Anatomy Physiology & Biochemistry
International Journal
ISSN: 2476-1400

Juniper PUBLISHERS
key to the Researchers

Research article
Volume 3 Issue 4 - October 2017
DOI: 10.19080/APBIJ.2017.03.555618

Anatomy Physiol Biochem Int J
Copyright © All rights are reserved by Dr. Chandan Badapanda

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 complicates 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 metagenomicic 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 metagenomtic or metatranscriptomic assembly tools i.e; IDBA-UD, MetaSPAdes, Megahit and CLC Genomic work bench designed 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 Bruijin 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-IDBI. 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 IDBA-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

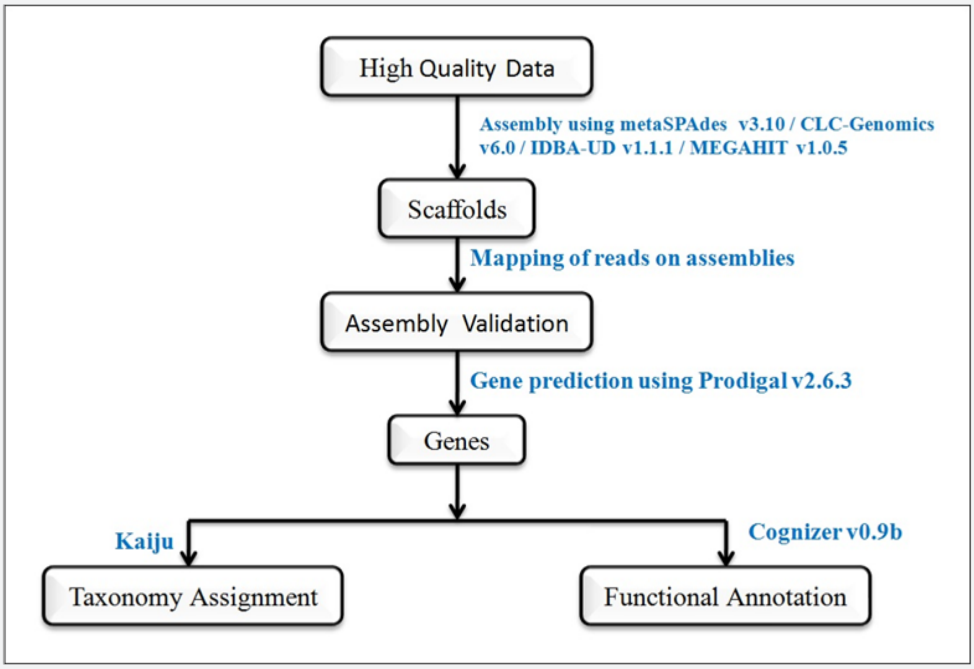
High Quality Data
Assembly using metaSPAdes v3.10 / CLC-Genomics v6.0 / IDBA-UD v1.1.1 / MEGAHIT v1.0.5
Scaffolds
Mapping of reads on assemblies
Assembly Validation
Gene prediction using Prodigal v2.6.3
Genes
Kaiju
Cognizer v0.9b
Taxonomy Assignment
Functional Annotation

Figure 1: Bioinformatics workflow for metatrascriptome assembly and downstream analysis is presented here. High quality reads were given as input to the four assemblers i.e. metaSPAdes v3.10, CLC-Genomics v6.0, IDBA-UD v1.1.1 and MEGAHIT v1.0.5 and then high quality reads were mapped back to each assembly to validate the assembly. Further, genes were predicted from each assembly using Prodigal v2.6.3 and the predicted genes were given as input to kaiju and cognizer for taxonomy and functional assignment.

A total of 3GB of metatranscriptomic data was generated in house and sequencing was performed using 2*150bp 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 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]. 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 1: List of four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) used for evaluation and estimation of metatranscriptomic dataset.

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

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.

Assembly Elements	MetaSPAdes	IDBA-UD	CLC	Megahit
No. of scaffolds	213685	20692	122289	40994
Total Scaffold length including gaps (in bp)	58123635	7925976	35244547	15775841
Average scaffold size (in bp)	271.95	383.04	284.73	384.83
Scaffold N50	255	370	282	366
Maximum scaffold size (in bp)	10570	23944	13683	24045
Minimum scaffold size (in bp)	200	200	200	200
Readmapback (%)	74.79%	79.67%	44.57%	53.84%

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 assemblies. The total size of assembly through metaSPAdes is highest, i.e; 58 Mb followed by CLC (35Mb) and MEGAHIT (15Mb) whereas 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.

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.

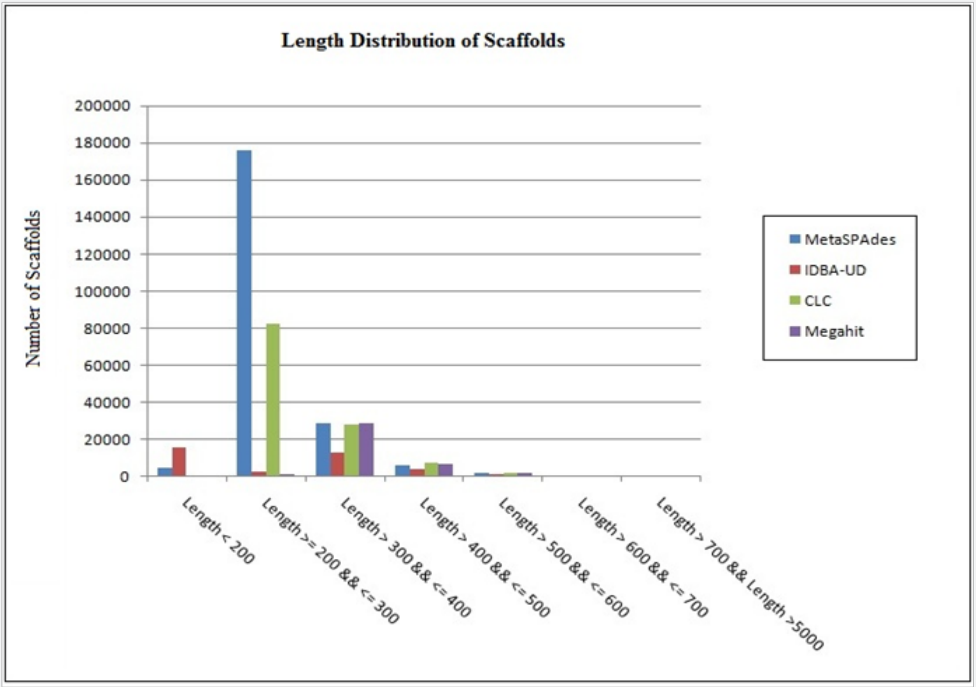

Figure 2: Assembly length distribution of all the four assemblies generated from metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT is represented in X-axis and their respective counts were provided in Y-axis. From this figure, it can be inferred that highest number of scaffolds were predicted from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

Percentage of read Mapping Back

High quality reads were mapped back to each assembly, to identify how much participated in constituting the assembly. 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 highested number of predicted genes were predicted obtained from metaSPAdes.

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

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.

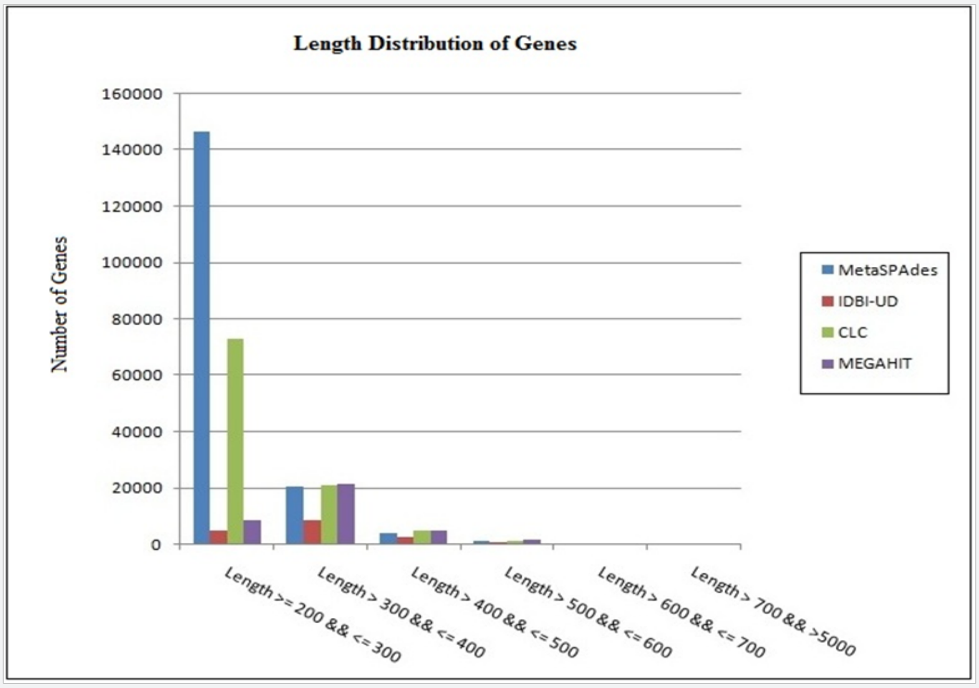

Figure 3: Gene length distribution from all the four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) is provided in this figure (Length distribution in X-axis and their counts in Y-axis). From this figure, it can be inferred that highest number of genes were predicted from metaSPAdes.

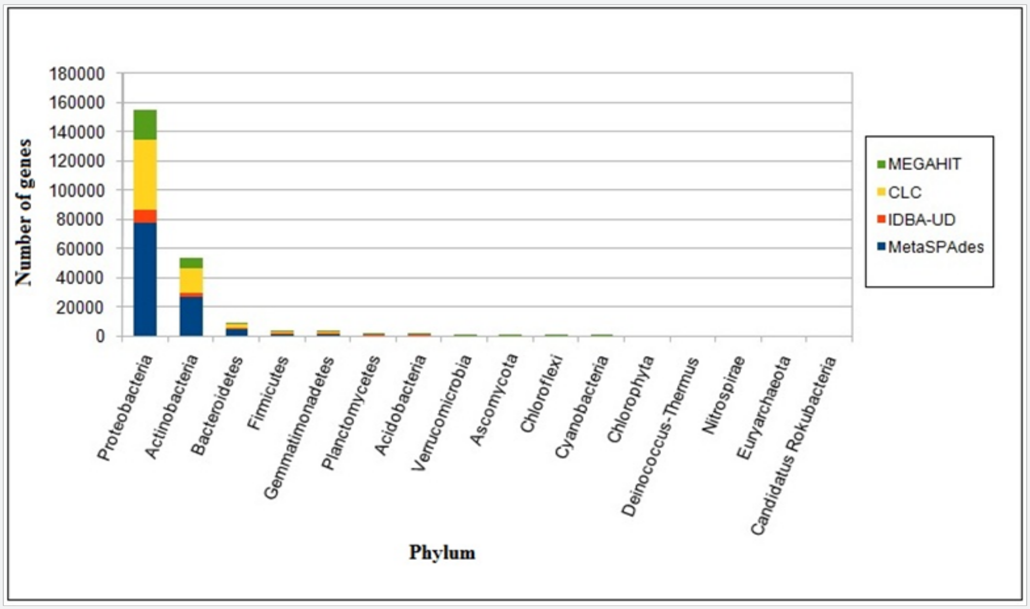

Figure 4: Distribution of top 15 phyla from all the four assemblers (metaSPAdes, IDBA-UD, CLC genomics workbench and MEGAHIT) and taxonomy was assigned through Kaiju is represented here. From this figure, Proteobacteria is the most abundant phylum in all the assembler whereas maximum number of taxonomic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

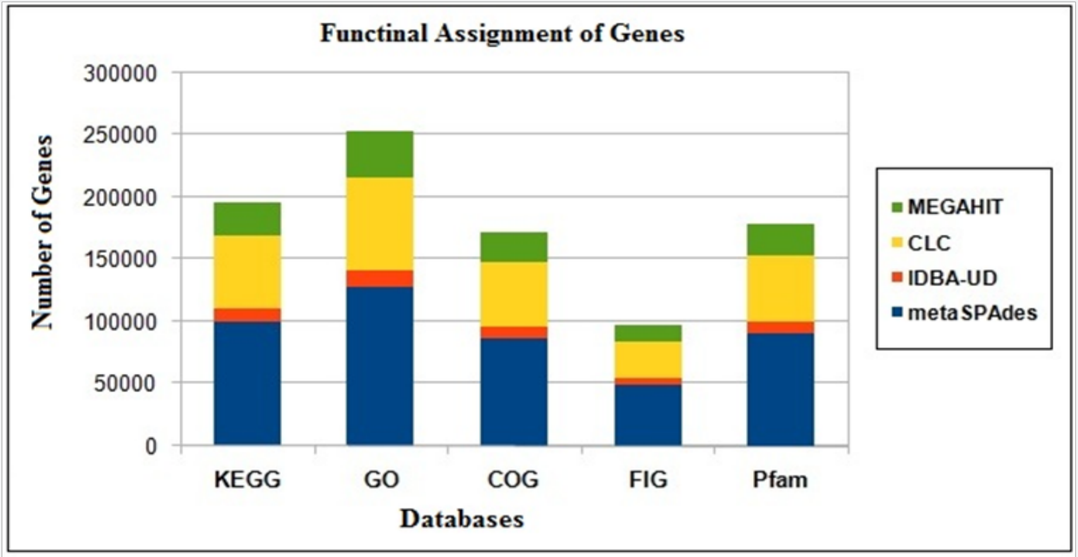

Figure 5: The functional annotation of genes from metaSPAdes assembler, IDBA-UD assembler, MEGAHIT assembler, annotated against GO, KEGG , Pfam, COG and FIG database using cognizer is provided above. From this figure, maximum annotation of genes were enriched with functional anotation obtained from metaSPAdes.

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 metatrancriptomic 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 mico-diversity, metaSPAdes is the best choice.

Acknowledgement

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

References

1. Scholz MB, Lo CC, Chain PS (2012) Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current opinion in biotechnology 23(1): 9-15.
2. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends in genetics 30(9): 418-426.
3. Sheng-Yong Niu, Jinyu Yang, Adam McDermaid, Jing Zhao, Yu Kang, et al. (2017) Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. Briefings in Bioinformatics.
4. Georgia Giannoukos, Dawn M Ciulla, Katherine Huang, Brian J Haas, Jacques Izard, et al. (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome biology 13(3): r23.
5. Yuzhen Ye, Haixu Tang (2015) Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. Bioinformatics 32(7): 1001-1008.
6. John Vollmers, Sandra Wiegand, Anne-Kristin Kaster (2017) Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective-Not Only Size Matters!. PLoS One 12(1): e0169662.
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): 555618. DOI: 10.19080/APBIJ.2017.03.555618.

(2017) metaSPAdes: a new versatile metagenomic assembler. Genome Research 27(5): 824-834.

8. Peng Y, Leung HC, Yiu SM, Chin FY (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27(13): i94-101.

9. Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31(10): 1674-1676.

10. Andrews S (2011) FastQC: a quality control tool for high throughput sequence data. Bioinformatics B. Cambridge, UK: Babraham Institute: 175-176.

11. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114-2120.

12. Chandan Badapanda, Suraj Mahendra Metha (2017) Advancing our understanding of the oxygen minimum zone microbial communities by an integrated metatranscriptomics approach. Meta Gene 14: 85–90.

13. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 11(1): 119.

14. Bose T, Haque MM, Reddy CV, Mande SS (2015) COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets. PLoS One 10(11): e0142102.

This work is licensed under Creative Commons Attribution 4.0 Licens

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

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.