FUNCTIONAL METAGENOMIC ANALYSIS OF THE HUMAN GUT MICROBIOME TO IDENTIFY NOVEL SALT TOLERANCE GENES

A Thesis Presented to the National University of Ireland Cork for the Degree of Doctor of Philosophy

by

Eamonn P. Culligan, M.Sc.

Alimentary Pharmabiotic Centre and School of Microbiology

University College Cork

February 2014

Head of Department: Professor Gerald F. Fitzgerald

Supervisors: Professor Colin Hill, Dr. Julian R. Marchesi, Dr. Roy D. Sleator

"As we train our senses and focus our disciplined imaginations on the vastness of space it seems to me that we are missing something close to home. There is more than the contempt of familiarity in the extraordinary circumstance that, in our eagerness to learn of life far beyond earth, we overlook the life on our own surface - on the surface of man himself."

- Theodor Rosebury, Life On Man, 1969.

TABLE OF CONTENTS

PAGE

DECLARATION

I hereby declare that the research presented in this thesis is my own work and effort and that it has not been submitted for any other degree, either at University College Cork or elsewhere. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

This work was completed under the guidance of Professor Colin Hill, Dr. Julian Marchesi and Dr. Roy Sleator at the Alimentary Pharmabiotic Centre and School of Microbiology, University College Cork.

Signature: ____

Date: ________________

Dedicated to the memory of my father, Noel Culligan

1952-2005

THESIS ABSTRACT

The ability to adapt to and respond to increases in external osmolarity is an important characteristic that enables bacteria to survive and proliferate in different environmental niches. When challenged with increased osmolarity, due to sodium chloride (NaCl) for example, bacteria elicit a phased response; firstly via uptake of potassium (K^+), which is known as the primary response. This primary response is followed by the secondary response which is characterised by the synthesis or uptake of compatible solutes (osmoprotectants). The overall osmotic stress response is much broader however, involving many diverse cellular systems and processes. These ancillary mechanisms are arguably more interesting and give a more complete view of the osmotic stress response. The aim of this thesis was to identify novel genetic loci from the human gut microbiota that confer increased tolerance to osmotic stress using a functional metagenomic approach. Functional metagenomics is a powerful tool that enables the identification of novel genes from as yet uncultured bacteria from diverse environments through cloning, heterologous expression and phenotypic identification of a desired trait. Functional metagenomics does not rely on any previous sequence information to known genes and can therefore enable the discovery of completely novel genes and assign functions to new or known genes.

Using a functional metagenomic approach, we have assigned a novel function to previously annotated genes; *murB*, *mazG* and *galE*, as well as a putative *brp*/*blh* family beta-carotene 15,15’-monooxygenase. Finally, we report the identification of a completely novel salt tolerance determinant with no current known homologues in the databases. Overall the genes identified originate from diverse taxonomic and phylogenetic groups commonly found in the human gastrointestinal (GI) tract, such as *Collinsella* and *Eggerthella*, *Akkermansia* and *Bacteroides* from the phyla *Actinobacteria*, *Verrucomicrobia* and *Bacteroidetes*, respectively. In addition, a number of the genes appear to have been acquired via lateral gene transfer and/or encoded on a prophage. To our knowledge, this thesis represents the first investigation to identify novel genes from the human gut microbiota involved in the bacterial osmotic stress response.

CHAPTER I

LITERATURE REVIEW

Metagenomics and Novel Gene Discovery: Promise and Potential for Novel Therapeutics

VIRULENCE. 2014 APR: 1;5(3): 399-412

Abstract

Metagenomics provides a means of assessing the total genetic pool of all the microbes in a particular environment, in a culture-independent manner. It has revealed unprecedented diversity in microbial community composition, which is further reflected in the encoded functional diversity of the genomes, a large proportion of which consist of novel genes. Herein, we review both sequence-based and functional metagenomic methods to uncover novel genes and outline some of the associated problems of each type of approach, as well as potential solutions. Furthermore, we discuss the potential for metagenomic bio-therapeutic discovery, with a particular focus on the human gut microbiome and finally, we outline how the discovery of novel genes may be used to create bio-engineered probiotics.

Keywords: metagenomics, functional metagenomics, novel gene discovery, gut microbiome, microbiota, bio-therapeutics, meta-biotechnology, bio-engineered probiotics

Abbreviations: DNA, deoxyribonucleic acid; bp, base pair; NGS, next generation sequencing; ATP, adenosine triphosphate; Gb, gigabase; CAZyme, carbohydrate-active enzyme; ORF, open reading frame; NCBI, National Centre for Biotechnology Information; PCR, polymerase chain reaction; PKS, polyketide synthase; KS$_{\beta,}$ ketosynthase beta/ chain length factor; MRSA, methicillin-resistant *Staphylococcus aureus*; VRE, vancomycin-resistant *Enterococcus*;

NRPS, non-ribosomal peptide synthetase; MHT, methyl halide transferase; HGT, horizontal gene transfer; TRACA, transposon-aided capture; 59-be, 59-base element; HMP, Human Microbiome Project; SSH, suppressive subtractive hybridisation; BAC, bacterial artificial chromosome; RBS, ribosome binding site; GFP, green fluorescent protein; Mb, megabases; SAM protein, S'-adenosylmethionine protein; MetaHit, Metagenomics of the Human Intestinal Tract; IBD, inflammatory bowel disease; STEC, shigatoxigenic-*E. coli*; ETEC, enterotoxigenic-*E. coli*; LPS, lipopolysaccharide; HIV, human immunodeficiency virus; CLA, conjugated linoleic acid; EGF, epidermal growth factor; CFU, colony forming unit; BSH, bile salt hydrolase; GABA, gamma-aminobutyric acid; GMO, genetically-modified organism; IL-10, interleukin 10.

Introduction

Metagenomics is a term that describes both a field of scientific research and a set of techniques that enables the culture-independent analysis of a microbial community in any environmental sample.[1] The field has grown exponentially in the last ten to fifteen years, allowing researchers unparalleled access to the "uncultured majority"[2] of our microbial counterparts in our environments and closer to home, on and in our own bodies. This review aims to provide a concise overview of the principal methods of metagenomic analysis and how metagenomics can be used for the discovery of novel genetic loci. We will also discuss the potential uses of such genes for the creation of bio-engineered probiotics, as well as the potential for the discovery of novel bio-therapeutics for human medicine. A wealth of information has been gained in the relatively short history of metagenomics and thus, it is timely to discuss the future directions of this ever expanding field of research.

It has been estimated that there are more than 10^{30} microbial cells on Earth; nine orders of magnitude greater than the number of known stars in the universe.[3, 4] It is estimated that a large proportion of microbes in many environments are, as yet, uncultured[5], while cultured prokaryotic representatives are overwhelmingly from just four phyla; the Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria.[6, 7] Metagenomics provides a means to access, characterise and quantify this untapped diversity of microbes and discover novel genes, metabolic pathways and important products with biotechnological, pharmaceutical and medical relevance.

Metagenomic analysis can take a sequence-based or a functional approach (or a combination of both) to study a complex microbial community, but always begins with the isolation of DNA from the environment of interest.

Sample DNA isolation

Careful and considered isolation of total metagenomic DNA is the key step to ensure the sample obtained is truly representative of the community as a whole and is of high-quality, high-purity and as unbiased as possible. Due to the obvious heterogeneity of environmental samples it is difficult to lyse cells of all species with equal efficacy. Harsh lysis methods may result in physical damage to DNA, whilst a more gentle approach is likely to result in an under-representation of DNA from more difficult to lyse Gram-positive and archaeal members of the community.[8] This is a particular issue with regard to recovery of DNA from low abundance species.

Different methods can be used to access the DNA, such as one or a combination of mechanical, enzymatic, chemical or temperature-based cell lysis. Extraction can either be direct or indirect. Direct extraction is carried out from the environmental sample of interest by cell lysis and subsequent separation of the DNA from sample particles and cell debris[9], while indirect extraction first involves the separation of cells from the environmental sample followed by subsequent lysis.[10, 11] There are numerous methods to extract metagenomic DNA that have been optimised to suit the nature of the particular sample in question and which take into consideration the physico-chemical conditions, microbial make-up and origin of the environmental sample.[11-14] For example, Salonen and co-workers[14] compared four different methods of metagenomic DNA extraction from a human faecal sample and found that the average DNA yield varied from approximately 40µg to 1000µg, depending on the method used, while variations of 10-20 fold between methods were observed for yields of archaeal DNA. Two methods were studied in further detail and a method using repeated bead-beating and sequential precipitation steps was deemed most suitable. Despite a lower initial DNA yield, it was noted that the effectiveness of cell breakage was the main factor affecting the final microbial composition and level of diversity.[14]

STRATEGIES TO IDENTIFY NOVEL GENES

Sequence-based metagenomics

Traditional metagenomic sequencing relied upon Sanger sequencing, which produces relatively long reads of greater than 700 base pairs (bp) with a low error rate.[15] However, the development of next-generation sequencing (NGS) technologies coupled with significant reductions in sequencing costs has resulted in a movement away from Sanger sequencing as the gold standard, primarily due to the relative high cost and intensive cloning required by this approach compared to NGS methods such as 454 and Illumina. One of the first large scale metagenomic projects employed Sanger sequencing to characterise the microbial population of the Sargasso Sea.[13] Over 1 billion bp of DNA was sequenced and more than 1.2 million novel genes identified, including over 700 that encoded bacterial rhodopsin proteins (proteorhodopsins). Proteorhodopsins are retinal-binding membrane proteins that function as light-activated proton pumps. At the time, proteorhodopsins had been recently discovered in bacteria using metagenomics and were previously thought to exist only in halophilic archaea (bacteriorhodopsins).[16] This discovery prompted a major rethink about carbon and energy flux in the world's oceans, as it became apparent that bacteria were able to harness light energy to generate ATP non-photosynthetically.[17]

Next generation sequencing (NGS) technologies

Ever improving sequencing technologies coupled with dramatically decreasing costs means that a “sequence everything” approach can be applied to diverse environmental samples, followed by assembly and annotation to gain a comprehensive picture of the abundance and encoded functional potential of the microbial community members. Some of the most widely used NGS platforms include the GS-FLX 454 pyrosequencer (Roche), MiSeq, HiSeq and Genome Analyser II platforms (Illumina), SOLiD system (Life Technologies/ Applied Biosystems), Ion Torrent and Ion Proton (Life Technologies) and the PacBio RS II (Pacific Biosciences). For detailed information and technical aspects of these technologies the reader is directed to comprehensive reviews elsewhere.[18-21]

Large-scale sequencing projects employing such technologies generate huge datasets that can consist of millions of gene sequences, and owing to the nature of metagenomes, a large proportion of these will be novel, but with no known function. This initial, untargeted sequencing approach can subsequently be used in combination with a more targeted approach which aims to identify a specific type of gene or gene family. It should be noted that we mean untargeted in relation to identifying specific genes initially, but focused rather on the overall community composition, phylogeny and/ or functional capacity. In addition, metagenome sequencing using NGS technologies followed by annotation, homology searches and phylogenetic clustering will uncover genes that are more divergent and more interesting than the consensus genes with known sequences used to design probes or primers for initial gene-focused studies.[22] For example, Hess and co-workers sequenced 286 gigabytes (Gb) of cow rumen metagenomic DNA using Illumina GAIIx and HiSeq 2000 platforms. They then searched for carbohydrate active enzyme (CAZymes) sequences bioinformatically and minimised the dependence on known CAZyme sequences by searching for individual functional domains of such enzymes rather than overall sequence similarity. This approach revealed 27,755 potential candidate genes from over 2.5 million predicted open reading frames (ORFs). Interestingly 24% of the genes were most similar to "hypothetical proteins" or "predicted proteins" in the non-redundant NCBI database. The sheer volume of candidate genes identified and the novelty of many sequences validates the initial high-throughput bioinformatic search. A sub-group of genes were subsequently amplified by PCR and the activity of their encoded proteins was assessed through functional heterologous expression.[23]

Gene targeting or Meta-gene analysis

Identification of novel genes for particular gene families or enzyme classes can be achieved by using DNA probes or PCR. Firstly a multiple sequence alignment of many sequences of the type of gene of interest is generated to identify the most conserved regions. Consensus primers are then designed and used to amplify target sequences from isolated metagenomic DNA. Subsequently, gene-specific primers can be designed and genome-walking can be carried out to retrieve the full gene sequence and those of neighbouring genes if desired.[24] This method is effective as large libraries do not need to be constructed and screened and novel genes can be identified irrespective of gene expression, which is a major hurdle to function-based screening (discussed later). Furthermore, full-length sequences are not required to identify genes initially, although they will have to be determined at a later stage.[25] This approach suffers from bias due to primer and probe design being limited to known sequences of similar gene family members as well as prejudices due to over-representation of dominant species in the metagenomic library. This will lead to PCR bias and less chance of amplification of genes from less abundant community members. Furthermore, it is unlikely that functionally homologous genes that have arisen due to convergent evolution will be identified.[26] While a significant limitation of this approach is its inability to uncover completely novel, non-homologous genes, this approach has nonetheless been used successfully to identify genes that encode biotechnologically and industrially relevant enzymes such as esterases, lipases and oxidoreductases.[9, 27, 28]

Of therapeutic relevance, a number of sequence-based screens have focused on identifying novel antimicrobial compounds. Type I and type II polyketide synthases (PKS) which are encoded on operons for biosynthesis of antibiotics and anti-cancer compounds are often the target genes. In one study, a soil metagenomic library was screened by PCR for type I PKS genes using primers that target regions flanking the highly conserved active site of the ketoacyl synthetase domain. Eleven novel type I PKS sequences were identified, with amino acid identity to known sequences ranging from 46-61%.[29] Feng and co-workers used degenerate primers to amplify ketosynthase beta chain length factor (KS$_{\beta}$) genes from a metagenomic library and used the amplicons as probes to identify PKS containing clones. Functional analysis of the clones revealed both known and unknown antimicrobial compounds, some of which had potent activity against methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin resistant *Enterococcus* (VRE).[30] In a recent extensive study, conserved domains within non-ribosomal peptide synthetase (NRPS) and PKS genes were used as targets to screen a >15 million member cosmid mega-library. Multiplex sequencing of PCR products, followed by clustering and BLAST analyses to identify sequences of most interest, revealed 18 gene clusters predicted to encode novel glycopeptide- and lipopeptide-like antibiotics, as well as anti-cancer and immunosuppressive compounds.[31] A PCR based approach was used to identify genes encoding OxyC enzymes, which are often found on biosynthetic gene clusters for glycopeptide antibiotics such as vancomycin and teicoplanin. In a twist on the previous examples, modifying enzymes found on these gene clusters were used to create 15 new sulphated glycopeptide antibiotic derivatives.[32, 33] Development of analogues, which could substitute for last-line antibiotics such as vancomycin, is an important research finding. Finally, Iwai *et al*, used a number of sets of degenerate primers for PCR, coupled with 454 pyrosequencing to identify genes encoding aromatic dioxygenases. This approach provides a more comprehensive view of the diversity of sequences present in the environment and this sequence information can be used to design probes to recover full-length genes subsequently.[34]

Data mining and "synthetic metagenomics"

An innovative approach to identify novel genes has been described in recent years, which involves mining existing sequence databases and/ or metagenomic datasets for sequences of interest, followed by chemical synthesis of selected genes.[35] The authors name this process "synthetic metagenomics" and successfully applied it to identify novel methyl halide transferase (MHT) enzymes, which are important in agriculture and industrial applications for more efficient production of biofuels. Eighty nine putative MHTs were identified, with amino acid identities to a known MHT as low as 18% and an average of 28% amino acid identity between sequences. The genes included 61 bacterial, 13 fungal, 1 archaeal and 14 from plants. The genes were codon optimised for heterologous expression in *Escherichia coli* and yeast cells and then chemically synthesised. Only 6% of the synthesised genes showed no MHT activity, which is remarkable considering only one was actually annotated as a MHT and only 55% were annotated as generic methyltransferases.[35] A similar study has also applied this approach to glycoside hydrolases.[36] Both demonstrate the utility of "synthetic metagenomics", which of course could be applied to any gene of interest.

Plasmid and integron capture

Horizontal gene transfer (HGT), mediated by mobile genetic elements such as plasmids and integrons play a crucial role in bacterial evolution, adaptation and survival.[37] Plasmids are likely to contain genes necessary for niche colonisation and encode functions important in that environment. Furthermore, genes encoding antibiotic resistance, virulence and antimicrobial production can often be found on plasmids. Developed to capture plasmids from the human gut mobile metagenome, the TRACA (transposon-aided capture) method identified a number of novel plasmids of both Gram-positive and Gram-negative origin.[38, 39] TRACA could be applied to any environment once sufficient quantities of DNA can be isolated and would provide a valuable means to identify novel mobile genes within specific environments. Indeed, TRACA has been recently used to identify plasmids from bacteria in the human oral cavity and activated sludge.[40, 41]

Integrons are gene recombination and expression systems. They encode an integrase protein and contain a integration site (*attI*) to capture genes from what are known as mobile gene cassettes which contain a conserved site-specific recombination sequence called a 59-be (base element).[42] By targeting conserved sequences associated with integron-gene cassettes using PCR one can retrieve full-length, novel genes from metagenomes in the absence of any other sequence information.[43] The overwhelming majority of integron-associated genes lack known homologues or cannot be assigned a function. A recent study found that 85% of *Treponema*-associated integron genes from the Human

Microbiome Project (HMP) dataset were of unknown function.[44] Furthermore, proteins with novel structure and folds have been identified among integron-gene cassettes which belong to as yet uncharacterised protein families.[45]

Suppressive Subtractive hybridisation

Suppressive subtractive hybridisation (SSH) is a PCR-based approach that can be used to compare and identify differences in the genomes of closely related bacterial species or differentially expressed mRNA transcripts within a cell.[46, 47] Differences between “driver” (control) and “tester” sequences are highlighted by the removal of sequences or transcripts of similar abundance. This approach has been used to identify novel sequences from complex metagenomic samples. Ninety-six bacterial and archaeal DNA fragments were identified from a rumen metagenome using this approach, of which 39 had no significant similarity to any database sequences.[48]

More recently, SSH has been adapted to recover full-length, novel genes from metagenomes.[49] Degenerate primers are first used to generate multiple target-gene amplicons (“driver DNA”), which are then used as hybridisation probes to capture full-length genes. “Driver DNA” is biotinylated and immobilised on streptavidin-coated magnetic beads. Metagenomic DNA (“tester DNA”) is then added to the coated beads to capture full-length genes of interest. This approach has greater specificity and less non-specific amplification compared to other PCR-based techniques, but does require the creation of a genomic DNA library. Nevertheless, this approach can retrieve numerous gene targets in a single reaction.[49]

Functional metagenomics

The second main area of metagenomic analysis is functional metagenomics. The oft-stated, broad aims of metagenomic analyses are to discover “who is there?” and “what are they doing?” Large scale NGS projects can tell us about the former, but is limited with regard to the latter, due to the dependence on known sequences already in the databases. Functional metagenomics on the other hand suffers no such reliance on previous sequence information and is a powerful approach with the potential to discover new functions for known genes and also, completely novel genes, gene families and their encoded proteins.

Functional metagenomics requires the creation of a metagenomic library containing fragments of community metagenomic DNA. The process follows the same general steps for the creation of a genomic DNA library; enzymatic digestion or random shearing of DNA, ligation to a suitable vector and transformation to a heterologous host.[50] One of the biggest technical challenges is constructing a library with sufficient coverage of the community metagenome and requires an extremely large number of clones, especially for complex environments. This is compounded further when one considers a library would need to contain 100- to 1,000-fold coverage in order to possess a significant proportion of clones carrying sequences form rare (<1%) community species.[51]

Nonetheless, it is achievable; cosmid megalibraries have been created from soil which contain >1.5 x 10^7 unique clones. At clone densities such as this, it is believed the library approaches saturation.[31]

Insert size and vector

Two types of library can be constructed depending on the type of screen to be performed and the desired library size. Libraries can be small-insert libraries, created using plasmids and containing inserts of <10kb, or large-insert libraries, created using fosmid (25-45kb inserts) or cosmid (15-40kb inserts) vectors or BAC (bacterial artificial chromosome; 100-200kb inserts) vectors.[52] Small insert libraries are maintained on high copy number plasmids with strong promoters, which are usually used in activity screens where a single gene is responsible for the activity. Large insert libraries are suitable for the identification of multi-gene encoded products, operons and entire biochemical pathways and usually utilise low-copy number or inducible vectors. Inducible vectors are advantageous in that the library may be stably maintained at low copy, but can be induced to high-copy for downstream applications such as transposon mutagenesis, as well as DNA extraction, cloning and sequencing. Large inserts also facilitate gene neighbourhood analysis and increase the probability of phylogenetic and taxonomic assignment of sequences.[53, 54]

Cloning host

To date, *E. coli* has been the cloning host of choice for the vast majority of metagenomic projects. *E. coli* possesses a number of desirable attributes that make it the host of choice; in depth knowledge of its physiology and biochemistry following decades of intensive research being primary among them. Furthermore, *E. coli*, (i) has a high transformation efficiency, (ii) is somewhat promiscuous with regard to the diversity of foreign expression signals it recognises, (iii) lacks genes for restriction modification and homologous recombination, and (iv) is capable of translating mRNA with diverse translation signals because the normal translational dependency on the degree of complementarity of the 3' terminus of the 16S RNA and the Shine-Dalgarno sequence, is not as strict in *E. coli*, which is more promiscuous with regard to complementary sequences recognised.[55, 56] Despite these advantages, *E. coli*, like any expression host, is unable to express all foreign DNA due to differences in the transcriptional, translational and post-translational machinery of the originating organism. For example, *cis*-acting factors such as a promoter and a ribosome binding site (RBS) compatible with the host machinery, as well as factors that may need to be supplied *in trans* by the host cell, such as chaperones, transcription factors or a compatible secretion system, are all potentially negative effectors of efficient expression.

Mathematical formulae have been developed to predict the chance of a given gene of interest being expressed in *E. coli*. Thirty-two sequenced bacterial genomes from three different phyla (Actinobacteria, Firmicutes and

Proteobacteria), as well as 10 Archaeal genomes were used in the analysis. Approximately 40% of genes in total were predicted to be functional in *E. coli*, but this ranged from 7% of Actinobacterial genes, to 73% of genes from *Firmicutes*.[57]

Hit rates

The hit rates (i.e. the chances of identifying a gene of interest through functional metagenomic expression) are traditionally quite low and can vary widely. Several factors influence the hit rate such as the source of the metagenomic DNA, the size of the gene of interest, its abundance in the metagenome and consequently the library, the vector system and host of choice, the screen itself and the ability of the host to successfully express the gene.[58] Hit rates have been shown to range from one positive hit per 2.7 megabases (Mb) of DNA screened to one per 3979.5 Mb.[58-60]

Strategies to improve heterologous expression

(a) Alternative or dual hosts

The use of different cloning hosts or the use of one host to maintain the library followed by transfer of the library to a different host (usually from a different Phylum or genera) for screening have been shown to be successful, whist demonstrating the “different host, different hit” effect.[55] Considering the environment of study in choosing a suitable host can increase the success of a functional metagenomic screen. *Streptomyces* species are commonly found in soil, are genetically amenable and produce a diverse range of medically relevant secondary metabolites such as antibiotics. Thus, numerous studies have employed *Streptomyces* species as screening hosts. McMahon and co-workers modified an integrative *E. coli* -*Streptomyces* cosmid vector to increase its recoverability and used it to create a cosmid metagenomic library. Initially the library was created in *E. coli* and then transferred via conjugation to a *Streptomyces lividans* mutant strain, which was defective in pigmented antibiotic production (increasing the ability to identify active clones). Screens for haemolytic activity and the production of secondary metabolites and pigments identified twelve biologically active clones due to functional expression of metagenomic DNA. The key observation however, is that none of the phenotypes were detectable in *E. coli*, demonstrating the utility of using a different host system.[61] A functional metagenomic screen to identify antibiotic production, altered colony morphology and pigmentation identified a number of clones expressing the desired traits using a cosmid library maintained in six different Proteobacterial hosts (*E.*coli, *Pseudomonas putida*, *Burkholderia graminis*, *Caulobacter vibrioides*, *Ralstonia metallidurans* and *Agrobacterium tumefaciens*). Interestingly, there was little overlap between the active clones, revealing substantial inter-Phylum variations in gene expression, even among more closely related species.[62] Other alternative hosts include *Bacillus subtilis*, *Rhizobium leguminosarum* and a dual *E. coli* - *Thermus thermophilus* system for screening extremophilic microorganisms.[63-65] Development of different hosts and molecular tools to make them genetically malleable will expand the repertoire and diversity of novel genes that can be recognised and heterologously expressed.

(b) Modified vectors

The use of alternative hosts is often coupled with the use or creation of novel expression vectors that can function in multiple hosts or maximise the chances of expression in said hosts. One example of a broad host-range fosmid and BAC vector is pRS44. Created for metagenomic screening, pRS44 contains two origins of replication, can be induced from low to high-copy with L-arabinose and contains an origin of transfer to allow conjugation to additional hosts, as well as a stabilisation element (*parDE*). A 20,000 member clone library with insert sizes of up to 200kb was successfully created and subsequently transferred to two other putative host organisms, *Pseudomonas fluorescens* and *Xanthomonas campestris*.[66] Similarly, Kakirde *et al*, (2010) created a BAC vector capable of replication and expression in diverse genera of Gram-negative bacteria such as *Escherichia*, *Salmonella*, *Enterobacter*, *Pseudomonas*, *Serratia* and *Vibrio*.[67] Recently, viral gene expression elements have been used to increase the expression and hit rate of metagenomic clones. Using phage T7 RNA polymerase to drive transcription and an inducible phage anti-termination protein to bypass many transcriptional terminators present on the insert, a 6-fold increase in the number of carbenicillin resistant clones identified in the screen was observed.[68] Furthermore, the use of a vector with dual-orientation promoters allows for bi-directional transcription and the possibility to significantly increase the hit rate. Such expression does not rely on the presence of a native, insert-borne promoter, or on the orientation of the cloned insert.[59]

(c) Codon optimisation

The majority of organisms have a particular preference for certain translation initiation codons and for overall codon usage, which is known as codon usage bias.[69] *E. coli* for example uses AUG as a start codon to initiate approximately 90% of translation, therefore non-AUG start codons such as GUG and UUG may not be efficiently recognised.[58] Foreign genes highly transcribed in *E. coli* have been shown to possess similar promoter sequences that bind the sigma factor RpoD (σ^{70}) of *E. coli*.[70] The effect of codon usage bias on expression has been investigated by introducing synonymous mutations to the third base position of the gene encoding green fluorescent protein (GFP). Expression levels between the 154 protein variants differed by up to 250-fold. Variable expression however, did not correlate with codon bias but rather, with the stability of mRNA folding near the ribosomal binding site. Codon bias on the other hand influenced the overall translation efficiency and cellular fitness.[71] Interestingly a recent study suggests (on the basis of analysing eleven diverse metagenomes) that regardless of phylogeny, microbes in the same environmental niche share common preferences for synonymous codon usage at the community level. These codon usage signatures differ between different metagenomes and can be used to predict functionally relevant genes for the community as a whole and also identify and characterise unknown genes with similar codon usage patterns to those genes.[72]

Knowledge of such issues and data mining large metagenomic datasets will enable the design of the most suitable expression systems with the greatest chance of success and also enable the creation of novel metagenome-derived synthetic genes or operons that are optimised for expression in *E. coli* or another relevant host.[35, 36] An overview of the main methods to identify novel genes through metagenomics is presented in Figure 1.

Figure 1.

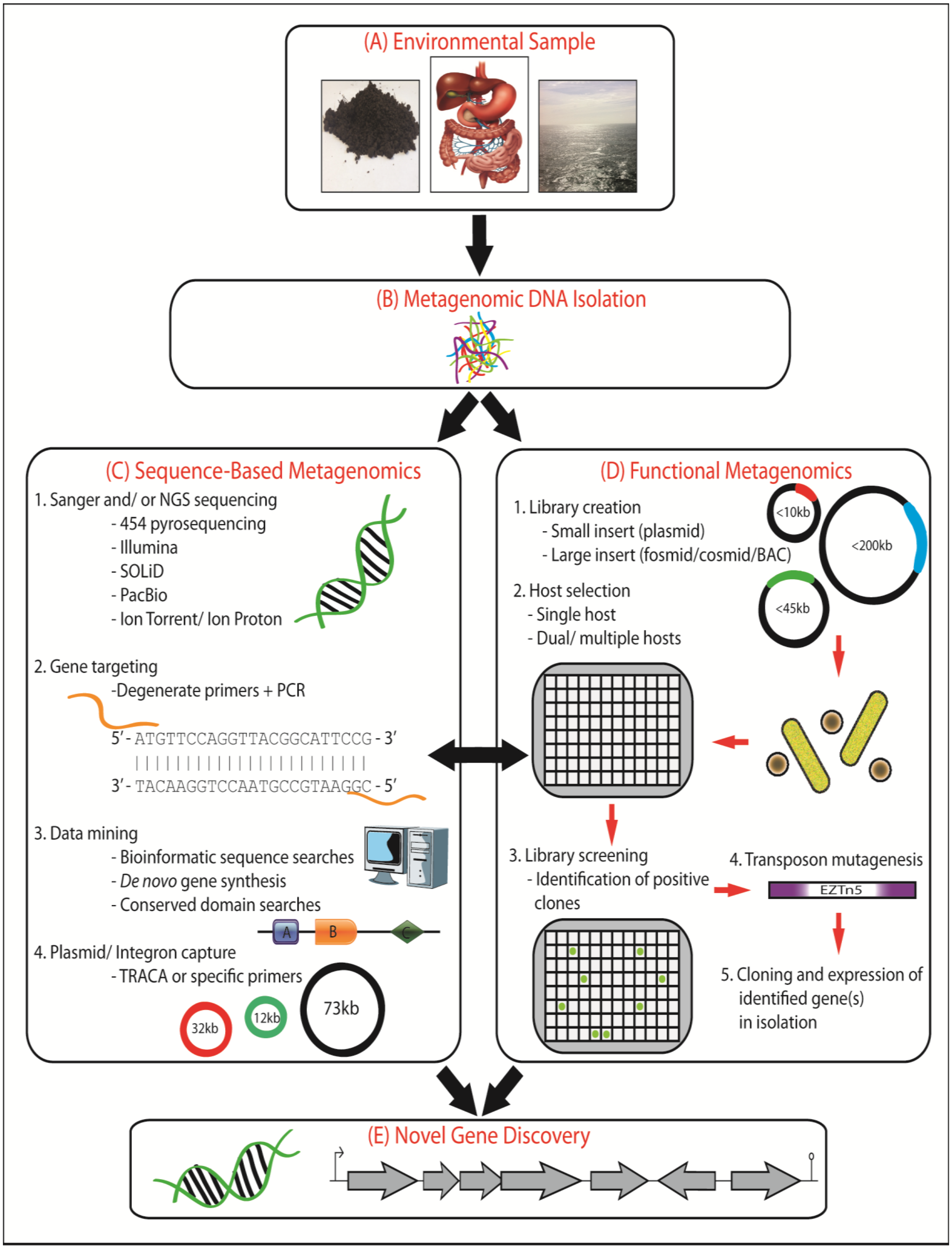

(A) Environmental Sample
(B) Metagenomic DNA Isolation
(C) Sequence-Based Metagenomics
1. Sanger and/ or NGS sequencing
- 454 pyrosequencing
- Illumina
- SOLiD
- PacBio
- Ion Torrent/ Ion Proton
2. Gene targeting
-Degenerate primers + PCR
5′- ATGTTCCAGGTTACGGCATTCCG - 3′
3′- TACAAGGTCCAATGCCGTAAGGC - 5′
3. Data mining
- Bioinformatic sequence searches
- *De novo* gene synthesis
- Conserved domain searches
A
B
C
4. Plasmid/ Integron capture
- TRACA or specific primers
32kb
12kb
73kb
(D) Functional Metagenomics
1. Library creation
- Small insert (plasmid)
- Large insert (fosmid/cosmid/BAC)
<10kb
<200kb
<45kb
2. Host selection
- Single host
- Dual/ multiple hosts
3. Library screening
- Identification of positive clones
4. Transposon mutagenesis
EZTn5
5. Cloning and expression of identified gene(s) in isolation
(E) Novel Gene Discovery

Figure 1. An overview of the main methods of novel gene discovery using metagenomics. A sample from the environment of interest (e.g. soil, human gastrointestinal tract, ocean) is collected (A), total metagenomic DNA is isolated directly or indirectly from the sample either by harsh or gentle lysis (B); metagenomic DNA is subsequently subjected to sequence-based analysis (C), which can include; creating a clone library followed by random Sanger sequencing or direct sequencing of metagenomic DNA using next generation (NGS) methods (C-1); Gene targeting using degenerate primers designed from conserved regions of homologous sequences followed by PCR amplification (C-2); Data mining of existing metagenomic sequence datasets for genes or conserved domains of interest, followed by complete synthesis of host-optimised genes (C-3); Capture of novel genes from plasmids using the TRACA (transposon-aided capture) method or integron-associated genes using PCR with primers that target conserved integration site sequences (C-4). Functional metagenomics (D) involves; The creation of a small- or large-insert library using a specific type of vector (plasmid, fosmid, cosmid or bacterial artificial chromosome (BAC)) depending on the needs of the user or aims of the project (D-1); Choosing a suitable heterologous host for expression of metagenomic DNA. The most widely used host is *E. coli*, but species of *Streptomyces* and *Bacillus*, for example, have also been utilised. The library is maintained in 96- or 384-well micro-titre plates and stored at -80ºC (D-2); Clones containing metagenomic DNA can be screened to identify a phenotype of interest by replicating all or a portion of the library into new micro-titre plates or onto agar

plates. Growth of library clones can be assessed for production of antimicrobials, pigmentation, altered morphology or increased resistance to various stressors relative to control strains carrying empty vectors (D-3). Transposon mutagenesis may be carried out on positive clones to identify the gene or genes of interest (D-4), which may be subsequently cloned as an isolated gene(s) to confirm the phenotype (D-5). Novel genes are discovered using sequence-based or functional metagenomic strategies, or a combination of both (E).

Functional metagenomic discovery of therapeutically relevant compounds

The majority of microbes in many environments are as yet uncultured and many will harbour the ability to produce therapeutically-relevant novel compounds. By utilising one or more of the strategies mentioned above to increase the chances of successful heterologous expression, there is potential to discover such novel compounds from uncultured members of well-known producers such as the Actinomycetes and also from novel microbial species. Nevertheless, novel compounds have been successfully discovered using functional metagenomics. The antibiotics Turbomycin A and B were discovered, somewhat serendipitously, following a functional screen for haemolytic clones in an *E. coli* metagenomic library. Turbomycin A had been reported previously as a fungal metabolite, but not bacterial, while Turbomycin B was a novel compound. The production of both antibiotics is proposed to occur via a single gene on the insert coupled with endogenous indole production from the *E. coli* host. Both antibiotics displayed broad-spectrum activity against a number of Gram-positive and Gram-negative pathogens.[73]

Gene clusters for the biosynthesis of novel compounds have often been found to be silent in both culturable organisms and also when cloned and expressed in a heterologous host. To overcome this, clones found to harbour PKS genes were subsequently screened for the presence of putative transcription factors in the neighbouring DNA, which usually tightly regulate such gene clusters. By cloning the genes encoding the transcription factors (under the control of the strong, constitutive promoter of the erythromycin resistance gene *ermE*) and subsequently transforming the corresponding PKS clone with the expression plasmid, a dramatic increase in expression was observed and a clone producing an antibacterial metabolite was identified. The compound, Tetarimycin A, is a Gram-positive specific antibiotic with potent activity against MRSA. The authors suggest that there are likely to be more members of this class of novel antibiotic encoded by similar gene clusters present in the environment.[74] A combination of sequence-based metagenomics and heterologous expression using a transcription factor (similar to the example above) has identified a compound with inhibitory activity against disease-related protein kinases, while another study identified a compound with potent anti-tumour activity.[75, 76]

Therapeutics from within: "bio-prospecting" the human microbiome

A microbiome is the totality of all microbes, their genes and interactions in a particular environment. The sequencing of the human genome was a historic scientific achievement and heralded exciting opportunities to decipher many of the intricacies of human diseases and conditions. It soon became apparent however, that studying the human genome and organism as a single entity gave an incomplete picture without considering the role of the human microbiome.[77, 78] Whilst the previous examples have dealt with the identification of novel genes that encode secondary metabolites such as antibiotics, the majority of these have been identified in the external environment. Large-scale metagenomic research projects, such as Metagenomics of the Human Intestinal

Tract (MetaHit), the Human Microbiome Project (HMP) and ELDERMET amongst others[79-83] have focused on looking within the human body, not only for novel genes and compounds, but also for novel species and for changes in the abundance and diversity of extant species of the human microbiome and how these may be “mined” or manipulated in a positive way for human health.

Thuricin CD was discovered from a screen of more than 30, 000 colonies from human faecal samples. Produced by *Bacillus thuringiensis*, thuricin CD is a novel bacteriocin with narrow-spectrum, potent antimicrobial activity against clinical isolates of *Clostridium difficile*. Importantly thuricin is as effective as the clinically-indicated antibiotics, vancomycin and metronidazole, in killing *C. difficile* and causes negligible “collateral damage” to other members of the gut microbiota in an *ex vivo* colon model, making it an attractive therapeutic compound to pursue for clinical application. While thuricin was not identified using a metagenomics approach, its discovery and its desirable properties indicate that applying sequence-based or functional metagenomic approaches (or a combination of both) could identify similar and also other novel antimicrobials.[84, 85] Indeed, similar to the *in silico* data-mining approach mentioned previously, gene clusters similar to that of thuiricin have been identified in genomic and metagenomic sequence datasets.[86] Using the amino acid sequences of the radical S’-adenosylmethionine (SAM) proteins (TrnC and TrnD) encoded on the thuricin gene cluster as driver sequences, 100 TrnC and 54 TrnD homologues were identified in the genomes of 112 unique microbial strains, none of which had been previously associated with bacteriocin production. Further analysis revealed 15 novel, putative thuricin gene clusters. In addition, 365 TrnC and 151 TrnD homologues were found in metagenomic datasets from diverse environments indicating the widespread presence of such genes in the environment. At the time of publication in 2011, the HMP dataset would have been unavailable, but it would be interesting to investigate the abundance of such gene clusters in the different anatomical human microbiomes, particularly the gut microbiome.

Discovery of novel antimicrobials has become of paramount importance due to significant increases in global antibiotic resistance and the emergence of multi-antibiotic resistant “superbugs”.[87] In a recently published annual report on infection and antimicrobial resistance, England’s’ Chief Medical Officer, Professor Dame Sally Davies, highlights the seriousness of the problem, makes a number of recommendations and states that, *“it (antimicrobial resistance) should be placed on the national risk register (specifically, the National Security Risk Assessment) and the Government should campaign for it to be given higher priority internationally, including collaborations to ensure the development of new antimicrobials and vaccines, such as Private Public Partnerships.”* [88] The full report can be downloaded at: https://www.gov.uk/government/publications/chief-medical-officer-annual-report-volume-2. Furthermore, overall societal costs due to antimicrobial resistant infections in the European Union, Iceland and Norway is estimated to be approximately €1.5 billion annually.[89] Research to identify novel antimicrobials or targets thereof has been scaled down or completely abandoned by some pharmaceutical companies.[90] Payne and co-workers give an insight from the perspective of a pharmaceutical company of the difficulties and costs associated with such research and the poor returns, both financial and of novel products.[91] This is compounded by the fact that only two new classes of antibiotics have been marketed since 1962.[92] Overall, these facts emphasise the need for novel and alternative strategies to combat antimicrobial resistance.

In a recent study, *Eggerthella lenta*, a common actinobacterial member of the gut microbiota, was shown to be solely responsible for the inactivation of the cardiac drug digoxin.[93] A highly up-regulated, two-gene (*cgr1* and *2*) operon was likely responsible for reducing digoxin to an inactive form, while arginine inhibited expression and strains lacking the operon were unable to reduce digoxin. Germ-free mice colonised with *E. lenta* had lower serum and urine digoxin compared to mice colonised with *E. lenta cgr*-negative strains. Furthermore, mice fed a high protein diet had increased serum and urine digoxin indicating a reduction in digoxin inactivation, likely through increased dietary arginine.[93]

Such studies open up new avenues for future therapeutic strategies. Profiling faecal samples for *E. lenta* in patients requiring digoxin could result in treatment interventions such as dietary supplementation with increased protein or arginine. The continual reduction in sequencing costs may see metagenomic sequencing become part of a personalised medicine treatment strategy or to ascertain the microbiome composition of subjects for clinical trials or before, during and after a particular drug regimen. Bacterial beta-glucuronidases of the human gut microbiota can reactivate CPT-11 (a chemotherapeutic agent used to treat colon cancer) in the human gut, causing severe diarrhea and limiting the dose intensification. Inhibitors of these beta-glucuronidases have been identified that could increase therapeutic efficacy.[94] The opposite of the *E. lenta* example could also be exploited; identification of a novel probiotic species that positively correlates with drug metabolism (such as activation of a pro-drug) or detoxification that could be co-administered as a probiotic with the drug. Metagenomics can increase our understanding of the structure, function, diversity and interactions of our microbiota which will aid the development of personalised medicine programs and the identification of novel drug targets and compounds. Furthermore, metagenomics in combination with *in vitro* assays and animal models may be used to identify specific microbes or combinations of microbes responsible for different drug interactions.[95] However, the inter-individual variability of the microbiome, even between healthy subjects remains a challenge for targeted therapies.[96, 97]

Bio-engineered probiotics

"Let food be thy medicine and medicine be thy food" – Hippocrates

There has been an increasing interest in the development of bio-engineered (also termed designer or recombinant) probiotics in recent years.[98-106] Probiotics are defined as "live microorganisms which when administered in adequate amounts confer a health benefit on the host." [107] Numerous research groups have engineered probiotic strains that successfully target specific pathogens and toxins, as well as developing probiotics as vaccine and drug delivery platforms.

Paton's group has used host-cell receptor mimicry to bio-engineer probiotics that bind and neutralise toxins produced by shigatoxigenic- and enterotoxigenic-*E. coli* (STEC and ETEC, respectively), as well as *Vibrio cholerae.* Recombinant bacteria expressing different oligosaccharide toxin receptor mimics in their lipopolysaccharide (LPS) outer core were engineered and subsequently shown to bind the toxins (shiga toxin and cholera toxin). This proved to be an extremely successful approach with up to 100% survival rates in murine models infected with virulent STEC or *V. cholerae* compared to controls.[108-110] A key advantage of this approach is that development of resistance is very unlikely as a selective pressure is not imposed on the pathogen.[111] Identification of genes that encode specific glycosyl transferases and sugar precursors are required to create different structural receptor mimics. Metagenomics may be useful to identify novel variants of such genes for the construction of diverse mimics that can target a broader range of pathogens and toxins. Other recombinant probiotics include those that exert immunomodulatory effects and target cancer and HIV (for a review, see Paton, Morona and Paton[112]), that modulate the fatty acid composition of host adipose tissue by producing *trans-10, cis12* conjugated linoleic acid (CLA), and that promote *in vitro* intestinal wound healing mediated by probiotic secretion of human epidermal growth factor (EGF).[113, 114]

The above examples demonstrate the diverse therapeutic potential of bio-engineered probiotics, which could be made even more diverse through the use of food-grade lactococci and lactobacilli for receptor mimicry[111] and the identification and cloning of novel, metagenome-derived genes in probiotic microbes.

Meta-biotechnology: application of metagenome-derived genes to bio-engineered probiotics

We have previously coined the term "meta-biotechnology" to describe the use of metagenomics as a robust method of identifying novel genes for use in biotechnology and the creation of bio-engineered probiotics.[115, 116] This is an extension of the patho-biotechnology concept which aims to exploit stress response, host evasion and virulence determinants of pathogenic microbes to enhance both physiological and technological robustness and stress resistance, as well as the clinical efficacy of probiotics.[101, 105, 106, 117] An increased ability to overcome the many host-imposed stresses of the gastrointestinal tract (low pH, bile, low iron concentrations and increased osmolarity) is a desirable trait in what are sometimes promising, but physiologically or technologically fragile probiotic strains.[118]

Listeria monocytogenes has been used as the model organism to successfully demonstrate proof-of-concept for patho-biotechnology. The *betL* gene (which encodes a betaine transporter involved in osmotic stress resistance and virulence[119-122]) was cloned in *Lactobacillus salivarius* UCC118, resulting in a significant increase in tolerance to numerous *in vitro* and technological stress conditions.[123] Furthermore, when expressed in *Bifidobacterium breve* UCC2003, *betL* also increased low pH and osmotic stress resistance *in vitro*, while also increasing gastrointestinal persistence and protecting against *L. monocytogenes* infection in a murine model.[124] Using a similar approach, the bile exclusion system (BilE) of *L. monocytogenes* was used to increase the bile tolerance and gastrointestinal persistence of both *L. lactis* NZ9000 and *B. breve* UCC2003, while *B. breve* alone reduced the listerial load in murine livers.[100]

The identification of similar genes to those mentioned above from the symbiotic gut microbiota using metagenomics could provide a large suite of genes for creating novel therapeutic probiotics and concerns regarding the use of genes from pathogens may be somewhat alleviated by these “self” genes. It is also likely that certain genes from the permanent inhabitants of the human gut microbiota will provide more robust protection against various GI stresses, as gut microbes will be better adapted for survival in this niche. Examples of genes discovered through metagenomics which could be used to create bio-engineered probiotics are outlined below.

The ability to colonise the gastrointestinal tract is one of a number of important characteristics for a probiotic strain to possess.[125] A functional metagenomic approach has been used to identify novel genes which conferred enhanced colonisation potential when expressed in *E. coli*.[126] An initial screen of a BAC library created from the murine gut metagenome identified a number of clones with enhanced ability to form biofilms. Further analysis of two clones in detail revealed four- and three-gene operons, respectively responsible for the phenotype. The operons were cloned in isolation and expressed in *E. coli*. The clones exhibited between 4- and 17-fold enhanced adherence to HT-29 human cell lines *in vitro*, as well as enhanced intestinal colonisation of antibiotic pre-treated mice, with approximately 6- to 28-fold more colony forming units (CFUs) recovered from plated intestinal homogenates.[126] Such genes could be especially useful for lactococcal-based engineered probiotics; while these strains are able to reach the intestine, they cannot usually colonise.[127]

The gastric juice in the stomach is the first major hurdle to overcome for potential probiotics. The pH of gastric juice ranges from pH 3 to pH 5 , but can be as low as pH 1.5 during fasting.[128] The ability to survive this initial challenge significantly influences microbial numbers that reach the intestine. Metagenomic analysis of extreme environments has proved fruitful in identifying novel stress resistance genes. Fifteen novel acid resistance genes from eleven clones were identified in a small-insert *E. coli* metagenomic library from the extremely acidic Tinto River.[129] Some clones were conferred with extreme acid resistance to pH 1.8, with up to 53% survival after 60 minutes treatment. The various genes responsible for the phenotype were both functionally and phylogenetically diverse and some could confer broad-host range acid resistance to *Pseudomonas putida* and *B. subtilis*, in addition to *E. coli*. Increasing the acid resistance of probiotics thus has the potential to enable greater persistence and enhanced therapeutic effect.

Recent work in our laboratory has identified novel salt tolerance genes from the human gut microbiota using functional metagenomics.[130] From a screen of over 20,000 clones and subsequent transposon mutagenesis and cloning, five genes were identified that could confer increased salt tolerance to *E. coli*. The ability to respond and adapt to increases in external osmolarity caused by salt and other solutes is an important survival determinant.[131] Like the examples with the *betL* salt tolerance locus mentioned above,[123, 124] these metagenome-derived genes could be used in a similar manner to increase the stress resistance of probiotic strains.

Bile resistance is also a desirable trait in a probiotic.[125] In addition to bile exclusion systems, some microorganisms possess bile salt hydrolase (BSH) genes which convert conjugated bile acids into their unconjugated form.[132] Bile acids, and consequently BSH, can influence host physiology in the form of lipid absorption and cholesterol homeostasis.[132] A comprehensive metagenomic study of the human gut microbiota revealed that BSH activity is highly conserved in this environment, being found in all major bacteria and archaeal divisions in the human gut. Furthermore, BSH activity increased bile resistance *in vitro* and gastrointestinal survival *in vivo*.[133] Consequently probiotics and BSH have been investigated as potential cholesterol-lowering therapeutics.[134, 135] Discovery of novel BSH genes and engineering probiotics with increased BSH activity may therefore be a viable and attractive therapeutic goal.

Finally, the gut-brain axis and the increasing evidence that the gut microbiota play a role in influencing processes in the brain has received much attention of late.[136-139] In addition, some microbes are capable of producing many neuroactive compounds such as gamma-aminobutyric acid (GABA), norepinephrine, serotonin and dopamine.[140] A recent study has shown that mice administered with a probiotic strain of *Lactobacillus rhamnosus* exhibited reduced anxiety and differential expression of GABA receptors in the brain. Animals that underwent vagotomy surgery (effectively cutting the main neural link between the gut and brain) did not display any of the effects seen in non-vagotomised animals when administered *L. rhamnosus*.[141] Such probiotics have recently been termed “psychobiotics”; a live organism that, when ingested in adequate amounts, produces a health benefit in patients suffering from psychiatric illness.[142] While much of the data is pre-clinical, it is certainly an exciting area to monitor in the future for further research and clinical trials. Perhaps metagenomics could be used to (a) identify genes from pathways that produce neuroactive compounds in the gut metagenome, (b) reveal the differences in the abundance of microbes capable of producing these compounds in the general population, (c) design functional assays to detect clones that produce known or novel neuroactive compounds or (d) create bio-engineered psychobiotics for specific psychiatric illnesses. Some of the potential applications of metagenomic novel gene discovery can be seen in Figure 2.

Figure 2.

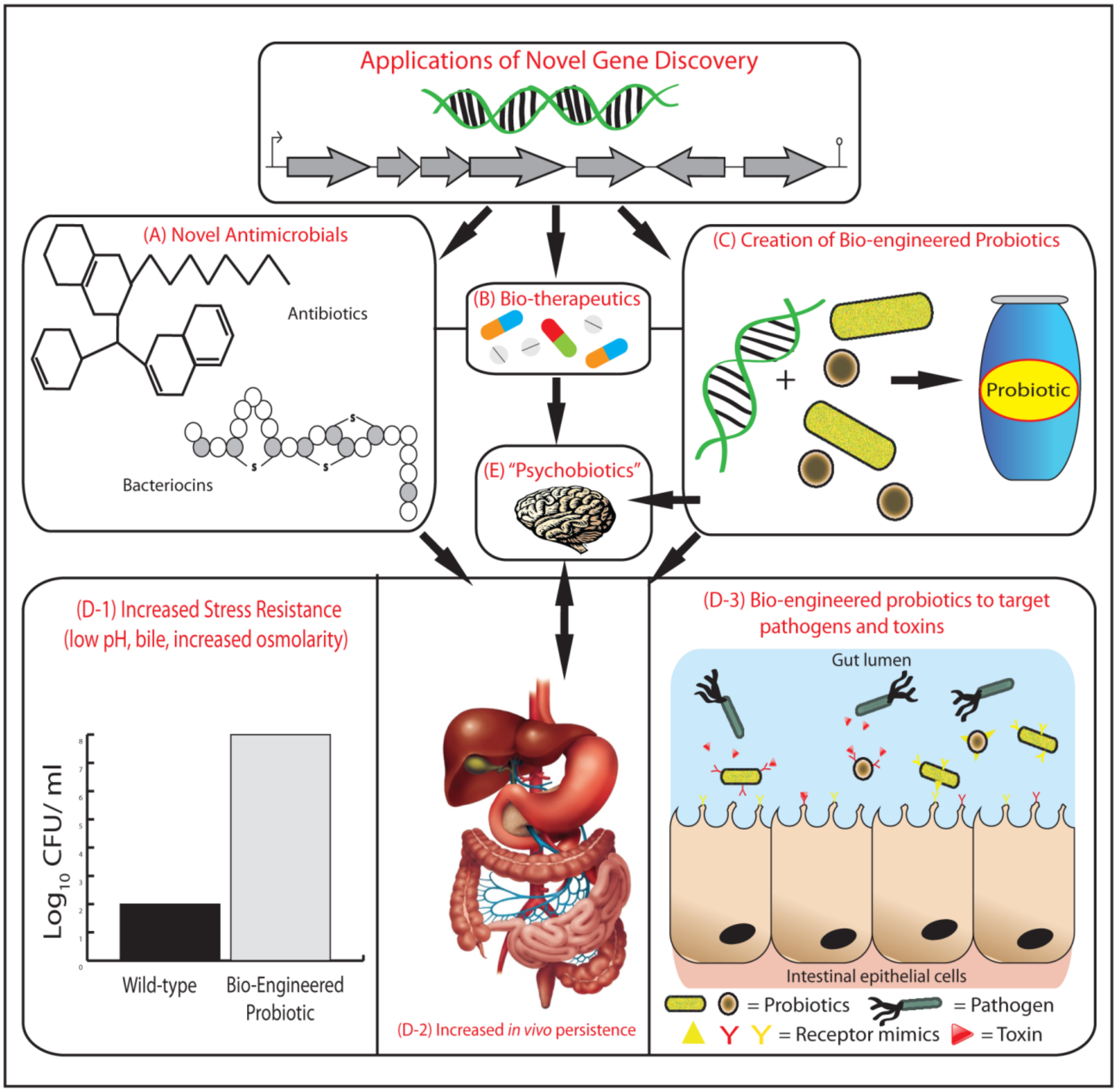

Applications of Novel Gene Discovery
(A) Novel Antimicrobials
Antibiotics
Bacteriocins
(B) Bio-therapeutics
(C) Creation of Bio-engineered Probiotics
Probiotic
(E) "Psychobiotics"
(D-1) Increased Stress Resistance
(low pH, bile, increased osmolarity)
Log$_{10}$ CFU/ ml
Wild-type
Bio-Engineered Probiotic
(D-2) Increased in vivo persistence
(D-3) Bio-engineered probiotics to target pathogens and toxins
Gut lumen
Intestinal epithelial cells
= Probiotics
= Pathogen
= Receptor mimics
= Toxin

Figure 2. Applications of novel genes discovered by metagenomics. Novel genes or gene clusters discovered through metagenomics will encode proteins for the production of novel antimicrobial compounds such as antibiotics and bacteriocins (A) and bio-therapeutics (B). Novel genes may also be used to create bio-engineered probiotic strains (C). Bio-engineered probiotics can be created that are more resistant host-associated stresses of the gastrointestinal tract such as low pH, bile and increased osmolarity (D-1), which will lead to

increased survival and persistence in the gastrointestinal tract (D-2). This will ultimately increase their therapeutic efficacy; using them to target specific pathogens or toxins by expressing receptor-mimic structures on their surface, thus preventing infection (D-3). The use of probiotics that have a health benefit to persons with a psychiatric illness have been termed “psychobiotics”. The creation of bio-engineered psychobiotics that can produce neuroactive compounds at an increased level or for a specific illness may be possible by identifying novel genes through metagenomics (E).

Biological Containment

The potential use of bio-engineered probiotics essentially involves the use of a genetically modified organism (GMO) and raises genuine concerns regarding what would be the deliberate release of a GMO into the environment when used as a bio-therapeutic. A number of containment strategies have been developed and can be categorised as either passive or active containment. With passive containment, an auxotrophic mutant is created which cannot grow without the exogenous provision of an essential metabolite. The most successful example of this is thymidine auxotrophy. Steidler *et al*, replaced the essential *thyA* gene (encoding thymidylate synthase) in *L. lactis*, with a gene encoding human interleukin 10 (IL-10) by chromosomal integration.[143] The mutant strain cannot grow in the absence of thymidine or thymine, which are absent in the environment, and cell death is induced by the phenomenon of “thymineless-death”.[144] This strain has been used in a human clinical trial for the treatment of IBD.[145] Active containment strategies are usually more complex and specific for the micro-organism in question, involving many gene deletions, tightly regulated gene expression and an inherent “suicide” system.[146, 147]

Conclusions and outlook

As we have outlined, metagenomic analyses enable comprehensive investigations of microbial communities and provide unprecedented access to the genetic diversity therein. An incredible amount of information has been gained in a relatively short time with regard to the taxonomic, phylogenetic and genetic novelty within diverse metagenomes. The development of new technologies and the continual reduction in sequencing costs will expand this at an ever increasing rate in the future, while the development of novel hosts and expression systems will increase hit rates and the variety of novel genes that can be discovered, thus helping to overcome some of the limitations of metagenomics. Development of novel and innovative screening assays will facilitate the discovery of previously unknown gene functions and novel therapeutic compounds. Greater emphasis will need to be placed on correct functional annotation of genes, as this already greatly lags behind the generation and deposition of metagenome sequences, which hampers subsequent analyses due to incorrect or absent database annotations. Synthetic biology is likely to come to the fore in the near future[148, 149], which will be coupled with decreasing costs and new opportunities to combine metagenomics and synthetic biology, possibly to synthesise optimised collections of genes mined from metagenome datasets for functional experiments. Identification of novel species or information from metagenomes will give clues to the metabolic and physiological requirements to enable the development of culture conditions to grow novel species. Bio-engineered probiotics face tough challenges if they are ever to be realised as clinical therapeutics; the creation and use of GMOs is a sensitive issue and acceptance by consumers would require undeniable demonstrations of biological containment, safety and ultimately, efficacy in large, double-blind, placebo controlled clinical trials.[150] Nevertheless, the extreme seriousness of increasing rates of antibiotic resistance and dissemination means that now is the time to think outside the box and intensively search out novel antimicrobials and therapeutics. It would be unfortunate to reach a stage where consumer acceptance of such novel therapies is based solely on a lack of alternatives. Metagenomics is likely to play a key role in identifying and developing such novel therapeutics.

Acknowledgements

The Alimentary Pharmabiotic Centre is a research centre funded by Science Foundation Ireland (SFI grant number 07/CE/B1368). We acknowledge the continued financial assistance of the Alimentary Pharmabiotic Centre, funded by Science Foundation Ireland. JRM acknowledges funding from The Royal Society which supports the bioinformatic cluster (Hive) at Cardiff University, School of Biosciences. RDS is an ESCMID Research Fellow and Coordinator of ClouDx-i an FP7-PEOPLE-2012-IAPP project

References

1. Sleator RD, Shortall C, Hill C. Metagenomics. Lett Appl Microbiol 2008; 47:361-6.

2. Rappe MS, Giovannoni SJ. The uncultured microbial majority. Annual review of microbiology 2003; 57:369-94.

3. Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, et al. Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol 2012; 30:513-20.

4. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 1998; 95:6578-83.

5. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological reviews 1995; 59:143-69.

6. Hugenholtz P. Exploring prokaryotic diversity in the genomic era. Genome biology 2002; 3:REVIEWS0003.

7. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 2013; 499:431-7.

8. Furrie E. A molecular revolution in the study of intestinal microflora. Gut 2006; 55:141-3.

9. Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, Streit WR. Prospecting for novel biocatalysts in a soil metagenome. Appl Environ Microbiol 2003; 69:6235-42.

10. Santosa DA. Rapid extraction and purification of environmental DNA for molecular cloning applications and molecular diversity studies. Molecular biotechnology 2001; 17:59-64.

11. Purohit MK, Singh SP. Assessment of various methods for extraction of metagenomic DNA from saline habitats of coastal Gujarat (India) to explore molecular diversity. Lett Appl Microbiol 2009; 49:338-44.

12. Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. Metagenomic comparison of direct and indirect soil DNA extraction approaches. Journal of microbiological methods 2011; 86:397-400.

13. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 2004; 304:66-74.

14. Salonen A, Nikkila J, Jalanka-Tuovinen J, Immonen O, Rajilic-Stojanovic M, Kekkonen RA, et al. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. Journal of microbiological methods 2010; 81:127-34.

15. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microbial informatics and experimentation 2012; 2:3.

16. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 2000; 289:1902-6.

17. Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. Proteorhodopsin phototrophy in the ocean. Nature 2001; 411:786-9.

18. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. Journal of biomedicine & biotechnology 2012; 2012:251364.

19. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012; 30:434-9.

20. Metzker ML. Sequencing technologies - the next generation. Nature reviews Genetics 2010; 11:31-46.

21. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. Journal of applied genetics 2011; 52:413-35.

22. Schloss PD, Handelsman J. Biotechnological prospects from metagenomics. Curr Opin Biotechnol 2003; 14:303-10.

23. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 2011; 331:463-7.

24. Kotik M. Novel genes retrieved from environmental DNA by polymerase chain reaction: current genome-walking techniques for future metagenome applications. Journal of biotechnology 2009; 144:75-82.

25. Tuffin M, Anderson D, Heath C, Cowan DA. Metagenomic gene discovery: how far have we moved into novel sequence space? Biotechnology journal 2009; 4:1671-83.

26. Singh B, Gautam SK, Verma V, Kumar M. Metagenomics in animal gastrointestinal ecosystem: Potential biotechnological prospects. Anaerobe 2008; 14:138-44.

27. Bell PJ, Sunna A, Gibbs MD, Curach NC, Nevalainen H, Bergquist PL. Prospecting for novel lipase genes using PCR. Microbiology 2002; 148:2283-91.

28. Knietsch A, Waschkowitz T, Bowien S, Henne A, Daniel R. Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on Escherichia coli. Appl Environ Microbiol 2003; 69:1408-16.

29. Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, Helynck G, et al. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl Environ Microbiol 2003; 69:49-55.

30. Feng Z, Kallifidas D, Brady SF. Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. Proc Natl Acad Sci U S A 2011; 108:12629-34.

31. Owen JG, Reddy BV, Ternei MA, Charlop-Powers Z, Calle PY, Kim JH, et al. Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. Proc Natl Acad Sci U S A 2013; 110:11797-802.

32. Banik JJ, Brady SF. Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proc Natl Acad Sci U S A 2008; 105:17273-7.

33. Banik JJ, Craig JW, Calle PY, Brady SF. Tailoring enzyme-rich environmental DNA clones: a source of enzymes for generating libraries of unnatural natural products. Journal of the American Chemical Society 2010; 132:15661-70.

34. Iwai S, Chai B, Sul WJ, Cole JR, Hashsham SA, Tiedje JM. Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. ISME J 2010; 4:279-85.

35. Bayer TS, Widmaier DM, Temme K, Mirsky EA, Santi DV, Voigt CA. Synthesis of methyl halides from biomass using engineered microbes. Journal of the American Chemical Society 2009; 131:6508-15.

36. Allgaier M, Reddy A, Park JI, Ivanova N, D'Haeseleer P, Lowry S, et al. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. PLoS One 2010; 5:e8812.

37. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 2005; 3:722-32.

38. Jones BV, Marchesi JR. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. Nat Methods 2007; 4:55-61.

39. Jones BV, Sun F, Marchesi JR. Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. BMC genomics 2010; 11:46.

40. Warburton PJ, Allan E, Hunter S, Ward J, Booth V, Wade WG, et al. Isolation of bacterial extrachromosomal DNA from human dental plaque associated with periodontal disease, using transposon-aided capture (TRACA). FEMS microbiology ecology 2011; 78:349-54.

41. Zhang T, Zhang XX, Ye L. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. PLoS One 2011; 6:e26041.

42. Recchia GD, Hall RM. Gene cassettes: a new class of mobile element. Microbiology 1995; 141 (Pt 12):3015-27.

43. Stokes HW, Holmes AJ, Nield BS, Holley MP, Nevalainen KM, Mabbutt BC, et al. Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. Appl Environ Microbiol 2001; 67:5240-6.

44. Wu YW, Rho M, Doak TG, Ye Y. Oral spirochetes implicated in dental diseases are widespread in normal human subjects and carry extremely diverse integron gene cassettes. Appl Environ Microbiol 2012; 78:5288-96.

45. Sureshan V, Deshpande CN, Boucher Y, Koenig JE, Stokes HW, Harrop SJ, et al. Integron gene cassettes: a repository of novel protein folds with distinct interaction sites. PLoS One 2013; 8:e52934.

46. Akopyants NS, Fradkov A, Diatchenko L, Hill JE, Siebert PD, Lukyanov SA, et al. PCR-based subtractive hybridization and differences in gene content among strains of Helicobacter pylori. Proc Natl Acad Sci U S A 1998; 95:13108-13.

47. Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, et al. Suppression subtractive hybridization: a method for generating

differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci U S A 1996; 93:6025-30.

48. Galbraith EA, Antonopoulos DA, White BA. Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. Environ Microbiol 2004; 6:928-37.

49. Meyer QC, Burton SG, Cowan DA. Subtractive hybridization magnetic bead capture: a new technique for the recovery of full-length ORFs from the metagenome. Biotechnology journal 2007; 2:36-40.

50. Simon C, Daniel R. Construction of small-insert and large-insert metagenomic libraries. In: Streit , WR and Daniel, R, ed(s). Metagenomics - Methods and ProtocolsNew York: Humana Press, 2010.

51. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annual review of genetics 2004; 38:525-52.

52. Ekkers DM, Cretoiu MS, Kielak AM, Elsas JD. The great screen anomaly--a new frontier in product discovery through functional metagenomics. Applied microbiology and biotechnology 2012; 93:1005-20.

53. Sleator RD. A Beginner's Guide to Phylogenetics. Microbial ecology 2013; 66:1-4.

54. Sleator RD. Phylogenetics. Arch Microbiol 2011; 193:235-9.

55. Gabor E, Liebeton K, Niehaus F, Eck J, Lorenz P. Updating the metagenomics toolbox. Biotechnology journal 2007; 2:201-6.

56. Boni IV. Diverse molecular mechanisms for translation initiation in prokaryotes. Molekuliarnaia biologiia 2006; 40:658-68.

57. Gabor EM, Alkema WB, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ Microbiol 2004; 6:879-86.

58. Uchiyama T, Miyazaki K. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol 2009; 20:616-22.

59. Lammle K, Zipper H, Breuer M, Hauer B, Buta C, Brunner H, et al. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. Journal of biotechnology 2007; 127:575-92.

60. Chung EJ, Lim HK, Kim JC, Choi GJ, Park EJ, Lee MH, et al. Forest soil metagenome gene cluster involved in antifungal activity expression in Escherichia coli. Appl Environ Microbiol 2008; 74:723-30.

61. McMahon MD, Guan C, Handelsman J, Thomas MG. Metagenomic analysis of Streptomyces lividans reveals host-dependent functional expression. Appl Environ Microbiol 2012; 78:3622-9.

62. Craig JW, Chang FY, Kim JH, Obiajulu SC, Brady SF. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. Appl Environ Microbiol 2010; 76:1633-41.

63. Li Y, Wexler M, Richardson DJ, Bond PL, Johnston AW. Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of Rhizobium leguminosarum and of Escherichia coli reveals different classes of cloned trp genes. Environ Microbiol 2005; 7:1927-36.

64. Angelov A, Mientus M, Liebl S, Liebl W. A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles. Systematic and applied microbiology 2009; 32:177-85.

65. Dobrijevic D, Di Liberto G, Tanaka K, de Wouters T, Dervyn R, Boudebbouze S, et al. High-throughput system for the presentation of secreted and surface-exposed proteins from Gram-positive bacteria in functional metagenomics studies. PLoS One 2013; 8:e65956.

66. Aakvik T, Degnes KF, Dahlsrud R, Schmidt F, Dam R, Yu L, et al. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. FEMS Microbiol Lett 2009; 296:149-58.

67. Kakirde KS, Wild J, Godiska R, Mead DA, Wiggins AG, Goodman RM, et al. Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. Gene 2011; 475:57-62.

68. Terron-Gonzalez L, Medina C, Limon-Mortes MC, Santero E. Heterologous viral expression systems in fosmid vectors increase the functional analysis potential of metagenomic libraries. Scientific reports 2013; 3:1107.

69. Johnston C, Douarre PE, Soulimane T, Pletzer D, Weingart H, Macsharry J, et al. Codon optimisation to improve expression of a Mycobacterium avium ssp. paratuberculosis-specific membrane-associated antigen by Lactobacillus salivarius. Pathogens and disease 2013; 68:27-38.

70. Warren RL, Freeman JD, Levesque RC, Smailus DE, Flibotte S, Holt RA. Transcription of foreign DNA in Escherichia coli. Genome research 2008; 18:1798-805.

71. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science 2009; 324:255-8.

72. Roller M, Lucic V, Nagy I, Perica T, Vlahovicek K. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res 2013.

73. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, et al. Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol 2002; 68:4301-6.

74. Kallifidas D, Kang HS, Brady SF. Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. Journal of the American Chemical Society 2012; 134:19552-5.

75. Chang FY, Brady SF. Cloning and characterization of an environmental DNA-derived gene cluster that encodes the biosynthesis of the antitumor

substance BE-54017. Journal of the American Chemical Society 2011; 133:9996-9.

76. Chang FY, Brady SF. Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. Proc Natl Acad Sci U S A 2013; 110:2478-83.

77. Feeney A, Sleator RD. The human gut microbiome: the ghost in the machine. Future microbiology 2012; 7:1235-7.

78. Sleator RD. The human superorganism - of microbes and men. Med Hypotheses 2010; 74:214-5.

79. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature 2011; 473:174-80.

80. Human_Microbiome_Project_Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012; 486:207-14.

81. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res 2007; 14:169-81.

82. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010; 464:59-65.

83. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature 2012; 488:178-84.

84. Rea MC, Dobson A, O'Sullivan O, Crispie F, Fouhy F, Cotter PD, et al. Effect of broad- and narrow-spectrum antimicrobials on Clostridium difficile and microbial diversity in a model of the distal colon. Proc Natl Acad Sci U S A 2011; 108 Suppl 1:4639-44.

85. Rea MC, Sit CS, Clayton E, O'Connor PM, Whittal RM, Zheng J, et al. Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against Clostridium difficile. Proc Natl Acad Sci U S A 2010; 107:9352-7.

86. Murphy K, O'Sullivan O, Rea MC, Cotter PD, Ross RP, Hill C. Genome mining for radical SAM protein determinants reveals multiple sactibiotic-like gene clusters. PLoS One 2011; 6:e20852.

87. Nature, Editorial. The antibiotic alarm. Nature 2013; 495:141.

88. Davies SC. Annual Report of the Chief Medical Officer: Volume Two, 2011; Infections and the rise of antimicrobial resistance Department of Health, London 2013.

89. European Medicines Agency. European Centre for Disease Prevention and Control. Joint technical report: the bacterial challenge—time to react. http://ecdc.europa.eu/en/publications/Publications/0909_TER_The_Bacterial_Challenge_Time_to_React.pdf 2009.

90. Alanis AJ. Resistance to antibiotics: are we in the post-antibiotic era? Arch Med Res 2005; 36:697-705.

91. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nature reviews Drug discovery 2007; 6:29-40.

92. Coates AR, Halls G, Hu Y. Novel classes of antibiotics or more of the same? British journal of pharmacology 2011; 163:184-94.

93. Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science 2013; 341:295-8.

94. Wallace BD, Wang H, Lane KT, Scott JE, Orans J, Koo JS, et al. Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. Science 2010; 330:831-5.

95. Haiser HJ, Turnbaugh PJ. Is it time for a metagenomic basis of therapeutics? Science 2012; 336:1253-5.

96. Roeselers G, Bouwman J, Venema K, Montijn R. The human gastrointestinal microbiota--an unexplored frontier for pharmaceutical discovery. Pharmacological research : the official journal of the Italian Pharmacological Society 2012; 66:443-7.

97. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. Science 2005; 308:1635-8.

98. Sleator RD. Probiotics -- a viable therapeutic alternative for enteric infections especially in the developing world. Discov Med 2010; 10:119-24.

99. Sleator RD, Hill C. Rational design of improved pharmabiotics. Journal of biomedicine & biotechnology 2009; 2009:275287.

100. Watson D, Sleator RD, Hill C, Gahan CG. Enhancing bile tolerance improves survival and persistence of Bifidobacterium and Lactococcus in the murine gastrointestinal tract. BMC Microbiol 2008; 8:176.

101. Sleator RD, Hill C. 'Bioengineered Bugs' - a patho-biotechnology approach to probiotic research and applications. Medical hypotheses 2008; 70:167-9.

102. Sleator RD, Hill C. New frontiers in probiotic research. Lett Appl Microbiol 2008; 46:143-7.

103. Sleator RD, Hill C. Designer probiotics: a potential therapeutic for Clostridium difficile? J Med Microbiol 2008; 57:793-4.

104. Sleator RD, Hill C. Battle of the bugs. Science 2008; 321:1294-5.

105. Sleator RD, Hill C. Patho-biotechnology; using bad bugs to make good bugs better. Sci Prog 2007; 90:1-14.

106. Sleator RD, Hill C. Patho-biotechnology: using bad bugs to do good things. Curr Opin Biotechnol 2006; 17:211-6.

107. FAO/WHO. Expert consultation on evaluation of health and nutritional properties of probiotics in food including milk powder with live lactic acid bacteria. FAO/WHO (Food and Agriculture Organization/World Health Organization) Cordoba, Argentina: WHO 2001

108. Paton AW, Morona R, Paton JC. A new biological agent for treatment of Shiga toxigenic Escherichia coli infections and dysentery in humans. Nat Med 2000; 6:265-70.

109. Focareta A, Paton JC, Morona R, Cook J, Paton AW. A recombinant probiotic for treatment and prevention of cholera. Gastroenterology 2006; 130:1688-95.

110. Paton AW, Jennings MP, Morona R, Wang H, Focareta A, Roddam LF, et al. Recombinant probiotics for treatment and prevention of enterotoxigenic Escherichia coli diarrhea. Gastroenterology 2005; 128:1219-28.

111. Paton AW, Morona R, Paton JC. Designer probiotics for prevention of enteric infections. Nat Rev Microbiol 2006; 4:193-200.

112. Paton AW, Morona R, Paton JC. Bioengineered microbes in disease therapy. Trends in molecular medicine 2012; 18:417-25.

113. Rosberg-Cody E, Stanton C, O'Mahony L, Wall R, Shanahan F, Quigley EM, et al. Recombinant lactobacilli expressing linoleic acid isomerase can modulate the fatty acid composition of host adipose tissue in mice. Microbiology 2011; 157:609-15.

114. Choi HJ, Ahn JH, Park SH, Do KH, Kim J, Moon Y. Enhanced wound healing by recombinant Escherichia coli Nissle 1917 via human epidermal growth factor receptor in human intestinal epithelial cells: therapeutic implication using recombinant probiotics. Infect Immun 2012; 80:1079-87.

115. Culligan EP, Hill C, Sleator RD. Probiotics and gastrointestinal disease: successes, problems and future prospects. Gut Pathog 2009; 1:19.

116. Culligan EP, Marchesi JR, Hill C, Sleator RD. Mining the human gut microbiome for novel stress resistance genes. Gut microbes 2012; 3:394-7.

117. Sleator RD, Hill C. Engineered pharmabiotics with improved therapeutic potential. Human vaccines 2008; 4:271-4.

118. Mattila-Sandholm T, Myllärinen P, Crittenden R, Mogensen G, Fondén R, Saarela M. Technological challenges for future probiotic foods. International Dairy Journal 2002; 12:173-82.

119. Sleator RD, Francis GA, O'Beirne D, Gahan CG, Hill C. Betaine and carnitine uptake systems in Listeria monocytogenes affect growth and survival in foods and during infection. J Appl Microbiol 2003; 95:839-46.

120. Sleator RD, Gahan CG, Abee T, Hill C. Identification and disruption of BetL, a secondary glycine betaine transport system linked to the salt tolerance of Listeria monocytogenes LO28. Appl Environ Microbiol 1999; 65:2078-83.

121. Sleator RD, Gahan CGM, O'Driscoll B, Hill C. Analysis of the role of betL in contributing to the growth and survival of Listeria monocytogenes LO28. Int J Food Microbiol 2000; 60:261-8.

122. Hoffmann RF, McLernon S, Feeney A, Hill C, Sleator RD. A single point mutation in the listerial betL sigma (A) -dependent promoter leads to improved osmo- and chill-tolerance and a morphological shift at elevated osmolarity. Bioengineered 2013; 4.

123. Sheehan VM, Sleator RD, Fitzgerald GF, Hill C. Heterologous expression of BetL, a betaine uptake system, enhances the stress tolerance of Lactobacillus salivarius UCC118. Appl Environ Microbiol 2006; 72:2170-7.

124. Sheehan VM, Sleator RD, Hill C, Fitzgerald GF. Improving gastric transit, gastrointestinal persistence and therapeutic efficacy of the probiotic strain Bifidobacterium breve UCC2003. Microbiology 2007; 153:3563-71.

125. FAO/WHO. Working Group on Drafting Guidelines for the Evaluation of Probiotics in Food. 2002.

126. Yoon MY, Lee KM, Yoon Y, Go J, Park Y, Cho YJ, et al. Functional screening of a metagenomic library reveals operons responsible for enhanced intestinal colonization by gut commensal microbes. Appl Environ Microbiol 2013; 79:3829-38.

127. Kimoto H, Nomura M, Kobayashi M, Mizumachi K, Okamoto T. Survival of lactococci during passage through mouse digestive tract. Canadian journal of microbiology 2003; 49:707-11.

128. Cotter PD, Hill C. Surviving the acid test: responses of gram-positive bacteria to low pH. Microbiol Mol Biol Rev 2003; 67:429-53, table of contents.

129. Guazzaroni ME, Morgante V, Mirete S, Gonzalez-Pastor JE. Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. Environ Microbiol 2013; 15:1088-102.

130. Culligan EP, Sleator RD, Marchesi JR, Hill C. Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. ISME J 2012; 6:1916-25.

131. Sleator RD, Hill C. Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. FEMS Microbiol Rev 2002; 26:49-71.

132. Ridlon JM, Kang DJ, Hylemon PB. Bile salt biotransformations by human intestinal bacteria. Journal of lipid research 2006; 47:241-59.

133. Jones BV, Begley M, Hill C, Gahan CG, Marchesi JR. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A 2008; 105:13580-5.

134. Jones ML, Tomaro-Duchesneau C, Martoni CJ, Prakash S. Cholesterol lowering with bile salt hydrolase-active probiotic bacteria, mechanism of action, clinical evidence, and future direction for heart health applications. Expert opinion on biological therapy 2013; 13:631-42.

135. Kumar M, Nagpal R, Kumar R, Hemalatha R, Verma V, Kumar A, et al. Cholesterol-lowering probiotics as potential biotherapeutics for metabolic diseases. Experimental diabetes research 2012; 2012:902917.

136. Cryan JF, Dinan TG. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. Nature reviews Neuroscience 2012; 13:701-12.

137. Barrett E, Ross RP, O'Toole PW, Fitzgerald GF, Stanton C. gamma-Aminobutyric acid production by culturable bacteria from the human intestine. J Appl Microbiol 2012; 113:411-7.

138. Collins SM, Bercik P. The relationship between intestinal microbiota and the central nervous system in normal gastrointestinal function and disease. Gastroenterology 2009; 136:2003-14.

139. Desbonnet L, Garrett L, Clarke G, Bienenstock J, Dinan TG. The probiotic Bifidobacteria infantis: An assessment of potential antidepressant properties in the rat. Journal of psychiatric research 2008; 43:164-74.

140. Lyte M. Probiotics function mechanistically as delivery vehicles for neuroactive compounds: Microbial endocrinology in the design and use of probiotics. BioEssays : news and reviews in molecular, cellular and developmental biology 2011; 33:574-81.

141. Bravo JA, Forsythe P, Chew MV, Escaravage E, Savignac HM, Dinan TG, et al. Ingestion of Lactobacillus strain regulates emotional behavior and

central GABA receptor expression in a mouse via the vagus nerve. Proc Natl Acad Sci U S A 2011; 108:16050-5.

142. Dinan TG, Stanton C, Cryan JF. Psychobiotics: A Novel Class of Psychotropic. Biological psychiatry 2013.

143. Steidler L, Neirynck S, Huyghebaert N, Snoeck V, Vermeire A, Goddeeris B, et al. Biological containment of genetically modified Lactococcus lactis for intestinal delivery of human interleukin 10. Nat Biotechnol 2003; 21:785-9.

144. Ahmad SI, Kirk SH, Eisenstark A. Thymine metabolism and thymineless death in prokaryotes and eukaryotes. Annual review of microbiology 1998; 52:591-625.

145. Braat H, Rottiers P, Hommes DW, Huyghebaert N, Remaut E, Remon JP, et al. A phase I trial with transgenic bacteria expressing interleukin-10 in Crohn's disease. Clin Gastroenterol Hepatol 2006; 4:754-9.

146. Kong W, Brovold M, Koeneman BA, Clark-Curtiss J, Curtiss R, 3rd. Turning self-destructing Salmonella into a universal DNA vaccine delivery platform. Proc Natl Acad Sci U S A 2012; 109:19414-9.

147. Bahey-El-Din M. Lactococcus lactis-based vaccines from laboratory bench to human use: an overview. Vaccine 2012; 30:685-90.

148. O' Driscoll A, Sleator RD. Synthetic DNA: The next generation of big data storage. Bioengineered 2013; 4:123-5.

149. Sleator RD. Digital biology: A new era has begun. Bioengineered 2012; 3:311-2.

150. Chakrabarty AM. Bioengineered bugs, drugs and contentious issues in patenting. Bioeng Bugs 2010; 1:2-8.

CHAPTER II

Functional Metagenomics Reveals Novel Salt Tolerance Loci from the Human Gut Microbiome

ISME J. 2012 Oct;6(10):1916-25

This Chapter may differ in layout from published manuscript as associated supplementary figures and tables have now been included in the main text.

Abstract

Metagenomics is a powerful tool that allows for the culture-independent analysis of complex microbial communities. One of the most complex and dense microbial ecosystems known is that of the human distal colon, with cell densities reaching up to 10^{12} per gram of faeces. With the majority of species in many environments considered as yet uncultured, there are an enormous number of novel genes awaiting discovery. In the current study, we conducted a functional screen of a metagenomic library of the human gut microbiota for potential salt-tolerant clones. Using transposon mutagenesis, three genes were identified from a single clone exhibiting high levels of identity to a species from the genus *Collinsella* (closest relative being *Collinsella aerofaciens*) (COLAER_01955, COLAER_01957 and COLAER_01981), a high GC, Gram positive member of the Actinobacteria commonly found in the human gut. The encoded proteins exhibit a strong similarity to GalE, MurB and MazG. Furthermore, pyrosequencing and bioinformatic analysis of two additional fosmid clones revealed the presence of an additional *galE* and *mazG* gene, with the highest level of genetic identity to *Akkermansia muciniphila* and *Eggerthella sp.* YY7918 respectively. Cloning and heterologous expression of the genes in the osmosensitive strain, *Escherichia coli* MKH13, resulted in increased salt tolerance of the transformed cells. It is hoped that the identification of atypical salt tolerance genes will help to further elucidate novel salt tolerance mechanisms, and will assist our increased understanding how resident bacteria cope with the osmolarity of the gastrointestinal tract.

Introduction

The ability to respond and adapt to changes in external osmolarity is a key determinant for bacterial survival and proliferation in various environmental niches (Sleator and Hill, 2002). Microorganisms are continually exposed to fluctuations and perturbations in osmolarity in their environment caused by rainwater, drought, salinity and changing solute concentrations. Both transient and symbiotic microorganisms that colonize the gastrointestinal (GI) tract are particularly susceptible to water loss due to osmotic stress (Gralla and Huo, 2008). It has been demonstrated that free water is not evenly distributed along the gut, but exists as pockets, meaning changes in osmolarity can be rapid (Schiller *et al.*, 2005). The elevated osmolarity of the upper small intestine (the equivalent of 0.3M NaCl (Chowdhury *et al.*, 1996)), represents an initial challenge to ingested microorganisms and the osmolarity is likely to be higher in the distal colon following further water absorption in the final stages of the digestive process. *Bacteroides fragilis* isolates from human stool samples have shown increased resistance to both NaCl and bile stress compared to isolates from blood or abscesses, indicating the potential importance of to such stress tolerance mechanisms in the gastrointestinal tract (Pumbwe *et al.*, 2007).

In general, bacteria respond to hyper-osmotic stress in a phased manner. Firstly, during the primary response, which is activated within seconds of osmotic up-shift, potassium (K^+) is rapidly transported into the cell (for a review see; (Epstein, 2003)). However, a more sophisticated osmotic stress response system is required once the cellular osmolality has been initially stabilized. The next phase, or secondary response, involves the uptake and/or synthesis of compatible solutes (also termed osmolytes or osmoprotectants). As their name suggests, these diverse compounds (Kempf and Bremer, 1998) are for the most part compatible with vital cellular functions and can restore cell volume and turgor without adversely affecting metabolism (Kunte, 2006). While the primary and secondary responses constitute the classical response to hyper-osmotic stress, the overall osmoadaptation strategy is much more complex, and involves a diverse range of cellular systems and processes seemingly unrelated to the primary and secondary responses. Identifying such diverse genes/proteins will provide us with a broader and more complete view of salt tolerance and the cellular response to salt-induced osmotic stress in bacteria. Indeed, many such ancillary mechanisms have been identified and range from two-component systems and proteolytic systems to numerous general stress and heat stress proteins and can also involve changes to the cell membrane (Gardan *et al.*, 2003; Kilstrup *et al.*, 1997; Lopez *et al.*, 2006; Piuri *et al.*, 2003; Sakamoto and Murata, 2002; Sleator and Hill, 2005; Wonderling *et al.*, 2004).

Since their inception, metagenomic strategies have led to the identification of numerous novel and diverse genes, enzymes and proteins from many diverse environments through sequence-based and/or functional approaches (Banik and Brady, 2008; Beja *et al.*, 2000; Gillespie *et al.*, 2002; Heath *et al.*, 2009; Lee *et al.*, 2007; Meilleur *et al.*, 2009). In principle, metagenomics can provide access to all of the genetic resources in a given environmental niche and as such is an extremely powerful tool to access the genomes of difficult-to-culture or unculturable microorganisms (Sleator *et al.*, 2008). A recent study identified novel genes from a pond water metagenome that increased resistance to salt-induced osmotic stress when expressed in *E. coli* (Kapardar *et al.*, 2010(b); Kapardar *et al.*, 2010(a)). While some studies have used functional metagenomics to screen for various phenotypes from the human gut environment (Gloux *et al.*, 2011; Kazimierczak *et al.*, 2008; Lakhdari *et al.*, 2010) to our knowledge this study is the first to identify genes that confer salt tolerance from the human gut microbiota.

Materials and methods

Bacterial strains and growth conditions

Bacterial strains and plasmids used in this study are listed in Table 1. Primers (Eurofins, MWG Operon, Germany) used in this study are listed in Table 2. *E. coli* EPI300::pCC1FOS (Epicentre Biotechnologies, Madison, WI, USA) was grown in Luria-Bertani (LB) medium (Maniatis, 1982) containing 12.5μg/ml chloramphenicol (Cm) and in 12.5μg/ml chloramphenicol plus 50μg/ml kanamycin (Kan) following EZ-Tn*5*™ transposon mutagenesis reactions. *E. coli* MKH13 was grown in LB medium and in LB medium supplemented with 20µg/ml Cm for strains transformed with the plasmid pCI372. LB media was supplemented with 1.5% agar when required. All overnight cultures were grown at 37°C with shaking.

Table 1. Bacterial strains, plasmids and transposon used in this study

Strain, plasmid or transposon	Genotype or characteristic(s)	Source or reference
Strains		
E. coli EPI300	F$^-$ *mcrA* Δ(*mrr-hsd*RMS-*mcrBC*) Φ80d*lacZ*ΔM15 Δ*lac*X74 *recA*1 *endA*1 *araD*139 Δ(*ara, leu*)7697 *galU galK* λ$^-$ *rpsL nupG trfA dhfr*; high-transformation efficiency of large DNA	Epicentre Biotechnologies, Madison, WI, USA
E.coli MKH13	MC4100*Δ*(*putPA*)101D(*proP*)2D(*proU*)	(Haardt *et al.*, 1995)
E. coli MKH13::pCI372-*murB*(3)	MKH13 containing pCI372 with *murB* gene from SMG 3 (similar to *Collinsella aerofaciens* ATCC 25986)	This study
E. coli MKH13::pCI372-*mazG*(3)	MKH13 containing pCI372 with *mazG* gene from SMG 3 (similar to *Collinsella aerofaciens* ATCC 25986)	This study
E. coli MKH13::pCI372-*galE*(3)	MKH13 containing pCI372 with *galE* gene from SMG 3 (similar to *Collinsella aerofaciens* ATCC 25986)	This study
E. coli MKH13::pCI372-*mazG*(5)	MKH13 containing pCI372 with *mazG* gene from SMG 5 (similar to *Eggerthella sp.* YY7918	This study
E. coli MKH13::pCI372-*galE*(25)	MKH13 containing pCI372 with *galE* gene from SMG 25 (similar to *Akkermansia muciniphila* ATCC BAA-835)	This study
Plasmids		
pCI372	Shuttle vector between *E. coli* and *L. lactis*, Cm^R	(Hayes *et al.*, 1990)
pCC1FOS	Fosmid cloning vector, Cm^R	Epicentre Biotechnologies, Madison, WI, USA
Transposon		
EZ-Tn*5* <*ori*V/ KAN-2>	Hyperactive Tn*5* transposon, Kan^R, inducible high copy origin of replication – *ori*V	Epicentre Biotechnologies, Madison, WI, USA

Metagenomic library construction

A metagenomic library which had been constructed previously (Jones and Marchesi, 2007) was used to screen for salt-tolerant clones. Briefly, metagenomic DNA was isolated from a human faecal sample (from a 26 year-old healthy Caucasian male) and from this a fosmid library (average insert size of approximately 40kb) was created using a Copy Control^{TM} fosmid library production kit (Epicentre biotechnologies) according to the manufacturer's instructions. A Genetix QPix 2 XT™ colony picking robot was used to transfer fosmid clones to 384-well micro-titer plates, which were stored at -80°C until needed.

Screening the metagenomic library

A total of 23,040 clones from the library (average insert size of ~40kb) were screened on LB agar supplemented with 6.5% NaCl for clones showing increased tolerance to NaCl compared to the *E. coli* EPI300 host strain (containing an empty pCC1FOS cloning vector). A Genetix QPix 2 XT™ colony picking/gridding machine was used to plate the clones onto Q-Trays (Genetix) containing LB agar supplemented with 12.5μg/ml Cm and 6.5% NaCl. Q-Trays were incubated at 37°C for 72 hours. Following incubation, likely salt-tolerant clones were replica plated onto LB containing 12.5μg/ml Cm and 6.5% NaCl and onto LB containing 12.5μg/ml but without NaCl (which served as a positive control).

Growth experiments

Cultures were grown overnight with shaking in LB broth. Subsequently, cells were harvested, washed in one quarter strength sterile Ringer's solution and resuspended in fresh LB broth. A 2% inoculum was sub-cultured in fresh LB broth containing the appropriate stress (i.e. 0-10% (w/v) NaCl, 0-8% (w/v) KCl, 0-90% (w/v) sucrose or 0-80% (v/v) glycerol as required) and 200µl was transferred to a sterile 96-well micro-titer plate (Starstedt Inc. Newton, USA). Plates were incubated at 37°C for 48 hours in an automated spectrophotometer (Tecan Genios) which recorded the OD 595nm every hour. After 48 hours the data was retrieved and analysed using the Magellan 3 software program. Results are presented as the average of triplicate experiments, with error bars being representative of the standard error of the mean (SEM).

DNA manipulations

Extraction of fosmids containing metagenomic DNA: 5ml of bacterial culture was grown overnight with 12.5μg/ml Cm. One millilitre of culture was used to inoculate 4 ml of fresh LB broth. To this, 5μl of 1000x Copy Control ™ Induction solution (Epicentre Biotechnologies) and 12.5μg/ml Cm were added. The mixture was incubated at 37°C for 5 hours with vigorous shaking (200-250 r.p.m.) to ensure maximum aeration. Cells were harvested from the whole 5ml of induced culture by centrifuging at 2100 x *g* for 12 minutes. Qiagen QIAprep® Spin mini-prep kit was used to extract fosmids as per manufacturer's instructions. PCR products were purified with a Qiagen PCR purification kit and digested with *XbaI* and *PstI* (Roche Applied Science), followed by ligation (T4 DNA ligase (Roche) for *mazG*(3) and *murB*(3); FastLink DNA ligase for *galE*(3)*, galE*(25) *and mazG*(5) (Epicentre Biotechnologies) as per manufacturer's instructions) to similarly digested plasmid pCI372. Electrocompetent *E. coli* MKH13 were transformed with the ligation mixture and plated on LB agar plates containing 20µg/ml Cm for selection. Colony PCR was performed on all resistant transformants using primers across the multiple cloning site (MCS) of pCI372 to confirm the presence and size of the insert.

Transposon mutagenesis

Transposon mutagenesis was carried out on SMG 3 according to the manufacturer's instructions using the EZTn-5 <*oriV*/ KAN-2> *in vitro* transposition kit (Epicentre Biotechnologies). *E. coli* EPI300 cells were transformed with the transposon reaction mixture and selected on plates containing Cm and Kan (12.5 and 50µg/ml, respectively). Subsequently, the transposon insertion clones were replica plated onto LB with and without 6.5% NaCl. Clones which grew on LB but not on LB + 6.5% NaCl indicates a likely insertion event in a gene involved in salt tolerance. Presumptive salt- tolerant knock-outs were grown overnight and a fosmid DNA extraction was performed. The extracted fosmid containing metagenomic DNA was subjected to sequencing from the ends of the transposon using the primers EZTn-FP-1 and EZTn-RP-1 (Table 2.). All sequencing was performed by GATC Biotech (Germany).

Sequencing and bioinformatic analysis

Fosmids SMG 1, 3, 5, 6 and 25 were fully sequenced and assembled by GATC Biotech (Germany) using the GS FLX (Roche) platform. Putative open reading frames were predicted using Softberry FGENESB bacterial operon and gene prediction software (Mavromatis *et al*, 2007). Retrieved nucleotide and translated amino acid sequences were functionally annotated by homology searches using the Basic Local Alignment and Search Tool (BLAST), to identify homologous sequences from the National Centre for Biotechnology Information (NCBI) website: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.

End sequencing was performed on all fosmid clones (SMG1-53) using T7 or M13 R primers and the taxonomic distribution of end sequences were assigned using the BLASTP program (Table 4).

Nucleotide sequences identified in this study were compared against the healthy individuals in the MetaHit dataset (Qin *et al.,* 2010) using the BLAST program to identify homologous sequences (>50% identity; e-value < $1^\wedge10^{-5}$). This data was used to determine the relative abundance of the identified genes in the data set and estimate the amount of DNA that would need to be screened (Mb) to return a hit to one of the genes.

In addition, BLASTP analysis was undertaken to identify the protein sequences in GenBank, which showed the highest identity to genes isolated here and which were shown to be responsible for the salt tolerance phenotype. The closets hits were aligned using ClustalW (Thompson *et al.*, 1994) in BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html). The aligned proteins were analysed using MEGA 5 (Tamura *et al.*, 2011) and the evolutionary history was inferred using the Neighbour-Joining method (Saitou and Nei, 1987) from the evolutionary distances which were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The bootstrap method (500 replicates) was used to test the percentage of replicate trees in which the associated taxa clustered together (Felsenstein, 1985) (Figure 5).

Table 2. Primers used in this study

Primer	Sequence (5' – 3')[a]
pCI372 FP	CGGGAAGCTAGAGTAAGTAG
pCI372 RP	CCTCTCGGTTATGAGTTAG
mazG(3) FP	AAAACTGCAGGCCCGTCGTTCCCGCAGTCTTAC
mazG(3) RP	GCTCTAGAATCTACGAGGGCGGCGCGTTC
murB(3) FP	AAAACTGCAGCCACCTCCTGGGCGATCTGCTTGAG
murB(3) RP	GCTCTAGACGACACACCGGACTGGGTTATCTGA
galE(3) FP	AAAACTGCAGATGGGTGTGCAGTCCGCCTC
galE(3) RP	GCTCTAGAGTCCCAACGATTTCCACGAACG
mazG(5) FP	AAAACTGCAGCTAAACAGGAGGCGAAGCTC
mazG(5) RP	GCTCTAGAGCAGAAGGCGTCAACGATA
galE(25) FP	GCTCTAGACCGGCTTAACAGCATTGATA
galE(25) RP	AAAACTGCAGGCTGCGTTGTCTTTCCAGTT
EZTn-FP-1	GCCAACGACTACGCACTAGCCAAC
EZTn-RP-1	GAGCCAATATGCGAGAACACCCGAGAA
T7	TAATACGACTCACTATAGGG
M13 R	CAGGAAACAGCTATGACC

[a] Restriction enzyme recognition sequences are underlined; FP= forward primer; RP= reverse primer.

Results

Screening the metagenomic library

In this study, a metagenomic fosmid library constructed from DNA isolated from human gut microbiota was screened for clones which conferred increased salt tolerance using a combined functional metagenomic, transposon mutagenesis and bioinformatic analysis approach. The cloning host is incapable of growth on LB agar supplemented with 6.5% NaCl, providing a positive selection for clones able to cope with elevated osmolarity. In total, 53 salt-tolerant clones were identified which were annotated SMG (salt metagenome) 1-53. Six clones (SMG 1-6) grew within 24 hours; while a further 47 CFUs appeared within 36 hours. Physiological growth experiments were conducted on host strain EPI300::pCC1FOS and SMG 1-6. All six clones showed increased salt tolerance relative to the control at 7% NaCl (*E. coli* EPI300::pCC1FOS host strain) (Figure 1). The clones were subjected to further study through transposon mutagenesis.

Figure 1.

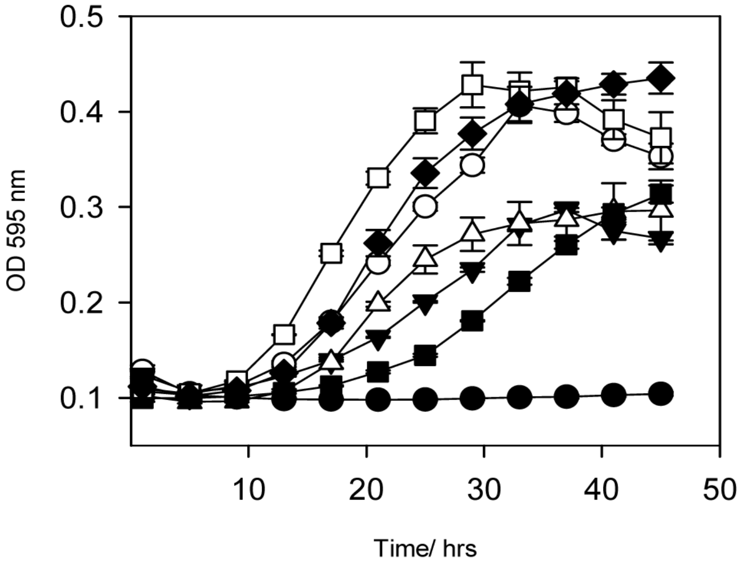


Figure 1: Growth in LB broth supplemented with 7% NaCl. Growth of *E. coli* EPI300::pCC1FOS host strain (● closed circle), SMG 1 (○ open circle), SMG 2 (▼ closed triangle), SMG 3 (△ open triangle), SMG 4 (■ closed square), SMG 5 (□ open square) and SMG 6 (◆ closed diamond).

Transposon mutagenesis

Transposon mutagenesis was carried out using the EZ-Tn*5*^{TM} <oriV/ KAN-2> insertion kit (Epicentre Biotechnologies). Transposon mutagenesis proved unsuccessful for SMG 1, SMG 5 and SMG 6, possibly due to difficulty in obtaining sufficiently high yields of fosmid DNA for *in vitro* transposon mutagenesis, so SMG 3 was chosen for further investigation. Transposon mutagenesis of SMG 3 yielded nine clones incapable of growth on 6.5% NaCl. Sequencing revealed transposon insertions in three distinct genes with high genetic identity (95-98%) to genes encoding hypothetical proteins from a member of the genus *Collinsella* (closest relative being C. *aerofaciens*). The genes; *galE*(3) (COLAER_01955), *murB*(3) (COLAER_01957) and *mazG*(3) (COLAER_01981), encode hypothetical proteins similar to UDP-glucose 4-epimerase (GalE), UDP-*N*-acetylenolpyruvoylglucosamine reductase (MurB) and nucleoside triphosphate pyrophosphohydrolase (MazG family protein) (see Table 3). While GalE has been tentatively linked to salt tolerance in previous studies (Bohringer *et al.*, 1995; Hengge-Aronis *et al.*, 1991; Nguyen *et al.*, 2004) we are unaware of any previous link between MurB or MazG and bacterial salt tolerance.

Table 3. Genes identified in this study

SMG clone	Putative gene	Method of identification	Top Blast hit	% Identity (aa)	e-value
SMG 3–3	*mazG*(3) COLAER_01981 (861 bp)	EZTn5 transposon mutagenesis	Hypothetical protein; *Collinsella aerofaciens,* Similar to *mazG* encoding a MazG family protein (nucleoside triphosphate pyrophosphohydrolase) (286 aa)	95 %	2e- 94
SMG 3–11	*murB*(3) COLAER_01957 (963 bp)	EZTn5 transposon mutagenesis	Hypothetical protein; *Collinsella aerofaciens,* Similar to *murB* encoding UDP-*N*-acetylenolpyruvoylglucosamine reductase (320 aa)	98%	5e- 83
SMG 3-17	*galE*(3) COLAER_01955 (1062 bp)	EZTn5 transposon mutagenesis	Hypothetical protein; *Collinsella aerofaciens,* Similar to *galE* encoding UDP glucose 4-epimerase (353 aa)	98%	3e- 175
SMG 5	*mazG*(5) EGYY_03530 (876 bp)	Bioinformatic analysis	Hypothetical protein; *Eggerthella sp.* YY7918, Similar to *mazG* encoding a MazG family protein (nucleoside triphosphate pyrophosphohydrolase) (291 aa)	61%	1e-117
SMG 25	*galE*(25) Amuc_1125 (990 bp)	Bioinformatic analysis	Hypothetical protein; *Akkermansia muciniphila,* Similar to *galE* encoding a UDP glucose 4-epimerase (329 aa)	96%	0.0

Abbreviations: aa = amino acid(s); bp = base pair(s).

Sequencing and bioinformatic analysis

End sequencing was carried out on all 53 SMG clones, using T7 or M13 R primers (GATC Biotech, Germany). Forty-nine clones were successfully end sequenced and taxonomically assigned based on Blast hits. Of these, the *Bacteroidetes* dominated, representing 57.14% of the sequences, followed by the *Actinobacteria* and *Proteobacteria* at 14.29% and 12.24% respectively. The *Verrucomicrobia* made up 8.16% of the sequences, while the *Firmicutes* were under-represented at 4.08%. The remainder of the sequences were unidentified and one sequence came from a member of the *Heterokontophyta* (*Blastocystis hominis*), a eukaryote (2.04% each) (Table 4).

Table 4. End sequencing data from SMG clones

Clone	Best hit microorganism	Top BLASTX hit	e-value	% coverage	% identity	Phylum
SMG 1	*Bacteroides thetaiotaomicron* VPI-5482	Hypothetical protein BT_1366	0.00E+00	100%	100%	*Bacteroidetes*
SMG 2	*Collinsella aerofaciens* ATCC 25986	Hypothetical protein COLAER_01951	1.00E-133	99%	96%	*Actinobacteria*
SMG 3	*Collinsella aerofaciens* ATCC 25986	Hypothetical protein COLAER_01951	2.00E-146	99%	88%	*Actinobacteria*
SMG 4	*Collinsella aerofaciens* ATCC 25986	Hypothetical protein COLAER_01951	2.00E-41	99%	95%	*Actinobacteria*
SMG 5	*Eggerthella sp.* YY7918	Aldehyde ferredoxin oxidoreductase	6.00E-120	99%	42%	*Actinobacteria*
SMG 6	*Bacteroides thetaiotaomicron* VPI-5482	Hypothetical protein BT_1366	0.00E+00	100%	100%	*Bacteroidetes*
SMG 7	*Bacteroides sp.* 4_3_47FAA	Two-component system; response regulator sensor kinase	1.00E-135	83%	98%	*Bacteroidetes*
SMG 8	*Bacteroides sp.* 3_1_40A	Hypothetical protein HMPREF9011_03915	3.00E-160	99%	99%	*Bacteroidetes*
SMG 9	*Alistipes putredinis* DSM 17216	Hypothetical protein ALIPUT_02368	0.00E+00	47%	72%	*Bacteroidetes*
SMG 10	*Paraprevotella xylaniphila* YIT 11841	Helicase protein	8.00E-143	99%	99%	*Bacteroidetes*
SMG 11	*Bacteroides sp.* 3_1_40A	Dipeptidyl aminopeptidase	0.00E+00	99%	99%	*Bacteroidetes*
SMG 12	*Collinsella aerofaciens* ATCC 25986	Hypothetical protein COLAER_01951	6.00E-48	92%	98%	*Actinobacteria*
SMG 13	*Bacteroides stercoris* ATCC 43183	Hypothetical protein BACSTE_02336	5.00E-80	91%	64%	*Bacteroidetes*
SMG 14	*Bacteroides sp.* D2	TonB-dependent receptor	1.00E-65	99%	55%	*Bacteroidetes*
SMG 15	n/a	No sequence data	n/a	n/a	n/a	n/a
SMG 16	*Pseudomonas syringae pv. aceris str.* M302273PT	Phage integrase, putative	7.00E-31	69%	47%	*Proteobacteria* (γ)
SMG 17	*Pseudomonas syringae pv. aceris str.* M302273PT	Phage integrase, putative	1.00E-32	50%	51%	*Proteobacteria* (γ)
SMG 18	n/a	No sequence data	n/a	n/a	n/a	n/a
SMG 19	*Prevotella copri* DSM 18205	Putative Tat pathway signal sequence	2.00E-26	59%	52%	*Bacteroidetes*

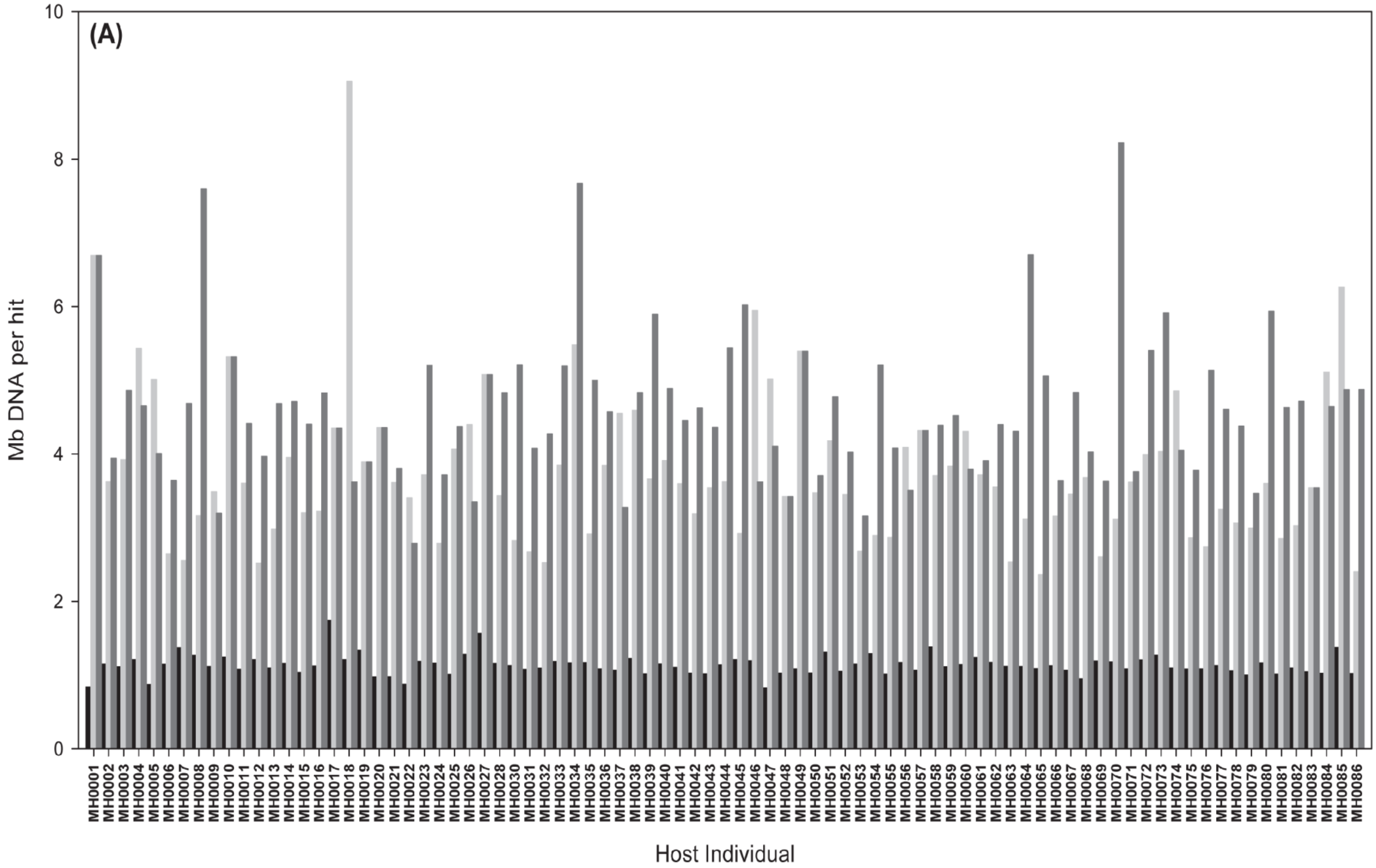
SMG 20	*Slackia heliotrinireducens* DSM 20476	Fe-S oxidoreductase, coproporphyrinogen III oxidase	2.00E-18	22%	70%	*Actinobacteria*
SMG 21	*Bacteroides caccae* ATCC 43185	Hypothetical protein BACCAC_01367	1.00E-126	58%	98%	*Bacteroidetes*
SMG 22	*Prevotella oulorum* F0390	Hypothetical protein HMPREF9431_01698	7.00E-61	58%	65%	*Bacteroidetes*
SMG 23	*Clostridium methylpentosum* DSM 5476	Hypothetical protein CLOSTMETH_00476	8.00E-19	97%	23%	*Firmicutes*
SMG 24	*Bacteroides eggerthii* 1_2_48FAA	TonB-dependent receptor	3.00E-123	100%	54%	*Bacteroidetes*
SMG 25	*Akkermansia muciniphila* ATCC BAA-835	DNA polymerase III, alpha subunit Amuc_0374	0.00E+00	99%	95%	*Verrucomicrobia*
SMG 26	*Bacteroides fragilis* 3_1_12	Excinuclease ABC subunit B	1.00E-34	73%	82%	*Bacteroidetes*
SMG 27	*Alistipes sp.* HGB 7	Glycosyl hydrolase, family 57	0.00E+00	97%	95%	*Bacteroidetes*
SMG 28	*Pseudomonas fluorescens* Pf-5	Efflux; ABC transporter ATP-binding protein	1.00E-72	79%	53%	*Proteobacteria* (γ)
SMG 29	Uncultured rumen bacterium	Beta-D-xylosidase/alpha-L-arabinosidase	1.00E-90	82%	66%	*Bacteroidetes*
SMG 30	*Bacteroides caccae* ATCC 43185	Hypothetical protein BACCAC_01367	2.00E-90	59%	98%	*Bacteroidetes*
SMG 31	n/a	No sequence data	n/a	n/a	n/a	n/a
SMG 32	*Lacinutrix sp.* 5H-3-7-4	Primosomal protein N'	9.00E-74	92%	44%	*Bacteroidetes*
SMG 33	*Capnocytophaga sp.* oral taxon 329 str. F0087	RND transporter, HAE1/HME family, permease protein	6.00E-84	98%	47%	*Bacteroidetes*
SMG 34	*Bacterium Ellin514*	Antibiotic biosynthesis monooxygenase	8.00E-41	54%	40%	*Verrucomicrobia*
SMG 35	*Alistipes sp.* HGB5	Transporter; major facilitator family protein (sugar phosphate permease)	9.00E-67	73%	56%	*Bacteroidetes*
SMG 36	*Bacteroides coprosuis* DSM 18011	Hypothetical protein Bcop_0579	5.00E-69	99%	48%	*Bacteroidetes*
SMG 37	*Pseudomonas fluorescens* SBW25	Putative integrase	1.00E-60	66%	61%	*Proteobacteria* (γ)
SMG 38	*Prevotella buccae* D17	Glycine dehydrogenase (decarboxylating), subunit 2	1.00E-57	35%	75%	*Bacteroidetes*
SMG 39	*Bacteroides sp.* 4_3_47FAA	Dipeptidyl peptidase IV	0.00E+00	99%	98%	*Bacteroidetes*
SMG 40	*Oxalobacter formigenes* OXCC13	Conserved hypothetical protein	3.00E-130	81%	79%	*Proteobacteria* (β)
SMG 41	*Alistipes sp.* HGB5	3-dehydroquinate dehydratase, type II	2.00E-16	45%	78%	*Bacteroidetes*

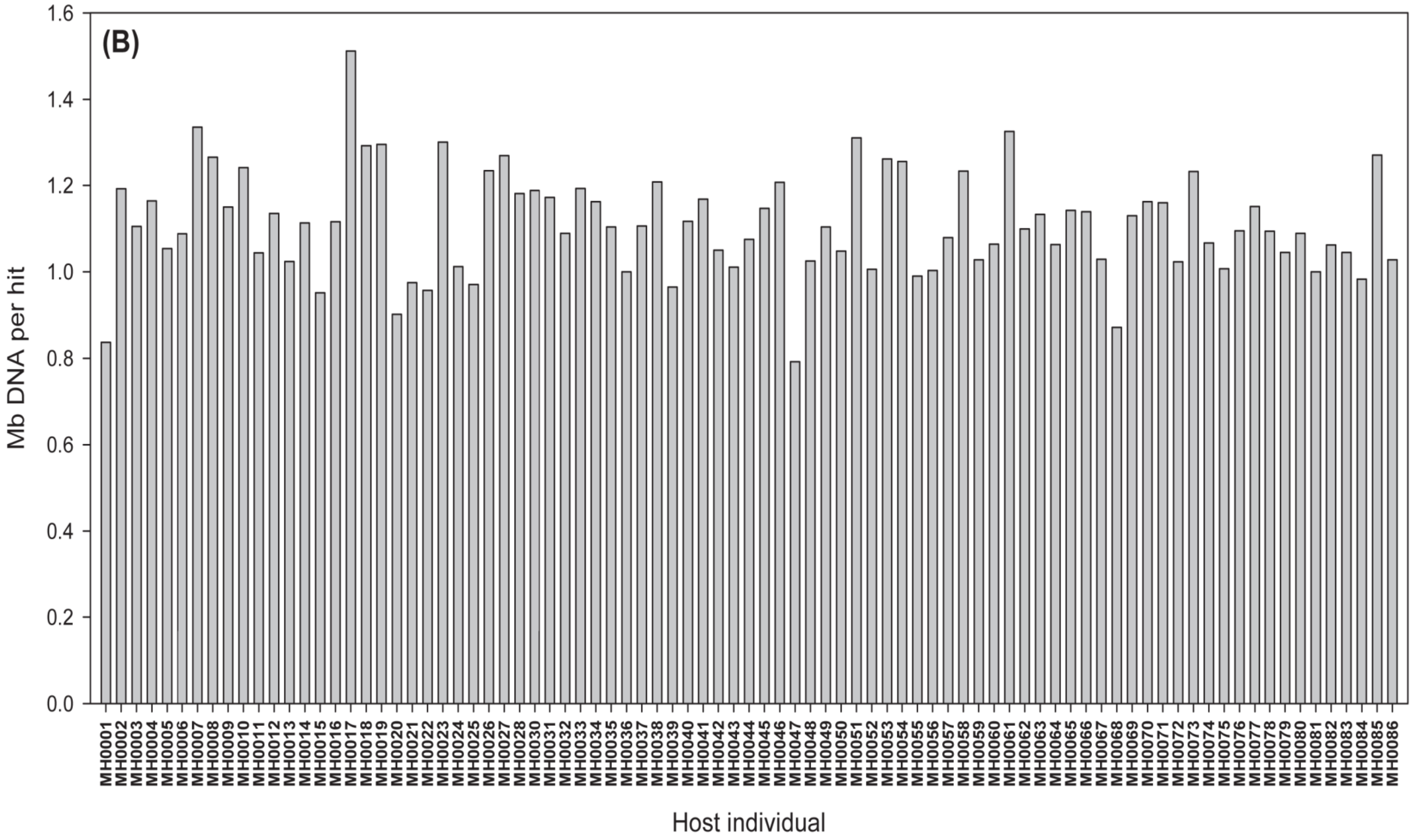
SMG 42	*Bacteroides sp.* 9_1_42FAA	V-type ATP synthase subunit E	2.00E-128	53%	100%	*Bacteroidetes*
SMG 43	*Eubacterium rectale* DSM 17629	RecA-family ATPase	3.00E-131	88%	99%	*Firmicutes*
SMG 44	n/a	No sequence data	n/a	n/a	n/a	n/a
SMG 45	*Blastocystis hominis*	Unnamed protein product	4.00E-17	91%	47%	*Heterokontophyta*
SMG 46	*Akkermansia muciniphila* ATCC BAA-835	Hypothetical protein Amuc_0150	7.00E-25	66%	36%	*Verrucomicrobia*
SMG 47	*Pseudomonas fluorescens* Pf0-1	GCN5-like N-acetyltransferase	2.00E-82	48%	81%	*Proteobacteria* (γ)
SMG 48	Unidentified	No significant similarity	n/a	n/a	n/a	n/a
SMG 49	*Bacteroides ovatus* SD CMC 3f	Conserved hypothetical protein	7.00E-94	98%	48%	*Bacteroidetes*
SMG 50	*Alistipes sp.* HGB5	Putative membrane protein	7.00E-45	52%	90%	*Bacteroidetes*
SMG 51	*Akkermansia muciniphila* ATCC BAA-835	Small multidrug resistance protein Amuc_0621	3.00E-39	32%	88%	*Verrucomicrobia*
SMG 52	*Bacteroides thetaiotaomicron* VPI-5482	Hypothetical protein (BT_1366)	0.00E+00	100%	100%	*Bacteroidetes*
SMG 53	*Bifidobacterium bifidum* PRL2010	TrpS Tryptophanyl-tRNA synthetase	1.00E-118	100%	99%	*Actinobacteria*

Full fosmid sequencing was performed on clones SMG 1, 3, 5, 6 and 25, revealing insert sizes of 36, 39, 42, 36 and 44kb respectively. The sequences have been submitted to GenBank; with following accession numbers: SMG 1 = JQ269596; SMG 3 = JQ269597; SMG 5 = JQ269598; SMG 6 = JQ269599; SMG 25 = JQ269600. SMG 1 and SMG 6 were found to have the same metagenomic insert, with the highest genetic identity to *Bacteroides thetaiotaomicron*. SMG 5 shared highest genetic identity to *Eggerthella sp.* YY7918 (a member of the high G+C Gram-positive *Actinobacteria*). However, it should be noted that BlastP analysis of the predicted proteins encoded by the SMG 5 genes revealed many variations in the associated species. While all but one corresponded to members of the phylum *Actinobacteria*, the genera were represented by various species of *Eggerthella, Slackia, Cryptobacterium* and *Gordonibacter*. This indicates that the SMG 5 insert represents DNA from a novel species from one of these genera or the occurrence of recombination events between species of these genera. SMG 25 shared highest genetic identity to *Akkermansia muciniphila* (a member of the Gram-negative *Verrucomicrobia*). Putative open reading frames were identified using FGENESB bacterial operon and gene prediction software from Softberry (Mavromatis *et al.*, 2007). BLAST analysis revealed the presence of a hypothetical protein (EGYY_03530) and a UDP glucose 4-epimerase encoded by putative *mazG* and *galE* genes respectively. These genes were present on SMG 5 and SMG 25 and were designated *mazG*(5) and *galE*(25) respectively. Having previously cloned and expressed homologous genes (*galE*(3) and *mazG*(3)) from SMG 3 in *E. coli* MKH13, resulting in an increased salt tolerance phenotype, it was decided to clone these bioinformatically identified *mazG*(5) and *galE*(25) genes into *E. coli* MKH13 also.

The nucleotide sequences of the genes identified were compared with the MetaHit dataset from the healthy individuals (Qin *et al.,* 2010) to determine the relative abundance of the genes among the gut metagenomes of the subjects. Homologues of all of the genes identified in this study were found to be present in all MetaHit samples (>50% identity; e-value <1^10^{-5}). The *galE* genes (*galE*(3) and *galE*(25)) were found to be much more abundant than both *mazG*(3) and *murB*(3), with a hit rate of approximately 1 per Mb of DNA screened, compared to approximately one per 4-6 Mb DNA for *mazG* and *murB* (Figure 2A, B and C).

Figure 2.


(A)
Mb DNA per hit
0
2
4
6
8
10
Host Individual
MH0001
MH0002
MH0003
MH0004
MH0005
MH0006
MH0007
MH0008
MH0009
MH0010
MH0011
MH0012
MH0013
MH0014
MH0015
MH0016
MH0017
MH0018
MH0019
MH0020
MH0021
MH0022
MH0023
MH0024
MH0025
MH0026
MH0027
MH0028
MH0030
MH0031
MH0032
MH0033
MH0034
MH0035
MH0036
MH0037
MH0038
MH0039
MH0040
MH0041
MH0042
MH0043
MH0044
MH0045
MH0046
MH0047
MH0048
MH0049
MH0050
MH0051
MH0052
MH0053
MH0054
MH0055
MH0056
MH0057
MH0058
MH0059
MH0060
MH0061
MH0062
MH0063
MH0064
MH0065
MH0066
MH0067
MH0068
MH0069
MH0070
MH0071
MH0072
MH0073
MH0074
MH0075
MH0076
MH0077
MH0078
MH0079
MH0080
MH0081
MH0082
MH0083
MH0084
MH0085
MH0086


(B)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Mb DNA per hit
MH0001
MH0002
MH0003
MH0004
MH0005
MH0006
MH0007
MH0008
MH0009
MH0010
MH0011
MH0012
MH0013
MH0014
MH0015
MH0016
MH0017
MH0018
MH0019
MH0020
MH0021
MH0022
MH0023
MH0024
MH0025
MH0026
MH0027
MH0028
MH0030
MH0031
MH0032
MH0033
MH0034
MH0035
MH0036
MH0037
MH0038
MH0039
MH0040
MH0041
MH0042
MH0043
MH0044
MH0045
MH0046
MH0047
MH0048
MH0049
MH0050
MH0051
MH0052
MH0053
MH0054
MH0055
MH0056
MH0057
MH0058
MH0059
MH0060
MH0061
MH0062
MH0063
MH0064
MH0065
MH0066
MH0067
MH0068
MH0069
MH0070
MH0071
MH0072
MH0073
MH0074
MH0075
MH0076
MH0077
MH0078
MH0079
MH0080
MH0081
MH0082
MH0083
MH0084
MH0085
MH0086
Host individual

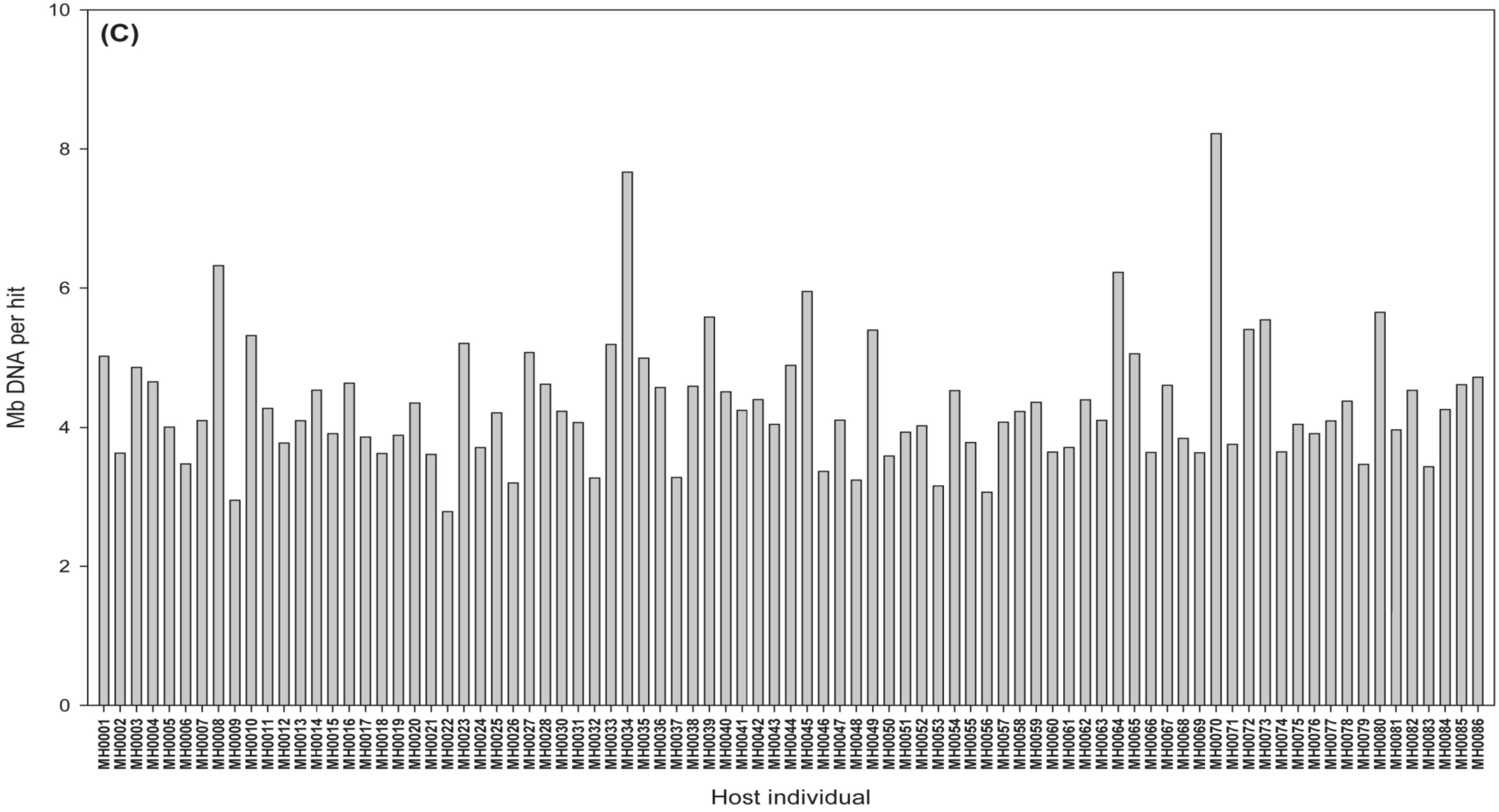
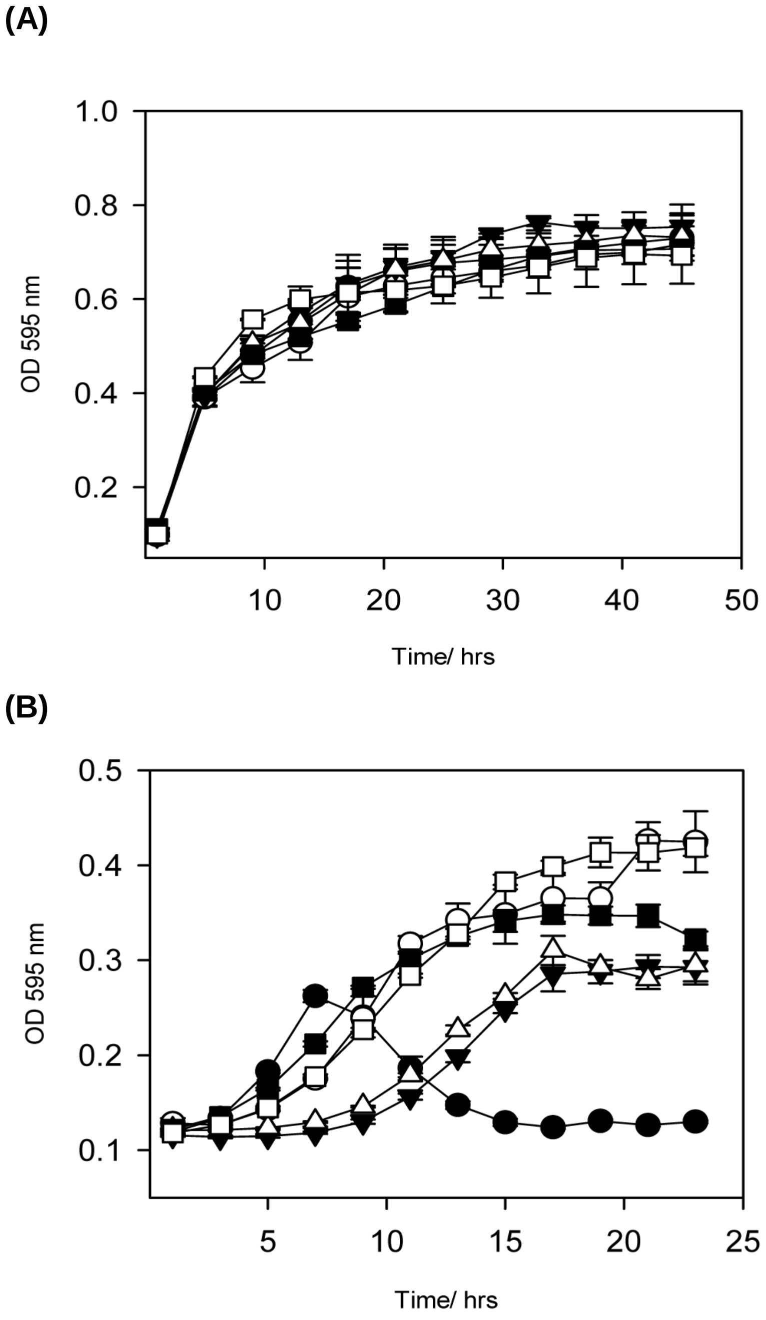


Figure 2. Relative abundance of the genes identified in this study when BLASTed against the healthy individuals in the MetaHit dataset. **(A)** *galE*(3) = black; *murB*(3) = light grey; *mazG*(3) = dark grey, **(B)** *galE*(25) and **(C)** *mazG*(5).

Growth experiments

Three genes were identified by transposon mutagenesis from SMG 3, namely *galE*(3), *murB*(3) and *mazG*(3) and two further genes were identified through bioinformatic analysis; *mazG*(5) from SMG 5 and *galE*(25) from SMG 25. The genes were cloned with some flanking DNA into the shuttle plasmid pCI372 and transformed into electro-competent *E. coli* MKH13. Growth experiments in LB and LB supplemented with NaCl or KCl (ionic osmotic stress) were conducted on MKH13::pCI372-*mazG*(3)*,* MKH13::pCI372-*murB*(3), MKH13::pCI372-*galE*(3), MKH13::pCI372-*mazG*(5) and MKH13::pCI372-*galE*(25). Each of the five transformed clones showed a statistically significant increase in salt tolerance (to both NaCl and KCl) compared to control strain MKH13 (Figure 3B and 3C), while no difference in growth was observed in LB alone (Figure 3A). Each of the five transformed clones was also tested in their ability to grow under conditions of non-ionic osmotic stress (i.e. sucrose or glycerol). An increased growth phenotype was not observed under these conditions for any of the clones (Figure 4), indicating these genes may confer a salt-specific protective effect.

Figure 3.



(A)
1.0
0.8
0.6
0.4
0.2
OD 595 nm
10
20
30
40
50
Time/ hrs
(B)
0.5
0.4
0.3
0.2
0.1
OD 595 nm
5
10
15
20
25
Time/ hrs

(C)

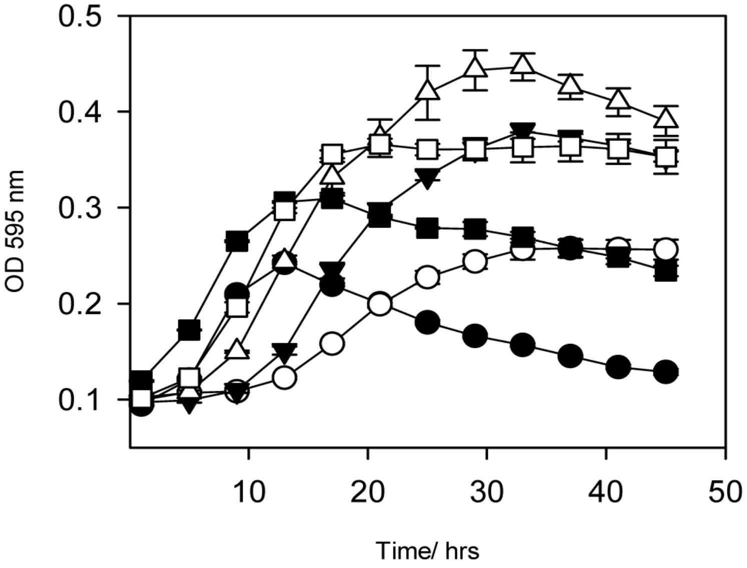


Figure 3. Growth in LB and LB supplemented with 3% NaCl or 4% KCl*. Growth of control strain MKH13::pCI372 (● closed circle), MKH13::pCI372-*galE*(3) (*P*<0.006) (*P*=0.0004) (○ open circle), MKH13::pCI372-*murB*(3) (*P*<0.0001) (*P*<0.0001) (△ open triangle), MKH13::pCI372-*mazG*(3) (*P*<0.0001) (*P*=0.0017) (▼ closed triangle), MKH13::pCI372-*galE*(25) (*P*=0.0002) (*P*=0.0012) (■ closed square) and MKH13::pCI372-*mazG*(5) (*P*=0.0003) (*P*<0.0001) (□ open square) in **(A)** LB broth, **(B)** LB broth supplemented with 3% NaCl and **(C)** LB broth supplemented with 4% KCl.

* The first *P*-value in parentheses represents growth in NaCl, while the second *P*-value represents growth in KCl.

Figure 4.

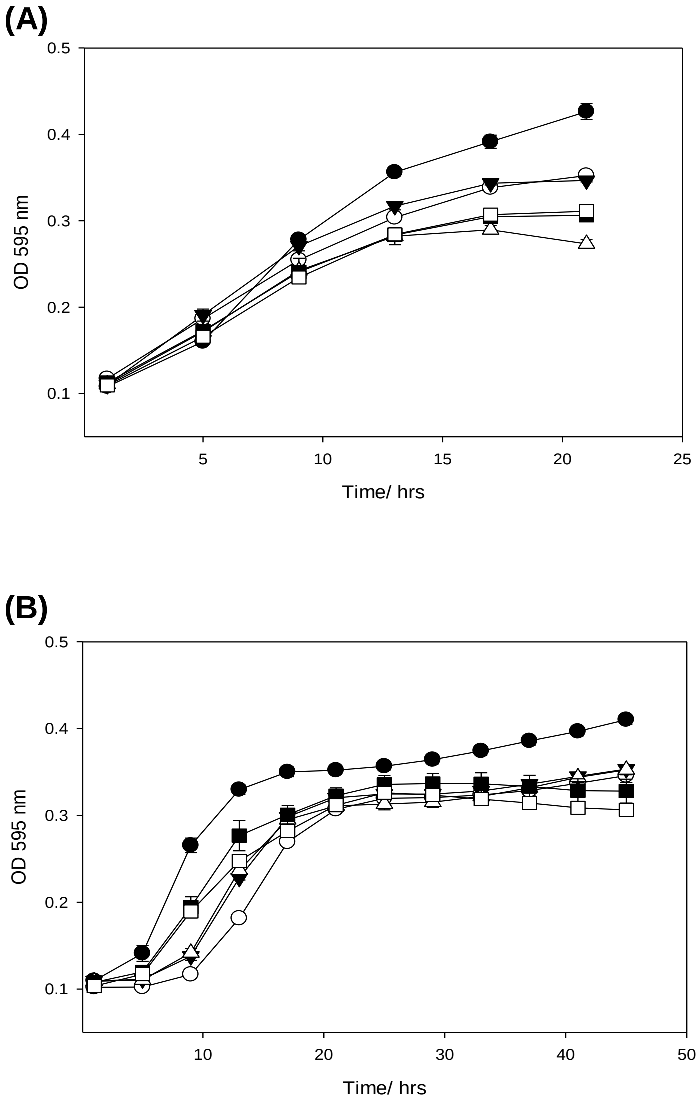
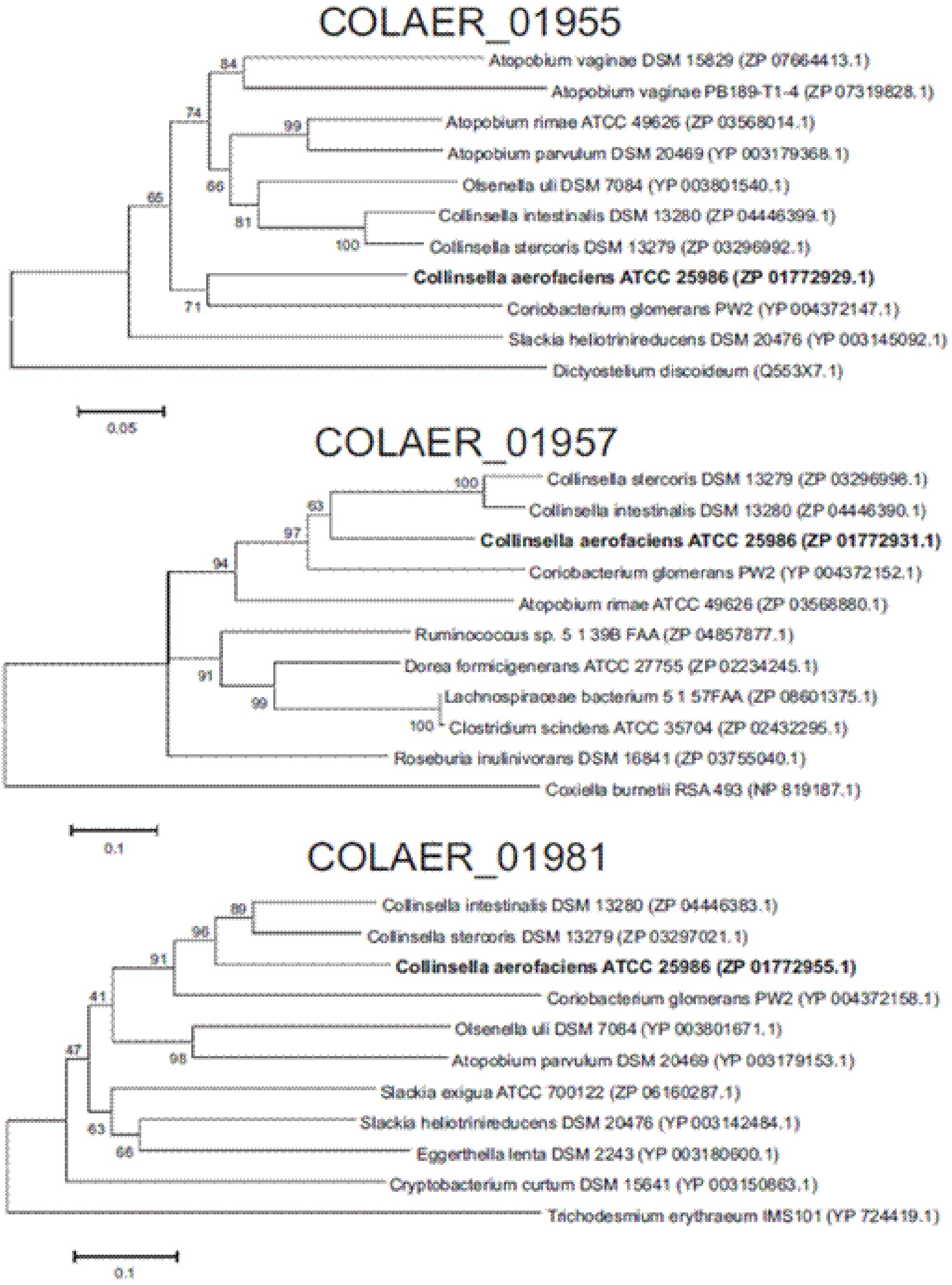


Figure 4. Growth of control strain MKH13::pCI372 (● closed circle), MKH13::pCI372-*galE*(3)*,* (○ open circle), MKH13::pCI372-*murB*(3), (△ open triangle), MKH13::pCI372-*mazG*(3), (▼ closed triangle), MKH13::pCI372-*galE*(25), (■ closed square) and MKH13::pCI372-*mazG*(5), (□ open square) in **(A)** LB supplemented with 20% glycerol, **(B)** LB broth supplemented with 30% sucrose.

Figure 5.

(A)


COLAER_01955
84
Atopobium vaginae DSM 15829 (ZP 07664413.1)
Atopobium vaginae PB189-T1-4 (ZP 07319828.1)
74
99
Atopobium rimae ATCC 49626 (ZP 03568014.1)
Atopobium parvulum DSM 20469 (YP 003179368.1)
66
Olsenella uli DSM 7084 (YP 003801540.1)
65
81
Collinsella intestinalis DSM 13280 (ZP 04446399.1)
100
Collinsella stercoris DSM 13279 (ZP 03296992.1)
Collinsella aerofaciens ATCC 25986 (ZP 01772929.1)
71
Coriobacterium glomerans PW2 (YP 004372147.1)
Slackia heliotrinireducens DSM 20476 (YP 003145092.1)
Dictyostelium discoideum (Q553X7.1)
0.05
COLAER_01957
100
Collinsella stercoris DSM 13279 (ZP 03296998.1)
63
Collinsella intestinalis DSM 13280 (ZP 04446390.1)
97
Collinsella aerofaciens ATCC 25986 (ZP 01772931.1)
94
Coriobacterium glomerans PW2 (YP 004372152.1)
Atopobium rimae ATCC 49626 (ZP 03568880.1)
Ruminococcus sp. 5 1 39B FAA (ZP 04857877.1)
Dorea formicigenerans ATCC 27755 (ZP 02234245.1)
91
Lachnospiraceae bacterium 5 1 57FAA (ZP 08601375.1)
99
100
Clostridium scindens ATCC 35704 (ZP 02432295.1)
Roseburia inulinivorans DSM 16841 (ZP 03755040.1)
Coxiella burnetii RSA 493 (NP 819187.1)
0.1
COLAER_01981
89
Collinsella intestinalis DSM 13280 (ZP 04446383.1)
96
Collinsella stercoris DSM 13279 (ZP 03297021.1)
91
Collinsella aerofaciens ATCC 25986 (ZP 01772955.1)
41
Coriobacterium glomerans PW2 (YP 004372158.1)
Olsenella uli DSM 7084 (YP 003801671.1)
47
98
Atopobium parvulum DSM 20469 (YP 003179153.1)
Slackia exigua ATCC 700122 (ZP 06160287.1)
Slackia heliotrinireducens DSM 20476 (YP 003142484.1)
63
66
Eggerthella lenta DSM 2243 (YP 003180600.1)
Cryptobacterium curtum DSM 15641 (YP 003150863.1)
Trichodesmium erythraeum IMS101 (YP 724419.1)
0.1

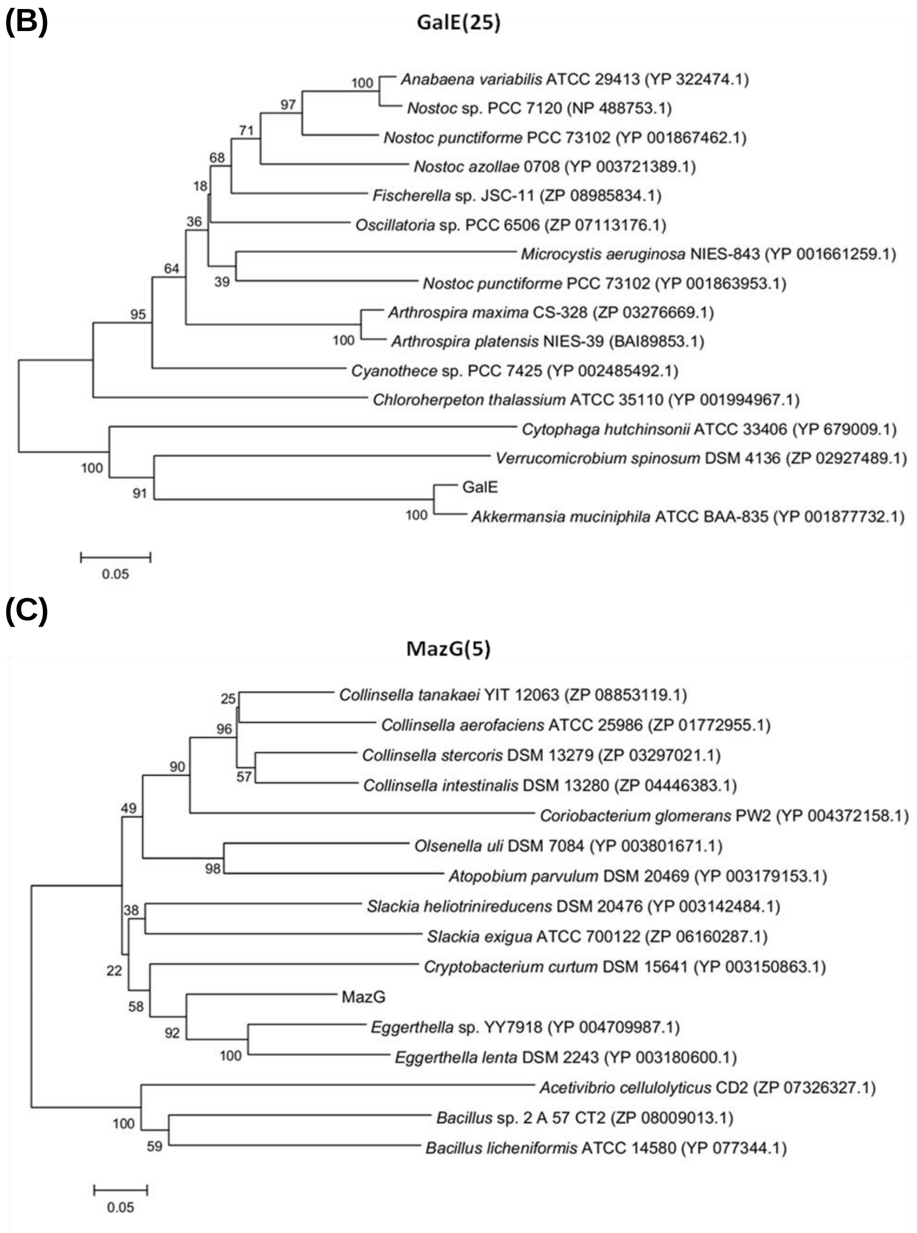

(B)
GalE(25)
Anabaena variabilis ATCC 29413 (YP 322474.1)
Nostoc sp. PCC 7120 (NP 488753.1)
Nostoc punctiforme PCC 73102 (YP 001867462.1)
Nostoc azollae 0708 (YP 003721389.1)
Fischerella sp. JSC-11 (ZP 08985834.1)
Oscillatoria sp. PCC 6506 (ZP 07113176.1)
Microcystis aeruginosa NIES-843 (YP 001661259.1)
Nostoc punctiforme PCC 73102 (YP 001863953.1)
Arthrospira maxima CS-328 (ZP 03276669.1)
Arthrospira platensis NIES-39 (BAI89853.1)
Cyanothece sp. PCC 7425 (YP 002485492.1)
Chloroherpeton thalassium ATCC 35110 (YP 001994967.1)
Cytophaga hutchinsonii ATCC 33406 (YP 679009.1)
Verrucomicrobium spinosum DSM 4136 (ZP 02927489.1)
GalE
Akkermansia muciniphila ATCC BAA-835 (YP 001877732.1)
100
97
71
68
18
36
64
39
95
100
100
91
100
0.05
(C)
MazG(5)
Collinsella tanakaei YIT 12063 (ZP 08853119.1)
Collinsella aerofaciens ATCC 25986 (ZP 01772955.1)
Collinsella stercoris DSM 13279 (ZP 03297021.1)
Collinsella intestinalis DSM 13280 (ZP 04446383.1)
Coriobacterium glomerans PW2 (YP 004372158.1)
Olsenella uli DSM 7084 (YP 003801671.1)
Atopobium parvulum DSM 20469 (YP 003179153.1)
Slackia heliotrinireducens DSM 20476 (YP 003142484.1)
Slackia exigua ATCC 700122 (ZP 06160287.1)
Cryptobacterium curtum DSM 15641 (YP 003150863.1)
MazG
Eggerthella sp. YY7918 (YP 004709987.1)
Eggerthella lenta DSM 2243 (YP 003180600.1)
Acetivibrio cellulolyticus CD2 (ZP 07326327.1)
Bacillus sp. 2 A 57 CT2 (ZP 08009013.1)
Bacillus licheniformis ATCC 14580 (YP 077344.1)
25
96
57
90
49
98
38
22
58
92
100
100
59
0.05

Figure 5. Phylogenetic analysis of identified genes. Phylogenetic analysis of **(A)** GalE(3) (COLAER_01955), MurB(3) (COLAER_01957), MazG(3) (COLAER_01981), **(B)** GalE(25) (Amuc_1125) and **(C)** MazG(5) (EGYY_03530) and related proteins. Phylogenetic analysis was performed using the program MEGA 5.0 (Tamura *et al.*, 2011). The bootstrap method (500 replicates) was used to test the percentage of replicate trees in which the associated taxa clustered together.

Discussion

Metagenomics has the potential to allow us to advance, or for the most part to begin, the study of the genetic complement of uncultured microbes. In the current study we used a combined functional metagenomic, transposon mutagenesis and bioinformatic strategy to screen a metagenomic library from the human gut microbiota for potential salt-tolerant clones and identified five genes (see Table 3.), namely *galE*(3) and *galE*(25), *mazG*(3) and *mazG*(5) and *murB*(3), involved in salt tolerance and likely to be important for life in the gut.

In the most comprehensive analysis to date of the genetic complement of the human gut microbiome, a cohort of over 1200 genes (termed range clusters) were identified which encode functions important for life in the gut (Qin *et al.*, 2010). These included genes which encoded proteins similar to those which we have identified in this study, namely; UDP glucose-4-epimerase, NTP pyrophosphohydrolase (for which MazG is a functional homologue), as well as a protein containing a tetrapyrrole methyltransferase domain and a MazG-like domain and UDP-*N*-acetylmuramate dehydrogenase (which is another name for MurB; UDP-*N*-acetylenolpyruvylglucosamine reductase) (Supplementary Table 10 from Qin *et al.*, 2010). The presence of the genes encoding these proteins in the enriched gene set leads us to conclude that they are important for survival in the gut and their putative role in adapting to fluctuating levels of osmotic stress in the gut. Also, homologues of the identified sequences from this study were found to be abundant in each of the individual metagenomes from the MetaHit dataset (Figure 2A, B and C). Furthermore, an analysis of human gut genomic and metagenomic datasets identified uncharacterized and novel protein families which are over-represented in the human gut (Ellrott *et al.*, 2010). Among these, and second on the list of over-represented protein families, are coiled-coil osmosensory transporters. The best characterized of the coiled-coil transporters is ProP of *E. coli*, which is responsible for the uptake of osmoprotectants such as proline and betaine during osmotic stress (Culham *et al.*, 1993). Such representative proteins were not found in any of the genomes analysed from microbes not associated with the human gut microbiome, indicating their importance to bacteria residing in the human gut. This observation also reinforces the idea that the human gut microbiota may employ novel strategies and possess novel mechanisms for osmoadaptation.

Transposon mutagenesis resulted in the identification of a gene (*galE*) from SMG 3 (*galE*(3)), encoding a UDP glucose 4-epimerase (GalE) (also termed UDP galactose 4-epimerase), while another *galE* gene was bioinformatically identified from SMG 25 (*galE*(25)). This enzyme catalyses the direct inter-conversion of UDP glucose and UDP galactose (Leloir, 1951), which are precursors involved in the synthesis of capsular polysaccharide (CPS) and the compatible solute trehalose as well as the synthesis of lipopolysaccharide (LPS) and membrane-derived oligosaccharides (MDO’s) in Gram negative bacteria and lipoteichoic acids in Gram positive bacteria (Fukasawa *et al.*, 1962; Giaever *et al.*, 1988; Grundling and Schneewind, 2007; Markovitz, 1977; Schulman and Kennedy, 1977; Seltman and Holst, 2002).

The *galE* gene is often found in an operon with the *galT* and *galK* genes for galactose metabolism. In SMG 3 (highest genetic identity to *C. aerofaciens*), the *galE* gene (COLAER_01955) is found not on a *gal* operon, but between two genes encoding hypothetical proteins similar to a peptidase and a 4-alpha-glucanotransferase (COLAER_01956 and COLAER_01953, respectively). COLAER_01957 (*murB*) is located 226 nucleotides downstream of COLAER_01956. *C. aerofaciens*' *galE* may still be involved in galactose metabolism as the organism can utilize galactose (Kageyama *et al.*, 1999). The *galE* gene from SMG 25 is not found on the galactose operon and also exhibits a different genomic arrangement to SMG 3. This gene, *galE*(25), which shares 96% identity to *A. muciniphila* (Amuc_1125), is located between a UDP-galactopyranose mutase homolog from *Chthoniobacter flavus* (52% identity) and a gene encoding a hypothetical protein similar to Amuc_1123 from *A. muciniphila*. Furthermore, the fosmid insert of SMG 25 has a vastly different genomic arrangement to that of *A. muciniphila*. This lack of synteny indicates that the fosmid insert is not from *A. muciniphila*, but from a related but as yet undiscovered species. Indeed, sequences representing at least eight uncharacterized species of the genus *Akkermansia* have been identified from different gut metagenomic libraries (van Passel *et al.,* 2011).. Interestingly, a UDP glucose 4-epimerase homolog has been identified as a putative gene involved in drought tolerance in rice by modulating the ability of rice roots to penetrate deeper into the substratum when exposed to drought conditions (Nguyen *et al.*, 2004).

A transposon insertion in the *galE* could potentially affect the cell in different ways; (1) by causing compositional changes in the LPS layer or in lipoteichoic acids of Gram negative and Gram positive bacteria, respectively, thus reducing the cell's ability to sense, respond to or resist various environmental stress conditions. (2) by disrupting the inter-conversion of UDP-glucose and UDP-galactose. UDP glucose may be a key molecule as it can be converted to the osmoprotectant trehalose (Giaever *et al.*, 1988), in addition to potentially modulating the expression of RpoS (which increases expression of a number of genes at high salt concentrations (Bohringer *et al.*, 1995; Hengge-Aronis *et al.*,1991).

Expressing this *galE* gene in *E. coli* may result in increased UDP glucose levels which could be converted to trehalose resulting in an osmoprotective effect. Padilla and co-workers increased the UDP glucose supply and consequently the levels of accumulated trehalose by expressing the *E. coli galU* gene in *Corynebacterium glutamicum* (Padilla *et al.*, 2004). In our study, the expression of both the *galE*(3) and *galE*(25) genes in the osmosensitive strain *E. coli* MKH13 resulted in a statistically significant increased salt (NaCl and KCl) tolerance phenotype (Figure 3B and 3C). Of the five genes identified and cloned, *galE*(3) had a lesser effect than the other four genes. MKH13::pCI372-*galE*(3) also had a prolonged lag phase. It has been demonstrated that cellular stress may be caused by imbalances and accumulation of certain intermediary metabolites, particularly in the case of amphibolic pathways (e.g. D-galactose pathway), as demonstrated in a *galE* mutant (Lee *et al.*, 2009). In the current study, supplying MKH13 with an additional plasmid encoded copy of *galE* may also cause an imbalance in an intermediary metabolite, leading to stress and ultimately a prolonged lag phase, which ends when the imbalance is corrected and homeostasis is restored.

Peptidoglycan is a major component of the bacterial cell wall and plays an important role in withstanding osmotic stress (van Heijenoort, 1996). MurB is essential for cell wall biosynthesis and is involved in a two-step process with UDP-*N*-acetylglucosamine enolpyruvyl tranferase (MurA) to form UDP-*N*-acetylmuramate, which is a building block for peptidoglycan (Sylvester *et al.*, 2001). The *murB* gene has been shown to be essential for normal growth in *Bacillus subtilis*, *E. coli* and *Staphylococcus aureus* (Matsuo *et al.*, 2003; Miyakawa *et al.*, 1972; Real and Henriques, 2006). In *S. aureus*, a mutation within the *murB* gene resulted in thermosensitive mutants which had thinner cell walls (Matsuo *et al.*, 2003). The *murB*(3) gene conferred an increased growth phenotype to *E. coli* MKH13 during both NaCl and KCl stress (Figure 3B and 3C). Disruption or deletion of the *murB* gene could make cells acutely sensitive to osmotic stress due to a reduction in cell wall integrity as well as causing a reduction in turgor pressure which is a driving force for cellular growth and division. Bacteria remodel the structure of their peptidoglycan in response to changes in environmental conditions (Quintela *et al.*, 1997; Vijaranakul *et al.*, 1995), which could be important in the gut by allowing for varying levels of rigidity or elasticity depending on the conditions in the immediate environment.

We have also identified two putative MazG family proteins (encoded by the *mazG* gene from SMG 3 and SMG 5 (*mazG*(3) and *mazG*(5), respectively). These are highly conserved proteins in bacteria and to date there has been no tangible link demonstrated between MazG and salt tolerance. MazG is a nucleoside triphosphate pyrophosphohydrolase (NTPase) which can hydrolyse (deoxy)ribonucleoside triphosphates ((d)NTPs) to their corresponding (deoxy)ribonucleoside monophosphates ((d)NMPs) and pyrophosphate (PPi) (Zhang and Inouye, 2002). It has been proposed that MazG plays a role in cellular “house-cleaning” by removing abnormal (d)NTP’s from nascent DNA strands (Galperin *et al.*, 2006), in addition to regulating programmed cell (PCD) death in *E. coli* (Gross *et al.*, 2006). It also regulates intracellular levels of (p)ppGpp, the main nutritional stress signal molecule involved in the stringent response, which has been shown to be an important response in *Campylobacter jejuni* during intestinal colonization (Stintzi *et al.*, 2005). By reducing (p)ppGpp levels, MazG plays a central role in maintaining cell viability during nutritional stress, which could be important in the gut during periods of intermittent availability of certain nutrients, while (p)ppGpp itself is required for growth during osmotic stress in *L. monocytogenes* (Okada *et al.*, 2002). Moreover, recent work has identified a role for MazG in the mycobacterial oxidative stress response and virulence (Lu *et al.*, 2010; Sassetti and Rubin, 2003). As outlined above MazG plays a number of different roles in different bacteria, but it is clear that stress response is a common theme. This study provides evidence for a novel role of MazG in salt tolerance. In the context of the gut environment, MazG may be important in removal of mutagenic nucleotides from growing DNA strands. Damage to bacterial DNA is likely due to exposure to genotoxic compounds such as nitrosamines and heterocyclic amines (Kurokawa *et al.*, 2007). In addition MazG could provide energy to the cell during stress through the hydrolysis of ATP. Due to the numerous roles of MazG in different stress conditions it may function as general stress response protein in the bacterial cell. More research and discovery of novel MazG proteins will help identify new physiological roles, substrates and precise mechanisms of action for these proteins in the myriad of stress conditions imposed on microorganisms.

In conclusion, we have identified five genetic loci involved in salt tolerance from within the human gut microbiome using a functional metagenomic approach. The genes represent two different homologs of *galE* and two of *mazG*, as well as a *murB* homolog, from three different species of the genus' *Collinsella, Akkermansia* and *Eggerthella*. The identification of three genes within approximately 40kb of metagenomic DNA from SMG 3 (highest genetic identity to *C. aerofaciens*) is relevant in that functionally related proteins are often co-located on the chromosome of prokaryotic genomes (Sleator and Walsh, 2010; Sleator, 2011). In addition to expanding our knowledge of salt tolerance mechanisms, this study may also facilitate the development of novel drug targets and related approaches to control resident gut microbiota (Sleator 2010a, Sleator 2010b). Ultimately some of the salt tolerance mechanisms identified might be used as part of a “patho-biotechnology” (Sleator and Hill, 2006) or “meta-biotechnology” (Culligan *et al.*, 2009) strategy for the design of improved probiotic cultures with greater resistance to process induced stresses (such as spray and freeze drying), as well as improved gut colonization (see (Sheehan *et al.*, 2006; Sheehan *et al.*, 2007; Watson *et al.*, 2008) for such examples). This will hopefully result in a broader and more comprehensive representation of salt tolerance mechanisms in this unique environment.

References

Banik JJ, Brady SF (2008). Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. *Proc Natl Acad Sci U S A* **105:** 17273-7.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al* (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902-6.

Bohringer J, Fischer D, Mosler G, Hengge-Aronis R (1995). UDP-glucose is a potential intracellular signal molecule in the control of expression of sigma S and sigma S-dependent genes in Escherichia coli. *J Bacteriol* **177:** 413-22.

Chowdhury R, Sahu G, Das J (1996). Stress response in pathogenic bacteria. *Journal of Biosciences* **21:** 149-160.

Culham DE, Lasby B, Marangoni AG, Milner JL, Steer BA, van Nues RW *et al* (1993). Isolation and sequencing of Escherichia coli gene proP reveals unusual structural features of the osmoregulatory proline/betaine transporter, ProP. *J Mol Biol* **229:** 268-76.

Culligan EP, Hill C, Sleator RD (2009). Probiotics and gastrointestinal disease: successes, problems and future prospects. *Gut Pathog* **1:** 19.

Ellrott K, Jaroszewski L, Li W, Wooley JC, Godzik A (2010). Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. *PLoS Comput Biol* **6:** e1000798.

Epstein W (2003). The roles and regulation of potassium in bacteria. *Prog Nucleic Acid Res Mol Biol* **75:** 293-320.

Felsenstein J (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783-791.

Fukasawa T, Jokura K, Kurahashi K (1962). A new enzymic defect of galactose metabolism in Escherichia coli K-12 mutants. *Biochem Biophys Res Commun* **7:** 121-5.

Galperin MY, Moroz OV, Wilson KS, Murzin AG (2006). House cleaning, a part of good housekeeping. *Mol Microbiol* **59:** 5-19.

Gardan R, Duche O, Leroy-Setrin S, Labadie J (2003). Role of ctc from Listeria monocytogenes in osmotolerance. *Appl Environ Microbiol* **69:** 154-61.

Giaever HM, Styrvold OB, Kaasen I, Strom AR (1988). Biochemical and genetic characterization of osmoregulatory trehalose synthesis in Escherichia coli. *J Bacteriol* **170:** 2841-9.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR *et al* (2002). Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68:** 4301-6.

Gloux K, Berteau O, El Oumami H, Beguet F, Leclerc M, Dore J (2011). A metagenomic beta-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proc Natl Acad Sci U S A* **108 Suppl 1:** 4539-46.

Gralla JD, Huo YX (2008). Remodeling and activation of Escherichia coli RNA polymerase by osmolytes. *Biochemistry* **47:** 13189-96.

Gross M, Marianovsky I, Glaser G (2006). MazG -- a regulator of programmed cell death in Escherichia coli. *Mol Microbiol* **59:** 590-601.

Grundling A, Schneewind O (2007). Genes required for glycolipid synthesis and lipoteichoic acid anchoring in Staphylococcus aureus. *J Bacteriol* **189:** 2521-30.

Haardt M, Kempf B, Faatz E, Bremer E (1995). The osmoprotectant proline betaine is a major substrate for the binding-protein-dependent transport system ProU of Escherichia coli K-12. *Mol Gen Genet* **246**: 783-786

Hayes F, Daly C, Fitzgerald GF (1990). Identification of the minimal replicon of *Lactococcus lactis* subsp. lactis UC317 plasmid pCI305. *Appl Environ Microbiol* **56:** 202-209

Heath C, Hu XP, Cary SC, Cowan D (2009). Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. *Appl Environ Microbiol* **75:** 4657-9.

Hengge-Aronis R, Klein W, Lange R, Rimmele M, Boos W (1991). Trehalose synthesis genes are controlled by the putative sigma factor encoded by rpoS and are involved in stationary-phase thermotolerance in Escherichia coli. *J Bacteriol* **173:** 7918-24.

Jones BV, Marchesi JR (2007). Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat Methods* **4:** 55-61.

Kageyama A, Benno Y, Nakase T (1999). Phylogenetic and phenotypic evidence for the transfer of Eubacterium aerofaciens to the genus Collinsella as Collinsella aerofaciens gen. nov., comb. nov. *Int J Syst Bacteriol* **49 Pt 2:** 557-65.

Kapardar R, Ranjan R, Puri M, Sharma R (2010(b)). Sequence analysis of a salt tolerant metagenomic clone. *Indian Journal of Microbiology* **50:** 212-215.

Kapardar RK, Ranjan R, Grover A, Puri M, Sharma R (2010(a)). Identification and characterization of genes conferring salt tolerance to Escherichia coli from pond water metagenome. *Bioresour Technol* **101:** 3917-24.

Kazimierczak KA, Rincon MT, Patterson AJ, Martin JC, Young P, Flint HJ *et al* (2008). A new tetracycline efflux gene, tet(40), is located in tandem with tet(O/32/O) in a human gut firmicute bacterium and in metagenomic library clones. *Antimicrob Agents Chemother* **52:** 4001-9.

Kempf B, Bremer E (1998). Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. *Arch Microbiol* **170:** 319-30.

Kilstrup M, Jacobsen S, Hammer K, Vogensen FK (1997). Induction of heat shock proteins DnaK, GroEL, and GroES by salt stress in Lactococcus lactis. *Appl Environ Microbiol* **63:** 1826-37.

Kunte HJ (2006). Osmoregulation in Bacteria: Compatible Solute Accumulation and Osmosensing. *Environmental Chemistry* **3:** 94-99.

Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A *et al* (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14:** 169-81.

Lakhdari O, Cultrone A, Tap J, Gloux K, Bernard F, Ehrlich SD *et al* (2010). Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-kappaB modulation in the human gut. *PLoS One* **5**.

Lee DG, Jeon JH, Jang MK, Kim NY, Lee JH, Kim SJ *et al* (2007). Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library. *Biotechnol Lett* **29:** 465-72.

Lee SJ, Trostel A, Le P, Harinarayanan R, Fitzgerald PC, Adhya S (2009). Cellular stress created by intermediary metabolite imbalances. *Proc Natl Acad Sci U S A* **106:** 19515-20.

Leloir LF (1951). The enzymatic transformation of uridine diphosphate glucose into a galactose derivative. *Arch Biochem Biophys* **33:** 186-90.

Lopez CS, Alice AF, Heras H, Rivas EA, Sanchez-Rivas C (2006). Role of anionic phospholipids in the adaptation of Bacillus subtilis to high salinity. *Microbiology* **152:** 605-16.

Lu LD, Sun Q, Fan XY, Zhong Y, Yao YF, Zhao GP (2010). Mycobacterial MazG is a novel NTP pyrophosphohydrolase involved in oxidative stress response. *J Biol Chem* **285:** 28076-85.

Maniatis T, Fritsch, E. F. & Sambrook, J (1982). *Molecular Cloning, A Laboratory Manual.*: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Markovitz A. (1977). *Surface carbohydrates of the prokaryotic cell*. W. I and Sutherland. (eds). Academic Press: New York, pp 415-462.

Matsuo M, Kurokawa K, Nishida S, Li Y, Takimura H, Kaito C *et al* (2003). Isolation and mutation site determination of the temperature-sensitive murB mutants of Staphylococcus aureus. *FEMS Microbiol Lett* **222:** 107-13.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC *et al* (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4:** 495-500.

Meilleur C, Hupe JF, Juteau P, Shareck F (2009). Isolation and characterization of a new alkali-thermostable lipase cloned from a metagenomic library. *J Ind Microbiol Biotechnol* **36:** 853-61.

Miyakawa T, Matsuzawa H, Matsuhashi M, Sugino Y (1972). Cell wall peptidoglycan mutants of Escherichia coli K-12: existence of two clusters of genes, mra and mrb, for cell wall peptidoglycan biosynthesis. *J Bacteriol* **112:** 950-8.

Nguyen TT, Klueva N, Chamareck V, Aarti A, Magpantay G, Millena AC *et al* (2004). Saturation mapping of QTL regions and identification of putative candidate genes for drought tolerance in rice. *Mol Genet Genomics* **272:** 35-46.

Okada Y, Makino S, Tobe T, Okada N, Yamazaki S (2002). Cloning of rel from Listeria monocytogenes as an osmotolerance involvement gene. *Appl Environ Microbiol* **68:** 1541-7.

Padilla L, Morbach S, Kramer R, Agosin E (2004). Impact of heterologous expression of Escherichia coli UDP-glucose pyrophosphorylase on trehalose and glycogen synthesis in Corynebacterium glutamicum. *Appl Environ Microbiol* **70:** 3845-54.

Piuri M, Sanchez-Rivas C, Ruzal SM (2003). Adaptation to high salt in Lactobacillus: role of peptides and proteolytic enzymes. *J Appl Microbiol* **95:** 372-9.

Pumbwe L, Wareham DW, Aduse-Opoku J, Brazier JS, Wexler HM (2007). Genetic analysis of mechanisms of multidrug resistance in a clinical isolate of Bacteroides fragilis. *Clin Microbiol Infect* **13:** 183-9.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C *et al* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464:** 59-65.

Quintela JC, de Pedro MA, Zollner P, Allmaier G, Garcia-del Portillo F (1997). Peptidoglycan structure of Salmonella typhimurium growing within cultured mammalian cells. *Mol Microbiol* **23:** 693-704.

Real G, Henriques AO (2006). Localization of the Bacillus subtilis murB gene within the dcw cluster is important for growth and sporulation. *J Bacteriol* **188:** 1721-32.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-25.

Sakamoto T, Murata N (2002). Regulation of the desaturation of fatty acids and its role in tolerance to cold and salt stress. *Curr Opin Microbiol* **5:** 208-10.

Sassetti CM, Rubin EJ (2003). Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A* **100:** 12989-94.

Schiller C, Frohlich CP, Giessmann T, Siegmund W, Monnikes H, Hosten N *et al* (2005). Intestinal fluid volumes and transit of dosage forms as assessed by magnetic resonance imaging. *Aliment Pharmacol Ther* **22:** 971-9.

Schulman H, Kennedy EP (1977). Identification of UDP-glucose as an intermediate in the biosynthesis of the membrane-derived oligosaccharides of Escherichia coli. *J Biol Chem* **252:** 6299-303.

Seltman G, Holst O (2002). Further cell wall components of Gram-positive bacteria. *Bacterial Cell Wall*. Springer, Germany
pp 131-161.

Shabala L, Bowman J, Brown J, Ross T, McMeekin T, Shabala S (2009). Ion transport and osmotic adjustment in Escherichia coli in response to ionic and non-ionic osmotica. *Environ Microbiol* **11:** 137-48.

Sheehan VM, Sleator RD, Fitzgerald GF, Hill C (2006). Heterologous expression of BetL, a betaine uptake system, enhances the stress tolerance of Lactobacillus salivarius UCC118. *Appl Environ Microbiol* **72:** 2170-7.

Sheehan VM, Sleator RD, Hill C, Fitzgerald GF (2007). Improving gastric transit, gastrointestinal persistence and therapeutic efficacy of the probiotic strain Bifidobacterium breve UCC2003. *Microbiology* **153:** 3563-71.

Sleator RD, Hill C (2002). Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS Microbiol Rev* **26:** 49-71.

Sleator RD, Hill C (2005). A novel role for the LisRK two-component regulatory system in listerial osmotolerance. *Clin Microbiol Infect* **11:** 599-601.

Sleator RD, Hill C (2006). Patho-biotechnology: using bad bugs to do good things. *Curr Opin Biotechnol* **17:** 211-6.

Sleator RD, Shortall C, Hill C (2008). Metagenomics. *Lett Appl Microbiol* **47:** 361-6.

Sleator RD (2010a). Probiotic therapy - recruiting old friends to fight new foes. *Gut Pathog* **2:** 5.

Sleator RD (2010b). Probiotics -- a viable therapeutic alternative for enteric infections especially in the developing world. *Discov Med* **10:** 119-124.

Sleator RD, Walsh P (2010). An overview of in silico protein function prediction. *Arch Microbiol* **192:** 151-5.

Sleator RD (2011). Phylogenetics. *Arch Microbiol* **193:** 235-239.

Stintzi A, Marlow D, Palyada K, Naikare H, Panciera R, Whitworth L *et al* (2005). Use of genome-wide expression profiling and mutagenesis to study the intestinal lifestyle of Campylobacter jejuni. *Infect Immun* **73:** 1797-810.

Sylvester DR, Alvarez E, Patel A, Ratnam K, Kallender H, Wallis NG (2001). Identification and characterization of UDP-N-acetylenolpyruvylglucosamine reductase (MurB) from the Gram-positive pathogen Streptococcus pneumoniae. *Biochem J* **355:** 431-5.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol*.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-80.

van Heijenoort J (1996). Murein synthesis. In: Neidhardt FC, Curtis, R., III., Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaecter, M. and Umbarger, H.E (ed). *Escherichia coli and Salmonella: cellular and molecular biology*. American Society for Microbiology: Washington, D.C. pp 1025-1034.

van Passel MW, Kant R, Zoetendal EG, Plugge CM, Derrien M, Malfatti SA *et al* (2011). The genome of Akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One* **6:** e16876.

Vijaranakul U, Nadakavukaren MJ, de Jonge BL, Wilkinson BJ, Jayaswal RK (1995). Increased cell size and shortened peptidoglycan interpeptide bridge of NaCl-stressed Staphylococcus aureus and their reversal by glycine betaine. *J Bacteriol* **177:** 5116-21.

Watson D, Sleator RD, Hill C, Gahan CG (2008). Enhancing bile tolerance improves survival and persistence of Bifidobacterium and Lactococcus in the murine gastrointestinal tract. *BMC Microbiol* **8:** 176.

Wonderling LD, Wilkinson BJ, Bayles DO (2004). The htrA (degP) gene of Listeria monocytogenes 10403S is essential for optimal growth under stress conditions. *Appl Environ Microbiol* **70:** 1935-43.

Zhang J, Inouye M (2002). MazG, a nucleoside triphosphate pyrophosphohydrolase, interacts with Era, an essential GTPase in Escherichia coli. *J Bacteriol* **184:** 5323-9.

Zuckerkandl E, Pauling L (1965). Evolutionary divergence and convergence in proteins. In: Bryson V and Vogel HJ (eds). *Evolving Genes and Proteins*. Academic Press: New York. pp pp. 97-166.

CHAPTER III

ADDENDUM TO CHAPTER II

Mining the Human Gut Microbiome for Novel Stress Resistance Genes

ADDENDUM TO: CULLIGAN EP, SLEATOR RD, MARCHESI JR AND HILL C. FUNCTIONAL METAGENOMICS REVEALS NOVEL SALT TOLERANCE LOCI FROM THE HUMAN GUT MICROBIOME. *ISME J. 2012 OCT;6(10):1916-25*

GUT MICROBES. 2012 JUL-AUG;3(4):394-7

Abstract

With the rapid advances in sequencing technologies in recent years, the human genome is now considered incomplete without the complementing microbiome, which outnumbers human genes by a factor of one hundred. The human microbiome, and more specifically the gut microbiome, has received considerable attention and research efforts over the past decade. Many studies have identified and quantified “who is there?”, while others have determined some of their functional capacity, or “what are they doing?” In a recent study, we identified novel salt-tolerance loci from the human gut microbiome using combined functional metagenomic and bioinformatics based approaches. Herein, we discuss the identified loci, their role in salt-tolerance and their importance in the context of the gut environment. We also consider the utility and power of functional metagenomics for mining such environments for novel genes and proteins, as well as the implications and possible applications for future research.

Introduction

Bacteria encounter numerous environmental stresses in various environments and the gastrointestinal tract is no exception.[1] This dynamic environment poses a set of challenges that both transient and symbiotic microorganisms must overcome in order to colonise and proliferate.[2] Low pH, bile acids, elevated osmolarity, iron limitation, intermittent nutrient availability and host immune factors are just some of the challenges faced in the gastrointestinal tract.[3] The ability to cope with rapid changes in external osmolarity is an important mechanism that allows microorganisms adapt to and colonise a given environmental niche.[4] The cellular response to hyper-osmotic stress is broad and involves a number of different processes such as potassium (K^+) uptake[5], compatible solute accumulation[6] and numerous ancillary systems.[7, 8]

The emergence of metagenomics as a key area of scientific research in recent years has transformed how we view ourselves as living organisms.[9, 10] Working with microbes in pure cultures is very reductive in terms of understanding microbial behaviour in complex ecological niches. Metagenomics can, in principle, allow us access the entire genetic complement of our associated microbiome without the need for classic microbiological culturing techniques. Despite recent advances in high-throughput anaerobic culturing techniques with gnotobiotic animal husbandry, which suggests that the human faecal microbiota consists largely of taxa and predicted functions that are represented in its cultured members,[11] functional metagenomics allows us to rapidly separate the 'wheat from the chaff'. Gaining insights about the functional capacity of the human gut microbiome was a key aim in our efforts to elucidate novel genetic loci conferring a salt tolerance phenotype. We employed a functional metagenomic screen of a human gut metagenomic library which resulted in the identification of five novel genes involved in salt tolerance.[12] An advantage of functional metagenomics is its ability to uncover completely novel functions for new or known genes without the need for any previous sequence information.

Figure 1.

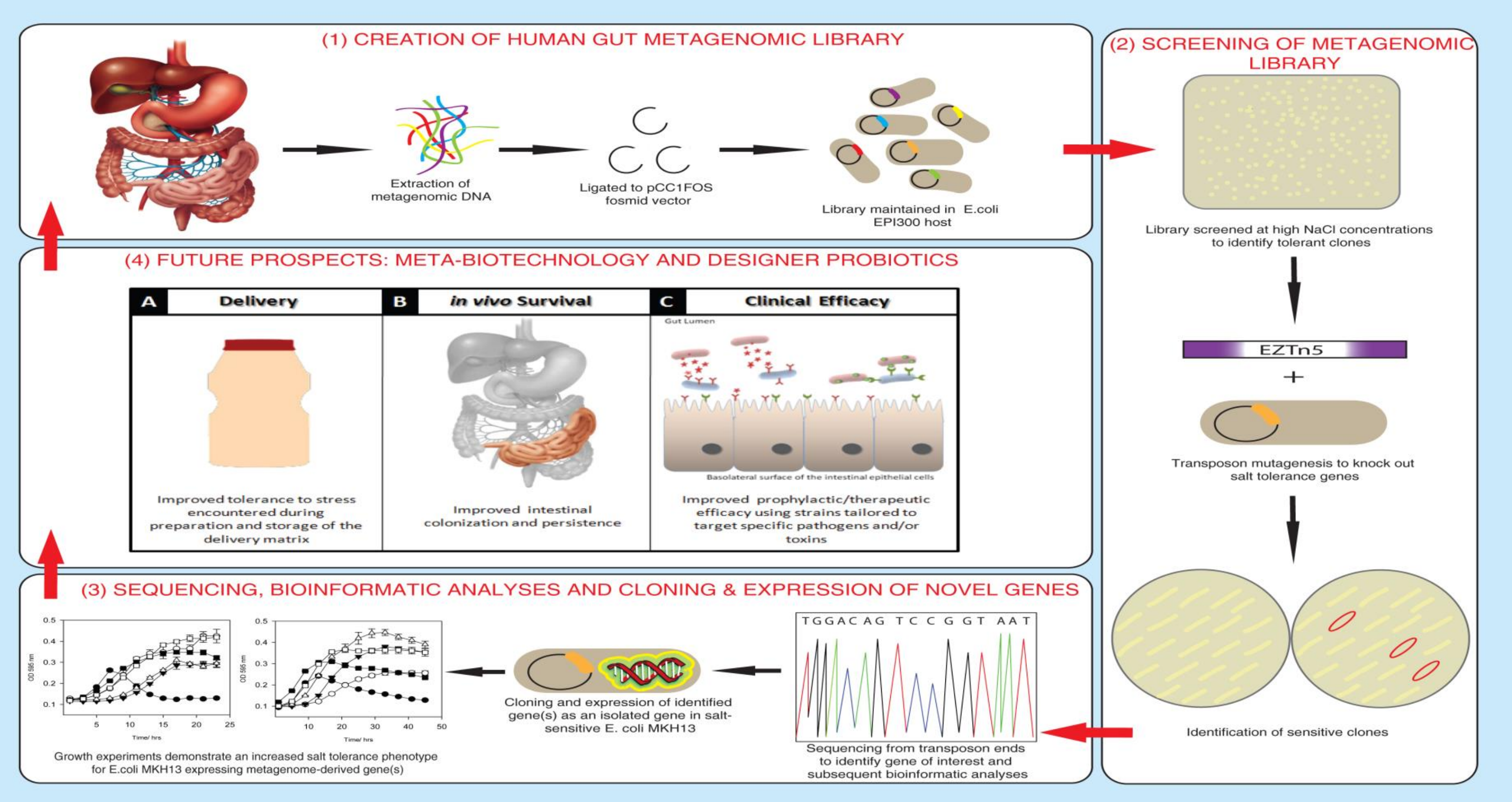

(1) CREATION OF HUMAN GUT METAGENOMIC LIBRARY
Extraction of metagenomic DNA
Ligated to pCC1FOS fosmid vector
Library maintained in E.coli EPI300 host
(2) SCREENING OF METAGENOMIC LIBRARY
Library screened at high NaCl concentrations to identify tolerant clones
EZTn5
+
Transposon mutagenesis to knock out salt tolerance genes
Identification of sensitive clones
(3) SEQUENCING, BIOINFORMATIC ANALYSES AND CLONING & EXPRESSION OF NOVEL GENES
TGGACAG TCCGGT AAT
Sequencing from transposon ends to identify gene of interest and subsequent bioinformatic analyses
Cloning and expression of identified gene(s) as an isolated gene in salt-sensitive E. coli MKH13
OD 595 nm
0.5
0.4
0.3
0.2
0.1
5
10
15
20
25
Time/ hrs
10
20
30
40
50
Growth experiments demonstrate an increased salt tolerance phenotype for E.coli MKH13 expressing metagenome-derived gene(s)
(4) FUTURE PROSPECTS: META-BIOTECHNOLOGY AND DESIGNER PROBIOTICS
A
Delivery
Improved tolerance to stress encountered during preparation and storage of the delivery matrix
B
in vivo Survival
Improved intestinal colonization and persistence
C
Clinical Efficacy
Gut Lumen
Basolateral surface of the intestinal epithelial cells
Improved prophylactic/therapeutic efficacy using strains tailored to target specific pathogens and/or toxins

Figure 1. An overview of novel gene discovery using functional metagenomics; from metagenomic library creation to novel therapeutics.

Novel salt tolerance loci identified

From our initial screen (the overall scheme is presented in Figure 1) of over 20,000 metagenomic clones, we identified 53 that could tolerate high sodium chloride (NaCl) concentrations. We termed these clones SMG 1-53 (salt metagenome). Through a combined transposon mutagenesis and bioinformatic strategy we identified five novel salt tolerance genes from clone SMG 3 (namely *galE*, *murB* and *mazG*) as well as additional *mazG* and *galE* genes from clone SMG 5 and SMG 25, respectively. Phylogenetic assignments revealed SMG 3 had the highest genetic identity to *Collinsella aerofaciens*, while SMG 5 and SMG 25 corresponded to *Eggerthella sp.* YY7918 and *Akkermansia muciniphila*, respectively. Each of the five genes was cloned separately and expressed in the osmosensitive strain *Escherichia coli* MKH13, which resulted in an increased tolerance to the ionic osmotic stressors NaCl and KCl (potassium chloride), but not to non-ionic stressors such as sucrose and glycerol, a finding which seems to suggest that these genes confer a salt-specific protective effect. While *E.* coli has been shown to up-regulate a set of genes in response to both ionic and non-ionic osmotic stress, it also regulates genes specific to each type of stress.[13] Furthermore each of the three genes was also found to be over-represented in the human gut metagenome and abundant among healthy subjects from the MetaHit dataset.[12, 14]

galE

The *galE* gene product (UDP glucose 4-epimerase) catalyses the inter-conversion of UDP glucose and UDP galactose and has been previously linked to the osmotic stress response through different mechanisms such as the production of the osmoprotectant trehalose or through cellular signalling.[15, 16] We theorise that *galE* may be important in maintaining the integrity of the lipopolysaccharide (LPS) in Gram negative or lipoteichoic acid (LTA) layers in Gram positive bacteria, making the cell more resistant to salt-induced osmotic stress. A recent functional metagenomic study identified a *galE* gene, that when cloned and expressed in *E. coli* conferred resistance to menadione, which can cause membrane damage through the generation of reactive oxygen species.[17] The authors believe the resistance is mediated through *galE* by increasing the permeability barrier of the cell. Such menaquinones are also found at significant concentrations in the human gastrointestinal tract.[18] The ability to form biofilms is likely to be critically important in the gut environment and for homeostasis within the community.[19] Furthermore, it has been shown that *galE* mutants have reduced ability to form biofilms, while *gal* mutants are defective in intestinal colonisation, both of which could be an important factor in the gastrointestinal tract.[20, 21]

murB

The *murB* gene is involved in the biosynthesis of peptidoglycan and the bacterial cell wall, which itself plays an important role in withstanding osmotic stress.[22] Disruption or deletion of the *murB* gene could make cells acutely sensitive to osmotic stress due to a reduction in cell wall integrity as well as causing a reduction in turgor pressure which is a driving force for cellular growth and division. Bacteria remodel the structure of their peptidoglycan in response to changes in environmental conditions [23], which could be important in the gut by allowing for varying levels of rigidity or elasticity depending on the conditions in the immediate environment. One of the more interesting functions of peptidoglycan, and a possible reason why genes for its synthesis are enriched among the human gut microbiota, is its stimulation of host immunity. Clarke *et al* (2010), have demonstrated peptidoglycan from the commensal microbiota modulate the innate immune system by improving neutrophil function even in the absence of infection.[24] The authors note that peptidoglycan can be translocated to the bloodstream, with concentrations at similar levels to those in faeces, indicating that there is constant peptidoglycan turnover among the microbiota and that immune stimulation by the microbiota can affect sites distal from the GI tract. Stimulated neutrophils demonstrated increased killing of the pathogenic bacteria *Streptococcus pneumoniae* and *S. aureus*.[24] This may indicate a co-evolution of a mutually beneficial arrangement by removing potentially harmful host pathogens and competitors to our symbiotic gut microbiota.

mazG

The *mazG* gene represents the most interesting of the identified genes, in that its possible mode of action in response to salt stress is not as immediately clear as *galE* or *murB.* As the latter two genes are related to outer membrane or cell wall functions, one can envisage how they could mediate resistance to external environmental stresses. MazG has been shown to play a role in different cellular processes, such as the removal of aberrant dNTP's from DNA strands,[25] as well as the oxidative[26] and nutritional stress responses.[27] Often found in association with the *mazEF* toxin-antitoxin (TA) system, the *mazG* gene product can delay programmed cell death and allow the cell to survive for longer periods under stress in the event that additional nutrients become available.[27] TA systems such as *mazEF* can induce cell death or arrest in response to various cellular stresses,[28] particularly those which induce DNA damage. MazG may delay apoptosis in salt-stressed cells as well as providing a mechanism to reduce or repair salt-induced damage to DNA. It has been speculated that individual TA systems may respond to specific stresses[29] and play a role in biofilm and persister cell formation, as well as having numerous other putative functions.[30] The development of persister cells allows for the survival of a small subpopulation through the death of the majority of the population.[31] Such an altruistic characteristic benefits long term survival and it seems, at the microscopic level, that the sacrifice of many for the good of a few may be a tenet of bacterial survival. Although the *mazG* genes identified in our study[12] are not located in the genomic neighbourhood of any obvious TA system, it is possible their encoded proteins could still regulate such TA systems at distant chromosomal locations or indeed regulate as yet unidentified, stress responsive genes or function as a general stress responsive protein itself.

Future perspectives

Whilst expanding our current knowledge on the diverse mechanisms employed by bacteria to overcome salt stress, the identification of novel genes may also assist in the development of novel drugs and drug targets as well as novel strategies to control some of the resident gut microbiota.[32, 33] Interestingly *murB* has been investigated as a possible target for novel antibiotics and antibacterial compounds[34] as it is exclusively found in bacteria and it has an important role in maintaining cellular integrity and viability. Furthermore, *galE* has been investigated as target for novel therapeutics against African sleeping sickness[35] and a *galE* mutant of *Salmonella enterica* serovar Typhi has been used to create an oral live attenuated vaccine for typhoid.[36] TA systems have also received attention as possible novel antibacterial drug targets.[37] MazG could also be a putative target, if disruption of its function could allow the toxin component of the TA system to cause cell death to certain bacterial populations.

Ultimately we would envisage that some of these novel salt tolerance genes could be used as part of a "meta-biotechnology"[38] strategy for the development of technologically more robust probiotic cultures and with greater ability to survive gastrointestinal transit and colonise the gut. Meta-biotechnology describes the "mining" of the human gut metagenome for novel genes for use in medicine, science and industry for the development of novel therapeutics[38] and is an extension of the patho-biotechnology[39, 40] concept. Patho-biotechnology has been used with success to engineer probiotic strains with increased stress resistance, improved gastrointestinal persistence and colonisation and therapeutic efficacy.[41]

Mining the metagenome is not limited to the identification of novel stress tolerance genes but may be used for the identification of novel antimicrobial compounds or genes that may be used for the development of designer probiotics which can be used to target specific pathogens or toxins.[42] New ways of thinking and alternative therapies are needed to control and combat pathogenic microorganisms in the era of increasing antibiotic resistance and emerging superbugs. The identification of such novel genes will broaden our understanding of salt tolerance in bacteria and uncover novel mechanisms employed for survival in the gastrointestinal tract as well as providing a platform for the development of novel biological therapeutics or novel drug targets.

References

1. Sleator RD, Hill C. Engineered pharmabiotics with improved therapeutic potential. Human vaccines 2008(a); 4:271-4.

2. Sleator RD, Hill C. New frontiers in probiotic research. Lett Appl Microbiol 2008(b); 46:143-7.

3. Louis P, O'Byrne CP. Life in the gut: microbial responses to stress in the gastrointestinal tract. Sci Prog 2010; 93:7-36.

4. Sleator RD, Hill C. Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. FEMS Microbiol Rev 2002; 26:49-71.

5. Epstein W. The roles and regulation of potassium in bacteria. Prog Nucleic Acid Res Mol Biol 2003; 75:293-320.

6. Kempf B, Bremer E. Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. Arch Microbiol 1998; 170:319-30.

7. Piuri M, Sanchez-Rivas C, Ruzal SM. Adaptation to high salt in Lactobacillus: role of peptides and proteolytic enzymes. J Appl Microbiol 2003; 95:372-9.

8. Sleator RD, Hill C. A novel role for the LisRK two-component regulatory system in listerial osmotolerance. Clin Microbiol Infect 2005; 11:599-601.

9. Sleator RD, Shortall C, Hill C. Metagenomics. Lett Appl Microbiol 2008(c); 47:361-6.

10. Sleator RD. The human superorganism - of microbes and men. Medical hypotheses 2010(a); 74:214-5.

11. Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, Dantas G, et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proc Natl Acad Sci U S A 2011; 108:6252-7.

12. Culligan EP, Sleator RD, Marchesi JR, Hill C. Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. The ISME Journal 2012.

13. Shabala L, Bowman J, Brown J, Ross T, McMeekin T, Shabala S. Ion transport and osmotic adjustment in Escherichia coli in response to ionic and non-ionic osmotica. Environ Microbiol 2009; 11:137-48.

14. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010; 464:59-65.

15. Bohringer J, Fischer D, Mosler G, Hengge-Aronis R. UDP-glucose is a potential intracellular signal molecule in the control of expression of sigma S and sigma S-dependent genes in Escherichia coli. J Bacteriol 1995; 177:413-22.

16. Leloir LF. The enzymatic transformation of uridine diphosphate glucose into a galactose derivative. Arch Biochem Biophys 1951; 33:186-90.

17. Mori T, Suenaga H, Miyazaki K. A metagenomic approach to the identification of UDP-glucose 4-epimerase as a menadione resistance protein. Biosci Biotechnol Biochem 2008; 72:1611-4.

18. Conly JM, Stein K. Quantitative and qualitative measurements of K vitamins in human intestinal contents. Am J Gastroenterol 1992; 87:311-6.

19. Macfarlane S, Dillon JF. Microbial biofilms in the human gastrointestinal tract. J Appl Microbiol 2007; 102:1187-96.

20. Ho TD, Waldor MK. Enterohemorrhagic Escherichia coli O157:H7 gal mutants are sensitive to bacteriophage P1 and defective in intestinal colonization. Infect Immun 2007; 75:1661-6.

21. Nakao R, Senpuku H, Watanabe H. Porphyromonas gingivalis galE is involved in lipopolysaccharide O-antigen synthesis and biofilm formation. Infect Immun 2006; 74:6145-53.

22. van Heijenoort J. Murein synthesis. In: Neidhardt FC, Curtis, R., III., Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaecter, M. and Umbarger, H.E, ed. Escherichia coli and Salmonella: cellular and molecular biology. Washington, D.C.: American Society for Microbiology, 1996:1025-34.

23. Vijaranakul U, Nadakavukaren MJ, de Jonge BL, Wilkinson BJ, Jayaswal RK. Increased cell size and shortened peptidoglycan interpeptide bridge of NaCl-stressed Staphylococcus aureus and their reversal by glycine betaine. J Bacteriol 1995; 177:5116-21.

24. Clarke TB, Davis KM, Lysenko ES, Zhou AY, Yu Y, Weiser JN. Recognition of peptidoglycan from the microbiota by Nod1 enhances systemic innate immunity. Nat Med 2010; 16:228-31.

25. Galperin MY, Moroz OV, Wilson KS, Murzin AG. House cleaning, a part of good housekeeping. Mol Microbiol 2006; 59:5-19.

26. Lu LD, Sun Q, Fan XY, Zhong Y, Yao YF, Zhao GP. Mycobacterial MazG is a novel NTP pyrophosphohydrolase involved in oxidative stress response. J Biol Chem 2010; 285:28076-85.

27. Gross M, Marianovsky I, Glaser G. MazG -- a regulator of programmed cell death in Escherichia coli. Mol Microbiol 2006; 59:590-601.

28. Hazan R, Sat B, Engelberg-Kulka H. Escherichia coli mazEF-mediated cell death is triggered by various stressful conditions. J Bacteriol 2004; 186:3663-9.

29. Wang X, Wood TK. Toxin-antitoxin systems influence biofilm and persister cell formation and the general stress response. Appl Environ Microbiol 2011; 77:5577-83.

30. Yamaguchi Y, Park JH, Inouye M. Toxin-antitoxin systems in bacteria and archaea. Annual review of genetics 2011; 45:61-79.

31. Erental A, Sharon I, Engelberg-Kulka H. Two programmed cell death systems in Escherichia coli: an apoptotic-like death is inhibited by the mazEF-mediated death pathway. PLoS biology 2012; 10:e1001281.

32. Sleator RD. Probiotic therapy - recruiting old friends to fight new foes. Gut Pathog 2010(b); 2:5.

33. Sleator RD. Probiotics -- a viable therapeutic alternative for enteric infections especially in the developing world. Discov Med 2010(c); 10:119-24.

34. Shapiro AB, Livchak S, Gao N, Whiteaker J, Thresher J, Jahic H, et al. A homogeneous, high-throughput-compatible, fluorescence intensity-based assay for UDP-N-acetylenolpyruvylglucosamine reductase (MurB) with nanomolar product detection. Journal of biomolecular screening 2012; 17:327-38.

35. Urbaniak MD, Tabudravu JN, Msaki A, Matera KM, Brenk R, Jaspars M, et al. Identification of novel inhibitors of UDP-Glc 4'-epimerase, a validated drug target for african sleeping sickness. Bioorganic & medicinal chemistry letters 2006; 16:5744-7.

36. Germanier R, Fuer E. Isolation and characterization of Gal E mutant Ty 21a of Salmonella typhi: a candidate strain for a live, oral typhoid vaccine. J Infect Dis 1975; 131:553-8.

37. Engelberg-Kulka H, Sat B, Reches M, Amitai S, Hazan R. Bacterial programmed cell death systems as targets for antibiotics. Trends in microbiology 2004; 12:66-71.

38. Culligan EP, Hill C, Sleator RD. Probiotics and gastrointestinal disease: successes, problems and future prospects. Gut Pathog 2009; 1:19.

39. Sleator RD, Hill C. Patho-biotechnology: using bad bugs to do good things. Curr Opin Biotechnol 2006; 17:211-6.

40. Sleator RD, Hill C. Patho-biotechnology; using bad bugs to make good bugs better. Sci Prog 2007; 90:1-14.

41. Sheehan VM, Sleator RD, Hill C, Fitzgerald GF. Improving gastric transit, gastrointestinal persistence and therapeutic efficacy of the probiotic strain Bifidobacterium breve UCC2003. Microbiology 2007; 153:3563-71.

42. Focareta A, Paton JC, Morona R, Cook J, Paton AW. A recombinant probiotic for treatment and prevention of cholera. Gastroenterology 2006; 130:1688-95.

CHAPTER IV

Metagenomic Identification of a Novel Salt Tolerance Gene from the Human Gut Microbiome which Encodes a Membrane Protein with Homology to a *brp*/*blh*-family Beta-Carotene 15,15’-Monooxygenase

SUBMITTED TO PLOS ONE

Abstract

The human gut microbiome consists of at least 3 million non-redundant genes, 150 times that of the core human genome. Herein, we report the identification and characterisation of a novel stress tolerance gene from the human gut metagenome. The locus, assigned *brpA*, encodes a membrane protein with homology to a *brp*/*blh*-family beta-carotene monooxygenase. Cloning and heterologous expression of *brpA* in *Escherichia coli* confers a significant salt tolerance phenotype. Furthermore, when cultured in the presence of exogenous beta-carotene, cell pellets adopt a red/orange pigmentation indicating the incorporation of carotenoids in the cell membrane.

Introduction

Metagenomics provides a culture-independent means to access and study the genetic content of all of the microorganisms in a particular environmental niche. Metagenomic analysis can be sequence-based or functional (or a combination of both). The development of faster, cheaper and more accurate next-generation sequencing (NGS) technologies has allowed new insights into microbial community structure and diversity and has led to the discovery of many novel genetic loci [1-4]. Functional metagenomics has also been utilised to identify many novel functions through cloning and heterologous expression of metagenomic DNA and subsequent phenotypic detection of a desired trait conferred on the cloning host. Some notable examples include genes encoding proteins of industrial, pharmaceutical and medical relevance such as lipases, esterases and novel antibiotics [5-8].

The human gut microbiome has become perhaps the most intensively studied environment using metagenomics [9,10]. Collectively, there are at least 150 times as many genes in the human gut microbiome than there are human genes in the genome, a large proportion of which are uncharacterised [11]. The ability to respond and adapt to external environmental stresses is key to microbial survival and it is possible to use metagenomics to identify novel mechanisms that enable such survival [12]. In the gastrointestinal (GI) tract microorganisms are faced with numerous challenges such as low pH, low iron concentrations, increased osmolarity, bile, immunity mechanisms and competing microbes [13,14]. Different sets of genes are activated in response to environmental cues [15]. Work in our lab is focused on genes that confer increased tolerance to osmotic stress [16]. The response to osmotic stress is broad and encompasses many diverse cellular processes and systems [17]. Metagenomics makes it possible to identify novel systems unrelated to the classical (and comprehensively studied) primary and secondary responses of potassium (K^+) uptake and osmoprotectant utilisation [18-20]. We have previously identified a number of novel salt tolerance loci from the human gut microbiota using a combination of functional metagenomic screening, next-generation sequencing and bioinformatic analyses [21-23].

In this study we report the identification of a novel salt tolerance gene from a human gut metagenomic library we have previously described [22]. An *in silico* analysis revealed the gene (which we have termed *brpA*) encoded a carotenoid modifying enzyme with homology to a *brp*/*blh*-family beta-carotene 15,15'-monooxygenase protein, which cleaves beta-carotene to two molecules of *all-trans* retinal (vitamin A aldehyde) [24,25]. Finally, we demonstrate that *brpA* confers an increased salt tolerance phenotype when heterologously expressed in *Escherichia coli.*

Materials and methods

Bacterial strains and growth conditions

Bacterial strains and plasmids used in this study are listed in Table 1. Oligonucleotide primers (synthesised by Eurofins, MWG Operon, Germany) are presented in Table 2. *E. coli* EPI300::pCC1FOS (Epicentre Biotechnologies, Madison, WI, USA) was cultured in Luria-Bertani (LB) medium containing 12.5μg/ml chloramphenicol (Cm) and in 12.5μg/ml chloramphenicol plus 50μg/ml kanamycin (Kan) following EZ-Tn*5*™ transposon mutagenesis. *E. coli* MKH13 was grown in LB and LB supplemented with 20μg/ml Cm for strains transformed with the plasmid pCI372. *E. coli* strains containing the pBAD expression vector were cultured in the presence of 100μg/ml ampicillin.

For growth in minimal media, strains were grown in M9 (Fluka) minimal salts supplemented with final concentrations of 0.4% glucose, 0.2% casamino acids, 2mM magnesium sulphate ($MgSO_4$) and 0.1mM calcium chloride ($CaCl_2$). When required, stock solutions of beta-carotene were added to media at a final concentration of 20μM. Growth media was supplemented with 1.5% agar for plate assays. All overnight cultures were grown with shaking at 37°C.

Table 1. Bacterial strains and plasmids

Strain, plasmid or transposon	Genotype or characteristic(s)	Source or reference
Strains		
E. coli EPI300	F$^-$ *mcrA* Δ(*mrr-hsd*RMS-*mcrBC*) Φ80d*lacZ*ΔM15 Δ*lac*X74 *recA*1 *endA*1 *araD*139 Δ(*ara, leu*)7697 *galU galK* λ$^-$ *rpsL nupG trfA dhfr*; high-transformation efficiency of large DNA	Epicentre Biotechnologies, Madison, WI, USA
SMG 6	EPI300 containing pCC1FOS fosmid with ~34kb of metagenomic DNA from human gut microbiome	This study
SMG 6-EZTn*5* #24	Transposon insertion in gene 24 (which precedes acyltransferase gene *atfA*)	This study
SMG 6-EZTn*5* #26	Transposon insertion in *brpA* gene	This study
SMG 6-EZTn*5* #34	Transposon insertion in acyltransferase gene (*atfA*)	This study
SMG 6-EZTn*5* #38	Transposon insertion in *brpA* gene	This study
E.coli MKH13	MC4100*Δ*(*putPA*)101D(*proP*)2D(*proU*)	[26]
E.coli MKH13::pCI372	MKH13 containing empty pCI372 plasmid	This study
E. coli MKH13::pCI372-*brpA*$_S$	MKH13 containing pCI372 with *brpA*$_S$ gene from SMG 6; “S” subscript denotes shorter predicted gene with TTG start codon	This study
E. coli MKH13::pCI372-*brpA*$_L$	MKH13 containing pCI372 with *brpA*$_L$ gene from SMG 6; “L” subscript denotes longer predicted gene with ATG Start codon	This study
E. coli MKH13::pCI372-*brpAatfA*	MKH13 containing pCI372 with *brpAatfA* genes from SMG 6	This study
E. coli EPI300::pBAD-*brpA*$_S$	EPI300 containing pBAD with *brpA*$_S$ gene from SMG 6	This study
E. coli EPI300::pBAD-*brpA*$_L$	EPI300 containing pBAD with *brpA*$_L$ gene from SMG 6	This study
E. coli EPI300::pBAD-*brpAatfA*	EPI300 containing pBAD with *brpA* and *atfA* genes from SMG 6	This study
Plasmids		
pCI372	Shuttle vector between *E. coli* and *L. lactis*, Cm^R	[27]
pCC1FOS	Fosmid cloning vector, Cm^R	Epicentre Biotechnologies, Madison, WI, USA
pBAD	L-arabinose inducible expression vector, Amp^R	Invitrogen, USA
Transposon		
EZ-Tn*5* <*ori*V/ KAN-2>	Hyperactive Tn*5* transposon, Kan^R, inducible high copy origin of replication – *ori*V	Epicentre Biotechnologies, Madison, WI, USA

Cm^R, Kan^R and Amp^R = chloramphenicol, kanamycin and ampicillin resistance respectively.

Table 2. Oligonucleotide primers

Primer	Sequence (5' – 3')[a]
pCI372 FP	CGGGAAGCTAGAGTAAGTAG
pCI372 RP	CCTCTCGGTTATGAGTTAG
pBAD FP	ATGCCATAGCATTTTTATC
pBAD RP	GATTTAATCTGTATCAGG
$brpA_S$ FP	AAAACTGCAGACCCAACACGATGCCATATT
$brpA_S$ RP	GCTCTAGATAACAGGGTGCGGTGATACA
brpAatfA FP	AAAACTGCAGTAGCGGCTGGATCGGTAGTA
brpAatfA RP	GCTCTAGAACCCAACACGATGCCATATT
$brpA_L$ FP	AAAACTGCAGGCCGAATATCAACCCAACAC
$brpA_L$ RP	GCTCTAGAAGGTATTTGTGCCTTGTGCT
EZTn-FP-1	GCCAACGACTACGCACTAGCCAAC
EZTn-RP-1	GAGCCAATATGCGAGAACACCCGAGAA

[a] Restriction enzyme cut-sites are underlined

(*PstI,* CTGCAG; *XbaI,* TCTAGA)

Construction and screening of the human gut metagenomic library

A previously constructed fosmid clone library, created from metagenomic DNA from the human gut microbiome isolated from a stool sample from a 26 year-old healthy Caucasian male [28] was used to screen for salt-tolerant clones. The library was screened using the protocol outlined by Culligan et al [22]. Briefly, a total of 23,040 library clones were screened on LB agar supplemented with 6.5% (w/v) NaCl using a Genetix QPix 2 XT™ colony picking/gridding robotics platform. Plates were incubated at 37°C for 2-3 days and checked periodically for growth of likely salt-tolerant clones.

Sequencing and bioinformatic analysis

The fosmid insert from clone SMG 6 was fully sequenced and assembled by GATC Biotech (Konstanz, Germany) using the GS-FLX 454 pyrosequencing (Roche) platform on a Titanium mini-run. Putative open reading frames were predicted using Softberry FGENESB bacterial operon and gene prediction software (www.softberry.com) and also GeneMark [29]. Retrieved nucleotide and translated amino acid sequences were functionally annotated by homology searches using the Basic Local Alignment and Search Tool (BLAST) to identify homologous sequences from the National Centre for Biotechnology Information (NCBI) website: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi. The following databases and tools were used to gain additional information on the BrpA protein: Conserved Domain Database (CDD), PROSITE motif search, SignalP 4.0, HMMER, TMHMM, HHPred, and Softberry BProm promoter search (www.softberry.com) [30-36].

The Fold and Functional Assignment System (FFAS03) is a profile-profile and fold recognition algorithm that can detect remote homology between proteins [37]. Profile-profile comparisons have increased sensitivity compared to sequence-sequence or profile-sequence algorithms. FFAS03 searches numerous databases including non-redundant (nr) NCBI, Global Ocean Sampling (GOS) from JCVI, PDB, SCOP, and COG, as well as numerous metagenome datasets including MetaHit [11] which contains over 3 million unique genes from the human gut microbiome. The BrpA protein sequence was submitted to the server to identify proteins with distant homology based on FFAS profiling or homologues by BLAST and PSI-BLAST against the databases and metagenome datasets. The FFAS03 server can be found at: http://ffas.burnham.org/ffas-cgi/cgi/document.pl

The Integrated Microbial Genomes and Metagenomes (IMG/M) [38] is a data management system for the comparative analysis of metagenome sequence data. IMG/M-HMP [39] specifically contains metagenome data from the Human Microbiome Project (HMP) [40]. It contains 748 metagenome datasets generated from sequencing samples from different body sites and also, tools for comparative analysis between hosted sequences and user supplied sequences. The BrpA protein sequence was used a query sequence to BLAST (1e-^{05} and 1e-^{50} maximum e-value cut-off) against all the available metagenomes from 17 body sites from the HMP dataset. The IMG/M-HMP server can be found at: http://www.hmpdacc-resources.org/cgi-bin/imgm_hmp/main.cgi

DNA manipulations and cloning

Induction of fosmids from low to high copy number was performed as per the manufacturer's instructions. The Qiagen QIAprep® Spin mini-prep kit was used to extract fosmids using the protocol outlined by manufacturer. The *brpA*$_L$*, brpA*$_S$ and *brpAatfA* genes were amplified using ReddyMix PCR mastermix (Thermo Scientific). PCR products were purified with a Qiagen PCR purification kit and digested with restriction enzymes *XbaI* and *PstI* (Roche Applied Science), followed by ligation using the Fast-Link DNA ligase kit (Epicentre Biotechnologies) to similarly digested plasmid pCI372. Electro-competent *E. coli* MKH13 were transformed with the ligation mixture and plated on LB agar plates containing 20µg/ml Cm for selection.

The pBAD TOPO® TA expression kit (Invitrogen, Carlsbad CA, USA) was used to clone the PCR products into the pBAD expression vector according to the manufacturer's instructions. The *brpA*$_L$, *brpA*$_S$ and *brpAatfA* genes were amplified as outlined above. The resulting plasmids, containing the genes of interest were electroporated into freshly competent *E. coli* EPI300 and plated on LB agar containing 100µg/ml of ampicillin.

Colony PCR was performed on resistant transformants using a using a gene and plasmid (pCI372 or pBAD) specific primer combination to confirm the presence and size of the insert. Inserts were sequenced to confirm the correct nucleotide sequence (GATC Biotech, Germany).

Growth experiments

Cultures were grown overnight in the relevant media (LB or M9 broth). Cells were subsequently harvested, washed in one quarter strength sterile Ringer's solution and re-suspended in fresh media. A 2% (v/v) inoculum was sub-cultured in fresh broth containing sodium chloride (NaCl), and 200µl was transferred to a sterile 96-well micro-titer plate (Starstedt Inc. Newton, USA). For minimal media experiments, filter-sterilised stock solutions of the osmoprotectants betaine, L-carnitine and L-proline were added to a final concentration of 1mM. Micro-titer plates were incubated at 37°C for 24-48 hours in an automated spectrophotometer (Tecan Genios) which recorded the OD 595nm every hour. The data was subsequently retrieved and analysed using the Magellan 3 software program.

Survival in high salt media in the presence and absence of 20µM beta-carotene was assessed by harvesting overnight cultures as above and sub-culturing in either 3% NaCl or 7% NaCl for MKH13 and EPI300 strains respectively. Cultures were incubated at 37°C both aerobically (with shaking) and anaerobically (static) for 48 hours. Subsequently, serial dilutions of cultures were made in one quarter strength sterile Ringers solution and plated on LB agar. Viable cells were enumerated and calculated as the number of colony forming units per millilitre (CFU/ ml).

Graphs (created using SigmaPlot 10.0) are presented as the average of triplicate experiments, with error bars being representative of the standard error of the mean (SEM).

Transposon mutagenesis

Transposon mutagenesis was carried out on SMG 6 using the EZTn-*5* <oriV/KAN-2> *in vitro* transposition kit (Epicentre Biotechnologies) in accordance with the manufacturer's instructions. *E. coli* EPI300 cells were transformed with the transposon reaction mixture and selected on plates containing Cm and Kan (12.5 and 50µg/ml, respectively). Transposon insertions in the regions of interest were confirmed by PCR. Regions containing the EZTn*5* transposon are approximately 1.9kb larger than the region covered by the primers. PCR products of the correct size were subjected to sequencing from the ends of the transposon using the primers EZTn FP-1 and RP-1 (Table 2) to confirm the location of transposon insertion. All sequencing was performed by GATC Biotech (Germany).

Results

Screening the human gut metagenomic library

Fifty-three salt-tolerant clones were identified from a screen of approximately 23,000 fosmid library clones. The clones were annotated as SMG (for Salt MetaGenome) 1-53. Six clones grew within 24 hours (SMG 1-6) and the remaining 47 grew over the following 24-48 hours. The focus of this study were clones SMG 1 and SMG 6, both of which were found to contain the same insert. SMG 6 was chosen for further analysis. Previous work has focused on clones SMG 3 and SMG 5 and SMG 25 [22,23]. End sequencing revealed that another clone, SMG 52, shared the same sequences at the 5’ and 3’ ends of the fosmid as SMG 1 and SMG 6. Furthermore, SMG 52 displayed a similar growth profile to SMG 1 and 6 when grown under sodium chloride (NaCl) stress and all three clones have a significant (*P* <0.0001 for all clones) growth advantage in the presence of 7% added NaCl compared to the EPI300 host strain carrying the empty fosmid vector (pCC1FOS) (Figure 1B). No difference in growth between any of the clones was observed in LB alone (Figure 1A). Further investigation involving pyrosequencing revealed SMG 52 contained the same insert as SMG 1 and SMG 6.

Figure 1.

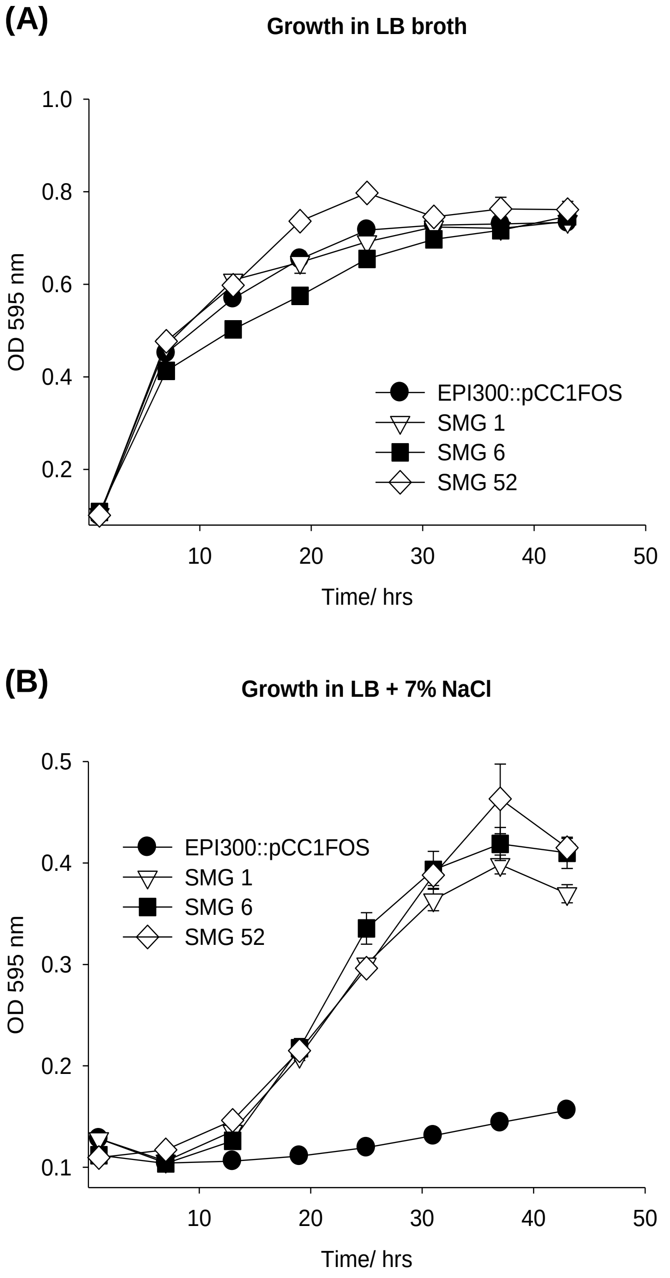


Figure 1. Growth of metagenomic clones SMG 1, 6 and 52 compared to EPI300 carrying an empty fosmid vector (pCC1FOS) in **(A)** LB broth and **(B)** LB broth supplemented with 7% NaCl.

Fosmid sequencing and bioinformatic analysis

The fosmid inserts from SMG 1, 6 and 52 were fully sequenced and assembled by GATC Biotech (Germany) using the GS-FLX Titanium mini run. All three inserts were found to be identical, sharing 100% nucleotide identity over the entire length of the fosmid insert (~34 kb). Gene prediction using FGENESB predicted the presence of thirty putative open reading frames (see Table 3.). Translated nucleotide sequences were subjected to BLASTP (maximum e-value cut-off of 1e-^{05}) analysis to identify homologous sequences in the database. The vast majority corresponded to proteins from the Gram-negative *Bacteroidetes* phylum, with amino acid identities ranging from 26% to 100%. Proteins with between 99%-100% amino acid identity corresponded to three species of *Bacteroides*, namely *Bacteroides thetaiotaomicron* VPI-5482, *Bacteroides sp.* 1_1_6 and *Bacteroides sp.* 1_1_14. The remainder corresponded to other members of the phylum *Bacteroidetes* from genera *Alistipes*, *Prevotella* and *Odoribacter*, as well to Gram-positive *Firmicutes* from the family *Lachnospiraceae* and genera *Clostridium* and *Veillonella*.

Functional assignment of the encoded proteins on SMG 6 based on homology searches using BLASTP revealed that gene 26 was predicted to encode a putative membrane protein, although none of the potential homologues identified shared greater than 30% amino acid identity (placing them in the “twilight zone” of evolutionary relatedness). This protein also shared sequence similarity with a *brp/blh*-family 15,15’-beta-carotene monooxygenase from *Prevotella marshii* DSM 16973 (28% identity over 254 amino acids) and to a proline symporter from *Bifidobacterium bifidum* BGN 4

(25% identity over 222 amino acids). Given that proline is an important osmoprotectant utilised by bacteria to counteract the deleterious effects of salt-induced osmotic stress [41,42], we elected to pursue this gene, which we have named *brpA,* for further study.

Table 3. List of putative proteins encoded on SMG 6 fosmid insert.

Gene #	Strand	Length (a.a)	BlastP top hit	Best hit organism (BlastP)	e-value	Query coverage	% ID (a.a)	% G+C
1	+	418	Hypothetical protein (BT_1366)	*Bacteroides thetaiotaomicron* VPI-5482	0.00E+00	100%	100%	48.05
2	-	124	Conserved hypothetical protein (DUF 3127)	*Bacteroides sp.* 1_1_6	6.00E-67	100%	100%	52.00
3	+	374	DNA polymerase III, chain beta (beta clamp superfamily)	*Bacteroides thetaiotaomicron* VPI-5482	0.00E+00	100%	100%	47.64
4	+	255	DNA polymerase III, epsilon chain (DDEHh exonuclease domain)	*Bacteroides sp* .1_1_6	4.00E-149	100%	100%	45.57
5	+	400	Phosphopantothenoylcysteine decarboxylase (Flavoprotein, DFP domains)	*Bacteroides sp.* 1_1_6	0.00E+00	100%	99%	49.38
6	+	555	DNA repair protein, RecN (ABC_RecN domains)	*Bacteroides sp.* 1_1_14	0.00E+00	100%	99%	50.66
7	+	247	tRNA/ rRNA methyltransferase (SpoU_sub_bind and methylase domains)	*Bacteroides thetaiotaomicron* VPI-5482	5.00E-143	100%	100%	47.98
8	+	568	Tetratricopeptide repeat family protein (Trypsin_2 and TPR domains)	*Bacteroides sp.* 1_1_6	0.00E+00	100%	99%	46.34
9	+	190	Putative transcriptional regulator (NGN_SP_UpxY domain; NusG)	*Bacteroides thetaiotaomicron* VPI-5482	4.00E-135	100%	99%	42.76
10	+	122	Transcriptional regulator (UpxZ domain)	*Bacteroides sp.* 1_1_14	2.00E-63	100%	97%	42.82
11	+	789	Capsule polysaccharide export protein (Poly_export and SLBB domains)	*Bacteroides thetaiotaomicron* VPI-5482	0.00E+00	100%	99%	48.90
12	+	378	Uncharacterised protein, putative chain-length determining	*Bacteroides faecis* CAG:32	0.00E+00	100%	71%	42.83
13	+	647	Capsular polysaccharide biosynthesis protein (CapD) (UDP_invert_4-6DH_SDR_e domain)	*Bacteroides sp.* 1_1_6	0.00E+00	98%	87%	44.24
14	+	410	UDP-glucose dehydrogenase (NAD_Gly3P_dh_N and UDPG_MGDP_dh domains)	*Bacteroides fragilis* HMW 610	0.00E+00	100%	82%	43.23
15	+	348	NAD dependent epimerase/ dehydratase (NADB_Rossmann superfamily)	*Bacteroides sp.* 4_1_36	3.00E-176	100%	84%	42.50
16	+	189	dTDP-4-dehydrorhamnose 3,5-epimerase (dTDP_sugar_isom domain)	*Bacteroides fragilis* CL07T00C01	3.00E-124	100%	90%	43.33
17	+	451	Alanine ABC transporter permease (DltB domain; MBOAT superfamily)	*Alistipes onderdonkii*	2.00E-136	97%	57%	37.20
18	+	292	Hypothetical protein Alfi_2997 (DUF 535 superfamily)	*Alistipes finegoldii* DSM 17242	1.00E-38	96%	30%	36.61
19	+	172	Hypothetical protein BF638R_1544 (RimL domain)	*Bacteroides fragilis* 638R	2.00E-67	99%	68%	39.35
20	+	368	Putative LPS biosynthesis transmembrane protein	*Bacteroides fragilis* 638R	4.00E-51	98%	36%	33.06

21	+	331	Hypothetical protein BSHG_0833	*Bacteroides sp.* 3_2_5	2.00E-83	100%	43%	31.02
22	+	479	Hypothetical protein, polysaccharide export (MATE_tuaB_like domain)	*Clostridium hathewayi* WAL-18680	6.00E-150	98%	46%	36.11
23	+	376	Hypothetical protein HMPREF9447_00823 (AHBA_syn domain; AAT_I superfamily)	*Bacteroides oleiciplenus* YIT 12058	0.00E+00	100%	81%	47.48
24	+	184	Putative acetyltransferase (LbetaH_MAT_like domain)	*Prevotella sp.* CAG:1092	6.00E-57	94%	55%	40.00
25	+	98	Hypothetical protein, putative acyltransferase	*Prevotella saccharolytica* F0055	7.00E-07	98%	37%	30.64
26	**+**	**271**	**Putative membrane protein**	***Prevotella sp.* CAG:873**	**6.00E-10**	**99%**	**26%**	**32.05**
27	+	288	Hypothetical protein HMPREF0994_06876 (ATP-grasp_tupA domain)	*Lachnospiraceae bacterium* 3_1_57FAA_CT1	2.00E-81	99%	70%	35.01
28	+	366	Uncharacterised protein BN814_01473, putative glycosyltransferase (GT1_ams_like domain)	*Veillonella sp.* CAG:933	1.00E-132	98%	54%	39.69
29	+	366	Putative glycosyltransferase	*Bacteroides thetaiotaomicron* VPI-5482	2.00E-53	98%	37%	40.05%
30	+	364	Hypothetical protein HMPREF9449_00933	*Odoribacter laneus* YIT 12061	0.00E+00	98%	69%	40.43

Abbreviations: aa = amino acid; %ID = % identity at amino acid level over the entire length of the protein; DUF = Domain of Unknown Function.

Features of SMG 6 and *brpA*/BrpA

The *brpA* gene is number 26 of the 30 predicted genes on SMG 6 (Fig. 2). It is predicted to be a lone open reading frame, preceded by and followed by a 7 and a 4 gene operon, respectively. It is flanked upstream and downstream by a number of genes predicted to encode proteins with acetyl-, acyl- and glycosyl-transferase activities. There are indications that *brpA* and a number of adjacent genes have been acquired through lateral gene transfer (LGT). The SMG 6 fosmid insert is ~34.26 kb and its overall %G+C content is 41.92%. The highest genetic identities of a large proportion of the genes are to *Bacteroides* species, with up to 100% identity in some cases. The %G+C content of genus *Bacteroides* ranges from 40-48%, with *B. thetaiotaomicron* VPI-5482, *Bacteroides sp.* 1_1_6 and *Bacteroides sp.* 1_1_14 all having a %G+C content of approximately 43% (Genomes Online Database, GOLD; http://www.genomesonline.org/). The %G+C content of the genes on the SMG 6 fosmid insert is illustrated in Figure 2A. Genes in the first half of the insert, up to and including gene 16, have a %G+C content of ~45%; similar to the average %G+C content of the genus *Bacteroides*. The second half of the insert displays a clear drop in %G+C content to ~37%. The %G+C content of some individual genes is also low, including *atfA* and *brpA* (Figure 2A), which share BLAST homology to low G+C Gram-positive bacteria, mainly from the Phylum *Firmicutes*.

The *brpA* gene was predicted to have different start codons using FGENESB depending on the settings used; the alternative start codon TTG (leucine) was predicted using “generic bacterial”, resulting in a 232 amino acid protein. Given that a number of the proteins on SMG 6 shared 100%

amino acid identity with *Bacteroides thetaiotaomicron* VPI-5482, it was also chosen as the closest organism for gene prediction and predicted an ATG (methionine) as the start codon, 117 base-pairs upstream of the predicted TTG start codon (encoding a protein of 271 amino acids). GeneMark was used for gene prediction as a comparison and it also predicted the same ATG as the start codon. A putative ribosome binding site (RBS) sequence (AGGTTT) was found ending seven base-pairs upstream of TTG, while a stronger RBS sequence (AGTAGG) ended 19 base-pairs upstream of the ATG start codon. Putative *E.* coli-type -10 and -35 promoter regions were detected using BProm (www.softberry.com) upstream of both putative start codons. Manual inspection of upstream sequences also revealed the presence of a near perfect *Bacteroidetes* -7/-33 promoter region (TAGGTTTG/TTTT; consensus TAnnTTTG/TTTG) [43,44] upstream of the TTG start codon and a GGTATTTG/TTTT at -14/-30 upstream of ATG. The predicted promoter sequences along with putative transcription factor binding sites can be seen in Figure 2C. A putative RpoS binding site is found upstream of the ATG start codon, while an OxyR binding sequence is predicted to be located upstream of the TTG start codon.

The BrpA protein was predicted to be a 30.9 kDa membrane protein with seven transmembrane regions as predicted with TMHMM (Figure 2D). BrpA has a predicted pI of 9.42 and is composed of ~46% hydrophobic amino acids, similar to other microbial Brp/Blh proteins (pI range 8.89-9.56 and 48-56% hydrophobic amino acids) [24]. No signal peptide sequence, conserved domains or sequence motifs were detected for BrpA. We also searched for motifs in the protein sequences homologous to BrpA from

BLAST. A lipocalin motif was detected in a hypothetical protein from *Clostridium sp* KLE-1755. Interestingly, lipocalin motifs are found in proteins that bind small hydrophobic molecules such as retinoids, carotenoids, lipids and steroids [45]. Table 4 shows the lipocalin motif and the corresponding motif identified in *Clostridium sp* KLE-1755. The BrpA amino acid sequence along with the top 10 BLAST homologues were aligned to identify conserved residues in these proteins. The residues that match the lipocalin motif are displayed in green and those that do not are in red (Table 4).

Figure 2.

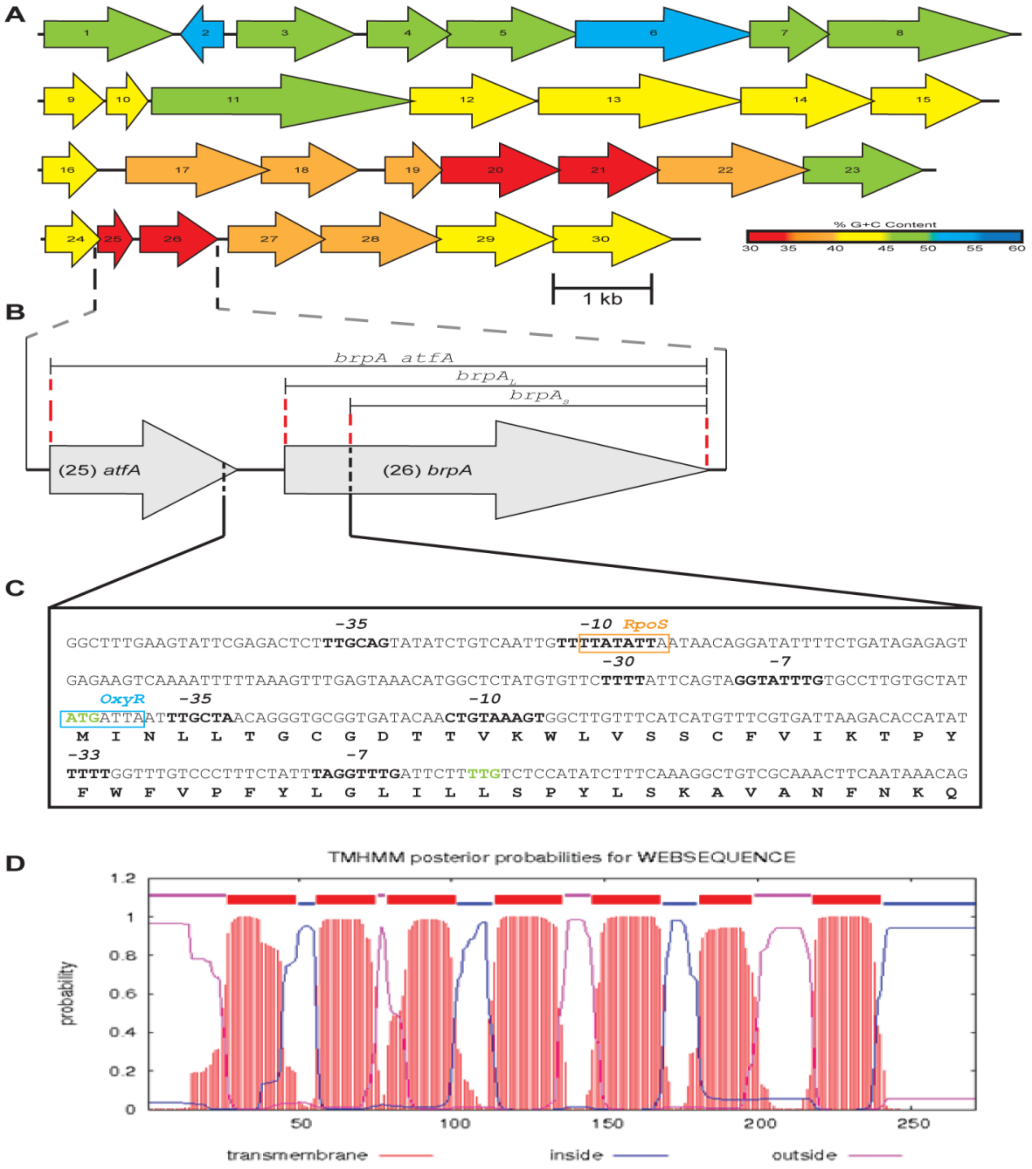

A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
% G+C Content
30 35 40 45 50 55 60
1 kb
B
brpA atfA
brpA_L
brpA_S
(25) atfA
(26) brpA
C
-35
-10 RpoS
GGCTTTGAAGTATTCGAGACTCTTTGCAGTATATCTGTCAATTGTTTTATATTAATAACAGGATATTTTCTGATAGAGAGT
-30
-7
GAGAAGTCAAAATTTTTAAAGTTTGAGTAAACATGGCTCTATGTGTTCTTTTATTCAGTAGGTATTTGTGCCTTGTGCTAT
OxyR
-35
-10
ATGATTAATTTGCTAACAGGGTGCGGTGATACAACTGTAAAGTGGCTTGTTTCATCATGTTTCGTGATTAAGACACCATAT
M I N L L T G C G D T T V K W L V S S C F V I K T P Y
-33
-7
TTTTGGTTTGTCCCTTTCTATTTAGGTTTGATTCTTTTGTCTCCATATCTTTCAAAGGCTGTCGCAAACTTCAATAAACAG
F W F V P F Y L G L I L L S P Y L S K A V A N F N K Q
D
TMHMM posterior probabilities for WEBSEQUENCE
probability
1.2
1
0.8
0.6
0.4
0.2
0
50
100
150
200
250
transmembrane
inside
outside

Figure 2. Bioinformatic analysis of SMG 6 and *brpA*/BrpA.

(A) Gene map of SMG 6 insert, displaying gene orientation and individual %G+C content indicated with a gradient colour bar. Gene numbers correspond to those in Table 3 and are drawn approximately to scale. (B) Focus on genes 25 (*atfA*) and 26 (*brpA*), showing the regions cloned for each construct. (C) Detailed view of putative ATG and TTG start codons of *brpA*, including upstream regions, as well as predicted promoter regions (highlighted in bold) and transcription factor binding site sequences (blue and orange boxes). (D) TMHMM prediction of seven transmembrane regions in BrpA.

Table 4. Putative lipocalin motifs in BrpA and its homologues

				[DENQG	X	[DENQ								
	LIPOCALIN MOTIF	[DENG]	{A}	STARK]	(0,2)	ARK]	[UVFY]	{CP}	G	{C}	W	[FYWLRH]	{D}	[LIVMTA]
Clostridium sp. KLE 1755	NPSRLAGAWYLVP	N	P	S		R	L	A	G	A	W	Y	L	V
BrpA SMG 6	LFSSMRDSIYLIPS	L	F	S		S	M	R	D	S	I	Y	L	I
Prevotella sp. CAG:873	DVHGALHSWWFVP	D	V	H		G	A	L	H	S	W	W	F	V
Prevotella buccalis ATCC 35310	VWQGMLDDSLFMF	V	W	Q		G	M	L	D	D	S	L	F	M
Prevotella sp. CAG:279	DVHSWLHSWAFVP	D	V	H		S	W	L	H	S	W	A	F	V
Prevotella marshii DSM 16973	PQTDFITSWSFLP	P	Q	T		D	F	I	T	S	W	S	F	L
Lachnospiraceae bacterium[1]	NPSQMADKWYLVP	N	P	S		Q	M	A	D	K	W	Y	L	V
Prevotella saccharolytica F0055	PQTDFITSWSFLP	P	Q	T		D	F	I	T	S	W	S	F	L
Clostridium nexile CAG:348	KPYQFANSSFIIL	K	P	Y		Q	F	A	N	S	S	F	I	I
Firmicutes bacterium CAG:24	NALTGRLGDFWNIVP	N	A	L	TG	R	L	G	D	F	W	N	I	V
Firmicutes bacterium CAG:65	GSDRIDGAVSLLL	G	S	D		R	I	D	G	A	V	S	L	L

[1] strain 3_1_57FAA_CT1

Table 4. A lipocalin motif was found in a homologue of BrpA from *Clostridium sp.* KLE1755. Lipocalin proteins can bind hydrophobic molecules such as carotenoids and retinoids. The top ten BLASTP homologues to BrpA were aligned to compare these protein sequences and identify putative lipocalin motifs. The consensus motif is displayed on the top row of Table 4. Residues that match the consensus are shown in green and mismatches are shown in red.

Due to low BLAST sequence identity, the FFAS03 server was used with the aim of identifying distant structural homologues to BrpA. The best structural homologues were an uncharacterised bacterial protein (COG 3274; acyltransferase) and a predicted membrane protein (COG 4763) with significant scores of -40.70 and -23.30 respectively. Interestingly the best structural homologue in the protein databank (PDB) was to an archaeal-type rhodopsin (3ug9), although the score of -9.43 did not reach significance (-9.50).

The IMG/M-HMP database which contains all metagenomic datasets encompassing 17 body sites from the Human Microbiome Project (HMP) was also screened for BrpA homologues. Using a combination of the most lenient and strictest search criteria (maximum e-value cut-off of 1e-05 and 1e-50) BrpA homologues were identified in the HMP datasets (Figure 3). In addition, there were 145 hits to the MetaHit dataset using BLAST on the FFAS03 server.

Figure 3.

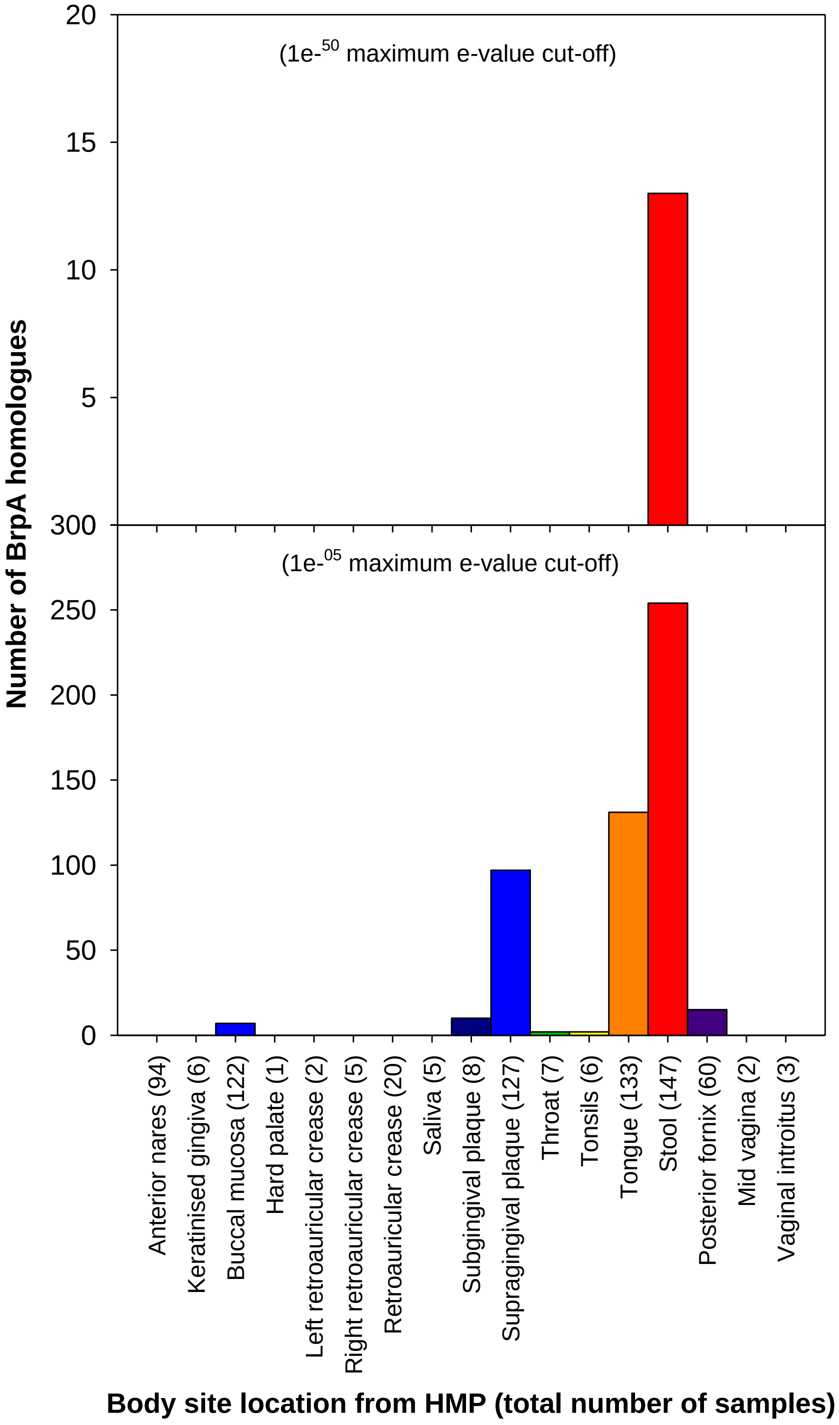


Figure 3. BrpA homologues identified when BLAST searched against Human Microbiome Project (HMP) datasets from 17 body sites at maximum e-value cut-off of **(A)** 1e^{-50} and **(B)** 1e^{-05}.

The *brpA* gene confers a salt tolerance phenotype when heterologously expressed in *Escherichia coli*

The *brpA* gene (gene 26) was cloned from both predicted start codons and expressed in *E. coli* MKH13. Both fragments increased the salt tolerance of MKH13 significantly. Cells expressing the larger fragment (*brpA*$_L$) had the most significant effect (*P* =0.0002) in the presence of 3% NaCl. Although cells expressing the smaller fragment (*brpA*$_S$) had a slower growth profile and a longer lag phase than the larger fragment (*brpA*$_L$), both exhibited a significant growth advantage compared to the *E. coli* MKH13 harbouring the empty plasmid pCI372 (*P* = 0.0039) (Figure 4B). The gene immediately upstream of *brpA* is predicted to encode a 98 amino acid putative membrane protein (putative acyltransferase), which we have named *atfA*, was also cloned in combination with *brpA* (*brpAatfA*). Both genes in combination did not increase the salt tolerance of MKH13 relative to *brpA*$_L$ alone, when grown in LB + 3% NaCl, but the increase in salt tolerance was significant (*P* =0.0002) (Figure 4B).

Figure 4.

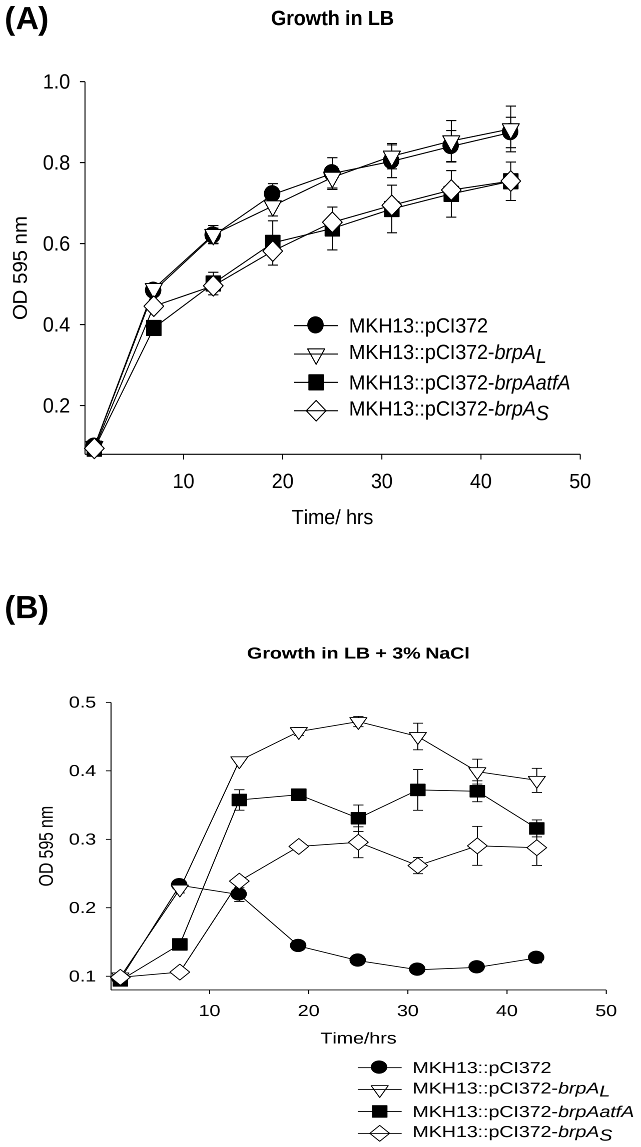


Figure 4. Growth of *E. coli* MKH13::pCI372 and *E.coli* MKH13 carrying a plasmid encoded copy of either *brpA*$_{L}$, *brpA*$_{S}$ or *brpAatfA* in **(A)** LB broth or **(B)** LB + 3% NaCl. All of the genes confer a significant salt tolerance phenotype to MKH13 relative to cells with an empty plasmid vector. All values are the average of triplicate experiments and error bars are representative of the standard error of the mean (SEM).

L-proline did not increase salt tolerance further

Once we had shown that the *brpA* gene could confer a salt tolerance phenotype when expressed in *E. coli*, we aimed to decipher the mechanism of action and thereby assign a function to the encoded protein. Given that BLASTP analysis of the BrpA sequence revealed homology to a proline symporter, growth curves were carried out in minimal media supplemented with L-proline and also other common osmoprotectants, betaine and L-carnitine (final concentration of 1mM). However, no growth advantage was seen in the presence of any of the added osmoprotectant compounds, suggesting that BrpA is not an osmoprotectant uptake system.

Functional annotation of *brpA*

BLASTP analysis also revealed that the BrpA protein exhibited homology to a *brp/blh*-family beta-carotene 15,15'-monooxygenase. Such proteins are related to bacteriorhodopsins [24], and are annotated as bacterio-opsin related protein (brp)/brp-like homologue (blh) protein. Brp/Blh proteins have been shown to have beta-carotene 15,15'-monooxygenase activity; cleaving beta-carotene into two molecules of *all-trans* retinal (vitamin A aldehyde) (Figure 5) [25]. The derived retinal is bound by a rhodopsin protein and cells expressing such proteins acquire an orange/red colour, indicative of the presence of retinal in the cell membrane [46-48]. Strains harbouring *brpA* were grown in the presence of beta-carotene and cell pellets were observed for the development of the characteristic red/orange colour. *E. coli* MKH13 cells carrying the *brpA* gene on the pCI372 plasmid did not show any obvious colour development (Figure 6).

Figure 5.

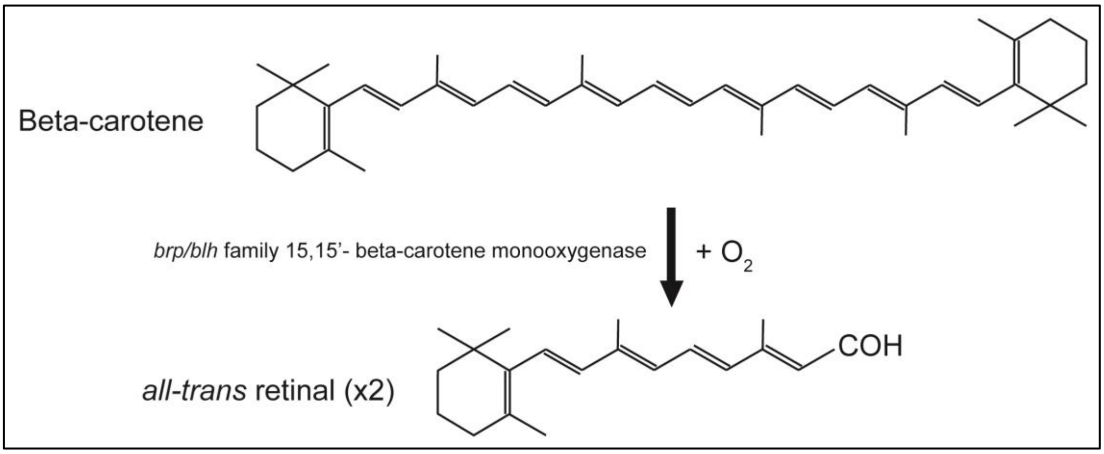


Figure 5. Representation of the reaction for the formation of retinal. Beta-carotene is cleaved at its central 15,15' bond by *brp* 15,15'- beta-carotene monooxygenase to form two molecules of *all-trans* retinal (Vitamin A aldehyde).

Figure 6.

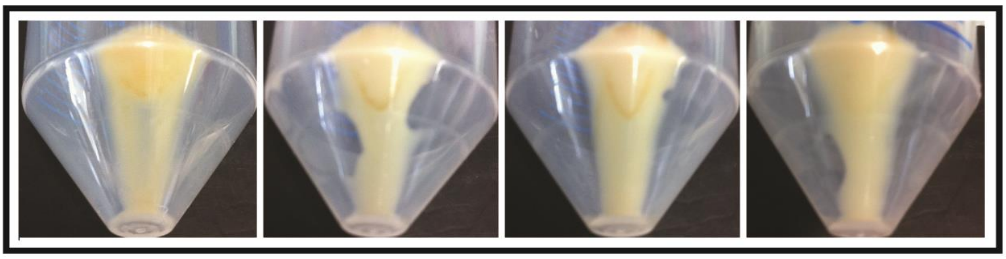


Figure 6. Appearance of cell pellets grown in LB supplemented with beta-carotene. From left to right: *E. coli* MKH13::pCI372, MKH13::pCI372-*brpA*$_L$, MKH13::pCI372-*brpA*$_S$ and MKH13::pCI372-*brpAatfA*.

Given that a number of previous studies have reported the use of an inducible vector to visualise pigmentation in cell pellets [46-50], we cultured the original fosmid clones (which can be induced due to Copy Control^{TM} capability of pCC1FOS fosmid vector) in the presence of beta-carotene and included an induction solution to induce the fosmid from low to high-copy number. The cell pellets developed an intense red/orange colour while cells with an empty vector did not (Figure 7). To confirm that the BrpA protein was responsible for this phenotype, we cloned *brpA* in isolation into the pBAD inducible expression vector and transformed it into *E. coli* EPI300 and repeated the growth experiments. Again, the cell pellets developed a distinctive a red/orange colour (Figure 8).

Figure 7.

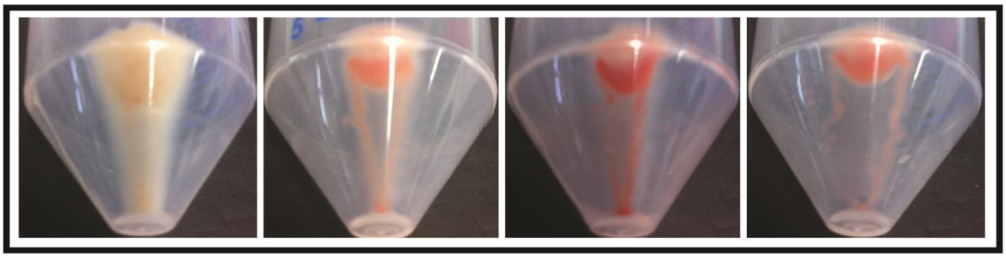


Figure 7. Appearance of cell pellets of clones grown in LB supplemented with beta-carotene and Copy Control^{TM} Induction solution (L-arabinose) (Epicentre Biotechnologies). From left to right: *E. coli* EPI300::pCC1FOS, SMG 1, SMG 6 and SMG 52**.**

Figure 8.

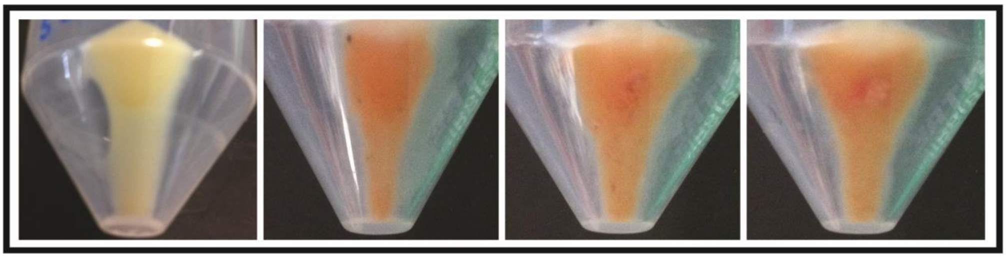


Figure 8. Appearance of cell pellets grown in LB supplemented with beta-carotene and L-arabinose (Sigma). From left to right: *E. coli* EPI300::pBAD, EPI300::pBAD-*brpA*$_S$, EPI300::pBAD-*brpA*$_L$, and EPI300::pBAD-*brpAatfA*.

brpA **also confers salt tolerance to** ***E. coli*** **EPI300**

The genes (*brpA*$_L$*, brpA*$_S$ *and brpAatfA*) were also cloned into the pBAD expression vector and transformed into *E. coli* EPI300. All of the transformed strains exhibited increased salt tolerance relative to the host containing the empty pBAD vector, although EPI300::pBAD-*brpA*$_S$, to a lesser extent, similar to our observations with MKH13 above (Figure 9B).

Figure 9.

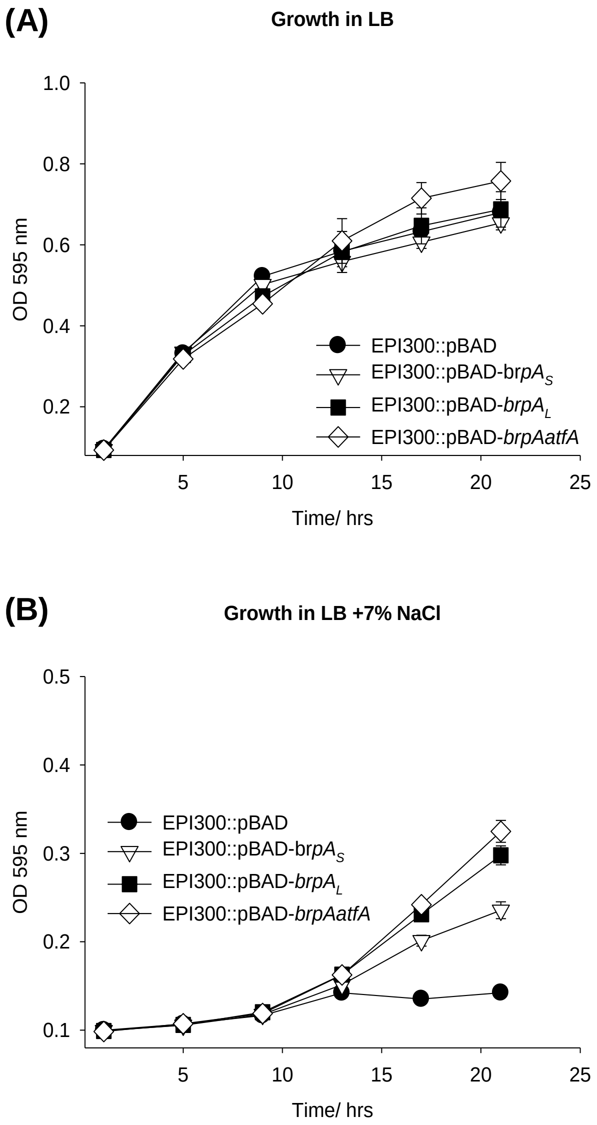


Figure 9. Growth of *E. coli* EPI300::pBAD and EPI300::pBAD-*brpA*$_S$ (*P* = 0.0008), EPI300::pBAD-*brpA*$_L$ (*P* = 0.0002) and EPI300::pBAD-*brpAatfA* (*P* = 0.0001) in **(A)** LB broth and **(B)** LB broth supplemented with 7% NaCl. All three strains had a statistically significant increased salt tolerance compared to EPI300 carrying an empty copy of the pBAD vector. Numbers in parentheses indicate significant *P* values (unpaired student *t*-test). All values are the average of triplicate experiments and error bars are representative of the standard error of the mean (SEM).

Effect of beta-carotene on survival in high-salt media

The effect of beta-carotene on survival of both *E. coli* MKH13 and EPI300 strains was assessed. Survival of strains carrying a plasmid-encoded copy of *brpA* was compared to controls (carrying an empty plasmid) in high-salt media (3% NaCl for MKH13 and 7% for EPI300) in the presence and absence of beta-carotene after a 48-hour period, both aerobically and anaerobically (Figure 10). Anaerobic conditions were tested due to the oxygen (O_2) requirement of beta-carotene 15,15’-monooxygenases for functionality. Beta-carotene did not provide an osmoprotective effect during salt stress to control strains or strains carrying a copy of the *brpA* gene under the conditions tested, however an increased salt tolerance phenotype was observed under both aerobic and anaerobic conditions.

Figure 10.

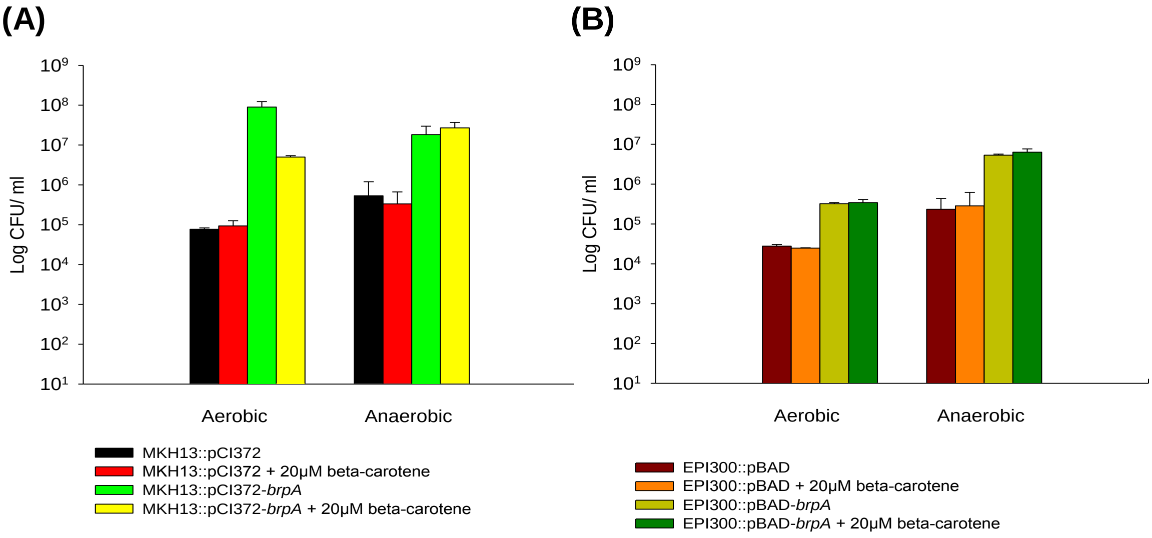


Figure 10. The effect of beta-carotene on the survival of MKH13 strains and EPI300 strains was assessed under aerobic and anaerobic conditions in (A) LB broth with 3% NaCl and (B) LB broth with 7% NaCl. Viable cells were determined by calculating the average CFU per millilitre after 48 hours. Results are representative of triplicate experiments and error bars are the standard error of the mean (SEM).

Transposon mutagenesis

Transposon mutagenesis was performed using the EZTn*5 in vitro* transposition system (Epicentre Biotechnologies) to create knock-out mutants of SMG 6. Clones harbouring a transposon insertion in the *brpA* and neighbouring genes were identified by PCR. The primer pair *brpAatfA* FP and RP were used to amplify this region, generating PCR products of ~1.4 kb in the absence of a transposon insertion and products of ~3.3 kb if the transposon was present (Figure 11A). Once positive clones were identified, the location of the transposon was confirmed by sequencing from the ends of the transposon. We identified four transposon mutants in SMG 6; namely 6-EZTn #24, #26, #34 and #38. The location of the transposon insertions are presented in Figure 11B. The aim was to identify clones that lacked pigmentation following transposition. Clones containing a transposon insertion do not display the same intense red pigmentation seen with SMG 6 and although there is visibly less pigmentation, some residual colour nevertheless remains (Figure 11C).

Figure 11.

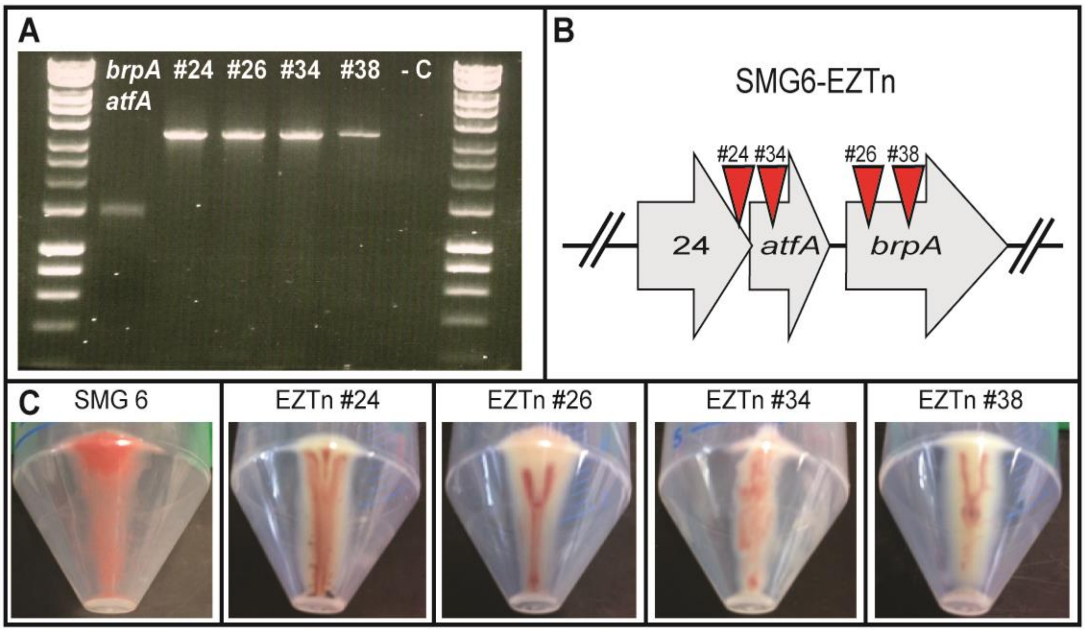


Figure 11. EZTn*5* transposon mutagenesis of SMG 6 was performed to identify mutants lacking pigmentation when grown in the presence of beta-carotene. Clones positive for a transposon in this region of SMG 6 fosmid insert were identified by PCR, with amplicons of ~3.3kb indicative of an insertion event (Figure 11A). Approximate locations of transposon insertions in relation to *brpA* and neighbouring genes are presented in Figure 11B. Appearance of cell pellets of SMG 6 and transposon insertion mutants (EZTn #24, #26, #34 and #38) following growth in the presence of beta-carotene (Figure 11C).

Discussion

In the current study we have identified and characterised a novel salt tolerance locus from the human gut microbiome. Functional assignment of its encoded protein, BrpA, using BLAST returned homologues mainly annotated as hypothetical or putative membrane proteins. The only clue to the possible function of the protein was that it also shared sequence similarity (albeit at <30%) to a proline symporter and a *brp/blh*-family beta-carotene 15,15'-monooxygenase. Sequence homologies of less than 30% are considered to be in the "twilight zone" and confidence of functional annotations diminishes below this threshold [51,52].

Growth experiments in minimal media supplemented with L-proline and other osmoprotectants had no effect on growth or salt tolerance. The gene, which we have termed *brpA*, possibly encodes a putative *brp/blh*-family beta-carotene 15,15'-monooxygenase. Such proteins have been shown to catalyse the conversion of beta-carotene into two molecules of *all-trans* retinal (vitamin A aldehyde) (Figure 5) [24,25]. Growth of the metagenomic clone SMG 6 in the presence of exogenous beta-carotene resulted in the cell pellets with a distinctive orange/red colour. A number of other studies have shown that bacterial cells expressing plasmid encoded beta-carotene biosynthesis genes in addition to a *brp/blh* gene and a proteorhodopsin (PR) encoding gene adopt a similar colour due to the cleavage of beta-carotene to retinal and subsequent binding of retinal by proteorhodopsins in the cell membrane [46-49]. The absence of any obvious PR encoding gene on SMG 6 therefore, does not explain the presence of colour in the SMG clones' cell pellet. Furthermore, when *bprA* was cloned in isolation the cell pellets still had pigmentation, indicating that *brpA* alone is sufficient to confer this phenotype. There are however a few possible explanations for the pigmentation; *in silico* analysis reveals that *brpA* is predicted to have acyltransferase activity (COG 3274), as is *atfA*, the gene immediately upstream of *brpA*. The *atfA* gene was cloned in combination with *brpA*, however expression of both genes together had no appreciable effect on the degree of pigmentation or salt tolerance observed. Carotenoids and retinoids are hydrophobic, lipophilic molecules. The majority of carotenoids are found embedded in the hydrophobic core of lipid membranes and in lipid globules and other hydrophobic environments [53,54]. Acylated carotenoids have been shown to be inserted in the membrane and the predicted acyltransferase activity of BrpA may explain the cell pellet pigmentation in the absence of a rhodopsin protein [55]. In *Staphylococcus aureus*, an acyltransferase is a key enzyme in the biosynthesis pathway for the orange carotenoid staphyloxanthin [56]. This enzyme was initially thought to carry out the final step in staphyloxanthin biosynthesis, although more recently it has been shown that it is actually the penultimate step [57]. The transfer of a polar acyl group or acyl-containing groups such as hydroxyl or keto groups to carotenoids would be likely to enable their interaction with phosphate head groups of lipids, thus anchoring them within membranes [55,58].

The presence of a lipocalin motif was identified in a BLAST homologue of BrpA. Lipocalin proteins can bind hydrophobic molecules such as carotenoids and retinoids. It seems unlikely however, that this is the case with BrpA since the motif is quite different and lacks the characteristic glycine-X-tryptophan (G-X-W) signature found in almost all lipocalins [59]. The BrpA protein has seven predicted transmembrane regions, a characteristic shared with rhodopsin proteins [60]. It has previously been suggested that Brp/Blh-like proteins may be multifunctional and both cleave beta-carotene and subsequently transport or bind the derived *all-trans* retinal, although this has not been demonstrated experimentally [25].

Four transposon mutants of SMG 6 were identified in this study using PCR. It was expected to obtain mutants that lack pigmentation when grown in the presence of beta-carotene. While there is a clear visible difference in the appearance of the cell pellets of the mutants compared to SMG 6, each of the mutants retain some level of pigmentation, albeit to a lesser degree and with diminished colour intensity. Transposon insertion in genes upstream of *brpA* (mutants #24 and #34) indicates a polar effect mediating the reduction in the degree of pigmentation. It is surprising that some pigmentation remains in clones containing a transposon within the *brpA* gene (mutants #26 and #38), indicating residual carotenoid accumulation, possibly due to acyltransferase activity of *atfA*.

The %G+C content of individual genes on SMG 6 drops as low as 30.64% for gene 25 (*atfA*), while its neighbouring gene, *brpA*, is 32.05%. In addition, only 12% of the top 100 BLASTP hits to BrpA are predicted to be from Gram-negative bacteria. The remaining 88% are represented in the main by proteins with similarity to the low G+C, Gram-positive *Firmicutes* phylum, mainly from the genera of *Clostridium, Enterococcus* and *Streptococcus* among others. Taken together, these observations suggest much of this region, including the especially low %G+C, *atfA* and *brpA* genes, were acquired through a LGT event [61,62]. Indeed, in support of this there is evidence that *brp*/*blh*-type genes, along with rhodopsins, undergo frequent LGT events [46,63-65]. In beta-carotene producing bacteria, only these two genes are required to produce retinal which is bound to the rhodopsin protein giving the recipient bacterium the ability to harvest light energy non-photosynthetically and convert it to chemical energy. Acquiring a rhodopsin gene in the gut would be somewhat redundant owning to the aphotic nature of the gut environment. A *brp/blh* beta-carotene monooxygenase however could be beneficial to break down dietary-derived beta-carotene.

There were two possible start codons predicted for the *brpA* gene using the FGENESB gene prediction program. Using the “bacterial generic” parameter as closest organism, a gene ($brpA_S$) encoding a 232 amino acid protein with the alternative initiation codon TTG (leucine) was predicted. Because a number of proteins encoded on SMG 6 shared 100% amino acid identity with *Bacteroides thetaiotaomicron* VPI-5482, this organism was also used as the “closest organism” parameter. Using *B. thetaiotaomicron* VPI-5482 as “closest organism” predicted a gene ($brpA_L$) encoding a 271 amino acid protein with an ATG (methionine) start codon. GeneMark also predicted ATG to be the start codon. Cloning and expression of the gene from both predicted start codons conferred salt tolerance to *E. coli*, although strains expressing the $brpA_L$ fragment had a shorter lag phase and reached a higher final OD. Initially, it seemed likely that ATG was the true start codon of *brpA*, however further manual inspection of the sequences upstream of both start codons revealed a characteristic *Bacteroides* -7/-33 promoter region preceding the TTG codon that deviated from the consensus by only one nucleotide. There is also a potential *Bacteroides*-type promoter upstream of ATG, but at position -14/-30 (GGTATTTG/TTTT). It therefore seems likely that TTG is the actual start codon in *Bacteroides*. Interestingly, previous studies have shown that the use of alternative initiation codons, other than ATG, is a common feature of osmotolerance genes in a number of gastrointestinal pathogens [12,17]. The increased salt tolerance phenotype of *brpA*$_L$ compared to *brpA*$_S$ may be due to the fact that ATG is the most commonly utilised codon to initiate translation (~90% of genes) in *E. coli* [66] and also the presence of strong RBS (AGUAGGU) upstream of the ATG start codon, which differs from the *E. coli* consensus RBS (AGGAGGU) by only one nucleotide. Taken together, the ATG start codon and strong *E. coli* RBS likely gives rise to more efficient levels of transcription and translation, as well as increased expression of *brpA* in *E. coli*, at least under the conditions tested in the current study. It is of course possible that the two protein types (long and short) are expressed under different environmental conditions, as was previously reported for the multi-stress resistance locus HtrA [67].

The presence of a putative RpoS binding site is predicted upstream of the ATG start codon of *brpA*. The alternative sigma factor (sigma 38) RpoS is the master regulator of the general stress response induced during stationary phase in *E. coli* and other Gram-negative bacteria [68]. In addition RpoS regulates the expression of a large number of genes in response to various stresses, including salt stress [69-71]. There is also a putative OxyR binding site in the upstream region of *brpA*. OxyR is a regulator of the oxidative stress response in many bacteria [72] and carotenoids can function as anti-oxidants and can increase resistance to oxidative stress [73,74]. It may be possible that the *brpA* gene is transcribed from two promoters under different environmental conditions, similar to the type of regulation seen with the osmoprotectant transporter ProP in *E. coli,* where the *proP* gene is transcribed from promoter 1 (P1) primarily in response to changes in osmolarity and from promoter 2 (P2) during stationary phase [75,76].

The BrpA amino acid sequence was used to BLAST search against all metagenomes from the HMP dataset at the lowest (1e-^{05}) and highest (1e-^{50}) e-value. Hits to BrpA were most abundant in the stool, supra-gingival plaque and tongue metagenome samples at the lowest e-value (Figure 3B). The majority of these hits had quite low percentage identities in the range of 25%-35%. When the e-value cut-off was increased to 1e-^{50} only 13 putative BrpA homologues were identified and only from the stool metagenome samples (Figure 3A) and would therefore appear to be a rare gene found in some strains of *Bacteroides thetaiotaomicron*, which is one of the most abundant species in the human gut microbiome, having been shown to comprise 6% of all bacteria among the human gut microbiota [77]. It is interesting that homologues of this gene are found most abundantly in body sites (tongue, sub- and supragingival plaque and gut lumen/stool) where the microbiota would encounter beta-carotene (i.e. from dietary sources).

Carotenoids have been shown to protect cells from various environmental stresses such as osmotic, oxidative and light as well as reinforcing and providing increased membrane rigidity [54,74,78-80]. In this study, beta-carotene however did not provide any further increase in salt tolerance under the conditions tested and therefore does not appear to function in an osmoprotective capacity. Acyltransferase enzymes have also been linked to various stress responses, including osmotic stress. For example, the acyltransferase HtrB, provides protection against and exhibits increased expression in response to heat, acid, oxidative and osmotic stress in *Campylobacter jejuni* and *Salmonella typhimurium* [81], while acyltransferases have also been linked to the stress response in *Pseudomonas putida* [82].

In the current study we have used a combined functional metagenomic and bioinformatic approach to identify a novel gene from the human gut microbiome that has not previously been linked to salt tolerance. The gene, *brpA*, encodes a protein with homology to a *brp/blh*-family beta-carotene 15,15'-monooxygenase. When expressed in *E. coli*, BrpA confers salt tolerance phenotype and cell pellets adopt a red/orange pigmentation when grown in the presence of exogenous beta-carotene.

Acknowledgements

The Alimentary Pharmabiotic Centre is a research centre funded by Science Foundation Ireland (SFI grant number 07/CE/B1368). We acknowledge the continued financial assistance of the Alimentary Pharmabiotic Centre, funded by Science Foundation Ireland. JRM acknowledges funding from The Royal Society which supports the bioinformatic cluster (Hive) at Cardiff University, School of Biosciences. RDS is an ESCMID Research Fellow and Coordinator of the EU FP7 IAPP project ClouDx-i.

References

1. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331: 463-467.

2. Iwai S, Chai B, Sul WJ, Cole JR, Hashsham SA, et al. (2010) Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. ISME J 4: 279-285.

3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

4. Culligan EP, Sleator RD, Marchesi JR, Hill C (2013) Metagenomics and novel gene discovery: Promise and potential for novel therapeutics. Virulence 5.

5. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, et al. (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol 68: 4301-4306.

6. Heath C, Hu XP, Cary SC, Cowan D (2009) Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Appl Environ Microbiol 75: 4657-4659.

7. Kallifidas D, Kang HS, Brady SF (2012) Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. J Am Chem Soc 134: 19552-19555.

8. Lammle K, Zipper H, Breuer M, Hauer B, Buta C, et al. (2007) Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. J Biotechnol 127: 575-592.

9. Feeney A, Sleator RD (2012) The human gut microbiome: the ghost in the machine. Future Microbiol 7: 1235-1237.

10. Sleator RD (2010) The human superorganism - of microbes and men. Med Hypotheses 74: 214-215.

11. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

12. Hoffmann RF, McLernon S, Feeney A, Hill C, Sleator RD (2013) A single point mutation in the listerial betL sigma (A) -dependent promoter leads to improved osmo- and chill-tolerance and a morphological shift at elevated osmolarity. Bioengineered 4.

13. Louis P, O'Byrne CP (2010) Life in the gut: microbial responses to stress in the gastrointestinal tract. Sci Prog 93: 7-36.

14. Sleator RD, Watson D, Hill C, Gahan CG (2009) The interaction between Listeria monocytogenes and the host gastrointestinal tract. Microbiology 155: 2463-2475.

15. Hill C, Cotter, P.D., Sleator, R.D. and Gahan, C.G.M. (2001) Bacterial stress response in *Listeria monocytogenes*: jumping the hurdles imposed by minimal processing. International Dairy Journal 12: 273-283.

16. Sleator RD, Hill C (2002) Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. FEMS Microbiol Rev 26: 49-71.

17. Sleator RD, Gahan CG, Hill C (2003) A postgenomic appraisal of osmotolerance in Listeria monocytogenes. Appl Environ Microbiol 69: 1-9.

18. Epstein W (2003) The roles and regulation of potassium in bacteria. Prog Nucleic Acid Res Mol Biol 75: 293-320.

19. Kempf B, Bremer E (1998) Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. Arch Microbiol 170: 319-330.

20. Kunte HJ (2006) Osmoregulation in Bacteria: Compatible Solute Accumulation and Osmosensing. Environmental Chemistry 3: 94-99.

21. Culligan EP, Marchesi JR, Hill C, Sleator RD (2012) Mining the human gut microbiome for novel stress resistance genes. Gut Microbes 3: 394-397.

22. Culligan EP, Sleator RD, Marchesi JR, Hill C (2012) Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. ISME J 6: 1916-1925.

23. Culligan EP, Sleator RD, Marchesi JR, Hill C (2013) Functional Environmental Screening of a Metagenomic Library Identifies *stlA*; A Unique Salt Tolerance Locus from the Human Gut Microbiome. PLOS ONE 8: e82985.

24. Kim YS, Kim NH, Yeom SJ, Kim SW, Oh DK (2009) In vitro characterization of a recombinant Blh protein from an uncultured marine bacterium as a beta-carotene 15,15'-dioxygenase. J Biol Chem 284: 15781-15793.

25. Peck RF, Echavarri-Erasun C, Johnson EA, Ng WV, Kennedy SP, et al. (2001) brp and blh are required for synthesis of the retinal cofactor of bacteriorhodopsin in Halobacterium salinarum. J Biol Chem 276: 5739-5744.

26. Haardt M, Kempf B, Faatz E, Bremer E (1995) The osmoprotectant proline betaine is a major substrate for the binding-protein-dependent transport system ProU of Escherichia coli K-12. Mol Gen Genet 246: 783-786.

27. Hayes F, Daly C, Fitzgerald GF (1990) Identification of the Minimal Replicon of Lactococcus lactis subsp. lactis UC317 Plasmid pCI305. Appl Environ Microbiol 56: 202-209.

28. Jones BV, Marchesi JR (2007) Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. Nat Methods 4: 55-61.

29. Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33: W451-454.

30. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.

31. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-37.

32. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225-229.

33. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-166

34. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33: W244-248.

35. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6: 175-182.

36. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, et al. (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112: 531-552.

37. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33: W284-288.

38. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: D534-538.

39. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, et al. (2012) IMG/M-HMP: a metagenome comparative analysis system for the Human Microbiome Project. PLoS One 7: e40151.

40. Human_Microbiome_Project_Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.

41. Hoffmann T, von Blohn C, Stanek A, Moses S, Barzantny H, et al. (2012) Synthesis, release, and recapture of compatible solute proline by osmotically stressed Bacillus subtilis cells. Appl Environ Microbiol 78: 5753-5762.

42. Sleator RD, Gahan CG, Hill C (2001) Identification and disruption of the proBA locus in Listeria monocytogenes: role of proline biosynthesis in salt tolerance and murine infection. Appl Environ Microbiol 67: 2571-2577.

43. Bayley DP, Rocha ER, Smith CJ (2000) Analysis of cepA and other Bacteroides fragilis genes reveals a unique promoter structure. FEMS Microbiol Lett 193: 149-154.

44. Vingadassalom D, Kolb A, Mayer C, Rybkine T, Collatz E, et al. (2005) An unusual primary sigma factor in the Bacteroidetes phylum. Mol Microbiol 56: 888-902.

45. Flower DR (1996) The lipocalin protein family: structure and function. Biochem J 318 (Pt 1): 1-14.

46. Martinez A, Bradley AS, Waldbauer JR, Summons RE, DeLong EF (2007) Proteorhodopsin photosystem gene expression enables

photophosphorylation in a heterologous host. Proc Natl Acad Sci U S A 104: 5590-5595.

47. Wang Z, O'Shaughnessy TJ, Soto CM, Rahbar AM, Robertson KL, et al. (2012) Function and regulation of Vibrio campbellii proteorhodopsin: acquired phototrophy in a classical organoheterotroph. PLoS One 7: e38749.

48. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289: 1902-1906.

49. Sabehi G, Loy A, Jung KH, Partha R, Spudich JL, et al. (2005) New insights into metabolic properties of marine bacteria encoding proteorhodopsins. PLoS Biol 3: e273.

50. von Lintig J, Vogt K (2000) Filling the gap in vitamin A research. Molecular identification of an enzyme cleaving beta-carotene to retinal. J Biol Chem 275: 11915-11920.

51. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85-94.

52. Sleator RD (2012) Proteins: form and function. Bioeng Bugs 3: 80-85.

53. Kerfeld CA, Sawaya MR, Brahmandam V, Cascio D, Ho KK, et al. (2003) The crystal structure of a cyanobacterial water-soluble carotenoid binding protein. Structure 11: 55-65.

54. Gruszecki WI, Strzalka K (2005) Carotenoids as modulators of lipid membrane physical properties. Biochim Biophys Acta 1740: 108-115.

55. Maresca JA, Bryant DA (2006) Two genes encoding new carotenoid-modifying enzymes in the green sulfur bacterium Chlorobium tepidum. J Bacteriol 188: 6217-6223.

56. Pelz A, Wieland KP, Putzbach K, Hentschel P, Albert K, et al. (2005) Structure and biosynthesis of staphyloxanthin from Staphylococcus aureus. J Biol Chem 280: 32493-32498.

57. Kim SH, Lee PC (2012) Functional expression and extension of staphylococcal staphyloxanthin biosynthetic pathway in Escherichia coli. J Biol Chem 287: 21575-21583.

58. Britton G (1995) Structure and properties of carotenoids in relation to function. FASEB J 9: 1551-1558.

59. Pevsner J (2009) Bioinformatics and Functional Genomics: 2nd Edition.: John Wiley & Sons Inc, Hoboken, New Jersey, USA.

60. Spudich JL, Sineshchekov OA, Govorunova EG (2013) Mechanism divergence in microbial rhodopsins. Biochim Biophys Acta.

61. Sleator RD (2013) A Beginner's Guide to Phylogenetics. Microb Ecol 66: 1-4.

62. Sleator RD (2011) Phylogenetics. Arch Microbiol 193: 235-239.

63. de la Torre JR, Christianson LM, Beja O, Suzuki MT, Karl DM, et al. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc Natl Acad Sci U S A 100: 12830-12835.

64. Sharma AK, Spudich JL, Doolittle WF (2006) Microbial rhodopsins: functional versatility and genetic mobility. Trends Microbiol 14: 463-469.

65. McCarren J, DeLong EF (2007) Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. Environ Microbiol 9: 846-858.

66. Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol 20: 616-622.

67. Stack HM, Sleator RD, Bowers M, Hill C, Gahan CG (2005) Role for HtrA in stress induction and virulence potential in Listeria monocytogenes. Appl Environ Microbiol 71: 4241-4247.

68. Battesti A, Majdalani N, Gottesman S (2011) The RpoS-mediated general stress response in Escherichia coli. Annu Rev Microbiol 65: 189-213.

69. Bohringer J, Fischer D, Mosler G, Hengge-Aronis R (1995) UDP-glucose is a potential intracellular signal molecule in the control of expression of sigma S and sigma S-dependent genes in Escherichia coli. J Bacteriol 177: 413-422.

70. Hengge-Aronis R, Klein W, Lange R, Rimmele M, Boos W (1991) Trehalose synthesis genes are controlled by the putative sigma factor encoded by rpoS and are involved in stationary-phase thermotolerance in Escherichia coli. J Bacteriol 173: 7918-7924.

71. Cheville AM, Arnold KW, Buchrieser C, Cheng CM, Kaspar CW (1996) rpoS regulation of acid, heat, and salt tolerance in Escherichia coli O157:H7. Appl Environ Microbiol 62: 1822-1824.

72. Chiang SM, Schellhorn HE (2012) Regulators of oxidative stress response genes in Escherichia coli and their functional conservation in bacteria. Arch Biochem Biophys 525: 161-169.

73. Clauditz A, Resch A, Wieland KP, Peschel A, Gotz F (2006) Staphyloxanthin plays a role in the fitness of Staphylococcus aureus and its ability to cope with oxidative stress. Infect Immun 74: 4950-4953.

74. Kelman D, Ben-Amotz A, Berman-Frank I (2009) Carotenoids provide the major antioxidant defence in the globally significant N2-fixing marine cyanobacterium Trichodesmium. Environ Microbiol 11: 1897-1908.

75. Xu J, Johnson RC (1997) Activation of RpoS-dependent proP P2 transcription by the Fis protein in vitro. J Mol Biol 270: 346-359.

76. Xu J, Johnson RC (1997) Cyclic AMP receptor protein functions as a repressor of the osmotically inducible promoter proP P1 in Escherichia coli. J Bacteriol 179: 2410-2417.

77. Bjursell MK, Martens EC, Gordon JI (2006) Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. J Biol Chem 281: 36269-36279.

78. Kim SH, Ahn YO, Ahn MJ, Lee HS, Kwak SS (2012) Down-regulation of beta-carotene hydroxylase increases beta-carotene and total carotenoids enhancing salt stress tolerance in transgenic cultured cells of sweetpotato. Phytochemistry 74: 69-78.

79. Mathews-Roth MM (1987) Photoprotection by carotenoids. Fed Proc 46: 1890-1893.

80. Mishra NN, Liu GY, Yeaman MR, Nast CC, Proctor RA, et al. (2011) Carotenoid-related alteration of cell membrane fluidity impacts Staphylococcus aureus susceptibility to host defense peptides. Antimicrob Agents Chemother 55: 526-531.

81. Phongsisay V, Perera VN, Fry BN (2007) Expression of the htrB gene is essential for responsiveness of Salmonella typhimurium and Campylobacter jejuni to harsh environments. Microbiology 153: 254-262.

82. Reva ON, Weinel C, Weinel M, Bohm K, Stjepandic D, et al. (2006) Functional genomics of stress response in Pseudomonas putida KT2440. J Bacteriol 188: 4079-4092.

CHAPTER V

Functional Environmental Screening of a Metagenomic Library Identifies *stlA*: A Unique Salt Tolerance Locus from the Human Gut Microbiome

PLOS ONE, 2013 DEC 12;8(12): E82985.

This Chapter may differ in layout from published manuscript as associated supplementary figures and tables have now been included in the main text.

Abstract

Functional environmental screening of metagenomic libraries is a powerful means to identify and assign function to novel genes and their encoded proteins without any prior sequence knowledge. In the current study we describe the identification and subsequent analysis of a salt-tolerant clone from a human gut metagenomic library. Following transposon mutagenesis we identified an unknown gene (*stlA*, for "salt tolerance locus A") with no current known homologues in the databases. Subsequent cloning and expression in *Escherichia coli* MKH13 revealed that *stlA* confers a salt tolerance phenotype in its surrogate host. Furthermore, a detailed *in silico* analysis was also conducted to gain additional information on the properties of the encoded StlA protein. The *stlA* gene is rare when searched against human metagenome datasets such as MetaHit and the Human Microbiome Project and represents a novel and unique salt tolerance determinant which appears to be found exclusively in the human gut environment.

Introduction

The human gastrointestinal (GI) tract is home to hundreds of bacterial species [1] which play an important and complex role in host health, metabolism and physiology [2]. This relatively diverse community is dominated by two bacterial phyla; the *Bacteroidetes* and *Firmicutes*, with most of the remaining microbes represented by members of the *Actinobacteria*, *Proteobacteria*, *Verrucomicrobia*, and *Fusobacteria* [3]. A significant proportion (estimates range from approximately 50-80%) of this bacterial community has thus far proved recalcitrant to traditional laboratory culture [3,4], although that number is constantly decreasing [5-7]. The emergence of culture-independent techniques such as metagenomics in the past 10-15 years has enabled researchers to study these “unculturable organisms” (although “as-yet uncultured” would be more accurate) through direct sequencing of metagenomic DNA or through cloning and functional expression in a heterologous host - an approach referred to as functional genomics [8,9].

The human GI tract imposes numerous stresses on its resident and transient microbiota [10]. The ability to adapt to and resist conditions such as low pH, bile acids, elevated osmolarity, nutrient limitation, host immune factors and competing microorganisms is a determining factor in niche colonisation and proliferation [11]. Our research focuses specifically on the osmotic stress response. Bacteria generally elicit a phased response when challenged in such a manner, firstly by the rapid accumulation of potassium (K^+) ions (primary response), followed by the synthesis or accumulation of osmoprotectant compounds (secondary response) [12-15]. A third mechanism is also employed which can involve a broad range of genes that are seemingly unrelated to the primary and secondary responses [16-19]. These atypical, ancillary systems are arguably more interesting and provide a more complete view of the cellular response to osmotic stress in different bacteria and may also identify specific strategies employed by specific bacteria in distinct environments.

Our aim was to identify novel genes encoding proteins that could confer a salt tolerance phenotype. It is hoped that the identification of atypical genes, which have not previously been linked to salt tolerance will help to broaden our understanding and possibly lead to the identification of novel and unusual systems that play as yet undefined roles in salt tolerance. While sequenced-based metagenomics can define the abundance and diversity of different bacteria within a given microbiome, it cannot enable researchers to assign novel functions to new or known genes. This task can only be achieved through functional screening of metagenomic libraries using activity-based assays. Approximately 30-40% of genes in a given genome will be annotated as hypothetical, conserved hypothetical or function unknown [20], while ~75% of functions important for life in the gut consist of uncharacterized orthologous groups and/or completely novel gene families [1], emphasising the significant degree of novelty that exists in these (meta)genomes.

A previous study from our group identified five genes (which were previously annotated) from the human gut microbiome, to which a novel function of salt tolerance could be assigned [21]. In the current study, we describe the identification of a gene with no currently known homologues.

Bioinformatic analysis suggests that the gene encodes a putative membrane protein, while transposon mutagenesis and subsequent cloning and heterologous expression of the gene revealed that it conferred a salt tolerance phenotype in *Escherichia coli*. This study illustrates the power of functional environmental screening of metagenomic libraries as means to identify and assign a function to as yet unknown genes and their encoded proteins.

Materials and methods

Bacterial strains and growth conditions

Bacterial strains and plasmids used in this study are listed in Table 1, while primers (Eurofins, MWG Operon, Germany) used are listed in Table 2.

E. coli EPI300::pCC1FOS (Epicentre Biotechnologies, Madison, WI, USA) was grown in Luria-Bertani (LB) medium containing 12.5μg/ml chloramphenicol (Cm) and in 12.5μg/ml chloramphenicol plus 50μg/ml kanamycin (Kan) following EZ-Tn*5* transposon mutagenesis reactions. *E. coli* MKH13 [22] and *Lactococcus lactis* MG1363 [23] were grown in LB and M17 (+0.5% glucose; GM17 media) media respectively. Media were supplemented with 20µg/ml Cm for strains transformed with the plasmid pCI372 [24]. Media were supplemented with 1.5% (w/v) agar when required. Overnight cultures of *E. coli* were grown at 37°C with shaking, while *L. lactis* cultures were grown statically at 30°C.

Table 1. Strains, plasmids and transposon used in this study.

Strain, plasmid or transposon	Genotype or characteristic(s)	Source or reference
Strains		
E. coli EPI300	F$^-$ *mcrA*Δ(*mrr-hsd*RMS-*mcrBC*) Φ80d*lacZ*ΔM15 Δ*lac*X74 *recA*1 *endA*1 *araD*139 Δ(*ara, leu*)7697 *galU galK* λ$^-$ *rpsL nupG trfA dhfr*; high-transformation efficiency of large DNA	Epicentre Biotechnologies, Madison, WI, USA
E. coli MKH13	MC4100*Δ*(*putPA*)101D(*proP*)2D(*proU*)	[22]
E. coli MKH13::pCI372	MKH13 containing pCI372 shuttle vector	This study
E. coli MKH13::pCI372-*stlA*	MKH13 containing pCI372 with *stlA* gene from human gut metagenomic library clone SMG 25	This study
L. lactis subspecies *cremoris* MG1363	Plasmid-free *Lactococcus* strain	[23]
L. lactis MG1363::pCI372	MG1363 containing pCI372 shuttle vector	This study
L. lactis MG1363::pCI372-*stlA*	MG1363 containing pCI372 with *stlA* gene from human gut metagenomic library clone SMG 25	This study
Plasmids		
pCI372	Shuttle vector between *E. coli* and *L. lactis*, Cm^R	[24]
pCC1FOS	Fosmid cloning vector, Cm^R	Epicentre Biotechnologies, Madison, WI, USA
Transposon		
EZ-Tn*5* <*ori*V/ KAN-2>	Hyperactive Tn*5* transposon, Kan^R, inducible high copy origin of replication – *ori*V	Epicentre Biotechnologies, Madison, WI, USA

Cm^R = chloramphenicol resistance; Kan^R = kanamycin resistance.

Table 2. Primers used in this study.

Primer	Sequence (5' – 3')^a
pCI372 FP	CGGGAAGCTAGAGTAAGTAG
pCI372 RP	CCTCTCGGTTATGAGTTAG
stlA FP	AAAACTGCAGTTCTGGCAGCAGTGATTTTG
stlA RP	GCTCTAGACGGTCGAGCAAGGTAATAGG
stlA-J FP	TGCTCTTCCGAAGCAGTCAG
stlA-J RP	AGCATATCGAAGACGGCCAG
stlA-OUT FP	CTGCTCTGTTGATGGGGTTT
stlA-OUT RP	CGGGCAACTACAAGGATGAT
stlA-IN FP	TATGGGAGGGGCTACTACGG
stlA-IN RP	ACCCAGTTGCCAAGCATATC
EZ-Tn FP-1	GCCAACGACTACGCACTAGCCAAC
EZ-Tn RP-1	GAGCCAATATGCGAGAACACCCGAGAA

^{a}Restriction enzyme recognition sequences are underlined; FP= forward primer; RP= reverse primer.

Construction and screening of metagenomic library

A previously constructed fosmid clone library [25,26], created from metagenomic DNA isolated from a faecal sample from a healthy 26 year old Caucasian male was used to screen for salt-tolerant clones. The library was screened as outlined previously [21]. Briefly, a total of 23,040 clones from the library were screened on LB agar supplemented with 6.5% (w/v) NaCl and 12.5μg/ml chloramphenicol using a Genetix QPix 2 XT™ colony picking/gridding robotics platform. Plates were incubated at 37°C for 2-3 days and checked periodically for growth of likely salt-tolerant clones.

Transposon mutagenesis

Transposon mutagenesis was carried out in accordance with the manufacturer's instructions, using the EZ-Tn*5* <*oriV*/ KAN-2> *in vitro* transposition kit (Epicentre Biotechnologies). *E. coli* EPI300 cells were transformed with the transposon reaction mixture and selected on plates containing Cm and Kan (12.5 and 50µg/ml, respectively). The transposon insertion clones were subsequently replica plated onto LB with and without 6.5% added NaCl. Clones which grew on LB but not on LB + 6.5% NaCl suggested a likely insertion event in a gene involved in salt tolerance. Presumptive salt-tolerant knock-outs were grown overnight and a fosmid DNA extraction was performed. The extracted fosmid containing metagenomic DNA was subjected to sequencing from the ends of the transposon using the primers EZ-Tn FP-1 and EZ-Tn RP-1 (Table 2.). All sequencing was performed by GATC Biotech (Konstanz, Germany).

DNA manipulations

Induction of fosmids from low to high copy number for downstream applications such as transposon mutagenesis and sequencing was performed as per manufacturer's instructions and as described previously [21]. The Qiagen QIAprep® Spin mini-prep kit was used to extract fosmids as per manufacturer's instructions. PCR products were purified with a Qiagen PCR purification kit and digested with restriction enzymes *XbaI* and *PstI* (Roche Applied Science), followed by ligation using the Fast-Link DNA ligase kit (Epicentre Biotechnologies) to similarly digested plasmid pCI372. Electro-competent *E. coli* MKH13 and *L. lactis* MG1363 were transformed with the ligation mixture and plated on LB and GM17 agar respectively, containing 20µg/ml Cm for selection. Colony PCR was performed on resistant transformants using a primer on the *stlA* gene (*stlA* FP) and a primer on the plasmid (pCI372 RP) to confirm the presence and size of the insert.

Detection of the *stlA* gene in metagenomic DNA isolated from human stool samples was attempted using PCR. Twenty samples from the ELDERMET study [27], which consisted of five community (healthy), five long stay (frail) old subjects and five healthy young and five young subjects with irritable bowel syndrome (IBS) were used as template DNA. Furthermore, five samples from healthy adults from a separate study {Knopp, 2010 #464} were also tested using the following primer pairs: *stlA* FP and *stlA* RP, *stlA*-J FP and *stlA*-J RP, *stlA*-OUT FP and *stlA*-OUT RP, *stlA*-IN FP and *stlA*-IN RP (see Table 2).

Growth experiments

Cultures were grown overnight in appropriate media. Cells were subsequently harvested, washed in one quarter strength sterile Ringer's solution and re-suspended in fresh broth. A 2% inoculum was sub-cultured in fresh broth containing the appropriate stress (i.e. sodium chloride (NaCl), potassium chloride (KCl), sucrose, glycerol, low pH or bile as required) and 200µl was transferred to individual wells of a sterile 96-well micro-titre plate (Sarstedt Inc. Newton, USA). Plates were incubated at 37°C (or 30° for *L. lactis* strains) for 24-48 hours in an automated spectrophotometer (Tecan Genios) which recorded the optical density at 595 nanometres (OD_{595nm}) every hour. For experiments using bile, uninoculated media containing bile were dispensed as blanks in the 96-well plate and their OD_{595nm} values were subtracted from the corresponding inoculated wells to give the OD_{595nm} for the microbial fraction. The data was subsequently retrieved and analysed using the Magellan 3 software program. Representative graphs were created using the Sigma Plot 10.0 software programme (Systat Software Inc, London, UK). Results are presented as the average of triplicate experiments, with error bars being representative of the standard error of the mean (SEM).

BIOLOG Phenotype Microarray (PM) Assay

The phenotype microarray (PM) osmolytes microplate (PM9) was used to compare the cellular phenotypes [29] of *E. coli* MKH13::pCI372 and MKH13::pCI372-*stlA* under 96 different conditions. The BIOLOG PM protocol for *E. coli* and other Gram-negative bacteria was followed for preparation of the different inoculating fluid (IF) solutions (IF-0 and IF-10;

supplied by BIOLOG) and inoculation of the PM plates. Briefly, isolated colonies were added to IF-0 fluid until a cell suspension of 42% T (transmittance) was achieved. This was subsequently diluted in IF-0 + dye mix A to achieve 85% T. Finally, this was diluted in IF-10 + dye mix A and 100ul was inoculated to each well of the PM 9 microplates. Plates were incubated at 37ºC and readings were taken over a 24 hour period using an automated plate reader (BIOTEK Synergy 2) which measured the absorbance at 590nm.

Sequencing and bioinformatic analysis

The fosmid insert from clone SMG 25 was fully sequenced and assembled by GATC Biotech (Konstanz, Germany) using the GS FLX (Roche) pyrosequencing platform on a titanium mini-run. Putative open reading frames were predicted using Softberry FGENESB bacterial operon and gene prediction software (available at www.softberry.com). Retrieved nucleotide and translated amino acid sequences were functionally annotated by homology searches using the Basic Local Alignment and Search Tool (BLAST) using a maximum e-value cut-off of 1e^{-03}, to identify homologous sequences from the National Centre for Biotechnology Information (NCBI) website: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.

The following databases and tools were used to gain additional information on the StlA protein: Expasy ProtParam server, Conserved Domain Database (CDD), PROSITE motif search, SignalP 4.0, HMMER, TMHMM, HHPred, Softberry BProm promoter search (www.softberry.com), SOPMA, SWISS

MODEL, iTasser and QUARK. Relevant information and results can be found in Table 4 [30-41].

The Fold and Functional Assignment System (FFAS03) is a profile-profile and fold recognition algorithm that can detect remote homology between proteins [42]. FFAS03 searches numerous databases including non-redundant NCBI protein sequence database (NCBI nr), Global Ocean Sampling (GOS) from JCVI (J. Craig Venter Institute), PDB (Protein Data Bank), SCOP (Structural Classification of Proteins), and COG (Clusters of Orthologous Groups), as well as numerous metagenome datasets (microbial metagenome samples from the Joint Genome Institute, human gut metagenome samples from the Hattori lab, human oral microbiome database from the Forsyth institute and GOS data from JCVI and CAMERA). Furthermore, FFAS03 searches against the MetaHit (Metagenomics of the Human Intestinal Tract) dataset [1], which contains over 3 million unique gene sequences from the human gut microbiome. The StlA protein sequence was submitted to the server to identify proteins with distant homology based on FFAS profiling or homologues by BLAST and PSI-BLAST (Position-Specific Iterated BLAST) against the databases and metagenome datasets. The FFAS03 server can be found at: http://ffas.burnham.org/ffas-cgi/cgi/document.pl

The Integrated Microbial Genomes and Metagenomes (IMG/M) [43] is a data management system for the comparative analysis of metagenome sequence data. IMG/M-HMP [44] specifically contains metagenome data from the Human Microbiome Project (HMP) [45]. It contains 748 metagenome datasets generated from sequencing samples from 17 different body sites (number of samples from each site are in brackets); anterior nares (94), keratinised gingiva (6), buccal mucosa (122), hard palate (1), retroauricular crease (20), left (2) and right retroauricular crease (5), palatine tonsils (6), saliva (5), throat (7), tongue dorsum (133), sub- (8) and supra-gingival plaques (127), mid-vagina (2), posterior fornix (60), vaginal introitus (3) and stool (147). It also provides tools for comparative analysis between hosted sequences and user supplied sequences. The StlA protein sequence was searched (maximum e-value cut-off of 1e^{-50}) against all the available metagenomes (748) from the HMP as well against all bacterial (9049), archaeal (323), viral/phage (2809) and eukaryotic genome (183) sequences (both assembled and draft), as well as all sequenced plasmids (1193) and all other non-human metagenome datasets (representing >1,300 non-human samples) from diverse environments, including terrestrial, aquatic and host-associated plants and animals (Supplementary File S1*) stored in the database at the time of writing (search date 19/09/13). The IMG/M-HMP server can be found at: http://img.jgi.doe.gov/cgi-bin/imgm_hmp/main.cgi. PhiSpy [46] was used to identify putative prophage genes and the boundaries of the putative prophage region on SMG 25.

* Supplementary File S1 can be accessed on the PLOS ONE wesite at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0082985;jsessionid=1EE719C721B576C1B180513AB3B44430

Taxonomic assignment of scaffolds

The scaffolds on which an *stlA* homologue was identified were subjected to BLASTX analysis (maximum e-value cut-off of 1e^{-50}). The BLASTX results were downloaded and imported to MEGAN 4 (Metagenome Analyser 4software program [47] for taxonomic assignment.

Results

Screening the metagenomic library

Screening approximately 23,000 clones from a human gut metagenomic library led to the identification of 53 clones which were designated as conferring salt-tolerance (i.e. facilitating growth on LB supplemented with 6.5% NaCl, a concentration which inhibits the growth of the cloning host carrying an empty fosmid vector). Six clones (annotated Salt MetaGenome; SMG 1-6) grew within 24 hours and the remaining 47 grew in the following 24-48 hours. SMG 25 represents one of the "late-bloomers" and was chosen at random for analysis. End sequencing of the fosmid insert revealed it shared highest genetic identity to species from the genus *Akkermansia*, namely *Akkermansia muciniphila* ATCC BAA-835*,* *Akkermansia muciniphila* CAG:154 and *Akkermansia sp.* CAG:344. *A muciniphila* ATCC BAA-835, the type strain [48,49], is a mucin degrading member of the Phylum *Verrucomicrobia* which is commonly found in the human gut microbiome [50]. Growth was monitored spectrophotometrically, by measuring the optical density at 595_{nm} (OD_{595nm}). SMG 25 was shown to have a significant (unpaired student *t*-test, *P* <0.0001) growth advantage in the presence of NaCl compared to the EPI300 host strain carrying an empty fosmid vector (pCC1FOS) (Figure 1).

Figure 1.

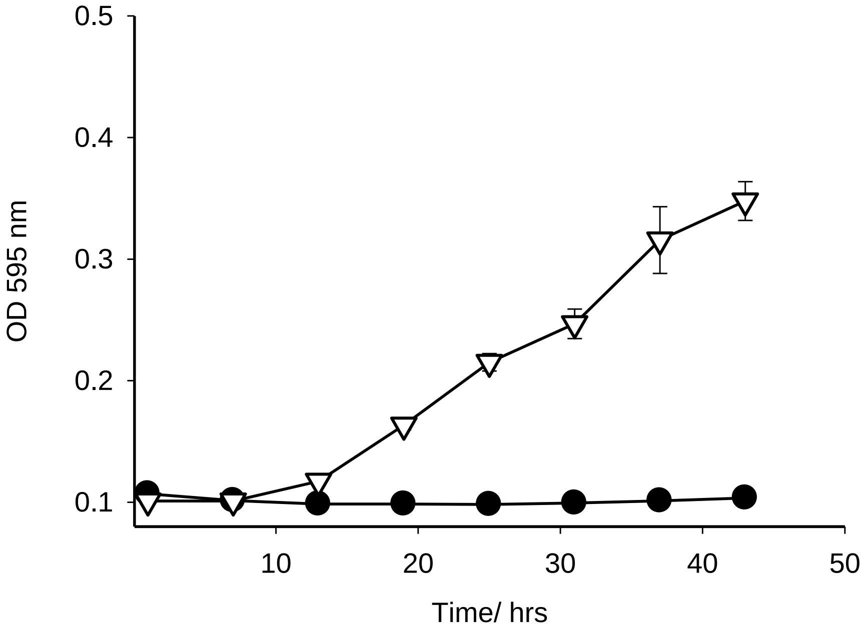


Figure 1: Growth in LB broth supplemented with 6.5% NaCl. Growth of *E. coli* EPI300::pCC1FOS host strain (● closed circle) and SMG 25 (▽) open triangle) (*P* <0.0001).

Fosmid sequencing and analysis

The fosmid insert from SMG 25 was fully sequenced and assembled by GATC Biotech (Konstanz, Germany) and was predicted to contain 45 putative open reading frames that encode proteins (see Table 3). Translated nucleotide sequences were subjected to BLASTP analysis to identify homologous sequences in the database. Twenty-six of the genes encoded proteins corresponding to different species of *Akkermansia* (ranging from 34-98% amino acid identity), but a sizeable proportion of the encoded proteins, approximately 27% (12/45) (highlighted in bold in Table 3) had no significant similarity to sequences in the database, indicating the presence of novel sequences. Overall, only 13 proteins could be assigned a putative function based on BLASTP searches, whilst the remaining genes encoded hypothetical or uncharacterised proteins. The full fosmid insert sequence of SMG 25 can be found in GenBank (accession number=JQ269600.1; gi=375342965). The G+C (guanine and cytosine) skew of the entire fosmid insert as well a picture of the G+C content of each individual gene are presented in Figure 7B and C respectively.

Table 3. List of putative proteins encoded on SMG 25 fosmid insert.

Protein	Length (aa)	Highest similarity (BLASTP)	Highest similarity organism (BLASTP)	E-value	% coverage	% ID (Length of similarity, aa)	Putative conserved domain(s)
1	555	Serine/ threonine protein kinase	*Akkermansia sp.* CAG:344	7.00E-53	50%	44% (285)	None
2	108	Hypothetical protein (Amuc_1368)	*Akkermansia muciniphila* ATCC BAA-835	1.00E-04	86%	34% (98)	None
3	**84**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
4	**160**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
5	**135**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
6*	**257**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
7	**157**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**DnaJ zinc-finger**
8	201	Hypothetical protein O71_18246	*Pontibacter sp.* BAB1700	2.00E-06	36%	39% (77)	DUF4339
9	153	Hypothetical protein HALAR_0188	Halophilic archaeon DL31	1.00E-04	53%	36% (83)	TM2
10	320	Ankyrin repeat protein	*Synergistetes bacterium SGP1*	2.00E-47	88%	45% (291)	Ankyrin repeat
11	73	Uncharacterized protein BN502_01474	*Akkermansia muciniphila* CAG:154	3.00E-04	54%	50% (40)	None
12	**338**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
13	283	Uncharacterised protein BN502_01467	*Akkermansia muciniphila* CAG:154	0.00+00	100%	94% (283)	DUF932
14	99	Uncharacterized protein BN502_01466	*Akkermansia muciniphila* CAG:154	1.00E-59	100%	94% (99)	None
15	52	Uncharacterized protein BN502_01465	*Akkermansia muciniphila* CAG:154	2.00E-22	100%	98% (52)	None
16	160	Uncharacterized protein BN502_01464	*Akkermansia muciniphila* CAG:154	2.00E-99	100%	91% (160)	None
17	43	Uncharacterized protein BN502_01463	*Akkermansia muciniphila* CAG:154	4.00E-04	95%	90% (41)	None
18	317	Hypothetical protein (Amuc_1352)	*Akkermansia muciniphila* ATCC BAA-835	9.00E-08	31%	40% (101)	None
19	**79**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
20	159	Phage-associated protein	*Rhizobium lupini* HPC(L)	3.00E-36	100%	52% (159)	DUF4065, GepA
21	264	Hypothetical protein EC2865200_1013	*Escherichia coli* 2865200	5.00E-26	67%	45% (181)	None
22	154	Uncharacterized protein BN502_01474	*Akkermansia muciniphila* CAG:154	2.00E-31	72%	56% (114)	None
23	**514**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
24	**129**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**Fimbrial OM usher protein**
25	**551**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
26	52	Hypothetical protein HALA3H3_770002	*Halomonas sp.* A3H3	2.00E-04	73%	63% (38)	None
27	657	H(+)-transporting two-sector ATPase	*Akkermansia sp.* CAG:344	0.00E+00	88%	92% (584)	TrkH superfamily
28	445	MATE efflux family protein (Amuc_1131)	*Akkermansia sp.* CAG:344	0.00E+00	95%	87% (445)	MATE, NorM
29	**49**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
30	**54**	**No significant similarity found**	**n/a**	**n/a**	**n/a**	**n/a**	**n/a**
31	278	Putative uncharacterized protein	*Akkermansia sp.* CAG:344	8.00E-56	100%	70% (279)	None
32	87	Putative uncharacterized protein	*Akkermansia sp.* CAG:344	3.00E-31	100%	83% (87)	None
33	186	Hypothetical protein (Amuc_1127)	*Akkermansia muciniphila* ATCC BAA-835	2.00E-61	81%	83% (153)	None
34	450	Dimethyladenosine transferase	*Akkermansia sp.* CAG:344	0.00E+00	99%	91% (448)	ksgA, NUDIX hydrolase
35	393	UDP-galactopyranose mutase	*Chthoniobacter flavus Ellin428*	7.00E-109	96%	52% (381)	GLF, NAD binding

36	329	UDP-glucose 4-epimerase	*Akkermansia muciniphila* ATCC BAA-835	0.00E+00	100%	96% (329)	UDP_G4E_1_SDR_e
37	55	Hypothetical protein (Amuc_1123)	*Akkermansia muciniphila* ATCC BAA-835	6.40E-03	85%	43% (39)	None
38	511	Hypothetical protein (Amuc_1124)	*Akkermansia muciniphila* ATCC BAA-835	0.00E+00	98%	86% (504)	Isoprenoid_C2_like
39	144	Sulphate transporter/anti-sigma factor antagonist	*Akkermansia muciniphila* ATCC BAA-835	4.00E-89	100%	90% (144)	STAS superfamily
40	453	Putative uncharacterized protein	*Akkermansia sp.* CAG:344	3.00E-180	99%	69% (454)	DUF2851
41	466	Glutamate decarboxylase	*Akkermansia muciniphila* CAG:154	0.00E+00	100%	91% (466)	AAT_I superfamily
42	1217	Outer membrane auto-transporter protein	*Akkermansia sp.* CAG:344	3.00E-96	100%	83% (1217)	Auto-transporter superfamily
43	142	Hypothetical protein (ANACAC_03730)	*Anaerostipes caccae DSM 14662*	4.00E-55	99%	69% (141)	NAT_SF domain
44	132	GCN5-related N-acetyltransferase	*Akkermansia sp.* CAG:344	8.00E-56	97%	81% (129)	NAF_SF domain
45	947	DNA polymerase III, alpha subunit	*Akkermansia muciniphila* CAG:154	0.00E+00	100%	95% (938)	DNA_polymerase_III

Abbreviations and symbols: aa (amino acids); n/a (not applicable); %ID (% identity at amino acid level); DUF (Domain of Unknown Function); OM (outer membrane); Asterisk (*) indicates *stlA* gene product. Text in bold indicates that no homologues for these gene products were found following BLAST searches of NCBI database

Transposon (EZ-Tn*5*) mutagenesis and cloning of the *stlA* gene

Transposon mutagenesis was performed on clone SMG 25 and a transposon insertion in gene 6 was identified which eliminated the growth advantage under osmotic stress; this locus (designated *stlA*) is predicted to encode a protein of 257 amino acids which, at the time of writing, has no homologues in the database. The transposon insertion was found to be between amino acid position 136 (alanine) and 137 (glutamine) of the protein. The *stlA* gene was cloned, along with some flanking DNA that was predicted to contain the native promoter region (predicted with BProm program; see Table 4 for details), into the shuttle plasmid pCI372 and transformed into *E. coli* MKH13 and *L. lactis* MG1363.

Growth experiments and BIOLOG phenotypic microarray

E. coli MKH13 cells transformed with a plasmid bearing a copy of the *stlA* gene were grown in LB broth containing various concentrations of NaCl (from 0-5% w/v added NaCl). It was observed that a statistically significant (unpaired student *t*-test) growth advantage was conferred upon the *stlA*$^+$ cells compared to control strain MKH13 carrying an empty plasmid in LB broth supplemented with both 3% (P =0.0019) and 4% NaCl (P <0.0001) (Figure 2B and C, respectively), while growth was similar in LB alone (Figure 2A).

Due to the uncharacterised and non-homologous nature of the *stlA* gene and its encoded protein, growth of control strain MKH13 and *stlA*$^+$ strains was compared under 96 different conditions using BIOLOGs phenotypic microarray (PM) technology [29] to identify possible further phenotypic changes. Strains were tested on BIOLOG plate PM9, which contains different osmotic stress conditions and osmolytes for analysis. In addition to NaCl, the results indicated *stlA*$^+$ had an increased growth phenotype in the presence of potassium chloride (KCl). Confirmatory growth curves were performed in LB broth supplemented with a concentration of 4% KCl. A statistically significant (unpaired student *t*-test) growth advantage was observed for *stlA*$^+$ in LB supplemented with 4% KCl (*P* <0.0001) compared to control strain MKH13 (Figure 2D).

Figure 2.

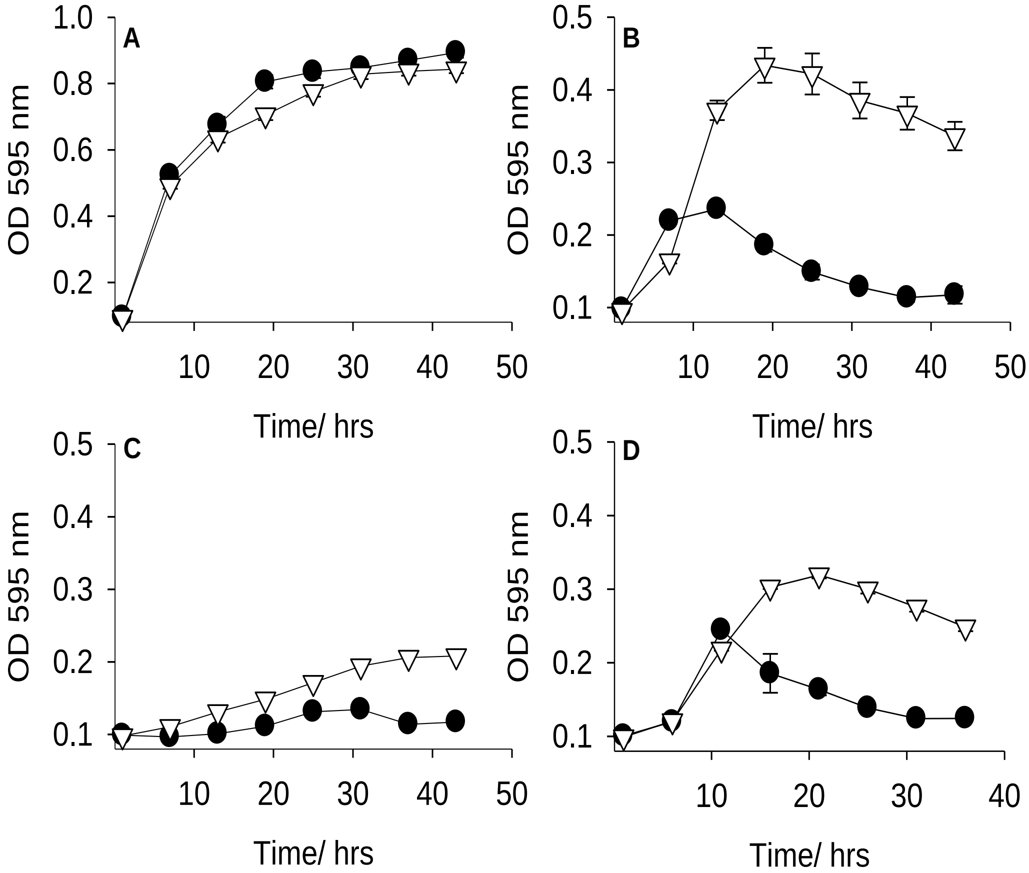


Figure 2. Growth in LB broth and LB broth supplemented with 3 or 4% NaCl. Growth of *E. coli* MKH13::pCI372 (●) and *E. coli* MKH13::pCI372-*stlA* (▽) in **(A)** LB broth, **(B)** LB broth + 3% NaCl (*P* =0.0470), **(C)** LB broth + 4% NaCl (*P* <0.0001) and **(D)** LB broth + 4% KCl (*P* <0.0001). *P*-values were determined using the student *t*-test (unpaired).

The *stlA* genes' ability to increase salt tolerance was also tested in a Gram positive host; *L. lactis* MG1363*.* There was no observable increase in salt tolerance in *L. lactis* MG1363 carrying a plasmid encoded copy of *stlA* compared to *L. lactis* carrying an empty copy of the plasmid, while a similar growth rate and final OD value was observed for both strains in GM17 broth alone (Figure 3A and 3B)

Figure 3.

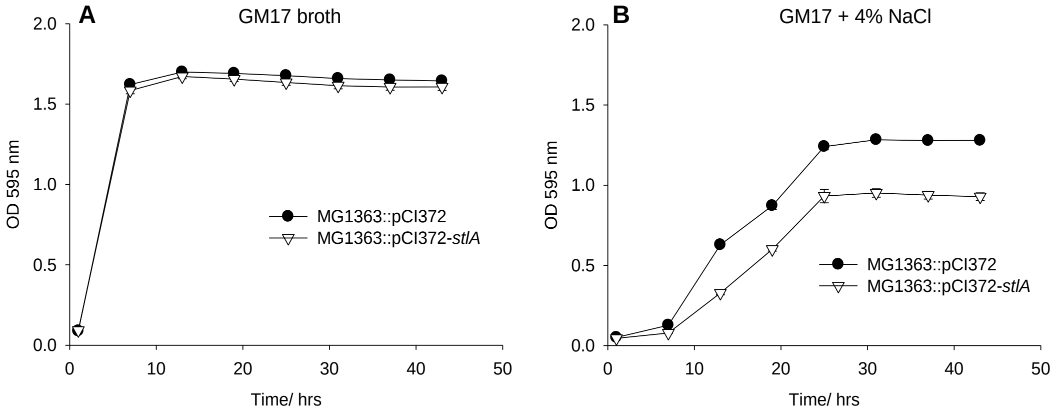
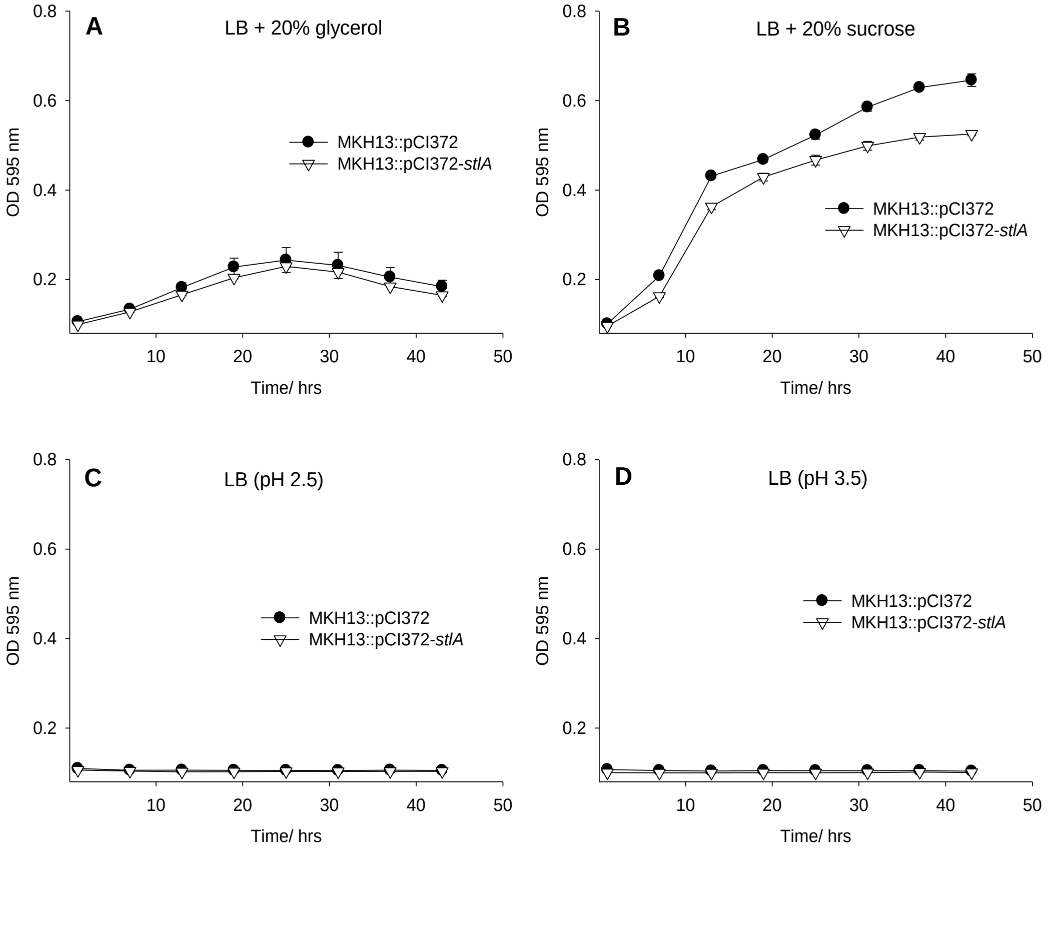


Figure 3. Growth of *L. lactis* MG1363::pCI372 and *L. lactis* MG1363::pCI372-*stlA* in GM17 broth and GM17 broth + 4% NaCl. The *stlA* gene did not provide a protective effect in a Gram-positive host under NaCl stress, which is noteworthy as the StlA protein is predicted to be inserted in the outer membrane. Results are presented as the average of triplicate experiments, with error bars being representative of the standard error of the mean (SEM).

Growth of both strains was also assessed under conditions of non-ionic osmotic stress (in the form of glycerol and sucrose), low pH and in both porcine and human bile, as all three stress conditions are commonly encountered in the GI tract. Growth of both strains was inhibited at pH 2.5 and pH 3.5, while no significant difference in growth was observed at pH 4.5 or pH 5.5, in the presence of sucrose or glycerol, or in the presence of either porcine or human bile (Figure 4).

Figure 4.



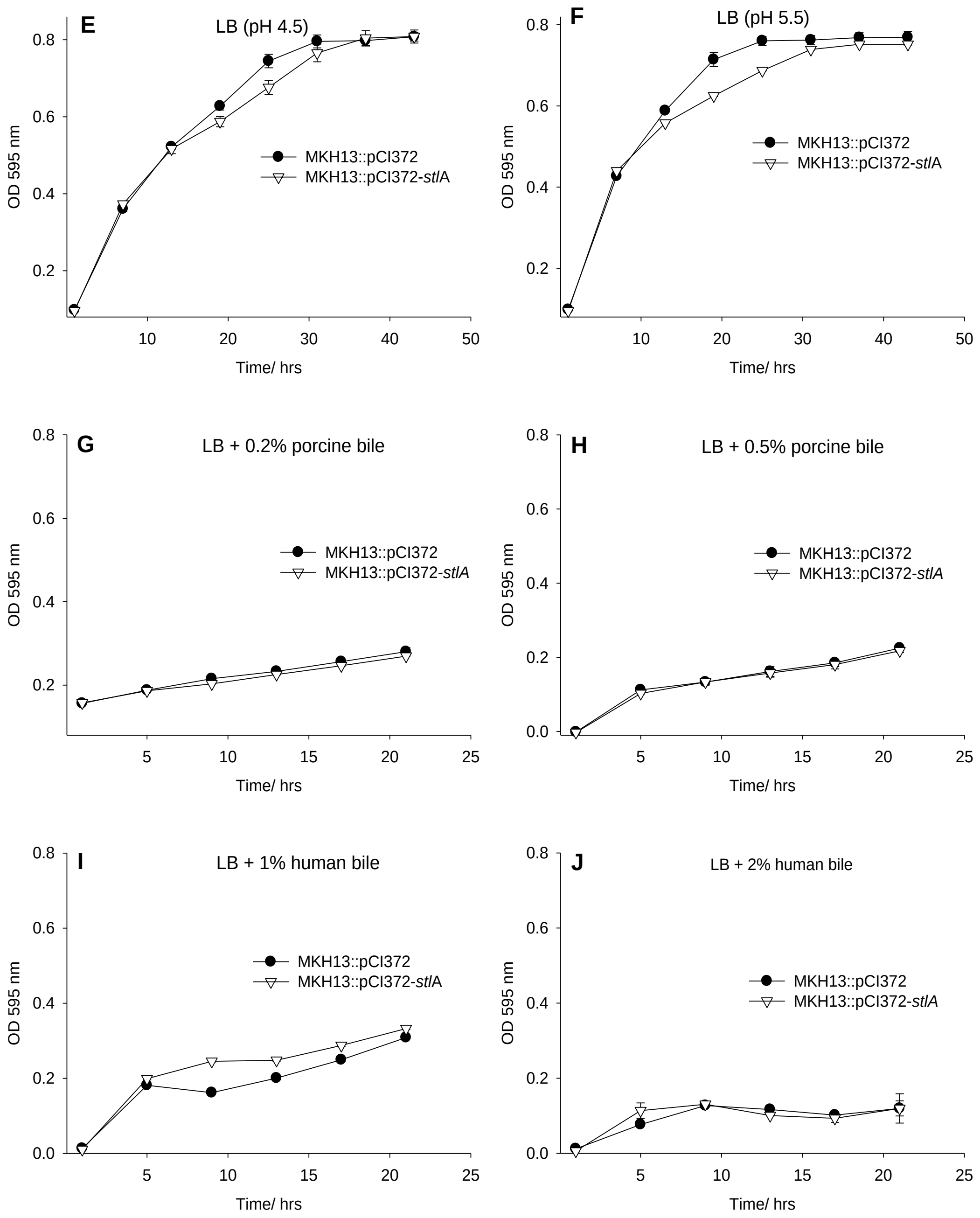


Figure 4. Growth of *E. coli* MKH13::pCI372 and *E. coli* MKH13::pCI372-*stlA* in LB broth supplemented with numerous stresses associated with the GI (gastrointestinal) tract, such as non-ionic osmotic stress, low pH and bile. A plasmid-encoded copy of the *stlA* gene did not confer increased tolerance to any of these stresses when expressed in *E. coli* MKH13. Results are presented as the average of triplicate experiments, with error bars being representative of the standard error of the mean (SEM).

Bioinformatic analysis of StlA

The databases and tools used to identify features of StlA are presented in Table 4 below, along with the results of the analyses. An illustration of the *stlA* gene and its associated features is presented in Figure 7D.

Table 4. Bioinformatic analysis of StlA protein sequence.

Database/ Tool used	Comment(s)	Feature(s) identified	Ref.
Expasy ProtParam	Allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered sequence	Molecular weight = 28.62 kDa; Theoretical pI = 6.39	[38]
Conserved domain database (CDD)	Protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins.	No conserved domains were detected	[34]
PROSITE motif search	Consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them	No motifs were detected	[36]
SignalP 3.0	Predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms	Predicted signal peptide at position 1-35	[31]
HMMER	Searches sequence databases for homologs of protein sequences, and for making protein sequence alignments	Predicted signal peptide at position 1-35; Four predicted TM regions and one disorder region	[32]
TMHMM	Prediction of transmembrane (TM) helices in proteins	Predicted four TM regions	[41]
HHPred	Homology detection & structure prediction by HMM-HMM (Hidden Markov Model) comparison	Detected outer membrane insertion C-terminal signal, OmpP85	[37]
BProm promoter search	Prediction of bacterial promoters	-10 box predicted 56 base pairs upstream of ATG start codon; TCTTATCAT; -35 box predicted 77 base pairs upstream of ATG start codon; TTGGCT	www.softberry.com
SOPMA	Secondary structure prediction	Alpha-helix; 166/257 residues = 64.6%; Extended strand; 26/257 residues = 10.1%; Beta-turn; 18/257 residues = 7.00%; Random coil; 47/257 residues = 18.3%	[33]
SWISS MODEL	Automated protein structure homology-modelling server	No similar or suitable template structures found	[30]
iTASSER	Protein structure and function predictions. 3D models are built based on multiple-threading alignments	All 5 predicted 3D models had a C-score of -3.41 or less which are below the -1.50 threshold for a high-confidence prediction of structure	[35,40]
QUARK	Algorithm for *ab initio* protein folding and protein structure prediction, using amino acid sequence only. Since no global template information is used in QUARK simulation, the server is suitable for proteins which are considered without homologous templates	Of the 10 predicted 3D models, the top template modelling (TM) score was 0.342 ±0.083, which is below the threshold of TM-score >0.50 for predicted correct fold	[39]

IMG/M-HMP analysis

The StlA protein sequence was screened against all available metagenomes from the human microbiome project (HMP) using BLASTP on the IMG/M-HMP website (http://img.jgi.doe.gov/cgi-bin/imgm_hmp/main.cgi) [45], which were sampled from 17 body sites giving a total of 748 samples. In addition, all available finished, permanent draft and draft genome sequences for *Bacteria*, *Archaea*, *Eukarya* and viruses/phages, as well as all available sequenced plasmids were searched using BLASTP for homologous sequences to StlA. There was no significant similarity for the StlA protein to any of the bacterial, archaeal, eukaryotic, viral or plasmid genomes/sequences, nor to any non-human associated metagenomes (over 1,300 samples from more than 200 metagenomes, see supplementary File S1*). The only similarity to StlA among the sampled microbiomes was to human stool microbiome samples, where 10 similar proteins from 8 different subjects (out of 100) (Table 5) were identified on different scaffolds. The date of the last search was on 19/09/13. The taxonomic assignment of the scaffolds can be seen in Figure 5.

* Supplementary File S1 can be accessed on the PLOS ONE wesite at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0082985;jsessionid=1EE719C721B576C1B180513AB3B44430

Table 5. Gene, scaffold and subject information from which *stlA* homologues were found in Human Microbiome Project (HMP) dataset.

Stool Microbiome Subject ID (Visit number)	Gene ID	Strand	Start Coordinate	End Coordinate	Length (bp)	Length (aa)	% ID to StlA (aa)	Gene Product Name (*stlA* homologue)	Scaffold Length (bp)	Scaffold GC %
N/A	***stlA* (from clone SMG 25; this study)**	**+**	**3193**	**3966**	**774**	**257**	**100% (257)**	**putative membrane protein**	**44331 (fosmid insert)**	**0.53**
*159753524 (2)	SRS053214_LANL_scaffold_17021__gene_42707	-	25802	26569	768	255	59% (237)	hypothetical protein	33560	0.5
*159753524 (3)	SRS077730_LANL_scaffold_24345__gene_72567	+	2793	3560	768	255	59% (237)	membrane protein	13529	0.49
‡764143897 (1)	SRS015217_WUGC_scaffold_30292__gene_65222	-	463	1236	774	257	82% (237)	membrane protein	5672	0.5
‡764143897 (2)	SRS051882_Baylor_scaffold_22757__gene_50812	-	1791	2564	774	257	82% (237)	membrane protein	7074	0.49
160643649 (1)	C2121591__gene_151559	+	1333	2157	825	274	89% (234)	membrane protein	5507	0.48
158944319 (1)	C3406971__gene_199744	-	248	1072	825	274	80% (234)	membrane protein	3122	0.52
159591683 (2)	SRS024549_LANL_scaffold_1815__gene_4559	-	4475	5248	774	257	82% (237)	membrane protein	10434	0.5
158337416 (2)	C2998990__gene_162710	+	340	972	633	211	81% (211)	hypothetical protein	974	0.45
765013792 (1)	SRS018656_WUGC_scaffold_544__gene_591	-	13762	14535	774	257	83% (237)	membrane protein	26364	0.51
159510762 (2)	SRS024075_LANL_scaffold_21370__gene_63545	-	21610	22320	711	236	82% (216)	hypothetical protein	35617	0.5

Table 5. Information for *stlA* homologues found in HMP dataset, including scaffold and subject of origin. Symbols (* and ‡) indicate detection of *stlA* homologue more than once from same subject. The StlA amino acid sequence was used to search against all 748 metagenome datasets from different body sites from the Human Microbiome Project (HMP) (1e-50 maximum e-value cut-off). Ten StlA homologues were identified in 8 different subjects. No StlA homologues were found in any other body site metagenome.

Figure 5. Taxonomic assignment of scaffold sequences from Human Microbiome Project on which an *stlA* homologue was found

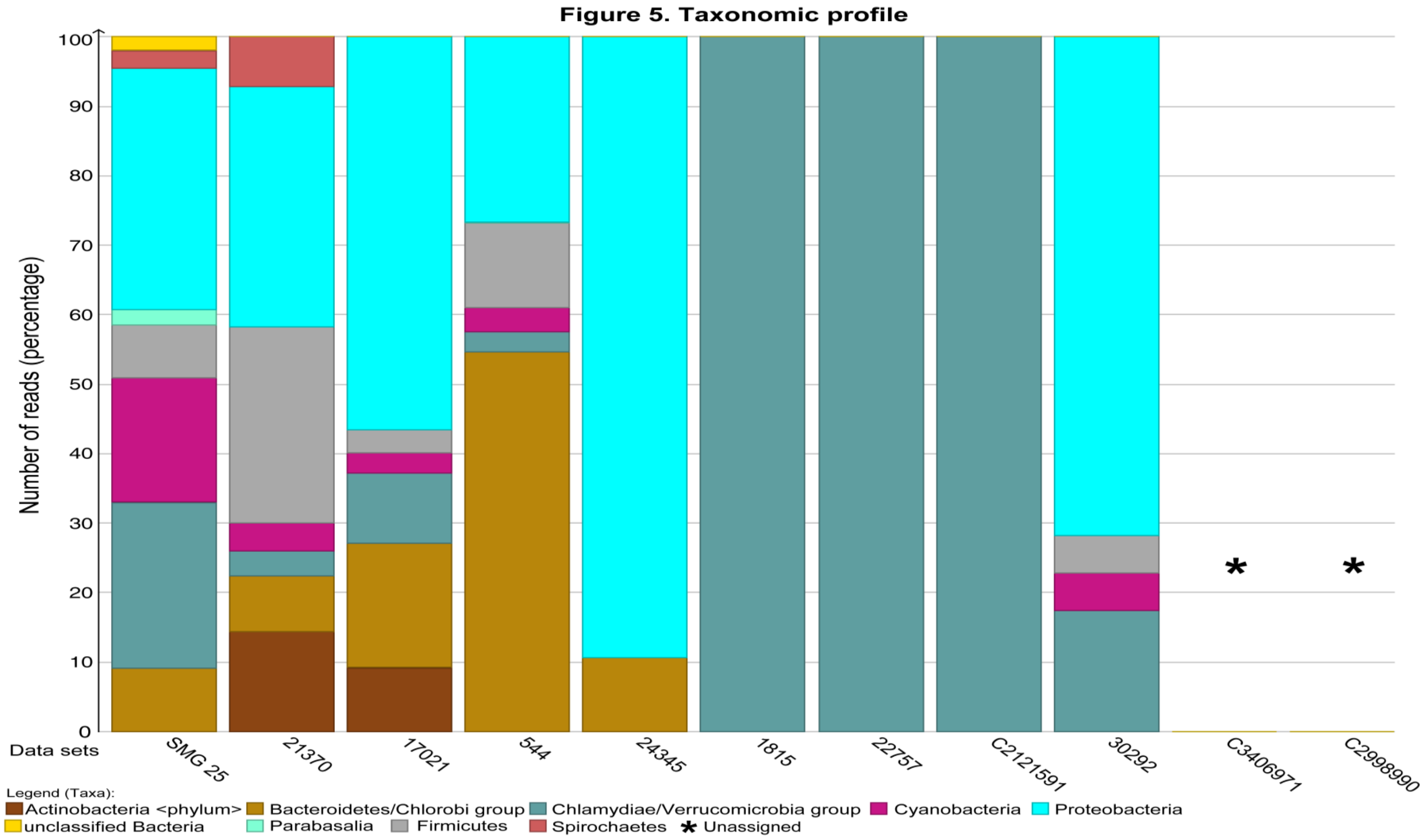
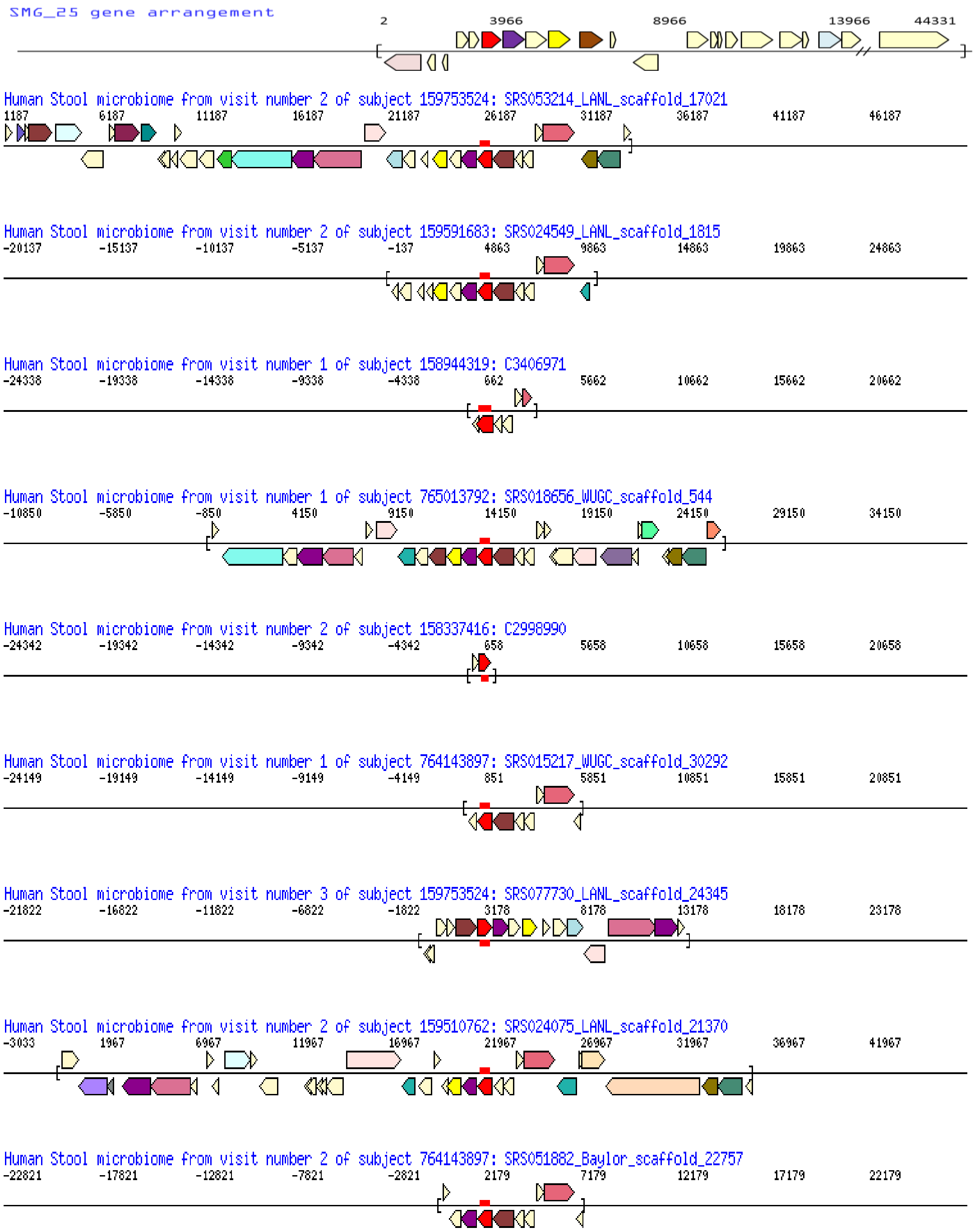


Figure 5. Taxonomic assignment of scaffold sequences from Human Microbiome Project on which an stlA homologue was found. Scaffold sequences were analysed using BLASTX. The BLASTX results were then downloaded and imported in MEGAN 4 software program which performed taxonomic assignment of each scaffold based on BLAST reads. Two of the shorter scaffolds indicated with an asterisk (*) could not be assigned any taxonomic classification.

The gene neighbourhoods around the genes homologous to *stlA* on each scaffold were investigated in an attempt to gain information on possible functions and conserved gene arrangements (Figure 6). The genes most commonly flanking the stlA homologues were on the same strand of DNA and encoded an ankyrin repeat protein (COG0666), a DnaJ class molecular chaperone with C-terminal zinc-finger domain (COG0484) and a predicted membrane protein (COG2314; Pfam05154 –TM2 domain). There are also a number of hypothetical proteins, for which no additional information is currently known. On two of the larger scaffolds, genes for a restriction modification system are present, as well as an integrase/site specific recombinase protein (COG4974; Pfam00589), indicating some of this region may have been acquired by lateral gene transfer (LGT) and may represent prophage DNA. A phage-associated protein is predicted to be encoded by gene 20 (designated “P” in Figure 7C) indicating the presence of a prophage on SMG 25 also. The fosmid insert of SMG 25 was analysed with PhiSpy [46] to identify possible prophage genes and the boundaries of the prophage region. PhiSpy predicted the prophage region to run from the start of gene 3 (nucleotide position 2024) to the end of gene 42 (nucleotide position 39972).

Figure 6.



SMG_25 gene arrangement
2
3966
8966
13966
44331
Human Stool microbiome from visit number 2 of subject 159753524: SRS053214_LANL_scaffold_17021
1187 6187 11187 16187 21187 26187 31187 36187 41187 46187
Human Stool microbiome from visit number 2 of subject 159591683: SRS024549_LANL_scaffold_1815
-20137 -15137 -10137 -5137 -137 4863 9863 14863 19863 24863
Human Stool microbiome from visit number 1 of subject 158944319: C3406971
-24338 -19338 -14338 -9338 -4338 662 5662 10662 15662 20662
Human Stool microbiome from visit number 1 of subject 765013792: SRS018656_WUGC_scaffold_544
-10850 -5850 -850 4150 9150 14150 19150 24150 29150 34150
Human Stool microbiome from visit number 2 of subject 158337416: C2998990
-24342 -19342 -14342 -9342 -4342 658 5658 10658 15658 20658
Human Stool microbiome from visit number 1 of subject 764143897: SRS015217_WUGC_scaffold_30292
-24149 -19149 -14149 -9149 -4149 851 5851 10851 15851 20851
Human Stool microbiome from visit number 3 of subject 159753524: SRS077730_LANL_scaffold_24345
-21822 -16822 -11822 -6822 -1822 3178 8178 13178 18178 23178
Human Stool microbiome from visit number 2 of subject 159510762: SRS024075_LANL_scaffold_21370
-3033 1967 6967 11967 16967 21967 26967 31967 36967 41967
Human Stool microbiome from visit number 2 of subject 764143897: SRS051882_Baylor_scaffold_22757
-22821 -17821 -12821 -7821 -2821 2179 7179 12179 17179 22179

Figure 6. Gene neighbourhood region of the *stlA* gene from SMG 25 compared with gene neighbourhoods from scaffolds with a *stlA* homologue. Homologues of *stlA* were identified through similarity searches (BLASTP; 1e-^{50} cut-off) to the Human Microbiome Project (HMP) datasets. Ten *stlA* homologues were identified and only from the stool microbiome. A legend describing putative gene functions is presented below.

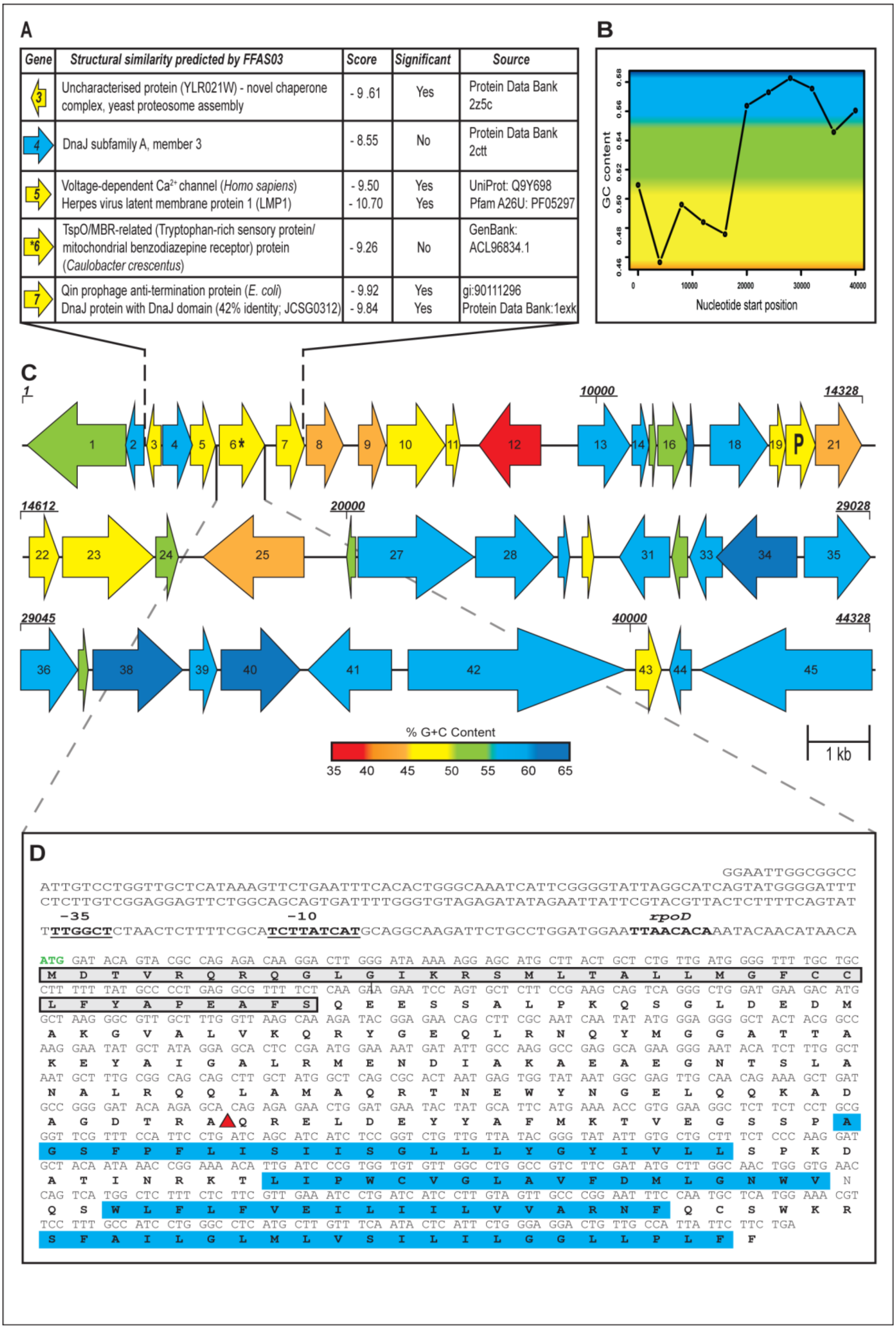
Legend: ▶ hypothetical/membrane protein (*stlA* and homologues); ▶ Hypothetical protein; ▶ COG0838 - NADH:ubiquinone oxidoreductase; ▶ COG0738 - Fucose permease; ▶ COG4974 - Site-specific recombinase, XerD; ▶ COG3550 - Uncharacterized protein related to capsule biosynthesis enzymes; ▶ COG3183 - Predicted restriction endonuclease; ▶ COG1451 - Predicted metal-dependent hydrolase; ▶ COG0610 - Type I site-specific restriction-modification system; ▶ COG0732 - Restriction endonuclease; ▶ COG0286 - Type I restriction-modification system methyltransferase subunit; ▶ COG1715 - Restriction endonuclease; ▶ COG3857 - ATP-dependent nuclease; ▶ COG2314 - Predicted membrane protein (TM2 domain); ▶ COG0484 - DnaJ-class molecular chaperone with C-terminal Zn finger domain; ▶ COG0666 - Ankyrin repeat protein; ▶ COG0515 - Serine/threonine protein kinase; ▶ COG4245 - Uncharacterized protein with von Willebrand factor (vWF) domain; ▶ COG4248 - Uncharacterized protein with protein kinase and helix-hairpin-helix DNA-binding domains; ▶ COG3943 - Virulence protein; ▶ COG2887 - RecB family exonuclease; ▶ COG0667 - Predicted oxidoreductase; ▶ COG1073 - Hydrolases of the alpha/beta superfamily; ▶ COG2207 - Transcriptional regulator, AraC-type DNA-binding domain-containing proteins; ▶ COG1373 - Predicted ATPase (AAA+ superfamily); ▶ COG1961 - Site-specific recombinase, DNA invertase Pin homologs; ▶ COG3177 - Filamentation induced by cAMP protein; ▶ COG4889 - Predicted helicase

FFAS03 analysis

The FFAS03 server [42] was used to detect distant homology and fold recognition to StlA. FFAS analysis was also carried out on the protein sequences encoded by the neighbouring genes of *stlA* on SMG 25, which also lacked any homologues in the databases (i.e. gene 3, 4, 5 and 7; gene 6 is *stlA*). The results of FFAS03 analysis are summarised in Figure 7A).

In addition to a profile-profile and a fold and functional assignment, the FFAS03 server also carries out a BLAST and PSI-BLAST search of the user sequence against numerous databases and metagenome datasets. The StlA protein was found to share significant similarity to a protein from two individuals from the MetaHit dataset [1]. These sequences corresponded to samples MH0011 (a healthy Danish female) and V1.CD-14 (a Spanish female with Crohn’s disease) which shared 60% identity (over 210 amino acids) and 82% identity (over 224 amino acids) respectively to StlA.

Figure 7.



A

Gene	Structural similarity predicted by FFAS03	Score	Significant	Source
3	Uncharacterised protein (YLR021W) - novel chaperone complex, yeast proteosome assembly	- 9 .61	Yes	Protein Data Bank 2z5c
4	DnaJ subfamily A, member 3	- 8.55	No	Protein Data Bank 2ctt
5	Voltage-dependent Ca^{2+} channel (*Homo sapiens*) Herpes virus latent membrane protein 1 (LMP1)	- 9.50 - 10.70	Yes Yes	UniProt: Q9Y698 Pfam A26U: PF05297
*6	TspO/MBR-related (Tryptophan-rich sensory protein/ mitochondrial benzodiazepine receptor) protein (*Caulobacter crescentus*)	- 9.26	No	GenBank: ACL96834.1
7	Qin prophage anti-termination protein (*E. coli*) DnaJ protein with DnaJ domain (42% identity; JCSG0312)	- 9.92 - 9.84	Yes Yes	gi:90111296 Protein Data Bank:1exk

D

```
                                                                          GGAATTGGCGGCC
ATTGTCCTGGTTGCTCATAAAGTTCTGAATTTCACACTGGGCAAATCATTCGGGGTATTAGGCATCAGTATGGGGATTT
CTCTTGTCGGAGGAGTTCTGGCAGCAGTGATTTTGGGTGTAGAGATATAGAATTATTCGTACGTTACTCTTTTCAGTAT
 -35               -10                                         rpoD
TTTGGCTCTAACTCTTTTCGCATCTTATCATGCAGGCAAGATTCTGCCTGGATGGAATTAACACAAATACAACATAACA

ATG GAT ACA GTA CGC CAG AGA CAA GGA CTT GGG ATA AAA AGG AGC ATG CTT ACT GCT CTG TTG ATG GGG TTT TGC TGC
 M   D   T   V   R   Q   R   Q   G   L   G   I   K   R   S   M   L   T   A   L   L   M   G   F   C   C
CTT TTT TAT GCC CCT GAG GCG TTT TCT CAA GAA GAA TCC AGT GCT CTT CCG AAG CAG TCA GGG CTG GAT GAA GAC ATG
 L   F   Y   A   P   E   A   F   S   Q   E   E   S   S   A   L   P   K   Q   S   G   L   D   E   D   M
GCT AAG GGC GTT GCT TTG GTT AAG CAA AGA TAC GGA GAA CAG CTT CGC AAT CAA TAT ATG GGA GGG GCT ACT ACG GCC
 A   K   G   V   A   L   V   K   Q   R   Y   G   E   Q   L   R   N   Q   Y   M   G   G   A   T   T   A
AAG GAA TAT GCT ATA GGA GCA CTC CGA ATG GAA AAT GAT ATT GCC AAG GCC GAG GCA GAA GGG AAT ACA TCT TTG GCT
 K   E   Y   A   I   G   A   L   R   M   E   N   D   I   A   K   A   E   A   E   G   N   T   S   L   A
AAT GCT TTG CGG CAG CAG CTT GCT ATG GCT CAG CGC ACT AAT GAG TGG TAT AAT GGC GAG TTG CAA CAG AAA GCT GAT
 N   A   L   R   Q   Q   L   A   M   A   Q   R   T   N   E   W   Y   N   G   E   L   Q   Q   K   A   D
GCC GGG GAT ACA AGA GCA CAG AGA GAA CTG GAT GAA TAC TAT GCA TTC ATG AAA ACC GTG GAA GGC TCT TCT CCT GCG
 A   G   D   T   R   A   Q   R   E   L   D   E   Y   Y   A   F   M   K   T   V   E   G   S   S   P   A
GGT TCG TTT CCA TTC CTG ATC AGC ATC ATC TCC GGT CTG TTG TTA TAC GGG TAT ATT GTG CTG CTT TCT CCC AAG GAT
 G   S   F   P   F   L   I   S   I   I   S   G   L   L   L   Y   G   Y   I   V   L   L   S   P   K   D
GCT ACA ATA AAC CGG AAA ACA TTG ATC CCG TGG TGT GTT GGC CTG GCC GTC TTC GAT ATG CTT GGC AAC TGG GTG AAC
 A   T   I   N   R   K   T   L   I   P   W   C   V   G   L   A   V   F   D   M   L   G   N   W   V   N
CAG TCA TGG CTC TTT CTC TTC GTT GAA ATC CTG ATC ATC CTT GTA GTT GCC CGG AAT TTC CAA TGC TCA TGG AAA CGT
 Q   S   W   L   F   L   F   V   E   I   L   I   I   L   V   V   A   R   N   F   Q   C   S   W   K   R
TCC TTT GCC ATC CTG GGC CTC ATG CTT GTT TCA ATA CTC ATT CTG GGA GGA CTG TTG CCA TTA TTC TTC TGA
 S   F   A   I   L   G   L   M   L   V   S   I   L   I   L   G   G   L   L   P   L   F   F
```

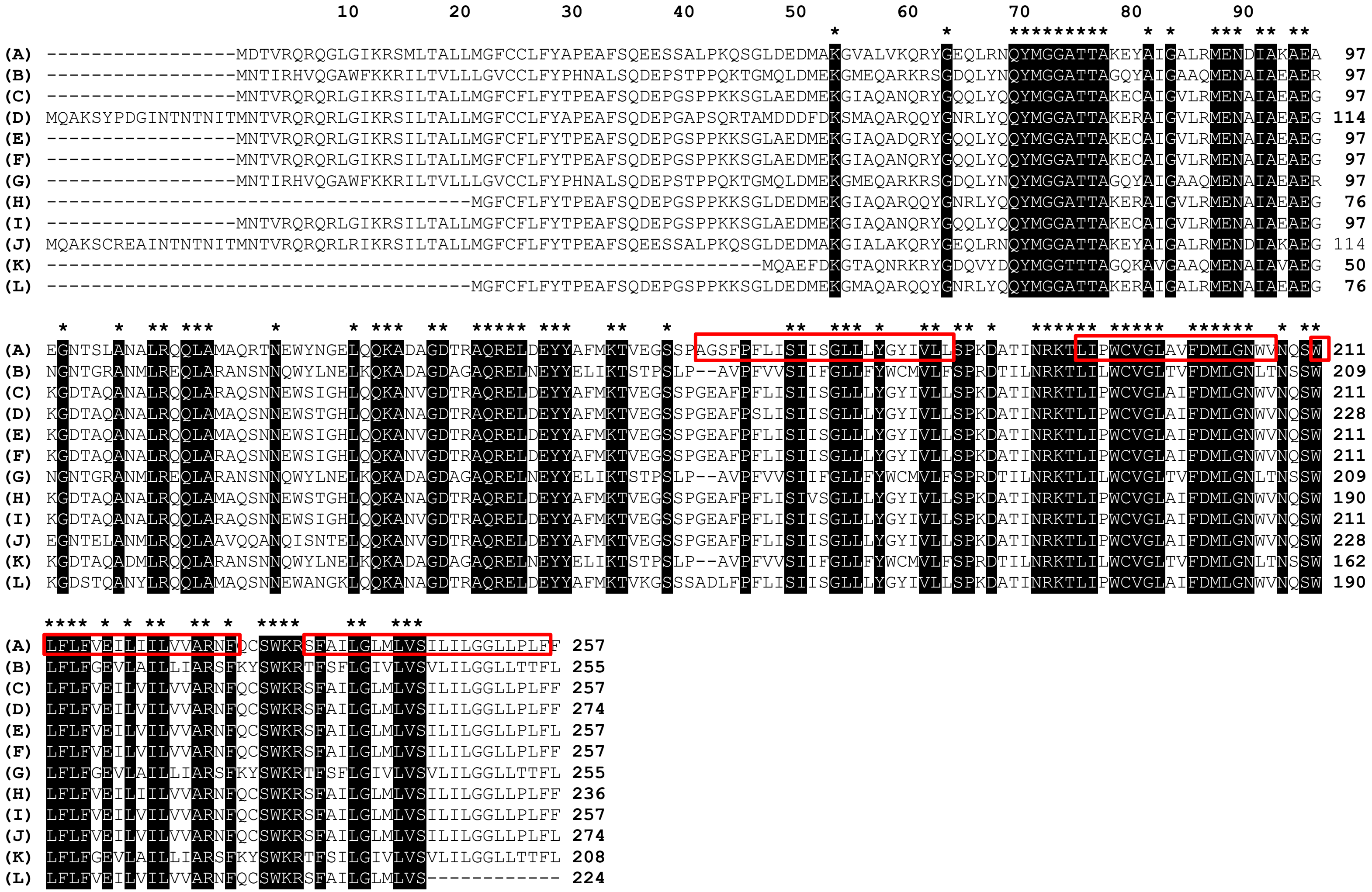
Figure 7. Bioinformatic analysis of SMG 25 fosmid insert.

(A) FFAS03 analysis of the StlA protein and the encoded proteins of flanking genes was performed to identify putative distant structural homologues. A score of -9.50 or lower is considered significant. **(B)** Representation of the G+C skew of the entire fosmid insert of SMG 25. **(C)** Representation of the gene arrangement on SMG 25. Gene lengths are approximately to scale and colour coding represents G+C content of each individual gene which can be determined from the G+C content gradient bar. The presence of a phage-associated gene and clear separation in G+C content over the length of the fosmid insert indicates much of this region may have been acquired via lateral gene transfer (LGT). Phage-associated gene is marked "P", while the *stlA* gene is indicated with an asterisk (*) symbol. Genes are numbered as indicated in Table 3 and as mentioned in the text. Numbering of some shorter genes has been excluded for clarity. Selected nucleotide positions (in base pairs) are displayed in bold italic font above genes. **(D)** A detailed view is presented of the nucleotide and amino acid sequence of the *stlA* gene and StlA protein respectively. The putative start codon is in green, while a 250 base-pair region upstream of this is shown to include putative -35 and -10 promoter regions (underlined) and a predicted *rpoD* transcription factor binding site (in bold). Amino acids surrounded by grey box indicate the predicted signal sequence of StlA and those highlighted in blue represent four transmembrane regions. The location of the EZTn*5* transposon insertion is indicated with a red triangle.

Detection of *stlA* in metagenomic DNA from human stool samples using PCR

The primer pair (*stlA* FP and *stlA* RP) initially used to amplify the *stlA* gene for cloning was unable to amplify PCR products in any of the metagenomic DNA samples (isolated from human stool microbiota), so a set of primers (*stlA*-J FP and *stlA*-J RP) were designed to amplify an internal fragment of the gene. This set of primers amplified numerous products of the correct size but these were found to be false positives following sequencing. An alignment was generated for StlA and homologous sequences from the stool microbiome from the HMP and MetaHit datasets to identify the most highly conserved regions (Figure 8) and different primer pairs were designed (*stlA*-OUT FP and RP; *stlA*-IN FP and RP). Two of the 25 metagenomic DNA samples tested (isolated from human faecal samples from ELDERMET [27] and another study {Knopp, 2010 #464}) generated PCR products of the correct size, which were confirmed to be *stlA* homologues following sequencing by using the *stlA*-IN FP and RP primer pair. One positive PCR product shared 72% nucleotide identity over approximately 300 base pairs (BLASTN versus stlA gene) and 64% identity (over 100 amino acids) using BLASTX, while one ELDERMET [27] sample was positive (community care/ healthy old) and confirmed by sequencing (87% identity over 339 nucleotides and 85% identity over 112 amino acids).

Figure 8. Multiple sequence alignment of StlA protein sequence with HMP and MetaHit homologues.



```
                              10        20        30        40        50        60        70        80        90
                                                                          *         *     *********   * *   *** ** **
(A) -----------------MDTVRQRQGLGIKRSMLTALLMGFCCLFYAPEAFSQEESSALPKQSGLDEDMAKGVALVKQRYGEQLRNQYMGGATTAKEYAIGALRMENDIAKAEA  97
(B) -----------------MNTIRHVQGAWFKKRILTVLLLGVCCLFYPHNALSQDEPSTPPQKTGMQLDMEKGMEQARKRSGDQLYNQYMGGATTAGQYAIGAAQMENAIAEAER  97
(C) -----------------MNTVRQRQRLGIKRSILTALLMGFCFLFYTPEAFSQDEPGSPPKKSGLAEDMEKGIAQANQRYGQQLYQQYMGGATTAKECAIGVLRMENAIAEAEG  97
(D) MQAKSYPDGINTNTNITMNTVRQRQRLGIKRSILTALLMGFCCLFYAPEAFSQDEPGAPSQRTAMDDDFDKSMAQARQQYGNRLYQQYMGGATTAKERAIGVLRMENAIAEAEG 114
(E) -----------------MNTVRQRQRLGIKRSILTALLMGFCFLFYTPEAFSQDEPGSPPKKSGLAEDMEKGIAQADQRYGQQLYQQYMGGATTAKECAIGVLRMENAIAEAEG  97
(F) -----------------MNTVRQRQRLGIKRSILTALLMGFCFLFYTPEAFSQDEPGSPPKKSGLAEDMEKGIAQANQRYGQQLYQQYMGGATTAKECAIGVLRMENAIAEAEG  97
(G) -----------------MNTIRHVQGAWFKKRILTVLLLGVCCLFYPHNALSQDEPSTPPQKTGMQLDMEKGMEQARKRSGDQLYNQYMGGATTAGQYAIGAAQMENAIAEAER  97
(H) --------------------------------------MGFCFLFYTPEAFSQDEPGSPPKKSGLDEDMEKGIAQARQQYGNRLYQQYMGGATTAKERAIGVLRMENAIAEAEG  76
(I) -----------------MNTVRQRQRLGIKRSILTALLMGFCFLFYTPEAFSQDEPGSPPKKSGLAEDMEKGIAQANQRYGQQLYQQYMGGATTAKECAIGVLRMENAIAEAEG  97
(J) MQAKSCREAINTNTNITMNTVRQRQRLRIKRSILTALLMGFCFLFYTPEAFSQEESSALPKQSGLDEDMAKGIALAKQRYGEQLRNQYMGGATTAKEYAIGALRMENDIAKAEG 114
(K) ----------------------------------------------------------------MQAEFDKGTAQNRKRYGDQVYDQYMGGTTTAGQKAVGAAQMENAIAVAEG  50
(L) --------------------------------------MGFCFLFYTPEAFSQDEPGSPPKKSGLDEDMEKGMAQARQQYGNRLYQQYMGGATTAKERAIGALRMENAIAEAEG  76

     *    *  ** ***     *      * ***  **  ***** ***   **   *          **  *** *   ** ** *   ****** *****  ******  * **
(A) EGNTSLANALRQQLAMAQRTNEWYNGELQQKADAGDTRAQRELDEYYAFMKTVEGSSPAGSFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAVFDMLGNWVNQSW 211
(B) NGNTGRANMLREQLARANSNNQWYLNELKQKADAGDAGAQRELNEYYELIKTSTPSLP--AVPFVVSIIFGLLFYWCMVLFSPRDTILNRKTLILWCVGLTVFDMLGNLTNSSW 209
(C) KGDTAQANALRQQLARAQSNNEWSIGHLQQKANVGDTRAQRELDEYYAFMKTVEGSSPGEAFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 211
(D) KGDTAQANALRQQLAMAQSNNEWSTGHLQQKANAGDTRAQRELDEYYAFMKTVEGSSPGEAFPSLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 228
(E) KGDTAQANALRQQLAMAQSNNEWSIGHLQQKANVGDTRAQRELDEYYAFMKTVEGSSPGEAFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 211
(F) KGDTAQANALRQQLARAQSNNEWSIGHLQQKANVGDTRAQRELDEYYAFMKTVEGSSPGEAFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 211
(G) NGNTGRANMLREQLARANSNNQWYLNELKQKADAGDAGAQRELNEYYELIKTSTPSLP--AVPFVVSIIFGLLFYWCMVLFSPRDTILNRKTLILWCVGLTVFDMLGNLTNSSW 209
(H) KGDTAQANALRQQLAMAQSNNEWSTGHLQQKANAGDTRAQRELDEYYAFMKTVEGSSPGEAFPFLISIVSGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 190
(I) KGDTAQANALRQQLARAQSNNEWSIGHLQQKANVGDTRAQRELDEYYAFMKTVEGSSPGEAFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 211
(J) EGNTELANMLRQQLAAVQQANQISNTELQQKANVGDTRAQRELDEYYAFMKTVEGSSPGEAFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 228
(K) KGDTAQADMLRQQLARANSNNQWYLNELKQKADAGDAGAQRELNEYYELIKTSTPSLP--AVPFVVSIIFGLLFYWCMVLFSPRDTILNRKTLILWCVGLTVFDMLGNLTNSSW 162
(L) KGDSTQANYLRQQLAMAQSNNEWANGKLQQKANAGDTRAQRELDEYYAFMKTVKGSSSADLFPFLISIISGLLLYGYIVLLSPKDATINRKTLIPWCVGLAIFDMLGNWVNQSW 190

    **** * * **  ** *  ****    **  ***
(A) LFLFVEILIILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFF 257
(B) LFLFGEVLAILLIARSFKYSWKRTFSFLGIVLVSVLILGGLLTTFL 255
(C) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFF 257
(D) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFF 274
(E) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFL 257
(F) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFF 257
(G) LFLFGEVLAILLIARSFKYSWKRTFSFLGIVLVSVLILGGLLTTFL 255
(H) LFLFVEILIILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFF 236
(I) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFF 257
(J) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVSILILGGLLPLFL 274
(K) LFLFGEVLAILLIARSFKYSWKRTFSILGIVLVSVLILGGLLTTFL 208
(L) LFLFVEILVILVVARNFQCSWKRSFAILGLMLVS------------ 224
```

Figure 8. Multiple sequence alignment of StlA protein sequence with HMP and MetaHit homologues. Black shading indicates regions of 100% amino acid identity. Putative transmembrane regions for StlA, predicted by TMHMM, are indicated with red boxes. Truncated or partial sequence fragments from HMP were not included (n=4). Information on the protein sequences (A) – (J) is indicated in the legend below.

(A) StlA protein sequence; **(B)** SRS053214_LANL_scaffold_17021__gene_42707;

(C) SRS024549_LANL_scaffold_1815__gene_4559; **(D)** C3406971__gene_199744;

(E) SRS018656_WUGC_scaffold_544__gene_591; **(F)** SRS015217_WUGC_scaffold_30292__gene_65222;

(G) SRS077730_LANL_scaffold_24345__gene_72567; **(H)** SRS024075_LANL_scaffold_21370__gene_63545;

(I) Baylor_scaffold_22757__gene_50812; **(J)** C2121591__gene_151559;

(K) MetaHit_MH0011_GL0108025[Complete]locus=scaffold6530_52:7938:8564

(L) MetaHit_V1_GL0100177 [Complete] locus=scaffold36986_1:2178:2888.

Discussion

Functional screening of metagenomic libraries has the power to reveal novel functions for known genes or to identify completely novel genes and proteins. In the present study we describe the identification of an unknown protein (annotated StlA) from the human gut microbiome, which lacks any current homologues in the databases. The encoding gene (*stlA*), when expressed in *E. coli*, conferred a salt tolerance phenotype and may represent a novel stress resistance gene found exclusively among the human gut microbiota. This builds on previous work by our group, where we identified a novel function (i.e. increased salt tolerance) for five previously annotated genes (*galE*, *mazG* and *murB*) when expressed in *E. coli* [21].

Sequencing of the full fosmid insert from SMG 25 revealed an interesting gene landscape (Table 3), with approximately 58% of the predicted genes encoding proteins which shared highest genetic identity to different species of *Akkermansia* and 27% having no homologues in the databases. The *Akkermansia*-associated proteins and the “unknown” proteins are interspersed with proteins associated with different phyla such as *Bacteroidetes/Chlorobi* group, *Synergistetes*, *Proteobacteria*, *Chlamydiae/ Verrucomicrobia* group and *Firmicutes*, as well as *Archaea*. The percentage identity at the amino acid level ranges from 36-69%, revealing a diverse range of proteins encoded within approximately 44kb of fosmid insert DNA (Table 3).

The G+C content of the entire fosmid insert is 52.97%, which is close to the average G+C content (55.8%) of the *A. muciniphila* genome [49]. The region from position 2024 (gene 3) to position 20148 (gene 26), which mainly consists of unknown genes or non-*Akkermansia*-associated genes has a lower G+C content of 47.97%. The region of the fosmid containing mainly *Akkermansia*-associated genes (from gene 27 at position 20120 to the end of the fosmid) has a G+C content of 56.73%, in line with the *A. muciniphila* genome (55.8%). A putative prophage region was predicted (using PhiSpy) to be present on SMG 25, running from gene 3 to 42 inclusive. It is difficult to say how reliable this prediction is because the criteria used by PhiSpy to predict prophage genes are strongly assisted by the degree of relatedness of the PhiSPy training genome sets and the genome/ DNA of the query organism [46]. Unfortunately PhiSPy does not contain an *Akkermansia* or Verrucomicrobial training genome, which would increase the predictive value of the result. However, by looking at the G+C skew of SMG 25 and the G+C content of each individual gene on SMG 25 (Figure 7B and C, respectively), it seems the prophage could indeed begin at gene 3, but it is possible that it ends somewhere between gene 23 and 26, as there is a clear difference in G+C content visible between this region and from gene 27 to 45 at the 3'-end of the fosmid (Figure 7B and C). Taken together, this data suggests that much of this region may have been acquired through LGT.

StlA is predicted to be a 257 amino acid, 28.62kDa membrane protein with four transmembrane regions. No conserved domains or motifs were detected, indicating the novelty of the protein. A signal peptide and a C-terminal outer membrane insertion signal are predicted to be present, suggesting that StlA may be exported to and inserted in the outer membrane. Furthermore, StlA possesses C-terminal phenylalanine residues, which are characteristic of, and highly conserved in, outer membrane proteins [51]. A detailed illustration of these features is presented in Figure 7D, along with putative promoter and transcription factor binding sites. The outer membrane itself is an important mediator to external stresses, serving as a permeability barrier and protecting the cell from compounds in the environment, while outer membrane proteins, specifically porins, play an significant role in the cellular responses to salt and osmotic stress [15,52,53]. It is noteworthy, given the likely location in the outer membrane, that StlA did not confer a salt tolerance phenotype on a Gram-positive host (*L. lactis*) (Figure 3).

Predictive 3D modelling was carried out with SWISS MODEL [30] and iTasser [35,40]. However, the results were not statistically significant, most likely due to the lack of any suitable template structure in the databases to build an appropriate model. *Ab initio* structure prediction was attempted using QUARK [39] as no template information is required and is thus suitable for proteins with no homologues. Again the results were not significant, but this is most likely due to the inherent difficulty and current limited ability of *ab initio* prediction. Successful cases of *ab initio* prediction have been limited to proteins of 100 residues or less and the fact remains that there are really no methods to predictively fold proteins of >200 amino acids without template modelling at present [39].

As no sequence-based homology for StlA could be determined with BLAST analysis, a more sensitive profile-profile comparison with FFAS03 [42] was used to detect remote homology through fold and structure recognition, as proteins with a similar structure or fold can have a common function in the absence of any sequence similarity. The highest score for StlA corresponded to a hypothetical protein from *Caulobacter crescentus*, which has a TspO/MBR domain. Members of this group are involved in transmembrane signalling and are located in the outer membrane [54,55]. They are associated with the major outer membrane porins (in prokaryotes) and with the voltage-dependent anion channel (in mitochondria), which links with the earlier observation that StlA may be inserted in the outer membrane. Such proteins have also been linked to desiccation stress in the bacterium *Bradyrhizobium japonicum* [56].

FFAS analysis of the encoded proteins in the gene neighbourhood of *stlA* on SMG 25 revealed some structural similarities to DnaJ and another type of molecular chaperone for the encoded proteins of genes 3, 4 and 7, while gene 5 encodes a protein with some structural similarity to a human voltage-gated calcium channel to which TspO has been linked, and it also shares a structural homology to Herpes virus latent membrane protein 1 (LMP 1). In addition to DnaJ, the predicted product of gene 7 also exhibited structural similarity to an anti-termination protein from the Qin prophage. This could indicate some of this region was acquired via integration of a phage into the host chromosome. The novelty of the sequences may point to an uncharacterized phage. It is also noteworthy that gene 20 on SMG 25 is predicted to encode a phage-associated protein, while on two of the larger scaffolds from the stool microbiome samples, a gene encoding a phage integrase protein is present, revealing a commonality of such genes in this region. An elegant study by Wang and co-workers, has demonstrated prophage DNA plays a significant role in host resistance to numerous stresses, including osmotic stress, through deletion of each of nine prophages in *E. coli* [57]. The phage-associated protein on SMG 25 shares

52% identity with a similar protein from *Rhizobium lupini* HPC(L) (100% coverage over 159 amino acids), Interestingly, this organism was recently sequenced following isolation from a saline desert soil [58]. *Rhizobium* species belong to the phylum *Proteobacteria* and, based on taxonomic assignment with MEGAN 4, proteobacterial sequences were found on all the larger scaffolds with an *stlA* homologue (Figure 5) and may indicate the origin of the phage. Furthermore, a number of genes on SMG 25 are predicted to encode proteins that share a high level of similarity to halophilic and halotolerant microorganisms (Table 3). For example, gene 8 and 9 are predicted to encode hypothetical proteins with similarity to *Pontibacter sp.* BAB1700 and a halophilic archaeon, respectively. *Pontibacter* species are halotolerant members of the phylum Bacteroidetes and have been isolated from saline and marine environments, while gene 26 is predicted to encode a protein with similarity to *Halomonas*, a genus of halophilic *Proteobacteria* with biotechnological and medical relevance [59-61]. It seems possible the phage originated in a “salty” environment such as saline soil, a salt lake, a solar saltern or marine ecosystem.

When compared against all the available samples from the HMP, homologues of the *stlA* sequence were found to be present only in stool microbiome samples. Furthermore, no homologous sequences were found in any bacterial, archaeal, eukaryotic or viral genome sequences, or in any sequenced plasmids. This indicates that *stlA* gene is extremely rare in the sequences tested and may be a gut-specific gene and present only in species of low abundance, as no homologues were found in any of the common or dominant members of the human gut microbiome. In addition, we could only detect *stlA* homologues by PCR in two of 25 metagenomic DNA samples isolated from stool. Gene neighbourhood analysis around the *stlA* homologues revealed they were most often found in combination with genes encoding DnaJ-type molecular chaperones (COG0484), an ankyrin repeat protein (COG0666) or a predicted membrane protein containing a TM-2 domain (COG02314), which is also similar to the gene organisation on SMG 25.

DnaJ-domain proteins are molecular chaperones that aid protein folding, prevent aggregation and repair damaged proteins following cellular stress [62]. They are members of the heat shock protein (Hsp) family, which have been shown to play important roles in the response to numerous stress conditions including osmotic stress and also can act as co-chaperones by stimulating the activity of other chaperones such as DnaK [63-65]. TM2 domain proteins are composed of a pair of alpha helices connected by a short linker. The function of this domain is unknown; however it occurs in a wide range of protein contexts. It occurs most often on its own or in tandem with another TM2 domain, but interestingly, the third most frequent association is with a DnaJ domain.

Ankyrin-repeat proteins are found across all three domains of life and modulate a number of diverse functions through protein-protein interactions [66]. The repeat has been found in proteins of diverse function [67,68] and these proteins have also been linked to cellular stress responses, including osmotic stress [69-71].

With information gained from gene neighbourhood analysis and distant structural homology we can speculate as to the mechanisms of salt tolerance conferred by *stlA*. Overall, *stlA* and its neighbouring genes share common features that categorise them as stress responsive and may therefore constitute a stress operon. Three of the five encoded unknown proteins share a distant structural homology to chaperones. These chaperones could play a role in protein disaggregation and folding following stress as outlined above, or they could guide StlA through the periplasm and assist in inserting it in the membrane, although the latter situation would require *E. coli* chaperones to function in a similar capacity when *stlA* is cloned in isolation. StlA itself, being a predicted membrane protein could act as a sensor to external stresses or indeed stabilise the outer membrane during stress. It is noteworthy that the most significant homology predicted by FFAS for StlA was to a TspO/MBR protein which is involved in membrane signalling and is associated with voltage-dependent anion channels in mitochondria.

In conclusion, we have identified a novel salt tolerance gene, *stlA*, from the human gut microbiome through functional screening of a metagenomic library. The gene is rare among the HMP and MetaHit datasets and has no bacterial, archaeal, viral, plasmid or eukaryotic homologues in the current databases. Furthermore, no homologues were found in any non-human metagenome datasets nor in any of the human microbiome datasets (HMP and MetaHit) other than stool, indicating it is gut specific and present in a novel species of low abundance. The *stlA* gene appears to be on a prophage, indicating it may have been acquired (along with some of its neighbouring genes) through a LGT event and may confer a competitive advantage to its particular host species under stressful conditions in the gut or if there is an absence of or deficiency in some of the classical osmotolerance systems, such as in *C. jejuni* [72].

Overall this study illustrates the utility of functionally screening metagenomic libraries to assign a function to a completely novel gene and its encoded protein and suggests that novel mechanisms of osmotolerance may exist in different environmental niches. Mining (gut) microbiomes and the development of more sensitive and innovative screening assays will facilitate the discovery of novel stress resistance genes, antibiotics, biopharmaceuticals and bio therapeutics for use in biotechnology, medicine and health [73-77].

Acknowledgements

We thank Dr. Susan Joyce for kindly providing human bile samples and for advice regarding bile experiments and also Professor Paul O'Toole and Jennifer Deane for kindly providing DNA from faecal microbiota from the ELDERMET study [27]. We thank Alan Lucid for assistance with PhiSpy.

References

1. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

2. Clemente JC, Ursell LK, Parfrey LW, Knight R (2012) The impact of the gut microbiota on human health: an integrative view. Cell 148: 1258-1270.

3. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.

4. Kovatcheva-Datchary P, Zoetendal EG, Venema K, de Vos WM, Smidt H (2009) Tools for the tract: understanding the functionality of the gastrointestinal tract. Therap Adv Gastroenterol 2: 9-22.

5. Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, et al. (2011) Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proc Natl Acad Sci U S A 108: 6252-6257.

6. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, et al. (2012) Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect.

7. Walker AW, Ince J, Duncan SH, Webster LM, Holtrop G, et al. (2011) Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J 5: 220-230.

8. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669-685.

9. Sleator RD, Shortall C, Hill C (2008) Metagenomics. Lett Appl Microbiol 47: 361-366.

10. Sleator RD, Watson D, Hill C, Gahan CG (2009) The interaction between Listeria monocytogenes and the host gastrointestinal tract. Microbiology 155: 2463-2475.

11. Louis P, O'Byrne CP (2010) Life in the gut: microbial responses to stress in the gastrointestinal tract. Sci Prog 93: 7-36.

12. Epstein W (2003) The roles and regulation of potassium in bacteria. Prog Nucleic Acid Res Mol Biol 75: 293-320.

13. Kempf B, Bremer E (1998) Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. Arch Microbiol 170: 319-330.

14. Kunte HJ (2006) Osmoregulation in Bacteria: Compatible Solute Accumulation and Osmosensing. Environmental Chemistry 3: 94-99.

15. Sleator RD, Hill C (2002) Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. FEMS Microbiol Rev 26: 49-71.

16. Kapardar R, Ranjan R, Puri M, Sharma R (2010(b)) Sequence analysis of a salt tolerant metagenomic clone. Indian Journal of Microbiology 50: 212-215.

17. Kapardar RK, Ranjan R, Grover A, Puri M, Sharma R (2010(a)) Identification and characterization of genes conferring salt tolerance to Escherichia coli from pond water metagenome. Bioresour Technol 101: 3917-3924.

18. Sakamoto T, Murata N (2002) Regulation of the desaturation of fatty acids and its role in tolerance to cold and salt stress. Curr Opin Microbiol 5: 208-210.

19. Sleator RD, Hill C (2005) A novel role for the LisRK two-component regulatory system in listerial osmotolerance. Clin Microbiol Infect 11: 599-601.

20. Bork P (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 10: 398-400.

21. Culligan EP, Sleator RD, Marchesi JR, Hill C (2012(a)) Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. ISME J 6: 1916-1925.

22. Haardt M, Kempf B, Faatz E, Bremer E (1995) The osmoprotectant proline betaine is a major substrate for the binding-protein-dependent transport system ProU of Escherichia coli K-12. Mol Gen Genet 246: 783-786.

23. Gasson MJ (1983) Plasmid complements of Streptococcus lactis NCDO 712 and other lactic streptococci after protoplast-induced curing. J Bacteriol 154: 1-9.

24. Hayes F, Daly C, Fitzgerald GF (1990) Identification of the Minimal Replicon of Lactococcus lactis subsp. lactis UC317 Plasmid pCI305. Appl Environ Microbiol 56: 202-209.

25. Jones BV, Marchesi JR (2007) Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. Nat Methods 4: 55-61.

26. Jones BV, Begley M, Hill C, Gahan CG, Marchesi JR (2008) Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A 105: 13580-13585.

27. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, et al. (2012) Gut microbiota composition correlates with diet and health in the elderly. Nature 488: 178-184.

28. Knopp S, Mohammed KA, Stothard JR, Khamis IS, Rollinson D, Marti H, Utzinger J (2010) Patterns and risk factors of helminthiasis and anemia in a rural and a peri-urban community in Zanzibar, in the context of helminth control programs. PLOS Negl Trop Dis (4):5 e 681.

29. Bochner BR (2009) Global phenotypic characterization of bacteria. FEMS Microbiol Rev 33: 191-205.

30. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22: 195-201.

31. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.

32. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-37.

33. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 11: 681-684.

34. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225-229.

35. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725-738.

36. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-166

37. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33: W244-248.

38. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, et al. (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112: 531-552.

39. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80: 1715-1735.

40. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9: 40.

41. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6: 175-182.

42. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33: W284-288.

43. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: D534-538.

44. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, et al. (2012) IMG/M-HMP: a metagenome comparative analysis system for the Human Microbiome Project. PLoS One 7: e40151.

45. Human_Microbiome_Project_Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.

46. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.

47. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-439.

48. Derrien M, Vaughan EE, Plugge CM, de Vos WM (2004) Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. Int J Syst Evol Microbiol 54: 1469-1476.

49. van Passel MW, Kant R, Zoetendal EG, Plugge CM, Derrien M, et al. (2011) The genome of Akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. PLoS One 6: e16876.

50. Derrien M, Collado MC, Ben-Amor K, Salminen S, de Vos WM (2008) The Mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. Appl Environ Microbiol 74: 1646-1648

51. Liu L, Li Y, Li S, Hu N, He Y, et al. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.

52. Kao DY, Cheng YC, Kuo TY, Lin SB, Lin CC, et al. (2009) Salt-responsive outer membrane proteins of Vibrio anguillarum serotype O1 as revealed by comparative proteome analysis. J Appl Microbiol 106: 2079-2085.

53. Wexler HM, Tenorio E, Pumbwe L (2009) Characteristics of Bacteroides fragilis lacking the major outer membrane protein, OmpA. Microbiology 155: 2694-2706.

54. McEnery MW, Snowman AM, Trifiletti RR, Snyder SH (1992) Isolation of the mitochondrial benzodiazepine receptor: association with the voltage-dependent anion channel and the adenine nucleotide carrier. Proc Natl Acad Sci U S A 89: 3170-3174.

55. Yeliseev AA, Kaplan S (1995) A sensory transducer homologous to the mammalian peripheral-type benzodiazepine receptor regulates photosynthetic membrane complex formation in Rhodobacter sphaeroides 2.4.1. J Biol Chem 270: 21167-21175.

56. Cytryn EJ, Sangurdekar DP, Streeter JG, Franck WL, Chang WS, et al. (2007) Transcriptional and physiological responses of Bradyrhizobium japonicum to desiccation-induced stress. J Bacteriol 189: 6751-6762.

57. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, et al. (2010) Cryptic prophages help bacteria cope with adverse environments. Nat Commun 1: 147.

58. Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52: 413-435.

59. Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14: 303-310.

60. Kotik M (2009) Novel genes retrieved from environmental DNA by polymerase chain reaction: current genome-walking techniques for future metagenome applications. J Biotechnol 144: 75-82.

61. Tuffin M, Anderson D, Heath C, Cowan DA (2009) Metagenomic gene discovery: how far have we moved into novel sequence space? Biotechnol J 4: 1671-1683.

62. Lund PA (2001) Microbial molecular chaperones. Adv Microb Physiol 44: 93-140.

63. Chintakayala K, Grainger DC (2011) A conserved acidic amino acid mediates the interaction between modulators and co-chaperones in enterobacteria. J Mol Biol 411: 313-320.

64. Prasad J, McJarrow P, Gopal P (2003) Heat and osmotic stress responses of probiotic Lactobacillus rhamnosus HN001 (DR20) in relation to viability after drying. Appl Environ Microbiol 69: 917-925.

65. Yang XX, Maurer KC, Molanus M, Mager WH, Siderius M, et al. (2006) The molecular chaperone Hsp90 is required for high osmotic stress response in Saccharomyces cerevisiae. FEMS Yeast Res 6: 195-204.

66. Al-Khodor S, Price CT, Kalia A, Abu Kwaik Y (2010) Functional diversity of ankyrin repeats in microbial proteins. Trends Microbiol 18: 132-139.

67. Bennett V, Chen L (2001) Ankyrins and cellular targeting of diverse membrane proteins to physiological sites. Curr Opin Cell Biol 13: 61-67.

68. Li J, Mahajan A, Tsai MD (2006) Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry 45: 15168-15178.

69. Chinchilla D, Merchan F, Megias M, Kondorosi A, Sousa C, et al. (2003) Ankyrin protein kinases: a novel type of plant kinase gene whose expression is induced by osmotic stress in alfalfa. Plant Mol Biol 51: 555-566.

70. Flint A, Sun YQ, Stintzi A (2012) Cj1386 is an ankyrin-containing protein involved in heme trafficking to catalase in Campylobacter jejuni. J Bacteriol 194: 334-345.

71. Seong ES, Choi D, Cho HS, Lim CK, Cho HJ, et al. (2007) Characterization of a stress-responsive ankyrin repeat-containing zinc finger protein of Capsicum annuum (CaKR1). J Biochem Mol Biol 40: 952-958.

72. Cameron A, Frirdich E, Huynh S, Parker CT, Gaynor EC (2012) The hyperosmotic stress response of Campylobacter jejuni. J Bacteriol.

73. Collison M, Hirt RP, Wipat A, Nakjang S, Sanseau P, et al. (2012) Data mining the human gut microbiota for therapeutic targets. Brief Bioinform.

74. Culligan EP, Hill C, Sleator RD (2009) Probiotics and gastrointestinal disease: successes, problems and future prospects. Gut Pathog 1: 19.

75. Culligan EP, Marchesi JR, Hill C, Sleator RD (2012(b)) Mining the human gut microbiome for novel stress resistance genes. Gut Microbes 3: 394-397.

76. Yang JY, Karr JR, Watrous JD, Dorrestein PC (2011) Integrating '-omics' and natural product discovery platforms to investigate metabolic exchange in microbiomes. Curr Opin Chem Biol 15: 79-87.

77. Delavat F, Phalip V, Forster A, Plewniak F, Lett MC, et al. (2012) Amylases without known homologues discovered in an acid mine drainage: significance and impact. Sci Rep 2: 354.

THESIS DISCUSSION

Functional metagenomics can provide a means to access novel genes from as yet uncultured bacteria and to subsequently assign a function to the encoded proteins. Since its emergence as a research field many novel genes encoding diverse proteins such as rhodopsins, esterases, lipases, proteases and antibiotics have been identified in diverse environments [1-6]. Metagenomic strategies have also been employed to study the human gut microbiota, leading to the discovery of novel genes which encode β-glucuronidases, bile salt hydrolases (BSH's), antibiotic resistance determinants, NF-$_{\kappa}$β modulators and proteins involved in intestinal colonisation [7-11]. Metagenomics has been used to identify novel salt tolerance genes in pond water [12,13]; however, no one to our knowledge has investigated the human gut microbiome as source of novel salt tolerance genes. With this in mind, a functional metagenomic approach was employed to identify novel genes from the human gut microbiota involved in the osmotic stress response.

The osmotic stress response involves a number of phases; firstly the uptake of potassium (K^+), followed by the synthesis or uptake of compatible solutes (osmoprotectants) comprise the primary and secondary responses, respectively [14-16]. Ancillary systems, which encompass a broad range of cellular processes, also play an important role in the global cellular osmotic stress response. Much research has focused on elucidating the mechanisms of the primary and secondary response, but comparatively less information is available regarding more diverse genes involved in the overall cellular response. Consequently, this thesis has focused on identifying novel ancillary components of the overall osmotic stress response, thus giving a broader and more comprehensive view of the range of genes involved.

Chapter II describes an initial screen of a human gut microbiota metagenomic library to identify salt-tolerant clones and the subsequent characterisation of a number of these clones. From a screen of over 20,000 clones, fifty-three salt tolerant clones were identified and annotated as SMG 1-53 (salt metagenome 1-53). Transposon mutagenesis identified five novel salt tolerance genes from clones SMG 3, SMG 5 and SMG 25. The genes shared homology to *murB*, *mazG* (x2) and *galE* (x2), which had not been previously linked to salt tolerance. The genes originate from species of common members of the gut microbiota such as *Collinsella*, *Akkermansia* and *Eggerthella*. Cloning and subsequent expression of each gene in isolation conferred a significantly increased salt tolerance phenotype to *Escherichia coli*. To our knowledge, this chapter represents the first study to identify such novel genes from the human gut microbiota.

Chapter IV details the *in silico* identification of a *brp*/*blh* family beta-carotene 15,15'-monoxygenase (BrpA) as a novel salt tolerance determinant. Initially pursued for further investigation because BrpA shared homology to a proline symporter, it later transpired that while it did confer a salt tolerance phenotype, BrpA (when expressed from an inducible expression vector) also resulted in cell pellets acquiring a red/orange colour when grown in the presence of exogenous beta-carotene, indicating incorporation of carotenoids in the cell membrane. Furthermore, the *brpA* gene appears to have been acquired via lateral gene transfer (LGT); having a low percentage G+C content (~32%) compared to many of the other genes on the fosmid insert. Finally, *brpA* homologues are found most abundantly in the stool, supra-gingival plaque and tongue metagenomes from the Human Microbiome Project (HMP) dataset.

In **Chapter V** we report the identification of *stlA* (salt tolerance locus A), a novel salt tolerance gene with no current known homologues in the non-redundant sequence databases. The *stlA* gene was identified following transposon mutagenesis of clone SMG 25. *E. coli* cells expressing *stlA* had a significantly higher tolerance to sodium chloride (NaCl) stress. An *in silico* analysis revealed *stlA* and a number of other genes on the SMG 25 fosmid insert were encoded on a prophage, possibly indicating the importance of these genes for niche adaptation. Furthermore, when searched against metagenomic datasets from the Human Microbiome Project (HMP) and MetaHit, *stlA* appears to be extremely rare, with only 10 homologues found between both datasets and present only in stool microbiome samples, indicating StlA may be a gut-specific protein and found in a species of low abundance. This is further emphasised by the fact that StlA homologues were not found in any other non-human metagenomic datasets (>1300 samples), nor in any finished, draft or permanent draft genome sequences from bacteria, archaea, viruses, eukaryotes or plasmids. StlA represents the first protein of its kind to be identified and functionally characterised.

Overall this thesis describes the identification of a number of novel salt tolerance genes from the human gut microbiota using a functional metagenomic approach and represents a significant contribution to the literature on the bacterial stress response. By employing a functional metagenomic approach, genes without any obvious connection to the osmotic stress response have been identified. We cannot rule out completely the possibility that the acquired salt tolerance is an artefact of expression of these genes in *E. coli*; having said that, we feel it is likely they also act as salt tolerance genes in their original host(s) and put forward what we feel are plausible theories on their mechanisms of action. It would be interesting to test the salt tolerance of the various host species in wild-type and knock-out mutant backgrounds. However, owing to the diversity and uncultured nature of some of these species, the development of the genetic tools to manipulate and test this was simply not feasible. This being apparent from the outset, the main aim was to identify novel genes that could confer increased salt tolerance to a heterologous host using a functional metagenomic approach. One must also keep in mind that some novel salt tolerance genes will have been missed by our screen, owing to the well documented expression issues for foreign DNA in *E. coli* due to incompatibilities of transcriptional, translational and post-translational machinery between heterologous host and the organism from which metagenomic DNA originated.

On a broader scale salt tolerance genes such as those identified during the course of this thesis could be used to increase both the processing and host-associated stress resistance of probiotic microorganisms, ultimately increasing gastrointestinal persistence and therapeutic efficacy. Some of these genes could perhaps be candidates for novel drug targets also. Indeed, as outlined in **Chapter III** (addendum to Chapter II), *murB* and *galE* as well as components of toxin-antitoxin (TA) systems such as *mazG* have been utilised or investigated as novel drug targets [17-20].

A recent study by Gaddy and co-workers [21] raises some interesting questions regarding salt intake and the microbiota. The study found that gerbils fed a high salt diet and infected with *cagA*$^+$ *Helicobacter pylori* developed gastric cancer at higher rates than controls. The *cagA* gene encodes a bacterial oncoprotein, CagA, and is upregulated in response to salt, both *in vitro* and *in vivo*. Considering that many virulence, virulence-associated and of course salt tolerance genes themselves are often upregulated in response to salt, it is interesting to speculate on the effect of a high dietary salt intake on the gut microbiota. For instance, does a high salt intake result in quantifiable changes to the composition of the gut microbiota or could high salt intake promote the growth of certain species or provide a cue to opportunistic pathogens to initiate a process of infection or inflammation under certain conditions? Previous studies have shown that fluctuations in osmolarity can trigger changes in gene expression in pathogenic bacteria such as *Salmonella* and *Listeria monocytogenes* signalling a transition to invasive infection [22,23]. Furthermore, changes in intestinal ion transport and more specifically increased sodium (Na^+) concentration in NHE3 (Na^+/H^+ exchanger isoform 3) deficient mice results in changes to the microbiota and increases in *Bacteroides thetaiotaomicron* [24], whilst deficiencies in sodium absorption and consequently increases in luminal osmolarity are associated with chronic gastrointestinal conditions such as IBD (inflammatory bowel disease) and thus have implications for the microbial community composition. Studies to examine the effect (if any) of high dietary salt intake on the composition of the gut microbiota and on the pathophysiology of the intestinal mucosa may be a worthwhile endeavour.

In conclusion, this thesis represents the first study to identify novel salt tolerance genes from the human microbiota and illustrates the utility of functional metagenomics for novel gene discovery in complex environmental niches. Overall, it provides a more comprehensive view of and expands our knowledge in relation to the diverse systems employed by bacteria in response to osmotic stress.

References

1. Banik JJ, Brady SF (2008) Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proc Natl Acad Sci U S A 105: 17273-17277.

2. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289: 1902-1906.

3. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, et al. (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol 68: 4301-4306.

4. Heath C, Hu XP, Cary SC, Cowan D (2009) Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Appl Environ Microbiol 75: 4657-4659.

5. Lee DG, Jeon JH, Jang MK, Kim NY, Lee JH, et al. (2007) Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library. Biotechnol Lett 29: 465-472.

6. Meilleur C, Hupe JF, Juteau P, Shareck F (2009) Isolation and characterization of a new alkali-thermostable lipase cloned from a metagenomic library. J Ind Microbiol Biotechnol 36: 853-861.

7. Gloux K, Berteau O, El Oumami H, Beguet F, Leclerc M, et al. (2011) A metagenomic beta-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. Proc Natl Acad Sci U S A 108 Suppl 1: 4539-4546.

8. Jones BV, Begley M, Hill C, Gahan CG, Marchesi JR (2008) Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A 105: 13580-13585.

9. Kazimierczak KA, Rincon MT, Patterson AJ, Martin JC, Young P, et al. (2008) A new tetracycline efflux gene, tet(40), is located in tandem with tet(O/32/O) in a human gut firmicute bacterium and in metagenomic library clones. Antimicrob Agents Chemother 52: 4001-4009.

10. Lakhdari O, Cultrone A, Tap J, Gloux K, Bernard F, et al. (2010) Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-kappaB modulation in the human gut. PLoS One 5.

11. Yoon MY, Lee KM, Yoon Y, Go J, Park Y, et al. (2013) Functional screening of a metagenomic library reveals operons responsible for enhanced intestinal colonization by gut commensal microbes. Appl Environ Microbiol 79: 3829-3838.

12. Kapardar R, Ranjan R, Puri M, Sharma R (2010(b)) Sequence analysis of a salt tolerant metagenomic clone. Indian Journal of Microbiology 50: 212-215.

13. Kapardar RK, Ranjan R, Grover A, Puri M, Sharma R (2010(a)) Identification and characterization of genes conferring salt tolerance to Escherichia coli from pond water metagenome. Bioresour Technol 101: 3917-3924.

14. Epstein W (2003) The roles and regulation of potassium in bacteria. Prog Nucleic Acid Res Mol Biol 75: 293-320.

15. Kempf B, Bremer E (1998) Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. Arch Microbiol 170: 319-330.

16. Kunte HJ (2006) Osmoregulation in Bacteria: Compatible Solute Accumulation and Osmosensing. Environmental Chemistry 3: 94-99.

17. Engelberg-Kulka H, Sat B, Reches M, Amitai S, Hazan R (2004) Bacterial programmed cell death systems as targets for antibiotics. Trends Microbiol 12: 66-71.

18. Germanier R, Fuer E (1975) Isolation and characterization of Gal E mutant Ty 21a of Salmonella typhi: a candidate strain for a live, oral typhoid vaccine. J Infect Dis 131: 553-558.

19. Shapiro AB, Livchak S, Gao N, Whiteaker J, Thresher J, et al. (2012) A homogeneous, high-throughput-compatible, fluorescence intensity-based assay for UDP-N-acetylenolpyruvylglucosamine reductase (MurB) with nanomolar product detection. J Biomol Screen 17: 327-338.

20. Urbaniak MD, Tabudravu JN, Msaki A, Matera KM, Brenk R, et al. (2006) Identification of novel inhibitors of UDP-Glc 4'-epimerase, a validated drug target for african sleeping sickness. Bioorg Med Chem Lett 16: 5744-5747.

21. Gaddy JA, Radin JN, Loh JT, Zhang F, Washington MK, et al. (2013) High dietary salt intake exacerbates Helicobacter pylori-induced gastric carcinogenesis. Infect Immun 81: 2258-2267.

22. Tartera C, Metcalf ES (1993) Osmolarity and growth phase overlap in regulation of Salmonella typhi adherence to and invasion of human intestinal cells. Infect Immun 61: 3084-3089.

23. Sleator RD, Clifford T, Hill C (2007) Gut osmolarity: a key environmental cue initiating the gastrointestinal phase of Listeria monocytogenes infection? Med Hypotheses 69: 1090-1092.

24. Engevik MA, Aihara E, Montrose MH, Shull GE, Hassett DJ, et al. (2013) Loss of NHE3 alters gut microbiota composition and influences Bacteroides thetaiotaomicron growth. Am J Physiol Gastrointest Liver Physiol 305: G697-71

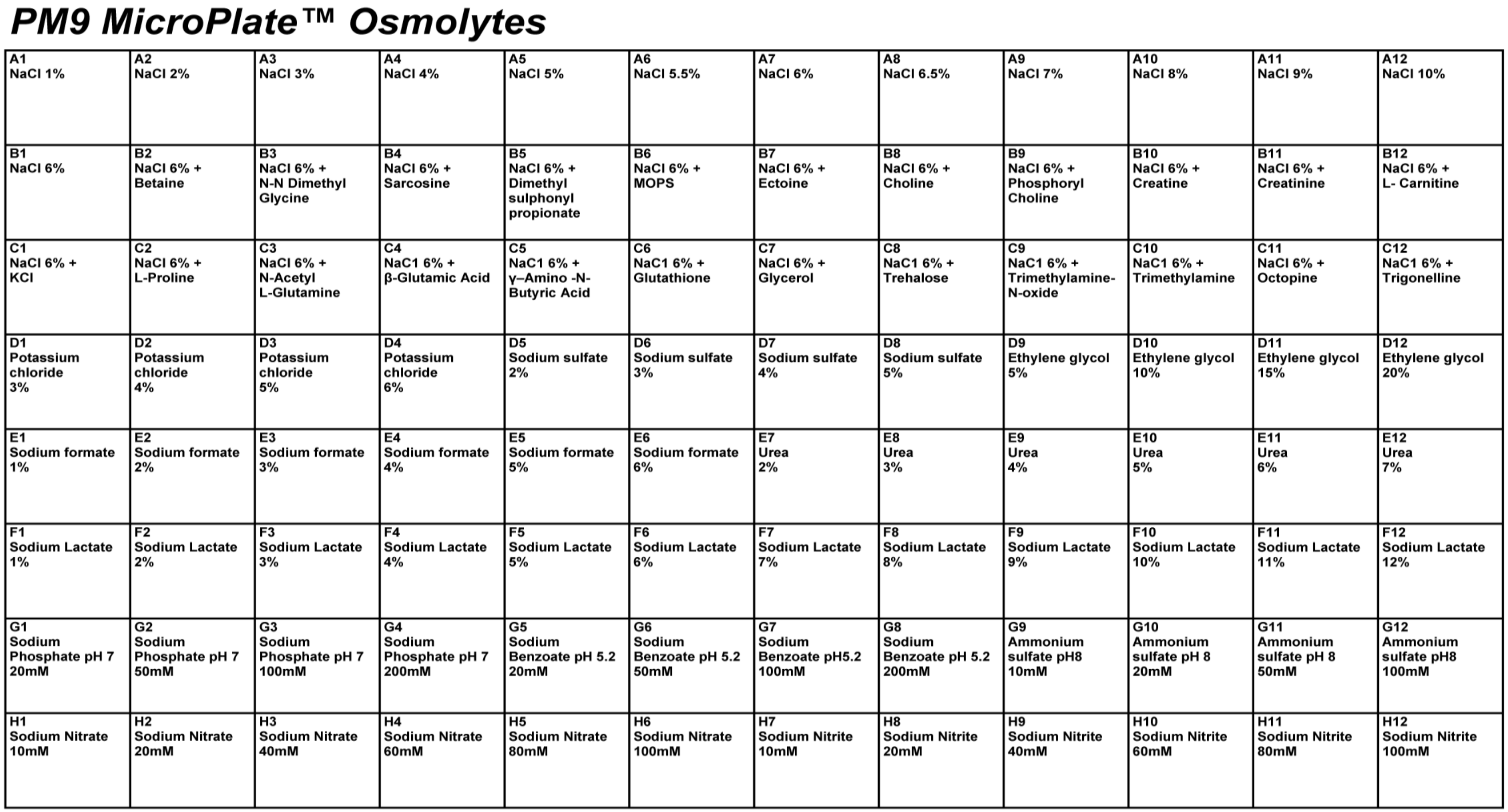
APPENDIX

FRONT. MICROBIOL. 2014, APR; 5:189

This appendix contains results of additional experiments carried out during the course of the thesis. The following work represents an initial characterisation of clone SMG 9, but not to the same extent as other clones detailed in the preceding chapters.

BIOLOG phenotypic microarray (PM) assay was used to identify possible osmotolerance-related phenotypes of SMG 9. Using plate PM 9 (Figure 1), which contains 96 different osmotic stress conditions and osmolytes, indicated that SMG 9 had the ability to utilise L-carnitine in the presence of 6% sodium chloride (NaCl), as revealed by the development of a purple colour in the corresponding well of plate PM9 (Figure 2A). No colour development was observed in plates inoculated with the control strain carrying an empty fosmid vector, *E. coli* EPI300::pCC1FOS (Figure 2B).

Figure 1. Layout of BIOLOG PM 9 osmolyte plate.



PM9 MicroPlate™ Osmolytes

A1 NaCl 1%	A2 NaCl 2%	A3 NaCl 3%	A4 NaCl 4%	A5 NaCl 5%	A6 NaCl 5.5%	A7 NaCl 6%	A8 NaCl 6.5%	A9 NaCl 7%	A10 NaCl 8%	A11 NaCl 9%	A12 NaCl 10%
B1 NaCl 6%	B2 NaCl 6% + Betaine	B3 NaCl 6% + N-N Dimethyl Glycine	B4 NaCl 6% + Sarcosine	B5 NaCl 6% + Dimethyl sulphonyl propionate	B6 NaCl 6% + MOPS	B7 NaCl 6% + Ectoine	B8 NaCl 6% + Choline	B9 NaCl 6% + Phosphoryl Choline	B10 NaCl 6% + Creatine	B11 NaCl 6% + Creatinine	B12 NaCl 6% + L- Carnitine
C1 NaCl 6% + KCl	C2 NaCl 6% + L-Proline	C3 NaCl 6% + N-Acetyl L-Glutamine	C4 NaC1 6% + β-Glutamic Acid	C5 NaC1 6% + γ–Amino -N-Butyric Acid	C6 NaC1 6% + Glutathione	C7 NaCl 6% + Glycerol	C8 NaC1 6% + Trehalose	C9 NaC1 6% + Trimethylamine-N-oxide	C10 NaC1 6% + Trimethylamine	C11 NaCl 6% + Octopine	C12 NaC1 6% + Trigonelline
D1 Potassium chloride 3%	D2 Potassium chloride 4%	D3 Potassium chloride 5%	D4 Potassium chloride 6%	D5 Sodium sulfate 2%	D6 Sodium sulfate 3%	D7 Sodium sulfate 4%	D8 Sodium sulfate 5%	D9 Ethylene glycol 5%	D10 Ethylene glycol 10%	D11 Ethylene glycol 15%	D12 Ethylene glycol 20%
E1 Sodium formate 1%	E2 Sodium formate 2%	E3 Sodium formate 3%	E4 Sodium formate 4%	E5 Sodium formate 5%	E6 Sodium formate 6%	E7 Urea 2%	E8 Urea 3%	E9 Urea 4%	E10 Urea 5%	E11 Urea 6%	E12 Urea 7%
F1 Sodium Lactate 1%	F2 Sodium Lactate 2%	F3 Sodium Lactate 3%	F4 Sodium Lactate 4%	F5 Sodium Lactate 5%	F6 Sodium Lactate 6%	F7 Sodium Lactate 7%	F8 Sodium Lactate 8%	F9 Sodium Lactate 9%	F10 Sodium Lactate 10%	F11 Sodium Lactate 11%	F12 Sodium Lactate 12%
G1 Sodium Phosphate pH 7 20mM	G2 Sodium Phosphate pH 7 50mM	G3 Sodium Phosphate pH 7 100mM	G4 Sodium Phosphate pH 7 200mM	G5 Sodium Benzoate pH 5.2 20mM	G6 Sodium Benzoate pH 5.2 50mM	G7 Sodium Benzoate pH5.2 100mM	G8 Sodium Benzoate pH 5.2 200mM	G9 Ammonium sulfate pH8 10mM	G10 Ammonium sulfate pH 8 20mM	G11 Ammonium sulfate pH 8 50mM	G12 Ammonium sulfate pH8 100mM
H1 Sodium Nitrate 10mM	H2 Sodium Nitrate 20mM	H3 Sodium Nitrate 40mM	H4 Sodium Nitrate 60mM	H5 Sodium Nitrate 80mM	H6 Sodium Nitrate 100mM	H7 Sodium Nitrite 10mM	H8 Sodium Nitrite 20mM	H9 Sodium Nitrite 40mM	H10 Sodium Nitrite 60mM	H11 Sodium Nitrite 80mM	H12 Sodium Nitrite 100mM

Figure 2.

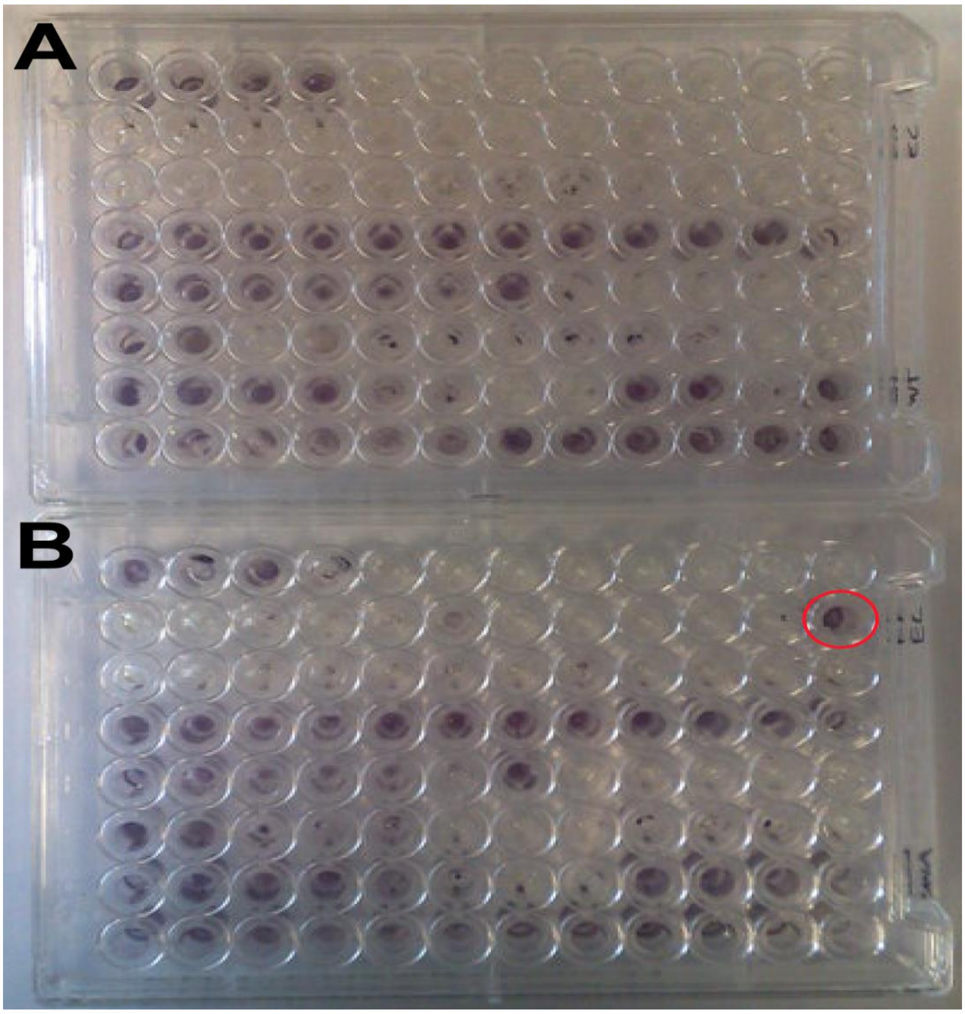


Figure 2. Appearance of PM 9 plates after incubation for 24 hours at 37°C. (A) Control EPI300::pCC1FOS and (B) SMG 9. PM plates measure cellular respiration colorimetrically via reduction of a tetrazolium dye with electrons from NADH generated during the process of respiration. Strongly metabolised substrates generate a more intense purple colour. Development of a strong purple colour can be seen in well B12 in Figure 2B (circled in red), which was inoculated with SMG 9, while no colour development is visible in B12 of the control plate. This indicates SMG 9 has a greater ability to transport and utilise L-carnitine compared to the EPI300::pCC1FOS host strain.

Figure 3.

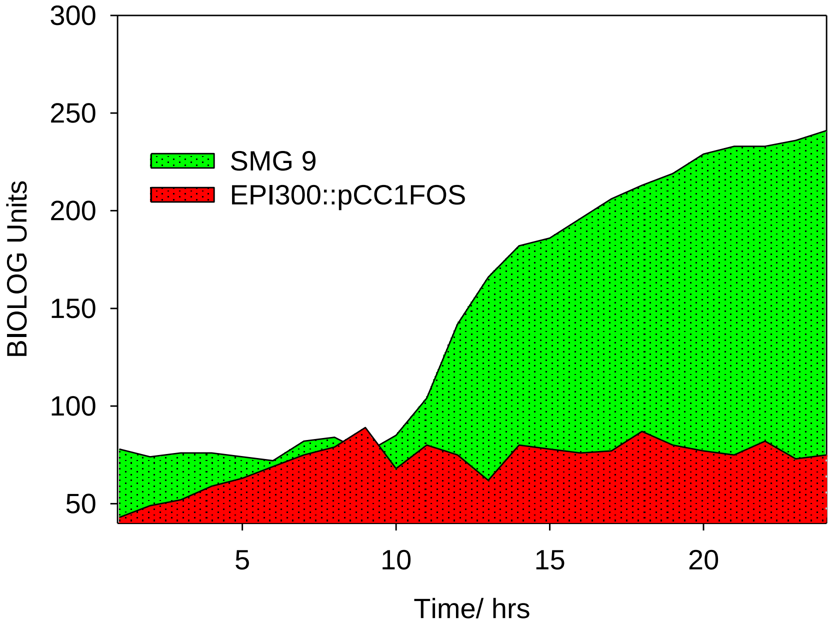


Figure 3. Kinetic data measured by BIOLOG Omnilog system.

Colour formation within each well was measured by BIOLOG's Omnilog machine, which produces a colour coded graph. Kinetic data from two clones can be compared. EPI300::pCC1FOS is shown in red and that from SMG 9 is shown in green. The green colour indicates more rapid metabolism by SMG 9 under the conditions in the well (6% NaCl + L-carnitine).

Growth experiments were performed in M9 minimal media containing a range of NaCl concentrations (0-6%) in the presence and absence of 1mM L-carnitine to investigate the phenotype further.

Figure 4.

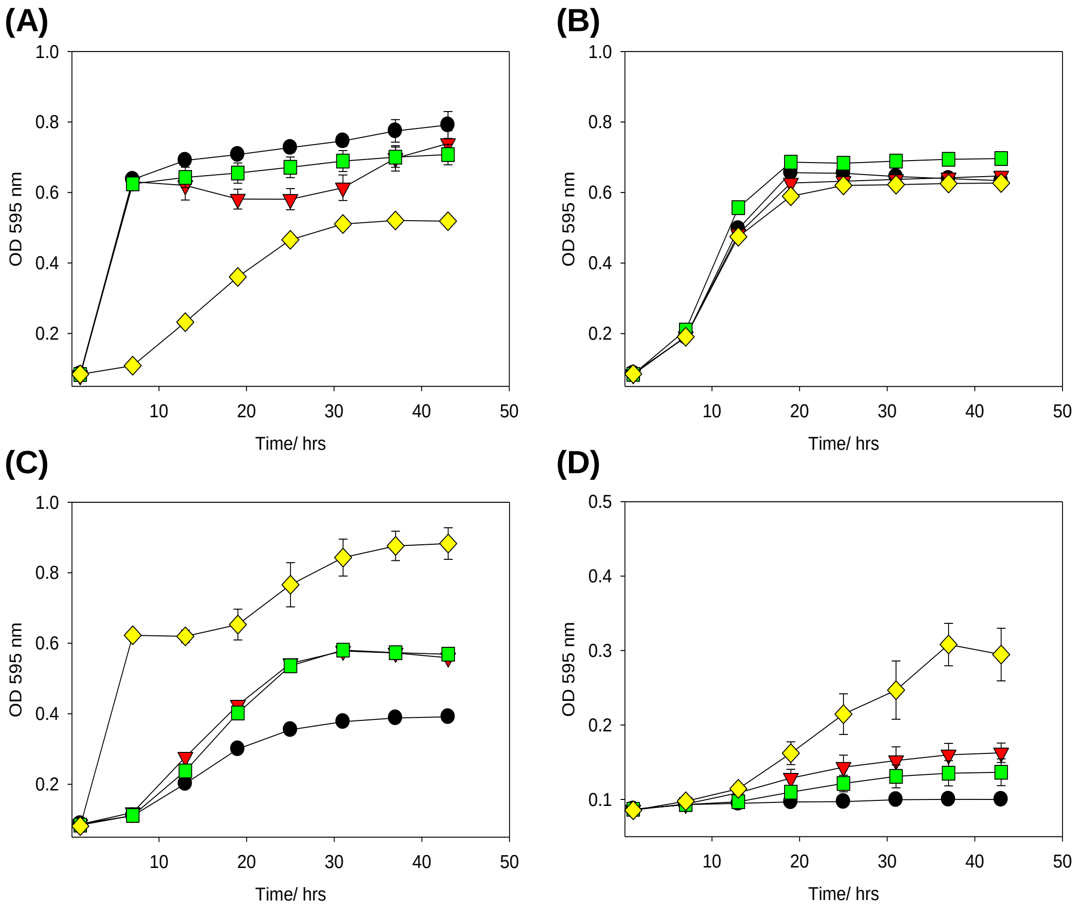


Figure 4. Growth of *E. coli* EPI300::pCC1FOS and SMG 9 in (A) M9 minimal media, (B) M9 minimal media + 4% NaCl, (C) M9 minimal media + 5% NaCl and (D) M9 minimal media + 6% NaCl.

Legend: *E. coli* EPI300::pCC1FOS (● black circle); SMG 9 (▼ red triangle); *E. coli* EPI300::pCC1FOS + 1mM L-carnitine (🟩 green square); SMG 9 + 1mM L-carnitine (◆, yellow diamond).

Figure 4A shows growth of both clones in M9 media in the presence and absence of 1mM L-carnitine. In the presence of L-carnitine, SMG 9 displays a growth defect, while growth is similar under all other conditions. In the presence of 4% NaCl there is no difference in growth between clones either in the presence or absence of L-carnitine (Figure 4B). At 5% NaCl however, SMG 9 has a significant growth advantage compared to EPI300::pCC1FOS both in the presence and absence of 1mM L-carnitine (Figure 4C). The positive effect of L-carnitine on the growth of SMG 9 is evident with cells entering logarithmic phase growth sooner and reaching a much higher final optical density (OD_{595nm}). A similar effect for L-carnitine is seen for 6% NaCl.

Sequencing of SMG 9 fosmid insert

The SMG 9 fosmid insert was sequenced using the GS-FLX platform (Roche) on a Titanium mini-run. Gene prediction with FGENESB (www.softberry.com) and functional annotation of predicted amino acid sequences using BLASTP (Table 1) was performed as described in the preceding chapters.

Table 1. Proteins predicted to be encoded on SMG 9 fosmid insert

Gene #	Putative Encoded Function	# a.a	Closest Hit Organism	% Coverage	e-value	% ID	Domains
1	ATP-dependent chaperone ClpB	554	*Bacteroides sp.* CAG:545	97%	0.00E+00	99%	COG0714; AAA ATPase
2	Preprotein translocase SecG subunit	121	*Bacteroides sp.* CAG:545	100%	3.00E-77	100%	SecG
3	Putative uncharacterized protein	187	*Bacteroides sp.* CAG:545	100%	3.00E-134	100%	None detected
4	Putative uncharacterized protein	177	*Bacteroides sp.* CAG:545	86%	3.00E-108	99%	LptE
5	**Transcriptional regulator**	**432**	***Bacteroides sp.* CAG:545**	**100%**	**0.00E+00**	**99%**	**AAA ATPase; sigma-54 interaction domain; HTH_8 bacterial regulatory protein, Fis family domain**
6	Putative uncharacterized protein	545	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	TadD
7	Putative uncharacterized protein	1015	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	SecD, SecF
8	OmpA/MotB domain protein	618	*Bacteroides sp.* CAG:545	71%	0.00E+00	99%	PD40, similar to WD40 domain
9	Putative uncharacterized protein	155	*Bacteroides sp.* CAG:545	100%	4.00E-106	100%	NfeD
10	uPF0365 protein AL1_06760	317	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	YdfA_immunity superfamily
11	Putative uncharacterized protein	142	*Bacteroides sp.* CAG:545	100%	1.00E-97	98%	Lipocalin_4
12	Subtilisin-like serine protease	678	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	Peptidase_S8_S53 superfamily
13	Uncharacterized protein	812	*Bacteroides sp.* CAG:545	99%	0.00E+00	99%	None detected
14	RagB/SusD family protein	547	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	Two SusD superfamily
15	Outer membrane receptor for ferrienterochelin and colicins	1068	*Bacteroides sp.* CAG:545	100%	0.00E+00	100%	Can_B2; Plug; OM channel; OMP_RagA_SusC
16	Alpha-L-fucosidase-like	513	*Bacteroides sp.* CAG:545	100%	0.00E+00	98%	COG3669; Alpha_L_fucos; F5_F8_Type_C
17	Putative uncharacterized protein	446	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	DHQ_FE-ADH (Dehydroquinate iron aldehyde dehydrogenase)
18	**Major facilitator superfamily MFS_1**	**487**	***Bacteroides sp.* CAG:545**	**100%**	**0.00E+00**	**100%**	**MFS; UhpC, sugar phosphate permease**
19	ThiF family protein	242	*Bacteroides sp.* CAG:545	100%	4.00E-174	98%	YgdL_like
20	Putative uncharacterized protein	649	*Bacteroides sp.* CAG:545	100%	0.00E+00	98%	Glyco_hydro_97
21	DNA mismatch repair protein MutS	894	*Bacteroides sp.* CAG:545	98%	0.00E+00	99%	PRK05399; MutS-I; MutS_II; MutS_III; ABC_MutS_1
22	Glycosyl transferase group 1	378	*Bacteroides sp.* CAG:545	100%	0.00E+00	99%	RfaG
23	No significant similarity found	68	*Bacteroides sp.* CAG:545	N/A	N/A	N/A	N/A
24	Hypothetical protein Fjoh_3657	162	*Flavobacterium johnsoniae* UW101	80%	4.00E-25	40%	AdkA, archaeal adenylate kinase

Functional assignment is based on BLASTP of amino acid sequences predicted by Softberry's FGENESB. Abbreviations: # a.a = number of predicted amino acids; %ID = % identity at the amino acid level; N/A = not applicable.

Following sequencing and functional annotation of predicted amino acid sequences on the SMG 9 fosmid insert no obvious L-carnitine transport or utilisation proteins could be identified. Transpsoson mutagenesis was attempted using the EZTn*5 in vitro* transposition system but this proved to be unsuccessful. Two genes (gene 5 and 18, indicated in bold in Table 1), which we felt may be likely to have a possible role in L-carnitine utilisation were cloned in isolation to examine the phenotype further. Gene 5 is predicted to encode a sigma^{54}-dependent transcriptional regulator, while gene 18 is predicted to encode a major facilitator superfamily (MFS) transporter. The genes were annotated as *sdtR* and *mfsT*, for sigma-dependent transcriptional regulator and major facilitator superfamily transporter, respectively. MFS transporters can have broad substrate specificity and are sometimes involved in the uptake of osmoprotectant compounds, while we reasoned that the transcriptional regulator could be positively or negatively regulating host EPI300 genes, contributing the L-carnitine-associated phenotype.

Figure 5.

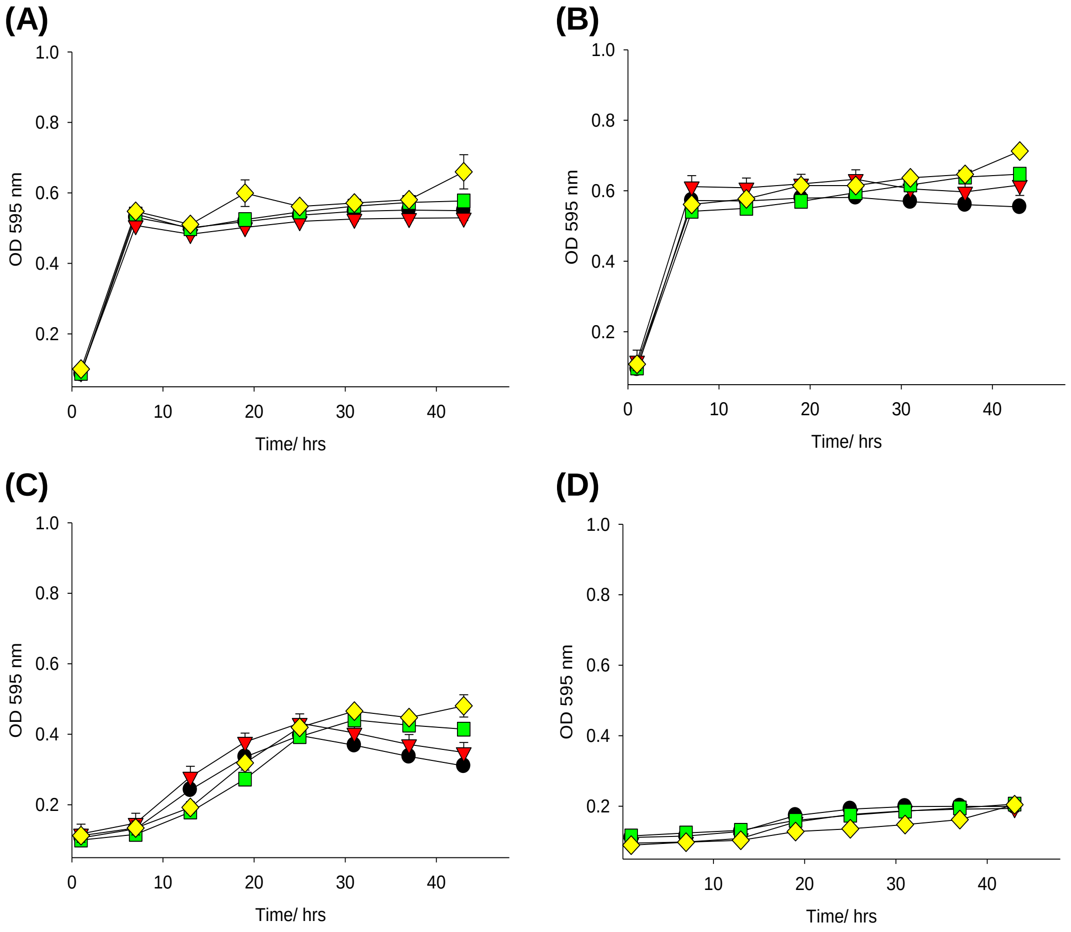


Figure 5. Growth of MKH13::pCI372 and MKH13::pCI372-*mfsT* in (A) M9 minimal media, (B) M9 minimal media + 2% NaCl, (C) M9 minimal media + 3% NaCl and (D) M9 minimal media + 4% NaCl.

Legend: *E. coli* MKH13::pCI372 (● black circle); MKH13::pCI372-*mfsT* (▼ red triangle); MKH13::pCI372 + 1mM L-carnitine (🟩 green square); MKH13::pCI372-*mfsT* + 1mM L-carnitine (◆, yellow diamond).

The *mfsT* gene did not confer salt tolerance to MKH13 and L-carnitine had no effect on growth under normal conditions (M9 minimal media) or in the presence of NaCl.

Figure 6.

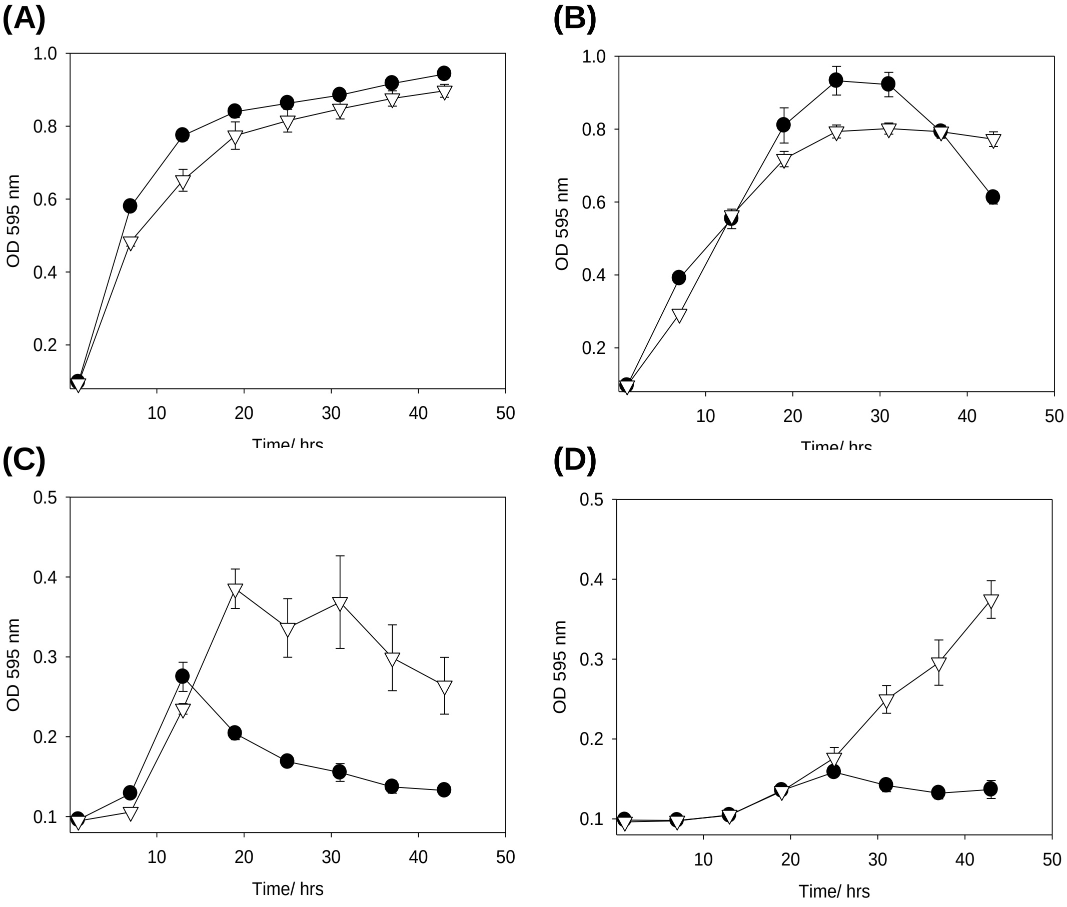


Figure 6. Growth in of *E. coli* MKH13::pCI372 (● black circle) and MKH13::pCI372-*sdtR* (▽ open triangle) in (A) LB broth and LB broth supplemented with (B) 2% NaCl, (C) 3% NaCl and (D) 4% NaCl.

Cloning and expression of *sdtR* in MKH13 conferred a significant salt tolerance phenotype when grown in at both 3% and 4% NaCl (Figure 6C and 6D, respectively) compared to MKH13::pCI372, while no difference in growth was observed in media lacking NaCl (Figure 6A).

Figure 7.

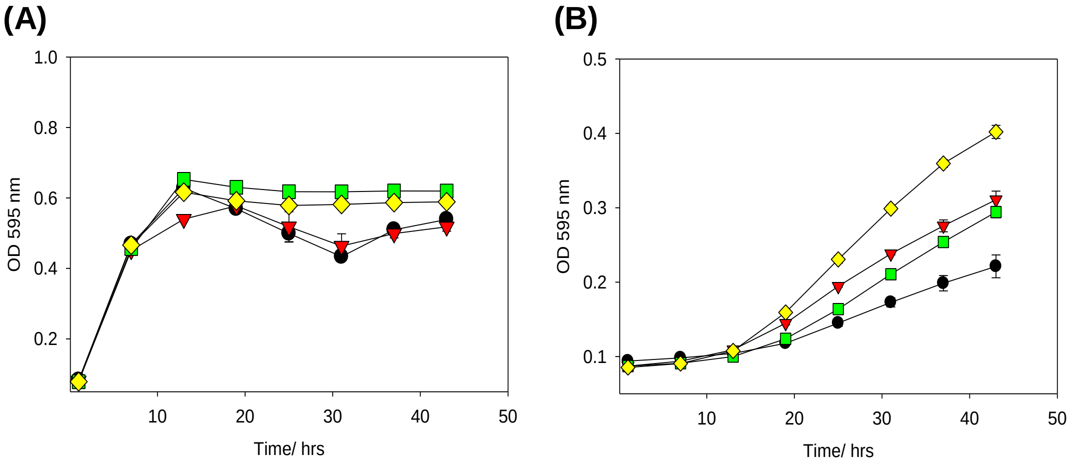


Figure 7. Growth of EPI300::pCI372 and EPI300::pCI372-*sdtR* in (A) M9 minimal media and (B) M9 minimal media + 6% NaCl.

Legend: *E. coli* EPI300::pCI372 (● black circle); EPI300::pCI372 + 1mM L-carnitine (▼ red triangle); EPI300::pCI372-*sdtR* (🟩 green square); EPI300::pCI372-*sdtR* + 1mM L-carnitine (◆, yellow diamond).

Cloning an expression of *sdtR* in EPI300 resulted in an increase in salt tolerance compare to wild-type EPI300 carrying an empty copy of the plasmid pCI372. Addition of 1mM L-carnitine increased the growth rate and final optical density of both strains, but its effect on the *sdtR*$^{+}$ strain was not significant relative to the EPI300::pCI372 control.

Figure 8.

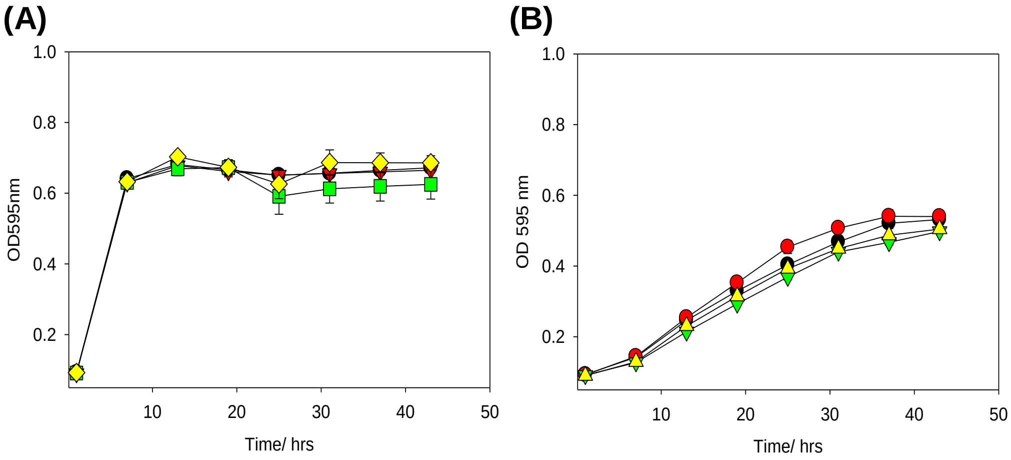


Figure 8. Growth of *E. coli* EPI300::pCC1FOS and SMG 9 in (A) M9 minimal media, (B) M9 minimal media + 5% NaCl.

Legend: *E. coli* EPI300::pCC1FOS (● black circle); SMG 9 (▼ red triangle); *E. coli* EPI300::pCC1FOS + 1mM L-carnitine (🟩 green square); SMG 9 + 1mM L-carnitine (◆, yellow diamond).

Further experiments revealed SMG 9 had lost the original phenotype seen in earlier experiments. Comparison of Figure 8A and 8B with Figure 4A and 4C, respectively illustrates this.

ACKNOWLEDGEMENTS

I've been lucky enough to work on a small aspect of what I feel is a fascinating area of science. This PhD has been an interesting, enjoyable (for the most part), challenging and sometimes stressful labour of love, but it has given me a greater appreciation of the microbial world and for the vast diversity and novelty that exists within it and indeed us. It would not however, have been possible without my supervisors giving me the opportunity carry out this research. So I'd like to sincerely thank Colin, Julian and Roy for your continuous support, encouragement and advice regarding experiments, writing papers and presentations and for your ever-present positivity and patience over the last few years. Your collective knowledge and enthusiasm for science is truly inspiring. Despite your significant respective flaws (Liverpool fan, Arsenal fan and Kerryman) I think things went pretty smoothly overall.

I would also like to take this opportunity to thank Cormac for his support and for our regular and insightful, but usually misguided and hopelessly optimistic discussions about ~~Fantasy Football~~ microbiology.

Undertaking a PhD requires a good support network, none more so than in the lab itself. I've worked with some fantastic people and made many, what I know will be, lifelong friends during my time here. You've all helped in many different ways and made my time here more enjoyable. So thanks to Ann, Aurelie, Avelino, Eileen, James, Jimmy, Joanne, Kiera, Mohammed, Neasa, Sinead, Tanya, Trev; my future "fantastico" wives Sarah and Ciara; all the crew in labs 4.25 and 335/337, especially Brian (Brad) Healy, my smoking buddy and go-to-guy for daily football conversations.

To G, Heather and Ruth you were all there from the start until almost the end, I couldn't have done it without you! G and Heather may have left and live on opposite sides of the world now (maybe I have that effect on people?), but I know when we meet again it will be just like old times. I seem to have had the opposite effect on Ruth as she now lives a bit too close for comfort and continues her effort to "feminise" our bachelor pad somewhat with soft *furnishings, candles and the like! Resistance is futile I fear. You were all a great support within the lab and your positivity and general craziness always brightened things up. Away from the lab we had many eventful nights out and I'll never forget our trips to Sweden, Holland, Galway (with an extended stop-over in Ballyhea that no one should ever have the misfortune to endure) and of course the trip to Edinburgh that started it all. Our level of attention and attendance at conference lectures could only increase after that.*

Thanks also to Susan Joyce for her help and advice and to Pat Casey for his expertise and lengthy Breaking Bad discussions. Thanks also to all the admin and technical staff in the Micro department and APC, especially Paddy, Dan, Maire, Margaret, Maurice, Liam, Carmel, Aine and Pat.

The majority of my postgrad days were spent living with a great group in "Lyndhurst". I think I've lived with more people in that house than I have worked with in the lab. We usually had one or two "characters" in residence, but the core members stayed largely the same; Liameen "The Chief" McCarthy, our landlord and leader, one of the most knowledgeable guys I know and also a spoofer supreme, but overall, one of life's all round good eggs! Your questions and antics were always (sometimes) a welcome distraction from the daily grind. Pity you never made that trip to Kilrush! Bridget you kept us all in line and your put-downs and story-telling are legendary, but most of all you were always up for a laugh and a bit devilment, mainly at poor Liameen's expense! Dave, many enjoyable hours were spent/wasted on Call of Duty (we never did finish that did we?), but thanks for probably saving my life and somehow getting me to A&E without crashing my car! Donny, I couldn't write this without giving you a special mention; a great man for a bit of slagging, lightning fast to remove the fuse from the plug of the kettle, a bad loser at chess, but always entertaining. Louise, it was good to have someone in touch with reality when living with a bunch of Corkonians consumed with their own sense of self-importance and if our respective science/nursing careers don't work out, I'm sure we could have *a future as a musical duet. Philip/ Phil/ P/ Pip/ Fox/ Dr.Phil, a man of many (similar) names, a lover of "exotic combinations" of foods, possessor of his own form of the English language, committed hurler, two-hour lunch and Home & Away enthusiast and a great friend. Always willing to oblige, offer support, encouragement and wisdom from past PhD experience and to provide temptation for a ramble and few sociables in Reardens or Bodega - much appreciated. I think I've mirrored your approach to education but taken a bit longer and observed the process, just to be safe... Hopefully I'll be on the Lyndhurst wall of fame soon too! Thanks to all Lyndhurst-associated friends also, Tadhg, Liam O'C, Joefus, Cillian, Catherine, Seamus, Power and of course the other Chief - Kieran O'C, I can't print your real nickname, but you know what it is.*

*I've been very fortunate to grow up with a great group of friends back home in Kilrush. There are too many to list but especially to Bean, Kelly, Alan, Domo, Darren, Mart, Cahill, Goggs, and Jimmy; there's nothing better than heading home for a few pints and ripping the p*ss out of each other, well mainly Bean, for a weekend (we all know he brings it on himself). Thank you all for your friendship over the years; you are like brothers to me.*

I'd like to thank all my extended family and family friends for their support since I was a young fella, Maureen, Mick, Tim, Belle, Pajoe, Mary, Sandra, Peggy and families and my favourite Godmother Mags and especially my uncle Senan.

Finally to my parents, my Mam Patsy and my late Father Noel, without your support; loving, moral and of course financial, I could never have made it this far or had the opportunity to receive such a fantastic education. The happiness and relief of finally finishing is tinged with a little sadness that Dad won't see this; he's probably the only person that would have actually read it cover to cover with genuine interest (apart from supervisors and examiners - hopefully!). You both sacrificed so much and always emphasised the importance of education, without ever putting *pressure on me. You always encouraged, reassured and supported me at every juncture and just told me to do my best. Mam you're one in a million and your unwavering support has got me through the most difficult times, from childhood to later childhood (i.e. my 30's, as I think you see it!).*

Cheers!

Eamonn.

"And that's the end of that chapter…!"

- Homer Simpson/Max Power, 1999

On Antonie van Leeuwenhoek's work in the 1670's:

…" But to find a whole world of microbes inside himself and in everyone around him, to discover that man is populated, was a cause for great excitement and for wonder and rejoicing. It was a discovery about ourselves, something spectacular; it turned a corner and opened a new vista."

- Theodor Rosebury, Life On Man, 1969.

THESIS PUBLICATIONS