

## Next Generation Sequencing-Data Analysis for Cellulose- and Xylan-Degrading Enzymes from POME Metagenome

(Analisis Data-Penjujukan Generasi Seterusnya bagi Enzim Selulosa dan Xilan Mendegradasi daripada Metagenom POME)

FARAH FADWA BENBELGACEM, MOHD NOOR MAT ISA, MUHAMMAD ALFATIH MUDDATHIR ABDELRAHIM, AFIDALINA TUMIAN, OUALID ABDELKADER BELLAG, ADIBAH PARMAN, IBRAHIM ALI NOORBATCHA & HAMZAH MOHD SALLEH\*

### ABSTRACT

*Metagenomic DNA library from palm oil mill effluent (POME) was constructed and subjected to high-throughput screening to find genes encoding cellulose- and xylan-degrading enzymes. DNA of 30 positive fosmid clones were sequenced with next generation sequencing technology and the raw data (short insert-paired) was analyzed with bioinformatic tools. First, the quality of 64,821,599 reverse and forward sequences of 101 bp length raw data was tested using Fastqc and SOLEXA. Then, raw data filtering was carried out by trimming low quality values and short reads and the vector sequences were removed and again the output was checked and the trimming was repeated until a high quality read sets was obtained. The second step was the de novo assembly of sequences to reconstruct 2900 contigs following de Bruijn graph algorithm. Pre-assembled contigs were arranged in order, the distances between contigs were identified and oriented with SSPACE, where 2139 scaffolds have been reconstructed. 16,386 genes have been identified after gene prediction using Prodigal and putative ID assignment with Blastp vs NR protein. The acceptable strategy to handle metagenomic NGS-data in order to detect known and potentially unknown genes is presented and we showed the computational efficiency of de Bruijn graph algorithm of de novo assembly to 21 bioprospect genes encoding cellulose-degrading enzymes and 6 genes encoding xylan-degrading enzymes of 30.3% to 100% identity percentage.*

*Keywords:* de Bruijn; de novo assembly; metagenomics; scaffold; SSPACE

### ABSTRAK

*Sebuah pangkalan data yang menyimpan DNA metagenom daripada efluen kilang minyak kelapa sawit telah dibina dan disaring dengan menggunakan kaedah penyaringan berskala besar untuk mencari enzim selulosa dan xilan. DNA daripada fosmid berklon positif telah disusun dengan menggunakan teknologi penjujukan berskala besar dan data mentah (dalam susunan pendek berpasangan) telah dianalisis dengan kaedah bioinformatik. Pertama, kualiti susunan 64,821,599 balikan dan ke depan sebanyak 101 bp panjang data mentah telah diuji menggunakan Fastqc dan SOLEXA. Kemudian, penyaringan data mentah dilakukan dengan memotong susunan yang berkualiti rendah dan pendek. Malah, vektor juga telah dikeluarkan dan susunan output telah diperiksa dan ditrim berulang kali sehingga set bacaan berkualiti tinggi diperolehi. Langkah kedua adalah himpunan de novo iaitu untuk menyusun semula 2900 contigs mengikut algoritma graf de Bruijn. Contigs awal sebelum himpunan telah diatur mengikut susunan, jarak antara contigs telah dikenal pasti berorientasikan SSPACE dengan 2139 perancah telah dibina. 16,386 gen telah dikenal pasti selepas kaedah peramalan gen menggunakan Prodigal dan penugasan ID putatif dengan Blastp vs protein NR. Strategi yang betul dalam mengendalikan data NGS-metagenom untuk mengesan gen-gen yang diketahui dan juga yang berpotensi tetapi masih belum diketahui telah ditunjukkan. Dalam kajian ini, kami menunjukkan kecekapan pengiraan komputer berdasarkan algoritma graf himpunan de Bruijn de novo kepada bioprospek 21 gen yang mengkodkan enzim selulosa dan 6 gen yang mengkodkan enzim xilan daripada 30.3% kepada 100% peratusan identiti yang serupa.*

*Kata kunci:* de Bruijn; himpunan de novo; metagenom; perancah; SSPACE

### INTRODUCTION

Metagenomic is a new strategy which has been invented to bioprospect untapped biocatalysts from unculturable microbiota by skipping the classic ways of laboratory cultivation (Kumar et al. 2015) For further improvement of biocatalysts discovery, high-throughput screening (HTS) tools are combined with metagenomic approach. Microtiter screening is one of the HTS methods which is widely

used to screen huge libraries in a short time (Mewis et al. 2013). In HTS, sensitive fluorogenic substrates are used to detect enzymes with the desired activity (Armstrong et al. 2015; Taupp et al. 2011). Finding the gene encoding the biocatalyst is a crucial part in the work flow and to get the full sequence of the gene, DNA sequencing needs to be carried out. Nowadays, Next Generation Sequencing (NGS) is adopted in metagenomic approaches due to its

capacity to merge huge data by covering big library size in one run and due to its high sensitivity and accuracy (Henson et al. 2012). NGS strategy overcomes the problem of the high cost and longtime of genome sequencing by Sanger sequencing approach; NGS allows for the shearing of the genome into sequences and re-assemble them after sequencing (Li et al. 2012).

Illumina is one of the second generation sequencing technologies which is based on synthesis principles and requires bridge amplification. In the amplification step, adaptor modified DNA strands hybridize to oligonucleotide anchor and the sequencing is performed by reversible dye terminator. Reads length varies between 50 and 250 bp and up to 3 billion reads as the output. The main advantage of Illumina technology is the high sequence yield while the two major disadvantages of this technology is the need of high DNA concentration and the high cost (Minakshi et al. 2014). The NGS-data analysis may be a dilemma if the wrong analysis strategy is followed. The data size should be considered during data-analysis work flow, where software to be used should be able to recover all the data. As Illumina sequencing strategy is based on the fragmentation of DNA to short reads before sequencing, an assembly step is compulsory to overlap the genome.

The assembly method depends on the type of sequenced DNA, where genomic DNA or transcriptome cDNA is not assembled in the same manner. The technology is able to handle DNA that is known or otherwise; the assembly of known DNA is based on homology alignment of reads with the DNA template termed reference genome assembly while the unknown method of assembly is called the *de novo* assembly where no homology information is available. In the current research, metagenomic DNA was isolated from palm oil mill effluent and fosmid DNA library was created and the DNA library was screened with cellulose and xylan analog substrates. High rated fosmids clones were isolated and sequenced with Illumina Hiseq. The flow-chart of NGS-raw data analysis starts with the pre-processing step to improve data quality and as the metagenomic DNA is a mixture of culturable and unculturable microbes, there are no reference data to be used as reference homology data for analysis. The novel strategy of *de novo* assembly overcomes this problem and makes the data analysis of unknown genomes possible. *de Bruijn* graph (DBG) was chosen instead of overlap-layout-consensus (OLC) due to its capacity to process massive amounts of NGS data and the fact that it involves breaking reads into even shorter k-mers. Due to the lack of information and protocols that clearly presents NGS-data analysis of functional metagenomics approach for the detection of special activities, this article aims to clarify and simplify the needed workflow.

## MATERIALS AND METHODS

### MATERIALS

The NGS data is obtained from metagenomic DNA library constructed from palm oil mill effluent (POME) microbiota

(FELDA Mempaga Mill, Bentong, Pahang, Malaysia). The DNA vector was pCC1FOS fosmid and the bacterial strain was *E. coli* EPI300T1R (Cat. No. CCFOS110 from Epicentre). The fosmid DNA was extracted using FosmidMAX™ DNA Purification Kit (Cat. No. FMAX046 from Epicentre). Fosmids were sequenced with Illumina Hiseq 2000 technology at Malaysia Genome Institute (MGI).

### METAGENOMIC DNA LIBRARY CONSTRUCTION AND HIGH-THROUGHPUT SCREENING

Metagenomic DNA library was constructed by shearing the POME metagenome after extraction to sequences of 35-45 kb size range and ligated to pCC1FOS fosmid and transfected to EPI300T1R with phage system. For this purpose, CopyControl™ Fosmid Library Production Kit with pCC1FOS™ Vector (Epicentre) was used following the manufacturer's instructions. The library was screened with three fluorogenic substrates: methylumbelliferyl-β-D-glucopyranoside (MUGlc), methylumbelliferyl-β-D-cellobioside (MUC) and chlorocoumarin-xylobioside (CCX) using a fluorescence microplate reader (Zerbino & Birney 2008). The 30 positives clones identified with the HTS were cultured in Luria Bertani (LB) media overnight with autoinduction solution and the fosmids were isolated with FosmidMAX™ DNA Purification Kit following the manufacturer's instructions. Thirty isolated fosmids were sequenced with Hiseq2000 strategy. The raw data of the sequencing is in fastq format and it was divided into 16 files of reverse and 16 files of forward sequences, each file containing 4,000,000 sequences of 101 bp length plus one the 17th file of forward and 17th file of reverse with 821,599 sequences of 101 bp length.

### NGS-DATA ANALYSIS PRE-PROCESSING

The workflow of NGS-data analysis is presented in Figure 1. In the first step of pre-processing, the raw quality was checked and the reads were trimmed accordingly. Quality control was done in both Command Line Interface (CLI) and Graphic User Interface (GUI). In GUI, FastQC (Version 0.11.5 released) was used; it is a tool provided by Babraham Institute which makes the quality control of high-throughput sequencing pipelines an easy matter. In CLI, SolexaQA was used (Boetzer et al. 2011). The Illumina sequence files were uncompressed first using the command:

```
gunzip R1.fq.gz
gunzip R2.fq.gz
```

Each fastq file reads were counted in TOTAL READS with the command:

```
find R1.fq -exec grep -c -H "@HWI-"
{} \; >> TOTAL_READS

find R2.fq -exec grep -c -H "@HWI-"
{} \; >> TOTAL_READS
```

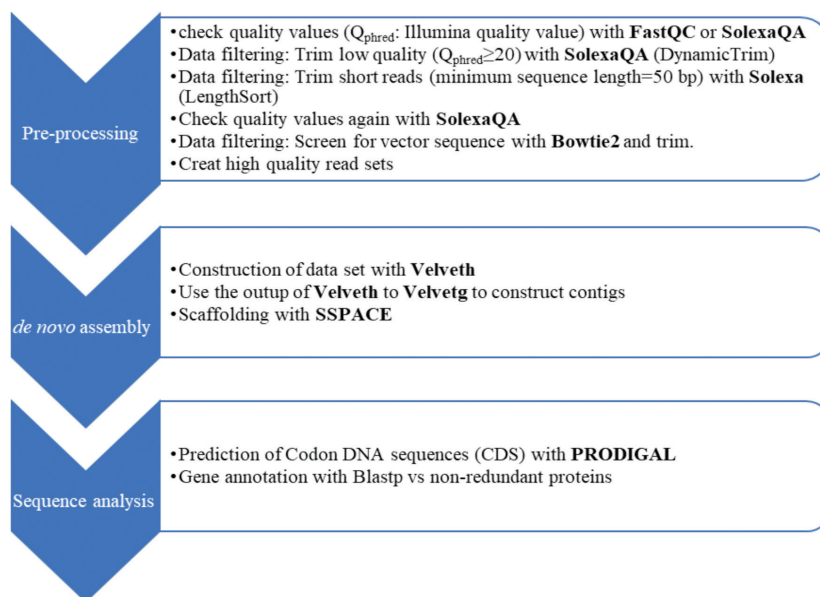


FIGURE 1. NGS-Data analysis workflow of 30 recombinant fosmid DNA. Pre-processing, *de novo* assembly and sequence analysis are the main steps of the presented work pipeline. In bold are the bioinformatic tools used in this pipeline

In SolexaQA, DynamicTrim command was used to trim the reads basing on quality values  $Q_{\text{phred}}$  sequences with  $Q_{\text{phred}}$  less than 20 were trimmed with the command:

```
SolexaQA++ dynamictrim -h 20 R1.fq
SolexaQA++ dynamictrim -h 20 R2.fq
```

LengthSort command was used to trim the reads based on sequence length by removing sequences shorter than 50 bp with the command:

```
SolexaQA++ lengthsort -l 50 R1.fq.
trimmed
SolexaQA++ lengthsort -l 50 R2.fq.
trimmed
```

Reads of TOTAL HIGH QUALITY READS of  $Q_{\text{phred}}$  more than 20 and sequences larger than 50 bp were counted with the command:

```
find R1.fq.trimmed.single -exec grep
-c -H "@HWI-" {} \; >> TOTAL_HIGH_
QUALITY_READS
find R2.fq.trimmed.single -exec grep
-c -H "@HWI-" {} \; >> TOTAL_HIGH_
QUALITY_READS
```

Bowtie2 was used to filter pCC1FOS sequences and the phiX spike used in Illumina sequencer as positive control with the command:

```
Bowtie2 -x ../bowtie_index/phix-174
-x ../bowtie2_index/pcc1fos R1.fq.
trimmed.single -un R1.cleaned.fq
```

```
Bowtie2 -x ../bowtie_index/phix-174
-x ../bowtie2_index/pcc1fos R2.fq.
trimmed.single -un R2.cleaned.fq
```

The reads of cleaned fastq files were counted and got into TOTAL CLEAN READS file with the command:

```
find R1.cleaned.fq -exec grep -c -H
"@HWI-" {} \; >>
TOTAL_CLEAN_READS
find R2.cleaned.fq -exec grep -c -H
"@HWI-" {} \; >>
TOTAL_CLEAN_READS
```

The cleaned sequences are paired or singletons, the paired-end sequences were then shuffled and the singletons reads were combined into a file using the followed two commands:

```
select_paired_fq.pl R1.cleaned.fq
R2.cleaned.fq R1.PE.fq R2.PE.fq
shuffleSequences_fastq.pl R1.PE.fq
R2.PE.fq reads.PE.shuffled.fq
cat R1.PE.fq.single R2.PE.fq.single
> reads.SG.fq
```

#### DE NOVO GENOME ASSEMBLY AND SCAFFOLDING

The second step is the *de novo* assembly of DNA sequences using Velvet (Hyatt et al. 2010). Velvet is a two step-process, first Velvet is used to prepare reads for Velvetg and the latter is to assemble the reads into contigs.

The Velvet program and all available options were first called up by simply typing:

```
velveth
velvetg
```

The VelvetOptimiser.pl script was used to optimize the Velvet parameters by finding automatically the optimal k-mer within the range 61-85 using the command:

```
VelvetOptimiser.pl -s 61 -e 85 -o
'-ins_length 500 -| ins_lentgh_sd
100' -t 2 -a -f '-fastq -shortPaired
\ reads.PE.shuffled.fq -short reads.
SG.fq'
```

where -s is the starting hash value, -e is the end hash value, -o is an extra Velveth options to pass through, -t is the maximum number of simultaneous Velvet instances to run, -a is to turn on Velvet's read tracking and amos file output and -f is the file section on the Velvet command line.

The optimal k-mer was 79 and the command used to run Velveth was:

```
velveth auto_79 79 -fastq -shortPaired
reads.PE.shuffled.fq -short reads.
SG.fq
```

To run the Velvetg the following command was used:

```
velvetg auto_79 -exp_cov auto
```

The orientation and arrangement of contigs or what is called scaffolding was performed using SSAGE-based Scaffolding of Pre-Assembled Contigs after Extension (SSPACE) (Li et al. 2012). The command used for scaffolding is:

```
SSPACE_Basic.pl -l libraries.txt -s
contigs.fa -x 0 -m 32 -o 20 -t 0 -k
5 -a 0.70 -n 15 -p 0 -v 0 -z 0 -g 0
-T 20 -b ref_genome.fa
```

where -l is library file containing two paired read files with insert size, error and orientation. -s is FASTA file containing contig sequences used for extension. Inserted paired reads are mapped to extended and non-extended contigs. -x indicates whether to extend the contigs of -s using paired reads in -l (-x 1=extension, -x 0=no extension, default -x 0). -m is minimum number of overlapping bases with the seed/contig during overhang consensus build up (default -m 32). -o is minimum number of reads needed to call a base during an extension (default -o 20). -t is to trim up to -t base(s) on the contig end when all possibilities have been exhausted for an extension (default -t 0). -k is minimum number of links (read pairs) to compute scaffold (default -k 5). -a is maximum link ratio between two best contig pairs. Higher values lead to least accurate scaffolding (default -a 0.7). -n is minimum overlap required between contigs to merge

adjacent contigs in a scaffold (default -n 15). -p is to make .dot file for visualisation (-p 1=yes, -p 0=no, default -p 0). -v is runs in verbose mode (-v 1=yes, -v 0=no, default -v 0). -z is minimum contig length used for scaffolding, it filters out contigs below this value (default -z 0). -g is maximum number of allowed gaps during mapping with Bowtie, it corresponds to the -v option in Bowtie and higher number of allowed gaps can lead to least accurate scaffolding (default -g 0). -T is to specify the number of threads in Bowtie, it corresponds to the -p/--threads option in Bowtie (default -T 1). -b is base name for output files (default -b standard\_output).

#### BACTERIAL GENE PREDICTION AND ANNOTATION

The output of SSPACE was used in PROkaryotic Dynamic programming Gene finding ALgorithm (PRODIGAL) to identify codon DNA sequences CDSs (Minakshi et al. 2014). The genes prediction was carried out with the command:

```
prodigal -a predictedGenes.faa -d
predictedGenes.ffn -f gff -I ref_
genome.fa -o predictedGenes.gff
```

where -a is to write protein translations to the selected file, -d is to write nucleotide sequences of genes to the selected file, -f is to select output format (gbk, gff or sco), -i is to specify FASTA/Genbank input file and -o is to specify output file.

The genes annotation process was carried out using the BLAST software against NCBI's non-redundant database. The first step consists on the creation of data base with the command:

```
Makeblastdb -in BLAST_db.ffn -dbtype
nucl
```

The BLAST then was run with the command:

```
blastp -db BLAST_db.ffn -query
predictedGenes.ffn -out \
blastpOutput.out -outfmt 0 -num_
threads 20
```

where the db is the BLAST data base, query is all genes file name, out is the output file name, outfmt is the alignment view options, 0 = pairwise and num\_threads are number of threads (CPUs) to use in blast search.

The output of the BLASTp result was given in pairwise alignment format blastpOutput.out. The output was converted into an easier and more readable tabular format with the command:

```
Onehitperquery_bpSearchIO+Strand.
pl blastpOutput.out \ blastpOutput.
parsed
```

The BLAST results and the predicted GFF into an annotated file that contains information from both with:



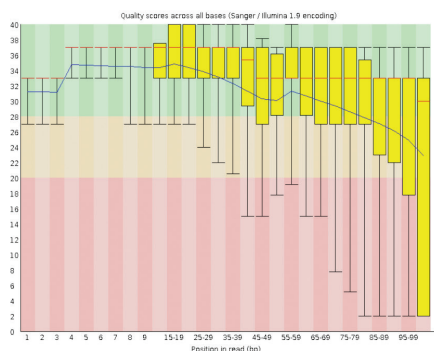
```
tabfile_creator_from_gff_and_blastparsed.sh
```

## RESULTS AND DISCUSSION

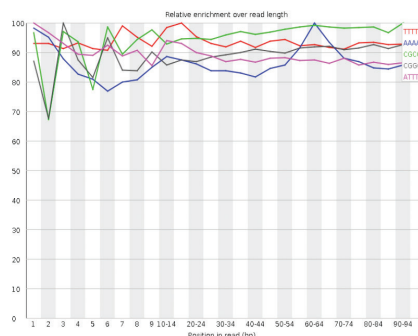
### PRE-PROCESSING STEP

Before starting the main analysis, a pre-processing step is extremely important to check the quality of the reads

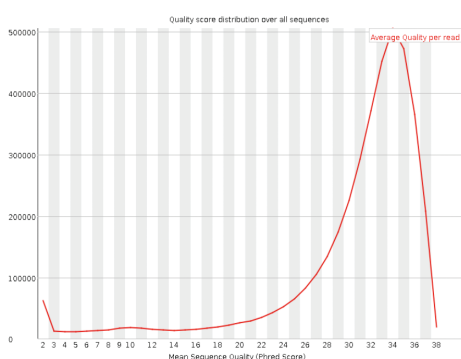
with SolexaQA or FastQC, reads quality statistics were represented and graphically visualised. Figure 2 is an example of how the FastQC report is given and in Figure 3, two different results of per base sequence quality where one is failed quality (A) and one is pass (B). Poor quality bases were removed with DynamicTrim based on BWA trimming algorithm, all basess with Qphred<20 were trimmed and the LengthSort algorithm was used to remove sequences less than 50 bp (Cox et al. 2010). At



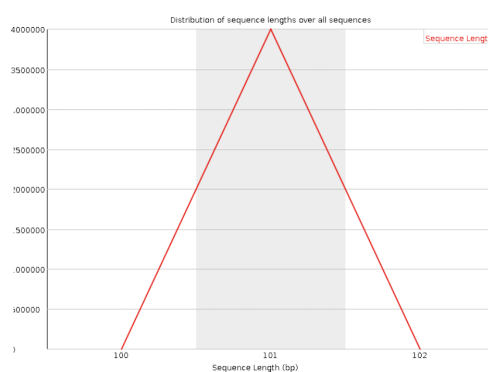
A- Per base sequence quality



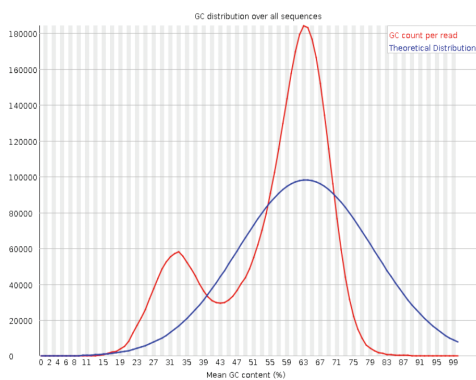
B- k-mer content



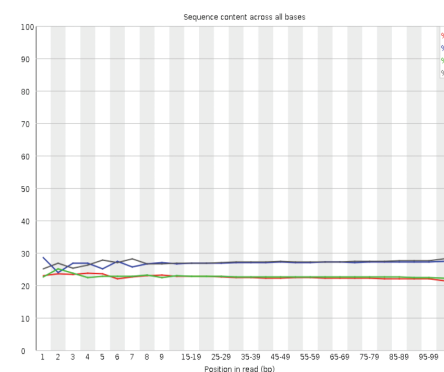
C- Per sequence quality score



D- Sequence length distribution



E- Per sequence GC content



F- Sequence content across all bases

FIGURE 2. Example of one of the FastQC reports of per base sequence quality, k-mer content, per sequence quality score, sequence length distribution, per sequence GC content and per base sequence content

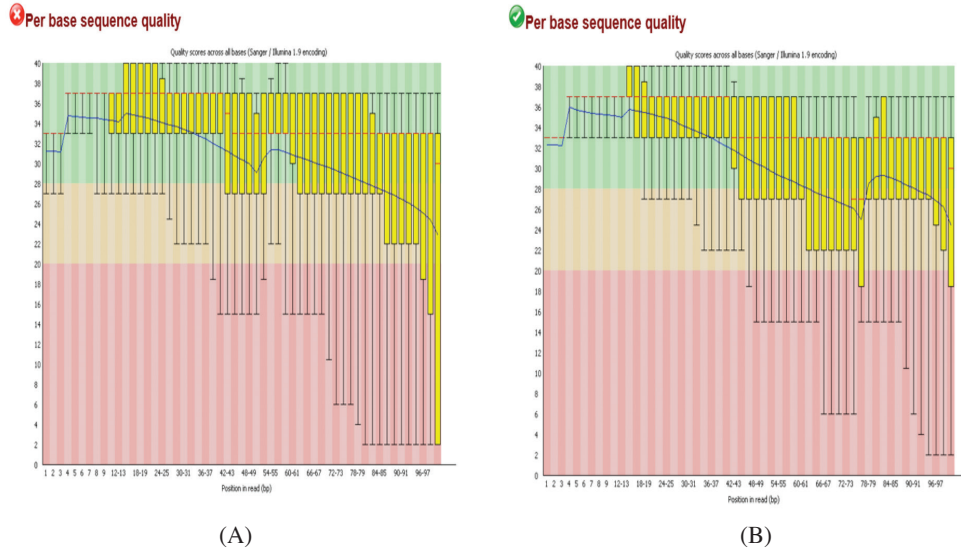


FIGURE 3. Examples of one failed per base sequence quality (A) and one pass per base sequence quality (B)

the end of this step, high quality reads were ready for downstream process.

DE NOVO ASSEMBLY

In parallel with sequencing technologies evolution, algorithms for genomes and transcriptomes assembly are being developed. The NGS-data is different from Sanger sequencing data; it is short DNA sequences instead of large DNA and it is characterized by the high coverage as well. These advantages impose difficulty to analyze this huge data where millions of reads need to be pieced together to re-construct the genome again, in a process called *de novo* assembly. As *de novo* assembly is different from reference genome assembly where the data is compared to known data base, algorithms have been developed

to assemble the genome. The first strategy of *de novo* assembly was based on the overlap-layout-consensus as Atlas and Celera. This algorithm represents each sequence as a node (N) and when an overlap is detected, the algorithm represents it as an arc between the two nodes. Euler assembler is another algorithm which is based on de Bruijn graph (DBG). In this latter algorithm, instead of arranging data on reads it is organized on k-mers (k is number of nucleotides). Velvet the assembler used in the presented work is also based on DBG where Zerbino and Birney (2008) have manipulated the algorithm to eliminate errors and resolve repeats issue imposed in previous assemblers. The k-mer 79 was given by Velvet optimizer to our data. 2900 contigs have been constructed. These contigs were the input of SSPACE algorithm use to

```
>scaffold1|size2838309
TGTCGCCATGCCATAGGGCCACCCGGGTGCCGGAATTTGCTTTTGTACGGCGTCCATGCTGAGTCGCGGATGGCGTTGGCAAGCTTTTACGAGAAGACATAACGAACTCCAGT
ACGGTTACAGTTGAGCGGCCAACATGCTCTCCAGCTTCTGCTGTCGACGGCGAACAGGCGAATGCTTCGGCGAGTTTCTCACCGCCATGGCGTCTGGTTGTCAGCCAGCGGAACCTCGCTTCG
GCCAGCGGAGCGGTTGTGTGAAGCCTTCGTAGACGGGTGAGCTTGCCTCAACCGGCTGCTGCTGTTTTTTCAGCTTCCACGAGGTTCCGGCGGATGTCAGACGATCGCAGCGGCCAGCG
CCAGAATTTGCTCAACCTTGGCAAGCTGGCCCATGATCACCGCTGGTAACGATGACGCTTGTATTCATAGATGGTACGCCAGCTTCCAGCCGGATCTGATCGGCATCGTAGTCGGCGG
CAGGCTGTTAGCCTGATACAGTCATAGATGCGGCGCAGAACGGCGAGATCAGATACACCGCGGCTTCGGCGCAGGCGGGCCGTCGAGCGAAAGAGAACAGCAGCGTCAGGTTGGTTGATGC
CTTCTTTTCCAGCTTTCGGCGGCTTGTATGCTTCCAGGTTGGAGGCCAGCTTGTATCAGAATGCGCGACTTGTGATGCTTGTGATACATGGGATGATGCTGCTTGTGATACATGGGATGATG
GCCGCGATCGAACGACGACGGGCATCCCTCGGTGAGAGATGCGGCCGGCAGCTTTTGGAGGATCTCGACGCCGATGTTGACCGCACTTATCGCTGGGCTTGTACGCTGGGTTCTTGTCTGC
CACCTTGGCGGGGCATATTCAGCGCTTCGGTATCAGCGCTTGTACTGCGGCGAGCGGGCGCTTTCAGGATCAGCGAAGGTTGGTGGTGGCTCTGGGGTTCGAAGTGGCGGATGGATTC
GATATCGCCGCTGTGGCGACCCGCTGCTGAGTGCCTTGAATGATTGATAAGCTCCTGACCAAGCTCAAGTGGGAAATGAGAACCTTTTTCATAGGGGGAACACACTG
CATCATCGGGTCGGGAATGAAATGCGCTTTTCTGATGTTGCGATGGATAAACAATGGATGCCAGTAAACCAGGCGGAGCGGGCTTCGCGCAGGGTAGCGGAAGGAACTGACTGGTG
GGTACCTGAGCGCGGGATGGCGCCCATCTGTAAGTGGCAAGACGGCTCAGGTTCAAAGGAGCGGGTACGCGGCTGAGAAGCGTCCAGTTCGGTTGGTATGAAATAGCCACCT
GGTGTGCTGTGACGCGCGTCACTTCCAGGTAAGGAGTTCGGGCCGTTTCAGTTGAGCTTGTCCAGCGCTGGCGTTGCTGATCCAGCTCGAAGAAAGTGGACTTGTGGCGCTGGCG
ATGTGACCCGTCACCGGGCGGCTCGGGGAAGCTTTCAGGCTTCGACGCTCTGCGGTTCTATTGACGCTCCGCACTGCGTGTGAGCTGCAATTTGCACTGATAGCGCGGCTGGCCGCGA
CTCGACATAACCGTGGCATTGCTGTTGCGAAACCGCAACATGGCCGCTCAGCAGGCCAGAGAGAGAGGTTACGGGAATTTTTCATAAGTCACTCTGACTGAAATATGTTGTAAGT
TGAGCGTATTCACCGCAAAATGAGAGACTAATACGCAATGAAATATCCCTTCTGCTGCGCTTGTATTTATCATTCTGGTGTGCGATAGAGCGGAATGTTTAAAGGCTATTAACGAAAA
TTATTTATGATGACGATCTTGTATCGTTAGGCTGAATTTATTAACGTGCAAGTGTAAAAACCGGGCGTATGATGAAATTTATATCATGATGATGATGATTTAAGATAGCGCTCTCATTTATGAA
GTATATTGCTTGTGTTGAGTACCTTGTGCTGATGATGATTTTGTGCGCACCACTCGGGAATTTAGTGTCAATTTTCTATTTTAAAGAGGAATTTGATGACTAGATATCCCTGGAC
GGGATTAATAATCAGCACCATCAAGAGCACCGGCTTCTCTATGCGCGGAGGGGTTGCATAAAACCGCTTCGCGGATCTCTATCGGGTTTCCAATTTGAAAGCAAACTCTGTGTAAACTT
TTTATCGCAATGGCCCTATGATACCTGACGGATGAAGGGGAGTTTCTCTGCGAAGGACGTTGGATTTAAATCGCGTCAAGATGCGGAAAGCGGGTTCGCAACATCCCAAGCTGGACAA
TAATATCCGCTGGCGTAGACACCTTCTCCCGCTGAAACGCTGACGCGAGGATATCCGCGCATATTAACAATTTGCCGAACCGCAATATCTCGGTGACGCGGGAAGCGTGAACCGCACCTGGG
ATGCGCTGAAGAAACCGGGCGGATCTGATCATCGCATCGGCAAAATCCCGACAGCGTGCAGGCCAAACCTGATGCTCGCAAGCTCAATTTCTGATGTCGCTATCGCTATCCCAACCTTCG
CGGCGAGAAAGCGGTGAGAACGAGACAAAAAAGGGTTACGCGCAGGCTGAACCTCCAGAGATGCCCCCTTTTACGGATTAGTTGCTCAGGTTGACGCTGATCGCTGATAGAAGCGG
TTGGCGGTGTCGGTATGTCACACCGCAAGGATCAGTGCAGACCGCTGCGATCTGCTGATGTCGGTATGCAAGAACGAGCGGCTGAACGACATCGTGGTGGTATCGCC
```

FIGURE 4. Output of scaffolding step with SSPACE. An example of one of the scaffolds with 2,838,209 bp

TABLE 1. Results of Blastp parsed top three. Each CDS has three closest genes represented by hit number 1, 2 and 3

QUERY_ NAME	HIT_ NUM	SCORE	BITS	%_ IDENTITY	E-VALUE	ACC_NUM	SUBJECT	IDENTITY	FRAME	QUERY_ LENGTH	MATCH_ LENGTH	QUERY_ START	QUERY_ END	SUBJECT_ START	SUBJECT_ END
scaffold size 2838309_1	1	174	71.6	100	5.00E-16	WP_072035785.1	hypothetical protein [ <i>Serratia marcescens</i> ]	35	0	35	35	1	35	1	35
scaffold size 2838309_1	2	174	71.6	100	5.00E-16	OJT43914.1	hypothetical protein BSR04_06080, partial [ <i>Serratia phymathica</i> ]	35	0	35	35	1	35	1	35
scaffold size 2838309_1	3	174	71.6	100	6.00E-16	KFB86345.1	ribulose-phosphate 3-epimerase, partial [ <i>Serratia grimesii</i> ]	35	0	35	35	1	35	1	35
scaffold size 2838309_2	1	1675	649	100	0	WP_063988882.1	transaldolase [ <i>Serratia marcescens</i> ]	316	0	316	316	1	316	1	316
scaffold size 2838309_2	2	1672	648	99.68354	0	WP_049211631.1	transaldolase [ <i>Serratia marcescens</i> ]	315	0	316	316	1	316	1	316
scaffold size 2838309_2	3	1671	648	99.68354	0	WP_019453873.1	MULTISPECIES: transaldolase [ <i>Serratia marcescens</i> ]	315	0	316	316	1	316	1	316
scaffold 2139 size 57_1	1	271	108	100	9.00E-28	WP_061551842.1	LysR family transcriptional regulator [ <i>Pseudomonas</i> sp. ABFPK]	52	0	52	52	1	52	13	64
scaffold 2139 size 57_1	2	271	108	100	9.00E-28	WP_063424180.1	LysR family transcriptional regulator [ <i>Pseudomonas putida</i> ]	52	0	52	52	1	52	13	64
scaffold 2139 size 57_1	3	271	108	100	9.00E-28	WP_060499365.1	LysR family transcriptional regulator [ <i>Pseudomonas</i> sp. NBRC	52	0	52	52	1	52	13	64

locate and orient contigs. SSPACE is a powerful tool to scaffold huge data in a short time. 2139 scaffolds have been constructed in our work. An example of the first scaffold found was 2,838,209 bp is shown in Figure 4.

## SEQUENCE ANALYSIS

After scaffolding the CDSs finding is possible with PRODIGAL. This program is designed to identify microbial genes in high sensitivity and avoiding false

```
>scaffold1|size2838309_1 # 2 # 106 # -1 #
ID=1_1;partial=10;start_type=ATG;rbs_motif=GGAG/GAGG;rbs_spacer=5-10bp;gc_cont=0.571
ATGTCTTCTCGTAAAGAGCTTGCCAACGCCATCCGCGCACTCAGCATGGACGCCGTACAAAAAGCAAATT
CCGGCCACCCGGGTGCCCTATGGGCATGGCGGAC
>scaffold1|size2838309_2 # 125 # 1075 # -1 #
ID=1_2;partial=00;start_type=ATG;rbs_motif=AGGAG;rbs_spacer=5-10bp;gc_cont=0.598
ATGAATCAATTAGACGCCTCAAGCAGCTGACCACGGTGGTCGCCGACAGCGGCATATCGAATCCATCC
GCCAGTTCGAACCCAGGACGCCACCACCAACCCTTCGCTGATCCTGAAAGCCGCGCGCTGCCGCAGTA
CAAAGCGCTGATCACCAGCGCTGGAATATGCCCGCGCCAGGGTGGCAGCAAGGAAACCCAGCTGATC
AACGCCAGCGATAAGCTGGCGGTCAACATCGGCGTCGAGATCCTCAAAAGCGTGCCGGGCCGCATCTCCA
CCGAAGTGGATGCCCGTCTGTTCGTCGATCGCGGCATGTGTGTCGCCAAAGCGCGTAAGCTGATCGCCAT
GTATCAAGAGCAAGGCATCGACAAGTCGCGCATTCTGATCAAGCTGGCCTCCACCTGGGAAGGCATCAAA
GCCGCCGAAGAGCTGGAAAAAGAAGGCATCAACACCAACCTGACGCTGCTGTTCTTTCGCTCAGGCC
GCGCCTGCGCCGAGCCGCGTGTATCTGATCTCGCCGTTTCGTCGCCCGCATCTATGACTGGTATCAGGC
TAAACAGCCTGCCCGGACTACGATGCCGATCAGGATCCGGCGTGAAGTCGGTCCGTACCATCTATGAA
TACTACAAGCGTCATCGTTACCAGACGGTGATCATGGCGCCAGCTTCCGCAAGGTTGAGCAAATTCGG
CGCTGGCCGGCTGCGATCGTCTGACCATCGCGCCGAACCTGCTGGAACAGCTGAAAAACAGCGACCAGCC
GGTTGAGCGCAAAGCTGACCCCGTCTACCGAAGGCTTCCACCAACCGGCTCCGCTGGCCGAAGCGGAGTTC
CGCTGGCTGCACAACCAGGACGCCATGGCGGTGGAGAACTCGCCGAAGGCATTCGCCTGTTTCGCCGTCG
ACCAGCAGAAGCTGGAAGACATGTTGGCCGCTCAACTGTAA
```

FIGURE 5. Output of PRODIGAL. Identified CDSs with their location in the scaffolds. In yellow highlight the information about the codon DNA sequence where scaffold is followed by a number, the two examples are from the same scaffold number 1 of 2,838,209 bp. \_1 is the first CDS found in this scaffold, 2 and 106 are the first and last nucleic bases numbers of the CDS

TABLE 2. Twenty one putative genes encoding cellulose-degrading enzymes and 6 putative genes encoding xylan-degrading enzymes from POME metagenomic DNA library

Enzyme	Class of enzyme	Identity percentage (%)	Microorganism
Cellulose-degrading enzymes	cellulase	100.0	<i>Staphylococcus hominis</i>
	glycosyl hydrolase family 5	100.0	<i>Pseudomonas</i> sp. NBRC 111130
	beta-glucosidase	99.7	<i>Serratia marcescens</i>
	endo-1,4-D-glucanase	99.7	<i>Serratia</i> sp. YD25
	endo-1,4-D-glucanase	99.6	<i>Serratia marcescens</i> EGD-HP20
	beta-glucosidase	99.6	<i>Pseudomonas</i> sp. NBRC 111130
	putative glycoside hydrolase	99.6	<i>Staphylococcus hominis</i>
	glycosyl hydrolase family 5	99.0	<i>Pseudomonas</i> sp. NBRC 111130
	glycoside hydrolase	98.8	<i>Serratia</i> sp. YD25
	beta-glucosidase-like glycosyl hydrolase	98.3	<i>Pseudomonas</i> sp. GM84
	cellulase M-like protein	98.3	<i>Streptococcus pneumoniae</i>
	endoglucanase	96.6	<i>Serratia marcescens</i>
	glycosyl hydrolase family 5	93.8	<i>Pseudomonas</i> sp. NBRC 111130
	beta-glucosidase-like glycosyl hydrolase	90.6	<i>Serratia marcescens</i> FG194
	endoglucanase-like subfamily M42 peptidase	86.7	<i>Pseudomonas syringae</i> pv. <i>syringae</i> B64
Xylan-degrading enzymes	beta-glucosidase	84.7	<i>Serratia rubidaea</i>
	endo-1,4-beta-glucanase	79.2	<i>Staphylococcus aureus</i> O46
	endoglucanase M	75.4	<i>Staphylococcus aureus</i>
	beta-glucosidase-like glycosyl hydrolase	68.1	<i>Xanthomonas gardneri</i> ATCC 19865
	endoglucanase	65.4	<i>Phyllobacterium</i> sp. YR531
	endoglucanase	63.3	<i>Pseudomonas aeruginosa</i>
	beta-1,4-xylanase	99.0	<i>Serratia marcescens</i>
	glycosyl hydrolase family 10	99.0	<i>Serratia</i> sp. YD25
	putative xylanase/chitin deacetylase	99.0	<i>Streptococcus pneumoniae</i>
	endo-1,4-beta-xylanase A precursor	92.1	<i>Cronobacter dublinensis</i> 1210
endo-1,4-beta-xylanase A	61.1	<i>Paenibacillus</i> sp. AD87	
endo-1,4-beta-xylanase A precursor	30.3	<i>Enhygromyxa salina</i>	



positives. The CDSs are presented in Word-Format and shown the sequence of the CDS and the location in the scaffolds (Figure 5). The function of each predicted genes has been assigned with blastp vs non-redundant proteins function in BLAST software. Two documents were given by BLAST, first file, blastp-parsed-top-3, presents the top three similar functions of each CDS (Table 1) and the second file, blastp-putativeID file, presents the putative ID of each protein found in the available database. In these results, 21 putative genes encoding cellulose-degrading enzymes and 6 putative genes encoding xylan-degrading enzymes have been identified identical to previously known genes (Table 2). These genes show different identity percentages from 30.3% to 100%. From a bioinformatician view point, one different nucleotide within the gene sequence makes it novel while some of the experts find that differences in hot spots or conserved domains regions may make the gene or the enzyme novel. From a scientist view point (Uwe T. Bornscheuer, personal communication) 'some scientists claim that if a paper has the phrase 'a novel enzyme', this translates into 'yet another enzyme'; for instance, if hundreds of glycosyl hydrolases have been described, many of these enzymes act on the same sugars with the same substrate specificity and cleavage site/preference. Strictly, that means: a novel enzyme must be an enzyme catalyzing a novel chemical reaction. This is however a rather rare event. Most scientists accept that a novel enzyme could also be an enzyme differing substantially from the well-described ones, e.g. lipase CAL-A differs substantially from lipase CAL-B (and most other lipases), so it is rather novel, i.e. thermostable, can catalyze acyltransfer in water and accepts tertiary alcohols. If our glycosyl hydrolases work at a totally different pH optimum, that could be interesting, is any cleaves a different bond in sugars, then for sure it is rather novel. If the sequence, structure and properties are very close to known enzymes, then this is not a 'novel enzyme'.

#### CONCLUSION

In NGS-data analysis, choosing the right work flow and pipeline is the most crucial part to succeed the work and to achieve the aim. In this work, the objective was to decode the genetic code and simplify the complicated raw data of NGS to come out with cellulose and xylan-degrading enzymes in POME metagenome and this was achieved and 27 (21 + 6) of known and potentially novel enzymes are found in our 30 positive fosmid clones. Our huge data is still holding a lot of other enzymes or other proteins that may be identified in the future.

#### ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Higher Education, Malaysia for financial support through

Fundamental Research Grant Scheme (grant no. FRGS 13-086-0327) and the Biotechnology Engineering Department, IUM for scientific equipment and facilities and the Malaysia Genome Institute (MGI) for technical support. We would also like to thank Prof. S.G. Withers (University of British Columbia, Canada) for sharing the original idea of the research project and the gift of a substrate used in the work and Prof. Dr. Uwe T. Bornscheuer from Ernst-Moritz-Arndt-Universität Greifswald, Biotechnology and Enzyme Catalysis Group, Greifswald, Germany, for consultation and valuable suggestions in the preparation of this manuscript.

#### REFERENCES

- Armstrong, Z., Mewis, K., Strachan, C. & Hallam, S.J. 2015. Biocatalysts for biomass deconstruction from environmental genomics. *Current Opinion in Chemical Biology* 29(18): 18-25.
- Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4): 578-579.
- Cox, M.P., Peterson, D.A. & Biggs, P.J. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11(1): 485.
- Henson, J., Tischler, G. & Ning, Z. 2012. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 13(8): 901-915.
- Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11(1): 119.
- Kumar, S., Krishnani, K.K., Bhushan, B. & Brahmane, M.P. 2015. Metagenomics: Retrospect and prospects in high throughput age. *Biotechnology Research International* 2015: 1-13.
- Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B. & Fan, W. 2012. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics* 11(1): 25-37.
- Mewis, K., Armstrong, Z., Song, Y.C., Baldwin, S.A., Withers, S.G. & Hallam, S.J. 2013. Biomining active cellulases from a mining bioremediation system. *Journal of Biotechnology* 167(4): 462-471.
- Minakshi, P., Ranjan, K., Brar, B., Ambawat, S., Shafiq, M., Alisha, A., Kumar, P., Ganesharao, J.V., Jakhar, S., Balodi, S., Singh, A. & Prasad, G. 2014. New approaches for diagnosis of viral diseases in animals. *Advances in Animal and Veterinary Sciences* 2(4S): 55-63.
- Taupp, M., Mewis, K. & Hallam, S.J. 2011. The art and design of functional metagenomic screens. *Current Opinion in Biotechnology* 22(3): 465-472.
- Zerbino, D.R. & Birney, E. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* 18(5): 821-829.

Farah Fadwa Benbelgacem, Oualid Abdelkader Bellag, Adibah Parman & Ibrahim Ali Noorbataha  
Bioprocess & Molecular Engineering Research Unit (BPMERU)  
Department of Biotechnology Engineering

Kulliyyah of Engineering  
International Islamic University Malaysia  
Jalan Gombak, 53100 Kuala Lumpur, Federal Territory  
Malaysia

Mohd Noor Mat Isa  
Malaysia Genome Institute  
Jalan Bangi, 43000 Kajang, Selangor Darul Ehsan  
Malaysia

Muhammad Alfatih Muddathir Abdelrahim & Afidalina Tumian  
Department of Computer Science  
Kulliyyah of Information and Communication Technology  
International Islamic University Malaysia  
Jalan Gombak, 53100 Kuala Lumpur, Federal Territory  
Malaysia

Hamzah Mohd Salleh\*  
International Institute for Halal Research and Training (INHART)  
International Islamic University Malaysia  
Jalan Gombak, 53100 Kuala Lumpur, Federal Territory  
Malaysia

\*Corresponding author; email: hamzah@iium.edu.my

Received: 30 May 2018

Accepted: 18 September 2018