Metasecretome phage display: A new approach for mining surface and secreted proteins from microbial communities

A thesis presented in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biochemistry at Massey University, Palmerston North, New Zealand

Milica Ciric

2014

Abstract

The microbial community residing in the reticulo-rumen degrades plant material to provide energy for its ruminant host. The key enzymes and proteins for plant fibre degradation are secreted from the microbial cells, and form part of the 'metasecretome' - the collection of cell-surface and secreted proteins that mediate important interactions between the microbiota and their rumen environment. Consequently, the metasecretome represents a valuable repository of bioactivities with potential applications in enhancing fibre digestibility and feed efficiency in ruminant animals, and in improving the depolymerisation of lignocellulosic feedstocks for biofuel production.

A new metasecretome phage display approach was developed in this thesis, with the aim to focus sequencing efforts on the metasecretome-encoding component of complex microbial community genomes (metagenomes). This was achieved by combining secretome-selective phage display at a metagenomic scale with next-generation sequence analysis. The ability of this approach to focus metagenome mining onto sequences encoding surface and secreted proteins from the highly fibrolytic rumen plant-adherent microbiota of a dairy cow has been assessed.

The metasecretome selection protocol efficiently enriched for a broad spectrum of metasecretome protein coding sequences, both in terms of the taxonomic and functional diversity, and the membrane-targeting signals present. This allowed *in silico* identification of functionally diverse surface and secreted carbohydrate-active enzymes (CAZymes). In particular, the metasecretome dataset was enriched for sequences encoding putative components characteristic of cellulosomes, the cell-surface multi-protein structures specialised for the degradation of plant fibre.

Over one-sixth of the putative CAZymes identified in the metasecretome dataset shared a low sequence similarity with putative CAZymes identified through previous genomic and metagenomic studies; hence this work has identified proteins that potentially have novel carbohydrate-active functions.

Affinity screening of the metagenomic phage display library on amorphous cellulose and arabinoxylan significantly enriched for a putative serine/threonine protein kinase. *In silico* analyses have not associated this protein with recognised carbohydrate binding functions, thus the observed binding may have not been carbohydrate specific.

Overall, the methodology developed in this thesis is applicable for the high-throughput metasecretome exploration and is complementary to existing strategies used for mining surface and secreted proteins of complex microbial communities.

Acknowledgments

I wish to express my sincere gratitude to my supervisors, Dr Dragana Gagic, Dr Christina Moon, Dr Jasna Rakonjac and Dr Graeme Attwood. This PhD project has been an intensive and at times bumpy journey, which I would have not completed without your support, guidance, encouragement and most importantly, patience.

I gratefully acknowledge AgResearch and Institute of Fundamental Sciences (IFS) for their financial support, Ministry of Business, Innovation and Employment for funding this project and Ruakura Animal Ethics Committee for granting animal ethics permission. I would also like to thank IFS, New Zealand Society of Animal Production and New Zealand Society for Biochemistry and Molecular Biology for financial assistance to present thesis work at Genomics of Energy & Environment annual JGI meeting in 2013.

I would also like to thank the following people for all their help: Roger Moraga Martinez for being a true bioinformatics guru; Dr Sinead Leahy, Dr Eric Altermann, Dr Chris Creevey and Dr Yanbin Yin for bioinformatics advice; Dr Garry Waghorn, Carrie Sang and Dong Li for help with cow rumen samplings; Dr Bill Kelly and Dr Adrian Cookson for providing feedback on various outputs from this thesis; Dr Samantha Noel for always being willing to share her wisdom on all things ruminant (and beyond) and Milivoje Gencic for around the clock IT and formatting advice.

I feel very fortunate that I had a chance to interact with friendly people passionate about science from AgResearch (especially its rumen microbiology team), The Hot Zone (former Helipad) lab, IFS and Massey University. All current and former student room occupants, especially Sonal and Tom, thank you for providing such great social environment and fun.

Special acknowledgment goes to my mum Nevenka for all her love and support, to whom I would like to dedicate this thesis.

Table of contents

Abstract.....iii

Acknowledgments..... v

Table of contents..... vii

List of tables xiii

List of figures.....xv

Abbreviations..... xvii

Chapter 1. Literature review 19

1.1 Introduction.....19

1.2 Rumen 20

1.2.1 Rumen, a continuous-flow fermenter for microbial fibre degradation 20

1.2.2 The rumen microbiota and its role in fibre digestion - an overview.....21

1.2.3 Bacterial diversity in the rumen 23

1.2.4 Mechanisms of attachment and degradation of plant material by ruminal fibrolytic bacteria 26

1.3 Carbohydrate-active enzymes involved in fibre degradation.....28

1.3.1 Unique aspects of grass cell wall architecture.....28

1.3.2 Strategies for cell wall digestion in the rumen 31

1.3.3 Carbohydrate Active Enzymes involved in fibre degradation.....33

1.3.3.1 Catalytic modules of cellulases 37

1.3.3.2 Catalytic modules of hemicellulases.....37

1.3.3.3 Catalytic modules of other fibre-degrading enzymes 39

1.3.3.4 Carbohydrate-binding modules of fibre-degrading enzymes 39

1.3.3.5 Cellulosomes 42

1.3.4 Metagenomic studies of fibre-degrading genes of rumen microbiomes 45

1.3.4.1 Metagenomics and next-generation sequencing technologies.....45

1.3.4.2 Metagenomic studies of fibre-degrading rumen microbial communities.....48

1.4 Metasecretome 51

1.4.1 Definition of the bacterial secretome 51
1.4.2 Secretion pathways of monoderm and diderm bacteria 52
1.4.2.1 Protein transport systems universal for all bacteria 53
1.4.2.2 Protein export systems specific to monoderm bacteria 55
1.4.2.3 Protein export systems specific to diderm bacteria 57
1.4.3 Secretion and membrane targeting signals and their prediction 60
1.4.3.1 Type I signal sequences 62
1.4.3.2 Type II signal sequences 63
1.4.3.3 Tat signal sequences 63
1.4.3.4 Type IV signal sequences 64
1.4.3.5 Transmembrane α-helices 64
1.4.3.6 Non-classical secretion 65
1.4.4 Methods to study the secretome 65
1.5 Phage display 68
1.5.1 The life cycle of Ff phage used for phage display 68
1.5.2 Principles and applications of phage display 71
1.5.3 Overview of the secretome-selective phage display system 75
1.6 Aims of the project 79
Chapter 2. Materials and Methods 81
2.1 Materials 81
2.1.1 Laboratory chemicals and enzymes 81
2.1.2 Buffers, solutions and media 81
2.1.2.1 Standard buffers and solutions 81
2.1.2.2 DNA-free water 82
2.1.2.3 OrangeG loading dye 82
2.1.2.4 Buffers used for rumen content fractionation 82
2.1.2.5 Buffers used for extraction and shearing of metagenomic DNA 82
2.1.2.6 Phage concentration, purification and disassembly solutions/buffers 82
2.1.2.7 Liquid and solid media 83

2.1.3 Bacterial strains, plasmids and phage 83
2.1.4 Oligonucleotides 84
2.1.5 Bioinformatic resources and software 84
2.2 Methods 87
2.2.1 Bacterial strains and growth conditions 87
2.2.2 Molecular biology methods 87
2.2.2.1 DNA extraction and purification 87
2.2.2.2 Agarose gel electrophoresis 88
2.2.2.3 DNA quantification 88
2.2.2.4 Preparation and transformation of electro-competent *E. coli* cells 88
2.2.2.5 Polymerase chain reaction (PCR) amplification 89
2.2.2.5.1 Bacterial colony PCR 89
2.2.2.5.2 PCR amplification of ssDNA for pyrosequencing 90
2.2.3 Phage protocols 91
2.2.3.1 Phage propagation 91
2.2.3.2 Preparation of PPs 92
2.2.3.2.1 PPs preparation by liquid method 92
2.2.3.2.2 PPs preparation by plate method 92
2.2.3.3 Purification and concentration of phage and PPs 93
2.2.3.4 Disassembled virion gel electrophoresis 93
2.2.3.5 Native virion gel electrophoresis 93
2.2.3.6 Enumeration of phage and PPs 94
2.2.3.6.1 Titration of phage and PPs 94
2.2.3.6.2 Phage and PPs quantification by densitometry 94
2.2.4 Rumen content fractionation and metagenomic DNA extraction 95
2.2.4.1 Rumen sampling 95
2.2.4.2 Rumen content fractionation 95
2.2.4.3 Metagenomic DNA extraction from rumen microbial plant-adherent fraction 97

2.2.5 Rumen metasecretome phage display 98
2.2.5.1 Construction of rumen microbial metagenomic shotgun libraries 100
2.2.5.2 Selection of metasecretome phage display library and isolation of ssDNA.. 101
2.2.5.3 Sequencing of pilot metasecretome phage display library 101
2.2.5.4 Testing conditions for the next-generation sequencing template preparation 102
2.2.5.5 Next-generation sequencing of the metasecretome..... 103
2.2.6 Bioinformatic analysis 103
2.2.6.1 Sequence analysis of the pilot metasecretome library inserts..... 103
2.2.6.2 *In silico* analysis of the next-generation sequencing metasecretome dataset 104
2.2.6.2.1 Rumen metasecretome unassembled and assembled datasets..... 104
2.2.6.2.2 Functional annotation and phylogenetic profile 106
2.2.6.2.3 CAZyme annotation and taxonomic assignment..... 108
2.2.6.2.4 Assessment of the novelty of putative CAZymes detected in the metasecretome dataset..... 109
2.2.6.2.5 Prediction of membrane-targeting signals in the metasecretome dataset 109
2.2.7 Affinity screening of the metagenomic shotgun library..... 110
2.2.7.1 Preparation, immobilisation and test-assays of complex carbohydrate substrates for panning 110
2.2.7.2 Affinity screening of the metagenomic shotgun phage display library on wheat arabinoxylan and amorphous cellulose 111
2.2.7.3 Sequence analysis of the affinity selected recombinant PPs 113
2.2.7.4 Wheat arabinoxylan-binding assay of affinity-selected recombinant PPs..... 114
Chapter 3. Metasecretome-selective phage display approach for mining the functional potential of a rumen plant-adherent microbial community ...115
3.1 Construction of rumen plant-adherent metagenomic libraries and metasecretome selection 115
3.2 Pilot metasecretome phage display library..... 119
3.2.1 Estimated enrichment of the secretome insert-containing recombinant library clones 119

3.2.2 Pilot metasecretome library secretion signal types, functional annotations and taxonomy....120
3.2.3 Overview of sections 3.1 and 3.2....123
3.3 Metasecretome characterisation by next-generation sequencing....124
3.3.1 Establishing a protocol for preparing the pyrosequencing template from the metasecretome phage display library....124
3.3.2 Preparation of metasecretome template for next-generation sequencing....125
3.3.3 Next-generation sequence analysis of the metasecretome phage display library....127
3.3.4 Prediction of common types of membrane-targeting signals in the putative metasecretome proteins in frame with pIII....129
3.3.5 Phylogenetic profile of the metasecretome dataset....131
3.3.6 Functional annotation of the metasecretome dataset....133
3.3.7 Diversity of CAZyme families captured by metasecretome selection....135
3.3.8 Abundance and phylogenetic diversity of cellulosome components predicted in the metasecretome dataset....141
3.3.9 Assessment of the novelty of CAZymes detected in the metasecretome dataset....145
3.3.10 Overview of section 3.3....147
3.4 Summary....147
Chapter 4. Affinity screening of the metagenomic shotgun phage display library from the rumen plant-adherent microbiome for proteins mediating interactions with complex carbohydrates....149
4.1 Optimisation of complex carbohydrate affinity screening assays....149
4.2 Affinity screening of the metagenomic phage display library for carbohydrate-binding proteins on RAC and AXYL....150
4.3 Characterisation of affinity-selected clones....155
4.4 Affinity-binding assays....159
4.5 Summary....161
Chapter 5. Discussion....163
5.1 New phage display approach to select for the metasecretome....163
5.2 Metasecretome characterisation by next-generation sequencing....166

5.2.1 Membrane-targeting signals and phylogenetic profile of the metasecretome 166

5.2.2 The metasecretome selection enriched putative proteins involved in carbohydrate transport and metabolism 168

5.2.3 The metasecretome selection captured diverse CAZymes 170

5.2.4 CAZyme families enriched in the metasecretome 171

5.2.5 Architecture of metasecretome ORFs with predicted multi-modular CAZyme organisation 173

5.2.6 Phylogenetic diversity of cellulosome components predicted in the metasecretome 174

5.2.7 Assessment of the novelty of CAZymes detected in metasecretome dataset 175

5.3 Affinity screening of the metagenomic shotgun phage display library for carbohydrate-binding proteins..... 177

5.4 Study limitations 180

Chapter 6. Conclusions and future directions 183

6.1 Conclusions..... 183

6.2 Future directions..... 184

Appendices..... 187

Appendix 1 187

Appendix 2 193

Appendix 3 226

References 229

List of tables

Table 1.1 Major CAZyme families involved in the degradation of plant cell wall polysaccharides. 35
Table 1.2. Summary of the available NGS platforms and their outputs. 46
Table 1.3. Profiles of genes encoding selected GH families and cellulosome domains in four different rumen metagenomes. 50
Table 2.1 *Escherichia coli* strains used in this study. 83
Table 2.2 Plasmids and phage used in this study. 84
Table 2.3 Oligonucleotide primers used in this study. 84
Table 2.4 Bioinformatic resources and software. 85
Table 2.5 Components of a PCR reaction mixture for colony PCR. 89
Table 2.6 Thermal profile of the colony PCR reaction. 89
Table 2.7 Components of a PCR reaction mixture for ssDNA amplification. 90
Table 2.8 Thermal profile of the PCR reaction (rapid amplification protocol). 91
Table 2.9 Overview of samples used for generation of the metasecretome and the metagenome datasets. 107
Table 3.1 Summary statistics for the unassembled metasecretome datasets. 128
Table 3.2 Summary statistics of the assembled metasecretome dataset. 128
Table 3.3 Comparison of the 20 most abundant CAZyme families in the metasecretome and metagenome datasets. 137
Table 3.4 Profiles of selected GH families and cellulosome domains in the metasecretome and the metagenome datasets in comparison with four published rumen metagenomes. 139
Table 4.1 Binding of pDJ01 vector-derived PPs to complex carbohydrate substrates. 150
Table 4.2 Enrichment of metagenomic phage display library PPs through four rounds of affinity panning on complex carbohydrates. 152
Table 4.3 Binding of metagenomic phage display library PPs over background through four rounds of affinity panning on complex carbohydrates. 153
Table 4.4 Distribution of 40 analysed inserts in regard to ORF frame status. 157
Table 4.5 Recovery of PPs in affinity binding assays on AXYL. 160
Table A1.1. Predicted membrane targeting signals and annotation of putative proteins in the metasecretome pilot library 187
Table A2.1 Putative carbohydrate-active enzymes and associated modules identified in the metasecretome and the metagenome dataset 193
Table A2.2 CAZy families predicted at higher frequency in the metasecretome compared to the metagenome dataset 216

Table A2.3 CAZy families predicted at lower frequency in the metasecretome compared to the metagenome dataset …….....…….....217
Table A2.4.Candidate putative proteins with predicted multi-modular organisation in the metasecretome dataset.....……...221
Table A3.1. Analysis of 40 clones selected from the rumen microbial plant-adherent metagenomic phage display library by affinity screening on complex carbohydrate substrates……...226

List of figures

Figure 1.1 Generalised structure of the primary plant cell wall. 29
Figure 1.2 The basic structural components of hemicellulose and the hemicellulases responsible for their degradation. 38
Figure 1.3 Schematic overview of the *Ruminococcus flavefaciens* 17 cellulosome. 44
Figure 1.4 Protein export systems of monoderm (Gram-positive) bacteria. 55
Figure 1.5 . Protein export systems of diderm (Gram-negative) bacteria. 57
Figure 1.6 Schematic representation of the structure of common cytoplasmic membrane-targeting signals. 61
Figure 1.7 Schematic representation of Ff filamentous phage. 68
Figure 1.8 Life cycle of filamentous bacteriophage in *Escherichia coli*. 70
Figure 1.9 Schematic drawing of 3+3 phagemid based phage display system. 72
Figure 1.10 Phage display library panning against an immobilised target. 74
Figure 1.11 Schematic overview of the bacterial secretome-selective phage display system. 76
Figure 2.1 Overview of the rumen content fractionation procedure. 96
Figure 2.2 Overview of the metasecretome library construction and selection. 99
Figure 2.3 Workflow overview of the *in silico* analysis. 105
Figure 3.1 Intact and mechanically sheared metagenomic DNA isolated from the plant-adherent rumen microbial community. 116
Figure 3.2 Overview of results of metasecretome library construction and selection. 117
Figure 3.3 Types of membrane-targeting sequences detected in the pilot metasecretome library. 121
Figure 3.4 Functional annotation of putative proteins in the pilot metasecretome library. 122
Figure 3.5 Taxonomic distribution of rumen microbial inserts from the pilot metasecretome library. 123
Figure 3.6 PCR amplification of metasecretome-enriched ssDNA. 126
Figure 3.7 PCR amplicons of the inserts of metasecretome-enriched PPs, processed by enzymatic and mechanical shearing to obtain the pyrosequencing template. 127
Figure 3.8 Common types of membrane-targeting signals detected in putative proteins in-frame with pIII in the metasecretome-enriched dataset. 130
Figure 3.9 Phylogenetic distribution of putative protein-coding genes in the metagenome and the metasecretome dataset. 132
Figure 3.10 Relative abundances of Pfams within the metagenome and metasecretome-enriched sequence datasets. 134
Figure 3.11 Comparison of dbCAN hits belonging to different CAZyme classes in the metasecretome and metagenome datasets. 136

Figure 3.12 Architecture of putative proteins with predicted multi-modular CAZyme organisation in the metasecretome dataset. 141
Figure 3.13 Frequency of cellulosome modules in three bovine rumen plant-adherent microbial datasets. 143
Figure 3.14 Phylogenetic diversity of the cellulosome modules predicted in the rumen metasecretome dataset. 144
Figure 3.15 Distribution of sequence identity of best BLASTP hits for CAZymes detected within the metasecretome dataset. 146
Figure 4.1 Recombinant phagemid profiles of the metagenomic library over four rounds of affinity panning on carbohydrate substrates. 154
Figure 4.2 Bacterial colony PCR of 380 clones selected from the metagenomic phage display library by affinity screening against complex carbohydrate substrates. 156
Figure A2.1. Overview of the frequencies of major putative CAZy families involved in the degradation of plant cell wall polysaccharides in the metasecretome and the metagenome dataset..... 220
Figure A2.2. Example alignment of putative multi-modular CAZyme and corresponding HMMs..... 224

Abbreviations

AXYL	insoluble wheat arabinoxylan
AA	Auxiliary Activities
BLAST	Basic Local Alignment Sequence Tool
bp	base pairs
C-	carboxyl-terminal
CAZyme	Carbohydrate Active enZyme
CBM	Carbohydrate Binding Module
CE	Carbohydrate Esterase
CFU	Colony Forming Units
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
ESP	EDTA/Sarkosyl/Protease
dNTP	deoxyribonucleoside triphosphate
Gb	Gigabase
GH	Glycoside Hydrolase
GT	Glycosyl Transferase
h	hour/hours
HMM	Hidden Markov Model
K_a	affinity constant
Kb	Kilobase
M	molar
Mb	Megabase
min	minute/minutes
m.o.i.	multiplicity of infection
N-	amino-terminal
NGS	next-generation sequencing
nt	nucleotide
OD	Optical Density
ORF	Open Reading Frame
PPs	Phagemid Particles
PFU	Plaque Forming Units
PCR	Polymerase Chain Reaction
PL	Polysaccharide Lyase
PFU	Plaque Forming Units

RAC	Regenerated Amorphous Cellulose
RT	Room Temperature
ssDNA	single-stranded DNA
SLH	S-Layer Homology
UV	Ultraviolet
v/v	volume per volume
wt	wild type
w/v	weight per volume