

**Identifying Novel Lignocellulosic  
Processing Enzymes from *Cellulomonas  
fimi* using Transcriptomic, Proteomic and  
Evolution Adaptive Studies**

**Nurashikin Ihsan**

**PhD**

**University of York**

**Biology**

**May 2017**

## Abstract

The declining reserves of fossil fuel twined with an increasing concern about the environmental consequences of burning these fuels and rising carbon dioxide levels, means that a more sustainable replacement is required. Lignocellulosic biomass is an attractive candidate that has been shown to be the best sustainable alternative source to produce bioethanol for liquid transportation fuels. It has enormous availability, is renewable and cost-effective. As an agricultural residue, it does not compete with food production. However, lignocellulosic biomass of plant cell walls is composed mainly of cellulose, hemicellulose and lignin, which are extremely resistant to digestion. Converting this biomass to useful products of fermentable sugars for bioethanol production has met with little success as harsh pretreatment and costly enzyme applications are required. An arsenal of enzymes and a synergistic mechanism are required to deconstruct recalcitrant lignocellulosic biomass for an efficient production of lignocellulosic bioethanol. To achieve this goal, this study used transcriptomic and proteomic approaches with the objective of identifying new genes and enzymes involved in lignocellulose degradation. This revealed that the only one AA10 of *Cellulomonas fimi* was among the highest enzymes identified during the degradation of cellulose. Another other 20 hypothetical proteins co-expressed with CAZymes have been identified including a potentially exclusively new *C. fimi*  $\beta$ -glucosidase (PKDP1) that contains a PKD-domain and oxidoreductase predicted function of PQQ-domain. A naturally mutagenized *C. fimi* population also was screened from an adaptive evolution experiment involving exposure to a wheat straw environment. One of the strains in the adaptive population (Strain-6) showed a higher association with wheat straw biomass, which may be an indication of the strategy that being used by the adapted strain to tackle obstinate substrates to sustain growth. These results show many new enzymes would be revealed from the *C. fimi* repertoire in order to have a better enzymatic cocktails for lignocellulose breakdown. For the future, this encourages a deeper understanding of lignocellulose deconstruction mechanisms by an orchestra of multiple enzymes in a bacterial system.

# List of Contents

Abstract.....	2
List of Contents.....	3
List of Tables .....	10
List of Figures.....	11
List of Appendices .....	14
Dedication .....	15
Acknowledgements .....	16
Author's declaration .....	18
1 Introduction .....	19
1.1 World's trend on water-food-energy demand .....	19
1.1.1 Global resources insecurity and climate change .....	19
1.2 Bioethanol as a biofuel .....	22
1.2.1 First generation of biofuel.....	23
1.2.2 Second generation of biofuel .....	23
1.3 Lignocellulosic biomass.....	24
1.3.1 Cellulose.....	25
1.3.2 Hemicellulose.....	26
1.3.3 Lignin.....	27
1.4 Lignocellulose digestion in nature .....	28
1.4.1 Aerobic lignocellulolytic bacteria.....	29
1.4.2 Glycosyl Hydrolases (GHs) .....	29
1.4.3 Lytic polysaccharide Monooxygenase (LPMO) .....	30
1.4.4 <i>Cellulomonas fimi</i> ATCC® 484™ .....	32
1.5 Aims of the project .....	36
2 General materials and methods .....	37
2.1 Chemical reagents, substrates, and organisms .....	37
2.1.1 Chemical reagents.....	37

2.1.2	Lignocellulose biomass .....	37
2.1.3	Organisms .....	38
2.2	Microbiology methods.....	38
2.2.1	Buffers.....	38
2.2.2	Media for bacterial growth.....	38
2.2.3	Agar plate and slant preparation .....	38
2.3	Molecular biology techniques.....	39
2.3.1	Gram-positive bacterial DNA extraction.....	39
2.3.2	Fungal cells <i>A. niger</i> disruption for DNA extraction.....	40
2.3.3	cDNA synthesis.....	40
2.3.4	Polymerase Chain Reaction (PCR).....	40
2.3.5	Agarose gel electrophoresis.....	41
2.3.6	Plasmid extraction .....	41
2.3.7	PCR clean-up .....	42
2.3.8	Nucleotide quantification .....	42
2.3.9	Sanger DNA Sequencing.....	42
2.3.10	In-Fusion™ cloning .....	42
2.3.11	DNA restriction digests .....	43
2.3.12	Transformation of competent cells .....	43
2.4	Protein methods .....	43
2.4.1	Bradford assay for protein quantification .....	43
2.4.2	Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) .....	43
2.4.3	Western blotting .....	44
2.4.4	Protein concentration by centrifugation .....	45
3	Growth and transcriptomic studies of <i>Cellulomonas fimi</i> .....	46
3.1	Introduction .....	46
3.2	Materials and methods.....	49
3.2.1	Experimental setup and growth study.....	49
3.2.2	Strain and growth conditions.....	50

3.2.3	Total RNA extraction .....	50
3.2.4	Enrichment of mRNA from total RNA .....	50
3.2.5	rRNA depletion using Ribo-Zero rRNA Removal Kit (Bacteria) and Ribo-Zero rRNA Removal Kit (Gram-positive) .....	51
3.2.6	Selective cDNA synthesis for rRNA depletion using NuGEN-Ovation® Universal RNA-Seq System .....	51
3.2.7	Illumina Sequencing libraries .....	52
3.2.8	Mapping .....	52
3.2.9	Transcript abundance and bioinformatics analysis.....	53
3.2.10	Differential expression analysis .....	53
3.3	Results.....	54
3.3.1	Growth study of <i>C. fimi</i> utilizing different substrates.....	55
3.3.2	Total RNA extraction from <i>C. fimi</i> .....	56
3.3.3	mRNA enrichment from total RNA .....	58
3.3.3.1	mRNA enrichment by rRNA depletion .....	58
3.3.3.2	RNA-Seq work and analysis using NuGEN® Ovation Universal RNA-Seq Sytem..	58
3.4	The transcriptome of <i>C. fimi</i> growing on five substrates.....	60
3.4.1	Expression of CAZy genes in <i>C. fimi</i> grown on polysaccharides and comparison with <i>C. fimi</i> grown on glucose.....	60
3.5	Expression of Carbohydrate Active Enzyme (CAZy)-encoding genes .....	61
3.6	Co-expression of upregulated CAZy- and non-CAZy encoding genes using differential analysis.....	66
3.7	Discussion .....	72
4	Proteomic Analysis of the Secretome of <i>Cellulomonas fimi</i> .....	76
4.1	Introduction .....	76
4.2	Materials and methods.....	79
4.2.1	Sample preparation .....	79
4.2.2	Protein quantification .....	80
4.2.2.1	Bradford assay and SDS-PAGE protein gel.....	80

4.2.3	Enzyme activity assays .....	80
4.2.3.1	Qualitative plate activity assay .....	80
4.2.3.2	Reducing sugar assay in protein solution .....	80
4.2.4	Label-free semi-quantitative proteomic analysis using mass spectrometry .....	81
4.2.4.1	1D mini gel electrophoresis and tryptic digestion for GelC-MS of <i>C. fimi</i> proteomic samples from culture supernatants .....	81
4.2.5	Mass Spectrometry Analysis of <i>C. fimi</i> proteomic samples from culture supernatants.....	82
4.2.5.1	Spectra analysis.....	82
4.2.5.2	Notes on Molar Fraction percentage (MFP) values .....	84
4.3	Results.....	85
4.3.1	Protein quantification .....	85
4.3.1.1	Sodium Dodecyl-Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE) .....	85
4.3.1.2	Bradford assay .....	87
4.3.2	Enzyme activity assay.....	88
4.3.2.1	Enzyme activity agar plate assay .....	88
4.3.2.2	Reducing sugar assay of protein solution .....	91
4.3.3	Label-free semi-quantitative proteomic analysis .....	92
4.3.4	Sequence-based analysis of the secreted proteins .....	93
4.3.4.1	Functional distribution and localization of identified proteins from <i>C. fimi</i> secretome.....	93
4.3.4.2	Secretion pathways of extracellular <i>C. fimi</i> proteins .....	94
4.3.5	Principal Component Analysis (PCA).....	95
4.3.6	Protein domain prediction.....	98
4.3.7	CAZymes secreted by <i>C. fimi</i> .....	100
4.3.8	Cellulases .....	100
4.3.9	A LPMO, additional glucanases and potential xylanases.....	101
4.3.10	Xylanases.....	101
4.4	Discussion .....	103

5	Attempted Characterization of a <i>Cellulomonas fimi</i> PKD-Domain Containing Protein (PKDP1).....	112
5.1	Introduction .....	112
5.1.1	Multidomain structure prediction of the <i>C. fimi</i> PKD-domain-containing protein, Celf_2278.....	112
5.1.2	Polycystic kidney disease I (PKD1) domain .....	114
5.1.3	Protective Antigen (PA) 14.....	116
5.1.4	A multidomain of Glucose/Sorbone Dehydrogenase (GSDH) and Pyrroloquinone quinone (PQQ) dependent domain .....	117
5.1.5	Aims of this chapter .....	119
5.2	Materials and methods.....	120
5.2.1	Isolation of native PKDP1 protein from supernatant and substrate-bound fractions induced with specific substrates in <i>C. fimi</i> culture.....	120
5.2.2	Weak anion exchange chromatography for native isolation of PKDP1.....	121
5.2.3	Expression of <i>C. fimi</i> PKDP1 in prokaryotic expression systems.....	121
5.2.4	Gene amplification and cloning .....	121
5.2.5	Expression of PKDP1 in <i>E. coli</i> strains .....	122
5.2.6	Larger scale expression culture of heterologous PKDP1 by <i>E. coli ArcticExpress</i> <sup>™</sup> .....	123
5.2.7	Purification of heterologously expressed PKDP1.....	123
5.2.8	Expression of <i>C. fimi</i> PKDP1 in an <i>Aspergillus niger</i> expression system .....	124
5.2.8.1	Cloning of a codon-optimized <i>C. fimi</i> PKDP1 gene into pIGF-pyrG expression system.....	124
5.2.9	Preparation of Hutner's trace elements .....	125
5.2.10	Preparation of minimal media .....	125
5.2.11	Preparation of protoplasts.....	125
5.2.12	Fungal transformation .....	127
5.2.13	Direct Colony Polymerase Chain Reaction (DCPCR) using DNA template from <i>A. niger</i> hyphae.....	128
5.2.14	Expression trials in <i>A. niger</i> system .....	128

5.3	Results.....	129
5.3.1	Attempted isolation of PKDP1 from <i>C. fimi</i> cultures .....	129
5.3.1.1	Growth on different substrates .....	129
5.3.1.2	Isolation trial of <i>C. fimi</i> PKDP1 using weak anion exchange chromatography .	130
5.3.2	Recombinant PKDP1 solubility study in <i>E. coli</i> BL21(DE3) .....	132
5.3.3	Expression trial of <i>C. fimi</i> PKDP1.....	133
5.3.4	Larger Scale Expression and Purification of <i>C. fimi</i> PKDP1 using <i>ArcticExpress</i> <sup>™</sup> cells .....	139
5.3.4.1	Result of larger protein expression.....	139
5.3.4.2	Result of protein purification by affinity column chromatography .....	140
5.3.5	Expression of codon-optimized recombinant <i>C. fimi</i> PKDP1 in <i>Aspergillus niger</i> .....	141
5.3.6	Solubility of PKDP1 protein.....	144
5.4	Discussion .....	148
5.4.1	Attempted isolation of native PKDP1 from <i>C. fimi</i> culture .....	148
5.4.2	PKDP1 expression in <i>Escherichia coli</i> .....	149
5.4.3	PKDP1 expression in <i>Aspergillus niger</i> .....	152
6	Adaptive Evolution of <i>Cellulomonas fimi</i> by Continuous Subculture on Wheat Straw ....	154
6.1	Introduction .....	154
6.2	Materials and methods.....	156
6.2.1	Bacterial strain, media, and cultivation .....	156
6.2.2	Strain construction.....	156
6.2.3	Analytical methods .....	158
6.2.3.1	Cell growth profiles and relative growth rate.....	158
6.2.3.2	Carbon dioxide evolution (CER) .....	158
6.2.3.3	Estimation of growth and bacterial residency by genomic DNA of <i>C. fimi</i> strains .....	158
6.2.3.4	Estimation of growth and bacterial residency by total protein.....	159
6.2.4	Scanning Electron Microscopy (SEM) analysis of wheat straw samples .....	160



6.2.5	Enzyme activity assays in culture supernatant .....	161
6.2.6	Compositional analysis of oligosaccharide from wheat straw degradation .....	162
6.2.7	Quantification of lignocellulose biomass residual after degradation .....	162
6.2.8	Statistical analysis .....	162
6.3	Results .....	163
6.3.1	Growth rates on wheat straw of <i>C. fimi</i> strains .....	163
6.3.2	Carbon Dioxide Evolution Rate (CER) of <i>C. fimi</i> .....	165
6.3.3	Isolation of <i>C. fimi</i> Genomic DNA (gDNA) from supernatant and biomass fractions .....	168
6.3.4	Total protein isolation from <i>C. fimi</i> culture supernatant and cell-bound to wheat straw .....	170
6.3.5	Qualitative visualisation of <i>C. fimi</i> bound to wheat straw using Scanning Electron microscopy (SEM) .....	172
6.3.6	Enzyme activity assay using culture supernatants .....	180
6.3.7	Polysaccharide content and composition .....	182
6.3.8	Quantification of residual wheat straw .....	184
6.4	Discussion .....	185
7	Final discussion .....	190
	Appendices .....	195
	Abbreviations .....	222
	References .....	225

# List of Tables

Table 1.1: Characterized <i>C. fimi</i> proteins involving in polysaccharide degradation as listed in the UniProt KB database.....	33
Table 2.1: Polymerase chain reaction components.....	40
Table 2.2: Polymerase chain reaction thermocycling conditions.....	41
Table 3.1: Experimental design and number of reads generated by RNA sequencing of fifteen <i>C. fimi</i> RNA-seq samples with triplicates for each condition. ....	61
Table 3.2: Number of upregulated CAZy-encoding genes for each comparison condition. ....	62
Table 3.3: The CAZy-encoding genes upregulated on each carbohydrate substrate using Differential Gene Expression (DGE) analysis. ....	64
Table 3.4: Identification of co-expressed genes which upregulated with CAZy-encoding genes from differential analysis of RNA-seq data. ....	67
Table 4.1: Known cellulase and xylanase families in <i>Cellulomonas fimi</i> . ....	76
Table 4.2: Importance of component variables from the <i>C. fimi</i> proteomic dataset analysed by Principal Component Analysis (PCA). ....	95
Table 4.3: Summary of predicted domains found in 1,4- $\beta$ -xylanase (Prot3, Celf_0574). ....	99
Table 4.4: Summary of detected cellulases and xylanases in <i>C. fimi</i> secretome grown on four types of carbon sources compared to the CAZy database of <i>C. fimi</i> enzymes (2017). ....	102
Table 5.1: PKDPs of <i>C. fimi</i> detected in the proteomic analysis using nanoLC-MS/MS. ....	113
Table 5.2: Primers used to amplify PKDP1 (Celf_2278) for cloning into pETFFP_3 vector with various solubility tags. ....	122
Table 5.3: Primers used to amplify codon-optimized PKDP1 and add the tag sequence into (Celf_2278) for cloning into pIGF-pyrG A. niger vector. ....	124
Table 5.4: Forward primers used for DCPCR of <i>C. fimi</i> PKDP1 gene in pIGF-pyrG vector.....	128
Table 5.5: Pre-computed score of <i>C. fimi</i> PKDP1 protein localization. ....	145
Table 5.6: Summary of advantages of E. coli strains for protein expression.....	150

# List of Figures

Figure 1.1: The world's trend demand and policy nexus. ....	19
Figure 1.2: Proportions of global greenhouse gas emissions by (A) economic sectors and (B) the projection of CO <sub>2</sub> emission. ....	20
Figure 1.3: Fires from open burning in India and Indonesia detected by NASA satellites. ....	21
Figure 1.4: Illustration of a plant cell walls. ....	25
Figure 1.5: Overview of cellulose structure. ....	26
Figure 1.6: Structure of lignin. ....	28
Figure 1.7: Schematics of microbial mechanisms of lignocellulose degradation. ....	30
Figure 3.1: Workflow of transcriptome analysis of <i>C. fimi</i> transcriptome grown on four types of carbon source. ....	49
Figure 3.2: Growth profiles of <i>C. fimi</i> grown on different carbon sources initiated using differently aged inocula of seeding cells. ....	55
Figure 3.3: Electrophoretic analysis of RNA from <i>C. fimi</i> using TapeStation® with standard RNA ScreenTape®.....	56
Figure 3.4: TapeStation® Electrophogram of total RNA harvested from five substrates of Day-3 <i>C. fimi</i> grown cultures. ....	57
Figure 3.5: Histogram indicating quality of RNA-Seq result with percentage of the reads mapped to the genome including the rRNA genes, coding sequences genes and the unmapped reads. ....	59
Figure 3.6: Differentially expressed transcripts between glucose and carbohydrate treatment after 3-days growth of <i>C. fimi</i> . ....	63
Figure 3.7: Venn diagram of the CAZy- and non-CAZy encoding genes that were significantly upregulated in each condition. ....	66
Figure 4.1: Workflow for secretome analysis of <i>C. fimi</i> grown on four types of carbon source.....	79
Figure 4.2: SDS-PAGE secretome profiles of <i>C. fimi</i> grown in 5 substrates with 3 biological replicates at 3 time points. ....	86
Figure 4.3: Protein concentration in the supernatant of day-3 <i>C. fimi</i> cultures grown on different substrates. ....	87
Figure 4.4: Detection of polysaccharide-degrading enzymes from supernatant of <i>C. fimi</i> culture in nutrient agar plates containing substrates stained with Congo Red. ....	89
Figure 4.5: Identification of polysaccharide-degrading activities in <i>C. fimi</i> culture by the activity plate assay. ....	90
Figure 4.6: Cellulase and xylanase activity in <i>C. fimi</i> cultures grown on different substrates.....	91

Figure 4.7: Distribution of extracellular proteins detected in day-3 <i>C. fimi</i> cultures. ....	92
Figure 4.8: Functional classification of proteins secreted by <i>C. fimi</i> grown on Avicel, xylan, wheat straw, and sugarcane bagasse. ....	93
Figure 4.9: Venn diagram showing the predicted distribution of four secretion pathways among the 71 identified extracellular proteins of <i>C. fimi</i> . ....	94
Figure 4.10: Scree plot from Principal Component Analysis (PCA) of proteomics from day-3 <i>C. fimi</i> culture supernatants from growth in Avicel, beechwood xylan, wheat straw, and sugarcane bagasse. ....	96
Figure 4.11: Biplot graph from PCA analysis of <i>C. fimi</i> proteomic dataset using two principal components. ....	97
Figure 5.1: Predicted domain organization of PKD-domain-containing proteins of <i>C. fimi</i> .....	113
Figure 5.2: Partial Sequence alignment of five PKDPs of <i>C. fimi</i> . ....	115
Figure 5.3: Chemical structure of pyrroloquinoline quinone (PQQ). ....	117
Figure 5.4: Proposed scheme of fungal synergies for oxidative degradation of cellulose. ....	119
Figure 5.5: Equipment setup for fungal protoplast collection. ....	126
Figure 5.6: SDS-PAGE analysis of protein extracted from supernatant and substrate-bound fractions induced by specific substrates in <i>C. fimi</i> culture. ....	129
Figure 5.7: Weak anion exchange separation of extracellular proteins of <i>C. fimi</i> grown in Avicel for 7 days. ....	131
Figure 5.8: Protein gel analysis of PKDP1 expression samples following dot blot screening of <i>C. fimi</i> candidate proteins. ....	133
Figure 5.9: Protein gel analysis of expressed PKDP1 following induction in <i>E. coli</i> BL21 (DE3).....	135
Figure 5.10: Protein gel analysis of expressed PKDP1 following induction in <i>E. coli</i> ArcticExpress™.....	136
Figure 5.11: Protein gel analysis of expressed PKDP1 following induction in <i>E. coli</i> SHuffle®.....	137
Figure 5.12: Protein gel analysis of expressed PKDP1 following induction in SHuffle®T7.....	138
Figure 5.13: Protein gel analysis of expressed PKDP1 following induction in <i>E. coli</i> ArcticExpress™ from 500 mL culture. ....	139
Figure 5.14: Protein gels analysis of affinity chromatography purification fractions of expressed PKDP1 in <i>E. coli</i> ArcticExpress™.....	140
Figure 5.15: Colonies of <i>A. niger</i> transformed with codon-optimized PKDP1 in pIGF-pyrG on AMMN agar plates after 4 days of incubation at 30°C .....	141
Figure 5.16: Confirmation of pIGF-pyrG:PKDP1 integrated in <i>A. niger</i> using direct colony PCR.....	142
Figure 5.17: SDS-PAGE gels loaded with unconcentrated supernatant from cultures of <i>A. niger</i> expressing recombinant PKDP1. ....	143

Figure 5.18: Transmembrane prediction of PKDP1 using membrane protein topology prediction based on Transmembrane hidden Markov model (TMHMM) prediction server. ....	145
Figure 5.19: The Kyte and Doolittle (278) hydrophobicity plot of PKDP1 from <i>C. fimi</i> . ....	147
Figure 6.1: Experimental design for adaptive evolution. ....	157
Figure 6.2: Comparison of growth profiles between <i>C. fimi</i> wild type and adapted strain-1 to strain-6 in 4-day time course. ....	164
Figure 6.3: Carbon dioxide evolution rate (CER) for wild type and 6 adapted strains of <i>C. fimi</i> grown on wheat straw.....	166
Figure 6.4: Carbon dioxide evolution comparison between <i>C. fimi</i> wild type and six independent one year-adapted strains over a 4-day time course. ....	167
Figure 6.5: Quantitative measurement and qualitative observation of gDNA isolated from Day-4 of the wild type and six population adapted strains of <i>C. fimi</i> from wheat straw culture....	169
Figure 6.6: Quantification of total protein isolated from supernatant and biomass fraction of wild type, strain-3 and strain-6 of 1-year adapted strains. ....	171
Figure 6.7(A): Surface images obtained by SEM on wheat straw after 4 days of aerobic degradation inoculated with the wild type of <i>C. fimi</i> in basal medium.....	173
Figure 6.7(B): Surface images obtained by SEM on wheat straw inoculated with strain-1 of <i>C. fimi</i> after 4 days of aerobic degradation in basal medium. ....	174
Figure 6.7(C): Surface images obtained by SEM on wheat straw inoculated with strain-2 of <i>C. fimi</i> after 4 days of aerobic degradation in basal medium. ....	175
Figure 6.7(D): Surface images obtained by SEM on wheat straw inoculated with strain-3 of <i>C. fimi</i> after 4 days of aerobic degradation in basal medium. ....	176
Figure 6.7(E): Surface images obtained by SEM on wheat straw inoculated with strain-4 of <i>C. fimi</i> after 4 days of aerobic degradation in basal medium. ....	177
Figure 6.7(F): Surface images obtained by SEM on wheat straw inoculated with strain-5 of <i>C. fimi</i> after 4 days of aerobic degradation in basal medium. ....	178
Figure 6.7(G): Surface images obtained by SEM on wheat straw inoculated with strain-6 of <i>C. fimi</i> after 4 days of aerobic degradation in basal medium. ....	179
Figure 6.8: Comparison of CMCase and xylanase activity in culture supernatant of wild type and adapted strain-1 to strain-6 of <i>C. fimi</i> . ....	181
Figure 6.9: Monosaccharide composition from the wheat straw degradation by wild type, strain-3 and strain-6 of <i>C. fimi</i> . ....	183
Figure 6.10: Mass loss of wheat straw after the degradation by <i>C. fimi</i> strains. ....	184

## List of Appendices

Appendix A: Predicted structural domains of proteins identified in <i>C. fimi</i> secretome.....	192
Appendix B: List of 71 proteins identified in <i>C. fimi</i> secretome grown in four types of carbon sources.....	200
Appendix C: Multiple sequences alignment of <i>C. fimi</i> PKD-domain containing proteins.....	206
Appendix D: Sequence of condon-optimized <i>C. fimi</i> PKDP1 protein synthesized by GeneArt® .....	210
Appendix E: pIGF-pyrG <i>A. niger</i> vector map.....	218

# Dedication

*This thesis is dedicated to my beloved children,  
who being my buffering system, my backbone,  
during the ups and downs and the bittersweets  
of this PhD journey,*

*Harith Hayyan & Sopheha Insheera;*

*This is for both of you and for our brighter future.*

# Acknowledgements

*In the name of God, the Most Beneficent the Most Merciful*

There are a number of people without whom this thesis might not have been written, and to whom I am greatly indebted.

To my mother and eldest sister, Hafifah and Khairunisa, who continue to be patiently and steadfastly in facing great challenges in your life years ago, and who have been a source of encouragement and inspiration to me throughout my life, a very special thank you for sacrifice a lot of things, also helped to fund and support me and my family through the last several months of my final labworks and writing stage in the UK. And also including my other siblings, Nurfarahin, Amirul Hakim and Izleen Azyzee for the myriad of ways in which, throughout my life, you have actively supported me in my determination to find and realise my potential, and to make this contribution to our world.

To my dear husband, Hairuddin who remains willing to engage with the struggle, and ensuing discomfort, of having a partner who refuses to accept any failures and in most of the time was emotionally-driven to continue this PhD journey until the finishing line. A very special thank you for your practical and emotional support as I added the roles of wife and then mother, to the competing demands of work, study and personal development.

Much loving thanks to dear Harith and Sophea, for being so supportive and indeed my main support - even when being 'without mum' was not the best event in most of the time. This work is for, and because of you and all the generations to come. It is dedicated to all our journeys in learning to thrive.

It is also dedicated to Dr. Laziana - friend, 'sister', colleague, and researcher - who knowingly and unknowingly- led me to an understanding of some of the subtler challenges to our ability to thrive. If our attempts to claim our right to speak our truth, and to unravel and follow the threads through which our oppression is maintained, and are instrumental in helping one other "Super Woman" from 'going over the edge' - in which sight we continuously live our lives - perhaps, it might be seen that your invaluable contribution to the attainment of many of the insights gained was worth it.

Loving thanks to my friends, project partners, and colleagues; Dr. Susie Bird, Dr. Nicola Oates, and Dr. Anna Alessi, who played such important roles along the journey, also to lovely Lynda Sainty and Dr. Alexandra Lanott, as we mutually engaged in making sense of the various challenges we faced and in providing encouragement to each other at those times when it seemed impossible to continue.

I offer my gratitude and appreciation to my supervisor, Prof Simon McQueen-Mason, for all your advice, guidance, calming influence, time to listen and for your understanding of my other commitments as an academic staff in UTM, and also as a mother of two. For the deft ways in which you challenged and supported me throughout the whole of this work - knowing when to push and when to let up. To Professor Neil Bruce and Professor Peter Young for your kind assistance, and generous time given for discussing research as a team in "Mining Compost" project.



I also greatly indebted special thanks to Dr Thorunn Helgason as one of the Training Advisory Panel (TAP) members, also to Professor Mike Brockhurst, for never under-estimate my potentials, and always help to boost my self-esteem to guide me in this research world mostly during our TAP meetings.

I offer special thanks to those who supported me in the mechanics of producing this thesis. Dr. Katrin Bießer and Dr. Clare Steele-King, for reading and rereading drafts; Luisa for giving hands and ideas as a sifoo of cloning work; Dr. Lesley Gilbert for kind motivation and help during preparation of RNA-Seq samples. To Laziana and Nahed for help with formatting and for helping me get 'unstuck' with this thesis on many occasions from the very first draft, to an extend helping me to print out the thesis for the viva.

To everyone in CNAP who have helped during my lab life. This has been an exceptionally enjoyable environment to learn and to work in, due to being filled with lovely people who take the time to support others around them. Leo Gomez, Joe, Rachael H., Rachael E., Emily, Dan, Dave, Louise, Maria M., Duong, Julia, Juliana, Daniel U., Giovanna, Fede and Aritha.

My appreciation also goes to Cikgu Salmah and Cikgu Saodah, who were among the first person who introduce me to science since in the primary school, Cikgu Nazli Abdullah, my biology teacher in the secondary school who has inspired me to continue and strive in life sciences stream, also Prof Noor Aini, and Assoc. Prof Dr Madihah Salleh, who seen my compassion in biology, and my determination to thrive and go beyond the edge.

Thank you to Ministry of Higher Education, Government of Malaysia; Universiti Teknologi Malaysia who provide the Academic Staff Bumiputra Training Scholarship (SLAB) through my 3.5 years of study. My special grateful appreciation to Radhika V Sreedhar Scholarship and Prof McQueen-Mason for the hardship funds in my last several months towards the end of my PhD journey. Personally, all the funds mean a lot to me and played as the main contributors of making this journey completed, successfully.

Most of all, I'm gratefully praise to the Greatest God, who make the impossible, possible.

Verily, with hardship, there is relief.

(Al-Insyirah [The Consolation], 94:6)

## **Author's declaration**

I declare that the work presented in this thesis is my own original work of research and I am the sole author, except where due reference has been given to collaborators and co-workers.

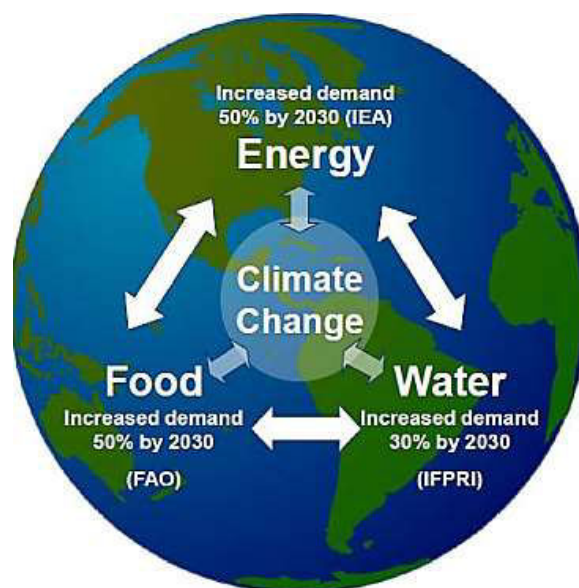
This work has not previously been presented for an award at this, or any other, University.

All sources are acknowledged as References.

# 1 Introduction

## 1.1 WORLD'S TREND ON WATER-FOOD-ENERGY DEMAND

The demand for water, food, and energy are predicted to increase by 40, 35, and 50 percent, respectively in the coming decades (1). This leads to the debate on 'resource scarcity' where the scientific findings suggest that humanity has exceeded the planetary boundaries and is threatening its own safety (2). Water, food and energy resources are tightly interconnected, forming a policy nexus (3,4) that is being discussed all over the globe by policy makers and scientists looking for solutions for sustainable development planning (Figure 1.1).



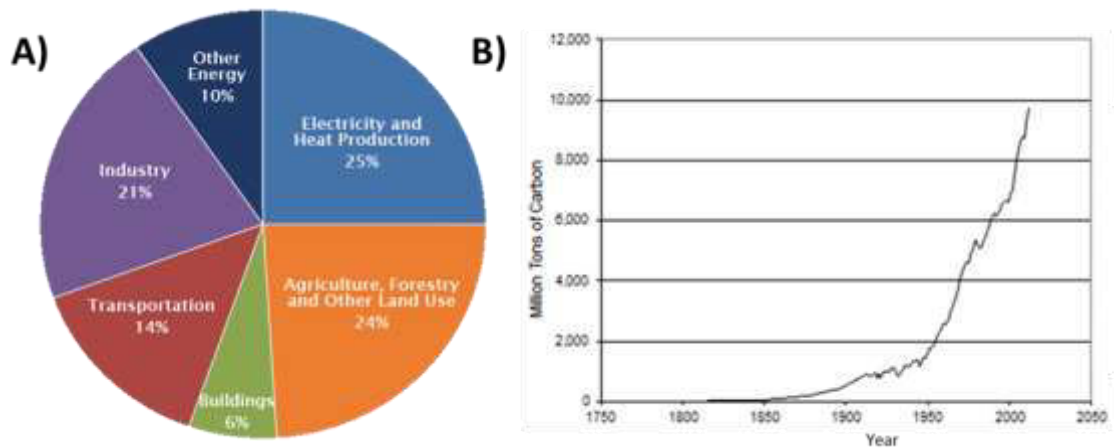
**Figure 1.1: The world's trend demand and policy nexus.**

It is predicted that by 2030 the world will need to produce around 50 per cent more food and energy, together with 30 per cent more fresh water, whilst mitigating and adapting to climate change. Illustration is reproduced from Beddington (5).

### 1.1.1 Global resources insecurity and climate change

Water is important for life and is a vital resource for the economy. It is also plays a fundamental role in the climate regulation cycle. Therefore, the management and protection of water resources is one of the keystones of environmental protection (4). Water insecurity caused by unmonitored development and environmental stress such as climate change may have a material impact on the economy. Climate change is the change in climate parameters such as regional temperature, precipitation, or extreme weather caused by increase in the greenhouse

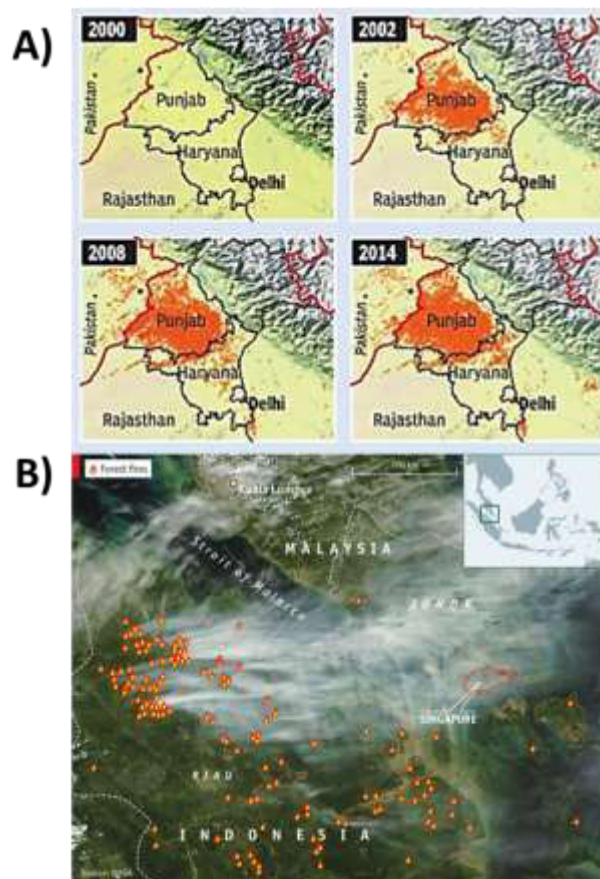
effect. It is significantly impacted by agriculture due to increasing water demand, limiting crop productivity and reducing water availability in areas where irrigation is most needed (4). Several other sectors can also cause climate change; e.g. from burning activities by the release of CO<sub>2</sub> from fossil fuel combustion and an underestimated source of greenhouse gases (GHG) emissions that is anticipated from tropical deforestation (6–8). The term Greenhouse gases refers to gases that contribute to the effect by absorbing infrared radiation (heat). The greenhouse effect is the process where the greenhouse gases (water vapors, CO<sub>2</sub>, methane, etc.) in the atmosphere absorb and re-emit heat being radiated from the Earth, hence trapping warmth that causes global warming (9). Global climate change is linked to the accumulation of greenhouse gases which causes concerns regarding the use of fossil fuels as the major energy source. To mitigate climate change while keeping energy supply sustainable, one proposal solution is to rely on the ability of microorganisms to use renewable resources for biofuel synthesis. Figure 1.2(A) shows the percentage of global greenhouse gas emissions from a study conducted by Intergovernmental Panel on Climate Change (IPCC) from 2010 (10). Electricity, heat production (25%) as well as agriculture, forestry and other land use (24%) contributed the most greenhouse emissions followed by transportation sector (14%). The increment production of GHG and black carbon emission are among the disadvantages of human activities that create a serious environmental concern. A study showed that the emission of CO<sub>2</sub> is projected to be increased since 1750 towards 2050 (Figure 1.2(B)).



**Figure 1.2: Proportions of global greenhouse gas emissions by economic sectors (A) and the projection of CO<sub>2</sub> emission (B).**

A) Six major economic sectors that use energy and produce GHG based on a global emissions study from 2010 reported in IPCC report, 2014 (10). B) Projection of global carbon dioxide emissions from fossil fuel burning since 1751 to 2012 (11).

Large scale tropical deforestation caused by burning activities in countries such as Indonesia as well as the burning of agricultural residues which occur particularly in developing countries e.g. in India and Vietnam cause toxic and severe air pollution (8). The National Aeronautics and Space Administration (NASA) revealed the severity of these activities from the satellite images taken from space (Figure 1.3). The maps revealed that stubble-burning was not widespread in 2000. However, the problem had grown alarmingly by 2002 and continues to be a major health hazard.



**Figure 1.3: Fires from open burning in India and Indonesia detected by NASA satellites.**

**A)** NASA satellite pictures reveal the evolution of paddy-stubble burning problem in Punjab, India since 2000. Each orange dot represents 1 sqkm<sup>2</sup> area where significant fires were observed. Each map shows cumulative farm fires from Oct 1 to Nov 10, each year. Images are reproduced from an article in 2015 by Amit Bhattacharya (12). **B)** Fires in Sumatra, Indonesia detected by NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) sensor in 2013. Fires set for deforestation/land clearing in Indonesia triggered health warnings in Singapore and most parts of Malaysia. Images courtesy of Google Earth and NASA, reproduced from an article from The Economist (13).

The NASA images are clear proof that widespread crop burning has become a major contribution to air pollution and GHG emission specifically in Asian countries. These are among the examples that emission of CO<sub>2</sub> and GHG contribute to the net carbon change in the atmosphere which trigger the climate instability and result in global warming (6). Black carbon emissions are a potential danger to human health and may cause premature deaths (14). The real scenarios are happening on the ground when the farmers burn fields to clear crop stubble left after harvest. However, as many farmers cannot afford to spend extra money to use a tractor and plough stubble into the earth to be decomposed, open-burning of the stubble became the fastest and cheapest option. To decompose the stubble, the farmers have to further invest in watering systems. As this process takes time, it is not favorably practiced. Furthermore, the quality of the stubble after being harvested using a machine is not usable as fodder and could not be recycled into cardboard (15).

The International Energy Agency (IEA) are targeting a 50% reduction of greenhouse gasses by 2050 (16). Several technologies for generating bioenergy to produce heat and power already exist, ranging from conventional solid wood heating installations for buildings to biogas digesters for power generation, to large-scale biomass gasifications, as well as the production of biofuels especially for transportation sector (16). Renewable sources for the generation of electricity and heat and can be produced from tidal and wind energies. However, these resources cannot be utilized as fuels; particularly liquid fuel for transportation. Therefore, the only way to produce sustainable renewable liquid fuels is through the use of renewable biological products to create biofuels.

## **1.2 BIOETHANOL AS A BIOFUEL**

Biofuels are produced by the conversion of biomass into liquids or gases, such as ethanol, lipids as biofuel precursors, biogas, or hydrogen, via biological or thermal processes. Bioethanol (CH<sub>3</sub>CH<sub>2</sub>OH) is a liquid biofuel which can be produced from several different feedstocks. Bioethanol can be used as a chemical in industrial applications or as fuel for energy generation; neat or blended with gasoline or diesel fuels. Biofuels can be broadly divided into first generation and second generation. Briefly, first generation bioethanol is mainly produced from edible crop feedstock by fermenting starch or sugars. The issue with first generation fuels is that their use of food commodities adds stress to world food security in an unsustainable manner. Second generation biofuels are produced from woody, non-food (lignocellulosic) plant biomass

such as crop residues or dedicated biomass crops. This is achieved via several pretreatments steps of the biomass, enzymatic hydrolysis and fermentation of the resulting sugars.

### **1.2.1 First generation of biofuel**

Currently, first generation biofuels are sourced from crops such as starch, sugar, vegetable oil as energy-containing molecules, or even animal fats processed by conventional methodologies. First generation biofuels offer benefits for reducing CO<sub>2</sub> emission and can aid to improve domestic energy security. Biodiesel (bio-esters), bioethanol and biogas are the examples of the first generation biofuels that have been categorized by its ability to be blended with petroleum-based fuels and combusted in existing internal combustion engines (17,18). The production of first generation biofuels is now commercially competitive with the largest ethanol producing countries, United States of America (USA) and Brazil being responsible for the production of  $54 \times 10^6$  and  $21 \times 10^6 \text{ m}^3$  in 2011, respectively (19). However, the source of feedstock raised concerns on the possible impact on biodiversity and land use; besides the competition with food crops (17). The disadvantage with these first generation biofuels is that they compete for resources with food commodities, adding to the stress on world food security brought about the growing global human population. This apparent conflict greatly limits the amount and sustainability of the biofuels that can be produced. One way in which the food security issue can be avoided is by producing biofuels from the woody non-food parts of crops and other residues.

### **1.2.2 Second generation of biofuel**

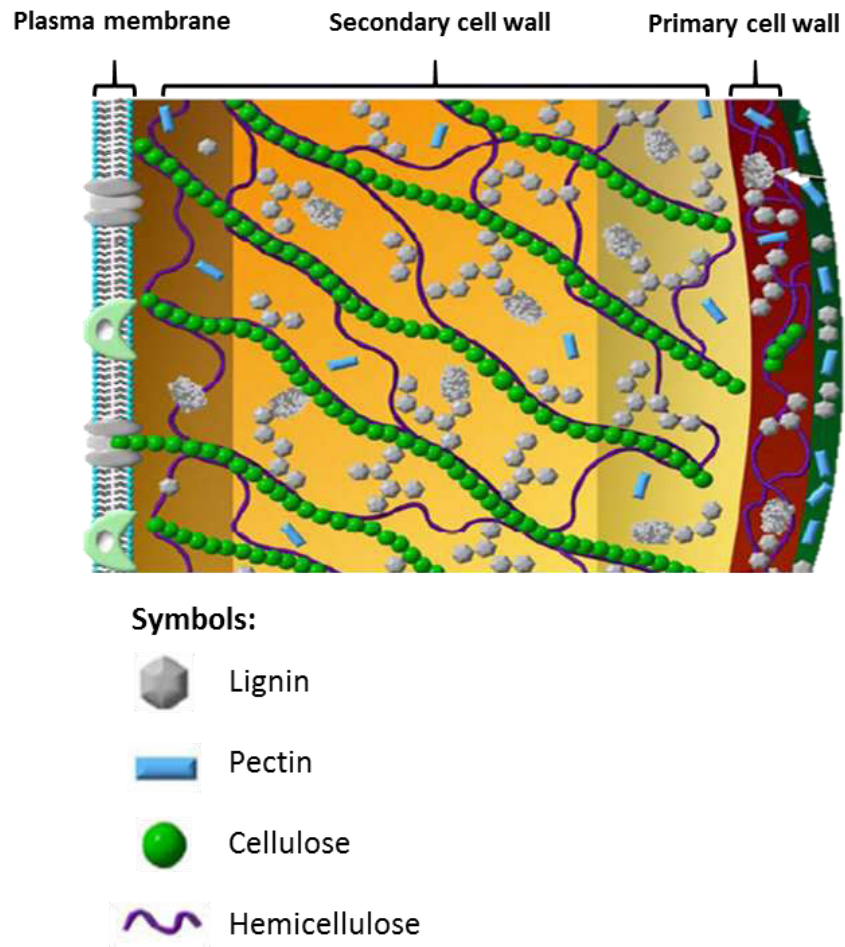
Second generation bioethanol can be produced by fermenting sugars from the lignocellulosic biomass of dedicated bio-energy crops e.g. miscanthus, or those from co-products such as cereal straw (20). Three major steps are involved in biomass-to-ethanol process; 1) biomass pretreatment and fractionation, 2) enzymatic hydrolysis of cellulosic fraction, and 3) fermentation of the derived sugars to ethanol. Many factors contribute to the overall costs of producing biomass derived ethanol, however, the feedstock cost has been reported to be among the highest (21). To reduce these costs, one of the possible ways forward is by making use of underutilised biomass materials such as wheat straw from agricultural farms. In England, there is a potential cereal straw supply of 5.27 million tons (Mt) from arable farm types; 3.82 Mt are currently used and 1.45 Mt currently chopped and incorporated (22). Approximately 10 Mt of cereals straw was generated from 3 million hectares of wheat, barley, and oats in 2015 (23,24). Of this, 75% of straw is used for animal bedding, 23% is chopped and recombined into

the soil, and 2% is used for the mushroom compost. Wheat straw is an example of lignocellulosic biomass which comes from an agricultural by-product by harvesting the cereal grains. There could be up to 1.4 million tons of wheat straw per annum available for the potential sectors such as the biofuel industry in the UK alone (25). The variations in regional straw yields ( $t\ ha^{-1}$ ) have a great effect on the England supply of straw and the potential amount of bioethanol that can be produced. This shows that commercially competitive substrates are available. However, biomass digestibility is still a major challenge. Thus, a few approaches still need to be improved; 1) to make biomass more digestible without compromising crops yield, and 2) to apply more effective pretreatments and enzymes for bioethanol conversion (26).

### **1.3 LIGNOCELLULOSIC BIOMASS**

Lignocellulosic biomass is an attractive resource for fuel and biochemical production due to its abundance in nature. Waste biomass, the stalks of agriculture crops such as wheat and paddy, corn stover and wood can serve as raw materials. One of the attractions of woody plant biomass, or lignocellulose, is that it is rich in polysaccharides that can be converted into sugars for fermentation. However, one of the reasons that lignocellulose is abundant is because it is hard to break down into simple sugars due to its naturally durable structure. The plant cell wall is a structure characterized by a network of polysaccharides, structural proteins, and phenolic compounds. This network of polymers protects the plant against external stresses and provides structural and mechanical support to plant tissues. It is biochemically resistant mainly due to the presence of polyphenols called lignin that serve as protection and natural barrier of the plant against hydrolytic enzymes produced by microorganism in nature (27). The chemical composition and mechanical properties make plant cell walls a rich source of chemicals and fermentable sugars for the production of biofuels as it is comprised of roughly 70% polysaccharides that can potentially serve as a source of fermentable sugars (28). Plant cell walls are classified as primary and secondary cell walls. Both are different in their physiological roles as well as their chemical composition. Primary cell walls are located around dividing and elongating cells which consists of a large proportion of polysaccharides (cellulose; 40-50%, hemicelluloses; 20-40%, and pectin; 20-30%). Secondary cell walls are made up of cross-linked hetero-matrix of cellulose, hemicelluloses, and lignin and are laid down on the interior of the primary cell walls (Figure 1.4). The relative abundant of these three polymers varies depending on the type of biomass (29).





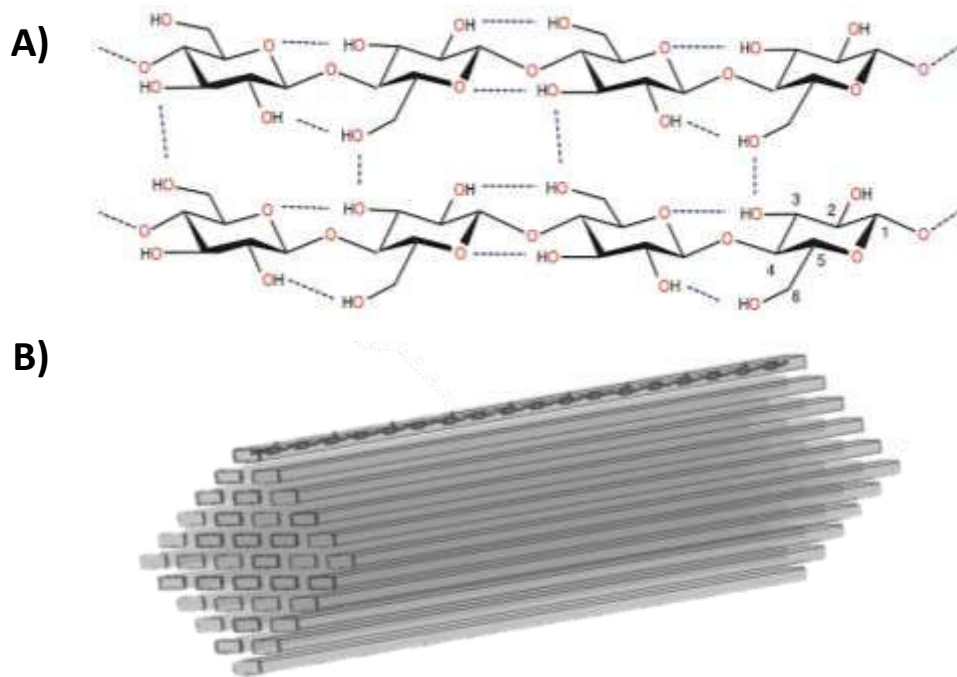
**Figure 1.4: Illustration of a plant cell walls.**

The features of the plant cell wall are shown. Relative thickness of the cell wall layers, relative abundance and specific localization of the various cell wall components, such as pectin, cellulose, hemicellulose, lignin and protein are illustrated. Image is reproduced from Achuyathan *et al.*, 2010 (30).

### 1.3.1 Cellulose

Cellulose is the main component of plant cell walls and the most abundant organic compound in terrestrial ecosystems. A linear cellulose polysaccharide consists of hundreds to over ten thousand  $\beta$ -1,4 linked glucose units (Figure 1.5A). The cellulose chains aggregate into microfibrils via hydrogen bonding and van der Waals interactions shows in Figure 1.5B (31,32). These microfibrils are crystalline, non-soluble, and challenging for enzymatic saccharification. Consecutive sugars along chains in crystalline cellulose are rotated by 180 degrees, meaning that the disaccharide (cellobiose) is the repeating unit. Cellulose tends to contain both well-ordered crystalline regions and disordered, more amorphous regions. While its recalcitrance to

enzymatic degradation may contribute problems, one big advantage of cellulose is its homogeneity. Complete depolymerization of cellulose yields just one product, glucose. Cellulose deconstruction is critical to ecosystem functioning and the global carbon cycle. Only selected lineages of fungi and bacteria have evolved the ability to efficiently degrade this highly recalcitrant substrate (33).



**Figure 1.5: Overview of cellulose structure.**

A) Cellulose chain (partial structure) consists of glucose monomers depicting an internal network of hydrogen bonds. The carbon numbering scheme is depicted on one glucosidic unit. Image is reproduced from Hemsworth *et al.*, 2013 (34). B) Simplistic sketch of a  $\beta$ -cellulose microfibril. Parallel cellulose chains aggregate into crystalline structures called microfibrils. Illustration is reproduced from Horn *et al.* (35).

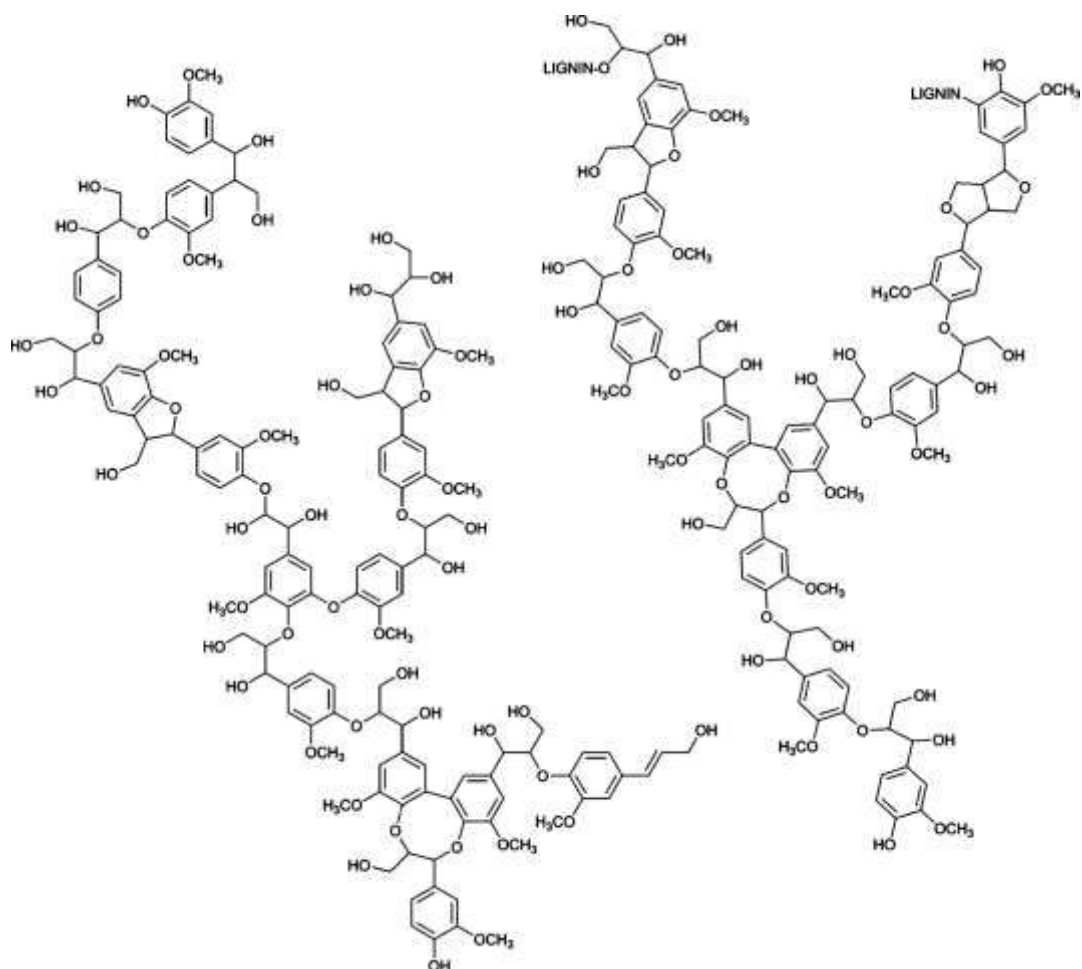
### 1.3.2 Hemicellulose

Hemicelluloses are a large group of polysaccharides found in the primary and secondary cell walls. Hemicelluloses are built up by pentoses (D-xylose, D-arabinose), hexoses (D-mannose, D-glucose, D-galactose) and sugar acids (36). These are including  $\beta$ -glucan, xylan, xyloglucan, arabinoxylan, mannan, galactomannan, arabinan and so on. The hemicelluloses found in cereal straws are largely represented as complex heteropolysaccharides with various degrees of branching of the  $\beta$ -1,4-linked xylopyranosyl main chain structure (37). Softwood contains mainly

glucomannans, while in hardwood xylans are most common. Hemicelluloses interconnect with other cell wall components through covalent bonds and secondary forces (38). Both the cellulose and hemicellulose can be broken down enzymatically into the component sugars which may be then fermented to ethanol. Multiple classes of enzymes are required for effective degradation of cellulose and hemicelluloses (39). The break down process involves enzymes like glycoside hydrolases, carbohydrate esterases, polysaccharide lyases, endo- hemicellulases and others, the concerted action of which hydrolyze glycosidic bonds, ester bonds and remove the chain's substituents or side chains. These include endo-1,4- $\beta$ -xylanase,  $\beta$ -xylosidase,  $\beta$ -mannanase,  $\beta$ -mannosidase,  $\alpha$ -glucuronidase,  $\alpha$ -L-arabinofuranosidase, acetylxylan esterase and other enzymes (40).

### 1.3.3 Lignin

While cellulose and hemicellulose are built from carbohydrates, the random structure in the tridimensional network inside the cell which consists of lignin is built up by oxidative coupling of three major C<sub>6</sub>-C<sub>3</sub> (phenylpropanoid) units, namely syringyl alcohol (S), guaiacyl alcohol (G), and *p*-coumaryl alcohol (H) (41,42). Lignins are highly branched, substituted, mononuclear aromatic polymers in the cell walls of certain biomass, especially woody species, and are often bound to adjacent cellulose fibers to form a lignocellulosic complex (Figure 1.6). This complex and the lignins alone are often quite resistant to conversion by microbial systems and many chemical agents. The lignin-hemicellulose complex surrounds the cellulose with which it is bound through extensive hydrogen bonding to form a supramolecular structure that protects the cellulose and is the reason for biomass recalcitrance (30). Lignin is one of the most abundant natural polymers expected to play an important role in the near future as a raw material for the production of bio-products. Large amounts of lignin are produced each year by the pulp and paper industry as by-products of delignification. The amount of lignin in plants vary widely, and is normally in the range of 20-30% by weight (43). Lignin is an aromatic hetero-biopolymer role as the constituent of an internal cell wall in all vascular plants including the herbaceous varieties. In the plant cell wall, hemicelluloses serve as a connection between lignin and cellulose and gives the whole cellulose-hemicelluloses-lignin network structure more rigidity besides 20 different types of bonds present within the lignin itself (44). Owing to its cross linking, lignin *in-situ* is usually insoluble in all solvents, unless it is degraded by physical or chemical treatments.



**Figure 1.6: Structure of lignin.**

Unlike most natural polymers, such as cellulose and starch, which consist of a single monomer and intermonomeric linkage, lignin is a network polymer made up by oxidative coupling of three major C<sub>6</sub>-C<sub>3</sub> (phenylpropanoid) units with many carbon-to-carbon and ether linkages, such as  $\beta$ -O-4, 4-O-5,  $\beta$ - $\beta$ ,  $\beta$ -1,  $\beta$ -5, and 5-5' (45). It is covalently linked to polysaccharides, forming a lignin-hemicellulose network made up of benzyl-ether, benzyl-ester, and phenyl-glycoside bonds (44). Image is reproduced from Crestini *et al.*, 2011 (46).

## 1.4 LIGNOCELLULOSE DIGESTION IN NATURE

Despite lignocellulose being a hard-to-digest structure, a range of animals and microbes can digest lignocellulosic biomass in nature. Animals such as termites (47,48), beetles (49,50) and marine wood borers (51,52) have evolved to live on a diet of lignocellulose. Microbes are the main source of lignocellulose digestion in these animal systems and also serve to turn over woody biomass in the environment. Filamentous fungi are major degraders of lignocellulosic biomass in the environment due to their ability to degrade lignin. This is mostly achieved

through the action of enzyme-mediated oxidative free radical attack of the lignin, exposing the polysaccharides for hydrolytic enzyme attack. Many biomass-degrading organisms secrete synergistic cocktails of individual enzymes with one or several catalytic domains per enzyme, whereas a few bacteria synthesize large multi-enzyme complexes (cellulosomes) which contain multiple catalytic units per complex (39,53). The cellulosomes present in obligate anaerobic microbes contain many catalytic units per individual complex, linked to a single carbohydrate binding module (CBM) bearing scaffoldin via cohesin–dockerin interactions (54,55). Although lignocellulolytic fungi such as *Aspergillus*, *Penicillium*, *Schizophyllum*, *Trichoderma*, *Phanerochaete* and *Sclerotium* species can secrete industrial quantities of extracellular enzymes, bacterial enzyme production can be more cost-efficient (56).

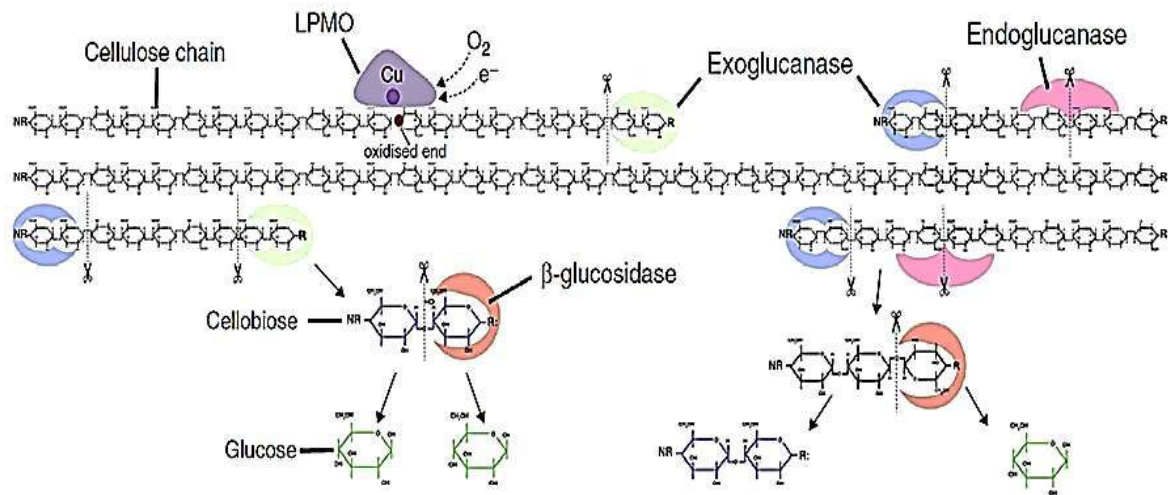
#### **1.4.1 Aerobic lignocellulolytic bacteria**

The rapid growth and multi-enzyme complexes with increased functionality and specificity ensure that the lignocellulolytic bacteria tolerate larger and more diverse environmental stresses during lignocellulose decomposition and occupy wider niches than filamentous fungi (57). A few bacterial species are currently known to degrade both cellulose and lignin. Among them are member of the genera *Pseudomonas* (order *Pseudomonadales*), *Streptomyces*, as well as *Cellulomonas* (order *Actinomycetales*) which are likely to employ extracellular laccases and peroxidases to attack lignin (42,58,59). With respect to recent trends in lignocellulose decomposition research, the broad studies conducted by scientists on laccases and peroxidases have identified that aerobic lignocellulolytic microbes exhibit free and complex enzymes synergy which require terminal or intermediate electron acceptors to support the decomposition under limited carbon source conditions (60).

#### **1.4.2 Glycosyl Hydrolases (GHs)**

In Nature, the enzymatic deconstruction of cellulose and hemicellulose is achieved by the orchestrated action of various carbohydrate-active enzymes (CAZymes), typically acting together as a cocktail with synergistic activities and modes of action (61) (see Figure 1.7). GHs are important enzymes that cleave glycosidic bonds that exist in cellulose and hemicellulose. The capacity of GHs are aided by polysaccharide esterases that remove methyl, acetyl and phenolic esters, permitting the GHs to break down hemicelluloses (62). Additionally, polysaccharides are depolymerised by the activity of polysaccharide lyases (PL) (63).

More recently, the action of lytic polysaccharide monooxygenases (LPMOs) has been shown to be critical for efficient cellulose hydrolysis by the oxidative cleavage of difficult to access glucans on the surface of crystalline cellulose microfibrils (64,65). Across the Tree of Life, the GH cocktail composition varies significantly in composition depending on the kingdom of the cellulolytic organism, the evolutionary pressure, and the environmental niche of the cellulolytic habitats (61). Lignocellulose-utilising creatures secrete some GHs, however most benefit from a mutualism relationship with their enzyme-secreting gut microflora, a particular example in termites. However, in shipworms the system consists of GH-secreting and LPMO-secreting bacteria that separate from the site of digestion, whereas, the isopod *Limnoria* solely relies on endogenous enzymes (51,52).



**Figure 1.7: Schematics of microbial mechanisms of lignocellulose degradation.**

Aerobic cell-free cellulase system employed by most of bacteria and fungi. Cellulose is digested via the synergistic interaction of individual GH and LPMO secreted enzymes. NR-, non-reducing ends; -R, reducing ends. Image is reproduced from Cragg *et al.* (51).

### 1.4.3 Lytic polysaccharide Monooxygenase (LPMO)

Lytic polysaccharide monooxygenases (LPMOs) are a type of enzyme which requires a reducing agent (either a small molecule reducing agent or cellobiose dehydrogenase), oxygen, and a copper (Cu) ion bound in the active site for activity (57,66,67). The glycoside hydrolases, pectate lyases, esterases and the new LPMOs are all often found as parts of multi-modular enzymes that contain substrate-targeting carbohydrate-binding modules (68). These enzymes are important for the decomposition of recalcitrant biological macromolecules such as chitin and plant cell wall

polymers (61,69). Since their discovery, LPMOs have become integral factors in the industrial utilization of biomass, especially in the sustainable generation of cellulosic bioethanol (70). LPMOs were originally designated as GH61 and CBM33, but now classified as Auxiliary Activity (AA) 9, AA10 and AA11 in the CAZy database (69,71,72). The reclassification of GH61 to AA family was based on the findings that although some GH61s appeared to have weak endoglucanase activity, enzymes from this group could enhance enzymatic depolymerization of cellulose into soluble sugars by GHs. The AA9 contains fungal enzymes and AA10 predominantly bacterial enzymes (73). 3D structural analyses of lytic polysaccharide monooxygenases of both bacterial AA10 (previously CBM33) and fungal AA9 (previously GH61) enzymes uncovered structures with  $\beta$ -sandwich folds containing an active site with a metal coordinated by an N-terminal histidine (68). LPMO are copper-containing enzymes (metalloenzymes) that depolymerize recalcitrant polysaccharides by breaking down the glycosidic bonds and direct oxidative attack on the carbohydrate polymer chains through a flat site with a centrally located copper atom (65). LPMOs cleave the polysaccharide chain by utilising the oxidative capacity of molecular oxygen to scission a glycosidic C-H bond. To break these bonds, LPMOs activate oxygen, in a reducing agent dependent manner, at a copper-containing active site known as the histidine brace (34,74,75). Working together with both canonical polysaccharide hydrolases and other electron transfer compounds, these enzymes significantly boost the deconstruction of polysaccharides into oligosaccharides. Consequently, they have real potential for improving the production of biofuels from lignocellulose sustainable sources.

#### 1.4.4 *Cellulomonas fimi* ATCC® 484™

At a biochemical level, one of the best understood cellulose-degrading bacterial systems is derived from *Cellulomonas fimi*. *C. fimi* is a Gram-positive coryneform bacterium, a group which includes a range of cellulolytic facultative anaerobes. The *C. fimi* genome encodes an array of glycosyl hydrolases (GHs) and Carbohydrate Active Enzymes (CAZymes) with similar numbers (176 CAZymes) to those found in other cellulomonads (*Cellulomonas Uda*, *Cellulomonas flavigena* and *Cellulomonas* sp. CS-1) but it has a slightly lower number of CAZymes compared to other cellulase-secreting bacteria such as *Fibrobacter succinogenes* (190 CAZymes), *Streptomyces coelicolor* (268 CAZymes), *Streptomyces bigichengensis* (276 CAZymes), and *Streptomyces davawensis* (337 CAZymes) (76). Despite the lesser number of CAZymes, previous studies reported its proficiency and capability to utilize cellulose by expressing extracellular cellulases which include exoglucanases (39,77–80), and endoglucanases (80,81) towards digestion of diverse set of carbohydrates including crystalline cellulose, *in vitro*. From the reported studies, 30 structures of proteins from *C. fimi* are available in the Protein Data Bank (PDB) and 10 well-characterized enzymes have been fully reviewed in Universal Protein Resource (UniProt KB) database (see Table 1.1) regarding to the mode of action of their catalytic and carbohydrate-binding module of actions towards various of polysaccharides (82–85). *C. fimi* is still of interest due to significant gaps in knowledge with regard to its ability to digest recalcitrant lignocellulose. Interestingly, the *C. fimi* genome not reveal any homology to typical cellulosome components such as scaffoldins, dockerins or cohesins which exists commonly in facultative anaerobes enzymatic systems (86). This is in contrast with other reports where *C. fimi* was reported to have a mutually exclusive approach by using both “secreted-enzyme” and “surface-enzyme” strategies during cellulose digestion other than the reported carbohydrate-binding proteins (87,88). This characteristic was only found in two cellulolytic facultative anaerobes bacteria including *C. fermentans* (86).



**Table 1.1: Characterized *C. fimi* proteins involving in polysaccharide degradation as listed in the UniProt KB database.**Reviewed *C. fimi* characterized proteins as curated in the UniProt KB database accessed in April, 2017.

UniProt ID	UniProt Entry name	Length	Protein name	Gene names	Catalytic activity	Protein family/CAZy	References
P14090	GUNC_CELFA	1,101	Endoglucanase C	cenC Celf_1537	Endohydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose, lichenin and cereal $\beta$ -D-glucans.	<b>CBM4.</b> Carbohydrate-Binding Module Family 4. <b>GH9.</b> Glycoside Hydrolase Family 9.	(89–92)
P50899	GUXB_CELFA	1,090	Exoglucanase B	cbhB cenE, Celf_3400	Hydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose and cellotetraose, releasing cellobiose from the non-reducing ends of the chains.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>GH48.</b> Glycoside Hydrolase Family 48.	(77,93,94)
P50400	GUND_CELFI	747	Endoglucanase D	cenD	Endohydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose, lichenin and cereal $\beta$ -D-glucans.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>GH5.</b> Glycoside Hydrolase Family 5.	(93)
P07984	GUNA_CELFI	449	Endoglucanase A	cenA	Endohydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose, lichenin and cereal $\beta$ -D-glucans.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>GH6.</b> Glycoside Hydrolase Family 6.	(39,89,95,96)

...continued

UniProt ID	UniProt Entry name	Length	Protein name	Gene names	Catalytic activity	Protein family/CAZy	References
P26255	GUNB_CELFI	1,045	Endoglucanase B	cenB	Endohydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose, lichenin and cereal $\beta$ -D-glucans.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>CBM3.</b> Carbohydrate-Binding Module Family 3. <b>GH9.</b> Glycoside Hydrolase Family 9.	(81,97)
P50401	GUXA_CELFA	872	Exoglucanase A	cbhA Celf_1925	Hydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose and cellotetraose, releasing cellobiose from the non-reducing ends of the chains.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>GH6.</b> Glycoside Hydrolase Family 6.	(93,98)
Q7WUL4	HEX20_CELFI	496	$\beta$ -N-acetylhexosaminidase	hex20 hex20A	Hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl- $\beta$ -D-hexosaminides.	<b>GH20.</b> Glycoside Hydrolase Family 20.	(99)
P54865	XYND_CELFI	644	Bifunctional xylanase/ deacetylase	xynD	Endohydrolysis of (1->4)- $\beta$ -D-xylosidic linkages in xylans.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>GH11.</b> Glycoside Hydrolase Family 11.	(100–103)

...continued

UniProt ID	UniProt Entry name	Length	Protein name	Gene names	Catalytic activity	Protein family/CAZy	References
P07986	GUX_CELFI	484	Exoglucanase/ xylanase	cex xynB	Hydrolysis of (1->4)- $\beta$ -D-glucosidic linkages in cellulose and cellotetraose, releasing cellobiose from the non-reducing ends of the chains.  Endohydrolysis of (1->4)- $\beta$ -D-xylosidic linkages in xylans.	<b>CBM2.</b> Carbohydrate-Binding Module Family 2. <b>GH10.</b> Glycoside Hydrolase Family 10.	(39,85,96,97,104–109)
Q7WUL3	NAG3_CELFI	564	$\beta$ -N-acetylglucosaminidase/b-glucosidase	nag3 nag3A	Hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl- $\beta$ -D-hexosaminides.  Hydrolysis of terminal, non-reducing $\beta$ -D-glucosyl residues with release of $\beta$ -D-glucose.	<b>GH3.</b> Glycoside Hydrolase Family 3.	(99)

## **1.5 AIMS OF THE PROJECT**

As discussed in 1.3, lignocellulosic biomass such as wheat straw is an attractive renewable and environmentally-friendly resource that could be used as a substrate for biofuel production. The rich-sugar content of lignocellulosic polysaccharides may provide better options and tackle food insecurity concerns. However, there is a need to find new enzymes to improve the degradation of such heterogeneous and recalcitrance biomass. The novel findings of LPMOs involvement in polysaccharide oxidation during the digestion process enlighten the potential of more interesting related enzymes to be discovered. *C. fimi* is a well-studied microorganism because of its inherent function of cellulases in degrading cellulose but very few studies involve lignocellulosic biomass digestion. Its sequenced genome provides information of potentially more new enzymes for degradation on biomass to be explored. The aims of this study are:

1. To characterise the transcriptomic response of *C. fimi* growing on wheat straw, sugarcane bagasse, Avicel and xylan (Chapter 3).
2. To characterise the secreted proteome of *C. fimi* growing on these substrates in order to identify candidate proteins for further studies (Chapter 4).
3. To clone, express and characterise recombinant target proteins (Chapter 5).
4. To explore the response of *C. fimi* adaptive evolution by continuous subculture on wheat straw (Chapter 6).

## 2. General materials and methods

### 2.1 CHEMICAL REAGENTS, SUBSTRATES, AND ORGANISMS

#### 2.1.1 Chemical reagents

Reagents and chemicals described in this thesis were purchased from Sigma-Aldrich (Poole, UK), Fisher Scientific (Loughborough, UK), Melford (Ipswich, UK), New England Biolabs (USA), GE Healthcare (London, UK), Promega (Southampton, UK), Qiagen (West Sussex, UK), Expedeon (Swavesey, UK), Clontech Laboratories (USA), Invitrogen (Paisley, UK), Cambio (Cambridge, UK), Cambridge Biosciences (Cambridge, UK) and Agilent (USA).

Oligonucleotide primers were synthesized and purchased from Integrated DNA Technology (Iowa, USA) and Eurofins Scientifics (Luxembourg).

All buffers were formulated in ultrapure water ( $18.2 \text{ M}\Omega \text{ cm}^{-1}$ ) followed by filtration. Ultra-Pure Water was obtained with an Elga PureLab Ultra water polisher (High Wycombe, UK). All growth media were autoclaved prior use.

#### 2.1.2 Lignocellulose biomass

The wheat straw used in the experiments was provided by a local Yorkshire UK farmer and was milled to 2 mm particles at the Biorenewable Development Centre (York, UK).

The sugarcane bagasse was provided by the Cosan Mill (Ibatè, SP, Brazil). The raw material, previously washed and roughly ground, was dried in a convection oven at  $60^\circ\text{C}$  for 24 h and further ground by knife milling into small pieces (625 mm x 188 mm avg.).

The xylan referred to throughout this thesis was beechwood xylan and therefore primarily glucuronoxylan, purchased from Sigma-Aldrich.

Phosphoric acid swollen cellulose (PASC) was prepared from Avicel (Sigma-Aldrich). Briefly, for every gram of Avicel used, 30 mL of 85% phosphoric acid was added. This was stirred for one hour, on ice, before 100 mL of acetone per gram of Avicel was added, filtered on a glass-filter funnel through MiraCloth® (Merck) and was washed three times with 100 mL ice cold acetone. This was washed again with 500 mL water until the pH was 6.5-7 and homogenized with a blender.

### 2.1.3 Organisms

The *Cellulomonas fimi* strain was purchased from American Type Culture Collection (ATCC), United States. Stock cultures of the bacterium were maintained at -80°C in NB medium containing 15% glycerol. Nutrient broth (NB) was inoculated with a single colony of *C. fimi* before the organism was transferred into basal medium in the experiment. NB was prepared by dissolving 13 g of NB powder in a final volume of 1 L of water. The solution was then autoclaved prior to use.

*Escherichia coli* strains *Gami*<sup>™</sup> (DE3) (Novazymes), *Stellar*<sup>™</sup> Competent cells (Clontech), *SHuffle*<sup>®</sup> and *SHuffle*<sup>®</sup> T7 (New England Biolab) were purchased for heterologous expression of proteins in bacteria. *Aspergillus niger* D15 was kindly received from Peter Punt (TNO, Netherlands) for heterologous expression of proteins in eukaryotes.

## 2.2 MICROBIOLOGY METHODS

### 2.2.1 Buffers

Phosphate buffer saline was prepared to x10 concentration (NaCl 80 g/L, KCl 2 g/L, Na<sub>2</sub>HPO<sub>4</sub> 14.4 g/L, KH<sub>2</sub>PO<sub>4</sub> 2.4 g/L) and diluted as necessary with H<sub>2</sub>O. After preparation the solution was pH adjusted to 7.3, autoclaved and stored at room temperature.

### 2.2.2 Media for bacterial growth

Cultures of *C. fimi* were grown on basal medium which contained: 1 g of NaNO<sub>3</sub>, 1 g K<sub>2</sub>HPO<sub>4</sub>, 0.5 g KCl, 0.5 g MgSO<sub>4</sub>·7H<sub>2</sub>O and 0.5 g yeast extract per litre of medium, pH 7.0 (Hitchner and Leatherwood, 1980) at 30°C with shaking at 180-200 rpm. Medium which was supplemented with either glucose (added to the autoclaved medium), avicel, beechwood xylan, wheat straw or sugar cane bagasse to the final concentration of 2 g/L.

### 2.2.3 Agar plate and slant preparation

Agar plates were prepared by the addition of 1% agar to the media of choice before autoclaving. The agar was then allowed to cool, before the addition of antibiotics, if appropriate and poured into 9 cm petri dishes.

## **2.3 MOLECULAR BIOLOGY TECHNIQUES**

### **2.3.1 Gram-positive bacterial DNA extraction**

DNA was extracted from *C. fimi* grown cultures using a phenol-chloroform method adapted from Cheng and Jiang (110). From the 25 mL grown culture, supernatant was transferred into 50 mL centrifuge tube. Bacterial pellet was separated from culture supernatant from different time points by centrifugation at 4,000 x g for 15 min at 4°C. Supernatant was removed and the cells pellet were collected into 1.5 mL Eppendorf tube by resuspended in freshly made sterile lysis buffer (0.5 mL of 0.5 mM Tris-EDTA buffer, 1 mg/mL lysozyme) before being incubated at 37°C for 30 min in warm bath. 30 µL of 10% SDS and 10 µL of 20 mg/mL proteinase K were added and mixed well during the lysis process. The mixture was incubated for another 2 h at 37°C.

After the lysis step, genomic DNA was isolated using a standard phenol/chloroform extraction followed by alcohol precipitation. For this step, 20 µL of 10% SDS and 130 µL of 5 M NaCl were added and mixed vigorously by hand. The extraction mixture was incubated for 30 min on ice before being centrifuge at 10,000 x g. Bacterial DNA that contained in the supernatant fraction was collected into 2 mL Eppendorf tube. For the purpose of protein removal, 700 µL of phenol was added into the supernatant and vortex before being centrifuge for 5 min at 10,000 x g. Supernatant was collected into a fresh Eppendorf tube and 700 µL of chloroform was added. The mixture was vortex and centrifuge at 5000 x g for 5 min.

The supernatant was collected into a fresh 1.5 mL Eppendorf tube to precipitate the nucleic acid. For this purpose, 1/10 volume of 3 M of sodium acetate (NaAc) pH 5.2 and 1 volume of isopropanol were added to the mixture. The mixture was mixed gently by inverting the Eppendorf tube several times to clump the DNA before been centrifuge at 10,000 x g for 10 min at 4°C. After the centrifugation, supernatant was removed and DNA pellet was washed by adding 500 µL of 70% cold ethanol. Sample was centrifuged again at 10,000 x g for 5 min at 4°C to remove the ethanol, and DNA pellet was air-dried for 2 h. To collect the DNA sample, 50 µL of nuclease-free water was added to the DNA pellet and resuspended by pipetting up and down.

The DNA samples were kept at -20°C for further processing. For the culture purity confirmation, a prokaryotic universal 16S forward and reverse primers were used to amplify the 16S ribosomal RNA gene using Polymerase Chain Reaction (PCR). The resulting 16S of *C. fimi* DNA amplicons were sent to GATC *Light-Run* Sequencing Service (GATC, Germany).

### 2.3.2 Fungal cells *A. niger* disruption for DNA extraction

Approximately 0.5 g of fungal biomass was harvested through Miracloth with 22-25  $\mu\text{m}$  pore size (CalBiochem, Merck KGaA, Germany) after growth in shake flasks with an appropriate media either in nutrient agar (NA) or Potato Dextrose Agar (PDA). The fungal biomass was ground under liquid nitrogen using a pestle and mortar, before 500  $\mu\text{L}$  Cetyltrimethylammonium Bromide (CTAB) extraction buffer was added, and the sample transferred to a 2 mL screw cap tube. To this, 800  $\mu\text{L}$  of Phenol/Chloroform/Isoamyl alcohol (25:24:1) mix was added and vortexed briefly to precipitate the DNA within the sample, an equal volume of ice-cold 100% isopropanol was added and incubated for 1 hour. DNA was pelleted by centrifugation at 13,000 rpm for 10 min and supernatant was removed without disturbing the pellet. The pellet was then washed with 80% ethanol, before being re-suspended in DNase free water. The DNA samples were kept at  $-20^{\circ}\text{C}$  before been used for PCR reaction as described in Section 2.3.3.

### 2.3.3 cDNA synthesis

cDNA was synthesised from RNA, that had been DNase treated with Room Temperature Stable (RTS) DNase kit (Mobio, UK) using the standard protocol described by the manufacturer. cDNA synthesis was performed using SuperScript II Reverse Transcriptase (Invitrogen) kit and the standard protocol with 100 ng of random hexamers (Thermo Scientific) per 20  $\mu\text{L}$  of reaction prepared.

### 2.3.4 Polymerase Chain Reaction (PCR)

PCR reactions were performed on a PTC-200 Peltier Thermal Cycler (MJ Research). Reactions were performed with Q5<sup>®</sup> High-Fidelity DNA Polymerase (NEB, UK) as per manufacturer's instructions. The PCR reaction mix and temperature cycling used are described in Table 2.1 and Table 2.2 for 20  $\mu\text{L}$  reactions.

**Table 2.1: Polymerase chain reaction components.**

Component	Volume	Final concentration
Nuclease free water	To 20 $\mu\text{L}$	N/A
5x Q5 <sup>®</sup> HF buffer	4 $\mu\text{L}$	1x
10 mM dNTPs	1 $\mu\text{L}$	200 $\mu\text{M}$
10 $\mu\text{M}$ forward primer	1 $\mu\text{L}$	0.5 $\mu\text{M}$
10 $\mu\text{M}$ reverse primer	1 $\mu\text{L}$	0.5 $\mu\text{M}$
Template	Depending the [DNA]	Depending the [DNA]
Q5 <sup>®</sup> polymerase	0.2 $\mu\text{L}$	0.4 units



**Table 2.2: Polymerase chain reaction thermocycling conditions.**

Primer annealing temperature was calculated using NEB Tm calculator (<http://tmcalculator.neb.com/#!/>)

Step	Temperature	Time
Initial denaturation	98°C	30 seconds
Denaturation	98°C	10 seconds
Annealing	As determined for each primer	30 seconds
Extension	72°C	30 seconds per kb
Cycling	Cycle between denaturation and extension	29 cycles
Final extension	72°C	10 min
Hold	4°C	N/A

### 2.3.5 Agarose gel electrophoresis

DNA fragments were separated by agarose gel electrophoresis. To prepare agarose gels 1% (w/v) agarose was dissolved, by microwave heating, in 0.5% Tris/Borate/EDTA (TBE) buffer. After cooling, 0.00005% ethidium bromide (EtBr) was added and the solution was poured into a cast and the well comb added. This was then left to set for 30 min, before being placed in an electrophoresis tank containing 0.5% TBE buffer and removal of the comb. Sample buffer (0.01%) was added to DNA samples and mixed with pipetting before being loading into each well, alongside a commercial DNA ladder. An electric current was then generated at 130 V for 30 min by a BioRad PowerPac 3000 to migrate the negatively charged polynucleotides towards the cathode. After completion UV illumination was used to visualise DNA bands using a UVItec gel documentation system.

### 2.3.6 Plasmid extraction

Plasmids were purified from *E. coli* cells that had been grown overnight in 5 mL of LB, shaking at 180 rpm in 37°C. Plasmids were extracted from these cells using QIAPrep Miniprep Kit (QIAGEN, USA) following the manufacturer's instructions.

### **2.3.7 PCR clean-up**

DNA from PCR reactions were purified according to manufacturer's instructions using Wizard PCR clean up kit and typically eluted into 30-50  $\mu\text{L}$  nuclease-free water depending the final concentration desired.

### **2.3.8 Nucleotide quantification**

DNA and RNA were quantified using NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific).

### **2.3.9 Sanger DNA Sequencing**

DNA was sequenced by GATC Biotech (Germany) using LIGHTRun Sanger sequencing service. Sample sent for this service contained 5  $\mu\text{L}$  of 80-100 ng of purified DNA in DNase-free water along with 5  $\mu\text{M}$  the appropriate primer. DNA fragments were sequenced using both forward and reverse primers to ensure complete coverage of the amplicon, reads were consolidated using BioEdit® software.

### **2.3.10 In-Fusion™ cloning**

All cloning was performed using Clontech's InFusion system. This system fuses the gene of interest and linearized vector by recognising 15 bp complementary regions at their ends. 15 bp overlaps on the gene of interest were added to the target gene, and destination the vector was linearised via PCR (Section 2.3.3). The cloning reaction was set up as per the manufactures instructions with 2  $\mu\text{L}$  In-Fusion® HD Enzyme Premix, 2  $\mu\text{L}$  PCR linearized vector, 1  $\mu\text{L}$  PCR fragment and 1  $\mu\text{L}$  Cloning Enhancer®, before being brought up to 10  $\mu\text{L}$  with DNase-free water. The reaction was then mixed, and incubated for 15 min at 37°C, followed by a 15-minute incubation at 50°C and placed on ice. The 2  $\mu\text{L}$  of this reaction was then immediately use for the transformation of competent cells, whilst the remaining was stored at -20 °C.

### **2.3.11 DNA restriction digests**

Typically, a reaction mix prepared for DNA digestion contained 7  $\mu\text{L}$  of DNA sample, 5 U of restriction enzyme, 1 x restriction buffer (recommended by manufacturer) in a total volume of 10  $\mu\text{L}$ . Samples were incubated for 60 min at room temperature as recommended by the manufacturer for the specific restriction enzymes being used.

### **2.3.12 Transformation of competent cells**

Competent cells were thawed on ice, whilst 2  $\mu\text{L}$  of In-Fusion<sup>®</sup> cloning reaction was pipetted into round-bottom 10 mL tubes. Once thawed, 50  $\mu\text{L}$  of cells were gently mixed with the cloning reaction and the tubes left on ice for a further 20 min. The cells were then heat-shocked for 45 seconds by placing them in a water bath, pre-heated to 42 °C before placing them back in ice for 2 min. Commercial Super Optimal broth with Catabolite repression (SOC) medium (Invitrogen) warmed to 37 °C, was then added to the transformation to a final volume of 500  $\mu\text{L}$ . The cells were left to shake (180 rpm) at 37°C for an additional hour before 100  $\mu\text{L}$  was spread onto an agar plate with an appropriate antibiotic.

## **2.4 PROTEIN METHODS**

### **2.4.1 Bradford assay for protein quantification**

The protein concentration in solutions was measured by Bradford (111) assay, performed in 96-well plate format. Briefly, 300  $\mu\text{L}$  samples diluted to be within the sensitivity range of the assay, were mixed with 10  $\mu\text{L}$  of Quick-Start<sup>®</sup> Bradford Protein Assay solution (BioRad), and then absorbance was measured after 10-minute incubation at room temperature. Concentrations were then determined through the comparison of known protein concentrations of bovine serum albumin (BSA).

### **2.4.2 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)**

SDS-PAGE experiments were performed using a discontinuous polyacrylamide gel system and sodium dodecyl sulfate (SDS) to denature the proteins, using a Mini-Protean Tetra cell apparatus

(Bio-Rad, USA). The resolving gel was composed of 10 mL of 1.5 M, Tris-HCl pH 9, 0.4 % (v/v) tetramethylethylenediamine (TEMED), 0.4 % (w/v) SDS, 4.2 mL H<sub>2</sub>O and 3.2 mL 40 % acrylamide/bis acrylamide. The gel solution was de-gassed and acrylamide polymerisation was induced by the addition of 0.1 mL 10 % (w/v) ammonium persulfate. The solution was quickly transferred into a pre-assembled gel apparatus and allowed to solidify. The stacking gel was composed of 5 mL stacking buffer (0.14 M Tris-HCl pH 6.8, 0.11 % (v/v) TEMED, 0.11 % (w/v) SDS) and 0.5 mL 40 % acrylamide/bis acrylamide. The gel solution was de-gassed and 0.05 mL 10 % (w/v) ammonium persulfate was added to aid gel polymerization. The solution was quickly transferred on top of the resolving solution in the pre-assembled gel apparatus, the well comb added and the gel was allowed to solidify.

Protein samples were mixed with appropriate volume of 2X loading buffer composed of 100 mM Tris-HCl pH 6.8, 20% (v/v) glycerol, 4% (w/v) SDS, 0.2 M dithiothreitol (DTT), 0.2% (w/v) bromophenol blue, boiled at 95 °C for 5 min and loaded into the wells of the stacking gel, located in an assembled tank filled with running buffer (25 mM Tris pH 8.3, 192 mM glycine, 0.1% (w/v) SDS). Protein samples were run alongside 10 µL pre-stained broad range protein marker to allow estimation of protein weights. Once the samples were loaded, gels were run at 100 V as samples moved through stacking gel and 200 V hereafter, until the marker eluted from the gel. To visualize the protein bands, the gel was washed twice with H<sub>2</sub>O and then stained with InstantBlue dye reagent (Expedeon Inc., USA) according to the manufacturer's recommendation.

### **2.4.3 Western blotting**

Protein samples ran on the low-concentration of agarose gel (8% or less) were blotted on to a nitrocellulose membrane (Protan™ BA85) using a Trans-Blot® wet transfer cell (Biorad) using a modified method. Wet transfer was chosen especially for large size of PKDP1 in order to minimize the prone to failure of blotting due to drying of the membrane using semi-dry transfer. A standard buffer for Tris/Glycine/SDS running buffer (25 mM Tris/190 mM glycine), with a low concentration of SDS (final concentration of 0.1%) to avoid the precipitation of protein in the gel that possibly hindering the transfer. An additional of methanol to a final concentration of 10% in the transfer buffer was to remove SDS and guard against protein precipitation. Transfer was done at 80 V for 20 min.

Once blotted, the membrane was stained with 0.1 % (w/v) poncean S in 5 % (v/v) acetic acid to check all protein bands had transferred to the membrane. The membrane was then incubated with blocking buffer (5 % (w/v) skimmed milk powder in 1x Tris-buffered saline-tween (TBST) buffer (50 mM Tris, 150 mM NaCl, 0.05 % Tween 20, pH 7.6)) for 75 min with gentle rocking. After blocking, the membrane was washed three times with 1x TBST buffer for 5 min with gentle rocking and then incubated with blocking buffer supplemented with anti-GST–Peroxidase Conjugate antibody (Sigma) at 1:10,000 dilution for 90 min. The membrane was then washed three times with 1x TBST buffer for 5 min with gentle rocking. Horseradish peroxidase (HRP) activity was detected using SuperSignal West Pico Chemiluminescent Substrate kit (Thermo Scientific) according to manufacturer's instructions.

#### **2.4.4 Protein concentration by centrifugation**

Protein solutions were concentrated via centrifugation using 20 mL Vivaspin20 10,000 MWCO (Sartorius) concentrators, with PES membrane. Samples were routinely spun in spin out bucket at 4,000 x g until the desired protein concentration had been reached.

## 3 Growth and transcriptomic studies of *Cellulomonas fimi*

### 3.1 INTRODUCTION

The decomposition of plant cell wall polysaccharides is one of important metabolic processes occur in nature environment. Animals and microbes live in limited carbon sources have been evolved to efficiently degrade lignocellulosic substrates by depolymerization mechanisms using their enzymes (61). *Cellulomonas fimi* is a free-living, Gram-positive, non-spore forming, facultative anaerobic rod. Soil is most common habitat of *C. fimi* that was isolated from. In the Carbohydrate active enzyme (CAZy) database (76) has listed variety of sugar active enzymes produce by *C. fimi* with at least two cellobiohydrolases (Cel6B, and Cel48A), and four endoglucanases (Cel6A, Cel9A, Cel9B and Cel5A) depending on the carbon source being supplied (112,113). In addition, *C. fimi* also encoded numbers of xylosidases including 11  $\alpha$ -xylosidases of GH3 family, and 7  $\beta$ -xylosidase of GH43 family. To date, total of 112 glycosyl hydrolases (GH) have been listed in CAZy identified in *C. fimi*. The cellulase system in *C. fimi* involved synergistic function which revealed by Mansfield and Meder (114). Given the capacity of this organism to grow efficiently with cellulose and xylan substrates and considering the complexity in chemical linkages within natural xylan, it is interesting to explore potentially new enzymes which may co-express and concerted to depolymerization of more complex native lignocellulosic substrates such as wheat straw and sugarcane in *C. fimi*. Furthermore, the genome of *C. fimi* has recently been fully sequenced (86) and this bacterium harbors at least 131 genes to encode CAZymes; either glycoside hydrolases, carbohydrate esterases, pectate lyases, and accessory protein of carbohydrate binding modules, excluding glycoside tranferases.

RNA sequencing (RNA-seq) is a method transforming how transcriptomes are studied and provides a highly sensitive read-out of responses in a particular system at the genome-wide level. Bacterial transcriptome analyses have been empowered by the development of deoxyribonucleic acid (DNA) microarrays and high-density tiling arrays (115,116). These consist of hundreds of thousands of DNA oligonucleotide probes representing both DNA strands of a particular genome. The microarray approach has been largely used to study bacterial transcriptomes and the major advantage of tiling arrays is that they can be used to interrogate the boundaries of the entire set of transcripts in a cell without taking account of genome

annotation (117,118). Consequently, the discovery of many new RNA molecules e.g. regulatory small RNAs (sRNAs) benefited from this technical advance including the comprehensive transcriptomes of *Bacillus subtilis* (115), *Halobacterium salinarum* (119) and *Escherichia coli* (120).

Although high-density tiling arrays were able to provide comprehensive information of transcriptome complexity in microbial cells without genome annotation, the array-based approach is limited by a high rate of noise background in the data due to signal saturation that leads to low dynamic range of detection, the incapacity to detect low copy number transcripts and cross-hybridization (121). The development of next-generation sequencing (NGS) and combination with mRNA enrichment methods have drastically increased the gene expression analytical capacity using high-throughput transcriptome sequencing techniques (122). The preference of this sequencing technique has been largely influenced by the cost and depth of sequencing (123). With regard to the conventional Sanger sequencing method, the main advantage offered by NGS methods is the cost-effective production of large volumes of sequence data useful for the identification and qualification of unusual transcripts even without early knowledge of a particular gene (124,125). Unlike a hybridization-based array approach, RNA-seq allows unambiguous mapping of transcripts to specific regions of the genome with single-base resolution with lower background noise (126). Besides, the accurate quantification of a transcriptome of known genes, RNA-seq also enables scientists to determine correct gene annotations, novel genes and RNAs, as well as expressed single-nucleotide polymorphisms with high levels of reproducibility (127–129). To date, four NGS systems are commercially available including the 1) Ion Torrent system, based on the use of a semiconductor-based sequencing technique; 2) the Illumina sequencing system, based on sequencing by synthesis; 3) the Roche 454 system, based on the pyrosequencing; and 4) the SOLiD system, based on sequencing by oligonucleotide ligation and detection (118). The first two techniques; Ion Torrent and Illumina sequencings have been attempted in the *C. fimi* RNA-seq work as presented in this chapter.

Prokaryotic RNA consist of 95% to 99% ribosomal RNA (rRNA) and only the rest e.g tRNA is comprised of useful messenger RNA (mRNA). Moreover, the instability of mRNAs with very short half-lives and the absence of a poly(A) tail at the 3'-end make bacterial transcriptome analysis challenging. Several methods are being used to remove the unwanted fractions (rRNA and tRNA) from the entire RNA samples, including, 1) terminator 5'-phosphate-dependent exonuclease treatment, and 2) strand-specific RNA-seq that mark the transcribed strand. Both methods have

been applied to *C. fimi* RNA samples in three separate experiments that revealed the most suitable method for this species.

The terminator 5'-phosphate-dependent exonuclease method is used to enrich the primary transcriptome by removing processed RNAs with a 5'-monophosphate end of unwanted strands e.g. rRNAs and tRNAs. Biotinylated probes by Ribo-Zero Epicentre® rRNA removal kits (Bacteria kit and Gram-positive kit) were tested that selectively bind rRNA and removed the unwanted transcripts. However, several drawbacks were identified when using this method and the quality of mRNA was compromised as the probes in the kits are not specifically matched to the rRNA strands of *C. fimi*. This becomes major problems with the GC-rich genome and gram-positive species such as *C. fimi*. However, the recently development of strand-specific RNA-seq analysis has improved the mRNA enrichment method by transcript marking approaches. The strand-specific marking could be done either by orientation-dependent adaptor ligation to the 5' and 3' ends of the RNA transcript by RNA ligation, or by chemical modifications using by bisulfite treatment (130) on RNA itself or dUTP incorporation on the second strand cDNA (131). The adaptor ligation approach was applied for the third attempt of *C. fimi* RNA-seq work to conserve the native total RNA from any destruction by chemical modification. Therefore, AnyDeplete, formerly known as Insert Dependent Adaptor Cleavage (InDA-C) technology from NuGEN® was used for targeted depletion of abundant transcripts i.e. rRNA using customized Gene-Specific Primers (GSPs) for this species. With these technologies, several bacterial RNA-seq studies demonstrated that bacterial transcription is not as simple as previously thought (118,132).

Taking advance from the current method for samples preparation of RNA-Seq experiment and the prior knowledge of fully sequenced genome of *C. fimi*, differential gene expression in this species was explored between conditions and range of carbohydrate substrates during bacterial growth using transcriptomic studies. In this chapter, the transcriptional changes of *C. fimi* associated with exposure to five substrates (glucose, Avicel, beechwood xylan, wheat straw and sugarcane bagasse) in laboratory cultures were investigated using NGS RNA-seq technology with the aim to identify any co-expressed genes and novel lignocellulosic processing enzymes.



## 3.2 MATERIALS AND METHODS

### 3.2.1 Experimental setup and growth study

During the growth of *C. fimi* in shake flasks, samples were collected from 25 mL cultures, throughout the growth cycle, with three biological replicates for each of three days of culture. Growth was monitored by measuring OD<sub>600</sub>; to avoid interference from the insoluble wheat straw and sugarcane bagasse, all samples were spun down with quick and slow setting in a centrifuge (500 rpm, 30s) before being transferred into a cuvette and optical density (OD) measurement. Sterile culture supernatants were obtained by centrifugation and filtration (0.22 µm). Proteins were desalted and concentrated using 10 kDa Amicon® Ultra-0.5 Centrifugal Filter Unit (UFC501024, Merck Milipore).

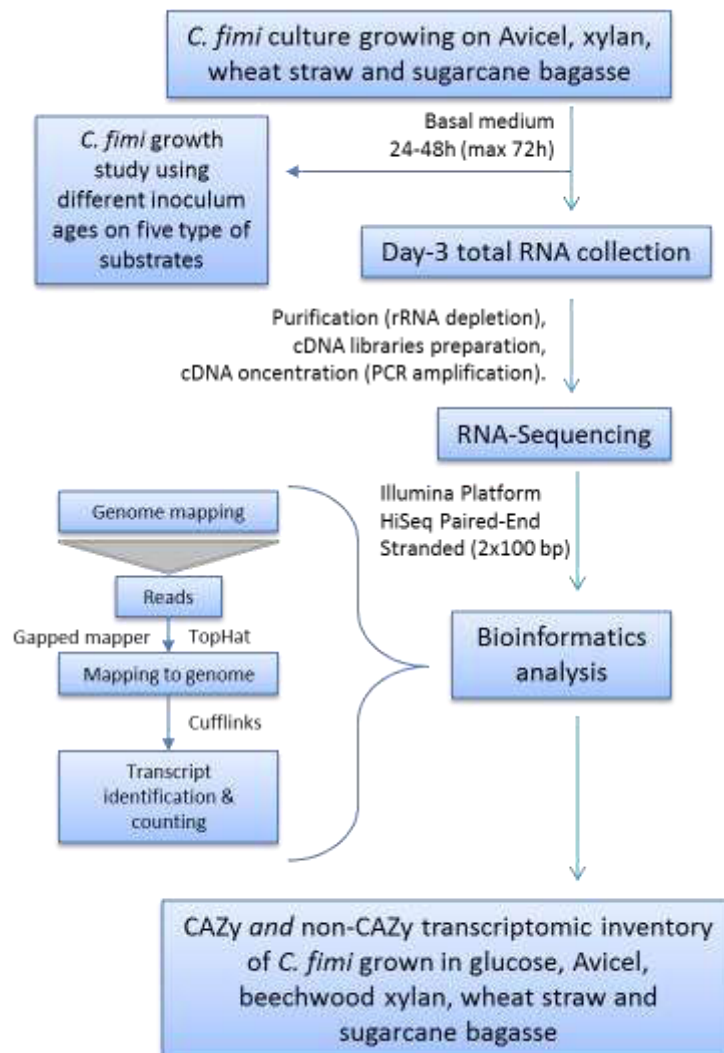


Figure 3.1: Workflow of transcriptome analysis of *C. fimi* transcriptome grown on four types of carbon source.

### 3.2.2 Strain and growth conditions

*C. fimi* ATCC® 484™ was used throughout this study. Details of growing conditions experimental set up and bacterial biomass collection are outlined in general materials and methods (Section 2.2.2). Cultures grown on glucose were also harvested at the same time point (day-3) as those for the other substrates conditions, 0.2% (w/v) of Avicel, beechwood xylan, wheat straw and sugarcane bagasse cultures. Care was taken to process the biomass as soon as possible, and all procedures were performed on ice, unless otherwise stated, with DEPC-treated solutions and sterile equipment to prevent RNA degradation.

### 3.2.3 Total RNA extraction

Total RNA was extracted from the bacterial pellet collected after 3-days of growth by centrifugation of 50 mL culture samples at 4000 x g for 15 min at 4°C. The cells were lysed initially with lysozyme for 30 min at 37°C and then RNA was extracted with TRIzol® reagent (Invitrogen). Total RNA was purified with Direct-Zol® columns (Zymo Research). The DNase treatment was performed using DNase RTS kit (MoBio), followed by an additional clean up step with RNA Clean and a Concentrator kit from Zymo Research. All RNA quality from three biological replicates of 3-days cultures were assessed for the quality of RNA by using RNA ScreenTape assay using the 2200 TapeStation system (Agilent Technologies).

### 3.2.4 Enrichment of mRNA from total RNA

For the RNA-Seq study, two different methods of mRNA enrichment were tested to reduce rRNA reads using kits from Epicentre® Ribo-Zero rRNA Removal Kits (Epicentre, Madison, WI, USA) and NuGEN-Ovation® Universal RNA-Seq system (NuGEN Technologies, Inc.; San Carlos, CA, USA).

### 3.2.5 rRNA depletion using Ribo-Zero rRNA Removal Kit (Bacteria) and Ribo-Zero Removal Kit (Gram-positive)

For this mRNA enrichment, briefly, 5 µg of total RNA sample (with 3 technical replicates) of *C. fimi* was purified from ribosomal RNA using RiboZero™ Bacteria kit (Epicentre) following the manufacturer's instructions. For the first and second RNA-seq trials, the Ion-Torrent whole transcriptome cDNA libraries were prepared according to the manufacturer's protocol in collaboration with Dr. Debs Rathbone, Biorenewable Development Centre (BDC). Sequencing was performed using the 318v2™ Chip of Ion Torrent Personal Genomic Machine (PGM) One Touch™ System.

### 3.2.6 Selective cDNA synthesis for rRNA depletion using NuGEN-Ovation® Universal RNA-Seq System

For the third attempt of RNA-seq work, the removal of rRNA by selective amplification and cDNA library preparation was performed according to the following NuGEN® protocol. Total of 300 ng intact RNA was combined with 3 µg of the random primers (Invitrogen; 3 µg/µl) in a final volume of 11 µl. The reaction was incubated at 70° C for 10 min and placed immediately on ice. The remaining reagents were added to the reaction in a final volume of 20 µl: 1x of first strand buffer (10x; Invitrogen), 10mM of DTT (0.1 M; Invitrogen), 0.5mM of dNTP mix (10 mM; Invitrogen), 20 U of SUPERase-In (20 U/µl; Ambion) and 200 U of SuperScript III (200 U/ µl; Invitrogen). The first strand reaction was incubated at 25°C for 10 min followed by 55°C for 60 min and then placed on ice. The second strand was synthesized by adding 1x of second strand buffer (5x; Invitrogen), 0.2 mM of dNTPs (10 mM; Invitrogen), 40 U of *E. coli* DNA polymerase I (10 U/µl; NEB, Ipswich, MA, USA), 10 U of *E. coli* DNA ligase (10 U/µl; NEB), 5 U of RNase H (5 U/µl; Invitrogen) to the first strand reaction (150 µl total volume). After 2 h at 16°C, the reaction was stopped by adding 10 µl of 0.5 M EDTA and purified using DNA Clean & Concentrator™-5 columns (ZymoResearch) according to the manufacturer's instructions.

For the Ovation® Universal RNA-Seq system, 0.5-1 mg of double-stranded DNA was used for library preparation in all samples (amplified from total RNA input of 100 ng). the cDNA was sheared (fragment size 100–500 bp) by sonication the Covaris S1 adaptive focused acoustics instrument (Covaris, Woburn, MA, USA) with duty cycle 5, intensity 3 and cycle/burst 200 for 180 s according to manufacturer's instructions. The sheared products were purified and concentrated with Agencourt AMPure XP beads (2x the reaction volume). The cDNA fragments were further treated following NuGEN® Ovation Prokaryotic RNA-seq System kit's instruction

for adaptor ligation which uses specific-designed primers selectively to avoid rRNA amplification. The cDNA then blunt-ended through an end-repair reaction and ligated to platform-specific double-stranded bar-coded adapters using library preparation kits from New England Biolabs (Ipswich, MA, USA). The end-repair, dA-tailing (for Illumina-based libraries), ligation of platform-specific adaptors and purification reactions required for library preparation and library amplification (10-30 cycles) steps were performed manually.

### 3.2.7 Illumina Sequencing libraries

The Leeds Institute of Molecular Medicine, Leeds, UK performed the RNA sequencing on an Illumina HiSeq platform. The final concentration of each bar-coded cDNA library (15 libraries) were standardized using elution buffer (Qiagen) and pooled in equimolar amounts of 12 nM to generate paired-end 100 bp reads in the case of NuGEN's RNA-seq system. These were then diluted to a final concentration of 10 pM, and the remainder of clustering process was conducted, and the library pool was run in a single lane of Genome Analyzer IIx (Illumina, Inc) for 100 cycles of each pair-end read before were demultiplexed. One base-pair mismatch per library was allowed, and reads were converted to a FASTq file.

### 3.2.8 Mapping

Reads were initially mapped to ribosomal RNA sequences using Bowtie (133) with default settings. Reads that mapped to ribosomal sequences were excluded from further analysis. In the case of paired-end Illumina reads, both pairs were removed if either pair mapped to rRNA. Ribosomal RNA sequences were acquired from GenBank (134). Remaining reads were mapped to the genome using TopHat v.1.1.3 (135) against the US Department of Energy (DOE) Joint Genome Institute (JGI) database (136) for the genome of *Cellulomonas fimi* NRS 133, ATCC 484.

### 3.2.9 Transcript abundance and bioinformatics analysis

Transcript abundance was determined from the TopHat alignment using a custom perl script and annotated transcripts from RefSeq. The annotated CDS to genes were then searched for carbohydrate active enzymes on CAZy database ([www.cazy.org](http://www.cazy.org)) and Blastp search for non-redundant (nr) protein sequence using Blastp on National Center for Biotechnology Information, NCBI database (<https://www.ncbi.nlm.nih.gov/>).

### 3.2.10 Differential expression analysis

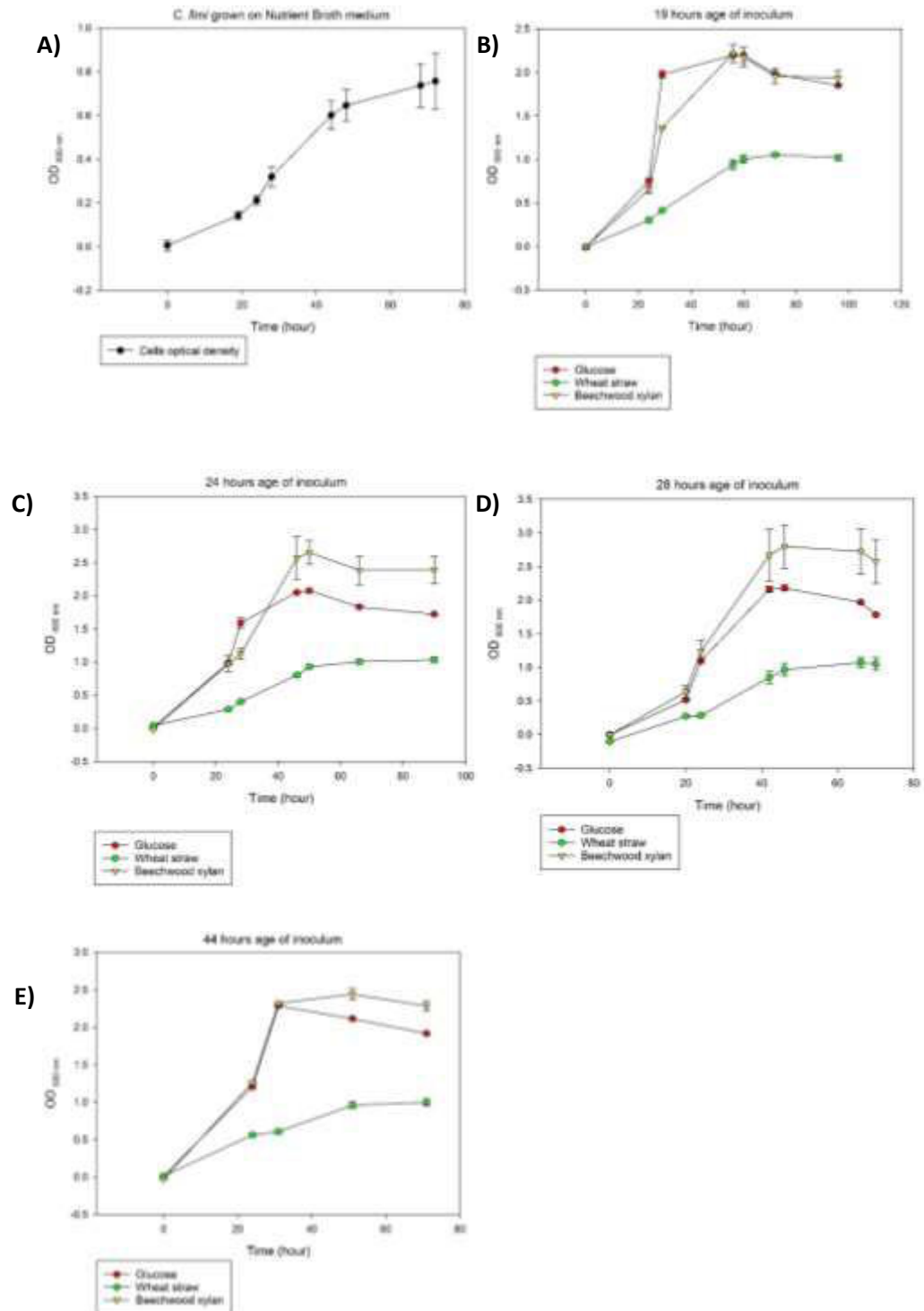
Differentially expressed transcripts were analyzed from Illumina sequenced data. Differential expression was assessed using transcript abundances as inputs to Blast2GO (137). The transcripts with an adjusted  $P > 0.05$  were considered to be differentially expressed. Transcripts called differentially expressed by pairwise differential expression analysis based on the software package “edgeR” which belongs to the Bioconductor project (138) featured in Blast2GO were separated into two groups: those upregulated and those downregulated in each sample condition. The analysis was performed on the Count per Million (CPM) of filtered reads, and were normalised using the package command. Differential gene expressions were then predicted between each set of conditions using an exact negative binomial test. Densities were fit to each group using R and were plotted against the density of all annotated RefSeq transcripts.

### **3.3 RESULTS**

#### **3.3.1 Growth study of *C. fimi* utilizing different substrates**

To identify the best culture seeding condition, a pre-cultured *C. fimi* was first grown in rich nutrient broth for 19, 24, 28, and 44 hours. Samples taken at these time points represented different stages of growth and were used to inoculate cultures growing on glucose, Avicel, beechwood xylan, wheat straw and sugarcane bagasse in fresh minimal medium for a maximum of 100 hours depending on the substrates. This was done for a first assessment and comparison of the effectiveness of differently aged inocula to initiate growth on those substrates.

Figure 3.2 shows that *C. fimi* could utilize glucose, beechwood xylan and milled wheat straw as a primary carbon source for growth. Rapid growth of *C. fimi* was observed on glucose and beechwood xylan with glucose as a better substrate for higher bacterial biomass. *C. fimi* was able to use wheat straw as a substrate but did not use it as efficiently as other substrates. However, there was little or no impact of using inoculum from different phases of growth and thus, the subsequent experiment was initiated using a 24-hour starter culture.

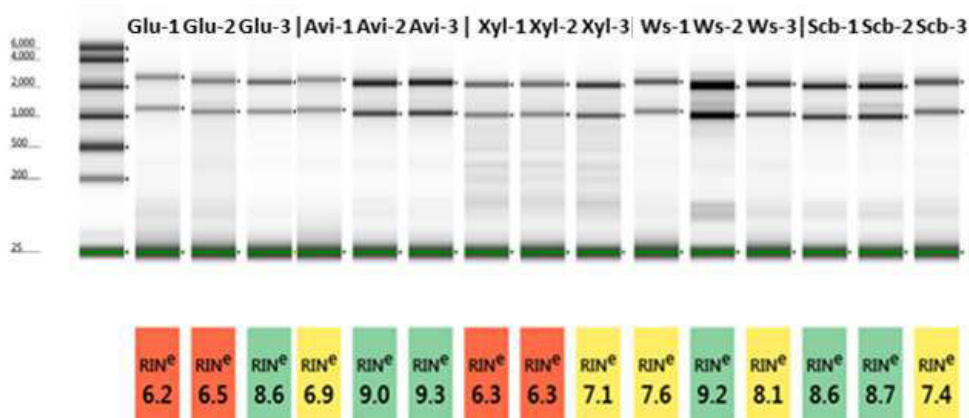


**Figure 3.2: Growth profiles of *C. fimi* grown on different carbon sources initiated using differently aged inocula of seeding cells.**

Figure 3.2(A) shows *C. fimi* growth profile when grown on nutrient broth. In subsequent experiments, *C. fimi* was grown on either 0.2% (w/v) of glucose, beechwood xylan and wheat straw using B) 19-hour, C) 24-hour, D) 28-hour, and E) 44-hour inocula.

### 3.3.2 Total RNA extraction from *C. fimi*

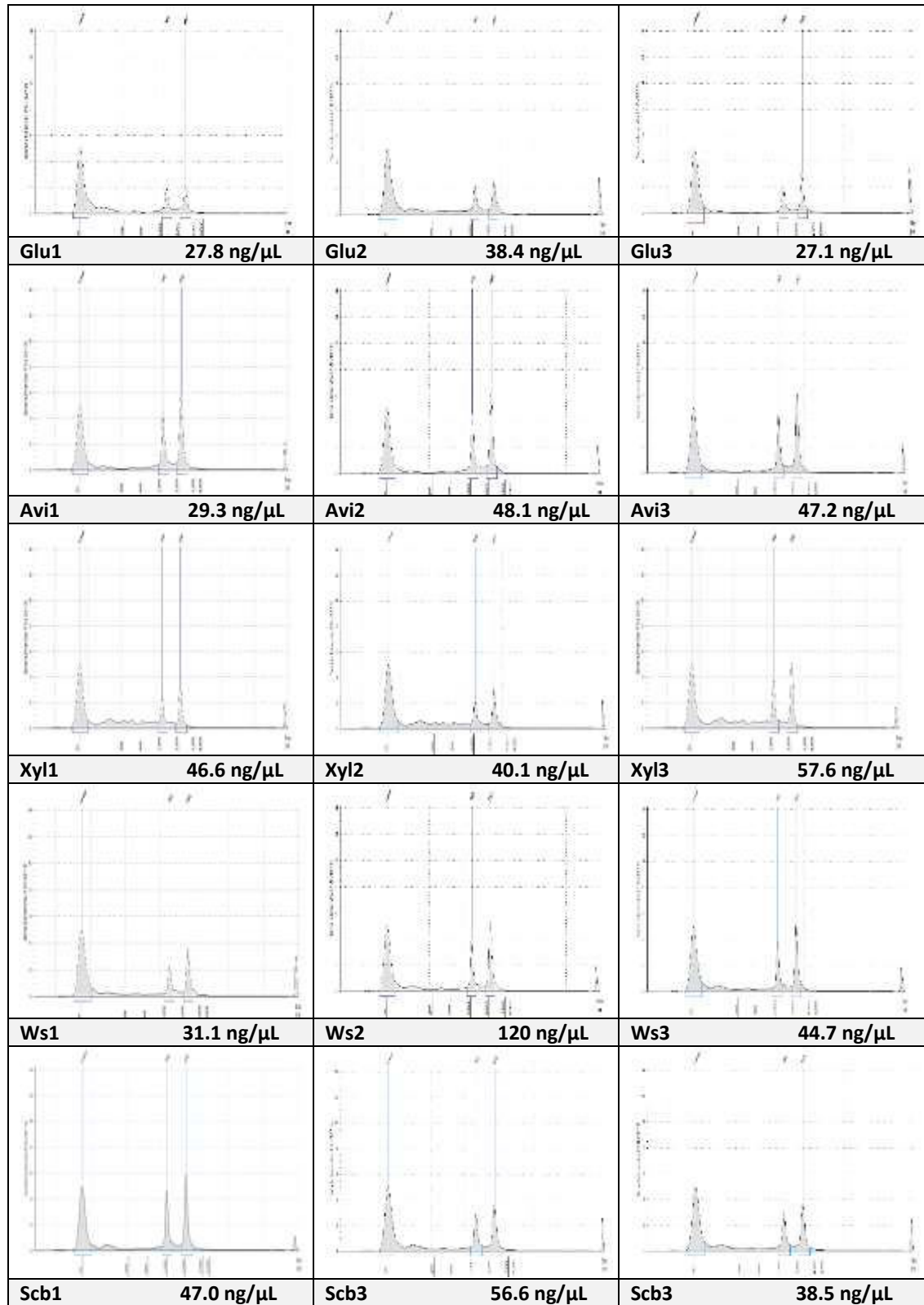
To assess changes in gene expression during growth on the various substrates, total RNA samples were isolated and purified for RNA-Seq analysis of the *C. fimi* transcriptome. The Day-3 RNA from *C. fimi* culture grown on glucose, Avicel, beechwood xylan, wheat straw, and sugarcane bagasse were chosen for RNA-Seq analysis based on the optimum growth indicated in the preceding experiment. RNA was successfully extracted from *C. fimi* growing on all five substrates. The quality of starting materials of total RNA extracted from *C. fimi* was analysed by Agilent 2200 TapeStation® using standard RNA ScreenTape®. Figure 3.3 represents the electrophoretogram from the TapeStation®, which indicated a high quality of cleaned and concentrated total RNA. The two bands between 1000 to 3000 nucleotides represented the 16S and 23S prokaryotic rRNA. The intense bands on the electrophoresis gel and sharp peaks of rRNAs on the TapeStation® electropherogram (Figure 3.4) indicate a good quality and quantity of total RNA. The RIN<sup>e</sup> number represents the ratio between both RNA peaks as classified in range 0 to 10, where 10 is the best quality. Total RNA from Avicel and sugarcane bagasse cultures obtained the best RIN<sup>e</sup> value. Furthermore, RNA from the insoluble wheat straw culture had a higher quality compared to both soluble substrates cultures of glucose and beechwood xylan that showed the lowest RIN<sup>e</sup> values of total RNA extracted. However, the low RIN<sup>e</sup> number in some of the samples may result from overloading the RNA samples and polysaccharides carried over from the culture media.



**Figure 3.3: Electrophoretic analysis of RNA from *C. fimi* using TapeStation® with standard RNA ScreenTape®**

The quality of total RNA isolated from Day-3 *C. fimi* culture grown on five substrates; glucose (Glu), wheat straw (Ws), sugarcane bagasse (Scb), Avicel (Avi) and beechwood xylan (Xyl) in 3 biological replicates analysed on the Standard RNA ScreenTape®. The first lane is a default standard RNA ladder.





**Figure 3.4: TapeStation® Electropherogram of total RNA harvested from five substrates of Day-3 *C. fimi* grown cultures.**

A) Glucose, B) Avicel, C) Beechwood xylan, D) Wheat straw, and E) Sugarcane bagasse.

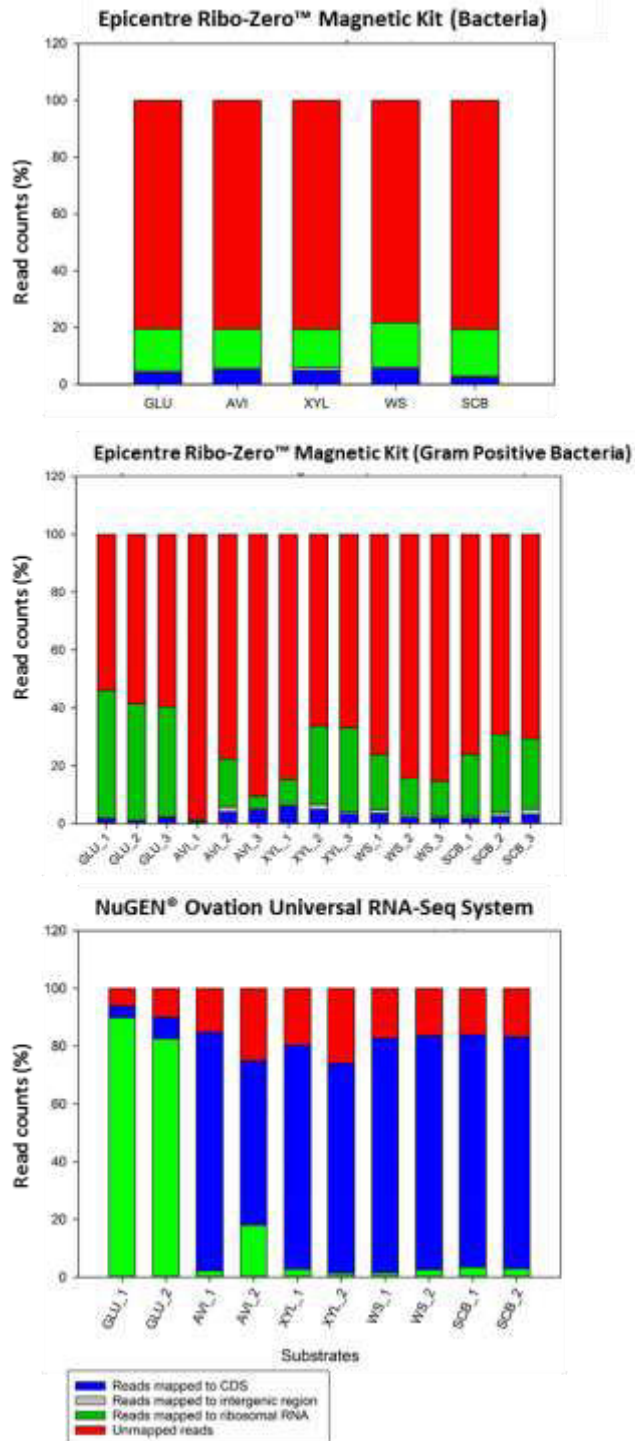
### 3.3.3 mRNA enrichment from total RNA

#### 3.3.3.1 mRNA enrichment by rRNA depletion

Removing rRNA from the RNA samples is a crucial part of the RNA-Seq workflow. Three procedures have been tested for this work. The first two experiments were using Room Temperature Stable (RTS) DNase-I followed by Ribo-Zero™ Magnetic Kit (Bacteria) and Ribo-Zero™ Magnetic Kit (Gram-Positive Bacteria), both from Epicentre®, respectively. The first two trials were unsatisfactory and showed the RNA samples were degraded and still contained high amount of ribosomal RNA due to unspecific rRNA probes. This resulted in low quality of RNA-Seq data from both experiments. Figure 3.5(A) shows for the first experiment that most of the reads from the pooled three biological replicates mapped to the rRNA. For the second experiment (Figure 3.5B), about 10 to 40% of the reads still contained high amount of rRNA. Unfortunately, as much as 45 - 98% of the reads could not be mapped to the *C. fimi* genome due to the very low quality of degraded mRNA samples.

#### 3.3.3.2 RNA-Seq work and analysis using NuGEN® Ovation Universal RNA-Seq System

The third attempt at mRNA purification used AnyDeplete (formerly known as InDA-C) probes of NuGEN® Ovation Universal RNA-Seq System. The kit provided gene-specific primers to remove the rRNA and allowed complete downstream RNA processing of cDNA libraries samples prior to the sequencing work. The obtained data was normalized into Pileup Per Kilobase per Millions of reads (PPKM) value to define the number of sequence tags covering each base of the genome. This dataset was filtered to only accept the genes which were present in two replicates at an expression rate higher than 10 ppkm. Figure 3.5C indicates that the removal of rRNA was successful in most of the samples with the percentage of reads mapped to CDSs were in range of 57.0% - 82.6% except for RNA from glucose-grown culture which only about 4.01 - 7.47% of reads were mapped to CDS compared to the total reads. However, the sequencing depth and number of reads obtained from the sequencing is appropriate to represent the gene expression in glucose cultures for which more than one million reads were yielded from both replicates (1.2 million reads for Glu1 and 1.8 million reads for Glu2 samples, respectively).



**Figure 3.5: Histogram indicating quality of RNA-Seq results with percentages of the reads mapped to the genome including the rRNA genes, coding sequences genes and the unmapped reads.**

rRNA depletion was performed using A) Epicentre® Ribo-Zero™ Magnetic Kit (Bacteria), B) Epicentre® Ribo-Zero™ Magnetic Kit (Gram Positive), and C) AnyDeplete™ gene specific primers from NuGEN® Ovation Universal RNA-Seq System. Abbreviations: Glu; Glucose, Avi; Avicel, Xyl; Beechwood xylan, WS; wheat straw, and SCB; sugarcane bagasse. Number after the substrate indicated the replication number.

### 3.3.4 The transcriptome of *C. fimi* growing on five substrates

The *C. fimi* genome is 4.27 Mb in size with 3845 predicted genes (86). Transcriptomes were sequenced from replicated independent cultures under 5 sets of conditions after growth for 3 days in the presence of glucose, Avicel, beechwood xylan, wheat straw and sugarcane bagasse as sole carbon sources in minimal media. The induction of genes involved in polysaccharide deconstruction were monitored after 3 days of bacterial growth on a specific substrate. Statistical tests were applied to identify all genes which were significantly differentially expressed (p-value <0.05 for all tests) between five conditions studied. RPKM values were calculated for each of the three biological replicates. The results shown in this study are from the combined mapping scores, reads filtered and only inductions showing a significant score in statistical test are discussed.

#### 3.3.4.1 Expression of CAZy genes in *C. fimi* grown on polysaccharides and comparison with *C. fimi* grown on glucose

Four classes of enzymes that mediate the degradation of plant cell wall carbohydrates are: the carbohydrate esterases (CEs), the polysaccharide lyases (PLs), the glycoside hydrolases (GHs), and the auxiliary activities (AAs). These enzymes are classified based on their primary amino acid sequence and related activity in the Carbohydrate Active Enzyme (CAZy) database ([www.cazy.org](http://www.cazy.org)). Analysis of *C. fimi* genome identified 10 CE-encoding genes representing 7 families, 6 PL-encoding genes representing 3 families, 112 GH-encoding genes representing 41 families, and 1 AA-encoding gene, representing 1 family. The family of AA encoded by the genome of *C. fimi* were formerly known as CBM33 which showed cleavage of crystalline chitin in an O<sub>2</sub>-dependent reaction. The AA family enzyme acts synergistically with hydrolytic cellobiohydrolases and endoglucanases and plays an important accessory role in enhancing lignocellulose degradation (139). RNA sequencing was performed on the Illumina HiSeq platform with 100bp of pair-end sequencing. The resultant reads were then filtered for rRNA contamination and trimmed. Percentage of CAZy-encoding genes were calculated for each treatment. The lowest and the highest CAZy-encoding genes expression were observed in glucose and Avicel cultures, respectively. After 3-days of *C. fimi* growth in each substrate, total percentage of CAZy families were identified (excluding accessory proteins (CBMs) and glycosyl transferases (GTs)). The average percentage of CAZy-encoding genes from the total reads in glucose, Avicel, Xylan, wheat straw and sugarcane bagasse cultures were 1.1%, 4.7%, 3.3%,

3.8%, 3.4%, respectively. Table 3.1 shows the experimental design of the RNA-seq experiments conducted with *C. fimi*, total reads of sequencing, total CAZy-encoding genes annotated and percentage of CAZy-encoding genes in each reads library.

**Table 3.1: Experimental design and number of reads generated by RNA sequencing of fifteen *C. fimi* RNA-seq samples with triplicates for each condition.**

Condition (Control/treatment)	Replicate	Reads after filtering	CAZy-encoding genes	Percentage of CAZy-encoding genes	Percentage average of CAZy-encoding genes
Glucose (control)	1	1,402,015.2	7486.1	0.5	1.1%
Glucose (control)	2	1,378,212.3	5879.9	0.5	
Glucose (control)	3	1,012,883.1	27676.6	2.7	
Avicel	1	962,559.6	32215.0	3.4	4.7%
Avicel	2	962,532.1	35062.7	3.6	
Avicel	3	1,318,274.0	85589.7	6.5	
Xylan	1	957,821.4	31720.8	3.3	3.3%
Xylan	2	957,355.4	31368.4	3.3	
Xylan	3	958,524.6	31358.3	3.3	
Sugarcane bagasse	1	961,268.5	33455.7	3.5	3.4%
Sugarcane bagasse	2	955,487.1	32592.5	3.4	
Sugarcane bagasse	3	959,955.0	32987.7	3.436	
Wheat straw	1	957,893.5	31370.9	3.275	3.8%
Wheat straw	2	1,006,083.5	46762.2	4.648	
Wheat straw	3	960,710.7	31956.2	3.326	

### 3.3.4.2 Expression of Carbohydrate Active Enzyme (CAZy)-encoding genes

Analysis of count data arising from RNA-seq work was performed using Blast2Go™, a standalone computer software designed to explore differential gene expression (DGE) based on the edgeR program (138). This analysis allows identification of differentially expressed genomic features (e.g. genes) in a pairwise comparison of two different experimental conditions. The software package edgeR (empirical analysis of DGE in R), implements quantitative statistical methods to evaluate significance of individual genes between two experimental conditions (treatment

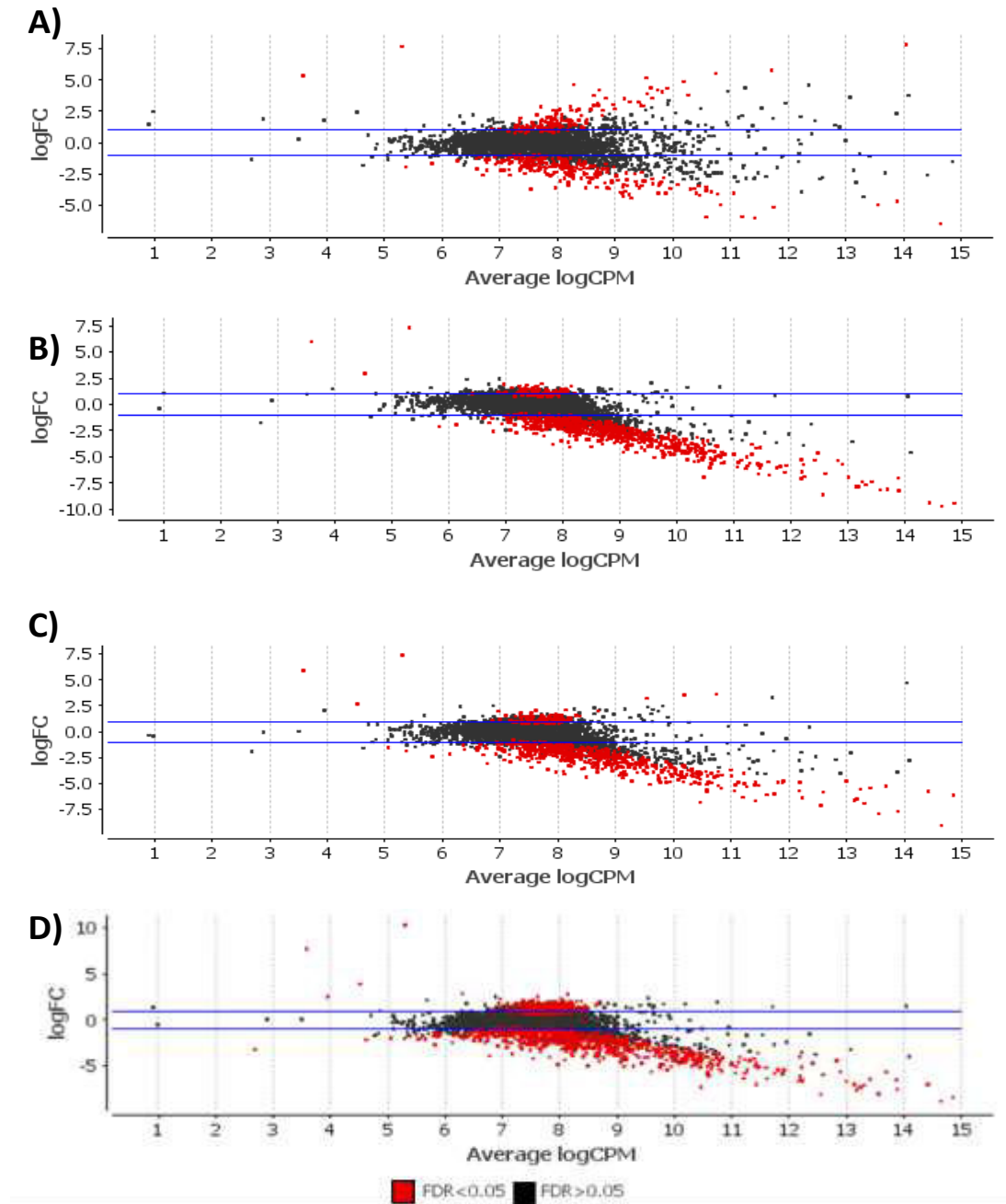
against control). From the DGE analysis, 134, 85, 113, and 292 of *C. fimi* genes have been upregulated in Avicel, xylan, wheat straw and sugarcane bagasse culture of 3-days as compared to growth on glucose. The patterns of differential expression in terms of transcript abundance from each treatment were evaluated using pairwise comparison analysis. The results of which are presented in smear plots in Figure 3.6. Among the upregulated genes, the CAZy-encoding genes representing different GH, CE, AA, and PL families that were transcriptionally upregulated by log-fold change (logFC) with  $\leq 0.05$  false discovery rate between condition of substrates are summarized in Table 3.2.

**Table 3.2: Number of upregulated CAZy-encoding genes for each comparison condition.**

CAZy-encoding genes were counted as present if their expression exceeded 1 RPKM in at least two of three biological replicates.

Condition	Upregulated CAZy- and accessory protein encoding genes						Grand total
	GH families	CE families	AA family	PL families	GT families	CBM families	
Avicel	13	3	1	1	0	2	20
Beechwood Xylan	4	0	0	0	0	1	5
Wheat straw	10	0	0	1	0	1	12
Sugarcane bagasse	16	0	0	1	2	3	22

The highest grand total of CAZy-encoding genes upregulated were in sugarcane bagasse culture (22 genes), and the lowest were in xylan culture (5 genes) compared to the other two substrates. The majority of upregulated CAZy-encoding genes were from GH families after 3-days incubation in all cultures. The highest number of carbohydrate esterases (CE1, CE2, CE3, and CE4) and the presence of one *C. fimi* auxiliary activity (AA10) were being identified to be significantly upregulated exclusively in Avicel-grown culture. The list of CAZy-encoding genes that had been upregulated in each substrate are presented in Table 3.4.



**Figure 3.6: Differentially expressed transcripts between glucose and carbohydrate treatment after 3-days growth of *C. fimi*.**

Visualization by MA plots of transcript expression profiles show A) Avicel versus glucose. B) Xylan versus glucose. C) Wheat straw versus glucose. D) Sugarcane bagasse versus glucose. Log of the fold changes (logFC) on the y-axis versus the average of the log of the CPM on the x-axis. The plot visualizes the differences between measurements taken for two samples (substrate treatment against negative control; glucose), by transforming the data onto M (log ratio) and A (mean average) scales generated by Blast2Go™ software based on EdgeR program. Transcripts that are identified as significantly differentially expressed at most 0.5% FDR, colored in red.

**Table 3.3: The CAZy-encoding genes upregulated on each carbohydrate substrate using Differential Gene Expression (DGE) analysis.**

Pairwise comparison method was applied to individually filtered and normalized libraries against the negative control (glucose culture) by Log<sub>2</sub>FC. The false discovery rate (FDR) was equal or less than 0.05. List of all upregulated CAZy-encoding genes in each library for A) Avicel, B) Beechwood xylan, C) wheat straw, and D) sugarcane bagasse are presented from the highest to the lowest order of fold-change.

**A) Upregulated CAZy-encoding genes on Avicel**

No.	Protein name	Protein domain	Log <sub>2</sub> FC	FDR
1	Celf_0270	AA10, CBM2	7.927	0.006
2	1,4-beta-cellobiosidase A	CBM2, GH6	5.853	0.009
3	Celf_0438	GH27, CBM13	4.485	0.008
4	Celf_1754	CE2, CBM2	3.868	0.003
5	Celf_3775	PL1	3.342	0.011
6	Celf_1913	GH74, CBM2	3.241	0.002
7	Celf_0374	GH11, CBM2, CE4, CBM2	3.054	0.027
8	Celf_0045	GH9, CBM2	2.763	0.008
9	Celf_0404	CE1, CBM2	2.663	0.005
10	Celf_1329	CE3, CBM2	2.661	0.008
11	Celf_0161	CBM50	2.156	0.003
12	Celf_0862	GH26, CBM23	1.930	0.003
13	Celf_2091	GH26	1.923	0.017
14	Celf_1186	GH13	1.889	0.044
15	Celf_1126	GH13, CBM48, GH13	1.809	0.004
16	Celf_0376	GH5, CBM46	1.549	0.007
17	Celf_3249	GH51	1.477	0.017
18	Celf_0403	CBM2	1.366	0.039
19	Celf_2053	GH3	1.319	0.027
20	Celf_2718	GH36	1.131	0.032

**B) Upregulated CAZy-encoding genes on xylan**

No.	Protein name	Protein domain	Log <sub>2</sub> FC	FDR
1	Celf_0161	CBM50	1.454	0.045
2	Celf_2053	GH3	1.424	0.008
3	Celf_0376	GH5, CBM46	1.196	0.031
4	Celf_2232	GH13	1.156	0.044
5	Celf_2718	GH36	1.119	0.019



**C) Upregulated CAZy-encoding genes on wheat straw**

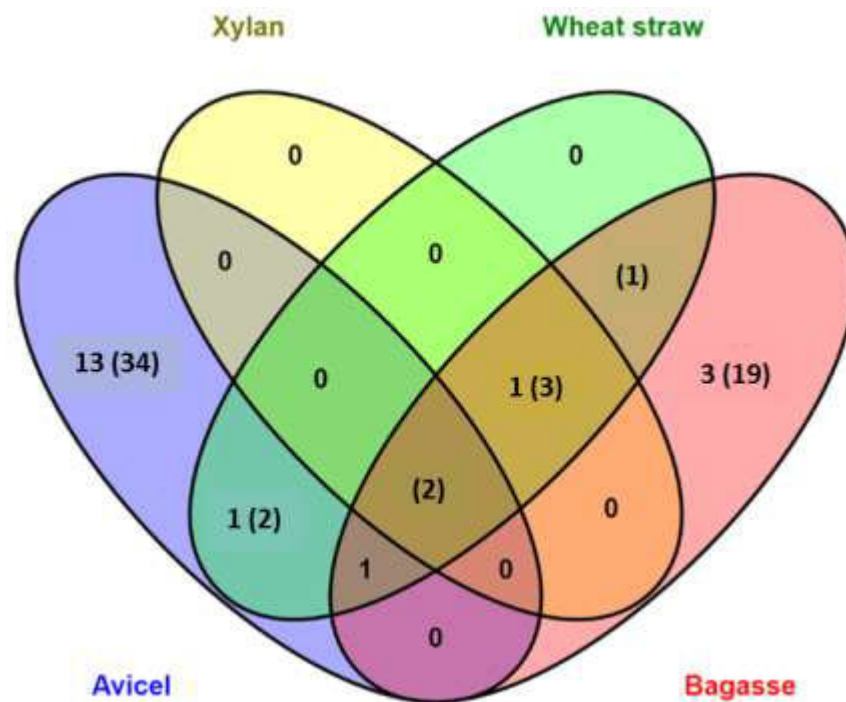
No.	Protein name	Protein domain	Log <sub>2</sub> FC	FDR
1	Celf_1913	GH74, CBM2	2.102	0.033
2	Celf_0161	CBM50	1.698	0.020
3	Celf_1729	GH10	1.589	0.031
4	Celf_2053	GH3	1.405	0.011
5	Celf_0862	GH26, CBM23	1.401	0.032
6	Celf_2232	GH13	1.316	0.022
7	Celf_0376	GH5, CBM46	1.309	0.019
8	Celf_1126	GH13, CBM48, GH13	1.239	0.048
9	Celf_3440	PL11	1.221	0.046
10	Celf_2714	GH13	1.123	0.032
11	Celf_2624	GH16	1.107	0.020
12	Celf_2718	GH36	1.064	0.029

**D) Upregulated CAZy-encoding genes on sugarcane bagasse**

No.	Protein name	Protein domain	Log <sub>2</sub> FC	FDR
1	Celf_1729	GH10	2.148	0.002
2	Celf_0067	GT2	2.122	0.010
3	Celf_1913	GH74, CBM2	2.097	0.024
4	Celf_1705	CBM4, CBM4, GH9	2.075	0.016
5	Celf_0161	CBM50	1.696	0.014
6	Celf_1482	GH43	1.683	0.048
7	Celf_2232	GH13	1.671	0.002
8	Celf_2053	GH3	1.587	0.003
9	Celf_0862	GH26, CBM23	1.546	0.013
10	Celf_3113	GH16, CBM13	1.540	0.005
11	Celf_2726	GH3	1.514	0.041
12	Celf_3440	PL11	1.453	0.012
13	Celf_0376	GH5, CBM46	1.408	0.008
14	Celf_0345	CBM2, PL11	1.383	0.044
15	Celf_2624	GH16	1.333	0.003
16	Celf_2714	GH13	1.314	0.008
17	Celf_3249	GH51	1.277	0.025
18	Celf_0398	CBM13, PL3	1.241	0.016
19	Celf_1126	GH13, CBM48, GH13	1.220	0.039
20	Celf_2784	GH31	1.200	0.003
21	Celf_2718	GH36	1.097	0.018
22	Celf_1719	GT4	1.035	0.012

### 3.3.4.3 Co-expression of upregulated CAZy- and non-CAZy encoding genes using differential analysis

The co-expression of upregulated CAZy and non-CAZy-encoding genes were further investigated from the previous analysis of differential expression analysis. A Venn diagram was generated to explore the genes that were being co-expressed with CAZymes in particular substrate supplied during 3-days *C. fimi* growth. The Venn diagram in Figure 3.7 is depicting the overlapping expression of a total of 80 significantly upregulated CAZy- and non-CAZy encoding genes across the four substrates. The diagram revealed that 47 genes were exclusively upregulated on Avicel, 22 genes were exclusively upregulated in sugarcane bagasse, 3 genes were commonly upregulated on Avicel and wheat straw, 1 gene was found to be expressed on Avicel, wheat straw and sugarcane bagasse, one gene was found to be expressed commonly in wheat straw and sugarcane bagasse, and 2 genes were commonly being upregulated in all conditions. The list of all co-expressed genes is presented in Table 3.5.



**Figure 3.7: Venn diagram of the CAZy- and non-CAZy encoding genes that were significantly upregulated in each condition.**

Venn diagram was generated using Venny™ online software (140). The list of upregulated genes from RNA-seq dataset were taken from differential expression analysis. The numbers of non-CAZy genes were indicated in the bracket.

**Table 3.4: Identification of co-expressed genes which upregulated with CAZy-encoding genes from differential analysis of RNA-seq data.****47 genes included exclusively in Avicel:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	Log <sub>2</sub> FC in Avicel
YP_004451802.1	Celf_0270-AA10,CBM2	CAZy	7.9
YP_004454685.1	CBM2,GH6-CBM2,GH6	CAZy	5.9
YP_004451699.1	hypothetical protein	Non-CAZy	4.7
YP_004451968.1	Celf_0438-GH27,CBM13	CAZy	4.5
YP_004452285.1	hypothetical protein	Non-CAZy	4.4
YP_004453792.1	Glucose/sorbose dehydrogenases	Non-CAZy	4.3
YP_004451712.1	hypothetical protein	Non-CAZy	4.3
YP_004454900.1	Cellobiohydrolase A (1,4-beta-cellobiosidase A)	CAZy	4.2
YP_004452494.1	hypothetical protein	Non-CAZy	4.0
YP_004453614.1	hypothetical protein	Non-CAZy	3.9
YP_004453273.1	Celf_1754-CE2,CBM2	CAZy	3.9
YP_004451816.1	NAD-dependent aldehyde dehydrogenases	Non-CAZy	3.6
YP_004451558.1	Cellobiohydrolase A (1,4-beta-cellobiosidase A)	CAZy	3.6
YP_004451719.1	hypothetical protein	Non-CAZy	3.6
YP_004454240.1	ABC-type sugar transport system, periplasmic component	Non-CAZy	3.3
YP_004455268.1	Celf_3775-PL1	Non-CAZy	3.3
YP_004451697.1	T4-like virus tail tube protein gp19.	Non-CAZy	3.3
YP_004452793.1	Beta-1,4-xylanase	CAZy	3.1
YP_004453611.1	Cation transport ATPase	Non-CAZy	3.1
YP_004453744.1	Response regulator with putative antiterminator	Non-CAZy	3.1
YP_004454061.1	Alkaline phosphatase	Non-CAZy	3.1
YP_004451906.1	Celf_0374-GH11,CBM2,CE4,CBM2	CAZy	3.1
YP_004453402.1	hypothetical protein	Non-CAZy	3.0
YP_004451716.1	hypothetical protein	Non-CAZy	2.9
YP_004451718.1	Response regulator containing a CheY-like receiver domain	Non-CAZy	2.9
YP_004453242.1	Membrane protease subunits	Non-CAZy	2.8

YP_004451583.1	Celf_0045-GH9,CBM2	CAZy	2.8
YP_004451910.1	Predicted glutamine amidotransferase	Non-CAZy	2.7
YP_004451700.1	hypothetical protein	Non-CAZy	2.7
YP_004453750.1	hypothetical protein	Non-CAZy	2.7
YP_004451934.1	Celf_0404-CE1,CBM2	CAZy	2.7
YP_004452296.1	hypothetical protein	Non-CAZy	2.7
YP_004452851.1	Celf_1329-CE3,CBM2	CAZy	2.7
YP_004453441.1	Endoglucanase	CAZy	2.6
YP_004451553.1	hypothetical protein	Non-CAZy	2.4
YP_004451551.1	hypothetical protein	Non-CAZy	2.3
YP_004451717.1	Periplasmic component of the biopolymer transport system	Non-CAZy	2.3
YP_004454228.1	ABC-type sugar transport system, periplasmic component	Non-CAZy	2.3
YP_004452262.1	hypothetical protein	Non-CAZy	2.3
YP_004451817.1	Uncharacterized protein conserved in bacteria	Non-CAZy	2.3
YP_004454543.1	FOG: PKD repeat (PKDP3, Celf_3039)	Non-CAZy	2.2
YP_004451696.1	Celf_0161-CBM50	CAZy	2.2
YP_004454192.1	hypothetical protein	Non-CAZy	2.1
YP_004453261.1	Putative peptidoglycan binding domain.	Non-CAZy	2.0
YP_004454064.1	Citrate synthase	Non-CAZy	2.0
YP_004453417.1	hypothetical protein	Non-CAZy	2.0
YP_004451695.1	Uncharacterized protein conserved in bacteria	Non-CAZy	2.0

**22 genes included exclusively in Bagasse:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	LogFC in bagasse
YP_004454623.1	Ribosomal protein L33	Non-CAZy	3.0
2506047258	hypothetical protein	Non-CAZy	2.7
YP_004454257.1	ABC-type sugar transport system, periplasmic component	Non-CAZy	2.4
YP_004452237.1	Response regulator containing a CheY-like receiver domain	Non-CAZy	2.2
YP_004454008.1	Predicted membrane protein	Non-CAZy	2.2
YP_004452591.1	Uncharacterized conserved protein	Non-CAZy	2.2
YP_004453248.1	Celf_1729-GH10	CAZy	2.1
YP_004452983.1	ABC-type nitrate/sulfonate/bicarbonate transport systems	Non-CAZy	2.1
YP_004454238.1	ABC-type sugar transport system, permease component	Non-CAZy	2.1
YP_004453278.1	PAS fold.	Non-CAZy	2.1
YP_004452977.1	Predicted amidohydrolase	Non-CAZy	2.1
YP_004454809.1	Predicted sugar isomerase	Non-CAZy	2.1
YP_004451605.1	Celf_0067-GT2	CAZy	2.1
YP_004454258.1	Ribose/xylose/arabinose/galactoside ABC-type transport systems	Non-CAZy	2.1
YP_004454307.1	Predicted nuclease of the RecB family	Non-CAZy	2.1
YP_004453244.1	ABC-type multidrug transport system, ATPase and permease components	Non-CAZy	2.1
YP_004453224.1	Celf_1705-CBM4,CBM4,GH9	CAZy	2.1
YP_004452213.1	Flagellar basal body-associated protein	Non-CAZy	2.1
YP_004452233.1	Methyl-accepting chemotaxis protein	Non-CAZy	2.1
YP_004454625.1	Ribosomal protein L31	Non-CAZy	2.0
YP_004454601.1	Nitrate reductase gamma subunit	Non-CAZy	2.0
YP_004453234.1	Enoyl-[acyl-carrier-protein] reductase (NADH)	Non-CAZy	2.0

**3 common genes in Avicel and Wheat straw:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	LogFC in Avi	LogFC in WS
YP_004453442.1	Cellobiohydrolase A (1,4-beta-cellobiosidase A)	CAZy	5.6	3.7
YP_004451702.1	Phage tail sheath protein FI	Non-CAZy	5.0	3.6
YP_004451701.1	T4-like virus tail tube protein gp19.	Non-CAZy	5.2	3.3

**1 common genes in Wheat straw and Bagasse:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	LogFC in WS	LogFC in SCB
YP_004454255.1	Sugar phosphate isomerases/epimerases	Non-CAZy	2.2	3.0

**4 common genes in Xylan, Wheat straw and Bagasse:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	LogFC in Xyl	LogFC in WS	LogFC in SCB
2506047404	hypothetical protein	Non-CAZy	3.1	2.8	4.0
YP_004452239.1	Response regulator	Non-CAZy	2.1	2.1	2.6
2506047552	Glycosyl hydrolases family 39.	CAZy	2.0	2.1	2.4
YP_004453519.1	hypothetical protein	Non-CAZy	2.0	2.0	2.2

**1 common gene in Avicel, Wheat straw and Sugarcane bagasse:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	LogFC in Avi	LogFC in WS	LogFC in SCB
YP_004453430.1	Celf_1913-GH74,CBM2	CAZy	3.2	2.1	2.1

**2 common genes in Avicel, Xylan, Wheat straw and Bagasse:**

Accession number	Protein name/Gene ID	CAZy/non-CAZy	LogFC in Avi	LogFC in Xyl	LogFC in WS	LogFC in SCB
2506047359	hypothetical protein	Non-CAZy	7.8	7.4	7.5	10.4
2506045288	Mitochondrial domain of unknown function (DUF1713).	Non-CAZy	5.5	6.1	6.0	7.8

From the RNA-seq differential gene analysis, 47 upregulated genes were identified being expressed exclusively in Avicel-induced culture with Log<sub>2</sub>FC cut-off value. From that total gene number, an AA10-encoding gene of *C. fimi* (Celf\_0270) which also contains a CBM2 domain was significantly the most upregulated CAZy-encoding gene with 7.9 fold-change compared to the glucose-induced culture. From the same condition, 16 hypothetical proteins were found to be co-expressed with CAZymes including another two uncharacterized proteins conserved in bacteria. A glucose/sorbone dehydrogenase (YP\_004453792.1) gene was also been found to be upregulated exclusively in Avicel-induced culture with a 4.3 fold-change compared to the glucose-induced culture. One out of five Polycystic Kidney Disease I (PKD)-domain containing proteins (Celf\_3039) that consists of PKD domain, and a Fibronectin III (FN3) domain in its structure was significantly upregulated with 2.2 fold-change in the Avicel-induced culture. The comparison of differential gene expression for sugarcane bagasse- and glucose-induced cultures showed 22 genes were exclusively being upregulated during *C. fimi* growth on sugarcane bagasse. These include one hypothetical protein (Gene ID: 2506047258) and another uncharacterized conserved protein (YP\_004452591.1) with 2.7 and 2.2 fold-changes, respectively. One of the CAZy-encoding genes also upregulated (2.1 fold-change) exclusively in this culture encodes a putative GH9 predicted to be a multimodular protein with two CBM4 domains in its structure (Celf\_1705).

There were four common upregulated genes expressed in xylan, wheat straw and sugarcane bagasse culture including two genes encoding hypothetical proteins. Both of these protein-encoding genes were expressed significantly higher in sugarcane bagasse culture than xylan, and wheat straw. The first gene (Gene ID: 2506047404) was upregulated 4-fold, and the second one (YP\_004453519.1) was upregulated 2.2 fold. Two genes were upregulated in all four (Avicel, xylan, wheat straw and sugarcane bagasse) culture conditions. One gene was predicted to encode a hypothetical protein (Gene ID: 2506047359) with 10.4-fold change in sugarcane bagasse culture, which was also upregulated in Avicel, xylan and wheat straw cultures with  $\geq 5$  fold-change. A GH74-encoding gene (Celf\_1913) predicted to consist of a CBM2 domain was the only one CAZy-encoding gene which was upregulated on three common substrates, i.e. Avicel (3.2 fold-change), wheat straw (2.1 fold-change) and sugarcane bagasse (2.1 fold-change) cultures.

### 3.4 DISCUSSION

There are no prior publications of transcriptomic data for *C. fimi*, and the work in this chapter may, therefore, represent the first such study. The response to growth on four polysaccharides, crystalline cellulose (Avicel) and plant cell wall hemicellulose (beechwood xylan), including two lignocellulose biomasses, ball-milled wheat straw and sugarcane bagasse were examined and compared against a glucose culture as an experimental control using a RNA-seq approach. The analysis presented here focused on identifiable carbohydrate active enzyme encoding genes, and those of unknown function showing co-expression with particular CAZyme genes in this study.

The third day of *C. fimi* growth on glucose and four other substrates was chosen as the time point for investigation based on the growth profiles study described in Chapter 3.1. as *C. fimi* grows more slowly on complex substrates than on rich medium. The active growth of *C. fimi* was indicated during its exponential phase between second and fourth day of incubation, however, due to the lengthy growth of *C. fimi* in certain culture e.g. insoluble wheat straw, the third day of culture was chosen to harvest the RNA materials. In addition, the latter stage of the time courses of microbial growth especially on the complex polysaccharides was expected to contain the material most recalcitrant to degradation, and therefore may induce the expression of novel enzymes.

The mRNA typically constitutes a very small fraction of the total RNA in bacterial cells and hence, rRNA subtraction is needed to allow enrichment of gene transcripts. *C. fimi* is a gram-positive bacterium with high GC-based content (74%) in its genome, therefore, hybridization of biotinylated probes using magnetic beads were applied using Epicentre® Ribo-Zero™ rRNA removal kit (Bacteria) and Ribo-Zero™ rRNA removal kit (Gram-Positive) for the first and second trials of mRNA enrichment.

rRNA removal was eventually achieved through a capture-based method that relies on the synthesis of first and second strand cDNA using a primer mix called Gene-Specific Primers (GSPs) i.e. using the Ovation Prokaryotic RNA-seq System by NuGEN™. The GSPs approach that is integrated in the kit is designed to selectively enrich the mRNA portion of the bacterial total RNA. The selective specific primers were designed against *C. fimi* 16S and 23S rRNA sequences from database then capture with probes called AnyDeplete, formerly known as Independent Adaptor Cleavage (InDA-C) to remove



the rRNA sequences in the library preparation workflow just before the final amplification steps. This method was successfully applied to *C. fimi* RNA-seq samples while also preserving the mRNAs relative abundance for subsequent Illumina deep sequencing. The highest observed CAZy-encoding genes expression was from cultures growing on Avicel, and the lowest was in glucose culture. Low level of mRNA from the genes encoding enzymes involved in the degradation of complex carbohydrates including hemicellulose and chitin, were present when the bacterium is cultivated on glucose-based medium and the few that are expressed are likely involved in growth and metabolism.

The highest percentage of CAZy-encoding genes was expressed in Avicel culture and this is in agreement with *C. fimi*'s well-known ability to degrade cellulose. On this substrate, the significantly highest fold-change was observed for AA10, the first LPMO found in *C. fimi* that previously was classified as carbohydrate binding module 33 (CBM33, now systematically called Auxiliary Activity 10, AA10) (69,139). Vaaje-Kolstad *et al.* were first to demonstrate the oxidative activity to CBM33 (141), which showed cleavage of crystalline chitin in an O<sub>2</sub>-dependent reaction (copper dependent oxygenase) (34). The presence of oxidoreductases has been reported in various cellulolytic bacteria and fungi, though an actual, physiological electron donor for LPMOs has not been unambiguously determined. The disruption of recalcitrant polysaccharide structures by an oxidative mechanism of action by AA10s provides an answer of how the initial attack on cellulose or chitin is effected especially by saprophytes (142). The oxidative function of AA10 works synergistically with hydrolytic cellobiohydrolases, endoglucanases and this significantly accelerates the degradation of polysaccharides into oligosaccharides (34,139).

Glucose sorbione dehydrogenase (GSDH), Celf\_2278 and Polycystic Kidney Disease (PKD)-containing protein (PKDP3), Celf\_3039 are among the non-CAZy genes appear to be co-expressed and highly upregulated exclusively in Avicel culture with 4.3 and 2.2 fold changes, respectively. The GSDH has previously been annotated as PKD-containing protein (PKDP1) with the existence of a multimodular domain in its structure. The reannotation of genes in NCBI database to the non-redundant (nr) reference sequence (RefSeq), this protein has 100% and 79% identities to a  $\beta$ -glucosidase of the *C. fimi* and *C. cellasea* genomes, respectively. Since this protein has not yet been characterized, for easy referral, it will be referred as PKD-domain containing protein 1 (PKDP1) here and in subsequent

chapters. Both of these proteins (PKDP1 and PKDP3) share similarity in the PKD domains in their predicted structures, but differ in additional domain, i.e. pyrroloquinoline-quinone (PQQ), and protective antigen (PA14) domains exclusively in PKDP1, and Fibronectin type 3 domain (FN3) only in PKDP3. A detailed discussion of each domain of PKDP1 is presented in Chapter 5. Apart from predicted non-CAZy proteins that co-expressed with other CAZymes, there were 16 hypothetical proteins and another 2 'uncharacterized protein conserved in bacteria' also being upregulated with  $\geq 2$  fold-change exclusively in Avicel cultures. It is an interesting observation as these proteins potentially may be involved and contribute to efficiency in cellulose depolymerisation, and further analysis on the predicted domain, protein localization, and pathway analysis by enrichment gene analysis could be investigated.

GH9 (Celf\_1705) and GH10 (Celf\_1729) are CAZymes that the genes of which were being highly upregulated amongst 22 genes exclusively in sugarcane bagasse culture. In Celf\_1705, there were three predicted domains; two domains of CBM4 and a catalytic domain of GH9. From the CAZy database, the two binding domains may be involved in binding to xylan,  $\beta$ -1,3/ $\beta$ -1,4/ $\beta$ -1,6-glucan and amorphous cellulose but not to crystalline cellulose. Based on Delmas *et al.* study (56), ball-milled wheat straw contains approximately 37% cellulose, 32% hemicelluloses, and 22% lignin, whereas the untreated sugarcane bagasse has been estimated to consist of 35% cellulose, 25% hemicellulose, and lignin 22% by Zhu *et al.*'s study in 2016 (27). Although celluloses have a high amorphous content that is usually more easily digested by enzymes, it is unclear whether the crystallinity index (CI) actually provides a clear indication of the digestibility of a cellulose sample (143). While in this case, Celf\_1705 with double CBM4 domains may indicate that the lower percentage of hemicellulose (25%) in sugarcane bagasse may contribute to the higher expression of this enzyme as the matrix polysaccharide is easier to be penetrated compared to the cellulose fraction of the wheat straw that contains higher (32%) of hemicellulose.

The putative Celf\_1729 has only one predicted catalytic GH10 domain which has been classified as an endo-1,4- $\beta$ -xylanase activity. Both enzymes, GH9 and GH10 were reported to be important especially for opening up the structure of lignocellulose biomass matrix. There was one hypothetical protein and another 'uncharacterized conserved protein' identified to be highly upregulated only in sugarcane bagasse culture. In cultures growing on xylan, wheat straw and sugarcane bagasse

cultures, two hypothetical proteins were detected to be co-expressed with CAZymes. Finally, the only putative GH74 (Celf\_1913) gene identified in *C. fimi* which also has CBM2 in its structure was expressed in Avicel, wheat straw and sugarcane bagasse cultures, but absent in xylan-induced medium. GH74 has been predicted to be endoglucanase/xyloglucanase (144). In Celf\_1913, there is a predicted domain of bacterial neuramidase 2 (BNR2) that may work as a sialidase, but this is unlikely to be the case as there is no known source of sialic acid in in sugarcane bagasse. This protein was also detected under the same specific condition in proteomic studies, which will be elaborated in the next chapter. As a conclusion, several uncharacterized predicted CAZy-encoding genes and a total of 20 hypothetical proteins with another 3 uncharacterized conserved proteins in bacteria that were being co-expressed with exhibited CAZymes, have been identified from this transcriptomic study. These proteins may need to be characterized to validate their involvement in lignocellulose break down mechanisms as they are potentially new CAZymes or processing enzymes from the *Cellulomonas fimi* repertoire.

## 4 Proteomic Analysis of the Secretome of *Cellulomonas fimi*

### 4.1 INTRODUCTION

*C. fimi* is a Gram-positive soil bacterium known for its ability to degrade cellulose and hemicellulose of plant cell walls (113). A significant array of carbohydrate active enzymes (CAZymes) is encoded in its genome (86). Based on the information of characterized *C. fimi* proteins deposited in the universal protein knowledgebase (UniProt KB) (145) as well as the extensive database of CAZymes ([www.CAZy.org](http://www.cazy.org)) (76), the secretion of known proteins involved in the breakdown of plant cell walls polysaccharides by *C. fimi* has been identified (Table 4.1). Glycoside Hydrolase family 3 (GH3) has notably the highest number of members (11) in *C. fimi* but only one auxiliary activity family 10 (AA10) has been identified in the system. GH9 is a family of enzymes that may have an endoglucanase/cellobiohydrolase as well as endo-xyloglucanase activities. This group of enzymes was formerly known as cellulase type E. On the other hand, GH10 is a family of enzymes that have more specific activity of endo-xylanases which previously classified as cellulase type F (76). A similar number of GH9 and GH10 are encoded in the *C. fimi* genome indicating the ability of the bacterium to degrade cellulose and hemicellulose specifically from plant cell walls.

**Table 4.1: Known cellulase and xylanase families in *Cellulomonas fimi*.**

Families of CAZymes are from the CAZy database (<http://www.cazy.org/Glycoside-Hydrolases.html>). The total GH number does not include pectin lyases.

CAZy family	CBM33 (AA10)	GH1	GH3	GH5	GH6	GH9	GH10	GH11	GH16	GH48	GH74	Total GH
<i>C. fimi</i>	1	1	11	3	4	4	5	1	3	1	1	109

Although many of the plant cell wall polysaccharide degrading enzymes of *C. fimi* have been well-studied for decades (104,146), not much is known about the ability of *C. fimi* particularly as a lignocellulose degrader. The AA10 identified in *C. fimi* is predicted to be involved in redox mechanism (64) for degrading cellulose which suggests that *C. fimi* uses a combination of hydrolytic and oxidative cleavage mechanisms for efficient biomass utilization.

Proteomic analysis is one of the methods to investigate the metabolic processes involved in an organism and has been improved over the last decades from qualitative identification of specific single proteins (147) to quantitative detection of hundreds to thousands of proteins by targeted shotgun proteomic technique using mass spectrometry (MS). Since the introduction of the MS-based approach in the proteomic studies (148), it has evolved into an advanced and powerful tool for investigating multi-complex protein samples, for instance during biological processes. Prior to MS-analysis, protein samples are treated either by tagging the peptide either with or without chemical labelling. The protein labelling approach using a protein biomarker (149–151) or an isobaric tag (iTRAQ) (152–154) are among familiar techniques for targeted proteomic analysis. Despite of that, label-free quantification (LFQ) has gained popularity in recent years due to its uncomplicated procedure, high dynamic range and robustness of method. This technique has benefited from the sophisticated computational algorithms during the peptides analysis process (155).

This chapter describes the application of the LFQ method to *C. fimi* secretome samples grown on soluble and insoluble substrates, which yielded high quality data. Proteomic analysis using LFQ has been applied to *C. fimi* secretome grown in Avicel, beechwood xylan, wheat straw and sugarcane bagasse cultures. Avicel is a commercially available microcrystalline cellulose, an insoluble linear polysaccharide comprised of  $\beta$ -(1,4) linked glucose monomers. Cellulose is the primary component of wood and considered the most abundant biopolymer in nature and is comprised of crystalline and amorphous parts. Another component of plant cell walls is hemicellulose with xylan as the major compound in plant cell walls of angiosperms (hardwood) and grasses. It is a heteropolymer that primarily consists of xylose, arabinose, galactose, glucuronic acid, either in glucuronoarabinoxylan or glucuronoxylan depending on the type of plant hemicellulose (37).

In this study, two insoluble lignocellulosic substrates from agricultural residues were also used. The milled wheat (*Triticum aestivum* L.) straw and sugarcane (*Saccharum sp.*) bagasse are carbon biopolymers mainly composed of cellulose, hemicellulose and lignin. Both lignocellulosic biomasses are attractive substrates for second generation biofuel production, as they complement and augment wheat and sugar productions rather than competing with food production (156) which helps avoiding change in land use and expansion of agricultural areas to meet growing energy demands (157). However, like other sources of lignocellulosic biomass, they are heterogeneous in nature due to the range of cells types and tissues in their structures (27,158,159) which has implications for saccharification efficiency (160,161).

To investigate the secretome of *C. fimi* for degrading cell wall polysaccharide components and the potential in digesting lignocellulosic biomass, this bacterium was grown on different types of substrates (Avicel, beechwood xylan, untreated wheat straw and untreated sugarcane bagasse) as the sole carbon source to investigate the secretome of *C. fimi* for degrading cell wall polysaccharide components and the potential in digesting lignocellulosic biomass. The secretomes from shake flask cultures were compared using bioinformatics analyses and putative new CAZy that are involved in plant cell wall degradation have been identified using nanoLC-MS/MS analysis.

## 4.2 MATERIALS AND METHODS

### 4.2.1 Sample preparation

During the growth of *C. fimi* in shake flasks, samples were collected from 25 mL cultures throughout the growth cycle, with three biological replicates for each three-day time points. Growth was monitored by measuring OD<sub>600</sub>; to avoid interference from the insoluble wheat straw and sugarcane bagasse, all samples were spun down in a centrifuge (500 x g, 30 s) before being transferred into cuvettes for OD measurements. For protein extraction, sterile culture supernatants were obtained by spin down the whole culture in 50 mL centrifuge tube, at 4000 x g. The collected samples were filtered using 0.22 µm PES membrane filter. Proteins were desalted and concentrated using 10 kDa Amicon® Ultra-0.5 Centrifugal Filter Unit (UFC501024, Merck Milipore).

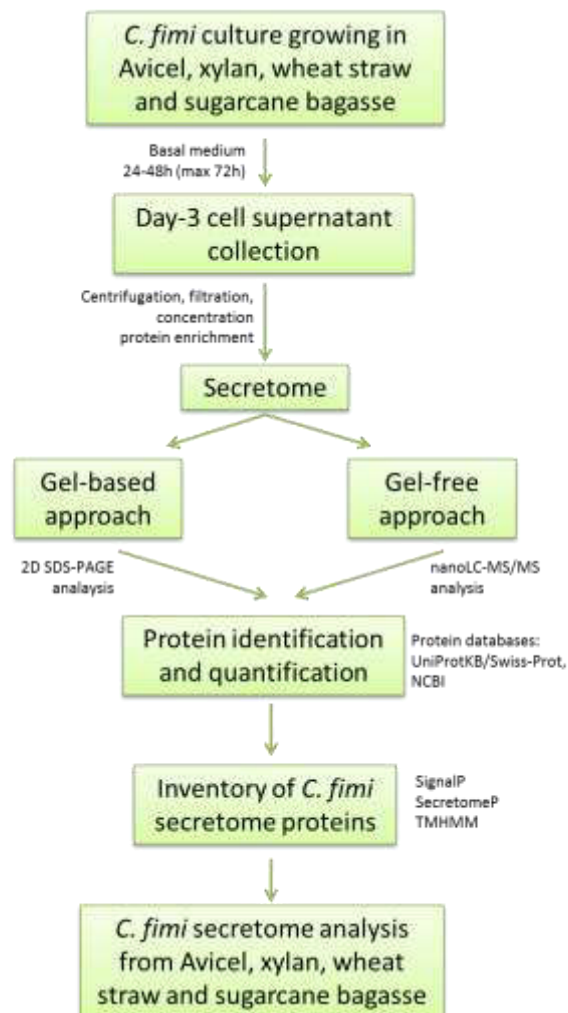


Figure 4.1: Workflow for secretome analysis of *C. fimi* grown on four types of carbon source.

## 4.2.2 Protein quantification

### 4.2.2.1 Bradford assay and SDS-PAGE protein gel

Protein samples were taken as described in 4.2.1. The protein concentration in the supernatants was measured by the Bradford assay (111), performed in 96-well plate format. The detailed protocol is described in Section 2.4.1. The proteins secreted by *C. fimi* growing on five substrates were separated by sizes using SDS-PAGE and ran on Mini-Protean Tetra cell apparatus (Bio-Rad, USA). The detailed protocol is described in Section 2.4.2.

## 4.2.3 Enzyme activity assays

### 4.2.3.1 Qualitative plate activity assay

Carboxymethyl (CM)-cellulose and beechwood xylan hydrolysis were examined using Congo Red dye (162). Undiluted supernatants from the culture were loaded onto 1.5% Difco agar plates containing 0.2% CM-cellulose in 50 mM potassium phosphate, pH 7.0. The plates were incubated overnight at 30°C, and then flooded with 2 mg/mL Congo Red solution. After 15 min, the Congo Red was rinsed off with distilled water and the plate was washed with 5% acetic acid to produce a contrasting background and enhanced visualization of the clearance zones. The control of the experiment was carried out using denatured protein samples and uninoculated culture supernatant as the negative control. The plate was then photographed.

### 4.2.3.2 Reducing sugar assay in protein solution

The ability of an enzyme to cleave polysaccharides and produce products with reducing ends was assessed after incubating the enzyme with 0.2% of appropriate polysaccharide substrate in 50 mM sodium phosphate at the desired pH and temperature. Unless otherwise stated, the pH used for assays was 7.0 and temperature 30°C. Before and after a set incubation time, 10 µL aliquots were mixed with p-hydroxybenzoic acid hydrazide (pHBAH), heated to 70°C for 10 min, and colour changes detected at 415 nm using a microtitre Tecan Safire2 plate reader (163). A stock solution of the appropriate monosaccharide diluted from 0.1 mg/mL to a 1 mg/mL was assayed to obtain a standard curve. One unit of exoglucanase/xylanase activity was defined as the generation of 1 nmol of sugar per min under these conditions. Glucose or xylose stock solution (1mg/mL) was serially diluted and assayed as the standard curve.



#### 4.2.4 Label-free semi-quantitative proteomic analysis using mass spectrometry

##### 4.2.4.1 1D mini gel electrophoresis and tryptic digestion for GelC-MS of *C. fimi* proteomic samples from culture supernatants

Samples from Day-3 *C. fimi* cultures grown on 4 substrates (Avicel, beechwood xylan, wheat straw and sugarcane bagasse) were collected for this analysis. Supernatant from the culture medium were concentrated 40 times using 20 mL Vivaspin20 10,000 MWCO (Sartorius) concentrators as described in Section 2.4.3. LC-MS/MS was performed to identify proteins in the supernatant fraction of the cultures. This analysis was performed by Mr. Adam Dowle of the Technology Facility at the University of York. A 26  $\mu$ L aliquot of the concentrated protein solution was taken from each sample and replicates for analysis. Each sample aliquot was mixed with 4  $\mu$ L of x10 sample reducing agent (Novex NuPAGE, NP0004) and 10  $\mu$ L of 4x sample buffer (Novex NuPAGE LDS, NP0007) before being heated at 70°C for 10 min. Heated samples were loaded onto the 10% gel (Novex NuPAGE 10% Bis-Tris Gel, NP0301BOX), with a blank (7.5  $\mu$ L water mixed with 2.5  $\mu$ L of 4x sample buffer) loaded between sample lanes. The sample was run into the gel in 800 mL of x1 running buffer (Novex NuPAGE MES SDS running buffer, NP000202) at a constant 200 V until the dye-front had fully run into the gel for about 6 min. Post running the gel was removed from the plastic cassette, washed with water and stained with SafeBlue protein stain (NBS-SB-1L). The gel was then destained with water for a 1 h before imaging.

For the in-gel tryptic digest, stained gel bands were excised and cut into approximately 1 mm pieces before transferring into LoBind Eppendorf tubes. The gel slices were washed with 0.5 M triethylammonium bicarbonate in 50% acetonitrile/50% dH<sub>2</sub>O (v/v: 200  $\mu$ L) for 20 min and repeated once. Supernatant was removed and gel slices were washed again with 200  $\mu$ L acetonitrile once for 5 min followed by drying in the speedvac for 20 min at medium setting. The samples were reduced with 200  $\mu$ L of dithioerythritol (DTE) solution (1.5 mg/mL of 10 mM DTE in aqueous 0.5 M triethylammonium bicarbonate) and incubated at 56°C for 1 h. The gel pieces were cooled down to room temperature and the supernatant was removed. The gel pieces were alkylated with 200  $\mu$ L iodoacetamide solution (9.5 mg/ml (50 mM) iodoacetamide in aqueous 0.5 M triethylammonium bicarbonate) followed by incubation in the dark at room temperature for 30 min. Protein digestion was performed overnight at 27°C. Peptides were then extracted with 50% aqueous acetonitrile, dried in a vacuum concentrate and resuspended in 0.1% aqueous trifluoroacetic acid.

#### 4.2.5 Mass Spectrometry Analysis of *C. fimi* proteomic samples from culture supernatants

Prepared samples as previously described were loaded onto a nanoAcquity UPLC system (Waters) equipped with a nanoAcquity Symmetry C<sub>18</sub>, 5 µm trap (180 µm x 20 mm Waters) and a nanoAcquity BEH130 1.7 µm C<sub>18</sub> capillary column (75 µm x 250 mm, Waters). The trap wash solvent was 0.1% (v/v) aqueous formic acid and the trapping flow rate was 10 µL/min. The trap was washed for 5 min before switching flow to the capillary column. The separation used a gradient elution of two solvents (solvent A: 0.1% (v/v) formic acid; solvent B: acetonitrile containing 0.1% (v/v) formic acid). The flow rate for the capillary column was 300 nL/min. Column temperature was 60°C and the gradient profile was as follows: initial conditions 5% solvent B (2 min), followed by a linear gradient to 35% solvent B over 20 min, then a wash with 95% solvent B for 2.5 min. The column was returned to initial conditions and re-equilibrated for 25 min before subsequent injections.

The nanoLC system was interfaced with a maXis LC-MS/MS System (Bruker Daltonics) with a CaptiveSpray ionisation source (Bruker Daltonics). Positive ESI-MS and MS/MS spectra were acquired using AutoMSMS mode. Instrument control, data acquisition and processing were performed using Compass 1.5 software (microTOF control, Hystar and Data Analysis, Bruker Daltonics). Instrument settings were: ion spray voltage: 1,400 V, dry gas: 3 L/min, dry gas temperature 150°C, ion acquisition range: *m/z* 50-2,200. AutoMSMS settings were: MS: 0.5 s (acquisition of survey spectrum), MS/MS (CID with N<sub>2</sub> as collision gas): ion acquisition range: *m/z* 350-1,500, 0.1 s acquisition for precursor intensities above 100,000 counts, for signals of lower intensities down to 1,000 counts acquisition time increased linear to 1.0 s, the collision energy and isolation width settings were automatically calculated using the AutoMSMS fragmentation table, 5 precursor ions, absolute threshold 1,000 counts, preferred charge states: 2-4, singly charged ions excluded. A single MS/MS spectrum with a 20 min gradient was acquired for each precursor and former target ions were excluded for 30 s.

##### 4.2.5.1 Spectra analysis

The resulting spectra obtained from the LC-MS/MS analysis were searched against the *Cellulomonas fimi* ATCC® 484™ sequence data. Peptide matches were filtered to accept only matches with an expect score of 0.05 or better. All identification was performed at the peptide level with protein assignments inferred from the peptides matched using Mascot program (Matrix Science LTd., version 2.4). This was locally run through the Bruker ProteinScape interface

(version 2.1). The search criteria were specified as follows; the instrument was selected as ESI-QUAD-TOF, trypsin was stated as the digestion enzyme, fixed modifications as carbamidomethyl (C), and variable modifications as oxidation (M). Peptide tolerance was 10 ppm, and MS/MS tolerance 0.1 Da. Results were filtered through 'Mascot Percolator' and adjusted to accept only peptides with an expect score of 0.05 or lower. An estimation of relative protein abundance was performed as described by Ishihama *et al.* (164), whereby an exponentially modified Protein Abundance Index (emPAI) is used to estimate the absolute abundance of proteins in LC-MS/MS experiments. The emPAI offers approximate, label-free, relative quantification of the proteins in a mixture based on protein coverage by the peptide matched in a database search result. This index is defined in equation 1 where  $N_{\text{observed}}$  is the total number of detected peptides, and  $N_{\text{observable}}$  is the theoretical number of observable peptides.

**Equation 4.1: Exponentially modified Protein Abundance Index (emPAI)**

$$\text{Protein Abundance Index (PAI)} = \frac{N_{\text{observed}}}{N_{\text{observable}}}$$

$$\text{Exponentially modified PAI (emPAI)} = 10^{\text{PAI}} - 1$$

In brief, the emPAI value reflects the fact that the amount of a protein present in a mixture relative to all other proteins in a mixture is best represented by the ratio of how much of the sequence of each protein is detected in an analysis, rather than by the number of peptides found for each protein (i.e. spectral count). The way emPAI values are calculated by the MScot program is based on an estimated value for  $N_{\text{observable}}$ . To make the data more useful, the emPAI value for each protein was converted into a molar fraction percentage (MFP) value using the formula described by Ishihama *et al.* (164) as shown below. The  $\Sigma$  (emPAI) is the sum of all individual emPAI values from a single LC-MS run dataset.

**Equation 4.2: Molar Fraction Percentage**

$$\text{Molar Fraction Percentage (mol \%)} = \frac{\text{emPAI}}{\Sigma (\text{emPAI})} \times 100$$

This value represents as a percentage how much of the total amount of protein in the sample is represented by a given protein.

The nucleotide sequence of each protein detected was retrieved and submitted to a Basic Local Alignment Search Tool (BLASTx) search (<http://blast.ncbi.nlm.nih.gov>). The highest ranking most matched BLAST hit was recorded for each protein. The UniProt identifier for each of the recorded BLAST assignments was obtained and recorded (<http://www.uniprot.org>). The UniProt identifier for each protein was used to select a suitable gene ontology term to allow protein to be grouped by function. If a protein was found to represent a glycosyl hydrolase (GH), polysaccharide lyase (PL), auxiliary activity (AA), or Carbohydrate Binding Module (CBM), the UniProt identifier for that protein was used to determine to which family the protein belonged to by searching the Carbohydrate-Active enzyme database (CAZy, <http://www.cazy.org>).

#### **4.2.5.2 Notes on Molar Fraction percentage (MFP) values**

There are often cases where a peptide is present in more than one protein in the database searched; in these cases, proteins with overlapping peptide sequences are grouped into families. Members in families contain at least one overlapping peptide sequence; in these cases, it is impossible to say from which protein the non-unique peptides are truly derived. As MFP values are calculated by dividing each individual emPAI value by the sum of all emPAI values in a sample, MFP values can only be used to make direct comparisons within that sample. The actual amount of protein one percent represents in each sample could be vastly different; therefore, differences must be discussed in terms of relative abundance.

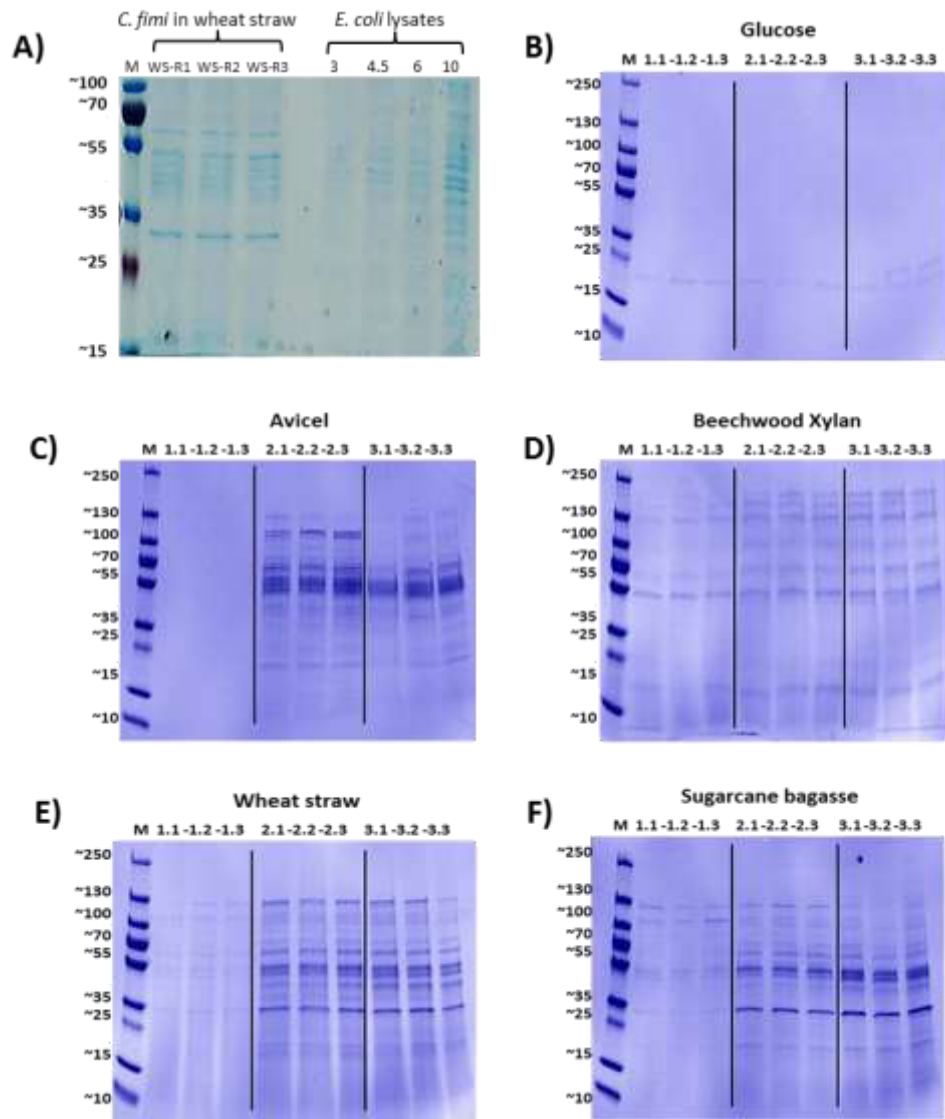
## 4.3 RESULTS

### 4.3.1 Protein quantification

#### 4.3.1.1 Sodium Dodecyl-Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

*C. fimi* was grown in shaking flasks containing glucose (negative control), Avicel, beechwood xylan, wheat straw and sugarcane bagasse as the sole carbon sources. To assess the range of extracellular enzymes produced by *C. fimi* growing on different substrates, protein samples from the extracellular medium were harvested at Day-1, Day-2 and Day-3 of cultures and analysed. The first step was to visualise the protein profiles of *C. fimi* to determine how much protein was present in the supernatant of the cultures. Figure 4.1A shows an example for estimation of protein concentration in the samples using SDS-PAGE gels and Coomassie stain. The overall amount of protein after staining produced in wheat straw culture seems most comparable to 10 mg of *E. coli* protein. This rough estimation would mean that 100  $\mu$ L of concentrated protein supernatants from *C. fimi* cultures contained at least approximately 100  $\mu$ g of protein for subsequent protein analysis.

The profiles of secreted proteins by *C. fimi* during the 3-day time course in individual substrates were visualised on SDS-PAGE gels. Different patterns of protein bands were seen from cultures on different substrates in a reproducible manner across three replicates. Minimum amounts of proteins were secreted during day-1 and increased towards day-3 in all cultures. Culture supernatants of *C. fimi* grown in glucose showed very limited protein secretion for the whole 3-day time course (Figure 4.2B). Contrarily, protein bands were observed on the gels after Coomassie staining with *C. fimi* supernatants from Avicel, xylan, wheat straw, and sugarcane bagasse cultures (Figure 4.2 C to 4.2F). Similar pattern of protein bands at a size range from 55 – 100 kDa were detected from Avicel and xylan supernatant cultures, with relatively more protein in Avicel cultures from day-2 and day-3 based on intensity of the protein bands (Figure 4-2C). However, from the xylan culture, unique protein bands appeared at higher molecular weight (130 – 200 kDa) that were not present at all in other culture samples. In the wheat straw and sugarcane bagasse culture supernatants, protein bands were visible approximately at 35 – 130 kDa. Despite this, proteins ranging from 70 to 130 kDa were secreted to relatively higher levels in the wheat straw culture than in the sugarcane bagasse culture.

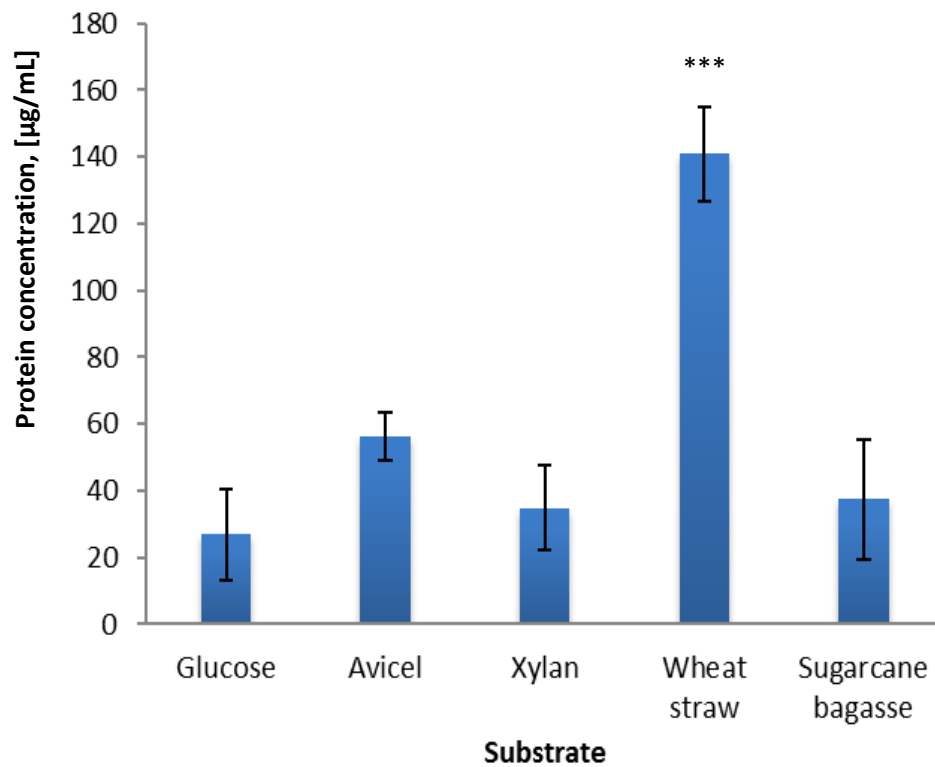


**Figure 4.2: SDS-PAGE secretome profiles of *C. fimi* grown in 5 substrates.**

**A)** Coomassie stained SDS-PAGE gel with 3 replicates of wheat straw supernatant (lane 1 to 3) from day-3 culture of *C. fimi* compared against a range of known concentrations of *E. coli* protein extracts producing a series of standards (lane 4 to 7) for the approximation of protein content. *E. coli* protein standards were quantified by Bradford method at 3, 4.5, 6, and 10 µg of protein. Profiles from day-1, day-2, and day-3 of secretome from the *C. fimi* grown in **B)** glucose, **C)** Avicel, **D)** beechwood xylan, **E)** wheat straw, and **F)** sugarcane bagasse with 3 biological replicates. The first 3 lanes on each gel represented triplicates from day-1 of the culture, followed by subsequent time points in triplicates. Fermentas PageRuler™ Plus Prestained Protein Ladder was used to approximate protein size in a range of 15 to 100 kDa.

#### 4.3.1.2 Bradford assay

Protein concentrations of day-3 cultures showed significantly the highest protein secretion in wheat straw supernatant followed by Avicel culture and all other substrates which showed statistically no difference to each other.



**Figure 4.3: Protein concentration in the supernatant of day-3 *C. fimi* cultures grown on different substrates.**

Protein concentration was determined by Bradford assay (130). A range of serially diluted Bovine Serum Albumin (BSA) was used to generate a protein standard curve. The mean of three biological replicates  $\pm$  SD is shown. Statistical analysis was performed using one-way ANOVA multiple comparisons test, \*\*\*  $P < 0.001$ .































### 4.3.2 Enzyme activity assay

#### 4.3.2.1 Enzyme activity agar plate assay

The activity of endo-/exo-glucanases and xylanases in the *C. fimi* supernatants from a 3-day time course were assayed on 0.2% (w/v) CM-Cellulose and xylan agar plates, respectively. Hydrolysis of these substrates on the agar plates were visualised by staining with Congo red with cleared zones providing a relative measurement of activity from the supernatants. All the culture supernatants samples had detectable enzyme activity against the substrates except the supernatant from the glucose culture, which only showed very faint clearing zones on CM-cellulose substrate agar plate (Figure 4.4).

To quantitatively determine enzyme activity from the plates, the approximate diameters of cleared zones were measured using ImageJ® software (165). Figure 4.5 shows the profiles of clearing zones produced by culture supernatants following the enzyme assay on plates during 3-day time course. The diameters of clearing zones increased with time in all samples compared to the glucose control for each day of the samples were taken. Highest exo-/endoglucanase activity on CM-cellulose was observed at day-3 as indicated by the biggest diameter of clearing zone from supernatant of *C. fimi* grown in wheat straw (Figure 4.5A). The clearing zones produced from the maximum exo-/endoxyylanase activity on day-3 of Avicel, wheat straw and sugarcane bagasse cultures were not significantly different from each other as shown in Figure 4.5B. Minimum activity of exo-/endoglucanase was detected on the beechwood xylan substrate when incubated with the supernatant of *C. fimi* grown in the glucose culture.

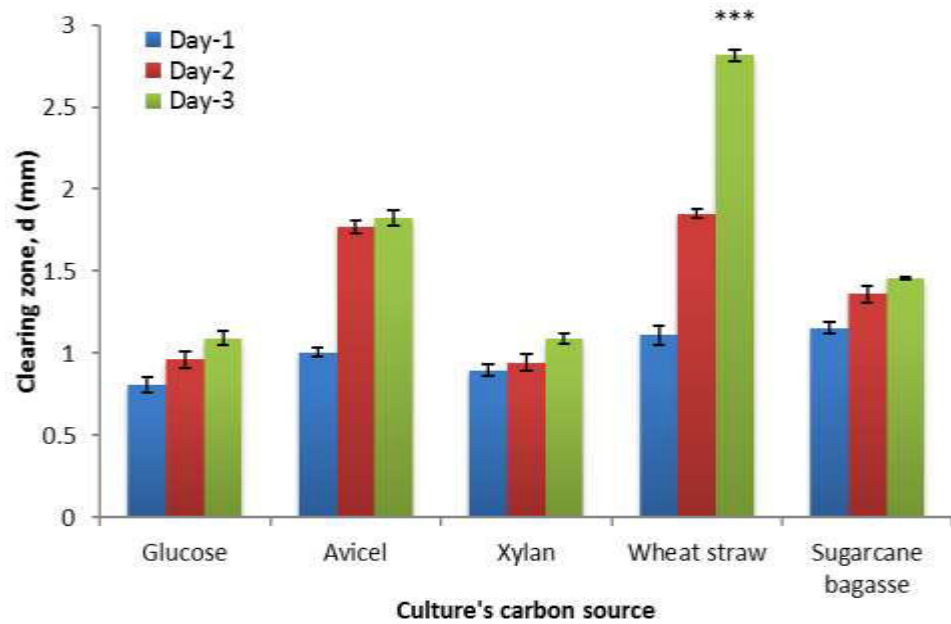


Plate assay substrate	Carboxyl-Methyl Cellulose (CM-Cellulose) activity plate assay					Xylan activity plate assay				
<i>C. fimi</i> Growth Media	Glucose	Avicel	Beechwood Xylan	Wheat straw	Sugarcane bagasse	Glucose	Avicel	Beechwood Xylan	Wheat straw	Sugarcane bagasse
Day-1 Secretome										
Day-2 Secretome										
Day-3 Secretome										

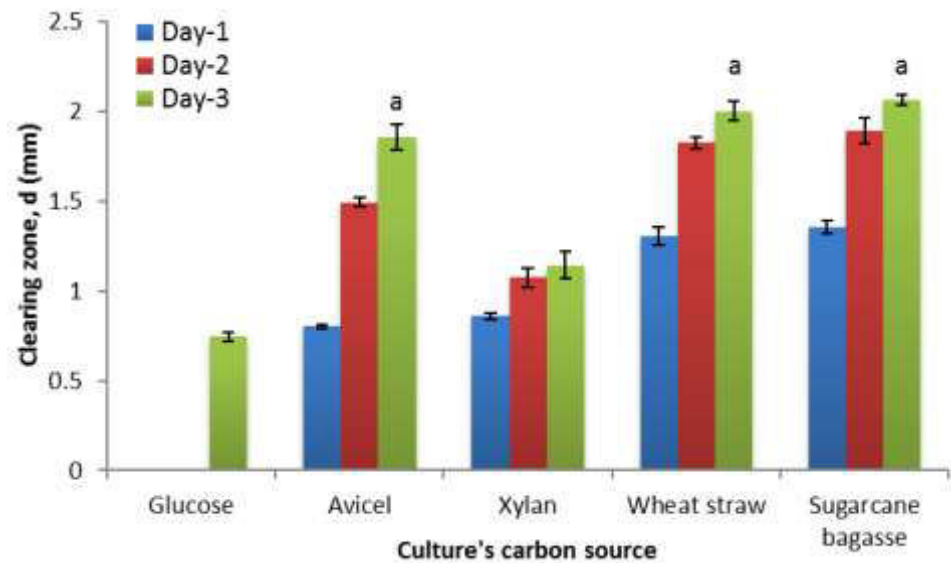
**Figure 4.4: Detection of polysaccharide-degrading enzymes from supernatant of *C. fimi* culture in nutrient agar plates containing substrates stained with Congo Red.**

CM-cellulose and beechwood xylan were used as substrates for the enzymatic reaction at a concentration of 0.2% (w/v), inoculated with 10  $\mu$ L of each biological replicate from unconcentrated *C. fimi* culture supernatant grown on different carbon sources. The plates were incubated at 30°C for 2 days before being stained with Congo Red to detect hydrolysis of the substrates.

## A) Agar plate substrate: CM-Cellulose



## B) Agar plate substrate: Beechwood xylan



**Figure 4.5: Identification of polysaccharide-degrading activities in *C. fimi* culture by the activity plate assay.**

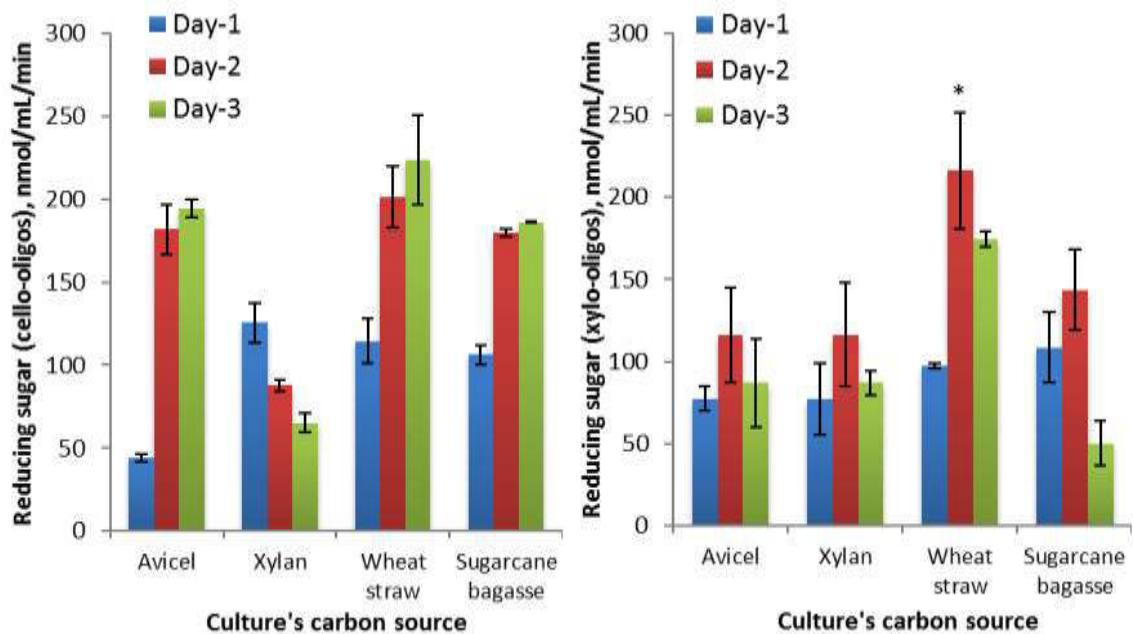
Unconcentrated culture supernatants of *C. fimi* (10  $\mu$ L) grown on different carbon sources from 3-day time courses were loaded onto 0.2% (w/v) A) CM-cellulose, B) Beechwood xylan agar plates. The mean of three biological replicates  $\pm$  SD is shown. Statistical analysis was performed using one-way ANOVA multiple comparisons test, \*\*\*  $P < 0.001$ .

### 4.3.2.2 Reducing sugar assay of protein solution

Biochemical analysis of the enzymatic activity was performed on *C. fimi* culture supernatants using p-hydroxybenzoic acid hydrazide (pHBAH) reagent which reacts with the reducing sugars and produces a colour that can be detected spectrophotometrically (163). Activity of cellulases releasing cello-oligomers during the breakdown of CM-cellulose was the highest in *C. fimi* supernatant grown in wheat straw but not significantly different compared to the other culture supernatants. In figure 4.6A, decreasing cellulase activity against CM-Cellulose was seen in *C. fimi* grown on the xylan over 3 days. In contrast, cellulase activity increased from day-2 and stayed high in cultures grown on Avicel, wheat straw, and sugarcane bagasse. The profile of xylanase activity during the 3-day time course was maximum on day-2 with significantly highest activity in wheat straw culture supernatants but decreased in day-3 in all culture supernatants (Figure 4.6B).

**A) CM-Cellulose reducing sugar assay**

**B) Beechwood xylan reducing sugar assay**



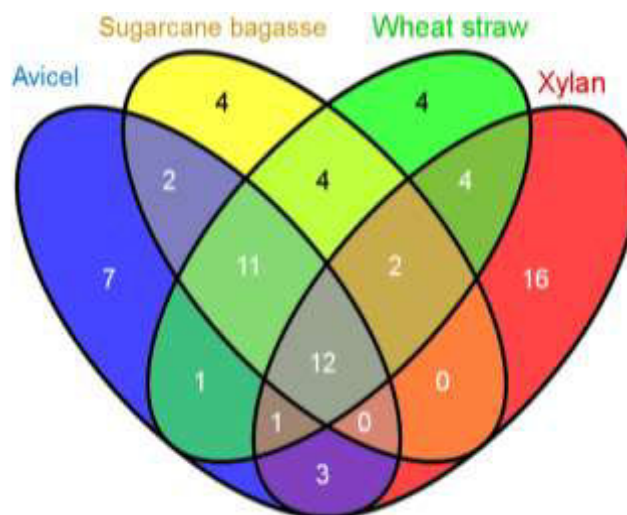
**Figure 4.6: Cellulase and xylanase activity in *C. fimi* cultures grown on different substrates.**

A) Cellulase, and B) Xylanase activity in cultures grown in the presence of 0.2% (w/v) of individual substrates; Avicel, xylan, wheat straw and sugarcane bagasse culture supernatant activities were estimated by calculating the release of reducing sugar after incubation on 1% CM-cellulose or xylan for 1 hour with supernatants. The mean of three biological replicate  $\pm$  SD is shown. Statistical analysis was performed using one-way ANOVA multiple comparisons test, \*  $P < 0.05$ .

### 4.3.3 Label-free semi-quantitative proteomic analysis

Protein samples for this analysis were taken on the Day-3 of culture as described in Section 4.2.4.1. The range and relative abundance of proteins in the *C. fimi* secretome grown on different substrates were evaluated using a label-free semi-quantitative proteomic approach as described under 4.2.4. From the analysis, approximately 30 unique peptides were identified for each secretome sample. This relatively small number of proteins is representative of a low complexity of the protein samples, which reflects the fact that only the secretome of *C. fimi* was examined which does not include intracellular proteins.

The Venn diagram in Figure 4.7 illustrates the commonalities and distinctions between proteins identified in the different samples. The same number of distinct proteins (4 proteins) was secreted on the lignocellulosic substrates wheat straw and sugarcane bagasse, whereas the highest number of unique proteins was observed in xylan (16 proteins). In general, the proteins identified in wheat straw and sugarcane bagasse culture samples were similar, but there were major differences between wheat straw and Avicel, and wheat straw and xylan. Out of the total 71 proteins identified, 12 were commonly secreted by *C. fimi* in all substrates.



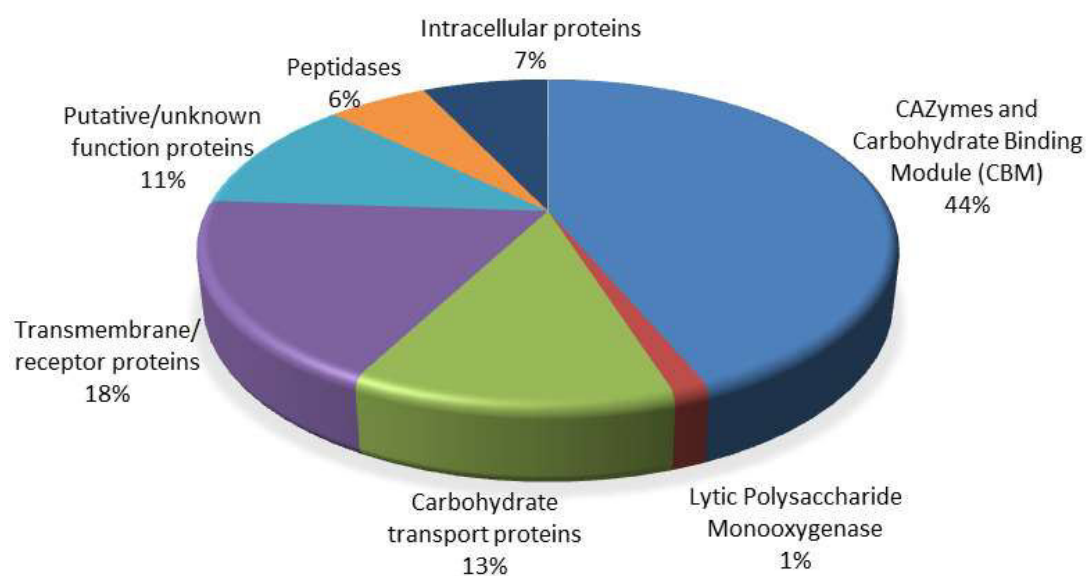
**Figure 4.7: Distribution of extracellular proteins detected in day-3 *C. fimi* cultures.**

The Venn diagram represents 71 unique secreted proteins identified from 20 min gradient LC-MS analysis of *C. fimi* grown on Avicel, beechwood xylan, wheat straw and sugarcane bagasse. The Venn diagram was built using an online web tool (140) (<http://bioinfogp.cnb.csic.es/tools/venny/>).

#### 4.3.4 Sequence-based analysis of the secreted proteins

##### 4.3.4.1 Functional distribution and localization of identified proteins from *C. fimi* secretome

The LC-MS/MS protein dataset of the secretome of *C. fimi* cultivated for 3-days on Avicel, xylan, wheat straw, and sugarcane bagasse was further analysed using available online databases. Figure 4.8 shows that from a total of 71 proteins identified on either one of the substrates, 58% are classified as CAZymes including 44% categorized as glycosyl hydrolases (GHs) and/or CBMs. The only LPMO annotated in the CAZy database for *C. fimi* was also identified in this dataset (1% of the identified proteins). Nine proteins were classified as transport proteins, mostly extracellular solute binding protein family 1 and 5 representing 13% in the overall samples. A further 18% proteins were categorized as transmembrane and/or receptor proteins which also consists of carbohydrate receptor domains. Several hypothetical putative and unknown function proteins have also been identified (11%). A low level of 7% intracellular proteins, however, contributed to the *C. fimi* secretome.

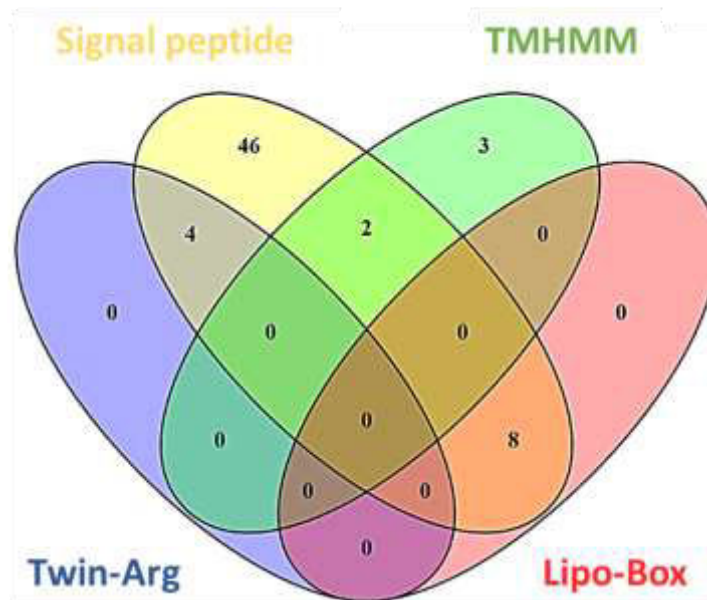


**Figure 4.8: Functional classification of proteins secreted by *C. fimi* grown on Avicel, xylan, wheat straw, and sugarcane bagasse.**

The proteins identified and molar percentage quantified on all test substrates with an unused protein score of  $\geq 2$  were considered for classification.

#### 4.3.4.2 Secretion pathways of extracellular *C. fimi* proteins

A Venn diagram in Figure 4.9 shows the communalities and unique secretion pathways (classical secretion pathway using signal peptide, and non-classical secretory pathway (SP); lipobox signal pathway (LipoP1.0), Twin-Arginine Translocation (TATfind), and Transmembrane Helices Hidden Markov Model (TMHMM)) characterized from all identified proteins of the secretome. From the total of 71 proteins, 60 proteins had a readily identifiable secretory signal with a cutoff value 0.45 for the cleavage site threshold indicating the presence of a signal peptide. Only 3 proteins exclusively contained TMHMM, whilst 12 proteins were predicted not only to comprise a classical secretion signal but also having the alternative of Twin-Arginine Translocation (Twin-Arg) or lipoprotein secretion (LipoBox) pathways.



**Figure 4.9: Venn diagram showing the predicted distribution of four secretion pathways among the 71 identified extracellular proteins of *C. fimi*.**

Predicted secretion pathways were identified using available database servers:

Signal peptide, SignalP4.1 (<http://www.cbs.dtu.dk/services/SignalP/>);

Transmembrane Helices (<http://www.cbs.dtu.dk/services/TMHMM/>);

Twin-Arginine Translocation (<http://signalfind.org/tatfind.html>); and

Lipoprotein translocation (<http://bioinformatics.biol.uoa.gr/PRED-LIPO/input.jsp>).

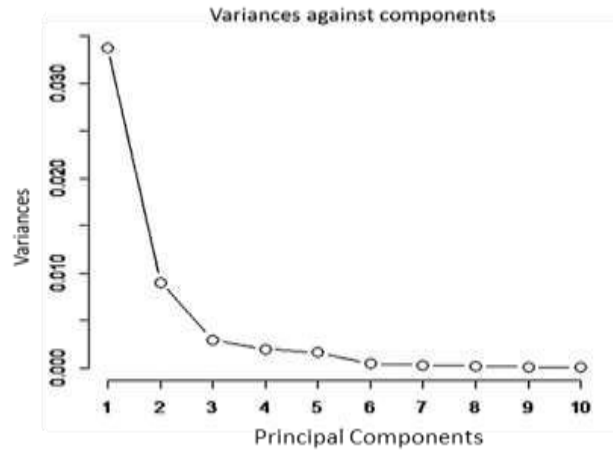


### 4.3.5 Principal Component Analysis (PCA)

A multivariate analysis was performed to reduce the dataset dimensionality of the protein lists and to extract the highly correlated proteins from the results of the LC-MS/MS analysis. For this purpose, a Principle Component Analysis (PCA) was applied to build a correlation matrix of identified proteins using `prcomp` function in R software (RStudio®, Inc. Version 0.98.507). Table 4.2 shows a set of a correlation matrix produced from the PCA where the proportion of variance is the total variance in each component. The first and second components were chosen to represent the total variation in the dataset that consist of 67% and 18%, respectively. The scree plot in figure 4.10 shows that two principle components (PCs) gave the highest percentage and therefore explained most of the variance among all the components which were thus taken into further analysis.

**Table 4.2: Importance of component variables from the *C. fimi* proteomic dataset analysed by Principal Component Analysis (PCA).**

	PC1	PC2	PC3	PC4	PC5	PC6
<b>Standard deviation</b>	0.1836	0.09505	0.05449	0.04468	0.04044	0.02067
<b>Proportion of Variance</b>	0.6668	0.17871	0.05873	0.03948	0.03235	0.00845
<b>Cumulative Proportion</b>	0.6668	0.84553	0.90426	0.94375	0.97609	0.98454
	PC7	PC8	PC9	PC10	PC11	PC12
<b>Standard deviation</b>	0.01744	0.01439	0.01094	0.00889	0.00844	2.72E-17
<b>Proportion of Variance</b>	0.00602	0.0041	0.00237	0.00157	0.00141	0.00E+00
<b>Cumulative Proportion</b>	0.99056	0.99466	0.99702	0.99859	1	1.00E+00



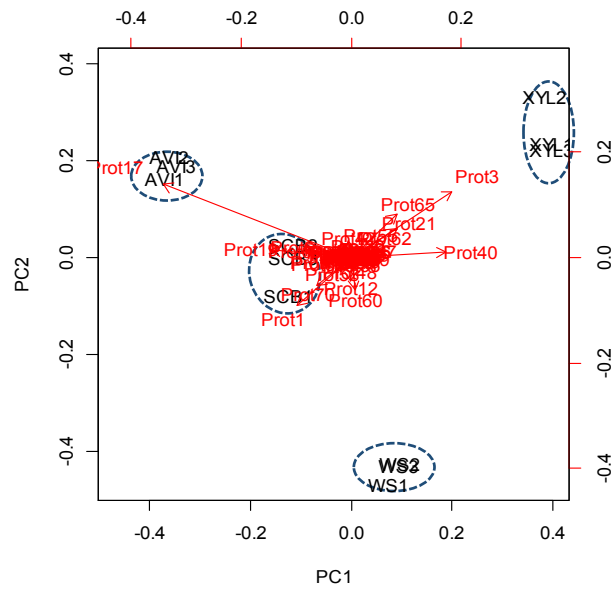
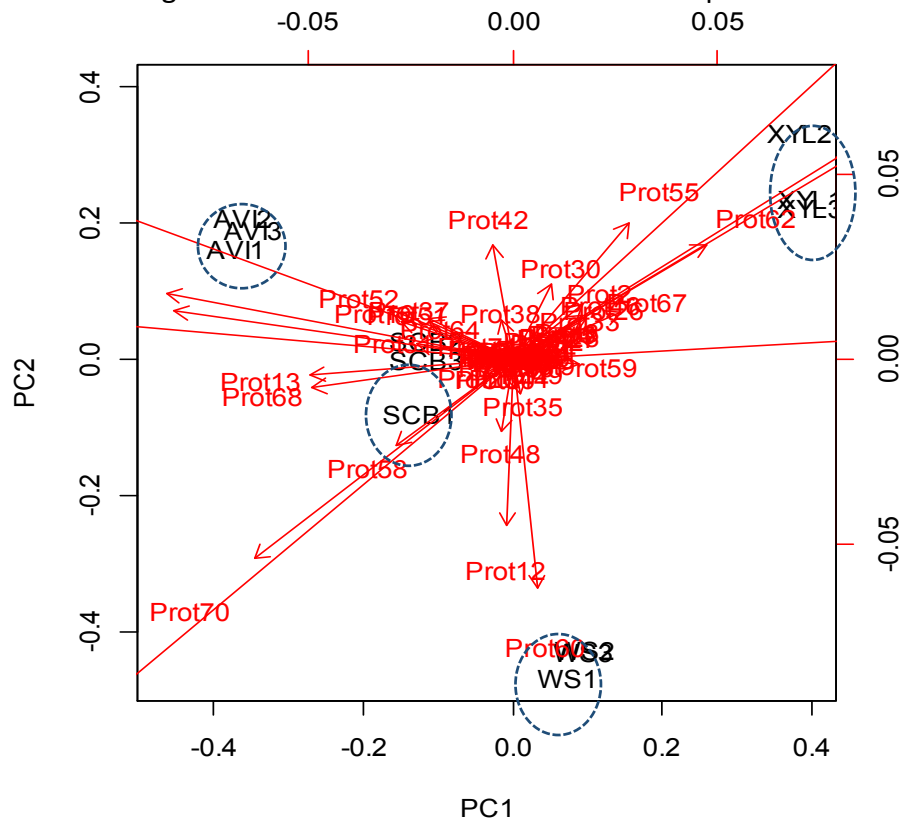
**Figure 4.10: Scree plot from Principal Component Analysis (PCA) of proteomics from day-3 *C. fimi* culture supernatants from growth in Avicel, beechwood xylan, wheat straw, and sugarcane bagasse.**

Variance explained by the top 10 principle components in 12 observations. Each circle represents the individual variance explained by the PC. The first two points (PC1) and (PC2) explained most of the variation in the observation and were hence selected to reduce the dimensionality of the dataset.

The distribution of each unique protein identified in the analysis is shown in a biplot graph (Figure 4.11). The distribution between two principal components has revealed the variation of extracellular proteins secretion by *C. fimi* grown in four different substrates. Three biological replicates of each condition clustered tightly indicating a good reproducibility of the data as marked in the blue circles. Figure 4.11(A) shows the overview of all plotted data loadings, and Figure 4.11(B) is a stretched scale of the biplot graph for a clearer data exploration.

Three types of endo-1,4- $\beta$ -xylanases; 1) Prot17; Celf\_3156, 2) Prot3; Celf\_0574, and 3) Prot40; Celf\_0088 stand out from the analysis. The first xylanase, Prot17 has high ordination in the Avicel and also sugarcane bagasse cultures, and the latter two xylanases (Prot3 and Prot40) were present in a larger number in the xylan culture. This distribution is shown by the arrows in figure 4.11(A). Further examination of the PCA biplot showed that two different modules of CBM2 were highly secreted in wheat straw culture labelled as Prot12 and Prot60. Two proteins, Prot1 (exoglucanase) and Prot70 (cellobiohydrolase, CbhB), were directed at the negative ordinate of PC1 and PC2 specifically in the sugarcane bagasse culture that indicates low secretion of these proteins in sugarcane bagasse culture. All the identified proteins were further analysed for detailed predicted protein domain structure presented in the next section.



A) Inset image of close distribution of extracellular proteins from *C. fimi*B) Rescale image from close distribution of extracellular proteins from *C. fimi*

**Figure 4.11: Biplot graph from PCA analysis of *C. fimi* proteomic dataset using two principal components.**

PCA analysis was performed to discover and summarize the pattern of intercorrelation of protein secretion among four different substrates supplied for *C. fimi* growth.

### 4.3.6 Protein domain prediction

To deepen the understanding of the types of proteins found in the *C. fimi* secretome, the annotation predicted protein sequences were scrutinised by BLAST searches and protein domain predictions. This was useful because the gene annotations provided were mostly taken automatically from the top BLAST hit and do not always give a clear indication of what a protein may do. This is particularly the case for *C. fimi* secreted proteins, which often have multiple domains. For example, a number of these proteins have multiple carbohydrate binding domains and glycosyl hydrolase domains in the same polypeptide, but the automated annotation software used will name it as the first BLAST hit, which will be to only one of these domains. Therefore, the Conserved Domain Sequence (CDS) of identified proteins in *C. fimi* secretome dataset was predicted using a protein alignment method in CDS NCBI Batch-Blast Database and the predicted protein sequences from the *Cellulomonas fimi* ATCC 484™ genome sequence. Protein domain predictions were also made using the CAZy database according to each protein family classification (Appendix A).

Proteins that were identified from the PCA analysis were further examined for their predicted domain structures. The endo-1,4- $\beta$ -xylanase (Celf\_3156) labelled as Prot17 in the PCA biplot that was present only in Avicel and sugarcane bagasse cultures consists of a GH10 and CBM13 domain. From the CAZy database, the GH10 domain has a predicted function as 1,4- or 1,3- $\beta$ -xylanase. A similar protein that contains a CBM13 proved to be able to bind xylan specifically through the arabinose residues. This has been demonstrated for the *Streptomyces lividans* xylanase A and arabinofuranosidase B. A putative sugar binding site contained in this protein has a conserved motif of Q-X-W that is present commonly in Ricin-B domains found in other xylanases (166,167). In addition, this domain also has the affinity to bind galacturonic acid (GalNAc) like a corresponding module of GalNAc transferase 4 in the same bacterium (76). However, the binding specificity of these protein modules (GH10 and CBM13) have not been established yet including for the one in *C. fimi*.

The other two proteins annotated as endo-1,4- $\beta$ -xylanases (Prot3 and Prot40) that were highly correlated with the arrows ordinated to the positive variance to the xylan substrate. The protein with the highest positive variance labelled as Prot3 (Celf\_0574) comprises complex multidomains and was found significantly most abundant on beechwood xylan, equally abundant on wheat straw and sugarcane bagasse but absent from Avicel cultures. Six domains were predicted in this protein including two domains each of CBM22 and CBM9, and a single

domain each of carbohydrate esterase family 3 (CE3) and GH10. From the CAZy depository, the summary of predicted function of each domain in 1,4- $\beta$ -xylanase (Prot3, Celf\_0574) is presented in Table 4.3.

**Table 4.3: Summary of predicted domains found in 1,4- $\beta$ -xylanase (Prot3, Celf\_0574).**

FAMILY	PREDICTED DOMAIN FUNCTION
CBM9	This module is approximately 170 residues and has been found so far only in xylanases. The cellulose-binding function was demonstrated by Boraston <i>et al.</i> study (168) that binds to amorphous cellulose, crystalline cellulose, and the insoluble fraction of oat spelt xylan.
CBM22	A xylan binding function has been demonstrated in several cases and affinity with mixed $\beta$ -1,3/ $\beta$ -1,4-glucans (76).
CBM NC	Carbohydrate-binding “non-classified” module that is not yet assigned to a family.
CE4	A characterized acetyl xylan esterase and been found to have an activity of chitin deactylase. The function of carbohydrate esterase (CE) is to catalyze the de-O or de-N-acylation of substituted saccharides (76).
GH10	Endo-1,4/endo-1,3- $\beta$ -xylanase

The other endo-1,4- $\beta$ -xylanase labelled as Prot40 (Celf\_0088) has a simpler domain structure with single GH10 and CBM2 domains. This protein module has been identified on all five substrates but was significantly the highest in beechwood xylan cultures.

The high level of Prot60 (Celf\_0403) was detected on wheat straw, and it was present in Avicel and sugarcane bagasse cultures, but not identified in xylan culture. Based on the NCBI database, this protein is described as a CBM2 which is the only domain contained in this protein. CBM2 domains are known to bind the crystalline cellulose and have been found in all *Cellulomonas* cellulase families with various domain combinations (86).

Another CBM2 protein which was detected at higher level on lignocellulosic substrates compared to Avicel is Prot12 (Celf\_1913). This protein has two other predicted domains in its structure which are a GH74 and a bacterial neuramidase 2 (BNR2) domain. GH74 is predicted to have a function of xyloglucanase (86), whereas additional BNR2 domain present in this protein is predicted of having a role as sialidase. The dual functions of this domain is predicted to be a

combination of carbohydrate binding and cleaving domain that may be common in microbes as observed in the sialidase from *Vibrio cholerae* (169). It is common not only to other bacterial and parasitic sialidases, but also to other glycosidases involved in pathogenesis (169). Copley *et al.* (170) detected a linear conserved motif of an Asp box in an  $\alpha$ -hairpin from 3D protein folds in the BNR2 domain, which was shown to have significant structural similarity to a single hairpin in the immuno-globulin-like (Ig-like)-domain of chitobiase. The occurrence of this motif in a non-homologous context is unknown but could be associated with the known *O*-glycosyl hydrolase activities of sialidases and chitobiases that occur frequently in proteins that act on, or interact with polysaccharides (170). The x-ray crystal structures of sialidase from *Micromonospora viridifaciens* showed an Ig-like fold that serves as a linker which is homologous to the galactose-binding domain of a fungal galactose oxidase (169).

A full list of CAZymes and other domain structure predictions of the above-described proteins and all identified proteins are summarized in Appendix A. Thirty-two proteins were annotated as CAZymes with various functions including glycosyl hydrolases, pectate lyases as well as carbohydrate-binding modules. NCBI-BLAST server was used to identify the similarity of proteins in *C. fimi* secretome against existing protein structures in Protein Data Bank (PDB). This search was carried out as one of the strategies to identify the best candidate proteins to further study using recombinant protein expression and characterization approaches.

#### 4.3.7 CAZymes secreted by *C. fimi*

Using the *C. fimi* genome data, numbers of CAZymes including the GHs in *C. fimi* were predicted. A summary of well-studied and characterized CAZy-GHs in *C. fimi* are listed and compared for their secretion level in the four substrates supplied in the cultures by a heat map in Table 4.3. The available extensive information of cellulase and xylanase families in CAZy database ([www.cazy.org](http://www.cazy.org)) of *C. fimi* was used to analyse their presence on all four substrates in the *C. fimi* secretome obtained from the nanoLC-MS/MS study (Appendix B).

#### 4.3.8 Cellulases

There were many known CAZymes expressed by *C. fimi* grown on the range of substrates used. Three cellulases, Cel6B/CbhA (Celf\_1925), Cel5A/CenD (Celf\_1924), Cel9B/CenC (endo- $\beta$ -1,4-glucanase C), and Cel48A (CbhB or CenE) were detected on Avicel, wheat straw, sugarcane bagasse but not in xylan. However, Cel6A/CenA (Celf\_3184) was identified in all the cultures

including the xylan and expressed to significantly highest levels in the Avicel culture. A cellulase enzyme that was exclusively found in wheat straw culture was Cel9A/CenB (Celf\_0019). A new GH9 member of *C. fimi* in CAZy (Cel9C/D) was not detected in any *C. fimi* supernatants, whereas, Celf\_1230 (Cel6C), a new member belonging to Cel6 family, was detected under all conditions with significantly highest levels on Avicel, followed by wheat straw and xylan, but was not present in sugarcane bagasse cultures. Three out of four GH6 proteins that belong to *C. fimi* were identified in this study, however, the fourth GH6 (Celf\_0233) was not detected under any conditions.

#### **4.3.9 A LPMO, additional glucanases and potential xylanases**

From this analysis, a number of potentially uncharacterized  $\beta$ -glucanases were identified such as the AA10, GH3, GH16, GH18, and GH64. The AA10 protein, formerly classified as CBM33 in CAZy, is the only LPMO. This LPMO has been detected on wheat straw, sugarcane bagasse and at highest levels in Avicel cultures, but not in xylan cultures. Several GHs observed to be present exclusively in the xylan cultures were the GH3 (Celf\_0583), GH16 (Celf\_3113), GH18 (Celf\_3161) and GH64 (Celf\_3330).

#### **4.3.10 Xylanases**

Theoretically, xylanases may be present in xylan culture supernatants rather than in cellulose (Avicel/CM-Cellulose) supernatant. However, a few exceptions have been observed from *C. fimi* secretome in this study. Xyn10B (Cfx) was predominantly present in the xylan culture, also detected in wheat straw and sugarcane bagasse cultures, but absent in Avicel culture. Meanwhile, Xyn10C (XylC) was not present in any cultures except of Avicel. A recently characterized xylanase XynE/Xyn10 (Celf\_3156) was not seen in xylan and wheat straw cultures, but identified in Avicel and sugarcane bagasse cultures (144,171). The only GH11 protein in *C. fimi* (Xyn11A) was found in xylan supernatant and also even at higher levels on all other substrates.

**Table 4.4: Summary of detected cellulases and xylanases in *C. fimi* secretome grown on four types of carbon sources compared to the CAZy database of *C. fimi* enzymes (2017).**

CAZy FAMILY	IDENTIFIED OF TOTAL PREDICTED	PROTEIN NAME	DESCRIPTION	AV I	XYL	WS	SCB
AA10	1/1	Celf_0270	Chitin-binding domain 3 protein				
GH3	1/11	Celf_0583	GH3 domain protein				
GH5	1/2	Celf_1924	GH5				
GH6	3/4	Celf_1230	1, 4- $\beta$ cellobiohydrolase (Cel6C)				
		Celf_1925	Exoglucanase A (GH6/CelB) Cellobiohydrolase A (cbhA)				
		Celf_3184	1, 4- $\beta$ cellobiohydrolase (GH6)				
GH9	4/4	Celf_0045	GH9				
		Celf_1705	GH9				
		cenC/Cel9B	Endoglucanase C				
		Celf_0019 / CenB	GH9				
GH10	3/5	Celf_0088	Endo-1,4- $\beta$ -xylanase				
		Celf_0574	Endo-1,4- $\beta$ -xylanase (Xyl10B, XynC)				
		Celf_3156	Endo-1,4- $\beta$ -xylanase				
GH11	1/1	Celf_0374	GH11				
GH16	1/3	Celf_3113	GH16				
GH48	0/1	N/D	N/D				
GH74	1/1	Celf_1913	Cellulose-binding family II				

Note: AVI; Avicel, XYL; Xylan, WS; Wheat straw, SCB; Sugarcane bagasse.

N/D; Not detected

Average molar %:



#### 4.4 DISCUSSION

*C. fimi*, a cellulolytic bacterium was grown on five different carbon sources over a 3-day time course. Glucose, Avicel, beechwood xylan, wheat straw, and sugarcane bagasse, respectively, were supplied to the medium. Glucose medium was used as a negative control as glucose should be taken up without further transformation. Avicel and xylan are distinct polysaccharides representing cellulose and hemicellulose fractions of secondary plant cell walls commonly used to characterise enzyme function and the latter two substrates represent complex lignocellulosic biomass from agricultural residues. From estimation of protein concentrations by Bradford colorimetric assay and polyacrylamide gels, it appears that different amounts of protein were secreted by *C. fimi* when growing on the five different substrates. A significantly higher amount of protein was detected in wheat straw and Avicel cultures towards day-3 of incubation. This is an unsurprising observation as *C. fimi* is a well-known cellulose degrader (86) and known to potentially digest more complex cellulose compounds such as lignocellulosic wheat straw (172). The different banding patterns of protein observed on the polyacrylamide gels indicate diverse profiles of digestive enzymes of particularly the cellulases (in range of 44.3 to 223.6 nmol/mL/min) and xylanases (in range of 50.1 to 216.4 nmol/mL/min) which were produced in response to each substrate, which has also been observed in other cellulosic degradation studies (56,144).

Increasing cellulase activity at later time points as identified by the reducing sugar assay from most of the cultures largely agreed with the plate enzyme activity assay. Bigger clearing zones over time were observed from all *C. fimi* cultures except the xylan cultures tested in the plate assay with CM-cellulose as the substrate. A significantly highest cellulase activity was detected from the wheat straw culture compared to the other substrates on day-3 showing a correlation between the clearing zones on plate activity assay as well as the amount of protein quantified in the Bradford assay. Cellulase activity was increased from day-1 to day-3 on Avicel, wheat straw, and sugarcane bagasse cultures suggesting that the cellulases/glycosyl hydrolases were present and recognized the substrates in synergistic mechanisms with Carbohydrate-Binding Modules (CBMs) which enabled the catalytic enzymes to progressively break down the polysaccharides (103,173,174). However, a decreasing pattern of cellulase activity was observed from day-1 to day-3 of *C. fimi* culture grown in xylan substrate. This implies the secretion of enzymes by *C. fimi* depending on the available substrate in the medium (113,139,175) as well as the recognition of decorated hemicellulose with polysaccharide side-chains in the hardwood/beechwood xylan such as arabinose and glucuronic acid (174,176). When the

degradation goes deeper into plant cell wall structure and the more complex hemicellulose-cellulose matrix of the cell wall is reached, and limited xylanase activity to degrade crystalline cellulose occurred hence the reducing activity of cellulase was observed towards the end of incubation.

A similar pattern of xylanase activity was detected in *C. fimi* secretome when grown on all four polysaccharide substrates except for undetected xylanase activity in the glucose culture (data was not shown). The low activity of xylanase degrading the polysaccharide at day-1 in all the cultures corresponded to hard-to-digest hemicellulose matrix being a recalcitrant substrate (63,177) or with the fact, that it takes some time to synthesize the enzymes to digest hemicellulose and to increase their levels from lower background expression. Xylose and other xylo-oligosaccharides were released at maximal level from all the cultures in day-2 (according to the reducing sugar assay) indicating xylanase activity and polysaccharide break down, with significantly highest activity in wheat straw cultures. The breakdown of xylan to reducing sugars of xylo-oligos was increased to about 2.3 times in day-2 compared to day-1 in wheat straw cultures (44.9 nmol/mL/min). However, it was not significantly different from the xylanase activity in sugarcane bagasse (1.3 times higher) and in Avicel (1.1 times higher) on the same day.

It is curious that xylanase activity was also increased on Avicel grown culture but this may reflect the fact that xylan is usually accompanied by cellulose in nature and the bacterium has evolved to “look” for one in the presence of the other (178,179). This activity also suggests that not only xylanase activity was detected during the assay but other bifunctional glucanases such as glycosyl hydrolase family 10 (GH10) in *C. fimi* secreted proteins such as cellobiohydrolases. This observation may suggest that GH10 may possess xylanase and cellobiohydrolase activity depending to the protein domains presence (180,181). However, Clarke *et al.* (182) revealed that xylan hydrolysis catalysed by a different type of *C. fimi* xylanase is mediated by a separate subset of multidomain enzymes to the xylanase containing distinct catalytic and cellulose-binding domains. Those enzymes were encoded by multiple genes and some of the *C. fimi* xylanases are not solely acting on xylan but can also act as exoglucanase that hydrolyse cellulose (182). A similar characteristic has been observed in an aerobic bacterium, *Pseudomonas fluorescens* subsp., *cellulosa* (183) where separate sets of cellulase and xylanase are lacking in cross-specificity yet the xylanase is responsible for the hydrolysis of cellulose and xylan (182,184). From these publications, *C. fimi* may possible secrete some xylanases that could have multiple functions toward the polysaccharides of plant cell wall. On the other hand, the reduced



activity of xylanase was observed in all day-3 cultures which could also be a sign of carbon catabolite repression in the system where enzyme expression was inhibited by certain level of break down products from xylan (185,186).

Biocomputational analyses of the polypeptides detected by nanoLC-MS/MS revealed that 88.7% of total identified proteins contained at least one of the secretion pathway characteristics (signal peptide, lipobox, twin-arginine translocation, or transmembrane helices, TMHMM) for the secretion to the outside of the cell (extracellular space) in *C. fimi*. Secreted proteins can be found in the presence of an N-terminal cleavable signal peptide sequence that is commonly around 15 to 30 amino acids long. This unique region does not show high sequence similarity but sharing a hydrophobic core flanked by N- and C-terminal regions with conserved amino acids at the subsides of -3 and -1 positions (158,159). The program used was SignalP4.1 which predicts the presence and location of signal peptide cleavage sites in amino acids sequences from different organisms (gram-positive, -negative, as well as eukaryotes) (160). A class of secretory proteins known as leaderless proteins is also exported from the cell without signal sequences through non-classical secretion pathways such as cell surface shedding or inclusion in exosomes and other secretory vesicles. Therefore, the proteomic dataset was further analysed using non-classical secretory pathway (SP) prediction database servers that are the lipobox signal pathway (LipoP1.0), Twin-Arginine Translocation (TATfind), and Transmembrane Helices Hidden Markov Model (TMHMM).

The LipoP1.0 server is a web-based server producing predictions of lipoproteins that discriminate between lipoprotein SPs, other SPs and N-terminal membrane helices particularly in Gram-negative bacteria (161). However, the LipoP1.0 server also has a good performance on sequences from Gram-positive bacteria (162). The TATFind software program is used to predict the TAT signal peptides which contain a highly conserved twin-arginine motif that allows for the secretion of folded proteins (163). Many prokaryotes use this pathway predominantly for the secretion of redox proteins, however, analyses of the predicted substrates in Dilks *et al.* (163) suggest that certain bacteria and archaea also secrete mainly non-redox proteins via the TAT pathway. The TMHMM database also used for prediction of transmembrane helices in proteins discriminates between solute and membrane proteins using membrane protein topology prediction with a method based on the Hidden Markov Model (164). A basic general method to locate the identified protein in the sample based on the bacterial genome sequence was also done by using the PSORTb 3.0 server for a prediction of gram-positive bacterial subcellular localization (165).

As much as 85% of total identified proteins (60/71 proteins) contained a signal peptide. Despite the presence of a signal peptide, prokaryotes also may possess several other additional alternative pathways such as twin-arginine translocation (TAT) and lipobox signal in bacteria (187,188). In prokaryotes, a class of secretory proteins known as leaderless proteins is also exported from the cell without signal sequences through non-classical secretion pathways such as cell surface shedding or inclusion in exosomes and other secretory vesicles (189). Taken together, this strongly suggests that a high proportion of discovered proteins in this study are extracellular proteins that are specifically secreted in response to the supplied substrates. The approximately 30 proteins identified in each sample replicate were considered a low number and is evidence of low complexity of the secreted enzyme cocktail with a minimum contamination of 7% from intracellular proteins (5 proteins). The remaining 11.3% of proteins showed unknown/multiple localization, or are present in the bacterial cytoplasm membrane.

The most differentially expressed proteins in this multivariate proteomic analysis were described by PCA analysis of the 71 expressed proteins, illustrated in a biplot graph. In the graph, arrows indicate the best ordination of each protein showing high secretion in the presence of specific or multiple types of polysaccharide substrates. Several proteins have been discovered, amongst them a cellobiohydrolase (Cel6C), and an auxiliary activity enzyme family 10 (AA10). The Cel6C (Celf\_1230) is a new member belonging to Cel6 family (190) and has only recently been biochemically characterized revealing additional endoglucanase activity (144). It is a cellobiohydrolase to unique *C. fimi* and has no close homologous in other actinobacteria as analysed by Blast search. Three out of four GH6s, (1) Celf\_1230; Cel6C, (2) Celf\_1925; CelB, and (3) Celf\_3184; CenA/Cel6A) of *C. fimi* were identified on multiple substrates but were predominantly present in Avicel. This possibly reflected the mixed activity of enzymes on cellobiose as well as xylan (144). The fourth GH6 of *C. fimi*, Celf\_0233 was not detected in any polysaccharide substrate used in this experiment and was also absent in CM-Cellulose and birch wood xylan cultures in the Wakarchuk *et al.* study (191).

Four uncharacterized proteins that have sequence similarity to GH3, GH16, GH18 and GH64 opened the opportunity for further exploration of *C. fimi* secretome in degrading cellulose and lignocellulosic polysaccharides. All four GHs were present exclusively in beechwood xylan cultures, which could be an indication of potentially novel xylanases from *C. fimi*. Yet, this is still to be confirmed by further characterizations as they were also detected in both supernatants of *C. fimi* when grown on CM-cellulose and birchwood xylan in the study of Warkarchuk *et al.* (144)

which may indicate a synergistic activity of the enzymes in degrading plant cell wall polysaccharides.

The sole LPMO of *C. fimi* is a copper-dependent AA10 that was formerly classified as CBM33. LPMOs catalyze the initial oxidative cleavage of recalcitrant polysaccharides after activation by an electron donor (142). The AA10 of *C. fimi* also contains a CBM2 domain. A recent study revealed the importance of specific carbohydrate binding domains for this particular LPMO as the deletion of the natural CBM2 in *C. fimi* AA10 and replacement with another CBM (CtCBM3a) from *Clostridium thermocellum* cellulosome scaffoldin CipA influenced the quantity of non-oxidized sugars during polysaccharide break down. This demonstrates the relevant function of CBMs which can modulate the mode of action of LPMOs (64).

The cellulases and xylanases of aerobic bacteria often have complex molecular architectures which consist of multiple CBMs, and occasionally several catalytic modules in addition to modules of unknown function (192). CBMs bind and attach the enzymes to the substrate of plant cell walls which brings the enzymes into close proximity and prolonged association with its recalcitrant substrate, hence increase the catalysis rate (64,103,174). A few *C. fimi* xylanases which are categorized as GH10 family containing CBMs have been identified in this proteomic study with different secretion level in different cultures including Celf\_0088 (GH10, CBM2), Celf\_0574 (CBM22, CE4, CBM22, GH10, CBMnc, CBM9), and Celf\_3156 (GH10, CBM13). From these three xylanases, two (Celf\_0088 and Celf\_0574) were highly secreted in beechwood xylan cultures, whereas Celf\_3156 was secreted in high amount only in Avicel and sugarcane bagasse cultures.

The CBM2 domain in Celf\_0088 works together with a catalytic domain of GH10 that degrades xylan and xylooligosaccharides by breaking down the glycosidic bonds. The CBM2 domain in this protein has been demonstrated in many cases to have a cellulose-binding function (39,102,103,193). Gilkes *et al.* (194) have shown that mutation of xylanase 10A of *C. fimi* by removal of CBM2a from the protein reduces the hydrolytic activity on cellulose which indicates the important function of CBM2 to enhances the enzyme function. However, several CBM2 modules have been shown to have affinity to mixed  $\beta$ -1,3/ $\beta$ -1,4-glucans and may also bind to chitin or xylan (102). This may explain the detection of xylanase (Celf\_0088) on Avicel as well as on wheat straw and sugarcane bagasse cultures. Additionally, McLean *et al.* (195) showed that CBM2 is one of the CBM families that recognizes different physical forms of prepared cellulose

The protein Celf\_0574 that consists of multicomplex structural domains showed the highest secretion level in beechwood xylan compared to the wheat straw and sugarcane bagasse cultures, and was not detected in the Avicel culture. This may be due to its high affinity toward the substrate with the presence of three xylan-binding modules (two domains of CBM22 and CBM9). The carbohydrate esterase (CE2) plays a role as acetyl xylan esterase that breaks down the ester bonds of polysaccharides. This observation is in agreement with the fact that the well characterized Xyl10B/XynC of *C. fimi* has a large xylanolytic substrate specificity and too much lesser extent cellulolytic activity with a very low secretion level in Avicel culture (195). Noternboom *et al.* (196) have shown consistent unique binding sites of CBM9 within xylan and amorphous cellulose architecture specifically to the reducing ends of sugars.

In this study, an uncommon observation has been made; the lacking expression of a *C. fimi* xylanase (Celf\_3156) in beechwood xylan and wheat straw cultures but its presence at high levels on the other two substrates (Avicel and sugarcane bagasse). The existence of CBM13 domain alongside with GH10 domain has been characterised in several proteins of microbes that bind to monosaccharides of hemicellulose side chains such as galactose (Gal; C4 epimer of glucose), N-acetylglucosamine (GalNAc; an amino acid sugar derivative of galactose), and mannose (Man, a hexose sugar found in hemicellulose) (76). CBM13 was first identified in several plant lectins e.g. ricin which binds to galactose residues. However, in lectins this module binds to mannose, which is a major sugar found in softwood hemicellulose. Lectins are ubiquitous proteins of non-immune origin that bind reversibly and specifically to carbohydrates (197). In contrast to classical CBMs, lectins are carbohydrate-binding proteins typically not appended to enzyme domains (and thus not generally classified in CAZy). The primary role of lectins is molecular/cellular recognition and not enzyme targeting to substrate (197,198).

As the source of xylan used in this *C. fimi* experiment was from beechwood (hardwood), the Celf\_3156 may not have been highly induced to be secreted in the particular culture. A couple of studies have demonstrated that xylanase A and arabinofuranosidase B from *Streptomyces lividans* bind to xylan also containing Ricin B Lectin (199,200). However, the rest of CBM13-protein containing domains in the CAZy have not yet been established for their binding specificity (71). In summary, three *C. fimi* endo-1,4-b-xylanases have been identified in different types of culture which provides an insight into the different mechanisms and strategies by particular xylanases. Each of them may bind and hydrolyze structurally different heteroxylans and xylo-oligosaccharides that are available in the different cultures.

Apart from the predicted GHs found in the *C. fimi* secretome from this study, several more proteins that are potentially involved in carbohydrate metabolism have been identified from their predicted domain structures. Four proteins that stand out for their remarkable domain arrangement and their secretion level in certain substrates are 1) Celf\_2339; Fibronectin type III (FN3) domain protein, 2) Celf\_1913; Cellulose-binding family II (CBM2) protein, 3) Putative hypothetical protein (Celf\_0121); no domain identified, and 4) Celf\_2278; Polycystic Kidney Disease (PKD)-domain containing protein.

The FN3-domain protein, Celf\_2339 is a large protein (210.72 kDa) with five repeats of the FN3 domain that was secreted exclusively in the beechwood xylan culture. This protein shares 61% identity by over 2000 amino acids to the other FN3 proteins in the NCBI database from the Blast search. This revealed the high conservation of the FN3 domains that were found in many other actinobacteria and 11 non-redundant sequences of *Cellulomonas* database entries. The multidomain FN3 arrangement has also been observed in *Clostridium thermocellum* Cbh9A (Cthe\_0413) where tandem FN3 domains have been shown to disrupt the surface of cellulose fibres (201). It is tempting to speculate that this protein could interact with polysaccharide substrates through the FN3 domains or may provide a scaffold for other secreted GHs (202) during the process of carbohydrate break down by *C. fimi*.

The second protein of interest is annotated as cellulose-binding family II (CBM2), Celf\_1913, which contains 918 amino acids and was identified predominantly in wheat straw and sugarcane bagasse, but not in beechwood xylan. This protein has 3 specific domains, apart from a CBM2 there is a GH74 domain and a unique bacterial neuramidase 2 (BNR2) domain. The GH74 is classified as the only one GH74 in the *C. fimi* genome, which is still uncharacterized. The potential activities for the GH74 protein family including endoglucanase, oligoxyloglucan reducing end-specific cellobiohydrolase, and xyloglucanase activities. The BNR2 is a type of neuraminidase which is a virulence factor for many other bacteria including *Bacteroides fragilis* and *Pseudomonas aeruginosa* and usually associated with sialidase activity (169). The presence of this protein at a high level in lignocellulosic substrates triggered attention whether this unusual protein arrangement would contribute to enhance the recognition and digestion of lignopolysaccharides.

One of the putative uncharacterized proteins identified in *C. fimi* secretome is Celf\_1021. This protein was found only in beechwood xylan culture. There is no single domain predicted to have similarity to this protein, however, based on bioinformatics exploration by a collaborator research group in Brazil, this protein has a predicted activity as chitinase based on the protein crystal structure and KEGG database analysis (203). However, the specific activity of this protein has not been characterized yet.

The Celf\_2778, a PKD-domain containing protein, (hereafter it will be referred to as PKDP1) has 836 amino acids and consists of several major domains including pyrroloquinoline-quinone (PQQ) domain, glucose sorbone dehydrogenase (GSDH), polycystic kidney disease domain 1 (PKD1), and protective antigen (PA14) domain. Based on the sequence identity of 26% to PQQ-domain proteins of other microbes in NCBI database, the *C. fimi* PQQ domain has been predicted to serve as a cofactor for a number of sugar and alcohol dehydrogenases in a limited number of bacterial species (204). Most of the characterized PQQ-dependent enzymes have multiple repeats of a specific amino acids sequence but this is not obvious for this this particular protein. The GSDH domain may be involved in oxidation processes that have a role as electron acceptor to the NAD(P)-dependent substrate (205). To understand and exploit the potential LPMOs, understanding the source of electrons is fundamental to bacterial physiology for biomass processing (142).

Another unusual domain in this protein is Polycystic kidney disease 1 (PKD1) domain which has also been identified in other microbial collagenases and chitinases (206). PKD1 has been fairly well studied and predicted potentially for protein-carbohydrate recognition (88). However, the specific role of this domain in carbohydrate metabolism particularly by *C. fimi* is still unclear. The third domain in this protein is a PA14 belonging to a GH3 superfamily domain and also has not been characterized yet. The PA14 domain may as well have a carbohydrate-binding function in polysaccharide digestion (207). In summary, the unique PKDP1 domains structure with a potential electron donor in PQQ domain at the N-terminus and a catalytic GH3 domain at C-terminus may suggest that this protein has a redox equivalent in its whole domain structure. This speculation warrants more evidence and characterization, therefore, the PKDP1 protein was selected for further study by recombinant protein production and characterization as presented in the next chapter. In conclusion, the study described in this chapter provides a more comprehensive and comparative

view of the capability of *C. fimi* in degrading cellulose, hemicellulose and the more complex and recalcitrant lignocellulosic biomasses of wheat straw and sugarcane bagasse. Several potentially new CAZymes have been identified and further characterization of these proteins would benefit a deeper knowledge and understanding especially of the mechanisms of lignocellulosic biomass degradation by this bacterium. Further investigation could be carried out to explore novel mechanisms to degrade lignocellulosic polysaccharides and the potential of *C. fimi* as lignin degrader.

## 5 Attempted Characterization of a *Cellulomonas fimi* PKD-Domain Containing Protein (PKDP1)

### 5.1 INTRODUCTION

#### 5.1.1 Multidomain structure prediction of the *C. fimi* PKD-domain-containing protein, Celf\_2278.

Studies were focused on *C. fimi* because of its demonstrated ability to digest cellulose containing substrates, and the large number of CAZy genes present in its genome. Many of the more obvious glycosyl hydrolase genes from *C. fimi* have been previously characterised. In order to identify possible new enzymes active on lignocellulosic substrates, proteomic studies were undertaken, as described in the preceding chapter. A PKD-domain-containing protein (UniProt ID: F4H2Q6, Protein ID: Celf\_2278) was identified as a protein of interest in the *C. fimi* genome, because of its interesting domain structure. Like many of the CAZymes in *C. fimi*, Celf\_2278, hereafter referred to as PKD domain-containing protein 1 (PKDP1) has a predicted multi-domain structure.

The predicted PKDP1 polypeptide is 836 amino acids long and has three identifiable protein domains. The N-terminal region shows sequence similarity to glucose/sorbone dehydrogenases, the central domain to polycystic kidney disease domains (PKD) that are typically involved in protein-protein or protein-carbohydrate interactions, and the C-terminal region shows homology to P14 domains that are found in some GHs such as glucosidases. PKDP1 has a predicted secretion signal at its N-terminus and as the protein is found in the extracellular medium, it is almost certainly an extracellular protein. This combination of features, and the fact that the PKDP1 is found in the proteome of *C. fimi* growing on lignocellulosic substrates suggests a potential role for the protein in lignocellulose degradation that might involve the generation and transfer of electrons for oxidative enzymes such as LPMOs. The *C. fimi* genome contains 5 different PKD domain-containing proteins. In the proteomic analysis of *C. fimi* grown on 4 lignocellulosic substrates, only 3 PKDPs were been identified (Table 5.1). The predicted domain structures of these proteins are presented in Figure 5.1.

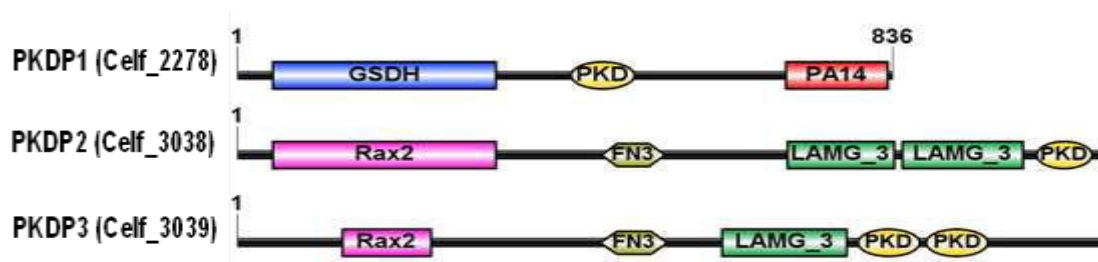


**Table 5.1: PKDPs of *C. fimi* detected in the proteomic analysis using nanoLC-MS/MS.**

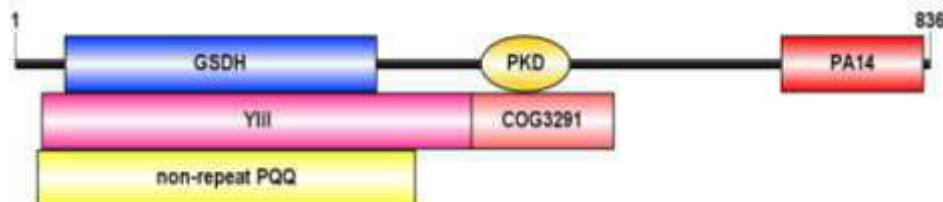
	PKDP in <i>C. fimi</i>	Avicel (Mol %)	Beechwood xylan (Mol %)	Wheat straw (Mol %)	Sugarcane bagasse (Mol %)
1	PKDP1; Celf_2278	0.95	0.91	0.7	0.72
2	PKDP2; Celf_3038	n.d	0.29	n.d	n.d
3	PKDP3; Celf_3039	0.07	0.07	n.d	n.d

\*n.d: Not detected

**A) Structural domain prediction of PKDP1, PKDP2, and PKDP3**



**B) Specific hit of multidomain structure prediction of PKDP1 (Celf\_2278) of *C. fimi***



**Figure 5.1: Predicted domain organization of PKD-domain-containing proteins of *C. fimi*.**

A) 3 different PKD-domain-containing proteins in *Cellulomonas fimi* ATCC 484 identified from *C. fimi* proteomic analysis. B) The detail of predicted conserved domains in PKDP1. The predicted domains were retrieved from amino acids sequences and blast search against NCBI Batch Web Conserve Domain (CD)-Search Tool (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). The domains were drawn to the closest approximate scale using Illustrator for Biological Sequences (IBS 1.0.1) (<http://ibs.biocuckoo.org/>).

### 5.1.2 Polycystic kidney disease I (PKD1) domain

Polycystic kidney disease (PKD) domain was originally found as a translated product of the human polycystin-1 gene. PKD expression leads to a genetic disorder that presents as abnormal cysts that develop and grow in the kidneys (208,209). The human PKD1 protein is a glycoprotein with multiple transmembrane domains and a cytoplasmic C-terminal domain (210). A study by Hughes *et al.* indicates that polycystin is an integral membrane protein involved in cell-cell/matrix interactions (210). Subsequently, PKD-like domains have been identified in many hydrolytic enzymes from marine bacteria, such as chitinases, cellulases and bacterial proteases (72,208,211–214). PKD-domains have also been found in archaeal surface layer proteins and glycoproteins (215). For example, two out of three domain types of surface layer proteins (SLPs) belonging to *Methanosarcina* were identified as PKD domains and it has been suggested that these proteins (that were adapted for cell surface functions and intercellular interactions in Archaea) later evolved to regulate cell-cell interactions in Metazoa (215). With advanced proteomics and enzymology approaches, more PKD domains have been found associated with single or multi existing, homologous or other domains, at the C-terminal domain of several proteases such as collagenase from *Clostridium* sp. and serine proteases from *Vibrio* sp. (216).

From a protein structural point of view, PKD domains have a  $\beta$ -sandwich fold that is common to a number of cell-surface protein modules with a distinct protein family (209,217). Madhuprakash *et al.* studied the role of the PKD domain for chitinase A from a proteobacterium, *Alteromonas* sp. Strain O-7 (AlChiA) (72). The analysis revealed that the “WDFGDG” sequence is a highly conserved unique domain region found within the PKD-domain (72) of this particular species. However, based on protein sequence alignment of five *C. fimi* PKDPs, this conserved sequence had a 100% match with the PKDP2 (Celf\_3038) protein sequence of *C. fimi*, but not with other putative *C. fimi* PKDPs. Further investigation on multiple sequences alignments of all five putative PKDPs of *C. fimi* revealed another similar conserved sequence that is found in AlChiA and PKDP2, but the second and third amino acids, aspartate (D) and phenylalanine (F) differed from the original sequence (Figure 5.2). The structure of the PKD domain contains an Immunoglobulin (Ig)-like fold that has previously been shown to form the ligand-binding sites in cell-surface proteins (209).

In the hydrolysis of crystalline chitin (an analogue to amorphous cellulose), the presence of a PKD domain in addition to chitin binding domain (ChtBD) is an added advantage during the hydrolysis process (211). Mutational studies carried by Orikoshi *et al.* revealed the important

role of PKD-domain in chitinase A from *Alteromonas* sp. Strain O-7 (AlChiA) (211). Two amino acids (W30 and W67) were deleted in the PKD-domain sequence which revealed that the domain is involved in the hydrolysis of powdered chitin by increasing hydrolysis efficiency. Furthermore, PKD- and Ig- domains were found to be linked between ChtBD and the catalytic domain in ChiC of *Pseudoalteromonas* sp. DL-6 providing further evidence for the involvement of this domain during the chitin hydrolysis (218).



**Figure 5.2: Partial Sequence alignment of five PKDPs of *C. fimi*.**

A) The conserved sequence “W--GDG” aligned with three PKDPs (PKDP1, PKDP2, and PKDP3) and the PKD-domain in proteins that are similar to AlChiA. B) The conserved sequence of “WDFGDG” in PKD-domain of AlChiA (72) has an exact match in the second PKD domain of PKDP2 (Celf\_3038) of *C. fimi*. The full sequence alignment of all five *C. fimi* PKDPs is presented in Appendix C.

A recent interesting study by Madhuprakash *et al.* (72) has shown an improvement in the catalytic efficiency of chitinase D from *Serratia proteamaculans* (SpChiD) in the degradation of insoluble chitin substrates as a result of fusing the auxiliary domains of either PKD-domain to the protein termini. The fusion of a PKD-domain to the C-terminal end of ChiD increased the overall catalytic efficiency of the fusion mutant ChiD-PKD by almost 2-fold. It has been suggested that PKD-domains may contribute to the flexibility of proteins and are important for effective

hydrolysis of crystalline chitin through an interaction between two aromatic residues and a chitin molecule (211). An earlier study of the Ig domain function in a psychrophilic chitinase from *Moritella marina* (MmChi60) was conducted by Malecki *et al.* in 2013 (219). Here the Ig-like domain was reported to give the protein a long reach over the chitin surface to increase the affinity between enzyme and substrate, particularly in cold environments (220).

### 5.1.3 Protective Antigen (PA) 14

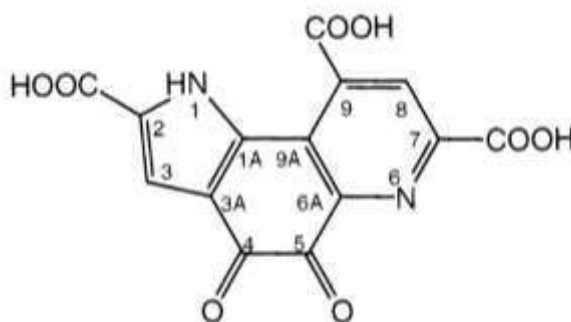
Based on bioinformatics analysis, the annotation of Celf\_2278 PKDP1 of *C. fimi* indicates that PA14 conserved domains have a sequence identity to bacterial  $\beta$ -glucosidases. This domain is named PA14 after its location in the PA20 pro-peptide and possesses a  $\beta$ -barrel architecture based on its crystal structure (204). The PA14 domain sequence is shared by a wide variety of bacterial and eukaryotic proteins but is not seen in any archaeal proteins (204,207,221–225). In the bacteria, this domain has also been found in glycosidases, glycotransferases, proteases, amidases, yeast adhesins, and also in bacterial toxins including anthrax protective antigen (226). The alignment of the PA14 sequence from PKDP1 reveals a conserved hydrophilic residues which indicate that PKDP1 may have a passive binding role despite also having a catalytic function within the GHs superfamily (204). The putative and characterized PA14-containing proteins are involved in carbohydrate binding and/or metabolism and include several glycoside hydrolase families including GH3 (207,224), GH10 (204), GH20 (204) and GH31 (223). Among several, the PA14-containing domains appear to be involved in adhesion and/or signaling, which is consistent with their ability to bind carbohydrate-containing ligands (207,221,222,224).

$\beta$ -glucosidases are classified in the carbohydrate active enzyme database as either GH1 or GH3 (227) and catalyse the hydrolysis of  $\beta$ -glycosidic bonds existing in disaccharides, oligosaccharides and alkyl or aryl  $\beta$ -glucosides thereby preventing the accumulation of cellobiose which would involve in the inhibition of endo- and exoglucanases (228,229). Crystallographic structure and homology modelling of PA14 domains by mutational approaches using a deletion and insertion technique has shed light of the function of this domain. Larsbrink and co-workers have shown the importance of a PA14 domain insert in the recognition of oligosaccharides by the extension of the active-site pocket in GH31 enzyme of *Cellovibrio japonicus* (223). An insertion and deletion mutation experiment of the PA14 domain of another type of  $\beta$ -glucosidase from rumen bacteria, the Bgxa1 ( $\beta$ -glucosidase/ $\beta$ -xylosidase/ $\alpha$ -arabinosidase) indicates that the multifunction activity of Bgxa1 is supported by the PA14 domain. This was shown by a comparison of reducing sugars released after birchwood hydrolysis by Bgxa1 and endoxylanase (Xyn10N8) was increased by 168% compared to Xyn10N8 alone. The result suggests that Bgxa1 acts synergistically with a

cellulase for the saccharification of polysaccharides (224). The PA14 domain has been shown to be involved in conferring substrate specificity in *Kluyveromyces marxianus* Bgxa1 and mutations in this domain decreased activity towards small oligomers, but did not affect the activity against longer chain polymers (207). The result of this deletion is a much more open, flexible active site in Bgxa1 (224). From the Blastp result, the PA14 domain sequence in *C. fimi* PKDP1 has 59% sequence identity to glycosyl hydrolase family 39 (GH39) from *Nonomuraea solani*. Based on the CAZy database, the known activities of GH39 are  $\alpha$ -L-iduronidase (EC 3.2.1.76) and  $\beta$ -xylosidase (EC 3.2.1.37). A further search in CAZy database has revealed another 3 predicted GH39 domain in *C. fimi* genome with protein's locus tags Celf\_1744, Celf\_2981, Celf\_3270.

#### 5.1.4 A multidomain of Glucose/Sorbone Dehydrogenase (GSDH) and Pyrroloquinone quinone (PQQ) dependent domain

Another interesting predicted domain that is found in PKDP1 is a multidomain of glucose/sorbone dehydrogenase (GSDH). This multidomain is a complex consisting of a GSDH  $\beta$ -propeller fold (Ylil) and a non-repeat pyrroloquinoline-quinone (PQQ) dependent domain (See Figure 5.3). Soluble glucose dehydrogenase (s-GDH; EC 1.1.99.17) is known to be a classical quinoprotein first identified in 1979 (230). s-GDH requires the cofactor pyrroloquinoline quinone (PQQ) to  $\beta$ -D-oxidize glucose to D-glucono- $\delta$ -lactone (231–236). PQQ is an aromatic heterocyclic anionic orthoquinone first identified as an enzyme cofactor in bacteria (237). In nature, PQQ (Figure 4.3) serves as an unconventional redox cofactor of membrane-associated dehydrogenases for a number of sugar and alcohol dehydrogenases in a limited number of bacterial species (204,205).



**Figure 5.3: Chemical structure of pyrroloquinone quinone (PQQ).**

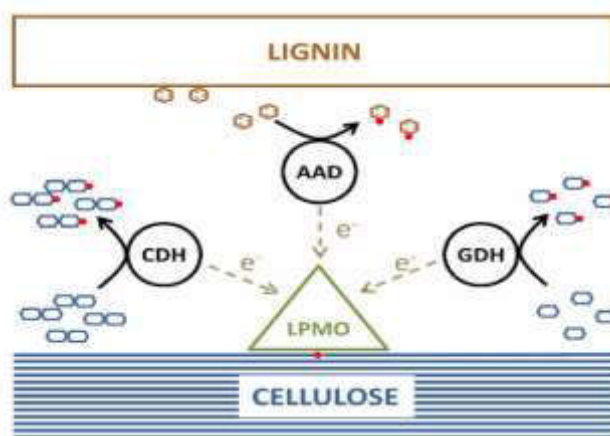
Atom nomenclature is indicated. Figure is reproduced from Oubrie, 1999 (232).

PQQ plays a significant role in quinoprotein dehydrogenases that oxidize numerous substrates, including D-sorbitol, D-gluconate and D-mannitol in the periplasm coupled to the respiratory chain (218). The oxidoreductase activity (GO:0016901) acts on the CH-OH group of the electron donor using quinone or similar compound as acceptor. This reaction may occur in the catalysis of a reduction-oxidation (redox) of CH-OH group acts as an electron donor and reduces a quinone or a similar acceptor molecule involving PQQ and glucose/sorbosone dehydrogenases (238). From the Blast search and sequence alignment on the NCBI database, the best characterised similar protein to the GSDH domain from *C. fimi* PKDP1 is a soluble glucose dehydrogenase (UniProt ID: P13650) with 23% identity at protein sequence level from *Acinetobacter calcoaceticus* (230,238). In the literature, the enzyme is a calcium-dependent homodimer that oxidises glucose to gluconolactone, and uses PQQ as a cofactor. The reaction mechanism comprises a general base-catalyzed hydride transfer. This suggests that PQQ-dependent enzymes use a mechanism similar to that of nicotinamide and flavin-dependent oxidoreductases (232). Similar enzymes have also been found in other prokaryotes; glucose dehydrogenase from *Sorangium cellulosum* (239), aldose dehydrogenase from *Escherichia coli* (240), and the thermophilic archaeon *Pyrobaculum aerophilum* (241), and L-sorbosone dehydrogenase from *Ketogulonicigenium vulgare* (242).

Recent studies in 2014-2015 have identified a novel eukaryotic PQQ-dependent oxidoreductase in a fungus: sorbone dehydrogenase from the basidiomycete *Coprinopsis cinerea* (CcSDH) belongs to a new auxiliary activity family in CAZy database (243,244). The biochemical characterization of CcSDH confirmed its PQQ-dependent activity; the enzyme strongly binds PQQ despite having a low amino acid sequence similarity with known PQQ-dependent enzymes. The discovery of a PQQ-dependent domain in CcSDH could form the basis for a new AA family in the CAZy database. This discovery also revealed the existence of many genes encoding homologous proteins in bacteria, archaea, amoebozoa and fungi. The phylogenetic tree of eukaryotic and prokaryotic quinoproteins shown in Matsumura *et al.*'s (243) work suggests that these quinoproteins may be members of a new family that is widely distributed not only in prokaryotes, but also in eukaryotes.

In 2016, Kracher and co-workers demonstrated that glucose dehydrogenase (GDH) has an ability to reduce quinone molecules (142). GDH was tested together with an LPMO to investigate the electron mediating process, and the study showed that LPMO catalyses the initial oxidative cleavage of recalcitrant polysaccharides after activation by an electron donor (142). This study

showed that the quinone reductases have been implicated in the recycling of quinones during the production of extracellular oxidizers by wood-rotting fungi may well play a role in catalyzing LPMO action. This recent discovery of a PQQ-containing dehydrogenase which belongs to the family AA12 contains a cytochrome domain that possibly transfers electrons generated in its AA12 DH domain to an LPMO (142). Another relevant study by Garajova *et al.* (245) has investigated the ability of AA3\_2 flavoenzymes from *Pycnoporus cinnabarinus*, to trigger oxidative cellulose degradation by AA9 LPMOs when secreted in lignocellulolytic culture. The results showed that glucose dehydrogenase and aryl-alcohol quinone oxidoreductases in the system are catalytically efficient electron donors for LPMOs, and display redox potentials compatible with electron transfer between partners as illustrated in Figure 5.4.



**Figure 5.4: Proposed scheme of fungal synergies for oxidative degradation of cellulose.**

Products of cellulose and lignin degradation are substrates for fungal dehydrogenases (CDH, GDH and AAQO) which provide electrons to LPMOs (Illustration is reproduced from Garajova *et al.*) (245). CDH; Cellobiose dehydrogenase, GDH; Glucose dehydrogenase, AAD; Aryl alcohol dehydrogenase, LPMO; Lytic Polysaccharide Monooxygenase. Red dots indicate di-phenols/quinones serve as redox mediators.

### 5.1.5 Aims of this chapter

PKDP1 of *C. fimi* is a large protein with interesting multimodular domains, but with low sequence identity to other characterized PPQ-dependent proteins. However, based on literature searches, this protein looks like a unique enzyme from *C. fimi* that might be involved in oxidoreductase/dehydrogenase activity. The unusual structure on PKDP1 suggests that this

protein may possess a synergistically redox equivalent between the specific protein domains and/or in the system of cellulose degradation by *C. fimi*. The availability of extracellular electron donor(s) is needed for the oxidative attack on polysaccharides, and the GSDH-PQQ predicted domain in *C. fimi* PKDP1 raised interest in further investigating of this protein by heterologous expression and biochemical characterization techniques. Exciting new findings for LPMO enzymes and their role in cellulose breakdown has triggered this research to look for another potentially new auxiliary activity in *C. fimi*. To achieve this target, heterologous expression of PKDP1 has been carried out in *Escherichia coli* and *Aspergillus niger* expression systems. Due to the specific challenges encountered during attempted the cloning and expression of PKDP1, native protein isolation has also been attempted as a prelude to further biochemical characterization.

## **5.2 MATERIALS AND METHODS**

### **5.2.1 Isolation of native PKDP1 protein from supernatant and substrate-bound fractions induced with specific substrates in *C. fimi* culture**

Proteins from *C. fimi*-induced culture with Avicel, PASC and Cellobiose were collected from the supernatant and substrate-bound fractions at Day-3 and Day-7. The samples were concentrated from 50 mL to 1 mL prior the analysis by SDS-PAGE as described in Section 2.4.2. The supernatant and bound-fraction were separated using centrifugation as described in Section 4.2.1. To selectively extract biomass-bound proteins, 0.5 gram of biomass collected from the cultures was boiled in 10 mL of 0.2% SDS for 5 minutes. Protein was then collected by centrifugation at 4,000 x g and the supernatant collected into a fresh 50 mL centrifuge tube. This was repeated three times, without heating, and with vigorous vortexing between each centrifuge step, to was the biomass of any remaining protein. Extracted protein was precipitated with five volumes of ice-cold acetone overnight at -20°C, before being centrifuged at 4,000 x g and the resulting pellet was washed with 80% ice-cold ethanol. The ethanol-protein mix was then centrifuged again, and the supernatant removed and the pellet air-dried. The protein was then solubilized in 5 mL of ultrapure water and quantified using the Bradford assay (Section 2.4.1).



## 5.2.2 Weak anion exchange chromatography for native isolation of PKDP1

A trial to isolate the PKDP1 was carried out based on the predicted isoelectric point (pI) value of 5.65 using a weak exchanger that contains positively charged groups called diethyl-aminoethyl groups (DEAE) (GE Healthcare). The theoretical pI value of *C. fimi* PKDP1 was computed using online tool on ExPASy. *C. fimi* was grown on 3 separate substrates; Avicel, phosphoric acid-swollen cellulose (PASC), and cellobiose to induce the *C. fimi* to secrete the PKDP1 into the culture medium. The culture was grown in 50 mL autoclaved standard basal medium at 30°C and shaking at 180 rpm according to the procedure described in Section 2.2.2). The cell-free culture supernatant after 7 days was collected for PKDP1 protein isolation using a diethyl-aminoethyl groups (DEAE) column from GE Healthcare column as detailed elsewhere (246). The concentrated fraction was then desalted and buffer exchanged into 20 mM Tris pH 8.0 containing 50 mM NaCl using a Zeba Desalt Spin column (Pierce Biotechnology). The desalted supernatant was purified through a 5-mL anion exchange DEAE column (GE Healthcare) using an Äkta 100 (GE Healthcare), equilibrated with 50 mM Tris pH 8.0, 50 mM NaCl. Proteins were eluted with a 0.05-1 M NaCl gradient in the same buffer at a flow rate of 1 mL/min and the elution was monitored through absorbance at 280 nm. Eluted fractions based on the peak of protein absorbance were analyzed by SDS/PAGE to confirm the presence of the desired native PKDP1 protein.

## 5.2.3 Expression of *C. fimi* PKDP1 in prokaryotic expression systems

Throughout this work, a number of methods for expressing PKDP1 protein in *E. coli* and isolating the native PKDP1 from *C. fimi* culture were evaluated in an attempt to obtain pure, soluble PKDP1 protein. Unless otherwise stated, suitable antibiotics were included at all stages of heterologous PKDP1 protein expression to preserve uniformity of the cultures, and to maintain plasmid selection pressure by using 30 µg/mL Kanamycin for *E. coli* strains BL21 and NEB *SHuffle* T7<sup>®</sup>, and the same concentration of these antibiotics with additional 20 µg/mL of Gentamycin for the *ArcticExpress*<sup>™</sup> expression host.

## 5.2.4 Gene amplification and cloning

In order to express the PKDP1 gene in *E. coli*, the coding sequence was amplified by PCR and cloned into pETFFP\_3 expression vector (GST-tagged vector) to help protein solubility. The Strataclone subcloning kit (Agilent, #240205) and DNA ligase (NEB, #M0202) were used as a cloning method. The peptide tags used in this work to aid purification were maltose-binding

protein (MBP-tag), Glutathione-S-transferase (GST-tag), glutathione fluorescence protein (GFP-tag), polyhistidine (HIS-tag) and Immuno-9 (Im9-tag). The plasmid vectors of all fusion protein constructs were created and kindly provided by Dr. Jared Cartwright (Protein Production Laboratory, Technology Facility, Department of Biology, University of York). The primers used are shown in Table 5.2. Plasmids containing the cloned PCR products were transformed into *E. coli* using a standard heat shock transformation step, after which bacterial colonies were formed. The correct sequence of the PKD gene was confirmed by agarose gel electrophoresis and Sanger DNA sequencing (LightRun, GATC). The plasmid DNA from these colonies was sequenced to verify that the intended gene had been inserted into the vector correctly in-frame and free from any coding errors incurred during amplification. A construct found to be wholly correct was taken forward for protein expression testing.

**Table 5.2: Primers used to amplify PKDP1 (Celf\_2278) for cloning into pETFFP\_3 vector with various solubility tags.**

Primer	Sequence
Celf_2278F (forward)	TCCAGGGACCAGCAATGGGGTTCTCCGAGTCGCTCG
Celf_2278R (reverse)	TGAGGAGAAGGCGCGTTATGGCCGCAGGCGGGTCGTC

### 5.2.5 Expression of PKDP1 in *E. coli* strains

From 5 constructs of PKDP1 created for solubility screening, the GST-tag construct was carried forward to further expression work. The GST-tag construct was transformed into 4 *E. coli* strains (BL21DE3, *ArcticExpress*<sup>™</sup>, *SHuffle*<sup>®</sup>, and *SHuffle T7*<sup>®</sup>). A fresh transformation was carried out for each protein expression experiment to minimise the potential for plasmid loss. Freshly transformed single colonies were used to inoculate starter cultures. These cultures were usually 10 mL in size, contained in 50 mL centrifuge tubes and used Lysogeny-Broth (LB) broth as a growth medium following to the manufacturer recipe (Sigma-Aldrich, Poole UK). The cultures were grown overnight at 37°C (except *ArcticExpress*<sup>®</sup>, where cells were grown at 30°C) with shaking at 180 rpm. The cultures were grown until their optical density at 600 nm (OD<sub>600</sub>) had reached desired value of 0.6-0.8 OD. A first trial of PKDP1 expression was carried out at 10 mL scale. To eliminate antibiotic degrading enzymes which may have built up in the medium of the starter culture, cells from starter cultures were harvested by centrifugation at 4000 x g and resuspended in fresh 50 mL growth medium. The all resuspended cells were used as an inoculum for larger amounts of growth media up to 500 mL. Larger cultures were grown at 37 °C, shaking at 180 rpm until their OD<sub>600</sub> had reached 0.6-0.8 OD. At this point cultures were induced to

express their plasmid encoded gene of interest by supplementing the growth media with isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) to a final concentration of either 0.5 or 1.0 mM to see if these concentrations would give a different effect in terms of protein expression. Once induced, cultures were then incubated at a different temperature to the growth phase, shaking at 180 rpm overnight after which cultures were chilled on ice, and their cells harvested by centrifugation at 4000 x g. The spent growth media was discarded and cell pellets were snap frozen in liquid nitrogen. Cells were stored at -80°C until processing.

### 5.2.6 Larger scale expression culture of heterologous PKDP1 by *E. coli ArcticExpress*<sup>™</sup>

The expression of heterologous PKDP1 was performed in 500 mL Luria-Bertani (LB) medium inoculated with 10 mL overnight culture of *ArcticExpress*<sup>™</sup> grown on LB broth with selected antibiotics of gentamycin (20  $\mu$ g/mL) and kanamycin (30  $\mu$ g/mL). On reaching an OD of 0.6, the culture was induced with 1.0 mM IPTG for an overnight incubation and shaken at 180 rpm at 13°C. Cell pellets were resuspended to 1 g mL<sup>-1</sup> with 20 mM Tris HCl pH 7.4 and 0.1 mM DTT. Sonication was carried out by ultrasonication for 3 x 30 s bursts at 4°C with 1 min interval using Vibra Cell VC375 (Cell<sup>™</sup> Sonic Materials). Cell lysates were fractioned into soluble and insoluble fractions by centrifugation at 15,000 x g for 25 min at 4 °C using a RC-5B centrifuge (Sorvall) equipped with an SS-34 rotor. If no purification was to be carried out these soluble and insoluble fractions were analysed by SDS-PAGE as described in Section 2.4.2.

### 5.2.7 Purification of heterologously expressed PKDP1

Heterologously expressed PKDP1 was purified using Glutathione-S-Tranferases (GSTrap FF, GE HealthCare) that prepacked with Glutathione Sepharose Fast Flow column for fast, convenient, one-step purification of glutathione S-transferase (GST) tagged proteins. Since the C-terminal of pIGF-pyrG vector also consists of His-tag, second purification column also been tested using HisTrap HP column, which was prepacked with Ni Sepharose High Performance and designed for simple, high-resolution purification of histidine-tagged proteins by immobilized metal ion affinity chromatography (IMAC). A polyhistidine tag is an amino acid motif in proteins that consists of at least five histidine (His) residues linked to the N- or C-terminus of the protein. A tag with six His residues was chosen because of its small size, strong metal ion binding, and ability to bind under denaturing conditions. Sepharose beads were washed twice with 10 x volume of binding buffer (1x PBS, pH 7.3) before incubation of the beads with cell lysate at 4°C for 2 h with end-over-end rotation. The bead mixture was then centrifuged at 500 x g, the supernatant collected (flow-through) and the beads washed three times with binding buffer by

centrifugation at 500 x g. Finally, elution buffer (50-80 mM Tris-HCl, 10 mM reduced glutathione, pH 8.0) was added and the mixture incubated at 4 °C for 10 min with end-over-end rotation before the eluted protein was collected by centrifugation at 500 x g. The flow-through, wash and eluates were analysed by SDS-PAGE and anti-GST immune-blotting to determine the presence and purity of the recombinant protein. The same method was applied for second purification trial of PKDP1 recombinant protein using His antibody column.

## 5.2.8 Expression of *C. fimi* PKDP1 in an *Aspergillus niger* expression system

### 5.2.8.1 Cloning of a codon-optimized *C. fimi* PKDP1 gene into pIGF-pyrG expression system

The gene of *C. fimi* PKDP1 without the signal peptide sequence was codon-optimized by GeneArt (LifeTechnologies) for protein expression in *A. niger*. The full sequence of codon-optimized PKDP1 gene is presented in Appendix D. The gene was amplified by PCR using gene-specific forward and reverse primers (Table 5.3) containing HpaI/XbaI sites (underlined) respectively for cloning.

**Table 5.3: Primers used to amplify codon-optimized PKDP1 and add the tag sequence into (Celf\_2278) for cloning into pIGF-pyrG *A. niger* vector.**

Primers were designed to incorporate overhangs for InFusion cloning, a molecular tag (histidine) for purification and the KEX2 cleavage site.

Primer	Sequence
PKD_optF	5' <u>CTAAAGCAAGTCTAGAAAGCGCGGCGGTGGCTCTAGAGTCCCCGCCGGCT</u> 3'
PKD_optR	5' <u>GTCGCGGTTCGACGTTAACGTGATGATGATGATGATGGTAACTTAGGGGCGGAGGCCGG</u> 3'

PCR reaction comprised 100 ng PKDdcp\_Aniger\_opt, 1× PCR buffer, 0.2 mM dNTPs, 0.2 μM of forward and reverse primers and 1 U Q5 HF DNA polymerase in a total of 25 μl reaction volume. PCR was performed using an initial denaturation step at 98°C for 5 min, followed by 30 cycles of denaturation at 95°C for 1 min, annealing at 65°C for 1 min, extension at 72°C for 1 min and a final extension at 72°C for 5 min. PCR reaction products were analysed by gel electrophoresis (1% agarose). A gel-purified and HpaI/XbaI-digested DNA fragment from carrier vector PMA-RQ containing the *C. fimi* PKDP1 gene was cloned into the HpaI/XbaI linearized pIGF expression vector containing the pyrG selection marker gene encoding the orotidine 5'-phosphate carboxylase involved in the biosynthesis of uracil. The pIGF-pyrG plasmid was kindly provided by Prof David Archer, The University of Nottingham, UK). The *A. niger* strain D15 was obtained

from a collection belonging to Prof Simon McQueen-Mason. The generated pIGF-pyrG: PKDP1 expression construct was transformed into *A. niger* as detailed in the Section 5.2.5. The cloning of PKDP1 was done in conjunction with Mrs Luisa Elias (CNAP, Department of Biology, University of York).

### 5.2.9 Preparation of Hutner's trace elements

Hutner's trace elements were prepared to the following compositions. Na<sub>2</sub>EDTA.2H<sub>2</sub>O; 50 g/L, ZnSO<sub>3</sub>7H<sub>2</sub>O; 22 g/L, H<sub>3</sub>BO<sub>3</sub>; 11.4 g/L, MnCl<sub>2</sub>4H<sub>2</sub>O; 0.506 g/L, FeSO<sub>4</sub>7H<sub>2</sub>O; 0.5 g/L, CoCl<sub>2</sub>6H<sub>2</sub>O; 0.16 g/L, CuSO<sub>4</sub>5H<sub>2</sub>O; 0.157 g/L, NH<sub>4</sub>6Mo<sub>7</sub>O<sub>24</sub>4H<sub>2</sub>O; 0.11 g/L. Trace elements were stored away from light at 4°C.

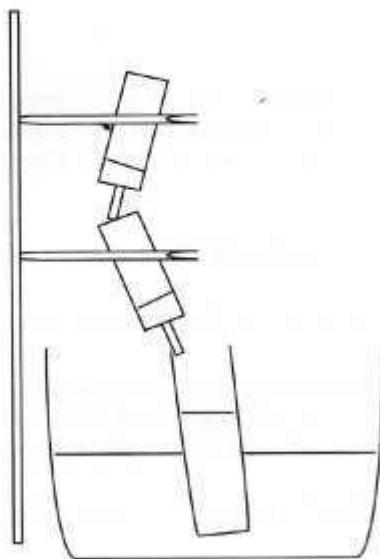
### 5.2.10 Preparation of minimal media

The minimal media contained KCl; 0.52 g/L, KH<sub>2</sub>PO<sub>4</sub>; 1.045 g/L, MgSO<sub>4</sub>; 1.35 g/L, NaNO<sub>3</sub>; 1.75 g/L, and 0.1 mL Hutner's trace elements.

### 5.2.11 Preparation of protoplasts

*A. niger* protoplasts were prepared using a modified method from Yelton *et al.* (247). Siliconized 1 L flasks containing 400 mL of minimal medium plus 4 mM filter-sterilized L-tryptophan were inoculated with 8 x 10<sup>8</sup> spores and shaken at room temperature for 18 h. The mycelium was harvested by filtration through Mira-Cloth (Calbiochem), washed with 0.6 M MgSO<sub>4</sub>7H<sub>2</sub>O, and blotted with blue paper towels for 10 min, and weighed in a sterile 50 mL Falcon tube. The fungal biomass was suspended by vigorous mixing using vortex in filter-sterilized ice cold osmotic medium (1.2 M MgSO<sub>4</sub>7H<sub>2</sub>O/20 mM MES, pH 5.8; 5 mL/g of mycelium), then transferred to a 250 mL flask, and placed on ice. Filter-sterilized solutions of p-glucuronidase (0.2 mL/g of mycelium) and Novozyme 234 (20 mg/mL in osmotic medium; 1 mL/g of mycelium) were added, and the cells were incubated on ice for 5 min in the presence of filter-sterilized 100 mg BSA to lyse the fungal cell wall. The total lysate was incubated at 30°C shaking initially for 30 min at 80 rpm, and then at 50 rpm. After two hours of incubation, 10 µL samples were viewed by microscopy at 30-minute intervals to monitor protoplast formation. Once protoplasts had become visible and detached from mycelium mass, they were purified through sterile polyallomer wool packed into two sterile 20 mL syringe barrels, with one tightly packed, and the other loosely.

The loosely packed syringe barrel was designed to filter out the larger fragments of mycelium, and the tightly packed to ensure that a minimal amount of mycelium fragments contaminated the collected protoplasts. The equipment was set up as in Figure 5.5. The protoplast mixture was poured slowly into the uppermost syringe barrel and allowed to pass through into the collection tube with gravity. The protoplast mixture was collected into 50 mL centrifuge tubes in 15 mL aliquots, and 30 mL of ice-cold NM buffer (1 M NaCl, 20 mM MES at pH 5.8) was added to each tube. These were then gently mixed by inversion and centrifuged at 2500 x g for 10 min at 4°C, after which the supernatant was discarded and the remaining pellet was carefully resuspended in 1 mL cold STC buffer (1.2 M sorbitol, 10 mM Tris base, 40 mM CaCl<sub>2</sub> at pH 7.5) by gentle agitation (vigorous pipetting is likely to destroy the protoplasts). The protoplast mixture was recentrifuged at 2000 x g for 10 min at 4°C, before the supernatant was removed and the whitish pellet of protoplasts was resuspended in 1 mL of ice cold STC buffer. Protoplasts were counted using a haemocytometer as described in Figure 4.6 and diluted with ice cold STC to a concentration of 5 to 10x10<sup>7</sup> protoplasts per mL.



**Figure 5.5: Equipment setup for fungal protoplast collection.**

Protoplasts were separated from mycelium mass of *A. niger* as illustrated above. Two syringe barrels were packed with polyallomer wool (glass-cotton wool), the uppermost barrel more loosely, and the protoplast mixture was allowed to pass through to the second barrel with gravity. The protoplasts were collected into 50 mL centrifuge tube chilled with ice.

#### **5.2.12 Fungal transformation**

Three microliter of vector (1 µg/µL) was mixed with 100 µL of freshly prepared protoplasts in a 15 mL centrifuge tube and incubated at room temperature for 25 min. Then, 200 µL of 60% PEG solution (polyethylene glycol (PEG) 5000/10 mM Tris, 50 mM CaCl<sub>2</sub>, at pH 7.5) was added drop-by-drop using microtip of pipette, and the tube was continuously agitated gently by hand. A further 1.1 mL of PEG was added gradually with mixing before tubes were filled to 10 mL total volume using ice cold STC. The tubes were then centrifuged at 2000 x g for 10 min at 4°C, and the supernatant was discarded. Pelleted cells were resuspended in 300 µL STC. An aliquot of cells (100 µL) were spread onto sorbitol-containing minimal agar plates. The agar plates were prepared with 10 mL salt solution (KCl 26 g/L, MgSO<sub>4</sub>·7H<sub>2</sub>O 26 g/L, KH<sub>2</sub>PO<sub>4</sub> 76 g/L, trace elements 10 mL/L), 6 g/L sodium nitrate, 10 g/L glucose, 218 g/L sorbitol, and 10 g/L agar, at pH 6.5. Plates were incubated at 20°C for three to five days until visible colonies were present. Colonies were transferred to fresh minimal media plates for three rounds of selection on LB agar plates without uridine which selects for the transformants.

### 5.2.13 Direct Colony Polymerase Chain Reaction (DCPCR) using DNA template from *A. niger* hyphae

The DCPCR was adapted from AlShahni *et al.* (248) to verify the transformants after 3 times propagation generations. Hyphae of *A. niger* were picked after 2 days incubated at 30°C and dissolved in an alkaline Tris-NaCl buffer (10 mM NaCl, 10 mM Tris HCl pH 7.5, 1 mM EDTA). The prepared buffer was mixed with 2 µL of 1 M NaOH with 98 µL buffer to make a spore solution. The spores were grown on minimal agar slants for 2 days to have whitish hyphae. Freshly picked hyphae from a plate were resuspended into 100 µL solution before being boiled at 100°C for 10 min. 1 µL of the solution with dissolved hyphae was added to a 10 µL PCR reaction using Q5® NEB standard PCR methods. Three internal forward primers were created for every 400 bp (see Table 5.4) and paired with one reverse primer contained HIS-tag colour-coded with red font); 5'GTCGCGTTCGACGTTAAC~~gtgatgatgatgatgatg~~GTTAACTTAGGGGCGGAGGCGG3'. These primers were used to confirm the recombinant of codon-optimized PKDP1 gene in the pIGF-pyrG vector.

**Table 5.4: Forward primers used for DCPCR of *C. fimi* PKDP1 gene in pIGF-pyrG vector.**

Primer	Forward primer	Expected size (bp)
1	5' CGCCGGCGACCTCCACGTC 3'	~1900
2	5' FCCGGCTCCGCCACCGACTG 3'	~1500
3	5' FTCCGACCCCGACGGCGGCA 3'	~900

### 5.2.14 Expression trials in *A. niger* system

The PKDP1 gene was expressed in 20 mL of ½ strength expression medium (75 g/L maltose, 30 g/L peptone, 0.5 g/L NaH<sub>2</sub>PO<sub>4</sub>·H<sub>2</sub>O, 7.5 g/L MgSO<sub>4</sub>·7H<sub>2</sub>O, 0.04 g/L Tween 80, 10 g/L MES, and 0.5 g/L). The pH of the medium was adjusted to 6.2 and autoclaved. 1 × 10<sup>6</sup>/mL spores were counted using haemocytometer and grown in 2 L shake flasks with 500 mL working volume at 3 tested temperatures, 20°C, 25°C and 30°C for a 6-day time course at 180 rpm. Culture supernatant was separated from the mycelium by filtration through Miracloth (Calbiochem), centrifuged at 4,000 g for 20 min, and clarified through filtration using filter funnel.

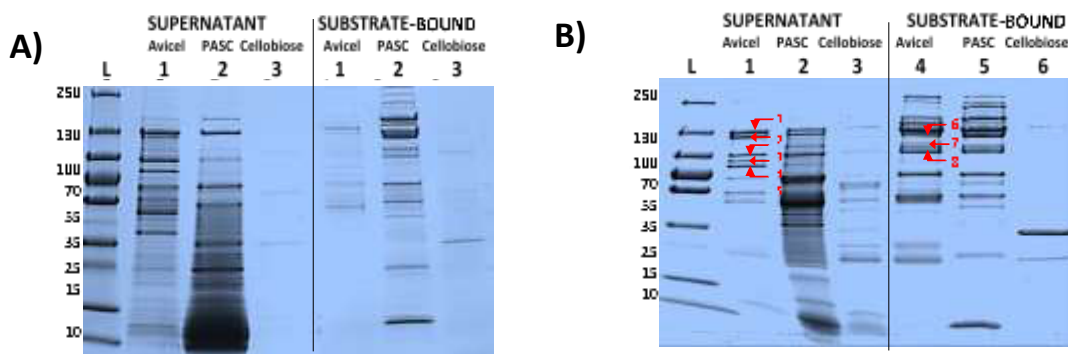


## 5.3 RESULTS

### 5.3.1 Attempted isolation of PKDP1 from *C. fimi* cultures

#### 5.3.1.1 Growth on different substrates

The protein in culture supernatant were concentrated and analysed on the SDS-PAGE gel as shows in Figure 5.6. Following the protein fingerprinting analysis, the spectra from the peptides were searched against 2 databases; the in-house database containing *C. fimi* PKDP1 sequence and also NCBI database. Most of the identified proteins were known to be related to the cellulose breakdown process and none of them matched the *C. fimi* PKDP1 which has predicted size of 86.5 kDa computed from ExpaSy, (online Bioinformatics tool). The protein bands 1 to 8 gave the following results; 1) single match to *C. fimi* endoglucanase C, 2 and 3) no significant matches, 4) Man26A/multifunctional nuclease, 5) Endoglucanase C/peptidase and a single match to multifunctional nuclease, 6) a single match to exoglucanase B/Man26A, finally, 7 and 8) were endoglucanase C. The isolation of PKDP1 from the Avicel inducing culture of *C. fimi* was unsuccessful, therefore, none of this work was taken forward for further analysis as all the identified proteins have been well-studied and characterized for their specific functions in cellulose degradation.

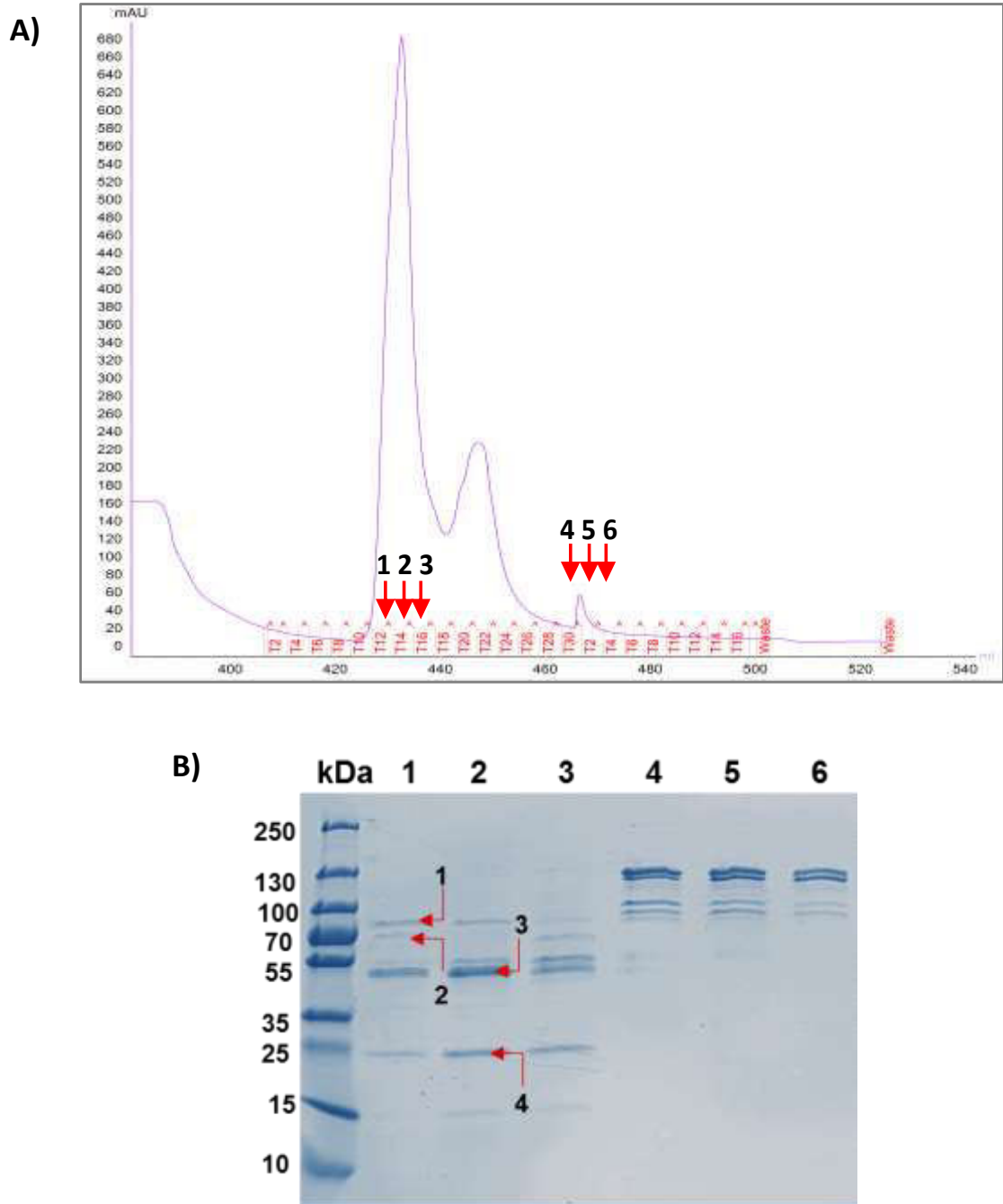


**Figure 5.6: SDS-PAGE analysis of protein extracted from supernatant and substrate-bound fractions induced by specific substrates in *C. fimi* culture.**

SDS-PAGE of A) day-3 and B) day-7 samples. Gels were loaded with 15  $\mu$ L protein samples from *C. fimi* substrate-induced culture. Lane L, ladder (ThermoScientific Pre-stained PageRuler S26619), Lane 1, supernatant fraction of Avicel-induced culture. Lane 2, supernatant fraction of PASC-induced culture. Lane 3, supernatant fraction of cellobiose-induced culture. Lane 4, substrate-bound fraction of Avicel-induced culture. Lane 5, substrate-bound fraction of PASC-induced culture. Lane 6, substrate-bound fraction of cellobiose-induced culture. Eight protein bands were selected for protein fingerprinting analysis from Day-7 samples.

### 5.3.1.2 Isolation trial of *C. fimi* PKDP1 using weak anion exchange chromatography

Two protein peaks detected by absorbance using anion exchange chromatography based on predicted pI value of 5.65 at 280 nm are shown in Figure 5.7A. Protein that bound to the column was collected following the linear gradient elution. Six fractions from the elution were collected and analysed using SDS-PAGE (Figure 5.7B). Multiple bands were observed following SDS-PAGE, suggesting several proteins were bound to the column and the isolation step may not produce a pure single protein. Four bands were selected for protein fingerprinting; two bands were in range of estimated size of *C. fimi* PKDP1, and two smaller size of protein bands were chosen to identify if they can be excluded by the purification based on the ionic charges. The spectra from the peptides were searched against an in house database containing *C. fimi* PKDP sequence and also UniProt database. Result from the analysis showed that none of selected bands matched to *C. fimi* PKDP1. Band 1 to 4 were identified as follows; 1 and 2) Exoglucanase B; 3) Endoglucanase D; and 4) had no significant matches to any protein. The protein spectra of band 4 were further searched against non-redundant RefSeq of *C. fimi* transcriptomic database. This protein spectrum was matched as Cfi\_0001.0003360 hypothetical protein (DUF4397).



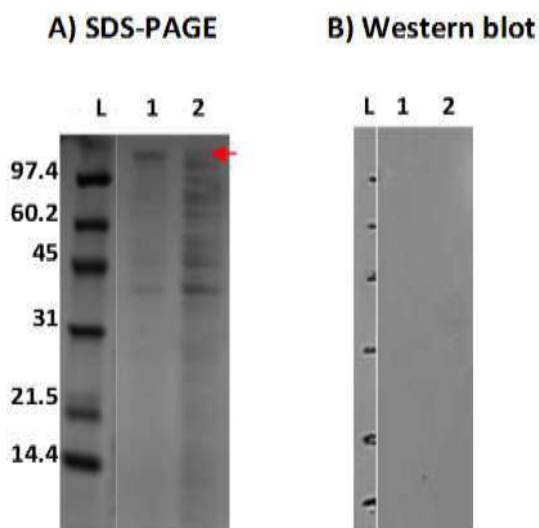
**Figure 5.7: Weak anion exchange separation of extracellular proteins of *C. fimi* grown in Avicel for 7 days.**

A) The chromatogram of protein purification, B) SDS-PAGE gel loaded with 6 selected fractions of purified *C. fimi* extracellular protein from Day-7 Avicel-induced culture. Four protein bands indicated by red arrows were cut out for trypsinolysis digestion prior to protein identification.

### 5.3.2 Recombinant PKDP1 solubility study in *E. coli* BL21(DE3)

The solubility screening for recombinant PKDP1 was carried out by cloning the PKDP1 protein gene into a TFFP of pET plasmid vector series which can be fused to the gene to generate C-terminal fusions to improve the protein solubility (249). The backbone of these vectors was using pETFPP\_0 comprised of HIS\*3cProtease where the \* site contains *Pst*I mutation. The soluble-tagged vectors screened in this work were pETFPP\_1; Histidine (HIS-3cProtease-ORF), pETFPP\_2; contains maltose-binding protein (HIS-MBP-3cProtease-ORF), pETFPP\_3; contains Glutathione-S-transferase (HIS-GST-3cProtease-ORF), pETFPP\_4; contains Immuno-9 (HIS-Im9-3cProtease-ORF) and pETFPP\_5; contains glutathione fluorescence protein (HIS-GFP-3cProtease-ORF). The solubility screening for protein expression work using these plasmid vectors was using an *E. coli* strain BL21DE3. This work was performed by Dr. Jared Cartwright in Protein Production Laboratory, Technology Facility, University of York using a dot-blot technique (250).

Among 5 solubility tags used for heterologous PKDP1 constructs, GST-tagged fusion protein (pETFPP\_3) indicated the presence of soluble protein from a preliminary dot-blot analysis on 96-well plates (data not shown). This protein construct was further analysed using SDS-PAGE (Figure 5.8A). An immunoblot assay using histidine antibody was performed but failed to show full-length protein in either the total or soluble extract (Figure 5.8B). The expression work was repeated using different *E. coli* strains as the expression host (a new line of BL21 DE3, Arctic Express<sup>®</sup>, NEB SHuffle<sup>®</sup> and NEB SHuffle<sup>®</sup>T7). The ArcticExpress<sup>®</sup> is an engineered *E. coli* strain derived from high performance Stratagene BL21-Gold cells to address the common bacterial gene expression hurdle of protein insolubility specifically by lowering the temperature following the induction for expression. The NEB SHuffle<sup>®</sup> and SHuffle T7<sup>®</sup> are chemically competent *E. coli* K12 cells engineered to export proteins containing disulphide bonds in the cytoplasm (251). The *C. fimi* PKD protein contains 2 potential disulphide bonds based on sequence prediction on DiANNA (<http://clavius.bc.edu/~clotelab/DiANNA/>), a web server for disulphide connectivity prediction (252). The western blot procedure was optimized in this study by slight modification of transfer buffer composition by increased the methanol concentration to 10% (v/v). Duration of blotting step on wet transfer also was increased to 30 min with 15 V current to help the blotting process of big fusion PKDP1 transfer onto the membrane.



**Figure 5.8: Protein gel analysis of PKDP1 expression samples following dot blot screening of *C. fimi* candidate proteins.**

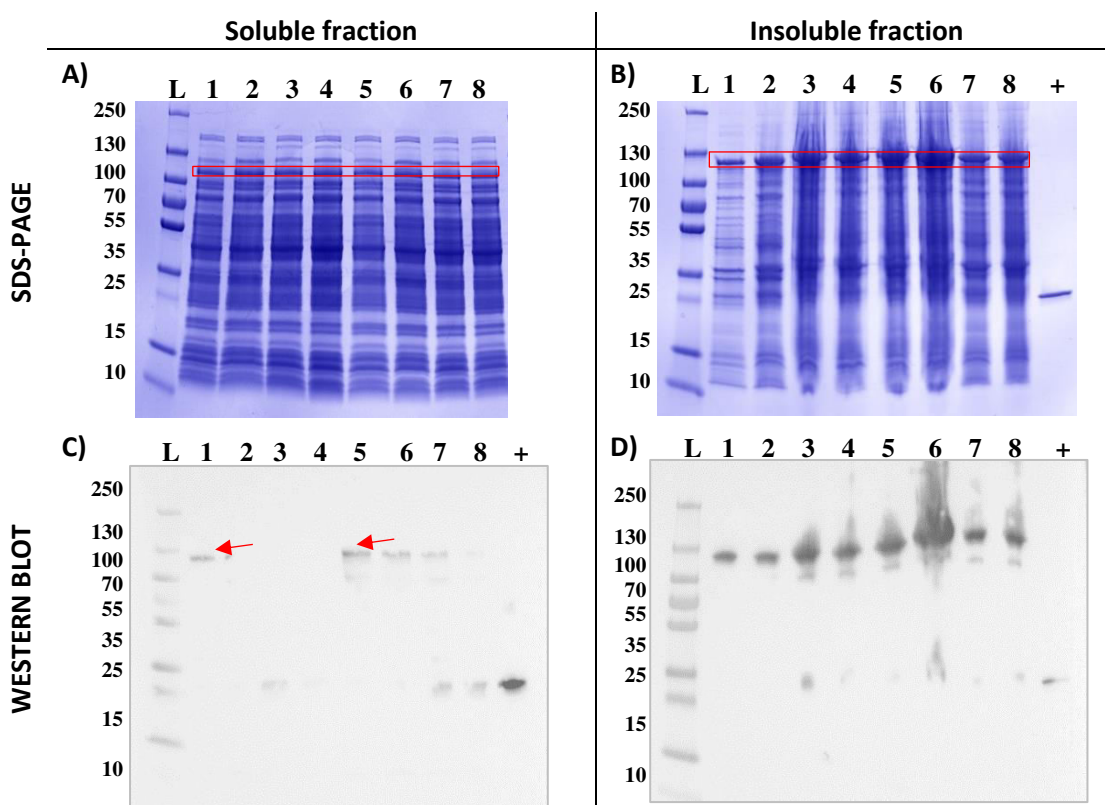
A) SDS-PAGE; and B) anti-HIS Western immunoblot assay of GST-tagged fusion PKDP1 following induction using *E. coli* BL21 (DE3) with pETFPP\_3 (GST-tagged vector) to help the solubility of the protein during the expression. Lane L, ladder (BioRad low range biotinylated protein standard (1610306). Lane 1, total protein fraction of expressed PKDP1 following induction. Lane 2, soluble protein fraction of expressed PKDP1 following induction.

### 5.3.3 Expression trial of *C. fimi* PKDP1

The PKDP1 expression trials were conducted and compared using 4 *E. coli* strains; BL21DE3, *ArcticExpress*<sup>™</sup>, *SHuffle*<sup>®</sup> and *SHuffle* T7<sup>®</sup>. Each strain has its own properties which give it an advantage for heterologous expression. The expressed proteins from the trials were assessed and indicated the presence of full length of GST-tagged PKDP1 at 111 kDa on SDS-PAGE gel for *E. coli* BL21(DE3) on Figure 5.9(A), *ArcticExpress*<sup>™</sup> on Figure 5.10(A), *SHuffle*<sup>™</sup> on Figure 5.11(A) and *SHuffle* T7<sup>™</sup> on Figure 5.12(A)). Results of immunoblotting with the GST antibody indicate the protein clearly was expressed using BL21DE3 cells grown at 10°C following induction either by 0.5 mM or 1.0 mM IPTG. A faint immune-detection of GST-tagged PKDP1 bands was also observed from the BL21DE3 cells grown at 15°C and 20°C following induction at the point when the culture OD was 0.8 by adding 0.5 mM IPTG (Figure 5.9C). A comparable result was observed from the *E. coli SHuffle*<sup>®</sup> and *SHuffle* T7<sup>®</sup> with no protein bands detected in the insoluble protein fraction. Only one soluble protein band was poorly detected from *SHuffle*<sup>®</sup> cell at the highest screened temperature of 25°C following 0.5 mM IPTG induction.

On the other hand, the *E. coli* *SHuffle* T7<sup>®</sup> consistently produced soluble recombinant PKDP1 at a broad range of expression temperatures at 10°C to 25°C with either 0.5 mM or 1.0 mM IPTG induction. Immunoblotting indicated that the *E. coli* *ArcticExpress*<sup>™</sup> was able to express the soluble PKDP1 at expression temperatures of 10°C or 15°C after the post-induction step (Figure 5.10C). This strain did not show any detectable effect of different IPTG concentration of 0.5 or 1.0 mM IPTG. Based on the assessment of small scale expression trials, *ArcticExpress*<sup>™</sup> strain was selected for larger expression of PKDP1.

*E. coli* BL21(DE3)



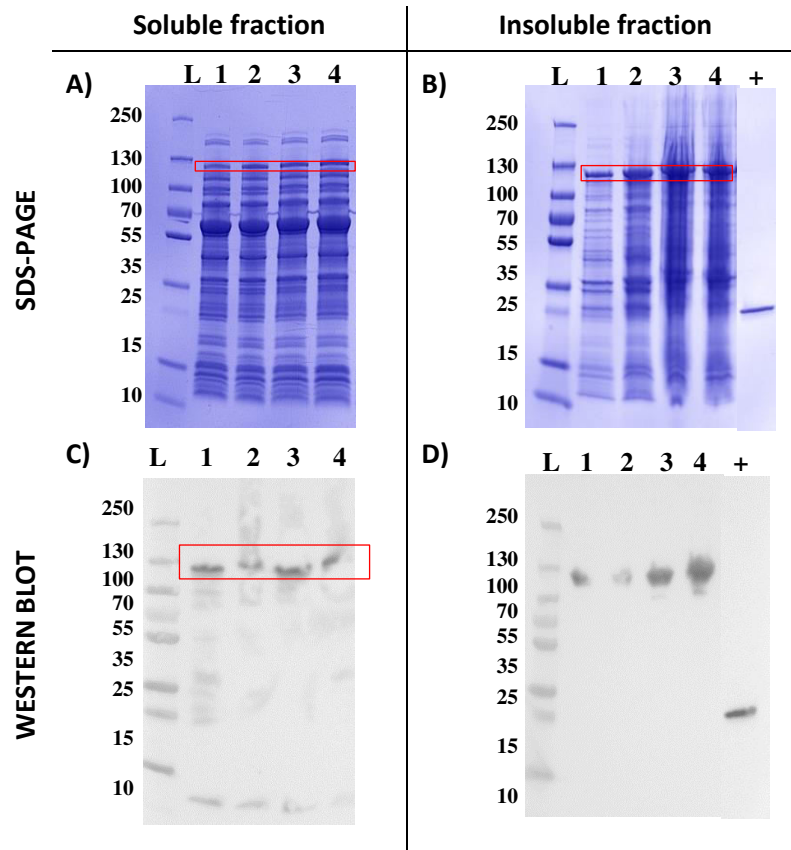
Legend of figures:

Lane	Temperature	IPTG concentration
L (Ladder)	n.a	n.a
1	10°C	1.0 mM
2	15°C	1.0 mM
3	20°C	1.0 mM
4	25°C	1.0 mM
5	10°C	0.5 mM
6	15°C	0.5 mM
7	20°C	0.5 mM
8	25°C	0.5mM
+ (GST-tag empty vector, positive control)	n.a	n.a

**Figure 5.1: Protein gel analysis of expressed PKDP1 following induction in *E. coli* BL21 (DE3).**

(A) Soluble protein fraction on SDS-PAGE. (B) Insoluble protein fraction on SDS-PAGE. (C) Soluble protein fraction by anti-GST Western immunoblot assay. (D) Insoluble protein fraction on anti-GST Western immunoblot assay. Two transformants of *E. coli* BL21 (DE3) cells were selected to express PKDP1 in combination of temperature and IPTG concentration. 15  $\mu$ L of protein samples were loaded onto the gel.

***E. coli ArcticExpress*<sup>™</sup>**



Legend of figures:

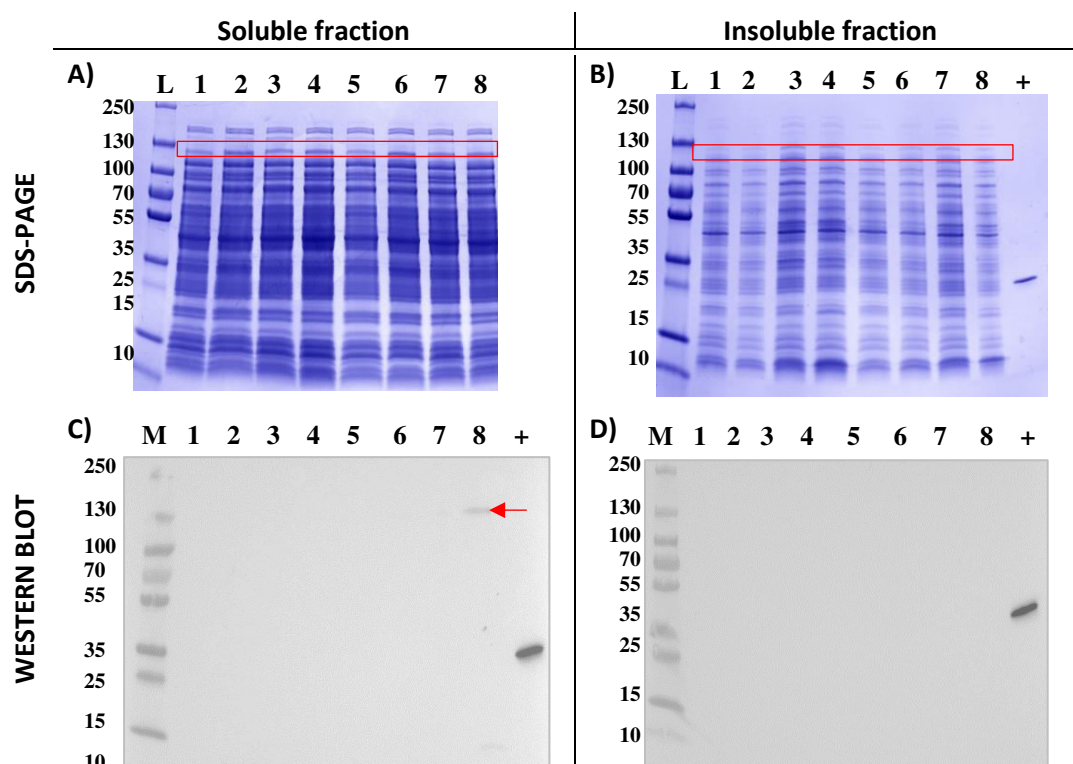
Lane	Sample/Ladder	Expression temperature	IPTG concentration
L	Ladder	n.a	n.a
1	Transformant-1	10°C	0.5 mM
2	Transformant-1	10°C	1.0 mM
3	Transformant-1	15°C	0.5 mM
4	Transformant-1	15°C	1.0 mM
+	GST positive control	n.a	n.a

**Figure 5.2: Protein gel analysis of expressed PKDP1 following induction in *E. coli ArcticExpress*<sup>™</sup>.**

(A) Soluble protein fraction on SDS-PAGE. (B) Insoluble protein fraction on SDS-PAGE. (C) Soluble protein fraction by anti-GST Western immunoblot assay. (D) Insoluble protein fraction on anti-GST Western immunoblot assay. A transformant of *E. coli ArcticExpress*<sup>™</sup> cell was selected to express PKDP1 in combination of temperature and IPTG concentration. 15 µL of protein samples were loaded onto the gel.



*E. coli* SHuffle® (NEB)



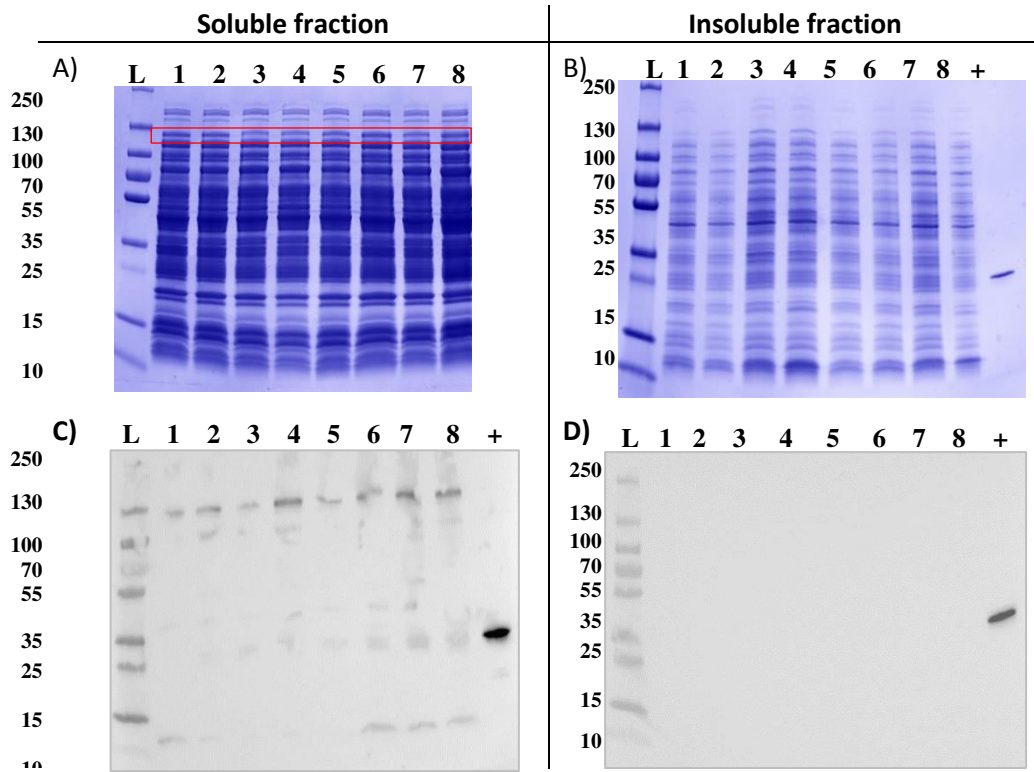
Legend of figures:

Lane	Temperature	IPTG concentration
L (Ladder)	n.a	n.a
1	10°C	1.0 mM
2	15°C	1.0 mM
3	20°C	1.0 mM
4	25°C	1.0 mM
5	10°C	0.5 mM
6	15°C	0.5 mM
7	20°C	0.5 mM
8	25°C	0.5mM
+ (GST-tag empty vector, positive control)	n.a	n.a

**Figure 5.1: Protein gel analysis of expressed PKDP1 following induction in *E. coli* SHuffle®.**

(A) Soluble protein fraction on SDS-PAGE. (B) Insoluble protein fraction on SDS-PAGE. (C) Soluble protein fraction by anti-GST Western immunoblot assay. (D) Insoluble protein fraction on anti-GST Western immunoblot assay. Two transformants of *E. coli* SHuffle® cells were selected to express PKDP1 in combination of temperature and IPTG concentration. 15 µL of protein samples were loaded onto the gel.

*E. coli* SHuffle® T7 (NEB)



Legend of figures:

Lane	Temperature	IPTG concentration
L (Ladder)	n.a	n.a
1	10°C	1.0 mM
2	15°C	1.0 mM
3	20°C	1.0 mM
4	25°C	1.0 mM
5	10°C	0.5 mM
6	15°C	0.5 mM
7	20°C	0.5 mM
8	25°C	0.5mM
+ (GST-tag empty vector, positive control)	n.a	n.a

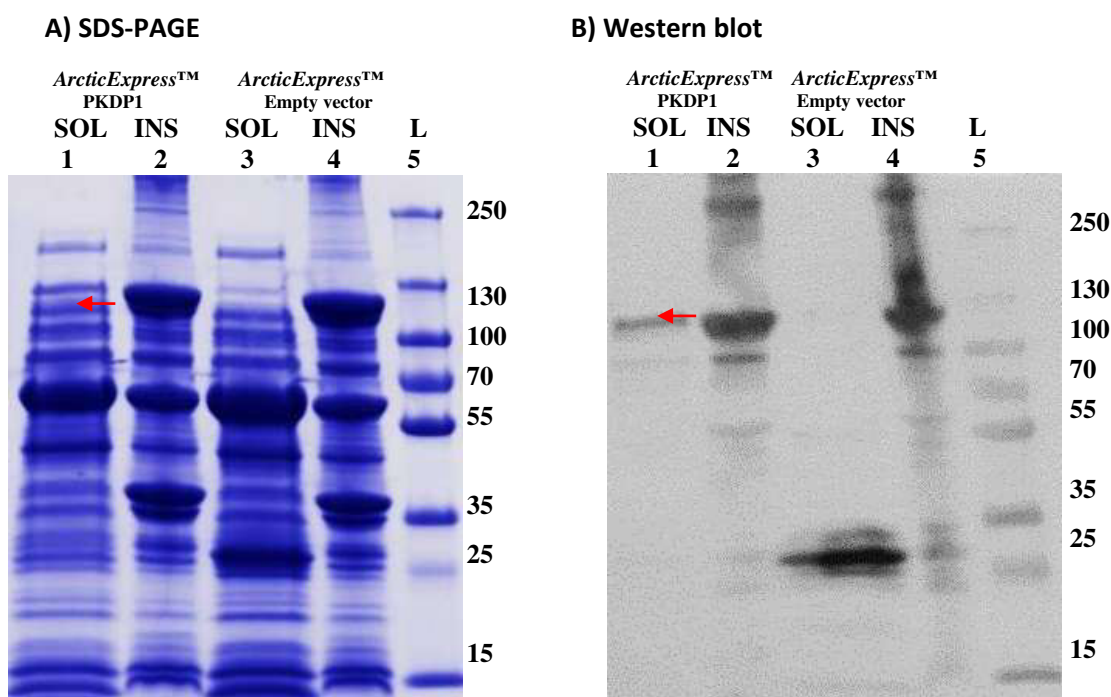
**Figure 5.2: Protein gel analysis of expressed PKDP1 following induction in SHuffle®T7.**

(A) Soluble protein fraction on SDS-PAGE. (B) Insoluble protein fraction on SDS-PAGE. (C) Soluble protein fraction by anti-GST Western immunoblot assay. (D) Insoluble protein fraction on anti-GST Western immunoblot assay. Two transformants of *E. coli* SHuffle T7® cells were selected to express PKDP1 in combination of temperature and IPTG concentration. 15 µL of protein samples were loaded onto the gel.

### 5.3.4 Larger Scale Expression and Purification of *C. fimi* PKDP1 using *ArcticExpress*<sup>™</sup> cells

#### 5.3.4.1 Result of larger protein expression

A larger culture was set-up to express and purify recombinant PKDP1 in *ArcticExpress*<sup>™</sup> cells. To confirm the expression of PKDP1, the protein samples were analysed by SDS-PAGE and Western blot as described below. A full-length of GST-fused PKDP1 was expressed in the culture as indicated at ~111 kDa in the Figure 5.13(A). The solubility of protein samples was confirmed by anti-GST immune-detection Western blot showed in Figure 5.13(B). This protein was used for the purification step presented in the next section.

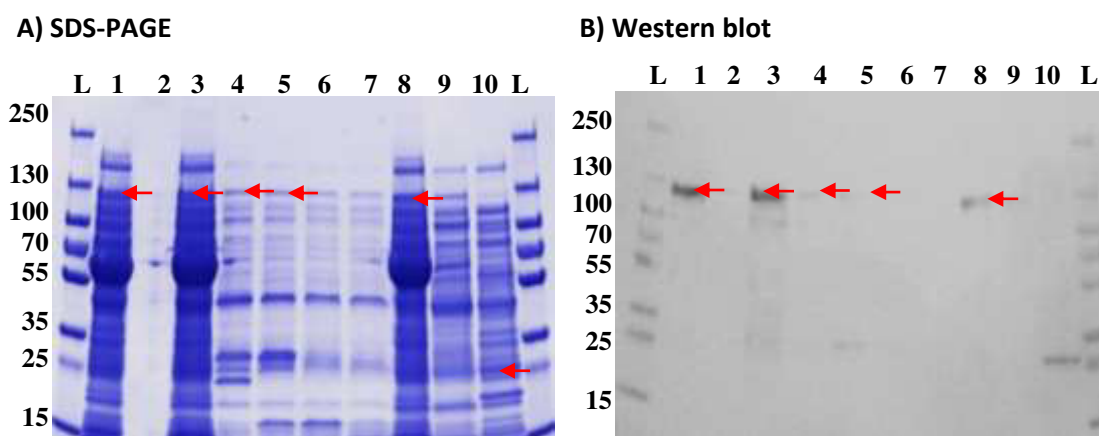


**Figure 5.1: Protein gel analysis of expressed PKDP1 following induction in *E. coli ArcticExpress*<sup>™</sup> from 500 mL culture.**

The PKDP1 expression was induced using 1.0 mM of IPTG in pre-cooled shaker at 13°C. A) Soluble and insoluble protein fractions on SDS-PAGE. B) The corresponding Western Blot. Lane 1 and lane 2, Soluble (SOL) and insoluble (INS) fractions of expressed PKDP1 by *ArcticExpress*<sup>™</sup>, respectively. Lane 3 and lane 4, Soluble (SOL) and insoluble (INS) fractions of expressed pETFFP\_3 empty vector by *ArcticExpress*<sup>™</sup>, respectively. Lane 5, Protein Ladder (ThermoScientific Pre-stained PageRuler S26619). 15  $\mu$ L of protein samples were loaded onto the gel.

### 5.3.4.2 Result of protein purification by affinity column chromatography

Since the PKDP1 was expressed with both oligo-histidine and GST affinity tags, both types of affinity columns were used in the trial to purify the expressed protein. Three attempts were made to purify recombinant PKDP1 using glutathione affinity chromatography, making use of the GST-tag. In all three cases, the recombinant protein failed to bind and was found in the flow through but not the glutathione eluted fractions, as confirmed by SDS-PAGE and Western analysis shown in Figure 5.14.



Legend of figures:

Lane	Fraction	Step	Column
1	Flow through from first purification column	Wash	1) GSTrap™ FF 5 mL
2	Elution from second purification column	Elute	2) GSTrap™ FF 1 mL
3	Flow through from second purification column	Wash	"
4	1 <sup>st</sup> elution from third purification column	Elute	3) HisTrap™ HP 1 mL
5	2 <sup>nd</sup> elution from third purification column	Elute	"
6	3 <sup>rd</sup> elution from third purification column	Elute	"
7	4 <sup>th</sup> fraction from third purification column	Elute	"
8	Flow through from third purification column	Wash	"
9	Wash residue from third purification column	Final wash	"
10	Positive control of GST-tagged protein supernatant	-	n.a

**Figure 5.2: Protein gels analysis of affinity chromatography purification fractions of expressed PKDP1 in *E. coli ArcticExpress*™.**

A) SDS-PAGE and B) Western blot of protein samples from the flow through, purified fractions, and wash residue obtained by affinity column purification. Three purification trials were applied to the cell free extracts of soluble protein expressed by *E. coli ArcticExpress*™ using GST-affinity column chromatography. The GSTrap™ Fast Flow 5mL CV, GSTrap™ Fast Flow 1mL CV, and HisTrap HP column 1mL CV were used for all the purification trials.

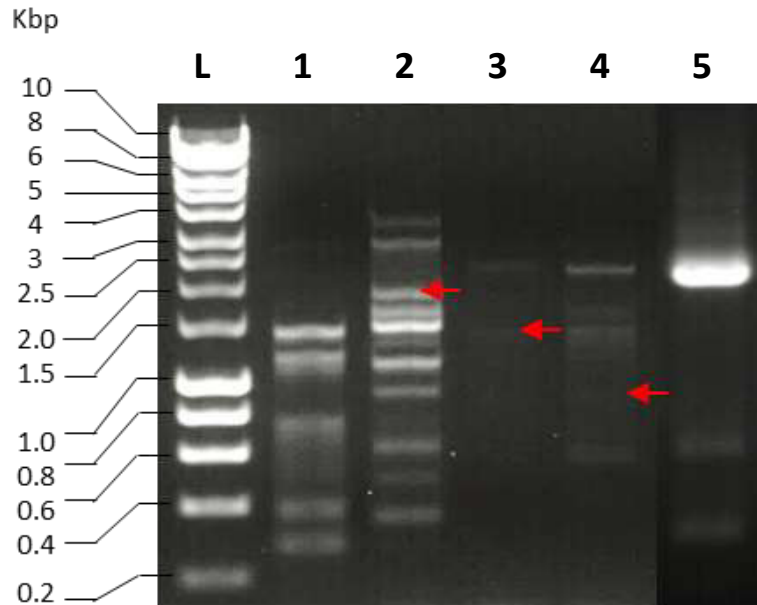
### 5.3.5 Expression of codon-optimized recombinant *C. fimi* PKDP1 in *Aspergillus niger*

A version of the PKDP1 coding sequence that had been codon optimised for *A. niger* was synthesized by GeneArt® (LifeTechnologies) and cloned into a pIGF vector (Appendix E) in order to produce the protein with an oligo-histidine (His-tag) at N-terminal for ease of purification. The codon-optimized *C. fimi* PKDP1 was cloned between XbaI/HpaI sites downstream of glucoamylase (*gla A*) open-reading frame (ORF). A Kex2 (K R G G G) cleavage site was added between the *gla A* ORF and the gene of interest by PCR to enable the protein to be cleaved in vitro and secreted into the medium.

Following PCR optimisation, a number of cloning strategies were used and the codon-optimized PKDP1 coding sequence was successfully sub-cloned into the pIGF vector in the *E. coli*. Transformation of the cloned DNA into *Aspergillus* protoplasts protoplasts was undertaken and 63 colonies were recovered (Figure 5.15). The gene coding for the orotidine-5-phosphate decarboxylase (253) (*pyrG*, from *Aspergillus oryzae*) was used as a selection marker. After transformation of *A. niger*, cells were selected for uridine prototrophy, confirming integration of the plasmid into the chromosome. After purification of the transformants, release of the selective pressure for the integrated plasmid was achieved by propagating the clones twice on *Aspergillus* Minimal Medium (AMM) agar slants containing 10 mM uridine. The direct colony-PCR products were analysed on agarose gels to confirm the size of amplified regions (Figure 5.16).



**Figure 5.3: Colonies of *A. niger* transformed with codon-optimized PKDP1 in pIGF-pyrG vector on AMMN agar plates after 4 days of incubation at 30°C .**



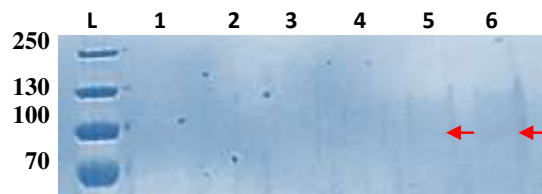
**Figure 5.1: Confirmation of pIGF-pyrG:PKDP1 integrated in *A. niger* using direct colony PCR.**

All PCR reactions showed unspecific amplified bands apart from the positive control although 3% of DMSO had been added. Lane 1, positive PCR reaction using forward and reverse primer to obtain the full size of pIGF-pyrG plasmid vector with expected size ~12 kb. However, the result on the gel shows off-target/unspecific binding of the primers during the PCR on this plasmid. Lane 2, PCR product using the first forward internal primer and reverse pIGF-pyrG primer (expected size ~1900 kb). Lane 3, PCR product using the second forward internal primer and reverse pIGF-pyrG primer (expected size ~1500 kb). Lane 4, PCR product using the third forward internal primer and reverse pIGF-pyrG primer (expected size ~900 kb). Lane 5, PCR product of positive control of wild type PKDP1 gene (expected size ~2.5 kb).

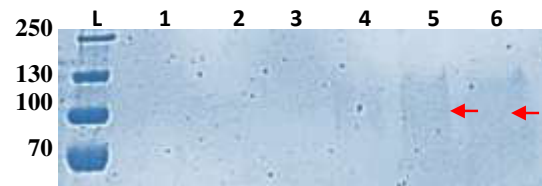


*A. niger* carrying the transgene was grown in expression medium for 6 days at 3 separate incubation temperatures (20°C, 25°C, and 30°C). Figure 5.17 shows protein samples analysed by SDS-PAGE gel from day-1 to day-6 of the cultures. Lanes 5 and 6 represented the expressed protein from day-5 and day-6 cultures showing some faint protein smears at a range of 70-130 kDa. To conclude this work, the protein samples were concentrated and analysed by SDS-PAGE (not shown). The protein band at approximate size between 70-100 kDa was cut out for protein identification by trypsinolysis and mass spectrometry. The result from protein identification matched to glucoamylase and succinyl-CoA ligase [ADP-forming]  $\beta$ -subunit protein which belongs to *A. niger*.

**A) 20°C**



**B) 25°C**



**C) 30°C**



**Figure 5.2: SDS-PAGE gels loaded with unconcentrated supernatant from cultures of *A. niger* expressing recombinant PKDP1.**

15  $\mu$ L aliquots of culture supernatant were loaded onto the gel for 6-day time course's on  $\frac{1}{2}$  strength expression media. A) Six protein samples from the culture representing each day, grown at 20°C. B) Six protein samples from the culture representing each day, grown at 25°C. C) Six protein samples from the culture of each day grown at 30°C. Lane 1, culture from day-1. Lane-2, culture from day-2. Lane-3, culture from day-3. Lane-4, culture from day-4. Lane-5, culture from day-5. Lane-6, culture from day-6.

### 5.3.6 Solubility of PKDP1 protein

PKDP1 was first identified in the *C. fimi* proteomic study which was based on soluble/extracellular protein samples. However, based on the literature of characterized glucose dehydrogenases (GDHs), GDHs exist in two quite different types which are a membrane-bound (m-GDH) and soluble-glucose dehydrogenase (s-GDH) (232). A couple of m-GDH have been characterized such as D-glucose dehydrogenase of *Gluconobacter suboxydans* (254) and pyrroloquinoline quinone (PQQ)-containing quinoproteins of *E. coli* (255) whereas, a s-GDH domain that contains a pivotal PQQ-containing quinoproteins that has been well described in *Acinetobacter calcoaceticus* (232,256).

Primary information from SignalP (an online tool for protein signal peptide prediction) and UniProt database (<http://www.uniprot.org/uniprot/F4H2Q6>) shows that *C. fimi* PKDP1 protein is predicted to have a signal peptide at 1 to 36 amino acids sequence, but, the prediction in UniProt is based on the UniProtKB automatic annotation system without a manual validation. Therefore, further analysis has been carried out on the peptides sequence of PKDP1 to investigate protein characteristics including protein subcellular localization (PSORTb), transmembrane protein prediction (THMM), protein hydrophobicity plot (using ProtParam) and protein hydrophobicity scale (ProtScale) provided by Swiss Institute Bioinformatics Expert Protein Analysis System Resources Portal (SIB ExpASY).

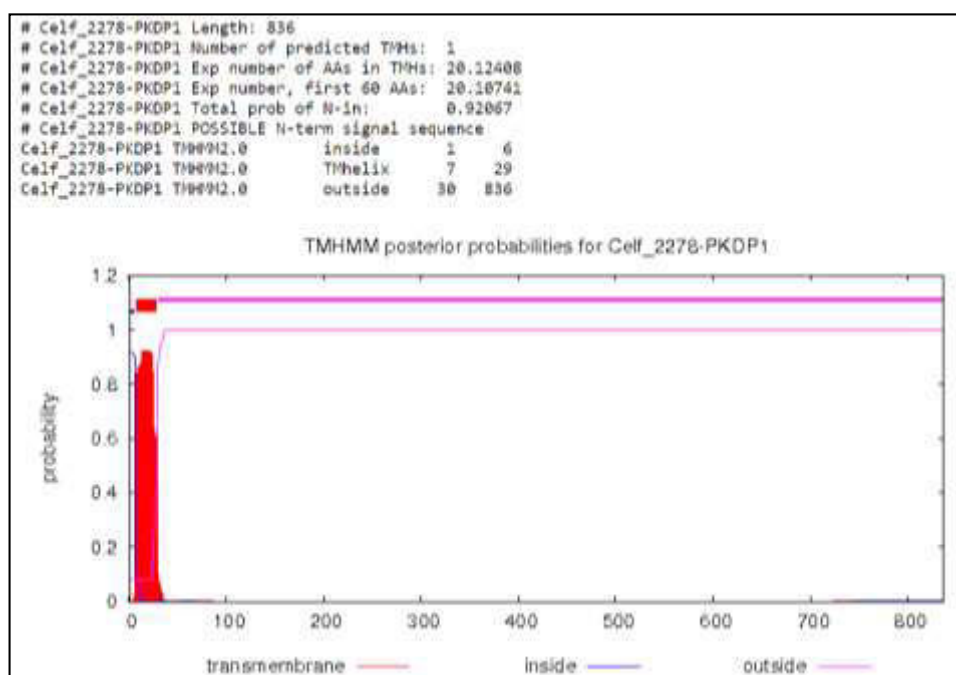
A prediction of *C. fimi* PKDP1 subcellular localization has been performed using the latest protein subcellular localizations (SCLs) database available online in PSORTdb (<http://db.psort.org/browse/genome?id=9904>). The database comprises 2 types of manually curated SCLs for proteins which have been experimentally verified (ePSORTdb), as well as pre-computed SCL predictions for deduced proteomes from bacterial and archaeal complete genomes that available from NCBI (cPSORTdb). Table 5.5 shows the summary of pre-computed protein localization scores (0 to 10, where 10 is the best prediction) of *C. fimi* PKDP1 from the both curations in PSORTdb database. The final score of PKDP1 localization is 7.21 that consequents the unknown status for prediction of location indicated that PKDP1 may having multiple location of its secretion.



**Table 5.1: Pre-computed score of *C. fimi* PKDP1 protein localization.**

Localization	Cytoplasmic membrane	Cell wall	Extracellular	Cytoplasmic	Final score
Score	7.21	1.45	1.34	0	7.21

The analysis has been further carried out by a protein transmembrane prediction using the Trans-Membrane Hidden Markov model (TMHMM) algorithm online tool available at (<http://www.cbs.dtu.dk/services/TMHMM>) (257). The TMHMM result suggests a possibility of PKDP1 having a Trans-Membrane Helix (TMH) in the N-terminal region at 7 to 29 amino acids (see Figure 5.18). However, a high score (0.92067) of N-best (the best score is 1) predicted that the first 30 amino acids of PKDP1 could be a signal peptide. The N-best algorithm is the sum over all paths through the model with the same location and direction of the helices of protein domain architecture.



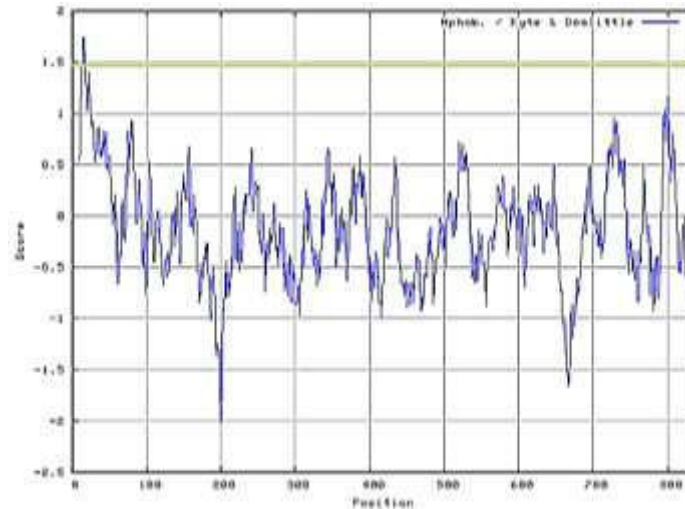
**Figure 5.3: Transmembrane prediction of PKDP1 using membrane protein topology prediction based on Transmembrane hidden Markov model (TMHMM) prediction server.**

The result shows a probability of a signal peptide at 1 to 6 amino acids, and a transmembrane helix located at 7 to 29 amino acids for PKDP1. The rest of the sequence is at the outside membrane (30 to 836 amino acids sequence). The expected number of amino acids in the predicted TMH is 20.125.

The score is larger than 18 and it is very likely to be a transmembrane protein or have a signal peptide. The expected number of amino acids in transmembrane helix in the first 60 amino acids of the protein is high but this result needs to be interpreted cautiously because of the possibility that TMH could be a signal peptide. Finally, the total probability of N-in described in the result is the total probability that the N-terminal is on the cytoplasmic side of the cell's membrane.

ProtParam (<http://web.expasy.org/cgi-bin/protparam/protparam>) computes various physico-chemical properties that can be deduced from a protein sequence including Grand Average of Hydropathy (GRAVY) (258). The GRAVY number of a protein is a measure of its hydrophobicity or hydrophilicity. The two measures are combined in a hydropathy scale or hydropathy index (259). Hydrophathy is a characteristic of a protein and will have a relatively high hydrophaty index if the protein is particularly hydrophobic (more tendency to water) or hydrophilic (tendency of nonpolar substance to aggregate in aqueous solution and exclude water molecules). The GRAVY value for a peptide or protein is calculated as the sum of hydropathy value (259) of all amino acids, divided by the number of residues in the sequence. In essence, a GRAVY score is the relative value for the hydrophobic residues of the protein. The hydropathy values range from -2 to +2 for most proteins, with the positively rated proteins being more hydrophobic (260,261). From the analysis, the GRAVY score for PKDP1 is -0.142. The value implies a low probability of PKDP1 to have much hydrophilic region as the more positive the value the more hydrophobic are the amino acids in the protein. Hydrophobic proteins are likely to be membrane bound, and therefore difficult to solubilize.

Finally, protein scale analysis estimates the hydrophobic regions profile produced by amino acid scale and using the proteomic tool available at <http://web.expasy.org/protscale/>. In this analysis, a PKDP1 hydrophobic plot was computed using a widely used and predefined scale by Kyte and Doolittle (259) for the detection of hydrophobic regions in proteins with a positive value. Several window size options for the analysis are available where Transmembrane domain characteristics could be identified with a window size of 19 for a calculated of hydrophobic value of just above 1.6. Figure 5.19 shows PDKP1 hydrophobicity plot where the standard cut-off scale for prediction of hydrophobic region is above 1.5. The hydrophobicity plot shows a correlation with previous TMHMM prediction where one short region of the amino acids at the N-terminal predicted contains several hydrophobic amino acids.



**Figure 5.4: The Kyte and Doolittle (259) hydrophobicity plot of PKDP1 from *C. fimi*.**

A window size of 19 was used to compute the plot for a total of 836 amino acids of *C. fimi* PKDP1. Most of the protein's peptide region is predicted to be non-hydrophobic with a standard cut-off score positive value of below 1.5.

In summary, from four analyses that were performed on the PKDP1 protein sequence, including the prediction of signal peptide (SignalP and UniProtKB), subcellular protein localization (PSORTb), protein transmembrane prediction (TMHMM), and protein hydrophobicity analyses (ProtParam for GRAVY score and ProtScale for hydrophobicity plot), it was postulated that PKDP1 may have multiple location following of its secretion. PKDP1 has been predicted to have a signal peptide at its N-terminal at 1-36 amino acids, however, from TMHMM analysis indicated that the stretch of signal peptide in PKDP1 may also act as a transmembrane region of the protein. Since the prediction of subcellular localization of PKDP1 remained to be unknown with a higher possibility to be in cytoplasmic membrane rather than secreted extracellularly, hydrophobicity property of PKDP1 has been examined. From ProtParam and ProtScale analyses, both results are in agreement with TMHMM and subcellular localization online database searches where the signal peptide of PKDP1 does not guarantee that this protein is freely secreted from the cell, as the hydrophathy measurement score of the PKDP1 suggests the existence of hydrophobic region in the protein sequence. Therefore, it could be postulated that PKDP1 protein may have multiple localization after its secretion from the cell.

## 5.4 DISCUSSION

The potential role of PKDP1 (Celf\_2278) in the breakdown of polysaccharides has been investigated by bioinformatics means and requires confirmation through biochemical characterization. The production and purification of recombinant PKDP1 in both bacterial and fungal expression systems, and the isolation of PKDP1 in its native form from *C. fimi* has been attempted.

### 5.4.1 Attempted isolation of native PKDP1 from *C. fimi* culture

Avicel is crystalline cellulose that consists of polymers of pure glucose. Avicel was chosen as one of the inducing substrates as it was shown that the PKDP1 was secreted highest by mole percentage in the Avicel culture (0.034%) when compared to three other substrates; beechwood xylan (0.007%), wheat straw (0.016%), and sugarcane bagasse (0.012%) from the proteomic work. PASC is a pretreated Avicel with 85% phosphoric acid to produce phosphoric acid swollen cellulose which results in a more amorphous (less crystalline substrate). Cellobiose is a disaccharide contains reducing sugar of two glucose molecules linked by  $\beta$ -1-4 and is the repeat disaccharide of cellulose. From the comparison of protein secretion pattern of extracellular and substrate-bound fractions induced by 3 substrates between the day-3 and day-7 of cultures, higher protein secretion was detected on the day-3 in the supernatant fraction and lower levels during day-7. The secreted proteins in both cultures showed more affinity towards the substrate at the later time point where more protein bands were observed on day-7 of substrate-bound fractions. The opposite pattern of protein secretion was seen for *C. fimi* grown on cellobiose, where relatively lower amounts of protein were secreted in both supernatant and cellobiose-bound fractions on the day-3, but increased by day-7. However, from the peptides identification of selected protein bands, none of the secreted protein was PKDP1. Another attempt to isolate native PKDP1 was using anion exchange chromatography. It is a form of ion exchange chromatography (IEX) that separates the proteins using a positively charged ion exchange resin with an affinity for the protein that having net negative surface charges. From a trial of isolation of PKDP1 using it's predicted isoelectric point (pI) value (5.65), culture of *C. fimi* induced with Avicel was tried to be purified using DEAE column, however, been unsuccessful. Attempts to isolate native PKDP1 from *C. fimi* cultures were made, but the low levels of protein present made it impractical to purify sufficient protein for further studies.

#### 5.4.2 PKDP1 expression in *Escherichia coli*

*E. coli* is a well-established, quick and cheap, simple prokaryotic expression system. It became a preferred host strain for protein expression as it generally give high yields for the majority of proteins (251). Furthermore, *E. coli* offers a range of vectors and strains for optimum protein expression. However, in some cases proteins are not expressed in a soluble form due to lack of proper protein folding (262). To overcome this challenge, a fusion protein technique has been applied using 5 solubility tags (-His, -GFP, -GST, -MBP, and -Im9), and the resulting proteins screened using dotblot technique in the TF Laboratory.

From the 96-wells plate dotblot expression screening, the most promising fusion protein construct for PKDP1 was the GST-tag. The protein was expressed in pETFPP\_3 vector with the tag located at the N-terminus followed by the PKDP1. The use of GST (26 kDa) as a fusion tag is desirable because it can act as a chaperone to facilitate protein folding, and frequently the fusion proteins can be expressed as a soluble protein rather than in inclusion bodies. The use of GST-fusion proteins also can be used to enable protein purification by affinity chromatography and use of site-specific protease recognition sequences located between the GST moiety and the target protein (263). However, despite showing full length expression of PKDP1 on the SDS-PAGE gel, the anti-GST immunoblot assay was failed to pick up any signal from the GST-fused PKDP1protein. This result could be explained because the PKD fusion-protein is very large (111 kDa) and may not have transferred efficiently during the blotting procedure. In this prokaryotic expression work, PKDP1 expression was attempted using four *E. coli* strains and varying two expression parameters; the expression temperature after induction at 10-25°C, and the concentration of IPTG (0.5-1.0 mM). The four *E. coli* strains that were used in this work are listed in Table 5.4, which summarizes their advantages in protein expression.

**Table 5.1: Summary of advantages of *E. coli* strains for protein expression**

<i>E. coli</i> strain	Description and advantage	Reference
BL21 (DE3) (Agilent®)	A good growth characteristic, among the fast growing bacteria and proven for simple culture handling. The utilization of the T7 RNA polymerase promoter to direct high-level expression is an advantage for this vector.	(264)
<i>ArcticExpress</i> ® (Agilent®)	Allows use of low-temperature cultivation strategy for the recovery of soluble protein by the help of <i>E. coli</i> chaperonins, which facilitate proper protein folding. Chaperonins bind and stabilize unfolded/partially folded proteins and lose activity at reduced temperatures.	(265,266)
<i>SHuffle</i> ™ (NEB)	Uses the advantage of a protease deficient B strain (chemically competent <i>E. coli</i> B cells) engineered to form proteins containing disulfide bonds in the cytoplasm	(251,267)
<i>SHuffle</i> ™ T7 (NEB)	A T7 expression K12 strain that has an enhanced capacity to correctly fold proteins with multiple disulfide bonds in the cytoplasm.	(268,269)

The most promising results were achieved with low temperature expression in *ArcticExpress*® (BL21DE3). The lowest tested temperature with successful soluble protein expression was at 10-15°C. Therefore, the PKDP1 gene was expressed at 13°C which was the middle point from the successful induction temperatures (10-15°C). By lowering the temperature during protein expression, the production rate of the protein was also significantly reduced, however the chaperonins present in the *ArcticExpress*® system confer improved protein processing at lower temperatures thus potentially increasing the yield of active, soluble recombinant protein.

Varying IPTG concentrations did not show any significant impact on expression levels between 0.5 or 1.0 mM, and therefore 1.0 mM IPTG was used. A larger culture for optimum expression condition was performed in a total of 1 L medium to produce enough protein for characterization work. Two affinity chromatography columns were used in the trial to purify the desired protein including GSTrap FF (GSH Sepharose) and HisTrap FF (Ni-NTA Sepharose). The Coomassie stained gels showed a full-length protein expression from the culture, and the right size of protein band (~111 kDa) was picked up by the GST-HRP antibody on western immunoblot assay suggesting a promising outcome.

Affinity chromatography is a conventional method used for protein purification where protein capture is performed through the molecular recognition between the tag and the ligand. However, in this experiment, the PKDP1 was ineffectively purified using this method. Shorter affinity peptides such as the polyhistidine tag (His-tag) used in this study is the most preferable option as they are less likely to interfere with the expression, structure and function of the target protein (270–272). Unfortunately, using this tag showed no improvement on PKDP1 protein solubility during the expression. Therefore, fusion protein approaches have been applied to aid soluble protein expression. Among five fusion proteins tested, Glutathione-S-transferase (GST)-fused to the PKDP1 protein generated soluble protein although GST fusion is considered as a large co-expressed protein that copurifies with the protein of interest. GST is a monomeric protein of 26 kDa (273,274). It acts as the ligand and also GST tag which acts both as a purification anchoring point as well as a stability/solubility enhancer (275,276). The elution condition for GST-fused/tag protein is mild because reduced glutathione can be employed as a competitive agent (274). However, if reducing conditions are compromised, the fusion protein can undergo oxidation aggregation due to the presence of four cysteine residues exposed at the surface of GST tag (277).

Consequently, fusion proteins with higher molecular weights than 100 kDa can lead to partially insoluble proteins (275) which mostly occurred in this work. As the diversity of proteins and their biochemical properties make the universal purification of proteins a challenge with most protein of interest usually lacking a suitable, specific and robust affinity ligand for capture on a solid matrix (278). Alternatively, the recent development of protein purification approach gives an option for the purification of fusion proteins which is based on inverse transition cycling (ITC). ITC which exploits the reversible soluble–insoluble phase transition behavior of the affinity tags and the desired fusion protein yield can be achieved with  $n$  ITC rounds (278).

### 5.4.3 PKDP1 expression in *Aspergillus niger*

*A. niger* is a fungal strain that has high protein secretion capacity even for large proteins (279). *C. fimi* PKDP1 is predicted to have at least 5 disulphide bridges according to protein sequence analysis on DiANNA, an online tool that designed to work with Neural Network based on cysteine pairs scoring with 81% accuracy (252). The *A. niger* expression system offers an ability to correctly fold secreted proteins containing disulphide bridges, and for post-translational modification that may be needed before protein folding which helps for the solubility of the protein. The lack of success in getting PKDP1 expressed in *Aspergillus* was disappointing.

*Aspergillus* sp. are recognized for their ability to secrete correctly folded extracellular proteins, but recent experience in the laboratory of Professors McQueen-Mason and Bruce indicate that the secretion of active proteins is not always achieved (280) as appears to be the case with PKDP1. *A. niger* is known to possess many endogenous proteases that lead to degradation of secreted proteins. However, the strain of *A. niger* used in these experiments has had most of these proteins deleted and this is unlikely to be the cause of the lack of PKDP1 expression.

In conclusion, both native isolation and recombinant approaches have failed to bulk up PKDP1 production for protein characterization, requiring further improvement to achieve the objective. The native isolation method could be enhanced by the addition of artificial electron acceptors that act as small molecule reductants (281) to reoxidize the cofactor (282). This strategy has been successfully applied during characterization of chitin-binding protein 21 (CBP21) in a chitin deacetylase from *Aspergillus nidulans* (141) to increase the solubility of longer chitin fragments by deacetylation. The presence of a reductant of ascorbic acid dramatically increased the efficiency of the reaction which enabled the breakdown of large crystalline  $\beta$ -chitin particles by CBP21 alone and the release of a range of oxidized products (141).

This is because the predicted glucose dehydrogenase and pyrroluloquinone-quinone domain (GDH-PQQ) in the PKDP1 may acts as an oxidoreductase that possibly has some redox enzyme-based electron systems (232). This type of protein is unable to utilize oxygen as the electron acceptor and instead transfer the electrons to various natural and artificial electron cofactors (283). Since the PKDP1 redox cofactors are still unknown and the need of cofactors by most of CAZy of *C. fimi* is not a major requirement (86), several artificial cofactors could be tested in the medium such as the ions



(Ca<sup>2+</sup>, Cu<sup>2+</sup>), ascorbate, sulfur-containing species, gallic acid, or pyrogallol (142,239). Recombinant expression particularly of PKDP1 in both bacterial and eukaryote systems has proven to be difficult and several key challenges remain to be overcome.

For the recombinant protein work, the predicted activity of PKDP1 as a putative  $\beta$ -glucosidase with two other interesting predicted domains is an advantage. The size of full-sequence multimodular PKDP1 protein is considered to be large (~86 kDa) compared to other commonly secreted recombinant proteins which are mostly in the range of 20-30 kDa. Additionally, many general limitations of *E. coli* as an organism for recombinant protein expression are well known including low production levels of soluble proteins and poor secretion ability (284). Therefore, it may be beneficial to express the individual domains of PKDP1 separately to study their activity. This is because it is generally easier to express a relatively smaller protein in any expression host (285). As an example from the literature, Li and the co-workers (286) have successfully expressed an active recombinant protein in *Pischia pastoris* of cellobiose hydrolase 3 (CBH 3) by its single domain separated from the other existing domains from *Chaetaomium thermophilum*. This suggests that a better expression rate can be achieved when working with protein which only have single catalytic domains, instead the more frequent three-domain proteins (comprised of a catalytic domain, linker and carbohydrate-binding domain) (286).

A gene deletion approach could be used to characterise the ability of *C. fimi* to digest or grow on cellulose. Clustered regularly interspaced short palindromic repeats (CRISPR) are an efficient gene editing technique that could be implemented in *C. fimi* to delete the PKDP1 gene to establish a new strain. CRISPR-Cas9 was first established for use in eukaryotic systems (287–289) however, more recently, many initiatives to develop a method particularly for prokaryotic genes have been initiated (290,291) and present a promising approach to validate the molecular and metabolic functions of specific genes in bacterial systems.

## 6 Adaptive Evolution of *Cellulomonas fimi* by Continuous Subculture on Wheat Straw

### 6.1 INTRODUCTION

In nature, lignocellulose deconstruction is mostly accomplished by heterotrophic fungi and bacteria (27,61,292). With a few exceptions, most animals that degrade cellulose accomplish the tasks with the assistance of microbes resident in their digestive tracts (293). The recalcitrance nature of lignocellulosic biomass is still the bottleneck of modern conversion in second generation biofuel processes (160,294). Therefore, both industrial and academic researchers have attempted to develop more economic routes for biofuel production over the past decades (61,295–297). In the industrial process, the recalcitrance of lignocellulose biomass is overcome by extreme chemical and physical pre-treatments, which break open the complex lignocellulosic structure so that it is more accessible to enzymatic digestion. However, these pretreatment steps also generate potential inhibitors of microbe's metabolism such as furan and furfural (297–299).

A number of soil bacteria have been identified that are able to oxidise lignin. *Cellulomonas fimi* is a well-studied cellulolytic actinomycete and mesophilic bacterium found in soil and is well known for its cellulolytic capability (77,84,175,192,300). The full sequence of *C. fimi*'s genome has been published recently revealing a wide range of putative and previously characterised Carbohydrate Active Enzymes (CAZy) (86). A wide range of *Cellulomonas* species have been found to have common occurrence in the various environments where cellulose decomposition occurs (112,301).

Adaptive evolution studies generally investigate the adaptation of microorganisms in the presence of a restraint over a large number of generations in order to observe evolutionary processes in real time. Adaptive evolution experiments have been carried out by researchers involving a range of target metabolisms in single species, and coevolution in bacteria (139,302–306) as well as in fungi (307,308).

A recent study has shown a successful adaptive evolution outcome, by generating a *Streptomyces* mutant strain which degrades cellulose using filter paper as a substrate (33). Yet, there has been no research to determine if any strains of bacteria including *C. fimi* have

the improved capability to digest plant cell wall components in complex lignocellulosic materials, such as wheat straw. Therefore, a research study was designed to investigate the growth of *C. fimi* on wheat straw in order to examine what adaptations might arise in response to this challenging substrate.

In this study, an adaptive evolution approach has been applied to the wild type *C. fimi* ATCC 484™ strain by periodic serial transfer into a fresh basal medium with untreated chopped wheat straw as a sole carbon source. This was done by subculturing 6 populations of *C. fimi* every 7 days for 52-weeks. Comparison analyses between the wild type and 6 populations of adapted strains were performed by the measurement of the growth rate, carbon dioxide evolution (CER), cells association to the biomass by gDNA and protein extractions, biomass sugar analysis, and quantification of residual biomass after degradation by adapted strains.

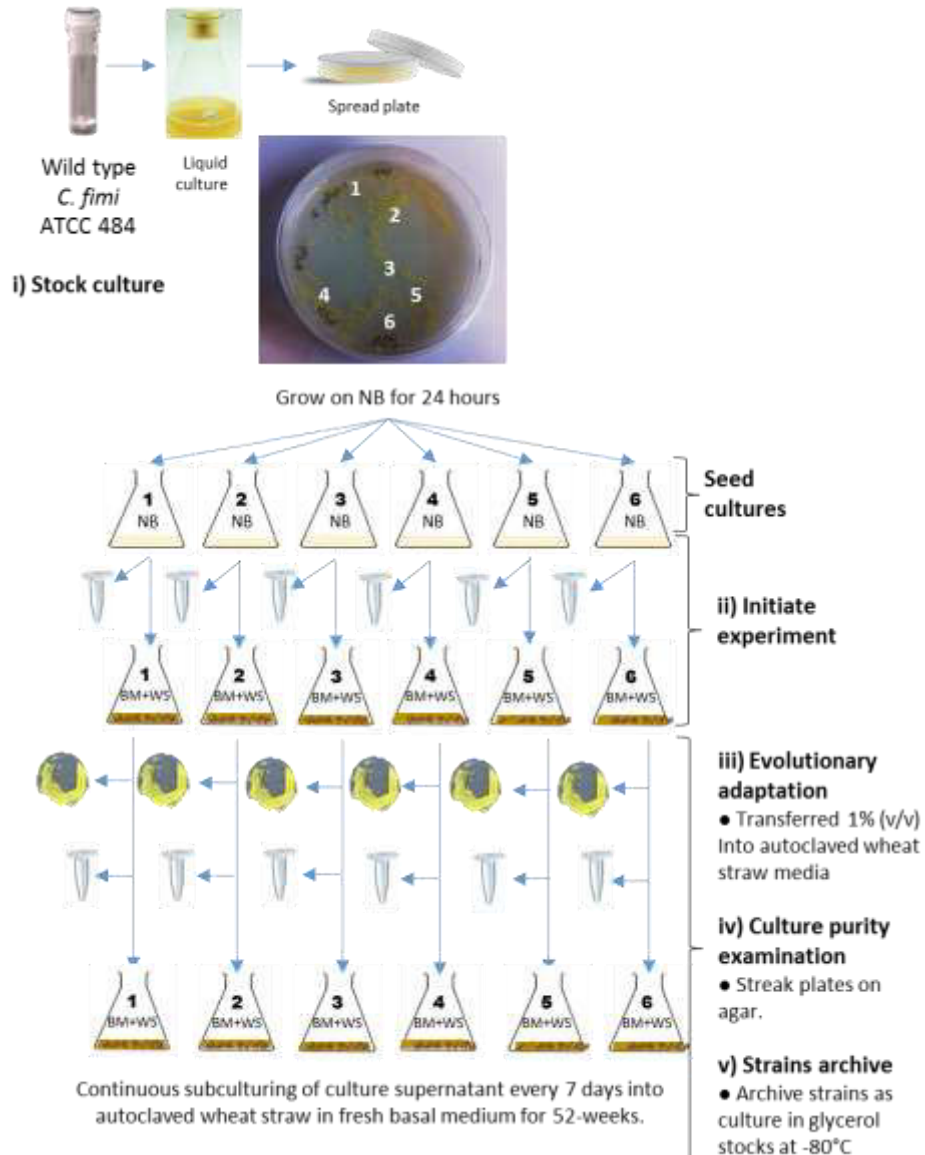
## **6.2 MATERIALS AND METHODS**

### **6.2.1 Bacterial strain, media, and cultivation**

The strains used in this experiment were the ATCC 484 (wild type) and 6 strains *Cellulomonas fimi* derived from a 52-weeks (one year) continuous adaptive evolution experiment as described in the following section. Double-autoclaved chopped wheat straw, 2% (w/v), supplemented with basal medium at pH 7, was used in all experiments. The composition of basal medium is described as in Section 2.2.2. All cultures were incubated at 30°C. Continuous subculture was carried out using 250 mL flasks with a working volume of 50 mL, and incubated on a rotary shaker at 200 rpm (Landert Motoren AG, Switzerland).

### **6.2.2 Strain construction**

The wild type of *Cellulomonas fimi* ATCC 484™ (American Type Culture Collection, USA) was grown using  $1 \times 10^{-5}$  diluted stock culture spreaded on Oxoid™ nutrient agar (ThermoFischer Scientific) at room temperature for 5 days to obtain a single colony. The pure colony was isolated and grown in liquid nutrient broth for 7 days allowing the bacterial growth to reach to early stationary phase. The generation of lignocellulose-adapted *C. fimi* strains was accomplished by 52-weeks serial transfer. The stages of strain construction are illustrated in Figure 6.1.



**Figure 6.1: Experimental design for adaptive evolution.**

i) A stock of wild strain was grown up in nutrient broth liquid culture before being streaked on an agar plate to obtain 6 pure single colonies. ii) Each pure colony was grown up in separate liquid culture. The experiment was initiated by transferring 1% (v/v) of cell biomass into 2% (w/v) autoclaved wheat straw in basal media. iii) Evolutionary adaptation experiments were started by transferring 1% (v/v) of each culture supernatant once weekly into fresh media for 52-weeks. iv) The purity of the grown cultures was checked on agar plates by streak plate technique in weekly basis. 16S genomic sequencing was carried out using universal prokaryotic 16S rRNA primer sequence and the whole *C. fimi* gDNA as a template by *LightRun*<sup>®</sup> sanger sequencing (GATC, Germany) to check culture integrity in quarterly period of time. v) Archives of strains were prepared in glycerol stock and kept at -80°C freezer for long-term storage.

## 6.2.3 Analytical methods

### 6.2.3.1 Cell growth profiles and relative growth rate

Cell growth profiles were determined from readings of optical density (OD) at 600 nm using Cary-50Bio UV-Visible Spectrophotometer (Agilent, USA) and counts of colony forming units (CFU). The number of CFU was determined in triplicate by plating 50  $\mu$ L of serially diluted ( $1 \times 10^{-5}$ ) culture supernatant on NB agar, followed by 5-day incubation at 30°C. This step was repeated using a 4-day time course, and the relative growth rate of each strain was calculated according to Dalgaard *et al.* (309).

### 6.2.3.2 Carbon dioxide evolution (CER)

Carbon dioxide measurements were determined in a 4-day time course by measuring the pressure at time point 0 ( $T_0$ ) and time point 1 ( $T_1$ ) (the initial and the end of a temporary 2 h anaerobic condition incubation, respectively) using a digital pressure gauge (DG1, Rototherm). Anaerobic conditions were achieved for 2 h by replacing the sponge bungs on the wheaton™ flasks with rubber bungs. CO<sub>2</sub> released during the anaerobic period was collected into exetainer® bottles. The gas samples were measured by gas chromatography (GC) using an HP-5 column equipped with Flame Ionization Detector (FID) and quantified using standard carbon dioxide.

### 6.2.3.3 Estimation of growth and bacterial residency by genomic DNA of *C. fimi* strains

Genomic DNA (gDNA) was extracted using a phenol-chloroform method (310) and purified using Genomic DNA Clean & Concentrator™-25 (D4064, ZymoResearch USA) according to the manufacturer's instructions. The insoluble wheat straw in the culture was separated from liquid supernatant through a filtration process using 200  $\mu$ m nylon mesh. A wheat straw biomass fraction of 0.5 g and liquid supernatant of 35 mL were collected into separate 50 mL tubes after the filtration process. The supernatant fraction was spun down by centrifugation at 4000 x g for 10 min to obtain the bacterial cell pellet. The cell pellet and biomass fractions were incubated at room temperature for 30 min at 37°C in 5 mL of 100  $\mu$ g/mL lysozyme in Tris-EDTA (TE) buffer pH 7 for cell lysis. Following cell lysis, the sample was spun for 10 min at 4000 x g to achieve separation of the phases, before the aqueous layer was removed to a

fresh 15 mL centrifuge tube. To the aqueous phase, chloroform:isoamyl alcohol (21:1) was added and this was spun and the aqueous phase transferred to a fresh tube. Precipitation of gDNA was obtained by adding an equal volume of ice-cold 100% isopropanol and incubated for 1 h. After centrifugation at 4000 rpm for 10 min, the resulting supernatant was discarded without disturbing the gDNA-containing pellet. The pellet was then washed with 80% cold ethanol and air-dried before being resuspended in Ambion® DNase free water (ThermoFisher). The quantification results obtained with nanodrop™ were compared with band intensities of high molecular weight gDNA visualized on ethidium bromide-stained, 1% (w/v) TAE agarose gels. Ten percent of the total corresponding eluate volumes were loaded onto the gels.

#### **6.2.3.4 Estimation of growth and bacterial residency by total protein**

Total protein content (expressed free protein, protein-bound to the biomass and intracellular proteins) of the culture was used as an indicator of growth on insoluble wheat straw material. Total protein was extracted in 50 mL centrifuge tubes by boiling 500 mg freeze-dried wheat straw in 10 mL of 0.2% SDS for 5 min to lyse all cells bound to the biomass. The protein supernatant was collected into a fresh 50 mL centrifuge tube by centrifugation at 4000 x g; this step was repeated two times without heating but with vigorous vortexing between each centrifugation step to wash the biomass and remove protein. Extracted protein was precipitated with five volumes of ice-cold acetone and incubated overnight at -20°C before centrifugation at 4000 x g. The resulting protein pellet was washed with 80% ice-cold ethanol. The ethanol-protein solution was then centrifuged again at 4000 x g after which the supernatant was discarded and the protein pellet was air-dried before being solubilized in 5 mL of ELGA ultrapure water. Protein concentration was quantified using the Bradford assay (Section 2.4.1) and SDS-PAGE (Section 2.4.2), respectively.

#### **6.2.4 Scanning Electron Microscopy (SEM) analysis of wheat straw samples**

The biomass-bound cells associated with the wheat straw were observed using a conventional SEM method (53). The isolated specimens of wheat straw fibres were obtained by filtration of wheat straw culture through 200 µm nylon mesh to separate the liquid and straw particles. Approximately, 300 µL level of wheat straw fibres in 1.5 mL Eppendorf tube were fixed with 2.5% glutaraldehyde in 100 mM phosphate buffer (pH 7.0) at 4°C. The fibre specimens were incubated on a rotator for 1 h before been immersed overnight at 4°C. After fixation, fibers were washed with phosphate buffer (3 x 20 min), followed by post fixation with 1% osmium tetroxide in phosphate buffer for 30 min to 1 h on ice. The fibers were washed again with phosphate buffer (3 x 20 min) before dehydration through a graded ethanol series (25, 50, 70, and 90%) once for each concentration for 20 min, and 3 times for 10 min using 100% of ethanol. Dehydrated fibers were washed once with 100% tetrabutyl alcohol and dried by vacuum freeze drying. Hexamethyldisilazane (HMDS) was used in the critical-point of drying step to remove the excess of ethanol and allow the specimens to air dry in a dry atmosphere. Prepared specimens were then mounted on SEM stubs and sputter-coated with gold plasma coater (SC-7640 Auto/Manual High Resolution Sputter Coater, Quorum Technologies Ltd.). Samples were viewed using a JSM-7600F SEM operating at 15 kV (JEOL Ltd., Tokyo, Japan). The preparation of specimens and SEM visualization was undertaken by Ms Joanne Marrison, Dr Anna Simon and Dr Karen Hodgkinson in the Imaging and Cytometry Laboratory, Department of Biology, University of York.



### 6.2.5 Enzyme activity assays in culture supernatant

The ability of enzymes to cleave polysaccharides and produce break down products with reducing ends was assessed according to Lever assay (163) that used p-hydrobenzoic acid (PAHBAH) solutions. PAHBAH reagent was prepared freshly on day of use by adding 0.761 g into 100 mL of 0.5 M NaOH. Preparation of sugars standard was done by pipetting into 1.5 mL microtubes 0, 10, 20, 30, 40, and 50  $\mu$ L of the 1 mg/mL standard of 1% glucose or xylose solutions, then each of the tube was made up to a final 50  $\mu$ L volume with distilled water. Next, 950  $\mu$ L of 50 mM pH 6.5 sodium phosphate buffer was added for a total volume of 1 mL. These standards mixture of tubes were not incubated at 37°C.

96-well plates were used by pipetting 150  $\mu$ L of PAHBAH mix into each well, including the standard sugars wells. 100  $\mu$ L of appropriate 1% substrate solution was added into fresh 1.5 mL microtubes in duplicate (for  $T_0$  and  $T_1$  measurements) for each sample. 50  $\mu$ L of enzyme preparation (culture supernatant) was added into the substrate solution and mixed by pipetting up and down, then immediately, 5  $\mu$ L of enzyme-substrate mixture was transferred into 150  $\mu$ L PAHBAH solution that been prepared earlier. This reaction gives an absorbance value for time 0 min ( $T_0$ ) as the negative control. Then, another reaction mixture was placed into 37°C waterbath, and incubated for 1 h. After 1 h of incubation, 100  $\mu$ L of incubated enzyme-substrate mixtures were transferred into 150  $\mu$ L PAHBAH mix. 5  $\mu$ L of each diluted standard was transferred into 150  $\mu$ L PAHBAH.

All samples and standards were then incubated at 70°C for 10 min to terminate the reaction. To read the absorbance after the enzymatic reaction, 150  $\mu$ L of sample was transferred into fresh microtitre plates and the absorbance was read at 405 nm Tecan Safire2 (Thermo Fisher) plate reader (163). From the glucose and xylose standard curves, the concentration of reducing sugars was calculated.

### **6.2.6 Compositional analysis of oligosaccharide from wheat straw degradation**

Sugars profiles after wheat straw degradation by *C. fimi* strains were determined by high performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) (DIONEX ICS - 3000, UK) using a CarboPac PA20 column with a 50 mM NaOH isocratic system and flow rate of 0.5 ml min<sup>-1</sup> at 30°C (27), after sequential hydrolysis using trifluoroacetic acid (TFA) following an alcohol precipitation (311). Five types of sugars including glucose, xylose, arabinose, galactose, rhamnose and three uronic acids; galacturonic acid, glucuronic acid and ferrulic acid were used as standards with mannitol as an internal standard. A standard curve was determined for each sugar and the trend line equation was used to calculate the amount of sugar content in g/L for each straw sample. The best fit line was calculated using Excel software. The linear equation obtained from the standard curve was used to find out the concentration of sample glucose in g/L.

### **6.2.7 Quantification of lignocellulose biomass residual after degradation**

The degradation of wheat straw by *C. fimi* (wild type and 6 adapted strains) in liquid culture was determined using a mass balance method after 10 days of culture incubation. Three biological replicates of wheat straw cultures from each strain were weighed before and after the 10-days culture incubation. Biomass samples were collected by filtration through 200 µm nylon mesh and washed with 40 mL of water twice, followed by three washes with 40 mL of 100% ethanol. The washed biomass samples were dried at 60°C for 2 days and then weighed. Results were compared against the negative control represented by wheat straw culture that had not been inoculated with the bacteria.

### **6.2.8 Statistical analysis**

The mean of 3 biological replicates represents the standard deviation (SD) in each quantitative analysis. Statistical analysis was performed using one-way ANOVA multiple comparisons test with significant value of  $P \leq 0.05$ .

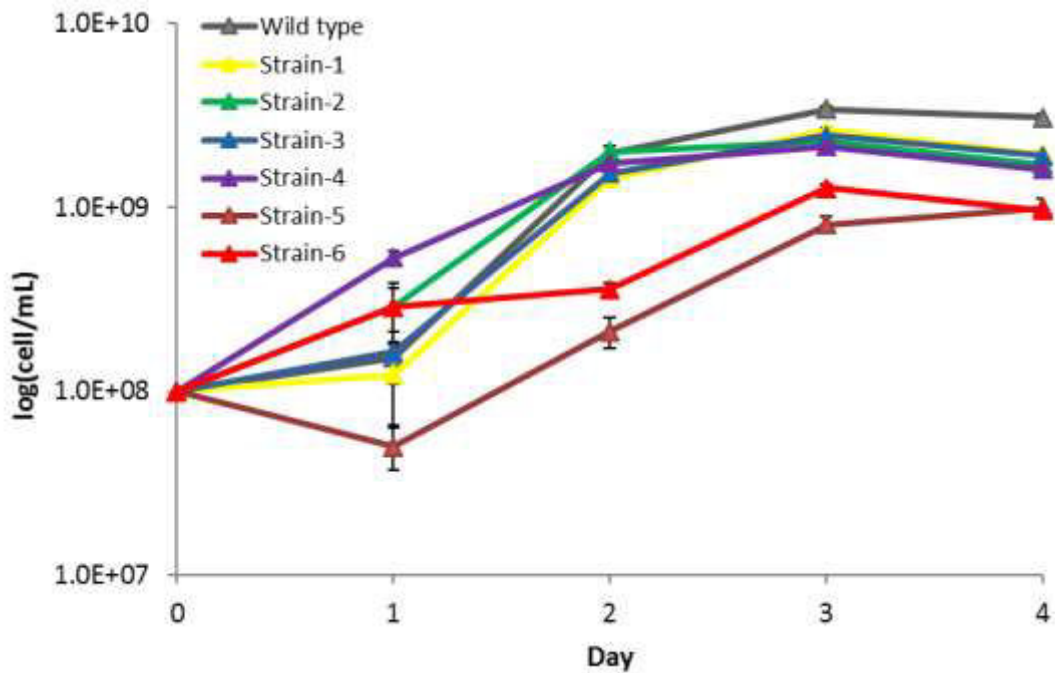
## 6.3 RESULTS

Six independently inoculated lines of *C. fimi* were initiated from a single colonies of the wild type strain, grown on rich media in liquid culture and then transferred to shake flasks containing 50 mL of minimal medium supplemented with 2% w/v of chopped wheat straw as the main carbon source. Flasks were incubated with shaking at 180 rpm at 30°C for one week, at which time an aliquot of the cells was removed and used to inoculate a fresh flask of wheat straw media. Cells from each line were also archived as glycerol stocks at the same time as the weekly transfers. This process was repeated weekly for 12 months, at which point the growth of the adapted cell lines on wheat straw was compared to that of the starting wild type cells.

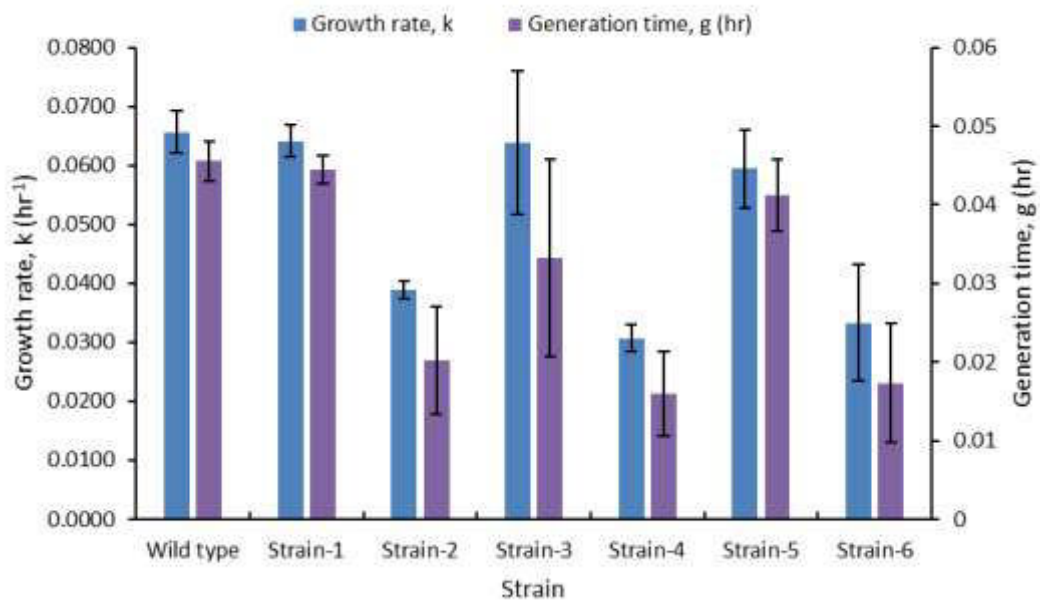
### 6.3.1 Growth rates on wheat straw of *C. fimi* strains

Growth profiles of *C. fimi* strains were monitored to measure the relative bacterial growth rate. Figure 6.2(A) shows the growth profiles of the wild type and 6 adapted strains of *C. fimi* in the wheat straw culture. All adapted strains showed a similar growth pattern and a comparable lag phase to the wild type in the first day of the culture except strain-4 and strain-5; these strains had a lag phase that was slightly faster or slower compared to the other strains, respectively. In all strains, exponential phase was reached slowly within 2 days. Strain-5 had the lowest OD throughout the incubation process, followed by the strain-6; the cell density in these two strains was significantly lower than the wild type throughout the 4-day time course, as measured from the supernatant. In Figure 6.2(B), the estimated generation time and growth rate were calculated between day-1 and day-3 during the exponential growth phase for all the cultures. The histogram shows that the wild type, strain-1, strain-3, and strain-5 had comparable growth rates;  $0.0657 \pm 0.04 \text{ h}^{-1}$ ,  $0.0641 \pm 0.003 \text{ h}^{-1}$ ,  $0.0639 \pm 0.012 \text{ h}^{-1}$ , and  $0.0594 \pm 0.007 \text{ h}^{-1}$ , respectively. Meanwhile, the generation time of strain-6 was 2-fold lower than the wild type, which was represented as the lowest cells count from the medium supernatant (but not representing the whole wheat straw culture).

**A) Growth profiles**



**B) Growth rates and generation times**



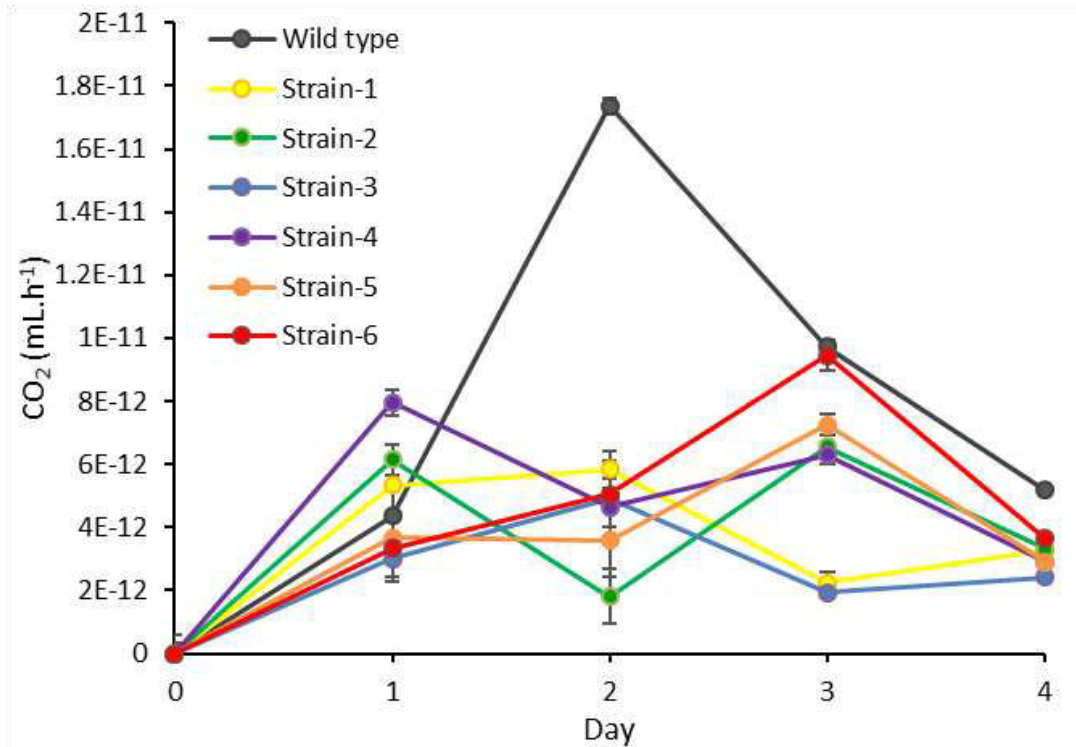
**Figure 6.2: Comparison of growth profiles between *C. fimi* wild type and adapted strain-1 to strain-6 in 4-day time course.**

**A)** Growth profile of *C. fimi* strains over a 4-day time course. **B)** Growth rates and generation times calculated during the exponential phase of *C. fimi* strains grown in wheat straw medium. The mean of four biological replicates  $\pm$  SD is shown.

### 6.3.2 Carbon Dioxide Evolution Rate (CER) of *C. fimi*

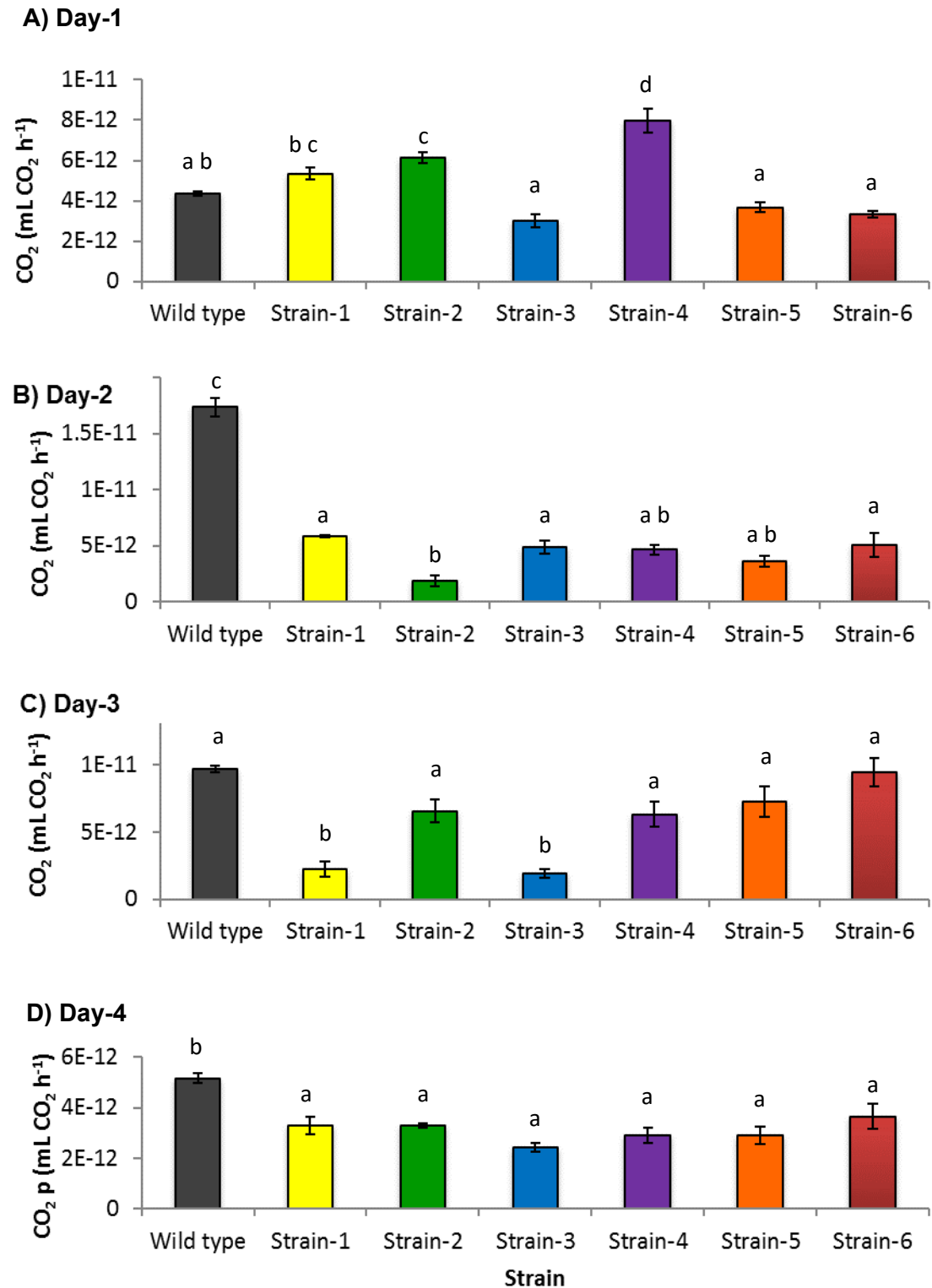
The amount of carbon dioxide (CO<sub>2</sub>) gas released in this experiment was measured as an alternative to measuring culture medium OD, as this would not account for cells that were anchored to the insoluble straw substrate. Figure 6.3 shows the CER profiles by 7 strains of *C. fimi* over a 4-day time course. A statistical analysis was performed using a one-way ANOVA test to compare the CER between each strain and data are presented as the bar plots in Figure 6.4.

On day-1 of incubation, strain-4 has the highest CER compared to the other 6 strains including the wild type. However, the CER of strain-4 dropped steadily throughout the 4 days of time course of the culture, which is statistically different to the other strains. The wild type showed a more rapid increase of CER from day-1 to day-2 of incubation compared to all six adapted strains of *C. fimi*. The CER of the wild type decreased immediately on the day-2 of incubation until the day-4 but was still significantly the highest compared to the other strains. Strain-1 and strain-3 exhibited a similar CER pattern where the CER of both strains increased from day-1 to day-2 and then started to drop to the lowest on day-3. Their CER remained unchanged in day-4. Strain-5 and strain-6 also possessed similar profiles; there was little change in the CER from day-1 to day-2, and both strains produced higher CO<sub>2</sub> on day-3 before drastically decreasing on the day-4.



**Figure 6.3: Carbon dioxide evolution rate (CER) for wild type and 6 adapted strains of *C. fimi* grown on wheat straw.**

Profiles of CER over a 4-day time course for *C. fimi* strains including wild type. The initial and final headspace pressure of the culture in the wheaton bottles during temporary anaerobic condition was measured using pressure gauge. CO<sub>2</sub> released during the anaerobic period was collected into exetainer® bottles. The gas samples were quantified by gas chromatography (GC) using a HP-5 column, and detected using a Flame Ionization Detector (FID). The mean of four biological replicates ± SD is shown.



**Figure 6.4: Carbon dioxide evolution comparison between *C. fimi* wild type and six independent one year-adapted strains over a 4-day time course.**

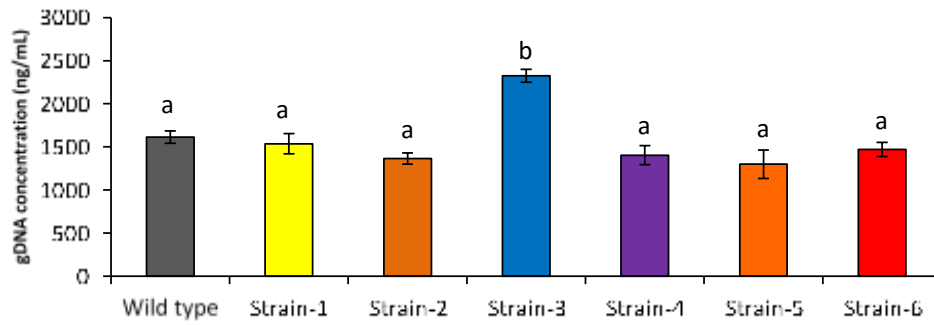
**A) Carbon dioxide evolution from Day-1; B) Day-2; C) Day-3; D) Day-4** incubated culture. The mean of four biological replicates  $\pm$  SD is shown. Results were statistically analysed by SigmaPlot 13 using One-way ANOVA of multiple comparisons test  $P < 0.05$ .

### 6.3.3 Isolation of *C. fimi* Genomic DNA (gDNA) from supernatant and biomass fractions

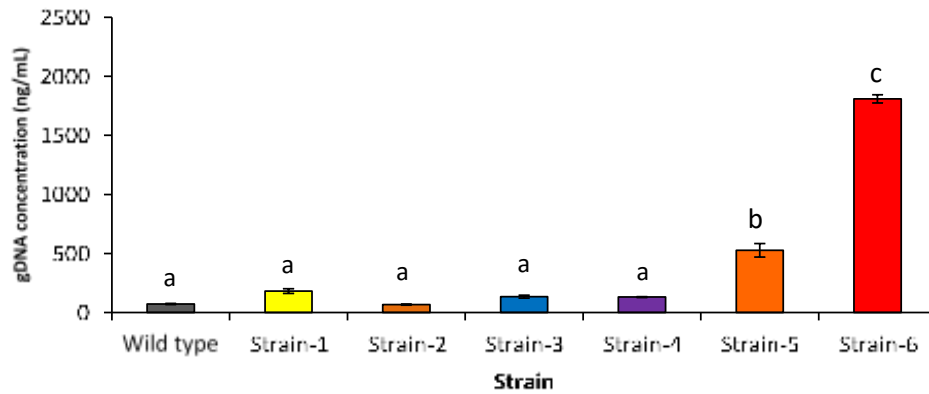
To provide an alternative measurement of the amount of cells from each line that were anchored to the straw biomass and in the culture medium, total genomic DNA of the wild type and adapted *C. fimi* strains were extracted from both supernatant and biomass-bound fractions using a standard phenol-chloroform method (312) after 4 days of aerobic culture incubation. From the supernatant fraction, most of the strains showed a similar amount of extracted DNA except the strain-3 gave significantly highest extracted gDNA ( $2323.7 \pm 75.2$  ng/mL). Whilst a high amount of gDNA was extracted from strain-5 ( $527.5 \pm 56.7$  ng/mL), strain-6 generated the most gDNA ( $1809.3 \pm 35.8$  ng/mL) from biomass-bound fraction compared of all strains (Figure 6.5B). The measurement was correlated by qualitative observation of high molecular weight bacterial gDNA on the TAE DNA agarose gels presented in Figure 6.5(C) and Figure 6.5(D).



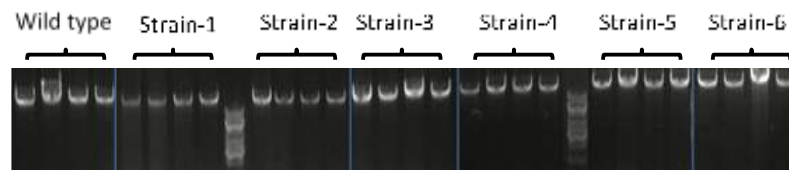
**A) Supernatant fraction**



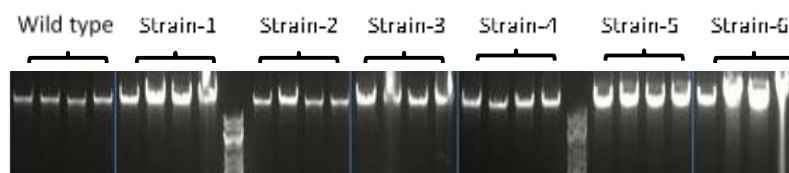
**B) Biomass-bound fraction**



**C) Supernatant fraction**



**D) Biomass-bound fraction**



**Figure 6.5: Quantitative measurement and qualitative observation of gDNA isolated from Day-4 of the wild type and six population adapted strains of *C. fimi* from wheat straw culture.**

NanoDrop® measurement of isolated gDNA from supernatant (A) and biomass fractions (B) of wild type and six adapted strains of *C. fimi* culture in 4 biological replicates, respectively. The mean of four biological replicates  $\pm$  SD is shown. Results were statistically analysed by SigmaPlot 13 using One-way ANOVA of multiple comparisons test  $P < 0.05$ . C) and D) 1% agarose gel loaded with 15  $\mu$ L of gDNA samples isolated from supernatant and biomass fractions of wild type and six adapted strains of *C. fimi* cultures in 4 biological replicates, respectively.

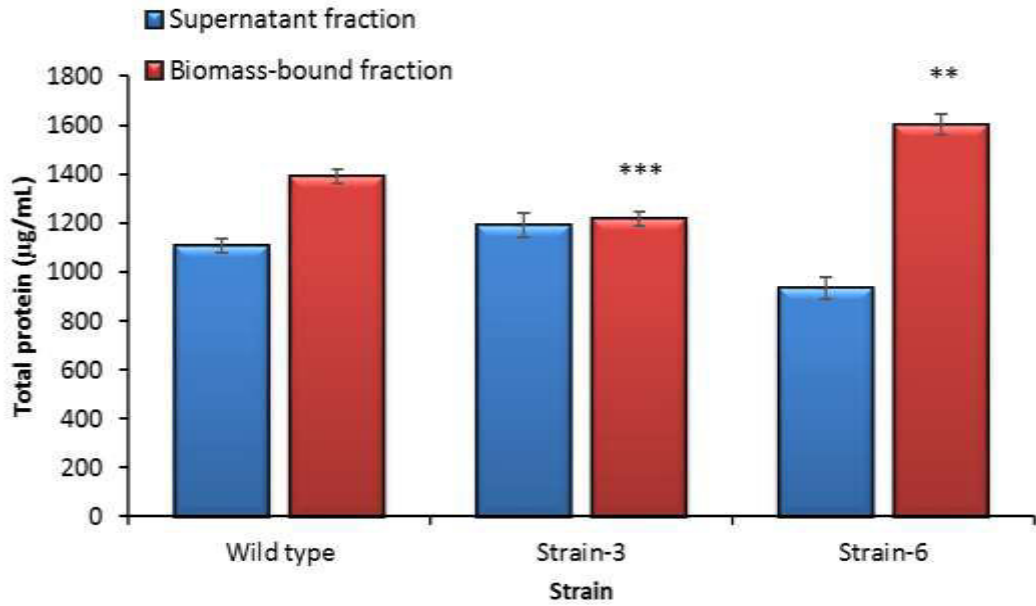
#### **6.3.4 Total protein isolation from *C. fimi* culture supernatant and cell-bound to wheat straw**

Strain-3 and strain-6 were selected for total protein analysis for a comparison against the wild type as they showed the highest amount of extracted DNA in the supernatant and in the biomass-bound fractions, respectively.

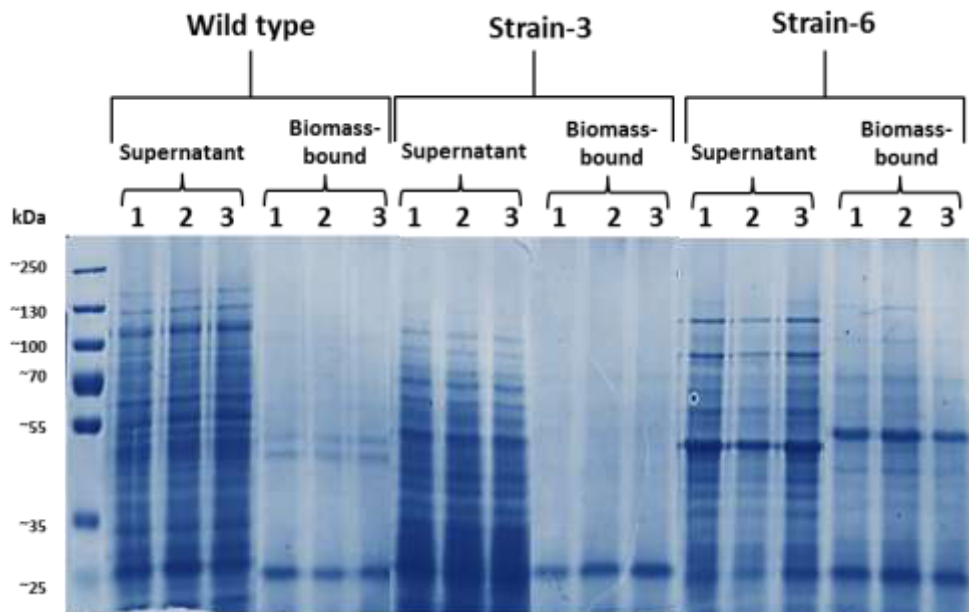
In the culture supernatant, there was no statistical difference observed in total protein concentration extracted in strain-3 and strain-6 compared to the wild type (Figure 6.6A) despite slight differences in the protein band intensity observed after SDS-PAGE (Figure 6.6B). Nonetheless, protein bands observed on the SDS-PAGE gels showed some variance in the size and intensity of the protein bands suggesting a variation of proteins might be secreted into the supernatant during wheat straw degradation by the selected strains.

However, in the biomass-bound fraction, the quantified total protein appeared to be strongly correlated with the intensity of protein bands on the gel. Total protein from strain-3 was significantly lower than the total protein from wild type, whereas strain-6 gave the highest total protein abundance compared to the wild type and strain-3. The SDS-PAGE gel showed a good reproducibility of the result from 3 biological replicates of the strain cultures.

**A) Protein Bradford assay**



**B) Protein SDS-PAGE gel**



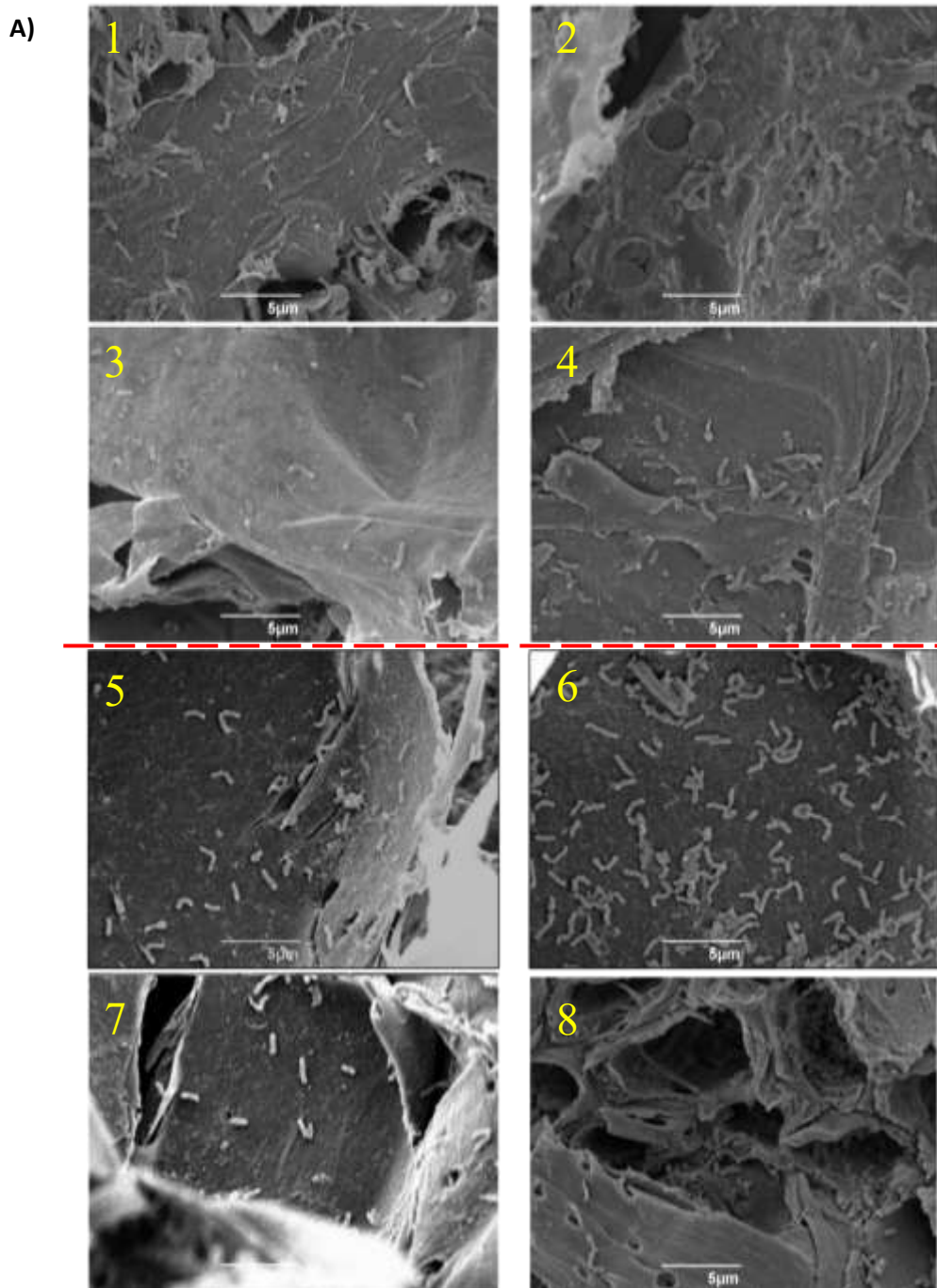
**Figure 6.6: Quantification of total protein isolated from supernatant and biomass fraction of wild type, strain-3 and strain-6 of 1-year adapted strains.**

**A)** Relative protein quantification using Bradford method. Mean of three biological replicates  $\pm$  SD shown. Results were statistically analysed by SigmaPlot 13 using One-way ANOVA of multiple comparisons with post-hoc Bonferroni test \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

**B)** SDS-PAGE loaded with 15  $\mu$ L of total protein isolated from supernatant and biomass fractions in three biological replicates of each culture. Protein was extracted using standard alcohol precipitation (section 5.2.1) followed by the Bradford assay (Section 2.4.1).

### **6.3.5 Qualitative visualisation of *C. fimi* bound to wheat straw using Scanning Electron microscopy (SEM)**

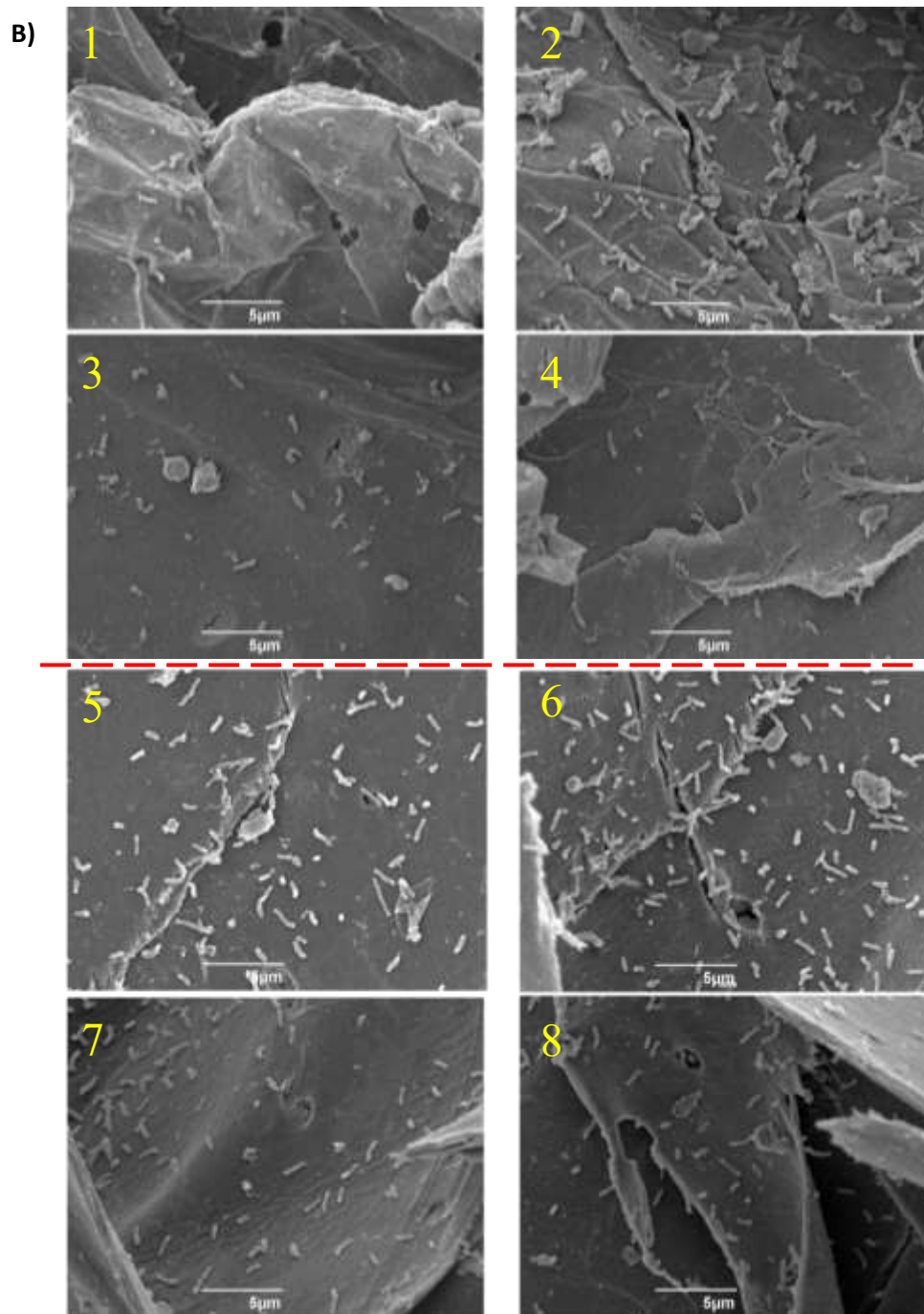
Scanning electron microscopy (SEM) was used to visualize the morphological characteristics of the interaction of wheat straw and *C. fimi* cells after 4 days of degradation. In all SEM observations, the wheat straw fibres appeared to be a relatively thin and multi-layered with some irregularities on the fibre surface. Putative bacterial cells can be seen on the wheat straw surfaces in most specimens of the strains. Figure 6.7(A) to figure 6.7(G) show the biomass surface characteristics of the wild type, and strain-1 to strain-6, after incubation respectively. Strain-1 and strain-4 were perceived to have a similar abundance of the cells associated on the wheat straw surface, where cells were scattered on the surface on most of wheat straw particles observed as in figure 6.7(B) and figure 6.7(E), respectively. Strain-5 and strain-6 were observed to have most obvious differences, with a greater abundance of attached cells apparent in the images as compared to other strains including the wild type (Figure 6.7(F) and figure 6.7(G)).



**Figure 6.7(A):** Surface images obtained by SEM on wheat straw after 4 days of aerobic degradation inoculated with the wild type of *C. fimi* in basal medium.

Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

The magnification is shown with the scale bars on the images.

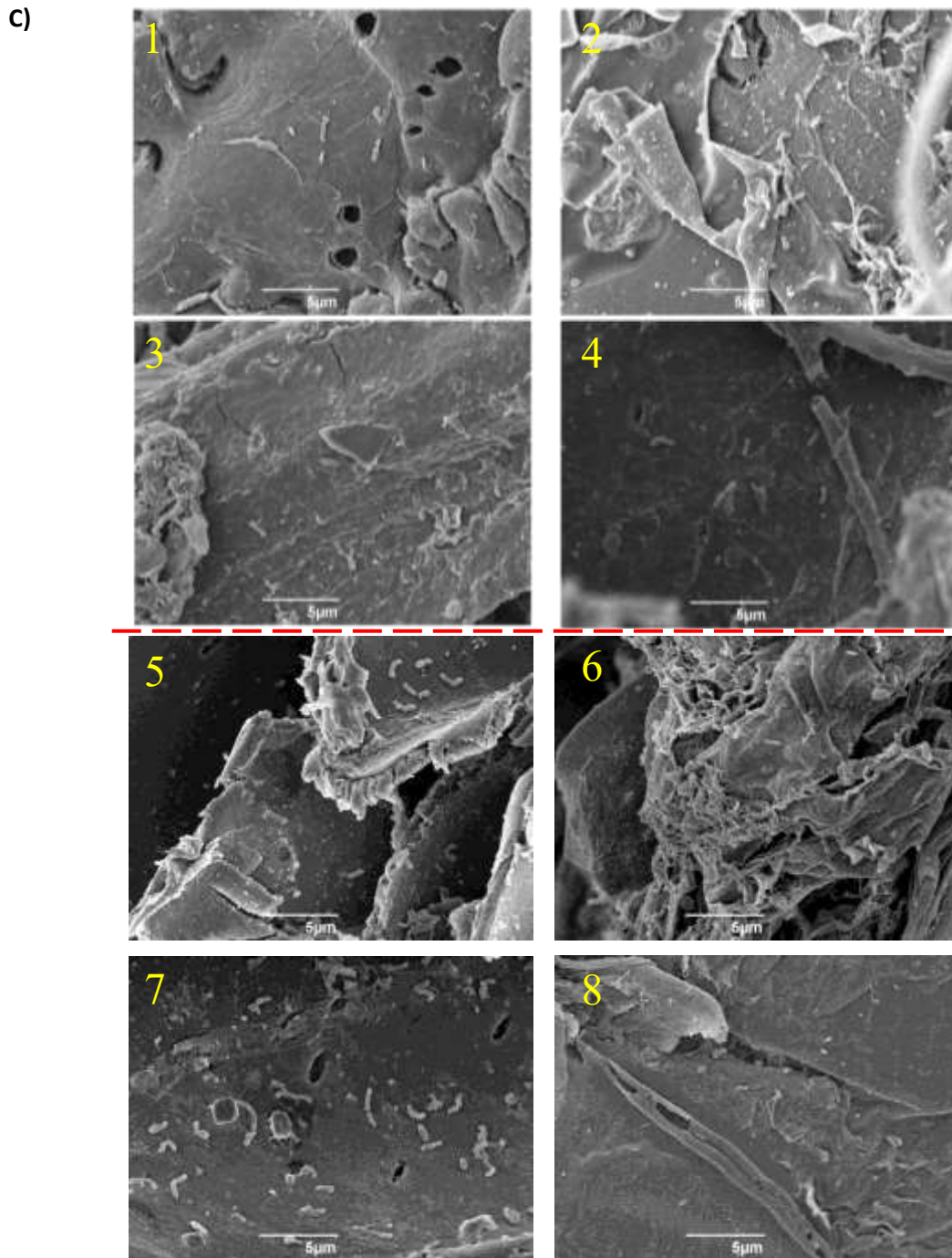


**Figure 6.7(B): Surface images obtained by SEM on wheat straw inoculated with strain-1 of *C. fimi* after 4 days of aerobic degradation in basal medium.**

Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

The magnification is shown with the scale bars on the images.

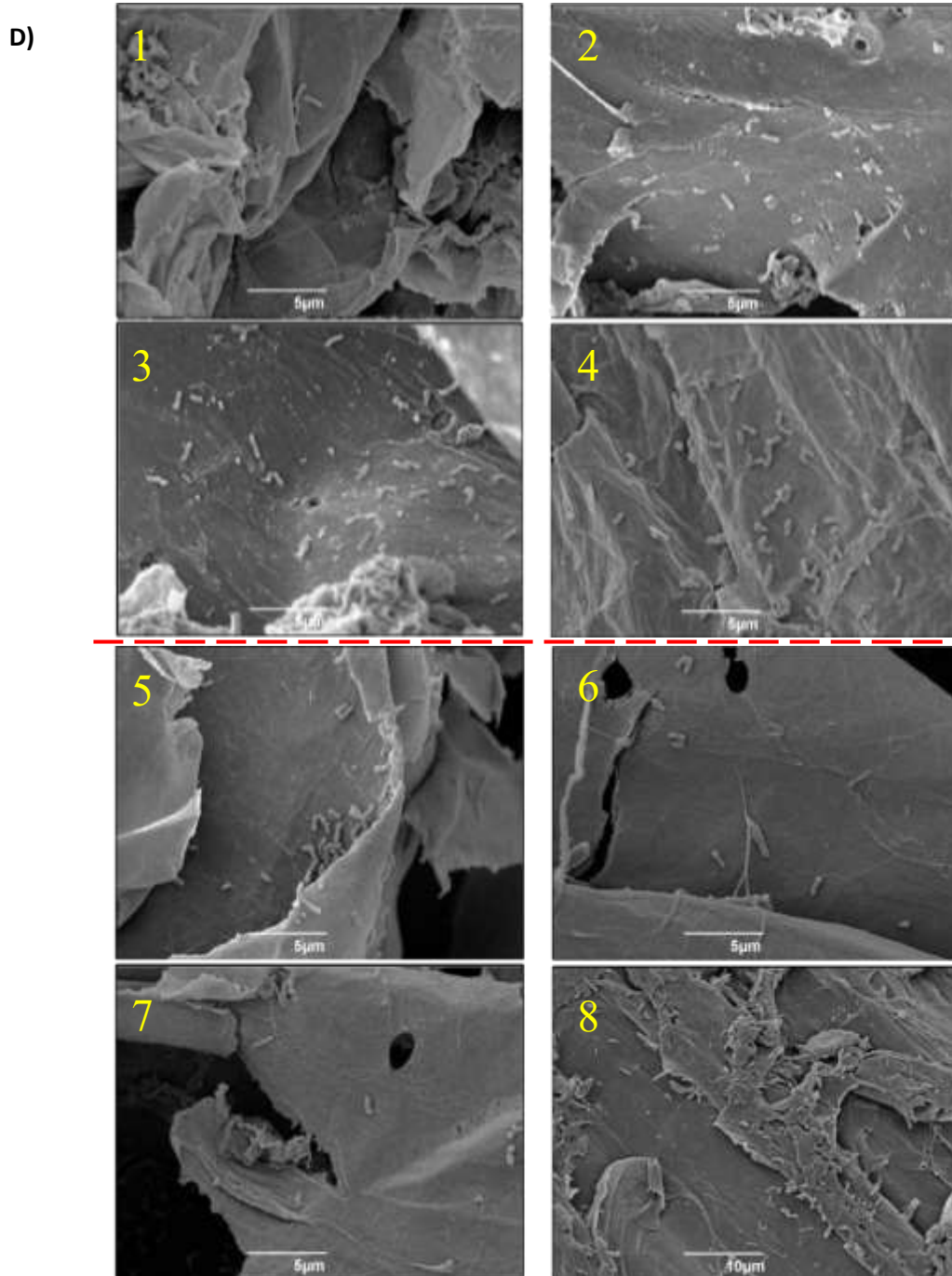




**Figure 6.7(C): Surface images obtained by SEM on wheat straw inoculated with strain-2 of *C. fimi* after 4 days of aerobic degradation in basal medium.**

Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

The magnification is shown with the scale bars on the images.

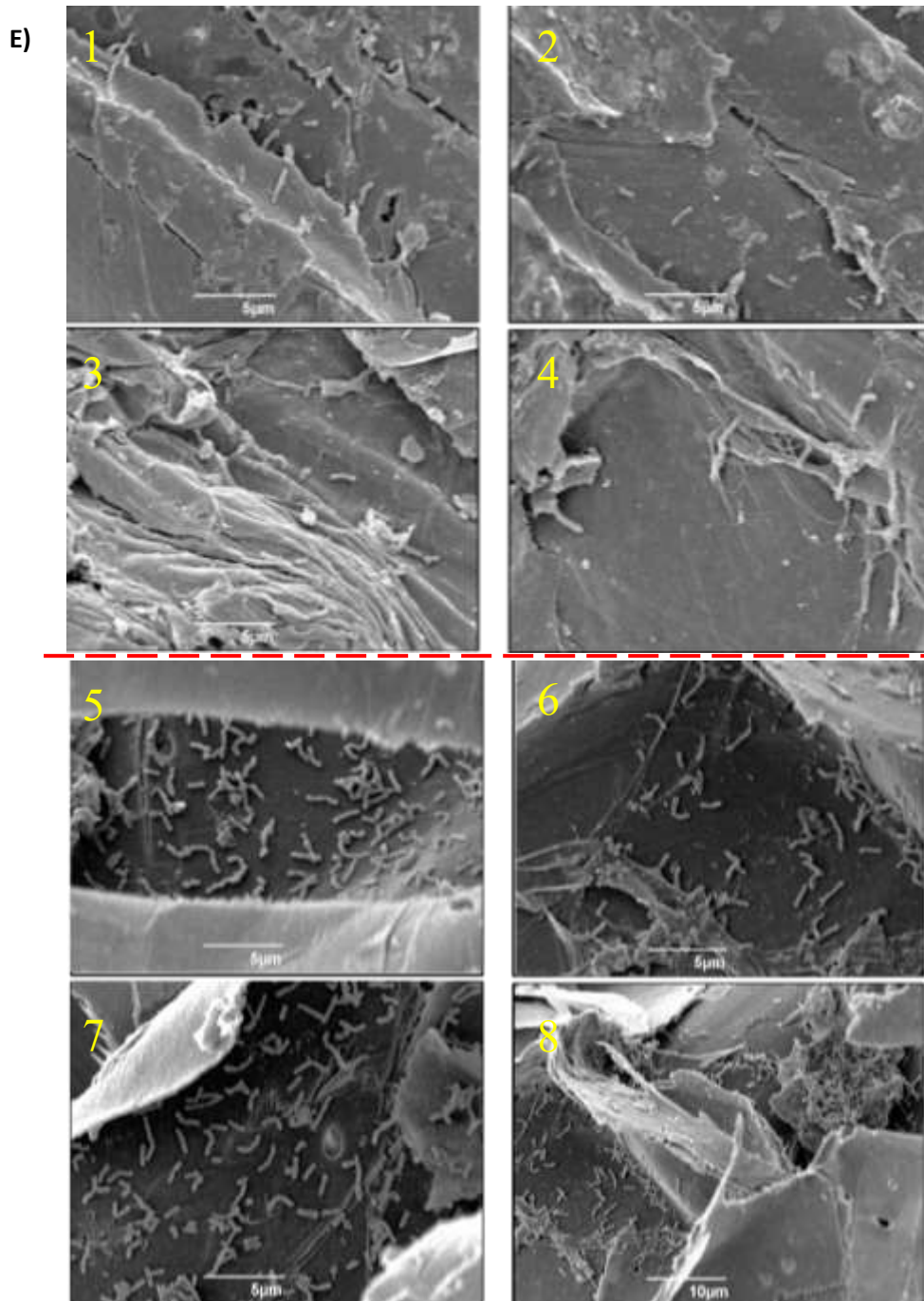


**Figure 6.7(D): Surface images obtained by SEM on wheat straw inoculated with strain-3 of *C. fimi* after 4 days of aerobic degradation in basal medium.**

Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

The magnification is shown with the scale bars on the images.

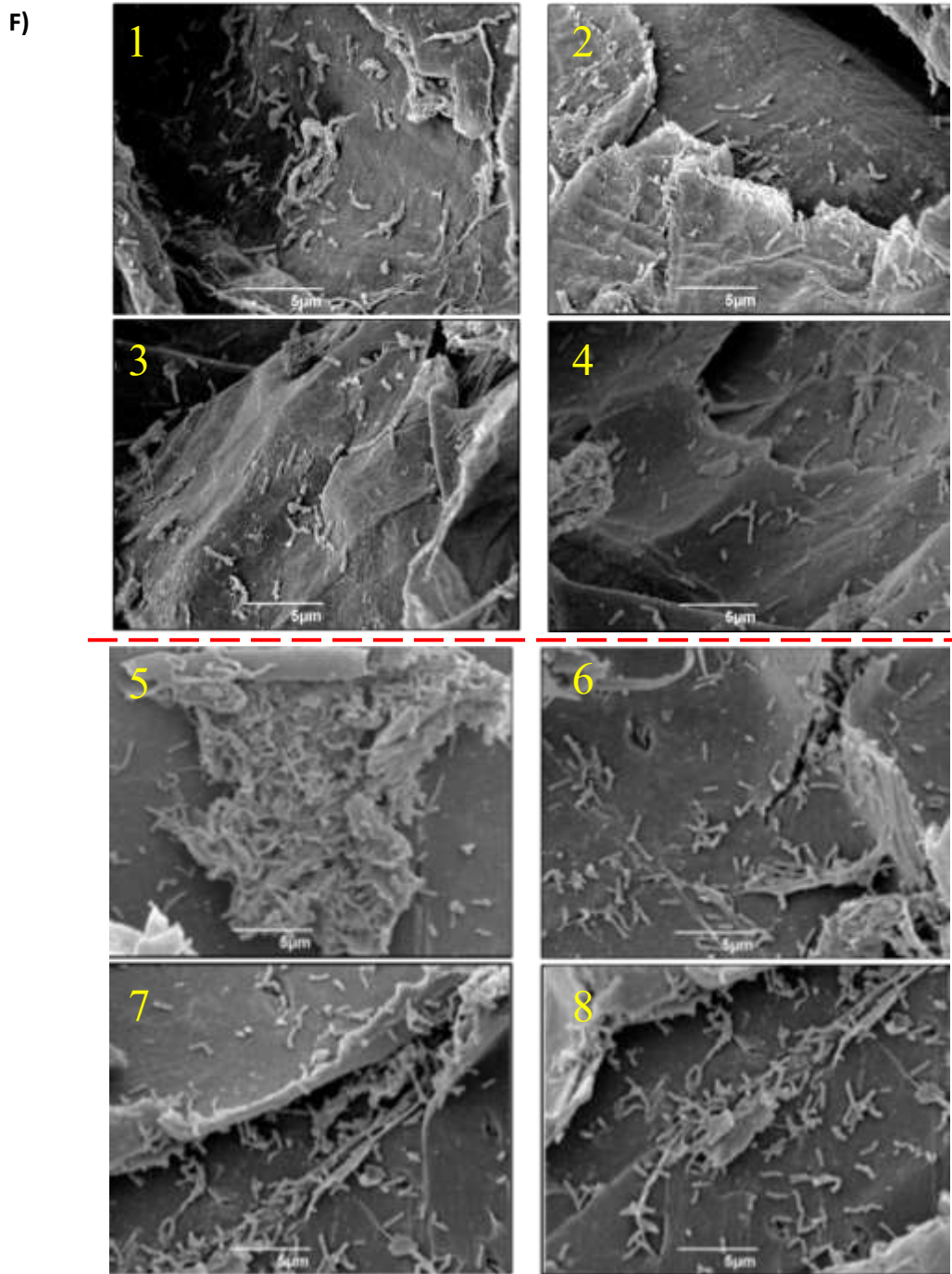




**Figure 6.7(E): Surface images obtained by SEM on wheat straw inoculated with strain-4 of *C. fimi* after 4 days of aerobic degradation in basal medium.**

Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

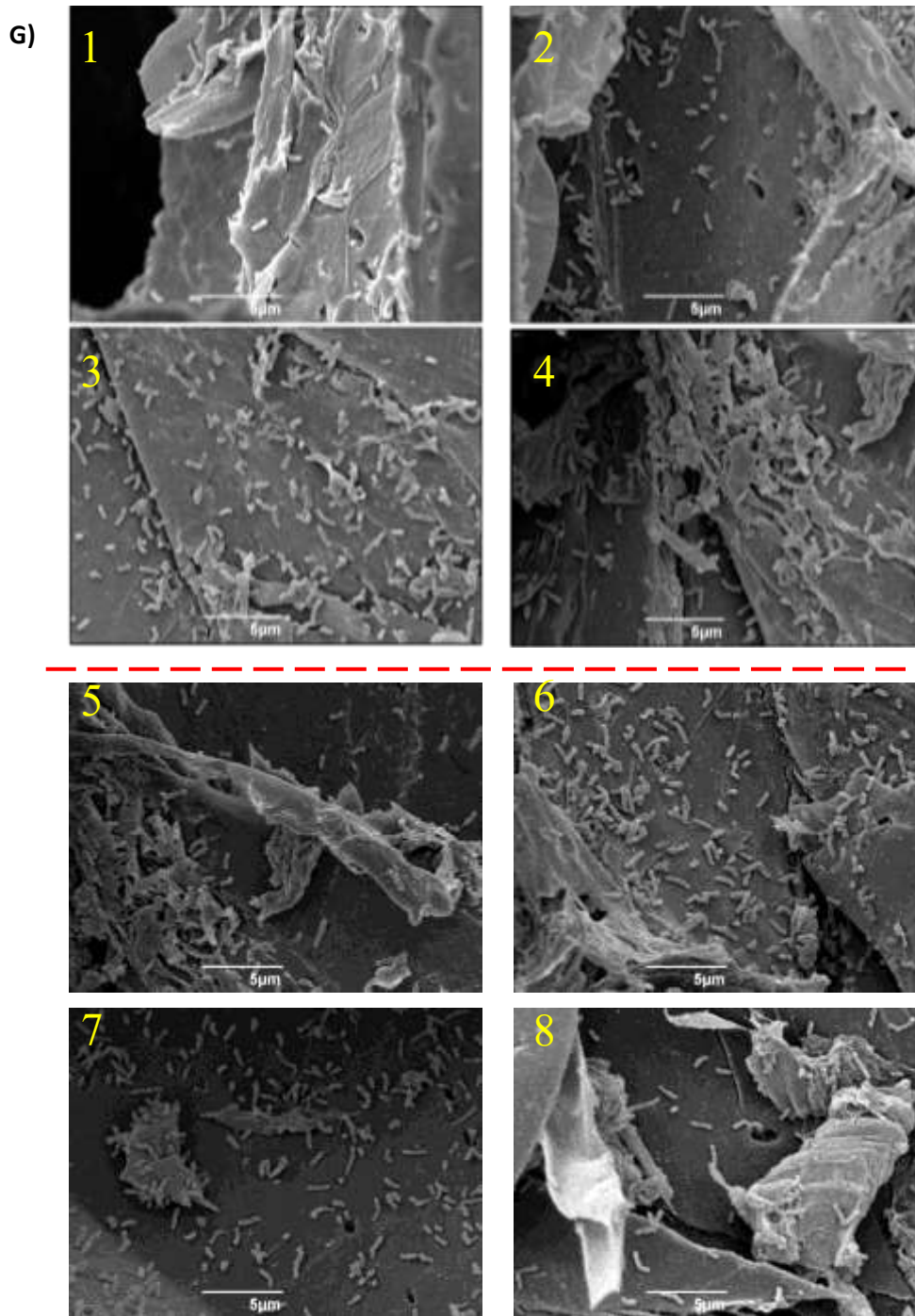
The magnification is shown with the scale bars on the images.



**Figure 6.7(F): Surface images obtained by SEM on wheat straw inoculated with strain-5 of *C. fimi* after 4 days of aerobic degradation in basal medium.**

Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

The magnification is shown with the scale bars on the images.



**Figure 6.7(G): Surface images obtained by SEM on wheat straw inoculated with strain-6 of *C. fimi* after 4 days of aerobic degradation in basal medium.**

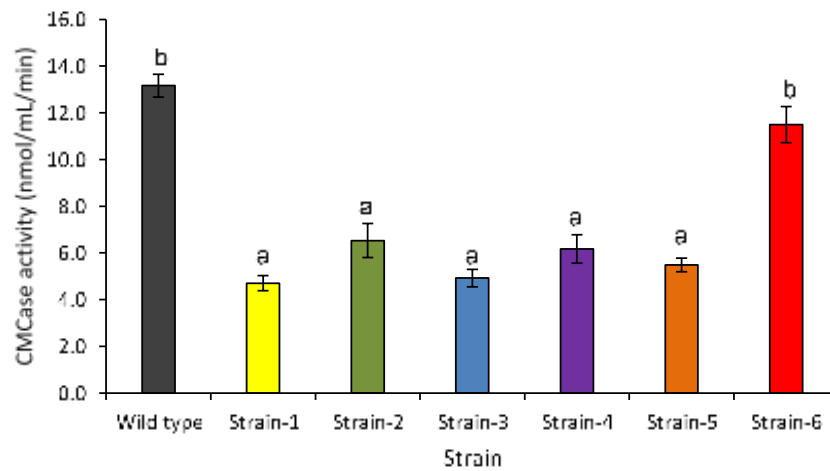
Image 1 to 4 represent 4 different areas of wheat straw particles from replicate 1. Image 5 to 8 represent 4 different areas of wheat straw particles from replicate 2.

The magnification is shown with the scale bars on the images.

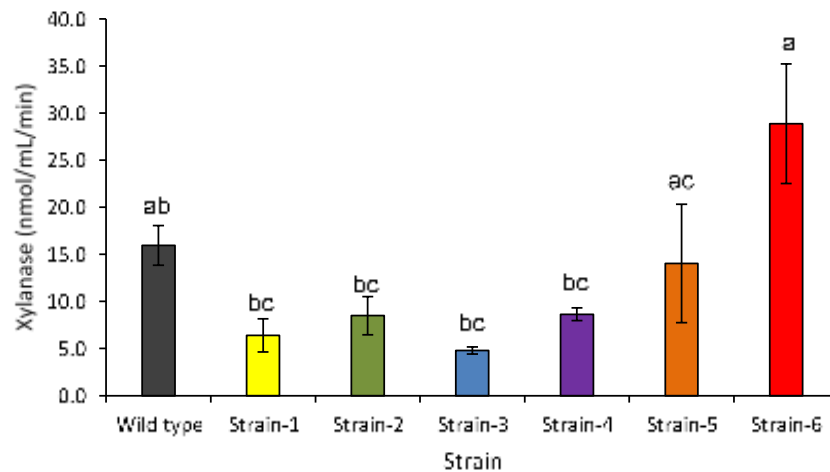
### 6.3.6 Enzyme activity assay using culture supernatants

In order to assess the relative soluble enzyme activity for each strain, aliquots of equal volume from culture supernatants were assayed for their ability to release sugars from either carboxymethyl cellulose (CMC) or xylan. The greatest amount of reducing sugar was observed after incubation on CMC from the wild type and strain-6 (Figure 6.8A). However, CMCase activity of the wild type and strain-6 were not statistically different to each other, with  $13.16 \pm 0.49$  and  $11.59 \pm 0.77$  nmol/mL/min respectively. Overall, this suggests that there was significantly lower CMCase activity in all strains versus the wild type except in strain-6. In terms of xylanase activity, enzyme activity on beechwood xylan was detected. (Figure 6.8B). Adapted *C. fimi* strains of strain-1, -2, -3 and -4 were observed to release similar amount of reducing sugars during the assay on xylan compared to the wild type. In spite of that, strain-5 and strain-6 were significantly at best equivalent to the wild type with big error bars with strain-6 displaying the highest xylanase activity across all the *C. fimi* strains.

**A) CMCase activity**



**B) Xylanase activity**



**Figure 6.8: Comparison of CMCase and xylanase activity in culture supernatant of wild type and adapted strain-1 to strain-6 of *C. fimi*.**

A) CMCase activity, and B) Xylanase activity were determined by release of reducing sugars in response to incubation of CMCase and xylanase substrates, CM-cellulose and beechwood xylan, with *C. fimi* culture supernatants, respectively. One unit of enzyme activity was defined as the amount of enzyme required to liberate one nanomole per milliliter equivalent of reducing sugars per minute. Mean of four biological replicates  $\pm$  SD shown. Result was statistically analysed by SigmaPlot13 using One-way ANOVA of multiple comparisons test  $P < 0.05$ .

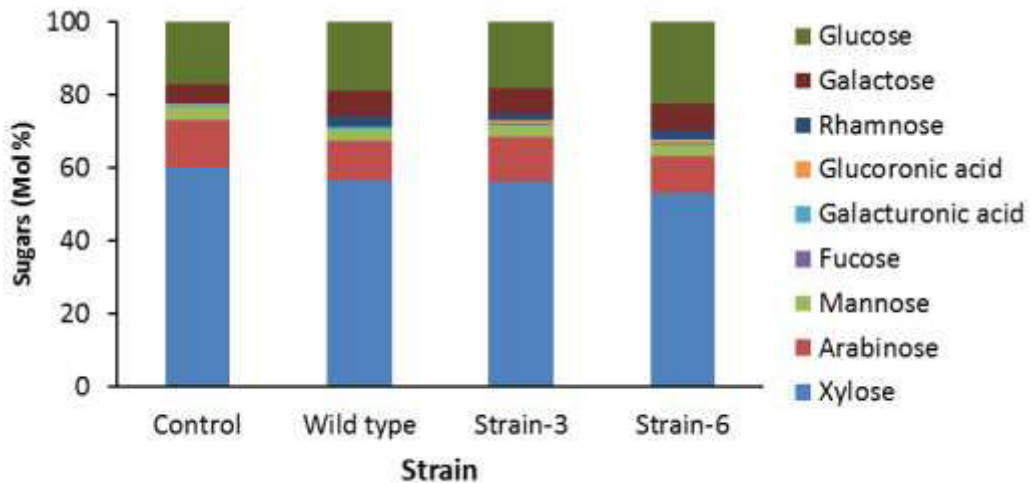
### 6.3.7 Polysaccharide content and composition

The composition of plant cell walls after the degradation of wheat straw in liquid culture was assessed by mild trifluoroacetic acid (TFA) digestion to release the monosaccharides from the non-crystalline polysaccharides in the biomass residue following growth of the adapted and wild type strains. Strain-3 and strain-6 were selected for this analysis because both strains showed the highest amount of extracted gDNA from the supernatant and biomass-bound fractions, respectively as presented in section 6.3.3. Figure 6.9(A) shows the profiles of components from hydrolyzed plant cell walls after the degradation by *C. fimi* strains grown on wheat straw. Statistical analysis was performed by a t-test pairwise comparison to the uninoculated wheat straw using a SigmaPlot 13 software (Figure 6.9B).

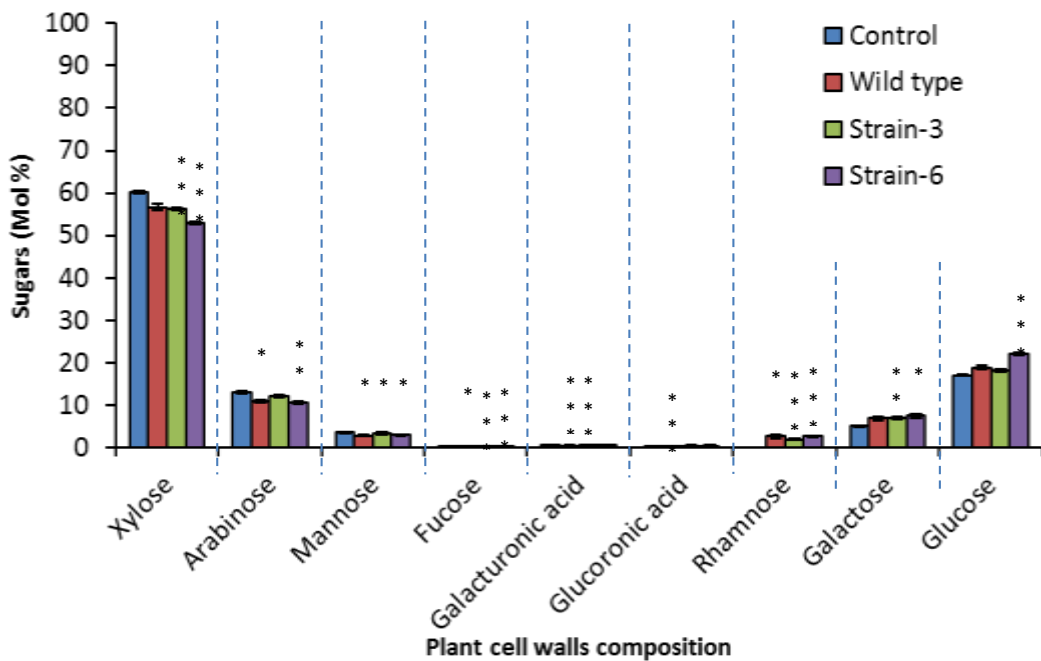
The quantities of glucose and galactose in biomass residues from growth with strain-6 were significantly higher than the uninoculated wheat straw (control), wild type and strain-3. Higher levels of glucuronic acid (GlcA) and galacturonic acid (GalA) were observed compared to the wild type and strain-3, however, they were not statistically different with the level of those monosaccharides in the negative control. Whilst significantly higher quantities of fucose and rhamnose were apparent from strain-6 compared to the negative control, comparable levels of the same monosaccharides were observed from the wild type and strain 6, and these two sugars are major components of the cell walls of *C. fimi* (313,314) and some of both sugars are likely to be derived from bacterial biomass. In comparison to the wild type, strain-3 and the control, strain-6 exhibits significantly the lowest quantities of xylose and arabinose. Mannose level from strain-6 ( $2.96 \pm 0.09$  Mol %) was significantly the lowest compared to the uninoculated wheat straw and strain-3, however, it was very close level to the wild type ( $2.89 \pm 0.17$  Mol %).



**A) Sugars composition of residual wheat straw after degradation by *C. fimi* strains**



**B) Comparison between *C. fimi* strains of sugars composition of residual wheat straw after degradation**

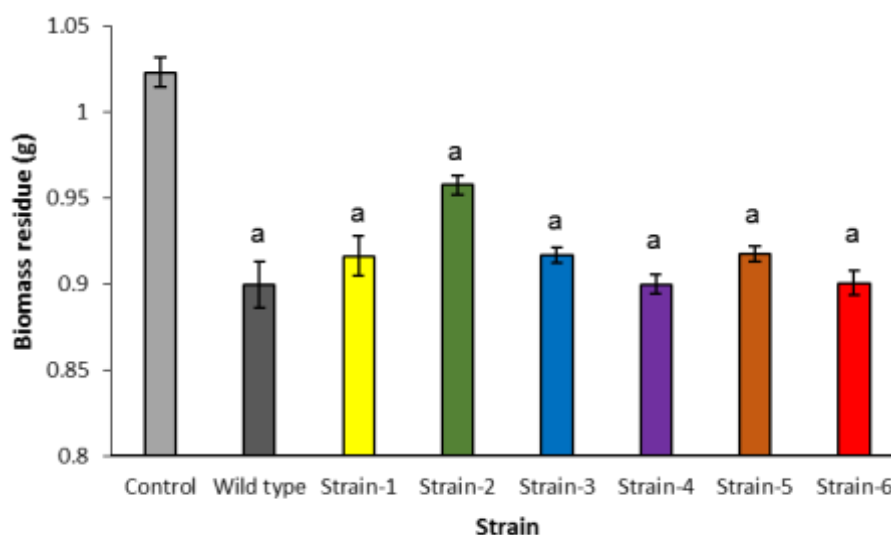


**Figure 6.9: Monosaccharide composition from the wheat straw degradation by wild type, strain-3 and strain-6 of *C. fimi*.**

**A)** The stacked column graphs represent concentration of 9 monosaccharides composition contained in degraded wheat straw by *C. fimi* strains. **B)** The clustered column graphs represent a comparison of the 9 sugars detected in wild type, strain-3 and strain-6 of residual biomass to the uninoculated wheat straw (control). Results were statistically analysed using SigmaPlot 13 by a pairwise t-test against the negative control of uninoculated wheat straw. Three biological replicates represent  $\pm$  SD shown. P<0.05, \*\* P<0.01, \*\*\* P<0.001.

### 6.3.8 Quantification of residual wheat straw

Figure 6.10 shows the total residual of wheat straw from the culture of the wild type and six *C. fimi* strains. The result was compared to the negative control represented by the wheat straw culture that not been inoculated with the bacteria. The overall quantification of biomass loss during the incubation of *C. fimi* strains in the wheat straw indicates a mass loss (approximately 10%) from total wheat straw supplied in the initial culture. However, in comparison between the wild type and six adapted strains of *C. fimi*, there was no significant difference in mass loss of the culture across all the strains.



**Figure 6.10: Mass loss of wheat straw after the degradation by *C. fimi* strains.**

The *C. fimi* strains were cultured with wheat straw as a sole carbon source in the basal medium for 10 days before the mass loss analysis was measured. The mean of four biological replicates  $\pm$  SD is shown. Results were statistically analysed by SigmaPlot 13 using one-way ANOVA of multiple comparisons with post-hoc Bonferroni test, \*  $P < 0.05$ .



## 6.4 DISCUSSION

Adaptive evolution is one of the established approaches for microbial adaptation (303,306,308) to promote higher fitness and to evaluate potential improved properties such as in degradation of plant biomass (193,308). Naturally mutagenised strains of *C. fimi*, a cellulolytic soil bacterium, were produced from 52-weeks continuous subculture in wheat straw medium, by a serial transfer method for every 7-days. In this experiment, six populations of *C. fimi* were evaluated for higher fitness in the wheat straw liquid culture. The strains were screened and compared to the wild type in order to get a better understanding of any changes during the adaptation period in wheat straw environment. Several properties were investigated including their growth characteristics, enzyme activities, and analysis of degraded wheat straw.

All strains were able to utilize wheat straw as a sole carbon source for their growth until four days of incubation. Most of the strains including the wild type demonstrated a long lag phase which took about 24 h before starting the exponential phase, compared to well-studied bacteria such as *E. coli* growing in a rich medium, where the lag phase is short which around 15 to 20 min (315,316). The lag in *C. fimi* cultures may be due to the fact that the average generation time of soil bacteria ranges from 1 to 2 days (317). This indicates that *C. fimi* strains required time to adapt and exploit the novel environment (318) where there is a limited accessibility to the substrate, particularly in the lignocellulosic wheat straw. After four days of culture incubation, there is a decline in the cells number of all *C. fimi* strains, where a stationary phase was defined as the point at when the number of cell counts did not show a significant difference over 24 h after 3 days of incubation. This may be a response to the depletion of accessible saccharides from the wheat straw, especially in the hemicellulose and amorphous cellulose parts of plant cell walls. The difficulty to breakdown the amorphous part of the matrix in lignocellulose polysaccharides (225) also may leads to a rate-limiting steps in biomass degradation by bacterial cells (59).

The generation time of strain-6 was 2-fold lower compared to the wild type as was estimated based on the cells density from the culture supernatant. Therefore, further analyses were done on the wheat straw biomass samples to have clearer ideas and investigate the physiological changes in the adapted strains compared to the wild type. The method of carbon dioxide evolution rate (CER) has previously been widely applied in ecotoxicological tests (319), investigation of microbial processes in soil (320), and waste water treatment (321).

This analysis was used to measure the aerobic respiratory and biodegradation rates by *C. fimi* strains during wheat straw degradation in liquid culture. All six adapted strains showed significantly lower CER compared to the wild type, except strain-4 in day-1. This observation was correlated with the faster growth of strain-4 compared to other strains during the lag phase in day-1 which may be a sign of rapid cells metabolism hence higher CO<sub>2</sub> produced in the culture.

Microbial metabolism evaluation by respiratory activity measurement maybe an ideal method for a soluble substrate culture (322); however, it also has several limitations (323,324). The CER principle needs to be carefully considered as *C. fimi* is a facultative anaerobic bacterium that can grow by producing energy by aerobic respiration in the presence of oxygen, but is also capable of switching to anaerobic respiration if oxygen is absent (39,86). Consequently, CO<sub>2</sub> is measured by collecting the CO<sub>2</sub> using a syringe into exetainer sample tubes. However, the liberation efficiency of CO<sub>2</sub> from the liquid phase in the wheaton bottle and the subsequent trapping efficiency of CO<sub>2</sub> into the exetainer tube may lead to lower measured biodegradation rates (325). This showed the dynamic digestion strategy of the bacteria depending to the environmental condition and substrate availability.

The results from the CFU and CER measured strains growth characteristics specifically from culture supernatant lead to further exploration of the adapted *C. fimi* strains in biomass-bound fraction. The highest significant quantity of gDNA was isolated from the biomass-bound fraction of strain-6 in comparison of all strains including the wild type. The results of total protein extraction also imply a significantly higher association of strain-6 cells on the biomass than in the supernatant. The association of *C. fimi* strains on the wheat straw biomass was further investigated by a three-dimensional (3D) visualization of degraded wheat straw particles using Scanning Electron Microscopy (SEM). Images of degraded biomass from strain-5 and strain-6 indicate that these strains were likely to have been clustered and clumped as a cells consortium on the wheat straw particles. These strains demonstrated a 'cell-to-biomass' adhesion pattern that may be a part of a biomass-degradation mechanism that is dissimilar to other strains, which may be producing 'free-form' of enzyme to degrade the biomass (53). This characteristic would be an advantage for plant cell wall recognition and binding for improvement of biomass degradation by cellulolytic bacteria (63,326).

Despite the variance in results between the wild type and strain-6 in terms of growth characteristics and variation of cell localization in the culture, the reducing sugar assay using CM-cellulose and beechwood xylan as the substrates did not show any significant difference between the two strains, indicating similar amounts of cellulase and xylanase activity in these strains, but values were significantly lower in the other strains. Two of the best studied cellulases of *C. fimi* are the CenA (endoglucanase) and Cex (the GH10 of bifunctional exoglucanase/xylanase) which both comprise carbohydrate binding-module family 2 (CBM2) (39,40). The function of CBM2 has been extensively studied and it was shown that this domain can act synergistically with catalytic domain, and independently as non-catalytic domain in disrupting cellulose fibres (79,327). Consequently, this reaction results in the release of fine cellulose particles, however, without any detectable hydrolytic activity (79,103).

The overall CMCase and xylanase activities measured by the release of reducing sugar after incubation with culture supernatant could be improved by more specific *para*-nitrophenol-linked substrates (*p*NP)-assays. Examples of disaccharide and oligosaccharide substrates that could be used including *p*NP- $\beta$ -glucose, -fucose, -galactose, as well as -cellotetraose and -cellopentaose. By using *p*NP-substrates in the assay, it may reveal the competency of each strain to break down a range of structurally different sugar polymers to monosaccharides residues, linkage position, or the chain length of the saccharified substrate (113).

Analysis of the composition of non-cellulosic polysaccharides in the straw biomass after growth of the bacterial strains revealed decreases in the levels of xylose and arabinose, both of which are components of the complex arabinoxylans that make up much of the wheat straw hemicellulose (328). This indicates that substantial amounts of xylanase must be produced by *C. fimi* in agreement with the enzyme activity assays discussed above. Interestingly, the relative levels of rhamnose, fucose and galactose all went up during growth of *C. fimi* on wheat straw, indicating either that there was preferential hydrolysis of these components or that new sugars are appearing. Where the hemicellulose component of plant cell walls consists of heteroglucans (xyloglucans) which includes glucuronoxylans, these most abundant non-cellulosic polysaccharide in hardwoods may contain rhamnose and galacturonic acid (329). Furthermore, the cell walls of gram-positive bacteria comprises of a ~40% thick polymer called peptidoglycan that is heteropolymer consisting a rigid glycan chains and cross-lined with flexible peptide substituents (330). Several studies on *Cellulomonas sp.* have identified the cell-wall sugars in this species are including rhamnose, galactose, glucose, xylose, and inositol (313,314).

A commonly used method to measure and assess the rate and extent of biomass biodegradation was performed using a mass balance experiment after the wheat straw degradation by *C. fimi* strains. About 10% mass loss of the wheat straw was relatively quantified within 10 days of degradation by *C. fimi* strains. A comparable result was observed across all the adapted strains and the wild type showing there was significantly no difference in terms of biodegradation rate among all strains. The mass balance of materials is a relatively simple and straight forward approach where the initial biomass supplied in the culture and final dried biomass after the degradation was quantified. However, it should be noted that these measurements are complicated by the conversion of part of the biomass into bacterial mass, and this will lead potentially to a slight underestimate in the amount of wheat straw that has been degraded.

Factors that can affect the rate of degradation such as nutrient supply, temperature, pH and moisture level have been kept consistent throughout the experiment, however, physiological factors of biomass e.g. the nature of biomass that can absorb moisture that lead to the alteration of material weight has proved to be a problematic in the method application. In addition, the degrading biomass also could disintegrate into smaller fragments hence causing difficulty in recovering the materials to determine the real weight loss in each sample. Crystalline cellulose is not present in uniform composition, and biomass residues also varied in their moisture content and particle size, therefore, method revision and optimisation was required. Wheat straw biomass can be first homogenized by sieve and separate to a similar preferable particle size (2 mm) and conditioned to a uniform moisture content by extracting the degraded wheat straw from the liquid culture using vacuum filtration through nylon mesh to remove the excess water.

Complex configurations of heterogeneous polysaccharides in plant cell walls are recalcitrant to degradation (35,63). Microbes have been found to secrete specific carbohydrate active enzymes (CAZymes) to break down the plant cell wall structures (71,331). More than a decade ago, studies revealed that CAZy are often aided by non-catalytic proteins such as carbohydrate-binding modules (CBMs) (87,88,174). The majority of CBMs contain members that target and play critical functions using diverse mechanisms in the recognition and cell attachment of plant cell wall components (193,195,332). The numbers of CBM families increased from 67 to 81 from 2013 to 2017 and indicates that more CBM families with diverse roles could potentially be found in other plant degrading systems, such as in *C. fimi*.

The results of the growth study, CO<sub>2</sub> evolution measurements and SEM observations indicate that strain-6 has evolved to bind more tightly to the biomass during growth. This hypothesis is supported by data indicating that the significantly highest quantity of bacterial gDNA and total protein were extracted from the biomass-bound fraction samples. Furthermore, high levels of monosaccharides (glucose, rhamnose and galactose) were detected from the compositional analysis of matrix polysaccharides using HPAEC following TFA hydrolysis from strain-6 samples. These sugars are potentially derived the breakdown of plant biomass and also the composition of bacterial cell walls (313). It could be an indicator of better breakdown of biomass and a sign of stronger adhesion of strain-6 onto the wheat straw.

The increased of adhesion evidence of strain-6 could be an advantage that signifies a better capability to perceive the multi layers of hemicellulosic decorated polysaccharides that are masked by pectin components and lignin (63,195). The recognition of intact plant cell wall hydrophobic surfaces supports the CAZy to anchor and bind the crystalline cellulose and uronic acid sugars (223,333). This recognition step is an important as a primary progression in the microbial response to the plant cell walls that increase enzyme-substrate proximity (333). Nevertheless, a higher affinity towards the wheat straw, as seen in strain-6, suggests that this strain may present an interesting candidate that may be encoding hydrophobic surface interacting proteins in potential roles of enhancing lignocellulose degradation.

These findings create more opportunities to further investigate the evolution of *C. fimi* strains during the adaptation of cellulolytic bacteria in degrading wheat straw. A molecular level of study with transcriptomic (305,334) and proteomic analyses may be beneficial to identify any new upregulated genes and protein expression involved in biofilm-, cell wall- and exopolysaccharide-related proteins, as well as lipoproteins (335) secretion in *C. fimi* strains. This data may provide clearer evidence of the evolutionary changes in microbial communities that may be driven by a set of abiotic factors (336,337), for example the limitation of carbon source in wheat straw culture and as one non “social” adaptation strategies of microbes. Whole genome re-sequencing could provide an information of any single-Nucleotide-Polymorphism (SNP) (338) occurred during the wheat straw adaptation process in *C. fimi*. A CBM-linked assay (339) to test if more cellulose-binding proteins been expressed in *C. fimi* adapted strains than the wild type, may also hold potential. This research would contribute new knowledge which may fill the research gap about the biofilm formation (340,341), or new classes of CBMs and LPMOs expressed during wheat straw degradation the *C. fimi*.

## 7 Final discussion

More efficient lignocellulosic hydrolyzing enzymes are currently in demand for the cost-effective pretreatment of substrates for biofuel production. Although there are many cellulose-degrading enzymes that have been functionally characterised from *Cellulomonas fimi*, it is hypothesized that there are still many important enzymes involved in plant cell walls and biomass breakdown to be explored from this recently genome-sequenced species. The overall objective of this study was to identify novel lignocellulose processing enzymes from *C. fimi* that could be of benefit for the production of biofuels from second generation feedstocks.

For that purpose, a growth experiments were set up comprising *in-vitro* cultures supplemented with laboratory plant cell walls polysaccharides (Avicel and beechwood xylan) and two types of agricultural biomass (wheat straw and sugarcane bagasse). From the growth profiles analysis (Chapter 3), *C. fimi* was able to thrive in such hostile conditions where the source of carbon is limited. The ability of *C. fimi* to grow on insoluble wheat straw and sugarcane bagasse is intriguing, as to date *C. fimi* has been tested to grow on different type of amorphous substrates such as carboxymethyl cellulose (CMC) (144,342) and phosphoric acid swollen cellulose (PASC) (342), hence growing *C. fimi* on naturally lignocellulose biomass in this work creates a new perspective of this microbe's lignocellulose-degrading potential.

This study has explored, for the first time, the response of *C. fimi* growing on different type of polysaccharides using an RNA-seq method. The rRNA removal was the most crucial step prior to the sequencing work. In this work, several attempts of efficient rRNA removal revealed the challenges of dealing with a gram-positive, and the GC-rich genome of *C. fimi*. The differences of rRNA primer base, strand locations, and the different lengths of rRNA amplicons between gram-negative and gram-positive bacteria make the specificity of the probes to capture the unwanted sequence of rRNA in a particular strain a primary importance. A specifically suitable method that may be exclusive from strain to strain is crucial to provide strand information especially for highly GC-rich genes that may lack sufficient thymidine nucleotides for dUTP incorporation.

Until now, several RNA-seq studies have been published for species with a range of different GC-contents, however, the RNA-seq work of *C. fimi* described here may have used the GC-richest strain (74%) so far, based on the available literature that evaluates rRNA depletion methods. In Giannoukos *et al.* study (343), strains tested for rRNA depletion methods were *Prochlorococcus marinus*, *Escherichia coli*, and *Rhodobacter sphaeroides* with 30%, 50%, and 69% GC-content, respectively. From the study, Ribo-Zero™ kit from Epicentre® has been proven to be the most efficient kit for mRNA enrichment of those species based on the degradation of processed RNA by 5'-3' exonuclease that specifically digests RNA species with a 5'-monophosphate end. A different approach using a combination of MICROBExpress/Ovation rRNA removal method based on hybridization capture of rRNAs by antisense oligonucleotides followed by pull down through binding to magnetic beads has been tested by Peano and her co-workers in 2013 (344). This combination method has been proven to be suitable for RNA-seq of the whole transcriptome of *Burkholderia thailandensis* (67.7% of GC content) and produced a 238-fold mRNA enrichment, with more than 90% transcripts sequenced. As for *C. fimi*, two trials using the degradation of rRNA approach failed to provide sufficient quality of mRNA materials for sequencing, and hence, there was inadequate transcripts coverage. However, by using a method similar to that used for *B. thailandensis* with gene-specific primers, the unwanted rRNA was successfully diminished resulting in high quality mRNA and better coverage of the transcriptome library.

In order to understand the regulation of components in the cells responding to the different types of substrates, a proteomic study through in-gel tryptic digestion followed by liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) has been performed and presented in Chapter 4. By this approach, the joint analysis of the transcriptomic and proteomic data provides useful insights that may not be deciphered from individual analysis of mRNA or protein expressions. For that purpose, the differential gene expression (DGE) method was applied to discover, from successful RNA-sequencing, novel genes of *C. fimi* that are being co-expressed with CAZymes.

From the analysis, the highest CAZy-encoding gene that exclusively expressed in Avicel culture has been identified as the only one auxiliary activity family 10 (AA10) of *C. fimi*. The enzyme was upregulated almost 10-fold change compared to the rest of other substrates. Among the list of co-expressed genes that have been found to be upregulated in Avicel culture, Glucose sorbione dehydrogenase (GSDH), Celf\_2278 and Polycystic Kidney Disease (PKD)-containing protein (PKDP3),

Celf\_3039 are among the non-CAZy genes that appear to be co-expressed in the same condition. The proteomic study presented in Chapter 4 showed that the result from principal component analysis (PCA) was consistent with RNA-seq DGE outcomes as the outstanding genes (GSDH/PKDP1 and PKDP3) were also detected in the *C. fimi* proteome. A better scope of low complexity of *C. fimi* secretome dataset is achieved where more than 80% of 71 unique proteins, identified in three biological replicates, were found to be extracellular proteins, based on existing tools using bioinformatics analysis compared to a similar study by Wakarchuk *et al.* (86) (where the 'leakage' of intracellular proteins contamination were significantly detected from over 600 total proteins of two biological replicates). An interesting finding observed from *C. fimi* secretome is that a significantly higher amount of protein was detected in wheat straw and Avicel cultures towards day-3 of incubation. This is an unsurprising observation as *C. fimi* is a well-known cellulose degrader and as such, may produce significantly higher enzymes during Avicel utilization. However, *C. fimi* could also, potentially, digest more complex cellulose compounds such as lignocellulosic wheat straw (172).

Apart from AA10 and another 20 hypothetical proteins that are upregulated with other CAZymes, the GSDH of PKDP1 are among the more interesting proteins identified from the *C. fimi* transcriptome and secretome grown on Avicel. Protein domain prediction analysis has provided more information about predicted protein function and triggered further investigation of PKDP1 protein, which has a predicted domain PA14 as  $\beta$ -glucosidase. A site-directed mutagenesis experiment on chitinase A (ChiA) gene of *Alteromonas* sp. by Orikoshi *et al.* in 2005 uncovered the existence of PKD-domains that participate in the effective hydrolysis of powdered chitin (211).

Another recent study by Suma and Podile (214) revealed several chitinases-accessory domains (including PKD domain) other than the catalytic GH18 domain (214) in *Stenotrophomonas maltophilia* chitinase (StmChiA and StmChiB) genes. Moreover, a study by Horn *et al.* (35) shows the presence of a PKD-domain in other naturally occurring CBM33-containing proteins, which are diverse in their module families but have been shown experimentally to contain a PKD-domain specific for the chitin substrate. Another domain in PKDP1 of *C. fimi* has 100% identity to  $\beta$ -glucosidases and aligned to a C-terminal protective antigen 14 (PA14) domain, but the potential activity of this protein remains a mystery. Furthermore, the predicted oxidoreductase molecular function of the PQQ-superfamily protein domain in the multimodular PKDP1 protein structure makes this protein a more interesting candidate for further characterization.



Therefore, in Chapter 5, several attempts for heterologous expression in both prokaryote (*E. coli*) and eukaryote (*A. niger*) systems, as well as an attempt for a native isolation of PKDP1 from the culture media, were undertaken. In the recombinant protein expression trials, several methods have been used including the fusion-protein technique and a different expression hosts approach, however, the limitation of recombinant protein solubility remains a challenge and warrants further investigation and trials. A study in 1997 by Rajoka and Malik (345) revealed the responses of four *Cellulomonas* strains (*C. biazotea*, *C. flavigena*, *C. cellasea*, and *C. fimi*) in terms of  $\beta$ -glucosidase synthesis from different substrates (monosaccharides, disaccharides, cellobiose, CMC, xylan,  $\alpha$ -cellulose, steam alkaline-treated wheat straw and bagasse) in culture media were significantly different. Furthermore, in all strains,  $\beta$ -glucosidase accumulated intracellularly and was mainly located in the periplasmic fractions of the cells. In most bacteria,  $\beta$ -glucosidase is synthesized but usually remains associated with the cell (345). This may be explained that product inhibition or induction for secretion of  $\beta$ -glucosidase may be less effective.

The adaptive evolution on *C. fimi* in Chapter 6 was performed to get a better understanding of any behavioural changes of the bacteria during the adaptation phase in a wheat straw environment. This is the first study to investigate the effect of naturally occurring mutagenesis to produce wheat straw-adapted strains from a 52-weeks continuous subculture in wheat straw medium, using a serial transfer method. In this experiment, six populations of *C. fimi* were evaluated for their adaptation to the wheat straw liquid culture. The investigation included an analysis of growth profiles and enzyme activities of each adapted strain, and an analysis of the resulting degraded wheat straw. Results from the adaptive evolution experiment indicate that strain-6 may have a better adhesion to the wheat straw biomass; this may be a preference that indicates a better capability by this strain to perceive the complex decorated hemicellulose of polysaccharides that are concealed by pectin and lignin components. This research provides a framework for the exploration of the evolution of *C. fimi* strains during the adaptation of cellulolytic bacteria in degrading wheat straw.

A future study investigating the strain adherence to the biomass including a molecular level of study with transcriptomic (305,334) analysis would be interesting, and may be beneficial in the identification of any new upregulated genes. Furthermore, proteomic analysis may reveal protein expression involved in biofilm-, cell wall- and exopolysaccharide-related proteins, as well as lipoproteins (335) secretion in the adapted strain, which would improve our understanding of


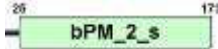

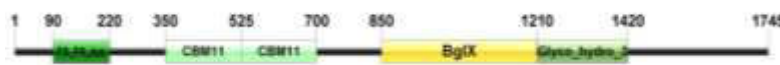

synergistic mechanism between multiple enzymes and accessory proteins. Whole genome re-sequencing could provide information of any single-Nucleotide-Polymorphism (SNP) (338) that occurred during the wheat straw adaptation process in *C. fimi*. Finally, a CBM-linked assay (339), to test if more cellulose-binding proteins are expressed in *C. fimi* adapted strains than the wild type, may also hold potential. In general, it seems that clearer evidence of the evolutionary changes in microbial adaptation may be driven by a set of abiotic factors such as the limitation of carbon source in wheat straw culture, and also as one of 'non-social' adaptation strategies of microorganisms (336,337). Although the ability of the bacterium to completely degrade lignocellulosic biomass e.g. wheat straw and sugarcane bagasse was still not ascertained, the studies presented in this thesis have gone some way towards enhancing the understanding of collective actions in the bacterium for an effective biomass-degradation system that requires the synergistic action of a large number and type of enzymes.

## Appendices










### Appendix A: Predicted structural domains of proteins identified in *C. fimi* secretome.

Diagram of protein domains were redrawn using Illustrator for Biological Sequences (IBS 1.0.1) software available at <http://ibs.biocuckoo.org/>.



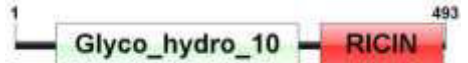



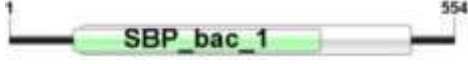



\*Due to limited space of layout, domains are not drawn to scale.

	PROTEIN DESCRIPTION	PROTEIN ID	CAZy DOMAINS	DOMAINS STRUCTURE
1	Exoglucanase	Exoglucanase	GH10, CBM2	
2	Membrane-flanked domain protein	Celf_0568	NOT CAZY	
3	Endo-1,4-beta-xylanase	Celf_0574	CBM22, CE4, CBM22, GH10, CBM9, CBM9	
4	Glycoside hydrolase family 3 domain protein	Celf_0583	CBM32, CBM11, CBM11, GH3	
5	Phosphate ABC transporter, periplasmic phosphate-binding protein	Celf_0591	NOT CAZY	











Appendix A continued...

6	Cellulose-binding family II	Celf_1754	CE2, CBM2	
7	Proteasome-associated ATPase	Proteasome-associated ATPase	NOT CAZY	
8	Extracellular ligand-binding receptor	Celf_1830	NOT CAZY	
9	Extracellular solute-binding protein family 5	Celf_1843	NOT CAZY	
10	Cobyrinic acid ac-diamide synthase	Celf_1876	NOT CAZY	
11	SCP-like extracellular	Celf_0742	NOT CAZY	
12	Cellulose-binding family II	Celf_1913	GH74, CBM2	
13	Glycoside hydrolase family 5	Celf_1924 (CenD)	GH5, CBM2	
14	Cobyrinic acid ac-diamide synthase	Celf_1938	NOT CAZY	









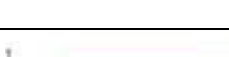

Appendix A continued...

15	Glycoside hydrolase family 16	Celf_3113	GH16, CBM13	
16	Alpha-N-arabinofuranosidase	Celf_3155	CBM13, GH62	
17	Endo-1,4-beta-xylanase	Celf_3156	GH10, CBM13	
18	Glycoside hydrolase family 18	Celf_3161	CBM2, GH18	
19	1, 4-beta cellobiohydrolase (CenA)	Celf_3184	CBM2, GH6	
20	Mannan endo-1,4-beta-mannosidase	Celf_0862	GH26, CBM23	
21	Extracellular solute-binding protein family 1	Celf_3272	NOT CAZY	
22	Glucan endo-1,3-beta-D-glucosidase	Celf_3330	GH64, CBM13	
23	Fatty acid desaturase	Celf_3356	NOT CAZY	
24	Lipoprotein	Celf_3360	NOT CAZY	









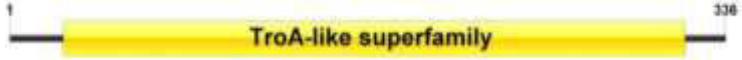
Appendix A continued...

25	Putative uncharacterized protein	Celf_2249	NOT CAZY	
26	LPXTG-motif cell wall anchor domain protein	Celf_3434	GH53, CBM61	
27	FG-GAP repeat protein	Celf_3440	PL11	
28	Extracellular repeat protein, HAF family	Celf_3445	NOT CAZY	
29	Alpha-1,6-glucosidase, pullulanase-type	Celf_1126	GH13, CBM48, GH13	
30	Peptidase S1 and S6 chymotrypsin/Hap	Celf_1127	NOT CAZY	
31	PKD domain containing protein	Celf_2278	NOT CAZY	
32	Fibronectin type III domain protein	Celf_2339	NOT CAZY	
33	Cell surface receptor IPT/TIG domain protein	Celf_3522	NOT CAZY	
34	NLPA lipoprotein	Celf_1210	NOT CAZY	

Appendix A continued...

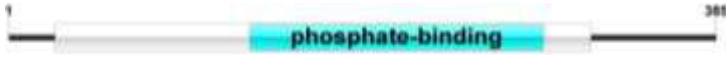






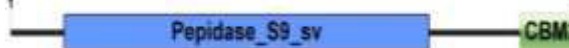

35	Glycoside hydrolase family 9	Celf_0019 (CenB)	GH9, CBM3, CBM2	
36	Glycoside hydrolase family 9	Celf_0045	GH9, CBM2	
37	1, 4-beta cellobiohydrolase	Celf_1230	GH6	
38	Extracellular solute-binding protein family 1	Celf_1290	NOT CAZY	
39	Extracellular solute-binding protein family 1	Celf_0084	NOT CAZY	
40	Endo-1,4-beta-xylanase	Celf_0088	GH10, CBM2	
41	Putative uncharacterized protein	Celf_0121	NOT CAZY	
42	Aminopeptidase Y	Celf_0132	NOT CAZY	
43	Putative F420-dependent oxidoreductase	Celf_1311	NOT CAZY	
44	Integral membrane sensor signal transduction histidine kinase	Celf_1318	NOT CAZY	

Appendix A continued...




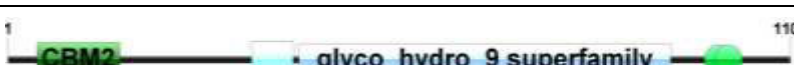





45	Pectate lyase	Celf_1340	CBM13,PL3	
46	Alkaline phosphatase	Celf_2549	NOT CAZY	
47	Endo-1,3(4)-beta-glucanase	Celf_3773	GH81, CBM6	
48	Pectate lyase/Amb allergen	Celf_3775	PL1	
49	Polynucleotide adenylyltransferase/metal dependent phosphohydrolase	Celf_3793	NOT CAZY	
50	Alpha/beta hydrolase fold protein	Celf_2590	NOT CAZY	
51	Chitin-binding domain 3 protein	Celf_0270	AA10, CBM2	
52	TAP domain protein	Celf_0295	NOT CAZY	
53	Periplasmic binding protein	Celf_1453	NOT CAZY	



Appendix A continued...

54	Putative ABC transporter substrate binding protein	Celf_1461	NOT CAZY	
55	Putative uncharacterized protein	Celf_2704	NOT CAZY	
56	Extracellular solute-binding protein family 1	Celf_2729	NOT CAZY	
57	Triacylglycerol lipase	Celf_0369	NOT CAZY	
58	Glycoside hydrolase family 11	Celf_0374	GH11,CBM2,CE4, CBM2	
59	Polar amino acid ABC transporter, inner membrane subunit	Celf_1572	NOT CAZY	
60	Cellulose-binding family II	Celf_0403	CBM2	
61	Cellulose-binding family II	Celf_0404	CE1, CBM2	
62	5'-Nucleotidase domain-containing protein	Celf_1653	NOT CAZY	

Appendix A continued...

63	Putative RNA polymerase, sigma-24 subunit, ECF subfamily	Celf_0474	NOT CAZY	
64	Putative uncharacterized protein	Celf_0523	NOT CAZY	
65	Peptidase S8 and S53 subtilisin kexin sedolisin	Celf_0539	NOT CAZY	
66	Glycoside hydrolase family 9	Celf_1705	CBM4, CBM4, GH9	
67	Extracellular solute-binding protein family 5	Celf_2906	NOT CAZY	
68	Endoglucanase C (cellobiohydrolase B)	cenC (Cbp120;CbhB; CenE)	CBM4, GH9 (Cel9B) GH48, CBM2	
69	Exoglucanase A (cellobiohydrolase A (CbhA;Celf_1925)	cbhA celf_1925	GH6, CBM2	
70	Exoglucanase B 4-β-cellobiohydrolase B CBP120 (Exocellobiohydrolase B)	cbhB celf_3400 (cbhB/cenE)	GH48/CBM2/ FN3	
71	Beta 1,4-xylanase	Beta 1,4-xylanase	GH10 (Cellulase F) Cfx	

APPENDIX B

Average molar %:



Appendix B: List of 71 proteins identified in *C. fimi* secretome grown in four types of carbon sources.

	UniProt Accession	Protein Name	CAZy	Avicel	Beechwood xylan	wheat straw	Sugarcane bagasse	Signal peptide or alternative secretion pathway	Uncharacterized /Potentially new CAZy in <i>C. fimi</i>
1	β-1,4-xylanase	B 1,4-xylanase	endo-1,4-β-xylanase (Cfx/Xyn10B)					Yes	
2	Celf_1925	Exoglucanase A (GH6/CelB)	Cellobiohydrolase A (cbhA)					Yes	
3	Celf_3400	Exoglucanase B	Cellobiohydrolase B (Cbp120;CbhB;CenE) (Cel48A)					Yes	
4	Celf_0019 / CenB	Glycoside hydrolase family 9	endo-β-1,4-glucanase B (CenB;Celf_0019) (Cel9A)					Yes	
5	Celf_0045	Glycoside hydrolase family 9	GH9, CBM2					Yes	
6	Celf_0084	Extracellular solute-binding protein family 1	Not in CAZy database					Yes	
7	Celf_0088	Endo-1,4-β-xylanase	GH10, CBM2					Yes	
8	Celf_0121	Putative uncharacterized protein	Not in CAZy database					Yes	Predicted as chitinase based on protein crystal structure

Appendix B continued..

9	Celf_0132	Aminopeptidase Y	Not in CAZy database					Yes	
10	Celf_0270	Chitin-binding domain 3 protein	AA10, CBM2					Yes	AA10 of <i>C. fimi</i>
11	Celf_0295	TAP domain protein	Not in CAZy database					Yes	
12	Celf_0369	Triacylglycerol lipase	Not in CAZy database					Yes	
13	Celf_0374	Glycoside hydrolase family 11	GH11, CBM2, CE4					Yes	
14	Celf_0403	Cellulose-binding family II	CBM2					Yes	
15	Celf_0404	Cellulose-binding family II	CE1, CBM2					Yes	
16	Celf_0474	Putative RNA polymerase	Not in CAZy database					No	
17	Celf_0523	Putative uncharacterized protein	Not in CAZy database					No	
18	Celf_0539	Peptidase S8 and S53 subtilisin kexin sedolisin	Not in CAZy database					Yes	
19	Celf_0568	Membrane-flanked domain protein	Not in CAZy database					No	
20	Celf_0574	Endo-1,4- $\beta$ -xylanase (Xyl10B, XynC)	CBM22, CE4, CBM22, GH10, CBMnc, CBM9					Yes	

Appendix B continued..

<b>21</b>	Celf_0583	Glycoside hydrolase family 3 domain protein	CBM32, CBM11, CBM11, GH3					Yes	Wakarchuk <i>et al.</i> , 2016
<b>22</b>	Celf_0591	Periplasmic phosphate-binding protein	Not in CAZy database					Yes	
<b>23</b>	Celf_0742	SCP-like extracellular	Not in CAZy database					Yes	
<b>24</b>	Celf_0862	Mannan endo-1,4- $\beta$ -mannosidase	GH26, CBM23					Yes	
<b>25</b>	Celf_1126	Alpha-1,6-glucosidase, pullulanase-type	GH13, CBM48, GH13					Yes	
<b>26</b>	Celf_1127	Peptidase S1 and S6 chymotrypsin/Hap	Not in CAZy database					Yes	
<b>27</b>	Celf_1210	NLPA lipoprotein	Not in CAZy database					Yes	
<b>28</b>	Celf_1230	1, 4- $\beta$ cellobiohydrolase (Cel6C)	GH6					Yes	
<b>29</b>	Celf_1290	Extracellular solute-binding protein family 1	Not in CAZy database					Yes	
<b>30</b>	Celf_1311	Putative F420-dependent oxidoreductase	Not in CAZy database					No	
<b>31</b>	Celf_1318	Integral membrane histidine kinase	Not in CAZy database					Yes	
<b>32</b>	Celf_1340	Pectate lyase	CBM13, PL3, CBM13					Yes	

Appendix B continued..

33	Celf_1453	Periplasmic binding protein	Not in CAZy database					Yes	
34	Celf_1461	Putative substrate binding protein	Not in CAZy database					Yes	
35	Celf_1572	Polar amino acid ABC transporter	Not in CAZy database					No	
36	Celf_1653	5'-Nucleotidase domain-containing protein	Not in CAZy database					Yes	
37	Celf_1705	Glycoside hydrolase family 9	CBM4, CBM4, GH9					Yes	
38	Celf_1754	Cellulose-binding family II	CE2, CBM2					Yes	
39	Celf_1830	Extracellular ligand-binding receptor	Not in CAZy database					Yes	
40	Celf_1843	Extracellular solute-binding protein family 5	Not in CAZy database					Yes	
41	Celf_1876	Cobyrinic acid ac-diamide synthase	Not in CAZy database					No	
42	Celf_1913	Cellulose-binding family II	GH74, CBM2					Yes	
43	Celf_1924	Glycoside hydrolase family 5	endo- $\beta$ -1,4-glucanase D (CenD;Celf_1924) (Cel5A)					Yes	
44	Celf_1938	Cobyrinic acid ac-diamide synthase	Not in CAZy database					Yes	

Appendix B continued..

45	Celf_2249	Putative uncharacterized protein	Not in CAZy database					Yes	
46	Celf_2278	PKD domain containing protein	Not in CAZy database					Yes	Interesting predicted domains
47	Celf_2339	Fibronectin type III domain protein	Not in CAZy database					Yes	Potentially new xylanase
48	Celf_2549	Alkaline phosphatase	Not in CAZy database					No	
49	Celf_2590	Alpha/ $\beta$ hydrolase fold protein	Not in CAZy database					No	
50	Celf_2704	Putative uncharacterized protein	Not in CAZy database					Yes	
51	Celf_2729	Extracellular solute-binding protein family 1	Not in CAZy database					Yes	
52	Celf_2906	Extracellular solute-binding protein family 5	Not in CAZy database					Yes	
53	Celf_3113	Glycoside hydrolase family 16	GH16, CBM13					Yes	Wakarchuk <i>et al.</i> , 2016
54	Celf_3155	Alpha-N-arabinofuranosidase	CBM13, GH62					Yes	
55	Celf_3156	Endo-1,4- $\beta$ -xylanase	GH10, CBM13					Yes	
56	Celf_3161	Glycoside hydrolase family 18	CBM2, GH18					Yes	Wakarchuk <i>et al.</i> , 2016
57	Celf_3184	1, 4- $\beta$ cellobiohydrolase (GH6)	endo- $\beta$ -1,4-glucanase A (CenA;Celf_3184) (Cel6A)					Yes	
58	Celf_3272	Extracellular solute-binding protein family 1	Not in CAZy database					Yes	

Appendix B continued..

59	Celf_3330	Glucan endo-1,3- $\beta$ -D-glucosidase	GH64, CBM13					Yes	Wakarchuk <i>et al.</i> , 2016
60	Celf_3356	Fatty acid desaturase	Not in CAZy database					No	
61	Celf_3360	Lipoprotein	Not in CAZy database					Yes	
62	Celf_3434	LPXTG-motif cell wall anchor domain protein	GH53, CBM61					Yes	
63	Celf_3440	FG-GAP repeat protein	PL11					Yes	
64	Celf_3445	Extracellular repeat protein, HAF family	Not in CAZy database					Yes	
65	Celf_3522	Cell surface receptor IPT/TIG domain protein	Not in CAZy database					Yes	
66	Celf_3773	Endo-1,3(4)- $\beta$ -glucanase	Cfx, CBM6					Yes	Wakarchuk <i>et al.</i> , 2016
67	Celf_3775	Pectate lyase/Amb allergen	PL1					Yes	
68	Celf_3793	Metal dependent phosphohydrolase	Not in CAZy database					No	
69	cenC/Cel9B	Endoglucanase C	GH9					Yes	
70	Exoglucanase	Exoglucanase	No match in CAZy database					Yes	
71	Proteosome-associated ATPase	Proteosome-associated ATPase	Not in CAZy database					No	



APPENDIX C

Appendix C: Multiple sequences alignment of *C. fimi* PKD-domain containing proteins.

CLUSTAL O(1.2.4) multiple sequence alignment

Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	----MTRSRGPVARSALLAAAALTFAGV-TLATPAAADTVPVAGTTVTVAADSLPTVQI	55
Celf_3039-PKDP3	MSLRPVADRSPVARTWAVLTSVALVAVGLVAGSAPAQADTAPPAGLPPTVAADALPTVQI	60
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	DGVANQQGVAGSKVVFVGGDFANARPAGSAGTNLVPRANLLAYDLTTGVLDAWAPDPNA	115
Celf_3039-PKDP3	DGVVWQQALSGNLVYAGGEFSNARPAGNAAGVGNVPRANLLRFQVRTGVLDPWAPNPMG	120
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	QVRAVAVSPDGSRVYVGGSFITTIAGQPRYRIAAFDAAATGALVSSFAPGVDSQVRAIAATD	175
Celf_3039-PKDP3	QVRAVVKSVDGSRIVYVGGSFITISGVARYRIAAFDVAVTGALITTFNAGANGQVRALAAATN	180
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	TAVYVGGFTTNANGEPKSKVAAFSAADGSLLPWAPSADNGSVSAMVVSQDRVQIVVGGSF	235
Celf_3039-PKDP3	TTVYAGGIFTQAGSSSRTRLAAFNASNGALLPWNPTVDDGVSVALTVAPDGGTVIVVGGNF	240
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	TTFNGSGNPGYGLARVDATTGATLPLAINGLIRNGSTQASITSVTSDATSFVYVYVGGF	295
Celf_3039-PKDP3	TSFNGSTAAPDGLARVDAVTGAALPFPAATSQIRNGGTNGSILGLTGDADNMYGVYVWGR	300
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	T-GNLEGTARGDWTGALVWVEDCHGDTYAAWPGPDVAVYAGHPHYCGNIGGFPQTEPW	354
Celf_3039-PKDP3	SGGTLEGVFAADWATGSIKRWIADCHGDSYAVH-AQGPVVYAASHHHYCGNVGGTPQNDW	359
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----	0
Celf_3038-PKDP2	TFYRGLAFSKSVTGTATADPYGYHNYAGQPTPSLLAWYPDINSGSYTGQGGPWTVTGND	414
Celf_3039-PKDP3	LFYRADAYTKPMRKLGREHLGYTNFEGTAAPMLNFYPLDGTGFTGQNGQPWAVTGDD	419
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	-----MPVPA	5
Celf_3038-PKDP2	QYLLYGGFEFTIVNKGQQGLARFARTDVAPNKQSPRLASATWPLAAASFQAGTVRLSWPA	474
Celf_3039-PKDP3	RYIVYGGFEFLNVNFRQQGLVRFVFAIPSIAPNDQGPFRFSGADFTPTARAIGAGTVRISWTA	479
Celf_2278-PKDP1	-----	0
Celf_2091	-----	0
Celf_1062	SAR-----HERPEH-----	14
Celf_3038-PKDP2	NYDRDNENLTYTVSRNGQVVHTVTAASVWKRQMGFLDTGLTPGASYTYRVKAVDPFGN	534
Celf_3039-PKDP3	NADFDNEDLTYTVLRNGQLPAQVKKSQIWMRPTQGVTDGTLTAGQTYTYAVRVSDPFGN	539
Celf_2278-PKDP1	-----MNRWRS-----LASVVTGALVAT	18
Celf_2091	-----	0
Celf_1062	-----PRRLAHARAVVPGPRRHGRRL--AGTLAL--V--GAVLAGGLVA	52
Celf_3038-PKDP2	QAITDQVQVTVSDASSTSAAYAQQVAADGATSLWRLGEPGGTTAFDHWGWSDLTLGTGVTR	594
Celf_3039-PKDP3	TVTSPSVTATANGSGAVSAYAQTVMADGASHLWRLGESSGTTALDTAGTDDGSGTGTALTR	599

Appendix C continued...

Celf_2278-PKDP1	GLAV-----AGTPATAAVP-----AGFSESLVASVPNPSAIAFTADGRML	59
Celf_2091	-----	0
Celf_1062	APAGAADPE--PF FVVLPDTQNYVSTSTNQATMGVQTQWIADHRDDLIAFVSVQVGD	109
Celf_3038-PKDP2	GAAGAIIGGDANAASTFDGTSTASGATSTPVQGPDTFAVEAWVRTTST-----RGGKI	646
Celf_3039-PKDP3	GTPGAIAGDSSSATTFAGTTNSRVVAPQREQNLNELSVEAWIRTTSN-----QGGWV	651
Celf_2278-PKDP1	VTQSGRLRVRTAAGTLLATPALDLASRLCTNSERGLLGVA TDPDPATRAIYLFYARTG	119
Celf_2091	-----	0
Celf_1062	VGVDTAEVQ-----W-----QRASQHMAVLDAAGVP	135
Celf_3038-PKDP2	VGFSNQ--Q-----TG--QSGSYDRHVYMDDTGRIT	673
Celf_3039-PKDP3	VGFGNSTSL-----TG--TSTSRDRQLYVDSAGRVQ	680
Celf_2278-PKDP1	TSCPTSQGGTPAGAPVNRVSRFVLGDDNLVDPASETVLLDGIAS PAGNHNAG---DLHV	175
Celf_2091	-----MHAMHWRAPARTL	13
Celf_1062	NAV-----LPGNHMDLTTGDA-----PLFRQYFPPSRYANASHWNGAAASY	176
Celf_3038-PKDP2	FGV-----YPGA-----VRTITTNASYNDDGQHHHVVASM	702
Celf_3039-PKDP3	FGA-----SPGQ-----NRTVRS PGAVNDGQHHHVVTM	709
Celf_2278-PKDP1	GKDG Y-LYVTTG-----DGGCDYRGDSGCGGONDASRDR---NVLLGKVL RVDRTT	222
Celf_2091	VRTR-----TRTHTRTRI-----RTG-----AGTGRSR---TAH	39
Celf_1062	G-----GHLG---QHV	184
Celf_3038-PKDP2	GAEGMKLFVDGRRRAGQRADTTYGQAYTYGNRI GGDNVGGWPNQPTSGFLAGSID--DVA	759
Celf_3039-PKDP3	GATGMALYVDGALVASRADNPSGRNFTGFWRIGGDSLSGWPNTPASNNFNGAID--EVA	766
	*	
Celf_2278-PKDP1	GAPAPG-----NPFLGT-----	234
Celf_2091	ATPRRRAGRVGLVALLATTAVLVLPVAPAAAAEPDLAPSGLLFGAYVAQRSAPSVEA	99
Celf_1062	VGPDVDRQNMDSFALFSAG-----GMDFLLLSL-----EM	215
Celf_3038-PKDP2	VYPAPLTVAQVEQHYTL SGR-----T-----VA	782
Celf_3039-PKDP3	VYFPPLSAG-----	775
	* *	
Celf_2278-PKDP1	-----GTASCR LAPAAPGTVCRETFAWG-----	257
Celf_2091	ATAAF EERVGRRELELRWYSRWDET-----MPPSG---LRASVAAGRTPVLSI--	144
Celf_1062	NAPDYALDWGRRVLAAYPERRA-----	237
Celf_3038-PKDP2	KPPAPADAYGKAVYDAGPELYWRLSEGSGTTAADS GPYGMPTYRGTF LRGGSGALSGVS	842
Celf_3039-PKDP3	-----	775
Celf_2278-PKDP1	-----LRNPFR	263
Celf_2091	--EPRTL AG-----TRLSWASVARGDHDARIRAQATGVASLGVPVLLA	185
Celf_1062	-----IVA	240
Celf_3038-PKDP2	NPAVRTLGTSNTLVASSQQFSNPTVYSQELWFRSTTTNGGKLI--GFGDKDTGVSGSYD	899
Celf_3039-PKDP3	-----	775
Celf_2278-PKDP1	FAFDPDASG-----TVFHVNDV GQNVWEEIDLGT PGADYGVWPVREGHC	306
Celf_2091	FHHEPELAGTHGTPAEFRAAWRRYVEV-FRAAGVTNVLWTWIVTPTVL-----	232
Celf_1062	THSYVDTAGGL-----TTQVAR-----ADGGNSGRAIWEELVRPSCS-----	277
Celf_3038-PKDP2	RHVYMETDGRL-----TFGTWGTVNTITTTASYNDSQWHLVATQGP-----	942
Celf_3039-PKDP3	-----	775
Celf_2278-PKDP1	AQTGSATDCGGALPAGMTNPIHDYGRSTGCGSITGGAFVPDGVWPAAYEGGYLFS DYNCG	366
Celf_2091	-----G---GT-----TGPTATDLYPGDDVVD-----	251
Celf_1062	-----VFLVVS GHFSE-----G	289
Celf_3038-PKDP2	-----DGMRLYVDGV-----	952
Celf_3039-PKDP3	-----	775
Celf_2278-PKDP1	RLMMLRGGTRTDVATGLGAHVHL---EF--GPWSGTQALY-YTTYANGGEIRR-----	413
Celf_2091	-RVGLD---AY-NWAGCAPGV-----PTAWRSMA TIAGPARDFA-----	285
Celf_1062	DLGEAR---RT-DTNACGDPVHAVLTDYQDRARGDGLRYYTFDPAAEIRATTYSPTL	345
Celf_3038-PKDP2	-LQG-----T-HPQTA AQAYNGYWRVFGDTTWGPQAWFA-----GTIDEV--AVYSSVL	997
Celf_3039-PKDP3	-----	775



Appendix C continued...

Celf_2091	-----RAHGKPLLLAEWGSVEDAADPGRKAAWLRE-----SLATLAS	322
Celf_1062	DRYETDADSSFTLPYAMTGVPPD-POPGLAAD-----TFGRRTLGGWGGADTGGGA	395
Celf_3038-PKDP2	DAQTVANHHA----LGTGGQLPN-Q--LPVAAFSTPTDLAVAFDATGSSDDPDGTVASYA	1050
Celf_3039-PKDP3	---TIAAHT----LGRGTGAPE-Q--PPVAAF TTATDGLKVTVDGTGSDPDGTVASYA	825
	. * * :	
Celf_2278-PKDP1	WQFGDGTDPDATTTTPTVQHTYAA-GTWATLRVRDPQGATSAAVTARIT-SGNTAPTAQI	517
Celf_2091	WPEVEGALYFHQHGTCPWNT-----DSTATTAFAAFEIANSPHTAHGRSAW	368
Celf_1062	WSVGGGSSRFVAVGGGTGQQSPAPGGTVTSTLGSVQSTADVTASVALDRIPSNAL-----	450
Celf_3038-PKDP2	WAFGDGATAT--GATPHTFAAGGTYTVTLTVTDDDGATAQVAHDVTVPPNQLPTAGF	1107
Celf_3039-PKDP3	WAFGDGATGT---GPTASRTYAAAGTYTVTLTVTDDDGNTAQTSHDVTV-----	871
	* * : : . . :	
Celf_2278-PKDP1	TS-----PAAGATF	526
Celf_2091	LRASTVLGGAPLAVAFDASRSAGAGTATGSGVTTWRLDPGDGTPTTGTGTPPAVLAHTY	428
Celf_1062	----YATVS-----GR-----V	458
Celf_3038-PKDP2	----TATPQDLTVAFDGGVSTDPD--GTITQHAWTFGDG--T---TGE---GPAPHTHY	1152
Celf_3039-PKDP3	-----	871
Celf_2278-PKDP1	VVGQTYT LSGSATDAQDGLPGSRLSNTV-----VRVHDQHTHPFL	567
Celf_2091	RTAGTFRARLDVTDAAAGTATDQRTVVAAPTLEVA-----	466
Celf_1062	VGSGDYGARLKVLATGAVQLHTERSGLTGGTLPGVSLTSGERLRVVRVQVEGTGP----	514
Celf_3038-PKDP2	AAGGDYEVTLTVTDDRGSSAQVVRTVTVVAPNQAPVAAF-DAVATDLTVAVDGSASTDAD	1211
Celf_3039-PKDP3	-----	871
Celf_2278-PKDP1	GPVTGSAVSFQAPGPEDLA-----	586
Celf_2091	-----	466
Celf_1062	TVVRVRAWEDGQPEPSTWQHTATDAPALQGAGGVRLMTYVVSSTTTGGALTVRWD-LLA	573
Celf_3038-PKDP2	GLTGHANDFGDGT-----TGTGATASHTYGAAGTYTVTLTVTDDGGETG	1256
Celf_3039-PKDP3	-----	871
Celf_2278-PKDP1	-----AAA-----	589
Celf_2091	-----RAVTTTGA-----ELVAVVDPDGLTGMVRF EWGTSTAYGAT	502
Celf_1062	TRIGAAPPPPPVNPQPPVASFVATDGLTVAVDASGSSDDPGIVVARSNAFGDGAT----	629
Celf_3038-PKDP2	TTTREVAVAAPVNPQGPVAAFATPHTHLAVAVDASASTDADGQVYAYANEFGDSAT----	1312
Celf_3039-PKDP3	-----TAPPPNPQPPVASFATATGLTVQVDAGASSDDPGTVTQRSNVDFGDGTT----	919
	: . :	
Celf_2278-PKDP1	-----NSHLRVTLTATDAQGATTTVTRDFLPRRVAATLATSP	626
Celf_2091	ATASVPAVTYAATTPVATGLAPSTTYAVRVTAATAAGTTTRSTTF-----	548
Celf_1062	ASGTTAARTYAA-----AGTYTVTLTVTDDDGAVASTARA-----	664
Celf_3038-PKDP2	ATGATASHTYAV-----DGYTITLVTTDDDGDTGTTTRE-----	1347
Celf_3039-PKDP3	ATGTTASHTYAA-----DGYTVTLTVTDDDGATAQTTRA-----	954
	: : . : * . :	
Celf_2278-PKDP1	AGRTLTVNGQVTGPTTVTSWAGFDLRLTVPSQRDAQGRTYELDGSWSDGS-TAARTWTT	685
Celf_2091	-----STAGPTTSTTWASGARTST-----TLGATVNPRLAATS AWFE	587
Celf_1062	-----VTVSAPPVGVLAADPFDRV-----TGGWG---AAATGGPWTV	700
Celf_3038-PKDP2	-----VTVAAPPADTFALDPFARV-----TGGWG---AADTGGAWSV	1383
Celf_3039-PKDP3	-----VTVTTPPADAPFAADAFARTV-----SGGWG---TADTGGAWSV	990
	* . * * : :	
Celf_2278-PKDP1	PASSTTLTATL-----GLRGLRAVYHDNADLTGAT-VTRIDPAVAFDWGLAAP	732
Celf_2091	WGPTAA--LGAGTAPVAL SAVGY-----DTLTTAVAGLAPGTTYHYRLVAR	632
Celf_1062	AGGASRFSVAAGAGAMQVPAGATL TAGL GAVSGTSDLTASLALASVPDGPPLY-ATLSGR	759
Celf_3038-PKDP2	TGGAANFSVAAGTGAMRVGTGAGYRLSSFLPVSSSTSDLTAKVALDVMPGTGAGTDL ELAGR	1443
Celf_3039-PKDP3	AGGATNFSVASGVGAMRVGTGAGFRLSGFLPVSSSTADVRVDVALDAMPTGGGTDL EVAGR	1050
	. : :	

Appendix C continued...

Celf_2278-PKDP1	VS-----GIGADTFSVRWSGSVVPRYS----QTY-----TF	759
Celf_2091	NAHG-----SV-----AGPVRT-VTTSR-----	649
Celf_1062	VVGGAD-LGARVKVLTGGAVQLHTE---RTGTV-LTGGTLPGVVLTPGARLRVRVQVQG	813
Celf_3038-PKDP2	TAGTTDGYRLRLKMLATGVVRRASLVGISAGSTTT-VAQVNVPLTYTAGQTLQVRLQVDG	1502
Celf_3039-PKDP3	TVGTTDGYRRLKMLSTGVVRRASIVGISAGTTTT-VAQVNVPLTYTAGQTLVSRFQADG	1109
.		
Celf_2278-PKDP1	ATTSODGVRLWVWVDTLVLDQNTNHSRRVDTGTVALTAGQAVPIVLEYFDGVRNAVAELRW	819
Celf_2091	-----	649
Celf_1062	TAPTTVRARAWLEGSPEPSTWQYAATD---GTAALQAAGSVREMSYLSSSATTGPVTVRW	870
Celf_3038-PKDP2	TGTTALRAKVVAAGTPEPAANTLQGTS---TTAALQVGGIGLSVYTSSTSTTLPLTARW	1559
Celf_3039-PKDP3	TGTTALRLKVNVPAGTPEPAANTLTGSS---TTAALQVAGGVGLSVYTSSTTTTLPLTARW	1166
Celf_2278-PKDP1	SSTSQASEIVPTTRLRP	836
Celf_2091	-----	649
Celf_1062	DDVLVTPA-G-----	879
Celf_3038-PKDP2	SELAARPV-VP-----	1569
Celf_3039-PKDP3	SQLTARPV-PA-----	1176

## APPENDIX D

Appendix D: Sequence of condon-optimized *C. fimi* PKDP1 protein synthesized by GeneArt®.

15ABMFFP\_1711803.gb

```

      ....|....| ....|....| ....|....| ....|....| ....|....|
          10      20      30      40      50
LOCUS CTAAATTGTA AGCGTTAATA TTTTGTAA AATCGCGTTA AATTTTGT
      ....|....| ....|....| ....|....| ....|....| ....|....|
          60      70      80      90     100
LOCUS AAATCAGCTC ATTTTAAAC CAATAGCCG AAATCGGCAA AATCCCTTAT
      ....|....| ....|....| ....|....| ....|....| ....|....|
         110     120     130     140     150
LOCUS AAATCAAAAG AATAGACCGA GATAGGGTTG AGTGGCCGCT ACAGGGCGCT
      ....|....| ....|....| ....|....| ....|....| ....|....|
         160     170     180     190     200
LOCUS CCCATTCGCC ATTCAGGCTG CGCAACTGTT GGAAGGGCG TTTCGGTGCG
      ....|....| ....|....| ....|....| ....|....| ....|....|
         210     220     230     240     250
LOCUS GGCCTCTTCG CTATTACGCC AGCTGGCGAA AGGGGATGT GCTGCAAGGC
      ....|....| ....|....| ....|....| ....|....| ....|....|
         260     270     280     290     300
LOCUS GATTAAGTTG GGTAACGCCA GGGTTTCCC AGTCACGACG TTGTAACG
      ....|....| ....|....| ....|....| ....|....| ....|....|
         310     320     330     340     350
LOCUS ACGCCAGTG AGCGCGACGT AATACGACTC ACTATAGGGC GAATTGGCGG
      ....|....| ....|....| ....|....| ....|....| ....|....|
         360     370     380     390     400
LOCUS AAGGCCGTCA AGGCCGATT CTAGAGTCCC TGCCGGCTTC TCCGAGTCCC
      ....|....| ....|....| ....|....| ....|....| ....|....|
         410     420     430     440     450
LOCUS TCGTCGCCTC CGTCCCCAAC CCTCCGCTA TCGCCTTAC CGCCGATGGC
      ....|....| ....|....| ....|....| ....|....| ....|....|
         460     470     480     490     500
LOCUS CGCATGCTCG TCACCCAGCA GTCCGGTCGC CTCCGCTCC GGACCGCCGC

```

```

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      510      520      530      540      550
LOCUS TGGCACCCTC CTCGCCACCC CTGCCCTCGA TCTCGCCTCC CGCCTCTGCA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      560      570      580      590      600
LOCUS CCAACTCCGA GCGCGGCCTC CTCGGCGTCG CCACCGATCC CGATCCCGCC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      610      620      630      640      650
LOCUS ACCCGCGCCA TCTACCTCTT CTACACCGCC CGCACCGGCA CCAGCTGCCC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      660      670      680      690      700
LOCUS CACCTCCCAG GGCGGCACCC CCGCTGGCGC CCCTGTCAAC CGCGTGTCCC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      710      720      730      740      750
LOCUS GCTTCGTCCT CGGCGACGAT AACCTCGTCG ATCCCGCCTC CGAAACCGTC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      760      770      780      790      800
LOCUS CTCCTCGATG GTATCGCCTC CCCTGCCGGC AACCACAACG CCGGCGATCT

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      810      820      830      840      850
LOCUS CCACGTCGGC AAGGATGGCT ACCTCTACGT CACCACCGGC GACGGCGGCT

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      860      870      880      890      900
LOCUS GCGATTACCG CGGCGATTCC GGCTGCGGCG GTGATAACGA TGCCTCCCGC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      910      920      930      940      950
LOCUS GATCGCAACG TCCTCCTCGG CAAGGTCTC CGCGTCGATC GCACCACCGG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
      960      970      980      990     1000
LOCUS CGCTCCCGCC CCTGGCAACC CCTTCCTCGG CACCGGCACC GCCAGCTGCC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
     1010     1020     1030     1040     1050
LOCUS GCCTCGCCCC TGCCGCTCCC GGAACCGTCT GCCGCGAAAC CTTCGCCTGG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
     1060     1070     1080     1090     1100
LOCUS GGCTCCGCA ACCCTTCCG CTTCGCCTTC GATCCCGATG CCTCCGGCAC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
     1110     1120     1130     1140     1150
LOCUS CGTGTTCCAC GTCAACGATG TCGGCCAGAA CGTCTGGGAG GAAATCGATC

```

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1160 1170 1180 1190 1200  
**LOCUS** TCGGCACCCC TGGCGCCGAT TACGGCTGGC CCGTCCGCGA GGGCCACTGC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1210 1220 1230 1240 1250  
**LOCUS** GCCCAGACCG GCTCCGCCAC CGATTGCGGC GGTGCCCTCC CTGCCGGCAT

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1260 1270 1280 1290 1300  
**LOCUS** GACCAACCCC ATCCACGATT ACGGCCGCTC CACCGGCTGC GGCTCCATCA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1310 1320 1330 1340 1350  
**LOCUS** CCGGCGGTGC CTTCTGCCCC GATGGCGTCT GGCCTGCCGC CTACGAGGGC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1360 1370 1380 1390 1400  
**LOCUS** GGCTACCTCT TCTCCGATTA CAACTGCGGT CGCCTCATGA TGCTCCGCGG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1410 1420 1430 1440 1450  
**LOCUS** TGGCACCCGC ACCGATGTCG CCACCGCCT CGGCGCTGCC GTCCACCTCG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1460 1470 1480 1490 1500  
**LOCUS** AGTTCGGCCC CTGGTCCGGC ACCCAGGCC TCTACTACAC CACCTACGCC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1510 1520 1530 1540 1550  
**LOCUS** AACGGCGGCG AGATCCGCCG CTTGGCCTAC ACCGGAACCG CCAACCGCAC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1560 1570 1580 1590 1600  
**LOCUS** CCCCACCGCC GCTCTCACCG CCTCCCCAC CTCCGGCGCT GCCCCCCTCA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1610 1620 1630 1640 1650  
**LOCUS** CCACCACCCT CGATGGCCGC GGTTCCTCCG ATCCCGATGG CGGCACCCTC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1660 1670 1680 1690 1700  
**LOCUS** ACCTACCTCT GGCAGTTCGG CGACGGCACC CCCGATGCCA CCACCACCAC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 1710 1720 1730 1740 1750  
**LOCUS** CCCCACCGTC CAGCACACCT ACGTGCCGG CACCTGGACC GCCACCCTGC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|

1760 1770 1780 1790 1800  
**LOCUS** GCGTCCGCGA TCCCCAGGGC GCCACCTCCG CCGCTGTGAC CGCCCGCATC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 1810 1820 1830 1840 1850  
**LOCUS** ACCTCCGGCA ACACCGCCCC CACCGCCCAG ATCACCTCCC CCGTGCCGG  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 1860 1870 1880 1890 1900  
**LOCUS** CGCTACCTTC GTCGTCGGC AGACCTACAC CCTCTCCGGC AGGCCACCG  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 1910 1920 1930 1940 1950  
**LOCUS** ATGCCCAGGA TGGCACCTC CCCGGCTCC GCCTCTCCTG GACCGTCGTC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 1960 1970 1980 1990 2000  
**LOCUS** CGGTCCACG ATCAGCACAC CCACCGTTC CTCGGCCCCG TTACCGGTC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2010 2020 2030 2040 2050  
**LOCUS** CGCTGTCTCC TTCCAGGCC CTGGCCCCGA GGATCTGGCC GCTGCCGCCA  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2060 2070 2080 2090 2100  
**LOCUS** ACTCCCACCT CCGCGTCACC CTCACCGCCA CCGACGCCA AGGCGTACC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2110 2120 2130 2140 2150  
**LOCUS** ACTACCGTGA CCCGCGATTT CCTGCCTCGC CGCGTCGCCG CCACCTCGC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2160 2170 2180 2190 2200  
**LOCUS** CACCTCTCCC GCTGGCCGA CCCTGACCGT CAACGGCCAG ACCGTCACCG  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2210 2220 2230 2240 2250  
**LOCUS** GCCCCACCAC CGTCACCAGC TGGGCCGGCT TCGATCTCCG CCTCACCGTC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2260 2270 2280 2290 2300  
**LOCUS** CCCTCCCAGC GCGACGCCA GGGCCGCACC TACGAGCTGG ATGGCTGGTC  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2310 2320 2330 2340 2350  
**LOCUS** CGATGGCTCC ACCGCCGCA CCCGCACCTG GACCACCCC GCCAGCTCA  
 .....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2360 2370 2380 2390 2400  
**LOCUS** CCACCTGAC CGCTACCCTC GGCTCCGCG GCCTGCGCGC CGTCTACCAC



.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2410            2420            2430            2440            2450  
**LOCUS**        GATAACGCCG ATCTCACCGG CGCCACCGTC ACCCGCATCG ATCCCGCCGT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2460            2470            2480            2490            2500  
**LOCUS**        CGCCTTCGAT TGGGGCCTCG CCGCTCCCGT GTCCGGCATC GGCGCCGACA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2510            2520            2530            2540            2550  
**LOCUS**        CCTTCTCCGT CCGCTGGTCC GGCTCCGTCG TCCCCGCTA CTCCAGACC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2560            2570            2580            2590            2600  
**LOCUS**        TATACCTTCG CCACCACCTC CGATGATGGC GTCCGCCTCT GGTTCGATGG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2610            2620            2630            2640            2650  
**LOCUS**        CACCCTGGTC ATCGATCAGT GGACCAACCA CTCTCGCCGC GTGGATACCG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2660            2670            2680            2690            2700  
**LOCUS**        GCACCGTCGC CCTCACCGCT GGCCAGGCCG TCCCACATCGT CCTCGAGTAC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2710            2720            2730            2740            2750  
**LOCUS**        TTCGACGGCG TCCGCAACGC CGTGGCCGAG CTGCGCTGGT CCTCCACCAG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2760            2770            2780            2790            2800  
**LOCUS**        CCAGGCCTCC GAGATCGTCC CCACCACCCG CCTCCGCCCC TGAGTTAACC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2810            2820            2830            2840            2850  
**LOCUS**        TGGGCCTCAT GGGCCTTCCG CTCACTGCCC GCTTTCCAGT CGGGAAACCT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2860            2870            2880            2890            2900  
**LOCUS**        GTCGTGCCAG CTGCATTAAC ATGGTCATAG CTGTTTCCTT GCGTATTGGG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2910            2920            2930            2940            2950  
**LOCUS**        CGCTCTCCGC TTCCTCGCTC ACTGACTCGC TGCCTCGGTT CGTTCGGGTA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 2960            2970            2980            2990            3000  
**LOCUS**        AAGCCTGGGG TGCCTAATGA GCAAAAGGCC AGCAAAAGGC CAGGAACCGT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3010 3020 3030 3040 3050  
**LOCUS** AAAAAAGGCCG CGTTGCTGGC GTTTTTCCAT AGGCTCCGCC CCCCTGACGA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3060 3070 3080 3090 3100  
**LOCUS** GCATCACAAA AATCGACGCT CAAGTCAGAG GTGGCGAAAC CCGACAGGAC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3110 3120 3130 3140 3150  
**LOCUS** TATAAAGATA CCAGGCGTTT CCCCTGGAA GCTCCCTCGT GCGCTCTCCT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3160 3170 3180 3190 3200  
**LOCUS** GTTCCGACCC TGCCGCTTAC CGGATACCTG TCCGCCTTTC TCCCTTCGGG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3210 3220 3230 3240 3250  
**LOCUS** AAGCGTGGCG CTTTCTCATA GCTCACGCTG TAGGTATCTC AGTTCGGTGT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3260 3270 3280 3290 3300  
**LOCUS** AGGTCGTTCG CTCCAAGCTG GGCTGTGTGC ACGAACCCCC CGTTCAGCCC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3310 3320 3330 3340 3350  
**LOCUS** GACCGCTGCG CCTTATCCGG TAACTATCGT CTTGAGTCCA ACCCGGTAAG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3360 3370 3380 3390 3400  
**LOCUS** ACACGACTTA TCGCCACTGG CAGCAGCCAC TGGTAACAGG ATTAGCAGAG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3410 3420 3430 3440 3450  
**LOCUS** CGAGGTATGT AGGCGGTGCT ACAGAGTTCT TGAAGTGGTG GCCTAACTAC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3460 3470 3480 3490 3500  
**LOCUS** GGCTACACTA GAAGAACAGT ATTTGGTATC TGCCTCTGCTG TGAAGCCAGT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3510 3520 3530 3540 3550  
**LOCUS** TACCTTCGGA AAAAGAGTTG GTAGCTCTTG ATCCGGCAAA CAAACCACCG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3560 3570 3580 3590 3600  
**LOCUS** CTGGTAGCGG TGGTTTTTTT GTTTGCAAGC AGCAGATTAC GCGCAGAAAA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 3610 3620 3630 3640 3650  
**LOCUS** AAAGGATCTC AAGAAGATCC TTTGATCTTT TCTACGGGGT CTGACGCTCA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3660 3670 3680 3690 3700  
**LOCUS** GTGGAACGAA AACTCACGTT AAGGGATTTT GGTCATGAGA TTATCAAAAA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3710 3720 3730 3740 3750  
**LOCUS** GGATCTTCAC CTAGATCCTT TTAAATTAAA AATGAAGTTT TAAATCAATC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3760 3770 3780 3790 3800  
**LOCUS** TAAAGTATAT ATGAGTAAAC TTGGTCTGAC AGTTACCAAT GCTTAATCAG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3810 3820 3830 3840 3850  
**LOCUS** TGAGGCACCT ATCTCAGCGA TCTGTCTATT TCGTTCATCC ATAGTTGCCT

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3860 3870 3880 3890 3900  
**LOCUS** GACTCCCCGT CGTGTAGATA ACTACGATAC GGGAGGGCTT ACCATCTGGC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3910 3920 3930 3940 3950  
**LOCUS** CCCAGTGCTG CAATGATACC GCGAGAACCA CGCTCACCGG CTCCAGATTT

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 3960 3970 3980 3990 4000  
**LOCUS** ATCAGCAATA AACCAGCCAG CCGGAAGGGC CGAGCGCAGA AGTGGTCTCT

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 4010 4020 4030 4040 4050  
**LOCUS** CAACTTTATC CGCCTCCATC CAGTCTATTA ATTGTTGCCG GGAAGCTAGA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 4060 4070 4080 4090 4100  
**LOCUS** GTAAGTAGTT CGCCAGTTAA TAGTTTGCGC AACGTTGTTG CCATTGCTAC

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 4110 4120 4130 4140 4150  
**LOCUS** AGGCATCGTG GTGTCACGCT CGTCGTTTGG TATGGCTTCA TTCAGCTCCG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 4160 4170 4180 4190 4200  
**LOCUS** GTTCCCAACG ATCAAGGCGA GTTACATGAT CCCCATGTT GTGCAAAAAA

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|  
 4210 4220 4230 4240 4250  
**LOCUS** GCGGTTAGCT CCTTCGGTCC TCCGATCGTT GTCAGAAGTA AGTTGGCCGC

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4260 4270 4280 4290 4300  
**LOCUS** AGTGTTATCA CTCATGGTTA TGGCAGCACT GCATAATTCT CTTACTGTCA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4310 4320 4330 4340 4350  
**LOCUS** TGCCATCCGT AAGATGCTTT TCTGTGACTG GTGAGTACTC AACCAAGTCA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4360 4370 4380 4390 4400  
**LOCUS** TTCTGAGAAT AGTGTATGCG GCGACCGAGT TGCTCTTGCC CGGCGTCAAT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4410 4420 4430 4440 4450  
**LOCUS** ACGGGATAAT ACCGCGCCAC ATAGCAGAAC TTTAAAAGTG CTCATCATTG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4460 4470 4480 4490 4500  
**LOCUS** GAAAACGTTT TCGGGGCGA AACTCTCAA GGATCTTACC GCTGTTGAGA

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4510 4520 4530 4540 4550  
**LOCUS** TCCAGTTCGA TGTAACCCAC TCGTGCACCC AACTGATCTT CAGCATCTTT

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4560 4570 4580 4590 4600  
**LOCUS** TACTTTTACC AGCGTTTCTG GGTGAGCAAA AACAGGAAGG CAAAATGCCG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4610 4620 4630 4640 4650  
**LOCUS** CAAAAAAGGG AATAAGGGCG ACACGGAAAT GTTGAATACT CATACTCTTC

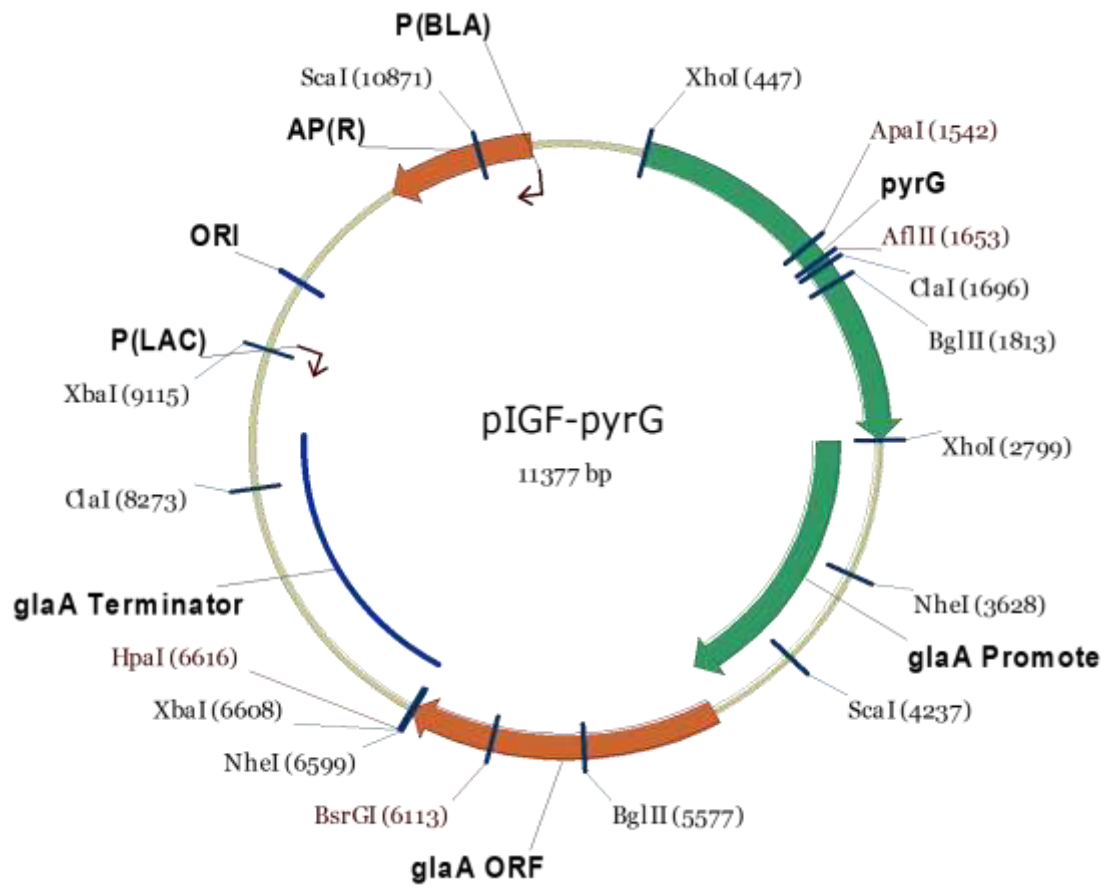
.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4660 4670 4680 4690 4700  
**LOCUS** CTTTTTCAAT ATTATTGAAG CATTTATCAG GGTTATTGTC TCATGAGCGG

.....|.....| .....|.....| .....|.....| .....|.....| .....|.....|  
 4710 4720 4730 4740 4750  
**LOCUS** ATACATATTT GAATGTATTT AGAAAAATAA ACAAATAGGG GTTCCGCGCA

.....|.....| .....|.....| .  
 4760 4770  
**LOCUS** CATTTCCCCG AAAAGTGCCA C

APPENDIX E

Appendix E: pIGF-pyrG *A. niger* vector map



## Abbreviations

AA	- Auxiliary Activity
BDC	- Biorenewable Development Centre
BLAST	- Basic Local Alignment Search Tool
BNR2	- Bacterial Neuramidase 2
CAZy	- Carbohydrate Active Enzyme
CAZymes	- Carbohydrate-active enzymes
CBM	- Carbohydrate binding module
CD	- Conserve Domain
CDS	- Coding DNA sequence
CER	- Carbon dioxide evolution rate
CEs	- Carbohydrate esterases
CFU	- Colony forming units
CI	- Crystallinity index
CO <sub>2</sub>	- Carbon dioxide
cm	- Centimeter
CPM	- Count per Million
DEAE	- Diethyl-Aminoethyl
DGE	- Differential Gene Expression
DNA	- deoxyribonucleic acid
ds cDNA	- double-stranded cDNA
DTE	- dithioerythriol
DTT	- dithiothreitol
emPAI	- exponentially modified Protein Abundance Index
EtBr	- ethidium bromide
FDR	- false discovery rate
FID	- Flame Ionization Detector
FN3	Fibronectin III
GalNAc	- N-acetylglucosamine

GFP	- glutathione fluorescence protein
GHG	- greenhouse gases
GHs	- glycoside hydrolases
GRAVY	- Grand Average of Hydropathy
GSDH	- Glucose sorbone dehydrogenase
GSPs	- Gene-Specific Primers
GST	- Glutathione-S-transferase
HMDS	- Hexamethyldisilazane
h	- hour
IEA	- International Energy Agency
InDA-C	- Insert Dependent Adaptor Cleavage
inDel	- insertion and deletion
IPCC	- Intergovernmental Panel on Climate Change
ITC	- inverse transition cycling
iTRAQ	- isobaric tag
LB	- Luria-Bertani
LFQ	- label-free quantification
logFC	- log-fold change
LPMOs	- lytic polysaccharide monooxygenases
µm	- micrometer
min	- minute(s)
MBP	- maltose-binding protein
MODIS	- Moderate Resolution Imaging Spectroradiometer
mRNA	- messenger RNA
MS	- mass spectrometry
NASA	- National America Space Association
NB	- Nutrient broth
NGS	- next-generation sequencing
nr	- non-redundant
OD	- optical density

PA14	- protective antigen
PASC	- Phosphoric acid swollen cellulose
PCA	- Principle Component Analysis
PCR	- Polymerase Chain Reaction
PDB	- Protein Data Bank
PGM	- Personal Genomic Machine
pHBAH	- p-hydrobenzoic acid
PKD	- Polycystic Kidney Disease
PL	- polysaccharide lyases
PPKM	- Pileup Per Kilobase per Millions of reads
PQQ	- pyrroloquinoline-quinone
RefSeq	- reference sequence
RNA-seq	- RNA sequencing
rRNA	- ribosomal RNA
RTS	- Room Temperature Stable
s	- seconds
SDS-PAGE	- Sodium dodecyl sulfate polyacrylamide gel
SEM	- Scanning electron microscopy
SLPs	- surface layer proteins
SOC	- Super Optimal broth with Catabolite repression
SP	- secretory pathway
sRNAs	- small RNAs
TBE	- Tris/Borate/EDTA
TEMED	- tetramethylethylenediamine
TFA	- trifluoroacetic acid
TMHMM	- Transmembrane Helices Hidden Markov Model
UniProt KB	- Universal Protein Resource



## References

- 1 Dudley B. British Petrol (BP) Energy Outlook 2017 Edition. 2017 [cited 29 April 2017]. Available from: <http://www.bp.com/en/global/corporate/energy-economics/energy-outlook.html>
- 2 Chung R. The Status of the Water-Food-Energy Nexus in Asia and the Pacific [Internet]. United Nations Economic and Social Commission for Asia and the Pacific. 2013 [cited 1 May 2017]. Available from: <http://www.unescap.org/resources/status-water-food-energy-nexus-asia-and-pacific>
- 3 Vogt KA., Patel-Weynand T., Shelton M., Vogt DJ., Gordon JC., Mukumoto CT. f, *et al.* Sustainability unpacked: Food, energy and water for resilient environments and societies. 1st ed. Sustainability Unpacked: Food, Energy and Water for Resilient Environments and Societies. Abingdon, Oxon: Earthscan; 2012. 1-306.
- 4 Waughray D. Water security: the water-food-energy-climate nexus: The World Economic Forum water initiative. Island Press; 2011. Pg 248.
- 5 Beddington J. Food, Energy, Water and The Climate: A Perfect Storm of Global Events? [Internet]. Government Office for Science. 2009 [cited 30 April 2017]. Available from: <http://webarchive.nationalarchives.gov.uk/20121206120858/http://www.bis.gov.uk/assets/goscience/docs/p/perfect-storm-paper.pdf>
- 6 Van Der Werf GR, Randerson JT, Giglio L, Collatz GJ, Mu M, Kasibhatla PS, *et al.* Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997-2009). *Atmos Chem Phys*. 2010; 10(23): 11707–35.
- 7 Chen B, Bai Z, Cui X, Chen J, Andersson A, Gustafsson Ö. Light absorption enhancement of black carbon from urban haze in Northern China winter. *Environ Pollut*. 2017; 221: 418–26.
- 8 Pearson TRH, Brown S, Murray L, Sidman G. Greenhouse gas emissions from tropical forest degradation: an underestimated source. *Carbon Balance Manag*. 2017; 12(1): 3.
- 9 Chen X. The Greenhouse Metaphor and the Greenhouse Effect: A Case Study of a Flawed Analogous Model. 2012; *Philos Cogn Sci*. (2): 105-14.
- 10 Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner K, Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow TZ and JC, (eds.) M, editors. IPCC, 2014: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA.: Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.; 2014.
- 11 UEA. News Release: CO2 emissions set to reach new 40 billion tonne record high in 2014. *Glob Carbon Proj*. 2014; (0): 8–10.
- 12 Bhattacharya A. Widespread crop burning began over dozen years ago. *The Times of India (Delhi)*. 2015; 1.

- 13 South-East Asia's smog: Unspontaneous combustion [Internet]. The Economists. 2013 [cited 30 April 2017]. Available from: <https://www.economist.com/news/asia/21580154-forest-fires-bring-record-levels-air-pollution-and-end-not-sight-unspontaneous>
- 14 Perera FP. Multiple Threats to Child Health from Fossil Fuel Combustion: Impacts of Air Pollution and Climate Change. *Environ Health Perspect.* 2016; 125(April): 141–8.
- 15 Aradhna Wal. Toxic Air: Young Farmers Vs Old Guard on Crop Burning in Punjab, Haryana [Internet]. News18, India. 2016 [cited 20 Apr 2017]. Available from: <http://www.news18.com/news/india/delhi-pollution-old-guard-in-punjab-haryana-fields-choking-efforts-to-clear-capital-air-1308314.html>
- 16 Richard T. L. Challenges in Scaling Up Biofuels Infrastructure. *Science.* 2010; 329 (5993): 793-796.
- 17 Naik SN, Goud V V., Rout PK, Dalai AK. Production of first and second generation biofuels: A comprehensive review. *Renew Sustain Energy Rev.* 2010; 14(2): 578–97.
- 18 Aro EM. From first generation biofuels to advanced solar biofuels. *Ambio.* 2016; 45(1): 24–31.
- 19 Komendantova N. and Pachauri S. Renewables 2012 Global Status Report [Internet]. International Institute for Applied Systems Analysis. 2012 [cited 30 Apr 2017]. Available from: <http://pure.iiasa.ac.at/10134/>
- 20 Glithero NJ, Ramsden SJ, Wilson P. Barriers and incentives to the production of bioethanol from cereal straw: A farm business perspective. *Energy Policy.* 2013; 59(100): 161–71.
- 21 Burkhardt S, Kumar L, Chandra R, Saddler J. How effective are traditional methods of compositional analysis in providing an accurate material balance for a range of softwood derived residues? *Biotechnol Biofuels.* 2013; 6(90): 1-10.
- 22 Glithero NJ, Wilson P, Ramsden SJ. Straw use and availability for second generation biofuels in England. *Biomass and Bioenergy.* 2013; 55: 311–21.
- 23 Yuill A. Sustainable Straw Combustion of Straw for Combined Heat and Power [Internet]. 2016. Natural Power. Available from: <https://www.naturalpower.com/sustainable-straw/>.
- 24 Wilson P, Glithero NJ, Ramsden SJ. Prospects for dedicated energy crop production and attitudes towards agricultural straw use: The case of livestock farmers. *Energy Policy.* 2014; 74: 101–10.
- 25 Roy J, Tucker GA, Sparkes DL. Wheat straw for biofuel production. [PhD project final report AHDB Cereals & Oilseed, RD-2007-3690] University of Nottingham; 2014.
- 26 Gomez LD, Steele-King CG, McQueen-Mason SJ. Sustainable liquid biofuels from biomass: the writings on the walls. *New Phytol.* 2008; 178(3): 473–85.
- 27 Zhu Z, Rezende CA, Simister R, McQueen-Mason SJ, Macquarrie DJ, Polikarpov I, *et al.* Efficient sugar production from sugarcane bagasse by microwave assisted acid and alkali pretreatment. *Biomass and Bioenergy.* 2016; 93: 269–78.

- 28 McNeil M, Darvill AG, Albersheim P. Structure of Plant Cell Walls: XII. Identification of Seven Differently Linked Glycosyl Residues Attached to O-4 of the 2,4-Linked L-Rhamnosyl Residues of Rhamnogalacturonan I. *Plant Physiol.* 1982; 70(6): 1586–91.
- 29 Chandra RP, Bura R, Mabee WE, Berlin A, Pan X, Saddler JN. Substrate Pretreatment: The Key to Effective Enzymatic Hydrolysis of Lignocellulosics? In: *Biofuels*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. 67–93.
- 30 Achyuthan KE, Achyuthan AM, Adams PD, Dirk SM, Harper JC, Simmons BA, *et al.* Supramolecular self-assembled chaos: Polyphenolic lignin’s barrier to cost-effective lignocellulosic biofuels. *Molecules.* 2010; 15(12): 8641–88.
- 31 Somerville C, Bauer S, Brininstool G, Facette M, Hamann T, Milne J, *et al.* Toward a Systems Approach to Understanding Plant Cell Walls. *Science (80).* 2004; 306 (5705): 2206-11.
- 32 Parthasarathi R, Bellesia G, Chundawat SPS, Dale BE, Langan P, Gnanakaran S. Insights into Hydrogen Bonding and Stacking Interactions in Cellulose. *J Phys Chem A.* 2011; 115(49): 14191–202.
- 33 Book AJ, Lewin GR, McDonald BR, Takasuka TE, Wendt-Pienkowski E, Doering DT, *et al.* Evolution of High Cellulolytic Activity in Symbiotic *Streptomyces* through Selection of Expanded Gene Content and Coordinated Gene Expression. Hillis DM, editor. *PLOS Biol.* 2016; 14(6): e1002475. <https://doi.org/10.1371/journal.pbio.1002475>
- 34 Hemsworth GR, Davies GJ, Walton PH. Recent insights into copper-containing lytic polysaccharide mono-oxygenases. *Current Opinion in Structural Biology.* 2013; 23: 660–668
- 35 Horn SJ, Vaaje-Kolstad G, Westereng B, Eijsink VG. Novel enzymes for the degradation of cellulose. *Biotechnol Biofuels.* 2012; 5(1):45.
- 36 Mohanram S, Amat D, Choudhary J, Arora A, Nain L, Bringezu S, *et al.* Novel perspectives for evolving enzyme cocktails for lignocellulose hydrolysis in biorefineries. *Sustain Chem Process.* 2013; 1(1): 15.
- 37 Bastawde KB. Xylan structure, microbial xylanases, and their mode of action. *World J Microbiol Biotechnol.* 1992; 8(4): 353–68.
- 38 Kumar R, Wyman CE. Strong cellulase inhibition by Mannan polysaccharides in cellulose conversion to sugars. *Biotechnol Bioeng.* 2014; 111(7): 1341–1353.
- 39 Duedu KO, French CE. Characterization of a *Cellulomonas fimi* exoglucanase/xylanase-endoglucanase gene fusion which improves microbial degradation of cellulosic biomass. *Enzyme Microb Technol.* 2016; 93: 113–21.
- 40 Woolridge EM. Mixed Enzyme Systems for Delignification of Lignocellulosic Biomass. *Catalysts.* 2014; 4:1–35.
- 41 Lan W, Lu F, Regner M, Zhu Y, Rencoret J, Ralph SA, *et al.* Tricin, a Flavonoid Monomer in Monocot Lignification. *Plant Physiol.* 2015; 167: 1284–95.
- 42 de Gonzalo G, Colpa DI, Habib MHM, Fraaije MW. Bacterial enzymes involved in lignin degradation. *J Biotechnol.* 2016; 236: 110–9.

- 43 Liu L, Ye XP, Womac AR, Sokhansanj S. Variability of biomass chemical composition and rapid analysis using FT-NIR techniques. *Carbohydr Polym.* 2010; 81(4): 820–9.
- 44 Davin LB, Patten AM, Jourdes M, Lewis NG. Lignins: A Twenty-First Century Challenge. In: *Biomass Recalcitrance*. Blackwell Publishing Ltd.; 2008. 213–305.
- 45 Xu F, Yu J, Tesso T, Dowell F, Wang D. Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: A mini-review. *Appl Energy.* 2013; 104: 801–9.
- 46 Crestini C, Melone F, Sette M, Saladino R. Milled Wood Lignin: A Linear Oligomer. *Biomacromolecules.* 2011; 12(11): 3928–35.
- 47 Sethi A, Scharf ME, Sethi A, Scharf ME. Biofuels: Fungal, Bacterial and Insect Degraders of Lignocellulose. In: eLS. Chichester, UK: John Wiley & Sons, Ltd; 2013.
- 48 Nimchua T, Thongaram T, Uengwetwanit T, Pongpattanakitsote S, Eurwilaichitr L. Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes. *J Microbiol Biotechnol.* 2012; 22(4): 462-469.
- 49 Scott JJ, Oh D-C, Yuceer MC, Klepzig KD, Clardy J, Currie CR. Bacterial protection of beetle-fungus mutualism. *Science.* 2008; 322(5898): 63.
- 50 Adams AS, Adams SM, Currie CR, Gillette NE, Raffa KF. Geographic variation in bacterial communities associated with the red turpentine beetle (*Coleoptera: Curculionidae*). *Environ Entomol.* 2010; 39(2): 406–14.
- 51 King AJ, Cragg SM, Li Y, Dymond J, Guille MJ, Bowles DJ, *et al.* Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proc Natl Acad Sci U S A.* 2010; 107(12): 5345–50.
- 52 Kern M, McGeehan JE, Streeter SD, Martin RNA, Besser K, Elias L, *et al.* Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc Natl Acad Sci U S A.* 2013; 110(25): 10189–94.
- 53 Liu Z, Ho S-H, Sasaki K, den Haan R, Inokuma K, Ogino C, *et al.* Engineering of a novel cellulose-adherent cellulolytic *Saccharomyces cerevisiae* for cellulosic biofuel production. *Sci Rep.* 2016; 6:24550. DOI: 10.1038/srep24550.
- 54 Bayer EA, Henrissat B, Lamed R. The Cellulosome: A Natural Bacterial Strategy to Combat Biomass Recalcitrance. In: *Biomass Recalcitrance*. Blackwell Publishing Ltd.; 2008. 407–35.
- 55 Gilbert HJ. Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol Microbiol.* 2007; 63(6): 1568–76.
- 56 Delmas S, Pullan ST, Gaddipati S, Kokolski M, Malla S, Blythe MJ, *et al.* Uncovering the Genome-Wide Transcriptional Responses of the Filamentous Fungus *Aspergillus niger* to Lignocellulose Using RNA Sequencing. *PLoS Genet.* 2012; 8(8): e1002875.
- 57 Woo HL, Hazen TC, Simmons BA, DeAngelis KM. Enzyme activities of aerobic lignocellulolytic bacteria isolated from wet tropical forest soils. *Syst Appl Microbiol.* 2014; 37(1): 60–7.

- 58 Levasseur A, Lomascolo A, Chabrol O, Ruiz-Dueñas FJ, Boukhris-Uzan E, Piumi F, *et al.* The genome of the white-rot fungus *Pycnoporus cinnabarinus*: a basidiomycete model with a versatile arsenal for lignocellulosic biomass breakdown. *BMC Genomics*. 2014; 15(1): 486.
- 59 Brown ME, Chang MCY. Exploring bacterial lignin degradation. *Current Opinion in Chemical Biology*. 2014; 19: 1-7.
- 60 DeAngelis KM, Sharma D, Varney R, Simmons B, Isern NG, Markillie LM, *et al.* Evidence supporting dissimilatory and assimilatory lignin degradation in *Enterobacter lignolyticus* SCF1. *Front Microbiol*. 2013; 4: 1–14.
- 61 Cragg SM, Beckham GT, Bruce NC, Bugg TD, Distel DL, Dupree P, *et al.* Lignocellulose degradation mechanisms across the Tree of Life. *Curr Opin Chem Biol*. 2015; 29: 108–19.
- 62 Van Den Brink J, De Vries RP. Fungal enzyme sets for plant polysaccharide degradation. 2011; 91: 1477–92.
- 63 Gilbert HJ, Knox JP, Boraston AB. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Current Opinion in Structural Biology*. 2013; 23: 669–677.
- 64 Crouch LI, Labourel A, Walton PH, Davies GJ, Gilbert HJ. The Contribution of Non-Catalytic Carbohydrate Binding Modules to the Activity of Lytic Polysaccharide Monooxygenases. *J Biol Chem*. 2016; 291(14): 7439–49.
- 65 Kjaergaard CH, Qayyum MF, Wong SD, Xu F, Hemsworth GR, Walton DJ, *et al.* Spectroscopic and computational insight into the activation of O<sub>2</sub> by the mononuclear Cu center in polysaccharide monooxygenases. *Proc Natl Acad Sci U S A*. 2014; 111(24): 8797–802.
- 66 Forsberg Z, Mackenzie AK, Sørliie M, Røhr ÅK, Helland R, Arvai AS, *et al.* Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing lytic polysaccharide monooxygenases. *Proc Natl Acad Sci U S A*. 2014; 111(23): 8446–51.
- 67 Tan T-C, Kracher D, Gandini R, Sygmund C, Kittl R, Haltrich D, *et al.* Structural basis for cellobiose dehydrogenase action during oxidative cellulose degradation. *Nat Commun*. 2015; 6: 7542.
- 68 Hemsworth GR, Henrissat B, Davies GJ, Walton PH. Discovery and characterization of a new family of lytic polysaccharide monooxygenases. *Nat Chem Biol*. 2013; 10(2): 122–6.
- 69 Busk PK, Lange L. Classification of fungal and bacterial lytic polysaccharide monooxygenases. *BMC Genomics*. 2015. 16:368 DOI 10.1186/s12864-015-1601-6
- 70 Langston JA, Shaghasi T, Abbate E, Xu F, Vlasenko E, Sweeney MD. Oxidoreductive cellulose depolymerization by the enzymes cellobiose dehydrogenase and glycoside hydrolase 61. *Appl Environ Microbiol*. 2011; 77(19): 7007–15.
- 71 André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M. CAZyme discovery and design for sweet dreams. *Current Opinion in Chemical Biology*. 2013; 19: 17-24.

- 72 Madhuprakash J, El Gueddari NE, Moerschbacher BM, Podile AR. Catalytic efficiency of chitinase-D on insoluble chitinous substrates was improved by fusing auxiliary domains. *PLoS One*. 2015; 10(1): e0116823. doi:10.1371/journal.pone.0116823
- 73 Beeson WT, Vu V V, Span EA, Phillips CM, Marletta MA. Cellulose degradation by polysaccharide monooxygenases. *Annu Rev Biochem*. 2015; 84: 923–46.
- 74 Hemsworth GR, Taylor EJ, Kim RQ, Gregory RC, Lewis SJ, Turkenburg JP, *et al*. The Copper Active Site of CBM33 Polysaccharide Oxygenases. *J Am Chem Soc*. 2013; 135(16): 6069–77.
- 75 Quinlan RJ, Sweeney MD, Lo Leggio L, Otten H, Poulsen J-CN, Johansen KS, *et al*. Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. *Proc Natl Acad Sci U S A*. 2011; 108(37): 15079–84.
- 76 Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*. 2014; 42(Database issue): D490-5. doi:10.1093/nar/gkt1178
- 77 Shen H, Gilkes NR, Kilburn DG, Miller RC, Antony R, Warren J. Cellobiohydrolase B, a second exo-cellobiohydrolase from the cellulolytic bacterium *Cellulomonas fimi*. *Biochem J*. 1995; 311: 67–74.
- 78 Tomme P, Kwan E, Gilkes NR, Kilburn DG, Antony R, Warren J. Characterization of CenC, an Enzyme from *Cellulomonas fimi* with Both Endo-and Exoglucanase Activities. *J Bacteriol*. 1996; 178(14): 4216–23.
- 79 Din N, Damude HG, Gilkes NR, Miller RC, Antony R, Warren J, *et al*. C1-Cx, revisited: Intramolecular synergism in a cellulase. 1994; 91: 11383–7.
- 80 White a, Withers SG, Gilkes NR, Rose DR. Crystal structure of the catalytic domain of the beta-1,4-glycanase cex from *Cellulomonas fimi*. *Biochemistry*. 1994; 33(42): 12546–52.
- 81 Meinke A, Braun C, Gilkes NR, Kilburn DG, Miller RC, Warren RA. Unusual sequence organization in CenB, an inverting endoglucanase from *Cellulomonas fimi*. *J Bacteriol*. 1991;173(1):308–14.
- 82 Poon DKY, Withers SG, McIntosh LP. Direct demonstration of the flexibility of the glycosylated proline-threonine linker in the *Cellulomonas fimi* xylanase Cex through NMR spectroscopic analysis. *J Biol Chem*. 2007; 282(3): 2091-100.
- 83 Tešić M, Wicki J, Poon DKY, Withers SG, Douglas DJ. Gas Phase Noncovalent Protein Complexes that Retain Solution Binding Properties: Binding of Xylobiose Inhibitors to the  $\beta$ -1, 4 Exoglucanase from *Cellulomonas fimi*. *J Am Soc Mass Spectrom*. 2007; 18(1): 64–73.
- 84 Le Nours J, Anderson L, Stoll D, Ståhlbrand H, Lo Leggio L. The structure and characterization of a modular endo- $\beta$ -1,4-mannanase from *Cellulomonas fimi*. *Biochemistry*. 2005; 44: 12700-708.
- 85 Notenboom V, Williams SJ, Hoos R, Withers SG, Rose DR. Detailed structural analysis of glycosidase/inhibitor interactions: Complexes of cex from *Cellulomonas fimi* with xylobiose-derived aza-sugars. *Biochemistry*. 2000 Sep 26;39(38):11553-63.

- 86 Christopherson MR, Suen G, Bramhacharya S, Jewell KA, Aylward FO, Mead D, *et al.* The genome sequences of *Cellulomonas fimi* and "*Cellvibrio gilvus*" reveal the cellulolytic strategies of two facultative anaerobes, transfer of "*Cellvibrio gilvus*" to the genus *Cellulomonas*, and proposal of *Cellulomonas gilvus* sp. nov. PLoS One. 2013; 8(1): e53954.
- 87 Jing H, Cockburn D, Zhang Q, Clarke AJ. Production and purification of the isolated family 2a carbohydrate-binding module from *Cellulomonas fimi*. Protein Expr Purif. 2009; 64(1): 63–68.
- 88 Hashimoto H. Recent structural studies of carbohydrate-binding modules. Cell Mol Life Sci. 2006; 63(24): 2954–67.
- 89 Moser B, Gilkes NR, Kilburn DG, Warren RAJ, Miller RC. Purification and Characterization of Endoglucanase C of *Cellulomonas fimi*, Cloning of the Gene, and Analysis of In Vivo Transcripts of the Gene. Appl Environ Microbiol. 1989; 55(10): 2480–7.
- 90 Coutinho JB, Moser B, Kilburn DG, Warren RA, Miller RC. Nucleotide sequence of the endoglucanase C gene (*cenC*) of *Cellulomonas fimi*, its high-level expression in *Escherichia coli*, and characterization of its products. Mol Microbiol. 1991; 5(5): 1221–33.
- 91 Johnson PE, Tomme P, Douglas G, Kilburn A, McIntosh and LP. Structure of the N-Terminal Cellulose-Binding Domain of *Cellulomonas fimi* CenC Determined by Nuclear Magnetic Resonance Spectroscopy. Biochemistry. 1996; 35 (45): 14381–94.
- 92 Brun E, Johnson PE, Creagh AL, Tomme P, Webster P, Charles A. Haynes A, *et al.* Structure and Binding Specificity of the Second N-Terminal Cellulose-Binding Domain from *Cellulomonas fimi* Endoglucanase C. Biochemistry, 2000; 39 (10), 2445–58.
- 93 Meinke A, Gilkes NR, Kilburn DG, Miller RC, Warren RAJ. Cellulose-binding polypeptides from *Cellulomonas fimi*: Endoglucanase D (CenD), a family A  $\beta$ -1,4-glucanase. J Bacteriol. 1993; 175(7): 1910–18.
- 94 Shen H, Tomme P, Meinke A, Gilkes NR, Kilburn DG, Warren RA, *et al.* Stereochemical course of hydrolysis catalysed by *Cellulomonas fimi* CenE, a member of a new family of beta-1,4-glucanases. Biochem Biophys Res Commun. 1994; 199(3): 1223–28.
- 95 Raymond Wong WKK, Gerhard B, Guo ZM, Kilburn DG, Anthony R, Warren J, *et al.* Characterization and structure of an endoglucanase gene *cenA* of *Cellulomonas fimi*. Gene. 1986; 44(2–3): 315–24.
- 96 Lakhundi SS, Duedu KO, Cain N, Nagy R, Krakowiak J, French CE. *Citrobacter freundii* as a test platform for recombinant cellulose degradation systems. Lett Appl Microbiol. 2016; 35–42.
- 97 Meinke A, Gilkes NR, Kilburn DG, Miller RC, Warren RA. Multiple domains in endoglucanase B (CenB) from *Cellulomonas fimi*: functions and relatedness to domains in other polypeptides. J Bacteriol. 1991; 173(22): 7126–35.
- 98 Meinke A, Gilkes NR, Kwan E, Kilburn DG, Warren RAJ, Miller RC. Cellobiohydrolase A (Cbha) from the cellulolytic bacterium *Cellulomonas fimi* is a  $\beta$ -1,4-exocellobiohydrolase analogous to *Trichoderma reesei* CBH II. Mol Microbiol. 1994; 12(3): 413–22.

- 99 Mayer C, Vocadlo DJ, Mah M, Rupitz K, Stoll D, Warren RAJ, *et al.* Characterization of a  $\alpha$ -N-acetylhexosaminidase and  $\alpha$ -N-acetylglucosaminidase-glycosidase from *Cellulomonas fimi*. FEBS J. 2006; 273(13): 2929–41.
- 100 Millward-Sadler SJ, Hall J, Black GW, Hazlewood GP, Gilbert HJ. Evidence that the Piromyces gene family encoding endo-1,4-mannanases arose through gene duplication. FEMS Microbiol Lett. 1996; 141(2–3): 183–8.
- 101 Simpson PJ, Bolam DN, Cooper A, Ciruela A, Hazlewood GP, Gilbert HJ, *et al.* A family IIb xylan-binding domain has a similar secondary structure to a homologous family II a cellulose-binding domain but different ligand specificity. Structure. 1999; 7(7): 853–64.
- 102 Simpson PJ, Xie H, Bolam DN, Gilbert HJ, Williamson MP. The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. J Biol Chem. 2000; 275(52): 41137–42.
- 103 Bolam DN, Xie H, White P, Simpson PJ, Hancock SM, Williamson MP, *et al.* Evidence for synergy between family 2b carbohydrate binding modules in *Cellulomonas fimi* Xylanase 11A. Biochemistry. 2001; 40(8): 2468–77.
- 104 O’Neill G, Goh SH, Warren RAJ, Kilburn DG, Miller RC. Structure of the gene encoding the exoglucanase of *Cellulomonas fimi*. Gene. 1986; 44(23): 325–30.
- 105 Tull D, Witherst SG, Gilkes NR, Kilburn DG, Antony J Warren R, Aebersold R. Glutamic acid 274 is the nucleophile in the active site of a “retaining” exoglucanase from *Cellulomonas fimi*. J Biol Chem. 1991; 266(24): 15621–5.
- 106 MacLeod AM, Lindhorst T, Withers SG, Warren RAJ. The Acid/Base Catalyst in the Exoglucanase/Xylanase from *Cellulomonas fimi* Is Glutamic Acid 127: Evidence from Detailed Kinetic Studies of Mutants. Biochemistry. 1994; 33(20): 6371–6.
- 107 Xu G-YY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, *et al.* Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. Biochemistry. 1995; 34(21): 6993–7009.
- 108 Notenboom V, Birsan C, Nitz M, Rose DR, Warren RAJ, Withers SG. Insights into transition state stabilization of the  $\beta$ -1,4-glycosidase Cex by covalent intermediate accumulation in active site mutants. Nat Struct Biol. 1998; 5(9): 812–8.
- 109 Notenboom V, Birsan C, Warren RAJ, Withers SG, Rose DR. Exploring the Cellulose / Xylan Specificity of the -1, 4-Glycanase Cex from *Cellulomonas fimi* through Crystallography and Mutation. 1998; 2960(97): 4751–8.
- 110 Cheng H-R, Jiang N. Extremely rapid extraction of DNA from bacteria and yeasts. Biotechnol. Lett. 2006; 28: 55–59.
- 111 Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976; 72(1–2): 248–54.
- 112 Stackebrandt E, Schumann P, Prauser H. The Family *Cellulomonadaceae*. In: The Prokaryotes. New York, NY: Springer New York; 2006.983–1001.
- 113 Gao J, Wakarchuk W. Characterization of five  $\beta$ -glycoside hydrolases from *Cellulomonas fimi* ATCC 484. J Bacteriol. 2014; 196(23): 4103–10.



- 114 Mansfield SD, Meder R. Cellulose hydrolysis – the role of monocomponent cellulases in crystalline cellulose degradation. *Cellulose*. 2003; 10(2): 159–69.
- 115 Rasmussen S, Nielsen HB, Jarmer H. The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol*. 2009 Sep;73(6):1043–57.
- 116 Van Vliet a. HM. Next generation sequencing of microbial transcriptomes: Challenges and opportunities. *FEMS Microbiol Lett*. 2010; 302:1–7.
- 117 Ragno S, Romano M, Howell S, Pappin DJC, Jenner PJ, Colston MJ. Changes in gene expression in macrophages infected with *Mycobacterium tuberculosis*: a combined transcriptomic and proteomic approach. *Immunology*. 2001; 104(1): 99–108.
- 118 Cho S, Cho Y, Lee S, Kim J, Yum H, Chang Kim S, *et al*. Current Challenges in Bacterial Transcriptomics. *Genomics Inf*. 2013; 11(2): 76–82.
- 119 Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, *et al*. Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol*. 2009; 5:285. doi: 10.1038/msb.2009
- 120 Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, *et al*. The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol*. 2009; 27(11): 1043–9.
- 121 Pinto AC, Melo-Barbosa HP, Miyoshi A, Silva A, Azevedo V. Review Application of RNA-seq to reveal the transcript profile in bacteria. *Genet Mol Res*. 2011; 10(3): 1707–18.
- 122 Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*. 2010; 11(1): 9–16.
- 123 Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014; 15(2): 121–32.
- 124 Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010; 11(1): 31–46.
- 125 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009; 10(1): 57–63.
- 126 Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. *Mol Syst Biol*. 2014;9(1):640–640.
- 127 Croucher NJ, Thomson NR. Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol*. 2010; 13(5): 619–24.
- 128 Philippe N, Boureux A, Bréhélin L, Tarhio J, Commes T, Rivals E. Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res*. 2009; 37(15): e104.
- 129 Güell M, Yus E, Lluch-Senar M, Serrano L. Bacterial transcriptomics: what is beyond the RNA hori-zome? *Nat Rev Microbiol*. 2011; 9(9): 658–69.
- 130 He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The Antisense Transcriptomes of Human Cells. *Science*. 2008; 322(5909): 1855-7.
- 131 Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitch S, *et al*. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*. 2009; 37(18).

- 132 Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, *et al.* A simple method for directional transcriptome sequencing using illumina technology. *Nucleic Acids Res.* 2009; 37(22): e148. doi: 10.1093/nar/gkp811.
- 133 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10(3): R25. (doi: 1186/gb-2009-10-3-r25)
- 134 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2009;37(Database): D26–31.
- 135 Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009; 25(9): 1105–11.
- 136 Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, *et al.* The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* 2014; 42(D1): D26–31.
- 137 Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 2008; 36(10): 3420–35.
- 138 Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010; 26(1): 139–40.
- 139 Book AJ, Yennamalli RM, Takasuka TE, Currie CR, Phillips GN, Fox BG. Evolution of substrate specificity in bacterial AA10 lytic polysaccharide monoxygenases. *Biotechnol Biofuels.* 2014; 7:109. doi: 10.1186/1754-6834-7-109. eCollection 2014.
- 140 Oliveros JC. Venny - An interactive tool for comparing lists with Venn's diagrams [Internet]. Venny 2.0. 2015 [cited 28 Dec 2016]. Available from: <http://bioinfogp.cnb.csic.es/tools/venny/index.html>.
- 141 Vaaje-Kolstad G, Westereng B, Horn SJ, Liu Z, Zhai H, Sørli M, *et al.* An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. *Science.* 2010; 330(6001): 219–22.
- 142 Kracher D, Scheiblbrandner S, Felice AKG, Breslmayr E, Preims M, Ludwicka K, *et al.* Extracellular electron transfer systems fuel cellulose oxidative degradation. *Science.* 2016; 352(6289): 1098-101.
- 143 Park S, Baker JO, Himmel ME, Parilla PA, Johnson DK. Cellulose crystallinity index: measurement techniques and their impact on interpreting cellulase performance. *Biotechnol Biofuels.* 2010; 3(1): 10.
- 144 Wakarchuk WW, Brochu D, Foote S, Robotham A, Saxena H, Erak T, *et al.* Proteomic analysis of the secretome of *Cellulomonas fimi* ATCC 484 and *Cellulomonas flavigena* ATCC 482. *PLoS ONE* 11(3): e0151186. <https://doi.org/10.1371/journal.pone.0151186>
- 145 UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2017; 45(D1): D158–69.
- 146 Henrissat B, Claeysens M, Tomme P, Lemesle L, Mornon JP. Cellulase families revealed by hydrophobic cluster analysis. *Gene.* 1989; 81(1): 83–95.

- 147 O'farrell PH. High Resolution Two-Dimensional Electrophoresis of Proteins. *J Biol Chem.* 1975; 250(10): 4007-21.
- 148 Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom.* 1994; 5(11): 976–89.
- 149 Gu Q, Yu LR. Proteomics quality and standard: From a regulatory perspective. *Journal of Proteomics.* 2014; 353-359. <https://doi.org/10.1016/j.jprot.2013.11.024>
- 150 Guest PC, Gottschalk MG, Bahn S. Proteomics: improving biomarker translation to modern medicine? *Genome Med.* 2013; 5(2):17. doi: 10.1186/gm421. eCollection 2013.
- 151 Crutchfield CA, Thomas SN, Sokoll LJ, Chan DW. Advances in mass spectrometry-based clinical biomarker discovery. *Clin Proteomics.* 2016; 13(1): 1. doi: 10.1186/s12014-015-9102-9
- 152 Hou S, Jones SW, Choe LH, Papoutsakis ET, Lee KH. Workflow for quantitative proteomic analysis of *Clostridium acetobutylicum* ATCC 824 using iTRAQ tags. *Methods.* 2013; 61(3): 269-76. doi: 10.1016/j.ymeth.2013.03.013. Epub 2013 Mar 22.
- 153 Manavalan A, Adav SS, Sze SK. ITRAQ-based quantitative secretome analysis of *Phanerochaete chrysosporium*. *J Proteomics.* 2011; 75(2):642-54. doi: 10.1016/j.jprot.2011.09.001. Epub 2011 Sep 13.
- 154 Wu WW, Wang G, Baek SJ, Shen RF. Comparative study of three proteomic quantitative methods, DIGE, cIcAT, and iTRAQ, using 2D gel- or LC-MALDI TOF/TOF. *J Proteome Res.* 2006; 5(3): 651-8.
- 155 Tuveng TR, Arntzen MØ, Bengtsson O, Gardner JG, Vaaje-Kolstad G, Eijsink VGH. Proteomic investigation of the secretome of *Cellvibrio japonicus* during growth on chitin. *Proteomics.* 2016; 1904-14.
- 156 Guilherme AA, Dantas PVF, Santos ES, Fernandes FAN, Macedo GR. Evaluation of composition, characterization and enzymatic hydrolysis of pretreated sugarcane bagasse. 2015; 32(1): 23-33. [dx.doi.org/10.1590/0104-6632.20150321s00003146](https://doi.org/10.1590/0104-6632.20150321s00003146)
- 157 Martinelli LA, Filoso S. Expansion of sugarcane ethanol production in Brazil: Environmental and social challenges. *Ecol Appl.* 2008; 18(4): 885–98.
- 158 Collins SR, Wellner N, Martinez Bordonado I, Harper AL, Miller CN, Bancroft I, *et al.* Variation in the chemical composition of wheat straw: the role of tissue ratio and composition. *Biotechnol Biofuels.* 2014; 7(1): 121.
- 159 Harper SHT, Lynch JM. The chemical components and decomposition of wheat straw leaves, internodes and nodes. *J Sci Food Agric.* 1981; 32(11): 1057–62.
- 160 Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, *et al.* Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science.* 2007; 315(5813): 804–7.
- 161 Li B, Fillmore N, Bai Y, Collins M, Thomson J a., Stewart R, *et al.* Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology.* 2014; 15:553. doi: 10.1186/s13059-014-0553-5

- 162 Teather RM, Wood PJ. Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from the bovine rumen. *Appl Environ Microbiol.* 1982; 43(4): 777–80.
- 163 Lever M. A new reaction for colorimetric determination of carbohydrates. *Anal Biochem.* 1972; 47(1): 273–9.
- 164 Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics.* 2005; 4(9): 1265–72.
- 165 Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods.* 2012; 9(7): 671–5.
- 166 Boraston AB, Tomme P, Amandoron EA, Kilburn DG. A novel mechanism of xylan binding by a lectin-like module from *Streptomyces lividans* xylanase 10A. *Biochem J.* 2000; 350: 933–41.
- 167 Pohleven J, Renko M, Magister Š, Smith DF, Künzler M, Štrukelj B, *et al.* Bivalent carbohydrate binding is required for biological activity of CNL, the LacdiNac (GalNAc $\beta$ 1–4GlcNAc)-specific lectin from basidiomycete *Clitocybe nebularis*. *The Jour. Of Biol. Chem.* 2012. 287; 10602-612.
- 168 Boraston AB, Creagh | A Louise, Alam MM, Kormos JM, Tomme P, Haynes CA, *et al.* Binding Specificity and Thermodynamics of a Family 9 Carbohydrate-Binding Module from *Thermotoga maritima* Xylanase 10A. *Biochemistry.* 2001; 40 (21), 6240–6247. doi: 10.1021/bi0101695
- 169 Gaskell A, Crennell S, Taylor G. The three domains of a bacterial sialidase: a  $\beta$ -propeller, an immunoglobulin module and a galactose-binding jelly-roll. *Structure.* 1995; 3(11): 1197–205.
- 170 Copley RR, Russell RB, Ponting CP. Sialidase-like Asp-boxes: sequence-similar structures within different protein folds. *Protein Sci.* 2001; 10(2): 285–92.
- 171 Kane S. The Degradation of Cellulosic Material by *Cellulomonas fimi* [PhD]. University of Edinburgh; 2015.
- 172 Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol.* 2001; 305(3): 567–80.
- 173 Das SP, Ghosh A, Gupta A, Goyal A, Das D. Lignocellulosic fermentation of wild grass employing recombinant hydrolytic enzymes and fermentative microbes with effective bioethanol recovery. *Biomed Res Int.* 2013; Article ID 386063, 14 pg.
- 174 Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J.* 2004; 382: 769–81.
- 175 Hekmat O, Florizone C, Kim Y-W, Eltis LD, Warren RAJ, Withers SG. Specificity fingerprinting of retaining beta-1,4-glycanases in the *Cellulomonas fimi* secretome using two fluorescent mechanism-based probes. *Chembiochem.* 2007; 8(17): 2125–32.

- 176 Pollet A, Delcour JA, Courtin CM. Structural determinants of the substrate specificities of xylanases from different glycoside hydrolase families. *Crit Rev Biotechnol*. 2010; 30(3): 176–91.
- 177 Xu F, Shi Y-C, Wang D. X-ray scattering studies of lignocellulosic biomass: a review. *Carbohydr Polym*. 2013; 94(2): 904–17.
- 178 Kaneko S, Ichinose H, Fujimoto Z, Kuno A, Yura K, Go M, *et al*. Structure and function of a family 10  $\beta$ -xylanase chimera of *Streptomyces olivaceoviridis* E-86 FXYN and *Cellulomonas fimi* cex. *J Biol Chem*. 2004; 279: 26619–26.
- 179 Thomas L, Joseph A, Gottumukkala LD. Xylanase and cellulase systems of *Clostridium* sp.: An insight on molecular approaches for strain improvement. *Bioresource Technology*. 2014; 158: 343–350.
- 180 Pérez-Avalos O, Sánchez-Herrera LM, Salgado LM, Ponce-Noyola T. A bifunctional endoglucanase/endoxylanase from *Cellulomonas flavigena* with potential use in industrial processes at different pH. *Curr Microbiol*. 2008; 57(1): 39–44.
- 181 Rashamuse KJ, Visser DF, Hennessy F, Kemp J, Roux-Van Der Merwe MP, Badenhorst J, *et al*. Characterisation of two bifunctional cellulase-Xylanase enzymes isolated from a bovine rumen metagenome library. *Curr Microbiol*. 2013; 66(2): 145–51.
- 182 Clarke JH, Laurie JI, Gilbert HJ, Hazlewood GP. Multiple xylanases of *Cellulomonas fimi* are encoded by distinct genes. *FEMS Microbiol Lett*. 1991; 67(3): 305–9.
- 183 Gilbert HJ, Jenkins G, Sullivan DA, Hall J. Evidence for multiple carboxymethylcellulase genes in *Pseudomonas fluorescens* subsp. *cellulosa*. *MGG Mol Gen Genet*. 1987; 210(3): 551–6.
- 184 Clarke JH, Davidson K, Gilbert HJ, Fontes CMGA, Hazlewood GP. A modular xylanase from mesophilic *Cellulomonas fimi* contains the same cellulose-binding and thermostabilizing domains as xylanases from thermophilic bacteria. *FEMS Microbiol Lett*. 1996; 139(1): 27–35.
- 185 Penttilä PA, Várnai A, Pere J, Tammelin T, Salmén L, Siika-aho M, *et al*. Xylan as limiting factor in enzymatic hydrolysis of nanocellulose. *Bioresour Technol*. 2013; 129: 135–141.
- 186 Salamanca-Cardona L, Ashe CS, Stipanovic AJ, Nomura CT. Enhanced production of polyhydroxyalkanoates (PHAs) from beechwood xylan by recombinant *Escherichia coli*. *Appl Microbiol Biotechnol*. 2014; 98(2): 831–42.
- 187 Buist G, Ridder ANJA, Kok J, Kuipers OP. Different subcellular locations of secretome components of Gram-positive bacteria. *Microbiology*. 2006; 152: 2867–74.
- 188 Sutcliffe IC, Harrington DJ. Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology*. 2002; 148: 2065–77.
- 189 Peabody MA, Laird MR, Vlasschaert C, Lo R, Brinkman FSL. PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res*. 2016;44(D1): D663–8.

- 190 Caspi J, Barak Y, Haimovitz R, Gilary H, Irwin DC, Lamed R, *et al.* *Thermobifida fusca* exoglucanase Cel6B is incompatible with the cellulosomal mode in contrast to endoglucanase Cel6A. *Syst Synth Biol.* 2010; 4(3): 193–201.
- 191 Miller RC, Langsford ML, Gilkes NR, Wakarchuk WW, Kilburn DG, Miller RC, *et al.* The Cellulase System of *Cellulomonas fimi*. *Microbiology.* 1984; 130(6): 1367–76.
- 192 Coutinho JB, Gilkes NR, Warren RAJ, Kilburn DG, Miller RC. The binding of *Cellulomonas fimi* endoglucanase C (CenC) to cellulose and Sephadex is mediated by the N-terminal repeats. *Mol Microbiol.* 1992; 6(9): 1243–52.
- 193 Moser F, Irwin D, Chen S, Wilson DB. Regulation and characterization of *Thermobifida fusca* carbohydrate-binding module proteins E7 and E8. *Biotechnol Bioeng.* 2008; 100(6): 1066–77.
- 194 Gilkes NR, Jervissi E, Henrissat B, Tekant B, Miller RC, Antony R, *et al.* The Adsorption of a Bacterial Cellulase to Crystalline Cellulose. *J Biol Chem.* 1992; 267(10): 6743–9.
- 195 Mclean BW, Boraston AB, Brouwer D, Sanaie N, Fyfe CA, Antony R, *et al.* Carbohydrate-binding Modules Recognize Fine Substructures of Cellulose. *J Biol Chem.* 2002; 277(52): 50245–54.
- 196 Notenboom V., Boraston A. B., Kilburn D. G., and, Rose D. R. Crystal Structures of the Family 9 Carbohydrate-Binding Module from *Thermotoga maritima* Xylanase 10A in Native and Ligand-Bound Forms. 2001; 40: 6248-56.
- 197 Varrot A, Basheer SM, Imberty A. Fungal lectins: structure, function and potential applications. *Curr Opin Struct Biol.* 2013; 23: 678–85.
- 198 Davies GJ, Henrissat B. Cracking the code, slowly: the state of carbohydrate-active enzymes in 2013 Editorial overview. *Curr Opin Struct Biol.* 2013; 23: 649–51.
- 199 Vincent P, Shareck F, Dupont C, Morosoli R, Kluepfel D. New  $\alpha$ -arabinosefuranosidase produced by *Streptomyces lividans*: Cloning and DNA sequence of the abfB gene and characterization of the enzyme. *Biochem J.* 1997; 322: 845–52.
- 200 Shareck F, Roy C, Yaguchi M, Morosoli R, Kluepfel D. Sequences of three genes specifying xylanases in *Streptomyces lividans*. *Gene.* 1991; 107(1): 75–82.
- 201 Kataeva IA, Seidel RD, Shah A, West LT, Li X-L, Ljungdahl LG. The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. *Appl Environ Microbiol.* 2002; 68(9): 4292–300.
- 202 Kim DY, Han MK, Park D-S, Lee JS, Oh H-W, Shin D-H, *et al.* Novel GH10 xylanase, with a fibronectin type 3 domain, from *Cellulosimicrobium* sp. strain HY-13, a bacterium in the gut of *Eisenia fetida*. *Appl Environ Microbiol.* 2009; 75(22): 7275–9.
- 203 Mello BL. Biophysical characterizations of *Cellulomonas fimi* hypothetical protein, Celf\_0121. [PhD]. University of Sao Paolo; 2017.
- 204 Rigden DJ, Mello L V, Galperin MY. The PA14 domain, a conserved all- $\beta$  domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends Biochem Sci.* 2004; 29(7): 335–9.
- 205 Duine JA. The PQQ story. *J Biosci Bioeng.* 1999; 88(3): 231–6.

- 206 Bauer R, Janowska K, Taylor K, Jordan B, Gann S, Janowski T, *et al.* Structures of three polycystic kidney disease-like domains from *Clostridium histolyticum* collagenases ColG and ColH. *Acta Crystallogr D Biol Crystallogr.* 2015; 71(3):565–77.
- 207 Yoshida E, Hidaka M, Fushinobu S, Koyanagi T, Minami H, Tamaki H, *et al.* Role of a PA14 domain in determining substrate specificity of a glycoside hydrolase family 3  $\beta$ -glucosidase from *Kluyveromyces marxianus*. *Biochem J.* 2010; 431(1): 39-49.
- 208 Chen M-H, Chen K-S, Hou J-W, Lee C-C, Huang J-S. Coexistence of autosomal dominant polycystic kidney disease and neurofibromatosis: report of a family. *Am J Nephrol.* 2002; 22(4): 376–80.
- 209 Bycroft M, Bateman A, Clarke J, Hamill SJ, Sandford R, Thomas RL, *et al.* The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. *EMBO J.* 1999; 18(2): 297–305.
- 210 Hughes J, Ward CCJ, Peral B, Aspinwall R, Clark K, San Millán JLL, *et al.* The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. *Nat Genet.* 1995; 10(2): 151–60.
- 211 Orikoshi H, Nakayama S, Hanato C, Miyamoto K, Tsujibo H. Role of the N-terminal polycystic kidney disease domain in chitin degradation by chitinase A from a marine bacterium, *Alteromonas* sp. strain O-7. *J Appl Microbiol.* 2005; 99(3): 551–7.
- 212 Wong CM, Wong KH, Chen XD, Ming Wong C, Hei Wong K, Dong Chen X. Glucose oxidase: natural occurrence, function, properties and industrial applications. *Appl Microbiol Biotechnol.* 2008; 78(6): 927–38.
- 213 Bauer R, Wilson JJ, Philominathan STL, Davis D, Matsushita O, Sakon J. Structural comparison of ColH and ColG collagen-binding domains from *Clostridium histolyticum*. *J Bacteriol.* 2013; 195(2): 318–27.
- 214 Suma K, Podile AR. Chitinase A from *Stenotrophomonas maltophilia* shows transglycosylation and antifungal activities. *Bioresour Technol.* 2013; 133: 213–20.
- 215 Jing H, Takagi J, Liu J, Lindgren S, Zhang R, Joachimiak A, *et al.* Archaeal surface layer proteins contain b-propeller, PKD, and b-helix domains and are related to metazoan cell surface proteins. *Structure.* 2002; 10(10): 1453-64.
- 216 Huang J, Wu C, Liu D, Yang X, Wu R, Zhang J, *et al.* C-terminal domains of bacterial proteases: structure, function and the biotechnological applications. *J Appl Microbiol.* 2016; 1–11.
- 217 Sandford R, Sgotto B, Aparicio S, Brenner S, Vaudin M, Wilson RK, *et al.* Comparative analysis of the polycystic kidney disease 1 (PKD1) gene reveals an integral membrane glycoprotein with multiple evolutionary conserved domains. *Hum Mol Genet.* 1997;6(9):1483-9.
- 218 Wang X, Chi N, Bai F, Du Y, Zhao Y, Yin H. Characterization of a cold-adapted and salt-tolerant exo-chitinase (ChiC) from *Pseudoalteromonas* sp. DL-6. *Extremophiles.* 2016; 20(2): 167–76.
- 219 Malecki PH, Raczynska JE, Vorgias CE, Rypniewski W. Structure of a complete four-domain chitinase from *Moritella marina*, a marine psychrophilic bacterium. *Acta Crystallogr Sect D Biol Crystallogr.* 2013; 69(5): 821–9.

- 220 Yang H, Liu L, Xu F. The promises and challenges of fusion constructs in protein biochemistry and enzymology. *Appl Microbiol Biotechnol.* 2016; 100(19): 8273–81.
- 221 Ouidir T, Jarnier F, Cosette P, Jouenne T, Hardouin J. Characterization of N-terminal protein modifications in *Pseudomonas aeruginosa* PA14. *J Proteomics.* 2014; 114(2): 214–25.
- 222 de Groot PWJ, Klis FM. The conserved PA14 domain of cell wall-associated fungal adhesins governs their glycan-binding specificity. *Mol Microbiol.* 2008; 68(3): 535–7.
- 223 Larsbrink J, Izumi A, Ibatullin FM, Nakhai A, Gilbert HJ, Davies GJ, *et al.* Structural and enzymatic characterization of a glycoside hydrolase family 31  $\alpha$ -xylosidase from *Cellvibrio japonicus* involved in xyloglucan saccharification. *Biochem J.* 2011; 436(3): 567–80.
- 224 Gruninger RJ, Gong X, Forster RJ, McAllister TA. Biochemical and kinetic characterization of the multifunctional  $\beta$ -glucosidase/ $\beta$ -xylosidase/ $\alpha$ -arabinosidase, Bgxa1. *Appl Microbiol Biotechnol.* 2014; 98(7): 3003–12.
- 225 Bayer EA, Shoham Y, Lamed R. Lignocellulose-decomposing bacteria and their enzyme systems. In: *The Prokaryotes: Prokaryotic Physiology and Biochemistry.* 2013; 215–266.
- 226 Goossens KVV, Ielasi FS, Nookaew I, Stals I, Alonso-Sarduy L, Daenen L, *et al.* Molecular mechanism of flocculation self-recognition in yeast and its role in mating and survival. *MBio.* 2015; 6(2): e00427–15.
- 227 Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J.* 1991; 309–16.
- 228 Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol Mol Biol Rev.* 2002; 66(3): 506–77.
- 229 Zhang Y-HP, Cui J, Lynd LR, Kuang LR. A transition from cellulose swelling to cellulose dissolution by o-phosphoric acid: evidence from enzymatic hydrolysis and supramolecular structure. *Biomacromolecules.* 2006 ;7(2): 644–8.
- 230 Duine JA, Jzn. JF, Van Zeeland JK. Glucose dehydrogenase from *Acinetobacter calcoaceticus*. *FEBS Lett.* 1979; 108(2): 443–6.
- 231 Anthony C. The structure of bacterial quinoprotein dehydrogenases. *Int. J. Biochem.* 1992; 24(1): 29–39.
- 232 Oubrie A, Tte H, Rozeboom J, Kalk KH, Olsthoorn AJJ, Duine JA, *et al.* Structure and mechanism of soluble quinoprotein glucose dehydrogenase. *EMBO J.* 1999; 18(19): 5187–94.
- 233 Mitchell AE, Jones AD, Mercer RS, Rucker RB. Characterization of pyrroloquinoline quinone amino acid derivatives by electrospray ionization mass spectrometry and detection in human milk. *Anal Biochem.* 1999; 269(2): 317–25.
- 234 Yoshida H, Araki N, Tomisaka A, Sode K. Secretion of water soluble pyrroloquinoline quinone glucose dehydrogenase by recombinant *Pichia pastoris*. *Enzyme Microb Technol.* 2002; 30(3): 312–8.



- 235 Matsushita, H. Toyama, M. Yamada, O K, Toyama H, Yamada M, Adachi O. Quinoproteins: structure, function, and biotechnological applications. *Appl Microbiol Biotechnol.* 2002; 58(1): 13–22.
- 236 Mathieu Y, Piumi F, Valli R, Aramburu JC, Ferreira P, Faulds CB, *et al.* Activities of secreted aryl alcohol quinone oxidoreductases from *Pycnoporus cinnabarinus* provide insights into fungal degradation of plant biomass. *Appl Environ Microbiol.* 2016; 82(8): 2411–23.
- 237 Rucker R, Chowanadisai W, Nakano M. Potential physiological importance of pyrroloquinoline quinone. *Altern Med Rev.* 2009; 14(3): 268-77.
- 238 Duine JA, Strampraad MJF, Hagen WR, de Vries S. The cooperativity effect in the reaction of soluble quinoprotein (PQQ-containing) glucose dehydrogenase is not due to subunit interaction but to substrate-assisted catalysis. *FEBS J.* 2016; 283(19): 3604–12.
- 239 Hofer M, Bönsch K, Greiner-Stöffele T, Ballschmiter M. Characterization and Engineering of a Novel Pyrroloquinoline Quinone Dependent Glucose Dehydrogenase from *Sorangium cellulosum* So ce56. *Mol Biotechnol.* 2011; 47(3): 253–61.
- 240 Southall SM, Doel JJ, Richardson DJ, Oubrie A. Soluble aldose sugar dehydrogenase from *Escherichia coli*: a highly exposed active site conferring broad substrate specificity. *J Biol Chem.* 2006; 281(41): 30650–9.
- 241 Sakuraba H, Yokono K, Yoneda K, Watanabe A, Asada Y, Satomura T, *et al.* Catalytic properties and crystal structure of quinoprotein aldose sugar dehydrogenase from hyperthermophilic archaeon *Pyrobaculum aerophilum*. *Arch Biochem Biophys.* 2010; 502(2): 81–8.
- 242 Miyazaki T, Sugisawa T, Hoshino T. Pyrroloquinoline quinone-dependent dehydrogenases from *Ketogulonicigenium vulgare* catalyze the direct conversion of L-sorbosone to L-ascorbic acid. *Appl Environ Microbiol.* 2006; 72(2): 1487–95.
- 243 Matsumura H, Umezawa K, Takeda K, Sugimoto N, Ishida T, Samejima M, *et al.* Discovery of a Eukaryotic Pyrroloquinoline Quinone-Dependent oxidoreductase belonging to a new auxiliary activity family in the database of carbohydrate-active enzymes. *PLoS One.* 2014; 9(8): e104851.
- 244 Takeda K, Matsumura H, Ishida T, Samejima M, Ohno H, Yoshida M, *et al.* Characterization of a novel PQQ-dependent quinohemoprotein pyranose dehydrogenase from *Coprinopsis cinerea* classified into auxiliary activities family 12 in carbohydrate-active enzymes. *PLoS One.* 2015; 10(2): e0115722.
- 245 Garajova S, Mathieu Y, Beccia MR, Bennati-Granier C, Biaso F, Fanuel M, *et al.* Single-domain flavoenzymes trigger lytic polysaccharide monooxygenases for oxidative degradation of cellulose. *Sci Rep.* 2016; (6): 28276. doi:10.1038/srep28276
- 246 Sun MZ, Zhang XY, Xin Y. Purification and characterization of an endo-D-arabinase produced by *Cellulomonas*. *Protein J.* 2012; 31(1) :51-8. doi: 10.1007/s10930-011-9374-5.
- 247 Yelton MM, Hamer JE, Timberlake WE. Transformation of *Aspergillus nidulans* by using a trpC plasmid. *Genetics.* 1984; 81: 1470–4.

- 248 Alshahni MM, Makimura K, Yamada T, Satoh K, Ishihara Y, Takatori K, *et al.* Direct Colony PCR of Several Medically Important Fungi Using Ampdirect<sup>®</sup> Plus. *Jpn J Infect Dis.* 2009; 62: 164–7.
- 249 Fogg MJ, Wilkinson AJ. New Methods for the Study of Protein–Nucleic Acid Interactions Higher-throughput approaches to crystallization and crystal structure determination. *Biochem. Soc. Trans.* 2008; 36: 771–775. doi:10.1042/BST0360771
- 250 Vincentelli R, Canaan S, Offant J, Cambillau C, Bignon C. Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal Biochem.* 2005; 346(1): 77–84.
- 251 Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, Berkmen M. SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microb Cell Fact.* 2012; 11: 56.
- 252 Ferrè F, Clote P. DiANNA: A web server for disulfide connectivity prediction. *Nucleic Acids Res.* 2005; 33(SUPPL. 2): 230–2.
- 253 Boeke JD, LaCroute F, Fink GR. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol Gen Genet.* 1984; 197(2): 345–6.
- 254 Ameyama M, Shinagawa E, Matsushita K, Adachi O. D-Glucose Dehydrogenase of *Gluconobacter suboxydans*: Solubilization, Purification and Characterization. *Agric Biol Chem.* 1981; 45(4): 851–61.
- 255 Yamada M, Elias MD, Matsushita K, Migita CT, Adachi O, Elias MD, *et al.* *Escherichia coli* PQQ-containing quinoprotein glucose dehydrogenase: its structure comparison with other quinoproteins. *Biochim Biophys Acta - Proteins Proteomics.* 2003; 1647(1–2): 185–92.
- 256 Olsthoorn AJJ, Duine JA. Production, characterization, and reconstitution of recombinant quinoprotein glucose dehydrogenase (soluble type; EC 1.1.99.17) apoenzyme of *Acinetobacter calcoaceticus*. *Arch Biochem Biophys.* 1996; 336(1): 42–8.
- 257 Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol.* 2001; 305(3): 567–80.
- 258 Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, Frishman D. Protein solubility: sequence based prediction and experimental verification. 2007;23(19):2536–42.
- 259 Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 1982; 157(1): 105–32.
- 260 Lolkema JS, Slotboom DJ. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Mol Membr Biol.* 1998; 15(1): 33–42.
- 261 White SH. Hydropathy Plots and the Prediction of Membrane Protein Topology. In: *Membrane Protein Structure.* New York, NY: Springer New York; 1994. 97–124.
- 262 Baneyx F, Mujacic M. Recombinant protein folding and misfolding in *Escherichia coli*. *Nat Biotechnol.* 2004; 22(11): 1399–408.

- 263 Harper S, Speicher DW. Purification of proteins fused to glutathione S-transferase. *Methods Mol Biol.* 2011; 681: 259–80.
- 264 Novick RP, Londono-Vallejo A, Harry E, Stragier P, Losick R. Genetic systems in *Staphylococci*. *Methods Enzymol.* 1991; 204(2): 587–636.
- 265 Porowińska D, Czarnecka J, Komoszyński M. Chaperones Are Necessary for the Expression of Catalytically Active Potato Apyrases in Prokaryotic Cells. 2014; 173(6): 1349-59. doi: 10.1007/s12010-014-0858-6.
- 266 Hartinger D, Heintl S, Schwartz HE, Grabherr R, Schatzmayr G, Haltrich D, *et al.* Enhancement of solubility in *Escherichia coli* and purification of an aminotransferase from *Sphingopyxis* sp. MTA144 for deamination of hydrolyzed fumonisins B 1. *Microb Cell Fact.* 2010; 9: 62. doi: 10.1186/1475-2859-9-62.
- 267 Bessette PH, Åslund F, Beckwith J, Georgiou G. Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proc Natl Acad Sci USA.* 1999; 96(24): 13703-08.
- 268 Chen J, Song JL, Zhang S, Wang Y, Cui DF, Wang CC. Chaperone activity of DsbC. *J Biol Chem.* 1999; 274(28): 19601–5.
- 269 Delisa MP, Tullman D, Georgiou G, Beckwith J. Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. *Proc Natl Acad Sci USA.* 2003; 100(10): 6115-20.
- 270 Hearn MTW, Acosta D. Applications of novel affinity cassette methods: use of peptide fusion handles for the purification of recombinant proteins. *J Mol Recognit.* 2001; 14(6): 323–69.
- 271 Hedhammar M, Alm T, Gräslund T, Hober S. Single-step recovery and solid-phase refolding of inclusion body proteins using a polycationic purification tag. *Biotechnol J.* 2006; 1(2): 187–96.
- 272 Terpe K. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol.* 2003; 60(5): 523–33.
- 273 Boyer TD. Special article the glutathione S-transferases: An update. *Hepatology.* 1989; 9(3): 486–96.
- 274 Smith DB, Johnson KS. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene.* 1988; 67(1): 31–40.
- 275 Frangioni JV, Neel BG. Solubilization and Purification of Enzymatically Active Glutathione S-Transferase (pGEX) Fusion Proteins. *Anal Biochem.* 1993; 210(1): 179–87.
- 276 Smith DB. Generating fusions to glutathione S-transferase for protein studies. *Methods Enzymol.* 2000; 326: 254–70.
- 277 Kaplan W, Husler P, Klump H, Erhardt J, Sluis-Cremer N, Dirr ' H. Conformational stability of pGEX-expressed *Schistosoma japonicum* glutathione S-transferase: A detoxification enzyme and fusion-protein affinity tag. *Protein Sci.* 1997; 6(2): 399-406.
- 278 Pina AS, Lowe CR, Cecilia A, Roque A. Challenges and opportunities in the purification of recombinant tagged proteins. *Biotechnol Adv.* 2014; 32: 366–81.

- 279 Meyer V, Minkwitz S, Schütze T, van den Hondel CAMJJ, Ram AFJ. The *Aspergillus niger* RmsA protein: A node in a genetic network? *Commun Integr Biol.* 2010; 3(2): 195–7.
- 280 Oates N. Mining Compost for Novel Lignocellulosic Enzymes from *Graphium* sp [PhD thesis]. University of York; 2017.
- 281 Isaksen T, Westereng B, Aachmann FL, Agger JW, Kracher D, Kittl R, *et al.* A C4-oxidizing Lytic Polysaccharide Monooxygenase Cleaving Both Cellulose and Cello-oligosaccharides. *J. Biol. Chem.* 2013; 289: 2632–42.
- 282 Oubrie A, Tte H, Rozeboom J, Dijkstra BW. Active-site structure of the soluble quinoprotein glucose dehydrogenase complexed with methylhydrazine: A covalent cofactor- inhibitor complex. *Proc Natl Acad Sci USA.* 1999; 96(21): 11787–91.
- 283 Ferri S, Kojima K, Sode K. Review of glucose oxidases and glucose dehydrogenases: a bird's eye view of glucose sensing enzymes. *J Diabetes Sci Technol.* 2011; 5(5): 1068–76.
- 284 Lambertz C, Garvey M, Klinger J, Heesel D, Klose H, Fischer R, *et al.* Challenges and advances in the heterologous expression of cellulolytic enzymes: a review. *Biotechnol Biofuels.* 2014; 7(1): 135.
- 285 Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Front Microbiol.* 2014; 5: 172.
- 286 Li Y-L, Li H, Li A-N, Li D-C. Cloning of a gene encoding thermostable cellobiohydrolase from the thermophilic fungus *Chaetomium thermophilum* and its expression in *Pichia pastoris*. *J Appl Microbiol.* 2009; 106(6): 1867–75.
- 287 Liu R, Chen L, Jiang YP, Zhou ZH, Zou G, Punt P, *et al.* Efficient genome editing in filamentous fungus *Trichoderma reesei* using the CRISPR/Cas9 system. *Cell Discov.* 2015; 1: 15007. doi:10.1038/celldisc.2015.7
- 288 Zhu Z, González F, Huangfu D. The iCRISPR Platform for Rapid Genome Editing in Human Pluripotent Stem Cells. *Methods Enzymol.* 2014; 546C: 215–50.
- 289 Liu Q, Gao R, Li J, Lin L, Zhao J, Sun W, *et al.* Development of a genome-editing CRISPR/Cas9 system in thermophilic fungal *Myceliophthora* species and its application to hyper-cellulase production strain engineering. *Biotechnol Biofuels.* 2017; 10(1): 1.
- 290 Xu T, Li Y, Shi Z, Hemme CL, Li Y, Zhu Y, *et al.* Efficient Genome Editing in *Clostridium cellulolyticum* via CRISPR-Cas9 Nickase. *Appl Environ Microbiol.* 2015; 81(13): 4423–31.
- 291 Mougias I, Bosma EF, de Vos WM, van Kranenburg R, van der Oost J. Next Generation Prokaryotic Engineering: The CRISPR-Cas Toolkit. *Trends Biotechnol.* 2016; 34(7): 575–87.
- 292 Hansen MAT, Ahl LI, Pedersen HL, Westereng B, Willats WGT, Jørgensen H, *et al.* Extractability and digestibility of plant cell wall polysaccharides during hydrothermal and enzymatic degradation of wheat straw (*Triticum aestivum* L.). *Ind Crops Prod.* 2014; 55: 63–69.
- 293 Wilson DB. Three microbial strategies for plant cell wall degradation. In: *Annals of the New York Academy of Sciences.* 2008; 1125: 289–97. doi: 10.1196/annals.1419.026.

- 294 Jeoh T, Ishizawa CI, Davis MF, Himmel ME, Adney WS, Johnson DK. Cellulase digestibility of pretreated biomass is limited by cellulose accessibility. *Biotechnol Bioeng.* 2007; 98(1): 112-22.
- 295 Bhalla A, Bansal N, Kumar S, Bischoff KM, Sani RK. Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes. *Bioresour Technol.* 2013; 128: 751–9.
- 296 Bartley LE, Ronald PC. Plant and microbial research seeks biofuel production from lignocellulose. *California Agriculture.* 2009; 63(4): 178-184.
- 297 Ragauskas AJ, Laser J, Sinskey AJ, Ragauskas A, Williams C, Davison B, *et al.* The Path Forward for Biofuels and Biomaterials. *Science.* 2006; 311(5760):484–9.
- 298 Kurosawa K, Laser J, Sinskey AJ. Tolerance and adaptive evolution of triacylglycerol-producing *Rhodococcus opacus* to lignocellulose-derived inhibitors. *Biotechnol Biofuels.* 2015; 8: 76. doi: 10.1186/s13068-015-0258-3
- 299 Rai PK, Singh SP, Asthana RK, Singh SP. Biohydrogen production from sugarcane bagasse by integrating dark- and photo-fermentation. *Bioresour Technol.* 2014; 152: 140-146.
- 300 Whittle DJ, Kilburn DG, Warren RA, Miller Jr. RC, Miller RC. Molecular cloning of a *Cellulomonas fimi* cellulose gene in *Escherichia coli*. *Gene.* 1982; 17(2): 139–45.
- 301 Poulsen OM, Petersen LW. Growth of *Cellulomonas* sp. ATCC 21399 on different polysaccharides as sole carbon source Induction of extracellular enzymes. *Appl Microbiol Biotechnol.* 1988; 29(5): 480-84.
- 302 Plucain J, Suau A, Cruveiller S, Médigue C, Schneider D, Le Gac M. Contrasting effects of historical contingency on phenotypic and genomic trajectories during a two-step evolution experiment with bacteria. *BMC Evol Biol.* 2016; 16: 86. doi: 10.1186/s12862-016-0662-8.
- 303 Martin M, Holscher T, Dragos A, Cooper VS, Kovacs and Akos T. Laboratory evolution of microbial interactions in bacterial biofilms. *J. of Bacteriol.* 2016. doi:10.1128/JB.01018-15
- 304 Charlesworth J, Eyre-Walker A. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol.* 2006; 23(7): 1348-56.
- 305 Rajaraman E, Agarwal A, Crigler J, Seipelt-Thiemann R, Altman E, Eiteman MA. Transcriptional analysis and adaptive evolution of *Escherichia coli* strains growing on acetate. *Appl Microbiol Biotechnol.* 2016; 100(17): 7777–85.
- 306 Großkopf T, Consuegra J, Gaffé J, Willison JC, Lenski RE, Soyer OS, *et al.* Metabolic modelling in a dynamic evolutionary framework predicts adaptive diversification of bacteria in a long-term evolution experiment. *BMC Evol Biol.* 2006; 16 (1): 163.
- 307 Schoustra S, Punzalan D. Correlation of mycelial growth rate with other phenotypic characters in evolved genotypes of *Aspergillus nidulans*. *Fungal Biol.* 2012; 116(5): 630–6.
- 308 Patyshakuliyeva A, Arentshorst M, Allijn IE, Ram AFJ, de Vries RP, Gelber IB. Improving cellulase production by *Aspergillus niger* using adaptive evolution. *Biotechnol Lett.* 2016; 38(6): 969–74.

- 309 Dalgaard P, Ross T, Kamperman L, Neumeyer K, McMeekin TA. Estimation of bacterial growth rates from turbidimetric and viable count data. *Int J Food Microbiol.* 1994; 23(3): 391–404.
- 310 Bollet C, Gevaudan MJ, De Lamballerie X, Zandotti C, De Micco P. A simple method for the isolation of chromosomal DNA from Gram positive or acid-fast bacteria. *Nucleic Acids Res.* 1955; 19(8): 1955.
- 311 Yu Ip CC, Manam V, Hepler R, Hennessey JP. Carbohydrate composition analysis of bacterial polysaccharides: Optimized acid hydrolysis conditions for HPAEC-PAD analysis. *Anal Biochem.* 1992; 201(2): 343–9.
- 312 Chomczynski P, Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nat Protoc.* 2006;1(2):581–5.
- 313 Hatayama K, Esaki K, Ide T. *Cellulomonas soli* sp. nov. and *Cellulomonas oligotrophica* sp. nov., isolated from soil. *Int J Syst Evol Microbiol.* 2013; (63): 60–5.
- 314 Shi Z, Luo G, Correspondence GW, Wang G. *Cellulomonas carbonis* sp. nov., isolated from coal mine soil. *Int J Syst Evol Microbiol.* 2012; 62: 2004-10.
- 315 Bremer H. Variation of generation times in *Escherichia coli* populations: its cause and implications. *J Gen Microbiol.* 1982; 128(12): 2865–76.
- 316 Powell E O. Growth Rate and Generation Time of Bacteria, with Special Reference to Continuous Culture. *J Gen Microbiol.* 1956; 15(3): 492-511.
- 317 Blet-Charaudeau C, Muller J, Laudelout H. Kinetics of Carbon Dioxide Evolution in Relation to Microbial Biomass and Temperature. *Soil Sci Soc Am J.* 1990; 54(5): 1324.
- 318 Lagaert S, Beliën T, Volckaert G. Plant cell walls: Protecting the barrier from degradation by microbial enzymes. *Seminars in Cell and Developmental Biology.* 2009; 20: 1064-73.
- 319 van Beelen P, Doelman P. Significance and application of microbial toxicity tests in assessing ecotoxicological risks of contaminants in soil and sediment. *Chemosphere.* 1997; 34(3): 455–99.
- 320 Dilly O. Regulation of the respiratory quotient of soil microbiota by availability of nutrients. *FEMS Microbiol Ecol.* 2003; 43(3): 375–81.
- 321 IAWQ Task Group on Respirometry. H, Spanjers H, International Association on Water Quality. G, Dold PL. Respirometry in control of the activated sludge process: principles. *International Association on Water Quality;* 1998. 48.
- 322 Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, *et al.* Species Interactions Alter Evolutionary Responses to a Novel Environment. Ellner SP, editor. *PLoS Biol.* 2012;10(5): e1001330.
- 323 Alfenore S, Molina-Jouve C. Current status and future prospects of conversion of lignocellulosic resources to biofuels using yeasts and bacteria. *Process Biochem.* 2016; 51(11): 1747-56.

- 324 Kovács R, Házi F, Csikor Z, Miháltz P. Connection between oxygen uptake rate and carbon dioxide evolution rate in aerobic thermophilic sludge digestion. *Ř Period Polytech Chem Eng.* 2007; 511: 17–22.
- 325 Mei C-F, Liu Y-Z, Long W-N, Sun G-P, Zeng G-Q, Xu M-Y, *et al.* A comparative study of biodegradability of a carcinogenic aromatic amine (4,4'-Diaminodiphenylmethane) with OECD 301 test methods. *Ecotoxicol Environ Saf.* 2015; 111: 123–30.
- 326 Zhang X, Rogowski A, Zhao L, Hahn MG, Avci U, Knox JP, *et al.* Understanding how the complex molecular architecture of mannan-degrading hydrolases contributes to plant cell wall degradation. *J Biol Chem.* 2014; 289, 2002-12. doi: 10.1074/jbc.M113.527770
- 327 Boraston AB, McLean BW, Guarna MM, Amandaron-Akow E, Kilburn DG. A family 2a carbohydrate-binding module suitable as an affinity tag for proteins produced in *Pichia pastoris*. *Protein Expr Purif.* 2001; 21(3): 417–23.
- 328 Ratnayake S, Beahan CT, Callahan DL, Bacic A. The reducing end sequence of wheat endosperm cell wall arabinoxylans. *Carbohydr Res.* 2014; 386. 23-32.
- 329 Harris PJ, Stone BA. Chemistry and Molecular Organization of Plant Cell Walls. In: Biomass Recalcitrance. Oxford, UK: Blackwell Publishing Ltd.; 2008. 61–93.
- 330 Sarvas M, Harwood CR, Bron S, Van Dijl JM. Post-translocational folding of secretory proteins in Gram-positive bacteria. *Biochimica et Biophysica Acta - Molecular Cell Research.* 2004; 16941-3: 311-327.
- 331 Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol Biofuels.* 2013; 6: 1.
- 332 Boraston AB, Nurizzo D, Notenboom V, Ducros V, Rose DR, Kilburn DG, *et al.* Differential Oligosaccharide Recognition by Evolutionarily-related  $\beta$ -1,4 and  $\beta$ -1,3 Glucan-binding Modules. *J Mol Biol.* 2002; 319(5): 1143–56.
- 333 Ries L, Pullan ST, Delmas S, Malla S, Blythe MJ, Archer DB. Genome-wide transcriptional response of *Trichoderma reesei* to lignocellulose using RNA sequencing and comparison with *Aspergillus niger*. *BMC Genomics.* 2013; 14(1): 541.
- 334 Rytioja J, Hildén K, Hatakka A, Mäkelä MR. Transcriptional analysis of selected cellulose-acting enzymes encoding genes of the white-rot fungus *Dichomitus squalens* on spruce wood and microcrystalline cellulose. *Fungal Genet Biol.* 2014 ;72: 91-8.
- 335 Chopra S, Ramkissoon K, Anderson DC. A systematic quantitative proteomic examination of multidrug resistance in *Acinetobacter baumannii*. *J Proteomics.* 2013; 84(0): 17–39.
- 336 Matulich KL, Martiny JBH. Microbial composition alters the response of litter decomposition to environmental change. *Ecology.* 2015; 96(1): 154–63.
- 337 Portillo F, Yashchuk O, Hermida É. Evaluation of the rate of abiotic and biotic degradation of oxo-degradable polyethylene. *Polym Test.* 2016; 53: 58–69.
- 338 Raeside C, Gaffé J, Deatherage DE, Tenailon O, Briska AM, Ptashkin RN, *et al.* Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *MBio.* 2014;5(5): e01377-14.

- 339 Kim H-D, Choi S-L, Kim H, Sohn JH, Lee S-G. Enzyme-linked assay of cellulose-binding domain functions from *Cellulomonas fimi* on multi-well microtiter plate. *Biotechnol Bioprocess Eng.* 2013; 18(3): 575–80.
- 340 Garrett TR, Bhakoo M, Zhang Z. Bacterial adhesion and biofilms on surfaces. *Prog Nat Sci.* 2008; 18(9): 1049–56.
- 341 Feng G, Cheng Y, Wang S-Y, Borca-Tasciuc DA, Worobo RW, Moraru CI. Bacterial attachment and biofilm formation on surfaces are reduced by small-diameter nanoscale pores: how small is small enough? *Nat Publ Gr.* 2015; 1. doi:10.1038/npjbiofilms.2015.22
- 342 French CE, Barnard DK, Fletcher E, Kane SD, Lakhundi SS, Liu C-K, *et al.* Synthetic Biology for Biomass Conversion. In: *New and Future Developments in Catalysis.* Elsevier. 2013; 115–40.
- 343 Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* 2012; 13(3): R23.
- 344 Peano C, Pietrelli A, Consolandi C, Rossi E, Petiti L, Tagliabue L, *et al.* An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb Inform Exp.* 2013; 3: 1.
- 345 Rajoka MI, Malik KA. Enhanced Production of Cellulases by *Cellulomonas* Strains Grown on Different Cellulosic Residues. *Folia Microbiol.* 1997; 42(1): 59–64.