Analysis of the Metatranscriptome of Microbial Communities by Comparison of Different Assembly Tools Reveals Improved Functional Annotation

Ruchi Rani and Chandan Badapanda*

Bioinformatics Division, Xcelris Labs Limited, India

Submission: October 09, 2017; Published: October 31, 2017

*Corresponding author: Dr. Chandan Badapanda, Bioinformatics Division, Xcelris Labs Limited, India, Tel-91-79-66092177.
Email: chandan.badapanda@xcelrislabs.com

Abstract
Assembling metatranscriptomic or metagenomic data is a challenging task considering huge amount of short read data generated through Next Generation Sequencing (NGS) platforms. Metagenomic assembly involves new computational challenges due to the uneven read coverage of bacterial strains present in the sample, similarity between different species and dissimilarities between closely related strains of the same bacteria. During recent times, a large diversity of specialized software tools are available for metagenomic or metatranscriptomic assembly. Nevertheless, choosing the most appropriate assembly methods can be rather challenging. Thus, we have chosen four highly cited metagenomic or metatranscriptomic assembly tools i.e. IDBA-UD, MetaSPAdes, MEGAHIT and CLC Genomics Workbench for this study. The validation of the assembly was performed on various parameters such as percentage of reads that were participated in the assembly, length distribution of scaffolds, assembly size and N50 Value. Further, taxonomic assignment was achieved through Kaiju and functional annotation of genes were executed through Cognizer. Based on the sensitivity of all the four assemblers towards the assembly size, length distribution, percentage of annotated genes obtained through Kaiju and Cognizer, the assembler tool MetaSPAdes outperforms the other assembly tools.

Introduction
Before the advent of Next Generation Sequencing (NGS) technology, data generation of uncultured species along with the analysis of microbial data was limited. Advancement in the sequencing technology has revolutionized the sequencing of individual genome as well as metagenome. NGS technology coupled with the development of algorithm for analysis of NGS data have increased our understanding of microbial community structure [1,2]. In metagenomic study, the information of all genes are used to interpret microbial identities up to the species or strain level [3] whereas, metatranscriptomic study reveals the gene expression patterns of active genes and their functionality in different pathways [4,5]. In both the pipeline (metagenomic and metatranscriptomic), it is important to assemble the reads into contigs which represents gene objects. However, there are various challenges associated with the assembly of metatranscriptome and metagenome data, which is addressed below:

a. Huge amount of data is being generated by NGS technology which are of short reads length, making it difficult to assemble [1,6].

b. The wide range of genomes present within a sample making it complicated to assemble [1].

c. Similarity between different species as they share highly conserved regions and also the dissimilarity between closely related strains of the same bacterial species, further make the assembly of metagenomic sample more complicated.

d. Bacterial strains present in the metagenome are considerably of uneven read coverage, results in fragmented assembly [7].

To overcome these problems associated with metagenomic assembly, two major approaches are commonly used i.e; Overlap layout consensus (OLC) and the de Bruijin graph approach. Both these methods use a data structure called a “graph” to represent all connections (edges) between all basic sequence elements, e.g. reads [5,6]. OLC approaches are highly suited for the assembly of long sequencing reads whereas the de Bruijin graph approach is good for assembly of short reads. However, de Bruijin graph approach is more erroneous over Overlap layout consensus (OLC). To remove error in assemblies, assemblers use a number of speculative approaches [1].

Here, in this study we have compared four highly cited metagenomic or metatranscriptomic assembly tools i.e; IDBA-UD, MetaSPAdes, Megahit and CLC Genomic work bench de-
signed for handling high throughput short read sequencing data. This publication intends to figure out the best assembler based on the challenges which are associated with metagenome or metatranscriptomic data.

**IDBA-UD**

IDBA (Iterative de Bruijn Graph De Novo Assembler) is a suite of different de Bruijn graph based assemblers, each dedicated for a specific task. There are two main module in IDBA: IDBA-UD and Meta-IDBA, which are used for assembly of metagenome and metatranscriptome. However, IDBA-UD performs better than Meta-IDBA. IDBA specifically designed to handle data with highly uneven sequencing depths. IDBA is comparatively memory and cost efficient assembler. IDBA-UD achieved its best performance by iterating k-mer from 20 to 100 [8].

**MetaSPAdes**

metaSPAdes first constructs the de Bruijn graph of all reads using SPAdes. MetaSPAdes has efficient assembly graph processing to address the micro-diversity challenges. For the assembly, SPAdes utilizes an iterative multi-k-mer approach similar to ID-BA-UD. The range of k-mer is from 21 to 128 bp. SPAdes and metaSPAdes accept a wide range of data types and formats in both compressed and uncompressed form [7].

**MEGAHIT**

MEGAHIT is based on succinct de Bruijn graph which is a memory efficient assembler. MEGAHIT uses a range of k-mer values; length is set between 15 to 127. There are many optional parameters present that may be chosen based on the requirement. It accepts single as well as paired-end reads in compressed and uncompressed fasta or fastq format. Multiple computational threads can be specified and optionally a graphical processing unit can be employed to increase computational power [9].

**CLC Assembler**

CLC Genomics Workbench software package (CLC bio) is a Qiagen package with a graphical user interface (GUI), which is commercial software. CLC is an integrated software package which can be used for a number of functions in genomics as well as proteomics. CLC bio also uses de Bruijn graph-based approach for assembly.

**Material and Methodology**

A total of 3GB of metatranscriptomic data was generated in house and sequencing was performed using 2*150 bp chemistry on Illumina platform. Quality of data was checked using FastQC (v0.11.5) [10] tool. Data was filtered for adapter sequences, low quality reads, bases having quality score ≤20 using Trimmomatic (v0.36) [11] to get a high quality data. The reads smaller than 40 nucleotides were discarded and remaining reads were subsequently used for assembly [12]. Assembly was performed using four different metagenomic assembler tools (metaSPAdes from SPAdes v3.10 [7], IDBA-UD v1.1.1 [8], CLC Genomics Workbench.
Anatomy Physiology & biochemistry international journal (APBIJ)

v6.0 [https://www.qiagenbioinformatics.com/] and MEGAHIT v1.0.5 [9]) at their default parameters. Genes were called from scaffolds obtained from the four metagenomic assembler using prodigal v2.6.3 [13]. Taxonomy were assigned to all predicted genes using kaiju web server (http://kaiju.binf.ku.dk/server). Function assignment of predicted genes was done using cognizer v0.9b [14]. Cognizer is a comprehensive annotation tool for metagenomic or metatranscriptomic dataset which perform homology search against five databases such as GO, KEGG, PFAM, COG and FIG. Recently our group has published the metatranscriptomic data analysis pipeline using kaiju and cognizer [12].

Table 1: List of four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) used for evaluation and estimation of metatranscriptomic dataset.

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Version</th>
<th>Release Date</th>
<th>Method</th>
<th>k-mer</th>
<th>Licence</th>
</tr>
</thead>
<tbody>
<tr>
<td>metaSPAdes</td>
<td>3.1</td>
<td>2016</td>
<td>de Bruijn graph approach</td>
<td>21-128</td>
<td>Open Access</td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>1.1.1</td>
<td>2012</td>
<td>de Bruijn graph approach</td>
<td>20-100</td>
<td>Open Access</td>
</tr>
<tr>
<td>MEGAHIT</td>
<td>1.0.5</td>
<td>2015</td>
<td>de Bruijn graph approach</td>
<td>15 - 127</td>
<td>Open Access</td>
</tr>
<tr>
<td>CLC Genomics Workbench</td>
<td>6</td>
<td>2008</td>
<td>de Bruijn graph approach</td>
<td>GUI</td>
<td>Commercial</td>
</tr>
</tbody>
</table>

Assembly Parameters:

All the assemblers used in this study, use de Bruijn graph approaches for assembly [6]. These assemblers use a defined k-mers length or k-mers are detected within the reads (Table 1). IDBA-UD was launched with read error-correction enabled as recommended in the manual for metagenomic data analysis [7]. CLC default parameters (Minimum contig length: 200, Automatic word size: Yes, Perform scaffolding: Yes, Mismatch cost: 2, Insertion cost: 3, Deletion cost: 3, Length fraction: 0.5, Similarity fraction: 0.8). MEGAHIT and metaSPAdes were used on their default setting.

Table 2: Assembly Statistics of four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) is provided here. From the table, it can be inferred that metaSPAdes has highest number of scaffolds with maximum scaffold length as compared to other three assemblers.

<table>
<thead>
<tr>
<th>Assembly Elements</th>
<th>MetapSPAdes</th>
<th>IDBA-UD</th>
<th>CLC Genomics Workbench</th>
<th>MEGAHIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of scaffolds</td>
<td>213685</td>
<td>20692</td>
<td>122289</td>
<td>40994</td>
</tr>
<tr>
<td>Total Scaffold length including gaps (in bp)</td>
<td>58123635</td>
<td>7925976</td>
<td>35244547</td>
<td>15775841</td>
</tr>
<tr>
<td>Average scaffold size (in bp)</td>
<td>271.95</td>
<td>383.04</td>
<td>284.73</td>
<td>384.83</td>
</tr>
<tr>
<td>Scaffold N50</td>
<td>255</td>
<td>370</td>
<td>282</td>
<td>366</td>
</tr>
<tr>
<td>Maximum scaffold size (in bp)</td>
<td>10570</td>
<td>23944</td>
<td>13683</td>
<td>24045</td>
</tr>
<tr>
<td>Minimum scaffold size (in bp)</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Readmapback (%)</td>
<td>74.79%</td>
<td>79.67%</td>
<td>44.57%</td>
<td>53.84%</td>
</tr>
</tbody>
</table>

Benchmarking

Scaffold Length Statistics:

In total 213685, 20692, 122289 and 40994 scaffolds were obtained with N50 value of 255, 370, 282 and 366 through metaSPAdes assembler, IDBA-UD assembler, CLC Genomic Workbench and MEGAHIT assembler respectively. Highest number of scaffolds were obtained through metaSPAdes assembler followed by CLC Genomic Workbench, whereas least number of scaffolds were obtained by IDBA-UD. The N50 value for metaSPAdes assembly is the highest among the four assemblers. The total size of assembly through metaSPAdes is highest, i.e; 58 Mb followed by CLC (35Mb) and MEGAHIT (15Mb) where-as only 7 Mb size of assembly obtained through IDBA-UD. The statistical elements of the assembly were calculated by in house perl script which is listed in Table 2 for all four assembler. The four assembler were further compared on the basis of scaffold length distribution. The scaffolds length distribution of assembly is represented in Figure 2 using in house shell script. Considering the above, metaSPAdes assembly was found to have comparable number of scaffolds, scaffold length distribution along with comparable assembly size as compared to the other three assemblers. The assembly obtained from the four assemblers were validated based on the percentage of read mapping back, gene calling and length distribution, taxonomic classification and functional annotations.

Figure 1 represents the bioinformatics pipeline implemented in this study.

Results and Discussion

Metatranscriptome Assembly:

After quality filtration, a total of 3Gb high quality reads were obtained and were given as input in all the four assembler i.e; metaSPAdes from SPAdes v3.10 [7], IDBA-UD v1.1.1 [8], CLC Genomics Workbench v6.0 [https://www.qiagenbioinformatics.com/], MEGAHIT v1.0.5 [9].

Table 2: Assembly Statistics of four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) is provided here. From the table, it can be inferred that metaSPAdes has highest number of scaffolds with maximum scaffold length as compared to other three assemblers.
How to cite this article: Ruchi R, Chandan B. Analysis of the Metatranscriptome of Microbial Communities by Comparison of Different Assembly Tools Reveals Improved Functional Annotation. Anatomy Physiol Biochem Int J. 2017; 3(4): 555618. DOI: 10.19080/APBIJ.2017.03.555618.

Figure 2: Assembly length distribution of all the four assemblies generated from metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT is represented in X-axis and their respective counts were provided in Y-axis. From this figure, it can be inferred that the highest number of scaffolds were predicted from metaSPAdes as compared to other three assemblers.

Percentage of read Mapping Back

High quality reads were mapped back to each assembly, to identify how much participated in constituting the assembly. The highest read mapping was obtained for IDBA-UD assembly (79.67%) followed by metaSPAdes assembly (74.79%), whereas least mapping was obtained from CLC Genomics Workbench (44.57%) followed by MEGAHIT (53.84%).

Table 3: Gene Statistics of the four assembler (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) is provided here. From the table, it can be inferred that highest number of predicted genes were predicted obtained from metaSPAdes.

<table>
<thead>
<tr>
<th>Statistical elements</th>
<th>MetaSPAdes</th>
<th>IDBA-UD</th>
<th>CLC</th>
<th>MEGAHIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of genes</td>
<td>172324</td>
<td>16770</td>
<td>102111</td>
<td>37517</td>
</tr>
<tr>
<td>Total gene length including gaps (in bp)</td>
<td>45606927</td>
<td>5886909</td>
<td>28213389</td>
<td>13196172</td>
</tr>
<tr>
<td>Average gene size (in bp)</td>
<td>264.65</td>
<td>351.03</td>
<td>276.11</td>
<td>351.73</td>
</tr>
<tr>
<td>gene N50</td>
<td>249</td>
<td>348</td>
<td>270</td>
<td>345</td>
</tr>
<tr>
<td>Maximum gene size (in bp)</td>
<td>2268</td>
<td>2670</td>
<td>2670</td>
<td>2670</td>
</tr>
<tr>
<td>Minimum gene size (in bp)</td>
<td>201</td>
<td>201</td>
<td>201</td>
<td>201</td>
</tr>
</tbody>
</table>

Gene Calling and Length Distribution

Genes from each assembly were predicted using Prodigal [13]. A total of 172324, 16770, 102111 and 37517 genes were predicted from the assembly of MetaSPAdes, IDBA-UD, CLC Genomics and MEGAHIT respectively. Table 3 represents the total number of genes predicted in each assembly, their size with N50 values. Figure 3 represents the length distributions of genes predicted from the four assemblers. The number of genes and their length distribution was found to be better in MetaSPAdes than in other three assemblers.
How to cite this article: Ruchi R, Chandan B. Analysis of the MetaTranscriptome of Microbial Communities by Comparison of Different Assembly Tools Reveals Improved Functional Annotation. Anatomy Physiol Biochem Int J. 2017; 3(4): 555618. DOI: 10.19080/APBIJ.2017.03.555618.

**Figure 3:** Gene length distribution from all the four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) is provided in this figure (Length distribution in X-axis and their counts in Y-axis). From this figure, it can be inferred that the highest number of genes were predicted from metaSPAdes.

**Figure 4:** Distribution of top 15 phyla from all the four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) and taxonomy was assigned through Kaiju is represented here. From this figure, Proteobacteria is the most abundant phylum in all the assemblers whereas maximum number of taxonomic assignment was obtained from genes generated through metaSPAdes assembler.

**Taxonomy Classification**

Kaiju web server was used to classify individual metatranscriptomic genes using a reference database comprising of microbial subset of the non-redundant protein nr database as used in case of NCBI BLAST. Kaiju classified 125154 genes out of 172312 genes assembled through metaSPAdes, 13070 genes out of 16770 genes from IDBA-UD assembly, 75304 genes out of 101056 genes from CLC assembly and 31190 genes out of 37517 genes of MEGAHIT assembly. The sample was enriched with bacteria followed by archaea, eukaryota, virus and unclas-
sified microbiota at domain level. At phylum level, Proteobacteria was found to be most abundant group in all the assemblies. In total, 122, 59, 68 and 101 different phyla were obtained in metaSPAdes assembler, IDBA-UD assembler, MEGAHIT assembler and CLC Genomic Workbench respectively. From the result, it can be inferred that highest number of microbial diversity assignment was obtained which was derived from metaSPAdes assembly. The distribution of top 15 phyla was plotted in Figure 4.

### Functional Annotation

Predicted genes were given as input in Cognizer [14] for their functional annotation. A total of 127952, 13583, 75005 and 36774 genes hit obtained against GO database; 100220, 10266, 58800 and 27332 genes hit against KEGG database; 90389, 9672, 53727 and 25000 genes hit against Pfam database; 87283, 9272, 51774 and 24149 got hit against COG database; 549156, 5297, 29025 and 13768 genes hit against FIG database for metaSPAdes assembler, IDBA-UD assembler, CLC Genomic Workbench and MEGAHIT assembler. Figure 5 represents the functional annotation of genes from metaSPAdes, IDBA-UD, MEGAHIT and CLC Genomics annotated against GO, KEGG, Pfam, COG and FIG database.

### Conclusion

The present study was carried out to evaluate and benchmark the four most cited metagenomic or metatranscriptomic assembly tools i.e. MetaSPAdes, IDBA-UD, Megahit and CLC Genomic workbench which work on de Bruijn graph based approach. The assembly output generated from MetaSPAdes was significantly improved in terms of assembly length distribution, assembly size, percentage of read mapping back and micro-diversity represented by the genes as compared to IDBA-UD, Megahit and CLC Genomic workbench. In conclusion, on the basis of sensitivity of the assembler towards assembly size, length distribution and capturing high micro-diversity, metaSPAdes is the best choice.

### Acknowledgement

We thank Xcelris Labs Limited for their encouragement to carry out this project.

### References

7. Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, Pavel A Pevzner
How to cite this article: Ruchi R, Chandan B. Analysis of the Metatranscriptome of Microbial Communities by Comparison of Different Assembly Tools Reveals Improved Functional Annotation. Anatomy Physiol Biochem Int J. 2017; 3(4): 555-618. DOI: 10.19080/APBIJ.2017.03.555618.


Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission
https://juniperpublishers.com/online-submission.php

This work is licensed under Creative Commons Attribution 4.0 License
DOI: 10.19080/APBIJ.2017.03.555618