COMPARATIVE METAGENOMICS ANALYSIS OF PALM OIL MILL EFFLUENT (POME) USING THREE DIFFERENT BIOINFORMATICS PIPELINES

ADIBAH PARMAN[1], MOHD NOOR MAT ISA[2], FARAH FADWA BENBELGACEM[1], IBRAHIM ALI NOORBATCHA[1] AND HAMZAH MOHD. SALLEH[3]

[1]Bioprocess & Molecular Engineering Research Unit (BPMERU),
Department of Biotechnology Engineering, Kulliyyah of Engineering,
International Islamic University Malaysia,
[2]Malaysia Genome Institute, Jalan Bangi, 43000 Kajang, Selangor, Malaysia
[3]International Institute for Halal Research and Training (INHART),
International Islamic University Malaysia,
P.O. Box 10, 50728 Kuala Lumpur, Malaysia.

**Corresponding authors: ibrahiman@iium.edu.my, hamzah@iium.edu.my*

(Received: 9th March 2018; Accepted: 25th Jan 2019; Published on-line: 1st June 2019)

https://doi.org/10.31436/iiumej.v20i1.909

ABSTRACT: The substantial cost reduction and massive production of next-generation sequencing (NGS) data have contributed to the progress in the rapid growth of metagenomics. However, production of the massive amount of data by NGS has revealed the challenges in handling the existing bioinformatics tools related to metagenomics. Therefore, in this research we have investigated an equal set of DNA metagenomics data from palm oil mill effluent (POME) sample using three different freeware bioinformatics pipelines' websites of metagenomics RAST server (MG-RAST), Integrated Microbial Genomes with Microbiome Samples (IMG/M) and European Bioinformatics Institute (EBI) Metagenomics, in term of the taxonomic assignment and functional analysis. We found that MG-RAST is the quickest among these three pipelines. However, in term of analysis of results, IMG/M provides more variety of phylum with wider percent identities for taxonomical assignment and IMG/M provides the highest carbohydrates, amino acids, lipids, and coenzymes transport and metabolism functional annotation beside the highest in total number of glycoside hydrolase enzymes. Next, in identifying the conserved domain and family involved, EBI Metagenomics would be much more appropriate. All the three bioinformatics pipelines have their own specialties and can be used alternately or at the same time based on the user's functional preference.

ABSTRAK: Pengurangan kos dalam skala besar dan pengeluaran data 'next-generation sequencing' (NGS) secara besar-besaran telah menyumbang kepada pertumbuhan pesat metagenomik. Walau bagaimanapun, pengeluaran data dalam skala yang besar oleh NGS telah menimbulkan cabaran dalam mengendalikan alat-alat bioinformatika yang sedia ada berkaitan dengan metagenomik. Justeru itu, dalam kajian ini, kami telah menyiasat satu set data metagenomik DNA yang sama dari sampel effluen kilang minyak sawit dengan menggunakan tiga laman web bioinformatik percuma iaitu dari laman web 'metagenomics RAST server' (MG-RAST), 'Integrated Microbial Genomes with Microbiome Samples' (IMG/M) dan 'European Bioinformatics Institute' (EBI) Metagenomics dari segi taksonomi dan analisis fungsi. Kami mendapati bahawa MG-RAST ialah yang paling cepat di antara ketiga-tiga 'pipeline', tetapi mengikut keputusan analisa, IMG/M mengeluarkan maklumat philum yang lebih pelbagai bersama peratus identiti yang lebih luas berbanding yang lain untuk pembahagian taksonomi dan IMG/M juga mempunyai bacaan tertinggi dalam hampir semua anotasi fungsional karbohidrat, amino asid, lipid, dan koenzima pengangkutan dan metabolisma malah juga paling tinggi dalam jumlah enzim hidrolase glikosida. Kemudian, untuk mengenal pasti 'domain' terpelihara dan keluarga yang terlibat, EBI metagenomics lebih bersesuaian. Ketiga-tiga saluran 'bioinformatics pipeline' mempunyai keistimewaan mereka yang tersendiri dan boleh digunakan bersilih ganti dalam masa yang sama berdasarkan pilihan fungsi penggun.

KEYWORDS: *bioinformatics pipeline; metagenomics analysis; MG-RAST; IMG/M; EBI; metagenomics; palm oil mill effluent*

1. INTRODUCTION

In the last few decades, metagenomics has become one of the crucial tools in mining the hidden microbial treasure without the use of conventional laboratory culture techniques. Metagenomics involves the study of genetic material extracted from the diverse microbial population of environmental samples. The early stage of genomics relied on the standard laboratory cultivation method which is insufficient to identify the entire microbial population as compared to metagenomics. Furthermore, the change in biotechnology development within this era, such as inexpensive next-generation sequencing (NGS) technologies, high throughput screening technique for metagenomics library and advances in bioinformatics tools, have left a huge impact in the field of metagenomics [1].

Illumina is the most widely used NGS platform in metagenomics studies. The Illumina system has advantages to other NGS platforms in terms of its high throughput sequencing at an economical price with high accuracy (> 99%) reads [2]. The Illumina platform could initially only produce a short-read sequence length which has gradually been improved to a readable length and consequently made it more popular compared to the other platforms in NGS tools [2]. This fast evolution by NGS technologies allows researchers to achieve more variety of data with a high level of detailed sequencing results. NGS has also been developed continuously and rapidly, starting from its launch in 2006, resulting in the accumulation of massive amounts of sequences [2]. Hence, several bioinformatics tools for metagenomics annotation are needed to accurately analyze the enormous amount of data.

Palm oil mill effluent (POME) is a colloidal suspension of the final stage effluent in the palm oil industry production. The composition of POME includes 95-96% water, 4-5% total solids, and 0.6-0.7% oil [3]. Besides that, the raw POME also contains significant concentrations of carbohydrates, proteins, nitrogenous compounds, lipids, and minerals that enable this effluent to be used in various biotechnological applications like fermentation media, production of antibiotics, bio-insecticides, polyhydroxyalkanoates (PHA), organic acids, enzymes, and hydrogen [3]. The present study attempts to make a comparative bioinformatics analysis of POME's sample metagenome constructed using different automated bioinformatics pipelines of MG-RAST, IMG/MER, and EBI Metagenomics to evaluate their accuracy in the taxonomical assignment and functional annotation for future research directed to the industrial application.

Metagenome Rapid Annotation using Subsystem Technology (MG-RAST) is a free web-based server with a fully automated system that provides sequence alignment, gene prediction, structural and functional annotation, comparative metagenomics and archiving services [4]. It was launched at the Argonne National Library in 2007 to address the computational needs of huge metagenomics data production analysis [5]. This bioinformatic pipeline is being used by researchers around the world with the analysis record of over 250,000 datasets with 100 tera-basepairs of DNA being successfully and completely analyzed to date [6]. Besides, it also has a graphical user interface (GUI) that allows the researcher to study the composition of microbial communities with their specific function [6]. MG-RAST is also one of the bioinformatics pipelines that allows the submission of raw sequence data in the **fastq**, **fasta**, and **sff** format which will then be normalized and processed until annotation is completed by several integrated bioinformatic tools [6].

The Integrated Microbial Genomes with Microbiomes (IMG/M) is quite similar to MG-RAST, which can also examine the taxonomy and function or metabolic potential of microbiomes [7]. IMG/M is a metagenomics data management system supported by the DOE-JGI metagenome annotation pipeline (MAP V.4) which allows the submission of **fasta** or **fastq** format assembled and unassembled 454, Illumina, and pacBio nucleotide sequences [8]. In early 2016, all these unassembled reads could no longer be accepted; meanwhile the sequence data generated outside JGI has been limited to the **fasta** format in assembled data formed only. Until now, IMG/M still supports the external submission of assembled genomes only with the condition that the metagenomes submission and metadata have to be registered with Genomes Online Database (GOLD) version 5 [9].

European Bioinformatics Institute (EBI) Metagenomics is an expanding metagenomics analysis and archiving resource that uses the European Nucleotide Archive (ENA) data scheme developed by the European Molecular Biology Laboratory (EMBL). ENA is needed for the initial submission and archiving purposes in a long-term period storage for reuse in the future [10]. Besides, EBI Metagenomics is a free web-based server that enables users to perform analysis on large scale platforms from Ion Torrent, Roche 454, and Illumina metagenomic sequence data [11]. Similar to MG-RAST and IMG/M, EBI Metagenomics also has an established standardized system and analysis pipeline that includes a variety of analytical and visualization tools in generating the analysis of taxonomic and functional features of user-submitted sequence [12].

A comparison of MG-RAST to Qualitative Insights into Microbial Ecology (QIIME) based on 16S rRNA method found QIIME to be more accurate in term of taxonomic assignment compared to MG-RAST [13]. When MG-RAST was compared to QIIME but with MOTHUR as an additional bioinformatic tool, the results showed that QIIME was again the fastest compared to the other two [14]. QIIME is a bioinformatic tool used by EBI Metagenomics Version 3.0 to perform the taxonomical annotation that currently is being replaced with MAPseq in EBI Metagenomics Version 4.0. Even though in previous research QIIME produced a better result, it lacks the facility to manage, store, and analyze the metagenomics data. In this work, we provide a comparative analysis of the metagenomics data using a fully automated bioinformatics pipeline that integrates the work of management, analysis, storage, and sharing of metagenomics projects [15] instead of analysis of 16S rRNA data only.

The three bioinformatics pipelines, namely MG-RAST, IMG/M, and EBI Metagenomics are chosen because within the existing web resources in metagenomic studies, the three bioinformatics tools are highly rated in terms of ease in data uploading, online user support availability, analysis spectrum, citation, and storage capacity [16]. Hence, this research will only focus on the metagenomics analysis by these three bioinformatic pipelines: MG-RAST, IMG/MER, and EBI Metagenomics. These three tools, especially MG-RAST, have been used repeatedly to analyze many metagenome sequencing datasets from a variety of sources. In the present work, we compare the analysis result of the same input data using the common three web-based automated bioinformatics pipelines of MG-RAST, IMG/MER, and EBI metagenomics to evaluate their accuracy in taxonomy and functional annotation on the microbial diversity and several functional genes contain in POME.

2. MATERIALS AND METHODS

2.1 Collection of Samples and Creation of Metagenomics Libraries

POME samples were collected from FELDA Palm Industries Sdn. Bhd. of KKS Mempaga, Pahang, Malaysia. After sample collection, the metagenomic DNA was extracted using Meta-G-Nome™ DNA Isolation Kit (Epicentre, U.S), sheared, end-repaired, ligated to pCC1FOS fosmid vector and phage-mediated transfection to surrogate host EPI300T1 resistant *E.coli* to perform the cloning of metagenomic DNA grown in LB agar with the recommended antibiotic. The libraries (>100,000) were constructed by preparing 384-well transparent microplates with LB broth and 20% glycerol and each colony were inoculated from LB agar to the microplates and of glycerol media and stored at -80 ºC.

2.2 High Throughput Screening and Next-Generation Sequencing (NGS)

Colonies from each plate were inoculated to 384-well microplates filled with LB broth with antibiotic and inducer (respective to their positions in the library plates). Screening buffer (potassium acetate) and lysis mix (10% Triton X-100, 100 mM Tris and 10 mM EDTA) were added to break open the cells and methylumbelliferyl-β-D-glucopyranoside (MUGlc), methylumbelliferyl-β-D-cellobioside (MUC) and chlorocoumarin-xylobioside (CCX) fluorogenic substrates were added to each well. After an overnight incubation, the plates were screened using a microplate reader for the presence of cellulose- and xylan-degrading enzymes. The relative fluorescence units given by the microplate reader were changed to robust *z*-score to select the high rated hits. Robust *z*-score was calculated for each microplate independently and results of all plates (109,834 clones) were combined to select the 100 high rated clones [17,18] and sent for Illumina HiSeq2000 next-generation sequencing (NGS) at the Malaysia Genome Institute (MGI), Bangi, Selangor.

2.3 Post-Processing of NGS Data

Initial raw data of NGS in the **fastq** format had undergone sequence quality trimming and short read removal by SolexaQA++ program [19]. The fosmid vector and internal control phiX sequences were removed using bowtie2 [20]· The high-quality sequences were assembled with Velvet based on the de Bruijn graph algorithm to organize the sequences in contigs. Velvet is a new strategy developed to merge very short reads in combination with read pairs to produce useful assemblies [21].

2.4 Metagenomics Analysis using Three Bioinformatics Pipelines

The same dataset of assembled data from POME's NGS results with **fasta** format was uploaded into three automated bioinformatics pipelines of MG-RAST, IMG/M, and EBI Metagenomics. These free web-based bioinformatic pipelines were automatically run after submission of the file with related metadata. The details on the uploading method, databases, and system included are summarized in Table 1.

2.4.1 MG-RAST

The **fasta** format file and metadata were uploaded in the upload segment of MG-RAST. The metadata was also uploaded using Microsoft Excel template prior to the submission. Throughout the process, the percentage of analysis and process done could be seen in the progress segment. In MG-RAST, the file had undergone the quality control for data hygiene first and before proceeding with feature identification of coding DNA sequence (CDS) using FragGeneScan. This database could predict the DNA coding region of higher than 75 bp. The similarity searches on taxonomic classification in MG-RAST were conducted using the BLAST-Like Alignment Tool (BLAT) that is used to find sequence hits in seven taxonomic categories. For functional annotation, m5nr was used by providing non-redundant integration of several databases like SEED, KEGG, Genbank, IMG, UniProt, and eggNOGs [6].

Table 1: Technical comparison of MG-RAST, IMG/M, and EBI Metagenomics

	MG-RAST	**IMG/M**	**EBI Metagenomics**
Analysis Duration	5 h 21 mins	5 days 22 h 46 mins	3 days
License	Open-source	Open-source	Open-source
Current Version	4.0.3	4.570	Pipeline Version 4.0
Website	http://metagenomics.anl.gov/	https://img.jgi.doe.gov/m/	https://www.ebi.ac.uk/metagenomics/
Association	-	GOLD & DOE-JGI MAP	ENA
Primary Usage	GUI	GUI	GUI
Feature Identification	rRNA & CDS (FragGeneScan)	CRISPR, tRNA, rRNA & CDS Prediction (GeneMark, Prodigal, MetaGeneAnnotator, FragGeneScan)	rRNA & CDS (FragGeneScan, Prodigal)
Homology-Based Algorithm	BLAT	BLAST, HMMER, USEARCH	-
Functional Annotation	UniProt, IMG/M, COGs, eggNOGs, KEGG, SEED, STRING	COGs, Pfam, KO, EC, MetaCyc, KEGG	Comparing CDS using InterProScan
Taxonomic Annotation	BLAT	BLAST	Comparing rRNA using MAPseq

2.4.2 IMG/M

To submit our own data for metagenomics analysis in IMG/M, we used the Integrated Microbial Genomes with Microbiomes for expert review (IMG/MER). IMG/M has been associated with GOLD v.5 and DOE's Joint Genome Institute (JGI). The initial step involved is by login to IMG/MER and then registering the project metadata at GOLD prior to uploading the data at JGI. After completion of data submission, the quality control (QC) pre-processing like trimming and removing sequences shorter than 150 bp would take place. Next, the genes prediction of CRISPR, tRNA, rRNA, and CDS was conducted using several databases like GeneMark, Prodigal, MetaGeneAnnotator, and FragGeneScan. Finally, the functional annotation was completed by associating the protein-coding genes with the COG, Pfam, KO, EC, MetaCyc, and KEGG [8].

2.4.3 EBI Metagenomics (EMG)

To analyze the data from POME's NGS results using EBI metagenomics, the data file needs to be uploaded to the European Nucleotide Archive (ENA) using either Webin Uploader, FileZilla Client-Server, or Aspera. In this research, FileZilla Client-Server was utilized, which is easier to use compared to the other methods and has been used widely by many researchers. After QC and filtering out ncRNA reads, the CDS of proteins were predicted using FragGeneScan for short reads and Prodigal for longer reads or assembled ones. In EBI Metagenomics, MAPseq was used for taxonomic classification by assigning taxonomy and OTU classification to rRNA sequence. In the meantime, InterProScan was also used for functional annotation by predicting the domains and classifying them into families using a compilation of several databases like Pfam, TIGRFAM, PRINTS, PROSITE patterns, and Gene3d. Finally the functional results were produced in Gene Ontology (GO).

3. RESULTS AND DISCUSSION

To ensure the similar type of dataset for data comparison, the file in **fasta** format with assembled form has been selected for this research. The assembled format is used instead of the raw sequence because one of the bioinformatics pipelines, IMG/M cannot process the unassembled reads if they are generated outside the Joint Genome Institute (JGI) [9]. The data was submitted to get a clear result related to the taxonomic and functional annotation involving POME DNA metagenome sample. Specifically, the results of major phylum and genus distribution existing in POME will represent the taxonomic classification. On the other hand, major functional annotation and glycoside hydrolase enzyme present in POME's metagenomics DNA results will represent the functional annotation part of this research. Overall, MG-RAST was found to be the quickest in 5 hours 20 minutes (Table 1) compared to IMG/M and EBI Metagenomics. However, QIIME was the quickest compared to MG-RAST and MOTHUR [14]. On the contrary, MG-RAST is the quickest in this research, even though QIIME was included in EBI Metagenomics system Version 3.0. In this condition, QIIME is not a stand-alone system, and EBI metagenomics also involves other bioinformatic tools like FragGeneScan, Prodigal, and InterProScan which explains the slowness of EBI Metagenomics containing QIIME compared to MG-RAST.

Figure 1(a) showed the results of major phylum distribution present in effluent of FELDA palm oil mill, Mempaga, Pahang by each bioinformatic pipelines. All three bioinformatic pipelines have similar dominant phylum of Proteobacteria and Firmicutes with MG-RAST 73.7% and 25.3%, IMG/M 48% and 28%, and EBI Metagenomics 67% and 33%, respectively. The difference can be seen in IMG/M ER with additional dominant phylum of Actinobacteria (12%), dsDNA viruses (5%), and Bacteroidetes (4%). In this phylum composition, IMG/M ER appeared to be more diverse and details compared to the other two bioinformatics pipeline. MG-RAST using a BLAT algorithm while IMG/M ER used BLAST which is the most commonly used bioinformatics tool for sequence similarity analysis. Besides that, IMG/M could provide the percent identities greater than 30% which ensured the larger amount of results produced, while the setting in MG-RAST is only for the percent identities of 60% and above.

For genus distribution result, the comparison table only involves MG-RAST and IMG/MER as shown in Fig. 1(b). EBI Metagenomics was excluded because no details genes distribution could be detected except *Staphylococcus*. The latest version of EBI Metagenomics Pipeline 4.0 uses MAPseq framework for taxonomic classification whichallows the metagenomic analysis of reference based rRNA only. Consequently, only small amount of 16S rRNA could be found in this sequence and finally bring the limitation of *Staphylococcus* as an output. On the contrary, the other two pipelines of MG-RAST and IMG/M, besides rRNA as reference-based, they also used coding DNA sequence (CDS) of protein as the reference which results thousands of CDS can be found from the data submission.

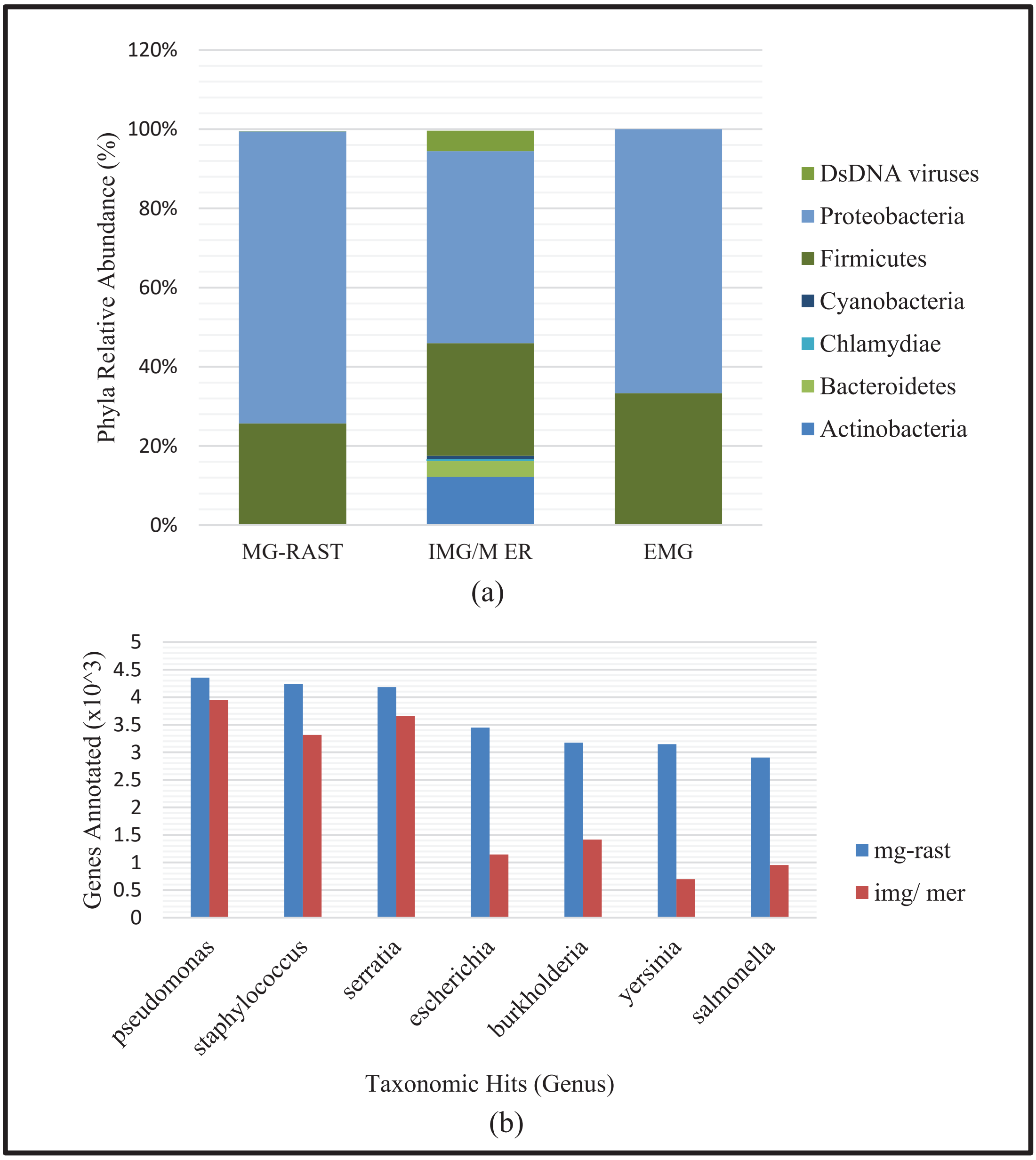

Fig. 1: (a) Phyla distribution of palm oil mill effluent (POME) computed by each tool and (b) Major genus distribution in POME analyzed by MG-RAST and IMG/M only.

The major genus found in MG-RAST in Fig. 1(b) is similar to IMG/MER which are *Pseudomonas, Staphylococcus, Serratia, Escherichia, Burkholderia, Yersinia,* and *Salmonella.* From the same figure too, the quantity of major genes found in all MG-RAST's genus is greater than IMG/M, and the total number of the genus found in MG-RAST initially also is greater than IMG/M. Soleimaninanadegani and Manshad's work on POME reported that they found *Bacillus*, *Micrococcus*, *Pseudomonas* and *Staphylococcus* genus in POME sample [22]. This finding explains the existence of *Pseudomonas* and *Staphylococcus* as one of the major genus found in POME while the other genus like *Serratia*, *Burkholderia*, *Yersinia*, and *Salmonella* are new findings in this research. Besides that, all the genus are identified under the family of Firmicutes and Proteobacteria which also explain the two families as the main families present in POME.

For functional annotation, several major categories of functional annotation can been extracted from the results obtained by each pipeline which are transport and metabolism of nucleotide, lipid, inorganic ion, coenzyme, carbohydrate, and amino acid as shown in Fig. 2. Raw POME contains a high concentration of carbohydrate, protein, nitrogenous compounds, lipids, and minerals [23], which explains why in all three pipelines' analysis of the metabolism that lipid, carbohydrate, amino acid, inorganic ion, and coenzyme are among the highest annotated reads of functional annotation. Once again, IMG/M has the highest number of annotated reads nearly in all the functional categories.

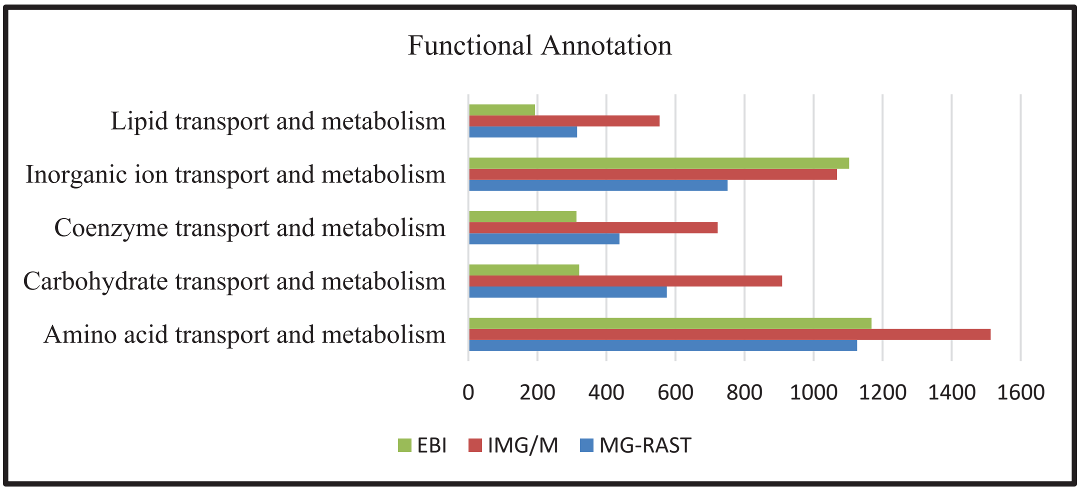

Fig. 2: Major functional annotation of palm oil mill effluent (POME) analyzed by MG-RAST, IMG/M, and EBI.

All the glycoside hydrolase enzymes analyzed by each pipeline are listed in Table 2 in order for future research related to the study of useful glycoside hydrolases related to POME. The POME glycoside hydrolase enzymes identified by bioinfomatics tools suggest that POME can be a potential local source for novel enzymes and future potential enzymes for industrial applications. The blank spaces of EBI Metagenomics part (Table 2) probably occurred because EBI metagenomics used InterProScan in functional annotation. InterProScan is a database that works by classifying the CDS into respected families and also by predicting any important domains and sites related to protein sequence [24]. Therefore, the output result of EBI Metagenomics will be more related to predicted domains and families rather than the protein or enzyme names.

Table 2: Total number of enzyme glycoside hydrolase (EC 3.2.1.-) in palm oil mill effluent (POME) identified by MG-RAST, IMG/M ER, and EBI Metagenomics

Enzyme ID	Enzyme Name	MG-RAST	IMG/M	EBI
EC 3.2.1.1	Alpha-amylase	1	2	5
EC 3.2.1.14	Chitinase	-	4	12
EC 3.2.1.141	Malto-oligosyltrehalose trehalohydrolase	-	2	-
EC 3.2.1.17	Lysozyme	-	1	-
EC 3.2.1.20	Alpha-glucosidase	1	3	-
EC 3.2.1.21	Beta-glucosidase	1	3	-
EC 3.2.1.23	Beta-galactosidase	2	2	7
EC 3.2.1.26	Beta-fructofuranosidase	3	3	-
EC 3.2.1.4	Cellulase	3	1	2
EC 3.2.1.52	Beta-N-acetylhexosaminidase	11	5	-
EC 3.2.1.85	6-phospho-beta-galactosidase	1	1	-
EC 3.2.1.86	6-phospho-beta-glucosidase	4	4	2
EC 3.2.1.93	Alpha,alpha-phosphotrehalase	1	2	2
		28	33	30

4. CONCLUSIONS

Throughout the results of taxonomic and functional annotation of palm oil mill effluent (POME) DNA metagenomics sample, IMG/M could be seen to provide a better analysis result compared to the other pipelines. IMG/M could give a diverse phylum result as it involves a much wider range of percent identity. Besides that, this bioinformatics pipeline also is the one with the highest in nearly all major functional annotation of carbohydrate, amino acid, lipid and coenzyme transport and metabolism as well as the highest in the total number of glycoside hydrolase enzymes contained in POME. On the other hand, the other two pipelines of MG-RAST and EBI metagenomics also have their own specialties. These two bioinformatic tools allow the submission of raw data compared with IMG/M, which is limited to the assembled one only. MG-RAST is also good for obtaining data analysis in a short period as it uses the BLAT system which is 500 times faster for nucleic acid sequence alignment and 50 times faster for protein sequence alignment than other famous existing tools including the one being used in IMG/M and EBI Metagenomics server [25]. Next, EBI metagenomics is more suitable for a deeper understanding of the protein families and domains involved in any specific sequence rather than focusing on taxonomical annotation. In conclusion, all the three bioinformatics pipelines have their own specialty that the researcher needs to know in detail so that the specific bioinformatic pipeline can be used based on individual functional interest

ACKNOWLEDGMENT

The authors would like to thank Ministry of Education, Malaysia for giving the financial support through the Fundamental Research Grant Scheme (FRGS) with project ID FRGS 13-086-0327. Besides that, the authors also deeply grateful to Malaysia Genome Institute (MGI) for the technical assistance in analyzing the metagenomic data of POME. The authors also would like to express their appreciation to Professor S.G. Withers and Dr Hong Ming (Chemistry Department, University of British Columbia, Vancouver, Canada) for providing substrates used for the high-throughput screening of this work.

REFERENCES

[1] Kumar S, Krishnani KK, Bhushan B, and Brahmane MP. (2015) Metagenomics : Retrospect and Prospects in High Throughput Age, vol. 2015.

[2] Mincheol K, Lee KH, Yoon SW, Kim BS, Chun J, and Hana Y. (2013) Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era, 11(3): 102-113.

[3] Abdullah N and Sulaiman F. (2013) The oil palm wastes in Malaysia, Biomass Now – Sustain. Growth Use, pp. 75-100.

[4] Meyer F, Paarmann D, Souza MD, Olson R, Glass EM, Kubal M, Edwards RA. (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes, 8: 1-8. https://doi.org/10.1186/1471-2105-9-386.

[5] Tang W, Bischof J, Desai N, Mahadik K, Gerlach W, Harrison T, Meyer F. (2014) Workload characterization for MG-RAST metagenomic data analytics service in the cloud. 2014 IEEE International Conference on Big Data (Big Data), 56-63. https://doi.org/10.1109/BigData.2014.7004394.

[6] Wilke A, Gerlach W, Harrison T, Paczian T, Trimble WL, & Meyer F. (2016) MG-RAST Manual for version 4, revision 1.

[7] Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Pillay M, Kyrpides NC. (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Research, 42(D1): 568-573. https://doi.org/10.1093/nar/gkt919.

[8] Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-espino D, Tennessen K, Kyrpides NC. (2016) The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v . 4). Standards in Genomic Sciences, 1: 1-5. https://doi.org/10.1186/s40793-016-0138-x.

[9] Chen IMA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Kyrpides NC. (2017) IMG/M: Integrated genome and metagenome comparative data analysis system. Nucleic Acids Research, 45(D1): D507-D516. https://doi.org/10.1093/nar/gkw929.

[10] Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-beltran A, Hunter C, Sansone S. (2014) EBI metagenomics — a new resource for the analysis and archiving of metagenomic data, 42: 600-606. https://doi.org/10.1093/nar/gkt961.

[11] Denise H, Mitchell A, Bucchini F, Cochrane G, Denise H, Hoopen P, Finn RD. (2016) EBI metagenomics in 2016 - An expanding and evolving resource for the analysis and archiving of metagenomic data EBI metagenomics in 2016 - An expanding and evolving resource for the analysis and archiving of metagenomic data, (November 2015). https://doi.org/10.1093/nar/gkv1195.

[12] Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Finn RD. (2017) EBI Metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Research, 46(D1): D726-D735. https://doi.org/10.1093/nar/gkx967.

[13] D'Argenio V, Casaburi G, Precone V, & Salvatore F. (2014) Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. Biomed Res Int, 325340. https://doi.org/10.1155/2014/325340.

[14] Plummer E, Twin J. (2015) A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. J. Proteomics & Bioinformatics, 8(12): 283–291. https://doi.org/10.4172/jpb.1000381.

[15] Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Iliopoulos I. (2015) Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. Bioinformatics and Biology Insights, 9: 75-88. https://doi.org/10.4137/BBi.s12462.

[16] Dudhagara P, Bhavsar S, Bhagat C, Ghelani A, Bhatt S, & Patel R. (2015) Web Resources for Metagenomics Studies, 13: 296–30.

[17] Malo N, Hanley JA, Cerquozzi S, Pelletier J, & Nadon R. (2006) Statistical practice in high-throughput screening data analysis. Nature Biotechnology, 24(2): 167-175. https://doi.org/10.1038/nbt1186.

[18] Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, Shamu CE. (2009) Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods, 6(8): 569-575. https://doi.org/10.1038/nmeth.1351.

[19] Cox MP, Peterson DA, & Biggs PJ. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics, 11: 485. https://doi.org/10.1186/1471-2105-11-485.

[20] Langmead B, & Salzberg SL. (2013) Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4): 357-359. https://doi.org/10.1038/nmeth.1923.Fast.

[21] Zerbino DR, Birney E. (2008) Velvet Manual Version 1.1. Genome Research, 18(5): 821-829. https://doi.org/10.1101/gr.074492.107.

[22] Soleimaninanadegani M, Manshad S. (2014) Enhancement of Biodegradation of Palm Oil Mill Effluents by Local Isolated Microorganisms. International Scholarly Research Notices, 1-8. https://doi.org/10.1155/2014/727049.

[23] Habib MAB, Yusoff FM, Phang SM., Ang KJ, & Mohamed S. (1997) Nutritional values of chironomid larvae grown in palm oil mill effluent and algal culture. Aquaculture, 158(1-2): 95-105. https://doi.org/http://dx.doi.org/10.1016/S0044-8486(97)00176-2.

[24] Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Mitchell AL. (2017) InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Research, 45(D1): D190-D199. https://doi.org/10.1093/nar/gkw1107.

[25] Kent WJ. (2002) BLAT — The BLAST -Like Alignment Tool. Genome Research, 12: 656-664. https://doi.org/10.1101/gr.229202.