

FUNCTIONAL METAGENOMIC APPROACH IN
IDENTIFY CELLULOSE-DEGRADING ENZYMES
FROM MALAYSIAN PALM OIL MILL EFFLUENT

BY

FARAH FADWA BEN BELGACEM

A thesis submitted in fulfilment of the requirement for the
degree of Doctor of Philosophy of Engineering
(Biotechnology Engineering)

Kulliyah of Engineering
International Islamic University Malaysia

OCTOBER 2020

ABSTRACT

Major source of enzymes is microbes and during cultivation step, more than 85% of them resist cultivation and their genetic patrimony is loss. This work sought to bioprospect cellulose-degrading enzymes from microbiota of the palm oil mill effluent (POME). To get access to this great microbial diversity and to discover the biocatalysts behind, this research has adopted metagenomic approach which technically escapes the cultivation step and screens the microbial DNA for desired enzymatic activity. Metagenomics consists of the creation and screening of metagenomic DNA libraries. *In vivo* identification of cellulose-degrading enzymes was carried out with high-throughput screening and *in silico* identification of genes encoded the enzymes was performed with algorithm-based methods while the results validation was executed by recombinant enzymes expression, purification and characterisation. Culture-enrichment strategy based on natural selection principle was used in the early stage to enhance the screening hit rate. Metagenomic DNA was extracted from enriched and non-enriched sample to construct 109,824 fosmids (4.49 Gb) metagenomic DNA library. In this library, pCC1FOS fosmid and *E. coli* EPI300T1^R are the vector and surrogate host of the cloning system, respectively. A high throughput functional metagenomic screen was developed and applied to search for cellulose-degrading enzymes within the library clones using 4-methylumbelliferyl- β -D-glucopyranoside (MUGlc) and 4-methylumbelliferyl- β -D-cellobioside (MUC) fluorogenic substrates. The screens were normalised using robust z-score and highest rated clones (100) were then selected. Their fosmids were isolated and sequenced with Hiseq (Illumina) of next-generation sequencing strategy. For quality control of the reads, SolexaQA and FastQC tools were used. Poor quality bases were removed with DynamicTrim algorithm, all bases with Qphred less than 20 were trimmed, and the LengthSort algorithm was used to remove sequences less than 50 bp. *de Bruijn* graph of *de novo* assembly algorithm has organised the reads on *k*-mers to build contigs, and Velvet optimiser has selected the optimum *k*-mers. These contigs were the input of SSPACE algorithm used to locate and orient contigs. Codon DNA sequences (CDS) were identified with PRODIGAL software. The genes identification was carried out following Blastp and SmartBLAST. Seventeen of bioprospected putative cellulose-degrading enzymes were cloned into pBAD-TOPO plasmid and expressed in TOP10 *E.coli* cloning. Enzymes were purified with HisTrap HP column with the aid of a FPLC system. Two putative glucanases and two putative β -glucosidases were then biochemically characterised for optimal pH and temperature in the presence of substrates MUGlc and MUC, *p*NPG and *p*NPC and CMC as well. In NGS-data analysis step, 4900 contigs and 3540 scaffolds were constructed. 42,247 CDS were detected and 96 potential cellulose-degrading enzymes were identified which evinces the richness of POME metagenome on biocatalysts. The protein sequences of 15 cellulose-degrading enzymes are 100% similar to protein sequences available in protein databases while 40 enzymes show (80-99%) similarities, 24 enzymes (60-79%), 14 enzymes (40-59%) and 3 enzymes show less than 40% similarity, this reflects the qualification of functional metagenomics to bioprospect untapped enzymes. The potential types of enzymes are 19.20% glucanases, 31.32% glucosidases and 46.48% glucoside hydrolases with cellulose-degrading enzyme conserved domains. For the 17 expressed enzymes, three different glycoside hydrolase

families, (enzyme 1, 2, 10 and 21) from GH3, (enzyme 6, 12 and 20) from GH5, (enzyme 3) from GH8 and other glycoside hydrolase families. Enzyme 3 is probably an example of untapped enzyme; it was active toward MUC and MUGlc. The optimum catalysis activity by enzyme 3 occurred at 50 °C and pH 4. For enzymes 4, 11 and 13, no enzymatic activity was detected due to low expression level. This research was very challenging but rewarding. It lays the foundation of diverse and untapped biocatalysts discovery. The bioprospected enzymes found in this metagenomic DNA library can be produced and optimised to be used in different industrial applications. In addition, the NGS-analysed-data can be used to study the diversity of POME.

ملخص البحث

الميكروبات هي المصدر الاساسي للانزيمات، وخلال مرحلة الاحياء، أكثر من 85 بالمئة من هذه الميكروبات تقاوم العيش في المختبر وبالتالي نفقد المعلومة الجينية الخاصة بها. سعى هذا البحث الى ايجاد انزيمات مفككة للسيليلوز من مجموع الميكروبات المتواجدة في مخلفات مصانع زيت النخيل. للتوصل الى هذا التنوع الميكروبي الهائل واكتشاف المحفزات الحيوية، تبنت هذه الدراسة منهج الميتاجينوميك الذي يتخلى عن خطوة الاحياء ويفحص الحمض النووي للانزيمات. يتضمن منهج الميتاجينوم انشاء وفحص مكتبات الحمض النووي. الكشف عن الانزيمات المفككة للسيليلوز *In vivo* تم بواسطة الفحص عالي المردودية، والكشف عن الجينات الأصلية لهذه الانزيمات *In silico* تم بوسائل لوغاريتمية، كما أن التؤكد من النتائج تم عن طريق انتاج الانزيمات، تنقيتها، ثم دراسة خصائصها. عملية تغذية الوسط المعتمدة على الانتقاء الطبيعي، استعملت في بادئ الأمر لتدعيم نتائج الفحص. استخلص الحمض النووي منقوص الريبوزوم من كل من العينات المدعمة والغير مدعمة لانشاء مكتبة ميتاجينومية عددها 109824 فوسميد (4.49 جيجابايت) ، في هذه المكتبة pCC1FOS هو الفوسميد الحامل وEPI300T1^R هي الخلية المضيفة لعملية الاستنساخ. الفحص الميتاجينومي العالي المردودية للبحث عن هذه الانزيمات، أستخدم كل من MUC و MUGlc كمادتين فاعليتين مفسفرتين. نتائج الفحص عدلت عن طريق حساب Robust z-score. من خلال هذه النتائج تم اختيار اعلى مئة خلية في الترتيب. في المرحلة الثانية، استخلصت فوسميدات هذه الخلايا المئة وتم قراءة حمضها النووي منقوص الريبوزوم بجهاز Hiseq (Illumina) الذي يعتبر من فئة الجيل الثاني. استخدم SolexaQA و FastQC كوسيلتين لترتيب البيانات. التسلسلات ذات الجودة الأقل من اللازمة يتم التخلص منها بواسطة وسيلة DynamicTrim algorithm والسلاسل مع اقل من Qphred20 تم التخلص منها مع السلاسل القصيرة أقل من 50bp عن طريق LengthSort algorithm. De Bruijn graph هو اللوغاريتم المستعمل لاعادة تسلسل الحمض النووي المنقوص الريبوزوم (de novo assembly) الى Kmers ومنها بناء Contigs. أما Velvet Optimiser عمل على اختيار ال Kmer الأمثل. SSPACE لوغاريتم

عمل على ترتيب ال Contigs وتوجيهها. سلسلة الكودون للحمض النووي منقوص الريبوزوم تم تحديدها عن طريق برنامج PRODIGAL. التعرف على الجينات بواسطة Blastp و SmartBLAST. سبعة عشر من الانزيمات المحتملة استنسخت في البلاسميد pBAD-TOPO للخلية البكتيرية TOP10 E.coli. تم تصفية الانزيمات بتقنية FPLC بواسطة HisTrap HP column. أنزيمان من نوع جلوكانيز وأنزيمان من نوع البيتا جلوكوزيديز تم تحديد خصائصهما البيوكيميائية من ال pH الأمثل ودرجة الحرارة المثلى في حضور كل من المواد المتفاعلة المفسرة MUC و MUGlc والمواد المتفاعلة الملونة pNPC و pNPG والمادة المتفاعلة CMC أيضا. من نتائج قراءة التسلسل بواسطة الجيل الجديد، شكلنا 4900 contigs و 3540 scaffolds. تم اكتشاف 42247 CDS منها 96 أنزيم مفكك للسيليلوز، مما يؤكد وفرة الانزيمات في عينة البحث POME. أوضحت نتائج التسلسل البروتيني أن السلاسل البروتينية ل 15 أنزيم مطابقة 100% لبروتينات متواجدة في قاعدة البيانات، 40 أنزيم مطابق بنسب تتراوح بين (80-99%)، 24 أنزيم مطابقة بنسب تتراوح بين (60-79%)، 14 أنزيم مطابق بنسب (40-59%) وثلاث أنزيمات مطابقة بأقل من 40%، وهذا ما يؤكد أيضا كفاءة الميتاجينوميك في اكتشاف الانزيمات الغير معروفة مسبقا. الانزيمات المكتشفة هي 12.20% جلوكانيز، 31.32% جلوكوزايديز، و 46.48% من مختلف أنواع الانزيمات المفككة للسيليلوز. من هذه الانزيمات 17 انزيم غير مكتشف تم تجربتها في المختبر، تنتمي الانزيمات 1، 2، 10 و 21 لعائلة GH3، وتنتمي الانزيمات 6، 12 و 20 لعائلة GH5 وينتمي الانزيم 3 لعائلة GH8. الانزيم 3 يحتمل أن يكون مثالا عن الانزيمات الغير مكتشفة سابقا حيث أنه أبدى فاعليته اتجاه المادتين الفعالتين MUC و MUGlc الفاعلية المفككة المثلى لهذا الانزيم 3 كان في درجة حرارة 50 درجة مئوية و pH. أما فيما يخص الانزيمات 4 و 11 و 13 فلم يتم رصد تفاعل كيميائي نظرا لمستوى الانتاج الضئيل. هذا البحث يعد صعبا ولكن مجزيا. فهذا يعد حجر أساس لاكتشاف العديد من الانزيمات المكتشفة والغير مكتشفة مسبقا. الانزيمات المكتشفة في هذه المكتبة الميتاجينومية يمكن تصنيعها وتحسينها لاستخدامها في صناعات مختلفة. كما أن البيانات المتحصل عليها من قراءة التسلسل للحمض النووي منقوص الريبوزوم تعتبر مادة غنية لدراسة التنوع في عينة POME.

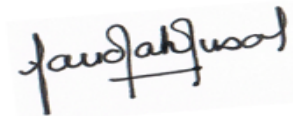
APPROVAL PAGE

The thesis of Farah Fadwa Ben Belgacem has been approved by the following:

Hamzah Mohd. Salleh
Supervisor

Ibrahim Ali Noorbacha
Co-Supervisor

Md. Zahangir Alam
Co-Supervisor



Faridah Yusof
Internal Examiner

Muhammad Mukram bin Mohamed Mackeen
External Examiner

Amir Feisal Merican
External Examiner

Fouad Mahmoud Mohamed Rawash
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Farah Fadwa Ben Belgacem

Signature

Date.....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

**DECLARATION OF COPYRIGHT AND AFFIRMATION OF
FAIR USE OF UNPUBLISHED RESEARCH**

**BIOPROSPECTING OF CELLULOSE-DEGRADING
ENZYMES FROM MALAYSIAN PALM OIL MILL EFFLUENT
BY FUNCTIONAL METAGENOMIC APPROACH**

I declare that the copyright holder of this thesis jointly owned by the student and
IIUM.

Copyright © 2020 Farah Fadwa Ben Belgacem and International Islamic University Malaysia. All
rights reserved.

No part of this unpublished research may be reproduced, stored in a retrieval
system, or transmitted, in any form or by any means, electronic, mechanical,
photocopying, recording or otherwise without prior written permission of the
copyright holder except as provided below

1. Any material contained in or derived from this unpublished research
may be used by others in their writing with due acknowledgement.
2. IIUM or its library will have the right to make and transmit copies (print
or electronic) for institutional and academic purposes.
3. The IIUM library will have the right to make, store in a retrieved system
and supply copies of this unpublished research if requested by other
universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM
Intellectual Property Right and Commercialization policy.

Affirmed by Farah Fadwa Ben Belgacem

.....
Signature

.....
Date

ACKNOWLEDGEMENTS

Firstly, I thank Allah, the almighty, for giving me the strength to carry on this research and for blessing me with many great people who have been my greatest support in both my personal and professional life.

I would like to seize this opportunity to express my deepest regards and gratitude to whom I dedicate this work, my dear father Mohammed Ben Belgacem, my lovely mother and my mother-in-law Zoura et Hanou, my beloved husband Oualid Abdelkader Bellag and my little angle Mohamed Khaled Alwalid, my pretty sisters Fairouz, Imen and Nesrine, my brothers Abdelkader and Akram, and to all my family members and friends who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

Finally, a special thanks to my supervision comity, Prof. Dr. Hamzah Mohd. Salleh, Prof. Dr. Ibrahim Ali Noorbacha and Prof. Dr. Md. Zahangir Alam for their continuous support, encouragement and leadership and to all the lecturers and researchers in IIUM, and for that, I will be forever grateful. I also would like to thank Prof. S.G. Withers (University of British Columbia, Canada) for kindly providing “chlorocoumarin xylobioside” needed for this project.

TABLE OF CONTENTS

Abstract	i
Arabic Abstract	iii
Approval Page.....	v
Declaration	vi
Copyright	vii
Acknowledgements.....	viii
List of Tables	xii
List of Figures	xiii
List of Abbreviations	xviii
CHAPTER ONE: INTRODUCTION	1
1.1 Background of the Study	1
1.2 Problem Statement and Research Significance	4
1.3 Research Objectives.....	6
1.4 Research Scope	6
1.5 Research Philosophy.....	7
1.6 Research Methodology	8
1.7 Thesis Organisation	11
CHAPTER TWO: LITERATURE REVIEW	12
2.1 Strategies to Obtain Enzymes With Unique Properties	12
2.1.1 Bioprospecting Existing Enzymes	14
2.1.2 Protein Engineering of Novel Enzymes.....	16
2.1.3 Database Search	19
2.1.4 <i>De novo</i> Design.....	20
2.2 Metagenomics, Novel and Powerful Method of Enzymes	
Bioprospecting	21
2.2.1 Principles of Metagenomics.....	28
2.2.2 Types of Metagenomics	29
2.2.2.1 Function-driven Analysis	29
2.2.2.2 Sequence-driven Analysis	30
2.2.3 Enrichment Strategy as Metagenomics Strengthened Tool	31
2.2.4 CopyControl Cloning System	33
2.3 High-Throughput Screening and Next Generation Sequencing as	
Complementary Methods for Enzymes Bioprospecting	37
2.3.1 Microtiter High-Throughput Screening for Enzyme Discovery ...	37
2.3.2 High-throughput Screening Data Analysis	39
2.3.2.1 Normalisation for Assay Variability.....	40
2.3.2.2 Normalisation for Systematic Errors	44
2.3.3 NGS-data Analysis for Gene Finding	46
2.3.3.1 Genome Assembly.....	48
2.3.3.2 De novo Genome Assembly: De Bruijn Graph	49
2.3.3.3 Removing Bubbles with the Tour Bus Algorithm.....	51
2.4 Cellulose-Degrading Enzymes	53

2.4.1 Endoglucanases	56
2.4.2 Exoglucanase	56
2.4.3 β -Glucosidase.....	57
2.4.4 Characterisation of Cellulose-degrading Enzymes.....	57
2.4.4.1 Temperature and pH	58
2.4.4.2 Substrate Specificity	62
2.5 Summary	65

CHAPTER THREE: MATERIALS AND RESEARCH METHODOLOGY 67

3.1 Introduction.....	67
3.2 Materials	69
3.2.1 Sample Collection	69
3.2.2 Chemicals and Reagents	70
3.2.3 DNA Extraction, Cloning and DNA Purification Kits	70
3.2.4 Sequencing Services	70
3.3 Methods	70
3.3.1 Metagenomic DNA Libraries Creation.....	70
3.3.1.1 POME Samples Enrichment.....	71
3.3.1.2 Metagenomic DNA Extraction.....	72
3.3.1.3 Fosmid CopyControl Cloning	74
3.3.1.4 Organization of metagenomic DNA library	81
3.3.2 High-Throughput Screening	81
3.3.2.1 Replication of Metagenomic DNA Library.....	83
3.3.2.2 Addition of screening assay mix	84
3.3.2.3 Fluorescence Reading of Assayed Clones.....	84
3.3.2.4 Data Normalisation and Results Representation	85
3.3.3 Fosmid DNA Isolation, NGS and Sanger Sequencing	85
3.3.3.1 Fosmids Isolation and DNA-Quality Control.....	85
3.3.3.2 Fosmid-ends Sanger Sequencing.....	87
3.3.3.3 D-pooled PCR.....	89
3.3.4 NGS-data Analysis for Genes Finding.....	90
3.3.4.1 NGS-data Analysis Pre-Processing	92
3.3.4.2 De Novo Genome Assembly and Scaffolding.....	94
3.3.4.3 Gene Prediction and Annotation.....	97
3.3.5 Enzymes Production, Purification and Biochemical Characterisation	98
3.3.5.1 Genes Expression Cloning and Enzymes Production with pBAD TOPO® TA Expression System	98
3.3.5.2 Enzymes Purification with Column Chromatography.....	111
3.3.5.3 Biochemical Characterisation of Purified Enzymes	113
3.4 Summary	115

CHAPTER FOUR: RESULTS AND DISCUSSION 118

4.1 Introduction.....	118
4.2 Enriched and Non-enriched POME MetagenomIC LIBRARIES	118
4.3 High-throughput Screening of Cellulose-Degrading Enzymes Discovery	122
4.3.1 Library Preparation and Screening	122
4.3.2 High-throughput Data Normalisation	125

4.3.2.1 Data normalisation Based On De Facto Negative Control.....	126
4.3.2.2 Data Normalisation Based on Median Absolute Deviation (MAD)	128
4.3.3 Selection of High-Rated Clones.....	129
4.4 Fosmid Isolation and Sequencing	132
4.4.1 Clones Auto-Induction and Fosmid Isolation	132
4.4.2 NGS and Sanger Sequencing	133
4.4.3 3D-pooled PCR.....	135
4.5 NGS-data Analysis	136
4.5.1 Quality Control and Data Filtering	136
4.5.2 <i>De Novo</i> Assembly into Contigs.....	140
4.5.3 Contigs Organisation into Scaffolds	141
4.5.4 PRODIGAL for Codon DNA Sequences (CDS) Finding.....	142
4.5.5 Blastp Against Non-Redundant Proteins for Gene Annotation	143
4.5.6 Selected Cellulose-degrading Enzymes	149
4.6 Cellulose-degrading Enzyme Expression and Characterisation.....	161
4.6.1 Cellulose-degrading Activity Assay on LB Agar-CMC.....	161
4.6.2 Cellulose-degrading Activity Assay with Fluorogenic Substrates.....	162
4.6.3 Characterisation of Cellulose-Degrading Enzymes	176
4.7 Summary.....	181
CHAPTER FIVE: CONCLUSION AND RECOMMENDATION.....	184
5.1 Conclusion	184
5.2 Main Contribution of the Study	186
5.3 Recommendation	186
BIBLIOGRAPHY	188
APPENDIX A	205
APPENDIX B	212
APPENDIX C	214
APPENDIX D	215
APPENDIX E	220
APPENDIX F	240
APPENDIX G.....	250
APPENDIX H.....	251

LIST OF TABLES

Table 2.1 Properties of selected metagenomic glycosyl hydrolases enzymes	25
Table 2.2 Summary of High Throughput Screening data normalisation methods	43
Table 3.1 Meta-G-Nome™ DNA Isolation Kit (Epicentre, USA) Contents	72
Table 3.2 Reading parameters of Tecan i-control of microplate reader Tecan infinite F200 Pro used in the high-throughput screening.	84
Table 3.3 Forward and reverse primers of the genes encoding putative cellulose-degrading enzymes.	100
Table 3.4 One Factor at Time (OFAT) design to optimise PCR based on three factors, magnesium ions concentration, annealing temperature and template DNA dilution	105
Table 3.5 Cloning reaction TOPO® reagents and preparation	107
Table 3.6 Optimisation of L-arabinose concentration following pBAD-TOPO expression protocol	110
Table 4.1 Relative fluorescence reading of 4-methylumbelliferone assay.	125
Table 4.2 Summary of <i>de novo</i> assembly, scaffolding and genes prediction output	143
Table 4.3 Results of blastp parsed top three.	145
Table 4.4 Summary of 17 selected cellulose-degrading enzymes from metagenomic DNA library	158
Table 4.5 Summary of enzymes purification and fluorogenic substrates assay of the selected putative cellulose-degrading enzymes	163

LIST OF FIGURES

Figure 1.1 Summary of some key research methods and their descriptions.	10
Figure 2.1 Multi-parameter footprint analysis (Lorenz & Eck, 2005). k_{cat} , catalytic reaction rate; k_{cat} , catalytic constant; K_m , Michaelis constant; U, unit.	13
Figure 2.2 Overview of strategies of enzymes identification or engineering with specific properties (Davids et al., 2013).	15
Figure 2.3 Workflow of functional metagenomic approach (Voget et al., 2003).	31
Figure 2.4 Functional metagenomic approach in the presence and absence of enrichment strategy and the effect of enrichment on enhancing the high-throughput screening results	33
Figure 2.5 pCC1FOS™ Vector Map (Epicentre Company, United States)	35
Figure 2.6 Overview of the CopyControl™ System (Epicentre Company, United States)	36
Figure 2.7 Production of a CopyControl™ Fosmid library and subsequent induction of clones to high-copy number (Epicentre Company, United States)	36
Figure 2.8 Fluorogenic substrates used in high-throughput screening of β -glucosidases and 1,4- β -glucanases.	38
Figure 2.9 schematic representation of the <i>de Bruijn</i> graph (Miller, 1959).	49
Figure 2.10 Example of Tour Bus error correction.	52
Figure 2.11 Structure of lignocellulose. In green is cellulose, in yellow is hemicellulose and in red in lignin (Henrissat, 1991)	55
Figure 2.12 Chemical structure (without showing hydroxyl groups) and enzymatic hydrolysis of glycosidic linkage. (each chemical unit is a glucose).	56
Figure 2.13 A bell-shaped curve showing the dependence of k_{cat}/K_M upon pH for hydrolysis of <i>o</i> -nitrophenyl β -xylobioside by <i>Bacillus circulans</i> xylanase at 25 °C, fit to the kinetic expression for two ionisable groups with apparent pK_a values of 4.6 and 6.7 (McIntosh et al., 1996).	59

Figure 2.14 A typical graph of enzyme activity as a function of temperature showing the apparent “optimum temperature” of an enzyme-catalysed reaction (Daniel et al., 2008).	61
Figure 2.15 The Equilibrium Model to explain the effects of temperature on enzymes. Eact is the active form of the enzyme, which is in equilibrium with the inactive form, Einact. Keq is the equilibrium constant describing the ratio of Einact/Eact whereas <i>kinact</i> is the rate constant for the Einact to X reaction; and X is the irreversibly inactivated form of the enzyme. The equation is not intended to imply that the conversion from Einact to X is a single step, or that X is a single species (Daniel et al., 2008).	62
Figure 3.1 Overview of experimental study	68
Figure 3.2 Three sites of the collected POME samples. A: fresh POME, B: cooled POME and C: anaerobic POME.	69
Figure 3.3 Principle steps of metagenomic DNA libraries construction from POME samples with and without enrichment.	71
Figure 3.4 Metagenomic DNA cloning to pCC1FOS and transformed into EPI300-T1 ^R host cell	75
Figure 3.5 Size selection of end-repaired metagenomic DNA with 1% LMP gel electrophoresis comparing to fosmid control DNA.	77
Figure 3.6 Workflow of metagenomic DNA library high-throughput screening for cellulose-degrading activity.	82
Figure 3.7 Workflow of the HTS of the metagenomic DNA libraries.	82
Figure 3.8 Fosmids DNA isolation and NGS sequencing	86
Figure 3.9 Nucleotide sequence of pCC1FOS from base 230 to 501.	88
Figure 3.10 Fosmid DNA sequencing with NGS and Sanger sequencing strategies.	88
Figure 3.11 DNA library screening with 3D-pooled PCR according to Ferrar and Dannison (2007)	90
Figure 3.12 DNA library screening with 3D-pooled PCR according to Intact Genomics Service	90
Figure 3.13 NGS-data analysis workflow of 30 recombinant fosmid DNA. Pre-processing, <i>de novo</i> assembly and sequence analysis are the main steps of the presented work pipeline. The bioinformatic tools used in this pipeline are shown in bold letters.	92

Figure 3.14 Outline of TOPO® cloning and expression of genes	99
Figure 3.15 Polymerase chain reaction parameters for the first group of genes with <i>Taq</i> DNA polymerase.	103
Figure 3.16 Polymerase chain reaction parameters for the second group of genes with <i>Taq</i> DNA polymerase.	104
Figure 3.17 Polymerase chain reaction parameters for the third group of genes with <i>Taq</i> DNA polymerase.	104
Figure 3.18 Polymerase chain reaction parameters for the fourth group of genes with <i>Taq</i> DNA polymerase.	105
Fig 3.19 pBAD-TOPO ^R vector map.	108
Figure 3.20 Flow chart of recombinant enzymes expression for FPLC purification	111
Figure 3.21 Flow-chart of designed strategy to troubleshoot the purification in case of experiment failure.	113
Figure 4.1 Metagenomic DNA concentration of non-enriched DNA extracted directly after sampling and from enriched samples after exposition step.	119
Figure 4.2 Large Petri-dishes (150 mm × 15 mm) of transformed <i>E. coli</i> EPI300-T1R plated on LB agar with 12.5 µg/mL chloramphenicol.	120
Figure 4.3 (A). Scatter plot of high-throughput screened libraries represented in relative fluorescence unit. (B). Scatter plot of high-throughput screened libraries after data normalisation represented in robust z -score.	127
Figure 4.4 (A) HTS-results of plate AE162 represented in z -scores based on mean and standard deviation. (B) HTS-results of plate AE162 represented in robust z -scores based on mean and standard deviation.	129
Figure 4.5 Scatter of robust z -score of high-throughput screened libraries. Red spots present the high rated clones chosen for sequencing	131
Figure 4.6 Example of DNA sequencing results (pass, accepted).	134
Figure 4.7 Example of DNA sequencing results (rejected).	135
Figure 4.8 One of the FastQC reports of per base sequence quality, k -mer content, per sequence quality score, sequence length distribution, per sequence GC content, per base sequence content, per base GC percentage, per base N-content, sequence duplication level	139

Figure 4.9 Examples of one failed per base sequence quality (A) and one passed per base sequence quality (B)	140
Figure 4.10 Output of scaffolding step with SSPACE. An example of one of the scaffolds with 2,838,209 bp	141
Figure 4.11 Output of PRODIGAL. Identified CDSs with their location in the scaffolds.	142
Figure 4.12 Cellulose-degrading enzymes found in the metagenomic libraries	148
Figure 4.13 Layout of the page reporting the conserved domains in gene 1.	153
Figure 4.14 Layout of the page reporting concise summary of the five best matches of gene 1.	154
Figure 4.15 Layout of the page reporting the conserved domains in gene 13	155
Figure 4.16 Layout of the page reporting concise summary of the five best matches of gene 13.	155
Figure 4.17 Layout of the page reporting sequences producing significant alignments of gene 13.	156
Figure 4.18 CMC-LB-agar plates of positive cellulose-degrading clones.	162
Figure 4.19 FPLC chromatogram of enzyme 3.	169
Figure 4.20 Fluorogenic substrates assay of enzyme 3.	170
Figure 4.21 FPLC chromatogram of enzyme 4.	171
Figure 4.22 Fluorogenic substrates assay of enzyme 4.	172
Figure 4.23 FPLC chromatogram of enzyme 11.	173
Figure 4.24 Fluorogenic substrates assay of enzyme 11.	173
Figure 4.25 FPLC chromatogram of enzyme 13.	174
Figure 4.26 fluorogenic substrates assay of enzyme 13.	175
Figure 4.27 Standard curve of different concentration of 4-MU fluorescence.	176
Figure 4.28 Incubation time monitoring for enzymes 3, 4, 11 and 13 assays. (A) is the case of enzyme 3 in the presence of MUC substrate. (B) is the case of enzyme 4 in the presence of MUC substrate. (C) is the case of enzyme 11 in the presence of MUGlc substrate. (D) is the case of enzyme 13 in the presence of MUGlc substrate.	178

Figure 4.29 Standard curve of different concentration of <i>p</i> NP.	179
Figure 4.30 Temperature dependence of activity of purified enzyme 3 with <i>p</i> NPC at indicated temperatures.	179
Figure 4.31 Enzyme 3 activity on <i>p</i> NPC at different pH.	180
Figure 4.32 fluorogenic substrate assay of eleven elution fractions of purified enzyme 3.	181

LIST OF ABBREVIATIONS

ATA	Amine transaminase
ASRA	Adaptive substituent reordering algorithm
BACs	bacterial artificial chromosomes
BLAST	Basic local alignment search tool
BG	β -Glucosidase
CAZy	Carbohydrate-active enzymes
CBH	Cellobiohydrolase
CMC	Carboxyl methyl cellulose
DBG	<i>de Bruijn</i> graph
DI	Digital imaging
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
DMSO	Dimethyl sulfoxide
EDTA	Ethylenediaminetetraacetic acid
EtBr	Ethidium bromide
EpPCR	Error-prone PCR
FACS	Fluorescent activated cell sorting
FPLC	Fast protein liquide chromatography
GH	Glycosyl Hydrolase
GFP	Green fluorescent protein
HTS	Hight-Throughput Screening
IVTC	<i>In vitro</i> compartmentalization
KCl	Potassium chloride
K ₂ HPO ₄	Potassium hydrogen phosphate
MAD	Median absolute deviation
MEGAWHOP	Megaprimer PCR of whole plasmid
MgCl ₂ .6H ₂ O	Magnesium chloride hexahydrate
MgSO ₄ .7H ₂ O	Magnesium sulfate heptahydrate
MUGlc	4-methylumbelliferyl- β -D-glucopyranoside
MUC	4-methylumbelliferyl- β -D-cellobioside
NaCl	Sodium chloride
Na ₂ CO ₃	Sodium carbonate
NGS	Next generation sequencing
<i>p</i> NPC	4-Nitrophenyl β -D-cellobioside
<i>p</i> NPG	4-Nitrophenyl β -D-glucuronide
OD	Optical density
PBD	Protein data bank
PCR	Polymerase chain reaction
PLP	Pyridoxal-5'-phosphate
POME	Palm oil mill effluent
QSAR	Quantitative structure-activity relationship
RBS	Ribosome binding site
RNase	Ribonuclease
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
TAE	Tris-acetate-EDTA

TE
UV

Tris-EDTA
Ultraviolet

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Various industries have utilised enzymes as part of their production due to their high selectivity, efficiency and rapidity in catalytic reactions. By using enzyme in industrial processes, it is possible to eliminate or at least prevent possible hazards caused by extreme parameters (including pressure, temperature, and pH) for chemical reactions to occur. In addition, lower energy consumption is possible as enzyme's optimum parameters are relatively mild. Enzyme usage also allows production of specific products only, therefore it is possible to avoid production of unwanted by-products that may be challenging and expensive to dispose or may even pose negative effect (Binod et al., 2013).

Lignocellulose is the major structural component of plant material; it is a sustainable and renewable resource available for use in biotechnological applications. Thus, cellulose-degrading enzymes are amongst the highly demanded industrial enzymes due to their applications in various industries including detergents, fermentation, food, textile, pulp and paper (Hill et al., 2006). Using cellulose-degrading enzymes in enzymatic hydrolysis of cellulosic materials into fermentable sugars for bioethanol production may reduce the process cost greatly as the processes may be improved to take place at mild conditions (temperature 45 to 50 °C and pH 4 to 6) (Kuhad et al., 2011). Cellulose-degrading enzymes are commonly synthesised by microorganisms such as fungi, bacteria and actinomycetes during their growth on cellulosic materials; these enzymes may be thermophilic, mesophilic, aerobic or

anaerobic. Several genera are more extensively studied compared to others including *Aspergillus*, *Cellulomonas*, *Clostridium*, and *Trichoderma* (Kuhad et al., 2010).

Microbial enzymes are normally obtained through cultivation and subsequent screening of pure strains of microorganisms. It is later found out that only 1-15% of microbial genomes are cultivable under laboratory conditions while more than 85% have never been studied before (Amann, 1995).

Techniques of specifically cloning environmental DNA involving the hereditary diagrams of whole microbial consortia what is called metagenome, furnish molecular grouping space that alongside cunning *in vitro* advancement technologies will collaborate synergistically to bring a limit of accessible succession space into biocatalytic application (Lorenz et al., 2002). The new method involves direct DNA isolation from an environmental sample followed by direct cloning for subsequent screening and product expression to allow unbiased genomic representation of the microbes.

Molecular techniques of environmental DNA cloning in synergy with *in vitro* microbiological technologies facilitate the biocatalytic application (Handelsman et al., 1998). This field is mostly similar to genome library development and screening, with the distinction that the cloned DNA does not start from a solitary known microorganism, but instead from the whole consortia in a special environment. Function-based metagenomics has developed as a strong strategy for DNA model approval and protein bioprospecting from natural and hand-engineered biological systems (Mewis et al., 2013).

The success of the cloning may be proven by positive presence of novel proteins through function-based screening where the clones' biological activity is monitored (Kumar et al., 2015). Screening may be performed through two

alternatives; sequence-based and activity or function-based screening. It is recommended to combine both methods to obtain the complete picture of the community. Sequence based screening allows the identification of the sequence of interest while functional screening provides identification of unknown and novel genes that might not be recognisable by only sequence based screening (Riesenfeld et al., 2004). It is a challenge to develop effective and sensitive functional screening where it previously relies on noticeable changes in colony morphology or the appearance of zymogram (Teather and Wood, 1982). High throughput screening is introduced where screens are conducted in 384-well plate format to increase the efficiency and comparability between samples. Cell lysis is also imposed in this method to overcome intracellular accumulation of enzyme activity. Substrate of interest (in this case fluorogenic substrate) may interact with the enzyme released from the cells and allow the reading of enzyme activity using fluorescence-based microplate readers (Taupp et al., 2011).

In this study, palm oil mill effluent (POME) was chosen because it is a rich habitat for microorganisms since it contains nutrients and growth factors. In the presence of significant quantity of cellulosic fruit residues, microorganisms secrete cellulose-degrading enzymes to adapt and survive in POME environment. Malaysia's POME samples were collected and enriched in the laboratory to enhance the screening of cellulose-degrading enzymes. After metagenomic DNA extraction, sequences of around 40 kb were cloned into a fosmid vector and transformed into a bacterial host, which is generally *Escherichia coli*. A high number of multiplications of *E. coli* with the inserted DNA created fosmid libraries of POME metagenomic DNA. The presence of cellulose-degrading enzymes was conducted with microtiter high-throughput screening of the four libraries using fluorogenic substrates in fluorescence-able

microplate reader. Only clones with the desired enzymes were able to metabolise the substrate and merge fluorescence to predict their presence. Due to the limited budget, only 100 high-rated clones were sequenced with next generation sequencing (NGS). Their DNA was sequenced to define the coding DNA sequence of each enzyme. Second cloning was carried out with those sequences into an expression vector. Finally, the enzymes were expressed, purified and characterised to discover more about the cellulose-degrading enzymes.

1.2 PROBLEM STATEMENT AND RESEARCH SIGNIFICANCE

The number of cellulose-degrading enzymes which are currently successfully utilised at the industrial level is considered low compared to the high demand due to the variety of cellulosic composition in plant biomass. In addition, the available enzymes in the market are considered expensive which imposes a real obstacle of the process. These problems indicate the need for further research and development of versatile and low-cost enzymes to grow progressively sustainable and financially competitive generation processes (Adrio & Demain, 2014). As microorganisms are the main source of enzymes, the real problem behind the lack of enzymes versatility and novelty is the weakness of classical methods to benefit the maximum possible of this available source (Kumar et al., 2015; Vavourakis et al., 2018). The evolution of bioinformatic techniques to study the biodiversity of microorganisms revealed the presence of unexpected numbers of unknown species with complicated phylogenetic connections which have been missed by traditional cultivation methods. Named “unculturable microorganisms”, are a prosperous source of biocatalysts if the right method of investigation is followed.