Identification and characterisation of novel glycan-binding bacterial adhesins encoded by the human gut microbial metagenome.

By:

Christy Agbavwe, BSc, MSc

A thesis submitted to the National University of Ireland, Galway for the degree of Doctor of Philosophy (Ph.D)

Discipline of Microbiology, School of Natural Sciences, College of Science, National University of Ireland, Galway

March 2017

Supervisors of Research: Dr. Aoife Boyd
Dr. Conor O'Byrne
Professor Lokesh Joshi

Table of Contents

Abstract i
Acknowledgements iii
List of Abbreviations iv
List of Figures vii
Chapter 1 Introduction: ix
Background 1
1.1 Anatomy and physiology of the human gastrointestinal tract 2
1.2 Bacterial colonisation of the GI tract and its diversity 3
1.2.1 Bacteroidetes 5
1.2.2 Firmicutes 6
1.3 Factors affecting the GI tract bacterial population 6
1.4 Glycans 8
1.4.1 Bacterial glycans and lectins 9
1.4.2 Glycan metabolism shapes the human gut microbiota 10
1.4.3 Glycans as legislators of host-microbial interactions 14
1.4.4 Mucus-binding proteins 16
1.4.4 Mucins 16
1.4.6 Mucin Glycosylation 18
1.5 Bacterial adhesins in host-microbe interaction 19
1.6 Mechanisms of bacterial adherence to host cells 20
1.6.1 P pili and Type 1 pili 21
1.6.2 Type IV pili 21
1.6.3 Curli 22
1.7 Bacterial adhesion in the human gastrointestinal tract 22
1.7.1 Surface-layer proteins as adhesins 23
1.8 Metagenomics 25
1.8.1 Sequence-driven analysis 25
1.8.2 Function-driven analysis 26
1.9 Functional screening of metagenomics libraries 28
1.10 Carbohydrate-based Microarrays 30
1.10.1 Mucin Microarray 30

1.10.2 Neoglycoconjugate Microarray (NGC) 31
1.11 Objective of the study 32
Chapter 2 Materials & Methods: 34
2.1 General microbiological techniques 35
2.1.1 Bacterial strains and plasmids 35
2.1.2 Culture media 36
2.1.2a Brain Heart Infusion (BHI) 37
2.1.2b Luria-bertani (LB) 37
2.1.2c M17 medium 37
2.1.2d Modified M17 medium (mGM17) 37
2.1.3 Media supplements 37
2.1.3a Antibiotics 37
2.1.4 Bacterial growth conditions. 38
2.1.4a General bacterial growth conditions. 38
2.1.4b Bacterial growth conditions for co-incubations. 38
2.1.5 Caco-2 cell culture conditions. 38
2.1.6 DNA agarose gel electrophoresis 39
2.1.7 Polymerase chain reaction (PCR). 39
2.1.8 Plasmid miniprep. 40
2.1.9 DNA purification using Wizard SV Gel/PCR Cleanup kit (Promega). 40
2.1.10 TOPO TA cloning of PCR products into pCR-XL-TOPO 40
2.1.11 Restriction endonuclease digestion. 41
2.1.12 Phosphatase treatment of vector DNA. 41
2.1.13 Ligations 41
2.1.14 Biofilm Assay 41
2.2 Preparation and transformation of competent cells 42
2.2.1 Preparation of electrocompetent cells of *Lactococcus lactis* 42
2.2.2 Transformation of *L. lactis* by high voltage electroporation 42
2.2.3 Preparation of electrocompetent cells of *Escherichia coli* 43
2.2.4 Transformation of *E. coli* by high voltage electroporation 43
2.3 Methods to evaluate adherence efficiency 43
2.3.1 Analysis of bacterial adherence 43
2.3.1a Enumeration of adherence efficiency 43

2.4 Metagenomics library preparation and selection methods..... 44
2.4.1 Fosmid library preparation..... 44
2.4.2 Shearing the metagenomics insert DNA..... 45
2.4.3 End-Repair of the metagenomic insert DNA..... 45
2.4.4 Size-Selection of the End-Repaired DNA..... 45
2.4.5 Recovery of the Size-Fractionated DNA..... 46
2.4.6 Packaging of CopyControl Fosmid Clone..... 47
2.4.7 Storage of metagenomics library in *E. coli*..... 47
2.4.8 Next Generation Sequencing..... 47
2.4.9 Bioinformatic analysis of clones..... 48
2.4.10 Selection of adherent library clones using a single round of selection..... 48
2.4.11 Selection of adherent library clones using multiple rounds of selection..... 48
2.5 Carbohydrate-based microarray characterization of putative adherent clones..... 53
2.5.1 Materials..... 53
2.5.2 Lectin microarray..... 54
2.5.3 Mucin microarray..... 54
2.5.4 Neoglycoconjugate (NGC) microarray..... 54
2.5.5 Preparation of bacterial fosmid clones for array analysis..... 55
2.5.6 Carbohydrate-based microarray data extraction and analysis..... 56
2.6 Expression of $MapA_{Ri}$ gene in *Lactococcus lactis* NICE system..... 56
2.6.1 PCR of $mapA_{Ri}$ gene..... 56
2.6.2 Construction of the recombinant NZ9000/pPTPi-$mapA_{Ri}$..... 56
2.6.3 Expression of $MapA_{Ri}$ protein in recombinant *L. lactis*..... 56
Chapter 3 Functional metagenomics approach to identify novel glycan binding bacterial adhesins encoded by the human gut metagenome..... 58
3.1 Introduction..... 59
3.1.1 Fosmid Metagenomic Library Construction..... 63
3.1.2 Small Fragment Metagenomic Library Construction..... 66
3.1.3 Transformation of small fragment library into *Lactococcus lactis*..... 67
3.2 Validation of two metagenomic DNA libraries..... 69
3.2.1 Characterization of microbial diversity of small fragment library..... 70
3.2.2 Evaluation of fosmid library diversity by restriction digestion..... 73
3.2.3 Characterization of microbial diversity of fosmid library..... 74
3.3 *In vitro* assay of bacterial adhesion onto mammalian epithelial cells..... 76

3.3.1 Analysis of Fosmid Library vs. EPI300 control strain without induction 78
3.3.2 Induction of the fosmid library does not increase adherence efficiency 79
3.3.3 Enrichment of adhesive clones of both metagenomic DNA libraries..... 81
3.3.4 Optimization of Control strains 86
3.3.5 Analysis of individual fosmid clones on 7 day old Caco-2 cells..... 88
3.3.6 Analysis of individual fosmid clones on 7 day old Caco-2 cells with..... 89
3.3.7 Analysis of individual fosmid clones on 3 week-old Caco-2 cells. 91
3.3.8 Analysis of individual fosmid clones on 3 week old Caco-2 cells 92
3.4 Discussion 95
Chapter 4 Characterization and bioinformatic analysis of adhesive cles identified by functional metagenomics 104
4.1 Introducti 105
4.1.1 Next neration Sequencing 105
4.1.2 NGC Chemistry 106
4.1.3 Pair-End Sequencing 107
4.1.4 Benefits of Next Generation Sequencing, NGS 107
4.2 Carbohydrate-based microarrays 108
4.2.1 Lectin Microarray Technology 109
4.2.2 Mucin Microarray 111
4.2.3 Neoglycoconjugate Microarray (NGC) 113
4.2.4 Advantages of Neoglycoconjugates 115
4.3 Biofilm Assay 116
RESULTS 118
4.4 Analysis of six putative adhesive fosmid clones (FC3, FC18, FC19, FC20, FC21 and FC22) 119
4.5 Next Generation Sequencing Data of six Fosmid clones 123
4.6 Bioinformatics Analysis of fosmid Clones FC3 and FC21 130
4.7 Rescue and re-transformation of fomid clones into *E. coli* host. 141
4.8 Comparison of fluorescent dye uptake 142
4.9 Lectin microarray results 144
4.10 Biofilm formation by gut metagenomic fosmid selected clones (FC21 & FC3) with adhesive capability. 147
4.11 Mucin Microarray results 149
4.12 Neoglycoconjugate Microarray Results 152

4.13 Discussion 164

Chapter 5 *In silico* analysis of the human gut metagenome identifies a putative bacterial adhesin ($MapA_{Ri}$) encoded by *Roseburia intestinalis*: 176

5.1 Introduction 177

5.2 Adherence factors in Lactobacillus 177

5.2.1 Extracellular mucus-binding protein, Mub 178

5.2.2 Lectin-like Mannose Specific Adhesin, Msa 180

5.2.3 Lactobacillus surface protein A, LspA 181

5.2.4 Starch binding proteins, SusD & SusC 181

5.2.5 Mucus adhesion promoting protein, MapA 183

5.3 Roseburia intestinalis 185

5.4 NICE, Nisin controlled gene expression system for *Lactococcus lactis* 186

5.5 Homologous sequences search using BLAST 189

Result 194

5.6 Mining human gut metagenomics database using 5 reference adhesins 195

5.7 BLAST search against 54 individual bacterial genomes using five reference adhesins. 195

5.7.1 BLAST homology search using MapA reference protein 196

5.7.2 BLAST homology search using Msa reference protein 205

5.7.3 BLAST homology search using Mub reference protein 205

5.7.4 BLAST homology search using SusD reference protein 206

5.7.5 BLAST homology search using LspA reference protein 207

5.7.6 Overview 207

5.8 *In silico* analysis, amplification and cloning of a putative MapA homolog, $MapA_{Ri}$, a putative L-Cystine ABC transporter from *Roseburia intestinalis* 208

5.9 *In Vitro* Adhesion Assays of nisin induced expressed (NICE) *L. lactis* NZ9000/pPTPi-$MapA_{Ri}$ recombinant clone. 213

5.10 Discussion 215

Chapter 6 General Discussion 220

6.1 Summary of Main findings 221

6.1.1 Functional screening of a metagenomic library 223

6.1.2 Cell surface glycosylation is altered for FC3 and FC21 226

6.1.3 Mucin binding signature of FC3 and FC21. 229

6.1.4 Identification of glycan-binding interactions of FC3 and FC21 230

6.1.5 No change in biofilm formation for FC3, FC21 and control strain 232
6.1.6 Adherence of *L. lactis* NZ9000/pPTPi-*mapA*$_{Ri}$ to Caco-2 cells 232
6.1.7 Advances in knowledge of gut-microbe interactions..... 233
6.2 The Challenges: Metagenomic Libraries 233
6.2.1 Leveraging existing libraries..... 235
6.2.2 Strategies to improve heterologous expression..... 238
6.3 The issue of *in vitro* adhesion 239
6.4 Carbohydrate-based Microarrays..... 241
6.5 Future work..... 242
6.5.1 Functional Screens 242
6.5.2 Diversity of gut microbiota and “omics” technology 242
6.5.3 Lectin binding signature of mammalian Caco-2 cells 243
6.6 Conclusion 245
References..... 247
Appendix..... 278

Abstract

Although there have been an increasing number of scientific publications describing the adherence of gut bacteria to components of the human intestinal mucosa, very little is known about the surface molecules mediating this adhesion and their individual receptors. In the current study, we describe the identification and subsequent analysis of putative glycan-binding factors from the human gut microbiome using a combination of functional metagenomics and bioinformatics-based approaches. A fosmid library of human gut microbiota in the surrogate host Phage T-1 Resistant EPI300™-T1^{R} *Escherichia coli* was constructed and screened for enhanced adherence capability. Two out of 42,000 fosmid clones, FC3 and FC21, exhibited enhanced adherence to Caco-2 cells in functional screens. DNA segments inserted into the FC3 and FC21 clones were distinct sequences of 24.6 kb and 8.1 kb, respectively. FC21 contained three functional genes and likely originates from the dominant commensal gut species *Bifidobacterium adolescentis*. Sequence analysis of FC3 revealed that the 24.6 kb insert is a fragment with no current known homologs in the database, suggesting that the insert DNA is derived from a microbe with an unknown genome sequence. When carbohydrate-based and lectin microarrays were used to characterize the carbohydrate binding specificities of FC3 and FC21, the lectin microarrays revealed that the host *E. coli* strains carrying FC3 and FC21 had altered cell surface glycosylation, while the mucin microarrays revealed that mucin binding pattern is not altered for the two fosmid clones as compared to the control, and finally neo-glycoconjugate microarrays revealed that FC3 and FC21 exhibited binding to specific glycans in the presence of arabinose and antibiotic. Five adhesins known to mediate adhesion to specific components of the human gastrointestinal tract were identified from the literature and used as reference proteins in BLASTp searches against the genomes of 54 of the most abundant species in the gut. A homologous protein to one of the reference adhesins, MapA$_{Ri}$, was subsequently identified in the gram-positive, butyrate-producing bacterium *Roseburia intestinalis M50/1*. MapA$_{Ri}$ was cloned into a *Lactococcus lactis* heterologous host and assessed for expression using the nisin controlled gene expression system (NICE)[1] and functionally screened to detect the adherence phenotype. The results indicate that induction of the recombinant strain *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ with nisin does not significantly increase its adherence to 7 day-old or 3 week-old Caco-2 cells suggesting that (a) MapA$_{Ri}$ does not function

as an adhesin, or (b) $MapA_{Ri}$ is not being expressed appropriately by the *L. lactis* NZ9000 heterologous host. The findings in this study demonstrate the power of functional screening but also raises significant questions about the usefulness of this approach and the sequence-based metagenomic approach in identifying glycan binding determinants encoded by the human gut metagenome. Identification of novel glycan binding genes which have not previously been linked to adhesion will help to broaden our understanding of host-microbe interaction and possibly lead to the identification of novel and unusual systems that play as yet undefined roles in adherence and perhaps ultimately in human gastrointestinal health.

Acknowledgements

Firstly, I am very grateful to my supervisors **Dr. Aoife Boyd**, **Dr. Conor O'Byrne** and **Professor Lokesh Joshi** for giving me the opportunity to take on this project and for their guidance and support throughout the years of my PhD. I would also like to extend my thanks to members of my GRC committee, **Dr. Cathal Seoighe** and **Dr. Gerard Wall** for their helpful advice and constructive criticisms throughout.

My sincere thanks to the technicians **Mike**, **Maurice**, **Ann** and **Katrina**. Thank you for your smiles and encouragement. I would like to say thank you to **Kityee**, **Dr. Estefanía Porca Belío** and **Sarah** for the coffee dates, laughter and venting sessions. Thank you to **Dr. Michelle Kilcoyne** and **Andrea Flannery** for all your help with microarray analysis. I would also like to thank **Dr. Kimon-Andreas Karatzas** who encouraged me during the early years of my PhD.

A very special thank you to my brothers **Otahre** and **Daniel** for all the jokes, laughter, love and patience. Finally, but by no means least, thanks goes to my amazing parents, **Rosemary** and **Charles**. Migwo Dad! Thank you for providing me with so many opportunities in life. Thank you Mom, you were by my side throughout this PhD, living every single minute of it. Without you, I would not have had the courage to embark on this journey and ultimately see this project through to the end. Thank you for your unwavering support, endless encouragement and unconditional love. I dedicate this thesis to you.

"It always seems impossible until it's done."

~ Nelson Mandela

List of Abbreviations

aa – Amino acid

Amp – Ampicillin

BAC – Bacterial artificial chromosome

BHI – Brain-heart infusion medium

BLAST - Basic Local Alignment Search Tool

BSA - Bovine serum albumin

CD – Crohn's disease

CFU – Colony forming units

CGH – Comparative genomic hybridisation

CHO – Chinese Hamster Ovary

Cm – Chloramphenicol

dH2O - Distilled water

DNA - Deoxyribonucleic acid

DMSO – Dimethyl sulfoxide

Ery – Erythromycin

EPO - Erythropoietin

EPEC – Enteropathogenic *E.coli*

F – Forward

FISH – Fluorescence *in situ* hybridization

FBS - Fetal bovine serum

FOS – Fructo-oligosaccharide

g – Gram

g – Gravity force

GIT – Gastrointestinal tract

GlcNac - N-acetyl-D-glucosamine

GalNAc - N-acetyl-D-galactosamine

h – Hour

HCl – Hydrochloric acid

HCE – Hierarchical clustering explorer

HPLC – High performance liquid chromatography

HMO – human milk oligosaccharides

HTS – High throughput sequencing technologies

HMW – high molecular weight

IBD – Inflammatory Bowel Disease

kb – Kilobase

kDa - kilodaltons

Kan - Kanamycin

LPxTG – C-terminal anchoring motif

LB – Luria Bertani

M – Molar

MapA – Mucus adhesion promoting protein

MRSA – methicillin-resistant staphylococcus aureus

mg – Milligram

min – Minute

ml – Millilitre

mM – Millimolar

mub – Mucus binding protein

µm – Micrometre

µF – Micro farad

µg – Microgram

MOI – Multiplicity of Infection

Msa – Lectin-like mannose specific adhesin

MSC – Multiple cloning site

NCBI – National Centre for Biotechnology Information

NSSF – No significant similarity found

NGC - Neoglycoconjugate

ng – Nanogram

OD - Optical density

ORF - Open reading frame

Pap – Pyelonephritis-associated pili

PenStrep - Penicillin streptomycin

PCR – Polymerase chain reaction

R – Reverse

rpm - Revolutions per minute

RBS – Ribosomal binding site

RNA – Ribonucleic acid

RT – Room temperature

SAP - Shrimp alkaline phosphatase

SrtA – sortase enzyme A

SCFA – Short chain fatty acids

SDS-PAGE - Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SUS – starch utilization system

s – Second

TrfA – plasmid replication initiator protein

TPA – tissue plasminogen activator

U – Units

UPEC – Uropathogenic *E.coli*

UC – Ulcerative colitis

V – Volt

v/v – Volume per volume

w/v – Weight per volume

List of Figures

Figure 1.1 Overview of the human gut……..……..3
Figure 1.2 Binding of bacterial glycans (carbohydrates) to host glycan binding proteins (lectins)…….…..9
Figure 1.3 Sources and chemical variation of glycans in the gut……..13
Figure 1.4 Diagram depicting representation of the components of the human intestinal mucosa and submucosa……...…23
Figure 1.5 Representation of the cell wall of a Gram-positive bacterium.....24
Figure 1.6 An overview of processes involved in the production of a Metagenomic library…..…....……..27
Figure 3.2 Schematic representation of formed cloning procedure as represented by Epicentre…….……63
Figure 3.3 pCC1FOS vector map……..…65
Figure 3.4 pTRKL2 vector map……...67
Figure 3.5 Evaluation of the metagenomic diversity of the small fragment library by restriction digest using the endonuclease enzyme *Eco*RI……...70
Figure 3.6 Evaluation of formed library metagenomic diversity by restriction digest of 7 random fosmid clones……...…73
Figure 3.7 *Vibrio parahaemolyticus* serves as a positive adherence control as it has been found to adhere highly to Caco-2 cells……77
Figure 3.8 No significant difference was found between the fosmid library and EPI300 control strain (without empty fosmid)……....79
Figure 3.9 Fosmid library induction with arabinose…….……81
Figure 3.10 *In vitro* selection and enrichment of adhesive clones from the small fragment metagenomic library……83
Figure 3.11 *In vitro* selection and enrichment of adhesive clones from the induced fosmid metagenomic library……...…84
Figure 3.12 Irreproducibility of highly adhesive clones on separate days……..86
Figure 3.13 EPI300 and EPI300 (pCC1FOS) display similar adherence levels…..87
Figure 3.14 Analysis of individual fosmid clones for their ability to adhere to 7 day old Caco-2 cells without arabinose induction…….89
Figure 3.15 Analysis of individual fosmid clones for their ability to adhere to 7 day old Caco-2 cells with arabinose induction……...90
Figure 3.16 Analysis of individual fosmid clones for their ability to adhere to 3-week old Caco-2 cells with arabinose induction……..92
Figure 3.17 Analysis of individual fosmid clones for their ability to adhere to 3 week old Caco-2 cells without arabinose induction…….94
Figure 4.1 The experimental setup to analyze bacterial glycosylation using the lectin microarray……...109
Figure 4.2 The image illustrates a miniature version of the lectin microarray…110
Figure 4.3 1% Agarose gel electrophoresis illustrating *Bam*HI restriction profiles of 6 fosmid clones…….120
Figure 4.4 Flowchart depicting the stepwise *de novo* assembly process…….124
Figure 4.5 ORF (Open Reading Frame) map of FC21 fosmid clone……...130
Figure 4.6 ORF (Open Reading Frame) map of FC3 fosmid clone…….134
Figure 4.7 Re-transformation of fosmid clones into *Escherichia coli* host.....142
Figure 4.9 Lectin microarray profile of FC3, FC21 and EPI300 control strain...145

Figure 4.10 Biofilm formation of FC3 and FC21 clones compared to the EPI300 control strain……148
Figure 4.11 Histogram representing mucin microarray profile of FC3, FC21 and EPI300 control strain……151
Figure 4.12 Histogram of neoglycoconjugate microarray profile of FC3with and without arabinose and antibiotic……...154
Figure 4.13 Histogram representing the neo-glycoconjugate microarray profile of FC21 in the absence and presence of arabinose and antibiotic……...157
Figure 4.14 Neo-glycoconjugate microarray profile of EPI300 (pCC1FOS) control strain in presence and absence of arabinose and antibiotic……159
Figure 4.15 Neo-glycoconjugate microarray profile of FC3, FC21 and EPI300 control strain in presence of antibiotic and arabinose……160
Figure 4.16 Histogram representing the differences in recognition of neo-glycoconjugates and glycoproteins by fluorescents-labelled bacterial strains FC3, EPI300 and FC21……..162
Figure 4.17 Clustering analysis of neoglycoconjugate triplicate data.....163
Figure 5.1 Domain architecture of the fully characterized LspA, Msa, and Mub lactobacilli adhesin proteins according to Pfam database……..180
Figure 5.2 Cluster of starch utilization genes……182
Figure 5.5 Amino acid sequence and domain architecture of the MapAprotein.185
Figure 5.6 pPTPi vector map……..189
Figure 5.7 Amino acid sequence alignment of the MapA protein from *Lactobacillus reuteri* 104R and the $MapA_{Ri}$ protein from *Roseburia intestinalis* M50/1……199
Figure 5.13 $mapA_{Ri}$ primer design……..……209
Figure 5.14 PCR amplification of $mapA_{Ri}$ from *Roseburia intestinalis* DNA…..210
Figure 5.15 Diagram of TOPO cloning vector…….……210
Figure 5.16 Restriction digestion of pTOPO::$MapA_{Ri}$…….…211
Figure 5.17 1% Agarose gel electrophoresis……..…..212
Figure 5.18 *In vitro* adhesion assay of recombinant *L. lactis* NZ9000/ pPTPi-MapA*Ri* on 7 day old Caco-2 cells……...…..214
Figure 5.19 *In vitro* assay of recombinant *L. lactis* NZ9000/ pPTPi-$MapA_{Ri}$ on 3 week old Caco-2 cells……...…215

Chapter 1:

Introduction

Background

The human gut microbiota is a complex, dynamic and diverse community which plays an important role in human health, nutrition, metabolism, immune function and physiology[2]. This field has become the subject of numerous and extensive research in recent years and our knowledge of the resident microbes and the full scope of their capacity is rapidly growing. It is estimated that the human gut harbours a complex community of over 100 trillion microbial cells with about 1000 bacterial species and 100 fold more genes than are found in the human genome[3]. This large community has been dubbed the hidden metabolic "organ" due to their immense impact on human well-being. It has now been established that our gut microbiome coevolved with us and that changes in the composition of the population can have major consequences, both beneficial and harmful, for human health[4]. Indeed, disruption of the gut microbial homeostasis (or dysbiosis) has been linked to diseases such as inflammatory bowel disease (IBD), diabetes, Crohn's disease (CD), Ulcerative colitis (UC) and obesity[5]. Gut homeostasis depends on a number of factors, particularly diet, which was shown to influence microbial composition and their metabolic activities. One of the main functions of the human gut microbiota in the healthy state is to degrade dietary carbohydrates that escaped digestion in the upper gastrointestinal tract (GI tract)[6]. Thus, the gut microbiota contributes nutrients and energy to the host via the fermentation of non-digestible dietary components in the large intestine.

Understanding the composition and functional capacity of the gut microbiome constitutes an enormous challenge. Fortunately, the role of the human gut microbiota in health and disease is becoming evident due to the advent of high throughput sequencing technologies (HTS) and other similar technologies[7]. Metagenomics provides insight into the genetic potential of various microbial communities and is able to identify novel proteins and biomolecules which can find application in industry, medicine and science[8]. Simple and complex glycans have long been known to play major metabolic, structural and physical roles in biological systems[9]. They are able to mediate a wide variety of biological roles due to their mass, shape, charge and other physical properties. Targeted microbial binding to host glycans has also been studied for decades[9]. The capacity of the human gut microbiota to adhere to the glycans on the surface of the gastrointestinal epithelial cells can be examined by using

functional metagenomic approaches, which can provide new information on novel glycan binding bacterial adhesins. Therefore, this thesis is devoted to investigating and characterizing novel glycan binding bacterial determinants encoded by the human gut metagenome.

1.1 Anatomy and physiology of the human gastrointestinal tract

The human gastrointestinal tract (GIT) is defined as an organ system that is responsible for transporting, digesting and absorbing consumed nutrients and discharging waste[10] (Figure 1.1). The GIT is estimated to be about 9 metres (30 feet) long and is divided into an upper and lower gastrointestinal tracts from the mouth to the anus[11]. The human GIT begins with the oral cavity where consumed food is mechanically digested and moistened by secreted saliva. The salivary glands secrete saliva which contains enzymes (e.g. amylase and lipase) that breakdown the dietary components in the oral cavity[12]. The chewed pieces of food particles are then swallowed via the oesophagus and down to the stomach through the peristaltic contraction of muscles. In the stomach, the masticated food is retained and further digested by protein-degrading enzymes and hydrochloric acid (HCl) before being sent to the small intestine. Protein components of the masticated food in the stomach are digested with enzymes, e.g. pepsin, which are activated at the low pH levels created by the presence of HCl. HCl also serves to destroy microorganisms ingested with the food. The stomach is connected to the small intestines by the duodenum which in turn is linked to the pancreas and the liver via the biliary tract. The pancreas produces pancreatic juice which consists of digestive enzymes such as trypsinogen, chymotrypsinogen, pancreatic lipase and amylase. The next part of the small intestine is the jejunum where the vast majority of absorption of nutrients takes place. The small intestine is also lined with microvilli which greatly enhance absorption of digested food. The last part of the small intestine that is connected to the large intestine is known as the ileum. Finally, the last segment of the lower GIT is the large intestine which is divided into the caecum, colon, rectum and anus. There are four sections that make up the colon; namely the ascending colon, the transverse colon, the descending colon, and the sigmoid colon. The colon is the central site of microbial colonisation and microbial activity including digestion of dietary components. Undigested and unabsorbed food residues are removed from the body through the process of defecation[13].

An important component of the GIT is the mucus layer which covers the epithelial cells of the stomach, small intestine and colon. The mucus is composed primarily of specific families of glycoproteins termed mucins and is generally very viscous. It serves to protect and lubricate the inner mucosa of the tract. It is composed of a well-defined outer "loose" layer and an inner "firm" layer attached to the epithelium of the stomach and large intestines[14]. The outer "loose" mucus layer contains an enormous number of bacteria, however, the firm inner layer is resistant to bacterial penetration and protects the epithelial cells from direct contact with bacteria[14]. Scientists have shown that mice lacking the glycoprotein MUC2 (mucin) secreted by specialised epithelial cells called goblets cells suffer spontaneous inflammation, highlighting the critical role of the mucosal barrier in host-bacteria homeostasis[15, 16].

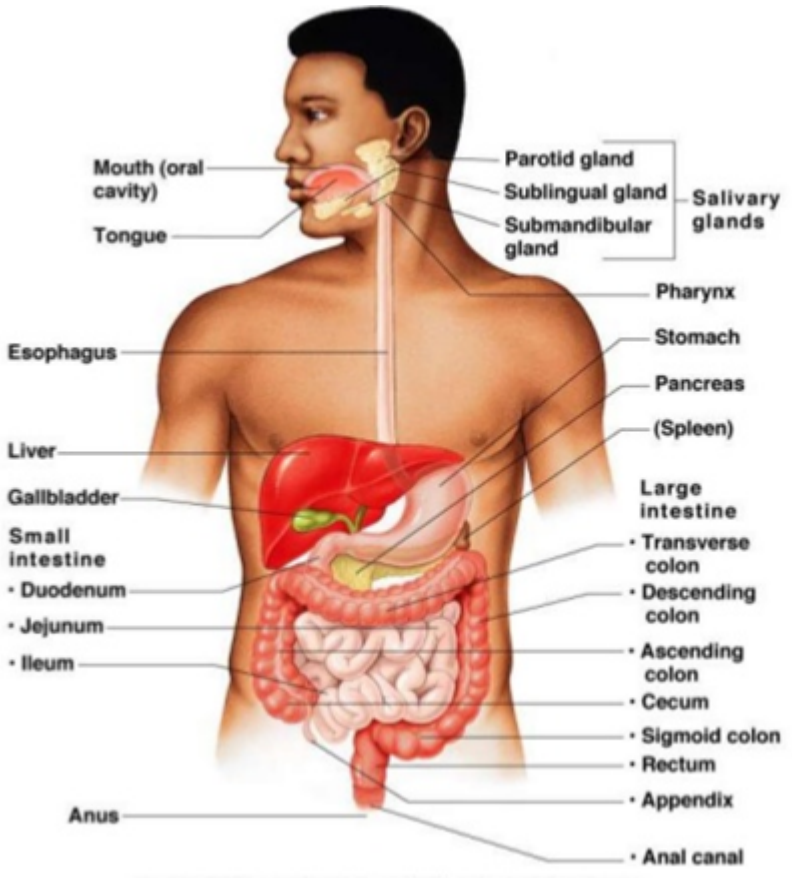


Figure 1.1 Overview of the human GI tract
Adapted from Pearson Education Inc., (2009)

1.2 Bacterial colonisation of the GI tract and its diversity

The human gastrointestinal tract has been shown to be relatively sterile *in utero*[17]. However, bacterial colonisation occurs immediately after birth and passage through

the birth canal[18]. Handling and feeding of the infant after birth leads to the establishment of a stable normal flora on the skin, oral cavity and gastrointestinal tract in approximately 48 hours after birth[19]. Infants delivered vaginally often acquire a microbiota that is similar to the mother's vaginal microbial community. The microbiota of babies delivered via Caesarean section resembles the general skin microbial population of their mothers[19]. Another factor that impacts the composition of the microbiota is the infant's feeding regime[17] that differs significantly between breast-fed infants and formula-fed infants[20]. Usually, after weaning, a more diverse and complex community becomes established in the infant GIT which resembles adult individuals[21, 22]. The bacterial distribution along the GIT increases from the upper GIT to the lower GIT. The stomach contains few bacteria due to the harsh and highly acidic environment. The jejunum has been estimated to contain approximately 10^5 CFU ml^{-1}. This increased number of bacteria is due to the higher pH, larger volume and slower peristaltic movements (longer retention time)[23] in the jejunum.

Studies have estimated that the human gut microbiota is composed of 1000 different species[24]. Enumeration and characterisation of cultured organisms is nowadays complemented with molecular profiling methods such as microarrays, quantitative PCR, high-throughput sequencing and microbial 16S rRNA[25]. These techniques have shed light on the composition and diversity of the predominant bacteria in the gut[26, 27]. It has been estimated that each individual carries at least 160 different bacterial species[28]. The predominant bacteria from the human gut belong to the phylum Bacteroidetes and to the low % G+C Firmicutes[27, 29, 30].

In the past, the gut microbiota was considered difficult to culture as 93% of the human gut bacterial 16S rRNA sequences corresponded to uncultured bacteria[31]. However, in 2011, Alan Walker and colleagues demonstrated that the most abundant phylotypes (>2%) are cultured at nearly 100%, which suggests that the majority of gut bacteria can be grown under laboratory conditions[30]. A similar conclusion was drawn by Goodman *et al.* (2011)[32]. The relative abundance of readily cultured phylotypes was estimated, followed by 16S rRNA analysis of complete faecal samples from healthy volunteers and was compared to the data derived from cultured samples. The results indicated that culturability was correlated with the taxonomic level. At the family-

level 89% phylotypes were readily cultured but at the species-level the proportion of cultured bacteria decreased to 56%.

1.2.1 Bacteroidetes

The phylum Bacteroidetes consists of three classes of Gram-negative, non-spore forming, anaerobic or aerobic, and rod shaped bacteria distributed widely in environments such as sea water, sediments, soil and the human gastrointestinal tract. The three classes include; Bacteroidia, Cytophagia and Flavobacteria. The Bacteroida class is often associated with the human gut microbiota and consists of several families. The members of the Bacteroidaceae are most frequently represented as part of the human gut microbiome. They are Gram-negative, pleomorphic, anaerobic bacteria that make up approximately 15-25% of the human colonic microbiota. They are known for their capacity to metabolize carbohydrate substrates[33, 34, 35] and formation of short chain fatty acids (SCFA) including lactate, succinate, acetate, formate and propionate as end products of bacterial fermentation. Studies have shown that some *Bacteroides* species are able to convert bile to metabolites, which are considered as co-carcinogens or mutagens[36]. As a results, species such as *B. vulgatus* and *B. stercoris* have been implicated with conferring a higher risk of colon cancer[37, 38]. A study by Sobhani and colleagues demonstrated that the bacterial diversity of colorectal cancer patients showed significantly higher levels of *Bacteroides/ Prevotella* than the controls[37]. The most widely studied member of this phylum is the *Bacteroides thetaiotaomicron* whose genome was sequenced by Xu and colleagues in 2003[39]. *B. thetaiotaomicron* is a prominent gut isolate that is able to degrade dietary glycans. *B. thetaiotaomicron* is adapted to a carbohydrate rich environment by the presence of multiple gene clusters in its genome that includes a multi-protein starch utilisation system (SUS)[39]. The SUS system enables the bacterium to efficiently bind and degrade starch[40].

1.2.2 Firmicutes

The Firmicutes are the most abundant phyla present in the human gut microbiota. They are mostly Gram-positive, low %G+C content bacteria that comprise approximately 60-70% of the colonic microbiota[32, 41, 27]. 16S rRNA analysis has demonstrated that Clostridia, Bacilli, Erysipelotrichi and Negativicutes classes are present in human faecal samples[41]. The Clostridia class is the most abundant and contains the order Clostridiales with the families Ruminococcaceae, Clostridiacae, Lachnospiraceae and Eubacteriaceae[41]. The order Clostridiales has been divided into several clostridial clusters on the basis of 16S rRNA sequencing[42] . The members of clostridial clusters IV and XIVa are the dominant groups in the human GI tract. Clostridium cluster IV is referred to as the *Clostridium leptum* group or Ruminococcaceae family with species such as: *Clostridium leptum, Cl. Sporosphaeroides, Faecalibacterium prausnitzii, Ruminococcus bromii, R. champanellensis* and *R. albus*. *F. prausnitzii* is a predominant species in this group which has been shown to be able to metabolize starch and inulin to form butyrate and D-lactate[43] and has also been shown to have anti-inflammatory properties based on studies in a colitis mouse model[44].

1.3 Factors affecting the GI tract bacterial population

The composition of the human gut microbiota can be influenced by dietary factors and lifestyle events (stress, ageing, disease, indigestible carbohydrates). To date, several diets, especially a Western lifestyle with a high consumption of meat and carbohydrates and a low consumption of vegetables, has been linked to common diseases such as atherosclerosis, inflammatory bowel syndrome, and colon cancer. The composition of the gut microbiota changes significantly from infancy through to adulthood and in the elderly. The first changes of the gut microbiota occur during early life with a decrease in the number of aerobes and facultative anaerobes and an increase in obligate anaerobic populations. Soon after the weaning of an infant, the gut microbiota gradually starts to resemble that of an adult's microbial community with the dominant bacteria from the phyla Firmicutes and Bacteroidetes[22]. In contrast to infants, significant changes occur in the gut microbiota of elderly people as their dietary habits and lifestyle changes. For example, elderly people are prone to reduced intestinal mobility, illness and medication treatment[45]. The ratio of Firmicutes vs. Bacteroidetes was discovered to be unusual in elderly individuals, with a higher

proportion of Bacteroidetes than in adults[46]. A significantly greater proportion of Enterobacteria was also found in elderly individuals compared to younger adults[47].

Diet is one of the most important environmental factors that has an impact on the microbial population in the GI tract. The dietary carbohydrates that do not get digested in the upper GI tract reach the colon and affect bacterial growth and their metabolic function. Researchers discovered that variations in the uptake of carbohydrates influenced the microbial composition and short-chain fatty acidy production (SCFA)[29]. A reduction in carbohydrate uptake often leads to a reduction in the number of butyrate-producing bacterial strains in the gut. Additionally, reduced carbohydrate consumption has been shown to influence bacterial homeostasis by increasing the pH of the colon. A higher colonic pH in turn reduces the population of butyrate-producers[48]. A good example of the effect of diet on gut microbiota was illustrated in research conducted comparing children from Burkina Faso and Europe. Scientists discovered that there was a higher ratio of Bacteroidetes in African children who consumed a fibre-rich diet[49]. Similarly, scientists observed that the proportion of Bacteroidetes in the faecal sample of obese subjects is significantly lower than their lean counterparts. The Bacteroidetes fraction increased when the obese humans were on a weight loss diet[50]. In fact, it has been proposed that the ratio of Bacteroidetes in a particular gut microbiota might serve as an obesity biomarker in the future.

Prebiotics are non-digestible food ingredients such as inulin and fructo-oligosaccharides (FOS) which are known to stimulate the growth of specific groups of gut bacteria. The supplementation of diet with prebiotics was shown to have a bifidogenic effect on infant, adult and elderly microbiota[51]. Additionally, supplementation of a diet with inulin stimulated *Faecalibacterium prausnitzii* species which are among the main butyrate-producers in the GI tract[52]. All in all, scientists were able to demonstrate that the modulation of energy metabolism and satiety was correlated with prebiotic supplementation and showed that food intake and glucose homeostasis can be regulated. Another important factor that influences gut microbiota composition is host genetics. Scientists have observed similar bacterial communities among related individuals such as twin siblings and their mother. Indeed, researchers demonstrated that the differences between monozygotic and dizygotic twins are not significant, thus the similarity in microbiota remains despite genetic differences

present in dizygotic twins[50]. Changes in the gut microbiota of each individual was shown to be short-lived and temporary, suggesting that each person has a stable and well-defined microbial core (three main enterotypes). As mentioned, diet is an important factor in influencing the microbial population of the GI tract. The following section will delve into the importance of glycans in the human gut microbiota.

1.4 Glycans

All cells from bacteria to human are covered in glycans [53]. Indeed, glycans decorate the surface of most organisms and living cells, creating a landscape of recognition sites and barriers that represent the first line of contact [54]. They have been shown to mediate the initial binding and recognition events of both immune cells and pathogens with their target cells and tissues[55]. The mucosal surface of the GI tract is the largest body surface in contact with the external environment that is covered in a total population of sugars and glycoconjugates. The surface of the intestinal epithelium is abundant with protein- and lipid-glycoconjugates which are important components of the intestinal mucosa. Glycans participate in almost every aspect of biology from sorting proteins to modulating cell differentiation and cell-cell interactions. Of the four fundamental building blocks (nucleic acid, amino acids, lipids, and glycans), glycans are the most diverse. The biochemistry of the various host and dietary glycans that enter the gut is exceptionally diverse[10].The modifications and biosynthesis of glycans are not dependent on template but are the result of multiple enzymatic activities. This dynamic ability to generate structural diversity in glycans facilitates the host's ability to accommodate rapid changes in the composition of the gut microbiota and the rapid evolution of individual species. Glycans are known as compounds consisting of a large number of monosaccharides linked glycosidically. They can be homo- or heteropolymers of monosaccharide residues, and can be linear or branched. Glycans can form linear or branched chains via an alpha or beta glycosidic linkages to any available hydroxyl of another monosaccharide[55]. These chains can be free (such as milk oligosaccharides), attached to proteins (glycoproteins, proteoglycans) or attached to lipids (glycolipids). The human genome only encodes a limited capacity to degrade glycans, typically those that contain one or two different linkages, namely starch, lactose and sucrose. In sharp contrast, the gut microbiota possesses the corresponding enzyme tools for depolymerizing complex glycan molecules into their component

sugars. The gut microbiota consists of species that are adept at foraging glycans and polysaccharides, including plant polysaccharides (starch, hemicellulose and pectin), animal-derived cartilage (glycosaminoglycans and N-linked glycans) and endogenous glycans from the host mucus (O-linked glycans)[10]. The next sections will explore bacterial glycans and lectins as well as the role of host and dietary glycans in shaping the human gut microbiota.

1.4.1 Bacterial glycans and lectins

The diverse carbohydrates on the surface of bacteria now serve to modulate numerous recognition processes in cell-cell interactions. At the bacteria-host level (Figure 1.2), bacterial glycans interact with host lectins to influence colonization and survival. Often the bacterial glycans are recognized as antigens by the host lectin of the innate and adaptive immune system [56]. Bacteria manipulate the immune response by exploiting the combinatorial potential of the carbohydrates on their surface [57]. They vary the epitopes present on their surface and sometimes mimic the epitopes present on the host surface. It is a well-studied phenomenon that pathogenic bacteria, such as hemolytic streptococci, often change their outer surface glycan profile to escape the host immune defense mechanism [58].

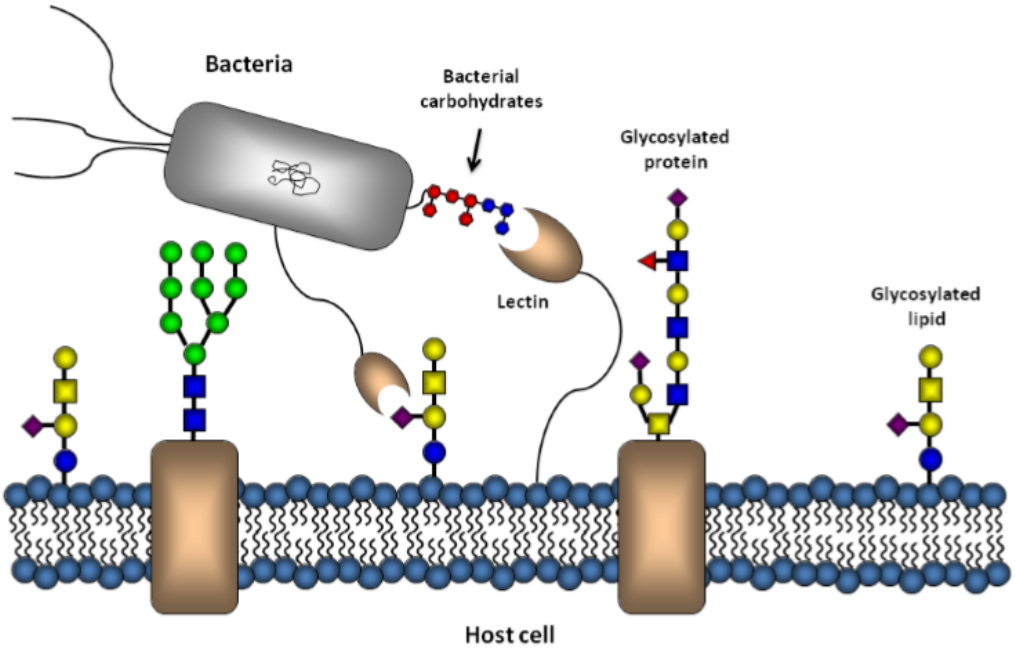


Figure 1.2 **Binding of bacterial glycans (carbohydrates) to host glycan binding proteins (lectins).** Bacterial glycans and host lectins play a critical role in mediating recognition during bacteria-host interactions. Bacterial glycans behave as antigens of the host immune response, whereas bacterial lectins effectuate attachment to host cells via recognition of host glycans. Diagram adapted from Ku-Lung Hsu, 2008[59].

In addition to producing surface glycans, bacteria are able to produce lectins on their surface which recognize glycans present on the host cell. Much like their surface glycans, bacteria can modulate the expression of their surface lectins producing cells that differ in their ability to bind host glycans [59]. Therefore, it can be said that in nature, the affinity and specificity of the bacteria-host recognition process is governed by the combined effect of multiple lectin-glycan & protein-protein interactions[58]. This suggests that analysis of the overall glycosylation pattern of a cell is paramount in understanding structure-function relationship of carbohydrates [59].

1.4.2 Glycan metabolism shapes the human gut microbiota

One of the major components that shape the composition and physiology of the human gut microbiota is the influx of glycans into the intestine from diet and host mucosal secretions[9] (Figure 1.3). Humans consume a wide variety of plant and animal-derived dietary glycans, most of which cannot be degraded by human encoded enzymes. Gut micro-organisms are able to produce a wide variety of enzymes that ferment dietary glycans into short chain fatty acids (SCFA) which serve as a nutrient source for the gut colonocytes and other epithelial cells[35]. Individual micro-organisms prefer different glycans. Therefore, selective consumption of glycans can influence the microbial composition that proliferate and persist in the human gut. This suggests that researchers can utilize dietary glycans as a non-invasive strategy to directly modulate the composition of bacterial species within the human gut[9, 60].

Specific members of the human gut microbiota that are able to degrade host glycans are found in mucus secretions[60]. The host endogenous glycans provide a constant source of nutrient for the microbiota especially during limited influx of dietary glycans. A large amount of host glycans are located in close proximity to the host tissue in the protective mucus layer. The ability of specific microorganisms to

penetrate and degrade mucus as a nutrient source allows them to exert an effect on colonic health, especially during a state of dysbiosis (abnormality in gut microbiota community composition)[61].

The process by which host and dietary glycans shape the gut microbiota is catalysed by changes in glycan availability from birth to adulthood. Studies indicate that the glycan composition of the human gut during the pre-and post-weaning period is an important factor that guides the establishment of the microbial community[62]. A variety of different glycan structures have been identified in human breast milk[63], namely lactose, glucose, galactose, *N*-acetylglucosamine, fucose, sialic acid and a mixture of complex human milk oligosaccharides (HMOs). HMOs are highly diverse glycans which have been shown to only be present in human breast milk[64] (Figure 1.3). Most HMOs are unable to be digested by human enzymes, which suggests that they evolved as natural prebiotics to guide the development of the infant gut microbiota by selectively feeding specific gut microbial species[65, 66, 67]. Higher proportions of Lactobacillus and Bifidobacterium are observed in infants that are exclusively fed breast milk, suggesting that they have co-evolved to occupy this niche. Consistent with this observation, researchers have shown that certain species of *Bifidobacterium* are able to directly metabolize HMOs[68, 69]. In contrast, formula-fed infants displayed lower abundance of Lactobacilli and Bifidobacteria while displaying an increased abundance of Clostridium, Bacteroides and Enterobacteriaceae[70, 71, 72]. These observations demonstrate that cow's milk-based formula selects for different micro-organisms during infancy and lacks the amount and diversity of oligosaccharides present in human milk[64].

When complex foods such as fruits, cereals and vegetables are introduced into the human infant diet, the composition of the microbiota shifts and microorganisms that prefer these glycans, such as the Gram-negative Bacteroidetes and Firmicutes become more prevalent[73, 71, 74]. Metagenomic studies have demonstrated the presence of genes for plant carbohydrate degradation prior to the introduction of solid food[74, 75]. The presence of glycan-adaptable species pre-weaning seems to suggest that the gut microbiota is prepared for the post-weaning dietary changes that inevitably occur after the infant is introduced to solid foods.

Gut microorganisms differ in the number of glycans they can degrade[73, 76]. For example, the Gram-negative gut symbiont *B. thetaiotaomicron* is able to degrade over a dozen different types of glycans[73, 68], while other species are only able to degrade one or a few[73]. Species with broad glycan-degrading capabilities that can shift their metabolism from host meal to meal are termed "generalists." In contrast, species that have narrower glycan degrading potential are termed "specialists". The disadvantage of specialists is that they may become extinct in a host if their preferred nutrient source is lacking. No gut microorganisms characterized today has the capacity to tackle all the glycan structures that enter the intestine. Host endogenous glycans (*O*-and *N*-linked glycans) are highly chemically diverse, with hundreds of different structures attached to a single mucin glycoprotein[77]. Degradation of host glycans requires mucosal bacteria that are able to produce numerous degradative enzymes to effectively utilize the glycans. It has been suggested that host glycans have evolved to be diverse in order to deter microbial species from evolving to be too efficient at harvesting mucosal glycans, thus protecting the integrity of the mucosal epithelial barrier.

Although, studies have shown that certain types of diets can shape the composition of the gut microbiota, supplementing the diet with specific glycans can also impact the type and abundance of a particular species. However, not all species that possess the capacity to degrade a given glycan will do so successfully *in vivo*. For example, inulin and certain fructo-oligosaccharides (FOS) increase the abundance of Bifidobacteria[51] although many *Bacteroides* species are also able to use these glycans[56]. Additionally, recent studies on human and animal feeding have shown that some gut microbial species are adept at degrading some forms of resistant starch (RS). A high consumption of some RS results in increases in the short chain fatty acid butyrate which is known to exert anti-inflammatory[78, 79, 78] and anti-tumorigenic benefits to the host[78 ,80, 81].

These variations in gut microbiota composition that arise from differing abilities of gut microorganisms to metabolize glycans could have profound implications for understanding both how the microbiota assembles over the span of a human lifetime and how transient community variations affect human health. Researchers can develop strategies to manipulate the gut microbiota function using prebiotic, probiotic or

pharmaceutical strategies. In order for this to be achieved, it is imperative to gain a deep insight into the molecular mechanisms involved in glycan-microbe interactions.

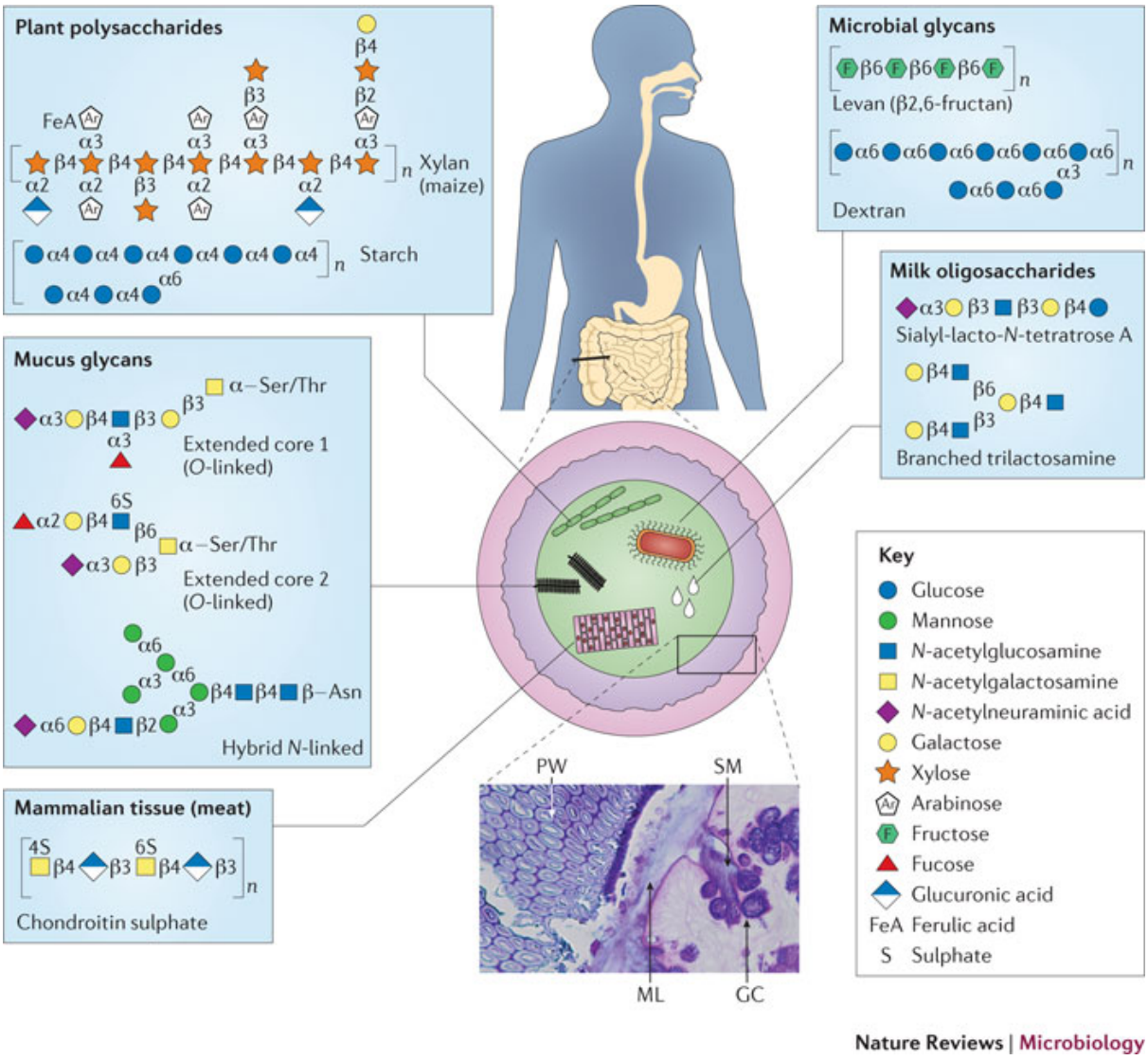


Figure 1.3 **Sources and chemical variation of glycans in the gut**

A cross-sectional view of the intestine depicting five different surces of glycans: dietary plants, dietary animal tissue, endogenous microorganisms, mucus and breast milk. The complexity of all possible glycans in each category is much more expansive than shown. Brackets at the end of horizontal glycan chains indicate that they may extend further with a similar linkage pattern. The inset in the upper left shows a section of germ free mouse colon with periodic acid-Schiff base and Alcian blue stains for various carbohydrates. The section is oriented similarly as the corresponding box in the gut illustration in the centre and highlights the locations of host mucus-secreting goblet cells (GC), secreted mucus (SM), the mucus layer (ML) and a fragment of plant cell wall (PW) located immediately adjacent to the mucus layer. Diagram adapted from Koropatkin *et al.,* 2012[10].

Regardless of the glycan substrate degraded by the gut microorganism (host glycan or dietary glycan), the host colonocytes benefit from the end result of the microbial metabolism by absorbing the short chain fatty acids (SCFA) produced. Butyrate, propionate and acetate are the three main types of short chain fatty acids produced. Butyrate that is produced in the colon is the preferred energy source of colonocytes and has been associated with suppressed growth of colonic tumors[78]. Acetate and propionate are absorbed into the bloodstream and travel to the liver where they are incorporated into lipid and glucose metabolism[82]. Notwithstanding absorption by the host, acetate also augments butyrate production by certain microorganism[83, 84] and prevents colonization of specific pathogens[85].

1.4.3 Glycans as legislators of host-microbial interactions

Glycans are involved in numerous aspects of cellular interactions with the host and microbe[9]. Studies have shown that initial recognition and binding at the cell surface is often mediated by glycans not only because of their structural diversity, but because of their high abundance and dependence on avidity[86]. Furthermore, glycans are ideal molecules for mediating binding at the cell surface due to their spatial organization in clustered saccharide patches which generates a unique topology for each protein[9]. This provides high specificity for glycan-protein interactions. Thus, in spite of the high abundance of glycans present on cells and tissues, glycan-mediated interactions occur only when the correct conformation or cluster is present[60]. Therefore, cells with similar glycan content can display unique clustered saccharide patches leading to differential recognition by glycan-binding proteins. The repertoire of glycans expressed by the host play a large role in determining whether a host-microbe relationship will be commensal, symbiotic or pathogenic[60]. The carbohydrate structures present on the surfaces of intestinal epithelial cells demonstrate great diversity, varying as a function of cell lineage, cellular location and developmental stage. For example, Galβ3GalNAc glycan structures are not present in the small intestine of mice until the conclusion of weaning. Similarly, structures recognized by the *Sambucus nigra* lectin are detectable in members of the mucus-producing goblet cell lineage early in postnatal life but not during adulthood[87]. This regional and developmental specificity of host glycan production suggests that the expression patterns of glycans may be linked to the spatial and temporal complexity of the intestinal microbiota[60].

The Gram-negative bacterium *Helicobacter pylori* found in the stomachs of more than half of the human population[88] can produce notable pathogenicity (e.g., chronic gastritis, duodenal ulcer and gastric adenocarcinoma) in only a subset of human hosts[87]. It has been suggested that *H. pylori* exists primarily as a commensal and emerges as a pathogen as a result of host, microbial and environmental factors. *In vivo* studies have shown that the *H. plyori* can behave as a commensal, pathogen or symbiont depending on the repertoire of glycans expressed in the gastric ecosystem of its host and by the microbe's ability to express the appropriate adhesin[60]. Just as glycan metabolism can shape the microflora, it is becoming increasingly clear that the host microflora has the capacity to shape the production of host glycans by modulating host cellular differentiation pathways in the intestinal mucosa[89]. For example, studies have shown that germ free mice do not produce the glycolipid fucosyl-asialo-GM_1 in their small intestine[90]. However, transitory expression of fucosyl-asialo-GM_1 is observed in these mice when they are colonized with a completely normal microflora. The appearance of the glycolipid has been attributed to the increased activity of a fucosyltransferase enzyme that adds fucose to asialo-GM_1[91,60].

Analysis of another mouse model demonstrated that specific host glycans are involved in establishing and maintaining a non-pathogenic, mutually beneficial relationship with at least one indigenous intestinal microbe; *Bacteroides thetaiotaomicron*. Studies have shown that when an adult germ-free mice is colonized with *B. thetaiotaomicron*, expression of Fucα1, 2Galβ glycans is induced in the ileum[92]. The ability of *B. thetaiotaomicron* to induce production of the fucosylated glycan is dependent on the density of *B. thetaiotaomicron* colonizing the gut. This suggests that *B. thetaiotaomicron* does not act by direct binding to the epithelium, but by means of a soluble bacterial factor[92]. Therefore, not only does *B. thetaiotaomicron* have the capacity to hydrolyze host-derived glycans, it is also capable of shaping the nature of glycans produced in its host intestinal epithelial cells. By serving as a nutrient source, the microbe induced the help of host glycans to establish a symbiotic host-bacterial relationship[92]. Furthermore, this ability of *B. thetaiotaomicron* to induce host glycan synthesis suggests that it might be a general strategy used by other members of the microflora involving other types of carbohydrate structures. Therefore, the host has evolved to synthesize highly structurally diverse glycans in part to evade pathogenic colonization but also to co-evolve symbiotic relationships with resident microbes.

1.4.4 Mucus-binding proteins

The human gastrointestinal tract is lined with a protective layer of mucus comprising of glycolipids and a complex mixture of large and highly glycosylated proteins or mucins (Figure 1.3). Mucus is simultaneously produced by goblet cells and degraded by proteases (human or bacterial origin) in the gastrointestinal tract[93]. A rapid turnover rate of mucus ensures that bacteria that adhere to the mucus have a short residence time in the gut. In this way, the mucus layer acts as a protective function against undesirable bacterial colonisation. Nonetheless, the mucus layer provides a habitat for commensal gut micro-organisms such as Lactobacilli[94,95]. Amongst the mucus adhesins identified and functionally characterized are the extracellular mucus-binding protein (Mub) of *Lactobacillus reuteri* 1063[96], the lectin-like mannose-specific adhesin (Msa) of *Lactobacillus plantarum* WCFS1[97], and the Mub of *Lactobacillus acidophilus* NCFM[98]. These three mucus binding proteins possess the same organization typical of cell surface proteins of Gram-positive bacteria (i.e, an N-terminal signal peptide targeting the protein for transport through the plasma membrane, and a C-terminal anchoring motif (LPxTG)) that is recognized by a family of enzymes called sortases for covalent attachment of the transported protein to the peptidoglycan of the bacterial cell wall[99, 100]. These three mucus-binding proteins share a similar mucus-binding domain (MUB) organization. This MUB domain, described as MucBP (MUCin-Binding Protein) in the Pfam database consist of a series of 50 amino acid residues in length and is found in a wide variety of bacterial proteins. To date, a total of 30 proteins containing three or more MUB domains have been identified in approximately ten lactic acid bacterial species[101].

1.4.5 Mucins

Mucins are the major component in the structure of mucus. One mucin molecule contains hundreds of heterogenous glycans. Mucins contain glycan rich domains that act as binding sites for commensal and pathogenic bacteria. To date, cDNA cloning has distinguished 20 genes of the human mucin family (MUC1, MUC2, MUC3, MUC3A, MUC3B, MUC4, MUC5AC, MUC5B, MUC6, MUC7, MUC8, MUC12, MUC13, MUC14, MUC15, MUC16, MUC17, MUC19, MUC20, MUC21) [102]. These genes have a common structural feature of tandem repeat domains consisting of amino acid repeats (in tandem) rich in PTS (Proline, Threonine, Serine) domains. These domains are saturated with hundreds of O-linked oligosaccharides [103]. Indeed, eight

core structures have been identified for mucin-type oligosaccharides O-linked to serine or threonine by an N-acetylgalactosamine (GalNAc) residue or GlcNAc, Gal and Fuc and the oligosaccharides can be terminated with sialic acid or sulfate group [104]. These family of high-molecular weight, heavily glycosylated proteins (Mucins) can be produced in the human gut as membrane bound (MUC1, MUC3, MUC4, MUC12, MUC13 and MUC17) or secreted mucins (MUC2, MUC5B, MUC5AC and MUC6). The membrane bound mucins are detained by the plasma membrane due to the presence of their hydrophobic membrane spanning domain. These extensively O-glycosylated transmembrane mucins have an extracellular domain that projects at least 800 nm above the cell surface and performs a crucial role in cell adhesion [105]. Meanwhile, the cytoplasmic region of the transmembrane mucin has been implicated in signal transduction [106]. Transmembrane mucins are also known to prevent colonization of pathogens by releasing their large extracellular domain on the cell surface as a decoy ligand for bacterial adhesins. The shedding of this extracellular domain limits the attachment of pathogens to other cell surface molecules and prevents invasion [107].

Secreted mucins form the bulk of the macromolecules of the epithelial surface mucus layer [108]. Secreted mucins form a protective mucous gel barrier that shields the epithelial cells lining the gastrointestinal tract from the luminal contents such as bile, proteases, toxins and commensal bacteria. This viscous, secreted mucous coating provides protection from damage to the epithelium but also alleviates the activation of both the innate and adaptive immune responses [104]. The attenuation of needless or excessive immune responses is crucial to preserving energy and maintaining homeostasis in GIT.

Of the secreted mucins, MUC2 has been connected to inflammation and cancer. Indeed Van der Sluis *et al*. (2006) have shown that Muc2 deficient mice spontaneously develop colitis inflammation and colorectal cancer. Muc2 has a large, O-glycosylated, centrally located PTS domain [109]. Muc2 inhibits the progression of inflammation in the GIT and subsequently prevents the development of intestinal tumours. Their research underscores the importance of a mucus layer in preserving the mutualistic and symbiotic relationship with the gut microbiota and host. The importance of the mucus gel layer for survival is further highlighted in the fact that these gel forming mucins have evolved and prevailed from early multicellular animals to humans. The

development of animal models with mutant mucin genes (muc2-) has been instrumental in aiding researchers to gain insight into the role of mucins in the maintenance of gut homeostasis [109].

1.4.6 Mucin Glycosylation

Glycosylation is one of the most prolific and important protein post-translational modifications. More than half of human proteins are decorated or glycosylated with different glycan chains. Glycosylation is important in many biological processes such as protein folding, conformational stability, cell division, cell growth and cell differentiation. Glycosylation also plays an important role as receptors on proteins to sense extracellular signals from surrounding cells or invading pathogens, consequently triggering an immunological response.

Mucins play a critical role in defense, microbial adhesion, immunomodulation, cancer and inflammation [110] of the gut microbiota. The glycan repertoire of a mucin determines the composition of commensal bacteria that inhabit a certain niche because they provide preferential bacterial-binding sites [111]. Indeed, the oligosaccharides on a mucin molecule can show a high degree of heterogeneity between species and at different locations within the human body. Mucin structures can exhibit high heterogeneity, even on the same molecule. The mucin oligosaccharides differ in chain length, residue linkages, residue composition and branching [112]. These highly dynamic structures provide binding sites for cell adhesion and participate in modulating the immune system. Mucin glycosylation alters the density of the substitutions along the mucin protein backbone as well as altering the distribution of specific structures and their presentation in space. All these alterations directly affect the function and bacterial binding capacity of each mucin molecule [113]. Mucin glycosylation can vary depending on cell lineage, developmental stage and tissue location. Moreover, the oligosaccharides present on the surface of the mucin molecule can also be influenced by health and disease status as well as growth, development, infection, cell differentiation, activation and neoplasia [102]. The glycosylation of mucins is important because it provides a source of ligands present in the mucus layer for bacterial adhesins, which recognize and bind to specific glycan structures [104]. This is a major reason why certain species form specific niches *in vivo* resulting in varied endogenous populations. It is possible that this differences in mucin glycosylation could account for the various disease outcomes between species and individuals [114].

1.5 Bacterial adhesins in host-microbe interaction

The concept of bacterial adhesion to host cells was first identified in 1908 when *Escherichia coli* was reported to bind to hemagglutinate animal cells by appendages which were later identified as multimeric pili[115]. Most commensal and pathogenic bacteria interacting with host cells express adhesive molecules on their surfaces that promote interaction with host cell receptors or with soluble macromolecules[115]. Bacterial adhesins are either assembled into complex polymeric organelle structures or linked to the cell surface as monomers or simple oligomers. Adhesin proteins are highly conserved with minor changes in the protein structure resulting in decreased or increased affinity for binding sugars[116]. They display extreme selectivity for their receptor molecule that they are able to recognize molecular motifs in a lock and key fashion similar to enzymes and immunoglobulins. As an example, *N. gonorrhoeae*[117] is a host specific pathogen that almost restrictedly infects humans; some diarrhoea causing *Escherichia coli* strains have adhesins that are restricted to pigs and humans only; lastly, *Escherichia coli* strains known to colonize the urinary tract have been shown to express specific fimbrial adhesins[118]. Adhesins behave like an address indicator for the microbe by targeting the bacterium to a specific tissue. This capacity of different bacteria to exhibit different host specificity and tissue tropism is determined by the specific interaction between their cell surface adhesins and the complementary glycan receptors on the host cell surface. The exact mechanisms by which bacterial adhesins interact with complementary receptors on host epithelial cells is still under investigation. To date, a large number of bacterial adhesins with distinct receptor specificities have been identified[101]. However, studies have revealed that some individual adhesins are able to rapidly modulate their receptor specificities[119].

The initial contact between adhesins and receptors involves biophysical and biochemical interactions between the host cell surface and the bacterial extracellular surface components[120]. One of the many functions of adhesins is to enable the bacteria to resist physical removal by sheer forces such as peristalsis in the human gastrointestinal tract. However, it has become increasingly clear that the bacterial adhesin-receptor binding event can activate complex signal transduction cascades in the host cell that can lead to the activation of innate host defences or the subversion of cellular processes facilitating bacterial colonization or invasion[121]. In many cases, this bacterial attachment to epithelial cells can also facilitate phagocytosis and clearing of

the bacteria[122]. Numerous pathogenic bacteria have counteracted this dilemma by expressing an antiphagocytic surface layer made of polysaccharides[123]. In the past, bacterial adhesive surface structures such as pili and fimbrial adhesins were the predominating adhesins studied[115]. However, a large number of monomeric surface-bound adhesive proteins have been identified and studied. The next section will describe in more detail the molecular strategies and adhesive mechanism used by bacteria to adhere to host cells.

1.6 Mechanisms of bacterial adherence to host cells

The mechanism of bacterial adherence to host cells may involve nonspecific or specific steps. In non-specific adherence, bacteria reversibly attach to the host eukaryotic cell surface. Hydrophobic interactions, electrostatic attractions, atomic and molecular vibrations, Brownian movement and recruitment, and trapping by biofilm polymers are involved in this adherence. In contrast, specific adherence involves the permanent formation of specific bonds between molecules that are complementary. Bacteria have evolved a very large arsenal of molecular strategies allowing them to target and adhere to host cells. One of the most well-studied bacterial adherence mechanism is the attachment mediated by pili or fimbriae. Pili are hair-like appendages that extend from the bacterial cell surface and allow the bacteria to make contact with host epithelial cells. The base of pili is anchored to the bacterial outer membrane, whereas the tip is usually an adherence factor that confers the binding specificity of the pili. Studies have shown that numerous pili are expressed on *Escherichia coli* and *Enterobacteriaceae*. The best characterized bacterial adhesin is the type 1 fimbrial FimH adhesin. The FimH adhesin is a two-domain, minor subunit protein at the tip of type I pili that enable *E. coli* to bind to D-mannose residues on the host eukaryotic cell surface. The bacterium synthesizes a precursor protein consisting of 300 amino acids then processes the protein by removing several signal peptides ultimately leaving a 279 amino acid protein. Mature FimH displayed on the bacterial surface as a component of type 1 fimbrial organelle. Minor changes in FimH adhesion may affect its binding affinity for sugars[116]. Genetic variation enables micro-organisms to develop adherence to different receptors. Pilus-associated adhesins have also been identified in a number of other bacteria such as *Pseudomonas spp*, *Vibrio spp*, Neisseria spp. and *Haemophilus spp*, and their receptors are rich in oligosaccharides.

Not all Gram-negative adhesins are associated with pili. For example, *Bordetella pertusiss* expresses two putative adhesins on its cell surface: filamentous hemagglutinin (FHA) and pretactin[124]. FHA mediates attachment to sulphated sugars on cell surface glucoconjugates, whereas pretactin mediates attachment to integrin-binding proteins. Further examples of non-pilus adhesins are the high molecular weight adhesion proteins (HMWI and HMWII) and immunogenic high molecular weight surface-exposed proteins (Hia) of *Haemophilus influenzae*.

Sections 1.6.1 – 1.6.3 will focus on the interaction between extended polymeric bacterial adhesins and the host. The best-characterized pilus structures, such as type 1 pili, P-pili, type IV pili, and curli in Gram negative bacteria are described.

1.6.1 P pili and Type 1 pili

Uropathogenic strains of *Escherichia coli* (UPEC) that colonize the urinary tract and are involved in kidney infections (acute pyelonephritis[125]), display P-pili (pyelonephritis-associated pili) at their surface. The tip of P-pili contains the adhesion factor PapG, that binds to glycosphingolipids of the kidney epithelial layer[126] and mediates binding to Gal-α(1-4) β-Gal moieties present in the globoseries of glycolipids on uroepithelial cells and erythrocytes[127] . Eleven genes that are organized in the Pyelonephritis-associated pili (*pap*) gene cluster are required for the expression and assembly of these organelles[128]. Some UPEC strains also possess Type I pili at their bacterial cell surface which bind specifically to D-mannosylated receptors, such as uroplakins of the bladder[129]. Type 1 pili are important determinants that are expressed in *E. coli* as well as in most members of the *Enterobacteriaceae* family that mediate binding to mannose-oligosaccharides[130, 121]. Studies have shown that Type 1 pili require approximately nine genes that are present in the type 1 gene cluster[121].

1.6.2 Type IV pili

Type IV pili are a class of polymeric adhesive surface structures that are expressed by different Gram-negative organisms such as *Neisseria spp*., *Pseudomonas aeruginosa*, pathogenic *Neisseria*, *Moraxella bovis*, *Dichelobacter nodosus*, *Vibrio cholerae* and enteropathogenic *E. coli* (EPEC) and Gram-positive organisms such as *Clostridium perfingens* and *Streptococcus sanguis*. These pili consists of thousands of copies of the major pilin which are produced in the bacterial cytoplasm and translocated across the inner membrane to be proteolytically processed. Once the pili is assembled and

aggregated on the bacterial surface, it has the ability to retract through the bacterial cell wall, while the pilus tip remains attached to its target surface, permitting a "twitching motility", flagellant-dependent mode of motility crucial for effective colonization of host surfaces[131]. Type IV pili has been implicated in a variety of functions, including adhesion to host cell surfaces, modulation of target cell specificity and bacteriophage adsorption.

1.6.3 Curli

Curli are thin, irregular, and highly aggregated surface structures that are produced by many clinical *E. coli* and *Salmonella enteritidis* isolates[132]. They are highly stable structures that often require extreme chemical treatment to depolymerize them (e.g., 90% formic acid). Curli are formed in a nucleation-dependent process in which the major subunit protein, CsgA is secreted across the inner membrane via the Sec leader. They mediate binding to a variety of host proteins including fibronectin, plasminogen[133] and human contact phase proteins[134]. Curli consists primarily of a 15.3-kDa protein termed CsgA, which exhibits more than 86% primary sequence similarity to its counterpart in *S. enteritidis*, AgfA. Curli are known to be notoriously sticky without demonstrating an evident ligand-binding specificity. Studies have shown that most commensal isolates of *E. coli* and *Salmonella* only express curli at room temperature. However, recent studies by Bian and colleagues demonstrate that a number of clinical *E. coli* urosepsis isolates also express curli at 37^{o}C, suggesting a role in pathogenicity[135].

1.7 Bacterial adhesion in the human gastrointestinal tract

Besides pili and fimbria, a plethora of different bacterial non-polymeric adhesins exist that recognize many different elements of host-cell surfaces and can be classified according to their targets in the human intestinal mucosa (i.e. mucus components, extracellular matrices) (Figure 1.4), according to their localization in the bacterial surface (Figure 1.5) (i.e. surface layer proteins) and/or according to the way they are anchored to the bacterial surface (i.e. sortase-dependent proteins).

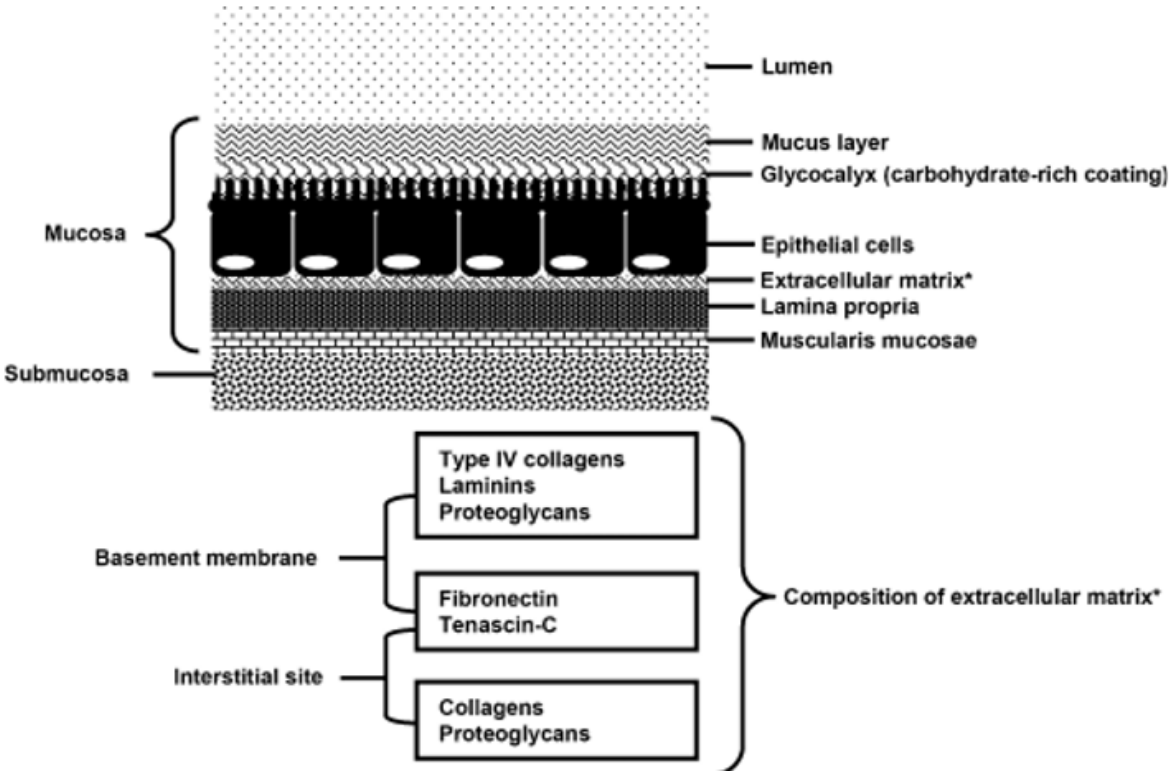


Figure 1.4 Diagram depicting representation of the components of the human intestinal mucosa and submucosa. Components of the extracellular matrix are indicated with an asterisk[101]. Diagram adapted from De Keersmaecker *et al*., 2007.

1.7.1 Surface-layer proteins as adhesins

A few well-characterized commensal S-layer proteins include CbsA of *Lactobacillus crispatus* JCM 5810[136, 137,] Slp of *Lactobacillus helveticus* R0052[138], SlpA of *Lactobacillus brevis* ATCC 8287[139, 140, 141] and SlpA of *L. acidophilus* NCFM[142]. S-layer (surface layer) is a part of the cell envelope found in almost all archea, as well as in many types of bacteria[143]. It consists of a monomolecular layer composed of identical proteins or glycoproteins. The S-layer structure is built via self-assembly and encloses the whole cell surface. Thus, the S-layer protein can represent up to 15% of the whole protein content of a cell[143]. Studies have shown that these proteins mediate adhesion to intestinal epithelial cells[141, 140, 142, 138], epithelial matrices[101, 139 141, 144] and to lipoteichoic acid of other species[137]. Further studies have shown that some of the S-layer proteins are effective in preventing adhesion of pathogenic bacteria to host epithelial cells[138]. Some S-layer proteins have been shown to exhibit similarities in their structre. For example, CsbA of *L. crispatus* JCM 5810 and Slp of *L. helveticus*

R0052 contain a bacterial S-layer protein domain that is often present in S-layer proteins.

Overall, bacterial adhesion in the human gut is paramount to maintaining the health of the host. There is increasing evidence that the human gut microflora is not only beneficial to the host but is essential for the host's proper development; in the differentiation and maturation of the intestinal tract and immune system[145,146]. A knowledge of how members of the normal human gut microflora adhere to their host is therefore important in understanding this symbiotic relationship and host-microbe interaction. Adhesion of a pathogenic micro-organism to the host epithelial cell is the first stage in any infectious disease and this truism provides another justification for studying bacterial adhesion in the human gut. Keen interest in bacterial virulence via adhesion is increasing rapidly as our collection of effective antibiotics dwindles owing to the development of resistance in major pathogens. Therefore, acquiring profound insight into the molecular mechanisms that govern bacterial adhesin-host glycan interactions will provide opportunities for scientists to influence the human gut ecosystem to improve health, physiology and nutrition. The next section will discuss current high throughput sequencing technologies and techniques that should provide insights into these glycan-microbe interactions.

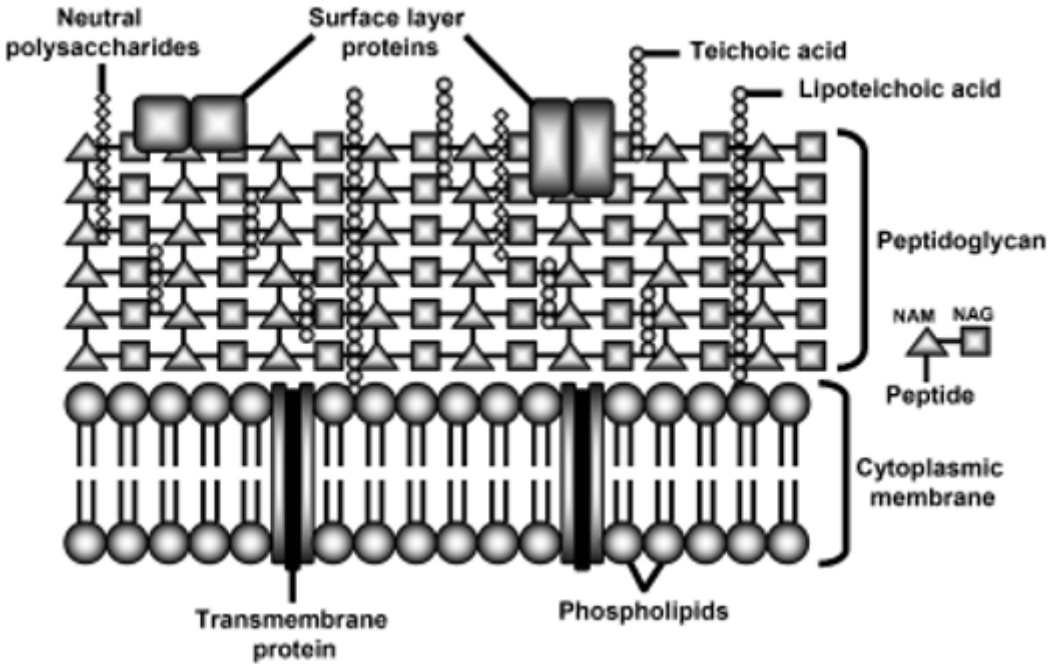


Figure 1.5 **Representation of the cell wall of a Gram-positive bacterium.** The bilipid cytoplasmic membrane is embedded with proteins and covered

by a multi-layered peptidoglycan shell decorated with neutral polysaccharides, LTAs, teichoic acids and surrounded by an outer envelope of S-layer proteins.[101] Diagram adapted from De Keersmaecker *et al*., 2007.

1.8 Metagenomics

The majority of the planet's diversity is comprised of uncultured microorganisms. Microorganisms represent two of the three main domains of life and consists of a vast diversity that is the result of billions of years of evolution[147]. To understand the genetic diversity, population structure and ecological roles of the gut microbiota, it is essential to perform culture independent methods. Metagenomics is the culture-independent, genomic analysis (functional and sequence-based) of the collective microorganisms contained in an environmental sample by direct extraction and cloning of DNA. It is a multi-step approach which requires sampling, sample processing, DNA extraction and data analysis based on the sequence or function. Sampling and DNA extraction are critical steps in metagenomics application as they impact downstream procedures. Metagenomics has the capacity to answer fundamental questions in microbial ecology. Many environments have been the focus of metagenomics, including soil, the oral cavity, aquatic habitats, hospital metagenome and faeces. The emergence of metagenomics techniques has made it possible for researchers to study "as-yet uncultured organisms" to identify numerous novel genes, enzymes and proteins from many diverse environments through direct sequencing of metagenomics DNA or through functional expression in a heterologous host [148].

1.8.1 Sequence-driven analysis

The sequence-based screening of a metagenomics library involves the identification of homology between sequenced clones and already characterized genes in the database. This approach is able to disclose genes of interests and catalogue the genetic potential, but will not detect fundamentally novel gene functions. A successful sequence-based approach depends primarily on sequencing effort and good microbial coverage of the sample of interest. The advent of next-generation high-throughput sequencing is able to produce a large number of fragmented pieces of sequences that can be assembled into longer contigs and then analysed. Next-generation sequencing analysis has improved greatly with the introduction of a variety of bioinformatics tools for gene analysis[149]. The second approach for a sequence-based analysis involves the

designing of DNA primers or probes which are derived from regions of already known genes or protein families[150] to retrieve specific genes from a pool of DNA. Instead of cloning all the extracted DNA, primers are designed specifically against an identified target gene. The advantage of using this sequence-driven approach is that it uses well-established and high throughput techniques, such as PCR and hybridization, and can be used for different targets. On the other hand, with this approach, already-known sequence types will be identified and only a fragment of the main target gene will be amplified. Despite this limitation, combining PCR detection of small conserved regions with genome sequencing at flanking regions makes it possible to obtain the entire gene[150].

1.8.2 Function-driven analysis

Functional metagenomics relies on the construction of a metagenomics library and the expression of the genes in a heterologous host (Figure 1.6). The main advantage of a function-based approach is the capacity to access previously unknown genes and their phenotypic traits, which may have applications in medicine, agriculture and industry[151]. Depending on the insert size, metagenomic libraries have been constructed using different cloning vectors such as plasmids (up to 15 kb), fosmids (both up to 42 kb) and bacterial artificial chromosomes (> 40 kb) (Figure 1.6). Small insert libraries are usually produced using plasmids as a cloning vector [152]. Small insert size libraries are employed to identify single genes (mostly enzymes) or small operons. They are usually constructed in *Escherichia coli* as a heterologous host, therefore the transformation efficiency is high (>10^{15} cfu/μg DNA). The main advantages of using plasmids is the high copy number and that the plasmid promoters and ribosome binding sites can be fused to the cloned DNA, thus the host transcription and translation systems can be used [153]. Large insert size libraries are produced to recover biosynthetic pathways and large clusters of genes involved in the synthesis of complex enzymes and antimicrobial compounds [154]. Fosmids have been used in multiple studies to produce large insert libraries because their copy number is low to ensure a high stability of the recombinant gene[155]. An advantage of using fosmids to generate a metagenomics library is that the "gene of interest," along with any genetic element in which it is embedded can be isolated. Additionally, DNA sequences that may indicate the phylogenetic origins of the original host bacteria can be contained in the fosmid clone [156].

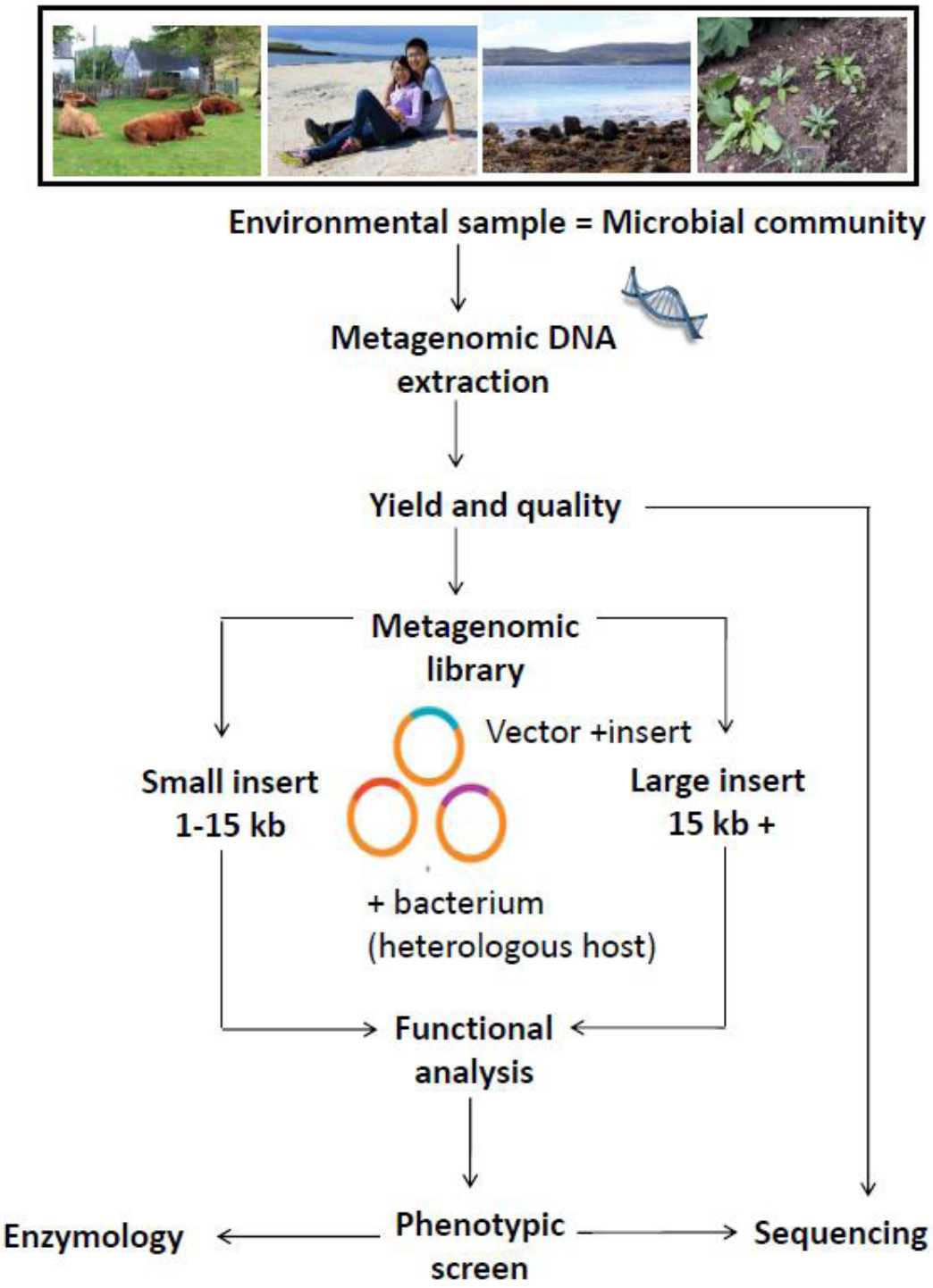


Figure 1.6 **An overview of processes involved in the production of a metagenomics library**. DNA is isolated from the cells in the sample and then fragmented, inserted into vectors, cloned. The vectors are introduced into host cells. After the metagenomics library, the metagenomes undergo function and sequence analysis. Diagram adapted from Dias *et al.,* 2014[157].

The frequency at which clones in metagenomics libraries express any given functional activity is relatively low. Indeed, in a search for lipolytic clones derived from German soil, Gottschalk and colleagues [158] were only able to find 1 clone out of 730,000 clones that showed activity. Similarly, in a library of DNA from North American soil, Rondon and colleagues[159] discovered that only 29 of a total of 25,000 clones expressed haemolytic activity. Overall, functional metagenomics analysis is currently one of the metagenomics screen that is able to isolate novel genes. However, the major disadvantage of this technique is that expression of the "gene of interest" is required. Since *E. coli* is still the most popular cloning host, genes that do not express in this gram negative organism will be lost [156]. Overcoming these problems requires the use of other cloning hosts other than *E. coli* [160].

As outlined, metagenomics analysis enables the comprehensive investigation of microbial communities and provides unprecedented access to the genetic diversity within these communities. The development of novel hosts and expression systems will increase discovery rate and the variety of novel genes that can be discovered.

1.9 Functional screening of metagenomics libraries

Functional metagenomics has only recently been applied to the study of the human commensal gut microbiota. This technique was originally proposed as a method to characterize the unculturable fraction of soil microbiota[161, 162, 159] and successfully used for years to characterize the functional diversity of microbes in a series of environments[161, 163]. The range of functional screens that can be performed using metagenomics libraries is immense. Functional metagenomics screening does not require direct culture of fastidious organisms. Instead, clone libraries are constructed by extracting and shearing DNA from a sample of a microbial community, then cloning the fragmented DNA into a relevant vector, and subsequently transforming the vector into an appropriate heterologous host[161] (Figure 1.6). Once a library has been constructed, it can be functionally screened by cultivation on a selective media or by employing a reporter system[161]. Using this approach, it is possible to identify genes that encode functions such as antibiotic resistance and metabolism of complex compounds. Subsequent sequencing and *in silico* analysis of the DNA from isolated clones provides additional information about the source of the genes and the putative mechanisms of action of their products[161]. The human gut microbiota is an interesting

environment to screen due to the immense diversity of its encoded genes. It encodes multiple critical functions impacting human health, such as metabolism of dietary substrates, prevention of pathogen invasion, immune system modulation and provision of a reservoir of antibiotic resistant genes accessible to pathogens[161].

Functional metagenomics screening has been successfully used in the discovery of new antibiotic resistance genes in the human gastrointestinal microbiota. The increasing prevalence of multidrug resistant bacteria in both hospitals and the community pose a growing threat to human health[164,165]. Novel antibiotic resistance genes have been identified in different environments including oral microbiota, soil microbiota, and moth gut flora[166,147]. In 2009, Sommer and colleagues demonstrated the power of metagenomic functional screens to identify novel antibiotic resistance genes in the faecal samples of two adults[167,161]. They generated metagenomics libraries with a total size of 9.3 Gb (gigabases) and an average insert size of 1.8 kb (kilobases) and screened these libraries for resistance against 13 different antibiotics, revealing 95 unique inserts representing a variety of known resistance genes as well as 10 novel beta-lactamase gene families[167].

In 2010, Tasse and colleagues performed a functional screen of a large human gut metagenomics library (156,000 *Escherichia coli* fosmid clones, covering in total 5.46 x 10^9 bp of metagenomics DNA, each clone comprising a 30-40-kb DNA insert) issued from the faeces of an individual who followed a fiber-rich diet, to easily isolate genes encoding enzymes that were able to break down raw and insoluble plant polysaccharide. The library was screened for the ability to hydrolyze five different polysaccharides. They were able to isolate 310 positive clones of which 8 maintained enzyme activity at pH 4.

Functional metagenomics screens have also been used to mine the salt tolerance genes from the human gut microbiota. The ability to respond and adapt to changes in external osmolarity is a key determinant for bacterial survival and proliferation in various environmental niches. Using transposon mutagenesis, Culligan and colleagues identified three genes from a single clone exhibiting high levels of identity to a species from the genus *Collinsella* and a high G+C, Gram-positive member of the Actinobacteria commonly found in the human gut[168].

In 2013, Yoon and colleagues constructed a bacterial artificial chromosome (BAC) library of murine bowel microbiota DNA in the surrogate host *Escherichia coli* DH10B and screened the library for enhanced adherence capability. Two out of the 5,472 DH10B clones exhibited enhanced capabilities to adhere to inanimate surfaces in functional screens[169]. The study revealed a genetic factor from unknown commensals that enhanced the ability of the bacteria to colonize the murine bowel[170].

As outlined, functional screening of metagenomics libraries has the power to reveal novel functions for known genes or to identify completely novel genes and proteins. Functional metagenomics is a technique that promises to expand our understanding of microbial community function, its impact on human health, and to provide novel targets for therapeutic development in coming years. The next sections will highlight the importance of characterizing bacteria using carbohydrate-based microarrays.

1.10 Carbohydrate-based Microarrays

1.10.1 Mucin Microarray

The mucosal epithelial tissue of the human gastrointestinal tract exposes an expansive surface area (400 m^2) to the exterior environment [102]. It constitutes the main route of access for viruses, archaea, yeast, protozoa and multicellular parasites that trigger disease in humans [104]. Exposure to the external environment (intestinal contents) renders the underlying mucosal epithelial tissue susceptible to microbial attack from the tens of trillions of resident microflora [171]. In response to the massive load of bacteria in the lumen, gut mucosal epithelial cells constitutively produce and secrete defensive compounds such as mucins (secreted or membrane bound), histatins, nitric oxide, cathelicidins, collectins, protegrins, antibodies and defensins [113]. Indeed, most epithelial cells are single cells and require very robust defense mechanisms to maintain the integrity of the epithelial barrier. One of the main defense barrier to the external environment in the GIT is the presence of a highly hydrated mucus layer [105]. This mucus layer coats the epithelial cells of the gastrointestinal, urinary and reproductive tracts of the human body. In the human GIT, the viscous mucus layer has a thickness of approximately 700 μm and protects the underlying epithelial cells against chemical, physical, enzymatic and mechanical injury [105]. This mucus layer is particularly efficient at trapping microorganisms and is continually removed by the peristaltic movement of the luminal contents. The thickness level of the mucus layer varies with

the region of the GIT, but is at its thickest in the colon and rectum [106]. The mucus layer is divided into a loose outer layer and an inner layer which is firmly attached to the underlying epithelial cells. Studies have shown that the presence of bacteria is restricted to the loose outer layer of mucus, while the inner layer is devoid of bacteria [172]. The exact mechanism by which these two mucus layers are formed is yet to be elucidated. Understanding the mechanism of colonization of bacteria to the mucus barrier is crucial because researchers have shown that a functional mucus layer is critical in maintaining the health of the host. Studies indicate that animal models with a depleted mucus layer develop spontaneous colitis [173]. Furthermore, some humans with ulcerative colitis also have a depleted mucus layer in their intestines enabling bacteria to penetrate into the underlying epithelial cells [174].

The mucus layer consists of a combination of complex molecules including; mucin glycoproteins (section 1.4.5), immunoglobulins, lipids, antimicrobial peptides and electrolytes [175]. Mucins are high molecular weight glycoproteins with a carbohydrate content accounting for up to 90% of their weight. It is this high carbohydrate content that aids in the general viscoelasticity and barrier function of the mucus layer [105]. On the other hand, the oligosaccharide structure of the mucins contribute to their physical and biological properties because of hydrophobicity, configuration and charge [104] (section 1.4.5).

Over the past several years, there has been rapid advances in metagenomics and sequencing technologies that have revealed the health consequences of altering the composition of the gut microbiota. Researchers have gained novel insights into the interactions between commensal and pathogenic bacteria to the mucus layer of the GIT. Mucus-microbial interactions in the GIT play an important role in determining the cross talk between host-microbe, however the adhesins and receptors involved in this relationship are unknown [176]. Although this subject has garnered widespread attention, our knowledge of the gut mucus-microbial interaction remains incomplete. As a result, there has been keen interest in elucidating how bacteria colonize mucus layers and interact with components of the gut mucus, such as mucin [113].

1.10.2 Neoglycoconjugate Microarray (NGC)

Scientists have come to recognize the crucial and informative role that carbohydrates play in human health and disease. Many host proteins are decorated with structurally

heterogeneous carbohydrates that play a fundamental role in protein function and interaction [177]. Carbohydrates play critical roles in numerous biological processes including fertility, viral infection, bacterial adhesion, immune response, immunity, immunodeficiency diseases and the nervous system[178]. This fact has spurred researchers to develop assays and technologies to exploit carbohydrate-protein interactions for therapeutic and diagnostic benefits. To cope with the keen and rapid interest in carbohydrate-protein interactions, scientist coined the term "glycobiology". Glycobiology is defined as the study of structure, biology and biosynthesis of carbohydrates (glycans & sugar chains) that are dispersed in nature [179]. This field has been growing significantly in interest and impact over the past decade. It is now known that viral infections require recognition of carbohydrates [178]. Oligosaccharide structures in erythropoietin have been shown to be important in its activity. Humans with the common flu contain glycans that play an important role in anchoring viral hemagglutinin glycoproteins.

1.11 Objective of the study

The overall aim of this project was to identify and characterize novel glycan-binding bacterial determinants encoded by the human gut metagenome. A culture-independent (metagenomic), functional metagenomics approach was chosen as one method of study. The first objective was to construct two types of human gut metagenomics libraries (small fragment library & fosmid library) for screening of novel adhesins encoded by the human gut metagenome using an *in vitro* adhesion assay of bacterial adhesion onto mammalian Caco-2 epithelial cells (human adenocarcinoma cells known to mimic gut epithelial cells). The second objective was to characterize the newly identified putative adhesive clones by interrogation onto three main types of carbohydrate-based microarray platforms; (i) carbohydrate-binding proteins (lectins) microarray to determine glycosylation patterns on the bacterial cell surface; (ii) natural mucin microarrays to determine mucin glycosylation and bacterial binding tropisms, (iii) and finally neo-glycoconjugate (NGC) microarrays to determine the binding affinity of bacteria to specific neo-glycoconjugates. Knowledge of the specific glycans that gut microbes bind will equip researchers with the possibility to develop dietary interventions such as anti-adhesion therapy, prebiotics, probiotics, and synbiotics to modulate the gut microbiota for the benefit of the host. A third objective was to identify a homologous glycan binding protein adhesin encoded by bacteria from the

human gut microbiota using an *in silico* based approach. The adhesin was PCR amplified and subsequently cloned into an appropriate expression vector and expressed using the Nisin controlled gene expression system of *Lactococcus lactis.* An *in vitro* adhesion assay onto Caco-2 cells was used to determine adhesive property of the homologous protein. The two approaches used in this study – functional metagenomics and bioinformatics – are complementary to one another. The bioinformatics approach is a targeted approach to identifying novel glycan-binding clones whereas the functional metagenomics approach is a blind approach used to identify unknown genetic determinants.

This project was developed with the collaboration of the Rowett Institute of Nutrition and Health (University of Aberdeen, Aberdeen, UK), School of Pharmacy and Biomolecular sciences (University of Brighton, Brighton, UK) and Bioscience Research Building (National University of Ireland, Galway, Ireland).

Chapter 2:

Materials and Methods

2.1 General microbiological techniques.

All chemicals and reagents were obtained from Sigma Aldrich, unless otherwise stated.

2.1.1 Bacterial strains and plasmids

The bacterial strains used in this study are presented in Table 2.1. Permanent stocks were made by centrifuging 5 ml of overnight culture and re-suspending in 3 ml of appropriate media supplemented with DMSO to a concentration of 7% (v/v). 1 ml aliquots were put in 2 ml cryovials and stored at -80^{o}C. Working cultures of the strains were streaked onto M17 agar plates (*L. lactis* MG1363 and *L. lactis* NZ9000) or LB agar (*E. coli*) and grown overnight at 37^{o}C. Plates were supplemented with selective antibiotics and stored at 4^{o}C. Plasmids used in this study are listed in Table 2.2.

Table 2.1 Bacterial strains used in this study

Strain	**Characteristics**	**Source**
Lactococcus lactis MG1363	Plasmid free strain, Lac-	Linares *et al*., 2010
Lactococcus lactis NZ9000	PepN::nisRnisK	Van Sinderen lab
EPI300™ *(Fosmid clone 18)*	Cm^{r}, pCC1FOS cloning vector carrying 24.6 kb fragment	This study
EPI300™ *(Fosmid clone 19)*	Cm^{r}, pCC1FOS cloning vector carrying 24.6 kb fragment	This study
EPI300™ *(Fosmid clone 3)*	Cm^{r}, pCC1FOS cloning vector carrying 24.6 kb fragment	This study
EPI300™ *(Fosmid clone 21)*	Cm^{r}, pCC1FOS cloning vector carrying 8.1 kb fragment	This study
EPI300™ *(Fosmid clone 22)*	Cm^{r}	This study
EPI300™*(*pCCIFOS^{TM}*)*	*F-mcrAD(mrr-hsdRMS-mcrBC)F80d lacZ D M15 lacx 74 recA 1 endA1 araD139D (ara, leu) 7697*	E. Culligan (UCC)

	galu galk]-rpsl nupG trfa dhfr ; high-transformation efficiency of large DNA, Cm^r, pCC1FOS cloning vector	
EPI300™-$T1^R$ *E. coli* strain	*F-mcrAD(mrr-hsdRMS-mcrBC)F80d lacZ D M15 lacx 74 recA 1 endA1 araD139D (ara, leu) 7697 galu galk]-rpsl nupG trfa dhfr ; high-transformation efficiency of large DNA*	Epicentre Biotechnologies
E. coli Top 10	*mcrA*, Δ(*mrr-hsd*RMS-*mcrBC*), Phi80*lacZ(del)M15*, Δ*lacX74*, *deoR*, *recA1*, *araD139*, Δ(*ara-leu*)7697, *galU*, *galK*, *rpsL(SmR)*, *endA1*, *nupG*	Invitrogen

Table 2.2 Plasmids used in this study

Plasmid	Characteristics	Reference
pTRKL2	EryR, *lacZ*, 6.4 kb	Sullivan *et al*., 1993
$pCC1FOS^{TM}$ vector	Fosmid cloning vector, Cm^r, 8.3 kb	Epicentre Biotechnologies
pPTPi	pnisA, tetK, shuttle vector, 6.8kb	Van Sinderen Lab
pCR-XL-TOPO	Kan^r, Amp^r, *lacZ*, *E.coli* cloning vector	Invitrogen
pPTPi::$MapA_{Ri}$	pPTPi vector carrying $MapA_{Ri}$	This study

2.1.2 Culture media

All media where stated were autoclaved at 121°C for 15 min in a Labo autoclave (Sanyo). Filter sterilisation was performed using a 0.22 μm syringe filter (Sartorius) and syringe (BD Plastipack).

2.1.2a Brain Heart Infusion (BHI)

BHI broth was prepared by adding 37 g of BHI broth powder (LabM) per 1000 ml dH_2O. Where 0.5M sucrose was required, 171.15g sucrose was added to media prior to autoclaving. BHI agar was prepared by adding 49 g BHI agar powder (LabM) per 1000 ml dH_20.

2.1.2b Luria-bertani (LB)

LB broth was prepared by adding 10g LB powder (Sigma) per 1000 ml dH_2O.
LB agar was prepared by adding 15g Agar No.2 (LabM) into 1000 ml.

2.1.2c M17 medium

M17 broth was prepared by adding 37.25 g of dehydrated M17 powder per 1000 ml dH_2O. To prepare GM17 agar, 15 g of M17agar was added to 1000 ml dH_2O supplemented with 50 ml (after autoclaving) 10% glucose.

2.1.2d Modified M17 medium (mGM17)

Modified M17 medium is prepared by adding 37.25 g of dehydrated M17 broth into 1000 ml dH_2O supplemented with 2.5% glycine, 0.5M sucrose and 10% glucose (after autoclaving or filter sterilising).

2.1.3 Media supplements

Depending on the strain or experimental condition, certain supplements were added to the media as required. Stocks of these supplements were made as outlined below:

2.1.3a Antibiotics

A stock solution of ampicillin (Amp) was prepared by dissolving 50 mg ampicillin sodium salts (Sigma) in 1 ml dH_2O. Chloramphenicol (Chl) was prepared by adding 50 mg chloramphenicol (Sigma) per 1 ml ethanol (70%). Erythromycin (Erm) was prepared by adding 50 mg erythromycin (Sigma) per 1 ml ethanol (70%). Kanamycin (Kan) was prepared by adding 20 mg kanamycin salt (Sigma) per 1 ml dH_2O. All antibiotic solutions were filter sterilised and stored at -20^{o}C. Antibiotics were added to media after sterilisation and once the media had reached 55^{o}C or below.

2.1.4 Bacterial growth conditions.

2.1.4a General bacterial growth conditions.

Escherichia coli strains were grown at 37°C in LB (section 2.1.2b) or BHI medium (section 2.1.2a) with shaking at 220 rpm, unless stated otherwise. *L. lactis* MG1363 and *L. lactis NZ9000* were grown at 30°C in M17 medium (section 2.1.2c) supplemented with 0.5% (w/v) glucose, under static conditions. *E. coli* EP1300 (pCC1FOS) (Epicentre Biotechnologies, Madison, Wi, USA) was grown in Luria-Bertani (LB) medium containing 12.5 μg ml^{-1} chloramphenicol (Cml).

2.1.4b Bacterial growth conditions for co-incubations.

For co-incubation with Caco-2, overnight cultures of bacterial strains were used to inoculate M17 or LB broth to an OD_{600} of 0.05. Cultures were incubated at 37°C *(E. coli*) or 30°C (*L. lactis*) until exponential phase was reached (OD_{600} of 0.3 to 0.8). 1 ml aliquots of exponential phase cultures were centrifuged at 10,000 X *g* for 2 min. Supernatants were discarded and pellets were washed with 1 ml Phosphate Buffered Saline (PBS). Centrifugation was repeated and pellets were re-suspended in 1 ml PBS. Suspensions of bacteria equivalent to 0.1 OD_{600} were prepared by dilution in PBS. 45 μl aliquots of 0.1 OD_{600} suspensions were added to each well of 7 day cultured Caco-2 in 24 well plates, resulting in a multiplicity of infection (MOI) of 10, assuming 450,000 cells per confluent well of Caco-2. Co-incubations were performed at 37°C in atmosphere with 5% CO_2.

2.1.5 Caco-2 cell culture conditions.

Caco-2 cells were cultured in DMEM complete (Dulbecco's Minimal Eagles Medium containing 20% (v/v) foetal bovine serum (FBS), 1 X non-essential amino acids, 20 mM L-glutamine, 100 U ml^{-1} penicillin and 100 μg ml^{-1} streptomycin) at 37°C, 5% CO_2. Cells were routinely cultured in T25 flasks for 5 to 7 days following seeding at a density of 100,000 cells in 5 ml DMEM complete. For co-incubation experiments, cells were seeded at 20,000 cells per well with 1 ml DMEM complete per well in 24 well tissue culture plates. Cells were cultured for 6 days, at which point DMEM complete was removed, and monolayers were washed with 1 ml PBS per well. 1 ml DMEM (DMEM complete, without addition of penicillin and streptomycin) was added per well and plates were returned to 37°C, 5 % CO_2 for a further 24 h. To allow for vector selection, antibiotics (section 2.1.3a) were included at the following

concentrations: for *L. lactis* of tetracycline 5 μg ml^{-1} and of chloramphicol 10 μg ml^{-1} and for *E. coli* of kanamycin 50 μg ml^{-1} and of tetracycline 10 μg ml^{-1} (section 2.1.3a).

2.1.6 DNA agarose gel electrophoresis.

To visualise DNA samples, a 1 % (w/v) agarose gel was made by adding agarose powder to a 1X TAE buffer (40 mM Tris base, 0.114 % (v/v) glacial acetic acid and 1 mM EDTA; pH 8.0). Gels were stained with SYBR Safe and poured into a casting tray. Cast gels were placed in a gel electrophoresis tank and covered with 1X TAE buffer. DNA samples were mixed with 5 μl of crystal 5X loading buffer (Bioline) and added to a formed gel well. 5 μl of Hyperladder I was added each side of the DNA samples to give a standard size marker for samples. The DNA sample was separated across the gel by running a current at 100 V for 45 min using a Powerpack Consort E132. DNA was then visualised by excitation of the SYBR safe dye on a UV transilluminator and images were captured using a G:BOX gel imager (Syngene) and GeneSnap software (Syngene).

2.1.7 Polymerase chain reaction (PCR).

PCR was routinely carried out using Biomix PCR mix (Bioline). Where high fidelity amplification was required (amplification of genes and generation of constructs for expression), High velocity Taq DNA polymerase (Bioline) was used. The PCR mixes contain appropriate concentrations of dNTPs, ATP, DNA polymerase and $MgCl_2$ at 4.0 mM. Reaction mixes were composed of 1X Biomix/High velocity Taq, 10 pmol forward primer, 10 pmol reverse primer, 0.5 μl template DNA and PCR grade H_2O to 25 μl. PCR templates routinely used included: 1:1,000 dilution of miniprep DNA; 50 μl resuspension of a bacterial colony. PCR was carried out using an Eppendorf Mastercycler Gradient thermocycler programmed as follows: Cell lysis/initial denaturation at 95°C for 5 min, denaturation at 95°C for 1 min, annealing at 50°C to 65°C (optimised for each primer pair), extension at 72°C for 30 s per 1,000 bp to be amplified. The PCR cycle was repeated at least 30 times, followed by a final extension step carried out for 10 min at the appropriate temperature. After amplification DNA was stored at -20^{o}C or used in a downstream application.

Table 2.3 PCR thermocycler conditions.

Step	Temperature (^{o}C)	Tim (min)
Cell lysis	94	2

Denaturing	94	1
Annealing	54.5	0.5
Elongation	72	<1 – 2[a]
Final elongation	72	5

[a]Varied depending on the fragment size being amplified

2.1.8 Plasmid miniprep.

Plasmids were isolated using 5 ml overnight cultures of *E. coli* grown in LB broth or 15 ml overnight cultures of *L. lactis* grown in M17 broth with the appropriate selective agent for vector retention. Cultures were centrifuged at 8,000 X *g* for 10 min. Pellets were resuspended in 250 µl buffer P1. 250 µl buffer P2 was added and incubated at room temperature for 5 min. 350 µl buffer N3 was added, the tubes were inverted and centrifuged at 16,000 X *g* for 10 min. Supernatants were then applied to Qiaprep Spin columns and centrifuged at 16,000 X *g* for 1 min. 500 µl buffer PB was added to each column and centrifugation was repeated. Columns were washed with 750 µl buffer PE and dried by centrifuging at 16,000 X *g* for 5 min. Plasmid/cosmid DNA was eluted with 50 µl PCR grade H_2O. The concentration of the plasmid DNA was determined by NanoDrop (Thermo Scientific).

2.1.9 DNA purification using Wizard SV Gel/PCR Cleanup kit (Promega).

DNA was purified from PCR reactions, vector digests and gel excisions using SV clean-up kit. Membrane binding buffer was added to the DNA sample in a 1:1 ratio. (Gel fragments were heated to 65°C for 10 min in order to melt the agarose). Samples were then applied to SV spin columns and incubated at room temperature for 1 min. Columns were centrifuged at 16,000 X *g* for 1 min and washed with 750 µl membrane wash buffer. The wash was repeated with 500 µl membrane wash buffer and columns were dried by centrifugation at 16,000 X *g* for 10 min. DNA was eluted with 30 µl PCR grade H_2O (Thermo Scientific).

2.1.10 TOPO TA cloning of PCR products into pCR-XL-TOPO

PCR products were generated as outlined in section 2.1.7. Where non-specific amplification occurred, the specific product of interest was isolated by gel electrophoresis (section 2.1.6) and purified from a gel excision as outlined in section 2.1.9. 2 µl PCR product, 1 µl salt solution and 2 µl of PCR grade H_2O were mixed in a micro-centrifuge tube. 1 µl of TOPO ready vector was added and reaction was again

mixed. Samples were incubated at room temperature for 20 min to allow for integration into the vector. Transformation was carried out as described in section 2.2.4

2.1.11 Restriction endonuclease digestion.

Restriction digests were carried out using either Fermentas Fast Digest or Promega restriction endonucleases. Digests were prepared by mixing 2 µl restriction enzyme (~20 U), 2.5 µl appropriate 10X enzyme buffer, 2.5 µl bovine serum albumin (BSA) and 18 µl of miniprep DNA (~200 ng μl^{-1}). The reaction mix was incubated at 37°C for 2 h. The efficiency of digestion was assessed by agarose gel electrophoresis (section 2.1.6). Where appropriate, gel fragments were excised and purified using the Promega SV Gel/PCR Cleanup kit (section 2.1.9). All purified digested samples were stored at -20^{o}C until required.

2.1.12 Phosphatase treatment of vector DNA.

Digested vector DNA was treated with shrimp alkaline phosphatase (SAP, Promega) prior to ligation. 3 µl SAP (3U), 3 µl 10X SAP buffer and 25 µl purified vector (~25 ng μl^{-1}) were mixed and incubated at 37°C for 30 min. SAP was inactivated by incubating at 65°C for 15 min.

2.1.13 Ligations

Column purified insert DNA was ligated into purified, phosphatase treated vectors as follows: 2 µl T4 ligase Buffer (Fermentas), 2 µl 10X ligase buffer, 20 ng vector and a 3:1 molar ratio of insert to vector (~30 ng) were mixed and completed to 20 µl with PCR-grade H_2O. The reaction mixture was incubated at 16°C overnight for 16 h and was either used directly in a transformation reaction or stored at -20^{o}C.

2.1.14 Biofilm Assay

An overnight culture was grown at 37^{o}C for 16-18 h. One millilitre of the overnight was centrifuged at 8000 × *g* for 6 min and the pellet was washed once in 1 ml PBS (pH 7.0). The supernatant was discarded and the washed pellet was resuspended in 1 ml PBS. Five microlitres of the washed cells were added to 5 ml of either BHI broth or DM supplemented with glucose to a final concentration of 50 mM and vortexed gently. 200 µl of this resuspension was transferred to a flat bottomed 96-well tissue culture plate (Sarstedt) with eight technical replicates used. Sterile media was added

to each plate as a control. The plate was subsequently incubated statically at 37^{o}C for the required time. After incubation, the OD_{595} nm was recorded using a Tecan Sunrise absorbance reader. The media was carefully removed from all wells using a pipette and each well was washed 3 times with 200 µl PBS. The plate was allowed to dry at room temperature for 45 min and 150 µl of a 1% (w/v) crystal violet solution was added to each well. The plate was incubated at 37^{o}C for 30 min and the crystal violet was removed. The plate was washed 4 times in 200 µl PBS and finally 160 µl 95% ethanol was added. The plate was incubated for a further 30 min at room temperature and the OD_{595} nm was recorded.

2.2 Preparation and transformation of competent cells

2.2.1 Preparation of electrocompetent cells of *Lactococcus lactis*

L. lactis MG1363 and *L. lactis* NZ9000 electrocompetent cells were prepared according to the modified protocol by Gerber and Solioz (2007). Cells were grown in GM17 medium supplemented with 2.5% (w/v) glycine and 0.5 M sucrose (mGM17) (section 2.1.2d). An aliquot (100 µl culture from a glycerol stock) was inoculated into 5 ml mGM17, and grown overnight at 30°C. Then 1 ml of this culture was inoculated into 10 ml mGM17 and grown overnight under the same conditions. An aliquot (10 ml) from the overnight culture was then inoculated into 100 ml of mGM17, until the OD_{600} reached 0.2-0.3. Cells were harvested by centrifugation (4,000 x *g*/ 4°C/10 min, rotor SS-34) in 50 ml cold sterile Falcon tubes. Subsequent steps were performed on ice with ice-cold buffers. Cells were washed firstly with 50 ml EP1 buffer, followed by washing with 25 ml EP2 and 50 ml EP1 buffer. Cells were gently re-suspended in 1 ml EP1 buffer and aliquoted (40 µl volumes) into 0.2 ml Eppendorf tubes which were stored at -80°C.

2.2.2 Transformation of *L. lactis* by high voltage electroporation

Transformation was performed using a BioRad Gene Pulser apparatus, set to 2.0 kV, 25 µF and 200 Ω. An aliquot of 50 µl frozen cells was thawed on ice. The pulse was applied and immediately 960 µl of modified pre-chilled GM17 medium (supplemented with 20 mM $MgCl_2$ and 2 mM $CaCl_2$) was added. Cuvettes were placed on ice for 5 minutes, and then the cell suspension was transferred into a 1.5 ml eppendorf tube which was incubated at 30°C for 2 hours, followed by plating on selective GM17 agar plates.

2.2.3 Preparation of electrocompetent cells of *Escherichia coli*

E. coli electrocompetent cells were prepared according to Sambrook and Russell (2001). An overnight starter culture was subcultured (1/100) into 200 ml of LB medium (in a 2 L flask) and grown at 37°C with vigorous shaking at 220 rpm until an OD_{600} of 0.3-0.5 was reached. Cells were harvested by centrifugation in 50 ml cold, sterile Falcon tubes (4,000 x *g* /4°C/10 min, Sorvall, rotor SS-34). Subsequent steps were performed on ice with ice-cold solutions. Cells were washed twice with 50 ml sterile water and once with 50 ml sterile 10% glycerol. They were gently re-suspended in 0.5 ml 10% glycerol, aliquoted to 0.2 ml Eppendorf tubes and stored at -80°C. For genomic and metagenomic library construction commercially available electro-competent cells were purchased.

2.2.4 Transformation of *E. coli* by high voltage electroporation

Transformation was performed using a BioRad Gene Pulser apparatus, set to 1.7 kV, 25 µF and 200 Ω. A 40 µl aliquot of frozen cells was thawed on ice. Purified ligation mix (10 ng of vector) was added to the cell suspension, mixed gently and transferred into pre-chilled 1 mm electroporation cuvettes (Ingenio). The pulse was applied and immediately 960 µl of pre-warmed (37°C) SOC medium was added. The cell suspension was transferred into 15 ml Falcon tubes and was incubated at 37°C for 1 h at 220 rpm, followed by plating on selective BHI agar plates.

2.3 Methods to evaluate adherence efficiency

2.3.1 Analysis of bacterial adherence

2.3.1a Enumeration of adherence efficiency.

Bacterial strains were prepared for co-incubation as described in section 2.1.4b. Inoculum count was determined by serial dilution of the inoculum followed by plating on LB agar (*E. coli*) or M17 agar (*Lactococcus lactis*). Bacteria were added to triplicate wells of Caco-2 at an MOI of 10. Co-incubation was carried out for 1 h 30 min. The medium was discarded and non-adherent bacteria were removed by washing monolayers three times with 1 ml PBS per wash. Caco-2 cells were lysed by incubation in PBS + 1% (v/v) Triton X-100 for 10 min. The lysate was serially diluted and plated on LB agar (*E. coli*) or M17 agar (*L. lactis*). Plates were incubated at 37°C overnight or 30^{o}C (*L. lactis*) under static conditions, and the numbers of colony forming units

(cfu) were recorded. The adherence efficiency was calculated as the number of adherent cfu expressed as a percentage of the number of cfu in the inoculum.

2.4 Metagenomics library preparation and selection methods.

2.4.1 Fosmid library preparation.

Construction of the clone library was carried out using pCC1FOS^{TM} Fosmid Library Production kit (Epicentre Biotechnologies) according to the manufacturer's instructions. The Copy Control Fosmid Library Production Kit produces a complete and unbiased primary fosmid library. The kit utilizes a novel strategy of cloning randomly sheared, end-repaired DNA. Shearing the DNA leads to the generation of highly random DNA fragments in contrast to more biased libraries that result from fragmenting the DNA by partial restriction digests. DNA from gut was purified, sheared to fragments of approximately 40 kb. The 40 kb fragments were end-repaired to produce blunt, 5′-phosphorylated ends. The desired size range of end-repaired DNA was isolated using low melting point (LMP) agarose gel electrophoresis. Next, the blunt-ended DNA was purified from the LMP agarose gel and ligated to the Cloning-Ready Copy Control pCC1FOS vector. The ligated DNA was then packaged into a lamda phage and plated on EPI300-T1^{R} cells and allowed to grow overnight. The Copy Control fosmid clones were picked and induced to high copy number using the Copy Control Fosmid Autoinduction Solution (L-Arabinose). The DNA was then purified for sequencing, fingerprinting, subcloning, and other applications. A healthy, 27 year old, female volunteer, consuming a Western diet (omnivorous), provided a fresh faecal sample for this study. The volunteer did not take any antibiotics or other drugs known to influence the faecal microbiota within the six month period prior to the study. The faecal sample was placed in a sealable tub with autoclave bag and all the air was removed before sealing with cellotape. The stool sample was homogenized with a sterile spoon before weighing out 10g aliquots into sterile 50 ml screw top falcon tubes. The aliquoted samples were further homogenized by adding 20 ml PBS into the parafilm sealed falcon tubes and vortexing. The slurry was then split between two 50 ml falcon tubes and centrifuged at 1000 x *g* for 5 min to remove large particles. The bacterial cells from the faecal material were physically separated from any contamination by host cells by layering according to density using Nycodenz gradient solution (1.3 g ml^{-1} TE) (Nycodenz Axis Shield 1002424). DNA from the bacterial cells was then isolated using conventional DNA extraction methods[180,181,182]. Fosmid

library construction was performed using the CopyControl cloning system described by CopyControl^{TM} Fosmid Library Production Kit, Epicentre[183].

2.4.2 Shearing the metagenomics insert DNA

At least 2.5 µg (at a concentration of 500 ng μl^{-1}) of faecal extracted metagenomic DNA was randomly sheared by passing it through a 200-µl small-bore pipette tip. Shearing the DNA into approximately 40-kb fragments led to the generation of highly random DNA fragments in contrast to more biased libraries that result from partial restriction endonuclease digestion. The DNA was aspirated and expelled from the pipette tip 50-100 times. 1-2 µl of the DNA was examined on a 20-cm agarose gel using the fosmid control DNA as a size marker. When 10% or more of the genomic DNA migrated with the Fosmid Control DNA, end-repair of the insert DNA was performed. The extent of shearing of DNA was tested by pulse field gel electrophoresis (PFGE) analysis (voltage and ramp times recommended by the manufacturer for separation of 10-to 100-kb DNA). If a PFGE was unavailable, the sample was run on a 20-cm long, 1% standard agarose gel at 30-35 V overnight. If the DNA was sill too large, the DNA was aspirated and expelled from the pipette tip an additional 50 times. 1-2 µl of this DNA was examined by agarose gel electrophoresis.

2.4.3 End-Repair of the metagenomic insert DNA

The sheared metagenomic DNA fragments were end-repaired into blunt-ended, 5′-phosphorylated DNA. The end-repair reaction was scaled as dictated by the amount of DNA available. The end-repair reaction was performed in 80 µl total reaction volume with 8 µl 10X End-Repair buffer, 8 µl dNTP MIX (2.5 mM), 8 µl ATP (10 mM), 4 µl End-Repair enzyme mix, up to 20 µg sheared insert DNA (approximately 0.5 µg ul^{-1}) and made up to 80 µl with sterile dH_2O. The reaction was incubated at room temperature for 45 minutes. Gel loading buffer was added to the reaction mixture and further incubated at 70^{o}C for 10 min to inactivate the end-repair enzyme mix.

2.4.4 Size-Selection of the End-Repaired DNA

A 1% LMP agarose gel was prepared in 1 X TAE. A wide comb was used to load sufficient DNA into the gel. DNA size markers were loaded into the outside lanes of the gel. 100 ng of Fosmid Control DNA was loaded into each of the inner adjacent lanes of the gel. The end-repaired insert DNA was loaded in the lane(s) between the

Fosmid Control DNA lanes. The samples were resolved by gel electrophoresis at room temperature overnight at a constant voltage of 30-35V. Following electrophoresis, the outer lanes of the gel containing the DNA size markers, the Fosmid Control DNA, and a small portion of the next lane that contains the randomly sheared end-repaired genomic DNA were excised. The cut-off sides of the gel were stained with SYBR Gold (Invitrogen), and visualized with UV light. The position of the desired size DNA in the gel was marked using a pipet tip or a razor blade. The gel was reassembled and and a gel slice that was 2-to 4-mm below the position of the Fosmid Control DNA was excised. The gel slice was transferred to a tared, sterile, screw-cap tube for extraction, either by using the GELase method, or other desired method for isolating DNA from agarose gels. The size of the tube to be used was dictated by the size and number of gel slices digested with GELase enzyme.

2.4.5 Recovery of the Size-Fractionated DNA

The weight of the gel slice(s) were determined by weighing the tared tubes. It was assumed that 1 mg of solidified agarose would yield 1 µl of molten agarose upon melting. The GELase 50X Buffer was heated to 45°C. LMP agarose was melted by incubating the tube at 70°C for 10-15 min. The tube was quickly transferred to 45°C.The appropriate volume of warmed GELase 50X Buffer was added to a 1X final concentration. 1 U (1 µl) of GELase Enzyme was carefully added to the tube for each 100 µl of melted agarose. The melted agarose solution was kept at 45°C and gently mixed. The solution was then incubated at 45°C for at least 1 h (overnight incubation is acceptable). The reaction was transferred to 70°C for 10 min to inactivate the GELase enzyme. 500-µl aliquots of the solution was put into sterile, 1.5-ml microfuge tube(s). The tube(s) were chilled in an ice bath for 5 minutes. The tubes were centrifuged in a microcentrifuge at maximum speed (>10,000 x *g*) for 20 min to pellet any insoluble oligosaccharides. The "pellet" was gelatinous, and translucent to opaque. The upper 90%-95% of the supernatant, which contains the DNA, was carefully removed and put into a sterile 1.5-ml tube. The DNA was precipitated 1/10 volume of 3M sodium acetate (pH 7.0) was added and mixed gently. 2.5 v of ethanol was added and the tube was capped and mixed by gentle inversion. Precipitation was allowed to proceed for 10 min at RT. The precipitated DNA was then centrifuged for 20 min in a microcentrifuge, at top speed (>10,000 x *g*). The supernatant was carefully aspirated from the pelleted DNA. The pellet was washed twice with cold, 70% ethanol.

After the second 70% ethanol wash, the tube was carefully inverted and the pellet was allowed to air-dry for 5-10 min (longer dry times will make resuspension of the DNA difficult). The DNA pellet was gently re-suspended in TE Buffer.

2.4.6 Packaging of CopyControl Fosmid Clone

50 ml of LB broth + 10 mM $MgSO_4$ + 0.2% Maltose was inoculated with 0.5 ml of the EPI300-T1^R overnight culture. The flask was shaked at 37^{o}C to an A_{600} of 0.8-1.0 (~2 h). The cells were stored at 4^{o}C until further use (cells can be stored up to 72 h at 4^{o}C if necessary). One tube of a MaxPlax Lambda Packaging Extract was thawed on ice for a standard 10-µl ligation reaction. When the extract was thawed, 25 µl (one-half) of the extract was immediately transferred to a second 1.5 ml microfuge tube and placed on ice. The remaining 25 µl of the MaxPlax Packaging Extract was returned to a -70^{o}C freezer. 10 µl of a ligation reaction was added to the 25 µl of thawed extracts.The solution was mixed by pipetting several times. The tubes were briefly centrifuged to get all the liquid to the bottom. The packaging reaction was incubated at 30^{o}C for 2 h. After the reaction was complete, the remaining 25 µl of MaxPlax Lambda Packaging Extract was added. The reaction was incubated for an additional 2 h at 30^{o}C. At the end of the second incubation, Phage Dilution Buffer (PDB) was added to a 1 ml final volume. 25 µl of chloroform was added and gently mixed. The titer of the phage particles was determined and the fosmid library plated.

2.4.7 Storage of metagenomics library in *E. coli*

Fosmid clones were stored in pools by harvesting all colonies from selective plates using a sterile spreader and resuspending them in sterile LB containing 12 µg ml^{-1} chloramphenicol and 10% glycerol. The suspension was mixed and stored in aliquots at -80^{o}C.

2.4.8 Next Generation Sequencing

The output from the NGS illumina HiSeq was condensed into several fastq files for each of the fosmid clones analysed. Assembly quality was assessed by observing the length of the contigs, number of contigs, number of reads being used and N50 value. In general, the more sequence in long contigs and fewer contigs (as long as most of the reads are being used) the better. The N50 value indicates that 50% of the assembly is in contigs of the size of the N50 value or greater, and again the higher the better. The NGS downstream anaysis involved an output in the form of fastq data (Figure

4.4). The fastq NGS data is imported into the assembly software Geneious. The fastq zip file was dragged or dropped into Geneious or using the *import* button in the Geneious menu (Figure 4.4). Geneious was then used to align and *de novo* assemble the reads into contigs. An assembly report is generated for each assembly performed. Once the contigs were generated by the assembler, they were analysed using BLASTn to determine if the sequence is host genomic (*Escherichia coli*) DNA, vector (pCC1FOS) DNA or insert DNA. In this way, genomic and vector DNA were gradually eliminated to leave only insert DNA from each fosmid clone analysed.

2.4.9 Bioinformatic analysis of clones

All sequence data were generated using NGS Illumina HiSeq sequencing platform (Auburn University) (section 2.4.8). The open reading frames were detected by using the ORF search tool provided by NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), Glimmer (Gene Locator and Interpolated Markov ModelER) and BASys (Bacterial Annotation System, http:www.basys.ca). Homology searches were run against the GenBank database using the BLASTP algorithms. Prosite (http://expasy.org/tools/scanprosite/) and PFAM (http://pfam.sanger.ac.uk/) databases were utilised for protein analysis (conserved domains and internal repeats prediction). SignalP (http://www.cbs.dtu.dk/services/SignalP/) was used to predict the presence and location of signal peptide cleavage sites in amino acid sequences from different clones. Multiple amino acid sequences alignments were prepared with ClustalW.

2.4.10 Selection of adherent library clones using a single round of selection.

The metagenomic library was prepared for co-incubation as described in section 2.1.4b. Following 90 min of co-incubation in triplicate wells of Caco-2, the medium was removed and the monolayers were washed as per adherence assay described in section 2.3.1. 40 clones were selected from isolated colonies and streaked on LB agar + chloramphenicol. The clones were screened by adherence assay on triplicate wells of Caco-2 as described in methods section 2.3.1a, to identify any clones exhibiting increased levels of adherence.

2.4.11 Selection of adherent library clones using multiple rounds of selection.

In order to optimise selection of adherent library clones, the selection process was adapted. The lysate from a single round of selection was added to 40 ml of LB broth

+ chloramphenicol (12.5 μg ml^{-1}). The flask was incubated at 37°C (*E. coli*) with shaking overnight. 1 ml overnight culture was taken and centrifuged at 10,000 X *g* for 2 min. Supernatant was discarded and the pellet was washed with 1 ml PBS. Centrifugation was repeated and the pellet was resuspended in 1 ml PBS. The OD_{600} of the cell suspension was adjusted to 0.2 in 1 ml. 45 μl aliquots were then added to triplicate wells of 7 day cultured Caco-2 (MOI = 10). The adherence assay was repeated as per section 2.3.1a. The lysate from the second round of selection was used to inoculate a fresh 40 ml aliquot of LB broth + chloramphenicol (12.5 μg ml^{-1}). The flask was incubated at 37°C with shaking overnight and selection was repeated. In total selection was carried out 4 times. Again 40 clones were selected from isolated colonies and streaked on LB agar + chloramphenicol (12.5 μg ml^{-1}). Finally, the clones were screened by adherence assay on duplicate wells of Caco-2 as described in methods section 2.3.1, to identify any clones exhibiting increased levels of adherence.

Table 2.4 List of neoglycoconjugates (NGC) and glycoproteins on the microarray

Abbrev	**Neoglycoconjugate**
Fetuin	Fetuin
ASF	Asialofetuin
XGlcbBSA	Glc-β-ITC-BSA
Ov	Ovalbumin
RB	RNAse B
bovXferrin	Transferrin, bovine
XGalbBSA	Gal-β-ITC-BSA
α-C	α-Crystallin from bovine lens
AGP	alpha-1-acid glycoprotein, human
GlcNAcBSA	GlcNAc-BSA
LacNAcBSA	LacNAc-BSA
XManaBSA	Man-α-ITC-BSA
LNFPIBSA	Lacto-*N*-fucopentaose I-BSA
LNFPIIBSA	Lacto-*N*-fucopentaose II-BSA
LNFPIIIBSA	Lacto-*N*-fucopentaose III-BSA
FucaBSA	Fuc-a-4AP-BSA
SLexBSA14	3'Sialyl Lewis x-BSA
6SuLexBSA	6-Sulfo Lewis x-BSA
3SuLexBSA	3-Sulfo Lewis x-BSA
6SuLeaBSA	6-Sulfo Lewis a-BSA
3SuLeaBSA	3-Sulfo Lewis a-BSA
BGABSA	Blood Group A-BSA
BGBBSA	Blood Group B-BSA
3SLNBSA	3'SialylLacNAc-BSA
GGGNHSA	Galα1,3Galβ1,4GlcNAc-HSA
M3BSA	Manα1,3(Manα1,6)Man-BSA
3SLexBSA3	3'Sialyl Lewis x-BSA

LexBSA	Lewis x-BSA
LNDHIBSA	Lacto-*N*-difucohexaose I-BSA
2FLBSA	2'Fucosyllactose-BSA
3SFLBSA	3'Sialyl-3-fucosyllactose-BSA
Gb4GBSA	Galb1,4GalBSA
H2HSA	H-Type 2-APE-HSA
3SLacHSA	3´-Sialyllactose-APD-HSA
LNnTHSA	Lacto-N-neotetraose-APD-HSA
RhaBSA	L-Rhamnose-Sp14-BSA
6SLacHSA	6´-Sialyllactose-APD-HSA
LeyHSA	Lewis y-tetrasaccharide-APE-HSA
FucbBSA	Fuc-β-4AP-BSA
LNTHSA	Lacto-N-tetraose-APD-HSA
GlobNTHSA	Globo-N-tetraose-APD-HSA
3LexHSA	Tri-Lex-APE-HSA
DiLexHSA	Di-Lewisx-APE-HSA
DFPLNHHSA	Difucosyl-para-lacto-N-hexaose-APD-HSA
SLNFVHSA	Sialyl-LNF V-APD-HSA
MMLNnHHSA	Monofucosyl, monosialyllacto-N-neohexaose-APD-HSA
3FLeyHSA	Tri-fucosyl-Ley-heptasaccharide-APE-HSA
SLNnTHSA	Sialyl-LNnT-penta-APD-HSA
GM1HSA	GM1-pentasaccharide-APD-HSA
aGM1HSA	Asialo-GM1-tetrasaccharide-APD-HSA
Ga3GBSA	Galα1,3Gal-BSA
Ga2GBSA	Gala1,2GalBSA,

Table 2.5 **Lectin panel.** Specificities are obtained from *Handbook of plant lectins: properties and biomedical applications*[184].

Abbreviati	Source	Species	Common name	General binding	Print
ALA,	Plant	*Artocarpus*	Jack fruit lectin	Gal, Gal-β-(1,3)-GalNAc	Gal
RPbAI	Plant		Black locust lectin	Gal	Gal
PA-I	Bacteria	*Pseudomon*	Pseudomonas lectin	Gal, Gal derivatives	Gal
SNA-II	Plant	*Sambucus*	Sambucus lectin-II	Gal/GalNAc	Gal
SJA	Plant	*Sophora*	Pagoda tree lectin	β-GalNAc	Gal
DBA	Plant	*Dolichos*	Horse gram lectin	GalNAc	Gal
GHA	Plant	*Glechoma*	Ground ivy lectin	GalNAc	Gal
SBA	Plant	*Glycine*	Soy bean lectin	GalNAc	Gal
VVA-B4	Plant	*Vicia*	Hairy vetch lectin	GalNAc	Gal
BPA	Plant	*Bauhinia*	Camels foot tree	GalNAc/Gal	Gal
WFA	Plant	*Wisteria*	Japanese wisteria	GalNAc/sulfated GalNAc	Gal
HPA	Animal	*Helix*	Edible snail lectin	α-GalNAc	Gal
GSL-I-A4	Plant	*Griffonia*	Griffonia isolectin I	GalNac	Gal
ACA	Plant	*Amaranthus*	Amaranthin	Sialylated/Gal-β-(1,3)-	Lac
ABL	Fungus	*Agaricus*	Edible mushroom	Gal-β(1,3)-GalNAc,	Lac
PNA	Plant	*Arachis*	Peanut lectin	Gal-β(1,3)-GalNAc	Lac
GSL-II	Plant	*Griffonia*	Griffonia lectin II	GlcNAc	GlcNAc
sWGA	Plant	*Triticum*	Succinyl WGA	GlcNAc	GlcNAc
DSA	Plant	*Datura*	Jimson weed lectin	GlcNAc	GlcNAc

STA	Plant	*Solanum*	Potato lectin	GlcNAc oligomers	GlcNAc
LEL	Plant	*Lycopersicu*	Tomato lectin	GlcNAc-β-(1,4)-GlcNAc	GlcNAc
Calsepa	Plant	*Calystegia*	Bindweed lectin	Man/Maltose	Man
NPA	Plant	*Narcissus*	Daffodil lectin	α-(1,6)-Man	Man
GNA	Plant	*Galanthus*	Snowdrop lectin	Man-α-(1,3)-	Man
HHA	Plant	*Hippeastru*	Amaryllis agglutinin	Man-α-(1,3)-Man-α-(1,6)-	Man
ConA	Plant	*Canavalia*	Jack bean lectin	Man, Glc, GlcNAc	Man
Lch-B	Plant	*Lens*	Lentil isolectin B	Man, core fucosylated,	Man
Lch-A	Plant	*Lens*	Lentil isolectin A	Man/Glc	Man
PSA	Plant	*Pisum*	Pea lectin	Man, core fucosylated	Man
WGA	Plant	*Triticum*	Wheat germ	NeuAc/GlcNAc	GlcNAc
MAA	Plant	*Maackia*	Maackia agglutinin	Sialic acid-α-(2,3)-linked	Lac
SNA-I	Plant	*Sambucus*	Sambucus lectin I	Sialic acid-α-(2,6)-linked	Lac
CCA	Animal	*Cancer*	California crab	O-acetyl sialic acids	Lac
PHA-L	Plant	*Phaseolus*	Kidney bean	Tri-and tetraantennary	Lac
PCA	Plant	*Phaseolus*	Leukoagglutinin	leukoagglutinin	Lac
PHA-E	Plant	*Phaseolus*	Scarlet runner bean	GlcNAc in complex	Lac
RCA-1/20	Plant	*Ricinus*	Kidney bean	Biantennary with bisecting	Lac
CAP	Plant	*Cicer*	Erythroagglutinin	Biantennary with bisecting	Lac
CAA	Plant	*Caragana*	Castor bean lectin I	Gal-β-(1,4)-GlcNAc	Gal
ECA	Plant	*Erythrina*	Chickpea lectin	Complex oligosaccharides	Lac
AAL	Fungi	*Aleuria*	Pea tree Lectin	Gal-β-(1,4)-GlcNAc	Lac
LTA	Plant	*Lotus*	Cocks comb/coral	Gal-β-(1,4)-GlcNAc	Lac
UEA-I	Plant	*Ulex*	Orange peel fungus	Fuc-α-(1,6), -α-(1,3)	Fuc
EEA	Plant	*Euonymous*	Lotus lectin	Fuc-α-(1,3)	Fuc
GSL-I-B4	Plant	*Griffonia*	Gorse lectin I	Fuc-α-(1,2)	Fuc
MPA	Plant	*Maclura*	Spindle tree lectin	α-Gal	Gal
GSL-I-B4	Plant	*Vigna*	Griffonia lectin I	α-Gal	Gal
MPA	Plant	*Marasmius*	Osage orange lectin	α-Gal	Gal
VRA	Plant	*Vigna*	Mung bean lectin	α-Gal	Gal
MOA	Fungus	*Marasmius*	Fairy ring	α-Gal	Gal

Table 2.6 **List of mucins and glycoproteins printed and printing conditions.** Mucins are colored according to species, e.g., bovine, green; equine, blue, red; human.

Code	Mucin source	Printed (mg ml^{-1})
M3	Bovine cervix	0.3
M4	Bovine cervix	0.5
M6	Equine stomach	0.25
M10	Ovine abomasum	0.25
M11	E12	0.5

M12	Ovine descending colon	0.25
M15	Bovine c-v	0.25
M18	Ovine spiral colon	0.5
M30	Ovine cervix (Suffolk)	0.5
M31	Ovine cervix (Belcare)	0.5
M32	Ovine cervix (Suffolk)	0.5
M33	Ovine cervix (Beclare)	0.5
M34	Chicken small intestine	0.25
M35	Ovine jejunum	0.5
M36	Ovine duodenum	0.25
M37	Porcine gastric	0.33
M39	Equine (pregnant) cervix	0.4
M41	Chicken large intestine	0.25
M48	Bovine c-v	0.5
M49	Bovine c-v	0.25
M52	Equine duodenum	0.3
M53	Equine trachea	0.5
M55	Deer jejunum	0.25
M56	Deer spiral colon	0.75
M57	Bovine abomasum	0.25
M58	Bovine duodenum	0.5
M59	Equine jejunum	0.25
M60	Equine left ventral colon	0.25
M61	Equine spiral colon	0.25
M62	Deer duodenum	0.5
M63	Bovine trachea	0.75

M64	Bovine endometrium	0.4
M65	Equine right ventral colon	0.25
M66	Equine dorsal colon	0.25
M67	Deer abomasum	0.25
M70	Chicken ceca	0.5
M72	LS174T	0.5
ASF	Asialofetuin	0.25
RB	RNase B	0.25
Fetuin	Fetuin	0.25
Xferrin	Transferrin	0.25
Ovomuc	Ovomucoid	0.25
PBST	PBS 0.01% Tween 20	0.25

2.5 Carbohydrate-based microarray characterization of putative adherent clones

2.5.1 Materials

Aldehyde ES microarray slides were from Pierce Biotechnology (Rockford, IL). Nexterion® Slide H microarray slides were purchased from Schott AG (Germany). Poly-L-lysine slides, BSA (cat. no. A7638, ≥99%), glycopyranosyl PITC and 4AP derivatives and goat anti-rabbit IgG labelled with Atto 633 were purchased from Sigma-Aldrich Co. (Dublin, Ireland). The BSA was periodate-treated25 and used for neoglycoconjugate synthesis and microarray slide blocking. The bicinchoninic acid (BCA) Protein Assay Kit and sulfosuccinimidyl-4-(Nmaleimidomethyl)cyclohexane-1-carboxylate (sulfo-SMCC) were from Pierce Biotechnology (Thermo Fisher Scientific Inc., Dublin, Ireland) and rabbit anti-cow albumin polyclonal antibody was from Dako (Glostrup, Denmark). Pure tetramethylrhodamine-(TRITC-) labelled lectins were from EY Laboratories, Inc. (San Mateo, CA). Glycoproteins and all other reagents were from Sigma-Aldrich Co. unless otherwise noted and were of the highest grade available.

2.5.2 Lectin microarray

A panel of lectins (Table 2.5) were printed on Nexterion® Slide H microarray slides in a 62% (+/-2%) humidity environment using a using a SciFLEXARRAYER S3 equipped with a 90 μm uncoated glass nozzle. Lectins were diluted to their print concentration of 0.5 mg ml^{-1} in phosphate buffered saline (PBS; 10 mM sodium phosphate, 137 mM NaCl, 2 mM KCl, 2 mM KH2PO4, pH 7.4) supplemented with 1 mM of their respective haptenic simple sugars to protect their carbohydrate recognition domains during conjugation to the slide surface. Each microarray slide was printed with eight replicate subarrays, with each lectin (probe) spotted in replicates of six. Slides were incubated in a humidity chamber overnight after printing to facilitate complete conjugation and were then blocked with 100 mM ethanolamine in 50 mM sodium borate, pH 8.0, washed four times in PBS-T for 2 min each, once with PBS and centrifuged dry (500 × *g*, 5 min). Printing and performance of the conjugated lectins was verified by incubation with fluorescently labelled glycoproteins. Microarray slides were stored dry with desiccant at 4°C until use.

2.5.3 Mucin microarray

Mucins (Table 2.6) and glycoproteins (probes) were dissolved in PBS, pH 7.4 (1.37 M NaCl, 0.027 M KCl, 0.02 M KH2PO4, and appropriate mixture of 0.1 M Na2HPO4 and NaH2PO4 for correct pH) and piezoelectrically printed onto Nexterion slide H microarray slides using a SciFLEXARRAYER S3 (Scienion AG, Germany) equipped with a 90 μm uncoated glass nozzle at 62% humidity (±2% tolerance). Probes were printed in replicates of six, approximately 1 nl per feature, 312 features per subarray. Slides were incubated in a humidity chamber overnight after printing to facilitate conjugation, and remaining functional groups were capped with 100 mM ethanolamine in 50 mM sodium borate, pH 8.0, for 1 h. Slides were washed three times in PBS with 0.05% Tween-20 (PBST), once in PBS, centrifuged dry (1500 rpm, 5 min), and stored at 4 °C with desiccant until use.

2.5.4 Neoglycoconjugate (NGC) microarray

Neoglycoconjugate (Table 2.4) (NGC) array slides were prepared as outlined in Kilcoyne *et al*. (2012)[57]. Poly-L-lysine slides were functionalised with sulfhydryl-reactive maleimide groups by incubation of the slide surface with 10 mM sulfo-SMCC prepared in PBS, pH 7.4 (137mM NaCl, 2.7 mM KCL, 2 mM, KH2PO4 and adjusted

to correct pH with Na2HPO4 and NaH2PO4) for 1 h at RT, in a humidity chamber. Functionalised slide were then washed twice in dH_2O, centrifuged dry (500 × *g*, 5 min) and stored at 4°C with desiccant until required. NGCs were prepared at a concentration of 1 mg ml^{-1} in PBS, pH 7.4, based on BCA assay (NGCs) or mass (glycoproteins) and printed at approximately 1 nL per feature on functionalised poly-L-lysine slides, or Nexterion Slide H microarray slides (Schott AG, Mainz, Germany), in humidity (62% ± 2%), using a SciFLEXARRAYER S3 (Scienion AG, Germany) equipped with a 90 µm uncoated glass nozzle. Each slide was printed with six subarrays, with each probe spotted in replicates of twelve per subarray. Slides were incubated in a humid atmosphere overnight after printing for complete conjugation. Functional groups on poly-L-lysine were deactivated or capped with 1.4 mM β-mercaptoethanol in PBS pH 7.4 for 1 h at RT and the Nexterion Slide H with 100 mM ethanolamine in 50 mM sodium borate, pH 8 for 1 h at RT. The slides were washed with PBS, pH 7.4 with 0.05% Tween-20 (PBS-T) three times and once with PBS. Slides were then centrifuges dry (500 × *g*, 5 min) and stored dry at 4°C with desiccant until required.

2.5.5 Preparation of bacterial fosmid clones for array analysis.

10 ml fosmid clones FC3, FC21 and control strain EPI300 (pCC1FOS) were grown overnight at 37°C for 24 h in the presence of L-arabinose and chloramphenicol. The clones were harvested, pelleted by centrifugation (5,000 x *g*, 5 min) and washed three times in Tris-buffered saline supplemented with Ca^{2+} and Mg^{2+} ions (TBS; 20 mM Tris-HCl, 100 mM NaCl, 1 mM CaCl2, 1 mM MgCl2, pH 7.2). Bacteria were diluted to an OD_{600} of 1.0 (~5 × 1010 cfu ml^{-1}) in TBS, and 1 ml of the bacterial suspension was pelleted by centrifugation and then resuspended in 0.5 ml of TBS. Bacteria were incubated with 20 µM SYTO 82 (Life Technologies, Carlsbad, CA) orange fluorescent cell-permeable nucleic acid dye (λex 541 nm, λem 560 nm) at 37°C for 1 h with rotation. After incubation, the fluorescently labelled cell suspension was washed seven times in TBS to remove excess dye, and finally resuspended in 0.5 ml of TBS with 0.05% Tween-20 (TBS-T) for immediate use on the microarrays. To determine the optimum SYTO 82 concentration for each strain, different concentrations of dye were added to the washed bacterial suspensions to give a final range of 5 - 100 µM. After incubation, 100 µl of the bacterial samples, both with and without post-staining wash steps, was loaded into 96 well black microtitre plate and fluorescence was measured on a SpectraMax M5e microplate reader (Molecular Devices, Inc., Berkshire, UK).

The optimal concentration was determined based on maximum fluorescence. Similar fluorescence intensities were noted when bacteria were incubated in TBS or PBS.

2.5.6 Carbohydrate-based microarray data extraction and analysis.

Raw intensity values were extracted from the image files using GenePix Pro v6.1.0.4 (Molecular Devices, Berkshire, U.K.) and a proprietary *.gal file using adaptive diameter (70-130%) circular alignment based on 230 μm features and exported as text to Excel (version 2007, Microsoft) where all data calculations were performed. Local background was subtracted and background-corrected median feature intensity (F543median-B543) was used for each feature intensity value. The median of twelve replicate spots per subarray was handled as a single data point for graphical and statistical analysis. Data intensities across three replicate microarray slides were normalised to the per-subarray total intensity mean and binding data was presented in histogram form of average intensity with standard deviation of three experimental replicates. The significance of inhibition data was evaluated using a standard Student's t-test (paired, two-tailed).

2.6 Expression of *$MapA_{Ri}$* gene in *Lactococcus lactis* NICE system

2.6.1 PCR of *$mapA_{Ri}$* gene

The *$mapA_{Ri}$* gene was amplified by PCR in the presence of *Roseburia intestinalis* genomic DNA and Roseburia forward primer 5'-cgggatcctgaactac-3' and Roseburia reverse primer 5'-cggaattctgttttaatac-3'. PCR was carried out (section 2.1.7, Table 2.3) under the following conditions: 30 cycles of denaturation at 94^{o}C for 1 min, annealing at 55^{o}C for 1 min, and extension at 72^{o}C for 1 min.

2.6.2 Construction of the recombinant NZ9000/pPTPi-*$mapA_{Ri}$*

The *$mapA_{Ri}$* gene was inserted into the *Escherichia coli*- (*E. coli*) *Lactococcus lactis* (*L. lactis*) shuttle vector pPTPi and the recombinant plasmid was transformed into *L. lactis* NZ9000 using voltage electroporation. The transformants were selected on plates containing 5 μg ml^{-1} of tetracycline.

2.6.3 Expression of *$MapA_{Ri}$* protein in recombinant *L. lactis*.

NZ9000/pPTPi-*$mapA_{Ri}$* was cultured overnight, then 400 μL of which was transferred, respectively, into 2 tubes of 10 ml GM17 liquid medium containing 5 μg ml^{-1} tetracycline. After NZ9000/pPTPi-*$mapA_{Ri}$* was cultured for 4 h, nisin stock was added,

respectively, to the final concentration 10 ng ml^{-1} and 20 ng ml^{-1} to induce the expression of $MapA_{Ri}$. NZ9000/pPTPi-*mapA*$_{Ri}$ was centrifuged after 4 h and the supernatant protein was collected and quantified using Bradford amount. $MapA_{Ri}$ expressed in *Lactococcus lactis* was detected by SDS-PAGE.

Chapter 3:

Functional metagenomics approach to identify novel glycan binding bacterial adhesins encoded by the human gut metagenome

3.1 Introduction

The human intestinal microbiota encodes multiple critical functions that have an impact on human health. To understand the full impact of this microbial community on human health, both the phylogenetic profile of human microbial communities and the functional capacity of their members must be characterized. Progress has been made towards these ends using direct bacterial culture, 16S RNA sequencing, shotgut metagenomics sequencing, PCR probing for specific genes, and chemical profiling of microbial metabolites. Thus far, these approaches have yielded incredible insights into the functional capacity of the gut microbiota.

The gut microbiota is a dynamic and complex community consisting largely of obligate anaerobes that are recalcitrant to standard cultivation techniques. Traditional estimates indicate that only 15-20% of the gastrointestinal microbiota are culturable, precluding direct characterization of the majority of bacterial species[98, 98, 185]. A recent report by Goodman et al. (2011)[32] demonstrated using high-throughput 16S sequencing in combination with extensive anaerobic culturing, that up to 56% of gastrointestinal microbial species are culturable[32]. Although this represents a dramatic improvement over standard culturing techniques, there remains a significant proportion of unculturable organisms that must be characterized. These unculturable microorganisms are often very diverse organisms and distantly related to the cultured microorganisms. This immense diversity of the encoded genes of the human gut microbiota necessitated the development of novel molecular, microbiological, and genomic tools[152]. Functional metagenomics is one such culture-independent technique, used for decades to study environmental microorganisms, but relatively recently applied to the study of the human commensal microbiota. Functional metagenomics neatly complements the aforementioned techniques currently used to characterize the human microbiota[161].

Functional metagenomics screens characterize the functional capacity of a microbial community, independent of identity to known genes, by subjecting the metagenome to functional assays in a genetically malleable host[161] (section 1.8.2). These screens were originally proposed as a method to characterize the unculturable fraction of soil microbiota[186, 159] and successfully used for years to characterize the functional diversity of microbes in a variety of environments [163]. This technique has relatively

recently been adapted to characterize functions of the human microbial communities, representing a cross-pollination between environmental microbiology and biomedical science.

Functional metagenomics screening method is based on clone libraries (Figure 3.2) containing metagenomics DNA from a microbial community, bypassing the need to directly culture fastidious organisms. Instead, clone libraries are constructed by extracting and shearing DNA from a sample of a microbial community, then cloning the fragmented DNA into a relevant vector, and subsequently transforming this vector into a suitable host strain[2] (Figure 3.2). Once a library is constructed, it can be functionally screened depending on the function of interest. The range of functional screens that can be performed with an environmental library is immense. As long as there is an assay for the function of interest and a bacterial heterologous host lacking that function, a functional screen is possible (Figure 1.6).

Functional screens most often comprise screening for a gain-of-function conferred by a cloned environmental DNA fragment. If the heterologous host is already proficient in the function of interest, then a mutant defective in the function is used for the gain-of-function screen. Several examples of such functional screens of environemtal libraries have been published, including the identification of a unique salt tolerant gene, *stl*A[168], and the identification of operons responsible for enhanced intestinal colonization by murine gut commensal microbes[169] (section 1.9). Using the functional metagenomics approach, it is possible to identify genes encoding a variety of functions such as antibiotic resistance, cell adhesion, metabolism of complex compounds, and modulation of eukaryotic cells. Subsequent sequencing and *in silico* analysis of the DNA inserts from isolated clones provides imformation about the source of the genes and the putative mechanisms of action of their products[161].

In this project, a functional metagenomic approach (section 1.8.2) was undertaken to construct metagenomic libraries for screening of novel proteoglycan binding elements encoded in the human gut microbial metagenome. Two types of libraries were generated; a small-insert library in plasmid vectors (less than 10 kb) and large-insert library in fosmid vectors (up to 40 kb). Small insert libraries are maintained on high copy number plasmids with strong promoters, which are usually used in activity screens where a single gene is responsible for the activity. Large insert libraries are

suitable for the identification of multigene encoded products, operons, and entire biochemical pathways and usually utilize low-copy number or inducible vectors. Inducible vectors are advantageous in that the library may be stably maintained at low copy, but can be induced to high copy for downstream applications.Both libraries were obtained from healthy adult faeces. An *in vitro* assay of bacterial adhesion onto Caco-2 epithelial cells (model for intestinal epithelium) was used to select for adherent clones. Caco-2 cells are used extensively in studies of bacterial adherence[187, 188]. Caco-2 cells express several markers that are characteristic of normal small intestinal villus cells and have played a major role in studies on the mechanisms of adherence and invasion of many bacteria[189]. Caco-2 cells thus provide a good system for studying not only mechanisms through which species in the normal microbiota adhere to the intestine, but also how these bacteria may interact with bacteria that compete in the same ecosystem[189]. Furthermore, an *in vitro* assay of bacterial adhesion onto Caco-2 cells was used as a standard adherence assay because previous studies by Letourneau and colleagues demonstrated the relative robustness of this assay in illustrating the adherence of a specific adhesin to Caco-2 cells[190]. The next sections will describe the construction and validation of two human gut metagenomics libraries.

RESULTS:

3.1.1 Fosmid Metagenomic Library Construction

A healthy 27 year old, female volunteer consuming a Western diet (omnivorous) provided fresh faecal samples for this study. The volunteer did not take any antibiotics or other drugs known to influence the faecal microbiota for six months prior to the study. A fosmid metagenomic library with the desired average insert size of 42 kb was constructed using the Copy Control- Fosmid Library Production system (section 2.4.1).

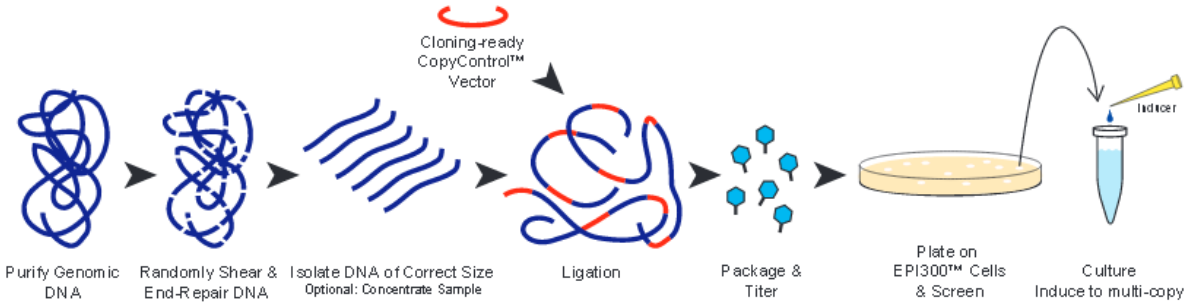


Figure 3.2 **Schematic representation of fosmid cloning procedure as represented by Epicentre**. Production of a Copy Control™ Fosmid library and subsequent induction of clones to high-copy number [183]. Diagram adapted from Strain *et al*., 2012.

The Copy Control Cloning System is based on a technology developed by Dr. Waclaw Szybalski that combines the clone stability afforded by single-copy cloning with the advantages of high yields of DNA obtained by "on-demand" induction of the clones to high-copy number[191]. The Copy Control pCC1FOS (Figure 3.3) fosmid vector contains both a single copy origin of replication as well as an *oriV* high-copy origin of DNA replication. Initiation of replication from *ori*V requires the *trf*A gene product that is supplied by a second system component. The Copy Control pCC1FOS vector is completely inactive in commonly used heterologous hosts, because they do not produce the TrfA replication protein upon which replication at *ori*V depends. To supply the TrfA protein, special heterologous hosts (Phage T-1 Resistant EPI300™-T1^{R} *E. coli* Plating Strain) were constructed, in which synthesis of copy-up TrfA mutant protein is very tightly controlled by the P_{araBAD} (P_{BAD}) promoter and AraC protein[191] (present in Phage T-1 Resistant EPI300™-T1^{R} *E. coli* Plating Strain). In this way, the system permits the conditional amplification of fosmid vectors (with or

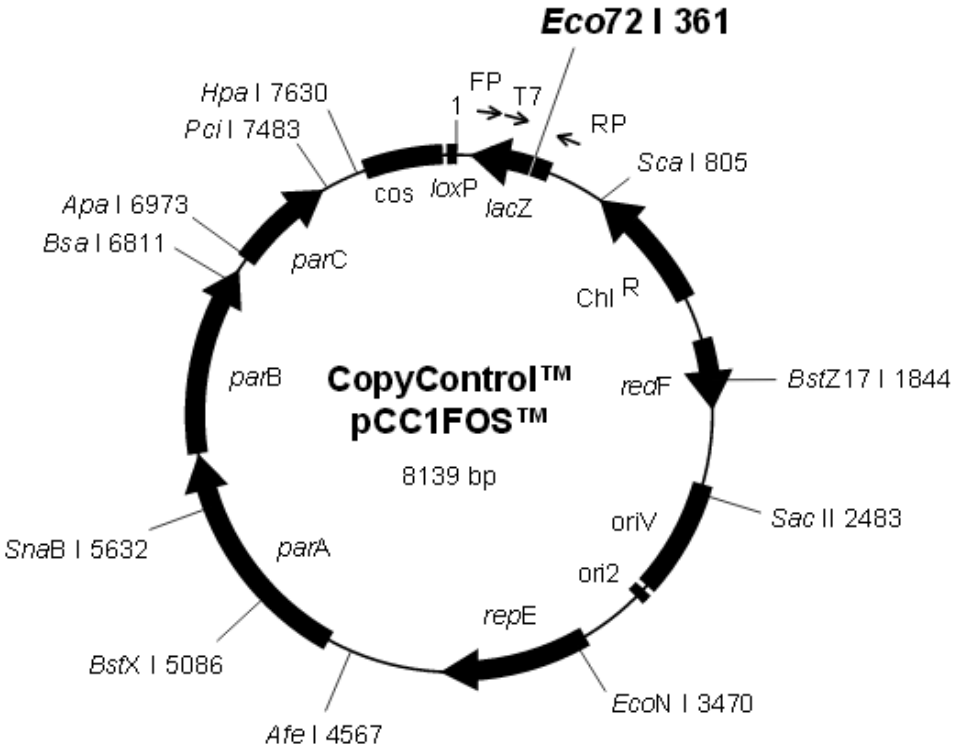
without) inserts consisting of the *ori*V vector and a host supplying (only upon induction by addition of arabinose) a copy-up mutant of TrfA protein. In this system, the *ori*V clone is maintained at single-copy level, but when the synthesis of the TrfA is induced by the addition of arabinose, DNA is amplified up to 100-fold[191]. Therefore, fosmid clones can be maintained at low copy number with high stability in the absence of arabinose, or induced to high copy level (10-200 copies per cell) by adding L-arabinose to the growth medium, which triggers *trf*A expression, activating *ori*V, resulting in an increase in copy number[183]. On demand induction of pCC1FOS fosmid clones can improve DNA yields for sequencing, fingerprinting, subcloning, *in vitro* transcription, and other applications[191]. Moreover, high copy number often results in increased gene expression, sometimes allowing the identification of hits in a library that otherwise (single-copy condition) would not have been detected.

$$N = \frac{\ln(1 - P)}{\ln (1-i/Gn)}$$

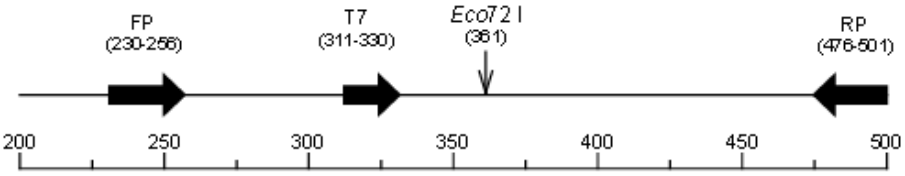
The above formula is used to determine the approximate number of clones required in a metagenomics library to obtain representative coverage of the microbiota in an ecosystem. N = number of clones required, P = probablitiy (usually 95% or 99%), i = average insert size, G = average genome size and n = number of different genomes. For example, the number of clones required to ensure a 99% probability of a given DNA sequence of *E. coli* (genome = 4.7 Mb) being contained within a fosmid library composed of 40-kb inserts is:

$N = \ln (1 - 0.99) / \ln (1 - [4 \times 10^4 \text{ bases} / 4.7 \times 10^6 \text{ bases}]) = -4.61 / -0.01 = 461$ clones

Based on the formula, a library consisting of 285,333 clones is required to obtain representative coverage of the gut microbiota, assuming 1000 species present in the human gut.



Note: Not all restriction enzymes that cut only once are indicated above.
See Appendix E for complete restriction information.
Primers are not drawn to scale.



FP = pCC1™/pEpiFOS™ Foward Sequencing Primer	5' GGATGTGCTGCAAGGCGATTAAGTTGG 3'
RP = pCC1™/pEpiFOS™ Reverse Sequencing Primer	5' CTCGTATGTTGTGTGGAATTGTGAGC 3'
T7 = T7 Promoter Primer	5' TAATACGACTCACTATAGGG 3'

Figure 3.3 **pCC1FOS vector map**. Features of the Copy Control pCC1FOS™ Vector include chloramphenicol resistance as an antibiotic selectable marker, *E. coli* F factor-based partitioning and single-copy origin of replication, *ori*V high-copy origin of replication, Bacteriophage lambda *cos* site for lambda packaging or lambda-terminase cleavage, Bacteriophage P1 *lox*P site for Cre-recombinase cleavage and Bacteriophage T7 RNA polymerase promoter flanking the cloning site. Diagram adapted from Strain *et al*., 2012[192].

3.1.2 Small Fragment Metagenomic Library Construction

The small fragment library was constructed by Szczepanska and Louis (personal communication) at the Rowett Institute of Nutrition and Health at the University of Aberdeen. A healthy, 27 year old, female volunteer, consuming a Western diet (omnivorous), provided a fresh faecal sample for this study. The volunteer did not take any antibiotics or other drugs known to influence the faecal microbiota for six months prior to the study. The desired average insert size was 5-10 kb, therefore commercially available kits applying chemical or mechanical cell disruption were used (QIAamp DNA stool kit QIAGEN, Extract Master Fecal DNA, Epicentre and Fast®DNA Spin kit for soil). A dried down sample of the metagenomic DNA was provided to us. The initial goal was to screen the library in a gram negative host *E. coli* HB101 and then transform it into the gram positive host, *Lactococcus lactis* MG1363. *E. coli* HB101 was chosen as a heterologous host because it possesses a deletion in the Type 1 fimbrial operon rendering the organism non-adherent. Moreover, *E. coli* HB101 is deficient in restriction endonucleases and recombinases making it an effective host for retention of heterologous DNA [193]. This strain has been used as a control for adhesion and as a host of expression of heterologous adhesins [194].

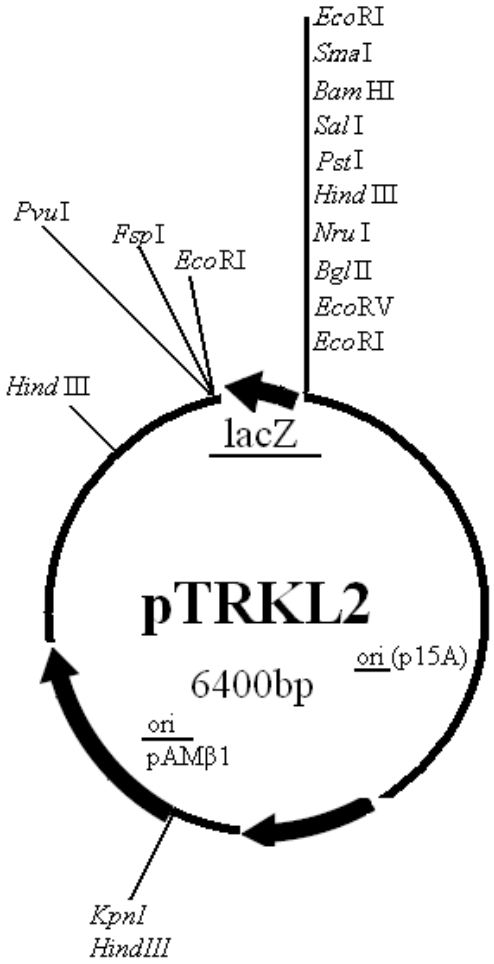


Figure 3.4 PTRKL2 vector map.
Low-copy-number shuttle cloning vector constructed by incorporating the *Escherichia coli* P15A plasmid origin of replication into the pAMPl-derived vectors. Structurally stable in *Lactococcus lactis* and *E. coli*[195]. Diagram adapted from O'Sullivan *et al*., 1993.

3.1.3 Transformation of small fragment library into *Lactococcus lactis* MG1363

Several alternative heterologous hosts have been used for metagenomic library screening. This project investigated the use of the Gram-positive bacterium *Lactococcus lactis* as an alternative heterologous host for a functional screening of the small fragment library. *L. lactis* is widely recognized as an attractive alternative heterologous host to the *E. coli* expression system[196]. It has been reported that there is a positive correlation between codon usage of an individual gene and the surrogate host [170]. Therefore, it was hypothesized that genes derived from the low %G+C gut

firmicutes would be expressed in a host such as *L. lactis* which is a Gram-positive low %G+C coccus belonging to the phylum Firmicutes [196]. *Lactococcus lactis* MG1363 was chosen in the present study to extend the expression host range for the functional screening of metagenomic libraries. *L. lactis* MG1363 is a plasmid-free strain [197] obtained through the sequential protoplasting and regeneration of *L. lactis* NCD0712, which led to the creation of strains that retain none of the plasmids (MG1363) or only Plp712 (MG1299)[197]. *L. lactis* MG1363 does not produce any extracellular proteases, which is beneficial if the adhesin product is secreted. As a result of these factors, *L. lactis* MG1363 has been employed as a cell factory for the production of macromolecules (bacteriocins), enzymes and metabolites [198]. *L. lactis* MG1363 is a model micro-organism used worldwide and alongside other lactic acid bacteria. It is classified as a "generally regarded as a safe" (GRAS) organism.

3.2 Validation of two metagenomic DNA libraries

Two metagenomic libraries were used in this study to characterise and identify novel glycan binding bacterial adhesins encoded by the human gut microbiota. Faecal metagenomic DNA used to construct both libraries was derived from a healthy female volunteer who had been on a western diet. The small fragment library was estimated to contain ~ 250,000 clones with insert fragment sizes ranging from 5-10-kb.The fosmid library was estimated to contain 42,000 clones with insert fragment sizes ranging from 25 to 45 kb.

To evaluate the metagenomic diversity of the small fragment library, 40 random clones were selected and digested with several restriction endonucleases (section 2.1.11) and then observed on a 1% agarose gel electrophoresis (section 2.1.6). Restriction digestion was used to verify that the clones contain genetically diverse DNA fragments based on the restriction patterns observed. Although, the restriction digest analysis did not give a 100% representation of the small fragment library, it did provide a glimpse into the type of genetic diversity present in the library.

An example of the *Eco*RI restriction digest of 4 out of the 40 clones is illustrated in Figure 3.5. The restriction endonuclease *Eco*RI cleaves two sites adjacent to the insert on the $pTRKL_2$ plasmid (Figure 3.4). Many of the remaining 36 clones exhibited distinct restriction endonuclease profiles when digested with *Eco*RI (data not shown). The results of the four digested clones (Figure 3.5; Lane 2, 4, 6 & 8) show 4 distinct banding patterns indicating the presence of 4 genetically distinct clones. Clones with exactly the same banding patterns are likely to contain identical inserts. Lanes 2, 4, 6 and 8 exhibit different numbers and sizes of DNA bands. The restriction profiles of all 4 clones show the unique $pTRKL_2$ vector band at 6400 bp (Figure 3.4). Lane 2 shows two bands at 1,200 bp (insert band) and 6400 bp (Vector band). Lane 4 shows three bands at 4,000 bp, 1350 bp and 6400 bp (Vector band). Lane 6 shows two distinct bands at 1750 bp and 6400 bp (Vector band). Finally, lane 8 shows two distinct bands at 3,300 bp and 6400 bp (Vector band). Although many of the 36 clones exhibited distinct restriction profiles, 39% of the 40 clones analysed had no insert (data not shown). This suggests that the vector was not digested properly during preparation of library, or that self ligation occurred during cloning, or that the insert was too small to detect, or that the insert was exactly the same size as the vector. Overall, based on the

agarose gel image (Figure 3.5) and other restriction digests (*Hind*III and *Eco*RI; data not shown) the small fragment library was deemed to be genetically diverse. Profiling by restriction digest represents an inexpensive means of comparing library clones in order to identify genetically distinct clones from a population.

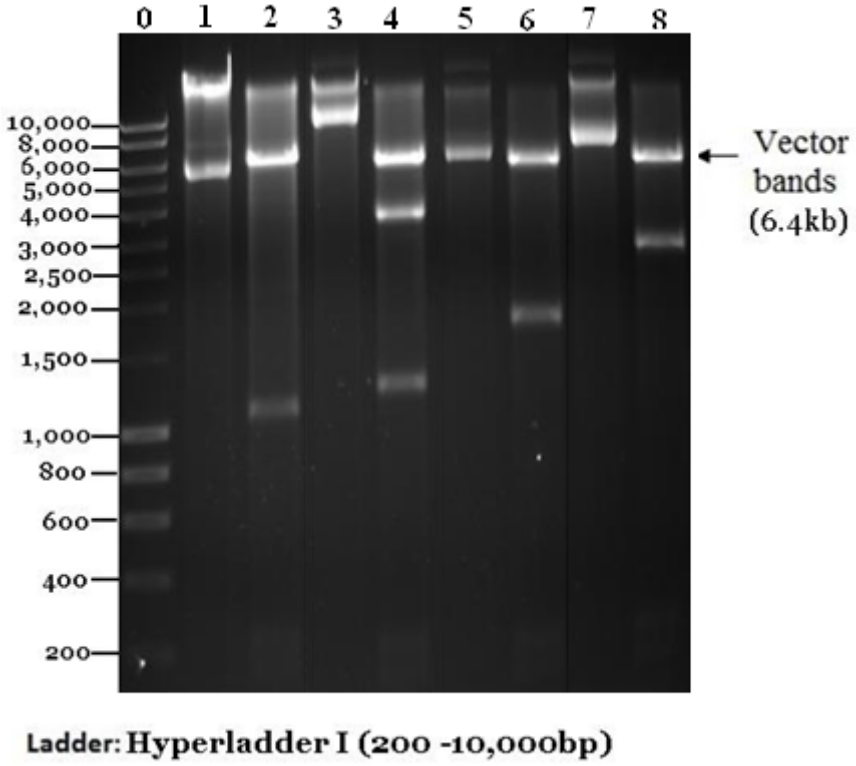


Figure 3.5 **Evaluation of the metagenomic diversity of the small fragment library by restriction digest using the endonuclease enzyme *Eco*RI**. 1% Agarose gel electrophoresis of restriction endonuclease activity (*Eco*RI) showing molecular weight marker in lane 0. Lane 1=Uncut clone 1, Lane 2=*Eco*RI cut clone 1, Lane 3= Uncut clone 2, Lane 4= *Eco*RI cut clone 2, Lane 5= Uncut cone 3, Lane 6=*Eco*RI cut clone 3, Lane 7 = Uncut clone 4, Lane 8 = *Eco*RI cut clone 4.

3.2.1 Characterization of microbial diversity of small fragment library by sequence analysis

To characterize the microbial diversity of the small fragment library, DNA sequences of 4 clones confirmed to have inserts were end-sequenced by the GATC Biotech Company (Table 3.1). The 4 clones here are not the same clones analysed in Figure 3.5. Good quality sequences were retrieved after end-sequencing of all 4 clones. The end sequences of each clone had an average length of 400 bp. Information regarding the sequence identity of the clones was determined by comparing each sequence to the

non-redundant (nr) protein sequences in the online database Basic Local Alignment Search Tool (BLAST). Before performing the BLAST searches, the pTLRK2 vector sequence were truncated from the rest of the nucleotide sequence for each of the four clones. The first column in Table 3.1 lists the plasmid clone query DNA sequence. The next column described the output from BLAST searches. The next two columns are associated with the statistics of the database search. BLAST uses statistics to sort through all the hits, shows only the best and explains why it is the best. The "Total Score" is a number generated when BLAST finds multiple, but not joined, sections of similarity between the query and the hit. If the Max Score is equal to the Total Score, then only a single alignment is present. If the Total Score is larger than the Max Score, then multiple alignments must be present and their individual scores have contributed to the Total Score. The Query Coverage is the percentage generated if BLAST can align all the nucleotides of the query against a hit, then that would be 100%. The maximum identity calculates the percentage identity between the query and the hit in a nucleotide-to-nucleotide alignment [199].

Table 3.1 **Characterization of microbial diversity of small fragment library by end-sequence analysis of 4 random small fragment library plasmid clones.** Tabular display of BLASTn hits**.** Abbreviations: n/a (not applicable) No Significant Similarity Found = indicates that no homologues for the gene product was found following BLAST searches of NCBI database.

Plasmid Clone	Description	Max. Score	Total Score	Query Cover	Max. Identity
1Forward	No Significant Similarity Found	n/a	n/a	n/a	n/a
1Reverse	PICI19r Cloning vector	80.5	80.5	4%	100%
2Forward	*Eubacterium rectale* DSM 17629 draft	1738	2029	99%	99%
2Reverse	*Roseburia intestinalis* M50/1 draft genome	1408	2497	95%	95%
3Forward	No Significant Similarity Found	n/a	n/a	n/a	n/a
3Reverse	*Bacteroides fragilis* NCTC 9343, complete genome	154	254	47%	95%
4Forward	No Significant Similarity Found	n/a	n/a	n/a	n/a

4Reverse	PICI19r Cloning Vector	89.8	89.8	5%	100%

In this study, two out of the four clones (Table 3.1, plasmid clones 2 & 3) were found to have sequences that were homologous (>95%) to the sequences from three common bacterial inhabitants of the human gut. The forward end-sequence of clone 2 was found to be 99% identical to a sequence from the gut bacterium *Eubacterium rectale.* The reverse end-sequence of this clone was found to be 95% identical to a sequence from *Roseburia intestinalis*. A possible explanation for obtaining hits from two different organisms with the same clone is that one of the sequences (either the forward or reverse) is conserved in both these organisms. *Eubacterium rectale* is a gram-negative, non-spore forming rod whose predominance in the gut is comparable to the dominant gram-negative inhabitants such as *Bacteroides thetaiotaomicron*. *Roseburia intestinalis* is a saccharolytic, butyrate-producing bacterium that was first isolated from human faeces. It is an anaerobic, gram-positive bacterium that is slightly curved and rod shaped [200].

The forward end-sequence of clone 3 was found to have "No Significant Similarity Found" to any of the sequences in the BLASTn database. One possible explanation for this is that the insert metagenomic DNA in these clones are novel sequences that have no homology to any of the sequences in the database. In contrast, the reverse end-sequence of clone 3 shows 95% sequence homology to the dominant gut bacterium, *Bacteroides fragilis*. *Bacteroides fragilis* is an obligate anaerobic, gram negative, rod shaped bacterium that is part of the normal microflora of the human gut[201]. Like all members of the Bacteroidetes phylum (section 1.2.1), it is an efficient glycan forager[201].

Both clones 1 and 4 are genetically different from each other yet exhibit "No Significant Similarity" for the forward end sequence and 100% identity to the cloning vector P1C19r. This suggests that the inserts in clone 1 and 4 have no homologs in the database. In conclusion, based on these four sequenced clones, it is possible to infer that the small fragment library is diverse.

3.2.2 Evaluation of fosmid library diversity by restriction digestion

To evaluate the genetic diversity of the fosmid library, restriction digestion of 7 random clones (each carrying ~ 40 kb inserts) was performed using *BamHI* restriction endonuclease. The digested fragments were visualized on a 1% agarose electrophoresis gel (Figure 3.6). Genetic diversity was determined by observing the restriction patterns of each clone on the gel. The more diverse the library, the more diverse the restriction pattern of each clone on the gel. All 7 clones in Figure 3.6 appear to have distinct and different restriction profiles. Overall, based on the agarose gel image (Figure 3.6) and other restriction digests (not shown) the fosmid library was determined to be genetically diverse. There are more banding patterns present in Figure 3.6 than in Figure 3.5 illustrating the large difference in insert size of these two libraries.

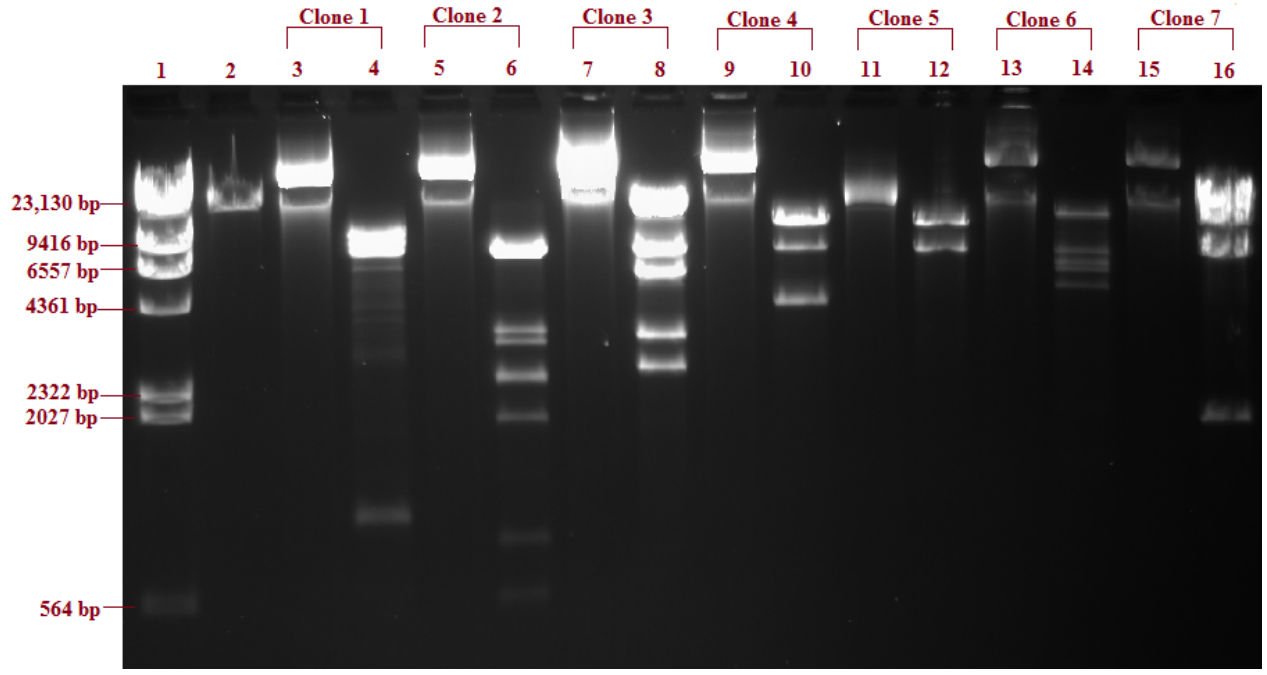


Figure 3.6 **Evaluation of fosmid library metagenomic diversity by restriction digest of 7 random fosmid clones;** 1% Agarose gel electrophoresis of restriction endonuclease activity. Lane 1 = DNA molecular weight marker II, Lane 2 = random, uncut clone 1, Lane 3 = uncut clone 1, Lane 4 = *BamHI* digested clone 1. The pCC1FOS vector band is located at 8139 bp. The same digestion pattern is observed for all 7 clones. Approximately 50 clones were digested with *Bam*HI, but these data are not shown. The vast majority of fosmid clones digested with *Bam*HI depicted different restriction profiles, confirming a diverse genetic library.

3.2.3 Characterization of microbial diversity of fosmid library by sequence analysis

To characterize the microbial diversity of the fosmid library, 10 random clones (distinct from the 7 clones' decribed in Figure 3.6) were end-sequenced and analysed by GATC Biotech Company (Table 3.2). Information regarding the sequence identity of the clones was determined by comparing each sequence to the non-redundant protein sequences in the online database Basic Local Alignment Search Tool (BLAST). Before performing the BLAST searches, the pCC1FOS vector sequence was truncated from the rest of the nucleotide sequence for all ten clones. Four (Fosmid clone 1, 3, 6 & 9) out of the 10 fosmid clones sequenced were shown to have homology with sequences from the dominant gut bacterium, *Bifidobacterium adolescentis*. Fosmid clone 5 and 10 showed homology with a sequence in a species of uncultured *sphingobacterium*. Uncultured *sphingobacterium* belongs to the dominant gut phylum Bacteroidetes and the class of *sphingobacteria.* An important feature of bacteria that belong to this class is the presence of high concentrations of sphingophospholipids in the cellular lipid components. They are gram-negative, non-fermentative bacilli, ubiquitous in nature and rarely involved in human infections [202].

Table 3.2 **Characterization of microbial diversity of fosmid library by sequence analysis** 10 random fosmid clones were end-sequenced and analysed using BLAST. Abbreviations: No Significant Similarity Found = indicates that no homologues for the gene product was found following BLAST searches of NCBI database.

Fosmid Clone	Description	Max. Score	Total Score	Query Cover	Max. Identity
1Forward	*Bifidobacterium adolescentis* ATCC 15703 DNA, complete genome	689	689	90%	82%
1Reverse	*Bifidobacterium adolescentis* ATCC 15703 DNA, complete genome	1136	1136	91%	89%
3Forward	*Bifidobacterium adolescentis* ATCC 15703 DNA, complete genome	1310	1310	89%	99%

4Forward	*Candidatus Snodgrassella* sp. T3 2 35043 genomic sequence	143	143	86%	99%
5Forward	Uncultured sphingobacteria bacterium, whole genome shotgun sequence	122	122	83%	97%
5Reverse	Uncultured sphingobacteria bacterium, whole genome shotgun sequence	159	159	72%	98%
6Forward	*Bifidobacterium adolescentis* ATCC 15703 DNA, complete genome	628	628	60%	99%
6Reverse	*Bifidobacterium adolescentis* ATCC 15703 DNA, complete genome	785	785	86%	98%
7Forward	Glycine max clone 11-5-136 amadillo/beta-catenin-like repeat protein amino acid transfer protein, and alpha-SNAP protein genes	538	699	32%	99%
7Reverse	No Significant Similarity Found	n/a	n/a	n/a	n/a
8Forward	*Candidatus Snodgrassella* sp. T3 2 35043 genomic sequence	147	147	8%	97%
8Reverse	*Escherichia coli* O104:H4 str.2009EL-2071, complete sequence	571	571	98%	93%
9Forward	*Bifidobacterium adolescentis* ATCC 15703 DNA, complete genome	364	364	79%	97%
9Reverse	Environmental Halophage Ehp-14, partial genome	161	161	7%	99%
10Forward	N/A	n/a	n/a	n/a	n/a
10Reverse	Uncultured sphingobacteria bacterium, whole genome shotgun sequence	141	141	12%	98%

According to Table 3.2, 32% of the query sequence (forward end-sequence) of clone 7 shows homology to a sequence from a genetically modified soybean (Glycine max) designated clone 11-5-136. The query sequence is showing homology to beta-catenin-like repeat protein, amino acid transfer protein and alpha-SNAP protein genes. In contrast to the forward sequence of clone 7, there is no significant similarity found for the reverse sequence. Overall, the data confirm the genetic diversity of both metagenomics libraries by restriction digestion and end-sequencing of several clones from each library. The next sections will describe the functional screening of the metagenomics library to identify adherent clones.

3.3 *In vitro* assay of bacterial adhesion onto mammalian epithelial cells

In order to gain a general understanding of the adherence properties of our control strain (Phage T-1 Resistant EPI300™-T1^{R} *E. coli* Plating Strain) and in order to confirm a positive control for adherence, a number of adhesion assays were performed. Adherence assays were carried out by incubating Caco-2 cells (Human Adenocarcinoma cells) with exponential phase *E. coli* bacteria at a Multiplicity of Infection (MOI) of 10 for 90 min[190] (section 2.3.1a). Exponential phase bacteria were used as high levels of protein are produced during this growth phase and also to ensure a high ratio of live: dead bacterial cells. An MOI of 10 was used as higher MOIs were found to cause rapid cell lysis in certain strains[203]. Adherent bacteria were selected as outlined in section 2.4.10 and were enumerated by spread plating.

To verify Caco-2 cells as a reliable *in vitro* adhesion model, an adhesion assay (section 2.3.1a) of *Vibrio parahaemolyticus* and EPI300 onto 7 day old Caco-2 epithelial cells was performed (Figure 3.7). *V. parahaemolyticus* has been previously shown to adhere highly to Caco-2 cells with an adherence efficiency of 19 % of the inoculum on 7 day old Caco-2 cells [203], 20-fold higher than the control strain (non-adherent *E. coli HB101*). In this study, *V. parahaemolyticus* displayed an adherence efficiency of 2.3% (Figure 3.7), seven fold higher than the EPI300 control strain (Figure 3.7). The lower adherence efficiency observed in our study as compared to the work of O'Boyle *et al*.,(2013)[203] may be explained by differences in culture conditions and buffer compositions. These findings demonstrate that *V. parahaemolyticus* is highly adherent to Caco-2 cells and serves as a good control for our adhesion assays. In addition, this experiment demonstrates the level of adherence of our control strain *E. coli* EPI300

(without pCC1FOS vector). As shown in Figure 3.7, EPI300 displays an adherence efficiency of 0.2%, seven fold lower than *V. parahaemolyticus*. All in all, we were able to demonstrate low adherence efficiency for our control strain as well as a good positive control for our *in vitro* adhesion assays. Having established a good positive control and a good baseline control strain, the next section will deal with adhesion assays of the Fosmid library and EPI300 strain to determine if they are statistically significantly different from one another. Additionally, the level of background adherence of the control strain against the fosmid library will be assessed.

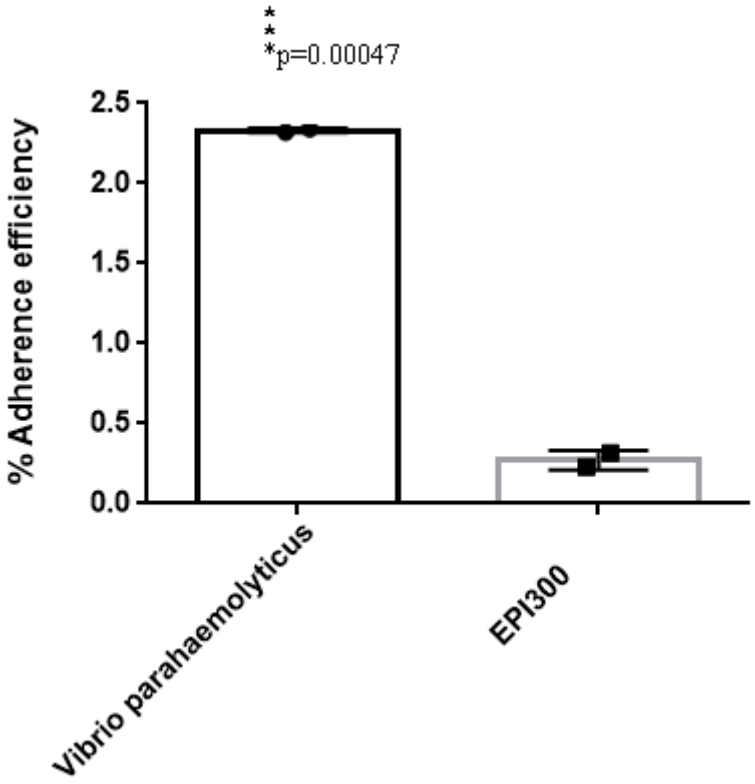


Figure 3.7 ***Vibrio parahaemolyticus* serves as a positive adherence control as it has been found to adhere highly to Caco-2 cells**. Percent adherence efficiency of the control strain EPI300 vs. *Vibrio parahaemolyticus* on 7 day old Caco-2 cells. The bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, ***, $P<0.001$. The squares and circles represent the standard deviation observed in *V. parahaemolyticus* (circles) and EPI300 (squares).

3.3.1 Analysis of Fosmid Library vs. EPI300 control strain without induction

One of the aims of this study was to determine if there was statistically significant differences between the adherence detected for the fosmid library and the control strain EPI300 WT (without empty fosmid vector). Additionally, we aimed to determine the level of background adherence by the EPI300 control strain. The assay was carried out as previously described (section 2.4.10) by performing five different experiments on 5 separate days.

As seen in Figure 3.8, the percent adherence efficiency of the control EPI300 strain varies significantly in each experiment on the 5 different days. On day 1, an adherence efficiency of 0.18% was obtained for the EPI300 strain. By the 5^{th} day, a value of 0.0023% (>78 fold decrease) was obtained for the EPI300 control strain. The adhesion of the fosmid library ranged between 0.22% (Day 1) to 0.008% (Day 5). These experiments demonstrate the high inter-experimental variation that can occur using *in vitro* adhesion assay of bacterial adhesion onto Caco-2 cells.

Interestingly, in four out of the five experiments, there was no significant difference between the fosmid library and the control. These results are not surprising as the likelihood of observing significant increase in the adherence efficiency of the library is quite low. Approximately, 0.01% of each genome of the estimated 420 individual genomes in the fosmid library is represented in each clone. The probability of each clone expressing an adhesin is low. Therefore, the probability of observing significant adherence efficiency in the fosmid library above that of the control strain is also low. Taken together, these results demonstrate that the fosmid library and EPI300 control strain are not statistically significantly different from one another. This trend is apparent on 4 of the 5 experiments performed and suggests that the results are reproducible. However, it is clear that the level of inter-experimental variation is high. As discussed in section 3.1.1, it is possible to increase gene expression of fosmid library clones by "on-demand" induction of the clones to high-copy number. In the next section, induction of the fosmid library was performed to determine levels of adherence efficiency as compared to the control.

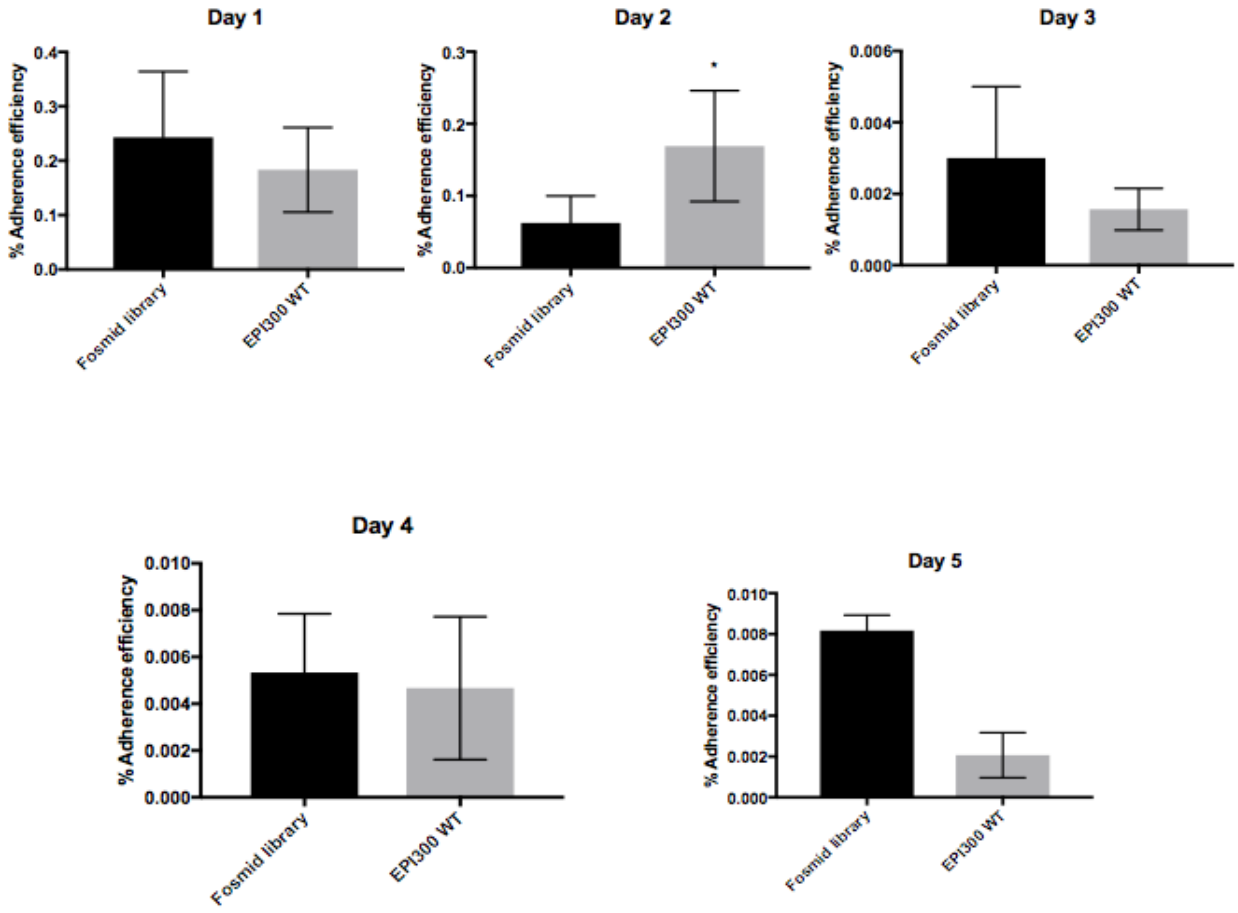


Figure 3.8 **No significant difference was found between the fosmid libray and EPI300 control strain (without empty fosmid).** Percent adherence efficiency of the fosmid library vs. EPI300 WT control strain on 7 day old Caco-2 cells without induction. The experiment was performed 5 times on 5 different days. The error bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, * $p<0.005$.

3.3.2 Induction of the fosmid library does not significantly increase adherence efficiency

To study the effect of induction on the adherence efficiency of the library and the control, assays with cultures grown in 1% arabinose on 7 day old Caco-2 cells were performed (2 experiments performed on 2 separate days). If there were adhesive clones present in the fosmid library, induction of the copy number of such clones was expected to increase expression of the adhesive phenotype. It was hypothesized that induction will increase expression of any adhesive clones and thus increase adherence in our fosmid library.

When comparing the fosmid library adherence in the presence (Figure 3.9) and absence of arabinose (Figure 3.8), the results indicate more adherence of both the control and the fosmid library in the presence of arabinose as compared to in its absence. However, when comparing the adherence of the fosmid library to the control strain in the presence of arabinose, induction of the fosmid library and control strain seems to decrease the adherence efficiency of the fosmid library on both Day 1 and Day 2 (Figure 3.9). Arabinose induction seems to inhibit the adherence of the fosmid library when compared to the control[191]. A possible explanation for this result (Figure 3.9) is that retention of multiple copies of a large foreign DNA molecule, coupled with high levels of expression increases the likelihood of toxicity in heterologous hosts[191, 192]. This explains why the EPI300 control strain is not showing similar inhibition of adherence because it lacks the pCC1FOS vector (Figure 3.3) and is unable to be induced.

Thus far, this study has demonstrated that *V. parahaemolyticus* is a good positive control (Figure 3.7) for adhesion assays, it has demonstrated that EPI300 is a good baseline control strain as it has low adherence efficiency (Figure 3.8). This study has also established that adhesion assays have high inter- and intra- experimental variation. The variation is both intrinsic and experimental. Sources of experimental variation will be discussed in the discussion section of this chapter. Moreover, this study has demonstrated that the fosmid library and the EPI300 control strain are not statistically significantly different from one another in their adherence to 7 day old Caco-2 cells. Finally, induction by arabinose may cause toxicity and stress to heterologous host harbouring the insert and thus affect adherence[157, 191]. In the next sections, a series of selection rounds of the two metagenomic libraries will be performed in order to isolate clones expressing functional adhesins.

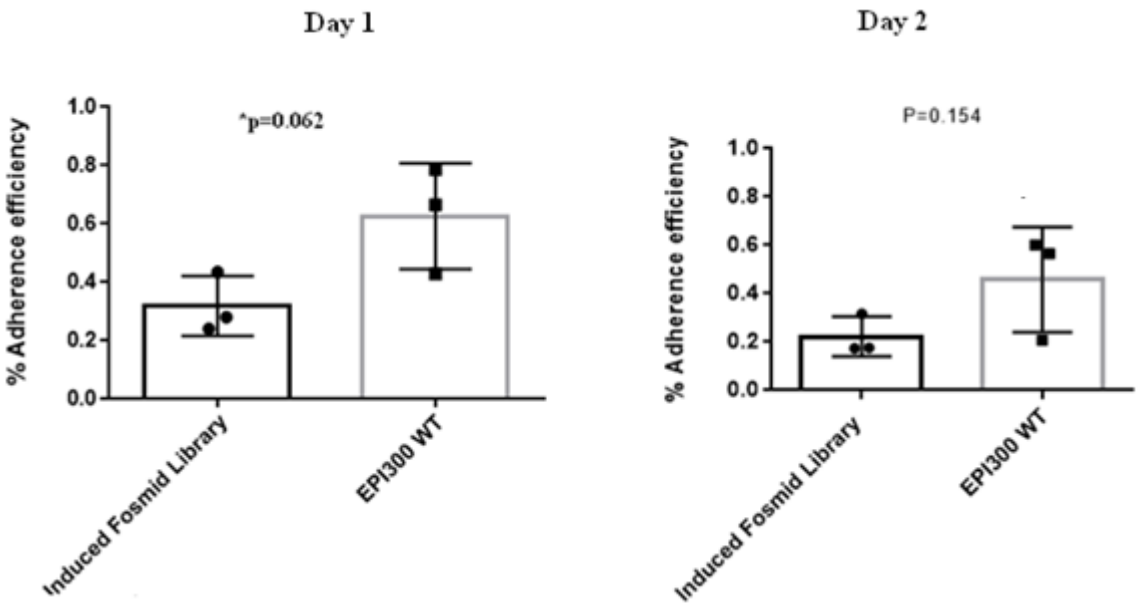


Figure 3.9 **Fosmid library Induction with Arabinose.** Percent adherence efficiency of the induced (+arabinose) fosmid library vs. EPI300 WT control strain on 7 day old Caco-2 cells (+arabinose). The experiment was repeated 2 times on 2 different days. The error bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, *, $P<0.05$.

3.3.3 Enrichment of adhesive clones of both metagenomic DNA libraries

Following preparation of the metagenomic library, a range of selections were carried out in order to isolate clones expressing functional adhesins. Selection of adhesive clones was performed by incubating the library with Caco-2 cells. After incubation, the cells were washed to remove non-adherent cells and the Caco-2 were lysed to enable plating of adhered clones within the lysate (section 2.3.1a). Caco-2 cells were grown for seven days prior to library selections. This resulted in the formation of polarised differentiated Caco-2 cell monolayers. This polarisation coupled with the formation of tight junctions led to the establishment of a monolayer which is highly representative of the human intestinal epithelium.

To identify adherent clones, multiple rounds of adhesion assays to selectively enrich adhesive clones was performed (section 2.4.11). *In vitro* adhesion assay selection rounds were performed by repeating the assay with the retrieved bacteria scraped from their respective agar petri dishes. The selection cycle was repeated by co-incubating Caco-2 cells with bacteria from the former selection round to produce newly bound bacteria, which was then used for further selection rounds until a significant

enrichment of adhesive clone was achieved. The number of adhesive clones should increase with every selection round. However, in some cases a significant enrichment of the library with adhesive clones was noticeable after the first rounds of selection. In this study, we performed five selection rounds. The majority of non-specific binding clones were lost in the first selection round. Over the following rounds, the numbers plateaued. Once specific binding clones were selected and transferred to the next selection round, they, hopefully, would not be lost in the selection step and the number of colonies would increase dramatically.

It was expected that the use of five rounds of selection would amplify the selection of adhesive clones, however after five rounds of selection the adherence efficiency of the small fragment metagenomic library was lower than that observed after a single round in all 3 wells (Figure 3.10). The adherence efficiency did increase with 2 rounds of selection (rounds 4 and 5 in wells 2 and 3). It may be the case that selection of adhesive clones was indeed amplified over two rounds of selection, but that subsequent rounds of selection resulted in outgrowth of truly adhesive clones by clones which were not actively adhesive but exhibited a more rapid growth rate than the adhesive population[203]. A similar pattern was observed in the control strain *E. coli* Top 10. After 5 rounds of selection, the adherence efficiency of the control *E. coli* Top 10 was lower than that observed after a single round in each of the three wells. The adherence efficiency did increase with two rounds of selection (rounds 4 and 5 in wells 1 and 3).

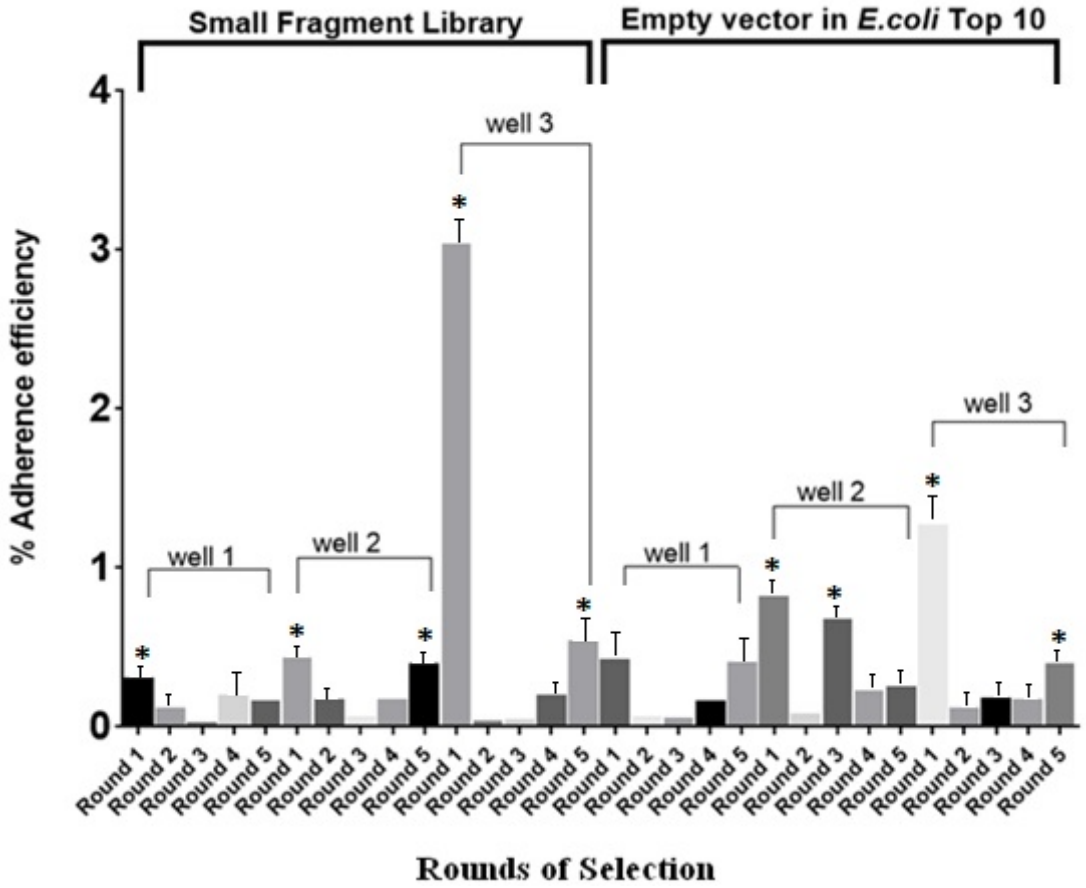


Figure 3.10 ***In vitro*** **selection and enrichment of adhesive clones from the small fragment metagenomic library.** The graph above represents the percent adherence efficiency of the small fragment library vs. *E. coli* TOP10 with an empty vector (pTRKL2) (Figure 3.4) on 7 day old Caco-2 cells. Five rounds of selection were carried out in an attempt to isolate adherent clones. This experiment was repeated several times (data not shown). The error bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, *, $P<0.05$.

Another possibility is that while adhesins may have been expressed in the initial selection process, the growth of retrieved bacteria may have resulted in the bacteria adapting so as to reduce expression of a potentially deleterious protein. This would result in selection of clones which may indeed carry adhesins but which have become incapable of expressing such proteins. Taken together, the use of multiple rounds of selection did not lead to the amplification selection towards adhesive clones, likely due to the introduction of bias towards rapidly growing clones. The same rounds of

selection were performed with the induced fosmid library and EPI300 control strain (Figure 3.11) with very similar results.

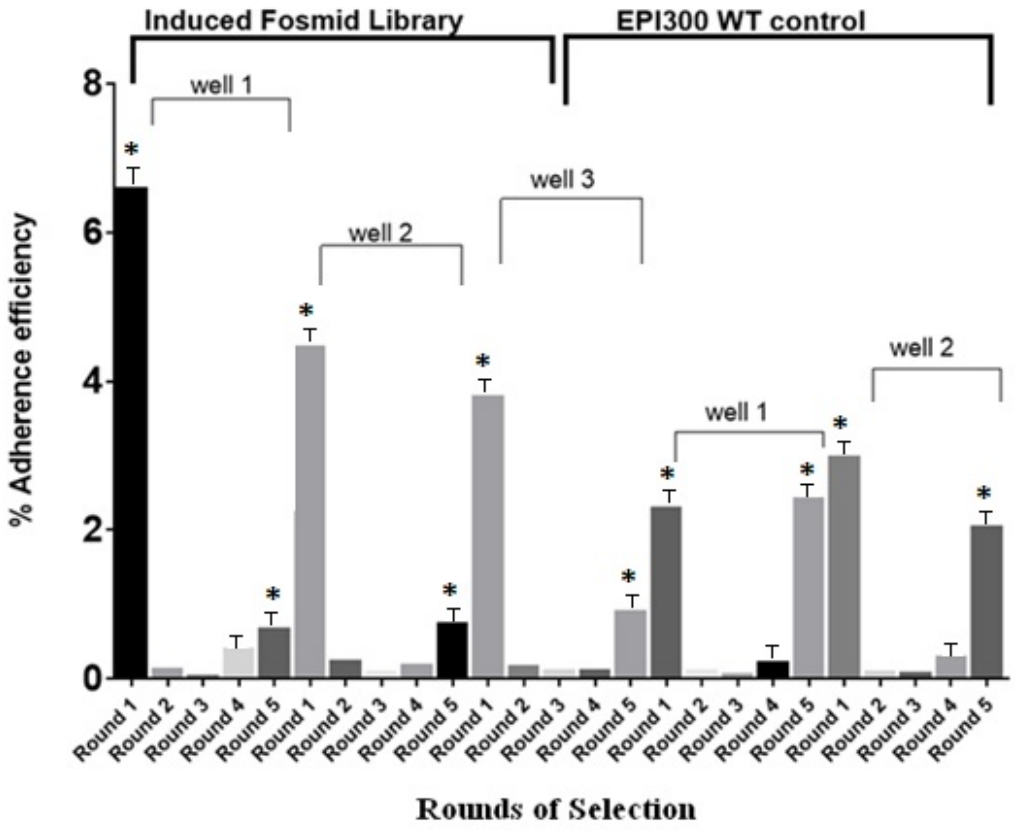
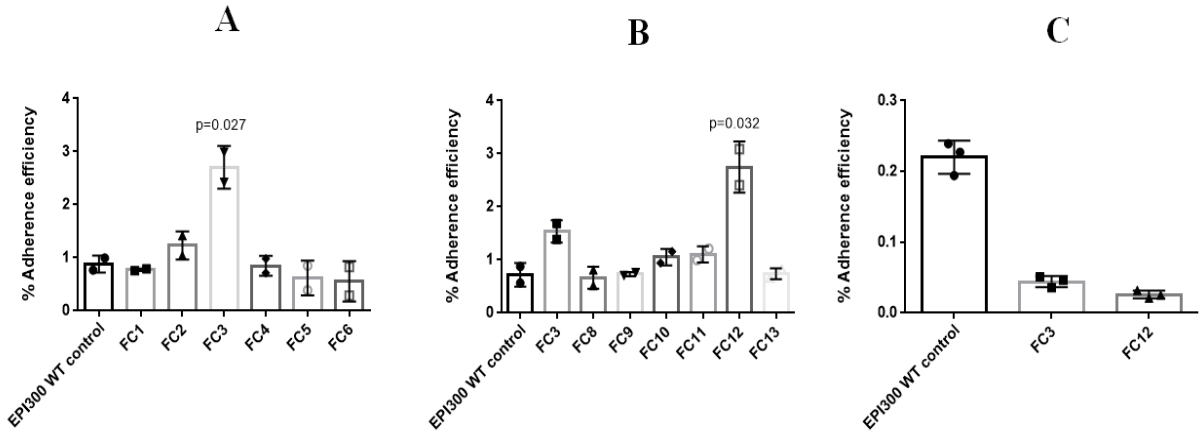


Figure 3.11 ***In vitro*** **selection and enrichment of adhesive clones from the induced fosmid metagenomic library**. The graph represents the percent adherence efficiency of the fosmid library vs. EPI300 control strain on 7 day old caco-2 cells. Five rounds of selection were carried out in an attempt to isolate adherent clones. This experiment was repeated several times (data not shown).

In vitro adhesion assay selection rounds were performed as previously described (section 2.4.11). Much like the results observed in Figure 3.10, after five rounds of selection, the adherence efficiency of the induced fosmid library was lower than that observed after a single round in all three wells (Figure 3.11). Again, this could be attributed to the fact that selection of adhesive clones could have been amplified over two rounds of selection, but that subsequent rounds of selection resulted in outgrowth of truly adhesive clones, by clones which were not actively adhesive but exhibited a more rapid growth rate that the adhesive population.

Figure 3.11 illustrates a gradual increase in adherence efficiency of the induced fosmid library from round 3 in each well to round 5. Unfortunately, this increasing adherence efficiency is also visible in the control clones. The improved recovery of library clones following four rounds of selection may have occurred due to removal of non-adherent clones by the selection process. The improved recovery however, could also be attributed to the removal of slow growing library clones by five successive rounds of growth. Initially, 21 fosmid clones were selected at random from colonies of the 5th round of selection and labelled FC1 to FC22 to be screened with Caco-2 cells for their adherence potential (Figure 3.12). The remaining colonies of the 5th round of selection were stored for further use at 4^{o}C. Of the 21 clones, 16 clones (FC1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,17) shown in Figure 3.12, only 2 clones (FC3 and FC12) exhibited relatively high adherence efficiencies as compared to the control strain (Figure 3.12 A & B). However, in a subsequent experiment these two clones did not have enhanced adherence (Figure 3.12 C).

It was difficult to determine if the induction of individual fosmid clones increased or decreased adherence efficiency. According to Figure 3.12, addition of arabinose seems to increase the adherence efficiency of fosmid clone FC3 in Figure 3.12A and fosmid clones FC3 and FC12 in Figure 3.12B. However, both FC3 and FC12 exhibited a 4-fold reduction in adherence compared to the control strain in Figure 3.12C. The same pattern was observed in Figure 3.12D with a 6-fold decrease in adherence compared to the control strain. Even the adherence efficiency of the control EPI300 strain changed dramatically between experiments.



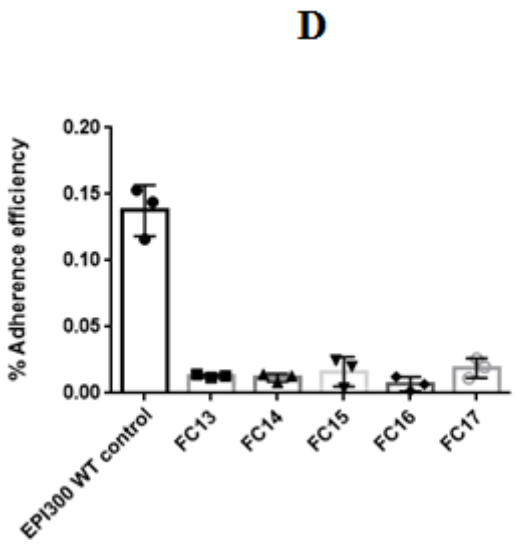


Figure 3.12 Irreproducibility of highly adhesive clones on separate days. Percent adherence efficiency of 16 individual, induced fosmid clones derived from the 5th round of selection of Figure 3.11 against the EPI300 control strain on 7 day old Caco-2 cells. As illustrated in Figure 3.12A, clone FC3 is statistically significantly different from the control strain (A)Adherence efficiency of EPI300 control and FC1-FC6 **(**B**)**Adherence efficiency of EPI300 control strain and FC3, FC8-FC13 (C) Adherence efficiency of EPI300 control strain and FC3 & FC12 (D) Adherence efficiency of EPI300 control strain and FC13-FC17. In experiment B, FC12 shows significant adherence compared to the control.

These experiments illustrate the level of variation present with *in vitro* adhesion assays performed under the same conditions. An obvious limitation in the previous adhesion assays was the lack of an appropriate control strain with an empty pCC1FOS vector. The next section will describe the optimization of control strains (EPI300 WT and EPI300 (pCC1FOS)) for future adhesion assays. Furthermore, the effect on adherence of both strains after induction with L-arabinose will be described.

3.3.4 Optimization of Control strains

According to the results indicated in Figure 3.13, EPI300 WT and EPI300 (pCCIFOS) are not significantly different from one another and can therefore be used as legitimate controls for future assays. These results validated the use of the EPI300 (without empty PCC1FOS vector) as a control strain in previous experiments (Figures 3.7, 3.8, 3.9, 3.10, 3.11 and 3.12). Arabinose induction seems to inhibit the adherence of both strains. In this study, both the EPI300 (pCC1FOS) and the EPI300 strain without vector were seemingly inhibited by arabinose. This suggests that the arabinose

molecules may somehow be binding to adhesin/receptors on the surface of the EPI300 bacteria and therefore blocking the sites for adhesion. Another possibility is that the presence of multiple copies of vector in the EPI300 strain may cause a deleterious effect on the heterologous host and affect adherence.

The next study deals with the screening of five randomly chosen individual fosmid clones from the 5th round of enrichment selection described in section 3.3.3 (Figure 3.11). These clones were investigated for their adherence capacities in the presence and absence of arabinose and on different days of Caco-2 cell differentiation (7 day and 3 week-old Caco-2 cells). Clones which displayed high levels of adherence in pure culture experiments were chosen for further experimental replicates. Clones with consistently increased adherence were selected for sequencing of insert DNA extremeties, followed by bioinformatic analysis to reveal potential adhesins.

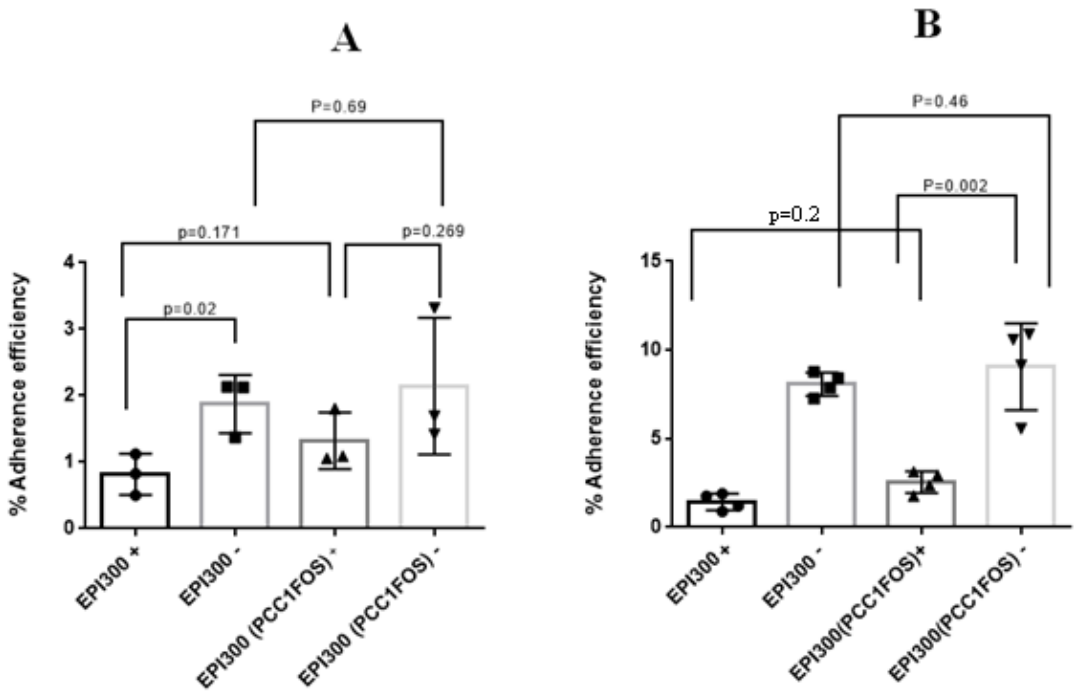


Figure 3.13 EPI300 WT and EPI300 (pCC1FOS) display similar adherence levels. Percent adherence efficiency of the control strain EPI300 WT vs. EPI300:pCC1Fos empty vector on 7 day old caco-2 cells (A) and 3 week old caco-2 cells (B). The error bars represent an individual experiment ran in triplicates (A) and quadruplicates (B). Significance was determined using Student's t-test. (+) = arabinose induction and (-) = no arabinose added.

3.3.5 Analysis of individual fosmid clones on 7 day old Caco-2 cells without induction

The functional screen (*in vitro* adhesion assay) was performed under both single-copy (Arabinose -) and copy-up (Arabinose +) conditions to increase the probability of successfully identifying adhesive clones. Granted, some clones can sufficiently express a protein in single-copy but become toxic to the host under copy-up conditions. Clones that have been induced to high-copy number occasionally lead to the loss of certain clones or the accumulation of insert deletions. Therefore, it was paramount that assays were performed on both single-copy clones and copy-up clones. Having established that EPI300 (pCC1FOS) is as good a control as EPI300 WT (Figure 3.13), EPI300 (PCC1FOS) was used as a baseline control in the next experiments.

Individual fosmid clones FC3, FC18, FC19, FC20 and FC21 were tested for their capacity to adhere to 7 day-old Caco-2 cells without arabinose induction. The experiment was performed three times on three separate days. The control strain maintained a consistent adherence efficiency below 0.3% on each day and across the three days (Figure 3.14). Interestingly, all clones, including the control maintained an adherence efficiency below 0.5% on Day 1 and Day 2. On Day 1, clones FC19 and FC3 were statistically significantly different from the control and exhibited a higher adherence efficiency than the three other clones (FC18, FC20, and FC21). On day 1, FC3 displayed the highest adherence efficiency of 0.455%. On Day 2, the highly adherent clones (FC3 and FC21) of Day 1 lost their significance but the control maintained a <0.3% adherence efficiency. On day 2, FC18 displayed the highest adherence efficiency at 0.40%. Day 3 showed statistically significant differences to the control for all clones (FC18, 19, 20, 21 and FC3). However, FC3 stood out with five-fold higher adherence (1.65%) efficiency than the control. Put together (average of 3 days), none of the clones were significantly different to the control. FC3 had a 3 fold higher adherence efficiency than the control. Overall, in spite of the inter-experimental variation, FC3 and FC19 showed statistical significance to the control on two separate occasions (Day 1 & Day 3). It was worth performing further analysis on these two clones.

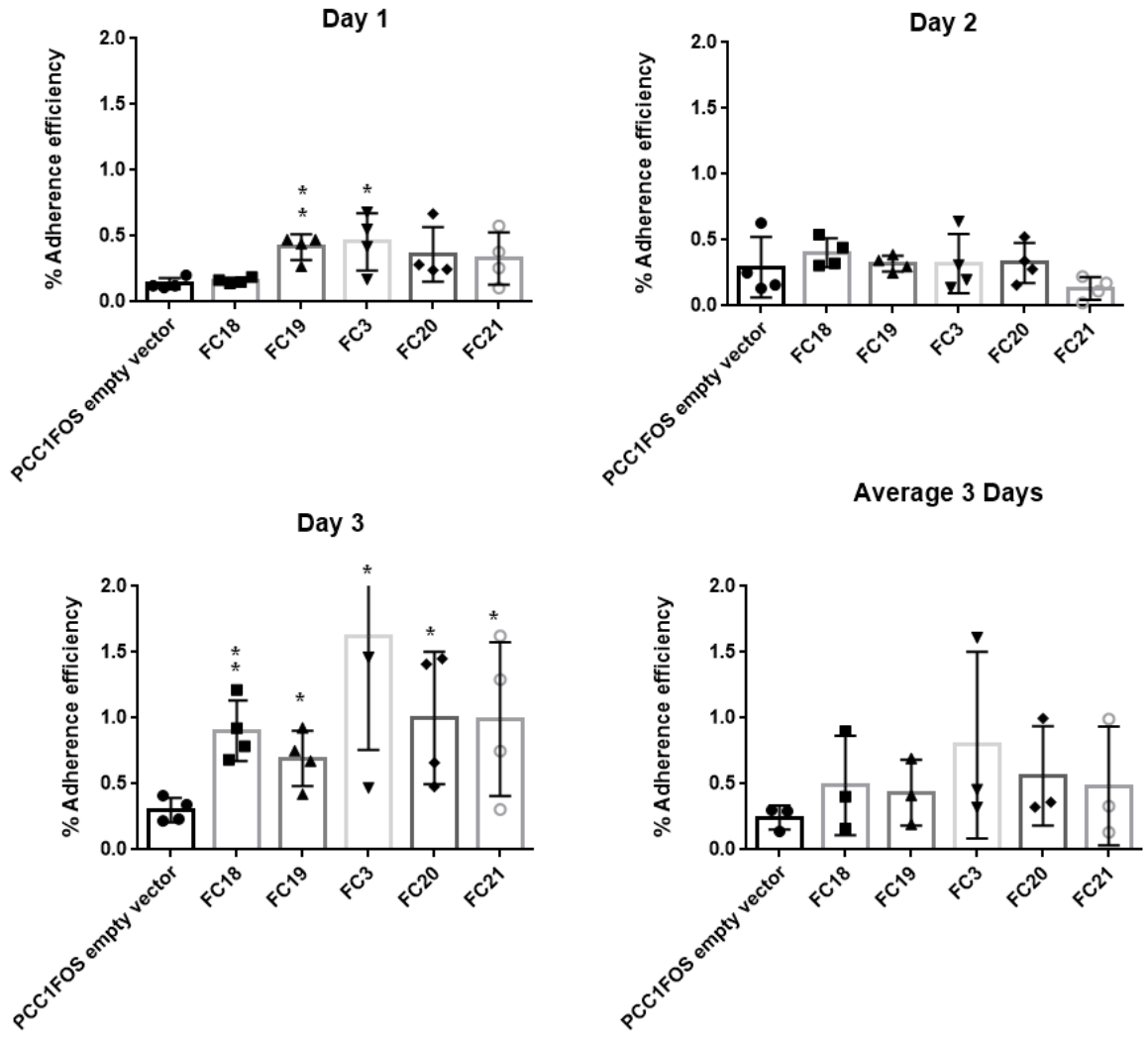


Figure 3.14 Analysis of individual fosmid clones for their ability to adhere to 7 day old Caco-2 cells without arabinose induction: Percent adherence efficiency of the fosmid clones FC18, FC19, FC3, FC20 and FC21 on (-arabinose) 7 day-old Caco-2 cells. The experiment was repeated 3 times on 3 different days. The average of 3 days was also plotted. The error bars represent an individual experiment ran in quadruplicates. Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$.

3.3.6 Analysis of individual fosmid clones on 7 day old Caco-2 cells with induction

The results of three individual adhesion assays performed on three separate days are depicted in Figure 3.15. The adherence efficiency for all the clones varied notably between the 3 days. Unlike the experiments in Figure 3.14, the control strain showed

more variation but always remained below 0.8%. On day 1, most of the clones maintained adherence efficiencies that were very similar to the control strain. Interestingly, only FC21 exhibited high standard deviation illustrating the level of intra-experimental variation in each of the wells. On day 2, all the clones were statistically significantly different from the control. Especially FC22 with a 5 fold higher (1.25%) adherence efficiency than the control (0.25%). On day three, the overall adherence efficiency of the clones increased with most of the clones maintaining statistical significance. FC3 showed a highly significant adherence efficiency of 29.2%, 38 fold higher than the control. When the average of all three experiments was determined none of the clones were significantly different from the control. Overall, the clones showed a high level of intra- and inter-experimental variation when tested on 7 day Caco-2 cells with arabinose.

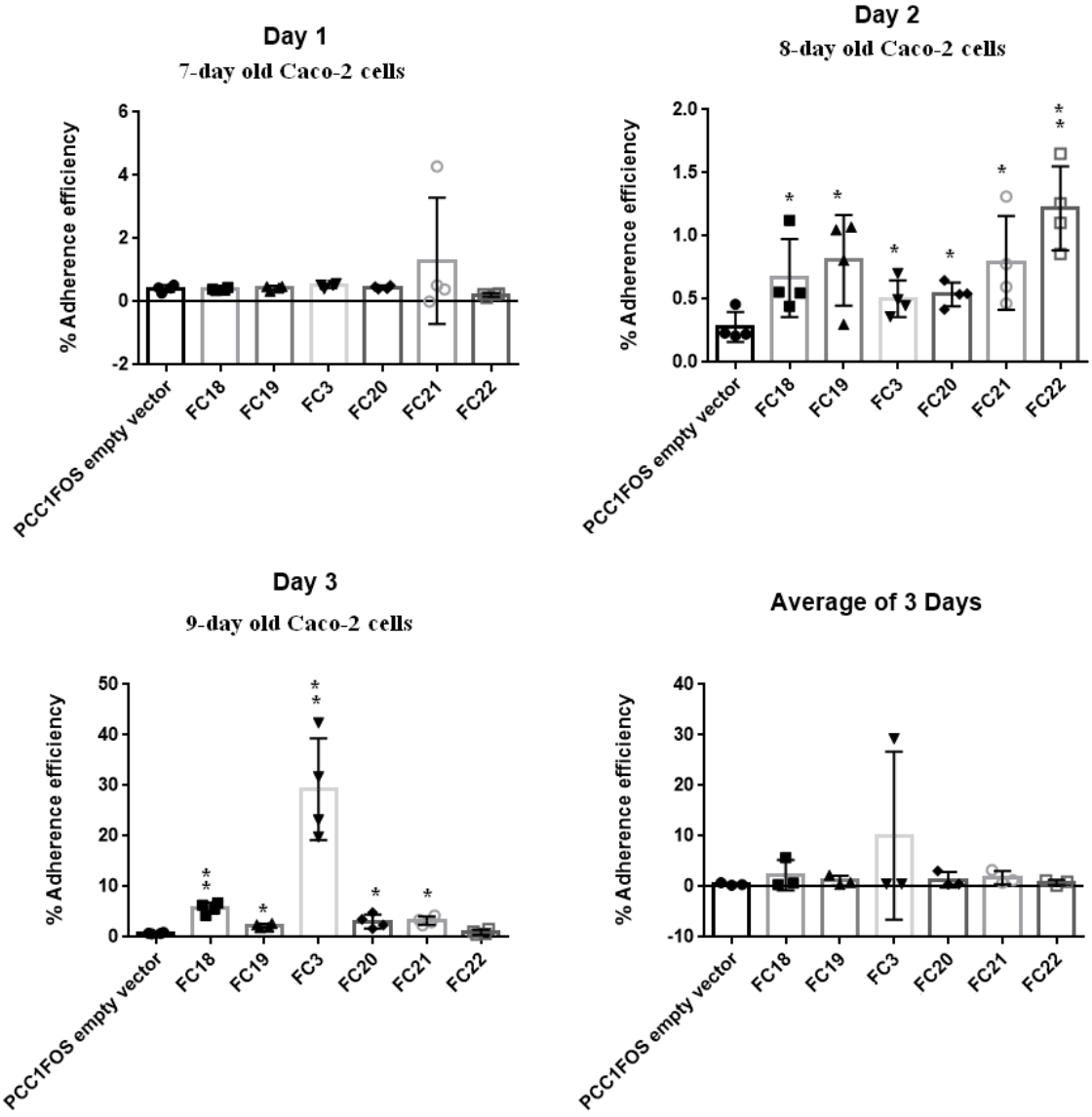


Figure 3.15 Analysis of individual fosmid clones for their ability to adhere to 7 day old Caco-2 cells with arabinose induction: Percent adherence efficiency of induced fosmid clones FC18, FC19, FC3, FC20 and FC21 on (arabinose +) 7 day old Caco-2 cells. The experiment was repeated 3 times on 3 different days. The average of 3 days was also plotted. The error bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$, ***.

3.3.7 Analysis of individual fosmid clones on 3 week-old Caco-2 cells with induction.

On Day 1 of Figure 3.16, all five clones were significantly different from the control. FC3 was the most adherent clone with a 12 fold (10.8%) higher adherence efficiency as compared to the control (0.9%). FC3 was followed closely by FC18. On day 2, only three clones were significantly different from the control, FC19, FC3 and FC21. The adherence efficiency of the control remained relatively consistent across the three days. The same consistency was observed in FC20 on all three days. On day 2, FC21 showed the highest adherence efficiency of 20% followed closely by FC19 at 18.0%. On day 3, FC18, FC3 and FC21 were statistically significantly different from the control. Once again, FC21 had the highest adherence efficiency at 13.0% followed closely by FC18 at 8.0%. Put together, two clones maintained significance; namely FC18 and FC3. FC21 was more adherent but showed a lot of standard deviation within the wells. Overall, fosmid clones FC18, FC21 and FC3 demonstrated enhanced capacity to bind to 3 week-old Caco-2 cells in the presence of arabinose.

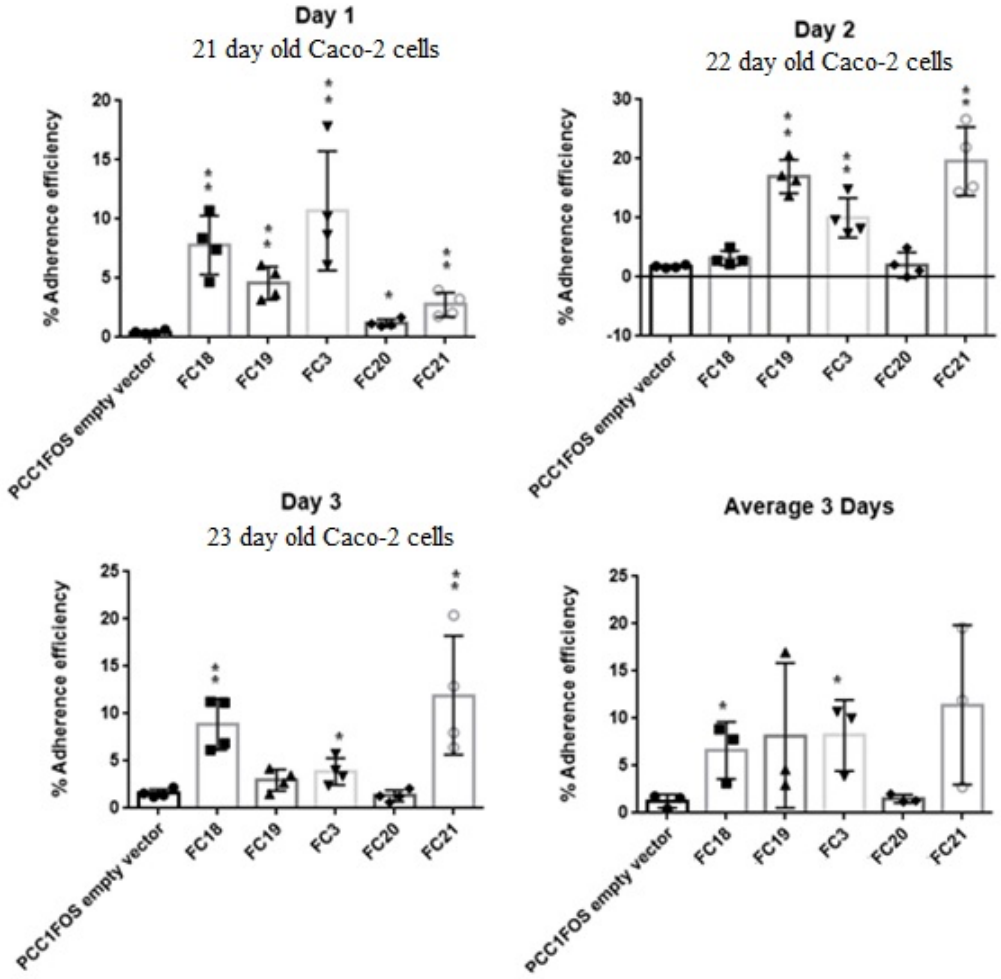


Figure 3.16 Analysis of individual fosmid clones for their ability to adhere to 3 week old Caco-2 cells with arabinose induction: Percent adherence efficiency of induced fosmid clones FC18, FC19, FC3, FC20 and FC21 on (arabinose +) 3 week old Caco-2 cells. The experiment was repeated 3 times on 3 different days. The average of 3 days was also plotted. The asterix depicts $p<0.05$. Two asterix means a very small p value. The error bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$.

3.3.8 Analysis of individual fosmid clones on 3 week old Caco-2 cells without induction

When the individual fosmid clones were tested on 3 week-old Caco-2 cells in the absence of arabinose, the results demonstrated high levels of inter-experimental variation (Figure 3.17). This was observed by comparing the values obtained for

EPI300 (pCC1FOS) in each experiment. On day 1 an adherence efficiency of 1.0% was observed for EPI300 (pCC1FOS), while on day 2, a value of 0.6% (2 fold decrease) was observed. Overall, EPI300 control strain remained at an adherence efficiency below 1.3% in all experiments.

On day 1, all the clones except FC3 displayed statistical significance in their adherence efficiencies compared to the control. FC18 had the highest adherence efficiency at 3.2%, 3 fold higher than the control. On day 2, the overall adherence efficiency of the control had reduced from 1% (day 1) to 0.6% (day 2). On day 2, none of the clones were significantly different from the control. However, on day 3, FC18 stood out as the only clone that was statistically significantly different from the control. The other clones displayed adherence efficiencies that were lower than the baseline control. Put together, none of the clones showed statistical significance to the control. However, FC18 showed slightly higher adherence efficiency than all the other clones at 1.8%. Overall, a correlation seemed to exist between the age of Caco-2 cells and adherence efficiency. The individual fosmid clones exhibited higher adherence efficiencies when screened on 3 week-old Caco-2 cells than 7 day old Caco-2 cells. This was demonstrated in the relatively high adherence efficiencies observed in Figure 3.16 as compared to Figure 3.17. According to Figure 3.14, the highest adherence efficiency of un-induced 7 day old Caco-2 cells was 1.65% on day 3. The highest adherence efficiency of arabinose induced 7 days old Caco-2 cells was 29.28% on day 3 (Figure 3.15). The same pattern can be seen with 3 week old Caco-2 cells. The highest adherence efficiency of un-induced 3 week Caco-2 cells (Figure 3.17) was 1.5% on day 3. The highest adherence efficiency of arabinose induced 3 week Caco-2 cells (3.16) was 20% on day 2. These results are in contrast to the inhibitory effect observed in arabinose induced heterologous hosts (Figure 3.9, page 81).

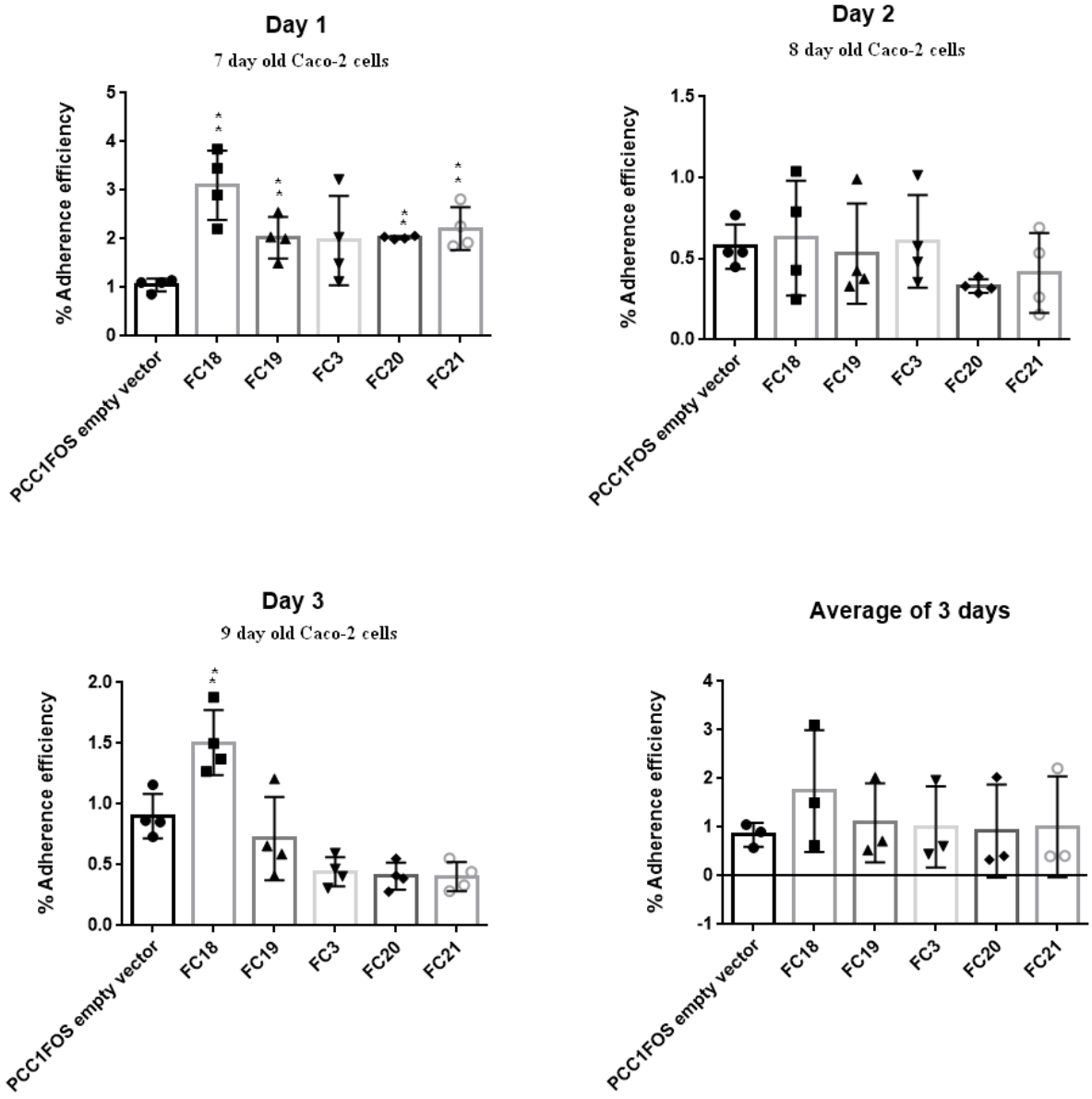


Figure 3.17 **Analysis of individual fosmid clones for their ability to adhere to 3 week old Caco-2 cells without arabinose induction:** Percent adherence efficiency of fosmid clones FC1, FC2, FC3, FC9 and FC10 on (arabinose -) 3 week old caco-2 cells. The experiment was repeated 3 times on 3 different days. The average of 3 days was also plotted. The error bars represent an individual experiment ran in triplicates. Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$.

Analysis of 16 individual fosmid clones for their ability to adhere to 7 day-old and 3 week old Caco-2 cells revealed that the adherence efficiency of individual clones was higher in 3 week old Caco-2 cells as compared to 7 day old Caco-2 cells in the presence

of arabinose. Moreover, the adherence efficiency of individual fosmid clones was higher when induced with arabinose than without arabinose. Analysis of the 16 individual fosmid clones for their ability to bind 7-day old and 3-week old Caco-2 cells led to the selection of 6 putative adhesive clones (FC3, FC18, FC19, FC20, FC21 and FC22) that demonstrated statistically significant increase in adherence to Caco-2 cells. Chapter four will describe the analysis and characterization of these clones in more detail.

3.4 Discussion

In this study, functional screening was based on an *in vitro* assay of adhesion onto Caco-2 cells of bacteria carrying a human gut metagenomic library. This adherence method has proven to be successful in studies investigating the adherence level of known and novel probiotics such as *lactobacillus* and *bifidobacterium* strains[204]. The vast majority of metagenomic libraries have been screened for activities such as enzyme activity, antibiotic activity and biofilm activity[156]. However, no metagenomic libraries have been screened for adherence using an *in vitro* adhesion assay[190].

One of the major limitations to functional screening is the appropriate expression of cloned DNA in the heterologous host *E. coli*. There are two main problems that can arise from cloning a foreign gene into *E. coli*; limitations due to the cloned gene sequence or to the *E. coli* host. The codon bias of the cloned gene may not be ideal for translation in *E. coli*. Although virtually all organisms use the same genetic code, each organism has a bias toward preferred codons [170]. This bias reflects the efficiency with which tRNA molecules in the organism are able to recognize the different codons. If the cloned gene contains a high proportion of disfavoured codons, *E. coli* tRNA may encounter difficulties in translating the gene, reducing the amount of protein that is synthesized[315, 313]. Additionaly, the cloned gene might contain sequences that act as termination signals in *E. coli*[394]. These sequences are perfectly normal in host cell, but in the bacterium they result in premature termination and a loss of gene expression[394,401].

A few other difficulties that may be encountered with *E. coli* as the host can come from inherent properties of the bacterium. The *E. coli* bacterium may not be able to fold the recombinant protein correctly and may be unable to synthesize disulphide bonds present in many proteins. *E. coli* might degrade the recombinant protein.

Exactly how it can recognize the foreign protein and thereby subject it to preferential turnover is not known. However there are other factors that can limit gene expression including transcription and translation initiation signals, protein-folding, post-translational modification, protein secretion or toxicity of the recombinant gene [151]. Adhesins may be expressed by the heterologous host but without the correct secretion machinery and subsequent cell surface localisation, the functionality of such proteins may be lost.

The main way to overcome many of these limitations is by utilizing alternative heterologous hosts. Previous studies have shown that alternative heterologous hosts can increase the diversity and efficiency of functionally screened metagenomic libraries[205]. Already, several studies have employed alternative hosts (*Bacillus subtilis*, *Lactococcus lactis*) for genomic and metagenomic library construction in order to increase the number and diversity of positive clones[206]. Carbohydrate active enzymes derived from different micro-organisms have been cloned in *Lacotococcus lactis*. The expression of rumen-derived-β-glucanase was examined in *L. lactis* by Ekinci *et al*. (1997) and showed detectable enzyme activity. Other carbohydrate active enzymes including xylanases and cellulases have also been studied using lactic acid bacteria as a heterologous host [207].

Craig and colleagues[208] describe the functional screening of a DNA soil library in various species of bacteria. *E. coli* produced two antibacterial active clones and zero pigmented clones, whereas *R. metallidurans* produced four positive clones for antimicrobial activity and more than 18 clones which tested positive for approximately three different types of pigmentation. Since *E. coli* is a member of the *gammaproteobacteria* and *R. metallidurans* is a member of the *betaproteobacteria*, it illustrates the importance of selecting a range of diverse hosts for expression of heterologous DNA from diverse ecosystems like the human gut.

Estimates show that only 40% of heterologous genes are readily expressed in *Escherichia coli*. 73% of bacterial genes originating from Firmicutes – the dominant phylum of human gut microbiota – were predicted to be expressed in *E. coli* [209] . The remaining 27% of genes could not be detected due to gene expression failure; as a result *Lactococcus lactis* MG1363 was used as an alternative host in this study. Researchers have found a positive correlation between codon usage of individual gene

and surrogate host [170]. It was hypothesised that genes from gut Firmicutes will be readily expressed in *L. lactis* since the %G+C content is similar [210].

This study attempted to establish the suitability of L. *lactis* as a heterologous host for functional screens. Several attempts were made to produce the small fragment library in *L. lactis* MG1363 but only a limited number of transformants were obtained restricting the preparation of a metagenomic library in our gram positive host. As mentioned, the major limitation in preparing the metagenomic library directly in *L. lactis* MG1363 was the transformation efficiency which was substantially different from *E. coli*. It has been shown that the transformation efficiency in *E. coli* can reach up to 1000-fold higher than in *L. lactis* [211]. The poor cloning efficiencies in *L. lactis* were also observed in a study conducted by Geertsma and Poolman (2007)[212]. The preparation of highly efficient electrocompetent *L. lactis* cells was shown to produce up to 10^8 CFU/µg DNA [213]. However, direct transformation into *L. lactis* produced as little as 10^2 CFU/µg DNA [214]. Taken together, although both bacteria (*L. lactis* and *E. coli*) were reported as successful heterologous hosts in a number of studies [215] in the present study *E. coli* was used as the only host for the functional screening of our metagenomic libraries.

In this study, the results of the *in vitro* adhesion assays were highly variable. From day to day, the percentage adhesion varied up to 4 fold. Therefore we used three to four replicates per experiment and planned several experiments to perform statistical analysis using student paired test. Although a large portion of the variation observed in the adhesion assays was intrinsic, there were many factors that were likely to influence the adhesion efficiency of our fosmid clones. Some of these factors were related to the mode of culture adopted *in vitro*; Multiplicity of Infection (MOI), buffer composition, incubation time, temperature, growth medium and growth stage. In most assays, bacteria are co-incubated for 1-3 hours with the epithelial cells[176]. The effect of the incubation time on the adhesion of bacteria strains has not been fully investigated. Scientists have found that the time of incubation is crucial in reducing assay variation [190]. It was imperative not to exceed 3 hours of co-incubation of bacteria with Caco-2 cells otherwise the secretion of toxins would cause cell death and bias the results [190]. In our study, a co-incubation time of 1 hour and 30 minutes was maintained for all assays.

Particularly important sources of variation were the MOI (multiplicity of infection) and the number of washes. The objective of the study was to maintain an MOI of 10 bacterial cells to 1 Caco-2 cell for *E. coli* cells. Higher MOI yielded higher variability and background and bacteria tended to stick to the plastic of the plate [190]. Lower MOI also yielded high variability. Once the appropriate MOI was chosen, it was imperative to consistently maintain and respect that MOI.

Another key source of variation was the washes after co-incubation of cell-associated bacteria. Care was taken so that the Caco-2 epithelial cells were not detached from the substratum. If the cells sloughed off then there would be less adhered bacteria in that well. The number of adhered bacteria is correlated with the number of washes and depends heavily on how the experimenter performs the washes. It was therefore important to employ the same number of washes for all wells and to repeat precisely the same procedure every time a wash was performed.

Bacterial growth phase will affect adhesion, primarily because different proteins will be expressed at different phases of growth. The growth stage of the bacteria has been observed to influence the adhesion significantly. It appears that log phase cells are more adhesive than early stationary phase[96]. The composition of the growth medium has been observed to influence the adhesion of several probiotic strains dramatically [190]. At the log phase, *E. coli* cells grow rapidly and are at the prime state to produce proteins. The *E. coli* cells are often harvested at the middle to late log phases for protein production. Harvesting cells later than the log phase may observe low protein yield and high protein degradation. Although the cell density of *E. coli* is highest at stationary phase, the cells are usually stressed. At the stationary phase, the medium's nutrients become limited and metabolic products accumulate to such a high level that they are inhibitory to the cell growth. Therefore, in this study libraries and individual clones were grown to log phase before co-incubation with Caco-2 cells.

Other ambient conditions such as temperature may also affect the percent adherence efficiency. Assays incubated at 37^{o}C have shown adherence yields greater than those obtained at 25^{o}C[216]. This is not surprising since 37^{o}C is close to the actual physiological temperature in the human gastrointestinal tract. Incubation conditions affect the outcome of an adhesion assay and should be optimized for physiological relevance. All assays in this study were incubated at 37^{o}C.

Studies indicated that the pH was able to impact adherence ability in *Candidatus albicans*[217]. While the pH of the luminal small gut is about 7, individual viscosity of the brush border environment is acidic due to the presence of mucus overlay and metabolic activity of intestinal cells and adhered bacteria; therefore it is worth performing adhesion tests at different pH values.

Another source of variation is the age and passage number of the Caco-2 cells. A potential disadvantage of using Caco-2 cells is that they are cancer cells (adenocarcinoma cells), which may or may not be different from the normal intestinal epithelial cells. Caco-2 cells are derived from colonic adenocarcinoma and express protein characteristics for both colonocytes and enterocytes immediately after confluence. Thereafter, the content of colonocyte specific proteins decreases. The expression level of P-glycoproteins and other transport proteins is known to vary significantly with the age of the Caco-2 cells [218]. This makes timing in the use of Caco-2 cells particularly important[219]. As observed in our results, fosmid clones behave differently on 7 day old Caco-2 cells than they do on 3 week old Caco-2 cells (Figure 3.15 & Figure 3.16). In general, we observed lower inter-experimental variation of clones tested on 3 week old Caco-2 cells than 7 day old Caco-2 cells. Moreover, fosmid clones grown in arabinose showed higher adherence efficiencies than those grown on 7 day old Caco-2 cells.

In our study, we conducted a functional screen of two metagenomic libraries of the human gut microbiota for potential adhesive clones and identified six potential clones (FC3, FC18, 19, 20, 21 and FC22) out of 42,000 clones likely to be involved in adhesion within the human gut. Studies have demonstrated that the hit rate of functional metagenomics screen is relatively low. This was not surprising since the frequency of metagenomic clones to express any given activity is often low. For example, in a search for lipolytic clones derived from German soil, only 1 in 730,000 clones showed activity [158]. In a library of DNA from North American soil, 29 of a total of 25,000 clones expressed haemolytic activity[220]. Brady and colleagues have described “hit rates” as low as 1 in 100,000 clones for the phenotypes of antibiotic activity and pigmentation. Low hit rates in metagenomic libraries commonly occur as a result of incompatibility between the cloned DNA and the heterologous host. The complexity inferred on metagenomic libraries as a result of their diversity also requires that a large number of clones must be screened to detect a functionally active clone.

This has the effect of limiting functional analysis to simple traits such as pigmentation and antimicrobial activity. This scarcity of active clones illustrates the necessity to develop efficient screens and selections for discovery of new activities and proteins.

Overall, this study demonstrated an effective means of rapidly profiling metagenomics library clones by restriction digestion and end-sequencing. To validate both libraries, two approaches were taken: end-sequencing of randomly selected clones and restriction digestion of randomly selected clones. Based on the digests and BLASTn analysis of end-sequenced clones, we determined that both libraries were genetically diverse. Interestingly, BLASTn analysis of a small fragment clone (plasmid clone 2 forward and reverse, Table 3.1) exhibited a forward end-sequence of 99% identity to the bacterium *Eubacterium rectale* and a reverse end-sequence of 95% identity to *Roseburia intestinalis*. A possible explanation for this observation is that one of the sequences is conserved in both these organisms. This is not surprising since phylogenetic studies have indicated that one of the most closely related species to *Eubacterium rectale* is *Roseburia intestinalis*[221]. *Eubacterium rectale* is reported to be one of the most abundant bacterial species in human faeces both from anaerobic cultivation and culture-independent analysis of 16S rRNA sequences[222]. It is one of the major species in the human colon that is responsible for butyrate formation. This is an important trait, as butyrate, which is one of the three major short chain fatty acids (SCFA) formed in the colon, is the preferred energy source for colonocytes and has a protective effect against colon disease. It would be interesting to isolate glycan-binding genetic determinants from this health promoting organism[222]. Another small fragment library clone in Table 3.1 exhibited a 95% similarity to the gut bacterium *Bacteroides fragilis*. Studies have shown that *B. fragilis* is capable of mediating powerful effects on the host immune system[223]. Troy and colleagues[146] recently characterized a model symbiosis factor, Polysaccharide A (PSA) produced by *B. fragilis*. PSA is capable of activating T cell-dependent immune responses that can affect both the development and homeostasis of the host immune system.

As indicated in Table 3.2, 4 out of the 10 fosmid clones analysed showed sequence honology to the dominant gut bacterium *Bifidobacterium adolescentis*. *B. adolescentis* is a normal inhabitant of the healthy human and animal intestines [53]. It is a Gram-positive, non-motile, often branched anaerobic bacterium. *Bifidobacterium* has been associated with a healthy microbiota and they are included in many food preparations

with associated health claims [224]. The forward end-sequence of fosmid clone 4 (99% identical) and fosmid clone 8 (97% identical) showed homology to a sequence from the species *Candidatus snodgrassella* (Table 3.2). Very little is known about *Candidatus snodgrassella*, except that it is a novel genera and species found in the gut microbiota of honey bees and bumblebees[225]. It was discovered in 2013 by Kwong and Moran, and named after Robert Evans Snodgrass, a pioneer in the study of insect physiology in the early 20th century[225].

This study demonstrated an effective means of characterizing microbial diversity of metagenomics libraries. The end-sequenced analysis of clones from the small fragment library (Table 3.1) demonstrated that the inserts belong to the two dominant phyla Firmicutes and Bacteroidetes. Similarly, end-sequened BLASTn results of clones from the fosmid library were ascribed to the phylum Proteobacteria, Bacteroidetes and Actinobacteria. Studies have shown that the four dominant bacterial phyla in the human gut are Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria.

An alternative strategy for library validation and diversity determination would have been the amplification and sequencing of 16S rRNA genes using the libraries' DNA as template[226]. Using 16S rRNA gene sequence analysis, Alinne Pereira de Castro and colleagues were able to identify the bacterial phyla in both their soil metagenomics libraries with a confidence of 95%[226].

Analysis of the adherence efficiency of the fosmid library and the EPI300 control strain without induction indicated that there was no statistically significant increase in adherence of the fosmid library (Figure 3.8). The experiment was performed in triplicates on five separate days. However, the addition of arabinose gave rise to an increase in adherence efficiency of both the control and the fosmid library (Figure 3.9). It was hypothesized that if there are adhesive clones present in the fosmid library, induction of such clones should increase expression of the adhesive phenotype of the fosmid library significantly. In spite of the increased adherence efficiency observed for both the fosmid library and control, the adherence efficiency of the fosmid library is not statistically different to the EPI300 control (Figure 3.9). In fact, the EPI300 control exhibits a 2-fold increase in adherence efficiency on both Day 1 and Day 2 as compared to the fosmid library. A possible explanation for the contrast in adherence

efficiency between the fosmid library and the wild-type EPI300 control strain is that the retention of multiple copies of large foreign DNA molecules coupled with high levels of expression increases the likelihood of toxicity in the heterologous hosts. This explains why the EPI300 wild-type control strain is not demonstrating inhibition of adherence because it lacks the pCC1FOS vector and is unable to be induced. A comparison of the results present in Figure 3.8 and 3.9 demonstrate that both the fosmid library and control strain exhibit increased adherence efficiency in the presence of arabinose. However, this increase in adherence of the fosmid library and control is not statistically significant when comparing the fosmid library to the control strain.

This study demonstrated that the enrichment of clones from both metagenomics libraries for amplification selection towards adhesive clones was not successful, likely due to the introduction of bias towards rapidly growing clones. Two possibilities exist; (i) the selection of adhesive clones was amplified over two rounds of selection, but then subsequent rounds of selection resulted in outgrowth of truly adhesive clones by clones which were not actively adhesive but exhibited a more rapid growth rate than the adhesive population; (ii) while putative adhesins may have been expressed in the initial selection process, the growth of retrieved bacteria may have resulted in the bacteria adapting so as to reduce expression of a potentially deleterious protein. This would result in selection of clones which may indeed carry adhesins but which have become incapable of expressing such clones. Having taken steps to ensure that background adherence would not interfere with selection of adherent clones, it appears that other issues may have caused poor recovery rates, such as incompatibility between metagenomics-derived fosmid library and the heterologous *E. coli* host.

As recounted in this chapter, adhesion assays were performed on individual fosmid clones selected from the 5th round of the multiple rounds of enrichment studies (section 3.3.3). Six fosmid clones (FC3, FC18, FC19, FC20, FC21, FC22) demonstrated significant adherence to 3-week old Caco-2 cells and 7 day old Caco-2 cells. These clones were selected for further characterization and analysis. For example, FC3, FC18,19,20 and 21 exhibited significant adherence to 7 day old Caco-2 cells in the absence of arabinose (Figure 3.14, Day 3) FC18 and FC19 indicated at least 2-fold increase in adherence as compared to the control. Both FC20 and FC21 exhibited a 3-fold increase in adherence as compared to the control strain. FC3 demonstrated the largest increase in adherence with a 5-fold increase as compared to the control. These

six clones also demonstrated significant adherence to 7 day old Caco-2 cells with arabinose induction. As illustrated in Figure 3.15, Day 2. Similar significant adherence are observed in their ability to bind 3 week-old Caco-2 cells with arabinose induction (Figure 3.16, Day 1). Finally, FC18, FC19, FC20 and FC21 demonstrated significant adherence to 3 week-old Caco-2 cells without arabinose induction. Based on the statistical significant adherence observed for these clones, further characterization and analysis are required.

In conclusion, the findings of this work serve to demonstrate an effective means of rapidly profiling metagenomics library clones by end-sequencing and restriction digestion to validate and characterize microbial diversity in both libraries. Using various strategies of enrichment and individual clone analysis, the work described in this chapter has led to the identification of 6 potential adhesive clones (FC18, FC19, FC3, FC20, FC21 and FC22) to be further characterised. These fosmid clones offer significant opportunity to develop understanding of the molecular mechanisms governing adherent processes. There were numerous technical difficulties to the approach caused by intrinsic variation in the adhesion assays. The following chapter will deal with the characterization and bioinformatics analysis of the putative adhesive clones identified by functional metagenomics.

Chapter 4:

Characterization and bioinformatic analysis of adhesive clones identified by functional metagenomics.

4.1 Introduction

In chapter 3, functional metagenomics screens were used to identify putative adherent clones from a fosmid metagenomic library consisting of 42,000 clones. The putative adherent clones exhibited enhanced capacity to adhere to differentiated Caco-2 cells *in vitro*. This chapter will describe the subsequent sequencing and *in silico* analysis of the DNA inserts from the isolated clones to provide information about the source of the genes and the putative mechanism of action of their products. Furthermore, this study will discuss carbohydrate-based microarrays and their use in profiling bacteria-carbohydrate interactions. These studies are presented in three sections covering three main types of carbohydrate-based microarray platforms; (i) carbohydrate-binding protein (lectin) microarray to determine glycosylation patterns on the cell surface of the bacteria (ii) natural mucin microarrays to determine mucin glycosylation and bacterial binding tropisms, and (iii) finally neoglycoconjugate (NGC) microarrays to determine the binding affinity of bacteria to specific NGC.

4.1.1 Next Generation Sequencing

Over the past decade, a new generation of sequencing technologies has provided the scientific community with unparalleled opportunities to analyse complex microbial communities[8,227]. Next Generation Sequencing (NGS) is a large-scale DNA sequencing technology where millions or billions of DNA strands are sequenced in parallel at high speed and low cost, producing significantly more data throughput and diminishing the need for conventional sequencing methods [228].

In 2004, NGS platforms became commercially available and started to radically transform biomedical inquiry. The rapid progress of NGS technology facilitated gut microbiome research, and precipitated the exploration of the genetic and functional diversity of uncultured gut microbial communities with affordable costs and high throughput[229]. NGS platforms are currently evolving into molecular microscopes making their way into many fields of biomedical research. Thus far, NGS has been instrumental in determining the role of the human microbiome in disorders like Inflammatory Bowel Syndrome, diabetes and obesity[230]. Scientists can now obtain a complete genomic catalogue of disease genes and in-depth insight into the differences amongst thousands of people to uncover pivotal genes that cause cancer,

schizophrenia, heart disease and autism[231]. Overall, NGS technology has the capacity to bring expansive change in biomedical and genetic research thus expanding our knowledge tremendously.

The advent of NGS platforms is opportune because there remain many complex genomic research questions that require in depth information beyond the capacity of traditional DNA sequencing technologies (i.e., Capillary Electrophoresis (CE), Sanger chain termination etc.). DNA sequencing has progressed tremendously since the use of two dimensional chromatography in 1970s and the Sanger chain termination method in 1977[232]. Using conventional Sanger sequencing method, the first human genome (3.2 million bases) took 15 years to sequence (published in *Science* and *Nature*, 2001) and cost approximately 3 billion dollars[233]. More than 20,000 large bacterial artificial chromosomes (BAC) clones each containing approximately 100kb fragments of the human genome were used[234]. Sanger sequencing is still used extensively today for routine sequencing applications and to validate NGS data [235]. Although these first generation platforms were considered high-throughput for their time, the introduction of NGS platforms surpassed and revolutionized sequencing technologies.

4.1.2 NGS Chemistry

The underlying concept behind Illuina Hiseq next generation sequencing is the catalytic incorporation of fluorescently labelled dNTPs (deoxy ribonucleotide triphosphates) by DNA polymerase into a DNA template strand during sequential cycles of DNA synthesis [227]. This process is very similar to the conventional capillary electrophoresis sequencing, however, instead of sequencing a single DNA fragment, NGS expands this process across thousands of millions of fragments in a parallel fashion [236]. Illumina sequencing by synthesis (SBS) chemistry is by far the most common next-generation sequencing (NGS) technology utilized in industry and worldwide. In this study, NGS (Illumina Hiseq) is utilized as an effective tool to elucidate the sequences of metagenomic fosmid clones. One of the advantages of the Illumina sequencing by synthesis (SBS) chemistry is that it provides high accuracy and high yield of error free reads [237].

4.1.3 Pair-End Sequencing

Illumina sequencing by synthesis technology supports both single-read and paired-end libraries. Over the years, NGS has advanced significantly with the introduction of pair-end sequencing. Pair end-sequencing is the sequencing of both ends of the DNA fragments in a sequencing library and aligning the forward and reverse reads as read pairs [238]. The major advantage of this read pair alignment is the ability to produce twice the number of reads for the same time and effort and higher accuracy of read alignments. A lot of these benefits are not possible with single reads [239]. The advantage of pair-end sequencing is that longer reads make it easier to assemble the reads. Currently, the vast majority of researchers use pair-end sequencing.

4.1.4 Benefits of Next Generation Sequencing, NGS

NGS platforms provide far more benefits than their conventional counterparts (CE and Sanger). Unlike Sanger sequencing, NGS is directly able to detect base variations (substitutions, re-arrangements (inversions & translocations), insertions, deletions) within a genome in a single experiment [233]. Previously, in order to detect mutations within a genome, dedicated assays such as fluorescence in situ hybridisation (FISH) for conventional karyotyping or comparative genomic hybridisation (CGH) microarrays to detect sub-microscopic mutations were used [235].

Another advantage of NGS is that it can be unselective not requiring pre-knowledge of the gene or locus under investigation. However, is also possible to amplify specific genes prior to sequencing, such as amplicon sequencing work. Sanger sequencing can be both selective and unselective if a metagenomic clone library is sequenced. NGS can be used to interrogate full genomes in order to discover new mutations or disease causing genes [237]. Moreover, NGS has enabled scientists to explore new areas of biological enquiry such as the sequencing of ancient genome samples, the characterization of ecological diversity and the widening scope of metagenomics analysis of environmentally derived samples [229]. The next section will describe the different kinds of carbohydrate-based microarrays in the characterisation of bacterial-glycan interactions.

4.2 Carbohydrate-based microarrays

Lectins are carbohydrate-binding proteins that differentiate sugars based on subtle differences in structure. Lectins are able to bind both monosaccharides and disaccharides with high specificity and in a reversible mode. They have been used for many years as an essential tool in the detection of glycans[240]. More recently, they have been used to profile glycosylation in disease states such as cancer [241]. This insight that glycans are important for most biological processes has led to the development of a rapid, comprehensive and high-throughput strategy for structural analysis of glycans on the surface of intact cells. Conventional procedures for the analysis of glycans include high performance liquid chromatography (HPLC), mass spectrometry (MS) and capillary electrophoresis[242]. These techniques require the prior liberation of the glycan from its core protein. Subsequently, these liberated glycans must be fluorescently labelled for detection and for improved separation during HPLC. This extraction and purification step can modify the structures of these glycans[243]. Chemical methods such as mass spectrometry often examine only a single carbohydrate motif, which is slow and limited by time, equipment and expertise. Moreover, the number of glycans that can be simultaneously analysed using these conventional techniques is restricted[243]. This is problematic since a global examination of glycans is an important pre-requisite to understanding the structure-function relationship of glycans. In short, traditional methods of studying glycans are often complicated, laborious, and time consuming [172].

As a result, a novel platform termed "lectin microarray" for the analysis of glycoproteins was developed in 2005. Unlike conventional methods, crude samples (e.g. intact cells) containing glycoproteins can be directly analysed without the need to liberate the glycans[244]. The lectins are displayed in a microarray format on the microarray that enables distinct, multiple binding interactions to be observed simultaneously. The advent of this novel technique has since attracted increasing interests from glycoscientists, microbiologists and scientist from other fields[245]. In this study, lectin microarrays were used to determine the glycosylation patterns on the surface of fosmid clones.

Bacteria are coated with a dense, variable and highly diverse layer of sugars on their surfaces. Their surfaces vary enormously depending on the phylogenetic group. This

diversity and variability can be problematic and challenging to analyse [246]. Indeed, the inherent qualities of carbohydrates promote structural diversity. They are structurally more complex than nucleic acids or proteins, varying in their linkage position, branching, residue ring size and non-sugar substituents such as phosphorylation[57]. This combinatorial capacity of carbohydrates is staggering and leads to the possibility of an enormous number of structural isomers [247]. It has been calculated that a hexamer can lead to more than 10^{12} structures [248]

Traditional techniques that use lectins to analyse bacterial glycosylation usually generate results that are difficult to interpret (subjective), not suitable for high-throughput and the assay often lacks sensitivity [184]. Keen interest in using lectins for high-throughput analysis led to the development of the Lectin Microarray. The next section will describe the lectin microarray technology in more detail.

4.2.1 Lectin Microarray Technology

Lectin microarrays are composed of a panel of natural lectins immobilized on a glass slide (Figure 4.1, Table 2.5). The microarrays containing the lectins can then be interrogated with glycosylated, fluorescently labelled samples (e.g. fluorescently labelled bacteria) creating a visual lectin binding pattern that imparts information about the glycans on the sample.

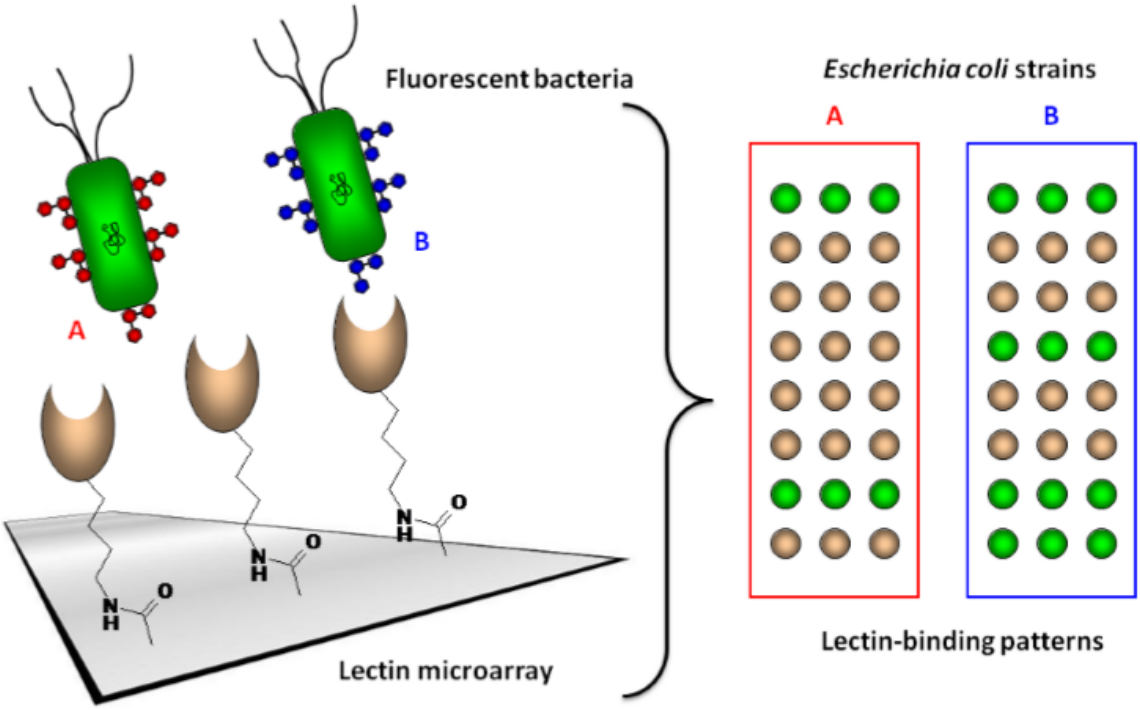


Figure 4.1 **The experimental setup to analyze bacterial glycosylation using the lectin microarray**. Fluorescently-labelled bacteria are bound to the array generating a visual binding pattern. The lectin-binding patterns can be used to compare glycosylation between strains (A & B) and provide some structural information about bacterial glycans. Diagram adapted from Ku-Lung Hsu, 2008[59].

In this study, a panel of 48 commercially-available plant lectins (Table 2.5) were immobilized onto activated glass slides (Nexterion Slide H microarray slides) using a SciFLEXARRAYER S3 (Scienion AG, Germany) by Dr. Michelle Kilcoyne's group, NUI Galway [57]. The majority of these lectins are able to recognize mammalian glycans. Surprisingly many lectins are useful in the analysis of human cells and glycoproteins, even though many of the probe lectins are derived from plant and micro-organisms (fungi) (Table 2.5)

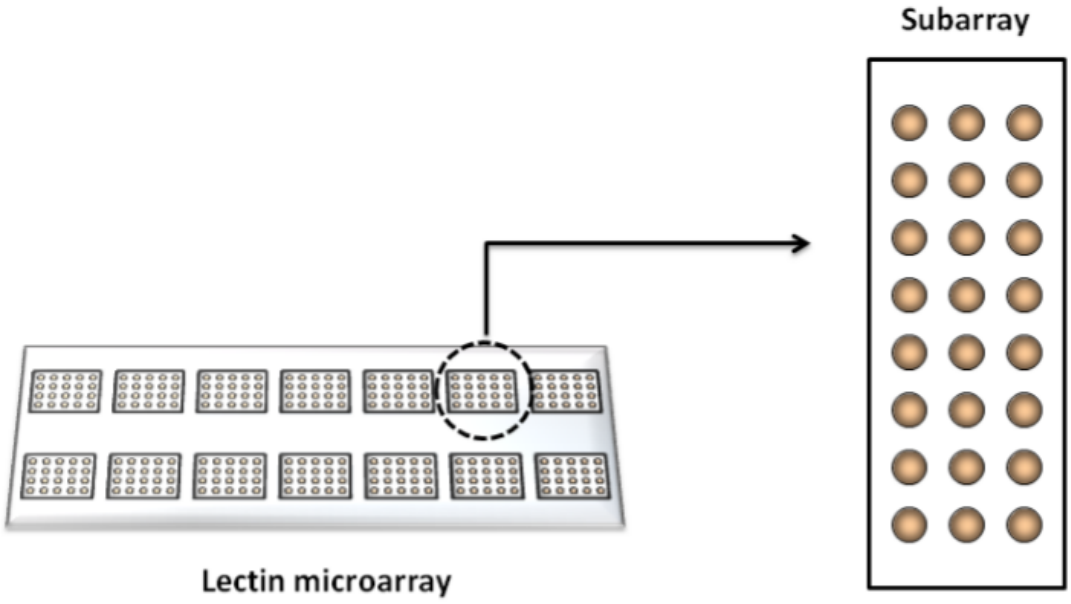


Figure 4.2 **The image illustrates a miniature version of the lectin microarray**. Each microarray slide was printed with 14 defined subarrays available for sample analysis. Each subarray contains the lectin panel with multiple replicate spots (replicates of 6) for each lectin. Diagram adapted from Ku-Lung Hsu, 2008[59].

4.2.2 Mucin Microarray

Thus far, we know that the capacity of gut bacteria to adhere to the mucus layer of the GIT is the first step to colonization and proliferation of the natural community. Insight into how bacteria interact with mucus could lead to the development of novel strategies to prevent infection and novel strategies to promote colonization by beneficial bacteria such as probiotics and other dietary interventions and therapeutics [204]. Fortunately, there has been a growth in the development of novel, improved tools to make such studies more attainable than previously before.

Traditionally, mucins were analyzed by interrogating lectins on porcine gastric mucins (PGM) using flow cytometry[249]. Tissue microarrays were synthesized using gastric biopsies and gall bladder samples to investigate mucin *in situ*. These techniques are not favorable for high throughput mucin analysis because they often require the use of large amounts of reagents and purified mucin. Artificial glycopolymers that mimic natural mucins (e.g. spatial positioning) were constructed by the Bertozzi group[58] and printed on a microarray. Additionally, synthetic mucins have also been synthesized and printed on microarray slides. The disadvantage of these techniques is that they are unable to provide an accurate picture of the binding capacity of carbohydrate binding proteins and whole bacteria in a physiological, biological context [103].

In the past, one of the main limitations in studying bacteria-mucin interactions was the accessibility of human and animal mucus. Another limitation is the inherent differences in the glycosylation of the mucins in different individuals. Finally, obtaining and purifying mucins is a costly and lengthy procedure that often yields low amounts of material. Many of these limitations were overcome by the advent of a novel high throughput mucin microarray platform [107]. A natural mucin microarray containing thirty seven mucins from the gastrointestinal and reproductive tracts of six different animal species (bovine, equine, porcine, chicken, ovine and deer) was available for this project. Thirty five of the mucins were purified from the mucosal surfaces of the GIT, reproductive and respiratory tracts of the six different animals, whereas two mucins were purified from human GIT cell lines (LS174T and E12) (Table 2.6). Mucin microarrays were developed because they require small amounts of purified mucins and simplify profiling of mucin glycosylation and the analysis of bacteria-mucin interactions in a HTP format.

The protein backbone of mucin was used to conjugate the mucin to a NHS functionalized hydrogel via their amine groups to facilitate optimal presentation and accessibility of the oligosaccharide on the mucins. An important advantage to the use of 3D hydrogel slides is that they eliminate the need for an initial blocking step. The absence of a blocking reagent is advantageous because whole bacteria may bind non-specifically to "sticky" molecules like BSA or other blocking agents [106].

After printing, the glycan profiles of the mucins present on the microarray were assessed with the use of lectins and antibodies. It was discovered that accessible glycan motifs varied according to localization of mucin origin and species [105]. Overall, Kilcoyne *et al.,*[57] were able to demonstrate that natural mucin microarrays are a vital and effective tool for profiling mucin glycan epitopes.

An advantage of the mucin microarray is that it enables researchers to perform high throughput and quantitative analysis of the interactions of fluorescently labeled bacteria with mucins. Most importantly, the mucin microarray enables efficient use of the limited quantities of mucins by increasing the number of binding experiments performed and optimizes 3D presentation of the glycan for efficient access of the bacteria to potential receptors [250]. Indeed, unlike traditional arrays with single glycan presentations, mucins contain hundreds of glycans giving them a "bottle brush" appearance. Finally, previous research has shown that protein-glycan bindings usually have low binding affinities. Mucins overcome this limitation due to the multimeric presentation of their glycans maximizing the potential for high affinity binding occurring [96]

One of the first experiments performed on the mucin microarray was to elucidate the mechanism of interaction of *Campylobacter jejuni* and *Helicobacter pylori* with mucus and mucins [58]. Both these strains are phylogenetically closely related and colonize different niches across the human GIT. Four strains of *H.pylori* and six strains of *C. jejuni* were fluorescently labelled with SYTO 82 and interrogated onto the natural mucin microarray. It was found that the strains of both *C.jejuni* and *H.pylori* bind to different sets of mucins. The natural mucin microarray were further used to elucidate the interaction of *Lactobacillus salivarius* and *Bifidobacteria longum*. Both these strains interacted with several mucins on the microarray to differing degrees. The pattern of their interaction was different and their binding tropism was not related

to species or location of the mucin. This further highlights the importance of specific mucin glycosylation [58].

Very recently, a human colonic mucin microarray was developed using mucin derived from patients with colon cancer and ulceritive colitis. The mucin microarray was constructed in a very similar manner to the natural mucin microarray [251] and was used to analyse the interactions of *A. muciniphila* and *Desulfovibrio spp.* Both *A. muciniphila* and *Desulfovibrio spp*. exhibited differences in their affinity for mucin for inflamed and non-inflamed colon [251].

Overall, this chapter presents the use of a novel, natural mucin microarray platform containing 37 purified mucins (Table 2.6) from the reproductive and gastrointestinal tract of six different species (equine, bovine, ovine, porcine, chicken and deer) as a constructive tool for characterizing the adhesive clones described in chapter 3. Specifically, the goal was to identify if the clones conferred any enhanced binding to known mucins. This could potentially lead to the identification of novel adhesin-glycan interactions that are important in the human gut and might help to explain the increased adhesion of these clones to Caco-2 cells that was described in chapter 3. Two functionally characterized adhesive clones (FC3 & FC21) were interrogated onto the mucin microarray to determine how they colonise and interact with the mucin on the arrays [102].

4.2.3 Neoglycoconjugate Microarray (NGC)

As mentioned earlier, the main foundation of glycobiology is based on carbohydrate-protein interactions. In nature, glycoconjugates are heterogeneous and complex which makes them difficult to use as tools in glycobiological research. They are difficult to obtain in large and sufficient quantities (accumulating natural oligosaccharides in pure state is very labour intensive) making it difficult to evaluate ligand-protein receptor recognition. This phenomenon of heterogeneity in natural glycoconjugates can be demonstrated by the oligosaccharide chains of erythropoietin (EPO) produced in Chinese Hamster Ovary (CHO) which have been shown to contain more than 20 kinds of oligosaccharide structures. Likewise, tissue-plasminogen activator (TPA) produced in CHO cells show a large amount of heterogeneity of oligosaccharide chains[178]. Increased heterogeneity of natural glycoconjugates makes it difficult to prepare synthetic carbohydrate analogues (structurally homogenous carbohydrate ligands) for

use in the study of carbohydrate binding molecules. In this study, the goal was to identify if the clones conferred any enhanced binding to known glycoproteins and neo-glycoconjugates.

To compensate for many of these limitations, synthetic glycoconjugates, also known as neoglycoconjugates, which contain carbohydrate residues with the main structural feature involved in the binding of the ligand, have been generated [252]. Scientists have coined the term "neoglycoconjugate" to depict the many areas of preparation and application of semi-natural and synthetic carbohydrate derivatives that are useful for scientific research and medical applications [178]. This conjugation of proteins, lipids or proteoglycans with carbohydrate derivatives is not a novel phenomenon. Already, in the 1980s Stockwell and Lee successfully attached hapten glycosides to proteins so as to raise antibodies [252]

Proteins are the most common medium on which to construct NGC. This is due to the fact that they are ideal for modification with carbohydrate derivatives because of the availability of their side chains [253]. Most are also soluble in aqueous solutions. Carbohydrate groups can be attached to the protein via chemical or enzymatic means. The most commonly used peptide side chain on proteins for such attachment is the amino group because of its abundance, accessibility and chemical reactivity [254]. Often, multiples of one type of monosaccharide or polysaccharide unit are attached to a single protein in order to increase binding affinity and biological activity. It is also possible to attach a single unit to the protein. Often, proteins such as enzymes, growth factors and immunoglobulins are used as media for constructing neoglycoproteins. The advantage of this is that a double agent is produced; a NGC that serves both a carbohydrate mediated function and the original biological function.

NGC microarrays were developed as a rapid and robust high throughput screening platform for carbohydrate-protein interactions. It has gradually been recognized as an essential tool in glycobiological research, drug discovery and medicine. Much like other carbohydrate-based microarrays, it permits the analysis of many combinations of carbohydrate structure and presentations [253]. It also increases the number of possible experiments to be performed with limited sample quantities. Today, it is possible to array a variety of glycan-containing molecules such as neoglycolipids, natural glycoproteins, neoglycoproteins, and carbohydrates with various linkers,

polysaccharides, glycosaminoglycans and mucins. The density and scaffolding of the carbohydrate ligand presented on the microarray depends largely on the molecule. Therefore, protein (lectins, adhesins) recognition and binding will be affected.

In this study, we used a microarray surface arrayed with mono-and di-saccharide neoglycoconjugates (NGCs), using bovine serum albumin (BSA) as a multivalent molecular scaffold, and glycoproteins for presentation of naturally occurring oligosaccharides (Table 2.4). The neoglycoconjugate slides were incubated with fluorescently-labelled whole bacterial cells. In addition, neoglycoconjugate analogues with two different common linkers (4-aminophenyl (4AP) and lisothiocyanatte (ITC) were included in the group to assess the influence of these linkers on protein-carbohydrate interaction.

In the past, bacteria-carbohydrate interactions were difficult and time-consuming to study, requiring the use of specialised equipment's and expertise (e.g. isothermal titration calorimetry ITC) and surface plasom resonance (SPR)). However, in 2004, Disney and Seeberger[255] interrogated *Escherichia coli* cells using carbohydrate microarrays. Since then, the interrogation of carbohydrate microarrays with whole cells has flourished [256].

4.2.4 Advantages of Neoglycoconjugates

The advent of neoglycoconjugates has provided numerous benefits to the fields of science and medicine. One of the main benefits of neoglycoconjugates is that they enable researchers to study carbohydrate-protein interactions. Indeed, the design and construction of neoglycoconjugates has provided us with glycoconjugates containing carbohydrate groups of known structure and guaranteed purity [257]. As mentioned earlier, natural carbohydrates are heterogeneous. Thus, it is very rare to obtain natural glycoconjugates with a single unique carbohydrate structure. Unlike natural glycoconjugates, neoglycoconjugates can be constructed and generated in large quantities [178]. Often the complete structure of the carbohydrate group in natural glycoconjugates is not incorporated into the neoglycoconjugate because the complete glycan structure is often not necessary for biological function. This makes it easier to produce neoglycoconjugates in mass quantities.

Although the process of isolating and purifying bioactive polysaccharides is laborious and time consuming, it is still possible to identify short oligosaccharide sequences or

even monosaccharides responsible for biological activity to use in the construction of effective neoglycoconjugates [257]. In fact, in most cases of carbohydrate recognition studied so far, the range of recognition on the carbohydrate is relatively short (mono-to-pentasaccharides)[253].

Sometimes the recognition marker may be a longer, more complex oligosaccharide. In this case, it is still possible to construct the corresponding neoglycoconjugate but the process may be more demanding. Overall, it is still less labor intensive to construct these neglycoconjugates than it is to use naturally derived glycoconjugates for experimnts due to their limited availability.

Neoglycoconjugates provide valuable physical properties such as solubility, ease of labelling and hydrodynamic size [257]. Another benefit of neoglycoconjugates lies in their ability to provide multivalency far above that found in natural glycoconjugates. Usually, the monovalent interaction between a single carbohydrate ligand and single receptor binding domain of a protein is very weak. Multivalency enhances binding of the target carbohydrate ligand by the carbohydrate-binding protein. Moreover, carbohydrate-binding proteins often possess two or more binding sites or assemble into a multivalent complex to enhance binding [257]. In turn, the carbohydrate ligand must contain appropriate orientation and spacing. A strong binding affinity effect can be generated when both lectin and ligand are multivalent [179].

4.3 Biofilm Assay

Over the years, research has shown that the human body is inhabited by a complex bacterial communities that are able to thrive as surface-attached and self-adherent aggregations known as biofilms. Biofilms mainly consist of bacteria lodged in an extracellular matrix of host and bacterial polysaccharides, proteins and DNA [258]. Often, the ability of bacteria to form biofilm is crucial to their capacity to adhere, persist and survive on surfaces [259]. Although biofilms are pervasive and can be found on many natural environments, understanding of gastrointestinal microbiota biofilm remains limited [260]. As a result, there has been a heightened interest in the role of the mucosal microbiota that occur in biofilms on the surface of the gut. Recent research has shown that the microbial community that makes up the biofilm layer in the gut plays a critical role in colorectal cancer [261].

Indeed, over the past years, there have been several publications demonstrating the presence of biofilms in the colonic microbiome of the human gut. Until recently, the only information obtained about the intestinal microbiota was derived from studies with human faeces and luminal material [262]. Now we know that the human intestine is lined with mucosal bacteria growing in biofilms on the surface lining the gut [263]. The colon mucosa is covered by a layer of mucus that protects the underlying epithelial cells. A rupture in the protective mucus layer will lead to the increased association between the mucosal microbial community and the colonic epithelial cells [264]. This increased access of mucosal microbiota with the epithelial cells invariably leads to the formation of biofilm. The effect of the direct contact of biofilm with epithelial cells leads to a disturbance in epithelial metabolism and facilitates the onset of colorectal cancer [265]. Unfortunately, little research has been conducted to study the factors that contribute to colorectal biofilms to date.

The development of biofilms occurs in five different steps; (1) the reversible aggregation of cells on a surface, (2) irreversible adhesion accelerated by flagella, type IV pilli and adhesins, (3) the development of micro-colonies (4) maturation of the biofilm and (5) the detachment of the cells and the dispersion into a new niche[266]. This study will focus on investigating the biofilm-forming capability of two putative adhesive clones (FC3 & FC21) as compared to a control strain EPI300 (pCC1FOS).

RESULTS:

4.4 Analysis of six putative adhesive fosmid clones (FC3, FC18, FC19, FC20, FC21 and FC22)

The six putative clones identified by functional metagenomics were cleaved with the restriction endonuclease *Bam*HI to estimate the relative size of the clone inserts and to investigate the diversity of all six clones by comparing their restriction profiles. The restriction endonuclease *Bam*HI was chosen to cleave the fosmid clones because it cleaves at two sites (location 353 and 365 bp) adjacent to the insert, which is located in the *lacZ* gene at nucleotide 361 of 8139 bp on the vector map (Figure 3.3, Chapter 3)[192].

As shown in Figure 4.3, fosmid clones FC18 (Lane 2), FC19 (Lane 4) and FC3 (Lane 6) appear to have the same restriction profiles when cleaved with *Bam*HI. These three clones also exhibited the same restriction profiles when cleaved with the restriction endonuclease *Eco*RI (data not shown). The 1% w/v agarose gel electrophoresis (Figure 4.3) allows the presence of the fosmid insert to be confirmed, the relative size of the fosmid insert to be estimated, and also establishes whether the clones are genetically distinct. As illustrated in Figure 4.3, the presence of the pCC1FOS vector (8139 bp) was confirmed for all clones except clone FC20 (lane 8). FC20 consistently showed no bands when cleaved with varying restriction endonucleases. The presence of the fosmid metagenomic DNA insert was confirmed for all clones except fosmid clones FC20 (Lane 8) and FC22 (Lane 12). As seen in lane 12, after cleavage with *Bam*HI, FC22 exhibited a single band at the expected size of the fosmid vector (8139 bp), but no insert band was evident. Based on the 1% agarose gel, the approximate size of each fosmid clone insert was determined. All six clones were run on an agarose gel electrophoresis with a high molecular weight ladder (>10,000 bp) (data not shown) Fosmid clones FC18, FC19 and FC3 were determined to have an insert of ~24.6 kb. Fosmid clone FC21 was determined to have an insert of ~8.1 kb.

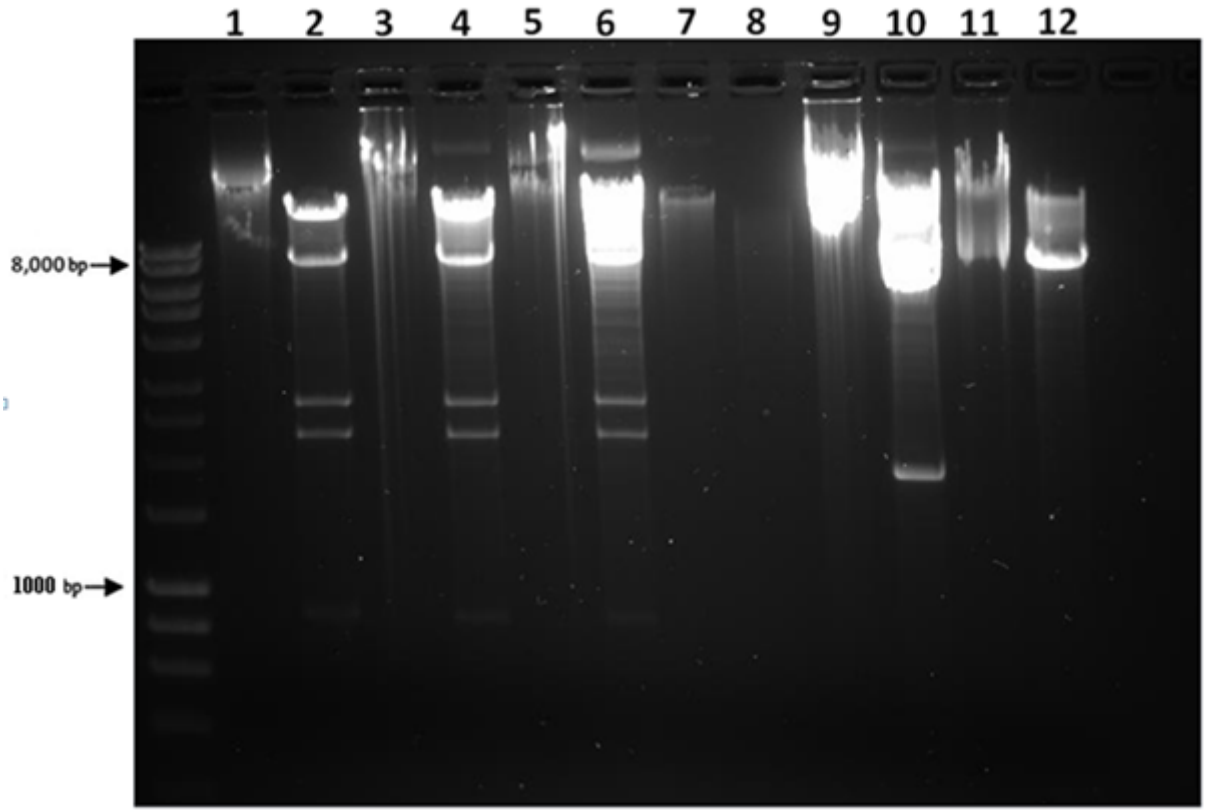


Figure 4.3 **1% Agarose gel electrophoresis illustrating *Bam*HI restriction profiles of 6 fosmid clones.** MW ladder, Lane 1 = uncut FC18, Lane 2 = cut FC18, Lane 3 = uncut FC19, Lane 4 = cut FC19, Lane 5 =uncut FC3, Lane 6 = cut FC3, Lane 7=uncut FC20, Lane 8 = cut FC20, Lane 9 = uncut FC21, Lane 10 = cut FC21, Lane 11 =uncut FC22, Lane 12= cut FC22

As fosmid clones FC18, FC19 and FC3 exhibited the same restriction pattern/profile when cleaved with *Bam*HI, it was hypothesized that all three fosmid clones were the same clone. To test this hypothesis, all six clones were end-sequenced. The results of the end-sequencing of all six clones are shown in Table 4.4. The Basic Local Alignment Search Tool (BLAST) was used to analyse the retrieved nucleotide sequences, using BLASTn. The vector sequences were clipped off for all retrieved sequences in all the clones before BLAST queries were performed.

As shown in Table 4.4, the BLASTn result of the FC21 forward and reverse sequences generated a 95% identical hit in the database to *Bifidobacterium adolescentis* ATCC 15703. *B. adolescentis* ATCC 15703 is a normal inhabitant of the gut. Colonization of *B. adolescentis* in the gut occurs very soon after birth[267]. Their population in the gut

remains relatively stable until late adulthood, where factors such as diet, stress, and antibiotics cause the community to decline[267].

No nucleotide sequences were retrieved for FC20 after end-sequencing. This is not surprising since the gel image (Figure 4.3, lane 7 and 8) of fosmid clone FC20 did not generate any bands when cleaved with several restriction endonucleases, suggesting that this putative clone did not carry a fosmid. The BLASTn result of the FC3 forward sequence generated a 99% identical hit to a vector Pig DNA sequence from clone WTSI for 32% of the query sequence. There was no match to the other 68% of the query. This same results are observed in the BLASTn output of the forward sequences of fosmid clones FC18 and FC19. Overall, fosmid clones FC18, FC19 and FC3 exhibited similar Blastn results. CLUSTALW sequence alignment of the retrieved clone sequences of FC3, FC18 and FC19 showed 96% similarity (data not shown).

The forward end sequence of FC22 generated a BLASTn output of "No significant similarity found." These results are surprising since we already observed that FC22 (Figure 4.3, Lane 12) only contains a vector band. The vector nucleotide sequence were truncated before performing a BLASTn search. The information provided by end-sequencing of the fosmid clones was not sufficient to make effective conclusions about the origins of the insert from each fosmid clone. In order to further confirm these results, all six fosmid clones were sent off to be sequenced by Next Generation Sequencing Illumina HiSeq 1500 to the Genomics & Sequencing lab at Auburn University, Alabama.

Table 4.4 **End-Sequence Result of Six Fosmid clones with subsequent BLAST analysis**. Table depicting end sequence results of six putative clones with subsequent bioinformatics analysis.

Fosmid Clones	**Description**	**Max Score**	**Total Score**	**Query Cover %**	**Max Identity %**
FC3 Forward Blastn	Pig DNA sequence from clone WTSI 1061-50I12 on chromosome Y, complete sequence	551	1421	32%	99%

FC3 Reverse Blastn	No Significant Similarity Found	/	/	/	/
FC18 Forward Blastn	Pig DNA sequence from clone WTSI 1061-50I12 on chromosome Y, complete sequence	551	1421	32%	99%
FC18 Reverse Blastn	No Significant Similarity Found	/	/	/	/
FC19 Forward Blastn	Pig DNA sequence from clone WTSI 1061-50I12 on chromosome Y, complete sequence	551	1421	32%	99%
FC19 Reverse Blastn	No significant Similarity Found	/	/	/	/
FC20 Forward	No nucleotide sequence was retrieved after sequencing	/	/	/	/
FC20 Reverse	No nucleotide sequence was retrieved after sequencing	/	/	/	/
FC21 Forward Blastn	*Bifidobacterium adolescentis* ATCC 15703 complete genome.	1306	1306	92%	95%
FC21 Reverse Blastn	*Bifidobacterium adolescentis* ATCC 15703 complete genome	752	752	81%	95%
FC22 Forward Blastn	No significant similarity found	/	/	/	/

FC22 Reverse Blastn	No significant similarity found	/	/	/	/

4.5 Next Generation Sequencing Data of six Fosmid clones.

In this study, six putative adhesive clones were sequenced using the NGS Illumina sequencing platform (HiSeq, 500 cycles, 100x2 Pair End sequencing). Each clone was sequenced with 100x coverage producing millions of reads for each fosmid. After sequencing was performed, the DNA sequences for each clone were *de novo* assembled using the software Geneious[268]. *De novo* genome assembly is performed without prior knowledge of the source of the DNA sequence length, composition or layout[269]. As a sequence assembler, Geneious is able to assemble and produce long contiguous pieces of sequence (contigs) from the generated reads[268]. Further bioinformatic analysis of the assembled sequences was then performed involving gene predictions and functional annotation.

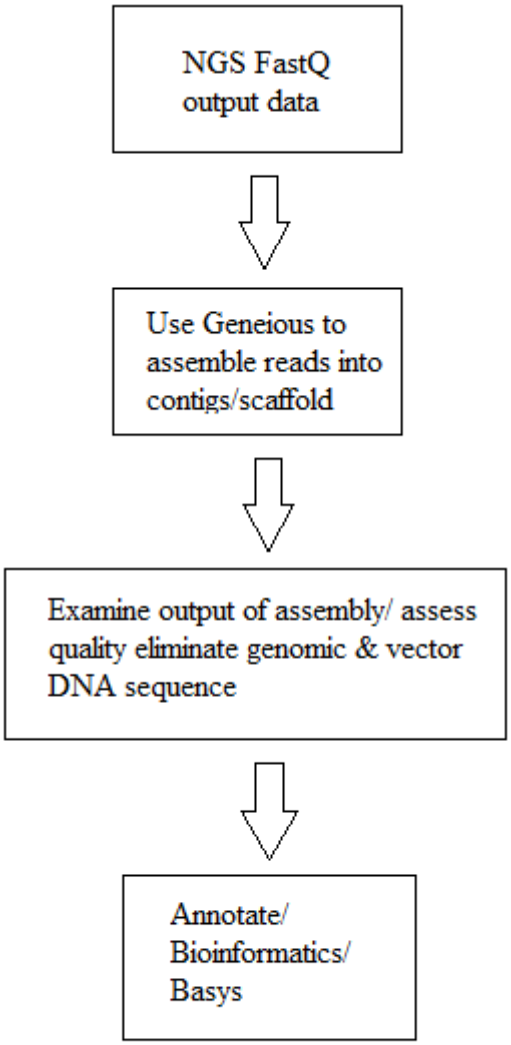


Figure 4.4 **Flowchart** depicting the stepwise *de novo* assembly process using Geneious software.

Once genomic and vector DNA were eliminated for each fosmid (section 2.4.8), insert contigs were re-assembled to produce the long, complete and contiguous insert. The assembled insert sequences for FC3, FC18 and FC19 were aligned using ClustalW. The results confirmed that all three clones were the same clone (Data not shown). Figure 4.4 is a flowchart that illustrates the stepwise *de novo* assembly process of the fosmid clones using the Geneious software. Once the full, contiguous insert sequence has been deciphered, bioinformatics tools such as BLAST[270], ClustalW[271] and BASys[272] are utilized to annotate the insert sequence.

Next generation sequencing analysis also permitted the direct comparison of the sequence of clones FC18, FC19 and FC3 which had earlier been suspected to be genetically related based on restriction analysis (Figure 4.3). From the end sequencing performed on the clones FC3 and FC21, it was demonstrated that FC3 contained a DNA insert that did not match any genomic entries in the DNA database (Table 4.4, FC3, FC18 & FC19). Therefore, it was not possible to determine the genes present in that clone and it was necessary to sequence the entire insert by NGS. Additionally, for the clone that did have a match in the database, it was desirable to confirm the presence of the genes present as predicted from the database entry (Table 4.4, FC21).

The Geneious software was able to *de novo* assemble a FC22 (Table 4.5) fastq file containing 33,769 reads in two minutes and fifty-five seconds. 24,053 of 33,769 (71%) reads were assembled to produce 5,886 contigs covering both fosmid and vector insert. The time required for the Geneious *de novo* assembler depended largely on the hardware, the size of the dataset, and the settings used for assembly. Often reducing the "sensitivity" setting led to a decrease in the stringency of the assembler and a reduction in the time required to complete the assembly. When there are multiple contigs produced, the assembly report also indicated the N50 statistic which is commonly used to measure the quality of an assembly. 71% of the available reads were assembled into contigs. All 5,886 contigs consisted of 100 bp or more. In fact, the minimum length of contig was 110bp. The median length was 130 bp and the mean length was 148 bp. The longest contig out of 5,886 consisted of 942 bp.The assembly report for the other FC22 fastq files are not shown, once assembled, a BLAST search was performed for the FC22 contigs to determine if the DNA was vector derived or chromosomal *E. coli* DNA. Contigs belonging to vector or chromosomal DNA were discarded. As observed in Figure 4.3, only vector DNA sequence was observed for FC22. No metagenomic insert DNA was present.

Table 4.10 illustrates the assembly report generated for part of the reads generated after sequencing FC21. Out of 164,428 reads, 155,041 were assembled to produce 13, 602 contigs. BLAST searches were performed with each contig to determine if they were metagenomics DNA, vector DNA or chromosomal *E. coli* DNA. The elimination of vector DNA and chromosomal DNA led to the discovery of the FC21 metagenomic DNA insert originating from *Bifidobacterium adolescentis* ATCC 15703.

Table 4.5 **Geneious assembly report/statistics of FC22.** Assembly report detailing the *de novo* assembly output for some of the reads generated for FC22

Assembly Report FC22				
24,053 of 33,769 reads were assembled to produce 5,886 contigs				
9,716 reads were not assembled				
Assembly Duration: 2 minutes and 55 seconds				
Statistics	**Unused Reads**	**All Contigs**	**Contigs >=100bp**	
Number of	9,716	5,886	5,886	
Min Length (bp)	110	110	110	
Median Length (bp)		130	130	
Mean Length(bp)	110	148	148	
Max Length (bp)	110	942	942	
N50 length (bp)		153	153	
Number of contigs >=N50		2,185	2,185	
Length Sum (bp)	1,068,760	876,615	876,615	

Table 4.6 **Geneious assembly report/statistics of FC3.** Assembly report detailing the *de novo* assembly output for some of the reads generated for FC3. 18,548 of 24,010 reads were assembled to produce 3,226 contigs.

Assembly Report FC3
18,548 of 24,010 reads were assembled to produce 3,226 contigs
5,462 reads were not assembled
Assembly duration: 13 minutes and 50 seconds

Statistics	Unused Reads	All Contigs	Contigs > =100bp	Contigs > =1000bp
Number of	5,462	3,226	3,226	4
Min Length (bp)	110	110	110	1043
Median Length (bp)		149	149	1328
Mean Length (bp)	110	169	169	1719
Max Length (bp)	110	3179	3179	3179
N50 Length (bp)		177	177	1567
Number of contigs > =N50		1101	1101	2
Length sum (bp)	600,820	546055	546055	6879

Table 4.7 **Geneious assembly report/statistics of FC3.** Assembly report detailing the *de novo* assembly output for some of the reads generated for FC3. 17,155 of 22,460 reads were assembled to produce 4,160 contigs.

Assembly Report FC3				
17,155 of 22,460 reads were assembled to produce 4,160 contigs				
5,305 reads were not assembled				
Assembly duration: 5 minutes and 23 seconds				
Statistics	**Unused Reads**	**All Contigs**	**Contigs > =100bp**	**Contigs > =1000bp**
Number of	5,305	4,160	4,160	3
Min Length (bp)	110	110	110	1,026
Median Length (bp)		141	141	1,051
Mean Length (bp)	110	160	160	1,296
Max Length (bp)	110	1,812	1,812	1,812
N50 Length (bp)		171	171	1,051

Number of contigs > =N50		1,478	1,478	2
Length sum (bp)	583,550	668, 259	668, 259	3,889

Table 4.8 **Geneious assembly report/statistics of FC3.** Assembly report detailing the *de novo* assembly output for some of the reads generated for FC3. 10,465 of 22,460 reads were assembled to produce 2,808 contigs.

Assembly Report FC3				
10,465 of 22,460 reads were assembled to produce 2,808 contigs				
11,995 reads were not assembled				
Assembly duration: 4 inutes and 49 seconds				
Statistics	**Unused Reads**	**All Contigs**	**Contigs > =100bp**	**Contigs > =1000bp**
Number of	11,995	2,808	2,808	1
Min Length (bp)	110	110	110	1,192
Median Length (bp)		127	127	1,192
Mean Length (bp)	110	145	145	1,192
Max Length (bp)	110	1,192	1,192	1,192
N50 Length (bp)		147	147	1,192
Number of contigs > =N50		1,056	1,056	1
Length sum (bp)	1, 319, 450	409,437	409,437	1,192

Table 4.9 **Geneious assembly report/statistics of FC3.** Assembly report detailing the *de novo* assembly output for some of the reads generated for FC3. 11,835 of 24,010 reads were assembled to produce 3,160 contigs.

Assembly Report FC3
11,835 of 24,010 reads were assembled to produce 3,160 contigs

12,175 reads were not assembled				
Assembly duration: 6 minutes and 39 seconds				
Statistics	**Unused Reads**	**All Contigs**	**Contigs =100bp**	**Contigs > =1000bp**
Number of	12,175	3160	3160	
Min Length (bp)	110	110	110	
Median Length (bp)		124	124	
Mean Length (bp)	110	144	144	
Max Length (bp)		816	816	
N50 Length (bp)		146	146	
Number of contigs > =N50		1190	1190	
Length sum (bp)	1,339,250	456,627	456,627	

Table 4.10 **Geneious assembly report/statistics of FC21.** Assembly report detailing the *de novo* assembly output for some of the reads generated for FC21. 155,041 of 164,428 reads were assembled to produce 13,602 contigs.

Assembly Report FC21				
155,041 of 164,428 reads were assembled to produce 13,602 contigs				
9,387 reads were not assembled				
Assembly duration: 25 minutes and 28 seconds (31 minutes and 47 seconds)				
Statistics	**Unused Reads**	**All Contigs**	**Contigs >=100bp**	**Contigs >=1000bp**
Number of	9,387	13,602	13,602	72
Min Length (bp)	110	110	110	1000
Median Length (bp)		190	190	1233
Mean Length (bp)	110	238	238	1302

Max Length (bp)	110	2795	2795	2795
N50 Length (bp)		271	271	1281
Number of contigs >=N50		3729	3729	30
Length Sum (bp)	1,032,570	3,240,775	3,240,775	93,756

In contrast to the sequencing output of FC22 (Table 4.10), 94% of the available reads of FC21 (Table 4.7) were assembled to produce 13,602 contigs, 72 contigs consisted of 1000bp or more. The minium length of the majority of the 13,602 was 110 bp. The maximum length consist of 2,795bp. Assembly of FC21 generated long contigs. The longer the contigs, the easier it was to put the sequence back together.

In comparing the assembly statistics for the three fosmid clones presented (Tables 4.5-4.10), FC21 is noticeable as having a higher assembly quality than the other assemblies because 94% of the reads were used in the assembly to produce 13,602 contigs. Furthermore, 72 contigs out of the 13,602 contigs were larger than 1000bp. The maximum length of the contigs generated from this assembly is 2795bp.

Overall, the NGS illumina sequenced fosmid clones confirmed the results obtained from both restriction digestion (Figure 4.3) and end-sequencing results (Table 4.4) of each of the 6 putative adhesive clones. The Geneious Assembler was successful in generating and assembling contigs that were re-assembled into the contiguous, complete DNA insert. Further downstream analysis using BLAST helped to corroborate the initial results from end-sequencing the six clones (Table 4.4) Fosmid clones FC3, FC18 and FC19 contained insert sequences with no homology to any organisms in the database and a CLUSTALW alignment of their complete sequences indicated that all three clones were identical.

4.6 Bioinformatics Analysis of fosmid Clones FC3 and FC21

After assembling the full, complete sequence of FC21, a BLASTn search of the insert was performed to confirm the results from end-sequencing (Table 4.4). As expected, the results correlated 100% with the results observed from end-sequencing fosmid clone FC21 (Table 4.4). The DNA insert from FC21 was shown to be 8.1kb in length

and to have a 100% similarity to a region of the genome sequence of gut inhabitant *Bifidobacterium adolescentis*.

Figure 4.5 **ORF map of FC21 fosmid clone.** FC21 insert DNA depicting the five genes present on the 8.1 kb insert. Only three genes are functional namely; BAD_0085, aroP and BAD_0086. The length and direction of the arrow indicates the relative size and transcriptional direction of each gene. The start and end positions refer to the start and end coordinates of the DNA on the *B. adolescentis* ATCC 1507 genome sequence. (Screen capture image, NCBI).

Table 4.11 **Table of FC21 ORF**. Five genes present on the 8.1 kb DNA insert of the FC21 fosmid clone with their predicted locations and domains. Table generated using BASys software.

Gene	Protein ID	Locus	Product	Location	Domains present	Start position	End position
Gene 1	YP_908946.1	Bad_0083	sialic acid-specific 9-O-acetylesterase	Cytoplasmic	DUF303	112996	114870
aroP	YP_908947.1	Bad_0084	aromatic amino acid transport protein AroP	Cell inner membrane; multi pass membrane protein	AnsP, prk10238, GABA_permease AA_permease	116537	115008
Gene 3	YP_908948.1	Bad_0085	hypothetical protein	Membrane	FtsX, MacB_PCD, DUFF2203 superfamily, SalY, Mt_ATP-synt_B superfamily	116725	119886

Gene 4	YP_908949 .1	Bad_0086	Na(+)/H(+) antiporter-like protein	Cell membrane; multi-pass membrane protein	KefB domain, Na+/H+ antiporter, PLN03159 Na_H exchanger	121179	119905
Gene 5	YP_908950 .1	Bad_0087	sortase-like protein	Membrane	Sortase_C_3, Uncharacterized_protein_YhcS, Sortase, SrtA	122116	122919

After deciphering the origin of the DNA insert of FC21, an attempt was made to predict the attribute of the clone that confered the adhesive phenotype of FC21. Both sialic acid-specific 9-0-acetylesterase and sortase-like genes were truncated and thus predicted to be non-functional genes (Figure 4.5). Bioinformatics analysis revealed that 1441bp out of 1875bp (77%) was truncated from sialic acid-specific 9-0-acetylesterase gene while 339bp of 804bp (42%) was truncated from the sortase-like protein gene. The hypothetical protein, Bad_0085 (large transmembrane protein possibly involved in transport) was the largest membrane bound gene and contained several putative conserved domains including FtsX, MacB_PCD, DUFF2203 superfamily and SalY (Table 4.11). FtsX-like permease family are a family of predicted permeases and hypothetical transmembrane proteins. The SalY conserved domain refers to ABC-type antimicrobial peptide transport system, permease component. Thus some of the domains present in BAD_0085 are involved with transport.

The sodium/proton antiporter protein, Bad_0086, contained conserved domains including KefB domain, PLN03159 and the sodium/hydrogen exchanger. The KefB domain is a Kef-type K+ transport system, membrane component of KefG which is involved in inorganic ion transport and metabolism. The sodium/hydrogen exchanger are key transporters in maintaining the pH of metabolizing cells. To date, the mechanisms of antiporter are still to be elucidated. The antiporters consist of 10-12 transmembrane regions at the amino terminus and a large cytoplasmic region at the carboxyl terminus. The aromatic amino acid, BAD_0084 is located in the inner cell membrane and is a multipass membrane protein. It contains conserved domains

including an AnsP domain that is involved in L-asparagine transport and related permeases. This domain is involved in amino acid transport and metabolism. Furthermore, BAD_0084 contains a GABA_permease domain that catalyses the translocation of 4-aminobutyrate (GABA) across the plasma membrane, with homologues expressed in Gram-negative and Gram-positive organisms. This permease is a highly hydrophobic transmembrane protein consisting of 12 transmembrane domains with hydrophilic N- and C-terminal ends. Although the three proteins are located in the cell membrane, analysis of the domains and motifs present within each gene did not yield any significant information that would be attributable to the adhesive phenotype of FC21.

As revealed in Table 4.11, each of the three functional genes present on FC21 are all involved in transport. This suggests that each gene is part of a gene cluster that encodes for similar polypeptides, or proteins, which collectively share a generalized function. Therefore, it was possible that all three genes together form an adhesive complex on the surface of FC21 that is responsible for the adherence phenotype.

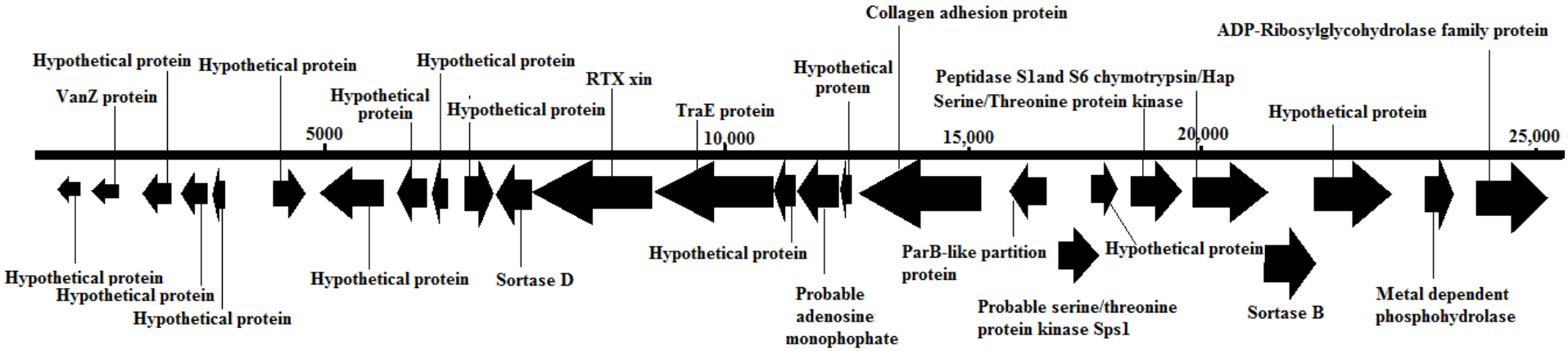


Figure 4.6 **ORF map of FC3 fosmid clone**.26 genes present on the 24.6 kb DNA insert of the FC3 fosmid clone with their predicted locations. The length and direction of the arrow indicates the relative size and transcriptional direction of each gene. (Image derived from SnapGene software)

Sequence analysis of FC3, FC18 and FC19 revealed that they each contain an identical 24.6 kb insert. A BLASTn search of the complete DNA insert sequence of FC3 clone exhibited no known homologs to strains in the database suggesting that the insert DNA is derived from a microbe with an unknown genome sequence. To annotate the complete 24.6 kb FC3 sequence, a web server for automated bacterial genome annotation was used. BASys (Bacterial Annotation System) is a web server that generates automated, extensive textual annotation and hyperlinked image output of bacterial genomic (chromosomal and plasmid) sequences[272]. BASys employs over 30 different programs to deduce approximately 60 annotation subfields for each gene, including the gene/protein name, COG function, GO function, potential paralogues and orthologues, molecular weight, isoelectric point, operon structure, signal peptides, transmembrane regions, 3D structure and a host of other reactions and pathways[272]. The full, complete, raw DNA sequence of FC3 was submitted as a FASTA-formatted file into the BASys web interface. Table 4.12 depicts a subset of the annotations generated by BASys for FC3; including gene number and name, product, cell location and domains present. The 24.6 kb FC3 DNA insert is predicted to contain 26 genes encoding 26 putative proteins located at different regions of the cell (Figure 4.6 & Table 4.12).

Of the 26 putative proteins predicted to be expressed by FC3, approximately five of them are noticeable as being potentially relevant for the adhesive phenotype observed for the FC3 strain. The first criteria for selection of putative adhesive proteins of the 26 proteins is the location of the protein. 11 out of 26 proteins are located on the cell membrane. The other 15 proteins were not taken into account because they are all located in the cytoplasm. The second criteria for selection of putative adhesive proteins in the FC3 insert was the conserved domain predicted to be present in the proteins. For example, Gene 11 is a putative sortase D protein which is located on the cell membrane. The presence of a gene encoding a putative Sortase D enzyme on the FC3 insert is relevant because sortases play a fundamental role in microbial adherence to host cells by anchoring certain sortase-dependent pili and adhesins to the cell wall of gram-positive bacteria[273]. Sortase substrate proteins that are attached to the cell walls of bacteria include enzymes, pilins and adhesion-mediating large surface glycoproteins. These proteins often play critical roles in colonization and virulence by bacteria[274]. Sortases are partitioned into distinct families called class A to F enzymes.

Very little is known about members of the class D sortase enzymes. However, they are known to be present in bacilli and have thus far been characterized in *B. anthracis*[275].

Table 4.12 The 26 predicted gene products encoded by fosmid clone FC3.The 26 putative predicted proteins that are predicted by BASys and other public software available.

Gene	Product	Cell Location	Domains Present
Gene1	Hypothetical protein	Cytoplasm	none
Gene2	Hypothetical protein	Cytoplasm	none
Gene3	VanZ protein	Cytoplasm	VanZ-like family protein
Gene4	Hypothetical protein	Cytoplasm	none
Gene5	Hypothetical protein	Membrane	none
Gene6	Hypothetical protein	Membrane	none
Gene7	Hypothetical protein	Cytoplasm	Sigma70_r4_2, PRK12546, RNA_polymerase_sigma_factor, RpoE
Gene8	Hypothetical protein	Cytoplasm	none
Gene9	Hypothetical protein	Cytoplasm	none
Gene10	Hypothetical protein	Cytoplasm	Cro/C1-type HTH domain, virulence-associated_protein_I, HTH_XRE, HipB, HTH_19
Gene11	Sortase D	Membrane	Sortase Superfamily
Gene12	RTX Xin	Membrane	NlpC/P60 family; Spr Cell-wall associated hydrolase, Phosphotantetheine attachment site.
Gene13	TraE protein	Cytoplasm	TnpV superfamily
Gene14	Hypothetical protein	Cytoplasm	LPLAT superfamily
Gene15	Probable adenosine	Cytoplasm	Fic superfamily

	monophosphate protein transferase y14H		
Gene16	Hypothetical protein	Cytoplasm	Fic superfamily
Gene17	Collagen Adhesion Protein	Membrane	Cna_B
Gene18	ParB-like partition protein	Cytoplasm	ParBC superfamily
Sps1 Gene19	Probable serine/threonine protein kinase	Membrane	PKc_like superfamily
Gene20	Hypothetical protein	Membrane	FHA superfamily
Gene21	Serine/threonine protein kinase	Membrane	PKc_like superfamily
Gene22	Peptidase S1 and S6 chymotrypsin	Membrane	FHA, Tyrpsin_2, DegQ, probable_periplasmic_serine_protease_do/HhoA-like
Gene23	Hypothetical protein	Membrane	none
Gene24	Sortase B	Membrane	Sortase superfamily
Gene 25	Metal dependent phosphohydrolase	Cytoplasm	Hdc superfamily
Gene26	ADP-Ribosylglycohdrolase	Cytoplasm	ADP ribosyl_GH superfamily

Gene 12 (Table 4.12) is predicted to express an exoprotein member of the RTX (repeats-in-toxin) family of toxins produced by Gram-negative bacteria, some of which are known to bind cell membranes and cause disruption of the permeability barrier, leading to efflux of cell contents. There are currently over 1000 known members with a variety of functions. The RTX family is defined by two main features; characteristic repeats in the toxin protein sequences and an extracellular secretion by the type I secretion system (T1SS).

FC3 is predicted to carry an adhesion protein gene (Cna_B; Collagen-binding surface protein Cna, B-type domain). Cna_B (Table 4.12, Gene 17) is a repeated B region domain found in the collagen-binding surface protein Cna in *Staphylococcus aureus*, as well as other related domains. Cna has a non-repetitive, collagen-binding A region, followed by instances of this B region repetitive unit. The B region does not itself bind collagen but has one to four 23 kDa repeat units (B1-B4) each with a prealbumin-like beta-sandwich fold of seven strands in two sheets with a Greek key topology. The Cna_B domain appears to form a stalk that presents the ligand binding domain away from the bacterial cell surface. Cna is a collagen-binding MSCRAMM (Microbial Surface Component Recognizing Adhesive Matrix Molecules), and is necessary and sufficient for *S. aureus* cells to adhere to cartilage[276].

FC3 is predicted to carry two different classes of sortase enzymes; sortase D and sortase B. Sortase B are widely distributed in Firmicutes and have primary sequences that are most closely related to the sortase B enzyme from the species *S. aureus*. In some bacteria, class B sortases attach haemoproteins to the cell wall. Sortases belonging to class B are also present in *Clostridum perfringens*, *Clostridium difficile* and *Enterococcus faecalis*.

Gene 21 is predicted to be a Serine/threonine protein kinase (STPKs) membrane protein. Serine/threonine-protein kinase are proteins that catalyse the phosphorylation of serine or threonine residues on target proteins by using ATP as phosphate donor. Such phosphorylation may cause changes in the function of the target protein by changing enzyme activity, cellular location, or association with other proteins. Protein phosphorylation plays a key role in most cellular activities and is a reversible process. The reverse process is catalysed by phosphoprotein phosphatases. Protein kinases share a conserved catalytic core common to both serine/threonine and tyrosine protein

kinases. The catalytic subunits of protein kinases are highly conserved, and several structures have been solved, leading to large screens to develop kinase-specific inhibitors for the treatment of a number of diseases. Recently, Herbert and colleagues demonstrated that a serine-threonine kinase (StkP) regulates expression of the Pneumococcal pilus and modulates bacterial adherence to human epithelial and endothelial cells in vitro[277].

Gene 20 is a predicted membrane bound hypothetical protein with a FHA superfamily domain[278]. The Forkhead-associated domain (FHA) is a phosphopeptide binding motif found in many regulatory (protein secretion, antibiotic resistance, transcription, peptidoglycan synthesis, metabolism and virulence) eukaryotic and prokaryotic proteins[279]. It displays specificity for phosphothreonine-containing epitopes but will also recognize phosphotyrosine with relatively high affinity[280]. ForkHead-associated domain are the key interacting partners of serine/threonine protein kinases (STPKs) that mediate the signals inside the cells emanating from the cognate kinases, which explains why these genes are next to one another on the FC3 DNA fragment. Gene 20 could possibly be an ABC transporter substrate of the adjacent gene 21. Gunjan Arora and colleagues have shown that most of the FHA domain containing proteins of *Mycobacterium tuberculosis* are phosphorylated by serine/threonine protein kinases[279]. Gene 23 is a hypothetical membrane protein flanked by a Sortase B membrane protein and a peptidase S1 and S6 chymotrypsin protein (Table 4.12). Based on the flanking genes, gene 23 could possibly be a substrate of sortase B or a protein involved in cellular signaling like peptidase S1 and S6 chymotrypsin. Gene 22 is a predicted membrane bound peptidase S1 and S6 chymotrypsin trypsin with FHA domain. FC3 contains an Sps1 gene predicted to encode a probable serine/threonine protein kinase like Gene 21. Bacterial serine/threonine protein kinases (STPKs) are known to regulate cell division by sensing and responding to specific signals in the host environment[279]. Genes 5 and 6 are predicted hypothetical membrane proteins that may encode a protein involved in adherence. Determining the exact gene responsible for the enhanced adherence of FC3 would require sub-cloning of fragments of FC3 or individual genes to assess functionality.

Further analysis of FC3 strongly suggests that the DNA fragment is derived from species belonging to the genus Clostridium. Consistent with this finding, a large portion of the predicted gene products were highly homologous to those of

Clostridium spp. Studies have indicated that clostridium are leading players in the maintenance of gut homeostasis[281]. Commensal Clostridia consist of gram-positive, rod-shaped bacteria in the phylum Firmicutes and make up a substantial part of the total bacteria in the gut microbiota. They start to colonize the intestine of breastfed infants during the first month of life and populate a specific region in the intestinal mucosa in close relationship with intestinal cells[281]. This position allows them to participate as crucial factors in modulating physiologic, metabolic and immune processes in the gut during the entire lifespan, by interacting with the other resident microbe populations, but also by providing specific and essential functions. In conclusion, we present here a bioinformatic analysis of two putative adhesive fosmid clones that may help explain the adhesive properties of the host strain.

4.7 Rescue and re-transformation of fomid clones into *E. coli* host.

Functional screening of two fosmid clones, FC3 and FC21, indicated that they exhibit significant adherence to 3 week-old caco-2 cells when induced with L-arabinose (Figure 4.7A + B). As illustrated in Figure 4.7A, FC3 exhibited a 20-fold increase in adherence as compared to the control (EPI300 (pCC1FOS)). FC21 exhibited a 5-fold increase in adherence efficicency as compared to the control. The results depicted in this experiment (Figure 4.7A) are nearly identical to the results observed for FC3 and FC21 in Figure 3.16D (average of 3 days). This is surprising since the results from *in vitro* adhesion assays have demonstrated high intrinsic and inter-experimental variation. To determine if the observed adherence capability was conferred by the metagenomics insert DNA and not the host chromosomal DNA, the FC3 and FC21 fosmids were rescued and re-transformed into a fresh *E. coli* host. The results indicated that both FC3 and FC21 are statistically significantly more adherent to 3 week-old Caco-2 cells than the control strain (Figure 4.7B). FC3 is 7-fold more adherent than the control and FC21 is approximately 8-fold more adherent than the control. FC3 and FC21 exhibit reproducible adherence, indicating that the adherence capability is likely attributed to specific genes on the inserts.

The next section of this study will describe the interactions of the putative FC3 and FC21 adhesive clones with glycans, mucins and lectins on microarrays. These studies are presented in three sections covering three main types of carbohydrate-based microarray platforms; (i) carbohydrate-binding protein (lectins) microarray which were profiled with whole bacteria to determine glycosylation patterns on the cell

surface of the bacteria, (ii) natural mucin microarrays which were profiled with whole bacteria to determine mucin glycosylation and bacterial binding tropisms, (iii) and finally neoglycoconjugate (NGC) microarrays profiled with whole bacteria to determine the binding affinity of bacteria to specific neoglycoconjugates (NGC).

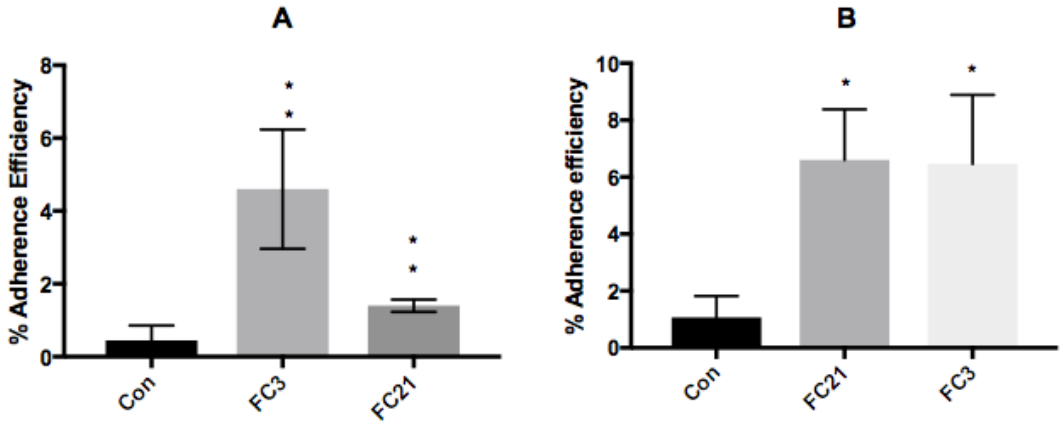


Figure 4.7 **Retransformation of hit fosmid clones into fresh *E. coli* host** (A) Fosmid clones FC3 and FC21 exhibited enhanced adherence to 3 week old Caco-2 cells when grown in the presence of antibiotic and L-arabinose (B) The fosmid clones (FC3, FC21 and control strain EPI300 (pCC1FOS)) were *re-transformed* into fresh host and exhibited a similar adhesive pattern when interrogated onto 3 week old Caco-2 cells when grown in the presence of antibiotic and arabinose. The graphs in A and B represent two separate experiments, each performed with triplicates. Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$.

4.8 Comparison of fluorescent dye uptake

One of the primary concerns in microarray technology is the efficient and equal fluorescent-labelling of the bacterial strains. The nucleic acid staining dye SYTO 82 was used to fluorescently label our fosmid clones. The SYTO family of dyes have been used extensively in many biological applications[282]. The SYTO dyes consist of a family of commercially available cyanine dyes. The interaction cyanine dyes with DNA is complex and is influenced by electrostatic, van der Waals, hydrophobic, and steric interactions, all of which are regulated by the dye's chemical structure[283]. The binding of the SYTO 82 dye appears to be cooperative and impacts the dye concentration and dye-to-base pair ratio. Unlike other cyanine dyes, the SYTO dyes are more hydrophobic. The SYTO dye binds DNA mainly on charge and primarily in the minor groove. Different SYTO dyes exhibit variations in fluorescent enhancement

on nucleic acid binding, excitation and emission spectra, DNA/RNA selectivity, binding mode, and binding affinity[284].

As a result, different concentration ranges for the SYTO dyes are suggested depending on the cell type. Differences in uptake of the SYTO 82 dye by the different clones is likely to occur because staining can be affected by the growth medium used, the cell density, the presence of other cell types and other factors. Fosmid clones FC3 and FC21 and control strain EPI300 (pCC1FOS) were fluorescently labelled with SYTO 82. It was necessary that all three bacterial strains were labelled in an equivalent manner to eliminate the possibility of data bias resulting from differences in bacterial uptake of the dye. The SYTO 82 dye was chosen because the absorption (567 nm) and emission (583 nm) maxima were compatible with the Cy3 emission filter (550-600 nm) of the Genepix 4100A microarray scanner available in the Kilcoyne lab, NUI Galway.

To determine the optimum SYTO 82 concentration for labelling each clone, different concentrations of dye were added to the washed bacterial suspensions to give a final range of 0 - 50μM. The optimal concentration for each strain was determined based on maximum fluorescence uptake (545 nm excitation, 590 nm emission) standardized to the optical density at 600 nm (OD_{600}) versus dye concentration for the bacterial fosmid clones used in this study. Dr. Michelle Kilcoyne's Lab determined that 20 μM concentration of SYTO 82 was chosen as this is the standard saturating concentration. For all subsequent labelling experiments, a concentration of 20 μM was used since additional amounts of dye would not enhance fluorescence staining. Next, a test to determine whether the dye uptake at 20 μM was equivalent between the three strains was performed. This was important since preferential labelling of one bacterial clone could skew results during the lectin microarray experiments. To compare the dye uptake between strains, a one-way analysis of variance (ANOVA) was used. This statistical test permits comparison of 3 or more groups with respect to dye uptake. Three separate replicates for each bacterial strain were used for the analysis (data generated by Dr. Michelle Kilcoyne's Lab). The results of the test (P =0.33, N =3) showed that the uptakes were not different.

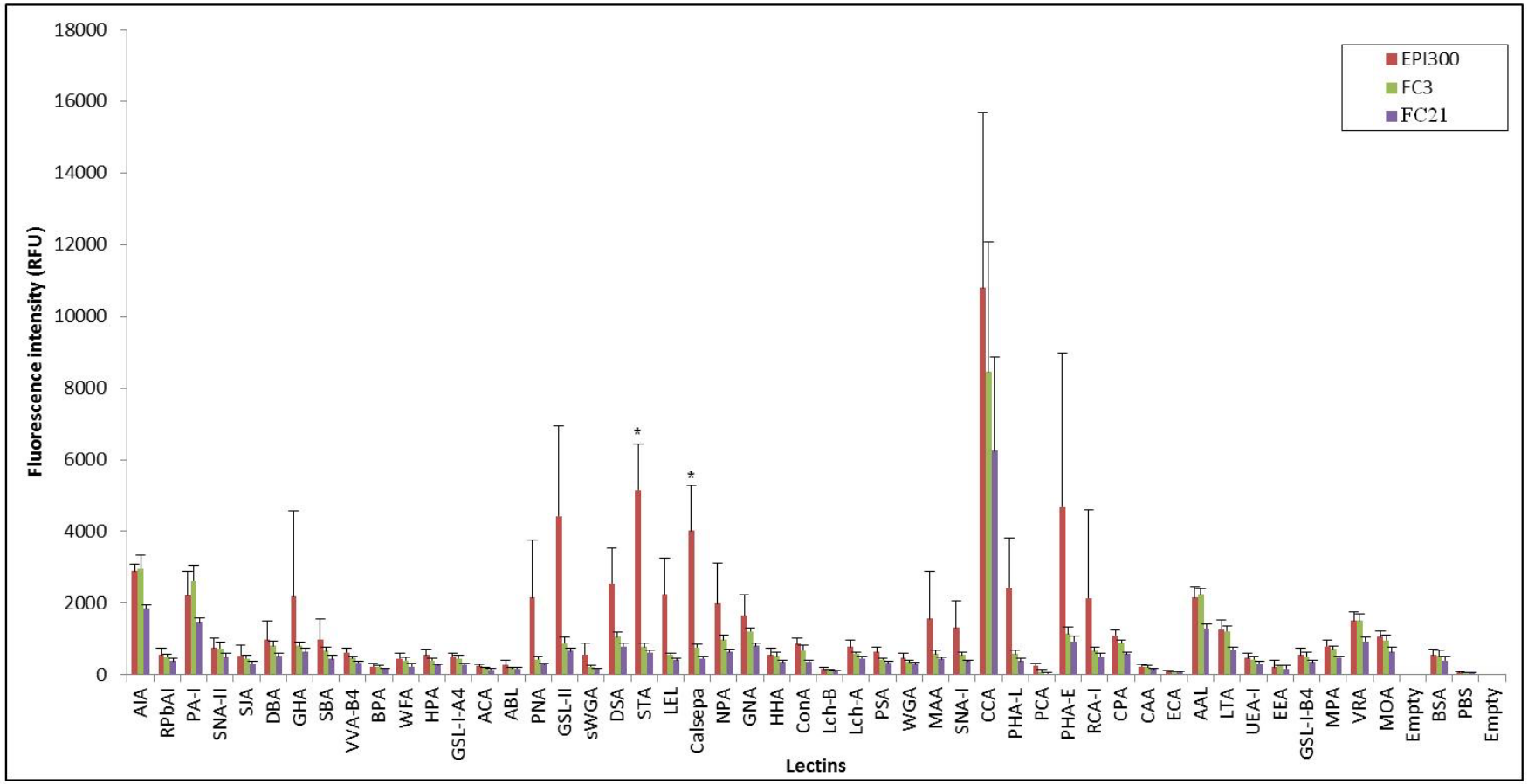
4.9 Lectin microarray results

Each microarray slide was printed with 14 replicate subarrays with each lectin (probe) spotted in replicates of six (Figure 4.2). Bacterial glycosylation of the FC3 and FC21 clones was investigated to determine if the presence of the insert had altered the cell surface glycans of the clones. Different genes being expressed on the clones could potentially yield different bacterial surface glycosylation patterns. In this study, it was important to investigate bacterial glycosylation because previous research has shown that some adhesins have either O-linked or N-linked cell wall anchored glycoproteins on their cell surface[285]. Glycosylation of an adhesin can protect the adhesin against premature degradation and can influence tethering of the adhesins to the bacterial cell surface[286].

The aim of this study was to profile the surface glycome of our adhesive fosmid clones FC3 and FC21 in comparison to the control strain EPI300 (pCC1FOS vector) using the lectin microarray. Each strain was grown overnight at 37^{o}C under aerobic conditions in the presence of L-arabinose and chloramphenicol to stationary phase. The results from the lectin microarray experiments showed distinct lectin binding patterns for the fosmid clones (FC3, FC21) as compared to the control strain EPI300 (pCC1FOS vector) (Figure 4.9A).

Both fosmid clones (FC3 & FC21) demonstrated reduced binding to plant lectins that were specific for GlcNAc (*N*-Acetylglucosamine) when compared to the control EPI300 (pCC1FOS) strain (Figure 4.9A). EPI300 (pCC1FOS) showed strong binding to lectins GSL-11, DSA, STA and LEL, while FC3 and FC21 only gave weak binding to these same lectins. Furthermore, EPI300 (pCC1FOS) showed stronger binding signals to lectins specific for bi-antennary (PHA-E) and tri- and tetra antennary (PHA-L) glycans coated with two to four GlcNAc branches linked to the core glycan in comparison with the fosmid clones. These results suggest that cell surface glycosylation of the fosmid clones had been altered as compared to the control. Bi- and tetra-antennary glycans are coated with two to four GlcNAc branches linked to the core glycan. Fosmid clones that lack significant amounts of GlcNAc residues on their cell surface will not be detected by lectins specific for GlcNAc.

A


Fluorescence intensity (RFU)
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
EPI300
FC3
FC21
*
*
AIA
RPbAI
PA-I
SNA-II
SJA
DBA
GHA
SBA
VVA-B4
BPA
WFA
HPA
GSL-I-A4
ACA
ABL
PNA
GSL-II
sWGA
DSA
STA
LEL
Calsepa
NPA
GNA
HHA
ConA
Lch-B
Lch-A
PSA
WGA
MAA
SNA-I
CCA
PHA-L
PCA
PHA-E
RCA-I
CPA
CAA
ECA
AAL
LTA
UEA-I
EEA
GSL-I-B4
MPA
VRA
MOA
Empty
BSA
PBS
Empty
Lectins

B

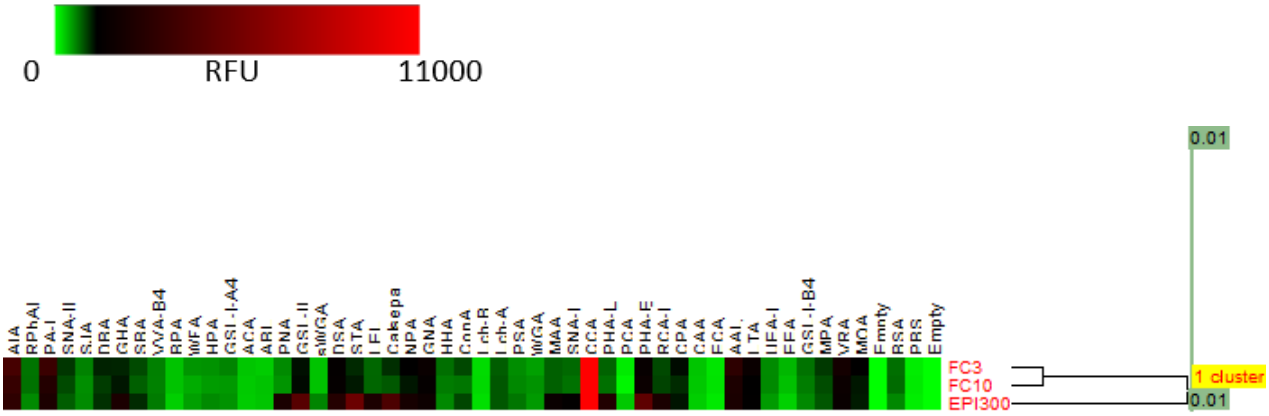


Figure 4.9 **Lectin microarray profile of FC3, FC21 and EPI300 control strain**. (A) Lectin microarray revealed altered binding profiles of fosmid clones FC3, FC21 and EPI300 (empty pCC1FOS vector) control. See Table 4.1 for lectin abbreviations. Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$. (B) Heat map with dendrogram of hierarchical clustering for technical and biological replicates was generated by using Cluster 3.0. Heat map of lectin microarrays of the three different strains (FC3, FC21 and EPI300) Clustering analysis of lectin binding profiles showed similarity in surface glycosylation of the two clones, with the control strain, EPI300, showing 1% surface glycosylation similarity to the two clones.

The weakened interactions of the fosmid clones with lectins specific for GlcNAc residues suggested that the metagenomic DNA insert present in both clones has significantly altered the bacteria's surface glycosylation as compared to the control strain EPI300 (pCC1FOS). The clustering analysis of all three bacterial strains supported the observation that both clones have an altered cell surface glycosylation as compared to the EP1300 (pCC1FOS) (Figure 4.9B) Given that microbial surface glycans have been shown to influence bacterial adhesion[287], these results suggests a role of glycosylation in the observed increased adherence of clones FC3 and FC21.

All of the bacterial strains (EPI300 (pCC1FOS), FC3 and FC21) showed very little binding to plant lectins specific for N-acetylgalactosamine (GalNAc) or galactose (e.g, SBA, BPA, VVA-B4, WFA, HPA and GSL-I-A4). This suggests that the surface sugar moieties specific for these lectins are generally absent from among EPI300 (PCC1FOS) control strain and fosmid clones (FC3 & FC21).

Only the control strain EPI300 (pCC1FOS) showed increased binding of the plant lectin PNA (peanut agglutinin) which is specific for Gal-beta-(1, 3)-GalNAc. Studies by Karen Giannasca and colleagues[288] demonstrated that a Galβ(1-3) GalNAc epitope recognized by PNA and located in the glycocalyx is involved in the early recognition events between *Salmonella Typhimurium* and Caco-2 cells[288]. The results suggests that both FC3 and FC21 have few Galβ (1-3) GalNAc residues on their cell surfaces as compared to the control.

The results further demonstrate a strong signal of all strains (EPI300 (pCC1FOS), FC3 and FC21) in binding to a sialic acid animal lectin (CCA) with high specificity for O-acetyl sialic acids. This suggests that all three strains contain relatively high amounts of O-acetyl sialic acids on their surface glycoconjugates. Of the known lectins which have been purified and characterized, few bind sialic acids. Wheat germ agglutinin (WGA) is one of the few plant lectins that bind to sialic acids. However, it also binds to oligosaccharides containing N-acetylglucosamine (GlcNAc) and N-acetylneuraminic acid (NeuAc). The weak binding of our bacterial cells with WGA seems to suggest that there are few oligosaccharides containing GlcNAc and NeuAc on the cell surface of the bacteria.

The lectin microarray results have produced important insights into the usefulness of this platform. The results support the fact that lectin microarrays can be used to examine the glycosylation of intact bacterial cells. The ability of this technology to promptly evaluate glycosylation empowers researchers to easily detect and monitor dynamic changes in bacterial surface glycans [243].

4.10 Biofilm formation by gut metagenomic fosmid selected clones (FC21 & FC3) with adhesive capability.

The putative adhesive clones FC3 and FC21 exhibited altered surface glycosylation as compared to the control strain EPI300 (pCC1FOS) when interrogated on lectin microarrays. Both fosmid clones exhibited reduced binding to lectins specific for GlcNac (N-Acetylglucosamine). In contrast, EPI300 (pCC1FOS) showed significantly stronger binding signals to lectins specific for bi-antennary (PHA-E) and tri- and tetra antennary (PHA-L) GlcNAc-containing glycans than the fosmid clones. Based on these results, we hypothesized that the clones FC3 & FC21 would exhibit enhanced biofilm forming capability as compared to the control strain. The initial

formation of biofilm involves physicochemical and electrostatic interactions between the surface and the bacterial envelope. Based on the nature of these interactions, the attachment can be transient or permanent[289]. Research has shown that cell surface proteins and polysaccharide adhesins can be critical for specific and non-specific attachment to surfaces and subsequent biofilm formation[266]. These capsular polysaccharides are also used by the bacteria to maintain the structural integrity of the biofilm. Thus, we hypothesized that altered glycosylation on the surface of the bacteria can impact biofilm formation.

Using a simple, biofilm assay performed on a 96 well plate, we assessed the three strains for their biofilm-forming capability. Before the experiment, each strain was grown in the presence of antibiotic (chloramphenicol) and inducer (arabinose) (the same conditions under which changes in surface glycan expression was detected). The biofilm formation was normalized using bacterial cell growth examined by measuring the OD_{600}. Phage T1-resistant EPI300-$T1^R$ cells that contain the empty vector pCC1FOS was used as a baseline control. Figure 4.10 shows the biofilm-forming capability of all three clones. Both FC3 and FC21 formed biofilms that were comparable to the control strain as crystal violet staining showed no significant difference in the biofilm formation between all three strains (Figure 4.10).

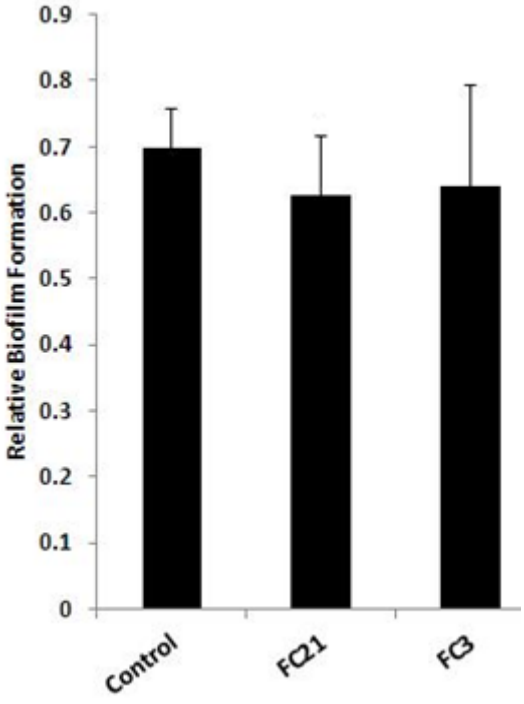


Figure 4.10 Biofilm formation of FC3 and FC21 clones compared to the EPI300 control strain. The fosmid clones were inoculated into 200 ul of LB plus chloramphenicol & arabinose placed in wells of 96-well plates and grown for 48 hours at 37°C. Quantification of crystal violet staining (OD_{540}) normalized to cell growth (OD_{600}) was used to represent each strain's biofilm robustness. Mean standard deviation (SD) are shown by each bar. The experiment was repeated three times with three replicates in each experiment.

In spite of the results observed with the lectin microarray (Figure 4.9A), neither of the strains exhibited enhanced capacity to adhere to an inanimate plastic surface compared to the control. This suggests that the decreased abundance of GlcNac on the surface of FC3 and FC21 does not play a role in biofilm formation in *E. coli*.

4.11 Mucin Microarray results

The lectin microarray results (Figure 4.9A) revealed that both FC3 and FC21 have altered glycosylation patterns on their cell surfaces. It was hypothesized that both clones would exhibit distinct binding patterns and species tropism to the mucins on the mucin microarray when compared to the control. Indeed research performed by Naughton *et al*.[290], showed that despite being closely related, *Campylobacter jejuni* and *Helicobacter pylori* exhibited very different binding patterns and mechanism of interactions with both mucus and mucins on a mucin microarray platform[290]. *C. jejuni* displayed a binding preference for chicken gastrointestinal mucins compared to mucins from other animals and preferentially bound mucins from specific avian intestinal sites. *H. pylori* bound to a number of animal mucins, including porcine stomach mucin.

Despite their different genetic contents (and altered surface glycosylation), the results in this study indicate that all three strains (FC3, FC21 and EP1300 (pCC1FOS)) bind all 35 different mucins samples (Table 2.6) in the same way (Figure 4.11A). They most intensely bound to the multivalent glycoprotein Asialofetuin and mucins from the human colon carcinoma-derived LS174T 204 cells (produce and secrete the mucins, MUC2) and rat duodenum. As seen in Figure 4.11A, fluorescent intensity units is increased for FC3 binding to mucins. However, given the low fluorescent

intensities of the clones and the large standard deviations, there is in fact, no significant difference in the binding of all clones to the glycoproteins and mucins.

A

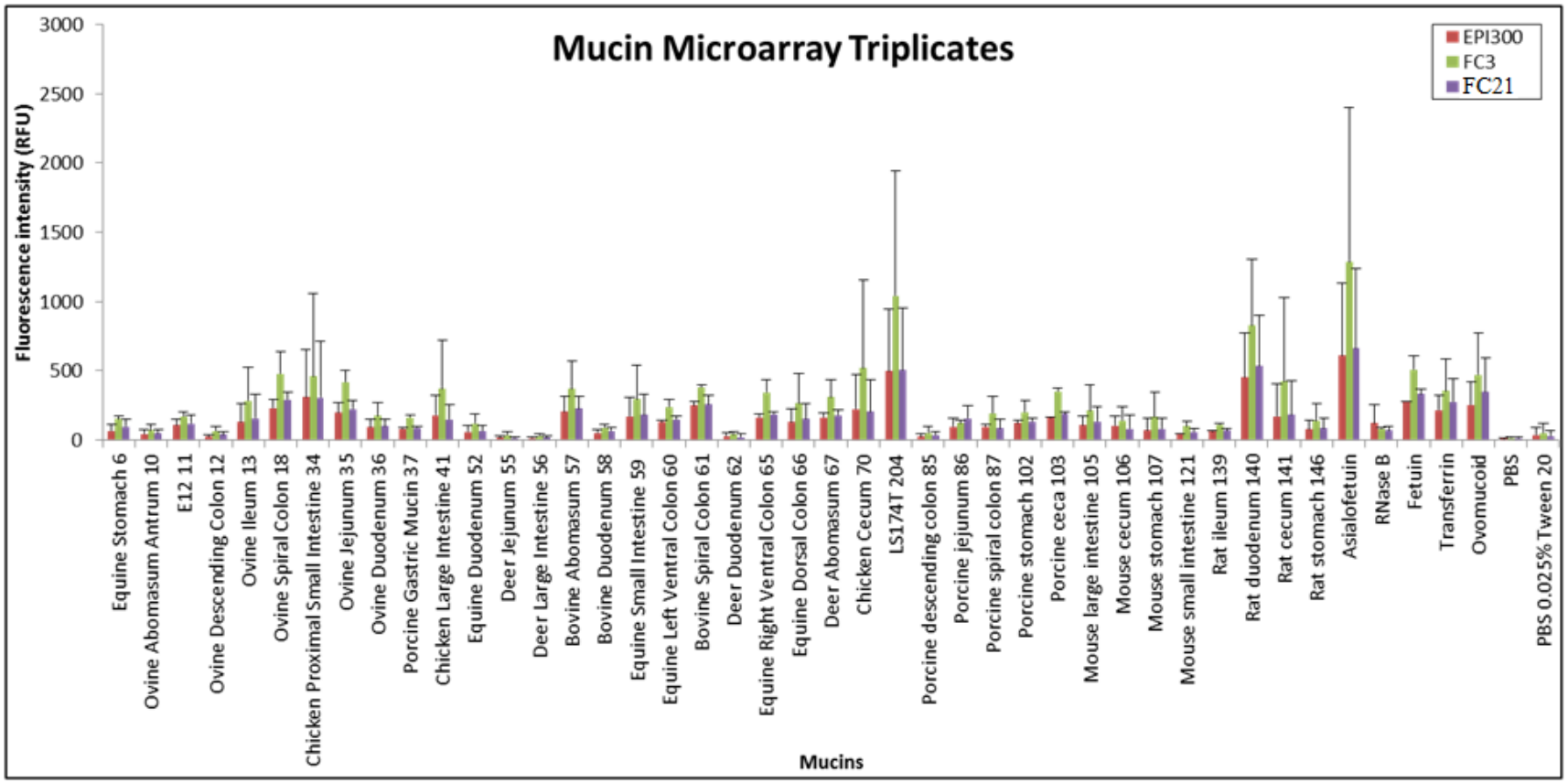
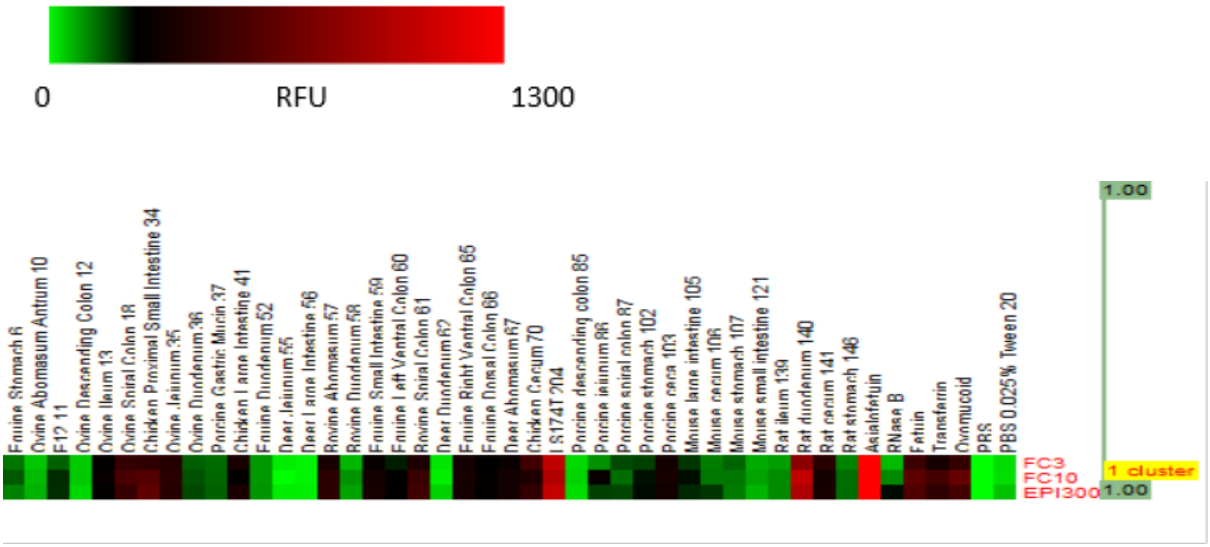


Figure 4.11 **Histogram representing mucin microarray profile of FC3, FC21 and EPI300 control strain.** (A) Histogram representing the mean fluorescence intensity from three replicate microarray slides of individual clones (FC3, FC21 and EPI300 (pCC1FOS)) binding to natural printed mucins. Human derived cell lines are E12 11 and LS174T 204. Error bars indicate the standard deviations of the means for three replicates. RFU, relative fluorescence units. (B) Clustering analysis of mucin triplicate data indicates 100% similarity in mucin binding between FC3, FC10 and EPI300.

B



Clustering analysis of the triplicate data indicated 100% similarity in mucin binding between FC3, FC21 and control strain EPI300 (pCC1FOS) (Figure 4.11B). Overall, the results indicate that there is no change in binding of the two fosmid clones compared to the control. A comparison of the colonization efficiencies with parental HT29-MTX (mucus-secreting) and HT29 (non-mucus-secreting) cells was performed. The next section will describe the interrogation of the fosmid clones (and control) onto neoglycoconjugate microarrays to determine their glycan binding specificities.

4.12 Neoglycoconjugate Microarray Results

In this study, we aimed to investigate the interactions of three whole bacterial strains (FC3, FC21 & control strain EPI300 (pCC1FOS)) with probes on a neoglycoconjugate microarray. All three bacteria strains were fluorescently labelled with SYTO 82 nucleic acid dye. The strains were grown both in the presence and absence of the inducer L-arabinose and the antibiotic marker chloramphenicol. The binding affinity of the control strain (EP1300 (pCC1FOS)) to the NGC is illustrated in Figure 4.14. Figure 4.14 is a histogram representing the mean fluorescent intensity from three replicate microarray slides of the control strain EPI300 (pCC1FOS) (grown in the presence and absence of arabinose inducer & antibiotic) binding to the printed probes. The presence of antibiotic and arabinose promoted a ~0.5 fold increase in the binding affinity of EPI300 (pCC1FOS) to the probes on the microarray. These microarray results are contrasted to the inhibitory effect of arabinose on the adherence of EPI300

(pCC1FOS) to 7 day-old Caco-2 cells observed in Figure 3.13. One of the reasons predicted for the observation in Figure 3.13 is that arabinose molecules bind to the bacterial cell surface ligands, thereby competitively blocking attachment of adhesins to epithelial cell receptors. Another hypothesis was that a high-copy number of the vector induced in the EPI300 host is deleterious to the cell and thus negatively impacts adherence. However, according to Figure 4.14, induction of EPI300 (pCC1FOS) with arabinose promotes neo-glycoconjugate binding. A possible reason for the observed difference in binding of EPI300 (pCC1FOS) to 7 day-old Caco-2 cells and neoglycoconjugates on a microarray is the way in which the glycans are presented in both assays. Strong evidence exists that the presentation of glycans within different formats can have a profound effect on the apparent-binding affinity and specificity of bacterial cells on the array[291]. It is not surprising that on the surface of a microarray, how a glycan is presented to the bacterial cell, the relative density, the local glycan packing and even the linker (which can affect packing) can alter what is observed. In this study, the neoglycoconjugate consisted of 40 glycoproteins and neoglycoconjugates (Table 2.4). To obtain the most accurate preliminary identification of bacterial specificity requires an array that covers a significant portion of the glycome. Expanding microarrays to cover the mammalian glycome is an enormous undertaking.

The presence of antibiotic and arabinose promoted a significant increase in the binding affinity of FC3 to the probes on the neoglycoconjugate microarray (Figure 4.12). These microarray results are similar to the effect of arabinose on the adherence efficiency of FC3 on 7 day-old Caco-2 cells (Figure 3.15, Day 2 & 3). These results suggest that the expression of the adherence factor encoded by the FC3 DNA insert is increased after induction by arabinose, therby increasing adherence to epithelial cells and neoglycoconjugates. Contrary to the observed results depicted in Figure 3.16, the addition of arabinose and antibiotic does not promote the increased adherence of FC21 with the probes on the microarray (Figure 4.13) However, without arabinose or antibiotic, fluorescence intensity is increased for both FC3 and FC21 compared to the control (Figure 4.16).

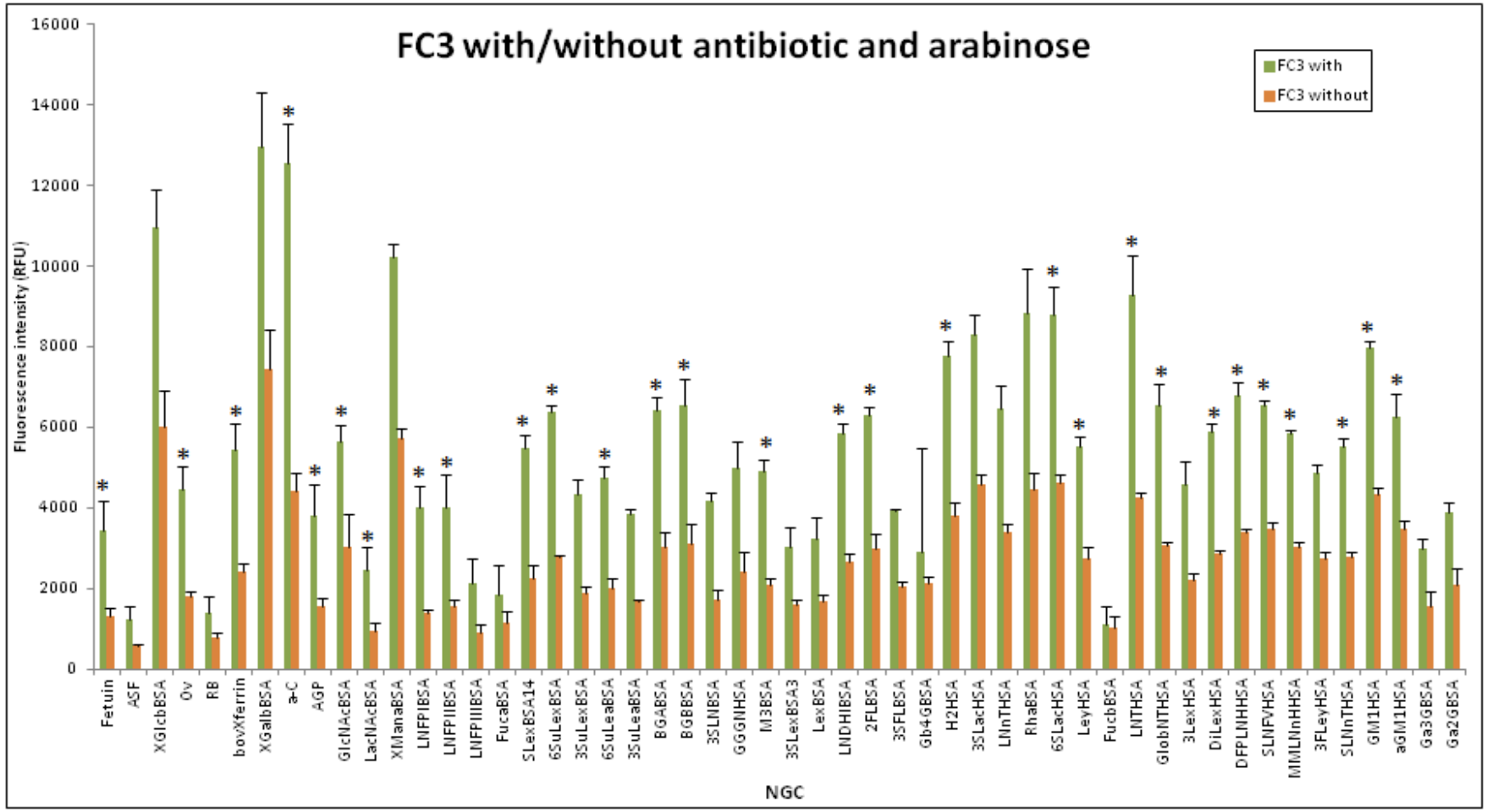


Figure 4.12 Histogram of neoglycoconjugate microarray profile of FC3 in the presence and absence of arabinose and antibiotic. Histogram representing the mean fluorescence intensity from three replicate microarray slides of FC3 (grown in the presence & absence of antibiotic and arabinose) binding to printed probes (glycoproteins and neoglycoconjugates). Significance was determined using Student's t-test, *, $P<0.05$; **, $P<0.01$.

The control strain EPI300 (pCC1FOS) shows a 0.99 fold-change in its binding affinity to XGlcbBSA (Glc-b-4AP-BSA) in the presence of arabinose (Figure 4.14). The presence of arabinose elicited a small increase of approximately 0.5 fold in the affinity of control strain EPI300 (pCC1FOS) to bind to most of the probes, especially Fetuin, ASF (Asialofetuin), Ov (Ovalbumin) and a-C (a-Crystallin from bovine lens). Unlike the control strain EPI300 (pCC1FOS), the presence of arabinose elicited negative fold changes in the binding affinity of fosmid clone FC21 to several of the probes on the microarray. For example, FC21 shows a fold change of ~ -0.21 in its binding affinity to LNFPIIIBSA (Lacto-N-fucopentaose-III-BSA) and ~ -0.28 to 3LexHSA (Tri-Lex-APE-HSA) in the presence of arabinose. These results indicate that the presence of L-arabinose can slightly inhibit the binding affinity of the FC21 clone to several glycoproteins and NGC.

In contrast to the control strain EPI300 (pCC1FOS) and FC21, fosmid clone FC3 showed an increase in binding affinity to the NGC probes in the presence of arabinose (Figure 4.12). For example, FC3 shows a 1.64 fold increase in its binding affinity to bovine fetuin in the presence of L-arabinose. FC3 exhibited binding specificity for (1) ovalbumin (Ov), (2) bovine transferrin (bovXferrin), (3) α-Crystallin (a-C) from bovine lens, (4) GlcNAc (GlcNAcBSA), (5) Lacto-*N*-fucopentaose I and (6) II (LNFPIBSA and LNFPIIBSA), (7) 3'Sialyl Lewis x-BSA (SLexBSA14), (8) 6-Sulfo Lewis x-BSA (6SuLexBSA), (9) 3-Sulfo Lewis x-BSA (3SuLexBSA), (10) Galα1,3Galβ1,4GlcNAc-HSA (GGGNHSA), (11) Manα1,3(Manα1,6)Man-BSA (M3BSA), (12) Tri-fucosyl-Ley-heptasaccharide-APE-HSA (3FLeyHSA), (13) Tri-Lex-APE-HSA (3LexHSA) and (14) 2'Fucosyllactose-BSA (3SFLBSA) (Figure 4.12)

The binding specificity of FC3 and FC21 to GlcNAcBSA (N-acetylglucosamine) probe is biologically relevant because it suggests that Caco-2 cells possibly contains GlcNAc residues that serve as epitopes for glycan-binding interactions with microbial adhesins. GlcNAc is well-known for supporting the human body's creation of a healthy mucus layer in the gut. Studies have demonstrated that GlcNAc helps support the growth of beneficial gut bacteria like *Bifidobacterium bifidum*. *N*-acetyl-glucosamine containing oligosaccharides were first identified 50 years ago as the

‘bifidus factor’, a selective growth substrate for intestinal bifidobacteria. Further studies demonstrate that GlcNAc may improve immune function in patients with multiple sclerosis. While N-acetylglucosamine might benfit anyone with digestive problems, it looks to be promising for people suffering from inflammatory bowel disease. Patients with conditions like Crohn’s disease and Ulceritive colitis have much thinner mucus barrier in the gastrointestinal tract. In recent study by Andy Zhu and colleagues (April 2015), patients with inflammatory bowel disease taking N-acetylglucosamine for 1 month had substantial improvement in their symptoms.

Moreover, FC3 exhibits binding specificity to the human milk oligosaccharide 2’Fucosyllactose. 2’Fucosyllactose is the most prevalent human milk oligosaccharide, making up 30% of all HMOs. Humans are unable to digest HMOs such as 2’Fucosyllactose, hence the majority of HMO’s reach the gut, where they serve as food for desirable gut bacteria.

FC3 showed low binding affinity to α-linked and β-linked Fuc with the aminophenyl linker in both types of growth conditions (Figure 4.12). Interestingly, FC3 showed low binding affinity to Asialofetuin (ASF) the multivalent glycoprotein that possesses nine LacNAc epitopes. Likewise, FC3 showed low binding affinity for the high mannose glycoprotein, RNAseb. Although FC3 binds minimally to Fetuin, there is a significant difference between the binding intensity of FC3 grown in inducer/antibiotic versus FC3 grown without inducer/antibiotic.

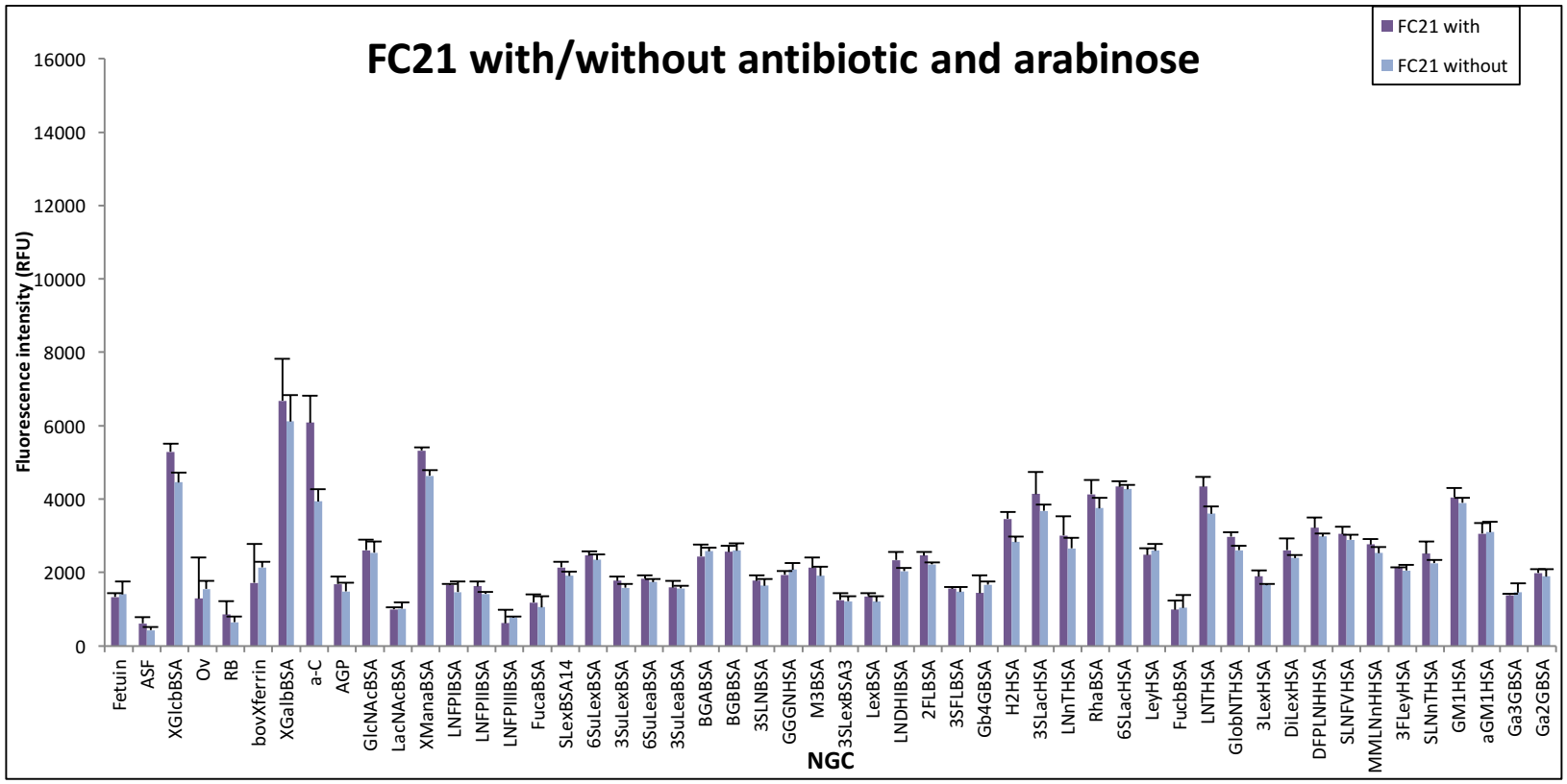


Figure 4.13 **Histogram representing the neo-glycoconjugate microarray profile of FC21 in the absence and presence of arabinose and antibiotic**. Histogram representing the mean fluorescence intensity from three replicate microarray slides of FC21 (grown in the presence & absence of chloramphenicol and arabinose) binding to printed neoglycoconjugates.

FC3 bound with very good intensity to β-Glc attached via a lithothiocyanate linker to the BSA backbone. Figure 4.12 illustrates very similar binding intensities of FC3 with Gal-β-ITC-BSA, Man- α-ITC-BSA and α-crystallin from bovine lens. So overall, FC3 showed a high binding affinity to the three monosaccharide NGC analogues (α-Man, β-Gal, and β-Glc) attached via a lithothyocyanate linker to BSA (Figure 4.12). In fact, the three monosaccharide NGC analogues (α-Man, β-Gal and β-Glc) printed with the ITC linker displayed high binding intensity than other neoglycoconjugates with FC21 (Figure 4.13) and EPI300 control strains (Figure 4.14)

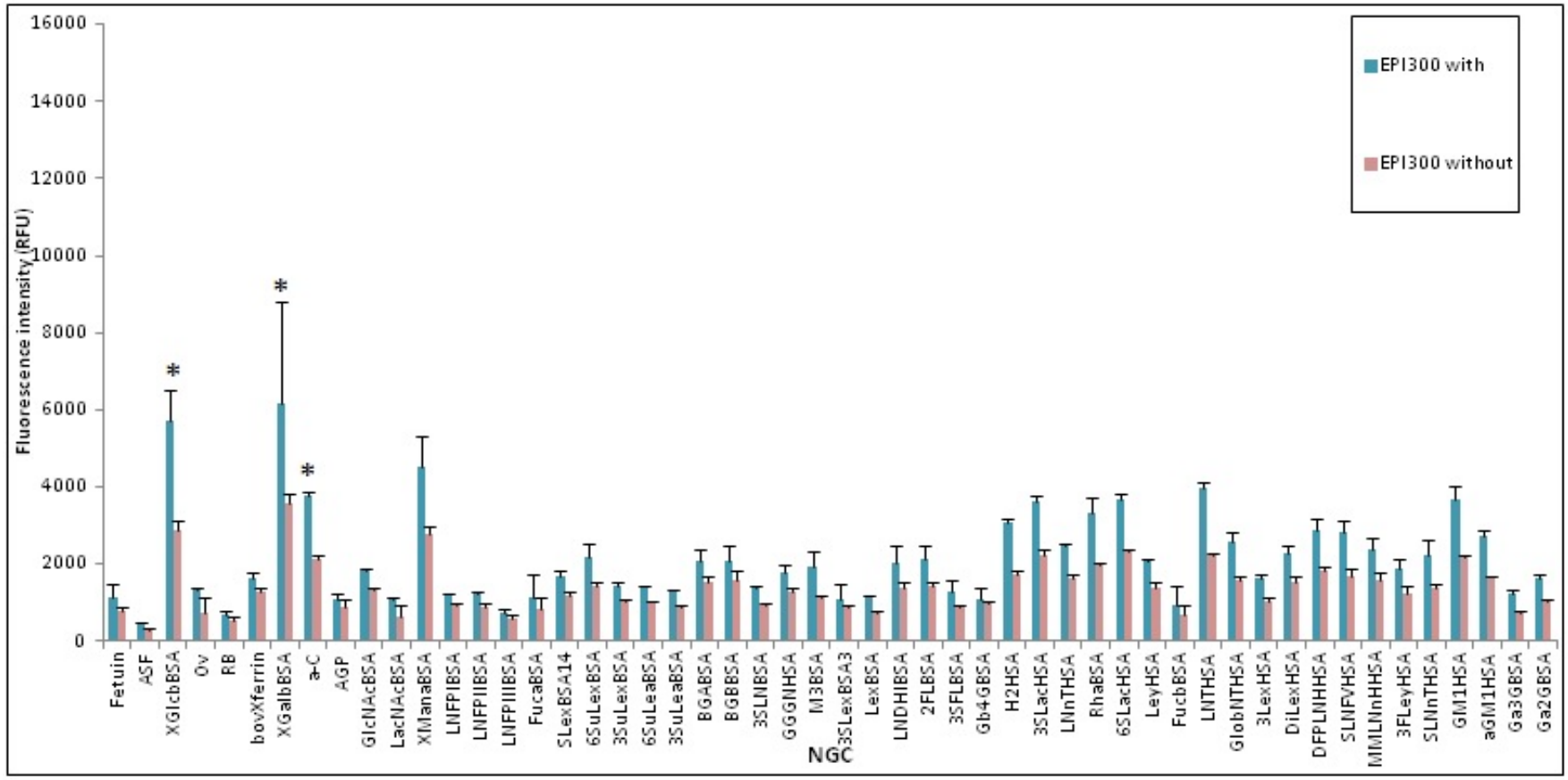


Figure 4.14 Neo-glycoconjugate microarray profile of EP1300 (pCC1FOS) control strain in presence and absence of arabinose and antibiotic. Histogram representing the mean fluorescence intensity from three replicate microarray slides of control strain EPI300 (grown in the presence and absence of antibiotic and arabinose) binding to printed neoglycoconjugates.

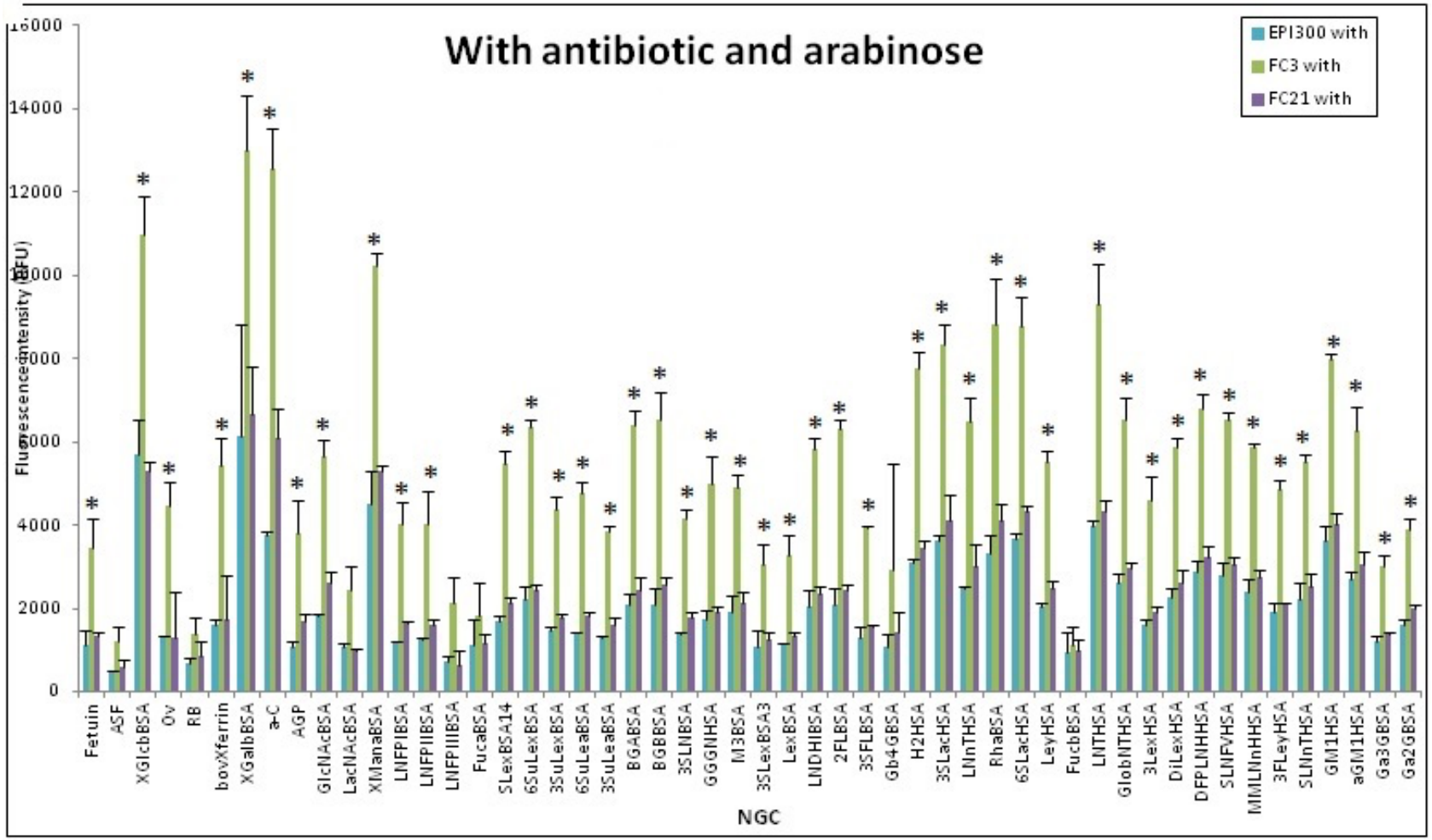


Figure 4.15 Neo-glycoconjugate microarray profile of FC3, FC21 and EPI300 control strain in presence of antibitic and arabinose. Histogram representing the mean fluorescence intensity from three replicate microarray slides of FC3, FC21 and EPI300 (grown in the presence of antibiotic and arabinose) binding to printed neo-glycoconjugates.

These microarray results indicate that without antibiotic or arabinose, fluorescence intensity is increased for clones FC3 and FC21 as compared to the control strain EPI300 (Figure 4.16). HCE clustering (Figure 4.17) revealed a 69% similarity in NGC binding of the three strains in the absence of antibiotic and arabinose. In the presence of antibiotic and arabinose, fluorescence intensity is primarily increased for clone FC3 (Figure 4.15). The presence of antibiotic and arabinose (induction and stabilisation of plasmid) promotes NGC binding for FC3. HCE clustering analysis shows 15% similarity in NGC binding for FC3 with FC10 and EPI300.

With arabinose and antibiotic present, FC3 demonstrated the most differences in NGC binding compared to the control EPI300 (pCC1FOS). Specifically, FC3 binds greater to 14 glycoconjugate (Figure 4.17); namely bovine fetuin, ovalbumin (Ov), bovine transferrin (bovXferrin), alpha-Crystallin (a-C) from bovine lens, GlcNAc (GlcNAcBSA), Lacto-*N*-fucopentaose I and II (LNFPIBSA and LNFPIIBSA), 3'Sialyl Lewis x-BSA (SLexBSA14), 6-Sulfo Lewis x-BSA (6SuLexBSA), 3-Sulfo Lewis x-BSA (3SuLexBSA), Gala1,3Galb1,4GlcNAc-HSA (GGGNHSA) Mana1,3(Mana1,6)Man-BSA (M3BSA), Tri-fucosyl-Ley-heptasaccharide-APE-HSA (3FLeyHSA), Tri-Lex-APE-HSA (3LexHSA) and 2'Fucosyllactose-BSA (3SFLBSA). (See red arrows Figure 4.17).

Deciphering the glycan specificity of FC3 provided insight into the types of glycoproteins and glycan epiotpes that may be present on the cell surface of 3 week-old Caco-2 cells which bind FC3. A good way to confirm the presence of specific glycoproteins or NGC on the surface of Caco-2 cells is to interrogate them on lectins with known glycan binding specificities.

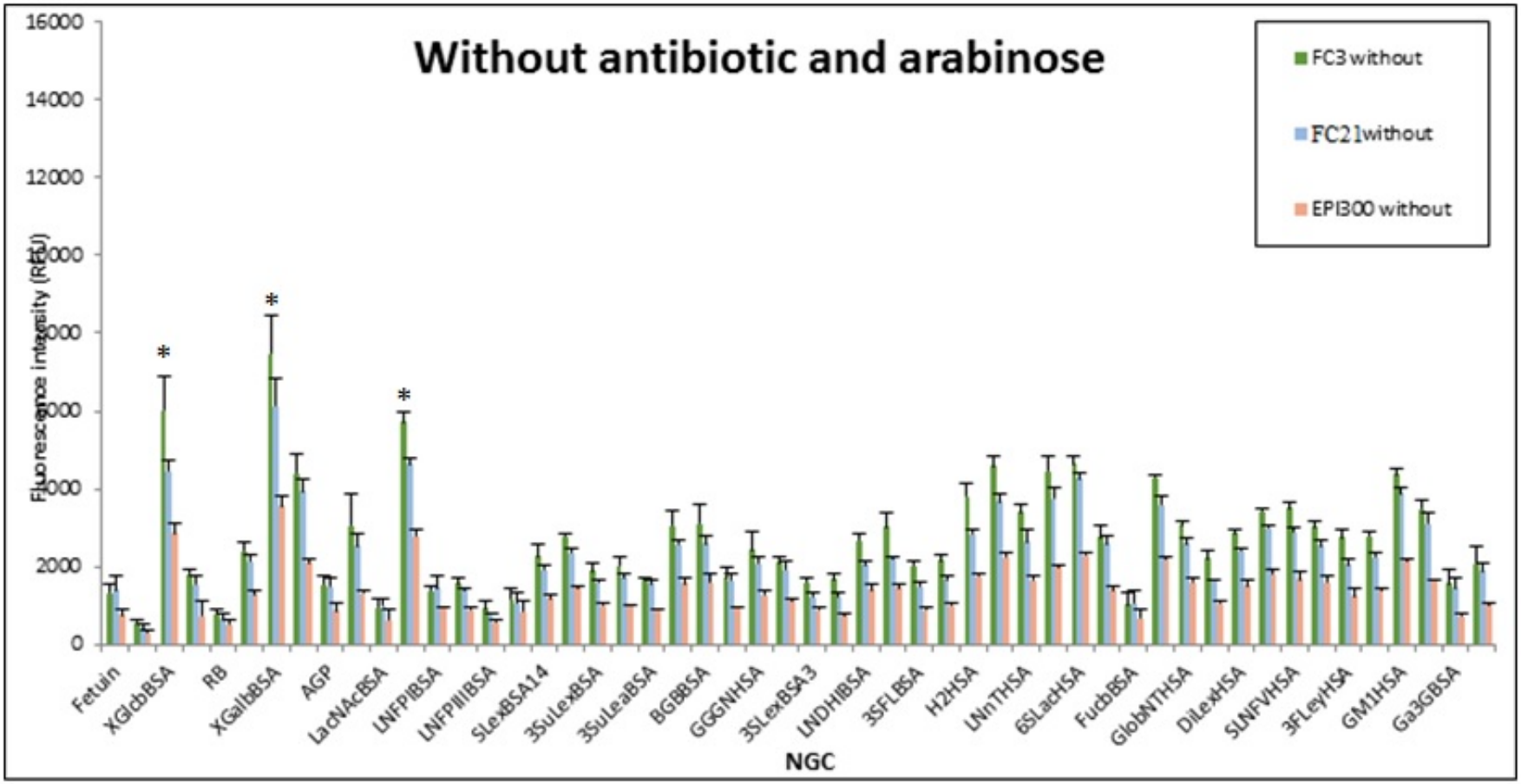


Figure 4.16 Histogram representing the differences in recognition of neoglycoconjugates and glycoproteins by fluorescently-labelled bacterial strains FC3, EPI300 and FC21 (grown in the absence of antibiotic and arabinose). The histogram represents the average of three replicate experiments and error bars are 1 standard deviation.

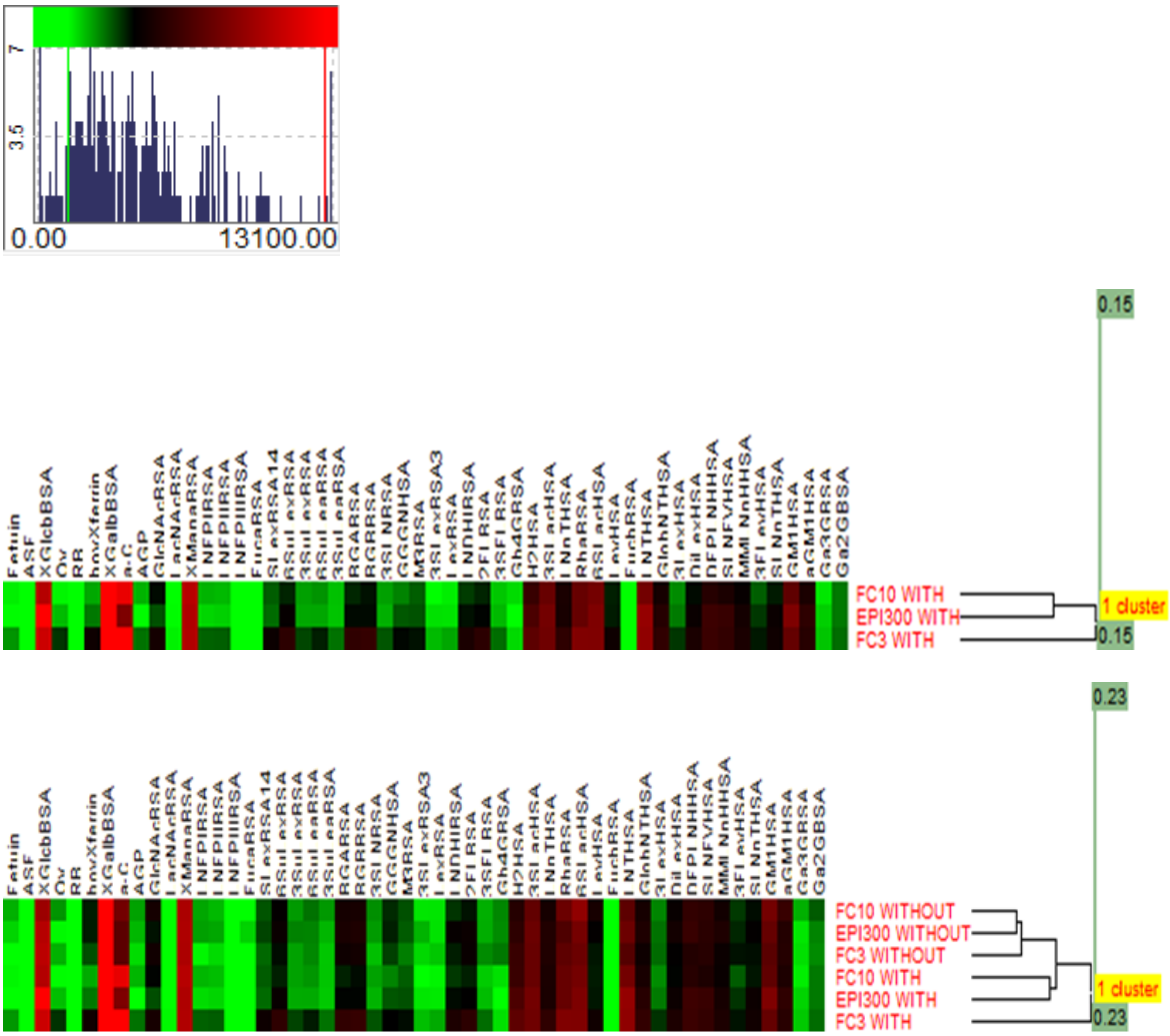


Figure 4.17 **Clustering analysis of neoglycoconjugate triplicate data**. The red arrows indicate the binding specificity of FC3.

In this study, whole bacterial cells (FC3, FC21 & EPI300 control) were interrogated on a neoglycoconjugate microarray to profile their carbohydrate binding specificities. Overall, our results show that FC3 demonstrated the highest binding affinity to the neoglycoconjugates in the presence and absence of antibiotic and inducer compared to FC21 and control strain EPI300 (pCC1FOS). In fact, the presence of antibiotic and inducer seem to promote neoglycoconjugate binding of FC3. In the absence of antibiotic and inducer, both FC3 and FC21 have a higher binding affinity to the neoglycoconjugates as compared to the control strain. Clustering analysis reveals a 69% similarity in binding for the two fosmid clones (FC3 & FC21) as compared to the control.

Overall, neoglycoconjugate-glycoprotein microarrays are useful strategies to characterize the glycan binding specificities of bacterial strains. In this work, we successfully characterized the glycan binding profile of two putative adhesive clones interrogated on NGC microarrays. Successfully characterizing the glycan binding specificities of FC3 and FC21 provided a glimpse of the type of glycans they bind on Caco-2 cells as well as an idea of the glycan landscape of the Caco-2 cells.

4.13 Discussion

Little is known of the molecular and structural mechanisms that mediate the colonization and persistence of the microbiota in the gut. Enquiry into microbial adhesion in the gut is arduous not only because gut microbes are mostly unculturable but also because of the absence of effective methods to preserve the intestinal mucus layer, where microbial communities are formed. Functional metagenomic screens, followed by Next Generation Sequencing and bioinformatics analysis, elucidated two putative clones (FC3 & FC21) associated with enhanced adherence to 3 week-old Caco-2 cells when grown in the presence of antibiotic and L-arabinose. Bioinformatic analysis of the complete insert sequence of FC3 and FC21 revealed that FC3 and FC21 clones are 24.6 kb and 8.1 kb and include 26 and 3 protein-coding open reading frames (ORFs), respectively. The DNA insert from FC21 was shown to have a 100% similarity to the gut inhabitant *Bifidobacterium adolescentis* ATCC 1507. The three functional genes present on the FC21 insert all play critical roles in transport. As shown in Figure 4.5, the three functional genes present on the FC21 DNA fragment are located on the cell membrane (Table 4.11). AroP is a transport protein that takes up tryptophan, phenylalanine and tyrosine. Consistent with its membrane location AroP is highly hydrophobic and apparently comprises two equivalent domains, each composed of six alpha-helical segments with membrane spanning potential[292]. AroP may be responsible for the enhanced adherence of FC21 as compared to the control strain because several studies have demonstrated the involvement of transport proteins in adherence[293]. For example the PEB1a protein of the gastrointestinal pathogen *Campylobacter jejuni* is known to play a key role in transport and is an important factor in host colonization[294]. PEB1a is homologous with periplasmic-binding proteins associated with ABC transporters and is able to bind L-aspartate and L-glutamate[295]. Therefore, in addition to its established role as an adhesin, the PEB1 protein also plays

a key role in the transport and utilization of aspartate and glutamate[294]. Merino and colleagues[296] demonstrated that the *mgtE* gene, which encodes a Mg^{2+} transport protein, is involved in the adherence of the water-borne bacteria *Aeromonas hydrophila*. *Aeromonas hydrophila* strains carrying mutations in *mgtE* showed a 50% reduction of *in vitro* adherence to HEp-2 cells and a decrease in biofilm formation of over 60% in comparison to the wild-type strain[296]. Similarly, a *glnQ* mutant of group B streptococcus showed decreased adherence to respiratory epithelial cells *in vitro* and decreased virulence *in vivo*, indicating the importance of the glutamine transporter in group B streptococcus virulence[297]. In *Agrobacterium tumefaciens*, a mutant with a transposon mutation in an operon showing homology to operons encoding ABC transporters was shown to be deficient in attachment to host cells[298]. Henrich and colleagues identified an adherence-associated lipoprotein P100 of *Mycoplasma hominis* and demonstrated that the P100 gene is organized with an operon structure containing genes putatively encoding the core domains of an ABC transporter[293].

The hypothetical protein, BAD_0085 of FC21 is also predicted to be an ABC transporter due to the types of conserved domains and motifs present in the gene. For example, BAD_0085 contains an FtsX domain. FtsX is an integral membrane protein that is often found in members of the ABC transporter family and is involved in cell division[299]. It is encoded in the same operon as signal recognition particle docking protein FtsY and FtsE. The precise function of FtsX is not yet known, but it is involved in sporulation[299]. BAD_0085 also contains a MacB-like periplasmic core domain found in a variety of ABC transporters.

Sodium-proton antiporters (Na^+/H^+) are ATP-independent membrane glycoprotein transporters that are involved in the regulation of intracellular pH, cell volume and the cellular response to hormones. All living cells maintain a sodium concentration gradient directed inward and a constant intracellular pH at around neutrality. Hence, all cells have Na^+ extrusion system(s) and homeostatic mechanisms controlling the proton circulation across the cytoplasmic membrane[300]. Reports implicating sodium-hydrogen antiporters in adherence are less forthcoming and suggest an indirect effect of this gene in adherence.

However, it is possible that the sodium-proton antiporter may be responsible for the altered cell surface glycosylation of FC3 and FC21 observed in Figure 4.9A. Studies

have demonstrated that the sodium-hydrogen antiporter exhibits a structural role in regulating the cortical cytoskeleton that is independent of its function as an ion exchanger. Denker and colleagues demonstrated that a sodium-hydrogen antiporter 1 directly binds to actin binding proteins (ERM) and regulates focal adhesion assembly, cell shape determination, and cortical cytoskeleton organization[301]. They discovered a novel role for sodium-proton antiporter 1 as an anchor for actin filaments that is mediated by actin binding proteins, and they suggested that this anchoring contributes to the organization of cortical actin filaments and the determination of cell shape[301]. Based on the results of these studies, it is possible that the Na^+/H^+ antiporter gene of FC21 may be involved in the alteration of cell surface glycosylation.

Overall, it is probable that *aroP*, the hypothetical protein and the Na^+/H^+ antiporter of FC21 likely form a complex that is a functional transporter and "putative" adhesin. Studies have demonstrated that numerous transporters are multimetic complexes[302,303]. FC3 was found to originate from an unknown Clostridium spp. species. A large portion of the FC3 predicted gene products were highly homologous to those of Clostridium spp. Predicted annotations of the 26 ORFs of FC3 (Table 4.12) revealed three membrane located hypothetical proteins (Table 4.12; gene5, gene6, gene 23) with no known domains or functions. It is possible that one or all three hypothetical proteins are adhesive factors that confer FC3 with the capacity to bind 3 week-old Caco-2 cells. Further functional metagenomic analysis on each individual hypothetical protein would provide some insight into their function. A possible strategy to characterize the function of these hypothetical proteins is by sub-cloning each gene into a shuttle vector and expressing the gene in an appropriate expression host for functional screens.

A prominent gene of FC3 that may be responsible for the enhanced adherence to Caco-2 cells is the putative collagen adhesion protein, gene 17 (Table 4.12). This gene is predicted to contain a Cna_B domain. Cna_B is a repeated B region domain that is found in *Staphylococcus aureus* collagen-binding surface protein Cna. The primary sequence of Cna has a non-repetitive collagen-binding A region, followed by repetitive B region. The B region contains one to four 23 kDa repeat units (B_1-B_4), depending on the strain of origin[276]. The affinity of the A region for collagen is independent of the B region. However, the B region assembly has been suggested to effectively provide the needed flexibility and stability of a "stalk" that projects the A region from the bacterial surface and thus facilitate bacterial adherence to collagen.

Indeed, the $B_1B_2B_3B_4$ repeats are packed in a zig-zag fashion, like an accordion; they might stretch and contract from the bacterial cell wall and thus aid in the projection of the A region away from the cell surface[276]. It is possible that in the event of proteolytic loss of a Cna A region resulting from a specific cleavage by host's extracellular defensive apparatus, the B-repeat units could be pressed into an adhesive process or other function necessary for bacterial survival. Moreover, the repeated number of these domains might reflect the added stability the bacteria would achieve by multiple anchoring[276].

The two predicted sortase genes (Sortase B and Sortase D) present on the FC3 insert are regarded as accessory sortases that either anchor their specific substrates to the bacterial cell wall or assemble cell surface pili[304]. Studies have revealed that many genes coding for accessory sortases are found in the same gene operons that encode their substrate proteins[305]. Therefore, it is likely that the substrate proteins of sortase B and sortase D are also encoded on the FC3 insert. A quick method to determine the importance of both sortase enzymes is by disrupting either the *srtB* and/or the *srtD* to determine if binding of FC3 to caco-2 cells is significantly reduced.

To characterize the molecular mechanisms utilized by FC3 and FC21 to adhere to host cells, three carbohydrate-based microarrays platforms were used. (i) carbohydrate-binding protein (lectins) microarray which were profiled with whole bacteria to determine glycosylation patterns on the cell surface of the bacteria (ii) natural mucin microarrays which were profiled with whole bacteria to determine mucin glycosylation and bacterial binding tropisms, (iii) and finally neoglycoconjugate (NGC) microarrays profiled with whole bacteria to determine the binding affinity of bacteria to specific neoglycoconjugates (NGC).

When interrogated on lectin microarrays, both fosmid clones (FC3 and FC21) exhibited reduced binding to plant lectins that are specific for GlcNAc (N-Acetylglucosamine) compared to the control EPI300 (pCC1FOS). GlcNAc is an interesting molecule because it is known to play important roles in both cell structure and cell signalling[306]. It is a key component of the bacterial cell wall peptidoglycan and therefore a reduced abundance of GlcNAc on the bacterial cell surface will affect the cell surface glycosylation. The reduced abundance of GlcNAc on the bacterial cell surface may increase the bacteria capacity to adhere to glycan receptors on host cells

(Caco-2 cell) by increasing the accessibility of protein adhesins that extend from the cell surface to bind to receptors on host cells. The physical hindrance of large GlcNAc residues on the surface of FC3 and FC21 is decreased thereby potentially increasing accessibility of cell surface adhesins to bind to glycan epitopes on Caco-2 cells.

Clearly, the genetic information encoded in the metagenomic inserts of FC3 and FC21 altered their cell surface glycosylation. This is not surprising as studies have indicated that genes present on a DNA can exert control over the assembly of macromolecular complexes on the cell surface. The use of lectin microarrays to examine the dynamic changes to *E. coli* bacterial glycosylation is not a new phenomenon. Studies indicate that lectin microarrays have been used to distinguish *E. coli* strains based on glycosylation.

It is possible that the presence of genes carried on the insert of the fosmid vector is sufficient to trigger altered surface glycosylation. Not only does the altered cell surface seem to increase adherence by exposing adhesins, some studies have demonstrated that in certain instances, the presence of certain cell surface O-glycans can limit the adherence of certain bacterial species. For example, Ricciuto and colleagues demonstrated that the mucin-rich environment of the intact corneal epithelium contributes to the prevention of *Staphylococcus aureus* infection[307]. Removal of mucin-O-glycosylation using the chemical primer benzyl-alpha-GalNAc resulted in increased adherence of parental strain RN6390 to apical human corneal-limbal epithelial cells and to biotinylated cell surface protein in static and liquid phase adhesion assays. Their results suggests that alteration of cell surface glycosylation from disease or trauma could contribute to higher risk of infection[307]. Likewise, the altered cell surface glycosylation of FC3 and FC21 led to increased adherence of these clones to 3 week-old Caco-2 cells.

The lectin-binding signals observed in Figure 4.9A may not always reflect the glycan composition of the bacterial cell wall. Some strains may simply be able to bind many different types of lectins and others may show limited lectin binding. There is the possibility that physical hindrance of the glycan binding sites on the bacteria may limit binding of the lectin. However, it is worth noting that lectin binding profiles reflect, in part, the content and structure of the bacterial cell wall polysaccharides. An effective way to have validated the specificity of the observed lectin (Figure 4.9) and NGC

(Figure 4.12-4.16) interactions on the microarray would be to perform carbohydrate inhibition assays.

As mentioned in section 4.9, FC3, FC21 and EP1300 (pCC1FOS) exhibited very strong signals in binding to a sialic acid animal lectin (CCA) with high specificity for O-acetyl sialic acids. These results suggest that all three strains contain relatively high amounts of O-acetyl sialic acids on their cell surface glycoconjugates. Studies have demonstrated that sialic acid contributes significantly to the structural properties of molecules, both in solutions and on the cell surfaces[86]. Sialic acid is an important regulator of molecular and cellular interactions; where it plays a dual role by masking recognition or serving as a recognition determinant. Studies have shown that several cell surface proteins that have been involved in bacterial adherence, such as integrins, fibronectin and plasminogen, are sialylated. Recent studies by Tong *et al*. (2002), demonstrated that modulation of sialic acid on host cell surface can reduce the adherence of the bacteria. The numerous data on sialic acid and its role in bacterial adherence suggest that sialic acid may be a potential therapeutic target molecule in bacterial adherence.

As a result of the altered cell surface glycosylation of FC3 and FC21 (Figure 4.9A), we hypothesized that FC3 and FC21 would produce distinct biofilm forming capability when compared to the control EPI300 (pCC1FOS) strain. As observed in Figure 4.10, FC3 and FC21 did not produce significantly different levels of biofilms than the control. In spite of the altered cell surface glycosylation of FC3 and FC21, these clones did not produce higher levels of biofilm forming capacity than the control strain. These results may suggest that, in spite of the altered cell surface glycosylation of FC3 and FC21, all three strains favour specific binding to biotic surfaces as opposed to non-specific binding to abiotic surfaces. Bacteria use different mechanisms to form biofilms (abiotic surface) and to adhere to glycan host epithelial cells (biotic surface). The low levels of biofilm forming capacity of FC3 and FC21 to abiotic surfaces suggests that these clones contain cell surface proteins that specifically bind to distinct glycan epitopes on Caco-2 cells.

One of the aims of this study was to interrogate a natural mucin microarray with putative fosmid adhesive clones (FC3 & FC21) to determine their differences in binding to mucin as compared to the control strain EPI300 (pCC1FOS). Results

indicate that the mucin binding pattern is not altered for the two fosmid clones as compared to the control (Figure 4.11A). Clustering analysis of the triplicate data indicates 100% similarity in mucin binding between FC3, FC21 and EPI300 (pCC1FOS) (Figure 4.11B). Overall, the results indicate that there is no novel binding of the two fosmid clones compared to the control. Conversely, when FC3 and FC21 were interrogated onto microarray consisting of glycoproteins and NGC probes, the fluorescence intensity increased for FC3 and FC21 compared to the control strain EPI300 (pCC1FOS) (Figure 4.15). HCE clustering revealed a 69% similarity in NGC binding of FC3 and FC21 in the absence of antibiotic or arabinose (Figure 4.17). In contrast, with arabinose and antibiotic the fluorescence intensity is primarily increased for clone FC3. HCE clustering analysis showed a 15% similarity in NGC binding for FC3 and FC21 and control EPI300 (pCC1FOS). Overall, the results indicated that the presence of antibiotic and arabinose (induction and stabilisation of the fosmid) promotes NGC binding of FC3 (Figure 4.12).

Possible future approaches to further characterize these fosmid clones should include the combinatorial use of mucus secreting cells and mucin microarrays. Naughton *et al*[290] used methotrexate adapted cell line HT29-MTX which secretes mucins into culture and HT29-MTX-E12 cells which form an adherent mucus layer to assess the effect of mucus and mucins on the interaction of *Campylobacter jejuni* and *Helicobacter pylori*. Indeed both the fosmid clones (FC3 & FC21) used in this study were selected on cultured Caco-2 cells. Caco-2 cells are human adenocarcinoma cells that mimic the epithelial cells of the intestine. However, Caco-2 cells do not produce a mucus layer overlying the cells[413]. As a result, the cells do not exactly mimic the conditions *in vivo*. It is this recognition that intestinal cell lines often used for such studies that prompted the development of gut-derived epithelial cells that secrete mucins into the supernatant[415]. These cells harbor an overlying adherent mucus layer and thus (more accurately mimic *in vivo* conditions)[408].

In the future, addition of a much larger variety of human mucin samples or natural mucin microarrays should be considered. As mentioned, only two human cell lines were presented on our mucin microarrays (Table 2.6 E12 & LS174T). The natural mucin microarray has the potential to uncover novel biologically relevant motifs in bacteria-host interactions and accompany the use of traditional *and* novel *in vitro* cell models, such as the mucus secreting cell lines. It is an effective glyco-profiling and

discovery tool flexible enough to suit a biological question under investigation. A further test on the lectin microarray platform would be the glycan profiling of extracts of Caco-2 epithelial cells. Angeloni and colleagues[179] immobilized Caco-2 cell extracts (non-differentiated Caco-2 cells, 7 days post seeding and differentiated 21 day old Caco-2 cells) and interrogated them with a series of plant lectins. Their results indicate that the cell glycosylation phenotype changes with increasing culture time or differentiated status, respectively. They discovered that alpha-2,3-linked sialic acid epitopes reduced from 7 days to 21 days in culture[308]. Studies like these help to explain the different adhesive profiles observed by our fosmid clones at varying Caco-2 differentiation stages (7 days vs 3 weeks). These changes can be attributed to the change in cell surface glycosylation of the Caco-2 cells with increasing culture time and differentiation status. The exact epitopes present on the Caco-2 cell surface that are recognized by FC3 and FC21 are yet to be fully elucidated. Studies that measure changes in the glycan landscape of Caco-2 cell membrane with varying bacterial infection have been performed. Using high resolution mass spectrometric techniques, Park and colleagues[309] performed a glycomic analysis to characterize glycosylation changes on epithelial cell surfaces upon prolonged contact with foreign and resident bacteria of the gut. Their results indicated that Caco-2 cell surface glycosylation is dominated by sialylated and fucosylated complex and hybrid glycans. When considering the relative intensities, however, high mannose glycans were among the most abundant. These results suggest that Caco-2 cell membranes have a large amount of terminal mannose residues, which may have functional significance in epithelial cells during infection[309]. During the course of infection, levels of bisecting and tri-antennary complex glycans were significantly altered. The most abundant glycan in the uninfected sample, a bisecting, monofucosylated, bisialylated complex glycan, decreased dramatically in signal post-infection, becoming suppressed by other high abundant glycans. An isomer of this glycan, which eluted at a later time, increased in abundance fifteen fold after infection. Terminal fucose and sialic acid residues on a glycan with more than two antennas may act as receptors for bacteria and be utilized as a source of energy, carbon, and nitrogen. Deficient glycan degrading enzyme activity of the bacteria led to an accumulation of certain oligosaccharide substrates on the cell surface. On average, 176 glycan compositions were identified in the uninfected sample and 166 compositions for the infected sample[309].

The advent of Lectin Microarray technology has generated numerous benefits in the study of microbe-host interactions. Few other techniques are able to study the large diversity of carbohydrate structures present on intact bacterial cell surfaces in a high-throughput fashion. In spite of its many attributes, the lectin microarray has several limitations. The technique is not quantitative neither does it allow for the determination of glycan structures like Mass Spectrometry [243]. That is, lectin microarrays do not accurately identify glycan structures, but rather obtain information about the functional glycans recognized by a group of lectins in a panel. The method is more efficiently applied for comparative purposes (e.g. differential profiling). Furthermore, the lectins used to generate the microarray dictate the range of carbohydrate structures that are analysed [240]. It is possible that some carbohydrate structures that are exclusive to bacteria may not be evaluated due to the absence of lectins that recognize these structures. There is a lack in the commercial availability of sugar binding proteins and lectins that are diverse in their recognition of sugar structures. All cellular glycomes are complex and dynamic in nature, so it is imperative that high density microarrays with a more diverse set of lectins are developed. Additionally, the addition of more human and animal lectins will empower this technology and reveal subtle differences in glycan structures on cell surfaces. Another major limitation of the lectin microarray technology is the discovery that the majority of plant lectins present on the array are obtained from natural sources. These lectins have undergone post-translational modifications with carbohydrates, leading to potential false positives due to binding by bacterial lectins [59]. In summary, the lectin microarray platform is a promising glycomics technology. The ability of this technology to rapidly assess dynamic cell surface glycosylation has enabled researcher to monitor and obtain vast information about glycomes. Rapid assessment of bacterial carbohydrates empowers us to review how bacteria are able to modulate their surface glycans to establish cell-cell interactions and host-gylcan interactions. This technology has not only enabled researchers to analyze mammalian cell surface signatures but also capture selected glycosylation defective cell lines.

Although neoglycoconjugate microarray technology has slowly become a critical tool for glycobiologists, there are still numerous challenges that remain. For example, previous use of the technology as a screening tool has yielded little to no binding.

Moreover, it has been found that different microarray platforms yield significantly different results.

There are many variables that affect probe presentation and interaction with proteins on the microarray. There are also many variables that contribute to the information extracted from a microarray experiment and thus to improving this technology. For example, the size, density and diversity of the carbohydrate immobilized on the microarray plays a significant role in the results. Molecular scaffolding and density of the carbohydrate ligands presented on a microarray varies depending on the molecule printed and consequently carbohydrate binding protein recognition will be affected, as multivalent presentation of carbohydrate ligands generally enhances the avidity of carbohydrate binding proteins[103]. The presentation of the glyconjugates on the array surface is a critical factor to recognition. Features such as spacing & orientation of carbohydrates, linker length and flexibility, ligand density all have major impact on recognition. There are several future improvements that could be implemented in microarray technology. For example, the diversity of glycans presented on the existing arrays could be increased. However, researchers have found it challenging to obtain a sizeable collection of carbohydrates in a format that is suitable for immobilization on the microarray. Carbohydrates such as glycosaminoglycans, glycopeptides and nonhuman glycans could be added to future microarrays. Also, slide surface chemistry can have a critical impact on 3D structure, capacity, background noise, spot morphology, presentation, reproducibility and interaction of the array glycan molecules.

FC3 exhibited binding specificity for (1) ovalbumin (Ov), (2) bovine transferrin (bovXferrin), (3) α-Crystallin (a-C) from bovine lens, (4) GlcNAc (GlcNAcBSA), (5) Lacto-*N*-fucopentaose I and (6) II (LNFPIBSA and LNFPIIBSA), (7) 3'Sialyl Lewis x-BSA (SLexBSA14), (8) 6-Sulfo Lewis x-BSA (6SuLexBSA), (9) 3-Sulfo Lewis x-BSA (3SuLexBSA), (10) Galα1,3Galβ1,4GlcNAc-HSA (GGGNHSA), (11) Manα1,3(Manα1,6)Man-BSA (M3BSA), (12) Tri-fucosyl-Ley-heptasaccharide-APE-HSA (3FLeyHSA), (13) Tri-Lex-APE-HSA (3LexHSA) and (14) 2'Fucosyllactose-BSA (3SFLBSA) (Figure 4.12). The binding specificity of FC3 to GlcNAcBSA (N-acetylglucosamine) probe is biologically relevant because it suggests that the human gastrointestinal tract may contains GlcNAc residues that serve as epitopes for glycan-binding interactions with microbial adhesins. GlcNAc is well-known for supporting

the human body's creation of a healthy mucus layer in the gut. Studies have demonstrated that GlcNAc helps support the growth of beneficial gut bacteria like *Bifidobacterium bifidum*. *N*-acetyl-glucosamine containing oligosaccharides were first identified 50 years ago as the 'bifidus factor', a selective growth substrate for intestinal bifidobacteria. Further studies demonstrate that GlcNAc may improve immune function in patients with multiple sclerosis. While N-acetylglucosamine might benfit anyone with digestive problems, it looks to be promising for people suffering from inflammatory bowel disease. Patients with conditions like Crohn's disease and Ulceritive colitis have much thinner mucus barrier in the gastrointestinal tract. In recent study by Andy Zhu and colleagues (April 2015), patients with inflammatory bowel disease taking N-acetylglucosamine for 1 month had substantial improvement in their symptoms.

Moreover, FC3 exhibited binding specificity to the human milk oligosaccharide 2'Fucosyllactose. 2'Fucosyllactose is the most prevalent human milk oligosaccharide, making up 30% of all HMOs. Humans are unable to digest HMOs such as 2'Fucosyllactose, hence the majority of HMO's reach the gut, where they serve as food for desirable gut bacteria. This finding that FC3 binds specifically to 2'Fucosyllactose probes on a microarray is relevant to understanding glycan-microbe interactions beneficial to human health because studies have indicated that 2'Fucosyllactose is able to influence intestinal epithelial cell maturation *in vitro*[310], inhibit *Campylobacter jejuni*-induced inflammation in the intestinal mucosa. HMOs can improve the inner layer of the human gut, boost the immune system, and may be essential nutrient for brain development in babies[311]. Due to its structure, 2'Fucosyllactose binds detrimental bacteria and toxins to prevent them from binding to the baby's gut, decreasing the risk of infection. Further studies have demonstrated that HMO's in both infants and adults are highly specific in the way they modulate the microbiota[312]. The primary impacts are increases in certain Bifidobacterium species and the reduction in several undesirable bacteria. FC3 demonstrates binding specificity to two other human milk oligosaccharides, namely Lacto-*N*-fucopentaose I and II.

Overall, the findings in this chapter demonstrated that restriction digestion and next generation sequencing of the 6 initial putative clones (Figure 4.3) revealed only two viable putative adhesins (FC3 & FC21). DNA segments inserted into the FC3 and

FC21 clones were 24.6 kb and 8.1 kb respectively. FC21 contains three functional genes and belongs to the dominant commensal gut species *Bifidobacterium adolescentis*. Sequence analysis of FC3 revealed that the 24.6 kb insert is a fragment with no current known homologs in the database. Analysis of specific genes present on FC3 highlight the presence of putative adhesive genes such as a collagen adhesion protein, sortase B and sortase D. Analysis of the genes on FC21 suggests that the three transporter genes may not be acting independently but are likely to form a complex that acts as a transporter (adhesin) unit. Lectin microarray analysis of FC3 and FC21 revealed that both clones have altered cell surface glycosylation (reduced GlcNAc residues) as compared to the control strain. This finding is significant because it suggests that a reduction of the GlcNAc residues on the surface of the clones could increase the access of adhesins on the cell surface to bind to receptors on the Caco-2 cell surface. Mucin microarray results indicate that there is no novel binding of the two fosmid clones to mucin as compared to the control strain. In fact, clustering analysis of the triplicate data indicates 100% similarity in mucin binding between FC3, FC21 and EPI300 (pCC1FOS). Finally, the fluorescence intensity for FC3 and FC21 increased as compared to the control strain when interrogated onto NGC microarrays. The presence of antibiotic and inducer promoted the NGC binding of FC3.

The next chapter of this study will describe the *in silico* analysis of a putative bacterial adhesin, $MapA_{Ri}$, encoded by *Roseburia intestinalis* of the human gut metagenome.

Chapter 5:

In silico analysis of the human gut metagenome identifies a putative bacterial adhesin ($MapA_{Ri}$) encoded by *Roseburia intestinalis*.

5.1 Introduction

Microorganisms comprise the major reserve for genetic diversity on earth. The study of DNA sequenced directly from an ecosystem is known as metagenomics. Metagenomics is a powerful tool that allows for the culture independent analysis of complex microbial communities such as the human gut microbiota[313, 314]. The majority of species in the gut are non-culturable which has created difficulties for scientists who study them[186, 315]. However, the recent advances in culture independent techniques have made it possible to identify the majority of the bacteria living in the gut and to compare the microflora composition of different individuals and species[186]. To advance our understanding of gut microbial adhesion and to elucidate potential glycan adhesins, we utilized an *in silico* database screening approach (sequence–based metagenomics) to identify novel adhesin-specific genes encoded by the human gut microbiome. Unlike functional screening of our metagenomics library, *in silico* analysis of the gut microbiota requires prior sequence knowledge to identify and assign putative functions to homologous genes and their encoded proteins.

We present here an *in silico* search for 5 adhesin homologous proteins from 54 individual gut microbial genomes using the basic local alignment search tool (BLAST). Homologous proteins were selected by taking into account sequence similarity, conserved motif, bit scores and expectation values; the efficiency of these screening strategies is discussed. As a result of the searches, a novel $mapA_{Ri}$ gene, encoding a putative adhesin is described.

The following sections will describe the five reference adhesins (four of which were derived from Lactobacillus) used to BLAST against the genomes of 54 individual gut microorganisms to identify homologous proteins. Furthermore, a description of the bioinformatics methodologies that permit the analysis of the metagenomes is also discussed.

5.2 Adherence factors in Lactobacillus

Lactobacilli are one of the indigenous micro-organisms living in the mammalian gastrointestinal tract and have the ability to adhere to mucosal sufaces[144, 316]. The adhesion of some lactobacilli to the gastrointestinal mucosa or mucus is thought to be of importance in the host to promote modulation of the intestinal immune system and

to exert inhibitory effects against pathogenic bacteria[224]. The proteins on the surface of bacteria are composed of a diverse group of molecules with critical roles such as transport, adhesion, signalling, invasion and interaction with the host immune system. In this study, three functionally characterized *Lactobacillus* proteins (Mub, Msa and LspA) that belong to the sortase-dependent protein family were utilized (Figure 5.1). Many Lactobacilli sortase-dependent proteins have the abilty to bind mucus components[101]. Just as lactobacilli are able to bind to the intestinal epithelium and mucus layer, studies have shown that the surface proteins of lactobacillus have been associated with binding to various extracellular matrices. The extracellular matrix (ECM) is a complex structure that forms a boundary around epithelial cells. It consists of proteins such as fibrinogen, collagen and laminin. Any one of these proteins can be shed into the underlying mucus layer, especially in the case of a damaged mucosa. When the mucosa is damaged, it exposes the ECM to unwelcome colonization by pathogens. The presence of lactobacilli which have the capacity to adhere to components of the ECM helps to competitively inhibit binding by pathogens.

One well studied ECM adhesin is the collagen binding protein (CnBP) of *Lactobacilli reuteri* NCIB 11951[317]. CnBP has the capacity to adhere and solubilize collagen. According to Pfam, CnBP contains a bacterial extracellular solute-binding domain. This particular domain is also present in the CnBP homologous protein in *Lactobacillus reuteri* 104R, MapA (Mucus adhesion promoting protein) (Figure 5.5). Interestingly, MapA has been shown to bind to Caco-2 cells (human adenocarcinoma cells) and mucus, even though it does not contain any known mucus-binding domains[101].

5.2.1 Extracellular mucus-binding protein, Mub

In 2002, Roos & Jonsson[316] identified an extracellular mucus-binding protein (Mub) of *Lactobacillus reuteri* 1063 (Figure 5.1C). They cloned and sequenced this cell surface protein after discovering that it binds to mucus components *in vitro*. Mub genes are found in all of the six genomes of *Lactobacillus reuteri* that are available[318]. The Mub protein family has been extensively studied. The binding of Mub to mucus components occurred in the pH range 3-7.4, with the maximum binding at pH 4-5 and was partially inhibited by the glycoprotein Fetuin. Roos & Jonsson were able to demonstrate the presence of the Mub protein on the cell surface of *Lactobacillus reuteri* 1063 by using affinity-purified antibodies against recombinant Mub in

immunofluorescence microscopy. They detected the presence of Mub proteins in the growth medium by using antibodies in a Western blot analysis[316].

Mub proteins consist of a 49-amino acid N-terminal secretion signal peptide, followed by a mature protein with a predicted molecular mass of 353 kDa. The Mub protein is one of the largest bacterial cell-surface proteins identified[316]. The protein is highly repetitive with two different types of MucBP (Mucin binding protein) amino acid repeats (Mub1 and Mub2) of roughly 200 amino acids, present in eight and six copies, respectively, and shown to be responsible for the adherence of the bacteria to intestinal mucus[96] (Figure 5.1C). Specifically, Mub consists of six copies of a type 1 repeat (Mub1) ranging from 183 to 206 amino acids in length[319]. The remaining eight copies of a type 2 repeat (Mub2) all consist of 184 amino acids in length with the exception of one which has 186 amino acids[319]. Previous research has shown that the six Mub1 repeats are quite diverse, whereas the Mub2 repeats demonstrate low sequence variation. Proteins that consist of Mub repeats are most often found in lactic acid bacteria (LAB), with the highest abundance exhibited in lactobacilli of the GIT[319]. The Mub repeat sequence was found to be highly similar to the mucin-binding protein (MucBP, PF06458) domain and a sequence identified in the *Listeria monocytogenes* strain[320, 321]. Further investigations of the MucBP homologous proteins in other organisms revealed that 10 bacterial species contain cell-surface proteins with amino acid sequences similar to the MucBP domain[322, 318, 323]. Recent research by Etzold *et al*. has demonstrated that MUB recognizes sialic acid residues in mucin chains[324].

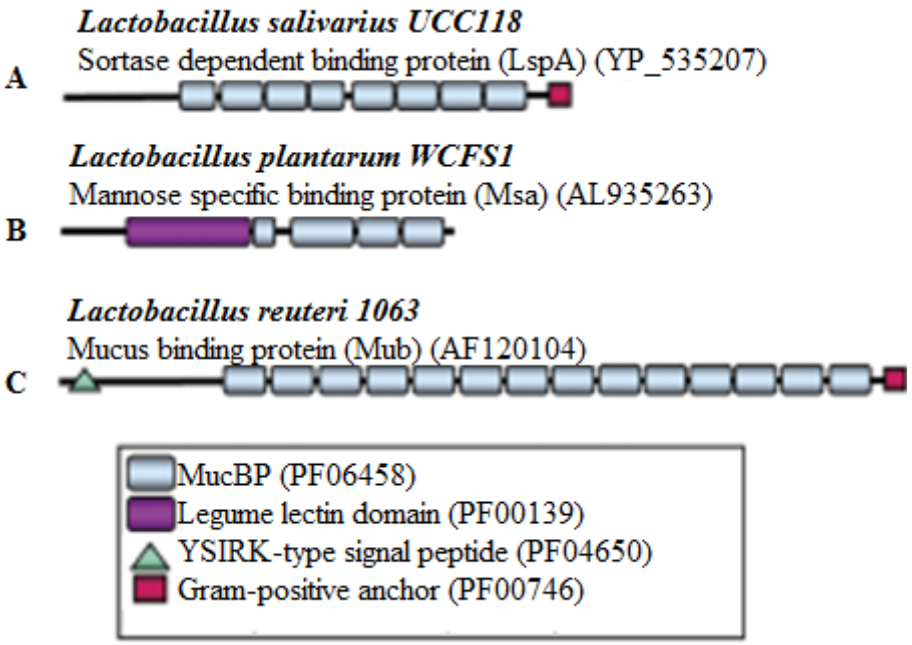


Figure 5.1 **Domain architecture of the fully characterized LspA, Msa, and Mub lactobacilli adhesin proteins according to Pfam database.** (A) Domain architecture of the LspA lactobacilli sortase-dependent adhesin. The YSIRK sequence feature represents a YSIRK-type signal peptide and 8 MucBP domains and a gram-positive anchor. (B) Domain architechture of the fully characterized Msa lactobacilli adhesin. The legume lectin domain with 4 MucBP domains and a gram-positive anchor. (C) Mub contains 14 MucBP domains and a gram-positive anchor.

5.2.2 Lectin-like Mannose Specific Adhesin, Msa

Lactobacillus plantarum WCFS1 is a facultative hetero-fermentative lactic acid bacterium that was originally isolated from human saliva[325]. *L. plantarum* is a highly flexible and adaptive species that is detected in many different environmental niches. Its chromosome encodes more than two hundred extracellular proteins, many of which are predicted to be bound to the cell envelope[325]. In 2005, Pretzer and colleagues[97] identified a protein of *Lactobacillus plantarum* WCFS1 that contains MUB domains (Figure 5.1B). The protein is involved in binding of mucus via mannose, which is a component of mucin glycosylation moieties [97] (Figure 5.1B). This mannose-specific adhesin (Msa) is a protein of 1,010 amino acid residues with conserved sequences that are highly homologous with sequences of the ConA lectin-like SasA domain and MucBP domain[321, 97].

Msa also contains a legume lectin domain. The leguminous lectins form one of the largest lectin families and resemble each other in their physicochemical properties but differ in their carbohydrate specificities[326]. They bind either glucose, mannose or galactose. Carbohydrate binding activity depends largely on the presence of both calcium and a transition metal ion[327]. The exact function of legume lectins is unknown, however they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens[328]. The ability of *L. plantarum* to bind mannose could be useful with regard to proposed probiotic characteristics such as colonization of the intestinal surface and competitive exclusion of pathogens[329].

5.2.3 Lactobacillus surface protein A, LspA

In 2006, Claesson and colleagues[209] performed a comparative bioinformatics analysis of four publicly available Lactobacillus genomes and the genome of *Lactobacillus salivarius* subp. *salivarius* strain UCC118 to identify both secreted and cell wall linked proteins. *Lactobacillus salivarius* strain UCC118 has several probiotic qualities such as resistance to acid and bile, production of a blood-spectrum bacteroicin, attenuation of induced arthritis in a mouse knockout model, alleviation of symptoms associated with mild-to-moderate Crohn's disease and adherence to the intestinal mucosa[330]. Claesson and colleagues identified 10 sortase-dependent surface proteins in *L. salivarius* UCC118[330]. One of these is called LspA (Lactobacillus surface protein A) (Figure 5.1A).

LspA mediates adhesion of the strain to human epithelial cells and mucus[209]. According to the Pfam database, LspA consists of eight mucus-binding domains (MucBP) (PF06458). LspA (LSL_0311) is a 1,209 amino acid protein that contains eight repeats of 79aa (Repetetive region 1 to repetitive region 7, R1 to R7) (Figure 5.1A)[331]. The R1 and R7 are the least conserved repeats and share 73% identity. On the other hand, R2 to R6 are more conserved with over 92% identity. According to the Pfam database, each of the 7 repeats is similar to mucus-binding domains (PF06458).

5.2.4 Starch binding proteins, SusD & SusC

Bacteroides thetaiotaomicron is a prominent gut microbe, gram-negative, obligate anaerobe that is able to effectively utilize polysaccharides as a source of carbon and energy[177]. Studies have confirmed that *B. thetaiotaomicron* produces cell associated enzymes that break down polysaccharides with the majority of the activity located in

the periplasm and cytoplasm[332]. An important first step to polysaccharide utilization by *B. thetaiotaomicron* is the binding of the polysaccharide to the cell surface before hydrolysis. The polysaccharide is first bound to a putative outer membrane receptor complex and then translocated to the periplasm, where the degradative enzymes are located. *B. thetaiotaomicron* uses this strategy during polysaccharide utilization to allow the bacterium to effectively sequester hydrolysis products and also to attach itself to a polysaccharide containing particle[332].

B. thetaiotaomicron contains a cluster of eight starch utilization (susABCDEFGR) genes (Figure 5.2). Studies have indicated that both SusC and SusD are surface exposed. When produced separately in intact *E. coli* cells, both SusC and SusD demonstrate accessibility to protease digestion[333]. Moreover, the amino acid sequence of SusC suggests that SusC might be a porin[334]. In this role, SusC would have to be surface exposed in order to admit the oligosaccharides into the periplasmic space. As for SusD (Figure 5.2), it is known to bind long-chain starch which would require that this protein be exposed on the surface as well.

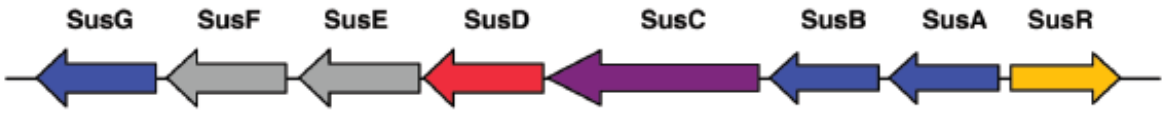


Figure 5.2 **Cluster of starch utilization genes.** SusCDEFG are localized at the cell surface and bind, degrade and import soluble starch molecules. Diagram not drawn to scale.

Studies have shown that the outer membrane protein SusC is not sufficient to bind starch alone. Rather, a combination of SusC and SusD form a complex and interact with one another on the cell surface to bind starch[335]. Therefore, in this study, it was hypothesized that SusD/SusC would behave like adhesins since they are involved in binding and are surface exposed.

Table 5.4 Five adhesins used as reference adhesins during *in silico* work

Adhesin	Full Name	Organism	Binding Target	N-terminal domain	C-terminal domain	Adhesive domains
MapA (263aa)	Mucus adhesion promoting protein	*Lactobacillus reuteri* 104R	Caco-2 cells and mucus	Sec Leader	No Hits in PFam	1 bacterial extracellular solute binding domain
Msa (1010aa)	Lectin-like mannose specific adhesion	*Lactobacillus plantarum* WCFS1	Mucus via mannose binding	KxYKxGKxW signal peptide	Gram positive Anchor LPXTG	1 legume lectin domain, 3 MucBP
Mub (3269aa)	Mucus binding protein	*Lactobacillus reuteri* 1063	Mucus components	YSIRK signal peptide	Gram positive anchor LPxTG	14 MucBP
LspA (1209aa)	Lactobacillus surface protein	*Lactobacillus salivarius* UCC118	Human epithelial cell lines	YSIRK signal peptide	Gram positive anchor LPQTG	8 MucBP
SusD (551aa)	Starch binding protein	*Bacteroides thetaiotaomicron* VPI-5482	Starch; long chain levans	N/A	N/A	1 SusD starch binding domain

5.2.5 Mucus adhesion promoting protein, MapA

In 2005, Miyoshi and his colleagues[336] established that the surface protein MapA is an adhesin located on the surface of the gut bacterium *Lactobacilli reuteri* 104R (Figure 5.5). MapA is a cell-surface protein with a molecular weight of 26kDa and a theoretical pI of 9.7 (Figure 5.5)[337]. MapA consists of a 25 amino acid sec sequence and a SBP_bac_3 extracellular solute binding protein domain (Figure 5.5B). They showed that some *L. reuteri* strains are able to bind to both mucus and epithelial cells (Caco-2 cells)[336]. They screened purified MapA protein onto Caco-2 cells (human

adenocarcinoma) and discovered that binding of the purified MapA to Caco-2 cells inhibited binding of the *L. reuteri* 104R in a concentration-dependent manner. Therefore, binding of *L. reuteri* 104R appears to be mediated to some extent by MapA adhesins which bind to receptor-like molecules on Caco-2 cells[337]. Their experiments also showed that although MapA is a key adhesin in the binding of *L. reuteri*, it is not the only one. Their competitive adhesion assay indicated that half the number of bacteria still bound to Caco-2 cells despite the saturating quantities of MapA as the competitor. Interestingly, these researchers identified two receptor molecules for MapA from Caco-2 cells using far-western analysis. The molecules were 90-kDa and 200-kDa respectively. Based on their hydrophobicity, Miyoshi *et al.,*[336] hypothesized that the molecules are located in the cell membrane fraction of Caco-2 cells.

Sequence alignments have demonstrated that MapA is a member of family III of the bacterial solute binding proteins[338], and like other members of this family, it is able to bind to polar amino acids, opines (cysteine, glutamine, arginine, histidine, lysine, octopine and nopaline) and transfer them to a membrane-located translocation complex. Open reading frames upstream of *mapA* potentially encode the other components of an ABC-type uptake system[339]. It is possible to extract the MapA protein from the cell surface with strong electrolyte solutions. Studies have demonstrated that MapA can be selectively removed from whole cells by using 5M LiCl or a low-pH buffer, indicating that it is likely to be anchored non-covalently to the cell surface[339]. Comparison of MapA to all proteins encoded by the *Escherichia coli* genome revealed that MapA is most similar to the L-cystine binding protein FLiY[340]. There is further evidence that MapA is necessary for maximal resistance to oxidative stress[337].

Although MapA is reported to mediate the binding of *L. reuteri* 104R to Caco-2 cells and mucus, no mucus-binding domains were detected according to the Pfam database. Bohle and colleagues discovered that MapA produces an antimicrobial peptide (AMP) degradation product called AP48-MapA[336].This finding suggested that MapA performs additional functions by conferring antimicrobial capacity to the *Lactobacillus reuteri* strain. The MapA/AP48-MapA system is an example of an intestinal protein giving rise to an antimicrobial peptide.

A

MKFWKKALLTIVALTVGTPAGITSVSAASSAVNSELVHKGELTIG**LEGTYSPYSYRKNNKLTGFEVDL**
GKAVAKKMGLKANFVPTKWDSLIAGLGSGKFDVVMNNITQTPERAKQYNFSTPYIKSRFALIVPTDSN
IKSLKNIKGKKIIAGTGTNNANVVKKYKGNLTPNGDFASSLDMIKQGRAAGTINSREAWYAYSKKNST
KGLKMIDVSSEQDPAKISALFNKKDTAIQSSYNKALKELQQDGTVKKLSEKYFGADITE

B

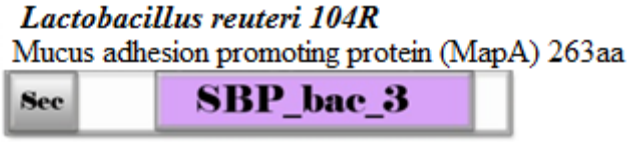


Figure 5.5 **Amino acid sequence and domain architecture of the MapA protein**. (A) MapA (accession no. AJ293860) is a mucus adhesion promoting protein (263aa) containing a *sec* leader sequence (underlined), the emboldened sequence represents the sequence deciphered by chemical sequencing of purified peptides (B) Domain architecture of the fully characterized *Lactobacillus reuteri* 104R adhesin MapA according to PFam database.

MapA shows high sequence similarity at the amino acid level (98%) to CnBP of *L. reuteri* NCIB 11951, a 29-kDa surface protein (p29) from *L. fermentum* RC-14 and BspA (CyuC) of *L. fermentum* BR11 (BR11). Cystine uptake C (CyuC) is a surface protein that is required for L-cystine uptake and oxidative defence and is an L-cystine binding protein[341]. Turner *et al*.,[342] showed that a CyuC mutant is incapable of binding to L-cystine, transporting L-cystine and is sensitive to oxygen and the superoxide generating reagent methyl viologen[342]. MapA participates in L-cystine uptake and oxidative defense and is likely to also be an L-cytsine binding protein[343]. Hung and colleagues[341] discovered that de-energized *L. reuteri* cells were able to bind radiolabelled L-cystine with a Kd of 0.2 μM. Mutant MapA cells were unable to bind L-cystine.

5.3 Roseburia intestinalis

*S*tudies on putative adhesins derived from *Roseburia intestinalis* are highly relevant to the medical and scientific fields due to the health-promoting effects of this butyrate producing microorganism[221]. *Roseburia intestinalis* is a saccharolytic, motile, Gram-positive, non-spore forming, anaerobic, low G + C-content, butyrate-producing bacterium that is abundant in the human gastrointestinal tract. Bacteria that ferment

carbohydrates (particularly undigested fiber) to produce short-chain fatty acids (SCFA) such as butyrate are considered to have health-promoting properties[344]. Butyrate is an important nutrient for colonocytes and acts as a signalling molecule with a critical role in cell differentiation and apoptosis[344]. Butyrate in particular is the preferred energy source of colonocytes and is believed to play an important part in maintaining colonic health in humans[345]. It is also believed to exert direct effects upon gene expression in mammalian cells through histone hyperacetylation and through interaction with butyrate response elements upstream of some genes[344]. Butyrate production *in vitro* of human faecal microflora is highly dependent on the growth substrate (energy source). Starch strongly induces the production of butyrate whereas polysaccharides such as pectin produce less butyrate and more acetate and propionate[345]. Therefore, the relative level of SCFA in the human gut provides a link between diet and colonic health.

Recently, several human colonic Roseburia species were confirmed to actively metabolize linoleic acid, forming either vaccenic acid or hydroxyl-18:1 fatty acid with the capacity to act as a precursor for health-promoting conjugated linoleic acid[345]. As a result, there is a keen interest in the role of species such as *Roseburia intestinalis* in the human gut. Surprisingly, little is known about the identities, physiologies and ecologies of the predominant butyrate producing bacteria from the human gut. Greater insight into the particular butyrate producing species of the human gut will yield more mechanistic understanding of the effects of diet on butyrate production in the colon.

5.4 NICE, Nisin controlled gene expression system for *Lactococcus lactis*

Nisin is an antimicrobial peptide that is widely used as a food preservative[346]. It is produced by several strains of *Lactococcus lactis*, that belong to the Class I bacteroicins called lantibiotics[198, 347]. *Lactococcus lactis* is a homofermentative bacterium that is used for the production of fermented milk products such as buttermilk, fermented butter and many varieties of soft and hard cheeeses[348]. Nisin is a relatively small (3.4 kDa), 34-amino acid, cationic, hydrophobic peptide with five characteristic rings formed by significant post-translational modification[198,349]. A cluster of 11 genes are involved in the biosynthesis of nisin and are transcriptionally arranged as nisABTCIPRKFEG[1, 349].

In order to improve the likelihood that the candidate adhesins would be expressed in different heterologous hosts, a shuttle vector pPTPi (Figure 5.6) capable of replicating in *Escherichia coli* and gram-positive bacteria (*Lactococcus lactis* NZ9000) containing nisin-inducible promoter (P*nis*A) and genes encoding NisR and NisK, the two-component signalling mechanism for activating transcription from P*nis*A in the presence of nisin[350] was utilized. This expression system based on the autoregulatory (quorum-sensing) properties of the *Lactococcus lactis* nisin gene cluster has become one of the most successful and widely used tools for regulated gene expression in Gram-positive bacteria[351,348]. Nisin induces the transcription of the "gene of interest" under control of the PnisA promoter via a two-component regulatory system, consisting of a histidine kinase (HK) and a response regulator (RR) [347]. Therefore, the combination of the nisA promoter, the nisRK regulatory genes and the externally added nisin (sub-inhibitory concentration, 0.1-5 ng ml^{-1}) controls the expression. Studies have indicated that increasing amounts of nisin yield a linear dose-response curve[1]. Thus, the NICE system can be used not only for on/'-off gene expression studies but also to modulate the amount of the target protein produced.

In this study, this inducible NICE (Nisin Controlled Gene Expression system) system was used as a one-plasmid system with an engineered host strain containing nisRK genes integrated in the chromosome (*Lactococcus lactis* NZ9000)[210]. *Lactococcus lactis* NZ9000 is a derivative of L. *lactis subsp. cremoris* MG1363, a plasmid-free progeny of the dairy starter strain NCDO712[347]. *L. lactis* NZ9000 is constructed by the integration of the genes nisK and nisR into the pepN gene of the MG1363. This integration of the nisRK genes into the chromosome (*L. lactis* NZ9000) often limits the system to specially designed host strains[198].

One of the many reasons why *L. lactis* is an excellent model species for expression and study of our candidate adhesins is because *L. lactis* is a gram positive bacterium with only one cellular membrane. Many of the target adhesins are derived from Gram positive organisms which would be presented appropriately on a Gram positive cell than a Gram negative *E. coli* cell. Nisin is an ideal molecule for induction because it is widely used in the food industry and is regarded as a food-grade inducer; the level of gene expression is controllable in a dynamic range of > 1000 fold which is directly dependent on the concentration of nisin used for induction[347]. The tightly regulated NICE system permits the cloning and induction of membrane proteins that could be

toxic for the cell. Indeed, expression of the desired gene is so tightly controlled that undetectable protein levels are observed in the uninduced state. This makes it possible to produce lethal proteins. The cells have weak proteolytic activity, the membrane proteins are easily solubilized with various detergents and *L. lactis* has only one membrane which permits direct functional studies with either a whole bacteria or isolated membrane vesicles[1]. To date, *L. lactis* has been used in the expression and functional analysis of a series of prokaryotic integral membrane proteins such as ATP-binding cassette (ABC) transporter, peptide transporters, mechano-sensitive channel, ABC efflux pumps, ATP/adenosine diphosphate transporters and major facilitator superfamily proteins[348].

Studies have also shown that *L. lactis* is effective and suitable for the expression of eukaryotic membrane proteins such as KDEL receptor and mitochondrial carriers[196]. These proteins were expressed at between 0.1 and 5% of all membrane proteins and remained functionally intact with the same characteristics exhibited in their natural environment[350]. Another important reason why *L. lactis* is suitable for the expression and surface exposure of proteins is because of its low extracellular proteinase activity. Currently, there are only two proteases recorded in lactococci, namely the cell wall anchored protease PrtP (200kD) and the housekeeping membrane-bound protease HtrA. PrtP is plasmid-encoded and absent from plasmid free host strains. HtrA can be mutated to help stabilize secreted proteins in lactococci[347].

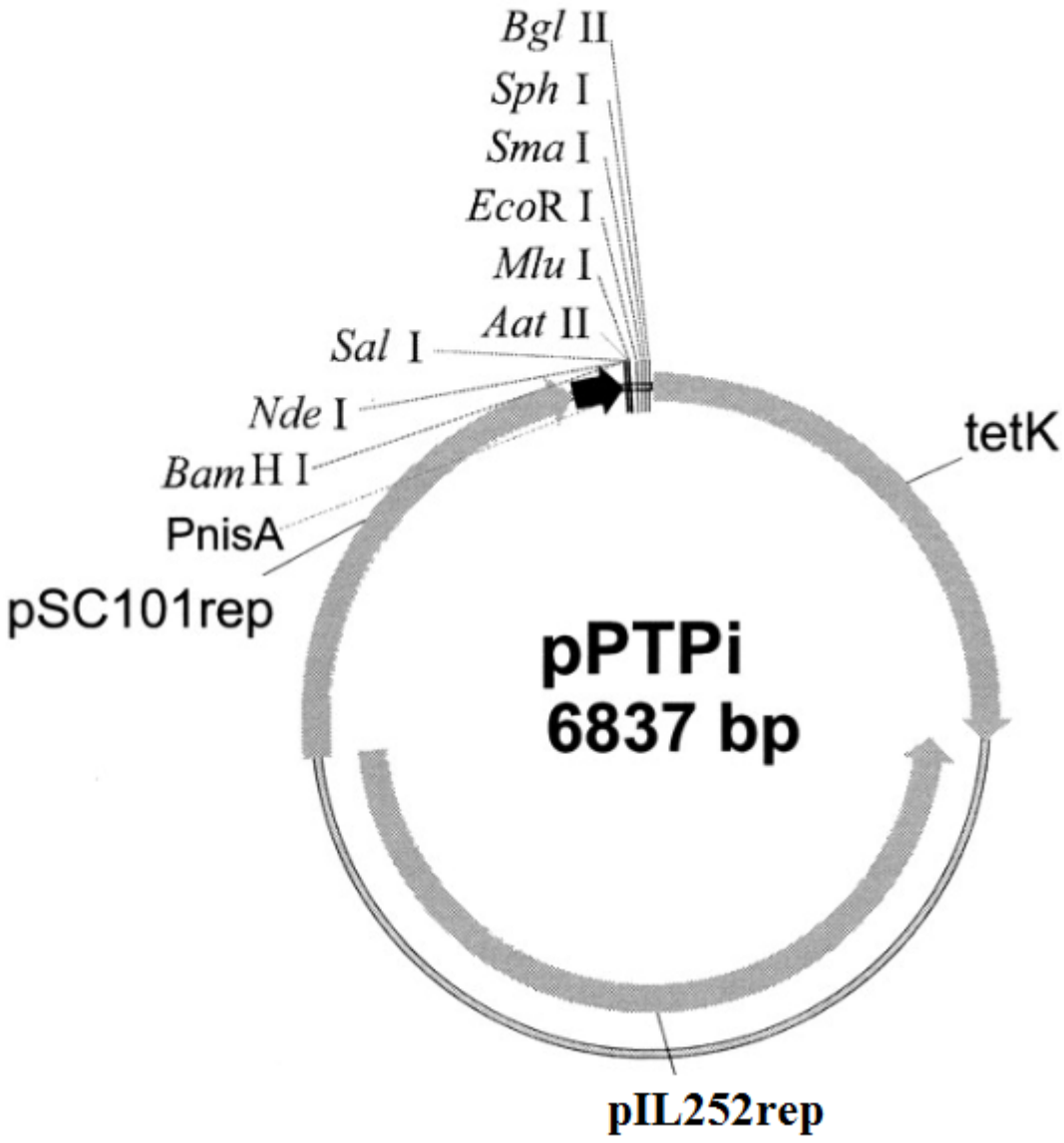


Figure 5.6 **pPTPi vector map**. pPTPi is a low-copy number *E.coli-L.lactis* shuttle vector for cloning Tc^r, P*nis*A, pPTP derivative. Diagram adapted from Douillard *et al.*, 2011[198].

5.5 Homologous sequences search using BLAST

In this study, the presence of homologous "adhesin" proteins in the bacterial genomes of 54 abundant gut species (Table 5.8) was investigated by performing a BLASTp (protein-protein BLAST) sequence similarity search of the corresponding reference adhesins (Table 5.4) against the non-redundant protein sequence database (National Centre for Biotechnology Information, NCBI) for each of the 54 organisms. The Basic Local Alignment Search Tool (BLAST) is a program that reports regions of local similarity and is able to produce reliable and accurate statistical estimates of protein

and DNA sequences that share significant similarity[352]. BLAST does not determine homology but provides data that may support an inference of homology. The data is provided in the form of percent sequence identity, scores to rank comparisons, statistics to help judge relevance and the alignments. Protein or DNA sequences that share significant similarity can be inferred to be homologous; that is to say they display common evolutionary ancestry[353].

The ability to detect sequence homology facilitates the identification of conserved domains that are shared by multiple genes and is an important step in predicting the function of proteins that have not been studied experimentally[353]. Generally, homologous proteins have similar structure because structures diverge much more slowly than their sequences over time[352]. However, homologous sequences do not always share significant sequence similarity. Many homologous protein alignments are not significant but are homologous based on statistically significant structural similarity or strong sequence similarity to an intermediate sequence[352]. The inference that two protein sequences are homologous does not ensure that every part of one protein sequence has a homolog in the other[352].

In this study, protein (or translated DNA sequences) sequences were used in the BLAST searches because they are more sensitive than DNA: DNA searches[354]. Studies have shown that DNA: DNA alignments have between 5-10 fold shorter evolutionary look-back time than protein: protein or translated DNA: protein alignments[352]. After approximately 200-400 million years of divergence, DNA: DNA alignments rarely detect homology. However, protein: protein alignments are capable of detecting homology in sequences that share a common ancestor 2.5 billion years ago (e.g. humans to bacteria). Furthermore, it has been shown that DNA: DNA alignments are less accurate than protein: protein statistics. DNA: DNA expectation values $<10^{-6}$ often occur randomly by chance; whereas protein: protein alignments with expectation values $<10^{-3}$ can reliably be used to infer homology[352].

The first step in performing the BLAST search was to identify an appropriately stringent set of BLAST parameters (e.g. bit score and e-value) to ensure (to the extent possible) that the homologs of the adhesins detected in the individual genomes were functional (and did not correspond to distantly related non-functional homologs of the reference genes). Notwithstanding this, it was important to ensure that the stringency

of the thresholds were not so high as to result in loss of distantly related homologues that share functionality with the reference adhesins. Bit score and Expectation value (statistical indicators of alignment) were used as thresholds in this study to identify homologous proteins in the genomes of 54 gut bacterial micro-organisms.

The Expectation value (E-value) provides information about the probability that a given sequence match occurred purely by chance. The lower the e-value, the less likely the database match is a result of random chance and therefore the more significant the match[354]. If e-value <1e-15, there should be confidence that the database match is a result of a homologous relationship. However, if the e-value is between 0.1 and 10, the match may be considered not significant but may hint at distant homology. An E-value of 10 means there is essentially no likelihood of true homology – it indicates the number of matches that would be expected to arise purely by chance in a given search. The e-value takes into account the specific size of the database (NCBI non-redundant subset "nr" which currently contains about 6.5 million proteins and 2.2 billion amino acids) used in the query as well as the length of the query sequence[352]. Therefore, as the database grows, the e-value for a given match will also increase. An alignment score found by searching a 10 million protein entry database will be 100-fold less significant than exactly the same score found in a search of 100,000 protein entry database. This, however, does not mean that the sequences are no longer homologous[352]. It simply means that homology will be harder to detect in the larger protein entry database. Therefore, sequences that share significant sequence similarity can be inferred to be homologous, but the absence of significant similarity does not imply non-homology. An expectation value threshold of <1e-15 and a bit score of 50 was used in this study. The choice of expectation value threshold was determined in an arbitrary manner but depended largely on query and database lengths. Identifying homologous sequences using BLAST required the adjustment of BLAST parameters to increase the sensitivity and specificity of the searches. At NCBI, a cut-off of <1e-6 is used as the default for internal processes using BLASTp search against the protein non-redundant subset of Genebank ("nr") database[352].

Unlike the expectation value, the bit score measures sequence similarity independent of the query sequence length and database size. It is normalized based on the raw pairwise alignment score. Thus, the higher the bit score, the more highly significant the match is. For an average length protein, a bit score of 50 is almost always

significant[352]. In this study, a colour code (similar to NCBI Blast web site) of blue for alignment scores between 40-50 bits, green for scores between 50-80 bits, pink for scores between 80-199 bits; and red for scores >=200 bits was used (Table 5.8 & 5.9). Excluding very long proteins and very large databases, 50 bits of similarity score will always be statistically significant and is a good rule-of-thumb for inferring homology in protein alignments[352].

In the past, investigators used "percent identity" as a criterion to describe similarity between two sequences. However, the percent identity between two sequences cannot be used to infer degree of homology because two sequences are either homologous or they are not. They cannot share low or high degrees of homology. The common rule of thumb was that two sequences are homologs if they are more than 30% identical over their lengths[352]. However, this 30% criterion misses many easily detected homologs. Though a 30% identical alignment over more than 100 amino acid residues implies statistical significance, many homologs are readily detected using an e-value cut off of <1e-10 that are not 30% identical. In this study, homologous proteins were readily detected using an expectation threshold of <1e-15 that were less than 30% identical. For example, a single homologous protein was detected in *Parabacteroides johnsonii* to the SusD query sequence (Table 5.8) in a BLASTp search. The homolog exhibited a 28% identity alignment spanning the SusD domain region of the target protein. This region of similarity represents the conserved domain (SusD domain) between both sequences. In another example, when 6,629 *Saccharomyces cerevisiae* proteins were compared to 20, 241 human proteins, 46.5% (3,084) of the yeast protein shared significant similarity with a human protein (E-value <1e-6), but only 67.4% (2,081) of those proteins were more than 30% identical[352]. This highlights the importance of the expectation value and bit score as a sensitive search criterion to detect homologous proteins that would otherwise have been overlooked. While percent identity is neither a very sensitive nor a reliable measure of sequence similarity, it is a reasonable substitute for measuring evolutionary distance once homology has been established. Percent identity is not linear with evolutionary distance, but it provides a useful approximation of evolutionary distance. Once two sequences are confirmed to be homologous, it is important to observe the sequence alignments and percent identities. For example, an identity percentage of 35% is valuable because 35% of the identity of the homologous protein may span a critical

part of the protein (i.e., a binding site, a shared domain or motif). Therefore, homologous sequences are usually similar over an entire sequence or domain, typically sharing 20-25% or greater identity for more than 200 amino acid residues. However, matches that are more than 50% identical in 20-50 amino acid regions occur frequently by chance and do not indicate homology. The following sections of this study will describe in detail the results obtained for BLAST homology searches using each of the five reference adhesins (Table 5.4).

RESULTS:

5.6 Mining human gut metagenomics database using 5 reference adhesins

The first step in identifying homologous adhesins in the human gut metagenome was to mine publicly available human gut metagenomics databases. The nucleotide sequences of the five reference adhesins (Table 5.4) were obtained from NCBI and used to perform a BLASTn (nucleotide-nucleotide) search against publicly available human gut metagenomics databases. Sequences of gut microbial fragments (referred to as contigs) obtained from the gut microbiomes of individuals belonging to diverse geographical locations were downloaded from http://www.bork.embl.de/Docu/Arumugam_et_al_2011/downloads.html. These metagenomic sequences that were derived from Dutch, American, Danish, French, Spanish, Japanese and Italian individuals, were previously analysed by Arumugal *et al* [28]. Gut metagenomics contigs corresponding to 116 European (Danish and Spanish) individuals, previously analysed by Qin *et al* [25], were also downloaded from http://gutmeta.genomics.org.cn/. Additionally, contigs corresponding to 90 gut metagenomes obtained from American individuals, sequenced as part of the Human Microbiome Project, were downloaded HMP-DACC website (http://www.hmpdacc.org/HMASM/). The gut metagenomics datasets from 30 Chinese individuals previously analysed by Hu *et al.*[355] were also downloaded from the NCBI SRA database. All files had to be downloaded individually and unzipped using a zip software. The community cyberinfrastructure for advanced microbial ecology research and analysis (CAMERA) database was consulted. Local attempts to screen the human gut metagenomics databases failed to return any matches as a result of software and computational limitation. Consequently, our search was narrowed to 54 individual gut microbial genomes to give a more focussed study.

5.7 BLAST search against 54 individual bacterial genomes using five reference adhesins.

In this study, five reference adhesins (Table 5.4) were used as query sequences in BLASTp (non-redundant (nr) protein sequences) searches against the individual genomes of 54 abundant gut bacterial microbes. The 54 abundant gut microbial genomes[25] were identified in works performed by Qin and colleagues (2013). Each reference adhesin was BLASTp against each of the 54 bacterial genomes and the output assembled into Table 5.8 (homologous proteins and bit scores) and Table 5.9

(homologous proteins and e-value). Some reference adhesins did not yield any output ("No significant similarity found") when BLAST against certain genomes. One possible explanation for this is that the query reference adhesin originates from a different taxa than the one in which it is being searched.

Overall, we were able to identify homologous sequences to all reference adhesins. A MapA homologous protein was selected for further characterization because the BLAST search returned a very strong match (high bit score, low e-value) to the *Roseburia intestinalis* genome. Studies have indicated that the butyrate producing *R. intestinalis* has numerous health promoting effects for the host and little is known about the mechanisms it uses to attach to the colonic mucosa.

5.7.1 BLAST homology search using MapA reference protein

Overall, we were able to infer homology for all the MapA matches depicted in Table 5.8 with bit scores above 80 and expectation values below 1e-15 (Table 5.9). The sequence alignments were analysed to determine if the homologous protein sequence is similar over a critical part of the query sequence such as a domain or binding site (Appendix 1-5). For example, although the bit scores of the MapA homologous matches in *Anaerotruncus colihominis* and *Clostridium scindens* are the lowest, respectively, we were still able to infer homology based on the low expectation values and the region of sequence alignment.

Four different homologous proteins were detected from four different strains as an output from the BLASTp search of the MapA query sequence against *Anaerotruncus colihominis*. Although the homologs exhibit relatively low expectation values and high bit scores, they did not exhibit amino acid sequence similarity throughout the entire length of the target MapA protein. They exhibited homology to portions of the MapA protein that are not actively involved in the adherence of MapA to Caco-2 cells or mucus (such as the extracellular solute binding domain) (Appendix 1). As a result, the four detected MapA homologs in *Anaerotruncus colihominis* were not pursued as they did not possess the conserved bacterial extracellular solute-binding domain (SBP_bac_3) region that is relevant for adherence of the MapA target protein to mucus and Caco-2 cells[101].

Statistically significant homologous MapA proteins were detected in 24 out of the 54 gut bacterial genomes. As described in section 5.2.5, MapA is an L-cystine-binding

protein that is involved in the uptake of L-cystine for oxidative defence of the cell. Intercellular thiols like L-cystine play critical roles in the regulation of cellular processes. Most bacteria possess L-cystine transporters (ABC transporters) that transport cystine from the periplasm to the cytoplasm[343]. Therefore, it was not surprising that homologous proteins in 24 of the 54 individual bacterial genomes with MapA query sequence were detected. In contrast, BLAST searches of the 54 individual bacterial genomes using the reference adhesins with MucBP domain proteins (Msa, Mub & LspA) did not detect homolgous proteins in the vast majority of individual genomes at the expectation threshold of 1e-15.

The biggest number of MapA homologous proteins were detected in *Escherichia coli* (100) and *Streptococcus thermophilus* (42). 100 different MapA homologous proteins were detected in 100 different *Escherichia coli* strains presumably due to the large number of *E. coli* genome sequences in the database. The top homologous hit in *Escherichia coli* for MapA is a Cystine ABC transporter substrate-binding protein (sequence ID: WP_047083357.1) with a bit score of 149 and an expectation value of 3e-42. This match exhibited a 35 percent sequence identity throughout the entire lentgth of the MapA target protein. All 100 homologous proteins exhibited 35 percent sequence identity throughout the entire length of the MapA target protein. This suggests that the bacterial extracellular solute binding domain is conserved in each homolog.

Five matches to the MapA query sequence were detected in the species *Bacteroides xylanisolvens* (Table 5.8, 5.9). The highest match exhibited a bit score of 100 and an e-value of 4e-25. Upon examining the alignment for local areas of high similarity, it was detected that the homolog retained the conserved bacterial extracellular solute binding domain (SBP_bac_3) present in the target protein. Therefore, the homologous protein was retained as relevant to this study. All homologous proteins with low expectation values, high bit scores and relevant sequence alignment were retained for further analysis.

Seven hits to the MapA query sequence was detected in *Bifidobacterium adolescentis*. The highest match exhibited a bit score of 87.8 and an expectation value of 1e-20. A 30% percent identity was detected for this hit, spanning 6-260aa residues of the MapA query sequence. This region of the MapA query sequence encompasses the vast

majority of the full protein including the bacterial extracellular solute binding protein (42-259aa).

Six matches to the MapA query sequence was detected in *Bifidobacterium longum*. The highest match exhibited a bit score of 143 and an expectation value of 2e-47. Based on the bit score and expectation value, homology can be inferred for this match. With careful inspection of the results, we can see that most of the amino acid residues of the top hit align to SPB_bac_3 (bacterial extracellular solute binding protein domain) region of the query sequence. Moreover, a 38% percent identity was detected for this match spanning 35-263aa residues of the MapA query sequence. This region excludes the signal peptide but includes the bacterial extracellular solute-binding protein domain. Four matches to the MapA query sequence was detected in *Bifidobacterium infantis*. The highest match exhibited a bit score of 143 and an expectation value of 6e-42. A 38% percent identity was detected for this match spanning 35-263aa residues of the MapA query sequence. Only 1 match to the MapA query sequence was detected in *Roseburia intestinalis* M501. The match exhibited a bit score of 176 (Table 5.8) at an expectation value of 4e-55 (Table 5.9). This single match exhibited a 42% sequence identity throughout the entire length of the MapA target protein (Appendix 1).

As observed in Table 5.8, proteins identified to the MapA sequence in Lactic acid bacteria (*Lactobacillus reuteri* 1063, *Lactobacillus plantarum* WCSF1, *Lactobacillus salivarius* UCC118) are mostly colour coded in red signifying very high bit scores (>=200) and consequently very low expectation values. These results are not unexpected since the MapA reference protein also originates from a Lactic acid bacteria (*Lactobacillus reuteri* 104R)[336]

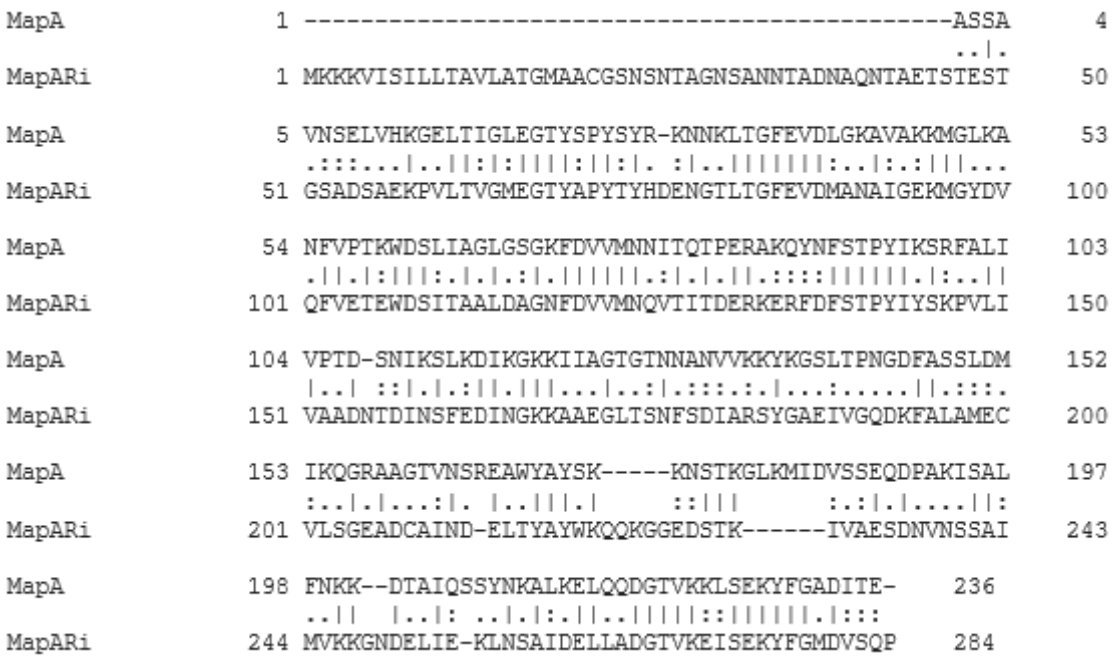


Figure 5.7 **Amino acid sequence alignment of the MapA protein from *Lactobacillus reuteri* 104R and the $MapA_{Ri}$ protein from *Roseburia intestinalis* M50/1**. The first 25 amino acid residues encode a sec leader sequence. Amino acid residues 42-258 encode the Bacterial extracellular solute binding protein domain (SBP_bac_3).

Table 5.8 ***In silico* adhesin search results against 54 of the most abundant microbial genomes of the human gut**. Candidates with high BLAST maximum scores (>200) are colored red. Candidates with 80-200 are colored in pink. NSSFID = No Significant Similarity Found in the Database. Hits with an e-value <1e-15 were considered putative adhesive homologs. The number of homologous proteins detected in each species are derived from multiple strains.

Gut bacterial species	MapA	Msa	Mub	SusD	LspA
Alistipes putredenis	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Anaerotruncus colihominis	4[83.2]	NSSFID	NSSFID	NSSFID	NSSFID
Bacteroides caccae	NSSFID	NSSFID	NSSFID	2[261]	NSSFID
Bacteroides capillosus	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Bacteroides dorei	NSSFID	NSSFID	NSSFID	14[346]	NSSFID
Bacteroides eggerthii	NSSFID	NSSFID	NSSFID	6[219]	NSSFID
Bacteroides finegoldii	NSSFID	NSSFID	NSSFID	6[285]	NSSFID
Bacteroides fragilis 3_1_12	NSSFID	NSSFID	NSSFID	2[293]	NSSFID
Bacteroides intestinalis	NSSFID	NSSFID	NSSFID	12[569]	NSSFID
Bacteroides ovatus	NSSFID	NSSFID	NSSID	18[576]	NSSFID
Bacteroides stercoris	NSFFID	NSSFID	NSSFID	6[270]	NSSFID
Bacteroides thetaiotaomicron	NSSFID	NSSFID	NSSFID	5[1150]	NSSFID
Bacteroides uniformis	NSFFID	NSSFID	NSSFID	23[564]	NSSFID
Bacteroides vulgatus	NSFFID	NSSFID	NSSFID	15[343]	NSSFID
Bacteroides xylanisolvens	5[100]	NSSFID	NSSFID	9[575]	NSSFID
Bifidobacterium bifidum	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium adolescentis	7[87.8]	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium longum	6[143]	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium infantis	4[143]	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium lactis	4[88.6]	NSSFID	NSSFID	NSSFID	NSSFID
Blautia hansenii	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Butyrivibrio crossotus	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium asparagiforme	1[100]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium leptum	1[110]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium nexile	1[85.9]	NSSFID	NSSFID	NSSFID	NSSFID

Clostridium scindens	3[82]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium sp. L2-50	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Collinsella aerofaciens	4[98.4]	NSSFID	NSSFID	NSSFID	NSSFID
Coprococcus eutactus	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Coprococcus comes SL7 1	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Dorea formicigenerans	2[89]	NSSFID	NSSFID	NSSFID	NSSFID
Dorea longicatena	14[87.8]	NSSFID	NSSFID	NSSFID	NSSFID
Enterococcus faecalis TX0104	5[98.6]	NSSFID	NSSFID	NSSFID	NSSFID
Escherichia coli	100[146]	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium hallii	7[99.8]	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium rectale M104/1	NSSFID	NSSFID	NSSFID	NSSFID	NSFFID
Eubacterium siraeum 703	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium ventriosum	3[102]	NSSFID	NSSFID	NSSFID	NSSFID
Faecalibacterium prausnitzii	3[149]	NSSFID	NSSFID	NSSFID	NSSFID
Holdemania filiformis	1[147]	NSSFID	NSSFID	NSSFID	NSSFID
Lactobacillus reuteri	41[530]	26[176]	67[6640]	NSSFID	NSSFID
Lactobacillus reuteri 1063	3[526]	5[126]	6[6640]	NSSFID	NSSFID
Lactobacillus plantarum WCFS1	3[257]	2[2041]	3[408]	NSSFID	2[149]
Lactobacillus salivarius UCC118	5[286]	NSSFID	NSSFID	NSSFID	1[2403]
Parabacteroides johnsonii	NSSFID	NSSFID	NSSFID	1[104]	NSSFID
Parabacteroides merdae	NSSFID	NSSFID	NSSFID	1[110]	NSSFID
Roseburia intestinalis M501	1[176]	NSSFID	NSSFID	NSSFID	NSSFID

Ruminococcus bromii L2-63	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus lactaris	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus obeum A2-162	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus sp. SR15	1[86.7]	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus torques L2-14	NSSFID	NSFFID	NSSFID	NSSFID	NSSFID
Streptococcus thermophilus	42[122]	1[169]	25[253]	NSSFID	NSSFID
Subdoligranulum variabile	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID

Table 5.9 **Homology of reference adhesins to 54 of the most abundant microbial genomes of the human gut**. Candidates with high BLAST maximum scores (>200) are colored red. Candidates with scores of 80-200 are colored in pink. NSSFID = No Significant Similarity Found in the Database. Hits with an e-value <1e-15 were considered putative adhesin homologs.

Gut bacterial species	MapA	Msa	Mub	SusD	LspA
Alistipes putredenis	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Anaerotruncus colihominis	4[2e-19]	NSSFID	NSSFID	NSSFID	NSSFID
Bacteroides caccae	NSSFID	NSSFID	NSSFID	2[9e-81]	NSSFID
Bacteroides capillosus	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Bacteroides dorei	NSSFID	NSSFID	NSSFID	14[3e-113]	NSSFID
Bacteroides eggerthii	NSSFID	NSSFID	NSSFID	6[4e-65]	NSSFID
Bacteroides finegoldii	NSSFID	NSSFID	NSSFID	6[2e-90]	NSSFID
Bacteroides fragilis 3_1_12	NSSFID	NSSFID	NSSFID	2[2e-93]	NSSFID
Bacteroides intestinalis	NSSFID	NSSFID	NSSFID	12[0.0]	NSSFID

Bacteroides ovatus	NSSFID	NSSFID	NSSFID	18[0.0]	NSSFID
Bacteroides stercoris	NSSFID	NSSFID	NSSFID	6[8e-34]	NSSFID
Bacteroides thetaiotaomicron	NSSFID	NSSFID	NSSFID	5[0.0]	NSSFID
Bacteroides uniformis	NSSFID	NSSFID	NSSFID	23[0.0]	NSSFID
Bacteroides vulgatus	NSSFID	NSSFID	NSSFID	15[5e-112]	NSSFID
Bacteroides xylanisolvens	5[4e-24]	NSSFID	NSSFID	9[0.0]	NSSFID
Bifidobacterium bifidum	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium adolescentis	7[1e-20]	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium longum	6[1e-44]	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium infantis	4[1e-44]	NSSFID	NSSFID	NSSFID	NSSFID
Bifidobacterium lactis	4[2e-21]	NSSFID	NSSFID	NSSFID	NSSFID
Blautia hansenii	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Butyrivibrio crossotus	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium asparagiforme	5[9e-26]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium leptum	1[9e-30]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium nexile	1[9e-21]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium scindens	3[2e-19]	NSSFID	NSSFID	NSSFID	NSSFID
Clostridium sp. L2-50	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Collinsella aerofaciens	4[5e-25]	NSSFID	NSSFID	NSSFID	NSSFID
Coprococcus eutactus	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Coprococcus comes SL7 1	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Dorea formicigenerans	2[1e-21]	NSSFID	NSSFID	NSSFID	NSSFID
Dorea longicatena	14[9e-21]	NSSFID	NSSFID	NSSFID	NSSFID
Enterococcus faecalis TX0104	5[2e-25]	NSSFID	NSSFID	NSSFID	NSSFID

Escherichia coli	100[3e-42]	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium hallii	7[1e-25]	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium rectale M104/1	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium siraeum 703	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Eubacterium ventriosum	3[6e-27]	NSSFID	NSSFID	NSSFID	NSSFID
Faecalibacterium prausnitzii SL3 3	3[5e-45]	NSSFID	NSSFID	NSSFID	NSSFID
Holdemania filiformis	1[3e-44]	NSSFID	NSSFID	NSSFID	NSSFID
Lactobacillus reuteri	41[0.0]	26[1e-50]	67[0.0]	NSSFID	NSSFID
Lactobacillus reuteri 1063	3[0.0]	5[1e-30]	6[0.0]	NSSFID	NSSFID
Lactobacillus plantarum WCFS1	3[4e-87]	2[0.0]	3[2e-115]	NSSFID	2[2e-37]
Lactobacillus salivarius UCC118	5[8e-99]	NSSFID	NSSFID	NSSFID	1[0.0]
Parabacteroides johnsonii	NSSFID	NSSFID	NSSFID	1[3e-24]	NSSFID
Parabacteroides merdae	NSSFID	NSSFID	NSSFID	1[3e-26]	NSSFID
Roseburia intestinalis M501	1[2e-55]	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus bromii L2-63	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus lactaris	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus obeum A2-162	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus sp. SR15	1[5e-21]	NSSFID	NSSFID	NSSFID	NSSFID
Ruminococcus torques L2-14	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID
Streptococcus thermophilus	42[6e-34]	1[2e-48]	25[7e-69]	NSSFID	NSSFID
Sobdoligranulum variabile	NSSFID	NSSFID	NSSFID	NSSFID	NSSFID

5.7.2 BLAST homology search using Msa reference protein

Msa is a lectin-like mannose specific adhesin (1010 aa) that originates from *Lactobacillus plantarum* WCFS1[356]. Homologous Msa proteins were detected in 4 of the 54 individual gut bacterial genomes using a BLASTp search with an expectation threshold of 1e-15 (Table 5.8). The Msa adhesin itself was detected as one of two matches with a bit score of 2041 in a BLASTp search of *Lactobacillus plantarum* WCSF1. The second hit, adherence-associated mucus-binding protein, LPXTG-motif cell wall anchor (sequence ID: WP_011101323.1), produced a bit score of 93.2 and an expectation value of 3e-20. It exhibited a 29% sequence identity spanning the 4 MucBP regions of the Msa query sequence. Therefore, the conserved MucBP sequence confers homology of this protein to Msa adhesin. Twenty-six different homologs were identified in 26 different *Lactobacillus reuteri* strains as the output for the BLASTp Msa query sequence (Table 5.8). The top hit exhibited a 59% identity spanning the 4 MucBP domains present in the Msa target gene (Appendix 2).

Five different homologs were identified from five different strains as an output from BLASTp of the Msa query sequence against *Lactobacillus reuteri* 1063. The top match in *Lactobacillus reuteri* 1063 produced a bit score of 126 and an expectation value of $1e^{-30}$. 31 percent sequence identity of the homolog spanned 3 MucBP domain repeats of the Msa query sequence. A single hit to the Msa query sequence was detected in *Streptococcus thermophilus,* hypothetical protein (sequence ID: WP_065473344.1) with a bit score of 169 (Table 5.8) and an expectation value of 2e-48 (Table 5.9). A 57% sequence identity spanning the 4 MucBP regions of the Msa query sequence (Appendix 2).

5.7.3 BLAST homology search using Mub reference protein

Mub is a mucus binding protein (3269 amino acids) that originates from *Lactobacillus reuteri* 1063[316]. Homologous Mub proteins were detected in 4 of the 54 individual gut bacterial genomes at an expectation threshold of 1e-15. The Mub adhesin itself was detected as a hit (sequence ID: AAF25576.1) with a bit score of 6640 amongst 5 other hits from five different *L. reuteri* strains. 3 different homologous matches from three different *L. plantarum* strains were observed in the BLASTp search of *Lactobacillus plantarum* WCSF1 using the Mub query sequence. The top match, mucus-binding protein (sequence ID: WP_011101486.1) produced a bit score of 408 (Table 5.8) and

an expectation value of 2e-115 (Table 5.9). A 30% sequence identity was observed spanning all 14 MucBP proteins in the query sequence. In contrast, one of the three matches of the Mub query in *Lactobacillus plantarum* WCSF1 (sequence ID: WP_011101323.1) exhibited a 32% sequence identity spanning only 2 MucBP domain regions (Appendix 3).

25 different homologous proteins were detected in 25 different *Streptococcus thermophilus* strains as the BLASTp output of the Mub query sequence against the *Streptococcus thermophilus* organism. The top match was a YSIRK signal domain/ LPXTG anchor domain surface protein (sequence ID: WP_064411033.1), which produced a bit score of 253 (Table 5.8) and an e-value of 7e-69 (Table 5.8). It exhibited a 31 percent sequence identity throughout the entire Mub protein, spanning only 7 of the 14 MucBP domain repeats of the Mub protein sequence.

5.7.4 BLAST homology search using SusD reference protein

SusD is a starch binding protein (551 amino acids) that originates from *Bacteroides thetaiotaomicron* VPI_5482. Of the five reference adhesins, only SusD is derived from a Gram-negative, non-lactic acid organism (Table 5.4). The four other adhesins are derived from different species of lactic acid (Gram-positive) bacteria (Table 5.4). SusD homologous proteins were detected in 14 of the 54 individual gut bacterial organisms. All 14 genomes belonged to Gram-negative Bacteroides and parabacteroides species. As expected, the highest BLASTp bit score match (1150) to the SusD query sequence was detected in the *Bacteroides thetaiotaomicron* VPI-5482 strain (Table 5.8). This match is indicative of the SusD query adhesin itself. The next highest bit score (576) match was detected amongst 18 different matches from 18 different strains of *Bacteroides ovatus*. As shown in Appendix 4, the hit with the highest bit score (Starch-binding outer membrane lipoprotein SusD) exhibited a 54% percent sequence identity throughout the entire SusD query amino acid sequence residues (1-550 amino acid residues).

The lowest BLASTp bit score match (104) to the SusD query sequence was detected in *Parabacteroides johnsonii*. In spite of the lower bit score and higher expectation value, the 28% percent sequence identity of RagB/SusD family nutrient uptake outer membrane protein (WP_008145839.1) spans the SusD domain region of the SusD query sequence (Appendix 4) indicating conservation of the SusD domain in the

homologous protein. Overall, based on the bit scores, expectation values and alignment region, homology was inferred for all the hits in the 14 Gram-negative gut micro-organisms indicating that the SusD domain is conserved in these.

5.7.5 BLAST homology search using LspA reference protein

LspA is a Lactobacillus surface protein A (1209 amino acids) derived from *Lactobacillus salivarius* UCC118. Only two of the 54 gut bacterial micro-organisms (*Lactobacillus plantarum* WCFS1 and *Lactobacillus salivarius* UCC118) produced matches against the BLASTp LspA query sequence at an expectation threshold of 1e-15 or below. As expected, the LspA adhesin itself was detected as a single match in *Lactobacillus salivarius* UCC118 with a bit score of 2403. A mucus-binding protein (seqID: WP_011101221.1) match was detected in *Lactobacillus plantarum* WCFS1 with a bit score of 149 (Table 5.8) and an expectation value of 2e-37 (Table 5.9). This protein exhibited a 25% percent sequence identity spanning all 8 MucBP (Mucin binding protein) regions of the LspA query sequence. In contrast, the second match in *Lactobacillus plantarum* WCFS1, mucus-binding protein (WP_011101804.1) (Appendix 5), exhibited a 31% sequence alignment identity spanning only 6 MucBP regions of the LspA query sequence. Therefore, homologous protein (WP_011101221.1) produced a higher bit score and spans all 8 MucBP domain regions of the LspA query sequence.

5.7.6 Overview

An important point considered in the BLAST searches was the availability of a fully annotated genome for each of the 54 individual gut genomes. Almost all of the 54 individual bacterial genomes considered possessed complete proteomes. The availability of a complete or incomplete proteome has a direct impact on whether a homologous protein will be detected and the number of possible homologs that can be identified. Only one species does not seem to be fully annotated with a complete proteome, *Blautia hansenii*. This may explain why no hits were detected in this organism using any of the 5 reference adhesins. Unsurprisingly we were unable to detect homologous proteins of gram positive adhesins in gram negative organisms with the exception of MapA in *B. xylanisolvens* and *E. coli*. Overall, the BLAST threshold was stringent enough to identify homologous proteins and ensure that the homologs detected were functional. A lot of distant homologs were not detected at the

chosen threshold which may explain the high number of NSSFID (no significant similarity found in the database) results obtained. Our choice of threshold was validated by the homologs identified. At a stringent expectation value of 1e-15, we were able to infer homology for all matches presented in Table 5.8. On the other hand, distantly related homologous sequences may fail to be detected because their similarity is not statistically significant. Therefore, we could have relaxed BLAST expectation value and identified more distant homologs. However, we did not do this because we wanted to ensure a high probability of functional match for subsequent studies.

At the end of the analysis, a candidate adhesin, here named *MapA*$_{Ri}$ (MapA homolog in *Roseburia intstinalis*; L-cystine ABC transporter, periplasmic L-cystine binding protein), was selected for functional analysis by cloning and subsequently expressing it in *Lactococcus lactis* using the nisin controlled gene expression system (NICE). The choice of a MapA homologous protein was based on two principal criteria. One of the contributors to the choice of MapA homolog in *Roseburia intestinalis* was the high bit score of 176 and low expectation value. Secondly, *Roseburia intestinalis* is a butyrate producing bacteria that confers numerous benefits to the host physiology, nutrition and health (section 5.3) and insights into its attachment mechanisms may have future applications in gastrointestinal health. The activity of the recombinant protein in *L. lactis* was analysed using an *in vitro* assay of bacterial adhesion to Caco-2 cells.

5.8 *In silico* analysis, amplification and cloning of a putative MapA homolog, *MapA*$_{Ri}$, a putative L-Cystine ABC transporter from *Roseburia intestinalis*.

The first step in cloning the putative MapA$_{Ri}$ adhesin was the *in silico* design of primers (Figure 5.13) before cloning into the expression shuttle vector (pPTPi). Appropriate restriction sites that are present on the plasmid (pPTPi) but absent from the target gene, were incorporated into the forward and reverse primers during primer design.

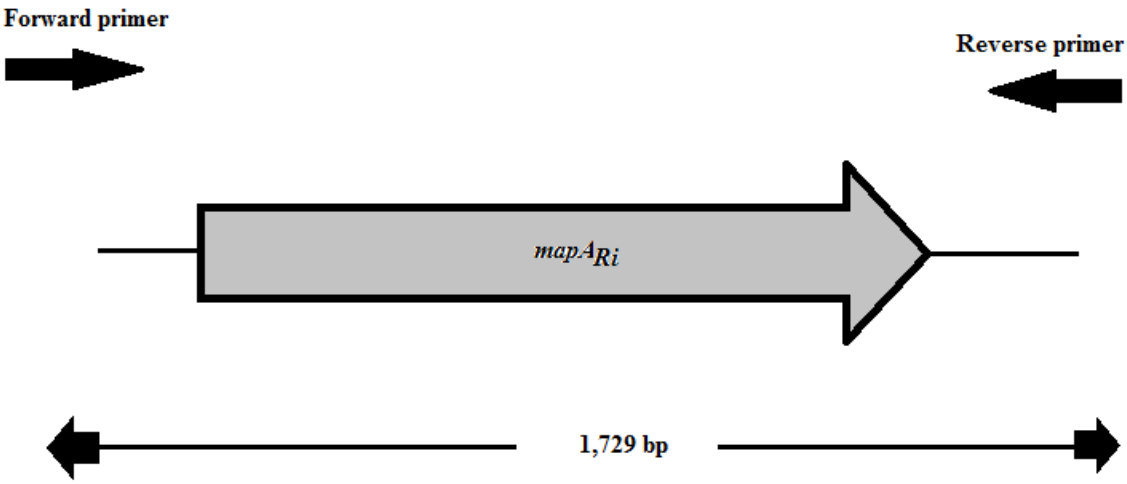


Figure 5.13 ***mapA*$_{Ri}$ primer design**. *MapA*$_{Ri}$ region demonstrating the designed upstream forward and downstream reverse primers.

As illustrated in Figure 5.13, a forward primer was designed 463bp upstream of the target gene and a reverse primer was designed 410bp downstream of the target gene. The primer was purposefuly designed to begin 463 bp upstream of the target gene in order to ensure the inclusion of all regulatory elements such as transcriptional elements upstream of the target gene. A *BamHI* restriction site was incorporated into the forward primer and an *EcoRI* restriction site was incorporated into the reverse primer. *Roseburia intestinalis* M50/1 was used as the template for the PCR amplification of *mapA*$_{Ri}$ (section 2.6.1). The final amplified PCR product of *mapA*$_{Ri}$ is 1,729 base pairs (Figure 5.14).

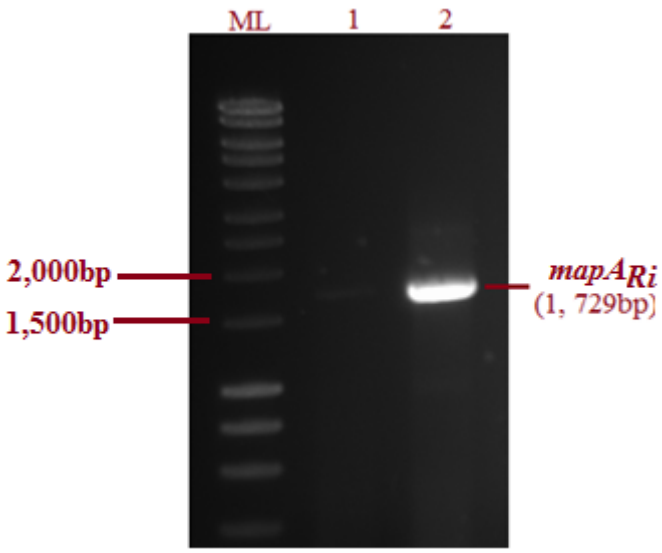


Figure 5.14 PCR amplification of *mapA*$_{Ri}$ from *Roseburia intestinalis* DNA template. 1% Agarose gel image of *mapA*$_{Ri}$ amplicon band in Lane 2 = 1,729bp. Lane ML = molecular ladder and Lane 1= faint *mapA*$_{Ri}$ band (section 2.6.1)

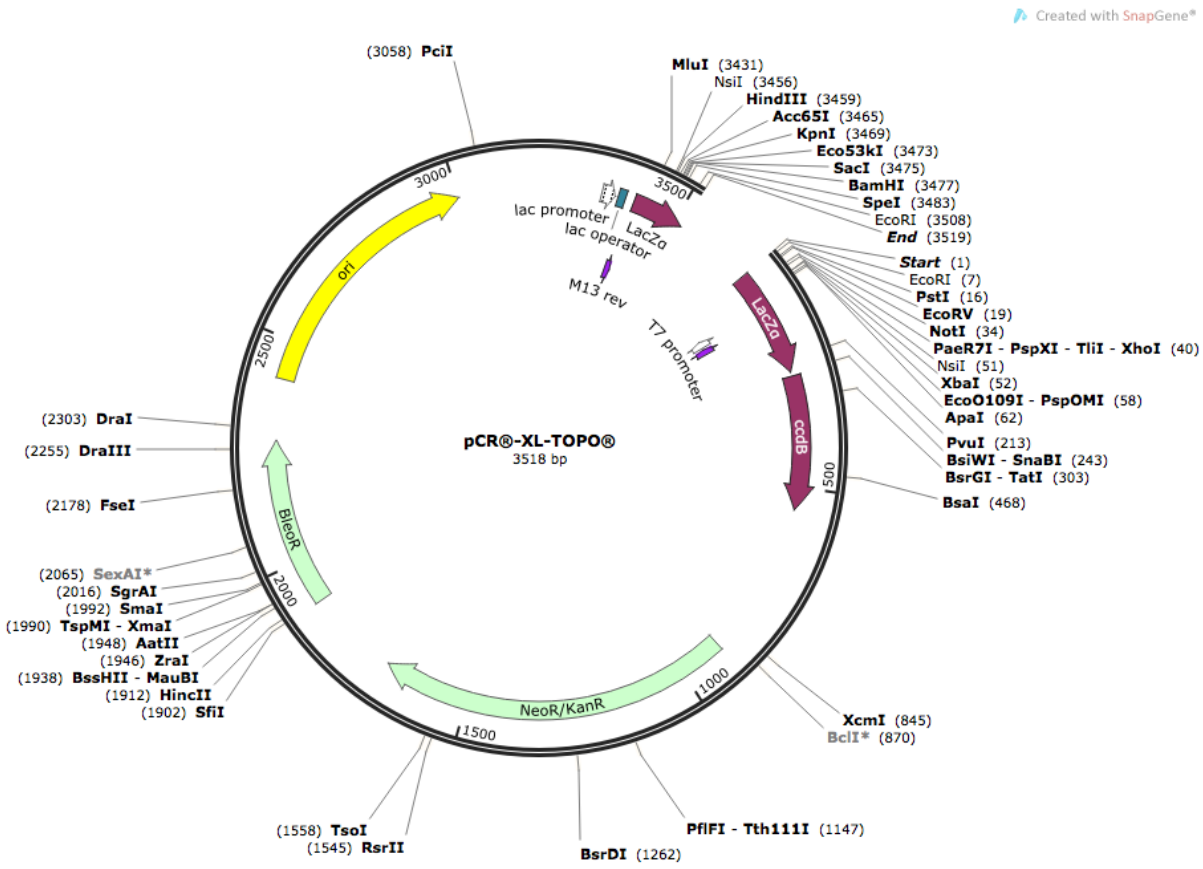


Figure 5.15 Diagram of TOPO cloning vector, pCR-XL-TOPO (section 2.1.10)

The TOPO cloning vector contains a multiple cloning site with *SacI* and *PstI* flanking the PCR product insertion site for easy excision and subcloning of the cloned insert (Figure 5.15). To effectively ligate $mapA_{Ri}$ gene to our cloning ready, linearized plasmid (pPTPi), it was important to ligate $mapA_{Ri}$ in the correct orientation to drive gene expression. The PCR product insert could be inserted into the plasmid in either of two orientations. One would put the $mapA_{Ri}$ gene in the correct orientation, and this would lead to the desired expression of the protein. However, the other orientation would lead to a non-expressed protein. Therefore, the TOPO clones were screened to determine which insert orientation was present in their plasmids. Before screening for the correct orientation of our $mapA_{Ri}$, we predicted the length of the restriction fragments in both the correct and incorrect orientations. The insert DNA was digested with a restriction enzyme (*NcoI*) that cuts only once within the $mapA_{Ri}$ sequence. Simultaneously, the TOPO plasmid was digested with *SacI* that cuts only once within the TOPO vector (Figure 5.15). The double-digested plasmid was run on an agarose gel (section 2.1.6). As predicted, cutting the pTOPO::MapA$_{Ri}$ clone with *NcoI* and *SacI* would yield a fragment of size 1096 bp band on the gel as well as a 4132 bp band (Figure 5.16A).

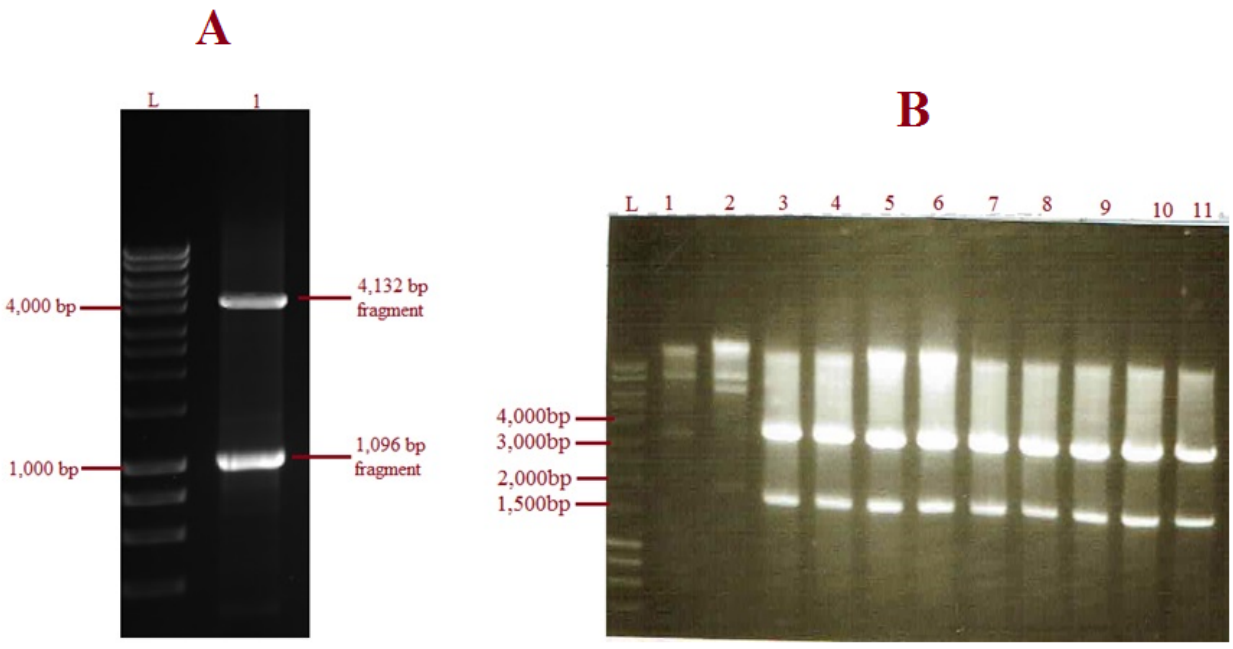


Figure 5.16 **Restriction digestion of pTOPO::MapA$_{Ri}$** (A) Restriction digest of the pTOPO::MapA$_{Ri}$ clone with *NcoI* and *SacI* yields two distinct bands of predicted sizes, 4,132bp and 1,096bp. This distinct digest pattern confirms that the MapA$_{Ri}$ insert in

TOPO is in the correct orientation required for gene expression to occur. Lane L = DNA marker (Hyperladder I). (B) The $MapA_{Ri}$ insert is excised from TOPO using *SacI* and *PstI*. The expected band sizes of 1,788 bp and 3,441 bp can be seen in lanes 3-11 on agarose gel.

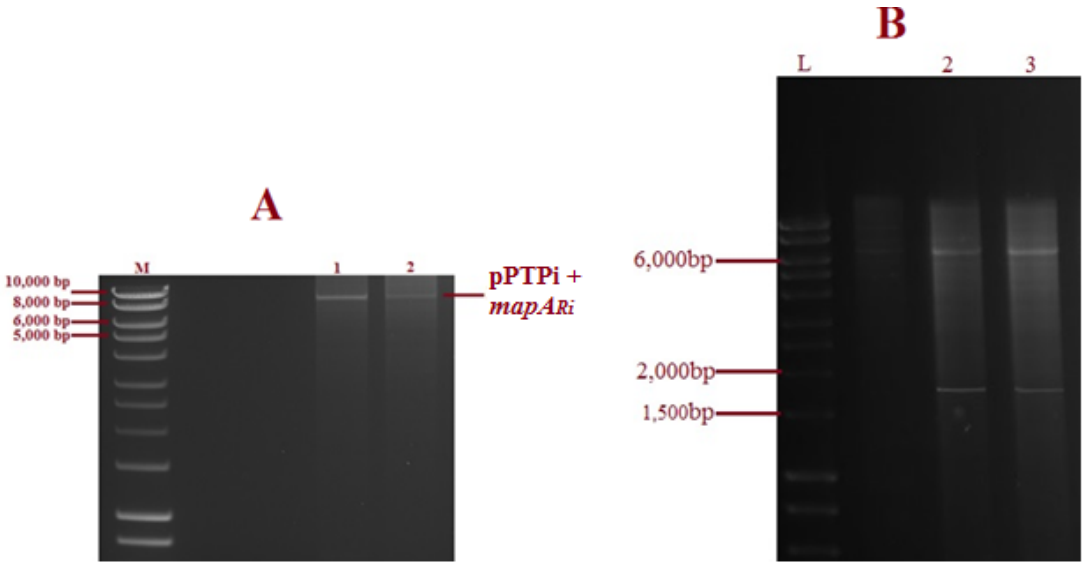


Figure 5.17 **1% Agarose gel electrophoresis** (A) A single restriction digest (*HindIII*) of the recombinant DNA, pPTPi ($MapA_{Ri}$) yields a single band of 8,556bp in lanes 1 and 2. Lane M = DNA marker (Hyperladder I) (B) A double-digest of the recombinant DNA using *BamHI* and *EcoRI* excised the $MapA_{Ri}$ insert from the pPTPi vector. The two distinct bands are visible on the B gel image. A vector band of 6,837 bp and an insert band of 1,729 bp are visible in lanes 2 and 3. Lane L = DNA marker (Hyperladder I)

After confirming the correct orientation of $mapA_{Ri}$, the insert was released by digesting the pTOPO::$MapA_{Ri}$ clone with *SacI* and *PstI* (Figure 5.16B). T4 DNA ligase was used to fuse the linearized pPTPi vector with the purified and cloning-ready $mapA_{Ri.}$ A standard ligation reaction (section 2.1.13) was performed with a recipient plasmid to insert ratio of approximately 1:3. After the ligation reaction, the ligation mixture was transformed into electrocompetent *E. coli* TOP10 cells (Section 2.2.4). Individual colonies were screened for successful ligations. The gene fragment, $mapA_{Ri}$, was inserted into the pPTPi plasmid, which was named pPTPi-$MapA_{Ri}$ (section 2.6.2). To identify whether the $mapA_{Ri}$ gene was inserted into pPTPi, PCR screening was performed on plasmid DNA of pPTPi::$mapA_{Ri}$ with the forward and reverse primers used to amplify $mapA_{Ri}$. Results showed that the expected DNA band (1,729 bp) of

the *mapA*$_{Ri}$ gene has been amplified. The DNA from a successful ligation was purified and a diagnostic restriction digest was performed with the enzymes used for cloning (*BamHI* and *EcoRI*). Figure 5.17B illustrates the digest run on an agarose gel with the two bands (vector and insert). Figure 5.17A illustrates the recombinant pPTPi::*mapA*$_{Ri}$ band at 8,556 bp in lanes 1 and 2. The results indicated that the target gene had been successfully recombined into pPTPi according to the previous design. The recombination expression system was named NZ9000/pPTPi-MapA$_{Ri}$. Electroporation was used to introduce the recombinant plamids (pPTPi::*mapA*$_{Ri}$) into *L. lactis* NZ9000 (section 2.6.2). The newly formed recombinant DNA clone (NZ9000/pPTPi-MapA$_{Ri}$) was used for functional analysis studies. MapA$_{Ri}$ protein expression was investigated in the recombinant strain by performing protein extraction and SDS-PAGE (section 2.6.3).

5.9 *In Vitro* Adhesion Assays of nisin induced expressed (NICE) *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ recombinant clone.

To determine if *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ recombinant clone is able to adhere to Caco-2 cells, an *in vitro* adhesion assay on 7 day old Caco-2 cells and 3 week-old Caco-2 cells (section 2.3.1a) was performed. 7 day old Caco-2 cells and 3 week-old Caco-2 cells differ in their level of differentiation as well as in the glycans and glycoproteins that are being expressed on their cell surfaces[190]. Previous adhesion assays have shown distinct differences in the level of adherence of metagenomic clones bound to 7 day old and 3 week-old Caco-2 cells (Sections 3.3.5 & 3.3.6).

Adherence assays were carried out by incubating Caco-2 cells with exponential phase bacteria at an MOI (multiplicity of infection) of 1000 for 1.5 h (section 2.3.1a). Exponential phase bacteria were used as high levels of protein are produced during this growth phase[357] and also to ensure a high ratio of live: dead bacterial cells. An MOI of 1000 was used as a suitable multiplicity required for efficient adherence for these bacteria.

The bar graph (Figure 5.18) represents the sum of three separate experiments each performed in triplicates. There was little difference between the adherence efficiency of *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ with nisin and *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ without nisin (Figure 5.18). The same pattern was observed with the adherence efficiency of *L. lactis* (pPTPi) with and without nisin. Overall, our results indicate that

L. lactis NZ9000/pPTPi-MapA$_{Ri}$ induced by nisin does not adhere to 7 day old Caco-2 cells to a greater extent than the control.

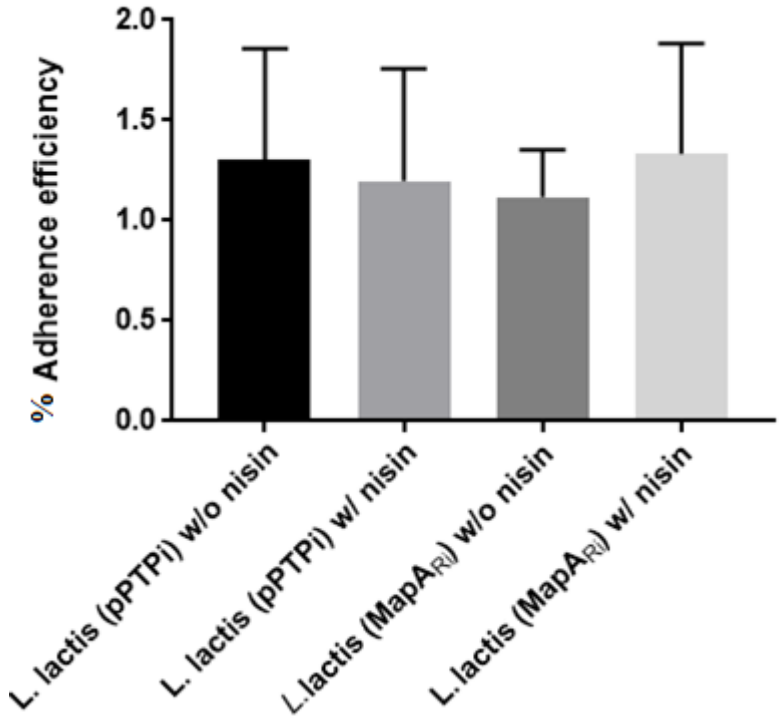


Figure 5.18 ***In vitro*** **adhesion assay of recombinant *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ on 7 day old Caco-2 cells**. Percent adherence efficiency of *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ and the control strain *L. lactis* (pPTPi) with and without nisin. The experiment was repeated three times on three different days. The error bars represent three individual experiments ran in triplicates. Significance was determined using Student's t-test.

A similar trend to the results in Figure 5.18 is also observed with 3 week-old Caco-2 cells (Figure 5.19). *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ induced with nisin does not significantly adhere to 3 week-old Caco- 2 cells. There is little difference between the adherence efficiencies of the controls and the recombinant strain. Overall, the induction of *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ with nisin did not significantly increase its adherence to 7 day old Caco-2 cells or 3 week-old Caco-2 cells. These results suggest that either (1) MapA$_{Ri}$ does not function as an adhesin, (2) or MapA$_{Ri}$ is not being expressed appropriately by the *L. lactis* NZ9000 heterologous host.

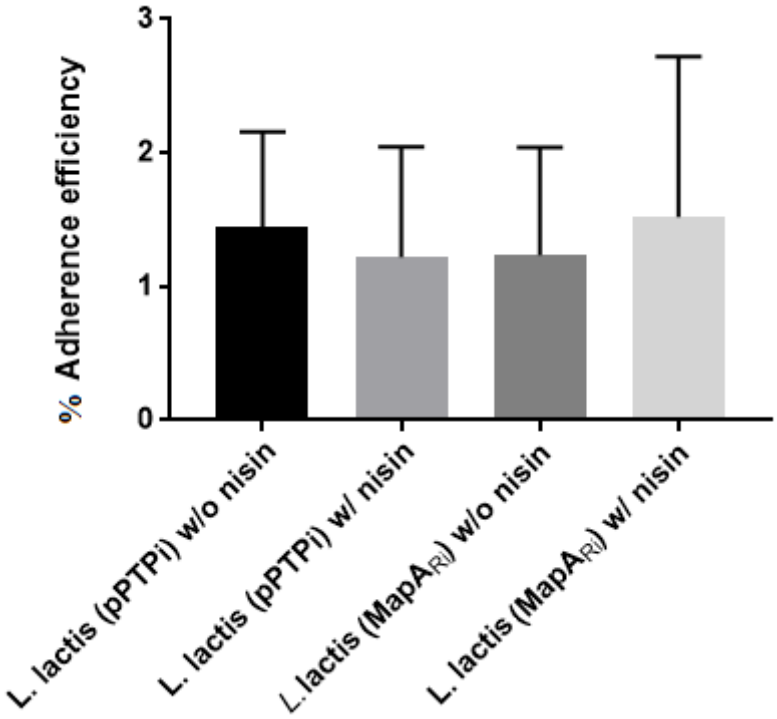


Figure 5.19 ***In vitro* adhesion assay of recombinant *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ on 3 week-old Caco-2 cells.** Percent adherence efficiency of *L. lactis* NZ9000/pPTPi-MapA$_{Ri}$ and the control strain with and without nisin on 3 week-old Caco-2 cells. The experiment was repeated 3 times on 3 different days. The error bars represent the standard deviation from the mean of three individual experiments each conducted in triplicates. Significance was determined using Student's t-test.

5.10 Discussion

The identification of homologous proteins in 54 specific gut microbial genomes was a successful alternative to mining human gut metagenomics databases online. When the human gut metagenomics databases failed to produce homologs, we decided to limit the BLAST search to 54 dominant gut microbial organisms[25]. Using this approach, we were able to detect several homologous proteins for each reference adhesin (Table 5.4). Homologous matches with high bit scores (>80) and very low expectation values (<1e-^{15}) reduced the chance of the match occurring randomly. Many of the detected homologous proteins possessed sequence similarity to crucial domains within the reference adhesins (MucBP, SBP_bac_3, SusD, and Lectin L-type) (Table 5.4). As mentoned previously, MapA is a 263 amino acid protein originating from *Lactobacillus reuteri* 104R. It contains one bacterial extracellular binding domain and is capable of binding to Caco-2 cells and mucus. Used as a query

sequence in BLAST searches (e-value <1e-^{15}, bit score >80), MapA query yielded homologous hits in 24 of the 54 gut bacterial genomes analysed. Of the 24 organisms with homologous matches, two of them originated from gram–negative organisms, namely *Bacteroides xylanisolvens* and *Escherichia coli*. The identification of MapA homologous proteins in 24 organisms of the 54 illustrates that the bacterial extracellular solute-binding domain is conserved in many gram-positive and some gram-negative species.

Although we selected just one MapA homologous protein (originating from *Roseburia intestinalis M50/1*) for further characterization, many other MapA homologous proteins could have been selected from the genomes of the 24 gut micro-organisms. For example, further characterization and analysis could have been performed with the homologous protein (*amino acid ABC transporter substrate-binding protein*) identified in *Faecalibacterium prausnitzii* with a bit score of 149 and an expectation value of 5e^{-45}. Over the past few years, an increasing number of studies have described the importance of this highly metabolically active commensal bacterium as a component of the healthy human microbiota. Changes in the abundance *Faecalibacterium prausnitzii* have been linked to dysbiosis in several human disorders[44]. Its low prevalence in many intestinal disorders, particularly in IBD patients, suggests its potential as an indicator of intestinal health. *F. prausnitzii* is a butyrate producer and has demonstrated anti-inflammatory effects *in vitro* and *in vivo* using a mouse colitis model making it a key member of the microbiota that may contribute to intestinal homeostasis. Modulation of *F. prausnitzii* abundance, for example using prebiotics and probiotics, might have prophylactic or therapeutic applications in human health. Therefore, characterization and analysis of a putative adhesin originating from *F. prausnitzii* could further knowledge and understanding of the health promoting effects of this important microbe.

Interestingly, a MapA homologous protein was identified in *Holdemania filiformis* at a bit score of 147 and an expectation value of 3e-^{44}. *Holdemania filiformis* was isolated from the faeces of healthy people. However, the significance or abundance of this organism in the intestinal microbiota remains unknown. Further characterisation and analysis of the homologous protein could shed light on the glycan-microbe interaction and adherence capacity of this gram-positive gut micro-organism.

MapA homologous proteins were also identified in *Bifidobacterium longum* and *Bifidobacterium infantis*. Studies have long extolled the benefits of Bifidobacterium species in the health of the human gut[358]. Additionally, MapA homologous proteins were identified in gram-positive commensal Clostridia species; namely *C. asparagiforme* and *C. leptum*. Contrary to popular notion, commensal clostridium species confer numerous health benefits to the host.They maintain overall gut function, as well as contribute in the unfavourable alteration of microbiota composition (dysbiosis) that has been implicated in several gastrointestinal disorders. The presence of a MapA homologous protein in the gram-negative *Bacteroides xylanisolvens* and clostridial species merits further analysis and characterization.

Msa (1 legume lectin domain and 4 MucBP domain), Mub (14 MucBP domain) and Lspa (8 MucBP domain) yielded homologous matches in 3 (Msa), 3 (Mub) and 1 (LspA) organisms respectively out of the 54 gut bacterial genomes analysed. The homologous proteins were identified in only gram-positive organisms demonstrating the MucBP mucus binding mechanism is not as common in gut as MapA binding mechanism. Moreover the MucBP mucin binding mechanism seems to be restricted to only a few species; Lactic acid bacteria.

SusD is not validated or characterized as a known adhesin. However, in this study, it was used as a reference protein to identify putative adhesins because it contains a starch binding outer membrane protein. Unsurprisingly, SusD homologous proteins were identified in 14 gram-negative organisms of the 54 gut organisms analysed (section 5.7.4).

The aim of this study was to clone the selected homologous protein into an expression vector for use in functional screens, we opted for a stringent expectation value threshold to increase the likelihood that detected homologs were functional. However, depending on the purpose of the search, it is possible to reduce the stringency of BLAST parameters to include distant homologous proteins.

After detecting numerous homologous proteins for each reference adhesin, a homologous protein hit (L-cystine ABC transporter, periplasmic substrate binding protein; dubbed $MapA_{Ri}$) of the MapA reference adhesin from *Roseburia intestinalis* M50/1 was chosen. As the MapA reference adhesin and $MapA_{Ri}$ (L-cystine ABC transporter, periplasmic substrate binding protein) are homologous they contain

similar domain content and structural structures (Figure 5.6). The MapA reference adhesin is the smallest reference adhesin (Table 5.4) with an amino acid length of 263 making it easier to work with using molecular cloning techniques. Furthermore, the MapA homologous hit in *Roseburia intestinalis* M50/1 exhibited the highest bit score (176) in a non-lactic acid bacteria. Another reason for the choice of $MapA_{Ri}$ is that *Roseburia intestinalis* M50/1 has received widespread attention for its benefit to human health, physiology and nutrition. Identifying an adhesin in this particular species could provide more insight into the health-promoting capacity of this important gut bacterium. BLAST is a powerful tool for detecting similarity. However, research has indicated that increased sensitivity can be obtained by using more sophisticated methods, such as more accurate algorithms like PSI-BLAST (position-specific-iterated BLAST) and the hidden Markov model (HMMs)[353]. Studies have shown that these methods have allowed for large improvements in the identification of homologous proteins compared to BLAST. A good principle of homology searching is to use more than one approach by trying different scoring matrices, databases, search parameters, and even algorithms[352].

The ability to mine the accumulating metagenomics data to uncover biological information is becoming increasingly accessible through various databases and annotation platforms. A considerable limitation in mining these datasets involved the lack of completely sequenced (complete proteome) genomes available in the databases. Some functional proteins are not being identified because of the dependence of current platforms on the already sequenced (well annotated) metagenomes of microorganisms. The existing metagenomic database data illustrates the resources available to microbiologists today. These resources are being constantly replenished by increasing data sets and sequenced gut metagenomes. Indeed, the currently available gut metagenomes represent only a fraction of their existing genetic potential. In the future, the continuous advances in sequencing technologies (and sequenced genomes, metagenomes) will generate not only more sequences, but longer sequences.

The recombinant strain *L. lactis* NZ9000/pPTPi-$MapA_{Ri}$ did not exhibit a statistically significant increase in adherence as compared to the control strain *L. lactis* (pPTPi). The age of the Caco-2 cells did not significantly impact adherence efficiency of the strains either. As a result, SDS-PAGE analysis was used to confirm the presence of

$MapA_{Ri}$ recombinant protein in *L. lactis* NZ9000. SDS-PAGE gel image quality was poor (data not shown). These results are surprising since inducible gene expression systems have proven to be very beneficial tools for achieving protein over production. *L. lactis* has been successfully used as a host organism for the overproduction of many heterologous proteins[359, 360].

The three main areas that require optimization of NICE for maximum protein yields are cell density, nisin-controlled induction & protein production and variables specific to the $MapA_{Ri}$ gene. Due to its fermentative metabolism, growth of *L. lactis* to cell densities far above 20 g l^{-1} dry biomass concentration are often not possible. In a simple culture of M17 medium, the maximum cell density that is often reached is about $0D_{600}$ = 3 (1 g l^{-1} dry cell mass). Growth stops when the pH reaches 5.0. However, with neutralization using NaOH or NH_4OH, the cell density can rise to $0D_{600}$ = 15 (5 g l^{-1} dry cell mass). To date, several attempts have been made to develop high cell density cultivation methods for lactic acid bacteria, but none have been applied to increase gene expression[196]. Researchers have shown that effective protein expression with the NICE system in *L. lactis* staunchly depends on the medium composition, the fermentation parameters and the amount of nisin added for induction. Studies have shown that the vigilant optimization of these key variables often leads to the notable increase in target protein yield[1].

Chapter 6:

General Discussion

6.1 Summary of Main findings

a) Functional metagenomic screening of a human gut fosmid library (42,000 clones) led to the identification of two putative fosmid clones, FC3 and FC21 that confer enhanced adherence capability to the host strain towards 3 week-old Caco-2 cells in the presence of antibiotic and L-arabinose inducer (Figure 4.7).

b) Next generation sequencing and bioinformatics analysis revealed that FC3 and FC21 clones are distinct sequences of 24.6 kb and 8.1 kb and include 26 and 3 protein-coding open reading frames (ORFs), respectively. The 8.1 kb FC21 fragment contains three functional genes and belongs to the dominant commensal gut species *Bifidobacterium adolescentis* (Figure 4.5, Table 4.11). Sequence analysis of FC3 revealed that the 24.6 kb insert is a fragment with no current known homologs in the database, suggesting that the insert DNA is derived from a microbe with an unknown genome sequence. A large portion of the FC3 predicted gene products were homologous to those of Clostridium spp. suggesting that FC3 originates from an unknown *Clostridium spp*. species (Table 4.12).

c) There are 11 putative membrane proteins encoded by the ORFs on the FC3 insert (Table 4.12) that could be possible adhesins responsible for the enhanced adherence to Caco-2 cells. Bioinformatic analysis of 5 of these 11 membrane proteins indicated possible links to adherence; (1) the putative collagen adhesion protein (Cna_B domain), (2) Sortase D, (3) Sortase B, (4) RTX xin and serine/threonine protein kinase. There are 4 hypothetical membrane bound proteins on FC3 that could also potentially encode adhesins.

d) All three transport genes present on the FC21 insert (Table 4.11, Figure 4.5) could possibly be part of a gene cluster which collectively share a generalized function. It is possible that all three genes together form an adhesive complex that is responsible for the enhanced adherence to Caco-2 cells. Nonetheless, studies have demonstrated the ability of cell surface located transport proteins to indirectly or directly influence adherence to host cells.

e) Lectin microarray analysis of FC3 and FC21 revealed that they have altered cell surface glycosylation as compared to the control EPI300 (pCC1FOS). FC3

and FC21 exhibited reduced binding to plant lectins that are specific for GlcNAc (N-Acetylglucosamine) compared to the control strain (Figure 4.9A).

f) Reduced GlcNAc residues on the cell surface of both FC3 and FC21 may influence the accessibility of previously hidden adhesins in binding to specific Caco-2 cell receptors.

g) According to section 4.10, both fosmid clones (FC3 & FC21) and the control (EPI300 (pCC1FOS)) produced similar levels of biofilm formation. Although FC3 and FC21 have altered cell surface glycosylation, they produced similar levels of biofilm formation as the control strain.

h) FC3 and FC21 did not exhibit altered mucin binding patterns compared to EPI300 (pCC1FOS). Clustering analysis of the mucin microarray data indicated 100% similarity in mucin binding of FC3, FC21 and control strain EPI300 (pCC1FOS) (Figure 4.11). This is not surprising since the Caco-2 cells don't produce mucin.

i) Neoglycoconjugate (NGC) analysis of FC3 and FC21 revealed that the average fluorescence intensity increased for FC3 and FC21 compared to the control strain. HCE clustering revealed a 69% similarity in NGC binding of FC3 and FC21 in the absence of antibiotic and L-arabinose (Figure 4.17). In contrast, with arabinose and antibiotic, the fluorescence intensity is primarily increased for clone FC3. HCE clustering analysis showed a 15% similarity in NGC binding for FC3 and FC21 and the control strain. Overall, the results indicate that the presence of antibiotic and arabinose promotes NGC binding of FC3 (Figure 4.15).

j) FC3 binds specifically to 14 neo-glycoconjugates on the microarray (Figure 4.17, red arrows), including 2'Fucosyllactose and N-acetylglucosamine (GlcNAc). Both 2'Fucosyllactose and N-acetylglucosamine play fundamental roles in gut glycan-binding microbe interactions as well as the health of the human gastrointestinal tract.

k) FC21 binds specifically to α_Crystallin (α_C), Asialofetuin (ASF), Lacto-N-neotetraose-APD-HAS (LNnTHSA), H-Type2-APE-HAS (H2HSA), RNAse B (RB) and Lacto-N-fucopentaose II-BSA (LNFPIIBSA).

l) The *in vitro* analysis of a putative bacterial adhesin ($MapA_{Ri}$) encoded by *Roseburia intestinalis* did not yield significant binding to both 7 day and 3-week old Caco-2 cells suggesting that (1) $mapA_{Ri}$ does not function as an

adhesin, (2) or $mapA_{Ri}$ is not being expressed appropriately by *L. Lactis* NZ9000 heterologous host.

6.1.1 Functional screening of a metagenomic library reveals clones with enhanced adherence to Caco-2 cells.

In this study, functional screening of a fosmid metagenomics library revealed two clones (FC3 and FC21) that exhibited enhanced adherence to 3 week old Caco- 2 cells in the presence of L-arabinose inducer. Both clones contained metagenomic DNA from two of the most abundant phyla in the gut; namely Actinobacteria and Firmicutes. Bioinformatic analysis of the complete insert sequence of FC3 and FC21 revealed that FC3 and FC21 clones are 24.6 kb and 8.1 kb and include 26 and 3 protein-coding open reading frames (ORFs), respectively. The 8.1 kb FC21 fragment contains three functional genes and belongs to the dominant commensal gut species *Bifidobacterium adolescentis*. Two of the three functional genes on FC21 have been previously annotated and play active roles in transport (Table 4.11). The hypothetical protein (BAD_0085) is predicted to encode a large transmembrane protein possibly involved in transport too.

FC21 encodes an aromatic amino acid transport protein *aroP*, also knowns as an ABC transporter permease, which promotes the transfer of amino acids from one side of a membrane to the other. Studies reveal that some ABC transporters can mediate or drive adhesion of bacterial cells to host epithelial cells[361, 362, 298, 363]. Jalalvand and colleagues[364] were the first to report an ABC transporter protein directly mediating bacterial adherence to host components. Their study revealed the mucosal pathogen *Haemophilus influenzae* (NTHi) ABC-transporter protein PF to be a novel laminin and cell-binding adhesin[364]. Similarly, one of the three functional proteins present on FC21 DNA could be exerting an influence on the adherence of FC21 to glycan epitopes on 3 week-old Caco-2 cells. Further sub-cloning and characterization is needed to establish an association between the three functional genes on FC21 and bacterial adherence to host cells.

FC21 also encodes a sodium-proton antiporter which regulates sodium concentrations and pH balance within cells[300]. These proteins convert the proton motive force to a sodium motive force for efflux of Na^+ ions. Interestingly, Deshpande and colleagues[365] reported the detection of an association between sodium-hydrogen antiporters and

bacterial adherence of *Campylobacter concisus* isolates to host cells. Searches for genes present in the *C. concisus* strains with high adherence and absent in those with low adherence identified the sodium-hydrogen antiporter NhaC. It is well known that the flagella of Campylobacter species, including *C. concisus*[366] play a major role in the adherence to host cells, and bacterial flagella are driven by a proton motive force[367]. Therefore, they hypothesized that the absence of NhaC from some *C. concisus* strains may influence the proton motive force, and thus, influence the strength of flagellar adherence to host microvilli. Overall, the authors conceded that further work was required to establish an association between sodium-proton antiporters and bacterial adherence to host cells[365]. A future perspective in this study would be to sub-clone the FC21 sodium-proton antiporter gene to confirm its functionality in adherence to Caco-2 epithelial cells.

Sub-cloning and functionally screening of each of the three genes on FC21 would reveal the gene responsible for the adherent phenotype of FC21. Although all three genes on FC21 have been previously annotated and shown to play roles in transport, a novel function of adherence could be assigned to the adherence-causing gene after further characterisation. Previous studies by Culligan and colleagues[168] identified five genes (which were previously annotated) from the human gut microbiome, to which a novel function of salt tolerance could be assigned based on the results of their functional screens[168]. However, as mentioned previously, it is highly likely that the three proteins on FC21 do not act as independent transporters but likely form a complex that is one functional transporter/adhesin.

The identification of an adherent FC21 clone with a metagenomic DNA fragment derived from *Bifidobacterium adolescentis* ATCC 15703 is of particular significance because this species is a known probiotic[368, 369]. *B. adolescentis* are gram-positive, non-motile anaerobic bacteria that are normal inhabitants of healthy human and animal intestinal tracts[267]. Colonization by *B. adolescentis* in the human gut occurs immediately after birth and tends to remain relatively stable until late adulthood. The ability to adhere to mucosal epithelial cells is one of the main selection criteria for probiotics[370, 371]. Adherent strains easily colonize the intestine, particularly the small intestine where flow rates are relatively high. Although probiotics do not colonize humans permanently, they need to persist in the gut long enough to exert beneficial properties to the host. Adhesive probiotic strains have been found to have a greater

effect on the host immune system than less adhesive strains[204]. Persistent probiotic strains may prevent the binding of some pathogenic organisms. Studies have shown that when probiotics have higher affinity for receptors than pathogens or when they are in higher concentrations than these, they are able to displace adhering pathogens[372]. Therefore, a major challenge of probiotic delivery to the host is the relatively transient colonization time in the gut. Elucidating novel bacterial-glycan binding capabilities of *B. adolescentis* could lead to the development of techniques to prolong persistence of probiotic strains *in vivo* so as to obtain the maximum benefits of these strains.

Sequence analysis of FC3 revealed that the 24.6 kb insert is a fragment with no current known homologs in the database, suggesting that the insert DNA is derived from a microbe with an unknown genome sequence. It is hoped that the identification of atypical genes which have not previously been linked to adherence will help broaden our understanding and possibly lead to the identification of novel and unusual systems that play as yet undefined roles in microbe-glycan adherence. Sequencing of the full fosmid insert (24.6 kb) of FC3 revealed an interesting gene landscape (Table 4.12), with approximately 40% of the predicted genes encoding proteins which shared the highest genetic identity to different species of Clostridium and 32% having no homologoues in the databases (data not shown). The Clostridium-associated proteins and the unknown proteins are interspersed with proteins associated with different phyla such as Firmicutes, Cyanobacteria and Bacteroidetes (data not shown). In the gut, commensal clostridia consist of gram positive, rod-shaped bacteria in the phylum Firmicutes that make up a substantial part of the total bacteria in the gut microbiota. Sub-cloning and functional screening of each of the 26 protein encoding genes present on FC3 may lead to the identification of a novel glycan binding adhesin found exclusively among the human gut microbiota. Interestingly, FC3 is predicted to encode two probable serine-threonine genes (Table 4.12). Studies indicate that a serine-threonine kinase demonstrated the ability to regulate expression of Pneumococcal pilus and modulate bacterial adherence to human epithelial and endothelial cells *in vitro*[277]. Moreover, FC3 encodes a putative collagen adhesion protein consisting of a Cna_B domain. This domain is found in the collagen-binding surface protein Cna of *Staphylococcus aureus*. This repeated B region domain does not mediate collagen binding, instead it appears to form a stalk that presents the adhesin ligand binding domain away from the bacterial cell surfaces[276]. Future sub-

cloning and functional metagenomic studies on FC3 will elaborate the role of each gene (26 protein coding genes) in conferring enhanced adherence capability to FC3 on 3 week-old Caco-2 cells.

The successful identification of FC3 and FC21 using functional metagenomics acknowledges the fact that the gut microbiome encodes adhesive factors of both known and unknown genes. It also highlights the fact that known and annotated membrane transport genes may behave as adhesins. Sub-cloning of hypothetical genes on FC3 and FC21 for further screening will provide more insight into the functional capacity of these hypothetical proteins so that they can be assigned functions and annotated in the future. The task of assigning novel functions can be achieved through functional screening of metagenomics libraries using activity-based assays. Studies indicate that approximately 30-40% of genes in a given genome are typically annotated as hypothetical, conserved hypothetical or function unknown[373], while ~75% of functions important for life in the gut consist of uncharacterized orthologous groups and/or completely novel gene families[25], highlighting the significant amount of novelty that exists in the gut metagenome. Mining gut microbiomes and the development of more sensitive and innovative screening assays will facilitate the discovery of novel adherence factor genes, antibiotics, biopharmaceuticals and biotherapeutics for use in biotechnology, medicine and health. The present study indicates that functional metagenomics is a useful tool in identifying novel genes. Functional metagenomics, unlike sequence-driven approaches, does not require that genes have homology to genes of known function and it offers the opportunity to add functional information to the nucleic acid and protein databases. The identification of FC3 of no known homologs in database highlights the diversty and novelty in gut microbiota.

6.1.2 Cell surface glycosylation is altered for FC3 and FC21

In this study, lectin microarrays were used for the glycomic analysis of intact fosmid clones FC3 and FC21 compared to a control strain. When interrogated on lectin microarrays, both fosmid clones (FC3 and FC21) exhibited reduced binding to plant lectins that are specific for GlcNAc (N-Acetylglucosamine) (Table 2.5) compared to the control EPI300 (pCC1FOS). One reason for the observed results in both clones is that FC3 and FC21 each have vectors carrying foreign metagenomic DNA inserts. The presence of metagenomics DNA inserts in the vector is sufficient to alter cell surface

glycosylation of both fosmid clones. The EPI300 control contains an empty vector and thus no alteration to cell surface glycans. These results corroborate the discovery that FC3 and FC21 exhibited enhanced adherence to 3 week-old Caco-2 cells when compared to the control strain. The results of the lectin microarrays suggests that FC3 and FC21 demonstrate enhanced adherence to Caco-2 cells due to alterations in their cell surface glycosylation. One of the ways in which the reduced abundance of GlcNAc residues on the bacterial cell surface of both clones may increase the bacteria's capacity to adhere to glycan receptors is by increasing the accessibility of bacterial cell surface adhesins to glycan epitopes on Caco-2 cells. The reduced GlcNAc residues could result in the exposure of previously shielded adhesin molecules. Complementary receptor and adhesin molecules must be accessible and arranged in such a way that they can contact and attach. Studies have demonstrated that alterations of the cell surface glycocalyx or glycosylation of a bacteria cell surface can regulate cell adhesion. Multiple studies using leukocytes have shown that decreases in glycocalyx thickness directly correlate with increased cell adhesion[374, 375]. Shedding and disruption of the glycocalyx increases the availability of ligands to bind to cell surface receptors. Similarly, based on experiments investigating the role of glycocalyx in leukocyte-endothelial cell adhesion, Mulivor and his colleague concluded that the glycocalyx serves as a barrier to adhesion and that its shedding during natural activation of endothelial cells may be an essential part of the inflammatory process[376]. Interestingly, Mitchell and colleague demonstrated that the sugar-rich glycocalyx coating expressed on the surface of cells can serve as a physical barrier to control the spacing and availability of sugar receptor-ligand interactions. Given that the glycocalyx layer can approach a thickness of 0.5 μm while receptors are mostly <100 nm in length, the glycocalyx can act to control receptor interactions with their respective ligands[377]. However, alterations in sheer stress and proteases can cause shedding and/or remodelling of the glycocalyx, increasing the number of available receptors to bind to adhesin ligands. Similarly, the reduction of the abundance of GlcNAc residues on the cell surface of FC3 and FC21 has increased the number of FC3 and FC21 cell surface adhesins available to bind to receptors on Caco-2 cells. This altered cell surface glycosylation may be attributed to (1) the stress of maintaining metagenomics DNA insert in the heterologous host or (2) the expression of metagenomics DNA genes that directly or indirectly alter cell surface glycosylation.

Lectin microarray technology can be adapted successfully to the analysis of intact bacteria to examine bacterial glycans in their native context on the cell surface. The analysis of whole cell bacteria avoids the destructive isolation procedures used in other analytical formats and allows direct analysis of bacteria after fluorescently labelling, without any further processing. The high-throughput aspect of lectin microarrays facilitates rapid assessment of bacterial glycosylation. The visual lectin binding pattern generated from fluorescent bacteria can be used to distinguish strains, providing a more quantitative and sensitive method for serotyping.

A future perspective is to confirm the specificities of these lectin-bacteria interactions by competitive carbohydrate inhibition studies. This would further confirm the glycan-binding specificities of the fosmid clones. Another future perspective of this study is the interrogation of a lectin microarray with cell extracts of 3 week-old Caco-2 cells to decipher the specific glycan epitopes present on the cells as ligands for the adherence factors of FC3 and FC21.

The advent of Lectin Microarray technology has generated numerous benefits in the study of microbe-host interactions. Few other techniques are able to study the large diversity of carbohydrate structures present on intact bacterial cell surfaces in a high-throughput fashion. In spite of its many positive attributes, the lectin microarray has several limitations. The technique is not quantitative neither does it allow for the determination of glycan structures like Mass Spectrometry[243]. That is, lectin microarrays do not accurately identify glycan structures, but rather obtain information about the functional glycans recognized by a group of lectins in a panel. The method is more efficiently applied for comparative purposes (e.g. differential profiling). Furthermore, the lectins used to generate the microarray dictate the range of carbohydrate structures that are analysed [240]. It is possible that some carbohydrate structures that are exclusive to bacteria may not be evaluated due to the absence of lectins that recognize these structures. There is a lack in the commercial availability of sugar binding proteins and lectins that are diverse in their recognition of sugar structures. All cellular glycomes are complex and dynamic in nature, so it is imperative that high density microarrays with a more diverse set of lectins are developed. Additionally, the addition of more human and animal lectins will empower this

technology and reveal subtle differences in glycan structures on cell surfaces. Another major limitation of the lectin microarray technology is the discovery that the majority of plant lectins present on the array are obtained from natural sources. These lectins have undergone post-translational modifications with carbohydrates, leading to potential false positives due to binding by bacterial lectins [59]. In summary, the lectin microarray platform is a promising glycomics technology. The ability of this technology to rapidly assess dynamic cell surface glycosylation has enabled researcher to monitor and obtain vast information about glycomes. Rapid assessment of bacterial carbohydrates empowers us to review how bacteria are able to modulate their surface glycans to establish cell-cell interactions and host-gylcan interactions. This technology has not only enabled researchers to analyze mammalian cell surface signatures but also capture selected glycosylation defective cell lines.

6.1.3 Mucin binding signature is the same for FC3 and FC21 compared to control.

The gastrointestinal tract of humans are rich sources of glycans[16, 254] due to the presence of highly glycosylated molecules in the mucus layer overlaying the epithelial cells. The expression and glycosylation of mucins differ depending on a number of factors, including the species, the location in the body, inflammation and the presence of microbes[378]. To further the study of bacterial interactions with complex glycans, a novel mucin microarray containing a wide range of natural mucins (Table 2.6), including those from a number of gastrointestinal sites in several animal species[249] was used to profile the interactions of FC3 and FC21 to mucins. Since FC3 and FC21 have altered cell surface glycosylation, it was hypothesized that both clones would exhibit distinct binding patterns and species tropism to the mucins on the microarray when compared to the control. Despite their altered cell surface glycosylation, results indicate that all three strains; FC3, FC21 and control bound all 35 mucins samples (Table 2.6) in the same way. The clustering analysis indicated 100% similarity in mucin binding between FC3, FC21 and EPI300 (pCC1FOS) (Figure 4.11B). This is not surprising as they were selected against a cell line that does not produce mucin.

Although FC3 and FC21 exhibited enhanced binding to 3 week-old caco-2 cells, Caco-2 cells do not form a mucus layer and therefore do not fully represent the intestinal epithelial cells *in vivo*. A mucin microarray analysis would be better preceded by

functional screens on epithelial cells that produce mucus (e.g., HT29-MTX-E12 cells) because mucins are the principal components of mucus. A combination of mucus-secreting cells, purified mucin as well as the novel mucin microarray platform would produce a broader picture of the mucin binding signatures of FC3 and FC21. New tools that enable scientists to study the interaction of bacteria with mucin oligosaccharides have become widely available. These include cell lines that produce adherent mucus layers[379, 380]. Naughton and colleagues[290] discovered that the production of mucus by HT29-MTX-E12 cells promoted higher levels of infection by *C. jejuni* and *H. pylori* than those for the non-mucus-producing parental cell lines.

The results of such screens would provide insight into how bacteria colonize mucosal surfaces. Understanding the molecular mechanisms that bacteria use to colonize the mucus layer is important, since such knowledge may suggest novel approaches for the prevention of colonization by pathogens or the encouragement of colonization by probiotics. An advantage of this technique is that binding chemistry of the microarray slides allows for the optimal presentation of the glycans, thereby maximizing the access of the clones to potential glycan receptors.

6.1.4 Identification of glycan-binding interactions of FC3 and FC21

FC3 and FC21 were interrogated on neo-glycoconjugate (NGC) microarrays to further characterize their glycan binding specificities. In the absence of antibiotic and L-arabinose, there is a 69% similarity in NGC binding of FC3, FC21 and the control strain (Figure 4.17) In the presence of antibiotic and arabinose, fluorescence intensity is primarily increased for clone FC3 (Figure 4.15). Therefore, the presence of antibiotic and arabinose (induction and stabilisation of plasmid) seems to promote NGC binding of FC3. FC3 bound specifically to (1) ovalbumin (Ov), (2) bovine transferrin (bovXferrin), (3) α-Crystallin (a-C) from bovine lens, (4) GlcNAc (GlcNAcBSA), (5) Lacto-*N*-fucopentaose I and (6) II (LNFPIBSA and LNFPIIBSA), (7) 3'Sialyl Lewis x-BSA (SLexBSA14), (8) 6-Sulfo Lewis x-BSA (6SuLexBSA), (9) 3-Sulfo Lewis x-BSA (3SuLexBSA), (10) Galα1,3Galβ1,4GlcNAc-HSA (GGGNHSA), (11) Manα1,3(Manα1,6)Man-BSA (M3BSA), (12) Tri-fucosyl-Ley-heptasaccharide-APE-HSA (3FLeyHSA), (13) Tri-Lex-APE-HSA (3LexHSA) and (14) 2'Fucosyllactose-BSA (3SFLBSA) (Figure 4.12) This raises the possibility that one of these glycans may be present on the cell surface of Caco-2 cells and mediating

the observed adhesion. Specifically, 3'Sialyl lewis X carbohydrate structures have been shown to be present on the surface of Caco-2 cells. The binding specificity of FC3 to GlcNAcBSA (N-acetylglucosamine) probe is biologically relevant because it suggests FC3 contains GlcNAc residues that serve as epitopes for glycan-binding interactions with microbial adhesins. GlcNAc is well-known for supporting the human body's creation of a healthy mucus layer in the gut. Studies have demonstrated that GlcNAc helps support the growth of beneficial gut bacteria like *Bifidobacterium bifidum*. *N*-acetyl-glucosamine containing oligosaccharides were first identified 50 years ago as the 'bifidus factor', a selective growth substrate for intestinal bifidobacteria. Further studies demonstrate that GlcNAc may improve immune function in patients with multiple sclerosis. While N-acetylglucosamine might benfit anyone with digestive problems, it looks to be promising for people suffering from inflammatory bowel disease. Patients with conditions like Crohn's disease and Ulceritive colitis have much thinner mucus barrier in the gastrointestinal tract. In recent study by Andy Zhu and colleagues (April 2015), patients with inflammatory bowel disease taking N-acetylglucosamine for 1 month had substantial improvement in their symptoms.

Moreover, FC3 exhibited binding specificity to the human milk oligosaccharide 2'Fucosyllactose. 2'Fucosyllactose is the most prevalent human milk oligosaccharide, making up 30% of all HMOs. Humans are unable to digest HMOs such as 2'Fucosyllactose, hence the majority of HMO's reach the gut, where they serve as food for desirable gut bacteria. This finding that FC3 binds specifically to 2'Fucosyllactose probes on a microarray is relevant to understanding glycan-microbe interactions beneficial to human health because studies have indicated that 2'Fucosyllactose is able to influence intestinal epithelial cell maturation *in vitro*[310], inhibit *Campylobacter jejuni*-induced inflammation in the intestinal mucosa. HMOs can improve the inner layer of the human gut, boost the immune system, and may be essential nutrient for brain development in babies[311]. Due to its structure, 2'Fucosyllactose binds detrimental bacteria and toxins to prevent them from binding to the baby's gut, decreasing the risk of infection. Further studies have demonstrated that HMO's in both infants and adults are highly specific in the way they modulate the microbiota[312]. The primary impacts are increases in certain Bifidobacterium species and the reduction in several undesirable bacteria. FC3 demonstrates binding

specificity to two other human milk oligosaccharides, namely Lacto-*N*-fucopentaose I and II. The ability to decipher the glycan-binding specificities of both FC3 and FC21 provides insight into the types of glycans and glycoproteins present on both Caco-2 cells and the human gastrointestinal epithelial cells. It also provides information on the type of adhesins present on the surface of the bacteria. This study demonstrated the usefulness of glycan microarrays in characterizing the glycan binding specificities of whole bacteria cells.

6.1.5 No change in biofilm formation for FC3, FC21 and control strain

Based on the observed enhanced adherence to 3 week-old Caco-2 cells and their altered cell surface glycosylation, it was hypothesized that FC3 and FC21 may exhibit altered biofilm formation as compared to the control strain. The results indicated that FC3 and FC21 exhibited similar levels of biofilm formation to that of the control strain at 0.7 (Figure 4.10). The weak biofilm formation of FC3, FC21 and EPI300 (pCC1FOS) suggest that the adherence factors present on the cell surface of FC3 and FC21 may be specific for receptors present on host cells rather than abiotic surfaces. Bacterial cells have developed a series of surface adhesins promoting specific or non-specific adhesion under various environmental conditions. While adhesion to abiotic surfaces is usually mediated by non-specific interactions, adhesion to biotic surfaces (e.g., Caco-2 cell surface) typically requires a specific receptor-ligand interaction[381]. Non-specific adhesins are primarily responsible for biofilm formation and bacterial adhesion to abiotic surfaces. The lack of differences in biofilm formation of FC3 and FC21 suggests that they only express adherence factors that promote specific adhesion to Caco-2 cell glycan epitopes. The next sections will highlight some of the main challenges, limitations, potential solutions and future perspectives of this study.

6.1.6 Adherence of *L. lactis* NZ9000/pPTPi-$mapA_{Ri}$ to Caco-2 cells is not different to control.

In this study, the nisin-controlled expression system of Lactococcus lactis was selected as a vector to express the $mapA_{Ri}$ gene derived from the butyrate-producing bacteria, *Roseburia intestinalis M50/1*. Food-grade recombinant *L. lactis* NZ9000/pPTPi-$mapA_{Ri}$ was successfully constructed. The results from the *in vitro* bacterial adhesion assay using the recombinant strain (*L. lactis* NZ9000/pPTPi-$mapA_{Ri}$) indicated no significant binding to both 7 day-old and 3 week-old Caco-2 cells in the presence and

absence of nisin (Figure 5.18 & 5.19). These results suggest that (a) the recombinant *L. lactis* NZ9000/pPTPi-*mapA*$_{Ri}$ does not function as an adhesion or (b) the recombinant *L. lactis* NZ9000/pPTPi-*mapA*$_{Ri}$ is not being expressed appropriately by the *L. Lactis* NZ9000 heterologous host.

6.1.7 Advances in knowledge of gut-microbe interactions

One of the main findings of biological significance in this study is the altered cell surface glycosylation observed on the surfaces of both FC3 and FC21. The abundance of cell surface GlcNAc residues on the surfaces of FC3 and FC21 was considerably less than the abundance of GlcNAc residues present on the surface of the control EPI300 strain (Figure 4.9A). As a result, both clones (FC3 & FC21) exhibited enhanced binding to Caco-2 epithelial cells (Figure 4.7). Large GlcNAc residues could potentially have blocked relevant adhesins from finding and binding to receptors on the surface of epithelial cells. Therefore, the reduction of GlcNAc residues on the surface of FC3 and FC21 may have enhanced binding to Caco-2 cells by increasing the accessibility of previously 'hidden' adhesins to receptors on the surface of Caco-2 cells. This may explain why EPI300 control strain with intact cell surface GlcNAc residues exhibits significantly lower adherence than the two clones (Figure 4.7)

Additional results in this study indicate that FC3 and FC21 have unique glycan specificities. This finding validates the utility of carbohydrate-based microarrays in identifying the main epitopes recognized by FC3 and FC21, showcasing the power of this technology to rapidly identify binders for unknown clones.

6.2 The Challenges: Metagenomic Libraries

The overall aim of this study was to identify and characterize novel glycan binding bacterial adhesins encoded by the human gut metagenome. A metagenomic library consisting of 42,000 clones in the surrogate host Phage T-1 Resistant EPI300™-T1^{R} *Escherichia coli* was generated from the fresh faecal sample of a female volunteer on a western diet. Although it was easy to obtain fresh faecal samples, the information obtained from them often does not give an accurate representation of the complete picture within the gut[382]. Several studies have shown that the small intestine of the gastrointestinal tract contains a very different abundance and composition of gut microbiota, with more dynamic variation compared to the distal colon. The microbiota composition of the colon are influenced greatly by the efficient degradation of

complex indigestible polysaccharides. In contrast, the composition of the microbiota of the small intestine is shaped by its capacity for the rapid import and conversion of small carbohydrate and the fast adaptation to nutrient availability. Therefore, faeces may not be an ideal representative of the GI tract. It mainly serves to exhibit a snapshot of the diversity within the large intestines[382].

Another potential limitation in this study was the size of the gut metagenomic libraries constructed. The library size required to obtain sufficient coverage of the metagenome of even the simplest community presents a significant challenge for screening. One of the main limitations of library construction (small fragment library and fosmid library) is that the library invariably represents only a portion of the human gut microbiota. Because members of a community are not equally abundant, it is likely that a metagenomic library of minimum coverage would only represent the genomes of the most abundant species. To obtain substantial representation of rare members (<1%) of the community, the library would likely need to contain 100- to 1000-fold coverage of the metagenome[147]. Cosmid megalibraries have been created from soil which contain >1.5 x 10^7 unique clones[383]. At clone densities such as this, it is believed the library approaches saturation[384]. Most of the reported large-insert metagenomics libraries contain fewer than 100,000 clones and are several orders of magnitude too small to capture the entire microbial diversity present in the complex, communities they represent. Although increasing the library size is a worthy goal, existing libraries have provided useful insights into the microbial ecology of several ecosystems in the absence of complete metagenome coverage.

In this work, two clones (FC3 and FC21) with enhanced adherence to Caco-2 cells were successfully identified from the fosmid metagenomic library. However, constructing metagenomic libraries from a complex ecosystem like the human gut has many technical challenges. The determination of target insert size, cloning vector, and minimum number of library genes is governed by the type of genes that are sought and the complexity of the microbial community. Based on the formula presented in section 3.1.1, an ideal, metagenomics library that consist of 285,333 fosmid clones (DNA insert = 40Kb) was required to obtain representative coverage of the gut microbiota, assuming ~ 1000 species in the human gut. This number is seven fold larger than the library constructed (42,000 clones) and used for functional screens in this study. We

generated a library of considerably less clones. It is possible that a larger library would have led to more positive hits during the functional screening assays.

Therefore, to obtain high metagenomics coverage, a large quantity and high quality of DNA samples are crucial. Although precautionary steps are implemented, studies indicate that human contaminants are discovered in 50%-90% of sequences[385]. The variation of the different DNA extraction kits between laboratories also has an impact on the assessment of the human gut microbiota[386] Moreover, the comparison of data across studies that use different DNA extraction methods is wrought with difficulties[387]. A significant limitation to the success of metagenomics studies is the quality of functional annotations of metagenomics sequence fragments. Unfortunately, a large portion of data cannot be assigned a function due to lack of closely related hits in reference databases[25].

Due to the limitations of metagenomics, it is necessary to combine metagenomics and other microbiome approaches including cultivation methods, with a study of the metagenomics in the intestinal microbiome. This will ensure that the results are more accurate and convincing[388,389,2]. In order to completely overcome the limitations of metagenomics, it is important to generate a unified microbial DNA extraction method, improve computational algorithms, and complete the reference databases[2].

Bioinformatic analysis of the complete insert sequence of FC3 and FC21 revealed that FC3 and FC21 clones are 24.6 kb and 8.1kb and include 26 and 3 protein-coding open reading frames (ORFs), respectively. The DNA insert sizes of both clones are considerably smaller in size than the expected 40 kb carried by fosmids. This suggests poor library quality such as inefficient lambda packaging of sheared DNA fragment and poor DNA quality. If the majority of clones in the library contained inconsistent insert shorter fragments of DNA this could help to explain low hit rates.

6.2.1 Leveraging existing libraries

Two fosmid clones out of 42,000 demonstrated enhanced adherence capability on 3 week-old Caco-2 cells. This was not surprising since hit rates are traditionally low and vary widely. For example, in a search for lipolytic clones derived from German soil, only 1 in 730,000 clones showed activity[158]. However, studies have shown that several factors influence the hit rate such as the source of metagenomics DNA, the size of the gene of interest, its abundance in the metagenome and consequently the library, the

vector system and host of choice, the screen itself and the ability of the expression host to successfully express the gene[390]. Consequently, improving the hit rate requires modification and optimization of many of the factors mentioned. There are techniques that could have improved the observed low hit rates. One of the major ways of improving hit rates is by developing versatile vectors for library construction. In this study, the human gut metagenomic library was constructed using the cloning-ready commercial vector pCC1FOS (Chapter 3; Figure 3.3). Numerous other metagenomics libraries from diverse environments have also been constructed using the pCC1FOS vector. Despite its popularity, pCC1FOS has several disadvantages that render the constructed metagenomics library less versatile than they could be. For example, pCC1FOS does not contain an *ori*T which would allow it to be effectively transferred by conjugation to other species that may be more suitable for heterologous expression. To achieve conjugation capabilities, several scientists have added the RK2 *ori*T to pCC1FOS[391, 392, 393]. Even after library construction has occurred, some researchers have retrofitted individual pCC1FOS –based clones with an *ori*T[392]. These modifications demonstrate the need for fosmid vectors such as pCC1FOS to include the *ori*T so that duplication of work is avoided. It is possible that transformation can be used to transfer libraries to other hosts, but only for recipient hosts that are accommodating to those techniques who will not reject DNA that has been synthesized in *E.coli* due to the presence of host-restriction-modification systems[394]. Another possibility is to modify the host strains by deleting restriction-modification genes.

Another evident disadvantage of the pCC1FOS vector is that the *trf*A gene is not incorporated on the vector. Consequently, species that would otherwise be able to use *ori*V are unable to replicate pCC1FOS. Unsurprisingly, the vast majority of metagenomics studies using pCC1FOS are performed in *E. coli* as the screening host. This represents a tremendous disadvantage for functional metagenomics because different clones can be isolated from the same metagenomics library when different screening hosts are used[394, 208, 395]. Recently, Cheng[394] and colleagues discovered that using the legume-symbiont *Sinorhizobium meliloti* as a host results in a much greater diversity of clones that *E. coli* when screening their corn field soil metagenomics library for beta-galactosidase activity. The importance of developing systems that allow for the functional screening in diverse expression hosts has been studied by numerous scientific researchers[390, 151, 396].

Numerous metagenomics libraries (pCC1FOS or derivatives) that have already been constructed can be screened in non-*E. coli* hosts. They are accessible to any RK2-compatible host if a copy of the *trf*A gene is also made available. Aakvik and colleagues[397] successfully incorporated the *trf*A gene into the chromosome of *Gammaproteobacteria* species *Pseudomonas fluorescens* and *Xanthomonas campestris* for screening of libraries constructed using a pCC1FOS derivative[397]. Moreover, Westenberg and colleagues[398] were able to incorporate *ara*C-P_{BAD}-*trf*A into the *E.coli* EL350 chromosome to grant copy number inducibility to the lambda Red recombineering strain[398]. Overall, the ability to incorporate trfA into RK2-compatible species is a relatively uncomplicated procedure to expand the range of expression hosts for existing pCC1FOS-based libraries.

A substitute method to incorporating the *trf*A gene into the desired expression host is to alter the vector for integration into the host genome, eliminating the requirement for trfA[394]. Angelov and colleagues[399] employed this strategy to integrate clones into a target locus in the genome of the thermophile *Thermus thermophilus* for functional screening, by altering pCC1FOS to include a selectable marker and regions for homologous recombination[394 ,399]. In general, chromosomal integration is probably less useful than maintaining the clone due to the difficulty in recovering the integrated DNA manipulation, including DNA sequence analysis when libraries have been screened[394].

It might be beneficial to assemble a databank of gut colonization-associated genes/operons that have demonstrated functions in the gut. For example, in this study, the proteins encoded by the two clones that we have detected (FC3 and FC21) could be termed *Clostrid-Bifido* adhesion niche factors until they have been studied further. The FC3 DNA insert could be studied in terms of proteomics and molecular modelling so that its function can be defined. As demonstrated by our results, fosmid metagenomics libraries will play an important role in building such a databank. One of the most important factors in identifying putative adhesins is the functional screening method performed. The next section will describe strategies to improve heterologous expression.

6.2.2 Strategies to improve heterologous expression

One of the most common problems encountered in functional metagenomics is the effective expression of all foreign DNA in the expression host. In this study, one of the major limitations was that the metagenomics library was only screened in *E. coli*. Several attempts were made to produce the small fragment library in *L. lactis* MG1363 but only a limited number of transformants were obtained restricting the preparation of a metagenomic library in our gram positive host. *E. coli* has been the expression host of choice for the vast majority of functional metagenomics projects. This is not surprising since *E. coli* possesses a number of desirable attributes that make it the host of choice. It has a high transformation efficiency, is somewhat promiscuous with regard to the diversity of foreign expression signals it recognizes, lacks genes for restriction modification and homologous recombination and is capable of translating mRNA with diverse translation signals[400]. Despite these many advantages, *E. coli*, just like any expression vector, is unable to express all foreign DNA because of differences in transcriptional, translational and posttranslational machinery of the originating organism[401]. Potentially negative effectors of efficient expression are *cis*-acting factors such as promoters and ribosome binding sites (RBS) which are not compatible with the host machinery. Moreover, factors that are supplied in trans by the host cell such as chaperones, transcription factors and a compatible secretion systems are all potential barriers to efficient expression. Fortunately, mathematical formulae have been developed to predict the chance of a given gene of interest being expressed in *E. coli*. Research suggests that approximately 40% of genes are predicted to be functional in *E. coli*.

Two ways to improve heterologous expression include the use of alternative or dual hosts and modified vectors. The use of different cloning hosts or the use of one host to maintain the library followed by transfer of the library to a different host (from a different phylum or genera) for screening have been shown to be successful[400]. Gabor and colleagues described this as the "different host, different hit" effect[400]. Coupled with the use of alternative hosts is the use or creation of novel expression vectors that can function in multiple hosts or maximize the chances of expression in these hosts[401]. For example, Kakirde and colleagues[154] created a BAC vector that is capable of replication and expression in diverse group of gram-negative bacteria such as *Salmonella*, *Pseudomonas*, *Serratia*, *Enterobacter* and *Escherichia coli*[154]. More

positive hits would have been identified in this study had we been successful in transfering the metagenomics libraries into *Lactococcus lactis*.

Recently, viral gene elements have been employed to increase the expression and hit rate of metagenomic clones. Using phage T7 RNA polymerase to drive transcription and an inducible phage anti-termination protein to bypass many transcriptional terminators present on the insert, Medina and colleagues[393] noted a six-fold increase in the number of carbenicillin-resistant clones identified in their screens. Furthermore, vectors with dual orientation promoters that allow for bidirectional transcription and the possibility to significantly improve the hit rates are currently being used. The benefit of this system is that it will not rely on the presence of a native, insert-borne promoter, or on the orientation of the cloned insert[402].

Another strategy to improve heterologous expression is codon optimization. The majority of organisms have a codon usage bias; particular preference for certain translation initiation codons and for overall codon usage[403]. For example, *E. coli* uses AUG as start codon for nearly 90% of translation, thus non-AUG start codons such as GUG and UUG may not be recognized and processed effectively[390]. Studies have indicated that foreign genes are only highly transcribed in *E. coli* because they possess similar promoter sequences that bind the sigma factor RpoD (σ^{70}) of *E. coli*.

An in-depth understanding of codon usage bias will enable the design of the most suitable expression systems with the greatest chance of success and enable the creation of novel metagenome-derived synthetic genes or operons that are optimized for expression in *E. coli* or other relevant host[404], [405]. Overall, functional metagenomics screens are known to exhibit low hit rates due to these types of molecular barriers. The identification of obstacles to cloning and screening will equip researchers with the capacity to develop new tools and technologies for functional metagenomics[406], providing us with greater capacity in terms of what we are able to access from functional screens[401].

6.3 The issue of *in vitro* adhesion

The functional metagenomic screen used to identify novel genes from an ecosystem is important. In this study, an *in vitro* assay of bacterial adhesion to Caco-2 mammalian cells was used to screen the human gut metagenomics library (section 3.1.1). One of

the challenges of *in vitro* adhesion assays is that little is known as to how well *in vitro* adhesion correlates with *in vivo* adhesion. Studies demonstrate that many factors interfere with mucosal adhesion, it is therefore difficult to extrapolate *in vitro* adhesion results reliably to *in vivo* situations in humans. Various adhesion models usually describe only one part of the intestinal mucosa. Most of the models used to assess the adhesion *in vitro* (such as Caco-2 cells) represent simplifications of *in vivo* conditions[204]. Furthermore, investigation of microbial adhesion in the gut is difficult not only because gut microbes are largely unculturable but also because of the lack of effective means to preserve the intestinal mucus layer, where microbial communities are formed[407,408]. In order to study the multi-factorial process of adhesion, a variety of *in vitro* model systems for routine adhesion experiments [e.g., Caco-2 or HT-29 human derived adenocarcinoma cells, immobilized intestinal mucus[316] , immobilized extracellular matrices[409] and detection methods for the quantitative measurement of adhesion [e.g. quantitative culturing[410], microscopical enumeration, radiolabelling[224], immunological detection and FISH[411]] have been developed. However, to gain a better insight into the molecular mechanisms of the complex bacterial adhesin-host glycan interactions, some difficulties still remain. For example, protocols for measuring bacterial adhesion are not yet standardized across studies. One way to improve the *in vitro* analysis of metagenomics clones is to use more than one model. For example, tissue culture cells or intestinal mucus can be used in pre-screening and whole tissue or organ culture can then be used as a subsequent refined model. Whole tissue or organ cultures would take into account the influence of the normal intestinal microbiota.

It would be beneficial to interrogate differentiated Caco-2 cells on lectin microarrays to decipher the relative types of glycan epitopes present on the Caco-2 cell surface. Angeloni *et al* (2005)[179] immobilized Caco-2 cell extracts (non-differentiated Caco-2 cells 7 days post seeding and differentiated 21 day old Caco-2 cells) and interrogated them with a series of plant lectins. The results indicate that the cell glycosylation phenotype changes with increasing culture time or differentiated status, respectively. They discovered that alpha-2,3-linked sialic acid epitopes reduced from 7 days to 21 days in culture[179]. Studies like these help to explain the different adhesive profiles observed by our fosmid clones at varying Caco-2 differentiation stages (7 days vs 3 weeks). The adhesive profiles of our fosmid clones on 7 day old caco-2 cells and 3 week old caco-2 cells are vastly different. These changes can be attributed to the

change in cell glycosylation of the Caco-2 cells with increasing culture time and differentiation status.

Another factor that is worth studying is the effect of pH on bacterial adherence *in vitro*. The pH has been shown to affect adherence ability in *Candidatus albicans*[217]. Indeed, the mechanism of the effect of pH value on adherence has not been reported in detail so far. While the pH of the luminal small gut is about 7, individual viscosity of the brush border environment is acidic due to the presence of mucus overlay and metabolic activity of intestinal cells and adhered bacteria; therefore it is worth doing adhesion tests at different pH values.

6.4 Carbohydrate-based Microarrays

As described in chapter 4, carbohydrate microarrays (lectin, mucin and neoglycoconjugate) were used to characterize the carbohydrate binding specificities of FC3 and FC21 because cell surface glycosylation patterns encode information implicated in adherent processes. The lectin microarray revealed that clones FC3 and FC21 have altered cell surface glycosylation. However, in spite of their altered cell surface glycosylation, both clones did not bind mucin differently from the control strain and each other. However, when interrogated on NGC microarrays, FC3 demonstrated enhanced binding to specific glycans in the presence of arabinose and antibiotic.

When positive data is seen on carbohydrate microarrays, it should be further validated by other assays to confirm the binding seen on the microarrays. For example, if a glycan is bound by adhesive clone, the same glycan structure can be obtained and tested by ELISA to show that the sample can bind the glycan in an alternate format. The carbohydrate microarray should be viewed as a screening technique and one of several experimental methods to prove the interactions of adhesive factors with glycan ligands. Another effective method to validate the specificity of observed interactions in carbohydrate microarrays would be to perform carbohydrate inhibition assays. In the future, the addition of a much larger variety of human mucin samples on natural mucin microarrays is a consideration. As mentioned, only two human cell lines were presented on our mucin microarrays.

The utility of the carbohydrate microarray is directly related to the number and variety of the glycans available on the printed surface for interrogation by the adhesive clones. Knowledge of the specificity of carbohydrate binding of FC3 and FC21 contributes to understanding its function, to define the paradigms by which adhesin-carbohydrate interactions mediate cell communication.

6.5 Future work

The results of the experiments performed in this study are a good foundation for future characterisation studies to elucidate bacterial-glycan binding mechanism of the gut. This project highlights a vast range of research possibilities that would be interesting to pursue and explore in the future:

6.5.1 Functional Screens

As a future perspective, mucin microarrays could have been explored further by examining the interactions of FC3 and FC21 with HT29-MTX-E12 cells, which harbour and adherent mucus layer, and compared colonization efficiencies with parental HT29-MTX (mucus-secreting) and HT29 (non-mucus secreting) cells. To determine the effect of secreted mucins and an adherent mucus layer on the interactions between host cells and FC3 and FC21, a comparison can be made of the colonization of 3 cell lines (HT29, HT29-MTX, and HT29-MTX-E12) with the two clones. We could explore the interaction of FC3 and FC21 with purified mucus and mucin from HT29-MTX-E12 cells. We could do this by using an antibody to detect bound bacteria on either mucus or purified mucin from HT29-MTX-E12 cells immobilized on a PVDF membrane. If the binding of FC3 and FC21 is abolished by treatment with sodium metaperiodate treatment, it will indicate that the clones were adhering to glycan epitopes.

6.5.2 Diversity of gut microbiota and “omics” technology

Metagenomics has been applied to soil and sea ecosystems for several years. Its application in the human gut microbiome is still in its infancy. It is important to note that the human gut harbours not only bacteria, but eukaryotes, fungi and viruses. To date, very few studies have been performed on viruses and eukaryotes of the gut using metagenomics. Therefore, metagenomics studies of the entire human gut microbiota has tremendous potential. Also the combination of other “omics” technologies such as

metatranscriptomics, metaproteomics and metabolomics to microbe-glycan interaction research will make it possible for researchers to identify new microbial diagnostic markers that will provide early diagnosis and novel treatments. Furthermore, the majority of current metagenomics data of the human gut microbiome comes from studies performed in North American and Europe. Fewer studies are performed in Asia, Africa or South America. This introduces a biased view of the gut microbiota. It is necessary to enhance our understanding of the human gut microbiota by investigating human populations from different countries, for longer periods, and include multiple age groups, and various disease stages[2].

6.5.3 Lectin binding signature of mammalian Caco-2 cells

In this study, the human cell line Caco-2 was used as an *in vitro* model for intestinal epithelial cells in the adhesion assay of our metagenomics library. A characteristic trait of Caco-2 cells is their spontaneous enterocyte-like differentiation in culture after cells reach confluence[188]. Although proliferation and differentiation of Caco-2 has been studied extensively, including quantitative and proteomic analysis[412, 413, 414], the associated changes in glycosylation that accompany Caco-2 cell differentiation have yet to be fully characterized. Thus far, glycosylation targeted studies have focused mainly on changes in glycosyltransferase activity and mRNA levels[309]. Studies show that upon differentiation, increased activity was observed for GlcNAc transferase II and V, which are involved in N-glycosylation[415], and beta-3-galactosyltransferase, alpha-2-fucosyltransferase, sialyltransferase, and besta-6-GlcNAc transferase, which are relevant to O-glycan synthesis[309]. Characterizing the glycosylation of differentiating Caco-2 cells is paramount to understanding bacterial-glycan interactions on Caco-2 cell surfaces.

A proposed future perspective that could yield great insight into the bacterial-glycan interactions of the gut microbiota is the lectin microarray profiling of 7 day old and 3 week old caco-2 cells to obtain a more global glycan analysis. Lectin array profiling of the surfaces of Caco-2 cells have already demonstrated that lectins which recognize branched fucose and alpha-2,6-sialic acid are more effective at Caco-2 cell binding[416]. However, the results of these studies provide a more qualitative indication of the presence of glycan, motifs on the cell surface, the complete composition or the relative amounts of individual structures cannot be distinguished. Furthermore, these methods do not provide information about underlying protein scaffold. Precise identification of

glycan compositions with structural detail and additional glycoproteomic analysis is necessary to adequately monitor changes in glycosylation patterns associated with cell differentiation.

In this study, 48 plant lectins on a lectin microarray were used to profile the bacterial cell surface glycosylation of our clones. Studies have indicated that plant lectin microarrays are able to recognize mammalian glycans.

Further work on Caco-2 cells would involve the use of mass spectrometry to analyse the full glycan structures present on Caco-2 cell surface. Recent advancements in mass spectrometry have overcome many limitations that were inherent to glycan profiling methodologies[417, 418]. Dayoung Park and colleagues[309] employed an MS-based analytical approach utilizing nano-LC separation with high resolution TOF MS for accurate detection of compounds to rapidly identify and quantify N-glycan alterations during Caco-2 cell differentiation. The high resolution TOF MS analysis provides accurate mass measurements and consequently, detailed and selective assignment of over 200 glycan compounds from a single injection[419]. They utilized membrane enrichment methods compatible with mass spectrometry to direct their analysis to the cell membrane compartment to identify the specific glycan features that accompany Caco-2 cell differentiation. They were able to identify the corresponding membrane-localized proteins, from which glycans were released. Monitoring changes in specific structures is important to identify the key differences in glycan landscape from 7 day old caco-2 cells to 3 week old caco-2 cells.

Dayoung Park and colleagues discovered that differentiation of Caco-2 cells after 7 days begins with the emergence of extended microvilli structures and detectable levels of intestinal alkaline phosphatase activity. Mass spectrometry analysis revealed that after 7 days, the Caco-2 cell surface glycomic profile begins to show marked preference toward complex type glycans.It is believed that glycan changes observed after 7 days is largely microvilli-associated. Hydrolases such as alkaline phosphatase, which are present on the apical border, are highly glycosylated[420] and their function and stability have been shown to be associated with the presented glycans[421, 422]. The percent of microvilli covering the cell surface continue to increase until the surface is entirely covered on day 21. Therefore, the presentation of complex type glycan structures on cell surfaces coincides with the development and dominance of

microvilli. Researchers were able to discriminate between differentiated and undifferentiated Caco-2 cells based on the abundance of high mannose type structures present on the cell surface of undifferentiated Caco-2 cells[423]. In comparison, as Caco-2 cells mature, a redistribution of the relative abundance of mannose on the cell surface was observed. Significant decreases in mannose type glycans were followed by increases in decorated complex type glycans.

Glycosylation and differentiation appear to be correlated with sialylation. Sialic acids are known to reside on the terminal ends of N-glycans. The increase sialylation of differentiated cell membrane glycans can be explained by the up-regulation expression of the ST6GAL1 gene, which encodes for a s sialyltransferase that adds sialic acid to galactose in an alpha-2,6-linkage. The preference for the cell to regulate the addition of alpha-2, 6-sialic acids on differentiated cells suggests a correlation between sialic acid linkage and differentiation. Studies have shown that the prominence of sialic acid residues on epithelia has functional significance on how the cell interacts with the external environment[424,425]. Another distinguishing feature of the mature glycosylation profile is elevated levels of the bisecting GlcNAc containing oligosaccharides. Increases were also observed in MGAT3 transcript from differentiated cells, which encodes for beta-1, 4-N-acetylglucosaminyltransferase III (GnT-III), responsible for the introduction of a bisecting GlcNAc.

6.6 Conclusion

This study aimed to further current understanding regarding the molecular interactions that govern gut bacterial-glycan adhesion, using functional metagenomics and carbohydrate-based microarrays. In chapter 3 it was observed that two clones were capable of adhering to 3 week-old Caco-2 cells with a 3 fold higher adherence efficiency than the control. Lectin microarray analysis indicated that the cell surface glycosylation of FC3 and FC21 were altered when compared with the control strain. The similar biofilm forming capability of FC3 and FC21 (Figure 4.10) suggests that their cell surface adherence factors are mainly specific for glycan epitopes present on the cell surface of Caco-2 cells. In spite of the altered cell surface glycosylation and glycan specificity, FC3 and FC21 did not bind differently to mucins on an array. HCE clustering analysis indicated 100% similarity in mucin binding of FC3, FC21 and the control. However, interrogation of FC3 and FC21 on a NGC array revealed that the

presence of antibiotic and arabinose promotes NGC binding of FC3 to specific glycans on the NGC array. The in silico analysis of the bacterial adhesion, $mapA_{Ri,}$ encoded by Roseburia intestinalis did not yield significant binding to both 7 day old Caco-2 cells (Figure 5.18) and 3-week old Caco-2 cells (Figure 5.19) suggesting that (1) $mapA_{Ri}$ does not function as an adhesin (2) or $mapA_{Ri}$ is not being expressed appropriately by *L. Lactis* NZ9000 heterologous host (3) $MapA_{Ri}$ was expressed but in undetectable protein amounts.

Overall, this study has furthered current knowledge in identifying novel genes encoded by the gut microbiota using functional metagenomics. This study has also demonstrated the glycan binding characteristics of two putative adherent clones and their interactions with mucin and NGC. The research conducted in this study illustrates the potential of functional metagenomic screening of metagenomics libraries as a means to identify and assign a function to as yet unknown genes and their encoded proteins. This study also illustrates the power of carbohydrtate-based microarrays in deciphering the encoded information implicated in bacterial-glycan adherent processes. Elucidating the mechanisms of bacterial-glycan binding interactions of the gut is an important pre-requisite for the educated selection and design of strategies to modulate the gut microbiota to benefit human health, nutrition and physiology.

REFERENCES

1. Xia, X., Fen, W., Xia, G. & Jiang, Y. The nisin-controlled gene expression system : Construction , application and improvements. *Biotechnology Advances* **24,** 2005–2007 (2006).

2. Wang, W. L. *et al.* Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology* 21(3): 803-814 (2015). doi:10.3748/wjg.v21.i3.803

3. Morgan, X. C. & Huttenhower, C. Chapter 12 : Human Microbiome Analysis. *PLOS Computational Biology* **8,** (2012).

4. Schreiner, A., Kao, J., Young, V. The gut microbiome in health and in disease. *Curr Opin Gastroeneterol* **31**(1)**:** 69–75 (2016).

5. Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M. & Owen, L. J. Dysbiosis of the gut microbiota in disease. *Microbial Ecology in Health and Disease* **1,** 1–9 (2015).

6. Bolam, D. N. & Sonnenburg, J. L. Mechanistic insight into polysaccharide use within the intestinal microbiota. *Gut Microbes* **2,** 86–90 (2011).

7. Hiergeist, A., Gläsner, J., Reischl, U. & Gessner, A. Analyses of Intestinal Microbiota : Culture versus Sequencing. *ILAR Journal* **56,** 228–240 (2015).

8. Oulas, A. *et al.* Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights* (2015). doi:10.4137/BBIi.Ss12462.

9. Varki, A. Biological roles of glycans. *Glycobiology* 1–47 (2016). doi:10.1093/glycob/cww086

10. Koropatkin, N. M., Cameron, E. a. & Martens, E. C. How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol.* **10,** 323–335 (2012).

11. Hao, W.-L. & Lee, Y.-K. Microflora of the gastrointestinal tract: a review. *Methods Mol. Biol.* **268,** 491–502 (2004).

12. Richard, J. Progress in gastroenterology development of the human gastrointestinal tract. A review. *Gastroenterology* **70,** 790–810 (1976).

13. Million, M. *et al.* Increased Gut Redox and Depletion of Anaerobic and Methanogenic Prokaryotes in Severe Acute Malnutrition. *Nat. Publ. Gr.* 1–11 (2016). doi:10.1038/srep26051

14. Johansson, M. E. V, Larsson, J. M. H. & Hansson, G. C. The two mucus layers of colon are organized by the MUC2 mucin , whereas the outer layer is a legislator of host – microbial interactions. *PNAS* **108,** 4659–4665 (2011).

15. Jonstrand, C. *et al.* Spontaneous Colitis in Muc2-Deficient Mice Reflects Clinical and Cellular Features of Active Ulcerative Colitis. *PLOS ONE* **9,** 1–12 (2014).

16. Hooper, L. V. Commensal Host-Bacterial Relationships in the Gut. *Science* **292,** 1115–1118 (2001).

17. Younes, J. A. & Knight, R. Microbiota restoration : Natural and supplemented recovery of human microbial communities Microbiota restoration : natural and supplemented recovery of human microbial communities. *Nat. Publ. Gr.* **9,** 27–38 (2010).

18. Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z. & Maria, G. HHS Public Access. *Trends Mol Med*. **21,** 109–117 (2015).

19. Rodrı, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial Ecology in Health and Disease* **1,** 1–17 (2015).

20. Nakamura, N. *et al.* Molecular Ecological Analysis of Fecal Bacterial Populations from Term Infants Fed Formula Supplemented with Selected Blends of Prebiotics. *Applied and Environmental Microbiology*. **75,** 1121–1128 (2009).

21. Kelly, D., King, T. & Aminov, R. immunity Importance of microbial colonization of the gut in early life to the development of immunity. *Mutation Research* **622**, 58-69 (2007). doi:10.1016/j.mrfmmm.2007.03.011

22. Spor, A., Koren, O. & Ley, R. Focus on Mucosal Microbiology. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Publ. Gr.* **9,** 279–290 (2011).

23. Walter, J. & Ley, R. The human gut microbiome: ecology and recent evolutionary changes. *Annu. Rev. Microbiol.* **65,** 411–29 (2011).

24. Hooper, L. V & Macpherson, A. J. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Publ. Gr.* **10**, 159-169 (2015).

25. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).

26. Qin, J. *et al.* Europe PMC Funders Group Europe PMC Funders Author Manuscripts A human gut microbial gene catalog established by metagenomic sequencing. *Nature* **464,** 59–65 (2013).

27. Nalin, R. *et al.* Towards the human intestinal microbiota phylogenetic core Towards the human intestinal microbiota phylogenetic core. *Environmental Microbiology* (2009). doi:10.1111/j.1462-2920.2009.01982.x

28. Arumugam, M. *et al.* Europe PMC Funders Group Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2013).

29. Duncan, S. H., Louis, P. & Flint, H. J. Cultivable bacterial diversity from the human colon. *Applied Microbiology* **44,** 343–350 (2007).

30. Walker, A. W. *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* **5,** 220–230 (2010).

31. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. a & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–20 (2005).

32. Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *PNAS* **15**, 6252-6257 (2011). doi:10.1073/pnas.1102938108/ /DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1102938108

33. Chassard, C., Delmas, E. & Lawson, P. A. Bacteroides xylanisolvens sp . nov ., a xylan- degrading bacterium isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology* **58**, 1008–1013 (2008). doi:10.1099/ijs.0.65504-0

34. Chassard, C. & Lawson, P. A. Bacteroides cellulosilyticus sp . nov ., a cellulolytic bacterium from the human gut microbial community. *International Journal of Systematic and Evolutionary Microbiology* **57**, 1516–1520 (2007). doi:10.1099/ijs.0.64998-0

35. Mirande, C., Kadlecikova, E., Matulova, M., Capek, P. & Forano, E. Dietary fibre degradation and fermentation by two xylanolytic bacteria Bacteroides xylanisolvens XB1A T and Roseburia intestinalis XB6B4 from the human intestine. *Journal of Applied Microbiology* **109,** 451–460 (2010).

36. Narushima, S., Itoh, K., Kuruma, K. & Uchida, K. Caecal Bile Acid Compositions in Gnotobiotic Mice Associated with Human Intestinal Bacteria with the Ability to Transform Bile Acids in 6 itro. *Microbial Ecology in Health and Disease* **11**, 55-60 (1999).

37. Sobhani, I. *et al.* Microbial Dysbiosis in Colorectal Cancer (CRC) Patients. *PLoS ONE* **6,** (2011).

38. Moore, W. E. C. & Moore, L. H. Intestinal Floras of Populations That Have a High Risk of Colon Cancer. *Applied and Environmental Microbiology* **61,** 3202–3207 (1995).

39. Xu, J. & Gordon, J. I. Honor thy symbionts. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 10452–9 (2003).

40. Mackenzie, A. K. *et al.* Two SusD-Like Proteins Encoded within a Polysaccharide Utilization Locus of an Uncultured Ruminant Bacteroidetes Phylotype Bind. *Applied and Environmental Microbiology* **78,** 5935–5937 (2012).

41. Auria, G. D. The Active Human Gut Microbiota Differs from the Total Microbiota. *PLoS ONE* **6,** (2011).

42. New, F., New, E. & Combinations, S. The Phylogeny of the Genus Clostridium : Proposal of Five New Genera and Eleven New Species Combinations. *International Journal of Systematic Bacteriology* **44**, 812-826 (1994).

43. Heinken, A. *et al.* Functional Metabolic Map of Faecalibacterium prausnitzii , a Beneficial. *Journal of Bacteriology* **196,** 3289–3302 (2014).

44. Watterlot, L. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *PNAS* **105,** (2008).

45. Monti, D., Satokari, R., Franceschi, C., Brigidi, P. & Vos, W. De. Through Ageing , and Beyond : Gut Microbiota and Inflammatory Status in Seniors and Centenarians. *PLoS ONE* **5,** (2010).

46. Claesson, M. J. *et al.* Composition , variability , and temporal stability of the intestinal microbiota of the elderly. *PNAS* **108,** 4587- 4591 (2011).

47. Mueller, S. *et al.* Differences in Fecal Microbiota in Different European Study Populations in Relation to Age , Gender , and Country : a Cross-Sectional Study. *Applied and Environmental Microbiology* **72,** 1027–1033 (2006).

48. Walker, A. W., Duncan, S. H., Leitch, E. C. M., Child, M. W. & Flint, H. J. pH and Peptide Supply Can Radically Alter Bacterial Populations and Short-Chain Fatty Acid Ratios within Microbial Communities from the Human Colon. *Applied and Environmental Microbiology* **71,** 3692–3700 (2005).

49. Filippo, C. De, Cavalieri, D., Di, M., Ramazzotti, M. & Baptiste, J. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *PNAS* **107,** 14691–14696 (2010).

50. Turnbaugh, P. J. *et al.* A core gut microbiome in obsese and lean mice. *Nature* **457,** 480–484 (2009).

51. Meyer, D. The bifidogenic effect of inulin and oligofructose and its consequences for gut health. *Eur. J. Clin. Nutr.* **63,** 1277–1289 (2009).

52. Bercik *et al.* The anxiolytic effect of Bifidobacterium longum NCC3001 involves vagal pathways for gut–brain communication. *Neurogastroenterol. motil* **23,** 1132–1139 (2012).

53. Turroni, F. *et al.* Bifidobacterium bifidum as an example of a specialized human gut commensal. *Front. Microbiol.* **5,** 437 (2014).

54. Trinchera, M., Zulueta, A., Caretti, A. & Dall'Olio, F. Control of Glycosylation-Related Genes by DNA Methylation: the Intriguing Case of the B3GALT5 Gene and Its Distinct Promoters. *Biology (Basel)*. **3,** 484–97 (2014).

55. Cohen, M. Notable Aspects of Glycan-Protein Interactions. *Biomolecules* **5**, 2056–2072 (2015). doi:10.3390/biom5032056

56. Luciano, A.CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet* **11,** 181–190 (2011).

57. Kilcoyne, M., Gerlach, J. Q., Kane, M. & Joshi, L. Surface chemistry and linker effects on lectin–carbohydrate recognition for glycan microarrays. *Anal. Methods* **4,** 2721 (2012).

58. Flannery, A., Gerlach, J. Q., Joshi, L. & Kilcoyne, M. Assessing Bacterial Interactions Using Carbohydrate-Based Microarrays. *Microarrays* **4,** 690–713 (2015). doi:10.3390/microarrays4040690

59. Ku-Lung, H. (2008) *A Systems Approach to Analyzing Bacterial Glycans and Glycan- Binding Proteins* (Dcotoral dissertation). Retrieved from University of Texas Libraries .

60. Hooper, L. V & Gordon, J. I. Glycans as legislators of host – microbial interactions : spanning the spectrum from symbiosis to pathogenicity. *Glycobiology* **11,** 1–10 (2001).

61. Johansson, M. E. V *et al.* Composition and functional role of the mucus layers in the intestine. *Cell. Mol. Life Sci.* **68,** 3635–41 (2011).

62. Hu, J., Nie, Y., Chen, J., Zhang, Y. & Wang, Z. Gradual Changes of Gut Microbiota in Weaned Miniature Piglets. *Frontiers in Microbiology* **7,** 1–15 (2016).

63. Ard, R. O. E. W. *et al.* A Strategy for Annotating the Human Milk Glycome. *J.Agric.Food Chem* **54**, 7471–7480 (2006).

64. Fuhrer, A. *et al.* Milk sialyllactose influences colitis in mice through selective intestinal bacterial colonization. *JEM* **207,** 2843-2854 (2010).

65. Medicine, E. *et al.* Survival of Human Milk Oligosaccharides in the Intestine of Infants. *Advances in Experimental Medicine and Biology* (2001). doi:10.1007/978-1-4615-1371-1

66. German, J. B., Freeman, S. L., Lebrilla, C. B. & Mills, D. A. Human Milk Oligosaccharides: Evolution, Structures and Bioselectivity as Substrates for Intestinal Bacteria. *Nestle Nutr Workshop Ser Pediatr Program.* **62**, 205-222 (2008). doi:10.1159/000146322.

67. Gnoth, M. J., Kunz, C., Kinne-saffran, E. & Rudloff, S. Human Milk Oligosaccharides Are Minimally Digested In Vitro 1. *J. Nutr*. 130: 3014–3020 (2000).

68. Marcobal, A., Barboza, M., Sonnenburg, ED., Pudlo, N., Martens, EC., Desai, P., Lebrilla C.B., Weimer, BC., Mills, D.A., German, J.B. Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways. *Cell Host Microbe* **10,** 507–514 (2012).

69. Miwa, M. & Horimoto, T. Cooperation of β -galactosidase and β - N - acetylhexosaminidase from bifidobacteria in assimilation of human milk oligosaccharides with type 2 structure. *Glycobiology* **20,** 1402–1409 (2010).

70. Favier, C. F., Vaughan, E. E., Vos, W. M. De & Akkermans, A. D. L. Molecular Monitoring of Succession of Bacterial Communities in Human Neonates. *Applied and Environmental Microbiology* **68,** 219–226 (2002).

71. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the Human Infant Intestinal Microbiota. *PLoS Biol* **5,** (2007).

72. Fallani, M. *et al.* Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology* **157**, 1385–1392 (2011). doi:10.1099/mic.0.042143-0

73. Salyers, A. A., Vercellotti, J. R., West, S. E. H. & Wilkins, T. D. Fermentation of Mucin and Plant Polysaccharides by Strains of Bacteroides from the Human Colon. *Applied and Environmental Microbiology* **33,** 319–322 (1977).

74. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *PNAS* **108**, 4578-4585 (2010). doi:10.1073/pnas.1000081107

75. Urokawa, K. K. *et al.* Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. *DNA Research* **14**, 169–181 (2007).

76. Salyers, A. A., West, S. E. H., Vercellotii, J. R. & Wilkins, T. D. Fermentation of Mucins and Plant Polysaccharides by Anaerobic Bacteria from the Human Colon. *Applied and Environmental Microbiology* **34,** 529–533 (1977).

77. Holm, J. M., Karlsson, H., Sj, H. & Hansson, G. C. A complex , but uniform O -glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC / MS n. *Glycobiology* **19,** 756–766 (2009).

78. Hamer, H. M. *et al.* Review article : the role of butyrate on colonic function. *Aliment Pharmacol Ther* **27**, 104–119 (2008). doi:10.1111/j.1365-2036.2007.03562.x

79. Segain, J. *et al.* Butyrate inhibits inflammatory responses through NF B inhibition : implications for Crohn ' s disease. *Gut* **34**, 397–403 (2000).

80. Mcintyre, A., Gibson, P. R. & Young, G. P. Butyrate production from dietary fibre and protection against large bowel cancer in a rat model. *Gut* **34**, 386–391 (1993).

81. Dronamraju, S. S., Coxhead, J. M., Kelly, S. B., Burn, J. & Mathers, J. C. Cell kinetics and gene expression changes in colorectal cancer patients given resistant starch : a randomised controlled trial. *Gut* **58**: 413-420 (2008) doi:10.1136/gut.2008.162933

82. Rombeau, J. L. & Kripke, S. A. Metabolic and Intestinal Effects of Short-Chain Fatty Acids. *Journal of Parenteral and Enteral Nutrition* **14**, 181–185 (1990)

83. Duncan, S. H., Barcenilla, A., Stewart, C. S., Pryde, S. E. & Flint, H. J. Acetate Utilization and Butyryl Coenzyme A (CoA): Acetate-CoA Transferase in Butyrate-Producing Bacteria from the Human Large Intestine. *Applied and Environmental Microbiology* **68,** 5186–5190 (2002).

84. Duncan, S. H. *et al.* Contribution of acetate to butyrate formation by human faecal bacteria. *British Journal of Nutrition* **91**, 915–923 (2004). doi:10.1079/BJN20041150

85. Taylor, T. Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature* **469**, 543-547 (2011). doi:10.1038/nature09646

86. Sakarya, S. & Öncü, S. Bacterial adhesins and the role of sialic acid in bacterial adhesion. *Med Sci Monit* **9,** 76–82 (2003).

87. Falk, P. E. R. G., Bry, L. & Holgerssont, J. A. N. lineage of FVB / N mouse

stomach results in production of Leb-containing glycoconjugates : A potential transgenic mouse model for studying Helicobacter pylori infection. *Proc. Natl. Acad. Sci.* USA **92,** 1515–1519 (1995).

88. Marshall, J. Barry."Helicobacter connections." NHMRC Heliobacter pylori Research Laboratory, QEII Medical Centre, Nedlands, WA 6009, Australia. 8 December . 2005. Nobel Lecture.

89. Falk, P. E. R. G., Hooper, L. V & Midtvedt, T. Creating and Maintaining the Gastrointestinal Ecosystem : What We Know and Need To Know from Gnotobiolog. *Microbiology and Molecular biology Reviews* **62,** 1157–1170 (1998).

90. Umesaki, Y. Interactions Between Epithelial Cells and Bacteria, Normal and Pathogenic. *Science* **276,** 964–965 (1997).

91. Setoyama, H. Segmented Filamentous Bacteria Are Indigenous Intestinal Bacteria That Activate Intraepithelial Lymphocytes and Induce MHC Class II Molecules and Fucosyl Asialo GM1 Glycolipids on the Small Intestinal Epithelial Cells in the Ex-Germ-Free Mouse. *Microbial. Immunol* **39,** 555–562 (1995).

92. Hooper, L. V, Midtvedt, T. & Gordon, J. I. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283–307 (2002). doi:10.1146/annurev.nutr.22.011602.092259

93. Deplancke, B. & Gaskins, H. R. Microbial modulation of innate defense : goblet cells and the intestinal mucus layer 1 – 3. *American Journal of Clinical Nutrition* **73,** (2001).

94. Servin, A. L. Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiology Review* **28,** 405–440 (2004).

95. H, f. Bifidobacteria and Lactobacilli exhibited different mitogenic activity on murine splenocytes. *International Journal of Probiotics and Prebiotics* **1,** 77–82 (2006).

96. Jonsson, H., Ström, E. & Roos, S. Addition of mucin to the growth medium triggers mucus-binding activity in different strains of Lactobacillus reuteri in vitro. *FEMS Microbiol. Lett.* **204,** 19–22 (2001).

97. Pretzer, G. *et al.* Biodiversity-Based Identification and Functional Characterization of the Mannose-Specific Adhesin of Lactobacillus plantarum Biodiversity-Based Identification and Functional Characterization of the Mannose-Specific Adhesin of Lactobacillus plantarum. *Journal of Bacteriology* **187** (17): 6128 (2005). doi:10.1128/JB.187.17.6128

98. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–9 (2006).

99. Navarre, W. W. & Schneewind, O. Surface Proteins of Gram-Positive Bacteria and Mechanisms of Their Targeting to the Cell Wall Envelope Surface Proteins of Gram-Positive Bacteria and Mechanisms of Their Targeting to the Cell Wall

Envelope. *Microbiol. Mol. Biol. Rev* **63** (1): 174 (1999).

100. Ton-that, H., Marraffini, L. A. & Schneewind, O. Protein sorting to the cell wall envelope of Gram-positive bacteria. *Biochimica et Biophysica Acta* **1694,** 269–278 (2004).

101. Vélez, M. P., De Keersmaecker, S. C. J. & Vanderleyden, J. Adherence factors of Lactobacillus in the human gastrointestinal tract. *FEMS Microbiol. Lett.* **276,** 140–8 (2007).

102. Bansil, R. & Turner, B. S. Mucin structure, aggregation, physiological functions and biomedical applications. *Curr. Opin. Colloid Interface Sci.* **11,** 164–170 (2006).

103. Pett, Christian. (2016) *Synthesis and Development of a Mucin Glycopeptide Microarray System for Evaluation of Protein- Interactions* (Doctoral dissertation).

104. Adikwu, M. U. Mucins and their potentials. *Trop. J. Pharm. Res.* **5,** 581–582 (2006).

105. Biochemistry, M. & Academy, B. S. *Mucus and mucins during gastrointestinal infections Nazanin Navabi Department of Medical Biochemistry and Cell biology*. (2014).

106. Kim, J. J. & Khan, W. I. Goblet cells and mucins: role in innate defense in enteric infections. *Pathog. (Basel, Switzerland)* **2,** 55–70 (2013).

107. Fogg, F. J. *et al.* Characterization of pig colonic mucins. *Biochem. J.* **316** (Pt 3), 937–942 (1996).

108. Fogg, F. J., Allen, A., Harding, S. E. & Pearson, J. P. The structure of secreted mucins isolated from the adherent mucus gel: comparison with the gene products. *Biochem. Soc. Trans.* **22,** 229S (1994).

109. Hollingsworth, M. A. & Swanson, B. J. Mucins in cancer: protection and control of the cell surface. *Nat. Rev. Cancer* **4,** 45–60 (2004).

110. Loy, C. "N-linked Glycosylation". (2012). Lecture.

111. Duarte, H. *et al.* Mucin-Type O-Glycosylation in Gastric Carcinogenesis. *Biomolecules* **6,** 33 (2016).

112. Bennett, E. P. *et al.* Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22,** 736–756 (2012).

113. Linden, S. K., Sutton, P., Karlsson, N. G., Korolik, V. & McGuckin, M. a. Mucins in the mucosal barrier to infection. *Mucosal Immunol.* **1,** 183–197 (2008).

114. Vesterlund, S., Karp, M., Salminen, S. & Ouwehand, A. C. Staphylococcus aureus adheres to human intestinal mucus but can be displaced by certain lactic acid bacteria. *Microbiology* **152,** 1819–26 (2006).

115. Dahlberg, S., Normark, S., Henriques-normark, B., Kline, K. A. & Fa, S. Review Bacterial Adhesins in Host-Microbe Interactions. *Cell Host and Microbe* (2009). doi:10.1016/j.chom.2009.05.011

116. Sokurenko, E. V, Courtney, H. S., Ohman, D. E., Klemm, P. E. R. & Hastyl, D. L. FimH Family of Type 1 Fimbrial Adhesins : Functional Heterogeneity due to Minor Sequence Variations among fimH Genes. *Journal of Bacteriology* **176,** 748–755 (1994).

117. Hung, M. *et al.* The Biology of Neisseria Adhesins. *Biology* **2**, 1054–1109 (2013). doi:10.3390/biology2031054

118. Johnson, J. R. Virulence Factors in Escherichia coli Urinary Tract Infection. *Clinical Microbiology Reviews* **4,** 80–128 (1991).

119. Friberg, N. *et al.* Factor H Binding as a Complement Evasion Mechanism for an Anaerobic Pathogen , Fusobacterium necrophorum 1. *J Immunol* **181**, 8624-8632 (2008). doi:10.4049/jimmunol.181.12.8624

120. Finlay, B. B. Common Themes in Microbial Pathogenicity Revisited. *Microbiology and Molecular biology Reviews* **61,** 136–169 (1997).

121. Hultgren, S. J. Bacterial Adhesins: Common Themes and Variations in Architecture and Assembly. *Journal of Bacteriology* **181,** 1059–1071 (1999).

122. Cdllb, C. R., Ishibashi, B. Y., Claus, S. & Relman, D. A. Bordetella pertussis. *J. Exp. Med*. **180,** 1225-1223 (1994).

123. Stokes, R. W. *et al.* The Glycan-Rich Outer Layer of the Cell Wall of Mycobacterium tuberculosis Acts as an Antiphagocytic Capsule Limiting the Association of the Bacterium with Macrophages. *Infection and Immunity* **72,** 5676–5686 (2004).

124. Scheller, E. V & Cotter, P. A. Bordetella filamentous hemagglutinin and fimbriae : critical adhesins with unrealized vaccine potential. *FEMS Pathogens and Disease* **73**, 1–9 (2015). doi:10.1093/femspd/ftv079

125. Leffler, H. & Svanborg-edrnn, C. Glycolipid Receptors for Uropathogenic Escherichia coli on Human Erythrocytes and Uroepithelial Cells. *Infection and Immunology* **34,** 920–929 (1981).

126. Roberts, J. A. *et al.* The Gal (al-4) Gal-specific tip adhesin of Escherichia coli P-fimbriae is needed for pyelonephritis to occur in the normal urinary tract. *Proc. Natl. Acad. Sci. US*A **91,** 11889–11893 (1994).

127. Leffler, H. & N, C. S. E. D. I. 3 . 1 . Glycosphingolipid composition of urinary sedi-. *FEMS Microbiology Letters* **8,** 127–134 (1980).

128. Hull, R. A., Gill, R. E., Hsu, P., Minshew, B. H. & Falkow, S. Construction and Expression of Recombinant Plasmids Encoding Type 1 or D-Mannose-Resistant Pili from a Urinary Tract Infection Escherichia coli Isolate. *Infection and Immunity* **33,** 933–938 (1981).

129. Lillington, J., Geibel, S. & Waksman, G. Biochimica et Biophysica Acta

Biogenesis and adhesion of type 1 and P pili. *Biochimica et Biophysica Acta* **1840,** 2783–2785 (2014).

130. Krogfelt, K. a, Bergmans, H. & Klemm, P. Direct evidence that the FimH protein is the mannose-specific adhesin of Escherichia coli type 1 fimbriae. *Infect. Immun.* **58,** 1995–8 (1990).

131. Burrows, L. L. Twitching Motility : Type IV Pili in Action. *Annu. Rev. Microbiol.* **66**, 493–520 (2012). doi:10.1146/annurev-micro-092611-150055

132. Collinson, S. K., Emody, L. & Kay, W. W. Purification and Characterization of Thin , Aggregative Fimbriae from Salmonella enteritidis. *Journal of Bacteriology* **173,** 4773–4781 (1991).

133. La, K. Bacterial Plasminogen Receptors : In Vitro Evidence for a Role in Degradation of the Mammalian Extracellular Matrix. *Infection and Immunity* **63,** 3659–3664 (1995).

134. Wick, M. J. O., Mo, M. & Olse, A. Curli , Fibrous Surface Proteins of Escherichia coli , Interact with Major Histocompatibility Complex Class I Molecules. *Infection and Immunity* **66,** 944–949 (1998).

135. Bian, Z., Brauner, A. & Li, Y. Expression of and Cytokine Activation by Escherichia coli Curli Fibers in Human Sepsis. *The Journal of Infectious Diseases* **181**, 602-12 (2000)

136. Marti, B. *et al.* Characterization of the Collagen-Binding S-Layer Protein CbsA of Lactobacillus crispatus. *Journal of Bacteriology* **182,** 6440–6450 (2000).

137. Antikainen, J., Anton, L., Sillanpää, J. & Korhonen, T. K. Domains in the S-layer protein CbsA of Lactobacillus crispatus involved in adherence to collagens , laminin and lipoteichoic acids and in self-assembly. *Molecular Microbiology* **2**, 381–394 (2002).

138. Hagen, K. E. *et al.* Surface-layer protein extracts from Lactobacillus helveticus inhibit enterohaemorrhagic Escherichia coli O157 : H7 ... Surface-layer protein extracts from Lactobacillus helveticus inhibit enterohaemorrhagic Escherichia coli O157 : H7 adhesion to epithelial cells. *Cellular Microbiology* **9**(2), 356-367 (2007). doi:10.1111/j.1462-5822.2006.00791.x

139. Vidgren, G., Pakkanen, I. P. R. & Lounatmaa, K. S-Layer Protein Gene of Lactobacillus brevis : Cloning by Polymerase Chain Reaction and Determination of the Nucleotide Sequence. *Journal of Bacteriology* **174,** 7419–7427 (1992).

140. Åvall-ja, S., Kyla, K. & Kahala, M. Surface Display of Foreign Epitopes on the Lactobacillus brevis S-Layer. *Applied and Environmental Microbiology* **68,** 5943–5951 (2002).

141. Hynönen, U., Westerlund-wikström, B., Palva, A. & Korhonen, T. K. Identification by Flagellum Display of an Epithelial Cell- and Fibronectin-Binding Function in the SlpA Surface Protein of Lactobacillus brevis. *Journal of Bacteriology* **184,** 3360–3367 (2002).

142. Buck, B. L., Altermann, E., Svingerud, T. & Klaenhammer, T. R. Functional Analysis of Putative Adhesion Factors in Lactobacillus acidophilus NCFM. *Applied and Environmental Microbiology* **71,** 8344–8351 (2005).

143. Sleytr, U. B., Schuster, B., Egelseer, E. & Pum, D. S-layers : principles and applications. *FEMS Microbiol Rev* **38**, 823-864 (2014). doi:10.1111/1574-6976.12063

144. Leeuw, E. De, Li, X. & Lu, W. Binding characteristics of the Lactobacillus brevis ATCC 8287 surface layer to extracellular matrix proteins. *FEMS Microbiol Lett* **260,** 210–215 (2006).

145. Wu, H. & Wu, E. The role of gut microbiota in immune homeostasis and autoimmunity © 2012 Landes Bioscience . Do not distribute . © 2012 Landes Bioscience . Do not distribute . *Gut microbes* **3,** 4–14 (2012).

146. Chung, H. *et al.* Gut Immune Maturation Depends on Colonization with a Host-Specific Microbiota. *Cell* **149,** 1578–1593 (2013).

147. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics : Genomic Analysis of Microbial Communities. *Annu. Rev. Genet.* **38**: 525-52 (2004). doi:10.1146/annurev.genet.38.072902.091216

148. Torres, A. G., Zhou, X. & Kaper, J. B. Adherence of Diarrheagenic Escherichia coli Strains to Epithelial Cells Minieview Adherence of Diarrheagenic Escherichia coli Strains to Epithelial Cells. *Infect. Immun.***73** (1): 18**,** (2005).

149. De Filippo, C., Ramazzotti, M., Fontana, P. & Cavalieri, D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief. Bioinform.* (2012). doi:10.1093/bib/bbs070

150. Handelsman, J. Metagenomics : Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev* **68,** 669–685 (2004).

151. Taupp, M., Mewis, K. & Hallam, S. J. The art and design of functional metagenomic screens. *Curr. Opin. Biotechnol.* **22,** 465–72 (2011).

152. Simon, C. & Daniel, R. Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* **77,** 1153–61 (2011).

153. Pookhao, N. *et al.* Genome analysis A two-stage statistical procedure for feature selection and com- parison in functional analysis of metagenomes. *Bioinformatics* 15; 31(2) 158–165 (2014).

154. Kakirde, K. S. *et al.* Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene* **475,** 57–62 (2011).

155. Edlund, A., Hårdeman, F., Jansson, J. K. & Sjöling, S. Active bacterial community structure along vertical redox gradients in Baltic Sea sediment. *Environ. Microbiol.* **10,** 2051–63 (2008).

156. Mullany, P. Functional metagenomics for the investigation of antibiotic resistance. *Virulence* **5,** 443–7 (2014).

157. Dias, R. *et al.*(2014) Metagenomics : Library construction and screening methods.

158. Gottschalk, G., Henne, A., Schmitz, R. A., Bo, M. & Daniel, R. Screening of Environmental DNA Libraries for the Presence of Genes Conferring Lipolytic Activity on Escherichia coli. *Applied and Environmental Microbiology* **66,** 3113–3116 (2000).

159. Rondon, M. R. *et al.* Cloning the Soil Metagenome : a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Applied and Environmental Microbiology* **66,** 2541–2547 (2000).

160. Mukai, K., Kawata-Mukai, M. & Tanaka, T. Stabilization of phosphorylated Bacillus subtilis DegU by DegR. *J. Bacteriol.* **174,** 7954–62 (1992).

161. Moore, A. M., Munck, C., Sommer, M. O. A. & Dantas, G. Functional metagenomic investigations of the human intestinal microbiota. *Frontiers in Microbiology* **2,** 1–8 (2011).

162. 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. *Chemistry & Biology* **5**, 245-249 (1998).

163. Ivanova, N. *et al.* Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450,** (2007).

164. Boucher, H. W. *et al.* Bad Bugs , No Drugs : No ESKAPE ! An Update from the Infectious Diseases Society of America. *IDSA Report* **2111,** 1–12 (2009).

165. Rd, B. *et al.* Deconvolution analysis to quantify autotrophic and heterotrophic respiration and their temperature. *New Phytologist* **188**, 10–11 (2010). doi:10.1029/2002GB001971.Wiant

166. Diaz-torres, M. L. *et al.* Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS Microbiol. Lett* **258**, 257-262 (2006). doi:10.1111/j.1574-6968.2006.00221.x

167. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2016).

168. Culligan, E. P., Sleator, R. D., Marchesi, J. R. & Hill, C. Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. *ISME J.* **6,** 1916–25 (2012).

169. Yoon, M. Y. *et al.* Functional screening of a metagenomic library reveals operons responsible for enhanced intestinal colonization by gut commensal microbes. *Appl. Environ. Microbiol.* (2013). doi:10.1128/AEM.00581-13

170. Yoon, M. Y. *et al.* Functional screening of a metagenomic library reveals operons responsible for enhanced intestinal colonization by gut commensal microbes. *Appl. Environ. Microbiol.* **79,** 3829–38 (2013).

171. Sakakushev, B. E. Enhanced Recovery after Surgery for Gastric Cancer. *J*

Gastroint Dig Syst **3,** 12–14 (2013).

172. Xu, J. *et al.* Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* **5,** e156 (2007).

173. Sun, J. *et al.* Therapeutic potential to modify the mucus barrier in inflammatory bowel disease. *Nutrients* **8,** 1–15 (2016).

174. Nagler-Anderson, C. Man the barrier! Strategic defences in the intestinal mucosa. *Nat. Rev. Immunol.* **1,** 59–67 (2001).

175. Anderson, R. C., Dalziel, J. E. & Gopal, P. K. The Role of Intestinal Barrier Function in Early Life in the Development of Colitis. *Colitis* 3–31 (2009).

176. Van Tassell, M. L. & Miller, M. J. Lactobacillus adhesion to mucus. *Nutrients* **3,** 613–36 (2011).

177. Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* **16,** 1105–1115 (2009).

178. Sasaki, N. & Toyoda, M. Glycoconjugates and related molecules in human vascular endothelial cells. *Int. J. Vasc. Med.* (2013).

179. Angeloni, S. *et al.* Glycoprofiling with micro-arrays of glycoconjugates and lectins. *Glycobiology* (2005). doi:10.1093/glycob/cwh143

180. Smith, B., Li, N., Andersen, A. S., Slotved, H. C. & Krogfelt, K. A. Optimising Bacterial DNA Extraction from Faecal Samples : Comparison of Three Methods. *The Open Microbiology Journal* **5** 14–17 (2011).

181. Nelson, E. A., Palombo, E. A. & Knowles, S. R. Comparison of methods for the extraction of bacterial DNA from human faecal samples for analysis by real-time PCR. *Applied Microbiology and Microbial Biotechnology* 1479–1485 (2010).

182. Tongeren, S. P. Van, Degener, J. E. & Harmsen, H. J. M. Comparison of three rapid and easy bacterial DNA extraction methods for use with quantitative real-time PCR. *Eur J Clin Microbiol Infect Dis* 30, 1053–1061 (2011). doi:10.1007/s10096-011-1191-4

183. Strain, P. CopyControl ™ Fosmid Library Production Kit with pCC1FOS ™ Vector CopyControl ™ HTP Fosmid Library Production Kit with pCC2FOS ™ Vector. (2012).

184. Van Damme J.M.Els *et al*. Handbook of Plant Lectins. Properties and Biomedical Applications.John Wiley and Sons Ltd, 1998. Print.

185. Langendijk, P. S. *et al.* Quantitative Fluorescence In Situ Hybridization of Bifidobacterium spp . with Genus-Specific 16S rRNA-Targeted Probes and Its Application in Fecal Samples. *Applied and Environmental Microbiology* **61,** 3069–3075 (1995).

186. Handelsman, J. Metagenomics : Application of Genomics to Uncultured

Microorganisms Metagenomics : Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev.***68** (4): 669 (2004).

187. Fogh, J. M. & Orfeo, T. One Hundred and Twenty-Seven Cultured Human Tumor Cell Lines Producing Tumors in. *J. Nat. Cancer. Inst.* **59,** 221–226 (1977).

188. Pignata, S., Maggini, L., Rea, A., Acquaviva, A. M. & Pansini, S. The Enterocyte-like Differentiation of the Caco-2 Tumor Cell Line Strongly Correlates with Responsiveness to cAMP and Activation of Kinase A Pathway.*Cell Growth and Differentiatio*n **5,** 967–973 (1994).

189. Greene, J. D. & Klaenhammer, T. R. Factors involved in adherence of lactobacilli to human Caco-2 cells. *Appl. Environ. Microbiol.* **60,** 4487–94 (1994).

190. Letourneau, J., Levesque, C., Berthiaume, F., Jacques, M. & Mourez, M. In vitro assay of bacterial adhesion onto mammalian epithelial cells. *J. Vis. Exp.* **1,** 3–6 (2011).

191. Wild, J., Hradecna, Z. & Szybalski, W. Conditionally Amplifiable BACs : Switching From Single-Copy to High-Copy Vectors and Genomic Clones. *Genome Research* **12**, 1434–1444 (2002). doi:10.1101/gr.130502.replication

192. Strain, P. & Strain, P. CopyControl ™ Fosmid Library Production Kit with pCC1FOS ™ Vector CopyControl ™ HTP Fosmid Library Production Kit with pCC2FOS ™ Vector. (2012).

193. Pallen, M. J., Lam, a C., Loman, N. J. & McBride, a. An abundance of bacterial ADP-ribosyltransferases--implications for the origin of exotoxins and their human homologues. *Trends Microbiol.* **9,** 302–7; discussion 308 (2001).

194. Markowicz, C., Olejnik-schmidt, A., Borkowska, M. & Schmidt, M. T. SpaCBA sequence instability and its relationship to the adhesion efficiency of Lactobacillus casei group isolates to Caco-2 cells. *Acta. Biochim. Pol* **61**(2) 341-7 (2014).

195. Sullivan, D. J. O. & Klaenhammer, T. R. High- and low-copy-number features for clone screening. *Gene* **137,** 227–231 (1993).

196. Kunji, E. R. ., Slotboom, D.-J. & Poolman, B. Lactococcus lactis as host for overproduction of functional membrane proteins. *Biochim. Biophys. Acta - Biomembr.* **1610,** 97–108 (2003).

197. Wegmann, U. *et al.* Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. *J. Bacteriol.* **189,** 3256–70 (2007).

198. Douillard, F. P., Mahony, J., Campanacci, V., Cambillau, C. & van Sinderen, D. Construction of two Lactococcus lactis expression vectors combining the Gateway and the NIsin Controlled Expression systems. *Plasmid* **66,** 129–35 (2011).

199. Lamolle, Guillermo." Basic Local Alignment Search Tool." Maestría

PEDECIBA en Bioinformática.

200. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell* **148,** 1258–1270 (2012).

201. Wexler, H. M. Bacteroides : the Good , the Bad , and the Nitty-Gritty. *Clinical Microbiology Reviews* **20,** 593–621 (2007).

202. Thudichum, Hornemann. Sphingolipid metabolism. UniversitatSpital Zurich. Sept. 2007. Lecture.

203. O'Boyle, N. & Boyd, A. Manipulation of intestinal epithelial cell function by the cell contact-dependent type III secretion systems of Vibrio parahaemolyticus. *Front. Cell. Infect. Microbiol.* **3,** 114 (2014).

204. Ouwehand, A. C. & Salminen, S. In vitro adhesion assays for probiotics and their in vivo relevance: a review. *Microb. Ecol. Health Dis.* **15,** 175–184 (2003).

205. Zhu, B., Wang, X. & Li, L. Human gut microbiome: the second genome of human body. *Protein Cell* **1,** 718–25 (2010).

206. Felczykowska, A., Bloch, S. K. & Nejman-faleńczyk, B. Metagenomic approach in the investigation of new bioactive compounds in the marine environment. *Biochimica Polonica* **59,** 501-505 (2012).

207. Liu, J. *et al.* Complexity of coupled human and natural systems. *Science* **317,** 1513–6 (2007).

208. Craig, J. W., Chang, F.-Y., Kim, J. H., Obiajulu, S. C. & Brady, S. F. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl. Environ. Microbiol.* **76,** 1633–41 (2010).

209. van Pijkeren, J.-P. *et al.* Comparative and functional analysis of sortase-dependent proteins in the predicted secretome of Lactobacillus salivarius UCC118. *Appl. Environ. Microbiol.* **72,** 4143–53 (2006).

210. Cavanagh, D. *et al.* Phages of non-dairy lactococci: isolation and characterization of ΦL47, a phage infecting the grass isolate Lactococcus lactis ssp. cremoris DPC6860. *Front. Microbiol.* **4,** 417 (2014).

211. Cao, P., Wang, L., Zhou, G., Wang, Y. & Chen, Y. Rapid assembly of multiple DNA fragments through direct transformation of PCR products into E. coli and Lactobacillus. *Plasmid* **76,** 40–46 (2014).

212. Eric, R., Geertsma, E. R. & Poolman, B. High-throughput cloning and expression in recalcitrant bacteria. *Nature Methods* **4**(9), 705-707 (2007). doi:10.1038/NMETH1073

213. Papagianni, M., Ambrosiadis, I. & Filiousis, G. Mould growth on traditional greek sausages and penicillin production by Penicillium isolates. *Meat Sci.* **76,** 653–7 (2007).

214. Geertsma, E. R., Groeneveld, M., Slotboom, D.-J. & Poolman, B. Quality control of overexpressed membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 5722–7 (2008).

215. Kunji, E. R. S. & Harding, M. Projection structure of the atractyloside-inhibited mitochondrial ADP/ATP carrier of Saccharomyces cerevisiae. *J. Biol. Chem.* **278,** 36985–8 (2003).

216. Granato, D. *et al.* Cell surface-associated lipoteichoic acid acts as an adhesion factor for attachment of Lactobacillus johnsonii La1 to human enterocyte-like Caco-2 cells. *Appl. Environ. Microbiol.* **65,** 1071–7 (1999).

217. Persi, M. A., Burnham, J. C. & Duhring, J. L. Effects of Carbon Dioxide and pH on Adhesion of Candida albicans to Vaginal Epithelial Cells. *Infection and Immunity* **50,** 82–90 (1985).

218. Coconnier, M. H., Klaenhammer, T. R., Kernéis, S., Bernet, M. F. & Servin, a L. Protein-mediated adhesion of Lactobacillus acidophilus BG2FO4 on human enterocyte and mucus-secreting cell lines in culture. *Appl. Environ. Microbiol.* **58,** 2034–9 (1992).

219. Ouwehand, A., Lee, Y., Puong, K., Ouwehand, A. C. & Salminen, S. Displacement of bacterial pathogens from mucus and CaCO-2 cell surface by lactobacilli Displacement of bacterial pathogens from mucus and Caco-2 cell surface by lactobacilli. *Journal of Medical Microbiology* **52**, 925-930 (2003). doi:10.1099/jmm.0.05009-0

220. Begum, I. Metagenomics and its application in soil microbial community studies : biotechnological prospects. *Journal of Animal & Plant Sciences* **6,** 611–622 (2010).

221. Duncan, S. H., Hold, G. L., Barcenilla, A., Stewart, C. S. & Flint, H. J. Roseburia intestinalis sp . nov ., a novel saccharolytic , butyrate-producing bacterium from human faeces. *International Journal of Systematic and Evolutionary Microbiology* **52**, 1615–1620 (2016).

222. Duncan, S. H. & Flint, H. J. Request for an Opinion Proposal of a neotype strain (A1-86) for Eubacterium rectale . Request for an Opinion. *International Journal of Systematic and Evolutionary Microbiology* **58**, 1735–1736 (2008). doi:10.1099/ijs.0.2008/004580-0

223. Manuscript, A. immune system. *International Journal of Systematic and Evolutionary Microbiology* **58**, 1735–1736 (2008).

224. Bernet, M. F., Brassart, D., Neeser, J. R. & Servin, a L. Adhesion of human bifidobacterial strains to cultured human intestinal epithelial cells and inhibition of enteropathogen-cell interactions. *Appl. Environ. Microbiol.* **59,** 4121–8 (1993).

225. Moran, N. a, Hansen, A. K., Powell, J. E. & Sabree, Z. L. Distinctive gut microbiota of honey bees assessed using deep sampling from individual worker bees. *PLoS One* **7,** e36393 (2012).

226. Pereira, A., Betania, D. C., Quirino, F., Allen, H. & Williamson, L. L. Construction and validation of two metagenomic DNA libraries from Cerrado soil with high clay content. *Biotechnol. Lett.* (2011). doi:10.1007/s10529-011-0693-6

227. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–45 (2008).

228. Illumina. An Introduction to Next-Generation Sequencing Technology Table of Contents. 1–16 (2015).

229. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

230. Siebold, Alex."Back to the Basics: Next-Generation Sequencing 101"Agilent Technologies. Oct.2013. Lecture.

231. Coarfa, C. & Milosavljevic, A. Pash 2.0: scaleable sequence anchoring for next-generation sequencing technologies. *Pac. Symp. Biocomput.* **113,** 102–113 (2008).

232. Rudy, G. & Development, P. A Hitchhiker ' s Guide to Next-Generation Sequencing Part 1 : Evolution of sequencing technologies as a research tool. *Analysis* (2000).

233. Hempel, M., Haack, T. B., Eck, S. & Prokisch, H. Next generation sequencing. *Monatsschrift Kinderheilkd.* **159,** 827–833 (2011).

234. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1842,** 1932–1941 (2014).

235. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).

236. Frese, K., Katus, H. & Meder, B. Next-Generation Sequencing: From Understanding Biology to Personalized Medicine. *Biology (Basel).* **2,** 378–398 (2013).

237. Leblanc, V. G. & Marra, M. A. Next-generation sequencing approaches in cancer: Where have they brought us and wherewill they take us? *Cancers (Basel).* **7,** 1925–1958 (2015).

238. Church, G. & Gilbert, W. Genomic sequencing. *Proc. Natl.*(1984).

239. Risca, V. I. & Greenleaf, W. J. Beyond the Linear Genome: Paired-End Sequencing as a Biophysical Tool. *Trends Cell Biol.* **25,** 716–719 (2015).

240. Hsu, K. & Mahal, L. K. Protocol. A lectin microarray approach for the rapid analysis of bacterial glycans. *Nat.Protoc.***1**(2), 543–549 (2006).

241. Fry, S. A. *et al.* Lectin microarray pro fi ling of metastatic breast cancers. *Glycobiology* **21**(8), 1060–1070 (2011).

242. Hirabayashi, J. Development of lectin microarray , an advanced system for glycan profiling. *Technol. Ind. Proj. Struct. Glycomics* **7,** 105–117 (2014).

243. Hsu, K., Pilobello, K. T. & Mahal, L. K. Analyzing the dynamic bacterial glycome with a lectin microarray approach. *Nat.Chem.Biol*. **2**(3), 153–157 (2006).

244. Hirabayashi, J., Yamada, M., Kuno, A., Tateno, H. Lectin microarrays: concept, principle and applications. *Chem Soc Rev*. **42**, 4443–4458 (2013). doi:10.1039/c3cs35419a

245. Narimatsu, H. *et al.* A strategy for discovery of cancer glyco-biomarkers in serum using newly developed technologies for glycoproteomics. *FEBS J.* **277,** 95–105 (2010).

246. Carbohydrates: Sugars, Saccharides, Glycans :Biochemistry. Chapter. Pearson Canada Inc. 2013

247. Transforming Glycoscience: A roadmap for the future. National Academy of Science. 2012

248. Laine, R. a. A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **4,** 759–767 (1994).

249. Kilcoyne, M. *et al.* Construction of a Natural Mucin Microarray and Interrogation for Biologically Relevant Glyco-Epitopes. *Anal. Chem*. **84,** 3330-3338 (2012).

250. Kilcoyne, M., Gerlach, J. Q., Kane, M. & Joshi, L. Surface chemistry and linker effects on lectin–carbohydrate recognition for glycan microarrays. *Anal. Methods* **4,** 2721 (2012).

251. Earley, H. *et al.* A preliminary study examining the binding capacity of Akkermansia muciniphila and Desulfovibrio spp., to colonic mucin in health and ulcerative colitis. *PLoS One* **10,** 1–14 (2015).

252. Yamamoto, K. Synthesis of Bioactive Glycoconjugates Using the Transglycosylation Activity of Microbial Endoglycosidase. *Synthesis (Stuttg)*. 2–10

253. Lee, Y. C. & Lee, R. T. Neoglycoconjugates: Preparation and Applications. 549 (1994).

254. Sonnenburg, J. L. *et al.* Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* **307,** 1955–9 (2005).

255. Disney, M. D., Disney, M. D. & Seeberger, P. H. The Use of Carbohydrate Microarrays to Study Carbohydrate-Cell Interactions and to Detect Pathogens The Use of Carbohydrate Microarrays to Study Carbohydrate-Cell Interactions and to Detect Pathogens. *Cell Chemical Biology* **11**, 1701-1707 (2017). doi:10.1016/j.chembiol.2004.10.011

256. Melorose, J., Perroy, R. & Careas, S. No Title No Title. *Statew. Agric. L. Use Baseline 2015* **1,** (2015).

257. Weikkolainen, Krista (2007) *Synthesis of Neoglycoconjugates and Oligosaccharides with Potential anti- Helicobacter pylori Activity* (Academic Dissertation).

258. Besciak, G. & Surmaez-Górska, J. Biofilm As a Basic Life Form of Bacteria. 1–8 (2011).

259. Adetunji, V. O. & Isola, T. O. Crystal violet binding assay for assessment of biofilm formation by Listeria monocytogenes and Listeria spp on wood, steel and glass surfaces. *Glob. Vet.* **6,** 6–10 (2011).

260. Zmantar, T., Kouidhi, B., Miladi, H., Mahdouani, K. & Bakhrouf, A. A Microtiter plate assay for staphylococcus aureus biofilm quantification at various pH levels and hydrogen peroxide supplementation. *New Microbiol.* **33,** 137–145 (2010).

261. Tram, G., Korolik, V. & Day, C. J. MBDS Solvent: An Improved Method for Assessment of Biofilms. *Adv. Microbiol.* **3,** 200–204 (2013).

262. Antony, Alina (2011). *Study of Biofilm forming capacity of pathogens involved in Chronic Rhinosinusitis* (Master of Philsophy).Auckland University of Technology.

263. Sonkusale, K. D. & Tale, V. S. Isolation and Characterization of Biofilm Forming Bacteria from Oral Microflora. *Int.J.Curr.Microbiol.App.Sci* **2,** 118–127 (2015).

264. Maldonado, N. & Ruiz, C. S. De. A simple technique to detect Klebsiella biofilm-forming-strains. Inhibitory potential of Lactobacillus fermentum CRL 1058 whole cells and products.*Curr. Res.*52–59 (2007).

265. Bueno, J. Microbial & Biochemical Technology Anti-Biofilm Drug Susceptibility Testing Methods: Looking for New Strategies against Resistance Mechanism. *J. Microb. Biochem Technol* **S3,** 1–9 (2014).

266. Macfarlane, S. & Dillon, J. F. Microbial biofilms in the human gastrointestinal tract. *J. Appl. Microbiol.* **102,** 1187–96 (2007).

267. Lee, J. & Sullivan, D. J. O. Genomic Insights into Bifidobacteria. *Microbiology and Molecular biology reviews* **74,** 378–416 (2010).

268. Kearse, M. *et al.* Geneious Basic : An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).

269. Technical support."De Novo Assembly Using Illumina Reads".October 2009: A4. Print.

270. Leung, Wilson. "An In-Depth Introduction to NCBI BLAST" 99–116 (2011)

271. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W : improving the

sensitivity of progressive multiple sequence alignment through sequence weighting , position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22,** 4673–4680 (1994).

272. Van Domselaar, G. H. *et al.* BASys: A web server for automated bacterial genome annotation. *Nucleic Acids Res.* 1(33), W455-9 (2005). doi:10.1093/nar/gki593

273. Suryadinata, R., Seabrook, S. A., Adams, T. E., Nuttall, S. D. & Peat, T. S. Structural and biochemical analyses of a Clostridium perfringens sortase D transpeptidase research papers. Acta. *Biol. Cryst* D**71**, 1505–1513 (2015). doi:10.1107/S1399004715009219

274. Marraffini, L. A., Dedent, A. C. & Schneewind, O. Sortases and the Art of Anchoring Proteins to the Envelopes of Gram-Positive Bacteria. *Microbiol. Mol. Biol. Rev*. **70,** 192–221 (2006).

275. Schuch, R., Pelzek, a J., Kan, S. & Fischetti, V. a. Prevalence of Bacillus anthracis-like organisms and bacteriophages in the intestinal tract of the earthworm Eisenia fetida. *Appl. Environ. Microbiol.* **76,** 2286–94 (2010).

276. Snodgrass, J. L. *et al.* Functional Analysis of the Staphylococcus aureus Collagen Adhesin B Domain. *Infection and Immunity* **67,** 3952–3959 (1999).

277. Herbert, J. A., Mitchell, A. M. & Mitchell, T. J. A Serine-Threonine Kinase (StkP) Regulates Expression of the Pneumococcal Pilus and Modulates Bacterial Adherence to Human Epithelial and Endothelial Cells In Vitro. *PLoS ONE* 1–17 (2015). doi:10.1371/journal.pone.0127212

278. Liang, X., Van Doren, S. Mechanistic Insights into Phosphoprotein-Binding FHA Domains. *Acc. Chem. Res* **41,** 991–999 (2010).

279. Arora, G. *et al.* Identification of Ser / Thr kinase and Forkhead Associated Domains in Mycobacterium ulcerans : Characterization of Novel Association between Protein Kinase Q and MupFHA. *PLOS Neglected Tropical Diseases* **8,** (2014).

280. Li, J., Lee, G., Doren, S. R. Van & Walker, J. C. The FHA domain mediates phosphoprotein interactions. *Journal of Cell Science* **4149,** 4143–4149 (2000).

281. Lopetuso, L. R., Scaldaferri, F., Petito, V. & Gasbarrini, A. Commensal Clostridia : leading players in the maintenance of gut homeostasis. *Gut Pathog.* **5,** 1 (2013).

282. Facts, Q. SYTO ® Orange Fluorescent Nucleic Acid Stains. 1–2 (2001).

283. Eischeid, A. C. SYTO dyes and EvaGreen outperform SYBR Green in real-time PCR. *BMC Res. Notes* **4,** 263 (2011).

284. Attila Tarnok. SYTO Dyes and Histoproteins Myriad of Applications. *Cytometry* **73**, 477–479 (2008). doi:10.1002/cyto.a.20588

285. Zhou, M. & Wu, H. Glycosylation and biogenesis of a family of serine- rich bacterial adhesins. *Microbiology* **155**, 317–327 (2009).

doi:10.1099/mic.0.025221-0

286. Nothaft, H. & Szymanski, C. M. N-Glycosylation : New Perspectives and Applications. *Journal of Biological Chemistry* **288,** 6912–6920 (2013).

287. Tytgat, H. L. P. & Lebeer, S. The Sweet Tooth of Bacteria : Common Themes in Bacterial. *Microbiology and Molecular Biology Reviews* **78,** 372–417 (2014).

288. Giannasca, K. T., Giannasca, P. J. & Neutra, M. R. Adherence of Salmonella typhimurium to Caco-2 Cells : Identification of a Glycoconjugate Receptor. *Infection and Immunity* **64,** 135–145 (1996).

289. Donlan, R. M. Biofilm Formation : A Clinically Relevant Microbiological Process. *Healthcare Epidemiology* **33,** (2001).

290. Naughton, J. A. *et al.* Divergent Mechanisms of Interaction of Helicobacter pylori and Campylobacter jejuni with Mucus and Mucins. *Infection and Immunity* **81,** 2838–2850 (2013).

291. Wang, L. *et al.* Cross-platform comparison of glycan microarray formats. *Glycobiology* **24,** 507–517 (2014).

292. Honore, N. & Cole, S. T. Nucleotide sequence of the aroP gene encoding the Escherichia coli K-12: homology with yeast transport general aromatic amino acid transport protein of proteins. *Nucleic Acids Research* **18,** 1440 (1990).

293. Henrich, B., Hopfe, M., Kitzerow, A. & Hadding, U. The Adherence-Associated Lipoprotein P100 , Encoded by an opp Operon Structure , Functions as the Oligopeptide-Binding Domain OppA of a Putative Oligopeptide Transport System in Mycoplasma hominis. *Journal of Bacteriology* **181,** 4873–4878 (1999).

294. Dodson, E. *et al.* A Bacterial Virulence Factor with a Dual Role as an Adhesin and a Solute-binding Protein : The A Bacterial Virulence Factor with a Dual Role as an Adhesin and a Solute-binding Protein : The Crystal Structure at 1 . 5 Å Resolution of the PEB1a Protein from the Food-borne Human Pathogen Campylobacter jejuni. *Journal of Molecular Biology* **372**, 160-171 (2007). doi:10.1016/j.jmb.2007.06.041

295. Blasersqli, J. PEB1 , the Major Cell-binding Factor of Cumpylobucter jejuni , Is a Homolog of the Binding Component in Gram-negative Nutrient Transport Systems. *Journal of Biological Chemistry* **268,** 18717-18725 (1993).

296. Merino, S., Gav, R., Maguire, M. E. & Toma, J. M. The MgtE Mg 2 ‡ transport protein is involved in Aeromonas hydrophila adherence. *FEMS Microbiology Letters* **198,** 189–195 (2001).

297. Tamura, G. S., Nittayajarn, A. & Schoentag, D. L. A Glutamine Transport Gene , glnQ , Is Required for Fibronectin Adherence and Virulence of Group B Streptococci. *Infection and Immunity* **70,** 2877–2885 (2002).

298. Matthysse, A. N. N. G., Yarnall, H. A., Young, N., Hill, C. & Carolina, N. Requirement for Genes with Homology to ABC Transport Systems for Attachment and Virulence of Agrobacterium tumefaciens. *Journal of*

Bacteriology **178,** 5302–5308 (1996).

299. Bajaj, R. *et al.* Biochemical characterization of essential cell division proteins FtsX and FtsE that mediate peptidoglycan hydrolysis by PcsB in Streptococcus pneumoniae. *Microbiology Open* **5**, 738-752 (2016). doi:10.1002/mbo3.366

300. Karpel, R., Alon, T., Glasers, G., Schuldiner, S. & Padan, E. Expression of a Sodium Proton Antiporter (NhaA) in. *Journal of Biological Chemistry* 266, 21753-21759 (1991).

301. Denker, S. P. *et al.* Direct Binding of the Na – H Exchanger NHE1 to ERM Proteins Regulates the Cortical Cytoskeleton and Cell Shape Independently of H ⅞ Translocation. *Molecular Cell.* **6,** 1425–1436 (2000).

302. Guo, Y., Smith, K. & Petris, M. J. Cisplatin Stabilizes a Multimeric Complex of the Human Ctr1. *Journal of Biological Chemistry* **279,** 46393–46399 (2004).

303. Kiesau, P. *et al.* Evidence for a Multimeric Subtilin Synthetase Complex. *Journal of Bacteriology* **179,** 1475–1481 (1997).

304. Spirig T *et al*. Sortase enzymes in Gram-positive bacteria. *Mol. Microbiol* **82,** 1044–1059 (2013).

305. Maresso, A. W., Chapa, T. J. & Schneewind, O. Surface Protein IsdC and Sortase B Are Required for Heme-Iron Scavenging of Bacillus anthracis. *Journal of Bacteriology* **188,** 8145–8152 (2006).

306. Bond, M. R. & Hanover, J. A. A little sugar goes a long way : The cell biology of O-GlcNAc. *JCB* **208,** 869–880 (2015).

307. Ricciuto, J., Heimer, S. R., Gilmore, M. S. & Argu, P. Cell Surface O-Glycans Limit Staphylococcus aureus Adherence. *Infection and Immunity* **76,** 5215–5220 (2008).

308. Angeloni, S., Ridet, J. L., Kusy, N. & Gao, H. Glycoprofiling with micro-arrays of glycoconjugates and lectins. *Glycobiology* **15,** 31–41 (2005).

309. Park, D. *et al.* Characteristic Changes in Cell Surface Glycosylation Accompany Intestinal Epithelial Cell (IEC) Differentiation : High Mannose Structures Dominate the Cell Surface Glycome of Undifferentiated Enterocytes. *Molecular & Cellular Proteomics* 2910–2921 (2015). doi:10.1074/mcp.M115.053983

310. Holscher, H. D., Davis, S. R. & Tappenden, K. A. Human Milk Oligosaccharides Influence Maturation of Human Intestinal Caco-2Bbe. *Journal of Nutrition* (2014). doi:10.3945/jn.113.189704.5

311. Bode, L. Human milk oligosaccharides : Every baby needs a sugar mama. *Glycobiology* **22,** 1147–1162 (2012).

312. Elison, E. *et al.* Oral supplementation of healthy adults with 2 ′ -O-fucosyllactose and lacto-N-neotetraose is well tolerated and shifts the intestinal microbiota. *British Journal of Nutrition* 116, 1356–1368 (2017). doi:10.1017/S0007114516003354

313. Cecchini, D. A. *et al.* Functional Metagenomics Reveals Novel Pathways of Prebiotic Breakdown by Human Gut Bacteria. *PLoS One* (2013). doi:10.1371/journal.pone.0072766

314. Barriuso, J., Barriuso, J., Valverde, J. R. & Mellado, R. P. Estimation of bacterial diversity using next generation sequencing of 16S rDNA : A comparison of different workflows Estimation of bacterial diversity using next generation sequencing of 16S rDNA : a comparison of different workflows. *BMC Bioinformatics* **12,** 473 (2011).

315. Schmeisser C, Steele, H, Streit. Metagenomics biotechnology with non culturable microbes. *Appl. Microbiol. Biotechnol.* **75**(5), 955-62 (2007).

316. Roos, S. & Jonsson, H. A high-molecular-mass cell-surface protein from Lactobacillus reuteri 1063 adheres to mucus components. *Microbiology* **148,** 433–42 (2002).

317. Wang B *et al.* Identification of a Surface Protein from Lactobacillus reuteri JCM1081 That Adheres to Porcine Gastric Mucin and Human. *Curr Microbiol* 57, 33–38 (2008). doi:10.1007/s00284-008-9148-2

318. Mackenzie, D. A. *et al.* Strain-specific diversity of mucus-binding proteins in the adhesion and aggregation properties of Lactobacillus reuteri. *Microbiology* **156**, 3368–3378 (2010). doi:10.1099/mic.0.043265-0

319. Mackenzie, D. A., Tailford, L. E., Hemmings, A. M. & Juge, N. Crystal Structure of a Mucus-binding Protein Repeat Reveals an Unexpected Functional Immunoglobulin Binding Activity. *The Journal of Biological Chemistry* **284,** 32444–32453 (2009).

320. Cossart, P. Listeria monocytogenes Surface Proteins : from Genome Predictions to Function. *Microbiology and Molecular Biology Reviews* **71,** 377–397 (2007).

321. Nishiyama, K., Sugiyama, M. & Mukai, T. Adhesion Properties of Lactic Acid Bacteria on Intestinal Mucin. *Microorganisms* 1–18 (2016). doi:10.3390/microorganisms4030034

322. Etzold, S. *et al.* Structural and molecular insights into novel surface-exposed mucus adhesins from Lactobacillus reuteri human strains. *Molecular Microbiology* **92,** 543–556 (2014).

323. Boekhorst, J., Helmer, Q., Kleerebezem, M. & Siezen, R. J. Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria. *Microbiology* **152,** 273–80 (2006).

324. Etzold, S. *et al.* Structural basis for adaptation of lactobacilli to gastrointestinal mucus. *Environmental Microbiology* **16,** 888–903 (2014).

325. Adlerberth, I. *et al.* A mannose-specific adherence mechanism in Lactobacillus plantarum conferring binding to the human colonic cell line HT-29. *Appl. Environ. Microbiol.* **62,** 2244–51 (1996).

326. Malik, S., Petrova, M. I., Imholz, N. C. E. & Verhoeven, T. L. A. High

mannose-specific lectin Msl mediates key interactions of the vaginal Lactobacillus plantarum isolate CMPG5300. *Nat. Publ. Gr.* 1–16 (2016). doi:10.1038/srep37339

327. Petrova, M. I. *et al.* The lectin-like protein 1 in Lactobacillus rhamnosus GR-1 mediates tissue-specific adherence to vaginal epithelium and inhibits urogenital pathogens. *Nat. Publ. Gr.* 1–15 (2016). doi:10.1038/srep37437

328. Petrova, M. I., Imholz, N. C. E., Verhoeven, T. L. A., Balzarini, J. & Els, J. Lectin-Like Molecules of Lactobacillus rhamnosus GG Inhibit Pathogenic Escherichia coli and Salmonella Biofilm Formation. *Scientific Reports* (2016). doi:10.1371/journal.pone.0161337

329. Boekhorst, J., Wels, M., Kleerebezem, M. & Siezen, R. J. The predicted secretome of Lactobacillus plantarum WCFS1 sheds light on interactions with its environment. *Microbiology* **152,** 3175–83 (2006).

330. Claesson, M. J. *et al.* Multireplicon genome architecture of Lactobacillus salivarius. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 6718–23 (2006).

331. Sengupta, R. *et al.* The Role of Cell Surface Architecture of Lactobacilli in Host-Microbe Interactions in the Gastrointestinal Tract. *Mediators of Inflammation* **2013,** (2013).

332. Salyers, A. A. Biochemical Analysis of Interactions between Outer Membrane Proteins That Contribute to Starch Utilization by Bacteroides thetaiotaomicron. *Journal of Bacteriology* **183,** 7224–7230 (2001).

333. Ravcheev, D. A., Godzik, A., Osterman, A. L. & Rodionov, D. A. Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron : comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics* 1–17 (2013).

334. Reeves, A. R., Wang, G. & Salyers, A. A. Characterization of Four Outer Membrane Proteins That Play a Role in Utilization of Starch by Bacteroides thetaiotaomicron. *Journal of Bacteriology* **179,** 643–649 (1997).

335. Shipman, J. a, Berleman, J. E. & Salyers, a a. Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. *J. Bacteriol.* **182,** 5365–72 (2000).

336. Miyoshi, Y., Okada, S., Uchimura, T. & Satoh, E. A Mucus Adhesion Promoting Protein, MapA, Mediates the Adhesion of Lactobacillus reuteri to Caco-2 Human Intestinal Epithelial Cells. *Biosci. Biotechnol. Biochem.* **70,** 1622–1628 (2006).

337. Bøhle, L. A., Brede, D. A., Diep, D. B., Holo, H. & Nes, I. F. Specific degradation of the mucus adhesion-promoting protein (MapA) of Lactobacillus reuteri to an antimicrobial peptide. *Appl. Environ. Microbiol.* **76,** 7306–9 (2010).

338. Tam, R. & Saier, M. H. Structural , Functional , and Evolutionary Relationships among Extracellular Solute-Binding Receptors of Bacteria. *Microbiological*

Reviews **57,** 320–346 (1993).

339. Turner, M. S., Timms, P., Hafner, L. M. & Giffard, P. M. Identification and Characterization of a Basic Cell Surface-Located Protein from Lactobacillus fermentum BR11. *Journal of Bacteriology* **179,** 3310–3316 (1997).

340. Mytelka, D. S. & Chamberlin, M. J. Escherichia coli fliAZY Operon. *Journal of Bacteriology* **178,** 24–34 (1996).

341. Hung, J., Cooper, D., Turner, M. S., Walsh, T. & Gi, P. M. Cystine uptake prevents production of hydrogen peroxide by Lactobacillus fermentum BR11 *FEMS Microbiology Letters* 227 **227,** 93–99 (2003).

342. Turner, M. S., Woodberry, T., Hafner, L. M. & Giffard, P. M. The bspA Locus of Lactobacillus fermentum BR11 Encodes an L -Cystine Uptake System. *Journal of Bacteriology* **181,** 2192–2198 (1999).

343. Atkins, H. L., Geier, M. S., Prisciandaro, L. D., Turner, M. S. & Howarth, G. S. Effects of a Lactobacillus reuteri BR11 Mutant Deficient in the Cystine-Transport System in a Rat Model of Inflammatory Bowel Disease. *Dig Dis Sci* 713–719 (2012). doi:10.1007/s10620-011-1943-0

344. Louis, P. & Flint, H. J. Minireview Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology* **0,** 1–13 (2016).

345. Lu, Y. *et al.* Short Chain Fatty Acids Prevent High-fat-diet-induced Obesity in Mice by Regulating G Protein- coupled Receptors and Gut Microbiota. *Nat. Publ. Gr.* **3,** 1–13 (2016).

346. Zacharof, M. P. & Lovitt, R. W. Bacteriocins Produced by Lactic Acid Bacteria A Review Article. *APCBEE Procedia* **2,** 50–56 (2012).

347. Kleerebezem, M., Beerthuyzen, M. M., Vaughan, E. E., Vos, W. M. D. E. & Kuipers, O. P. Controlled Gene Expression Systems for Lactic Acid Bacteria : Transferable Nisin-Inducible Expression Cassettes for Lactococcus , Leuconostoc , and Lactobacillus spp. *Applied and Evironmental Microbiology* **63,** 4581–4584 (1997).

348. Mierau, M. I. & Kleerebezem, M. 10 years of the nisin-controlled gene expression system (NICE) in Lactococcus lactis. *Appl Microbiol Biotechnol* 68, 705–717 (2005). doi:10.1007/s00253-005-0107-6

349. Vos, D. Characterization of the nisin gene cluster nisABTCZPR of Lactococcus lactis Requirement of expression of the nisA and nisZ genes for development of immunity. *Eur. J. Biochem*. **291,** 281–291 (1993).

350. Dodd, H. M., Horn, N., Giffard, C. J. & Gasson, M. J. A gene replacement strategy for engineering nisin. *Microbiology* **142**, 47–55 (1996).

351. Vos, W. M. De. Autoregulation of Nisin Biosynthesis in Lactococcus lactis by Signal Transduction. *Journal of Biological Chemistry* **270,** 27299–27304 (1995).

352. Pearson, W. An Introduction to Sequence Similarity (‘ Homology ’) Searching.

Curr. Protoc. Bioinformatics 1–9 (2014). doi:10.1002/0471250953.bi0301s42.An

353. Boekhorst, J. & Snel, B. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC Bioinformatics* **7,** 1–7 (2007).

354. Friedrich, A. *et al.* Blast sampling for structural and functional analyses. *BMC Bioinformatics* **17,** 1–17 (2007).

355. Hu, Y. *et al.* genes in a large cohort of human gut microbiota. *Nature Communications* 4:2151 (2013). doi:10.1038/ncomms3151

356. Kleerebezem, M. *et al.* Complete genome sequence of Lactobacillus plantarum WCFS1. *PNAS* (2002).

357. Maier, R. M. Bacterial Growth. *Environmental Microbiology* 37–54 (2009).

358. Turroni, F., Berry, D. & Ventura, M. Bifidobacteria and their role in the huan gut microbiota. *Frontiers in Microbiology* (2017) doi:10.3389/978-2-88945-100-5

359. Ugenholtz, J. E. H. & Altrich, D. I. H. High-Level Expression of Lactobacillus β -Galactosidases in Lactococcus lactis Using the Food-Grade , Nisin-Controlled Expression System NICE. *J. Agric. Food Chem.* 2010, 2279–2287 (2010). doi:10.1021/jf902895g

360. Morello, E., Llull, D., Miraglio, N., Langella, P. & Poquet, I. Lactococcus lactis , an Efficient Cell Factory for Recombinant Protein. *J Mol Microbiol Biotechnol*.14: 48–58 (2008). doi:10.1159/000106082

361. Gabbianelli, R. *et al.* Role of ZnuABC and ZinT in Escherichia coli O157 : H7 zinc acquisition and interaction with epithelial cells. *BMC Microbiology* 11:36 (2011).

362. Rendon, M. *et al.* Characterization of SP41 , a surface protein of Brucella associated with adherence and invasion of host epithelial cells Characterization of SP41 , a surface protein of Brucella associated with adherence and invasion of host epithelial cells. *Cellular Microbiology* **8**, 1877-1887 (2007). doi:10.1111/j.1462-5822.2006.00754.x

363. Yang, M., Johnson, A. & Murphy, T. F. Characterization and Evaluation of the Moraxella catarrhalis Oligopeptide Permease A as a Mucosal Vaccine Antigen. *Infection and Immunity* **79,** 846–857 (2011).

364. Jalalvand, F. *et al.* Haemophilus in fl uenzae Protein F Mediates Binding to Laminin and Human Pulmonary Epithelial Cells. *JID* **207,** 803–813 (2013).

365. Deshpande, N. P., Kaakoush, N. O., Wilkins, M. R. & Mitchell, H. M. Comparative genomics of Campylobacter concisus isolates reveals genetic diversity and provides insights into disease association. *BMC Genomics* **14,** 1 (2013).

366. Kaakoush, N. O. *et al.* The Pathogenic Potential of Campylobacter concisus Strains Associated with Chronic Intestinal Diseases. *PLoS ONE* **6,** (2011).

367. Manson, M. D., Tedesco, P. A. T., Berg, H. C., Haroldt, F. M. & Drift, C. V. A. N. D. E. R. A protonmotive force drives bacterial flagella Microbiology. *Proc. Nati. Acad. Sci. USA* **74,** 3060–3064 (1977).

368. State, N. C. Effect of Probiotic Supplements of Lactobacillus acidophilus and BiJidobacteriurn adolescentis 2204 on P-glucosidase and P-glucuronidase Activity in the Lower Gut of Rats Associated with a Human Faecal Flora. *Microbial Ecology in Health and Disease*. **2,** 223–225 (1989).

369. Annan N.T. Encapsulation in alginate-coated gelatin microspheres improves survival of the probiotic Bifidobacterium. *Food Research International* **41**, 184-193 (2008). doi:10.1016/j.foodres.2007.11.001

370. Dunne, C. *et al.* In vitro selection criteria for probiotic bacteria of human origin : correlation with in vivo findings. *The American Journal of Clinical Nutrition*. 1 – 4. **73,** 386–392 (2001).

371. Dash, S. K. Selection criteria for probiotics Focus on Dietary fibres - Pre / Probiotics. *Pre/Probiotics* 26–28 (1979).

372. Lee, Y. K. *et al.* Quantitative Approach in the Study of Adhesion of Lactic Acid Bacteria to Intestinal Cells and Their Competition with Enterobacteria. *Applied and Environmental Microbiology*, **66,** 3692–3697 (2000).

373. Bork, P. Powers and Pitfalls in Sequence Analysis. *European Molecular Biology Laboratory (EMBL)* 398–400 (2000).

374. Sabri, S. *et al.* Glycocalyx modulation is a physiological means of regulating cell adhesion. *Journal of Cell Science* **1600,** 1589–1600 (2000).

375. Ponsonnet, L. *et al.* Adhesion-related glycocalyx study : quantitative approach with imaging-spectrum in the energy ¢ ltering transmission electron microscope. *FEBS Letters* **429,** 1–6 (1998).

376. Mulivor, A. W. & Lipowsky, H. H. Role of glycocalyx in leukocyte-endothelial cell adhesion. *Am J Physiol Heart Circ Physiol* 283, 1282–1291 (2002).

377. Physiol, P. C. The Role of Cell Glycocalyx in Vascular Transport of Circulating Tumor Cells. *Am. J. Physiol. Cell. Physiol.* (October (2013). doi:10.1152/ajpcell.00285.2013

378. Moran, a P., Gupta, a & Joshi, L. Sweet-talk: role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract. *Gut* **60,** 1412–25 (2011).

379. Alemka, A. *et al.* Probiotic Colonization of the Adherent Mucus Layer of HT29MTXE12 Cells Attenuates Campylobacter jejuni Virulence Properties *Infection and Immunity* **78,** 2812–2822 (2010).

380. Dolan, B., Naughton, J., Tegtmeyer, N., May, F. E. B. & Clyne, M. The Interaction of Helicobacter pylori with the Adherent Mucus Gel Layer Secreted by Polarized HT29-MTX-E12 Cells. *PLoS ONE* **7,** (2012).

381. Ducret, A., Hardy, G. G. & Brun, Y. V. negative bacteria. *Microbiol Spectr*. **3,** 1–45 (2015).

382. Marchesi, J. R. *et al.* The gut microbiota and host health : a new clinical frontier. *Gut* 1–10 (2015). doi:10.1136/gutjnl-2015-309990

383. Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* **2,** 1297–305 (2007).

384. Owen, J. G. *et al.* Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *PNAS* **110,** 11797–11802 (2013).

385. Human, T. & Project, M. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).

386. Kennedy, N. A. *et al.* The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. *PLoS ONE* **9,** 1–9 (2014).

387. Wesolowska-andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* 2:19, 1–11 (2014). doi:10.1186/2049-2618-2-19

388. Guo, M. *et al.* Combination of Metagenomics and Culture-Based Methods to Study the Interaction Between Ochratoxin A and Gut Microbiota. *Toxicological Sciences* **141,** 314–323 (2014).

389. Mitra, S. *et al.* Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. *BMC Genomics* **14,** S16 (2013).

390. Miyazaki, K. & Science, A. I. Uchiyama T , Miyazaki K .. Functional metagenomics for enzyme discovery : challenges to efficient screening . *Curr Opin in Biotechnol* **20**, 616-622 (2009). doi:10.1016/j.copbio.2009.09.010

391. Strand, T. A., Lale, R., Degnes, K. F., Lando, M. & Valla, S. A New and Improved Host-Independent Plasmid System for RK2-Based Conjugal Transfer. *PLoS ONE* **9,** 1–6 (2014).

392. Buck, J. D. *Physiological Effects of Heterologous Expression of Proteorhodopsin Photosystems* (Doctoral of dissertation) Biological Engineering Massachusetts Institute of Technology (2012).

393. Medina, C., Limo, M. C., Santero, E. & Terro, L. fosmid vectors increase the functional. *Scientific reports* 3:1107 (2012) doi:10.1038/srep01107

394. Lam, K. N., Cheng, J., Engel, K., Neufeld, J. D. & Charles, T. C. Current and future resources for functional metagenomics. *Front. Microbiol.* (2015). doi:10.3389/fmicb.2015.01196

395. Martinez, A. *et al.* Genetically Modified Bacterial Strains and Novel Bacterial Artificial Chromosome Shuttle Vectors for Constructing Environmental Libraries and Detecting Heterologous Natural Products in Multiple Expression Hosts. *Appl. Environ. Microbiol*. **70,** 2452–2463 (2004).

396. Ekkers, D. M., Cretoiu, M. S., Kielak, A. M. & Elsas, J. D. Van. The great screen anomaly — a new frontier in product discovery through functional metagenomics. *Appl. Microbiol. Biotechnol* **93**, 1005–1020 (2012). doi:10.1007/s00253-011-3804-3

397. Degnes, K. F. & Schmidt, F. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiology* **296**, 149-158 (2009). doi:10.1111/j.1574-6968.2009.01639.x

398. Westenberg, M., Bamps, S., Soedling, H., Hope, I. A. & Dolphin, C. T. Escherichia coli MW005 : lambda Red-mediated recombineering and copy-number induction of oriV -equipped constructs in a single host. *BMC Biotechnology* 10: 27 (2010).

399. Angelov, A., Mientus, M., Liebl, S. & Liebl, W. A two-host fosmid system for functional screening of (meta) genomic libraries from extreme thermophiles. *Systematic and Applied Microbiology* **32,** 2008–2010 (2009).

400. Gabor, E., Niehaus, F. & Ag, B. Updating the metagenomics toolbox Updating the metagenomics toolbox. *Biotech. Journal* **2**, 201-206 (2007). doi:10.1002/biot.200600250

401. Culligan, E. P., Sleator, R. D., Marchesi, J. R. & Hill, C. Metagenomics and novel gene discovery Promise and potential for novel therapeutics. *Virulence* **5**, 399–412 (2014).

402. Zipper, H. *et al.* Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *Journal of Biotechnology* **127,** 575–592 (2007).

403. Johnston, C. *et al.* Codon optimisation to improve expression of a Mycobacterium avium ssp . paratuberculosis- specific membrane-associated antigen by Lactobacillus salivarius. *Pathogens and Disease* **68**, 27–38 (2013). doi:10.1111/2049-632X.12040

404. Society, A. C. & Francisco, S. Synthesis of Methyl Halides from Biomass Using Engineered Microbes Synthesis of Methyl Halides from Biomass Using Engineered. *Journal of the American Chemical Society* **131** (2009). doi:10.1021/ja809461u

405. Allgaier, M. *et al.* Targeted Discovery of Glycoside Hydrolases from a Switchgrass-Adapted Compost Community. *PLoS ONE* **5,** (2010).

406. Engel, K. *et al.* Meeting Report : 1st International Functional. *Standard in Genomic Sciences* **8**, 106–111 (2013). doi:10.4056/sigs.3406845

407. Bollinger, R *et al.* Biofilms in the normal human large bowel : fact rather than fiction *Gut* **56**(10), 1481–1482 (2007).

408. Strugala, V. *et al.* Colonic mucin: methods of measuring mucus thickness. *Proceedings of the Nutrition Society* **62**, 237-243 (2003).

409. Janoli, G. N., Ax, L. A. R. S. T., So, E. L. S. & Richter, C. A. R. O. L. E. S.

Binding of Lactobacillus reuteri to Fibronectin Immobilized on Glass Beads. *Zbl. Bakt*. **528,** 519-528 (1992).

410. Mack, D. R. *et al.* in vitro by inducing intestinal mucin gene expression. *American Journal of Physiology* **5**, 941-950 (1999).

411. Wolfaardt, G. M. & Dicks, L. M. T. Adhesion of Lactobacillus plantarum 423 and Lactobacillus salivarius 241 to the intestinal tract of piglets , as recorded with fluorescent in situ hybridization (FISH), and production of plantaricin 423 by cells colonized to the ileum. *Journal of Applied Microbiology* **100,** 838–845 (2006).

412. Stierum, R. *et al.* Proteome analysis reveals novel proteins associated with proliferation and differentiation of the colorectal cancer cell Proteome analysis reveals novel proteins associated with proliferation and. *Biochimica et Biophysica* **1650**, 73-91 (2003). doi:10.1016/S1570-9639(03)00204-8

413. Sinnett, D. Subcellular proteomics of cell differentiation : Quantitative analysis of the plasma membrane proteome of Caco-2 cells Quantitative analysis of the plasma membrane. *Proteomics* **7**, 2201-2215 (2007). doi:10.1002/pmic.200600956

414. Acta, B. & Hospital, S. H. Transcriptome changes during intestinal cell differentiation. *Biochimica et Biophysica* **1589,** 160-167 (2002).

415. Simon-assmann, P. In vitro models of intestinal cell differentiation. *Cell Biology and Toxicology* (2007). doi:10.1007/s10565-006-0175-0

416. Tao, S. C. *et al.* Lectin microarrays identify cell-specific and functionally significant cell surface glycan markers. *Glycobiology* (2008). doi:10.1093/glycob/cwn063

417. Zaia, J. NIH Public Access. *Chem. Biol*. **15,** 881–892 (2009).

418. North, S. J., Hitchen, P. G., Haslam, S. M. & Dell, A. NIH Public Access. *Curr. Opin. Struct. Biol.* **19,** 498–506 (2010).

419. Williams, C. C. & Wu, L. D. Isomer-specific chromatographic profiling yields highly sensitive and specific potential N- glycan biomarkers for epithelial ... *J. Chromatogr. A* **1279,** 58–67 (2014).

420. Gillian, M., Sjostrom, H. & Noren, O. Topology and quaternary structure of pro-sucrase / isomaltase and final-form sucrase / isomaltase. *Biochem. J.* **237,** 455–461 (1986).

421. Wetzel, G., Heine, M., Rohwedder, A. & Naim, H. Y. Impact of glycosylation and detergent- resistant membranes on the function of intestinal sucrase-isomaltase Impact of glycosylation and detergent-resistant membranes on the function of intestinal sucrase-isomaltase. *Biol. Chem*. **390**, 545-549 (2009). doi:10.1515/BC.2009.077

422. Quaronis, A. Posttranslational Regulation of Sucrase-Isomaltase Expression in Intestinal Crypt and Villus Cells . *The Journal of Bilological Chemistry* **264**, 20000–20011 (1989).

423. Fujitani, N. *et al.* Total cellular glycomics allows characterizing cells and streamlining the discovery process for cellular biomarkers. *PNAS* (2013). doi:10.1073/pnas.1214233110

424. Schauer, R. Sialic acids as regulators of molecular and cellular interactions. *Curr. Opin. Struct. Biol.* 1–8 (2009). doi:10.1016/j.sbi.2009.06.003

425. Varki, A. Sialic acids in human health and disease Ajit. *Trend Mol. Med* **14**(8)**,** 351–360 (2009).

Appendix

Appendix 1. BLAStp hits to MapA query sequence against 54 gut micro-organisms. The table illustrates the sequence ID and amino acid alignment region of each homologous match for a specific gut organism.

Organism	BLAST Hits to MapA Query Sequence	Sequence ID	Region of MapA homology	MapA query amino acid residue (aa) range
Anaerotruncus colihominis	ABC transporter arginine-binding protein [Anaerotruncus colihominis]	CUP67114.1	Lig_chan-Glu_bd	37-169
	ABC transporter arginine-binding protein [Anaerotruncus colihominis]	WP_040342269.1	Lig_chan-Glu_bd	37-169
	ABC transporter arginine-binding protein [Anaerotruncus colihominis]	WP_070097947.1	Lig_chan-Glu_bd	46-169
	ABC transporter arginine-binding protein [Anaerotruncus colihominis]	EDS12629.1BA	Lig_chan-Glu_bd	46-169
Bacteroides xylanisolvens	Basic amino acid ABC transporter substrate-binding protein [Bacteroides xylanisolvens]	WP_055237269.1	SBP_bac_3	36-262
	Amino acid ABC transporter substrate-binding protein [Bacteroides xylanisolvens]	WP_055236521.1	SBP_bac_3	44-262
	Putative solute-binding protein precursor [Bacteroides xylanisolvens]	CUP20128.1	SBP_bac_3	42-262
	Hypothetical protein [Bacteroides xylanisolvens]	WP_055237596.1	SBP_bac_3	39-257
	Amino acid ABC transporter substrate-binding protein [Bacteroides xylanisolvens]	WP_055236939.1	SBP_bac_3	42-258
	Amino acid ABC transporter substratebinding protein [Bifidobacteria adolescentis]	WP_041777387.1	SBP_bac_3	6-260
	Putative ABC-type amino acid transport system periplasmic component [Bifidobacteria adolescentis ATCC 15703]	BAF40037.1	SBP_bac_3	6-260
	Glutamine ABC transporter permease [Bifidobacterium adolescentis]	WP_033499166.1	SBP_bac_3	7-258

Bifidobacterium adolescentis	Glutamine ABC transporter permease [Bifidobacterium adolescentis]	WP_003808718.1	SBP_bac_3	7-258
	Glutamine ABC transporter permease [Bifidobacterium adolescentis]	WP_055680406.1	SBP_bac_3	7-258
	Glutamine ABC transporter permease [Bifidobacterium adolescentis]	WP_041777273.1	SBP_bac_3	7-258
	Similar to glutamine ABC transporter (binding and transport protein)[Bifidobacterium adolescentis ATCC 15703]	BAF39358.1	SBP_bac_3	7-258
Bifidobacterium longum	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_060620799.1	SBP_bac_3	35-263
	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_012577147.1	SBP_bac_3	35-263
	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_065474456.1	SBP_bac_3	35-263
	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_008782667.1	SBP_bac_3	35-263
	MULTISPECIES: amino acid ABC transporter substrate-binding protein [Bifidobacterium]	WP_032684004.1	SBP_bac_3	35-263
	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_013141051.1	SBP_bac_3	35-263
Bifidobacterium infantis	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_060620799.1	SBP_bac_3	35-263
	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_012577147.1	SBP_bac_3	35-263
	Amino acid ABC transporter substrate-binding protein [Bifidobacterium longum]	WP_065474456.1	SBP_bac_3	35-263
	MULTISPECIES: amino acid ABC transporter substrate-binding protein [Bifidobacterium]	WP_008782667.1	SBP_bac_3	35-263
	ABC-type amino acid transport/signal transduction systems, periplasmic component/domain protein [Bifidobacterium animalis subsp. lactis B420]	AFJ16989.1	SBP_bac_3	9-261

Bifidobacterium lactis	amino acid ABC transporter substrate -binding protein [Bifidobacterium animalis]	WP_004218550.1	SBP_bac_3	9-261
	Bacterial extracellular solute-binding s. 3 family protein [Bifidobacterium animalis subsp. lactis CECT 8145]	CDL71827.1	SBP_bac_3	36-252
	ABC transporter substrate-binding protein [Bifidobacterium animalis]	WP_004218779.1	SBP_bac_3	36-252
Clostridium asparagiforme	ABC transporter substrate-binding protein [Clostridium asparagiforme]	WP_007719436.1	SBP_bac_3	26-262
Clostridium leptum	Amino acid ABC transporter substrate-binding protein [Clostridium leptum]	WP_003532176.1	SBP_bac_3	15-262
Clostridium nexile	MULTISPECIES: ABC transporter substrate-binding protein [Tyzzerella]	WP_004613944.1	SBP_bac_3	12-262
Clostridium scindens	ABC transporter substrate-binding protein [Clostridium scindens]	WP_025643718.1	SBP_bac_3	42-258
	ABC transporter substrate-binding protein [Clostridium scindens]	WP_004606333.1	SBP_bac_3	42-258
	Amino acid ABC transporter substrate-binding protein [Clostridium scindens]	WP_004607876.1	SBP_bac_3	7-262
Collinsella aerofaciens	Amino acid ABC transporter substrate-binding protein [Collinsella aerofaciens]	WP_055309752.1	SBP_bac_3	36-256
	Amino acid ABC transporter substrate-binding protein [Collinsella aerofaciens]	WP_022094421.1	SBP_bac_3	36-256
	Amino acid ABC transporter substrate-binding protein [Collinsella aerofaciens]	WP_006234383.1	SBP_bac_3	36-256
	MULTISPECIES: amino acid ABC transporter substrate-binding protein [Collinsella aerofaciens]	WP_035138867.1	SBP_bac_3	36-256
Dorea formicigenerans	Hypothetical protein [Dorea formicigenerans]	WP_005337487.1	SBP_bac_3	7-262
	Amino acid ABC transporter substrate-binding protein [Dorea formicigenerans]	WP_005337434.1	SBP_bac_3	42-262
	ABC transporter substrate-binding protein [Dorea longicatena]	WP_055180967.1	SBP_bac_3	42-256

Dorea longicatena	ABC transporter substrate-binding protein [Dorea longicatena]	WP_006426768.1	SBP_bac_3	42-256
	ABC transporter substrate-binding protein [Dorea longicatena]	WP_055281231.1	SBP_bac_3	42-256
	ABC transporter substrate-binding protein [Dorea longicatena]	WP_028086265.1	SBP_bac_3	42-256
	ABC transporter substrate-binding protein [Dorea longicatena DSM 13814]	WP_044920568.1	SBP_bac_3	7-262
	Hypothetical protein [Dorea longicatena]	WP_006428668.1	SBP_bac_3	7-262
	Amino acid ABC transporter substrate-binding protein [Dorea longicatena]	WP_055182572.1	SBP_bac_3	7-262
	Phage head-tail adapter protein [Dorea longicatena]	EDM62058.1	SBP_bac_3	7-262
	Amino acid ABC transporter substrate-binding protein [Dorea longicatena]	WP_049940493.1	SBP_bac_3	42-262
	Amino acid ABC transporter substrate-binding protein [Dorea longicatena]	WP_049947079.1	SBP_bac_3	42-262
	L-cystine-binding protein tcyA precursor [Dorea longicatena]	WP_055180507.1	SBP_bac_3	52-260
	Hypothetical protein [Dorea longicatena]	CUO51197.1	SBP_bac_3	52-260
	Hypothetical protein [Dorea longicatena]	WP_028086805.1	SBP_bac_3	52-260
	Hypothetical protein [Dorea longicatena]	WP_006428363.1	SBP_bac_3	52-260
Enterococccus faecalis TX0104	ABC transporter, substrate-binding protein, family 3 [Enterococcus faecalis TX0104]	EEI12735.1	SBP_bac_3	42-262
	Amino acid ABC transporter amino-acid-binding protein [Enterococcus faecalis]	WP_010713855.1	SBP_bac_3	42-262
	ABC transporter, permease protein [Enterococcus faecalis TX0104]	EEI10544.1	SBP_bac_3	33-254
	MULTISPECIES : glutamine ABC transporter substrate-binding protein [Bacilli]	WP_002381368.1	SBP_bac_3	33-254
	MULTISPECIES : glutamine ABC transporter permease [Bacilli]	WP_002355770.1	SBP_bac_3	24-260

Escherichia coli	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_047083357.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_063091033.1	SBP_bac_3	5-262
	MULTISPECIES : cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_040079243.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_046201444.1	SBP_bac_3	5-262
	Cysteine binding periplasmic protein [Escherichia coli]	WP_053294188.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	APK15967.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001639629.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_053880423.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_038339390.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_060634210.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001317901.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_012311600.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_047625202.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherchia coli]	WP_062873022.1	SBP_bac_3	5-262
	Cystine-bindng periplasmic protein [Escherichia coli]	WP_069914621.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001643606.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_032609574.1	SBP_bac_3	5-262

	Cystine binding periplasmic protein [Escherichia coli]	WP_024194793.1	SBP_bac_3	5-262
	Cysteine binding periplasmic protein [Escherichia coli OK1357]	WP_069185376.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024240712.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_040234760.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	CTW09376.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli]	ESK36174.1	SBP_bac_3	5-262
	Bacterial extracellular solute-binding s, 3 family protein [Escherichia coli MP021017.10]	WP_024192602.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli]	EF185938.1	SBP_bac_3	5-262
	MULTISPECIES: cysteine-binding periplasmic protein [Escherichia coli]	OJS89335.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_047660002.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	WP_032238953.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate binding protein [Escherichia coli]	WP_001353139.1	SBP_bac_3	5-262
	Cysteine-binding periplasmic protein [Escherichia coli KTE233]	WP_024191042.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_062894301.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	KGM81482.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_032163926.1	SBP_bac_3	5-262
	Cystine-binding periplasmic protein [Escherichia coli]	WP_001371765.1	SBP_bac_3	5-262
	Cysteine-binding periplasmic protein [Escherichia coli]	WP_063024787.1	SBP_bac_3	5-262

	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024202009.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_042029065.1	SBP_bac_3	5-262
	MULTISPECIES: cysteine ABC transporter substrate-binding protein [Enterobacteriacie]	WP_0121358998.1	SBP_bac_3	5-262
	Cysteine-binding periplasmic protein [Escherichia coli HVH 3(4-7276001)]	KYZ96195.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001726157.1	SBP_bac_3	5-262
	Putative periplasmic binding transport protein [Escherichia coli FRIK1996]	WP_024189630.1	SBP_bac_3	5-262
	MULTISPECIES: cystine-binding periplasmic protein [Enterobacteriaceae]	WP_012602280.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_060773473.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	EIE37190.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001296168.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024238520.1	SBP_bac_3	5-262
	MULTISPECIES: cystine ABC transporter substrate-binding protein [Enterobacteriaceae]	WP_014640806.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_032249859.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_044078643.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_053276058.1	SBP_bac_3	5-262
	Cystine-binding periplasmic protein [Escherichia coli OK1180]	WP_033545505.1	SBP_bac_3	5-262

	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001559844.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli J53]	E1064100.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001302033.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	ELD90797.1	SBP_bac_3	5-262
	Cystine-binding periplasmic protein [Escherichia coli]	WP_032218390.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024181879.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_053893623.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024195788.1	SBP_bac_3	5-262
	Cystine-binding periplasmic protein [Escherichia coli]	WP_001295643	SBP_bac_3	5-262
	MULTISPECIES: cystine ABC transporter substrate-binding protein [Proteobacteriae]	WP_001374794.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001369124.1	SBP_bac_3	5-262
	Cystine-binding periplasmic protein [Escherichia coli]	WP_059332905.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	WP_024216544.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	WP_069897575.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024787871.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_024187504.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli]	WP_023142110.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	EQV84877.1	SBP_bac_3	5-262

	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_044723144.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_002855922.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_053295300.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	EQN13087.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli KTE47]	ELE72282.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli ECA-727]	ELG26918.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	CTS64340.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	AAN80794.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	CTU02332.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	ERA00958.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	EQ093819.1	SBP_bac_3	5-262
	Cystine transporter subunit [Escherichia coli]	EQY61050.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli MS 196-1]	WP_032319232.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli UMEA 3323-1]	WP_000779810.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli MS O157:H7 str. EC4501]	EST81900.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli KOEGE 70 (185a)]	WP_024190803.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli KTE78]	WP_012578930.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli KTE25]	EQ088492.1	SBP_bac_3	5-262

	Cystine binding periplasmic protein [Escherichia coli KTE64]	ESL37423.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	EDU83971.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein precursor [Escherichia coli CFT073]	WP_024194820.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli HVH 139 (4-3192644)]	EDV61663.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli HVH 46 (4-2758776)]	WP_001480520.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli UMEA 3718-1]	ETJ22723.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli HVH 53 (4-0631051)]	WP_000494876.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_932191624.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli DORA A 5 14 21]	WP_059326480.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_044717717.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli UMEA 3240-1]	WP_042199241.1	SBP_bac_3	5-262
	Cystine ABC transporter substrate-binding protein [Escherichia coli]	WP_001358499.1	SBP_bac_3	5-262
	Cystine binding periplasmic protein [Escherichia coli HVH 33 (4-2174936)]	ERA06094.1	SBP_bac_3	5-262
Eubacterium halli	Basic amino acid ABC transporter substrate-binding protein [Eubacterium hallii]	WP_005351069.1	SBP_bac_3	1-257
	ABC transporter substrate-binding protein [Eubacterium hallii]	WP_022170732.1	SBP_bac_3	44-256
	Glutamine ABC transporter substrate-binding protein [Eubacterium hallii]	WP_055182378.1	SBP_bac_3	44-256

	Glutamine ABC transporter substrate-binding protein [Eubacterium hallii]	WP_005346454.1	SBP_bac_3	44-256
	Hypothetical protein [Eubacterium hallii]	WP_055182831.1	SBP_bac_3	7-256
	ABC transporter substrate-binding protein [Eubacterium hallii]	EEG37285.1	SBP_bac_3	7-256
	ABC transporter, substrate-binding protein, family 3 [Eubacterium hallii]	WP_022170102.1	SBP_bac_3	7-256
Eubacterium ventriosum	Hypothetical protein [Eubacterium ventriosum]	WP_005338969.1	SBP_bac_3	36-259
	ABC transporter, substrate-binding protein, family 3 [Eubacterium ventriosum ATCC 27560]	WP_040446182.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein [Eubacterium ventriosum]	EDM51593.1	SBP_bac_3	8-263
Faecalibacterium prausnitzii SL3/3	Amino acid ABC transporter substrate-binding protein , PAAT family (TC 3.A.1.3.-)[Faecalibacterium prausnitzii SL3/3]	CBL01004.1	SBP_bac_3	40-262
	Amino acid ABC transporter substrate-binding protein , PAAT family (TC 3.A.1.3.-)[Faecalibacterium prausnitzii SL3/3]	CBL01019.1	SBP_bac_3	37-259
	Amino acid ABC transporter substrate-binding protein , PAAT family (TC 3.A.1.3.-)[Faecalibacterium prausnitzii SL3/3]	CBL01671.1	SBP_bac_3	40-262
Holdemania filiformis	Amino acid ABC transporter substrate-binding protein [Holdemania filiformis]	WP_006059692.1	SBP_bac_3	34-262
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	CAC05301.1	SBP_bac_3	1-263
	Mucus adhesion promoting protein [Lactobacillus reuteri]	WP_003675946.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_003666279.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_35169574.1	SBP_bac_3	1-263

Lactobacillus reuteri	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_042746500.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_003667327.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_0019252231.1	SBP_bac_3	1-263
	Collagen-binding protein [Lactobacillus reuteri]	WP_035154961.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_019252998.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_003670552.1	SBP_bac_3	1-263
	ABC type amino acid transport/signal transduction system, periplasmic component/domain [Lactobacillus reuteri]	CUR39694.1	SBP_bac_3	1-263
	Mucus adhesion promoting protein [Lactobacillus reuteri]	WP_0658676461	SBP_bac_3	28-263
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_0655336581.1	SBP_bac_3	39-258
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	ADN22849.1	SBP_bac_3	39-260
	Amino acid ABC transporter substrate-binding protein {Lactobacillus reuteri]	AGZ84812.1	SBP_bac_3	39-260
	High affinity cystine binding protein [Lactobacillus reuteri]	AAC45332.1	SBP_bac_3	1-263
	ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_003666411.1	SBP_bac_3	20-260
	ABC transporter substrate-binding protein {Lactobacillus reuteri]	WP_003676256.1	SBP_bac_3	39-258
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_019252155.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_042746477.1	SBP_bac_3	20-260

	ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_03155083.1	SBP_bac_3	20-260
	ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_019252400.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_0351558321.1	SBP_bac_3	20-260
	ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_003670451.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_042746792.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_065532738.1	SBP_bac_3	20-260
	ABC transporter, substrate-binding protein, family 3 [Lactobacillus reuteri SD2112]	WP_003676073.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_003666657.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_003667610.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	CUR39405.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_065867875.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_035169630.1	SBP_bac_3	20-260
	Amino acid ABC transporter, amino acid-binding protein [Lactobacillus reuteri]	WP_003667453.1	SBP_bac_3	20-260
	glutamine ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_041816976.1	SBP_bac_3	20-260
	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	AE157954.1	SBP_bac_3	20-260
	ABC transporter, substrate-binding protein, family 3 [Lactobacillus reuteri CF48-3A]	WP_019253286.1	SBP_bac_3	20-260

Lactobacillus reuteri 1063	Amino acid ABC transporter substrate-binding protein [Lactobacillus reuteri]	WP_035154231.1	SBP_bac_3	1-263
	ABC transporter substrate binding protein [Lactobacillus reuteri]	WP_042746157.1	SBP_bac_3	39-258
	Amino acid ABC transporter substrate-binding protein [Lactobacillus plantarum]	CUR42875.1	SBP_bac_3	20-260
Lactobacillus plantarum WCSF1	Amino acid ABC transporter substrate-binding protein [Lactobacillus plantarum]	WP_035160181.1	SBP_bac_3	1-263
	Amino acid ABC transporter substrate-binding protein [Lactobacillus plantarum]	EE165863.1	SBP_bac_3	1-263
	glutamine ABC transporter substrate-binding protein [Lactobacillus plantarum]	WP_00345826.1	SBP_bac_3	57-258
Lactobacillus salivarius UCC118	Amino acid ABC transporter substrate-binding protein [Lactobacillus salivarius]	WP_003704326.1	SBP_bac_3	12-263
	glutamine ABC transporter substrate-binding protein [Lactobacillus salivarius]	WP_011476146.1	SBP_bac_3	45-258
	Polar amino acid ABC transporter substrate-binding protein [Lactobacillus salivarius]	WP_011475829.1	SBP_bac_3	44-158
	glutamine ABC transporter substrate-binding protein [Lactobacillus salivarius]	WP_011475829.1	SBP_bac_3	22-258
	glutamine ABC transporter substrate-binding protein [Lactobacillus salivarius]	WP_004563836.1	SBP_bac_3	57-258
Roseburia intestinalis M50/1	L-cystine ABC transporter periplasmic L-cystine-binding protein [Roseburia intestinalis]	WP_006857763.1	SBP_bac_3	15-236
Ruminococcus sp. SR15	Amino acid ABC transporter substrate-binding protein, PAAT family (TC 3.A.1.3.-) [Ruminococcus sp. SR1/5]	CBL212744.1	SBP_bac_3	7-256
	Glutamine ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_011227163.1	SBP_bac_3	37-260
	Glutamine ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_065424491.1	SBP_bac_3	37-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_011227361.1	SBP_bac_3	5-257

Streptococcus thermophilus	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_002953233.1	SBP_bac_3	5-262
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_059257414.1	SBP_bac_3	5-257
	ABC transporter substrate-binding protein [Streptococcus thermophilus]	WP_065973340.1	SBP_bac_3	5-262
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus TH1435]	ETE40947.1	SBP_bac_3	5-262
	L-cystine ABC transporter, periplasmic cystine-binding protein TcyA [Streptococcus thermophiles]	SCB62612.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_011688725.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_011225460.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	AKH34564.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_053042623.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus IF8CT]	EWM57738.1	SBP_bac_3	6-258
	ABC transporter substrate binding protein [Streptococcus thermophilus TH985]	EWM58229.1	SBP_bac_3	5-262
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_014621153.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_0530426224.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_011225465.1	SBP_bac_3	6-258
	ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_002952911.1	SBP_bac_3	37-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_002952516.1	SBP_bac_3	6-258

	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_064355633.1	SBP_bac_3	6-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus M17PTZA496]	ETW91508.1	SBP_bac_3	21-258
	ABC transporter substrate binding protein [Streptococcus thermophilus JIM 8232]	CCC19135.1	SBP_bac_3	66-258
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_041827091.1	SBP_bac_3	1-260
	ABC transporter substrate binding protein [Streptococcus thermophilus CNRZ1066]	AAV63184.1	SBP_bac_3	1-260
	L-cystine ABC transporter, periplasmic cystine-binding protein TcyA [Streptococcus thermophiles]	SCB63640.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_065972543.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_041828322	SBP_bac_3	1-260
	Polar amino acid ABC uptake transporter substrate binding protein [Streptococcus thermophiles LMG 18311]	AAV61257.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus TH1436]	ETE40295.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_014608653.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_014727671.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus TH1435]	ETE40470.1	SBP_bac_3	1-260
	Polar amino acid ABC uptake transporter substrate binding protein [Streptococcus thermophilus]	WP_014608653.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_014727671.1	SBP_bac_3	1-260

	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus M17PTZA496]	ETE40470.1	SBP_bac_3	1-260
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_002951750.1	SBP_bac_3	1-260
	ABC-type amino acid transport/ periplasmic component/domain protein [Streptococcus thermophilus MTCC 5460]	WP_002951750.1	SBP_bac_3	48-157
	ABC transporter substrate binding protein [Streptococcus thermophilus TH1436]	WP_0116815531.1	SBP_bac_3	22-262
	Amino acid ABC transporter substrate binding protein [Streptococcus thermophilus TH982]	ETW88370.1	SBP_bac_3	66-258
	ABC transporter substrate binding protein [Streptococcus thermophilus]	WP_071417332.1	SBP_bac_3	22-262
	ABC transporter substrate binding protein [Streptococcus thermophilus]	EWM61894.1	SBP_bac_3	22-262
	ABC transporter substrate binding protein [Streptococcus thermophilus TH1477]	WP_003214423.1	SBP_bac_3	22-262

Appendix 2. BLAStp hits to Msa query sequence against 54 gut micro-organisms. The table illustrates the sequence ID and amino acid alignment region of each homologous match for a specific gut organism.

Organism	BLAST hits to Msa Query Sequence	Sequence ID	Msa region of homology	Msa amino acid residue range
	Internalin, putative (LPTXG motif) [Lactobacillus reuteri]	CUR40724	3 MucBP domains	806-1010, 707-822
	Cell wall surface anchor family protein [Lactobacillus reuteri]	OJI10619.1	2 MucBP domains	707-822
	MucBP binding domain protein [Lactobcillus reuteri]	CUR38029.1	2 MucBP domains	646-1010
	MucBP binding domain protein [Lactobcillus reuteri]	WP_063164336.1	2 MucBP domains	723-915

Lactobacillus reuteri	Cell wall surface anchor family protein [Lactobacillus reuteri]	WP_003675299.1	2 MucBP domains	646-1010
	Signal peptide [Lactobacillus reuteri]	CUR43781.1	2 MucBP domains	723-915
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_016496451.1	2 MucBP domains	646-1010
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_020843133.1	2 MucBP domains	723-915
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_066036024.1	2 MucBP domains	646-1010
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_035156718.1	2 MucBP domains	723-915
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_065533047.1	2 MucBP domains	790-958
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_035153250.1	2 MucBP domains	726-1010
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_003665712.1	2 MucBP domains	790-958
	LPXTG-motif cell wall anchor domain protein [Lactobacillus reuteri]	WP_019253808.1	2 MucBP domains	798-958
	Hypothetical protein [Lactobacillus reuteri]	WP_035169032.1	2 MucBP domains	849-1010
	Hypothetical protein [Lactobacillus reuteri]	WP_035150883.1	2 MucBP domains	602-915
	Hypothetical protein [Lactobacillus reuteri]	WP_035161497.1	2 MucBP domains	827-1009
	Putative mucin binding protein [Lactobacillus reuteri ATCC 53608]	WP_065533287.1	2 MucBP domains	602-915
	Mucus binding protein precursor [Lactobacillus reuteri ATCC 53608]	CCC03122.1	2 MucBP domains	602-915
	Mucus binding protein [Lactobacillus reuteri ATCC 53608]	WP_00140377.1	2 MucBP domains	602-915
	Hypothetical protein [Lactobacillus reuteri]	AAF25576.1	2 MucBP domains	827-1009
	LPXTG-motif cell wall anchor domain protein [Lactobacillus reuteri]	CUU133321.I	2 MucBP domains	849-1010

	Hypothetical protein [Lactobacillus reuteri]	WP_066035528.1	2 MucBP domains	827-1009
	LPXTG-motif cell wall anchor domain protein [Lactobacillus reuteri]	WP_042746230.1	2 MucBP domains	827-1009
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_035156688.1	2 MucBP domains	827-1010
	Hypothetical protein [Lactobacillus reuteri]	WP_003664761.1	2 MucBP domains	827-1010
Lactobacillus reuteri 1063	MucBP binding domain protein [Lactobacillus reuteri]	WP_065867349.1	2 MucBP domains	646-1010
	Hypothetical protein [Lactobacillus reuteri]	WP_019251579.1	2 MucBP domains	723-915
	Putative mucin binding protein [Lactobacillus reuteri ATCC 53608]	CCC03122.1	2 MucBP domains	602-915
	Mucus binding protein precursor [Lactobacillus reuteri ATCC 53608]	AAF25576.1	2 MucBP domains	602-915
	Mucus binding protein [Lactobacillus reuteri ATCC 53608]	CUU13332.1	2 MucBP domains	602-915
Lactobacillus plantarum WCFS1	Adhesion [Lactobacillus plantarum]	WP_011101323.1	Sigal peptide, Lectin-type domain and three MucBP domains	1-1010
	Adherence-associated mucus binding protein, LPXTG-motif cell wall anchor [Lactobacillus plantarum]	WP_011102023.1	3 MucBP domains	602-954
Streptococcus thermophilus	Hypothetical protein [Streptococcus thermophilus]	WP_065973399.1	3 MucBP domains	806-1010, 707-822

Appendix 3 BLAStp hits to Mub query sequence against 54 gut micro-organisms. The table illustrates the sequence ID and amino acid alignment region of each homologous match for a specific gut organism.

Organism	BLAST hits to Mub Query Sequence	Sequence ID	Mub region of homology	Mub amino acid residue range
Lactobacillus reuteri	Mucus binding protein precursor [Lactobacillus reuteri ATCC 53608]	AAF25576.1	14 MucBP domains	1-3269
	Mucus binding protein [Lactobacillus reuteri ATCC 53608]	CUU13332.1	14 MucBP domains	1-3269
	Putative mucing binding protein [Lactobacillus reuteri ATCC 53608]	OJI10377.1	14 MucBP domains	1-2039, 1648-3269, 900-2407
	Hypothetical protein [Lactobacillus reuteri]	CCC03122.1	14 MucBP domains	1-1368, 2384-3183,900-1736, 1289-2104, 1832-2656, 2016-2840, 1648-3269
	Chain A, Type 2 Repeat Of the mucus binding protein Mub from Lactobacillus reuteri	WP_035161497.1	14 MucBP domains	1695-1878, 2063-2246,1511-1694,1879-2062,2247-2430, 2431-2614, 2615 - 2798, 1326-1510, 2799-2980,796-945
	Chain A, crystal structure of Mub-rv	3157A	14 MucBP domains	2799-2981, 947-1133, 2615-2797, 557-748, 750-945, 1511-1693, 1879-2061,2247-2429, 2431-2613, 1695-1877, 2063-2245
	Hypothetical protein [Lactobacillus reuteri]	4MT5A	14 MucBP domains	1-1368
	Mucus-binding protein [Lactobacillus reuteri]	WP_065533285.1	6 MucBP domains	2384-3183
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_003664760.1	6 MucBP domains	900-1736
	YSIRK signal domain/ LPXTG anchor domain surface protein [Lactobacillus reuteri]	WP_020843313.1	6 MucBP domains	1648-2472

	Mucus-binding protein [Lactobacillus reuteri]	WP_019251482.1	6 MucBP domains	1289-2104
	Hypothetical protein [Lactobacillus reuteri]	WP_066035527.1	6 MucBP domains	2016-2840
	Hypothetical protein [Lactobacillus reuteri]	WP_019253693.1	6 MucBP domains	1832-2656
	LPXTG-motif cell wall anchor domain protein [Lactobacillus reuteri]	WP_020843314.1	6 MucBP domains	2315-3185,2683-3186, 595-1515,1507-2067,2059-2619,471-1481,1691-2251,2611-3183,2427-3025,1476-1699..etc
Lactobacillus reuteri 1063	Mucus binding protein precurson [Lactobacillus reuteri ATCC 53608]	AAF25576.1	14 MucBP domains	1-3269
	Mucus binding protein [Lactobacillus reuteri ATCC 53608]	CUU13332.1	14 MucBP domains	1-3269
	Putative mucin binding protein [Lactobacillus reuteri ATCC 53608]	CCC03122.1	14 MucBP domains	1-2039
	Hypothetical protein [Lactobacillus reuteri]	WP_035161497.1	6 MucBP domains	1648-3269
	Chain A, Crystal structure Of Mub-rv	4MTSA	8 MucBP domains	900-2407
	MucBP binding domain protein [Lactobacillus reuteri]	WP_003675299.1	14 MucBP domains	1-1368
Lactobacillus plantarum WCFS1	Mucus binding-protein [Lactobacillus plantarum]	WP_011101486.1	14 MucBP domains	1865-2980, 1130-2217, 552-1665
	Mucus binding-protein [Lactobacillus plantarum]	WP_011102047.1	14 MucBP domains	2232-2980, 1864-2585, 1496-2234, 2048-2769, 745-1498, 942-1665, 2600-3184
	Adhesin [Lactobacillus plantarum]	WP_011101323.1	2 MucBP domains	845-1130

Streptococcus thermophilus	YSIRK signal domain/ LPXTG anchor domain surface protein [Streptococcus thermophilus]	WP_064411033.1	7 MucBP domains	1-746, 1295-1481,2607-2769,2423-2585,1503-1665,1871-2033,2239-2401
	YSIRK signal domain/ LPXTG anchor domain surface protein [Streptococcus thermophilus]	WP_064355458.1	14 MucBP domains	1-746, 1295-1481, 2607-2769, 242-2585, 1503-1665, 1871-2033, 2239-2401
	Cell surface protein [Streptococcus thermophilus]	EWM59458.1	14 MucBP domains	1-746, 1320-1481, 2607-2769, 2423-2585, 1503-1665, 1871-2033, 2239-2401
	YSIRK signal domain/ LPXTG anchor domain surface protein [Streptococcus thermophilus]	WP_024704016.1	14 MucBP domains	1-746, 1320-1481, 2607-2769, 2423-2585,1503-1665, 1871-2033, 2239-2401
	Putative cell surface protein [Streptococcus thermophilus]	AKH34802.1	14 MucBP domains	1-746, 1320-1481,2607-2769, 2423-2585, 1503-1665, 1871-2033, 2239-2401

Appendix 4 BLAStp hits to SusD query sequence against 54 gut micro-organisms. The table illustrates the sequence ID and amino acid alignment region of each homologous match for a specific gut organism.

Organism	BLAST hits to SusD Query Sequence	Sequence ID	SusD region of homology	SusD amino acid resiude range
Bacteroides caccae	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides caccae]	WP_055170775.1	SusD domain	2-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides caccae]	WP_005677618.1	SusD domain	2-550
	Hypothetical protein HMPREF1065 02751 [Bacteroides dorei CL03T12C01]	EIY36834.1	SusD domain	14-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides dorei]	WP_032951073.1	SusD domain	14-551
	Hypothetical protein HMPREF1063 02482 [Bacteroides dorei CL03T12C01]	EIY25736.1	SusD domain	14-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides dorei]	WP_032942989.1	SusD domain	14-551
	MULTISPECIES: RagB/SusD family nutrient uptake outer membrane protein [Bacteroides]	WP_032936085.1	SusD domain	14-551
	SusD family protein [Bacteroides dorei DSM 17855]	EEB25287.1	SusD domain	14-551

Bacteroides dorei	Membrane protein [Bacteroides dorei]	ALA76006.1	SusD domain	14-551
	Membrane protein [Bacteroides dorei]	AII67022.1	SusD domain	14-551
	MULTISPECIES: RagB/SusD family nutrient uptake outer membrane protein [Bacteroides]	WP_007841981.1	SusD domain	14-551
	SusD family protein [Bacteroides dorei DSM 17855]	EEB25292.1	SusD domain	14-551
	MULTISPECIES: RagB/SusD family nutrient uptake outer membrane protein [Bacteroides]	WP_005641955.1	SusD domain	14-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides dorei]	WP_007841624.1	SusD domain	14-551
	MULTISPECIES: RagB/SusD family nutrient uptake outer membrane protein [Bacteroides]	WP_007836714.1	SusD domain	14-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides dorei]	007840827.1	SusD domain	14-551
Bacteroides eggerthii	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides eggerthii]	WP_004291797.1	SusD domain	14-551
	Acyl-coenzyme A thioesterase 11[Bacteroides eggerthii 1 2 48FAA]	EFV31061.1	SusD domain	21-551
	MULTISPECIES: RagB/SusD family nutrient uptake outer membrane protein [Bacteroides]	WP_005641955.1	SusD domain	63-551
	Hypothetical protein HMPREF1016 03596 [Bacteroides eggerthii 1 2 48FAA]	EFV28197.1	SusD domain	63-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides eggerthii]	WP_004292516.1	SusD domain	9-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides eggerthii]	WP_004290209.1	SusD domain	9-550
Bacteroides finegoldii	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides finegoldii]	WP_055279545.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides finegoldii]	WP_007755008.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides finegoldii]	WP_007763852.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides finegoldii]	WP_007758865.1	SusD domain	6-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides finegoldii]	WP_007755958.1	SusD domain	6-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides finegoldii]	WP_055278193.1	SusD domain	6-550

Bacteroides fragilis 3_1_12	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides fragilis]	WP_032541748.1	SusD domain	1-550
	SusD family protein [Bacteroides fragilis 3_1_12]	EFR54720.1	SusD domain	21-550
Bacteroides intestinalis	SusD family protein [Bacteroides intestinalis DSM 17393]	EDV04003.1	SusD domain	1-550
	Starch binding outer membrane lipoprotein SusD [Bacteroides intestinalis]	WP_044155062.1	SusD domain	6-550
	Starch binding outer membrane lipoprotein SusD [Bacteroides intestinalis]	WP_021968266.1	SusD domain	6-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides intestinalis]	WP_061437914.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides intestinalis]	WP_061433851.1	SusD domain	1-550
	SusD family protein [Bacteroides intestinalis]	KXT54890.1	SusD domain	15-551
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacteroides intestinalis]	WP_052340627.1	SusD domain	22-551
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacteroides intestinalis]	WP_025725855.1	SusD domain	1-551
	SusD family protein [Bacteroides intestinalis]	KXT54679.1	SusD domain	1-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides intestinalis]	WP_044154509.1	SusD domain	1-551
	SusD family protein [Bacteroides intestinalis DSM 17393]	EDV07661.1	SusD domain	1-551
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacteroides intestinalis]	WP_007211493.1	SusD domain	15-551
	Starch-binding outer membrane lipoprotein SusD [Bacteroides ovatus]	WP_004299755.1	SusD domain	15-551
	Starch-binding outer membrane lipoprotein SusD [Bacteroides ovatus]	WP_004305041.1	SusD domain	15-551
	Starch-binding outer membrane lipoprotein SusD [Bacteroides ovatus]	SDB75688.1	SusD domain	15-551
	Starch-binding outer membrane lipoprotein SusD [Bacteroides ovatus]	WP_004310515.1	SusD domain	15-551
	MULTISPECIES : starch-binding outer membrane lipoprotein SusD [Bacteroides ovatus]	WP_022198967.1	SusD domain	15-551
	Starch-binding outer membrane lipoprotein SusD [Bacteroides ovatus]	SDH87047.1	SusD domain	15-551

Bacteroides ovatus				
Bacteroides stercoris	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides stercoris]	WP_005654647.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides stercoris]	WP_016661434.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides stercoris]	WP_060386044.1	SusD domain	6-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides stercoris]	WP_005655694.1	SusD domain	6-551
	SusD family protein [Bacteroides stercoris]	KWR57214.1	SusD domain	6-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides stercoris]	WP_060385269.1	SusD domain	24-551
Bacteroides thetaiotaomicron VP1-5482	MULTISPECIES : starch-binding outer membrane lipoprotein SusD [Bacteroides]	WP_008767005.1	SusD domain	1-551
	Chain A, B. Thetaiotaomicron SusD with Alpha-Cyclodextrin	3CK7 A	SusD domain	26-551
	Chain A, B. Thetaiotaomicron SusD	3CKC A	SusD domain	26-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides thetaiotaomicron]	WP_011109369.1	SusD domain	22-544
	MULTISPECIES : starch-binding outer membrane lipoprotein SusD [Bacteroides]	WP_004295296.1	SusD domain	63-551
	Starch-binding outer membrane lipoprotein SusD [Bacteroides uniformis]	WP_016273189.1	SusD domain	6-550
	MULTISPECIES : Starch-binding outer membrane lipoprotein SusD [Bacteroides uniformis]	WP-009036582.1	SusD domain	6-550

Bacteroides uniformis	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_057089225.1	SusD domain	6-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_057252892.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_35449523.1	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_005826347.1	SusD domain	1-547
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_044469067.1	SusD domain	1-550
	Hypothetical protein C801 03967 [Bacteroides uniformis dnLKV2]	EDS05370.1	SusD domain	7-550
	MULTISPECIES : Starch-binding outer membrane lipoprotein SusD [Bacteroides uniformis]	WP_005833437.1	SusD domain	4-547
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_057281716.1	SusD domain	7-550
	SusD family [Bacteroides uniformis]	CUP78796.1	SusD domain	153-551
	Hypothetical protein [Bacteroides uniformis]	WP_070101181.1	SusD domain	153-551
	SusD family protein [Bacteroides uniformis]	CUP36557.1	SusD domain	63-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_016274441.1	SusD domain	63-551
	SusD family protein [Bacteroides uniformis]	CUP66345.1	SusD domain	63-551
	MULTISPECIES : Starch-binding outer membrane lipoprotein SusD [Bacteroides uniformis]	WP_005826118.1	SusD domain	9-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	WP_061412322.1	SusD domain	9-550
	MULTISPECIES : Starch-binding outer membrane lipoprotein SusD [Bacteroidales]	WP_061412322.1	SusD domain	9-550
	MULTISPECIES : Starch-binding outer membrane lipoprotein SusD [Bacteroides]	WP_005641955.1	SusD domain	9-550
	SusD family protein [Bacteroides uniformis]	WP_004295296.1	SusD domain	6-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides uniformis]	KXT34379.1	SusD domain	7-550
	Hypothetical protein HMPREF1072 02714 [Bacteroides uniformis CL03T00C23]	WP_061411947.1	SusD domain	9-550
	MULTISPECIES : Starch-binding outer membrane lipoprotein SusD [Bacteria]	EIY73745.1	SusD domain	9-550

Bacteroides vulgatus	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacteroidales]	WP_0345233611	SusD domain	4-550
	Putative lipoprotein [Bacteroides vulgatus PC510]	EFG19623.1	SusD domain	4-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_032944578.1	SusD domain	4-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_005849638	SusD domain	1-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_011965159.1	SusD domain	1-550
	Chain A, Crystal Structure of SusD superfamily protein (Yp 001298690. 1) From Bacteroides Vulgatus ATCC 8482 At 2.00 A Resolution	3JYS A	SusD domain	73-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_01671143.1	SusD domain	7-550
	SusD, outermembrane protein [Bacteroides vulgatus]	ALK84216.1	SusD domain	226-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_012055744.1	SusD domain	63-551
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacteroidales]	WP_005641955.1	SusD domain	63-551
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_032952938.1	SusD domain	143-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_011965061.1	SusD domain	143-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_057279223.1	SusD domain	143-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides vulgatus]	WP_011964774.1	SusD domain	153-550
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacterodies]	WP_005840742.1	SusD domain	153-550
Bacteroides xylanisolvens	MULTISPECIES: starch-binding outer membrane lipoprotein SusD [Bacteroides]	WP_008024454.1	SusD domain	1-550
	SusD family [Bacteroides xylanisolvens XB1A]	CBK69610.1	SusD domain	1-550
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacterodies]	WP_008016762.1	SusD domain	1-550
	Putative lipoprotein [Bacteroides xylanisolvens SD CC 2a]	EFF59313.1	SusD domain	1-337
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacterodies]	WP_008024544.1	SusD domain	6-550

	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides xylanisolvens]	WP_004314446.1	SusD domain	404-550
	Starch binding associating with outer membrane protein [Bacteroides xylanisolvens]	SEA43044.1	SusD domain	94-550
	RagB/SusD family nutrient uptake outer membrane protein [Bacteroides xylanisolvens]	WP_008023866.1	SusD domain	94-550
	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacterodies]	WP_004316159.1	SusD domain	130-550
Parabacteroides johnsonii	RagB/SusD family nutrient uptake outer membrane protein [Parabacteroides johnsonii]	WP_008145839.1	SusD domain	139-550
Parabacteroides merdae	MULTISPECIES : RagB/SusD family nutrient uptake outer membrane protein [Bacteroidales]	WP_005641955.1	SusD domain	63-551

Appendix 5 BLAStp hits to LspA query sequence against 54 gut micro-organisms. The table illustrates the sequence ID and amino acid alignment region of each homologous match for a specific gut organism.

Organism	BLAST hits to LspA Query Sequence	Sequence ID	LspA region of homology	LspA amino acid residue range
Lactobacillus plantarum WCSF1	Mucus-binding protein [Lactobacillus plantarum]	WP_011101221.1	All 8 MucBP domains	492-1194
	Mucus-binding protein [Lactobacillus plantarum]	WP_011101804.1	6 MucBP domains	726-1118
Lactobacillus salivarius UCC118	Mucus-binding protein [Lactobacillus salivarius]	WP_011475677.1	All 8 MucBP domains	1-1209