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

List of tables

List of figures

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

Chapter 1. Literature review

1.1 Introduction

The rumen is the fermentative forestomach of ruminant animals, and is often described as the engine of the New Zealand economy. As of June 2012, products derived from ruminant animals (including dairy, meat and wool) represented over 50% of New Zealand's total commodity export and contributed over $19 billion to the economy . Accordingly, the pastoral agricultural industry is heavily reliant on ruminants, and in 2012, New Zealand had 3.7 million beef and 6.4 million dairy cattle, 31.3 million sheep and 1.1 million deer [1].

The diet of New Zealand ruminants is predominantly based on pasture and contains a large proportion of fibre [2]. Dietary fibre is mainly composed of plant cell wall polysaccharides, and cannot be degraded by the animal's own digestive enzymes, and instead, its degradation is performed by a complex microbial community residing within the rumen. Digestibility of the lignocellulosic component of fibre represents the main limiting factor for increasing ruminant productivity. Secreted and cell-surface associated fibrolytic enzymes and accessory proteins produced by rumen microbes have a central role in initiating fibre degradation and therefore are of significant interest.

Such proteins have potential applications in agriculture, biomass waste management and biofuel production, and can also contribute to a better understanding of the mechanisms underlying fibrolytic processes. They also represent potential targets for improving the enzymatic performance, lowering production costs and obtaining enzymes tailored for specific application through genetic engineering and directed evolution [3-6].

Rumen microorganisms have been used, since the early 1980s, as a source of enzymatic activities involved in the degradation of lignocellulose. The traditional approach to tap into this resource is *via* the cultivation of rumen microbes and screening for individual strains with the desired activities. Considering that only a limited fraction of the rumen microbial diversity is currently represented by microbes in culture [7-9], complex microbial communities are being explored using culture-independent molecular approaches, including metagenomics [10]. Metagenomic approaches for novel bioactive discovery include shotgun sequencing of community DNA to determine the genetic potential of the community, or targeted functional screens of libraries constructed from community DNA [11]. Structural and functional genomic studies of cultured fibre-degrading microorganisms, in combination with metagenomic studies

of different biomass-degrading ecosystems, have revealed new approaches for fibre degradation and binding [12-25].

The rumen represents one of the most complex microbial ecosystems, and its complexity, in combination with current limits of high-throughput sequencing platforms typically results in low coverage of the complete metagenomic gene pool, even in a large scale metagenomic investigations [26]. Genes encoding secreted and surface fibrolytic enzymes and accessory proteins represent a fraction of the secretome protein encoding genes in the rumen bacterial community, which comprise 10 – 30% of the total coding capacity of bacterial genomes [27, 28]. Focusing sequencing efforts onto the subset of metagenomic sequences encoding secreted and surface proteins is predicted to increase the chance of detecting rare genes that have not yet been captured by traditional metagenomic approaches and could lead to expanding the known repertoire of catalytic and accessory fibrolytic activities from a highly complex microbial community.

1.2 Rumen

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

Ruminant animals possess a foregut fermentation strategy, and comprise nearly 200 species of herbivorous mammals belonging to six different families. The majority of ruminant species are members of the families Bovidae and Cervidae, found natively across a range of climates and habitats on all continents except Antarctica and Australia. Domesticated ruminants, such as cattle, sheep, goats, water buffalos, yaks and reindeer, sustain the livelihoods of hundreds of millions of people worldwide [29, 30].

The defining characteristic of ruminants is the possession of the rumen. This term commonly refers to the first two compartments of the animal's four-chambered stomach, the reticulum and the rumen. The rumen is followed in the digestive tract by the omasum, and the true acid stomach, known as the abomasum, which is similar in structure and function to the stomach of monogastric animals [31]. Ruminant animals have evolved a unique symbiotic relationship with a complex microbial community residing in their rumen. Members of this community have key roles in initial feed digestion, and the release of energy locked in the plant cell walls which are indigestible to the ruminant host. About 55 – 65% of digestion takes place in the rumen, 20 – 30% in the small intestine and 5 – 15% in the large intestine. The microbiota are responsible for digestion in the rumen and large intestine, while host enzymes enable abomasal and small intestine digestion [2].

The rumen is a large organ that can hold around 60 – 120 kg of content (digesta) in cows and 6 – 8 kg digesta in sheep, with daily digesta flow of 120 – 200 L for cows and 10 – 30 L for sheep. Digesta in the rumen typically contains only 9 – 18% of dry matter and is buffered by copious amounts of saliva [2, 31]. The rumen acts as an anaerobic, continuous-flow fermenter for microbial fibre digestion. It provides optimal conditions for maintaining high microbial densities of predominantly obligate, and some facultative, anaerobes and facilitates microbial digestion of plant material [31, 32]. Turnover rate of digesta in the rumen is diet-dependant and is approximately 30 h. A stable temperature of 39ºC is maintained in the rumen and may rise slightly after feeding, when fermentation is maximal [33]. Acidic end-products of carbohydrate fermentation are buffered with bicarbonates from saliva to maintain an optimal rumen pH for fermentation (between 5.8 – 6.8) [34]. Physically, the rumen contents are mixed by regular contractions and ruminated (regurgitated and chewed) several times, which assists microbial fibre degradation and the reduction of plant fragments to a size where they can pass, together with the liquid phase, out of the rumen [2, 35].

The end-products of the feed digestion in the rumen are: volatile fatty acids (VFAs), such as acetate, propionate, butyrate and valerate; peptides, amino acids and nucleic acids (mainly from degraded bacteria); saturated dietary and microbial long-chain fatty acids. Gases (methane and carbon-dioxide) are also produced from rumen fermentation, but are expelled by eructation [2]. VFAs are a major energy source, and together with microbial proteins, provide 70 – 85% of the nutrients absorbed by the animal, and are therefore of key importance for nutrition and productivity [2, 36].

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

The rumen microbiota is phylogenetically diverse, and is comprised of bacteria (10^{10} – 10^{11} cells/mL of rumen contents), methanogenic archaea (10^{7} – 10^{9} cells/mL), protozoa (10^{4} – 10^{6} cells/mL), fungi (10^{3} – 10^{6} cells/mL) and bacteriophage (10^{9} – 10^{10} particles/mL). The relative concentrations of these microbes are highly dependent on diet composition [37-40]. The majority of rumen microbes are obligate anaerobes, but a significantly smaller fraction of facultative anaerobes have an important role in maintaining anaerobic conditions in the rumen by rapidly metabolizing oxygen that is introduced with ingested plant material.

Bacteria are the most numerous and diverse group among the rumen microbiota, and are responsible for the majority of the digestion in the rumen. Based on their substrate preference, bacteria belonging to diverse functional groups are present, such as fibrolytic, proteolytic, amylolytic, lipolytic, ureolytic, tanninolytic and acetogenic bacteria. Fibrolytic bacteria are the

main suppliers of VFAs and microbial proteins for the host and their role in fibre digestion will be discussed in detail in section 1.2.3.

Protozoa, large unicellular eukaryotes, digest both plant particles and bacteria, and are known to establish ecto- and endo- symbiotic associations with methanogens for the disposal of the hydrogen resulting from feed fermentation. Protozoa can account for up to two thirds of the total rumen microbial biomass [40], and are represented by entodiniomorphs belonging to the family Ophryoscolecidae (dominated by *Entodinium caudatum*) and the holotrichs from the order Vestibuliferida (dominated by *Isotricha prostoma* and/or *Dasytricha ruminantium*) [38]. It is thought that entodiniomorphs are directly involved in the degradation of plant fragments, including complex plant cell wall carbohydrates, while holotrichs do not have an active role in the fibre degradation and depend on starch granules and soluble polysaccharides [41]. Although protozoa are not considered to provide a significant nutrient supply to the host, they have a limited role in digestion of higher-quality forages with more soluble carbohydrate and plant polysaccharide degradation [42-44].

Anaerobic fungi can constitute 8 – 20% of total microbial biomass in the rumen [39, 45] and, in New Zealand ruminants, major phylogenetic groups belong to genera *Piromyces*, *Neocallimastix, Caecomyces* and *Orpinomyces* [46]. Rumen fungi have high fibrolytic potential [47], as well as the potential for the degradation of the more recalcitrant plant cell walls in forages due to the unique ability of some rumen fungi to degrade lignified cell walls [48, 49]. For this reason, fungi have an important role in the initiation of plant fibre degradation, by disrupting the fibre structure and allowing access of fibrolytic enzymes from non-fungal members of rumen microbial community [50].

Rumen methanogenic archaea are dominated by species belonging to *Methanobrevibacter*, *Methanomicrobium* and Thermoplasmatales-affiliated uncultured group, termed Rumen Cluster C [51], and constitute about 0.3 – 3.3% of the rumen microbiota [52]. Rumen methanogens can metabolise fermentation end-products into methane, preventing their accumulation within the rumen [53]. Rumen methanogens are mainly hydrogenotrophic, and they have an important role in eliminating the inhibitory effects of hydrogen on microbial fermentation by providing the main pathway for its disposal [51, 54]. A significant proportion (between 2% and 12%) of total energy derived from fermentation in the rumen can be lost through the production of enteric methane, which is emitted from the ruminant host [55, 56].

The current understanding of the role of the large and dynamic bacteriophage community within the rumen is limited. It is thought that phage have an impact on maintaining a continual balance of microbial communities, and reducing the efficiency of feed utilisation in ruminants through extensive bacterial lysis [57, 58].

1.2.3 Bacterial diversity in the rumen

The techniques used to examine rumen microbial diversity and abundance have evolved from historically used culture-based techniques to modern culture-independent molecular techniques, such as phylogenetic analyses of small-subunit ribosomal RNA (*rrs*) gene sequences, DNA fingerprinting techniques such as denaturing gradient gel electrophoresis (DGGE) and mass DNA sequencing of phylogenetically informative loci [40].

Diet, host genetics, developmental stage of the ruminant, animal management, geographic and temporal factors, as well as different environmental niches (microenvironments) within the rumen are among factors known to influence rumen microbial community structure and add to the large global diversity of the rumen microbiome. Analysis of *rrs* gene sequences amplified from different ruminant species, and animals fed different diets across broad geographic regions have helped in expanding our understanding of the phylogenetic diversity within the rumen [7]. It was estimated that in order to achieve 99.9% coverage of the global rumen microbial diversity at the species level, a minimum of ~80,000 bacterial and ~25,000 archaeal *rrs* gene sequences, from samples obtained under different conditions known to influence rumen microbial community structure, would need to be analysed.

Depending on rumen microenvironments (environmental niches) they occupy, bacteria can be classified into several groups [59, 60]. Bacteria loosely and firmly associated with feed particles comprise up to 75% of the metabolically active rumen microbial community, and are crucial for plant cell wall breakdown and feed digestion [59, 61, 62]. Bacteria associated with the rumen epithelium, as well as bacteria attached to the surface of the protozoa and fungal sporangia (~1% of total rumen population) have a minor role in feed digestion. Free-living (planktonic) bacteria associated with rumen liquid phase, such as non fibre-adherent cells and daughter cells released from the cell division process on the colonised fibre, comprise 20 – 30% of the total community microbes. Bacteria from the liquid rumen fraction have indirect roles in fibre degradation [61, 63], possibly by cross-feeding on products of plant polysaccharide decomposition by adherent rumen microbes [64].

The significant impact of diet, microenvironment and animal variation on rumen bacterial community structure has been well documented. Differences in community composition are greater between animals fed different diets [19, 42, 62, 65-68] than between rumen microbes from different microenvironments (e.g. bacteria loosely and firmly associated with feed particles, free-living bacteria associated with rumen liquid) within an animal [62, 69]; and both factors have a greater impact on the microbial diversity compared to intraspecific animal to animal variation when the same diet and microenvironment are compared [70]. Significant differences in community composition were observed when bacterial populations were compared between any two rumen microbial microenvironments within the same diet, but

in contrast, differences were not observed when the same rumen microbial fraction from different individual animals on a same diet was compared [70]. Studies by Larue *et al*. (2005) [62] and Kong *et al*. (2010) [70] showed that dietary factors affect community structure in the rumen of sheep or cow and that, within the same diet, the sub-communities tightly attached to feed have more complex and diverse profiles as compared to sub-communities that were either loosely attached to the feed, or planktonic.

It is thought that the complexity of the feed, especially its structural carbohydrates and secondary compounds, favours microbial diversity [26], and prompt responses in rumen bacterial diversity have been demonstrated for animals fed contrasting forage diets. For example, a greater diversity of rumen bacterial communities has been observed in animals fed with the structurally more complex Bermudagrass hay diet compared to wheat forage [42] or, in the case where two high-fibre diets were compared, in animals fed more nutritionally complex alfalfa hay compared to animals fed triticale straw, with a relatively restricted nutrient content [70]. However, although highly responsive to dietary changes at the species level, the rumen microbial ecosystem appears to be stable at the phylum, class and even order level and returns to the original state after being affected by external factors, such as dietary intervention [71].

The feed-adherent bacterial community is distinct, and more diverse compared to those in the loosely-attached and planktonic fractions [62, 70] and is dominated by the Firmicutes [62, 72]. Analysis of prokaryotic and fungal diversity in the rumens of cows consuming a forage diet, based on sequence analysis of *rrs* gene loci, demonstrated a significant difference between the bacterial community profiles of the liquid and solid rumen content fractions [73]. However, the diversity of archaea and fungi did not appear to differ between the two fractions. This study also showed that the solid fraction of the rumen was dominated by members of the order Clostridiales, mainly *Butyrivibrio* and *Blautia*, while *Prevotella* and *Tannerella* from the order Bacteroidales were overrepresented in the liquid fraction. Other studies have also shown that members of genus *Prevotella* are more prevalent in the liquid fraction of pasture-fed cows and Bermudagrass hay- or wheat-fed steers, while members of genus *Butyrivibrio* are more abundant in the solid fraction of pasture-fed cows and Bermudagrass hay- or wheat-fed steers [8, 42, 70].

The extent of bacterial diversity found in the rumen is currently unknown, and estimations made in different studies place the number of bacterial species-level operational taxonomic units between 400 to over 5000 [7, 8, 19, 22, 70, 73-76]. Based on molecular studies of large-scale 16S rRNA gene libraries, it is estimated that only a small proportion (perhaps as low as 10%) of the rumen microbial diversity is currently represented by microbes in culture [7-9]. The cultivated rumen microorganisms do not necessarily represent the numerically dominant or functionally significant members of the rumen microbiome, and this is a major barrier for defining the biology, functional roles and interactions in this microbial ecosystem.

The phyla distribution observed in the rumen appears to be universal for different gut environments, including hindgut fermenters and monogastric animals [18]. A study of the global diversity of the rumen microbiome, estimated through the analysis of the publicly available bacterial and archaeal *rrs* gene sequences from globally distributed domesticated and wild ruminant animals, identified 19 bacterial phyla, while almost all archaeal sequences were assigned to only one phylum, the Euryarchaeota [7]. The phyla Firmicutes (dominated by the clostridial families, Lachnospiraceae and Ruminococcaceae), Bacteroidetes (dominated by Prevotellaceae) and Proteobacteria (with representatives from all five classes) accounted for up to 90% of all bacterial assignments, and most studies, regardless of experimental conditions, diet or rumen contents fraction, identified these phyla as being dominant in the rumen [7, 8, 60, 68-70, 76, 77].

The major differences in rumen communities are seen at the genus and species levels. However, there is mounting evidence for existence of a core rumen microbiome, which is consistently present in different animals, microbial fractions and diets, but exhibits high variability in abundance across samples. The core bovine rumen microbiome is comprised of 25 – 32 genera, including *Acetivibrio*, *Barnesiella*, *Butyrivibrio*, *Fibrobacter*, *Oscillibacter*, *Paraprevotella*, *Prevotella*, *Pseudobutyrivibrio*, *Ruminococcus*, *Succiniclasticum* and *Treponema* [65, 70, 73, 75, 78, 79], as well as genera belonging to three large groups of unclassified (and mainly uncultured) Clostridiales, Lachnospiraceae and Ruminococcaceae [7]. Most studies identified *Prevotella* as the predominant genus within the rumen microbiome [7, 42, 65, 71, 73, 75, 78]. The abundance of this genus has been confirmed by real-time PCR quantification, however the majority of *Prevotella* species are still uncultured [80-82].

A limited number of fibrolytic rumen bacteria have been identified. Cultivation and genomic studies identified the Gram-negative *Fibrobacter succinogenes* [15] and Gram-positive *Ruminococcus flavefaciens* [12] and *Ruminococcus albus* [83] as cellulose degrading specialists. The Gram-negative *Prevotella* [13] is a generalist capable of degrading and utilizing many different polysaccharides, but not cellulose, while the *Butyrivibrio/Pseudobutyrivibrio* assemblage [14], which is prominent in the rumen of animals on high fibre diets [84, 85], are considered to be a primary hemicellulose-degrading group [26, 86]. Based on population quantification using real-time PCR, it is becoming clear that some of the uncultured bacteria are not only as abundant as known hemicellulolytic bacteria but also might have an important role in ruminal feed digestion [82].

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

The reduction of plant particle size by the ruminant host and microbial enzymatic hydrolysis of plant cell wall polysaccharides are the crucial steps in ruminant digestion [87]. Chewing during eating and rumination results in the significant reduction of feed particle size, thus increasing substrate hydration and exposed surface area (by $10^4 – 10^6$ fold), and physically disrupts the protective cutin layer of forage, exposing the more digestible interior layers of the plant cell wall [88, 89]. Rumen bacteria generally start digestion from the luminal side (interior) of disrupted plant cells, only after lignified cells in the feed have been physically ruptured [90]. It is well established that physical or chemical pre-treatment of the feed increases the digestibility of fibre in the rumen compared to untreated substrates [86]. Pre-treatment creates more adhesion sites and production of well-defined pits in the colonised tissue increases subsequent degradation. The cell walls are degraded from the inside out, starting from the lumen of broken cell, towards the outer layer of the secondary cell wall and lignified middle lamella [91]. In contrast, non-lignified cell walls can be degraded by rumen bacteria both from the exterior and interior, and bacteria can digest adjoining cells *via* digestion of intervening walls [92].

Feed particles are fermented in the rumen until they are small enough, typically millimetre-size (depending on host species), to pass through the reticulo-omasal orifice to the omasum. A relatively long retention time of feed particles in the rumen is enabled by the large rumen size and omasal laminae that trap large feed particles and flush them back to the rumen. This allows prolonged enzymatic access and hydrolysis of the cell wall polysaccharides by microbial fibrolytic enzymes [93]. Approximately one third of the cells within the typical forage particle size leave the rumen undigested because of their inaccessibility to the rumen microbes [94]. The extent of feed degradation depends on the digestion rate and rate of passage of solids from the rumen. Prolonged retention of digesta in the rumen (critical for maximizing the plant cell wall cellulose and hemicellulose digestion) and faster rumen emptying allowing greater feed intake (which results in a net increase of digested nutrients) are two opposing forces affecting the ruminant's digestion and productivity [95].

Fibrolytic bacterial species demonstrate different specificities for binding and colonisation of fibre in the rumen. This species-specific mode of adhesion to plant cell walls reduces competition between fibrolytic bacteria [59]. The strategy of strong and specific adhesion of rumen microbes to the plant material increases digestion of the plant cell wall polysaccharides and has several advantages: enzymes are concentrated on the substrate; other microbes are excluded from the site of hydrolysis; bacteria and their enzymes are protected from

ruminal proteases, and bacteria attached to feed particles have up to three times longer retention time in the rumen as compared to planktonic bacteria [96].

Ingested forage is primarily colonised by planktonic cells or through direct physical contact of the incoming forage and feed that is already colonised. Colonisation is a relatively rapid process due to the high densities of bacteria and plant material in the rumen [59, 91]. Initial non-specific adhesion of bacteria to unprotected sites of the substrate, through van der Waals forces, and conformation to the substrate shape and wedging into feed cavities is enabled by components of the bacterial glycocalyx, and happens shortly after contact with the substrate [96, 97]. Specific adhesion is mediated by interaction of ligands or adhesins on bacterial cell surfaces (such as fimbriae or pili, glycosylated epitopes of the bacterial glycocalyx and carbohydrate binding modules of fibrolytic enzymes and cellulosome complexes) with the fibrous substrate. During initial cell wall polysaccharide digestion, ruminal bacteria specifically attached to feed fibre receive signals stimulating the expression of inducible ligands, adhesins and fibrolytic enzymes. For example, a model describing the formation of a *Fibrobacter succinogenes* biofilm on a cellulose surface proposes that surface proteins (fibro-slime domain-containing proteins and pilins forming type IV pili) might play a role in attachment of the bacterial cell to the substrate, thus mediating close contact of cellulases and hemicellulases associated with the cell membrane with plant cell wall polysaccharides and initiating their deconstruction [98, 99]. Out of 25 identified proteins from the outer membrane fraction of cellulose-grown *F. succinogenes*, 16 were up-regulated by growth on cellulose compared to when grown on glucose [98]. Both the non-specific and specific phases of bacterial adhesion depend on factors related to: bacteria (such as age, glycocalyx condition and microbial competition), substrate (factors such as the presence of the protective plant tissue cuticle, surface area, hydration status and ionic charge) and environment (factors such as temperature, pH, presence of cations and soluble carbohydrates) [59]. In the last phase of adhesion, adhered bacteria proliferate to create colonies on potentially digestible sites of forage particles.

Several studies of the dynamics of the early stages of colonisation of fresh forage by fibrolytic bacteria indicate that a diverse community of ruminal bacteria colonises forage rapidly, with initial attachment accomplished within minutes of feed ingestion, and with the adherent community size stabilizing after 15 min [60, 72]. Colonisation is immediately followed by production of exopolymeric substances involved in biofilm formation in the rumen and it was shown that it reaches a maximum after 1 h and 4 h for perennial ryegrass leaf and stem as substrates, respectively [100, 101]. Huws and colleagues (2013) [101] demonstrated a biphasic bacterial colonisation of perennial ryegrass *in sacco* in the cow rumen, with primary colonisers starting to detach after approximately 2 h, and their partial replacement by secondary colonisers as early as 2 – 4 h after introduction of the ryegrass in the rumen. A transition in bacterial 16S rRNA gene diversity was detected by DGGE, suggesting that distinct bacterial communities are

involved in these two stages. *Prevotella* contributed to more than 50% of the total detected bacterial 16S rRNA at all time points examined (0 – 24 h). However, the majority of *rrs* gene sequences belonged to uncultured and unknown bacteria, and could not be classified beyond the family level, so the study did not provide insights into the functional roles of each distinct population.

It was initially hypothesised that the spatial organisation and attachment of secondary colonisers is influenced solely by soluble nutrients or microbial end products released by primary colonisers maintaining their population [63]. A gene-centric metagenomic analysis, used to assess the prevalence of enzymes involved in polysaccharide degradation, led to a model of fibre degradation in which bacteria able to remove the easily available side chains of complex plant polysaccharides are initial fibre colonisers, which are subsequently replaced by cellulose and xylan degraders [19].

1.3 Carbohydrate-active enzymes involved in fibre degradation

1.3.1 Unique aspects of grass cell wall architecture

Plant dry matter is, in general, primarily composed of fibre (55 – 70%) and protein (15 – 22%), with soluble sugars, nucleic acids, lipids, minerals and secondary compounds making up the remainder [2]. In the context of ruminant nutrition, the term 'fibre' is often used to refer to the plant cell walls [102]. Fibre is a feed component with the slowest and most variable digestion in the ruminant's diet and its quantity, structure and chemical composition significantly differs between plant species, and within a plant both over time, and in different tissues. As an estimate of the cell wall concentration in feed, neutral detergent fibre (NDF) content is determined using a detergent fibre analysis system established by Van Soest [103]. The daily NDF intake capacity of high-producing dairy cows on a forage based diet is around 1.2% of their body weight [104]. Fibre content and its digestibility are the major limiting factors for the intake potential and energy availability from feed in ruminant animals, and have great impact on ruminant productivity [105].

The plant cell wall is a complex structure, containing up to three layers (middle lamella, primary and secondary cell wall), providing structural support and protection to the plant cell. The primary cell wall is a flexible and extensible layer of the growing cell, within which a thick secondary cell wall is deposited in mature plants. The outermost layer, the middle lamella, lies at the interface of adjacent plant cells. Plant cell walls are mainly composed of complex carbohydrates, such as cellulose and matrix polysaccharides (hemicelluloses and pectins), a phenolic component lignin, small amounts of structural proteins and some minor secondary

compounds. Hemicellulose, together with lignin, forms a protective sheet around the cellulose [106]. It cross-links cellulose *via* hydrogen bonds into a robust network inside the plant cell wall, which covalently binds to lignin and is embedded in a pectin matrix [107]. The generalised structure of the plant cell wall is represented in Figure 1.1.

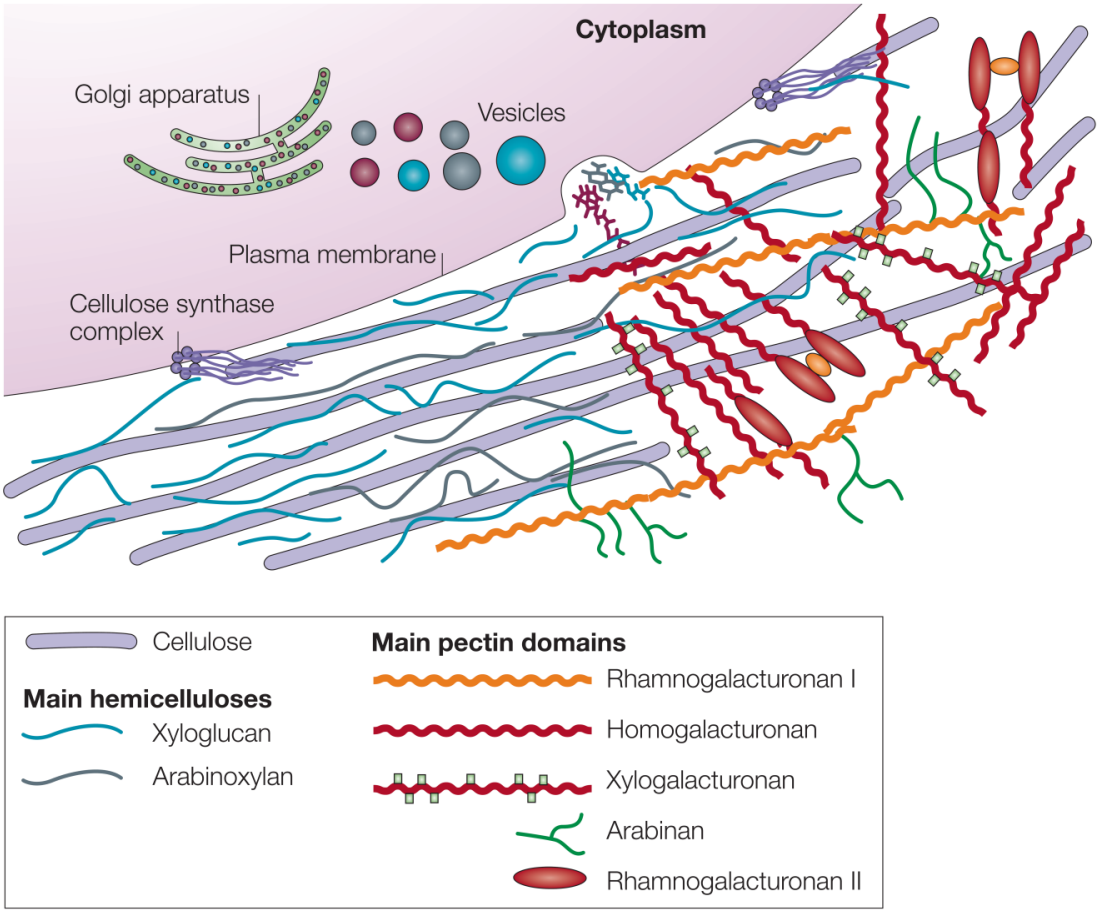

Figure 1.1 Generalised structure of the primary plant cell wall.

The primary plant cell wall consists of cellulose microfibrils (purple rods), hemicelluloses and pectins. The hemicellulose–cellulose network is shown on the left part of the cell wall scheme without pectins for clarity. The symbols used to represent the main plant cell wall structural polysaccharides are represented in the box below the diagram. The predominant hemicellulose in most plant species is xyloglucan, with smaller amounts of hemicelluloses such as arabinoxylans and mannans (not shown) also present. The main pectin polysaccharides include rhamnogalacturonan I and homogalacturonan, with smaller amounts of xylogalacturonan, arabinan, arabinogalactan I (not shown), and rhamnogalacturonan II. Pectin domains are covalently linked together (orange circles) and bind to xyloglucan by covalent and non-covalent bonds. In addition, neutral pectin polysaccharides (green) can also bind cellulose. Figure taken from [108] with permission.

Cellulose, the most abundant form of fixed carbon on Earth, is comprised of cellulose microfibrils, formed through bundling and crystallisation of dozens of linear chains of

β-(1,4)-linked D-glucan subunits. The cellulose microfibrils are tightly and parallelly packed into insoluble crystalline regions *via* hydrogen bonding and hydrophobic stacking. Approximately every 100 – 150 nm along the microfibrils there are weak points, where less compact, staggered microfibrils form soluble amorphous regions [109]. The relative proportion of crystalline to amorphous cellulose varies during the cell wall formation and maturation and between cell types. In naturally occurring cellulose, 40 – 90% has a crystalline structure, with the rest being amorphous [110]. In addition to crystalline and amorphous regions, variable nanodomains comprised of different amorphous structures and several crystalline allomorphs, are also present in the cellulose structure [111].

Hemicelluloses and pectins are extremely heterogeneous groups of complex matrix polysaccharides, comprised of different 5- and 6-carbon monosaccharide units. These polymers can bear various chemical modifications (usually acetylations and methylations), adding to the complexity of fibre. Hemicelluloses account for approximately one third of all renewable organic carbon on Earth [112] and include complex branched heteropolymers (such as xyloglucans, arabinoxylans, glucuronoarabinoxylans and mannans), with variable structure, degree and pattern of substitution that can bind tightly to the cellulose surface [108]. Basic structural components of hemicellulose are represented in Figure 1.2. Arabinoxylan consists of a (1,4)-linked β-D-xylan backbone, substituted with O-2, O-3 or O-2,3 linked α-L-arabinose residues and with attached ferulic acid esters, acetyl groups or glucuronic acid residues (in glucuronoarabinoxylans, GAX) [113]. Their twisted chains and great number of substituents prevent the tight packing of these molecules into a crystalline array. In contrast, ribbon-like mannan molecules comprised of linear arrays of β-1,4 linked mannose units pack together, but less tightly than cellulose molecules [114]. The backbone structure of xyloglucans is similar to cellulose, but on average, three quarters of backbone glucose residues are decorated with xylose and these side chains can also carry serially appended galactose and fucose residues.

The predominant pectins in plant cell walls are homogalacturonan and rhamnogalacturonans I and II, but xylogalacturonan, galactans, arabinans, as well as other pectic polysaccharides can be present in the matrix. All pectic polysaccharides are rich in α-D-galacturonic acid and can have branches (e.g. xylose in xylogalacturonan); additional backbone sugar residues (e.g. rhamnose in rhamnogalacturonan I or at least 12 different glycosyl residues linked together by more than 20 different glycosidic linkages in rhamnogalacturonan II); modifications (e.g. methyl esterified carboxyl groups in homogalacturonan and xylogalacturonan) and can exist with multimeric organisation (e.g. rhamnogalacturonans II are present in plant cell walls as borate cross-linked dimers) [115].

Lignins, a large group of highly branched phenylpropanoid polymers, are mainly built from monolignols (coniferyl alcohol, sinapyl alcohol and minor amounts of p-coumaryl alcohol), methoxylated to various degrees in the form of guaiacyl, syringyl and

p-hydroxyphenyl units, respectively. In addition, phenolic acid and other hydroxycinnamic acids, such as ferulic acids can be also present in cell wall lignin [116].

Ruminant animal production systems throughout the world rely heavily on forages and crop residues as the major feed resource. The diet of New Zealand ruminants consists mainly of ryegrass (*Lolium perenne*) with 10 – 20% of the forage legume white clover (*Trifolium repens*), supplemented with silage and specialist crops to fill feed gaps during winter and summer [2, 117]. Grass type II cell walls are a major source of dietary fibre for ruminants and they differ significantly from type I cell walls, found in dicots and non-commelinoid monocots [118], in relative abundance and cross-linking of hemicelluloses, as well as in the amounts of pectins, phenolic compounds and proteins [119]. In grass primary cell walls, cellulose fibres are encased in a network of GAX. High levels of mixed linkage glucans and hydroxycinnamates are present, with very little pectin and structural proteins. Adjacent GAX molecules are cross-linked by hydroxycinnamates, contributing to the indigestibility of this cell wall fraction in grasses [120]. Secondary cell walls, deposited inside the primary cell walls comprise at least 50% of the cell mass in both stems and leaves. They are largely composed of cellulose and GAX with fewer side chains than in the primary cell walls, which results in stronger hemicellulose-cellulose interactions. Grass secondary cell walls also have substantial amounts of lignin, which fills the pores between polysaccharides and contains much higher amounts of hydroxycinnamates than dicots [119]. On the other hand, legume cell walls are pectin rich [92] and hydroxycinnamic acids are virtually absent [102].

1.3.2 Strategies for cell wall digestion in the rumen

Bacterial, fungal and protozoal members of complex microbial communities in the rumen and from other fibre-degrading ecosystems (e.g. soils, swamps, seawater sediments, thermal and volcanic springs, sewage sludge, compost heaps) use universal biomass-degrading strategies. Their cellulolytic and hemicellulolytic enzymes are built from combinations of the same basic modules belonging to different families of a large and diverse group of carbohydrate-active enzymes (CAZymes). These enzymes contain at least one function-defining catalytic domain and one or more accessory modules (e.g. carbohydrate binding modules, scaffoldins), assisting or modifying the primary hydrolytic action of the enzyme [64]. The efficient breakdown of substrates as chemically complex as plant cell walls require the coordinated activity of a plethora of enzymes that can attack the array of linkages that are found within and between the plant cell wall polysaccharides and their closely associated polymers, pectin and lignin.

Fibre digestion in the rumen is not optimal due to the limited efficiency of lignocellulosic degradation (supported by the finding that fibre recovered from faeces is still fermentable) and energy losses *via* methanogenesis. It is known that feed efficiency in ruminants can be improved through methane mitigation strategies and through strategies for enhancing fibre degradability, such as mechanical and chemical pre-treatment of the feed and reduction of plant lignin content through genetic manipulations of plant cell walls [86, 121].

The rumen microbiota has the metabolic potential to hydrolyse nearly the entire spectrum of plant polysaccharides [19], but the extent of the actual digestibility of cell wall polysaccharides in the rumen is limited by their rate of digestion. Pectins are generally more rapidly digested compared to cellulose and hemicellulose, both in the case of microbial digestion of polysaccharides isolated from the cell wall matrix *in vitro*, and embedded in intact cell walls *in vivo* [91, 122]. The cell-wall polysaccharide degradability of both grasses and legumes is negatively affected by lignin concentration, hydrophobicity and its cross-linking with hemicellulose, as well as by shifts in polysaccharide composition during plant maturation [105, 123-125]. Lignin composition and structure, on the other hand, are less likely to affect digestibility [105]. The presence of pectin in forage cell walls can dramatically restrict the access of xylanases and cellulases to their substrates, resulting in inefficient lignocellulose degradation [126].

In contrast to hydrolytic enzyme systems involved in carbohydrate degradation, that can act both aerobically and anaerobically, lignin depolymerisation requires oxygen. Lignolysis (also known as 'enzymatic combustion') involves several non-specific oxidative ligninolytic enzyme systems and their presence is limited to filamentous prokaryotes and some fungi [127]. For this reason, the degradation of lignified cell walls is inefficient in the rumen, and represents a major rate-limiting step for the anaerobic degradation of lignocellulose [64].

In order to convert insoluble complex carbohydrate components of the cell wall into soluble products that can be transported into the cell, fibre-degrading microorganisms use extracellular enzymes that are either secreted, or associated with the cell surface [114]. These microorganisms have evolved multiple enzymatic strategies, with enzymes and auxiliary proteins involved in plant cell walls hydrolysis occurring in any of the following states: as single or multi-functional enzymes, free or directly attached to the bacterial surface, and/or as a part of intricate, cell-surface anchored multi-enzyme complexes, known as cellulosomes [5]. Many biomass-degrading organisms secrete synergistic cocktails of individual enzymes with one or several catalytic domains per enzyme, whereas only a limited number of anaerobic bacteria and fungi are known to synthesise cellulosomes, containing multiple catalytic units per complex. It has been hypothesised that energetic constraints of anaerobic environments have exerted greater selective pressure for the evolution of the highly efficient cellulosome machinery in anaerobic bacteria and fungi, while free enzyme systems are the hallmark of

aerobic bacteria and fungi [128]. Based on information from a limited number of sequenced cellulose-degrading bacteria, it appears that clostridial species use all four strategies, and in these organisms cellulosomal organisation is predominantly found. However, all of the currently sequenced representatives belonging to the classes Bacilli, Actinobacteria, Gammaproteobacteria and Chloroflexi use a free enzyme strategy [64].

It has been proposed that multifunctional recruitment of enzymes/catalytic modules in close spatial proximity enables the optimisation of their synergistic interaction, as well as substrate targeting through scaffold-borne CBMs [128]. Mutants of *Clostridium thermocellum* which lack the scaffoldin CipA from their cellulosomes due to a transposon insertion into the *cipA* gene, show a 15-fold reduced activity against crystalline cellulose compared to a wild-type (wt). In contrast, the activity of the same *cipA* mutant against the soluble beta-glucans remains intact, suggesting that recruitment of multiple cellulases into the cellulosome complex is crucial for hydrolysis of crystalline substrates [129]. A highly efficient cellulose-degrading anaerobic bacterium *Caldicellulosiruptor bescii* (formerly classified as *Anaerocellum thermophilum* [130]), produces free multifunctional enzymes, and it is thought that synergistic interactions between catalytic domains are enhanced through their physical linkage [131].

1.3.3 Carbohydrate Active Enzymes involved in fibre degradation

Carbohydrate Active Enzymes (CAZymes) are an extremely variable group of proteins involved in the synthesis, degradation and modification of glycoconjugates, oligosaccharides and polysaccharides. CAZymes have evolved divergently from a limited number of progenitors by acquiring novel specificities at the substrate and product level. In general, it has been estimated that 1 – 3% of genes within the publicly available genomes of prokaryotes and eukaryotes encode CAZymes [132]. Moreover, organisms specialised in cell wall degradation are known to dedicate a higher proportion of their genome's coding capacity to encode larger repertoire of CAZymes than average. For example, the genome of the human gut bacterium *Bacteroides thetaiotaomicron* devotes 6.6% of its coding capacity to CAZymes, while *Butyrivibrio proteoclasticus* B316$^{(T)}$ (formerly classified as *Clostridium proteoclasticum* [133]), a highly xylanolytic rumen bacterium, present in high numbers in the rumen of animals consuming mainly pasture or grass silage diet, dedicates over 6% of its coding capacity to CAZymes [14, 134]. *In silico* analyses indicate that one third of polysaccharide degrading enzymes from strain B316 are secreted and it is proposed that they are involved in the initial degradation of insoluble plant polysaccharides (such as xylan, starch and pectin), while the complete degradation of substituted and non-substituted xylooligomers is catalysed by various intracellular enzymes [14].

The CAZy family classification system was introduced in 1991 by Henrissat and colleagues [135-138] in an effort to classify a limited number of glycoside hydrolases with cellulolytic activity, and it was subsequently extended to all CAZymes. Unlike the historically used Enzyme Commission (EC) and IUBMB (*International Union of Biochemistry and Molecular Biology*) nomenclature that are based on substrate specificity, the CAZy family classification approach is based on amino acid sequence similarity and it integrates structure and mechanistic features of enzymes, as well as evolutionary relationships between enzymes [138]. This system often groups CAZymes with different substrate specificity in 'poly-specific' families, based on conservation of three-dimensional structure, catalytic amino acid residues and stereo chemical mechanism [139]. CAZy families were originally created following a hydrophobic cluster analysis from a limited number of available sequences [135] and were later complemented by gapped BLAST- and HMM (Hidden Markov Model)- based sequence similarity approaches.

CAZy families are based on information available from biochemically and structurally characterised proteins; for some of these, enzymatic activity has been experimentally assessed on a range of soluble and insoluble carbohydrate substrates. All sequences corresponding to the catalytic and binding modules of carbohydrate-active enzymes are grouped in the 'high-quality' BLAST library and families are subsequently populated with various genomic and metagenomic sequences deposited in public databases and their function is inferred based on their significant sequence similarity with family members [138].

The most comprehensive specialised, knowledge-based resource, containing continuously curated and updated genomic, structural and biochemical information on CAZymes is the CAZy database [138, 140]. As of September 2013, the database lists over 340 protein families grouped into 6 classes: glycoside hydrolases (GHs, 132 families), glycosyltransferases (GTs, 94 families), polysaccharide lyases (PLs, 22 families), carbohydrate esterases (CEs, 16 families), auxiliary activities (AAs, 10 families) and associated carbohydrate-binding modules (CBMs, 67) [141]. The enzymatic repertoire of the CAZy database has recently been expanded with an 'auxiliary activities' class, with members acting on lignin, in order to systematically describe range of enzyme mechanisms necessary for lignin depolymerisation [127].

In terms of plant cell wall degradation, the most important catalytic CAZy modules belong to GH, CE and PL classes, but can also involve AAs. Appended non-catalytic CBMs are necessary for targeting of CAZymes to their carbohydrate substrates [64]. GHs catalyse the hydrolysis and transglycosylation of glycosidic bonds in carbohydrates and glycoconjugates, resulting either in net retention, or inversion of the anomeric stereochemistry of the substrate. The three-dimensional structures (protein folds) of GHs are better conserved than their amino acid sequences, resulting in further grouping of some families into clans (GH-A to GH-N)

[138]. CEs remove ester-based modifications of carbohydrates, such as acetyl, feruloyl and cinnamoyl groups in xylan, thus facilitating the action of GHs on complex polysaccharides [142]. PLs use a β-elimination mechanism to catalyse the cleavage of uronic acid-containing polysaccharides, such as glycosaminoglycans and pectin, to generate terminal hexenuronic acid residue and a new reducing end. The catalytic mechanism of these enzymes is complementary to the hydrolytic mechanism employed by GHs for polysaccharide breakdown [143-145]. AAs include lytic polysaccharide monooxygenase, such as the recently re-classified GH61 (now AA9) and CBM33 (now AA10) and other redox enzymes involved in lignin degradation [127, 146].

Most cellulases and the majority of the hemicellulases belongs to GH class of enzymes, while some of the hemicellulolytic and pectinolytic enzymes belong to CE and PL classes, respectively (Table 1.1).

Table 1.1 Major CAZyme families involved in the degradation of plant cell wall polysaccharides.

Main fibre-degrading enzymes		**Function**	**CAZy family**
Cellulases	Endo- and exo-glucanases	cleave internal bonds randomly or cleave cellobiose residues from either end of cellulose chain	GH5, GH6, GH7 (exclusively found in fungi), GH8, GH9, GH12, GH44, GH45, GH48 [a], GH74, GH124
	β-glucosidases	cleave glucose from non-reducing end of cellobioses and cellooligodextrins	GH1, GH3, GH116
Hemicellulases (hemicellulose backbone degrading enzymes)	Xylanases	cleave β-1,4 bond of xylan backbone releasing xylooligomers	GH10 and GH11 (exclusive xylanases); GH5 and GH8 (only some members of the family)
	Mannanases	cleave β-1,4 bond of mannan releasing mannan oligomers	GH5, GH26, GH113
	Lichenases	cleave β-1,4-glycosidic bonds in 3-*O*-substituted glucose of 1,3-1,4- β-glucans	GH8, GH16, GH17
Hemicellulases (hemicellulose side-chain	β- xylosidases	cleaves exo β-1,4 bond of xylooligomers releasing xylose	GH3, GH30, GH39, GH43, GH52, GH54, GH116, GH120

Main fibre-degrading enzymes		Function	CAZy family
degrading enzymes)	α-glucuronidases	cleave α-1, 2 bond between glucuronic acid side chain substitutions releasing glucuronic acid	GH67, GH115
	α-L-arabinofuranosidases	cleave arabinans at *O*-2 and *O*-3 positions on xylan back bone	GH43, GH51, GH54, GH62, GH127
	1,4-β-mannosidases	cleaves exo β-1,4 bond of mannan oligomers releasing mannose	GH1, GH2 and GH5
	Acetyl xylan esterases	cleaves acetyl side chain substitutions releasing acetic acid	CE1, CE2, CE3, CE4, CE6, CE7, CE12
	Cinnamoyl and feruloyl esterases	cleaves hydroxycinnamic acids side chain substitutions releasing *p*-coumaric and ferulic acids	CE1
Pectin-degrading enzymes	various pectate-degrading enzymes	cleave main chain and branches in pectic polysaccharides	PL1, PL3, PL9, PL10 (pectate lyases); CE8 (pectin methyl esterase); CE12 (pectin acetyl esterase), GH53 (endo-1,4-β-galactanase); PL11 (rhamnogalacturonan lyase), CE12 (rhamnogalacturonan acetyl esterase)
Degrading ester linkage between hemicellulose and lignin	4-*O*-methyl-glucuronoyl methylesterase	cleave ester linkage between 4-*O*-methyl-D-glucuronic acid and alcohol group	CE15

[a] The founding member of the GH48 CAZy family is a predominant component of the *Clostridium thermocellum* cellulosome and members of this family have been consistently found in each new discovered cellulosome, as well as part of free and multifunctional cellulases. Based on information from [64, 147].

1.3.3.1 Catalytic modules of cellulases

Cellulose, despite its structural simplicity is an extremely tough substrate to degrade. It is insoluble, has crystalline regions, and in plant cell walls, is entangled with hemicelluloses and lignin. Cellulases are secreted or surface associated *O*-glycosyl-hydrolases that hydrolyse cellulosic β-1,4-glucosidic bonds. This group of enzymes is very heterogeneous and has been traditionally divided into three functional types: endocellulases, exocellulases and processive endocellulase.

Endocellulases (endoglucanases, EC 3.2.1.4) have catalytic domains with an 'open' active site in the form of a groove or a cleft, enabling them to randomly bind to several sugar units in the interior of long cellulose molecules [148]. Besides endocellulases, this 'open' active site structure is common among other endo-acting polysaccharidases, such as xylanases, α-amylases, lysozymes, chitinases, β-1,3- and β-1,3-1,4- glucanases [132].

Exocellulases (cellobiohydrolases, EC 3.2.1.91) have tunnel-like active sites, enabling them to accommodate the substrate chain only at its terminus and sequentially cleave cellobiose residues from either the non-reducing or the reducing end of the cellulose chain [149].

Processive endocellulases have, like all endocellulases, an open active site cleft, but they contain a CBM lacking conserved aromatic amino acid residues, causing weak binding to the substrate. The CBM, necessary for the processive activity of the enzyme, is rigidly attached to the C-terminus of the catalytic domain and the cellulose polymer can bind simultaneously to both domains [150]. Processive enzymes are thought to be a key component affecting the efficiency of hydrolysis [64].

In addition, the activity of β-glucosidase (EC 3.2.1.21), which converts cellobiose and cellooligodextrins to glucose monomers *via* the hydrolysis of terminal non-reducing glucose residues, is necessary for complete degradation of cellulosic substrate. The major GH families with cellulolytic activities are listed in Table 1.1.

1.3.3.2 Catalytic modules of hemicellulases

Hemicelluloses, found at the lignin-cellulose interface in the plant cell walls, are variable in structure and the majority of the hemicellulosic polymers are either insoluble or are closely associated with the insoluble cellulose matrix and lignin. For this reason, they require the coordinated action of many secreted or surface associated hemicellulases for complete degradation [106, 107, 151, 152]. GHs, such as enzymes that can cleave the hemicellulose backbones (e.g. xylanases (EC 3.2.1.8) and β-mannanases (EC 3.2.1.78), and those that can degrade side chains or short end products (e.g. β-mannosidases (EC 3.2.1.25), α-L-

arabinofuranosidases (EC 3.2.1.55), α-L-arabinanases (EC 3.2.1.99), α-D-glucuronidases (EC 3.2.1.139) and β-xylosidases (EC 3.2.1.37), act synergistically. In addition, hemicellulolytic esterases, including acetyl xylan esterases (EC 3.1.1.72) and feruloyl esterases (EC 3.1.1.73) are needed for complete degradation of hemicelluloses [107]. The main enzymes involved in hemicellulose degradation are represented in Figure 1.2.

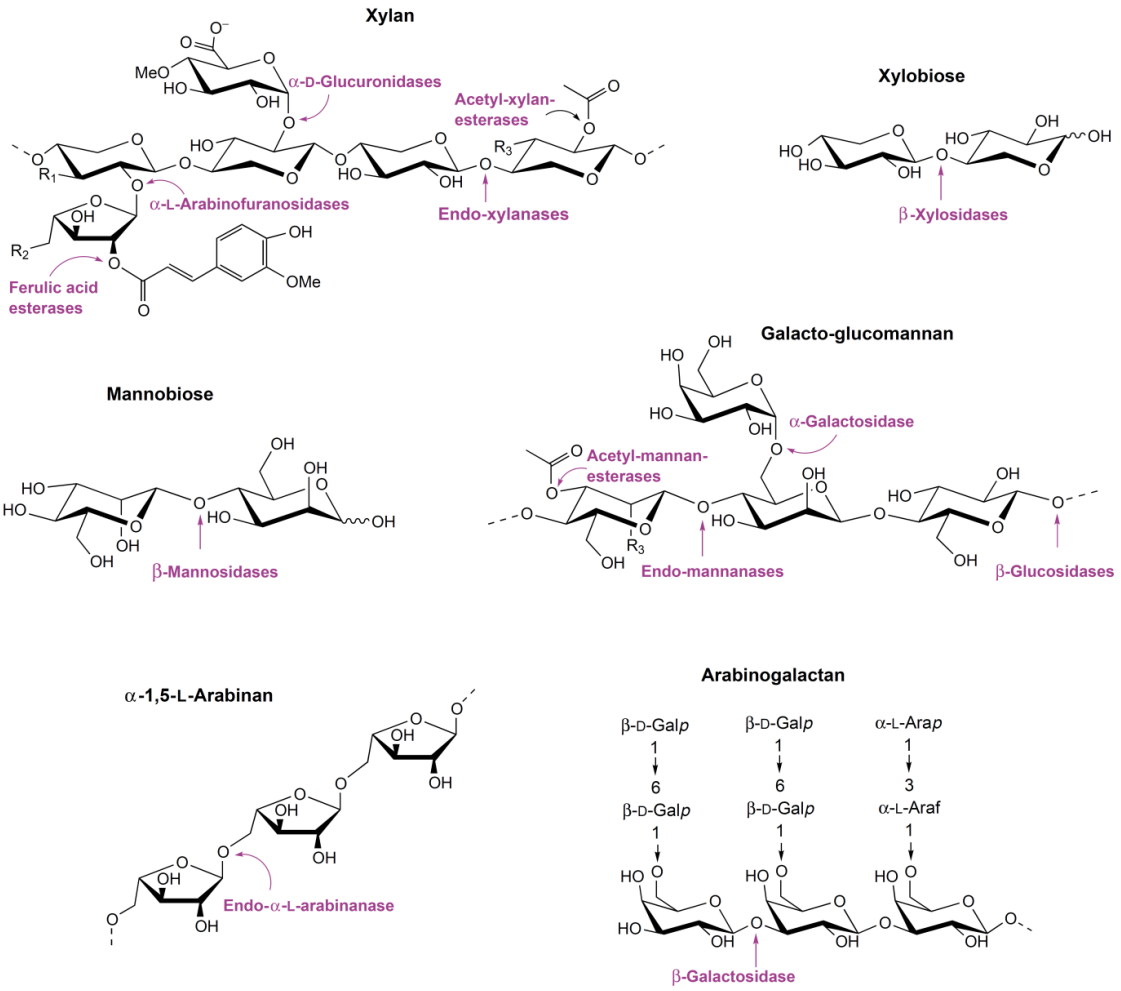

Figure 1.2 The basic structural components of hemicellulose and the hemicellulases responsible for their degradation.

Hemicellulose xylan, composed of D-xylopyranosyl units linked by β-1,4-glycosidic bonds, is hydrolysed by endo-xylanases, while the resulting disaccharides xylobioses are hydrolysed by β-xylosidases. The xylan backbone is modified with various side chains; the abundance and linkage types of these substitutions vary between xylans from different sources. In hardwood, modifications include 4-*O*-methyl-D-glucuronic acid linked to the xylose units *via* α-1,2-glycosidic bonds and acetic acid that esterifies the xylose units at the *O*-2 or *O*-3 positions and these modifications are removed by α-D-glucuronidases and acetyl-xylan esterases, respectively. Non-acetylated softwood xylans, in addition to uronic acids, have L-arabinofuranose residues attached to the main chain by α-1,2 and/or α-1,3-glycosidic linkages and these modifications are removed with α-L-arabinofuranosidases. In grasses, arabinofuranoses are esterified with *p*-coumaric and/or ferulic acid and these phenolic substitutions are removed by hydroxycinnamoyl

esterases. β-mannan based polymers are also major components of hemicelluloses. Their backbone is made of β-1,4-linked mannose residues alone or of randomly distributed mannose and glucose residues and hemicellulases such as β-mannosidases, endo-mannanases and β-glucosidases are involved in their breakdown. Galactoglucomannans also contain α-1,6-linked galactose side chains, and the *O*-2 and *O*-3 of the mannose units can be substituted with acetate groups and these modifications are cleaved off by α-galactosidases and acetyl-mannan esterases. Arabinan and arabinogalactans are generally classified as hemicelluloses, although they originate from the galacturonan or from cell wall glycoproteins. The arabinan backbone, comprised of α-1,5-linked L-arabinofuranosyl units, is hydrolysed by endo-α-L-arabinanases and can be further decorated with α-1,2- and α-1,3-linked L-arabinofuranosides. Arabinogalactans are composed of β-1,3-linked galactose residues and their backbone is hydrolysed by β-galactosidases. The backbone is substituted with β-1,6-linked galactose units and α-1,3-linked L-arabinofuranosyl or arabinans side chains. Abbreviations: Me, methyl; R1, OH or arabinofuranosyl group; R2, OH or ferulic acid; R3, OH or acetyl group. Figure taken from [107] with permission.

Hemicellulolytic systems are inducible and under carbon catabolite regulation [152]. Many microorganisms produce a range of hemicellulolytic enzymes, with diverse structures and specific activities, as well as isoenzymes, thus increasing the efficiency and extent of hemicellulose degradation. The major GH families with hemicellulolytic activities are listed in Table 1.1.

1.3.3.3 Catalytic modules of other fibre-degrading enzymes

Catalytic modules of CE and PL families complement the activities of cellulases and hemicellulases. Among 22 currently known PL families, only five are relevant for the degradation of pectins in the plant cell wall (Table 1.1). In contrast, the majority of the members of CE families are involved in the degradation of cell wall polysaccharides, mainly in cleaving non-sugar substituents in side chains (e.g. acetic, ferulic or *p*-coumaric acids). Most members of CE families involved in fibre-degradation act as acetyl xylan esterases (Table 1.1). Only members of the CE1 family possess the ability to cleave feruloyl and cinnamoyl bonds, which is crucial for separation of xylan from lignin and its more effective degradation [64].

1.3.3.4 Carbohydrate-binding modules of fibre-degrading enzymes

Carbohydrate binding activity is not exclusively found in CAZymes. For this reason, CBMs are defined as contiguous amino acid sequences with a discrete fold (modules) within carbohydrate-active enzymes (with a few exceptions of CBMs in cellulosomal scaffoldins and rare independent putative CBMs). This requirement sets this class of carbohydrate-binding

proteins apart from other non-catalytic sugar-binding proteins, such as lectins and sugar transport proteins [153].

The main sites of polysaccharide-CBM interactions are aromatic tryptophan or tyrosine amino acid residues exposed on the CBM surface, which form hydrophobic stacking interactions with non-polar faces of sugar rings, resulting in strong and stabilizing van der Waals interactions. It seems that the orientation of aromatic amino acid residues determines ligand specificities of the CBM families. Stacking interactions can be aided by hydrogen bonds between side chains of polar amino acid residues and sugar ligand [154, 155].

Amino acid similarity-based CAZy classification groups CBMs into families with a notable variation in substrate specificity and consequently with miscellaneous biological functions. For example, members of the CBM6 family can bind a wide range of substrates such as β-1,3-glucans, β-1,4-glucans and β-1,3-1,4-mixed linkage glucans, amorphous and insoluble cellulose and xylan [156], while the CBM2 family encompasses two subfamilies showing specificity for only cellulose (CBM2a) or only xylan (CBM2b). It is proposed that different ligand specificities within this family are based on an Arg/Gly polymorphism, which results in different orientations of one of the surface tryptophans involved in protein-carbohydrate interactions [154].

Another classification, based on the topology of their ligand binding sites and related to their ligand specificity, groups CBMs into three types (types A – C) [157]. Type A ('surface-binding') CBMs have a planar architecture of the binding sites, containing mainly aromatic amino acid residues and are complementary to the flat surfaces presented by insoluble cellulose or chitin crystals to which they bind. Type A CBMs show little or no affinity for soluble carbohydrates. Type B ('glycan-chain-binding') CBMs have evolved binding site topographies which can interact with the individual glycan chains, rather than with crystalline surfaces. Their extended carbohydrate binding sites are in form of grooves and clefts, have variable depth and are able to accommodate a few individual sugar units of the polymeric ligand. Aromatic amino acid residues also have a key role in ligand binding. The orientation of the aromatic amino acid residues, together with hydrogen bonds, define the affinity and ligand specificity of type B CBMs. Type C ('small sugar-binding') CBMs optimally bind to mono-, di- or tri-saccharides due to steric restrictions in their binding site. The network of hydrogen bonds between type C CBM-containing proteins and their ligands is more extensive than in type B modules, which is consistent with their lectin-like binding properties.

Type A and B CBMs are present in plant cell wall degrading enzymes. Type A modules (e.g. members of CBM families 1, 2a, 3, 5 and 10) can be appended to a variety of GHs, while type B modules (e.g. members of CBM families 2b, 4, 6, 15, 17, 20, 22, 27, 28, 29, 34 and 36) are appended to the cognate catalytic modules of fibre degrading enzymes (e.g. cellulases, xylanases, mannanases) targeting the enzyme to their specific polysaccharide substrate. Type C

CBMs, particularly CBM13 and CBM32, are prevalent in bacterial GHs and GTs and toxins that attack eukaryotic cell surfaces or glycans from extracellular matrices [157, 158].

The CBMs affect the degradative capacity of their associated catalytic modules through a proximity effect, determination of substrate affinity and selectivity, targeting function and surface/interfacial modifications of the substrate [153]. By the proximity effect, CBMs increase enzyme concentrations on the substrate surface, thus increasing catalytic activity on the substrate. Substrate affinity and selectivity is determined mainly by orientation of aromatic amino acid residues in the binding site and, to a lesser extent, by hydrogen bond formation. The targeting function is even more subtle than this crude partitioning of enzymes to different plant cell walls polysaccharides, and enables fine tuning of substrate recognition. Different CBMs that bind to the same polysaccharide substrate display clear differences in specificity, targeting hydrolytic enzymes to specific regions of the cell wall. Surface/interfacial modifications of the substrate can result from an increased surface charge after binding of the CBM, which disrupts the double-layer repulsive forces between negative charges of cellulosic surfaces [153]. Non-hydrolytic fibre disruption (amorphogenesis-inducing [159]) activity of CBMs, resulting in increased surface area, reduction of fibre acidity and surface polarity and leading to disorganisation and enhanced availability of polysaccharide chains [155, 160], was demonstrated for substrates such as cellulose [161], starch [162] and chitin [163].

Hervé *et al*. (2010) [164] demonstrated that CBMs use targeting and proximity effects to promote the enzymatic deconstruction of polysaccharides in intact plant cell walls. It was previously observed that crystalline cellulose-binding CBMs may be found in enzymes that hydrolyse non-cellulose substrates, such as pectins or xylans. In these enzymes, cellulose-binding CBMs were shown to enhance the degradation of pectic homogalacturonan by PLs, while xylan-binding CBMs enhanced the removal of side chains from arabinoxylan by arabinofuranosidases, whereas both the cellulose and xylan binding CBMs potentiated the capacity of xylanases to degrade xylan in cell walls. Appended CBMs most likely potentiated the hydrolytic action of cognate catalytic module through recognition and binding of unrelated, non-substrate polysaccharide, bringing the catalytic module into proximity with their substrate within plant cell walls [164].

Carbohydrate-CBM interactions have relatively low affinity, with an affinity constant (K_a) $<10^6\ M^{-1}$ for cellulose, and even lower for hemicellulose [157]. In contrast, K_a of the high affinity binding interactions, such as the strongest known non-covalent interaction between biotin and avidin (or streptavidin), ($K_a \sim 10^{15}\ M^{-1}$) [165] or cohesin-dockerin pairing ($K_a \sim 10^9 - 10^{12}\ M^{-1}$) [166] are significantly higher. However, avidity resulting from multivalent interactions between CBMs and their ligand can compensate for these weak interactions in nature and can be a result of multiple carbohydrate-binding motifs in a single CBM (e.g. CmCBM6-2 from *Cellvibrio mixtus* endoglucanase 5A [156]) or tandemly arrayed multiple

CBMs, frequently found in GHs (e.g. three CBM6 modules in a *Clostridium stercorarium* thermophilic xylanase [167]). Increasing substrate affinities of CBMs is an important target for cellulose modification and short peptides with the ability to bind cellulose are of special interest since they can be used as tandem or clustering motifs for building up artificially designed CBMs with higher substrate affinities [168].

1.3.3.5 Cellulosomes

In the early 1980s, Lamed and Bayer proposed the term 'cellulosome' to describe the discrete multi-enzyme organisation of the cellulose-degrading complex from the anaerobic bacterium *Clostridium thermocellum* [169-171]. Cellulosome systems and their organisation have been extensively studied in other Gram-positive anaerobes belonging to clostridial and ruminococcal species [64]. Initially, it was believed that the cellulosomal complexes exclusively degrade crystalline cellulosic substrates, but it was soon recognised that bacterial cellulosomes contain an array of CAZyme modules (such as GHs, PLs and CEs) with hemicellulase and even pectinase activities [172-174]. Moreover, additional types of enzymes (e.g. peptidases), serpins, and putative structural proteins also appear to be components of cellulosomes.

The cellulosomal organisation of fibre-degrading enzymes was found in anaerobic fungi from genera *Neocallimastix*, *Piromyces* and *Orpinomyces*, as well as in many representatives of Clostridiales [64, 175]. For some of these bacteria (*Bacteroides cellulosolvens* ATCC35603, *Clostridium thermocellum* ATCC27405 and DSM1313, *C. cellulolyticum* H10 ATCC 35319, *C. josui*, *C. clarifavum* DSM19732, *Clostridium sp.* BNL 1100) a recent re-classification to genus *Ruminiclostridium* [176] and family Ruminococcaceae [177] has been proposed. Cellulosomal organisation was also found in other Ruminococcaceae (*Acetivibrio cellulolyticus* CD2, *Ruminococcus flavefaciens* FD-1 and 17) and Clostridiaceae (*C. cellulovorans* 743B, non-cellulolytic bacterium *C. acetobutylicum* ATCC824 and EA2018) [64, 175].

The cellulosome is an intricate extracellular multisubunit complex with an average molecular mass of 2 MDa, involved in the efficient degradation of crystalline cellulose and other associated polysaccharides [175, 178]. The main cellulosomal components are large non-catalytic polypeptides, termed scaffoldins [179], that spatially integrate CBMs and catalytic domains of enzymes and other cellulosomal components into a single functional entity. The scaffoldin subunits are composed of multiple copies of cohesin modules, often in combination with other modules. Scaffoldins may play multiple roles, such as: integration of dockerin module-bearing catalytic subunits into the cellulosomal complex through interaction with its complementary cohesin modules; anchoring of the cellulosome to the cell wall by virtue of its C-terminal dockerin or surface (S)-layer homology (SLH) module, and targeting of the

fibrolytic bacterium to the carbohydrate substrate by its CBM. The high-affinity cohesin-dockerin interaction (K_a ~10^9 – 10^{12} M^{-1}) is involved in the specific integration of polysaccharide hydrolases into the cellulosome complex, determining the supramolecular architecture of the entire complex and providing its stability [175, 180, 181].

Dockerins are domains 70 amino acid residues in length, present in a single copy at the C-terminus of cellulosomal enzymes and are comprised of two duplicated segments (repeats), each containing a motif about 22 amino acid residues in size, spaced by a linker [182, 183]. The first 12 amino acid residues of each repeat are similar to the eukaryotic EF hand motif and are highly conserved. However, this homology is restricted only to the calcium-binding loop, containing the highly conserved calcium-binding amino acid residues asparagine or aspartate. This is consistent with the calcium-dependence of the cohesin-dockerin interactions [184, 185].

Cohesin modules, approximately 140 amino acid residues in length, are usually tandemly repeated between 4 and 11 times in scaffoldin proteins [186]. Several types of cohesin modules have been distinguished on the basis of their structure and binding specificity. Dockerins are, by definition, designated to be of same type as cohesins they are interacting with. Type I cohesins can be found in the 'primary' (enzyme-integrating) scaffoldins of the majority of described cellulosomes [187], while type II cohesins, originally discovered in a group of non-catalytic, cell-surface 'anchoring' scaffoldins of *C. thermocellum*, have been also detected in cellulosomes of *A. cellulolyticum* and *B. cellulosolvens* [188]. Type III cohesins have been described only in *Ruminococcus flavefaciens* [189]. Regardless of type, comparative studies of cohesin domains indicate that a common structural feature of the cohesin fold is a nine-stranded β-sandwich with an overall 'jelly roll' fold [190]. Both type II and type III cohesins may have several additional structural elements, such as α-helix and two 'β-flaps'. The crystal structure of the type III ScaE cohesin from *R. flavefaciens* showed that the additional α-helix is enveloped by an extensive N-terminal loop, a feature not seen in any other known cohesins [191-193]. It has been suggested that these structural differences between different types of cohesins have a role in the type-specificity of cohesin-dockerin interactions.

Cohesins and dockerins are well conserved between species [128]. Species-specificity of the cohesin-dockerin interaction, reported for several pairs of clostridial species, depends on conserved amino acid residues at positions 11 and 12 of dockerin repeats [166, 194, 195] and the combination of both segments is important for target recognition [196]. However, several exceptions to this rule are known. For example, Cel9D-Cel44A or Xyn11A dockerin modules of *C. thermocellum* can interact with cohesins from *C. josui* [196, 197]

Cohesins and dockerins were long considered the signature domains for identification of the cellulosome-producing microbes. However, the extensive bioinformatic mining of genomes of Bacteria, Archaea and primitive eukaryotes identified putative non-cellulosomal cohesin and dockerin modules in proteins with unknown or non-fibrolytic predicted functions,

suggesting that the cellulosomal paradigm may be the exception rather than the rule for the use of these modules in the three domains of life [198].

Although it was initially proposed that the main cultured representatives of rumen cellulolytic bacteria could use cellulosome-like complexes for the initial adhesion to, and degradation of substrates [59], these structures appear to be scarce. The cellulosomal organisation and anchoring of fibre-degrading enzymes *via* scaffoldins has only been implicated for several strains of *R. flavefaciens* [189, 190, 193, 199-201]. In contrast, *R. albus* [202] and *F. succinogenes* [15, 203] seem to use a direct cell-surface anchoring of their enzymes.

The cellulosome system of *R. flavefaciens* strain 17 is very elaborate (Figure 1.3). Its particular feature is a large number of cohesin-dockerin specificities and, in contrast with clostridial systems, not all enzyme-bound dockerins have the same binding specificity, enabling a more ordered arrangement of enzymes in the cellulosome complex [204]. The multiple scaffoldin-encoding *sca* gene cluster consists of five genes (*scaA-E*), encoding proteins bearing one or more cohesins. The cluster also includes the *cttA* gene, which encodes a cell wall-anchored substrate-binding protein containing two putative CBMs [205]. The cellulosome is anchored to the cell wall covalently *via* a single cohesin (ScaE), and this is mediated through the sortase mechanism. ScaE binds to the dockerin at the C-terminus of the large scaffoldin, ScaB, and its cohesins accommodate C-terminal dockerins of the smaller scaffoldin ScaA. The ScaA contains three cohesins, which can bind various dockerin-containing enzyme subunits.

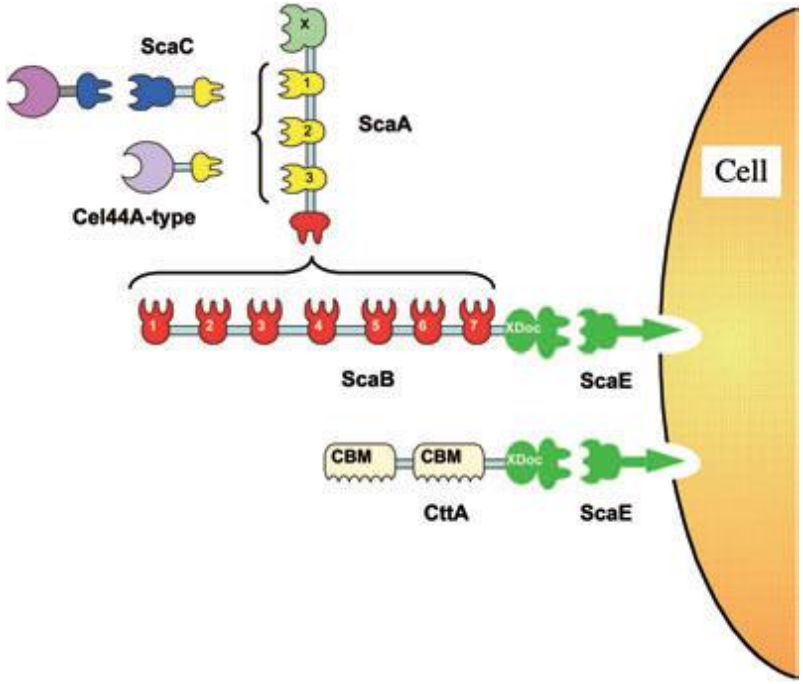

Figure 1.3 Schematic overview of the *Ruminococcus flavefaciens* 17 cellulosome.

The scaffoldin ScaB, and CBM-containing protein CttA, are bound to the bacterial cell wall-anchoring ScaE cohesin *via* the conserved XDoc dyad. The seven ScaB cohesins (red) interact

with the ScaA dockerins (red), thus increasing the number of components incorporated into the cellulosome. The ScaA cohesins (yellow) bind directly to a group of Cel44A-like enzyme containing dockerin (yellow) or alternatively, they bind to the dockerin (yellow) from ScaC scaffoldin. ScaC has a divergent cohesin type (blue) that recognises and incorporates a different set of dockerin containing enzymes and other components into the cellulosome. Figure taken from [175] with permission.

1.3.4 Metagenomic studies of fibre-degrading genes of rumen microbiomes

1.3.4.1 Metagenomics and next-generation sequencing technologies

Microorganisms represent a diverse and largely untapped resource for the discovery of novel genes, bioactive molecules and new biocatalysts that may be applied to improve industrially relevant processes [206, 207]. Traditional approaches to explore complex microbial communities *via* the cultivation of microorganisms and screening for individual strains with the desired phenotype(s) fail to sample this extraordinary diversity, since only a limited fraction of microbes are represented by cultures. The nature of these complex microbial communities is being investigated through culture-independent approaches, collectively referred to as metagenomics, for the analysis of the collective microbial genomes (metagenomes) recovered directly from environmental samples [208]. Metagenomic methods enable significantly higher resolution and throughput in assessment of microbial community composition and function compared to traditional approaches.

Metagenome analysis strategies may be sequence-based and/or function-based [209] and encompass approaches such as: deep sequencing of amplicons derived from phylogenetically informative genes and regions within metagenomic DNA (e.g.16S rRNA genes) to assess microbial community composition; direct shotgun sequencing of metagenomic DNA to determine microbial diversity and uncover the functional potential of the community, and targeted functional screens of expression libraries constructed from metagenomic DNA [210, 211].

Development of next-generation sequencing (NGS) technologies in the last decade allowed massively parallel sequencing of hundreds of thousands to hundreds of millions of DNA templates. Using this methodology, high-throughput sequencing of genomes or amplicons derived from environmental microbiomes has become feasible [212]. This approach is affordable, as it can produce thousands of sequence reads in a single run at a fraction of the cost of traditional dye-terminator sequencing. Furthermore, sample-specific ‘barcoding’ of templates combined with massively parallel sequencing allows sequencing of multiple samples in the same run. All NGS platforms (Table 1.2) that are currently widely used in genomics and

metagenomics require preparation of libraries of clonally amplified templates, introducing amplification artefacts and library preparation errors [213]. Emerging single-molecule sequencing technologies aim to overcome these issues through elimination of the amplification step. A summary of available NGS platforms and their outputs is provided in Table 1.2.

Table 1.2. Summary of the available NGS platforms and their outputs.

Platform	Company	Technology	Instrument	Read length [a]	Reads per run [b]	Sequence yield per run [b]	Reported accuracy [c]
454	454 Life Sciences (Roche)	emPCR, (pyro) sequencing by synthesis	GS Junior/GS Junior +	400-700 bases	0.07-0.1 million	35-70Mb	99% [d]
			GS FLX Titanium XLR70	450-600 bases	0.7-1 million	450 Mb	99.995% [e]
			GS FLX Titanium XL+	700-1000 bases	1 million	700 Mb	99.997% [e]
Illumina	Illumina	reversible-terminator dye sequencing by synthesis technology	MiSeq	2x250 bp (reagent kit v2) and 2x300 bp (reagent kit v3) [a]	12-15 million (v2) and 22-25 million (v3)	0.3-15 Gb	> 75% (2x250 bp) and >70% (2x300 bp) [f]
			NextSeq500	2x150 bp [a]	130-400 million	30-120 Gb	>75% [f]
			HiSeq2500	2x150 bp (rapid) and 2x125 bp (high-output) [a]	600 million (rapid) and 4 billion (high-output)	150-180 Gb (rapid) and 0.9-1Tb (high-output)	>80% (rapid) and >75% (high-output) [f]
			HiSeqX	2×150 bp	3 billion	1.6-1.8 Tb	>75% [f]
SOLiD	Life Technologies	emPCR, Sequencing by Oligonucleotide Ligation and Detection (SOLiD), two base encoding system	5500xl SOLiD System	75 bases and 2x60 bp	2.8-4.8 billion	180-300Gb	99.99% [g]
HeliScope	Helicos biosciences [h]	true Single Molecule Sequencing (tSMS), reversible dye terminator	HeliScope Sequencing System	25-55 bases	0.6-1 billion	21-35 Gb	>99.995% [a]
Ion Torrent	Life Technologies	emPCR, ion semiconductor sequencing	Ion Proton System	up to 200 bases	60-80 million	up to 10Gb	>99%
			Ion PGM System [i]	200 or 400 bases	0.4-5.5 million [j]	30Mb-1Gb (200 base read) or 60 Mb-2 Gb (400 base read) [i]	>99% [k]

Platform	Company	Technology	Instrument	Read length [a]	Reads per run [b]	Sequence yield per run [b]	Reported accuracy [c]
PacBio	Pacific Biosciences	Single Molecule Real Time (SMRT) sequencing	PacBio RS II (P4-C2 and P5-C3 chemistry available)	5,500 (P4-C2) and 8,500 (P5-C3) bases on average	50,000 reads	275 Mb (P4-C2) and 375 Mb (P5-C3)	>99.999% (for P4-C2)
Nanopore	Oxford Nanopore Technologies	Single strand sequencing, changes in ion current generated by DNA passing through a nanopore	GridION and MinION (in a customer-testing phase)	> 5,500			~96%

Summary of the announced and commercially available NGS platforms and their outputs as of December 2013. [a] Read length is presented in bases for single-end reads (template sequenced from one end) and in bp for paired-end reads (template sequenced from both ends); [b] Dependant on instrument and chemistry used; [c] Accuracy is reported differently for each platform. When a quality score (Q-score) is used, it refers to a prediction of the probability of an error in base calling and Q20, Q30 and Q40 correspond to an accuracy of 99%, 99.9% and 99.99%, respectively; [d] Q20 read length of 700 bases (99% accuracy at 700 bases and higher for preceding bases); [e] Consensus accuracy at 15x coverage; [f] The percentage of bases > Q30 is averaged across the entire run; [g] Accuracy is based on sequencing control synthetic beads, and reference-free data analysis; [h] Helicos no longer sells instruments, but conducts sequencing through a service centre; [i] Ion 314, 316 and 318 Chip v2 available; [j] Depending on the chip used; [k] Consensus accuracy at 20x read coverage, aligned/measured Q30 accuracy of the majority of bases out to 250 bases.

Multiple benchmarking studies, comparing performance of different NGS platforms in terms of their throughput, sequencing read length, error rate, cost and run time [214-219], clearly demonstrate that each platform has unique combination of strengths and constraints, and consequently, its own application niche. Both SOLiD and Ilumina platforms generate the highest sequencing yield per run with low error rates, mainly contributad by substitution errors. Illumina had achieved substantial recent improvements in extending the read length and lowering the consensus error rate. Single molecule real-time sequencing on the Pacific Biosciences (PacBio) platform currently generates the longest reads, that are, however, highly prone to non context-specific insertion and deletion (indel) errors. In contrast to chain-termination technologies, pyrosequencing (Roche 454) and semiconductor sequencing (Ion Torrent) are prone to indels in homopolymer regions.

Compared to other two platforms comonly used in metagenomics (Illumina and Ion Torrent), the 454 sequencing platform delivers the longest read lengths with the lowest per base error. However, systematic artefacts intrinsic to the 454 sequencing platform may affect the accuracy of downstream sequence analyses. Gomez-Alvarez *et al*. (2009) [220] reported an over-abundance (11-35%) of raw reads representing artificial replicates in several published and original shotgun metagenomic pyrosequencing datasets. These replicates belong either to

clusters of identical reads (duplicates) or to clusters of reads begining at the same 5' position that can vary in length or contain sequence discrepancies [221]. Replicates result in overestimation of certain genes' and taxons' abundance detected in metagenomic pyrosequencing datasets. Conversely, filtering of the replicates in the dataset raise the issue of their underestimation, due to inability to distinguish natural from artificial replicates [220, 221]. In addition, systematic errors in homopolymeric tracts, which cannot be mitigated by increasing the depth of coverage or assembling reads, may lead to frame-shifts in coding regions, thus affecting accuracy of functional annotations of pyrosequencing datasets [213].

A hybrid approach, combining sequencing data generated on multiple platforms is a strategy of choice not only for obtaining high-quality reference genomes, but also for metagenomic studies [222].

1.3.4.2 Metagenomic studies of fibre-degrading rumen microbial communities

In the metagenomic era, complex, biomass-degrading rumen microbial communities represent a valuable genetic resource for discovery of novel fibrolytic activities from a large number of yet uncultured species, with potential applications in agriculture and biofuel production, but that are also important for elucidating mechanisms underlying fibre degradation [19, 26]. The unique rumen bacterial metagenome has been estimated to be around 2.7 Gb in size and encode approximately 2.7 million genes [223]. The construction of clone libraries of size sufficient to represent the entire rumen metagenome is not feasible. The rise of low-cost, high-throughput next-generation sequencing (NGS) technologies has allowed generation of extensive rumen microbial gene catalogues, which are key for understanding microbial functions and their interactions with feed and ruminant host. However, the cataloguing process is influenced by both technical limitations of current NGS sequencing platforms and limitations of bioinformatic tools used for sequence analysis. These limitations include low metagenome coverage; overprediction of large numbers of incomplete genes due to short reads obtained with current sequencing platforms, poor assemblies and lack of 'metagenome-specific' tools for biologically meaningful functional analysis. For this reason, cataloguing rumen microbial functions through metagenomics and genomic studies will require considerable sequencing and bioinformatic efforts in the future [26, 211].

Recent culture-independent studies of metagenomes of several herbivores harbouring plant-biomass degrading communities, such as the wood degrading termite [16], Tammar wallaby [20], Svalbard reindeer [23], yak [24] and cow [19, 22] have provided new insights into the potential mechanisms and diversity of protein domains and CAZy families associated with gut microbial fibre degradation.

In 2009, Brulc and colleagues [19] were the first to apply high throughput metagenomic shotgun sequencing to the fibre adherent and liquid rumen microbiomes from steers fed a grass-legume hay diet. Starting from total of 104 Mb of raw metagenomic DNA sequence, the coding potential for fibre-degrading enzymes was obtained from unassembled sequence datasets, *via* similarity-based searches against the CAZy database. In contrast to the abundance of hits to 35 different putative GH families mainly implied in digestion of side chains and oligosaccharides, a surprisingly low numbers of CBM and dockerin hits and no hits to cohesins were detected in these datasets. The apparent under-representation of CBMs, dockerins and cohesins among the CAZyme hits may have been due to lower abundance of genes encoding these modules in the metagenome dataset. However, due to the short pyrosequencing read lengths used in this study (averaging ~100 bp), it was shown that the relatively shorter dockerin and CBM modules are underrepresented in sequence datasets in similarity-based searches [19].

In 2011, Hess *et al.* [22] published a metagenomic study with considerable depth of sequence coverage, which focused on mining the switchgrass-adherent microbiome of the cow rumen for biomass-degrading genes. Starting from 268 Gb of raw metagenomic DNA sequence, the assembled and annotated data allowed the detection of 27,755 putative carbohydrate-active genes. The number of reported carbohydrate-active genes was 5-fold greater than previously reported in three metagenomic studies of different biomass-degrading microbiomes [16, 19, 20]. However, only a limited fraction of the total CAZyme hits were to cohesin- and dockerin-specific Pfam domains. In order to minimise a bias towards the detection of candidate CAZymes with overall similarity to known enzymes, putative open reading frames (ORFs) predicted in assembled dataset were screened against Pfam HMMs representing 68 GH catalytic domains and 22 CBMs. The ability to detect putative CAZymes with limited overall sequence identity to enzymes already deposited in public databases was assessed, resulting in 43% of putative CAZymes in this study sharing less than 50% amino acid sequence identity to proteins deposited in the NCBI nr database.

Metagenomic surveys of genes encoding the fibrolytic capacities of rumen microbiomes of Svalbard reindeer [23] and yak [24] have also contributed to cataloguing the diversity of these genes in microbiomes of different ruminants, and have provided a glimpse into the remarkable complexity of the rumen microbiome and its functions in different ruminant hosts. Comparisons of the profiles of genes encoding selected GH families targeting plant structural polysaccharides in four rumen metagenomes is represented in Table 1.3.

Table 1.3. Profiles of genes encoding selected GH families and cellulosome domains in four different rumen metagenomes.

Predominant activity of GH family members		Bovine fibre-adherent [a,b] [19]	Bovine switchgrass-adherent [c,d] [22]	Svalbard reindeer fibre-adherent [a,d] [23]	Yak whole rumen [c,d] [24]
Cellulases					
GH5	cellulases	20	1451	287	1302
GH6	endoglucanases	0	0	0	0
GH7	endoglucanases	0	1	0	0
GH9	endoglucanases	17	795	109	767
GH44	endoglucanases	0	0	5	0
GH45	endoglucanases	0	115	0	13
GH48	cellobiohydrolases	1	3	5	32
Total cellulases		38	2365	406	2114
Endohemicellulases					
GH8	endoxylanases	7	329	35	174
GH10	endo-1,4-β-xylanases	16	1025	190	2664
GH11	xylanases	1	165	8	244
GH12	xyloglucanases	0	0	0	0
GH26	β-mannanases and xylanases	16	369	153	537
GH28	galacturonanases	9	472	120	244
GH53	endo-1,4-β-galactanases	51	0	125	1066
Total endohemicellulases		100	2360	631	4929
Xyloglucanases					
GH16	xyloglucanases	1	483	116	563
GH74	endoglucanases and xyloglucanases	0	0	44	0
Total xyloglucanases		1	483	160	563
Debranching enzymes					
GH51	α-L-arabinofuranosidases	184	0	488	0
GH54	α-L-arabinofuranosidases	4	0	23	111
GH62	α-L-arabinofuranosidases	0	1	0	0
GH67	α-glucuronidases	0	120	74	1090
GH78	α-L-rhamnosidases	93	1260	313	426
Total debranching enzymes		281	1381	898	1627
Oligosaccharide-degrading enzymes					
GH1	β-glucosidases	31	253	122	331
GH2	β-galactosidases	527	1436	716	942
GH3	β-glucosidases	497	2844	844	5448
GH29	α-L-fucosidases	79	939	268	899
GH35	β-galactosidases	27	158	39	468
GH38	α-mananosidases	46	272	116	90
GH39	β-xylosidases	7	315	76	159

Predominant activity of GH family members		Bovine fibre-adherent [a,b] [19]	Bovine switchgrass-adherent [c,d] [22]	Svalbard reindeer fibre-adherent [a,d] [23]	Yak whole rumen [c,d] [24]
GH42	β-galactosidases	35	374	95	207
GH43	arabino/xylosidases	176	0	787	2313
GH52	β-xylosidases	0	0	2	0
Total oligosaccharide -degrading enzymes		1425	6591	3065	10857
Total number of GHs detected in the study		2720	27755	5160	37563
Cellulosome domains					
Cohesins		0	80	52	51
Dockerins		1	188	92	516
Raw sequence information		0.08	268	0.5	0.09

Data are presented in the format used in [224], with GH families targeting plant structural polysaccharides grouped according to their major functional role in the plant fibre degradation. [a] Unassembled metagenome; [b] Combined metagenome datasets of three individual animals; [c] Assembled metagenome; [d] Metagenome obtained from pooled rumen samples of two animals.

1.4 Metasecretome

1.4.1 Definition of the bacterial secretome

The term 'secretome', coined by Tjalsma *et al*. in 2000 [225], was originally proposed to refer to both the secreted proteins and components of the protein secretion machineries in bacteria. Today, the secretome is broadly described as a subset of the bacterial proteome, containing the extracellular proteome (exoproteome), released to the extracellular milieu and the surface-associated proteome, either exposed to the bacterial surface or intrinsic to the external side of plasma membrane and the cell wall, but excluding integral membrane proteins and proteins intrinsic to the internal side of the plasma membrane [226, 227].

Secretome proteins (e.g. receptors, transporters, adhesins, complex cell structures, secreted enzymes, toxins and virulence factors) allow bacteria to interact with, and adapt to their environment. Bacterial secretory proteins are known to be involved in processes such as: provision of nutrients through recognition; binding, degradation and uptake of complex extracellular molecules; communication between bacterial cells; detoxification of the environment; attachment to host cells and signal transduction; while in pathogenic bacteria they also play critical roles in virulence and immunogenicity [228-232]. Secretome proteins have been reported to occupy 10 – 30% of the total coding capacity of bacterial genomes [27, 28]. Bacterial lipoproteins typically represent 2% of the bacterial proteome [233].

In this thesis, the term ‘metasecretome’ will be used to describe a collection of secreted, surface and integral membrane proteins (secretome) [234] derived from environmental microbial communities.

1.4.2 Secretion pathways of monoderm and diderm bacteria

The number of membranes is essential in the context of protein transport and for this reason terms monoderm and diderm bacteria are preferentially used over the terms Gram-positive and Gram-negative bacteria, which can be ambiguous and can refer to the cell staining properties, organisation of cell envelope or taxonomic group [226]. The cell envelope of monoderm bacteria consists of a single, cytoplasmic membrane and a cell wall, comprised of a thick peptidoglycan layer cross-linked with different molecules, such as capsular polysaccharides, cell wall teichoic acids and proteins [235]. In contrast, diderm bacteria are enveloped by inner (cytoplasmic) and outer membranes. The presence of two membranes defines an additional subcellular compartment (the periplasmic space), containing a thin meshwork of peptidoglycans. Some monoderms also have a distinctive thin granular layer (inner wall zone) between the membrane and the mature cell wall, equivalent to the periplasmic space in diderm bacteria [236].

In order to be anchored to the cell surface or released into the extracellular milieu, secretome proteins must be translocated across one or more biological membranes [237]. Transport of proteins into or across biological membranes (translocation), catalysed by membrane-bound proteinaceous transport machineries, is a universal event in the protein secretion mechanism, and it can occur several times during the course of secretion [226].

Once a secreted protein is translocated across the outermost membrane, it can remain anchored (covalently or non-covalently associated with cell-wall components in monoderm bacteria or outer membrane components in diderm bacteria); assemble into macromolecular structures on the cell surface (flagella, pili), be injected into host cells, or released to the extracellular milieu.

A remarkable array of protein export systems have been described in monoderm and diderm bacteria. Descriptive names are used in the nomenclature of systems involved in protein translocation across cytoplasmic membranes of both diderm and monoderm bacteria, while an alphanumerical system has been adopted for naming protein secretion systems of diderm bacteria [226].

1.4.2.1 Protein transport systems universal for all bacteria

Systems that are universally involved in protein translocation across the cytoplasmic membrane, and encoded in both monoderm and diderm bacteria are: the conserved general secretion (Sec) system, YidC insertase, the twin-arginine translocation (Tat) system and hole-forming pathway *via* holins [226].

The Sec system is a major secretory pathway for protein insertion into the inner membrane, and is conserved in all eubacteria. It is also ubiquitous in archaea, and the membranes of eukaryotic endoplasmic reticulum and chloroplasts. This system also plays a key role in further transport of some proteins into the periplasmic space, outer membrane (e.g. lipoproteins and beta barrel proteins), or their assembly into the surface-associated structures (e.g. pili subunits). Furthermore, some of the components of the specialised secretion systems in diderms and their substrates (proteins transported via these secretion systems) are initially transported across the inner membrane by the SecYEG translocon.

In bacteria, the Sec system is composed of the SecYEG translocon and three major accessory systems that target the secretome proteins to the translocon: SecB/A, SRP/FtsY and YidC. SecYEG is an evolutionarily conserved heterotrimeric protein complex, and its SecY subunit forms an hourglass-shaped aqueous protein transport channel embedded in the inner membrane [27, 232]. The translocon transiently interacts with different proteins during the transport process (e.g. SecA, FtsY, SecDF). SecA, a post-translational pathway motor protein accepts the substrate protein delivered by the cytosolic targeting factor SecB, and pushes it through the translocon in a stepwise and ATP-dependant manner [238]. FtsY, the SRP-receptor, occupies the ribosome binding site of SecY until its displacement by the translating ribosome during co-translational targeting [27]. The membrane-integrated SecDF chaperone provides proton-motive force to power ATP-independent protein translocation through the SecYEG channel [239].

The signal sequence of the transported protein is first inserted from the cytosolic side into the membrane *via* the positively charged N-terminus. Routing into or across the membrane is aided by recognition of the N-terminal hydrophobic signal sequence of the transported proteins. Integral inner membrane proteins and some exported proteins with highly hydrophobic signal peptides are targeted co-translationally by the signal recognition particle (SRP), while proteins that should reside in the periplasm, outer membrane or outside the cell envelope are targeted post-translationally by the SecA/B pathway [235, 240]. The soluble part of the protein is transported across the membrane *via* interaction between SecA and SecYEG and the signal sequence is eventually cleaved-off by a membrane embedded signal peptidase. In the case of membrane proteins, a more hydrophobic signal sequence, that serves as a membrane-anchor and does not contain a signal peptidase cleavage site, is recognised by the SRP. SRP binds

hydrophobic stretches [241] and α-helical transmembrane domains close to the N terminus of the secretome protein [242]. The hydrophobic α-helical transmembrane domains are first inserted into SecYEG and then moved into the lipid bilayer through a lateral gate of the translocation pore [243].

The YidC insertase is present in all bacteria, many archaea, as well as mitochondria and chloroplasts [27], and is used as an alternate inner-membrane insertion mechanism that can act both independently or in cooperation with SecYEG [244-246]. The SecYEG translocon and YidC insertase recognise the N-terminal hydrophobic signal sequence through cooperation with cytosolic partner proteins and translocate unfolded proteins. Inner membrane proteins use YidC as an alternate insertion site *via* targeting through the SRP/FtsY pathway, which recognises exported proteins independently of the downstream integration site. Site-directed cross-linking showed that the C-terminus of YidC is in contact with SRP, the SRP receptor and ribosomal proteins [242, 243, 247]. It was previously thought that only small inner membrane proteins could use this route, but Welte *et al*. (2012) [242] have recently demonstrated that YidC can also integrate more complex and larger proteins, such as the multi-spanning membrane proteins mannitol permease or TatC. At present, it seems that only SecA-dependent multi-spanning membrane proteins are exclusively integrated into the cytoplasmic membrane by SecYEG, but not by YidC, suggesting that SecA can interact with the SecYEG translocon, but not with the YidC insertase. Most membrane proteins are recognised by the SRP, and then targeted to the next available insertion site (either SecYEG translocon *via* the SRP receptor or FtsY). This prevents all translocons being occupied by co-translational insertion, in which protein synthesis, a much slower process than protein transport, is the rate-limiting step [242].

The Tat system is present in most bacteria, archaea and chloroplasts, and transports fully folded and even multimeric proteins containing ‘twin-arginine' amino-acid sequence motifs across the inner membrane. Depending on the bacterial group, Tat translocases are comprised of only 2 – 3 integral membrane protein subunits, providing a transmembrane route big enough for the passage of structured macromolecular substrates, while maintaining the membrane barrier intact [235, 248]. TatABC-type translocases are typical for diderms, while the minimal TatAC translocases are typical for monoderm bacteria [27]. Instead of using a general signal sequence recognition protein in the cytosol, like the other two universal transport systems, the Tat system exploits specific signal-sequence binding chaperones encoded by the same operon as the cognate Tat substrate [249].

The hole-forming pathway exports fully folded peptidoglycan hydrolases (endolysins) from the cytosol to the cell wall *via* holins, small transmembrane proteins that can integrate into the bacterial cytoplasmic membrane and form large oligomeric flexible pores. Holins, and their target endolysins, are encoded by a wide variety of monoderm and diderm bacteria and their

lytic phage, and mediate bacterial cell lysis in programmed cell death and phage infection [250-252].

1.4.2.2 Protein export systems specific to monoderm bacteria

In addition to universal secretion systems, monoderm bacteria possess Wss (WXG100 secretion systems), accessory Sec systems (SecA2-only and SecA2/SecY2 export pathways), flagella export apparatus (FEA), the fimbrilin-protein exporter (FPE) and Sec-dependent sortases. In monoderm bacteria, secreted proteins can have different fates (Figure 1.4). They are transported across the cytoplasmic membrane *via* the Sec and YidC pathways and then secreted into the extracellular milieu *via* SecYEG, Tat, holin or Wss, in addition to being attached (covalently or non-covalently) to the cell wall using the sortase or assembled into the cell surface appendages *via* Sec pathway (e.g. cellulosomes or pili), *via* FPE (e.g. competence pseudo-pili) or *via* FEA (e.g. flagellae).

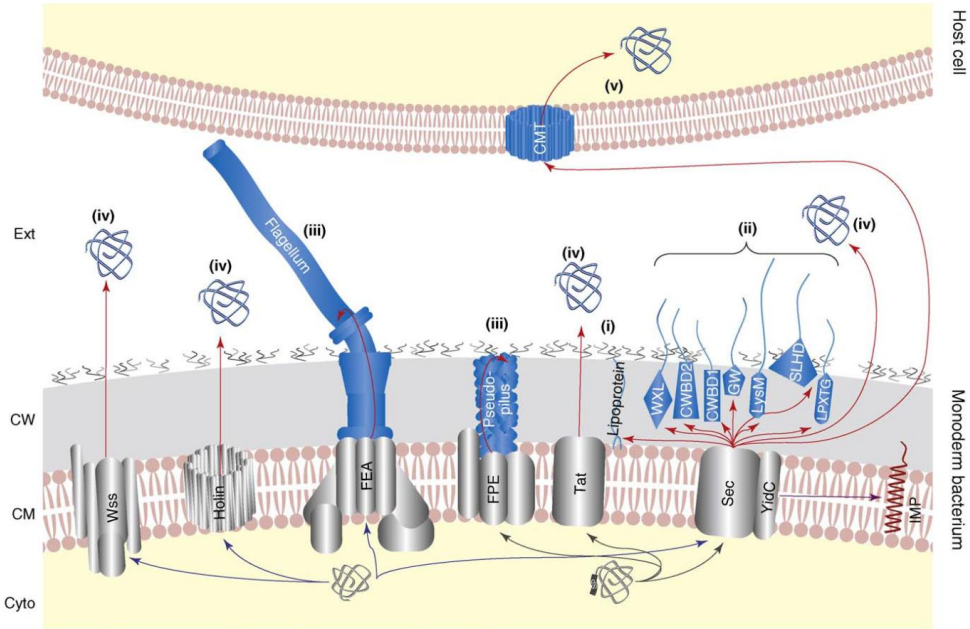

Figure 1.4 Protein export systems of monoderm (Gram-positive) bacteria.

Secreted proteins of monoderm bacteria can be: (i) anchored to the CM (e.g. lipoproteins); (ii) attached to the CW either covalently (e.g. LPXTG proteins) or non-covalently (e.g. proteins containing LysM, GW, CWBD1, CWBD2, SLHD or WXL motifs); (iii) part of cell-surface structure, such as cellulosomes or pili (their subunits are secreted *via* Sec), competence pseudo-pili (assembled *via* FPE) or flagella (assembled *via* FEA); (iv) released into the Ext *via* Sec, Tat, holin or Wss; or (v) translocated into a host cell after secretion of a cholesterol-dependent cytolysin (*via* Sec), which integrates into the host cell membrane permitting transport of secreted effectors in a process called cytolysin-mediated translocation (CMT). Secreted proteins are represented in blue. Black arrows indicate routes of proteins targeted to the CM exhibiting

an N-terminal signal peptide, whereas blue arrows are routes used by proteins lacking such a signal peptide. Red arrows are related to secretion, and violet arrows are related to integration of IMP. Abbreviations: Cyto, cytoplasm; CM, cytoplasmic membrane; IMP, integral membrane proteins; CW, cell wall; OM, outer membrane; Ext, extracellular milieu. Figure taken from [226] with permission.

The FEA and FPE of monoderm bacteria share homology with type 2 and type 3 protein-secretion systems, respectively, described in diderms [226]. Sortases are transpeptidases involved in the covalent anchoring of surface proteins containing a C-terminal LPXTG or LPXTG-like motif and are transported *via* the Sec translocon across the cytoplasmic membrane [253]. Sortases have been traditionally described as protein secretion systems unique to monoderm bacteria. However, genes encoding sortase homologues have been identified across eubacteria and even in the Archaea [253, 254]. Also, sortases do not have an active role in translocation across the membrane, *sensu stricto*, but act as post-translational architects involved in the maturation of the bacterial surface by recognizing and cleaving off C-terminal sorting signals and mediating protein anchoring to peptidoglycans [226]. Other cell wall associated proteins, such as S-layer proteins which self-assemble into para-crystalline lattices on the surface of some Gram-positive bacteria (typically bacilli and clostridia) and S-layer-associated proteins, which either act as a scaffold or an enzyme, do not possess LPXTG motifs. These proteins have specific SLH domains instead, enabling non-covalent tethering of these proteins to secondary cell wall polysaccharides at the bacterial surface [255]. It was demonstrated that export of an S-layer protein in *Clostridium difficile* is dependent on SecA2, and it is possible that this additional system is universally involved in the export of SLH proteins in other bacteria [256].

The Wss and the secondary Sec2 export pathways are unique to monoderms. Wss, also called Ess (ESX secretion system), mediates the secretion of small proteins, such as virulence factors belonging to the WXG100 (named because of WXG motif and length of around 100 amino acid residues), or exclusively mycobacterial PE/PPE protein families [235].

Some, mainly pathogenic, monoderm bacteria, possess accessory Sec systems (Sec2 pathway), composed of an additional SecA protein (SecA2-only system) or additional SecA2 and SecY2 proteins (SecA2/SecY2 system). Unlike the housekeeping, conserved general Sec (‘Sec1’) pathway, which handles the export of the majority of proteins, the Sec2 systems are involved in the transport of a limited number of substrates, mainly virulence factors [235]. Few substrates of the accessory SecA2-only systems of pathogenic bacterial species have been identified so far [257, 258]. Some of these substrates lack a signal sequence (e.g. superoxide dismutases SodA and MnSod, catalase-peroxidases KatG, fibronectin-binding protein FbpA) and some do not (e.g. autolysin p60). For this reason, SecA2 has been implied in both signal peptide dependent (classical) and non-classical secretion [259]. It is thought that translocation of

these substrates across the cytoplasmic membrane, mediated by either SecA2 alone or in cooperation with SecA, is through the SecYEG pore. At present, it is unclear which features distinguish substrates exported *via* the Sec pathway from SecA2-only substrates, and how are these substrates recognised by A2. The accessory SecA2/SecY2 systems export large serine-rich glycoproteins that serve as adhesins in monoderm bacteria [260, 261].

1.4.2.3 Protein export systems specific to diderm bacteria

Due to the added complexity of their cell envelope, at least two additional systems for targeting proteins to the outer membrane and eight additional systems for secretion of proteins outside of the cell have been described in diderm bacteria (Figure 1.5).

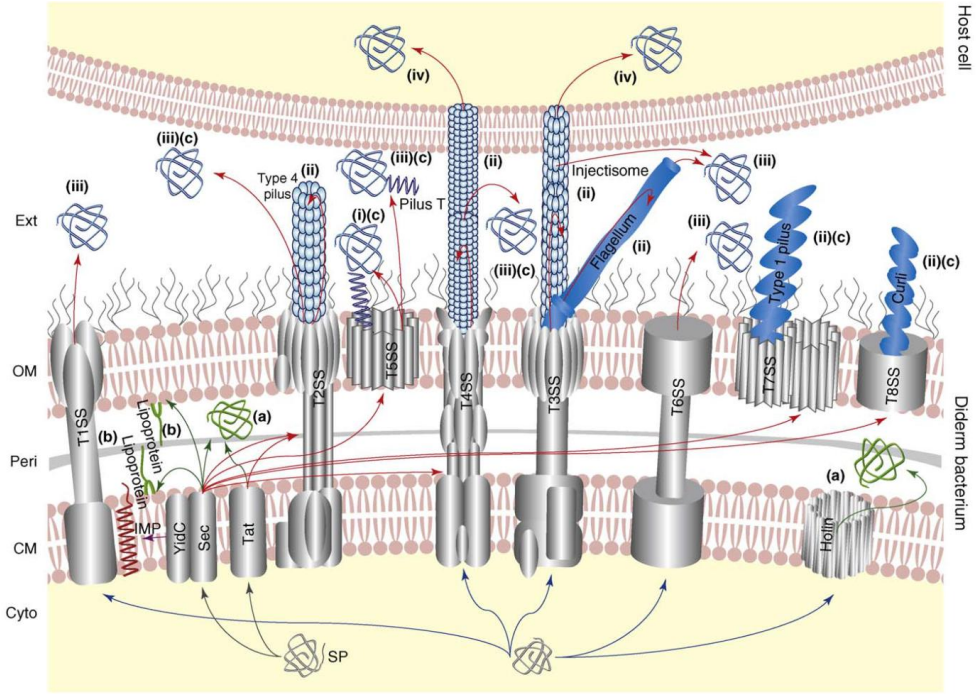

Figure 1.5 . Protein export systems of diderm (Gram-negative) bacteria.

Secreted proteins of diderm bacteria can be: (i) localised on the cell-surface when anchored to the OM *via* Omp8 (this is including some autotransporter proteins of T5SS); (ii) part of cell-surface appendages such as flagella (exported *via* T3bSS) or pili-like structures (type 4 pilus exported *via* T2SS, injectisome exported *via* T3aSS, pilus T exported *via* T4SS, type 1 pilus exported *via* the chaperone–usher pathway [T7SS] or curli exported *via* the extracellular nucleation-precipitation pathway [T8SS]); (iii) released extracellularly *via* T1SS – T6SS; or (iv) directly injected into a eukaryotic or bacterial host cell *via* T3aSS or T4SS. In diderm bacteria, exported proteins can be (a) released into the periplasm *via* systems for protein translocation across CM (i.e. Sec, Tat or holins); (b) lipoproteins anchored either to the CM or to the periplasmic side of the OM *via* Lol system; or (c) further subjected to a second translocation event across the OM *via* the T2SS, T4SS, T5SS, T7SS or T8SS. Black arrows indicate routes of

N-terminal signal peptide containing proteins targeted to CM, whereas blue arrows indicate routes used by proteins lacking such a signal peptide. Red arrows are related to secretion, violet arrows are related to integration of IMP and green arrows are related to export (which is not synonymous with secretion in diderm bacteria). Secreted and exported proteins are represented in blue and green, respectively. Abbreviations: Cyto, cytoplasm; CM, cytoplasmic membrane; IMP, integral membrane proteins; Peri, periplasm; OM, outer membrane; Ext, extracellular milieu. Figure taken from [226] with permission.

After Sec- or Tat- dependent translocation across the inner membrane, outer membrane-specific lipoproteins and unfolded β-barrel proteins are targeted to the outer membrane *via* the Lol pathway and β-barrel assembly machinery (BAM) pathway, respectively [232].

Secreted proteins targeted to the extracellular milieu, or to another cell, can be exported out of the cell directly, or by a two-step secretion process *via* type 1 – 6 secretion systems (T1SS – T6SS). In addition, the chaperone-usher system (CU or T7SS); the extracellular nucleation-precipitation mechanism (ENP or T8SS) system; as well as type IV (T4PS) and tight-adherence (Tad) piliation systems are dedicated to exporting different types of pili subunits across the outer membrane.

The direct (Sec pathway-independent) secretion in diderms is through a contiguous tunnel spanning two membranes and the periplasm *via* the T1SS, T3SS, T4SS and T6SS systems. Two-step secretion processes involves protein export to the periplasm by the Sec or, less often, the Tat pathways, followed by export across outer membrane *via* T2SS, T5SS, T7SS or T8SS systems. T1SS and T5SS are relatively simple systems involving few proteins, while T2SS, T3SS, T4SS and T6SS are complex structures composed of large number of subunits, and spanning the entire bacterial cell envelope.

Type 1 secretion systems are capable of delivering large, unfolded proteins (up to 800 kDa), such as adhesins, proteases and toxins into a host cell. The T1SS depends on two inner membrane proteins (ABC transporter and adaptor protein) and outer membrane component TolC, forming a barrel in the outer membrane and a 10 nm long periplasmic tunnel [262, 263]. The ABC (ATP-binding cassette) transporter-dependant systems are found in all kingdoms of life and are involved in the ATP-powered single-step secretion of various proteinaceous and non-proteinaceous substrates [264].

Type 2 secretion systems, a major secretion pathway in diderms, export substrates such as exotoxins, cellulases and other enzymes, across the outer membrane [265]. Type 2 secretion is facilitated by the assembly of the pseudo-pilus structure in the periplasm, which is thought to act as a piston that pushes secreted protein through the secretin complex in the outer membrane [266, 267].

Type 3 secretion systems export effector proteins of pathogenic bacteria directly from the bacterial cytoplasm into the cytoplasm of the eukaryotic host cell *via* a molecular syringe complex called the injectisome [268]. This system is structurally and evolutionary related to the

flagellar-assembly machinery [269] and the building blocks/components of flagellar filament (e.g. flagellin) are exported by the basal body, which is evolutionarily related to the T3SS.

Type 4 secretion systems are widely distributed among prokaryotes and consist of a multi-subunit pilus-like protein channel spanning the bacterial cell envelope. They transport proteins and protein-bound DNA into eukaryotic or bacterial cell in a process dependant on the direct contact with a target cell. T4SS are diverse and include: conjugation systems that mediate horizontal gene transfer, effector translocator systems that deliver proteins or other macromolecules to eukaryotic cells and DNA release/uptake systems for DNA import or export from or to the extracellular milieu [270, 271].

Type 5 secretion systems include autotransporters and 2-partner systems (Tps) and are involved in the secretion of variety of proteins conferring virulence traits, such as adhesion, autoaggregation, invasion, biofilm formation and cytotoxicity, in a two-step, Sec-dependant process [272].

Type 6 secretion systems have been relatively recently described [273] and have since been predicted in approximately one quarter of all sequenced Gram-negative bacteria. Some bacterial species harbour multiple and distinct T6SSs, thus making it probably the most widespread specialised secretion system in diderms [274]. T6SS consist of a syringe-like macromolecular nanomachine in the bacterial envelope that injects target eukaryotic and prokaryotic cells with effector proteins belonging to two different classes [275]. Evolved VgrG proteins are both integral component of secretion system and effector molecules implicated in binding and modification of actin filaments, cell adhesion and chitosan degradation, while the second class of effectors are classical toxins [274].

Type 7 secretion system (the CU system) is involved in the secretion and assembly of CU pili (e.g. type 1, P and S pili) and fimbrial structures (e.g. 'non-pilus' Afa/Dr adhesins) [276]. These linear multisubunit non-flagellar adhesive appendages on the outer membrane of Gram-negative bacteria are involved in attachment and biofilm formation, cell motility and host immune system evasion [277].

The extracellular nucleation-precipitation pathway (ENP or T8SS) assembles pilin CsgA (curlin) into curli pili, first described in enterophatogenic *Salmonella* spp. and *E. coli* and involved in biofilm formation and host cell adhesion and invasion [276].

Type IV piliation system (T4PS), a most common pilus biogenesis pathway, is involved in assembly of the type IV pili (T4P), common to many Gram-negative pathogens, but also observed in archaea and Gram-positive bacteria. For example, in *Ruminococcus albus* 20, it was shown that T4P mediate adhesion to cellulose [278]. Type IV pili are formed at the cytoplasmic membrane by homopolymerisation of pilin subunits and the pilus tip can contain adhesive subunits that attach to the receptor, enabling bacteria to move while the pilus retracts. Besides the 'typical' role of pili in biofilm formation and host cell adhesion, type IV pili also mediate

phage infection and transduction, DNA uptake during transformation, twitching motility and even extracellular electron transfer and some of these functions are enabled by the retracting ability of pili [276].

Based on phylogenetic analysis, some components of the T2SS (such as ATPases, type IV signal peptidases and secretins) are homologous to the type IV (T4PS) and tight-adherence (Tad) piliation systems and an alphanumerical classification into subtypes (T2SSa – c for classical T2SS, T4PS and Tad, respectively) has been proposed [237, 279]. On the other hand, the system for assembly and secretion of filamentous phage is distantly structurally related to T2SS and T4PS. A well conserved superfamily of secretins are key component involved in macromolecular transport in four evolutionarily ancient secretion systems (T4PS, T2SS, T3SS and filamentous phage assembly) [280, 281]. T4SS are related to conjugation systems, while T3SS are structurally similar to the flagella and Hpr pilus export apparatus [226].

1.4.3 Secretion and membrane targeting signals and their prediction

The first stage in the process of the export of secretome proteins is sorting and targeting of proteins to the cytoplasmic membrane, followed by membrane crossing and maturation/release of the translocated protein. The sorting process, through which proteins are routed to their specific subcellular compartments, is based on localisation information contained in a short amino acid sequence that acts as a protein sorting signal [282]. Blöbel and Sabatini pioneered the concept of these intrinsic signals (‘zip codes’) governing protein traffic, transport and localisation in the cell [283]. Discrimination between secreted and cytoplasmic proteins is based on the presence of membrane-targeting sequences, such as signal sequences and transmembrane α-helices that are recognised by distinct secretory pathway-associated molecular chaperones and are necessary for a correct targeting to the translocation pathway.

Signal sequences are N-terminal and usually cleavable peptides present in the secretory pre-proteins. Their function and overall structure are conserved in all domains of life; however, these peptides lack a primary sequence homology even within a species [284]. Several types of signal sequences have been described: type I (classic) signal sequence, type II (lipoprotein) signal sequence, Tat signal sequence, type IV (pseudopilin-like) signal sequence and bacteriocin/pheromone signal sequence (Figure 1.6). Based on hydrophobicity and charge, most signal sequences have a conserved overall tripartite organisation consisting of a hydrophobic core (h-domain), flanked by hydrophilic positively charged N-terminal region (n-domain) and a polar C-terminal region (c-domain) with cleavage/retention sites [284-286]. However, the signal sequences of proteins (such as type IV and bacterocin/pheromone signal sequences) targeted for

minor signal sequence pathways (SS2 and SS1 respectively), do not precisely follow such a structural outlay [285].

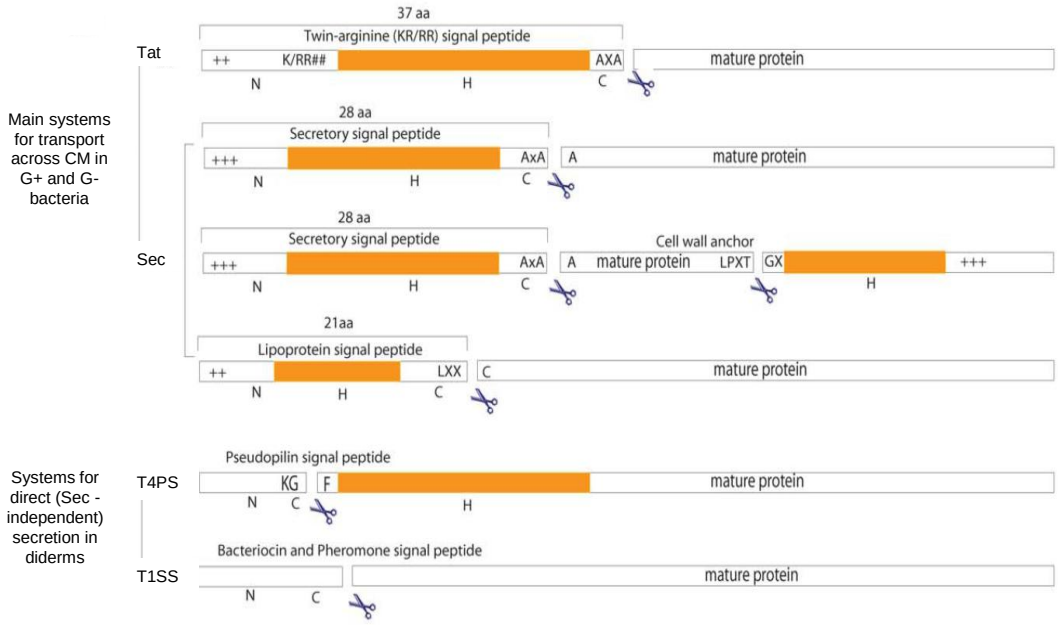

Figure 1.6 Schematic representation of the structure of common cytoplasmic membrane-targeting signals.

Most known signal sequences (ss) exported *via* Tat- and Sec-dependent pathways, main systems universally involved in protein translocation across the CM in both monoderm (G+) and diderm (G-) bacteria, have a conserved tripartite structure. These ss consist of: a positively charged 'N domain' (with positively charged amino acid residues such as lysine and/or arginine represented with +); a hydrophobic 'H domain' (represented in orange) and a polar 'C domain' (containing the signal peptidase cleavage signals). Consensus sequences of cleavage/retention signals are represented on the scheme: the AxAA type I SPase cleavage site; the L-x-x-C (lipobox) type II SPase cleavage site and the AxA Tat-substrate cleavage site. The LPxTG-type motif is a C-terminal sorting signal which is involved in the covalent attachment of proteins to the peptidoglycans of the cell wall. Unlike Tat- and Sec-dependent ss, ss of proteins targeted for direct (Sec pathway-independent) secretion in diderms, such as T4PS (also known as T2SSb) for type IV pilin assembly and T1SS for transport of pheromones/bacteriocins, do not follow the N-H-C structure. Abbreviations: CM, cytoplasmic membrane; G+, Gram-positive; G-, Gram-negative; ss, signal sequence. Figure adapted from [227] with permission.

Signal sequences are usually removed during or shortly after their translocation across the membrane by several types of membrane-associated signal peptidases (SPases), which also have a role in quality control and regulated turnover of exported proteins [287]. In bacteria, non-lipoprotein precursor proteins that are translocated through the Sec and Tat-pathways [288] are proteolytically processed by a 'general' type I signal peptidase (SPaseI) [289]. Processing of the lipoprotein signal sequences is performed by type II lipoprotein signal peptidase (SPaseII). The

lipoproteins are transported across the inner membrane in a Sec-dependant manner [290]. The Tat-dependent export of lipoproteins has only been demonstrated in streptomycetes [287, 291]. The prepilin signal peptidase (SPaseIV) is responsible for processing proteins containing type IV signal sequence, such as pilins and related pseudopilins, that have mainly Sec-dependant export across the inner membrane [292-294].

1.4.3.1 Type I signal sequences

Type I signal sequences are in the 18 – 30 amino acid residues length range. The average length of Sec-dependent signals in Gram-positive bacteria (32 amino acid residues) is higher than in Gram-negative bacteria (24 amino acid residues) [295]. The positively charged n-domain has a role in interaction of the signal sequence with the cytoplasmic membrane surface and has been implicated in the electrostatic interactions with membrane phospholipids. These interactions aid protein secretion across the membrane by orienting the helix with the N-terminus in the cytoplasm, which is required for signal peptidase cleavage and correct orientation of the whole secreted/membrane protein. Reduction in the number of positive charges in this region results in decreased rates of transport, but this phenomenon can sometimes be compensated for by an increased hydrophobicity of the h-domain [284, 296, 297]. A hydrophobic core of 10 – 15 amino acids has a role in signal sequence insertion into the lipid bilayer from the SecYEG translocon/channel and this region must maintain a minimal length and 'hydrophobic density' to function properly. It was shown that insertions of polar or charged amino acid residues into the h-region result in secretion defects [284]. Both the h- and n-regions are critical structural elements recognised by SRP or SecA components of Sec machinery. The binding affinity of SRP for signal sequences increases with the hydrophobicity of the h-region, whereas the interaction with SecA increases with the number of positive charges in the n-region [298]. The c-region contains the only conserved amino acid sequence motif in the Sec signal sequence, necessary for recognition and cleavage of the signal peptide from the mature secretory protein by SPaseI during or shortly after translocation [287]. Structural studies have shown that conserved small amino acids at the –3 and –1 positions relative to the cleavage site (AxAA) form a cleavage signal recognised by the SPase binding pocket [285, 286, 299]. In contrast to single SPase I (LepB) in Gram-negative bacteria, many Gram-positive bacteria contain more than one SPaseI that differ to some extent in their substrate preference and cannot completely substitute each other. The majority of SPases have a conserved Ser-Lys catalytic dyad, with the exception of several bacterial species belonging to Actinobacteria, Firmicutes (e.g. *Bacillus*, *Clostridium*, *Eubacterium* and *Ruminococcus*) and Mollicutes that employ a conventional Ser-His-Asp catalytic triad or a Ser-His catalytic dyad [287].

1.4.3.2 Type II signal sequences

Bacterial lipoproteins, a functionally diverse class of membrane anchored proteins, contain type II signal sequences with a relatively well conserved stretch of amino acids in their c-domains at the positions −3 to +1 relative to the cleavage site (L[A/S][G/A]C), called the lipobox [290]. Maturation of lipoprotein precursors takes place on the periplasmic side of the inner membrane and this post-translational modification enables their association with the cytoplasmic membrane of monoderms or inner or outer membrane of diderms. The sequence of events in lipoprotein modification in Gram-negative bacteria is: addition of diacylglycerol at a conserved +1 cysteine residue of the pre-lipoprotein by phosphatidylglycerol/prolipoprotein transferase (Lgt); cleavage of the signal peptide by lipoprotein-specific signal peptidase type II (Lsp), an aspartic acid protease that recognises modified cysteine residue within the lipobox, and N-acylation of this residue by phospholipid/apolipoprotein transacylase (Lnt) [290, 300]. This sequence is less clear in Gram-positive bacteria and it has been demonstrated that Lsp of *Listeria monocytogenes* is able to cleave non-lipidated lipoprotein precursors [301], suggesting that lipoprotein modification might not always occur in Gram-positive bacteria and that lipidation by Lgt is not a prerequisite for the activity of Lsp. In contrast, aminoacylation is essential for the Lol-dependent release in Gram-negative bacteria [302]. The differentiation between targeting to the inner and outer membrane in diderms is achieved by the presence or absence of the 'Lol avoidance signal', typically an aspartate residue at the +2 position, which inhibits the recognition of lipoproteins by LolCDE, thus causing their retention on the inner membrane [302, 303].

1.4.3.3 Tat signal sequences

Tat signal sequences also share a tripartite organisation, but are on average longer (26 – 58 amino acid residues) than Sec-dependant signals due to longer n- and h- regions (13 – 20 amino acid residues) [295]. In addition, these signal sequences have a less hydrophobic h-region compared to Sec signal sequences and a conserved pattern of amino acids located at the interface of the n- and h-regions, including an almost invariant twin-arginine motif, found 2 – 30 amino acid residues from the N-terminus in prokaryotes. The twin-arginine motif pattern was originally described as S/T-R-R-x-F-L-K in bacteria [304], but it is now thought, based on larger number of known bacterial Tat substrate sequences, that motif is simpler Z-R-R-x-Φ-Φ, where Z stands for any polar residue and Φ for hydrophobic amino acid residues [282, 299]. The signal peptidase recognition motif (A-X-A) is recognised by SPaseI. Site-directed

mutagenesis of residue in the RR pair [305], as well as naturally occurring Tat signal peptides lacking an arginine in twin pair [306], suggested that paired arginine amino acid residues are not an absolute requirement for the Tat signal peptide targeting into the Tat pathway. Moreover, introduction of RR-motif alone in the typical SecA-dependent signal peptide PelB showed not to be sufficient for Sec-independent transport, suggesting existence of further determinants of Tat-specificity [307]. The short polar c-region, apart from the SPaseI cleavage site (AxA), often contains a basic amino acid residue, which is not essential for targeting to Tat pathway [305], but was proposed to function as a 'Sec-avoidance' signal [308-310].

1.4.3.4 Type IV signal sequences

The type IV signal sequence is short (usually only 5 – 8 amino acids long, but longer signal sequences, up to 25 amino acids have been reported), positively charged and followed by a highly hydrophobic domain of approximately 20 amino acids [292, 311]. The precursors of type IV pilins and related pseudopilins are specifically processed by type 4 prepilin peptidases (TFPP), aspartic acid proteases also responsible for N-methylation of the phenylalanine at position +1 relative to the cleavage site required for prepilin maturation. The consensus sequence of the TFPP recognition motif is K-G-(F/M)-T-L-(I/L)-E and mutational analyses showed that glycine at the –1 position is required for complete processing of the preprotein [287, 312]. Unlike SPases I and II, TFPP cleave signal sequence from prepilins and pseudopilins at the cytoplasmic surface of bacterial membranes and cleavage takes place between the n-region of the signal peptide and the h-region of mature protein [313, 314].

1.4.3.5 Transmembrane α-helices

A transmembrane α-helix, a hydrophobic segment usually composed of 15 – 30 amino acids, is recognised by SRP and can also be used as membrane targeting signal on its own [315]. These helices serve as membrane anchors of integral membrane proteins. Besides (uncleaved) transmembrane domains, proteins known to be retained at the membrane can also have signal peptide-like sequences that act as uncleaved signal anchors are found in some N-anchored proteins. These membrane-targeting sequences have a typical tripartite structure, with highly similar n- and h domains, and even a SPase processing site [316]. No apparent sequence pattern is obvious when multiple sequences of experimentally validated N-anchored and secreted proteins are aligned [285].

1.4.3.6 Non-classical secretion

Some secreted bacterial proteins for which information on the secretory route is currently lacking or are exported *via* well-known routes, but lack an identifiable N-terminal or C-terminal targeting signal or other conserved motifs governing secretion process are termed 'non-classically' secreted proteins [231]. These proteins were first identified in eukaryotes (e.g. human interleukin 1β and thioredoxin, both of which are secreted without any known signal sequence) but were subsequently also observed in bacteria. However, non-classical secretion does not describe a single secretion pathway and the status of non-classically secreted proteins is not necessary permanent, since knowledge of new pathways and targeting motifs is constantly expanding. One example for this are mycobacterial ESX1-5 export systems, initially considered non-classical but now it is known that the substrates of all five systems possess a universal seven amino acid residues long C-terminal secretion signal, containing a conserved YxxxD/E motif, with additional signal(s) that are implicated in providing system specificity [317].

Substrates of T6SS and holin-mediated secretion systems, some SecA2-dependant substrates (e.g. superoxide-dismutase of *M. tuberculosis*) and some proteins released into extracellular environment via outer membrane vesicles (e.g. ClyA cytotoxin of *E. coli*) seem to currently fit the description of non-classically secreted proteins [232, 250, 318, 319]. In addition, some proteins with a well-established cytoplasmic localisation are also found extracellularly and they form a subset of multifunctional proteins able to perform multiple unrelated functions which are not partitioned into different protein domains and thus termed 'moonlighting proteins' [320]. In bacteria, multiple conserved proteins involved in the cell stress response or metabolic regulation (e.g. glycolytic and other metabolic enzymes and molecular chaperones) exhibit moonlighting activity implicated in virulence of several human pathogens (including *Streptococcus pyogenes, S. pneumoniae, Staphylococcus aureus, Mycobacterium tuberculosis* and *Helicobacter pylori*) [321].

1.4.4 Methods to study the secretome

Mining bacterial secretomes is important for a range of applications, including identification of novel enzymes, understanding of bacterial adhesion and their interactions with the environment, investigating pathogenic mechanisms, epitope mapping and identification of new vaccine candidates. Secretomes are traditionally studied *in vitro* using biochemical approaches and *in silico* using bioinformatic tools. Surface display screening methods and reporter fusion systems [322-327], as well as phage-display based systems [328-331], described

in more detail in sections 1.5.2 and 1.5.3, have also been used for screening, identifying and characterising secretome proteins.

Secretomes are studied *in vitro* using high-resolution separation (2D gel electrophoresis and/or liquid chromatography) of secreted or extracted membrane proteins, coupled with mass spectrometric methods for the identification of peptides and proteins in the sample [332]. Biochemical approaches for elucidating the secretome of a microorganism allow direct functional characterisation of identified proteins; however, they are very tedious and limited only to cultivable cells. Furthermore, construction of a proteome map of surface-associated and membrane proteins can be hindered by technical limitations of protein extraction from the membrane. In the absence of experimental data, a secretome can be deduced from a completely sequenced genome *in silico*, using bioinformatic tools for the prediction of secretome proteins based on their specific conserved features. The disadvantages of *in silico* secretome analysis is that it can be only applied to organisms with sequenced genomes; its accuracy depends on prediction algorithm performance, as well as on genomic annotation accuracy, and to improve the identification of secretome proteins, genomic predictions need to be integrated with transcriptomics and proteomics data [333]. Functional analysis of predicted secretomes requires many secretome proteins to be expressed and purified individually. In addition, as the secretome represents only a portion of the bacterial proteome, obtaining the complete genome sequences of multiple bacterial strains in order to identify their secretome-encoding genes is inefficient [234]. The task of predicting the metasecretomes of complex environmental microbial communities is even more challenging. This is due to current limitations in the identification of complete genes *via* sequence-based metagenomics approaches from low-coverage metagenomic assemblies derived through next-generation sequencing of complex environmental microbial communities, often containing closely related microbial species [22, 334].

Computational methods for secretome protein prediction are based on weight matrices, sequence alignment or machine learning algorithms, and can be roughly grouped into global tools for subcellular protein localisation prediction, and specialised tools for the prediction of signal sequences [333, 335]. More sophisticated machine learning algorithms, based on neural networks and decision trees, support vector machines, Bayesian networks, HMMs or their combination, are now more prominently used for discriminating secreted and non-secreted proteins [285]. During the training phase, typical signal and non-signal peptides are presented to the algorithm, and a classification model is subsequently built. Tools for signal sequence prediction such as SignalP [336], LipoP [337], TMHMM [338], PRED-LIPO [339], PRED-TAT [340], SecretomeP [259], and tools for subcellular protein localisation prediction, such as PSORTb [341] or TargetP [342] belong in this class.

In a benchmark study by Choo *et al*. [343], 13 commonly used prediction tools, belonging to all three types of prediction computational methods, were evaluated for their

accuracy, specificity and sensitivity in discriminating between signal and non-signal peptides in test datasets containing eukaryotic, as well as combinations of eukaryotic and bacterial sequences. Machine learning tools outperformed the alignment-based and matrix-based tools, since these heavily depend on the regular tuning of their model parameters to reflect updates in the sequence database, and SignalP was the leading tool. Most tools were able to successfully distinguish secretory from non-secretory proteins, but lacked accuracy in determining the location of the signal peptide cleavage site.

In prokaryotes, classically secreted proteins can be predicted based on recognition of the tripartite structure of their N-terminal, cleavable signal peptides, and conserved amino acid residues at the –3 and –1 positions relative to the cleavage sites. In addition to these, the lipobox of type II signal sequence and the Tat motif in Tat signal sequences are highly amenable to identification by bioinformatic tools, while transmembrane α-helices can be identified based on their hydrophobicity [233, 335]. Recognition of the SPaseIV cleavage motif is not sufficient for the accurate detection of type IV signal sequences, since these have no tripartite structure like other Sec-dependant substrates. It was demonstrated that the specificity of searches for type IV pilin-like proteins may be enhanced by including additional search requirements, such as the presence of 14 sequential uncharged amino acid residues immediately after the cleavage motif or presence of a single transmembrane helix within 50 amino acid residues of the N-terminus, since true pilins contain only one transmembrane helix, typically close to the cleavage motif [314].

Cleavable N-terminal signal peptides of secreted proteins are readily distinguishable from the longer hydrophobic N-terminal transmembrane helices of transmembrane proteins. In contrast, their discrimination from uncleaved N-terminal signal anchors, tethering some of these Sec-exported proteins to the membrane, is often problematic [285, 299, 333]. However, tools such as SignalP 4.0 are trying to overcome this challenge by combining predictions of transmembrane protein topology with signal sequence identification.

Optimal sensitivity and specificity for *in silico* identification of secretome proteins is achieved by combining multiple prediction tools, while the accuracy of the prediction can be improved by further examination of proteins classified as non-secreted using homology-based searches with BLAST to determine whether these proteins are homologous to known secretory proteins of related organisms [333].

1.5 Phage display

1.5.1 The life cycle of Ff phage used for phage display

In 1985, George Smith introduced the phage display technique [344]. After inserting a foreign DNA fragment between the signal sequence and the mature portion of M13 filamentous phage gene III, encoding one of the virion's structural proteins, he demonstrated that fusion proteins could be displayed in an immunologically accessible form on the infectious particles. These virions could be also affinity-purified from a large background of wt phages using an antibody directed against the foreign sequence. Genetically highly similar group of filamentous phage, M13, f1 and fd, that infect *E. coli via* its F pilus and reproduce without killing their bacterial host, termed F-pilus specific (Ff) filamentous phages, have since been widely used in phage display technology and nanotechnology [345].

Like other filamentous bacteriophage of the genus *Inovirus*, Ff are in the form of long and thin filaments containing a circular single stranded DNA (ssDNA) genome [346]. The Ff phage genome, 6.4 Kb in size, encodes 11 proteins (pI – pXI) essential for genome replication, its packaging into virions and virion release in a secretion-like process [347, 348]. The Ff virions are composed of approximately 2700 copies of major coat protein pVIII, tiled in a helical arrangement, and 5 copies of each of the four minor coat proteins, arranged in pairs (pVII-pIX and pIII-pVI) and forming a 'cap' at each end of the filament (Figure 1.7).

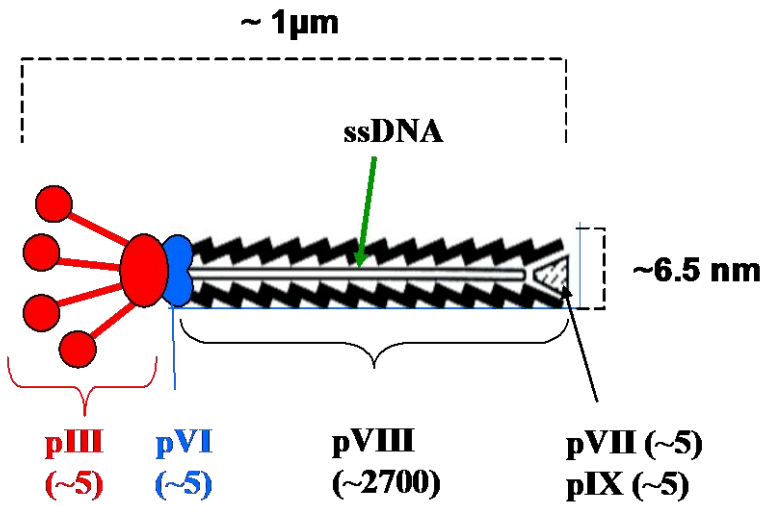

Figure 1.7 Schematic representation of Ff filamentous phage.

Major virion protein pVIII and minor virion proteins pIII, pVI, pVII, and pIX are represented and approximate number of copies of each virion protein is indicated. Figure taken from [349] with permission.

The phage protein pIII mediates infection of the host through binding of its two N-terminal domains (N2 and N1) to the primary and secondary Ff phage receptors of *E. coli* [350]. Interaction between the N2 domain of pIII and the tip of the F pilus (primary receptor) causes a structural rearrangement between the two N domains and release of the N1 domain, exposing its TolA-binding site. This site interacts with periplasmic domain of the TolA protein, a component of the cytoplasmic membrane complex TolQRA (secondary receptor). Interactions between phage pIII and host receptors result in integration of virion coat proteins into the inner membrane and entry of the phage ssDNA into the cytoplasm of the *E. coli* host. A negative (–) strand is synthesised by the host RNA polymerase using positive strand (+) phage ssDNA (also called the infective form) as a template. The resulting dsDNA form of Ff phage genome is referred to as the replicative form (RF). The RF serves as a template for replication of the phage (+) strand, as well as for transcription of phage genes. Early in the phage infection, the newly synthesised (+) strands are used as a template to create more copies of RF, thus increasing the synthesis of phage encoded proteins. Newly synthesised phage coat proteins (pVIII, pIII, pVI, pVII and pIX) and proteins forming the phage assembly/export complex (pI, pXI and pIV) are translocated *via* the Sec-dependant pathway and integrated into the host membranes, while proteins involved in replication (pII, pX), and formation of the packaging substrate (pV), remain in the cytoplasm of the host. In the host cell, the phage ssDNA genome replicates episomally by a rolling-circle mechanism, one strand at a time. The pII protein nicks the supercoiled dsDNA at the (+) strand origin, creating a primer for the rolling circle synthesis of (+) strand and after one round of replication, the newly synthesised (+) strand is cleaved off by pII and its ends are ligated. Later in the phage infection, when the concentration of the phage proteins has increased, (+) strands are coated with dimers of the ssDNA-binding protein pV in the cytoplasm. The ssDNA-pV complex represents a packaging substrate for the phage assembly and is targeted to the phage assembly/export complex [345].

The Ff phage assembly starting from the pVII-pIX ‘cap’ [351], and its export from the cell resembles the T2SS secretion in diderm bacteria. A periplasm-spanning transport complex involved in phage assembly and phage export consists of pI and pXI proteins, forming an assembly complex in the cytoplasmic membrane and 14 pIV monomers, forming a large gated pore in the outer membrane [352, 353]. The ssDNA-pV complexes are brought to the phage assembly/export complex in the membrane, where the packaging signal, the only exposed segment of ssDNA in the form of hairpin loop, interacts with pVII, pIX and pI. As the ssDNA genome passes through the phage assembly/export membrane complex, pV dissociates and is replaced by pVIII, forming a filamentous tube around the phage genome. When the phage ssDNA is completely coated with pVIII, pVI and pIII are added to the end of phage, resulting in the release of newly assembled phage from the host membrane. If either of the two distal ‘cap’ proteins are absent, phage filaments continue to elongate while remaining tethered to the

membrane, resulting in pili-like structures protruding from the surface of infected cells [354, 355]. The life cycle of Ff filamentous phage is represented in Figure 1.8.

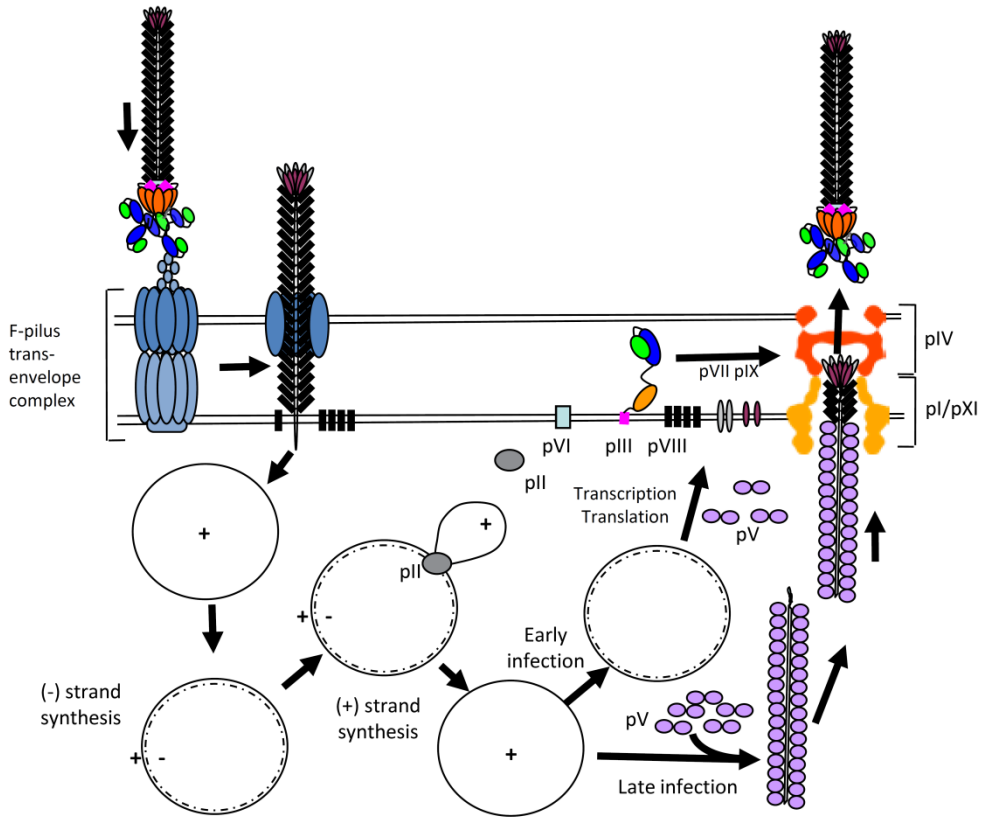

Figure 1.8 Life cycle of filamentous bacteriophage in *Escherichia coli*.

Upon infection of *E. coli* by binding of the tip of the phage to the tip of the F pilus (represented in the left top corner of the scheme), the major phage coat protein pVIII integrates into the inner membrane, while the phage ssDNA (+) strand enters into the cell of the host. In the early stage of infection, the minus (–) strand is synthesised by host proteins, starting from the (+) strand origin of replication, resulting in the dsDNA. In the second stage of infection, phage protein II (pII) nicks the supercoiled dsDNA at the (+) strand origin, creating a primer for the rolling circle synthesis of the (+) strand and after one round, the product is cleaved off by pII, and the ends ligated. In the early infection, the (+) strand is recycled for replication to create RF that is used as a template for transcription, followed by the translation into phage proteins. In the late infection, the (+) ssDNA genome is coated by pV and this complex is the substrate for packaging into the virions. Newly synthesised pI/pXI phage proteins form multimers in the inner membrane while a pIV multimer forms a large gated channel in the outer membrane and together they form the machinery for assembly and export of virions. An exposed hairpin loop at the tip of pV/ssDNA packaging substrate interacts with pI, and then with distal phage tip proteins pVII and pIX to initiate virion assembly in the periplasm. As ssDNA is extruded through the *E. coli* envelope, pVIII packs on around it and in the end pIII/pVI release phage from membranes. Figure taken from [345] with permission.

1.5.2 Principles and applications of phage display

Phage display is a molecular technique for the expression and display of peptides or proteins encoded by a population of foreign variant DNA sequences on the phage surface. Peptides and polypeptides with specific binding properties can be selected from a vast number of displayed variants by using a high-throughput screening process based on binding affinity [356]. The physical link between the phenotype and genotype of the protein displayed on the phage surface; high phage replication capacity and affinity selection through rounds of selective enrichment and amplification are the major underlying principles of phage display technology. This technique has been used to display libraries of random peptides, protein domains or whole proteins [357, 358], with applications including molecular evolution, convergent and directed evolution of peptides and proteins [359-362], cDNA expression screening [363], analysis of protein-ligand interactions [356], antibody phage display [364] and synthesis and assembly of nanowires from magnetic, semiconducting [365] or electrode materials [366].

The two key elements of phage display are the libraries of DNA variants encoding peptides or proteins, and the phage vehicles on which these sequences are expressed [356]. Ff filamentous phage (M13, fd, f1) are the most commonly used vehicles for phage display, although T4 [367], T7 [368, 369] and λ [370, 371] phage display systems have also been developed. Ff phage are excellent cloning vehicles, because their size is not constrained by the size of the cloned DNA [356]. Furthermore, they are resistant to a wide range of pH (2 – 10), temperatures and detergents and have high productivity, giving rise to titres up to 10^{13} virions per mL of host culture. On the other hand, the main limitation in the use of the non-lytic Ff phage vehicle is that all of the phage coat components are exported *via* the Sec-dependant pathway through the bacterial inner membrane prior to their assembly into mature phage particles. As a result, only proteins capable of such export may be displayed, which depends not only on presence of adequate membrane-targeting signals, but also their sequence, length and folding characteristics [369].

A display of polypeptides is achieved through the translational fusion of a population of variant DNA sequences encoding foreign peptides or proteins, to genes encoding any of the five types of Ff filamentous phage coat proteins, located either in the phage genome or on a phagemid vector [356, 372]. The different coat proteins have relative merits as fusion partners, regarding the number of fusion proteins displayed per phage particle, and effects of the expressed fusion proteins on phage viability and stability [356, 373]. A minor Ff phage coat protein pIII (present in 5 copies per virion), and the major coat protein, pVIII (present in 2,700 copies per virion), are the most frequently used display platforms [347]. pVIII is the fusion partner of choice when a large number of short peptides are to be displayed, while pIII is a suitable fusion partner for smaller numbers of larger proteins [356].

There are two general types of phage display systems: phage-based and phagemid-based, which can be distinguished based on the vector system used. It was observed that the direct approach of cloning libraries of variant DNA sequences as a fusion with the chosen coat protein gene within the phage genome, although advantageous regarding simplicity and number of displayed fusion proteins, could lead to compromised phage viability, stability and infectivity. This is because no wt version of coat protein, which is used as the phage display platform, is present, leading to recombination that eliminates inserts in order to improve the efficiency of phage assembly or infection. To overcome these problems, two hybrid systems for displaying proteins on the pIII and pVIII platforms were created [356, 372, 373].

The first system, known as type 33 or 88, is phage-based. In this system, the phage genome carries two copies of gene III (*gIII*) or gene VIII (*gVIII*) and foreign DNA is inserted in one of the copies, yielding phages with a mixture of wt and fusion proteins on the same phage particle. The second system, known as type 3+3 or 8+8, is phagemid-based (Figure 1.9).

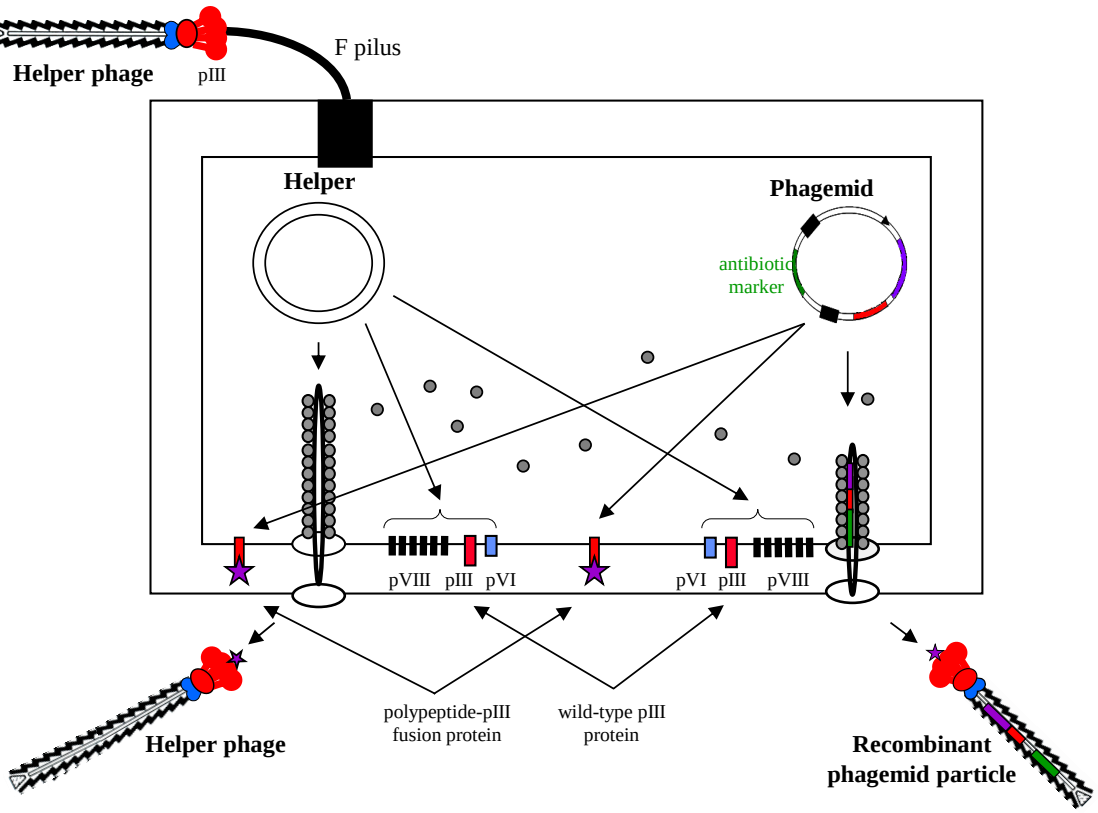

Figure 1.9 Schematic drawing of 3+3 phagemid based phage display system.

The helper phage genome encodes all proteins necessary for the replication and packaging of recombinant phagemids into virions. In the presence of the helper phage, the phagemid is packaged into recombinant phagemid particles. Virions produced in this system may incorporate either wild-type pIII derived from the helper phage (red lollipop-like structure) or the polypeptide-pIII fusion protein derived from the recombinant phagemid vector (red lollipop-like structure with a purple star). Adapted from [349] with permission.

Sequences encoding fusion proteins are carried by phagemids (plasmids bearing an Ff phage origin of replication and a packaging signal). Phagemids are replicated as plasmids in a host that is not infected with Ff phage. The production of phagemid-containing phage particles (recombinant phagemid particles) can only be achieved after infection of the phagemid-carrying host cells with a helper phage, which provides the replication and packaging machinery necessary for the formation of the phage particles. This procedure is known as 'phage-rescue'. Typically, the helper phage has a defective phage origin of replication and/or a defective packaging signal. This modification leads to production of phagemid particles at a 10- to 100-fold excess over the helper phage, because phagemid ssDNA is preferentially replicated and packaged over the helper phage genome.

In the phagemid-based phage display platform, the average number of displayed fusion proteins is reduced due to competition for incorporation between the wt and fusion coat proteins. The valence of recombinant fusion can be increased by the use of modified helper phage, such as M13Δ*gIII* and R408Δ*gIII*, that have a complete deletion of the gene encoding the wt pIII protein [374]. An additional refinement of phage display vectors is the insertion of peptide tags (c-myc or E-tag) followed by an *amber* stop codon (TAG) downstream of the insert (please see section 1.5.3 and Figure 1.11 for an example). This enables the display of the fusion protein in the *amber* suppressor strain and production and purification of a soluble foreign protein when phages are propagated in an appropriate non-suppressing strain.

Phage display enables expression of libraries of variant nucleotide sequences into a library of variant peptides or proteins displayed on the phage surface. These phage display libraries can be subsequently screened for desirable binding properties using an affinity selection procedure (Figure 1.10), usually referred to as panning or bio-panning [356].

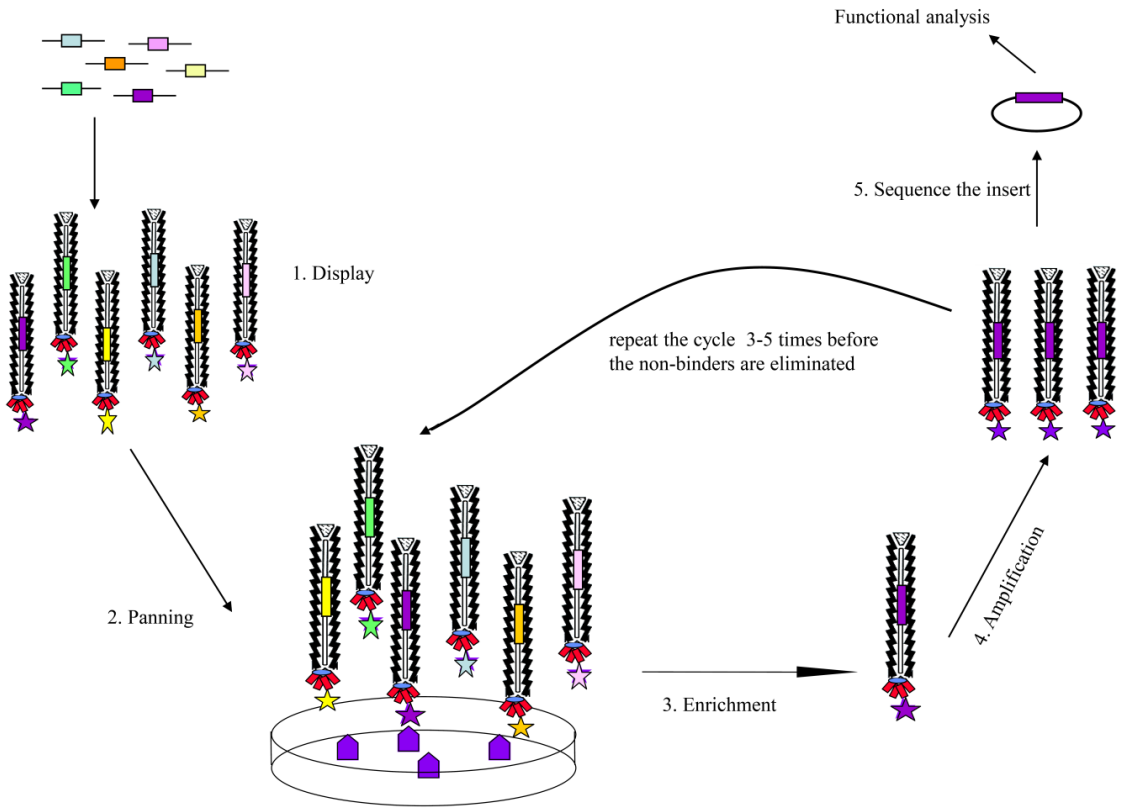

Figure 1.10 Phage display library panning against an immobilised target.

1. A library of variant DNA sequences is constructed using phage- or phagemid- based phage display systems. This results in the display of variant (poly)peptides (depicted as stars of different colour) on the surface of the recombinant phagemid particles. **2**. The phage library is exposed to immobilised target molecules (purple pentagons) to capture specific binders (i.e. the purple star displaying phage). **3**. The non-binding virions are washed off, although some non-specific ('background') binding may occur. Bound phagemid particles are eluted by conditions that disrupt interactions between displayed (poly)peptide and target. **4**. Eluted virions are amplified in a suitable bacterial host and the resulting amplified phage population is greatly enriched for recombinant phage clones displaying (poly)peptide variants with the capacity to bind the target molecules. An additional 3 – 5 rounds of panning (steps 2 – 4) result in a clonal phagemid population that binds the target. 5. Captured putative binders can be identified by sequencing and functionally analysed. Figure taken from [349] with permission.

During panning, recombinant phages are exposed to ligands of interest, usually immobilised on a solid support, in order to selectively capture a phage displaying the binding peptide. Phage display systems are highly flexible, and affinity selection can be performed, besides proteins, against other immobilised inorganic [375-377] and organic molecules [168, 378, 379] or cells *in vitro* [380, 381], as well as *in vivo* [382-385]. Through successive rounds of binding, washing, elution and amplification, the original highly diverse phage population (phage libraries displaying up to 10^{10} peptide or protein variants [372]) is rapidly enriched for the phage library clones with specificity for the binding target molecule. Due to some unspecific binding

occurring between phage and components of system used for ligand immobilisation, at least three rounds of panning (typically 3 – 5) are needed to eliminate 'background' non-specific binding and enrich monoclonal recombinant phage populations with the desired ligand specificity. However, if several binding variants are displayed in the library, variants capable of high-affinity interactions with the target will outcompete low-affinity binders during the course of panning [345]. Displayed proteins can be identified by sequencing the inserted DNA encoding the displayed peptide, and can be easily purified and subsequently functionally analysed. Sequencing of individual candidate binders enriched through panning to determine the corresponding peptide sequence of encoded putative target-binding peptides is one of the most labour-intensive in DNA panning. This is also rate-limiting step in discovery of different binding variants.

Next generation sequencing technologies have been increasingly applied for in depth identification of the diversity of phage binding variants enriched after one or two rounds of panning on ligand of interest [386-390]. This allows, besides high-throughput identification of a numerous binding variants, overcoming a problem of competition between high-affinity and low-affinity binders. Dias-Neto and colleagues (2009) [386] adapted pyrosequencing for deep-sequencing of amplicons derived from phage ssDNA using primers flanking the insert-*gIII* fusion. When the GC content, codon usage and amino acid frequencies, the frequency of homopolymers as well as the overlap of peptides observed were compared between phage sequences derived using Sanger sequencing and pyrosequencing technologies, the authors did not observe a significant difference between the two sequencing approaches. DiNiro *et al.* (2010) [387] combined two rounds of panning of a cDNA phage display library (to achieve an optimal balance between the high numbers of positive clones and broad diversity) and pyrosequencing, which enabled at least two orders of magnitude increase in the number of identified selected clones compared to traditional affinity screening. The authors identified a 'landscape' of binding variants from the phage display library and, based on the ranking of the most frequently selected clones detected through the high-throughput sequencing, they estimated that at least 1000 clones need to be picked and analysed using traditional screening approach in order to pick the top five most frequent clones after two rounds of selection with 99% probability.

1.5.3 Overview of the secretome-selective phage display system

Mining of bacterial secretomes is of importance for a range of applications. However, as described in section 1.4.4, traditional approaches to analyse the secretome can be inefficient.

In 2007, a new system for the direct selection, expression and display of the bacterial secretome was published by Jankovic *et al*. [234].

This approach allows secretome-encoding genes (as well as the corresponding displayed secretome proteins) to be specifically selected before sequence and functional analysis. It also demonstrated superior performance in the enrichment and display of secretome proteins from Gram-positive bacteria in an *E. coli* host, compared to other phage-display based systems. The bacterial secretome-selective phage display system depends on two components: a secretome-selective phagemid vector (pDJ01) that contains a pIII cloning cassette without a signal sequence [234], and a helper phage (VCSM13d3) with the entire pIII coding sequence deleted [374] (Figure 1.11).

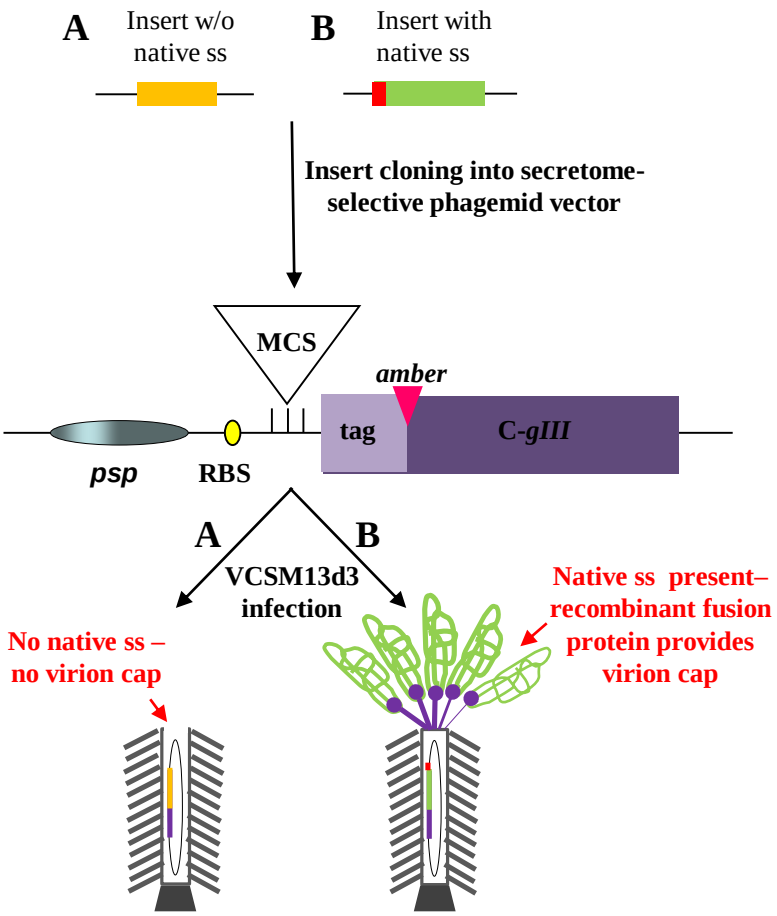

Figure 1.11 Schematic overview of the bacterial secretome-selective phage display system.

The display cassette of the secretome-selective phagemid vector contains: filamentous phage infection-inducible promoter (*psp*), ribosome-binding site (RBS), multiple cloning site (MCS), the sequence encoding common peptide tag c-myc (tag) followed by a single *amber* stop codon (*amber*) and the sequence encoding C-terminal domain of phage protein pIII (C-gIII), which is

used as a display platform. This vector does not have a signal sequence (ss). Infection of *E. coli* cells harbouring recombinant secretome-selective vectors with cloned inserts without (A) or with (B) native ss with a helper phage (VCSM13d3) with the entire pIII coding sequence deleted results in generation of incomplete recombinant virions without the pIII cap at one end (A) or complete virions with recombinant fusion proteins providing a cap (B).

The display cassette of pDJ01 contains the filamentous phage infection-inducible *psp* (phage-shock protein) promoter, that has a very low level of basal expression and decreases the potential toxic effects of foreign protein-pIII fusions on the host cell [391, 392]. The promoter is followed by the ribosome-binding site, multiple cloning site, the sequence encoding a commonly used c-myc peptide tag followed by a single *amber* stop codon and the sequence encoding the C-terminal domain of phage protein pIII, which is used as a display platform in this system. Incorporation of pIII in the virion is necessary for forming a cap on one end of the newly assembled phagemid particles, rendering virions resistant to the detergent sarcosyl. This vector also contains a chloramphenicol resistance gene and a plasmid (ColE1) origin of replication, which enables plasmid replication to medium-copy numbers, and ameliorates the potential toxicity of inserts to the host. The phage intergenic sequence in the vector, containing the f1 origin of replication and packaging signal, enables ssDNA replication and packaging into recombinant phagemid particles (PPs) in the presence of helper phage. In contrast to other phage display vectors, pDJ01 does not have a signal sequence. Thus, when the helper phage VCSM13d3 is used to assemble the PPs, library inserts without native signal sequence produce only defective virions that are sensitive to the sarcosyl detergent. In contrast, inserts that contain a signal sequence, or other motifs that can mediate targeting of the N-terminus of the fusion to the *E. coli* membrane, are assembled into the detergent-resistant virions [234].

The *E. coli* host strain TG1 (*amber*-suppressor *supE* strain), used for library construction and screening, contains a *supE* mutation that allows *amber* stop codon to be read as glutamic acid. In TG1 strain, ORFs cloned in frame with sequences downstream of MCS are translated through the TAG stop codon into the in-frame gIII-coding sequence and displayed as a pIII-fusion protein on the surface of the phage. Switching a host strain from TG1 to a non-suppressing *E. coli* strain allows the c-myc tagged soluble secretome proteins (without pIII moiety) to be expressed and easily detected and purified using the c-myc peptide (affinity) tag [234].

The phage display system for the direct selection and display of secretome proteins was successfully applied at a single genome scale to *Lactobacillus rhamnosus* and *Mycobacterium tuberculosis*. Moreover, sequence analysis and affinity screenings of the resulting phage display secretome libraries allowed characterisation of surface proteins with functions of interest [393-395]. This technology coupled with the high-throughput sequencing has a potential to mine

entire microbial communities, where cultivation-independent methods are required to enable discovery and functional characterisation of their products.

1.6 Aims of the project

The aims of my PhD project are: (i) to apply secretome-selective phage display to the bovine rumen plant-adherent microbial community to mine for novel fibre-degrading and binding activities. This was achieved by (ii) analysing a metasecretome phage display library *via* pyrosequencing, to focus sequencing efforts on the metasecretome-encoding component of the metagenome, and (iii) functionally identifying activities associated with the binding of complex polysaccharide substrates in the metasecretome phage display library, through standard phage display affinity selection.

My research questions are:

- Can metasecretome display enrich for and efficiently display functionally and taxonomically diverse low-abundance surface and secreted proteins encoded by rumen microbial communities?
- Can the next-generation sequencing of rumen microbial metasecretome libraries enable such proteins to be identified at a higher frequency relative to the standard shotgun metagenome sequencing?
- What is the diversity of fibre degrading and binding proteins in the rumen plant-adherent microbial fraction uncovered by metasecretome phage display, and how does this diversity compare to those found through other function- and sequence- based metagenomic approaches? What proportion of secreted CAZymes and other components involved in the ruminal fibre degradation identified through metasecretome phage display is novel?
- Can the standard phage display affinity selection procedures be used at a metagenomic scale to identify fibre-binding components of the rumen microbial metasecretome (such as bacterial cellulosomes and other CAZymes)?

Chapter 2. Materials and Methods

2.1 Materials

2.1.1 Laboratory chemicals and enzymes

General laboratory chemicals were manufactured by Sigma-Aldrich (USA), BDH (USA), Thermo Fisher Scientific (USA), Life Technologies (USA), Merck (USA) and Scharlau (Spain). Microcrystalline cellulose and insoluble wheat arabinoxylan were manufactured by Sigma-Aldrich (USA) and Megazyme (Ireland), respectively. BD Difco™ 2×YT broth formulation and BD Difco™ Agar were manufactured by Becton-Dickinson (USA). Bacteriological agar was obtained from Oxoid (UK). All antibiotics were purchased from Sigma-Aldrich (USA).

Restriction endonucleases, DNA polymerases and DNA modifying enzymes were obtained from the following companies: Life Technologies (USA); Roche Applied Science (Germany); Takara Bio Inc. (Japan); Affimetryx (USA); Promega (USA) and Thermo Fisher Scientific (USA). SeaPlaque® GTG® Agarose and AgarACE™ agarase were manufactured by Cambrex Bio Science Rockland, Inc. (USA) and Promega (USA), respectively. All custom oligonucleotides were synthesised by Integrated DNA Technologies (USA).

2.1.2 Buffers, solutions and media

2.1.2.1 Standard buffers and solutions

Standard buffers, solutions, media and antibiotic stocks were prepared as described in Sambrook *et al.* [396] and sterilised either by autoclaving at 121°C for 20 min, or by microfiltration through a sterile Millex®-GP Syringe Filter Units (EMD Millipore, USA) containing a fast flow and low protein binding Millipore Express (PES) membrane with a 0.22 µm pore size. Buffers, solutions and media were stored at room temperature (RT) unless stated otherwise. Antibiotic stock solutions were stored at –20°C.

2.1.2.2 DNA-free water

Ultrapure water was collected from a Milli-Q Integral Water Purification System (EMD Millipore, USA) and sterilised by autoclaving, followed by microfiltration (see section 2.1.2.1). The water was then irradiated with UV light (254 nm, 6 W) for 8 h.

2.1.2.3 OrangeG loading dye

OrangeG loading and tracking dye for nucleic acid gel electrophoresis was prepared as a 6× stock containing 20 mM Tris-HCl (pH 7.6), 0.4% 1-phenylazo-2-naphthol-6,8-disulfonic acid disodium salt (orangeG sodium salt), 30% glycerol, 60 mM ethylenediaminetetraacetic acid (EDTA; pH 8.0) and 0.1 % sodium dodecyl sulfate (SDS). Dye stocks were stored at 4°C short-term or at –20°C long-term.

2.1.2.4 Buffers used for rumen content fractionation

RM02 buffer contained 0.15% (w/v) potassium chloride, 0.14% (w/v) potassium di-hydrogen phosphate and 0.06% (w/v) ammonium sulphate. Dissociation buffer contained 0.1% (w/v) sodium pyrophosphate and 0.5% Tween 20 (pH 6.8).

2.1.2.5 Buffers used for extraction and shearing of metagenomic DNA

The lysis buffer contained 1% (w/v) sodium lauroyl sarcosinate (sarkosyl), 0.2% (w/v) sodium deoxycholate, 10 mM Tris-HCl (pH 8.0), 50 mM sodium chloride (NaCl), 0.1 M EDTA and 1 mg mL^{-1} lysozyme. The EDTA/Sarkosyl/Protease (ESP) buffer contained 20 mM EDTA, 0.5% (w/v) sarkosyl and 0.013 AU protease (Qiagen, The Netherlands). Both buffers were prepared fresh before use.

The shearing buffer contained 55 mM Tris-HCl (pH 8.0), 15 mM magnesium chloride ($MgCl_2$) and 25% glycerol.

2.1.2.6 Phage concentration, purification and disassembly solutions/buffers

The PEG/NaCl phage precipitation solution was prepared as a 6× stock, containing 30% (w/v) polyethylene glycol (PEG) 8000 and 3 M NaCl and sterilised by autoclaving. The loading

buffer for disassembled virion gel electrophoresis contained 2.5× BlueJuice™ Gel Loading Buffer (Life Technologies, USA) and 1% (w/v) SDS.

2.1.2.7 Liquid and solid media

Double Yeast extract Tryptone (2×YT) media was prepared from the BD Difco™ 2×YT broth formulation according to manufacturer's protocol. 2×YT agar also contained 1.5% (w/v) bacteriological agar (Oxoid, UK), while 2×YT soft (top) agar contained 0.6% (w/v) BD Difco™ agar unless indicated otherwise. When required, antibiotics were added to 2×YT broth at the following concentrations: 25 µg mL^{-1} chloramphenicol (Cm), 60 µg mL^{-1} ampicillin (Amp) and 30 µg mL^{-1} kanamycin (Kan) to make 2×YTCm25, 2×YTAmp60 and 2×YTKan30 media or plates, respectively.

Double-layer Cm plates were prepared by overlaying 2×YTCm25 plates (prepared from 21 mL 2×YT agar supplemented with antibiotic in a Petri dish) with 9 mL of chloramphenicol-free 2×YT agar shortly before use.

M9 minimal agar, prepared as described in [396], was supplemented with 0.4% (w/v) glucose and 0.01% thiamine. SOB and SOC media were prepared as described in [396].

2.1.3 Bacterial strains, plasmids and phage

Bacterial strains, plasmids and phage used in this study are listed in Tables 2.1 and 2.2.

Table 2.1 *Escherichia coli* strains used in this study.

***E. coli* strain**	**Genotype**	**Reference**
TG1	*supE thi-1 Δ(lac-proAB) Δ(mcrB-hsdSM)5* (r_K^- m_K^-) [F' *traD36 proAB lacI*^{q}*ZΔM15*]	Stratagene, USA
K1976	TG1//pJARA112	[374]
K1931	TG1//pDJ01	[349]
MS pilot 1 – 90	TG1 transformed with secretome-selected single-stranded pDJ01-derived recombinant phagemids containing metagenomic DNA inserts from the fibre-adherent fraction of the rumen microbiome (pilot metasecretome-enriched plasmid library clones)	This study
RAC 1 – 60 RAC 101 – 200 RAC 261 – 340	TG1 transformed with secretome-selected pDJ01-derived recombinant phagemids containing metagenomic DNA inserts obtained from the fibre-adherent fraction of the rumen microbiome, selected after panning on regenerated amorphous cellulose.	This study

E. coli strain	Genotype	Reference
AXYL 61 – 100 AXYL 201 – 260 AXYL 341 – 380	TG1 transformed with secretome-selected pDJ01-derived recombinant phagemids containing metagenomic DNA inserts derived from the fibre-adherent fraction of the rumen microbiome, selected after panning on arabinoxylan.	This study

Table 2.2 Plasmids and phage used in this study.

Plasmid and phage	Description/genotype	Reference	Antibiotic marker [a]
pJARA112	Contains *gIII* under the control of the *psp* (phage shock protein) promoter	[374]	Amp
pDJ01	Phagemid phage display vector for selective expression of bacterial secretome proteins as pIII fusions	[234]	Cm
MS 1 – 90	pDJ01-derived recombinant phagemids isolated from pilot metasecretome-enriched plasmid library clones	This study	Cm
VCSM13	interference-resistant helper phage derived from M13KO7 helper phage	Stratagene, USA	Kan
VCSM13d3	VCSM13 *ΔgIII*	[374]	Kan

[a] Antibiotic marker, antibiotic resistance marker; Amp, Ampicillin; Cm, Chloramphenicol; Kan, Kanamycin.

2.1.4 Oligonucleotides

Oligonucleotides (primers) used for sequencing and PCR reactions are listed in Table 2.3.

Table 2.3 Oligonucleotide primers used in this study.

Primer name	Sequence (5' – 3')	Application
pspF03	ATGTTGCTGTTGATTCTTCA	sequencing
pspR03	TGCCTTTAGCGTCAGACTGTAGC	sequencing
PCRF2	GCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCA	PCR
PCRR2	GGCGACATTCAACCGATTGAGGGAGGGAAGGT	PCR

2.1.5 Bioinformatic resources and software

Bioinformatic resources and software used in this thesis are listed in Table 2.4.

Table 2.4 Bioinformatic resources and software.

Resource	Application	Source	Reference
Vector NTI® Advance 11	DNA sequence display and analysis	Life Technologies, USA	[397]
SeqClean	Trimming and validation of DNA sequences by screening for various contaminants, low quality and low complexity sequences	http://sourceforge.net/projects/seqclean/	[398]
dbCAN 3.0	HMM-based database for automated CAZyme annotation	http://csbl.bmb.uga.edu/dbCAN/	[399]
IMG/M ER	Functional annotation and expert review of unpublished metagenomic data sets	https://img.jgi.doe.gov/cgi-bin/m/main.cgi	[400]
GS *De Novo* Assembler 2.7	*De novo* assembly of the 454 pyrosequencing data	Roche Applied Science, Germany	[401]
CD-HIT	Sequence clustering	http://weizhong-lab.ucsd.edu/cd-hit/	[402]
GETORF	Finding and extracting open reading frames	http://emboss.bioinformatics.nl/cgi-bin/emboss/getorf	[403]
FUJIFILM Science Lab Image Gauge 4.0	Image analysis and quantification of image data	Fujifilm, Japan	
Basic Local Alignment Search Tool (BLAST)	Finding regions of local similarity between sequences	http://blast.ncbi.nlm.nih.gov/Blast.cgi	[404]
SignalP 4.1	Prediction of signal peptides and their cleavage sites in proteins from Gram-positive and Gram-negative prokaryotes and eukaryotes	http://www.cbs.dtu.dk/services/SignalP/	[336]
TMHMM 2.0	Prediction of transmembrane helices in proteins	http://www.cbs.dtu.dk/services/TMHMM/	[338]
LipoP 1.0	Prediction of lipoprotein signal peptides in Gram-negative bacteria	http://www.cbs.dtu.dk/services/LipoP/	[337]

Resource	Application	Source	Reference
SecretomeP 2.0	Prediction of non-classical protein secretion	http://www.cbs.dtu.dk/services/SecretomeP/	[259]
PRED-LIPO	Prediction of lipoprotein signal peptides in Gram-positive bacteria	http://bioinformatics.biol.uoa.gr/PRED-LIPO/	[339]
PRED-TAT	An HMM-based prediction of Twin-Arginine Translocation (Tat) signal peptides and their cleavage sites	http://www.compgen.org/tools/PRED-TAT/	[340]
PilFind 1.0	Prediction of type IV pilin-like signal peptides and their prepilin peptidase cleavage sites in bacterial Gram-positive and Gram-negative bacteria	http://signalfind.org/pilfind.html	[314]
PSORTb 3.0	Bacterial localisation prediction tool	http://www.psort.org/psortb/	[341]
InterProScan	Scanning sequences for matches against the InterPro collection of protein signature database	http://www.ebi.ac.uk/Tools/pfa/iprscan/	[405]
CAZymes Analysis Toolkit (CAT)	Analysis and annotation of CAZymes	http://mothra.ornl.gov/cgi-bin/cat.cgi	[406]
KnotInFrame	Prediction of ribosomal –1 frameshift sites	http://bibiserv.techfak.uni-bielefeld.de/ knotinframe/	[407]
Scanner And Reporter Of Target-Unrelated Peptides (SAROTUP)	Suite of web tools for finding possible target-unrelated peptides from panning results	http://immunet.cn/sarotup/	[408]

2.2 Methods

2.2.1 Bacterial strains and growth conditions

E. coli strain TG1 was used to construct the shotgun metagenomic phage display libraries, as a host for propagating wild type (wt) VCSM13 helper phage, and for transformation with ssDNA of a pilot metasecretome phage display library.

The complementing *E. coli* strain K1976 was obtained by transforming TG1 strain with plasmid pJARA112, which contains full length *gIII* under the control of the phage infection-inducible promoter, *psp* [374]. Strain K1976 was used to propagate and titre VCSM13d3, a *gIII*-deletion (*ΔgIII*) helper phage.

All *E. coli* strains were propagated at 37°C in liquid media (2×YT or SOC) with aeration, or on solid media (2×YT or M9 minimal). Media were supplemented with antibiotics as required (section 2.1.2.7). Bacterial strains were stored short-term at 4°C on solid media in Petri dishes. All bacterial strains were stored long-term at –80°C as 1 mL stocks derived from fresh overnight cultures, supplemented either with 7% (v/v) dimethyl sulfoxide (DMSO) or 15% (v/v) glycerol.

2.2.2 Molecular biology methods

2.2.2.1 DNA extraction and purification

For plasmid DNA extraction from *E. coli*, commercial kits based on alkaline lysis methods were used according to the manufacturers' protocols. An AxyPrep Plasmid Miniprep Kit (Corning, USA) was used for small-scale plasmid isolation. A PureLink™ HiPure Plasmid Filter Midiprep Kit (Life Technologies, USA) was used if larger quantities of plasmid were needed. The E.Z.N.A.® M13 DNA Mini Kit (Omega Bio-Tek, USA) was used to purify ssDNA from phage particles.

For purification of DNA fragments from agarose gels, AxyPrep DNA Gel Extraction (Corning, USA) and QIAquick Gel Extraction (Qiagen, The Netherlands) kits were used as per the manufacturers' protocols.

Metagenomic DNA was extracted from bovine rumen plant-adherent microbiota as described in section 2.2.4.3. Purified DNA was stored at –20°C until required.

2.2.2.2 Agarose gel electrophoresis

The gels contained 0.8 – 1% (w/v) agarose and the 1× final concentration of SYBR® Safe DNA Gel Stain (Life Technologies, USA) in 1× TAE buffer. DNA samples were mixed with OrangeG DNA Loading Dye (1× final) prior to loading, except for the DNA size standard (1 Kb Plus DNA ladder; Life Technologies, USA). Gels were run in 1× TAE buffer at 5 V cm^{-1} in a Wide mini-Sub® Cell GT horizontal electrophoresis system (Bio-Rad Laboratories, Inc., USA) for 90 min or in an Owl Horizontal A2 Large Gel System (Thermo Fisher Scientific, USA) for 30 min. DNA was visualised using a UV transilluminator and digitally photographed using a Nikon D700 digital camera with Kodak Gel Logic 200 Imaging System software (Eastman Kodak, USA). A Safe Imager™ 2.0 Blue Light Transilluminator (Life Technologies, USA) was used to visualise DNA for gel extraction.

2.2.2.3 DNA quantification

Fluorometric and spectrophotometric methods were used to measure the concentration, and estimate the purity of DNA samples. Fluorescence-based quantification assays were performed using a Qubit® dsDNA BR Assay Kit and Qubit® ssDNA Assay Kit with the Qubit® 2.0 Fluorometer (Life Technologies, USA) according to the manufacturer's instructions. A NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, USA) was used for spectrophotometric quantification, where the DNA concentration was calculated from absorbance at 260 nm wavelength.

2.2.2.4 Preparation and transformation of electro-competent *E. coli* cells

Electro-competent *E. coli* cells used in this work were prepared using a modified protocol by Jacobsson *et al*. [329], with the exception of the TG1 electro-competent cells used to generate the large shotgun library, that were purchased from Stratagene (USA).

Electro-competent cells were transformed by electroporation using 0.1 cm gap electroporation cuvettes and a Gene Pulser® II Electroporation System (Bio-Rad), under the following conditions: 1.8 kV, 25 μF, 200 Ω, unless stated otherwise. Transformed cells were transferred to a tube containing 950 μL of freshly prepared SOC medium and incubated for 1 h at 37°C with rotary agitation. Ten-fold serial dilutions of the transformed cells were plated on 2×YTCm25 agar and incubated overnight at 37°C to allow colonies to form before counting.

2.2.2.5 Polymerase chain reaction (PCR) amplification

2.2.2.5.1 Bacterial colony PCR

Recombinant plasmids were analysed by PCR using DNA released from resuspended bacterial colony cells as a template. Primers pspF03 and pspR03 (Table 2.3) used in these reactions were complementary to the vector sequences flanking the multiple cloning site (MCS). The reverse primer pspR03 anneals 137 nt downstream of the cloning site, resulting in amplification of the vector sequence encoding the c-myc peptide tag and the first 30 amino acid residues of the C-terminal domain of pIII.

To prepare the cell suspension for PCR, individual colonies were picked from the transformation plates using sterile toothpicks and resuspended in 100 µL 1×TBS (25 mM Tris, 137 mM NaCl, 3mM KCl, pH 7.4) and this suspension was used at 1/10 volume of the colony PCR reactions (e.g. 5 µL in a 50 µL reaction, see Table 2.5). Colony PCR amplifications were carried out using Platinum® Taq DNA Polymerase High Fidelity (Life Technologies, USA) in the Mastercycler®pro system (Eppendorf, Germany). The components and the thermal profile of this PCR reaction are listed in Tables 2.5 and 2.6, respectively.

Table 2.5 Components of a PCR reaction mixture for colony PCR.

Component	**Volume per reaction (50 µL)**	**Final concentration**
10× High Fidelity PCR Buffer	5 µL	1×
10 mM dNTP mixture	1 µL	0.2 mM each
50 mM $MgSO_4$	2 µL	2 mM
200 pM sense primer pspF03 (section 2.1.4)	0.125 µL	0.5 pM
200 pM antisense primer pspR03 (section 2.1.4)	0.125 µL	0.5 pM
Cell suspension in 1×TBS	5 µL	N/A
Platinum® Taq HiFi polymerase (5U μL^{-1})	0.2 µL	0.02 U μL^{-1}
DNA-free PCR water (section 2.1.2.2)	36.55 µL	N/A

Table 2.6 Thermal profile of the colony PCR reaction.

PCR step		**Temperature**	**Time**
1	Initial denaturation	94ºC	5 min
2	Denature	94ºC	30 s
3	Anneal	55°C	30 s
4	Extend	68ºC	2 min*

PCR step		Temperature	Time
5	Repeat steps 2 – 4 total 30×		
6	Final extension	68ºC	7 min

* Extension time recommended by manufacturer: 1 min per Kb of DNA template

The amplicons obtained after completion of the bacterial colony PCR were analysed by agarose gel electrophoresis and/or used as templates for sequencing. The PCR products that served as sequencing templates were further purified using the NucleoSpin® Gel and PCR Clean-up (Macherey-Nagel, Germany) and USB® ExoSAP-IT® PCR product clean-up (Affimetryx, USA) kits according to the manufacturers' protocols.

The cell suspensions that were used in the PCR reactions were further clonally purified by streaking onto the selective plates (2×YTCm25). These purified transformants were used as a source of recombinant plasmids (section 2.2.2.1) for further analyses.

2.2.2.5.2 PCR amplification of ssDNA for pyrosequencing

PCR amplification of the recombinant phagemid ssDNA pool, derived from the large-scale shotgun metagenomic library after metasecretome selection, was performed in order to amplify the secretome-selected inserts as a template for pyrosequencing. The hot-start PrimeSTAR® Max DNA Polymerase (Takara Bio Inc., Japan) was used for amplification in the Mastercycler®pro system (Eppendorf, Germany). Primers PCRF2 and PCRR2 (see section 2.1.4) were designed to anneal to the phagemid vector sequences 361 bp upstream and 367 bp downstream of the library insert, respectively.

The components and the thermal profile of this PCR reaction are listed in Tables 2.7 and 2.8, respectively.

Table 2.7 Components of a PCR reaction mixture for ssDNA amplification.

Component	Volume per reaction (50 µL)	Final concentration
PrimeSTAR Max Premix (2×) [a]	25 µL	1×
20 pM sense primer PCRF2 (section 2.1.4)	1.25 µL	0.5 pM
20 pM antisense primer PCRR2 (section 2.1.4)	1.25 µL	0.5 pM
ssDNA template	2.5 µL	5 – 50 pg μL^{-1}
DNA-free PCR water (section 2.1.2.2)	20 µL	N/A

[a] 2× premix contains the PrimeSTAR® Max DNA Polymerase, reaction buffer, 2 mM Mg^{2+} and dNTP mixture

Table 2.8 Thermal profile of the PCR reaction (rapid amplification protocol).

PCR step		Temperature	Time
1	Denature	98ºC	10 s
2	Anneal	55°C	5 s
3	Extend	72ºC	25 s [a]
4	Repeat steps 1 – 3 total 35×		

[a] Extension time recommended by manufacturer: 5 s per Kb of DNA template

2.2.3 Phage protocols

In the phagemid/helper phage system used in this work, virions containing helper phage genomes are referred to as a phage, whereas virions containing phagemid or recombinant phagemid genomes are referred to as Phagemid Particles (PPs).

2.2.3.1 Phage propagation

VCSM13 (*gIII*$^+$) and VCSM13d3 (Δ*gIII*) helper phage have been used in this work. VCSM13 phage was propagated and titred on strain TG1 (see section 2.2.1), whereas the VCSM13d3 phage stocks were generated on complementing strain K1976 (see section 2.2.1). The VCSM13 phage was used for production of infectious recombinant PPs for panning and affinity binding. The VCSM13d3 phage was used for selection of inserts encoding the secretome protein-c-myc-pIII translational fusions. In the VCSM13d3-mediated display, the secretome protein-c-myc-pIII fusions are displayed in five copies, i.e. the display is 'polyvalent', in contrast to a single fusion per virion when the wt helper phage, VCSM13, is used.

All helper phage stocks were derived from a single plaque, using a plate-based (plate) method, by mixing 10^5 – 10^6 phage from a resuspended plaque with 0.2 mL of the appropriate host strain overnight culture and 2.5 mL of soft agar, and poured over the appropriate solid medium in a Petri dish. The plates were incubated overnight at 37°C and the phage were subsequently extracted from the lawn by overlaying 5 mL of 2×YT media and incubating with slow rotary agitation for 1 h at RT. The extracted phage were collected and separated from the cells by centrifugation at 13,200×g for 20 min at 4 °C, followed by heating at 65 °C for 10 min to kill the remaining viable cells. Phage was additionally purified and concentrated as described in section 2.2.3.2.

Phage stocks of the volumes over 100 mL were obtained using a liquid-based (liquid) method. An exponentially growing *E. coli* culture ($OD_{600\ nm}$ ~0.2) was infected by a phage stock (obtained from a single plaque using the plate method) at the multiplicity of infection (m. o. i.) of 50 phage per cell and incubated for 30 min at 37 °C without agitation, then 6 h with shaking at 300 rpm. The cells were pelleted by centrifugation at 13,200×g for 20 min at 4 °C. The supernatant (phage stock) was collected and the remaining bacterial cells were killed by heating at 65°C for 10 min. Phage stocks were additionally purified and concentrated as described in section 2.2.3.3..

2.2.3.2 Preparation of PPs

In those experiments where PPs were produced, the liquid method was mainly used (see section 2.2.3.2.1), except for PPs derived from the large shotgun metagenomic library that were prepared by plate method (see section 2.2.3.2.2).

2.2.3.2.1 PPs preparation by liquid method

Cultures of *E. coli* cells containing recombinant phagemids or phagemid vectors (100 – 200 mL) in the exponential phase of growth ($OD_{600\ nm}$ ~0.2) were infected at an m. o. i. of 50 with appropriate helper phage and incubated at 37ºC without shaking for 30 min (for VCSM13) or 1 h (for VCSM13d3).

Infected cells were pelleted by centrifugation at 3,300×g for 10 min at RT to remove the remaining unabsorbed helper phage, resuspended in an equal volume of fresh media and incubated for 6 h at 37ºC with aeration. Host cells were removed by centrifugation at 13,200×g for 20 min at 4 ºC, the supernatant was collected and the PPs were concentrated and purified as described in section 2.2.3.3.

2.2.3.2.2 PPs preparation by plate method

To prepare PPs derived from the large metagenomic library in pDJ01, library cells were, after an overnight amplification, diluted 100-fold into 100 mL of 2×YTCm25 media and grown until exponential phase. Cells in the exponential phase of growth ($OD_{600\ nm}$ ~0.2) were infected with the VCSM13d3 helper phage and harvested by centrifugation as described in section 2.2.3.2.1. The resulting pellet was mixed with 40 mL of soft agar. Agarose-embedded cells were poured over 16 double-layer Cm plates and incubated overnight at 37°C [352]. Both

the soft agar and the double-layer plates were prepared as described in section 2.1.2.7 except that molecular biology grade agarose was used instead of bacteriological agar. PPs were extracted from each plate by adding 5 mL of 2×YT media on top of the soft agar surface and incubating with rotary agitation for 1 h at RT. Aliquots of the media containing PPs collected from the individual plates were pooled together and PPs were purified and concentrated (see section 2.2.3.3).

2.2.3.3 Purification and concentration of phage and PPs

The 2×YT medium containing extracted virions was filtered through a sterile Millex®-HV Syringe Filter Units (EMD Millipore, USA) containing a very low protein binding Durapore® (PVDF) membrane with a 0.45 µm pore size to eliminate any remaining bacterial cells. Virions were further concentrated (100 – 200×) by precipitation in pre-chilled phage precipitation buffer (PEG/NaCl; section 2.1.2.6) for 1 h on ice, pelleted at 13,200×g for 30 min at 4 ºC and resuspended in 1 mL 10 mM Tris-HCl (pH 7.6). Virions were titred as described in section 2.2.3.6 and stored at 4ºC short-term or in 7% DMSO at –80ºC long-term.

2.2.3.4 Disassembled virion gel electrophoresis

Disassembled virion gel electrophoresis was used to detect total viral ssDNA (the sum of encapsulated and free viral DNA). Prior to electrophoresis, virions were mixed with SDS-containing buffer (see section 2.1.2.6) in a 3:1 volume ratio (virions to buffer) and disassembled by incubation at 70°C for 20 min. Disassembled virions were directly loaded onto a gel containing 0.6% (w/v) agarose in 1× TAE buffer and subjected to agarose gel electrophoresis as described in section 2.2.2.2. Total viral ssDNA was visualised by post-staining with a solution of 0.5 µg mL^{-1} ethidium bromide (EtBr).

2.2.3.5 Native virion gel electrophoresis

Native virion agarose gel electrophoresis allows detection of the free and encapsulated viral DNA, and was used to analyse the stability of recombinant virions and to verify sarkosyl selection. Samples were loaded onto 0.4% agarose gels in BlueJuice™ Gel Loading Buffer (Life Technologies, USA) and electrophoresis was performed at 1 V cm^{-1} in Wide mini-Sub® Cell GT horizontal electrophoresis system (Bio-Rad) for 17 h. Free (not encapsulated) phage ssDNA was detected by soaking the gel in a solution of 0.5 µg mL^{-1} EtBr. To detect

encapsulated phage ssDNA, virions were subsequently disassembled by soaking the gel in 0.2 M NaOH for 1 h, followed by neutralisation in 0.45 M Tris (pH 7.1) for 15 min and soaking again in EtBr solution to visualise the DNA that became exposed after the virion disassembly step by NaOH.

2.2.3.6 Enumeration of phage and PPs

2.2.3.6.1 Titration of phage and PPs

The phage and PPs were titred using a quick 'drop' method for titre estimation, or a full 'plate' method for increased accuracy.

In the quick 'drop' method, 10 µL drops of serially diluted phage were placed onto a soft agar layer containing 200 µL of TG1 or K1976 overnight culture. Plaques that developed in the area of the absorbed drops were counted to estimate the approximate number of phage. For accurate titres, the appropriate volume of a dilution, such that 200 – 300 plaques per plate are obtained (as estimated by quick 'drop' method), was mixed with 200 µL of TG1 or K1976 overnight culture and 2.5 mL of soft agar. The mixture was poured over pre-warmed plates (in triplicate) and incubated overnight at 37°C.

The helper phage were titred as described above, and the titre was expressed as the number of plaque-forming units per mL (PFU mL^{-1}). VCSM13 phage was titred on TG1 cells and VCSM13d3 phage on the complementing strain, K1976.

The infectious PPs were titred in triplicate on TG1 as described above, except that instead of counting plaques, the titre was determined through transduction of the Cm resistance marker (Cm^R), carried by PPs, into the host strain. For this, specially prepared double-layer Cm plates (see section 2.1.2.7) were used to allow in-agar infection of the TG1 strain and expression of transduced Cm^R marker by indicator strain TG1 prior to antibiotic exposure [352]. The PP titre was expressed as colony-forming units per mL (CFU mL^{-1}).

2.2.3.6.2 Phage and PPs quantification by densitometry

Non-infectious phage and PPs were quantified by densitometry based on the amount of encapsulated DNA [409]. Phage ssDNA, released from the SDS-disassembled virions and subjected to disassembled virion agarose electrophoresis (see section 2.2.3.4), was stained with EtBr and quantified densitometrically. Given that densitometric measurement of band density in agarose gels is based on comparison with a standard curve generated using DNA of known

concentrations, every gel contained a series of two-fold dilutions of purified helper phage ssDNA (10 – 360 ng) as standards. The gel was digitally photographed as described in section 2.2.2.2 and densitometric analysis was performed using the Science Lab 2001 Image Gauge software version 4.0 (Fujifilm, Japan) and Excel (Microsoft, USA).

A second-degree polynomial function was used for standard curve fitting over the series of standard data points. Conversion of the calculated amount of ssDNA in the samples into the number of virions *via* number of genome equivalents was based on molecular weight of ssDNA genome, calculated from the size of phage or phagemid DNA and its base composition. The number of genome equivalents corresponds to the number of genomes in a particular PP sample, which may correspond to a single PP, if a single genome is packaged into a particle, or a unit of a PP corresponding to a single genome, if multiple genomes are sequentially packaged into a long PP. The population of PPs obtained using VCSM13d3 helper phage in the presence of functional pIII expressed from the phagemid typically demonstrates a wide distribution of PP lengths, containing from one to several genomes packaged into a single virion, whereas the PPs obtained using the VCSM13 are typically 95% single-genome PPs and 5% of double-genome PPs [354].

2.2.4 Rumen content fractionation and metagenomic DNA extraction

2.2.4.1 Rumen sampling

Rumen sampling was conducted in May 2009 at a DairyNZ research farm (Lye Farm, Hamilton, New Zealand) under the animal ethics permission number AE 11483 granted by the Ruakura Animal Ethics Committee. A sample of the whole rumen content (digesta) was obtained from a fistulated Friesian dairy cow that was a part of normal production herd grazing *ad libitum* on a ryegrass/clover pasture. The herd was supplemented with pasture silage (around 10% of the recommended daily intake per animal) during the sampling period. Approximately 1.5 kg of rumen contents was collected in the morning (two hours post-feeding), immediately placed in a bucket gassed with carbon dioxide to minimise the effect of oxygen on the sample, and transported to the laboratory for further processing.

2.2.4.2 Rumen content fractionation

A protocol for partitioning the rumen microbial fraction that is tightly adherent to plant biomass (plant-adherent fraction) from the planktonic (liquid) and loosely attached (associated)

microbial fractions was modified from Larue *et al.* [62]. A schematic overview of the rumen content fractionation procedure is represented in Figure 2.1.

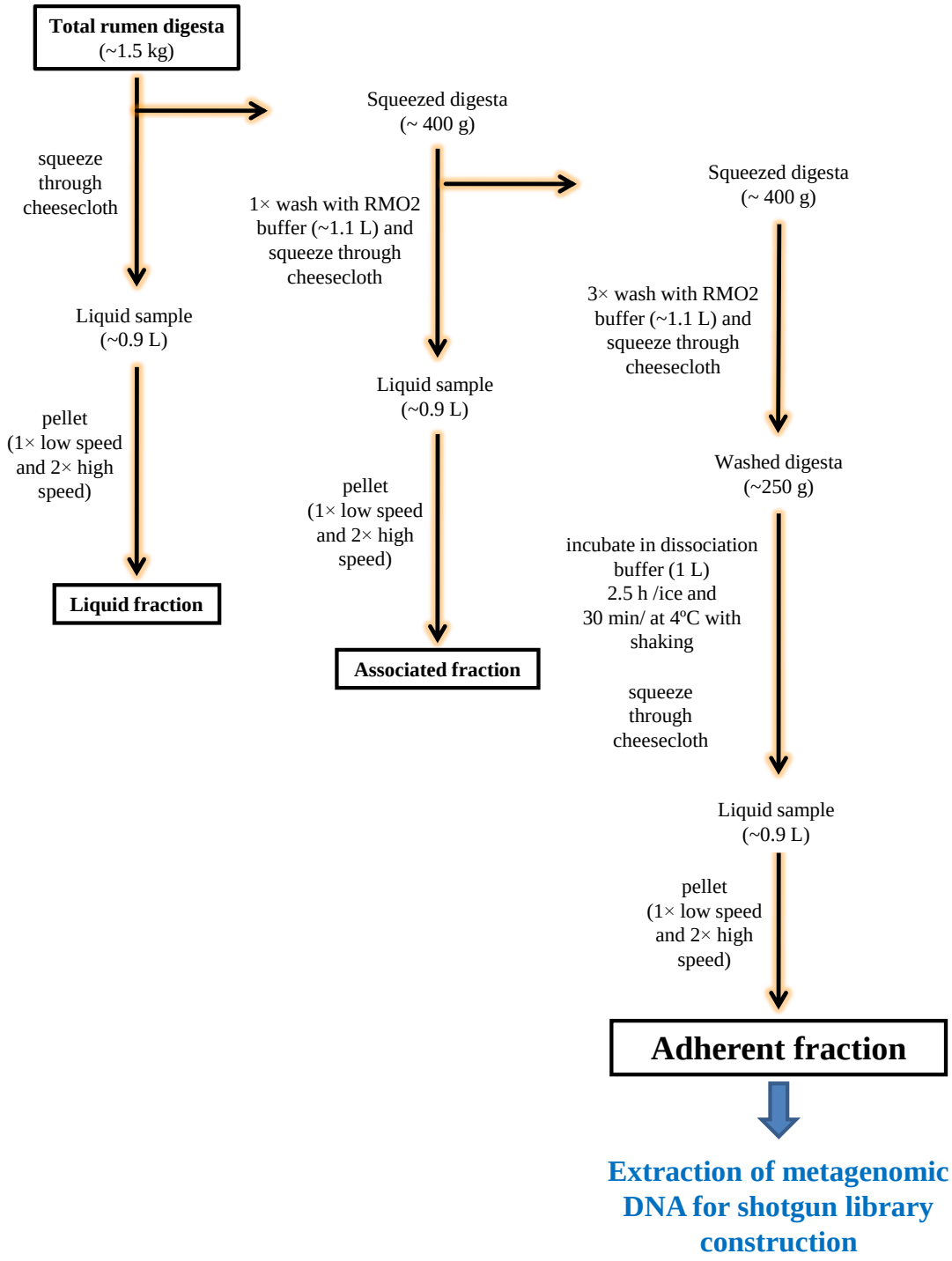

Figure 2.1 Overview of the rumen content fractionation procedure.

Briefly, the freshly collected rumen content (~1.5 kg) was squeezed through double-layered cheesecloth. The liquid fraction of the rumen content (0.9 L) was processed to isolate microbial cells defined as the 'planktonic' rumen microbial fraction. Plant debris was first

removed from the liquid fraction of the rumen content by low-speed centrifugation at RT (350×g for 15 min), followed by harvesting the microbial cells from the supernatant by two high-speed centrifugation steps at 10,000×g for 10 min. Approximately 400 g of the solid fraction of the rumen content that remained in the cheesecloth (squeezed digesta) was washed by resuspending in 1066 mL RM02 buffer (see section 2.1.2.4) at RT, and squeezing the liquid again through a double-layer cheesecloth. The microbial sub-community defined as the 'plant-associated' rumen microbial fraction was obtained from this first wash by the same procedure as described above for the planktonic microbial fraction. A further three washes of the squeezed digesta were performed using RM02 buffer as described above.

To obtain the 'plant-adherent' rumen microbial fraction, the microbial cells were chemically detached from 250 g of the washed and squeezed digesta by incubation in 1 L of dissociation buffer (see section 2.1.2.4) for 2.5 h on ice, followed by incubation at 4ºC for 30 min with shaking at 290 rpm. The fluid obtained after squeezing the digesta through the cheesecloth was centrifuged at 350×g for 15 min at RT to remove the small plant debris, followed by the centrifugation at 10,000×g for 10 min to pellet the microbial cells. The supernatant was discarded and the microbial pellet was washed once in 5 mL RM02 buffer, centrifuged at 10,000×g for 10 min to collect the cells, followed by decanting the supernatant and snap-freezing the cells in liquid nitrogen.

The weight of obtained microbial fractions and samples of digesta from different phases of fractionation process was recorded. The frozen bacterial pellet was stored at –85°C until DNA extraction.

2.2.4.3 Metagenomic DNA extraction from rumen microbial plant-adherent fraction

The protocol for extracting metagenomic DNA from the rumen microbial plant-adherent fraction was modified from Stein *et al.* [410]. The microbial cell pellet from the plant-adherent fraction (2 g) was split into five samples and each was separately embedded into 0.7 mL of 1% low-melting-temperature agarose (SeaPlaque® GTG® Agarose, Cambrex Bio Science Rockland, Inc., USA) and incubated in a syringe for 10 min on ice. Samples were extruded into 10 mL of lysis buffer (see section 2.1.2.5) and incubated for 2.5 h at 37°C, followed by incubation in 40 mL ESP buffer (see section 2.1.2.5) for 17 h at 55°C to inactivate nucleases present in the sample. After addition of fresh ESP buffer (20 mL) to each sample and incubation at 55°C for 1 h, three washes with TE buffer [containing 10 mM Tris-HCl (pH 8.0) and 1 mM EDTA] were performed and the remaining proteases from the ESP buffer were inactivated at 70°C for 15 min. To digest the agarose, samples were incubated overnight at 37°C

with 15 U of AgarACE™ agarase (Promega, USA). Residual insoluble oligosaccharides were removed by centrifugation and the supernatant, containing crude lysate released from the agarose, was subjected to DNA extraction using phenol:chloroform:isoamyl alcohol (in 25:24:1 ratio, v/v). After pooling together the five samples derived from 2 g of the microbial cell pellet, metagenomic DNA was concentrated using a Vivaspin sample concentrator (100 KDa cut-off; GE Healthcare Lifesciences, USA). Pulsed-field gel electrophoresis was used to confirm the integrity and quality of the isolated metagenomic DNA (Figure 3.1).

2.2.5 Rumen metasecretome phage display

A schematic overview of the metasecretome library construction and selection is presented in Figure 2.2.

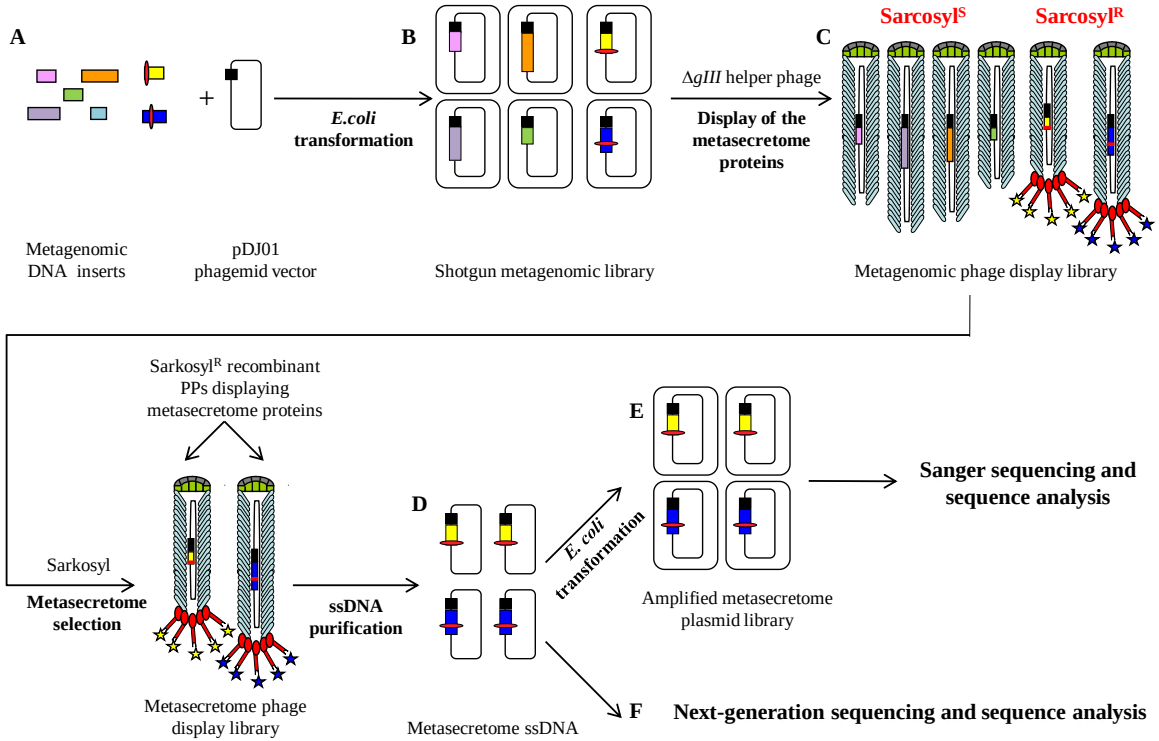

Figure 2.2 Overview of the metasecretome library construction and selection.

A. Shotgun metagenomic library construction. Some metagenomic inserts contain endogenous signal sequences, represented by red ovals. **B**. Shotgun metagenomic library infection with the gIII-deleted helper phage VCSM13d3. **C**. Metagenomic phage display library contains a mix of virions capped by insert-pIII fusion proteins (signal sequence-positive clones) that are resistant to sarkosyl (Sarkosyl^R) and uncapped virions (signal sequence-negative clones) that are sensitive to sarkosyl (Sarkosyl^S). Sarkosyl resistance is used as a basis for selection. **D**. Single stranded DNA (ssDNA) purified from Sarkosyl^R virions after the selection was used to obtain the amplified metasecretome plasmid library for preliminary assessment of metasecretome diversity by Sanger sequencing (**E**) and for in-depth assessment of metasecretome selection by next-generation sequencing (**F**).

2.2.5.1 Construction of rumen microbial metagenomic shotgun libraries

Two shotgun metagenomic libraries were constructed: i) a small pilot library for the preliminary assessment of rumen microbial metasecretome selection and ii) a large library for the metasecretome characterisation by next-generation sequencing. Both libraries were constructed from mechanically sheared metagenomic DNA isolated from the rumen plant-adherent microbial fraction, and cloned into the secretome-selective phagemid vector pDJ01 [234].

Around 150 µg of high molecular weight metagenomic DNA was split into ten aliquots, mixed with shearing buffer (see section 2.1.2.5), and sheared in disposable medical nebulisers (Unomedical Inc., USA) by subjecting the samples to the pressure of 10 psi for 1 min. Nebulisation was followed by size fractionation, de-salting and concentration using a Vivaspin sample concentrator (100 KDa cut-off; GE Healthcare Lifesciences, USA). The ends of the sheared metagenomic DNA (72 µg) were repaired and 5'-phosphorylated using an enzyme cocktail containing 16 U T4 DNA Polymerase (Roche Applied Science, Germany), 76 U Klenow Enzyme (Roche Applied Science, Germany) and 190 U USB OptiKinase™ (Affimetryx, USA). The repaired DNA was purified by phenol:chloroform:isoamyl alcohol (25:24:1) extraction, followed by ethanol precipitation, and the DNA pellet was resuspended in 150 µL of 10 mM Tris-HCl (pH 8.0). The phagemid vector pDJ01 was prepared by digesting with *Sma*I restriction endonuclease (Roche Applied Science, Germany) and the 5' ends of the linearised vector were dephosphorylated using rAPid Alkaline Phosphatase (Roche Applied Science, Germany). The linearised and dephosphorylated vector was purified from the remaining uncut vector by gel purification using a QIAEX II Gel Extraction Kit (Qiagen, The Netherlands).

To optimise conditions for library construction, a number of test ligations using different molar ratios of vector and the sheared and end-repaired metagenomic DNA (the insert) were performed using T4 ligase (Roche Applied Science, Germany). Test ligations were desalted using agarose cones as described in [411], transformed into *E. coli* TG1 competent cells by electroporation (as described in section 2.2.2.4), and the number of transformants was determined. The ratio that generated the largest number of transformants was used for the construction of primary libraries.

Approximately 19 µg of the end-repaired metagenomic DNA inserts were ligated to 6.5 µg of pDJ01 vector using T4 ligase. After phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation, the ligated DNA was washed with 70% (v/v) ethanol and dissolved in 75 µL of sterile deionised water.

A total of 2 µg of ligated and purified metagenomic DNA was transformed into the electrocompetent *E. coli* cells as described in section 2.2.2.4, to obtain the pilot shotgun library. The remaining ligation mixture was used for 27 separate transformations into electrocompetent TG1 purchased from Stratagene to generate a large shotgun library. These 27 TG1 aliquots were separately transformed and processed through the whole metasecretome selection procedure and the template preparation for the next-generation sequencing.

To estimate the primary shotgun library size, aliquots from each transformation were plated on 2×YT Cm_{25} plates. The remaining portion of each transformation mixture was mixed with 9 mL of 2×YT Cm_{25} broth and incubated for 8 h at 37°C with aeration to amplify the library aliquots. Amplified aliquots of the master shotgun metagenomic library were frozen at –80°C in 7% DMSO, apart from 1 mL which was immediately used for secretome selection.

2.2.5.2 Selection of metasecretome phage display library and isolation of ssDNA

Starting from a 1 mL aliquot of the overnight culture containing amplified shotgun library clones (described in section 2.2.5.1), PPs were produced using the plate method and concentrated (see section 2.2.3.1.2). A modified protocol [234] for the direct selection of the secretome phage display library was applied to PPs produced from both shotgun metagenomic libraries (the small pilot library and 27 large shotgun library aliquots) separately. To eliminate structurally unstable virions derived from non-secretome library clones, PPs were incubated in 0.1% (w/v) sarkosyl for 10 min at RT. The ssDNA released from defective virions was removed by incubation with 200 U of DNaseI (Roche Applied Science, Germany) in the presence of 5 mM $MgCl_2$ for 1 h at RT, followed by addition of EDTA (to final concentration of 25 mM). Sarkosyl-resistant recombinant virions were precipitated by PEG/NaCl solution for 3 h at RT and resuspended in 10 mM Tris-HCl (pH 7.6). This was followed by heating at 75°C for 10 min to inactivate DNaseI prior to extraction of ssDNA from sarkosyl-resistant PPs using E.Z.N.A.® M13 DNA Kit according to the manufacturer's recommendations.

The ssDNA isolated after the secretome selection was either transformed into electrocompetent *E. coli* cells as described in section 2.2.2.4 (and individual clones were analysed by sequencing), or was subjected to processing for the 454 pyrosequencing.

2.2.5.3 Sequencing of pilot metasecretome phage display library

Recombinant phagemid DNA from 90 randomly selected pilot library transformants was purified using AxyPrep™ Plasmid Miniprep Kit (Axygen, USA). All inserts were sequenced using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems,

USA) and analysed on an ABI3730 DNA analyser at the Massey Genome Service (Massey University, New Zealand). Inserts were sequenced from the 3' end only, using the primer pspR03 (Table 2.3) complementary to the pIII-coding sequence of the vector, in order to identify the insert-c-myc-pIII junction and determine the frame of the insert-containing ORF relative to pIII. If the start of the ORF was not reached, a sequencing reaction using the pspF03 forward primer (Table 2.3) complementary to the pDJ01 vector sequence upstream of the MCS was performed.

2.2.5.4 Testing conditions for the next-generation sequencing template preparation

The subject of this sequencing project is a metasecretome-selected library containing metagenomic DNA inserts (ranging in size between 0.7 and 5 Kb) in a phagemid vector (3.1 Kb). For this reason, using the whole recombinant phagemid as a sequencing template would result in a large amount of unnecessary sequencing of the vector. To overcome this problem, the template for shotgun sequencing was generated by PCR amplification of insert and flanking vector sequence from recombinant phagemid ssDNA, based on the concept described in [386]. A range of amplicon sizes were expected, with a large proportion over the 0.8 Kb size. DNA fragments in the 0.6 – 0.8 Kb size range are preferentially used by the Macrogen Inc. sequencing facility (Seoul, Korea) as a template for the 454 pyrosequencing, since longer DNA fragments do not amplify well in the emulsion-based PCR (emPCR). For this reason, it was necessary to fragment the obtained amplicons to the 0.6 – 0.8 Kb size range before subjecting them to shotgun sequencing. To investigate various fragmentation methods, a test mixture of amplicons that cover the expected size range of the inserts (0.7 – 5 Kb), obtained from the pilot library recombinant phagemids using the same pair of primers as for the pyrosequencing template preparation, was sheared enzymatically and mechanically using different conditions, and examined by agarose gel electrophoresis to monitor the resulting size range.

The amplicon test-mix was incubated with several concentrations of *Alu*I restriction endonuclease (0.05 – 10 U of enzyme per µg DNA) and varying incubation times (30 sec - 2 h), followed by heat inactivation at 65°C for 20 min. Partial digestion with DNaseI (0.05 – 1 U of enzyme per µg DNA) in the presence of Mn^{2+} (10 mM) at 37°C for 30 sec and 1 min was also tested, followed by DNaseI inactivation with EDTA (final concentration of 50 mM) and heat inactivation at 75°C for 10 min.

Mechanical shearing of the test-mix (containing 10% glycerol) was performed by nebulisation, subjecting a sample in a disposable nebuliser (Life Technologies, USA) to the pressure of 35 psi for 1, 2, 3, 4, 6, 8 and 10 min or by passing the sample through a 0.8 mm bore needle 5, 25, 50, 75 and 100 times.

2.2.5.5 Next-generation sequencing of the metasecretome

To produce the template for the next-generation sequencing of the metasecretome, amplicons were generated in 27 separate PCR reactions (as described in section 2.2.2.5.2), using the secretome-selected ssDNA as a template and primers PCRF2 and PCRR2 (Table 2.3). After the amplification was completed, the PCR reactions were pooled and subsequently divided into portions, and fragmented using five different conditions: 1 min *Alu*I digestion; 3 h *Alu*I digestion, 6 min nebulisation at 35 psi; 6 min nebulisation at 35 psi followed by 1 min *Alu*I digestion and 6 min nebulisation followed by 3 h *Alu*I digestion. *Alu*I was used at 5 U per µg of DNA and the reactions were incubated at 37°C, followed by heating at 65°C for 20 min to inactivate the enzyme. Nebulisation was performed on ice in a disposable nebuliser (Life Technologies, USA).

The sheared amplicons obtained under five conditions described above were separately purified and concentrated using NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel, Germany). DNA was quantified in each sheared sample by fluorometry, using a Qubit® dsDNA BR Assay Kit (Life Technologies, USA). The template for next-generation sequencing, containing 12.5 µg DNA in total, was prepared by mixing equal amounts (2.5 µg) of DNA templates prepared under each of the five conditions described above.

The next-generation shotgun sequencing (1/2 plate) using Titanium chemistry on a 454 GS FLX instrument (Roche Applied Science, Germany) was carried out by the Macrogen Inc. sequencing facility (Seoul, Korea). The sequencing service provider completed the template preparation according to the Rapid Library Preparation Method Manual (Roche Applied Science), starting from the second, fragment end-repair step of the protocol.

2.2.6 Bioinformatic analysis

2.2.6.1 Sequence analysis of the pilot metasecretome library inserts

The sequences obtained from sequencing the inserts of 90 randomly selected pilot metasecretome library clones (described in section 2.2.5.3) were analysed using the Vector NTI® Advance 11 software package. ORFs in translational fusion with *c-myc-g*III were identified using the Vector NTI ORF finder in combination with manual inspection.

Functional annotation of putative proteins encoded by the pilot metasecretome library ORFs was based on the best BLASTP hit against NCBI non-redundant (nr) protein sequence

database. Functional categories have been assigned to hits with an E-value of <1e-05. The taxonomic assignment of inserts at the genus level was based on the best BLASTX hits with an E-value of <1e-05 and a query coverage of >30%. Types of membrane-targeting signals in putative proteins, longer than 24 amino acid residues and in frame with the c-myc-pIII encoded by vector, were predicted using a range of available algorithms (SignalP 4.1 [336], TMHMM 2.0 [338], LipoP 1.0 [233, 337], PredLipo [339], SecretomeP 2.0 [259], PilFind 1.0 [314] and PRED-TAT [340]) using the default settings and cut-off values.

The efficiency of selection was estimated by comparing the frequency of the secretome insert-containing recombinant phagemids after selection with the theoretically expected frequency before selection. The theoretical expected frequency of the secretome insert-containing recombinant phagemids before selection was calculated as described in [329]. The calculation was based on the average proportion of the secretome ORFs in bacterial genomes, the probability of a correct insert orientation and a probability of the putative protein encoded by the insert to be in the correct frame for display. The frequency of the secretome insert-containing recombinant phagemids after the selection was calculated as the proportion of inserts containing an ORF encoding putative protein in frame with c-myc-pIII and a predicted membrane-targeting signal, as compared to all sequenced inserts.

2.2.6.2 *In silico* analysis of the next-generation sequencing metasecretome dataset

2.2.6.2.1 Rumen metasecretome unassembled and assembled datasets

Generation of the rumen metasecretome datasets used throughout this work is described below and the descriptive name of each dataset is represented in a bold font. A workflow overview of the metasecretome dataset *in silico* analysis is represented in Figure 2.3.

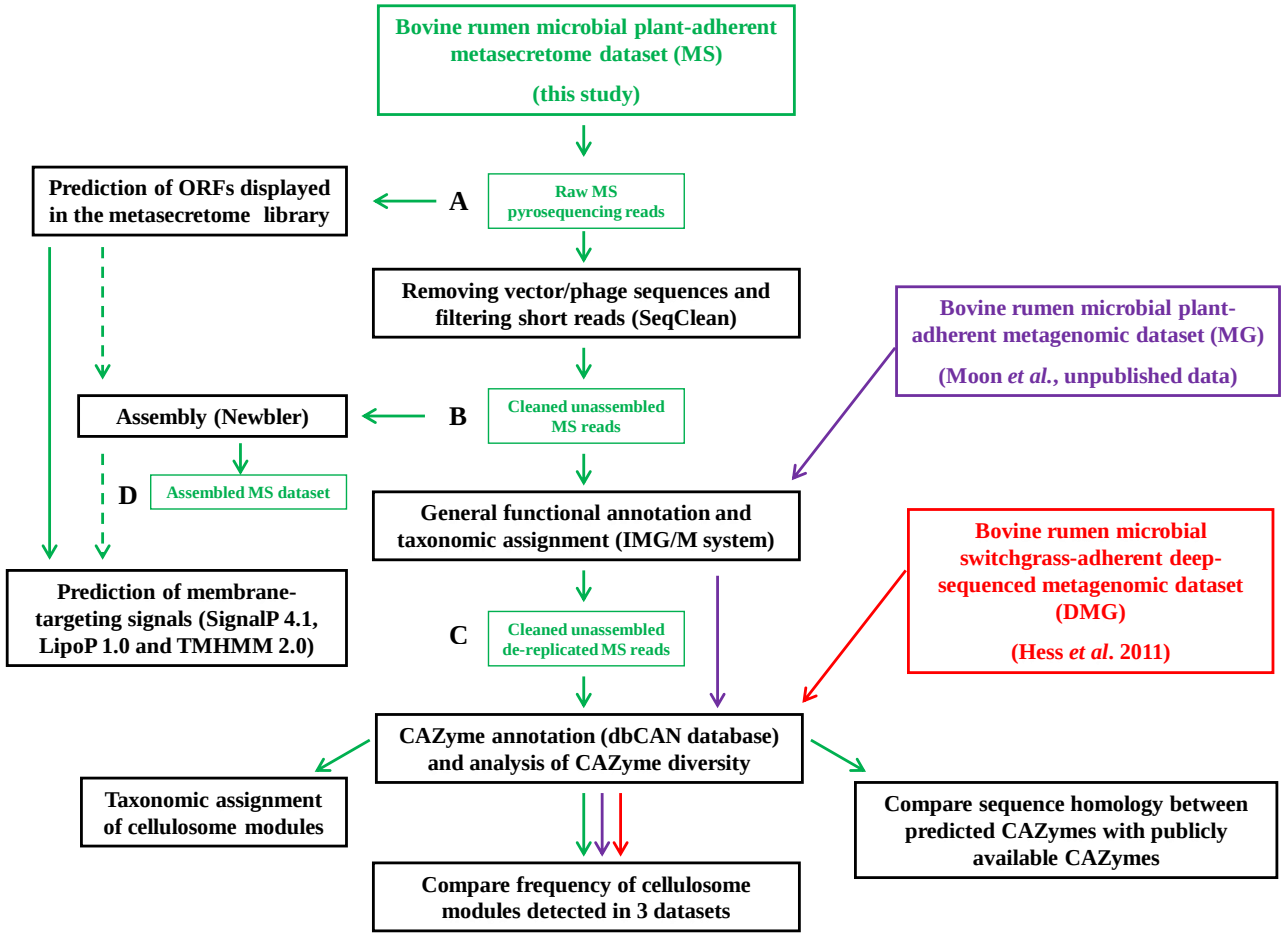

Figure 2.3 Workflow overview of the *in silico* analysis.

The metasecretome dataset (this thesis) was compared to the two bovine rumen microbial plant-adherent shotgun metagenomic datasets: Moon *et al.* (unpublished) and Hess *et al*. (2011) [22].

The raw metasecretome pyrosequencing reads (Figure 2.3, dataset A) were obtained by sequencing the metasecretome-enriched DNA template obtained by secretome selection, amplification and amplicon fragmentation (as described in sections 2.2.5.2 and 2.2.5.5, respectively), using the Roche 454 GS FLX Titanium platform. These raw reads were trimmed using a SeqClean script [398] to remove the sequences of pDJ01 vector and VCSM13d3 helper phage, and to filter out the low-quality and short reads. The minimum percent identity for an alignment with a contaminant was 85% and the minimum length cut-off for cleaned sequences was 100 bp, with the rest of parameters with default values. The resulting **cleaned unassembled metasecretome reads** (Figure 2.3, dataset B) were used as an input into the Integrated Microbial Genomes with Microbiome samples (IMG/M) system [400], hosted by the U.S. Department of Energy Joint Genome Institute [412]. Briefly, in the IMG/M pipeline, the following steps were performed prior to annotation: removal of commonly occurring discrepancies in the input sequence files; trimming of the low-quality regions; masking of the low-complexity regions, and de-replication. During de-replication, duplicate sequences were detected at 95% sequence identity and if the terminal 3 bp of detected duplicates were identical, only one representative sequence was retained. The resulting **cleaned unassembled de-replicated metasecretome reads** (Figure 2.3, dataset C) were used in all subsequent analyses within the IMG/M pipeline.

The cleaned unassembled metasecretome reads, obtained as described above, were assembled using a GS Roche *De Novo* Assembler [401], version 2.7, with the overlap requirements of a minimum length of 40 bp and a minimum identity 100%. To predict the ORFs in this **assembled metasecretome dataset** (Figure 2.3, dataset D)**,** the GETORF package [403] was used, and the minimal ORF length was set at 90 bp (minimal length of putative protein 30 amino acid residues).

2.2.6.2.2 Functional annotation and phylogenetic profile

A bovine plant-adherent rumen microbial metagenome dataset of two New Zealand cows on a fresh pasture diet (Moon *et al.*, unpublished) was included in the bioinformatics analyses to provide a reference point for comparison to the metasecretome data. Overview of samples used for generation of metasecretome and metagenome datasets is provided in Table 2.9.

Table 2.9 Overview of samples used for generation of the metasecretome and the metagenome datasets.

Dataset	Metasecretome	Metagenome
Sampled animal	Friesian dairy cow	Friesian dairy cow
Number of sampled animals	1	2
Diet	Ryegrass/clover pasture supplemented with 10% silage	Ryegrass-dominant pasture
Sampling time	May 2009	February 2008
Sampling location	Hamilton, New Zealand	Palmerston North, New Zealand
Sampling technique	Rumen fistula	Rumen fistula
Sample type	Whole rumen content	Whole rumen content
Microbial fraction	Plant-adherent	Plant-adherent
Fractionation method	Modified from [62]	[62]
DNA extraction method	Modified from [410]	ZR Fecal DNA Kit (ZymoResearch, USA), no bead beating
Sequencing platform	Roche 454 GS FLX (Titanium)	Roche 454 GS
Amount of sequencing	Half plate	Full plate, one plate per cow
Sequencing centre	Macrogen, South Korea	University of Otago, New Zealand
Number of analysed reads [a]	153,002	609,709 [c]
Average read length (bp)	362	226 [c]
Total sequence information (Mb) [b]	55	145 [c]

[a,b] Number of reads and total sequence information is provided for reads after processing via IMG/M pipeline. [c] Combined pyrosequencing datasets of two cows.

The metagenome dataset was processed and automatically annotated *via* IMG/M system [400] using the same parameters as for the metasecretome dataset. The metasecretome and metagenome sequences are deposited in the NCBI BioProject database [413] (accession ID PRJNA244109), while two annotated datasets can be accessed through the IMG/M website [412] under taxon objects ID 3300000332 and 3300000524, respectively.

Protein-coding genes were predicted in the metasecretome and metagenome datasets *via* the IMG/M system [400] using integrated Genemark, Prodigal, Metagene and FragGeneScan ORF calling/*ab initio* gene prediction programs. IMG/M-generated functional annotation associated putative proteins in the metasecretome and metagenome datasets with protein families and domains based on COG (Clusters of Orthologous Groups of proteins) clusters and functional categories, Pfams (Protein families), KO (KEGG Orthology) terms and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and EC numbers [400, 414]. The phylogenetic profiles of the two datasets were produced in the IMG/M pipeline by computing the distribution of the best BLASTP hits of putative proteins against the non-redundant reference database containing sequences from public IMG genomes and the KEGG database. Taxonomic assignments have been reported at the phylum level with 30% BLASTP identities cut-off. This low percent cut-off was necessary because the accuracy of taxonomic assignment

via similarity-based search depends on the availability of reference data, which is limited for metagenome samples.

2.2.6.2.3 CAZyme annotation and taxonomic assignment

Putative proteins, encoded by ORFs predicted in the metasecretome and metagenome datasets *via* the IMG/M pipeline, were downloaded using the JGI portal. Putative proteins from metasecretome and metagenome datasets (encoded by 222,960 and 671,876 ORFs, respectively), as well as putative proteins encoded by 2,547,270 predicted ORFs from the published bovine rumen microbial switchgrass-adherent metagenome dataset [22], were subjected to automated annotation and assignment to families of carbohydrate-active enzymes (CAZymes) using the dbCAN database CAZy family-specific HMMs (release 3.0, based on the CAZy database as of March 2013) [399]. The hmmscan functionality of the HMMER 3.0 software package [415] was used to search dbCAN database for homologous CAZyme domains (modules) in three datasets and the output was parsed using the following cut-off values: for an alignment length >80 amino acid residues, an E-value <1e-05; otherwise an E-value <1e-03. The HMM-based CAZyme annotation was chosen over the sequence similarity-based approaches to minimise dependence of the identification of putative CAZymes on the sequence similarity shared with known CAZymes and to allow the annotation of putative novel CAZyme sequences. To remove duplicates, all dbCAN hits were clustered at a 100% sequence identity threshold using the CD-HIT algorithm [402].

To search for putative proteins with multi-modular CAZyme organisation in the metasecretome dataset, only clustered putative proteins with non-overlapping hits to multiple CAZyme HMMs (previously parsed based on an E-value <1e-05 for an alignment length >80 amino acid residues and an E-value <1e-03 for shorter alignment lengths) that had query coverage >30% for each putative domain, were further inspected. The start-end position consistency between the HMM and the query protein and hit 'strength' for each domain were taken into account for prediction of multi-modular organisation of putative CAZymes. Hit strength was judged based on the conditional E-value (<<1 and smaller than individual E-value). For each 'hit' (an alignment between the query sequence and the domain represented by HMM) produced by hmmscan, a bit score, an individual E-value (iE-value) and a conditional E-value (cE-value) were produced [416]. An iE-value indicates the statistical significance of the best hit (from the top hits list) to the individual domain and describes the expected number of non-homologous hits to that particular domain to occur in the whole database by chance. A cE-value indicates the statistical significance of each domain under assumption that the target sequence is a true homologue and describes the expected number of additional hits that would

be found with obtained bit score among the set of sequences reported in the top hits list by chance. Small cE-value (<<1), smaller than the corresponding iE-value, indicates confidence that particular region of a query is 'hitting' a true homologous domain and can also be used to decide the statistical significance of weak-scoring hits to additional domains.

The clustered hits to cellulosome-associated modules were further analysed. The family level taxonomic assignments of putative proteins within the metasecretome dataset with hits to putative cellulosome modules were made based on the best BLASTP hit against the NCBI nr protein database with a bit-score over a threshold of 40 for cohesin and SLH module-containing proteins and 35 for dockerin-module containing proteins. The taxonomic assignments were then manually curated using recent bacterial classification proposals [176, 417-421].

2.2.6.2.4 Assessment of the novelty of putative CAZymes detected in the metasecretome dataset

To assess the ability to annotate novel putative CAZymes in the metasecretome dataset using an HMM-based approach, putative proteins containing predicted CAZyme modules, clustered at 100% sequence identity and >30 amino acid residues long, were compared with proteins deposited in the NCBI nr protein database and with relevant full-length putative CAZyme proteins downloaded from the dbCAN website using BLASTP with default parameters. Putative CAZymes deposited in the dbCAN database represented, as of May 2012, an approximately three times larger collection of CAZyme homologues compared to the NCBI nr database [399]. This specialised collection contains putative CAZyme sequences (4,073,867 as of May 2013) from CAZy [138], NCBI (nr and environmental nr) databases [423] and UniProt [424] databases; multiple metagenome datasets (JGI [425], CAMERA [426], BGI-gut [427] and cow rumen metagenome [22]); a small number of sequences from the human gut microbiome [428] and plant genomes [429] that are not available in GenBank. The best BLASTP hits were extracted and further analysed.

2.2.6.2.5 Prediction of membrane-targeting signals in the metasecretome dataset

To extract ORFs encoding polypeptides in frame with the c-myc tag and pIII, ORFs were first found and extracted from raw metasecretome pyrosequencing reads using GETORF [403] with the default parameters. Next, a TBLASTN search using the 15 amino acid residues long c-myc tag as a query was performed to locate ORFs encoding putative proteins that are in frame with c-myc (and pIII). After removing chimeric sequences with more than one c-myc match, translated ORFs with >93% c-myc sequence identity (allowing up to one mismatch to

accommodate potential sequencing errors) were extracted from the raw metasecretome pyrosequencing dataset. To remove duplicates, extracted translated ORFs in frame with c-myc were clustered at 100% sequence identity threshold using the CD-HIT algorithm [402].

Distinct putative proteins in frame with the c-myc tag and pIII, and longer than 24 amino acid residues, were analysed to predict the presence of membrane-targeting signals using the following software packages: SignalP 4.1 [336] with a D-score cut-off that reproduces the sensitivity of SignalP 3.0 and no N-terminal truncation of the input sequence, as well as TMHMM 2.0 [338] and LipoP 1.0 [233, 337] with default settings.

Putative proteins without predicted signal sequences were subjected to a BLASTP search against a local database of putative proteins encoded by metasecretome contig ORFs to find longer homologous putative proteins. ORFs were predicted in the assembled metasecretome dataset as for raw reads, using GETORF [403] with the default parameters. A BLASTP search was performed with default parameters for putative proteins without predicted membrane-targeting signals that were over 30 amino acid residues long, while parameters optimised for queries shorter than 30 amino acid residues were applied for putative proteins without predicted signal sequences, which were between 25 – 30 amino acid residues in length. Finding longer homologous putative proteins, some of which may contain the N-terminal sequence encoding a membrane-targeting signal, could increase the probability of detecting membrane-targeting signals in partial ORFs encoding putative proteins in frame with c-myc and pIII. Best BLASTP hits were extracted and subjected to prediction of membrane-targeting signals as described above. Finally, a predicted membrane-targeting signal was assigned to the putative proteins encoded by unassembled metasecretome sequences (without a previously detected membrane-targeting signal) based on the sequence similarity to the best BLASTP homologue in the assembled metasecretome dataset.

2.2.7 Affinity screening of the metagenomic shotgun library

2.2.7.1 Preparation, immobilisation and test-assays of complex carbohydrate substrates for panning

Several complex carbohydrate substrates: cellulose (in three forms), oat spelt xylan, insoluble wheat arabinoxylan, and ryegrass fibre (neutral detergent fraction) were tested as substrates (baits) by assays to identify those that result in low non-specific binding of PPs, as a low background is essential for the success of selection. One of the cellulose baits was Whatman No.1 paper (GE Healthcare Lifesciences, USA) in discs, containing 2.65 mg cellulose per disc. It was placed in polypropylene microtubes; the washing and binding was performed by

moving the disc from one tube to another using sterile tweezers. Around 100 mg of other carbohydrate substrates: microcrystalline cellulose (Sigma-Aldrich, USA), oat spelt xylan (Sigma-Aldrich, USA), insoluble wheat arabinoxylan (Megazyme, Ireland), regenerated amorphous cellulose and neutral detergent fraction of ryegrass were each placed in empty disposable polypropylene columns (10 mL) containing a polyethylene frit to retain the substrate, while allowing draining of the liquid when the stopper is removed (PD-10 empty columns; GE Healthcare Lifesciences, USA). All carbohydrate substrates, apart from the amorphous cellulose and the neutral detergent fibre fraction of ryegrass, were weighed and placed directly onto the column in powder form. A regenerated amorphous cellulose slurry, containing 20 g cellulose L^{-1}, was prepared as described in [430], supplemented with 0.2% (w/v) of sodium azide and stored at 4°C. The neutral detergent fraction of ryegrass, representing the insoluble component of plant cell walls (fibre), was prepared as described in [103].

The pDJ01 vector-derived PPs, used for the assays, were produced using VCSM13 helper phage, purified and titred (see section 2.2.3).

All substrates were blocked overnight with 2% (w/v) bovine serum albumin (BSA) in 1× TBS, and conditions described in section 2.2.7.2 were used in the binding assays. The number of PPs in the 'input' (mixed with the substrate at the start of the assay) and the 'output' (eluted from the substrate at the end of the assay) was determined by titration. The ratio of the total number of the output to input PPs served as a measure of 'background' PPs that bind non-specifically to the bait and the immobilisation matrices/vessels.

2.2.7.2 Affinity screening of the metagenomic shotgun phage display library on wheat arabinoxylan and amorphous cellulose

To identify recombinant PPs displaying hemicellulose- and cellulose- binding proteins, the metagenomic shotgun phage display library was screened using an affinity enrichment protocol. The two substrates, AXYL (100 mg) or 5 mL RAC slurry (20 mg mL^{-1}) were separately immobilised using columns as described in section 2.2.7.1. The substrate was washed once with 5 mL 1×TBST (1×TBS containing 0.05% (v/v) Tween 20) and blocked overnight with 5 mL of 2% (w/v) BSA in 1×TBS with gentle mixing at 4°C. The substrate was washed the following morning three times with 5 mL of 1×TBST buffer.

The PPs derived from the 27 separate aliquots of metagenomic library in pDJ01 vector (MG1 – MG27 PPs) were produced with the VCSM13 helper phage. This helper phage encodes a wt pIII, which competes for incorporation into the PPs with the phagemid-encoded pIII fusions, resulting in infectious PPs with monovalent display of the fusion proteins. Given that the vector does not encode a signal sequence, the display depends on the presence of an

endogenous signal sequence encoded by the metagenomic inserts. The PPs were produced using the liquid method (see section 2.2.3.1.1), starting from 27 separate 2 mL aliquots of the amplified shotgun metagenomic library (in the form of *E. coli* cells containing phagemids replicating from their plasmid origins; from 7% DMSO stocks), obtained as described in section 2.2.5.1. The PPs were precipitated in PEG/NaCl solution and purified as described in section 2.2.3.2. Twenty-seven aliquots of MG PPs were pooled together and the MG1 – 27 PPs mix was titred as described in section 2.2.3.3.1. Control PPs derived from vector pDJ01 were prepared as described in section 2.2.7.1.

MG1 – 27 PPs (sample) and pDJ01 PPs (control) were separately resuspended in 5 mL of 1×TBS supplemented with 0.2% BSA 1 h prior to addition to the column-immobilised pre-blocked substrate. Approximately 1×10^{12} PPs of both sample and control were separately mixed with the two carbohydrate substrates (in triplicate for each substrate) and incubated for 3 h at RT with continual rocking. Unbound PPs were removed by washing the substrate nine times with 1×TBST, followed by the last wash with 1×TBS. After binding and every washing step, the column was centrifuged at 1,000×g for 2 min to drain the liquid, whereas the substrate was retained in the column by the frit.

The bound sample and control PPs were eluted from the substrate under three different conditions: acidic, basic and by direct on-substrate infection of the *E. coli* TG1 host cells. For the acidic and basic elution, the mixture of the substrate and PPs were incubated at RT after the last wash with 920 µL acidic elution buffer (0.1 M glycine-HCl, pH 2.2) for 30 min or with 0.9 mL alkaline elution buffer (20 mM NaOH/100 mM NaCl, pH 9.2) for 10 min, respectively. Elution by pH extremes was followed by neutralisation of acidic eluate with 80 µL of unbuffered 1 M Tris or by neutralisation of basic eluate with 100 µL of 1 M Tris-HCl (pH 7.0). The neutralised acidic and basic eluates were used to separately infect 10 mL of exponentially growing TG1 culture for 30 min at 37°C, apart from 20 µL of each eluate that was used for titration of PPs as described in section 2.2.3.3.1. For elution with host (*E. coli*) cells, samples were mixed with 10 mL of exponentially growing TG1 culture and incubated for 30 min at 37°C without shaking to allow infection. An aliquot of 100 µL was taken for titration and the remainder of the infected TG1 was subjected to the amplification of eluted library pool.

Following the 30 min infection at 37°C, the transduced cells from each of the three modes of elution were mixed with 90 mL of fresh 2×YTCm25 medium and incubated at 37°C for 8 h with aeration to amplify the eluted library pool. An aliquot of amplified cells (1 mL) was subsequently infected with the VCSM13 helper phage to produce PPs using the liquid method (see section 2.2.3.1.1) that were used as an input for the next round of panning.

In total, four rounds of panning of the metagenomic phage display library on complex carbohydrate substrates were performed. Enrichment of the recombinant PPs displaying

potential substrate-binding fusions was monitored through increase of the output/input ratio over the four rounds of panning.

Furthermore, plasmid profiles of enriched library pools after each round of panning were monitored by agarose gel electrophoresis. Disappearance of a smear of recombinant plasmids, seen in the library pool prior to panning due to random insert distribution, and the appearance of discrete bands were taken as indicators of an enrichment of particular recombinant phagemids relative to the rest of the library.

The most prominent plasmid bands, identified after the first and last round of panning, were purified from the agarose gel and used to separately transform electrocompetent TG1 cells as described in section 2.2.2.4. Bacterial colony PCR was carried out on 20 randomly picked colonies from each transformation to examine the insert size distribution, as described in section 2.2.2.5.1. The recombinant phagemids containing inserts of different sizes were further analysed by DNA sequencing.

2.2.7.3 Sequence analysis of the affinity selected recombinant PPs

Taxonomic assignments of sequenced inserts obtained from affinity selected clones were based on the best BLASTX hit (E-value <1e-05 and query coverage >30%). Putative proteins in frame with the c-myc and pIII, encoded by affinity selected inserts, were annotated based on their best BLASTP hit (E-value <1e-05 and query coverage >30%) against the NCBI nr protein database. The putative proteins longer than 24 amino acid residues were further analysed for the presence of the membrane-targeting signals as described in section 2.2.6.1. The putative proteins were also inspected for putative CAZyme modules using a HMM-based search *via* the dbCAN web server and a BLASTP search against CAZymes in the dbCAN and CAT databases with default cut-offs.

Phage display recombinant clones that do not specifically bind to the target of interest, but are rather selected due to a competitive advantage (preferential replication and assembly during amplification phases of the panning protocol) or due to a substrate-unrelated binding (caused by the affinity of displayed peptide for the matrices used for immobilisation of the target or various commonly used blocking agents, such as bovine serum albumin or skim milk powder), are often isolated in affinity screens. A suite of web tools called SAROTUP [408], designed to identify peptides that provide competitive advantage, as well as those providing substrate-unrelated binding, was used for scanning of translated inserts, purified and sequenced from recombinant phagemid clones obtained after the affinity screening of the metasecretome library on complex carbohydrates. Short regions of putative proteins (40 amino acid residues

long with 10 amino acid residues long overlapping region) were analysed *via* the SAROTUP database [408].

To investigate whether –1 programmed ribosomal frameshift (PRF) events could have contributed to inserts encoding putative proteins out of frame with pIII being displayed (and thus captured by affinity selection), all inserts were inspected for putative signals stimulating –1 PRF using the KnotInFrame algorithm [407].

2.2.7.4 Wheat arabinoxylan-binding assay of affinity-selected recombinant PPs

The binding of five individual recombinant PPs that were selected from panning the metagenomic shotgun phage display library on AXYL (described in section 2.2.7.2) was assayed on the same substrate; the PPs derived from empty vector (pDJ01) were used as a control. The PPs were produced with the VCSM13 helper phage using the liquid method, precipitated, purified and titred as described in sections 2.2.3.2 and 2.2.3.3.1.

The affinity-binding assay was performed under the same binding and washing conditions that were used in the metagenomic library panning (section 2.2.7.2), except that elution was performed only by direct infection of the TG1 host cells. The output/input ratio of the recombinant PPs was compared to that of the vector control after one round in order to determine whether the displayed proteins have an increased affinity for the substrate.

Chapter 3. Metasecretome-selective phage display approach for mining the functional potential of a rumen plant-adherent microbial community

The reticulo-rumen is the fermentative forestomach of ruminant animals, and it harbours a complex microbial ecosystem. Plant fibre-degradation is the most prominent feature of the rumen microbiome, and this activity allows ruminants to utilise the lignocellulosic components of the forage plant cell walls as an energy source, and has tremendous impact on ruminant nutrition and productivity [431].

The activities of rumen microbes required for utilising feed material are mediated *via* surface, secreted and transmembrane proteins, known collectively as the secretome, or the metasecretome at the scale of the whole microbial community. Surface proteins are thought to mediate the adherence of microbes to feed material, while secreted enzymes and carbohydrate binding proteins are responsible for the initial steps of the plant fibre degradation [60, 63]. For this reason, it is expected that the rumen metasecretome represents a valuable repository of novel proteins involved in lignocellulolytic bioactivities.

One of the aims of this project was to develop a new metasecretome phage display approach that could specifically enrich for genes encoding the metasecretome proteins at the scale of the complex microbial community. To maximise the probability of identifying the metasecretome proteins involved in fibre degradation, a plant-adherent fraction of the rumen microbial community from a pasture-fed cow, considered to be rich in fibrolytic activities [60, 63], was used as a source of metagenomic DNA for construction of the metasecretome phage display libraries.

The metasecretome phage display was combined with next-generation sequencing and sequence-based metagenomic analyses to explore the metasecretome of the bovine rumen plant-adherent microbial community captured through selection, especially the diversity of hemicellulose-degrading and carbohydrate-binding activities.

3.1 Construction of rumen plant-adherent metagenomic libraries and metasecretome selection

The inserts for the metasecretome library were prepared from the microbial fraction that was tightly adherent to plant biomass, derived from the whole rumen contents of one dairy cow (Friesian, fed mainly ryegrass/clover pasture) as described in section 2.2.4.2. High molecular

weight metagenomic DNA was isolated from the adherent microbial fraction (section 2.2.4.3) and its integrity and quality were confirmed by pulsed-field gel electrophoresis (Figure 3.1 A). To prepare the metagenomic DNA for construction of shotgun phage display libraries in a phagemid vector, the metagenomic DNA was further mechanically sheared, end-repaired and size-fractionated as described in section 2.2.5.1. The resulting fragments ranged in size from 0.7 to 5 Kb (Figure 3.1 B).

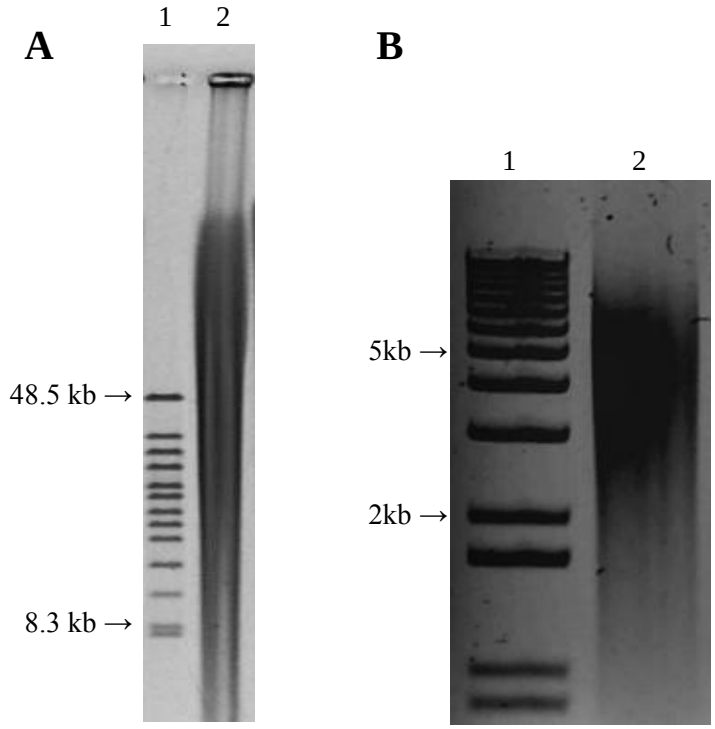

Figure 3.1 Intact and mechanically sheared metagenomic DNA isolated from the plant-adherent rumen microbial community.

A. Pulsed-field gel electrophoresis (PFGE) of intact metagenomic DNA (performed by Dong Li, AgResearch), lanes: 1, PFGE standard (BioRad); 2, intact metagenomic DNA. **B**. Agarose gel electrophoresis of fragmented metagenomic DNA, lanes: 1, 1 Kb Plus DNA ladder (Life Technologies); 2, mechanically sheared (10 psi for 1 min) and size-fractionated metagenomic DNA.

The fragmented and repaired DNA was used to prepare a shotgun metagenomic phage display library, which was further subjected to a secretome selection protocol (Figure 3.2). The secretome selection is based on the requirement that a membrane-targeting sequence be fused to the virion protein pIII, encoded by a phage display phagemid vector pDJ01, in order to assemble structurally stable PPs (resistant to sarkosyl), when a VCSM13d3 helper phage (carrying a *gIII* deletion) is used. The membrane-targeting signal is supplied within the library inserts derived from the secretome proteins, provided they are in-frame with the vector-encoded pIII that does not have a signal sequence.

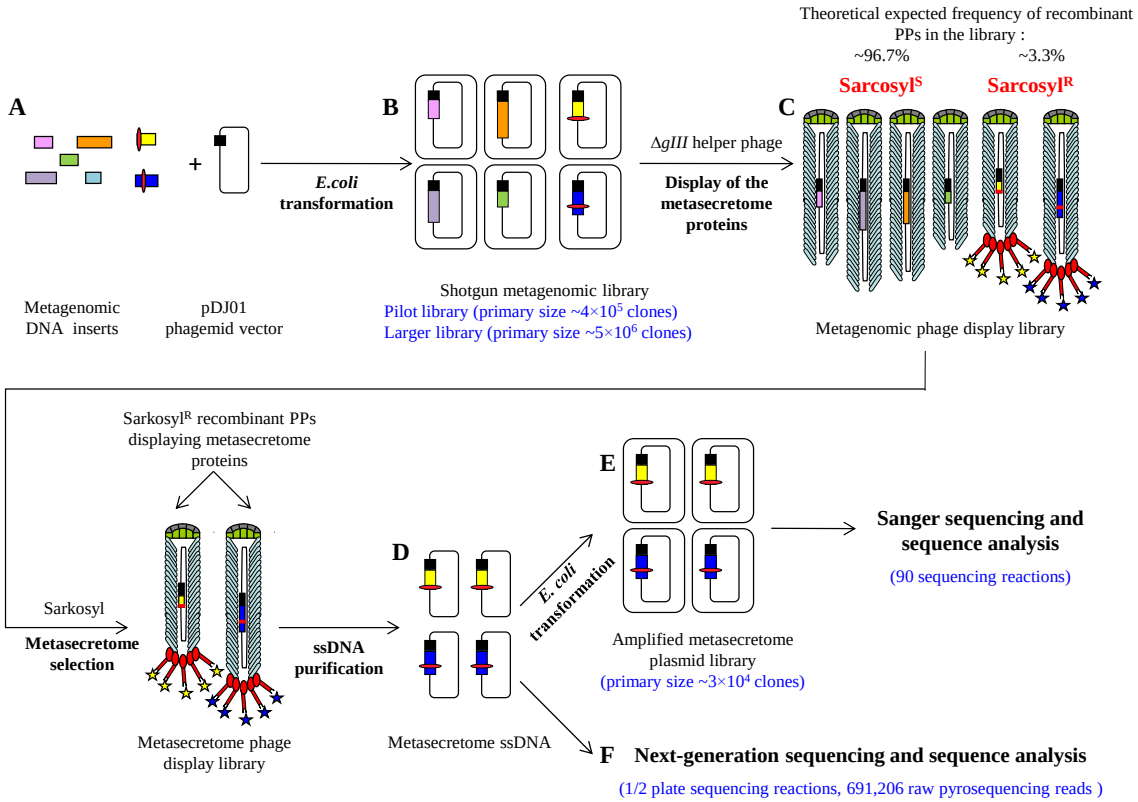

Figure 3.2 Overview of results of metasecretome library construction and selection.

The numbers in brackets refer to data obtained in this work. **A**. Shotgun metagenomic libraries were constructed by cloning metagenomic DNA into the pIII-cloning cassette of the pDJ01 phagemid vector, which does not contain a signal sequence. Note that only some metagenomic inserts contain endogenous signal sequences, which are represented by red ovals. **B**. Recombinant phagemids replicate as plasmids inside the cells, and in the presence of the helper phage, they are packaged as recombinant virions. **C**. After infection of the shotgun metagenomic library with the *gIII*-deleted helper phage VCSM13d3, the resulting metagenomic phage display library contains a mix of virions capped by insert-pIII fusion proteins (signal sequence-positive clones), that are resistant to sarkosyl (Sarkosyl^R), and uncapped virions (signal sequence-negative clones), that are sensitive to sarkosyl (Sarkosyl^S). Sarkosyl resistance is used as a basis for selection. **D**. Single-stranded DNA (ssDNA) purified from Sarkosyl^R particles after selection was used to transform *E. coli* to obtain an amplified metasecretome plasmid library for preliminary assessment of metasecretome diversity by Sanger sequencing of clone inserts (**E**), and as a template for in-depth assessment of metasecretome selection by next-generation shotgun sequencing using the 454 GS FLX Titanium platform (**F**).

The secretome-selective phage display has been used previously at a genomic scale, for *Lactobacillus rhamnosus* [234] and *Mycobacterium tuberculosis* [394]. However, it has never

been used on the scale of a complex microbial community. It was therefore essential to probe its versatility and stringency in selecting the secretome proteins from diverse bacterial species. A pilot library was initially constructed for the development and optimisation of the secretome display technology on a metagenomic scale, and for preliminary assessment of the metasecretome plasmid library diversity by Sanger sequencing.

First, a pilot shotgun metagenomic library, containing around 4×10^5 primary clones (before secretome selection) and, subsequently, a larger shotgun metagenomic library, containing around 5×10^6 primary clones were constructed by ligation of fragmented and end-repaired metagenomic DNA into the secretome-selective phagemid vector pDJ01. The library ligations were transformed into the *E. coli* strain TG1 and amplified using the plasmid origin of replication as described in section 2.2.5.1 (Figure 3.2 A-B).

Both pilot and large metagenomic shotgun libraries were separately subjected to the secretome selection protocol, as described in section 2.2.5.2. To initiate the secretome-selection aimed at enriching the metagenomic inserts encoding the secretome proteins (referred to in this thesis as the metasecretome), each amplified primary shotgun metagenomic library was infected with the VCSM13d3 helper phage. The VCSM13d3 helper phage provided proteins for replication of the phagemid from the phage origin of replication, and packaging into recombinant PPs, except for pIII, which had to be supplied from the phagemid in order to assemble sarkosyl-resistant ($Sarkosyl^R$) PPs. The assembled PPs were purified and subjected to sarkosyl selection as described in section 2.2.5.2 (Figure 3.2 C). The two-step elimination of sarkosyl-sensitive recombinant PPs consisted of a sarkosyl-induced release of recombinant phagemid ssDNA that lacked a membrane-targeting signal in frame with pIII, followed by elimination of the released DNA with DNaseI. The efficiency of these two steps of the selection process were monitored by native gel electrophoresis (as described in section 2.2.3.5), and it was verified that defective ($Sarkosyl^S$) particles were disrupted during sarkosyl selection (data not shown).

The ssDNA, enriched for recombinant phagemids containing metagenomic inserts that encode the metasecretome proteins in-frame with vector-encoding pIII, was isolated from the $Sarkosyl^R$ PPs (Figure 3.2 D). The pilot library ssDNA was used to transform *E. coli*, followed by sequence analysis of the inserts of a sample of 90 individual transformants using the Sanger sequencing approach, to assess the presence and types of membrane-targeting signals, as well as the taxonomic diversity of the selected putative metasecretome proteins (Figure 3.2 E). For in-depth assessment of the DNA sequences encoding putative metasecretome proteins captured through selection, ssDNA purified from the $Sarkosyl^R$ PPs obtained from the larger shotgun metagenomic library was used as a template for pyrosequencing (Figure 3.2 F).

3.2 Pilot metasecretome phage display library

3.2.1 Estimated enrichment of the secretome insert-containing recombinant library clones

For a secretome protein-encoding insert to be selected, it had to fulfil two conditions: i) to be translationally fused (i.e. in-frame) with the vector-encoded phage protein, pIII (which is devoid of a signal sequence); ii) to encode a membrane-targeting signal, which could target the insert-pIII fusion to the inner membrane of *E. coli*. These criteria were used to assess the enrichment for secretome-encoding inserts in the pilot library. Inserts containing ORFs encoding putative proteins in frame with pIII, and longer than 24 amino acid residues [432] were identified, and the presence of membrane-targeting signals was predicted using a suite of algorithms as described in section 2.2.6.1.

Some of the ORFs, containing the same point of fusion to gIII, were isolated multiple times among the analysed recombinant clones (e.g. one distinct putative pilin-encoding ORF was carried by 21.1% of all examined pilot library clones). Among the 90 inserts analysed, 55 distinct metasecretome ORFs were identified. Out of the 90 inserts, 85 (94.4%) contained 53 distinct ORFs encoding putative secretome proteins with typical membrane-targeting sequences in frame with pIII. One insert out of the remaining five (5.6%) contained an ORF encoding a polypeptide in frame with pIII that was shorter than 24 amino acid residues and was considered 'background'. The remaining four inserts contained identical ORFs without typical membrane-targeting sequences. Further analysis using the SecretomeP 2.0 software package [259], which discriminates between non-classically secreted proteins that lack the typical membrane-targeting signals and cellular proteins based on amino acid composition, secondary structure and disordered regions, gave a score of <0.5, indicating that putative protein encoded by this ORF is most likely a cytoplasmic protein. This was confirmed by a BLASTP search, which showed that the protein encoded by this ORF is homologous to a conserved hypothetical protein with a predicted cytoplasmic localisation. Therefore, these inserts were considered to be 'background', where the original defective PPs were not eliminated by selection.

Based on the average proportion of secretome ORFs in bacterial genomes (~20%) [27, 28], and the probability of the insert being in the appropriate orientation (50%) and fused with gene *gIII* to create an in-frame protein fusion with pIII (33.3%), it was expected that only ~3.3% of the inserts in the primary metagenomic library, prior to the secretome selection, encode secretome proteins. The frequency of secretome protein-encoding recombinant phagemids after selection was calculated as the proportion of inserts containing ORFs with predicted membrane-targeting signal compared to all sequenced inserts 85/90 (94.4%). The frequency of

recombinant phagemids encoding secretome proteins, obtained after selection, divided by the expected frequency before selection, gave an enrichment of 29-fold, indicating a high stringency of selection which had eliminated the majority of recombinant phagemids containing non-secretome encoding inserts.

3.2.2 Pilot metasecretome library secretion signal types, functional annotations and taxonomy

A complete list of the membrane-targeting signals predicted for the 55 putative proteins identified in the pilot metasecretome phage display library, their functional annotation and taxonomic assignments, is presented in Appendix 1.

The majority of ORFs (35) were predicted to encode type I signal sequences. Eight ORFs were predicted to encode N-terminal transmembrane α-helices (potential N-terminal membrane anchors), while six ORFs were predicted to encode either multiple or single internal transmembrane α-helices. Type II or lipoprotein signal sequences (predicted in three ORFs) and a single type IV (pilin-like) signal sequence made up the remainder. No Tat signal sequences were identified. The types of membrane-targeting signals predicted from the pilot metasecretome phage display library ORFs are presented in Figure 3.3.

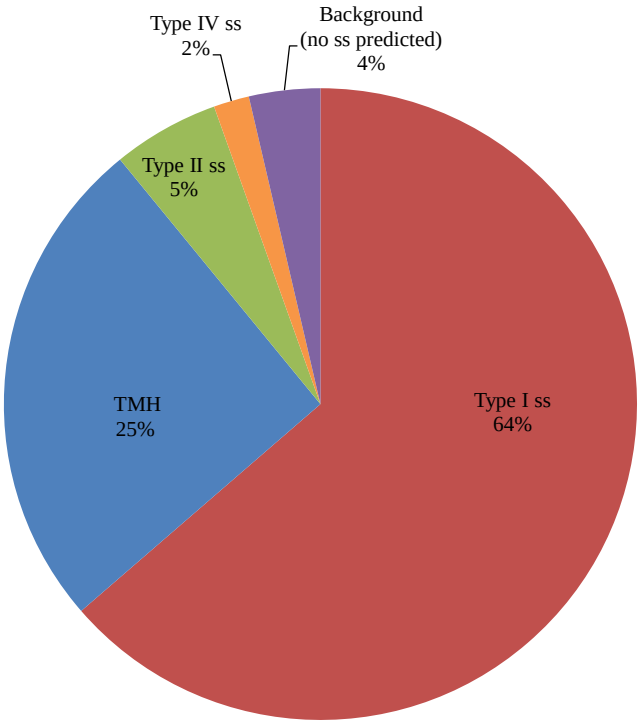

Figure 3.3 Types of membrane-targeting sequences detected in the pilot metasecretome library.

Abbreviations: Type I ss, classical signal sequence; Type II ss, lipoprotein signal sequence; Type IV ss, pilin-like signal sequence; TMH, N-terminal or internal transmembrane α-helix/helices; background, ORFs encoding putative proteins without a predicted membrane-targeting signal/non-classical secretion, or ORFs encoding putative proteins and peptides ≤24 amino acid residues, that were excluded from the analysis as they were deemed too short to be informative.

Functional annotation of the 55 putative proteins identified in the pilot library, based on their best BLASTP hit against the NCBI nr protein sequence database with an E-value <1e-05 and a query coverage >30%, resulted in a large number of assignments to hypothetical (30) and conserved hypothetical proteins (17). Five ORFs were annotated to encode putative enzymes (three of which may be involved in carbohydrate metabolism), while the remaining three ORFs were predicted to encode putative proteins involved in host/microbe interactions, signal transduction and sporulation (Figure 3.4).

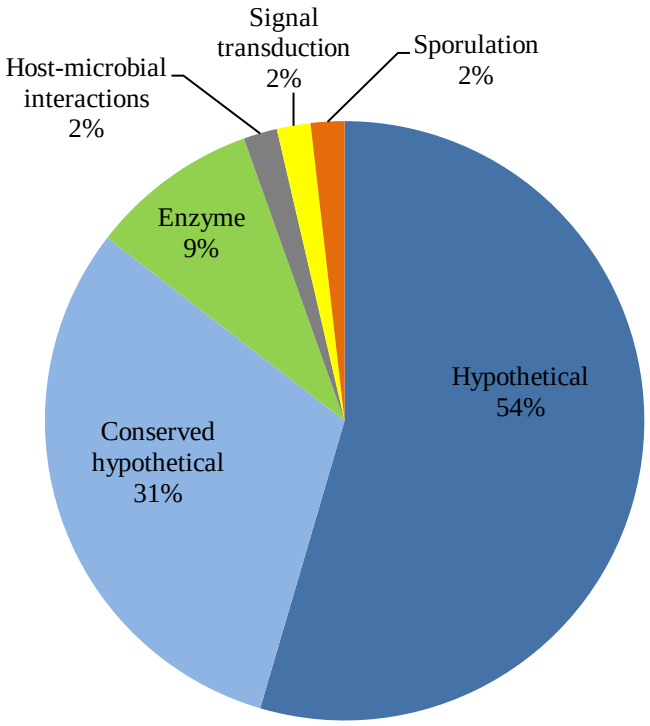

Figure 3.4 Functional annotation of putative proteins in the pilot metasecretome library.

Functional annotation of putative proteins was based on best BLASTP hit against NCBI nr protein database (E-value <1e-05 and query coverage >30%) and, in case their best BLASTP hits were to hypothetical proteins, putative proteins were annotated as conserved hypothetical. Putative proteins without significant similarity to proteins in the NCBI nr database at set E-value cut-off were annotated as hypothetical.

To identify the organisms from which the metasecretome library inserts were derived, taxonomic assignments of the metasecretome inserts were made based on their best BLASTX hits against the NCBI nr protein sequence database with an E-value of <1e-05 and a query coverage of >30%. The most abundant assignments were to the genera *Prevotella* (13%), *Clostridium* (10%), *Butyrivibrio* (7%), *Ruminococcus* (6%), *Bacteroides* (6%) and *Fibrobacter* (4%); genus-level assignments could not be made for 50% of the inserts analysed. At the phylum level, the majority of inserts were assigned to Bacteroidetes (23%), Firmicutes (22%) and Fibrobacteres (5%). The taxonomic representation with regard to Gram-staining properties was 28% of inserts from Gram-negative bacteria, 22% from Gram-positive bacteria and 50% were unclassified (Figure 3.5).

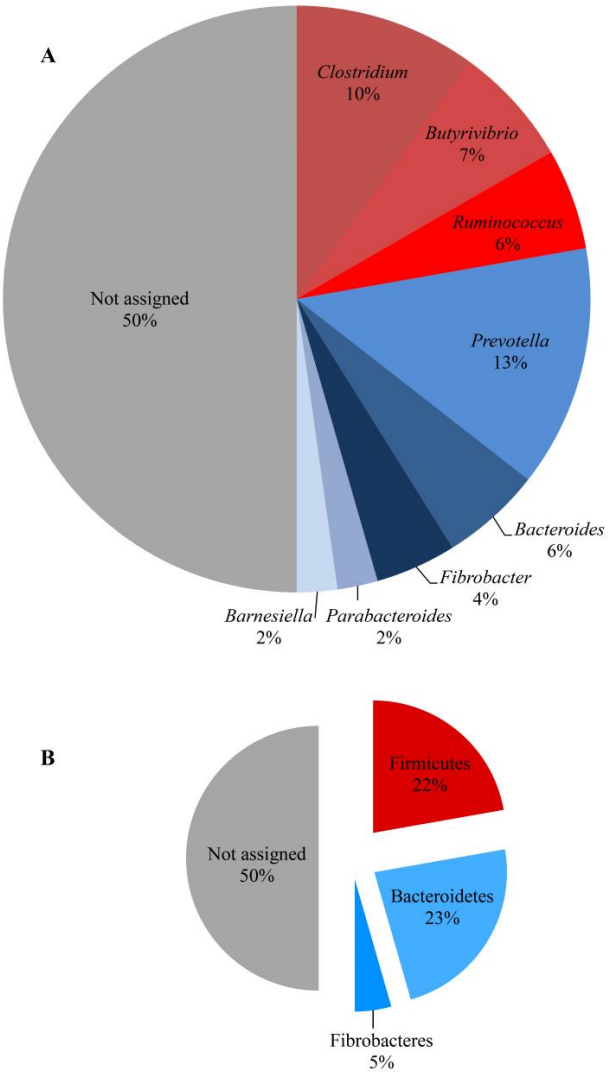

Figure 3.5 Taxonomic distribution of rumen microbial inserts from the pilot metasecretome library.

Taxonomic categories were assigned to inserts based on their best BLASTX hit (E-value <1e-05 and query coverage >30%) at the genus (**A**) and phylum level (**B**). 'Not assigned' corresponds to inserts that had best BLASTX hits below the thresholds used.

3.2.3 Overview of sections 3.1 and 3.2

Small scale metasecretome selection in combination with sequencing of limited number of rumen microbial plant-adherent metasecretome library clones (pilot study) demonstrated high rate of enrichment for inserts encoding metasecretome proteins (96%) containing diverse types of membrane-targeting signals. Functional and taxonomic diversity of the metasecretome library inserts captured through selection was representative of the rumen microbiome. However, observable bias towards representation of Gram-negative bacteria, as well as overrepresentation

of certain clones in the library, renders approach used in the pilot study unsuitable for larger scale analysis of the rumen metasecretome.

In conclusion, the pilot study enabled the development and optimisation of the secretome-selective phage display protocols and showed that the secretome selection at the metagenomic scale is feasible, provided that the primary library size be upscaled and bottlenecking of selection by transformation of *E. coli* with single-stranded DNA avoided.

3.3 Metasecretome characterisation by next-generation sequencing

To improve the representation of the rumen microbial metasecretome captured through selection over that seen in the pilot library (section 3.2), a shotgun metagenomic library (5×10^6 primary clones; twelve-fold larger in comparison to the pilot library) was constructed and subjected to sarkosyl/DNaseI secretome selection (Figure 3.2 A-D). However, in contrast to the pilot library, the upscaled metasecretome phage display library inserts were analysed by next-generation sequencing for in-depth assessment of the diversity of the selected metasecretome (Figure 3.2 F).

This approach was chosen to allow for more sequence information available from the library to be extracted and analysed, and to bypass the observed low transformation efficiency of ssDNA (data not shown). In addition, this approach avoids a bias in replication and assembly of PPs observed in the pilot experiment, which resulted in some highly overrepresented metasecretome library clones (e.g. clones encoding putative pilin fragment).

3.3.1 Establishing a protocol for preparing the pyrosequencing template from the metasecretome phage display library

The upscaled ligation, amplification and selection were carried out in 27 separate aliquots to avoid domination of fast-assembling and fast-replicating recombinant phagemids (as described in section 2.2.5.1). The starting DNA sample to analyse the metasecretome phage display library inserts by next-generation sequencing was the metasecretome-enriched ssDNA pool, purified from the secretome-selected (Sarkosyl^R) phage display library PPs. To eliminate most of the vector sequences from this sample for pyrosequencing using the Roche 454 GS FLX Titanium platform, the inserts were amplified by PCR. The primers were designed to anneal 361 bp upstream and 367 bp downstream of the cloning site (insert ends), to ensure the inclusion of shorter inserts (<600 nt) for sequencing, given that DNA fragments in the 0.6 – 0.8 Kb size range are preferentially used as a sequencing template for the 454 pyrosequencing. The PCR

amplification of the recombinant phagemid inserts using these primers gave products in the size range from 0.7 to 5 Kb, and thus, these required further fragmentation to provide the optimal size range for sequencing.

Different DNA fragmentation approaches have different limitations and biases, which can be carried through to the subsequent steps of the sequencing protocol, and affect the sequencing output. For this reason, a range of conditions for both enzymatic and mechanical fragmentation were trialled on a test-mix of amplicons ranging from 0.7 to 5 Kb in size (obtained from the pilot library recombinant phagemids using the same pair of primers as for pyrosequencing template preparation) as described in section 2.2.5.4, to enrich for fragments in the 0.6 – 0.8 Kb size range, and in a manner that allows adequate representation of the starting material.

Prolonged mechanical shearing in nebulisers led to extensive sample loss *via* aerosols and did not result in breakdown of fragments shorter than 1.5 Kb, while the needle shearing was not suitable for the relatively small size range required. Digestion with *Alu*I restriction endonuclease (with expected average frequency of its 4 nucleotide long recognition site AGTC once in every 256 nucleotides) yielded fragments mainly in the 0.5 – 0.85 Kb size range, which is suitable as a sequencing template. However, complete digestion with *Alu*I removed the vector flanks due to the presence of three *Alu*I recognition sites in the approximately 730 bp long vector sequence flanking the insert. As this would result in excluding very short fragments during the size-fractionation step of the pyrosequencing template preparation by the sequencing facility, other strategies for template fragmentation were investigated. DNaseI digestion in the presence of Mn^{2+} ions was, unlike *Alu*I digestion, non-specific in the terms of the cleavage target, but resulted in a wider range of fragment sizes (0.1 – 3 Kb) and showed low reproducibility, hence it was not used further.

Based on these observations, a combination of mechanical shearing by nebulisation and enzymatic digestion with *Alu*I restriction endonuclease was applied to prepare the pyrosequencing template (as described in section 2.2.5.5). Mechanical shearing was applied to ensure that the long fragments (>1.5 Kb) were randomly broken, while the *Alu*I digestion was applied under two different conditions, to ensure both complete and partial digestion of the fragments resistant to mechanical shearing and promote generation of a pool of overlapping fragments falling in the size range suitable for pyrosequencing.

3.3.2 Preparation of metasecretome template for next-generation sequencing

The template for next-generation sequencing was prepared by PCR amplification followed by mechanical shearing and *Alu*I fragmentation under the conditions established in

section 3.3.1 (also see section 2.2.5.5 for experimental details). The PCR amplification was performed separately for 27 aliquots of metasecretome-selected ssDNA (derived from the original 27 separate ligation reactions) to minimise potential biases in PCR amplification within each independent reaction.

The amplicons from the 27 separate PCR reactions were subjected to gel electrophoresis. Dominant bands, possibly stemming from the PCR amplification bias, were noticeable in each of the fractions, but the banding patterns of the dominant bands in each of the fractions were different (Figure 3.6).

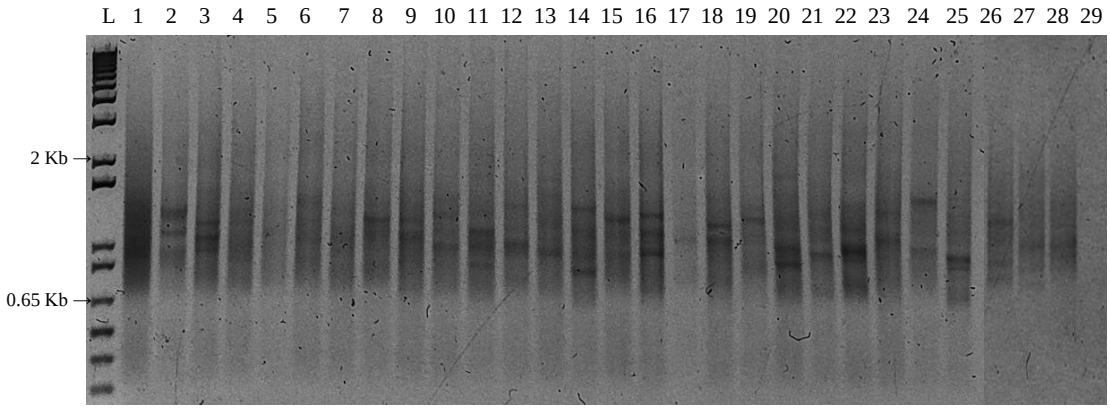

Figure 3.6 PCR amplification of metasecretome-enriched ssDNA.

PCR amplification was performed separately for 27 aliquots of metasecretome-enriched ssDNA using a pair of primers designed to anneal 361 bp upstream and 367 bp downstream of the insert. Lanes: L, 1 Kb Plus DNA ladder (Life Technologies); 1, a pool of all 27 PCR reactions; 2 – 28, PCR amplification reactions 1 – 27; 29, PCR reaction without a template (PCR negative control).

Sequencing of the gel-extracted DNA from two randomly selected dominant bands (~2 Kb and ~1.5 Kb in size) in one of the PCR reactions showed the presence of multiple chromatogram traces, confirming that the single band on a gel contains multiple amplicons.

Amplicons from the 27 separate PCR reactions were pooled before further processing to minimise the loss of material in subsequent steps. Pooled amplicons ranged in size between 0.7 and 5 Kb (with the majority <2 Kb) and were further fragmented enzymatically (by partial or complete *Alu*I digestion), by mechanical shearing, or by combined mechanical and enzymatic fragmentation, under the conditions described in section 2.2.5.5. The template for pyrosequencing was obtained by combining equal amounts of DNA fragmented under all five fragmentation conditions (Figure 3.7 A and B). Two prominent bands visible in *Alu*I treatments, 200 bp and 350 bp in length, correspond to completely digested vector sequence flanking the inserts (Figure 3.7 C).

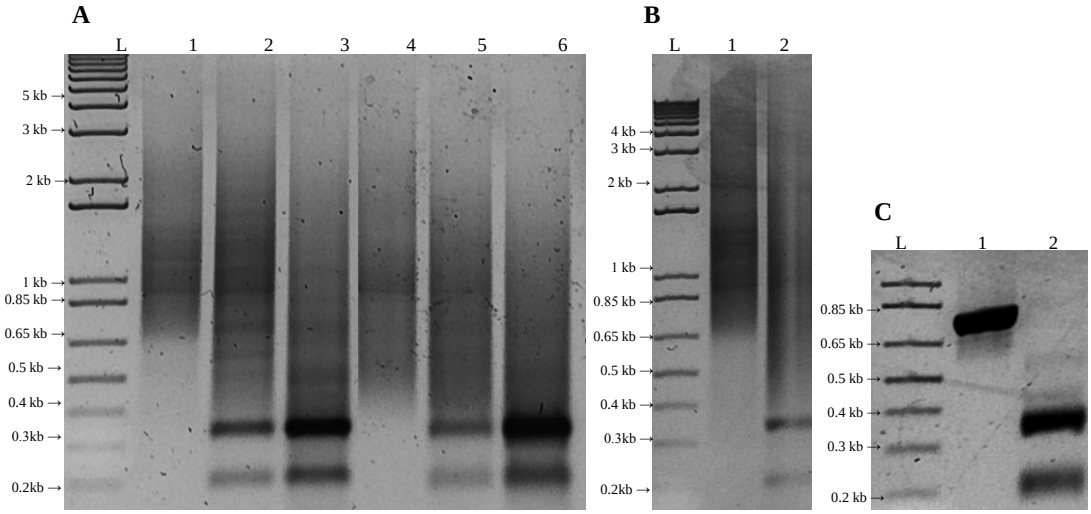

Figure 3.7 PCR amplicons of the inserts of metasecretome-enriched PPs, processed by enzymatic and mechanical shearing to obtain the pyrosequencing template.

Samples were loaded onto 1% agarose gel containing SYBR® Safe DNA Gel Stain (Life Technologies). A. Metasecretome amplicons before and after enzymatic and mechanical shearing. Lanes: L, 1 Kb Plus DNA ladder (Life Technologies); 1, metasecretome PCR amplicons; 2 – 6 metasecretome PCR amplicons subjected to various fragmentation conditions: 2, *Alu*I digestion for 1 min; 3, *Alu*I digestion for 3 h; 4, mechanical shearing for 6 min; 5, mechanical shearing followed by 1 min *Alu*I digestion; 6, mechanical shearing followed by 3 h *Alu*I digestion. B. Comparison of starting metasecretome amplicons and final pyrosequencing template. Lanes: L, 1 Kb Plus DNA ladder; 1, metasecretome PCR amplicons; 2, pyrosequencing template (pooled fractionated metasecretome PCR amplicons). C. Completely *Alu*I-digested vector sequence. Lanes: L, 1 Kb Plus DNA ladder; 1, PCR amplicons derived from vector pDJ01 (PCR positive control); 2, *Alu*I digested PCR positive control.

3.3.3 Next-generation sequence analysis of the metasecretome phage display library

Shotgun sequencing of the metasecretome-enriched DNA sample, obtained from the metasecretome phage display library, yielded 691,206 raw metasecretome pyrosequencing reads (Figure 2.3, dataset A). Trimming off the vector and helper phage sequences, and subsequent filtering out the low-quality and short reads (<100 bp), resulted in 492,198 cleaned unassembled metasecretome reads (Figure 2.3, dataset B) that were either processed *via* the IMG/M pipeline [400] or assembled (Figure 2.3, dataset D). The processing and quality filtering of the cleaned metasecretome reads in the IMG/M system, including de-replication step, resulted in 153,002 cleaned de-replicated unassembled metasecretome reads (Figure 2.3, dataset C). Approximately two thirds of raw reads were lost in the process of de-replication.

Summary statistics of the unassembled and assembled metasecretome datasets are represented in Table 3.1 and Table 3.2, respectively.

Table 3.1 Summary statistics for the unassembled metasecretome datasets.

	Raw MS pyrosequencing reads [a]	**Cleaned unassembled MS reads [b]**	**Cleaned unassembled de-replicated reads [c]**
Number of reads	691,206	492,198	153,002
Average read length (bp)	548	322	362
Total sequence information (Mb)	379	158	55

Abbreviations: MS, rumen microbial plant-adherent metasecretome. [a] Raw reads were obtained by shotgun sequencing of the metasecretome-enriched DNA from the rumen plant-adherent microbial fraction with the Roche 454 GS FLX Titanium platform using half a plate. [b] Cleaned reads were obtained by trimming the phagemid vector and helper phage sequences, and filtering out short reads (<100 bp) using SeqClean. [c] Cleaned unassembled reads after processing, including trimming (resulting in 492,197 reads), low complexity filtering (resulting in 491,963 reads) and de-replication (resulting in 153,502 reads) *via* the IMG/M pipeline.

Assembling cleaned metasecretome reads (Figure 2.3, dataset D), using GS Roche *De Novo* Assembler version 2.7 [401] as described in section 2.2.6.2.1, resulted in 3,574 contigs containing 1.7 Mb of consensus sequence (Table 3.2).

Table 3.2 Summary statistics of the assembled metasecretome dataset.

Read status [a]	**Number of reads**	**% of total reads used by assembler**
Assembled [b]	205,339	41.72
Partial [c]	186,930	37.98
Singleton [d]	54,433	11.06
Repeat [e]	119	0.02
Outlier [f]	45,371	9.22
Too short [g]	0	0.00
Total reads used by assembler	492,192	100
Total bases in consensus sequence (Mb)	1.74	
Number of contigs	3,574	
Number of large contigs (>500bp)	1,400	
N50 (bp) [h]	771	
Largest contig (bp)	2,225	

[a] Read status reported by the GS Roche *De Novo* Assembler for assembly of cleaned metasecretome reads with the overlap requirements of a minimum length of 40bp and a minimum identity of 100%. [b] Assembled, the read is fully incorporated into the assembly. [c] Partial, only part of the read meeting the set overlap requirements was included in the assembly. [d] Singleton, the read did not overlap with any other reads in the input and was not incorporated in the assembly. [e] Repeat, the read is deemed to be from repeat regions and was not incorporated in the assembly. [f] Outlier, the read was identified by the assembler as problematic and was not incorporated in the assembly. [g] Too short, the read was too short to be used in the assembly. [h] N50, a half of all bases from the input dataset reside in contigs of this size or longer.

Around 41.7% of the reads were fully and 38% were partially incorporated into the assembly. Success of assembling reads into longer contigs (>500bp) was limited. A high proportion of replicates in the metasecretome dataset led to replicates 'collapsing' into contigs that are not much longer than unassembled reads and leaving high number of unassembled 'singletons'. For this reason, the assembled metasecretome dataset was used only for the prediction of membrane-targeting signals, while the cleaned de-replicated unassembled metasecretome dataset was used in all subsequent analyses.

3.3.4 Prediction of common types of membrane-targeting signals in the putative metasecretome proteins in frame with pIII

The pyrosequencing template was derived by fragmentation of PCR products amplified from the pool of metasecretome-enriched recombinant phagemid ssDNA. These amplicons contained, in addition to the recombinant phagemid insert, around 360 bp of vector sequence flanking each side of the insert. Due to the random nature of shearing, it is expected that a proportion of the pyrosequencing reads contain insert sequence only, while a proportion contain also vector sequence downstream of the insert (including the c-myc tag and pIII-encoding region) or upstream of the insert. Out of the 691,206 raw reads, we have identified 283,574 reads that contained the vector sequence downstream of the insert as described in section 2.2.6.2.5 and 278,430 of these reads encoded putative proteins in frame with pIII. Clustering of putative 'in-frame' proteins at 100% sequence identity threshold to remove duplicity resulted in 85,239 distinct putative proteins, 75,059 (~88%) of which were longer than 24 amino acid residues. These putative proteins were analysed for the presence of the three types of membrane-targeting signals (type I and type II signal sequences, and transmembrane α-helices), that were most commonly detected in the pilot library as described in section 2.2.6.2.5.

Because of the fragmentation step in the pyrosequencing template preparation, it is likely that for some partial ORFs in frame with *c-myc* and *gIII*, the N-terminal part of the ORF,

encoding the membrane-targeting signal, is found on a separate read after fragmentation. For this reason, membrane-targeting signal prediction was based on a sequential combination of direct predictions for 'in-frame' putative proteins and predictions assigned based on sequence similarity between 'in-frame' putative proteins without detected membrane-targeting signals, and longer homologous putative proteins in the metasecretome assembled dataset (as described in section 2.2.6.2.5).

Around 46% of the putative proteins had a predicted type I signal sequence; 2% had type II (lipoprotein) signal sequences and 11% had one or more transmembrane helices. For around 28% of the putative proteins, none of these three types of membrane-targeting signals were detected, while for 13% of the putative proteins, the presence of membrane-targeting signals could not be determined by either direct or indirect methods, due to the lack of significant similarity to predicted proteins encoded by the assembled metasecretome dataset (Figure 3.8).

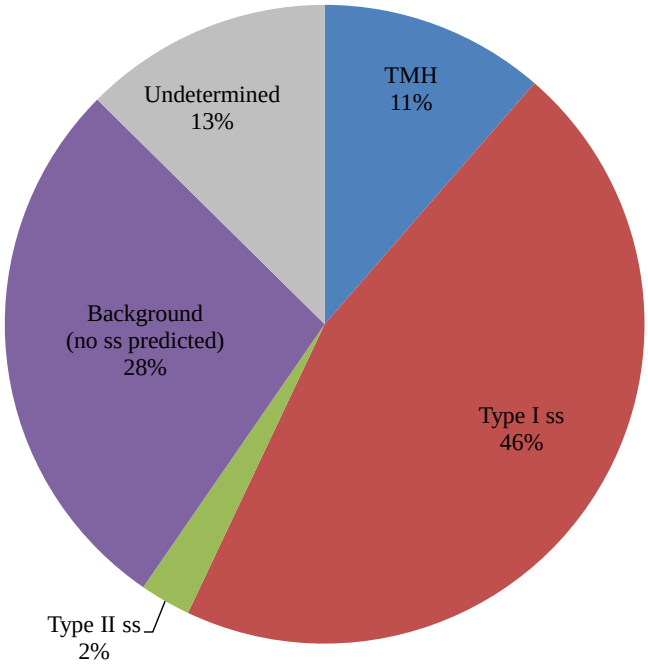

Figure 3.8 Common types of membrane-targeting signals detected in putative proteins in-frame with pIII in the metasecretome-enriched dataset.

Abbreviations: ss, signal sequence; Type I ss, classical signal sequence; Type II ss, lipoprotein signal sequence; TMH, N-terminal or internal transmembrane α helix/helices; Background, putative proteins without predicted membrane-targeting signals; Undetermined, putative proteins without significant similarity to predicted proteins encoded by the assembled metasecretome dataset.

Due to experimental procedure used to process the DNA template for pyrosequencing, the proportion of putative proteins with membrane-targeting signals in the pyrosequencing dataset is not a true measure of the stringency of selection prior to shearing.

3.3.5 Phylogenetic profile of the metasecretome dataset

In addition to the cleaned unassembled metasecretome reads, a metagenome dataset (Moon *et al.*, unpublished) was included in the bioinformatics analyses to provide a reference point for comparison to the metasecretome dataset (Table 2.9). This dataset was obtained by shotgun sequencing of the total metagenomic DNA from the plant-adherent rumen microbial communities of two New Zealand cows on a similar diet to the cow used for the metasecretome library analysis, using the Roche 454 GS FLX platform (one plate per cow sample). Both metasecretome and metagenome datasets were processed and automatically annotated *via* IMG/M system [400] using the same parameters (as described in section 2.2.6.2.2) and their putative functions and phylogenetic profiles were compared between the two datasets.

The phylogenetic distribution of protein-coding genes in the two datasets was also obtained *via* the IMG/M system using a similarity-based binning approach (see section 2.2.6.2.2). The accuracy of this approach depends on the availability of appropriate reference data, which is currently limited for rumen microbial metagenome samples. For this reason, the phylogenetic distribution of the best BLASTP hits of the protein-coding genes predicted in the metasecretome and metagenome dataset was reported at the level of phyla, with a 30% identity cut-off (Figure 3.9).

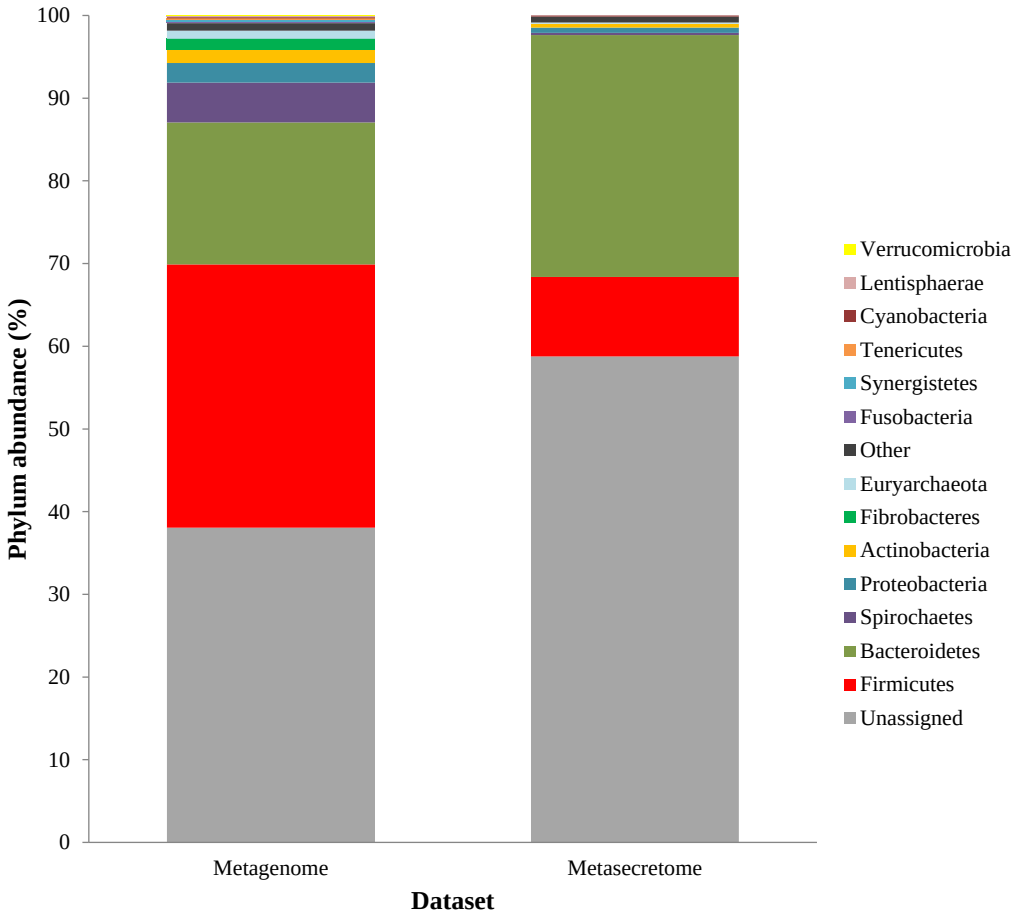

Figure 3.9 Phylogenetic distribution of putative protein-coding genes in the metagenome and the metasecretome dataset.

The taxonomic assignments, at the phylum level, were based on the distribution of the best BLASTP hits at a 30% amino acid sequence identity threshold for protein-coding genes predicted in metagenome (N=671,876) and metasecretome (N=222,960) datasets. Each section of the stacked columns represents the percentage of total protein-coding genes assigned to the corresponding phylum. The section labelled 'Other' contains putative protein-coding genes assigned to a phylogenetic group with low abundance in the dataset (<0.1%), while the section labelled 'Unassigned' corresponds to putative protein-coding genes with best BLASTP hit below 30% identity cut-off.

At the domain level, the majority of assignments in the metagenome and metasecretome datasets were to Bacteria (60.7% and 40.9%, respectively), Archaea (1% and 0.2%, respectively) and Eukaryota (0.2% and 0.1%, respectively). In both datasets, of all the sequences assigned to Eukaryota, approximately one third were most similar to fungi and around 12% – 14% to plants. Assignments to viruses were rare (0.06% and 0.004%, respectively), while around 38.1% of putative protein-coding genes in the metagenome and 58.8% in the metasecretome dataset remained unassigned.

At the phylum level (Figure 3.9), the main assignments in the metasecretome dataset were to Bacteroidetes (29.2%) and Firmicutes (9.6%), with minor contributions from different divisions of the Proteobacteria (0.64%), Actinobacteria (0.45%), Spirochaetes (0.27%), Euryarchaeota (0.16%) and Cyanobacteria (0.14%). In contrast, in the metagenome dataset, Firmicutes (31.9%) was the predominant phylum, followed by Bacteroidetes (17.2%), Spirochaetes (4.8%), Proteobacteria (2.4%), Actinobacteria (1.6%), Fibrobacteres (1.3%), Euryarchaeota (1%), and a larger number of other phyla, each with less than 1% of assigned reads.

3.3.6 Functional annotation of the metasecretome dataset

To determine the putative functions enriched in the metasecretome library, functional annotations of protein coding genes in the metasecretome and metagenome shotgun sequencing datasets were compared. Annotation *via* the IMG/M system [400] resulted in 35% and 49% Pfam [433] assignments of the total protein coding genes in the metasecretome and metagenome datasets, respectively. In this subset of protein-coding genes, 32% and 44% could be further categorised into COG-based functional categories in the metasecretome and metagenome datasets, respectively.

Functional annotations based on Pfams [433] assigned to predicted protein-coding genes and classified into COG-based functional categories were compared between the two datasets to look at the effect of the metasecretome enrichment on the abundance of each of the functional categories (Figure 3.10). The ‘carbohydrate transport and metabolism’ category (Figure 3.10, bar G) had the most assignments for both the metagenome dataset (10.6%) and the metasecretome datasets (19.4%). Metasecretome phage display also enabled enrichment of proteins predicted to be involved in the ‘cell wall/membrane/envelope biogenesis’ processes (Figure 3.10, bar M) and peptides with ‘unknown function’ (Figure 3.10, bar S). In contrast, the functional categories of ‘replication, recombination and repair’ (Figure 3.10, bar L); ‘translation, ribosomal structure and biogenesis’ (Figure 3.10, bar J); ‘amino acid transport and metabolism’ (Figure 3.10, bar E) and ‘coenzyme transport and metabolism’ (Figure 3.10, bar H), comprised mainly of intracellular proteins, were relatively underrepresented in the metasecretome dataset.

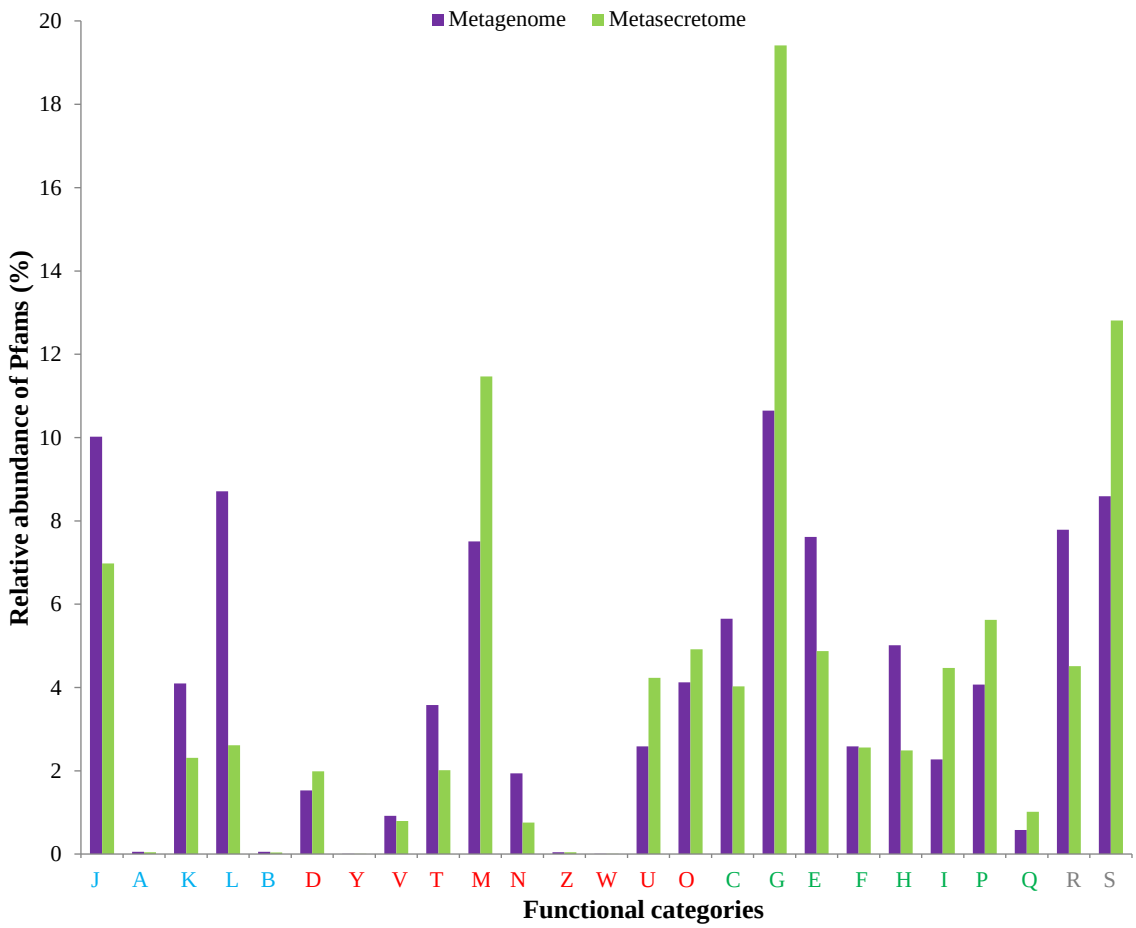

Figure 3.10 Relative abundances of Pfams within the metagenome and metasecretome-enriched sequence datasets.

Relative abundances of the IMG/M-annotated COG-based functional categories for protein family (Pfam) conserved domains assigned to the predicted protein coding genes within the metagenome (purple bars) and metasecretome (green bars) sequence datasets are shown. Abbreviations for the functional categories, grouped by general functional role: **Information storage and processing (blue font)**: J, Translation, ribosomal structure and biogenesis; A, RNA processing and modification; K, Transcription; L, Replication, recombination and repair; B, Chromatin structure and dynamics; **Cellular processes and signalling (red font)**: D, Cell cycle control, cell division, chromosome partitioning; Y, Nuclear structure; V, Defence mechanisms; T, Signal transduction mechanisms; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; Z, Cytoskeleton; W, Extracellular structures; U, Intracellular trafficking, secretion and vesicular transport; O, Posttranslational modification, protein turnover, chaperones; **Metabolism (green font)**: C, Energy production and conversion; G, Carbohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; **Poorly characterised (grey font)**: R, General function prediction only; S, Function unknown.

3.3.7 Diversity of CAZyme families captured by metasecretome selection

Protein-encoding genes from the metasecretome and metagenome datasets (222,960 and 671,876 ORFs, respectively), predicted *via* IMG/M pipeline [400], as well as from the published bovine rumen microbial switchgrass-adherent deep-sequenced metagenome (DMG) dataset (2,547,270 ORFs), obtained from microbiota adhered to switchgrass after 72 h of its incubation in the rumen of two fistulated cows [22], were subjected to automated annotation for CAZyme domains. Annotation was performed using the CAZy family-specific HMMs (333 HMMs, as of May 2013) [399], built from signature domain regions of CAZy families and three signature cellulosomal modules from the most complete set of metagenomic CAZyme genes deposited in the dbCAN database [399].

The analysis identified 12,565 putative CAZyme hits in the metasecretome dataset, with a significant match to at least one catalytic domain or associated module belonging to 196 different CAZy families, while the analysis of the metagenome dataset identified 21,823 hits belonging to 318 CAZy families. A complete list of hits in both datasets is presented in Appendix 2 (Table A2.1).

The diversity of families of catalytic, auxiliary and carbohydrate-binding CAZyme modules and cellulosome components captured by the metasecretome selection was compared with CAZyme families present in the metagenome dataset.In the metasecretome dataset, cellulosome components (dockerins and cohesins), as well as the GH and CE enzyme classes were predicted at higher frequencies compared to the metagenome dataset. In contrast, GTs and AAs occurred at lower frequencies in the metasecretome dataset, whereas overall frequencies of CBMs and PLs were comparable between two datasets (Figure 3.11).

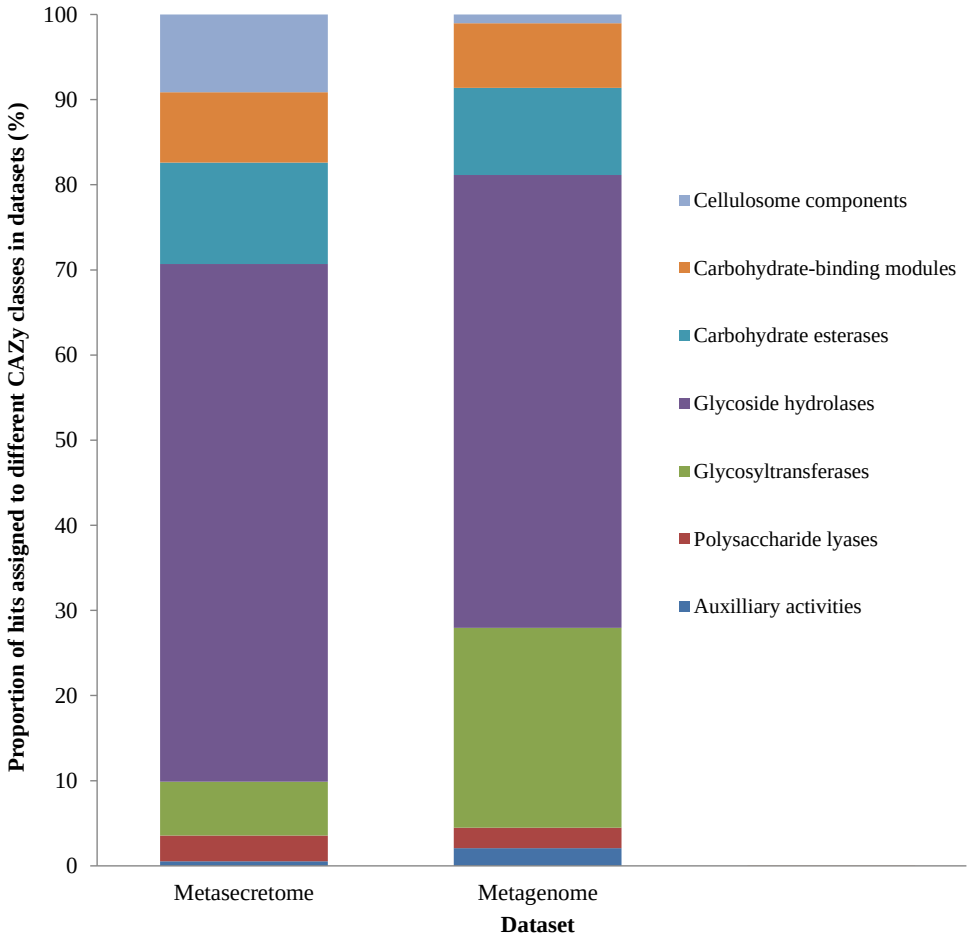

Figure 3.11 Comparison of dbCAN hits belonging to different CAZyme classes in the metasecretome and metagenome datasets.

Proportions of dbCAN hits belonging to different CAZy families of catalytic, auxiliary and carbohydrate-binding and cellulosome components were compared between metasecretome (N=12,565) and metagenome (N=21,823) datasets derived from the plant-adherent rumen microbial community fractions.

Both the metasecretome and metagenome datasets contained sequences encoding an assortment of putative oligosaccharide-degrading enzymes, hemicellulases and debranching enzymes, as well as a suite of CEs responsible for removal of esterified moieties bound to hemicelluloses. Comparison of the 20 most abundant families in each dataset is presented in Table 3.3. CAZyme families predicted at higher and lower frequency in the metasecretome compared to the metagenome dataset are presented in Appendix 2 (Table A2.2 and Table A2.3, respectively). An overview of the occurrence of hits to the major CAZy families involved in the degradation of plant cell wall polysaccharides in the metasecretome and metagenome datasets is also presented in Appendix 2 (Figure A2.1).

Table 3.3 Comparison of the 20 most abundant CAZyme families in the metasecretome and metagenome datasets.

Metasecretome			**Metagenome**		
CAZY family	Count	% of all dbCAN hits	CAZY family	Count	% of all dbCAN hits
dockerin	1,049	8.3	GT2	1,461	6.7
GH43	1,032	8.2	GH2	1,070	4.9
GH2	828	6.6	GH43	772	3.5
GH97	574	4.6	GH13	742	3.4
GH3	455	3.6	GT4	714	3.3
CE8	427	3.4	GH3	649	3.0
GH25	423	3.4	GT41	505	2.3
GH92	387	3.1	CE10	501	2.3
CE1	356	2.8	GH31	496	2.3
GH127	340	2.7	CE1	487	2.2
GH53	291	2.3	GT35	472	2.2
GH5	268	2.1	GH94	460	2.1
GT41	266	2.1	GH5	395	1.8
CE7	248	2.0	GH51	390	1.8
GH31	230	1.8	GH97	381	1.7
GH30	228	1.8	GH77	373	1.7
GH105	208	1.7	GH36	348	1.6
CE10	195	1.6	GH95	326	1.5
GH36	194	1.5	GH10	291	1.3
CBM67	183	1.5	GH115	278	1.3
Total	8,182	65.1	Total	11,111	50.9

Comparison of the metasecretome and metagenome GH profiles showed that ORFs encoding putative enzymes that act on oligosaccharides, belonging to the GH2, GH3 and GH43 families; cellulase GH5; pectin-degrading GH53 and debranching enzymes GH127, all in the top 20 CAZyme families, as well as endohemicellulase GH26 and xyloglucanases (GH16 and GH74) occurred at higher frequency in the metasecretome dataset. Also, putative GH97 (α-glucosidase), GH25 (lysozyme), GH92 (mannosidase), GH127 (β-L-arabinofuranosidase), GH53 (endo-β-1,4-galactanase), GH30 (oligosaccharide-degrading enzyme) and GH105 (unsaturated rhamnogalacturonyl hydrolase), were all found in the 20 most abundant putative CAZyme families encoded by the metasecretome dataset. Putative endoglucanases (GH74 and GH124) were also abundant as compared to the metagenome. In contrast, debranching enzymes (GH51, GH67, GH78 and GH 115) and endohemicellulase GH10, as well as putative enzymes

found in the 20 most abundant metagenomic CAZymes hits (GH13, GH77, GH94) occurred at a lower frequency in the metasecretome dataset.

Sequences encoding putative acetyl xylan esterases (families CE1, CE3 and CE7) and putative pectin methylesterase (family CE8) were more abundant in the metasecretome dataset as compared to the metagenome dataset.

CBM67 (binding to L-rhamnose), was found among the 20 most abundant families in the metasecretome dataset, and was present at higher frequency compared to the metagenome dataset. CBM67 domains are usually associated with GH78 catalytic modules; however, GH78 was not detected in large numbers in the metasecretome dataset and was less abundant in the metasecretome compared to the metagenome. The putative CBM40 (sialic acid binding), CBM16 (cellulose and glucomannan binding) and CBM61 (binding to β-1,4-galactan), as well as the starch-binding CBM20 and CBM26, and cellulose-binding CBM30 also occurred at a higher frequency in the metasecretome relative to the metagenome.

Hits to several families of GHs and CBMs that are typically found in large numbers in fungi (GH6, GH7, GH47, GH61, GH72 and CBM1) [19] were either absent, or only few hits could be detected in both datasets. This is consistent with low numbers of assignments to fungi in the taxonomic analysis of these datasets.

Glycosyltransferases, the enzymes which assemble glycans (glycoproteins, glycolipids, oligosaccharides), were present in much lower frequency in the metagenome compared to the metasecretome (6.3% and 23.5%, respectively). In contrast, a prominent enrichment of sequences encoding putative dockerin and cohesin modules was observed in the metasecretome dataset. These modules are typically components of the complex carbohydrate-degrading cell-surface bound multi-enzyme complexes, cellulosomes, that have been rarely detected in previously published metagenomic studies of the cow rumen [19, 22].

The GH profile of the metasecretome dataset was similar to other reported bovine metagenomes, except that GH53 (exclusive β-1,4-galactanase), responsible for degradation of galactans and arabinogalactans, and GH43 (various oligosaccharide degrading enzymes) seem to be present in abundance (Table 3.4). In contrast, putative cohesin and dockerin modules were present at higher frequency compared to reports from previous rumen microbiome studies.

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.

Predominant activity of GH family members		Bovine fibre-adherent [a] metasecretome (this study)	Bovine fibre-adherent [a,b] metagenome (this study)	Bovine fibre-adherent [a,c] [19]	Bovine switchgrass-adherent [d,e] [22]	Svalbard reindeer fibre-adherent [a,e] [23]	Yak whole rumen [d,e] [24]
Cellulases							
GH5	cellulases	268	395	20	1451	287	1302
GH6	endoglucanases	0	3	0	0	0	0
GH7	endoglucanases	0	0	0	1	0	0
GH9	endoglucanases	109	242	17	795	109	767
GH44	endoglucanases	12	9	0	0	5	0
GH45	endoglucanases	0	11	0	115	0	13
GH48	cellobiohydrolases	0	4	1	3	5	32
Total cellulases		389	664	38	2365	406	2114
Endohemicellulases							
GH8	endoxylanases	7	98	7	329	35	174
GH10	endo-1,4-β-xylanases	77	291	16	1025	190	2664
GH11	xylanases	5	20	1	165	8	244
GH12	xyloglucanases	0	1	0	0	0	0
GH26	β-mannanases and xylanases	160	121	16	369	153	537
GH28	galacturonanases	129	206	9	472	120	244
GH53	endo-1,4-β-galactanases	291	126	51	0	125	1066
Total endohemicellulases		669	863	100	2360	631	4929
Xyloglucanases							
GH16	xyloglucanases	83	84	1	483	116	563
GH74	endoglucanases and xyloglucanases	68	46	0	0	44	0
Total xyloglucanases		151	130	1	483	160	563
Debranching enzymes							
GH51	α-L-arabinofuranosidases	69	390	184	0	488	0
GH54	α-L-arabinofuranosidases	1	9	4	0	23	111
GH62	α-L-arabinofuranosidases	0	2	0	1	0	0
GH67	α-glucuronidases	18	134	0	120	74	1090
GH78	α-L-rhamnosidases	3	190	93	1260	313	426
Total debranching enzymes		91	725	281	1381	898	1627
Oligosaccharide-degrading enzymes							
GH1	β-glucosidases	2	111	31	253	122	331
GH2	β-galactosidases	828	1070	527	1436	716	942
GH3	β-glucosidases	455	649	497	2844	844	5448
GH29	α-L-fucosidases	26	147	79	939	268	899

Predominant activity of GH family members		Bovine fibre-adherent [a] metasecretome (this study)	Bovine fibre-adherent [a,b] metagenome (this study)	Bovine fibre-adherent [a,c] [19]	Bovine switchgrass-adherent [d,e] [22]	Svalbard reindeer fibre-adherent [a,e] [23]	Yak whole rumen [d,e] [24]
GH35	β-galactosidases	2	86	27	158	39	468
GH38	α-mananosidases	2	22	46	272	116	90
GH39	β-xylosidases	2	74	7	315	76	159
GH42	β-galactosidases	2	47	35	374	95	207
GH43	arabino/xylosidases	1032	772	176	0	787	2313
GH52	β-xylosidases	1	4	0	0	2	0
Total oligosaccharide -degrading enzymes		2352	2982	1425	6591	3065	10857
Total number of GHs detected in the study		7639	11606	2720	27755	5160	37563
Cellulosome domains							
Cohesins		52	27	0	80	52	51
Dockerins		1049	121	1	188	92	516
Raw sequence information (Gb)		0.38	180	0.08	268	0.5	0.09

Data are presented in the format used in [224], with GH families targeting plant structural polysaccharides grouped according to their major functional role in the plant fibre degradation.
[a] Unassembled metagenome; [b] Combined metagenomes of two individual animals; [c] Combined metagenomes of three individual animals; [d] Assembled metagenome; [e] Metagenome obtained from pooled rumen samples of two animals.

In a subset of 33 individual metasecretome ORFs with multiple, non-overlapping hits to CAZyme modules (with an E-value <1e-05 for an alignment length >80 amino acid residues and an E-value <1e-03 for shorter alignment lengths), 8 ORFs that had query coverage >30% for each putative CAZyme module were further inspected as described in section 2.2.6.2.3. Six candidate putative proteins with predicted multi-modular CAZyme organisation were identified using this approach (Figure 3.12 and Table A2.4 of the Appendix 2). An example alignment of putative multi-modular CAZyme and corresponding HMMs is provided in Appendix 2 (Figure A2.2).

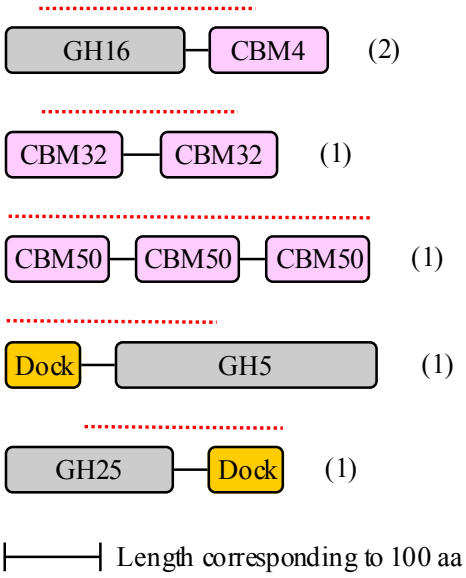

Figure 3.12 Architecture of putative proteins with predicted multi-modular CAZyme organisation in the metasecretome dataset.

Putative proteins predicted in the metasecretome dataset *via* the IMG/M pipeline were subjected to automated CAZyme annotation, followed by the manual inspection of putative proteins with non-overlapping hits to multiple CAZyme HMMs (with an E-value <1e-05 for an alignment length >80 amino acid residues and an E-value <1e-03 for shorter alignment lengths) that had query coverage >30% for each putative domain. Predicted co-located catalytic and associated modules are shown as coloured boxes and labelled by their family designations. Black line connecting modules corresponds to the fraction of the putative protein in the metasecretome dataset that did not have a significant hit to the represented CAZyme modules (not to scale). A dotted horizontal red line above each combination of modules (HMMs) indicates the HMM fraction covered in an alignment with the putative metasecretome protein. The number of predicted individual multi-modular proteins with particular CAZyme module combination is represented in brackets. Grey boxes, modules belonging to the GH (glycoside hydrolase) families; pink boxes, CBMs (carbohydrate-binding modules); yellow boxes, Dock (single dockerin repeats). The lengths shown for the CAZyme modules are based on the lengths of their HMMs in the dbCAN database. The scale bar represents 100 amino acid residues.

3.3.8 Abundance and phylogenetic diversity of cellulosome components predicted in the metasecretome dataset

Hits to the ‘signature’ cellulosome-associated modules: cohesins, dockerins and SLH (S-layer homology) domains in the metasecretome dataset were further analysed after clustering at 100% sequence identity using the CD-HIT algorithm [402] to remove duplicity. Resulting distinct putative cellulosome modules (44 cohesins, 499 dockerins and 34 SLH) were further analysed. Around 6.3% of the total clustered CAZyme hits in the metasecretome dataset were to dockerin modules (complete or partial). Dockerins are usually around 70 amino acid residues in

length, and contain two repeats, each with a non-EF hand calcium binding motif of about 22 amino acid residues in length, separated by a linker sequence [182, 183]. Around 4.5% of the dockerin hits in the metasecretome dataset were partial (to a single dockerin repeat), 1.7% were complete (to both dockerin repeats) and 0.1% were to a single dockerin repeat in combination with another CAZyme module. Around 0.6% and 0.4% of the clustered CAZyme hits were to two other putative cellulosome-associated modules, cohesins and SLH domains, respectively.

The frequencies of the three 'signature' cellulosome modules predicted in the metasecretome and metagenome datasets were compared with those in the DMG, the most extensive published bovine rumen metagenome dataset currently available [22], also using predictions *via* dbCAN HMMs. Due to the extent of sequencing undertaken (268 Gb raw sequence), at least five times more predicted CAZyme-encoding genes (27,755) were originally reported for this dataset, as compared to all previously published lower-depth rumen microbial metagenomic studies combined and over one half of CAZyme-encoding genes were complete [22].

Consequently, the total number of distinct putative CAZyme modules identified using an HMM-based approach, obtained after clustering of all hits at 100% sequence identity using CD-HIT [402], was much smaller in the metasecretome and metagenome datasets (7,978 and 21,607, respectively) than in the DMG dataset (123,223). However, comparison of the frequencies of cellulosome module encoding sequences in the three datasets showed a prominent enrichment for sequences encoding putative cohesin and dockerin modules in the metasecretome dataset compared to both the metagenome and DMG datasets (Figure 3.13).

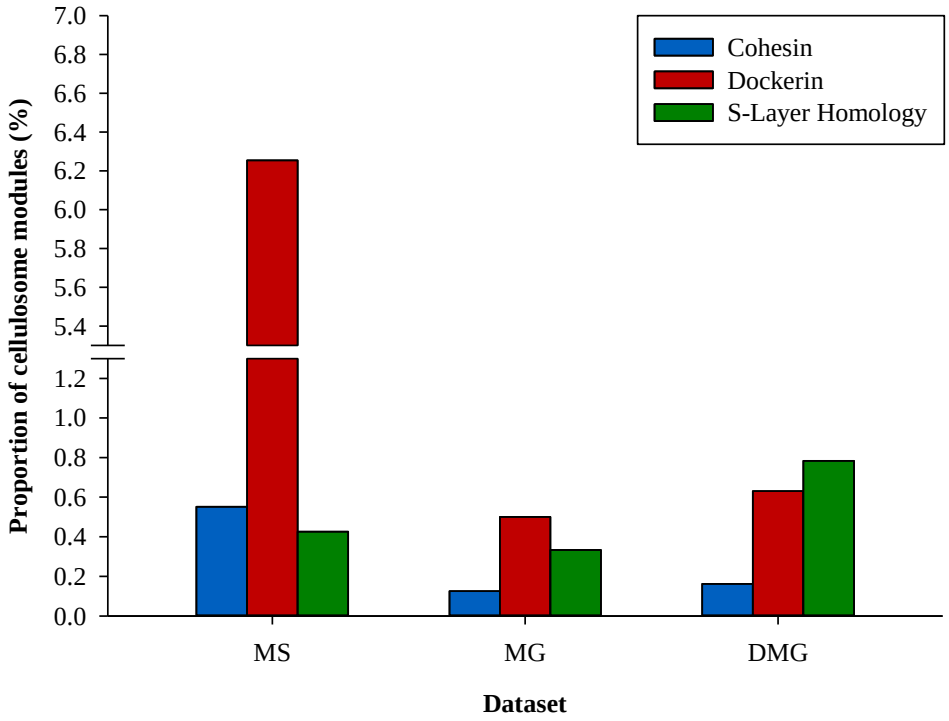

Figure 3.13 Frequency of cellulosome modules in three bovine rumen plant-adherent microbial datasets.

Frequency of three putative distinct 'signature' cellulosome modules: cohesins (blue); dockerins (red) and surface S-layer homology (SLH) domains (green) in three datasets: MS, metasecretome dataset (N=7,978); MG, metagenome dataset (N=21,607) and published deep-sequenced metagenome (DMG) dataset (N=123,223). Both the MS and MG dataset were derived from the plant-adherent rumen microbial community fraction isolated from fistulated pasture-grazing dairy cows, while the DMG dataset was derived from the bovine switchgrass-adherent microbial community [22].

The family-level taxonomic assignment (see section 2.2.6.2.3) of putative proteins in the metasecretome dataset with predicted cohesin, SLH and the complete dockerin domains (containing two putative dockerin repeats) allowed the estimation of the phylogenetic diversity of these modules (Figure. 3.14).

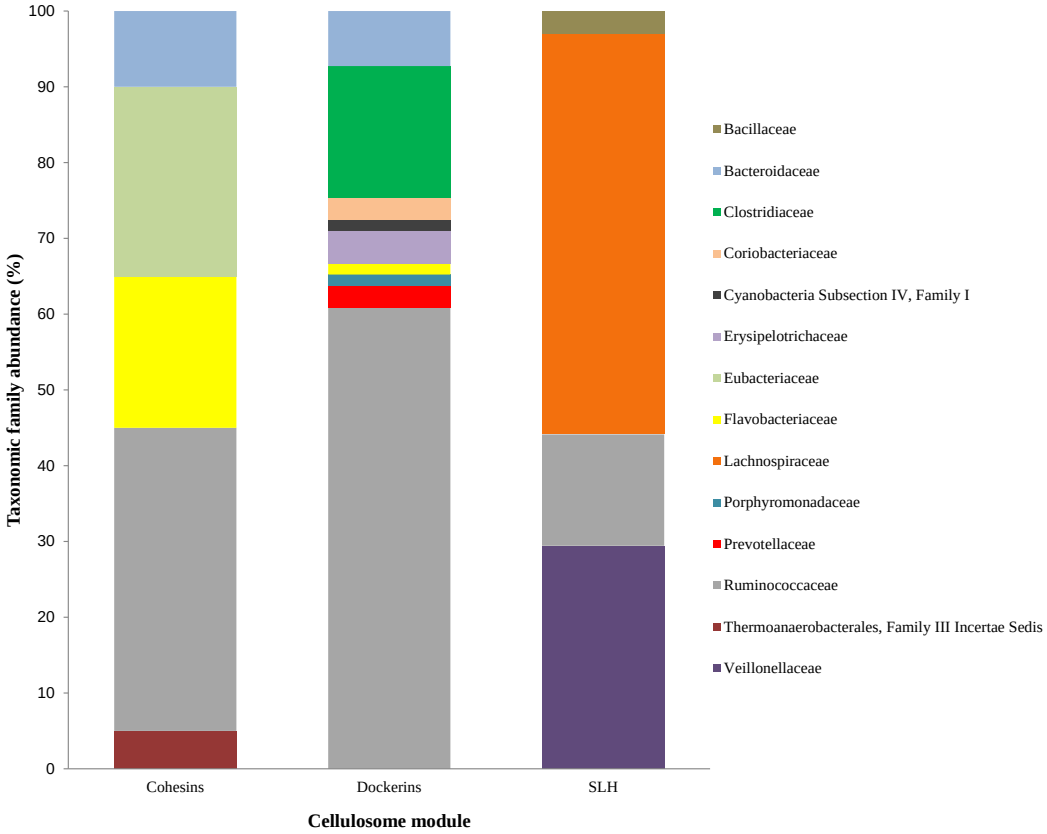

Figure 3.14 Phylogenetic diversity of the cellulosome modules predicted in the rumen metasecretome dataset.

Putative proteins from the metasecretome dataset that were predicted to contain 'signature' cellulosome modules: cohesins, dockerins and SLH domains were compared to the NCBI nr protein database using BLASTP. Family-level taxonomic assignments were made for the host organism of the best BLASTP hit with bit-score >40 for cohesins (N=20) and SLH (N=34) and >35 for dockerins (N=403), taking into account the recent taxonomic classification proposals for these organisms. The chart shows the abundance of each family for each cellulosome module. For the dockerin data, only sequences that contained two putative dockerin repeats (N=69) are shown.

Around two thirds of the cohesin module-containing sequences were assigned to the Firmicutes [including Ruminococcaceae (40%) and Eubacteriaceae (25%)], with the remaining modules assigned to Bacteroidetes [Flavobacteriaceae (20%) and Bacteroidaceae (10%)]. The vast majority of dockerin-containing sequences were assigned to the Firmicutes [including Ruminococcaceae (61%) and Clostridiaceae (17%)], while Bacteroidetes representation was mainly within the Bacteroidaceae (7.3%), and Prevotellaceae (2.9%). Among the best BLASTP hits, many were to species that have been previously reported as cellulosome-producers, such as *Acetivibrio cellulolyticus*, *Clostridium acetobutylicum*, *C. cellulolyticum* (recently, a re-classification as *Ruminiclostridium cellulolyticum* has been proposed [176]), *C. josui* (a re-classification as *Ruminiclostridium josui* has been proposed [176]), *C. thermocellum*

(a re-classification as *Ruminiclostridium thermocellum* has been proposed [176]), *Ruminococcus albus* and *R. flavefaciens* [64]. In contrast, 91% of putative SLH domains were assigned to Firmicutes (including 53% to Lachnospiraceae, 29% to Veillonellaceae and 15% to Ruminococcaceae).

3.3.9 Assessment of the novelty of CAZymes detected in the metasecretome dataset

Automated annotation of CAZymes *via* the dbCAN database was used not only to identify CAZyme-encoding sequences enriched in the metasecretome dataset after selection, but also to identify putative novel CAZymes that have a limited overall sequence identity to known proteins, and therefore may potentially possess divergent functional properties.

Briefly, the sequences of 7,910 distinct putative proteins longer than 30 amino acid residues, containing putative CAZyme modules from the metasecretome dataset were compared with proteins deposited in the NCBI nr database, and with relevant full-length putative CAZyme proteins deposited in the dbCAN database. The collection of putative CAZymes available through the dbCAN database (4,073,867) was estimated (in 2012) to have three times as many CAZyme homologues compared to the NCBI nr database [399]. This specialised dbCAN-deposited collection contains putative CAZymes from CAZy [138], NCBI (nr and env nr) [423] and UniProt [424] databases; multiple metagenome datasets (JGI [425], CAMERA [426], BGI-gut [427] and the deep-sequenced metagenome of cow rumen switchgrass-adherent microbiota [22]); as well as a small number of sequences from the human gut microbiome [428] and plant genomes [429] that are not available in GenBank.

Over one quarter (26%) of putative proteins with predicted CAZyme modules in the metasecretome dataset showed less than 50% of sequence identity, while around 65% shared 50 - 90% sequence similarity with proteins deposited in the NCBI nr protein database (Figure 3.15, red line).

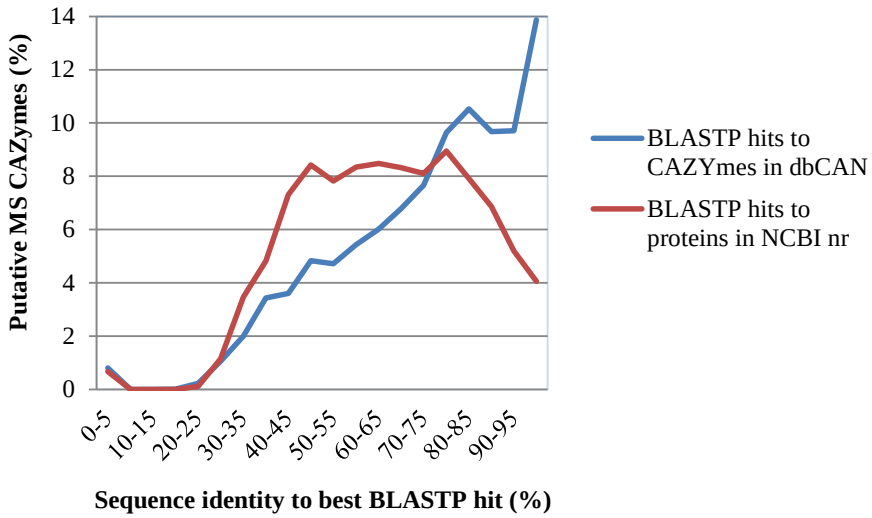

Figure 3.15 Distribution of sequence identity of best BLASTP hits for CAZymes detected within the metasecretome dataset.

Distribution of sequence identity is represented for best BLASTP hits of distinct putative CAZymes, detected within the metasecretome dataset (N=7,910), against a collection of full-length putative CAZymes deposited in the dbCAN database (blue) and NCBI nr protein database (red).

When compared to the dbCAN-deposited collection of putative CAZymes, around 16% of putative proteins with predicted CAZyme modules within the metasecretome dataset showed <50% identity (Figure 3.15, blue line). Around 60% of putative CAZymes within the metasecretome dataset shared 50 - 90% sequence similarity with dbCAN-deposited putative CAZymes. Putative hits to cohesins and complete dockerins modules within the metasecretome dataset shared limited sequence similarity (41 - 70% for cohesins, and 27 - 59% for dockerins) with putative cohesins and dockerins deposited in the dbCAN database.

Around 89% of the putative CAZymes predicted in this study had a significant match (E-value <1e-05) to the putative CAZymes deposited in the dbCAN database. Of these, the majority were homologous to putative CAZymes from the published cow rumen DMG dataset (66.7%) [22] and human gut microbial metagenome sequences (13.7%) [427]. Compared to the putative CAZymes from the DMG dataset, 19.4% of all CAZymes detected in the metasecretome dataset were highly similar (>90% amino acid sequence identity) and 3.6% of these were identical to CAZymes from DMG.

Results indicate that HMM-based CAZyme annotation of sequences in the metasecretome dataset allowed detection of both putative CAZymes sharing limited sequence similarity and those highly homologous to CAZymes observed in previous sequence and function based (meta)genomic studies. However, the effect of systematic errors (indels) in homopolymeric tracts causing frame-shifts in coding regions, produced by the 454 sequencing

platform with a frequency of around 1%, on the accuracy of functional annotations in the metasecretome dataset, cannot be excluded.

3.3.10 Overview of section 3.3

Metasecretome selection, applied in combination with NGS to enrich and explore the metasecretome of bovine rumen plant-adherent microbiota, increased the frequency of detection of the putative proteins involved in the carbohydrate transport and metabolism relative to that found in the plant-adherent metagenome. The strategy of using a combination of the metasecretome-selective phage display, that enriched for surface and membrane proteins and an HMM-based CAZyme annotation allowed detection of ORFs encoding diverse putative catalytic and binding modules of fibrolityc enzymes, including GH catalytic modules (assigned to 85 GH families) accompanied by a variety of CBMs (belonging to 38 CBM families), CEs (13 families) and PLs (10 families).

Of particular interest is identification of a large number of putative cellulosome components (dockerins and cohesins) in the metasecretome dataset, a large proportion of which shared homology with the cellulosome components found in the Ruminococcaceae family. Compared to other published bovine rumen metagenomic datasets, hits to putative cellulosome modules were detected in the metasecretome dataset in a higher proportion and were obtained from very limited amount of sequence information (55Mb corresponding to 379 Mb of raw sequence). This demonstrates the power of the metasecretome selection approach for enrichment and mining of ‘rare’ secretome proteins, scarcely represented in the shotgun metagenomic datasets.

3.4 Summary

The study described here was the first application of the secretome-selective phage display on a metagenomics scale. Using first a pilot library in combination with Sanger sequencing, then a large-scale library in combination with the next-generation sequencing, this method was shown to be successful in cloning, selection and analysis of the secretome portion of a complex metagenome.

To determine the putative functions enriched in the metasecretome library, metasecretome sequence data was compared to a metagenome dataset derived from the plant-adherent rumen microbial fraction of two New Zealand cows on a pasture-based diet.

Comparison of taxonomic assignments of putative protein-coding sequences between the metasecretome and metagenome datasets showed a notably higher representation of metasecretome inserts from Gram-negative bacteria in the metasecretome dataset, which is expected given that the *E. coli* membrane targeting machinery is used for the secretome selection step.

Most of the functional assignments in both the metasecretome and metagenome datasets were to putative proteins involved in carbohydrate transport and metabolism (19.4% and 10.6%, respectively) and this was also the most prominently enriched COG functional category after selection. The metasecretome selective approach also enabled enrichment of genes encoding putative proteins involved in the 'cell wall/membrane/envelope biogenesis' processes and peptides with 'unknown function', while several predicted functional categories, comprised mainly of intracellular proteins, were underrepresented in the metasecretome dataset.

Putative cellulosome components (dockerins and cohesins), GH, CE and PL enzyme classes occurred at higher frequency, while GTs and AAs occurred at lower frequency in the metasecretome compared to the metagenome dataset. Metasecretome selection allowed the detection of an abundance of putative cohesin and dockerin modules (0.6% and 6.3% of all CAZyme hits clustered at 100% sequence identity). These modules were not detected in such high proportions in the metagenome and published deep-sequenced metagenome dataset [22] using the same database and search parameters. The majority of the metasecretome inserts predicted to encode dockerin and cohesin modules showed strong homology to sequences from members of the Ruminococcaceae.

Around 16% of putative CAZyme-encoding metasecretome ORFs were very divergent from the relevant full-length putative CAZyme proteins deposited in the dbCAN database. This indicates that metasecretome selection in combination with HMM-based CAZyme annotation allowed detection of some CAZyme-encoding genes that have not been observed through previous (meta)genomics efforts.

Chapter 4. Affinity screening of the metagenomic shotgun phage display library from the rumen plant-adherent microbiome for proteins mediating interactions with complex carbohydrates

For plant cell wall degradation, among the most important carbohydrate-active enzymes are GHs and CEs, both of which are typically modular in nature. In addition to their catalytic domains, they can contain other functional modules, such as CBMs, which facilitate the targeting of the enzymes to their carbohydrate substrates. CBMs affect the degradative capacity of their cognate catalytic modules through proximity effects and determine substrate affinity and selectivity, *via* targeting functions and through surface/interfacial modifications of the substrate [153].

The aim of the second part of this project was to explore the use of the phage display approach to identify surface and secreted proteins of the plant-adherent rumen microbiota that are involved in binding to complex carbohydrates. To achieve this, affinity screening of the recombinant PPs from the larger metagenomic shotgun phage display library (described in Chapter 3) was performed, using the complex carbohydrates, cellulose and hemicellulose, as baits.

4.1 Optimisation of complex carbohydrate affinity screening assays

To establish the conditions for affinity screening of the metagenomic phage display library for complex carbohydrate-binding PPs, several complex carbohydrate substrates were initially tested as baits by affinity binding assays, to identify those that result in low non-specific ('background') binding of PPs. These substrates were: cellulose (Whatman paper disc, microcrystalline cellulose and regenerated amorphous cellulose), oat spelt xylan, insoluble wheat arabinoxylan, and the neutral detergent fraction of ryegrass (corresponding to the cellulosic and hemicellulosic components of the plant cell wall). Substrates were immobilised onto empty disposable polypropylene columns and a binding assay was carried out with the pDJ01 vector-derived PPs as described in section 2.2.7.1. The amount of input PPs (mixed with the substrates at the start of the assay), and the amount of output PPs (recovered from the substrates at the end of the assay) was determined by titration, as described in section 2.2.3.6.1. The PPs were recovered from the substrates under three different elution conditions: acidic

(pH 2.2), basic (pH 9.2) and by direct on-substrate infection of *E. coli* TG1 host cells (see section 2.2.7.1 and 2.2.7.2). The average output/input ratios of PPs from all three elution conditions were calculated (Table 4.1).

Table 4.1 Binding of pDJ01 vector-derived PPs to complex carbohydrate substrates.

Substrate [a]	**What**	**MCC**	**RAC**	**XYL**	**AXYL**	**NDF**	**Plastic**
Average output/input ratio [b]	5×10^{-4}	2×10^{-4}	4×10^{-5}	2×10^{-6}	2×10^{-7}	5×10^{-4}	4×10^{-6}

[a] Substrate: What, Whatman paper No. 1; MCC, microcrystalline cellulose; RAC, regenerated amorphous cellulose; XYL, oat spelt xylan; AXYL, insoluble wheat arabinoxylan; NDF, neutral detergent fraction of ryegrass; Plastic, polypropylene walls of empty columns used for substrate immobilisation. [b] Average output/input ratio from three different elution conditions: acidic (low pH), basic (high pH) and by on-substrate infection of *E. coli*. Output/input ratio was calculated based on the total number of vector-only PPs (pDJ01 PPs) in the eluate (output) and total number of PPs used as input.

Substrates with the lowest output/input ratio among the tested cellulosic and hemicellulosic substrates, indicating the lowest levels of non-specific 'background' binding of virions, were regenerated amorphous cellulose (RAC), xylan and arabinoxylan (AXYL). Thus, to represent cellulose, RAC was chosen to be used as a substrate for affinity screening of the metagenomic phage display library, and affinity binding assays. To represent hemicellulose, AXYL was chosen over xylan as it not only had a lower background binding, but it is more representative of the xylan structure found in ryegrass, which was the predominant component of the diet of the animal from which the phage display libraries were derived.

4.2 Affinity screening of the metagenomic phage display library for carbohydrate-binding proteins on RAC and AXYL

Proteins from the cow plant-adherent rumen microbiome, displayed on the surface of metagenomic phage display library virions, were affinity screened on AXYL and RAC to find those that are potentially involved in binding to complex carbohydrates.

The complete shotgun metagenomic library (primary size ~5×10^6 clones), constructed from the metagenomic DNA from rumen plant-adherent microbiota (described in Chapter 3), was used to produce PPs for affinity screening. This strategy was chosen instead of affinity screening of the metasecretome phage display library PPs, as transformation of *E. coli* with the

metasecretome-enriched ssDNA (post secretome selection) is of extremely low efficiency, and would likely result in biases in the metasecretome library representation.

To produce infectious PPs required for multiple rounds of the panning protocol, the master shotgun metagenomic library was used as the starting material. This library was produced by transforming the library ligation products into *E. coli* in 27 separate transformation reactions, with recombinant phagemids amplified from the plasmid origin of replication in *E. coli*, and each amplified transformation aliquot was stored separately at –80°C in 7% DMSO, as described in section 2.2.5.1. Each aliquot was used to seed a separate overnight culture, followed by mass-infection with the wt ($gIII^+$) helper phage VCSM13 to initiate replication from the phagemid f1 origin, and enable packaging into PPs. The PPs were expected to have a monovalent display of the fusion proteins, dependent on the presence of the endogenous signal sequences encoded by the recombinant phagemid inserts. The PPs obtained after the VCSM13 infection of 27 metagenomic library cultures were pooled (MG1-27 PPs), purified by PEG/NaCl precipitation, and titrated as described in sections 2.2.3.2 and 2.2.3.3. The resulting master rumen plant-adherent metagenomic phage display library (5×10^{12} PPs/mL) was stored in aliquots in 7% DMSO at –80˚C for later use in affinity screening procedures. The PPs derived from the 'empty' vector pDJ01, produced and purified in parallel with the library PPs, were used as a negative control.

Approximately 1×10^{12} PPs from the rumen metagenomic phage display library, and pDJ01 PPs were used as the inputs in each round of affinity enrichment on AXYL and RAC (as described in section 2.2.7.2). Putative binders to carbohydrate substrates were identified through four rounds of panning. In each library-panning round, the unbound PPs were removed by extensive washing, followed by elution of substrate-bound PPs using the three different conditions: acidic, basic and on-substrate infection of the host cells. The PPs eluted by all three conditions were amplified in the *E. coli* TG1 host strain. The amplified PPs were concentrated and purified and used in the subsequent round of panning. The output/input ratios for metagenomic PPs and pDJ01 PPs (control) were determined for each round of panning (Table 4.2).

Table 4.2 Enrichment of metagenomic phage display library PPs through four rounds of affinity panning on complex carbohydrates.

Substrate [a]	Elution method	Output/input ratio [b]				Enrichment [c]
		1st round	2nd round	3rd round	4th round	4th/1st round
RAC	Acidic	7.5×10^{-4}	5.7×10^{-5}	8×10^{-6}	8.7×10^{-6}	0.01×
RAC	Basic	6×10^{-4}	3.3×10^{-5}	4.6×10^{-6}	1.2×10^{-4}	0.2×
RAC	TG1	2.2×10^{-3}	2.9×10^{-4}	2.7×10^{-4}	6×10^{-4}	0.3×
AXYL	Acidic [d]	1.1×10^{-10}	1.5×10^{-10}	1.3×10^{-9}	ND	ND
AXYL	Basic [d]	5×10^{-9}	2.5×10^{-8}	1.7×10^{-9}	1.9×10^{-9}	ND
AXYL	TG1	6×10^{-7}	6.9×10^{-6}	2.6×10^{-5}	1.3×10^{-5}	22×

[a] Substrate: RAC, regenerated amorphous cellulose; AXYL, insoluble wheat arabinoxylan. [b] Output/input ratio was calculated for each round of library panning on each substrate by dividing the total number PPs in the eluate (output) with the total number of PPs used as input. Output and input PP numbers were determined by titration. [c] Enrichment was calculated by dividing the output/input ratios calculated for the 4th round of panning with the output/input ratio calculated for the 1st round of panning. [d] Enrichment was not considered for PPs eluted from AXYL under acidic conditions. ND, not determined.

An increase in the output/input ratio is an indication of successful selection due to amplification of phage library members with adequate complex carbohydrate binding properties within the phage population. The observed lack of increase in the number of PPs eluted from the RAC under all three elution conditions after the fourth round of panning compared to the first round of panning suggests that no specific enrichment has taken place, or that the displayed proteins in selected recombinant PPs have a relatively low affinity for substrate, hence their binding cannot be detected through an increase in the output titre.

A 22-fold increase in the number of PPs recovered from AXYL by elution with host cells after the fourth round relative to the first round of panning was observed, indicating that an interaction may have occurred between PPs and the substrate. Enrichment of PPs eluted from AXYL under acidic and basic condition was not considered due to extremely low output titres and observed toxicity to the *E. coli* host. The toxicity could have been caused by cumulative effect of compounds released into the eluate during exposure of the hemicellulosic substrate to the acidic or basic elution buffers. Products of hemicellulose hydrolysis (organic acids and aldehydes), deacetylation and di-ferulate bond cleavage are documented in the literature to inhibit the growth of *E. coli* to different extents [434-436].

The PPs eluted from AXYL with TG1 host cells showed considerable, and relatively consistent, increases in binding over the background control (pDJ01 PPs) through four rounds of panning (Table 4.3). In contrast, non-specific adsorption of library PPs to the cellulosic

substrate might have contributed to the observed level of their binding to RAC, which is similar to the level observed for empty vector-derived PPs that are not displaying any peptides (pDJ01 PPs).

Table 4.3 Binding of metagenomic phage display library PPs over background through four rounds of affinity panning on complex carbohydrates.

Substrate [a]	Elution method	pDJ01 PPs output/ input ratio [b]	Fold binding of metagenomic phage display library PPs above background (vector-only pDJ01 PPs) [c]			
			1st round	2nd round	3rd round	4th round
RAC	Acidic	2.8×10^{-6}	267	20	3	0.3
RAC	Basic	2×10^{-6}	300	17	2	59.7
RAC	TG1	2×10^{-4}	11	2	1	3
AXYL	Acidic [d]	1×10^{-10}	1.1	2	13	ND
AXYL	Basic [d]	4×10^{-9}	1.2	6	0.4	ND
AXYL	TG1	1×10^{-8}	60	690	2600	1300

[a] Substrates: RAC, regenerated amorphous cellulose; AXYL, insoluble wheat arabinoxylan; [b] Output/input ratio was calculated based on the total number of PPs in the eluate (output) and total number of PPs used as input for each round of panning for metagenomic phage display library PPs, or for a single round of panning for the pDJ01-derived PPs. [c] Fold binding over background was calculated as the ratio between output/input for recombinant metagenomic phage display library PPs relative to the output/input ratio for the pDJ01-derived PPs for each round of panning. [d] Binding over background was not considered for PPs eluted from AXYL under acidic and basic conditions. ND, not determined.

High background binding of filamentous phage virion components, other than the displayed peptide, has been reported previously to contribute to non-specific binding of PPs to the cellulosic substrate [379].

As another indicator of individual recombinant clone enrichment within the library, phagemid profiles of the amplified metagenomic library pools from each round of panning were monitored by agarose gel electrophoresis (Figure 4.1). Disappearance of the smear of recombinant phagemids due to the random insert size distribution, and appearance of discrete bands, is an indication of enrichment of particular recombinant phagemids (candidate carbohydrate-binders) relative to the rest of the library.

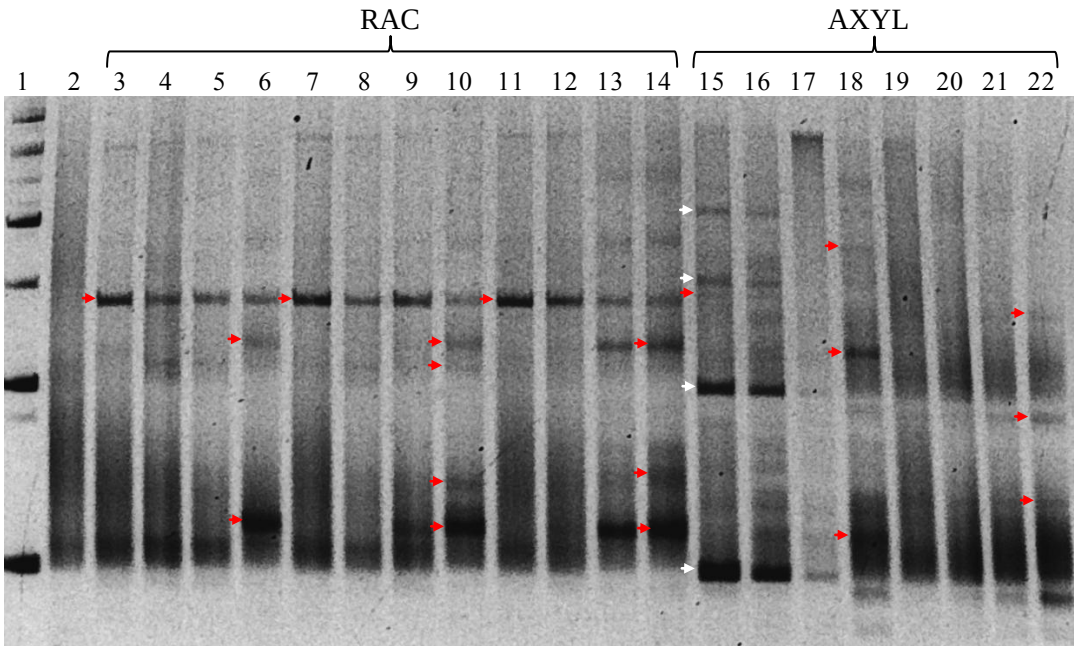

Figure 4.1 Recombinant phagemid profiles of the metagenomic library over four rounds of affinity panning on carbohydrate substrates.

The recombinant phagemid library pool (in the form of closed circular dsDNA) was purified from aliquots of the metagenomic library amplified overnight after each round of panning. Phagemid DNA bands were visualised by staining with EtBr after agarose gel electrophoresis. Abbreviations: RAC, regenerated amorphous cellulose; AXYL, insoluble wheat arabinoxylan; MG1 – 27, master rumen plant-adherent metagenomic phage display library; A, acidic elution; B, basic elution; T, elution with host *E. coli* TG1; 1 – 4, first to fourth round of panning. Lanes: 1, vector pDJ01; 2, MG1 – 27 before panning; 3, RAC A1; 4, RAC A2; 5, RAC A3; 6, RAC A4; 7, RAC B1; 8, RAC B2; 9, RAC B3; 10, RAC B4; 11, RAC T1; 12, RAC T2; 13, RAC T3; 14, RAC T4; 15, AXYL B1; 16, AXYL B2; 17, AXYL B3; 18, AXYL B4; 19, AXYL T1; 20, AXYL T2; 21, AXYL T3; 22, AXYL T4. Red arrows indicate DNA bands purified from preparative gels; white arrows correspond to pDJ01 phagemid vector bands.

The appearance of different sets of discrete recombinant phagemid bands during the library panning rounds on RAC compared to AXYL substrate was indicative of selection of different phagemid populations in the two affinity enrichment experiments. Recombinant phagemid DNA corresponding to three bands from the first round of panning and nine bands from the fourth round of panning on RAC, eluted under all three elution conditions, were extracted from slices of a preparative agarose gel. Additionally, DNA corresponding to a single band from the first round of panning and six bands from the last round of panning on AXYL, eluted with basic elution buffer and host TG1 cells, respectively, were extracted from agarose gel slices (Figure 4.1, red arrows). The DNA purified from each agarose gel slice was used to separately transform electrocompetent TG1 cells and the resulting transformants (20 from each of the 19 transformations) were analysed.

4.3 Characterisation of affinity-selected clones

To examine the insert size distribution of the affinity-selected phagemid clones, bacterial colony PCR was performed (as described in section 2.2.2.5.1) on 20 randomly picked colonies from each of the 19 transformations. In total, 380 independent colonies of transformants were analysed (Figure 4.2). The insert sizes ranged from 0.3 – 2.5 Kb. Forty randomly selected recombinant phagemids (Figure 4.2, red arrows: 5 from the first round of panning and 19 from the fourth round of panning on AXYL, and 7 from first round of panning and 9 from fourth round of panning on RAC) contained inserts of varying sizes, and were further analysed by DNA sequencing.

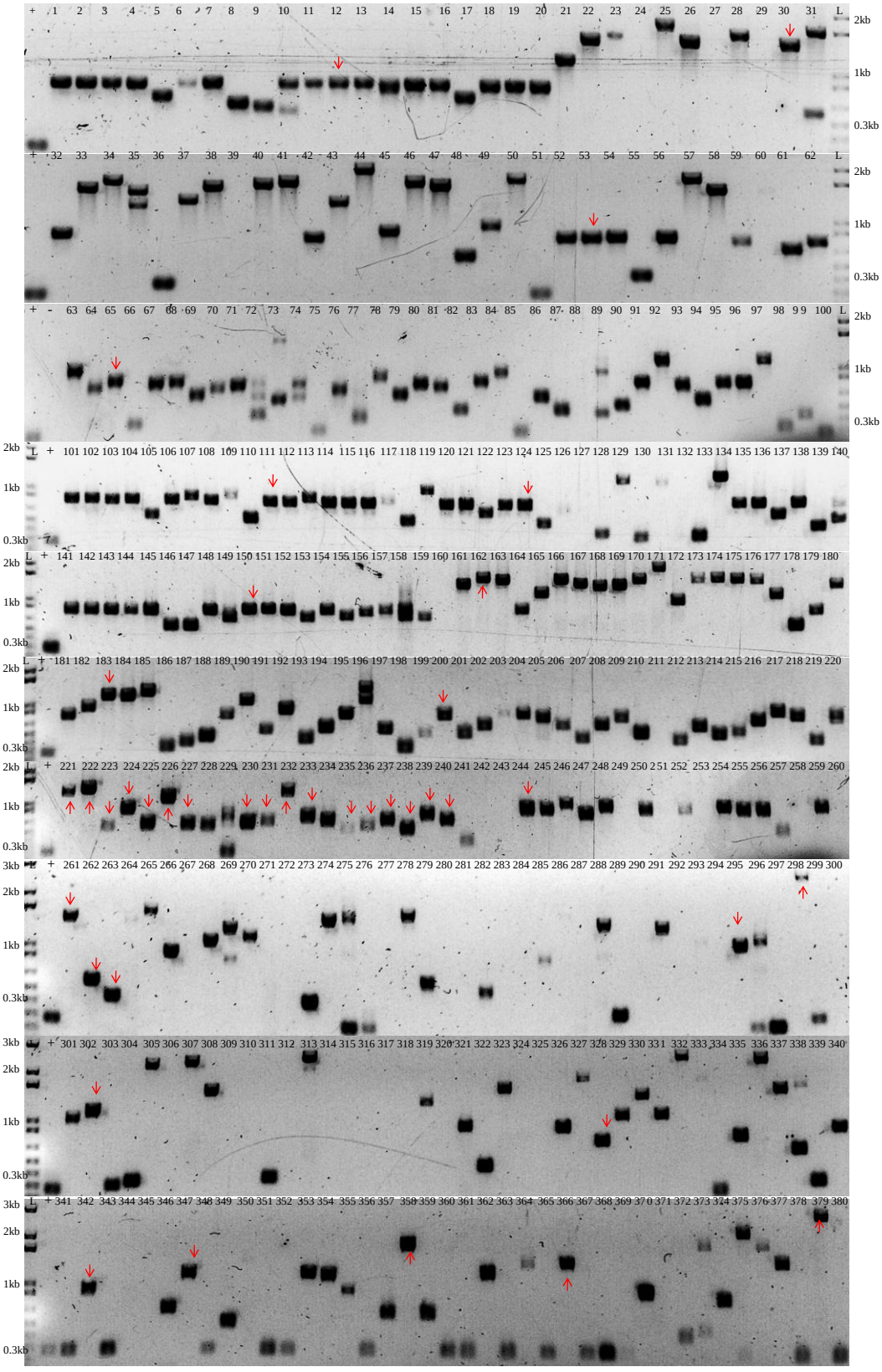

Figure 4.2 Bacterial colony PCR of 380 clones selected from the metagenomic phage display library by affinity screening against complex carbohydrate substrates.

Lanes 1 – 380 correspond to PCR amplicons from randomly picked colonies obtained through transformations with 19 extracted DNA bands indicated in Figure 4.1. Abbreviations: RAC,

regenerated amorphous cellulose; AXYL, insoluble wheat arabinoxylan; A, acidic elution; B, basic elution; T, elution with host *E. coli* TG1; 1 – 4, first to fourth round of panning. Lanes: 1 – 60 RAC T4; 61 – 100 and 361 – 380 AXYL B4; 101 – 140 RAC A4; 141 – 200 and 301 – 320 RAC B4; 201 – 260 AXYL T4; 261 – 280 RAC A1; 281 – 300 RAC B1; 321 – 340 RAC T1; 341 – 360 AXYL B1; L, 1 Kb Plus DNA ladder (Life Technologies); +, PCR positive control (PCR amplicon from TG1 colony transformed with vector pDJ01); –, PCR negative control. Red arrows indicate sequenced inserts.

Sequencing was performed with an antisense primer (pspR03; see section 2.1.4) complementary to the pDJ01 vector sequence downstream of the insert. This primer anneals within the vector *gIII* gene, encoding virion protein pIII that serves as a platform for display of peptides in fusion with pIII. Thus, sequencing with pspR03 allows the joint between the insert and the vector in the insert-*gIII* fusion to be determined. If the 5' end of the fusion ORF was not reached, sequencing with a sense primer (pspF03; see section 2.1.4) was performed. The 40 sequenced recombinant phagemid inserts were analysed as described in section 2.2.7.3, and a detailed overview is represented in Appendix 3.

Of the 40 sequenced inserts, 24 contained ORFs in frame with *gIII*. Of these, 13 inserts contained distinct ORF encoding putative proteins longer than 24 amino acid residues. Eleven inserts contained 10 short distinct ORFs encoding putative peptides and proteins (≤24 amino acid residues long) that were considered 'background' (Table 4.4). A large proportion of inserts (40%) contained distinct ORFs out of frame with *gIII*, that were not expected to be captured by affinity selection, and they also represent 'background'.

Table 4.4 Distribution of 40 analysed inserts in regard to ORF frame status.

ORF status	**RAC 1st**		**RAC 4th**		**AXYL 1st**		**AXYL 4th**	
	Count	%	Count	%	Count	%	Count	%
In frame [a]	3	42.9	1	11.1	2	40	7	36.8
Background [b] (≤24 aa)	3	42.9	2	22.2	2	40	4	21.1
Background [c] (Out of frame)	1	14.3	6	66.7	1	20	8	42.1
Total	7	100	9	100	5	100	19	100

Abbreviations: RAC, regenerated amorphous cellulose; AXYL, insoluble wheat arabinoxylan; aa, amino acid residues. [a] In frame, ORFs encoding putative proteins >24 aa in frame with vector-encoded pIII; [b] Background (≤24 aa), ORFs encoding putative proteins and peptides ≤24 aa in frame with vector-encoded pIII; [c] Background (Out of frame), 'background' inserts containing ORFs encoding putative proteins and peptides that are not in frame with vector-encoded pIII; Total, total ORFs analysed per round.

Selection for 'background' inserts containing 'out of frame' ORFs during panning procedure has been reported previously [437-439]. It has been hypothesised that some of these 'out of frame' constructs may have a selective advantage, and programmed ribosomal frameshifting (PRF) events have been implicated [440, 441]. PRF is a translational recoding mechanism involving ribosomal slippage that leads to a switch of reading frame in –1 and +1 directions, or translation past stop codons when a ribosome encounters a specific signal in the mRNA sequence [442]. Strong stimulators of –1 PRF are a combination of a heptameric 'slippery sequence' of the type X XXY YYZ (where X represents any nucleotide, Y represents A or U and Z represents A, U, or C), a 1 – 15 nucleotide long spacer sequence, and a sequence that can form a stable secondary hairpin structure (pseudoknot).

To investigate whether the putative proteins encoded by 'out of frame' ORFs could have been brought in frame with pIII due to PRF events, resulting in their display and capture by affinity selection, inserts were inspected for putative sites that can stimulate –1 PRF using the KnotInFrame algorithm [407]. Around 31% of all inserts that contained ORFs out of frame with *gIII* had predicted –1 frameshift sites. However, none of the –1 PRF events stimulated from these putative sites could have led to a frameshift into the pIII frame and, for this reason, could have not contributed to display of putative proteins encoded by the 'out of frame' ORFs.

Putative proteins longer than 24 amino acid residues in frame with pIII were further analysed for the presence of membrane-targeting signals as described in section 2.2.7.3 (Appendix 3). None of the analysed putative proteins captured in the first and the last round of panning on RAC were predicted to contain membrane-targeting signals. In contrast, among the 19 putative proteins captured in the last round of panning on AXYL, two contained putative type I signal sequences, and one contained a predicted N-terminal transmembrane helix. In addition, four putative proteins (two captured in first, and two captured in the last round of panning on AXYL), that did not contain a signal sequence, or transmembrane helices, had SecretomeP 2.0 [259] scores higher than 0.5, indicating their possible secretion *via* non-classical pathways. Based on this analysis, five recombinant phagemids were selected for the affinity binding assays to AXYL.

To explain the high proportion of putative proteins without predicted membrane-targeting signals, or shorter than 24 amino acid residues, putative proteins encoded by ORFs in frame with *gIII* were scanned for motifs conferring possible competitive advantage and/or substrate-unrelated binding as described in section 2.2.7.3. Around 30% of all putative proteins in frame with pIII from the first round and 46% from the fourth round of panning on both substrates were predicted to confer possible propagation advantage and/or putative motifs for binding to unrelated targets such as plastic or IgG class of antibodies, which are present as low-level contaminants in the BSA used for substrate blocking during panning (Appendix 3). When putative proteins containing these motifs from the first and last round of panning are

compared, the overall proportion of putative proteins encoded by background ORFs increased from 10% in the first, to 23% in the last round, while the proportion of putative proteins longer than 24 amino acid residues was the same.

Functional annotation of the 13 identified ORFs encoding putative proteins longer than 24 amino acid residues in frame with pIII, based on best BLASTP (E-value <1e-05) against the sequences in the NCBI nr protein database, resulted in a large number of assignments to conserved hypothetical (5) and hypothetical (3) proteins. One ORF (AXYL236) encoded a putative protein involved in protein transport across the membrane, while four ORFs (AXYL225, RAC261, RAC263 and AXYL342 were predicted to encode putative enzymes (Appendix 3). No putative CBMs were identified through a HMM-based search *via* dbCAN web server and BLASTP searches against CAZymes in the dbCAN and CAT databases.

4.4 Affinity-binding assays

Affinity-binding assays were carried out (as described in section 2.2.7.4) to compare the recovery of PPs derived from five recombinant phagemids (4 from the last and one from first round of panning on AXYL) to control (pDJ01 vector-containing PPs) using AXYL as bait. Three of the selected recombinant phagemids contained ORFs encoding putative proteins with predicted membrane-targeting signals (AXYL223 and AXYL232 type I signal sequence and AXYL221 one N-terminal transmembrane helix). The other two selected recombinant phagemids contained ORFs (AXYL225 and AXYL342) encoding putative proteins that were indicated as being secreted *via* non-classical secretion pathways.

Phagemid particles were derived from five affinity selected recombinant clones with wt and Δ*gIII* helper phage, and subjected to native virion gel electrophoresis (section 2.2.3.5) to assess their packaging into intact PPs. Packaging into non-defective PPs with Δ*gIII* helper phage, based on the endogenous signal sequence encoded by recombinant phagemid ORFs, could only be confirmed for two recombinant phagemids with ORFs encoding putative proteins predicted to contain type I signal sequences (data not shown).

Affinity-binding assays on AXYL with infectious PPs produced with wt helper phage were performed in a single round and under the same conditions that were previously used in metagenomic library panning, with the exception that all PPs were eluted with host TG1 cells. The fold recovery of tested PPs over the control (pDJ01 PPs) was calculated and is presented in Table 4.5.

Table 4.5 Recovery of PPs in affinity binding assays on AXYL.

Tested PPs ID	ORF length (aa)	Protein description [a]	Predicted ss [c]	Fold recovery of tested PPs (over vector-only control)
AXYL221	101	Hypothetical protein [b]	TMH	1×
AXYL223	112	Hypothetical protein [b]	Type I ss	3×
AXYL225	98	Concanavalin A-like lectin/ Glucanases [*Prevotella* sp. CAG:474]	Possible non-classical secretion	13×
AXYL232	77	Secreted protein [*Bacteroides* sp. CAG:545]	Type I ss	10×
AXYL342	216	Serine/threonine protein kinase [*Veillonella* sp. oral taxon 780]	Possible non-classical secretion	400×

Abbreviations: aa, amino acid residues; ss, signal sequence. [a] Protein description is based on comparison of the putative proteins encoded by five tested recombinant phagemid ORFs with the sequences in the NCBI nr protein database by BLASTP using default parameters. Annotation is based on best BLASTP hit with an E-value of <1e-05. [b] Putative proteins that had no hit with an E-value of <1e-05 were annotated as hypothetical. [c] Membrane-targeting signals were predicted as described in section 2.2.7.3: Type I ss, classical ss; TMH, transmembrane α-helix; Possible non-classical secretion, secretion *via* non-classical pathways indicated by SignalP 2.0 score >0.5.

Phagemid particles produced from four affinity-selected recombinant clones (AXYL221, AXYL223, AXYL225 and AXYL232) had 1 – 13 fold higher recovery compared to the control PPs derived from pDJ01, indicating potential binding to the substrate. However, because of the intrinsic titration error (estimated standard error of the mean ± 32.5% [386]), the observed differences may not be due to increased binding. AXYL342 PPs, produced from the affinity-selected recombinant phagemid containing ORF predicted to encode putative protein sharing 37% sequence identity with a serine/threonine protein kinase (E-value = 2e-48), had an approximate 400-fold higher recovery compared to the control on AXYL, indicating a relatively strong interaction with the immobilised substrate (Table 4.4). The putative protein encoded by AXYL342 also contained predicted protein motifs conferring a possible propagation advantage and/or substrate-unrelated binding to IgG class of antibodies. Contamination with IgG is present at low level (up to 0.05%) in the BSA used for substrate blocking during panning and it is possible that the observed binding of AXYL342 PPs might have been due to substrate-unrelated binding.

4.5 Summary

In the search for rumen microbial proteins involved in binding to complex carbohydrates, recombinant PPs from a rumen plant-adherent metagenomic phage display library were affinity screened using RAC and AXYL substrates as bait.

Through four rounds of panning, a small (22-fold) enrichment in the number of metagenomic PPs recovered by TG1 host cell elution from AXYL in the last round, relative to the first round of panning was observed. This, together with an observed increased binding over background (vector-only control), indicates that specific, but low-affinity interactions may have occurred between PPs and AXYL.

Five recombinant PPs, produced from four clones obtained after the fourth round of panning and one clone obtained after first round of panning on AXYL, were assayed for affinity binding to AXYL. The assay showed a 400-fold increase in binding of one of the purified recombinant PPs (AXYL342) compared to the vector-only control, indicating strong binding. The putative serine/threonine protein kinase encoded by AXYL342 ORF was predicted to be secreted *via* non-classical secretion. This protein also contained putative protein motifs that confer a potential propagation advantage and/or substrate-unrelated binding to the IgG class of antibodies, present as low-level contaminants in BSA preparations. Thus, it presently cannot be excluded that the high level of observed binding of the AXYL342 PPs might have been due to substrate-unrelated binding during the panning and affinity binding assays.

Chapter 5. Discussion

Improving the digestive process of ruminant animals, or the processes of biofuel production from lignocellulosic biomass, requires an understanding of the enzymatic depolymerisation of plant cell wall structural carbohydrates. The majority of the information currently available has been generated from the studies of individual microbes and their enzymatic capability. However, in nature the breakdown of plant polysaccharides is initiated by microbial consortia and their secreted enzymes. Microbial consortia are much more difficult to study than individual species and the recent development of high-throughput sequencing and associated metagenomic techniques has opened up new opportunities to begin to understand the complex process of microbial polysaccharide breakdown.

In this study, secretome-selective phage display technology was used to focus next-generation sequencing and sequence analysis to the metasecretome-encoding portion of the rumen plant adherent microbial metagenome. This is, to my knowledge, the first report of selective sequence analysis as a method to focus on the sequences encoding secreted proteins from a metagenome.

5.1 New phage display approach to select for the metasecretome

A pilot library was initially constructed for the development and optimisation of the secretome display technology on a metagenomic scale, and for the preliminary assessment of selection stringency and diversity of enriched metasecretome library clone inserts. Membrane targeting signals were detected in around 94% of the sequenced library inserts, which is a higher proportion than reported for the secretome phage display of two Gram-positive bacteria, *L. rhamnosus* (84%) [234] and *M. tuberculosis* (70%) [394].

The estimated enrichment of the metasecretome protein-encoding recombinant library clones after selection (29 fold) indicates a high stringency of selection, and that the majority of recombinant phagemids containing non-secretome protein-encoding inserts were eliminated. The proportion of metasecretome ORFs encoding putative proteins predicted to contain membrane targeting signals in the pilot metasecretome library (~96%) was comparable to those reported for secretome ORFs of both *L. rhamnosus* HN001 and *M. tuberculosis* obtained after sarkosyl selection (98%) [234, 394].

Type I signal sequences were predominantly detected in the pilot library pIII fusions (64%). This is not surprising, taking into account that the targeting of wt pIII to the inner membrane of Gram-negative bacterial host depends on the presence of the type I signal

sequence, which is cleaved off by SPaseI, releasing the N-terminus from the membrane, and leaving the C-terminus anchored in the inner membrane, which is the topology that is required for assembly of pIII into a detergent-resistant virion cap [345]. The remaining 32% of the pilot library ORFs encoding putative proteins in frame with pIII contained other types of Sec-dependent membrane-targeting sequences (transmembrane α-helices, type II and type IV), which is consistent with the dependence of phage display on the SecYEG translocation pathway. Similar to reports for the secretome-selective phage display of two Gram-positive bacteria, no Tat signal sequences were detected. This is due to the known inability of conventional phage display systems to present proteins that are substrates for the Tat translocation pathway [443] characteristic for fully folded proteins, most likely because protein-pIII fusion typically folds in the oxidising environment of the *E. coli* periplasm, in contrast to the Tat-dependent proteins that fold in the reducing environment of the cytoplasm [444]. The distribution of membrane targeting signals observed in the pilot library indicates that different types of Sec-dependent signals, are successfully processed by various *E. coli* membrane or periplasmic proteases, to achieve the topology of pIII required for the virion assembly and release (periplasmic N-terminus, and a single C-terminal membrane anchor) [445]. Therefore, the diversity of rumen plant adherent microbial metasecretome sampled through selection was not limited only to type I signal-containing sequences.

Functional annotations of putative proteins encoded by ORFs in frame with *gIII* showed that these proteins are involved in diverse functions, such as putative enzymatic functions (including carbohydrate metabolism), host/microbial interactions, signal transduction and sporulation, as well as large proportion of hypothetical and conserved hypothetical proteins.

The taxonomic assignment of the pilot secretome library showed that even with the limited number of analysed clones, the inserts were assigned to eight different genera. Seven genera, mainly representatives of the core bovine rumen microbiome [7, 8, 60, 68-70, 76, 77], belonged to predominant rumen phyla Firmicutes and Bacteroidetes (representing 23% and 22% of analysed sequences, respectively). The remaining genus (*Fibrobacter*), representing 5% of the analysed sequences, and belonging to the minor core rumen phylum Fibrobacteres, was found to represent only 0.8% of analysed bacterial *rrs* sequences in the study of global ruminal phylogenetic diversity [7]. The proportion of inserts assigned to the ruminal cellulolytic bacterial genus *Fibrobacter* (5%) was unexpected. Studies investigating the abundance of three predominant culturable species of ruminal cellulolytic bacteria (*F. succinogenes*, *R. flavefaciens*, and *R. albus*) using species-specific 16S rRNA oligonucleotide hybridisation probes, placed *F. succinogenes* at 0.1% of total bacterial population in the sheep rumen and 0.2 – 0.3% of total ruminal bacterial population in cows fed diets that varied in the source of forage and the concentration of fibre [446, 447]. It is possible that Fibrobacteres are overrepresented simply by chance, due to the low number of analysed pilot library clones.

However, a recent study of plant adherent rumen community composition of four cows (including the animal from which plant-adherent microbial fraction was obtained for construction of shotgun metagenomic libraries described here) over four seasons using molecular methods, placed the average abundance of phylum Fibrobacteres in plant-adherent microbial fraction from these samples at 2.4% [448]. This, in combination with the *E. coli* host used in the metasecretome selection process being likely to favour phage display of fragments from Gram-negative bacteria, could have contributed to observed proportion of this genus. Half of all analysed sequences did not show significant similarity to any sequences in the NCBI nr protein database.

We have estimated that the rumen bacterial metasecretome size is around 600 Mb, based on an estimated bacterial metagenome size of around 3 Gb [223], and the average coding capacity of bacterial genomes devoted to metasecretome proteins (20%). A large number of library clones would be needed to represent such large metasecretome diversity. Based on coverage of the *L. rhamnosus* HN001 secretome achieved using the same secretome-selective phage display system (approximately 44%), it was estimated that similar coverage of metasecretome proteins from a complex microbial community, comprised of up to 100 species, could be achieved with a starting primary library of 10^8 clones, and analysis of approximately 50,000 clones after selection [234]. Taking into account the higher bacterial diversity in a complex rumen microbial community, conservatively estimated at 500-1000 species, the size of the primary metagenomic library, as well as the number of sequenced metasecretome inserts after selection would have to be at least 5-10-fold higher in order to achieve around 40% coverage of the rumen microbial metasecretome.

In the pilot metasecretome library, some of the ORFs were isolated multiple times among the analysed clones. Cloning and transformation biases in combination with observed low transformation efficiency of ssDNA could have all contributed to the reduced insert diversity. To improve the representation of metasecretome fraction captured through selection, the secretome selection protocol was applied on a larger (5×10^6) primary shotgun library. Furthermore, the selected metasecretome library inserts were directly sequenced using the high-throughput pyrosequencing, bypassing the bottleneck of transforming the secretome ssDNA.

5.2 Metasecretome characterisation by next-generation sequencing

5.2.1 Membrane-targeting signals and phylogenetic profile of the metasecretome

This thesis developed protocols for template preparation for next-generation sequencing starting from metasecretome-enriched ssDNA, as well as data analysis in order to extract the information on the secretome proteins within the selected metasecretome library. Template preparation included a PCR-amplification step using ssDNA isolated from sarkosyl-resistant PPs, followed by mechanical and enzymatic fragmentation. Fragments obtained in this way were used directly in the standard shotgun pyrosequencing protocol, at the point of end-repair and size-fractionation of metagenomic fragments.

In order to investigate the presence of three types of membrane-targeting signals commonly observed in the pilot metasecretome phage display library, it was important to first identify sequences in-frame with vector-encoded pIII in the metasecretome dataset. Identification of 'in-frame' sequences in the raw sequence dataset (before removal of the vector sequences) was achieved by identifying the short vector sequence encoding peptide tag, preceding the vector-encoded pIII using BLAST. For prediction of membrane-targeting signals, a sequential combination of direct predictions for putative proteins in-frame with pIII and predictions assigned based on sequence similarity between 'in-frame' putative proteins without detected membrane-targeting signals, and longer homologous putative proteins in the metasecretome assembled dataset was applied. Around 59% of the distinct putative proteins contained one of the three putative membrane-targeting signals commonly observed in the pilot library (type I or type II signal sequences, and transmembrane helices). For around 28% of the putative proteins, none of the three types of membrane-targeting signals were detected, while for the remainder, the presence of the membrane-targeting sequences could not be determined, due to the short length of sequence and lack of significant similarity to longer putative proteins in the assembled metasecretome dataset. In comparison, around 11.8% of the putative proteins in the metagenome dataset contained type I and type II signal peptides, while around 9.2% contained predicted transmembrane helices.

Because of the fragmentation step in the sequencing template preparation, it is likely that for some partial ORFs that are in frame with *gIII*, the N-terminal part of the ORF encoding the membrane-targeting signal, is found on a separate sequencing read. Due to the high complexity of the metagenomic DNA before selection, and subsequent complexity of the enriched metasecretome-encoding DNA after selection, a low depth of sequence coverage is expected for the volume of sequencing performed in this study. This has led to a limited fraction of reads that were fully assembled into contigs (42%), resulting in a high proportion of ORFs

lacking significant similarity to longer homologous metasecretome contig ORFs, some of which might encode an N-terminal signal sequence.

The proportion of ORFs encoding putative proteins with membrane targeting signals in the pyrosequencing dataset is not a measure of efficiency of selection prior to shearing. However, it can indicate that, at a larger scale, selection was not as stringent as in the pilot library, where almost 95% of putative proteins possessed one of the three types of membrane targeting signals analysed in the metasecretome pyrosequencing dataset. This result could have been a consequence of technical issues with the secretome selection step in some of the 27 library aliquots, which were separately subjected to the secretome selection protocol. One such problem could have been incomplete DNase digestion of the ssDNA released from the sarkosyl-disassembled defective PPs that contained the non-secretome inserts, the key step in eliminating the non-secretome protein-encoding recombinant phagemids.

The majority of taxonomic assignments of the predicted protein-coding sequences in the metasecretome dataset were to Bacteria (40.9%), while around 59% of the protein-coding genes remained unassigned. Such a high percentage of taxonomically unassigned protein-coding genes is expected, even for the 'shallow' level of binning, since the accuracy of similarity-based binning approaches used to produce the taxonomic profile of the metasecretome-enriched dataset depends on the availability of reference data, which is currently very limited for environmental microbial communities, such as that from the rumen. However, initiatives such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project [449] and the Hungate 1000 project [450] aim to provide genome sequences from organisms that are not currently represented and systematically fill the gaps in representation of taxonomic groups for which little genomic data is available, and the availability of such data will greatly enhance our ability to make taxonomic assignments.

Archaeal assignments (0.2%) were predominantly to the Euryarchaeota phylum, with some contribution from the Crenarchaeota phylum, and this taxonomic distribution is in agreement with what was observed in study of phylogenetic diversity census of ruminal microbiomes [7]. Archaea typically comprise 0.5 - 3% of the rumen microbiome [451] and a low taxonomic assignment (0.2%) of the metasecretome ORFs to this microbial domain is most likely further diminished by the poor association of Archaea with the plant adherent fraction of the rumen microbial community, as well as a lack of reference gene information due to the very limited volume of archaeal sequences available in the databases. Finally, archaeal signal sequences have a hybrid character, combining a archaeal-specific composition of hydrophobic region with a bacterial-like charge distribution, and a eukaryal-like cleavage site, hence they may be poorly processed by the *E. coli* SecYEG translocon or the signal peptidase [452]. Nevertheless, preliminary experiments (D. Gagic, personal communication) showed that archaeal sequences successfully guide archaea-pIII protein fusions to the virion. Of all the

sequences assigned to Eukaryota, approximately 28% were most similar to fungi and around 14% to plants, which may reflect the presence of low levels of plant and fungal material contamination within the plant-adherent microbiome samples, and is consistent with the sampled environment.

At the phylum level, the main bacterial assignments in the metasecretome dataset were to Bacteroidetes (29%) and Firmicutes (10%), with minor contributions from different divisions of Proteobacteria, Actinobacteria, Spirochaetes and Cyanobacteria. In contrast, in the metagenome dataset, the Firmicutes were predominant (31.9%), followed by Bacteroidetes (17.2%), and a small number of assignments were made to a much larger number of phyla. In the metasecretome dataset, besides reduced diversity at the phylum level, the main taxonomic assignments are in agreement with predominant phyla found in the metagenome dataset and in the rumen [7]. However, numerical relations do not reflect the phyla abundances usually detected in the bovine rumen microbiome using molecular methods. The prevalence of Firmicutes has been reported in many studies of ruminant bacteria [7, 8, 70, 77], although some studies identified Bacteroidetes as the most prevalent phylum [71, 75, 453] in the bovine rumen. Based on a number of 16S rRNA gene based studies of bacterial diversity in the rumen, the estimated abundance of Gram positive and Gram negative bacterial species in the rumen is 58% and 42%, respectively (B. Kelly, personal communication) and a similar proportion was observed in [7]. Notably, the higher representation of sequences from Gram-negative bacteria in the metasecretome dataset might be due to higher efficiency of the secretome-selective phage display system to recognise Gram-negative membrane targeting signals, which was also indicated by taxonomic representation of the secretome inserts in the pilot library.

A higher proportion of putative protein-coding sequences from the metasecretome dataset (59%) was unassigned, as compared to sequences from the metagenome dataset (38%). This observation is consistent with the secretome enrichment, since the secretome fraction of the proteome typically contains a much larger proportion of unique proteins compared to the rest of the proteome, but could have also been contributed by the ability of the selection procedure to enrich for sequences encoding the putative metasecretome from rare phylotypes within the rumen microbial communities, that are otherwise inaccessible *via* low-depth metagenomic studies.

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

Functional annotation of the metasecretome was compared to that of a metagenome dataset, obtained by shotgun sequencing of the total plant-adherent rumen microbial

metagenomic DNA of two New Zealand cows fed a similar diet to the cow from which the metasecretome dataset was derived.

A study by Noel (2013) [448], comparing rumen plant-adherent bacterial community composition of four New Zealand cows (including the cow from which the metasecretome dataset was derived) representative of the national dairy herd on a pasture-based diet sampled five times over one year, using DGGE profiles of the 16S rRNA gene V3 region and pyrosequencing of the 16S rRNA gene V1-V3 region amplicons, demonstrated remarkable similarity between all the animals. In that work, only 0.1-0.6 % OTUs were estimated to be unique to each animal. Principal coordinate analysis of pyrosequences showed no clustering due to animal variation and a weak clustering of animals related to the sampling month, and the plant-adherent microbial communities appeared to be very similar among animals fed a similar diet. Although minor seasonal variations are expected between the adherent microbial communities from which the metasecretome and metagenome datasets were derived, observations from the Noel study [448] were the basis for selection of the particular metagenome dataset for comparison with the metasecretome.

The profile of COG-based functional categories, detected at over 5% of their relative abundance in the metagenome dataset, is similar to the COGs profile reported for another rumen metagenome [454]. High relative abundances of genes encoding putative proteins with unknown and general function, as well as proteins predicted to be involved in carbohydrate transport and metabolism, and amino acid metabolism, have been also reported in metagenomic surveys of other fibre-degrading gastrointestinal microbiomes [427, 455, 456].

Comparison of the COG-based functional categories between the metasecretome and metagenome datasets revealed that the most abundant functional category in both datasets was ‘carbohydrate transport and metabolism’ (19.4% and 10.7%, respectively), which is consistent with the involvement of the rumen plant-adherent microbiome in initiating carbohydrate metabolism. This functional category was also the most prominently enriched after selection. A comparable enrichment of genes encoding putative proteins belonging to this category (23%) was reported for fosmid library clones containing metagenomic DNA fragments from human gut microbiota after applying a multi-step functional screening for carbohydrate-active enzymes involved in the degradation of dietary fibre [17]. The aim of both approaches was to focus the sequencing effort onto metagenomic DNA fragments enriched in genes encoding the putative proteins involved in microbial fibre digestion. However, the single-step secretome selection presented in this thesis led to a greater coverage of rumen metasecretome (55 Mb) compared to the published fosmid approach (1 Mb), which relied on laborious high-throughput functional screening for enzymatic activities.

Metasecretome phage display also enabled the enrichment of genes with putative protein products predicted to be involved in the ‘cell wall/membrane/envelope biogenesis’,

processes known to require transport across the membrane(s) [457], as well as the enrichment of peptides with unknown function. Proteins of unknown function are generally overrepresented in the secretome fraction of bacterial genomes [230], and their enrichment is consistent with enrichment of the metasecretome. The functional categories of 'replication, recombination and repair' and 'translation, ribosomal structure and biogenesis', were the most under-represented in the metasecretome dataset, reflecting the predominantly intracellular localisation of proteins belonging to these functional categories.

5.2.3 The metasecretome selection captured diverse CAZymes

The plant-adherent rumen microbiome is specialised for the initial degradation of plant fibre through the action of surface-associated and secreted enzymes. In accordance, the metasecretome-selective phage display approach considerably enriched genes encoding secretome proteins in the 'carbohydrate transport and metabolism' functional category. The mining effort was focused on genes encoding proteins from this functional category (CAZymes and cellulosome components) in order to explore the diversity and potential novelty of putative proteins involved in fibre degradation, captured through metasecretome selection in combination with next-generation sequencing.

For annotation of the putative CAZyme and cellulosome modules in three sequence datasets derived from bovine plant-adherent microbial communities: plant-adherent metasecretome and metagenome and a published switchgrass-adherent deep-sequenced metagenome (DMG) f, an approach based on dbCAN database CAZyme family-specific HMMs [399] was chosen over the sequence homology-based, or Pfam domain-based annotation approaches. These dbCAN HMMs are derived from the signature domain regions of the most complete collection of metagenomic CAZyme genes [399]. Annotation based on sequence similarity shared with existing known CAZymes has the greatest accuracy, but limits the ability to annotate CAZymes with low sequence identity and potentially divergent biochemical properties [22]. On the other hand, CAZy family annotation using the Pfam protein families database decreases both the specificity and sensitivity of annotation, because not all of the existing CAZy families are represented by Pfam models [406]. Annotation of CAZymes in the published DMG dataset *via* dbCAN HMMs detected a greater diversity of putative catalytic and accessory CAZyme modules (belonging to 332 CAZy families) and with five-fold higher frequency compared to the originally reported Pfam-domain based annotation [22].

The diversity of sequences encoding putative catalytic, auxiliary and carbohydrate-binding modules and cellulosome components, belonging to 196 CAZyme families captured by

the metasecretome selection was compared to the diversity of sequences encoding putative modules belonging to 318 CAZyme families present in the plant-adherent metagenome dataset.

Both datasets contained sequences encoding putative oligosaccharide-degrading enzymes, cellulases, hemicellulases and debranching enzymes, as well as enzymes involved in the removal of ester linkages in hemicelluloses and in the degradation of pectins.

Comparison of frequencies with which CAZyme hits were observed in three sequence datasets subjected to dbCAN analysis (metagenome, metasecretome and published DMG) showed that the metasecretome selection allowed detection of ~2.3 putative CAZyme modules for every 10 Kb of analysed sequence, in contrast to ~1.5 in the metagenome and ~0.7 in the DMG datasets. The frequency of detection of putative partial CAZyme-encoding genes in the metasecretome dataset is at least tenfold higher than average frequency (~0.2 CAZyme encoding genes /10 Kb) observed from shotgun metagenomic sequencing from rumen [23] and gut environments [458]. The average number of detected putative CAZyme-encoding genes per 10 Kb of analysed sequence in the metasecretome dataset is also higher than detected through functional screening strategies of fosmid libraries containing metagenomic DNA from plant-adherent rumen microbiome (~1.2 putative CAZyme module/10 Kb) [459] and the number reported in the two-step functional screening of metagenomic inserts from human gut (~1 putative CAZyme module/10 Kb) [17]. Observed differences might be contributed, besides selection, by higher density of CAZymes harboured in the fibre-adherent microbial fraction of cow rumen compared to other rumen and gut environments [18].

The glycoside hydrolase/glycosyl transferase (GH/GT) ratio has been used previously to estimate the enrichment of catabolic genes [17]. Metasecretome selection led to 4-5 fold higher GH/GT ratio (9.6) than those observed in the metagenome and DMG datasets (2.3 and 2 respectively), and in functional screening of plant-adherent rumen metagenome (1.9) [459]. The plant-adherent rumen microbiome is expected to have, on average, a high proportion of its metasecretome dedicated to the catabolism of carbohydrates.

The results presented here indicate that the metasecretome-selection strategy targeted to the plant adherent microbial metagenome, known to have a diverse and rich coding potential for fibrolytic activities, was powerful in focusing the sequencing onto metagenomic DNA fragments rich in sequences encoding diverse putative CAZyme modules, particularly catabolic genes.

5.2.4 CAZyme families enriched in the metasecretome

Putative cellulosome components (dockerins and cohesins), as well as carbohydrate-active enzyme classes GHs, CEs and PLs, were predicted to occur at higher frequencies in the

plant-adherent metasecretome dataset compared to the plant-adherent metagenome dataset. The selectivity of the method for secretome proteins was particularly apparent when the abundance of GTs, enzymes involved in biosynthesis of various glycans (glycoproteins, glycolipids, oligosaccharides) and 'signature' cellulosomal modules, cohesins and dockerins, were compared to corresponding groups in a metagenome dataset. The relatively lower representation of GTs in the metasecretome (6.3%) compared to the metagenome (23.5%) is consistent with the predominantly cytoplasmic localisation of GTs [460]. Conversely, sequences encoding putative proteins containing cohesin and dockerin domains were represented at higher frequency in the metasecretome compared to the metagenome dataset. This is consistent with cohesin and dockerin-containing proteins being secreted, or membrane-bound to form cellulosomes, as described for several anaerobic bacteria, notably *C. thermocellum* (for which a re-classification to genus *Ruminiclostridium* [176] within the family Ruminococcaceae [177] has been proposed) and *C. cellulovorans*, and *R. flavefaciens* FD1 [204, 461].

Putative CAZyme-encoding gene fragments that were more frequent in the metasecretome compared to the metagenome dataset belong to 34 families with diverse functions, 18 of which are found within the top 25 hits in the metasecretome dataset, with a frequency >1%. The sequences encoding putative dockerin module and dockerin repeat containing proteins, which have been scarcely observed in other rumen microbial metagenomic studies [19, 22, 23], have been detected with 12.5-fold higher frequency in the metasecretome compared to the metagenome dataset. Hits to putative dockerin modules and repeats were also the most prevalently detected hits in the metasecretome dataset (8.3%), indicating that enrichment of gene fragments encoding these putative modules allowed their successful detection. In contrast, putative cohesin modules, also represented with higher frequency in the metasecretome compared to the metagenome dataset (3 fold), were not detected as frequently as shorter dockerins.

Interestingly, some of the generally low abundant (0.1 – 0.9%) sequences encoding putative CAZyme domains in the metasecretome dataset are nevertheless enriched relative to the metagenome dataset. For example, CBM16, CBM38, GT94 and CBM40 were significantly enriched between 6.8 and 10.6 fold in the metasecretome relative to the metagenome dataset.

Other enriched predicted CAZyme-encoding ORFs included representatives of CAZyme families involved in degradation of major fibre components, such as cellulase (GH124), endohemicellulases (GH53, GH26), xyloglucanases (GH16, GH74), side chain degrading enzymes (acetyl xylan esterase CE3, CE7 and CE1 and β-L-arabinofuranosidase GH127), oligosaccharide degrading enzymes (β-xylosidases GH3, GH30, GH43 and 1,4-β-manosidase GH2) and pectin-degrading enzymes (GH53, CE8).

Some of the ORFs moderately enriched in the metasecretome dataset encode putative GH43, GH26, GH3, GH105 modules, considered important in degrading cellulose, xylan and

polygalacturonan [459]. These modules have been detected in polysaccharide utilisation loci (PULs) mainly of uncultured Bacteroidetes, as well as in *Prevotella* from metagenomic DNA originating from the rumen [23, 24, 459, 462] and gut [20] microbiomes. It is tempting to speculate that the selection procedure, successful in enriching for the metasecretome genes from Bacteroidetes, might have led to detection of rare and otherwise inaccessible components of the metasecretome belonging to PULs of Bacteroidetes.

Some putative CAZyme modules enriched in the metasecretome dataset are involved in carbohydrate metabolism that is not associated with fibre degradation. For example, hits to the GH25 family are abundant, representing 3.4% of CAZyme hits in the metasecretome dataset (enriched 2.9 fold relatively to the metagenome dataset). GH25 enzymes are lysozymes involved in degradation and remodelling of the carbohydrate backbone of bacterial peptidoglycan and in the release of phage progeny at the end of the phage lytic cycle and both of these are membrane-associated functions.

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

Because of the large number of incomplete gene fragments in the metasecretome dataset, co-localisation of only a limited number of partial sequences encoding putative CAZyme modules in the same polypeptide was detected. Some of the predicted associations are well-documented in the CAZy database and can be searched *via* the CAZymes Analysis Toolkit (CAT) web service [406]. The association of GH16 with CBM4 is common in bacterial CAZymes with broad functional and substrate specificity. Multiple co-located CBMs, sometimes associated with enhanced affinity for carbohydrate substrates through the cooperative binding of tandemly repeated modules, have been reported in GHs (e.g. tandem CBM4 from *Cellulomonas fim*i Cel9B [463], linked CBM10 and CBM2 from *Pseudomonas fluorescens* Xyn10A [464], duplicated family 2b CBMs from *C. fimi* Xyn11A [465], three CBM6 in the *Clostridium stercorarium* thermophilic xylanase [167], CBM40 and CBM32 in the *C. perfringens* sialidase [466]). Multi-modularity of CBM32, in a combination with catalytic GH module or independent, predicted for two putative proteins in the metasecretome dataset, has been commonly observed in proteins from *C. perfringens* and *Saccharophagus degradans*. A tandemly repeated CBM50 modules associated with GH25 are typically found in bacterial lysozymes [467]. An independent association of multiple CBM50 modules without GH module is not typical for bacteria, but has been reported for secreted proteins of several species of plant pathogenic fungi and it was proposed that these proteins bind surface-exposed hitin, acting as effectors in evading recognition by plant immune receptors [468].

Putative proteins containing dockerin repeat in combination with GH5 or GH25, catalytic modules that are not usually found in cellulosomes, have also been predicted in the metasecretome dataset. *In silico* analyses of the genome of cellulosome-producing *Ruminococcus flavefaciens* FD-1 identified twelve GH5-encoding ORFs containing also putative dockerin coding sequence and predicted signal peptides, indicating secretion of these proteins [12].

5.2.6 Phylogenetic diversity of cellulosome components predicted in the metasecretome

Metasecretome selection led to the enrichment of putative dockerin and cohesin modules, 'signature' components of the complex carbohydrate-degrading cell-surface bound multi-enzyme complexes - cellulosomes. Secreted and membrane-bound cellular localisation has been described for cohesin- and dockerin- containing cellulosomal components of several anaerobic bacteria, notably *Clostridium thermocellum* (a re-classification as *Ruminiclostridium thermocellum* has been proposed [176]), *C. cellulovorans*, and *R. flavefaciens* FD1 [204, 460, 461], and predicted for non-cellulosomal bacterial proteins containing cohesins and dockerins [198].

A striking difference in comparison with reports from previous rumen microbiome studies and the plant-adherent metagenomes used for comparison lies in the presence of a high proportion of cohesin and dockerin modules that have been rarely detected in previously published metagenomic studies of the cow rumen microbiome [19, 22]. For example, comparison of the abundance of cellulosome-associated modules in the metasecretome dataset, with those in a switchgrass-adherent bovine rumen microbial deep-sequenced metagenome DMG dataset [22], predicted using the same database and search parameters [399], showed a prominent enrichment for cohesin and dockerin modules. Other published rumen metagenomic datasets have detected even lower proportions of cellulosomal modules [19, 23, 24] than in the switchgrass adherent dataset.

The majority of the putative dockerin and cohesin modules in the metasecretome dataset showed strong homology to sequences from members of the Ruminococcaceae family [176]. This finding is consistent with the taxonomic affiliations of known cultivated cellulosome producing-bacteria, which also predominantly belong to the Ruminococcaceae [175]. This suggests that within the plant-adherent rumen microbial fraction, members of the Ruminococcaceae also have the greatest potential to produce cellulosome-like structures. A number of cohesins (10%) and dockerins (7.25%) were also assigned to the Bacteroidaceae (with hits to *Bacteroides nordii*), suggesting the potential for this family to produce

cellulosomes. However, there currently are no reports of cellulosome-producing organisms from this family. One of the earliest reported cellulosome producers, *Bacteroides cellulosolvens* [469], is now recognised as a member of the Ruminococcaceae [177] and its re-classification to *Ruminiclostridium cellulosolvens* has been proposed [176].

In the metasecretome dataset, almost 18% of ORFs encoding putative dockerin modules were most similar to sequences from members of the Clostridiaceae, but curiously, no cohesins were identified from this taxon. In total, 44 sequences with hits to distinct cohesin domains were detected in the metasecretome dataset and almost 500 sequences had hits to one or both dockerin repeats (69 of which were to both dockerin repeats). The HMM within dbCAN specific for a cohesin module (168 amino acid residues) is longer than that for a single dockerin repeat (22 residues) [399]. Thus, with the metasecretome library ORFs being of small sizes, the probability of *in silico* detection of partial cohesin-encoding sequences was lower than that for identifying dockerins.

Interestingly, putative cohesin modules from *R. albus* have been detected in the metasecretome dataset. However, to date, cohesin domain-encoding genes have not been identified in the available genome sequence from cultured *R. albus* strain 8, even though this putative cellulosome producer contains many genes encoding dockerin-containing enzymes [175]. It was thus speculated that closely related organisms may produce cognate cohesin bearing scaffoldins that would enable use of the dockerin-containing enzymes produced by *R. albus* 8 [175].

A small number of dockerin and cohesin module-containing sequences in the metasecretome dataset appeared to be associated with a number of bacterial families that are not known to produce cellulosomes, such as the Coriobacteriaceae, Erysipelotrichaceae and Porphyromonadaceae. It is thus uncertain whether these are from cellulosome-producing organisms. Alternatively, they may be associated with proteins that mediate roles uncharacteristic of the conventional polysaccharide-degrading cellulosomal function, such as proteolysis (proteases), oxidative reduction (peroxidases) or dephosphorylation (phosphatases) [198]. It has been proposed that in complex ecosystems, different organisms could use cohesin and dockerin modules for interspecies cell-cell adhesion or that these domains may evolve into roles that are unrelated to binding interaction [198].

5.2.7 Assessment of the novelty of CAZymes detected in metasecretome dataset

Conventional sequence homology-based CAZyme annotation has a bias towards the identification of candidates similar to known enzymes. For this reason, this approach is not

suitable for discovering novel CAZyme related proteins with low sequence identity and potentially divergent properties from environmental metagenomes [22, 399].

The ability of metasecretome selection in combination with automated HMM-based CAZyme annotation to discover putative carbohydrate active enzymes with limited overall amino acid sequence identity to known proteins was assessed.

Distinct putative CAZyme-encoding ORFs (longer than 30 amino acid residues) were compared to proteins deposited in the NCBI nr database in order to compare with the 'CAZyme discovery rate' reported for the published DMG dataset. Over one quarter (26%) of putative CAZyme-encoding ORFs in the metasecretome dataset showed less than 50% of amino acid sequence identity with sequences deposited in the NCBI nr database. In contrast, 43% of putative CAZyme-encoding genes detected in the DMG dataset was reported to have less than 50% amino acid sequence identity to putative proteins deposited in the NCBI nr database at that time [22]. The higher estimated novelty reported for the DMG dataset was achieved with 700 fold greater sequencing effort compared to this study (268 Gb of raw sequence data compared to 379 Mb in the metasecretome dataset) and with over half of originally detected CAZyme-encoding genes estimated to be complete.

Similarity searches were also performed against the specialised collection of putative proteins with annotated CAZyme domains, containing sequences from CAZy and other databases, as well as the collection of CAZymes from different metagenome projects. In 2012, this collection had about three times more CAZyme homologues compared to the NCBI nr database [399]. Over 60% of putative CAZymes identified in the rumen plant-adherent microbial fraction using metasecretome phage display shared limited similarity (50 – 90%) and over one-sixth shared less than 50% of amino acid sequence identity with dbCAN-deposited putative CAZymes available from other studies. All putative cohesin and complete dockerins modules identified in this study shared limited amino acid sequence similarity (41 – 70% for cohesins and 27 – 59% for dockerins) with putative cohesin and dockerin sequences deposited in the dbCAN database.

Among putative CAZymes predicted in this study with a significant match (E-value <1e-05) to dbCAN database entries, the majority were homologous to putative CAZymes from the published cow rumen metagenome dataset [22] and human gut microbial metagenome sequences (around 67% and 14%, respectively) [427]. Around 3.6% of putative CAZymes identified in this study shared 100% amino acid sequence identity with CAZymes predicted from the DMG dataset. This is comparable with findings from functional metagenomic mining of the fibre-adherent microbiome from cow rumen, where 84% of all putative proteins from that study were homologous to proteins predicted in the DMG dataset (reported at E-value <1e-10) [459], with 4.5% of identified putative CAZymes sharing 100% amino acid sequence identity to putative proteins in the DMG dataset.

It is indicated that the secretome selection had captured a mixture of potentially novel CAZyme-encoding sequences, suggesting that the developed approach is complementary to existing strategies used for exploring the functional potential of complex microbial communities. This also suggests that, if the sequencing of metasecretome selected genes, in combination with HMM-based CAZyme annotation, is applied on a larger scale, allowing a higher coverage of selected metasecretome, this could be an efficient strategy for mining the metasecretomes of biomass-degrading complex microbial communities for novel putative genes, and shed light on fibre-degrading strategies in the rumen.

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

Secreted and surface-associated fibrolytic enzymes and accessory proteins, produced by complex rumen plant-adherent microbial communities, have a central role in initiating fibre degradation and providing energy for the ruminant host. Accessory CBMs have a key role in binding associated catalytic domains to the substrate and targeting the degradative capacity of catalytic modules through a proximity effect [153].

The development of an efficient metasecretome-selection platform for the enrichment of genes encoding relatively lowly abundant surface and secreted proteins of complex microbial communities, and their *in silico* characterisation, is an important step towards functional analysis of metasecretome proteins encoded by these genes. Functional characterisation of (poly)peptide libraries displayed on the surface of filamentous phage typically involves biopanning of the phage display library against a target of interest, allowing affinity selection of phage clones that bind to the target and binder identification through sequencing. Biopanning of combinatorial phage display libraries of the random heptapeptides on cellulose has previously been successfully used to identify and characterise cellulose-binding peptide motifs [168, 379].

In this thesis, the suitability of a standard biopanning procedure for affinity selection of the phage-displayed plant-adherent rumen metasecretome proteins potentially involved in binding to amorphous cellulose and arabinoxylan was examined. It was recognised that it is likely that, due to the limited primary library size (before selection) and relatively low transformation efficiency of metasecretome-enriched ssDNA after the selection, the diversity of proteins displayed in the metasecretome phage display library is significantly reduced. For this reason, the PPs of the metagenomic phage display library (before the selection), produced using a wt helper phage to enable infectivity of library virions, were chosen for affinity screening, to identify proteins potentially involved in binding to the complex carbohydrate substrates of interest.

However, this library also had limitations. The primary metagenomic library size before selection was ~5×10^6 clones and the 'effectively displayed' size and diversity of secretome proteins fused to pIII is predicted to be ~3.3% (up to $1.6x10^5$ displayed protein variants). Furthermore, the display is monovalent, with typically 10% of PPs displaying one copy of the pIII fusion encoded by the phagemid, and 90% of particles displaying the helper phage-encoded wt pIII, due to a preferential assembly of the latter into the PPs [470]. The wt helper phage had to be used to allow amplification of PPs between the successive rounds of panning, given that the pDJ01 phagemid vector encodes only the C-terminal domain of pIII, necessary for packaging and release of recombinant PPs [354], and lacks the receptor-binding N1 and N2 domains, required for infection [345]. (This design was necessary to avoid resistance of phagemid-containing cells to helper phage infection, which was mediated by the N1 and N2 pIII domains).

After four rounds of panning using arabinoxylan and amorphous cellulose as baits, the inserts from 40 recombinant phagemids were sequenced. One third of the sequenced inserts carried ORFs encoding putative proteins in frame with pIII, and were further analysed. The remaining inserts carried 'background' ORFs that were either too short to be further analysed or out of frame with pIII. ORFs out of the frame with *gIII*, were analysed for the frameshifting signals that could have placed them in frame with pIII [407], however none contained a predicted frameshift into the pIII frame. Hence, they could not have been displayed on the surface of the PPs.

Functional annotation of the candidate low-affinity binders (13 ORFs encoding polypeptides in frame with pIII) resulted in a large proportion of assignments to conserved hypothetical or hypothetical proteins (69%) and putative enzymes. None of the analysed ORFs captured in the first and the last round of panning on regenerated amorphous cellulose encoded putative proteins with a classical or non-classical membrane-targeting or secretion motifs, indicating that these ORFs most likely represent background, and are not displayed on the surface of the phage.

Five monoclonal recombinant PPs (four from the last, and one from the first round of panning on arabinoxylan), selected based on the presence of putative membrane targeting signals or implicated in the non-classical secretion, were subjected to affinity binding assays to determine whether the displayed proteins have an increased affinity for the substrate.

Four tested clonally purified PPs (displaying putative proteins AXYL221, AXYL223, AXYL225 and AXYL232), showed only 1-13 fold higher recovery compared to the vector control and these differences might be contributed by a low affinity of binding to the substrate, or a titration error. In contrast, AXYL342 PPs, displaying a putative serine/threonine protein kinase fragment, were recovered from arabinoxylan with a 400-fold increase in bound PPs as compared to the vector control. No prediction of the exact function could have been derived for

putative protein encoded by the AXYL342 ORF, since it showed only a limited amino acid sequence identity to proteins deposited in the NCBI nr database. A DELTA-BLAST search showed strong homology (4e-75) between AXYL342 putative protein and a catalytic domain of a putative eukaryotic serine/threonine protein kinase. However, AXYL342 shared only limited amino acid sequence identity with kinases in the NCBI nr database. This, together with membrane localisation of its best prokaryotic homologue (2e-48) suggests that AXYL342 could encode a catalytic domain of putative eukaryote-like serine/threonine kinase [471]. *In silico* analyses have not associated this protein with recognised carbohydrate binding functions, indicating that the observed binding may have not been carbohydrate-specific. In addition, putative protein motifs conferring a potential propagation advantage and/or substrate-unrelated binding to the IgG class of antibodies, present as low-level contaminants in BSA preparations, were predicted for AXYL342. Thus, it presently cannot be excluded that the high level of observed binding of the AXYL342 PPs might have been due to substrate-unrelated binding during the panning and following affinity binding assay and would require further investigation.

Overall, the affinity screening of the metagenomic shotgun phage display library PPs against amorphous cellulose did not identify any binding domains, whereas the arabinoxylan screen identified a putative serine/threonine protein kinase as a candidate arabinoxylan-binding protein. However, substrate-unrelated binding could not be excluded for this putative protein. Affinity screening of the metagenomic phage display library on amorphous cellulose and arabinoxylan as baits was not efficient, probably due to several factors discussed below.

First, the copy number of displayed proteins is low due to competition between the secretome protein fusions to pIII with the wt pIII from the helper phage. The monovalent display is not suitable for affinity-selection of CBMs that mostly have a low affinity for their cognate substrates (K_a ~10^2 – 10^6 M^{-1}) [157]. Avidity resulting from multivalent interactions between tandem or arrayed CBMs and their ligand that can compensate these weak interactions in nature would be presumably rare in the library, due to the relatively small insert sizes. Polyvalent phage display approach could not been taken in this screen, since a full-length pIII from the helper phage is required for amplification of the selected particles through several rounds of panning, that is necessary to separate real binders from the background of non-specific binders.

Second, the display of metasecretome proteins depends on the presence of endogenous membrane-targeting signals encoded by library inserts, typically found at their N-terminus. Hence, the display is biased towards the N-terminal domains. Relatively short insert sizes observed after affinity panning (with the majority <500 nt) have contributed to a decreased number of complete domains of the secretome proteins. This could have resulted in the exclusion of CBM domains located near the C-terminus or, in the case of incomplete CBMs, preclusion of domain folding and formation of a binding surface for its carbohydrate substrate.

Third, immobilisation of carbohydrate substrates using columns could have contributed to a high background, by increasing the probability of selecting non-specific plastic binders on a column frit. In the case of panning on the RAC, unspecific binding of PPs to the slurry-like substrate, increasing the background, was very likely.

Fourth, the pIII fusions that are targeted to the membrane often result in longer generation times of *E. coli* cells in which they are expressed. The recombinant library phagemids that did not contain inserts encoding membrane-targeting signals fused to pIII would have had a growth advantage during the amplification steps between the panning rounds, outcompeting the clones encoding low-affinity substrate binders which were only marginally enriched during the binding/elution steps of the panning.

Overall, a high non-specific background binding in combination with growth advantage during amplification had very likely vastly increased the number and variety of background recombinant clones over the 'low-affinity' binders. Consequently, the analysis of only 40 recombinant clones by sequencing may have not been sufficient to identify the low-affinity binders that had been enriched, but were not dominant in the eluted library pools. The next-generation sequencing technologies have been increasingly applied for in-depth identification of the diversity of inserts following the affinity-screening after one or two rounds of panning on ligand of interest, allowing high-throughput identification of numerous binding variants and overcoming a problem of competition between the 'background' and the true binders [386-390].

5.4 Study limitations

The main limitations of the microbial metasecretome enrichment and characterisation are likely contributed by limitations inherent to phage display/secretome-selective phage display and particular experimental choices made in this study.

Prominent bias towards representation of Gram-negative bacteria observed in the metasecretome phage display library might be due to the limited ability of *E. coli* host SecYEG translocon and accessory proteins to recognise and export membrane-targeting signals from Gram-positive bacteria and Archaea.

The second major limitation is the low coverage of the complex microbial metasecretome. At the time of metasecretome dataset generation (late 2011), the 454 GS FLX platform with Titanium chemistry offered significantly longer sequencing reads and highest per base accuracy compared to other available platforms (Illumina and IonTorrent). Choosing the pyrosequencing platform required a trade-off to be made between long read length, crucial for removal of vector sequence included in the NGS template preparation from metasecretome-enriched ssDNA, and a modest amount of sequencing data that can be obtained from a half

plate. In addition, systematic artefacts intrinsic to the 454 sequencing platform in combination with two rounds of PCR, one to prepare the sequencing template from metasecretome-enriched ssDNA, followed by the standard emulsion PCR to add sequencing adaptors, most likely heavily contributed to the loss of over two thirds of raw reads in the de-replication process. The high level of 'replication' observed in the metasecretome dataset might have also been contributed by a 'bottleneck' in the starting metagenome diversity, inherent to the construction of metagenomics libraries, whose primary sizes are typically up to 10^6 different recombinant clones. Unlike the classical shotgun metagenomic approach, in which DNA is extracted from environmental microbial communities, fragmented and directly sequenced, metasecretome phage display requires construction of a shotgun metagenomic library in *E. coli* to allow assembly of metagenomic phage particles upon infection of the host with helper phage, as these PPs represent a substrate for the downstream secretome selection. Construction of a shotgun metagenomic library of sufficient primary size to fully represent the extremely complex rumen metagenome is not feasible. This leads not only to reduction in the starting metagenome diversity, but also may allow over-representation of certain library clones harbouring non-deleterious metagenomic inserts during their propagation.

Third, a relatively small insert size resulted in a large number of partial gene fragments in the metasecretome dataset that hindered both functional annotations and taxonomic assignments of the reads, as well as the detection of membrane-targeting signals. In addition, both accuracy of functional and taxonomic annotations and relative gene abundances might be affected by indels introduced in homopolymeric tracts by the 454 platform.

Another caveat is in regards to the limited power of comparison between metasecretome and metagenome datasets obtained from different animals under different experimental conditions, due to undetermined level of variability in each sample (e.g. sampling season, geographical location, and different rumen content fractionation and DNA extraction methods). For this reason, caution is needed in interpretation of these comparisons.

In addition, the experimental design of the affinity screening described in this thesis might not be well suited for identification of carbohydrate binding domains displayed on the phage surface. This might be contributed by monovalent display of fusions that do not allow for selection of low-affinity intereactions, such as those reported between many CBMs and their carbohydrate substrates, and difficulties with immobilisation of carbohydrate bait.

Chapter 6. Conclusions and future directions

6.1 Conclusions

The research presented in this thesis demonstrates that it is possible to focus sequencing efforts onto the subset of metagenomic sequences encoding secreted and surface proteins of a complex microbial community, by combining the secretome-selective phage display at a metagenomic scale with next-generation sequencing. When applied to the rumen plant-adherent microbial community, the developed metasecretome phage display approach successfully selected for genes encoding functionally and taxonomically diverse secretome proteins. The prominent enrichment of sequences encoding a subset of low-abundance surface and secreted proteins relevant for fibre-degradation allowed for the *in silico* identification of functionally diverse CAZymes, and a large number of cellulosome module-containing proteins. This was achieved at a scale that would otherwise require much larger sequencing efforts had the metagenome been sequenced directly.

Putative proteins from the rumen plant-adherent microbial community, involved in the binding and degradation of plant fibre, uncovered using the metasecretome phage display approach, represent a mixture of potentially novel CAZyme-encoding genes and genes encoding putative homologues of the CAZymes observed in previous sequence- and function- based metagenomic studies. This suggests that the approach presented in this thesis is complementary to existing strategies used for exploring the coding potential of complex microbial communities. Applied in a more targeted way, and on a larger sequencing scale, the metasecretome-selective phage display technology has potential to further expand the known repertoire of fibre degrading and binding capabilities from complex microbial communities.

The obtained rumen microbial metasecretome phage display library is a valuable resource that has allowed *in silico* identification of diverse putative proteins involved in or associated with fibre degradation. The functional identification of proteins involved in binding to complex polysaccharide substrates of interest by affinity screening was trialled, leading to the identification of one arabinoxylan-binding candidate. However, substrate-unrelated binding could not be excluded for this candidate. Further work is required to confirm the binding and to up-scale the analysis of selected library pools through next-generation sequencing in order to identify the complete complement of substrate-binding proteins enriched by affinity screening of the library.

6.2 Future directions

The discovery of microbial genes encoding novel lignocellulolytic enzymes (especially hemicellulases), allowing an efficient breakdown of plant biomass, as well as accessory proteins (e.g. scaffoldin components) aiding fibre degradation, is of considerable biotechnological interest [25]. These genes and their protein products are likely to contribute to a better understanding of mechanisms underlying fibrolytic processes and have potential applications in agriculture, industry, biomass waste management and biofuel production. In addition, they represent possible targets for improving the enzymatic performance, lowering production costs and obtaining enzymes tailored for a specific application through genetic engineering and directed evolution [3-6].

The rumen microbial metasecretome phage display library obtained from this work is a valuable genomic resource of diverse putative proteins involved in fibre degradation and binding in the rumen plant-adherent microbial fraction. At present, this library is being used to explore the diversity and roles of the identified putative cellulosome module-containing proteins through functional studies. Cohesin-dockerin pairing is crucial for cellulosome assembly and the way cellulosome architecture determines the recruitment of enzyme components, which contributes to their synergistic activities. Understanding these interactions, and identifying novel cellulosome modules that can be used as building blocks for 'custom-made' cellulosomes, is highly sought after to enable their implementation in biotechnological and nanotechnological applications, such as engineering display of designer cellulosomes on microbial cell surfaces to improve biofuel production processes [6]. Using the sequence information obtained from the metasecretome library, genes encoding five cohesins that share less than 50% amino acid sequence similarity with cohesins deposited in the dbCAN database have been successfully amplified from the metagenomic DNA of the plant-adherent microbial community. The cohesin domains have been expressed so they can be used as bait in affinity screening of the metasecretome phage display library to identify their cognate dockerins (D. Gagic, personal communication). Panning conditions have been tested using a functionally characterised cohesin-dockerin pair from *Ruminococcus flavefaciens* FD-1, namely one of the cohesin modules from the functionally characterised scaffoldin ScaA as a bait, and recombinant PPs displaying its cognate dockerin from the GH family 44 enzyme, Cel44A, on their surface [461].

The metasecretome library constructed in this study will be used in the future for exploring putative novel fibre degrading and binding domains and proteins of interest, using the existing sequence information for expression and display of the candidate genes, and their functional characterisation using standard enzymatic and/or carbohydrate binding assays.

In addition to expanding the known repertoire of fibre degrading and binding capabilities, phage display-based metasecretome selection could also be used to identify novel

candidate genes for potential biotechnological applications in a targeted fashion. Lignocellulosic biomass degradation into the individual component sugars for the production of biofuels, driven by microbial enzymes, is a promising alternative approach to the currently used mechanical and chemical procedures for saccharification, which are relatively expensive and inefficient [472]. For example, metasecretome libraries could be constructed from other rumen microbial communities that are tightly adherent to the feedstocks used for production of lignocellulosic bioethanol (miscanthus, switchgrass and corn stover). These metasecretome libraries could be used to directly express soluble metasecretome proteins (without the pIII fusion partner), by exploiting features of the secretome-selective phagemid vector (e.g. the sequence encoding the c-myc peptide tag is followed by a single amber stop codon, allowing the expression of the protein-c-myc fusion in a suppressor-negative *E. coli* host strain) [234]. Alternatively, metasecretome libraries could be subjected to affinity screening on substrates relevant for biofuel production processes to enrich for candidate binders. The biochemical activities of soluble metasecretome proteins or expressed candidate genes encoding novel putative secreted and cell-surface CAZymes and other proteins of importance could be verified by enzymatic screening using assays developed for a panel of model carbohydrate substrates.

Information obtained from the metasecretome phage display libraries could be further improved in terms of quantity and quality, by optimising individual steps in the library construction and selection. For example, to avoid the domination of short inserts in the library, stringent removal of shorter metagenomic fragments prior to their cloning into phagemid vector is required, or construction of the fosmid-based phage display vectors that would accommodate only large inserts. Furthermore, to increase the amount of metasecretome-enriched DNA available for sequence analysis and/or metasecretome library construction and bypass PCR, multiple-displaced amplification using the φ29 DNA polymerase could be used [461] for the preparation of DNA for the library construction and/or for next-generation sequencing template. Finally, coverage of the enriched metasecretome pool could be maximised by a hybrid sequencing approach, combining sequencing data generated on multiple platforms (e.g. using a combination of Illumina and 454 platforms for parallel sequencing of different-sized DNA fragments).

In order to improve screening and identification of metasecretome phage display library variants with carbohydrate binding affinity, polyvalent display in combination with next generation sequencing of PPs enriched after single round of panning on carbohydrate baits, chemically immobilised onto a matrix, could be applied.

The metasecretome phage display method presented in this thesis offers a complementary approach to existing strategies used for exploring the secretome of complex microbial communities. It can be applied for the simultaneous enrichment and screening of secretome proteins with functions of interest of any complex microbial community, bridging the

gap between sequence-based and function-based metasecretome gene discovery. Its scope is not limited to catalytic and accessory fibrolytic activities and can be applied to expand the known repertoire of various surface and secreted functions involved in interactions of microbes with their environment, animal/plant host, or other microbes.

Appendices

Appendix 1

Table A1.1. Predicted membrane targeting signals and annotation of putative proteins in the metasecretome pilot library.

Clone ID [a]	Number of clones with identical insert	Insert length (bp)	Putative protein length (aa)	Predicted membrane-targeting signal [b]	SignalP 4.1 mean D-score and predicted ss cleavage site position (G-)	SignalP 4.1 mean D-score and predicted ss cleavage site position (G+)	TMHMM 2.0 number and predicted position of TMH	LipoP 1.0 location prediction (G-)	PRED-LIPO location and ss cleavage site location prediction (G+)	PILFIND 1.0 prediction (G+ and G-)	SecretomeP 2.0 score (G-)	SecretomeP 2.0 score (G+)	Predicted signal sequence/anchor and cleavage sites [c]	Putative protein description based on best BLASTP hit [d]	Taxonomic assignment based on best BLASTX hit [e]	Accession number
MS1	1	245	27	Type I ss	0.338	0.455 (23-24)	0	Cyt	Signal (1-21)	FALSE-pattern not present	0.11	0.04	MKKIICITLMMALCLVSPHKAFA–IDF	hypothetical protein	*Prevotella*	KF790700
MS2	2	267	63	Type I ss	0.687 (27-28)	0.814 (27-28)	1 TMH (7-26), possible N-terminal ss	SpI	Signal (1-27)	FALSE-pattern not present	0.05	0.76	MRNTRRIFLAIALVIAMVFTMGTLALA–ADG	conserved hypothetical protein	*Clostridium*	KF790701
MS3	1	351	52	N-terminal TMH	0.122	0.24	1 TMH (12-34), possible N-terminal ss	TMH	Membrane (10-29)	FALSE-pattern not present	0.12	0.91	MFTKLLRRNLI*MMLIPVLVIQILII FMVIQLGL*LPA	hypothetical protein	*Ruminococcus*	KF790702
MS4	1	613	63	Type I ss	0.73 (29-30)	0.689 (29-30)	1 TMH (12-29), possible N-terminal ss	Cyt	Cyt	FALSE-pattern not present	0.84	0.97	MRTNTIKRKTLLAFVLTIAGLVAGQSAWA–TGD	hypothetical protein	Genus not assigned	KF790703
MS5	1	478	33	Type I ss	0.425 (19-20)	0.259	0	SpI	Signal (1-27)	FALSE-pattern not present	0.09	0.02	MKRLTYLFLVLFWALSAIA–QDK	hypothetical protein	Genus not assigned	KF790704
MS6	1	722	235	Type I ss	0.589 (18-19)	0.385	0	SpI	Signal (1-18)	FALSE-pattern not present	0.74	0.23	VHLYIPAGVTLTAIGANA–AGT	hypothetical protein	Genus not assigned	KF790705
MS7	1	1223	347	Type I ss	0.783 (24-25)	0.603 (24-25)	0	SpI	Signal (1-24)	FALSE-pattern not present	0.09	0.07	LKKRAALPAVLCALILGILACARA–DTP	NHL repeat protein	*Ruminococcus*	KF790706

Clone ID [a]	Number of clones with identical insert	Insert length (bp)	Putative protein length (aa)	Predicted membrane-targeting signal [b]	SignalP 4.1 mean D-score and predicted ss cleavage site position (G-)	SignalP 4.1 mean D-score and predicted ss cleavage site position (G+)	TMHMM 2.0 number and predicted position of TMH	LipoP 1.0 location prediction (G-)	PRED-LIPO location and ss cleavage site location prediction (G+)	PILFIND 1.0 prediction (G+ and G-)	SecretomeP 2.0 score (G-)	SecretomeP 2.0 score (G+)	Predicted signal sequence/anchor and cleavage sites [c]	Putative protein description based on best BLASTP hit [d]	Taxonomic assignment based on best BLASTX hit [e]	Accession number
MS8	1	716	101	Type I ss	0.852 (18-19)	0.588 (20-21)	0	SpI	Signal (1-18)	FALSE-pattern not present	0.94	0.1	MKKLMIFALMLLSMGTQA–QT–ARQ	Acetylesterase/ GDSL-like lipase/ acylhydrolase	*Prevotella*	KF790707
MS9	2	815	495	N-terminal TMH	0.383	0.305	1 TMH (27-45), possible N-terminal ss	SpI	Membrane (27-44)	FALSE-pattern not present	0.87	0.86	MHALNDKRHQPVGSSGNARNDVARWK*YALLACCMMALMMSPM MWA*QQP	conserved hypothetical protein	*Barnesiella*	KF790708
MS10	2	290	89	Type I ss	0.399	0.456 (32-33)	1 TMH (10-27), possible N-terminal ss	Cyt	Cyt	FALSE-pattern not present	0.03	0.81	MMVVDHTRTYRTIFFIFLISALVFCFSKKADA–KDL	conserved hypothetical protein	*Clostridium*	KF790709
MS11	1	666	77	Type I ss	0.764 (20-21)	0.711 (20-21)	0	SpI	Signal (1-22)	FALSE-pattern not present	0.54	0.07	MKKLLLILFTIHCSCVPAVA–SEQ	α-N-acetyl glucosaminidase	*Prevotella*	KF790710
MS12	1	311	68	Type I ss	0.887 (20-21)	0.714 (20-21)	0	SpI	Signal (1-20)	FALSE-pattern not present	0.58	0.22	MKRILFISILCLLAVSGALA–QKP	alpha-amylase	*Bacteroides*	KF790711
MS13 [f]	1	265	264	Multiple TMH	0.139	0.142	3 TMH (10-30; 37-59; 63-85), possible N-terminal ss	TMH	Membrane (38-59)	FALSE-pattern not present	0.28	0.98	**MHAP**EDIMSSLTD*YLWAFLIGGA ICTVGQVLMSL*TRLTPA*RILVLFVT SGVVLTALGLYSPVV*EAG*GAGATV PLTGFGYALATGAIEGA*KTE	stage V sporulation protein AE	*Clostridium*	KF790712
MS14	2	751	91	N-terminal TMH	0.144	0.232	1 TMH (21-43), possible N-terminal ss	TMH	Membrane (21-42)	FALSE-pattern not present	0.04	0.66	MSLISFIRYDIKRLFGHGKT*AILA VLSPIPVLLLFASFLIPFL*SAD	hypothetical protein	*Fibrobacter*	KF790713
MS15	2	341	99	Internal TMH	0.112	0.128	1 TMH (39-61), possible N-terminal ss	Cyt	Membrane (44-65)	FALSE-pattern not present	0.41	0.91	MWKDEEREQIEQAPADDEGQELTDSVSTAEDGDKASRK*KLLLSTG LPVLLTIIFFLIPLFF*KTP	hypothetical protein	Genus not assigned	KF790714
MS16 [f,g]	4	314	104	Background (no ss)	0.098	0.091	0	Cyt	Cyt	FALSE-pattern not present	0.19	0.28		conserved hypothetical protein	*Butyrivibrio*	KF790715
MS17	2	826	82	Type I ss	0.319	0.482 (35-36)	1 TMH (13-35), possible N-terminal ss	TMH	Signal (1-35)	FALSE-pattern not present	0.09	0.65	VRRFIRTGAGKTALFMLFVLSVIVLALSATGIVGA–VEY	hypothetical protein	*Clostridium*	KF790716
MS18	2	791	107	Type I ss	0.761 (22-23)	0.594 (27-28)	0	SpI	Signal (1-27)	FALSE-pattern not present	0.12	0.55	MTNRKVKIITAVITCCMFCATA–VDIYA–SSL	histidine kinase	*Ruminococcus*	KF790717

Clone ID [a]	Number of clones with identical insert	Insert length (bp)	Putative protein length (aa)	Predicted membrane-targeting signal [b]	SignalP 4.1 mean D-score and predicted ss cleavage site position (G-)	SignalP 4.1 mean D-score and predicted ss cleavage site position (G+)	TMHMM 2.0 number and predicted position of TMH	LipoP 1.0 location prediction (G-)	PRED-LIPO location and ss cleavage site location prediction (G+)	PILFIND 1.0 prediction (G+ and G-)	SecretomeP 2.0 score (G-)	SecretomeP 2.0 score (G+)	Predicted signal sequence/anchor and cleavage sites [c]	Putative protein description based on best BLASTP hit [d]	Taxonomic assignment based on best BLASTX hit [e]	Accession number
MS19	1	799	90	Type I ss	0.560 (26-27)	0.672 (26-27)	0	SpI	Signal (1-26)	FALSE-pattern not present	0.37	0.08	MKEMRKSLLFLSFCLLCACGPR GGKA–VTD	conserved hypothetical protein	*Bacteroides*	KF790718
MS20	1	900	167	N-terminal TMH	0.136	0.214	1 TMH (36-58), possible N-terminal ss	Cyt	Membrane (36-55)	FALSE-pattern not present	0.21	0.84	MNEQDYMKITNSIDERFVAEYQ TSAKVHDLTGRRR*ISIGIAVAVLV AVMIPAGVFAYN*QIT	hypothetical protein	*Butyrivibrio*	KF790719
MS21	1	935	144	Type I ss	0.744 (19-20)	0.45 (19-20)	0	SpI	Signal (1-20)	FALSE-pattern not present	0.82	0.39	MRKLTFVCCLMTLWTQALY–ADD	conserved hypothetical protein	*Prevotella*	KF790720
MS22	3	948	83	Type I ss	0.566 (37-38)	0.395	1 TMH (21-38), possible N-terminal ss	SpI	Cyt	FALSE-pattern not present	0.14	0.83	MQQPPPEVAPHIDERMNRYRLT LLLAAALLACGAAFA–QEP	tPR repeat-containing protein	*Prevotella*	KF790721
MS23 [h]	1	299	34	Type I ss	0.26	0.361	1 TMH (7-26), possible N-terminal ss	Cyt	Membrane (9-29)	FALSE-pattern not present	0.09	0.96	MNKTHP*RILILLLAVCLVLPLTEAF S*GCR	hypothetical protein	Genus not assigned	KF790722
MS24	19	865	174	Type IV ss	0.11	0.225	1 TMH (15-37), possible N-terminal ss	Cyt	Membrane (11-39)	TRUE-pattern found, TMH (15-37)	0.47	0.91	MIFTGEKKDKGFTL*VELAAVIAIM VVMTGVLTLSLSV*IYN	pilin	Genus not assigned	KF790723
MS25	1	902	170	Type I ss	0.414	0.539 (27-28)	1 TMH (7-29), possible N-terminal ss	SpI	Signal (1-27)	TRUE-pattern found, TMH (15-37)	0.10	0.64	MRKMMRYFVRMAIPVLVIVCC AFNLHA–LEL	conserved hypothetical protein	*Bacteroides*	KF790724
MS26	2	246	77	Type I ss	0.4	0.444 (43-44)	1 TMH (30-47), possible N-terminal ss	Cyt	Cyt	FALSE-pattern not present	0.06	0.64	VGRNLDIAQAFEYVVCGPLYRK EALVMKNIALVSLLTASVAFA–QIP	conserved hypothetical protein	*Fibrobacter*	KF790725
MS27	1	700	136	Type I ss	0.391	0.574 (24-25)	1 TMH (5-24), possible N-terminal ss	SpI	Signal (1-26)	FALSE-pattern not present	0.90	0.69	MRKHILTALLVMLTVVTFGQQ WVA–IKS	conserved hypothetical protein	Genus not assigned	KF790726
MS28	1	606	43	Type I ss	0.740 (22-23)	0.550 (20-21)	0	SpI	Signal (1-22)	FALSE-pattern not present	0.07	0.03	MKSIIKTTLLALTICLPALA–SA–HDI	hypothetical protein	*Parabacteroides*	KF790727
MS29	1	805	90	Internal TMH	0.149	0.224	1 TMH (45-67), possible N-terminal ss	Cyt	Membrane (48-65)	FALSE-pattern not present	0.90	0.97	LTEIQSSTAAMPNSTPKPWPRSG SSTHTTTATPSRTCQGTRNMR*IK LSSIGIFTAFLMALSIPITQG*CES	hypothetical protein	*Parabacteroides*	KF790728
MS30	1	843	268	Type II ss	0.631 (23-24)	0.42	1 TMH (12-34), possible N-terminal ss	SpII (24-25)	Signal (1-35)	FALSE-pattern not present	0.66	0.39	MIRRMGKVVRILVNFLIGMLLL AS–CTQ	conserved hypothetical protein	Genus not assigned	KF790729

Clone ID [a]	Number of clones with identical insert	Insert length (bp)	Putative protein length (aa)	Predicted membrane-targeting signal [b]	SignalP 4.1 mean D-score and predicted ss cleavage site position (G-)	SignalP 4.1 mean D-score and predicted ss cleavage site position (G+)	TMHMM 2.0 number and predicted position of TMH	LipoP 1.0 location prediction (G-)	PRED-LIPO location and ss cleavage site location prediction (G+)	PILFIND 1.0 prediction (G+ and G-)	SecretomeP 2.0 score (G-)	SecretomeP 2.0 score (G+)	Predicted signal sequence/anchor and cleavage sites [c]	Putative protein description based on best BLASTP hit [d]	Taxonomic assignment based on best BLASTX hit [e]	Accession number
MS31	1	605	204	N-terminal TMH	0.152	0.387	1 TMH (47-69), possible N-terminal ss	Cyt	Membrane (51-69)	FALSE-pattern not present	0.53	0.95	MPPPVKRTVTKNKVNVKKASV PKTVVKPSVKVENKTKVEDNSL NKD*LVMRSILLVSYTLIIVFLVIGFV* DSL	conserved hypothetical protein	Genus not assigned	KF790730
MS32	2	215	44	Type I ss	0.677 (22-23)	0.495 (22-23)	0	SpI	Signal (1-25)	FALSE-pattern not present	0.31	0.37	MKKSIAILTTLLLLGGTICPFA–GEI	hypothetical protein	Genus not assigned	KF790731
MS33 [h]	1	867	37	N-terminal TMH	0.245	0.39	1 TMH (13-35), possible N-terminal ss	Cyt	Cyt	FALSE-pattern not present	0.07	0.97	MNMMISGKKKNC*IPVLRRMAAV LLSAGILLGMCFP*AA	hypothetical protein	*Clostridium*	KF790732
MS34	2	820	55	Type I ss	0.615 (35-36)	0.704 (35-36)	1 TMH (13-35), possible N-terminal ss	SpI	Signal (1-35)	FALSE-pattern not present	0.79	0.94	LFKMKKSHIKGRFMAMAAAAA MTCSFTYGTLPVSA–AET	hypothetical protein	Genus not assigned	KF790733
MS35	1	338	60	Type I ss	0.647 (19-20)	0.537 (20-21)	0	SpI	Membrane (5-26)	FALSE-pattern not present	0.58	0.08	MMRKIALTVLLAASQLTMS–A–QFFG	conserved hypothetical protein	Genus not assigned	KF790734
MS36	1	691	48	Type I ss	0.881 (19-20)	0.487 (10-11)	1 TMH (5-24), possible N-terminal ss	SpI	Signal (1-24)	FALSE-pattern not present	0.06	0.91	MKKLSFLMLV–LALAFPINA–SVT	hypothetical protein	Genus not assigned	KF790735
MS37	1	717	55	Type I ss	0.707 (26-27)	0.556 (26-27)	0	SpI	Signal (1-26)	FALSE-pattern not present	0.79	0.24	MKTKDFSKKLLVALTLMLALA MNAKA–QKS	hypothetical protein	*Prevotella*	KF790736
MS38	2	553	146	Type II ss	0.448 (21-22)	0.244	0	SpII (16-17)	Lipoprotein (1-16)	FALSE-pattern not present	0.50	0.17	MQKIYVLAGAALIMAACGGSE–KEY	membrane protein	*Prevotella*	KF790737
MS39 [f]	1	242	80	Internal TMH	0.134	0.194	1 TMH (49-71), possible N-terminal ss	Cyt	Membrane (49-65)	FALSE-pattern not present	0.60	0.89	**MPP**KALLKPATRTKQWLSDVV QSLIKRDNEAVKAQILNDQGYR KALFKASR*LALKLGMGTVAFAIS GYLGAAYL*AIE	hypothetical protein	Genus not assigned	KF790738
MS40 [h]	1	764	35	Type I ss	0.338	0.381	1 TMH (7-25),possible N-terminal ss	SpI	Signal (1-24)	FALSE-pattern not present	0.14	0.95	LMIKRP*AALILAGIMALLAVPAFG A*DEI	hypothetical protein	Genus not assigned	KF790739
MS41 [f]	1	299	99	Multiple TMH	0.155	0.171	2 TMH (10-32; 45-67), possible N-terminal ss	TMH	Membrane (45-65)	FALSE-pattern not present	0.08	0.93	**MPL**MILPMTMQE*MAGVTSDAMT NAFSFLFIAIVLK*AVNDNQKLSN KQ*IMVITILGTCIGMCKMVYIPLLL* LLL	conserved hypothetical protein	Genus not assigned	KF790740

Clone ID [a]	Number of clones with identical insert	Insert length (bp)	Putative protein length (aa)	Predicted membrane-targeting signal [b]	SignalP 4.1 mean D-score and predicted ss cleavage site position (G-)	SignalP 4.1 mean D-score and predicted ss cleavage site position (G+)	TMHMM 2.0 number and predicted position of TMH	LipoP 1.0 location prediction (G-)	PRED-LIPO location and ss cleavage site location prediction (G+)	PILFIND 1.0 prediction (G+ and G-)	SecretomeP 2.0 score (G-)	SecretomeP 2.0 score (G+)	Predicted signal sequence/anchor and cleavage sites [c]	Putative protein description based on best BLASTP hit [d]	Taxonomic assignment based on best BLASTX hit [e]	Accession number
MS42	1	866	140	Type II ss	0.652 (20-21)	0.608 (25-26)	0	SpII (17-18)	Signal (1-24)	FALSE-pattern not present	0.45	0.22	MKNRHIFAIAAVLMAIT–CKA	hypothetical protein	Genus not assigned	KF790741
MS43	1	887	136	Type I ss	0.841 (28-29)	0.505 (30-31)	1 TMH (13-31), possible N-terminal ss	SpI	Signal (1-28)	FALSE-pattern not present	0.05	0.59	MIYKLILRNMRKVVFGAVALFM AFVSNA–QV–LDP	conserved hypothetical protein	*Prevotella*	KF790742
MS44	1	864	43	Type I ss	0.832 (23-24)	0.617 (30-31)	1 TMH (7-24), possible N-terminal ss	SpI	Signal (1-28)	FALSE-pattern not present	0.27	0.92	MTMTKRFFMAVFAVCLVALST LA–DDKQAEA–TFS	hypothetical protein	*Bacteroides*	KF790743
MS45	1	388	133	Type I ss	0.173	0.430 (26-27)	0	Cyt	Cyt	FALSE-pattern not present	0.85	0.6	MHAPSSCPQSRSHRSCICPTPPC APA–CTP	hypothetical protein	Genus not assigned	KF790744
MS46 [h]	2	87	25	Type I ss	0.338	0.392	0	Cyt	Cyt	FALSE-pattern not present	0.11	0.03	MKKIVRTIAAMQAAVMILTSTA FAE	hypothetical protein	Genus not assigned	Not available [i]
MS47	1	485	52	Internal TMH	0.285	0.324	1 TMH (20-42), possible N-terminal ss	SpI	Signal (1-39)	FALSE-pattern not present	0.46	0.95	LRISFLRRVSDMNGRFRRF*ICFT AVAAALLILAGAGAGAYSY*GDD	hypothetical protein	*Butyrivibrio*	KF790745
MS48	1	618	64	N-terminal TMH	0.39	0.373	1 TMH (15-37), possible N-terminal ss	SpI	Signal (1-36)	FALSE-pattern not present	0.47	0.91	MIKLRFYRLLLTKN*YIFMKRVLF MLAGVLMFTAVSAQ*ENG	hypothetical protein	*Prevotella*	KF790746
MS49	1	816	228	Type I ss	0.734 (25-26)	0.479 (25-26)	1 TMH (5-27), possible N-terminal ss	SpI	Signal (1-25)	FALSE-pattern not present	0.47	0.81	MRTSHISFFLTSLLLHFLMPAFLS A–TTTC	Ser/Thr phosphatase family protein	*Bacteroides*	KF790747
MS50	1	259	41	Type I ss	0.864 (23-24)	0.708 (23-24)	1 TMH (5-23), possible N-terminal ss	SpI	Signal (1-23)	FALSE-pattern not present	0.39	0.93	MKKLLALVLALVLSLGVVSFAS A–EEP	hypothetical protein	Genus not assigned	KF790748
MS51	1	643	114	N-terminal TMH	0.353	0.257	1 TMH (21-38), possible N-terminal ss	SpI	Membrane (21-38)	FALSE-pattern not present	0.94	0.97	MNTTYEHTKRLAQQTITTLR*LTI ALLLMLVLGSTSVWG*QDY	hypothetical protein	Genus not assigned	KF790749
MS52	1	420	41	Type I ss	0.599 (21-22)	0.495 (21-22)	1 TMH (7-26), possible N-terminal ss	SpI	Signal (1-21)	FALSE-pattern not present	0.08	0.96	MTMKKIFTFVFCALLATAAWA–DVI	hypothetical protein	Genus not assigned	KF790750
MS53 [h]	1	818	37	Type I ss	0.169	0.274	1 TMH (7-29), possible N-terminal ss	Cyt	Lipoprotein (1-21)	FALSE-pattern not present	0.13	0.97	MKKRHLIIQLIFMLIVFIIPA–CSP	hypothetical protein	*Clostridium*	KF790751

Clone ID [a]	Number of clones with identical insert	Insert length (bp)	Putative protein length (aa)	Predicted membrane-targeting signal [b]	SignalP 4.1 mean D-score and predicted ss cleavage site position (G-)	SignalP 4.1 mean D-score and predicted ss cleavage site position (G+)	TMHMM 2.0 number and predicted position of TMH	LipoP 1.0 location prediction (G-)	PRED-LIPO location and ss cleavage site location prediction (G+)	PILFIND 1.0 prediction (G+ and G-)	SecretomeP 2.0 score (G-)	SecretomeP 2.0 score (G+)	Predicted signal sequence/anchor and cleavage sites [c]	Putative protein description based on best BLASTP hit [d]	Taxonomic assignment based on best BLASTX hit [e]	Accession number
MS54	1	888	427	Type I ss	0.807 (24-25)	0.634 (28-29)	0	SpI	Signal (1-28)	FALSE-pattern not present	0.33	0.17	LFMKKLIAAVTGAALAFMAFSP QA–EVLA–DSA	beta-lactamase	*Ruminococcus*	KF790752
MS55 [g]	1	315	15	Background (short ORF)	ND	ND	ND	ND	ND	ND	ND	ND	ND	hypothetical protein	Genus not assigned	Not available [i]
	90 [j]	641 [k]														

Abbreviations: aa, amino acid residues; G-, Gram-negative bacteria; G+, Gram-positive bacteria; D-score, discrimination score; ND, not determined. [a] The rumen metasecretome (MS) recombinant clones are numbered in the order of identification. [b] Predicted membrane-targeting signal types: ss, signal sequence; Type I ss, classical ss; Type II ss, lipoprotein ss; Type IV ss, pilin-like ss; TMH, transmembrane α-helix; Background, putative proteins without predicted membrane-targeting signal or shorter than 24 aa (the latter were not analysed for presence of signal sequence due to their short size). [c] Amino acid residues in italics correspond to the sequence of the predicted transmembrane α helix, – corresponds to predicted cleavage sites for SPaseI or SPaseII and these sites are represented for both G- and G+ bacteria if different; underlined sequence corresponds to Type IV ss. [d] Putative protein description, based on best BLASTP hit against NCBI nr protein database, is provided only for the hits with an E-value of <1e-05. Putative proteins with best BLASTP hit below this threshold are described as hypothetical. [e] Taxonomic assignments at the genus level, based on the best BLASTX hit of pilot library inserts encoding corresponding putative proteins, are provided only for the hits with an E-value of <1e-05 and a query coverage >30%. Genus was not assigned to inserts with best BLASTX hit below this threshold. Since 55 distinct ORFs were identified among 90 analysed inserts, this column does not reflect the taxonomic distribution of all analysed inserts (90). [f] Vector pDJ01 provided start codon, putative protein part encoded by the vector is represented in bold font. [g] Putative proteins with no ss, transmembrane helices or non-classical predicted protein secretion. [h] Putative proteins with D-score below SignalP 4.1 'sensitive' (SignalP-3.0 compliant) cut-off (0.42 for both G- and G+ bacteria) due to the short sequence length. Membrane-targeting signals identified in these putative proteins were assigned as Type I ss (rather than N-terminal membrane anchor), based on sequence similarity to the best BLASTP homologue that already carries the predicted ss in the aligned portion of the sequence. Membrane-targeting signals of short putative proteins with no known homologues were assigned as N-terminal membrane anchor, but are possible Type I ss. [i] Putative protein sequence too short for deposition. [j] Total number of inserts analysed. [k] Average insert size (bp).

Appendix 2

Table A2.1 Putative carbohydrate-active enzymes and associated modules identified in the metasecretome and the metagenome dataset.

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
AA1		laccase/p-diphenol:oxygen oxidoreductase/ferroxidase (EC 1.10.3.2); ferroxidase (EC 1.10.3.-); laccase-like multicopper oxidase (EC 1.10.3.-)	0	0.000	2	0.009	0.000
AA2	cd00692	manganese peroxidase (EC 1.11.1.13); versatile peroxidase (EC 1.11.1.16); lignin peroxidase (EC 1.11.1.14); peroxidase (EC 1.11.1.-)	1	0.008	0	0.000	0.000
AA3	COG2303	cellobiose dehydrogenase (EC 1.1.99.18); glucose 1-oxidase (EC 1.1.3.4); aryl alcohol oxidase (EC 1.1.3.7); alcohol oxidase (EC 1.1.3.13); pyranose oxidase (EC 1.1.3.10)	2	0.016	148	0.678	0.023
AA4		vanillyl-alcohol oxidase (EC 1.1.3.38)	0	0.000	12	0.055	0.000
AA5	self-built	oxidase with oxygen as acceptor (EC 1.1.3.-); galactose oxidase (EC 1.1.3.9)	1	0.008	23	0.105	0.076
AA6	TIGR01755	1,4-benzoquinone reductase (EC. 1.6.5.6)	50	0.398	169	0.774	0.514
AA7		glucooligosaccharide oxidase (EC 1.1.3.-)	0	0.000	10	0.046	0.000
AA8	self-built	iron reductase	3	0.024	79	0.362	0.066
AA9 (formerly GH61)	pfam03443	copper-dependent lytic polysaccharide monooxygenases (LPMOs); cleavage of cellulose chains with oxidation of various carbons (C-1, C-4 and C-6)	6	0.048	7	0.032	1.489
AA10 (formerly CBM33)	pfam03067	copper-dependent lytic polysaccharide monooxygenases (LPMOs); act on chitin and cellulose	4	0.032	1	0.005	6.947
Total AAs			67	0.5	451	2.1	
CBM1	pfam00734	Binding to cellulose and chitin	1	0.008	1	0.005	1.737
CBM2	pfam00553	Binding to cellulose, chitin and xylan.	1	0.008	49	0.225	0.035
CBM3	pfam00942	Binding to cellulose and chitin	1	0.008	17	0.078	0.102
CBM4	pfam02018	Binding to xylan, β-1,3-glucan, β-1,3-1,4-glucan, β-1,6-glucan and amorphous cellulose	43	0.342	109	0.499	0.685
CBM5	pfam02839	Binding to chitin	0	0.000	4	0.018	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
CBM6	pfam03422	Binding to amorphous cellulose, β-1,4-xylan, β-1,3-glucan, β-1,3-1,4-glucan, and β-1,4-glucan	41	0.326	169	0.774	0.421
CBM8	self-built	Binding to cellulose	0	0.000	4	0.018	0.000
CBM9	pfam06452	Binding to cellulose	9	0.072	22	0.101	0.711
CBM11	pfam03425	Binding to beta-1,4-glucan and beta-1,3-1,4-mixed linked glucans	0	0.000	9	0.041	0.000
CBM12	pfam02839	Binding to chitin	4	0.032	34	0.156	0.204
CBM13	pfam00652	Binding to xylan, GalNAc and mannose	21	0.167	67	0.307	0.544
CBM14	pfam01607	Binding to chitin	0	0.000	3	0.014	0.000
CBM15	pfam03426	Binding to xylan and xylooligosaccharides	2	0.016	2	0.009	1.737
CBM16	pfam02018	Binding to cellulose and glucomannan	86	0.684	22	0.101	6.789
CBM17	pfam03424	Binding to amorphous cellulose, cellooligosaccharides and derivatised cellulose	7	0.056	7	0.032	1.737
CBM19	pfam03427	Binding to chitin	0	0.000	1	0.005	0.000
CBM20	pfam00686	Binding to granular starch and cyclodextrins	56	0.446	43	0.197	2.262
CBM22	pfam02018	Binding to xylan, mixed β-1,3/β-1,4-glucans and thermostabilising effect	10	0.080	35	0.160	0.496
CBM23	pfam03425	Binding to mannan	0	0.000	5	0.023	0.000
CBM24	self-built	Binding to α-1,3-glucan (mutan)	1	0.008	4	0.018	0.434
CBM25	pfam03423	Binding to starch	4	0.032	4	0.018	1.737
CBM26	self-built	Binding to starch	49	0.390	38	0.174	2.240
CBM27	pfam09212	Binding to mannan	0	0.000	5	0.023	0.000
CBM28	pfam03424	Binding to non-crystalline cellulose, cellooligosaccharides, and beta-(1,3)(1,4)-glucans	0	0.000	1	0.005	0.000
CBM29	self-built	Binding to mannan and glucomannan	0	0.000	9	0.041	0.000
CBM30	pfam02927	Binding to cellulose	35	0.279	19	0.087	3.199
CBM31	pfam11606	Binding to beta-1,3-xylan	0	0.000	3	0.014	0.000
CBM32	pfam00754	Binding to galactose, lactose, polygalacturonic acid and LacNAc (β-D-galactosyl-1,4-β-D-N-acetylglucosamine)	66	0.525	82	0.376	1.398
CBM34	cd02857	Binding to granular starch	0	0.000	46	0.211	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
CBM35	pfam03422	Binding to xylan, decorated soluble mannans, mannooligosaccharides and β-galactan	55	0.438	100	0.458	0.955
CBM36	pfam03422	Binding to xylans and xylooligosaccharides (calcium-dependent)	0	0.000	12	0.055	0.000
CBM37	smart00060	Binding to xylan, chitin, microcrystalline, phosphoric-acid swollen cellulose and more heterogeneous substrates, such as alfalfa cell walls, banana stem and wheat straw	25	0.199	85	0.389	0.511
CBM38	self-built	Binding to inulin	16	0.127	4	0.018	6.947
CBM39	self-built	Binding to beta-1,3-glucan, lipopolysaccharide and lipoteichoic acid	0	0.000	4	0.018	0.000
CBM40	pfam02973	Binding to sialic acid	98	0.780	16	0.073	10.638
CBM41	pfam03714	Binding to alpha-glucans amylose, amylopectin, pullulan, and oligosaccharide fragments derived from these polysaccharides	0	0.000	10	0.046	0.000
CBM42	pfam05270	Binding to arabinofuranose (present in arabinoxylan)	0	0.000	2	0.009	0.000
CBM43	pfam07983	Binding to β-1,3-glucan	1	0.008	1	0.005	1.737
CBM44	pfam00801	Binding to cellulose and xyloglucan	2	0.016	12	0.055	0.289
CBM45	self-built	Binding to starch	0	0.000	2	0.009	0.000
CBM46	pfam03442	Binding to cellulose	0	0.000	11	0.050	0.000
CBM47	pfam00754	Binding to fucose	2	0.016	1	0.005	3.474
CBM48	pfam02922	Binding to glycogen	41	0.326	116	0.532	0.614
CBM49	pfam09478	Binding to crystalline cellulose	0	0.000	2	0.009	0.000
CBM50	pfam01476	Binding to chitopentaose	83	0.661	143	0.655	1.008
CBM51	pfam08305	Binding to galactose and to blood group A/B-antigens	2	0.016	12	0.055	0.289
CBM52	pfam10645	Binding to beta-1,3-glucan	0	0.000	4	0.018	0.000
CBM53	pfam03423	Binding to starch	0	0.000	3	0.014	0.000
CBM54	self-built	Binding to xylan, yeast cell wall glucan and chitin	1	0.008	16	0.073	0.109
CBM56	self-built	Binding to β-1,3-glucan	4	0.032	69	0.316	0.101
CBM57	pfam11721	Binding to Glc2-N-glycans	2	0.016	10	0.046	0.347
CBM58	self-built	Binding to maltoheptaose	6	0.048	15	0.069	0.695
CBM59	self-built	Binding to mannan, xylan, and cellulose	0	0.000	6	0.027	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
CBM60	self-built	Binding to xylan	8	0.064	5	0.023	2.779
CBM61	self-built	Binding to β-1,4-galactan	65	0.517	50	0.229	2.258
CBM62	self-built	Binding to the galactose moieties found on xyloglucan, arabinogalactan and galactomannan	1	0.008	12	0.055	0.145
CBM63		Binding to cellulose	0	0.000	7	0.032	0.000
CBM64	self-built	Binding to cellulose	1	0.008	3	0.014	0.579
CBM65		Binding to a range of β-glucans (significant preference for xyloglucan)	0	0.000	4	0.018	0.000
CBM66	self-built	Binding to terminal fructoside residue of fructans	5	0.040	29	0.133	0.299
CBM67	self-built	Binding to L-rhamnose	183	1.456	76	0.348	4.182
Total CBMs			1038	8.3	1656	7.6	
CE1	pfam00756	acetyl xylan esterase (EC 3.1.1.72); cinnamoyl esterase (EC 3.1.1.-); feruloyl esterase (EC 3.1.1.73); carboxylesterase (EC 3.1.1.1); S-formylglutathione hydrolase (EC 3.1.2.12)	356	2.833	487	2.232	1.270
CE2	cd01831	acetyl xylan esterase (EC 3.1.1.72).	15	0.119	132	0.605	0.197
CE3	cd01833	acetyl xylan esterase (EC 3.1.1.72).	117	0.931	67	0.307	3.033
CE4	pfam01522	acetyl xylan esterase (EC 3.1.1.72); chitin deacetylase (EC 3.5.1.41); chitooligosaccharide deacetylase (EC 3.5.1.-); peptidoglycan GlcNAc deacetylase (EC 3.5.1.-); peptidoglycan N-acetylmuramic acid deacetylase (EC 3.5.1.-).	46	0.366	153	0.701	0.522
CE5	pfam01083	acetyl xylan esterase (EC 3.1.1.72); cutinase (EC 3.1.1.74)	0	0.000	9	0.041	0.000
CE6	pfam03629	acetyl xylan esterase (EC 3.1.1.72).	24	0.191	91	0.417	0.458
CE7	pfam05448	acetyl xylan esterase (EC 3.1.1.72); cephalosporin-C deacetylase (EC 3.1.1.41).	248	1.974	160	0.733	2.692
CE8	pfam01095	pectin methylesterase (EC 3.1.1.11).	427	3.398	138	0.632	5.374
CE9	cd00854	N-acetylglucosamine 6-phosphate deacetylase (EC 3.5.1.25); N-acetylglucosamine 6-phosphate deacetylase (EC 3.5.1.80)	2	0.016	214	0.981	0.016

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
CE10	COG0657	arylesterase (EC 3.1.1.-); carboxyl esterase (EC 3.1.1.3); acetylcholinesterase (EC 3.1.1.7); cholinesterase (EC 3.1.1.8); sterol esterase (EC 3.1.1.13); brefeldin A esterase (EC 3.1.1.-)	195	1.552	501	2.296	0.676
CE11	pfam03331	UDP-3-0-acyl N-acetylglucosamine deacetylase (EC 3.5.1.-)	0	0.000	56	0.257	0.000
CE12	cd01821	pectin acetylesterase (EC 3.1.1.-); rhamnogalacturonan acetylesterase (EC 3.1.1.-); acetyl xylan esterase (EC 3.1.1.72)	49	0.390	137	0.628	0.621
CE13	pfam03283	pectin acetylesterase (EC 3.1.1.-)	1	0.008	4	0.018	0.434
CE14	pfam02585	N-acetyl-1-D-myo-inosityl-2-amino-2-deoxy-α-D-glucopyranoside deacetylase (EC 3.5.1.89); diacetylchitobiose deacetylase (EC 3.5.1.-); mycothiol S-conjugate amidase (EC 3.5.1.-)	10	0.080	23	0.105	0.755
CE15	self-built	4-O-methyl-glucuronoyl methylesterase (EC 3.1.1.-)	9	0.072	59	0.270	0.265
CE16	cd01846	acetylesterase (EC 3.1.1.6) active on various carbohydrate acetyl esters	0	0.000	4	0.018	0.000
Total CEs			1499	11.9	2235	10.2	
GH1	pfam00232	β-glucosidase (EC 3.2.1.21); β-galactosidase (EC 3.2.1.23); β-mannosidase (EC 3.2.1.25); β-glucuronidase (EC 3.2.1.31); β-D-fucosidase (EC 3.2.1.38); phlorizin hydrolase (EC 3.2.1.62); exo-β-1,4-glucanase (EC 3.2.1.74); 6-phospho-β-galactosidase (EC 3.2.1.85); 6-phospho-β-glucosidase (EC 3.2.1.86); strictosidine β-glucosidase (EC 3.2.1.105); lactase (EC 3.2.1.108); amygdalin β-glucosidase (EC 3.2.1.117); prunasin β-glucosidase (EC 3.2.1.118); raucaffricine β-glucosidase (EC 3.2.1.125); thioglucosidase (EC 3.2.1.147); β-primeverosidase (EC 3.2.1.149); isoflavonoid 7-O-β-apiosyl-β-glucosidase (EC 3.2.1.161); hydroxyisourate hydrolase (EC 3.-.-.-); β-glycosidase (EC 3.2.1.-)	2	0.016	111	0.509	0.031
GH2	COG3250	β-galactosidase (EC 3.2.1.23) ; β-mannosidase (EC 3.2.1.25); β-glucuronidase (EC 3.2.1.31); mannosylglycoprotein endo-β-mannosidase (EC 3.2.1.152); exo-β-glucosaminidase (EC 3.2.1.165)	828	6.590	1070	4.903	1.344

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH3	pfam00933	β-glucosidase (EC 3.2.1.21); xylan 1,4-β-xylosidase (EC 3.2.1.37); β-N-acetylhexosaminidase (EC 3.2.1.52); glucan 1,3-β-glucosidase (EC 3.2.1.58); glucan 1,4-β-glucosidase (EC 3.2.1.74); exo-1,3-1,4-glucanase (EC 3.2.1.-); αlpha-L-arabinofuranosidase (EC 3.2.1.55).	455	3.621	649	2.974	1.218
GH4	pfam02056	maltose-6-phosphate glucosidase (EC 3.2.1.122); α-glucosidase (EC 3.2.1.20); α-galactosidase (EC 3.2.1.22); 6-phospho-β-glucosidase (EC 3.2.1.86); α-glucuronidase (EC 3.2.1.139)	12	0.096	50	0.229	0.417
GH5	pfam00150	chitosanase (EC 3.2.1.132); β-mannosidase (EC 3.2.1.25); cellulase (EC 3.2.1.4); glucan 1,3-β-glucosidase (EC 3.2.1.58); licheninase (EC 3.2.1.73); glucan endo-1,6-β-glucosidase (EC 3.2.1.75); mannan endo-β-1,4-mannosidase (EC 3.2.1.78); endo-β-1,4-xylanase (EC 3.2.1.8); cellulose β-1,4-cellobiosidase (EC 3.2.1.91); β-1,3-mannanase (EC 3.2.1.-); xyloglucan-specific endo-β-1,4-glucanase (EC 3.2.1.151); mannan transglycosylase (EC 2.4.1.-); endo-β-1,6-galactanase (EC 3.2.1.164)	268	2.133	395	1.810	1.178
GH6	pfam01341	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91)	0	0.000	3	0.014	0.000
GH8	pfam01270	chitosanase (EC 3.2.1.132); cellulase (EC 3.2.1.4); licheninase (EC 3.2.1.73); endo-1,4-β-xylanase (EC 3.2.1.8); reducing-end-xylose releasing exo-oligoxylanase (EC 3.2.1.156)	7	0.056	98	0.449	0.124
GH9	pfam00759	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91); β-glucosidase (EC 3.2.1.21); exo-β-glucosaminidase (EC 3.2.1.165)	109	0.867	242	1.109	0.782
GH10	pfam00331	endo-1,4-β-xylanase (EC 3.2.1.8); endo-1,3-β-xylanase (EC 3.2.1.32)	77	0.613	291	1.333	0.460
GH11	pfam00457	xylanase (EC 3.2.1.8)	5	0.040	20	0.092	0.434
GH12	pfam04616	alpha-1,3-L-neoagarooligosaccharide hydrolase (EC 3.2.1.-); alpha-1,3-L-neoagarobiase / neoagarobiose hydrolase (EC 3.2.1.-)	0	0.000	1	0.005	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH13	pfam00128	α-amylase (EC 3.2.1.1); pullulanase (EC 3.2.1.41); cyclomaltodextrin glucanotransferase (EC 2.4.1.19); cyclomaltodextrinase (EC 3.2.1.54); trehalose-6-phosphate hydrolase (EC 3.2.1.93); oligo-α-glucosidase (EC 3.2.1.10); maltogenic amylase (EC 3.2.1.133); neopullulanase (EC 3.2.1.135); α-glucosidase (EC 3.2.1.20); maltotetraose-forming α-amylase (EC 3.2.1.60); isoamylase (EC 3.2.1.68); glucodextranase (EC 3.2.1.70); maltohexaose-forming α-amylase (EC 3.2.1.98); maltotriose-forming α-amylase (EC 3.2.1.116); branching enzyme (EC 2.4.1.18); trehalose synthase (EC 5.4.99.16); 4-α-glucanotransferase (EC 2.4.1.25); maltopentaose-forming α-amylase (EC 3.2.1.-); amylosucrase (EC 2.4.1.4); sucrose phosphorylase (EC 2.4.1.7); malto-oligosyltrehalose trehalohydrolase (EC 3.2.1.141); isomaltulose synthase (EC 5.4.99.11); amino acid transporter	172	1.369	742	3.400	0.403
GH14	pfam01373	beta-amylase (EC 3.2.1.2)	0	0.000	3	0.014	0.000
GH15	pfam00723	glucoamylase (EC 3.2.1.3); glucodextranase (EC 3.2.1.70); alpha,alpha-trehalase (EC 3.2.1.28)	0	0.000	5	0.023	0.000
GH16	cd00413	xyloglucan:xyloglucosyltransferase (EC 2.4.1.207); keratan-sulfate endo-1,4-β-galactosidase (EC 3.2.1.103); endo-1,3-β-glucanase (EC 3.2.1.39); endo-1,3(4)-β-glucanase (EC 3.2.1.6); licheninase (EC 3.2.1.73); β-agarase (EC 3.2.1.81); κ-carrageenase (EC 3.2.1.83); xyloglucanase (EC 3.2.1.151)	83	0.661	84	0.385	1.716
GH17	pfam00332	glucan endo-1,3-beta-glucosidase (EC 3.2.1.39); glucan 1,3-beta-glucosidase (EC 3.2.1.58); licheninase (EC 3.2.1.73); beta-1,3-glucanosyltransglycosylase (EC 2.4.1.-)	0	0.000	2	0.009	0.000
GH18	pfam00704	chitinase (EC 3.2.1.14); endo-β-N-acetylglucosaminidase (EC 3.2.1.96); xylanase inhibitor; concanavalin B; narbonin	17	0.135	80	0.367	0.369
GH19	cd00325	chitinase (EC 3.2.1.14)	0	0.000	2	0.009	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH20	pfam00728	β-hexosaminidase (EC 3.2.1.52); lacto-N-biosidase (EC 3.2.1.140); β-1,6-N-acetylglucosaminidase) (EC 3.2.1.-); β-6-SO3-N-acetylglucosaminidase (EC 3.2.1.-)	2	0.016	103	0.472	0.034
GH23	cd00254	lysozyme type G (EC 3.2.1.17); peptidoglycan lyase (EC 4.2.2.-) also known in the literature as peptidoglycan lytic transglycosylase	14	0.111	107	0.490	0.227
GH24	cd00737	lysozyme (EC 3.2.1.17)	1	0.008	26	0.119	0.067
GH25	pfam01183	lysozyme (EC 3.2.1.17)	423	3.366	256	1.173	2.870
GH26	pfam02156	β-mannanase (EC 3.2.1.78); β-1,3-xylanase (EC 3.2.1.32)	160	1.273	121	0.554	2.297
GH27	pfam02065	α-galactosidase (EC 3.2.1.22); α-N-acetylgalactosaminidase (EC 3.2.1.49); isomalto-dextranase (EC 3.2.1.94); β-L-arabinopyranosidase (EC 3.2.1.88)	18	0.143	89	0.408	0.351
GH28	pfam00295	polygalacturonase (EC 3.2.1.15); exo-polygalacturonase (EC 3.2.1.67); exo-polygalacturonosidase (EC 3.2.1.82); rhamnogalacturonase (EC 3.2.1.-); endo-xylogalacturonan hydrolase (EC 3.2.1.-); rhamnogalacturonan α-L-rhamnopyranohydrolase (EC 3.2.1.40)	129	1.027	206	0.944	1.088
GH29	pfam01120	α-L-fucosidase (EC 3.2.1.51)	26	0.207	147	0.674	0.307
GH30	COG5520	glucosylceramidase (EC 3.2.1.45); β-1,6-glucanase (EC 3.2.1.75); β-xylosidase (EC 3.2.1.37); β-fucosidase (EC 3.2.1.38); β-glucosidase (3.2.1.21); endo-β-1,6-galactanase (EC:3.2.1.164)	228	1.815	93	0.426	4.258
GH31	pfam01055	α-glucosidase (EC 3.2.1.20); α-1,3-glucosidase (EC 3.2.1.84); sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); α-xylosidase (EC 3.2.1.-); α-glucan lyase (EC 4.2.2.13); isomaltosyltransferase (EC 2.4.1.-)	230	1.830	496	2.273	0.805

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH32	pfam00251	invertase (EC 3.2.1.26); endo-inulinase (EC 3.2.1.7); β-2,6-fructan 6-levanbiohydrolase (EC 3.2.1.64); endo-levanase (EC 3.2.1.65); exo-inulinase (EC 3.2.1.80); fructan β-(2,1)-fructosidase/1-exohydrolase (EC 3.2.1.153); fructan β-(2,6)-fructosidase/6-exohydrolase (EC 3.2.1.154); sucrose:sucrose 1-fructosyltransferase (EC 2.4.1.99); fructan:fructan 1-fructosyltransferase (EC 2.4.1.100); sucrose:fructan 6-fructosyltransferase (EC 2.4.1.10); fructan:fructan 6G-fructosyltransferase (EC 2.4.1.243); levan fructosyltransferase (EC 2.4.1.-)	98	0.780	198	0.907	0.860
GH33	cd00260	sialidase or neuraminidase (EC 3.2.1.18); trans-sialidase (EC 2.4.1.-); 2-keto-3-deoxynononic acid sialidase (EC 3.2.1.-)	25	0.199	33	0.151	1.316
GH34	pfam00064	sialidase or neuraminidase (EC 3.2.1.18)	6	0.048	1	0.005	10.421
GH35	pfam01301	β-galactosidase (EC 3.2.1.23); exo-β-glucosaminidase (EC 3.2.1.165)	2	0.016	86	0.394	0.040
GH36	COG3345	α-galactosidase (EC 3.2.1.22); α-N-acetylgalactosaminidase (EC 3.2.1.49); stachyose synthase (EC 2.4.1.67); raffinose synthase (EC 2.4.1.82)	194	1.544	348	1.595	0.968
GH37	pfam01204	alpha,alpha-trehalase (EC 3.2.1.28)	0	0.000	11	0.050	0.000
GH38	pfam01074	α-mannosidase (EC 3.2.1.24) ; mannosyl-oligosaccharide α-1,3-1,6-mannosidase (EC 3.2.1.114); mannosyl-oligosaccharide α-1,3-mannosidase (EC 3.2.1.-)	2	0.016	22	0.101	0.158
GH39	self-built	α-L-iduronidase (EC 3.2.1.76); β-xylosidase (EC 3.2.1.37)	2	0.016	74	0.339	0.047
GH42	pfam02449	β-galactosidase (EC 3.2.1.23)	2	0.016	47	0.215	0.074
GH43	pfam04616	β-xylosidase (EC 3.2.1.37); β-1,3-xylosidase (EC 3.2.1.-); α-L-arabinofuranosidase (EC 3.2.1.55); arabinanase (EC 3.2.1.99); xylanase (EC 3.2.1.8); galactan 1,3-β-galactosidase (EC 3.2.1.145)	1032	8.213	772	3.538	2.322
GH44	self-built	endoglucanase (EC 3.2.1.4); xyloglucanase (EC 3.2.1.151)	12	0.096	9	0.041	2.316
GH45	pfam02015	endoglucanase (EC 3.2.1.4)	0	0.000	11	0.050	0.000
GH46	cd00978	chitosanase (EC 3.2.1.132)	0	0.000	1	0.005	0.000
GH47	pfam01532	alpha-mannosidase (EC 3.2.1.113)	0	0.000	5	0.023	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH48	pfam02011	endoglucanase (EC 3.2.1.4); chitinase (EC 3.2.1.14); cellobiohydrolase (EC 3.2.1.91); endo-processive cellulases (EC 3.2.1.-); [reducing end] cellobiohydrolase (3.2.1.-)	0	0.000	4	0.018	0.000
GH49	pfam03718	dextranase (EC 3.2.1.11); isopullulanase (EC 3.2.1.57); dextran 1,6-alpha-isomaltotriosidase (EC 3.2.1.95)	0	0.000	1	0.005	0.000
GH50	self-built	beta-agarase (EC 3.2.1.81)	0	0.000	24	0.110	0.000
GH51	COG3534	α-L-arabinofuranosidase (EC 3.2.1.55); endoglucanase (EC 3.2.1.4)	69	0.549	390	1.787	0.307
GH52	pfam03512	β-xylosidase (EC 3.2.1.37)	1	0.008	4	0.018	0.434
GH53	pfam07745	endo-β-1,4-galactanase (EC 3.2.1.89)	291	2.316	126	0.577	4.011
GH54	pfam09206	α-L-arabinofuranosidase (EC 3.2.1.55); β-xylosidase (EC 3.2.1.37)	1	0.008	9	0.041	0.193
GH55	self-built	exo-β-1,3-glucanase (EC 3.2.1.58); endo-β-1,3-glucanase (EC 3.2.1.39)	124	0.987	33	0.151	6.526
GH56	pfam01630	hyaluronidase (EC 3.2.1.35); chondroitin hydrolase (EC 3.2.1.-)	0	0.000	3	0.014	0.000
GH57	pfam03065	alpha-amylase (EC 3.2.1.1); 4-alpha-glucanotransferase (EC 2.4.1.25); alpha-galactosidase (EC 3.2.1.22); amylopullulanase (EC 3.2.1.41); branching enzyme (EC 2.4.1.18)	0	0.000	54	0.247	0.000
GH58	pfam12217	endo-N-acetylneuraminidase or endo-sialidase (EC 3.2.1.129)	2	0.016	2	0.009	1.737
GH59	pfam02057	galactocerebrosidase (EC 3.2.1.46)	13	0.103	16	0.073	1.411
GH62	pfam03664	alpha-L-arabinofuranosidase (EC 3.2.1.55)	0	0.000	2	0.009	0.000
GH63	PRK10137	processing alpha-glucosidase (EC 3.2.1.106); alpha-1,3-glucosidase (EC 3.2.1.84); alpha-glucosidase (EC 3.2.1.20)	1	0.008	10	0.046	0.174
GH64	self-built	beta-1,3-glucanase (EC 3.2.1.39)	0	0.000	7	0.032	0.000
GH65	pfam03632	α,α-trehalase (EC 3.2.1.28); maltose phosphorylase (EC 2.4.1.8); trehalose phosphorylase (EC 2.4.1.64); kojibiose phosphorylase (EC 2.4.1.230); trehalose-6-phosphate phosphorylase (EC 2.4.1.-);	2	0.016	20	0.092	0.174
GH66	self-built	cycloisomaltooligosaccharide glucanotransferase (EC 2.4.1.-); dextranase (EC 3.2.1.11)	1	0.008	19	0.087	0.091
GH67	COG3661	α-glucuronidase (EC 3.2.1.139); xylan α-1,2-glucuronidase (EC 3.2.1.131)	18	0.143	134	0.614	0.233

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH68	pfam02435	levansucrase (EC 2.4.1.10); beta-fructofuranosidase (EC 3.2.1.26); inulosucrase (EC 2.4.1.9)	0	0.000	1	0.005	0.000
GH70	pfam02324	dextransucrase (EC 2.4.1.5); alternansucrase (EC 2.4.1.140); reuteransucrase (EC 2.4.1.-); alpha-4,6-glucanotransferase (EC 2.4.1.-)	0	0.000	5	0.023	0.000
GH71	pfam03659	alpha-1,3-glucanase (EC 3.2.1.59)	0	0.000	2	0.009	0.000
GH72	pfam03198	beta-1,3-glucanosyltransglycosylase (EC 2.4.1.-)	0	0.000	1	0.005	0.000
GH73	pfam01832	peptidoglycan hydrolase with endo-β-N-acetylglucosaminidase specificity (EC 3.2.1.-)	3	0.024	67	0.307	0.078
GH74	smart00602	endoglucanase (EC 3.2.1.4); oligoxyloglucan reducing end-specific cellobiohydrolase (EC 3.2.1.150); xyloglucanase (EC 3.2.1.151)	68	0.541	46	0.211	2.567
GH75	pfam07335	chitosanase (EC 3.2.1.132)	2	0.016	2	0.009	1.737
GH76	pfam03663	α-1,6-mannanase (EC 3.2.1.101)	6	0.048	25	0.115	0.417
GH77	pfam02446	amylomaltase or 4-α-glucanotransferase (EC 2.4.1.25)	100	0.796	373	1.709	0.466
GH78	pfam05592	α-L-rhamnosidase (EC 3.2.1.40)	3	0.024	190	0.871	0.027
GH79	self-built	β-glucuronidase (EC 3.2.1.31); β-4-O-methyl-glucuronidase (EC 3.2.1.-); heparanase (EC 3.2.1.-); baicalin β-glucuronidase (EC 3.2.1.167)	1	0.008	2	0.009	0.868
GH80	cd00978	chitosanase (EC 3.2.1.132)	1	0.008	3	0.014	0.579
GH81	pfam03639	endo-beta-1,3-glucanase (EC 3.2.1.39)	0	0.000	5	0.023	0.000
GH82	COG5434	ι-carrageenase (EC 3.2.1.157)	5	0.040	19	0.087	0.457
GH83	pfam00423	neuraminidase (EC 3.2.1.18)	1	0.008	4	0.018	0.434
GH84	pfam07555	N-acetyl beta-glucosaminidase (EC 3.2.1.52); hyaluronidase (EC 3.2.1.35)	0	0.000	14	0.064	0.000
GH85	pfam03644	endo-beta-N-acetylglucosaminidase (EC 3.2.1.96)	0	0.000	6	0.027	0.000
GH86	self-built	β-agarase (EC 3.2.1.81)	1	0.008	4	0.018	0.434
GH87	self-built	mycodextranase (EC 3.2.1.61); α-1,3-glucanase (EC 3.2.1.59)	5	0.040	20	0.092	0.434
GH88	pfam07470	δ-4,5 unsaturated beta-glucuronyl hydrolase (EC 3.2.1.-)	2	0.016	33	0.151	0.105
GH89	pfam05089	α-N-acetylglucosaminidase (EC 3.2.1.50)	2	0.016	37	0.170	0.094
GH90	pfam09251	endorhamnosidase (EC 3.2.1.-)	0	0.000	5	0.023	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH91	self-built	inulin lyase [DFA-I-forming] (EC 4.2.2.17); inulin lyase [DFA-III-forming] (EC 4.2.2.18); dIota;-fructofuranose 1,2':2,3' dianhydride hydrolase [DFA-IIIase] (EC 3.2.1.-)	0	0.000	10	0.046	0.000
GH92	pfam07971	mannosyl-oligosaccharide α-1,2-mannosidase (EC 3.2.1.113); mannosyl-oligosaccharide α-1,3-mannosidase (EC 3.2.1.-); mannosyl-oligosaccharide α-1,6-mannosidase (EC 3.2.1.-); α-mannosidase (EC 3.2.1.24); α-1,2-mannosidase (EC 3.2.1.-); α-1,3-mannosidase (EC 3.2.1.-); α-1,4-mannosidase (EC 3.2.1.-)	387	3.080	114	0.522	5.896
GH93	self-built	exo-α-L-1,5-arabinanase (EC 3.2.1.-)	3	0.024	4	0.018	1.303
GH94	COG3459	cellobiose phosphorylase (EC 2.4.1.20); cellodextrin phosphorylase (EC 2.4.1.49); chitobiose phosphorylase (EC 2.4.1.-); cyclic β-1,2-glucan synthase (EC 2.4.1.-)	66	0.525	460	2.108	0.249
GH95	self-built	α-1,2-L-fucosidase (EC 3.2.1.63); α-L-fucosidase (EC 3.2.1.51)	174	1.385	326	1.494	0.927
GH96	self-built	α-agarase (EC 3.2.1.158)	1	0.008	5	0.023	0.347
GH97	pfam10566	α-glucosidase (EC 3.2.1.20); α-galactosidase (EC 3.2.1.22)	574	4.568	381	1.746	2.617
GH98	pfam08306	endo-β-galactosidase (EC 3.2.1.-)	6	0.048	20	0.092	0.521
GH99	self-built	glycoprotein endo-α-1,2-mannosidase (EC 3.2.1.130)	2	0.016	19	0.087	0.183
GH100	pfam04853	alkaline and neutral invertase (EC 3.2.1.26)	0	0.000	3	0.014	0.000
GH101	self-built	endo-alpha-N-acetylgalactosaminidase (EC 3.2.1.97)	0	0.000	2	0.009	0.000
GH102	pfam03562	peptidoglycan lytic transglycosylase (EC 3.2.1.-)	0	0.000	6	0.027	0.000
GH103	TIGR02283	peptidoglycan lytic transglycosylase (EC 3.2.1.-)	0	0.000	5	0.023	0.000
GH104	cd00736	peptidoglycan lytic transglycosylase (EC 3.2.1.-)	0	0.000	2	0.009	0.000
GH105	pfam07470	unsaturated rhamnogalacturonyl hydrolase (EC 3.2.1.-)	208	1.655	188	0.861	1.922
GH106	self-built	α-L-rhamnosidase (EC 3.2.1.40)	118	0.939	184	0.843	1.114
GH107	self-built	sulfated fucan endo-1,4-fucanase (EC 3.2.1.-)	0	0.000	4	0.018	0.000
GH108	pfam05838	N-acetylmuramidase (EC 3.2.1.17)	0	0.000	1	0.005	0.000
GH109	pfam01408	α-N-acetylgalactosaminidase (EC 3.2.1.49)	8	0.064	147	0.674	0.095
GH110	self-built	α-galactosidase (EC 3.2.1.22); α-1,3-galactosidase (EC 3.2.1.-)	4	0.032	14	0.064	0.496

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH112	pfam09508	lacto-N-biose phosphorylase or galacto-N-biose phosphorylase (EC 2.4.1.211); D-galactosyl-1,4-L-rhamnose phosphorylase (EC 2.4.1.-)	1	0.008	37	0.170	0.047
GH113	self-built	β-mannanase (EC 3.2.1.78)	1	0.008	15	0.069	0.116
GH114	pfam03537	endo-α-1,4-polygalactosaminidase (EC 3.2.1.109)	2	0.016	18	0.082	0.193
GH115	self-built	xylan α-1,2-glucuronidase (3.2.1.131); α-(4-O-methyl)-glucuronidase (3.2.1.-)	53	0.422	278	1.274	0.331
GH116	pfam04685	acid β-glucosidase (EC 3.2.1.45); β-glucosidase (EC 3.2.1.21); β-xylosidase (EC 3.2.1.37)	2	0.016	19	0.087	0.183
GH117	pfam04616	α-1,3-L-neoagarooligosaccharide hydrolase (EC 3.2.1.-); α-1,3-L-neoagarobiase / neoagarobiose hydrolase (EC 3.2.1.-)	18	0.143	21	0.096	1.489
GH118	self-built	β-agarase (EC 3.2.1.81)	6	0.048	5	0.023	2.084
GH119	self-built	α-amylase (EC 3.2.1.1)	12	0.096	22	0.101	0.947
GH120	pfam07602	beta-xylosidase (EC 3.2.1.37)	0	0.000	21	0.096	0.000
GH121	self-built	β-L-arabinobiosidase (EC 3.2.1.-)	44	0.350	24	0.110	3.184
GH122	COG4697	alpha-glucosidase (EC 3.2.1.20)	0	0.000	1	0.005	0.000
GH123	self-built	glycosphingolipid beta-N-acetylgalactosaminidase (EC 3.2.1.-)	0	0.000	13	0.060	0.000
GH124	self-built	endoglucanase (EC 3.2.1.4)	132	1.051	16	0.073	14.329
GH125	pfam06824	exo-α-1,6-mannosidase (EC 3.2.1.-)	15	0.119	28	0.128	0.930
GH126	pfam02156	beta-mannanase (EC 3.2.1.78); beta-1,3-xylanase (EC 3.2.1.32)	0	0.000	5	0.023	0.000
GH127	pfam07944	β-L-arabinofuranosidase (EC 3.2.1.-)	340	2.706	208	0.953	2.839
GH128	pfam11790	β-1,3-glucanase (EC 3.2.1.39)	1	0.008	5	0.023	0.347
GH129	self-built	α-N-acetylgalactosaminidase (EC 3.2.1.49)	3	0.024	19	0.087	0.274
GH130	pfam04041	1-β-D-mannopyranosyl-4-D-glucopyranose:phosphate α-D-mannosyltransferase (EC 2.4.1.281); β-1,4-mannooligosaccharide phosphorylase (EC 2.4.1.-)	93	0.740	121	0.554	1.335

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GH131	pfam01055	alpha-glucosidase (EC 3.2.1.20); alpha-1,3-glucosidase (EC 3.2.1.84); sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); alpha-xylosidase (EC 3.2.1.-); alpha-glucan lyase (EC 4.2.2.13); isomaltosyltransferase (EC 2.4.1.-)	0	0.000	1	0.005	0.000
GH132	self-built	activity on β-1,3-glucan (curdlan); activity on laminarioligosaccharides; transglycosylation activity	1	0.008	2	0.009	0.868
Total GHs			7639	60.8	11606	53.2	
GT1	COG1819	UDP-glucuronosyltransferase (EC 2.4.1.17); 2-hydroxyacylsphingosine 1-beta-galactosyltransferase (EC 2.4.1.45); N-acylsphingosine galactosyltransferase (EC 2.4.1.47); flavonol 3-O-glucosyltransferase (EC 2.4.1.91); indole-3-acetate beta-glucosyltransferase (EC 2.4.1.121); sterol glucosyltransferase (EC 2.4.1.173); ecdysteroid UDP-glucosyltransferase (EC 2.4.1.-); zeaxanthin glucosyltransferase (EC 2.4.1.-); zeatin O-beta-glucosyltransferase (EC 2.4.1.203); zeatin O-beta-xylosyltransferase (EC 2.4.2.40); limonoid glucosyltransferase (EC 2.4.1.210); sinapate 1-glucosyltransferase (EC 2.4.1.120); anthocyanin 3-O-galactosyltransferase (EC 2.4.1.-); anthocyanin 5-O-glucosyltransferase (EC 2.4.1.-); anthocyanidin 3-O-glucosyltransferase (EC 2.4.1.115); dTDP-beta-2-deoxy-L-fucose: alpha-L-2-deoxyfucosyltransferase (EC 2.4.1.-); UDP-beta-L-rhamnose: alpha-L-rhamnosyltransferase (EC 2.4.1.-); UDP-glucose: 4-hydroxybenzoate 4-O-beta-glucosyltransferase (EC 2.4.1.194); flavonol L-rhamnosyltransferase (EC 2.4.1.159); salicylic acid beta-glucosyltransferase (EC 2.4.1.-)	0	0.000	21	0.096	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT2	pfam00535	cellulose synthase (EC 2.4.1.12); chitin synthase (EC 2.4.1.16); dolichyl-phosphate β-D-mannosyltransferase (EC 2.4.1.83); dolichyl-phosphate β-glucosyltransferase (EC 2.4.1.117); N-acetylglucosaminyltransferase (EC 2.4.1.-); N-acetylgalactosaminyltransferase (EC 2.4.1.-); hyaluronan synthase (EC 2.4.1.212); chitin oligosaccharide synthase (EC 2.4.1.-); β-1,3-glucan synthase (EC 2.4.1.34); β-1,4-mannan synthase (EC 2.4.1.-); β-mannosylphosphodecaprenol-mannooligosaccharide α-1,6-mannosyltransferase (EC 2.4.1.199); α-1,3-L-rhamnosyltransferase (EC 2.4.1.-)	72	0.573	1461	6.695	0.086
GT3	pfam05693	glycogen synthase (EC 2.4.1.11)	1	0.008	82	0.376	0.021
GT4	pfam00534	sucrose synthase (EC 2.4.1.13); sucrose-phosphate synthase (EC 2.4.1.14); α-glucosyltransferase (EC 2.4.1.52); lipopolysaccharide N-acetylglucosaminyltransferase (EC 2.4.1.56); GDP-Man α-mannosyltransferase (EC 2.4.1.-); 1,2-diacylglycerol 3-glucosyltransferase (EC 2.4.1.157); diglucosyl diacylglycerol synthase (EC 2.4.1.208); digalactosyldiacylglycerol synthase (EC 2.4.1.141); trehalose phosphorylase (EC 2.4.1.231); phosphatidylinositol α-mannosyltransferase (EC 2.4.1.57); UDP-Gal α-galactosyltransferase (EC 2.4.1.-); UDP-Xyl α-xylosyltransferase (EC 2.4.2.-)	99	0.788	714	3.272	0.241
GT5	cd03791	UDP-Glc: glycogen glucosyltransferase (EC 2.4.1.11); ADP-Glc: starch glucosyltransferase (EC 2.4.1.21); NDP-Glc: starch glucosyltransferase (EC 2.4.1.242); UDP-Glc: α-1,3-glucan synthase (EC 2.4.1.183) UDP-Glc: α-1,4-glucan synthase (EC 2.4.1.-)	8	0.064	239	1.095	0.058
GT6	pfam03414	α-1,3-galactosyltransferase (EC 2.4.1.87); α-1,3 N-acetylgalactosaminyltransferase (EC 2.4.1.40); α-galactosyltransferase (EC 2.4.1.37); globoside α-N-acetylgalactosaminyltransferase (EC 2.4.1.88).	2	0.016	12	0.055	0.289

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT7	pfam02709	lactose synthase (EC 2.4.1.22); beta-N-acetylglucosaminyl-glycopeptide beta-1,4-galactosyltransferase (EC 2.4.1.38); N-acetyllactosamine synthase (EC 2.4.1.90); beta-1,4-N-acetylglucosaminyltransferase (EC 2.4.1.-); xylosylprotein beta-4-galactosyltransferase (EC 2.4.1.133)	0	0.000	9	0.041	0.000
GT8	pfam01501	lipopolysaccharide alpha-1,3-galactosyltransferase (EC 2.4.1.44); UDP-Glc: (glucosyl)lipopolysaccharide alpha-1,2-glucosyltransferase (EC 2.4.1.-); lipopolysaccharide glucosyltransferase 1 (EC 2.4.1.58); glycogenin glucosyltransferase (EC 2.4.1.186); inositol 1-alpha-galactosyltransferase (galactinol synthase) (EC 2.4.1.123); homogalacturonan alpha-1,4-galacturonosyltransferase (EC 2.4.1.43); UDP-GlcA: xylan alpha-glucuronyltransferase (EC 2.4.1.-)	0	0.000	107	0.490	0.000
GT9	pfam01075	lipopolysaccharide N-acetylglucosaminyltransferase (EC 2.4.1.56); heptosyltransferase (EC 2.4.-.-)	0	0.000	66	0.302	0.000
GT10	pfam00852	galactoside alpha-1,3/1,4-L-fucosyltransferase (EC 2.4.1.65); galactoside alpha-1,3-L-fucosyltransferase (EC 2.4.1.152); glycoprotein alpha-1,3-L-fucosyltransferase (EC 2.4.1.214)	0	0.000	20	0.092	0.000
GT11	pfam01531	GDP-L-Fuc: galactoside α-1,2-L-fucosyltransferase (EC 2.4.1.69); GDP-L-Fuc: β-LacNac α-1,3-L-fucosyltransferase (EC 2.4.1.-)	1	0.008	97	0.444	0.018
GT12	pfam00535	[N-acetylneuraminyl]-galactosylglucosylceramide N-acetylgalactosaminyltransferase (EC 2.4.1.92)	0	0.000	9	0.041	0.000
GT13	pfam03071	alpha-1,3-mannosyl-glycoprotein beta-1,2-N-acetylglucosaminyltransferase (EC 2.4.1.101)	0	0.000	22	0.101	0.000
GT14	pfam02485	β-1,3-galactosyl-O-glycosyl-glycoprotein β-1,6-N-acetylglucosaminyltransferase (EC 2.4.1.102); N-acetyllactosaminide β-1,6-N-acetylglucosaminyltransferase (EC 2.4.1.150); protein O-β-xylosyltransferase (EC 2.4.2.26)	2	0.016	37	0.170	0.094
GT15	pfam01793	glycolipid 2-α-mannosyltransferase (EC 2.4.1.131); GDP-Man: α-1,2-mannosyltransferase (EC 2.4.1.-)	3	0.024	5	0.023	1.042

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT16	pfam05060	alpha-1,6-mannosyl-glycoprotein beta-1,2-N-acetylglucosaminyltransferase (EC 2.4.1.143)	0	0.000	2	0.009	0.000
GT17	pfam04724	β-1,4-mannosyl-glycoprotein β-1,4-N-acetylglucosaminyltransferase (EC 2.4.1.144)	3	0.024	15	0.069	0.347
GT18	self-built	α-1,3(6)-mannosylglycoprotein β-1,6-N-acetyl-glucosaminyltransferase (EC 2.4.1.155)	1	0.008	0	0.000	
GT19	pfam02684	lipid-A-disaccharide synthase (EC 2.4.1.182)	3	0.024	72	0.330	0.072
GT20	pfam00982	alpha,alpha-trehalose-phosphate synthase [UDP-forming] (EC 2.4.1.15)	0	0.000	11	0.050	0.000
GT21	cd02520	UDP-Glc: ceramide β-glucosyltransferase (EC 2.4.1.80)	1	0.008	14	0.064	0.124
GT22	pfam03901	Dol-P-Man α-mannosyltransferase (EC 2.4.1.-)	2	0.016	7	0.032	0.496
GT23	self-built	N-acetyl-beta-D-glucosaminide alpha-1,6-L-fucosyltransferase (EC 2.4.1.68)	0	0.000	13	0.060	0.000
GT24	cd06432	UDP-Glc: glycoprotein α-glucosyltransferase (EC 2.4.1.-)	1	0.008	1	0.005	1.737
GT25	pfam01755	lipopolysaccharide beta-1,4-galactosyltransferase (EC 2.4.1.-); beta-1,3-glucosyltransferase (EC 2.4.1.-); beta-1,2-glucosyltransferase (EC 2.4.1.-); beta-1,2-galactosyltransferase (EC 2.4.1.-)	0	0.000	8	0.037	0.000
GT26	pfam03808	UDP-ManNAcA: β-N-acetyl mannosaminuronyltransferase (EC 2.4.1.-); UDP-ManNAc: β-N-acetyl-mannosaminyltransferase (EC 2.4.1.-); UDP-Glc: β-1,4-glucosyltransferase (EC 2.4.1.-)	1	0.008	67	0.307	0.026
GT27	cd02510	polypeptide α-N-acetylgalactosaminyltransferase (EC 2.4.1.41)	4	0.032	61	0.280	0.114
GT28	pfam04101	1,2-diacylglycerol 3-beta-galactosyltransferase (EC 2.4.1.46); 1,2-diacylglycerol 3-beta-glucosyltransferase (EC 2.4.1.157); UDP-GlcNAc: Und-PP-MurAc-pentapeptide beta-N-acetylglucosaminyltransferase (EC 2.4.1.227)	0	0.000	121	0.554	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT29	pfam00777	sialyltransferase (EC 2.4.99.-); beta-galactoside alpha-2,6-sialyltransferase (EC 2.4.99.1); alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase (EC 2.4.99.3); beta-galactoside alpha-2,3-sialyltransferase (EC 2.4.99.4); N-acetyllactosaminide alpha-2,3-sialyltransferase (EC 2.4.99.6); (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase (EC 2.4.99.7); alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase (EC 2.4.99.8); lactosylceramide alpha-2,3-sialyltransferase (EC 2.4.99.9)	0	0.000	4	0.018	0.000
GT30	pfam04413	CMP-β-KDO: α-3-deoxy-D-manno-octulosonic-acid (KDO) transferase (EC 2.4.99.-)	1	0.008	46	0.211	0.038
GT31	pfam01762	N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase (EC 2.4.1.149); Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase (EC 2.4.1.122); fucose-specific beta-1,3-N-acetylglucosaminyltransferase (EC 2.4.1.-); globotriosylceramide beta-1,3-GalNAc transferase (EC 2.4.1.79); chondroitin synthase (beta-1,3-GlcUA and beta-1,4-GalNAc transferase (EC 2.4.1.175); chondroitin beta-1,3-glucuronyltransferase (EC 2.4.1.226); chondroitin beta-1,4-N-acetylgalactosaminyltransferase (EC 2.4.1.-)	0	0.000	4	0.018	0.000
GT32	pfam04488	alpha-1,6-mannosyltransferase (EC 2.4.1.-); alpha-1,4-N-acetylglucosaminyltransferase (EC 2.4.1.-); alpha-1,4-N-acetylgalactosaminyltransferase (EC 2.4.1.-);GDP-Man: inositol-phosphorylceramide transferase (EC 2.4.1.-); UDP-Gal: b-galactoside a-1,4-galactosyltransferase (EC 2.4.1.-); UDP-Gal: lactose/N-acetyl-lactosamine a-1,4-galactosyltransferase (EC 2.4.1.-)	0	0.000	56	0.257	0.000
GT33	cd03816	GDP-Man: chitobiosyldiphosphodolichol beta-mannosyltransferase (EC 2.4.1.142)	0	0.000	3	0.014	0.000
GT34	pfam05637	UDP-Gal: galactomannan alpha-1,6-galactosyltransferase (EC 2.4.1.-); UDP-Xyl: xyloglucan alpha-1,6-xylosyltransferase (EC 2.4.2.39); alpha-1,2-galactosyltransferase (EC 2.4.1.-)	0	0.000	4	0.018	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT35	pfam00343	glycogen or starch phosphorylase (EC 2.4.1.1)	67	0.533	472	2.163	0.247
GT37	pfam03254	galactoside 2-L-fucosyltransferase (EC 2.4.1.69)	0	0.000	4	0.018	0.000
GT38	pfam07388	polysialyltransferase (EC 2.4.-.-)	4	0.032	4	0.018	1.737
GT39	pfam02366	Dol-P-Man: protein α-mannosyltransferase (EC 2.4.1.109)	6	0.048	62	0.284	0.168
GT40	cd04186	beta-1,3-galactofuranosyltransferases (EC 2.4.1.-)	0	0.000	27	0.124	0.000
GT41	COG3914	UDP-GlcNAc: peptide β-N-acetylglucosaminyltransferase (EC 2.4.1.94)	266	2.117	505	2.314	0.915
GT42	pfam06002	CMP-NeuAc alpha-2,3-sialyltransferase (EC 2.4.99.-)	0	0.000	7	0.032	0.000
GT43	pfam03360	beta-glucuronyltransferase (EC 2.4.1.135); UDP-Xyl: xylan beta-1,4-xylosyltransferase (EC 2.4.2.-)	0	0.000	2	0.009	0.000
GT44	pfam04488	UDP-Glc: α-glucosyltransferase (EC 2.4.1.-); UDP-GlcNAc: α-N-acetylglucosaminyltransferase (EC 2.4.1.-)	1	0.008	7	0.032	0.248
GT45	pfam00535	α-N-acteylglucosaminyltransferase (EC 2.4.1.-)	4	0.032	30	0.137	0.232
GT46	self-built	This family has been deleted until a convincing sufficient biochemical characterisation is found for a member of this family	0	0.000	6	0.027	0.000
GT49	self-built	beta-1,3-N-acetylglucosaminyltransferase (EC 2.4.1.-)	0	0.000	2	0.009	0.000
GT50	pfam05007	Dol-P-Man alpha-1,4-mannosyltransferase (EC 2.4.1.-)	0	0.000	5	0.023	0.000
GT51	pfam00912	murein polymerase (EC 2.4.1.129)	93	0.740	188	0.861	0.859
GT52	pfam07922	alpha-2,3-sialyltransferase (EC 2.4.99.4); alpha-glucosyltransferase (EC 2.4.1.-)	0	0.000	2	0.009	0.000
GT53	pfam04602	UDP-L-Ara: α-L-arabinosyltransferase (EC 2.4.2.-)	1	0.008	4	0.018	0.434
GT54	pfam04666	UDP-GlcNAc: α-1,3-D-mannoside β-1,4-N-acetylglucosaminyltransferase (EC 2.4.1.145)	1	0.008	2	0.009	0.868
GT55	pfam09488	GDP-Man: mannosyl-3-phosphoglycerate synthase (EC 2.4.1.217)	0	0.000	2	0.009	0.000
GT56	pfam07429	TDP-Fuc4NAc: lipid II Fuc4NAc transferase (EC 2.4.1.-)	1	0.008	6	0.027	0.289
GT57	pfam03155	Dol-P-Glc: alpha-1,3-glucosyltransferase (EC 2.4.1.-)	0	0.000	1	0.005	0.000
GT58	pfam05208	Dol-P-Man: dolichol pyrophosphate-mannose alpha-1,3-mannosyltransferase (EC 2.4.1.130); Dol-P-Man: dolichol pyrophosphate-Man5GlcNAc2 alpha-1,2-mannosyltransferase (EC 2.4.1.130)	0	0.000	1	0.005	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT59	pfam04922	Dol-P-Glc: Glc2Man9GlcNAc2-PP-Dol alpha-1,2-glucosyltransferase (EC 2.4.1.-)	0	0.000	3	0.014	0.000
GT60	pfam11397	UDP-GlcNAc: polypeptide alpha-N-acetylglucosaminyltransferase (EC 2.4.1.-); UDP-GlcNAc: hydroxyproline polypeptide alpha-N-acetylglucosaminyltransferase (EC 2.4.1.-)	0	0.000	3	0.014	0.000
GT61	pfam04577	beta-1,2-xylosyltransferase (EC 2.4.2.38)	0	0.000	4	0.018	0.000
GT62	pfam03452	alpha-1,2-mannosyltransferase (EC 2.4.1.-); alpha-1,6-mannosyltransferase (EC 2.4.1.-)	0	0.000	4	0.018	0.000
GT63	self-built	UDP-Glc: DNA beta-glucosyltransferase (EC 2.4.1.27)	0	0.000	7	0.032	0.000
GT64	pfam09258	UDP-GlcNAc: heparan alpha-N-acetylhexosaminyltransferase (EC 2.4.1.224)	0	0.000	5	0.023	0.000
GT65	pfam10250	GDP-Fuc: protein O-alpha-fucosyltransferase (EC 2.4.1.-)	0	0.000	1	0.005	0.000
GT66	pfam02516	Dol-PP-α-oligosaccharide: protein β-oligosaccharyltransferase (EC 2.4.1.119)	16	0.127	73	0.335	0.381
GT67	PTZ00210	UDP-Gal: phosphoglycan β-1,3 galactosyltransferase 1 (SCG1) (EC 2.4.1.-)	2	0.016	0	0.000	0.000
GT68	pfam10250	GDP-Fuc: protein O-alpha-fucosyltransferase (EC 2.4.1.-)	0	0.000	1	0.005	0.000
GT70	self-built	UDP-GlcA: β-glucuronosyltransferase (EC 2.4.1.17)	1	0.008	4	0.018	0.434
GT71	pfam11051	alpha-mannosyltransferase (EC 2.4.1.-)	0	0.000	1	0.005	0.000
GT72	pfam11440	UDP-Glc: DNA alpha-glucosyltransferase (EC 2.4.1.26)	0	0.000	1	0.005	0.000
GT73	PRK09822	CMP-beta-KDO: alpha-3-deoxy-D-manno-octulosonic-acid (KDO) transferase (EC 2.4.99.-)	0	0.000	18	0.082	0.000
GT74	self-built	alpha-1,2-L-fucosyltransferase (EC 2.4.1.69)	0	0.000	7	0.032	0.000
GT75	pfam03214	UDP-Glc: self-glucosylating beta-glucosyltransferase (EC 2.4.1.-)	0	0.000	3	0.014	0.000
GT76	pfam04188	Dol-P-Man: alpha-1,6-mannosyltransferase (EC 2.4.1.-)	0	0.000	10	0.046	0.000
GT77	pfam03407	α-xylosyltransferase (EC 2.4.2.39); α-1,3-galactosyltransferase (EC 2.4.1.37); arabinosyltransferase (EC 2.4.2.-); arabinosyltransferase (EC 2.4.2.-)	1	0.008	0	0.000	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT78	self-built	GDP-Man: α-mannosyltransferase (mannosylglycerate synthase) (EC 2.4.1.-)	1	0.008	2	0.009	0.868
GT79	self-built	GDP-D-Ara: phosphoglycan α-1,2-D-arabinopyranosyltransferase 1 (EC 2.4.2.-)	1	0.008	4	0.018	0.434
GT80	pfam11477	beta-galactoside alpha-2,6-sialyltransferase (EC 2.4.99.1); beta-galactoside alpha-2,3-sialyltransferase (EC 2.4.99.4)	0	0.000	3	0.014	0.000
GT81	PRK13915	NDP-Glc: glucosyl-3-phosphoglycerate synthase (EC 2.4.1.-); NDP-Man: mannosyl-3-phosphoglycerate synthase (EC 2.4.1.-);ADP-Glc: glucosyl-2-glycerate synthase (EC 2.4.1.-)	0	0.000	59	0.270	0.000
GT82	pfam06306	UDP-GalNAc: beta-1,4-N-acetylgalactosaminyltransferase (EC 2.4.1.-)	0	0.000	8	0.037	0.000
GT83	COG1807	undecaprenyl phosphate-alpha-L-Ara4N: 4-amino-4-deoxy-beta-L-arabinosyltransferase (EC 2.4.2.-); dodecaprenyl phosphate-beta-galacturonic acid: lipopolysaccharide core alpha-galacturonosyl transferase (EC 2.4.1.-)	3	0.024	74	0.339	0.070
GT84	pfam10091	cyclic beta-1,2-glucan synthase (EC 2.4.1.-)	0	0.000	9	0.041	0.000
GT85	pfam12250	β-D-arabinofuranosyl monophosphoryldecaprenol: galactan α-D-arabinofuranosyltransferase (EC 2.4.2.-)	2	0.016	6	0.027	0.579
GT87	pfam09594	polyprenol-P-Man: α-1,2-mannosyltransferase (EC 2.4.1.-)	5	0.040	7	0.032	1.241
GT88	self-built	UDP-Glc: alpha-glucosyltransferase (EC 2.4.1.-)	0	0.000	4	0.018	0.000
GT89	self-built	β-D-arabinofuranosyl-1-monophosphoryldecaprenol : arabinan β-1,2-arabinofuranosyltransferase (EC 2.4.2.-)	3	0.024	9	0.041	0.579
GT90	smart00672	UDP-Xyl: (mannosyl) glucuronoxylomannan/galactoxylomannan beta-1,2-xylosyltransferase (EC 2.4.2.-)	0	0.000	9	0.041	0.000
GT91	pfam12141	β-1,2-mannosyltransferase (EC 2.4.1.-)	1	0.008	3	0.014	0.579
GT92	pfam01697	UDP-Gal: N-glycan core alpha-1,6-fucoside beta-1,4-galactosyltransferase (EC 2.4.1.-); UDP-Gal: beta-galactoside beta-1,4-galactosyltransferase (EC 2.4.1.-)	0	0.000	8	0.037	0.000
GT93		UDP-GluA : a-glucuronyltransferase (EC 2.4.1.-) involved in GAG polymerisation	0	0.000	5	0.023	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
GT94	self-built	GDP-Man: GlcA-β-1,2-Man-α-1,3-Glc-β-1,4-Glc-α-1-PP-undecaprenol β-1,4-mannosyltransferase (2.4.1.251)	108	0.860	22	0.101	8.526
Total GTs			793	6.3	5126	23.5	
PL1	smart00656	pectate lyase (EC 4.2.2.2); exo-pectate lyase (EC 4.2.2.9); pectin lyase (EC 4.2.2.10)	73	0.581	97	0.444	1.307
PL2	pfam06917	pectate lyase (EC 4.2.2.2); exo-polygalacturonate lyase (EC 4.2.2.9)	0	0.000	3	0.014	0.000
PL3	pfam03211	pectate lyase (EC 4.2.2.2)	0	0.000	11	0.050	0.000
PL4	self-built	rhamnogalacturonan lyase (EC 4.2.2.-)	65	0.517	34	0.156	3.320
PL5	pfam05426	alginate lyase (EC 4.2.2.3)	0	0.000	2	0.009	0.000
PL6	self-built	alginate lyase (EC 4.2.2.3); chondroitinase B (EC 4.2.2.4)	5	0.040	19	0.087	0.457
PL7	pfam08787	alginate lyase (EC 4.2.2.3); alpha-L-guluronate lyase (EC 4.2.2.11)	0	0.000	4	0.018	0.000
PL8	pfam02278	hyaluronate lyase (EC 4.2.2.1); chondroitin AC lyase (EC 4.2.2.5); xanthan lyase (EC 4.2.2.12); chondroitin ABC lyase (EC 4.2.2.20)	0	0.000	8	0.037	0.000
PL9	self-built	pectate lyase (EC 4.2.2.2); exopolygalacturonate lyase (EC 4.2.2.9); thiopeptidoglycan lyase (EC 4.2.2.-)	2	0.016	89	0.408	0.039
PL10	pfam09492	pectate lyase (EC 4.2.2.2)	23	0.183	30	0.137	1.332
PL11	self-built	rhamnogalacturonan lyase (EC 4.2.2.-); exo-unsaturated rhamnogalacturonan lyase (EC 4.2.2.-)	141	1.122	138	0.632	1.775
PL12	pfam07940	heparin-sulfate lyase (EC 4.2.2.8)	1	0.008	19	0.087	0.091
PL13	self-built	heparin lyase (EC 4.2.2.7)	1	0.008	2	0.009	0.868
PL14	self-built	alginate lyase (EC 4.2.2.3); exo-oligoalginate lyase (EC 4.2.2.-); beta-1,4-glucuronan lyase (EC 4.2.2.14)	0	0.000	4	0.018	0.000
PL15	pfam07940	oligo-alginate lyase (EC 4.2.2.-)	0	0.000	3	0.014	0.000
PL16	pfam07212	hyaluronan lyase (EC 4.2.2.1)	0	0.000	3	0.014	0.000
PL17	pfam07940	alginate lyase (EC 4.2.2.3)	0	0.000	2	0.009	0.000
PL18	pfam08787	alginate lyase (EC 4.2.2.3)	0	0.000	8	0.037	0.000
PL20	self-built	endo-β-1,4-glucuronan lyase (EC 4.2.2.14)	1	0.008	0	0.000	0.000

CAZy family	dbCAN HMM source	Known activities from CAZY database	MS count	MS abundance (%)	MG count	MG abundance (%)	MS(%)/MG(%) ratio
PL21	pfam07940	heparin lyase (EC 4.2.2.7); heparin-sulfate lyase (EC 4.2.2.8); acharan-sulfate lyase (EC 4.2.2.-)	2	0.016	2	0.009	1.737
PL22	TIGR02800	oligogalacturonate lyase / oligogalacturonide lyase (EC 4.2.2.6)	68	0.541	46	0.211	2.567
Total PLs			382	3.0	524	2.4	
cohesin	pfam00963	The cohesin-dockerin interaction is the crucial interaction for complex formation in the cellulosome	52	0.414	27	0.124	3.345
dockerin	pfam00404	The cohesin-dockerin interaction is the crucial interaction for complex formation in the cellulosome	1049	8.349	121	0.554	15.057
SLH	pfam00395	S-layer homology domain, anchoring cellulosome onto the bacterial cell surfaces	46	0.366	77	0.353	1.038
Total cellulosome			1147	9.1	225	1.0	
Grand total			12565	100	21823	100	

Abbreviations: MS, rumen microbial plant-adherent metasecretome dataset; MG, rumen microbial plant-adherent metagenome dataset.

Table A2.2 CAZy families predicted at higher frequency in the metasecretome compared to the metagenome dataset.

CAZy family	Known activities from CAZY database	MS(%)/MG(%) ratio
dockerin	The cohesin-dockerin interaction is the crucial interaction for complex formation in the cellulosome	15.1
GH124	endoglucanase (EC 3.2.1.4)	14.3
CBM40	Binding to sialic acid	10.6
GT94	GDP-Man: GlcA-β-1,2-Man-α-1,3-Glc-β-1,4-Glc-α-1-PP-undecaprenol β-1,4-mannosyltransferase (2.4.1.251)	8.5
CBM38	Binding to inulin	6.9
CBM16	Binding to cellulose and glucomannan	6.8
GH55	exo-β-1,3-glucanase (EC 3.2.1.58); endo-β-1,3-glucanase (EC 3.2.1.39)	6.5
GH92	mannosyl-oligosaccharide α-1,2-mannosidase (EC 3.2.1.113); mannosyl-oligosaccharide α-1,3-mannosidase (EC 3.2.1.-); mannosyl-oligosaccharide α-1,6-mannosidase (EC 3.2.1.-); α-mannosidase (EC 3.2.1.24); α-1,2-mannosidase (EC 3.2.1.-); α-1,3-mannosidase (EC 3.2.1.-); α-1,4-mannosidase (EC 3.2.1.-)	5.9
CE8	pectin methylesterase (EC 3.1.1.11).	5.4
GH30	glucosylceramidase (EC 3.2.1.45); β-1,6-glucanase (EC 3.2.1.75); β-xylosidase (EC 3.2.1.37); β-fucosidase (EC 3.2.1.38); β-glucosidase (3.2.1.21); endo-β-1,6-galactanase (EC:3.2.1.164)	4.3
CBM67	Binding to L-rhamnose	4.2
GH53	endo-β-1,4-galactanase (EC 3.2.1.89)	4.0
cohesin	The cohesin-dockerin interaction is the crucial interaction for complex formation in the cellulosome	3.3
PL4	rhamnogalacturonan lyase (EC 4.2.2.-)	3.3
CBM30	Binding to cellulose	3.2
GH121	β-L-arabinobiosidase (EC 3.2.1.-)	3.2
CE3	acetyl xylan esterase (EC 3.1.1.72).	3.0
GH25	lysozyme (EC 3.2.1.17)	2.9
GH127	β-L-arabinofuranosidase (EC 3.2.1.-)	2.8
CE7	acetyl xylan esterase (EC 3.1.1.72); cephalosporin-C deacetylase (EC 3.1.1.41).	2.7
GH97	α-glucosidase (EC 3.2.1.20); α-galactosidase (EC 3.2.1.22)	2.6
GH74	endoglucanase (EC 3.2.1.4); oligoxyloglucan reducing end-specific cellobiohydrolase (EC 3.2.1.150); xyloglucanase (EC 3.2.1.151)	2.6
PL22	oligogalacturonate lyase / oligogalacturonide lyase (EC 4.2.2.6)	2.6
GH43	β-xylosidase (EC 3.2.1.37); β-1,3-xylosidase (EC 3.2.1.-); α-L-arabinofuranosidase (EC 3.2.1.55); arabinanase (EC 3.2.1.99); xylanase (EC 3.2.1.8); galactan 1,3-β-galactosidase (EC 3.2.1.145)	2.3
GH26	β-mannanase (EC 3.2.1.78); β-1,3-xylanase (EC 3.2.1.32)	2.3
CBM20	Binding to granular starch and cyclodextrins	2.3
CBM61	Binding to β-1,4-galactan	2.3
CBM26	Binding to starch	2.2
GH105	unsaturated rhamnogalacturonyl hydrolase (EC 3.2.1.-)	1.9
PL11	rhamnogalacturonan lyase (EC 4.2.2.-); exo-unsaturated rhamnogalacturonan lyase (EC 4.2.2.-)	1.8
GH16	xyloglucan:xyloglucosyltransferase (EC 2.4.1.207); keratan-sulfate endo-1,4-β-galactosidase (EC 3.2.1.103); endo-1,3-β-glucanase (EC 3.2.1.39); endo-1,3(4)-β-glucanase (EC 3.2.1.6); licheninase (EC 3.2.1.73); β-agarase (EC 3.2.1.81); κ-carrageenase (EC 3.2.1.83); xyloglucanase (EC 3.2.1.151)	1.7
GH2	β-galactosidase (EC 3.2.1.23) ; β-mannosidase (EC 3.2.1.25); β-glucuronidase (EC 3.2.1.31); mannosylglycoprotein endo-β-mannosidase (EC 3.2.1.152); exo-β-glucosaminidase (EC 3.2.1.165)	1.3
CE1	acetyl xylan esterase (EC 3.1.1.72); cinnamoyl esterase (EC 3.1.1.-); feruloyl esterase (EC 3.1.1.73); carboxylesterase (EC 3.1.1.1); S-formylglutathione hydrolase (EC 3.1.2.12)	1.3
GH3	β-glucosidase (EC 3.2.1.21); xylan 1,4-β-xylosidase (EC 3.2.1.37); β-N-acetylhexosaminidase (EC 3.2.1.52); glucan 1,3-β-glucosidase (EC 3.2.1.58); glucan 1,4-β-glucosidase (EC 3.2.1.74); exo-1,3-1,4-glucanase (EC 3.2.1.-); αlpha-L-arabinofuranosidase (EC 3.2.1.55).	1.2

Abbreviations: MS, rumen microbial plant-adherent metasecretome dataset; MG, rumen microbial plant-adherent metagenome dataset.

Table A2.3 CAZy families predicted at lower frequency in the metasecretome compared to the metagenome dataset.

CAZy family	Known activities from CAZY database	MS(%)/MG(%) ratio
CBM34	Binding to granular starch	0.000
CE11	UDP-3-0-acyl N-acetylglucosamine deacetylase (EC 3.5.1.-)	0.000
GH120	beta-xylosidase (EC 3.2.1.37)	0.000
GH50	beta-agarase (EC 3.2.1.81)	0.000
GH57	alpha-amylase (EC 3.2.1.1); 4-alpha-glucanotransferase (EC 2.4.1.25); alpha-galactosidase (EC 3.2.1.22); amylopullulanase (EC 3.2.1.41); branching enzyme (EC 2.4.1.18)	0.000
GT1	UDP-glucuronosyltransferase (EC 2.4.1.17); 2-hydroxyacylsphingosine 1-beta-galactosyltransferase (EC 2.4.1.45); N-acylsphingosine galactosyltransferase (EC 2.4.1.47); flavonol 3-O-glucosyltransferase (EC 2.4.1.91); indole-3-acetate beta-glucosyltransferase (EC 2.4.1.121); sterol glucosyltransferase (EC 2.4.1.173); ecdysteroid UDP-glucosyltransferase (EC 2.4.1.-); zeaxanthin glucosyltransferase (EC 2.4.1.-); zeatin O-beta-glucosyltransferase (EC 2.4.1.203); zeatin O-beta-xylosyltransferase (EC 2.4.2.40); limonoid glucosyltransferase (EC 2.4.1.210); sinapate 1-glucosyltransferase (EC 2.4.1.120); anthocyanin 3-O-galactosyltransferase (EC 2.4.1.-); anthocyanin 5-O-glucosyltransferase (EC 2.4.1.-); anthocyanidin 3-O-glucosyltransferase (EC 2.4.1.115); dTDP-beta-2-deoxy-L-fucose: alpha-L-2-deoxyfucosyltransferase (EC 2.4.1.-); UDP-beta-L-rhamnose: alpha-L-rhamnosyltransferase (EC 2.4.1.-); UDP-glucose: 4-hydroxybenzoate 4-O-beta-glucosyltransferase (EC 2.4.1.194); flavonol L-rhamnosyltransferase (EC 2.4.1.159); salicylic acid beta-glucosyltransferase (EC 2.4.1.-)	0.000
GT10	galactoside alpha-1,3/1,4-L-fucosyltransferase (EC 2.4.1.65); galactoside alpha-1,3-L-fucosyltransferase (EC 2.4.1.152); glycoprotein alpha-1,3-L-fucosyltransferase (EC 2.4.1.214)	0.000
GT13	alpha-1,3-mannosyl-glycoprotein beta-1,2-N-acetylglucosaminyltransferase (EC 2.4.1.101)	0.000
GT28	1,2-diacylglycerol 3-beta-galactosyltransferase (EC 2.4.1.46); 1,2-diacylglycerol 3-beta-glucosyltransferase (EC 2.4.1.157); UDP-GlcNAc: Und-PP-MurAc-pentapeptide beta-N-acetylglucosaminyltransferase (EC 2.4.1.227)	0.000
GT32	alpha-1,6-mannosyltransferase (EC 2.4.1.-); alpha-1,4-N-acetylglucosaminyltransferase (EC 2.4.1.-); alpha-1,4-N-acetylgalactosaminyltransferase (EC 2.4.1.-);GDP-Man: inositol-phosphorylceramide transferase (EC 2.4.1.-); UDP-Gal: b-galactoside a-1,4-galactosyltransferase (EC 2.4.1.-); UDP-Gal: lactose/N-acetyl-lactosamine a-1,4-galactosyltransferase (EC 2.4.1.-)	0.000
GT40	beta-1,3-galactofuranosyltransferases (EC 2.4.1.-)	0.000
GT73	CMP-beta-KDO: alpha-3-deoxy-D-manno-octulosonic-acid (KDO) transferase (EC 2.4.99.-)	0.000
GT8	lipopolysaccharide alpha-1,3-galactosyltransferase (EC 2.4.1.44); UDP-Glc: (glucosyl)lipopolysaccharide alpha-1,2-glucosyltransferase (EC 2.4.1.-); lipopolysaccharide glucosyltransferase 1 (EC 2.4.1.58); glycogenin glucosyltransferase (EC 2.4.1.186); inositol 1-alpha-galactosyltransferase (galactinol synthase) (EC 2.4.1.123); homogalacturonan alpha-1,4-galacturonosyltransferase (EC 2.4.1.43); UDP-GlcA: xylan alpha-glucuronyltransferase (EC 2.4.1.-)	0.000
GT81	NDP-Glc: glucosyl-3-phosphoglycerate synthase (EC 2.4.1.-); NDP-Man: mannosyl-3-phosphoglycerate synthase (EC 2.4.1.-);ADP-Glc: glucosyl-2-glycerate synthase (EC 2.4.1.-)	0.000
GT9	lipopolysaccharide N-acetylglucosaminyltransferase (EC 2.4.1.56); heptosyltransferase (EC 2.4.-.-)	0.000
CE9	N-acetylglucosamine 6-phosphate deacetylase (EC 3.5.1.25); N-acetylglucosamine 6-phosphate deacetylase (EC 3.5.1.80)	0.016
GT11	GDP-L-Fuc: galactoside α-1,2-L-fucosyltransferase (EC 2.4.1.69); GDP-L-Fuc: β-LacNac α-1,3-L-fucosyltransferase (EC 2.4.1.-)	0.018
GT3	glycogen synthase (EC 2.4.1.11)	0.021
AA3	cellobiose dehydrogenase (EC 1.1.99.18); glucose 1-oxidase (EC 1.1.3.4); aryl alcohol oxidase (EC 1.1.3.7); alcohol oxidase (EC 1.1.3.13); pyranose oxidase (EC 1.1.3.10)	0.023
GT26	UDP-ManNAcA: β-N-acetyl mannosaminuronyltransferase (EC 2.4.1.-); UDP-ManNAc: β-N-acetyl-mannosaminyltransferase (EC 2.4.1.-); UDP-Glc: β-1,4-glucosyltransferase (EC 2.4.1.-)	0.026
GH78	α-L-rhamnosidase (EC 3.2.1.40)	0.027
GH1	β-glucosidase (EC 3.2.1.21); β-galactosidase (EC 3.2.1.23); β-mannosidase (EC 3.2.1.25); β-glucuronidase (EC 3.2.1.31); β-D-fucosidase (EC 3.2.1.38); phlorizin hydrolase (EC 3.2.1.62); exo-β-1,4-glucanase (EC 3.2.1.74); 6-phospho-β-galactosidase (EC 3.2.1.85); 6-phospho-β-glucosidase (EC 3.2.1.86); strictosidine β-glucosidase (EC 3.2.1.105); lactase (EC 3.2.1.108); amygdalin β-glucosidase (EC 3.2.1.117); prunasin β-glucosidase (EC 3.2.1.118); raucaffricine β-glucosidase (EC 3.2.1.125); thioglucosidase (EC 3.2.1.147); β-primeverosidase (EC 3.2.1.149); isoflavonoid 7-O-β-apiosyl-β-glucosidase (EC 3.2.1.161); hydroxyisourate hydrolase (EC 3.-.-.-); β-glycosidase (EC 3.2.1.-)	0.031
GH20	β-hexosaminidase (EC 3.2.1.52); lacto-N-biosidase (EC 3.2.1.140); β-1,6-N-acetylglucosaminidase) (EC 3.2.1.-); β-6-SO3-N-acetylglucosaminidase (EC 3.2.1.-)	0.034
CBM2	Binding to cellulose, chitin and xylan.	0.035
GT30	CMP-β-KDO: α-3-deoxy-D-manno-octulosonic-acid (KDO) transferase (EC 2.4.99.-)	0.038
PL9	pectate lyase (EC 4.2.2.2); exopolygalacturonate lyase (EC 4.2.2.9); thiopeptidoglycan lyase (EC 4.2.2.-)	0.039
GH35	β-galactosidase (EC 3.2.1.23); exo-β-glucosaminidase (EC 3.2.1.165)	0.040
GH112	lacto-N-biose phosphorylase or galacto-N-biose phosphorylase (EC 2.4.1.211); D-galactosyl-1,4-L-rhamnose phosphorylase (EC 2.4.1.-)	0.047
GH39	α-L-iduronidase (EC 3.2.1.76); β-xylosidase (EC 3.2.1.37)	0.047
GT5	UDP-Glc: glycogen glucosyltransferase (EC 2.4.1.11); ADP-Glc: starch glucosyltransferase (EC 2.4.1.21); NDP-Glc: starch glucosyltransferase (EC 2.4.1.242); UDP-Glc: α-1,3-glucan synthase (EC 2.4.1.183) UDP-Glc: α-1,4-glucan synthase (EC 2.4.1.-)	0.058

CAZy family	Known activities from CAZY database	MS(%)/MG(%) ratio
AA8	iron reductase	0.066
GH24	lysozyme (EC 3.2.1.17)	0.067
GT83	undecaprenyl phosphate-alpha-L-Ara4N: 4-amino-4-deoxy-beta-L-arabinosyltransferase (EC 2.4.2.-); dodecaprenyl phosphate-beta-galacturonic acid: lipopolysaccharide core alpha-galacturonosyl transferase (EC 2.4.1.-)	0.070
GH42	β-galactosidase (EC 3.2.1.23)	0.074
AA5	oxidase with oxygen as acceptor (EC 1.1.3.-); galactose oxidase (EC 1.1.3.9)	0.076
GH73	peptidoglycan hydrolase with endo-β-N-acetylglucosaminidase specificity (EC 3.2.1.-)	0.078
GT2	cellulose synthase (EC 2.4.1.12); chitin synthase (EC 2.4.1.16); dolichyl-phosphate β-D-mannosyltransferase (EC 2.4.1.83); dolichyl-phosphate β-glucosyltransferase (EC 2.4.1.117); N-acetylglucosaminyltransferase (EC 2.4.1.-); N-acetylgalactosaminyltransferase (EC 2.4.1.-); hyaluronan synthase (EC 2.4.1.212); chitin oligosaccharide synthase (EC 2.4.1.-); β-1,3-glucan synthase (EC 2.4.1.34); β-1,4-mannan synthase (EC 2.4.1.-); β-mannosylphosphodecaprenol-mannooligosaccharide α-1,6-mannosyltransferase (EC 2.4.1.199); α-1,3-L-rhamnosyltransferase (EC 2.4.1.-)	0.086
GH89	α-N-acetylglucosaminidase (EC 3.2.1.50)	0.094
GT14	β-1,3-galactosyl-O-glycosyl-glycoprotein β-1,6-N-acetylglucosaminyltransferase (EC 2.4.1.102); N-acetyllactosaminide β-1,6-N-acetylglucosaminyltransferase (EC 2.4.1.150); protein O-β-xylosyltransferase (EC 2.4.2.26)	0.094
GH109	α-N-acetylgalactosaminidase (EC 3.2.1.49)	0.095
CBM56	Binding to β-1,3-glucan	0.101
GH88	δ-4,5 unsaturated beta-glucuronyl hydrolase (EC 3.2.1.-)	0.105
GT27	polypeptide α-N-acetylgalactosaminyltransferase (EC 2.4.1.41)	0.114
GH8	chitosanase (EC 3.2.1.132); cellulase (EC 3.2.1.4); licheninase (EC 3.2.1.73); endo-1,4-β-xylanase (EC 3.2.1.8); reducing-end-xylose releasing exo-oligoxylanase (EC 3.2.1.156)	0.124
GT39	Dol-P-Man: protein α-mannosyltransferase (EC 2.4.1.109)	0.168
CE2	acetyl xylan esterase (EC 3.1.1.72).	0.197
CBM12	Binding to chitin	0.204
GH23	lysozyme type G (EC 3.2.1.17); peptidoglycan lyase (EC 4.2.2.-) also known in the literature as peptidoglycan lytic transglycosylase	0.227
GH67	α-glucuronidase (EC 3.2.1.139); xylan α-1,2-glucuronidase (EC 3.2.1.131)	0.233
GT4	sucrose synthase (EC 2.4.1.13); sucrose-phosphate synthase (EC 2.4.1.14); α-glucosyltransferase (EC 2.4.1.52); lipopolysaccharide N-acetylglucosaminyltransferase (EC 2.4.1.56); GDP-Man α-mannosyltransferase (EC 2.4.1.-); 1,2-diacylglycerol 3-glucosyltransferase (EC 2.4.1.157); diglucosyl diacylglycerol synthase (EC 2.4.1.208); digalactosyldiacylglycerol synthase (EC 2.4.1.141); trehalose phosphorylase (EC 2.4.1.231); phosphatidylinositol α-mannosyltransferase (EC 2.4.1.57); UDP-Gal α-galactosyltransferase (EC 2.4.1.-); UDP-Xyl α-xylosyltransferase (EC 2.4.2.-)	0.241
GT35	glycogen or starch phosphorylase (EC 2.4.1.1)	0.247
GH94	cellobiose phosphorylase (EC 2.4.1.20); cellodextrin phosphorylase (EC 2.4.1.49); chitobiose phosphorylase (EC 2.4.1.-); cyclic β-1,2-glucan synthase (EC 2.4.1.-)	0.249
CE15	4-O-methyl-glucuronoyl methylesterase (EC 3.1.1.-)	0.265
GH29	α-L-fucosidase (EC 3.2.1.51)	0.307
GH51	α-L-arabinofuranosidase (EC 3.2.1.55); endoglucanase (EC 3.2.1.4)	0.307
GH115	xylan α-1,2-glucuronidase (3.2.1.131); α-(4-O-methyl)-glucuronidase (3.2.1.-)	0.331
GH27	α-galactosidase (EC 3.2.1.22); α-N-acetylgalactosaminidase (EC 3.2.1.49); isomalto-dextranase (EC 3.2.1.94); β-L-arabinopyranosidase (EC 3.2.1.88)	0.351
GH18	chitinase (EC 3.2.1.14); endo-β-N-acetylglucosaminidase (EC 3.2.1.96); xylanase inhibitor; concanavalin B; narbonin	0.369
GT66	Dol-PP-α-oligosaccharide: protein β-oligosaccharyltransferase (EC 2.4.1.119)	0.381
GH13	α-amylase (EC 3.2.1.1); pullulanase (EC 3.2.1.41); cyclomaltodextrin glucanotransferase (EC 2.4.1.19); cyclomaltodextrinase (EC 3.2.1.54); trehalose-6-phosphate hydrolase (EC 3.2.1.93); oligo-α-glucosidase (EC 3.2.1.10); maltogenic amylase (EC 3.2.1.133); neopullulanase (EC 3.2.1.135); α-glucosidase (EC 3.2.1.20); maltotetraose-forming α-amylase (EC 3.2.1.60); isoamylase (EC 3.2.1.68); glucodextranase (EC 3.2.1.70); maltohexaose-forming α-amylase (EC 3.2.1.98); maltotriose-forming α-amylase (EC 3.2.1.116); branching enzyme (EC 2.4.1.18); trehalose synthase (EC 5.4.99.16); 4-α-glucanotransferase (EC 2.4.1.25); maltopentaose-forming α-amylase (EC 3.2.1.-); amylosucrase (EC 2.4.1.4); sucrose phosphorylase (EC 2.4.1.7); malto-oligosyltrehalose trehalohydrolase (EC 3.2.1.141); isomaltulose synthase (EC 5.4.99.11); amino acid transporter	0.403
CBM6	Binding to amorphous cellulose, β-1,4-xylan, β-1,3-glucan, β-1,3-1,4-glucan, and β-1,4-glucan	0.421
CE6	acetyl xylan esterase (EC 3.1.1.72).	0.458
GH10	endo-1,4-β-xylanase (EC 3.2.1.8); endo-1,3-β-xylanase (EC 3.2.1.32)	0.460
GH77	amylomaltase or 4-α-glucanotransferase (EC 2.4.1.25)	0.466
AA6	1,4-benzoquinone reductase (EC. 1.6.5.6)	0.514
CE4	acetyl xylan esterase (EC 3.1.1.72); chitin deacetylase (EC 3.5.1.41); chitooligosaccharide deacetylase (EC 3.5.1.-); peptidoglycan GlcNAc deacetylase (EC 3.5.1.-); peptidoglycan N-acetylmuramic acid deacetylase (EC 3.5.1.-).	0.522
CE10	arylesterase (EC 3.1.1.-); carboxyl esterase (EC 3.1.1.3); acetylcholinesterase (EC 3.1.1.7); cholinesterase (EC 3.1.1.8); sterol esterase (EC 3.1.1.13); brefeldin A esterase (EC 3.1.1.-)	0.676
GT41	UDP-GlcNAc: peptide β-N-acetylglucosaminyltransferase (EC 2.4.1.94)	0.915

Abbreviations: MS, rumen microbial plant-adherent metasecretome dataset; MG, rumen microbial plant-adherent metagenome dataset.

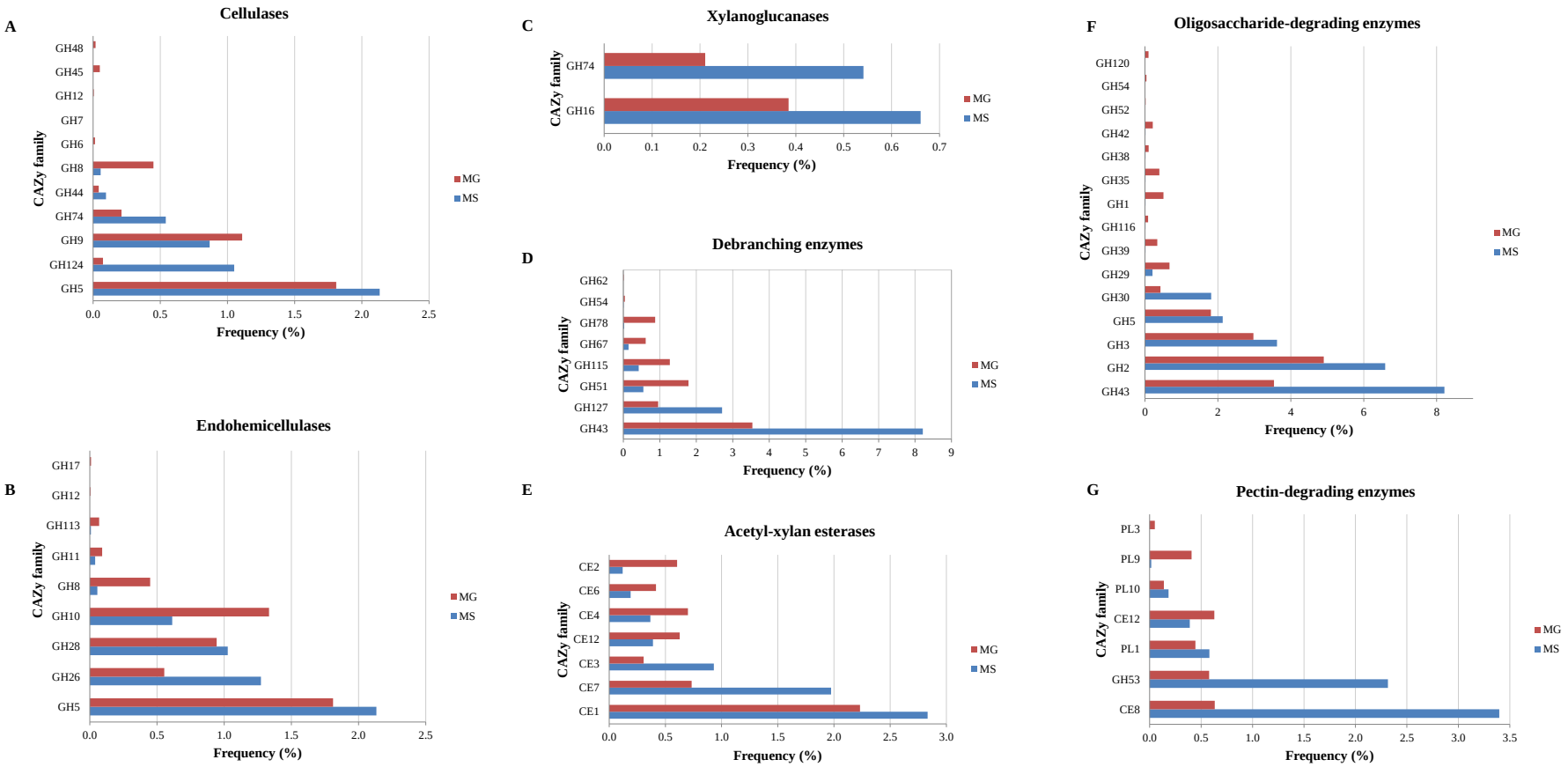

Figure A2.1. Overview of the frequencies of major putative CAZy families involved in the degradation of plant cell wall polysaccharides in the metasecretome and the metagenome dataset.

Abbreviations: MS, rumen microbial plant-adherent metasecretome dataset; MG, rumen microbial plant-adherent metagenome dataset; GH, glycoside hydrolase; CE, carbohydrate esterase; PL, polysaccharide lyase. GHs, CEs and PLs are grouped according to their major functional role in degradation of of plant cell wall polysaccharides [64].

Table A2.4. Candidate putative proteins with predicted multi-modular organisation in the metasecretome dataset.

Query ID [a]	Query length (aa)	Target (dbCAN HMM)	Target length (aa)	iE-value	Bit score	cE-value	Target start	Target end	Target covered fraction	Query start	Query end	Query covered fraction	Query sequence	Best DELTA BLASTP hit [b]
RMSec_ 22195141	198	CBM32	124	9.20E-10	33.4	2.80E-12	37	109	0.58	26	104	0.39	AFKKMIDEVFRKNLAEKA RRQTKGDETVIDFGKPTTF NRFLVEEDIRYGQRVKKFK LEAEVDGQWQQLKDALVE NGDGLTTIGHRRIICFPTVN	hypothetical protein [Prevotella sp. P6B4]
		CBM32	124	7.50E-08	27.2	2.30E-10	32	81	0.40	142	193	0.26	ATKLRFTIVDTKCGPIIKKL GVYLAPELTADIPDAGEKK SSNLHLFFSSPTQMMIDWD TEQTITSFRYLPPQESKDGT VTHYTLWASTDWTNWTK LASGEFSNV	
RMSec_ 22291211	260	CBM50	40	1.90E-13	45.0	5.80E-16	1	40	0.98	72	114	0.16	MSFYYYETGANESIYDIAA KLGVTKDYIIKNNPSAADG IENGMTLYFPVSESNASKP AKQETVVAPATANTHVVE QGETLYGLAKRYGVTIDEL	glycosyl hydrolase family 25 [Clostridium sp. ATCC 29733]
		CBM50	40	4.80E-05	18.1	1.40E-07	1	33	0.80	154	185	0.11	ILANPGSENGIKIGQKLNIP SANATASSTTNANQQVRN AFENAASPATVAVPQGSDP VFHTLQAGESIYSLAKQYN SSIEGIITANPGLKPEEYTQ	
		CBM50	40	2.40E-05	19.0	7.30E-08	4	28	0.60	219	243	0.09	GAKVKVVPNIALPFIYERT GRRNYKYEAGRGETFATIA AANGITEEELKAANPTEKN VKKGKNHHSAQAL	
RMSec_ 22522671	275	GH16	189	1.70E-49	163.3	1.00E-51	36	189	0.81	2	198	0.71	RDGKLVLKAIKTQKDGKD YYTSGKVTGQNKTDFQYG	laminarinase [Ruminococcaceae

Query ID [a]	Query length (aa)	Target (dbCAN HMM)	Target length (aa)	iE-value	Bit score	cE-value	Target start	Target end	Target covered fraction	Query start	Query end	Query covered fraction	Query sequence	Best DELTA BLASTP hit [b]
													KVVVSAKVPEGQGLWPAI WMMPKDESFYGQWPKCG EIDIMESLGNDTTTSYSTIH YGEPHAEQQGTIVKEGAER FSAKFHEYSVEWEPGEMR	bacterium AB4001]
		CBM4	126	5.00E-06	21.5	3.00E-08	1	49	0.38	222	275	0.19	FYTDGELVLTVNDWFTAV QGEDDKPYPAPFNQPFFVQ MNLAVGGNWAREILMLST DFSKAEFEIDYVRVYQKPS YDTNVKKPEKQYGKADAT GNFIRNGNFKKAEKLDDD VDWKFLLFNGGVGAAEIK NGEIVITSSNCGTEEYSVQ	
RMSec_ 24462541	194	GH16	189	3.60E-16	54.6	2.2E-18	100	189	0.46	1	121	0.63	ESLGNDTTKSYSTIHYGEP HAEQQGTIVKEGADSFSAK FHEYSVEWEPGEMRFYTD GELVLTVNDWFTAVQGRR QALSCTFQPTILRSDETSQS	laminarinase [Ruminococcaceae bacterium AB4001]
		CBM4	126	0.00017	16.5	1.00E-06	1	44	0.34	145	193	0.25	EATGPGNPDSTTDFSKAEF EIDYVRVYQKPSYDTNVK KPEKQYGQADSTGNFIRNG DFKKTEKLDDDVDWKFLL FNGGVGAAEIKDGEIVITSS NCGTEE	
RMSec_ 23376741	126	GH25	177	1.10E-08	29.9	6.60E-11	84	149	0.37	1	67	0.52	LPLVVDVEDWSNDEQIKD ERTQQHLDAMLDNLRSRG HKVMIYTNGDGYKKYIKN	glycosyl hydrolase [Rhizobium sp. OR191]
		dockerin	21	8.70E-05	16.7	5.20E-07	1	16	0.71	102	117	0.12	GQININLWLCSFRQPDRTIK IRRAMKRLLSLAVFISCVL	

Query ID [a]	Query length (aa)	Target (dbCAN HMM)	Target length (aa)	iE-value	Bit score	cE-value	Target start	Target end	Target covered fraction	Query start	Query end	Query covered fraction	Query sequence	Best DELTA BLASTP hit [b]
													VASSKIIGDVNGDEKISAA DVATLVSYLIDDNI	
RMSec_ 24393731	202	dockerin	21	6.90E-06	20.2	8.30E-08	1	21	0.95	27	47	0.10	VNLDDVKLLRDYLTTEKE ELAVPANGDLDGDNVITAI DLTLLKRGILEGKYSSGDG PAADETAMEFVKHIKLGW NLGNTFDAQIKGAYQSPQ	Endoglucanase [Ruminococcus champanellensis]
		GH5	275	3.90E-20	66.9	4.60E-22	10	107	0.35	82	202	0.59	QAETAWGNPQTSKAMIDA IKAAGFNTVRVPVSWGEK MNSNYEIDTAWMNRVQE VVDYVIDNDMYCILNIHHD NGYESNGQFVAPSCPYFYP TSAHYEHSEKFVTAVWTQ V	

Abbreviations: MS, rumen microbial plant-adherent metasecretome dataset; MG, rumen microbial plant-adherent metagenome dataset.; RMSec, rumen metasecretome; aa, amino acid residues; dbCAN HMM, dbCAN database CAZyme-specific Hidden Markov Model; iE-value, independent E-value; cE-value, conditional E-value. Putative proteins in the metasecretome dataset (clustered at 100% sequence identity) with multiple, non-overlapping hits to CAZyme-specific HMMs (E-value <1e-05 for an alignment length >80 aa and an E-value <1e-03 otherwise and query coverage > 30%) were inspected for 'start-end' position consistency between the HMM and the query protein and hit 'strength' for each domain, as described in section 2.2.6.2.3.

[a] Query IDs were randomly assigned to putative RMSec proteins predicted in the IMG/M pipeline. [b] Annotation based on best DELTA-BLAST hit in the NCBI nr protein database.

A.

```
>RMSec_22522671
                    RDGKLVLKAIKTQKDGKDYYTSGKVTGQNKTDFQYGKVVVSAKVPEGQGLWPAIWMMPKDESFYGQWPKCGEIDIMESLGNDTTTSYSTIHYGEPHAEQQ
                    GTIVKEGAERFSAKFHEYSVEWEPGEMRFYTDGELVLTVNDWFTAVQGEDDKPYPAPFNQPFFVQMNLAVGGNWAREILMLSTDFSKAEFEIDYVRVYQK
                    PSYDTNVKKPEKQYGKADATGNFIRNGNFKKAEKLDDDVDWKFLLFNGGVGAAEIKNGEIVITSSNCGTEEYSVQ
```

B.

```
  == domain 1    score: 163.3 bits;  conditional E-value: 1e-51
        GH16.hmm  36  dgnLvlratresta.k..ytsgriesk..fsfkygrvearaklpkgsglwpafwllgdd......wpasgEiDiiEvvgn.pnkvhqtlhytnrg 118
                      dg+Lvl+a +++++ k  ytsg+++ +  ++f+yg+v ++ak+p+g+glwpa+w++++d      wp++gEiDi+E++gn +++ +++t+hy+++
  RMSec_22522671   2  DGKLVLKAIKTQKDgKdyYTSGKVTGQnkTDFQYGKVVVSAKVPEGQGLWPAIWMMPKDesfygqWPKCGEIDIMESLGNdTTTSYSTIHYGEPH 96
                      899***9999998866999******97677*****************************99*******************747777******9744 PP

        GH16.hmm 119  ts.ga.....saeess.dasddfhtYgveWtpdeitwyvDgelvrtvt.................pfdqpfylilnlavgg...........kgp. 178
                      ++ ++     +    + ++s++fh Y+veW+p+e+++y Dgelv tv+                 pf+qpf++ +nlavgg           +++
  RMSec_22522671  97  AEqQGtivkeG----AeRFSAKFHEYSVEWEPGEMRFYTDGELVLTVNdwftavqgeddkpypaPFNQPFFVQMNLAVGGnwareilmlstDFSk 187
                      43321344321....247999*****************************99999***********999**************999988877543332 PP

        GH16.hmm 179  aemevdyvrvy 189
                      ae+e+dyvrvy
  RMSec_22522671 188  AEFEIDYVRVY 198
                      46********9 PP


  == domain 2    score: 21.5 bits;  conditional E-value: 3e-08
        CBM4.hmm   1  nlikngdFdeg.....lagWtlyessgakatfsvedgelkvtitnggenrwdvq 49
                      n+i+ng+F++         W +   +g+ + +++++ge+ +t +n g+++++vq
  RMSec_22522671 222  NFIRNGNFKKAeklddDVDWKFLLFNGGVGAAEIKNGEIVITSSNCGTEEYSVQ 275
                      89********999996667**********************************9 PP
```

C.

```
RMSec_22522671               rDGKLVLKAIKTQKDgKdyYTSGKVTGQnkTDFQYGKVVVSAKVPEGQGLWPAIWMMPKDesfygqWPKCGEIDIMESLGNdTTTSYSTIHYGEPHAEqQ
#=GR RMSec_22522671 PP       5899***9999998866999******97677*****************************99*******************747777******97444332
#=GC PP_cons                 .899***99999988.6..9******97..7*****************************9......**************7.7777******974443.2
#=GC RF                      .xxxxxxxxxxxxxx.x..xxxxxxxxx..xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx......xxxxxxxxxxxxxx.xxxxxxxxxxxxxx.x

RMSec_22522671               GtivkeGAeRFSAKFHEYSVEWEPGEMRFYTDGELVLTVNdwftavqgeddkpypaPFNQPFFVQMNLAVGGnwareilmlstDFSkAEFEIDYVRVYqk
#=GR RMSec_22522671 PP       1344321247999*************************99999***********999**************99998887754333246********9**
#=GC PP_cons                 1.....12.7999*************************9.............99**************...........333.46********9..
#=GC RF                      x.....xx.xxxxxxxxxxxxxxxxxxxxxxxxxxxxx.............xxxxxxxxxxxxxxx...........xxx.xxxxxxxxxx..

RMSec_22522671               psydtnvkkpekqygkadatgnfirngnfkkaekldddvdwkfllfnggvgaaeikngeivitssncgteeysvq
#=GR RMSec_22522671 PP       **************************************************************************
#=GC PP_cons                 ..........................................................................
#=GC RF                      ..........................................................................
```

```
RMSec_22522671                rdgklvlkaiktqkdgkdyytsgkvtgqnktdfqygkvvvsakvpegqglwpaiwmmpkdesfygqwpkcgeidimeslgndtttsystihygephaeqq
#=GR RMSec_22522671 PP        ****************************************************************************************************
#=GC PP_cons                  ....................................................................................................
#=GC RF                       ....................................................................................................

RMSec_22522671                gtivkegaerfsakfheysvewepgemrfytdgelvltvndwftavqgeddkpypapfnqpffvqmnlavggnwareilmlstdfskaefeidyvrvyqk
#=GR RMSec_22522671 PP        ****************************************************************************************************
#=GC PP_cons                  ....................................................................................................
#=GC RF                       ....................................................................................................

RMSec_22522671                psydtnvkkpekqygkadatgNFIRNGNFKKAeklddDVDWKFLLFNGGVGAAEIKNGEIVITSSNCGTEEYSVQ
#=GR RMSec_22522671 PP        ********************89********999996667*********************************9
#=GC PP_cons                  ......................89********9.....667*********************************9
#=GC RF                       ......................xxxxxxxxxxx.....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Figure A2.2. Example alignment of putative multi-modular CAZyme and corresponding HMMs.

A. Sequence of a putative metasecretome dataset protein RMSec_22522671 with predicted multi-modular CAZyme organisation. Putative protein regions with hits to two homologous domains: GH16 (amino acid residues 2-198) and CBM4 (amino acid residues 222-275) are highlighted in yellow and green, respectively. B. Alignment between homologous regions of the RMSec_22522671 sequence (query) and two CAZyme-specific HMMs (targets), produced by hmmscan, is represented in a Stockholm format. Target, consensus and query sequences are represented in the first three lines of an alignment, while the bottom line contains posterior probability (PP), indicating expected accuracy for each aligned residue. PP scores 0-9 represent accuracy intervals of 0-5%; 5-15%; 15-25%; 25-35%; 35-45%; 45-55%; 55-65%; 65-75%; 75-85% and 85-95%, respectively, while * represent accuracy interval of 95-100%. Two different notations are used for the "match" and the "insert" columns of an alignment. In a "match" column of an alignment, residues are represented in upper case and a '-' character represents gaps (which are treated as deletions relative to consensus). In an "insert" column of an alignment, residues are represented in lower case, and a '.' character is used as padding to accommodate insertions. A '+" character represents conservative substitutions in a consensus sequence. C. Alignment between full length query RMSec_22522671 sequence and each of the two target HMMs, produced using hmmalign, is represented in a Stockholm format. Query sequence, query PP (GR PP) and consensus PP (GC PP_cons), calculated as the arithmetic mean of the per-residue PP in each column, are represented in the first three lines of an alignment. Bottom line of an alignment contains the reference coordinate (GC RF) annotation, with an x marking each column considered to be consensus.

Appendix 3

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.

Sequence ID [a]	Substrate and panning round [b]	ORF status [c]	Putative protein length (aa)	Putative protein description [d]	Predicted membrane-targeting signal [e]	Motifs conferring possible competitive advantage and/or substrate-unrelated binding [f]
RAC12 [g]	RAC 4th	Background (Out of frame)				
RAC30	RAC 4th	Background (≤24 aa)				None
RAC53 [g]	RAC 4th	Background (Out of frame)				
RAC65	RAC 4th	In frame	179	Conserved hypothetical protein [*Bacteroides* sp. CAG:770]	No ss detected	Propagation advantage
RAC111 [g]	RAC 4th	Background (Out of frame)				
RAC124 [g]	RAC 4th	Background (Out of frame)				
RAC150	RAC 4th	In frame	29	Hypothetical protein	No ss detected	None
RAC162	RAC 4th	Background (Out of frame)				
RAC183	RAC 4th	Background (≤24 aa)				Propagation advantage and suspected binder to immunoglobulin Fc region
RAC200 [g]	RAC 4th	Background (Out of frame)				
AXYL221	AXYL 4th	In frame	101	Hypothetical protein	N-terminal TMH	None
AXYL222	AXYL 4th	In frame	264	Conserved hypothetical protein [*Eggerthella* sp. CAG:298]	Possible non-classical secretion	Propagation advantage
AXYL223	AXYL 4th	In frame	112	Hypothetical protein	Type I ss	None
AXYL224	AXYL 4th	Background (Out of frame)				
AXYL225	AXYL 4th	In frame	98	Concanavalin A-like lectin/glucanase [*Prevotella* sp. CAG:474]	Possible non-classical secretion	None
AXYL226	AXYL 4th	Background (Out of frame)				
AXYL227	AXYL 4th	Background (Out of frame)				
AXYL230	AXYL 4th	Background (Out of frame)				
AXYL231	AXYL 4th	Background (Out of frame)				
AXYL232	AXYL 4th	In frame	77	Conserved hypothetical [*Bacteroides* sp. CAG:545]	Type I ss	Propagation advantage
AXYL233 [h]	AXYL 4th	Background (≤24 aa)				Propagation advantage
AXYL235	AXYL 4th	Background (Out of frame)				
AXYL236	AXYL 4th	In frame	127	Preprotein translocase subunit SecA [Clostridiales genom sp. BVAB3 str. UPII9-5]	No ss detected	None

Sequence ID [a]	Substrate and panning round [b]	ORF status [c]	Putative protein length (aa)	Putative protein description [d]	Predicted membrane-targeting signal [e]	Motifs conferring possible competitive advantage and/or substrate-unrelated binding [f]
AXYL237	AXYL 4th	Background (Out of frame)				
AXYL238	AXYL 4th	Background (≤24 aa)				Propagation advantage and suspected plastic binder
AXYL239	AXYL 4th	Background (Out of frame)				
AXYL240 [h]	AXYL 4th	Background (≤24 aa)				Propagation advantage
AXYL244	AXYL 4th	Background (≤24 aa)				None
RAC261	RAC 1st	In frame	400	Diguanylate cyclase (GGDEF) domain protein [*Clostridium* sp. CAG:127]	No ss detected	Propagation advantage
RAC262	RAC 1st	In frame	63	Conserved hypothetical protein [*Clostridium* sp. CAG:352]	No ss detected	Propagation advantage
RAC263	RAC 1st	In frame	47	Oxidoreductase [*Prevotella ruminicola* 23]	No ss detected	Propagation advantage
RAC295	RAC 1st	Background (Out of frame)				
RAC298	RAC 1st	Background (≤24 aa)				None
RAC302	RAC 1st	Background (≤24 aa)				None
RAC328	RAC 1st	Background (≤24 aa)				None
AXYL342	AXYL 1st	In frame	216	Serine/threonine protein kinase [*Veillonella* sp. oral taxon 780]	Possible non-classical secretion	Propagation advantage and suspected binder to unrelated antibodies
AXYL347	AXYL 1st	In frame	295	Conserved hypothetical protein [*Prevotella saccharolytica*]	Possible non-classical secretion	Propagation advantage and suspected plastic binder
AXYL358	AXYL 1st	Background (≤24 aa)				None
AXYL366	AXYL 1st	Out of frame				
AXYL379	AXYL 1st	Background (≤24 aa)				Propagation advantage and suspected binder to immunoglobulin Fc region

Abbreviations: aa, amino acid residues; RAC, regenerated amorphous cellulose; AXYL, wheat arabinoxylan, 1st, first round of panning; 4th, fourth round of panning. [a] IDs of the 40 sequenced amplicons correspond to IDs of 40 bacterial colony PCR amplicons selected for sequencing (marked with red arrows in Figure 4.2). [b] Substrate and panning round from which 40 sequenced metagenomic phage display library clones were selected by affinity screening. [c] ORF status: In frame, ORFs encoding putative proteins >24 aa in frame with vector-encoded pIII; Background (≤24 aa), ORFs encoding putative proteins and peptides ≤24 aa in frame with vector-encoded pIII; Background (out of frame), 'background' inserts containing ORFs encoding putative proteins out of frame with vector-encoded pIII. [d] Putative protein description based on best BLASTP hit against NCBI nr protein database is provided only for the best BLASTP hits with an E-value of <1e-05 and a query coverage of >30%. Putative proteins with best BLASTP hit below this threshold are described as hypothetical. [e] Membrane-targeting signals predicted for putative proteins in frame with pIII and >24 aa: ss, signal sequence; Type I ss, classical ss; TMH, transmembrane α-helix; Possible non-classical secretion, secretion *via* non-classical pathways indicated by SignalP 2.0 score >0.5. [f] Short overlapping regions (40 aa long with 10 aa

long overlapping sequence) of all putative proteins in frame with pIII were scanned for motifs conferring possible competitive advantages and/or substrate-unrelated binding using the SAROTUP database [408]. [g,h] Multiple inserts containing single distinct ORF.

References

1. **Statistics New Zealand Agricultural Production Statistics** [**http://www.stats.govt.nz/browse_for_stats/industry_sectors/agriculture-horticulture-forestry/AgriculturalProduction_final_HOTPJun12final.aspx**]
2. Waghorn GB, JL Kolver, ES: **Principles of feeding value.** In *Pasture and Supplements for Grazing Animals*. pp. 35-60: NZSAP; 2007:35-60.
3. Juturu V, Wu JC: **Microbial xylanases: engineering, production and industrial applications.** *Biotech Adv* 2012, **30:**1219-1227.
4. Kuhad RC, Gupta R, Singh A: **Microbial cellulases and their industrial applications.** *Enzyme Res* 2011, **2011:**1-10.
5. Himmel ME, Xu Q, Luo Y, Ding S-Y, Lamed R, Bayer EA: **Microbial enzyme systems for biomass conversion: emerging paradigms.** *Biofuels* 2010, **1:**323-341.
6. Vazana Y, Moraïs S, Barak Y, Lamed R, Bayer EA: **Designer cellulosomes for enhanced hydrolysis of cellulosic substrates.** *Methods Enzymol* 2012, **510:**429-452.
7. Kim M, Morrison M, Yu Z: **Status of the phylogenetic diversity census of ruminal microbiomes.** *FEMS Microbiol Ecol* 2011, **76:**49-63.
8. Edwards J, McEwan N, Travis A, John Wallace R: **16S rDNA library-based analysis of ruminal bacterial diversity.** *Antonie van Leeuwenhoek* 2004, **86:**263-281.
9. Pers-Kamczyc E, Zmora P, Cieślak A, Szumacher-Strabel M: **Development of nucleic acid based techniques and possibilities of their application to rumen microbial ecology research.** *J Anim Feed Sci* 2011, **20:**315-337.
10. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms.** *Microbiol Mol Biol Rev* 2004, **68:**669-685.
11. Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC: **Comparative metagenomics of microbial communities.** *Science* 2005, **308:**554-557.
12. Berg Miller ME, Antonopoulos DA, Rincon MT, Band M, Bari A, Akraiko T, Hernandez A, Thimmapuram J, Henrissat B, Coutinho PM: **Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of *Ruminococcus flavefaciens* FD-1.** *PLOS ONE* 2009, **4:**e6650.
13. Purushe J, Fouts DE, Morrison M, White BA, Mackie RI, Coutinho PM, Henrissat B, Nelson KE: **Comparative genome analysis of *Prevotella ruminicola* and *Prevotella bryantii*: insights into their environmental niche.** *Microb Ecol* 2010, **60:**721-729.
14. Kelly W, Leahy S, Altermann E, Yeoman C, Dunne J, Kong Z, Pacheco D, Li D, Noel S, Moon C, et al: **The glycobiome of the rumen bacterium *Butyrivibrio proteoclasticus* B316(T) highlights adaptation to a polysaccharide-rich environment.** *PLOS ONE* 2010, **5**.
15. Suen G, Weimer PJ, Stevenson DM, Aylward FO, Boyum J, Deneke J, Drinkwater C, Ivanova NN, Mikhailova N, Chertkov O: **The complete genome sequence of *Fibrobacter succinogenes* S85 reveals a cellulolytic and metabolic specialist.** *PLOS ONE* 2011, **6:**e18814.
16. Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560-565.

17. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, Cantarel BL, Coutinho PM, Henrissat B, Leclerc M, et al: **Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes.** *Genome Res* 2010, **20:**1605-1612.
18. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA: **Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis.** *Nat Rev Microbiol* 2008, **6:**121-131.
19. Brulc JM, Antonopoulos DA, Berg Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P: **Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases.** *Proc Natl Acad Sci U S A* 2009, **106:**1948-1953.
20. Pope P, Denman S, Jones M, Tringe S, Barry K, Malfatti S, McHardy A, Cheng J-F, Hugenholtz P, McSweeney C: **Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores.** *Proc Natl Acad Sci U S A* 2010, **107:**14793-14798.
21. Duan C-J, Feng J-X: **Mining metagenomes for novel cellulase genes.** *Biotechnol Lett* 2010, **32:**1765-1775.
22. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, et al: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331:**463-467.
23. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, Eijsink VG: **Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci.** *PLOS ONE* 2012, **7:**e38571.
24. Dai X, Zhu Y, Luo Y, Song L, Liu D, Liu L, Chen F, Wang M, Li J, Zeng X, et al: **Metagenomic insights into the fibrolytic microbiome in yak rumen.** *PLOS ONE* 2012, **7:**e40430.
25. Weimann A, Trukhina Y, Pope PB, Konietzny SG, McHardy AC: ***De novo* prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta) genomes.** *Biotechnol Biofuels* 2013, **6:**24.
26. Morgavi DP, Kelly WJ, Janssen PH, Attwood GT: **Rumen microbial (meta)genomics and its application to ruminant production.** *Animal* 2012, **7:**184-201.
27. Kudva R, Denks K, Kuhn P, Vogt A, Müller M, Koch H-G: **Protein translocation across the inner membrane of Gram-negative bacteria: the Sec and Tat dependent protein transport pathways.** *Res Microbiol* 2013, **164:**505-534.
28. Wallin E, Heijne GV: **Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.** *Protein Sci* 1998, **7:**1029-1038.
29. Hackmann T, Spain J: **Ruminant ecology and evolution: perspectives useful to ruminant livestock research and production.** *J Dairy Sci* 2010, **93:**1320-1334.
30. Clauss M, Hume I, Hummel J: **Evolutionary adaptations of ruminants and their potential relevance for modern production systems.** *Animal* 2010, **4:**979.
31. Hungate R: **The rumen microbial ecosystem.** *Annu Rev Ecol Evol Syst* 1975, **6:**39-66.
32. Kingston-Smith AH, Edwards JE, Huws SA, Kim EJ, Abberton M: **Plant-based strategies towards minimising 'livestock's long shadow'.** *Proc Nutr Soc* 2010, **69:**613-620.
33. Dale H, Stewart R, Brody S: **Rumen temperature. Temperature gradients during feeding and fasting.** *Cornell Vet* 1954, **44:**368.
34. Kolver ES, de Veth MJ: **Prediction of ruminal pH from pasture-based diets.** *J Dairy Sci* 2002, **85:**1255-1266.

35. Russell JB, Rychlik JL: **Factors that alter rumen microbial ecology.** *Science* 2001, **292:**1119-1122.
36. Van Houtert M: **The production and metabolism of volatile fatty acids by ruminants fed roughages: a review.** *Anim Feed Sci Technol* 1993, **43:**189-225.
37. Stewart C, Flint H, Bryant M: **The rumen bacteria.** In *The rumen microbial ecosystem.* Edited by Hobson PN, Stewart CS: Springer 1997: 10-72.
38. Williams AG, Coleman G: **The rumen protozoa.** In ***The Rumen Microbial Ecosystem***. Edited by Hobson PN, Stewart CS: Springer Netherlands; 1997: 73-139
39. Orpin C, Joblin K: **The rumen anaerobic fungi.** In *The rumen microbial ecosystem.* 2nd edition. Edited by Hobson PN, Stewart CS: Springer Netherlands; 1997: 140-195.
40. Sirohi SK, Singh N, Dagar SS, Puniya AK: **Molecular tools for deciphering the microbial community structure and diversity in rumen ecosystem.** *Appl Microbiol Biotechnol* 2012, **95:**1135-1154.
41. Bonhomme A: **Rumen ciliates: their metabolism and relationships with bacteria and their hosts.** *Anim Feed Sci Technol* 1990, **30:**203-266.
42. Pitta DW, Pinchak WE, Dowd SE, Osterstock J, Gontcharova V, Youn E, Dorton K, Yoon I, Min BR, Fulford J: **Rumen bacterial diversity dynamics associated with changing from bermudagrass hay to grazed winter wheat diets.** *Microb Ecol* 2010, **59:**511-522.
43. Devillard E, Bera-Maillet C, Flint H, Scott K, Newbold C, Wallace R, Jouany J, Forano E: **Characterization of XYN10B, a modular xylanase from the ruminal protozoan *Polyplastron multivesiculatum*, with a family 22 carbohydrate-binding module that binds to cellulose.** *Biochem J* 2003, **373:**495-503.
44. Béra-Maillet C, Devillard E, Cezette M, Jouany JP, Forano E: **Xylanases and carboxymethylcellulases of the rumen protozoa *Polyplastron multivesiculatum*, *Eudiplodinium maggii* and *Entodinium* sp.** *FEMS Microbiol Lett* 2005, **244:**149-156.
45. Rezaeian M, Beakes GW, Parker DS: **Distribution and estimation of anaerobic zoosporic fungi along the digestive tracts of sheep.** *Mycol Res* 2004, **108:**1227-1233.
46. Kittelmann S, Naylor GE, Koolaard JP, Janssen PH: **A proposed taxonomy of anaerobic fungi (Class Neocallimastigomycetes) suitable for large-scale sequence-based community structure analysis.** *PLoS ONE* 2012, **7:**e36866.
47. Akin DE, Borneman WS: **Role of rumen fungi in fiber degradation.** *J Dairy Sci* 1990, **73:**3023-3032.
48. Bootten T, Joblin K, McArdle B, Harris P: **Degradation of lignified secondary cell walls of lucerne (*Medicago sativa* L.) by rumen fungi growing in methanogenic co-culture.** *J Appl Microbiol* 2011, **111:**1086-1096.
49. Joblin K, Naylor G, Odongo N, Garcia M, Viljoen G: **Ruminal fungi for increasing forage intake and animal productivity.** In *FAO/IAEA International Symposium on Sustainable Improvement of Animal Production and Health, Vienna, Austria, 8-11 June 2009*. Food and Agriculture Organization of the United Nations (FAO); 2010: 129-136.
50. Liggenstoffer AS, Youssef NH, Couger M, Elshahed MS: **Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores.** *ISME J* 2010, **4:**1225-1235.
51. Janssen PH, Kirs M: **Structure of the archaeal community of the rumen.** *Appl Environ Microbiol* 2008, **74:**3619-3625.
52. Lin C, Raskin L, Stahl DA: **Microbial community structure in gastrointestinal tracts of domestic animals: comparative analyses using rRNA-targeted oligonucleotide probes.** *FEMS Microbiol Ecol* 1997, **22:**281-294.

53. Attwood G, Altermann E, Kelly W, Leahy S, Zhang L, Morrison M: **Exploring rumen methanogen genomes to identify targets for methane mitigation strategies.** *Anim Feed Sci Technol* 2011, **166:**65-75.
54. McAllister T, Newbold CJ: **Redirecting rumen fermentation to reduce methanogenesis.** *Anim Prod Sci* 2008, **48:**7-13.
55. Johnson KA, Johnson DE: **Methane emissions from cattle.** *J Animal Sci* 1995, **73:**2483-2492.
56. Kujawa M: **Energy partitioning in steers fed cottonseed hulls and beet pulp.** Ph. D Dissertation, Colorado State University, Fort Collins, Colo, USA, 1994.
57. Swain RA, Nolan JV, Klieve AV: **Natural variability and diurnal fluctuations within the bacteriophage population of the rumen.** *Appl Environ Microbiol* 1996, **62:**994-997.
58. Klieve AV, Swain RA: **Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry.** *Appl Environ Microbiol* 1993, **59:**2299-2303.
59. Miron J, Ben-Ghedalia D, Morrison M: **Invited review: adhesion mechanisms of rumen cellulolytic bacteria.** *J Dairy Sci* 2001, **84:**1294-1309.
60. Koike S, Yoshitani S, Kobayashi Y, Tanaka K: **Phylogenetic analysis of fiber-associated rumen bacterial community and PCR detection of uncultured bacteria.** *FEMS Microbiol Lett* 2003, **229:**23-30.
61. Craig WM, Broderick GA, Ricker DB: **Quantitation of microorganisms associated with the particulate phase of ruminal ingesta.** *J Nutr* 1987, **117:**56-62.
62. Larue R, Yu Z, Parisi VA, Egan AR, Morrison M: **Novel microbial diversity adherent to plant biomass in the herbivore gastrointestinal tract, as revealed by ribosomal intergenic spacer analysis and *rrs* gene sequencing.** *Environ Microbiol* 2005, **7:**530-543.
63. McAllister T, Bae H, Jones G, Cheng K: **Microbial attachment and feed digestion in the rumen.** *J Anim Sci* 1994, **72:**3004.
64. Bayer EA, Shoham Y, Lamed R: **Lignocellulose-decomposing bacteria and their enzyme systems.** In *The Prokaryotes.* Springer; 2013: 215-266.
65. Callaway T, Dowd S, Edrington T, Anderson R, Krueger N, Bauer N, Kononoff P, Nisbet D: **Evaluation of bacterial diversity in the rumen and feces of cattle fed different levels of dried distillers grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing.** *J Animal Sci* 2010, **88:**3977-3983.
66. Durso LM, Harhay GP, Smith TP, Bono JL, DeSantis TZ, Harhay DM, Andersen GL, Keen JE, Laegreid WW, Clawson ML: **Animal-to-animal variation in fecal microbial diversity among beef cattle.** *Appl Environ Microbiol* 2010, **76:**4858-4862.
67. Shanks OC, Kelty CA, Archibeque S, Jenkins M, Newton RJ, McLellan SL, Huse SM, Sogin ML: **Community structures of fecal bacteria in cattle from different animal feeding operations.** *Appl Environ Microbiol* 2011, **77:**2992-3001.
68. Tajima K, Arai S, Ogata K, Nagamine T, Matsui H, Nakamura M, Aminov RI, Benno Y: **Rumen bacterial community transition during adaptation to high-grain diet.** *Anaerobe* 2000, **6:**273-284.
69. Kim M, Morrison M, Yu Z: **Phylogenetic diversity of bacterial communities in bovine rumen as affected by diets and microenvironments.** *Folia Microbiol* 2011, **56:**453-458.

70. Kong Y, Teather R, Forster R: **Composition, spatial distribution, and diversity of the bacterial communities in the rumen of cows fed different forages.** *FEMS Microbiol Ecol* 2010, **74:**612-622.
71. Li RW, Wu S, Li W, Li C: **Perturbation dynamics of the rumen microbiota in response to exogenous butyrate.** *PLOS ONE* 2012, **7:**e29392.
72. Edwards J, Huws S, Kim E, Kingston Smith A: **Characterization of the dynamics of initial bacterial colonization of nonconserved forage in the bovine rumen.** *FEMS Microbiol Ecol* 2007, **62:**323-335.
73. Fouts DE, Szpakowski S, Purushe J, Torralba M, Waterman RC, MacNeil MD, Alexander LJ, Nelson KE: **Next generation sequencing to define prokaryotic and fungal diversity in the bovine rumen.** *PLOS ONE* 2012, **7:**e48289.
74. Krause D, Plaizier J, Attwood G: **The advances in dietary protein and carbohydrate nutrition and the microbial diversity and genomics of the rumen bacteria.** In *Proceedings of the 7th International Symposium on the Nutrition of Herbivores*. 2007.
75. Jami E, Mizrahi I: **Composition and similarity of bovine rumen microbiota across individual animals.** *PLOS ONE* 2012, **7:**e33306.
76. Sundset MA, Præsteng KE, Cann IK, Mathiesen SD, Mackie RI: **Novel rumen bacterial diversity in two geographically separated sub-species of reindeer.** *Microb Ecol* 2007, **54:**424-438.
77. Tajima K, Aminov RI, Nagamine T, Ogata K, Nakamura M, Matsui H, Benno Y: **Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries.** *FEMS Microbiol Ecol* 1999, **29:**159-169.
78. Sitao Wu RLB, Weizhong Li, Congjun Li, Erin E. Connor, and Robert W. Li: **The bacterial community composition of the bovine rumen detected using pyrosequencing of 16S rRNA genes.** *Metagenomics* 2012, **1:**1-11.
79. de Menezes AB, Lewis E, O'Donovan M, O'Neill BF, Clipson N, Doyle EM: **Microbiome analysis of dairy cows fed pasture or total mixed ration diets.** *FEMS Microbiol Ecol* 2011, **78:**256-265.
80. Stevenson DM, Weimer PJ: **Dominance of *Prevotella* and low abundance of classical ruminal bacterial species in the bovine rumen revealed by relative quantification real-time PCR.** *Appl Microbiol Biotechnol* 2007, **75:**165-174.
81. Bekele AZ, Koike S, Kobayashi Y: **Genetic diversity and diet specificity of ruminal *Prevotella* revealed by 16S rRNA gene-based analysis.** *FEMS Microbiol Lett* 2010, **305:**49-57.
82. Kim M, Yu Z: **Quantitative comparisons of select cultured and uncultured microbial populations in the rumen of cattle fed different diets.** *J Anim Sci Biotechnol* 2012, **3:**28.
83. Suen G, Stevenson DM, Bruce DC, Chertkov O, Copeland A, Cheng J-F, Detter C, Detter JC, Goodwin LA, Han CS: **Complete genome of the cellulolytic ruminal bacterium *Ruminococcus albus* 7.** *J Bacteriol* 2011, **193:**5574-5575.
84. Reilly K, Attwood G: **Detection of *Clostridium proteoclasticum* and closely related strains in the rumen by competitive PCR.** *Appl Environ Microbiol* 1998, **64:**907-913.
85. Paillard D, McKain N, Rincon MT, Shingfield KJ, Givens DI, Wallace RJ: **Quantification of ruminal *Clostridium proteoclasticum* by real-time PCR using a molecular beacon approach.** *J Appl Microbiol* 2007, **103:**1251-1261.
86. Krause DO, Denman SE, Mackie RI, Morrison M, Rae AL, Attwood GT, McSweeney CS: **Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics.** *FEMS Microbiol Rev* 2003, **27:**663-693.

87. Van Soest P: *Nutritional ecology of the ruminant: ruminant metabolism, nutritional strategies, the cellulolytic fermentation and the chemistry of forages and plant fibers.* O & B Books 1982.
88. Russell JB, Muck RE, Weimer PJ: **Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen.** *FEMS Microbiol Ecol* 2009, **67:**183-197.
89. Weimer PJ, Russell JB, Muck RE: **Lessons from the cow: what the ruminant animal can teach us about consolidated bioprocessing of cellulosic biomass.** *Bioresour Technol* 2009, **100:**5323-5331.
90. Engels F: **Some properties of cell wall layers determining ruminant digestion.** In *Physico-Chemical Characterisation of Plant Residues for Industrial and Feed Use.* Edited by Chesson A, Orskov E: Springer 1989: 80-87.
91. Weimer P: **Why don't ruminal bacteria digest cellulose faster?** *J Dairy Sci* 1996, **79:**1496-1502.
92. Engels FM, Jung H-JG: **Alfalfa stem tissues: impact of lignification and cell length on ruminal degradation of large particles.** *Anim Feed Sci Technol* 2005, **120:**309-321.
93. Mackie RI: **Mutualistic fermentative digestion in the gastrointestinal tract: diversity and evolution.** *Integr Comp Biol* 2002, **42:**319-326.
94. Wilson J, Mertens D: **Cell wall accessibility and cell structure limitations to microbial digestion of forage.** *Crop Sci* 1995, **35:**251-259.
95. Jung H-JG, Samac DA, Sarath G: **Modifying crops to increase cell wall digestibility.** *Plant Sci* 2012, **185:**65-77.
96. Pell A, Schofield P: **Microbial adhesion and degradation of plant cell walls.** In *Forage cell wall structure and digestibility.* Edited by Jung HG, Buxton DR, Hatfield RD, Ralph J: ASA-CSA-SSSA; 1993: 397–423
97. Roger V, Fonty G, Komisarczuk-Bony S, Gouet P: **Effects of physicochemical factors on the adhesion to cellulose avicel of the ruminal bacteria *Ruminococcus flavefaciens* and *Fibrobacter succinogenes* subsp. *succinogenes*.** *Appl Environ Microbiol* 1990, **56:**3081.
98. Jun H-S, Qi M, Gong J, Egbosimba EE, Forsberg CW: **Outer membrane proteins of *Fibrobacter succinogenes* with potential roles in adhesion to cellulose and in cellulose digestion.** *J Bacteriol* 2007, **189:**6806-6815.
99. Gong J, Egbosimba EE, Forsberg CW: **Cellulose-binding proteins of *Fibrobacter succinogenes* and the possible role of a 180-kDa cellulose-binding glycoprotein in adhesion to cellulose.** *Can J Microbiol* 1996, **42:**453-460.
100. Mayorga OL, Huws S, Kim E, Kingston-Smith A, Newbold C, Theodorou M: **Microbial colonisation and subsequent biofilm formation by ruminal microorganisms on fresh perennial ryegrass.** *Microbiol Ecol Health Dis* 2007, **19:**26.
101. Huws S, Mayorga O, Theodorou M, Onime L, Kim E, Cookson A, Newbold C, Kingston-Smith A: **Successional colonization of perennial ryegrass by rumen bacteria.** *Lett Appl Microbiol* 2013, **56:**186-196.
102. Jung H-JG: **Forage digestibility: the intersection of cell wall lignification and plant tissue anatomy.** In *International Advances in Ruminant Nutrition Research in Brazil*. 2011: 137-160.
103. Van Soest Pv, Robertson J, Lewis B: **Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition.** *J Dairy Sci* 1991, **74:**3583-3597.

104. Mertens D: **Impact of NDF content and digestibility on dairy cow performance.** *WCDS Advances in Dairy Technology* 2009, **21:**191-201.
105. Jung H, Allen M: **Characteristics of plant cell walls affecting intake and digestibility of forages by ruminants.** *J Animal Sci* 1995, **73:**2774-2790.
106. Khandeparker R, Numan MT: **Bifunctional xylanases and their potential use in biotechnology.** *J Ind Microbiol Biotechnol* 2008, **35:**635-644.
107. Shallom D, Shoham Y: **Microbial hemicellulases.** *Curr Opin Biotechnol* 2003, **6:**219-228.
108. Cosgrove DJ: **Growth of the plant cell wall.** *Nat Rev Mol Cell Biol* 2005, **6:**850-861.
109. Wilson DB: **Three microbial strategies for plant cell wall degradation.** *Ann N Y Acad Sci* 2008, **1125:**289-297.
110. Hildén L, Johansson G: **Recent developments on cellulases and carbohydrate-binding modules with cellulose affinity.** *Biotechnol Lett* 2004, **26:**1683-1693.
111. Ruel K, Nishiyama Y, Joseleau J-P: **Crystalline and amorphous cellulose in the secondary walls of *Arabidopsis*.** *Plant Sci* 2012, **193:**48-61.
112. Prade R: **Xylanases: from biology to biotecnology** *Biotech Genet Eng Rev* 1995, **13:**100-131.
113. Niño-Medina G, Carvajal-Millán E, Rascon-Chu A, Marquez-Escalante JA, Guerrero V, Salas-Muñoz E: **Feruloylated arabinoxylans and arabinoxylan gels: structure, sources and applications.** *Phytochem Rev* 2010, **9:**111-120.
114. Warren R: **Microbial hydrolysis of polysaccharides.** *Annu Rev Microbiol* 1996, **50:**183-212.
115. O'Neill MA, Ishii T, Albersheim P, Darvill AG: **Rhamnogalacturonan II: structure and function of a borate cross-linked cell wall pectic polysaccharide.** *Annu Rev Plant Biol* 2004, **55:**109-139.
116. Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W: **Lignin biosynthesis and structure.** *Plant Physiol* 2010, **153:**895-905.
117. Waghorn G, Clark D: **Feeding value of pastures for ruminants.** *New Zeal Vet J* 2004, **52:**320-331.
118. Carpita N: **Structure and biogenesis of the cell walls of grasses.** *Annu Rev Plant Physiol Plant Mol Biol* 1996, **47:**445-476.
119. Vogel J: **Unique aspects of the grass cell wall.** *Curr Opin Plant Biol* 2008, **11:**301-307.
120. Grabber JH, Ralph J, Lapierre C, Barrière Y: **Genetic and molecular basis of grass cell-wall degradability. Lignin-cell wall matrix interactions.** *C R Biol* 2004, **327:**455-465.
121. Gill M, Smith P, Wilkinson J: **Mitigating climate change: the role of domestic livestock.** *Animal* 2010, **4:**323-333.
122. Hatfield R, Weimer P: **Degradation characteristics of isolated and *in situ* cell wall lucerne pectic polysaccharides by mixed ruminal microbes.** *J Sci Food Agr* 1995, **69:**185-196.
123. Jung H, Casler M: **Maize stem tissues: impact of development on cell wall degradability.** *Crop Science* 2006, **46:**1801.
124. Jung H, Engels F: **Alfalfa stem tissues.** *Crop Science* 2002, **42:**524-534.
125. Grabber JH: **How do lignin composition, structure, and cross-linking affect degradability? A review of cell wall model studies.** *Crop Science* 2005, **45:**820-831.
126. Blake A, McCartney L, Flint J, Bolam D, Boraston A, Gilbert H, Knox J: **Understanding the biological rationale for the diversity of cellulose-directed**

carbohydrate-binding modules in prokaryotic enzymes. *J Biol Chem* 2006, **281:**29321.

127. 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:**1-14.
128. Fontes CM, Gilbert HJ: **Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates.** *Annu Rev Biochem* 2010, **79:**655-681.
129. Zverlov VV, Schwarz WH: **Bacterial cellulose hydrolysis in anaerobic environmental subsystem - *Clostridium thermocellum* and *Clostridium stercorarium*, thermophilic plant-fiber degraders.** *Ann N Y Acad Sci* 2008, **1125:**298-307.
130. Yang S-J, Kataeva I, Wiegel J, Yin Y, Dam P, Xu Y, Westpheling J, Adams MW: **Classification of 'Anaerocellum thermophilum'strain DSM 6725 as Caldicellulosiruptor bescii sp. nov.** *International journal of systematic and evolutionary microbiology* 2010, **60:**2011-2015.
131. Yang S-J, Kataeva I, Hamilton-Brehm SD, Engle NL, Tschaplinski TJ, Doeppke C, Davis M, Westpheling J, Adams MW: **Efficient degradation of lignocellulosic plant biomass, without pretreatment, by the thermophilic anaerobe *Anaerocellum thermophilum* DSM 6725.** *Appl Environ Microbiol* 2009, **75:**4762-4769.
132. Davies G, Gloster T, Henrissat B: **Recent structural insights into the expanding world of carbohydrate-active enzymes.** *Curr Opin Struct Biol* 2005, **15:**637-645.
133. Moon C, Pacheco D, Kelly W, Leahy S, Li D, Kopecny J, Attwood G: **Reclassification of *Clostridium proteoclasticum* as *Butyrivibrio proteoclasticus* comb. nov., a butyrate-producing ruminal bacterium.** *Int J Syst Evol Microbiol* 2008, **58:**2041.
134. Attwood GT, Reilly K, Patel BKC: ***Clostridium proteoclasticum* sp. nov., a novel proteolytic bacterium from the bovine rumen.** *Int J Syst Bact* 1996, **46:**753-758.
135. Henrissat B, Claeyssens M, Tomme P, Lemesle L, Mornon J: **Cellulase families revealed by hydrophobic cluster analysis.** *Gene* 1989, **81:**83-95.
136. Henrissat B: **A classification of glycosyl hydrolases based on amino acid sequence similarities.** *Biochem J* 1991, **280:**309-316.
137. Henrissat B, Bairoch A: **New families in the classification of glycosyl hydrolases based on amino acid sequence similarities.** *Biochem J* 1993, **293:**781-788.
138. Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics.** *Nucl Acids Res* 2009, **37**.
139. Henrissat B, Davies GJ: **Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics.** *Plant Physiol* 2000, **124:**1515-1519.
140. Cantarel BL, Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. : **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics.** *Nucl Acids Res* 1999, **37:**D233-D238.
141. **Carbohydrate-active enzymes (CAZy) database.** 2013.
142. Tenkanen M, Eyzaguirre J, Isoniemi R, Faulds CB, Biely P: **Comparison of catalytic properties of acetyl xylan esterases from three carbohydrate esterase families.** In *Applications of Enzymes to Lignocellulosics. Volume* 855. Edited by Mansfield SD, Saddler JN: ACS Publications; 2003: 211-229: *ACS Symposium Series*].
143. Yip VL, Withers SG: **Breakdown of oligosaccharides by the process of elimination.** *Curr Opin Chem Biol* 2006, **10:**147-155.

144. Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho P, Henrissat B: **A hierarchical classification of polysaccharide lyases for glycogenomics.** *Biochem J* 2010, **432:**437-444.
145. Garron M-L, Cygler M: **Structural and mechanistic classification of uronic acid-containing polysaccharide lyases.** *Glycobiology* 2010, **20:**1547-1573.
146. Dimarogona M, Topakas E, Christakopoulos P: **Cellulose degradation by oxidative enzymes.** *Comput Struct Biotechnol J* 2012, **2**.
147. Juturu V, Wu JC: **Insight into microbial hemicellulases other than xylanases: a review.** *J Chem Technol Biotechnol* 2012, **88:**353-363.
148. Juy M, Amrt A, Alzari P, Poljak R, Claeyssens M, Béguin P, Aubert J: **Three-dimensional structure of a thermostable bacterial cellulase.** *Nature* 1992, **357:**89-91.
149. Rouvinen J, Bergfors T, Teeri T, Knowles J, Jones T: **Three-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*.** *Science* 1990, **249:**380.
150. Sakon J, Irwin D, Wilson D, Karplus P: **Structure and mechanism of endo/exocellulase E4 from *Thermomonospora fusca*.** *Nat Struct Mol Biol* 1997, **4:**810-818.
151. Numan MT, Bhosle NB: **α-L-arabinofuranosidases: the potential applications in biotechnology.** *J Ind Microbiol Biotechnol* 2006, **33:**247-260.
152. Collins T, Gerday C, Feller G: **Xylanases, xylanase families and extremophilic xylanases.** *FEMS Microbiol Rev* 2005, **29:**3-23.
153. Shoseyov O, Shani Z, Levy I: **Carbohydrate binding modules: biochemical properties and novel applications.** *Microbiol Mol Biol Rev* 2006, **70:**283.
154. Simpson P, Xie H, Bolam D, Gilbert H, Williamson M: **The structural basis for the ligand specificity of family 2 carbohydrate-binding modules.** *J Biol Chem* 2000, **275:**41137.
155. Guillén D, Sánchez S, Rodríguez-Sanoja R: **Carbohydrate-binding domains: multiplicity of biological roles.** *Appl Microbiol Biotechnol* 2010, **85:**1241-1249.
156. Henshaw J, Bolam D, Pires V, Czjzek M, Henrissat B, Ferreira L, Fontes C, Gilbert H: **The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities.** *J Biol Chem* 2004, **279:**21552-21559.
157. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Carbohydrate-binding modules: fine-tuning polysaccharide recognition.** *Biochem J* 2004, **382:**769-781.
158. Hashimoto H: **Recent structural studies of carbohydrate-binding modules.** *Cell Mol Life Sci* 2006, **63:**2954-2967.
159. Arantes V, Saddler JN: **Access to cellulose limits the efficiency of enzymatic hydrolysis: the role of amorphogenesis.** *Biotechnol Biofuels* 2010, **3**.
160. Pinto R, Moreira, S, Mota M, Gama, M: **Studies on the cellulose-binding domains adsorption to cellulose.** *Langmuir* 2004., **20**.
161. Din N, Gilkes N, Tekant B, Miller R, Warren R, Kilburn D: **Non-hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulase.** *Nat Biotechnol* 1991, **9:**1096-1099.
162. Southall SM, Simpson PJ, Gilbert HJ, Williamson G, Williamson MP: **The starch-binding domain from glucoamylase disrupts the structure of starch.** *FEBS Lett* 1999, **447:**58-60.
163. Vaaje-Kolstad G, Houston DR, Riemen AH, Eijsink VG, van Aalten DM: **Crystal structure and binding properties of the *Serratia marcescens* chitin-binding protein CBP21.** *J Biol Chem* 2005, **280:**11313-11319.

164. Hervé C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP: **Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects.** *Proc Natl Acad Sci U S A* 2010, **107:**15293-15298.
165. Green NM: **Avidin.** *Adv Protein Chem* 1975, **29:**85-133.
166. Jindou S, Soda A, Karita S, Kajino T, Beguin P, Wu JH, Inagaki M, Kimura T, Sakka K, Ohmiya K: **Cohesin-dockerin interactions within and between *Clostridium josui* and *Clostridium thermocellum*: binding selectivity between cognate dockerin and cohesin domains and species specificity.** *J Biol Chem* 2004, **279:**9867-9874.
167. Boraston AB, McLean BW, Chen G, Li A, Warren RAJ, Kilburn DG: **Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase.** *Mol Microbiol* 2002, **43:**187-194.
168. Serizawa T, Iida K, Matsuno H, Kurita K: **Cellulose-binding heptapeptides identified by phage display methods.** *Chem Lett* 2007, **36:**988-989.
169. Bayer EA, Kenig R, Lamed R: **Adherence of *Clostridium thermocellum* to cellulose.** *J Bacteriol* 1983, **156:**818-827.
170. Lamed R, Setter E, Bayer EA: **Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*.** *J Bacteriol* 1983, **156:**828-836.
171. Lamed R, Setter E, Kenig R, Bayer E: **Cellulosome: a discrete cell surface organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-binding and various cellulolytic activities.** In *Biotechnology and Bioengineering Symposium*. 1983.
172. Morag E, Bayer E, Lamed R: **Relationship of cellulosomal and noncellulosomal xylanases of *Clostridium thermocellum* to cellulose-degrading enzymes.** *J Bacteriol* 1990, **172:**6098-6105.
173. Kosugi A, Murashima K, Doi RH: **Xylanase and acetyl xylan esterase activities of XynA, a key subunit of the *Clostridium cellulovorans* cellulosome for xylan degradation.** *Appl Environ Microbiol* 2002, **68:**6399-6402.
174. Tamaru Y, Doi RH: **Pectate lyase A, an enzymatic subunit of the *Clostridium cellulovorans* cellulosome.** *Proc Natl Acad Sci U S A* 2001, **98:**4125-4129.
175. Bayer EA, Lamed R, White BA, Flints HJ: **From cellulosomes to cellulosomics.** *Chem Rec* 2008, **8:**364-377.
176. Yutin N, Galperin MY: **A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia.** *Environ Microbiol* 2013, **15:**2631-2641.
177. Ludwig W, Schleifer K-H, Whitman W: **Revised road map to the phylum Firmicutes.** In *Bergey's Manual® of Systematic Bacteriology.* Edited by Vos P, Garrity G, Jones D, Krieg N, Ludwig W, Rainey F, Schleifer K-H, Whitman W: Springer New York; 2009: 1-13
178. Shoham Y, Lamed R, Bayer EA: **The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides.** *Trends Microbiol* 1999, **7:**275-281.
179. Bayer EA, Morag E, Lamed R: **The cellulosome-a treasure-trove for biotechnology.** *Trends Biotechnol* 1994, **12:**379-386.
180. Wu JD, Newcomb M, Sakka K, Wall J, Harwood C, Demain A: **Cohesin-dockerin interactions and folding.** In *Bioenergy.* Edited by Wall JD, Harwood CS, Demain AL: ASM Press; 2008: 107-113.

181. Bayer EA, Belaich J-P, Shoham Y, Lamed R: **The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides.** *Annu Rev Microbiol* 2004, **58:**521-554.
182. Tokatlidis K, Dhurjati P, Béguin P: **Properties conferred on *Clostridium thermocellum* endoglucanase CelC by grafting the duplicated segment of endoglucanase CelD.** *Protein Eng* 1993, **6:**947-952.
183. Salamitou S, Raynaud O, Lemaire M, Coughlan M, Beguin P, Aubert J-P: **Recognition specificity of the duplicated segments present in *Clostridium thermocellum* endoglucanase CelD and in the cellulosome-integrating protein CipA.** *J Bacteriol* 1994, **176:**2822-2827.
184. Choi SK, Ljungdahl LG: **Structural role of calcium for the organization of the cellulosome of *Clostridium thermocellum*.** *Biochemistry* 1996, **35:**4906-4910.
185. Chauvaux S, Beguin P, Aubert J, Bhat K, Gow L, Wood T, Bairoch A: **Calcium-binding affinity and calcium-enhanced activity of *Clostridium thermocellum* endoglucanase D.** *Biochem J* 1990, **265:**261-265.
186. Shoseyov O, Takagi M, Goldstein MA, Doi RH: **Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A.** *Proc Natl Acad Sci U S A* 1992, **89:**3483-3487.
187. Tavares GA, Béguin P, Alzari PM: **The crystal structure of a type I cohesin domain at 1.7 Å resolution.** *J Mol Biol* 1997, **273:**701-713.
188. Leibovitz E, Beguin P: **A new type of cohesin domain that specifically binds the dockerin domain of the *Clostridium thermocellum* cellulosome-integrating protein CipA.** *J Bacteriol* 1996, **178:**3077-3084.
189. Ding S-Y, Rincon MT, Lamed R, Martin JC, McCrae SI, Aurilia V, Shoham Y, Bayer EA, Flint HJ: **Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*.** *J Bacteriol* 2001, **183:**1945-1953.
190. Alber O, Noach I, Lamed R, Shimon LJ, Bayer EA, Frolow F: **Preliminary X-ray characterization of a novel type of anchoring cohesin from the cellulosome of *Ruminococcus flavefaciens*.** *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2008, **64:**77-80.
191. Noach I, Lamed R, Xu Q, Rosenheck S, Shimon LJ, Bayer EA, Frolow F: **Preliminary X-ray characterization and phasing of a type II cohesin domain from the cellulosome of *Acetivibrio cellulolyticus*.** *Acta Crystallogr D Biol Crystallogr* 2003, **59:**1670-1673.
192. Noach I, Frolow F, Jakoby H, Rosenheck S, Shimon LW, Lamed R, Bayer EA: **Crystal structure of a type-II cohesin module from the *Bacteroides cellulosolvens* cellulosome reveals novel and distinctive secondary structural elements.** *J Mol Biol* 2005, **348:**1-12.
193. Alber O, Noach I, Rincon MT, Flint HJ, Shimon LJ, Lamed R, Frolow F, Bayer EA: **Cohesin diversity revealed by the crystal structure of the anchoring cohesin from *Ruminococcus flavefaciens*.** *Proteins* 2009, **77:**699-709.
194. Pages S, Belaich A, Belaich JP, Morag E, Lamed R, Shoham Y, Bayer EA: **Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain.** *Proteins* 1997, **29:**517-527.
195. Mechaly A, Yaron S, Lamed R, Fierobe H-P, Belaich A, Belaich J-P, Shoham Y, Bayer EA: **Cohesin-dockerin recognition in cellulosome assembly: experiment versus hypothesis.** *Proteins* 2000, **39:**170-177.

196. Sakka K, Sugihara Y, Jindou S, Sakka M, Inagaki M, Sakka K, Kimura T: **Analysis of cohesin–dockerin interactions using mutant dockerin proteins.** *FEMS Microbiol Lett* 2011, **314:**75-80.
197. Sakka K, Kishino Y, Sugihara Y, Jindou S, Sakka M, Inagaki M, Kimura T, Sakka K: **Unusual binding properties of the dockerin module of *Clostridium thermocellum* endoglucanase CelJ (Cel9D-Cel44A).** *FEMS Microbiol Lett* 2009, **300:**249-255.
198. Peer A, Smith SP, Bayer EA, Lamed R, Borovok I: **Noncellulosomal cohesin- and dockerin- like modules in the three domains of life.** *FEMS Microbiol Lett* 2009, **291:**1-16.
199. Salama-Alber O, Gat Y, Lamed R, Shimon LJ, Bayer EA, Frolow F: **Crystallization and preliminary X-ray characterization of a type III cohesin-dockerin complex from the cellulosome system of *Ruminococcus flavefaciens*.** *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2012, **68:**1116-1119.
200. Vodovnik M, Duncan SH, Reid MD, Cantlay L, Turner K, Parkhill J, Lamed R, Yeoman CJ, Miller MEB, White BA: **Expression of cellulosome components and type IV pili within the extracellular proteome of *Ruminococcus flavefaciens* 007.** *PLOS ONE* 2013, **8:**e65333.
201. Brulc JM, Yeoman CJ, Wilson MK, Berg Miller ME, Jeraldo P, Jindou S, Goldenfeld N, Flint HJ, Lamed R, Borovok I: **Cellulosomics, a gene-centric approach to investigating the intraspecific diversity and adaptation of *Ruminococcus flavefaciens* within the rumen.** *PLOS ONE* 2011, **6:**e25329.
202. Ezer A, Matalon E, Jindou S, Borovok I, Atamna N, Yu Z, Morrison M, Bayer EA, Lamed R: **Cell surface enzyme attachment is mediated by family 37 carbohydrate-binding modules, unique to *Ruminococcus albus*.** *J Bacteriol* 2008, **190:**8220-8222.
203. Brumm P, Mead D, Boyum J, Drinkwater C, Gowda K, Stevenson D, Weimer P: **Functional annotation of *Fibrobacter succinogenes* S85 carbohydrate active enzymes.** *Appl Biochem Biotechnol* 2011, **163:**649-657.
204. Rincon MT, Ding S-Y, McCrae SI, Martin JC, Aurilia V, Lamed R, Shoham Y, Bayer EA, Flint HJ: **Novel organization and divergent dockerin specificities in the cellulosome system of *Ruminococcus flavefaciens*.** *J Bacteriol* 2003, **185:**703-713.
205. Rincon MT, Cepeljnik T, Martin JC, Barak Y, Lamed R, Bayer EA, Flint HJ: **A novel cell surface-anchored cellulose-binding protein encoded by the *sca* gene cluster of *Ruminococcus flavefaciens*.** *J Bacteriol* 2007, **189:**4774-4783.
206. Cowan DA: **Microbial genomes-the untapped resource.** *Trends Biotechnol* 2000, **18:**14-16.
207. Cowan D, Meyer Q, Stafford W, Muyanga S, Cameron R, Wittwer P: **Metagenomic gene discovery: past, present and future.** *Trends Biotechnol* 2005, **23:**321-329.
208. Handelsman J, Rondon M, Brady S, Clardy J, Goodman R: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.** *Chem Biol* 1998, **5:**R245-R249.
209. Ekkers DM, Cretoiu MS, Kielak AM, van Elsas JD: **The great screen anomaly-a new frontier in product discovery through functional metagenomics.** *Appl Microbiol Biotechnol* 2012, **93:**1005-1020.
210. Sharma P, Kumari H, Kumar M, Verma M, Kumari K, Malhotra S, Khurana J, Lal R: **From bacterial genomics to metagenomics: concept, tools and recent advances.** *Indian J Microbiol* 2008, **48:**173-194.
211. Prakash T, Taylor TD: **Functional assignment of metagenomic data: challenges and applications.** *Brief Bioinform* 2012, **13:**711-727.

212. Nikolaki S, Tsiamis G: **Microbial Diversity in the Era of Omic Technologies.** *Biomed Res Int* 2013, **2013:**1-15.
213. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ: **High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.** *Nat Rev Microbiol* 2012, **10:**599-606.
214. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms.** *Nat Biotechnol* 2012, **30:**434-439.
215. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC Genomics* 2012, **13:**341-354.
216. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J: **Updating benchtop sequencing performance comparison.** *Nat Biotechnol* 2013, **31:**294-296.
217. Pallen MJ: **Reply to "Updating benchtop sequencing performance comparison".** *Nature biotechnology* 2013, **31:**296-296.
218. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems.** *BioMed Research International* 2012, **2012:**1-11.
219. Schirmer M, D'Amore L, Hall N, Quince C (Eds.): **Error profiles for next generation sequencing technologies**: EMBnet. journal; 2013.
220. Gomez-Alvarez V, Teal TK, Schmidt TM: **Systematic artifacts in metagenomes from complex microbial communities.** *ISME J* 2009, **3:**1314-1317.
221. Niu B, Fu L, Sun S, Li W: **Artificial and natural duplicates in pyrosequencing reads of metagenomic data.** *BMC Bioinformatics* 2010, **11:**187-198.
222. Robasky K, Lewis NE, Church GM: **The role of replicates for error mitigation in next-generation sequencing.** *Nat Rev Genet* 2013**:**56-62.
223. Attwood G, McSweeney C: **Methanogen genomics to discover targets for methane mitigation technologies and options for alternative H_2 utilisation in the rumen.** *Aust J Exp Agric* 2008, **48:**28-37.
224. Allgaier M, Reddy A, Park JI, Ivanova N, D'haeseleer P, Lowry S, Sapra R, Hazen TC, Simmons BA, VanderGheynst JS: **Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community.** *PLOS ONE* 2010, **5:**e8812.
225. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM: **Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome.** *Microbiol Mol Biol Rev* 2000, **64:**515-547.
226. Desvaux M, Hébraud M, Talon R, Henderson IR: **Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue.** *Trends Microbiol* 2009, **17:**139-145.
227. Zhou M, Theunissen D, Wels M, Siezen R: **LAB-Secretome: a genome-scale comparative analysis of the predicted extracellular and surface-associated proteins of Lactic Acid Bacteria.** *BMC Genomics* 2010, **11:**651.
228. Walsh C: **Molecular mechanisms that confer antibacterial drug resistance.** *Nature* 2000, **406:**775-781.
229. Tjalsma H, Antelmann H, Jongbloed JD, Braun PG, Darmon E, Dorenbos R, Dubois J-YF, Westers H, Zanen G, Quax WJ: **Proteomics of protein secretion by *Bacillus subtilis*: separating the "secrets" of the secretome.** *Microbiol Mol Biol Rev* 2004, **68:**207-233.

230. Antelmann H, Tjalsma H, Voigt B, Ohlmeier S, Bron S, van Dijl JM, Hecker M: **A proteomic view on genome-based signal peptide predictions.** *Genome Res* 2001, **11:**1484-1502.
231. Wooldridge K (Ed.). **Bacterial secreted proteins: secretory mechanisms and role in pathogenesis**. UK: Caister Academic Press; 2009.
232. Dalbey RE, Kuhn A: **Protein traffic in Gram-negative bacteria: how exported and secreted proteins find their way.** *FEMS Microbiol Rev* 2012, **36:**1023-1045.
233. Rahman O, Cummings SP, Harrington DJ, Sutcliffe IC: **Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria.** *World J Microb Biot* 2008, **24:**2377-2382.
234. Jankovic D, Collett MA, Lubbers MW, Rakonjac J: **Direct selection and phage display of a Gram-positive secretome.** *Genome Biol* 2007, **8:**R266.
235. Freudl R: **Leaving home ain't easy: protein export systems in Gram-positive bacteria.** *Res Microbiol* 2013.
236. Zuber B, Haenni M, Ribeiro T, Minnig K, Lopes F, Moreillon P, Dubochet J: **Granular layer in the periplasmic space of gram-positive bacteria and fine structures of *Enterococcus gallinarum* and *Streptococcus gordonii* septa revealed by cryo-electron microscopy of vitreous sections.** *J Bacteriol* 2006, **188:**6652-6660.
237. Desvaux M, Dumas E, Chafsey I, Hebraud M: **Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure.** *FEMS Microbiol Lett* 2006, **256:**1-15.
238. Nijeholt JAL, Driessen AJ: **The bacterial Sec-translocase: structure and mechanism.** *Philos Trans R Soc Lond B Biol Sci* 2012, **367:**1016-1028.
239. Tsukazaki T, Mori H, Echizen Y, Ishitani R, Fukai S, Tanaka T, Perederina A, Vassylyev DG, Kohno T, Maturana AD: **Structure and function of a membrane component SecDF that enhances protein export.** *Nature* 2011, **474:**235-238.
240. Beck K, Wu L-F, Brunner J, Müller M: **Discrimination between SRP-and SecA/SecB-dependent substrates involves selective recognition of nascent chains by SRP and trigger factor.** *EMBO J* 2000, **19:**134-143.
241. Luirink J, Heijne Gv, Houben E, Gier J-Wd: **Biogenesis of inner membrane proteins in *Escherichia coli*.** *Annu Rev Microbiol* 2005, **59:**329-355.
242. Welte T, Kudva R, Kuhn P, Sturm L, Braig D, Müller M, Warscheid B, Drepper F, Koch H-G: **Promiscuous targeting of polytopic membrane proteins to SecYEG or YidC by the *Escherichia coli* signal recognition particle.** *Mol Biol Cell* 2012, **23:**464-479.
243. Dalbey RE, Wang P, Kuhn A: **Assembly of bacterial inner membrane proteins.** *Annu Rev Biochem* 2011, **80:**161-187.
244. Chen M, Xie K, Jiang F, Yi L, Dalbey RE: **YidC, a newly defined evolutionarily conserved protein, mediates membrane protein assembly in bacteria.** *Biol Chem* 2002, **383:**1565-1572.
245. Beck K, Eisner G, Trescher D, Dalbey RE, Brunner J, Müller M: **YidC, an assembly site for polytopic *Escherichia coli* membrane proteins located in immediate proximity to the SecYE translocon and lipids.** *EMBO Rep* 2001, **2:**709-714.
246. Sachelaru I, Petriman NA, Kudva R, Kuhn P, Welte T, Knapp B, Drepper F, Warscheid B, Koch H-G: **YidC occupies the lateral gate of the SecYEG translocon and is sequentially displaced by a nascent membrane protein.** *J Biol Chem* 2013, **288:**16295-16307.

247. Wang P, Dalbey RE: **Inserting membrane proteins: the YidC/Oxa1/Alb3 machinery in bacteria, mitochondria, and chloroplasts.** *Biochim Biophys Acta* 2011, **1808:**866-875.
248. Palmer T, Berks BC: **The twin-arginine translocation (Tat) protein export pathway.** *Nat Rev Microbiol* 2012, **10:**483-496.
249. Jack RL, Buchanan G, Dubini A, Hatzixanthis K, Palmer T, Sargent F: **Coordinating assembly and export of complex bacterial proteins.** *EMBO J* 2004, **23:**3962-3972.
250. Reddy BL, Saier Jr MH: **Topological and phylogenetic analyses of bacterial holin families and superfamilies.** *Biochim Biophys Acta* 2013, **1828:**2654-2671.
251. Wang I-N, Smith DL, Young R: **Holins: the protein clocks of bacteriophage infections.** *Annu Rev Microbiol* 2000, **54:**799-825.
252. Young R, Wang I-N, Roof WD: **Phages will out: strategies of host cell lysis.** *Trends Microbiol* 2000, **8:**120-128.
253. Spirig T, Weiner EM, Clubb RT: **Sortase enzymes in Gram-positive bacteria.** *Mol Microbiol* 2011, **82:**1044-1059.
254. Pallen MJ, Lam AC, Antonio M, Dunbar K: **An embarrassment of sortases–a richness of substrates?** *Trends Microbiol* 2001, **9:**97-101.
255. Hendrickx AP, Budzik JM, Oh S-Y, Schneewind O: **Architects at the bacterial surface-sortases and the assembly of pili with isopeptide bonds.** *Nat Rev Microbiol* 2011, **9:**166-176.
256. Fagan RP, Fairweather NF: ***Clostridium difficile*** **has two parallel and essential Sec secretion systems.** *J Biol Chem* 2011, **286:**27483-27493.
257. Rigel NW, Braunstein M: **A new twist on an old pathway–accessory secretion systems.** *Mol Microbiol* 2008, **69:**291-302.
258. Feltcher ME, Braunstein M: **Emerging themes in SecA2-mediated protein export.** *Nat Rev Microbiol* 2012, **10:**779-789.
259. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC Microbiol* 2005, **5:**58.
260. Takamatsu D, Bensing BA, Cheng H, Jarvis GA, Siboo IR, López JA, Griffiss JM, Sullam PM: **Binding of the *Streptococcus gordonii* surface glycoproteins GspB and Hsa to specific carbohydrate structures on platelet membrane glycoprotein Ibalpha.** *Mol Microbiol* 2005, **58:**380-392.
261. Bensing BA, Sullam PM: **An accessory sec locus of *Streptococcus gordonii* is required for export of the surface protein GspB and for normal levels of binding to human platelets.** *Mol Microbiol* 2002, **44:**1081-1094.
262. Koronakis V, Eswaran J, Hughes C: **Structure and function of TolC: the bacterial exit duct for proteins and drugs.** *Annu Rev Biochem* 2004, **73:**467-489.
263. Koronakis V, Sharff A, Koronakis E, Luisi B, Hughes C: **Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export.** *Nature* 2000, **405:**914-919.
264. Rees DC, Johnson E, Lewinson O: **ABC transporters: the power to change.** *Nat Rev Mol Cell Biol* 2009, **10:**218-227.
265. Filloux A, Hachani A, Bleves S: **The bacterial type VI secretion machine: yet another player for protein transport across membranes.** *Microbiology* 2008, **154:**1570-1583.
266. Reichow SL, Korotkov KV, Hol WG, Gonen T: **Structure of the cholera toxin secretion channel in its closed state.** *Nat Struct Mol Biol* 2010, **17:**1226-1232.

267. Cianciotto NP: **Type II secretion: a protein secretion system for all seasons.** *Trends Microbiol* 2005, **13:**581-588.
268. Kubori T, Matsushima Y, Nakamura D, Uralil J, Lara-Tejero M, Sukhan A, Galán JE, Aizawa S-I: **Supramolecular structure of the *Salmonella typhimurium* type III protein secretion system.** *Science* 1998, **280:**602-605.
269. Saier Jr MH: **Evolution of bacterial type III protein secretion systems.** *Trends Microbiol* 2004, **12:**113-115.
270. Fronzes R, Christie PJ, Waksman G: **The structural biology of type IV secretion systems.** *Nat Rev Microbiol* 2009, **7:**703-714.
271. Alvarez-Martinez CE, Christie PJ: **Biological diversity of prokaryotic type IV secretion systems.** *Microbiol Mol Biol Rev* 2009, **73:**775-808.
272. Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D: **Type V protein secretion pathway: the autotransporter story.** *Microbiol Mol Biol Rev* 2004, **68:**692-744.
273. Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ: **Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system.** *Proc Natl Acad Sci U S A* 2006, **103:**1528-1533.
274. Kapitein N, Mogk A: **Deadly syringes: type VI secretion system activities in pathogenicity and interbacterial competition.** *Curr Opin Microbiol* 2013, **16:**52-58.
275. Coulthurst SJ: **The type VI secretion system–a widespread and versatile cell targeting system.** *Res Microbiol* 2013, **164:**640-654.
276. Proft T, Baker E: **Pili in Gram-negative and Gram-positive bacteria-structure, assembly and their role in disease.** *Cell Mol Life Sci* 2009, **66:**613-635.
277. Fronzes R, Remaut H, Waksman G: **Architectures and biogenesis of non-flagellar protein appendages in Gram-negative bacteria.** *EMBO J* 2008, **27:**2271-2280.
278. Rakotoarivonina H, Jubelin G, Hebraud M, Gaillard-Martinie B, Forano E, Mosoni P: **Adhesion to cellulose of the Gram-positive bacterium *Ruminococcus albus* involves type IV pili.** *Microbiology* 2002, **148:**1871-1880.
279. Economou A, Christie PJ, Fernandez RC, Palmer T, Plano GV, Pugsley AP: **Secretion by numbers: protein traffic in prokaryotes.** *Mol Microbiol* 2006, **62:**308-319.
280. Pelicic V: **Type IV pili: *e pluribus unum*?** *Mol Microbiol* 2008, **68:**827-837.
281. Trindade MB, Job V, Contreras-Martel C, Pelicic V, Dessen A: **Structure of a widely conserved type IV pilus biogenesis factor that affects the stability of secretin multimers.** *J Mol Biol* 2008, **378:**1031-1039.
282. Nakai K: **Protein sorting signals and prediction of subcellular localization.** *Adv Protein Chem* 2000, **54:**277-344.
283. Blobel G, Sabatini DD: **Ribosome-membrane interaction in eukaryotic cells.** In *Biomembranes. Volume* 2: Springer; 1971: 193-195.
284. Rusch SL, Kendall DA: **Interactions that drive Sec-dependent bacterial protein transport.** *Biochemistry* 2007, **46:**9665-9673.
285. Zhou M, Boekhorst J, Francke C, Siezen RJ: **LocateP: genome-scale subcellular-location predictor for bacterial proteins.** *BMC Bioinformatics* 2008, **9:**173.
286. Driessen AJ, Nouwen N: **Protein translocation across the bacterial cytoplasmic membrane.** *Annu Rev Biochem* 2008, **77:**643-667.
287. Dalbey RE, Wang P, van Dijl JM: **Membrane proteases in the bacterial protein secretion and quality control pathway.** *Microbiol Mol Biol Rev* 2012, **76:**311-330.

288. Lüke I, Handford JI, Palmer T, Sargent F: **Proteolytic processing of *Escherichia coli* twin-arginine signal peptides by LepB.** *Arch Microbiol* 2009, **191:**919-925.
289. Auclair SM, Bhanu MK, Kendall DA: **Signal peptidase I: cleaving the way to mature proteins.** *Protein Sci* 2012, **21:**13-25.
290. Okuda S, Tokuda H: **Lipoprotein sorting in bacteria.** *Annu Rev Microbiol* 2011, **65:**239-259.
291. Thompson BJ, Widdick DA, Hicks MG, Chandra G, Sutcliffe IC, Palmer T, Hutchings MI: **Investigating lipoprotein biogenesis and function in the model Gram-positive bacterium *Streptomyces coelicolor*.** *Mol Microbiol* 2010, **77:**943-957.
292. Arts J, van Boxtel R, Filloux A, Tommassen J, Koster M: **Export of the pseudopilin XcpT of the *Pseudomonas aeruginosa* type II secretion system via the signal recognition particle-Sec pathway.** *J Bacteriol* 2007, **189:**2069-2076.
293. Francetic O, Buddelmeijer N, Lewenza S, Kumamoto CA, Pugsley AP: **Signal recognition particle-dependent inner membrane targeting of the PulG pseudopilin component of a type II secretion system.** *J Bacteriol* 2007, **189:**1783-1793.
294. Peabody CR, Chung YJ, Yen M-R, Vidal-Ingigliardi D, Pugsley AP, Saier MH: **Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella.** *Microbiology* 2003, **149:**3051-3072.
295. Fekkes P, Driessen AJ: **Protein targeting to the bacterial cytoplasmic membrane.** *Microbiol Mol Biol Rev* 1999, **63:**161-173.
296. Hikita C, Mizushima S: **The requirement of a positive charge at the amino terminus can be compensated for by a longer central hydrophobic stretch in the functioning of signal peptides.** *J Biol Chem* 1992, **267:**12375-12379.
297. Gennity J, Goldstein J, Inouye M: **Signal peptide mutants of *Escherichia coli*.** *J Bioenerg Biomembr* 1990, **22:**233-269.
298. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, et al: **The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific.** *PLoS Biol* 2007, **5**.
299. Natale P, Brüser T, Driessen AJ: **Sec- and Tat- mediated protein secretion across the bacterial cytoplasmic membrane-distinct translocases and mechanisms.** *Biochim Biophys Acta* 2008, **1778:**1735-1756.
300. Nakayama H, Kurokawa K, Lee BL: **Lipoproteins in bacteria: structures and biosynthetic pathways.** *FEBS J* 2012, **279:**4247-4268.
301. Baumgärtner M, Kärst U, Gerstel B, Loessner M, Wehland J, Jänsch L: **Inactivation of Lgt allows systematic characterization of lipoproteins from *Listeria monocytogenes*.** *J Bacteriol* 2007, **189:**313-324.
302. Fukuda A, Matsuyama S, Hara T, Nakayama J, Nagasawa H, Tokuda H: **Aminoacylation of the N-terminal cysteine is essential for Lol-dependent release of lipoproteins from membranes but does not depend on lipoprotein sorting signals.** *J Biol Chem* 2002, **277:**43512-43518.
303. Masuda K, Matsuyama S-i, Tokuda H: **Elucidation of the function of lipoprotein-sorting signals that determine membrane localization.** *Proc Natl Acad Sci U S A* 2002, **99:**7390-7395.
304. Berks BC: **A common export pathway for proteins binding complex redox cofactors?** *Mol Microbiol* 1996, **22:**393-404.
305. Stanley NR, Palmer T, Berks BC: **The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in *Escherichia coli*.** *J Biol Chem* 2000, **275:**11591-11596.

306. Hinsley AP, Stanley NR, Palmer T, Berks BC: **A naturally occurring bacterial Tat signal peptide lacking one of the 'invariant' arginine residues of the consensus targeting motif.** *FEBS Lett* 2001, **497:**45-49.
307. Brüser T, Deutzmann R, Dahl C: **Evidence against the double-arginine motif as the only determinant for protein translocation by a novel Sec-independent pathway in *Escherichia coli*.** *FEMS Microbiol Lett* 1998, **164:**329-336.
308. Bogsch E, Brink S, Robinson C: **Pathway specificity for a delta pH-dependent precursor thylakoid lumen protein is governed by a 'Sec-avoidance' motif in the transfer peptide and a 'Sec-incompatible' mature protein.** *EMBO J* 1997, **16:**3851-3859.
309. Bogsch EG, Sargent F, Stanley NR, Berks BC, Robinson C, Palmer T: **An essential component of a novel bacterial protein export system with homologues in plastids and mitochondria.** *J Biol Chem* 1998, **273:**18003-18006.
310. Wexler M, Bogsch EG, Klösgen RB, Palmer T, Robinson C, Berks BC: **Targeting signals for a bacterial Sec-independent export system direct plant thylakoid import by the ΔpH pathway.** *FEBS Lett* 1998, **431:**339-342.
311. LaPointe CF, Taylor RK: **The type 4 prepilin peptidases comprise a novel family of aspartic acid proteases.** *J Biol Chem* 2000, **275:**1502-1510.
312. Dupuy B, Deghmane A, Kheir M: **Type IV prepilin peptidase.** In *Handbook of proteolytic enzymes.* 3rd edition. Edited by Rawlings ND, Salvesen G. New York: Academic Press 2013.
313. Pohlschroder M, Ghosh A, Tripepi M, Albers S-V: **Archaeal type IV pilus-like structures-evolutionarily conserved prokaryotic surface organelles.** *Curr Opin Microbiol* 2011, **14:**357-363.
314. Imam S, Chen Z, Roos DS, Pohlschröder M: **Identification of surprisingly diverse type IV pili, across a broad range of Gram-positive bacteria.** *PLOS ONE* 2011, **6:**e28919.
315. White SH, von Heijne G: **How translocons select transmembrane helices.** *Annu Rev Biophys* 2008, **37:**23-42.
316. Tjalsma H, van Dijl JM: **Proteomics-based consensus prediction of protein retention in a bacterial membrane.** *Proteomics* 2005, **5:**4472-4482.
317. Daleke MH, Ummels R, Bawono P, Heringa J, Vandenbroucke-Grauls CM, Luirink J, Bitter W: **General secretion signal for the mycobacterial type VII secretion pathway.** *Proc Natl Acad Sci U S A* 2012, **109:**11342-11347.
318. Braunstein M, Espinosa BJ, Chan J, Belisle JT, R Jacobs W: **SecA2 functions in the secretion of superoxide dismutase A and in the virulence of *Mycobacterium tuberculosis*.** *Mol Microbiol* 2003, **48:**453-464.
319. Lloubes R, Bernadac A, Pommier S: **Non classical secretion systems.** *Res Microbiol* 2013, **164:**665-663.
320. Huberts DH, van der Klei IJ: **Moonlighting proteins: an intriguing mode of multitasking.** *Biochim Biophys Acta* 2010, **1803:**520-525.
321. Henderson B, Martin A: **Bacterial virulence in the moonlight: multitasking bacterial moonlighting proteins are virulence determinants in infectious disease.** *Infect Immun* 2011, **79:**3476-3491.
322. Lee SY, Choi JH, Xu Z: **Microbial cell-surface display.** *Trends Biotechnol* 2003, **21:**45-52.

323. Chen W, Georgiou G: **Cell-surface display of heterologous proteins: from high-throughput screening to environmental applications.** *Biotechnol Bioeng* 2002, **79:**496-503.
324. Georgiou G, Stathopoulos C, Daugherty PS, Nayak AR, Iverson BL, Curtiss III R: **Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines.** *Nat Biotechnol* 1997, **15:**29-34.
325. Liu R, Yang C, Xu Y, Xu P, Jiang H, Qiao C: **Development of a whole-cell biocatalyst/biosensor by display of multiple heterologous proteins on the *Escherichia coli* cell surface for the detoxification and detection of organophosphates.** *J Agric Food Chem* 2013, **61:**7810-7816.
326. Åvall-Jääskeläinen S, Lindholm A, Palva A: **Surface display of the receptor-binding region of the *Lactobacillus brevis* S-layer protein in *Lactococcus lactis* provides nonadhesive lactococci with the ability to adhere to intestinal epithelial cells.** *Appl Environ Microbiol* 2003, **69:**2230-2236.
327. Wernérus H, Ståhl S: **Biotechnological applications for surface-engineered bacteria.** *Biotechnol Appl Biochem* 2004, **40:**209-228.
328. Rosander A, Bjerketorp J, Frykberg L, Jacobsson K: **Phage display as a novel screening method to identify extracellular proteins.** *J Microbiol Methods* 2002, **51:**43-55.
329. Jacobsson K, Rosander A, Bjerketorp J, Frykberg L: **Shotgun phage display-selection for bacterial receptins or other exported proteins.** *Biol Proced Online* 2003, **5:**123-135.
330. Rosander A, Guss B, Pringle M: **An IgG-binding protein A homolog in *Staphylococcus hyicus*.** *Vet Microbiol* 2011, **149:**273-276.
331. Bjerketorp J, Rosander A, Nilsson M, Jacobsson K, Frykberg L: **Sorting a *Staphylococcus aureus* phage display library against *ex vivo* biomaterial.** *J Med Microbiol* 2004, **53:**945-951.
332. Yang X-Y, Lu J, Sun X, He Q-Y: **Application of subproteomics in the characterization of Gram-positive bacteria.** *J Proteomics* 2012, **75:**2803-2810.
333. Caccia D, Dugo M, Callari M: **Bioinformatics tools for secretome analysis.** *Biochim Biophys Acta* 2013, **1834:**2442-2453.
334. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT: **Individual genome assembly from complex community short-read metagenomic datasets.** *ISME J* 2011, **6:**898-901.
335. Goudenège D, Avner S, Lucchetti-Miganeh C, Barloy-Hubler F: **CoBaltDB: complete bacterial and archaeal orfeomes subcellular localization database and associated resources.** *BMC Microbiol* 2010, **10:**88.
336. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions.** *Nat Methods* 2011, **8:**785-786.
337. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A: **Prediction of lipoprotein signal peptides in Gram-negative bacteria.** *Protein Sci* 2003, **12:**1652-1662.
338. Krogh A, Larsson BÈ, Von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a Hidden Markov Model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.

339. Bagos PG, Tsirigos KD, Liakopoulos TD, Hamodrakas SJ: **Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model.** *J Proteome Res* 2008, **7:**5082-5093.
340. Bagos PG, Nikolaou EP, Liakopoulos TD, Tsirigos KD: **Combined prediction of Tat and Sec signal peptides with Hidden Markov Models.** *Bioinformatics* 2010, **26:**2811-2817.
341. Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26:**1608-1615.
342. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300:**1005-1016.
343. Choo KH, Tan TW, Ranganathan S: **A comprehensive assessment of N-terminal signal peptides prediction methods.** *BMC Bioinformatics* 2009, **10:**S2.
344. Smith G: **Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface.** *Science* 1985, **228:**1315.
345. Rakonjac J, Bennett NJ, Spagnuolo J, Gagic D, Russel M: **Filamentous bacteriophage: biology, phage display and nanotechnology applications.** *Curr Issues Mol Biol* 2011, **13:**51.
346. Marvin D, Hohn B: **Filamentous bacterial viruses.** *Bacteriol Rev* 1969, **33:**172.
347. Russel M, Model P: **Filamentous phage.** *The Bacteriophages* 2006**:**146–160.
348. Rakonjac J, Conway JF: **Bacteriophages: self-assembly and applications.** In *Molecular Bionanotechnology.* Edited by Rehm B: Horizon Bioscience; 2006: 153-190.
349. Jankovic D: **Direct selection and phage display of the *Lactobacillus rhamnosus* HN001.** *PhD thesis.* Massey University, IMBS; 2008.
350. Deng L-W, Malik P, Perham RN: **Interaction of the globular domains of pIII protein of filamentous bacteriophage fd with the F pilus of *Escherichia coli*.** *Virology* 1999, **253:**271-277.
351. Endemann H, Model P: **Location of filamentous phage minor coat proteins in phage and in infected cells.** *J Mol Biol* 1995, **250:**496-506.
352. Russel M: **Protein-protein interactoins during filamentous phage assembly.** *J Mol Biol* 1993, **231:**689-697.
353. Russel M: **Moving through the membrane with filamentous phages.** *Trends Microbiol* 1995, **3:**223-228.
354. Rakonjac J, Feng J, Model P: **Filamentous phage are released from the bacterial membrane by a two-step mechanism involving a short C-terminal fragment of pIII.** *J Mol Biol* 1999, **289:**1253-1265.
355. Rakonjac J, Model P: **Roles of pIII in filamentous phage assembly.** *J Mol Biol* 1998, **282:**25-41.
356. Willats WGT: **Phage display: practicalities and prospects.** *Plant Mol Biol* 2002, **50:**837-854.
357. Kay B, Winter J, McCafferty J (Eds.): **Phage display of peptides and proteins: a laboratory manual**: Academic Press, Inc.; 1996.
358. Pande J, Szewczyk M, Grover A: **Phage display: concept, innovations, applications and future.** *Biotechnol Adv* 2010, **28:**849-858.
359. Fernandez-Gacio A, Uguen M, Fastrez J: **Phage display as a tool for the directed evolution of enzymes.** *Trends Biotechnol* 2003, **21:**408-414.

360. Uchiyama F, Tanaka Y, Minari Y, Tokui N: **Designing scaffolds of peptides for phage display libraries.** *J Biosci Bioeng* 2005, **99:**448-456.
361. Kay B, Kasanov J, Knight S, Kurakin A: **Convergent evolution with combinatorial peptides.** *FEBS Lett* 2000, **480:**55-62.
362. Devlin J, Panganiban L, Devlin P: **Random peptide libraries: a source of specific protein binding molecules.** *Science* 1990, **249:**404.
363. Jestin J: **Functional cloning by phage display.** *Biochimie* 2008, **90:**1273-1278.
364. Carmen S, Jermutus L: **Concepts in antibody phage display.** *Brief Funct Genomics Proteomic* 2002, **1:**189.
365. Mao C, Solis DJ, Reiss BD, Kottmann ST, Sweeney RY, Hayhurst A, Georgiou G, Iverson B, Belcher AM: **Virus-Based Toolkit for the Directed Synthesis of Magnetic and Semiconducting Nanowires.** *Science* 2004, **303:**213-217.
366. Nam KT, Kim D-W, Yoo PJ, Chiang C-Y, Meethong N, Hammond PT, Chiang Y-M, Belcher AM: **Virus-Enabled Synthesis and Assembly of Nanowires for Lithium Ion Battery Electrodes.** *Science* 2006, **312:**885-888.
367. Efimov VP, Nepluev IV, Mesyanzhinov VV: **Bacteriophage T4 as a surface display vector.** *Virus Genes* 1995, **10:**173-177.
368. Rosenberg A, Griffin K, Studier FW, McCormick M, Berg J, Novy R, Mierendorf R: **T7Select® Phage Display System: A powerful new protein display system based on bacteriophage T7.** In *Newsletter of Novagen, Inc*, vol. 6; 1996.
369. Danner S, Belasco JG: **T7 phage display: a novel genetic selection system for cloning RNA-binding proteins from cDNA libraries.** *Proc Natl Acad Sci U S A* 2001, **98:**12954-12959.
370. Maruyama IN, Maruyama HI, Brenner S: **Lambda foo: a lambda phage vector for the expression of foreign proteins.** *Proc Natl Acad Sci U S A* 1994, **91:**8273-8277.
371. Sternberg N, Hoess RH: **Display of peptides and proteins on the surface of bacteriophage lambda.** *Proc Natl Acad Sci U S A* 1995, **92:**1609-1613.
372. Paschke M: **Phage display systems and their applications.** *Appl Microbiol Biotechnol* 2006, **70:**2-11.
373. Sidhu S: **Phage display: increasing the rewards from genomic information.** *Drug Discov Today* 2001, **6:**936.
374. Rakonjac J, Jovanovic G, Model P: **Filamentous phage infection-mediated gene expression: construction and propagation of the *gIII* deletion mutant helper phage R408d3.** *Gene* 1997, **198:**99-103.
375. Naik RR, Brott LL, Clarson SJ, Stone MO: **Silica-precipitating peptides isolated from a combinatorial phage display peptide library.** *J Nanosci Nanotechnol* 2002, **2:**95-100.
376. Frascione N, Codina-Barrios A, Bassindale AR, Taylor PG: **Enhancing in vitro selection techniques to assist the discovery, understanding and use of inorganic binding peptides.** *Dalton Trans* 2013, **42:**10337-10346.
377. Liu Y, Mao J, Zhou B, Wei W, Gong S: **Peptide aptamers against titanium-based implants identified through phage display.** *J Mater Sci: Mater Med* 2010, **21:**1103-1107.
378. Dennissen MA, Jenniskens GJ, Pieffers M, Versteeg EM, Petitou M, Veerkamp JH, van Kuppevelt TH: **Large, tissue-regulated domain diversity of heparan sulfates demonstrated by phage display antibodies.** *J Biol Chem* 2002, **277:**10982-10986.

379. Guo J, Catchmark JM, Mohamed MNA, Benesi AJ, Tien M, Kao T-h, Watts HD, Kubicki JD: **Identification and characterization of a cellulose binding heptapeptide revealed by phage display.** *Biomacromolecules* 2013, **14:**1795-1805.
380. Watters JM, Telleman P, Junghans RP: **An optimized method for cell-based phage display panning.** *Immunotechnology* 1997, **3:**21-29.
381. Pavoni E, Vaccaro P, Anastasi AM, Minenkova O: **Optimized selection of anti-tumor recombinant antibodies from phage libraries on intact cells.** *Mol Immunol* 2014, **57:**317-322.
382. Johns M, George A, Ritter M: ***In vivo* selection of sFv from phage display libraries.** *J Immunol Methods* 2000, **239:**137-151.
383. Pasqualini R, Ruoslahti E: **Organ targeting *in vivo* using phage display peptide libraries.** *Nature* 1996, **380:**364-366.
384. Arap W, Kolonin MG, Trepel M, Lahdenranta J, Cardó-Vila M, Giordano RJ, Mintz PJ, Ardelt PU, Yao VJ, Vidal CI: **Steps toward mapping the human vasculature by phage display.** *Nat Med* 2002, **8:**121-127.
385. Kolonin MG, Sun J, Do K-A, Vidal CI, Ji Y, Baggerly KA, Pasqualini R, Arap W: **Synchronous selection of homing peptides for multiple tissues by *in vivo* phage display.** *FASEB J* 2006, **20:**979-981.
386. Dias-Neto E, Nunes DN, Giordano RJ, Sun J, Botz GH, Yang K, Setubal JC, Pasqualini R, Arap W: **Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis.** *PLOS ONE* 2009, **4:**e8338.
387. Di Niro R, Sulic A-M, Mignone F, D'Angelo S, Bordoni R, Iacono M, Marzari R, Gaiotto T, Lavric M, Bradbury AR: **Rapid interactome profiling by massive sequencing.** *Nucl Acids Res* 2010, **38:**e110-e110.
388. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois M, Fischer N: **By-passing in vitro screening—next generation sequencing technologies applied to antibody display and *in silico* candidate selection.** *Nucl Acids Res* 2010, **38:**e193-e193.
389. Hoen PA, Jirka SM, ten Broeke BR, Schultes EA, Aguilera B, Pang KH, Heemskerk H, Aartsma-Rus A, van Ommen GJ, den Dunnen JT: **Phage display screening without repetitious selection rounds.** *Anal Biochem* 2012, **421:**622-631.
390. Mathonet P, Ullman CG: **The application of next generation sequencing to the understanding of antibody repertoires.** *Front Immunol* 2013, **4**.
391. Model P, Jovanovic G, Dworkin J: **The *Escherichia coli* phage-shock-protein (psp) operon.** *Mol Microbiol* 1997, **24:**255-261.
392. Beekwilder J, Rakonjac J, Jongsma M, Bosch D: **A phagemid vector using the *E. coli* phage shock promoter facilitates phage display of toxic proteins.** *Gene* 1999, **228:**23-31.
393. Gagic D, Wen W, Collett MA, Rakonjac J: **Unique secreted-surface protein complex of *Lactobacillus rhamnosus*, identified by phage display.** *MicrobiologyOpen* 2013, **2:**1-17.
394. Liu S, Han W, Sun C, Lei L, Feng X, Yan S, Diao Y, Gao Y, Zhao H, Liu Q, et al: **Subtractive screening with the *Mycobacterium tuberculosis* surface protein phage display library.** *Tuberculosis (Edinb)* 2011, **91:**579-586.
395. Liu S, Han W, Sun C, Lei L, Feng X: **Identification of two new virulence factors of *Mycobacterium tuberculosis* that induce multifunctional CD4 T cell responses.** *J Mycobac Dis* 2013, **3:**2161-1068.1000125.

396. Sambrook J, Russell DW: *Molecular cloning: a laboratory manual*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2001.
397. Lu G, Moriyama EN: **Vector NTI, a balanced all-in-one sequence analysis suite.** *Brief Bioinform* 2004, **5:**378-388.
398. **SeqClean script** [**http://seqclean.sourceforge.net/**]
399. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation.** *Nucl Acids Res* 2012, **40:**W445-451.
400. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, et al: **IMG/M: the integrated metagenome data management and comparative analysis system.** *Nucl Acids Res* 2012, **40:**D123-129.
401. **GS Roche De Novo Assembler** [**http://www.454.com/products/analysis-software/**]
402. Huang Y, Niu B, Gao Y, Fu L, Li W: **CD-HIT suite: a web server for clustering and comparing biological sequences.** *Bioinformatics* 2010, **26:**680-682.
403. Rice P, Longden I, Bleasby A: **EMBOSS: the European molecular biology open software suite.** *Trends Genet* 2000, **16:**276-277.
404. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
405. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucl Acids Res* 2005, **33:**W116-W120.
406. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC: **CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database.** *Glycobiology* 2010, **20:**1574-1584.
407. Theis C, Reeder J, Giegerich R: **KnotInFrame: prediction of -1 ribosomal frameshift events.** *Nucl Acids Res* 2008, **36:**6013-6020.
408. Huang J, Ru B, Li S, Lin H, Guo F-B: **SAROTUP: scanner and reporter of target-unrelated peptides.** *J Biomed Biotechnol* 2010, **2010:**1-7.
409. Nelson FK, Friedman SM, Smith GP: **Filamentous phage DNA cloning vectors: a noninfective mutant with a nonpolar deletion in gene III.** *Virology* 1981, **108:**338-350.
410. Stein J, Marsh T, Wu K, Shizuya H, DeLong E: **Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon.** *J Bacteriol* 1996, **178:**591.
411. Atrazhev A, Elliott J: **Simplified desalting of ligation reactions immediately prior to electroporation into *E. coli*.** *Biotechniques* 1996, **21:**1024.
412. **The Integrated Microbial Genomes with Microbiome samples (IMG/M) webpage** [https://img.jgi.doe.gov/cgi-bin/m/main.cgi]
413. **NCBI BioProject database** [**http://www.ncbi.nlm.nih.gov/bioproject/**]
414. **Metagenome Annotation Standard Operating Procedure for IMG** [https://img.jgi.doe.gov/mer/doc/MetagenomeAnnotationSOP.pdf]
415. Eddy SR: **A new generation of homology search tools based on probabilistic inference.** In *Genome Inform*. World Scientific; 2009: 205-211.
416. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucl Acids Res* 2011**:**367-383.
417. Carlier J-P, Bedora-Faure M, K'ouas G, Alauzet C, Mory F: **Proposal to unify *Clostridium orbiscindens* Winter et al. 1991 and *Eubacterium plautii* (Séguin 1928) Hofstad and Aasjord 1982, with description of *Flavonifractor plautii* gen. nov.,**

comb. nov., and reassignment of *Bacteroides capillosus* to *Pseudoflavonifractor capillosus* gen. nov., comb. nov. *Int J Syst Evol Microbiol* 2010, **60:**585-590.

418. Downes J, Dewhirst FE, Tanner AC, Wade WG: **Description of *Alloprevotella rava* gen. nov., sp. nov., isolated from the human oral cavity, and reclassification of *Prevotella tannerae* Moore et al. 1994 as *Alloprevotella tannerae* gen. nov., comb. nov.** *Int J Syst Evol Microbiol* 2013, **63:**1214-1218.
419. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** *Nucl Acids Res* 2007, **35:**7188-7196.
420. Krieg NR, Ludwig W, Euzéby J, Whitman WB: **Phylum XIV. Bacteroidetes phyl. nov.** In *Bergey's Manual® of Systematic Bacteriology. Volume* 4. 2nd edition. New York: Springer; 2010: 25-469.
421. Schleifer K-H: **Phylum XIII. Firmicutes Gibbons and Murray 1978, 5 (Firmacutes [sic] Gibbons and Murray 1978, 5).** In *Bergey's Manual® of Systematic Bacteriology. Volume* 3: Springer; 2009: 19-1317.
422. **dbCAN database download** [**http://csbl.bmb.uga.edu/dbCAN/download/dbCAN-all-domains.txt**]
423. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucl Acids Res* 2012, **40:**D130-D135.
424. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database* 2011.
425. **JGI's Metagenome Program** [**http://jgi.doe.gov/our-science/science-programs/metagenomics/**]
426. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: a community resource for metagenomics.** *PLoS Biol* 2007, **5:**e75.
427. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464:**59-65.
428. Gordon JI, Ley RE, Wilson R, Mardis E, Xu J, Fraser CM, Relman DA: **Extending our view of self: the human gut microbiome initiative (HGMI).** 2005.
429. **Phytozome Project** [**http://www.phytozome.net/**]
430. 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:**644-648.
431. Jung H, Buxton D, Hatfield R, Ralph J: *Forage cell wall structure and digestibility.* American Society of Agronomy, Inc.; 1993.
432. von Heijne G, Abrahmsèn L: **Species-specific variation in signal peptide design: implications for protein secretion in foreign hosts.** *FEBS Lett* 1989, **244:**439-446.
433. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J: **The Pfam protein families database.** *Nucl Acids Res* 2012, **40:**D290-D301.
434. Zaldivar J, Martinez A, Ingram LO: **Effect of selected aldehydes on the growth and fermentation of ethanologenic *Escherichia coli*.** *Biotechnol Bioeng* 1999, **65:**24-33.
435. Barthelmebs L, Diviès C, Cavin J-F: **Expression in *Escherichia coli* of native and chimeric phenolic acid decarboxylases with modified enzymatic activities and method for screening recombinant *E. coli* strains expressing these enzymes.** *Appl Environ Microbiol* 2001, **67:**1063-1069.

436. Zaldivar J, Nielsen J, Olsson L: **Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration.** *Appl Microbiol Biotechnol* 2001, **56:**17-34.
437. Cárcamo J, Ravera MW, Brissette R, Dedova O, Beasley JR, Alam-Moghé A, Wan C, Blume A, Mandecki W: **Unexpected frameshifts from gene to expressed protein in a phage-displayed peptide library.** *Proc Natl Acad Sci U S A* 1998, **95:**11146-11151.
438. Jacobsson K, Frykberg L: **Phage display shot-gun cloning of ligand-binding domains of prokaryotic receptors approaches 100% correct clones.** *Biotechniques* 1996, **20:**1070-1081.
439. Jacobsson K, Frykberg L: **Cloning of ligand-binding domains of bacterial receptors by phage display.** *Biotechniques* 1995, **18:**878-885.
440. Song L, Mandecki W, Goldman E: **Expression of non-open reading frames isolated from phage display due to translation reinitiation.** *FASEB J* 2003, **17:**1674-1681.
441. Goldman E, Korus M, MandeckiI W: **Efficiencies of translation in three reading frames of unusual non-ORF sequences isolated from phage display.** *FASEB J* 2000, **14:**603-611.
442. Farabaugh PJ: **Programmed translational frameshifting.** *Microbiol Rev* 1996, **60:**103.
443. Paschke M, Höhne W: **A twin-arginine translocation (Tat)-mediated phage display system.** *Gene* 2005, **350:**79-88.
444. Thie H, Schirrmann T, Paschke M, Dübel S, Hust M: **SRP and Sec pathway leader peptides for antibody phage display and antibody fragment production in *E. coli*.** *Nat Biotechnol* 2008, **25:**49-54.
445. Duguay AR, Silhavy TJ: **Quality control in the bacterial periplasm.** *Biochim Biophys Acta* 2004, **1694:**121-134.
446. Koike S, Kobayashi Y: **Development and use of competitive PCR assays for the rumen cellulolytic bacteria: *Fibrobacter succinogenes, Ruminococcus albus* and *Ruminococcus flavefaciens*.** *FEMS Microbiol Lett* 2001, **204:**361-366.
447. Weimer P, Waghorn G, Odt C, Mertens D: **Effect of diet on populations of three species of ruminal cellulolytic bacteria in lactating dairy cows.** *J Dairy Sci* 1999, **82:**122-134.
448. Noel S: **Cultivation and community composition analysis of plant-adherent rumen bacteria.** Massey University, IMBS; 2013.
449. **Genomic Encyclopedia of Bacteria and Archaea project** [**http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/**]
450. **The Hungate 1000 project** [**http://www.hungate1000.org.nz/**]
451. Ziemer CJ, Sharp R, Stern MD, Cotta MA, Whitehead TR, Stahl DA: **Comparison of microbial populations in model and natural rumens using 16S ribosomal RNA-targeted probes.** *Environ Microbiol* 2008, **2:**632-643.
452. Bardy SL, Eichler J, Jarrell KF: **Archaeal signal peptides-a comparative survey at the genome level.** *Protein Sci* 2009, **12:**1833-1843.
453. An D, Dong X, Dong Z: **Prokaryote diversity in the rumen of yak (*Bos grunniens*) and Jinnan cattle (*Bos taurus*) estimated by 16S rDNA homology analyses.** *Anaerobe* 2005, **11:**207-215.
454. Li RW, Connor EE, Li C, Baldwin V, Ransom L, Sparks ME: **Characterization of the rumen microbiota of pre-ruminant calves using metagenomic tools.** *Environ Microbiol* 2012, **14:**129-139.

455. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.** *DNA Res* 2007, **14:**169-181.
456. Zhu L, Wu Q, Dai J, Zhang S, Wei F: **Evidence of cellulose metabolism by the giant panda gut microbiome.** *Proc Natl Acad Sci U S A* 2011, **108:**17714-17719.
457. Silhavy TJ, Kahne D, Walker S: **The bacterial cell envelope.** *Cold Spring Harb Perspect Biol* 2010, **2**.
458. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP: **A core gut microbiome in obese and lean twins.** *Nature* 2008, **457:**480-484.
459. Wang L, Hatem A, Catalyurek UV, Morrison M, Yu Z: **Metagenomic insights into the carbohydrate-active enzymes carried by the microorganisms adhering to solid digesta in the rumen of cows.** *PLOS ONE* 2013, **8:**e78507.
460. Woodward R, Yi W, Li L, Zhao G, Eguchi H, Sridhar PR, Guo H, Song JK, Motari E, Cai L: ***In vitro*** **bacterial polysaccharide biosynthesis: defining the functions of Wzy and Wzz.** *Nat Chem Biol* 2010, **6:**418-423.
461. Jindou S, Borovok I, Rincon MT, Flint HJ, Antonopoulos DA, Berg ME, White BA, Bayer EA, Lamed R: **Conservation and divergence in cellulosome architecture between two strains of *Ruminococcus flavefaciens*.** *J Bacteriol* 2006, **188:**7971-7976.
462. Duan CJ, Xian L, Zhao GC, Feng Y, Pang H, Bai XL, Tang JL, Ma QS, Feng JX: **Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens.** *J Appl Microbiol* 2009, **107:**245-256.
463. Tomme P, Creagh AL, Kilburn DG, Haynes CA: **Interaction of polysaccharides with the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC. Binding specificity and calorimetric analysis.** *Biochemistry* 1996, **35:**13885-13894.
464. Ponyi T, Szabó L, Nagy T, Orosz L, Simpson PJ, Williamson MP, Gilbert HJ: **Trp22, Trp24, and Tyr8 play a pivotal role in the binding of the family 10 cellulose-binding module from *Pseudomonas* xylanase A to insoluble ligands.** *Biochemistry* 2000, **39:**985-991.
465. Bolam DN, Xie H, White P, Simpson PJ, Hancock SM, Williamson MP, Gilbert HJ: **Evidence for synergy between family 2b carbohydrate binding modules in *Cellulomonas fimi* xylanase 11A.** *Biochemistry* 2001, **40:**2468-2477.
466. Boraston AB, Ficko-Blean E, Healey M: **Carbohydrate recognition by a large sialidase toxin from *Clostridium perfringens*.** *Biochemistry* 2007, **46:**11352-11360.
467. CAZy database [**http://www.cazy.org/**]
468. O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, Damm U, Buiate EA, Epstein L, Alkan N: **Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses.** *Nature Genet* 2012, **44:**1060-1065.
469. Lamed R, Morag E, Mor-Yosef O, Bayer EA: **Cellulosome-like entities in *Bacteroides cellulosolvens*.** *Curr Microbiol* 1991, **22:**27-33.
470. O'Brien PM, Aitken R (Eds.): **Antibody phage display: methods and protocols**. Totowa, New Jersey: Humana Press Inc.; 2002.
471. Pereira SF, Goss L, Dworkin J: **Eukaryote-like serine/threonine kinases and phosphatases in bacteria.** *Microbiol Mol Biol Rev* 2011, **75:**192-212.
472. Rubin EM: **Genomics of cellulosic biofuels.** *Nature* 2008, **454:**841-845.