

**Functional and Metagenomic Analysis of the Human Tongue  
Dorsum using Phage Display**

**Submitted for the degree of Doctorate of Philosophy**

**Samantha Easton**

**September 2009**

**University College London**

**Department of Structural and Molecular Biology**

**Gower Street, London, WC1E 6BT**

**Declaration:**

I, Samantha Easton, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

<b>Contents</b>	<b>Page number</b>
<b>Table of Contents</b>	<b>3</b>
<b>List of Figures</b>	<b>7</b>
<b>List of Tables</b>	<b>9</b>
<b>Abbreviations and Measurements</b>	<b>10</b>
<b>Abstract</b>	<b>11</b>
<b>Chapter 1: Introduction</b>	<b>13</b>
General introduction	14
Friendly bacteria	16
Bacterial colonization	16
Microbial communities	17
Oral cavity, saliva and tongue dorsum	19
Host cell biology	21
Bacterial architecture	23
Adhesins	25
Bacteria-host interactions	25
Types of binding interaction	26
Innate microbial recognition	28
Bacterial research	29
Metagenomics for community analysis	31
Constructing a metagenomic library	33
Phage display	35
Phage life cycle	36
Phage display application	38
Affinity selection – Biopanning	40
Panning ligand – IgA	42
Panning ligand – Fibronectin	43
Aim of this project	45
<b>Chapter 2: Materials &amp; Methods</b>	<b>46</b>
(a) Media	47
(b) Solutions	47
Loading buffer	47
$\lambda$ Pst agarose gel ladder	47
40% PEG (polyethylene glycol)/2.5M NaCl	48
Sodium carbonate buffer	48
PBS and PBS-T	48
10% CTAB solution	48
TE buffer with RNase	48
TE + 20% glycerol	49
BCIP/NBT	49
(c) Strains, Bacteriophage and Vectors	49
Preparation of glycerol stocks	49
Phagemid vector isolation	50
Phagemid vector digestion and dephosphorylation	51
(d) Bacterial DNA Sampling, Extraction and Preparation	51
Tongue scraping protocol	51

Sample storage	51
DNA extraction by CTAB method	52
Ethanol precipitation	52
(e) Phage Display Library Production	53
DNA fragmentation and repair	53
DNA visualisation using agarose gels	53
Ligation and purification	54
Preparation of electrocompetent cells	54
Preparation of chemically competent cells	54
Electroporation	55
Chemical transformation	55
Plasmid extraction (Miniprep)	55
(f) Phagemid library conversion	56
Enumeration of recombinants	56
(g) Biopanning	56
(h) Sequencing	57
Primers	57
<i>In silico</i> analysis	58
(i) Antibody Screening	58
Isolation of phage supernatant	58
Antibody screening against anti Poly-His	58
Antibody screening against anti c-Myc and alkaline phosphatase conjugate	59
(j) SDS-PAGE	59
(k) pET Vector Expression	60
Individual primer design	60
Individual clone amplification	60
TOPO TA cloning	60
pET vector expression	61
(l) Adhesion Assays	61
(m) 16S rRNA Gene Diversity Analysis	62
16S rRNA gene Amplification	62
Cloning of 16S rRNA gene amplified DNA	62
16S rRNA gene data analysis	63

### **Chapter 3: Constructing a Phage Display Library** **65**

Introduction	66
DNA sampling and concentration	66
DNA fragmentation	68
Determination of optimum ligation conditions	69
Transformation	71
Insert frequency and size following transformation	72
Conversion of phagemid to phage display	73
PEG precipitation	74
Additional rinse	74
Final library considerations	75
Discussion	76
DNA extraction method	77
Library construction	79

<b>Chapter 4: Panning &amp; Antibody Screening</b>	<b>81</b>
Introduction	82
Panning procedure	83
Panning results	
(a) IgA	84
(b) Fibronectin	85
Bioinformatic screening of fusion proteins	85
Antibody screening	88
Antibody screening colour change results	
(a) IgA	89
(b) Fibronectin	90
(c) BSA	91
Discussion: Panning considerations	94
Antibody screening	96
<b>Chapter 5: Individual Protein Analysis</b>	<b>98</b>
Introduction	99
Final 18 clones	99
pET expression	103
Adhesion assays	105
Individual proteins	113
<b>Chapter 6: 16S rRNA Diversity Analysis</b>	<b>116</b>
Introduction	117
Initial analysis	117
Full 16S rRNA gene analysis	118
16S rRNA gene data analysis	118
16S rRNA gene analysis results	119
Novel bacteria	120
Discussion	123
<b>Chapter 7: Additional Work - pQR492</b>	<b>127</b>
Introduction	128
Investigating pQR492	128
<i>In silico</i> analysis	131
<b>Chapter 8: Discussion</b>	<b>149</b>
Introduction	150
Phage display	151
Panning	154
Antibody screening	156
Individual proteins	157
Binding affinity confirmation	159
16S rRNA gene analysis	160
Comparison of 16S rRNA gene and phage display analyses	161
pQR492	163
Future experiments	164

Conclusion	164
<b>Reference List</b>	<b>166</b>
<b>Appendix 1: Sample Sheet for Volunteers</b>	<b>180</b>
<b>Appendix 2: Restriction Map of <math>\lambda</math>Pst Ladder</b>	<b>182</b>
<b>Appendix 3: IgA Panning BLAST Results</b>	<b>183</b>
<b>Appendix 4: FN Panning BLAST Results</b>	<b>191</b>
<b>Appendix 5: BSA Panning BLAST Results</b>	<b>203</b>
<b>Appendix 6: 16S rRNA Gene Phylogenetic Tree</b>	<b>212</b>
<b>Appendix 7: pQR492 Text Sequence</b>	<b>214</b>
<b>Appendix 8: Final 18 Clones: InterProScan Analysis</b>	<b>217</b>

<b>List of Figures</b>	<b>Page number</b>
<b>Chapter 1: Introduction</b>	
Figure 1 Phylogeny of the Living World	15
Figure 2 Biofilm development	18
Figure 3 Diagrams of the Human Tongue Dorsum	21
Figure 4 Stratified Epithelia	23
Figure 5 Gram Positive Cell Wall	24
Figure 6 Gram Negative Cell Wall	24
Figure 7 Construction of a Metagenomic Library	34
Figure 8 Structure of Filamentous Phage	35
Figure 9 M13 Life Cycle	37
Figure 10 Phagemid Displaying Fusion Peptides on Gene VIII	38
Figure 11 Filamentous Phage Genes and Gene Products	39
Figure 12 pG8H6 Phagemid	40
Figure 13 Panning Process	41
Figure 14 Schematic representations of IgA1 and S-IgA	42
Figure 15 Schematic representation of Fibronectin molecule	44
<b>Chapter 3: Production of a Phage Display Library</b>	
Figure 1 DNA concentration following treatment with and without isopropanol	67
Figure 2 Production of a phage display library	67
Figure 3 Sonication of DNA sample in 5 second intervals	68
Figure 4 DNA fragmentation by partial digestion	69
Figure 5 Ligation reactions A and B	70
Figure 6 Phagemid vector pG8H6 containing metagenomic DNA	72
Figure 7 Library insert size before and after conversion	73
<b>Chapter 4: Panning &amp; Antibody Screening</b>	
Figure 1 Panning process	82
Figure 2 Schematic representation of panning a phage display	83
Figure 3 Numbers of bound phage after 3 rounds of panning	84
Figure 4 pG8H6 sequence showing poly-His and c-Myc tags	86
Figure 5 Schematic representation of steps taken following antibody screening	88
Figure 6 (a) and (b). Antibody screening results from phage display library panned against IgA.	89
Figure 7 (a) and (b). Antibody screening results from phage display library panned against FN.	90
Figure 8 (a) and (b). Antibody screening results from phage display library panned against BSA.	91
<b>Chapter 5: Individual Protein Analysis</b>	
Figure 16 Schematic drawing of the phagemid pG8H6	104
Figure 2 Adhesion assay results from control experiment	107
Figure 3 Adhesion assay results with clone 30	108
Figure 4 Adhesion assay results with clone 44	109
Figure 5 Adhesion assay results with clone 39	110

Figure 6 Adhesion assay results with clone 59	111
Figure 7 InterProScan result of clone number 59	111

### **Chapter 7: Additional Work - pQR492**

Figure 1 pQR492 restriction map	129
Figure 2 pQR492 showing <i>Pst</i> I sites	130
Figure 3 Sequencing of pQR492 insert DNA	131
Figure 4 Diagram showing potential ORF's of pQR492	133
Figure 5 Possible gene content of pQR492	135
Figure 6 Possible amino acid sequence of pQR492	135
Figure 7 Alignment of putative ORF's of pQR492 with cauri_0414	139
Figure 8 <i>Corynebacterium aurimucosum</i> genome fragment	140
Figure 9 Results of Pfam search of cauri_0414 of <i>C. aurimucosum</i>	141
Figure 10 Results of Pfam search of ORF's 1 – 4 of pQR492	141
Figure 11 InterProScan analysis of ORF 1	142
Figure 12 Potential ribosome binding sites of ORF 1	143
Figure 13 Potential ribosome binding sites of ORF 2	144
Figure 14 Potential ribosome binding sites of ORF 3	144
Figure 15 Potential ribosome binding sites of ORF 4	145
Figure 16 Clustalw alignment of ORF 5 with TRCF	146
Figure 17 InterProScan analysis of ORF 5	147
Figure 18 InterProScan analysis of full length TRCF	147
Figure 19 Potential ribosome binding sites of ORF 5	147



<b>List of Tables</b>	<b>Page number</b>
<b>Chapter 1: Introduction</b>	
Table 1 Salivary Components	20
Table 2 Human Cell Membrane Constituents	22
Table 3 Human Membrane Components	27
<b>Chapter 2: Materials &amp; Methods</b>	
Table 1 Growth media	47
Table 2 Names and Sources of Bacterial Strains	49
Table 3 Names and Sources of Bacterial and Phagemid Vectors	50
Table 4 List of Primers for Sequencing	57
Table 5 Primers for pET vector expression	61
<b>Chapter 3: Production of a Phage Display Library</b>	
Table 1 Number of colonies for ligations with various insert: vector ratios	70
Table 2 Electroporation transformation frequency	71
Table 3 Titres of 3 stages during PEG precipitation	74
Table 4 Electroporation efficiency using Invitrogen ECC	75
<b>Chapter 4: Panning &amp; Antibody Screening</b>	
Table 1 Initial sequence analysis of panning experiments	87
Table 2 Antibody screening experiment results	93
<b>Chapter 5: Individual Protein Analysis</b>	
Table 1 Table of 18 clones	100
Table 2 Predicted protein products of final 18 recombinant clones	113
<b>Chapter 6: 16S rRNA Diversity Analysis</b>	
Table 1 Similarity of 15 'novel' phylotypes to public database	120
<b>Chapter 7: Additional Work - pQR492</b>	
Table 1 tBLASTx homology of pQR492	132

## Abbreviations and symbols

<b>aa</b>	amino acid
<b>Amp</b>	ampicillin
<b>BLAST</b>	basic local alignment search tool
<b>BSA</b>	bovine serum albumin
<b>bp</b>	base pairs
<b>CFU</b>	colony forming units
<b>DNA</b>	deoxyribonucleic acid
<b>ECC</b>	electrocompetent cells
<b>FN</b>	fibronectin
<b>g</b>	grams
<b>Gb</b>	gigabase pairs
<b>HSA</b>	human serum albumin
<b>Ig</b>	immunoglobulin
<b>kb</b>	kilobase pairs
<b>Mb</b>	megabase pairs
<b>MCS</b>	multiple cloning site
<b>ml</b>	millilitres
<b>MOI</b>	multiplicity of infection
<b>OD</b>	optical density
<b>ORF</b>	open reading frame
<b>PCR</b>	polymerase chain reaction
<b>PEG</b>	poly-ethylene glycol (8000)
<b>pH</b>	$-\log_{10} [\text{H}^+]$
<b>RBS</b>	ribosome binding site
<b>RDP</b>	ribosomal database project
<b>rpm</b>	revolutions per minute
<b>rRNA</b>	ribosomal ribonucleic acid
<b>SDS-PAGE</b>	sodium dodecyl sulphate polyacrylamide gel electrophoresis
<b>U</b>	unit
<b>µl</b>	microlitres
<b>°C</b>	degrees centigrade

## Abstract

It is well established that mixed microbial communities contain organisms which have not been studied by conventional culture-based methods. In the human oral cavity this number is estimated at around 50%. Commensal bacteria develop and maintain an intimate relationship with human cells without triggering proinflammatory mechanisms and this study aims to explore this by searching for bacterial proteins which facilitate binding to the human tongue dorsum and wider oral cavity.

Metagenomic DNA from the human tongue dorsum of 9 volunteers was extracted and a phage display library created, to our knowledge the first to incorporate metagenomic DNA. Phage display is an elegant molecular technique involving fusion of fragmented DNA to a phagemid coat protein, such that inserted DNA is encoded by the phage and displayed on the phage surface. The affinity selection technique panning, then exploited the natural affinity and specificity of the fusion proteins to identify bacterial binding proteins using, in this case, three ligands: IgA, Fibronectin and BSA. IgA is of special interest to this group as it interacts with bacterial proteins and is poised to respond to bacterial numbers in human secretions such as saliva. Proteins from panning were analysed *in silico*, however, the majority were discarded due to the presence of stop codons in the protein sequences. Remaining phagemid displaying fusion proteins of interest were assessed for function and binding assays carried out to confirm binding specificity.

Due to the biased nature of phage display library production, a 16S rRNA gene analysis was also carried out in order to assess metagenomic DNA diversity prior to library construction. Because phage display was used successfully by colleagues with the genomes of single organisms, it was believed that including metagenomic DNA in a phage display library would cast a wide net over the tongue dorsum allowing capture of many more binding proteins occurring in this environment from a wide range of bacteria.

## Acknowledgements

There are many that I should thank and acknowledge but none are more important to me than my family. Their faith in me to succeed has always been an inspiration and any achievements I have made is owed as much to their unwavering support as it is to any contribution of mine. **Mum, Dad, Craig, Grans and Deys**, for everything, **Thank You!**

Several people are worthy of thanks for helping me manoeuvre through the sleep-depriving waters of my PhD. **Professor John Ward** is, without doubt, the best supervisor I could have asked for. Honest and kind, intelligent without condescension, he is the gold standard by which all other supervisors should be judged. **Professor Brian Henderson** has always been full of (umpty-tumpty) ideas and encouragement for this project, which he believed in from the start. Thanks also to everyone who gave me money: BBSRC, Wellcome, SfAM, SGM, UCL.

**Emma Stanley** became a valuable ally and a true friend, and I am glad we took every opportunity to put the world to rights. **Kathrin Schulze-Schweifing** brought brightness into our lab, and my appreciation goes to her for sharing the pain that is phage display with me. **Steph Hunter** is the sensible voice of guidance, and that encouragement will not be forgotten.

Without the following people, I would be a more miserable, deluded and cynical version of myself: **Martin (Jackie!) Cowie**, my best friend and security blanket for 10 long years. **Jenny Ritchie** became Jenny Marr during the final stretch of my PhD, and is my oldest and dearest friend. **Mag Murphy**, my partner in crime in Edinburgh, is an angel in every sense of the word. **Charlotte Hamilton**: HWU class of 2004!!! All of my colleagues from UCL, some of whom made me laugh every day, sometimes at them but mostly with them, notably: **Will Bryant** (guilty of numerous Dad jokes), **Steven Branston**, **Lewis Dartnell** and **Michael Hanley**.

### Planet Earth

*Planet Earth, my home, my place  
A capricious anomaly in the sea of space  
In my veins I've felt the mystery  
Of corridors of time, books of history  
Life songs of ages throbbing in my blood  
I've danced the rhythm of the tide and flood  
Your misty clouds, your electric storm  
Were turbulent tempests in my own form  
I've licked the salt, the bitter, the sweet  
Of every encounter, of passion, of heat  
Your riotous colour, your fragrance, your taste  
Have thrilled my senses beyond all haste  
In your beauty I've known the how  
Of timeless bliss, this moment of now.*

MJJ, 1958 - 2009

***Ladies and Gentlemen, I am now locked up in a handcuff that has taken a British mechanic five years to make. I do not know whether I am going to get out of it or not, but I can assure you I am going to do my best.***

- Harry Houdini, London Hippodrome  
Saint Patrick's Day, 1904.

---

## **Chapter 1**

### **Introduction**

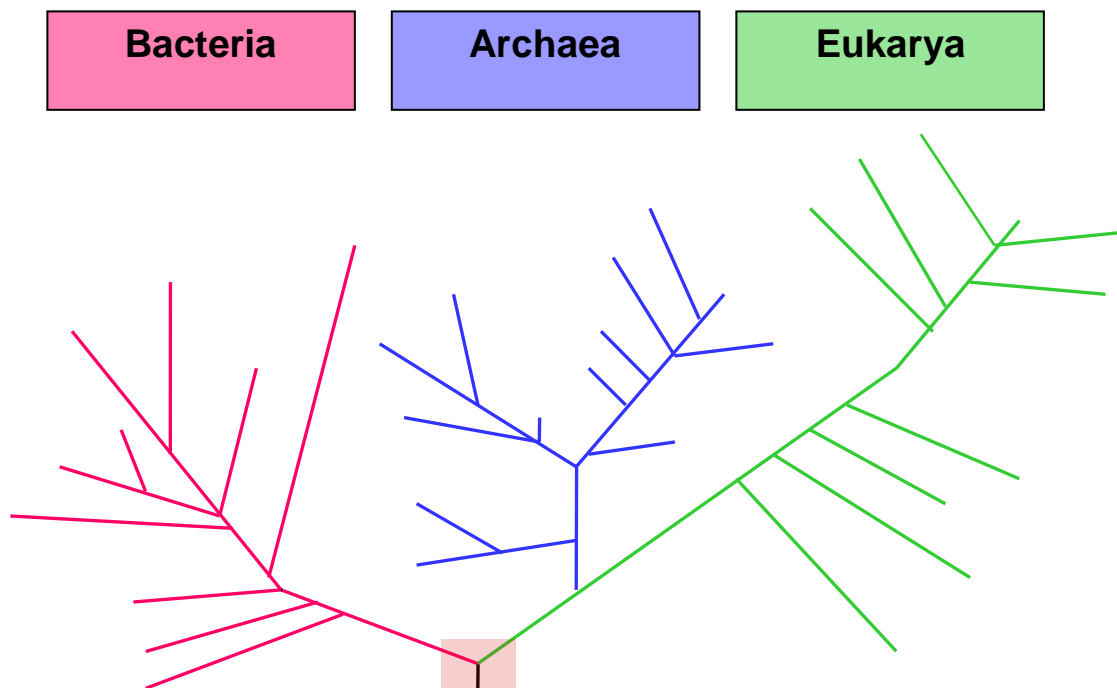
---

## General Introduction

The natural cycle of life and death, growth and multiplication by organisms great and small, is relentlessly shaped by microbial metabolic activity and physiological diversity, which has remained largely unexplored for many years. The first use of the microscope by Antonie van Leeuwenhoek almost 400 years ago prompted the realisation that a whole world of unseen organisms existed unexplored. The notion that more bacteria are present in a sample than the number appearing on culture plates is known today as the ‘great plate count anomaly’ (Handelsman, 2004) and alluded to a large microbial diversity long before molecular investigations became commonplace. However, in the 19<sup>th</sup> century, pure-culture techniques underpinned the interrogation of single bacteria and attempts to classify different organisms was based on specific nutritional requirements.

In 1884, Danish bacteriologist Christian Gram developed a method which distinguished the two major bacterial classes based on the physical and chemical properties of their cell wall. The Gram stain is now a common, and often the first, bacteriological technique used when rapid identification of bacterial infections is needed. Some bacterial cell walls are resistant to the decolorisation by acids used in the Gram stain, for example *Mycobacteria* spp., and these are called acid-fast bacteria, classified in 1938 (Gordon & Hagan, 1938).

However intriguing the microbiological world appeared to certain scientists, the view that large animals were of paramount importance was reinforced by evolutionary biologists when, in 1969, Whittaker developed a ‘Five Kingdoms’ evolutionary view of life, including animals, plants, fungi, protists and bacteria. This representation heavily insinuated what most people believed, that life on Earth was dominated by eukaryotes. A revolution in the study of microbial diversity came in the 1970’s, which turned bacterial evolution and the existing variety therein, upside down. Genetic information and taxonomy complement each other in the identification of living organisms, where taxonomic information is not robust enough as a standalone resource (Chu *et al*, 2006). Carl Woese’s pioneering 16S rRNA sequence and analysis study (1987) containing representative sequences from all known phylogenetic domains facilitated the construction of a sequence-based phylogenetic tree, illustrating relationships between previously unlinked organisms. Three domains exist in the current Tree of Life: Eukarya (eukaryotes), Bacteria and Archaea (**Figure 1**). This pictorial representation of evolutionary biology highlights that all previous research to record and describe biological species and relationships, had been directed at a tiny portion of microbial species (Pace, 1997). Now of course, ribosomal RNA (rRNA) genes are commonly used as ‘DNA barcodes’, and PCR amplification and analysis of this small subunit gene is commonly used to quantify species diversity.



**Figure 1 Phylogeny of the Living World.** This universal phylogenetic tree is derived from comparative sequencing of 16S rRNA. The evolutionary distance between two groups of organisms is proportional to the cumulative distance between the end of the branch and the node that joins the two groups. *Data obtained from the Ribosomal Database Project (Madigan et al, 2003).*

The interaction of bacterial communities associated with a human host can have direct implications for health (Jones & Marchesi, 2007). Public awareness of the beneficial properties of bacteria has increased in recent years, but so too has the implication of pathogenic bacteria in disease and the apparent rise of antibiotic resistant bacteria incessantly captures the public imagination. The acknowledgement of the potential consequence of large scale antibiotic resistance has triggered a surge in research interest and funding to find novel treatment options for bacterial disease. Some research interest has also focussed on bacterial proteomics – or reverse genomics (Brotz-Osterhelt *et al.*, 2005) - for the production of novel antibiotics for use in pharmaceuticals (Yoneyama & Katsumata, 2006). Primarily, public perception and understanding of micro-organisms is basic, but undeniably tarnished, governed by fears of infectious disease, antibiotic resistance and ‘superbugs’: views reinforced to a degree by tabloid ignorance and scare-mongering. Indeed, the rich diversity of bacteria is only partially understood, but now it is vital to gain functional information on bacterial proteins and how they interact with human factors.

Given the scale of human bacterial cargo this project aims to discover more about the methods by which bacteria, in particular the commensal microbiota of the human tongue dorsum, adhere to human surfaces. The human oral cavity supports a huge number of bacterial

passengers with very few ill effects, and it is of enormous interest to the scientific community to discover how this is possible.

### **Friendly bacteria**

Most micro-organisms not only co-exist in a co-operative symbiosis with mammals and other bacteria, but through their natural metabolic pathways, enable the persistence of eukaryotic life as we know it, by encoding essential metabolic functions that humans have not evolved for themselves (Ley *et al.*, 2008). Molecular studies of microbial populations at specific sites in the human body have highlighted the beneficial effect these communities can invoke. A major site of interest is the human gastro-intestinal tract (GI), and comparisons between germfree and conventional animal models have clearly shown the gut microbial community to have considerable influence on host biochemistry, physiology, immunology and low-level resistance to gut infections (Gordon & Pesti, 1971; Walter, 2008), possibly by counteracting the adhesion mechanisms used (Collado *et al.*, 2008).

The beneficial effects of indigenous microbiota are no longer disputed however clarification is still needed regarding the methods by which bacteria and host survive together without more persistent infections. For example, the presence of certain bacterial components is known to trigger particular medical conditions, such as Inflammatory Bowel Disease (IBD). IBD causes misery and discomfort for thousands of people, and a study by Lodes *et al.* in 2004 set out to identify the bacterial antigens responsible for the pathogenic nature of IBD; found to be flagellins. Determining the individual factors responsible for such immune responses may illuminate the reasons why bacteria are mostly tolerated on human cells and tissues.

### **Bacterial Colonization**

Microbes do not sit passively on host surfaces. Micro-organisms interact with mammalian hosts in a number of ways, most of which are still poorly understood. Three types of host-bacteria and inter-bacteria interaction are recognised in mammals. The first is mutualism – where both members of the association benefit from the others' presence. The second is commensalism – where one member benefits, but the other is unaffected. The third is parasitism – where one member benefits at the expense of the other (Wilson, 2005a). The predominant form of symbiosis between micro-organisms and their warm-blooded hosts appears to be mutualism, and in most literature these are referred to as the normal microbiota – or commensals.

Bacterial colonization of mucosal surfaces is thought to require the evasion of local host immune responses (Kilian, 2003), although it is conceivable that the host mediates its



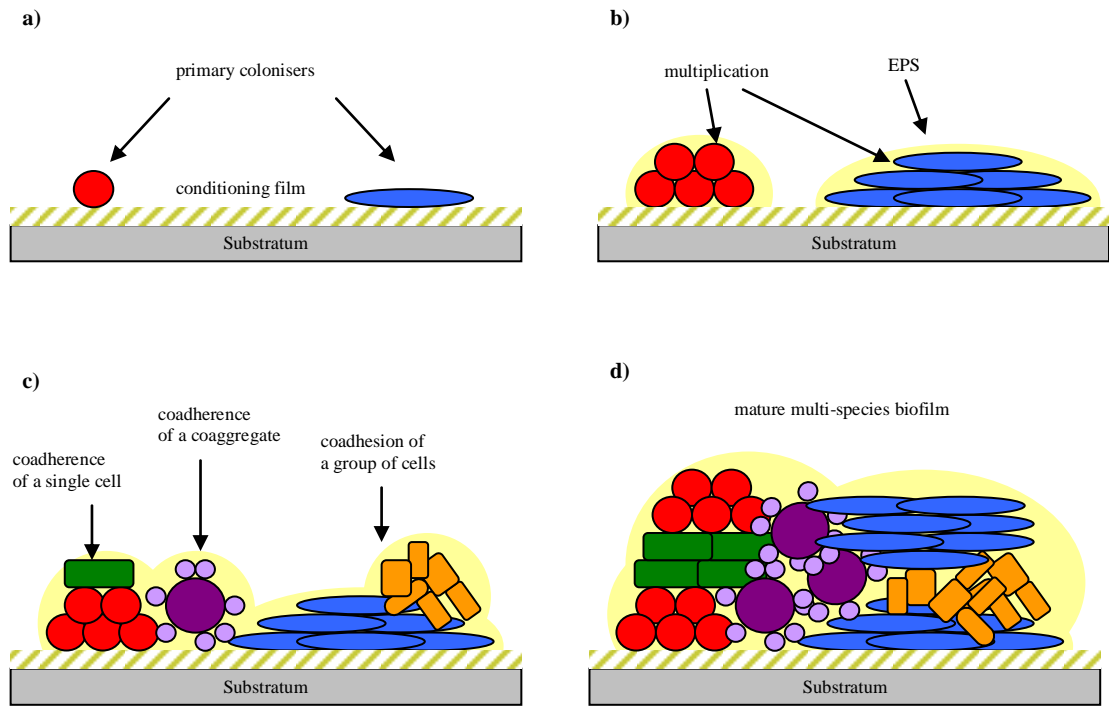
reaction if the bacteria are beneficial, and facilitates specific adherence mechanisms which initiate bacterial attachment to a human cell. Attachment is a priority for organisms who benefit from staying in a specific location. This adherence procedure is enabled by bacterial cell-surface binding proteins, many of which are known and characterised as adhesins; proteins which recognise and bind specific receptors on human cells (Azzazy & Highsmith, 2002). Commensal and pathogenic bacteria all express adhesive abilities, and suitable locations will always be efficiently colonised, ultimately reaching a healthy and dynamic equilibrium, or climax community. Several studies have implied that the commensal microbiota act as a protective barrier to invading pathogens; providing competition for space and nutrients and limiting pathogen growth and multiplication (Madigan *et al.*, 2003). The niche-specificity of bacteria, in association with their ability to breach mucosal barriers and invade a host, distinguishes pathogenic from commensal organisms (Wizemann *et al.*, 1999). Some bacterial diseases are associated with the commensal microbiota, but these are thought to result from a failure of normal immune regulatory mechanisms rather than aggressive behaviour on the part of the bacteria themselves (Wade, 2002).

### **Microbial communities**

The mixed microbial communities which populate oral surfaces can contain between 34 and 72 species per person (Aas *et al.*, 2005) and this assortment of species often leads to biofilm development (**Figure 2**). Growth of any biofilm is in stages, as each bacteria engages in signalling and/or metabolic interactions with those around it (Egland *et al.*, 2004), culminating in a dynamic 3D structure of cooperative microorganisms (Kara *et al.*, 2007). This process probably takes around 8 hours on the tooth surface, but could be less on the tongue dorsum since the biofilm would probably never be entirely removed.

Metabolic collaboration is widespread between species during the development of stable and resilient biofilm communities. Various species of *Streptococcus* are first to colonise oral surfaces, and the colonization of the tooth surface is particularly well understood (Kolenbrander *et al.*, 2002). Primary colonizers use surface bound receptors of adsorbed salivary proteins, meanwhile catabolising carbohydrates to produce shorter chain organic acids. *Veillonella* species routinely accompany streptococcal colonisation, probably due to their dependence on *Streptococci* to catabolise sugars so the Veillonellae can ferment the resulting organic acids (Palmer *et al.*, 2006). This kind of interaction provides the basis for such cooperative symbioses.

Inter-bacterial binding, or coaggregation (Rickard *et al.*, 2003), promotes bacterial accumulation on surfaces that have already been colonised, a trait common to *Veillonella* sp.. Bacteria which join a biofilm community in the later stages of development can act as bridging



**Figure 2 Biofilm development** a) primary colonisers of the tongue surface might include Streptococcal species, which attach through specific or non-specific interactions with the components of the organic conditioning film, b) multiplication of the primary colonisers leads to the formation of microcolonies, the increased output of which changes the environmental conditions of the early biofilm, c) secondary colonisers attach to the primary colonisers by the specific process of coadherence of either single cells, coaggregates (interspecies binding which has occurred separate from the biofilm environment) or groups of cells, d) coadhered cells become part of the larger multispecies biofilm community and the biofilm is said to be mature. Adapted from Rickard *et al.*, 2003.

species, and may be able to form interactions with many different kinds of bacteria. The biofilm grows in size and complexity until it is mature or parts of it are dispersed by natural or mechanical means.

On the tooth surface, *Fusobacterium nucleatum* is such a bridging organism, coaggregating with a wide range of oral bacteria, causing it to be known as a secondary coloniser and linking it with the development of periodontal disease. The work of Edwards, (2007) found that coaggregation of *Streptococcus cristatus* to *Fusobacterium nucleatum* not only promotes the survival of *S. cristatus* in saliva, but enables its carriage into host cells as a passenger when bound to *F. nucleatum* (Edwards *et al.*, 2006). Thus, it is clear that adhesion is being used as a survival mechanism in the oral environment.

Being part of a biofilm has many advantages for bacteria. Biofilms concentrate/trap food and metabolic by-products as substrates for microbial growth (Kierek-Pearson & Karatan, 2005). Biofilms are less susceptible to antimicrobial agents and are structurally and

functionally organised by countless antagonistic and synergistic microbial interactions. More subtle cell-signalling can lead to coordinated gene expression in the community. As a result, a multispecies biofilm can possess a combined metabolic activity and efficiency that is greater than a single species population. Other advantages include gene transfer, enhanced pathogenicity and an increased host range (Marsh, 2005).

### **Oral cavity, saliva and tongue dorsum**

In humans, the oral cavity is generally agreed to be home to around 750 transient bacterial species (Jenkinson & Lamont, 2005). Huge disparity in the commensal microbiota occurs between individuals, resulting from general heterogeneity within the human population such as diet and oral health, and in the various anatomic features of the oral cavity such as gaps between the teeth, salivary flow rate and depth of tongue crypts, all of which may influence the survival of certain microbes over others (Roldan *et al.*, 2003; Paster *et al.*, 2006). With a mean surface area of 215cm<sup>2</sup> (Collins & Dawes, 1987), teeth, keratinized and non-keratinized surfaces make up about 20%, 30% and 50% of the oral cavity respectively (Mager *et al.*, 2003). The presence of keratin on the filiform papillae of the tongue dorsum (Dresselhuis *et al.*, 2008) provides a tough impermeable barrier. Studies which focus on the tongue microbiota specifically (Kazor *et al.*, 2003; Riggio *et al.*, 2008) are rare in comparison to those on dental plaque and associated diseases, such as periodontitis and gingivitis (Vitorino *et al.*, 2006). The precise role of the tongue microbiota with regards to human health and disease is not clear – leaving a gap in the knowledge in this particular area.

Local physiochemical conditions in the human oral cavity are conducive to rapid bacterial proliferation. The ubiquity of saliva has many positive functions such as diluting carbohydrates which, when metabolised by bacteria, produce acids promoting the development of periodontal disease (Marcotte & Lavoie, 1998). Saliva also promotes the flow of nutrients around the oral cavity and the component lysozyme causes aggregation and removal of unattached microbes (Wilson, 2005c).

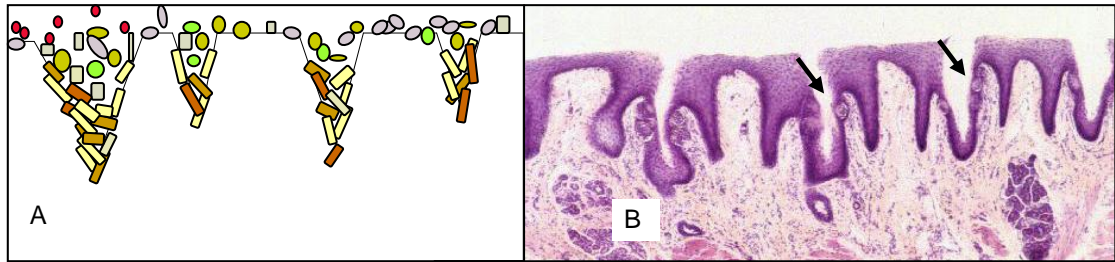
Additionally, salivary components act as ligands for bacterial adhesion (**Table 1**). These components, particularly Fibronectin and other extracellular membrane (ECM) constituents, have been major sources of bacterial binding proteins (MSCRAMMs) over the last few years (Weiss *et al.*, 2000; Christie *et al.*, 2002; Goo *et al.*, 2006; Mullen *et al.*, 2007; Edwards *et al.*, 2007). Salivary amylase binds the oral bacteria *S. gordonii*, *S. mitis* and *S. oralis* (Helmerhorst & Oppenheim, 2007), which may promote or prevent adherence within the oral cavity. Secreted IgA (S-IgA), the main immunoglobulin found in saliva, enhances the antimicrobial activity of other salivary components, such as lactoferrin (sequesters iron in the environment), agglutinin (mediates bacterial aggregation and adherence) and mucins

<i>Component</i>	<i>Range of concentration in saliva(<math>\mu\text{g/ml}</math>)</i>
Albumin	25
Fibronectin	0.2 – 2
IgG	15 - 30
Lactoferrin	0.4 – 7
Proline rich Proteins	0 – 180
Secreted IgA	5 - 58

**Table 1 Salivary Components. Adapted from Scannapieco, F.A., 1994.**

(protective role), all of which assist in managing the bacterial load present (Marcotte & Lavoie, 1998). The presence of S-IgA limits the adherence of bacteria to surfaces by producing antibodies against the offending bacterium, such as oral streptococci (Vudhichamnong *et al.*, 1982), *Candida albicans* (Williams & Gibbons, 1972) and the fimbriae of members of the *Enterobacteriaceae* (Tratmont *et al.*, 1980). It does this by disrupting the stereochemical and non-specific interactions required by bacteria to bind to a surface, resulting in a reduction in the number of possible adhesive interactions. Additionally, S-IgA agglutinates bacteria, promoting clearance from the oral cavity (Liljemark *et al.*, 1979). The complex structure of the hydrophobic tongue dorsum and the nature of salivary components mean that saliva-protein emulsions are readily adsorbed (Dresselhuis *et al.*, 2008). Secreted IgA bound to the bacterial surface can enhance clearance from the oral cavity or adherence to oral surfaces or other bacteria (Scannapieco *et al.*, 1989).

The tongue dorsum offers a far different environment for bacterial life than the hard, non-shedding surfaces of the teeth (Mager *et al.*, 2003). The papillary structure of the keratinized tongue dorsum (du Toit, 2003) has an extended surface area due to the presence of deep crypts and grooves between the filiform papillae (**Figure 3**). This irregular surface facilitates the accumulation of saliva, other bacteria and particles of food where unregulated bacterial growth leads to the development of anaerobic microhabitats (Roldan *et al.*, 2003; Wilson, 2005c). The individual filiform papillae are covered in a layer of interdigitating epithelial cells in various stages of exfoliation. The anterior surface of each papilla is covered in micro-organisms, which can be found between epithelial cells up to four layers down (Brady *et al.*, 1975), where bacteria may then come into contact with FN in the ECM. A build up of bacterial by-products commonly leads to halitosis (bad breath), and the tongue surface is the principal source of oral malodour, the target of many recent studies (Haraszthy *et al.*, 2007;



**Figure 3** Diagrams of the Human Tongue Dorsum. A & B both depict the highly papillated structure of the tongue dorsum, where the presence of deep furrows produces a highly convoluted and therefore increased surface area, perfect for bacterial proliferation. Diagram A is a schematic illustration of the proposed bacterial load on this surface, with anaerobic microbes proliferating at the bottom of the deep crypts. Diagram B is a histological slice of a murine tongue dorsum, stained with haematoxylin and eosin, which clearly shows the dark epithelium covering the papilla and the taste buds (arrows).

Riggio *et al.*, 2008). The general anaerobic tendency of microbes living on the tongue dorsum is due in part to the development of thick biofilms. Biofilm progression results in the development of a wide range of redox potentials as it matures and becomes denser. More bacteria become bound to those in the vicinity – a process known as coaggregation – resulting in the maturation of a dense microbial mat. During this process the biofilm becomes increasingly anaerobic as microbial activity utilises trapped oxygen and the increasing thickness of the biofilm reduces the ability of oxygen to permeate it (Marcotte & Lavoie, 1998).

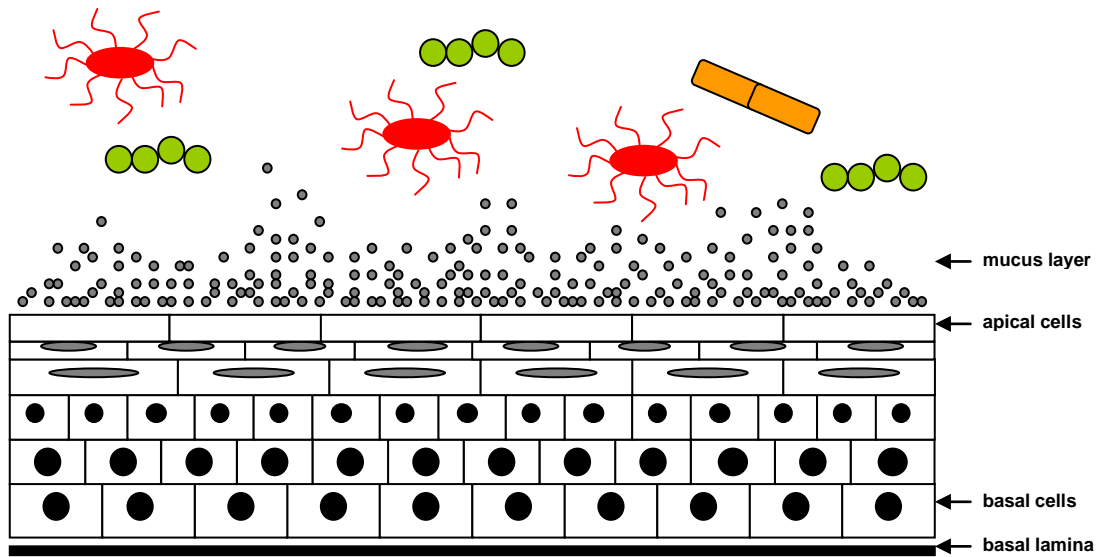
### Host cell biology

Clearly, bacterial adhesion is a complex process and in order to unravel the interactive strings mediating bacterial adhesion to host tissues a thorough knowledge of the cell and tissue biology of the target is essential (Ofek *et al.*, 2003b). The constituents of mammalian cell membranes are varied and complex (**Table 2**), however all membranes share universal features: (i) membrane lipids are organised in a planar bilayer configuration, consisting of glycerolphospholipids, sphingolipids and sterols in a fluid state (ii) the bilayer contains integral proteins including glycolipids, glycoproteins and proteoglycans (iii) a combination of weak interactions, hydrogen bonding and hydrophobic affinities bind other proteins and glycoproteins to the membrane surface, these are known as peripheral components, and (iv) the presence of a mucous blanket or cell coat comprising carbohydrate rich materials on the surface of epithelial cells provides protection to the cells and interaction opportunities with bacterial adhesins.

<i>Membrane Constituent</i>	<i>Characteristics</i>
<i>Integral Proteins</i>	
Integrins	Glycosylated, membrane spanning heterodimers, >20 identified, bind to basolateral surfaces of epithelial cells and the basal lamina. Also bind to ECM components and act as receptors for a number of pathogens
Proteoglycans	Heparan sulfate proteoglycan family contains 5 classes, syndecan and glypican are the main components. Some are transmembrane and some ECM components, but all bind a variety of bacteria, viruses and parasites including <i>Streptococcus</i> spp.
Cadherins	E, P and N-cadherin are most widely expressed and all contain a single transmembrane domain, a large extracellular domain and a cytoplasmic domain that interacts with the cytoskeleton.
Glycocalyx	Complex and highly glycosylated, it is comprised of integral glycoproteins, proteoglycans and glycolipids, as well as a coat of non-secreted mucins anchored in the cell membrane. The high density and diversity of saccharides provides multiple receptors for lectin-bearing bacteria.
<i>Peripheral Proteins</i>	
ECM – Collagens	Three types, I, II and III, exist of the most abundant proteins in animals. Consist of a helical arrangement of $\alpha$ chains, polymerized in a staggered manner.
ECM – Fibronectins	Ubiquitous adhesive glycoproteins composed of type I, II and III repeats which bind to various receptors such as collagen, proteoglycans and bacteria.
ECM – Laminins	Basement membrane adhesive glycoproteins which bind to collagen, heparin sulfate and integrins. Interacts with <i>E. coli</i> and viridans and group A streptococci.
ECM – Vitronectin	Part of the complement system, this glycoprotein binds to collagen and heparin and is a receptor for group A streptococci.
Cell Coat	Consisting of a layer of glycoprotein, it often includes mucins, which inhibit or promote bacterial adhesion to host mucous membranes. It is the first point of contact for bacteria colonising host tissues.

**Table 2 Human Cell Membrane Constituents**

The first point of contact for a bacterium on a mammalian tissue is most likely to be the epithelial (skin) cells or those of the gastrointestinal (GI) tract. Epithelial cells are keratinised, the fibrous nature of which provides a tough, impermeable barrier. Structural cells of this type exist on the tongue dorsum and the skin (**Figure 4**) because they provide protection against physical damage as well as bacterial colonization. This epithelium has three basic characteristics: (i) the cell layer has a free apical surface where the superficial cells are keratinised and coated with a thick mucous layer; (ii) contiguous cells are joined by junctional



**Figure 4 Stratified Epithelia.** Stratified epithelia commonly have up to 20 layers of cells. The basal lamina consists of collagen, laminin, proteoglycans, fibronectin and other ECM components. Cells are joined by junctional complexes and the apical cell layer is not only protected by keratin, but also by a thin mucous blanket.

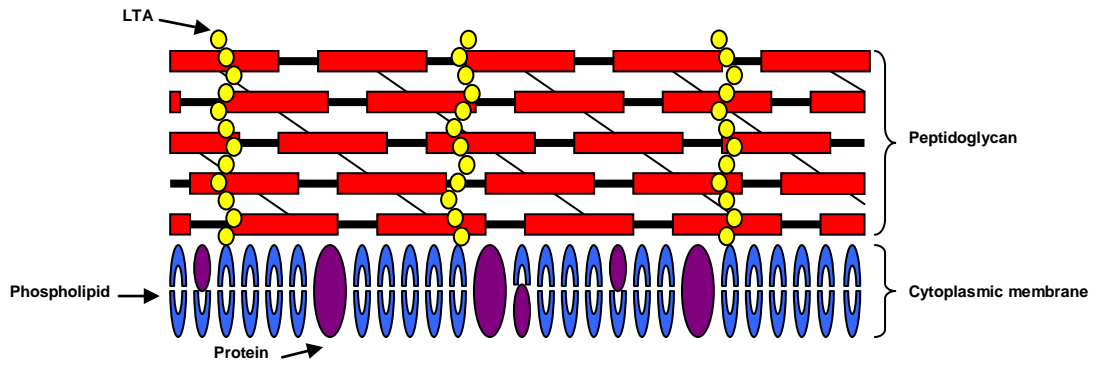
complexes and (iii) the cell layer is attached basally to an extracellular matrix specialisation termed the basal lamina. Epithelia such as the tongue dorsum are known as ‘stratified’, having up to 20 layers of cells. Only the basal layer is attached to the basal lamina. Below the basement membranes (basal side) is the lamina propria, containing the connective tissues, of which fibronectin and collagen interact to form a structurally sound extracellular protein network – the extracellular matrix (ECM) (Geiger *et al.*, 2001). This network contains a range of cell types, such as macrophages and fibroblasts, as well as neural and vascular constituents.

ECM macromolecules, which also include vitronectin and laminin, can underlie the epithelia at a considerable distance from the cell surface. However, they do come into contact with membrane receptors in many instances, making them a point of contact for adhering microbes (Ofek *et al.*, 2003b). Bacteria commonly encode and use fibronectin binding proteins (Fnbp’s) to facilitate binding to human cells (Jonsson *et al.*, 1991; Joh *et al.*, 1998).

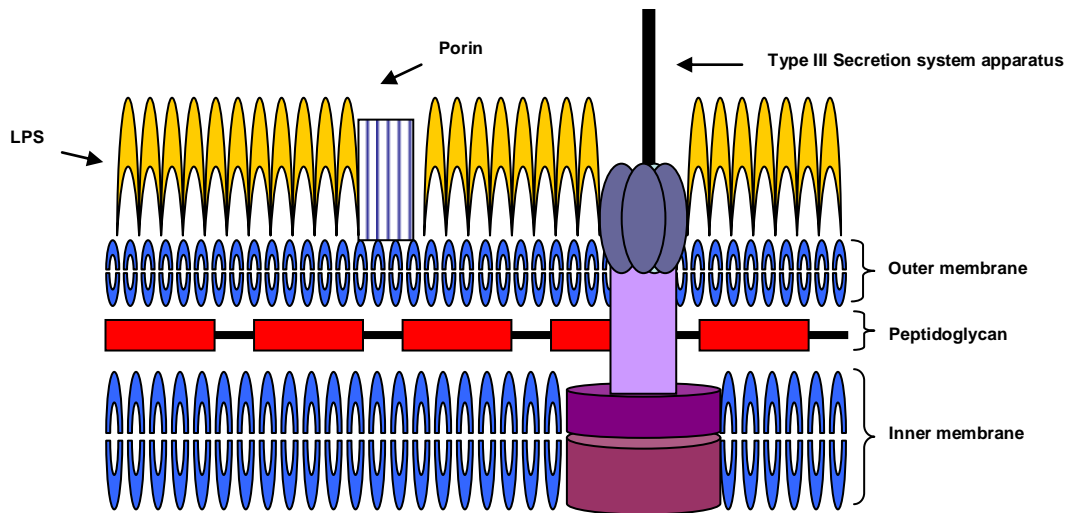
### Bacterial architecture

When extracting mixed microbial DNA directly from a sample, the cell wall heterogeneity of the microbes contained within must be considered. A common feature of both Gram positive (**Figure 5**) and Gram negative (**Figure 6**) bacteria is the peptidoglycan-containing network which gives shape to the bacterial cell and absorbs trauma. In Gram

positive bacteria the peptidoglycan layer is 20-40 nm thick in order to withstand the high (20-25 atm) osmotic pressure inside the Gram positive cell (Koch *et al.*, 1982), in comparison to the 5-10 nm layer in Gram negatives. Although peptidoglycan is heterogeneous, it does not appear to function as a ligand for bacteria-bacteria adhesion, instead providing a support for adhesin presentation and stabilization (Nikaido *et al.*, 1996).



**Figure 5** Gram positive cell wall containing teichoic acid and a thick (20-40 nm) peptidoglycan membrane needed to withstand the osmotic pressure within the cell. Cell wall anchored proteins, transmembrane proteins, surface proteins and glycolipids are all involved in binding. (LTA= Lipoteichoic Acid).



**Figure 6** Gram negative cell wall is made up of LPS and protein and contains thin layers of peptidoglycan. The outer layer of the cell wall contains porins for specific types of molecule. Gram negative bacteria bind using fimbriae, outer membrane proteins (OMP), glycolipids (LPS) and other unknown surface structures.



Gram positives also contain teichoic acid, where Gram negatives have an outer membrane (OM) covering the thin peptidoglycan layer which also contains lipopolysaccharide (LPS), fimbria, porins and type III secretion systems (Ofek *et al.*, 2003c). The means by which adhesins are anchored to the bacterial cell surface are different for Gram-positive and Gram-negative bacteria. In Gram-positives cell-wall anchored proteins, transmembrane proteins, surface proteins (lipoteichoic acid) and glycolipids (internalin) are all involved in initiating or maintaining the binding process. In Gram-negative bacteria this role is carried out by fimbriae, outer membrane proteins (OMP), glycolipids (LPS) and other unknown surface structures (Ofek *et al.*, 2003c). It is important to understand architectural differences between Gram positive and Gram negative bacteria during metagenomic analyses as DNA extraction techniques can result in unequal or non-lysis of certain species, lending a bias to subsequent libraries produced from the material (Hattori & Taylor, 2009).

### **Adhesins**

The expression of bacterial binding proteins requires complex processes moderated by coordinated mechanisms, which differ for each adhesin and between bacterial species (Ofek *et al.*, 2003c). One example is the binding proteins of *Staphylococcus aureus*, one of the many virulence factors this organism encodes. Some are expressed throughout the growth cycle, like Clumping Factor A (ClfA), whereas most are only expressed during the early exponential and mid exponential growth phases, when cell density is low (Lowy, 1998). Fibronectin binding proteins (FnBP's) are also examples and include FnBPA (Signas *et al.*, 1989), FnBPB (Jonsson *et al.*, 1991) and protein A (Heinrichs *et al.*, 1996). As growth progresses and bacterial cell density increases, the expression of adhesins is downregulated.

Of the various types of bacterial binding proteins, all must be presented or anchored to the bacterial cell surface in some way before colonisation can occur. Some binding structures and components are common between Gram positive and Gram negative bacteria, but again, the stage of growth and even the growth media used can have a drastic effect on proteins expressed at any time (Ofek *et al.*, 2003c).

### **Bacteria-host interactions**

In order for a bacterium to attach itself anywhere, an energy barrier must be overcome – regardless of the mechanism of adhesion (Ofek *et al.*, 2003a). Both bacteria and adhesive substrata on human cells are negatively charged and these repulsive forces are overcome by hydrophobic determinants, leading to weak, reversible adhesion. These non-specific interactions are mediated by surface proteins such as LTA (G+ve) and type IV pili (G-ve)

(Ofek *et al.*, 2003a). Because repulsive forces increase in proportion to the diameter of the approaching particles (Ofek *et al.*, 2003a), fimbriae or other small bacterial components are an effective way of overcoming the barrier. Maximum repulsion between the adhering substrata occurs at 50 nm, and at 2 nm complementary binding relations such as lectins, hydrophobic or electrostatic interactions come into play. The stabilizing effect of hydrophobic sites solidifies the connection by maintaining stereospecific interactions (Ofek *et al.*, 2003c).

The second stage of bacterial adhesion involves stereospecific interactions with complementary receptors on host surfaces through ionic interactions, hydrogen bonding and/or hydrophobic effects. For example, in Group A Streptococci, hydrophobic lipoteichoic acid (LTA) mediates initial binding to fibronectin, but the second step is mediated by fibronectin binding protein (FBP), which brings the bacterium through the repulsion barrier set up by the respective negative charges, thus increasing binding strength (Ofek *et al.*, 2003a).

Any binding event involves a complex cascade of molecular interactions between the ligand and receptor molecules. A successful binding event between bacteria and host can initiate complex signal transduction cascades in the host which may (i) activate innate host defences, (ii) initiate changes to host cellular processes which facilitate bacterial colonization, or (iii) exaggerate microbial pathogenicity by activation of gene expression within the bacterium (Soto & Hultgren, 1999). Although binding events are ubiquitous, the process is still not explicitly understood. What is known however is that adherence to host structures such as the ECM by microbes is a key factor in determining virulence and can lead to the progression of infectious disease (Fine *et al.*, 2005).

### **Types of binding interaction**

Known proteins involved in adhesion are highly conserved (Abraham *et al.*, 1988), and the lack of variation is thought to be due to analogous host receptors (Wizemann *et al.*, 1999). Due to the ubiquity of bacterial binding, it is likely that uncharacterised binding proteins are commonly used for colonisation, and taking a metagenomic approach provides an opportunity to identify some of them.

Attachment of a bacterial cell to a ligand is mediated by three main categories of adhesive interaction, of which bacterial lectins are the most common type among Gram negative and Gram positive bacteria (Wizemann *et al.*, 1999). Lectins can be bound to the bacterium or to the mucosa being colonized. In Gram negative bacteria, lectins incorporate fimbriae, pili and other outer membrane (OM) components; and in Gram positives lectins are found in the peptidoglycan matrix. Lectins are classified by sugar specificity and the carbohydrate structure of an individual lectin is incredibly difficult to study as the sugar specificity includes both the primary sugar specificity and fine sugar specificity. An example of

lectin binding is given by Ruhl *et al.*, 1996, where interactions between bacterial lectins and various carbohydrate side chains of IgA were detected in species such as *E. coli*, *S. gordonii* and *A. naeslundii*. One of the best understood mechanisms of bacterial adherence is that mediated by cell surface pili or fimbriae, and much is known about them (Wizemann *et al.*, 1999; Skerker & Berg, 2001).

Human cell membranes have specific molecules associated with them (**Table 3**), which can function as ligands for commensal bacterial colonization (Kazor *et al.*, 2003). These are usually proteins, glycoproteins and polysaccharides (Kolenbrander, 1993), and these protein-protein interactions include binding of Extracellular Matrix (ECM) components such as fibronectin, collagen, vitronectin (Patti *et al.*, 1994) and laminin (Fine *et al.*, 2005) to the Fibronectin Binding Proteins (FnBP's) of bacteria. For example, the adhesins FimA and SsaB have affinity for salivary glycoprotein on the tooth surface and use this attraction to colonize the oral cavity (Schennings *et al.*, 1993).

<i>Type of membrane component</i>	<i>Cell type</i>	<i>Adhesin</i>	<i>Bacterium</i>
Glycolipids	Uroepithelial cells	P fimbriae	<i>E. coli</i>
Glycoproteins:			
- Integrins	M cells, enterocytes	IpaB, IpaC	<i>Shigella</i>
- Heparan sulphate, proteoglycans	Epithelial cell	Opa proteins	<i>N. gonorrhoeae</i>
- E-cadherin	M cells, enterocytes	Internalin A	<i>L. monocytogenes</i>
Glycolipoprotein	Mast cells	Type 1 fimbriae	<i>E. coli</i>

**Table 3 Human membrane components and associated adhesins.**

Hydrophobins are the third and least well characterised category of known adhesive interactions, but are the most common way for bacteria to overcome initial repulsive forces between components. Hydrophobins are not really bonds, but a tendency for apolar molecules to associate with other apolar molecules rather than water. Hydrophobins include any surface component which promotes hydrophobicity and adhesion to surfaces, some such as streptococcal or staphylococcal proteins are covalently bound to the cell wall (Csh A) where others are part of the outer membrane like lipids or fimbriae (Doyle *et al.*, 2000).

Clearly, bacteria have access to plethora of routes for gaining a foothold in the human body, but microbes do not sit passively on human surfaces ignored by the human immune

system, but are recognised by the innate immune system and can induce localized adaptive responses without eliciting destructive inflammation (Lu *et al.*, 2006).

### **Innate microbial recognition**

It is currently accepted that part of the interaction between host and microbial cells is mediated through the intercellular signalling proteins cytokines which mediate bacterial homeostasis. These molecules interact locally with bacteria or their released products and become amplified, signalling microbial presence (Henderson *et al.* 1996). However, commensal and pathogenic bacteria can produce molecules which induce both pro- and anti-inflammatory cytokines thus allowing microbes to exert more control over the host response than previously thought (Henderson *et al.*, 1996). Clearly, the innate immune system recognises the presence of bacteria but what remains unclear is to what extent this is occurring and how – or if – the immune system recognises and copes with commensal and pathogenic organisms differently. Similarly, do bacteria recognise when a host immune system is depressed and use this chance to invade?

It appears that pathogens are treated differently by the host. Both commensal and pathogenic organisms express Micro-organism Associated Molecular Patterns (MAMP's) which are recognised by host pattern recognition receptors (PRR's), but the proinflammatory response is limited only to pathogens, so avoiding excessive and detrimental inflammatory responses (Sirard *et al.*, 2006). Commensals avoid initiating an immune response by either down-regulating the host response themselves or by altering MAMP expression or structure to avoid detection.

Bacteria are also equipped to take advantage of opportunities provided by the host with many microbes expressing binding proteins which mediate adhesion to specific areas, like the ECM. Microbial Surface Components Recognising Adhesive Matrix Molecules (MSCRAMMS) are molecules on the microbial cell surface which recognise, with high affinity and specificity, an ECM ligand such as collagen, laminin, FN and fibrinogen (Patti *et al.*, 1994). The host can mediate bacterial attack by upregulating genes with defense functions such as *SPLUNC1* from humans and rodents, which is closely related to LPS binding protein (LBP), known to enhance proinflammatory signals in response to bacterial LPS. More interestingly this group (LeClair *et al.*, 2004) located a related but novel protein *splunc5*, unique to rodents, which is expressed solely on the tongue epithelium, suggesting that its placement has been adapted to meet specific needs of the innate immune system (LeClair *et al.*, 2004).

There is clearly enormous diversity in the mechanisms used by bacteria to bind and communicate with host cells and tissues. By choosing two panning ligands expressed

commonly in the oral cavity it was anticipated that, over the course of this project, some of these complex interactions might be identified and deciphered in more detail.

### **Bacterial Research**

Microbiology research now depends less heavily on the maintenance of individual bacteria in pure culture, as it is now realised that very few micro-organisms exist in pure culture in nature, instead thriving as broad, multifaceted communities in select niches (Lu *et al.*, 2006).

In humans, bacterial cells outnumber human cells by a factor of 10 (Wilson, 2005b) – and encode at least 100 times as many genes as the host genome (Jones & Marchesi, 2007). Prokaryotes clearly demonstrate such metabolic and physiological diversity that they can successfully populate any environment (Handelsman, 2004; Steele & Streit, 2005). The human body contains a huge diversity of unique colonization sites however the complex nature of host-bacteria signalling, host immune molecules and the range of local conditions, means that these constant interactions are still poorly understood.

The development of polymicrobial communities in humans has stimulated intense debate regarding their function, and the reaction of the host immune system in response to multiple bacteria, all pumping out extracellular products. This area of research is receiving a great deal of attention however the complexity of these interspecies interactions is making clear answers a challenge to find. Understanding the mechanisms by which a host is able to survey and discriminate between commensal and pathogenic bacteria, is crucial to a more complete appreciation of the host-microbe relationship (Lu *et al.*, 2006), and therefore new insight into pathogenic mechanisms.

Calling bacteria culturable or unculturable refers to the ability of bacteria to be cultured in the laboratory. Unculturable bacteria are those which are known to be present, for example through direct counts under a microscope, but which do not grow on conventional culture media under standard conditions (Wade, 2002). Current arguments on this subject are extensive, mainly falling into two categories, those who see ‘unculturable’ as inappropriate terminology (Rappe & Giovannoni, 2003; Stevenson *et al.*, 2004; Cowan *et al.*, 2005) debating instead that, with more detailed information on bacterial growth/substrate requirements, suitable culture conditions will be found (Hugenholz & Pace, 1996). Others note that media for bacterial growth can now be supplemented in virtually endless combinations, and that culture-based research will continue to result in only *a* cultivable fraction of a community being expressed, rather than *the* cultivable fraction (Ritz, 2007).

## Chapter 1: Introduction

A significant proportion of known organisms are associated with human health and disease, and it is prudent to suppose that a proportion of non-culturables will also be attributable to disease involvement (Aas *et al.*, 2005). Therefore the most up to date skills and knowledge should be used to assess the prospective variety and resources contained within.

Years of focus on pure culture enrichment of micro-organisms on selective media has resulted in the generation of a vast quantity of functional information concerning certain organisms. It is popularly quoted that less than 1% of all bacteria are cultivable using known techniques (Torsvik *et al.*, 1990a; Amann *et al.*, 1995; Pace, 1997), but this number is probably an underestimate (Prosser & Embley, 2002). Actually, the number of cultivable bacteria differs depending on the environmental niche. Soils are said to contain around 0.1% cultivable bacteria, seas are less cultivable with 0.01%, probably because much of it is inaccessible. More familiar sites, such as the oral cavity are quoted as between 50% (Paster *et al.*, 2001; Jenkinson & Lamont, 2005) and 60% (Kolenbrander *et al.*, 2002) culturable.

There are several explanations for the general reluctance of bacteria to be maintained in pure culture. Commonly, microbes depend on the provision of additional factors for growth and replication, such as the by-products of bacterial fermentation. The absence of a mixture of bacteria may be inhibitory to others (Tringe *et al.*, 2005). Culture media, which often contains artificial substrates, is regularly used to cultivate micro-organisms from a sample in the laboratory. This media may be toxic and, in its homogeneity, may be lacking in some unknown factor(s) required for growth. Furthermore, the production of inhibitory substances by more dominant bacteria in a mixed culture may also affect the growth of others (Wade, 2006).

Cells in the unculturable majority may be either **i.** species which are already known and characterised, but for whom standard culture conditions are not suitable or **ii.** unknown species which require further development of a suitable culture method/media (Handelsman *et al.*, 1998). An example of **i.** are ammonia oxidising bacteria from terrestrial and aquatic nitrifiers, challenging to purify due to incredibly slow growth and a low yield when cultured in the laboratory (Prosser & Embley, 2002). However difficult it may be, reaching pure culture status is still a desirable and fundamental objective in order to determine physical characteristics and to confirm the existence of an individual organism and its viability within an environment (Torsvik *et al.*, 1990a). Pure culture techniques complement molecular approaches as both disciplines contribute crucial pieces of the microbial genomic jigsaw. Culture-independent methods have become more widely used and have contributed greatly to the information available in public sequence data banks. Assessing the DNA contained within an entire microbial community is now possible by undertaking a metagenomic analysis. By extracting the entire complement of microbial genes in a mixed community, an opportunity is presented to interrogate all microbial genomes present without suffering the potential losses of a cultivation-based approach (Nichols, 2007).

Many metagenomic-based analyses have been carried out over the last 10 years, in a variety of environments including sea water (Wexler *et al.*, 2005; Venter *et al.*, 2004; Rusch *et al.*, 2007; Yooseph *et al.*, 2007; Kennedy *et al.*, 2007; D'Costa *et al.*, 2007; Wijffels, 2008), marine sponges (Daniel, 2004), microalgae (Ellis *et al.*, 2003) and drinking water (Schmeisser *et al.*, 2003). A great deal of attention has been paid to soil metagenomes since the likelihood of identifying uncultured and previously unknown microbes is high, and the environment is a rich source of new antibiotics and biotechnologically important enzymes (Torsvik *et al.*, 1990b; Gillespie *et al.*, 2002; Joseph *et al.*, 2003; Liles *et al.*, 2003; Lee *et al.*, 2004; Ginolhac *et al.*, 2004; Voget *et al.*, 2003; Gabor *et al.*, 2004; Fierer *et al.*, 2007; Hong *et al.*, 2007).

Closer to home, in particular the oral cavity (Paster *et al.*, 2001; Munson *et al.*, 2002; Mager *et al.*, 2003; az-Torres *et al.*, 2003; Roldan *et al.*, 2003; Aas *et al.*, 2005; Jenkinson & Lamont, 2005; Paster *et al.*, 2006; Spencer *et al.*, 2007, Riggio *et al.*, 2008) and the gut (Backhed *et al.*, 2004; Gill *et al.*, 2006; Turnbaugh *et al.*, 2006; Kurokawa *et al.*, 2007) have been extensively studied, in the hope of finding a cure or cause for some bacterial diseases such as Crohn's disease (Peterson *et al.*, 2008) and periodontal disease (Kaplan *et al.*, 2009). More unusual mixed bacterial communities have also been studied to gain insights into how they operate, such as an acid mine biofilm (Tyson *et al.*, 2004), termite guts (Warnecke *et al.*, 2007), bovine rumen (Ferrer *et al.*, 2005), human faeces (Brietbart *et al.*, 2003), honey bee colonies (Cox-Foster *et al.*, 2007) and the Uranian deep sea hypersaline anoxic basin (Ferrer *et al.*, 2005).

### **Metagenomics for Community Analysis**

The overwhelming complexity with which microbial systems operate in natural ecosystems, and the limitations posed by traditional culture-dependent analysis of micro-organisms, has forced the rapid development of new technologies which circumvent the need for bacterial culture. Metagenomics is one such approach and entails the functional and DNA analysis of all genomes present in an environmental sample (Handelsman *et al.*, 1998). In theory, a metagenomic library is representative of all genes present in a mixed microbial sample, however this depends of course on the robustness of DNA extraction and cloning methodologies used (Steele & Streit, 2005).

Ribosomal RNA (rRNA) is a structural molecule common to all micro-organisms, so there is no codon usage divergence. It is extremely valuable for rapid investigation of the microbial constituents in an environmental sample using the Polymerase Chain Reaction (PCR) (Pace, 1997). The study of microbial diversity based on 16S rRNA studies – termed phylogenetics – is based on analysis of the highly conserved 16S rRNA gene, which accounts for only 0.05% of the microbial genome (Steele & Streit, 2005). The synthesis of 'universal'

PCR primers amplifies rRNA genes from the DNA of all organisms present in a sample and, when followed by cloning and sequencing, can generate a huge quantity of environmental data regarding sample diversity (Streit & Schmitz, 2004). A phylotype is defined, in 16S rRNA terms, as clones which have > 98.5% identity. Those with < 98.5% similarity to previously defined clones are considered to represent a new species (Paster *et al.*, 2006). Care must be taken however since high levels of 16S rRNA gene sequence similarity between strains belonging to the same genus do not automatically indicate membership of the same species (Yassin *et al.*, 1996). Clearly, prokaryotic physiological and metabolic diversity cannot be assessed in its entirety using 16S rRNA gene analysis alone (Steele & Streit, 2005), and it is recommended for use in conjunction with other investigations, for example shotgun (randomly fragmented) libraries involving cloning of sample DNA into a suitable vector, and screening of clones for conserved genes by hybridization or multiplex PCR. Expression of specific traits can also be sought, such as enzyme activity or antibiotic production (Kang *et al.*, 2009) or the shotgun library can be randomly sequenced.

The main problem with interrogating entire microbial communities is the volume of sequence data which must be generated in order to gain a representative view of the sample (Steele & Streit, 2005). As an example of this, the first major attempt at a ‘whole community’ sequencing effort was initiated and funded by Craig Venter (Venter *et al.*, 2004). The group obtained samples of seawater from the Sargasso Sea in the Bermuda Triangle, known for its low nutrient levels, and sequenced as much DNA as possible. They sequenced 1.045 billion base pairs and found 1800 genomic species, 148 belonging to new bacterial phylotypes. Overall, the group recovered 1.2 million new genes. One of the interesting findings of this research was the detection of 782 new rhodopsin-like photoreceptors and the discovery that they can be attributed to a wide range of bacteria including the CFB group (*Cytophaga – Flavobacterium – Bacteroides group*). The phylogenetic identification of all species present in the samples is far from complete (Handelsman, 2004) and the vast quantity of data generated in this project will take many years to fully assess, but will provide a valuable insight into prokaryotic species and functional diversity in the Sargasso Sea.

An exciting potential of metagenomics is that it allows community-wide assessment of metabolic and biogeochemical function, but large scale sequencing efforts such as the Sargasso Sea – and the generation of sequence data *en masse* – entail years of annotation and analysis before the input becomes applicable for functional comparisons. A good example of a ‘complete’ habitat analysis is Tysons’ work on the biofilm community of the Richmond Mine (Tyson *et al.*, 2004). The extreme environmental conditions present in the mine means that the community structure here is basic, consisting of 3 bacterial species and one archaeal species. The bacteria form a pink floating biofilm on the surface of the mine water, which has a pH of between 0 and 1 and a constant temperature of 42°C. High levels of Fe, Zn, Cu and As in the



mine water, results in domination of this small community by extremely specialised species able to cope with these harsh conditions including *Leptospirillum*, *Sulfobacillus* and *Acidomicrobium*, whose genetic analysis reveal a multitude of genes centred on maintaining a non-toxic intracellular environment (Handelsman, 2004).

Tyson's basic community lent itself to complete DNA cloning and sequence analysis, leading to closure of the bacterial genomes dominating that niche. As a result, Tyson's work matched the bacterial phylotypes to functional roles within the biofilm community structure, facilitating deduction of interactive trends utilised to optimise survival. This comparison highlights that less complex communities are conducive to genome closure and therefore gene function assignment, as their simple nature presents fewer assembly puzzles (Nichols, 2007).

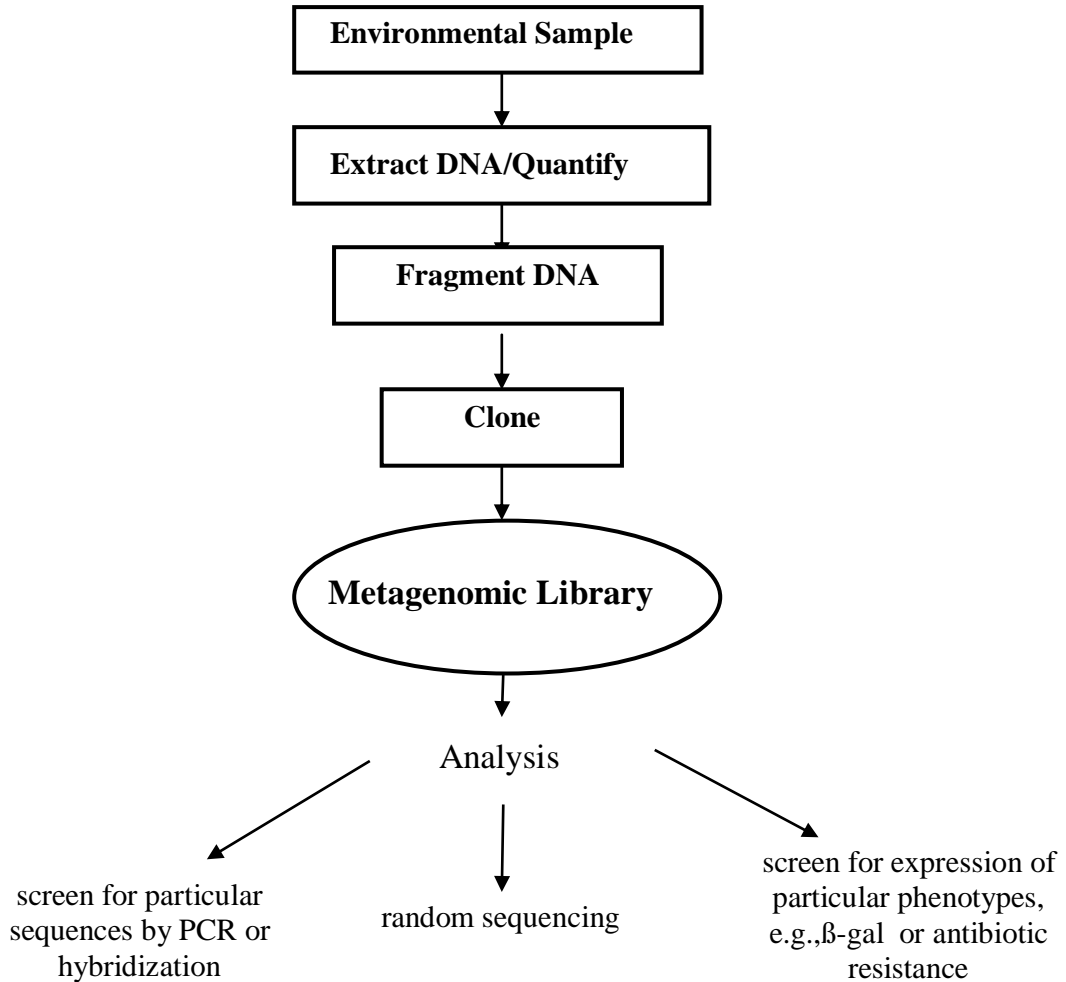
A major benefit of enormous metagenomic studies like Venters' is the rapid development of new software which simplifies cataloguing, storage and processing of the extraordinary volume of data generated. Perhaps more importantly, these advances enable data evaluation and assessment of complex communities at the system level (Schmeisser *et al.*, 2007). As sequencing capacity has increased however, the capacity to annotate this information has not. Between 2004 and January 2007, over 2 billion base pairs were deposited in databases of major metagenomic projects, eclipsing the entire 764 Mb of previously sequenced genomes, but until the ORF's are annotated with biological functions to provide context the information is not particularly useful (Harrington *et al.*, 2007).

### **Constructing a Metagenomic Library**

Cultivating bacteria gives microbial ecologists a context in which to investigate theoretical molecular findings, and provides more direct access to environmental genomes (Nichols, 2007). However meticulously formulated, culture conditions can never completely replicate the dynamic heterogeneity of a natural environment and so using cultivability studies and metagenomics as complementary techniques will allow contextualisation of phenotypic, taxonomic and genomic information from a particular habitat (Ritz, 2007).

There are several stages involved in metagenomic library construction (**Figure 7**). Sample collection, including adequate treatment and storage is followed by whole community DNA extraction. Extraction approaches differ depending on the analysis method to be used at the next stage. For example, if constructing a large fragment library, care must be taken to minimise shear damage during extraction. Following extraction, DNA fragmentation is normally by sonication or restriction digest to the appropriate size for further analysis. At this stage, it is common to clone fragments into a variety of vectors including pUC18/19, and conduct initial analysis by shotgun cloning and sequencing. Vectors designed for large fragment propagation can also be used such as BAC (Bacterial Artificial Chromosome) and

fosmids. Plating onto selective media allows selection of DNA fragments containing certain traits, e.g. protease activity, antibiotic resistance and lipase activity (Rhee *et al.*, 2005). As depicted in **Figure 7**, these are all options for locating interesting genes which display a searchable phenotype.



**Figure 7 Construction of a Metagenomic Library.** Metagenomic libraries allow the isolation of all genetic information present in an environmental sample. This can be analysed using a range of techniques, but completely circumvents the need for bacterial culture, resulting in the potential capture of unknown bacteria.

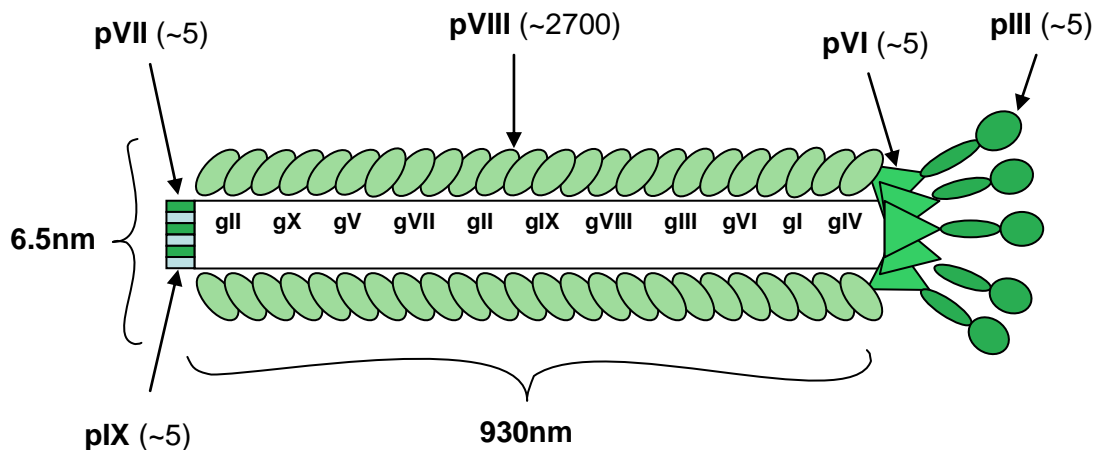
In specific human environment projects like this one, it is important to understand how novel proteins could interact with the host. In order to study bacteria-host interactions at the molecular (protein-ligand) level, the molecular technique Phage Display will be employed, along with the affinity selection stage Biopanning.

## Phage Display

Rather than undertaking a major sequencing and assembly project, the aim of this study is to use functional genomic tools to identify a few protein candidates for sequencing and characterisation, which demonstrate binding to human ligands. Exploiting naturally occurring protein-ligand interactions in an efficient screening process can result in the identification of those with highest affinity and specificity (Jacobsson *et al.*, 1997). One of the first genetic tools leading to the full realisation of the value of protein-ligand interactions was pioneered by George Smith in 1985 and is called Phage Display (Smith, 1985).

Phage display is used for many purposes. As a natural selection procedure it is useful for generating targets for drug discovery (Benhar, 2001; Trepel *et al.*, 2002; Gnanasekar *et al.*, 2004), epitope mapping (Matthews *et al.*, 2002) and for screening antibodies (Prinz *et al.*, 2004). Antibodies were one of the first proteins to be displayed on a phage surface (McCafferty *et al.*, 1990), and the isolation of monoclonal antibodies has been one of the most successful applications of phage display to date (Hoogenboom *et al.*, 1998).

This project utilises filamentous bacteriophage (**Figure 8**) for the display of bacterial proteins, but all phage carry intrinsic commercial value in their own right. Commonly quoted as the most abundant biological entity on the planet, bacteriophages are estimated to total  $10^{31}$  virus particles (Brussow & Hendrix, 2002). Due to the bacteria-killing activity of some phage, and to the diminishing power of antibiotics to treat disease, phage therapy is becoming big business (Alisky *et al.*, 1998; Miedzybrodzki *et al.*, 2007; Capparelli *et al.*, 2007) and several important studies have been carried out on the possibility of developing phage as an alternative to antibiotics (Weber-Dabrowska *et al.*, 2000; Wagenaar *et al.*, 2005).



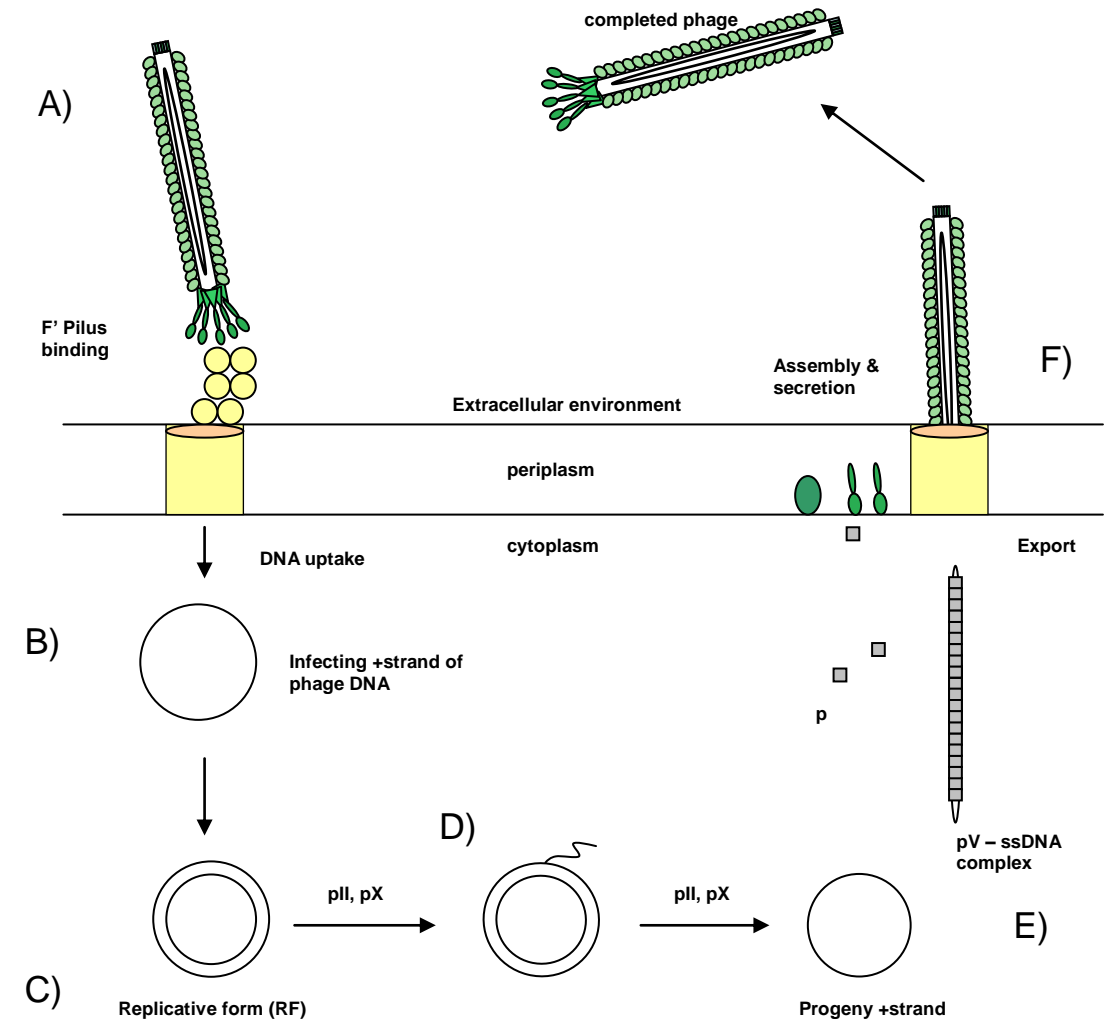
**Figure 8** Structure of filamentous phage showing all coat proteins, their location on the phage and all genes (adapted from Mullen *et al.*, 2006). Figures in brackets signify the number of copies of each structural protein on the phage surface.

Bacteriophage are used for phage display due to their natural ability to infect bacterial cells, and because they can incorporate foreign DNA into their circular genome and transport them into a bacterial cell during infection (Smith & Petrenko, 1997). Filamentous phage display allows assembly in, and secretion from, an infected bacterium without compromising the host cell membrane (Mullen *et al.*, 2006). *Escherichia coli* cells infected with such bacteriophage become a factory for phage production, as the host machinery is commandeered to generate phage virions.

### Phage life cycle

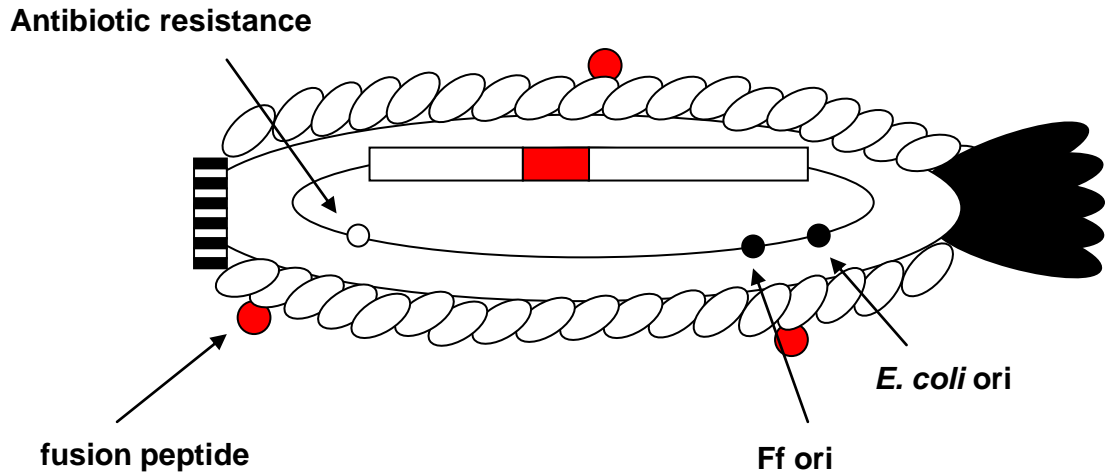
The life cycle of M13 filamentous phage has been described in particular detail (**Figure 9**). Coat protein pIII initiates the infection process by binding to the F pilus of *E. coli*. Pilus retraction brings the phage into intimate contact with the bacterial membrane where pIII binds to TolA, an *E. coli* membrane protein (Reichmann & Holliger, 1997). Upon contact, the phage transfers its single stranded (ss) genome into the host cell and the coat proteins insert into the bacterial outer membrane (Nakamura *et al.*, 2003). The phage genome is immediately converted by bacterial enzymes to double stranded (ds) DNA to create the replicative form (RF). The RF replicates using the method of rolling circle replication where the build up of single strand (ss) binding protein pV sequesters the displaced strand leading to the production of plus-strand copies of phage DNA coated by pV, preventing conversion to dsDNA. Structural proteins pVIII, pVII, pXI, pVI and pIII spontaneously insert into the inner membrane of the bacterium as they are synthesised and viral particles are assembled from the pV-coated ss genomes. During assembly, pV is removed from the ssDNA and the phage coated in pVIII, then pVII and pIX (Kehoe & Kay, 2005). The virion is extruded from the bacterial cell through a membrane pore and pIII and pVI are added to the end at this stage. Infected *E. coli* cells can grow and divide indefinitely, albeit at half the rate of uninfected bacteria (Kehoe & Kay, 2005).

Phage display involves the expression of proteins on the surface of filamentous phage by splicing foreign inserts into the genome (Azzazy & Highsmith, 2002). Phage display differs from conventional expression systems in that the foreign DNA is spliced into the gene for a major coat protein (**Figure 10**), resulting in the fusion of the foreign amino acids of the insert to the endogenous amino acids of the coat protein. This insert is then replicated along with the *E. coli* host and expressed as a protein - displayed - on the phage surface. By inserting potential peptide-encoding sequences into the cloning site, a peptide library is produced, containing several billion protein sequences encoded from heterologous foreign insert DNA.



**Figure 9 M13 life cycle.** A) a bacteriophage approaches an *E. coli* cell and attaches to its F' pili using coat protein pIII and by retracting the pilus, the bacteriophage is brought into close contact with the bacterium. The TolQRA (3 protein) complex participates in channel formation (Reichmann & Holliger, 1997), where the phage ss genome is brought into the cell, and the phage coat proteins are inserted into the membrane. B) Conversion of the ss phage genome by the bacterium results in the ds replicative form in C). D) Replication of the RF is by the rolling circle method, leading to E) plus-strand copies of phage DNA which are coated with pV to prevent conversion to dsDNA. At this stage, there will be a build up of structural phage proteins in the bacterial membrane and viral particles are assembled as they pass into the membrane. F) Following addition of the phage coat proteins, the assembled phage is extruded from the bacterial cell through a membrane pore.

Resulting libraries are subsequently screened to purify specific phage-encoded sequences using affinity selection between protein and ligand (Azzazy & Highsmith, 2002) – a process known as biopanning. Clearly, the success of a phage display and panning experiment is dependent on the quality, size and range of the initial library (Sidhu, 2001).

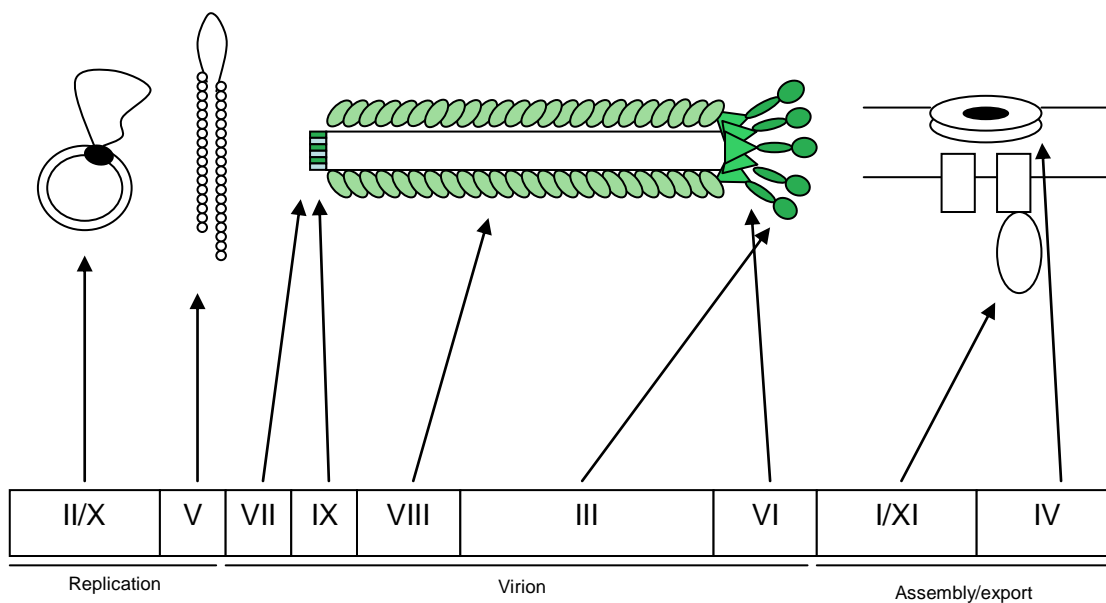


**Figure 10 Phagemid Displaying Fusion Peptides on Gene VIII.** The red block in the phagemid genome depicts inserted foreign DNA, which is then translated by the host machinery and expressed as a fusion protein (red circle) on the surface of the converted phage coat protein.

### Phage display application

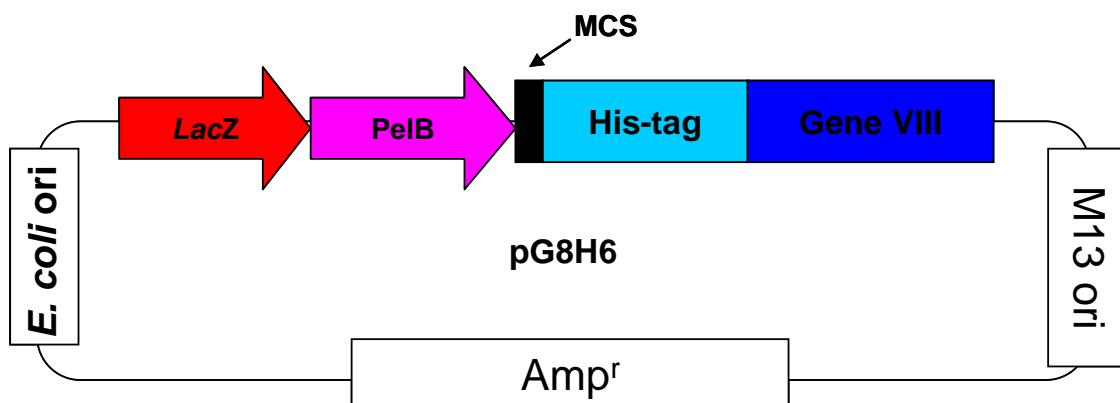
The power of phage display lies in its ability to (i) maintain a physical link between the displayed protein and the DNA sequence encoding it, i.e., between phenotype and genotype (**Figure 10**) (Christensen *et al.*, 2001), and (ii) functionally screen huge libraries containing billions of unique peptides and proteins (Russel, 2007a) for specific traits.

Filamentous phage fd, f1 and M13 are used for phage display. They are almost identical in structure and biology (Smith & Petrenko, 1997; Marvin *et al.*, 2006), 1  $\mu\text{m}$  in length and around 10nm wide, with a protein coat encasing the ss DNA genome of 11 genes (Sidhu, 2001). They are specific for bacteria carrying F-pili (such as *E. coli*) and are known as Ff phage (F- specific filamentous) (Marvin *et al.*, 2006). Of the 11 genes, depicted in **Figure 11**, 5 code for coat proteins – one (p8) is the major coat protein and the other 4 are minor coat proteins (p3, p6, p7 and p9).



**Figure 11 Filamentous Phage f1 (M13/fd) genes and gene products. Gene VIII encodes the major coat protein, the protein used for display in this project. Gene II encodes p2 which initiates replication by host proteins. P10 is required for the switch to ssDNA accumulation. Gene V encodes the ssDNA binding protein p5. Genes VII and IX encode two small proteins that are first to exit the cell during assembly. Gene VIII encodes the major coat protein p8, and genes III and VI encode p3 and p6, located at the end of the virion. These proteins are responsible for termination of assembly, virion release and infection. Gene I encodes p1 and p11 which are essential cytoplasmic membrane proteins. Gene IV encodes p4, a multimeric outer membrane protein channel through which the phage exits the bacterium (Russel, 2007a).**

All 5 confer structural stability and two in particular are used for phage display – p3 and p8. Gene III encodes protein 3 (p3) and is present in 3-5 copies per phage. The function of p3 is host cell recognition and infection, and it is the largest of the coat proteins at 406 aa (Sidhu, 2001). Gene VIII, which encodes protein 8 (p8) is present in ~2700 copies per phage (Smith & Petrenko, 1997) and is 50 aa in size. P8 molecules are arranged in a repeating helical array, with exposed N termini on the surface and the C termini concealed at the core (Sidhu, 2001). Because each phage particle contains several thousand p8 molecules, fused peptides are displayed in a polyvalent format (Russel, 2007a), resulting in polyvalent display of fusion proteins. The remaining genes encode proteins required for viral replication and assembly (Sidhu, 2001). Phage do not accept peptides longer than 6 amino acids so for display of larger peptides phagemid vectors are used (**Figure 12**). Phagemid are hybrids of phage and plasmid vectors (Mullen *et al.*, 2006), and contain a modified version of gene VIII which, upon infection with a helper phage, hybrid capsids containing the fusion protein are assembled and dispersed in an otherwise wild-type capsid (Cesareni, 1992). Phagemid do not contain an active phage double stranded origin of replication and lack all of the genes required to make a complete phage. Helper phage infection is therefore required following replication in *E. coli* as a double-stranded plasmid, which deliver a native copy of all other proteins required for



**Figure 12 pG8H6 Phagemid.** pG8H6 (2.6kb) contains a His-tag inserted between the PelB leader sequence and the MCS. It introduces a ribosomal slippage, so down-regulating the fusion peptide. This enhances progeny phage production as *E. coli* viability deteriorates when too many fusion peptides are waiting in the membrane for assembly.

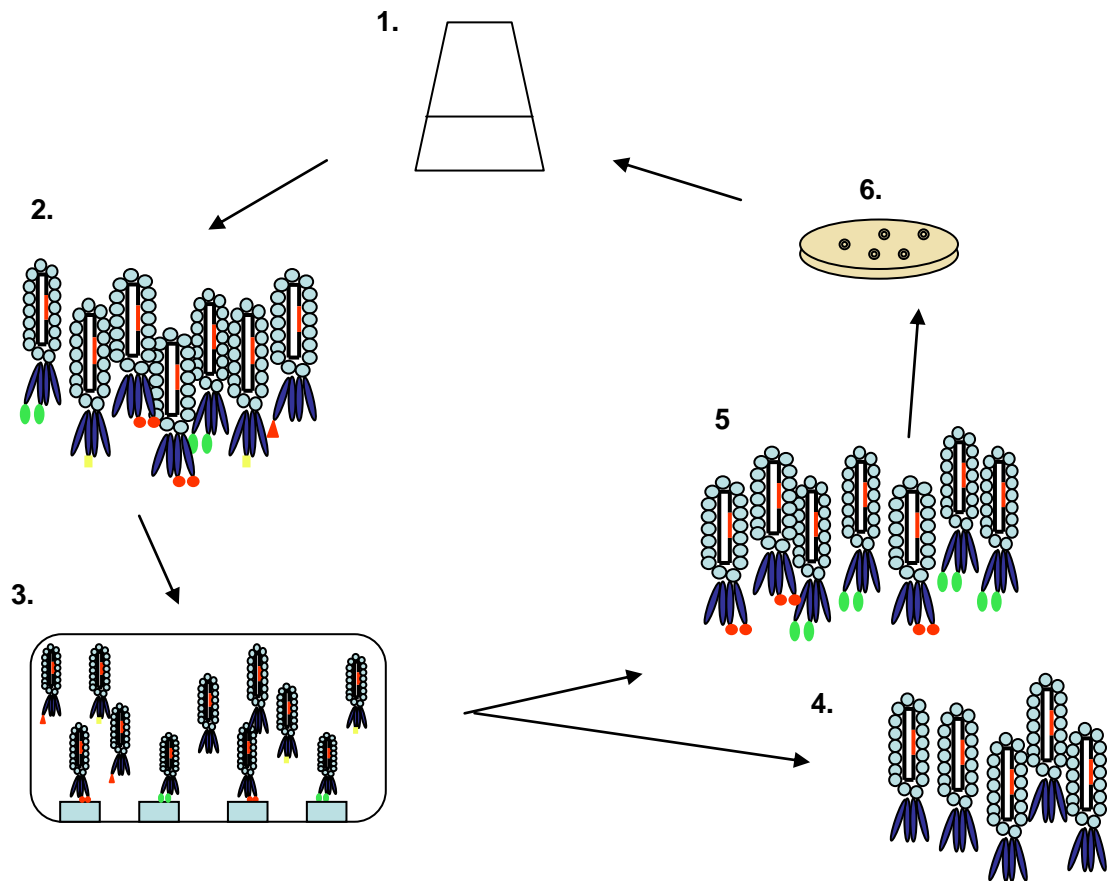
replication and packaging (Russel *et al.*, 1986), avoiding phage instability as a result of the expression of overly large proteins in fusion to every copy of gene VIII (Sidhu, 2001).

Phagemid often incorporate specific features which make them valuable for phage display technology such poly-His tags for ease of expression. Many phagemid use the *lacZ* promoter to drive expression of the coat protein fusion. To display the gene VIII product, the *lacZ* promoters' catabolic repressor (glucose) is simply removed, which allows generation of the fusion product and polyvalent phage particles (Hoogenboom *et al.*, 1998).

### Affinity selection – Biopanning

Once phage display library construction is complete, the natural specificity and affinity of fusion proteins displayed in the library can be exploited to search for genes or fragments of interest. Panning (**Figure 13**) is an affinity purification technique which involves immobilization of a ligand on tubes or plates, then phage library incubation with the target long enough to allow the formation of protein-ligand bonds (Smith & Scott, 1993). The minority of phage whose displayed peptides bind to the target are retained while the non-specifically bound phage are removed by three vigorous washing steps. The hugely enriched bound phage are eluted and, still infective, are propagated by infecting log phase *E. coli*. This infection produces a hugely amplified eluate which feeds directly into the next round of selection. The first round of panning normally contains proteins with a range of binding affinities, and therefore a range of binding proteins with varying degrees of specificity. Three to six rounds of panning are recommended for optimum enrichment of clones which bind tightly to the target





**Figure 13 Panning Process.** (1) Grow phage library, (2) Purify phage, (3) Incubate in tube with immobilised antigen or ligand, (4) Remove unbound or weakly bound phage by washing, (5) Elute bound phage, infect host bacteria and grow colonies, (6) Pick colonies and analyse the selected DNA/protein using public databases and *in silico* tools.

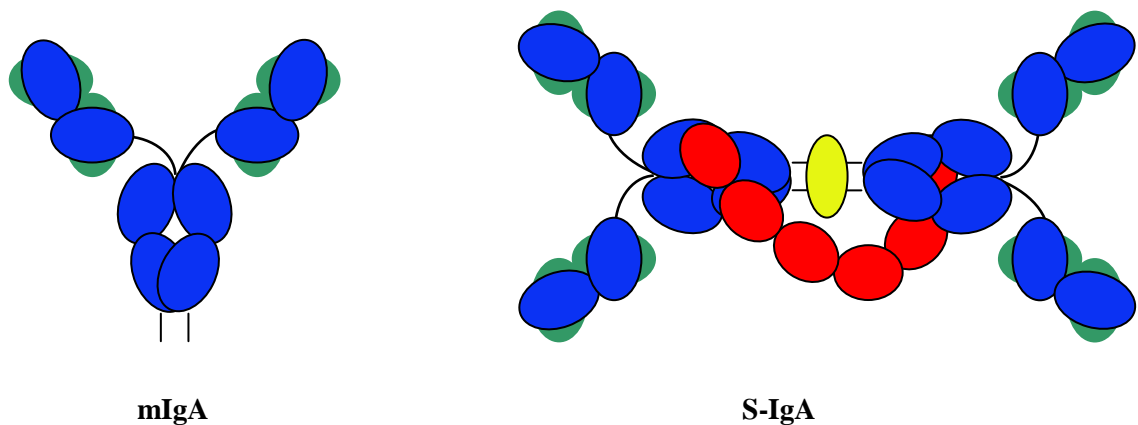
protein (Smith & Petrenko, 1997), theoretically resulting in a phage pool dominated by one or more specific binders. This has been the case with previous phage display studies which have made libraries from pure culture DNA (Jacobsson & Frykberg, 1996; Mullen *et al.*, 2007) however, incorporating metagenomic DNA in phage display libraries could lead to rapid identification of many more binding proteins. As yet, there are no studies which have used metagenomics and phage display in tandem.

Specific binding phage from panning are analysed individually (Smith & Petrenko, 1997), most commonly by sequencing the insert following excision from the phagemid. *In silico* analysis of the encoded DNA and amino acid sequences can lead to a tentative identification of the bound peptide, its function and the bacteria it originated from based on homology to public sequence databases such as the Genbank database (Benson *et al.*, 2008), part of NCBI (National Centre for Biotechnology Information: [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)).

**Panning ligand – IgA**

Mucosal surfaces like the oral cavity produce secretions containing high levels of immunoglobulins (Ig) which protect from bacterial attack. The predominant adaptive immune factor in secretions is secreted IgA (S-IgA), which is produced in daily quantities which far exceed the combined production of all other Ig isotypes (Kilian, 2003; Woof & Mestecky, 2005). As the most concentrated immunoglobulin in the oral cavity, IgA is not bound to any surface but functions as a bacterial receptor (Ahl & Reinholdt, 1991) and protective component of relevant surfaces and tissues.

IgA is locally produced by plasma cells, plentifully located in the mucosal subepithelium (Brandtzaeg & Johansen, 2005) and is actively transported to the surface by selective receptor-mediated transepithelial transport (Woof & Mestecky, 2005). IgA is highly heterogeneous in external secretions and exists as monomers (mIgA) and polymers (pIgA), which are made up of dimers and tetramers of mIgA linked by small polypeptides called J chains. 50-90% of pIgA is associated with the secretory component (SC), an extracellular part of the Ig receptor, linked to the Fc portion of the molecule (Woof & Mestecky, 2005). The most important distinguishing feature of S-IgA (**Figure 14**) is the presence of the associated glycoprotein SC, bound by J chain, which is not only the receptor for transepithelial transport of polymeric S-IgA to mucosal secretions, but increases S-IgA stability and protects it from proteolysis (Bruce *et al.*, 1989).



**Figure 14** Schematic representations of monomeric IgA and S-IgA. S-IgA is a double-Y shaped molecule of 2 monomers joined at the Fc region. Heavy chains are shown in blue, light chains in green, J chain in yellow, SC in red. For clarity, carbohydrates have not been included. Adapted from Woof & Mestecky (2005).

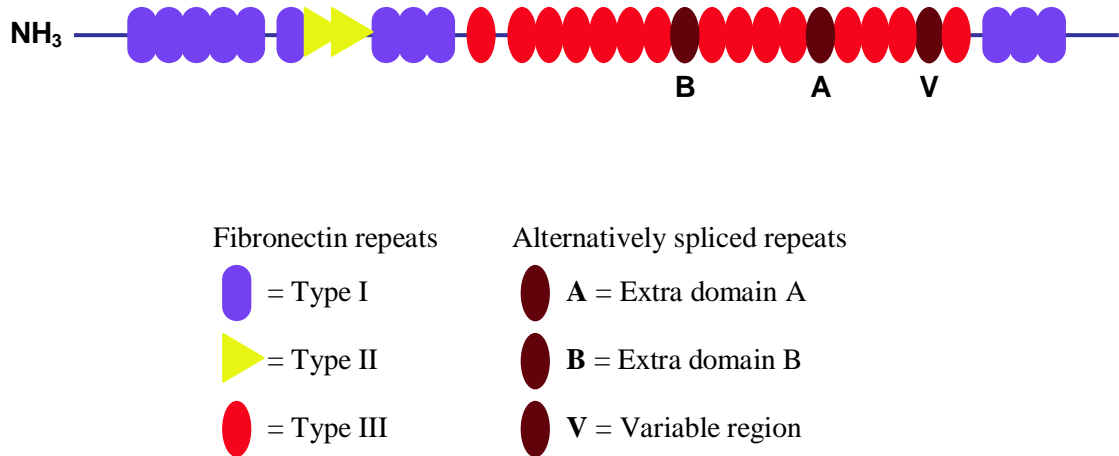
Polymeric IgA with J chain bound by SC at the cell surface is internalized into vesicles and transported through the cell. Following vesicle fusion with the cell membrane, the complex is released as S-IgA with the SC component remaining bound and intact (Bruce *et al.*, 1989). IgA dimers contain multiple antigen binding sites - four for IgA dimers, and eight for IgA tetramers - and inhibit bacterial adherence to epithelia by imparting a negatively charged coating, repelling them from prime colonisation surfaces (Woof & Mestecky, 2005). Additionally, S-IgA mediates protection from bacterial disease by direct killing, agglutination, inhibition of invasion, inhibiting microbial adherence and opsonisation for uptake by phagocytes (Bruce *et al.*, 1989). pIgA and S-IgA are also extremely adept at neutralising the activity of viruses, enzymes and toxins through epitope binding which they may do alone or in concert with other host defence mechanisms (Ahl & Reinholdt, 1991; Russel & Mestecky, 2002).

*S. pneumoniae* expresses surface proteins that bind S-IgA through the secretory component (Hammerschmidt *et al.*, 1997). This surface protein (SpsA) of *S. pneumoniae* that binds to SC is expressed by 2/3 of strains and conserved between different serotypes (Hammerschmidt *et al.*, 1997). Other bacteria, notably *Streptococcus pyogenes*, bind the Fc part of human IgA as part of their virulence capacity using the IgA binding proteins Arp4 and Sir22 (Johnsson *et al.*, 1994 & 1999).

Although in general, immunoglobulins are resistant to bacterial degradation, some pathogenic microbes produce proteases that cleave IgA molecules in the hinge region, thus removing their protective effect (Ahl & Reinholdt, 1991). Of the two IgA subclasses present in the oral cavity, IgA1 is far more commonly produced than IgA2, at between 70 - 95%. The addition of a 13aa stretch at the hinge region of IgA1 confers added flexibility to the molecule but also makes it more susceptible to IgA1 proteases from the commensal Gram positive bacteria of the oral cavity (Kilian., 2003). The resistance of IgA to degradation indicates how important immunoglobulins are in maintaining the homeostasis of the oral cavity (Dumas *et al.*, 1987). Secretory IgA coats gut microbiota in healthy individuals and mucosal IgA has been shown to bind commensal bacteria specifically suggesting that the commensal flora drives the secretory IgA system (Pickard *et al.*, 2004). Since binding events are so common, IgA a clear choice for inclusion in panning experiments.

### **Panning ligand - Fibronectin**

Fibronectin is a ~450 kDa modular glycoprotein involved in extracellular matrix (ECM) interactions as well as cell adhesion, migration, growth and differentiation, and found on the surface of mammalian cells (Pankov & Yamada, 2002) (**Figure 15**). It is a well characterised constituent of the tongue basement membrane zone (Couchman *et al.*, 1979), and



**Figure 15 Schematic representation of Fibronectin molecule. From: Fibronectin at a Glance, Pankov & Yamada (2002). The image here is cellular FN although FN is also found in plasma. Type I modules are ~40 amino acids long, type II are ~60 amino acids long and type III ~90 amino acids long. Type I and II contain 2 disulphide bonds each, where type III contains none (Pankov & Yamada, 2002). The main binding activity is in the N-terminal end, consisting of five type I molecules (Joh *et al.*, 1998).**

although arising from a single gene, can appear in up to 20 variants which are generated by including or excluding some of the Type I, II or III repeats.

FN plays a crucial role in the insoluble extracellular matrix (ECM), where it acts as a ligand for integrin receptors – linking the ECM with the intracellular cytoskeleton - and many other biologically active molecules including heparin, collagen and fibrin. FN matrix assembly – or fibrillogenesis – is the process of creating and depositing FN fibrils into the ECM. Fibrils are aggregates of FN, where the molecule self-associates in line with binding sites along its length. FN is one of the largest multi-domain proteins to have been studied in great detail. As such it is well known to serve as a ligand for bacterial adhesion.

Many bacteria use FN to gain a foothold onto human cells and tissues. In particular, streptococci and staphylococci encode fibronectin binding proteins, often in a specific N-terminal 30 kDa region of the FN molecule (Williams *et al.*, 2002), and a large body of work exists on FNBP's (Mitchell, 2008; Mullen *et al.*, 2008). In particular, FNBPA and FNBPB of *S. aureus* are known to mediate adherence to endothelial and epithelial cells, which suggests their importance as virulence factors (Peacock *et al.*, 1999; Dziewanowska *et al.*, 1999).

In a general sense, bacterial ECM binding proteins are termed MSCRAMMS (Patti *et al.*, 1994) and FN binding proteins share structural similarity to other cell wall proteins of Gram positive bacteria (Joh, 1999). FN can also act as a bridge between bacterial binding

proteins and other, less accessible, host cell components such as integrins (Dziewanowska *et al.*, 2000), bound by FN as part of the ECM.

FN binding has been associated with the propensity to establish colonisation of host tissues, and this highly multivalent binding is mediated by several high affinity binding sites (Schwarz-Linek *et al.*, 2003; Meenan *et al.*, 2007). It is becoming apparent that the majority of bacteria which colonise mammals express FNBP's (Schwarz-Linek *et al.*, 2004), the majority of which bind to the N terminus of the FNI modules 1 to 5 (Schwarz-Linek *et al.*, 2006). Many staphylococci and streptococci express FN binding proteins (Joh *et al.*, 1998), notably the 100 kDa FNBPA (Signas *et al.*, 1989) and FNBPB (Jonsson *et al.*, 1991) of *S. aureus* which are similar in structure to other FNBP's, and which facilitate mechanically resistant colonisation of host tissues (Mitchell, 2008). The FNBP's of *S. aureus* and *S. pyogenes* all contain the cell wall anchoring motif LPX[T,S,A]G and a short positively charged C-terminal intracellular tail (Schwarz-Linek *et al.*, 2006).

Although less information is available on FNBP's of Gram negative bacteria, what is known is that they demonstrate structural and functional resemblances to the known Gram positive FNBP's of *S. aureus* and *S. pyogenes* (Raibaud *et al.*, 2006). BBK32 from *Borrelia burgdorferi*, the causative agent of Lyme disease was the first example of such a protein. In contrast, Mullen *et al.*, (2008) identified a novel FNBP in the Gram negative bacteria *P. multocida*, the gene (PM1665) has homologous proteins in all other sequenced members of the *Pasteurellaceae*. The group found that PM1665 did not bind to the N terminal 30 kDa or 45 kDa fragments many other FNBP's bind, but instead bound to the 120-kDa central cell binding segment which mediates active adhesion of FN to cell surface integrins. The nature of the binding site suggests a novel mechanism of binding action.

Undoubtedly, far more is known about FN-binding proteins than IgA binding proteins, but the tendency of oral bacteria like *Streptococcus* sp. to bind to FN extends the likelihood that the phage display library in this project will lead to the identification of more FN binding proteins.

### **Aim of this project**

This project had two main aims:

1. To investigate the use of metagenomic DNA in a phage display library
2. To test the hypothesis that an enormous variety of binding events are taking place on the human tongue and that they can be investigated further with this combination of molecular tools.

---

**Chapter 2**  
**Materials and Methods**

---

**Media, Solutions and Strains****(a) Media**

All media was prepared and sterilized by autoclaving at 121°C and 15 p.s.i. for 20 min unless otherwise stated.

<i>Growth Media</i>	<i>Broth</i>	<i>Agar Plates</i>
Nutrient Broth (Oxoid)	13g/L	+ 4g agarose
Nutrient Broth No. 2 (Oxoid)	25g/L	+ 4g agarose
Luria-Bertani	10g Tryptone, 10g Yeast Extract (Difco), 5g Sodium Chloride, 1mg/ml Sodium Hydroxide (BDH)	n/a
NB2 + 4% Glucose	25g/L NB2 + 4g Glucose (BDH)	+ 4g agarose
DNase Agar (Oxoid)	39g/L	n/a

**Table 1 Growth media****Ampicillin-containing media**

A stock solution of 100 milligrams per millilitre ampicillin (Sigma, St. Louis, MO, USA) was made up in distilled water, filter sterilized and stored at -20°C. It was added to autoclaved nutrient agar (<50°C) to a concentration of 100µg/ml and poured into 9cm Petri dishes (Western Laboratory, Hampshire, England).

**(b) Solutions**

This is a note of general solutions and mixtures used during the course of this project. All solutions were prepared and sterilized by autoclaving at 121°C, 15 p.s.i. for 20 min unless otherwise stated.

**Loading buffer for agarose gels**

Four grams of sucrose (BDH, Poole, England) was added to a sterile universal tube (20ml). Two millilitres of 0.5M EDTA and 1ml bromophenol blue (BPB, 1.5mg/ml) were added and the solution was made up to 10ml with sterile distilled water. The loading buffer was not sterilized.

**λPst agarose gel ladder**

One hundred micrograms (µg) of bacteriophage λ DNA (New England Biolabs, Hertfordshire, London, UK) was digested fully with the restriction enzyme *Pst*I (New England Biolabs). After 3 hours incubation at 37°C, digestion was verified as complete by testing an aliquot on an

agarose gel. When complete, 450 µl of Loading Buffer was added. The ladder was stored at -20°C until required. The ladder produced clear fragments between 5 kb and 200 bp (**Appendix 2**) and was included with every DNA gel for uniform comparison.

#### **40% (w/v) PEG (polyethylene glycol)/2.5M NaCl**

2.5M NaCl was prepared by measuring 73.55 grams NaCl (BDH, Poole, England) into a 1L Duran bottle and adding distilled water up to 1L. The solution was sterilised by autoclaving. 160g PEG 8000 (Sigma, St Louis, MO, USA) was added to a fresh 400ml Duran bottle and 2.5M NaCl was added up to 400ml. This solution was stirred on a heated magnetic stirrer for 2 hours, or until the PEG was completely dissolved.

#### **Sodium carbonate buffer**

Two starting solutions were required to make this buffer. 100mM Sodium Carbonate (Bicarbonate) buffer was prepared by adding 1.682g Sodium Carbonate (BDH, Poole, England) to 200 ml distilled water. 100 mM Sodium Carbonate anhydrous was prepared by adding 2.138g Sodium Carbonate anhydrous (BDH, Poole, England) to 200 ml distilled water. A volume of 4 ml Sodium Carbonate anhydrous was added to 46 ml Sodium Carbonate in a sterile Duran bottle to make 100 mM Sodium Carbonate Buffer, pH 9.4.

#### **Phosphate buffered saline (PBS) and PBS-T (PBS + 0.05% Tween 20)**

PBS, used for panning, was prepared by making a solution of 137 mM NaCl, 2.7 mM KCl (BDH, Poole, England), 10 mM Na<sub>2</sub>HPO<sub>4</sub> and 2 mM KH<sub>2</sub>PO<sub>4</sub> in a 1L Duran bottle. PBS-T (pH 7.4) was prepared by adding 500µl Tween20 (Sigma, St Louis, MO, USA) to 1L PBS after autoclaving.

#### **10% (w/v) CTAB solution**

In order to make up this solution for DNA extraction, five grams of CTAB powder (Hexadecyltrimethyl ammonium bromide) (Sigma, St Louis, MO, USA) was added to 50ml 0.7M NaCl solution and autoclaved.

#### **TE buffer with RNase**

For this DNA extraction solution, a final concentration of 10 mM EDTA solution and 100 mM Tris-HCl (pH 8.2) was in 1L of distilled water. Forty millilitres of TE buffer was poured into a 50 ml Falcon and 50 µl (10mg/ml) RNaseA (Fermentas, Hanover, MD, USA) added before use. This solution was prepared when required and the excess was not stored.



**TE + 20% (v/v) glycerol**

This phage eluate recovery solution contained a final volume of 10 mM EDTA solution and 100 mM Tris-HCl (pH 8.2) in 800 ml distilled water. Glycerol was added to 1L and the solution autoclaved.

**BCIP/NBT (5-bromo-4-chloro-3-indolyl phosphate/nitro blue tetrazolium)**

This colour change agent used in antibody screening experiments was prepared by adding separately, to water, 50 mg/ml BCIP and 10 mg/ml NBT. When ready to use, 33 µl BCIP and 330 µl NBT were added to substrate buffer.

**(c) Strains, Bacteriophage and Vectors: source, preparation and storage**

All bacterial strains and bacteriophage used are given in **Table 2**. All phagemid vectors and general cloning vectors used are given in **Table 3**.

<i>Strain/Isolate</i>	<i>Genotype</i> <sup>(Source)</sup>
<i>Escherichia coli</i> TG1	[K12;Δ( <i>lac-proAB</i> ) <i>supE thi hsdD5/F'</i> ( <i>traD36 proA+proB+lacIg lacZΔM15</i> )] <sup>1</sup>
<i>Escherichia coli</i> JM107	<i>endA1, glnV44, thi-1, gyrA96, hsdR17 (R<sub>K</sub>-m<sub>K</sub><sup>+</sup>)λ<sup>-</sup>, supE44, relA1, λ, Δ(<i>lac-proAB</i>), [F', <i>traD36, proAB</i><sup>+</sup> <i>lacI</i><sup>q</sup>, ZΔM15]<sup>2</sup></i>
<i>Escherichia coli</i> DH5α	F <sup>-</sup> <i>endA1, glnV44, thi-1, recA1, relA1, gyrA96, deoR, nupG, θ80dlacZΔM15 Δ(<i>lacZYA-argF</i>) U169, hsdR17 (r<sub>K</sub>-m<sub>K</sub><sup>+</sup>), λ<sup>-2</sup></i>

1. Eastman Dental Institute, Grays Inn Road, London, WC1X 8LD

2. University College London, Gower Street, London, WC1E 6BT.

**Table 2 Names and Sources of Bacterial Strains****Preparation of glycerol stocks**

From agar plates: 25% glycerol (v/v) was prepared by adding 50ml Glycerol (BDH, Poole, England) to 150ml distilled water and autoclaving. Four millilitres of the 25% glycerol solution was dispensed onto a fresh overnight agar plate and the bacteria were suspended into solution by lightly scraping the surface of the plate with an inoculating loop. The suspension was removed from the plate by pipetting and dispensed into a sterile 5ml Falcon tube (Falcon, Becton Dickinson Labware Europe, France). The suspension was stored at -20°C.

From liquid culture: 50% glycerol (v/v) was prepared by adding 100ml Glycerol (BDH) to 100ml distilled water and autoclaving. Two millilitres of 50% glycerol solution was dispensed into a 5ml tube (Falcon) and 2ml of a fresh liquid bacterial culture was added to it. The suspension was mixed thoroughly before storage at -20°C.

<i>Vector</i>	<i>Source</i>
pG8SAET phagemid	Gene VIII fusion vector, <i>E. coli</i> origin, MCS ( <i>Sna</i> BI), E-tag (for enrichment of clones with an ORF), Amp <sup>r</sup> <sup>1</sup>
pG8H6 phagemid	Gene VIII fusion vector, <i>E. coli</i> origin, <i>lacZ</i> , PelB leader sequence, His-tag, MCS ( <i>Sma</i> I), Amp <sup>r</sup> <sup>1</sup>
pUC 19 cloning vector	pBR322 based cloning vector, pMB1 origin, MCS (nt 397-454 inverted), <i>lacZ</i> $\alpha$ , Amp <sup>r</sup> <sup>2</sup>

**Table 3 Names and Sources of Bacterial and Phagemid Vectors Used. 1. Eastman Dental Institute, Grays Inn Road, London, WC1X 8LD. 2. New England Biolabs, Hertfordshire, England, UK.**

### Phagemid vector isolation

Phagemid were isolated using the Qiagen Midiprep Kit (West Sussex, England) according to manufacturers' instructions. Briefly, 50 ml LB broth in a 500 ml conical flask was inoculated with one colony of pG8SAET or pG8H6 from a fresh overnight colony grown at 37°C and 200 rpm in NB2 broth plus 100 µg/ml ampicillin. The culture was centrifuged at 2,500 x g for 15 min, the supernatant discarded and the pellet resuspended in 250 µl Buffer P1 containing RNase. Buffer P2 (NaOH/SDS) 250 µl, was added to lyse the bacteria under alkaline conditions which denatures chromosomal and plasmid DNA. Buffer N3 (acetic acid) 350 µl, was added to neutralise the lysate and allow rapid renaturation of the plasmid DNA, in addition to creating a high salt environment. This eases plasmid DNA binding to the extraction column membrane and allows precipitation of denatured proteins, chromosomal DNA, cellular debris and SDS, but not precipitation of the plasmid DNA. Samples were left in the syringe for 10 min for precipitation and syringed into an equilibrated Midi column which was left to empty by gravity flow. Buffer QC (Isopropanol) 10ml, was used to wash the column and the DNA was eluted with Buffer QF, 5ml, both allowed to travel through the column by gravity flow. The eluted DNA was precipitated by mixing with isopropanol at room temperature (22°C) and filtering through a QIAprecipitator into a waste bottle. The DNA trapped in the precipitator was washed with 70% (v/v) ethanol and eluted using 1 ml Buffer TE and a syringe. The

concentration of vector DNA following the procedure was determined using a Nanodrop™ (ND-1000 Spectrophotometer) and found on average to be between 100 – 200 ng/μl.

### **Phagemid vector digestion and dephosphorylation**

The phagemid DNA extracted from the Midiprep Kit (Qiagen) was sufficiently pure for immediate digestion by restriction enzymes following extraction. Therefore, aliquots of pG8SAET or pG8H6 DNA were digested using the restriction enzymes *Sna*BI and *Sma*I respectively (both New England Biolabs). pG8SAET required 3 hours digestion at 37°C whereas pG8H6 digestion by *Sma*I required incubation at 25°C for 3 hours. Complete digestion was verified by testing a 10 μl aliquot of digested vector against an aliquot of undigested vector on an agarose gel. Once verified, the restriction enzyme was inactivated by heating in a Thermomixer™ (Microcentrifuge, Hamburg, Germany) at 80°C for 20 min. In order to prevent recircularisation of the vector it was necessary to dephosphorylate it. To do this, Antarctic Phosphatase Buffer (New England Biolabs) was added to an aliquot of vector after digest, to a concentration of no less than 10%, this is due to Antarctic Phosphatases' requirement for Zn<sup>2+</sup>. Antarctic Phosphatase (NEB) was added to 3U per reaction (1U/μl). The sample was incubated for 30 min at 37°C, and heat treating at 65°C for 5 min denatured the enzyme. Further purification was not needed.

### **(d) Bacterial DNA Sampling, Extraction and Preparation**

#### **Tongue scraping protocol**

Volunteers from the research group were asked to provide tongue scraping samples provided that they had not taken antibiotics for 6 months previous to sampling. Each volunteer was given a sample sheet with sampling instructions (see Appendix 1), 12 universal tubes each containing 10 ml sterile Dulbecco's PBS (Sigma), a pack of sterile tissues (VWR, Spartanburg, SC, USA) and a toothbrush (Boots, UK). Sampling instructions were, briefly, swab excess saliva from the tongue dorsum using a sterile tissue, gently so as not to remove loosely attached micro-organisms. Brush the tongue dorsum with the toothbrush as vigorously as is comfortable for one minute, dislodging bacteria from the toothbrush periodically by shaking in an aliquot of sterile PBS. The sample was frozen immediately at -20°C until needed for further processing, and the toothbrush was stored at -20°C until next use.

#### **Sample storage in isopropanol**

According to Torsvik *et al*, 1990, the yield of DNA from an environmental sample can be increased by storing the sample in isopropanol before extraction. This principle was used by Torsvik on soil bacteria although the mechanism by which it works is not clear. Therefore, tests were carried out in the current study on tongue bacterial samples. Briefly, 12 tongue scraping

samples were thawed and centrifuged at 2,500 x *g* for 10 min at 4°C and the supernatant removed. Five millilitres of isopropanol (VWR) was added to each sample and the pellets resuspended and left at -20°C for 1 to 7 days. Aliquots were removed and the DNA extracted after each 24 hour period to check the optimum duration for isopropanol treatment. Following the isopropanol testing, all remaining bacterial samples were stored in isopropanol for 7 days before DNA removal using the CTAB extraction method (Bailey, 1995).

### **DNA extraction by CTAB method**

Each 5 ml tongue scraping sample was dispensed into four 1.5ml Microcentrifuge tubes (Trefflab, Degersheim, Switzerland), centrifuged for 1 minute at 8,000 x *g* and the supernatant removed. Pellets were resuspended in 500µl fresh lysis buffer, prepared before each extraction (20µg/ml Proteinase K (New England Biolabs) in 0.5% SDS (w/v) (BDH)), and maintained at 55°C for 30 min, with gentle shaking at 10 min intervals throughout. 100 µl of prewarmed NaCl (5M) and 80 µl CTAB solution (10%) were added and the temperature increased to 65°C for a further 10 min. Addition of 680 µl isoamyl alcohol:chloroform (1:24) (Sigma) and subsequent shaking formed an emulsion which, when centrifuged for 10 min at 8,000 x *g*, separated into 3 distinct layers. The top aqueous layer was removed to a clean sterile microcentrifuge tube and 360 µl isopropanol added which precipitated the DNA during storage at 4°C overnight. After centrifugation at for 10 min, and removal of excess isopropanol, DNA pellets were washed by adding 300 µl 70% (v/v) ethanol and centrifuged for a further 10 min at 8,000 x *g*. After removal of excess ethanol, the pellets were allowed to air dry and the pellet resuspended overnight in 50 µl TE buffer containing RNase at room temperature. At this stage, the average concentration of DNA recovered was around 1000 ng/µl by Nanodrop™ (ND-1000 Spectrophotometer). All samples were extracted separately, and stored as individual volunteer samples at -20°C. Genomic DNA fragment size was determined using Pulsed Field Gel Electrophoresis (PFGE) (Gene Navigator Pulsed Field System, Pharmacia, LKB) using a MidRange II PFG Marker and a Low Range PFG Marker (both NEB) and was shown to be between 15 and 40 kb.

### **Ethanol Precipitation**

Ethanol precipitation was used to concentrate DNA samples following enzymatic treatments. Briefly, one tenth reaction mixture volume in 5M NaCl was added and mixed, followed by two volumes of absolute ethanol (BDH). After 1 hour precipitation at -20°C, samples were centrifuged at 8,000 x *g* for 10 min and supernatant gently removed. Samples were washed with 200 µl 70% ethanol, centrifuged again and the supernatant discarded. After air drying completely, the DNA was resuspended in 50 µl EB or sterile distilled water if needed for sequencing.

### **(e) Phage Display Library Production**

#### **DNA fragmentation and repair**

Before fragmentation of total metagenomic DNA, 50 µl was removed from each of the nine individual DNA samples and combined in a 1.5 ml microcentrifuge tube for processing.

Total DNA was either fragmented using a sonicator or by partial restriction digest. Both approaches are described here. For sonication, 450 µl of bacterial DNA was fragmented using an MSE Soniprep 150 at 14 amplitude microns, in a 1.5 µl microcentrifuge tube. In order to get an initial idea of the time required for suitable fragment generation from sonication, the full aliquot was sonicated for 10 seconds, then a small volume removed to a separate labelled tube, the remainder sonicated for another 5 seconds and another small volume removed, and so on, until the last 50-75 µl had been sonicated for 30 seconds. An aliquot of each sample was tested on an agarose gel against λPst ladder and the optimum sonication time confirmed as 15-18 seconds for a 450 µl sample volume. After sonication, exposure to shearing forces results in DNA fragments with overhanging ends. These must be repaired in order to optimise ligation into the blunt end vector. This was carried out using dNTP's (10mM) and DNA polymerase (~18U) (both New England Biolabs), which were added to an aliquot of fragmented DNA and incubated at 37°C for 30 min. DNA polymerase was inactivated by heating to 65°C for 10 min.

For restriction digest, bacterial DNA was fragmented using blunt end restriction enzymes *RsaI*, *AluI* and *HaeIII* (both New England Biolabs). Ten µl of each enzyme was added to 450 µl of DNA, along with buffer 4 (New England Biolabs), the solution mixed and incubated at 37°C. Initially, to determine the optimum time for digestion, aliquots were removed after every 10 min of digestion following an initial incubation of 20 min, and tested on an agarose gel. The optimum digestion time was found to be 20 min for 450 µl of sample. The enzymes were inactivated by heating in a Thermomixer at 80°C for 20 min.

#### **DNA visualisation using agarose gels**

Tris Borate Electrophoresis Buffer (TBE 1X) was made up using pre-measured sachets, according to manufacturers' instructions. To prepare one gel, 100 ml of 1 X TBE Buffer was mixed with 1 gram of high grade agarose (Invitrogen) and microwaved for 2 min, shaking every 10 seconds after the solution reached boiling temperature. After heating, 5 µl ethidium bromide (EtBr, Fisher, Leicester, England) was added and stirred, the agarose was poured onto a glass slide sealed with autoclave tape, a 20 well comb fitted, and the gel left at room temperature to set. All agarose gels were run at 150 V for 2.5 – 3 hours. DNA bands were subsequently visualised using the Gene genius Bio imaging system and Genesnap software (version 6.05) from Syngene.

### **Ligation and purification**

Ligations between blunt ended vector and inserts require a high concentration of insert DNA compared to vector. Ligations were set up in 10µl volumes using a 3-fold concentration excess of insert to vector, determined by Nanodrop™. Before adding T4 DNA ligase and ligase buffer (both New England Biolabs), a few microlitres of the insert-vector mixture were removed as a control, then the ligations incubated at room temperature (22°C) for 2 hours or at 4°C overnight. After incubation an aliquot was removed and added to 10µl Loading Buffer, which was run alongside the control aliquot on an agarose gel to test the ligation efficiency.

In a successful ligation, removal of salt and other impurities is necessary prior to electroporation. The PCR Cleanup Kit from Qiagen is recommended for this by Jacobsson (2003), and was carried out according to manufacturers' instructions. Purified DNA was stored at -20°C until further use.

### **Preparation of electrocompetent cells**

Before beginning, 4L shake flasks were autoclaved with distilled water, and the water discarded before preparing LB. Care is taken when preparing electrocompetent cells (ECC) because traces of detergents or chemicals notably reduce electroporation efficiencies. An overnight culture of *Escherichia coli* was grown from an overnight colony, and 200 µl dispensed into one or two 4L shake flasks containing 1L sterile NB2 broth. Growth took several hours at 37°C with vigorous shaking (300 rpm) and was checked regularly using a handheld spectrophotometer (Biorad) and growth stopped once cells reached OD 0.7 or 0.8. The flasks were chilled for 30 min on ice, and the cells retained at 4°C for the remainder of the protocol. The cells were centrifuged in sterile chilled Sorvall bottles in a Sorvall Ultracentrifuge RC6+ using an SLA-1500 rotor at 3,000 x g at 4°C for 10 min. The supernatant was discarded and the cells washed by gently resuspending the pellet in sterile distilled water (sdH<sub>2</sub>O). This centrifugation - washing step was repeated a further 3 times, resulting in a reduction in the ionic strength of the cell suspension. The pellet was resuspended in 40 ml of ice cold, sterile 10% (v/v) glycerol and centrifuged in 40 ml Sorvall tubes using an SS-34 rotor at 3,000 x g for 15 min. Each pellet was resuspended in 2.5 ml ice cold, sterile 10% glycerol and dispensed into 1.5ml precooled microcentrifuge tubes in 50 µl aliquots. The cells were stored immediately at -80°C until further use.

### **Preparation of chemically competent cells**

A 5 ml overnight culture of *Escherichia coli* TG1 was added to 20 ml NB2 with 20 mM MgCl<sub>2</sub> in a 50 ml Falcon tube (Becton Dickinson Labware Europe, France) and left to grow for 60 min at 37°C with shaking (200 rpm). The cells were retrieved by centrifugation in a bench-top centrifuge (Microcentrifuge) at 2,500 x g for 10 min at 4°C. The pellet was maintained on ice and resuspended in 2ml ice-cold sterile 75mM CaCl<sub>2</sub> 15% glycerol. Aliquots

of 200  $\mu\text{l}$  were dispensed into pre-chilled 1.5ml microcentrifuge tubes and either used immediately or stored at  $-80^{\circ}\text{C}$  until required.

### **Electroporation**

Electroporations were carried out to maximise the efficiency at which phagemid with inserts were taken up into *E. coli*. The ligation mixture was added to the ECC and electroporated using a Biorad Micropulser at 2,400 volts. Immediately following application of the electric current, the cells were placed into 10 ml prewarmed ( $37^{\circ}\text{C}$ ) NB2 and allowed to grow at  $37^{\circ}\text{C}$ , with shaking at 200 rpm, for 2 – 2.5 hours. Aliquots (100  $\mu\text{l}$ ) of  $10^0$ ,  $10^{-1}$  and  $10^{-2}$  dilutions were then plated onto prewarmed NB2 + Ampicillin plates and grown at  $37^{\circ}\text{C}$  overnight. The 10 ml cultures were refrigerated at  $4^{\circ}\text{C}$  until the library was checked for inserts by counting the numbers of ampicillin resistant colonies.

### **Chemical transformation**

An aliquot (2  $\mu\text{l}$ ) of ligation was added to each tube of cells, mixed and left on ice for 30-45 min. The ligations were heat-shocked at  $42^{\circ}\text{C}$  for 45 seconds in a Thermomixer (Microcentrifuge) and returned to ice for a further 4 min, before adding each aliquot to 5 ml pre-warmed NB2. After growth for 2 hours at  $37^{\circ}\text{C}$  with shaking (200 rpm), aliquots of 100  $\mu\text{l}$  were plated onto pre-warmed ( $37^{\circ}\text{C}$ ) NB2 + Ampicillin plates and control plates of NB2 only, and grown overnight at  $37^{\circ}\text{C}$ . Transformation frequency was determined the following day by counting Ampicillin resistant colonies.

### **Plasmid extraction (Miniprep)**

In order to test the insert frequency of each library, between 20 and 40 colonies were picked from the plates, inoculated separately into 4 ml NB2 and grown overnight, at 200 rpm and  $37^{\circ}\text{C}$ . The DNA was extracted using the Qiagen Miniprep Kit to produce suitable purity plasmid DNA for subsequent digest and sequencing. Briefly, the overnight cultures were dispensed into 1.5 ml Microcentrifuges and centrifuged at  $8,000 \times g$  for 1 minute. The pellet was re-suspended in 250  $\mu\text{l}$  buffer P1 containing RNase A, followed by the addition of 250  $\mu\text{l}$  buffer P2 (NaOH/SDS) for lysing bacteria under alkaline conditions which denatures plasmid DNA. The lysate was neutralised by adding 350  $\mu\text{l}$  of buffer N3. Samples were centrifuged for 10 min at  $8,000 \times g$  and the supernatant transferred to a QIAprep spin column. Samples were centrifuged for 1 min at  $8,000 \times g$  and the flow through discarded. Discarding flow through at every stage, 500  $\mu\text{l}$  buffer PB was added to bind DNA to the membrane and centrifuged for one minute, and then the DNA was washed with 750  $\mu\text{l}$  buffer PE and centrifuged twice to remove residual ethanol. To elute DNA from the QIAprep spin column, 100  $\mu\text{l}$  EB buffer or sterile distilled water (SDW) was added and the columns left to stand for 1 min. The column was centrifuged for 1 min at  $8,000 \times g$  to elute DNA from the column membrane.

Phagemid were then digested with restriction enzymes (*NcoI* and *EcoRI* were used for pG8SAET, and *PstI* and *XhoI* were used for pG8H6) to release the insert, and the sample was gel electrophoresed to check for a visual representation of insert frequency.

### **(f) Phagemid library conversion**

All cells from successful transformations were combined (If cells had previously been in storage at 4°C, they were resuscitated by growing at 37°C with shaking (200 rpm) for 1 hour before combining). The pooled cells were decanted into 50ml Falcon tubes and centrifuged in a chilled rotor (4°C) at 2,500 x g for 10 min before resuspension in 10 ml sterile NB2. After resuspension, all cells were combined and helper phage R408 (Promega, Southampton, UK) added to a multiplicity of infection (MOI) of 20. After static incubation at 37°C for 30 min to allow phage attachment, the infection was added to 100 ml sterile NB2 in a 1L flask containing 100 µg/ml Ampicillin, and grown overnight at 37°C, 200 rpm. Cells were separated by centrifugation in a Sorvall Ultracentrifuge RC6+ using SLA-1500 rotor at 10,000 x g for 15 min, 4°C, and the supernatant decanted into sterile Sorvall bottles containing 40% PEG/2.5M NaCl. After precipitation at 4°C overnight, phage were removed by high speed centrifugation in 40 ml Sorvall bottles in an SS-34 rotor at 15,000 x g for 20 min, 4°C, and the pellet resuspended in 1 ml TE Glycerol. The tubes were rinsed with a further 1 ml TE Glycerol, and this additional volume added to the first. A 20 µl aliquot of the library was removed for enumeration and the rest stored at -80°C until required for Panning.

### **Enumeration of recombinants**

Ten microlitres (µl) of 10<sup>0</sup> to 10<sup>-8</sup> phage dilutions in 20 mM Tris HCl were added to 200 µl of log phase *E. coli* TG1 and left to attach statically at 37°C for 30 min. After incubation, 100 µl of each infection was plated onto NB2 plates containing 100 µg/ml ampicillin and grown overnight at 37°C. After 16-24 hours growth, colonies were counted and the titre calculated.

### **(g) Biopanning**

All centrifugation steps in the Biopanning protocol were carried out using a Sorvall Ultracentrifuge RC6+ (Beckmann). The rotors and centrifugation speeds are specified at each stage.

The ligands IgA, FN and BSA (all from Sigma) were diluted into 2.4 ml 100 mM sodium carbonate buffer (pH 9.4) and added to a Nunc Immuno Tube (Gibco), then sealed with Parafilm™ and rolled overnight at room temperature (22°C). The tubes were washed with 3 changes of PBS-T (PBS + 0.05% Tween 20) by adding 4 ml PBS-T, sealing tube with Parafilm and rolling at room temperature for 10 min. Blocking free binding sites on the tubes was carried out by adding 4 ml PBS-T + 2% BSA and rolling for 2 hours at room temperature.



Tubes were washed 4 times with PBS-T as described, and 2 ml titred phage display library added. After rolling for 2 hours, excess phage was decanted and retained at 4°C for use against a different ligand, and the tubes were washed with 6 changes of PBS-T. Bound phage were eluted by adding 1 ml elution buffer (Glycine/HCl pH 2.2) and rolling for 10 min at room temperature. After removal to 500 µl 1M Tris.HCl (pH 7.5) to neutralise the acid, 20 µl eluate was enumerated. The remaining eluted phage (all, or an aliquot, depending on the titre) were added to 5-8 ml log phase *E. coli* TG1 in NB2 + 2% glucose, and left to attach statically for 60 min at 37°C. After attachment, the cells were grown at 37°C for a further 60 min, then helper phage R408 was added to an MOI of 20, and left to attach statically at 37°C for 30 min. The infection was inoculated into 190 ml NB2 containing 100 µg/ml ampicillin and grown overnight at 37°C, 200 rpm. Cells were recovered by centrifugation using an SLA-1500 rotor at 7,000 x g at 4°C for 20 min. Supernatant was decanted into a sterile Sorvall bottle containing 40% PEG/2.5M NaCl and left at 4°C overnight to precipitate the phage. Phage were recovered by centrifugation using an SS-34 rotor at 15,000 x g at 4°C for 20 min, and resuspended in 500 µl PBS. The recovered phage were enumerated then used to begin the second round of panning (starting at the beginning) to further amplify the binding protein sequences.

For enumeration of phage titre the same protocol was used as detailed for phagemid library conversion with the exception of using NB2 plates containing 4% Glucose and 100 µg/ml ampicillin.

#### (h) Sequencing

All DNA was sent to the Wolfson Institute for Biomedical Research (WIBR) for sequencing in 15 µl aliquots per reaction at a concentration of 100 ng/µl. See [www.ucl.ac.uk/wibr/services/dna](http://www.ucl.ac.uk/wibr/services/dna)

#### Primers

Primers used in sequencing reactions for both phagemid vectors, TOPO\_TA (16S analysis) and pUC19 are shown in **Table 4**. Custom primers were supplied to WIBR at 2-5pmoles/µl, allowing 6µl per reaction. All primers were ordered from Operon (Operon Biotechnologies GmbH, Cologne, Germany).

<i>Vector</i>	<i>Forward</i>	<i>Reverse</i>
pG8SAET	5'-AGGTACACTTATATCTGG-3'	5'-CCGCTTTTGC GGGATCGTCAC-3'
pG8H6	5'-TTGCCTACGGCAGCCGCTGAA-3'	5'-TGCGGCCCCATTCAGATCCTC-3'
TOPO TA	5'-AGAGTTTGATCMTGGCTCAG-3'	5'-TACCTTGTTACGACTT-3'
pUC19	5'-GTTTTCCAGTCACGAC-3'	5'-GGAAACAGCTATGACCATG-3'

**Table 4 List of Primers for Sequencing**

### **Analysis *in silico***

Sequencing results were returned in text format and as .ab1 files, viewed using Vector NTI software (Invitrogen). Text sequences from forward reaction were pasted into a new word document and the signal sequence deleted (from CCCC of the *Sma*I cloning site). The remaining sequence was pasted into CLUSTALW's multiple sequence alignment program (<http://align.genome.jp/>) along with the end sequence, which begins GGG from the other end of the *Sma*I cloning site. If the sequence did not extend as far as the c-Myc tag, the reverse sequence was entered (reverse complement) into CLUSTALW and the c-Myc tag located from that end. At the c-Myc end, the vector sequence was highlighted from GGG to GAAT then the remaining vector deleted. The entire sequence and up to GAC of the vector was copied (so that the amino acids VQVD appear when in frame), and pasted into either BCM search launcher in EXPASY tools ([www.expasy.ch/tools/](http://www.expasy.ch/tools/)) or another program which allows 6 frame translation of DNA sequences, such as [http://molbiol.ru/eng/scripts/01\\_13.html](http://molbiol.ru/eng/scripts/01_13.html).

With the reading frames translated, +1, +2 and +3 were checked for the amino acids VQVD at the c-Myc end, giving the reading frame in which the protein is translated. The amino acid sequence was pasted into a new word document, so that for each insert sequence both a Word file containing the trimmed DNA sequence of the insert, and a Word file containing the translated amino acid sequence was generated. Both the amino acid and DNA sequences were used to search for homology to other DNA or proteins using the Basic Local Alignment Search Tool (BLAST) algorithm, provided free online by the National Centre for Biotechnology Information (NCBI), found at <http://www.ncbi.nlm.nih.gov/blast/>. The two main searches used were blastp (protein-protein) and tblastx (translated query vs. translated database).

### **(i) Antibody Screening**

#### **Isolation of phage supernatant**

Three hundred colonies from 3<sup>rd</sup> round panning eluate titre plates were transferred individually into 3ml NB2, infected with helper phage R408, then grown overnight at 37°C with shaking. Phage supernatant was recovered by centrifugation at 2,500 x g in an Microcentrifuge benchtop centrifuge for 5 min. Phage supernatant was removed from the bacterial pellet and transferred into sterile microcentrifuge tubes, then stored at 4°C until needed. Bacterial pellets were stored at -20°C until needed for analysis of the insert by miniprep.

#### **Antibody screening against anti Poly-His**

Five microlitres of each phage supernatant was spotted onto nitrocellulose paper and allowed to dry at room temperature. The nitrocellulose strips were immersed in 20 ml 5% (w/v) skimmed milk powder in Tris Buffered Saline (TBS) and blocked with gentle agitation for 60 min, after which time they were washed with three rinses of TTBS (TBS plus 0.05% Tween-20

v/v), 5 min per wash. After washing, the filters were immersed in 20 ml Anti Poly-His (diluted 1:2000 in TBS), and incubated with gentle agitation for 60-120 min. Following a further three washes in TTBS, 5 min per wash, the filter was treated with BCIP/NBT substrate solution for 10 min at room temperature, or until colour development occurred.

### **Antibody screening against anti c-Myc and alkaline phosphatase conjugate**

5µl of each phage supernatant was spotted onto nitrocellulose paper and allowed to dry at room temperature. The nitrocellulose strips were immersed in 20 ml 5% skimmed milk powder in Tris Buffered Saline (TBS) and blocked with gentle agitation for 60 min, after which time they were washed with three rinses of TTBS (TBS plus 0.05% Tween-20), 5 min per wash. After washing, the filters were immersed in 10 ml c-Myc (diluted 1:5000 in TBS), and incubated with gentle agitation for 60-120 min. Following a further three washes in TTBS, 5 min per wash, the filter was incubated with the secondary antibody, IgG Alkaline Phosphatase Conjugate (diluted 1:30,000), and incubated for a further 60 min with gentle agitation. After 3 standard washes in TTBS, the filter was treated with BCIP/NBT substrate solution for 10 min at room temperature, or until a colour change was observed.

### **(j) SDS-PAGE and Western Blotting**

#### **Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis of Proteins**

Proteins were resolved using SDS-PAGE. SDS solubilises proteins and coats them in a negative charge, resulting in electrophoretic mobility being based purely on size. Samples were prepared for SDS-PAGE analysis by adding 10 µl of 2X SDS gel loading buffer (50 mM Tris Cl [pH 6.8], 100 mM dithiothreitol, 2% w/v SDS, 0.1% BPB), 10% glycerol) to a 10 µl sample and boiling for 5 min at 95°C. Gels were made fresh on the day of the experiment and consisted of a resolving gel containing up to 18 % acrylamide for resolving proteins between 6 and 50 kDa (1.5M Tris pH 8.8, 10% SDS, 10% ammonium persulphate, 8 µl TEMED). The addition of 10 % ammonium persulphate (APS) and Tetramethylethylenediamine (TEMED) catalysed the polymerisation of the gels. The resolving gel mixture was cast by pipetting into the space between the glass plates. A layer of ddH<sub>2</sub>O was carefully placed on top to avoid curving of the gel surface. Once polymerisation was complete, the water layer was removed and any unpolymerized acrylamide was washed away. A stacking gel containing 5 % acrylamide was cast on top with a comb added to form loading wells (contents as in resolving gel except Tris used in 1M of pH6.8). Following polymerisation of the stacking gel, the comb was removed and wells were washed with tris-glycine running buffer to remove any unpolymerised acrylamide. The gel apparatus was assembled according to the manufacturer's instructions and Tris-glycine electrophoresis buffer (25 mM Tris Base, 250 mM Glycine, 0.1% SDS) added to the top and bottom reservoirs. Following the addition of 15µl of each sample (heated at 100°C for 3 min), and a broad range protein marker (2-212 kDa, NEB), empty wells were filled with 1

x SDS gel-loading buffer. SDS-PAGE was performed using Biorad Mini-PROTEAN 3 system and a continuous voltage of 8 V/cm was applied to the gel for 40 min, or until BPB reached the bottom of the resolving gel. The gel was then carefully removed and stained.

To stain, the gel was immersed in 5 volumes of Coomassie Brilliant Blue (Severn Biotech Ltd, UK) and slowly agitated on a rocking platform for 4 hours at room temperature. Excess stain was then removed from the gel by soaking in destain (500 ml methanol; 400 ml dH<sub>2</sub>O; 100 ml glacial acetic acid), on a rocking platform for at least 4 hours at room temperature. Destain solution was changed whenever necessary during this period. Gel staining with Coomassie did not successfully highlight protein bands, either due to the stain being insufficiently sensitive or because of an over-vigorous destain step. Sypro Ruby Red (Biorad, USA) was subsequently tried as it is more sensitive, and this involved staining overnight on a rocking platform. Stained gels were photographed using Gene genius Bio imaging system and Genesnap software (version 6.05) from Syngene.

### **(k) pET Vector Expression**

#### **Individual primer design**

In order to amplify DNA from the final 18 clones, individual primers were required since each insert had a unique reading frame and were designed to include *NdeI* and *EcoRI* restriction sites to enable seamless cloning into pET. The individual primers are detailed in **Table 5**.

#### **Individual clone amplification**

Polymerase Chain Reaction (PCR) was carried out to amplify insert sequences from phagemid vector for individual study. PCR was performed in 0.5µl Microcentrifuge tubes in a total reaction volume of 50 µl. Reactions comprised 1 µl template DNA, 2 µl of 5 µM each oligonucleotide primer and 45 µl PCR supermix (Invitrogen, UK). Reactions were prepared on ice. The PCR was performed for 30 cycles of 95°C, 5 min (1st cycle only) 95°C, 30 s; 54 - 68°C, 30 s; 70°C, 60 seconds; 70°C, 5 min (last cycle only), using a Techne TC-512 gradient thermal cycler. The PCR product was quality and size checked using 1% (w/v) agarose gel electrophoresis.

#### **TOPO TA cloning**

Ten microlitres of each PCR product were cloned into the TA cloning vector pCR4-TOPO (Invitrogen, UK) according to the manufacturer's instructions. The ligation mixture was transformed into JM 107 chemically competent *Escherichia coli*. Transconjugants were detected on LB agar supplemented with 50 µg / ml kanamycin. Successful clones were harvested by Qiagen miniprep, the inserts excised using restriction enzymes *NdeI* and *XhoI* and

the digests run on a 0.8% agarose gel. Inserts were clearly present as distinct bands and were recovered using Qiagen Gel Extraction Kit according to manufacturers instructions.

<i>Universal Primer</i>	5' – CCG TTT GAT CTC GAG GTC GAC C – 3'
<i>Clone Number</i>	<i>Individual Primer</i>
1	5' – CAT ATG CCG CTG TTG GTG CTC CTG G – 3'
2	5' – CAT ATG CCG CTC TAT CGC AGG GAA TG – 3'
11	5' – CAT ATG CTA AAT CTT TTG GAA CTG AAA GC – 3'
16	5' – CAT ATG CAA CGT CAC GCT ATA GAA CTA G – 3'
17	5' – CAT ATG TAT ATT ATT TCT GCT AGC CTC TAT G – 3'
19	5' – CAT ATG AAA TCC TGG AAC TTC CAG GAC G – 3'
20	5' – CAT ATG CTG GGC GTG GAG AAC CTG TAC G – 3'
22	5' – CAT ATG GCA GAG GCA GGA CAT ATC GAG G – 3'
27	5' – CAT ATG GAG GGA ACT CCT CCA GAA AAT AG – 3'
36	5' – CAT ATG CTT ATT TTT CTT TTG GGA TTA G – 3'
39	5' – CAT ATG GTG ATG GCT GTT CAC CGC ATG – 3'
42	5' – CAT ATG CTT TAT CTC ATG ACT GCA AAA TC – 3'
44	5' – CAT ATG CCG CAC ACG GTG TCA GCG TCC G – 3'
52	5' – CAT ATG ATC GGT ATC GTT AAA GGG GGG – 3'
58	5' – CAT ATG TCA ACT TTA ATG ATA GGT ATG GAA A – 3'
59	5' – CAT ATG CTT ATT TTA GGT AGA ATA AAC TAT – 3'
60	5' – CAT ATG GTG ATT CTT GGC TTG ATT TTC TTT – 3'

**Table 5 Primers for pET vector expression**

### **pET Vector Expression**

Inserts from gel extraction were ligated into pET21 vector pET21b, pre-digested with *NdeI* and *XhoI*, using T4 DNA ligase (NEB). No successful ligations were ever made so pET vector expression was not taken further.

### **(I) Adhesion assays**

In order to test the binding specificity of each individual fusion protein for the panning ligand that first identified it, individual populations of pure phage were required. To do this, 1 µl of phagemid containing the insert of interest was transformed into *E. coli*. Helper phage R408 was added to an MOI of 20 and the infection inoculated into 190 ml NB2 containing 100

µg/ml ampicillin and grown overnight at 37°C, 200 rpm. Cells were removed by centrifugation using an SLA-1500 rotor at 7,000 x g at 4°C for 20 min. Phage supernatant was decanted into a sterile Sorvall bottle containing 40% PEG/2.5M NaCl and left at 4°C overnight to precipitate. Phage were recovered by centrifugation using an SS-34 rotor at 15,000 x g at 4°C for 20 min, and resuspended in 500 µl TB containing 20% glycerol. Recovered phage were enumerated using the same protocol as used for phagemid library conversion and panning eluate reenumeration, using dilutions of phage supernatant to infect 100 µl of log phase *E. coli*.

Nunc Immunoplates coated with 100 µl 0.1 mg/ml ligand were incubated at 4°C overnight. Following removal of the ligand, wells were thoroughly washed with 200 µl PBS (x5) and 200 µl PBS-T (x5) to remove all unbound ligand. One hundred microlitres of the previously amplified and recovered phage population at  $1 \times 10^9$  were added to each well and left to bind to the immobilized ligand for 90 min. Following excess phage removal, the washing steps were repeated; 200 µl PBS (x5) and 200 µl PBS-T (x5) to remove all unbound phage. Bound phage were eluted in 100 µl elution buffer (glycine/HCl pH 2.2) and neutralised with 50 µl Tris.HCl (pH 7.5). Dilutions of eluted phage were used to infect log phase *E. coli* and enumerated as previously described. Higher titres were expected where fusion proteins showed stronger binding affinity to the panning ligands which led to their initial identification.

### **(m) 16S rRNA Gene Diversity Analysis**

#### **16S rRNA Gene Amplification**

Polymerase Chain Reaction (PCR) was carried out using universal primers 27f-CM (5' – AGA GTT TGA TCM TGG CTC AG – 3') and 1492r (5' – TAC CTT GTT ACG ACT T – 3'). PCR was performed in 0.5µl Microcentrifuge tubes in a total reaction volume of 50 µl. Reactions comprised 1 µl template DNA, 2 µl of 5 µM each oligonucleotide primer and 45 µl PCR supermix. All reactions were prepared on ice.

The PCR was performed in the initial analysis for 30 cycles of 95°C, 5 min (1st cycle only) 95°C, 45 s; 46 -54°C, 60 s; 72°C, 1.5 min; 72°C 15 min (last cycle only), using a Techne TC-512 gradient thermal cycler. The positive control samples contained all the PCR reagents together with *Escherichia coli* DNA. The PCR product was quality and size checked using 1% (w/v) agarose gel electrophoresis. In the full 16S analysis, the PCR was performed using the same settings but annealing at 46°C and for 10 cycles only.

#### **Cloning of 16S rRNA gene amplified DNA**

Ten microlitres of each PCR product was cloned into the TA cloning vector pCR4-TOPO (Invitrogen, UK) according to the manufacturer's instructions. Two microlitres of the ligation mixture was transformed into TOP 10 chemically competent *Escherichia coli*. Transconjugants were detected on multiple LB agar plates supplemented with 100 µg/ml

Ampicillin, 10 mM IPTG and 40 µg/ml X-Gal. Following overnight growth at 37°C, 380 white colonies were selected, plasmid miniprep (Qiagen) and 20 µl dispensed into 96-well plates at a concentration of 5 µM for sequencing at the Comparative Genomics Centre, part of University College London.

Purified PCR-amplified 16S rRNA fragments were sequenced using the Big Dye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) using a 3730xl Capillary sequencer, also manufactured by Applied Biosystems. The universal M13 forward (5'– GTT TTC CCA GTC ACG AC –3') and reverse (5'– GGA AAC AGC TAT GAC CAT –3') primers were used. Each reaction (well) contained 10 µl volume, made up of 2 µl DNA, 0.5 µl Big Dye reagent, 2 µl of 5X buffer supplied with the kit, 0.32 µl of 5 µM primer, and the rest made up with water. The cycle sequencing program was 95 °C for 60s (1<sup>st</sup> cycle only), then 25 cycles of denaturation at 95 °C for 10s; annealing at 50 °C for 5s and extension at 60 °C for 4 min.

### 16S rRNA data analysis

Once sequenced, the 16S forward and reverse files containing ~900 bp of sequence, were aligned using BioEdit Sequence Alignment Editor v7.0.9 in FASTA format. The sequences were then aligned using the Greengenes database (DeSantis *et al.*, 2006a) using NAST (Nearest Alignment Space Termination) (Desantis *et al.*, 2006b), which outputs the MSA (Multiple Sequence Alignment) in the standard format of 7682 characters per sequence, allowing similar loci to be located in similar positions in subsequent batches. In this analysis, 4 out of 380 sequences did not meet the match requirements, being either less than the minimum length of 1250 bp, or sharing less than 75% identity to the template sequence. These sequences were removed from further analysis.

Due to the presence of genomic data from different origins containing the conserved 16S rRNA gene, amplification by PCR is known to introduce hybrid molecules which can distort the results of a diversity analysis (Liesack *et al.*, 1991). The percentage of chimeric sequences in this 16S rRNA gene analysis was checked using the greengenes Bellerophon (version 3) server (Huber *et al.*, 2004). In other studies (Kazor *et al.*, 2003; Aas *et al.*, 2005) the number of chimeric sequences was between 1 and 15%. This analysis located 42 chimeric sequences (11%), which were below the 97% threshold BLAST similarity and less than 1250 bp match to the Core Set of sequences. The species identification of chimeric sequences was not obtained and these 42 were removed from further study.

Using the NAST aligned sequences, minus chimeras, the remaining 333 sequences were classified using the Simrank interface which finds similarity between query and database in terms of the number of unique 7-mer count present in either query or database. Sequence diversion from near-neighbours was calculated using the DNAML option of DNADIST (PHYLIP package). The reference sequences used for classification were non-chimeric

## Chapter 2: Materials & Methods

(divergence ratio  $<1.10$ ) and taxonomic analysis was conducted using the RDP taxonomic nomenclature.



---

## **Chapter 3**

### **Results and Discussion:**

### **Constructing a Phage Display Library**

---

## Introduction

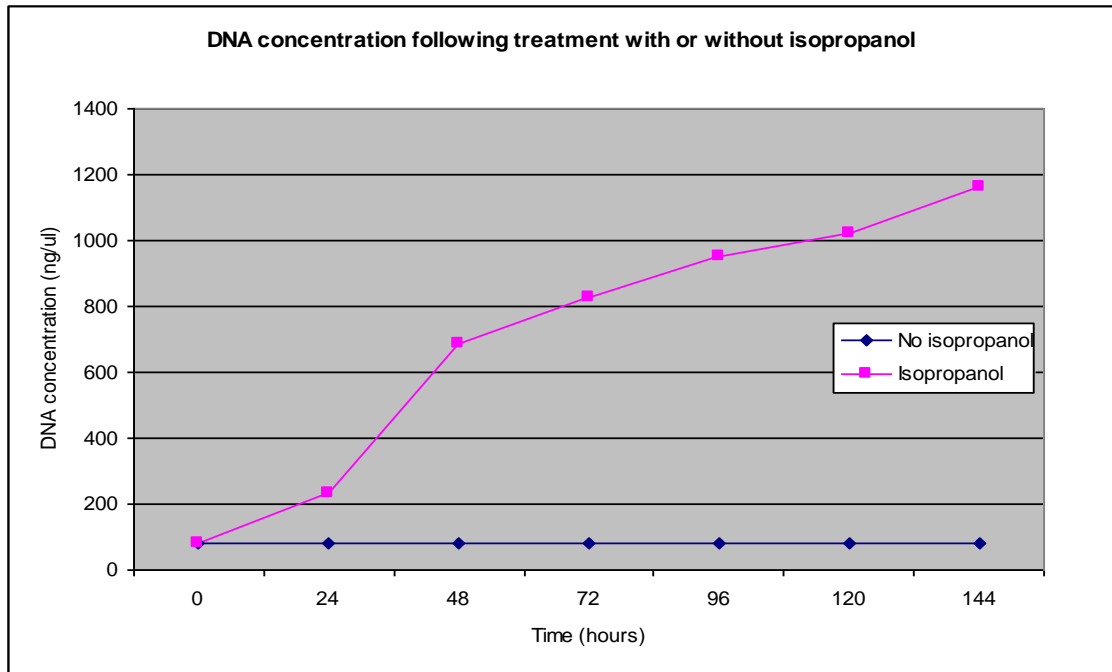
The ligands IgA and Fibronectin (FN) are important components of the human cellular or immune system and as such are known to associate with the commensal bacterial cargo. The aim of this project was to investigate novel binding mechanisms through the combination of metagenomics and phage display. Taking a metagenomic approach allowed the extraction of all genomic material in the samples, and the production of a phage display library containing representative sequences from (theoretically) all bacteria present. This chapter details the steps from DNA extraction to phage display library production.

## DNA sampling and concentrations

Nine volunteers were asked to provide scrapings from the tongue dorsum for this study. The volunteer data sheet is shown in **Appendix 1**. All samples were collected in 5 ml of PBS after vigorous brushing with a sterile toothbrush.

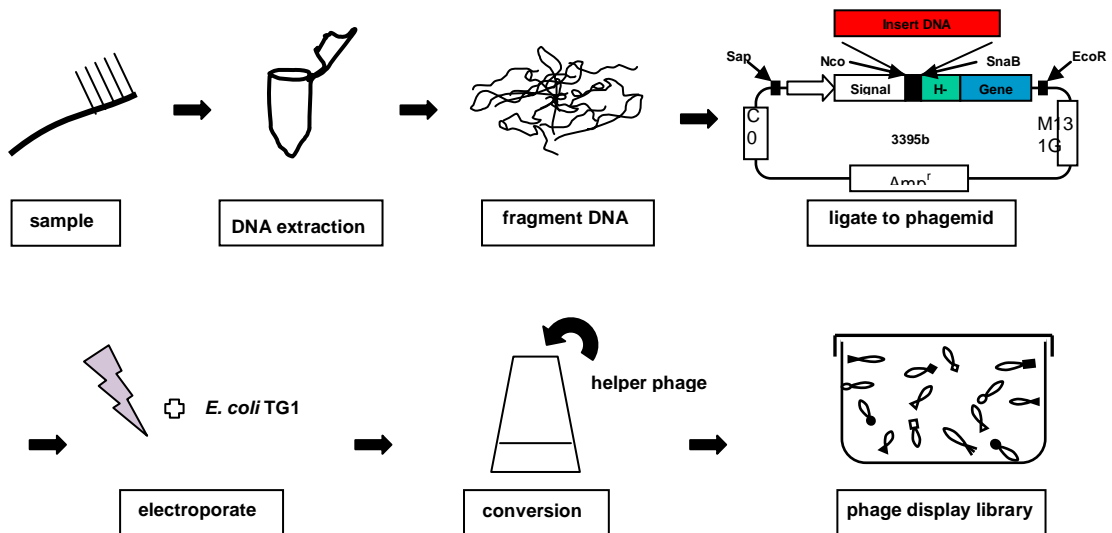
Previous studies mention that resuspending a mixed microbial sample in isopropanol may increase the DNA concentration upon extraction (Torsvik *et al*, 1990). The method used by Torsvik in 1990 to achieve this increase was not described explicitly in the publication; therefore isopropanol treatment of tongue samples was tested over varying amounts of time to assess the increase in DNA concentration (**Figure 1**). Briefly, a fresh 5 ml tongue-scrape sample in PBS from one of the 9 volunteers was centrifuged and the pellet resuspended in 5 ml isopropanol and stored at -20°C for between 1 and 6 days. Following this treatment, DNA was extracted using the CTAB protocol and DNA concentration measured by Nanodrop™

All Nanodrop™ readings were taken in triplicate as the instrument does not provide a completely accurate value. Further, it was realised that the DNA concentration between samples would vary depending on the number of bacterial cells in the sample from day to day, and in order to minimise this effect, samples from one volunteer were used exclusively and vortexed just before use. The mechanism by which isopropanol increases DNA concentration prior to extraction is unclear, however these data clearly demonstrate that an increase did occur. We hypothesised that isopropanol was either penetrating the bacterial cells therefore disrupting DNA/protein interactions, or that it was removing lipid contamination similar to the method discussed by Stadler & Hales, 2002. Whatever the mechanism, following this experiment all subsequent DNA samples were stored in isopropanol for 7 days, as detailed in Chapter 2, page 51.



**Figure 1** DNA concentration following treatment with and without isopropanol. Blue line signifies the DNA concentration with no prior isopropanol treatment, which did not change over 6 days. Pink line illustrates the increase in DNA concentration over 6 days.

Following DNA extraction from all samples using the CTAB protocol, 200  $\mu$ l from samples 1 - 9 were combined and the metagenomic DNA concentration determined by Nanodrop™ to be in the region of 1000 ng/ $\mu$ l. This pooled DNA was used to construct a phagemid library, and a pictorial overview of this is shown in **Figure 2**.

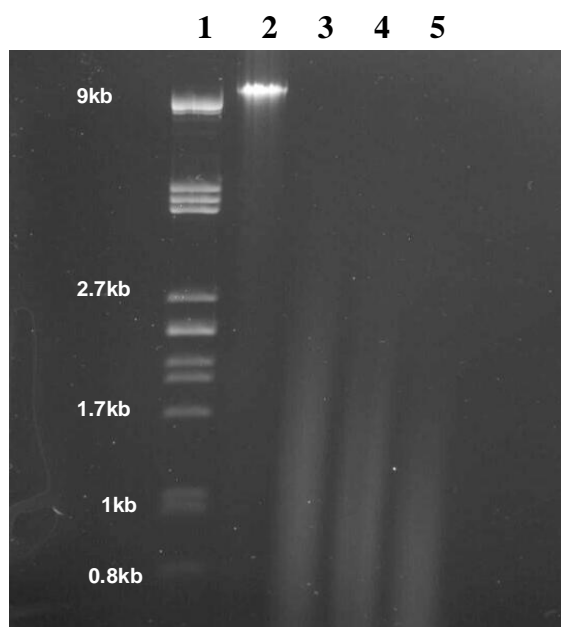


**Figure 2** Production of a phage display library. Following sample extraction, genomic DNA was fragmented and ligated into the phagemid vector pG8H6. Successful ligations were transformed into *E. coli* cells and converted to a phage display library, which was stored at  $-80^{\circ}\text{C}$  for subsequent panning.

**DNA fragmentation**

Metagenomic DNA was fragmented by two different methods to identify the optimal method for library production; sonication, the traditional method of fragmenting metagenomic DNA; and partial restriction digest using three blunt-end restriction enzymes.

Sonication of metagenomic DNA was carried out as described in Chapter 2, page 53. Briefly, DNA was fragmented for 5, 10 and 15 seconds, blunt-ended using dNTPs and DNA polymerase, and the fragments visualised using 1% (w/v) agarose gel electrophoresis (**Figure 3**). Phage display is commonly used with DNA fragments between 400 bp – 1 kb; therefore the optimum sonication time from this gel is clearly 15 seconds as most of the larger DNA fragments at this time point are around 1 kb.



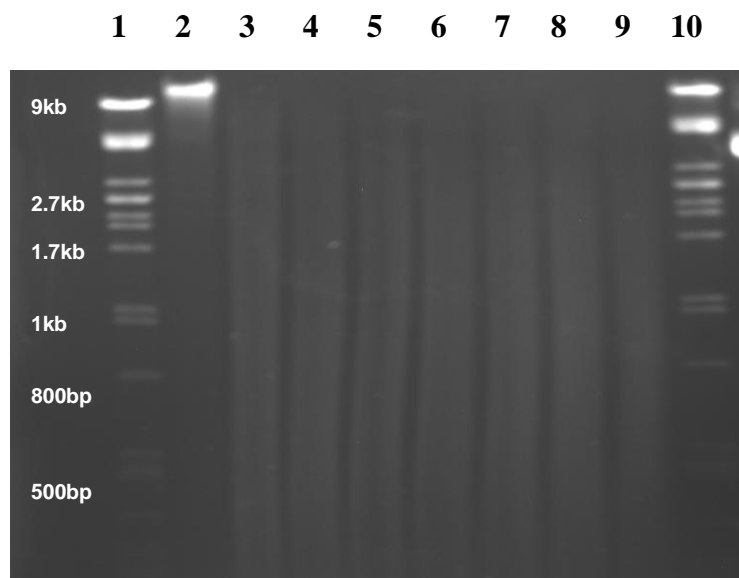
**Figure 3** Sonication of DNA sample in 5 second intervals. (1)  $\lambda$ Pst marker with corresponding sizes, (2) 1  $\mu$ l of unfragmented DNA sample at 1000 ng/ $\mu$ l, (3) DNA after 5 second sonication, (4) DNA after 10 second sonication, (5) DNA after 15 second sonication.

DNA fragmentation by partial digestion was carried out as described in Chapter 2, page 53. Briefly, three blunt-end producing restriction enzymes (*HaeII*, *AluI* and *RsaI*) were added to metagenomic DNA for the purpose of partial digest. Each enzyme digests different DNA bases, some of which occur more frequently than others in an attempt to make the fragmentation as random as possible. Aliquots were removed every 10 minutes from the 20 minute time point onwards (**Figure 4**).

There appears to be no real difference in fragment size range, except that the largest fragments at each time point decrease in size as incubation time increases. For this project,

### Chapter 3: Constructing a Phage Display Library

fragments between 500 bp and 1 kb were thought able to contain enough of a bacterial gene, such that binding affinity for a ligand would be retained, facilitating detection during panning and for further downstream processing. Fragments of 500 – 1500 bp were achieved by 20 minutes, therefore 20 minutes incubation was used for future restriction digests.



**Figure 4 DNA fragmentation by partial digestion with *AluI*, *RsaI* and *HaeIII*** (1)  $\lambda$ Pst DNA marker, (2) 1  $\mu$ l of untreated DNA at 1000 ng/ $\mu$ l, (3) DNA after 20 minutes restriction digest, (4) after 30 minutes digestion, (5) after 40 minutes digestion, (6) after 50 minutes digestion, (7) after 60 minutes digestion, (8) after 70 minutes digestion, (9) after 80 minutes digestion, and (10)  $\lambda$ Pst DNA marker.

#### **Determination of optimum ligation conditions**

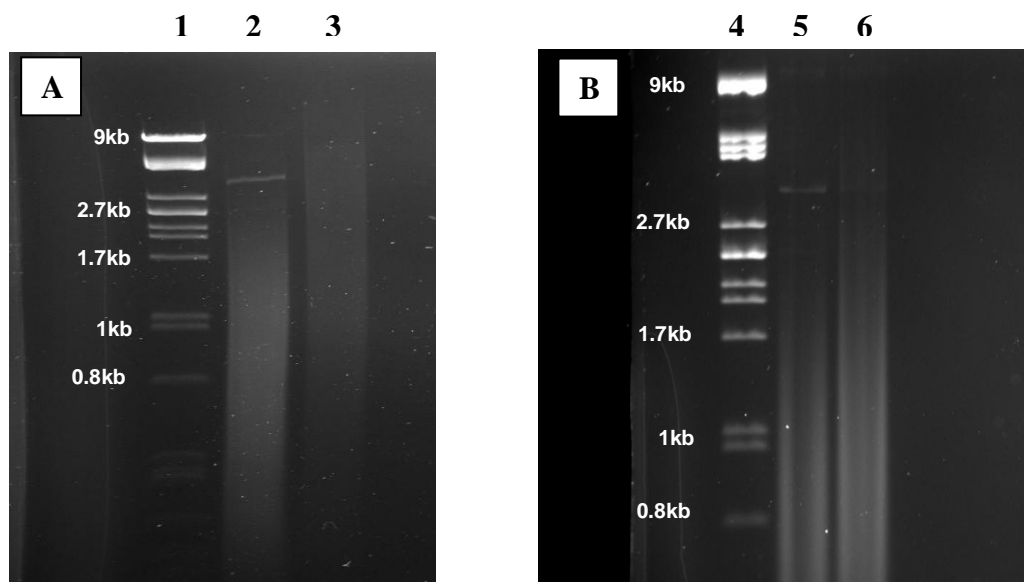
Phagemid vector (pG8H6) DNA, extracted using the Qiagen MidiPrep Kit, described in Chapter 2, page 50, was linearized with restriction enzyme *SmaI* and dephosphorylated to prevent self-ligation. Using sonicated and partially digested metagenomic DNA, different ratios of insert and vector concentration (in ng/ $\mu$ l, from Nanodrop™) were tested to determine the optimum insert: vector ratio in 10  $\mu$ l ligation volumes. Briefly, a variety of ligations were set up using insert and phagemid vector pG8H6 DNA diluted to equal concentrations. Two microlitres of each ligation reaction was then transformed into electrocompetent *E. coli* cells and grown in 3 ml of NB2 broth. Following 2 hours growth, 100  $\mu$ l from each culture was plated on NB2 agar plates containing 100  $\mu$ g/ml ampicillin and the number of colonies following 16-24 hours incubation was counted (**Table 1**).

		<i>Insert (μl)</i>		
		1	2	3
<i>Vector (μl)</i>	1	81	389	1306
	2	103	436	1123
	3	96	237	563

**Table 1** Number of colonies for ligations with various insert: vector ratios

Given that both ratios of two or three parts insert to one part vector produced the greatest number of colonies on ampicillin agar plates, a 3:1 insert to vector ratio was used in all subsequent ligations.

Ligations were prepared in a 20 μl volume. Before adding T4 ligase, 2-3 μl was removed for comparison on an agarose gel (shown in lane 2, **Figure 5A**). After overnight incubation, the ligation success was visualised using 1% agarose gel electrophoresis. Ligation reactions where the vector joined more successfully with DNA fragments occurred more often when using DNA fragmented with blunt-end restriction enzymes (**Figure 5B**).



**Figure 5** Ligation reactions A and B. **Figure A**; (1) λPst DNA marker, (2) control: vector band pG8H6 and smear of metagenomic DNA prior to ligation, (3) ligation reaction following overnight incubation; vector band has disappeared illustrating ligation to the metagenomic DNA and resulting in an overall increase in fragment size. **Figure B**; (4) λPst DNA ladder. Ligation reactions containing pG8H6 phagemid vector and (5) sonicated DNA, and (6) DNA fragmented by 3 blunt-end restriction enzymes.

These ligations also resulted in higher numbers of transformants in the subsequent library, and to those transformants containing a higher percentage of metagenomic DNA inserts (data not shown). This may be due to the extra step required when using sonicated DNA, prior to ligation into a blunt-end vector. After sonication, exposure to shearing forces results in DNA fragments with overhanging ends which must be repaired (with dNTP's and DNA polymerase) to optimise ligation into the blunt-end vector. The efficiency of ligation really relies on a high degree of end repair however the efficiency of this step is only substantiated by successful ligation reactions. Using blunt-end producing restriction enzymes bypassed this repair step completely, meaning that all DNA fragments were blunt-ended and therefore able to successfully ligate into the phagemid vector.

### Transformation

Prior to electrochemical transformation, ligation reactions were purified of excess salt using the Qiagen PCR Cleanup Kit. Chemical transformations were used in the initial stages of phagemid library production, but were substituted for electrochemical transformation when testing was over. While testing and optimising phagemid library production electrocompetent *E. coli* cells (EEC) were made in the laboratory to minimise costs. However, the efficiency of homemade electrocompetent cells can vary from batch to batch and because the highest quality library was required for phagemid conversion, 10 aliquots of electrocompetent *E. coli* cells (EEC) from Invitrogen were used to produce the final library.

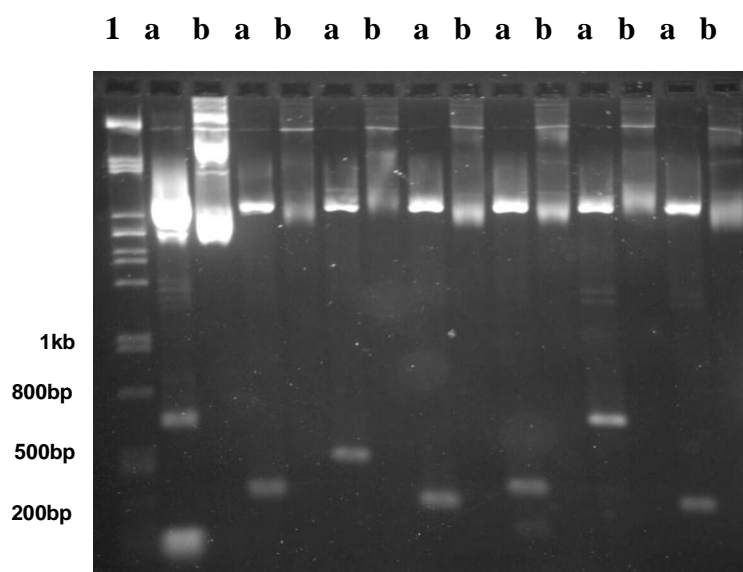
In order to get the highest number of transformants in the phagemid library in the early stages, different ligation volumes were tested in 50  $\mu$ l aliquots of EEC. As shown in **Table 2**, increasing the ligation volume from 1  $\mu$ l to 2  $\mu$ l resulted in an increase in the number of transformants per millilitre from 660 to 1490, an increase which did not continue with additional ligation mixture. When transforming the final phagemid library 2  $\mu$ l of ligation was used in each 50  $\mu$ l aliquot of ECC.

Volume plated out ( $\mu$ l)	<i>Ligation volume (<math>\mu</math>l) in 50<math>\mu</math>l of ECC</i>			
	1	1.5	2	2.5
100 $\mu$ l	66	84	149	145
<b>Transformants/ml</b>	<b>660</b>	<b>840</b>	<b>1490</b>	<b>1450</b>

**Table 2** Electroporation transformation frequency.

**Insert frequency and size following transformation**

Although the phagemid vector had been treated with phosphatase to reduce recircularisation, and because the efficiency of *E. coli* to take up plasmids is not 100%, it was necessary to check the insert frequency before converting the phagemid library to phage display. Briefly, 40 individual colonies from ampicillin agar plates were grown overnight in 3 ml NB2, then the plasmids isolated by plasmid miniprep. Ten microlitre aliquots of plasmid DNA were digested with restriction enzymes *NcoI* and *XhoI* to release the insert, and the digested DNA electrophoresed through a 1% agarose gel. **Figure 6** clearly shows the phagemid library containing a range of inserts between 350 bp and 800 bp in size.



**Figure 6** Phagemid vector pG8H6 containing metagenomic DNA inserts. (1)  $\lambda$ Pst DNA marker, (a) DNA samples after splicing with restriction enzymes *NcoI* and *XhoI*, clearly showing inserts, (b) DNA samples before restriction digest.

A metagenomic DNA insert must be present in the phagemid genome in order that (upon conversion) phage coat protein 8 displays a fusion protein. A high frequency of metagenomic DNA inserts at this stage was necessary to achieve a high diversity of fusion proteins following conversion. It was thought that more than 80% phagemid containing inserts would generate a library of adequate complexity for panning. Prior to conversion, this library contained inserts in 93% of the phagemid tested.

Having established the complexity of the phagemid library, it was important to pinpoint the origin of that DNA, since a phagemid library containing more human DNA than bacterial would be of limited interest for this project. To do this, 40 colonies from a phagemid library spread on ampicillin agar plates were plasmid-extracted and sequenced and the results

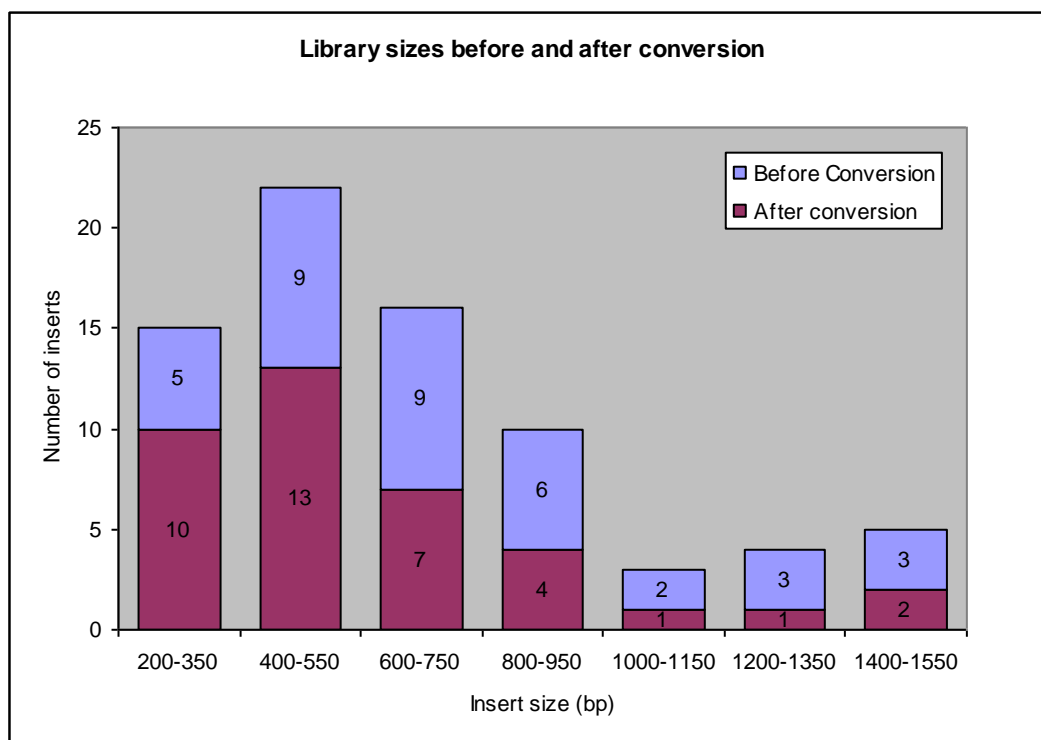


analysed by BLAST search. From 40 colonies, five contained human DNA (12%) and the remainder contained bacterial DNA of mixed homology to the database (data not shown). It is prudent to mention here that none of the inserts tested were identical, which means that the use of restriction enzymes to fragment metagenomic DNA produces fewer than one clone in 40 that is repeated.

### Conversion of phagemid to phage display

Initially, test phagemid libraries were converted to phage display to optimise the process before converting the final library. Briefly, cells from successful transformations were combined, centrifuged and resuspended in sterile NB2 broth. Helper phage R408 were added to an MOI of 20 and incubated statically at 37°C, then incubated overnight with shaking in 100 ml NB2 broth containing ampicillin. Following centrifugation to remove *E. coli* cells, the phage supernatant was PEG precipitated overnight then recovered by high speed centrifugation.

Because it was important to maintain large inserts for downstream expression, the sizes of 50 were checked following conversion by excising inserts from miniprep DNA using restriction enzymes *NcoI* and *XhoI*, and visualised using 1% (w/v) agarose gel electrophoresis. Metagenomic DNA insert sizes are shown in **Figure 7** before and after library conversion.



**Figure 7** Library insert size before and after conversion. The converted library (maroon bars) generally contains smaller inserts than the phagemid library prior to conversion (lilac bars).

## Chapter 3: Constructing a Phage Display Library

Prior to conversion, the test library contained 93% inserts (37/40) at an average of 685 bp in size, and contained  $7 \times 10^4$  transformants (theoretically  $6.5 \times 10^4$  inserts) so was suitable for conversion to a phage display library. Following conversion, the number of inserts remained stable at 38/40 however, the average insert size decreased from 685 bp to 567 bp.

Insert frequency did not drop following conversion, perhaps indicating that either phage are better at expressing smaller inserts, or smaller phagemid package more quickly resulting in more being made per cell, and this numerical dominance means they are seen more often when analysing individual colonies.

### PEG precipitation

The protocol for conversion described in Chapter 2, page 56 is a revised version of the original which was used for the test conversions, and contains a TE glycerol stage which was later removed. The purpose of this stage in the protocol was as a mid-way point in reducing the total volume of PEG precipitated supernatant from 100 ml to 50 ml, before final recovery in 1 ml, however it was felt that phage were being lost between these points. This conversion process was tested throughout to monitor phagemid losses at each stage (**Table 3**).

The titre of the practice library was  $1.8 \times 10^6$  CFU. Stage 2 involved centrifuging resuspended phage in 50 ml TE glycerol only to centrifuge and resuspend again in 1 ml. The titre of Stage 2 supernatant was low at  $5.2 \times 10^4$  which meant that phage losses were not high enough to impact the final library size, however it was decided that this intermediate step was not required to concentrate the phage library so it was removed in all subsequent conversions.

### Additional rinse

Following conversion of the test library, it was eluted in 1 ml TE glycerol. A phage display library should be at least  $1 \times 10^{10}$  CFU/ml for panning (Mullen *et al*, 2006) so it was important that the phage display library titre was as high as possible before amplification.

<i>Stage</i>	<i>1</i>	<i>2</i>	<i>3</i>
<i>Titre (total)</i>	$5 \times 10^7$	$3 \times 10^6$	$4.85 \times 10^7$

**Table 3** Titres of 3 stages during PEG precipitation to account for phage losses. Stage 1 is the titre of the whole 100 ml of raw supernatant following bacterial cell removal. Stage 2 is the titre of phagemid present in the rinse stage to gauge phage losses, and stage 3 is the final library following concentration to 1 ml, without the addition of the rinse.

However, due to the viscosity of PEG it was difficult to get all the elution out in 1 ml. Therefore, an additional 1 ml rinse of TE glycerol was added following removal of the first 1 ml, to check how many phage remained in the centrifuge tube after one rinse. In the test library the first 1 ml titre was  $5 \times 10^7$  and the 1 ml rinse titre was  $3 \times 10^6$ . Clearly, the additional rinse did recover a significant number of phage so it was decided to keep the additional rinse of 1 ml in every subsequent phagemid conversion.

### Final Library Considerations

The quality of electrocompetent cells used to transform ligated DNA is perhaps the most important factor in producing large libraries according to Woiwode *et al.*, 2003. For this project, 10 aliquots of Invitrogen TG1 electrocompetent cells were bought to produce the final phagemid library before converting to phage, using the optimised conversion process. Before conversion, optimal ligation volume was tested in duplicate transformations with 1  $\mu$ l, 1.5  $\mu$ l and 2  $\mu$ l (Table 4).

	<i>Ligation volume (<math>\mu</math>l) in 50<math>\mu</math>l of electrocompetent cells</i>		
	<i>1<math>\mu</math>l</i>	<i>1.5<math>\mu</math>l</i>	<i>2<math>\mu</math>l</i>
<i>Transformants/ml</i>	<b><math>6.2 \times 10^4</math></b>	<b><math>6.7 \times 10^4</math></b>	<b><math>7 \times 10^4</math></b>

**Table 4** Electroporation efficiency using Invitrogen electrocompetent cells. Results shown are the number of transformants per millilitre, where each transformation was grown in 5 ml NB2.

The number of transformants per millilitre increases from an average of  $6.7 \times 10^4$  to  $7 \times 10^4$  as ligation volume is increased from 1.5  $\mu$ l to 2  $\mu$ l, whereas between 1  $\mu$ l and 1.5  $\mu$ l the number of transformants per ml increased from  $6.2 \times 10^4$  to  $6.7 \times 10^4$ , a greater increase in transformants without a large increase in the ligation volume required. For this reason, 1.5  $\mu$ l of ligation was used in each transformation, which were then incubated in 5 ml nutrient broth and plated onto selective media to check the number of transformants in the library. The final phagemid library of 50 ml contained  $3.5 \times 10^6$  transformants which, at 93% insert frequency, meant the library contained  $3.25 \times 10^6$  inserts. Multiplying the average insert size following conversion (567 bp) by the theoretical number of inserts in the library ( $3.25 \times 10^6$ ) means that the phage display library could contain in the region of 1,842 Mb of DNA (1.8 Gb). If the average bacterial genome is ~5 Mb in size, that would imply that the phage display library could contain in the region of 369 average genomes, providing good coverage of the human

tongue microbiota, thought to contain between 12 and 27 genomes per person (Kazor *et al.*, 2003).

Other groups have produced phage display libraries with the fragmented genomic DNA of individual bacteria, ranging from  $9.2 \times 10^6$  transformants (Jacobsson & Frykberg, 1995) to  $9 \times 10^7$  (Williams *et al.*, 2002). This phagemid library was felt to be comparable in number to previous studies so it was converted to phage. Following phage recovery by high speed centrifugation, two rinses of 1 ml TE glycerol ensured near total recovery of the phage display library at a final titre of  $1.7 \times 10^{11}$  CFU. This titre is comparable with that of Jacobsson & Frykberg who, in their 1995 and 1996 papers, achieved library titres of  $2.6 \times 10^{10}$  and  $2 \times 10^{10}$  respectively.

#### Discussion

Metagenomics involves the extraction, cloning and analysis of the entire genetic complement of a mixed microbial habitat which can be used for diversity analysis, or functional assessments of microbial life within that environment. Making sure the DNA containing this diversity remains complete and unbiased throughout library construction is a difficult task, and as such every method for the assessment of metagenomic library diversity has inherent drawbacks.

Previous metagenomic studies have enlisted a variety of methods for interrogating microbial samples. Stable Isotope Probing (SIP) (Radajewski *et al.*, 2003) is particularly useful for accessing metabolically active organisms; however this method can be limited by incomplete labelling. DNA microarrays allow high-throughput robotic screening for the target of multiple gene products (Wu *et al.*, 2001) and, although this began as an expensive option, as it becomes more heavily used it is becoming cheaper and therefore more accessible. Shotgun sequencing of 16S rRNA, used in a landmark paper by Tyson *et al.* in 2004, allowed complete sequence closure of a simple bacterial community. Shotgun sequencing was also used to sequence viruses by Brietbart *et al.*, 2003. Pyrosequencing, a next-generation sequencing technology, was developed by Margulies *et al.* in 2005 as a high-throughput alternative to capillary sequencing which uses emulsion PCR to amplify individual DNA strands coating hundreds of thousands of beads. The individual beads are separated onto individual fibre optic strands and sequenced, each 7 hour run giving around 100 Mb of sequence data in 250bp chunks per sample (Mardis *et al.*, 2007). This technology was used by Cox-Foster, D.L. (2007) in a study of honey bee colony collapse disorder, and is now becoming a more convenient method of bulk sequencing of environmental metagenomes (Margulies *et al.*, 2005).

Subtractive cloning, PCR, fluorescent in situ hybridization (Harmsen *et al.*, 2002), terminal restriction fragment length polymorphism (T-RFLP) (Nagashima *et al.*, 2003),

membrane assays (Matsuki *et al*, 2002), cosmid (Courtois *et al*, 2003) and fosmid libraries (Nesbo *et al*, 2005) are all alternative methods of interrogating bacterial samples, however none of these alternatives present the opportunity to functionally screen the (partially unknown) proteome in the same way as phage display does. Although the process of arriving at a phage display library suitable for panning took around 2 years, the information contained within it, and its functional screening capacity, may shed more light on the binding proteins of tongue bacteria and the variety of interactions going on in this area.

### **DNA Extraction Method**

When extracting metagenomic DNA it is important to maintain two factors as much as possible: sample diversity and DNA concentration. Options to increase the amount of DNA in the samples, for example sample cultivation in saliva (Foster & Kolenbrander, 2004) were unsuitable for this project since this could alter and reduce sample diversity as culturable bacteria thrive at the expense of others (Schmeisser *et al*, 2007). Daniel, 2005, agreed that enrichment steps like this can have a negative impact on sample diversity but could be useful when particularly high quality DNA is needed or when carrying out SIP analysis on metabolically active community members.

Various methods exist for DNA extraction from an environmental sample and although it generally produces DNA of high quality and purity, the CTAB protocol was chosen since it was rapid and had been used previously to successfully extract high quality metagenomic DNA. CTAB is a direct method of extraction, known to introduce less bias than indirect methods, such as those involving prior cultivation or cell separation (Courtois *et al*, 2003; Kauffmann, 2004).

When isolating DNA from environmental samples for metagenomic studies, Schmeisser *et al* (2007) pinpointed 3 issues which should be taken into account:

1. DNA should come from the broadest host range possible and must represent the original microbial community. To achieve this in this project, samples from 9 volunteers were combined to create a more diverse library than using individual samples.
2. Unintentional mechanical shearing of genomic DNA should be avoided as much as possible; although this has more serious implications for large fragment library construction rather than phage display library construction.
3. DNA should be free from contamination which could interfere with downstream processing such as restriction, ligation and transformation. One of the main benefits of using the CTAB protocol was that it produced high concentration, high purity DNA which was ready for use immediately (Bailey, 1995).

### Chapter 3: Constructing a Phage Display Library

Clearly, the eventual success of metagenomic library production and screening depends on a combination of factors: sample diversity and composition; collection, extraction and storage of the sample; the host and vector systems used for cloning and expression; and the screening strategy itself (Daniels, 2005). To get the highest concentration of DNA possible from the tongue bacterial samples, cells were resuspended in isopropanol and stored at -20°C for one week which increased the DNA concentration of the final extracted DNA. Although isopropanol treatment was used by Torsvik (1990), the exact mechanism by which isopropanol treatment increases DNA concentration is not entirely clear. It is possible that the solvent acts to disrupt the cell membrane, lysing the cells and then separating lipids and nucleic acids for subsequent purification. With environments such as soil, DNA extraction techniques must avoid concentrating matrix compounds from the soil itself (Daniels, 2005) so isopropanol treatment may not be appropriate for all environmental samples. In the oral cavity, one of the main issues is human DNA contamination which, if not removed before DNA extraction from the environmental sample, would continue to appear in the phagemid and phage display libraries. In order to try to reduce the presence of human DNA prior to DNA extraction, a crude method of repeatedly freeze-thawing the samples disrupted the delicate osmotic balance of the human cells, exposing the DNA to bacterial DNases prior to their own lysis. This 'freeze-thaw' method was chosen over an alternative but tricky cell-separation, which would have removed human cells but also, potentially, the bacterial cells associated with them, potentially resulting in a reduction in sample diversity. Human DNA concentration following the freeze-thaw cycle method was checked by end-sequencing 40 shotgun clones, which identified <10% human DNA (results not shown). This number is low enough to merit library construction with this DNA, and it was not thought necessary to extract DNA from samples which had not been through the freeze-thaw cycles.

Ultimately, diversity of a metagenomic DNA sample depends on the bacterial makeup of the original sample. The CTAB protocol was chosen without knowing whether this method was equally suitable for the various organisms present on the tongue surface (Archaea, G+ve and G-ve bacteria). It is likely that this method has resulted in unequal lysis of cells, which is probably unavoidable given the variety of bacterial cell wall architecture. Some microbes, such as *Mycoplasma* spp., have very delicate cell walls which are likely to lyse in the early stages of extraction. Gram positive bacterial cells like *Arthrobacter* and *Rhodococcus* can also be inefficiently lysed as their cell wall architecture makes them more resistant to lysis (Kauffmann, 2004). Rapid cell lysis may have resulted in the release of DNase which could damage other DNA in the sample, and also the early release of genomic DNA may have resulted in excessive mechanical shearing as the protocol progressed. Taking these points into consideration, extracting metagenomic DNA is clearly likely to introduce bias towards bacterial cells that are in the middle of the lysis spectrum. It is important to realise that from sampling,

each processing stage will make the library less representative of the original environment. This does not mean that it has less value only that care must be taken when drawing conclusions about it.

#### **Library Construction**

Prior to conversion to phage, the phagemid library contained  $3.5 \times 10^6$  transformants, where 93% of those phagemid contained metagenomic DNA. This meant that the library had a complexity (number of inserts) of  $3.25 \times 10^6$  and, given an average insert size following conversion of 567bp, that it could potentially hold 1,842 Mb of DNA. Given an average bacterial genome size of 5Mb (between 0.6 – 10Mb), the phagemid library could represent 369 whole bacterial genomes. However, the presence of insert DNA in the phagemid does not guarantee that a recombinant protein will be displayed on the phage surface. It is estimated that only 1 in 18 phage will contain an insert that is in the correct orientation and in-frame with the promoter and gene VIII and therefore display a recombinant protein (Jacobsson et al, 2003), which leaves a theoretical 20.5 (369/18) genomes represented by the phage display library after conversion. Because the extracted metagenomic DNA contains several unknowns, such as the number of species and the abundance of each species within the sample, it would be almost impossible to tell what representation of the metagenome was present in the library.

Vector choice is extremely important in phage display. Previously, the minor coat protein III has been used for monovalent display (1 – 5 fusion proteins); however occasionally, because there are only 5 copies of the protein per phage, and overexpression is tightly controlled, phagemid would contain no fusion proteins at all (Jacobsson & Frykberg, 1995). Fusion to the major coat protein 8 facilitates multivalent display – fusions on more than one of the ~3000 copies of p8 – which allows identification of a wider range of proteins including those with lower affinity to the ligand but, due to the polyvalency of display, higher avidity. This avidity effect allows identification of a wider range of proteins than protein III display and should result in a wide variety of interesting bacterial binding proteins.

Choosing the correct bacterial host can also have an effect on the resulting library. *E. coli*, the microbiologist's equivalent of a white mouse, has limited capabilities when it comes to expressing environmental libraries. *E. coli* continues to carry out endogenous activities, potentially diverting resources away from the production of expression compounds, for example antibiotic resistance. Some bacterial inserts in the library will never be expressed in *E. coli*, as is thought to be the case with Actinomycete and other high G+C genes (Strohl, 1992), the promoters for which *E. coli* does not recognise (Kauffmann, 2004). Bacterial hosts which have the inherent ability to express the gene clusters needed for small molecule manufacture have since been developed. Martinez *et al* (2004) has extended the range of bacterial hosts to

### Chapter 3: Constructing a Phage Display Library

include the Actinomycete *Streptomyces lividans*, known for its expression of heterologous polyketides (Kieser *et al*, 2000), and *Pseudomonas putida*, a soil organism which not only produces a wide range of secondary metabolites, but which has now been developed for straightforward genetic manipulation. The Martinez group found that increasing the range of vectors and host strains facilitated high throughput screening, therefore increasing the likelihood of capturing the numerous and diverse natural products contained within.

Not only is protein expression dependant on the bacterial host used, but it also depends whether the sequence and folding characteristics of the encoded protein are compatible with transport through the bacterial inner membrane and display on the phage surface. Often, cytosolic proteins are incompatible with this translocation process, impairing cDNA library display (Slootweg, 2006). Lytic phage display, in the case of T7 phage applications, bypasses this completely since the phage are released from the bacterium by bacterial lysis instead of relying on the secretory mechanism (Slootweg, 2006). DNA fragments encoding external membrane proteins could be more readily expressed on the phage surface over intracellular proteins, which do not normally exist in an oxidising environment such as that on the outside of a cell. Although this is a true bias, it works in favour of this project as bacterial proteins involved in binding events with human tissue or immune components are likely to be externally distributed as cell-surface molecules/ extracellular proteins, and could therefore be stable on the phage surface.

Phage display is an extremely flexible molecular tool which allows functional screening for virtually any trait one might wish to search for. The combination of phage display with metagenomics to seek out binding proteins is a novel approach although functional analysis using phage display libraries of individual bacteria was previously successful and resulted in the identification of 4 genes encoding potential adhesins, none of which were previously proposed to code for adhesins (Mullen *et al*, 2007).

The major strengths of phage display are the creation of large libraries, the level of control over binding conditions, the link between phenotype and genotype to establish interactions, and the variety of freedom regarding the screening agent from various proteins to whole cells (Slootweg, 2006). It has been in use for over 20 years and the following chapters detail results from the unique combination of metagenomics and phage display.



---

## **Chapter 4**

### **Results and Discussion: Panning and Antibody Screening**

---

## Introduction

Using phage display, the natural specificity and affinity of fusion proteins on the phage surface can be exploited by screening against ligands (in this study; IgA, fibronectin (FN) and BSA) to identify protein-protein interactions. Genes or coding fragments can then be identified by sequencing. The procedure of affinity selection is called bio-panning (panning) and was used in an attempt to locate a range of bacterial binding proteins that may be involved in facilitating bacterial binding to the human tongue dorsum and oral cavity. The panning process is depicted in Figure 1, which shows a phage display library containing a range of fusion proteins (coat protein 3 depicted for clarity).

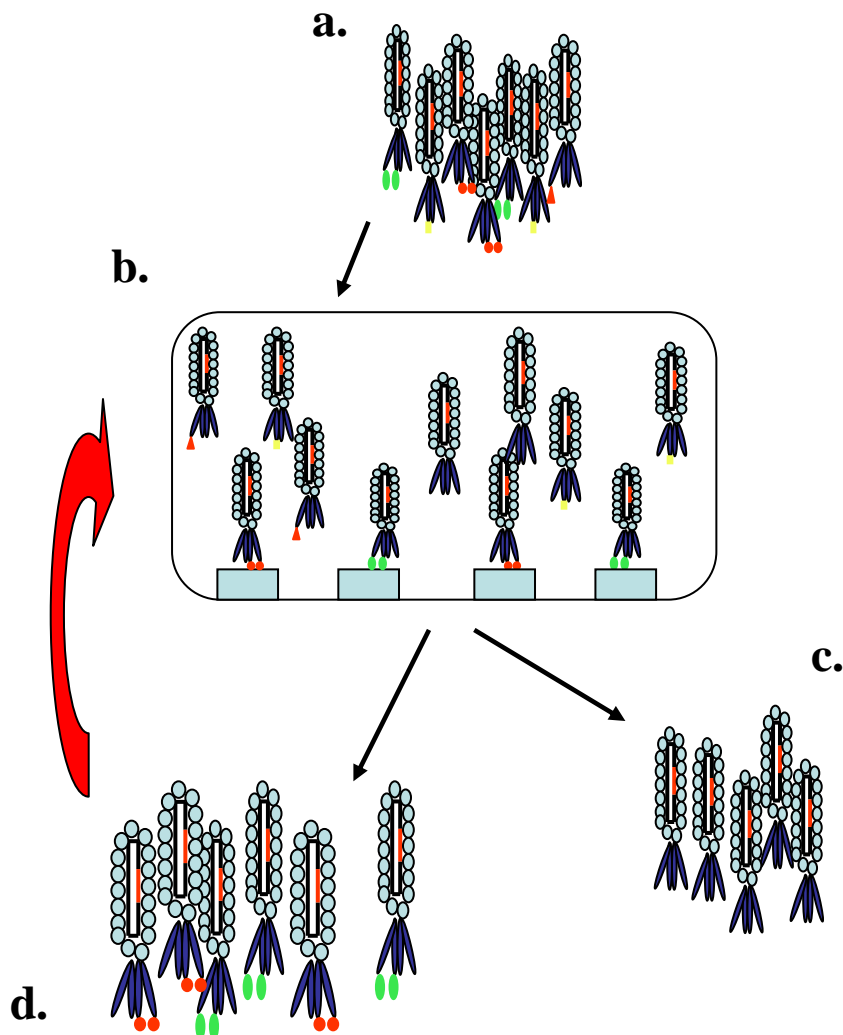
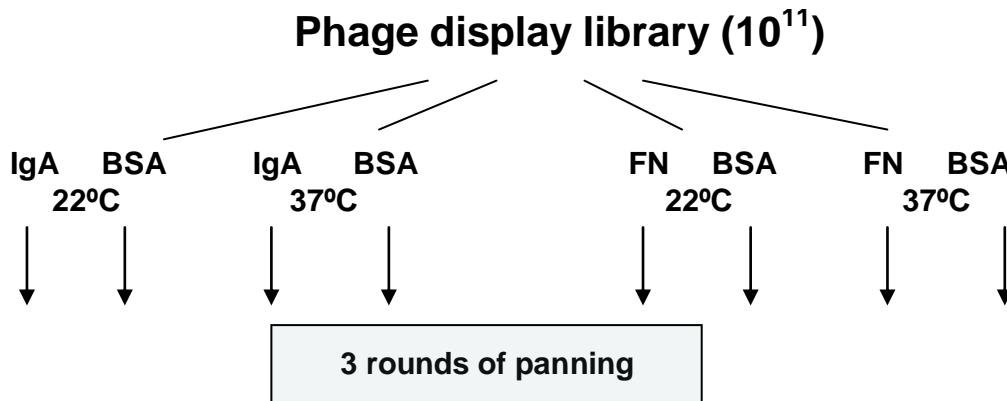


Figure 1 Panning process. (a) phage displaying various fusion proteins (green and red ellipsoids and spheres) are added to immunotubes containing immobilized ligand. (b) phage displaying fusion proteins with affinity adhere to the ligand and are retained, while (c) non-binders are washed off. (d) these specific phage are eluted and amplified before being added to the next round of panning.

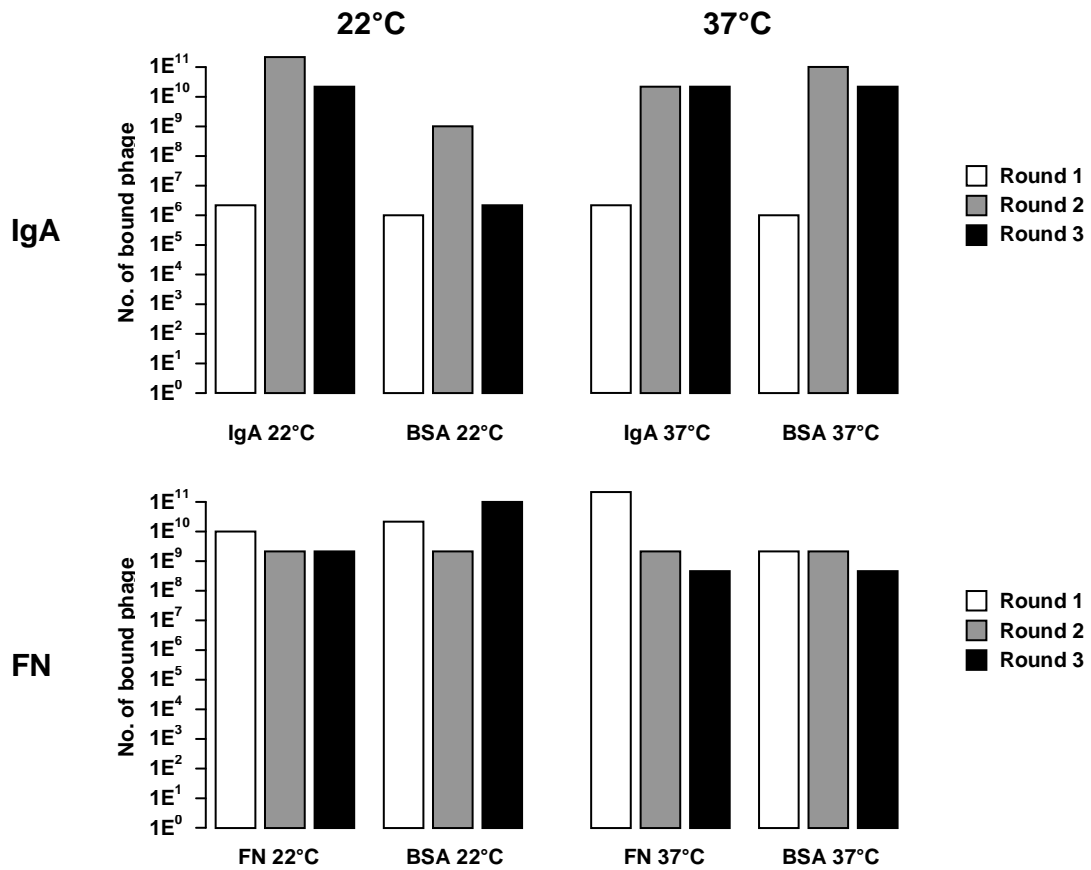
### Panning Procedure

The procedure for panning a phage display library is described in Chapter 2, page 56 and **Figure 2** provides a pictorial overview of the panning arrangement used in this project. Briefly, tubes were coated overnight with 0.5 mg of the ligands IgA, FN or BSA in PBS. Tubes were subsequently blocked with BSA for 2 hours and washed extensively with PBS-T (PBS containing 0.05% Tween20) then PBS. The  $1.7 \times 10^{11}$  phage display library constructed in Chapter 3 was diluted to  $1 \times 10^{10}$  CFU/ml and 2 ml added to each tube, which were incubated at either 22°C or 37°C for 2 hours. Unbound phage were removed and the tubes washed extensively with PBS-T then PBS as before. Bound phage were eluted in 1ml glycine buffer (pH 2.1) and neutralised with 0.5ml Tris-HCl (pH 8.0). Bound phage were enumerated upon elution by *E. coli* infection and then amplified for use in the second round of panning.



**Figure 2** Schematic representation of the panning of a phage display library at 2 temperatures. Each of the four panning experiments used BSA as a control, and both the IgA and Fibronectin experiments were carried out at 22°C and 37°C to determine whether a wider range of proteins might be more easily identified at a temperature closer to that of the original environment.

To amplify the phage for the next round of panning, eluted phage were added to log phase  $F^+$  containing *E. coli* cells and grown for 1 hour. Superinfections with helper phage R408 at an MOI of 20 were then grown overnight in 200 ml NB2 containing ampicillin. Phage were recovered by PEG precipitation and pellets resuspended in 2 ml TE glycerol. Three rounds of panning were used, each with an elution stage and an amplification stage, apart from the third round where the eluate was analysed directly without amplification. The titres from each stage are shown in **Figure 3**.



**Figure 3** Numbers of bound phage following 3 rounds of panning. Phage were enumerated by infecting log phase *E. coli* TG1 then counting colonies present on ampicillin selective agar.

Previously, Mullen *et al*, 2007 reported that libraries containing fragmented DNA from a single bacterial species showed clear enrichment for specific clones by an increase in phage titre from round 1 to 3. In the Mullen report, only half of the panning experiments showed enrichment, however, none of the panning experiments from this thesis, depicted in **Figure 3**, show this enrichment. This could be because metagenomic DNA is so diverse that enrichment for one specific clone would require more than 3 panning rounds and as such, demonstrates a wide variety of proteins which show binding affinity for the ligands used.

## Panning results

### (a) IgA

Two millilitres of phage display library ( $2 \times 10^{10}$  CFU total), was panned against IgA with a BSA control, at 22°C and 37°C (**Figure 3**). Following the first round of panning, a total of  $2.7 \times 10^6$  phage particles bound to IgA tubes at both temperatures, whereas a slightly lower number of  $1 \times 10^6$  phage particles bound to BSA at both temperatures. Upon amplification, an increase of at least three logs was produced of each phage population for the next round of

panning. In the second round of panning,  $2 \times 10^{11}$  and  $1.8 \times 10^{10}$  phage particles bound to IgA at 22°C and 37°C respectively. For the BSA tubes,  $4 \times 10^9$  and  $1 \times 10^{11}$  phage particles bound at 22°C and 37°C respectively. Upon amplification of the second round eluate, a small increase was observed in phage present. In the third round of panning,  $4.12 \times 10^{10}$  phage particles bound to IgA at both temperatures which, taken with the round 1 binding numbers, gives the impression that no temperature based enrichment is taking place for IgA binding proteins. In the BSA tubes,  $2.2 \times 10^6$  and  $3.64 \times 10^{10}$  phage particles bound at 22°C and 37°C respectively, which show a consistent high number of phage binding at 37°C in contrast to 22°C. This could be indicative of an increase in efficiency in domain folding.

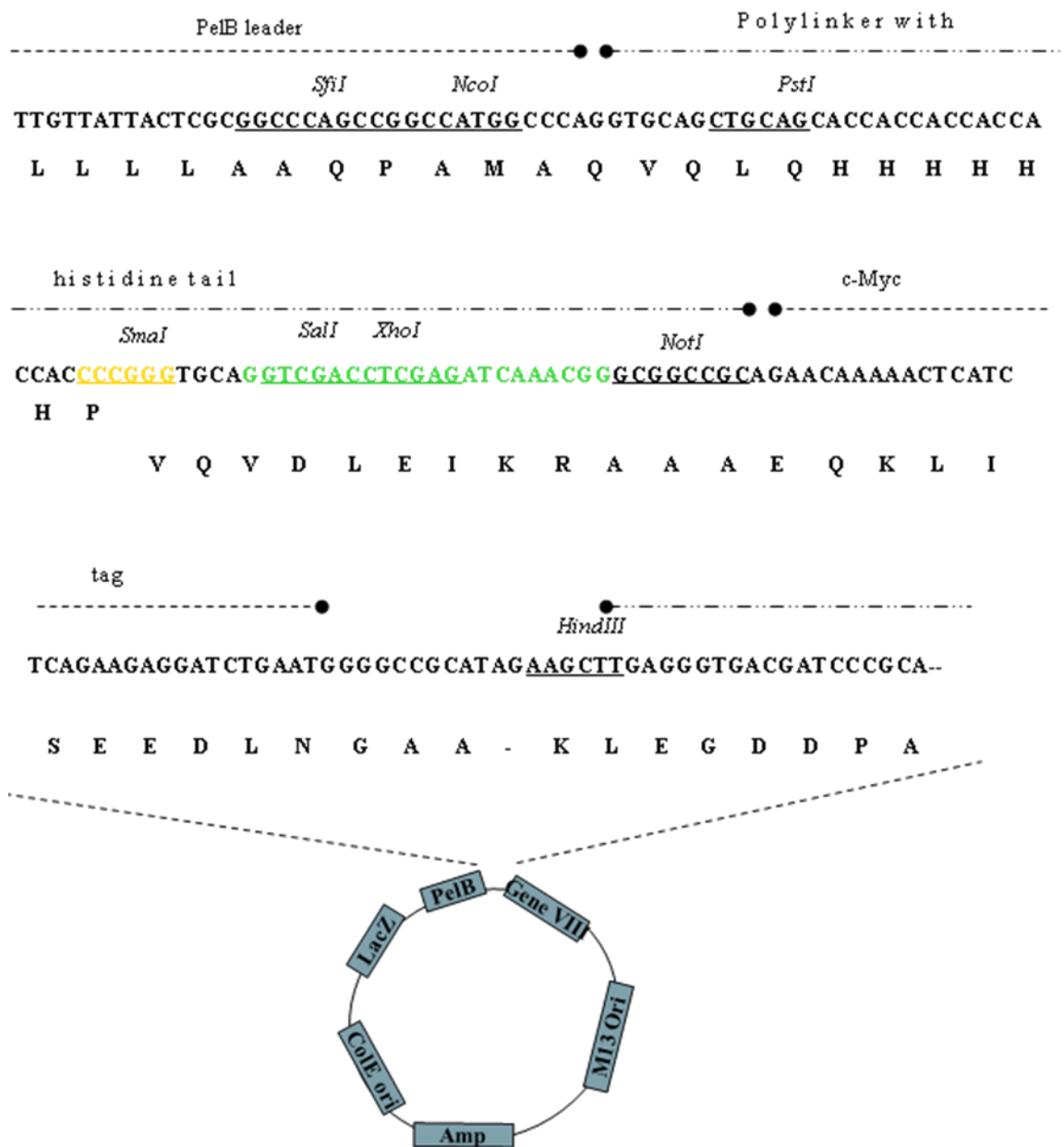
### **(b) Fibronectin**

Two millilitres of phage display library, containing  $2 \times 10^{10}$  CFU, was panned against Fibronectin (FN) with a BSA control, at 22°C and 37°C (Figure 3). Following the first round of panning, a total of  $2.4 \times 10^9$  and  $2.1 \times 10^9$  phage particles bound to FN at 22°C and 37°C respectively. In the BSA tubes,  $5.8 \times 10^9$  and  $4.1 \times 10^9$  phage particles bound at 22°C and 37°C respectively. These phage particles are showing an extremely high affinity for FN after only one round of panning. This result is understandable because FN is a known ligand for many different bacterial binding proteins (Joh *et al.*, 1998; Schwarz-Linek *et al.*, 2004) whereas no such comparable body of work exists for IgA binding proteins. Alternatively the high amounts of phagemid seen binding at the early rounds could include non-specifically bound phagemid particles. Due to the already high titre of the eluate, amplification added only one log to each phage population. In the second round of panning, both ligands at both temperatures bound phage in low ( $10^9$ ) numbers. There is a possibility that many of the binding phage from the first round had low affinity for the ligand and were replaced in the second round by fewer, more strongly binding phage, or those with more fusion proteins on the surface, contributing to the avidity effect. Due to the low titres from the second round eluate, amplification added one log to both BSA phage populations and two logs to FN populations. In the third round of panning,  $2.4 \times 10^9$  and  $5.2 \times 10^8$  phage particles bound to FN at 22°C and 37°C respectively. In the BSA tubes,  $1.3 \times 10^{11}$  and  $1.8 \times 10^8$  phage particles bound at 22°C and 37°C respectively, which is the opposite result to that in round 1 of the FN panning experiments, and the BSA results within the IgA experiments.

### **Bioinformatic Screening of Fusion Proteins**

Two hundred colonies from the 3rd round panning eluates were individually analysed by excising the insert and sequencing it. Vector bases were trimmed *in silico* from both ends of the raw sequence files, leaving the insert sequence and the first four amino acids of the vector -

VQVD. It was important to determine the amino acid sequence of the final protein product picked up by the panning experiments to begin to piece together potential roles for the bacterial proteins binding to human ligands. Because each sequence could have been ligated into the vector in any reading frame, and in the forward or reverse orientation, and because the insert sequence should remain in the same frame all the way through, the VQVD sequence acted as an identifier of the correct reading frame when the insert sequences were translated into all 6 frames (**Figure 4**). Proteins which appeared to remain in frame from start to finish were retained for closer analysis and all others were discarded.



**Figure 4** pG8H6 phagemid sequence showing the poly-His and c-Myc tags. Inserts are spliced into the *Sma*I site shown in yellow and the amino acids VQVD, below the yellow text, were used for *in silico* frame selection.

This analysis was supposed to show that the panning eluate contained a selection of binding proteins from which individuals could be analysed in more detail. One hundred clones each from the IgA and FN panning eluates were sequenced and a summary of findings are shown in **Table 1**.

Many of the amino acid sequences ending in VQVD contained one or more stop codons in all 3 forward reading frames, meaning that these inserts would not form a continuous protein and should therefore not appear on the phagemid surface. These clones should not have displayed any binding capacity above that of the phagemid alone. An effort was made to ensure the stop codons were not simply sequencing artefacts by resequencing following PCR cleanup using the Qiagen PCR Cleanup Kit. Although some stop codons were expected, the very high number in some clones in all 3 forward reading frames was not; especially since the phagemid vector pG8H6 contains a poly-His tag which is designed to allow frameshifting into the correct frame.

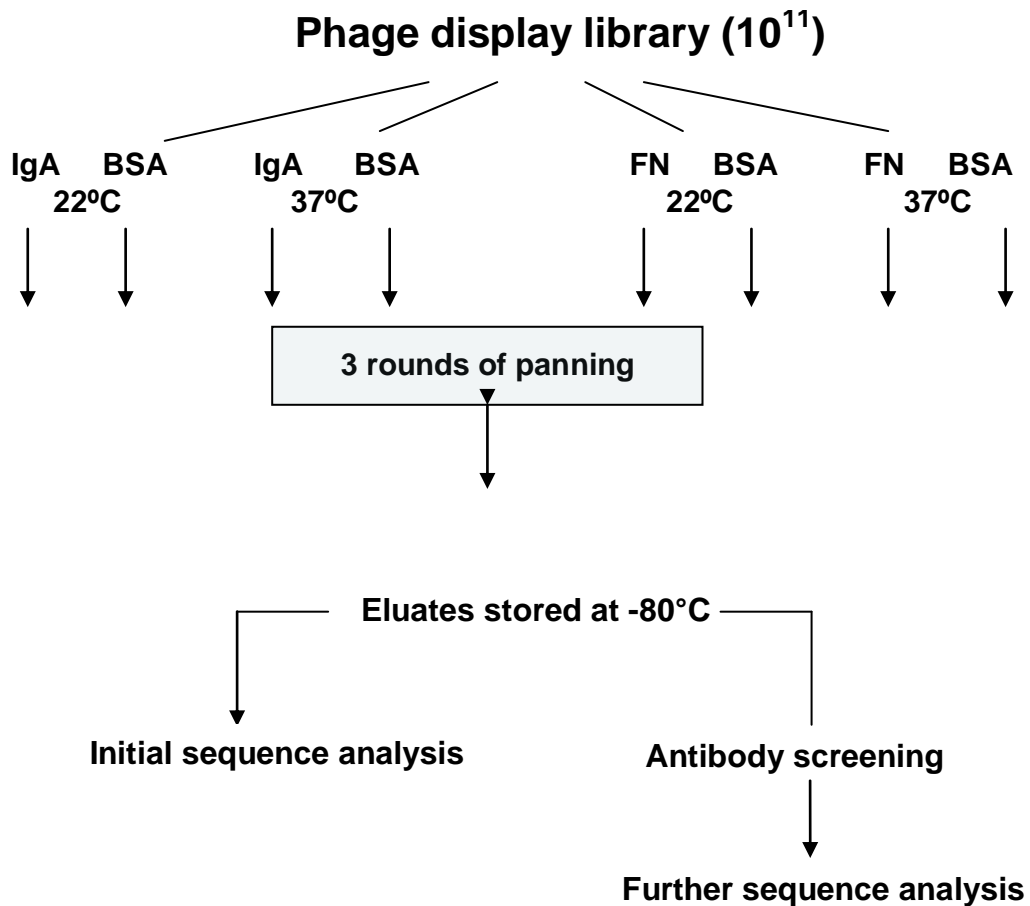
	<i>IgA</i>	<i>FN</i>
<i>Number of clones sequenced</i>	100	100
<i>Number of good quality sequences</i>	31	78
<i>Number of clones with/without stop codons</i>	24/7	72/6

**Table 1 Initial sequence analysis of panning experiments. One hundred clones were initially sequenced from each ligand, however many of the reactions encountered secondary structure and were not able to provide sequence data. On translation to amino acids using the VQVD sequence of the vector as a guide, it was clear that a huge number of clones contained between one and 20 stop codons in the insert sequence, and should not result in fusion proteins on the phage surface. The clones without stop codons were taken from this experiment and added to those from antibody screening.**

Because this initial analysis (**Figure 5**, arrow to ‘initial sequence analysis’) only identified 13 proteins suitable for individual protein analysis and expression, it was decided to analyse a much larger number of clones from the panning eluate using a different method. Antibody screening was introduced as a rapid screening step to identify and eliminate from further analysis those clones which did not express both the poly-His and c-Myc tags, and therefore a legitimate fusion protein. One hundred clones from each ligand IgA, FN and BSA were screened; 300 in total (shown on **Figure 5**, arrow to ‘antibody screening’). It was thought that a larger number of clones (300 as opposed to 200 in the initial analysis) as well as the rapid elimination of ‘decoys’ by antibody screening, would facilitate identification of many more legitimate proteins that could be studied in more detail.

**Antibody Screening**

Antibody screening analysis of 300 clones for the presence of the c-Myc and poly-His tags was carried out as described in Chapter 2, page 58. Briefly, each of the 300 individual phagemid were transformed into *E. coli*, converted to phage, and then superinfected with helper phage and the phage containing protein fusions recovered by PEG precipitation. These supernatants were spotted individually on nitrocellulose and screened separately for the c-Myc tag and the poly-His tail, present at either end of the insert (**Figure 4**).



**Figure 5** Schematic representation of steps taken following antibody screening. Following 3 rounds of panning an initial sequence analysis was carried out. During this analysis, most of the clones contained stop codons so this line of enquiry was taken no further. Another small aliquot was taken from the panning eluate and used to analyse 300 clones through antibody screening, looking specifically to eliminate those clones which did not express the c-Myc and poly-His tags.

Four outcomes were possible following this screening process:

- (i) A positive result for both tags (desired outcome)
- (ii) Positive poly-His, negative c-Myc



(iii) Negative c-Myc, positive poly-His

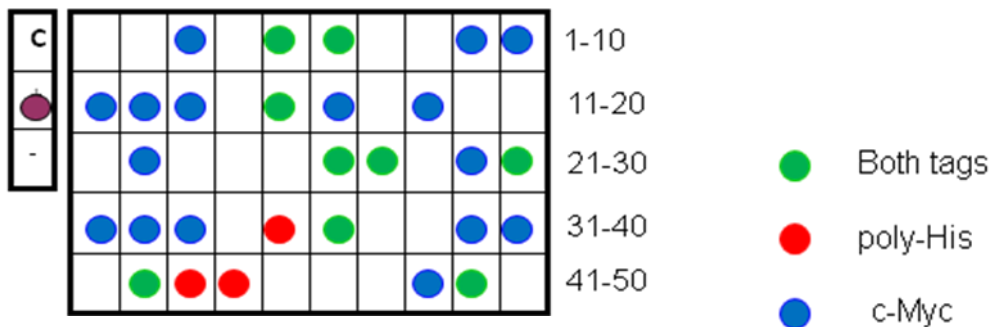
(iv) A negative result for both tags.

The presence of the tag combined with the BCIP/NBT (5-bromo-4-chloro-3-indolyl phosphate/nitro blue tetrazolium) substrate solution gave a colour change denoting tag presence (**Figure 6, Figure 7 and Figure 8**). If an individual clone showed a positive result for both tags, it was retained for further study. Similarly, any showing a negative result for both tags were not studied further. A positive result for c-Myc was important because it implied that translation of the inserted protein continued into the vector in the correct frame. However, it was conceded that a sufficiently large inserted protein may – by its size or folding domains - have blocked the tag, preventing detection by antibody screening as the c- Myc tag was closest to the phagemid coat protein and sandwiched between the inserted peptide/protein and the coat protein. Therefore, with one or two exceptions, recombinants with a negative result for c-Myc were excluded from further study. However, recombinants with a negative result for the poly-His tag were not immediately disregarded because it was considered possible that the inserted DNA sequence could contain an internal Shine-Dalgarno sequence, promoter and signal sequence, meaning the poly-His tag would not be present on the phagemid even if it did carry a fusion protein. Therefore, although most of the recombinants analysed did contain poly-His, those which did not, but which tested positive for c-Myc, were also taken to the next level of screening. This selection process reduced the number of recombinants in analysis from 300 to 221.

**Antibody screening colour change results**

(a) IgA

(a) 37°C



(b) 22°C

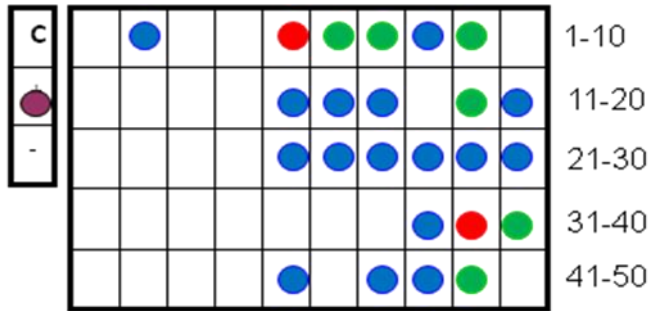
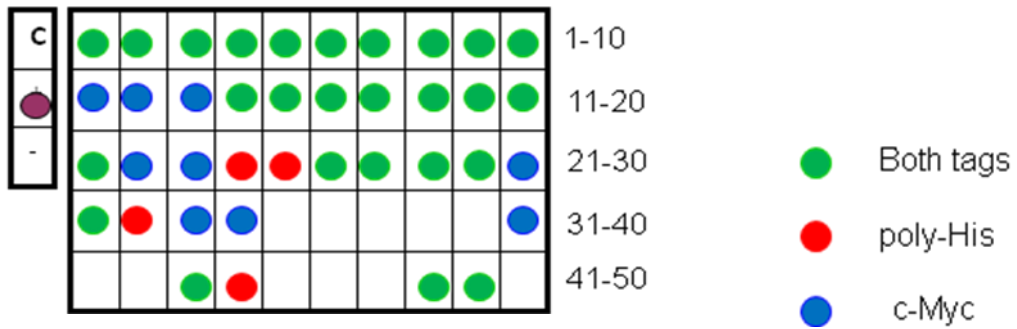


Figure 6 (a) and (b). Antibody screening results from phage display library panned against IgA.

(b) Fibronectin

(a) 37°C



(b) 22°C

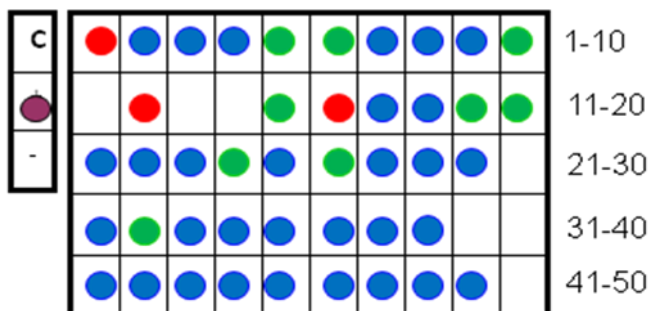
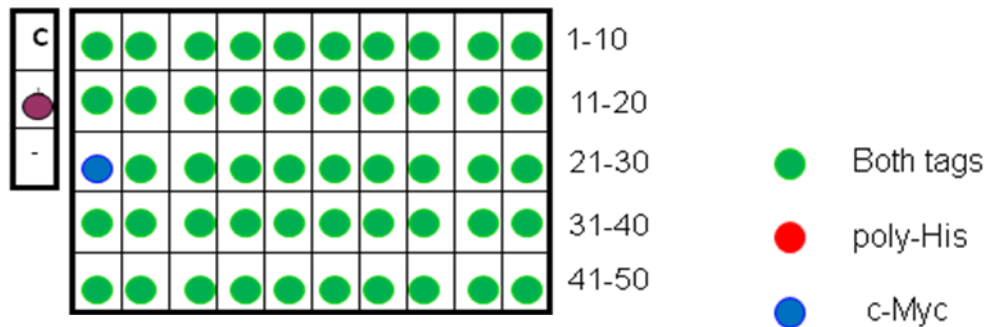


Figure 7 (a) and (b). Antibody screening results from phage display library panned against FN.

(c) BSA

(a) 37°C



(b) 22°C

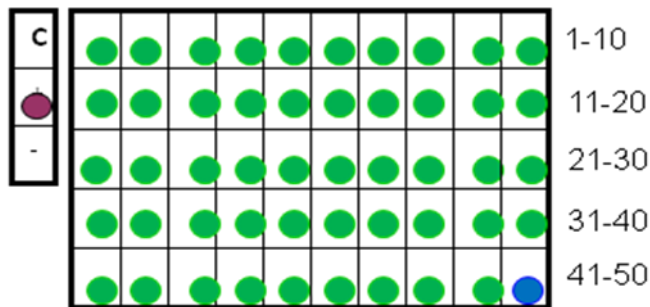


Figure 8 (a) and (b). Antibody screening results from phage display library panned against BSA.

From **Figure 6**, **Figure 7** and **Figure 8** there is clear disparity in tag presence depending on the ligand used for panning. IgA binding clones had the lowest tag presence of the three ligands, followed by FN, with BSA displaying the highest number of tags, almost 100%. This could be because of the size of the fusion proteins; in the BSA panning experiment they were generally much smaller than those in either the IgA or FN experiments and smaller proteins tended to contain fewer stop codons in the protein sequence in at least one frame (see Appendix 3, 4 & 5) so would have been more likely to express both tags with the fusion.

Recombinants containing tags were sequenced resulting in the elimination of some clones, including those containing no insert (10) and – from BLASTn searches – those containing human DNA (33). Unfortunately, there were also 34 instances where the presence of secondary structure meant that the sequencing reaction itself was unsuccessful and, following previous futile attempts to make these reactions work, these 34 clones were also discarded. This left 144 recombinants for the next level of analysis (BSA=66, FN=50, IgA=28).

## Chapter 4: Panning & Antibody Screening

The following level of screening involved a more detailed *in silico* examination of each cloned insert. Again, each insert was translated to the reading frame ending in the amino acid sequence VQVD, which signifies the changeover from insert back into the amino acids of the vector. Because all clones not displaying the poly-His and c-Myc tags had been removed from further study, the remaining clones should have been free of stop codons and able to be displayed on the phage surface. Unfortunately, a large number of the remaining 144 clones (69) testing positive for tags still contained multiple stop codons in all 3 frames (**Table 2**).

Occasionally, two fragments of metagenomic DNA would ligate to each other, prior to insertion in the phagemid vector, which was easily spotted from BLAST searches where the pictorial representation of the DNA sequence being compared clearly contained two different coloured sequences, and matches to completely different proteins. These clones were left aside. Additionally, as a final vetting process, some proteins were considered too short for the subsequent individual analysis, i.e., those less than 50 amino acids in length, were also removed. This final step removed 55 clones, leaving 17. The 13 clones which came from the initial analysis were included (see **Table 1**), plus an additional 3 clones from a previous lab project, which brought the total to 33 clones.

From these 33 clones, a shortlist of 18 was made for further analysis in order of insert size and interest value. Insert size was the crucial deciding factor in which of the 33 clones to include in the shortlist because, in order to carry out the intended binding studies with the individual proteins, the fusion protein should be of a sufficient size to retain some folding ability or domain structure. Ranking in terms of 'interest value' was based on BLAST results where proteins or sequences with no known homology to the database, or with homology to a hypothetical protein, were classed as potentially more interesting since they could be novel proteins. These 18 clones are described in more detail in Chapter 5.

	(a)	(b) Tag presence			(c) sequence data		(d) Following translation		
	Clones analysed	Both	c-Myc only	Remaining clones	No sequence data	Human DNA	Number of clones containing stop codons	Clones containing < 50 amino acids	Remaining clones
<i>IgA</i>	100	15	31	46	16	2	20	5	3
<i>FN</i>	100	35	40	75	10	19	28	10	8
<i>BSA</i>	100	98	2	100	21	12	21	40	6
<i>Overall</i>	<b>300</b>	<b>148</b>	<b>73</b>	<b>221</b>	<b>47</b>	<b>33</b>	<b>69</b>	<b>55</b>	<b>17</b>

**Table 2 Antibody screening experiment results.** As illustrated in column (a) 100 clones from each panning experiment were analysed by antibody screening for the presence of the poly-His and c-Myc tags. (b) shows the results of the antibody screening experiment, with some clones showing a positive result for both tags and some for c-Myc only. The remaining 221 clones were sequenced (c), where ‘no sequence data’ means that either the clone had no insert or the sequencing reaction was unsuccessful. Column (d) shows the results following *in silico* translation into the correct reading frame (containing VQVD entering the vector), where clones containing stop codons in all 3 forward frames were removed from further study. In addition, clones containing less than 50 amino acids were also removed, leaving 17 clones for individual analysis.

## Discussion

### Panning considerations

There are many considerations when panning a phage display library which have the potential to affect the outcome:

1. Choice of ligands: One of the main aims of this project was to locate binding proteins responsible for interactions between bacteria and the human tongue dorsum, so the choice of FN and IgA are supportive of this goal as they are known to interact with bacteria in the oral cavity. FN especially is known to interact with a plethora of bacteria in its role as a part of the ECM. It was anticipated that studying the binding interactions of IgA against a range of bacterial proteins would unearth some previously unseen communication between the two, which would be interesting since far fewer IgA binding bacterial proteins have been identified than FN. Because the oral cavity contains many host-derived proteins (Table 1, Introduction page 31), a wide variety of ligands could have been used for panning, and could still be used in future to pan the existing library.
2. Number of panning rounds: Typically 3 rounds are used with libraries produced from the genomes of distinct bacterial phylotypes (Mullen *et al.*, 2007). However, because metagenomic DNA was used in this project and resulted in a wide variety of binding proteins, even after 3 panning rounds, panning for more than 3 rounds may have only acted to narrow the range of proteins to those with highest affinity to the ligands. Additional rounds of panning were not added however, since more rounds could have resulted in the loss of bacterial proteins with lower expression levels or of those with a lower affinity to the ligand, but still novel or intriguing to study. Phage enrichment described by Mullen (2007) and previously discussed in this chapter refers to enrichment of a particular phage containing a particular fusion protein, however it seems unlikely that enrichment of a single fusion protein would occur from a metagenomic library using only 3 panning rounds.
3. Frameshifting: Commonly, inserts identified through panning are not in frame with the pG8H6 vector sequence (Jacobsson *et al.*, 1995; Jacobsson & Frykberg, 1996; Carcamo *et al.*, 1998). It appears there is a selection for inserts out of frame (usually frame +1 or -1) and ribosomal slippage corrects the frameshift during translation. It was for this reason the pG8H6 vector was constructed with the poly-His tag in front of the insert sequence (Jacobsson and Frykberg, 1996) and a stretch of 5 adenosines in the c-Myc tag, which should have resulted in fusions in frame. However, by selecting an out-of-frame fusion, the phage display system can mediate fusion protein expression levels by frameshifting to downregulate expression. This means that expression does not always remain high enough for display at the phage surface (Jacobsson *et al.*, 2003) meaning the expression levels of some proteins are too low to enable binding or detection (Jacobsson & Frykberg, 1998).

This is why it is common to see a high titre of bound phage following the first round of panning, and a lower titre in the second round; the system has introduced a frameshift to allow the production of viable phage which may have reduced expression levels such that these fusion proteins are no longer affinity-selectable (Jacobsson & Frykberg, 1996).

4. Using two panning temperatures: Incubations and washes are normally carried out at 22°C as cooler temperatures could prevent dissociation of the bound phage (Woiwode *et al.*, 2003). The hypothesis was that using a panning temperature closer to that of the human body would identify a range of proteins which are normally produced at that temperature. It was hoped that two distinct groups of binding proteins would be found, with perhaps more identified closer to body temperature than 22°C. This may still be a useful experiment when panning a library produced from a single species, however, no distinction was observed in the bound proteins between the two panning temperatures in this project. This is probably due to the diversity of the library and also that the number of clones picked out for analysis was too small to provide sufficient data for conjecture on content of the entire library.
5. Valency of display: This point relates to the number of times the fusion protein appears on the phage surface and is important because of its impact on the ability to discriminate binding proteins with varying affinity for the ligand. Polyvalent display allows the expression of multiple fusion proteins on coat protein 8 which, along with high affinity proteins (strength of a single bond), allows weaker binding clones to be identified where polyvalency causes a strong interaction (high avidity – combined strength of multiple bonds). In contrast, phage display using phage coat protein 3 allows between 1 - 5 fusions per phagemid, so selection is based purely on affinity and is mainly used in studies where only the tightest binding variants are required (Russell *et al.*, 2004). Currently, it is not known how many fusions appear per phage or how much that number differs between phage populations or within the same phage population.

Of the 5 points above, choice of ligand is the principal factor in dictating the binding proteins recovered from the panning process. Both FN and IgA are known to interact with bacteria by binding to them or being bound by them. They are very different ligands in terms of structure, size and purpose, and it was hoped that this variety of function would facilitate identification of a range of binding proteins.

As well as using BSA as blocking agent in the FN and IgA panning experiments, BSA was used as a control for checking protein binding between panning experiments. However, the tertiary structure of BSA is similar to Human Serum Albumin (HSA) (Geisow & Beaven, 1977) and, because many bacterial proteins bound to it, panning against BSA acted less like a control and more like a separate panning experiment. In plasma, HSA is one of the most

dominant proteins, along with immunoglobulins (Ig) (Johansson *et al.*, 2001), and this has led to the isolation of proteins that specifically bind albumin by several bacterial phylotypes such as *Streptococcus pyogenes* (Frick *et al.*, 1994). The use of BSA in panning is discussed in more detail in Chapter 8.

### **Antibody screening**

Antibody screening was used to rapidly eliminate phagemid displaying fusion proteins which contained stop codons and therefore were not in frame with the vector. The resulting fusion proteins should therefore have been uninterrupted with a true binding affinity which could be studied further in individual analyses. Ending up with 17 clones out of 300 analysed fits with Jacobssons' statement that only 1 in 18 clones will be in the correct orientation and in frame however, the presence of both tags on a clone clearly was not indicative of a legitimate binding protein. Out of the 148 clones displaying both tags, many seemed not to contain a legitimate ORF at all, even though they had been enriched through 3 rounds of panning and displayed the tag indicating fusion to protein 8. Looking back at **Table 1**, at the very first analysis of the panning eluate, 13 clones satisfied all the requirements shown in **Table 2**. Following at least 6 weeks of phage supernatant precipitation and antibody screening, 17 additional clones met the same requirements.

Aside from identifying potential positive clones for individual analysis, antibody screening could also indicate how efficient the phage are at producing fusion proteins, a more intense colour meaning more tags, i.e. more protein. It was decided to use 5 µl of phage supernatant for the antibody screening experiment; however this volume would not have contained equal numbers of phage particles, or the same number of fusion proteins per phage since these features depend on the phage population. It is difficult to quantify the amount of fusion protein produced by a phage simply by making an objective decision based on a colour change. This is made more difficult by the potential of each individual phage particle to display a different number of fusion proteins (increased avidity), which depend on how well the fusion becomes incorporated into the phage coat and its behaviour on the phage surface.

Frameshifting is frequently mentioned as an issue by other groups using phage display (Carcamo *et al.*, 1998; Jacobsson *et al.*, 2003), and the enduring presence of stop codons in the coding sequence was very common in this project. In many instances, the DNA sequences of the peptide encoding genes alone were sufficient to draw some conclusions regarding the types of organisms in the sample. However, this project intended to use some individual proteins for expression and eliminating those containing stop codons left few clones with which to take the planned individual analysis forward.



## Chapter 4: Panning & Antibody Screening

In order to still be present following 3 rounds of panning, the clones containing stop codons must have had some binding capacity and, until sequencing, they appeared to be legitimate binding proteins. Much time was wasted during the analysis of these clones in order to reach a sufficient amount to take forward to pET vector expression.

From the antibody screening data (**Figure 6**, **Figure 7** and **Figure 8**) it is quite clear that most of the BSA-binding clones should express both tags, where FN-binding clones contain fewer and IgA-binding clones contain fewer again. Diminishing tag presence may be due to the insert size where, as in the BSA-binding clones, smaller fusion proteins facilitate expression and display of both tags in concert, which could be due to smaller proteins containing fewer stop codons. Larger proteins like those more commonly found binding to FN or IgA, with bulky folding domains could block the tag, negating detection.

In **Table 2**, the 17 successful clones resulting from antibody screening are clearly outnumbered compared to the 55 clones which were smaller than 50 amino acids (column d). As expected, the phagemid system preferentially expresses proteins which are smaller than 50 amino acids, and because the largest proteins were the priority of this study it is likely that some interesting small proteins have been completely bypassed. In particular, some FN binding proteins are known to share a similar repeated modular architecture to the FN molecule, and some of these repeated domains are around 40 – 50 amino acids in size. Binding proteins like these could be among the 55 small proteins found.

Because there is no power over which fragments are cloned into the phagemid vector during metagenomic library construction, there were probably millions of ‘decoy’ sequences in the wrong frame and containing multiple stop codons present in the library prior to panning. Gene 8 was used for fusion in this project because the polyvalent aspect of display should have resulted in selection of weaker binding interactions (high avidity) as well as strong (high affinity). There is a possibility that it resulted in many non-specific interactions being mistakenly analysed through high avidity clones. Individual proteins are discussed in more detail in Chapter 5.

---

## **Chapter 5**

### **Results and Discussion: Individual Protein Analysis**

---

## Introduction

As described in Chapter 4, antibody screening and *in silico* analysis of a selection of recombinant clones, affinity purified through 3 rounds of panning, led to the final shortlisting of 18 proteins. These 18 clones resulted in the translation of novel/interesting proteins that bacteria may use to facilitate binding to the human tongue dorsum. This chapter looks at the final 18 in more detail, and discusses plans for their analysis according to size and 'interest value', i.e. if the insert coded for a previously unidentified protein. Larger inserts, likely to contain more of the gene of interest, were thought likely to retain folding and binding properties similar to the full length protein, and these were prioritised in the list (Table 1).

One of the targets of this project was to express one or more of these 18 proteins using the pET vector expression system. Successful expression was to be followed by periplasmic extraction, and then individual assessment to check binding affinity for panning ligands, along with other traits. With 18 potential proteins and the likelihood of a huge time investment, clones were taken forward in order, the largest and most interesting first. This chapter describes the 18 shortlisted clones in more detail and outlines the pET expression studies and the subsequent adhesion assays which became necessary following continuous pET expression failure.

## Final 18 Clones

A list of 18 clones identified from panning a metagenomic phage display library was compiled in order of priority, based on apparent interest and the size of the protein product. This list is shown in **Table 1**. The 18 shortlisted proteins were found to contain homology to some interesting proteins in the NCBI BLAST database, regardless of the species level match. Certain putative proteins were more interesting than others, for example number 36 shows very low identity (52%) to an outer membrane transport protein, or number 42 shows some homology to a pili biogenesis protein. These are exactly the types of proteins that were expected from the phage display/panning process; likely to be oriented on the bacterial cell surface where they could potentially act as receptors for human cells or salivary components.

Certain types of bacteria and adhesins were expected to result from this metagenomic analysis of the tongue surface. Due to the high concentration of S-IgA in the oral cavity (Scannapieco *et al.*, 1994), organisms encoding IgA proteases, common colonisers of human mucus membranes, would certainly be expected. These include *Prevotella* sp., several *Neisseria* sp., and some *Streptococcus* sp., including *pneumoniae*, *sanguinis*, *oralis* and *mitis*, all commonly found in the oral cavity (Kilian, 2003) and present in this study.

<i>Clone Number</i>	<i>Binding Ligand</i>	<i>Tag presence</i>	<i>tBLASTx result</i>	<i>Protein sequence</i>
36	BSA	BOTH	<i>A. pleuropneumoniae</i> OM receptor protein (Fe transport) E=7e-21, id=45/85 (52%)	LIFLLGLDAPLTDIWKIGNNISTGFRNPTASEMYFSFEHPAGNWIPN PDLKAEQALNQSIYQAEHLLGSFGLTFYHTRYKNLLTEQUESTYKK RNPYYNAYSASYGQQGVQVD
59	IgA	n/a <sup>1</sup>	<i>F. nucleatum</i> chloride channel protein E=2e-46, id= 99/100 (99%)	LILGRINYNNWFFELLAKFFAGVVLGIGAGLSLGREGPSVQLGSYVG YGASKILKTDTVERNYLLTSGSSAGLSGAFGAPLAGVMFSIEEIHK YLSGKLLI
1	IgA	BOTH	<i>A. odontolyticus</i> hypothetical protein E=1e-29, id= 62/62 (100%)	PLLVLLVDPIVSGGNASEADAGHEIAARVWRVGSDDLTAGVDVPAP GTQVGLAPEIACGHCAPCTSGRSNVCANMRLFGTGVDG
39	IgA	n/a <sup>2</sup>	<i>G. kaustophilus</i> twitching motility protein Expect = 0.001, Id = 17/27 (62%)	VMAVHRMISLFPGEQQEERSQISQVLRAVICQRLLRWNKKFITIRD ILLNTHAVANLIRTRKEPQIISIQETQLPMKTLEMGVQVD
42	IgA	n/a <sup>2</sup>	<i>N. meningitides</i> FtsK DNA translocase E=1e-16, id=48/85 (56%)	LYLMTAKSSKTQTKKRASTKPAAKPTTRKSAKTQTQADNKVSQR LKAAKELQKNEEKARPEHVVNILINDALWLFGLVITIYLGVD
58	IgA	n/a <sup>1</sup>	<i>S. usitatus</i> radical SAM domain protein E=3e-14, id=35/97 (36%)	STLMIGMETDTVESIRQIPDIEEIGVDVPRYNILTPYPGTPFYEQLK AENRLLTRDWYYYDTETVVFQPKNMSPATLQEEFYKLWQDTFTY KRIFK
44	FN	n/a <sup>2</sup>	<i>Rothia</i> sp. Aspartate/ornithine binding domain E=1e-13, id=32/39 (82%)	PHTVSASADNNALMTCWSRERIKSGDAWDNASPSRPESDSGWRC ASSVKSNDASIVRVRCSARLVESSTSSNTLSRK
52	FN	n/a <sup>2</sup>	<i>A. metalliredigens</i> IM component E=2e-09, id=29/74 (39%)	IGIVKGGLAGFSTPSIDRWLSRLIDLVLGFPMVIAIAFIGIMGPSITN VIISLCITKWAELYALITRGLVVVEKVFYRH

Chapter 5: Individual Protein Analysis

22	FN	BOTH	<i>N. meningitides</i> hypothetical protein E=4e-10, id=39/72 (54%)	AEAGHIEAAFQLAGCLFENHENEQDLAIAVEYLKQAARAGHPYAR YNLLQLQENNGAEVETLISAYQELAE EGLVP
30	BSA	BOTH	<i>A. odontolyticus</i> hypothetical protein E=0.012, id=18/29 (65%)	ERRRMAEYLASPGYDHVMHVVRARFMAGNYYDL CAGVCRDFA NTVGNFNIDRGVADGHWTRPTRRRRHGIGLGLGVQVD
60	IgA	n/a <sup>1</sup>	<i>S. mutans</i> putative ABC transporter	VILGLIFFLDTRLGQAYIATGDNSDMAKSF GINTDRMELMGLVISN GIIALSGALMAQQE GYADASRGIGVIV
17	FN	c-myc	<i>C. botulinum</i> PstC, ABC-type phosphate transport system. E=2e-08, id= 33/63 (52%)	YIISASLYVSLLSLIWALPLGIGTSVGLSLGVSPRIRQFCLSTIDMIAG IPSVIVGFIGLAVVVP
19	FN	BOTH	<i>A. odontolyticus</i> hypothetical protein E=4e-27, id= 55/63 (87%)	KSWNFQDAGIGMAAINAYHSHPEVALARGFTPC EENNWARTFHP YAPLVAGKRVAIIGHFPFAGVQVD
20	FN	BOTH	<i>A. odontolyticus</i> hypothetical protein E=3e-22, id= 50/53 (94%)	LGVENLYEAANTPLIGFLNNAIRAKELFFRDRDYIVDAGEILIVDEH TGRVLP
27	BSA	BOTH	No similarity	EGTPPENRDGTCRVLVLPRVQPPAGRLHGRQWLHEGRRGFFLGV R VSEKARTRKATR
2	IgA	BOTH	<i>P. gingivalis</i> glycogen synthase E=4e-10,	RSIAGNDKELYTYMDAYDGDQMARELGVEAKHEVEKLA AHKAR TVMPAALPVIASAPAGVQVD
16	FN	c-myc	<i>H. influenzae</i> PP-loop superfamily ATPase E=1e-18, id= 47/57 (82%)	QRHAIELEKARWIAKDLGVKQTLIDTSVIKSITHNALMDANADIEQ KDGELPNTFVD
11	FN	BOTH	<i>B. fragilis</i> Fe-S-cluster redox enzyme E=1e-08, id= 28/51 (54%)	LNLLELKAVAKEFGMPAFTGGQMAKWLYIQHVTTIDEMTNISKNN REKLKA

**Table 1** Table of 18 clones identified from panning experiments against the ligands IgA, Fibronectin or BSA, following 3 rounds of affinity selection. Alongside the clone number is information regarding the panning ligand where affinity was shown and the presence or absence of the Poly-His or c-Myc tags. Data from tBLASTx searches are shown and the resulting protein product as identified from the search database. BLAST searches are accurate as of 25<sup>th</sup> June 2009.

<sup>1</sup>No tag data is available for these clones because they resulted from a parallel project to the present thesis, supervised by the author, and have been included in this study.

<sup>2</sup>No tag data is available for these clones as they resulted from panning experiments carried out prior to antibody screening.

In general, IgA binding proteins of streptococci belong to the M protein family, and are dimeric coiled-coil molecules that can bind more than one Ig at a time. IgA binding proteins bind to various parts of S-IgA, but some bind to the Fc part of human IgA and are particularly expressed by pathogenic bacteria, mostly G<sup>+</sup>ves. Although not all IgA binding proteins are related, some regions have been identified that bind to IgA including the Sir22 binding protein (Johnsson *et al.*, 1999) and Arp4 (Johnsson *et al.*, 1994), which both contain a 29 residue binding region. None of the putative IgA binding proteins identified from this study included this binding region. IgA proteases cleave IgA1 at the hinge region leaving the Fab fragments free to be bound by other bacteria, commonly seen in plaque bacteria, which enables them to evade the host immune response. Several *Streptococcus* sp. express surface proteins which bind IgA molecules in an antibody-independent manner, i.e., through the secretory component (SC) (Kilian, 2003), for example the protein SpsA of *S. pneumoniae* that binds SC is expressed by 2/3 of strains and is conserved between serotypes (Hammerschmidt *et al.*, 1997). None of the proteins identified from the IgA panning eluate showed any homology through the NCBI database to IgA binding proteins. The Fc part of human IgA also binds commonly to Group A Streptococci, and the IgA binding proteins which facilitate interaction are members of the M-protein family (Johnsson *et al.*, 1999).

The oral colonising streptococci are well known for their adhesion properties to the extracellular matrix (ECM) and in particular fibronectin (FN), a major component of the tongue dorsum which is also present in soluble form in saliva. FN-binding proteins from the shortlisted 18 proteins would also be expected to share similar characteristics to known FN-binding proteins. The best-characterized FN-binding proteins from streptococci and staphylococci share a similar mosaic architecture (similar, in fact, to FN), the ligand-binding domain consisting of tandem repeats of a 45 amino acid long unit which binds to the 29-kDa N-terminal region of FN (Joh *et al.*, 1999). In fact, a recently described adhesin of *Campylobacter jejuni*, named fibronectin-like protein A contains fibronectin type III domains (Flanagan *et al.*, 2009). These cell surface proteins possess an N-terminal signal peptide for *sec*-dependent secretion and an LPXTG C-terminal motif for covalent anchorage to cell-wall peptidoglycan (Navarre & Schneewind, 1999). Of the 8 putative FN binding proteins identified from this study, all contained an N-terminal signal peptide according to the SignalP search tool from expasy tools ([www.expasy.ch/tools/](http://www.expasy.ch/tools/)). None however, contained the LPXTG C-terminal motif within the length of the protein sequence, which is probably because the full length gene was cut short by the phage display library construction process. Again, none of the proteins investigated from the FN panning eluate showed any homology to known FN binding proteins, however, as previously mentioned, phage display does not operate without bias and, for whatever reason, either the known FN binding proteins were not represented in the library at all, or were not in the small number of clones analysed.

Known albumin binding proteins have been found to express a conserved 45 aa stretch within a 52 aa region (Jacobsson *et al.*, 1997), which is homologous with albumin binding modules found in other bacterial cell surface proteins. Again, none of the 18 proteins shortlisted from the panning experiments detailed in previous chapters showed any identity to this region.

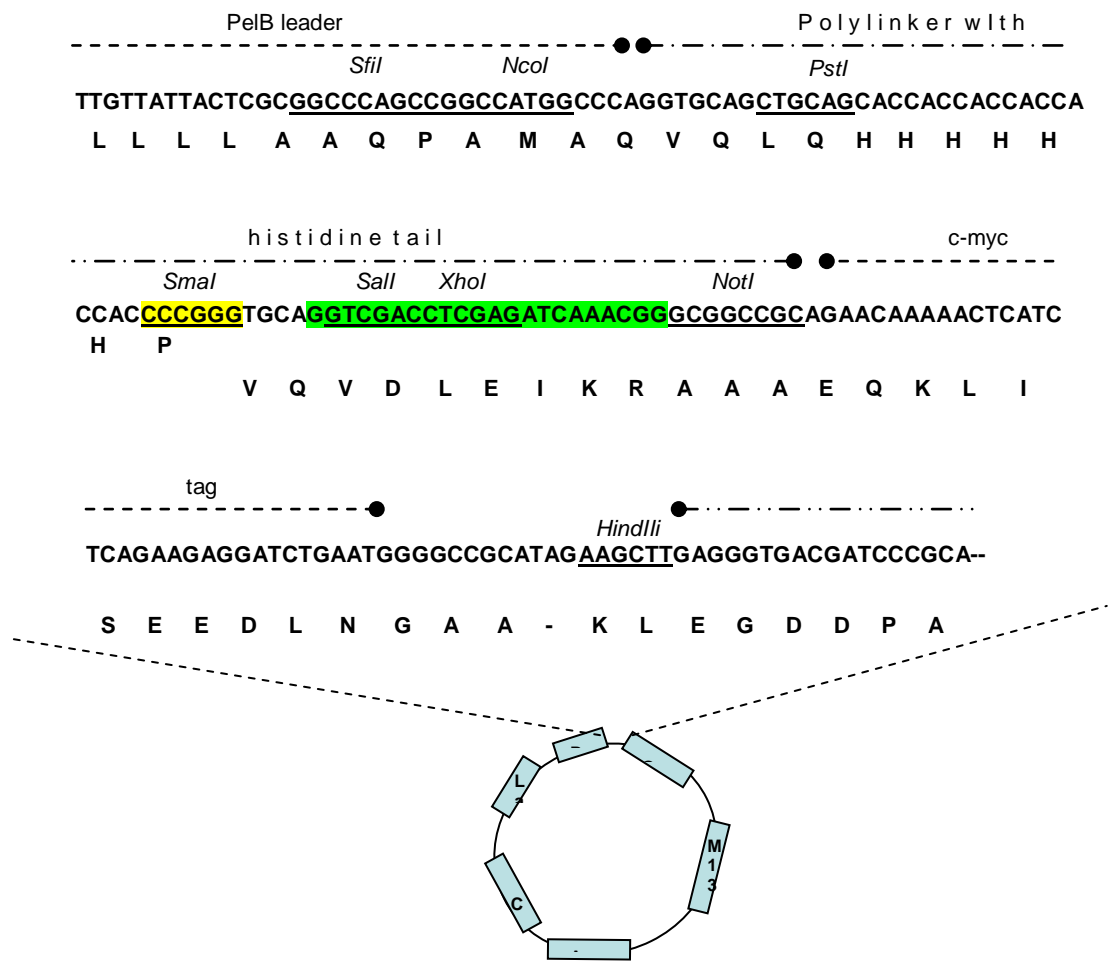
### **pET Expression**

A wide variety of pET vectors are available, all derived from the commonly used *E. coli* cloning vector pBR322. The vector used in this project is pET 21b, a transcription vector which expresses target RNA and provides translation signals. The main reason for choosing pET 21b is because it has an expanded multiple cloning site in all three reading frames, allowing the inserts of interest to be cloned in by precise primer design to match the reading frame required for expression.

Before cloning the DNA fragment of interest into pET21b for expression, each insert required PCR amplification, involving the design of individual primers for each of the 18 clones. The primers were designed to lift out the insert from the phagemid vector pG8H6 complete with restriction enzyme sites needed for easy cloning and expression into the pET vector expression system. **Figure 1** illustrates the main features of the pG8H6 phagemid vector and how the primers were designed to facilitate cloning into pET.

PCR amplification was performed in 0.5 µl Eppendorf tubes in a total reaction volume of 50 µl. Reactions comprised 1 µl template DNA (phagemid containing insert preparation), 2 µl of 5 µM each oligonucleotide primer and 45 µl PCR supermix. All reactions were prepared on ice. The PCR was performed for 30 cycles of 95°C, 5 minutes (1st cycle only) 95°C, 30 s; 54 - 68°C, 30 s; 70°C, 60 seconds; 70°C, 5 minutes (last cycle only), using a Techne TC-512 gradient thermal cycler. The PCR product was quality and size checked using 1% agarose gel electrophoresis. All inserts were successfully PCR amplified, with the exception of clone number 58 which, despite several attempts, was not amplified. In the interest of time, clone 58 was left aside.

In the following stages, clones were taken forward in more manageable groups of 4. In the interest of simplicity, inserts were cloned first into TOPO TA (Invitrogen) with the aim of using the introduced *Nde*I and *Xho*I sites from PCR amplification to ease cloning into pET21b. However, the following problems were encountered with cloning into TOPO: **i.** ligation of amplified insert into TOPO TA was not successful, (and this stage was not able to be checked on a gel due to the tiny volume of vector present), and/or; **ii.** transformation of ligation was not successful (no colonies present on kanamycin agar plates). If ligation and transformation were both successful then TOPO clones containing the insert were grown up individually in nutrient



**Figure 1** Schematic drawing of the phagemid pG8H6 showing main features and sequence of cloning site used for primer design. Individual primers were made for each clone at the N terminal end of the insert, keeping the protein product in frame once expressed and incorporating an *NdeI* site. From the C- terminal end of the insert, a universal primer was designed which spans the *XhoI* site, in the same frame for cloning into pET21b. Yellow highlighted area is *SmaI* restriction site used for cloning insert sequences. Green highlighted area illustrates the sequence which was used as the universal primer incorporating *XhoI* site.

broth and a high concentration plasmid preparation (Qiagen Plasmid Prep Kit) was digested with restriction enzymes *NdeI* and *XhoI* before quality checking on a 1% agarose gel and extracting the insert using Qiagen Gel Extraction Kit. Several issues were encountered at this stage including; **iii.** unsuccessful restriction digest (not able to be checked until a 1% agarose gel was run); **iv.** there was not enough insert DNA to set up a ligation following gel extraction.

If the restriction digest was successful, and recovery from gel extraction gave an adequate concentration of insert to set up ligations, then the insert was cloned into pET21b (pre-digested with *NdeI* and *XhoI* – another stage which could not be quality checked) and



transformed into electrocompetent BL-21 *E. coli* cells. Problems at this stage were; **v.** unsuccessful ligation and/or; **vi.** unsuccessful transformation. Due to these issues, none of the remaining 18 clones were able to be expressed using the pET expression system in this project. Apart from the problems mentioned, expression cloning in *E. coli* gives accessibility to fewer proteins than originally thought since many expression signals do not function in this organism (Gabor *et al.*, 2004).

The problems encountered with pET vector expression may turn out to be more easily solved than initially thought. Phage display fuses inserted proteins with a structural protein which provides support to the fusion protein (protein 8), which may in some cases, enable folding of the fusion into its native state. pET vector expression linearises the peptide insert into a plasmid with no flexibility, and relies on the proteins ability to fold completely by itself upon expression. This may not be a suitable method of expression for small(ish) proteins like those in the final 18. An alternative might have been glutathione-S-transferase (GST) fusions, another IPTG-inducible expression vector, because of its large tag (26kDa) and the GST-fusions high affinity for glutathione.

### **Adhesion assays**

Due to the persistent problems with pET expression, and the necessity of proving binding of at least some of the panning clones to their respective ligands in the time remaining for the project, it was decided to use adhesion assays, used previously with phage display libraries constructed from 4 members of the *Pasteurellaceae* by Mullen at the Eastman Dental Institute (Mullen *et al.*, 2007). Mullen used adhesion assays with success against 10 different ligands to quantify the affinity of certain fusion proteins for specific ligands. Those recombinants with the strongest affinity for a ligand bound in numbers exceeding 25,000 phage per well. Those ligands which did not demonstrate affinity bound in numbers of around 100 bound phage per well. One of the most exciting aspects of adhesion assays is that, as in the Mullen study, recombinants can be tested against similar ligands to determine binding specificity. For example, the recombinants identified from panning against FN were tested more thoroughly by Mullen against the 30 kDa, 45 kDa and 110 kDa fragments of FN, as well as the whole molecule (~440 kDa), and similar molecules such as fibrinogen. Such experiments provide a great deal of additional information about fusion proteins which have not yet been characterised and could provide useful information in the present study, for example, on the high number of recombinants showing affinity to IgA which could also be tested for affinity to IgG and other immune factors.

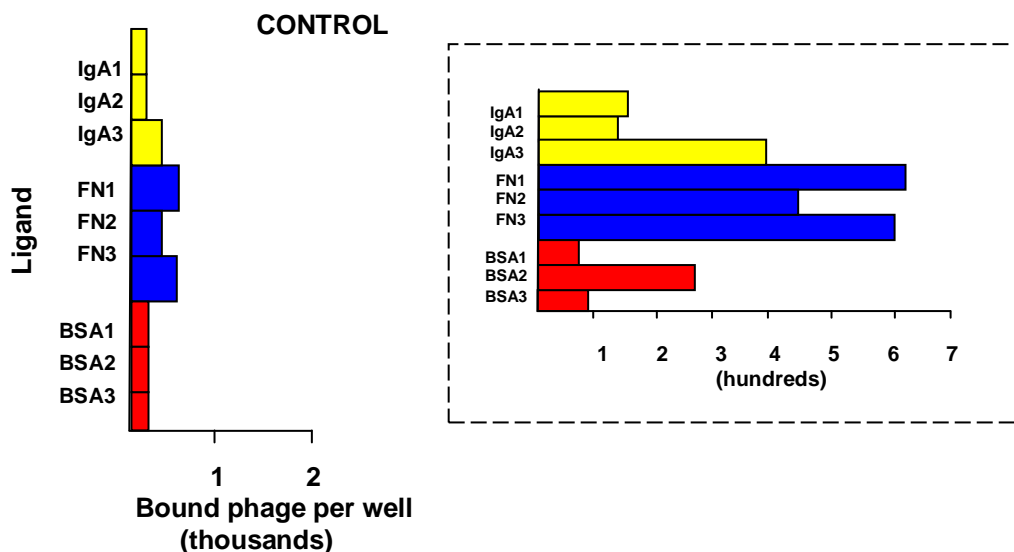
The purpose of adhesion assays is that upon incubating a pure sample of phage displaying the fusion proteins of interest with a variety of separate ligands, obvious and strong

affinity should be shown to one in particular (preferably the ligand identified from the panning process). Indeed, the adhesion assay procedure is very similar to that of panning. Briefly, Nunc Immunoplates were coated with 100µl 0.1 mg/ml ligand and left to bind overnight. Following removal of the ligand and thorough washing a phage population were added ( $1 \times 10^9$  per well) and left to bind to the immobilized ligand for 90 minutes. Following excess phage removal and thorough washing steps, the bound phage – still infective – were eluted and used to infect *E. coli*, then equal volumes plated onto selective agar. Counting colonies resulting from the adhesion assays gave an indication of binding affinity to the ligands chosen; in this project IgA, FN and BSA. Adhesion assays were carried out as follows: Nunc Maxisorp plates coated with immobilized ligand were thoroughly rinsed and 100 µl of  $1 \times 10^9$  recovered phage was added. After incubation and rinsing, bound phage were eluted in low pH glycine buffer and diluted for adding to log phase *E. coli*. Infections were incubated for 30 minutes before plating on NB2 agar containing ampicillin and bound phage per well calculated from colonies following overnight growth.

The adhesion assay process was not as straightforward as initially imagined. Firstly, each clone population required transformation, conversion to phage, amplification, PEG precipitation and recovery in order to reach an adequate volume of fusion phage at high concentration. From the list of 18 clones, 30, 39, 44 and 59 were amplified and recovered to  $1 \times 10^{11}$  CFU, then diluted to a concentration of  $1 \times 10^9$  CFU required for adhesion assays. Other clones 36, 1, 42 and 58 either did not transform well, or were unsuccessful at the conversion stage. Additionally, no blocking agent was used in the first or second adhesion assay; under normal circumstances BSA would be used, however some interesting clones from the panning experiment bound to BSA so it was decided to use it as one of the ligands for study, instead of the control it was intended as. The inclusion of a blocking agent is discussed later.

In the first adhesion assay all agar plates (showing bound phage per well) showed the same number of colonies for all ligands and all dilutions. This was caused by *E. coli* cells which developed ampicillin resistance, which happened in the first adhesion assay and again in the second. There seemed to be no obvious explanation for this resistance, since the *E. coli* cells used were isolated from an ampicillin sensitive stock which had been used previously in multiple experiments and all aliquots stored at  $-80^\circ\text{C}$ .

Recognising continuity issues but in order to carry out a repeat experiment in the short time remaining, strain JM107 (also F') was swapped for ampicillin sensitive *E. coli* TG1 cells for phagemid infection. At least one clone identified from each panning ligand was tested in the adhesion assays alongside the empty phagemid vector pG8H6 as a control. Using the above protocol and the 4 phage supernatants 30, 44, 59 and 39, the results are summarised in the following figures; **Figure 3**, **Figure 4**, **Figure 5** and **Figure 6**, with the control results shown in



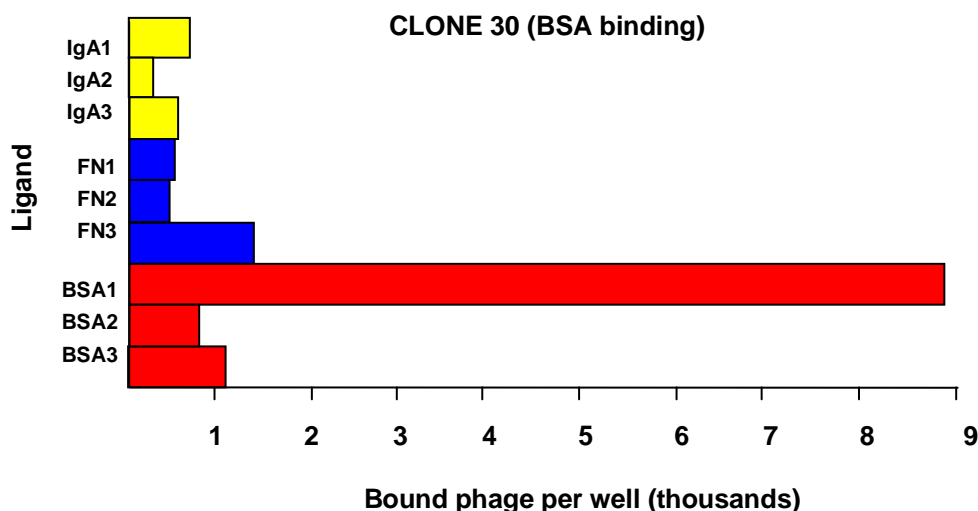
**Figure 2** Adhesion assay results from control experiment. The phagemid pG8H6 with no insert was used to test basal binding levels to each ligand; IgA, FN and BSA. For each ligand experiments were carried out in triplicate, numbered 1, 2 and 3 on the chart. For ease of comparison, bound phage per control well are presented in thousands (main picture) and also in hundreds (in box), due to the low control numbers. The number of bound phage per well is calculated from plating neat and 10x diluted samples onto NB2 agar + ampicillin.

**Figure 2.**

**Figure 2** shows the adhesion assay control results. All the following figures use the same scale (bound phage per well in thousands), however because the control results were low, it was thought appropriate to also present them on a scale showing bound phage in hundreds. From the figure it is clear that basal binding of filamentous phage without fusion proteins is very low, although the interaction of phage with the ligand FN was a little above the average. Taking an average of these control figures shows basal IgA binding at 320 bound phage per well, FN binding at 550 and BSA binding at 160 bound phage per well. These figures were averaged because the three experiments showed similar numerical results. This was not the case in some of the other adhesion assays discussed over the remainder of this chapter.

**Figure 3** shows the results of testing the binding levels of clone 30 to the 3 ligands used in adhesion assays; IgA, FN and BSA. Because clone 30 was originally identified binding to BSA in the panning experiments (Chapter 4), it was expected to bind preferentially to BSA in this experiment.

From **Figure 3**, the highest number from the BSA wells (BSA 1) is almost 8 times higher than the other 2 wells, which both show a similar binding affinity to each other, slightly higher than that of clone 30 for FN. The obvious outlier in this case is BSA1 with almost 9000 bound phage per well. Disregarding BSA1 for the moment and taking only BSA2 and BSA3

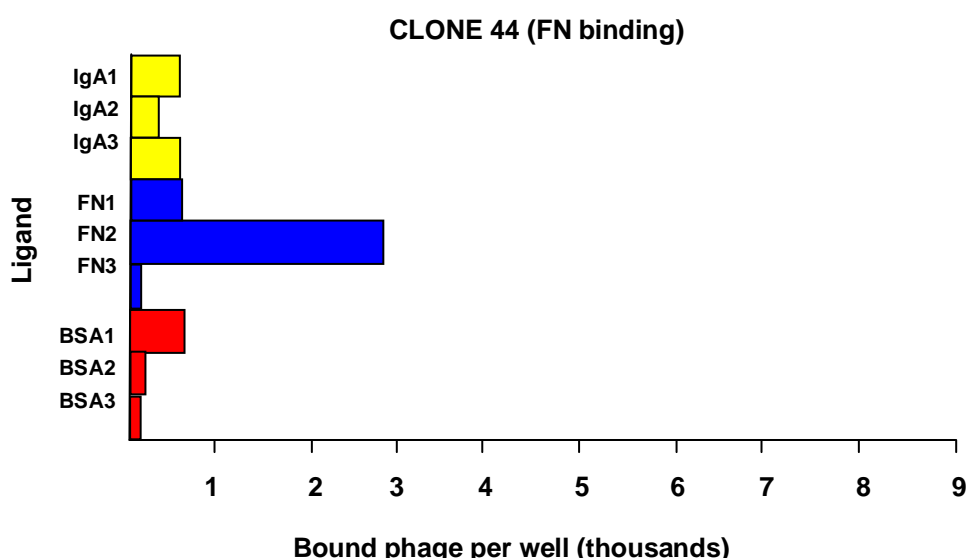


**Figure 3** Adhesion assay results with clone 30. Binding levels of clone 30 was tested against each ligand; IgA, FN and BSA. For each ligand experiments were carried out in triplicate, numbered 1, 2 and 3 on the chart. The number of bound phage per well is calculated from plating neat and 10x diluted samples onto NB2 agar + ampicillin.

into consideration, one could say that clone 30 binds better to BSA than to either IgA or FN, and that clone 30 clearly binds better than the control which had an average of 160 bound phage per well compared to a minimum of 800 in experiment BSA 3. These numbers may not prove that clone 30 is a BSA binding protein but they do strengthen the case for further investigation of this clones specific binding capacity.

The aberrant numbers of binding clones, such as that seen in BSA1 in **Figure 3**, could be due to the tendency of phage to aggregate in high numbers. Filamentous phage are extremely long and thin, and these hair-like structures can align closely in a linear fashion, groups of which could easily be picked up with a pipette and end up in one of the experimental wells. Combined with the sensitive titre method used to detect even low phage numbers, this could easily result in some quantitative errors. The structure of FN could also explain why filamentous phage appear to ‘stick’ to it. FN has already been discussed in Chapter 1 as a well known ligand for bacterial adhesion. This large, multi-domain protein is made up of 29-31 modules of 3 types of molecules stacked end to end, and is well placed to act as a ligand for all sorts of interactions. For this reason, adding these long structures, FN and filamentous phage, together, even if binding does not take place there could still be a degree of entrapment evident in the experimental numbers, and this is what is believed to be the cause of the elevated control numbers to FN. This experiment should be repeated with some minor variations to reduce or prevent phage aggregation, which are discussed at the end of this chapter; however there was no time to make these amendments within the timeframe of this project.

Clone number 44, which contained a putative FN-binding protein (**Figure 4**), bound in similar numbers to IgA (average ~ 450 bound phage per well) and BSA (average ~ 400). Again, there is a very high figure (FN2) in the experimental numbers, although the difference between highest and lowest is not as much as noted previously in **Figure 3**. If the experiment FN2 is discounted clone 44 would then appear to bind all 3 ligands in quite similar numbers. When compared with control results, clone 44 does not appear to bind FN or BSA any better than the control did, but does appear to bind IgA twice as strongly as the empty vector, and this potential binding affinity could be investigated further with additional adhesion assays incorporating changes discussed later.

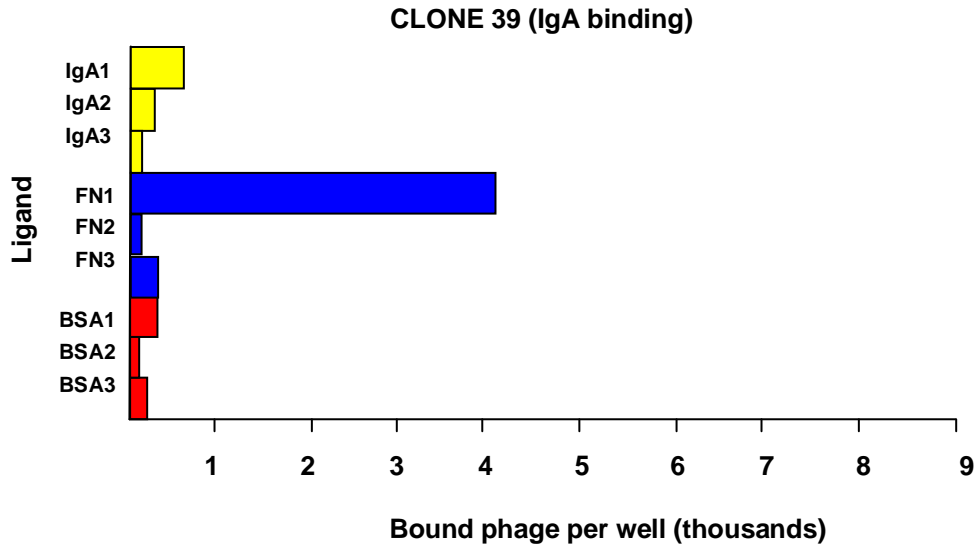


**Figure 4** Adhesion assay results with clone 44. Binding levels of clone 44 was tested against each ligand; IgA, FN and BSA. For each ligand experiments were carried out in triplicate, numbered 1, 2 and 3 on the chart. The number of bound phage per well is calculated from plating neat and 10x diluted samples onto NB2 agar + ampicillin.

Fusion proteins displayed on the phage surface that do not bind back to the respective ligand in adhesion assays can be explained like this: phage enrichment takes place on the promise that bound phage will remain bound during the washing process of panning prior to phage elution. Phage that are specifically bound, but not tightly bound (perhaps through low affinity for the ligand, a low expression level, or a low number of fusion proteins on the surface) will be removed during the washing steps, leaving phage which may not be bound through a specific interaction but which are bound to the ligand through the presence of many fusions on the phage surface (avidity effect).

Clones 39 and 59 were both identified from IgA panning experiments and show very high identity to proteins from organisms normally resident in the oral cavity; *Veillonella dispar*

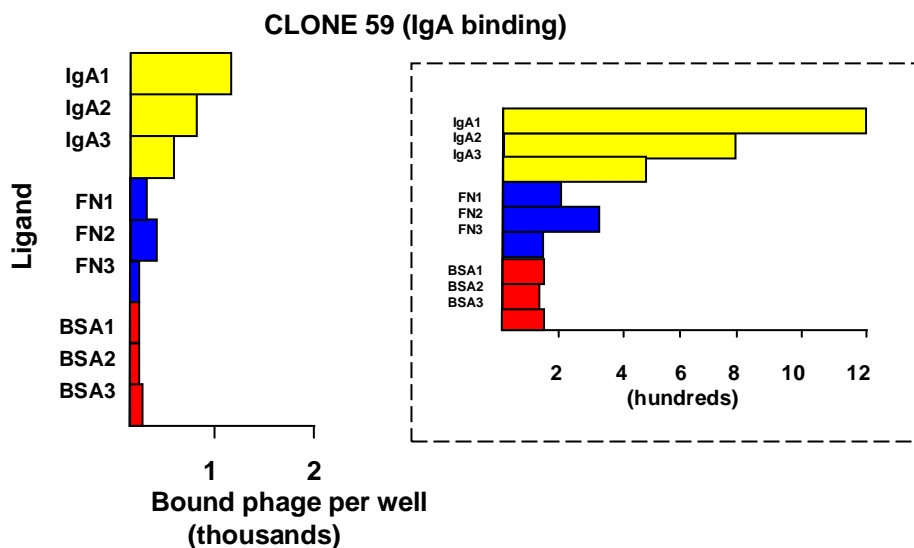
(96%) and *Fusobacterium nucleatum* (99%) respectively (**Table 2**). However, the adhesion assays showed that perhaps the binding affinity of clone 39 was not as strong as its recovery from the third round of panning suggested (**Figure 5**).



**Figure 5** Adhesion assay results with clone 39. Binding levels of clone 39 was tested against each ligand; IgA, FN and BSA. For each ligand experiments were carried out in triplicate, numbered 1, 2 and 3 on the chart. The number of bound phage per well is calculated from plating neat and 10x diluted samples onto NB2 agar + ampicillin.

Clone 39 bound to IgA in very similar numbers (average ~ 350 bound phage per well) as the control pG8H6 to IgA (average ~ 250), indicating no real binding affinity for IgA at all. Although there is an instance of apparently strong binding to FN1, as previously discussed, this is more likely to be due to phage aggregation, since the other two FN wells showed consistently low binding affinity. BSA and FN (excluding FN1) binding numbers for clone 39 are on a par with those in the control experiment so it appears that clone 39 binds no better than the control to any of the ligands.

In contrast, clone 59 (**Figure 6**) bound to FN, BSA and the control pG8H6 in very similar numbers, but importantly, bound in consistently higher numbers to IgA. In order to show the binding of clone 59 more clearly, bound phage per well have also been presented on a hundred scale. This experiment is interesting because the average bound phage to the FN wells is slightly lower than in the control experiment, and bound phage to IgA wells is almost 3 times higher than the control results (average of 800 bound phage per well compared to 320). The consistency of the results for this experiment also makes them seem more believable as there



**Figure 6** Adhesion assay results with clone 59. Binding levels of clone 59 was tested against each ligand; IgA, FN and BSA. For each ligand experiments were carried out in triplicate, numbered 1, 2 and 3 on the chart. The number of bound phage per well is the average calculated from neat and 10x diluted samples onto NB2 agar + ampicillin.

are no extreme figures present to skew the results. For these reasons, clone 59 is the most reasonable suggestion for closer analysis of binding affinity. A closer analysis of clone 59 was carried out using SignalP of expasy tools to search for a signal sequence, and an InterProScan search of the protein sequence to identify any conserved topological domains (**Figure 7**). The SignalP program identified a predicted signal peptide (probability 0.991) for cleavage at amino acid residue 9. The InterProScan analysis located the signal sequence along with one transmembrane domain and a chloride channel protein. The InterProScan analysis for all 18 clones is in **Appendix 8**.

unintegrated		
<a href="#">PTHR11689:SF2</a>		CHLORIDE CHANNEL PROTEIN 5 LONG ISOFORM
SignalP		signal-peptide
tmhmm		transmembrane_regions

**Figure 7** InterProScan result from analysis of the protein sequence of clone number 59.

Even though all the adhesion assays were performed in triplicate, at the same time using the same buffers, the presence of outlying experimental numbers did plant seeds of doubt that the results could be fully trusted. On the other hand some affinity clearly was shown,

especially in the cases of clone 30 and clone 59, for the ligands which identified them in the panning experiments, and these clones could be investigated more thoroughly for binding affinity and specificity. Clone 39 would possibly be best left aside as it appears not to show affinity to IgA however clone 44 could be tested for affinity to IgA, instead of FN.

In cases where there are aberrant numbers of bound phage per well, this could be remedied by a variety of approaches. Firstly, fewer phage could be applied to the well; a lower phage titre means that there is less opportunity for aggregation, and this alteration alone could give more consistent results. Secondly, adding more Tween-20 to the buffer used for phage dilution could help to disperse some of the phage aggregates already present and prevent others forming. A similar approach is to use a higher concentration of salt in the dilution buffer to prevent and disperse phage aggregates.

The value of using monovalent phage has already been discussed (Chapter 4) as having only one fusion protein on the phage surface allows binding affinity to be separated from avidity; the effect of multiple bonds on binding strength. Although it may not have any effect on phage aggregation in solution, monovalent display might help to clarify binding affinities for a particular fusion protein for different ligands. In binding clones, such as 59, which show clear preference for one ligand, the fusion protein could be excised from the pG8H6 vector and used in a gene III vector, which would allow testing of the binding affinity of between 1-5 fusion proteins. With polyvalent phage display (as with pG8H6), there is no way to tell how many fusion proteins each phage produces, or whether that number varies within a phage population, and by how much. Another available option, time permitting would be GST fusions, commonly used to confirm suspected interactions between a probe protein and unknown targets. This approach was used successfully by colleagues at the Eastman Dental Institute (Mullen *et al.*, 2008) however there was not enough time to integrate GST fusions into this project.

Although adhesion assays are useful for determining the affinity of fusion proteins for a certain ligand, they are less useful for determining binding strength when using phagemid where proteins can appear on the phage surface more than once. Using a phage display vector which allows protein display in a polyvalent fashion means that direct quantification of binding strength cannot be calculated, unlike using gene III for monovalent display which, in reality can allow between 1 and 5 fusions, depending on the vector. Because of this uncertainty in the number of fusion proteins on a phage and in a phage population, the binding affinity of each fusion protein is most effective when compared against ligands, and not against other fusion phage populations (Deshayes *et al.*, 2002). That is why one cannot say, for example, that fusion protein 59 binds more strongly to IgA than clone 30 because there may be more fusion proteins displayed on the surface of clone 59 than clone 30, so the apparent 'high affinity' is purely due



to a higher avidity effect of multiple fusion proteins. Similarly, at first glance of the pictorial results clone 44 appears to have very weak binding affinity for any of the ligands but it would not be a first choice for further investigation because it bound to the ligands in lower numbers than the other clones, because this could be due to fewer fusion proteins presented on the phage surface. It would not be first choice for further investigation simply because it did not show a clear preference for one ligand over the others.

### Individual Proteins

Whilst attempting to find out more about the physical properties of the fusion proteins, more research was carried out on the homology of these proteins to others in the public databases, which are updated on a daily basis. To do this, the BLASTp function of the NCBI website was used, which is a protein – protein homology database. In addition to locating potential protein functions, it was useful to find homologous sequences in the database to the final 18 clones, as none had shown any homology to BLASTn, a nucleotide – nucleotide comparison. The results of this analysis are shown in **Table 2**.

<i>Clone number</i>	<i>Size of predicted protein (amino acids)</i>	<i>Residue identity (%)</i>	<i>Positives (%)</i>	<i>Predicted protein product</i>
36	105	49/102 (48%)	67/102 (65%)	Putative OM haemoglobin receptor [ <i>Neisseria meningitides</i> ]
59	100	99/100 (99%)	100/100 (100%)	chloride channel protein [ <i>Fusobacterium nucleatum</i> ]; Gene ID:992636 FN 1727
1	83	62/62 (100%)	62/62 (100%)	hypothetical protein ACTODO_01254 [ <i>Actinomyces odontolyticus</i> ATCC 17982]
39	83	81/83 (97%)	82/83 (98%)	hypothetical protein VEIDISOL_00956 [ <i>Veillonella dispar</i> ATCC 17748]
42	84	76/80 (95%)	77/80 (96%)	hypothetical protein NEISUBOT_00167 [ <i>Neisseria subflava</i> NJ9703]
58	97	96/97 (98%)	96/97 (98%)	radical SAM domain-containing protein [ <i>Fusobacterium</i> sp. 2_1_31]
44	79	-	-	No similarity to protein database
52	79	60/64 (93%)	62/64 (96%)	ABC-type dipeptide/oligopeptide/nickel transport system, permease component [ <i>Veillonella parvula</i> DSM]

22	76	75/76 (98%)	75/76 (98%)	Sell repeat family protein [ <i>Neisseria flavescens</i> SK114]
30	76	18/29 (65%)	24/29 (82%)	hypothetical protein ACTODO_00462 [ <i>Actinomyces odontolyticus</i> ATCC 17982]
60	73	63/73 (86%)	69/73 (94%)	ABC transporter permease protein [ <i>Streptococcus gordonii</i> str. Challis] GENE ID: 5598624 SGO_0857
17	65	33/63 (52%)	46/63 (73%)	phosphate ABC transporter permease [ <i>Clostridium botulinum</i> B1 str. Okra] GENE ID: 6148283 CLD_0364
19	64	55/63 (87%)	59/63 (93%)	hypothetical protein ACTODO_00837 [ <i>Actinomyces odontolyticus</i> ATCC 17982]
20	53	50/53 (94%)	52/53 (98%)	IISP family type II secretory pathway protein SecA [ <i>Actinomyces odontolyticus</i> ATCC 17982]
27	57	-	-	No similarity to protein database
2	59	30/40 (75%)	34/40 (85%)	glycosyltransferase family alpha- glycosyltransferase [ <i>Parabacteroides</i> <i>distasonis</i> ATCC 8503]; GENE ID: 5307154 BDI_2004
16	57	47/57 (82%)	52/57 (91%)	predicted PP-loop superfamily ATPase [ <i>Haemophilus influenzae</i> R3021]
11	51	40/51 (78%)	42/51 (82%)	hypothetical protein PREVCOP_00536 [ <i>Prevotella copri</i> DSM 18205]

**Table 2 Predicted protein product of final 18 recombinant clones from panning which showed binding affinity to ligands IgA, FN or BSA. This table shows the homology between the proteins in the most likely reading frame (dictated by tBLASTx) to those in the protein database. Recombinant clones are in order of priority for pET vector expression or adhesion assay experiments. Search results accurate at 29<sup>th</sup> June 2009.**

**Table 2** includes 7 clones which contained hypothetical proteins from three of the well known oral microflora such as *Actinomyces*, *Veillonella* and *Prevotella* sp. It was thought that, through the use of pET vector expression systems and adhesion assays, more information could be gained on the specific ability of these proteins to facilitate bacterial binding to the tongue dorsum. In particular, three clones (1, 39 & 42) were picked up from panning against IgA, the

affinity between the tongue bacteria and IgA being of particular interest to this group. The table also contains two clones which showed no similarity to the protein database (BLASTp), or the nucleotide database (BLASTn). One of these, number 27 (BSA binding), showed no similarity at all to tBLASTx (translation into all 6 reading frames), a more stringent checking system, where number 44 (FN binding) contained a small region with homology to an aspartate binding domain.

What is also interesting from the table is the presence of 4 hypothetical proteins, with reasonable similarity and identity scores, supposedly from the oral commensal *Actinomyces odontolyticus*. According to Hallberg *et al* (1998), *A. odontolyticus* is the most prevalent member of the *Actinomyces* sp. on the tongue surface. The Hallberg group tested the binding of *A. odontolyticus*, and *A. naeslundii* genospecies 1 and 2 to N-acetyl-[beta]-D-galactosamine and acidic proline-rich proteins and found that both genospecies 1 and 2 of *A. naeslundii* bound N-acetyl-[beta]-D-galactosamine with 100% efficiency where *A. odontolyticus* showed only 10% efficiency. Upon binding to acidic proline-rich proteins, *A. naeslundii* genospecies 1 and 2 bound with 25% and 100% efficiency respectively, where *A. odontolyticus* bound with only 15% efficiency. The group observed that *A. odontolyticus* bound in 89% of cases to unknown structures on the surface of streptococci isolates, highlighting that those organisms which bind to buccal and dental mucosa exhibit alternative specificities to those that enable colonisation of the tongue surface (Hallberg *et al.*, 1998), suggesting that a great variety of adhesive mechanisms have yet to be discovered in this area, not only to human components. Although the fimbrial and host adhesive receptors of *A. odontolyticus* have not been fully investigated, *Actinomyces* sp. are well known for inter- and intra- generic coaggregations with streptococci, and interactions with the oral epithelia in general (Gibbons, 1989), and therefore it is not surprising to see the *Actinomyces* sp., and mainly *A. odontolyticus*, present in the current study alongside *Streptococcus* sp.

---

## **Chapter 6**

### **Results and Discussion:**

### **16S rRNA Gene Diversity Analysis**

---

## Introduction

For many years, much of microbiological diversity could not be accessed due to the inability of culture methods to sustain many bacteria. Since 1990, 16S ribosomal RNA (rRNA) gene analysis has become the single most common method of studying mixed microbial communities, and is often used to complement other molecular techniques such as cosmid (Courtois *et al.*, 2003) and fosmid libraries (Nesbo *et al.*, 2005), fluorescent *in situ* hybridisation (FISH; Harmsen *et al.*, 2002) and fingerprinting methods such as denaturing gradient gel electrophoresis (DGGE; Peterson, 2007) and terminal restriction fragment length polymorphisms (T-RFLP; Case *et al.*, 2007); the combination of which provides different viewpoints into community structure and function.

The inherent bias in using any PCR-based technology is well known, as is the case with many other techniques – whether culture based or not. 16S rRNA analysis is used principally to differentiate between operational taxonomic units (OTU's) and usually a similarity of 97-98% is deemed sufficient to identify two sequences as belonging to the same OTU (Gevers *et al.*, 2005), although even 99% identity can still represent sufficient diversity within a species (Case *et al.*, 2007).

In order to assess the diversity of the human tongue dorsum in a manner comparable to other studies (Kazor *et al.*, 2003; Aas *et al.*, 2005; Haraszthy *et al.*, 2007; Riggio *et al.*, 2008), 16S rRNA phylogenetic analysis was undertaken. In this project, it was felt that the phage display library was unlikely to have provided a representative view of the level of microbial diversity in the original sample, leading to partiality towards certain bacterial phylotypes or bacterial proteins. This chapter details the analysis of 16S rRNA genes from the human tongue dorsum and comparisons between the 16S rRNA and phage display analyses.

## Initial analysis

The 16S rRNA was PCR amplified initially using 30 cycles for testing purposes, then using 10 cycles for the 16S rRNA used for the analysis, described in Chapter 2, page 62. Briefly, 30 cycles were carried out using the twofold degenerate primer 27f-CM (5'– AGA GTT TGA TCM TGG CTC AG -3') and 1492r (5' - TAC CTT GTT ACG ACT T -3') (Frank *et al.*, 2008), and a gradient of annealing temperatures between 46°C and 54°C, based on the melting temperature of the primers, and the previous annealing temperature of 48°C used by Frank *et al.*, 2008. The resulting PCR products were checked using 1% agarose gel electrophoresis. The band showing a clear PCR product of the correct size (roughly 1500 bp) with high intensity was used for cloning using the TOPO TA cloning kit, according to manufacturer's instructions. Successful colonies were identified using blue/white screening and

21 white colonies were sequenced to confirm the presence of the 16S rRNA gene. Sequencing was carried out using the primers M13 forward (5'– GTT TTC CCA GTC ACG AC –3') and reverse (5'– GGA AAC AGC TAT GAC CAT –3') at a concentration of 5 µM. Sequence data was searched for homology to the Ribosomal Database Project sequence collection (Cole *et al.*, 2009) and the top match taken as the search result.

From the 21 white colonies, 4 were *Streptococcus* and 4 were *Veillonella* (2 *V. dispar* and 2 uncultured). Three were identified as *Prevotella* sp. and 3 as uncultured bacterium from unknown sources. Two were identified as *Haemophilus*, one uncultured and one *H. parainfluenzae*, and one each of the genus *Neisseria*, and *Clostridiales*, and one each of the phylotypes *Fusobacterium peridonticum*, *Rothia mucilaginosa* and *Bulleidia moorei*. The preliminary results showed the same types of bacteria in similar proportions to previous studies carried out on the tongue dorsum (Aas *et al.*, 2005; Kazor *et al.*, 2003) and on the strength of this data, the full 16S analysis of 380 clones was carried out.

### **Full 16S rRNA analysis**

Following the successful 16S analysis trial, a slight variation of the aforementioned PCR reaction was set up using 10 cycles instead of 30, with the aim of retaining as much of the original diversity as possible, at the optimum temperature established in the test (49°C). The PCR product was tested using 1% agarose gel electrophoresis and then cloned, as before, using the TOPO TA cloning kit, according to manufacturer's instructions. From the blue/white screening plates, 380 colonies were hand-picked and minipreped using the Qiagen Plasmid Spin Kit, and the DNA diluted to 10 ng/µl for 16S rRNA analysis, carried out at the Comparative Genomics Centre, part of University College London. The primers used for sequencing were M13 forward and reverse, detailed in the 'Initial Analysis' section. Primers were synthesized commercially by Eurofins MWG Operon ([www.eurofinsdna.com](http://www.eurofinsdna.com)) and were based on the primers tested by Frank *et al.*, 2008.

### **16S rRNA data analysis**

Once sequenced, the 16S forward and reverse files were aligned using BioEdit Sequence Alignment Editor v7.0.9 in FASTA format to form the contiguous 16S rRNA sequence of around 1592 bp. Contigs were then specifically aligned for use in the Greengenes database (DeSantis *et al.*, 2006a) using NAST (Nearest Alignment Space Termination) (Desantis *et al.*, 2006b), which outputs the MSA (Multiple Sequence Alignment) in the standard format of 7,682 characters per sequence, allowing similar loci to be located in similar positions in subsequent batches. During this analysis, 4 out of 380 sequences did not meet the

match requirements, being either less than the minimum length of 1250 bp, or sharing less than 75% identity to the template sequence. These sequences were removed from further analysis.

Due to the presence of genomic data from different origins containing the conserved 16S rRNA gene, amplification by PCR is known to introduce hybrid molecules which can distort the results of a diversity analysis (Liesack *et al.*, 1991). The percentage of chimeric sequences in this 16S rRNA gene analysis was therefore checked using the greengenes Bellerophon (version 3) server (Huber *et al.*, 2004). In other studies (Kazor *et al.*, 2003; Aas *et al.*, 2005) the number of chimeric sequences was between 1 and 15%. This analysis located 42 chimeric sequences (11%), which were below the 97% threshold BLAST similarity and contained fewer than 1250 matching base pairs to the Core Set of sequences. Species identification of chimeric sequences was not obtained and these 42 were removed from further study.

Using the NAST aligned sequences, minus chimeras; the remaining 333 sequences were classified using the Simrank interface which finds similarity between query and database in terms of the number of unique 7-mer count present in either query or database. Sequence diversion from near-neighbours was calculated using the DNAML option of DNADIST (PHYLIP package). The reference sequences used for classification were non-chimeric (divergence ratio <1.10) and taxonomic analysis was conducted using the RDP taxonomic nomenclature. In the identification of closest relatives, all sequences were compared to >100,000 sequences in the RDP database. The similarity cut-off used for species differentiation was 98%, and 16 of these clones did not meet this cut-off. Those 16 with a higher level of differentiation are shown on the phylogenetic tree by only the clone number (**Appendix 6**).

### **16S rRNA gene analysis results**

From the 333 sequences, the majority (60.4%) were from the phylum *Firmicutes*. The remaining classification revealed 19.8% *Bacteroidetes*, 12.9% *Proteobacteria*, 4.2% *Actinobacteria*, 2.1% *Fusobacteria* and 0.6% TM7, a novel phylum for which there are no cultured representatives. The classification indicated 26 genera in total and the two most abundant, both from the phylum *Firmicutes*, were *Veillonella* (26.7%) and *Streptococcus* OTU's (24.6%). Although the oral cavity is known to harbour Archaea, most commonly of the genus *Methanobrevibacter*, PCR amplification of Archaeal rDNA requires specific primers which were not included in this analysis (Lepp *et al.*, 2004).

### Novel bacteria

During this analysis, 15 clones fell short of the 98% similarity cut-off, and in this situation, these clones were tentatively classed as potentially novel phylotypes. These 16 clones are analysed in more detail in **Table 1**.

In all BLAST searches where the top match did not suggest a species name for the sequence, the second match did. In clone 9 the highest species homology was to *Rothia mucilaginosa*; clone 60 was most similar to *Haemophilus parainfluenzae*; clone 233 was most similar to *Prevotella* ‘Oral Taxon 299’; and clone 307 was most similar to *Streptococcus parasanguinis*. Clone 327 however showed very low homology to an uncultured *Veillonella* species, but this similarity was at least half way down the BLAST ranking table. This discovery could make clone 327 of particular interest as perhaps an as yet uncultured *Veillonella* species, or an unknown member of a related species.

From the RDP classification, clones number 22, 23 and 70 (in bold typeface) all have between 92 – 98% homology to the same uncultured bacterium clone FIU\_KM\_MD\_004, which shares some similarity with an oral *Prevotella* species. An alignment of these 3 rRNA sequences using clustalw showed that, although showing homology to the same entry in the BLAST database and therefore potentially from the same organism, the sequences do not completely align. Out of the 1500 base pairs aligned, there are 132 bp (8.8%) which do not align between clones 22 and 23, and 134 bp (8.9%) non-alignment between clones 23 and 70, whereas clones 22 and 70 appeared more closely related with only a 103 bp non-alignment (6.8%). Of course, one can not tell from this information whether or not clones 22, 23 and 70 belong to the same organism or species but given the positioning of these 3 clones on the phylogenetic tree (**Appendix 6**), it is probable that they all belong to the *Prevotella* sp. Given that clones 22, 23 and 70 were sufficiently different, they were submitted to Genbank as separate entries.



<i>Clone number</i>	<i>Number of bases available for BLAST</i>	<i>Matching bases</i>	<i>Sequence identity (%)</i>	<i>Accession number</i>	<i>Closest species match</i>	<i>Additional information regarding environmental origin</i>
9	1671	1461/1510	96.7	GQ398416	Uncultured bacterium clone A_S_01_77	Healthy Chinese oral cavities FJ470589
<b>22</b>	<b>1611</b>	<b>1428/1505</b>	<b>94.9</b>	<b>GQ398417</b>	<b>Uncultured bacterium clone</b>	Clone from cystic fibrosis patient EU670056
<b>23</b>	<b>1613</b>	<b>1402/1510</b>	<b>92.8</b>	<b>GQ398418</b>	<b>As 22</b>	As 22
45	1622	1465/1514	96.8	GQ398419	<i>Haemophilus</i> sp. oral clone JM053	
47	1608	1473/1510	97.5	GQ398420	<i>Prevotella salivae</i> strain EPSA11	
60	1629	1464/1522	96.2	GQ398421	Uncultured bacterium clone A_D_01_61	
71	1614	1464/1519	96.4	GQ398423	Uncultured <i>Porphyromonas</i> clone 302E06	
137	1598	1441/1506	95.7	GQ398424	<i>Prevotella</i> sp. oral clone ID019	
189	1647	1484/1538	96.5	GQ398425	<i>Streptococcus</i> sp. F1	
233	1601	1458/1493	97.7	GQ398426	Uncultured bacterium clone P1D1-678	Healthy Chinese oral cavities (as clone 9) FJ470432
273	1640	1472/1553	96.0	GQ398427	<i>Streptococcus parasanguinis</i>	
284	1584	1384/1495	92.6	GQ398428	Uncultured <i>Campylobacter</i> sp. clone 202B08  (oral)	

Chapter 6 – R & D: 16S rRNA Diversity Analysis

296	1621	1459/1519	96.1	GQ398429	<i>Prevotella</i> 'Oral Taxon 299'	
307	1642	1478/1545	95.7	GQ398430	Uncultured bacterium clone SJTU_F_10_28	Clones from human tracheal aspirates EF511999
327	1634	1482/1539	96.3	GQ398431	Uncultured bacterium clone Oh_3137A9A	Isolates of Bartonella positive fleas EU137432

**Table 1 Similarity of 15 'novel' phlotypes to public access database. Clones in bold typeface were thought to be homologues. Clones were submitted to GenBank using BankIt (<http://www.ncbi.nlm.nih.gov/BankIt/>) on 17<sup>th</sup> July 2009.**

## Discussion

The 16S rRNA gene contains a few fundamental properties which have facilitated its use as a conventional molecular marker such as its essential function, ubiquity and conserved nature. 16S rRNA analysis was included in this project for several reasons. Firstly, to assess the initial diversity of the bacterial metagenomic DNA used to make the phage display library. Secondly, it allows the use of that diversity as a benchmark for assessing the diversity of the phage display library. Thirdly, it allows a comparison between this and other similar studies on the human tongue dorsum. Lastly, it provides some context into the range of bacteria being identified by the panning process and allows speculation on phage display limitations. It should be noted that this study does not attempt to infer which phylotypes were the most prevalent (i.e., present in the most people), as most comparable studies have done. In this project, 16S rRNA analysis was simply carried out to try to gauge microbial diversity in the tongue metagenomic DNA samples used to create the phage display library. The expectation was that a representative analysis of the tongue microbiota by 16S rRNA analysis would provide some information on the types of bacteria that could be expected in the phage display library, and of course, highlight any notable exceptions.

The primers used in this study were tested previously by Frank *et al.*, 2008, where the 27f – 1492r primer set was found to bind commonly to most bacteria. The 27f primer has also been commonly used in many previous studies assessing tongue diversity (Kazor *et al.*, 2003; Paster *et al.*, 2006, Riggio *et al.*, 2008) and in those which assessed the specificity of certain primers to bacteria from different environments (Flanagan *et al.*, 2007). However, different primer sets were used by Aas *et al.*, 2005 (D88f and E94r) when assessing oral microbial diversity, and clearly primer choice will result in some bias in the resulting diversity analysis. When attempting to assess true diversity of a sample, it appears that culture dependent and independent methods should still be used in concert (Pratten *et al.*, 2003).

The bias towards certain species over others in a PCR-based analysis has been attributed to preferential amplification of low G + C templates (Polz, 1998). This intrinsic bias is reportedly reduced by using fewer amplification cycles (Suzuki, 2008), as any initial inequality in the sample, perhaps resulting from the DNA extraction method used, will be increased exponentially as the cycles progress. In order to retain as much of the original sample diversity through to the PCR end product, ten cycles were used in the current study as opposed to either 30 (Kazor *et al.*, 2003; Aas *et al.*, 2005) or 33 (Riggio *et al.*, 2008). Although the increase in preserved diversity is probably not large (Pratten *et al.*, 2003), in this project it has resulted in the location of 112 phylotypes from 5 bacterial phyla, with a further 16 phylotypes which showed < 98% similarity to the RDP database, and can tentatively be classified as novel. In a comparable study carried out by Kazor in 2003, ninety-two phylotypes were identified

from 6 bacterial phyla (they classed the Clostridia as a separate phylum), from a sample size of 11 patients, in comparison to 9 volunteers used in the present thesis. The number of species in the present study was 44, not including the additional 16 ‘novel’ clones, which may belong to one or more species. In similar studies, using 30 PCR cycles, the following numbers were observed from the tongue dorsum; 8 patients, 84 species (Haraszthy *et al.*, 2007); 5 patients, 39 species (Aas *et al.*, 2005); 32 patients, 78 species (Riggio *et al.*, 2008). Of course, the combination of primer set and number of amplification cycles, plus many other contributing factors, will cause some of these variations in final figures. From these numbers, the combination of the 27f-CM – 1492r primer set with a reduced number of 10 PCR cycles has allowed the identification of a comparable number of species on the tongue dorsum to that found in other studies.

In the present study the two most common genera, making up just over half of the total sample and both from the phylum Firmicutes, were *Veillonella* (26.7%) and *Streptococcus* (24.6%). The most common were 13 varieties of uncultured *Veillonella*, making up 18% of the total. This was followed by 18 varieties of uncultured *Streptococcus* at 12% and 6 uncultured *Prevotella* at 7%. The most common species found were *Veillonella dispar* ATCC 17748 (5.4%), *Neisseria mucosa* AJ239279 (4.2%), and an uncultured *Veillonella* Ax3\_690 (3.9%). All of the isolates found were generally accepted members of the normal human oral cavity.

Points of differentiation between the present study and others in terms of diversity are summarised as follows. In the study conducted by Riggio (2008), 5 – 10 % of clones were made up of various *Actinomyces* sp., where in the present study *Actinomyces* constituted 0.01% of all clones. This could be due to the primer set used by Riggio (27f and 1387r) which potentially recognises Actinomycete 16S rRNA better than the primers used in this thesis. They also found that between 7 and 12 % of clones were *Lysobacter*-type species, not previously found in great numbers on the tongue surface, and not present in the current analysis. They also found *Streptococcus salivarius* in 6% of clones and *Veillonella dispar* in 5.5%, the latter figure agreeing almost exactly with the present findings. The group also identified 9 known *Streptococcus* species, making up 13 - 16 % of total numbers, along with numerous unknown *Streptococci* making up a further 3 – 6 %. The present study identified 5 known *Streptococcus* species making up 8%, but the majority were unknown or uncultured species, making up 14.7%. The appearance of a higher percentage of unknown/uncultured species may be due to reduced stringency of the primers during PCR, or it could be due to the reduced number of amplification cycles used, which may suggest that the original environment contains many more unknown or uncultured species than first thought. The same scenario is true with the identified *Veillonella*: where the Riggio study located 4 known *Veillonella* species (between 11 and 14 % of total), and 1.5 % unknown species. In the present study, 3 known species (7.5 %) and 13 unknown or uncultured, making up 19.2 %, were identified. The differences in these

figures could be real or they could be due to reasons such as errors during sequencing. A similarity cut-off of 98% was used in this study which, in the 1250bp of sequence used for classification, is only 25 base pairs, so any sequencing errors at all could result in the *appearance* of higher sample diversity than was actually present. In addition, the melting temperature of the 27f-CM primer used in the present study was 54 °C, and the annealing temperature found to give the sharpest band of the correct size was 49 °C. The low annealing temperature in comparison to the high primer melting temperature could result in reduced specificity, as primers are known to hybridize to template DNA in conditions of lower stringency (Pratten *et al.*, 2003).

The results of the present study align with those of Kazor (2003), in that only 40% of clones were identifiable as known species. This is directly comparable with the present study where 37% (122/333) were identified as known species. Although the Kazor study was split into halitosis and healthy samples, by far the most common genus identified were various known and unknown *Streptococcus* species, with no *Veillonella* presence in the top 4 common species in either healthy or halitosis groups. This is where Kazor and the present study differ, as *Veillonella* was the most abundant species found here with over 26% presence. The Kazor study may not have picked up *Veillonella* species with great frequency due to the primers they chose, as a similar study by Tanner *et al.*, 2006, also used the same primers as Kazor and did not even place *Veillonella* in the 40 most prevalent species, their study being heavily biased toward *Streptococcus* sp.

The results described in this thesis are probably more representative of true microbial diversity at the tongue surface they are in agreement with culture based and DNA-DNA hybridization studies, such as that of Faveri *et al.*, 2006; Donaldson *et al.*, 2005; and Mager *et al.*, 2003, who identified *Veillonella* sp. as the most prevalent, and *Prevotella* as the second most prevalent. What is important to remember however is that the human oral microbiota varies from person to person, and also between geographic locations, depending on diet and various other factors. The abundance of one genera or species over another may simply be a manifestation of local variation.

One study which stands alone from all others in this field is that of Keijser *et al.*, 2008, who identified 19,000 phylotypes in the oral cavity as a whole, by analysing saliva and supragingival plaque from 71 individuals. This group targeted the 16S hypervariable region which is flanked directly by well conserved regions that can be used for PCR amplification. Using high throughput pyrosequencing which provides reads of around 100 base pairs, the group generated almost 200,000 reads and allocated a new phylotype to each sequence containing a single base discrepancy. A phylotype is defined, in 16S rRNA terms, as clones which have > 98.5% identity. Those with < 98.5% similarity to previously defined clones are considered to represent a new species (Paster *et al.*, 2006). The Keijser analysis was driven by

the hypothesis that even the now-common use of molecular techniques for assessing microbial diversity, such as 16S rRNA analysis, was still missing low abundance phylotypes, and set out to analyse ‘true’ microbial diversity by analysing thousands of clones. However, the problem is that a single base pair discrepancy does not automatically signify a new phylotype or species, and it is known that very similar bacteria can have significant differences between 16S rRNA sequences. It has been demonstrated previously, that high levels of 16S rRNA gene sequence similarity between strains belonging to the same genus do not indicate membership of the same species (Yassin *et al.*, 1996). Further, a single base discrepancy can easily be introduced through sequencing errors, leading to a misrepresentation in the studies final numbers.

In **Table 1**, the 16 ‘novel’ clones identified during 16S rRNA analysis are described in more detail. In this study so far, 63% of clones were from known genera but unknown species and 57% had never been cultured. These figures agree with other studies which have estimated the number of uncultured representatives in the oral cavity at between 50% (Paster *et al.*, 2001; Jenkinson & Lamont, 2005) and 60% (Kolenbrander *et al.*, 2002). There does not appear to be much information on the number of ‘novel’ species found on the tongue surface or oral cavity but the number of ‘novel’ species or phylotypes in this study could be as much as 4%. This number may seem rather low in comparison to other environments which appear to continually identify novel species, however given the amount of research carried out on the oral cavity microbiota, perhaps a figure of 4% is reasonable.

---

**Chapter 7**

**Additional Work: pQR492**

---

## Introduction

Due to the overwhelming complexity of natural microbial ecosystems, and the limitations posed by traditional culture-dependent analysis of micro-organisms, a variety of approaches are now routinely used which circumvent the need for bacterial culture. Currently, it is standard practice to interrogate a bacterial sample by sequencing the microbial genomes contained within it. The most straightforward way of doing this is to create a library of randomly fragmented DNA spliced into a cloning vector. These so-called ‘shotgun libraries’ are also useful because they allow the clones to be screened for specific traits, such as enzyme production, and clones which show interesting characteristics can then be sequenced.

In 2005, a *Bam*HI shotgun library was produced by Professor John Ward in the universal cloning vector pUC19, containing the fragmented microbial genomes of organisms from the human tongue. Upon sequencing the ends of 50 clones, they were used to query tBLASTx database for homologous sequences. This analysis revealed that some of the recombinants, in particular pQR492 and pQR494, contained DNA fragments worthy of closer analysis. From the BLAST comparison – carried out in 2005 - clone pQR492 showed homology to a putative glycopeptide synthesis gene from the predominantly oral genus *Actinobacillus* and pQR494 showed homology to a tetracycline resistance gene, a trait commonly found in the oral cavity (az-Torrez *et al.*, 2003). This chapter discusses work which was started at the very beginning of the project and picked up at various points during and at the end of the project, with the aim of obtaining full sequences of pQR492 and, time permitting, pQR494.

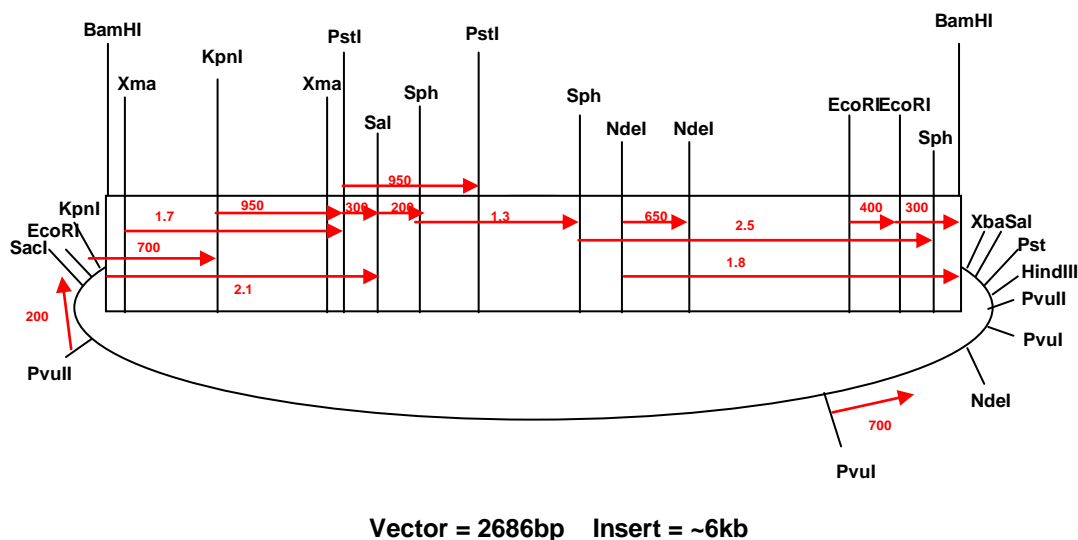
## Investigating pQR492

Clone pQR492 contains a ~6kb bacterial DNA fragment, inserted into pUC19 (2.6kb), using the *Bam*HI cloning site. Starting with around 700 base pairs (bp) of sequence from the initial sequence analysis, the NCBI BLASTn algorithm confirmed that the insert was indeed from a bacterial source (by not matching to any part of the human genome, which has been sequenced in its entirety), and also noted some homology to a potential glycopeptide synthesis gene using the tBLASTx search function. Because of this interesting potential function pQR492 was chosen for further analysis however the BLAST search carried out in 2005 did not produce the same match in a BLAST search carried out in July 2009. At this time, the first and last 800 base pairs of pQR492 showed homology to a transcription repair coupling factor and a putative membrane protein, and these similarities are discussed in more detail in this chapter.

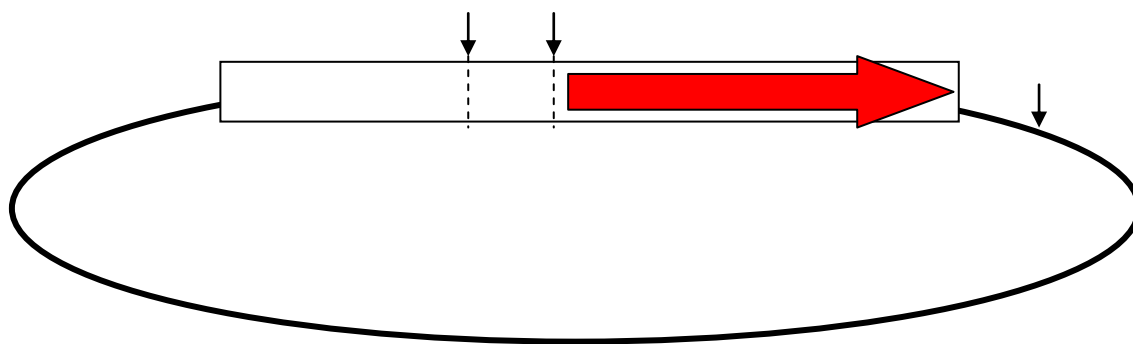
The first stage in analysing pQR492 in more detail was to produce a restriction map of the fragment. Briefly, pQR492 was transformed into *E. coli* Top10 F’ and the plasmid extracted



using the Qiagen Plasmid Maxi Kit according to the manufacturer's instructions, which provided 1 ml of plasmid DNA. Sequencing returns approximately 800 bp of good quality DNA sequence each time, so a basic restriction map was constructed using single and double digests (**Figure 1**), and plasmids where deletions were created then the deleted recombinant sequenced. During this process it became clear that most restriction sites were clustering to the left side of the insert, leaving big gaps to the right. In order to access sequence at the right side, subcloning was introduced at that end only. From **Figure 2**, three *PstI* sites are clear, 2 in the insert itself, 950 bp apart, and one in the vector just after the end of the insert sequence, so the *PstI* restriction sites were used to separate the 6 kb insert into 3 fragments. To do this, 300  $\mu$ l of plasmid preparation was digested using the restriction enzyme *PstI* at 37°C for 3 hours to ensure complete digestion, which cut pQR492 into 3 easily identifiable fragments; 950 bp, 3.5 kb and just over 4kb (**Figure 2**). The digested plasmid was then run on a 0.8% agarose gel, and the 3.5 kb DNA fragment excised and extracted using the Qiagen Gel Extraction Kit, according to the manufacturer's instructions.



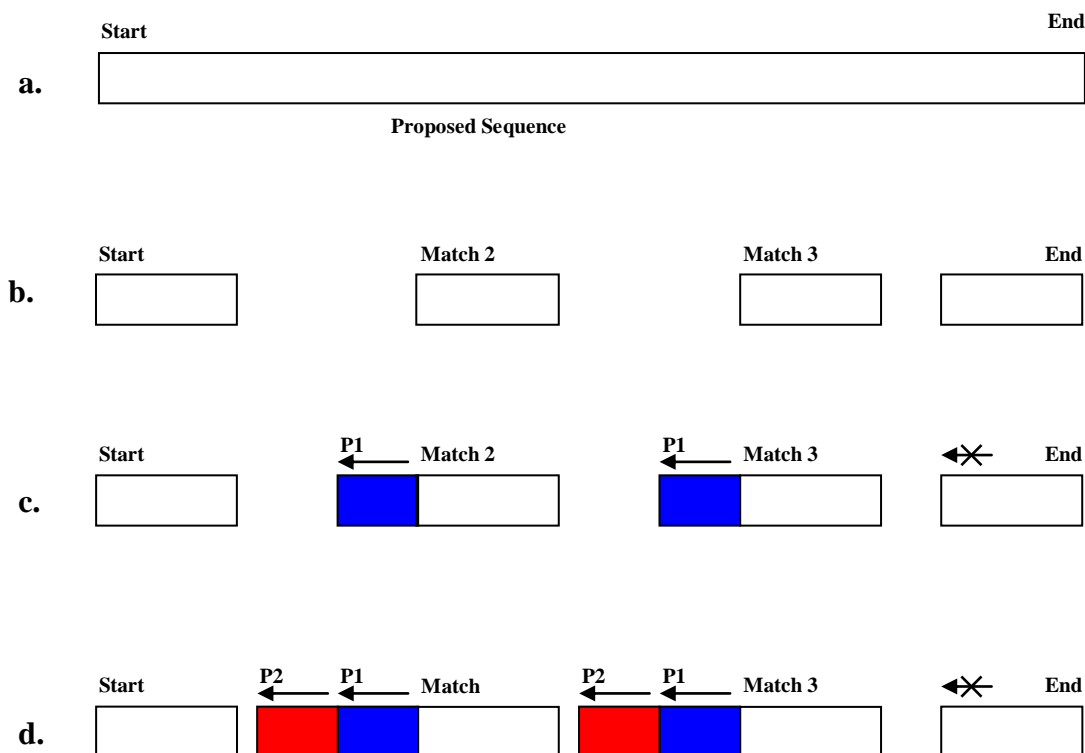
**Figure 1** pQR492 restriction map. Using single and double digests, and creating deletions within the insert with sequencing, this restriction map was compiled. The arrows denote the distance between restriction sites at the beginning and the end of the arrow.



**Figure 2** pQR492 showing *Pst*I sites (black arrows) used for extraction of a 3.5 kb part of the insert needed for subcloning. Red block arrow illustrates target DNA sequence for separate analysis.

Following extraction, the 3.5 kb fragment was split into 3 aliquots and digested individually using the restriction enzymes *Hae*III, *Alu*I and *Sau*3A, resulting in DNA fragments of around 50 – 100 bp, which were cloned back into pUC19. Thirty-six of the resulting recombinants were sequenced and the sequences aligned using multiple alignment software (Geneious), by a visiting Nuffield student who produced the resulting alignment ‘match 3’, a contiguous alignment of 900 bp to the right of the insert. All of the preceding and subsequent analysis of pQR492 was carried out by the author. The situation at this stage is shown in **Figure 3**, diagram b, which illustrates all sequenced parts of the 6 kb insert to that point. Around 700 bp at the start and end of the insert came from the initial exploratory sequencing which checked whether the recombinant was worth studying further. ‘Match 2’ was a contiguous sequence of around 1.8 kb from the restriction mapping exercise, placed around the left hand side of the 6 kb fragment. ‘Match 3’ was known to be from the right hand side of the fragment, although where exactly was not known. ‘Match 2’ did not align with the start sequence or with ‘match 3’ and ‘match 3’ did not align with the end sequence. The non-alignment was mainly due to a shortage of overall sequencing of the 6 kb fragment, having only reached 3850 bp of sequence at this time. At this point it was decided that primer walking was the best approach to close the gaps in this sequence.

When attempting primer walking for the first time primers were designed from both ends of each aligned fragment of DNA. As an example, primers were designed for the left hand side of ‘match 2’ coming out towards the start and into the alignment, and also from the right hand side of ‘match 2’ out towards ‘match 3’ and into the alignment. However, only the reverse complement primers were ever successful. It was supposed that primer binding failure was due to errors in the 6 kb insert sequence, leading to the design of ineffectual primers. This approach gave the scenario depicted in **Figure 3**, diagram c which, though the primers produced a further 2 kb of sequence, still did not allow alignment between the start, end, or either ‘match’ sequence. Primer walking was continued and the new sequences added together until all DNA fragments aligned, leaving the insert size at 6,318 bp, agreed by agarose gel electrophoresis.



**Figure 3 Sequencing of pQR492 insert DNA.** a) full 6kb sequence, b) areas of alignment (match 2 and match 3) along with the start and end sequences (3850bp in total), c) primer walking (blue area) added 2080bp (5930bp total), d) further primer walking from the blue sequences (in red) added 1796bp (7726bp total). The insert was known from agarose gels to be 6kb in size so the lack of alignment was thought to be due to some bad quality sequence at the C terminal end of match 3, match 2 and start.

### *In silico* analysis

With the full sequence of pQR492, the whole 6 kb was used to check homology to sequences held in public databases. An NCBI BLAST search using both BLASTn and tBLASTx showed no significant similarity with the full length sequence. Since the full sequence did not generate any ‘hits’ with tBLASTx, possibly because using the full 6.3 kb made any short areas of homology seem insignificant, it was split into 4 blocks of 1575 bp which were used separately to query the database. The 4 shorter fragments did match others in the database and the results are shown in **Table 1**.

The first quarter of pQR492 shows homology to a different protein than the remaining three quarters. One could assume that this was due to the joining of insert fragments during the ligation process, prior to cloning into pUC19 – similar to that seen in the phage display library - but if this was the case an internal *Bam*HI site would indicate the point of ligation since this restriction enzyme was used to make the library. Instead, the general low level homology to the organisms *Arthrobacter* sp. and *Corynebacterium aurimucosum* could indicate that the

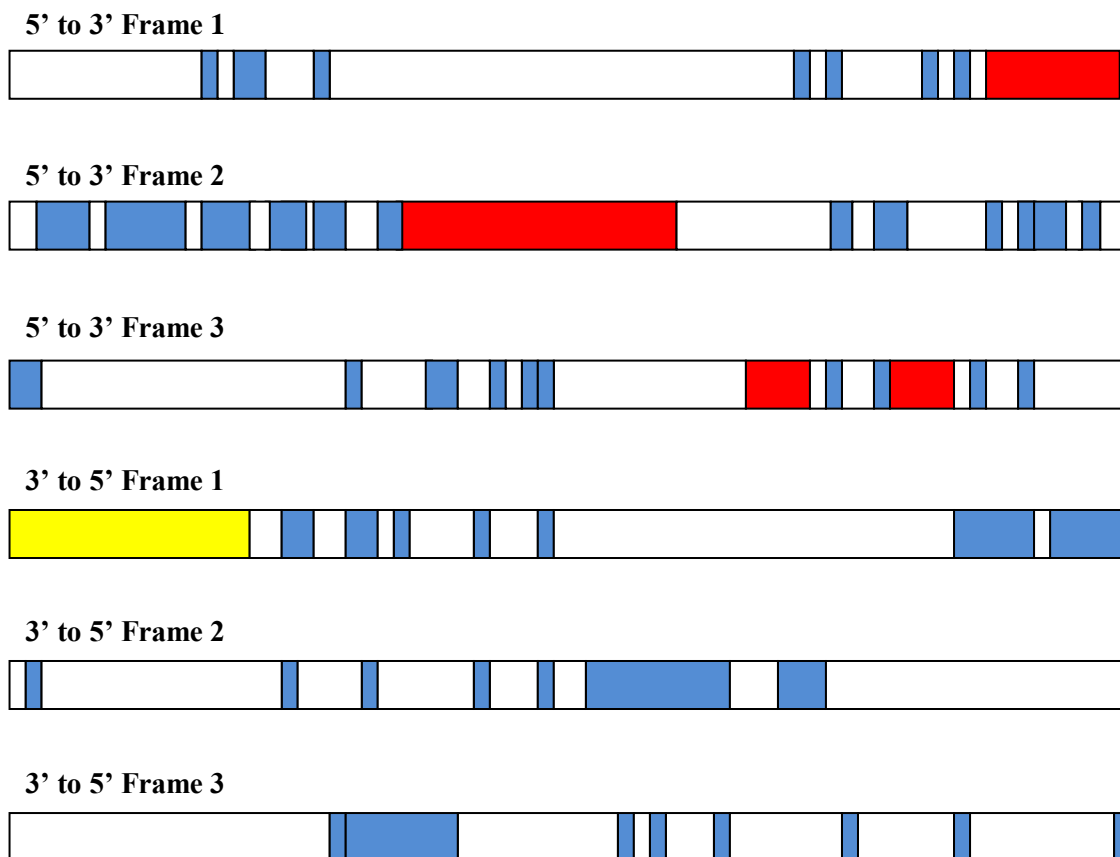
fragment in pQR492 is similar to both organisms but is actually neither. The fact that homology exists to these two high G+C Gram positive organisms however does indicate that the origin of pQR492 is also likely to be a high G+C Gram positive organism.

<i>Insert part</i>	<i>Description of tBLASTx hits</i>
0 – 1593 bp	Arthrobacter sp. FB24, transcription repair coupling factor, identity 65/97 (67%), Positives 82/97 (84%), 3e-97, 3-4 large (100aa) matches over the query sequence.
1594 – 3177 bp	Several very low identity (35%) short (50 – 80 aa) matches to <i>Corynebacterium aurimucosum</i> across this fragment
3178 – 4751 bp	Several very low identity (35%) short (50 – 80 aa) matches to <i>Corynebacterium aurimucosum</i> across this fragment
4752 – 6318 bp	As above but slightly higher identity (60%) fragments of 150 – 250 amino acids across the sequence, to <i>Corynebacterium aurimucosum</i>

**Table 1 tBLASTx homology of pQR492 using short fragments of the whole 6.3 kb sequence.**

Three of the four fragments in **Table 1** show low levels of homology to *Corynebacterium aurimucosum*, an organism commonly found on the tooth surface (Aas *et al.*, 2005) and therefore also a likely member of the tongue microbiota. *Arthrobacter* sp. are more commonly found in soil and are also members of the Actinomycete branch of the Gram positive bacteria. The BLAST searches seemed to suggest that three quarters of the pQR492 sequence contained a large part, if not all, of an Actinomycete gene. The next stage in this analysis was to identify open reading frames in the pQR492 sequence. The open reading frame (ORF) finder program from NCBI ([www.ncbi.nlm.nih.gov/projects/gorf/orfig.cgi](http://www.ncbi.nlm.nih.gov/projects/gorf/orfig.cgi)) identified potential ORF's of pQR492, which are shown in **Figure 4**.

ORF's coloured in red are those which showed homology to *Corynebacterium aurimucosum* or *Actinomyces odontolyticus*, another high G+C organism similar to *C. aurimucosum* also found in the oral cavity. The ORF depicted in yellow is a putative 430 amino acid transcription repair coupling factor (TRCF) with 95% sequence identity to *Rothia mucilaginosa*, another high G+C Actinomycete. From **Figure 4**, the placement of the red ORF's indicated that they may in fact form a single continuous protein. Even single base errors during sequence alignment of the pQR492 fragment could be responsible for the appearance of the protein as 4 separate ORF's. Although an effort was made to achieve 3 to 4 times sequence coverage of the 6 kb insert, there are one or two instances where short areas may only have been sequenced once. In order to identify whether these ORF's do in fact form a single protein, it was important to find their closest homologues in the public databases and use them as a basis for comparison.



**Figure 4** Diagram showing ORF's present in pQR492. Frames are depicted on this diagram with all ORF's of 20 amino acids or more, identified using ORF Finder ([www.ncbi.nlm.nih.gov/projects/gorf/orfig.cgi](http://www.ncbi.nlm.nih.gov/projects/gorf/orfig.cgi)). ORF's depicted in red are those identified by BLASTp to have homology to protein cauri\_0414 from *Corynebacterium aurimucosum* or *Actinomyces odontolyticus*, which is very similar to *C. aurimucosum*. The ORF in yellow has 95% identity to a transcription repair coupling factor from *Rothia mucilaginosa*, and the position of these ORF's within 492 creates the possibility that there may be 2 genes in the fragment.

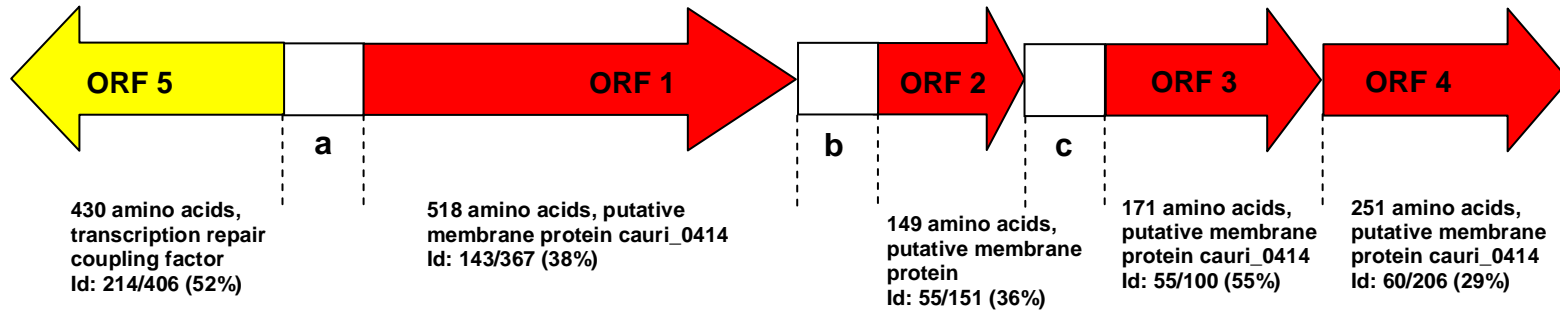
The largest continuous open reading frame is in frame +2 and is 518 amino acids long; subsequently referred to as ORF 1. BLASTp identified homology of ORF1 with a putative membrane protein (cauri\_0414) from *Corynebacterium aurimucosum* ATCC 700975 with 143/367 (38%) identity, 206/367 (56%) positives, and an expect value of 1e-64. *Corynebacterium aurimucosum* is a novel species discovered in 2002 and the family of bacteria are known for their appearance in isolates of clinical and veterinary importance (Yassin *et al.*, 2002). The genus belongs to the Actinomycetes, common in the oral cavity and on mucosal surfaces.

Frame three on the 5' to 3' strand (Figure 4) contains two fragments (in red) with similarity to a hypothetical protein from the oral commensal *Actinomyces odontolyticus*; a 149 amino acid protein (hereafter called ORF 2) and a 171 amino acid protein (hereafter referred to as ORF 3). Although the highest similarity was to *A. odontolyticus* for these two fragments, the second BLASTp match was again to cauri\_0414 of *C. aurimucosum*, indicating that the

fragment captured during cloning may encode some or all of a related protein. At the end of frame 1 on the 5' to 3' strand there is another ORF with homology to *C. aurimucosum* (subsequently known as ORF 4). This putative 251 amino acid protein goes right to the end of pQR492, with no stop codon before the *Bam*HI site of the pUC19 vector, which means that the protein has been cut short by the cloning process.

The 4 ORF's with homology to *C. aurimucosum* all have various degrees of similarity to the putative membrane protein cauri\_0414. Taking ORF's 1 - 4 together gives 1,090 amino acids, very close to the full length cauri\_0414 protein which is 1,213 amino acids long, suggesting that not only do ORF's 1 - 4 belong to the same protein, but that they could align with the cauri\_0414 gene, the closest homologue. The insert of 6 kb may appear in different frames in parts of the sequence where alignments done by hand are incorrect by one or two bases, which may not have been obvious at the time but which could be manifested as a frameshift. Because the protein represented by ORF's 1-4 is shorter than cauri\_0414, one could presume that the end of the protein is missing where fragmentation during cloning has occurred. **Figure 5** outlines the proposed layout of genes contained in pQR492 and indicates areas **a**, **b** and **c** which, on the current sequence, lie between ORF's 1, 2, and 3. **Figure 6** depicts all areas of pQR492 identity to the BLAST database along with amino acid level homology and illustrates the areas of frameshifting between ORF's.

Areas of sequence outside of the proposed ORF's were used individually to query BLAST for homology to cauri\_0414. Box **a**, a potential promoter region for ORF1 and ORF5, had no similarity to the BLAST database and box **c** showed very low level identity to fragments of human and zebrafish DNA. This lack of homology suggests that the gene fragment cloned into pQR492 has not yet been characterised. Box **b** did contain two short regions of homology to the *C. aurimucosum* genome, although not to the cauri\_0414 protein itself. Again, these short areas of low homology appear to suggest that the fragment contained within pQR492 belongs to a similar kind of organism to *C. aurimucosum*, and is likely to be a high G+C organism. This is exactly the type of organism that researchers hope to find when doing a metagenomic, non-culture based analysis. The low identity to BLAST indicates no record of this particular protein in the NCBI databases, although its homology to a membrane protein suggests a possible function. That this exact protein does not exist in the database could indicate that the original organism has not yet been characterised. Perhaps the organism has particular nutritional requirements which prevent it being easily cultured, or it may be present in very low numbers in the oral cavity, making it difficult to detect in some molecular based analyses which recognise abundant organisms more effectively than those in the minority.



**Figure 5** Outline of the possible gene structure of the genomic DNA fragment cloned into pQR492. Yellow arrow depicts ORF5 which showed homology to the Mfd gene of *K. rhizophila* and the red arrows depict ORF's 1 – 4, all with homology to *C. aurimucosum*. Direction of arrows indicates direction of translation. White block regions, boxes a, b and c, depict areas between ORF's, some of which also showed homology to *C. aurimucosum*. Box (a) could contain an area consisting of two promoters to ORF1 and ORF5, however checking for homology with the tBLASTx database did not reveal any similarity to either protein in this 211 amino acid region. Box (b) contains 175 amino acids, which contain two short homologous BLAST matches to the *C. aurimucosum* bacterium already identified from this study, and these areas of homology are shown in the sequence alignment in Figure 11 (below). Box (c) also contains 175 amino acids but a BLAST search with this sequence reveals homology to human and zebrafish DNA, rather than bacterial.

492 -1 **ASSPVA**PPTSPTLSDIGDEIADAI SQNELAGAAADELLDADTVLETIEWWSLKAELAQATTQDITRFGSDSMRLQGLDIPAVADAEASTDWAAALFEENTAVLDEARARVREPEVN  
P + T +DIGDEIADAI SQNELAGAAADELLDADTVLETIEWWSLK ELAQ TTQDITRFGSDSMRLQGLDIPAVADAEASTDWAAALFEENTAVLDEARARVREPEVN  
PTANVTYTDIGDEIADAI SQNELAGAAADELLDADTVLETIEWWSLKDELAQT TTTQDITRFGSDSMRLQGLDIPAVADAEASTDWAAALFEENTAVLDEARARVREPEVN

492 -1 **IIMSGAPLLETLTNMSEILLPTLSEMGEVYIGGAIKELMAAAAPYDAKLRAARSMVEPTILLERCPLLTLETLEEGGSLTRQDAVSFHRMEDLEDGFFELRVPTTATPPFVDIIGG**  
IIMSGAPLLETLTNMSEILLPTLSEMGEVYIGGAIKELM AAAPYDAKLRAARSMVEPTILLERCPLLTLETLEEGGSLTRQDAVSFHRMEDLEDGFFELRVPTTATPPFVDIIGG  
IIMSGAPLLETLTNMSEILLPTLSEMGEVYIGGAIKELMAAAAPYDAKLRAARSMVEPTILLERCPLLTLETLEEGGSLTRQDAVSFHRMEDLEDGFFELRVPTTATPPFVDIIGG

492 -1 **RVAYEGRKAVLDVRSYAADNLGRVVDKFPYEEGRVLHVPELKEIGTVIPQIVARVPAIVVQPRKAAEGTMARLVQLRRGVTSDRPSLREHPLTEWAPFLAIDAAPLYSRLAAALDE**  
RVAYEGRKAVLDVRSYAADNLGRVVDKFPYEEGRVLHVPELKEIGTVIPQIVARVPAIVVQPR EGTMARLVQLRRGVTSDRPSLREHPLTEWAPFLAIDAAPLYSRLAAALDE  
RVAYEGRKAVLDVRSYAADNLGRVVDKFPYEEGRVLHVPELKEIGTVIPQIVARVPAIVVQPRSSPEGTMARLVQLRRGVTSDRPSLREHPLTEWAPFLAIDAAPLYSRLAAALDE

Chapter 7: Additional Work – pQR492

← ORF 5

492 -1 AQRDTPAILLSLAEATTDRVATSIDAILAAHTGAVAGILTRASRETPQASAATRIAAWSSLQNTNLLPHLPAELPQATAKRTTSVPEGSDPTKEQPPQAPPPSSHITPISRHRNPA  
 AQRDTPAILLSL E TT RVA SIDAILAAHTGAVAGILTRASRE P+ASAATRIAAWSSLQNTNLLPHL  
 AQRDTPAILLSLVEPTTGRVAASIDAILAAHTGAVAGILTRASREAPKASAATRIAAWSSLQNTNLLPHL

ORF 1 →

492 +2 Promoter \*KRM  
 492 +3 Region \*  
 492 -1 DSHIQR\*// 159 AMINO ACIDS // \*

492 +2 TAFTAAPLTPPTPASAATGTHDGSSDAAAASCYEVKQVNPSSASGTYWMLYTPQSGPAQFYCDQETDGGGWVMIGRGREGWTESYNGTGDPNQLHQNP TGPSAFTPVQLPANTV  
 +P + A T DGSS D+AAASC+ +KQ NP A +G+YW L TPQM P +F+CDQE DGGGWVMIGRGREGW G GD + L P+ F P QLP T+  
 SPAAQDTPAVVTRDGSSPDRAAASCWAIKQDNPDAKNGSYWILLTPQMAPQEFFCDQEMDGGGWVMIGRGREGWDRYPAGQGDISALTSRDRTPADFAPAQLPTKTI

492 +2 DALLNGIKPQDLDPDGMRLHRAHNARGTQWQNVVQRPQTEQWTWAMSYGQRWGTVKFTGAGI----NRTAHMGRHASEMAPGITSSVRFANPNQGYQIGFAYGALVNFGENP  
 D LL+G +L +GMR+ RA N GT+WQ ++ + E W+WA+S ++F+ +RT +A + GI S+ R + Y+IGF YG G ++  
 DGLLSGQHINELDEGMRVVRATNNSGTRWQTADIKPQRMENWSWALS-AEDPALFRFDNGPLWYRADRTDRFMGNAIGLR-GIDISTTR-----ARQYKIGFGYGPWKRVGRALP

492 +2 DSYIYHKRGSAGYSIPFTQVFLRPKLTQRDNLNFSQIGSSSAAS-NRRALPNSYTMPVRWRTSEQTGTGKKNEMNTYVQAITQVGDVFTGGDFKYVESAGGERVDQSYLAGYNVD  
 DS+IY G A ++P+ +++LRP+L+ D F I + A +S++ P +W + TG+ E N VQ A Q DTVF GG+F E+A GE + ++ +A ++  
 DSEIYLRPLANADAPFEPiPLRPLLEVTDSAFVSISSPTQWGVGTGAL-TSRSTPGNWPTGADSFTGSTNEYN--VQAFQKDDTVFVGGNFTAAENHAGESLPRTAVAAFDAT

← ORF 1 KGQYV\*GVRLLACAARFKLS

492 +3  
 492 +2 SGELVRSFRPTFNGQIKALKALPNNRLALPNNRLALVASSPRVMARRSTTSSPFWTQPARSTARGIFSQRVMRCSSGEDLLVQDGYLYIGVTS<sup>\*</sup> \*  
 +GE+ R F +GQ+KAL LPN +L + \* \*  
 TGEVRRDFAVDLDGQVKALLVLPNGKLLI



Chapter 7: Additional Work – pQR492

ORF 2 →

492 +3 NGAVDWNWRPNFNGTVNGITAASDNSTVHAAGYFTE **LNNQRAFRLLAALNGSDASNIRWEWEP**SLKLNITDRIVYAFQFDVQDAGSTVWVTAGADHLIANYSKNGYGRISTAI SKYGG  
 +G D +W P FNGTV ++D +AAGYFT+ +S GG  
 DGVPDRSWNPEFNGTVVDTDVSADGGRFYAAGYFTK MSSNGG

492 +3 **DWQDLHLSGNTIYGACHCGDVLFE**GSTGYHTYWKESKAVHRMRLVAAF~~DKDS~~GEVVGEFSPVLK**GASGYGVWESFVDS**RGNLWVGGDINRS-LGANGEQRTVGFARFAPRDVTAPS  
 DQ + +G Y +CHC + ++ S + T ++ V A+D +G ++GEFSP + G++ G W F+ G LW GGD S Q GF R+ +D AP  
 DVQTIASNGEVTYASCHCNENAYQDSYSWPT RIQWVGAWDAKTGKQLGEFSPYMLGSNNGGGWSLFIAEDGALWAGGDFGTGSRTNLTTAQWNGGFVRYPAQDREAPR

← ORF 2

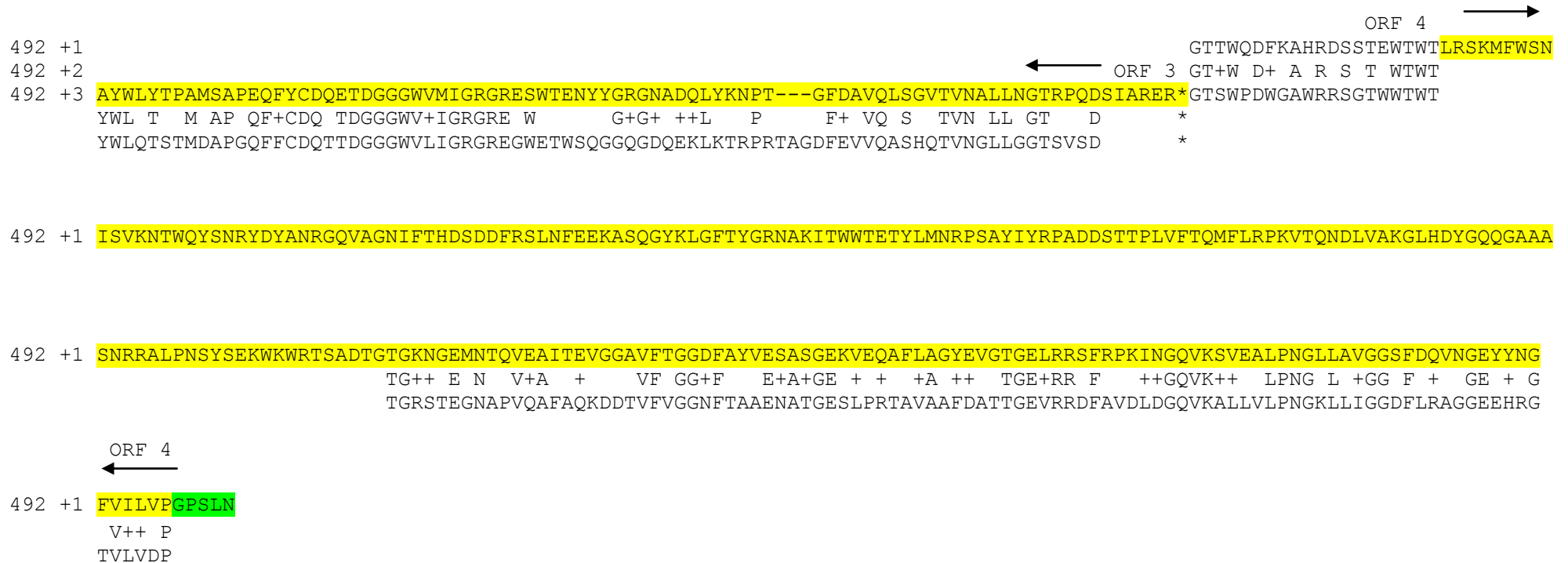
492 +3 **TPSNLSV-QRDG**STDKLSW**SGVRESGARYQVLR**DDRVIATVSGT**SYEVEHTDGARYYVRSIDASENFSASTGAAQA\***VRLLI\*FL\*AAAWFGLCRVCGVRVRRRCTRGVCP  
 P ++ Q T L+W+ ++ A Y+VLRDDR+AT V RY+VR++D + N SA+T A A\*  
 VPDKVTFNQSTAKTVGLTWAEASDA-ASYEVLRRDDR+VATSISPRATVPRGGDDRYFVRAVDEAGNRSATTHVAVA\*

492 +3 \*FWFGPVFPVCYLFAPVAAMMGWDSLRRMGCFRPF\*WAES\*SVVFLRCWYDRGNFTDF\*RSASYGCVATVGFVREGWGVFIQFPITTLTDSGYRLAVSAVHNASPLPRCLVT

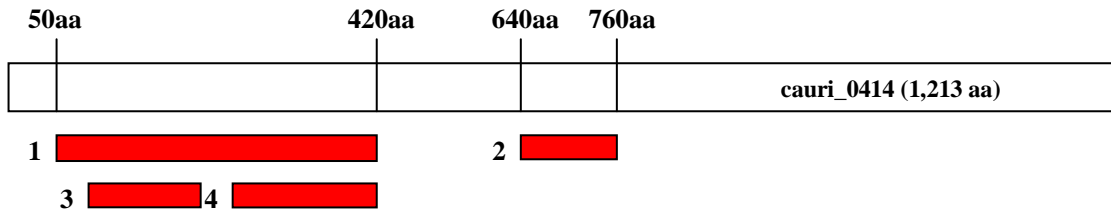
ORF 3 →

492 +3 MLCLVWGGCCCVNCSFLLCAG\*GL\*ERQ**MN**FLPFTRGGRA**QGAASSASEGSRVASRSGSRALGAAAA**SFAMVAASLGP**IASGAQAADARYYDGSSSERAAA**SCWEVK**QNNPRGKSG**  
 DG S+ AAASCWE+KQN+PR ++G  
 DGLSAATAAASCWEIKQNDPRSQNG

Chapter 7: Additional Work – pQR492



**Figure 6** Sequence alignment of the entire amino acid sequence of pQR492 and all areas of homology to the BLAST database. Top line is the probable full amino acid sequence of pQR492. Open reading frames located by the ORF Finder program of NCBI are highlighted in yellow with arrows marking the beginning and end. pUC19 bases at the beginning and end of the 492 sequence are shown highlighted in green. Asterisks denote a stop codon and frameshifts are illustrated by the continuation of the amino acid sequence on a higher or lower text line, and frame is shown by +1, +2 or +3 at the left hand side of the sequence. The promoter region between ORF's 1 and 5 contained multiple stop codons in all 6 frames, making the true amino acid sequence difficult to identify.



**Figure 7 Alignment of putative ORF's from pQR492 with cauri\_0414 of *C. aurimucosum* using clustalw multiple sequence alignment program.**

Knowing that much of the pQR492 sequence, whether contained within a single ORF or not, did share similarity with cauri\_0414, and given the similar lengths of the two proteins, a clustalw alignment was used to identify exactly where the ORF's 1 – 4 aligned on the cauri\_0414 gene. To check this, the 4 ORF's were individually aligned with the full amino acid sequence of cauri\_0414, shown in **Figure 7**.

**Figure 7** illustrates the sequence similarity shared between ORF's 1, 2, 3 and 4 and cauri\_0414. ORF's 3 and 4 share sequence similarity with ORF 1 and the corresponding region on the cauri\_0414 sequence, which gives the impression that the protein encoded by pQR492 could fold back on itself and is in fact not related to cauri\_0414. As the pQR492 text sequence was aligned in a piecemeal fashion over a period of about 3 years, there was a small possibility that one or two fragments had been misplaced or repeated in the sequence, which could have led to the scenario in **Figure 7**. To check for overlapping text in the pQR492 sequence a DNA dot blotting program ([www.vivo.colostate.edu/molkit/dnadot](http://www.vivo.colostate.edu/molkit/dnadot)) was used, which checked for areas of repeated sequence. According to this program there were no such repeated regions in the nucleotide sequence of pQR492, meaning that instead of repeated regions due to human errors during sequencing or alignment, the nucleotide sequence in the original organism and the corresponding protein structure contains repeated domains which are a legitimate part of the domain architecture. That there is evidence of repetition in the raw text sequence of pQR492, but there is in the protein structure could indicate that, as the organism has evolved, the DNA sequence has changed but the protein structure has conserved these repeated domains which may serve a functional purpose. The same dot blotting program was used to check if the closest homologue to ORF's 1-4, cauri\_0414, contained any regions of repetition similar to that seen in pQR492, but none were found. This finding posed an interesting scenario where, instead of seeing the expected alignment of ORF's 1 – 4 from pQR492 along the length of cauri\_0414, there appeared to be a region of similarity between the first and second half of pQR492, which could functionally differentiate it from the cauri\_0414 protein. To see if the ORF arrangement around cauri\_0414 was similar to pQR492, the genes in the immediate vicinity were checked.

According to Genbank the full length of the putative membrane protein cauri\_0414 is 3,639 bp and lies immediately downstream of another putative membrane protein, cauri\_0413, of 1,587 bp and upstream of a smaller hypothetical protein, cauri\_0415, of 567 bp (**Figure 8**). This section lies between 450 and 455 KB on a 2.79 Mb genome. The 567 bp protein, cauri\_0415, does not appear downstream of ORF 1 in pQR492, or indeed anywhere in the pQR492 insert sequence. The level of homology between cauri\_0414 and ORF 1 is very low however, because of the general level of homology spread across the pQR492 fragment to *Corynebacterium aurimucosum*, and since pQR492 does not contain the 1,213 amino acid protein cauri\_0414, it may contain a similar protein from a related organism.

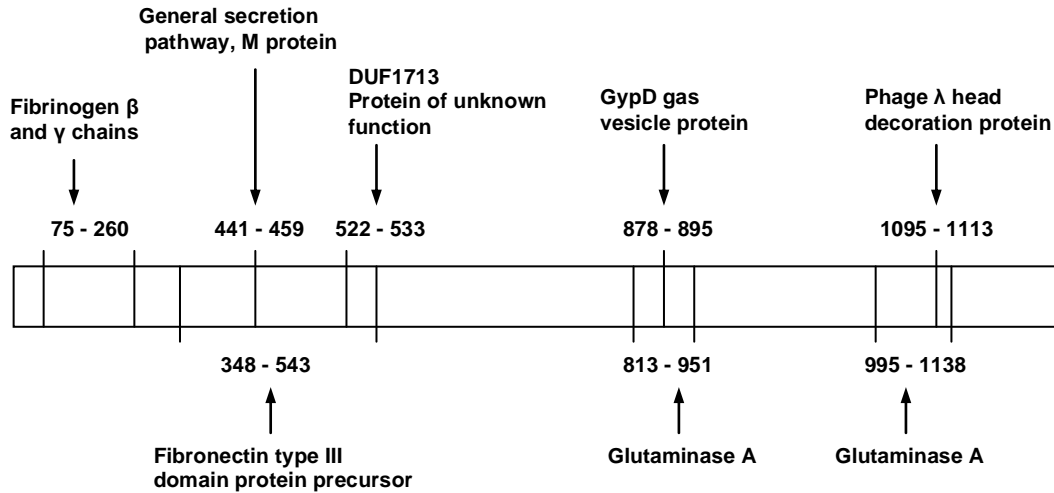


**Figure 8** *Corynebacterium aurimucosum* genome fragment showing the location of the 3,639 bp cauri\_0414 gene, showing the highest homology to ORF 1, the largest in pQR492, at 1,557 bp.

What is clear from this analysis so far is that pQR492 contains a fragment of microbial genome showing various (low) degrees of homology to *Corynebacterium aurimucosum* ATCC 700975 and *Actinomyces odontolyticus*. Both these organisms belong to the phylum Actinobacteria and share some characteristics. They are both Gram positive rods capable of anaerobic metabolism, and are both constituents of the skin flora. These organisms are both classed as high G+C content with *A. odontolyticus* having a 65% G+C content and *C. aurimucosum* at 60% G+C content. *Corynebacterium aurimucosum* ATCC 700975 has been fully sequenced so the distinct lack of full homologous sections between pQR492 and the NCBI database could imply that an unknown representative of the *Corynebacteria* or *Actinomycetes* has been isolated in this case.

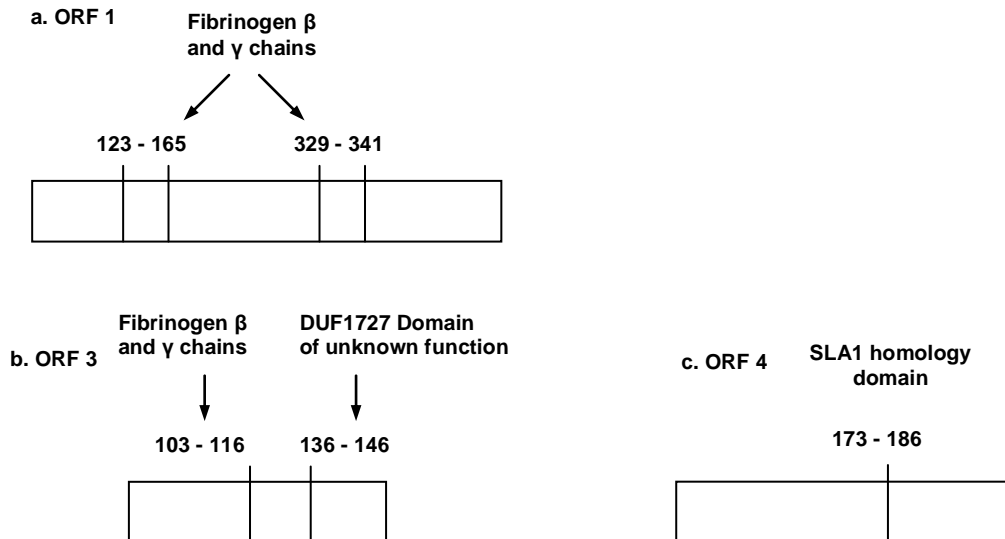
To find out more about the functional role of domains contained within cauri\_0414 and the ORF's of pQR492 the Pfam database was used, provided by the Sanger Institute ([www.pfam.sanger.ac.uk](http://www.pfam.sanger.ac.uk)). The Pfam database is a large collection of protein domain families. Each family is represented by multiple sequence alignments and hidden Markov models (HMMs). There are two levels of quality to Pfam families: Pfam-A and Pfam-B. Pfam-A entries are derived from the underlying sequence database. Pfam-B families are un-annotated and of lower quality as they are generated automatically from the non-redundant clusters of the latest ADDA release. Although of lower quality, Pfam-B families can be useful for identifying functionally conserved regions when no Pfam-A entries are found (Finn *et al.*, 2008). Pfam search results were used to compare domains present between cauri\_0414 and ORF's 1 – 4.

From cauri\_0414, five Pfam A entries and 3 Pfam B entries were found, and are illustrated in **Figure 9**.



**Figure 9** Results of Pfam search of gene cauri\_0414 of *C. aurimucosum*. Pfam A results are shown above the bar depicting the entire 1,213 amino acid protein of *C. aurimucosum*, and Pfam B results are shown below the bar in regions denoted by the vertical lines.

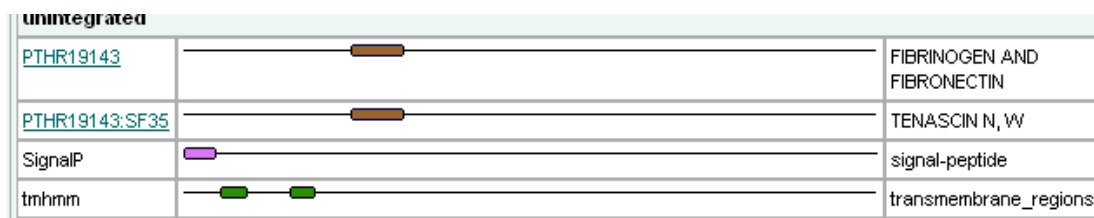
With this information, ORF's 1 to 4 of pQR492 were also compared for shared domain architecture with cauri\_0414 or similar proteins. These results are shown in **Figure 10**.



**Figure 10** Results of Pfam search of ORF's 1 to 4 of pQR492 for domain architecture similar to cauri\_0414. ORF's are depicted by the white rectangles and vertical lines illustrate where the architecture has been identified. All matches are to Pfam A and therefore are depicted above the white rectangles. ORF 2 has not been depicted here because it contains no Pfam A or B matches.

Constructing Pfam diagrams helped to locate areas of similar domain architecture between the ORF's of pQR492 and the nearest BLASTp homologue, cauri\_0414, with the aim of identifying a potential role of the protein encoded by pQR492. **Figures 9 and 10** show clear repetition of the fibrinogen beta and gamma chains, C-terminal globular domains, between 75 and 260 amino acids on cauri\_0414, and ORF's 1 and 3. ORF 4 contains architecture similar to the SLA1 domain which is thought to function as an endocytic adaptor. The presence of fibrinogen related domains could, if also present on the surface of neighbouring bacteria, act as a focal point for some binding interaction. Clearly, the presence of fibrinogen related domains in both pQR492 and cauri\_0414 was the source of previous sequence identity, but does this mean that these two proteins share a similar function?

Because of its similarity to cauri\_0414, a putative membrane protein, the topology of pQR492 was checked to highlight potential membrane spanning domains. This analysis was carried out using InterProScan of expasy tools ([/www.ebi.ac.uk/Tools/es/cgi-bin/iprscan/iprscan.cgi](http://www.ebi.ac.uk/Tools/es/cgi-bin/iprscan/iprscan.cgi)), which combines different signature recognition members native to the InterPro member databases into one resource (Zdobnov & Apweiler, 2001). **Figure 11**, which represents ORF 1, contains a signal peptide followed by two transmembrane regions, and a fibronectin/fibrinogen binding domain. The presence of a signal peptide and a transmembrane domain indicates that this protein could be targeted to the membrane in the original organism, therefore strengthening its position as a potential membrane protein as initially indicated from its homology with cauri\_0414.



**Figure 11 InterProScan analysis of ORF 1. This diagram illustrates the presence of a signal peptide and two transmembrane domains which was not identified in any previous analysis and which could indicate that this protein is targeted and anchored to the bacterial cell membrane. The presence of two transmembrane domains could indicate that the protein is either intracellular or extracellular.**

Regions of pQR492 shared domain architecture with the first half of cauri\_0414 in the fibronectin/fibrinogen like domains, indicating a similar functional role for both of these proteins, so the other half of the cauri\_0414 protein was checked for topology which could confirm or deny further similarity to pQR492. The program InterProScan did not identify any further regions of homology to pQR492 other than the fibronectin/fibrinogen domains already identified by the Pfam analysis, and also did not identify any regions which could differentiate

the role of cauri\_0414 further from pQR492. This analysis means that so far, protein cauri\_0414 is the closest homologue to ORF's 1 – 4 of pQR492 and there is no additional information available to differentiate them further.

It was initially thought that ORF's 1-4 could be part of the same protein, potentially one closely related to cauri\_0414, and that sequencing or alignment errors could have resulted in the frameshifts present between the ORFs. However, these ORFs could each represent a separate gene and if this is the case they should each be preceded by a ribosome binding site (RBS). Ribosome binding sites consist of a Shine-Dalgarno (SD) sequence (4 - 9 bases long), positioned 3 – 11 bases upstream of the initiation codon, usually ATG (Winnacker, 1987) but which can also be GTC and CTG in high G+C organisms like that partially cloned into pQR492. Although much is known about RBS's in *E. coli*, there were no guarantees that RBS's from the organism pQR492 originated from would follow the same rules. Gene recognition is made much more difficult by the interruption of reading frames by frameshifting, which can either be biological or caused by sequencing or alignment errors, which was a possibility in this case. The pQR492 text sequence (**Appendix 7**) was checked by eye for RBS's because computer programs for this purpose may not recognise the gene recognition features present in pQR492, many being based on the gene start sites of *E. coli*, where the inserted fragment in pQR492 probably originated from a high G+C organism. Two potential RBS's were identified upstream of ORF 1 and these are illustrated in **Figure 12**.

```

      ↓
GAAAGCGCCGCATGCCCATCATGTCTGCTAGATGGGGTGCTTTCGCCTCGCTTGTACTGCTTTACTG
      →
GCGACCTCTTACTCTTACTATCAACCCTCCGGTTTTTCAGTATTTACTGCGGTTTTTCTAGTGCGCTAA
ATTCACCGTATCTTTCGATGTTTCTCCGGAACGTCGCATCAGAAAGACGCATATGACAACCACATA
  
```

**Figure 12 Potential ribosome binding sites of ORF 1. Potential SD sequences are underlined and the potential start codons are shown in bold typeface. The vertical arrow indicates the ORF start position as proposed by the ORF Finder program and the horizontal arrow indicates direction of translation.**

Although ORF's can be predicted with some degree of accuracy using current online programs, gene start sites are somewhat more difficult for a computer program to pin down. Because the pQR492 sequence was a manageable size, the 5 ORF's were checked for RBS's by hand. The first potential RBS in ORF 1 is the most likely true start site. The SD sequence GAAAG is followed 6 bases downstream by the start codon **ATG**, which also indicates the proposed start position of ORF 1, illustrated by the vertical arrow. However, there is a second potential RBS 160 bases downstream of the first which is again comprised of the SD sequence GAAAG followed by **ATG**. The validity of start site prediction can only be tested

experimentally; therefore what is said here is by no means without question, however given the predicted ORF start position the first potential RBS appears to be the most probable.

Carrying out the same analysis on the nucleotide sequence at the beginning of ORF 2 again identified two potential RBS's, illustrated in **Figure 13**.

```

CTGGCAGGATCTGCACCTGAGCGGTAACACCATTTACGGCGCGTGCCACTGCGGTGACGTCCTCTTTGA
GGTTCTACCGGTTACCACACCTACTGGAAGGAATCGAAAGGCGGTTACCGCAATGCGCCTGGTCGCGG

```

**Figure 13 Potential ribosome binding sites of ORF 2. Potential SD sequences are underlined and the potential start codons are shown in bold typeface. The vertical arrow indicates the ORF start position as proposed by the ORF Finder program and the horizontal arrow indicates direction of translation.**

The first potential RBS of ORF 2 consists of the SD sequence AGGA followed by the translational start site **CTG** 7 bases downstream. Although this potential RBS fits the 'rules', i.e., the SD sequence (4 - 9 bp) is positioned 3 - 11 bases upstream of the start codon, there is another potential RBS further downstream which both fits the rules and is closer to the proposed ORF start position. This potential RBS, consisting of the SD sequence AAGG, is followed 11 bases downstream by the start codon **ATG** and, because it is closest to the proposed ORF start position (indicated by the vertical arrow), appears the most likely RBS of the two illustrated for ORF 2.

The same analysis carried out on ORF 3 identified only one potential RBS at the proposed ORF start position, illustrated in **Figure 14**.

```

      ↓
AGGTCTCTAGGAGAGACAAATGAATTTCTTCCGTTACGCGTGGAGGG

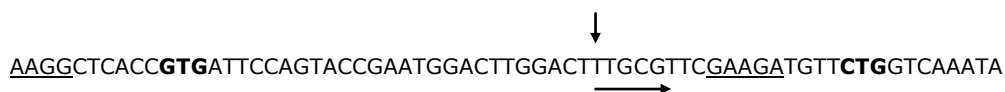
```

**Figure 14 Potential ribosome binding sites of ORF 3. Potential SD sequences are underlined and the potential start codons are shown in bold typeface. The vertical arrow indicates the ORF start position as proposed by the ORF Finder program and the horizontal arrow indicates direction of translation.**

With the SD sequence AGGAG followed 6 bases downstream of the translational start site **ATG**, and in the same position as the proposed ORF start site, it appeared that this was the only sequence capable of acting as the RBS for ORF 3.

Repeating the process for ORF 4 revealed two potential RBS's, illustrated in **Figure 15**.





**Figure 15 Potential ribosome binding sites of ORF 4. Potential SD sequences are underlined and the potential start codons are shown in bold typeface. The vertical arrow indicates the ORF start position as proposed by the ORF Finder program and the horizontal arrow indicates direction of translation.**

Although there is only one potential RBS close to the proposed ORF start site (vertical arrow), the situation is not as clear cut as that for the other ORF's. The nearest proposed ORF start site is 9 bases upstream of the SD sequence, which is followed 4 or 5 bases (depending on whether the SD sequence is GAAGA or GAAG) downstream by the translational start site **CTG**. Although this is probably the most likely RBS for ORF 4, the other one should also be considered which begins 38 bases upstream of the proposed ORF start site, consisting of the SD sequence AAGG followed by the translational start site **GTG** 6 bases downstream. Since both of these potential RBS's fit the 'rules', the most likely of the two is probably that closest to the proposed ORF start site.

The RBS analysis showed that each of the 4 ORF's with similarity to cauri\_0414 of *C. aurimucosum* could contain its own RBS and could therefore be individual genes. They could also be part of the same operon since they share homology to the same membrane protein, cauri\_0414.

Going back now to **Figure 4**, which showed all the ORF's identified in pQR492, the first ORF in frame 1 (in yellow and now known as ORF 5) in 3' to 5' direction was also analysed in some detail. Because this chapter was the last to be written prior to thesis submission, an update in the BLAST database on the 5<sup>th</sup> August 2009 provided this match to a *R. mucilaginosa* transcription repair coupling factor (TRCF), where previously the highest homology was to the same protein from the organism *Kocuria rhizophila* at 52% identity. The conserved nature of TRCF's means they are invariable between species, which could imply that although ORF 5 shows 95% identity to the *R. mucilaginosa* TRCF, this does not necessarily mean pQR492 contains 6.3 kb of the *R. mucilaginosa* genome. The entire TRCF gene which appeared in the BLAST database on the 4<sup>th</sup> August 2009 was 1,249 amino acids in length, far longer than ORF 5 at 430 amino acids. To find out what part of the TRCF gene was present in pQR492, both TRCF and ORF 5 were aligned using clustalw, illustrated in **Figure 16** which confirmed that pQR492 contained the first 430 amino acids of the TRCF gene. This meant that locating a probable RBS for this gene within the 211 amino acid region between the start of ORF 5 and the start of ORF 1 was quite likely. As expected, no stop codon was present before the *Bam*HI site of the pUC19 vector, which confirms that the full length version of the gene cloned into pQR492 was probably shortened due to fragmentation during the cloning process.

```

ORF 5  RQIHSDAPNRHRSIPTIHSSPPAQPPQAKTPDSGEPVSTTRKATAQPLEAPLHPLLNTLN
trcf   -----MSTTRKATAQPLEAPLHPLLNTLN
          :*****

ORF 5  QLSSWAAIRTAASAQPTERSARTLIGAVAGRHAALIADISTAVRDTTAEALSLIIAPTDR
trcf   QLSSWAAIRTAASAKPAERSARTLIGAVAGTHAALIADISAAVRGTTPEVLSLIIAPTDR
          *****:*.***** *****:***.*.*.*****

ORF 5  QAEDLAAAALRSYLPAADIALFPAWETLPHERLSPRSDTVGRRLQVLRAMTGEAAKRPQVV
trcf   QAEDLAAAALRSYLPAADIALFPAWETLPHERLSPRSDTVGRRLQVLRAMTGEPSRPQVV
          *****.:*****

ORF 5  IAPVRAVIQPIVTGIEKLEPVHLVEGEEYPFKDVVRGLNDAAYS SRVDLVAKRGEYAVRGG
trcf   IAPVRAVIQPIVTGIEKLEPVHLVRGEEYPFKDVVRGLNDAAYS SRVDLVAKRGEYAVRGG
          *****.******

ORF 5  IIDVFPPTATTPVRLEFFGDELDEMRFHFSVADQRTLSGGEELTELTLPCRELLITPEVM
trcf   IIDVFPPTATTPVRLEFFGDELDEMRFHFSVADQRTLSGGEELTELTLPCRELLITPEVM
          *****

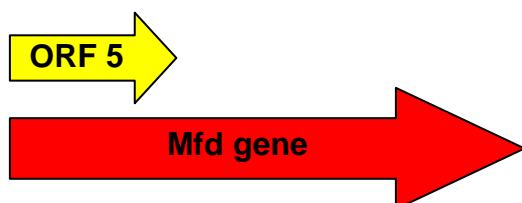
ORF 5  SRAARLKADYPAAAAMLEKIAGGIYVEGMESLTPLLIESMNTLTELTPAGSMI INVEPER
trcf   SRAARLKADYPAAATMFEKIAGGIYVEGMESLTPLLIESMNTLTELTPAGSMI INVEPER
          *****:*.*****

ORF 5  VRARAEDLVATNEEFLLAAAWDTSAEADAVAPIDLGLRMSDSGFRTIDQTTAQALEAKLS
trcf   VRARAEDLVATNEEFLLAAAWDTSAEADAVAPIDLGLRMSDSGFRTIDQTTAQALEDKLS
          *****:*** **

ORF 5  WWEITELVTDADLLEDAAAGALENQSIAADAIEDGIDSLTPS-----
trcf   WWEITELVTDADLLEDAAAGALENQSIAADAIEDGIDTYTVNATPATAFNNGSVERMLSQVG
          *****: *

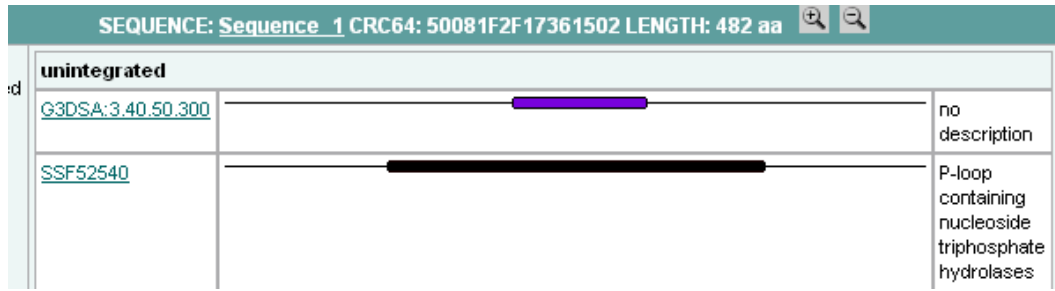
ORF 5  -----
trcf   DLIQQQWTVLALTNRGSTDRLIDL FHS GEGAPAVPAARRTSLEADPAGDLEHGIVEVCE

```



**Figure 16** Clustalw alignment of ORF 5 with transcription repair coupling factor, part of the Mfd superfamily of helicases. The full length transcription repair coupling factor (TRCF) is 1,249 amino acids in length (full length not shown on the clustalw alignment) and is illustrated by the red arrow under the sequence alignment. The yellow arrow illustrates ORF 5, contained within pQR492, and therefore depicts how much of the TRCF gene is contained within pQR492.

Because of the high sequence identity of ORF 5 to the TRCF, it was important to identify if these genes (or this gene, if they were one and the same) played a similar functional role in the environment. To find this information, InterProScan of expasy tools ([/www.ebi.ac.uk/Tools/es/cgi-bin/iprscan/iprscan.cgi?](http://www.ebi.ac.uk/Tools/es/cgi-bin/iprscan/iprscan.cgi?)) was used to try to predict the topology of the, albeit, shortened gene captured in ORF 5 (**Figure 17**).



**Figure 17 InterProScan analysis of ORF 5. This analysis identified a region containing a P-loop containing nucleoside triphosphate hydrolases, a region which is included twice in the full length version of the TRCF gene.**



**Figure 18 InterProScan analysis of the full length TRCF, identified as the closest homologue to ORF 5. The P-loop is clearly present in 3 repeated domains of the 230 amino acid conserved region of the AAA family of proteins.**

The InterProScan search of ORF 5 identified a phosphate binding loop which are often contained in ATP and GTP-binding proteins (Saraste *et al.*, 1990). According to Snider *et al.*, (2008) the family of proteins which contain this highly conserved region are called ATPases Associated with diverse cellular Activities, or AAA, and they often perform chaperone-like functions that assist in the assembly, operation, or disassembly of protein complexes. The common conserved module of the AAA proteins is 230 amino acids long, which appears to be the part identified on the InterProScan diagram in **Figure 17**, shown by the thick black line. An InterProScan search of the full length TRCF gene showing high identity to ORF 5 located 3 of these highly conserved P-loops in the 1,249 amino acid sequence, suggesting that ORF 5 contains the first of these three (**Figure 18**). From this analysis it would seem likely that pQR492 does contain the first 430 amino acids of a TRCF from a high G+C organism.

With confirmation that ORF 5 contained the start of a gene, an investigation could begin into possible gene start sites. Potential RBS's have already been discussed for ORF 1 in **Figure 12** so what follows is a discussion of the potential RBS's of ORF 5.



**Figure 19 Potential RBS's of ORF 5. The top arrow denotes the proposed translational start site by the ORF Finder program, and the horizontal arrow denotes direction of translation.**

From the diagram, there are three potential RBS's followed by translational start sites. The first potential SD sequence GAAG is followed 13 bases downstream by the alternative start codon **CTG**. Although 13 bases could be considered stretching the rules regarding the usual number of bases between a RBS and start codon (usually 3-11), this RBS was included because the **CTG** codon itself was identified by the ORF Finder program to be the start of ORF 5 (shown by the vertical arrow), which means it could be the true RBS. Further upstream is the potential SD sequence GAAAAG which is followed by the start codon **CTG**. Again, the strength of this sequence as a potential RBS is debatable because there are only three bases between the RBS and start codon. The final potential RBS has the SD sequence AGGAGA followed by the alternative start codon **GTG**. Although there are only four bases between the proposed RBS (underlined) and the start codon, the RBS could in fact incorporate fewer bases than those underlined. For example, canonical guidelines could be met if the RBS consisted of any of these base combinations: AGAGG; GAGG; AGAG, which could leave up to 6 bases between the RBS and the start codon. There are limitations with all the RBS's suggested for ORF 5 which means it is almost impossible to choose the most likely without experimental confirmation.

Although potential RBS's have been analysed thoroughly for ORF's 1 - 5, a similar analysis of potential promoters would be far more difficult and less accurate. Because promoter regions can differ between bacterial phylotypes, and because there was still no confirmation of the species level of the organism cloned into pQR492, analysis of potential promoters in box **a** of **Figure 5** was not carried out. A study of the high G+C organism *Corynebacterium glutamicum* promoters carried out in 2003 by Patek *et al.*, identified common promoter motifs in the -10 and -35 positions which were found to be substantially less conserved than those found in *E. coli* and *B.subtilis*.

---

## **Chapter 8**

### **Discussion**

---

## Introduction

**The present study details the successful production of the first phage display library to use metagenomic DNA; previous phage display projects have all used the genomes of single organisms. Combining metagenomics and phage display presents an opportunity to include a multitude of bacterial protein fragments with a functional screening process of elimination. In this project, metagenomics was the ‘broad net’, cast to allow detection of a larger proportion of the significant binding events occurring on the tongue dorsum than had previously been identified. The incredible molecular instrument phage display comprises many production stages, all with prospective hindrances, invariably making it a difficult technique to master. This chapter will discuss the difficulties and setbacks encountered during the current project, and address methods of resolution.**

In general, there are many examples of bacterial proteins binding to host cells, but very little data regarding the substrata they have affinity for. The approach taken in this thesis turned this around by using the host ligand as the bait for which to ‘fish’ for bacterial proteins that interact with them. The hypothesis was that tongue bacteria employ a host of proteins, far beyond those already known, that facilitate binding to the tongue surface and oral cavity, and that employing a combination of molecular techniques would help to identify a great deal more of them. A variety of proteins from different oral bacteria were successfully identified in this study through homology to the public sequence databases, and by some functional data. However it may have been possible to achieve greater results by amending and modifying some of the decisions made throughout the project.

Using metagenomic DNA to create the phage display library is a novel method of searching a mixed microbial community for binding proteins. But by its diverse nature, the DNA may have been the source of issues that were not envisaged at the start of the project. In previous phage display publications which use the genomes of single bacteria, typically one or two proteins are identified by 3 panning rounds of successful enrichment, the identification of one or more clones containing overlapping regions of the same gene. The intention in the present study was to use this same enrichment process to identify, by the 3<sup>rd</sup> panning round, a large pool of different genes encoding proteins responsible for bacterial adhesion. This did not happen for several reasons. Firstly, and due to the inclusion of metagenomic DNA, phage display library diversity was high which, although this is a positive aspect, did mean that each panning round was crowded with an enormous variety of proteins. Secondly, and in concert with the first point, using only 3 rounds of panning did not significantly reduce the large variety of proteins identified from the panning process which therefore, and thirdly, resulted in the subsequent *in silico* analysis being insufficient to cover a representative number of clones in the library. The resultant diversity of the third round panning eluate could be explained either by the presence of many true binding interactions, or the presence of a variety of specific and non-

specific interactions, the latter being the most likely explanation. In order to find out, further experiments, like adhesion assays, on pure preparations of individual clones must be carried out to identify how the clone of interest behaves.

Because of these unforeseen considerations, the range of proteins identified (and still awaiting identification) was much wider than expected and more numerous than the group had time to investigate. Perhaps this range of binding proteins actually goes some way to proving the original hypothesis; that the tongue microbiota do encode an enormous variety of binding proteins which facilitate association with the tongue surface and surrounding host proteins.

### **Phage Display**

The use of the pG8H6 phagemid vector system in this project was based on the success of our colleagues at the Eastman Dental Institute in their discovery of several new genes encoding potential adhesins when using phage display libraries made from single organisms (Williams *et al.*, 2002; Mullen *et al.*, 2007). However, the phagemid system has many issues, occasionally these are individual to each project, but in many cases they are not. Because genomic DNA is randomly fragmented before cloning into the pG8H6 phagemid vector and both must be in frame for expression, ribosomal slippage is often required in either the poly-His or c-Myc regions in order to produce the inserted protein in the correct frame with both tags and therefore the coat protein 8. However this slippage could be detrimental to protein production, an issue noted more than once (Jacobsson & Frykberg, 1995 & 1996; Carcamo *et al.*, 1998). Jacobsson and Frykberg are known for their work on phage display vectors and found this problem of frameshifting was extremely common. In 1996 they developed the pG8H6 vector, used in this project, with a deliberate ribosomal slippage sequence (poly-His) inserted in front of the fusion protein, so that the sequence would frameshift naturally to end up fused to the coat protein. Jacobsson found that all of their resulting inserts were out of frame with the phagemid vector, either +1 or -1 in all cases, which led them to suggest that the system was selecting *against* clones in the correct frame. Frameshifting downregulates expression of the fusion protein; although it should remain high enough for display at the phage surface (Jacobsson *et al.*, 2003) although, unfortunately sometimes, this expression level may be too low to enable binding or detection (Jacobsson & Frykberg, 1998). For this reason, the group went on to design the phagemid vector pG8SAET, which contains an E-Tag in frame with gene 8. Because the E-tag is out of frame with the signal sequence until a foreign insert is spliced in, this restores the reading frame in 1-in-18 clones. Following successful panning, the number of clones in the correct frame will then increase leading to, they claimed, almost 100% correct clones. The metagenomic DNA in the current project was used in both pG8H6 and pG8SAET, however the pG8H6 library came to fruition first and because phage display libraries take so long to make, library construction in pG8SAET was discontinued.

The pG8H6 system has the added limitation that fusions are to the C-terminal end of the displayed polypeptide which may impair identification of binding domains located in the extreme C-terminal end of the foreign protein. Cramer and Suter (1993) solved this problem by creating the phagemid pJuFo that contains gene III fused to the Jun-gene fragment and Fos-gene after which the foreign DNA is inserted. Thus, through the interaction between Jun and Fos, the foreign polypeptide is displayed at the phage surface (Jacobsson *et al.*, 2003). The variety of phagemid vectors available demonstrates the flexibility with which phage display can be used. Perhaps pG8H6 was not the optimal phagemid to use in the current study, however there was no guarantee that others would have performed better.

The problem of frameshifting is common in phagemid libraries, noted by Carcamo (1998) who, although using coat protein III for display, noticed 46% of clones contained 'unusual sequences'; in other words they contained either stop codons or frameshifting, resulting in an out-of-frame protein. So-called 'unusual sequences' were easier for the Carcamo group to spot as the library was constructed from a single synthetic stretch of 145 nucleotides. In the same study the group also noted that the frequency of inserts appearing out of frame was linked to the ligand used for selection. This appears also to be the case in this project where proteins identified from the IgA panning experiment were mostly out of frame and in the BSA panning experiment were, with few exceptions, short (around 30 amino acids) and contained both tags, and therefore in frame. It would appear on the basis of this evidence that there is a selection for inserts which are out of frame, which are then corrected by the inherent ribosomal slippage feature of the phagemid vector during the translation process. However in some cases, particularly the IgA panning eluate, this slippage did not operate sufficiently to allow expression of many of the proteins from the library.

The choice of coat protein 8 for peptide display appeared straightforward at the beginning of this study as protein 8 – present in ~2700 copies - allows polyvalent display of fusion proteins, therefore opening up the possibility of identifying weaker binding interactions as well as strong from the panning process. This was thought to be a useful trait which would allow identification of proteins which did not demonstrate strong affinity for the ligands used but which may still have been involved in the binding process. However, the likelihood of identifying any fusion proteins with strong affinity was probably lowered by incorporating metagenomic DNA with the gene 8 phagemid vector. The reasoning behind this is that an increased variety of recombinant phage going into the panning experiments will have resulted in the elution of an increased variety of fusion proteins. Because the pool of 'interesting' proteins is now so large, the number of proteins from the panning eluate able to be analysed individually became a much lower proportion of the entire eluate pool. Some of these 'interesting' proteins demonstrated non-specific and weak affinity to the ligand which, because of a higher number of fusions on the phage surface, would bind tightly through increased



avidity. Indeed, this did happen with clones 30, 39 and 44, analysed with adhesion assays. It appears that the failure of these clones to bind back to the original ligand refuted their binding specificity in the first place. Panning works on the premise that non-specific clones are washed away, leaving those with a specific binding interaction. In reality however, washing may remove those clones which are specific, but which are weakly bound due to a reduced expression level, or those which have few fusions on the phage surface. These bound but non-specific 'decoy' sequences in this project absorbed a significant time investment in sequencing and analysis before finding that they demonstrate very little ligand specific binding, or that the inserted DNA sequence contained a frameshift or stop codons. It was estimated by Cesareni (1992) that there could be between 100 – 1000 fusions per phage, which means that the ability of a weakly binding fusion protein to appear as one with strong affinity, simply due to increased avidity, is very high (Cesareni, 1992). However, there is some disagreement regarding the number of fusion proteins appearing on the phage surface, with Clark and March stating in 2006 that larger proteins (25 kDa) can be displayed less than once per phage.

The Carcamo study also agrees with the Jacobsson work, stating that for some of their 'unusual sequences', expression of the inserted gene (b-gal) did occur, indicating that successful translation of clones containing frameshifts (either +1 or -1) is not a rare event. It would appear that a similar trend occurred in this study, where out of 253 clones analysed in antibody screening experiments (with inserts), 69 (27 %) contained multiple stop codons in more than one frame (Chapter 4, Table 2), but apparently still showed preferential binding to their respective ligand, which enabled enrichment following 3 rounds of panning. A spot check of 20 DNA sequences from the antibody screening experiments showed that around 20% of sequences with multiple stop codons in one frame had at least one other frame free of stop codons. This means that in these 20% of sequences it is likely that, following the frameshift, the sequence was still capable of producing a fusion protein. Unfortunately, due to time constraints, none of these proteins were analysed further.

One reason why the presence of these frameshifted sequences was so destructive was that they were numerically significant among recombinants analysed individually, contributing to the analysis time without resulting in a protein of interest which could be tested further. With hindsight, using coat protein 3 for monovalent (1 to 5 fusions) display rather than coat protein 8 for polyvalent display may have facilitated the identification of the most strongly binding proteins. Combining monovalent display (protein 3) with metagenomics, though eliminating some of the 'interesting' proteins with weaker binding affinity, would probably have resulted in fewer eluted proteins. A gene III based library would not have solved the problems of frameshifting as Carcamo (1998) described, and it would still have been subject to lower expression levels by the phage system therefore potentially missing out more proteins, but it would have resulted in fewer overall clones which should all have stronger affinity for their

respective ligands. So far, only 500 recombinants from the phage display library ( $10^{11}$ ) constructed during this project have been explored, leaving many more potential binding proteins unidentified.

In filamentous phage display, phage particles are assembled in the cytoplasmic membrane and secreted from the infected host without disrupting the integrity of the cell membrane (Mullen *et al.*, 2006). However, the very basis of this life cycle imposes limitations on peptide display, where some proteins assembled to form the hybrid capsid protein may have inherent properties which prevent the correct transfer through the lipid bilayer of the *E. coli* inner membrane (Castagnoli *et al.*, 2001). Some fusion proteins, whether in their native or unfolded state, may be too bulky to make it through the narrow pore for phage extrusion created by phage proteins 1, 4 and 11. It is also possible that some fusion proteins are toxic to the *E. coli* host and if a reduction in expression level continues to result in toxicity to the cell, frameshifting may occur and upon closer analysis, the fusion protein may appear to have no continuous ORF (Carcamo *et al.*, 1998), a situation which occurred in this project. Most likely, the selection for inserts out of frame is a natural way for the system to optimise the production of viable phage with recombinant proteins on the surface.

Fusion proteins may not make it through the bacterial cell membrane for other reasons. High G+C organisms contain promoters which may not be recognised by *E. coli* and therefore these DNA inserts may not be produced in this system. Phage display does not allow the total expression of a mixed bacterial proteome on the surface of filamentous phage and, as previously discussed, only 1-in-18 fusions will appear in frame on the phage surface. Proteins that are only produced under specific conditions would also be missed along with those that are quickly degraded and do not accumulate in the cell. Additionally, genes shortened by fragmentation prior to library construction may result in shortened proteins which are no longer able to fold in their native conformation. Unfolded proteins, if they make it out of the bacterial cell on the phage surface, are less resistant to proteolysis and may therefore never appear in the panning eluate.

Alternative phage display systems to filamentous phage do exist, in particular exploiting the lytic phage, whose life cycle involves lysing the host bacterial cell membrane. In the cases of, for example, T4 or T7 phage display, these virulent phage assemble in the cytoplasm and lyse the bacterial membrane, negating the need for secretion through the host cell membrane. These systems are being trialled by our group at the moment but they have their own complexities and problems associated with them.

### **Panning**

One of the errors made during this project was to use BSA as a 'control' ligand in panning experiments. Jacobsson and Frykberg discussed its successful use in many

publications; however they used it in phage display libraries containing pure *Staphylococcus aureus* DNA as they were under the impression that *S. aureus* did not bind to BSA. In the panning experiment in this thesis, the control was in place so that fusion proteins identified from the ('more important') FN or IgA experiments could be compared with those from the BSA experiment, and if any isolates were identified from both FN and BSA, or IgA and BSA, they clearly demonstrated no specific affinity to one or the other. Although there was no problem with this reasoning, the enormous diversity of recombinants in the panning eluate, which did not become clear until after the panning process and analysis was completed, meant that, in order to find out if a protein of interest also bound to the control, all clones from the BSA panning eluate would have to be analysed – a feat certainly not achievable during the time span for this project. The additional factor not considered prior to panning was that BSA could also be a ligand suited to bacterial interaction, possibly due to its structural similarities to HSA, the most abundant protein in human blood plasma known for its ability to bind to hydrophobic molecules.

There are many successful phage display studies in publication (using the genomes of single bacteria), all of which identified at least one bacterial protein from a variety of overlapping clones from the phage display library, and the same was expected from this project with the proviso that many more clones would require analysis. Because of the small proportion of clones analysed from each panning eluate, expecting to find overlapping clones was short-sighted at best. This was one of the issues thrown up by the combination of phage display and metagenomics which was not expected in the initial stages of the project. From a 2 ml elution following 3 rounds of panning (at  $1 \times 10^{11}$  CFU), 10  $\mu$ l of this eluate, enriched with potential bacterial binding proteins, was used to infect *E. coli* and spread on agar plates. Due to the enormous time investment required to analyse the clones, a maximum of 500 was ever investigated in sufficient detail to tentatively allocate nomenclature or function. This means that a further  $9.9 \times 10^{10}$  clones remained in the library for analysis, clearly illustrating that the analysis performed on the 500 clones provides only the most cursory indication of the potential bacterial proteins contained within each panning eluate. It is of course realised that of these  $9.9 \times 10^{10}$  clones, only 1 in 18 are in the correct orientation and in frame ( $5.5 \times 10^9$ ). Because of this and the labour intensive nature of analysis, it is unlikely that a great deal more progress could have been made by analysing more clones in the time allowed for this particular project. The panning eluates remain at UCL however, and are a valuable resource for more research into the proteins with affinity for FN or IgA. The original library also remains, which could be used to pan against other ligands.

With the benefit of hindsight, two alterations could have been made to the panning experiments which may have enhanced their success. Using any animal protein as a blocking agent could throw up unseen problems so perhaps the best solution would have been to use

either a plant derived protein (soy protein) or no blocking agent at all. Blocking binding to everything but the ligand of interest may have caused a small reduction in the overall number of fusion proteins retrieved from the analysis. Secondly, panning for 4, 5 and perhaps 6 rounds may have reduced the variety of fusion proteins eluted from panning, providing a more manageable binding 'proteome' for each ligand. Of course, additional rounds of panning would also have acted to exclude many other fusion proteins from the analysis, as more numerous phage were amplified further, some of these not specific binders at all, but being retained by the ligand due to the avidity effect.

A plethora of potential panning ligands could have been incorporated in this study, which would have provided very different results. This could include other salivary proteins (Introduction Table 1), other Immunoglobulins and mucins. Any human components used as ligands will further the current knowledge as to how and why bacteria and humans exist in synergy. In addition, the option of panning against other bacteria would provide an insight into inter-bacterial communication such as that already explored by other groups (Egland *et al.*, 2004). However, the results of this project do highlight a warning to other investigators; using metagenomics and phage display together will provide an enormous amount of material which needs to be analysed further. It is advisable to simplify experiments as much as possible, by using only one ligand perhaps, or by only looking for the strongest possible binders in a population.

### **Antibody screening**

Antibody screening was carried out following panning to rapidly eliminate clones which did not display in fusion with the screening tags, c-Myc and Poly-His, and therefore the coat protein 8. The elimination of clones not in fusion with the tags would therefore avoid mass sequencing all 300 recombinant clones and concentrate attention on those which would contain inserts expressed in frame with the vector and (theoretically) be displayed on the phage surface.

Antibody screening initially seemed to solve many problems as a rapid screening step which would accurately pinpoint the 1 in every 18 clones supposedly in the correct orientation and in frame with the phagemid vector. Once it was realised that the poly-His tag may not be identified in clones where the insert contained an inherent promoter and signal sequence and that one or both tags could be blocked by a large insert, antibody screening began to throw up more questions than it answered. It became apparent after antibody screening that some recombinants containing ORF's showed a negative result for one or both tags, and that some recombinants which showed a positive result for both tags contained multiple stop codons in all 3 frames, and therefore should not logically have given a positive result. This was particularly trying when, following sequencing, some 'positive' recombinants did not even contain an

insert, so the effort made to get these clones to the sequencing stage was completely in vain. The antibody screening results (Chapter 4, Table 2) clearly demonstrated the inherent library issues such as frameshifting and stop codons; clones affected by this were already present in the phage display library before panning and were seen in the antibody screening results table. Because the antibody screening results made no sense at the time, all 300 clones were eventually sequenced to try to find a logical explanation. As mentioned previously, clones containing an ORF should have provided a positive result for both tags, although it was possible that the tags might not appear for the reasons stated above. However, there appeared to be no possible explanation for clones containing no insert, or an insert in the ‘wrong’ frame as coat protein 8, to provide a positive result for tags on the phage surface.

Antibody screening however did bring to light the difference in number and size of fusion proteins identified from the BSA panning experiment in comparison to FN and IgA. The majority of the BSA eluate recombinants included both tags (Chapter 4, Figure 8 (a) and (b)) and were around 30 amino acids in size. Of the 77 proteins identified from panning against BSA, most of which were in frame with gene 8 of the vector, 61% started with the bases CCC, CCA or CCG. It could be possible that frameshifting in BSA binding proteins is more successful. This could be because, since most of the proteins binding to BSA were very small, that these proteins can cross the *E. coli* inner membrane more efficiently, and so appear in the BSA panning eluate in higher numbers. This would also benefit them during the multiple rounds of amplification during panning. Conversely, fusion phage with binding affinity to the ligands IgA or FN may be less common, resulting in more availability for ‘decoy’ phagemid (in the wrong frame) to bind to the ligand.

### **Individual Proteins**

None of the 18 shortlisted proteins shared any homology to the known albumin, FN or IgA binding domains tested. Of course, there are probably scores of bacterial binding proteins which do not fit any previously noted canonical binding sequence, for example FBP54 and SDH in *S. pyogenes* and PavA in *S. pneumoniae* (Christie et al, 2002). One possibility which was not accounted for in our studies was the likelihood that some proteins in the phage display library would not bind immobilized FN/IgA, only the soluble version, which is a real possibility but probably outside the scope of this projects remit for investigation.

Eight of the 18 proteins identified following panning of the phage display library were ‘hypothetical proteins’ or were unrecognised by public databases (Chapter 5, Table 3). This result fitted with the hypothesis that a variety of unknown bacterial proteins are partially responsible for interacting with host ligands. It is possible that some bacterial genes from little-studied organisms still have no function associated with them. In most metagenomic studies, even with closure, around 40% of genes cannot be assigned a function (Nichols 2007). The

result that almost half of the proteins identified in this study have no known function or do not match public databases means that the human tongue dorsum remains a potential source of novel bacterial proteins and/or binding mechanisms. In Jacobsson and Frykberg (1996) the complete genome of *S. aureus* was panned against IgG and binding clones containing overlapping inserts were used to identify the gene responsible for binding as protein A, a well known IgG binding ligand. A comprehensive analysis with a clear outcome was not possible with the phage display library constructed in this project for many reasons. One reason is that metagenomic DNA from the human tongue dorsum is not equally or fully represented in the public databases. As a result, many of the proteins which demonstrated binding affinity during panning still have no known function assigned to them. For these ‘unknown’ proteins, functional data was needed.

Although not previously discussed, mostly because there were no ‘results’ as such, is that protein gels were attempted during this project, with the aim of continuing to Western blotting. The protein gels were used initially to try to visualise the fusion phage compared to phagemid only, which would have provided information regarding the size of the fusion protein and the amount of protein each phagemid produced, but would not have revealed phage numbers, or exactly how many copies of the fusion appeared on the phage surface. The protein gels were problematic in that the bands on the gel were not able to be clarified, and no fusion proteins could ever be visualised in comparison to the control. This may have been because the protein gels were of such low quality, and/or that the staining itself was not clear (using Sypro Ruby or Coomassie Brilliant Blue). It is also possible that the gel resolution was not high enough to allow visualisation of the smaller fusion proteins (between 50 and 100 amino acids). Although incrementally higher proportions of acrylamide were used in successive gels to capture the fusion proteins, they were never visualised and it was thought illogical to proceed with Western blotting experiments.

A major part of the latter stages of this project was intended to be taken up by pET vector expression. The minimal set of requirements for gene expression includes the presence of a promoter for transcription, and a ribosome binding site (rbs) in the -20 to -1 region upstream of the start codon for initiation of translation. Both sites must be suitable for the expression machinery of the bacterial host cell, and this suitability may have been a major contributor to the downfall of pET vector expression in this project. Besides these *cis*-acting DNA sequences, the formation of an active protein may also rely on *trans* factors that need to be provided by the host organism such as special transcription factors, inducers, chaperones, cofactors, protein-modifying enzymes, or proper secretion machinery. Whether or not essential *trans* factors are present in the host is in most cases difficult or even impossible to predict (Gabor *et al.*, 2004). None of the proteins attempted in pET vector expression were successful; all failing at different stages, however it was considered unlikely that all of these proteins were

toxic to the *E. coli* cell, however with so many alternative issues to consider, it would be difficult to identify exactly why the proteins did not express. The main reason for the failure of pET expression is that there simply was not enough time to optimise the process for each clone individually. With 18 potential proteins to express and time running out, priority shifted to an attempt to prove specific binding of at least one clone to its panning ligand.

### **Binding affinity confirmation**

Adhesion assays in Chapter 5 provided preliminary data on the binding affinity of recombinant clones to the panning ligands which first identified them. The 8 recombinants chosen for antibody screening were taken from the top of the list (Chapter 5, Table 1) as the fusion proteins were larger and should theoretically have retained some folding ability over smaller proteins. However, due to time constraints, this did mean that the remaining 10 recombinants on the list were not involved in any further analysis to confirm binding specificity. Only 4 out of the 8 recombinants were tested in binding assays, the other 4 either did not transform into *E. coli* or were unsuccessful at the conversion stage. It seemed unlikely that some recombinants would not amplify in pure phage populations as they had done previously during panning, but there did not seem to be an adequate explanation why some individual clones did not respond to amplification. In all 4 failed cases, the *E. coli* host cells did not grow following either transformation with the phagemid or conversion to phage.

In the binding assay, clone 39 did not show preferential binding for the panning ligand which identified it (IgA), apparently showing greater binding affinity for FN. Ostensibly, this apparent shift in affinity seems unlikely given the specific nature of protein-ligand interactions. The 83 amino acid protein displayed on clone 39 may be showing greater affinity for FN than IgA in adhesion assays, but it is possible that, had analysis of more clones from the FN panning eluate taken place, clone 39 would appear more often in that population than in the IgA population. It is also possible that clone 39 had a high avidity reaction with IgA, and was retained in that population over the FN population.

From the 4 binding assays carried out, one clone in particular did indicate that it could have affinity to the ligand which identified it from panning. Clone 59 showed clear reproducible binding affinity to IgA, without any outlying results, making it a promising clone for further experiments, potentially testing binding strength. Unfortunately, experiments with clones 39, 44 and 30 gave some aberrant results which, because these experiments were the last ones carried out prior to thesis writing, left no time to incorporate changes to the method. Even without further experiments, there are clear indications that fusion proteins identified from panning experiments do show preferential binding to the respective ligands. It is worth noting the potential reason for phage identified by panning not showing affinity for the original ligand. Phage enrichment takes place on the premise that bound phage will remain bound during the

washing process of panning prior to phage elution. Phage that are specifically bound but not tightly bound (perhaps through low affinity or low expression levels) will be removed during the washing steps, including those which adhere to the plastic or blocking material, leaving phage which may not be specifically bound but which have high affinity through polyvalent avidity or other interactions.

Other alternatives were available which could have been used to identify binding specificity such as surface plasmon resonance (SPR) – commonly used to measure adsorption of a solute (in this case, our protein of interest) onto a surface coated with the corresponding ligand, FN or IgA (Mullen *et al*, 2008). GST fusions were also an option, previously discussed in Chapter 5. Alternatives like this were considered but due to a lack of time only adhesion assays were used.

### **16S rRNA gene analysis**

The common nature of 16S rRNA gene analysis makes it a valuable comparative tool in genomics. 16S rRNA gene analysis was used in the present study to corroborate findings from the phage display library regarding bacterial proteins from panning, and to help identify any shortcomings in the phage display library such as a reduction in bacterial diversity through each stage and into the panning process. However, the phage display process provided a much less comprehensive analysis of bacterial proteins than was first imagined; making the 16S rRNA gene analysis in this study very important, in both assessing the range of bacterial species from the initial metagenomic DNA sample and providing a basis for comparison between the present study and others in the area. Although it did not provide any functional information, population inferences regarding species were helpful for tentative agreement on the species variety between the whole sample (16S rRNA) and variety following phage display and panning.

The 16S rRNA gene analysis in the present study (Chapter 6) was comparable with most studies carried out on or including the human tongue microbiota, in that most studies have varied results, mostly due to employing different extraction methods with a variety of primers and occasionally changing the number of PCR cycles. The results from this study were compared closely to research carried out by Riggio (2008) and Kazor (2003) in particular, which showed differences in the abundance of certain species. The metagenomic DNA used in this project consisted mainly of *Veillonella* (26.7 %) and *Streptococcus* (24.6 %), both well-known oral genera, but did not recognise a major contribution of *Actinomyces* or *Lysobacter*, which the Riggio (2008) study identified as important parts of the tongue microbial community.

Leaving aside individual discretions which can occur for many reasons, the 16S rRNA analysis provided a reasonable assessment of metagenomic DNA diversity. Results from phage display functional analysis and 16S rRNA diversity analysis were then compared.



### Comparison of 16S and phage display analyses

In order to try to assess the level of diversity in the phage display library, sequence data from each stage of the phage display and panning process were compiled. Sequence data from the 500 panning clones analysed were used to infer the diversity of the phage display library and compared against the 16S rRNA analysis results to identify points of disparity. For example, a species identified from 16S rRNA analysis as being present in the original metagenomic library, but which was not seen in the panning analysis of 500 clones could be explained in several ways. Perhaps, as in the case of high G+C organisms, promoter regions were not recognised by the *E. coli* host and therefore these organisms (and their related proteins) were not represented in the phage display library. Alternatively, a bacterial phylotype may have been present in very low numbers in the metagenomic DNA, notably those bacteria present in low numbers in the bacterial community, the 16S rRNA amplification of which could have resulted in identification in the diversity analysis, but which may have appeared in a very low proportion of phage display clones. It is well known that oral bacteria can produce biofilms, which form in layers consisting of primary colonisers which bind to the host, and secondary/tertiary colonisers which bind to the primary colonisers and some late arrivals. As the most abundant genus from 16S rRNA analysis, making up almost 27% of clones sequenced, *Veillonella* sp. was identified surprisingly little from phage display and panning clones, and appears only twice in the final 18 clones for further analysis. *Veillonella* is a Gram negative oral commensal which is normally found in synergistic association with Gram positive *Streptococcus* sp., where it utilizes the lactate produced by *Streptococcus* for growth. Incidentally, *Streptococcus*, present at 25% according to 16S rRNA analysis, was also identified in lower numbers than expected from sequence data of the 500 clones and appears only once in the final 18 clones, surprising since *Streptococcus* sp. are known to encode many FN binding proteins. The low presence of *Veillonella* sp. in the panning eluate could be explained by the fact that *Veillonella* acts as a secondary or tertiary coloniser of the tongue surface, and the low appearance of *Streptococcus* could indicate that these two organisms fill a similar ecological role on the tongue dorsum. Although colonisation on the tooth surface is known in detail, the same is not known for the tongue, and the species which act as early and late colonisers to the tooth surface may not on the tongue. On the tooth, proline rich proteins (PRP's) and statherin coat enamel which are recognised by the early colonisers (mostly *Streptococci*), which have the distinct advantage of the expression of multiple adhesins for receptor recognition (Jenkinson & Lamont, 2005). Even though, on epithelial surfaces, mucin and agglutinins are the substrates for bacterial interaction, the multiple adhesins of *Streptococcus* sp. should enable them to carry out the same ecological role all over the oral cavity.

It was hoped that data from successive panning rounds would provide information regarding the enrichment of certain types of proteins (for example, membrane proteins or transport proteins) as round 1 progressed to round 3. Indeed, in the shortlisted 18 clones there are 3 supposed ABC transporters, a SecA component for protein transport and a chloride channel, as well as several hypothetical proteins which could have the same roles. Interestingly, most of the sequence data showed mid to low (30 – 60%) homology to the NCBI public database using BLASTn and tBLASTx (only 25 - 36% showed any similarity to the BLASTn database), which makes it very difficult to identify trends imposed by library construction or panning. This information in itself seems unusual. The oral cavity supposedly consists of 50% cultured bacteria, a high number given the low percentage of cultured bacteria from other environments, so the expectation was that BLAST homology of the panning clones would show a clear trend regarding the function of the supposed binding proteins. Given the number of hypothetical proteins, this was not possible.

The appearance of only 16 ‘novel’ phylotypes from a 16S rRNA analysis of 333 clones (Chapter 6, Table 1), would suggest that much of the bacterial species present in the phage display library are indeed from known bacteria, suggesting that perhaps full genome sequences are not yet completed or fully annotated, so clones identified during this analysis may not appear. Further, it is not impossible that certain genes used to facilitate bacterial binding as yet have no functional data associated with them. Of course, the similarity cut-off used for the 16S rRNA analysis in this project was 98% where others have used 97%, and that extra 1% may have drastically increased the number of ‘novel’ phylotypes from this analysis.

An additional comparative observation is that *Actinomyces* sp., present at 0.01% according to the 16S rRNA analysis, appeared several times in the phage display analysis, making up 4 of the shortlisted 18 clones. This increase implies that either *Actinomyces* sp. are present at higher numbers in the metagenomic DNA but are not suitably amplified by PCR, or that *Actinomyces* sp. express proteins with very strong affinity for the ligands FN and IgA. The first explanation seems more likely, although this reasoning was contradicted by Riggio, 2008, who found 5 – 10% *Actinomyces* sp. in their metagenomic DNA sample, but as a consequence of using different primers. However, the very presence of *Actinomyces* in the phage display library refutes the earlier comment that the promoters of high G+C organisms (like *Actinomycetes*) are not recognised by *E. coli* hosts, as clearly some of them are.

Other, more detailed conclusions are difficult to draw. It was hoped that most panning clones would clearly identify bacteria and proteins responsible for binding, allowing sensible trends to be noted. For example, in a hypothetical library, the oral bacterium/pathogen *Fusobacterium nucleatum*, identified as a minor player from 16S rRNA analysis would be expected to decrease in numbers following panning since it is not normally found binding to the tongue dorsum, and is in fact a known periodontal pathogen (Kaplan *et al.*, 2009).

From this comparison of molecular techniques using the same source of metagenomic DNA clearly phage display libraries do not incorporate the representative mixed microbial community as fusion proteins. The omission of two key genera, *Veillonella* and *Prevotella*, from the phage display library, but which were identified as major players during the 16S rRNA diversity analysis, could be attributed to any of the phage display nuances previously discussed including vector choice, coat protein used for fusion or choice of host bacterium, or could simply be due to the ecological role of the organisms as secondary colonisers, which would mean these organisms do not normally adhere to FN or IgA.

### **pQR492**

The analysis of a 6 kb bacterial DNA fragment required taking a different analytical approach where importance was placed on investigating the role of the genes contained within. This analysis was a more complete version of that which could eventually have been carried out on the individual proteins identified from panning, and a great deal of information was found relating to pQR492, mostly because the inserted fragment was larger than any of the panning proteins. The organism from which pQR492 originated was probably a high G+C bacterium, most likely an Actinomycete given the homology of every BLAST analysis carried out on pQR492. The estimation that pQR492 contained 4 ORF's within the same operon was probably correct, given that the 4 ORF's (1-4) appeared to serve very similar functions and shared homology to the same protein, *cauri\_0414* of *C. aurimucosum*- although it appeared that the protein contained within pQR492 differed in its folding or domain architecture given the presence of repeated domains which were not present in *cauri\_0414*. Although gene start site prediction is an inexact science, some possibilities were identified for RBS's for each ORF 1 – 4. Whether pQR492 does in fact contain a membrane protein, the supposed function of *cauri\_0414*, the closest homologue to ORFs 1-4 of pQR492, is not known however the presence of a signal sequence and two transmembrane domains in ORF 1 suggests that this protein may indeed play some role at the cell surface.

The part of transcription repair coupling factor (TRCF) also identified in pQR492 contained a highly conserved AAA module (ATPases associated with diverse cellular activities). Although AAA modules have many diverse roles, one of these is in the bacterial membrane where proteins of this type are involved in protein degradation and unfolding. The analysis on pQR492 taken together suggests that the ORF's 1 – 5 could all play a role at the cell membrane.

### **Future experiments**

Given more time, the following experiments would be a useful continuation of the work set out in this thesis:

From the existing clones identified during this project:

- 1) Continue with binding assays of remaining proteins from panning.
- 2) Attempt pET expression with the most promising binding clones from binding assays. If pET expression continues to fail, GST vector expression would be a logical next step.

With the existing phage display library:

- 3) Pan the existing library against FN and IgA again, using a plant-based protein as a control
- 4) Pan the existing library using for up to 6 panning rounds

### **Conclusions**

This project brings together several molecular biology techniques including shotgun cloning (Chapter 7), 16S rRNA diversity analysis (Chapter 6), metagenomics and phage display (Chapter 3) with which to interrogate the microbiota residing on the tongue dorsum. In phage display, combining diverse ligands such as IgA and Fibronectin against a low stringency selection process was designed to provide as many bacterial binding proteins for individual study as possible. The use of a protein 8 based phagemid vector was based on the success of colleagues; however the combination of only 1 in 18 clones being in the correct orientation and in frame plus the use of metagenomic DNA to make the library, did cause the entire project to become rather unfocussed due to the very large number of clones from the panning eluate. The gene 8 based system increased target binding avidity through polyvalent display, which, along with the metagenomic DNA inserts, vastly increased the number and variety of 'binding clones'. Analysis of these clones became a gargantuan task, and meant that only a fraction of the panning eluate was ever studied. The issues encountered could not have been foreseen and the analysis which was carried out revealed a wide range of binding proteins which could be used by bacteria to bind to the human tongue dorsum and interact with the immediate environment.

The tongue dorsum has previously been included in dental studies due to its tendency to harbour periodontal pathogens, and the relationship between their presence on the tongue and transmission to other areas of the mouth has fascinated dental researchers. However, the tongue dorsum is a fascinating environment on its own, unique in the human body with an indigenous microbiota unlike any other area. Fitting with this and other dedicated studies, it is

an ideal environment to search for bacterial binding proteins due to the importance of bacterial adhesion in this area, and the mechanisms the human body uses to try to thwart such adhesion. Of course, the conditions under which binding interactions were investigated during this study are in no way representative of the conditions under which these interactions would normally take place. However, the potential of a huge variety of bacterial proteins to bind human ligands has been presented which deserves further investigation.

Through the panning of a metagenomic phage display library, several distinct and interesting proteins have been identified, as well as some 'hypothetical' proteins which may have novel functions involved in facilitating bacterial binding to the human ligands IgA and FN. Much more research into the specific nature of each interaction is needed to shed light into the field of bacteria-human interaction. This thesis highlights that there are numerous microbial species currently using a variety of proteins of unknown function to interact with humans to facilitate binding. The diversity of these interactions has been insinuated through the present study and the existence of the phage display libraries will allow far more detailed analysis to be carried out on these and other proteins. Furthermore, the combination of phage display and metagenomics could provide many other opportunities to identify bacterial proteins responsible for interaction.

---

## **Reference List**

---

## Reference List

- Aas,J.A., Paster,B.J., Stokes,L.N., Olsen,I., and Dewhirst,F.E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* *43*, 5721-5732.
- Abraham,S.N., Sun,D., Dale,J.B., and Beachey,E.H. (1988). Conservation of the D-mannose-adhesion protein among type 1 fimbriated members of the family Enterobacteriaceae. *Nature* *336*, 682-684.
- Ahl,T. and Reinholdt,J. (1991). Detection of immunoglobulin A1 protease-induced Fab alpha fragments on dental plaque bacteria. *Infect. Immun.* *59*, 563-569.
- Alisky,J., Iczkowski,K., Rapoport,A., and Troitsky,N. (1998). Bacteriophages show promise as antimicrobial agents. *J. Infect.* *36*, 5-15.
- Amann,R.I., Ludwig,W., and Schleifer,K.H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* *59*, 143-169.
- az-Torres,M.L., McNab,R., Spratt,D.A., Villedieu,A., Hunt,N., Wilson,M., and Mullany,P. (2003). Novel tetracycline resistance determinant from the oral metagenome. *Antimicrob. Agents Chemother.* *47*, 1430-1432.
- Azzazy,H.M. and Highsmith,W.E., Jr. (2002). Phage display technology: clinical applications and recent innovations. *Clin. Biochem.* *35*, 425-445.
- Backhed,F., Ding,H., Wang,T., Hooper,L.V., Koh,G.Y., Nagy,A., Semenkovich,C.F., and Gordon,J.I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl. Acad. Sci. U. S. A* *101*, 15718-15723.
- Bailey,M. (1995). Extraction of DNA From the Phylosphere. In: *Nucleic Acids in the Environment: Methods and Applications*, ed. J.Trevors and Elsas JDvBerlin: Springer-Verlag, 89-109.
- Benhar,I. (2001). Biotechnological applications of phage and cell display. *Biotechnol. Adv.* *19*, 1-33.
- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., and Sayers,E.W. (2009). GenBank. *Nucleic Acids Res.* *37*, D26-D31.
- Breitbart,M., Hewson,I., Felts,B., Mahaffy,J.M., Nulton,J., Salamon,P., and Rohwer,F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* *185*, 6220-6223.
- Brotz-Oesterhelt,H., Bandow,J.E., and Labischinski,H. (2005). Bacterial proteomics and its role in antibacterial drug discovery. *Mass Spectrom. Rev.* *24*, 549-565.
- Brussow,H. and Hendrix,R.W. (2002). Phage genomics: small is beautiful. *Cell* *108*, 13-16.
- Capparelli,R., Parlato,M., Borriello,G., Salvatore,P., and Iannelli,D. (2007). Experimental phage therapy against *Staphylococcus aureus* in mice. *Antimicrob. Agents Chemother.* *51*, 2765-2773.
- Carcamo,J., Ravera,M.W., Brissette,R., Dedova,O., Beasley,J.R., am-Moghe,A., Wan,C., Blume,A., and Mandeck,W. (1998). Unexpected frameshifts from gene to expressed protein in a phage-displayed peptide library. *Proc. Natl. Acad. Sci. U. S. A* *95*, 11146-11151.
- Castagnoli,L., Zucconi,A., Quondam,M., Rossi,M., Vaccaro,P., Panni,S., Paoluzi,S., Santonico,E., Dente,L., and Cesareni,G. (2001). Alternative bacteriophage display systems. *Comb. Chem. High Throughput. Screen.* *4*, 121-133.
- Cesareni,G. (1992). Peptide display on filamentous phage capsids. A new powerful tool to study protein-ligand interaction. *FEBS Lett.* *307*, 66-70.
- Childers,N.K., Bruce,M.G., and McGhee,J.R. (1989). Molecular mechanisms of immunoglobulin A defense. *Annu. Rev. Microbiol.* *43*, 503-536.

- Christensen,D.J., Gottlin,E.B., Benson,R.E., and Hamilton,P.T. (2001). Phage display for target-based antibacterial drug discovery. *Drug Discov. Today* 6, 721-727.
- Christie,J., McNab,R., and Jenkinson,H.F. (2002). Expression of fibronectin-binding protein FbpA modulates adhesion in *Streptococcus gordonii*. *Microbiology* 148, 1615-1625.
- Chu,K.H., Li,C.P., and Qi,J. (2006). Ribosomal RNA as molecular barcodes: a simple correlation analysis without sequence alignment. *Bioinformatics*. 22, 1690-1701.
- Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam,S.A., McGarrell,D.M., Garrity,G.M., and Tiedje,J.M. (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* 33, D294-D296.
- Collado,M.C., Isolauri,E., and Salminen,S. (2008). Specific probiotic strains and their combinations counteract adhesion of *Enterobacter sakazakii* to intestinal mucus. *FEMS Microbiol. Lett.*
- Couchman,J.R., Gibson,W.T., Thom,D., Weaver,A.C., Rees,D.A., and Parish,W.E. (1979). Fibronectin distribution in epithelial and associated tissues of the rat. *Arch. Dermatol. Res.* 266, 295-310.
- Courtois,S. *et al.* (2003). Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl. Environ. Microbiol.* 69, 49-55.
- Cowan,D., Meyer,Q., Stafford,W., Muyanga,S., Cameron,R., and Wittwer,P. (2005). Metagenomic gene discovery: past, present and future. *Trends Biotechnol.* 23, 321-329.
- Cox-Foster,D.L. *et al.* (2007). A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318, 283-287.
- Cramer,R. and Suter,M. (1993). Display of biologically active proteins on the surface of filamentous phages: a cDNA cloning system for selection of functional gene products linked to the genetic information responsible for their production. *Gene* 137, 69-75.
- D'Costa,V.M., Griffiths,E., and Wright,G.D. (2007). Expanding the soil antibiotic resistome: exploring environmental diversity. *Curr. Opin. Microbiol.* 10, 481-489.
- Daniel,R. (2004). The soil metagenome--a rich resource for the discovery of novel natural products. *Curr. Opin. Biotechnol.* 15, 199-204.
- Daniel,R. (2005). The metagenomics of soil. *Nat. Rev. Microbiol.* 3, 470-478.
- DeLong,E.F. (1997). Marine microbial diversity: the tip of the iceberg. *Trends Biotechnol.* 15, 203-207.
- DeSantis,T.Z., Jr., Hugenholtz,P., Keller,K., Brodie,E.L., Larsen,N., Piceno,Y.M., Phan,R., and Andersen,G.L. (2006). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* 34, W394-W399.
- DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P., and Andersen,G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069-5072.
- Deshayes,K., Schaffer,M.L., Skelton,N.J., Nakamura,G.R., Kadkhodayan,S., and Sidhu,S.S. (2002). Rapid identification of small binding motifs with high-throughput phage display: discovery of peptidic antagonists of IGF-1 function. *Chem. Biol.* 9, 495-505.
- Donaldson,A.C., McKenzie,D., Riggio,M.P., Hodge,P.J., Rolph,H., Flanagan,A., and Bagg,J. (2005). Microbiological culture analysis of the tongue anaerobic microflora in subjects with and without halitosis. *Oral Dis.* 11 Suppl 1, 61-63.
- Doyle,R.J. (2000). Contribution of the hydrophobic effect to microbial infection. *Microbes. Infect.* 2, 391-400.
- Dresselhuis,D.M., Stuart,M.A., van Aken,G.A., Schipper,R.G., and de Hoog,E.H. (2008). Fat retention at the tongue and the role of saliva: adhesion and spreading of 'protein-poor' versus 'protein-rich' emulsions. *J. Colloid Interface Sci.* 321, 21-29.



- du Toit,D.F. (2003). The tongue: structure and function relevant to disease and oral health. *SADJ*. 58, 375-3.
- Dumas,C., Champagne,A., and Lavoie,M.C. (1987). Proteolytic activity of bacteria isolated from the oral cavities of BALB/c mice toward salivary proteins. *J. Dent. Res.* 66, 62-64.
- Dziewanowska,K., Carson,A.R., Patti,J.M., Deobald,C.F., Bayles,K.W., and Bohach,G.A. (2000). Staphylococcal fibronectin binding protein interacts with heat shock protein 60 and integrins: role in internalization by epithelial cells. *Infect. Immun.* 68, 6321-6328.
- Edwards,A.M., Grossman,T.J., and Rudney,J.D. (2006). *Fusobacterium nucleatum* transports noninvasive *Streptococcus cristatus* into human epithelial cells. *Infect. Immun.* 74, 654-662.
- Edwards,A.M., Grossman,T.J., and Rudney,J.D. (2007). Association of a high-molecular weight arginine-binding protein of *Fusobacterium nucleatum* ATCC 10953 with adhesion to secretory immunoglobulin A and coaggregation with *Streptococcus cristatus*. *Oral Microbiol. Immunol.* 22, 217-224.
- Egland,P.G., Palmer,R.J., Jr., and Kolenbrander,P.E. (2004). Interspecies communication in *Streptococcus gordonii*-*Veillonella atypica* biofilms: signaling in flow conditions requires juxtaposition. *Proc. Natl. Acad. Sci. U. S. A* 101, 16917-16922.
- Ellis,R.J., Morgan,P., Weightman,A.J., and Fry,J.C. (2003). Cultivation-dependent and -independent approaches for determining bacterial diversity in heavy-metal-contaminated soil. *Appl. Environ. Microbiol.* 69, 3223-3230.
- Faveri,M., Feres,M., Shibli,J.A., Hayacibara,R.F., Hayacibara,M.M., and de Figueiredo,L.C. (2006). Microbiota of the dorsum of the tongue after plaque accumulation: an experimental study in humans. *J. Periodontol.* 77, 1539-1546.
- Ferrer,M., Golyshina,O.V., Chernikova,T.N., Khachane,A.N., Martins,D.S., V, Yakimov,M.M., Timmis,K.N., and Golyshin,P.N. (2005). Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem. Biol.* 12, 895-904.
- Fierer,N. *et al.* (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* 73, 7059-7066.
- Fine,D.H., Velliyagounder,K., Furgang,D., and Kaplan,J.B. (2005). The *Actinobacillus actinomycetemcomitans* autotransporter adhesin Aae exhibits specificity for buccal epithelial cells from humans and old world primates. *Infect. Immun.* 73, 1947-1953.
- Finn,R.D. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res.* 36, D281-D288.
- Flanagan,J.L. *et al.* (2007). Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with *Pseudomonas aeruginosa*. *J. Clin. Microbiol.* 45, 1954-1962.
- Flanagan,R.C., Neal-McKinney,J.M., Dhillon,A.S., Miller,W.G., and Konkel,M.E. (2009). Examination of *Campylobacter jejuni* putative adhesins leads to the identification of a new protein, designated FlpA, required for chicken colonization. *Infect. Immun.* 77, 2399-2407.
- Foster,J.S. and Kolenbrander,P.E. (2004). Development of a multispecies oral bacterial community in a saliva-conditioned flow cell. *Appl. Environ. Microbiol.* 70, 4340-4348.
- Frank,J.A., Reich,C.I., Sharma,S., Weisbaum,J.S., Wilson,B.A., and Olsen,G.J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl. Environ. Microbiol.* 74, 2461-2470.
- Frick,I.M., Akesson,P., Cooney,J., Sjobring,U., Schmidt,K.H., Gomi,H., Hattori,S., Tagawa,C., Kishimoto,F., and Bjorck,L. (1994). Protein H--a surface protein of *Streptococcus pyogenes* with separate binding sites for IgG and albumin. *Mol. Microbiol.* 12, 143-151.
- Gabor,E.M., de Vries,E.J., and Janssen,D.B. (2004). Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environ. Microbiol.* 6, 948-958.

- Geiger,B., Bershadsky,A., Pankov,R., and Yamada,K.M. (2001). Transmembrane crosstalk between the extracellular matrix--cytoskeleton crosstalk. *Nat. Rev. Mol. Cell Biol.* 2, 793-805.
- Geisow,M.J. and Beaven,G.H. (1977). Physical and binding properties of large fragments of human serum albumin. *Biochem. J.* 163, 477-484.
- Gill,S.R., Pop,M., Deboy,R.T., Eckburg,P.B., Turnbaugh,P.J., Samuel,B.S., Gordon,J.I., Relman,D.A., Fraser-Liggett,C.M., and Nelson,K.E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355-1359.
- Gillespie,D.E., Brady,S.F., Bettermann,A.D., Cianciotto,N.P., Liles,M.R., Rondon,M.R., Clardy,J., Goodman,R.M., and Handelsman,J. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* 68, 4301-4306.
- Ginolhac,A. *et al.* (2004). Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl. Environ. Microbiol.* 70, 5522-5527.
- Gnanasekar,M., Rao,K.V., He,Y.X., Mishra,P.K., Nutman,T.B., Kaliraj,P., and Ramaswamy,K. (2004). Novel phage display-based subtractive screening to identify vaccine candidates of *Brugia malayi*. *Infect. Immun.* 72, 4707-4715.
- Goo,S.Y., Lee,H.J., Kim,W.H., Han,K.L., Park,D.K., Lee,H.J., Kim,S.M., Kim,K.S., Lee,K.H., and Park,S.J. (2006). Identification of OmpU of *Vibrio vulnificus* as a fibronectin-binding protein and its role in bacterial pathogenesis. *Infect. Immun.* 74, 5586-5594.
- Gordon,H.A. and Pesti,L. (1971). The gnotobiotic animal as a tool in the study of host microbial relationships. *Bacteriol. Rev.* 35, 390-429.
- Gordon,R.E. and Hagan,W.A. (1938). The Classification of Acid-Fast Bacteria. II. *J. Bacteriol.* 36, 39-46.
- Hallberg,K., Hammarstrom,K.J., Falsen,E., Dahlen,G., Gibbons,R.J., Hay,D.I., and Stromberg,N. (1998). *Actinomyces naeslundii* genospecies 1 and 2 express different binding specificities to N-acetyl-beta-D-galactosamine, whereas *Actinomyces odontolyticus* expresses a different binding specificity in colonizing the human mouth. *Oral Microbiol. Immunol.* 13, 327-336.
- Hammerschmidt,S., Talay,S.R., Brandtzaeg,P., and Chhatwal,G.S. (1997). SpsA, a novel pneumococcal surface protein with specific binding to secretory immunoglobulin A and secretory component. *Mol. Microbiol.* 25, 1113-1124.
- Handelsman,J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669-685.
- Handelsman,J., Rondon,M.R., Brady,S.F., Clardy,J., and Goodman,R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245-R249.
- Haraszthy,V.I., Zambon,J.J., Sreenivasan,P.K., Zambon,M.M., Gerber,D., Rego,R., and Parker,C. (2007). Identification of oral bacterial species associated with halitosis. *J. Am. Dent. Assoc.* 138, 1113-1120.
- Harmsen,H.J., Raangs,G.C., He,T., Degener,J.E., and Welling,G.W. (2002). Extensive set of 16S rRNA-based probes for detection of bacteria in human feces. *Appl. Environ. Microbiol.* 68, 2982-2990.
- Harrington,E.D., Singh,A.H., Doerks,T., Letunic,I., von,M.C., Jensen,L.J., Raes,J., and Bork,P. (2007). Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc. Natl. Acad. Sci. U. S. A* 104, 13913-13918.
- Hattori,M. and Taylor,T.D. (2009). The human intestinal microbiome: a new frontier of human biology. *DNA Res.* 16, 1-12.

- Heinrichs, J.H., Bayer, M.G., and Cheung, A.L. (1996). Characterization of the sar locus and its interaction with agr in *Staphylococcus aureus*. *J. Bacteriol.* *178*, 418-423.
- Helmerhorst, E.J. and Oppenheim, F.G. (2007). Saliva: a dynamic proteome. *J. Dent. Res.* *86*, 680-693.
- Hemm, M.R., Paul, B.J., Schneider, T.D., Storz, G., and Rudd, K.E. (2008). Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol. Microbiol.* *70*, 1487-1501.
- Henderson, B., Poole, S., and Wilson, M. (1996). Microbial/host interactions in health and disease: who controls the cytokine network? *Immunopharmacology* *35*, 1-21.
- Hong, K.S., Lim, H.K., Chung, E.J., Park, E.J., Lee, M.H., Kim, J.C., Choi, G.J., Cho, K.Y., and Lee, S.W. (2007). Selection and characterization of forest soil metagenome genes encoding lipolytic enzymes. *J. Microbiol. Biotechnol.* *17*, 1655-1660.
- Hoogenboom, H.R., De Bruine, A.P., Hufton, S.E., Hoet, R.M., Arends, J.W., and Roovers, R.C. (1998). Antibody phage display technology and its applications. *Immunotechnology.* *4*, 1-20.
- Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics.* *20*, 2317-2319.
- Hugenholtz, P. and Pace, N.R. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol.* *14*, 190-197.
- Jacobsson, K. and Frykberg, L. (1995). Cloning of ligand-binding domains of bacterial receptors by phage display. *Biotechniques* *18*, 878-885.
- Jacobsson, K. and Frykberg, L. (1996). Phage display shot-gun cloning of ligand-binding domains of prokaryotic receptors approaches 100% correct clones. *Biotechniques* *20*, 1070-1071.
- Jacobsson, K., Jonsson, H., Lindmark, H., Guss, B., Lindberg, M., and Frykberg, L. (1997). Shotgun phage display mapping of two streptococcal cell-surface proteins. *Microbiol. Res.* *152*, 121-128.
- Jacobsson, K., Rosander, A., Bjerketorp, J., and Frykberg, L. (2003). Shotgun Phage Display - Selection for Bacterial Receptors or other Exported Proteins. *Biol. Proced. Online.* *5*, 123-135.
- Jenkinson, H.F. and Lamont, R.J. (2005). Oral microbial communities in sickness and in health. *Trends Microbiol.* *13*, 589-595.
- Joh, D., Speziale, P., Gurusiddappa, S., Manor, J., and Hook, M. (1998). Multiple specificities of the staphylococcal and streptococcal fibronectin-binding microbial surface components recognizing adhesive matrix molecules. *Eur. J. Biochem.* *258*, 897-905.
- Joh, D., Wann, E.R., Kreikemeyer, B., Speziale, P., and Hook, M. (1999). Role of fibronectin-binding MSCRAMMs in bacterial adherence and entry into mammalian cells. *Matrix Biol.* *18*, 211-223.
- Johnsson, E., Areschoug, T., Mestecky, J., and Lindahl, G. (1999). An IgA-binding peptide derived from a streptococcal surface protein. *J. Biol. Chem.* *274*, 14521-14524.
- Jones, B.V. and Marchesi, J.R. (2007). Accessing the mobile metagenome of the human gut microbiota. *Mol. Biosyst.* *3*, 749-758.
- Jonsson, K., Signas, C., Muller, H.P., and Lindberg, M. (1991). Two different genes encode fibronectin binding proteins in *Staphylococcus aureus*. The complete nucleotide sequence and characterization of the second gene. *Eur. J. Biochem.* *202*, 1041-1048.
- Joseph, S.J., Hugenholtz, P., Sangwan, P., Osborne, C.A., and Janssen, P.H. (2003). Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl. Environ. Microbiol.* *69*, 7210-7215.
- Kang, H.Y., Kim, J., Seol, S.Y., Lee, Y.C., Lee, J.C., and Cho, D.T. (2009). Characterization of conjugative plasmids carrying antibiotic resistance genes encoding 16S rRNA methylase,

- extended-spectrum beta-lactamase, and/or plasmid-mediated AmpC beta-lactamase. *J. Microbiol.* *47*, 68-75.
- Kaplan,C.W., Lux,R., Haake,S.K., and Shi,W. (2009). The *Fusobacterium nucleatum* outer membrane protein RadD is an arginine-inhibitable adhesin required for inter-species adherence and the structured architecture of multispecies biofilm. *Mol. Microbiol.* *71*, 35-47.
- Kara,D., Luppens,S.B., van,M.J., Ozok,R., and Ten Cate,J.M. (2007). Microstructural differences between single-species and dual-species biofilms of *Streptococcus mutans* and *Veillonella parvula*, before and after exposure to chlorhexidine. *FEMS Microbiol. Lett.* *271*, 90-97.
- Kauffmann,I.M., Schmitt,J., and Schmid,R.D. (2004). DNA isolation from soil samples for cloning in different hosts. *Appl. Microbiol. Biotechnol.* *64*, 665-670.
- Kazor,C.E., Mitchell,P.M., Lee,A.M., Stokes,L.N., Loesche,W.J., Dewhirst,F.E., and Paster,B.J. (2003). Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. *J. Clin. Microbiol.* *41*, 558-563.
- Kehoe,J.W. and Kay,B.K. (2005). Filamentous phage display in the new millennium. *Chem. Rev.* *105*, 4056-4072.
- Keijsers,B.J., Zaura,E., Huse,S.M., van,d., V, Schuren,F.H., Montijn,R.C., Ten Cate,J.M., and Crielaard,W. (2008). Pyrosequencing analysis of the Oral Microflora of healthy adults. *J. Dent. Res.* *87*, 1016-1020.
- Kennedy,J., Marchesi,J.R., and Dobson,A.D. (2007). Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Appl. Microbiol. Biotechnol.* *75*, 11-20.
- Kierek-Pearson,K. and Karatan,E. (2005). Biofilm development in bacteria. *Adv. Appl. Microbiol.* *57*, 79-111.
- Kilian,M. (2003). Bacterial Immunoglobulin-Evading Mechanisms: Ig-Degrading and Ig-Binding Proteins. In: *Bacterial Evasion of Host Immune Responses*, ed. B.Henderson and P.Oyston Cambridge University Press, 81-102.
- Koch,A.L., Higgins,M.L., and Doyle,R.J. (1982). The role of surface stress in the morphology of microbes. *J. Gen. Microbiol.* *128*, 927-945.
- Kolenbrander,P.E., Andersen,R.N., Blehert,D.S., Eglund,P.G., Foster,J.S., and Palmer,R.J., Jr. (2002). Communication among oral bacteria. *Microbiol. Mol. Biol. Rev.* *66*, 486-505, table.
- Kolenbrander,P.E. and London,J. (1993). Adhere today, here tomorrow: oral bacterial adherence. *J. Bacteriol.* *175*, 3247-3252.
- Kurokawa,K. *et al.* (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* *14*, 169-181.
- LeClair,E.E., Nomellini,V., Bahena,M., Singleton,V., Bingle,L., Craven,C.J., and Bingle,C.D. (2004). Cloning and expression of a mouse member of the PLUNC protein family exclusively expressed in tongue epithelium. *Genomics* *83*, 658-666.
- Lee,S.W., Won,K., Lim,H.K., Kim,J.C., Choi,G.J., and Cho,K.Y. (2004). Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl. Microbiol. Biotechnol.* *65*, 720-726.
- Lepp,P.W., Brinig,M.M., Ouverney,C.C., Palm,K., Armitage,G.C., and Relman,D.A. (2004). Methanogenic Archaea and human periodontal disease. *Proc. Natl. Acad. Sci. U. S. A* *101*, 6176-6181.
- Ley,R.E. *et al.* (2008). Evolution of mammals and their gut microbes. *Science* *320*, 1647-1651.
- Liesack,W., Sela,S., Bercovier,H., Pitulle,C., and Stackebrandt,E. (1991). Complete nucleotide sequence of the *Mycobacterium leprae* 23 S and 5 S rRNA genes plus flanking regions and their potential in designing diagnostic oligonucleotide probes. *FEBS Lett.* *281*, 114-118.

- Liles, M.R., Manske, B.F., Bintrim, S.B., Handelsman, J., and Goodman, R.M. (2003). A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl. Environ. Microbiol.* *69*, 2684-2691.
- Liljemark, W.F., Bloomquist, C.G., and Ofstehage, J.C. (1979). Aggregation and adherence of *Streptococcus sanguis*: role of human salivary immunoglobulin A. *Infect. Immun.* *26*, 1104-1110.
- Lodes, M.J., Cong, Y., Elson, C.O., Mohamath, R., Landers, C.J., Targan, S.R., Fort, M., and Hershberg, R.M. (2004). Bacterial flagellin is a dominant antigen in Crohn disease. *J. Clin. Invest.* *113*, 1296-1306.
- Lowy, F.D. (1998). *Staphylococcus aureus* infections. *N. Engl. J. Med.* *339*, 520-532.
- Lu, Y., Iyoda, S., Satou, H., Satou, H., Itoh, K., Saitoh, T., and Watanabe, H. (2006). A new immunoglobulin-binding protein, EibG, is responsible for the chain-like adhesion phenotype of locus of enterocyte effacement-negative, shiga toxin-producing *Escherichia coli*. *Infect. Immun.* *74*, 5747-5755.
- Madigan, M.T., Martinko, J.M., and Parker, J. (2003). *Human-Microbe Interactions*. In: *Brock Biology of Microorganisms* New Jersey: Prentice Hall, 727-754.
- Mager, D.L., Ximenez-Fyvie, L.A., Haffajee, A.D., and Socransky, S.S. (2003). Distribution of selected bacterial species on intraoral surfaces. *J. Clin. Periodontol.* *30*, 644-654.
- Malik, P., Terry, T.D., Gowda, L.R., Langara, A., Petukhov, S.A., Symmons, M.F., Welsh, L.C., Marvin, D.A., and Perham, R.N. (1996). Role of capsid structure and membrane protein processing in determining the size and copy number of peptides displayed on the major coat protein of filamentous bacteriophage. *J. Mol. Biol.* *260*, 9-21.
- Marcotte, H. and Lavoie, M.C. (1998). Oral microbial ecology and the role of salivary immunoglobulin A. *Microbiol. Mol. Biol. Rev.* *62*, 71-109.
- Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* *24*, 133-141.
- Margulies, M. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* *437*, 376-380.
- Marsh, P.D. (2005). Dental plaque: biological significance of a biofilm and community life-style. *J. Clin. Periodontol.* *32 Suppl 6*, 7-15.
- Martinez, A., Kolvek, S.J., Yip, C.L., Hopke, J., Brown, K.A., MacNeil, I.A., and Osburne, M.S. (2004). 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.
- Marvin, D.A., Welsh, L.C., Symmons, M.F., Scott, W.R., and Straus, S.K. (2006). Molecular structure of fd (f1, M13) filamentous bacteriophage refined with respect to X-ray fibre diffraction and solid-state NMR data supports specific models of phage assembly at the bacterial membrane. *J. Mol. Biol.* *355*, 294-309.
- Matsuki, T., Watanabe, K., Fujimoto, J., Miyamoto, Y., Takada, T., Matsumoto, K., Oyaizu, H., and Tanaka, R. (2002). Development of 16S rRNA-gene-targeted group-specific primers for the detection and identification of predominant bacteria in human feces. *Appl. Environ. Microbiol.* *68*, 5445-5451.
- Matthews, L.J., Davis, R., and Smith, G.P. (2002). Immunogenically fit subunit vaccine components via epitope discovery from natural peptide libraries. *J. Immunol.* *169*, 837-846.
- McCafferty, J., Griffiths, A.D., Winter, G., and Chiswell, D.J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* *348*, 552-554.
- Meenan, N.A., Visai, L., Valtulina, V., Schwarz-Linek, U., Norris, N.C., Gurusiddappa, S., Hook, M., Speziale, P., and Potts, J.R. (2007). The tandem beta-zipper model defines high

- affinity fibronectin-binding repeats within *Staphylococcus aureus* FnBPA. *J. Biol. Chem.* 282, 25893-25902.
- Miedzybrodzki,R., Fortuna,W., Weber-Dabrowska,B., and Gorski,A. (2007). Phage therapy of staphylococcal infections (including MRSA) may be less expensive than antibiotic treatment. *Postepy Hig. Med. Dosw. (Online. )* 61, 461-465.
- Mitchell,G., Lamontagne,C.A., Brouillette,E., Grondin,G., Talbot,B.G., Grandbois,M., and Malouin,F. (2008). *Staphylococcus aureus* SigB activity promotes a strong fibronectin-bacterium interaction which may sustain host tissue colonization by small-colony variants isolated from cystic fibrosis patients. *Mol. Microbiol.* 70, 1540-1555.
- Mullen,L.M., Nair,S.P., Ward,J.M., Rycroft,A.N., and Henderson,B. (2006). Phage display in the study of infectious diseases. *Trends Microbiol.* 14, 141-147.
- Mullen,L.M., Nair,S.P., Ward,J.M., Rycroft,A.N., Williams,R.J., and Henderson,B. (2007). Comparative functional genomic analysis of Pasteurellaceae adhesins using phage display. *Vet. Microbiol.* 122, 123-134.
- Mullen,L.M., Nair,S.P., Ward,J.M., Rycroft,A.N., Williams,R.J., Robertson,G., Mordan,N.J., and Henderson,B. (2008). Novel adhesin from *Pasteurella multocida* that binds to the integrin-binding fibronectin FnIII9-10 repeats. *Infect. Immun.* 76, 1093-1104.
- Munson,M.A., Pitt-Ford,T., Chong,B., Weightman,A., and Wade,W.G. (2002). Molecular and cultural analysis of the microflora associated with endodontic infections. *J. Dent. Res.* 81, 761-766.
- Nagashima,K., Hisada,T., Sato,M., and Mochizuki,J. (2003). Application of new primer-enzyme combinations to terminal restriction fragment length polymorphism profiling of bacterial populations in human feces. *Appl. Environ. Microbiol.* 69, 1251-1262.
- Nakamura,M., Tsumoto,K., Kumagai,I., and Ishimura,K. (2003). A morphologic study of filamentous phage infection of *Escherichia coli* using biotinylated phages. *FEBS Lett.* 536, 167-172.
- Navarre,W.W. and Schneewind,O. (1999). Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol. Mol. Biol. Rev.* 63, 174-229.
- Nesbo,C.L., Boucher,Y., Dlutek,M., and Doolittle,W.F. (2005). Lateral gene transfer and phylogenetic assignment of environmental fosmid clones. *Environ. Microbiol.* 7, 2011-2026.
- Nichols,D. (2007). Cultivation gives context to the microbial ecologist. *FEMS Microbiol. Ecol.* 60, 351-357.
- Nikaido H *et al.* (1996). Outer Membrane. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. Umbarger HE Washington DC: ASM Press, 29-47.
- Ofek,I., Hasty,D., and Doyle,R. (2003a). Adhesins As Cell Surface Structures. In: *Bacterial Adhesion to Animal Cells and Tissues* Washington: ASM Press, 64-85.
- Ofek,I., Hasty,D., and Doyle,R. (2003b). Basic Concepts in Bacterial Adhesion. In: *Bacterial Adhesion to Animal Cells and Tissues* Washington: ASM Press, 1-18.
- Ofek,I., Hasty,D., and Doyle,R. (2003c). Target Tissues for Bacterial Adhesion. In: *Bacterial Adhesion to Animal Cells and Tissues* Washington: ASM Press, 43-62.
- Pace,N.R. (1997). A molecular view of microbial diversity and the biosphere. *Science* 276, 734-740.
- Palmer,R.J., Jr., Diaz,P.I., and Kolenbrander,P.E. (2006). Rapid succession within the *Veillonella* population of a developing human oral biofilm in situ. *J. Bacteriol.* 188, 4117-4124.
- Pankov,R. and Yamada,K.M. (2002). Fibronectin at a glance. *J. Cell Sci.* 115, 3861-3863.

- Paster,B.J., Boches,S.K., Galvin,J.L., Ericson,R.E., Lau,C.N., Levanos,V.A., Sahasrabudhe,A., and Dewhirst,F.E. (2001). Bacterial diversity in human subgingival plaque. *J. Bacteriol.* *183*, 3770-3783.
- Paster,B.J., Olsen,I., Aas,J.A., and Dewhirst,F.E. (2006). The breadth of bacterial diversity in the human periodontal pocket and other oral sites. *Periodontol.* *2000.* *42*, 80-87.
- Patti,J.M., Allen,B.L., McGavin,M.J., and Hook,M. (1994). MSCRAMM-mediated adherence of microorganisms to host tissues. *Annu. Rev. Microbiol.* *48*, 585-617.
- Peacock,S.J., Foster,T.J., Cameron,B.J., and Berendt,A.R. (1999). Bacterial fibronectin-binding proteins and endothelial cell surface fibronectin mediate adherence of *Staphylococcus aureus* to resting human endothelial cells. *Microbiology* *145 ( Pt 12)*, 3477-3486.
- Peterson,C. (2007). Denaturing gradient gel electrophoresis (DGGE). *J. Vis. Exp.* 164.
- Peterson,D.A., Frank,D.N., Pace,N.R., and Gordon,J.I. (2008). Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. *Cell Host. Microbe* *3*, 417-427.
- Pickard,K.M., Bremner,A.R., Gordon,J.N., and MacDonald,T.T. (2004). Microbial-gut interactions in health and disease. Immune responses. *Best. Pract. Res. Clin. Gastroenterol.* *18*, 271-285.
- Polz,M.F. and Cavanaugh,C.M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* *64*, 3724-3730.
- Pratten,J., Wilson,M., and Spratt,D.A. (2003). Characterization of in vitro oral bacterial biofilms by traditional and molecular methods. *Oral Microbiol. Immunol.* *18*, 45-49.
- Prinz,D.M., Smithson,S.L., and Westerink,M.A. (2004). Two different methods result in the selection of peptides that induce a protective antibody response to *Neisseria meningitidis* serogroup C. *J. Immunol. Methods* *285*, 1-14.
- Prosser,J.I. and Embley,T.M. (2002). Cultivation-based and molecular approaches to characterisation of terrestrial and aquatic nitrifiers. *Antonie Van Leeuwenhoek* *81*, 165-179.
- Quan,C.P., Berneman,A., Pires,R., Avrameas,S., and Bouvet,J.P. (1997). Natural polyreactive secretory immunoglobulin A autoantibodies as a possible barrier to infection in humans. *Infect. Immun.* *65*, 3997-4004.
- Radajewski,S., McDonald,I.R., and Murrell,J.C. (2003). Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. *Curr. Opin. Biotechnol.* *14*, 296-302.
- Rappe,M.S. and Giovannoni,S.J. (2003). The uncultured microbial majority. *Annu. Rev. Microbiol.* *57*, 369-394.
- Rhee,J.K., Ahn,D.G., Kim,Y.G., and Oh,J.W. (2005). New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. *Appl. Environ. Microbiol.* *71*, 817-825.
- Rickard,A.H., Gilbert,P., High,N.J., Kolenbrander,P.E., and Handley,P.S. (2003). Bacterial coaggregation: an integral process in the development of multi-species biofilms. *Trends Microbiol.* *11*, 94-100.
- Riechmann,L. and Holliger,P. (1997). The C-terminal domain of TolA is the coreceptor for filamentous phage infection of *E. coli*. *Cell* *90*, 351-360.
- Riggio,M.P., Lennon,A., Rolph,H.J., Hodge,P.J., Donaldson,A., Maxwell,A.J., and Bagg,J. (2008). Molecular identification of bacteria on the tongue dorsum of subjects with and without halitosis. *Oral Dis.* *14*, 251-258.
- Ritz,K. (2007). The Plate Debate: Cultivable communities have no utility in contemporary environmental microbial ecology. *FEMS Microbiol. Ecol.* *60*, 358-362.
- Roldan,S., Herrera,D., and Sanz,M. (2003). Biofilms and the tongue: therapeutical approaches for the control of halitosis. *Clin. Oral Investig.* *7*, 189-197.

- Rusch,D.B. *et al.* (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS. Biol.* 5, e77.
- Ruhl,S., Sandberg,A.L., Cole,M.F., and Cisar,J.O. (1996). Recognition of immunoglobulin A1 by oral actinomyces and streptococcal lectins. *Infect. Immun.* 64, 5421-5424.
- Russel,M. (2007). Introduction to Phage Display and Phage Biology. In: *Phage Display: A Practical Approach* Oxford University Press, 1-26.
- Russel,M. (1995). Moving through the membrane with filamentous phages. *Trends Microbiol.* 3, 223-228.
- Russel,M., Kidd,S., and Kelley,M.R. (1986). An improved filamentous helper phage for generating single-stranded plasmid DNA. *Gene* 45, 333-338.
- Russell,M.W. and Mestecky,J. (2002). Humoral immune responses to microbial infections in the genital tract. *Microbes. Infect.* 4, 667-677.
- Saraste,M., Sibbald,P.R., and Wittinghofer,A. (1990). The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* 15, 430-434.
- Scannapieco,F.A. (1994). Saliva-bacterium interactions in oral microbial ecology. *Crit Rev. Oral Biol. Med.* 5, 203-248.
- Scannapieco,F.A., Bergey,E.J., Reddy,M.S., and Levine,M.J. (1989). Characterization of salivary alpha-amylase binding to *Streptococcus sanguis*. *Infect. Immun.* 57, 2853-2863.
- Schennings,T., Heimdahl,A., Coster,K., and Flock,J.I. (1993). Immunization with fibronectin binding protein from *Staphylococcus aureus* protects against experimental endocarditis in rats. *Microb. Pathog.* 15, 227-236.
- Schmeisser,C., Steele,H., and Streit,W.R. (2007). Metagenomics, biotechnology with non-culturable microbes. *Appl. Microbiol. Biotechnol.* 75, 955-962.
- Schmeisser,C. *et al.* (2003). Metagenome survey of biofilms in drinking-water networks. *Appl. Environ. Microbiol.* 69, 7298-7309.
- Schwarz-Linek,U., Hook,M., and Potts,J.R. (2004). The molecular basis of fibronectin-mediated bacterial adherence to host cells. *Mol. Microbiol.* 52, 631-641.
- Schwarz-Linek,U., Hook,M., and Potts,J.R. (2006). Fibronectin-binding proteins of gram-positive cocci. *Microbes. Infect.* 8, 2291-2298.
- Schwarz-Linek,U. *et al.* (2003). Pathogenic bacteria attach to human fibronectin through a tandem beta-zipper. *Nature* 423, 177-181.
- Shaw,G.C., Wu,M.Y., Lee,T.R., and Hsu,C.W. (2005). The influence of nucleotide sequences at and near ribosome-binding site on translational efficiency of the *Bacillus subtilis* rho gene. *Biochim. Biophys. Acta* 1729, 10-13.
- Sidhu,S.S. (2001). Engineering M13 for phage display. *Biomol. Eng* 18, 57-63.
- Signas,C., Raucci,G., Jonsson,K., Lindgren,P.E., Anantharamaiah,G.M., Hook,M., and Lindberg,M. (1989). Nucleotide sequence of the gene for a fibronectin-binding protein from *Staphylococcus aureus*: use of this peptide sequence in the synthesis of biologically active peptides. *Proc. Natl. Acad. Sci. U. S. A* 86, 699-703.
- Sirard,J.C., Bayardo,M., and Didierlaurent,A. (2006). Pathogen-specific TLR signaling in mucosa: mutual contribution of microbial TLR agonists and virulence factors. *Eur. J. Immunol.* 36, 260-263.
- Skerker,J.M. and Berg,H.C. (2001). Direct observation of extension and retraction of type IV pili. *Proc. Natl. Acad. Sci. U. S. A* 98, 6901-6904.
- Slotweg,E.J., Keller,H.J., Hink,M.A., Borst,J.W., Bakker,J., and Schots,A. (2006). Fluorescent T7 display phages obtained by translational frameshift. *Nucleic Acids Res.* 34, e137.



- Smith,G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228, 1315-1317.
- Smith,G.P. and Petrenko,V.A. (1997). Phage Display. *Chem. Rev.* 97, 391-410.
- Smith,G.P. and Scott,J.K. (1993). Libraries of peptides and proteins displayed on filamentous phage. *Methods Enzymol.* 217, 228-257.
- Soto,G.E. and Hultgren,S.J. (1999). Bacterial adhesins: common themes and variations in architecture and assembly. *J. Bacteriol.* 181, 1059-1071.
- Spencer,P., Greenman,J., McKenzie,C., Gafan,G., Spratt,D., and Flanagan,A. (2007). In vitro biofilm model for studying tongue flora and malodour. *J. Appl. Microbiol.* 103, 985-992.
- Stadler,F. and Hales,D. (2002). Highly-resolving two-dimensional electrophoresis for the study of insect proteins. *Proteomics.* 2, 1347-1353.
- Steele,H.L. and Streit,W.R. (2005). Metagenomics: advances in ecology and biotechnology. *FEMS Microbiol. Lett.* 247, 105-111.
- Stevenson,B.S., Eichorst,S.A., Wertz,J.T., Schmidt,T.M., and Breznak,J.A. (2004). New strategies for cultivation and detection of previously uncultured microbes. *Appl. Environ. Microbiol.* 70, 4748-4755.
- Streit,W.R. and Schmitz,R.A. (2004). Metagenomics--the key to the uncultured microbes. *Curr. Opin. Microbiol.* 7, 492-498.
- Strohl,W.R. (1992). Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res.* 20, 961-974.
- Suzuki,M.T. and Giovannoni,S.J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* 62, 625-630.
- Suzuki,K. and Fagarasan,S. (2008). How host-bacterial interactions lead to IgA synthesis in the gut. *Trends Immunol.*
- Torsvik,V., Goksoyr,J., and Daae,F.L. (1990a). High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* 56, 782-787.
- Torsvik,V., Salte,K., Sorheim,R., and Goksoyr,J. (1990b). Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. *Appl. Environ. Microbiol.* 56, 776-781.
- Tratmont,E.C., J.Ciak, J.Boslego, D.G.McChesney, C.C.Brington, and W.Zollinger (1980). Antigenic specificity of antibodies in vaginal secretions during infection with *Neisseria gonorrhoeae*. *J. Infect. Dis.* 142, 23-31.
- Trepel,M., Arap,W., and Pasqualini,R. (2002). In vivo phage display and vascular heterogeneity: implications for targeted medicine. *Curr. Opin. Chem. Biol.* 6, 399-404.
- Tringe,S.G. *et al.* (2005). Comparative metagenomics of microbial communities. *Science* 308, 554-557.
- Turnbaugh,P.J., Ley,R.E., Mahowald,M.A., Magrini,V., Mardis,E.R., and Gordon,J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031.
- Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S., and Banfield,J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37-43.
- Venter,J.C. *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66-74.
- Vitorino,R., de Morais,G.S., Ferreira,R., Lobo,M.J., Duarte,J., Ferrer-Correia,A.J., Tomer,K.B., Domingues,P.M., and Amado,F.M. (2006). Two-dimensional electrophoresis study of in vitro pellicle formation and dental caries susceptibility. *Eur. J. Oral Sci.* 114, 147-153.

- Voget,S., Leggewie,C., Uesbeck,A., Raasch,C., Jaeger,K.E., and Streit,W.R. (2003). Prospecting for novel biocatalysts in a soil metagenome. *Appl. Environ. Microbiol.* *69*, 6235-6242.
- Vudhichamnong,K., Walker,D.M., and Ryley,H.C. (1982). The effect of secretory immunoglobulin A on the in-vitro adherence of the yeast *Candida albicans* to human oral epithelial cells. *Arch. Oral Biol.* *27*, 617-621.
- Wade,W. (2002). Unculturable bacteria--the uncharacterized organisms that cause oral infections. *J. R. Soc. Med.* *95*, 81-83.
- Wade,W. Unculturable Oral Bacteria. Plenary Session. 158th meeting of the Society of General Microbiology, 1-4-2006. Ref Type: Conference Proceeding
- Wagenaar,J.A., Van Bergen,M.A., Mueller,M.A., Wassenaar,T.M., and Carlton,R.M. (2005). Phage therapy reduces *Campylobacter jejuni* colonization in broilers. *Vet. Microbiol.* *109*, 275-283.
- Walter,J. (2008). The ecological role of lactobacilli in the gastrointestinal tract: Implications for fundamental and biomedical research. *Appl. Environ. Microbiol.*
- Wan,C., Fiebig,T., Schiemann,O., Barton,J.K., and Zewail,A.H. (2000). Femtosecond direct observation of charge transfer between bases in DNA. *Proc. Natl. Acad. Sci. U. S. A* *97*, 14052-14055.
- Warnecke,F. *et al.* (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* *450*, 560-565.
- Weber-Dabrowska,B., Mulczyk,M., and Gorski,A. (2000). Bacteriophage therapy of bacterial infections: an update of our institute's experience. *Arch. Immunol. Ther. Exp. (Warsz.)* *48*, 547-551.
- Weiss,E.I., Shanitzki,B., Dotan,M., Ganeshkumar,N., Kolenbrander,P.E., and Metzger,Z. (2000). Attachment of *Fusobacterium nucleatum* PK1594 to mammalian cells and its coaggregation with periodontopathogenic bacteria are mediated by the same galactose-binding adhesin. *Oral Microbiol. Immunol.* *15*, 371-377.
- Wexler,M., Bond,P.L., Richardson,D.J., and Johnston,A.W. (2005). A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environ. Microbiol.* *7*, 1917-1926.
- Wijffels,R.H. (2008). Potential of sponges and microalgae for marine biotechnology. *Trends Biotechnol.* *26*, 26-31.
- Williams,R.C. and Gibbons,R.J. (1972). Inhibition of bacterial adherence by secretory immunoglobulin A: a mechanism of antigen disposal. *Science* *177*, 697-699.
- Williams,R.J., Henderson,B., Sharp,L.J., and Nair,S.P. (2002). Identification of a fibronectin-binding protein from *Staphylococcus epidermidis*. *Infect. Immun.* *70*, 6805-6810.
- Wilson,M. (2005a). The Oral Cavity and Its Indigenous Microbiota. In: *Microbial Inhabitants of Humans* Cambridge University Press, 318-374.
- Wilson,M. (2005b). Distribution and Nature of the Indigenous Microbiota. In: *Microbial Inhabitants of Humans* Cambridge University Press, 2-11.
- Winnacker EL (1987). Expression Vectors in Prokaryotes. In: *From Genes to Clones: Introduction to Gene Technology*, Neustadt: VCH, 239-313.
- Wizemann,T.M., Adamou,J.E., and Langermann,S. (1999). Adhesins as targets for vaccine development. *Emerg. Infect. Dis.* *5*, 395-403.
- Woese,C.R. (1987). Bacterial evolution. *Microbiol. Rev.* *51*, 221-271.
- Woiwode,T.F., Haggerty,J.E., Katz,R., Gallop,M.A., Barrett,R.W., Dower,W.J., and Cwirla,S.E. (2003). Synthetic compound libraries displayed on the surface of encoded bacteriophage. *Chem. Biol.* *10*, 847-858.

- Woof,J.M. and Mestecky,J. (2005). Mucosal immunoglobulins. *Immunol. Rev.* 206, 64-82.
- Wu,L., Thompson,D.K., Li,G., Hurt,R.A., Tiedje,J.M., and Zhou,J. (2001). Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* 67, 5780-5790.
- Yassin,A.F., Brzezinka,H., Molitor,E., and Schaal,K.P. (1996). Rapid chemotaxonomic diagnosis of human tuberculosis. *Zentralbl. Bakteriol.* 284, 466-473.
- Yoneyama,H. and Katsumata,R. (2006). Antibiotic resistance in bacteria and its future for novel antibiotic development. *Biosci. Biotechnol. Biochem.* 70, 1060-1075.
- Yooseph,S. *et al.* (2007). The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS. Biol.* 5, e16.
- Zdobnov,E.M. and Apweiler,R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 17, 847-848.

---

## **Appendices**

---

## Appendix 1 Sample Sheet for Volunteers

### Tongue Metagenomic Study

Dear ..... You are number .....

**Included:**        **1 x toothbrush**  
                      **1 x 10ml PBS (sterile)**  
                      **1 x pack sterile tissues**  
                      **1 x instruction sheet**

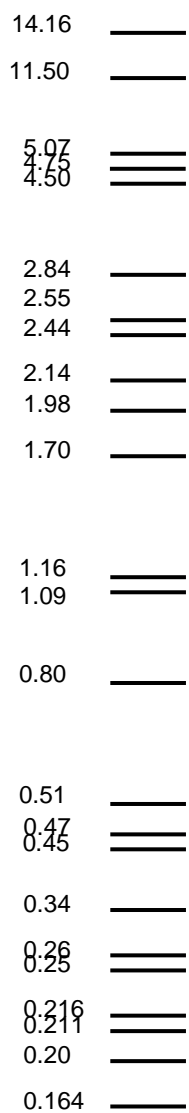
### Instructions for sampling

1. Swab tongue lightly with sterile tissue, **for 5 seconds maximum**. Aim to remove salivary contamination from the tongue surface by blotting lightly, not rubbing.
2. Dip toothbrush in PBS and brush tongue surface as vigorously as is comfortable for 15 seconds. Concentrate on the middle of the tongue, where the bacterial mat is thickest.
3. Use PBS to shake off some of the bacteria and repeat the brushing step a further 3 times (four x 15 second repeats altogether).
4. After last brush, shake the toothbrush in the PBS to dislodge as many bacteria as possible.
5. After sampling, put the toothbrush back in its plastic wrapper and store at -20°C.
6. Try to avoid the toothbrush coming into contact with any other area of your mouth at any time during the sampling.
7. Try not to close your mouth between repeats, as doing so will introduce fresh saliva to the tongue surface.
8. Sample collection should be carried out every Wednesday and Friday for 3 or 4 weeks.
9. Keep your own toothbrush (put your name on it!)
10. Date samples after collection and leave in Sam's -20°C freezer.

### Points to note:

- The samples you provide will be treated in such a way that will attempt to destroy all mammalian cells and DNA within the sample. The bacterial information that remains will be pooled and treated as a whole metagenomic library from human tongue material, so no individual inferences will be able to be drawn from resulting data.
- It is preferable to leave at least 2 hours between eating and collecting your sample. This is due to removal of the bacteria colonising the tongue surface by the movement of food and saliva around the mouth.
- If you routinely brush the tongue dorsum as part of your oral hygiene regime, this will significantly reduce the volume of bacteria on the tongue surface. On sample collection days, please try to avoid brushing your tongue in the morning, if you forget, just leave sample collection to as late in the day as possible.

## Appendix 2 Restriction Map of $\lambda$ Pst Ladder



	Protein BLAST	tblastx	Tag	ORF size (aa)	Amino Acid Sequence
F1	No similarity	Bacteroides thetaiotaomicron valyl-tRNA synthetase Expect = 9e-49, Id = 84/152 (55%)		157aa	YHFNSLQVSHFNDYTVVSLVVVLTLDYVQVELNLWSNVLLIAYGSHTADN RLNLLNSLDEL*FLLAWCIKFEFVTHDTLVILALIDVLPQLLSNERHEWMQ HLQE*IEEFKGC FV GELVDWFAIFWLNHLQVPA*EFVPEEAVNSHQSF*NT IGVQVD
F2	No similarity	Uncultured bacterium genomic sequence Expect = 0.058, Id = 14/26 (53%)	Cmyc	48aa	FGTSEKGVHNYADVKASACTLSMSTSVSKGFQ*SKSERPSAPGGVQVD
F4	CONSERVED DOMAIN Arthrobacter aureescens PtrB putative protease II Expect= 2e-19	Arthrobacter aureescens TC1 putative protease II oligopeptidase family protein Expect = 8e-24		186aa	DRRPHVRHAVGRRRRPGRRRKLTRPAGAPAAI**PASLFDTLELGVPTAP AHPHRHPAAQERGRRRHTPFTNP*LNERGGSMAEQNLTPPVPKKVEHRR EHHGDVFDHYEWLDRDKESEEVLNLYLKAEAETEAVTADQQPLRESIFNE IKGQRVGNREEQPLRDDHGTARRDVQVVAGVQGVQVD
F5	Phenylalanyl-tRNA synthetase beta subunit Magnetospirillum Expect=0.58	Listeria welshimeri serovar cysS (cysteinyl-tRNA synthetase) Expect = 6e-10, Id = 22/30 (73%)	polyH	102aa	PASTSASPPQPNAPRSCSPPTVTPPARRAPWLRRASRLIQRQLDYLHEWS ASIHQKKLKEIRLSEETYDSNLQHLNKTKRSI*TD*RRKSEHVCVRGVQVD
F6	No similarity	Mycobacterium smegmatis alcohol dehydrogenase, Expect = 1e-07 Id = 25/53 (47%)	BOTH	88aa	RCWCSW*TLLFRVAMHQKMLLATKSPRAYGALEATSPQESTFPLRARRS DSRPRLPADTAPPARVGALTCARTCAYSERGWMEGVQVD
F9	Porphyromonas gingivalis glycogen synthase, putative Expect=4e-10	Porphyromonas gingivalis glycogen synthase Expect = 2e-12, Id = 31/44 (70%)	BOTH	63aa	RSIAGNDKELYTYMDAYDGDQMARELGVEAKHEVEKLAHKARTVMP AALPVIASAPAGVQVD
F10	Hypothetical protein Leishmania braziliensis Expect=3.5	One match only -Bos taurus Expect = 3.7, Id = 12/15 (80%),	Cmyc	23aa	SSTGTRSTTALRRSWKRSGVQVD
F11	Histidine Kinase Burkholderia xenovorans Expect=16	Dinoroseobacter shibae 60 kDa inner membrane insertion protein Expect = 6.9, Id = 12/16 (75%)		19aa	RGRTVRVPCPTEVAGVQVD
F12	No insert				
F13	pyruvate dehydrogenase complex Listeria monocytogenes (id 9/11 81%), Expect=28	No similarity		16aa	ARTALVASDIPGVQVD
F14	No similarity	Burkholderia mallei formate		41aa	SS*GVSTPER*SCDLR*PCTPGTEKQEGDVTIRVLVGVQVD

IgA Antibody Screening Results

		dehydrogenase accessory protein Expect = 0.21, Id = 13/22 (59%)			
<b>F15</b>	No similarity	Human clone (e=0.48) from chr 6. id =8/15 (51%)	Cmyc	100aa	PCCSHCFLEANFLLWACSLGLA*ASSLF*PILLVDFGPLSWWLLGPMYLTG WL*KWAVLTMYSLL*RSNGIYKVMKYYYVLVCIQLFLFSWLSSVGVQVD
<b>F16</b>	hypothetical protein A. odontolyticus Expect=0.001	Acidovorax sp. serine O-acetyltransferase. Expect = 5.0 Id = 12/18 (66%)	Cmyc	21aa	AFETDREPGEDLRSIEGVQVD
<b>F17</b>	No similarity	Streptococcus sanguinis DNA repair protein radC. Expect = 0.003 Id = 18/30 (60%)	Cmyc	35aa	NFWLLMICSELINSSL*ILRPNSITARSSMGVQVD
<b>F18</b>	No similarity	2 parts showing homology to same sequence = Thermobifida fusca regulatory protein, MerR Expect = 0.004, Id = 15/27 (55%)		55aa	LEKALELEDASNPPEVNAFLLAHPLDVEQALHVTHILVATPPLLTMRNPQG VHVD
<b>F19</b>	No similarity	Veillonella dispar RNA polymerase B subunit (rpoB) gene Expect = 8e-150 Id = 289/293 (98%)	BOTH	443aa	P***E*CTSRTSKPARSRDKPPGPRADKRRLL*VNSANGFVWSMNWDNWE PKNSLIAATTGRMLINA*GVIDSIS*IVIRSRTTRSIRVKPIRNWFCNNSPTQR RRRLPK*SISSV*PAPSIRLSK*EMLAISSRVTVR*SNGKLQLLQITLDSTPS LFTM*NSAKSPSNTWLFATISSCSSPTCTPASRITSPVSGSTIG*ANV*PNIR *RQPSFLFNL*RPTRAKS*RRASKNKLSNKLADSTVGGSPGRNFL*ISTNA SSAD*VVSFSNVRTMRSS*IGKPSSSVLPFQRGQGCFF*YPNLMHVPI*LV IYGYGQYEPLRFRSHRFQARSMHNG*ESLWNP**TVTHLIDFLYYSKTP*R AYHLGYDTHVLHH***CCYQSSCSR*ILLDL*CSVFLDHYLPTIYLHKW CCVCSITIFTFINIILRLT*GVQVD
<b>F20</b>	sodium dependent transporter, Expect=3.5	Human DNA sequence Expect = 4e-53 Id = 81/81 (100%)	Cmyc	113aa	SPTAAKIKPSAVDTGVAALTGATVAPTALFLFCLPPTSATTMCLWVPPQL ELAFPSTRAPFRDTSYSLQTPGLPLDQTTLRSRGLGSCSTSPKFNTFAPPC PPRSHLGVQVD
<b>F21</b>	no significant similarity	Burkholderia sp. histidine kinase, Expect = 0.068, Id = 18/31 (58%), Anaeromyxobacter transcriptional regulator, TetR family Expect = 0.13, Id = 19/38 (50%)		46aa	TGRRTRHIAAADQSCRASSFAVAAMIRSARAFEASRLAVIVGVQVD
<b>F22</b>	unnamed protein product homo sapiens Expect=3.3	Candidatus Desulforudis audaxviator, DNA polymerase III, Expect = 1e-17, Id = 33/60 (55%)		79aa	PHYEYI*SHSLRSKRASQILHLQWTLRCHY**NLV*RIPVAP*GRILFSYPSH DALNLLGLPRCWDYRLEPLRLGVQVD
<b>F23</b>	chromosome 14 ORF 93 homo	Streptococcus sanguinis DNA		142aa	RGS*TASRRERRRNPAAVPRRHRTALPGQPRQPGFCHHARLIFVFLVET

IgA Antibody Screening Results



	sapiens Expect=4e-06	recombination protein RmuC, Expect = 2e-18, Id = 44/45 (97%)			GFHHVGVQAPRSSRRRATPDPARSRGSAACVSTPPGRRSSERPRPESLQTYGILI*PIPTPC*CHLHSFGC*AF*SQQTRSSRGLIRYLGVQVD
<b>F25</b>	no significant similarity	Clostridium botulinum str. Loch Maree, PTS system, fructose family Expect = 3e-60 Id = 55/76 (72%)	Cmyc	247aa	PRAIRRGARGPSAAFSGSSETGADICEAKNPATNAPKNPDGVYANTSLFRLRILGSARIPPTIPTINPGLSAILRPMNPARIGNIILNEILPK*NNTSAYLL*L GSSGLNELIPQINDIAIRIPPATTTNGSILLTPSIRCL*V*RQMLSSPAESSASS APS*L*IGALPSKIV*ISSFGLLIPSATFVIIVGRPLKRATSTFLSAATHIPAALS ISSLVSLFSTPI*PFVSTLTSSPISLGVQVD
<b>F26</b>	polysaccharide biosynthesis protein CapD Expect=8e-06	Neisseria meningitidis pilin glycosylation protein, Expect = 2e-133 Id = 191/276 (69%) <u>Matches all over sequence</u>	Cmyc	305aa	LAHDPACDSRSPQTDPRHHLRCRSVRPPVVGSHQTG*RIFSHCLCR*Q PKNPTYRHL*PCRPQPQRNPADQPLRRPQNPAAGDSKIFYARRTQRHHP RSIQMRSPDHSGHERFSRRQNQRQLIEKNLCGRFARPRPCDTAP*IDECRH QRQSRDGDRRRLYRLRTLPSDSLPSDQTAIV*IVRICPVQHRQRIARNPS SARQPSRSRTAFGLGSKQRT*QHEDLSRRHRLSCCGLQTRPYGRVQHH RRHSKQVRHTLLRTGCRRRSTFVLISTDKAVRPTNTMGASKRMGVQ VD
<b>F27</b>	no similarity	Acidobacteria TPR repeat protein, Expect = 6.7, Id = 13/29 (44%),	Cmyc	50aa	LFEHLTDF*KCSVSTPPSSEPPPPNLA*EPKN*DIARKV*S*LSLGVQVD
<b>F28</b>	flavodoxin from Fusobacterium nucleatum Expect=2e-10	Nocardioides sp phospholipid/glycerol acyltransferase Expect = 2e-10, Id = 26/49 (53%)	Cmyc	163aa	PPPEGPGRAVNKAQLNRHVRLHRDAIEARLPKVARRARPLRRDAQREGV GFANAPCHLVHQARDVLGTLHNPYIALFATAGVPPQMEHAKQSLINAAA CLPEGVVPVDTFICQGGKVDPKVIEMMYKMFPGHSHGQSADRDARHKQ AAVHPNEDDFKAGVQVD
<b>F29</b>	no similarity	Clostridium beijerinckii IMP dehydrogenase Expect = 7e-06, Id = 21/28 (75%)	Cmyc	33aa	LSIDCEPNMKEIPPSLASATAIVSFDTGVQVD
<b>F30</b>	no similarity	Streptococcus thermophilus LMD-9, 7,8- dihydro-6-hydroxymethylpterin- pyrophosphokinase Expect = 2e-12, Id = 30/34 (88%)	Cmyc	51aa	SRGFSWQLSL*GKARFKKSA*FVSPQAAVS*YEDTALTSVSGKSFKGVQVD
<b>F31</b>	no similarity	Homo sapiens BAC clone RP13-744A23 from 7, Expect = 2.7, Id = 16/46 (34%)		104aa	PCSAFMISLASTCLNWLAKSLRTSPRASVPKRFKPFNCSTRSLSGTVSS ALASVFLALGSSS*VAGASTGTSLADATSSPCRLGISPSNVKPAFIKGVQVD
<b>F33</b>	CONSERVED DOMAIN	Campylobacter curvus		320aa	PSLMTVILFLGGFQLLTIGILGKYVVGKIFMETKKRP*V*IA*IRQVVESCPPE

IgA Antibody Screening Results

	DraG glycosyl transferase Streptococcus gordonii Expect=7e-11	phosphomethylpyrimidine kinase Expect = 2e-25 Id = 56/80 (70%)			KITRACCFIVNS*FC*RGRFANRVI*RLTDYNGFRL*SRHFIFYIA*L*AST P*NQNFKTCCAPLLRDCIKKALLPIR*PCLPAPFPFCSVYF*HCLQVYPHYS GFCRSGCLSAWR*MRWTVVCWHGSSATWVLLNSKSYQEAVESALSVKGS KNLPVVVSTLAAAYYGLSQIPKDWVNALAGKEEVREIAIEWQMRWLD*E RNLYRNQFFILYLTKSILLYKRRILKEARLSSRAFLLTNNAVNSLLIFYIGQ VYKV*DSGVQVD
<b>F35</b>	hypothetical protein Expect=4.5	No similarity		22aa	PALPPDISEPIRGKVNQVQVD
<b>F36</b>	no similarity	Clostridium perfringens Transketolase Expect = 1e-16, Id = 37/43 (86%)		48aa	CKHSSFPYLTFKFSITENGIYTIISIIDFTC*SHTTSCRNPLGVQVD
<b>F37</b>	1 match – hypothetical protein, Expect=7.5	Moorella thermoacetica UDP-N- acetylmuramoylalanyl-D-glutamate--2,6- diaminopimela... Expect = 1e-87 Id = 94/196 (47%) but several matches to diff parts of seq, some with ID of 68%		353aa	PSNVSPRNATNTYPD*IVRVSDISLNCKLTSPLHVPPVASIISCNVNSAILD LLKSLSSLFVSIKMNIFIL*NLIIFMTLARNNNNIACFSIMNCMIDCLSTIDN HLIRCTITKCFYTYLYLNNLYRIFSARIIGCNINSITIFCCNASHNWAFCSE TITTATKYDDNAIFFNSLCCF*NIL*TIRCMCIYNYSKILPFLD*FKSTWYRT KCFQSRINGFH*NTFL*TSPNGSQCIHIK*TRSIHRYRITFTIINSILAHLSRH MNIGCI*GTIS*CTIRNSWSCRFFNHTMTIRIVNIDYSAFTSFNLIIGCTNLFK HLHLACSSSRFCGNPNDLASNRKYGVQVD
<b>F38</b>	phosphate ABC transporter, ATPase subunit, Expect=0.6	Helicobacter pylori type I restriction enzyme M protein, Expect = 0.005 Id = 16/21 (76%)	Cmyc	26aa	PRVRPVRGLGVSAPVREAARGVQVD
<b>F40</b>	no similarity	Methylobacterium radiotolerans, RNA- directed DNA polymerase Expect = 1e-18, Id = 44/87 (50%)	BOTH	106aa	LFHAA*QASS*P*RSQIYNKERISRT*RRLTFTFSVTIKPVTSCRLPRRCKWV FSSFSRKP*RSTIAFTCCTSSPCGRRISVSELKVRSSA*RKVVIPCAAGVQVD
<b>F42</b>	No insert				
<b>F45</b>	tartrate dehydratase subunit alpha Wolinella succinogenes, Expect=2.9	Streptococcus pneumoniae strain g394 surface protein PspC (pspC) Expect = 9e-10, Id = 24/42 (57%)	Cmyc	92aa	RIRVEVPRPPNPHTGHHHTRLLQEIPRITNPQPTSNHRTIGHVHAKTVASH PDAKVATSTTSGCASLTTV*APRSAPL*SVPSGRLGVQVD
<b>F46</b>	BioW protein Bacillus amyloliquefaciens, Expect=0.023	B.sphaericus bioXWF operon genes, Expect = 5e-56 Id = 88/170 (51%)		245aa	RPLSLPHPHPPRPKKKESGRARGRVFFVRSEIDRESVIEYLEKAPVLVQRRN NV*VF*RTA*S*NQQS*FADLKRILPSRCSAREAR**RILDDGF**LFRTYL*S SCHRRGCKRGSTIWYWIRRVSTCVWNISIIYRA*KCISKI*EY*KSSRI*YRL YG*CRNFGYCR*EYYL*RRLKSC*YH*WLSFK*SLYKGV*S*RCRGIKV PTETSRSRYTKTHRY*WCL*HGWRVYCTIR*GVQVD
<b>F48</b>	ABC transporter, ATP- binding protein Neisseria	Neisseria meningitidis ABC transporter, ATP-binding protein	Cmyc	39aa	SIFRKMTVEQNIRAILEISMKDKSRIDAELEKLLGVQVD

IgA Antibody Screening Results

	meningitides Expect=5e-09	Expect = 1e-10 Id = 31/34 (91%)			
<b>F51</b>	cytochrome c biogenesis protein, transmembrane region, Expect=68	No similarity		17aa	ALQEHRVGGVVLGVQVD
<b>F54</b>	no similarity	Bifidobacterium adolescentis preprotein translocase SecY subunit Expect = 1e-04, Id = 15/18 (83%)		38aa	LLVIASKRGITTRGSWIIIEA*CTA*FPAQRCLRVQVD
<b>F57</b>	CHORISMATE BINDING isochorismate synthase Bacillus cereus Expect=2e-06	Bacillus cereus subsp. cytotoxis isochorismate synthase Expect = 5e-08 Id = 21/41 (51-59%)		45aa	LTCLDGARAIHPTPAICGAPTEKALDLIRQLESFERRYFGGVQVD
<b>F59</b>	NO INSERT		Cmyc		
<b>F60</b>	GTP-binding protein EngA Rhodospirillum, Expect=16	No similarity	cmyc	20aa	PGLHRLMDA*PHAKWGVQVD
<b>F61</b>	no similarity	Bifidobacterium longum conserved hypothetical protein with phosphoesterase domain Expect = 5e-05 , Id = 23/41 (56%)	Cmyc	46aa	TLGRESEHVLPRDRVVRAQVLDHAEGARLAVLPGSPEIIDEGVQVD
<b>F63</b>	No insert				
<b>F65</b>	CONSERVED DOMAIN murD, UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase Streptomyces coelicolor Expect=2e-23	Streptomyces coelicolor putative UDP-N-acetylmuramoylalanine-D-glutamate ligase Expect = 2e-27 Id = 50/93 (53-80%)	BOTH	102aa	ADHLNWHGSLEAYAADKAKVYANTKLAAIYDLASEAALKMVQEADVRE GCRAIGLSRAVPEISQFGIVDGAIVDRAVPLRNKNAQIVAELEDLAHLGV QVD
<b>F66</b>	hypothetical protein Actinomycetes odontolyticus, Expect=2e-27	Listeria monocytogenes CoA-dependent propionaldehyde dehydrogenase Expect = 2e-22 Id = 53/71 (74%)	Cmyc	250aa	SGMSEGEFAATSETARLQYANGFFSGGHDLIIGKRHYVDGATEAHQLAGS GTLTPEEHEQYTAGVLFVSSAGFVYAKLSTQFANRVVNIDAYASDEQNK AALRIKY*N*TLH*IKHQV*KTLLQSVKLVIMV*HYMNYQPMVSSVL*RQV RIQQKH*YVIPSVCWWPRYR*PGRHLHGQWMNPLRW*RLESFEAHHVRHF AGRRIYVRASQCANVGFVGLAVPSAALLVTRITRCLGAEAGSFSSGVQVD
<b>F68</b>	CONSERVED DOMAIN	Fusobacterium nucleatum Outer membrane	Cmyc	201aa	QYLISQGVSSNRIVANGFGSSNPIASNATQEGRQANRRVEVRILPAQ*SQ**I

IgA Antibody Screening Results

	OmpA Outer membrane protein Fusobacterium nucleatum, Expect=2e-15	protein Expect = 3e-17, Id = 41/48 (85%)			IFSKQLSNYLLPLIIVKKIINNIS*VIYSH**LN*KWLEQ*SRSSLYYYNISNF SIALFLNSSLNFG*ASCLRPAPPPARAPRSAPRSPPPCASPSSTCWRS*FRWR FGSSTAPSTRITTPRSWCPS*FPWRRPTLPRTPLAGVQVD
<b>F72</b>	no insert				
<b>F73</b>	No insert				
<b>F74</b>	no similarity	Renibacterium salmoninarum aspartyl-tRNA synthetase Expect = 0.009, Id = 17/17 (100%) <u>Only 6 hits altogether</u>		22aa	PGTRGRSCP*STR*GRSGVQVD
<b>F76</b>	no similarity	Bifidobacterium longum NusB antitermination protein Expect = 0.042, Id = 14/30 (46%)	BOTH	39aa	PRLRSRSFRSRLGEWRTTCAASTP*FR RTPACPGVQVD
<b>F80</b>	no similarity	Acidothermus cellulolyticus transcription elongation factor GreA Expect = 5e-22, Id = 46/66 (69%)	BOTH	97aa	PAKEGRQPFRVRRFHHQGRFRQPQRCAAFLRVPGYGSR*GTARCERHGSR RSRANRLGEPRLRGEQSQDDAPRRDP*AHLVRETHLVADHLGVQVD
<b>F81</b>	CONSERVED DOMAIN RnR_3 anaerobic ribonucleoside triphosphate reductase Listeria monocytogenes Expect=4e-21	Lactobacillus reuteri ribonucleoside-triphosphate reductase class III catalytic... Expect = 9e-18 Id = 38/73 (52%)	Cmyc	272aa	PFGKWHPFENQYR*LLAPLTILLNLDDDLSETNLATCQFLAIVANLLGLFG GDETKTKENANKDARLFSTFRDLEAGEVSRFYALQRLPERVSGAHVSGDIH FHDLDYTPVPGMFCMLVDLPFILSREDFPIGNTRVNRVRSVETATDLIPQI AAQVSVGHSFLGLTLHLLSLLCCA WGNLSLVYSPSPRREYLEPHSISAYPSS PHSHFINQGCWAFSHSSHPAREWGFTKLAKQRCYRCYELWYNFQDTRAKT IDATVDAIKAGVQVD
<b>F82</b>	no similarity	Chlamydomonas reinhardtii flagellar associated protein Expect = 1.4, Id = 14/22 (63%) <u>3 matches only</u>	Cmyc	29aa	SDRSWRAVAPA*ATARGKSPRRKGVQVD
<b>F83</b>	no similarity	Streptococcus gordonii aspartate aminotransferase Expect = 2e-31, Id = 58/64 (90%)	Cmyc	69aa	SDRLEIGTNSFPL*TKRP*IS*ARMRTSCSTAQFPMKSSSRE*IIPVGLDGEL RTNTLLLSVRGVQVD
<b>F85</b>	no similarity	Porphyromonas gingivalis NADH:ubiquinone oxidoreductase, Na translocating, A subunit Expect = 6e-07, Id = 22/27 (81%)	polyH	32aa	P*ASRRSLPC**APCG*SAYRPDLRGGVQVD
<b>F86</b>	no similarity	Human DNA Neisseria meningitidis oxidoreductase, short-chain dehydrogenase/reductase family Expect = 3e-26, Id = 42/57 (73%)	BOTH	273aa	WRRSSSSPRHPPCSPCSSPEPPDSPGPQRGPTLSSRPLGSGQLGGAATGAAC REQSQDGD RAGHPFWKCLHFQAGEGRQWVAGSAGPPWLLVILKRKPKRL LRRLPLLPHRPRASSSGSTPSATPMQG*RAQGR TACTSSPPR PASRVISLRTIS RHVRASSTRSSTTSRLRTSKPFSSDDCKIYIWPDADYLNQDLIFPNKDILCFS

IgA Antibody Screening Results

					NVSSFTGAGSSR*TFMPAKVWATGMISMELTFIRAGR**SHKAVSAISCGWINSTFS*TSGVQVD
<b>F88</b>	hypothetical protein Salmonella enterica Expect=5e-11	Expression vector pHA-SP Expect = 2e-55 Id = 87/101 (86%)		181aa	ANPVTSGCCQWR*VVSYRVGLKTIVTG*GAAVGLNGGFVHTAQLGANDL HRTEIPTA*AMRKRHPSRREKGGQVSRKRQPPPPHEGGSRGKLQNLNPI GTMQMWL*QGKKNSRHGGKSY*NMPAQRPFCSNVLGYVSSAYKSNITK ICGYH*KVESDSRLSVMLKYDLSL*GVQVD
<b>F91</b>	hypothetical protein Burkholderia, Expect=0.34	Actinoplanes sp. acarbose (acb) gene cluster, Expect = 2.7 Id = 12/17 (70%) <u>6 matches only</u>		25aa	LLRAAVEKRIRRRKEFVDREGVQVD
<b>F92</b>	no similarity	No similarity	BOTH	32aa	PGPFTTS*L*PSSRIPSPFLSCSLALAGVQVD
<b>F93</b>	no similarity	Rhodococcus sp. C4-dicarboxylate transporter Expect = 0.18, Id = 14/23 (60%)	polyH	66aa	RVRMGAPDESDAPIHPPQRGDYGINRETIRSSDQPSSL*PMSLGEIGRARFRL IVGITHARGVQVD
<b>F94</b>	ATP-dependent protease La 1 Bacillus thuringiensis Expect=1e-06	Desulfitobacterium hafniense ATP- dependent protease La Expect = 6e-37 Id = 66/108 (61%)	polyH	292aa	PRRCRRSAWR*AGTCVSYPLIHGRVVPYHRLGLSRL*RMRGTR*PGMGIC KSSMMPMRCSPSTNLSGSRP*ARMSPLTPTFDERLMISGTVITPFPFRQSA SRILRPFTCDGNIRPIGQRCINPNFDKA*KCINKTI*KNSIP*WCRVNFH*RCIR RNR**SITAKNRCSWFACYH*KGDETRYVRSSFHAGSN*VHRKS*IRAEHR* TYTQK*GRPRHPVEIVIRRNNSFLCRNEFLFII*IPGGSMNLLEIGIPTVPLRGM VYPNIVIHLDIGRDKSIKAVEGVQVD
<b>F96</b>	No insert				
<b>F97</b>	not bacterial	Homo sapiens chromosome 16, Expect = 3e-28 Id = 52/53 (98%)		180aa	PEFAFKNHSIAGFLWHHFTSEVLVCLLIHYLLRILFMPGVVAHACNPSTLE GHHCYCRCYCRRYCSRHCRCYRRCRCYSRRCHCWNRCRYCSRRCRYSRC RRSPNSRWARRRASRWG*ASRWARQRTAERRSSRSARAWARCCAPCG*T RPSTGSRRPAGSRSPASCSRSTRPGVQVD
<b>F99</b>	OmpA family protein Pseudomonas fluorescens Expect=11e	Arthrobacter sp. acetyl-coenzyme A synthase Expect = 4e-06, Id = 18/25 (72%)	BOTH	30aa	RNQPCTP*WRQ*ESNRRCGFFHTVRGVQVD

IgA Antibody Screening Results

	Protein BLAST	tblastx	Tag	ORF size (aa)	Amino Acid Sequence
1	No similarity	Human DNA sequence Expect=2e-14	polyH	28aa	PMCIRPQCTQNSAQTGRHPGTHTFPGGVQVD
2	CONSERVED DOMAIN FTCD_N Streptococcus gordonii glutamate formiminotransferase Expect = 4e-07, Id = 26/27 (96%)	Streptococcus gordonii glutamate formiminotransferase Expect = 3e-08 Id = 25/25 (100%)	Cmyc	26aa	ERINRELGIPIFLYEDAATRPERKNLGVQVD
3	putative p150 [Homo sapiens] Expect = 3e-16, Id = 42/61 (68%)	Human DNA sequence, Expect 2e-56, Id=100%	Cmyc		LHIKLNRN*LIKGSERSLQ*KLQNIGEINLTGYKI*KDIPCSWTERINIVKM VILRKAI*RFNVILIKIPMTFFAEIEKTILKFMWNHKRP*IAGGVQVD
4	No similarity	Thermobifida fusca cobalamin-5'-phosphate synthase Expect = 0.75 Id = 13/20 (65%)	Cmyc	27aa	LLGIILDERLDGDRQGVVALGDDLGQGVQVD
5	hypothetical protein [Bacteroides fragilis Expect = 7e-09, Id= 28/51 (54%)	Bacteroides fragilis conserved hypothetical protein Expect = 2e-10, Id = 28/50 (56%)	BOTH	52aa	LNLLELKAVAKEFGMPAFTGGQMAKWLYIQHVTTIDEMTNISKNNREK LKAGVQVD
6	No similarity	Burkholderia xenovorans Rhs family protein Expect = 3e-04, Id = 19/31 (61%)	BOTH	124aa	PQNRNNVAGCYADSLHRSIGFAV*TAGARSDRRVHHFSDDREKMIESP EK*KVCTCRS*AAAYIRRLGCYANLITFNF*NIDS**ILLTF*NFRLG*LFL*V EIFSY*ICCSKDLIHFEKQGVQVD
7	Methylocella silvestris outer membrane protein assembly complex, YaeT protein Expect = 28, Id = 12/19 (63%)	Human DNA sequence ch9 Expect = 1e-10 Id = 23/23 (100%)	Cmyc	24aa	HQAGLGTSDSFSISVAVTESTFKGVQVD
8	Aspergillus fumigatus DNA damage repair protein (Rad9) Expect = 0.032, Id = 26/67 (38%)	1 <sup>st</sup> half is bacterial Arthro bacter aureus phosphoribosylformylglycinamide synthase II Expect = 2e-15, Id = 46/54 (85%) 2 <sup>nd</sup> half Homo sapiens chromosome 17, Expect = 2e-48, Id= 85/86 (98%)	Cmyc	195aa	FKTRCIQCI*AI*GTDKGDECRTFKFTTSTFCSEFHQYDNGTHG**IRNNIRY NGHRRGRHRPRHHLHGRPPDRRHGPAALRRRRPPRHRCARRPRRRLRRR RLRQLPGPFSASAIRQLEVRAVFSSFFHLPFSCAK*CYEGG*MLKVYFFFL I*KRQLTVNSRQFSKHTLIRDSGYEEESVVGMRWMTLESEQLHLGVQVD
9	No insert				
10	Xanthomonas oryzae amidase/aminoacylase/peptidase family protein	Homo sapiens chromosome 5 Expect = 2e-90, Id = 142/142 (100%)	BOTH	248aa	LSRWHHHDQ*VFIFYTLLFLRSVCLMNVLSPELSYYNLYTCPSSVYRCVA *WDLRKAD*AEEMGSSL*LKRFRMRLLVMAAPALLSTLAAGAGSCNEW RIGNSISGLVRGESRLKKLWIEDETLFQNCPLGSEKGGQSSVNGAVAPLQ

### FN Antibody Screening Results

	Expect = 0.40, Id = 23/91 (25%)				IRCQLVIEIGNGVSESAENEYFSVTRVNLVGSGLLRDQCPQRTELGITIRG HCTGVTDQFVENRNVAYEVHGQSLAVHIAQVDANLRPDLEHLSISVIEG VQVD
12	Haemophilus influenzae PittGG methionyl-tRNA formyltransferase Expect = 4e-30, Id = 62/85 (72%)	Haemophilus influenzae SUN protein Expect = 3e-144, Id = 165/212 (77%)	polyH	330aa	PISFLQLTDEQNEQTLKVYAAAVLPHVDKPAAGTILSVDDKGIQVATKE GVLNLLQLQPAGKKPMSVQDFLNGRADWFQVGKVLG*WHFNAKKQK NRPHFRYALLLLK*FCRF*IKVSLYQCCFQKCNRR*NHRIYLYRKSPLV FFAYYL*V*KIL*KNY*ISH*RVKPASCTACYWWDCTNYFICVCPMLRLW MKW*MPQNH*NRIVFVVWLMVYCVASYGNKRIFLL*INIGKRFILNGL* INSKLLIRIGVKLLRRITKSHQCGCELTNNKIIRKLTRTLLEEQR*QRLNVK IHMLYV*LNAFCLETAEF*TRFGDSSGSQCSVVGQVD
13	No similarity	Clostridium tetani oxygen- independent coproporphyrinogen III oxidase Expect = 2e-06, Id = 26/41 (63%)		66aa	PCRASNAVY*S*L*CFISRF**LFLQFLVDSRKGSSVKTDLPKLIIF*VPYIV LNSRCNIFGC*MGVQVD
14	transcriptional activator Cryptococcus neoformans Expect = 4.9, Id = 10/13 (76%)	No similarity		19aa	PTARFGVRPASMSPSISEGVQVD
15	Pedobacter 3-ketoacyl-(acyl- carrier-protein) reductase Expect = 1.1, Id = 12/14 (85%)	Homo sapiens chromosome 13 Expect = 1e-07, Id = 21/21 (100%)	BOTH	22aa	LAPSLRFAGILKVLVLELPGVQVD
16	CONSERVED DOMAIN Yersinia mollaretii Pyrrolidone-carboxylate peptidase Expect = 9e-05, Id = 18/25 (72%)	Citrobacter koseri hypothetical protein Expect = 3e-06, Id = 18/25 (72%)	polyH	103aa	PLRAQFCYRASNFLPFEMCTFMMVRENDRFDQRELSSRLIV*RRCA*RA AIGEQ*AACGTCSVR*AACGACGVQLRCGGKMKKILVTGFEPFGGEKIN PAWEGVQVD
17	No similarity	Pseudomonas stutzeri membrane protein Expect = 0.001, Id = 17/36 (47%)	Cmyc	39aa	CW*TESIVVAFTRKSHQH*Q*NCTNDRNNTNKHKTARTGVQVD
18	No similarity	(4 matches) Nocardioides sp. glycosyl transferase, group 1 Expect = 1.0, Id = 14/18 (77%)	cmyc	38aa	PNAAPRRHPSSLGAWAAGVACPAARAPSAGQCADLAGVQVD
19	Streptococcus gordonii str. Challis Replication initiation and membrane attachment protein (DnaB) superfamily	Streptococcus gordonii Replication initiation and membrane attachment protein Expect = 3e-16, Id = 39/78 (50%)	BOTH	96aa	PT*SGSSSRSTTPRRWVASVDQGEKQYILAQILNHLNIGFPQLLLAFDRLI AMGLLDLYEEEVGITIQLHAPLASEEFFSNAVFKRLLKIGEKGVQVD

FN Antibody Screening Results

	Expect = 1e-14, Id= 40/82 (48%)				
20	No similarity	Corynebacterium glutamicum hypothetical protein Expect = 1e-04, Id = 18/31 (58%)	BOTH	34aa	PGFGRVDRGVERAIRTDEDVLADDHRSDVEDGQGVQVD
21	human		Cmyc	156aa	PPSGC*WSPSLALGVLASWPFILSHAALALATLVYLVYAVRVGVVYAFM LTSFFLVAVVWFQSPSRVVIAGIVPIVAIAEVFAYAEGEPLWSGCYLALVGI FVGMARWRMERSRERLRRQDAIQRAKAAERARMSTDLHDILGHSLIGI TMISELGVQVD
22	CONSERVED DOMAIN PRK 10263 No similarity	Arthrobacter aurescens 2-isopropylmalate synthase Expect = 2e-91, Id = 134/199 (67%)	Cmyc	271aa	HQCSDQHADWSHQYQPNQC*STQYGYTPNGSADQQQAGTHPRRCSTAS RWSSRRPQRQYEPEQSLRTQPEHDAANPDRDPASAGTPYRPTASSHE *SPHRYDPGREDQSAGTEPPSKPHPRSYQPQQRYAPEPQSPS**RPGENQR TR*TPGHHHTANARAPSAGCSTQQSYGYRQYQRNQSADQYPGCTCSC** E*A*SWNISLPCSIREVEN*GNSTRY*K*VYP*IIS**VN*AS*SSI*ILSSITRY YRGYFMYI*L*T*IL*GGVQVD
23	Lactobacillus plantarum amylopullulanase, collagen – lipoprotein repeat Expect = 9e-15, Id = 55/109 (50%)	Lactobacillus salivarius conserved hypothetical protein Expect = 1e-11, Id = 33/60 (55%)	Cmyc	316aa	FSNCSPi*SN*EIS*KTRKTN*KTCC*YSES*IQLHEEIRRSCVESW*IQLR ERQRW*RD*NRKQ*C*R*DQVLSS*IQTWSRRYLILSCRRSER*RGWS*V R*DGCYSWSCCDEGR*SINSYFTNA*RY*V*QQGNTANTANTANAANTA NTANTANAANTANTANTANTANTTTANTANTINTTNAANTANTANTA NTANTANTANTTTANTANTTTANTTTANTTTNAANTANTANTGNTANIRKT* RS*VAKYW*TI*IWSCCTWCCAWSRRSRFGCKTKKT*RLRLR*HQYFDV EITILALK*RGFARKSKGVQVD
24		human	BOTH	157aa	HFEPLSVEKTKRMIAFACLGLDSCDPLELELDAKLQFVDGMTTKEIQKTL VQTAIEKVISKKDDGFGNQVNKMNQDWQFVAARLFLFDLYKEAAITRR YKAFGYGNFPNLVHMLVEEKKYADFFVTEYTADELQELGDYIKPKRDY LFNYEGLKLLGVQVD
25	Leishmania braziliensis DNAJ- domain transmembrane-like protein Expect = 0.16, Id = 18/27 (66%)	No similarity	cmyc	25aa	L**AAAAPARGARTRTAGATSTAGGVQVD
26	CONSERVED DOMAIN COG O313 Actinomyces odontolyticus hypothetical protein (methyltransferase) Expect = 4e-21, Id = 51/53 (96%)	Caulobacter sp. TonB-dependent receptor Expect = 2e-16, Id = 24/31 (77%)	BOTH	183aa	AKGGARVVFVSDAGMPTVSDPGFRLARAAIEAGVPLSVLPGPSAPLVAL ALSGRLTFKAAKVQCYLVAQFAAHKYT*GIPLIPCCIVFIPTRVIHRICTN WCRRILFTKESPFYTN*ETWRAVVDFIGLTTTHIAYPNSSRATFFSRCFSND INNAHRIRTI*SRSSAFQHFDGTGDRTCWNRGVQVD
27	Clostridium thermocellum	Clostridium thermocellum	Cmyc	183aa	VYKKSNGKDKIIGFYRYLMEKYNLEIRSTIDNFENEDSFIGFVYSDGDGL

FN Antibody Screening Results



	hypothetical protein Expect = 1e-05, Id = 46/172 (26%)	hypothetical protein Expect = 7e-05, Id = 12/33 (36%)			GDFFKNIKKVFIAADKRKVPNIEEYKFLKSFSAILDEVTKESLAETLNEI FKGTDSEDPKRWGEFLIVGGDDVCAVFDPTLAIIEISVKTQKKFEDTMEA RMSELSKKFTDSKLEEGICKVKITSSSGVVIGVQVD
28	No similarity	Human DNA synthetic construct Expect=0.01	Cmyc	56aa	PYSLTRLFA*EW*VSTVGSAMPLRSSSSVGSSRRLRRRVFRPIRDSFQAPR ERHVG VQVD
29	No similarity	Similarity to human DNA and Bacillus cereus, same E value??	Cmyc	66aa	FLLAIKLLKKSSNSWSAAFVIGVSAIIFVAESTLFFVVSATTVLVESAG STFFLLQLVVRLIIGVQVD
31	Anabaena variabilis Protein of unknown function Expect = 2.2, Id = 29/84 (34%) Only 2 matches	Porphyromonas gingivalis glutamine cyclotransferase-related protein Expect = 9e-12, Id = 19/34 (55%)	Cmyc	191aa	PLPRSTTRTPCPT*PTVTRRLSTASPADSRMLSCPACWTSRRAPTPLSTAE SATRQPWRRWLRSTPVTCTSSSWRTAADRTARPART*RTRPTTSGAS*T DMPSR*NCAPSPPTISRIPN*GLTAVGKFFA*Y*APNAQESSLF*YPVAQ VGTPQSSASKQTKSTPTVSFLSACN*RANSSITATPLGVQVD
32	No similarity	Homo sapiens BAC clone Expect = 1e-74, Id = 118/118 (100%)	BOTH	119aa	AENLRVTVAPSGNHL*LKARPQPSVN*EFLCQLFNFQGLLC*NQMSKR*N GSAVSRKCQ*HTQ*QKETKL*KATWKILTLY*ISPFQFKK*GFVFSWSLFK LEPFIKGP*NLFIMEMPGVQVD
33	No similarity	Bacteroides thetaiotaomicron beta- hexosaminidase precursor Expect = 1e-39, Id= 62/97 (63%)	Cmyc	100aa	ITVQEFFHTMGEMAL*LGYIF*VMLLHLLTIRTSLPCAALCLVSSDVDIL RREELHYLFQHVLHEGIGGFLTYTEVRLCIRFACT*QLWVGINLIAMGV QVD
34	No similarity	Homo sapiens BAC clone RP11- 511H23 from 7, Expect = 3e-27, Id = 48/48 (100%)	Cmyc	47aa	RIRTKE**D*KKYAAFYLISLWITEFHLHT*TILC*IA*ERSSSSNWGVQVD
35	Homo sapiens ch21 similar to seven transmembrane helix receptor Expect = 5.5, Id = 33/103 (32%)	Homo sapiens chromosome 21 beta-site APP-cleaving enzyme 2 Expect = 0.70, Id = 14/27 (51%)	Cmyc	184aa	PTCKVRRGSYVEKPHVCREATSLVAFRNTCGFSSIGVLPFGWLVRFLTS PPSLHTPHRYEYNSMSTNLFTRRAMLSGGTLLGLGALLAACGFLILQIIL RLRFRAHLSAPLSHWFL*ASSSRESCYRSNLGIHDVTLASRVREYCLSRG ERYVIYDNTSSYSHRFI*RQPP*CVCSID**PGVQVD
36	No similarity	Bifidobacterium adolescentis anthranilate synthase component I Expect = 4e-62, Id = 112/170 (65%)	cmyc	174aa	PRRPSRKPC*RSPPRPTGRSNRPRRTRRAPRARNHR*AGWHEAWA RPTGCPQMSQPGHQARQHPRAARSPWRKCAECERNA*VP*A*RPRPT RDRRPSTGRFEPGPRA*RAPTAPSGPREARQPGDGPRPNPRREGATLRSD TS*RGVVQSSRGSRETSQQPQNDRPGVQVD

FN Antibody Screening Results

37	No similarity	Bacteroides vulgatus putative GTPase, ThdF family Expect = 1e-40, Id = 80/110 (72%)	Cmyc	111aa	LCQ*SGK*FPVSLLCQP*HRG*YHEYLKQSLSPHVTSNSRALICPHLSCQ*WLLALHASVRYLFQTT**APRYARVLFQPTPSTPLYQQTLPAPHDRRSRALTQVGTSSSRGVQVD
38	No similarity	Human chromosome 4, Expect=3.1	Cmyc	68aa	GAGILLIILL*PDNIIPAAAPKAILLNEPPVGILGFSLINILLALPVISSKAFC PISAASSCWSAIGVQVD
39	No similarity	Human DNA Expect = 3e-85, Id = 80/82 (97%)		139aa	PALLDWGSSPG*YPAECFPTWFHSPHHFQVHQSDVDLVFVSHSPIFLGGFA HFFLFFFL*TLLASFHSFHLPLLIPFLPVDRIGS*GFCILHVLEPWFSAPS APLSTSLYWLF*LYILLNFFQSFQLLCLWFECPPVGVQVD
40	No similarity	Streptomyces griseus subsp. Griseus conserved hypothetical protein Expect = 0.027, Id = 15/25 (60%)		53aa	SSTLVPWSHWSPWCASSGSPSSGWQDARTPSTRVPPKASLKLSS*PGPA CEGVQVD
41	Aspergillus fumigatus serine/threonine protein kinase, Expect = 4.8, Id = 11/15 (73%)	Streptomyces griseus subsp. griseus putative guanosine pentaphosphate synthetase Expect = 4e-07, Id = 23/25 (92%)	Cmyc	26aa	PMR*ARTLLASSPSRSPSTRSARSSGVQVD
42	Chlamydomonas reinhardtii hypothetical protein Expect = 3.5, Id = 15/31 (48%)	Possible human sequence, E value=3	Cmyc	26aa	PPSHGGDDVCNVPTALSAWHGDAPWGVQVD
43	Neisseria meningitidis membrane fusion protein / antibiotic resistance efflux pump component Expect = 4e-22, Id = 53/56 (94%)	Streptococcus pneumoniae cation-transporting ATPase, E1-E2 family Expect = 8e-68, Id= 114/151 (75%)	Cmyc	226aa	PICHSFEACEFHVPFPLARHLQSLLSQFALQSHGLLQHYLQ*A*PF*DHV FSI*Q*LLEHFP*GHPARQLSLDHLRIGRQLFLLSQAIQIVFDNHVE*Q DHLRVVHSPTRLVYH**SQSHLFHARLGSFPLGKG*FLVQMHDKQ*LEP RSRRDMLLPCVMLHDQTTAKRSAEAGVKSAQAIAIKSAGISLNRSRITAPI SGFIGQSKVSEGTLNLSGDTTVLGVQVD
44	No similarity	Solibacter usitatus binding-protein-dependent transport systems inner membrane Expect = 0.079, Id= 17/45 (37%)	Cmyc	57aa	R*WCSRSVDVLDLIGVHGGHSLVGAWACCGRAIASGVPLNREGSRRSH GGALADGWGVQVD
45	Hypothetical protein homo sapiens Expect=0.02	Homo sapiens chromosome 8 Expect = 5e-61, Id = 94/94 (100%)	Cmyc	95aa	PV*EMLLL*PTLDLQPVVPPAVVILSWKPSALTISSQVL*EQFSAGALILG K*CPWNPVLTPTRRHLCHHR*GG*RGTLTQGAKAGCQGS*PSLGVQVD
46	No similarity	Haemophilus influenzae conserved hypothetical protein – predicted ATPase	Cmyc	58aa	NVTL*N*KKPVGSPKISA*NKH*LIRPSSNLLPIMP*WMPMQILNKKTVNY LIPL*MGVQVD

### FN Antibody Screening Results

		Expect = 7e-23, Id = 47/57 (82%)			
47	Clostridium botulinum phosphate uptake ABC transporter, PhoT family, permease protein Expect = 2e-08, Id= 33/66 (50%)	Clostridium botulinum phosphate uptake ABC transporter, PhoT family, permease Expect = 9e-12, Id = 33/62 (53%)	Cmyc	68aa	YIISASLYVSLLSLIWALPLGIGTSVGLSLGVSPRIRQFCLSTIDMIAGIPSVI VGFIGLAVVVPGVQVD
48	Homo sapiens ring finger protein 170 Expect = 3e-08, Id = 18/23 (78%)	MACACA MULATTA BAC clone CH250-327L24 from chromosome 13, Expect = 8e-09, Id= 23/23 (100%)	cmyc	25aa	PGQAQWLTPVILALWKVKAGGSLEGVQVD
49	No similarity	Campylobacter concisus transformation system protein Expect = 1e-39, Id = 75/88 (85%)	Cmyc	120aa	LRSWRLD*RILPPLLPKTALIIV*KTNLLGVLGLMPLVIYRQGENLLWKV KIFSKFLKCKFG*IFIDKKS VHLVL*VDELFDLRVAIEAINFKAL*LSYQRA WLA*NLRAGFTKCL*RKGVKVD
50	chromodomain helicase DNA binding protein 4 Expect = 0.72, Id = 29/76 (38%)	Azospirillum brasilense plasmid 90 Expect = 0.088, Id = 12/28 (42%)		167aa	SQSEHQAL*NAKVPKQGTGYSIMSLSPEVVRHKPTSLTRISSGDVLRARR RVITFIPQNDITLEGHTCLLTFLAARLSLAQHGLLRSLQHLLQFPLLPPLS ATTHLLPSSTFPVSPPALRTPSSSPFRRSTRSSSRLVAQVAVPTRLLVVP VLS*PV*SPSRKGVQVD
51	No similarity	Homo sapiens genomic DNA, chromosome 4 Expect = 9e-80, Id = 128/128 (100%)	BOTH	205aa	SKAQQ*QGCPQWH*RD*LDCSTCCMPGTPLLQQAASFPQESDLHSHH AC*TSGHPTLFLDLPPCD**TTECIPANLDGLFSLP*FSIG*SLIILLFKILHIL IVLSFL*AFSNITYKFLRQLILVKGLL*MEMFVSPQNSYIE*SPNP*YGIFGD RTSNSTTKEAI*G*MHKGEALL*QD*CFIRRDSRELSLVHTH*GKGVQV D
52	No similarity	Magnetospirillum magneticum Molecular chaperone Expect = 0.043, Id = 14/28 (50%)	BOTH	37aa	RS*KRTCRRGSARPRQPGYPCR*RRPGACS*WCS*GVQVD
53	No similarity	Myxococcus xanthus cell cycle protein, FtsW/RodA/SpoVE family Expect = 0.002, Id = 18/32 (56%)	BOTH	75aa	RRGPSFRPASTSYSPASSSSSTGSRRLTRPASGPG*PTR*RAGPRRRSR TRSQPAGSRSSPARLCTRWRFRGVQVD
54	Actinomyces odontolyticus hypothetical protein Expect = 4e-27, Id = 55/63 (87%)	Moorella thermoacetica ATCC 39073, Protein of unknown function Expect = 9e-05, Id = 23/61 (37%)	BOTH	64aa	KSWNFQDAGIGMAAINAYHSHPEVALARGFTPCEENNWARTFHPYAPL VAGKRVAIIGHFFAGVQVD
55	Acaryochloris marina acriflavin resistance protein, putative	Clavibacter michiganensis tryptophan synthase beta subunit	BOTH	15aa	PRTRFPPASTTRVSGVQVD

FN Antibody Screening Results

	Expect =21, Id= 9/11 (81%)	Expect = 1.0, Id=14/14 (100%)			
56	No insert				
57	No similarity	Fusobacterium nucleatum subsp. nucleatum Protein translocase subunit secY Expect = 2e-112, Id = 181/189 (95%)	BOTH	243aa	RKDTRGANRQEPLVCQGC*CSYCSYKKAYSWIINKPNRL*PKTNSSSLN NSKCYCQVSCPLSYSISSRFPFFLNFFKLRNNRYK*TYNY*CINVRDNP** EYRYPCEGSSRKHIYITKYVTFRSHRSKSIHIYTRSRN*SSYSGN*EH**CE NNSFSKFRNFYTT*SRAKFLH*SHIYTYLRLILFITFCSEKTFNFNYFNCS TFCFNSSLRLRLSGGSVDRLTN*LRDLAAGTYARAQGVQVD
58	No similarity	Propionibacterium acnes preprotein translocase SecA subunit Expect = 2e-21, Id = 41/52 (78%)	BOTH	54aa	WAWRTCTRLRTPRSSASSTTRSVRRCSSGIATTSWTPVRSLSSTSTRVA SCRGVQVD
59	Chlamydomonas reinhardtii predicted protein Expect = 1.5, Id = 16/28 (57%)	Bordetella petrii enoyl-CoA hydratase Expect = 7.0, Id = 10/13 (76%)	BOTH	18aa	LRGGRVPVAPHPALALPGVQVD
60	Neisseria meningitidis putative fructose-1,6- bisphosphate aldolase Expect = 1e-67, Id=124/129 (96%)	Neisseria meningitidis fructose-1,6- bisphosphate aldolase Expect = 3e-60, Id = 117/117 (100%)	BOTH	164aa	EFCFNFLSFVNAVGIAYICLDGKFFGGFTA AVGDADVVAAVEEFPHIP VVMHQDHGASPDVCQRSIQLGFSSVMMDGSLMEDGKTPSSYEYNVNAT RTVVNFSHACGVSVEGEIGVLGNLETGEAGEEDGVGAVGKLSHDQMLT SVEDAVRFVKDTGVDALGVQVD
62	One match, not bacterial	Strep at LHS, human at end	cmcy	174aa	IFQHFCLFSI*KVLTNSCRGHLNRRKKP*FQGFDDVK*LFVANHYKGFPTL SNLSFILNQLLAIFDKTNCFLQYRASL*LTKLY*FVPILHGLDLSETSLAHA RAIADKAPPQWINLMPTKEGAPALRQ*GQTL*RSPYIIRFSIRCNRWNIY RLIYITKGF*SFHHERWLKGVQVD
63	No similarity	Anaeromyxobacter ABC transporter related Expect = 0.081, Id = 17/32 (53%)	Cmyc	34aa	CVVLPSTVRVTLVASARWFTAETTWSRVMPATEGVQVD
64	Pseudomonas entomophila P-type ATPase, Mg <sup>2+</sup> ATPase transporter Expect = 0.48, Id= 25/79 (31%)	Homo sapiens PAC clone RP4-765G7 from 7, Expect = 2e-94, Id= 141/141 (100%)	BOTH	142aa	LNRDFS*LCFYTANREAPET*WASVGEWQQFLKSCQCLSASTLVSARE ESTPPKEKKSANKTSAARLFCGGWHRDGLLSANLCRAYSQPLRKPFGHS SSPTPPTVFQPLAFCGVEGFFLPFFSLPQFFLVSLRP*PQGKGGVQVD
65	One match - mammalian	Pelobacter propionicus PpiC-type peptidyl- prolyl cis-trans isomerise Expect = 0.73, Id = 13/23 (56%)	BOTH	40aa	PNGAPPSMCMSSSRDRRRPMP*TRTPLSPAGGSPLQPSVGVQVD

FN Antibody Screening Results

66	No similarity	Streptococcus gordonii conserved hypothetical protein Expect = 0.001, Id = 12/24 (50%)	BOTH	94aa	CLYYSRGNLHSLGE*RSNLHSPRFSFGSP*PILSRLNHQMNSPSD*LHHRIF LKALCKRFLQHRSHYFQL*IHLQSCPQP*DFSQID*IQI*QGVQVD
67	Rhodopseudomonas palustris hypothetical protein Expect = 0.15, Id = 18/29 (62%)	Kineococcus radiotolerans ribosomal protein S7/translation elongation factor G Expect = 2e-10, Id= 30/53 (56%)	BOTH	63aa	LIKT*KQRSPTMFHSTHVFTGETICRRPAQSLVFTSPA VVRERAVGLSHLV HVLTA LHS GTEGVQVD
68	Methylobacterium sodium/hydrogen exchanger Expect = 1.7, Id = 22/61 (36%)	Bacteroides vulgatus 4-hydroxythreonine-4-phosphate dehydrogenase Expect = 3e-32, Id = 55/95 (57%)	BOTH	277aa	FLRRYT*FNLNLVIEDREQVKPSILCF*CIFNNS E VCFNIECIAMIGGYFRR TIDDGRAEFQHLWFSKGLKDKLITNAVRVSVDCY TDSFILVHIIISCIVLF YSVSFLSFTFLSSFFINLLCR*EQTASHVKHVGFGFFTDHVHFFVNGDPA DQLFVFDNRRGDQVITFKRLRRFFHVFGTEAHDIGGRPRCLPPSESMG ALVSGKRRQSRTCGRYGLNGCFRRPPVCGNLLQHPREMNTSPPPPPLH RGFP*ALPEDLLRQLQSTPVSQGVQVD
69	No similarity	Clostridium difficile chorismate synthase (5-enolpyruvylshikimate-3-phosphate Expect = 2e-05, Id = 29/83 (34%)	BOTH	84aa	*TVFDANKIWIA*SN**TCNYDFN*TTGGVADLF*LHYKFFYFNTFICIEH NNWLCCDVTCKCFHIWMNHIYRSILNHSHLF*MGVQVD
70	FTCDC superfamily conserved domain - Porphyromonas gingivalis hypothetical protein Expect = 0.003, Id = 21/40 (52%)	Clostridium tetani formiminotetrahydrofolate cyclodeaminase Expect = 6e-06, Id= 23/34 (67%)	BOTH	37aa	LTAATGAALAEMVANLTFGKKGYEEVQSEMEELQTKGVQVD
71	Neisseria meningitidis hypothetical protein Expect = 1e-10, Id = 42/78 (53%)	Neisseria meningitidis conserved hypothetical protein Expect = 1e-14, Id = 41/74 (55%)	BOTH	77aa	AEAGHIEAAFQLAGCLFENHENEQDLAIAVEYLKQAARAGHPYARYNL LQLQENNGAEVETLISAYQELAEGLVPGVQVD
72	No similarity	Streptococcus sanguinis CbiG protein, putative Expect = 2e-27, Id = 58/85 (68%)	Cmyc	148aa	SPLLLVSSWQQFSSLLAFTCWFTREFSGTAIGFFSRKAISPTTRKATRGTT DPQH*STGSKPRLSVCHLF*GRTSSSS**VSSIGICEEDCRCRKCCTC*CGP CK*RKCADPTVRTKRCNICTW*IR*QLLNIFKEK*ICYTLLDLGVQVD
73	No similarity	Streptococcus salivarius isolate 8 DNA polymerase III (dnaE) Expect = 2e-21, Id = 43/43 (100%)	Cmyc	44aa	PLTRKRFRVWRVTSSESSDLHSQNNALYVKELQDNQTLQCQHLLGVQVD
74	1 match – not bacterial	Some matches human, some bacterial	polyH	85aa	*QTRSSIRRWARA WRSQFHRTSLRVMTRSAWSSISSVSPAPRQQGIAHL

FN Antibody Screening Results

		E value =1			SPDRPSRESRRA*ASRRVGHWETC*DG*PPENRTGVQVD
75	No similarity	Chromohalobacter salexigens glucose-methanol-choline oxidoreductase Expect = 0.021, Id= 17/33 (51%)	polyH	272aa	LRS*LVTQIPPSLTLTCVTSIPRDMNAIAYRIADTSRTDMAGSRLDQVLRV LRPFFMSFPQCRWHAASAGTGASSRGHTPANHMDNSITRSTDTCKVSSR VNVLYGTIPPLRRIWRT*PHRLTPAHHLRYPSTTSRPF*ELVKRKPKK RGRQA*KVPLIFLFLESIALWKRKRLIPENLQIKPPRSSWSRSSLNAWA VELPVRSSSNPTKEYTREYASPLPLPFSP*DSLRRTKVYSDAPRDNSSVE SPKDSVLTPLHGTPAVAPGVQVD
76	Actinomyces odontolyticus hypothetical protein Expect = 9e-05, Id= 39/90 (43%)	Bacillus amyloliquefaciens YqjI Expect = 2e-88, Id = 90/129 (69%)	BOTH	241aa	*IKFAGNVSEY*YANALENVGTGIPARTAVTTTSRTASW*DSVTYKK YLSNNKFFN*GSRNASVIFCKNCARMIHPARKILAI*FKSQLNSSLAT RI*ANPCA*LMIKYNPLR*CSIKA*RSEFVNSGILVAANSLLAATRSSFN VDTKRANTDSVINGSGTPKSNALCEAHLVPLFPAVSMITSTIGRPVSGSF FVKISAVISIR*PNSTGTALVPGLAAIAIVAGRRLVGVQVD
78	hypothetical protein Bacteroides stercoris Expect = 21, Id = 9/11 (81%)	Homo sapiens genomic DNA, chromosome 11 Expect = 2e-09, Id = 25/25 (100%)	BOTH	25aa	PSS*YHHLGG*GSTV*RMQTSALAGVQVD
79		Bacteroides vulgatus transcriptional regulator, involved in iron uptake Expect = 2e-09, Id = 18/33 (54%)	BOTH		First 300bp
80	hypothetical protein Actinomyces odontolyticus Expect = 4e-10, Id = 33/34 (97%)	Rhodococcus beta-glucosidase Expect = 5e-06, Id = 22/32 (68%)	Cmyc	35aa	SSVPALVNGYLTGQGGAAAMLVDVLTGVVNPVSGRLGVQVD
81	No insert				
82	Human zinc finger domain	Equal matches of mammalian and bacterial DNA – e values of ~0.5	polyH	40aa	IVPSEIARAPEIFTHRGTFISPVIGSRILCVSP*PPASRGVQVD
83	No similarity	Not bacterial	Cmyc	49aa	LC**ETPFSHYHGPHR*S*WSPHYIHNSYCIQCCSHYHSAQSLCHLC*GVQVD
84	Human	Human chromosome 7	Cmyc	232aa	LINPVYCHLKFTNLMYEKWLIVSLAFSIL*VGVSILSYD*R*LLKVCFAG YLLT*LGENL**NFTKI*SFFVN*Y*SYIVKIFQNM*RTKLICCVTYNYF*I NLVI*DISDCIFFFGKSTMIFYLQSIQK*SCTI*RQYILKYLSVFLSILL*T

FN Antibody Screening Results

					LFTYP*SLITQYIVDIILNKPLNVRSTWSTW*NLVCTEKCKN*PGMVSHT CNPSYSGG*SRRSLEPGRQRLQWVQVD
85	multi-sensor signal transduction histidine kinase Methanoculleus marisnigri Expect = 0.51, Id = 22/66 (33%)	Not bacterial		145aa	LLVVHLDGDHQ*WPNPPQRGARRPRCPCRGSRARRAGRRQEDLKGRPNRW EPDHTPKRTFDSLPA*ST*YLPDGRPMSVHDALPIYPLIRFLLFFFTSKLIS DELLSDQWSFLDGKTL*SLHFGGICSL*PLRPLAPS*PRGPHLSFFRVQVD
86	RNA binding protein Chlamydomonas reinhardtii Expect = 3.4, Id = 24/68 (35%) 1 match only	Not bacterial		79aa	SMRSDCGRMLGVSGRVCAGSG*GSGSSWRDIGVSLCATFSLCAAVSLCG AVVGAAT*FAK*VSLA*TADSLRGASAFCTGVQVD
87	No similarity	Haemophilus influenzae PittEE, phosphoribosylamine--glycine ligase Expect = 1e-56, Id= 79/81 (97%)		110aa	NNG*SNLFILSKFH*GRDICRRNIHVHFV*FQCGTCV*ARCDKNFIC*WRLC CFPC*GVFTT*AISYDEDVHVIVLAYA*FKKVRLFFP*VIPPL*RG*GEI*TI LNLV*GVQVD
88	1 match, not bacterial	Streptococcus thermophiles calcium transporter P-type ATPase Expect = 6e-86, Id = 82/98 (83%)		277aa	QQSTYSYVCKLVDQCFKCSILLYGYFFPFWLGHWCSSFGDGACPSGGRHP LVAKGPATLCP*PWLIGS*ALELSTSSSRAADDAGWRCGNHCDR*GCF WDRGSRKCYRRD*LDPV*LHACVWGGH*SRRTSCHCNHRS*LF*GDDNPC QTEFDRS*ITSG*NTWFNRNHRI**NRYFDHEPND*TSVYQR*IAKLSK* NCC*QQYSSCHELCQ*YQGRPIW*N*LGIQKLLWYSLVWTTTFD*VREV FEG*ASCG*IAI*L*S*ALVHGHLRGD*GALTRGVQVD
89	No insert				
90	2 short matches – not bacterial	Lactobacillus acidophilus glycolate oxidase Expect = 8e-57, Id = 60/89 (67%)	cmc	159aa	RTKGFDT*CYKR*GR*CERI**GVSGQHSRGDEGNR*CSSSGPRL*NIWREV LFADYDACFFASK*GRKGWQETDGSVCRGGFGAKCFKLGRHGTE*RICR NSSRC*DS*EDNKAFHES**DSR*DKVC*RARGNRCRY*YR*PCAWNRWQ IRCCRRISSRGVQVD
91	Bacteroides fragilis tyrosine type site-specific recombinase	Matches for both mammalian and bacterial e- values around 0.3		139aa	PPVPPGRPERK*WVPGTREAHPQ*QEPPEQPAVRRGIRYAPPPK*PAGRGG*P AHRRALRNGQGQ*QKYIRIDPFADYKAE*LP*HRRYLTTEELQRL*LTPI IDRQFERARQLPSSW*WGVVTVIGIDLGFRI*ELCATWGVQVD

FN Antibody Screening Results

	Expect = 0.010, Id = 21/59 (35%)				
92	Mycobacterium sp DNA polymerase IV. Expect = 2.2, Id = 15/37 (40%)	Pseudomonas aeruginosa ammonium transporter AmtB Expect = 0.022, Id = 21/51 (41%)		62aa	LTGTARALRIGKEVIKAILPINPDDIRSLAKLGKNMVGKLSAHGLNTAAD VADLANTAKHLGVQVD
93	Leishmania major hypothetical protein, Expect = 52, Id = 8/9 (88%)	No similarity	BOTH	16aa	PAAPRP*RLPLICMAGVQVD
94	1 match, not bacterial	Fusobacterium nucleatum Inorganic pyrophosphatase Expect = 8e-43, Id = 77/80 (96%)	polyH	90aa	IISPAANPATIA*NIPISNPRYLIPKEAIIVNNPAILVPISFACTLLPSSVLTV KVAIIEATIPKADINKGAAT*GVASSTGNKKAKVIVGVQVD
95	No similarity	Candidatus Desulforudis UDP-N-acetylenolpyruvoylglucosamine reductase Expect = 1e-18, Id = 42/102 (41%)		118aa	P*KKPY*RNVLVHVCENRSCFVIIRHLKLVLLIYLLNLRRWMS*ALRYA LFTNYKCL*LLLVAAPIFW*KTVVFAVLSYRFDI*HKS WIAMIMYYALVL VIC*KMLLNLLGKTGYLGVQVD
96	No similarity	Not bacterial		88aa	YYIIMLQETYFAQLFILALSIY*FKHIRIICFIKIILL**IDSKISSFVRSHKITL* I*YFLSC*ILLISLCHR*KLHTTLLHRIGVQVD
97	No insert				
98	No similarity	Frankia sp. ATP-dependent helicase HrpA Expect = 4e-04, Id = 19/31 (61%)	BOTH	48aa	PVTPQNAGTTDIPTLMAFDEGGLDEQFAGCGGVYAGVLYGAVGVDGQ GVQVD
100	No similarity	Campylobacter concisus putative lipoprotein Expect = 6e-36, Id = 67/73 (91%)		73aa	RR*CNRQDRVRNLAFLA*KVSALQVYLKIKFVG*TALNLVQLHHFCYQ ASLLCYNLARLFGYFFLNQKEGVQVD

FN Antibody Screening Results



	<b>Protein BLAST</b>	<b>tblastx</b>	<b>Tag(cmyc polyH or BOTH)</b>	<b>ORF size (aa)</b>	<b>Amino Acid Sequence</b>
<b>1</b>	Moorella thermoacetica Amino acid permease-associated region Expect = 6.9, Id = 12/21 (57%)	Matches to mammalian and bacterial , e-values all over 1	BOTH	28aa	PRPPHRPSPLHAPSPPHRPSPPHEASSGVQVD
<b>2</b>	No similarity	Nostoc sp. NADH dehydrogenase subunit 4 Expect = 0.081, Id = 13/26 (50%)	BOTH	37aa	PRDVTPRDQVMPIWGIHRWDSYTWYDAGADQSKTEAGVQVD
<b>3</b>	Yarrowia lipolytica hypothetical protein Expect = 12, Id = 11/15 (73%)	No similarity	BOTH	15aa	PRAHGLNGFPWANAGVQVD
<b>4</b>	Alkaliphilus metalliredigens phospholipase D/Transphosphatidylase Expect = 4.9, Id = 11/19 (57%)	No similarity	BOTH	16aa	LMDWYFVPKLSEFPHGVQVD
<b>5</b>	Rhodospirillum rubrum RNA methyltransferase TrmH, Expect =12, Id = 11/17 (64%)	Magnetospirillum magneticum Flp pilus assembly protein, ATPase CpaF Expect = 7.0, Id = 11/13 (84%)	BOTH	15aa	PRGAGATGRAPRPPGVQVD
<b>6</b>	Most likely to be mammalian	Homo sapiens chromosome 8, Expect = 1e-08, Id = 22/22 (100%)	BOTH	24aa	PQTFPCHSLPCLSSWWLDLSSSQGVQVD
<b>7</b>	Only 3 matches - Geobacter uraniireducens acetyl-CoA acetyltransferase Expect = 9.0, Id = 17/57 (29%)	Chlamydomonas reinhardtii NAD malic enzyme Expect = 0.054, Id = 12/22 (54%)	BOTH	59aa	LRELLQKIETAHAEFWYYLVYNRLPDGYTG VNGFTKEEE DSFSVYEFLKKQEHA KLHEGVQVD
<b>8</b>	Actinomyces odontolyticus hypothetical protein Expect = 1e-06, Id = 27/27 (100%)	Clostridium tetani excinuclease ABC subunit A Expect = 8e-07, Id = 24/27 (88%)	BOTH	28aa	LIATLERLRDLGNTLIVVEHDEETMEAGVQVD
<b>9</b>	No similarity	Azorhizobium caulinodans putative glycogen debranching protein Expect = 1e-04, Id = 20/29 (68%)	BOTH	32aa	PPCPT*RTWASPPSSCSPSTLSATSPSSPNEG VQVD
<b>10</b>	Bacillus clausii para-aminobenzoate synthase component I Expect =38, Id = 12/20 (60%)	No similarity	BOTH	18aa	PRLSAYGVRQTRSLPKSGVQVD
<b>11</b>	Heliobacterium modesticaldum conserved hypothetical protein Expect = 21, Id = 9/10 (90%)	Pseudomonas putida uI gene, rsaL gene and uR gene Expect = 6.6, Id = 10/16 (62%)	BOTH	18aa	PSWSWLTAPPSATEHSPGVQVD
<b>12</b>	mammalian	Homo sapiens FOSMID clone from chromosome 10,	BOTH	18aa	PREALSPRKWGQAQLLQGVQVD

BSA Antibody Screening Results

		Expect = 4e-04, Id = 17/17 (100%)			
13	No similarity	Akkermansia muciniphila Expect = 4e-06, Id = 22/33 (66%)	BOTH	32aa	PGFGRVDRGVERAIRTDEDVLADDHRSDVEDGQGVQVD
14	Salinispora arenicola Fibronectin type III domain protein Expect = 8.7, Id = 13/29 (44%)	4 matches, all mammalian e-values over 2	BOTH	24aa	PSSWLWLLPFPSSGSSEITRISGVQVD
15	Heliobacterium modesticaldum conserved hypothetical protein Expect = 21, Id = 9/10 (90%)	Pseudomonas putida uI gene, rsaL gene and uR gene Expect = 6.6, Id = 10/16 (62%)	BOTH	18aa	PSWSWLTAPPSATEHSPGVQVD
16	chitinase [uncultured bacterium] Expect = 21, Id = 9/9 (100%)	Homo sapiens chromosome 19 Expect = 4e-06, Id = 20/20 (100%)	BOTH	21aa	PVRSLLGLGIP*LPSGADDKGVQVD
17	No insert				
18	No similarity	Renibacterium salmoninarum ATCC hypothetical protein Expect = 3e-06, Id = 20/32 (62%)	BOTH	36aa	LSPASQHARDRHRLEFPAARTNQGADDAEFGVARGVQVD
19	mammalian	Homo sapiens chromosome 15, Expect = 4e-06, Id = 19/19 (100%)	BOTH	20aa	PLRHRSGQLYQWAWAEKDPGVQVD
20	No similarity	No similarity	BOTH	54aa	PFLSSLPIAPAPAPPMKPPAPAPPRPAAPAAPAAPLPPAP VVPVTPEAPAGVQVD
21	No similarity	Porphyromonas gingivalis hypothetical protein Expect = 6e-15, Id = 34/55 (61%)	BOTH	57aa	LRQKAGIGIDELHRTAGAEAEADDVSIHRVIAEARDVSL KEVFGLAGGELIVPQGVQVD
22	Trypanosoma cruzi hypothetical protein Expect = 1.1, Id = 19/50 (38%)	No similarity	BOTH	26aa	LCSWGVSLSFQYLQRVLELVPESDVGVQVD
23	No similarity	Kocuria rhizophila 30S ribosomal protein S13 Expect = 8e-10, Id = 28/31 (90%)	BOTH	41aa	PSSCLQRCAWDPCGYARWSWCADRGPAAGGVGAPTGKAG TGGVQVD
24	No similarity	No similarity	BOTH	26aa	PHRSYFDHIFTGTVGMFQSSSEDIVEEVGVQVD
25	Burkholderia phymatum 3-carboxy- cis,cis-muconate cycloisomerase Expect = 12, Id = 10/12 (83%)	No similarity	BOTH	14aa	PIWSPRATLSQAMGVQVD
26	Pseudomonas putida general secretion pathway protein L Expect = 2.0, Id = 17/39 (43%)	Human chromosome 9 100%id, Expect = 2e-06	BOTH	20aa	PRLQSRWKAEGGEQPWGAAGVQVD

BSA Antibody Screening Results

27	Actinomyces odontolyticus hypothetical protein Expect = 1e-10, Id = 20/20 (100%)	No similarity	BOTH	21aa	LTTLMERLAAAPVQVESPRMGVQVD
28	No similarity	Human chromosome 11, 100% id, Expect= 5e-12	BOTH	28aa	LLLFYCGFPMSAPLSRTCIPPDPDTPDPGVQVD
29	No similarity	Equal matches mammalian and bacterial. E values not less than 0.1	BOTH	35aa	PKIHHAPRADETREPTMTDRLIAPENADVPRTRREGVQVD
30	Burkholderia phymatum molybdenum cofactor synthesis domain protein Expect = 39, Id = 10/12 (83%)	No similarity	BOTH	15aa	LFGHTPQPAEGTAGGVQVD
31	Mammalian	Homo sapiens chromosome 17, Expect = 0.003, Id = 14/14 (100%)	BOTH	15aa	LPKWWDYRREPLHPGVQVD
32	mammalian	3 matches only. All to - Medicago truncatula Expect = 5.1, Id(64%)	BOTH	23aa	PGVILPIDRHQHRHVLPAQDPGGEQVD
33	Mammalian	No similarity	BOTH	24aa	PSSWLWLLPSSSGSSETTQTSGVQVD
34	Burkholderia pseudomallei sulfate adenylyltransferase subunit 2 Expect =0.018, Id = 14/18 (77%)	Actinobacillus succinogenes sulfate adenylyltransferase, small subunit Expect = 7e-04, Id = 21/23 (91%)	BOTH	25aa	PSIWKALTASASNTTESLPPENSKGVQVD
35	Synechococcus sp glycosyl transferase, WecB/TagA/CpsF family Expect = 4.8, Id= 7/7 (100%)	2 matches only - Verminephrobacter eiseniae transglutaminase domain protein Expect = 2.0, Id= 10/19 (52%)	BOTH	22aa	PGRWWRDHHDYTHAAEAGPFGGVQVD
37	No similarity	No similarity	BOTH	29aa	P*TRPSSPTSTTSRPATVSPSASTRATPGVQVD
38	Arthrobacter aurescens ferric enterobactin transport system permease protein Expect = 0.83, Id = 13/17 (76%)	No similarity	BOTH	22aa	PRGTLSNRLFEIFSTPVAPPPGVQVD
39	Actinomyces odontolyticus hypothetical protein Expect = 1e-10, Id = 20/20 (100%)	No similarity	BOTH	39aa	LTTLMERLAAAPVQVESPRMGVQVD
40	No similarity	Anaeromyxobacter sp. CHR domain containing protein Expect = 0.29, Id = 14/31 (45%)	BOTH	49aa	PGTAVPHVNPLIGASAGMPTYGTEGLRAEQRVPRWHAKA PSALDCDRQGVQVD
41	mammalian	Human DNA, Expect=2e-08 id 100%	BOTH	25aa	PLGFWSKVLPSADNDSPFERQPLGVQVD

BSA Antibody Screening Results

42	Verminephrobacter eiseniae Tfp pilus assembly protein tip-associated adhesin Expect =16, Id = 13/21 (61%)	Bifidobacterium longum LacZ Expect = 3.7, Id = 12/17 (70%)	BOTH	19aa	PWSSPTKTPSSRSPDSSTGVQVD
43	1 match - hypothetical protein Victivallis vadensis Expect = 6.7, Id = 15/34 (44%)	No similarity	BOTH	37aa	PRVTGQRLCLHPSANA VDAVDAVDAVDPVDTAGAAEGVQVD
44	No similarity	Symbiobacterium thermophilum UDP-N-acetylglucosamine pyrophosphorylase Expect = 0.11, Id = 15/25 (60%)	BOTH	34aa	LLPWHYAH*VVPWGDGLDDFAGRPPDLDLMSPTGVQVD
45	No similarity	Not bacterial	BOTH	28aa	PRVPFLSPHFSPNEEMAEPMDITNDPGVQVD
46	Thermofilum pendens Hrk 5 hypothetical protein Expect = 4.9, Id = 11/18 (61%)	Sulfurimonas denitrificans ATP synthase F1, alpha subunit Expect = 0.69, Id = 14/15 (93%)	BOTH	17aa	PPALWRVRASTSPSPGVQVD
47	Clostridium scindens hypothetical protein Expect = 7e-08, Id = 24/30 (80%)	Carboxydotherrmus hydrogenoformans DNA polymerase III, Gram-positive type Expect = 8e-08, Id = 23/27 (85%)	BOTH	29aa	PRIPKSELQRYREGLIIGSACEAGELYQGVQVD
48	Natronomonas pharaonis predicted transporter -predicted permease Expect =29, Id= 10/15 (66%)	No similarity	BOTH	16aa	PITWLTPYRGLPSDEGVQVD
49	Cryptococcus neoformans var. neoformans hypothetical protein Expect = 6.5, Id = 11/15 (73%)	Homo sapiens chromosome 15, Expect = 4e-06, Id = 19/19 (100%)	BOTH	20aa	PLRHRSGQLYQWAWAEKDPGVQVD
50	Ralstonia eutropha Secretion protein HlyD Expect = 8.8, Id = 11/19 (57%)	Mesorhizobium tianshanense Expect = 9.9, Id = 10/17 (58%)	BOTH	19aa	PLGWYWPYPIGSGAGSPQGVQVD
52	no similarity	Human DNA Expect=4e-21, 82% id	BOTH	70aa	LIALPLGTWIGHTGRGPREAAPIVWG*FF*VNLPIN*KNKG DGTGWIRLLARHYMASAGDKASVVPVSLGVQVD
54	3 short matches - Stenotrophomonas maltophilia hypothetical protein Expect = 0.26, Id = 21/50 (42%)	Human chromosome 14 Expect=4e-30	BOTH	156aa	PAMTADIK*ADAASSSKAHITAPRQFLFF*S*LLHTSGFFMC KTNCY*IIVKARLGKEQ*VSHGGDAFNDISANEIALHGFTY **LSSNSHVLLSACH*RPSVTSNCQPRAGTSSSAIPTLRAPT GSRPPRWTPPLSSSVIRPQMAAAGS*SVVGVQVD
55	not bacterial	No similarity	BOTH	18aa	PTWRCAPHASSASPPRGVQVD
56	hypothetical protein	Clavibacter michiganensis conserved	BOTH	23aa	LLEWACDASSQWSLPTLRLSEEQGVQVD

BSA Antibody Screening Results

	Actinomyces odontolyticus Expect = 3e-07, Id = 17/22 (77%)	hypothetical protein Expect = 2.3, Id = 14/22 (63%)			
57	no similarity	Clostridium phytofermentans binding-protein-dependent transport systems inner membrane Expect = 1e-21, Id = 43/60 (71%)	BOTH	62aa	SIVTKMNTIITSERPEPRFQFEVVVNSCSIILPIKYILPPPSMF EIAKVVSAGTNTMVMPLGVQVD
58	no similarity	Pseudomonas putida protein of unknown function DUF535 Expect = 6e-14, Id = 24/61 (39%)	BOTH	201aa	LFSLI*FLATLSSNRMLILLS*NISQ*LF*VWL*IFMNLHYHLI K*FIASFVLTTF*FDFFNWLAIFFQWNFVKSRIL*KFATTFL P*RIEFRHFSTMNTY*MCICIKTFIRYSIDCFYTAFFSNQSK*I YNKVFWSITKHLFGSYFCII*TILIALNRTNPHGFFPIKTFTK SEISMIYTF*F*C***C*QAFFTLPGVQVD
59	hypothetical protein Actinomyces odontolyticus Expect = 3e-06, Id = 25/38 (65%)	Mycobacterium sp. aspartyl-tRNA synthetase Expect = 9e-16, Id = 34/36 (94%)	BOTH	78aa	PDGSMQCHQCQNSVVAVIDSQDRIVFTHYASNEGDPV AACSRPTWAPW*CPVAHPSAAPWTLGRNGLSSAALRG VQVD
60	no similarity	Thermobifida fusca carbamoyl-phosphate synthase large subunit Expect = 1e-21, Id = 44/72 (61%)	BOTH	81aa	PR**GPRPWVLPVRRVPPRTCAPSAPGCRCDAAQKRYQRS PGS*SPFLPYGRWAGRSLVYRCRRSR*TPCPMWAYPWTHP GVQVD
61	no similarity	Renibacterium salmoninarum Trk system potassium uptake protein Expect = 8e-21, Id = 40/45 (88%)	BOTH	46aa	TSTPQESSNIHSFLGLFTRATVRGTPNSVLASREITRFTLSSP VAGVQVD
62	no similarity	Thiobacillus denitrificans ribonucleotide reductase Expect = 6e-11, Id = 28/50 (56%)	BOTH	199aa	PSFAPSRLQ*L*YTACRRSSQEGER*CRSPS*ASVSNFLS*TD QMPGLIDHI**SVKLILCELYYISIRLELFLIVKVKFVKL*P MEGALILLFLGLCWLVVSVKSLVSLMALLLQLFVENDI* TIALPTMHCPIHVPVKKVQSLIRAPGPPEVLQDRSEEARAH LEPEGDLALGAPVTLHQPDLDDGPG*PVVGVQVD
64	no similarity	Pseudomonas fluorescens Secretion protein HlyD Expect = 2.7, Id = 13/24 (54%)	BOTH	30aa	LI*AR*FMRKPRWT*GSSAPEPVRPGAAGVQVD
65	hypothetical protein Actinomyces odontolyticus	Rhodococcus sp. glucose-6-phosphate 1-dehydrogenase	BOTH	17aa	LWYTGNYKDPETGEREGVQVD

BSA Antibody Screening Results

	Expect = 3e-08, Id = 17/19 (89%)	Expect = 7.0, Id = 10/14 (71%)			
<b>66</b>	hypothetical protein Actinomyces odontolyticus Expect = 2e-09, Id = 20/21 (95%)	Anaeromyxobacter sp. cell surface receptor IPT/TIG domain protein Expect = 7.0, Id = 10/16 (62%)	BOTH	22aa	FQKAGITSVTLPLSLRDIKDEGVQVD
<b>67</b>	no similarity	Streptomyces avermitilis putative DNA processing Smf-family protein Expect = 0.43, Id = 14/26 (53%)	BOTH	119aa	LLPVSPLRPGCALPCFSLCGCTWCAFLGAPRRTRVSENHYT HLGSK*QLFAKQLPYEIEK*ERVS*TLRSRPNDGSPHTPPHS SSPAASPPSAAPPNSFPPQHCSRSPALPPCPPQQGVQVD
<b>68</b>	no insert				
<b>69</b>	no similarity	Uncultured bacterium clone Expect = 1e-19, Id = 39/71 (54%)	BOTH	78aa	RCTRIPSSDCTARMSSSVSRGFQKSNMVRSSVASNFTLPSM R*GWMRSSGRKIP*RKEVMKRCSCANRKS SSGVQVD
<b>70</b>	hypothetical protein Actinomyces odontolyticus Expect = 0.012, Id = 19/29 (65%)	Streptomyces coelicolor A3(2) putative ATP/GTP binding protein Expect = 0.12, Id = 16/48 (33%)	BOTH	76aa	ERRRMAEYLASPGQYDVMHVVRARFMAGNYDLCAG VCRDFANTVGNFNIDRGVADGHWTRPTRRRRHGIGLGLG VQVD
<b>71</b>	no similarity	Streptococcus sanguinis SK36 Hypothetical protein Expect = 3e-06, Id = 19/29 (65%)	BOTH	65aa	PIAILNKFGPRKIKMKSPQALPESAK*EINNTR*IMDK*KAT ERRMEVNLIFVG*F*IDAELKPGVQVD
<b>72</b>	no similarity	Acidothermus cellulolyticus 11B strain isochorismatase hydrolase Expect = 1e-12, Id = 28/37 (75%)	BOTH	38aa	RLLGRAPRRVSADRRHAGLAY*SGRSLFGDPRFRGHVGV QVD
<b>73</b>	no simialrity	Bifidobacterium longum NCC2705, LysR-type transcriptional regulator Expect = 1e-04, Id = 16/31 (51%)	BOTH	64aa	PRRLAGCCALV*YSPRSSHLRGDVNLLM*G*MARTGRPSR ASKPTSRLTSRTGTQQDAK*SSRGVQVD
<b>74</b>	sulfate adenylyltransferase Burkholderia pseudomallei Expect = 0.018, Id = 14/18 (77%)	Actinobacillus succinogenes 130Z, sulfate adenylyltransferase, Expect = 7e-04, Id = 21/23 (91%)	BOTH	25aa	PSIWKALTASANTTESLPENSKGVQVD
<b>75</b>	shikimate kinase Clostridium butyricum Expect = 2e-05, Id = 22/39 (56%)	Clostridium beijerinckii Shikimate kinase Expect = 3e-06, Id = 21/36 (58%)	BOTH	40aa	PVFLGSKKGVVIATGGGVIKRRENIDIYKENGFIIFLDRGVQ VD
<b>76</b>	no insert				
<b>79</b>	Neisseria gonorrhoeae	Neisseria gonorrhoeae FA 1090,	BOTH	148aa	EGLWVEEGVTFADLKAVFTDFIRFFERDDLQVRF RPSFFP

BSA Antibody Screening Results

	phenylalanyl-tRNA synthetase subunit alpha Expect= 6e-65, Id= 117/119 (98%)	putative phenylalanine tRNA synthetase, Expect = 1e-65, Id = 120/123 (97%)			FTEPSAEIDIMGENGKWLEVGCGMVHPNVLKNVDIDPEK YTGFAGFGLDRFAMLRYNVNDLRLFFDNDLNFLKQFK*IS DGLLIVRLKSE*ILLISSDKNLFQTGVQVD
<b>80</b>	no similarity	Salinispora arenicola 125 bp at 5' side: hypothetical protein 84 bp at 3' side: FAD-dependent pyridine nucleotide-disulphide oxidoreductase Expect = 0.043, Id = 19/32 (59%)	BOTH	38aa	PS*EPTSCASHEDRCESPSKRVRPQFRVTDVMPCARGVQ VD
<b>83</b>	hypothetical protein Actinomyces odontolyticus Expect = 5e-06, Id = 29/42 (69%)	Streptococcus thermophilus replication initiator protein Expect = 2e-50, Id = 49/65 (75%)	BOTH	344aa	LILGTLFGSTVLVEEAFSCPARQLLADSITTRMPCSCRRCA ARRGRHHCADADR*RCRVRRRSASGGGRGEGSVMVVS LPRALERGRVRVRILPWRCHSDRVPDDRARLAGGRRARG G G*GAYPSRGRVRAFPRRLPRGRRQSARSSKFSQSLT*KLAF LTASNPIHLRQRPK*RAIGKNQ**QRITQILSIERRRYIEKYH SISTRISKSSRFIKTRKIYLYQH*QINSNL***GPIT*TNIS*ITK ITR*YNSRI*YSRKSRRLFYSL*NDKRR*YKSNFLSTIGLW WLSSYTPRLQPKLTKRI*GVTSVEADYELRSIKSIRN*TVTI RPCF*YL**GVQVD
<b>84</b>	No insert				
<b>87</b>	no similarity	Frankia sp. Phosphate ABC transporter, permease protein PstC Expect = 0.017, Id = 14/22 (63%)	BOTH	34aa	LQWLQQWLGGAPIADLLTRGLGGENGTYLTKQGVQVD
<b>91</b>	Haemophilus influenzae outer membrane protein, PittII Expect = 6e-53, Id = 98/110 (89%)	Mannheimia succiniciproducens unknown Expect = 2e-87, Id = 89/115 (77%)	BOTH	217aa	FAGGDRSVRGYGYKKISPKNKEGKLVGGSRLVTGSLEYQ YQVYPNWWGAVFADTGLAADAYKANELRYGAGFGVVRW ASPVGAIKFDIATPIRDKDNSKNIQFYIGLGAEI*GMTMSEQ EKQPDNQTTPVKKKTKRILCVGSAVIFVPVLGLVTAL SFDSGQRAIQLADKMLDLSIEQVSGGLQDGLVLENLRFQ TTGVDVALPKTRLQLNLGVQVD
<b>96</b>	no similarity	Clostridium difficile 630 conserved hypothetical protein Expect = 2e-06, Id = 27/69 (39%)	BOTH	156aa	PQ*RRTSYYNQGNL*TR*KGEGSINIL*VARN*QP*MLKRR RLLF*NLNVLN*EINH*SINILNNIITKRRIALLHCICGKKPM AFVGLKKMAYCTSKVVVSAIHSYYLHLLVRMRSSSMVYC VQKNGL*KINYHSFSKVLARL*KNGWRNYAQQGVQVD

BSA Antibody Screening Results

<b>98</b>	similar to human phospholipid transfer proteins	Bacteroides fragilis putative glycosyltransferase Expect = 8e-22, Id = 29/55 (52%)	BOTH	118aa	L*LL*KRFRCKRTVASIHKDNLITSRLLQSFVHSKVKSSIRF ALYLHDMPILSLIGLTLISLSYTD CIVCRGSINDEMFMNIV LHKDTV*RWL*NVLRIIGYCDYRKCDHRWKLLRGVQVD
<b>99</b>	Outer membrane receptor proteins, mostly Fe transport Actinobacillus pleuropneumoniae Expect = 7e-21, Id = 45/85 (52%)	Neisseria meningitidis conserved hypothetical protein Expect = 2e-26, Id = 49/102 (48%)	BOTH	109aa	LIFLLGLDAPLTDIWKIGNNISTGFRNPTASEMYFSFEHPAG NWIPNPD LKAEQALNQSIYIQAEHLLGSFGLTFYHTRYKN LLTEQESTYKKRNPYYNAYSASYGQQGVQVD
<b>100</b>	no similarity	not bacterial	BOTH	43aa	CHPPFALHPPPRPRIIFP*KKKGAALRATAFAFFAWGTSSA RRVQVD



bacterium SRMC-53-10  
uncultured bacterium; P1D1-517  
uncultured bacterium; P1D1-727  
uncultured bacterium; AYRV2-138  
Neisseria perflava; U15  
Neisseria mucosa; M5  
Neisseria sp. J01

uncultured bacterium; P2D11-603  
uncultured bacterium; FIU\_KM\_MD\_004  
22  
70  
uncultured bacterium; nbw788f12c1  
uncultured bacterium; P2D1-502  
Prevotella melaninogenica (T)  
Prevotella histicola; T05-04  
uncultured bacterium; rRNA247  
Prevotella histicola; N19-30  
uncultured Bergeyella sp.; 450a  
137  
uncultured Prevotellaceae bacterium; 301C01(oral)  
uncultured bacterium; P3D1-597  
Prevotellaceae bacterium P4P\_62 P1  
Bacteroides cf. forsythus oral clone BU063  
47  
Prevotella salivae (T); JCM 12084; EPSA11  
233  
296  
uncultured bacterium; P1D1-514  
Prevotella pallens; 8792  
Prevotella pallens; 9423

Streptococcus infantis; CCUG 39817  
uncultured bacterium; SJTU\_C\_03\_56  
uncultured bacterium; 014B-B9  
uncultured bacterium; LY03  
Streptococcus salivarius; RKA5  
human oral bacterium C23  
uncultured bacterium; 2-002-all

uncultured bacterium; Ax3\_690  
Veillonella parvula; ATCC 10790  
Veillonella sp. NVG 24cf  
Veillonella sp. NVG 84cf  
uncultured bacterium; SJTU\_F\_11\_45  
Veillonella dispar; ATCC 17748  
uncultured bacterium; B2\_052  
uncultured Veillonella sp.; EHFS1\_S01e  
uncultured bacterium; FD03C07  
uncultured bacterium; FIU\_KM\_MD\_005  
uncultured bacterium; FIU\_KM\_MD\_012  
uncultured bacterium; nbw777a1c1  
uncultured bacterium; nbw790b10c1  
uncultured bacterium; FIU\_KM\_MD\_002  
Veillonella atypica; ATCC 17744  
uncultured bacterium; rRNA069  
uncultured Veillonella sp.; 59-8-23  
uncultured Veillonella sp.; KLONG06  
327  
uncultured bacterium; SJTU\_F\_05\_25  
Veillonella parvula (T); DSM 2008  
uncultured bacterium; P4D1-570  
uncultured bacterium; nbw827e08c1  
uncultured Bacteroidetes bacterium; S15B-MN34  
71

uncultured bacterium; Oh\_3127A5C  
uncultured bacterium; P2D1-728  
uncultured bacterium; P2D1-760

uncultured bacterium; Ax3\_475  
uncultured bacterium; nbw791a05c1  
uncultured bacterium; nbw825b01c1  
uncultured bacterium; P3D1-469  
45

Haemophilus sp. oral clone JM053  
Haemophilus parainfluenzae; CCUG 12836  
60

uncultured bacterium; Z775  
uncultured bacterium; Y132  
uncultured bacterium; P5D1-466  
Abiotrophia para-adiacens; TKT1

uncultured Streptococcus sp.; 401B03(oral)  
uncultured bacterium; FC01G09  
uncultured Streptococcus sp.; G15-006-B07  
Streptococcus sp. S16-11  
Streptococcus infantis; ATCC 700779  
Streptococcus sp. F1  
uncultured bacterium; FD02H05  
uncultured Streptococcus sp.; 2P-4-1-G19  
uncultured Streptococcus sp.; 2P-3-2-C13  
189  
uncultured bacterium; JSC7-29  
uncultured bacterium; P5D1-660  
Streptococcus infantis (T); GTC849  
273

Streptococcus oralis; CIP 105158  
307  
uncultured bacterium; nbw740f01c1  
Streptococcus oralis; CIP 104985  
uncultured bacterium; BF0001D086  
Streptococcus australis; ATCC 700641  
uncultured bacterium; nbw824b11c1  
Streptococcus parasanguinis; mother C3; 5C3  
Streptococcus sp. EO2001-01  
uncultured bacterium; P2D1-692  
Streptococcus sp. oral clone AG008  
uncultured bacterium; P2D1-735  
uncultured bacterium; NS03  
Streptococcus parasanguinis; 85-81  
Gemella sanguinis (T); 2045-94

uncultured bacterium; JPL2-3  
uncultured bacterium; LG25  
Peptostreptococcus stomatis (T); W2278  
Oribacterium sinus; F0268

uncultured bacterium; BF0002C015  
Eubacterium sp. oral clone DO016  
uncultured bacterium; nbw816h06c1  
Eubacterium sp. Smarlab BioMol-2301231; Smarlab BioMol 230  
uncultured bacterium; A\_S\_01\_63  
284

Fusobacterium periodonticum; ChDC F314  
Fusobacterium periodonticum; ChDC F312  
uncultured Leptotrichia sp.; 303A12(oral)

uncultured bacterium; nbw828b02c1  
uncultured bacterium; nbu189a11c1  
uncultured bacterium; NH01  
uncultured bacterium; A\_S\_01\_77  
uncultured bacterium; P4D1-488  
Rothia mucilaginosa; J04  
9

Chlamydia trachomatis; B/TW-5/OT

Scale: 0.1

## Appendix 7: pQR492 text sequence

>ORF 5

GGGGGGGGGGCTCGTCGCGCTCTGAGGATGCCGTTGAATGCGGTTGCGGGGGTGG  
CGTTGACGGTGTTAAGCTGTCGATGCCATCTTCGATGGCGTCAGCGATGCTCTGGT  
TTTCGAGTGCGCCTGCGGCGGCATCTTCGAGCAGGTCTGCGTCGGTGACGAGTTCG  
GTGATTTCCACAGCTGAGCTTGGCTTCGAGTGCCTGTGCGGTGGTTTGGTTCGAT  
GGTGCGGAATCCGCTGTCGCTCATGCGGAGCTGTCCGAGGTCGATGGGTGCTACTG  
CGTCGGCTTCTGCGGAGGTGTCCAGGCGGCGGCGAGGAATTCTTCGTTGGTGGCG  
ACGAGGTCTTCGGCGCGGGCGCGCACACGCTCGGGTTCGACGTTAATGATCATGGA  
CCCAGCCGGCAACAGCTCGGTGAGGGTGTTTCATGGATTTCGATGAGTAGCGGGGTG  
AGTGATTCATGCCTTCGACGTAGATTCCGCCGGCAATCTTCTCAAGCATGGCTGC  
GGCTGCCGGGTAGTCGGCTTTCAGTCGTGCGGCGCGGCTCATGACCTCGGGGGTGA  
TGAGCAGTTCGCGGCAGGGCAGCAGGGTGAGTTCGGTGAGTTCCTCCCGCTG  
AGGGTGCGCTGGTCGGCGACGGAGAAGTGGCGCATTCGTCGAGTTCGTCGCCGA  
AGAATTCGAGGCGTACGGGGGTGGTGGCGGTGGGCGGGAAGACGTCGATAATGCC  
GCCGCGTACAGCGTATTCGCCGCGCTTGGCAACCAGGTCTACGCGGGAGTATGCGG  
CGTCGTTGAGACCGCGCACACATCCTTGAAGGGGTATTCTTCGCCCGCAACCAGG  
TGCACGGGTTCAAGTTTTTCGATGCCGGTGACGATGGGCTGGATGACCGCGCGTAC  
GGGTGCGATAACGACCTGGGGGCGTTTAGCAGCTTCCCGGTCATGGCGCGCAGTA  
CCTGCAGGCGGCGGCCTACGGTGTGCGGAGCGCGGTGAGAGGCGTTCGTGGGGCAG  
GGTTTCCAGGCGGGGAAGAGCGCAATGTGCGGCGGCGGGCAGGTAGGAGCGCAGG  
GCTGCGGCGAGGTCTTCTGCTTGACGGTTCGGTGGGGGCGATGATGAGGCTGAGCG  
CTTCGGCGGTGGTATCGCGTACTGCGGTGGAGATGTCTGCAATGAGCGCGGCGTGG  
GTTCCGGCGACGGCGCCGATGAGGGTGCCTGCGCTACGTTCCGGTGGGCTGTGCGG  
AGGCGGCGGTACGGATCGCGGCCAGCTGCTGAGTTGGTTGAGGGTGTTTCAGGAG  
GGGGTGACG

>1290-1924

GGGTGCTTCGAGGGGCTGTGCAGTAGCTTTTCGGGTGGTGTACGCGGTTCTCCTG  
AGTCTGGGGTCTTCGCTTGCGGCGGTTGAGCTGGTGGAGGGCTGCTATGGATGGTC  
GGTATAGAGCGGTGCCGGTTTGGGGCGTCCGAGTGAATTTGCCTTTATTTTCATGGT  
GCTTTTTGTGCGTTTTTGGGGTGTTTTTTGCGGCTTTTCCATTGTATGACTTGTCTGCT  
TTTTCTGGCATTTTTCTTGTATTCTCTGTGATGGAATATATTCTTTTTGCACTCG  
GTTGTGCTTTTCGTTGGGTGATATTTGAGTATGGTTTTCTTCCGTGTCCGTTCCGTGTC  
CACAACGTTTCATCACATGGGGTGCAAATCACACATCCACGGGCATCTTTGCACTTT  
GTGCACCATTATTTTACAGCTTAATCCTGTACCCTAATCATTAACTTAGAGAAA  
ACCAGACAGTATATCCCCTTCCGGATTCTGTTTACCTCTTCCAACCAACCTCAGGTT  
TCAGTATTTACTGCGTTTTTCTAGTGCCTAAATTCACCGTATCTTTCGATGTTTCT  
TCCGGAACGTCGCATCAGAAAGACGCATATGACACTACATACGAACAATCCGC  
TGAAAGC

>ORF1 1925-3481

GCATGCCCATCATGTCTGCTAGATGGGGTGCTTTCGCCTCGCTTGTACTGCTTTACT  
GGCGACCTCTTACTCTTACTATCAACCCTCCGGTTTTTCAGTATTTACTGCGGTTTTTC  
TAGTGCCTAAATTCACCGTATCTTTCGATGTTTCTTCCGGAACGTCGCATCAGAAA  
GACGCATATGACAACCACATACACGAACAAACCCGCTGAAAAGCGCCATGCCCAT  
TCATGGTCTGCTAAGAAGTGGGGTGCTTTCGCACTCGCTTTTGGTACTGCTTTTACT  
GCGGCACCTCTTACTCCTACTACTCCCGCTCGGCTGCGACCGGAACCCATGACGG  
TTCCAGCAGCGATAAGGCGGCTGCTTCTGTTACGAGGTCAAGCAGGTTAATCCTT  
CCGCTCCTCTGGCACTTACTGGCTGTACACCCCGCAGATGAGCGGCCCGGCGCAG  
TTCTACTGCGATCAGGAACTGATGGCGGCGGCTGGGTCATGATTGGTTCGCGGTTCG  
TGAGGGTTGGACCGAGTCTTACAACGGCACCGGCGACCCGAACCAGCTGCATCAG  
AACCCACGGGCCCTTCTGCTTTTACCCCGGTCCAGCTTCCCGCAAACACTGTTCGAT  
GCTCTGCTGAACGGCATTAAAGCCTCAGGACCTTCCCGATGGCATGCGCTGCACCG  
TGCCACAATGCAAGGGGCACCCAGTGGCAGAACGTTTATGTTTCAGCGTCTCAGAA  
CTGAGCAGTGGACCTGGGCGATGAGCTACGGTCAGCGTTGGGGCACCGTGAAGTT  
CACCGGTGCCGCATTAACCGCACCGCTCATATGGGTCGTCATGCTTCGGAGATGG  
CTCCGGGTATTACCACAGTTCTGTTTCGTTTTCTTCGCAACCCGAACCAGGGCTACC

AGATTGGTTTTGCGTACGGTGCCTGGTGAACCTTCGGTAATGAGAATCCGGATTCA  
TACATTTACCACAAGCGTGGTTCCGCTGGTACTCCATCCCCTTACTCAGGTGTTCC  
CTGCGTCCGAAGCTGACTCAGCGTGACTTGAACCTTCGCAAATTGGTTCTAGCTCT  
GCCGCTAGCAACCGTCGTGCTCTGCCGAATAGCTACACGATGCCCGTTTCGTTGGCG  
CACCAGTGAGCAGACTGGCACCGGCAAGAAGAACGAGATGAACACCTACGTTACG  
GCTATTACTCAGGTGGGCGACACCGTCTTACCAGGTGGCGACTTCAAGTACGTGGA  
GTCGGCTGGTGGCGAGCGCGTGGACCAGTCGTACCTGGCGGGCTATAACGTGGATT  
CGGGCGAACTGGTTCGCTCTTCCGTCGACTTTCAACGGTCAGATTAAGGCTCTG  
AAGGCTCTGCCGAATAACCGTCTGGCGTTGGTGGCGAGTTCACCGAGGGTAATGGC  
GAGAAGGTCAACCACTTCGTATTCTGGACGCAACCACCGGCCAGATCGACCGCA  
CGTGGGATCTTCAGTCAGCGCGTAATGCGATGCAGTTCAGGTGAAGACCTGCTGGT  
TCAGGACGGTTACCTGTACATTGGCGTAACTTCACCCATG

>3481 - 4010

TAAAGGGCAATACGTCTAAGGCGTACGCTTACTCGCGTGTGCCGCTCGTTTTAAGC  
TCTCGAACGGTGCAGTGGATTGGAACCTGGCGCCCGAACTTTAACGGTACGGTCAAC  
GGTATTACCGCTGCGTCCGATAACTCGACTGTTACGCTGCGGGTACTTCACTGA  
GCTGAATAACCAGCGTGCTTTCGCTGAGCTGAACATTACTGACCGTATTGTG  
ACATTAGGTGGGAGTGGGAGCCTTCGCTGAAGCTGAACATTACTGACCGTATTGTG  
TACGCGTTCAGTTCGATGTTACAGGATGCTGGCTCGACCGTGTGGACTGCCGGTGC  
TGACCACCTGATTGCTAACTACTCGAAGAACGGTTACGGCCGTAATTCACCTGCGA  
TTTCTAAGTATGGTGGCGACTGGCAGGATCTGCACCTGAGCGGTAACACCATTTAC  
GGCGCGTGCCACTGCGGTGACGTCTCTTTGAGGGTTCACCGGTTACCACACCTA  
CTGGAAGGAATCGAAGGCGGTTAC

>ORF 2 4011 - 4460

CGCATGCGCCTGGTTCGCGGCGTTTGATAAGGACAGCGGCGAGGTTGTGGGCGAGT  
TCAGCCCGGTTCTGAAGGGCGCTAGCGGCTACGGTGTCTGGGAGTCCCTCGTGGAT  
TCTCGCGCAACCTGTGGGTTGGTGGCGACATTAACCGTTCGCTGGGTGCTAATGG  
TGAGCAGCGCACTGTCGGTTTTGCTCGTTTTGCTCCTCGCGATGTGACCGCCCCGTC  
GACTCCGTCGAATCTGTCGGTTCAGCGTGATGGCTCGACCGATAAGCTGTCTTGGT  
CTGGTGTTCGTGAGAGCGGTGCTCGCTACCAGGTGCTGCGTGATGACCGTGTGATT  
GCTACGGTCTCGGGCACCAGCTACGAGGTTGAGCATACTGATGGCGCTCGCTACTA  
TGTGCGTTCATCGATGCGTCTGAGAACTTCTCGGCTTCCACGGGAGCTGCTCAGG  
C

>4461 - 4988

TTAGGTCAGATTGTTGATCTAGTTTCTTTGAGCCCGCCGCTGGTTCGGGCTCTGCAG  
GGTTTGCGGGTCCGGGTCAGGCGGCGGTGCACCCGGGGTGTATGCCCTCTATAGT  
TTTGGTTCGGTCCGGTTTTCCAGTTTGTATCTCTTCGCCCTGTGGCGGCTATGA  
TGGGTTGGGATTCATTGCGTCGGATGGGTTGCTTCCGTCCTTTCCGATGATGGGCTG  
AATCTTGATCTGTGGTGTCTTTCGCTGTTGGTACGACAGAGGAACTTCACAGATT  
TTTAGCGTTCGCTTCGTACGGGTGCGTTGCCACTGTTGGCTTCGTGCGTGAGGGGT  
GGGGCGTATTCATTCAATTTCCGATAACTACACTCACTGATTCCGGCTACAGACTC  
GCTGTTAGCGCAGTTCATAACGCTTCCCCCTTACCCCGATGCCTGGTCACTATGCTG  
TGCCTTGTGTGGGGTGGGTGTTGCTGTGTGAATTGCAGTTTTCTGCTGTGTGCCGGT  
TGAGGTCTCTAGGAGA

>ORF 3 4989 - 5504

GACAAATGAATTTTCTTCCGTTACGCGTGGAGGGCGCGCGCAGGGTGCTGCTTCC  
TCCGCTTCTGAGGGCTCTCGGTAGCTTCTCGCTCCGGTCCCGTGTCTTGGGTGCG  
GCGGCTGCTTCTTTTGCATGGTTGCGGCTTCTTTGGTCCGATTGCTTCCGGCGCG  
CAGGCTGCTGATGCACGCTACTATGACGGTTCGTCGAGCGAGCGCGCGGCTGCGA  
GCTGCTGGGAGGTTAAGCAGAATAATCCGCGTGGCAAGAGCGGCGCGTACTGGCT  
GTACACCCCTGCGATGAGTGTCTTCTGAGCAGTTCATTGCGATCAGGAACTGATG  
GCGGCGGCTGGGTTATGATTGGCCGCGGTCGTGAGAGCTGGACTGAGAACTACTA  
CGGTCGCGGTAATGCCGATCAGCTGTATAAGAACCCGACCGGCTTTGATGCGGTGC  
AGCTCTCCGGCGTGACCGTGAACGCGCTACTGAACGGTACCCGTCCGCAGGATTCC  
ATCGCGCGAGAA

5504 - 5565

CGTTGAGGGCACTACCTGGCAGGATTTTAAGGCTCACCGTGATTCCAGTACCGAAT  
GGACT  
>ORF 4  
TGGACTTTGCGTTCGAAGATGTTCTGGTCAAATATTTTCGGTCAAAAACACCTGGCA  
GTACAGTAACCGTTATGACTACGCCAACCGTGGTCAGGTGGCTGGTAACATCTTCA  
CCCATGATTCTGATGACTTCCGTTCCCTGAACTTCGAAGAGAAGGCATCGCAGGGC  
TACAAGTTGGGCTTCACCTACGGTCGCAACGCCAAGATTACATGGTGGACCGAGAC  
CTACCTGATGAACCGTCCTTCGGCGTATATTTACCGCCGGCGGATGATTTCGACCA  
CTCCCCTGGTATTTACTCAGATGTTCTTGCGCCCGAAGGTGACTCAGAACGATTTG  
GTGGCTAAGGGTCTGCATGATTATGGCCAGCAGGGCGCTGCGGCAAGTAACCGCC  
GCGCGCTACCCAATAGCTACTCCGAGAAGTGGAAGTGGCGTACCAGCGCTGATAC  
CGAACC GGTAAGAACGGCGAGATGAACACCCAGGTTGAGGCGATCACCGAGGTC  
GGTGGCGCTGTCTTTACCGGTGGCGATTTTCGCGTATGTTGAGTCGGCAAGCGGCGA  
GAAGGTTGAGCAGGCTTTCTTGGCTGGTTACGAGGTGGGTACCGGCGAGCTGCGCC  
GTTCTTCCGCCCCAAGATTAACGGTCAGGTGAAGTCGGTTGAGGCTCTGCCGAAC  
GGCCTGCTTGCTGTGGGTGGTTCTTTTCGACCAGGTGAACGGCGAGTACTACAACGG  
TTTTGTAATTCTGGTCCCCGGACCGAGCTTGAAT

## Appendix 8: InterProScan Analysis

### Clone 1







Frame +1/-3

PLLVLLVDPIVSGGNASEADAGHEIAARVWRVGSDLTAGVDVPAPGTQVGLAPEIACG  
HCAPCTSGRSNVCANMRLFGTGVDG

#### SignalP analysis

```
>Sequence length = 70
# Measure Position Value Cutoff signal peptide?
max. C 24 0.711 0.52 YES
max. Y 24 0.430 0.32 YES
max. S 21 0.960 0.97 NO
mean S 1-23 0.495 0.51 NO
D 1-23 0.462 0.45 YES
# Most likely cleavage site between pos. 23 and 24: AGA-CC
Prediction: Non-secretory protein
Signal peptide probability: 0.438
Max cleavage site probability: 0.290 between pos. 23 and 24
```

**tBLASTx** : *Methylobacterium nodulans* ORS 2060, Alcohol dehydrogenase GroES domain protein. Expect = 4e-07, Identities = 23/46 (50%).

SEQUENCE: <u>Sequence 1</u> CRC64: 173F0AF11C959C60 LENGTH: 83 aa		
<b>InterPro</b> <a href="#">IPR011032</a> Domain 	<b>GroES-like</b>	
	<a href="#">SSF50129</a> 	GroES-like
<b>InterPro</b> <a href="#">IPR013154</a> Domain 	<b>Alcohol dehydrogenase GroES-like</b>	
	<a href="#">PF08240</a> 	ADH_N
noIPR unintegrated	<b>unintegrated</b>	
	<a href="#">G3DSA:3.90.180.10</a> 	no description
	SignalP 	signal-peptide

### Clone 2

frame +2

> - 63 codons

RSIAGNDKELYTYMDAYDGDQMARELGVEAKHEVEKLAHKARTVMPAALPVIASA  
PAGVQVD

### SignalP analysis

```
>Sequence length = 70
# Measure Position Value Cutoff signal peptide?
max. C 24 0.620 0.52 YES
max. Y 36 0.305 0.33 NO
max. S 38 0.992 0.92 YES
mean S 1-35 0.888 0.49 YES
D 1-35 0.596 0.44 YES
# Most likely cleavage site between pos. 35 and 36: ATA-CC
Prediction: Signal peptide
Signal peptide probability: 0.989
Max cleavage site probability: 0.904 between pos. 23 and 24
```

No InterProScan hits for this protein

tBLASTx : *Porphyromonas gingivalis* ATCC 33277, glycogen synthase  
Expect = 3e-12, Identities = 31/44 (70%).

### Clone 11

```
frame +2
> - 51 codons
LNLLLELKAVAKEFGMPAFTGGQMAKWLYIQHVTTIDEMTNISKNNREKLKA
```

### SignalP analysis

```
>Sequence length = 70
# Measure Position Value Cutoff signal peptide?
max. C 34 0.464 0.52 NO
max. Y 34 0.564 0.32 YES
max. S 29 0.981 0.97 YES
mean S 1-33 0.818 0.51 YES
D 1-33 0.691 0.45 YES
# Most likely cleavage site between pos. 33 and 34: AAA-AG
>Sequence
Prediction: Signal peptide
Signal peptide probability: 0.995
Max cleavage site probability: 0.429 between pos. 31 and 32
```

No InterProScan hits were found for this protein

tBLASTx : *Bacteroides fragilis* NCTC 9343, conserved hypothetical protein  
Expect = 2e-10, Identities = 28/50 (56%).

## Clone 16

frame +2

> - 57 codons

QRHAIELEKARWIAKDLGVKQTLIDTSVIKSITHNALMDANADIEQKDGELPNTFVD

### SignalP analysis

>Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	19	0.676	0.52	YES
max. Y	19	0.394	0.33	YES
max. S	18	0.996	0.92	YES
mean S	1-18	0.954	0.49	YES
D	1-18	0.674	0.44	YES

# Most likely cleavage site between pos. 18 and 19: AGA-AC

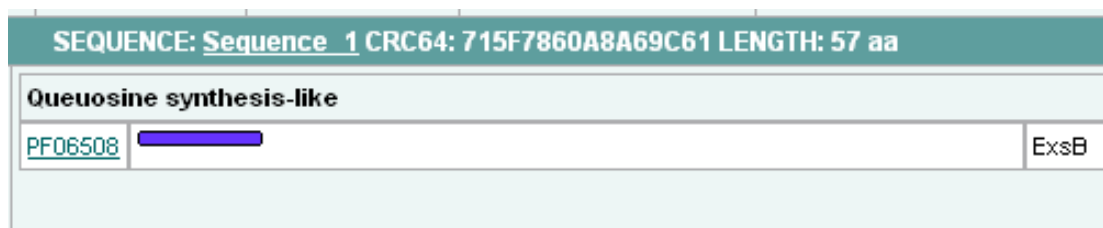
Prediction: Signal peptide

Signal peptide probability: 0.997

Max cleavage site probability: 0.353 between pos. 24 and 25

*tBLASTx*: *Aggregatibacter aphrophilus* NJ8700. WD-40 repeat, putative ExsB protein

Expect = 7e-24, Identities = 48/57 (84%).



## Clone 17

frame +2

> - 65 codons

YIISASLYVSLLSLIWALPLGIGTSVGLSLGVSPRIRQFCLSTIDMIAGIPSVIVGFIGLAVV  
VP

### SignalP analysis

>Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	29	0.278	0.52	NO
max. Y	29	0.297	0.32	NO
max. S	1	0.973	0.97	YES
mean S	1-28	0.863	0.51	YES
D	1-28	0.580	0.45	YES

# Most likely cleavage site between pos. 28 and 29: GTA-TC


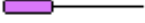

>Sequence

Prediction: Signal peptide

Signal peptide probability: 0.794

Max cleavage site probability: 0.490 between pos. 31 and 32

*tBLASTx*: *Clostridium botulinum* B1 str. Okra, phosphate uptake ABC transporter, PhoT family, permease, Expect = 1e-11, Identities = 33/62 (53%).

SEQUENCE: <u>Sequence_1</u> CRC64: 1E64121C1A6823C5 LENGTH: 65 aa		
<b>Binding-protein-dependent transport systems inner membrane component</b>		
<a href="#">PS50928</a>		ABC_TM1
<b>unintegrated</b>		
SignalP		signal-peptide
tmhmm		transmembrane_regions

### Clone 19

>Frame +3

KSWNFQDAGIGMAAINAYHSHPEVALARGFTPCEENNWARTFHPYAPLVAGKRVAII  
GHFPFAGVQVD

### SignalP analysis

>Sequence

length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	22	0.521	0.52	YES
max. Y	22	0.372	0.32	YES
max. S	9	0.958	0.97	NO
mean S	1-21	0.602	0.51	YES
D	1-21	0.487	0.45	YES

# Most likely cleavage site between pos. 21 and 22: AGG-AC



Prediction: Signal peptide

Signal peptide probability: 0.730

Max cleavage site probability: 0.455 between pos. 24 and 25

**tBLASTx** : *Desulfitobacterium hafniense* Y51, hypothetical protein

Expect = 1e-06, Identities = 17/29 (58%).

SEQUENCE: <u>Sequence_1</u> CRC64: EF48B8EE02EEEA5C LENGTH: 65 aa		
<b>InterPro</b>	<b>Protein of unknown function DUF364</b>	
<a href="#">IPR007161</a>	<a href="#">PF04016</a>	DUF364
Family		
InterPro		
		



## Clone 20

frame +1. This sequence translates into frames +1 and +2 without stop codons, but a tBLASTx search picks out frame +1 as being *A. adontolyticus*.

> - 53 codons

LGVENLYEAANTPLIGFLNNAIRAKELFFRDRDYIVDAGEILIVDEHTGRVLP

### SignalP analysis

Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	35	0.661	0.52	YES
max. Y	35	0.420	0.32	YES
max. S	30	0.848	0.97	NO
mean S	1-34	0.476	0.51	NO
D	1-34	0.448	0.45	NO

# Most likely cleavage site between pos. 34 and 35: ACA-CC  
Prediction: Signal peptide  
Signal peptide probability: 0.786  
Max cleavage site probability: 0.504 between pos. 22 and 23

**tBLASTx** : *Bifidobacterium animalis subsp. lactis* DSM 10140, preprotein translocase subunit SecA, Expect = 2e-21, Identities = 40/53 (75%).

SEQUENCE: <u>Sequence_1</u> CRC64: 176A7374A3317F27 LENGTH: 53 aa		
<b>SecA preprotein cross-linking region</b>		
<a href="#">PF01043</a>		SecA_PP_bind
<a href="#">SSF81767</a>		Pre-protein crosslinking domain of SecA
<b>SecA motor DEAD</b>		
<a href="#">PSS51196</a>		SECA_MOTOR_DEAD

## Clone 22

frame +3

> - 76 codons

AEAGHIEAAFQLAGCLFENHENEQDLAIAVEYLKQAARAGHPYARYNLLQLQENNGA  
EVETLISAYQELAEGLVP

### SignalP analysis

>Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	17	0.161	0.52	NO
max. Y	17	0.204	0.33	NO
max. S	2	0.965	0.92	YES
mean S	1-16	0.861	0.49	YES
D	1-16	0.533	0.44	YES

# Most likely cleavage site between pos. 16 and 17: ACA-TA  
Prediction: Signal peptide  
Signal peptide probability: 0.530

Max cleavage site probability: 0.282 between pos. 22 and 23

**tBLASTx**: *Neisseria meningitidis* 053442, conserved hypothetical protein.

Expect = 2e-14, Identities = 41/74 (55%).

SEQUENCE: <u>Sequence 1</u> CRC64: F2124C90E4B8994B LENGTH: 76 aa		
<b>Sel1-like</b>		
<a href="#">PF08238</a>		Sel1
<a href="#">SM00671</a>		SEL1
<b>Tetratricopeptide-like helical</b>		
<a href="#">G3DSA:1.25.40.10</a>		no description
<b>unintegrated</b>		
<a href="#">SSF81901</a>		HCP-like

## Clone 27

Frames +1 and +3 have no frameshifts in them. According to tBLASTx the frame of all hits is +3.

> - 57 codons

EGTPPENRDGTCRVLVLPVQPPAGRLHGRQWLHEGRRGFFLGVRVSEKARTRKATR

## SignalP analysis

Sequence	length = 70
# Measure	Position Value Cutoff signal peptide?
max. C	65 0.498 0.52 NO
max. Y	65 0.197 0.33 NO
max. S	19 0.999 0.92 YES
mean S	1-64 0.568 0.49 YES
D	1-64 0.383 0.44 NO

# Most likely cleavage site between pos. 64 and 65: ACA-AC

Prediction: Signal peptide

Signal peptide probability: 1.000

Max cleavage site probability: 0.527 between pos. 21 and 22

No InterProScan hits reported for this clone.

**tBLASTx**: *Chlamydomonas reinhardtii* strain CC-503 cw92 mt+, NAD malic enzyme

Expect = 0.065, Identities = 12/22 (54%).

### Clone 30

Frame +3

ERRRMAEYLASPQGYDHVMHVVRARFMAGNYDLCAGVCRDFANTVGNFNIDRGV  
ADGHWTRPTRRRRHGIGLGLGVQVD

#### SignalP analysis

Sequence		length = 70		
# Measure	Position	Value	Cutoff	signal peptide?
max. C	23	0.540	0.52	YES
max. Y	36	0.261	0.32	NO
max. S	34	0.778	0.97	NO
mean S	1-35	0.352	0.51	NO
D	1-35	0.307	0.45	NO

Sequence

Prediction: Non-secretory protein

Signal peptide probability: 0.259

Max cleavage site probability: 0.228 between pos. 22 and 23

No interProScan hits reported for this clone.

**tBLASTx** : *Ralstonia solanacearum* strain IPO1609, transmembrane protein  
Expect = 0.044, Identities = 15/31 (48%).

### Clone 36

Frame +1

LIFLLGLDAPLTDIWKIGNNISTGFRNPTASEMYFSFEHPAGNWIPNPDLKAEQALNQSI  
YIQAHELLGSFGLTFYHTRYKNLLTEQESTYKKRNPYYNAYSASYGQQGVQVD

#### SignalP analysis

Sequence		length = 70		
# Measure	Position	Value	Cutoff	signal peptide?
max. C	24	0.588	0.52	YES
max. Y	47	0.258	0.32	NO
max. S	5	0.935	0.97	NO
mean S	1-46	0.664	0.51	YES
D	1-46	0.461	0.45	YES

# Most likely cleavage site between pos. 46 and 47: GGA-AG

Prediction: Signal peptide

Signal peptide probability: 0.739

Max cleavage site probability: 0.300 between pos. 23 and 24

**tBLASTx** : *Neisseria meningitidis* hmbR pseudogene, strain 30931  
Expect = 3e-26, Identities = 49/102 (48%).

SEQUENCE: <u>Sequence 1</u> CRC64: 40305AE8DF5A54A3 LENGTH: 109 aa		
<b>TonB-dependent receptor, beta-barrel</b>		
<a href="#">G3DSA:2.40.170.20</a>		no description
<b>unintegrated</b>		
SignalP		signal-peptide
<a href="#">SSF56935</a>		Porins

### Clone 39

frame +1, as dictated by tBLASTx.

> - 82 codons

VMAVHRMISLFPGEQQEEIRSQISQVLRAVICQRLLRWNKKFITIRDILLNTHAVANLIR  
TRKEPQIISIQETQLPMKTLEM

Sequence	length = 70
# Measure	Position Value Cutoff signal peptide?
max. C	45 0.470 0.52 NO
max. Y	45 0.521 0.32 YES
max. S	40 0.989 0.97 YES
mean S	1-44 0.622 0.51 YES
D	1-44 0.571 0.45 YES

# Most likely cleavage site between pos. 44 and 45: ACA-AC  
Prediction: Signal peptide

Signal peptide probability: 0.865

Max cleavage site probability: 0.566 between pos. 27 and 28

No InterProScan hits reported for this clone. No obvious structural similarity to the PilT protein.

**tBLASTx** : *G. kaustophilus* PilT-like protein, pili biogenesis. E=0.0001, 17/27 (62%) id.

## Clone 42

frame +2

> - 83 codons

LYLMTAKSSKTQTKKRASTKPAAKPTTRKSAKTQTQADNKVSQRLKAAKELQKNEEK  
KARPEHVVNLIINDALWLFGLVITIYL

### SignalP analysis

Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	22	0.673	0.52	YES
max. Y	22	0.184	0.33	NO
max. S	37	0.998	0.92	YES
mean S	1-21	0.943	0.49	YES
D	1-21	0.564	0.44	YES

# Most likely cleavage site between pos. 21 and 22: AAA-AT  
Prediction: Signal peptide  
Signal peptide probability: 1.000  
Max cleavage site probability: 0.317 between pos. 21 and 22

No InterProScan hits reported for this clone.

**tBLASTx** : *Neisseria meningitidis* 053442, cell division protein FtsK  
Expect = 1e-14, Identities = 36/45 (80%)

## Clone 44

frame +3

> - 79 codons

PHTVSASADNNALMTCWSRERIKSGDAWDNASPSRPPESDSGWRCASSVKSNDAIVR  
VRCSSARLVESSTSSNTLSRK

### SignalP analysis

>Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	34	0.661	0.52	YES
max. Y	34	0.538	0.32	YES
max. S	30	0.983	0.97	YES
mean S	1-33	0.603	0.51	YES
D	1-33	0.570	0.45	YES

# Most likely cleavage site between pos. 33 and 34: ACA-AC  
Prediction: Signal peptide  
Signal peptide probability: 0.960  
Max cleavage site probability: 0.679 between pos. 33 and 34

No InterProScan hits reported for this clone.

**tBLASTx** : *Rothia* sp. T40-1 Mef, aspartate/ornithine binding domain, IS30 transposase family protein, and tet(W) genes. Expect = 1e-13, Identities = 32/39 (82%).

## Clone 52

frame +3

> - 79 codons

IGIVKGGLAGFSTPSIDRWLSRLIDLVLGFPNMVIAIAFIGIMGPSITNVIISLCITKWAEY  
ALITRGLVVVEKVFYRH

>Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	18	0.277	0.52	NO
max. Y	18	0.318	0.33	NO
max. S	41	0.989	0.92	YES
mean S	1-17	0.943	0.49	YES
D	1-17	0.631	0.44	YES




# Most likely cleavage site between pos. 17 and 18: AAA-GG

Prediction: Signal peptide

Signal peptide probability: 0.535

Max cleavage site probability: 0.109 between pos. 31 and 32

**tBLASTx** : *Methylovorus* sp. SIP3-4, binding-protein-dependent transport systems inner membrane. Expect = 6e-14, Identities = 32/67 (47%)

SEQUENCE: <a href="#">Sequence 1</a> CRC64: EEB6487FFE22223F LENGTH: 79 aa	
<b>InterPro</b> <a href="#">IPR000515</a> Family	<b>Binding-protein-dependent transport systems inner membrane component</b> <a href="#">PF00528</a>  <a href="#">BPD_transp_1</a>
<b>InterPro</b> 	
noIPR unintegrated	<b>unintegrated</b> <a href="#">tmhmm</a>  <a href="#">transmembrane_regions</a>

## Clone 58

Frame +1

STLMIGMETDTVESIRQIPDIIIEIGVDVPRYNILTPYPGTPFYEQQLKAENRLLTRDWYY  
YDTETVVFQPKNMSPATLQEEFYKLWQDTFTYKRIFK

### SignalP analysis

Sequence length = 70

# Measure	Position	Value	Cutoff	signal peptide?
max. C	26	0.332	0.52	NO
max. Y	41	0.209	0.33	NO
max. S	12	0.984	0.92	YES
mean S	1-40	0.857	0.49	YES
D	1-40	0.533	0.44	YES

# Most likely cleavage site between pos. 40 and 41: AAA-GC

Prediction: Non-secretory protein

Signal peptide probability: 0.451

Max cleavage site probability: 0.128 between pos. 21 and 22

No InterProScan hits reported for this clone.

**tBLASTx** : *Leptotrichia buccalis* DSM 1135, Expect = 5e-51, Identities = 82/97 (84%),

## Clone 59

Frame +1

LILGRINYNNWFFELLAKFFAGVVLGIGAGLSLGREGPSVQLGSYVGYGASKILKTDTVE  
RNYLLTSGSSAGLSGAFGAPLAGVMFSIEEIHKYLSGKLLI

# Measure	Position	Value	Cutoff	signal peptide?
max. C	29	0.548	0.52	YES
max. Y	29	0.343	0.33	YES
max. S	19	0.994	0.92	YES
mean S	1-28	0.950	0.49	YES
D	1-28	0.646	0.44	YES



# Most likely cleavage site between pos. 28 and 29: ACA-AC

Prediction: Signal peptide

Signal peptide probability: 0.991

Max cleavage site probability: 0.411 between pos. 28 and 29

**tBLASTx** : *F. nucleatum* chloride channel protein. E=2e-46, id= 99/100 (99%)

SEQUENCE: <u>Sequence_1</u> CRC64: BF72F4CD53F6677A LENGTH: 100 aa		
<b>InterPro</b> <a href="#">IPR001807</a> Family 	<b>Chloride channel, voltage gated</b>	
	<a href="#">PR00762</a>	CLCHANNEL
	<a href="#">PTHR11689</a>	CHLORIDE CHANNEL
	<a href="#">PF00654</a>	Voltage_CLC
<b>InterPro</b> <a href="#">IPR014743</a> Domain 	<b>Chloride channel, core</b>	
	<a href="#">G3DSA:1.10.3080.10</a>	no description
	<a href="#">SSF81340</a>	Clc chloride channel
noIPR unintegrated	<b>unintegrated</b>	
	<a href="#">PTHR11689:SF2</a>	CHLORIDE CHANNEL PROTEIN 5 LONG ISOFORM
	SignalP	signal-peptide
	tmhmm	transmembrane_regions

## Clone 60

Frame +1

VILGLIFFLDTRLGQAYIATGDNSDMAKSFINTDRMELMGLVISNGIIALSGALMAQQ  
EGYADASRGIGVIV

### SignalP analysis




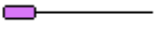
# Measure	Position	Value	Cutoff	signal peptide?
max. C	32	0.358	0.52	NO
max. Y	32	0.319	0.32	NO
max. S	29	0.952	0.97	NO
mean S	1-31	0.500	0.51	NO
D	1-31	0.410	0.45	NO

Prediction: Non-secretory protein

Signal peptide probability: 0.282

Max cleavage site probability: 0.131 between pos. 28 and 29

**tBLASTx:** *Streptococcus gordonii str. Challis* ABC transporter, permease protein. Expect = 6e-34, Identities = 63/73 (86%)

SEQUENCE: <u>Sequence 1</u> CRC64: 359E69B3784B3078 LENGTH: 73 aa	
<b>InterPro</b> <a href="#">IPR001851</a> Family  	<b>Bacterial inner-membrane translocator</b>  PF02653 <span style="float: right;">BPD_transp_2</span>
noIPR unintegrated	<b>unintegrated</b> SignalP  <span style="float: right;">signal-peptide</span>