supported by a pair-end reads if the pair-end reads can be aligned (by exact match) to the path with the distance between the aligned positions matching the insert distance of the pair-end reads (with up to 10% error). T-IDBA performs depth-first search from a vertex with zero in-degree to a vertex with zero out-degree in decreasing order of support of the branches. In practice, instead of performing a complete depth-first search, T-IDBA reports at most 3 potential isoforms for each zero in-degree node in each connected component (note that 3 is a parameter set to be set by the user).

2.3 Expected Sensitivity of T-IDBA

Given a length-n mRNA R with t length-w exons, i.e. $n = tw$. If the coverage of pair-end length-l reads on R is c and the error rate of each nucleotide in a read is e, we can evaluate in the following the probability that R can be reconstrcuted by T-IDBA as one contig. Thus, we can conclude that most mRNAs can be reconstructed by T-IDBA with high probability as long as the coverage of the mRNA exceeds a certain threshold.

In order to reconstrcut R, all $k_{\min}$-mers of R must exist in the de Bruijn graph when $k = k_{\min}$, i.e. every $k_{\min}$-mer must be sampled at least m times with no error. Since there are $s = cn/(2l)$ pair of reads sampled from R and the probability that a particular $k_{\min}$-mer contains in a pair of read is $2(l - k_{\min} + 1)/(n - l + 1)$, the probability that a $k_{\min}$-mer of R is sampled j times is

$$\binom{s}{j}\left(\frac{2(l-k_{\min}+1)}{n-l+1}\right)^{j}\left(1-\frac{2(l-k_{\min}+1)}{n-l+1}\right)^{s-j}$$

However, some of the sampled $k_{\min}$-mers may contain error. Since the probability that a $k_{\min}$-mer contains error is $p_{k_{\min}} = 1-(1-e)^{k_{\min}}$, the probability that all $n - k_{\min} + 1$ $k_{\min}$-mers of R exist in the de Bruijn graph when $k = k_{mod}$ is

$$\left[\sum_{j=m}^{s}\binom{s}{j}\left(\frac{2(l-k_{\min}+1)}{n-l+1}\right)^{j}\left(1-\frac{2(l-k_{\min}+1)}{n-l+1}\right)^{s-j}\left(\sum_{q=m}^{j}\binom{j}{q}p_{k_{\min}}^{j-q}\left(1-p_{k_{\min}}\right)^{q}\right)\right]^{n-k_{\min}+1} \quad (1)$$

If R exists in the de Bruijn graph when $k = k_{mod}$, R exists in a connected component if and only if for each pair of adjacent exons, there is no branch (with probability $1 - p_b$) or the branches can be resolved by pair-end reads. Similar to (1), the probability that α pair-end reads connecting two adjacent exons are sampled is

$$\binom{s}{j}\left(\frac{(l-k_{\mathrm{mod}}+1)}{n-l+1}\right)^{j}\left(1-\frac{(l-k_{\mathrm{mod}}+1)}{n-l+1}\right)^{s-j}$$

And the probability that two k_{mod}-mers, one from each one of pair-end reads, are sampled correctly is $p_c = (1 - p_{k_{\mathrm{mod}}}^{l-k_{\mathrm{mod}}+1})^2$, the probability that R exists in a connected component is

$$\left[(1-p_b) + p_b \sum_{j=\alpha}^{s} \binom{s}{j} \left(\frac{(l-k_{\min}+1)}{n-l+1}\right)^{j} \left(1-\frac{(l-k_{\min}+1)}{n-l+1}\right)^{s-j} \left(\sum_{q=\alpha}^{j} \binom{j}{q} p_c^{q} (1-p_c)^{j-q}\right)\right]^{t-1} \quad (2)$$

By multiplying the two probabilities given in (1) and (2), we can calculate the sensitivity of finding an mRNA with t length-m exons and coverage c. For example, given a length-2500 mRNA with 5 length-500 exons and 30X coverage, if the error rate is 1% and the threshold $m = 4$, $\alpha = 5$ the mRNA can be reconstucted with probability 0.69.

3 Experimental Results

3.1 Simulated data

We test our transcriptome assembler T-IDBA on mouse genes. The reference mRNA of all known mouse genes from UCSC (mm9, NCBI build 37) are used to generate the simulated sequencing reads. There are 26,989 genes and 49,409 isoforms in this dataset. About 60% of these genes have only one isoform and 0.2% of these genes have more than 10 isoforms. For the simulation, we first randomly generate the expression level for each isoform and sequencing reads are then sampled uniformly in each mRNA according to expression level. We consider the following three different distributions of expression level to show the performance of T-IDBA without expression level and with the last two distributions to capture the property of real data [18, 19].

(1) Equal: the expression level of each mRNA is set to 1.
(2) Uniform: the expression level of each mRNA is generated according to a uniform distribution in [0,1].
(3) Log Normal: a number r is generated according to a normal distribution $N(0, 1)$ and the expression level of each mRNA is set to e^r.

The sequencing reads are sampled with read length = 75, error rate = 1%, insert distance = 250. Based on expression levels, the number of reads of each mRNA is calculated by setting the total number of reads to 78M (about 50x depth on average).

The sensitivity and precision of T-IDBA and Abyss [11] are compared for the three simulated datasets of different distributions. Since only those isoforms above a certain expression level can be reconstructed from the reads, only the mRNAs with sequencing depths larger than 30x are considered for sensitivity evaluation.

For T-IDBA, we evaluate the performances of the output at three stages of T-IDBA to show the effect of each stages.

I. Single-end stage: contigs of the graph G at $k = k_{\mathrm{mod}}$ (Step 1 of Algorithm 1, without using pair-end information).
II. Pair-end stage: contigs of the graph G' at $k = k_{max}$ (Step 4 of Algorithm 1, using pair-end information to extend the de Bruijn graph).
III. Full stage: final results from T-IDBA.

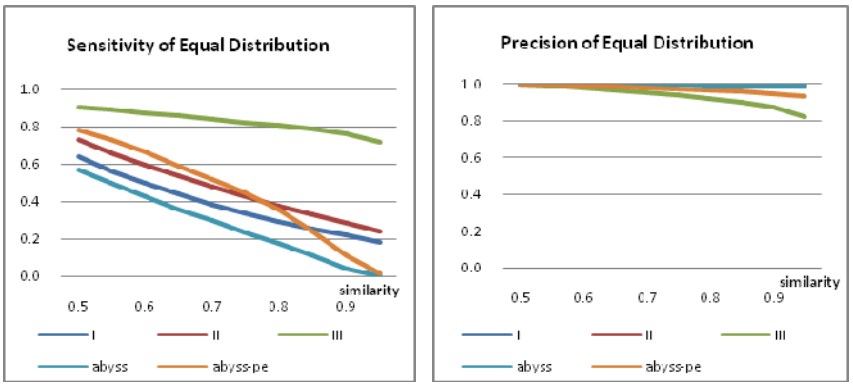

Figure 2. Experimental results of equal distribution dataset

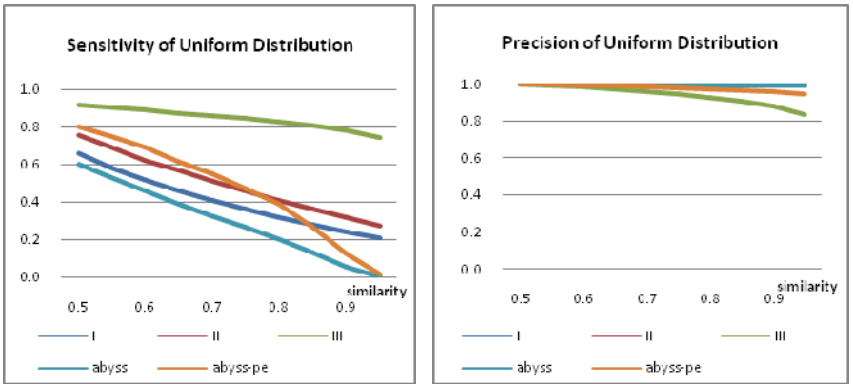

Figure 3. Experimental results of uniform distribution dataset

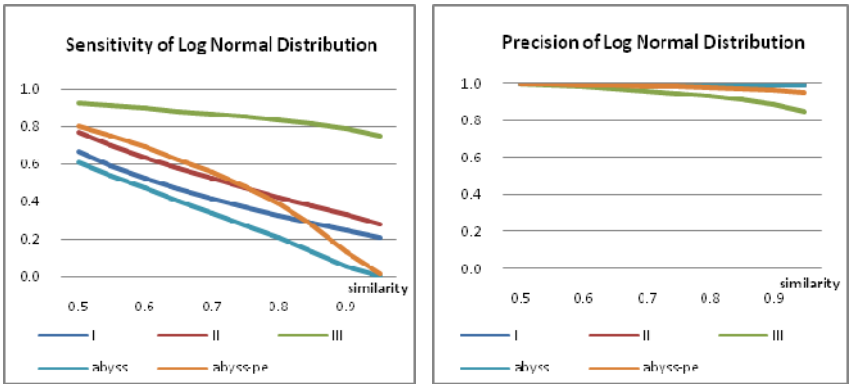

Figure 4. Experimental results of log normal distribution dataset

An isoform is said to be found if a contig can be aligned to the isoform with similarity exceeding a threshold. We compare sensitivity and precision under different levels of similarity. Similarly, a contig is considered as correct if it can be globally aligned to some part of a reference mRNA with the given similarity.

Note that the total number of correctly reconstructed mRNAs may not be the same as the total number of correct contigs, because an mRNAs may be separated into more than one contigs in the graph due to the gap or repeat problem. In this case, the contigs might be correct, while the corresponding mRNA is not considered as reconstructed.

For T-IDBA, k_{min}, k_{mod}, k_{max} are set to 25, 50, 90 repectively, while the value of k is set to 50 for Abyss. Figures 2, 3 and 4 show sensitivity and precision under different similarity settings. In all cases, the sensitivities of all algorithms drop when the similarity threshold increases, because higher similarity requires the algorithms to reconstruct a larger portion of the transcripts. Using traditional de novo assembly method, the repeats in different isoforms of the same gene are very difficult to resolve. Even with pair-end information, only a small portion of the isoforms can be found correctly. The reason is that if two contigs are supported by many pairs of reads, they are usually merged to form a large contig by the genome assembler. But in the case of transcriptome assembly, it is insufficient to merge them together directly, because two contigs can be connected in more than one way.

In Figure 2, T-IDBA has the highest sensitivity, especially when the similarity is more than 80%. For 95% similarity, only 0.43% and 1.43% mRNAs can be found in the single-end and pair-end versions of Abyss, while the 3 stages of T-IDBA (single-end, pair-end and full version) have sensitivity values of 18.54%, 24.08% and 72.10% respectively, which demonstrates the effectiveness of each stage of T-IDBA. Note that there should be 10113 components in the error-free de Bruijn graph with $k = 90$. Using pair-end information to increase the k value to 90, T-IDBA can find 11100 components from simulated reads, which is more than expected number of components because of the gap problem which breaks some components into more than one component. Path finding stage can further reconstruct isoforms from the same gene and improves the sensitivity to 72.10% with 33241 out 46409 mRNAs reconstructed from scratch. Figures 3 and 4 show a similar trends for different distribution assumptions. There are 32130 and 23123 expressed isoforms in uniform and log normal distribution data set. T-IDBA can reconstruct 74.47% and 75.05% of them for 95% similarity respectively, while those values of Abyss are 1.78% and 1.89 % respectively.

The precision performance of the algorithms on the three datasets have similar trend (Figures 2, 3 and 4). The single-end version of Abyss and the single-end and pair-end stages for T-IDBA all have nearly 99% precision as all contigs are correct but short. The pair-end Abyss and T-IDBA introduce a small number of errors by connecting incorrect contigs together. After applying path finding algorithm at the final stage, the precision of T-IDBA further drops to about 83%, because of some wrong combinations of contigs. When compared to IsoInfer [7], which makes use of a reference genome and additional annotated information, the performance of the full stage of T-IDBA is similar in terms of sensitivity and precision even without a reference genome. (IsoInfer's sensitivity and precision are around 77.4% and 81.3%, respectively for simulated log normal distribution data on the mouse.)

3.2 Real data

The RNA sequencing reads (152 millions 76-base pair-end reads)of embryonic stem cells in [9] are used to evaluate our assembly algorithm. We applied our assembler and Abyss on this dataset using the same parameters as for the simulated data. Since only mRNAs with coverage depth more than 30 can be reconstructed, we aligned the reads using BLAT [20] to the mRNA reference database and found that there are 2,835

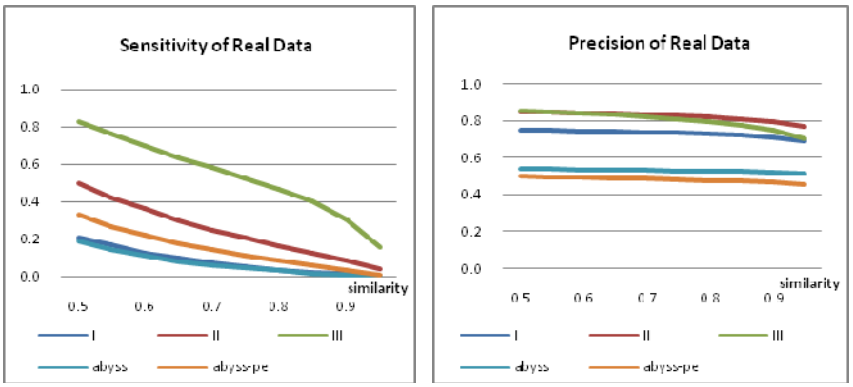

Figure 5. Experimental results of real data.

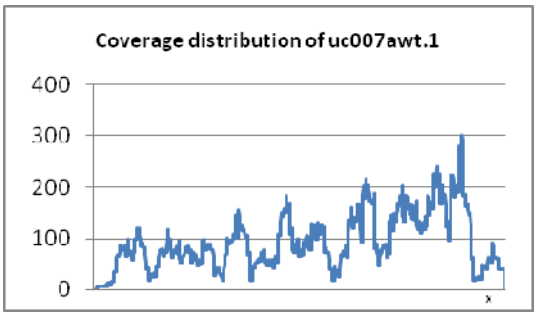

Figure 6. The coverage distribution of mRNA uc007awt.t in real data.

mRNAs[4] with coverage higher than 30 and at least 80% of the region reconstructed by more than 4 reads.

For real data, the performance of both T-IDBA and Abyss drops (Figure 5). It may be caused by the noise which is not well understood (e.g. from intron regions of the gene) or the log normal distribution of reads which may not be able to capture the real property of the transcriptome data. In all cases, T-IDBA outperforms Abyss as more isoforms can be reconstructed by T-IDBA. When similarity is set to 80%, the sensitivity and precision of T-IDBA are 46.7% and 79.7%, compared with only 8.6% and 47.9%, respectively, for Abyss.

There might be another explanation for the poorer performance for real data. Some parts of an exon are not reconstructed by enough reads thus the corresponding mRNA will be reconstructed by more than one contig. The non-uniform distribution of reads within one mRNA will also cause problems in T-IDBA as shown in Figure 6. Some parts of the mRNAs have very low coverage. It is unlikely that we can reconstruct these mRNAs using only one contig. If two contigs are allowed to cover one mRNA, the sensitivity of T-IDBA and Abyss can be increased to 69% and 19%, respectively, and this matches with our mathematical analysis as given in Section 2.3. Compared with Scripture [9],

[4] In [9], they showed that 15,352 known genes are found and 13,362 of them are significant expressed. The difference between these figures and ours is because they assume that a gene is found as long as there are enough reads covering the junctions (exon boundaries) of the gene instead of requiring the whole mRNA to be covered with enough coverage while in our case, we want to recover the whole mRNA sequence, thus our coverage requirement is higher.

which uses reference genomes and can reconstruct 78% of the expressed isoforms, the performance of T-IDBA looks reasonably good.

4 Conclusions

We observed that the de Bruijn graph of transcriptomes can be decomposed into small connected components if k is large enough. Our T-IDBA algorithm, which captures the merits of all k values in between k_{min} and k_{max} with pair-end information, can decomposite the graph generated from the transcriptome sequencing reads into many connected components, each of which contains very few mRNAs. Since most of the isoforms from the different genes will fall into different components, the isoform reconstruction becomes easier for each component. A heuristic-based isoform finding algorithm, based on maximizing the number of pair-end support, is used to generate the most possible isoforms. The performance of T-IDBA outperforms Abyss for both simulated and real data and matched with theoretical analysis. However, the performance of T-IDBA for real data is not as good as that for simulated data (although it is still a lot better than that of Abyss) due to the non-uniform read distribution in an mRNA. Further analysis on real data should be performed to build better model of the error and expression level distribution so as to have a more robust and accurate de novo transcriptome assembler.

Acknowledgments

We would like to acknowledge Liu Zhihua for his help on programming T-IDBA and Dr. M.Y. Chan for improving the readability of this paper.

References

1. Graveley BR: Molecular biology: power sequencing. *Nature* 2008, 453(7199):1197-1198.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008, 320(5881):1344-1349.
3. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9):1105-1111.
4. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A *et al*: Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A* 2009, 106(9):3264-3269.
5. Jiang H, Wong WH: Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 2009, 25(8):1026-1032.
6. Lacroix V, Sammeth M, Guigo R, Bergeron A: Exact Transcriptome Reconstruction from Short Sequence Reads. In: *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*. Karlsruhe, Germany: Springer-Verlag; 2008: 50-63.
7. Feng J, Li W, Jiang T: Inference of Isoforms from Short Sequence Reads (Extended Abstract). In: *RECOMB: 2010; Lisbon*; 2010.
8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010, 28(5):511-515.

9. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 2010, 28(5):503-510.
10. Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008, 18(5):821-829.
11. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009, 19(6):1117-1123.
12. Pevzner PA, Tang H, Waterman MS: An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 2001, 98(17):9748-9753.
13. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 2008, 18(5):802-809.
14. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE *et al*: De novo transcriptome assembly with ABySS. *Bioinformatics* 2009, 25(21):2872-2877.
15. Jackson BG, Schnable PS, Aluru S: Parallel short sequence assembly of transcriptomes. *BMC Bioinformatics* 2009, 10 Suppl 1:S14.
16. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA: Splicing graphs and EST assembly problem. *Bioinformatics* 2002, 18 Suppl 1:S181-188.
17. Yu P, Henry L, Yiu SM, Francis YLC: IDBA- A Practical Iterative de Bruijn Graph De Novo Assembler. In: *RECOMB: 2010; Lisbon*; 2010.
18. Alter MD, Rubin DB, Ramsey K, Halpern R, Stephan DA, Abbott LF, Hen R: Variation in the large-scale organization of gene expression levels in the hippocampus relates to stable epigenetic variability in behavior. *PLoS One* 2008, 3(10):e3344.
19. Konishi T: Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics* 2004, 5:5.
20. Kent WJ: BLAT--the BLAST-like alignment tool. *Genome Res* 2002, 12(4):656-664.