The auxiliary replicons of *Butyrivibrio proteoclasticus*

A Thesis presented in fulfilment of the Doctorate of Philosophy degree at Massey University, Palmerston North, New Zealand.

Carl Yeoman
2009

Table of Contents

List of Figures

List of Tables

Abstract

Butyrivibrio proteoclasticus B316^T is the most recently described species of the *Butyrivibrio* / *Pseudobutyrivibrio* assemblage and now the first to have its genome sequenced. The genome of this organism was found to be spread across four replicons: a 3.5 Mb major chromosome and three additional large replicons: 186, 302 and 361 Kb in size. This thesis describes the sequencing, analysis, annotation and initial characterisation of all three *B. proteoclasticus* auxiliary replicons. Most significantly, these analyses revealed that the 302-Kb replicon is a second chromosome. This small chromosome, named BPc2, encodes essential systems for the uptake and/or biosynthesis of biotin and nicotinamide adenine mononucleotide, as well as the enzymes required for utilisation of fumarate as the terminal electron acceptor during anaerobic respiration, none of which are found on the main chromosome. In addition, BPc2 contains two complete rRNA operons, a large number of enzymes involved in the metabolism of carbohydrates, nitrogen and fatty acids. In contrast to BPc2, both megaplasmids appear largely cryptic, collectively encoding 421 genes not previously described in public databases. Nevertheless, only the 186-Kb, but not 361-Kb megaplasmid, could be cured from *Butyrivibrio proteoclasticus* B316^T. The largest megaplasmid has a copy number of 5, while all other replicons are present at a copy number of 1. %GC content and codon usage analyses strongly suggests that all three auxiliary replicons have co-resided with the major chromosome for a significant evolutionary period. Moreover, the replication machineries of these three replicons are conserved. Interestingly, a survey of a number of *Butyrivibrio* / *Pseudobutyrivibrio* species revealed that the megaplasmids are widespread in this assemblage, however these other large plasmids do not show concordance with their 16*S* rRNA phylogeny and appear distinct to those of *B. proteoclasticus* B316^T.

A microarray analysis of gene expression in a co-culture experiment between *B. proteoclasticus* and the important ruminal methanogen, *Methanobrevibacter ruminantium* M1, revealed a potentially mutualistic interspecies interaction. In this relationship *M. ruminantium* appears to provide *B. proteoclasticus* with glutamate, essential to the final step of NAD^+ biosynthesis, while *B. proteoclasticus* appears to provide *M. ruminantium* with formate, hydrogen and carbon dioxide, each important substrate for methanogenesis.

Acknowledgements

It is a privilege and an honour to acknowledge the many great people who have guided me throughout this project in one way or another. But more importantly have helped to shape me into the person that I am today.

To my supervisors Dr Graeme Attwood, Dr William (Bill) Kelly, and Dr Jasna Rakonjac, it is impossible to put into words the sincere gratitude I have for you all. I certainly wouldn't have lasted long without your influence and direction.

To the great people of AgResearch for the practical and financial support, without either this project would not have been possible.

To the kind people of Meat and Wool New Zealand, in particular Alan Frazer, who have funded, essentially, my life throughout not only this Ph.D. but the preceding MSc work.

To the Tertiary Education Commission for their financial support with this Ph.D.

To Dr Keith Joblin, I am eternally greatful to you and Robert Skipp who collectively played a role in bringing me into the AgResearch family and in doing so keeping me in the field of science.

To the Rumen microbiology laboratory, particularly, Dr Sinead Leahy for your help with the assembly, Dr Eric Altermann for your help with all things computational, Zhan-hou Kong and Sam Taylor for your help with the microarrays, Dong Lee for being my go to guy, Diana Pacheco for being my go to gal, Dr Adrian Cookson, Dr Ron Ronimus, Dr Peter Janssen, Graham Naylor, Dr Christina Moon, Carrie Sang, Rechelle Perry, Gemma Henderson, Catherine Tootill, and the ex-pats Dr Lucy Skillman, Dr Nicola Walker, Dr Mathew Nicholson, Dr Karen Olsen, Dr Hassan Husein, Paul Evans, Sam Noel and Nikki Kenters for your friendship.

To my parents Trish and Jim Yeoman thanks for all your support and encouragement.

To my beautiful wife Casey Norris-Yeoman thanks for sticking by me and supporting me while I pursue my dreams it has been a long and tough road but hopefully will be worth it in the long run. Love always!

Dedication

This thesis is dedicated to the most beautiful girls in the world my daughters Summer Ashlee Pamela Yeoman and Sienna Caitlyn Estelle Yeoman.

Whatever you need, whenever you need it, I'll always be there for you both!!

Abbreviations

AWGS	Alan Wilson Centre Genome Sequencing
BAC	Bacterial Artificial chromosome
BER	BLAST-extend-repraze
BSA	Bovine Serum Albumin
bp	Base pair
CDS	Coding sequence
Contigs	Contiguous sequences
DR	Direct repeat
dso	Double-stranded origin
FDR	False discovery rate
g	Gravity
GH	Glycosyl hydrolase
HMM	Hidden Markov-model
IR	Inverted repeat
IVR	Inverse repeat
Kb	Kilobase pair
Mb	Megabase pair
Mpf	Mating pair formation complex
nt	Nucleotide(s)
OD	Optical Density
ORF	Open reading frame
PARP	Poly(ADP-ribose) polymerase
PCB	Polychlorinated biphenyl
PCR	Polymerase chain reaction

PFGE	Pulsed-field gel-electrophoresis
pI	Isoelectric point
polIII	DNA polymerase III
RC	Rolling circle
RE	Restriction endonuclease
rRNA	Ribosomal ribonucleic acid
sso	Single-stranded origin
TA	Toxin-Antitoxin
tRNA	Transfer ribonucleic acid
qPCR	Quantitative real-time PCR

1 Literature review

1.1 Ruminant animals

Ruminant animals comprise 184 species from 74 genera within 6 families of the kingdom *Animalia.* Sheep, cattle, deer and goats are the most important to New Zealand‟s economy (Wensvoort, 2002). In a calendar year New Zealand commercially raises approximately 84 million ruminant animals spread across 45,729 farms (Pink, 2005, Bascand, 2007). In 2007 products derived from ruminant animals (meat, dairy and wool) were responsible for $16.15 billion of New Zealand‟s exports. This accounts for more than 48% of New Zealand‟s export market and approximately 17% of the country‟s gross domestic product (Gudgeon, 2007).

Aside from two exceptions in the-family *Tragulidae* (Chevrotains and Mouse Deer, which both have a three-chambered stomach) all ruminants have a four-chambered stomach (Clarke, 1968) consisting of (in order of passage) the reticulum, rumen, omasum, and the abomasum (Hungate, 1988). Ruminant animals obtain their food purely from herbaceous sources. This however is potentially problematic as the celluloses, hemicelluloses and lignins, which are major components of the primary and secondary plant cell walls, are indigestible by higher eukaryotes including ruminant animals (Chesson *et al.*, 1986, Hungate, 1988).

1.2 Cellulose, Hemicellulose and Lignin

Cellulose and lignin are the two most abundant organic polymers on earth. Cellulose comprises a linear polymer of D-glucose units connected by β(1→4)-glycosidic bonds. It forms the major component of both the primary and secondary cell walls of all plants (Klemm *et al.*, 2005). Lignin is a large hydrophobic polyphenolic, macromolecule derived from the dehydrogenative polymerisation of one or more of three phytochemicals: *p*-coumaryl alcohol, coniferyl alcohol, and/or sinapyl alcohol (Suhas *et al.*, 2007). Hemicellulose, a xylan-based plant polymer comprising 37-48% of a plants primary cell wall, is thought to form bridges between cellulose and lignin via ester linkages (Chesson *et al.*, 1986).

Cellulose, hemicellulose, and the non-digestible polyphenolic polymer lignin have, through necessity, been evolutionarily selected as components of plant structural cell walls due to their degradation-resistant structures (Hungate, 1988). Their structural strength ensures plant leaves and stems are supported in the air (or water), preventing collapse of water-conducting cells and ensuring access of chlorophyll cells to sunlight (Hungate, 1988). Ruminant animals, like all higher eukaryotes, lack the necessary enzymes to breakdown these structural carbohydrates and instead rely on a vast array of enzymes produced by a variety of rumen-inhabiting micro-organisms (Hungate, 1988). Therefore a mutualistic relationship has evolved between ruminants and rumen micro-organisms, in which the animal provides a eutrophic environment in the form of the reticulo-rumen, and the microbes provide most of the nutrients required by the animal through products of fermentation (Barcroft *et al.*, 1944, Elsden, 1946). However, the ruminant‟s microbiome typically degrades just 20-70% of plant structural carbohydrates (Varga & Kolver, 1997). This inefficiency of fibre degradation is believed to limit the potential productivity of ruminant animals (Brink & Fairbrother, 1994). Current estimates suggest that every percent increase in ruminant productivity would be worth approximately $160 million to the New Zealand economy (Frazer, 2005, Gudgeon, 2007). Consequently, the microbial breakdown of cellulose and hemicellulose has received a great deal of research attention (Krause *et al.*, 2003).

1.3 The reticulo-rumen

The rumen and reticulum act as a microbial fermentative chamber occurring prior to the acid stomach (pre-peptic). The reticulo-rumen (henceforth referred to as the rumen) is inhabited by bacteria, archaea, fungi, phage and protozoa (Bryant & Small, 1956, Chagan *et al.*, 1999, Trinci *et al.*, 1988, Naylor, 1998, Ambrozic *et al.*, 2001). The typical size of the rumen ranges from 2.4 to 12 litres for sheep and 90 to 120 litres for cattle (Habel, 1975).

The rumen is anoxic, and thus is largely inhabited by anaerobes and facultative anaerobes. Any O_2 introduced with forage is quickly metabolised by the facultative anaerobes and reduced by fermentation by-products such as sulphide (Hungate, 1988). The conditions within the rumen have evolved to maximise the growth of suitable anaerobic microbes. This is because unlike aerobic micro-organisms which break

down carbohydrates to mostly CO_2 and H_2O (products of little nutritional value), many anaerobic micro-organisms ferment carbohydrates to give volatile fatty acid (VFA) products which are used by the ruminant host (Hungate, 1988). These VFA products include acetic acid, propionic acid and butyric acid that are produced in a molar ratio of 63:21:16 respectively, although this can vary with diet (Hungate, 1966).

These and other acid products of fermentation, which in other situations would reach levels inhibitory to microbial growth, are neutralised by constant secretion of sodium bicarbonate-concentrated saliva from the animal‟s submaxillary and sublingual glands. This allows the pH of the rumen to be maintained at 5.7 - 7.3 (Wolin, 1981). The temperature is maintained around 39 °C although this may rise slightly following feeding, when fermentation is maximal (Dale *et al.*, 1954). Muscles in the walls of the rumen constantly undergo a series of coordinated contractions which serve to mix the rumen contents (Hungate, 1988). The combination of the aforementioned factors allows the rumen to achieve microbial concentrations far in excess of those commonly attainable by *in-vitro* batch culture techniques. These high microbial densities are necessary to provide the enzymatic activities required to break down the indigestible portions of forage at a rate useful to the ruminant (Hungate, 1988).

The VFAs produced during microbial fermentation are subsequently absorbed through the rumen wall. Large papillae lining the rumen wall greatly increase its surface area, enhancing VFA absorption. The VFA butyric acid is specifically important for the development of these papillae as well as the musculature of the rumen wall (Hungate, 1988). There are a number of bacterial species known to be capable of producing butyric acid through fermentation, such as *Butyrivibrio fibrisolvens*, *Clostridium kluyveri*, or *Eubacterium limosum* (Barker *et al.*, 1945, Muller *et al.*, 1981, Miller & Jenesel, 1979). The major bacterial genus responsible for butyric acid production in the rumen is *Butyrivibrio* (Miller & Jenesel, 1979).

1.4 *Butyrivibrio*

The bacterial genus *Butyrivibrio* was first described by Bryant and Small (1956) as anaerobic, butyric acid-producing, curved rods. These organisms are small, typically 0.4 – 0.6 μm by 2 – 5 μm and motile via a single, usually polar or sub-polar, monotrichous flagellum. They are commonly found singly or in short chains but it is not unusual for them to form long chains. Despite historical descriptions as Gram-negative (Bryant & Small, 1956), their cell walls contain derivatives of teichoic acid (Cheng & Costerton, 1977) and electron microscopy indicates that bacteria of this genus have a Gram-positive cell wall type (Beveridge, 1990, Cheng & Costerton, 1977). Their Gram-negative appearance by Gram staining is thought to be due to thinning of their cell walls (to 12 to 18 nm) as they reach stationary phase (Beveridge, 1990).

Butyrivibrio species are common in the rumen of cows, deer and sheep under a variety of diets (Bryant & Small, 1956), where they are involved in a number of ruminal functions of agricultural importance in addition to butyrate production (Miller & Jenesel, 1979). These include fibre degradation, protein breakdown, biohydrogenation of unsaturated fatty acids, and the production of microbial inhibitors (Hunter *et al.*, 1976, Blackburn & Hobson, 1962, Kalmokoff & Teather, 1997, Kepler *et al.*, 1966, Dehority & Scott, 1967, Polan *et al.*, 1964). Of particular importance to ruminant digestion, and therefore productivity, is their role in the breakdown of hemicellulose-lignin complexes (Morris & Van Gylswyk, 1980, Dehority & Scott, 1967).

Butyrivibrio species are metabolically versatile and are able to ferment a wide range of sugars (Stewart *et al.*, 1997) and cellodextrins (Russell, 1985). Some strains are able to break down cellulose (Shane *et al.*, 1969) although this ability is often lost during *in-vitro* culturing. Most isolates are amylolytic (Cotta, 1988) and are able to degrade xylan by producing xylanase (Hespell *et al.*, 1987, Sewell *et al.*, 1988) and acetyl xylan esterase enzymes (Hespell & O'Bryan-Shah, 1988, Lin & Thomson, 1991).

A number of genes encoding glycoside hydrolases (GH) have been identified in *Butyrivibrio* species including: Endocellulase (GH-family 5 and 9); β-Glucosidase (GH-family 3); Endoxylanase (GH-family 10 and 11); β-xylosidase (GH-family 43); and α-amylase (GH-family 13) enzymes. Several carbohydrate binding modules (CBM) have also been identified that are predicted to bind glycogen (CBM-family 48); xylan or chitin (CBM-family 2); and starch (CBM-family 26; Krause *et al*., 2003, Cantarel *et al*., 2008).

The *Butyrivibrio* genus encompasses over 60 strains that were originally confined to the species *Butyrivibrio fibrisolvens* based on their phenotypic and metabolic characteristics. However, phylogenetic analyses based on 16*S* ribosomal RNA (rRNA) gene sequences has divided the genus *Butyrivibrio* into six species (Fig. 1.1, (Kopecny *et al.*, 2003). These species include the rumen isolates *Butyrivibrio fibrisolvens, B. hungatei, B. proteoclasticus, Pseudobutyrivibrio xylanivorans, P. ruminis* and the human isolate *B. crossotus.* The families *B. fibrisolvens, B. crossotus, B. hungatei* as well as *B. proteoclasticus* all belong to the *Clostridium* sub-cluster XIVa (Fig. 1.1; Willems *et al*., 1996).

1.6 *Butyrivibrio proteoclasticus* B316^T

Butyrivibrio proteoclasticus B316^T was first isolated and described by Attwood *et al*. (1996). It was originally assigned to the genus *Clostridium* based on its similarity to *Clostridium aminophilum*, a member of the *Clostridium* sub-cluster XIVa. Further analysis has shown that it is more appropriately placed within the genus *Butyrivibrio* and consequently the organism was recently reclassified as *Butyrivibrio proteoclasticus* B316^T (Moon *et al.*, 2008). Within this genus its 16*S* rDNA sequence is most similar to, but distinct from, *B. hungatei* (Fig. 1.1).

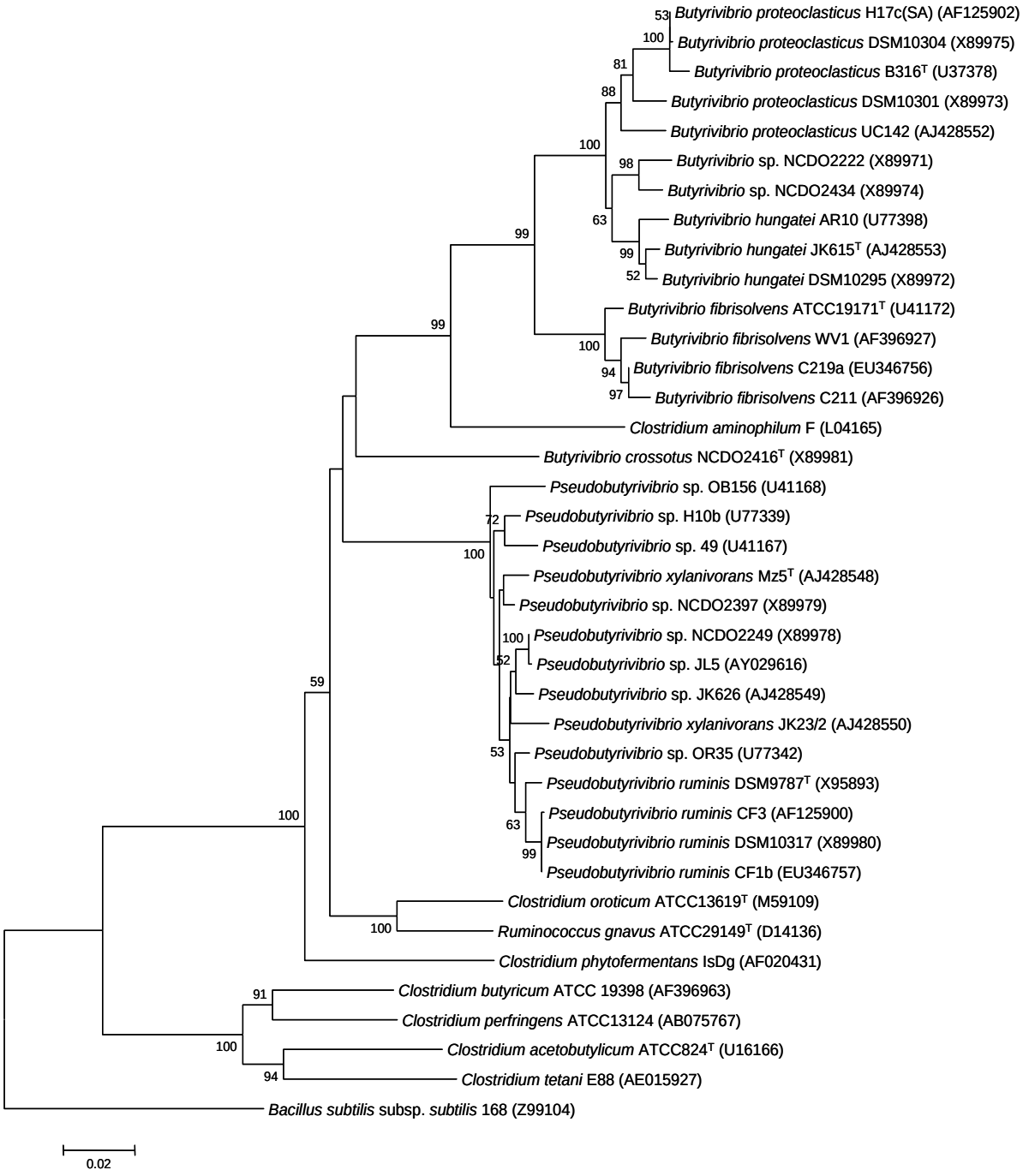


Figure 1.1 Phylogenetic placement of species within the *Butyrivibrio / Pseudobutyrivibrio* assemblage according to alignment of 16*S* rRNA gene sequences.
16*S* rRNA gene phylogeny using near full length sequences from clostridial cluster I and subcluster XIVa strains, with *Bacillus subtilis* subsp. *subtilis* as an outgroup. The phylogeny was inferred by Neighbor-Joining with distances calculated using the Kimura 2-parameter method, implemented in MEGA4 (Tamura *et al.*, 2007). There were a total of 1525 positions in the final dataset and all positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons. The rate variation among sites was modeled with a γ distribution (shape parameter = 1). The optimal tree with the sum of branch lengths = 0.95873951 is shown. Percentage bootstrap values were calculated from 10000 replicates, and are shown at nodes if greater than 50%. GenBank accession numbers are shown in brackets. The bar represents 0.02 nucleotide substitutions per site. Taken with permission from (Moon *et al.*, 2008).

B. proteoclasticus is present in the rumen at significant concentrations, which have been shown to range from 2.01 x 10^6/ml to 3.12 x 10^7/ml as estimated by competitive PCR (Reilly and Attwood 1998) or 2.2% to 9.4% of the total bacterial DNA within the rumen, as estimated by real time PCR (Paillard *et al.*, 2007). *B. proteoclasticus* cells are anaerobic, slightly curved rods, commonly found singly or in short chains, but it is not unusual for them to form long chains. They possess a single sub-terminal flagellum, but unlike other *Butyrivibrio* species they are not motile. They are ultrastructurally Gram-positive, although as with all *Butyrivibrio* species, they stain Gram-negative (Attwood *et al.*, 1996).

B. proteoclasticus does not produce indole, ammonia, catalase, lipase, or lecithinase, or reduce nitrate, however it is highly proteolytic and xylanolytic. Recently, *B. proteoclasticus* has been shown to have a role in biohydrogenation, converting linoleic acid to steric acid (Wallace *et al.*, 2006). The %G+C content of *B. proteoclasticus* was originally reported at 28%, however reanalysis of its %G+C content and analysis of its genomic DNA sequence has shown that it is 40.02% (Attwood *et al*., 1996, Moon *et al*., 2008).

1.7 Genome sequencing

The first complete sequence of a genome was that of bacteriophage ΦX174 completed and published in 1977 (Sanger *et al.*, 1977a). Eighteen years later the first bacterial genome sequence was published; that of *Haemophilus influenzae*, an opportunistic pathogen implicated in a number of human diseases (Fleischmann *et al.*, 1995).

To date, most genomes have been sequenced using the Chain-termination (or Sanger) method (Sanger & Coulson, 1975, Sanger *et al.*, 1977b) in combination with a „shotgun" cloning approach (Anderson, 1981). Collectively, these processes are referred to as shotgun sequencing. However during the course of this thesis a novel method of DNA sequencing, array-based pyrosequencing, has been developed. Pyrosequencing is faster and less expensive than the Chain-termination method and also forgoes the need for cloning (Margulies *et al.*, 2005). Pyrosequencing was first described in 1996 (Ronaghi *et al.*, 1996), and the development of the array-based

method has led to its emergence as a rapid platform for genome-scale DNA sequencing projects (Margulies *et al.*, 2005). This technology is becoming the predominant choice of genome (and metagenome) sequencing efforts. Array based pyrosequencing is currently limited by small sequence reads (200-400 bp) that may cause downstream problems in the assembly of highly repetitive genomes.

In 1998 Brent Ewing, Phil Green and colleagues described Phred, a base calling program capable of determining the reliability of each base-call by analysis of the corresponding trace file (Ewing & Green, 1998, Ewing *et al.*, 1998). It has since been the corner stone of evaluating sequence integrity. Ewing and colleagues identified four parameters of trace files that best indicated an incorrect base call: 1) inconsistent peak spacing 2+3) a high ratio of the largest uncalled peak to the largest called peak using both a window of 7 and a window of 3 surrounding peaks, and 4) the peak resolution (the number of bases between the scrutinised base and the nearest unresolved base) (Ewing & Green, 1998). The resulting analysis of these factors is fed into a complicated algorithm that assigns a quality score to each base. This quality score (q) relates to the error probability (p) by the following equation:

$$q = -10 \times \log_{10}(p)$$

Where for example a Phred quality score of 40 indicates a base has a 1 in 10,000 possibility of being incorrectly called (Ewing & Green, 1998).

Modern genome sequencing consists of two distinct phases the first being shotgun sequencing and assembly. This results in a collection of contiguous sequence reads (contigs), each being separated by a sequencing or physical gap. A sequencing gap refers to a gap that is captured but not sequenced. This results from the cloning of large fragments of DNA that exceed the maximal length of quality sequence obtainable through the sequencing of each of the clone‟s ends. Physical gaps exist where a stretch of DNA has not been cloned.

The second phase of modern genome sequencing is known as finishing, this is the process where the aforementioned gaps are closed and any low quality, typically less than Phred 40, portions of the sequence are enhanced.

The sequencing of bacterial genomes and subsequent analysis provides a wealth of information about an organism, such as its evolutionary origins (Martinez-Vaz *et al.*, 2005), lifestyle strategies (Backhed *et al.*, 2005) and metabolic capabilities (Arp *et al.*, 2007). It allows potential drug targets, if required, to be identified (Dutta *et al.*, 2006). Additionally whole genome comparisons can reveal information on microbial diversity (Venter *et al.*, 2004), the origins and limits of life on earth (Koonin, 2003) or the potential for life in other planetary systems (Cavicchioli, 2002).

This information generated may benefit, amongst others, biotechnology companies (Podar & Reysenbach, 2006), biomedical (Brudno *et al.*, 2007), agricultural (Yu *et al.*, 2005), and pharmaceutical researchers (Roses *et al.*, 2007) and provide novel drugs (Duenas-Gonzalez *et al.*, 2008), energy generating-, system enhancing- and bioremediation- technologies (Durre, 2007, Butler *et al.*, 2007) .

At the time of writing this literature review around 700 whole genomes have been sequenced, while another approximately 1,800 genomes are in progress worldwide (genomes online database, GOLD website). The majority (1574 as of May 2007) of genome sequencing projects are of bacteria. Most of these (49%) are of biomedical interest, with just 6% of bacterial genome projects of agricultural interest funded worldwide (Kyrpides, 1999, Bernal *et al.*, 2001, Liolios *et al.*, 2006).

Bacterial genomes sequenced to date range in size from the 160 kb genome of the psyllid endosymbiont *Carsonella ruddii* (Nakabachi *et al.*, 2006), to the 13 Mb genome of the myxobacterium *Sorangium cellulosum* (Schneiker *et al.*, 2007). Bacterial genomes may be entirely contained within a single chromosome (Schneiker *et al.*, 2007) or spread across multiple chromosomes (Chain *et al.*, 2006) and/or plasmids (Brom *et al.*, 2000, Casjens *et al.*, 2000).

1.8 Gene prediction, annotation and analysis

The purpose of genome sequencing is to identify genes and functional ribonucleic acids (RNAs) and predict their function through annotation. In its most basic sense, a gene or an open reading frame (ORF) constitutes a length of DNA commencing with a start codon (typically ATG although occasionally this can be substituted for TTG or GTG; Clark & Marcker, 1966) followed by a variable number of nucleotide triplets (known as codons) each encoding a single amino acid, finishing with a stop codon (TAA, TAG or TGA; Kohli & Grosjean, 1981). Prokaryotic genomes are far less complex than those of eukaryotes and therefore gene identification and annotation is comparatively easier. Intergenic regions account for very little of the total genome (typically 10 - 20%; Salzberger *et al*. 1998). Prokaryotic open reading frames (ORFs) are intronless and short, typically averaging around 1 Kb (Ochman & Davalos, 2006). This reduced complexity gives some predictability allowing bioinformatics tools to identify key features of the genome (such as ORFs) with more reliable accuracy.

However not every stretch of DNA preceded by a start codon and finishing with a stop codon is actually transcribed. Therefore an ORF without any evidence of transcription, function or significant similarity to any gene previously sequenced is described as encoding a hypothetical protein. The proportion of hypothetical proteins varies considerably between bacterial genomes ranging from 0.2% (Shigenobu *et al.*, 2000) to 53.4% (Casjens *et al.,* 2000; average 15.3%; Fukuchi & Nishikawa, 2004). Organisms with the fewest hypothetical proteins tend to be those with reduced genomes such as endosymbionts or the parasitic *Mycoplasma genitalium*. While the number of hypothetical proteins tends to increase with the phylogenetic distance between the sequenced organism and its closest sequenced relative (Fukuchi & Nishikawa, 2004).

Currently the best known and most commonly used software packages for predicting ORFs are GeneMark (Borodovsky & McIninch, 1993, Besemer & Borodovsky, 2005) and GLIMMER (Gene locator and interpolated Markov modeller; Delcher *et al.*, 1999). These programs detect more than 91% and 97% of ORFs from prokaryotic genomes, respectively, without manual curation (Salzberg *et al.*, 1998). GeneMark and GLIMMER first identify all potential ORFs larger in size than a specified threshold, typically set at 150 bp. ORFs smaller than this have a higher probability of

occurring by chance (Kellis *et al.*, 2003). This probability increases with %G+C content (Skovgaard *et al.*, 2001). These programs then determine the probability that each nucleotide of an ORF, and collectively the entire ORF, is in a coding (or non-coding) region using Markov models built by a training set of known genes from the organism (Salzberg *et al.*, 1998). This ability to differentiate coding from non-coding regions is based on the finding that genomes contain genome-specific signatures (Karlin *et al.*, 1997). These signatures vary between coding and non-coding regions (Sandberg *et al.*, 2003). Such signatures include di-nucleotide, tri-nucleotide and tetra-nucleotide frequencies (Karlin *et al.*, 1997), %G+C content (Chargaff, 1951), codon usage (Grantham *et al.*, 1980a, Grantham *et al.*, 1980b) and amino acid usage bias (Sueoka, 1961). GLIMMER and GeneMark differ in the length of the Markov-chains. GLIMMER does not have a fixed length (Delcher *et al.*, 1999), whereas GeneMark uses a 5^{th} order model (Borodovsky & McIninch, 1993). GeneMark additionally refines a retained ORF's start codon by searching for ribosome-binding sequences upstream of potential start codons (Lukashin & Borodovsky, 1998).

Ribosome binding sequences (RBS) are typically centered approximately 8 - 13 nucleotides upstream of a true start codon (Shine & Dalgarno, 1975). However, in some genes, typically those transcribed from alternative start codons (GTG or TTG), there appears to be a larger gap between the RBS and the start codon (Kozak, 1983). Ribosome binding sites typically have a sequence such as AGGA or GAGG, and is complimentary to the 3' end of the 16*S* rRNA chain (Shine & Dalgarno, 1975).

Several comparative analyses of genome sequences to date have identified a theoretical minimal set of between 169 and 206 genes essential for cellular life (Gil *et al.*, 2004, Koonin, 2000). While over-annotation, particularly of small ORFs, makes it difficult to accurately determine an average gene size (Skovgaard *et al*., 2001, Fukuchi & Nishikawa, 2004, Huynen & Snel, 2000) it is generally considered to be approximately 1 Kb (Ochman & Davalos, 2006). However, a single gene can exceed 20 Kb (Reva & Tummler, 2008). The largest prokaryotic genes described to date were found in *Chlorobium chlorochromatii* CaD3 and encode proteins of 36, 806 and 20, 647 amino acids (Reva & Tummler, 2008).

Genome analysis has shown the presence of a number of inactivated genes also, known as pseudogenes (Jacq *et al.*, 1977, Vanin, 1985). Pseudogenes arise through the accumulation of mutations that disrupt and ultimately degrade their functional predecessors. These are commonly described in genome analyses. It is believed that all traces of pseudogenes are likely to completely erode over time. This erosion is due to prokaryotes having a mutational bias favouring deletions over insertions (Ochman & Davalos, 2006).

Following the identification and refinement of ORFs, they can be analysed and ascribed a putative function. A number of bioinformatics tools exist to facilitate this analysis. The most popular include Basic Local Alignment Search Tool (BLAST; Altschul *et al*., 1997), Clusters of Orthologous Groups (COG; Tatusov *et al*., 1997), SignalP (Bendtsen *et al.*, 2004), LipoP (Juncker *et al.*, 2003), Transmembrane HMM (TmHMM; Krogh *et al*., 2001) and Interpro (Apweiler *et al.*, 2001). BLAST compares the similarity of a nucleotide or amino acid sequence to a sequence database (typically GenBank; Benson *et al*. 2008) using a heuristic alignment approach (Altschul *et al.*, 1997). Targeting of the encoded proteins to secretory systems can be predicted by SignalP and LipoP. SignalP uses neural network and hidden Markov model (HMM) algorithms to predict the classical signal peptidase I cleavage sites, and LipoP uses HMMs to predict signal peptidase II cleavage sites, characteristic of lipoproteins. The transmembrane topology can also be predicted by HMM, using TmHMM (Krogh *et al.*, 2001). Specific targeting signatures, such as carboxy-terminal cell wall anchoring in Gram-positive bacteria, via LPXTG-like motifs (Navarre & Schneewind, 1994), recognised by the protein Sortase (Ton-That *et al.*, 1999), can be predicted by specific HMMs available through PFam or TIGRFam libraries (Boekhorst *et al.*, 2005). Interpro integrates the major hidden Markov model searches: PFam (Sonnhammer *et al.*, 1997) and TIGRFam (Haft *et al.*, 2001). It also includes other protein signature databases, including Uniprot (Apweiler *et al.*, 2004), Prosite (Bairoch, 1991), ProDom (Corpet *et al.*, 1998), Smart (Schultz *et al.*, 2000, Ponting *et al.*, 1999, Schultz *et al.*, 1998), Panther (Thomas *et al.*, 2003), PIRSF (Wu *et al.*, 2004), Superfamily (Gough *et al.*, 2001), SCOP (Murzin *et al.*, 1995), CATH (Pearl *et al.*, 2003), Swiss Model, MOD Base (Peitsch, 1996), MSD (Golovin *et al.*, 2004), Gene 3D and Sprint. Collectively Interpro allows the identification of structural and

functional domains using primary, secondary and tertiary structural prediction and HMM analyses.

Following such analysis each ORF is ascribed a putative function. This should be descriptive but conservative and, where possible, consistent with the type species (this being *Bacillus subtilis* for Gram-positive organisms). Several publications have pointed out the confusion that is often caused by inconsistencies in the assignment of gene names (Wang *et al.*, 2005, Wyman *et al.*, 2004, Mitchell *et al.*, 2003). Konforti (2007) recently suggested that a gene name often reflects the laboratory where the genome was annotated rather than its function or even the organism it is derived from. Various methods have arisen to attempt to circumvent such inconsistencies, such as gene ontologies (Smith *et al.*, 2005). Gene Ontologies are a finite list of descriptors that define a gene by its molecular function(s), the biological process(es) where these functions take place and what cellular componentry it resides in, or belongs to. Collectively all above bioinformatics analyses can be managed on a genomic scale using an interface such as MANATEE (TIGR, 2001) or GAMOLA (Altermann & Klaenhammer, 2003).

1.9 Plasmids

Plasmids are classically defined as covalently-closed, circular, double stranded DNA molecules. However, more recently plasmids have been described existing as linear entities (Hirochika *et al.*, 1984). Plasmids exist, and are replicated, independently of the bacterial chromosome. All plasmids contain a replication origin (*oriR*) allowing them to be replicated (reviewed in Section 5.3). Many also contain genes that are beneficial to their host by conferring drug resistance (R plasmids) (Allignet *et al.*, 1998), bacteriocin production (Gamon *et al.*, 1999), enhanced virulence (Crespi *et al.*, 1992), enhanced symbiotic- (Innes *et al.*, 1988) or metabolic capabilities (Igloi & Brandsch, 2003).

Like other accessory genetic elements of bacterial genomes (e.g. phage, transposons or genomic islands) plasmids are capable of being transmitted, not only vertically during bacterial cell division but also horizontally by conjugation (reviewed in Section 1.13), transformation, or transduction (Garzon *et al.*, 1995, Andrup, 1998, Hanahan *et al.*, 1991). Various mobilisation methods, including transposition, can

move specific genes between the host chromosome and plasmids, as well as transfer groups of previously non-contiguous genes from different replication units into a single replicon (Amabile-Cuevas & Chicurel, 1992, Cohen, 1976). During this process additional genes may be co-transferred in a phenomenon known as "DNA hitchhiking". Therefore plasmids are typically thought to provide a mobile pool of genetic information.

Several plasmids, all with no apparent function (cryptic plasmids), have previously been described from *Butyrivibrio* species. These include pRJf1 and pRJf2 from *B. fibrisolvens* OB157 (now designated *Pseudobutyrivibrio ruminis* OB157), pBF1 from *B. fibrisolvens* AR10 (now designated *B. hungatei* AR10) and pOM1 from *B. fibrisolvens* Bu49 (now designated *P. xylanivorans* Bu49) (Hefford *et al.*, 1997, Kobayashi *et al.*, 1995, Ware *et al.*, 1992, Hefford *et al.*, 1993).

Plasmids come in a variety of sizes. Those described to date range in size from 846 bp (Nesbo *et al.*, 2006) to over 2 Mb (Salanoubat *et al.*, 2002). Plasmids greater than 100 Kb in size are termed "megaplasmids" (Bartosik *et al.*, 2002).

1.10 Megaplasmids

The first megaplasmid was described in *Agrobacterium tumefaciens* by Van Larebeke *et al*. in 1974. The term megaplasmid was not, however, coined until 1981 when Rosenberg *et al* (Rosenberg *et al.*, 1981) described pSYMA in *Rhizobium meliloti* (now known as *Sinorhizobium meliloti*). Although the prefix „mega" was likely intended to illustrate the fact that pSYMA exceeded 1 Mb pair in size, the term was adopted by many describing large plasmids (Amils *et al.*, 1998, Niazi *et al.*, 2001). It was eventually defined as a term to describe a plasmid greater than 100 Kb in size (Bartosik *et al.*, 2002).

The identification and characterisation of megaplasmids has previously been limited by technical challenges associated with working with large plasmids. Megaplasmids do not separate readily from chromosomal DNA in procedures designed to purify smaller plasmids (Grindley *et al.*, 1973, Currier & Nester, 1976, Sobral *et al.*, 1991). They are too large to resolve by conventional agarose gel-electrophoresis (Schwartz & Cantor, 1984, Mathew *et al.*, 1988), and are easily damaged by standard DNA

extraction procedures. However, techniques such as the preparation of DNA in agarose plugs to protect it from mechanical shear (Goering & Winters, 1992) and pulsed-field gel-electrophoresis (PFGE), which allows the resolution of large DNA fragments (Schwartz & Cantor, 1984), have emerged. These have led to optimised methods for detecting and sizing large plasmids (Barton *et al.*, 1995). Currently 209 megaplasmids have been identified and sequenced, as listed by GenBank in more than 30 genera of bacteria (NCBI, 2008, Woodsmall & Benson, 1993). These include an unculturable mycoplasma-like organism (Neimark & Kirkpatrick, 1993), several halophilic archaea (Ng *et al.*, 1998, Anton *et al.*, 1995, Ferrer *et al.*, 1996), and a radiation-resistant bacterium (White *et al.*, 1999).

There does not appear to be any phylogenetic correlation associated with megaplasmid distribution, nor does there appear to be any correlation with the total genome size of the host organism (Egan *et al.*, 2005). The majority of megaplasmids identified and sequenced to date derive from γ- and α-Proteobacterial classes (29% and 32% respectively) but this is likely to be biased by the disproportionately large number of Proteobacterial genomes (52% of all bacterial genomes) that have been sequenced to date (Liolios *et al.*, 2006). The vast majority of megaplasmids identified thus far appear to be hosted by bacterial species that interact with a host (Egan *et al.*, 2005).

Most bacterial megaplasmids have been found in Gram-negative organisms. At the commencement of this project, only ten (of 49 total) sequenced megaplasmids were hosted by Gram-positive organisms. These organisms included *Bacillus sphaericus* (Liu & Fan, 1991), *Rhodococcus erythropolis* (Stecker *et al.*, 2003), *Nocardia opaca* (Kalkus *et al.*, 1990), *Bacillus anthracis* (Okinaka *et al.*, 1999), *Streptomyces rochei* (Mochizuki *et al.*, 2003), *Streptomyces coelicolor* (Bentley *et al.*, 2004) *Gordonia westfalica* (Broker *et al.*, 2004), *Pediococcus acidilactici (Halami et al., 2000), Deinococcus radiodurans* (White *et al.*, 1999) and *Clostridium acetobutylicum* (Nolling *et al.*, 2001). The sequences of a further 24 megaplasmids from Gram-positive bacteria were deposited in during the tenure of this thesis, mostly from the family Actinobacteria. This included pRHL1, from *Rhodococcus* sp. RHA1, which had been described previously (Shimizu *et al.*, 2001) and is currently the largest Gram-positive megaplasmid described at 1.1 Mb.

Overall the megaplasmids identified to date range in size from a little over 100 kb to 2.1 Mb (Salanoubat et al., 2002, NCBI, 2008, Woodsmall & Benson, 1993). Many provide their host organism with distinctive and significant bacterial traits, such as root nodulation (Rosenberg *et al.*, 1981), heavy metal resistance (Taghavi *et al.*, 1997, Ng *et al*., 1998), rubber degradation (Broker *et al.*, 2004), sugar utilisation (Halami *et al.*, 2000), bacteriocin production (Halami *et al.*, 2000), and chemolithoautotrophic utilization of gases, for example utilisation of CO (carboxidotrophy), H_2 (hydrogenotrophy) and CO_2 under aerobic conditions (Fuhrmann *et al.*, 2003).

Gloria del Solar and Montero de-Espinosa (2000) proposed that plasmids are likely to confer a slight metabolic burden to the host. Shuvaev and Brilkov (Shuvaev & Brilkov, 2007) have further proposed that the acquisition of a plasmid would slow the cell cycle duration, due to the increased competition for nascent *dnaA* mRNAs at the ribosome. As the amount of DnaA protein per chromosomal origin (*oriC*), known as the initiating mass, is a major factor in determining cell cycle duration (Donachie, 1968, Boye & Nordstrom, 2003), and the amount of DnaA protein present is dependent on the completion of *dnaA* mRNA translation at the ribosome, Del Solar and Espinosa (2000) propose that the time taken to reach this initiation mass correlates with the time taken for *dnaA* to be translated to DnaA at the ribosome. Considering the inevitable competition between *dnaA* mRNAs and other mRNAs at the ribosome, plasmid-encoded mRNAs would further dilute *dnaA* mRNA and consequently increase the time taken to reach the initiating mass. Further, this effect would increase proportionally with the plasmid size and copy number due to the net increase of transcribed mRNAs. Investigations of the growth rates of *E. coli* containing various R plasmids (Zund & Lebek, 1980) found that in most cases *E. coli* strains possessing more than one plasmid, a plasmid greater than 80 Kb, or a plasmid with a copy number >20 resulted in a 15 to 50% increase in generation time, although the increase in generation time was not proportional to the plasmid size. Several studies have since shown that, if a plasmid is maintained by a host for sufficient evolutionary time, both the plasmid and the host organism will evolve to minimise or eliminate this cost (Heuer *et al.*, 2007, Dahlberg & Chao, 2003). This reduction in cost appears to be accelerated if the replicon is conjugatively transferable (Heuer *et al.*, 2007). Southamer and Kooijman (1993) used mathematical modelling to propose

that there is a benefit to an organism in spreading its DNA over several independently replicating units. They suggest that by doing so, a high surface to volume ratio may be maintained, positively influencing the population growth rate. The uptake of nutrients is proportional to the surface area of the cell, while maintenance costs are proportional to the volume, and growth of the individual cell continues until DNA replication of the genome is completed. Therefore, cells with a single large chromosome should be bigger at the moment of cell division resulting in a less favourable surface to volume ratio, negatively influencing the population growth rate. This would further help to explain why in many bacterial species disused genes are easily lost. However, large plasmids provide a means to retain these less used genes, which are beneficial only under certain circumstances, by allowing their storage in low-copy extra-chromosomal replicons. In support of this hypothesis is the common finding that megaplasmids, much like conventional plasmids, encode genes which are required occasionally, rather than consistently (Moreno, 1998) and the selective pressures for retaining these genes appear to be much more relaxed (Dryselius *et al.*, 2007). Megaplasmids may also allow an organism to differentially regulate large sets of genes to fit the requirements of a specific environment by means such as variable copy number (Dryselius *et al.*, 2007).

Megaplasmids are all thought to exist at a low copy number due the steric limitations of the cell. Claesson *et al*. have shown using fluorescent real time PCR that pMP118 of *Lactobacillus salivarius* exists at a copy number of 4.7 +/- 0.6 per chromosome (Claesson *et al.*, 2006). Further, Titok *et al*. (Titok *et al.*, 2003) have shown that the use of a replication origin from a large (~90 kb) plasmid of *Bacillus subtilis* in a vector gives the resulting vector a copy number of approximately 6 units per chromosome. Megaplasmids have been reported in a supercoiled circular (Lopezgarcia *et al.*, 1994) or linear form (Saeki *et al.*, 1999, Kalkus *et al*., 1990, Krum & Ensign, 2001). Most descriptions of linear megaplasmids are based on their ability to run into a pulsed-field gel (which without enzymatic linearization precludes the entry of relaxed circular DNA). Megaplasmids, like smaller plasmids, are generally referred to as accessory elements implying that an organism may exist in their absence. Yet, a number of "incurable" megaplasmids exist (Ng *et al*., 1998, Bartosik *et al*., 2002) raising the question "when is a megaplasmid a miniature, minor or secondary chromosome?".

1.11 Miniature, Minor or Secondary Chromosomes

There is some debate in the literature as to what can be defined as a chromosome. Schwartz and Friedrich (2001) used the size of the plasmid as the sole determinant to differentiate between a plasmid and a secondary-chromosome, while Bartosik *et al*. (2002) have proposed that a megaplasmid becomes a secondary-chromosome when it contains genes essential for growth in minimal media. However, to designate a gene as 'essential' is not always straightforward (Egan *et al.*, 2005). Ng *et al* (1998) similarly have proposed that a secondary-chromosome contains essential genes, so that the host can not be cured of it. Jumas-Bilak *et al*. (1998) argued that regardless of dispensability, a chromosome must possess characteristic chromosomal markers such as ribosomal RNA genes (*rrn*) or heat shock proteins (*hsp*). However, Krawiec and Riley (1990) have further argued that *rrn* or *hsp* genes are often (if not always) not unique to the megaplasmid possessing them, and duplicate copies are found on the chromosome. While the number of *rrn* operons, as a general rule, may appear to have an inverse correlation with generation time (Cole & Saint Girons, 1994), it appears possible that duplicates may be lost. This is supported by analysis of organisms that now possess only a single *rrn* operon, for example *Bradyrhizobium japonicum* (Kundig *et al.*, 1995). Therefore, until it can be shown that the *rrn* or *hsp* genes of a megaplasmid are essential for the normal growth of the organism, defining a megaplasmid as a secondary-chromosome may be incorrect. More recently Ochman provided a bioinformatically measurable definition of a chromosome, proposing that a chromosome should possess a unique copy of one of the genes described in the bacterial minimal gene set (Koonin, 2000). This is based on the premise that a chromosome represents "the ancestral genetic material" (Ochman, 2002). To summarise, it seems a plasmid (megaplasmid or otherwise) should be defined as consisting of a particular sample of the bacterial genome that is required occasionally, rather than continually. A chromosome should, however, possess unique genes required for housekeeping and are therefore indispensable. This definition would apply irrespective of the replicon size, thus uncritical acceptance that even a smaller auxiliary replicon is a plasmid may be incorrect (Kolsto, 1997, Egan *et al*., 2005).

At present, 30 bacterial species have been described as possessing two (or three) chromosomes, based on one or more of the aforementioned definitions. These are

largely of the β-Proteobacterial class (56%) particularly of the genus *Burkholderia* (33%). Aside from *Deinococcus radiodurans* R1, all bacterial species currently described as possessing multiple chromosomes are Gram-negative.

Secondary chromosomes that have been described range in size from 300 Kb (Bulach *et al.*, 2006) to 3.6 Mb (Copeland *et al.*, 2007), but are always smaller than the main chromosome (Ochman, 2002). All currently described secondary chromosomes have a very similar %G+C to that of the main chromosome, implying a long evolutionary co-existence within the same host (Egan *et al.*, 2005). The majority of genes deemed 'essential' are typically found on the main or major chromosome. In contrast, a larger proportion of genes with unknown function are found on the secondary chromosome (Egan *et al.*, 2005). Dryselius *et al*. have hypothesized that the major chromosme over time has become evolutionarily self-stabilising and the relative inflexibility of this replicon is the reason why most essential genes are retained there and not translocated to a secondary chromosome.

1.12 Plasmid Replication

Plasmids must replicate to allow their vertical transmission; failing this they would simply be lost from the host during cell replication and division. Therefore all plasmids contain a distinct locus (typically 500 bp-3 kb), known as the origin of replication (o*riR*) at which plasmid replication is initiated. The o*riR* not only allows initiation of plasmid replication but it also regulates the rate of initiation, allowing the plasmid to be maintained at a defined copy number per cell (Abeles *et al.*, 1995). Plasmid copy numbers range from one to over a thousand per cell (Schroeter *et al.*, 1988). In all plasmids studied to date the copy number is controlled by either an antisense RNA transcript (Osborn *et al.*, 2000) or by the binding of replication (Rep) proteins to regulatory sequences near the initiation sites (iterons; Abeles *et al.,* 1995).

In the RNA-binding mechanism, an anti-sense RNA targets an overlapping and complementary (sense) RNA, transcribed from the opposite strand. This target RNA would otherwise be used as a primer for DNA replication, or as a messenger for a Rep protein (De Wilde *et al.*, 1978, Muesing *et al.*, 1981). The base pairing is formed between complementary unpaired loops which are formed in both RNAs by

secondary folding (Brady *et al.*, 1983). This type of regulation, known as an "inhibitor-target mechanism"(Novick, 1987), is employed by many smaller plasmids which use basic replicons as well as several large conjugative plasmids of the incompatibility group F (IncF).

In the iteron-binding mechanism a trans-acting replication initiation/control protein (Rep) binds to a series of typically 18-22 bp direct repeats (iterons) situated near the *oriR* (Hitoshi *et al.*, 1999). Rep proteins use ATP to separate the double stranded DNA, then use their DNA helicase domains (or a host encoded DNA helicase during rolling circle replication; considered below) to partially unwind the plasmid (Kornberg *et al.*, 1978, Arai *et al.*, 1981). The partially unwound (relaxed) state of the double helix at the plasmid *oriR* allows the replication machinery to assemble and commence transcription. However, when Rep is in excess, it forms dimers. These dimers bind to inverse repeats in the *rep* gene promoter sequence preventing its own transcription (Diaz-Lopez *et al.*, 2003).

The trans-acting feedback regulation of the o*riR* by either the inhibitor-target or iteron-binding mechanism means that two plasmids sharing common elements involved in their control and partitioning cannot coexist and be propagated stably in the same host. This feature has lead to the grouping of plasmids based upon their incompatibility (Scaife & Gross, 1962). Yet, there are exceptions to this rule; the plasmids may contain multiple replication origins. This feature is common to the IncF group of large plasmids (Bergquist *et al.*, 1986). Plasmids regulated by the inhibitor-target mechanism may also become compatible by nucleotide changes in the overlapping transcript.

Following the initiation of replication, plasmids are replicated by either a Rolling-Circle (RC), a Strand Displacement or a Theta (θ) mechanism. RC replication was first seen in the 4 Kb plasmid pT181 from *Staphylococcus aureus* (Koepsel *et al.*, 1985). Despite being most prevalent in Gram-positive bacteria, and used by the *Butyrivibrio* plasmid pOM1 (Hefford *et al.*, 1997), the mechanism appears to be limited to small, high copy number plasmids and is yet to be observed in a plasmid greater than 10 Kb. This mechanism is, therefore, unlikely to be used by megaplasmids or secondary chromosomes. RC replication is a unidirectional,

asymmetric process (del Solar *et al.*, 1998) and is initiated by the incorporation of a nick in a site known as the double-stranded origin (*dso*) by a plasmid-encoded Rep protein (Marsin *et al.*, 2000). The *dso* is found exclusively in the plus, or Crick, strand of the plasmid (del Solar *et al.*, 1998). The resulting free 3'-OH end is used as a primer for leading strand synthesis (Lechner & Richardson, 1983). This process is thought to involve chromosome-encoded replication machinery such as DNA polymerase III and a single-strand binding protein (McInerney & O'Donnell, 2004). Along with the plasmid-encoded Rep protein, this is known as the replisome. As replication proceeds, the Crick strand is displaced. Replication continues until the replisome reaches the *dso* (Khan *et al.*, 1988). Finally the Rep protein catalyzes a DNA strand transfer reaction (Lechner *et al.*, 1983). This process releases both a double-stranded DNA molecule, consisting of the parental minus, or Watson, strand and the newly synthesized positive strand, and a single stranded DNA molecule consisting of the parental Crick strand (Khan *et al.*, 1988). The single-stranded molecule serves as a template for minus (Watson) strand synthesis. This process commences from the single-stranded origin (*sso*), distinct from the *dso*, but typically in the vicinity of the *dso* (Birch & Khan, 1992).

The Strand displacement mechanism, seen in some incompatibility group Q plasmids (del Solar *et al.*, 1998), requires two symmetrical and adjacent single strand origins (*sso*) present on opposite strands (Honda *et al.*, 1992). A rep protein binds to iterons in an adjacent AT-rich region and unwinds the duplex exposing the two *sso* as single-stranded entities. A plasmid-encoded primase primes the *sso* and replication proceeds continuously displacing the complimentary strand (Honda *et al.*, 1992). This results in two dsDNA molecules, each possessing a parental strand.

Theta (θ) replication can be either uni- or bi-directional (Frere *et al.*, 1993, Bruand *et al.*, 1993), involving a leading and a lagging strand. The leading strand is synthesized continuously, while the lagging strand is synthesized in discontinuous (Okazaki) fragments, which are later joined by DNA ligase (Inselburg & Oka, 1975). Theta replication can begin from a single or multiple o*riR*s. Theta o*riR*s commonly contain a number of general features aside from the Rep protein, including an adjacent AT-rich region containing direct, inverse and/or inverted repeats and one or more *dnaA* boxes. The *dnaA* boxes facilitate the binding of a host encoded DnaA, a chromosomal

replication initiator protein (Park & Chattoraj, 2001). Theta replication is seen in the *Butyrivibrio* plasmids pRJf1 and pRJf2 (Hefford *et al.*, 1997).

Analysis of the replication origins of all megaplasmids sequenced at the time of writing this literature review shows that most (>95%) contain genes encoding Rep proteins of the RepA or RepB-family (70%) and/or Partitioning (Par) proteins, such as ParA or ParB (82%) typically in close proximity to one another and, where defined, the replication origin. Rep proteins are thought to be exclusively encoded by plasmids, but are common to secondary chromosomes (NCBI). This may suggest that most secondary chromosomes have evolved from large or megaplasmids (Heidelberg *et al.*, 2000). Conversely genes encoding Par proteins are common to both plasmids and the major chromosome (Gerdes *et al.*, 2000). A comparison of Par proteins encoded by various plasmids to those encoded by major chromosomes shows that their phylogeny is distinct. Par proteins encoded by secondary chromosomes tend to show a phylogeny similar to that observed for plasmids (Gerdes *et al.*, 2000). Despite these apparent differences in replication machinery the replication of major and secondary chromosomes has been shown, at least in the case of *Vibrio cholerae,* to be co-ordinated (Egan *et al.*, 2004). Further, Egan *et al*. (2004) also found the initiation of the main and secondary chromosomes replication occurs simultaneously. Conversely, megaplasmids, like regular plasmids, are thought to autonomously regulate their initiation of replication as their copy number decreases (del Solar *et al.*, 1998) such as following cell division.

The best characterised *Butyrivibrio* o*riR*s are those of the small cryptic plasmids pRJf1, pRJf2 and pOM1. Each of these plasmids has two ORFs, both of which are required for their self replication and maintenance. These ORFs encode similar proteins in all three plasmids, potentially indicating a conserved requirement at the o*riR*. The first of these ORFs encodes an acidic plasmid recombinase (Pre) protein. The second encodes a Rep protein. This o*riR* composition is also seen in other rumen microbial plasmids such as pRRI2 of *Prevotella ruminicola* (Mercer *et al.*, 2001, Hefford *et al*., 1997, Kobayashi *et al*., 1995, Hefford *et al*., 1993).

1.13 Plasmid conjugative transfer

The ability of plasmids to be transferred between species is a common feature, and has been reported in several plasmids obtained from rumen bacteria including *Enterococcus faecium*, *Prevotella ruminicola*, and *Ruminobacter amylophilus* (Ogata *et al.*, 1999, Shoemaker *et al.*, 1992, Laukova *et al.*, 1990) as well as plasmid pOM1 from *Butyrivibrio fibrisolvens* Bu49 (Francia *et al.*, 2004, Hefford *et al.*, 1997). Transmissible plasmids may be separated into two groups based upon their transfer capacity: i) Conjugative plasmids, which encode all the apparatus necessary for both conjugation and transfer; ii) mobilisable plasmids, which encode the minimal transfer origin (o*riT*) and the mobilisation (*mob*) genes, but require a conjugative plasmid for co-mobilisation (Francia *et al.*, 2004). Conjugative transfer requires the intimate contact between the cell surfaces of a donor and a recipient cell. In Gram-negative bacteria this is established by specific sex pili. However, the mechanism of achieving cell-cell contact in Gram-positive bacteria has yet to be completely elucidated. Intraspecies transfer in *Enterococcus* is known to involve heat stable, protease-sensitive sex pheromones, and in *Bacillus thuringiens* protease-sensitive donor-recipient coaggregates are formed (Andrup *et al.*, 1993).

The transfer of DNA is thought to be mediated by two protein complexes, the relaxosome and the mating pair formation (mpf) complex, which are connected by a TraG-like coupling protein (Gilmour *et al.*, 2003). The relaxosome is a multiprotein complex which assembles at the o*riT* (Pansegrau *et al.*, 1990). It is composed of both plasmid and chromosome-encoded proteins (Furste *et al.*, 1989, Lanka & Wilkins, 1995). The relaxase (commonly a TraA, Mob or Pre protein) catalyses the cleavage of a phosphodiester bond at a specific site within the o*riT*, known as *nic* (Grohmann *et al.*, 2003). Inverted repeats are found adjacent to *nic* sites (Pansegrau *et al.*, 1988). It is thought that a stem-loop (hairpin) secondary structure allows the relaxase to recognise the *nic* site and catalyse the cleavage of a specific di-nucleotide pair within the unpaired region of the loop (Grohmann *et al.*, 2003). Following cleavage, the relaxosome becomes covalently linked by a tyrosyl residue to the 5' terminus of the cleaved strand. Cleaved DNA can then be unwound for transfer by a plasmid-encoded helicase (Lanka & Wilkins, 1995). Single-strand transfer is thought to occur by a replicative RC-type mechanism, until termination of a round of transfer is achieved by

the cleaving-joining activity of the relaxase linked to the 5' end of the transferring strand (Lanka & Wilkins, 1995).

The plasmid-encoded mpf complex facilitates the transfer of the DNA from donor to a competent recipient (Lessl *et al.*, 1993). In Gram-negative bacteria, conjugal mpf complexes are a subset of type-IV secretion systems (Salmond, 1994). Type IV secretion systems have the capacity to transfer protein or DNA-protein complexes intercellularly (Christie, 2001). In Gram-positive bacteria the complex is known to include an ATPase, mating channel proteins and a coupling protein (usually TraG, TrwB or TraD) (Grohmann *et al*., 2003, Francia *et al*., 2004). Conjugative transfer machinery has been used in the development of plasmid vectors (Lyras & Rood, 1998, Nallapareddy *et al.*, 2006), providing an alternative transformation procedure.

1.14 Plasmid vectors

Cloning vectors allow researchers to transfer and/or manipulate DNA within a compatible host organism. Researchers have previously used these techniques to express a fungal xylanase in *Butyrivibrio fibrisolvens* strains OB156 and H17c (Krause *et al.*, 2001, Xue *et al.*, 1997). While virus and transposon vectors exist, the most widely used vectors are plasmids (Zacher *et al.*, 1980, Sargent *et al.*, 2004). Genetic manipulation of rumen bacteria has been limited in comparison to those from other environments, despite a commonly held belief that genetic manipulation could provide a key opportunity to enhance animal productivity (Krause *et al.*, 2003). The limitation is due to the restricted availability of useful vectors and effective transfer and selection systems (Hefford *et al.*, 1997). Today several vectors have been developed for use in *Butyrivibrio* species (for example pBHerm, pSMerm1 or pYK4; Beard *et al*., 1995; Hefford *et al*., 1997; Kobayashi *et al*., 1998), which can be used to transform the bacterium by either electroporation or conjugation (Clark *et al.*, 1994, Kobayashi *et al.*, 1998). However, due, at least in-part, to the phylogenetically diverse nature of the *Butyrivibrio / Pseudobutyrivibrio* assemblage their host range tends to be limited to only a subset of all *Butyrivibrio* species. The most common choice of selectable marker in *Butyrivibrio* species is the erythromycin resistance gene, which has been shown to be effective (Kobayashi *et al.*, 1995), while several others, such as ampicillin or clindamycin have not (Ware *et al.*, 1992). However, the choice of an

antibiotic resistance marker depends on the target organism‟s innate ability to metabolise the corresponding antibiotic. *B. proteoclasticus* has been shown to be susceptible to the antibiotics ampicillin (10 μg/ml), tetracycline (10 μg/ml), chloramphenicol (10 μg/ml), gentamicin (10 μg/ml), and monensin (2.5 μg/ml) and resistant to streptomycin (Attwood *et al.*, 1996).

1.15 Shuttle Vectors

Vectors can be constructed which replicate in more than one species. These are known as shuttle vectors. Shuttle vectors allow a gene (or multiple genes) of interest to be shuttled between, and manipulated within, multiple hosts. This commonly includes *E. coli,* a host that is well characterised and has a short generation time, allowing easy manipulation and characterisation of the recombinant constructs. The same construct can then be introduced and replicated in the native system which is typically less well characterised and has a longer generation time. Shuttle vectors typically contain multiple origins of replication, each functional in the particular host where the propagation of the recombinant construct is planned. Alternatively, broad host range plasmids (promiscuous plasmids) enable the development of shuttle vectors functioning from a single o*riR* (Schofield *et al.*, 2003) in multiple host organisms. The *B. fibrisolvens* plasmids, pRJf1, and pOM1, have both been used successfully to develop *Butyrivibrio* / *E.coli* shuttle vectors (Beard *et al*., 1995, Hefford *et al*., 1997), using an erythromycin selectable marker from the *Enterococcus faecalis* plasmid pAMβ1, However, previous attempts to introduce these shuttle vectors into *B. proteoclasticus* have been unsuccessful.

2. Materials and Methods

2.1 Materials

2.1.1 Agarose

Four different grades of agarose were used including Low melt agarose (Progen, Heidelberg, Germany). Pulsed-field certified low melt agarose (Bio-Rad, Hercules, CA, USA), UltraPure™ agarose (Invitrogen, Carlsbad, CA, USA) and Pulsed-field certified agarose (Bio-Rad).

2.1.2 Antibiotics

The antibiotics ampicillin, chloramphenicol, tetracycline and novobiocin were supplied by Sigma-Aldrich (St. Louis, MO, USA).

2.1.3 Bacterial Strains

Bacterial strains used in this work and their source are listed in Table 2.1

Table 2.1 Bacterial strains used

Genus	Species	Strain	Source
Butyrivibrio	*crossotus*	DSM 2876	Rod Mackie, UIUC[3]
	fibrisolvens	D1^T	Rod Mackie, UIUC[3]
		WV1	AgResearch[2]
		C211	AgResearch[1]
	hungatei	JK615	Jan Kopecny, IAPG[4]
		DSM 10295	DSMZ[6]
		A38	Rod Mackie, UIUC[3]
		AR10	Ron Teather, LRC[7]
		C130a	AgResearch[1]
		C219	AgResearch[1]
	proteoclasticus	B316^T	AgResearch[1]
		UC142	Jan Kopecny, IAPG[4]
	unspeciated	JK619	Jan Kopecny, IAPG[4]
Escherichia	*coli*	DH5α	Paul Rainey, UA[8]
		S17-1 λπ	Paul Rainey, UA[8]
Methanobrevibacter	*ruminantium*	M1 (DSM 1093)	DSMZ[6]

Pseudobutyivibrio	*ruminis*	DSM9787	DSMZ[6]
		CF3	Burk Dehority, OARDC[5]
		CF1b	Burk Dehority, OARDC[5]
	xylanivorans	DSM 10317	DSMZ[6]
		$Mz5^T$	Jan Kopecny, IAPG[4]
		Ce52	AgResearch[1]

[1]AgResearch Ltd, Grasslands Research Centre, Palmerston North, New Zealand.

[2]AgResearch Ltd, Wallaceville Research Centre, Upper Hutt, New Zealand

[3]University of Illinois at Urbana-Champaign, IL, USA.

[4]Institute of Animal Physiology and Genetics, Prague, Czech Republic.

[5]Ohio Agricultural Research and Development Center, Ohio State University, Wooster, OH, USA.

[6]Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (German Collection of Microorganisms and Cell Cultures), Germany

[7]Lethbridge Research Centre, Lethbridge, Canada

[8]University of Auckland, School of Biological Sciences, Auckland, New Zealand

2.1.4 Buffers and solutions

2.1.4.1 Acridine Orange Curing Solution

110 μM resazurin (L. Light and Co. Ltd., Coinbrook, England) was added to distilled water (dH_2O) and boiled for 5 min to facilitate the removal of dissolved oxygen (O_2). It was then allowed to cool to room temperature while gassing with 100% CO_2. Gassing continued until the resaurin turned colourless, indicating the solution was anoxic. 1.9 mM acridine orange (MERCK., Darmstadt, Germany) was added and the solution was then capped and transferred to an anaerobic chamber. Inside the chamber the solution filter sterilised using a 0.22 μm filter (Millipore, Cork, Ireland).

2.1.4.2 Acriflavine Curing Solution

Solution was prepared as described above (2.1.4.1) using 2 mM acriflavine (Sigma-Aldrich) in place of acridine orange.

2.1.4.3 Alkaline Lysis Solutions I, II and III

Alkaline lysis solutions I, II and III were prepared as previously described (Sambrook & Russell, 2001). Components of each solution (listed Table 2.2) were dissolved separately in dH_2O. The solution was autoclaved at 15 psi for 20 min.

Table 2.2 Components for alkaline lysis solutions

Chemical	Concentration
Solution I	
Ethylenediaminetetraacetic acid (EDTA)	10 mM
Glucose	50 mM
Tris–Hydrochloride (Tris-HCl)	25 mM
Solution II	
Sodium dodecyl sulfate (SDS)	35 mM
Sodium hydroxide (NaOH)	0.2 M
Solution III	
Glacial acetic acid	1.92 M
Potassium acetate	3 M

2.1.4.4 Ammonium Acetate Solution (5M stock)

5 mM ammonium acetate (VWR International Ltd.) was dissolved in dH_2O and the resulting solution was filter sterilised.

2.1.4.5 CE Buffer

CE buffer was made as described by Martin and Dean (1989). 50 mM sodium carbonate and 25 mM EDTA (both supplied by VWR International Ltd.) were combined in dH_2O, adjusted to pH 10.0 using 6 N NaOH and then autoclaved.

2.1.4.6 Chloroform

Chloroform was supplied in analytical grade by VWR International Ltd.

2.1.4.7 Colony Hybridisation Lysis Solution

250 mM sucrose (VWR International Ltd.) and 10 mM tris-HCL (Invitrogen) were dissolved in dH_2O, adjusted to pH 7.5 and then autoclaved.

2.1.4.8 Conjugation Buffer

0.01 % Bacto-peptone (Becton, Dickinson and Co., Sparks, MD, USA), 145 mM Sodium chloride (NaCl; VWR International Ltd.) and 9 mM sodium thioglycollate (Sigma-Aldrich) were dissolved in dH_2O and autoclaved.

2.1.4.9 Denaturation Solution

1.5 M NaCl and 0.5 M NaOH were dissolved in dH_2O and autoclaved.

2.1.4.10 Diethylpyrocarbonate (DEPC)

DEPC was supplied by Sigma-Aldrich.

2.1.4.11 Depurination Solution

250 mM HCl was added to sterile dH_2O.

2.1.4.12 Deoxynucleoside Triphosphates (dNTP) Solution

dNTPs (dATP; dCTP; dGTP; and dTTP), were each supplied at a concentration of 100 mM by Invitrogen and were mixed together in equimolar concentrations and diluted 5 fold in milliQ water to give 5 mM stock solutions. Stock solutions were stored at -20 ºC until required.

2.1.4.13 EC Buffer

100 mM EDTA, 1 M NaCl, 35 mM N-lauryl-sarcosine (Sigma-Aldrich) and 6 mM trizma base (Invitrogen) were dissolved in dH_2O, adjusted to pH 7.6 with NaOH and autoclaved.

2.1.4.14 EDTA-Sarkosyl Solution

0.5 M EDTA and 35 mM N-lauryl sarcosine were dissolved in dH_2O, adjusted to pH 8.0 with NaOH, then autoclaved.

2.1.4.15 EDTA Solutions

The appropriate concentration of EDTA (0.5 M or 0.125 M) was dissolved in dH_2O and adjusted to pH 8.0, then autoclaved.

2.1.4.16 Electroporation Buffer

Electroporation buffer was prepared as previously described (Beard *et al*., 1995). 110 μM resazurin (L. Light and Co. Ltd., Coinbrook, England) was boiled in dH_2O and then cooled under O_2-free CO_2 until anaerobic. 300 mM sorbitol and 1 mM dithioerythritol (DTT; both supplied by Sigma-Aldrich) were added and gassing continued for 30 min. The solution was transferred into an anaerobic chamber and filter sterilised.

2.1.4.17 Ethanol

Ethanol was supplied in analytical grade by VWR International Ltd**.** as either 95% (v/v) or absolute (minimum 99.7% v/v). The ethanol was used at either 70% (diluted from the 95% stock) or at absolute concentration.

2.1.4.18 Ethidium Bromide

Ethidium bromide was supplied as a 10 mg/ml stock solution, diluted in water, by Invitrogen.

2.1.4.19 Ethidium Bromide Curing Solution

Solution was prepared as described above (2.1.4.1) using 1.27 mM ethidium bromide in place of acridine orange.

2.1.4.20 Glycerol Solutions (10% and 60%)

1.1 M (for 10 % solution) or 6.6 M (for 60% solution) of Glycerol (VWR International Ltd.) was mixed with dH_2O and boiled for 5 min. The solution was cooled to room temperature under O_2-free CO_2 and then autoclaved. It was then left to equilibrate for >72 h in an anaerobic chamber.

2.1.4.21 Hybridisation Solution (Southern)

Hybridisation solution was supplied by GE Healthcare (Uppsala, Sweden) as part of the AlkPhos Direct Labelling and Detection system. The composition is unknown as it is proprietary knowledge of the company.

2.1.4.22 IPTG Stock Solution

0.8 M isopropyl-β-D-thiogalactopyranoside (IPTG; Sigma-Aldrich) was dissolved in dH_2O and filter sterilised then stored in 1 ml aliquots at -20 ºC.

2.1.4.23 Isopropanol

Isopropanol (Propan-2-ol) was supplied in analytical grade by VWR International Ltd.

2.1.4.24 Liquid Nitrogen

Liquid nitrogen was supplied by BOC (Palmerston North, NZ.).

2.1.4.25 Microarray Pre-hybridisation Buffer

38 mM bovine serum albumin (BSA; Sigma-Aldrich), 3.5 mM sodium dodecyl sulphate (SDS; VWR International Ltd.) and 5 x saline sodium citrate (SSC; Ambion., Austin, TX, USA) were dissolved in sterile milliQ water, in the order listed, then filter sterilised.

2.1.4.26 Microarray Wash Solutions

The components of each Wash solution (listed in Table 2.3) were dissolved in sterile milliQ water then filter sterilised.

Table 2.3 Microarray wash solution compositions

Chemical	Concentration
Wash solution 1	
SDS	10%
SSC	2×
Wash solution 2	
SSC	1×
Wash solution 3	
SSC	0.1×

2.1.4.27 Mineral Solution

45 mM ammonium sulphate ($(NH_4)_2SO_4$), 10 mM calcium chloride dihydrate ($CaCl_2.2H_2O$), 10 mM magnesium sulphate 7-hydrate ($MgSO_4.7H_2O$), 200 mM NaCl and 45 mM potassium dihydrogen orthophosphate (KH_2PO_4; all supplied by VWR International Ltd.) were dissolved in dH_2O and autoclaved.

2.1.4.28 NaCl Solution

5 M NaCl was dissolved in dH_2O and autoclaved.

2.1.4.29 Neutralisation Solution

1.5 M NaCl and 0.5 M trizma base were dissolved in dH_2O and autoclaved.

2.1.4.30 Novobiocin Curing Solution

Solution was prepared as described above (2.1.4.1) with 820 µM novobiocin (Sigma-Aldrich) in place of acridine orange.

2.1.4.31 Pfennigs Heavy Metal Solution

84 µM cobalt chloride 6-hydrate ($CoCl_2.6H_2O$), 14 µM copper chloride (CuCl), 1.3 mM EDTA, 1 mM manganese chloride tetrahydrate ($MnCl_2.4H_2O$) 8 µM nickel (II) chloride 6-hydrate ($NiCl_2.6H_2O$), 0.5 mM orthoboric acid ($B(OH)_3$), 12 µM sodium molybdenum oxide dihydrate ($NaMoO_4.2H_2O$), 35 µM zinc sulphate 7-hydrate ($ZnSO_4.7H_2O$; all supplied by VWR International Ltd.) and 0.72 mM iron (II) sulphate 7-hydrate ($FeSO_4.7H_2O$; MERCK) were dissolved in dH_2O and stored at room temperature.

2.1.4.32 Phenol

UltraPure buffer-saturated phenol was supplied by Invitrogen.

2.1.4.33 Phenol:Chloroform Solution

Phenol and chloroform were combined in a 1:1 ratio (v/v).

2.1.4.34 Phenol :Chloroform:Isoamyl Alcohol Solution

Phenol, chloroform and isoamyl alcohol (Invitrogen) were combined in a ratio of 25:24:1 (v/v/v).

2.1.4.35 PMSF Stock Solution

100 mM phenylmethylsulfonyl fluoride (PMSF; Kneisel, 1968) was dissolved in isopropanol by shaking. The dissolved solution was filter sterilised and stored at -20 ºC. A working solution was made by diluting the PMSF stock solution 100:1 in TE (10:1) buffer.

2.1.4.36 Potassium Acetate Solution (1M)

1 M potassium acetate (VWR International Ltd.) was dissolved in dH_2O and adjusted to pH 7.5 with 2 M acetic acid. The solution was filter sterilised and stored at -20 °C in 1 ml aliquots.

2.1.4.37 R1 Salts Solution

R1 salts solution was made as described by Schaefer et al. (1980). 190 mM $CaCl_2.2H_2O$, 27 mM EDTA, 0.62 M $MgCl_2.4H_2O$ (all supplied by VWR International Ltd.) and 18 mM $FeSO_4.7H_2O$ were combined in 1 part pfennig‟s heavy metal solution (2.1.4.31) and 5 parts dH_2O and stored at room temperature

2.1.4.38 Reducing Agent

70 mM L-cysteine-HCL (VWR International Ltd.) was dissolved in ¼ of the final volume of dH_2O and adjusted to pH 10.0 with NaOH. $Na_2S.9H_2O$ was added rapidly along with the remaining dH_2O. The solution was boiled under O_2-free N_2 and subsequently aliquoted into 50 ml serum bottles flushed with O_2-free N_2. The solution was then autoclaved.

2.1.4.39 Rumen Fluid

Rumen fluid was collected from 14-18 h fasted rumen-canulated cattle. The liquid was centrifuged twice at 20,000 x gravity (*g*), with the pelleted material being discarded each time. The resulting liquid was stored at -20 ºC until required.

2.1.4.40 Saline-EDTA Solution

10 mM EDTA and 150 mM NaCl were dissolved in dH_2O and adjusted to pH 8.0 then autoclaved.

2.1.4.41 SDS Solution (20% w/v)

0.7 M SDS was dissolved in dH_2O and adjusted to pH 7.2 then filter sterilised.

2.1.4.42 SDS Curing Solution

Solution was prepared as described above (2.1.4.1) with 350 mM SDS in place of acridine orange.

2.1.4.43 Saline Sodium Citrate (20 ×)

Saline sodium citrate (SSC) for work pertaining to the microarray analysis was supplied by Ambion. All other SSC was prepared as described below:

0.75 M NaCl and 75 mM sodium citrate (L. Light and Co. Ltd.) were dissolved in dH_2O and adjusted to pH 7.0 with 1 M HCl. The solution was then autoclaved.

2.1.4.44 Sodium Acetate Solution (3M)

Sodium acetate (3 M final concentration, VWR International Ltd) was dissolved in DEPC-treated milliQ water and adjusted to pH 5.2 using glacial acetic acid (VWR International Ltd.), then autoclaved.

2.1.4.45 SSPE Solution

1 mM EDTA, 180 mM NaCl and 10 mM NaH_2PO_4 were combined in dH_2O and adjusted to pH 7.7 then autoclaved.

2.1.4.46 STE Buffer

1 mM EDTA, 100 mM NaCl and 10 mM Trizma base were dissolved in dH_2O and adjusted to pH 8.0 before autoclaving. The buffer was stored at 4 ºC.

2.1.4.47 TAE Buffer (50x Stock Solution)

950 mM acetic acid, 50 mM EDTA and 2 M trizma base were dissolved in dH_2O and adjusted to pH 8.0 before autoclaving. A working solution (1×) was made by diluting the stock solution 50:1 in dH_2O.

2.1.4.48 TBE Buffer (5× Stock Solution)

445 mM $B(OH)_3$, 10 mM EDTA and 445 mM trizma base were dissolved in dH_2O and autoclaved. A working solution (0.5 ×) was made by diluting the stock solution 10:1 in dH_2O.

2.1.4.49 TE Buffers

Components of each TE buffer (listed table 2.4) were dissolved in dH_2O and adjusted to pH 8.0 before autoclaving.

Table 2.4 TE buffer components

Chemical	Concentration
TE 10/0.1	
EDTA	0.1 mM
Trizma base	10 mM
TE 10/1	
EDTA	1 mM
Trizma base	10 mM
TE 10/100	
EDTA	100 mM
Trizma base	10 mM

2.1.4.50 TES Buffer

1 mM EDTA, 250 mM sucrose and 10 mM trizma base were dissolved in dH_2O, and adjusted to pH 7.5 before autoclaving.

2.1.4.51 Trace Element Solution

Nitrilotriacetic acid was dissolved in dH_2O and adjusted to pH 6.5 using 1 M KOH. The remaining components (listed Table 2.5) were added and the solution was stored at 4 ºC.

Table 2.5 Trace element solution components

Chemical	Concentration
Aluminum potassium sulfate dodecahydrate ($KAl(SO_4)_2.12H_2O$; VWR International Ltd.)	42 µM
$B(OH)_3$	162 µM
$CaCl_2.2H_2O$	680 µM
Cobalt sulphate 7-hydrate ($CoSO_4.7H_2O$; VWR International Ltd.)	640 µM
Copper (II) sulphate pentahydrate ($CuSO_4.5H_2O$; Ajax Chemical International Pty Ltd Sydney, Australia)	40 µM
$FeSO_4.7H_2O$	360 µM
Manganese sulphate dehydrate ($MnSO_4.2H_2O$; VWR International Ltd.)	2.24 mM
$MgSO_4.7H_2O$	12 mM
NaCl	17 mM
$NiCl_2.6H_2O$	100 µM
Sodium molybdate dihydrate ($Na_2MoO_4.2H_2O$; VWR International Ltd.)	40 µM
Sodium selenite pentahydrate ($Na_2SeO_3.5H_2O$; Hopkin and Williams Ltd. (Essex, England))	1 µM
$ZnSO_4.7H_2O$	625 µM

2.1.4.52 TRIzol

TRIzol (a guanidine-isothiocyante reagent used to stabilise RNA) was supplied by Invitrogen.

2.1.4.53 Vitamin Solution

Distilled water was boiled then allowed to cool to room temperature under O_2-free CO_2. All components (listed Table 2.6) were added and O_2-free CO_2 was allowed to bubble through the solution for a further 20 min before the solution was filter sterilised into Hungate tubes in 10 ml aliquots. Reducing agent (0.1 ml) was added to each Hungate tube and the resulting solution was stored at -20 ºC.

Table 2.6 Vitamin solution components

Chemical	Concentration
Biotin	8 µM
D-Ca-pantothenate	23 µM
Folic acid	4.5 µM
Lipoic acid	24 µM
Nicotinic acid	40 µM
p-Aminobenzoic acid	36 µM
Pyridoxine-HCl	50 µM
Reducing Agent	1%
Riboflavin	13 µM
Thiamine-HCl dihydrate	15 µM
Vitamin B_{12}	75 nM

2.1.4.54 Volatile Fatty Acid (VFA) Solution

114 mM butyric acid, 28 mM *iso*-butyric acid, 24 mM n-valeric acid, 24 mM *iso*-valeric acid (all supplied by Sigma-Aldrich), 24 mM DL-2-methyl butyric acid, 202 mM propionic acid (both supplied by MERCK) and 710 mM acetic acid were combined in dH_2O and adjusted to pH 7.5 with 1 M NaOH then stored at -20 ºC.

2.1.4.55 Wash Solution For Pulsed-Field Gel-Electrophoresis

2 M NaCl and 20 mM trizma base were dissolved in dH_2O and adjusted to pH 7.6 with 6 M HCl then autoclaved.

2.1.4.56 X-Gal

50 mM 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal; Sigma-Aldrich) was dissolved in dimethylformamide (DMF) and stored in an aluminium foil-covered 15 ml falcon tube at -20 ºC.

2.1.4.57 Xylan Solution

5 % Oat spelts xylan was dissolved in BY+ medium (2.1.9.1) and the solution was autoclaved. 1% reducing agent was added and the solution was stored at room temperature until use (within 1 week).

2.1.5 Enzymes

2.1.5.1 Calf Intestinal Alkaline Phosphatase

Calf intestinal alkaline phosphatase (CIP; Morton, 1950) was supplied at a concentration of 10,000 units/ml by Biolab Ltd (Auckland, New Zealand) and originally sourced from New England Biolabs (Beverley, MA, USA).

2.1.5.2 Lysozyme

Lysozyme (Flemming, 1922), was supplied by Boehringer Mannheim GmbH (Mannheim, Germany). Lysozyme was used from a 25 mg/ml stock solution made as follows**;** 125 mg of lysozyme was dissolved in 5 ml TE (10/1) buffer and filter sterilised then stored at -20 ºC.

2.1.5.3 Proteinase K

Proteinase K (Roelcke & Uhlenbruck, 1969) was supplied by Roche Diagnostics (Basel, Switzerland) and was used from a 10 mg/ml stock solution made as follows; 50 mg of proteinase K was dissolved in 5 ml TE (10/1) buffer and filter sterilised then stored at -20 ºC

2.1.5.4 Restriction Endonucleases

All Restriction endonucleases (REs), their reaction buffers and bovine serum albumin (BSA) were supplied by New England Biolabs.

2.1.5.5 Ribonuclease A (RNaseA)

RNaseA was supplied by Sigma-Aldrich and used from a 10 mg/ml stock solution made as described below: 50 mg of RNase A was combined with 5 ml 10mM Tris-HCl buffer. The solution was boiled for 15 min to degrade any contaminating DNase, and then allowed to cool to room temperature. The solution was filter sterilised and stored at -20 ºC.

2.1.5.6 T4 DNA ligase

T4 DNA ligase was supplied as Ready-To-Go™ ampules of lysophilized T4 DNA ligase, buffer and ATP (sufficient for a 20 µl reaction) by GE Healthcare (Uppsala, Sweden).

2.1.5.7 T4 DNA polymerase

T4 DNA polymerase was supplied by New England Biolabs at a concentration of 3,000 units/ml.

2.1.6 Gel migration size standards

All Gel migration standards are listed in Table 2.7

Table 2.7 Gel migration standards

Standard	Range	Use*	Supplier
1Kb+ ladder	100 bp – 12 Kb	GE	Invitrogen
Lambda (λ) ladder	48.5 Kb – 727.5 Kb	PFGE	New England Biolabs
Mid range marker I	15 Kb – 242.5 Kb	PFGE	New England Biolabs
Low range marker	2.03Kb – 194Kb	PFGE	New England Biolabs

* GE: indicates use in conventional Gel Electrophoresis, PFGE: indicates use in Pulsed-field gel-electrophoresis only.

2.1.7 Glassware

Beakers and Conical Flasks were supplied in a range of sizes by Biolab Limited (Auckland, New Zealand) and Kimble Chase Life Science and Research Products LLC (Vineland, NJ, USA) respectively. Hungate Tubes (10 ml and 15 ml sizes) were supplied by Kimble Chase Life Science and Research Products LLC. Schott Bottles were supplied in a range of sizes by Biolab Ltd (Auckland, New Zealand) and originally sourced from Schott Duran® (Mainz, Germany). Serum bottles were supplied in both 50 ml and 125 ml sizes by Wheaton Science Products (Millville, NJ, USA).

2.1.8 Laboratory equipment

The autoclave used for steriliation of media and equipment was a high pressure steam sterilizer model ES-315 supplied by TOMY (Hempstead, NY, USA). Nucleic acids were measured, by spectrophotometry, using an Agilent model 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) or a NanoDrop® model ND-1000, (NanoDrop Technologies, Inc. Wilmington, DE, USA). Five types of centrifuge were used in the work described; they are listed below (Table 2.8). Centrifuge tubes were supplied in a variety of different capacities as listed below (Table 2.9).

Table 2.8 Centrifuge specifications

Brand	Vessels spun	Max RCF ($\times g$)	Temp Control	Supplier
MiniSpin+ personal microcentrifuge	0.6ml and 1.5ml Eppendorf tubes	14,100	No	Eppendorf (Hamburg, Germany)
Biofuge Fresco	1.5ml Eppendorf tubes	16,000	Yes	Heraeus (Hanau, Germany)
Minor	Hungate tubes	3,000	No	MSE (London, UK)
IEC Centra	15 ml and 50 ml Falcon tubes	3,500	Yes	Thermo Fisher Scientific, Inc. (Waltham, MA, USA)
Sorvall® Evolution RC	50 ml Oakridge tubes and *250 ml centrifuge tubes	48,000	Yes	Thermo Fisher Scientific, Inc.

* The rotor housing the 250 ml centrifuge tubes could only be centrifuged at a maximum of 27,500× g

Table 2.9 Centrifuge tubes and suppliers

Brand	Capacity	Supplier
Eppendorf tubes	0.6 ml 1.5 ml	Eppendorf
Falcon tubes	15 ml 50 ml	Becton, Dickinson and Co. (Sparks, MD, USA)
Oakridge tubes	50 ml	Thermo Fisher Scientific, Inc.
Wide mouth centrifuge bottles	250 ml	Nalgene (Rochester, NY, USA)

The electrophoresis unit used for Pulsed-Field Gel-Electrophoresis was a CHEF-DR® III system powered by a Powerpac Basic (Bio-Rad). Electroporation was carried out using a Gene pulser ™ system supplied by Bio-Rad. Gas chromatography (GC) was carried out using an Aerograph 660 supplied by Wilkens Instruments and Research, Inc. (Walnut Creek, CA, USA). Gels (both conventional and pulsed-field) were digitally photographed and documented using a Gel Logic 200 Imaging System supplied by Eastman Kodak Company (Rochester, NY, USA). The gel tanks used for conventional agarose gel-electrophoresis were a wide Mini-sub® cell GT and for Pulsed-Field Gel-Electrophoresis, a CHEF electrophoresis cell was used (both supplied by Bio-Rad). The scanner used for microarrays was a GenePix® Professional 4200 Scanner supplied by MDS Analytical Technologies (Sunnyvale, CA, USA). An Olympus Vanox AHBT3 microscope was used for microscopy (Olympus America Inc., Center Valley, PA, USA). The shakers, incubators and waterbaths used are listed below (Table 2.10).

Table 2.10 Shakers, incubators and water baths

Brand	Shaker	Max Temp	Supplier
Orbitec XL	Yes	n/a	Infors-HT (Bottmingen, Switzerland).
Thermotec 2000	No	300 ºC	Contherm (Wellington, NZ)
Precision Incubator	No	100 ºC	Contherm
Hybridisation oven	Yes	100 ºC	GE Healthcare
Julabo F10 upright waterbath	No	0 ºC – 105 ºC	Labortechnik GMBH (Seelbach, Germany)
Grant Y28 waterbath	No	Room Temp – 100 ºC	Grant Instruments Ltd. (Cambridge, UK)

A warm room, which housed the orbital shakers, was set at 39 ºC and heat was dispersed throughout the room with a fan. Conventional Polymerase Chain Reactions (PCRs) were carried out in a Px2 Thermal Cycler (ThermoHybaid, Sedanstr, Germany), while Real time PCR reactions used a Lightcycler Series II (Roche Diagnostics, Basel, Switzerland). The pH meter used to measure and adjust the pH of media and solutions was a PHM62 pH Meter (Radiometer, Copenhagen, Denmark). Test tube heating was carried out in a Test Tube Heater SHTD (Stuart Scientific, Stone, Staffordshire, UK).

N_2 and CO_2 gases were supplied at food grade while H_2 was supplied either as a pure gas or as an 80:20 (v/v) mixture with CO_2 (BOC, Auckland, NZ.). All gases were scrubbed to remove traces of O_2 by passing the gas over heated (400 ^oC) copper filings in an inline furnace. The copper filings were routinely re-reduced by flushing with H_2 for several minutes.

The Anaerobic glove box is manufactured by Coy laboratory products INC. (Grass Lake, MI, USA) and maintains a 95% CO_2: 5% H_2 environment but would likely conatin trace amounts of N_2, which is used to flush incoming materials of O_2.

Parafilm M was supplied by American National Can (Chicago, IL, USA). Pipettors used during experimental work were Finnpipette Digital Variable Range pipettors supplied by Thermo Fisher Scientific (Waltham, MA, USA). Pipette tips were supplied in 10 µl, 200 µl, 1 ml, 5 ml and 10 ml sizes by Quality Scientific Plastics (Petaluma, CA, USA). Certified RNase/DNase-free pipette tips were supplied in 10 µl, 200 µl and 1 ml sizes by Sorenson Bioscience, Inc. (Salt Lake City, UT, USA). Microscope slides and coverslips were manufactured by Esco and supplied by Biolab Limited (Auckland, New Zealand). The counting chamber used for cell enumeration was a Thoma Slide etched with a 2.5 μm^2 grid that was 20 µm deep (Webber Scientific International, Ltd., Teddington, England).

2.1.9 Media

All broth media were prepared as described below unless otherwise specified. Where solidified media were used, 1.5% Bacteriological Agar was added to the described media.

2.1.9.1 BY+ medium

All components listed below (Table 2.11), except L-Cysteine-HCL and Vitamin solution, were dissolved in 1L of dH_2O. The solution was boiled until Resazurin turned from red to colourless, indicating the medium was anaerobic. The medium was allowed to cool on ice under O_2-free CO_2 to room temperature. L-Cysteine-HCL was then added and 96 ml aliquots of the medium were distributed into 125ml serum

bottles under O_2-free CO_2 and autoclaved. The Vitamin solution was added to each bottle just prior to use

Table 2.11 BY+ medium

Chemical	Volume
$(NH_4)_2SO_4$	250 mg
Bacto-Yeast extract (Fort Richard Laboratories Ltd., Auckland, NZ)	2 g
$CaCl.2H_2O$	130 mg
Dipotassium phosphate (K_2HPO_4)	1 g
KH_2PO_4	0.5 g
L-Cysteine-HCL	0.5 g
$MgSO_4.7H_2O$	200 mg
NaCl	1 g
Resazurin	1 g
Rumen fluid	300 ml
Sodium bicarbonate ($NaHCO_3$; VWR International Ltd.)	5 g
Trace Element solution	10 ml
Vitamin solution	1%

2.1.9.2 DM Arabinose medium

All ingredients listed below (Table 2.12), except Trace element solution, Vitamin solution, L-arabinose and Reducing agent, were dissolved in 1L of dH_2O. The solution was boiled until anaerobic, then cooled on ice to room temperature under O_2-free CO_2. L-arabinose and Trace element solution were then added. The medium was adjusted to pH 6.8 with 1 M KOH and distributed in 9 ml aliquots, into 15 ml, O_2-free CO_2-flushed Hungate tubes and autoclaved. Reducing agent and Vitamin solution were subsequently added to the cooled media, via a sterile CO_2-flushed 1 ml syringe.

Table 2.12 DMA medium

Chemical	Volume
Bacto-casamino acids (Becton, Dickinson and Co., Sparks, MD, USA)	5 g
Hemin	1 ml
K_2HPO_4	240 mg
L-arabinose	2 g
$MgSO_4.7H_2O$	250 mg
Mineral solution	40 ml
Na_2CO_3	4 g
R1 Salts	1 ml
Reducing agent	1%
Resazurin	1 mg
Trace element solution	10 ml
VFA solution	10 ml
Vitamin solution	1%

2.1.9.3 GYT medium

GYT media was prepared as described by Tung and Chow (1995). 0.5 g Bacto-tryptone, 0.25 g Bacto-yeast extract and 20 ml Glycerol were dissolved in 180 ml of dH_2O the medium was autoclaved.

2.1.9.4 Luria-Bertani (LB) medium

10 g Bacto-tryptone, 5 g Bacto-yeast extract and 10 g NaCl were dissolved in 1 L of dH_2O, the medium was adjusted to pH 7.0 and then autoclaved.

2.1.9.5 M704 medium

M704 media was prepared as described by the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ; DSMZ, 1993) All ingredients listed below (Table 2.13), except sodium carbonate, L-cysteine-HCL, and sodium sulphide, were dissolved in 1L dH_2O. The medium was boiled until anaerobic then cooled on ice under O_2-free CO_2 to room temperature. Sodium carbonate, sodium sulphide and L-cysteine-HCL were then added, the media was adjusted to pH 6.7 – 6.8 and 9 ml aliquots of the medium distributed, under O_2-free CO_2 into Hungate tubes and autoclaved.

Table 2.13 M704 medium

Chemical	Volume
Bacto-peptone (BBL)	2 g
Bacto-yeast extract	2 g
Cellobiose	0.5 g
Glucose	0.5 ml
Glycerol	0.5 g
Hemin	1 mg
K_2HPO_4	0.3 g
L-Cysteine-HCl	250 mg
Maltose	0.5 g
Mineral solution	75 ml
Na_2CO_3	4 g
$Na_2S.9H_2O$	250 mg
Resazurin	1 mg
Rumen fluid	150 ml
Starch, soluble	0.5 g
VFA solution	3.1 ml

2.1.9.6 RGM medium

All ingredients listed below (Table 2.14), except Reducing agent, were dissolved in 500 ml dH_2O and boiled until anaerobic. The medium was allowed to cool on ice to room temperature under O_2-free CO_2 and adjusted to pH 6.8. The medium was distributed, under O_2-free CO_2 in 9 ml aliquots, into Hungate Tubes and autoclaved. Reducing Agent was added following autoclaving.

Table 2.14 RGM medium

Chemical	Volume
Bacto-tryptone	1.5 g
Bacto-yeast extract	1 g
Glucose	2 ml
Hemin	0.5 ml
K_2HPO_4	150 mg
Mineral solution	25 ml
Na_2Co_3	0.8 g
R1 salt solution	0.5 ml
Reducing agent	1%
Resazurin	0.5 mg
VFA solution	5 ml

2.1.9.7 SOC medium

4 g Bacto-tryptone, 1 g Bacto-yeast extract, 720 µl Glucose, 0.5 g $MgSO_4.7H_2O$, 100 mg NaCl and 37 mg Potassium chloride (KCl; VWR International Ltd.) were dissolved in 200 ml of dH_2O then autoclaved.

2.1.10 Microarrays

Oligonucleotide probes (70mers) were designed against the draft *B. proteoclasticus* $B316^T$ and *Methanobrevibacter ruminantium* genomes sequences using ROSO software (Reymond *et al*., 2004). Oligonucleotides were synthesized and supplied at 100 µM concentration by Illumina (San Diego, CA, USA). Microarrays were printed onto epoxy-coated slides (Corning, Lowell, MA, USA) at the Invermay campus of AgResearch in Dunedin, using an ESI robot (Engineer Service Inc., Toronto, Ontario, Canada). Each probe was printed in spots 100 µm in diameter, containing approximately 0.26 µM of probe. Each spot probe was replicated within each slide between 3 and 7 times and spots were positioned randomly. Control spots included 321 blank spots, eight oligonucleotide probes designed to *Arabidopsis thaliana* ORFs (listed Table 2.15) and ten probes of random oligonucleotide sequence. Hybridisation-ready microarrays were stored desiccated at room temperature.

Table 2.15 Arabidopsis control probes

	NCBI reference	Gene name	Protein encoded
1	U39449	Act1	Actin gene
2	X14312	Cra1	12s seed storage protein
3	AB008103	AtERF-1	Ethylene responsive element
4	X13611	Ats1A	Ribulose 1,5-bisphosphate
5	AF042196	Arf8	Auxin responsive factor 8
6	X53435	Aux2-11	Auxin inducible gene
7	M84701	Tub3	β-3-tubulin
8	X64460	Lhb1B2	Photosystem II chlorophyll a/b binding-protein

2.1.11 Software

All software used is listed below (Table 2.17)

2.1.12 Vectors

All vectors used are listed below (Table 2.16)

Table 2.16 Vectors

Vector	Host	Size	Supplier
pJIR1456	*Enterococcus faecalis*	6857 bp	Department of Microbiology, Monash University, Victoria, Australia
Topo-TA	*E. coli*	3931 bp	Invitrogen
pCC1BAC	*E. coli*	8128 bp	Epicentre (Madison, WI, USA)

Table 2.17 Software details

Software	Application	Source	Reference(s)
Artemis (v7.0)	Genome sequence viewing and analysis	http://www.sanger.ac.uk/Software/Artemis/	(Rutherford *et al*., 2000)
Bioconductor (v 2.1)	Microarray analysis	http://bioconductor.org/biocLite.R biocLite()	(Gentleman *et al*., 2005)
Basic local alignment search tool (BLAST)	Heuristic alignment of query sequence to sequence database	AgResearch bioinformatics	(Altschul *et al*., 1997)
Cath	Identification of sequence and structure based motifs and folds	Incorporated in the interpro search engine	(Pearl *et al*., 2003)
Chromas (v 1.45)	Viewing sequence chromatograms	http://www.technelysium.com.au/chromas14x.html	(McCarthy, 2003)
Clustal X (v 2)	Global sequence alignment	ftp://ftp.ebi.ac.uk/pub/software/clustalw2.	(Thompson *et al*., 1997)
Cluster of orthologous genes (COG)	Functional classification based on orthollogous relationships	Incorporated in GAMMOLA and interpro search engines	(Tatusov *et al.*, 2000, Tatusov *et al*., 1997)
Dialign	Pair-wise sequence alignment	http://www.gsf.de/biodv/dialign.html	(Morgenstern, 2004)
Einverted (EMBOSS)	Identification of DNA secondary structures	http://www.interactive-biosoftware.com/embosswin/embosswin.html.	(Olson, 2002)
GAMMOLA	Automated genome sequence annotation	Supplied by the programmer Eric Altermann.	(Altermann & Klaenhammer, 2003)
Gene 3D	Identify structural relationships of query amino acid sequence to CATH database	Incorporated in the interpro search engine	(Buchan *et al.*, 2003, Yeats *et al.*, 2006)
GLIMMER	Gene prediction	Supplied as part of the GAMMOLA package	(Salzberg *et al*., 1998)
Interpro	Amino acid sequence analysis	AgResearch bioinformatics	(Apweiler *et al*., 2001)
LipoP	Identifies signal peptidase II	Incorporated in the MANATEE, GAMMOLA and interpro search engines	(Juncker *et al*., 2003)
MANATEE	Genome sequence viewing and analysis	http://manatee.sourceforge.net/downloads.shtml.	(TIGR, 2001)

Software	Application	Source	Reference(s)
Molecular biological (MB) DNA analysis	DNA analysis	http://www.molbiosoft.de/html/downloads.htm	(Simakov, 2006)
MOD Base	Comparative protein modelling	Incorporated in the interpro search engine	(Pieper *et al*., 2006)
Macromolecular structure database (MSD)	Comparative protein modelling	Incorporated in the interpro search engine	(Golovin *et al*., 2004)
Panther	Protein functional domain serching	Incorporated in the interpro search engine	(Thomas *et al*., 2003)
Prints-S	Searches sets of protein motifs	Incorporated in the interpro search engine	(Attwood *et al*., 2000)
Protein information resource superfamily (PIRSF)	Compares 1° protein sequence to protein superfamily database	Incorporated in the interpro search engine	(Wu *et al*., 2004)
Protein families (Pfam)	Searches for motifs in 1° protein sequence	Incorporated in the MANATEE, GAMMOLA and interpro search engines	(Sonnhammer *et al*., 1997)
Protein domains (ProDom)	Protein domain identification	Incorporated in the interpro search engine	(Corpet *et al*., 1998)
Prosite	Identifies protein domains, families and functional sites	Incorporated in the interpro search engine	(Bairoch, 1991)
R (v 2.6.2)	Statistical programming environment	http://cran.stat.auckland.ac.nz/	(Ihaka & Gentleman, 1996)
rJava (v 0.5-1)	Interface between java and R	http://cran.r-project.org/web/packages/rJava/index.html	(Urbanek, 2007)
RNAfold	Prediction of DNA and RNA 2° structures	http://www.tbi.univie.ac.at/~ivo/RNA/index.html	(Hofacker *et al*., 1994)
Structural classification of proteins (Scop)	Classifies proteins based on structural modelling	Incorporated in the interpro search engine	(Murzin *et al*., 1995)
Sequin	Submission of sequences to Genbank	http://www.ncbi.nlm.nih.gov/Sequin/download/seq_win_download.html	(Benson *et al*., 2008)
SignalP (v 3.0)	Predicts signal peptidase I cleavage sites	Incorporated in the MANATEE, GAMMOLA and interpro search engines	(Bendtsen *et al*., 2004)

Software	Application	Source	Reference(s)
Simple modular architecture research tool (Smart)	Identifies protein domains in 1° protein sequence	Incorporated in the interpro search engine	(Ponting *et al.*, 1999)
Splitstree (v 4)	Phylogenetic tree analysis	http://www.splitstree.org/	(Huson, 1998)
Staden (v 1.6.0)	DNA sequence assembly	http://sourceforge.net/project/showfiles.php?group_id=100316andpackage_id=107815andrelease_id=361767	(Staden, 1996, Staden *et al.*, 2000)
Superfamily	Compares HMM"s to database of proteins with known structure	Incorporated in the interpro search engine	(Gough *et al.*, 2001)
Swiss Model	Compares proteins based on structure modelling	Incorporated in the interpro search engine	(Peitsch, 1996)
The institute of genomic research protein families (TIGRfam)	Searches for HMM"s in 1° protein sequence	Incorporated in the MANATEE, GAMMOLA and interpro search engines	(Haft *et al.*, 2001)
Tinn-R (v 1.17.2.4)	Programming editor for R	http://www.sciviews.org/Tinn-R/.	(Faria & Grosjean, 2005)
TmHMM	Predicts transmembrane spanning regions in 1° protein sequence	Incorporated in the MANATEE, GAMMOLA and interpro search engines	(Krogh *et al.*, 2001)
tRNA scan-SE	Search DNA sequence data for tRNAs	http://lowelab.ucsc.edu/tRNAscan-SE/	(Lowe & Eddy, 1997)
Universal protein resource (Uniprot)	Protein sequence analysis	Incorporated in the interpro search engine	(Apweiler *et al.*, 2004)
Vector NTI (v 9)	DNA sequence display and analysis	http://informaxinc.com/site/us/en/home/Products-and-Services/Applications/Cloning/Vector-Design-Software.html (Dynamic license purchased by AgReseach)	(InforMax, 2001)

2.2 Methods

2.2.1 Growth Conditions

All *Butyrivibrio* and *Pseudobutyrivibrio* species, including *B. proteoclasticus,* were cultured anaerobically in 15 ml Hungate Tubes containing 10 ml of Medium 704 (*Butyrivibrio* spp. Medium, (DSMZ, 1993) at 39 ºC unless otherwise specified. Where possible, cultures were continued for short periods before fresh cultures were revived from frozen stocks to avoid *in-vitro* culture-biased evolution (Papadopoulos *et al*., 1999). *Methanobrevibacter ruminantium* cultures were grown anaerobically in BY+ media as described below (2.2.11.1). *Escherichia coli* cells were grown on LB plates (Bertani, 1951, Bertani, 2004) at 37ºC, unless otherwise stated.

2.2.2 Culture purity

Culture purity was verified by wet mounts and Gram-stains. *Methanobrevibacter ruminantium* wet mounts were additionally analysed by fluorescence microscopy. (Typical results are shown Appendix I, Fig. A1 and Fig. 6.3).

2.2.2.1 Wet mounts

Wet mounts were made using a modification of the method described by Pfennig and Wagener (1986). Slides were coated 2ml of prewashed sterile 2% (w/v) agarose. A drop (~25 µl) of the culture was extracted using a sterile 1 ml syringe and needle, and dispersed upon the surface of the slide. A coverslip was laid firmly across the culture and the resulting slide was subsequently examined under a microscope.

2.2.2.2 Gram stain

Gram stains were performed as originally described by Gram (1884). A small volume (~100 µl) of the culture was extracted using a sterile 1 ml syringe and needle, spread on the surface of a glass slide and allowed to dry next to an ignited Bunsen burner. Once dry, the slide was passed through the flame of the Bunsen burner several times to heat-fix the cells. The slide was stained sequentially with Crystal violet solution (10% (w/v) in ethanol) for 1 min, Iodine solution (0.3% (w/v) iodine and 0.7% (w/v) potassium iodine in water) for 1 min, acetone until the slide was decolourised and finally with Safranine solution (2.5% (w/v) in ethanol). Each treatment step was

followed by rinsing with water. The slide was blotted dry with tissue paper and viewed using a microscope.

2.2.3 Growth curves

Triplicate 500 µl samples of fresh overnight culture were used to inoculate Hungate Tubes containing 9.5 ml of fresh, pre-warmed, sterile M704 medium (5% inoculum). The initial optical density (OD) was taken using an Ultrospec 1100 Pro spectrophotometer (GE Healthcare, City State Country) at a wave length of 600 nm as previously described (Lambrecht, 1966, Beard *et al*., 1995). Cells were incubated at their optimal temperatures (unless otherwise specified), which for *Butyrivibrio* and *Pseudobutyrivibrio* species is approximately 39 ºC, consistent with the temperature of the rumen environment (Dale *et al*., 1954). OD_{600} was measured at regular intervals until stationary phase was reached. For *B. proteoclasticus* $B316^T$ and *M. ruminantium,* total cell numbers were also determined by Thoma slide counts.

2.2.3.1 Thoma slide counts

The edges of a coverslip were moistened with sterile dH_2O and pressed firmly onto a Thoma slide covering the counting chamber. Cultures were briefly mixed using an L46 vortex (Labinco Breda, Netherlands). A small volume was removed using a sterile 1 ml syringe and introduced to the edge of the coverslip and allowed to diffuse beneath it until the counting chamber was saturated. Cells were then viewed and enumerated using a Reichert Diavar microscope. Cells were counted from a minimum of 15 secondary squares, 5 each from, at least, 3 primary squares of the Thoma slide or until more than 200 cells had been counted.

The total cell number was estimated using the following equation:

$$\text{Cell density (cells / ml)} = \frac{\text{Number of cells counted}}{\text{Volume of Grid analysed (ml)}}$$

Each secondary square upon the grid was 2.5×10^{-3} mm^2 and 0.02 mm deep giving it a total volume of 5×10^{-5} mm^3. As 1 ml is equivalent to 1 cm^3 = 1000 mm^3, therefore the volume of each secondary square analysed was 5×10^{-8} ml.

2.2.4 Pulsed-field gel-electrophoresis (PFGE)

2.2.4.1 DNA extraction for PFGE

DNA was extracted using a modification of a procedure described by Thomas Eckhardt (1978). Most modifications of the procedure were introduced to circumvent an apparent non-specific endonuclease activity associated with several *Butyrivibrio* species. Cultures were grown to approximately mid-exponential phase ($\Delta OD_{600=}$~ 0.4) and cultures were heated to 70 ºC for 10 mins to destroy heat-sensitive nucleases. Cells were transferred to 1.5 ml Eppendorf tubes and harvested by centrifugation at 5,000 × *g* for 5 min using a MiniSpin Plus Personal microcentrifuge. The supernatant was discarded and the harvested cells were subsequently resuspended in 1 ml CE buffer (Martin & Dean, 1989) and harvested again by centrifugation at 5,000 × *g* for 5 min. This CE buffer wash was repeated and the cells were washed once in Wash solution before being resuspended in 150 µl of Wash solution. The cell suspension was mixed 1:1 with molten 2% pulsed-field certified low melt agarose (Bio-Rad) in 0.125 M EDTA that had been equilibrated to 50ºC. The resulting cell solution was dispensed into 10-well reusable plug moulds (Bio-Rad) and allowed to solidify at room temperature. Once solidified, plugs were transferred to universal bottles (up to 4 per universal) containing 4 ml of EC buffer containing 1 mg/ml Lysozyme. Plugs were left on an Orbitech XL flask shaker for 18 to 24 h at 37 ºC to allow cell lysis to occur. The EC buffer-lysozyme solution was replaced with 4 ml of 0.5 M EDTA-Sarkosyl solution containing 0.5 mg/ml of Proteinase K. Plugs were incubated on the flask shaker for 16 to 24 h at 37 ºC. Fresh EDTA-Sarkosyl/Proteinase K solution was added and the plugs were again incubated on the flask shaker for 16 to 24hr at 37 ºC. The EDTA-Sarkosyl/Proteinase K solution was replaced with TE buffer (10/1) containing 1 mM PMSF to eliminate residual Proteinase K activity. Plugs were incubated in this solution on the flask shaker for 2 h at 37 ºC. Plugs were then washed once in TE buffer (10/100) and stored at 4 ºC in the same solution.

2.2.4.2 Restriction Endonuclease digestion of pulsed-field gel plugs

The appropriate Pulsed-Field plugs were cut to ~2.5 mm using a sterile scalpel and carefully transferred to a sterile Eppendorf tube containing 1 ml of TE buffer (10/0.1) and allowed to equilibriate for 1 h. The TE buffer (10/0.1) was removed using a 1 ml pipetter and sterile 1 ml pipette tip. The plug was immersed in an excess of the

appropriate RE buffer and, where necessary, BSA and allowed to equilibriate for 1 h. The RE buffer was removed and a fresh 100 µl of the same solution was added along with 0.2 µl of the appropriate RE. The plug was incubated in this solution at the RE‟s optimal temperature for 16 to 24 h. The RE buffer was again removed and 500 µl of TE buffer (10/100) was added and kept at 4 ºC until analysis by Pulsed-Field Gel-Electrophoresis.

2.2.4.3 Pulsed-field gel-electrophoresis

Pulsed-Field Certified agarose (Bio-Rad) was mixed with 0.5 × TBE buffer to give a final concentration of 1%. The solution was boiled to dissolve the agarose and then allowed to cool to 50 ºC in a Grant Y28 waterbath (Grant Cambridge, England). Gels were cast in either a 14 × 13 cm casting stand (Bio-Rad) with a 1.5 mm thick 10-well comb (Bio-Rad), or a 20.25 x 14 cm casting stand (Bio-Rad) with a 1.5 mm thick 15-well comb (Bio-Rad). Occasionally the 20.25 × 14 cm casting stand was cast in the opposite orientation using the 1.5 mm thick 10-well comb (Bio-Rad). The gel was allowed to cool to room temperature and solidify. Agarose-plugs were cut to ~2.5 mm using a sterile scalpel and carefully transferred to the wells of the solidified gel along with an appropriate DNA size ladder. The plugs were embedded in the gel with molten 1% Pulsed-Field Certified agarose in 0.5× TBE buffer. The gel was transferred to a CHEF Electrophoresis cell (Bio-Rad) containing 2 L of 0.5% TBE buffer. The TBE buffer within the Chef Electrophoresis cell was maintained at 14 ºC by circulation through a Model 1000 Mini Chiller (Bio-Rad) using a variable speed pump (Bio-Rad) set at 95%. The DNA fragments were separated using a Chef-DRIII® system applying 5.5 volts/cm at 120º angles. Pulse switch times and electrophoretic duration were dependent upon the expected size range of the DNA fragments.

2.2.5 Genome sequencing and gap closure

The genome of *B. proteoclasticus* B316^T was sequenced to a theoretical genome coverage of 9 fold, as part of a genome sequencing project undertaken by the Rumen Microbial Genomics team at AgResearch. DNA libraries were constructed using a random shotgun cloning approach as previously described (Fleischmann *et al*., 1995). Genomic DNA was randomly disrupted and separated by agarose gel-electrophoresis. DNA fragments in the 2 to 4 Kb range were isolated and used to generate a small insert plasmid library, while fragments of ~40 Kb were isolated and used to produce a fosmid library (Kim *et al*., 1992). Clones resulting from both small and large insert libraries were recovered and sequenced using high throughput Sanger sequencing in capillary DNA sequencing machines (Agencourt Biosciences Corporation, MA, USA). DNA sequences were aligned to find overlaps and assembled into 313 contiguous sequences (contigs) using the STADEN software package (Staden, 1996). Physical and sequencing gaps were closed as described in chapter 3 using various protocols decribed below.

2.2.5.1 DNA extraction

B. proteoclasticus genomic DNA was extracted using the method described by Saito and Miura (1963). Cultures (10 ml) were harvested at mid-exponential phase (OD_{600}= ~0.4) by centrifugation at 5,000 × *g* for 5 min. The cell pellet was resuspended in 500 µl Saline-EDTA solution and harvested by centrifugation as above. The cell pellet was resuspended in 500 µl of Saline-EDTA solution containing 1 mg/ml Lysozyme and 20 µg/ml RNaseA. This solution was incubated for >1 h at 37 ºC to allow the Lysozyme to facilitate cell lysis, and the RNaseA to subsequently degrade RNA. SDS solution (27 µl) was added along with 11 µl of Proteinase K solution to give a final concentrations of 1% (w/v) and 200 µg/ml, respectively. The solution was incubated for 1.5 h at 60 ºC to allow the Proteinase K to degrade cellular protein and inactivate nucleases. TE buffer (10/1) (200 µl) was added and the organic lysate was removed by Phenol:Chloroform extraction.

2.2.5.2 Phenol:Chloroform extraction

One volume of room temperature buffer-satrated phenol was added to the solution containing the nucleic acid to be purified. The solution was mixed by inversion of the tube before precipitated protein and other organic material was pelleted by centrifugation at 13,000 × *g* for 5 min. The supernatant containing the nucleic acids was transferred to a fresh Eppendorf tube and mixed with an equal volume of Phenol:Chloform:Isoamyl Alcohol mixture and centrifuged at 13,000 × *g* for 5 min. The supernatant was transferred to a fresh Eppendorf tube and mixed with an equal volume of chloroform to remove residual phenol. After centrifugation at 13,000 × *g* for 5 min, the supernatant was transferred to a fresh Eppendorf tube and concentrated by ethanol precipitation, as described below (2.2.5.4).

2.2.5.3 Nucleic acid quantification

Nucleic acids concentration and purity was determined by spectrophotometrically using the NanoDrop ® ND-1000. Absorbances were determined at λ=230, 260 and 280 and nucleic acid concentration was determined using the Beer-Lambert equation modified to use an extinction coefficient of 50 ng-cm/ml for DNA and 40 ng-cm/ml for RNA. The purity of the nucleic acid samples were determined by analysis of the A_{260}/A_{280} and A_{260}/A_{230} ratios, where an A_{260}/A_{280} ratio of ~1.8 for DNA or ~2.0 for RNA indicates a pure sample, where a higher ratio indicate protein or Phenol contamination and where a lower ratio indicates copurified contaminants such as polysaccharides.

2.2.5.4 DNA concentration

Nucleic acids to be concentrated were mixed with 0.1 volumes 5 M Ammonium Acetate and 3 volumes of absolute Ethanol. The mixture was left for 16 to 24 h at -20 ºC and then centrifuged at 16,000 × *g* for 30 min at 4 ºC. The supernatant was carefully removed using a pipette to avoid disturbing the pellet. The pellet was washed in 200 µl of 70% Ethanol and centrifuged again at 16,000 × *g* for 15 min at 4 ºC. The supernatant was again carefully removed using a pipette and the Eppendorf tube was inverted and allowed to dry at 37 ºC for 30 min. The pellet was then dissolved in an appropriate volume of either sterile dH_2O, TE buffer (10/1) or 10 mM Tris-HCl, depending on the subsequent application.

2.2.5.5 Primer design

Primers were designed using Vector NTI (InforMax, 2001) at least 75 bp from the end of a contiguous sequence in high quality regions (> Phred 40). Where possible primer-design was constrained to fulfil the following criteria:

- 50 ºC $< T_m <$ 65 ºC
- Primer length 18 – 24 bp
- 40 < GC % < 60
- No more than 2 nucleotide repeats
- No secondary structures
- No palindromes

  } Particularly in the 3‟ region of each primer

- No sequence identity between primer pairs
- <5% difference in %GC between primer pairs
- <5% difference in T_m between primer pairs

The primers designed were compared to the most recent version of the *B. proteoclasticus* genome using BLASTN to ensure specificity and synthesized at 50 nM scale by Invitrogen. Primers were supplied as desalted and lysophilized pellets. Primers were redissolved at a concentration of 100 µM in either sterile MilliQ water for quantitative real time PCR (qPCR) primers or in TE buffer (10/1) for all other applications. Redissolved primers were stored at -20 ºC.

2.2.5.6 Polymerase chain reactions

Polymerase chain reactions (PCRs) were performed in 50 µl reaction volumes using a Touchdown PCR protocol, to limit the need for reaction optimisation. Constituents of the reactions (listed Table 2.18) were added to dH_2O in a 0.2ml thin wall PCR tube (Quality Scientific Plastics), in the order listed, on ice. The solution was mixed using a micropipetter and subsequently transferred to a PX2 Thermal Cycler.

The reaction was then cycled through a 3 min denaturation at 95 ºC, followed by 10 cycles of:

- 94 ºC for 45s
- 65 – 55 ºC for 45s (temperature decreasing by 1 ºC per cycle)
- 72º for 2 min

The reaction was then cycled through 20 further cycles of:

- 94 ºC for 45s
- 55 ºC for 35s
- 72 ºC for 2 min

The reaction then underwent a 5 min elongation step at 72 ºC and was subsequently cooled to 4 ºC until examined by agarose gel-electrophoresis, as described below (2.2.5.12).

Table 2.18 PCR master mix

Chemical	Concentration
10× PCR buffer	1×
DNA	1 µg
dNTPs	0.2 mM
$MgCl_2$	2.5 mM
Platinum® Taq	1.25 units
Each primer	5 µM

2.2.5.7 Multiplex PCR

Multiplex PCR Primers were designed as described above (2.2.5.5) and labeled using the notation shown below (Fig 2.1):

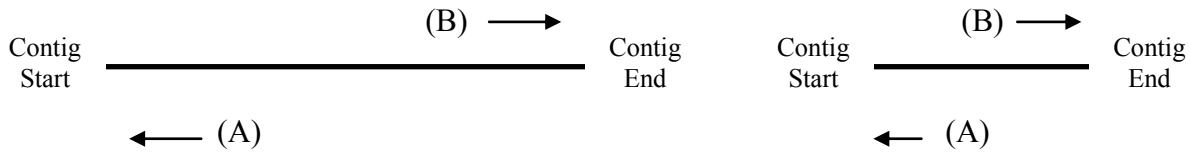


Figure 2.1 Multiplex PCR notation.

The primers for one contig end (e.g. p300c107a) were combined with an equimolar amount of each opposing primer (e.g. p300c108b, p300c109b, p300c110+b and p300c112b) and run through 30 cycles of touchdown PCR, as described above (2.2.5.6). This process was repeated for both ends of each contig.

2.2.5.8 Inverse PCR

Inverse PCR (iPCR) was completed using a protocol adapted from Ochman *et al*. (1988). RE recognition sites located within the terminal 3 Kb of each contig (or super contig) end were determined using Vector NTI software. Several REs, most with low %G + C, heximeric recognition sites were chosen so that resulting fragments were likely to be less than the amplification limits of conventional PCR. Primers were then designed, divergently, to both the terminal end of the contig (or supercontig) and as close as possible to the RE recognition site. Both primers were also designed to ensure they were not within 75bp of the contig end or the RE recognition site to ensure that sequence generated from the priming site to the contig terminus or the RE site could be recognised and trimmed. Inverse PCR primers that faced towards the contig end were labeled using the same notation as described above (2.2.5.7) and in many cases were the same primers designed for multiplex PCR. Primers that faced towards the known restriction site were labeled as shown below (Fig 2.2):

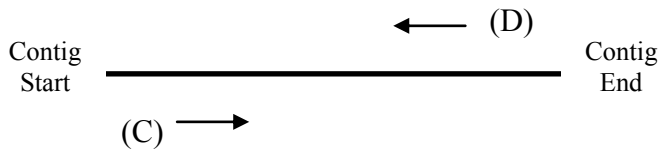


Figure 2.2 iPCR notation.

Genomic DNA (20 µg) was digested with the appropriate RE as described below (2.2.5.9) to produce each iPCR library. The solution was then concentrated to a total volume of 20 µl by ethanol precipitation. The resulting digested DNA was then re-circularized by DNA ligation as described below (2.2.5.10). The resulting re-circularized DNAs carrying the fragments of interest were subsequently amplified by PCR as described above (2.2.5.6). The resulting products were analysed by agarose

gel-electrophoresis as described in Section 2.2.5.12. Where possible, the resulting amplicons for each contig end were purified, (Section 2.2.5.13 or 2.2.5.14), and sequenced (Section 2.2.5.15).

2.2.5.9 Restriction Endonuclease digestion of DNA

Reaction constituents (listed Table 2.19) were combined, in the order listed, in sterile dH_2O to give a total volume of 50 µl. The reaction was mixed and incubated at the RE‟s optimal temperature (See suppliers instructions; New England Biolabs) for 12 to 16 h. The reaction was then heated to 65 ºC for 20 min to denature the RE and halt the reaction.

Table 2.19 RE reaction constituents

Chemical	Concentration
10× RE buffer	1×
DNA	* x µg
BSA	1× (as required)
Restriction enzyme	10 units

* The amount of DNA used in each digestion was typically 1 to 20 µg

2.2.5.10 DNA ligation

DNA ligation reactions were performed using Ready-To-Go™ T4 DNA Ligase (GE Healthcare). The DNA to be ligated (20µl), was suspended in either dH_2O or TE buffer (10/1), was added to a tube of Ready-To-Go™ T4 DNA Ligase (GE Healthcare). Each tube contains 6 Weiss units of T4 DNA ligase, 1× T4 DNA Ligase buffer and 0.1 mM ATP. The reaction was incubated at 16 ºC for 12 to 16 h and subsequently heated to 70 ºC for 10 min to inactivate the T4 DNA Ligase.

2.2.5.11 Long range PCR

Longrange PCR was undertaken using the Eppendorf TripleMaster system. The reaction was prepared as described in Section 2.2.5.6 using constituents described below (Table 2.20). The reaction was then cycled through a 3 min denaturation at 93 ºC, followed by 10 cycles of:

- 93 ºC for 15 s
- 65 to 55 ºC for 30 s (temperature decreasing by 1 ºC per cycle)
- 68 ºC for 21 min

The reaction was then cycled through a further 25 cycles of:

- 93 ºC for 15 s
- 55 ºC for 30 s
- 72 ºC for 21 to 29 min (time increasing by 20 s per cycle)

The reaction was subsequently cooled to 4 ºC until examined by agarose gel-electrophoresis.

Table 2.20 Long range PCR reaction constituents

Chemical	Concentration
10× Tuning buffer	1 ×
DNA	2 µg
dNTPs	0.5 mM
Triple master enzyme mix	2 units
Each primer	10 µM

2.2.5.12 Agarose gel-electrophoresis

Agarose (Invitrogen, Carlsbad, CA, USA) or Low Melt Agarose (Progen, Heidelburg, Germany) was mixed with 1× TAE buffer to give a final concentration of 1.5% or 1% (w/v) respectively. The solution was boiled to dissolve the agarose and allowed to cool to 50 ºC in a Grant Y28 waterbath (Grant, Cambridge, England). Gels were cast in a 15 x 10 cm casting stand (Bio-Rad) with either a 1.5 mm thick 15-well comb (Bio-Rad), or a 1.5 mm thick 20-well comb (Bio-Rad). Gels were allowed to cool to room temperature and solidify, after which the gel tray was transferred to a wide Mini-sub® Cell GT (Bio-Rad) and immersed in 1× TAE buffer. Unless otherwise stated, 9 µl of each PCR-amplified product was placed on to a small square of Parafilm M and combined with 1 µl of 10× BlueJuice™ (Invitrogen, Carlsbad, CA, USA) and carefully transferred to a separate well of the solidified gel. Additionally, 8 µl of 1Kb+ ladder (Invitrogen, Carlsbad, CA, USA) was added to unused wells of the

gel, typically bordering the sample wells. The DNA fragments were separated using a Powerpac Basic (Bio-Rad) applying 80 volts (5 $^{1}/_{3}$ volts/cm) for 60 min. The gel was removed from the gel apparatus and stained with Ethidium bromide solution for 40 min. The gel was washed in water for 10 min and analysed by ultraviolet light transillumination at λ=590nm and photographed using a Gel Logic 200 system.

2.2.5.13 Agarose gel elution

The DNA sample to be recovered was subjected to agarose gel-electrophoresis, as described in Section 2.2.5.12 using 1% low melt agarose. The DNA fragment of the correct size was cut from the gel using a sterile scalpel and transferred to a pre-tared Eppendorf tube and weighed (Explorer Scales, Ohaus Corporation, Pinebrook, NJ, USA). One µl of GELase buffer^{TM} (50× stock; Epicentre, Madison, WI, USA) was added for every 50 mg weight of the gel slice. The gel slice was melted by incubation at 70 ºC in a SHTD Test Tube Heater (Stuart Scientific) for 3 min for every 200 mg weight of the gel slice. The molten agarose was then transferred to a second test tube heater and equilibrated to 45 ºC for at least 2 min per 200 mg of the gel slice. Half a unit of GELase^{TM} enzyme (Epicentre) were added per 200 mg of the gel slice and the solution was incubated for 2-3 h. The DNA within the solution was extracted and concentrated, as described in Section 2.2.5.4 except 1 volume and 4 volumes of room temperature Ammonium acetate and Ethanol, respectively, were used and incubation and centrifugation were completed at room temperature to avoid the co-precipitation of oligosaccharides derived from digestion of the agarose gel.

2.2.5.14 PCR clean up

DNA fragments resulting from PCR reactions were purified using a QIAquick® PCR purification kit (QIAGEN, Hilden, Germany). PB buffer (5 volumes) was mixed with the product of each PCR reaction using a 1 ml micropipetter and sterile pipette tip. The solution was applied to a QIAquick spin column and centrifuged for 60 s at 13,000 × *g*. During this process, DNA of 0.1 to 10 Kb should bind to the column, while other constituents of the reaction should flow through. The flow–through liquid was discarded and 750 µl of PE buffer was applied to the column and centrifuged for 60 s at 13,000 × *g*. This washing step was repeated, discarding the flow–through liquid each time. The column was again subjected to centrifugation for an additional 1 min at 13,000 × *g* to remove any residual PE buffer. Sterile dH_2O (30 µl) was added

to the centre of the QIAquick column and this was incubated at room temperature for 5 min to enhance the elution of the purified DNA. The eluate was transferred to a fresh sterile Eppendorf tube by centrifugation at 13,000 × *g* for 60 s.

2.2.5.15 DNA sequencing

All DNA sequencing reactions, other than those described above for genome sequencing, were conducted by the Allan Wilson Centre Genome sequencing service. This service included fluorescent labeling of products using the BigDye™ Terminator Version 3.1, a Ready Reaction Cycle Sequencing kit cycle by sequencing PCR, subsequent removal of unincorporated fluorescent dideoxy NTPs (ddNTPs) by cleanup and precipitation of products and capillary separation on an ABI3730 Genetic Analyzer, (Applied Biosystems Inc., City State, USA). Results were returned as ABI tracefiles which were analysed using Chromas software.

2.2.6 Gene finding and annotation

ORFs were identified by GLIMMER 2.0 (TIGR, 1999, Salzberg *et al*., 1998, Delcher *et al.*, 1999) using a 90 bp minimum ORF size threshold. Ribosome binding sequences were predicted by RBS Finder (Suzek *et al*., 2001) and refined by manual inspection. Start codons were predicted (Suzek *et al*., 2001) and refined by analysis of their proximity to ribosome binding sequences as previously described (Shine & Dalgarno, 1975) and by compatibility with BLAST alignment data that minimized or eliminated overhanging query sequence. Assigned ORFs were analysed using BLASTP (Altschul *et al*., 1997), PFam (Sonnhammer *et al*., 1997), TIGRFam (Haft *et al*., 2001), COG (Tatusov *et al*., 2000, Tatusov *et al*., 1997), SignalP (Bendtsen *et al*., 2004), LipoP (Juncker *et al*., 2003) and TmHMM (Krogh *et al*., 2001) searches using the MANATEE (TIGR, 2001) and/or GAMOLA (Altermann & Klaenhammer, 2003) interfaces. Other Hidden Markov Models including Uniprot (Apweiler *et al*., 2004), Prosite (Bairoch, 1991), ProDom (Corpet *et al*., 1998), Smart (Schultz *et al.*, 2000, Ponting *et al*., 1999, Schultz *et al.*, 1998), Panther (Thomas *et al*., 2003), PIRSF (Wu *et al*., 2004), Superfamily (Gough *et al*., 2001), SCOP (Murzin *et al*., 1995), CATH (Pearl *et al*., 2003), Swiss Model, MOD Base (Peitsch, 1996), MSD (Golovin *et al*., 2004), Gene 3D and Sprint were detected by manual analysis of each ORF using Interpro (Apweiler *et al*., 2001). The following criteria were used to exclude ORFs from further analysis: ORFs smaller than 50 amino acids without significant match to

any previously described gene, no trusted hits to known HMMs, large deviation of the ORFs DNA sequence from the replicon‟s overall average G + C content, overlap of the ORF with a larger ORF. Putative functions were assigned to ORFs based upon the results of the bioinformatic analyses. In general, assigned functions were cautiously descriptive, and where possible the functions were identified by the appropriate Gene Ontologies and/or Enzyme Commission (EC) number(s). The designation “putative” was used to precede the description of ORFs that did not show significant similarity to a functionally characterised homolog. Significant similarity was deemed to be a BLASTP alignment with at least 30% sequence identity and 50% sequence similarity across at least 70% of both protein sequences. Where alignments of larger ORFs (>500 amino acids) were over full length, or near full length, the constraints of 30% sequence ID and 50% sequence similarity were relaxed marginally, although no less than 25% and 40% for sequence identity and sequence similarity, respectively. ORFs displaying no significant sequence similarity to any previously described gene were described as “hypothetical” proteins while those displaying significant sequence similarity to a gene of unknown function were described as “conserved hypothetical” proteins. Additional evidence for an ORF function were incorporated into the annotations during the course of this thesis. For example, where proteomic evidence demonstrated expression of a hypothetical protein, its annotation was upgraded to “uncharacterised” protein. The presence of a Signal Peptidase I cleavage site, as predicted by SignalP, resulted in the descriptor “secreted”, while a Signal Peptidase II cleavage site, predicted by LipoP, in the absence of other evidence, resulted in the descriptor ”lipoprotein”. The prediction of two or more transmembrane helices resulted in the descriptor “transmembrane”, while the prediction of an LPXTG motif resulted in the descriptor “membrane-bound”. ORFs were numbered clockwise from the ORF encoding the replication initiation protein RepB and nucleotides were numbered from the ribosome binding site of the RepB protein.

2.2.7 DNA sequence analysis

DNA secondary structures, including inverse-, inverted- and direct-repeats, were detected using Einverted (Olson, 2002, Rice *et al.*, 2000) and through alignment of each replicons sequence against itself using an in-house standalone BLASTN program (Altschul *et al*., 1997). The G + C content, and Codon Usage were determined using Artemis (Mural, 2000).

2.2.8 Amino acid sequence analysis

Protein isoelectric points were determined using MANATEE (TIGR, 2001) or a stand alone version of MB DNA analysis (Simakov, 2006).

2.2.9 Plasmid curing

Attempts to cure *B. proteoclasticus* of its co-resident replicons were based on previously described techniques successfully used in the curing of megaplasmids (Morrison *et al.*, 1983, Kojic *et al.*, 2005, Sadowsky & Bohlool, 1983). These included growth in M704 media at higher than optimal temperatures (Morrison *et al*., 1983) or addition of Novobiocin (Kojic *et al*., 2005), SDS (El-Mansi *et al*., 2000), Acriflavine (Mesas *et al*., 2004), Ethidium Bromide (Mattarelli *et al*., 1994) or Acridine Orange (Sadowsky & Bohlool, 1983). Fresh *B. proteoclasticus* cultures were grown, in triplicate, in increasing concentrations of each curing agent (or increased temperature) to identify the point at which the organism could no longer grow. With the exception of SDS, which failed to inhibit growth up to 5% (w/v), maximal sublethal concentrations were determined for each agent (or temperature). As each condition had a different affect on the *B. proteoclasticus* growth rate, generation time was determined for the maximal sublethal concentration of each condition. *B. proteoclasticus* cultures were grown for approximately 20 generations at the respective maximal sublethal concentration or temperature. Cultures were diluted 10^{-5} to 10^{-6} to give single colonies and 100 µl aliquots of cells were plated onto M704 Medium plates in an anaerobic chamber and grown for 24 to 36 h at 37 ºC. Colonies were replica plated and checked for the presence of each of the 4 replicons by colony hybridisation as described below (Section 2.2.9.1). Colonies which appeared to lack one or more of the replicons were subsequently verified by pulsed-field gel-electrophoresis, as described above in Section 2.2.4.

2.2.9.1 Colony hybridisation

Colony hybridisations were performed using a modified method of that described by Cocconcelli *et al*. (1991). Amersham Hybond-N+ nylon membranes (GE Healthcare, Uppsala, Sweden) were cut into circles to fit inside a 90 x 15 mm Petri dish (Biolab, Auckland, NZ). The dry membranes were gently placed onto the agar surface, containing the colonies of interest, and pressed lightly but evenly with a sterile L-

shaped plastic rod (Biolab, Auckland, NZ). After 1 min the membrane was removed with sterile forceps, and placed colony side down onto a fresh M704 plate and again pressed lightly but evenly with a sterile L-shaped plastic rod to give a replica plate. After a further minute the membrane was removed with sterile forceps, and placed colony side up onto Grade 1 filter paper (Whatman, Kent, UK) saturated with Colony Hybridisation lysis solution supplemented with 5 mg/ml lysozyme. The colonies were incubated for 5 to 10 min at room temperature and transferred to a series of separate blotting papers saturated with, (in order), 0.5% SDS solution for 3 min, Denaturation solution for 5 min, and two lots of Neutralisation solution for 5 min each. The membrane was then floated in 2× SSPE buffer for 5 min prior to being immersed in the same solution by gentle agitation. The released DNA was then dried and fixed to the membrane by baking for 2 h at 80 ºC in an oven.

2.2.9.2 Southern hybridisation

Amersham Hybond-N+ nylon membranes containing the DNA(s) of interest were probed for sequence similarity using the AlkPhos Direct labelling and detection system (GE Healthcare, Uppsala, Sweden). Probes were derived from PCR reactions, as described above (Section 2.2.5.6), using primers designed, as described above (Section 2.2.5.5), to target genes or loci of interest. PCR products were purified, as described in Section 2.2.5.14 and diluted to 20 ng/µl. Ten µl of the diluted PCR product was denatured by boiling for 5 min and rapidly cooling on ice for 5 min. The denatured DNA solution was labeled with alkaline phosphatase by incubation at 37 ºC with 10 µl of the reaction buffer, 2 µl of the labelling reagent and 10 µl of the Cross linker solution for 30 min. The membrane containing the DNA(s) of interest were pre-hybridised inside a hybridisation bottle (GE Healthcare, Uppsala, Sweden) at 62.5 ºC within a hybridisation Oven set at 60 rotations per min in 0.15 ml/cm^2 of pre-warmed hybridisation buffer containing 0.5 M NaCl and 4% (w/v) Blocking Reagent (GE Healthcare, Uppsala, Sweden) for 30 min. Following pre-hybridisation, 10 ng/ml of labeled probe was added and hybridisation was allowed to continue for 16 to 20 h at 62.5 ºC in a hybridisation Oven set at 30 rotations per min. The membrane was removed and washed with an excess volume of primary wash buffer pre-warmed to 62.5 ºC in a Shaking oven set to 30 strokes/min for 10 min. The membrane was then washed, as described, again in a fresh volume of primary wash buffer. The membrane was subsequently transferred to a fresh container and washed twice at room

temperature in secondary wash buffer on a shaker set to 30 strokes/min for 5 min. The membrane was covered in 30 to 40 µl/cm^2 of CDP-star Detection Reagent (Tropix Inc, Bedford, MA, USA) and incubated at room temperature for 5 min. The membrane was drained of excess Detection Reagent, wrapped in cling film and inside a darkroom was placed DNA side up in a film cassette with 20.3 x 25.4 cm BioMax XAR autoradiography film (Eastman Kodak Company, Rochester, NY, USA). The film was exposed, typically for 2 h, before being removed and developed by immersion in a tray of KODAK GBX developer and replenisher solution for 5 min. The film was then transferred to a tray of water and rinsed for 30 s with moderate agitation and transferred to a third tray filled with KODAK GBX fixer and replenisher solution and moderately agitated for 5 min. The film was finally returned to the tray of water and rinsed for 2 min with moderate agitation before being allowed to air dry. All development steps were performed at room temperature. Following satisfactory detection the membranes were stripped of probe by incubation in 0.5% (w/v) SDS solution at 60 °C for 60 minutes. The membranes were then rinsed in 100 mM Tris pH 8.0 for 5 minutes at room temperature, wrapped in cling film, and stored in at 4°C.

2.2.9.3 Culture storage

B. proteoclasticus strains devoid of pCY186 were stored in 10ml Hungate tubes containing 5 ml M704 media supplemented with 20% (v/v) glycerol at -85 ºC.

E. coli strains carrying shuttle vector constructs were stored using cryopreservation beads (Technical Service Consultants Ltd, Heywood, Lancashire, England) at -85 ºC.

2.2.10 Determining plasmid copy number

Using a method previously described by Lee *et al*. (2006), the copy number of each *B. proteoclasticus* auxiliary replicon was determined by absolute quantitative realtime PCR (qPCR). PCR primers were designed as described in Section 2.2.5.5 except they were designed to amplify a DNA region of no more than 1 Kb, as specified (Roche, 2003), for enumeration accuracy. qPCR primers were designed to amplify intergenic regions specific to each of the four replicons (BP_probe 1fp/BP_probe1rp, for the major chromosome; p360_probe1.bwp / p360_probe1.gwp for pCY360; p300_probe2.bwp / p300_probe2.gwp for BPc2; and p190_probe1.bwp / p190_probe1.gwp for pCY186; for primer details see Appendix II). Specificity of

each of the selected primer sets was tested by conventional PCR, as described in Section 2.2.5.6, and the resulting amplicons were screened by agarose gel-electrophoresis, as described in Section 2.2.5.12). The resulting amplicons were purified, as described in Section 2.2.5.14 and cloned into Topo-TA vectors and transformed into *E. coli* as described in Section 2.2.10.2. Single white colonies were picked and subjected to plasmid purification, as described in Section 2.2.10.3. The resulting plasmids (named Topo-BP_p1, Topo-pCY360_p1, Topo-BPc2_p2 and Topo-pCY186_p1 respectively) were screened for authenticity by PCR, using primers specific for their respective inserts and the PCR products were analysed by agarose gel-electrophoresis. The DNA concentration of each plasmid midi-prep was quantified (in triplicate) as described in Section 2.2.5.3. Using this information, the copy number of each TOPO-vector containing the respective sequence of interest was determined using the following equation (Whelan *et al*., 2003):

$$= \frac{6.02 \times 10^{23}(\text{copies/mole}) \times \text{DNA amount (grams)}}{\text{Total Vector length (bp)} \times 660\ (\text{grams/mole/bp})}$$

Each Topo-derived plasmid was subsequently diluted to give a stock solution containing 1×10^9 copies/µl. For each plasmid, a 10-fold dilution series ranging from 1×10^9 copies/µl to 1×10^5 copies/µl was made and used as a reference for Realtime qPCR. The PCR conditions for each primer set were optimised for temperature and MgCl concentration.

Quantitative PCR was performed on each Topo-derived plasmid, in triplicate, as described in Section 2.2.10.4, using its respective primers. The threshold cycle (C_T) was determined for each plasmid dilution and the C_T values plotted against the Log_{10} of the initial template copy number. The slope of the line fitted to these points, as determined by linear regression, was used to derive the qPCR amplification efficiency (*E*) from the following equation:

$E = 10^{-1/\text{slope}} - 1$ (Rasmussen, 2001)

Each Real-time qPCR was performed using a LightCycler (Roche Diagnostics, Bassel, Switzerland) and subsequent analysis was performed with LightCycler

software (Version 3.5; Roche Diagnostics, Bassel, Switzerland). The C_T was determined using the „Fit points" method, as described in the original protocol (Lee *et al*., 2006), with the „arithmetic" baseline adjustment, due to the background variation between the purified standard curves and the unknowns. The specificity of the qPCR reactions was determined by melting curve analysis that was performed from 65 – 95 ºC with a transition rate of 0.10 ºC / second and continuous fluorescence data acquisition using the LightCycler software (Version 3.5; Roche). To ensure the correct product was amplified, resulting qPCR reactions were analysed by Agarose gel-electrophoresis, as described (2.2.5.12).

2.2.10.1 Preparing electrocompetent cells

Electrocompetent *E. coli* cells were prepared as previously described (Sambrook & Russell, 2001). An overnight *E. coli* culture was streaked onto LB medium, to give discreet colonies, and incubated overnight at 37º C. A single *E. coli* colony was taken from the plate using a sterile loop and used to inoculate 50 ml of fresh sterile LB broth in a 250 ml conical flask. The culture was grown overnight at 37 ºC on an Orbitech XL flask shaker set to 250 rpm. Following incubation the culture was divided in half and used to inoculate two 500 ml aliquots of pre-warmed (37º C) LB broth, each in 2 L conical flasks. Both flasks were incubated at 37 ºC on an Orbitech XL flask shaker set to 250 rpm. OD was measured using an Ultrospec 1100 Pro spectrophotometer (GE Healthcare) at a wavelength of 600 nm initially and subsequently every 20 to 30 min until a ΔOD_{600} of 0.4 was reached. The flasks were transferred to a waterbath containing an ice/water slurry and incubated, with occasional swirling, for 30 min. The cultures were transferred to pre-chilled (-20ºC) sterile 250 ml centrifuge tubes and cells were pelleted by centrifugation at 1000 × *g* for 15 min at 4º C. The supernatant was discarded and cells were washed in 250 ml of pre-chilled (4ºC) sterile MilliQ water. The cells were again harvested by centrifugation, as described above for 20 min. The supernatant was discarded and cells were washed in 250 ml of pre-chilled (4º C) 10% Glycerol solution, as described above. The harvested cells were then resuspended in 10 ml of the 10% Glycerol solution and transferred and pooled in two pre-chilled (4º C) sterile Oakridge tubes. The cells were pelleted as described above and supernatant was carefully removed using a P200 micropipetter and sterile pipette tip. The pelleted cells were resuspended in 1 ml of ice-cold GYT Medium. A sample of the cells was diluted 1:100 and its OD

measured using the Nanodrop at λ=600nm. If the OD_{600} was significantly above 1.2 (optimal reading OD_{600} 0.8 – 1.2, equivalent to approximately 2.5×10^{10} +/- 0.5×10^{10} cells/ml) the suspended cells were diluted accordingly with GYT Medium. Cells were stored in 0.6 ml Eppendorf tubes in 40 µl aliquots at -85º C.

Electrocompetent *B. proteoclasticus* cells were prepared as previously described (Beard *et al*., 1995). *B. proteoclasticus* was inoculated into 40 ml M704 Medium in a 50 ml serum bottle and cultured for 12 to 16 h at 37º C. The culture was transferred to a waterbath containing an ice/water slurry and incubated, with occasional swirling, for 30 min. The cells were transferred to Oakridge tubes and pelleted by centrifugation at 5,000 × *g* for 5 min at 4º C. The supernatant was discarded and cells were resuspended in 40 ml of Electroporation buffer. The cells were again pelleted and then resuspended in 80 µl of Electroporation buffer. Aliquots (40 µl) were transferred to sterile 0.6 ml Eppendorf tubes and stored at -85º C. All steps, except centrifugations, were performed in an anaerobic chamber.

2.2.10.2 TOPO cloning

Purified DNA was combined with 1 µl of salt solution in 3 µl of sterile water in a 0.6 ml Eppendorf tube. The tube was briefly vortexed and then centrifuged. One µl of the Topo vector was added directly to the mixture and the ligation reaction was incubated at room temperature for 5 min. The reaction was placed on ice and 20 µl of OneShot® TOP10 chemically competent *E. coli* cells (Invitrogen) were added. The cell/Topo-vector mixture was incubated on ice for 20 min and heat-shocked at 42 ºC in a test tube heater for 30 seconds. SOC medium (250 µl at room temperature) was immediately added to the cells and transferred to a sterile 15 ml Falcon tube and incubated for 1 h on a flask shaker set to 250 rpm at 37ºC. Aliquots (100 µl) of the cultures were plated on LB medium containing 50 µg/ml ampicillin and 40 mg/ml X-Gal. Plates were incubated over night at 37 ºC and resulting white colonies were selected.

2.2.10.3 Plasmid mini preparations

Plasmid mini preparations were carried out as previously described (Sambrook & Russell, 2001). Two ml of the appropriate medium (LB for *E. coli*, M704 for *B. proteoclasticus*) containing 10 µg/ml of chloramphenicol in a 15 ml Falcon tube was inoculated with the transformed bacteria. The culture was incubated at 37 ºC

overnight on a flask shaker set to 200 rpm and 1.5ml of the culture was transferred to a sterile Eppendorf tube and cells pelleted by centrifugation at 13,000 × *g* for 1 min at 4º C. The supernatant was removed using a P200 micropipetter and sterile pipette tip. The cell pellet was washed in 375 µl of pre-chilled (4 ºC) STE buffer and mixed with 100 µl of pre-chilled Alkaline lysis solution I using a vortex mixer set to maximum. The mixture was combined with 200 µl of room temperature Alkaline Lysis solution II and mixed by several rapid inversions of the Eppendorf tube. Pre-chilled (4º C) Alkaline lysis solution III (150 µl) was added and mixed by several rapid inversions. The tube was incubated on ice for 4 min and the precipitated organic matter was pelleted by centrifugation at 13,000 × *g* for 5 min at 4º C. The supernatant was transferred to a fresh Eppendorf tube containing 500 µl of phenol and subjected to phenol:chloroform extraction and the plasmid DNA subsequently concentrated as described in Section 2.2.5.4.

2.2.10.4 Realtime qPCR

Realtime qPCR reactions were performed in 20 µl volumes using the Roche Diagnostics (Basel, Switzerland) Fast Start Kit. Constituents of the reaction were added in the order listed (Table 2.21) to sterile MilliQ water in a sterile 0.6 ml Eppendorf tube on ice. The solution was mixed using a microopipettor and a sterile pipette tip and subsequently transferred to a Lightcycler® capillary. The capillaries were centrifuged at 700 × *g* for 15 s and transferred to a Lightcycler II Realtime PCR Cycler (Roche Diagnostics, Basel, Switzerland). The reaction was cycled through a 10 min denaturation at 95º C, followed by 40 cycles of:

- 95 ºC for 15 s
- 60 ºC for 15 s
- 72 ºC for 35 s
- Fluorescence measurement

The reaction was cooled to 65 ºC and progressively heated to 95 ºC with fluorescence measurements being taken every 0.1 ºC increase for melting curve analysis. The reactions were cooled to 4 ºC until examination by agarose gel-electrophoresis, as described in Section 2.2.5.12. Data acquired from the fluorescence measurements were analysed using the LightCycler software (Version 3.5) (Roche Diagnostics,

Basel, Switzerland) and exported to a Microsoft Excel Spreadsheet to determine copy numbers.

Table 2.21 Realtime qPCR reaction constituents

Chemical	Concentration
FastStart DNA Master SYBR Green mix	2 µl
dNTPs	0.2 mM
$MgCl_2$	* x mM
DNA	** 2 µl
Each primer	5 µM

* An optimised concentration of $MgCl_2$ was used.

** Concentration varied, however a volume of 2 µl of each standard and unknown was used

2.2.11 Co-culture vs Mono-culture microarray analysis

2.2.11.1 Growth conditions

M. ruminantium cultures were grown anaerobically at 39 ºC on an orbital shaker set to 100 strokes/min in BY+ medium (Section 2.1.9.1). Cultures were pumped to 25 psi with 80% H_2:20% CO_2 gas initially and again when H_2 was utilized, as determined by pressure testing with a 5 ml syringe and headspace analysis by gas chromatography. The growth curves of *M. ruminantium* and *B. proteoclasticus* cultures were determined by Thoma slide enumeration, as described above (2.2.3). Cultures were grown to mid-exponential phase and then total RNA was extracted as described below (2.2.11.4). This phase of growth was deemed to be the most appropriate as not only is this point where RNA is thought to be most abundant, but due to the dynamic nature of the rumen environment, stationary phase is unlikely to occur, at least for any significant length of time. For the microarray experiment, 18 *M. ruminantium* cultures were divided evenly into six lots each comprising three 100 ml cultures. Two days following the second pumping of cultures with the hydrogen:carbon dioxide gas, estimated to be approximately mid-growth phase, all *M. ruminantium* cultures were flushed with O_2-free CO_2 gas until hydrogen was not detectible by gas chromatography (Appendix III, Fig. A3) and then supplemented with 4 ml of 5% Xylan solution to give a final concentration of 0.2%. Half of the cultures were then inoculated with 0.5 ml of a late exponential phase *B. proteoclasticus* culture and half were pumped to 25 psi with 80% H_2:20% CO_2 gas. A further three 100 ml volumes of

pre-warmed (39ºC) BY+ medium supplemented with 0.2% Xylan solution were also inoculated with 0.5 ml of the late exponential phase *B. proteoclasticus* culture. The inoculations and processing of each replicate set (1 set = 3x 100 ml *M. ruminantium* mono-cultures, 1x 100 ml *B. proteoclasticus* mono-culture *and* 3x 100 ml *M. ruminantium* inoculated with *B. proteoclasticus* co-cultures) were staggered by two hours so RNA was extracted at approximately the same phase of growth. Mid-exponential phase was estimated by the growth curve (Appendix III, Fig. A3) to occur after approximately 510 min (8½ h). To ensure growth was within the expected parameters, Thoma slide counts were taken after 5, 7 and 8 h, as well as immediately prior to extraction.

2.2.11.2 Microarray analysis

All materials used in the extraction and subsequent processing of RNA were treated to eliminate RNases, as described in Section 2.2.11.3. The RNA was purified using an RNeasy midi kit (Qiagen, Hilden, Germany), as described in Section 2.2.11.5, to remove DNA, xylan and any other contaminants. The RNA quality was assessed using the Bioanalyzer RNA 6000 nano assay, as descibed in Section 2.2.11.6. The RNA was transcribed using the SuperScript™ Indirect cDNA Labeling system (Invitrogen, Carlsbad, CA, USA) following the manufacturer‟s instructions (see Section 2.2.11.8). The resulting cDNA was purified (Section 2.2.11.9). The relative quantities of RNA contributed by each organism to the co-culture samples were then determined by qPCR (as described in Section 2.2.11.10). qPCR primers were designed, using methods described in Section 2.2.10, against the *B. proteoclasticus* butyryl-CoA dehydrogenase (bcd) gene (using primers bcdqfp and bcdqrp), and the *M. ruminantium* gene encoding N^5,N^{10}-methenyl-H_4MPT cyclohydrolase (mch) (using primers mchqfp and mchqrp). Homologues of both genes have previously been shown to be constitutively expressed in closely related species (Reeve *et al.*, 1997, Asanuma *et al.*, 2005). The mono-culture RNAs were then combined appropriately using the schedule shown in Appendix III (Table A3). The cDNAs were concentrated as described in Section 2.2.11.9 and labeled as described in Section 2.2.11.11. The labeled cDNAs were appropriately combined (the co-culture sample was labeled with Cy3 and combined with non-co-culture sample labeled with Cy5, and vice-versa for each of the three biological replicates) and hybridised to the microarrays.

2.2.11.3 Avoiding RNase contamination

To prevent RNase contamination during RNA extraction and subsequent processing, the following precautions were taken. MilliQ water was treated with 0.1% (v/v) diethylpyrocarbonate (DEPC) (Sigma-Aldrich, St. Louis, MO, USA). A portion of the MilliQ water containing DEPC was used to soak Eppendorf tubes. DEPC-treated water or tubes were allowed to incubate at room temperature overnight before being autoclaved to inactivate traces of DEPC; Metal spatulas and mortar and pestles were baked in an oven at 210 °C for 20 h. Centrifuge tubes and Agilent 2100 Bioanalyzer electrodes were cleaned thoroughly with RNaseZap (Ambion, Austin, TX, USA), then rinsed with DEPC-treated water. Centrifuge tubes were subsequently autoclaved. Only certified RNase-free pipette tips were used.

2.2.11.4 RNA extraction

Cultures were pooled into pre-chilled (-20º C) 250 ml centrifuge tubes and pelleted at 10,000 × *g* for 5 min at 4º C. The supernatant was discarded and the resulting cell pellet was resuspended in 10 ml of BY+ medium (+ 0.2% Xylan) and 20 ml of RNAprotect reagent (Qiagen). The culture was transferred to a 50 ml Oakridge tube and incubated for 5 min at room temperature. The cells were again pelleted by centrifugation for 10 min at 5,000 × *g*. The pellet was dried and frozen using liquid nitrogen. The frozen pellet was ground with a sterile pre-chilled (-20º C) mortar and pestle under liquid nitrogen. The ground sample was then resuspended in an excess of TRIzol reagent and subsequently transferred in 1 ml aliquots to Eppendorf tubes and incubated at 20 ºC in a waterbath for 5 min to permit complete dissociation of nucleoprotein complexes. Chloroform (200 µl) was added to each tube and mixed vigorously and incubated at 20 ºC in a waterbath for 3 min. The sample was centrifuged at 12,000 × *g* for 15 min at 4 ºC to separate the TRIzol-Chloroform and protein phases from the aqueous phase containing the RNA. The aqueous phase was transferred to a sterile Oakridge tube and mixed with 0.5 volumes isopropanol. This mixture was incubated at 20 ºC in a waterbath for 10 min to precipitate the RNA, which was subsequently pelleted by centrifugation at 12,000 × *g* for 10 min at 4º C. The supernatant was removed and discarded using a P200 micropipetter and sterile pipette tip The RNA pellet was washed in 5 ml of 75% ethanol before being pelleted by centrifugation at 12,000 × *g* for 10 min at 4º C. The ethanol was removed and discarded using a P1000 micropipetter and sterile pipette tip and allowed to air dry on

ice for 5 min. Each RNA pellet was resuspended in 1000 µl of DEPC-treated milliQ water and the RNA concentration was quantified by spectrophotometry.

2.2.11.5 RNA purification

The extracted RNA was purified using an RNeasy Midi kit (Qiagen, Hilden, Germany). All the centrifugation steps were carried out at 4º C. buffer RLT (4 ml) containing 1% (v/v) β-mercaptoethanol, was added, along with 2.8 ml of absolute ethanol. The sample was transferred to an RNeasy Midi column (Qiagen), placed in a 15 ml centrifuge tube, and centrifuged at 5,000 × *g* for 5 min. The flow-through volume was discarded and 4 ml of RW1 buffer was added to the column. The column and centrifuge tube were centrifuged at 5,000 × *g* for 5 min. The flow-through volume was discarded and the column was washed twice by the addition of RPE buffer and subsequent centrifugation at 5,000 × *g* for 2 min for the first wash and 5 min for the subsequent wash. The column was transferred to a fresh 15 ml centrifuge tube and 250 µl of DEPC-treated water was added to elute the RNA. The column was incubated at room temperature for 5 min and centrifuged at 5,000 × *g* for 3 min. The elution process was repeated to ensure all purified RNA was extracted from the column. The resulting 500 µl of suspended RNA was transferred to an Eppendorf tube and the RNA concentration was quantified by spectrophotometry (Appendix IV, Table A3).

2.2.11.6 RNA quality analysis

Each RNA sample was diluted to 100 ng/µl with DEPC-treated water and analysed by RNA 6000 Nano assay (Agilent Technologies, Santa Clara, CA, USA). RNA 6000 Nano gel matrix (550 µl) was filtered by centrifugation through a spin filter at 1,500 × *g* for 10 min. Aliquots (65 µl) of the filtered gel were placed into 0.5 ml Eppendorf tubes, and excess aliquots were stored at 4 ºC for no longer than 4 weeks. Dye concentrate (1 µl) was thoroughly mixed with a 65 µl filtered gel matrix aliquot using a Labinco L46 Vortex and then centrifuged at 13,000 × *g* for 10 min at room temperature. An RNA 6000 Nano LabChip® (Agilent Technologies, Santa Clara, CA, USA) was placed on the Chip Priming Station, and 9 µl of the gel-dye mixture was loaded into each of the three wells marked with a G, beginning with the well marked in **bold**. RNA 6000 Nano marker (5 µl, Agilent Technologies, Santa Clara, CA, USA) was added to the ladder well and all 12 sample wells. The RNA 6000 ladder and each

RNA sample were denatured at 70 °C for 2 min, cooled on ice for 1 min, then added in 1 µl volumes to the appropriate wells. The RNA Chip was vortexed at 2,400 strokes/min for 1 min, using an IKA vortex and subsequently loaded and run in an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The resulting data were acquired as both an electropherogram and electrophoretic gel-image translation.

2.2.11.7 Concentrating RNA or cDNA samples

The RNA was gently mixed with 0.1 volumes of 3M sodium acetate, 5 µg glycogen and 3 volumes of ice-cold absolute ethanol. The mixture was incubated at -85 °C for 30 min, and centrifuged at 16,000 × *g* for 30 min at 4 °C. The supernatant was discarded and the remaining RNA pellet was washed twice with 500 µl of ice-cold 70% ethanol. The supernatant was carefully discarded and the RNA pellet was allowed to air dry on ice for 5 min. The RNA pellet was redissolved in an appropriate volume of DEPC-treated TE buffer (10:1).

2.2.11.8 First strand cDNA synthesis

The total RNA samples (suspended in 17 µl of DEPC-treated TE buffer (10:1)) were mixed with the random hexamers in a 1.5 ml Eppendorf tube and incubated at 70 °C for 5 min to denature the RNA. The mixture was rapidly cooled on ice for 2 min and the remaining constituents (Table 2.22) were subsequently added. The reaction was incubated at 46 °C for 3 h. 1 M NaOH solution (15 µl) was added to the reaction and RNA was hydrolysed by incubation at 70 ºC for 10 min. 1 M HCl (5 µl) was added to neutralise the pH and finally 20 µl of 3 M sodium acetate was added.

Table 2.22 First strand cDNA synthesis reaction constituents

Chemical	Concentration
Total RNA	~10 µg
Random hexamers	0.5 µg
First-strand buffer	6 µl
0.1 M Dithiothreitol (DTT)	1.5 µl
dNTP	1.5 µl
RNaseOUT	40 units
SuperScript™ III RT	800 units

2.2.11.9 cDNA purification

Each cDNA sample was mixed with 500 µl of Loading buffer using a 1 ml micropipetter and sterile pipette tip. The sample was transferred to a S.N.A.P.™ column (Invitrogen, Carlsbad, CA, USA) placed in a 2 ml collection tube and centrifuged at 13,000 × *g* for 1 min. The column was washed twice with 700 µl of Wash buffer by addition and subsequent centrifugation at 13,000 × *g* for 1 min. A further centrifugation was performed at 13,000 × *g* for 1 min to eliminate any remaining traces of the Wash buffer, which contains ethanol. The column was then transferred to a fresh 2 ml centrifuge tube and 50 µl of DEPC-treated water was added to elute the RNA. The column was incubated at room temperature for 5 min and then centrifuged at 13,000 × *g* for 1 min. The elution process was repeated to ensure all purified cDNA was extracted from the column. All centrifugation steps during the first strand cDNA purification were carried out at room temperature. The column, collection tube and all buffers were supplied as part of the cDNA labelling purification kit (Invitrogen, Carlsbad, CA, USA).

2.2.11.10 Enumeration of organisms in co-culture

B. proteoclasticus‟ and *M. ruminantium*‟s relative mRNA contribution to the co-culture was enumerated by qPCR analysis of the copy numbers of *bcd* (*B. proteoclasticus*) and *mch* (*M. ruminantium*). 100ng of each cDNA sample (*B. proteoclasticus* in Mono-culture, *M. ruminantium* in Mono-culture, and *B. proteoclasticus* and *M. ruminantium* in coculutre) was serial 10 fold diluted in MilliQ water to 10^{-4}. The undiluted sample and each of the 10 fold dilutions were analysed by real time qPCR, as described in Section 2.2.10.4, using 2.5 mM $MgCl_2$ and the appropriate primer combination (bcd primer pair for *B. proteoclasticus* and mcd primer pair for *M. ruminantium*). Each qPCR run consisted of the Mono-culture dilutions (primed with the bcd primer pair, *B. proteoclasticus*; or the mch primer pair, *M .ruminantium*) and replicates of the co-culture dilutions (one replicate primed with the bcd primer pair the other replicate primed with the mch primer pair). Additionally a blank reaction, containing no cDNA, and three negative reactions, containing each purified mRNA sample, was included. The results were analysed and organism ratios in the co-culture samples determined using the mono-cultures as a reference (Table 4.4).

2.2.11.11 cDNA labelling

Dimethyl sulphoxide (DMSO, 45 µl) was mixed with both Cy3 and Cy5 dye vials (GE Healthcare, Uppsala, Sweden) using a Labinco L46 vortex. Each cDNA sample was divided into equal volumes and 5 µl of the appropriate Dye-DMSO solution was added to each cDNA sample (one volume with Cy3, one volume with Cy5) and incubated in the dark for 2 h, after which the labeled cDNA was purified as described in Section 2.2.11.9. The labeled cDNA was concentrated, (2.2.11.7) and resuspended in 10 µl of DEPC-treated MilliQ water.

2.2.11.12 Microarray hybridisation

Microarray slides were transferred to pre-warmed (50º C) microarray prehybridisation buffer and incubated at 50 °C for 30 min. The microarray slide was rinsed by dipping the slide several times into two 50 ml volumes of 0.22 µm (Millipore) filter-sterilised MilliQ water in 50 ml Falcon tubes. The slide was subsequently dipped into 50 ml of isopropanol in another 50 ml falcon tube and air dried. Once dry, the slides were transferred into hybridisation chambers (Corning, Lowell, MA, USA) and lifter cover slips (Erie Scientific, Portsmouth, NH, USA) were carefully laid over the probe areas. Samples to be compared (e.g. Cy3 Co-culture versus combined Cy5 individual mono-cultures) were combined in 0.2 ml PCR tubes. The combined sample was denatured by incubation at 95 °C for 10 min and mixed with 60 µl of pre-warmed (68° C) Slide Hyb buffer #1 (Ambion, Austin, TX, USA) using a P200 micropipetter and sterile RNase/DNase-free pipette tip. The mixture was loaded onto the pre-hybridised slide at the edge of the cover slip and allowed to diffuse beneath it. Sterile MilliQ water (30 µl) was added to the wells at each end of the hybridisation chambers to ensure the chambers did not dry out during hybridisation. The hybridisation chambers were then sealed and incubated in a water bath at 50 °C for 24 h. Following hybridisation, the cover slip was carefully removed and microarray slides were transferred sequentially to each of the microarray Wash solutions (1 to 3) pre-warmed (50º C) in aluminium foil-covered Falcon tubes. The microarray slides were incubated for 7 min in each Wash solution with extremely vigorous shaking by hand. Following the third wash the slide was transferred to a fresh 50 ml Falcon tube containing a small amount of lint-free paper towel at the bottom and centrifuged for 4 min at 1,500 × *g* at 20º C. The slide was transferred to a fresh 50 ml Falcon tube and dried by incubation in a 37 °C

vacuum oven (Contherm) with the lid removed for 20 min and stored in the dark until microarray scanning.

2.2.11.13 Data acquisition

Microarray slides were scanned using a GenePix® Professional 4200 scanner and GenePix Pro 6.0 software (Molecular Devices, Sunnyvale, CA, USA). The excitation wavelengths and emission filters were selected for optimal analysis of Cy3 and Cy5 dyes (excitation at 532 and 635 nm and emission detection between 655 to 695 nm and 550 to 600 nm, respectively). An image of the entire microarray slide was acquired using a low-resolution (40 µm/pixel) preview scan. The arrayed DNA portion of the slide was identified and an area encompassing these features was selected for scanning at high resolution (5 µm/pixel). The Photo Multiplier Tube (PMT) gains of each channel were adjusted between test scans, to balance the Cy5 and Cy3 signals, until a red to green ratio of 0.95 to 1.05 was achieved to ensure sufficient signal was obtained from each fluorophore. The software was set to scan each line twice, extracting the average value to maximise the signal to noise ratio (Axon Instruments, 2004). Each slide was scanned at a high, medium and low PMT gain settings (PMT gain levels and red to green ratios are shown Appendix III, Table A5; Microarray scan images are shown Appendix III, Fig. A4). Three PMT level scans were performed to ensure expression profiles of genes above the saturation level (65,000) or below the noise level (10,000) were included. The gene names associated with each spot on the microarray were uploaded and traced via a grid with parameters set to overlay the features of the scanned microarray (grid settings Appendix III, Table A6). Features were aligned to the grid automatically using the automatic GenePix Pro 6.0 software function and then refined manually. Some features on the array were not uniformly printed, therefore the „find irregular features" option was selected. Each feature was checked and any poor regions were flagged for exclusion before signal data were extracted using the option „analyse" resulting in a GenePix results (GPR) file.

2.2.11.14 Quality control and normalisation

To ensure the integrity of the acquired data and to determine the appropriate method of normalisation, the data were analysed using the limma package in Bioconductor (Smyth, 2005). Bioconductor was operated through the R platform, using the R editor Tinn-R (R code used for this analysis is shown Appendix IV, P1). To determine the scan level (high, medium or low) with the greatest proportion of features within the useable range (1,000 – 65,000) and thereby provide the base for subsequent analysis, box plots of the $\log_2$ intensities of each feature were produced for each scan level (Fig. 4.15). The signal to noise ratio was analysed by comparing the foreground and background intensities by calculating the intensity of each feature and that of the region around that feature (outside of a 2 pixel circular exclusion zone). To determine if there was any hybridisation spatial bias, the foreground and background densities were plotted individually for each of the 16 slide scans to get an overall impression of the data (Figure 4.16). To ensure there was no bias to either fluorophore‟s detection through the increasing signal intensities an MA plot was produced (Fig. 4.17). An MA plot is a scatterplot of the log ratio of Cy5 to Cy3 ($\log_2$(Cy5) - $\log_2$(Cy3); M) versus the average intensity for each feature ($^1/_2$($\log_2$Cy5+Cy3; A).

Density plots were produced to show the proportion of features at each intensity and the signal distribution across a microarray slide (Figure 4.18). Boxplots were produced to give an overview of the range of values observed for each of the spot types (*B. proteoclasticus*, *M. ruminantium*, controls (*Arabidopsis* and random oligomers), the constitutively expressed genes *dnaK* and *frhB* and blanks; data shown Fig 4.20).

The blank spots and those flagged as being of poor quality were removed from subsequent analysis. Normalisation was performed within each microarray using the „print-tip loess‟ method to remove any spatial trends within the slide. This method fits a loess smoother to each of the 32 print-tip blocks on each slide and was selected to remove the bias introduced from the different pins of the spotting robot. As some spatial variation is seen in the backgrounds which is not reflected in the foregrounds it was determined that background subtraction would unfairly bias the results therefore this was not performed. To determine the efficacy of normalisation, density plots were repeated following normalisation and compared to the raw data (Fig. 4.19) and

boxplots were produced of the intensity distribution of each feature by category (4.20).

2.2.11.15 Statistical analysis

Statistical data analyses were carried out using the limma package of the Bioconductor software. Data were $\log_{10}$ transformed prior to analysis and the mean difference was calculated using this scale, resulting in a log ratio for each feature. The log ratios for each probes replicates (biological and technical) were averaged prior to further analysis. Each feature was then adjusted for the technical slide replication by determining the average correlation between technical replicates using the „duplicateCorrelation" function, and accounting for this when fitting the linear model. Empirical Bayes moderated t-statistics (Robbins, 1956, Smyth, 2004) were then calculated for each gene. The moderated t-statistics were adjusted to give an FDR (False Discovery Rate; (Benjamini & Yekutieli, 2005) value. Genes with an up- or down-regulation of 2 fold or greater and an FDR value < 0.05 were deemed to be statistically significant.

3 Sequencing of *B. proteoclasticus* B316^T auxiliary replicons

3.1 Introduction

The genome of *B. proteoclasticus* B316^T was sequenced to a theoretical genome coverage of 9 fold, as part of a genome sequencing project undertaken by the Rumen Microbiology Genomics team at AgResearch. DNA libraries were constructed using a random shotgun cloning approach as previously described (Fleischmann *et al*., 1995). Genomic DNA was randomly disrupted and separated by gel-electrophoresis. DNA fragments in the 2 to 4 Kb range were isolated and used to generate a small insert plasmid library in *E. coli*, while fragments of ~40 Kb were isolated and used to produce a fosmid library (Kim *et al*., 1992). Clones resulting from both libraries were recovered and sequenced using a high throughput sequencing technology. Sequencing was completed by Agencourt Biosciences Corporation (Beverly, MA, USA), who also aligned the resulting sequences to find overlaps and assembled them into contiguous sequences (contigs). Prior to the commencement of this project the assembled genome was subjected to a single round of primer walking. Collectively these efforts resulted in 313 contigs and 3,939,787 bp of non-redundant sequence. This DNA sequence information is henceforth referred to as the Phase I genome sequence data. Pulsed-feild gel-electrophoresis of uncut whole-genomic DNA revealed the genome to be spread across 4 independent replicons including the major chromosome and three auxiliary replicons predicted from their mobilities within the gels to be of sizes 360, 300 and 190 Kb.

This chapter describes the identification of contigs pertaining to each of *B. proteoclasticus'* auxiliary replicons, their finishing, and confirming the integrity of each sequence.

3.2 Identifying episomal contigs within the Phase I genome sequence

To identify contigs containing DNA sequence of the auxiliary replicons, all generated contigs were searched for the presence of genes related to plasmid replication (reviewed in Section 1.12). This was achieved using an automated annotation of GLIMMER-identified ORFs present in all of the *B. proteoclasticus* contigs of the Phase I sequence, and an in-house BLASTP script that utilised the Phase I contigs as

its search database. This approach allowed ORFs to be traced back to their parent contigs. Five ORF"s encoding replication initiation (Rep) proteins (MANATEE ORFs 403, 1831, 1851, 2233 and 3092) and 7 ORFs encoding plasmid partitioning (Par) proteins (MANATEE ORFs 401, 1397, 1398, 3093, 3138, 3769 and 3772) were found spread across 5 contigs (Contigs 3, 67, 68, 112 and 240) (Table 3.1). To exclude chromosomally-encoded Par proteins, the phylogeny of all Par proteins identified was analysed. The amino acid sequence of each of the *B. proteoclasticus*" Par proteins, along with two chromosomal based ParA homologues (chromosome partitioning protein ParA [AAO34754] (*Clostridium tetani* E88) and Sporulation initiation inhibitor protein SoJ [P37522] (*Bacillus subtilis*)), two chromosomally located ParB homologues (chromosome partitioning protein ParB [AAO34755] (*Clostridium tetani* E88) and the Stage 0 sporulation protein J SpoOJ [P26497] (*Bacillus subtilis*)), two plasmid based ParA homologues (chromosome partitioning related protein [YP_209675] (pBCNF5603 of *C. perfringens*) and ATPase, ParA-family [AAM26200] (pXO2-39 of *B. anthracis*)) and two plasmid based ParB homologues (ParB protein [BAD90620] (pBCNF5603 of *C. perfringens*) and ParB-like nuclease domain protein [AAS44682] (pBc10987 of *B. cereus* ATCC10987)) were aligned using ClustalX (Thompson *et al*., 1994). The alignment was then used to construct a tree using Splitstree (Huson, 1998). Analysis of this tree revealed the primary sequence of two *B. proteoclasticus* Par proteins (MANATEE ORFs 1397 and 1398) clustered with their chromosomal counterparts (ParB and ParA respectively) and were therefore most likely be of chromosomal origin (Figure 3.1). Both Par proteins were encoded upon contig 68 and subsequently this contig was eliminated as belonging to one of the three auxiliary replicons.

Self-alignment of Contig 67, (360,097 bp), showed it had a 326 bp overlap at its ends (Figure 3.2) and therefore assembled as a circular molecule of 359,574 bp, consistent with the estimated size of the largest auxiliary replicon, now designated pCY360.

A second contig (Contig 112, 141,980 bp) containing *repB* and *parA* related genes was known, from end sequencing of the longer fosmid clones, to join together five other contigs (Contigs 107, 108, 109, 110 and 111). Collectively these contigs represented 280,809 bp of non-redundant sequence (Table 3.2). They were assigned to the 300 Kb replicon, which is now designated BPc2. This assignment was based on

them not belonging to pCY360 and their collective size already exceeding that of the 186 Kb replicon.

Two further contigs (Contigs 3 and 240) were found to contain *repB* genes. Contig 240 additionally contained genes encoding both ParA and ParB plasmid partitioning proteins. Both of these contigs along with Contigs 4 and 5 were known from end sequencing of the fosmid clones to be linked and collectively represent 100,907 bp of non-redundant sequence. They were assigned to the 186 Kb replicon, now designated pCY186 (Table 3.2).

Table 3.1. Identification of contigs encoding plasmid replication-related proteins

MANATEE ORF number	Automated annotation	Contig number	Position on contig (start..end)
Replication Initiation (Rep) proteins			
ORF00403	Initiator RepB protein	c67	22626..21449
ORF01831	Initiator RepB protein	c3	30231..29131
ORF01851	Initiator RepB protein	c3	51758..50970
ORF02233	Initiator RepB protein	c240	1637..396
ORF03092	Initiator RepB protein	c112	87177..88321
Plasmid partitioning (Par) proteins			
ORF00401	ATPase, ParA-family	c67	24815..24051
ORF01397	ParB-family	c68	32581..33480
ORF01398	ATPase, ParA-family	c68	31768..32529
ORF03093	ATPase, ParA-family	c112	83705..82881
ORF03138	ATPase, ParA-family	c112	27618..28373
ORF03769	ATPase, ParA-family	c240	4905..4114
ORF03772	ParB-family	c240	3639..2650

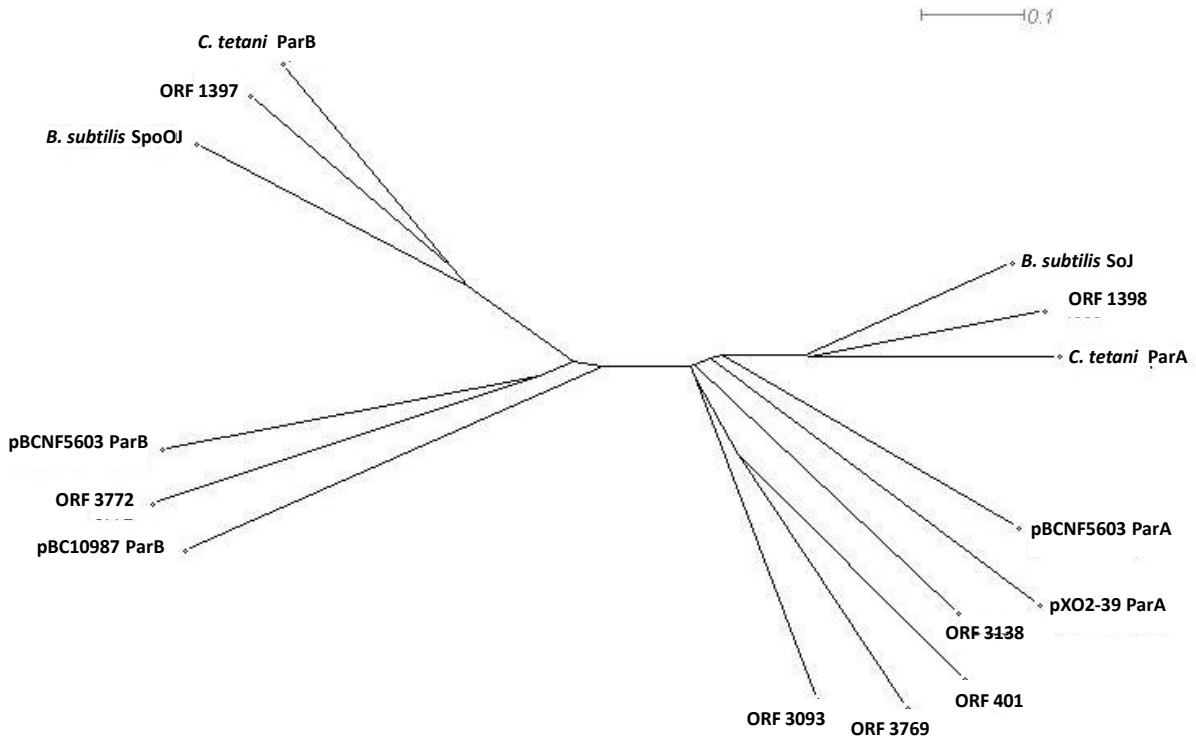


Figure 3.1. Phylogenetic placement of Par proteins. Phylogenetic tree of partitioning (Par) proteins identified in the Phase I genome sequence of *B. proteoclasticus* and known chromosomal (*B. subtilis* and *C. tetani*) and plasmid (pBCNF5603, pBc10987 and pXO2-39) ParA (Soj and ParA) and ParB (SpoJ and ParB) homologs. ORF 1398 clusters with chromosomal ParA-family proteins and ORF1397 clusters with chromosomal ParB-family proteins. Alignments were made using ClustalX and the tree was constructed with Splitstree.

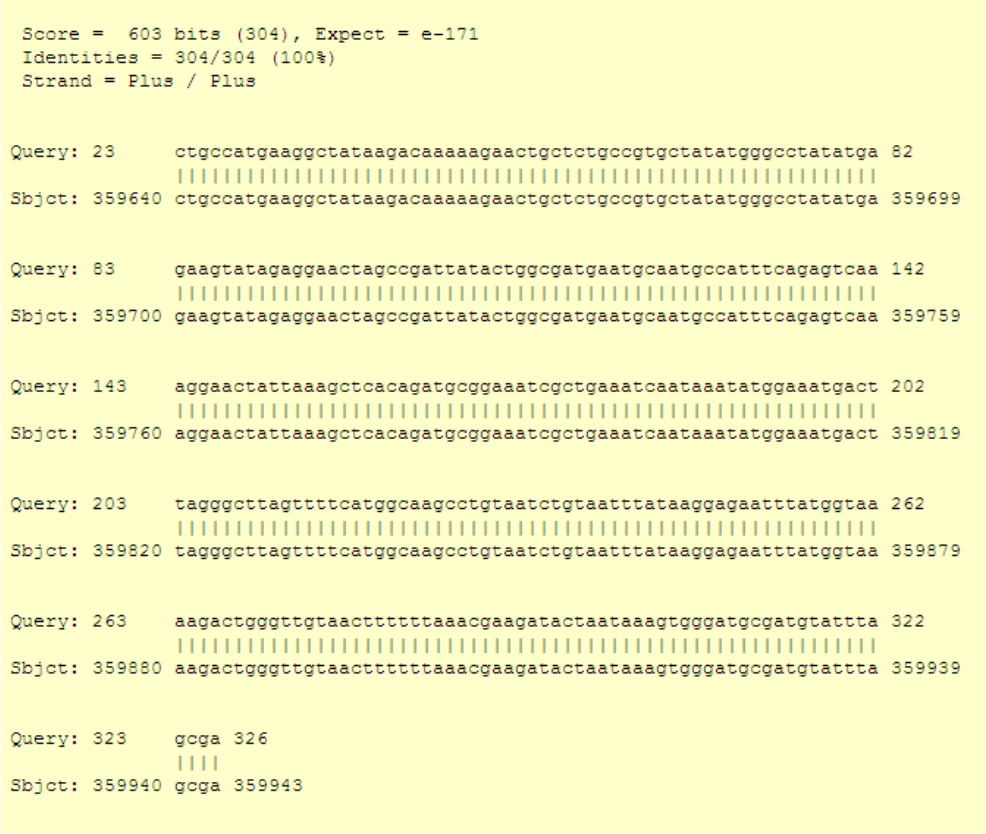


Figure 3.2. Self-alignment of Contig 67 reveals overlapping ends. Alignment of Contig 67 with itself by BLASTN shows that the ends overlap by 326 bp, suggesting it is complete and forms a circular structure.

3.3. Gap closure of episomal DNAs

Unlike pCY360, the two smaller replicons did not assemble into single contigs in the Phase I data. The sequence at the ends of these contigs contained an over-representation of transposase genes, with 3 of the 5 transposase-genes initially identified in BPc2 and pCY186 contigs being found at contig ends. A gene encoding a 16*S* ribosomal RNA was located at one end of Contig 107 and a 23*S* ribosomal RNA-encoding gene was found at one end of Contig 108, both contigs belonging to BPc2.

A genome closing strategy using inverse PCR closed the gap between Contig 110 and Contig 111 prior to the commencement of this Ph.D, resulting in the Supercontig 110+. To close the remaining sequence and physical gaps, various methods including multiplex, inverse, and long range PCRs, and the screening of BAC libraries

generated to aid genomic assembly, were used. The results are described below and summarised in Table 3.3.

3.3.1 Multiplex PCR

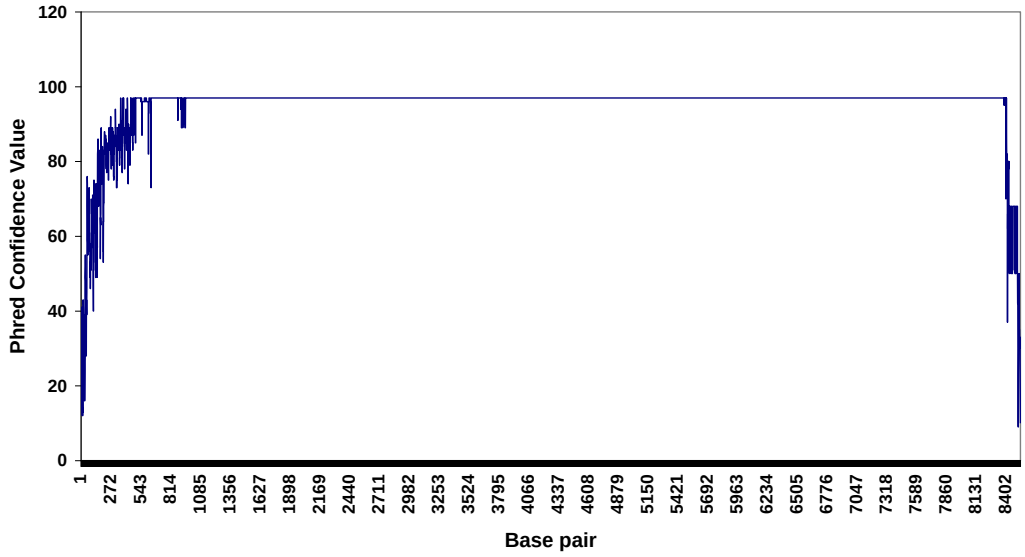
PCR primers were designed to amplify DNA from each of the remaining 5 contig ends (Contigs 110/111 forming the single Supercontig 110+) of BPc2 and the 4 contig ends of pCY186. As contig ends typically have poor sequence quality (Blakesley *et al*., 2004), the Phred scores of the assigned bases at each end were analysed (Figure 3.3). Primers were then designed within regions with a Phred score greater than 40. Using these primers multiplex PCR was performed (Section 2.2.5.7), whereby each forward primer was combined with a pooled equimolar concentration of each reverse primer, and vice-versa. All 10 BPc2 reactions and 8 pCY186 reactions were subjected to 30 rounds of touchdown PCR and the resulting product(s) were screened by agarose gel-electrophoresis. The resulting product sizes estimated from DNA mobility within the gel allowed several of the BPc2 contigs to be tentatively ordered and gap sizes to be estimated. Contig 109 was estimated to be 3.5 Kb from Supercontig 110+, which was in turn estimated to be less than 100 bp from Contig 112, and Contig 112 was estimated to be 1.5 Kb from Contig 107. No amplification product was identified by amplification from either end of Contig 108. However, given the contig ordering determined by the multiplex PCR, and the likely presence of an rRNA operon between Contig 107 and Contig 108, it was believed Contig 108 was flanked by Contigs 107 and 109. Further, given the size estimated for BPc2 was 300 Kb, it was predicted that ~14 Kb of unknown sequence was distributed in the two sequencing gaps between Contigs 107, 108 and 109.

Table 3.2 Contigs assigned to each replicon.

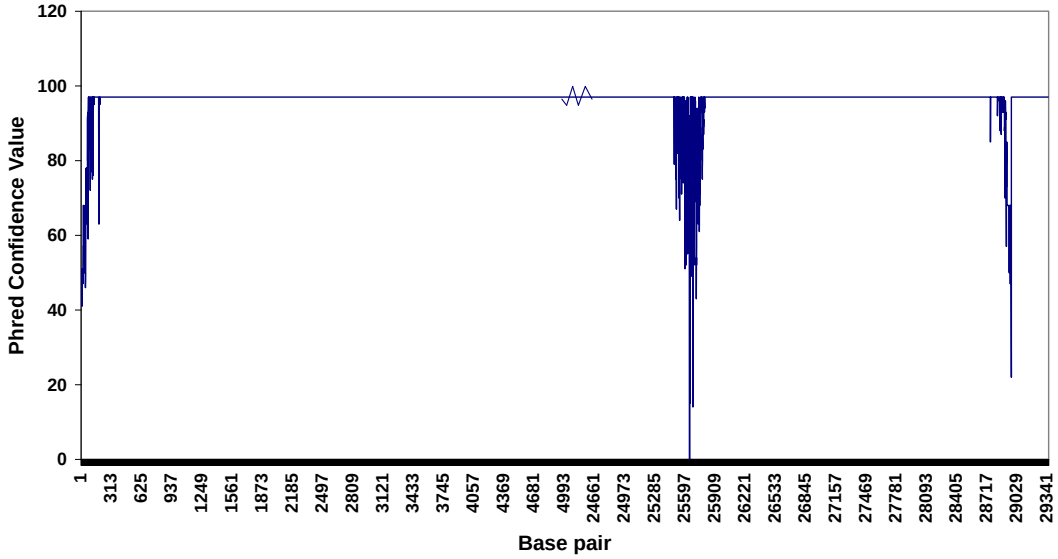
Contig	Size (bp)
pCY360	
67	359,574
BPc2	
107	8,534
108	29,356
109	79,029
110+	21,827
112	141, 979
pCY186	
3	94,386
4	27,404
5	11,877
240	6,521

Despite several multiplex PCR attempts using two different sets of designed primers, no products were amplified from any of the pCY186 contig ends or Contig 108 of BPc2. The gaps that were successfully amplified were sequenced, using the primer combinations identified through multiplex PCR to have complementarity (e.g. BPc2_112b – BPc2_107a). It is possible to isolate and purify the appropriate bands from the multiplex agarose gel, however, due to the nature of multiplex PCR, the reaction would have potentially been unbalanced and this has previously been shown to decrease PCR fidelity (Pierce *et al.*, 2005, Thein & Wallace, 1993).

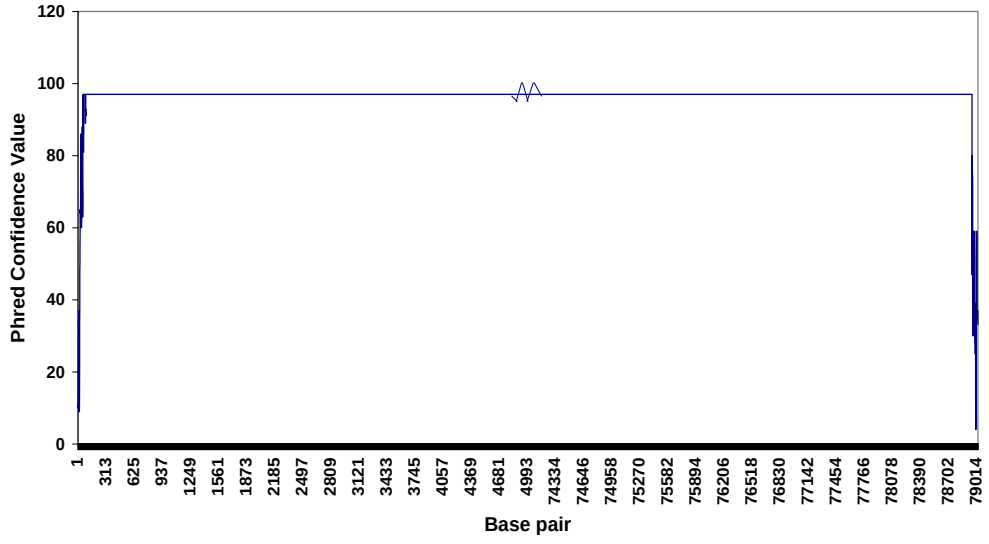
Contig 107



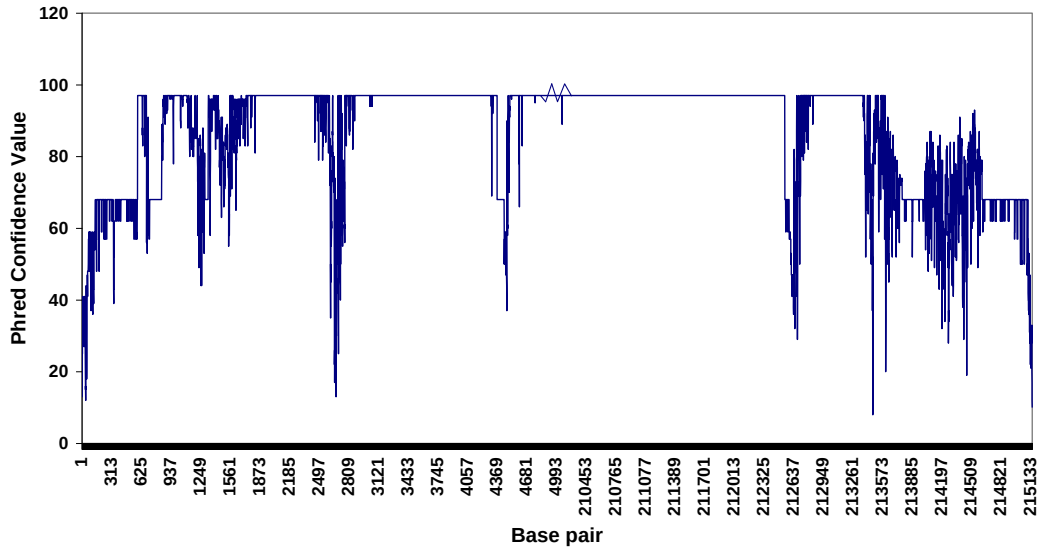
Contig 108



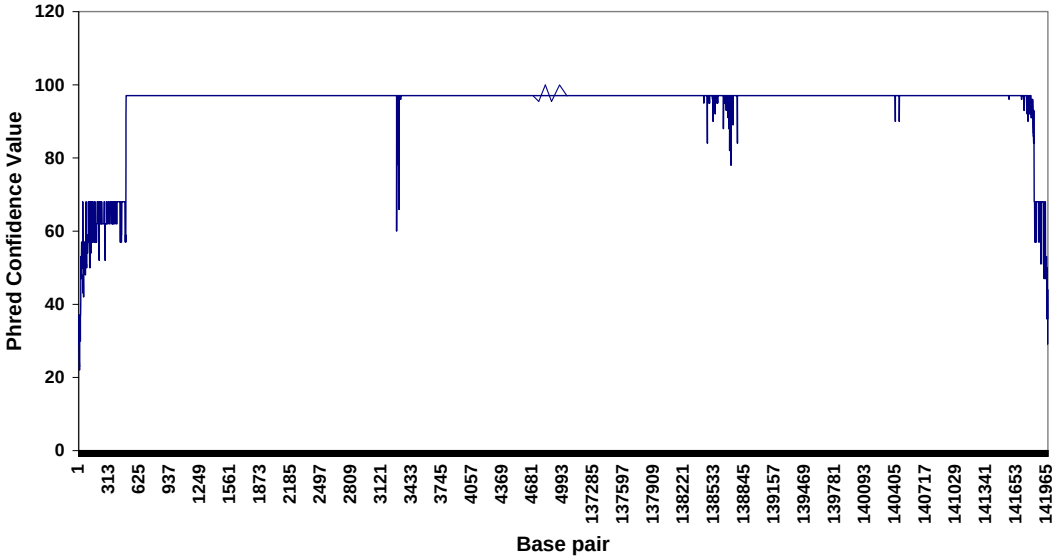
Contig 109



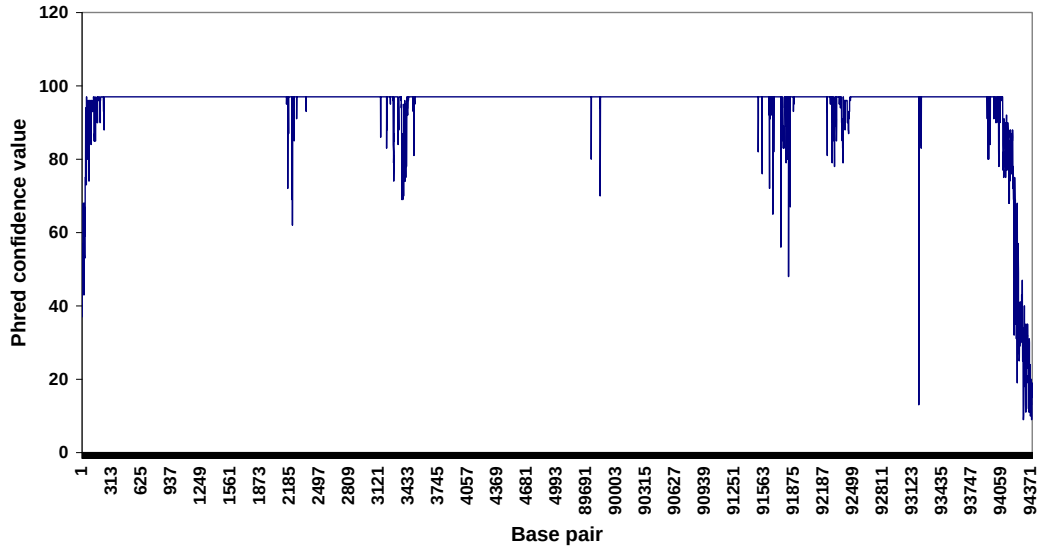
Contig 110/111



Contig 112



Contig 3



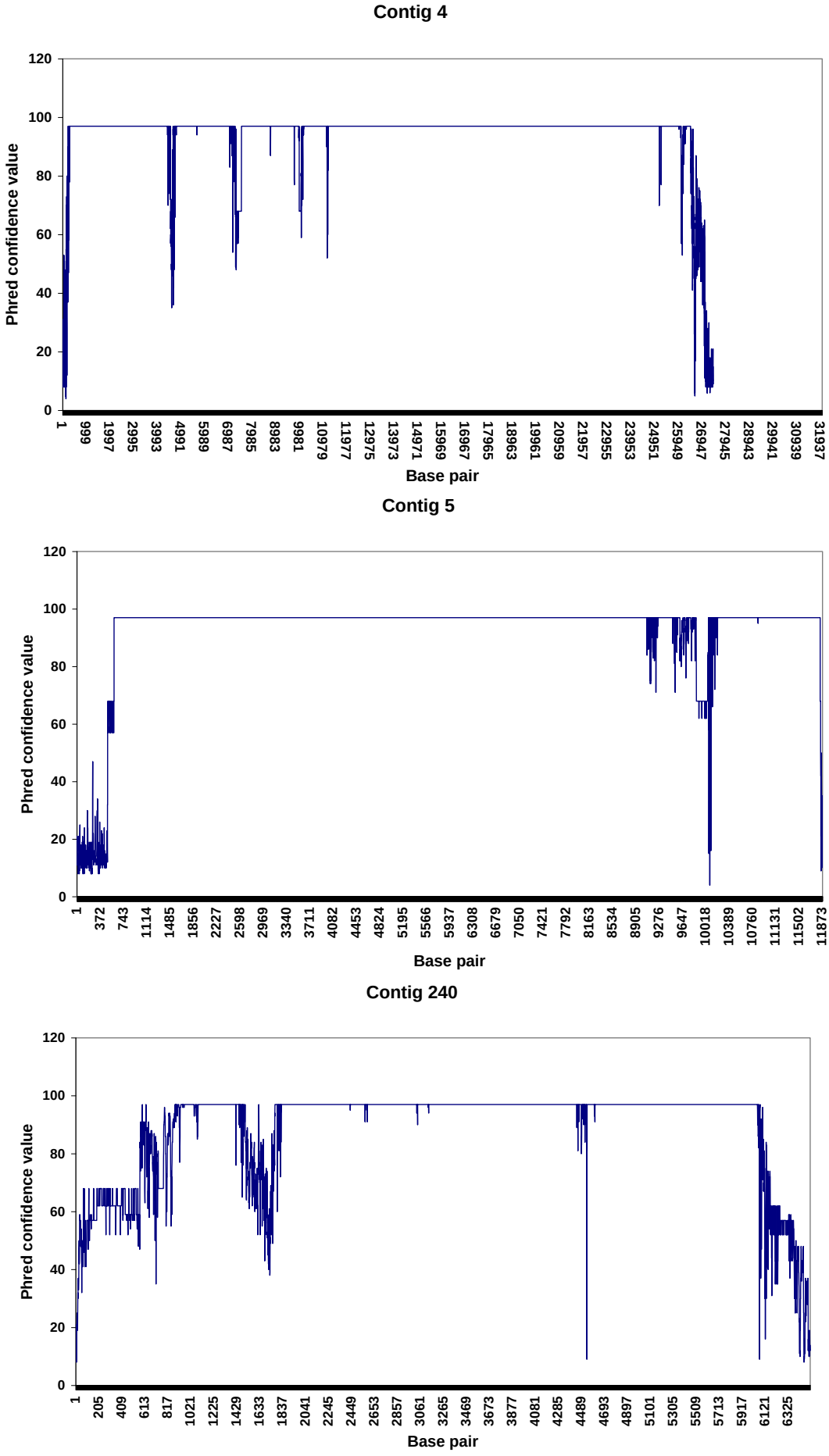


Fig. 3.3 Phred quality scores of contig ends. The Phred quality scores of each base pair for contigs associated with auxiliary replicons are shown. For larger contigs, only 5000 bp at each of the contig ends are shown.

3.3.2 BAC clone screening

The second approach to close the gaps remaining in the DNA sequences of the BPc2 and pCY186 replicons was the screening of the BAC clone library. Initial efforts to construct BAC clones with insert sizes ~250 Kb and later ~150 Kb had resulted in unstable clones that failed to provide useful sequencing data. Therefore the BAC library was constructed with no selection for insert size, resulting in an average insert size of approximately 30 Kb. Analysis of the BAC clone end-sequencing data revealed 2 useful BAC clones. The first clone extended from Contig 240 to Contig 105, a contig not previously associated with pCY186. The second BAC clone spanned from Contig 4 across Contig 105 and into novel sequence. Collectively the 1289 bp of new sequence derived from these BAC end-sequences connected Contig 240 with Contig 4 via Contig 105.

3.3.3 Conventional and long-range PCR

The primer pairs that were found to complement each other in the multiplex PCR analysis, and those corresponding to contigs (such as Contig 108) whose order relative to other contigs could be deduced, were tested by conventional PCR. The gaps that could not be amplified using conventional PCR (e.g. between Contig 107 and Contig 108) were subjected to long-range PCR using the Eppendorf TripleMaster system (Hamburg, Germany). Long-range PCR allowed primers to be nested more medial within a contig, allowing PCR primers to be better optimised for the conditions described in Section 2.2.5.5. In addition the sequencing and physical gaps of pCY186, where contig order was uncertain, were tested in all potential combinations. The resulting PCR products were detected and their size determined by gel electrophoresis and subsequently purified and sequenced. The resulting sequence trace files were manually analysed for quality and assembled. Where sequencing results only partially spanned the PCR product, new primers were designed upstream from either extremity of the newly derived sequence and a further round of sequencing was performed using the PCR product derived from the initial reaction. This process, known as primer walking, was repeated until the gap was closed with sequence of satisfactory quality. Conventional PCR contributed to the closure of the gaps between the BPc2 Contigs 109 and 110+, 110+ and 112, and 112 and 107. Long-range PCR successfully contributed to the closure of the gaps between contigs 107 and 108, 108 and 109, and 109 and 110+ of BPc2 and contigs 4 and 5 of pCY186.

3.3.4 Inverse PCR (iPCR)

A partial gap of BPc2 and two of the four pCY186 gaps were not able to be closed by the methods described above. To capture the remaining regions iPCR was used. Restriction endonuclease (RE) recognition sites occurring near the ends of each remaining contig (or supercontig) were identified using Vector NTI (Fig. 3.4). The REs *Eco*RI, *Hin*dIII, *Cla*I, *Bsp*HI, *Nde*I and *Pst*I were selected for use (all but *Pst*I, which has a high %G+C recognition sequence, satisfying the criteria described in Section 2.2.5.8). The resulting products were analysed by gel electrophoresis and, where successful, purified and sequenced. iPCR contributed to the closure of the gaps between Contigs 109 and 110+ of BPc2, and Contigs 3 and 240 of pCY186.

3.3.5 454 sequencing

Following the sequencing efforts described above, a single physical gap (Contig 5 to Contig 3) from the pCY186 replicon remained unclosed. Closing this physical gap, along with a number of physical gaps from the chromosome, could not be achieved. A new approach, array-based pyrosequencing of whole genomic DNA, was used to provide further sequence information to help in the closure of these gaps and consequently the genome. The array-based pyrosequencing was completed by 454 Life Sciences (Branford, CT, USA) and resulted in approximately 14-fold genome coverage. A quality score of Q40 (equivalent to Phred 40) or greater was assigned to 98.98% of all nucleotides produced. This sequence coverage was sufficient to close the remaining gap in the pCY186 replicon.

3.3.6 Reassembly

Following the closure of the entire genome, all DNA sequences were realigned to identify sequence overlaps and reassembled using the STADEN package (Staden, 1996). The pCY360 replicon was found to contain an extra 1823 bp missed during the original assembly, giving it a total size of 361,397 bp. This was largely due to two near-identical ORFs encoding transposases that resided adjacent to one another which were assembled in the original assembly as a single ORF (ORFs 90 and 91). BPc2 was found to be 302,357 bp in the final assembly, approximately 2 Kb (0.7%) larger than originally estimated. pCY186 was found to be 186,326 bp, approximately 3.5 Kb (2%) smaller than initially estimated.

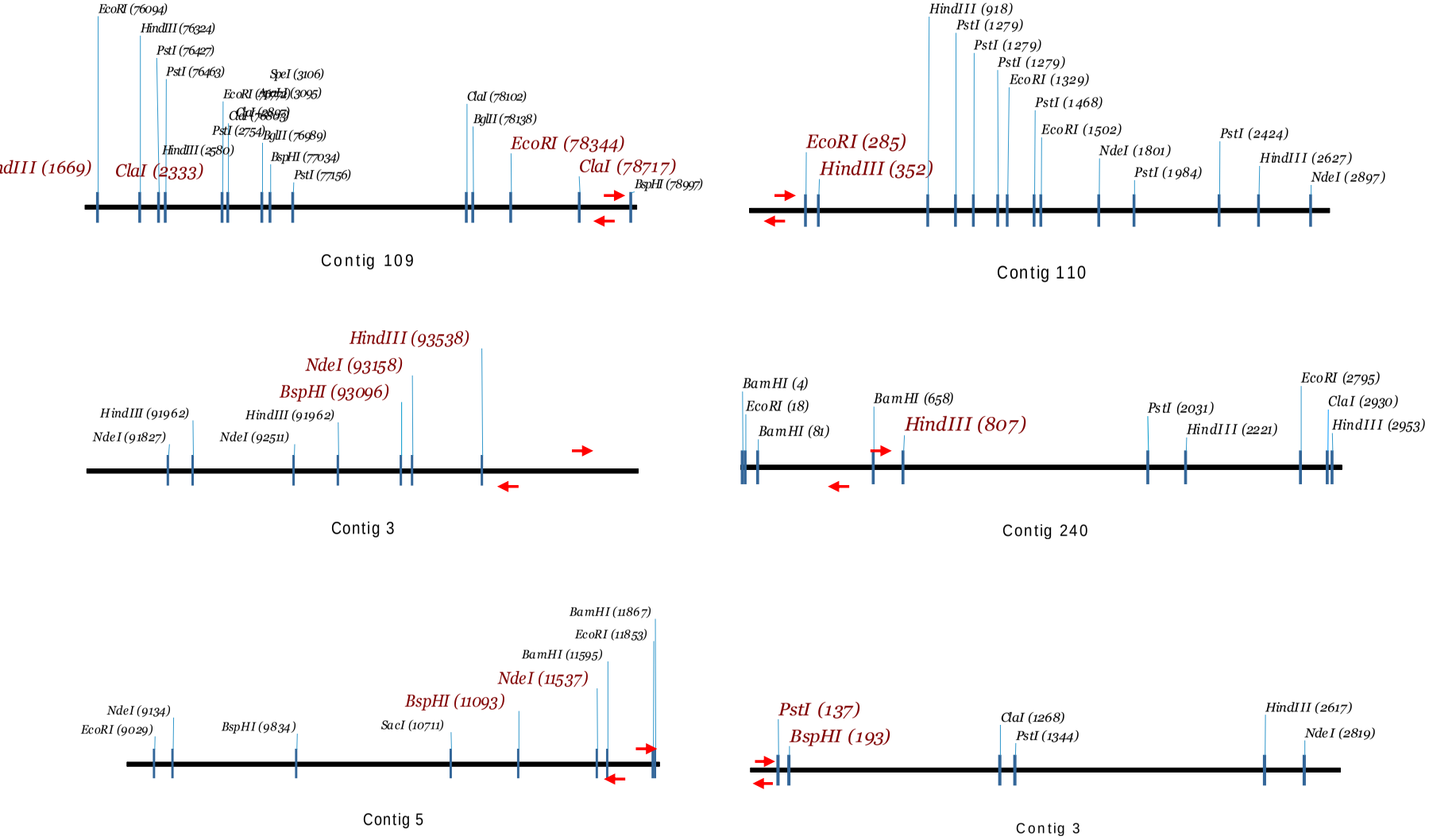


Figure 3.4 Restriction enzymes and PCR primer locations in contig ends as used for iPCR. The positions of RE recognition sites identified near contig ends are shown. Those REs selected for iPCR library construction are shown in red. The approximate locations of primers used for sequencing of iPCR products are shown as red arrows. Figures were constructed using Vector NTI.

Table 3.3. Sequence and physical gaps within replicon DNAs and their methods of closure.

Contig Gap	Size (bp)	PCR			BAC clone screening
		Conventional	Long-range	Inverse	
BPc2					
107 _ 108	3,855	-	+	**n/a**	-
108 _ 109	4,284	-	+	**n/a**	-
109 _ 110+	2,998	+	+	+	**n/a**
110+ _ 112	51	+	**n/a**	**n/a**	**n/a**
112 _ 107	1,081	+	**n/a**	**n/a**	**n/a**
pCY186					
3 _ 240	1,148	-	-	+	-
240 _ 4	1,289	-	**n/a**	**n/a**	+
4 _ 5	4,671	-	+	**n/a**	-
5 _ 3	39,053	-	-	-	-

All gaps initially present in the sequences of BPc2 and pCY186, their size and techniques used in their closure are shown. Techniques that contributed to the closure of a gap are indicated by a +, while those that were attempted but did not contribute to the closure of a gap are indicated by a -. Techniques not attempted are indicated as n/a.

The pCY360 replicon was found to be over-represented in the sequencing data derived from the small plasmid and larger fosmid sequencing libraries, accounting for 20% of all sequenced clones. Both BPc2 and pCY186 were under-represented in these libraries, accounting for 7.9% and 6.8% of all sequenced clones, respectively. The final sequence coverage of each replicon was 16.7, 6.7 and 5.7 fold for pCY360, BPc2 and pCY186, respectively. The distribution of this coverage was not uniform (Fig. 3.5) with certain regions of the BPc2 and pCY186 replicons being significantly under represented.

3.4 Confirmation of two ribosomal RNA operons

Preliminary annotation of the BPc2 replicon sequence data suggested a ribosomal RNA (rRNA) operon was likely encoded between Contig 107 and Contig 108. The BPc2 complete sequence revealed two rRNA operons, separated by a small (~4 Kb) region. To verify the integrity of the dual rRNA operons, whole genomic DNA was digested with I-*Ceu*I, a restriction endonuclease specific for the 23*S* rRNA gene, *rrl*, and separated by PFGE (Fig. 3.6a). A band of approximately 9 Kb, consistent with the distance expected between the *rrl* genes of the two rRNA operons was detected. To confirm this product was derived from BPc2, the digested bands were transferred to a nitrocellulose membrane and Southern hybridised with a labeled probe derived from the 4 Kb of intervening sequence. The resulting blot (Fig. 3.6b) confirmed that the digested product was derived from BPc2.

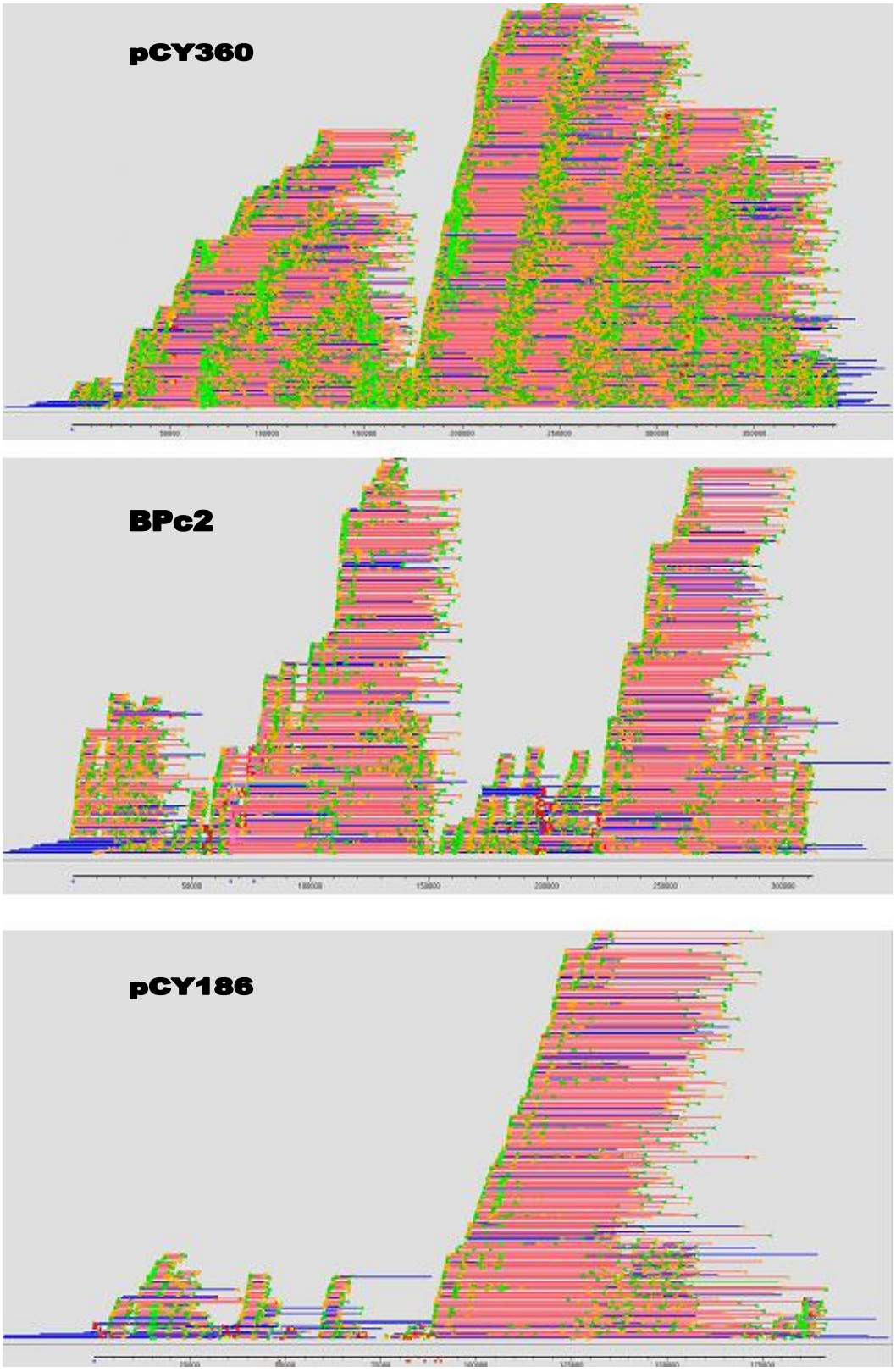


Fig. 3.5. Final sequence coverage of each replicon. All sequencing reactions are shown, as they were assembled upon each auxiliary replicon. Sequencing reactions are shown as either pink lines (sequence derived from both forward and reverse reactions of each clone or amplicon) or blue lines (sequence derived from either forward or reverse reactions of clone or amplicon). Figure was produced using Staden following sequence assembly.

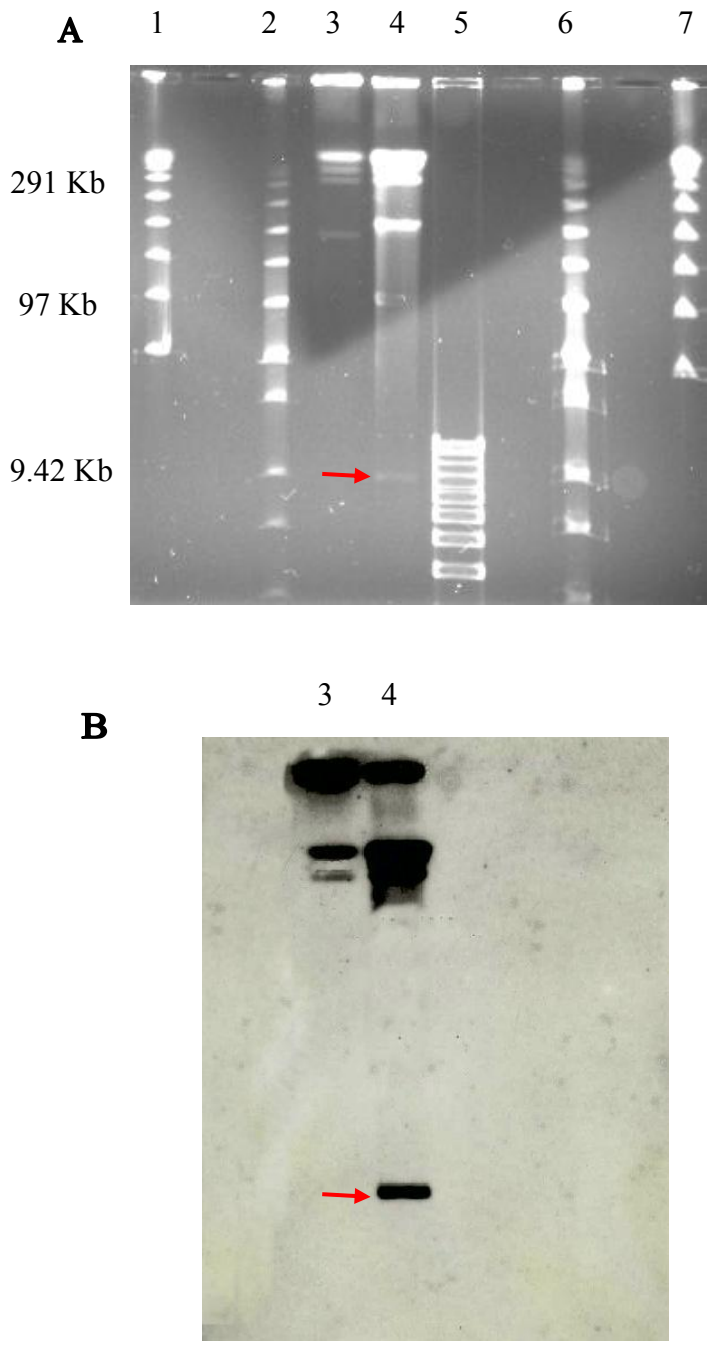


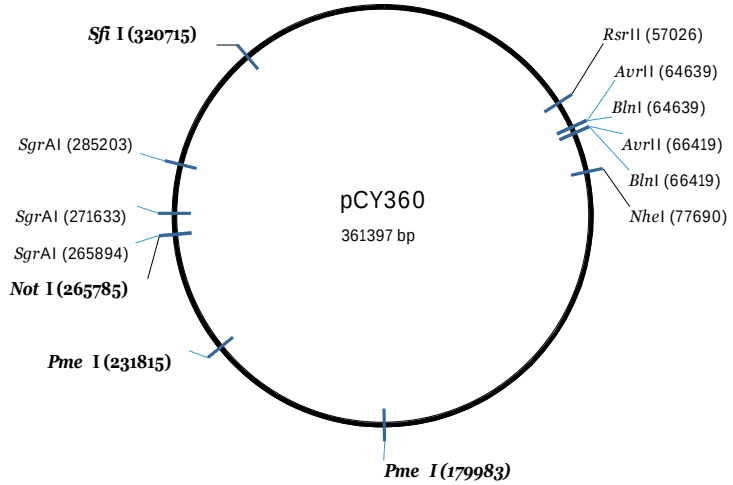
Fig. 3.6. Confirmation of two rRNA operons in BPc2. A) Pulsed-field gel electrophoresis of whole genomic DNA digested with I-*CeuI* (site present in rRNA operon; lane 4) or undigested (lane 3) alongside the λ (lanes 1 and 7), λ low range ladder (lanes 2 and 6) and 1 Kb+ (lane 5) ladder. Estimated band sizes for lanes 3 and 4 are (top to bottom left to right excluding lanes and the irresolvable raction immediately beneath the well) 360 Kb, 300 Kb and 190 Kb (lane 3) and 300 Kb, 200 Kb, 100 Kb and 9 Kb (lane 4). B) Southern blot of the above gel using as a probe a labeled PCR product corresponding to the 4 Kb sequence predicted to separate the two rRNA operons of BPc2. Red arrow indicates the 9 Kb band corresponding to the inter-RNA spacer. Hybridisation is also seen to both the major chromosome and BPc2 in lane 3.

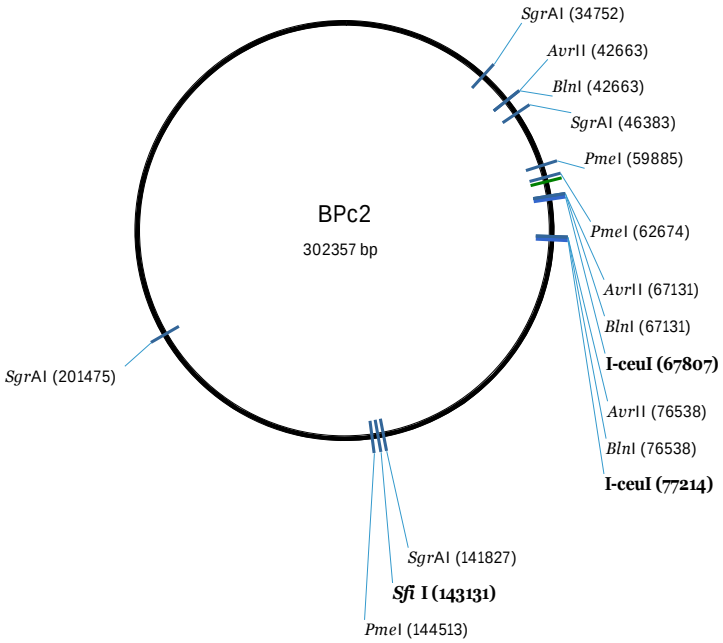
3.5 Confirmation of assembly

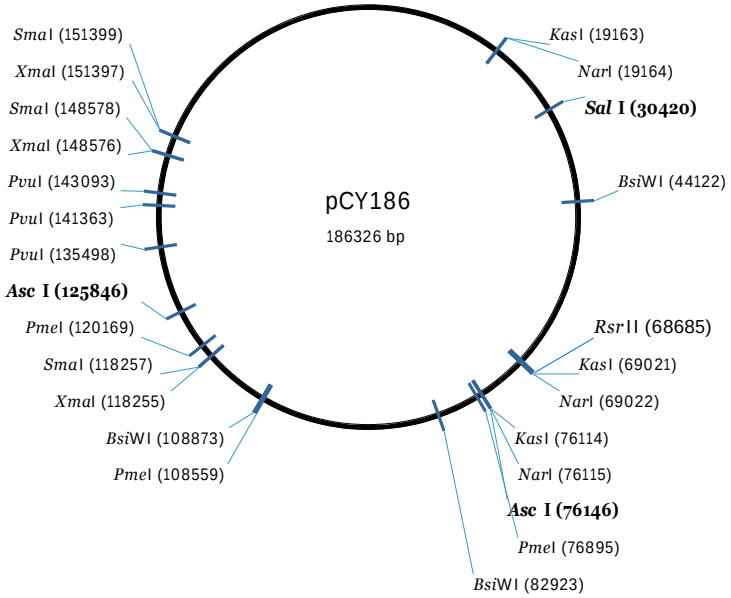
To ensure the assemblies of each of the auxiliary replicons were correct, the recognition sites of several REs were detected within each sequence using Vector NTI (InforMax, 2001). Only REs that were present 3 or fewer times were selected, to enable the resulting DNA bands to be discernable via PFGE (Fig. 3.7a). The restriction enzymes *Not*I, *Sfi*I and *Pme*I were used for pCY360, I-*Ceu*I and *Sfi*I for BPc2 and *Asc*I and *Sfi*I for pCY186. *Xba*I was also used as it was predicted not to cut the pCY360 or pCY186 replicons, but cuts the remainder of the genome at a high frequency.

Whole genomic DNA extracts of *B. proteoclasticus* B316^{T} were digested with each of these REs in single and double digests. The resulting digested genomic DNAs were separated by PFGE using reduced pulse durations to ensure that the smallest fragments would not migrate beyond the length of the gel. The resulting gel was transferred to a nitrocellulose membrane and probed with labeled amplicons of regions within fragments produced by cleavage with each selected RE. Analyses of the resulting Southern blots (Fig. 3.7b) were consistent with the *in-silico* restriction maps produced for each replicon, supporting the sequence assembly in each case.

A


Sfi I (320715)
Rsr II (57026)
Avr II (64639)
Bln I (64639)
Avr II (66419)
Bln I (66419)
Nhe I (77690)
Sgr AI (285203)
Sgr AI (271633)
Sgr AI (265894)
Not I (265785)
Pme I (231815)
pCY360
361397 bp
Pme I (179983)


Sgr AI (34752)
Avr II (42663)
Bln I (42663)
Sgr AI (46383)
Pme I (59885)
Pme I (62674)
Avr II (67131)
Bln I (67131)
I-ceu I (67807)
Avr II (76538)
Bln I (76538)
I-ceu I (77214)
Sgr AI (141827)
Sfi I (143131)
Pme I (144513)
Sgr AI (201475)
BPc2
302357 bp


Sma I (151399)
Xma I (151397)
Sma I (148578)
Xma I (148576)
Pvu I (143093)
Pvu I (141363)
Pvu I (135498)
Asc I (125846)
Pme I (120169)
Sma I (118257)
Xma I (118255)
Bsi WI (108873)
Pme I (108559)
Kas I (19163)
Nar I (19164)
Sal I (30420)
Bsi WI (44122)
Rsr II (68685)
Kas I (69021)
Nar I (69022)
Kas I (76114)
Nar I (76115)
Asc I (76146)
Pme I (76895)
Bsi WI (82923)
pCY186
186326 bp

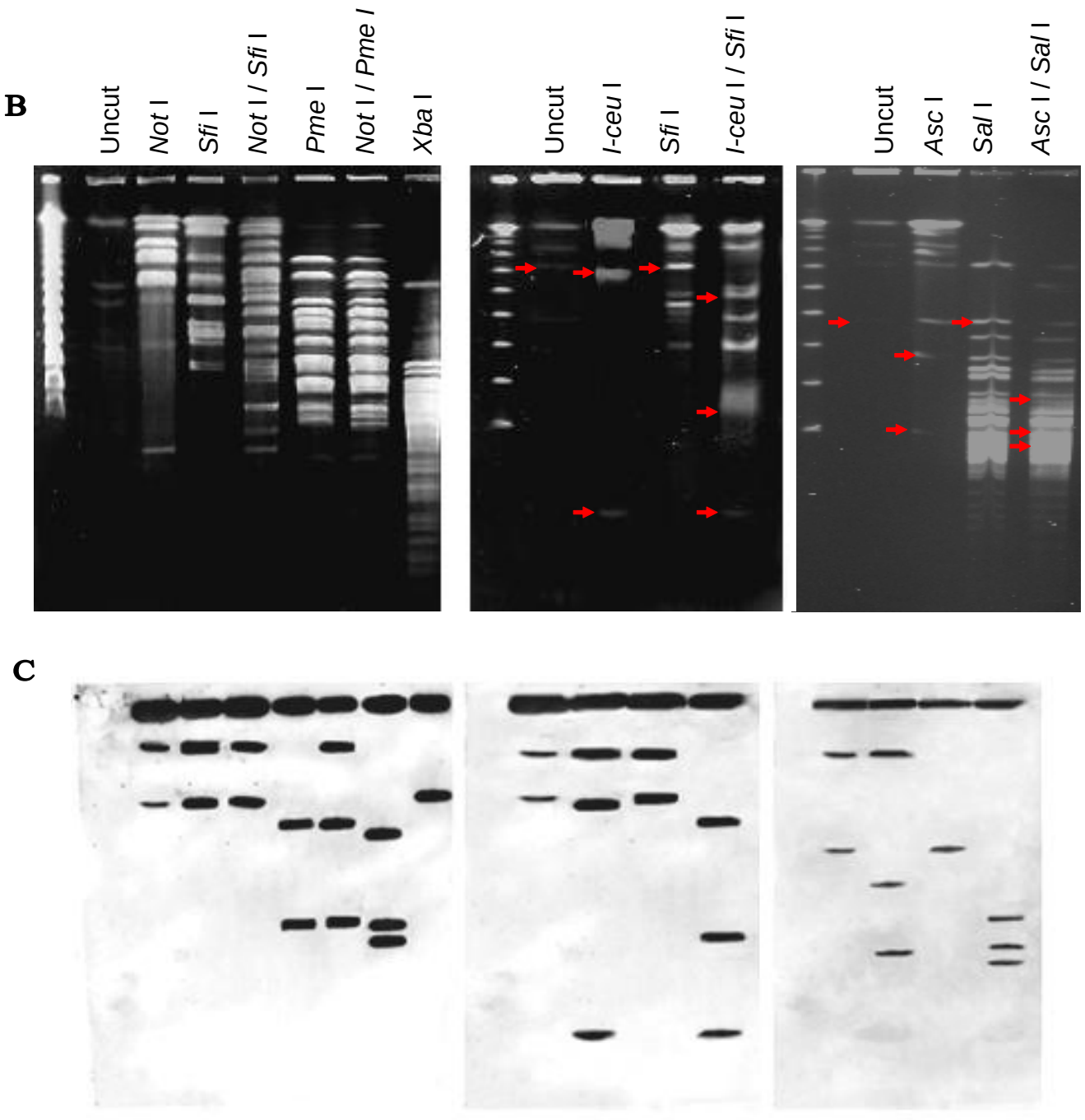


Fig. 3.7. Restriction mapping of replicons. A) The RE recognition sites were mapped *in-silico* for each auxiliary replicon using Vector NTI. The bolded sites were selected to experimentally test the final sequence assembly of each replicon. B) PFGE of RE digests of whole genomic DNAs. Expected DNA band sizes are indicated by arrows and are as follows (rounded to nearest Kb; top to bottom left to right): 361, 361, 361, 306, 55, 309, 52, 275, 52, 34 and 361 Kb; 302, 293, 9, 302, 227, 66 and 9 Kb; 186, 136, 50, 186, 90, 50 and 46 Kb C) Southern blot analysis of RE digests. Membranes were probed with products derived from regions within each expected restriction fragment of each replicon. Left panel pCY360; middle panel BPc2; right panel pCY186. Each lane is labeled with the corresponding RE used. The unlabeled first lane corresponds to the λ ladder - where each band indicates an approximately 50 Kb size increase.

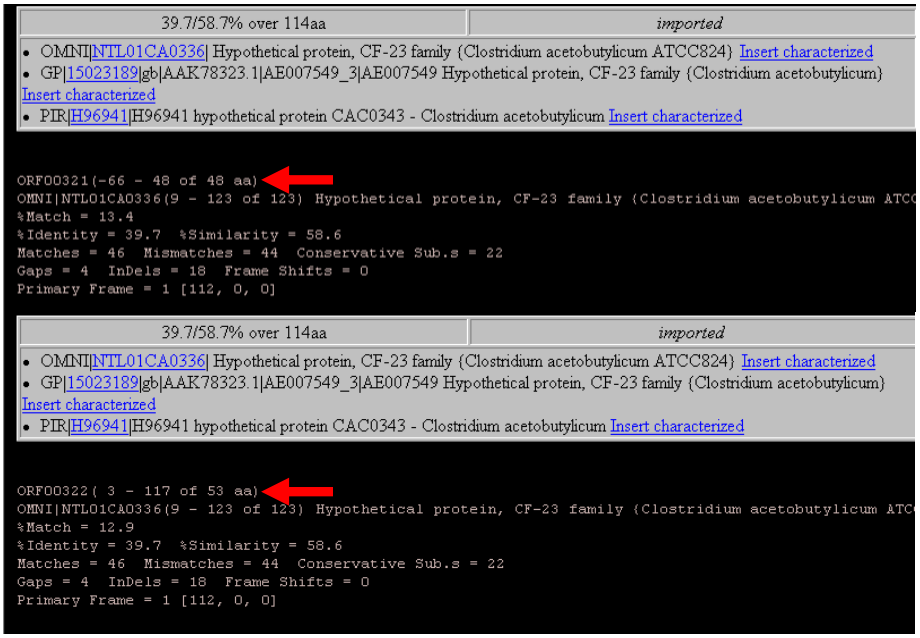
3.6 Quality control of replicon DNA sequences

Analysis of the pCY360 assembly found just 17 base pairs below the Phred 40 confidence value, all within a single region of ~2.5 Kb. This region was PCR amplified, column purified and sequenced using three internal primers. The quality of BPc2 and pCY186 sequences were assessed following array-based pyrosequencing. As pyrosequencing does not result in trace files they could not be assessed by Phred analysis.

Through annotation of the pCY360 sequence (described in Chapter 4) two regions, were identified to contain potential sequence errors. ORF 188 contained a potential frameshift, identified when a BLAST Extend Repraze (BER) alignment (TIGR, 2001) with a resolvase gene from the *Enterococcus faecium* transposase Tn1546 revealed a sudden loss of sequence similarity at amino acid 192, but regained similarity in the -2 frame at the theoretical amino acid 274.7 (Figure 3.8). ORF 336 contained a potential indel that resulted in a premature stop codon. This was identified when a BER alignment with a hypothetical protein, CF23, from *Clostridium acetobutylicum* revealed significant sequence similarity extending beyond ORF 336 and into ORF 337. Similarly, a BER alignment of ORF 336 with the same hypothetical protein revealed significant sequence similarity extending upstream into ORF 335 (Figure 3.9). Primers were designed surrounding each region of interest (the DNA sequence corresponding to amino acids 192 to 275 of ORF 188 and the carboxyl terminus of ORF 336 through the intergenic region and across the amino terminus of ORF 337), and their sequence integrity was analysed by sequencing of the resulting purified, PCR amplified product. In both cases re-sequencing of these regions confirmed the initial sequence as being correct.

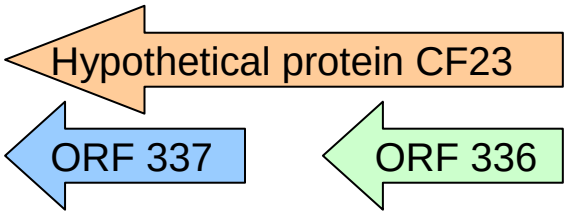
Figure 3.8. Re-sequencing of potential frame shift in ORF 187. A) Alignment of ORF 188 (formerly numbered ORF 171 in a previous draft sequence version) with a resolvase from the *Enterococcus faecium* transposase Tn1546. The alignment reveals a loss of sequence similarity at amino acid 192, however the similarity continues in the -2 frame at the theoretical amino acid 274.7. Primers were designed surrounding this region to determine the integrity of this frame shift. B) An alignment of re-sequencing data with the original sequence. The region pertaining to the potential frame shift (as indicated) was shown to be correct.

A



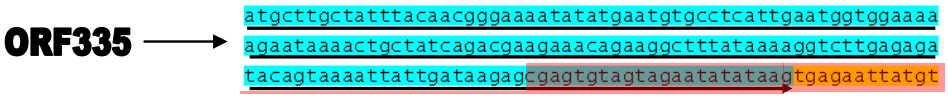
Sequencing data

B

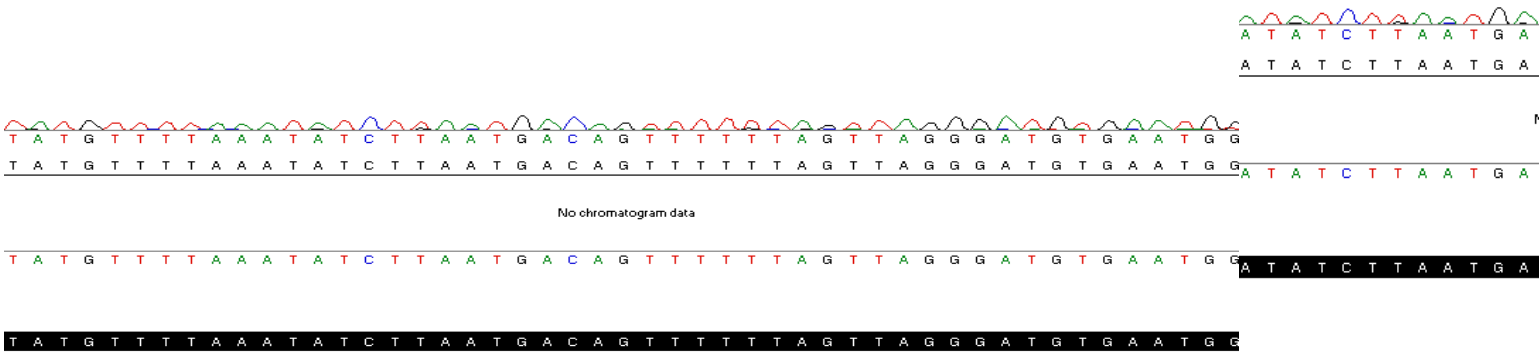


Phase 1 sequence

C



Sequencing data



confirms the stop codon as being correct.

3.7 Discussion

The means of extracting contigs pertaining to certain extrachromosomal elements, such as megaplasmids, from a whole genome sequence has not previously been described in the literature. Methods for „binning" species-specific contigs derived from metagenomic sequencing efforts exist, but these are based on the identification of genomic signatures such as di-, tri-, tetra-, and/or penta-nucleotide frequencies and a knowledge of that organism"s genome composition (Chan *et al*., 2008). At the outset of this project, nothing was known about the auxiliary replicons of *B. proteoclasticus*. Based on the size of the auxiliary replicons it was assumed they were likely megaplasmids. Using the fundamental differences in replicative apparatus between plasmids and chromosomes, a method was derived that allowed the separation of most contigs associated with these auxiliary elements from those of the major chromosome. Genes encoding plasmid replication initiation (Rep) proteins, found in 70% of all megaplasmids currently sequenced, and Partitioning (Par) proteins, found in 82% of all megaplasmids currently sequenced (and present in approximately 25% of megaplasmids not found to encode a Rep protein; reviewed in Section 1.12) were identified. Unlike *rep* genes, *par* genes are found on the major chromosomes of many bacterial species. However, the amino-acid sequence of Par proteins from major chromosomes possess a distinct phylogeny from those derived from plasmids or secondary chromosomes (Gerdes *et al*., 2000). The presence of a gene encoding either a Rep protein, or a Par protein that did not conform to major-chromosome type Par phylogeny, provided a useful indicator of a contig derived from one of the auxiliary replicons. Using this technique, four contigs were identified as likely being derived from the auxiliary replicons of *B. proteoclasticus*. A further 7 contigs were found to be associated with the four Rep- or Par- encoding contigs, as determined by the analysis of fosmid clone terminal sequences.

The largest auxiliary replicon, pCY360, was found to be represented as a single circular contig in the Phase I sequence data. The reason for this is likely to be related to its over-representation in the clone library, where it accounted for 20% of all the clones sequenced. This sequence bias toward pCY360 resulted in an under-representation of the other replicons in the clone library. This in-turn likely exacerbated the sequencing difficulties and complicated the closure of the physical

gaps of BPc2 and pCY186. The non-uniform coverage of these replicons by cloned sequences (Fig. 3.5), which was particularly evident in pCY186 but was also seen in BPc2, suggests the bias to pCY360 alone does not explain the inability to derive their entire sequences in the Phase I data. It has been observed that regions of some microbial genomes (particularly Gram-positive organisms) are not well represented in random DNA libraries (Fraser *et al*., 2002). This can be attributed to regions being difficult to clone, difficult to sequence or difficult to assemble. Cloning bias is a well recognised but poorly understood phenomenon (Carraro *et al.*, 2004, Goldberg *et al.*, 2006, Forns *et al.*, 1997). It can be caused by extracting DNA in the exponential phase, where regions surrounding the origin of replication can be significantly over-represented (Frangeul *et al*., 1999). This problem can be exacerbated by large, low copy number plasmids, whose replication may be asynchronous from that of the chromosome. It is also believed AT-rich (%G+C poor) regions may be less amenable to the cloning process (Frangeul *et al*., 1999). Sequencing difficulties are often related to the presence of secondary structures, %G+C rich regions, large fluctuations in %G+C content or stretches of alternating purine and pyrimidine residues. All of these features have been reported to cause DNA polymerases to stall and arrest (Ishikawa *et al.*, 2003, Sun *et al.*, 2006b, Sherman & Gefter, 1976, Weaver & DePamphilis, 1984). Transposase-encoding genes are known to be surrounded by complex secondary structures (Kleckner, 1981) and their over-representation at contig ends suggests they contribute to the curtailing of sequencing reactions. Genome assembly difficulties are typically related to repetitive elements (Touchman *et al.*, 2007, Gioia *et al.*, 2006, Sun *et al*., 2006b). Ribosomal RNA operons are largely invariant within a genome (Coenye & Vandamme, 2003) and therefore commonly cause difficulties in the assembly process (Carraro *et al*., 2004). They also form significant secondary structures and can therefore be potentially difficult to sequence. In this regard, the presence of two rRNA operons in the BPc2 sequence is likely to have caused the gap found between Contig 107 and Contig 108.

The process of gap closure was made easier for BPc2 by the ability to order contigs and estimate gap sizes by multiplex PCR. The closure of gaps was achieved by conventional, long range and iPCR methods, although no method alone was sufficient to close all gaps. In all cases, primers were designed in regions of high quality sequence (Phred scores > 40) and no closer than 75 bp to contig ends. The latter

parameter was set to address the need to discard the 5' and 3' portions of a sequencing reaction due to their poor quality. This is a well described imperfection resulting from Sanger sequencing, gel-electrophoresis and trace file processing (Ewing *et al*., 1998). The problem occurs at the 5' extremity where the migration of very short fragments is influenced more by the specific nucleotide sequence and the attached dye than the fragments size. This, along with unreacted dye-primer and dye-terminator molecules, typically results in peaks which have a high "noise" background, are unevenly spaced and are often overlapping. Toward the 3' end a similar effect is seen although this is attributable to less accurate trace processing as the relative mass differences between successive fragments decreases. The trace file also becomes less distinguishable from the background as the number of labeled fragments of that size decrease (Ewing *et al*., 1998).

Long-range PCR was attempted for gaps that could not be captured, or closed, by conventional PCR. This approach allowed the capture of gaps that exceeded the upper-limits of conventional PCR extension. Due to the addition of a more processive DNA polymerase, it is also able to enhance amplification over difficult stretches of DNA, such as those described above (McDowell *et al*., 1998). The Eppendorf TripleMaster system uses Taq DNA Polymerase mixed with an undisclosed, highly processive, thermostable polymerase that has 3‟ to 5‟ proofreading (exonuclease) activity, along with an unspecified polymerase-enhancing factor that is claimed to provide an extremely high extension rate and high proofreading-assisted fidelity. Additionally, the buffer is claimed to reduce template and amplicon degradation, caused by depurination, due to its unique ability to maintain pH at high temperatures (Eppendorf, 2000). Long range PCR led, or contributed to the capture of 4 gaps, 3 of which neared the theoretical upper-limits of conventional PCR. Inverse PCR, a useful technique that requires no knowledge about the down-stream DNA sequence (Benkel & Fong, 1996), contributed to the closure of 2 gaps. However, it was best used as a final resort due to the laborious and potentially repetitive nature of the protocol (Benkel & Fong, 1996).

Unlike BPc2, the contig order of pCY186 was not discernible via the multiplex approach. It therefore presented a much greater technical challenge that ultimately required the use of array-based pyrosequencing to solve. The decision to utilise array-

based pyrosequenceing in the closure of the *B. proteoclasticus* genome was based on the speed and cost-effectiveness of this technology compared to continued PCR-based gap closure techniques, which in many cases, including the pCY186 gap between Contigs 3 and 5, were consistently resulting in failure. The combined capacity of the Sanger and pyrosequencing methods to close genome sequences has previously been demonstrated (Goldberg *et al*., 2006). Array-based pyrosequencing provides a method of obtaining high coverage of short sequence reads without the need for cloning (Edwards *et al.*, 2006, Turnbaugh *et al.*, 2006). It thereby eliminates cloning bias and, through the short sequence reads, helps to mitigate the negative effects of complex secondary structures on the sequencing reaction (McMurray *et al*., 1998). This technique resulted in the completion of the final gap of the pCY186 replicon. All sequences were finished to a Phred 40 (pCY360) or Q40 (BPc2 and pCY186) standard, both equivalent to a maximum of 1 potential error per 10,000 sequenced bases. The RE mapping and Southern blot analysis of each replicon provided strong evidence that the three bands observed in pulsed-field gels of intact whole-genomic DNA corresponded with the DNA sequences derived from the sequencing process. It also corroborated the final assembly of each replicon being correct. To further validate the derived sequences, anomalies identified during annotation were re-sequenced. In both cases of identified sequence anomaly, the initial sequence was found to be correct. The presence of two rRNA operons was verified by digestion with the RE I-*Ceu*I followed by Southern blot analysis. I-*Ceu*I, derived from a mobile intron in the chloroplast 23*S* ribosomal RNA (*rrl*) gene of *Chlamydomonas eugametos*, recognises and cuts at a 26-bp site in the *rrl* gene of many bacteria (Liu *et al*., 1993). It is commonly used in conjunction with PFGE to map rRNA operons. Collectively the RE mapping and Southern blotting data show that the sequencing data is robust.

3.8 Summary

The *B. proteoclasticus* genome is spread across four independent replicons that include the 3.5 Mb major chromosome and three auxiliary replicons ranging in size from 186 to 361 Kb. Using the fundamental differences between the replicative apparatus of plasmids and major chromosomes, the sequence contigs of the auxiliary replicons were extracted from the Phase I genome sequence. In some cases, multiplex PCR enabled the ordering of contigs and the estimation of the corresponding gap sizes. The combination of conventional, long range and inverse PCR techniques, along with the screening of sequences derived from a BAC library, was sufficient to close most of these gaps. However, one large gap remained after these methods and was eventually closed by array-based pyrosequencing. RE mapping, along with Southern blotting, confirmed the sequences as belonging to the auxiliary replicons observed in the pulsed-field electrophoresis gels and provided strong support for the final assembly of each replicon‟s sequence. Sequence anomalies identified through annotation, such as potential indels, frame shifts or the presence of two rRNA operons upon BPc2, were re-evaluated and found to be consistent with the original sequence data, further supporting the sequence of each replicon as being robust.

4 pCY360

4.1 Introduction

The *B. proteoclasticus* genome is spread across 4 independent replicons, a chromosome and 3 large auxiliary replicons ranging in size from 186 kb to 361 kb. The largest of these auxiliary replicons, designated pCY360, accounts for almost 10% of the entire *B. proteoclasticus* genome and is the fourth largest megaplasmid currently reported in a Gram-positive bacterium. Many characterised megaplasmids have been found to confer significant traits to the host organism. In Gram-positive hosts, megaplasmids have been found to confer traits such as the ability to degrade polychlorinated biphenyl (PCB) compounds (Shimizu *et al*., 2001) or atrizine (Mongodin *et al*., 2006). Large extra-chromosomal DNAs of approximately 250 x 10^6 daltons have previously been observed in a number of *B. fibrisolvens* strains (Teather, 1982), however, their large size precluded detailed analysis of their structures or functions at that time. Thus the complete sequencing of the *B. proteoclasticus* pCY360 megaplasmid is the first detailed analysis of these large extrachromosomal DNAs from *Butyrivibrio* or *Pseudobutyrivibrio* species.

This chapter describes the detailed analysis of the pCY360 megaplasmid, and the attempted elucidation of its function and origin.

4.2 Sequence analysis of pCY360

The complete sequence of the pCY360 megaplasmid is 361,397 bp in length, which is consistent to within 0.04% of the size estimated from PFGE. The pCY360 megaplasmid has a %G+C content of 38.95% (39.55% in coding regions, 36.03% in non-coding regions), which is similar to that of the major chromosome (overall %G+C is 40.2%). GLIMMER analysis (Salzberg *et al*., 1998) identified 390 potential ORFs, however 20 ORFs were eliminated because they encoded proteins smaller than 50 amino acids, lacked a significant match to any previously described gene, gave no hits to any Hidden Markov model (HMM), lacked a discernable ribosome binding sequence, and one or both of the following: a greater than 10% deviation from the overall average %G+C content of the replicon, or overlap with a larger ORF. Additionally, 19 ORFs were identified manually that appeared to have been missed by the GLIMMER analysis. This gives a total of

389 ORFs likely to be genuine protein coding regions in pCY360 (listed Appendix V). The predicted ORFs have an average size of 743 bp, although the majority of ORFs are much smaller (median size 515bp; Fig. 4.1). This gives pCY360 a gene density of 1.26 genes / kb and gene coverage of 80% of the megaplasmid. The average ORF size appears to be greater between ORFs 205 and 318 (Fig. 4.1), this corresponds to the second third of the megaplasmid. Almost all (89%) of the pCY360 ORFs are located on the Watson strand of the DNA molecule and analysis of third-position GC skew shows the replicon has no discernable GC skew (Fig. 4.2).The distribution of predicted start codons is typical, with 97% ATG, 2% GTG and 1% TTG. The proteins encoded by pCY360 ORFs have an average isoelectric point (pI) of 6.06 (median 5.02; Fig. 4.1). The majority of these predicted proteins (296, 76%) have no significant sequence similarity to any previously described genes, while a further 26 (6.7%) match proteins of unknown function and are described as conserved hypothetical proteins. pCY360 encodes 31 (8%) proteins predicted by trans-membrane Hidden Markov Models (TmHMMs) to span the membrane more than once and a further 35 (9%) contain signal peptide sequences. During the course of this thesis, 13 of the predicted pCY360 proteins were identified by tandem mass spectroscopy analysis of liquid chromatography fractions from 1-dimensional protein gels (Dunne, in preparation). These identifications resulted in 4 hypothetical proteins being reclassified as uncharacterised proteins and three conserved hypothetical proteins being reclassified as uncharacterised conserved proteins.

4.3 pCY360 origin of replication

A site spanning nucleotides 352,152 to 2,253 was identified as likely possessing the origin of replication (*oriR*) (Fig. 4.2). The region encodes a putative replication initiation protein, repB (ORF1) and a plasmid partitioning protein, parA (ORF388) separated by a hypothetical protein (ORF389). Two large inverted repeats (IRs) of 89 bp (IR1) and 24 bp (IR2) were found within this region. The divergent nature of the ORFs surrounding IR1 suggests that this region is likely to contain the *oriR*. Within this region an 18 bp imperfect direct-repeat (DR1) and 7 potential *dnaA* boxes were also identified. Moreover, a 366 bp A+T rich (13.6% G+C) region spanning both IR1, DR1 and two of the potential *dnaA* boxes, was identified that bears 48% nucleotide identity to the putative *oriR* of *B. fibrisolvens* plasmid,

pRJF1 (Hefford *et al*., 1993). Lying outside of this central region are three ORFs. ORF5 is a XerC-family integrase/recombinase and ORFs14 and 15, represent a putative Pin domain toxin-antitoxin (TA) operon which, along with *repB* and *parA*, are implicated in plasmid retention. Collectively these genes appear to constitute the pCY360 plasmid „backbone".

4.4 Conjugative transfer-related proteins

Several ORFs related to conjugative transfer and Type IV pili formation are encoded by pCY360. Most of these ORFs are clustered in a 35 Kb region spanning nucleotides 201,461- 236,288 (Fig. 4.2). The ORFs identified include members of both the mating pair formation (mpf) complex (ORF238, TraE; ORF249, Type IV pilus subunit; ORF264, ATPase involved in pili biogenesis), and the relaxosome (ORF172, putative MobA and ORF292 putative recD/TraA helicase), as well as the coupling protein TraG (ORF242). Two lipoproteins (ORFs 253 and 257), 4 novel proteins containing signal peptides (ORFs 245, 246, 247 and 252) and 6 novel proteins predicted by a TmHMM (Krogh *et al*., 2001) to have two or more trans-membrane spanning domains (ORFs 239, 240, 241, 244, 248 and 259) are also found in the central 35 Kb region. One of the signal peptide-containing proteins (ORF247) matches over a short region to a structural lytic transglycosylase from *Enterobacter* phage P1 (29% ID over 90 aa). Two signal peptidases (*traF*, ORF232 and *sipB*, ORF279) also reside either side of the putative conjugative transfer region. Both the *mobA* and *recD*/*traA* helicase ORFs (ORFS 172 and 292) potentially encode relaxase proteins. To distinguish which ORF more likely fulfilled this function, or to determine if this function may have been redundant, both ORFs were searched for motifs conserved in relaxases from other Gram-positive bacteria (Grohmann *et al*., 2003). An alignment of the aa sequence encoded by ORFs 172 and 292 was made with those of pMRC01 (*Lactococcus lactis*), pRE25 (*Enterococcus faecalis*) and pSK41 (*Staphylococcus aureus*; Fig. 4.3). This revealed that Motif I, (characterised by a conserved tyrosine residue), was present in the recD/TraA helicase but not in MobA. Motif III, (characterised by two conserved histidine residues) was absent from both potential relaxases. A number of transfer origin (*oriT*) candidates were identified by searching the pCY360 sequence for the Gram positive core *oriT* consensus sequence (Grohmann *et al*., 2003; Fig 4.4).

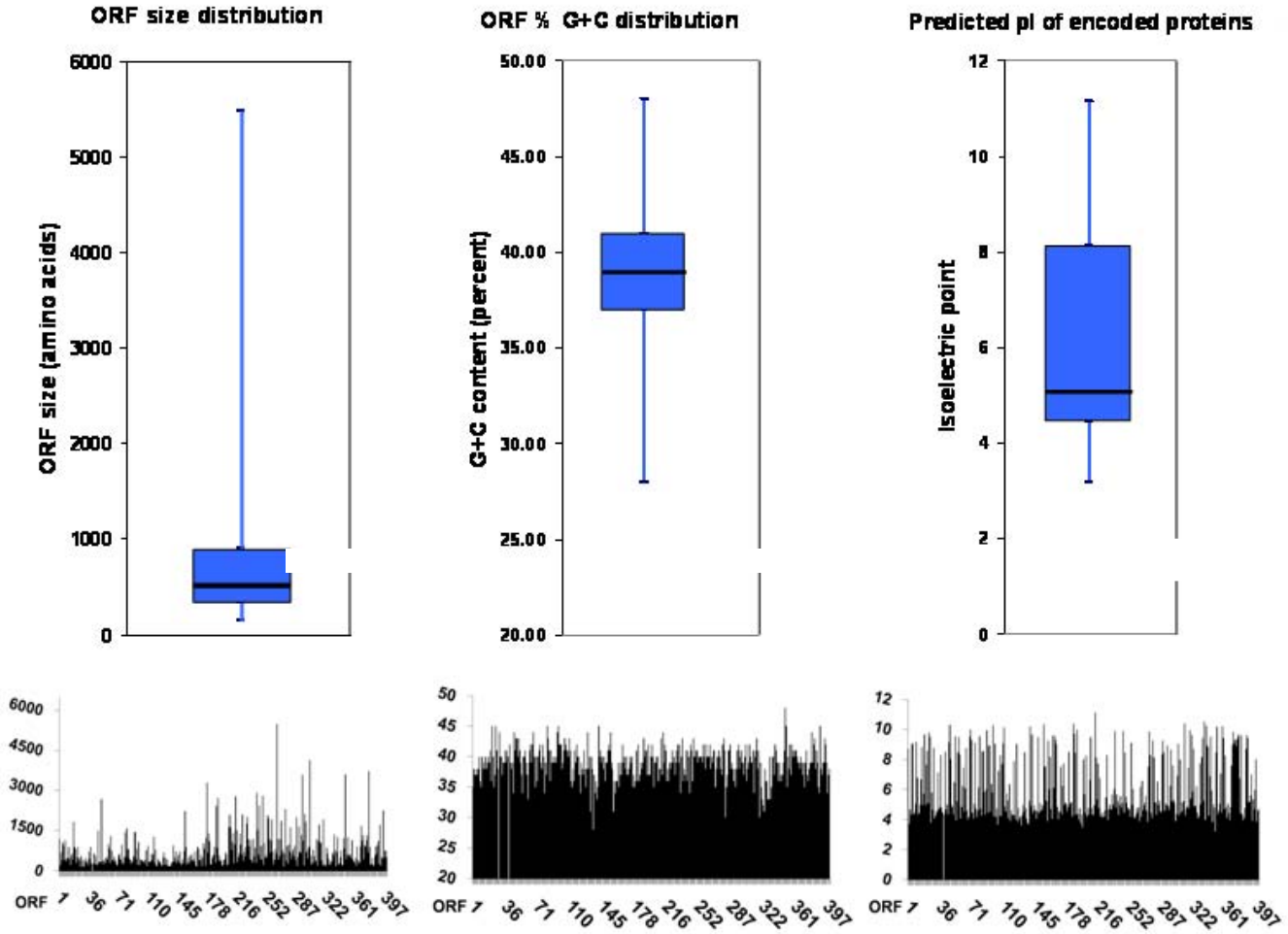


Figure 4.1. ORF and encoded protein composition . Box plots show size, %G + C and isometric point distribution for each quartile of the data. The histograms show the distribution of features from ORF 1 – 389. Figures generated in MS office Excel.

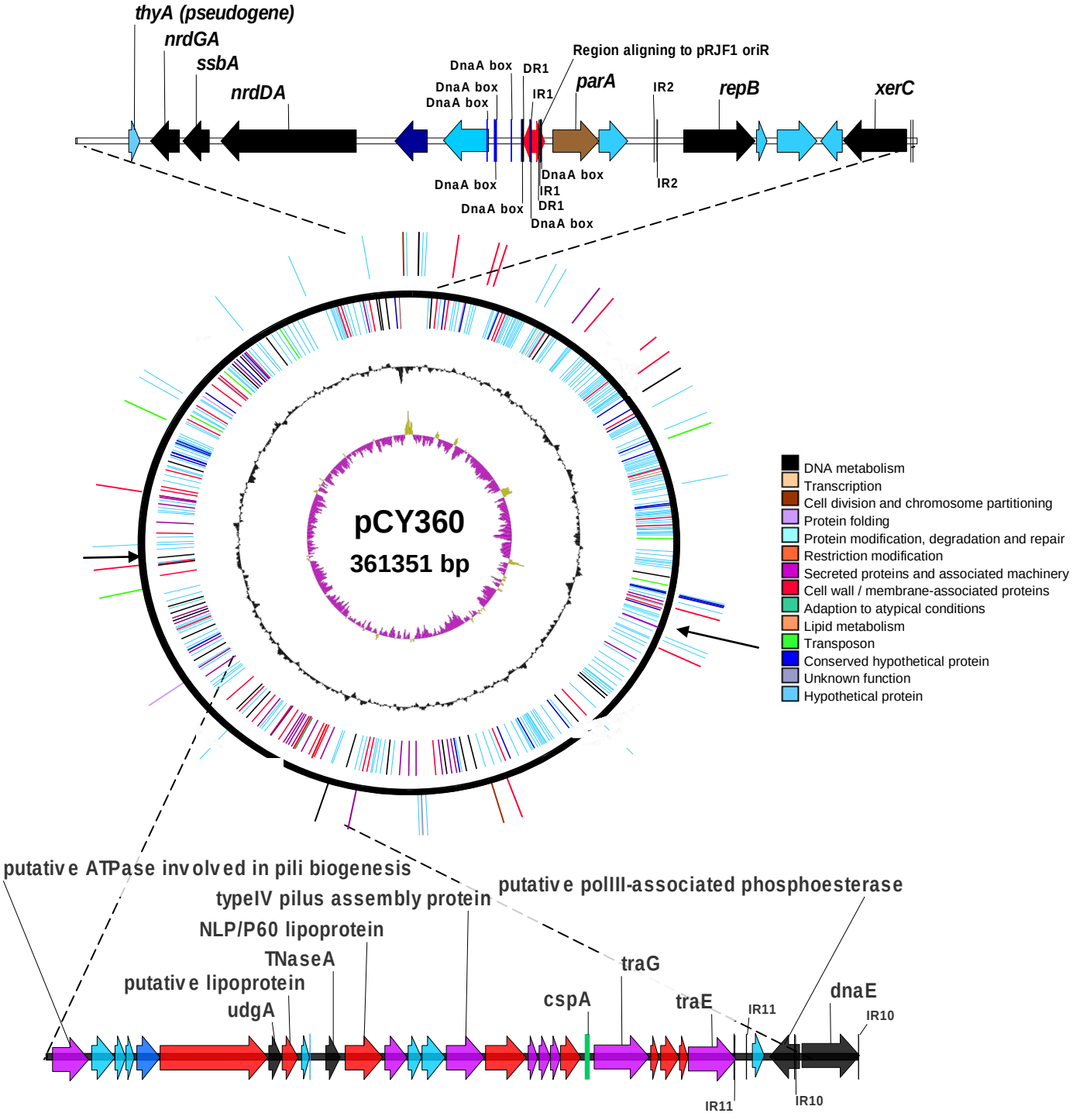


Figure 4.2. Map of pCY360 predicted ORF function. The circles from outside to inside indicate: genes transcribed from the Crick strand; genes transcribed from Watson strand; the deviation of %G+C from the mean; and GC skew. Individual genes are colour-coded according to their predicted COG functional role category. In cases where a gene falls into more than one role category the most descriptive role category is assigned. The putative *oriR*-containing loci (top) and central conjugative transfer loci (bottom) are enlarged. DnaA boxes, the direct repeat (DR1) and inverted repeats (IR1, IR2, IR10 and IR11) are shown as well as the region aligning to the putative origin of the *B. fibrisolvens* plasmid pRJF1. The positions of the potential relaxase encoding genes are indicated with arrows. Figures were generated using Vector NTI.

```
pmrc01          MAIFHMNFSNISAGKGRSAVASAS---YRSGEKLYSEMENKTYFYN-RSVMPESFILLPE 56
pRE25           -----MTIAKRENGK-RSLIAMAS---YRSGEKLYSELYEKTNLYNHRTVKPEAFILKPD 51
pSK41           MAMYHFQNKFVSKANGQSATAKSA---YNSASRIKDFKENEFKDYS-NKQCDYSEILLPN 56
mobA            --------------------MRAT---MRNSRSGYAKHNDRDFDIN----------SPD 26
TraA/RecD       MEKFCGKFIRSESIIGTYFILESKGKIYKCFGIAPELYPGTPIVVTGKAEFEDSKNITVE 60

pmrc01          NA-PEWAKDRQKLWNEVEAVDRKVNSRYAKEFNVALPIELSEDEQKELLTEYVQKIFVDK 115
pRE25           YV-PNEFLDRQTLWNKMELAEKSPNAQLCREVNVALPIELNNSDQRMLIEDFVKDNFVNE 110
pSK41           NA-DDKFKDREYLWNKVHDVENRKNSQVAREIIIGLPNEFDPNSNIELAKEFAES-LSNE 114
mobA            ND-GHIDADRQYLNTYWHCYG----------------------------------EE 48
TraA/RecD       NIGLDTTDDQKMIYFLSGRCFKKVGMVAARRIVESLNKRTKESGLTSYGELKYEDIQECL 120

pmrc01          GMVADVAIHRDHDENPHAHVMLTNRPFNADGSWGQKAKKEYILDENGNKTYTANGHARSR 175
pRE25           GMIADVAIHRDDENNPHAHIMLTMREVDSEGNILNKSHRIPKLDENGNQIFNEKGQRVTV 170
pSK41           GMIVDLNIHKINEENPHAHLLCTLRGLDKNNEFEPKRKGNDYIR---------------- 158
mobA            KLTFAQAEKKYYEEHYKASLEESNR---------------------------------- 73
TraA/RecD       KEAKTPKEAALKITLALTGIQERSRLYEKIKEFGGRLPDAESLYKKYFTGAWQAFKNDPY 180
```

Figure 4.3. Conserved sequence motifs in Gram-positive relaxases. Alignment of potential relaxases proteins (MobA and RecD/TraA) from pCY360 with relaxases from other Gram-positive bacterial plasmids (pMRC01, pSK41 and pRE25). Conservation of Motif I in the putative RecD/TraA helicase protein from pCY360 is shown highlighted in yellow but is not present in the putative MobA. Neither RecD/TraA nor MobA showed conservation of Motif III (highlighted in green).

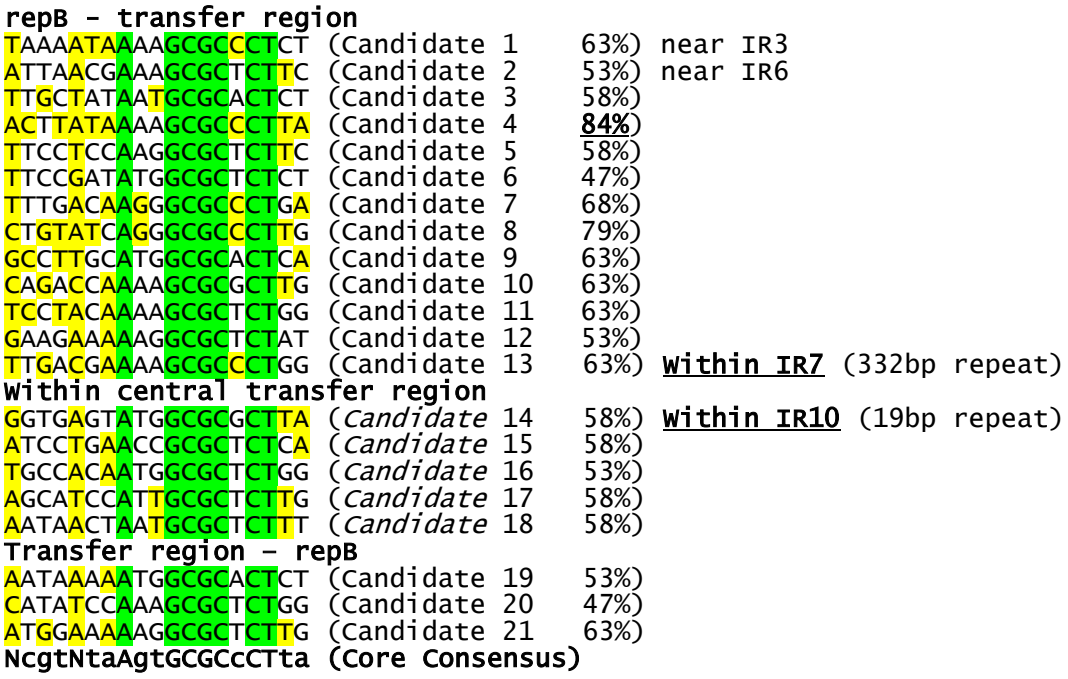


Figure 4.4. The o*riT* candidate sequences. All pCY360 sequences conforming to the conserved nucleotides of the core consensus *oriT* sequence determined by Grohmann *et al* (2003) are shown. Absolutely conserved nucleotides are highlighted in green, strongly conserved nucleotides are highlighted in yellow. The position of each sequence is shown relative to the central 35 Kb transfer region and their % identity to the consensus sequence is shown in brackets. Their position relative to inverted repeats is also indicated where relevant.

Five of these candidate sequences, each containing all of the absolutely conserved nucleotides used to derive this core consensus sequence, are present in the central 35 Kb conjugative transfer region. An *oriT* candidate was also found upstream of each of the potential relaxases mobA (ORF 172) and recD/TraA (ORF292) (labeled Candidates 12 and 19, respectively). As all previously identified *oriT*'s are found adjacent to inverted repeats, the proximity of each *oriT* candidate to such secondary structures was evaluated. Two of the candidates, one within the central transfer region (Candidate 13) and one approximately 50 Kb upstream of this region (Candidate 14) were found to be within inverted repeats (IR10 and IR7 respectively). A further two candidates (Candidates 1 and 2) were found to be in close proximity to IR3 and IR6.

4.5 The predicted impact pCY360 on the membrane and extracellular environment of *B. proteoclasticus*

In addition to the 31 proteins predicted to span the membrane two or more times, pCY360 additionally encodes 53 proteins predicted to span the membrane once and several proteins that are likely to influence the topology of the cell membrane. ORF 18 is predicted to encode an ATP-dependent zinc metalloproteinase, FtsH, ORF 311 encodes a putative membrane bound di-guanylate cyclase and ORFs 335 and 339 both encode putative sortase B proteins. In addition to the 35 proteins predicted to be secreted by either a signal peptidase I (32) or signal peptidase II (3) cleavage sites, pCY360 also encodes two signal peptidase I proteins (ORFs 232 and 279). Collectively 119 (31%) of the pCY360 replicon may contribute to sensing and responding to the extracellular environment.

4.6 pCY360 contains genes described in the Minimal Gene Set

Of the 67 plasmid ORFs which are able to be assigned a putative function, 8 genes (12%) are found in the Bacterial Minimal Gene Set (Koonin, 2000; listed Table 4.1). However, none of these ORFs are unique within the *B. proteoclasticus* genome, each having at least one paralogue encoded on the *B. proteoclasticus* B316^T 3.5 Mb major chromosome. Several of these ORFs (ORFs 383, *nrdGA*; 385, *nrdDA*; and 384, *ssbA*), all have functions predicted to be involved in DNA metabolism and are found clustered near the predicted *oriR* of pCY360 (Fig. 4.2). In addition, a truncated thymidylate synthase pseudogene (ORF 382) with strong

similarity to the N-terminal ~60 amino acids of characterised thymidylate synthases from *E. coli* (40% aa identity; Belfort *et al*., 1983) and *Bacillus subtilis* (43% aa identity; Tim and Borriss, 1995), is also found within this cluster.

4.7 Predicted contributions to enzymatic pathways

The pCY360 replicon encodes 13 ORFs that can be ascribed to enzymes defined by Enzyme Commission (EC) numbers (Table 4.2). ORF 211, a putative poly(ADP-ribose) polymerase (PARP) and ORF 254, a thermonuclease, are uniquely encoded within the genome by pCY360. PARP, an enzyme typically encoded exclusively by eukaryotic organisms, shows a weak (28% identity, 43% similarity) but full length alignment to the PARP of *Microscilla marina* ATCC 23134 along with a weak hit to the PFam describing the poly(ADP-ribose) polymerase, catalytic domain (PF00644).

4.8 Transposases

A total of 11 transposase genes were identified within the pCY360 sequence. Each can be assigned to one of three transposase families; IS4 (ORFs 90, 91 288 and 327), IS200 (ORF362) and IS605 (ORFs 95, 136, 151, 331, 343 and 361). Five of the pCY360 transposases share 22-34% amino acid identity to transposases found on the major chromosome (ORFs 136, 151, 331, 343 and 361). Three of these transposases also share 30-34% amino acid identity to transposases found on BPc2 (ORFs 331, 343 and 361).

Table 4.1. pCY360 genes belonging to the Bacterial Minimal Gene Set.

ORF*	Size (aa)	Putative function	Best Blast match	E-value	% ID	GenBank Accession
18	972	ATP-dependent Zinc metalloproteinase, FtsH	*Symbiobacterium thermophilum*	2 e-85	39%	YP_077024
235	215	DNA polymerase III α subunit	*Microbulbifer degradans*	2 e-97	30%	ZP_00314789
258	408	Uracil DNA glycosylase, udgA	*Fusobacterium nucleatum*	2 e-59	51%	ZP_00143950
338	210	Putative trigger factor protein	*Nocardia farcinica*	1 e-8	24%	YP_117541
346	608	Guanylate kinase, gmkA	*Clostridium acetobutylicum*	6.4 e-21	39%	AAK78279
383	751	Anaerobic ribonucleoside triphosphate reductase activating protein, nrdG	*Chromobacterium violaceum*	1 e-105	34%	AAQ60084
384	147	Single-strand binding protein, ssb	*Clostridium thermocellum*	1 e-27	46%	ZP_00504272
385	176	Anaerobic ribonucleoside triphosphate reductase, nrdD	*Fusobacterium nucleatum*	1 e-40	46%	AAL94518
382	61	Thymidylate synthase, thyA (Pseudogene)	*Streptococcus pneumoniae*	1 e-10	64%	ZP_00404657

* The ORF number, size, proposed function, species for which the best BLASTP match was obtained, the corresponding E-value, the percent amino acid identity and accession numbers are shown for each ORF of the Bacterial Minimal Gene Set.

4.9 Phylogenetic relationship of *repB* genes

Previously plasmids have been grouped based on their ability to co-exist in the same cell, in what are known as incompatibility classes (Novick & Hoppensteadt, 1978). Replication initiator proteins, such as RepB, are integral to the initiation of plasmid replication and regulation of its copy number (del Solar *et al*., 1998). Their regulation of the plasmid copy number is fundamental to determining plasmid incompatibility groups. To obtain an insight into the evolutionary history and replicative mechanisms of *B. proteoclasticus*‟s replicons a phylogenetic tree was constructed. Rep proteins from pCY360, BPc2 and pCY186 were aligned with closely related Rep proteins (as determined by heuristic alignments using BLASTP; pMBO-1 *Moraxella bovis*; pag6, *L. lactis*; pmvscs1, *Mannheimia varigena*; pjr2, *Pasteurella multocida*; pyc, *Yersinia pestis*; phs129, *Haemophilus somnus*; pBM19, *Bacillus methanolicus*; pmd136, *Pediococcus pentosaceus;* and phcm1, *Salmonella enterica*), Rep proteins from plasmids hosted by phylogenetically related- (pRJF1, pRJF2 and pOM1, *B. fibrisolvens*; pSOL1, pCLI, pmcf1, Clostridia), or rumen-inhabiting-bacteria (pRAM4, pONE429, pONE430), other megaplasmids (Ti, pSOL1, pMOL30, phcm1 and pSYMA), characterised rolling circle- (pOM1, pTA1060, pC194) and theta- (pLS32, pWV02 and pAMbeta1) replicating plasmids and plasmids that utilise iteron- (pLS32 and pAMbeta1) or RNA-binding mechanisms (ColIB-P9 and R1) (Fig 4.5). The amino acid sequence of a MobA protein from *Clostridium butyricum* was used as an outgroup. Mob proteins, like Rep proteins, encode a site-specific topoisomerase-like activity and have been shown to be distantly related (Ilyina & Koonin, 1992). Sequences were aligned using DIALIGN2, which was selected for the alignment, due to the number of insertions and deletions identified between RepB sequences, which makes them less suited to programs like CLUSTAL that utilise a global alignment method. DIALIGN searches for all ungapped pair-wise aligned regions and identifies the set of non-overlapping conserved blocks which have the highest similarity score. The alignment was then filtered to remove regions with little or no conservation. A tree was constructed using Split Decomposition, a program designed in 1992 by Bandelt and Dress (Bandelt & Dress, 1992) to identify if groups are related, without assuming the data is bifurcating and forms a tree. The Rep protein from pCY360, along with those from the co-residing BPc2 and pCY186 replicons formed a clade, distinct from other plasmid Rep proteins (Fig.

4.5a). The Rep proteins from the auxiliary replicons of *B. proteoclasticus* appeared closer to plasmids that utilise an iteron-binding mechanism than an RNA-binding mechanism. The most closely related Rep proteins were found in the phylogenetically-related theta-replicating plasmids from *B. fibrisolvens* (pRJF1 and pRJF2) and the *Salmonella enterica* megaplasmid phcm1.

4.10 Codon usage

The codon usage preferences of each of *B. proteoclasticus*‟ auxiliary replicons were compared to that of the 3.5 Mb major chromosome (Fig 4.6a) to look for possible translational amelioration. The analysis revealed only minor differences in codon usage frequencies between each auxiliary replicon and the major chromosome. A further comparison between each replicon, *B. fibrisolvens* (the most closely related species for which sufficient sequence is available for analysis), and the *B. proteoclasticus* major chromosome shows that pCY360 (average deviation from chromosome 0.18%), BPc2 (average deviation 0.09%) and pCY186 (average deviation 0.17%) were more similar in their codon usage pattern to the *B. proteoclasticus* major chromosome than any of the *B. proteoclasticus* replicons were to the most closely related species for which sufficient sequence was available for such analysis, *B. fibrisolvens* (average deviation 0.35%; Fig 4.6b).

Table 4.2 pCY360 ORFs that encode enzymes assigned an EC number

ORF*	Putative function	Best Blast match	E-value	E.C number
42	Ribonuclease H	*Clostridium tetani*	5 e-36	3.1.26.4
150	NAD-dependent DNA ligase	*Thermoanaerobacter tengcongensis*	5 e-94	6.5.1.2
162	Serine / Threonine protein phosphatase	*Mycobacterium flavescens*	2 e-4	3.1.3.16
211	Putative poly(ADP-ribose) polymerase*	*Drosophilla melanogaster*	2 e-9	2.4.2.30
232	Signal peptidase I	*Clostridium perfringens*	4 e-21	3.4.21.89
235	DNA polymerase III, α-subunit	*Pseudomonas aeruginosa*	3 e-89	2.7.7.7
236	Putative DNA pol III-associated phosphoesterase	*Synechocystis sp.*	9 e-23	3.1.3.15
254	Thermonuclease*	*Staphylococcus epidermidis*	2 e-23	3.1.31.1
265	Serine / Threonine protein phosphatase	*Clostridium acetobutylicum*	8 e-19	3.1.3.16
279	Signal peptidase I	*Oceanobacillus iheyensis*	5 e-14	3.4.21.89
346	*gmkA*	*Clostridium acetobutylicum*	6 e-21	2.7.4.8
383	*nrdGA*	*Fusobacterium nucleatum*	4 e-39	1.97.1.4
385	*nrdDA*	*Chromobacteria violaceum*	1 e-105	1.17.4.2

*The ORF number, proposed function and EC number are shown. ORFs present only on pCY360, are indicated with an asterisk.

A

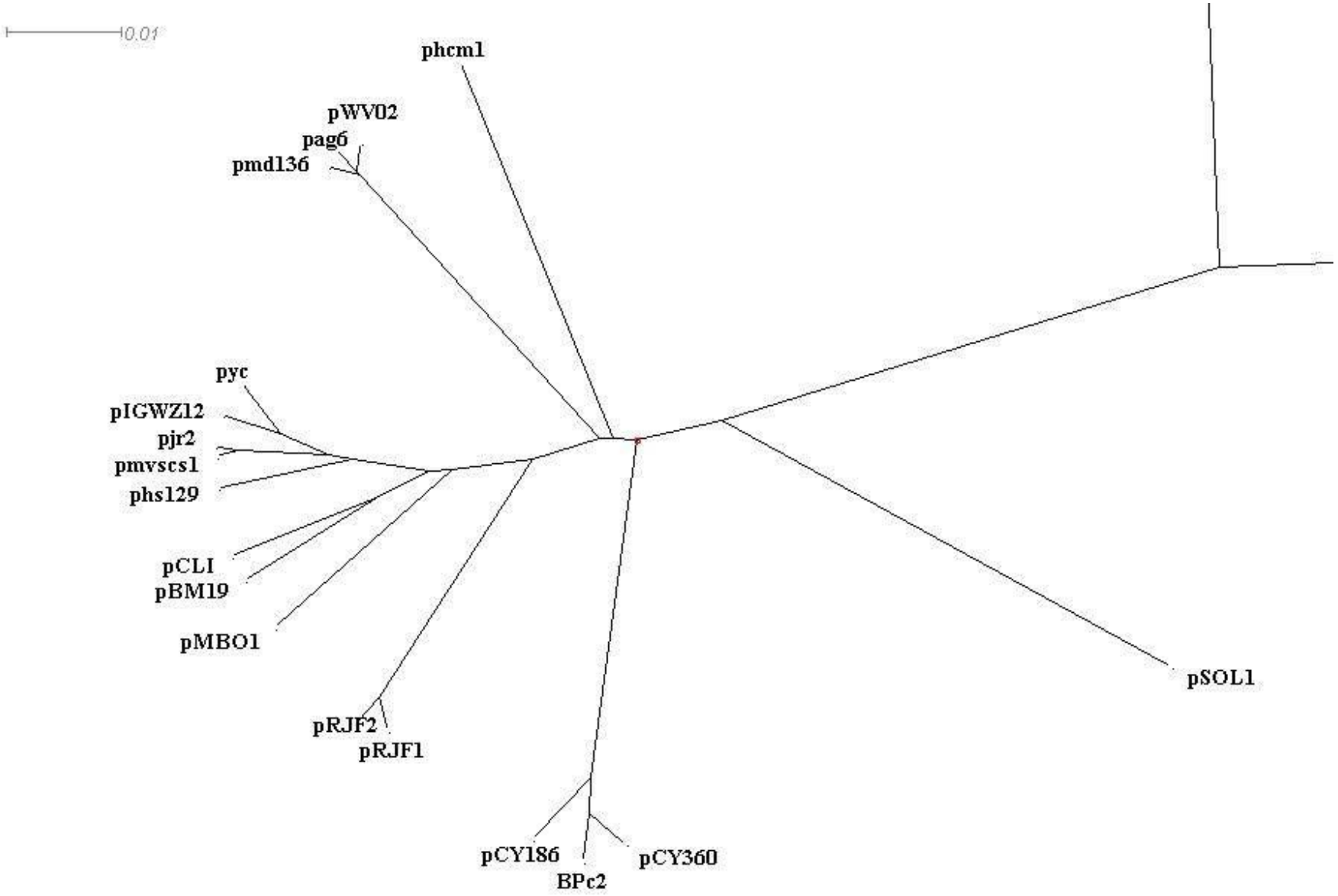
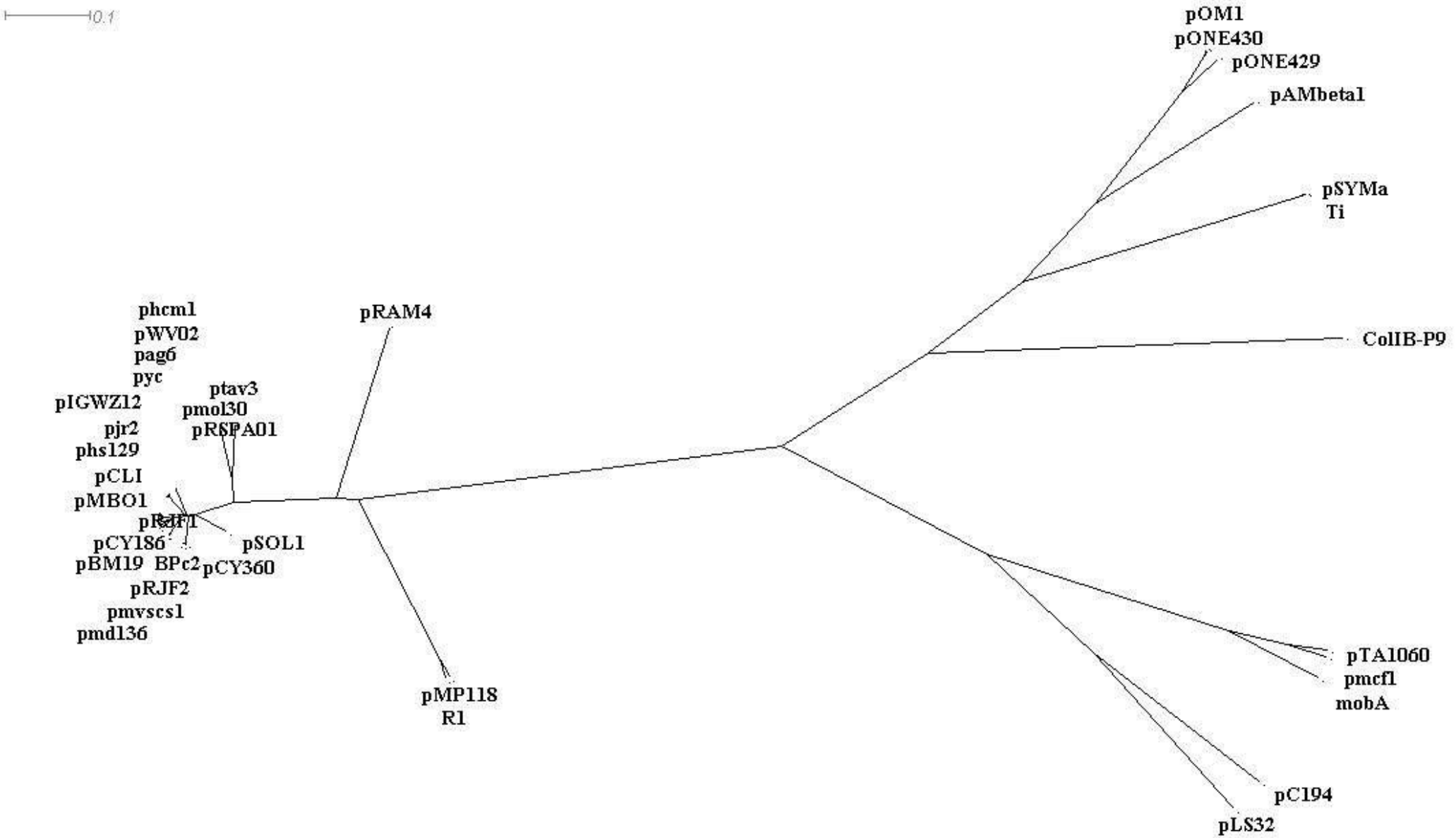


Fig. 4.5 Phylogenetic tree of RepB proteins. RepB amino acid sequences were aligned using Dialign (Morgenstern, 2004). The tree was constructed from this alignment using Splitstree4 (Huson, 1998). RepB amino acid sequences for comparison were from the most closely related RepB sequences as determined by BLASTP scores, megaplasmid sequences, plasmids hosted by phylogenetically- and environmentally-related bacteria, rolling circle- and theta- replicating plasmids and plasmids that utilise iteron- or RNA- binding mechanisms The complete unrooted tree is shown (B) along with an amplified view of the Rep-proteins that cluster closest to those of *B. proteoclasticus* (A). Scale bar shows 1 substitution per 100 bases (A) or per 1Kb (B).

B


0.1
pOM1
pONE430
pONE429
pAMbeta1
pSYMa
Ti
ColIB-P9
pTA1060
pmcf1
mobA
pC194
pLS32
pMP118
R1
pRAM4
phcm1
pWV02
pag6
pyc
pIGWZ12
ptav3
pmol30
pRSPA01
pjr2
phs129
pCLI
pMBO1
pRJF1
pCY186
pSOL1
pBM19
BPc2
pCY360
pRJF2
pmvscs1
pmd136

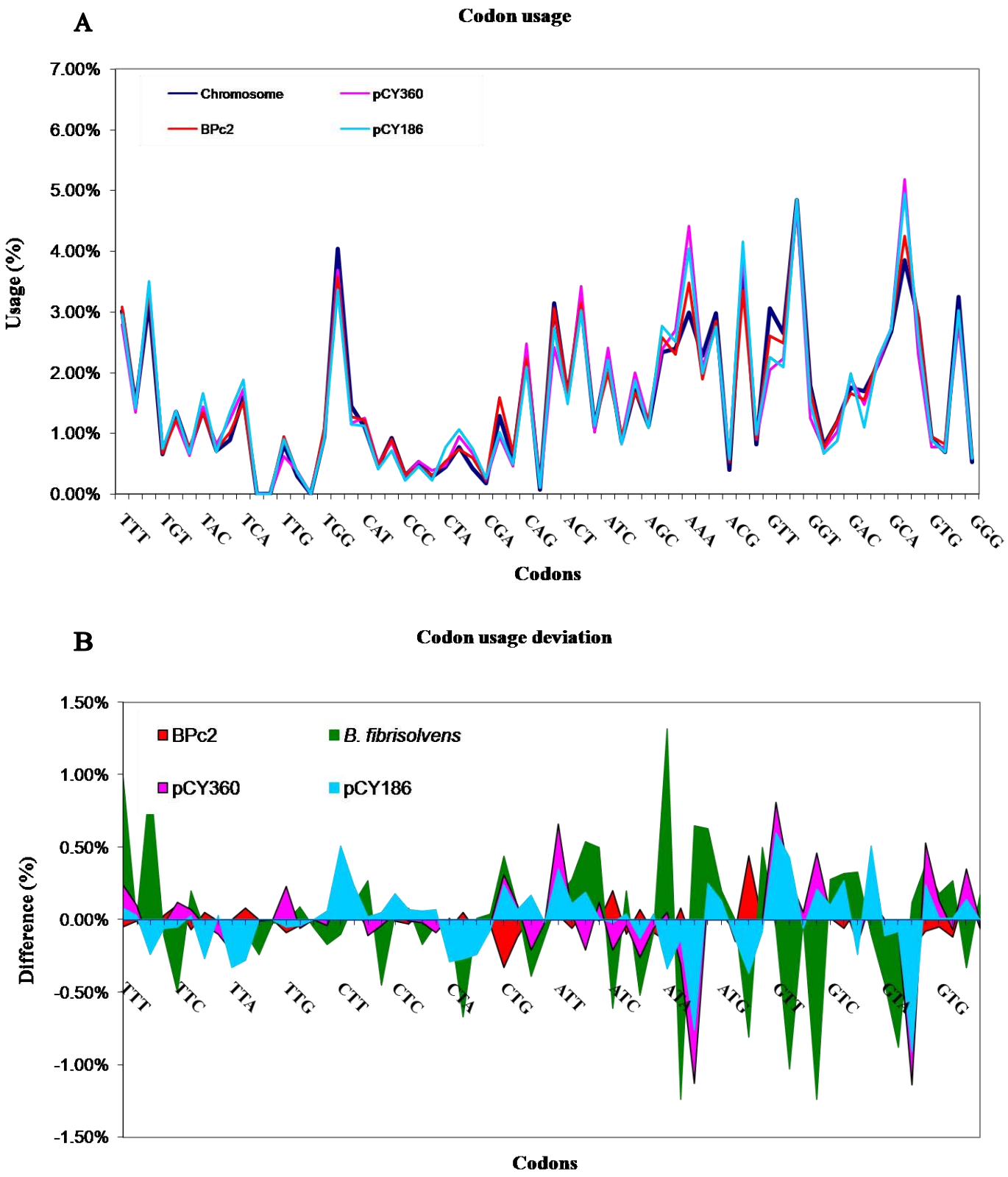


Figure 4.6. Codon usage comparison. (A) Codon usage of pCY360 (pink), BPc2 (red) and pCY186 (light blue) is compared to the 3.5 Mb major chromosome (dark blue) of *B. proteoclasticus* B316^{T}. (B) The deviation in codon usage of pCY360, BPc2, pCY186 and *B. fibrisolvens* (green) to that of the *B. proteoclasticus* 3.5 Mb major chromosome (baseline) is shown.

4.11 Attempts to 'cure' *B. proteoclasticus* of pCY360

Attempts were made to „cure" *B. proteoclasticus* B316^T of the pCY360 replicon using growth at the maximum sub-lethal temperature or growth in media supplemented with maximal sub-lethal levels of novobiocin, acriflavine, ethidium bromide or acridine orange.

4.11.1 Determining maximal sub-lethal levels

Sub-lethal levels of each of the curing agents were determined by growth, in triplicate, in increasing concentrations of each curing agent or increased temperature (Table 4.3). SDS was also initially selected as a curing agent, however, the anionic surfactant failed to inhibit growth up to a final concentration of 5% (w/v), and so was excluded from further analysis. Maximum sub-lethal limits were determined as being growth at 45 ºC or growth in media supplemented with: 5 µg/ml novobiocin; 1 µg/ml acriflavine; 500 ng/ml ethidium bromide; or 10 µg/ml acridine orange.

4.11.2 Determining generation time under curing conditions

Generation times were determined for *B. proteoclasticus* under each curing condition (Fig. 4.7), as described in Section 2.2.11. Generation times were determined to be: 38 min for growth at 45º C; 625 min, novobiocin; 475 min, acriflavine; 305 min, ethidium bromide; and 300 min, acridine orange. The maximum cell density was equivalent to the wild type strain when grown in ethidium bromide (ΔOD_{600} ~1.3), slightly reduced when grown in novobiocin or acriflavine (ΔOD_{600} ~0.8 - 0.9, respectively) and significantly reduced when grown at 45 ºC or in acridine orange (ΔOD_{600}=0.35 and 0.075 respectively).

4.11.3 Evaluation of megaplasmid loss

Following growth for 20 generations under each condition, 5,688 colonies (1,107 treated with novobiocin; 1,016 with acriflavine, 1,198 with ethidium bromide; 1,156 with acridine orange; 1,211 grown at 45º C) were screened by colony hybridisation using the pCY360 specific probe pCY360_probe1, designed to target a 743bp, largely intergenic, region of the megaplasmid. All colonies examined were detected by the Southern probe, suggesting no *B. proteoclasticus* cells screened were cured of the pCY360 megaplasmid.

Table 4.3. Maximum sub-lethal limits of plasmid-curing agents

Acridine Orange																												
Concentration (µg/ml)	0.05			0.1			0.5			1			5			10			12			20			40			100		
Tube	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Growth(+/-)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-

Acriflavine																												
Concentration (µg/ml)	0.5			1			2			3			5			10			20			30			40			50		
Tube	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Growth(+/-)*	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Ethidium Bromide																								
Concentration (µg/ml)	0.05			01			0.5			1			2			5			10			15		
Tube	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Growth (+/-)	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

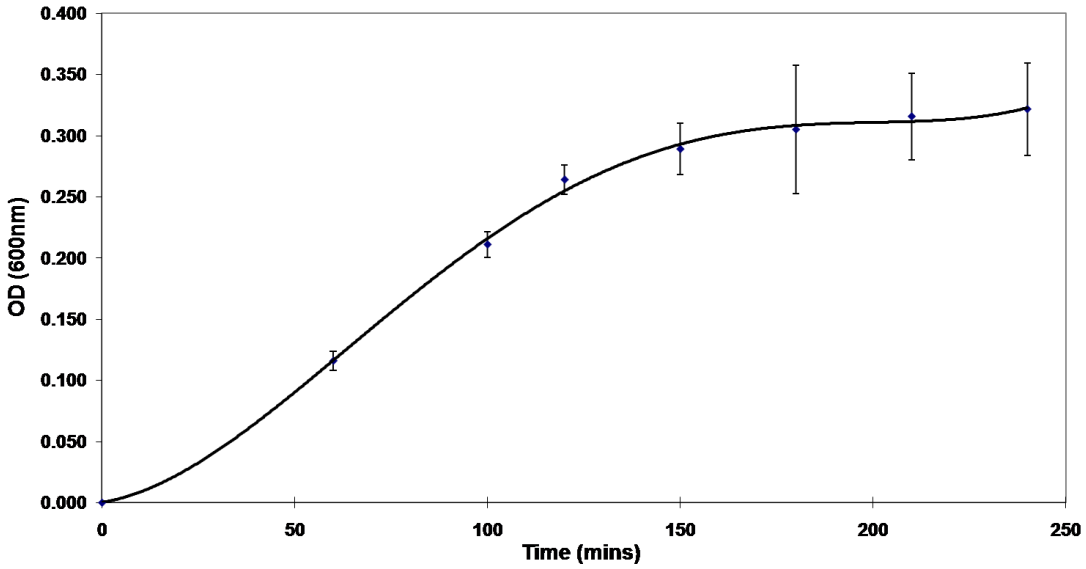
Novobiocin																					
Concentration (µg/ml)	1			5			6			7			10			20			40		
Tube	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Growth (+/-)	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

SDS																												
Concentration (µg/ml)	1			5			10			20			50			100			1000			5000			20,00			50,000		
Tube	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Growth (+/-)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

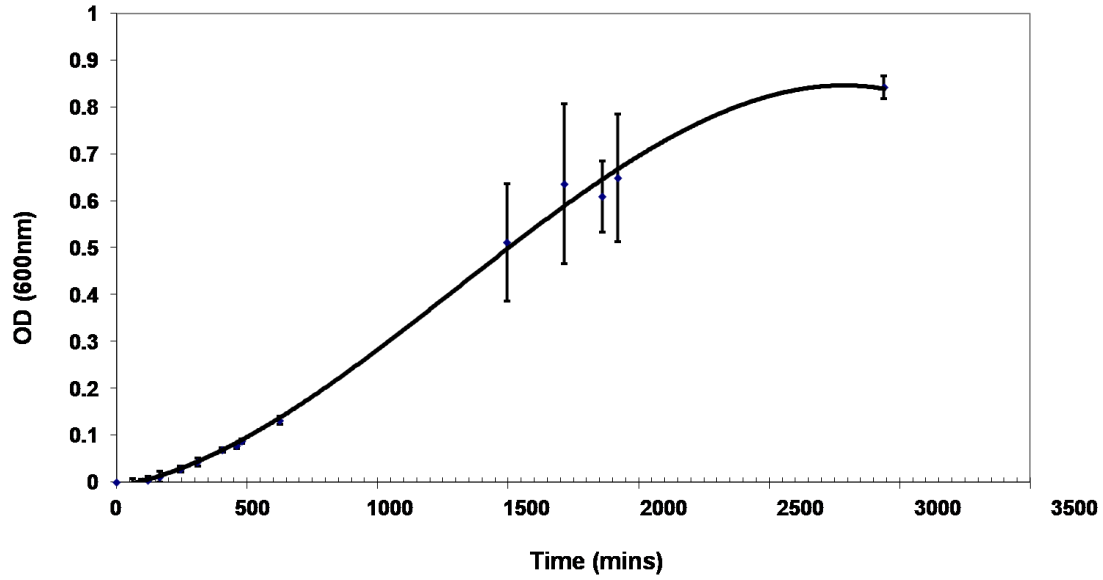
Elevated Temperature																					
Temperature	39°C			41°C			43°C			45°C			46°C			47°C			50°C		
Tube	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Growth (+/-)	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-

* The table shows growth (+) or absence of growth (-) at increasing concentrations of curing agent or growth temperature. Cultures were grown in triplicate and growth was determined by a change in absorbance at 600 nm and subsequently verified by microscopy.

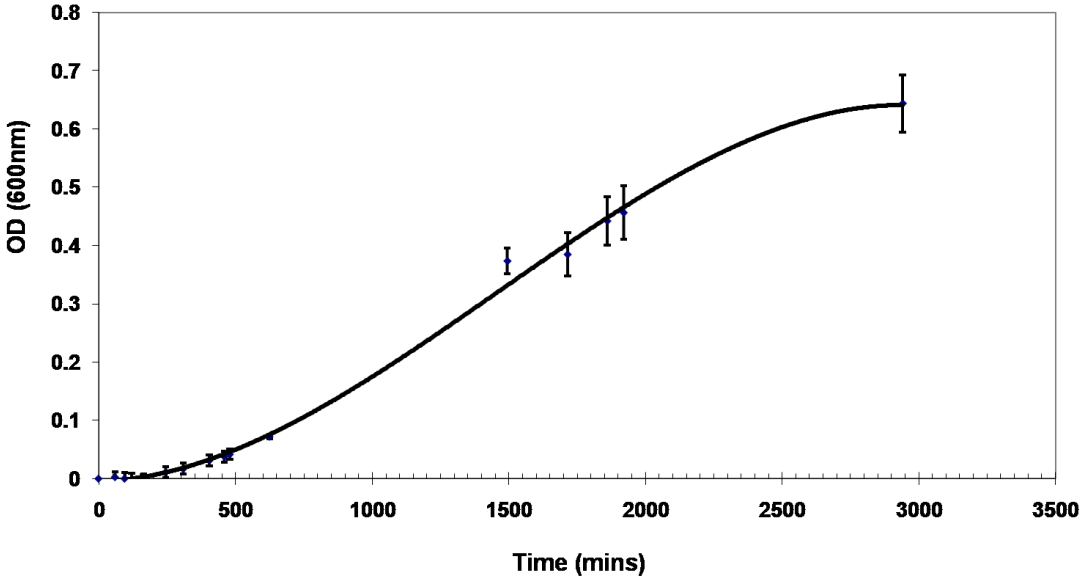
Growth of *B. proteoclasticus* at 45°C



Growth of *B. proteoclasticus* in M704 supplemented with 1ug/ml acriflavine



Growth of *B. proteoclasticus* in M704 supplemented with 5ug/ml novobiocin



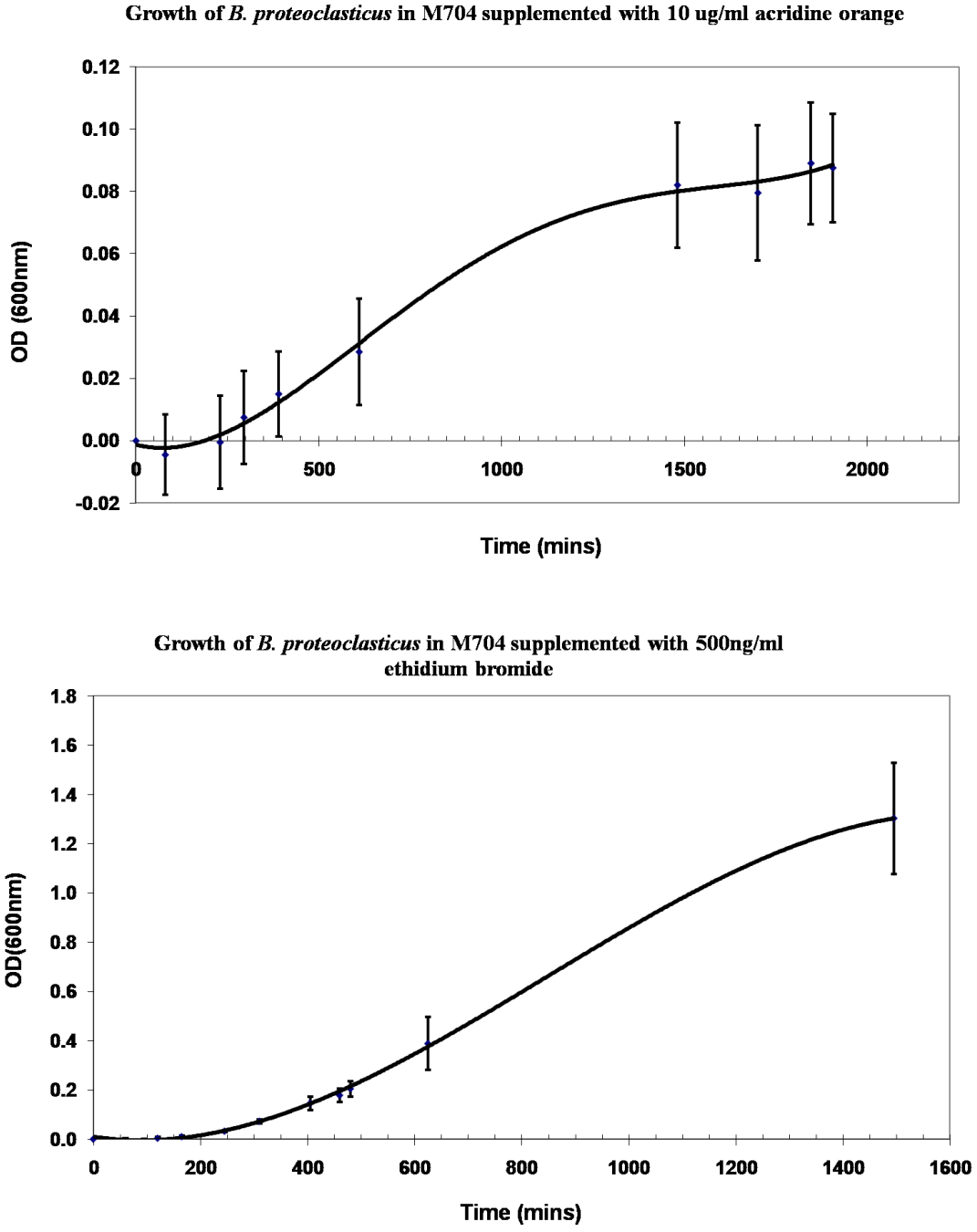


Figure 4.7. Growth curves of *B. proteoclasticus* exposed to maximal sub-lethal limits of curing agents or temperature. Cells were grown as described (Section 2.2.11) and growth was measured by change in optical density at 600 nm. Experiments were carried out in triplicate and Y-error bars indicate margin of error for 99% confidence.

4.12 Copy number of pCY360

The copy number of pCY360 was determined using a quantitative real time PCR (qPCR) method described by Lee *et al*. (2006) as described in Section 2.2.12.

4.12.1 qPCR optimisation

As the amount of error in the final quantification is inversely proportional to the amplification efficiency of the PCR reaction (Tichopad *et al*., 2002), it was important to first optimise each reaction. The LightCycler machine is not capable of performing temperature gradient PCR, therefore the T_m of each primer set was optimised by conventional PCR using a PX2 Thermal Cycler with an annealing temperature gradient ranging from 55 ºC – 65 ºC. The run conditions and reaction components were set to replicate a qPCR reaction in the LightCycler, as described in Section 2.2.12.4. Following the PCR reactions, the products were analysed by agarose gel electrophoresis. The resulting gel was pictured and product intensity was determined using the KODAK 1D Image analysis software (Pizzonia, 2001). From this analysis the optimal temperature for both specificity and optimal amplification efficiency was determined as being 60 ºC (Fig. 4.8a).

The reactions were then optimised for $MgCl_2$ concentration. Replicate reactions were set up as above with between 1 to 6 mM $MgCl_2$. The PCR run conditions and reaction components were again set to replicate a qPCR reaction in the LightCycler, described in Section 2.2.12.4 with a single annealing temperature of 60 ºC. Following the PCR reactions, the products were analysed by gel electrophoresis. The resulting gel was pictured and product intensity was determined as described above. From this analysis the optimal $MgCl_2$ concentration using an annealing temperature of 60 ºC, ensuring both specificity and optimal amplification efficiency, was determined for both reactions as being 3 mM (Fig. 4.8b).

4.12.2 Ensuring reaction integrity

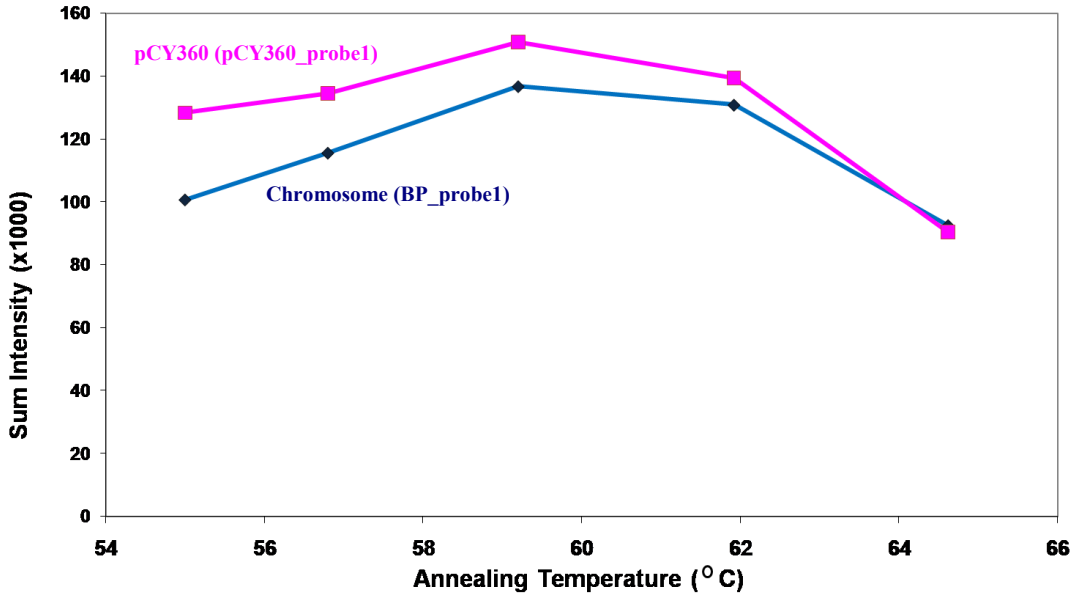
To ensure the specificity of the reactions following completion of the qPCR cycle, melting curve analysis and agarose gel electrophoresis of the PCR products were performed. Melting curve analysis revealed a single peak for both the BP_probe1 and pCY360_probe1 reactions occurring at 85.08 ºC and 84.82 ºC respectively (Fig 4.10a). Subsequent analysis by gel electrophoresis revealed it to be consistent with

the expected size of the BP_probe1 (703bp) and pCY360_probe1 (743bp) products (Figure 4.10b). Standard curves (Fig 4.9a) of the threshold cycles vs. the $\log_{10}$ of the replicon copy numbers for each standard were determined as being -3.420 and -3.362 for Topo-BPp1 and Topo-pCY360p1 standards respectively. This equates to amplification efficiencies for the reactions of 0.96 and 0.98 respectively. The standard curves were both linear in the range tested ($1 \times 10^5 - 1 \times 10^9$ copies / µl; $R^2 > 0.99$). Standard curves of the threshold cycles vs. the $\log_{10}$ of the fold dilution for each target (Fig. 4.9b) were also determined to ensure they were approximately equal. Slopes were determined to be 3.383 and 3.432 for Topo-BPp1 and Topo-pCY360p1 standards respectively. This equates to an amplification efficiency of 0.98 and 0.96 respectively. The standard curves were both linear in the range tested ($1 \times 10^{-1} - 1 \times 10^{-5}$ DNA extract / reaction; $R^2 > 0.99$).

4.12.3 Absolute copy numbers

Analysis revealed the replicon to be present in $5.23 \pm 0.982 \times 10^{10}$ copies in the tested sample (99% confidence) per $1.3 \pm 0.175 \times 10^{10}$ copies of the chromosome (99% confidence). This equates to a copy number of 4.02 ± 0.93 copies of pCY360 per chromosome (99% confidence). This is consistent with the cloning bias toward pCY360 observed during sequencing (Fig. 3.6).

A



B

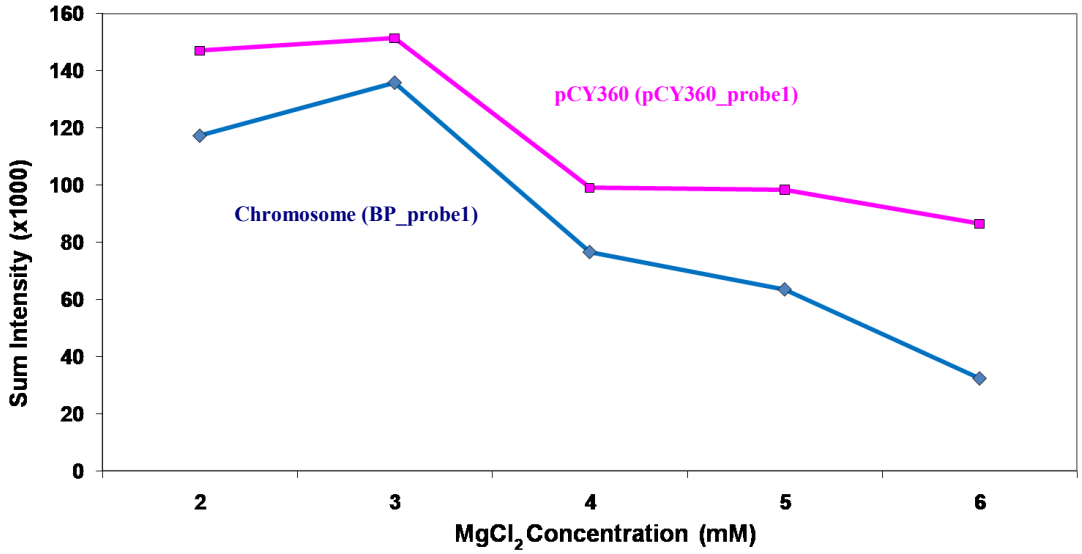


Figure 4.8. Optimisation of qPCR reactions. qPCR reactions were optimised for temperature (A) and $MgCl_2$ concentration (B) for both BP-probe1 (blue) and pCY360-probe1 (pink). Optimal conditions were determined as those producing the greatest amount of the desired qPCR product. This was determined by measuring the sum of the pixel intensity from the region of interest (ROI) using KODAK 1D Image analysis software.

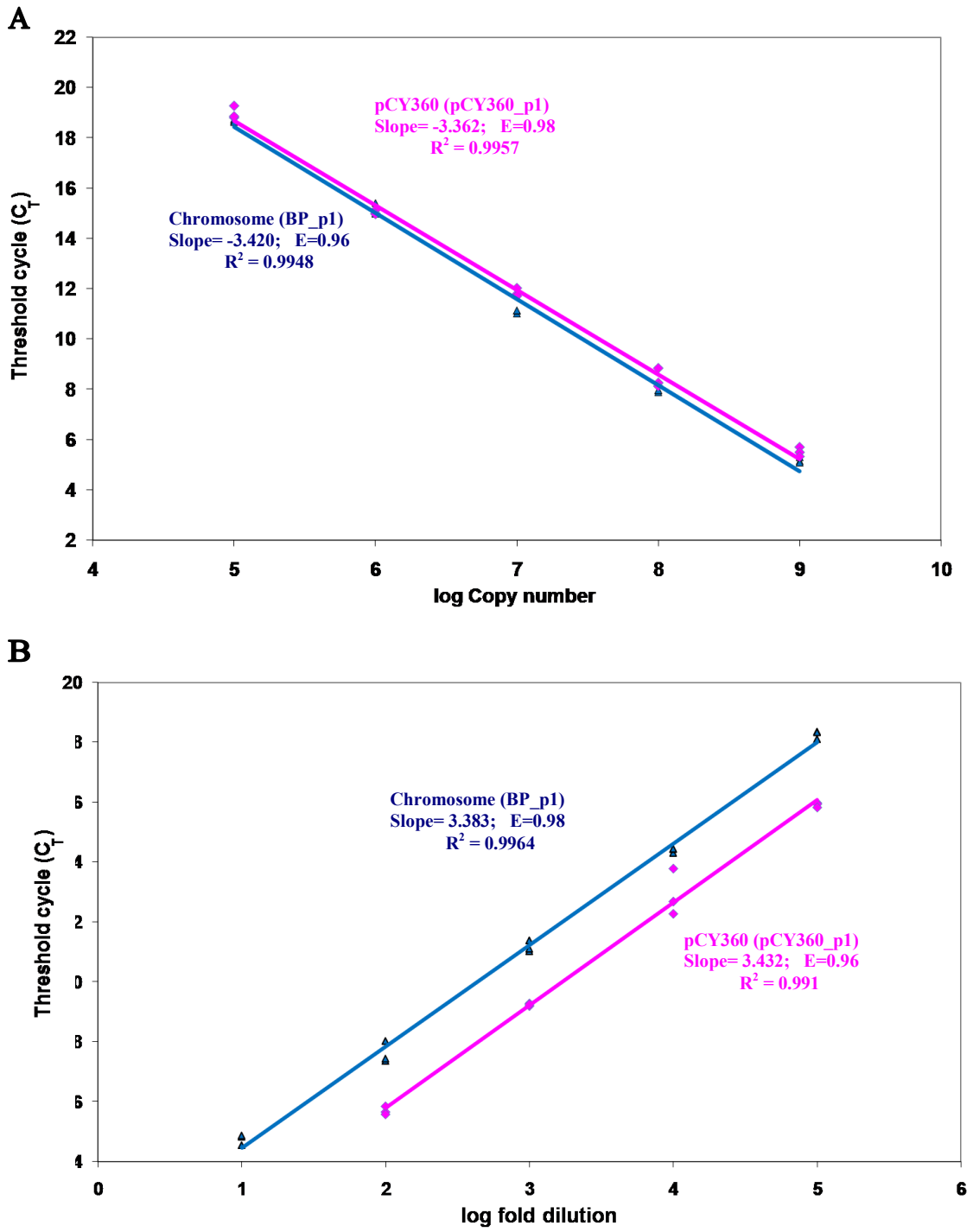


Figure 4.9. Standard curves for qPCR reactions. The slope of each threshold cycle vs. $\log_{10}$ of the copy number in each standard (A) or threshold cycle vs. $\log_{10}$ fold dilution (B) is shown. BP_p1 is shown in blue and pCY360_p1 is shown in pink. The slope, efficiency (E), and correlation coefficient (R^2) are shown.

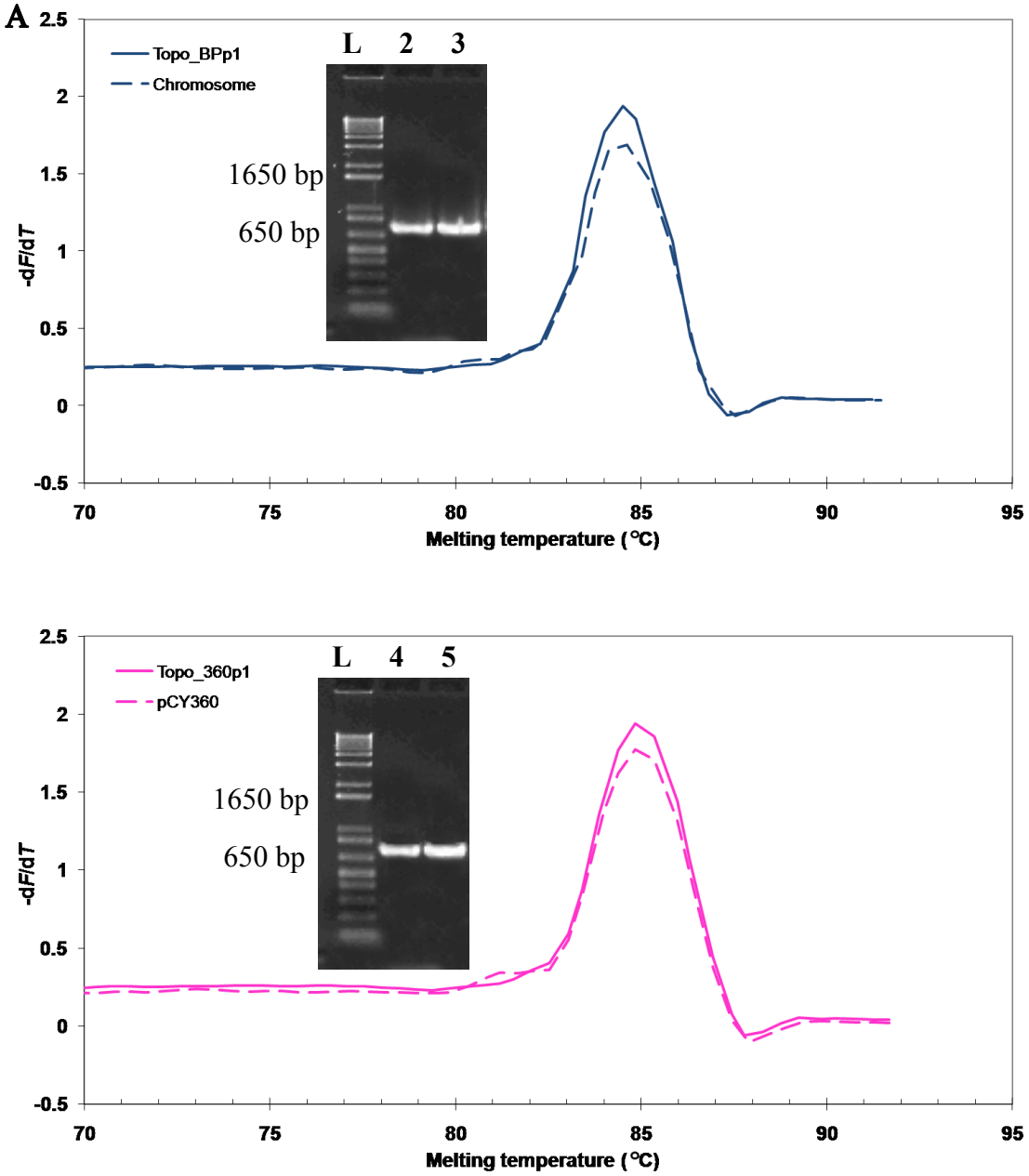


Figure 4.10. Confirmation of qPCR amplification specificities of the chromosome primer set (A) and pCY360 primer set (B). Melting peaks were examined for each set of primers using both the *B. proteoclasticus* total DNA preparation (broken line) and the quantified Topo-standards (solid line). Inset: Gel-electrophoresis of the qPCR products was performed. Lanes: L, the 1Kb+ ladder, lane 2-5 PCR-products (703 bp, Chromosome_probe1 and 743 bp pCY360_probe1) obtained using as templates either quantified Topo-recombinant plasmids carrying the target sequence (Chromosome probe 1, lane 2 and pCY360 probe 1, lane 4) or the total DNA preparation (lanes 3 and 5).

4.13 Microarray analysis of pCY360 gene expression in co-culture with the rumen methanogen, *Methanobrevibacter ruminantium*.

As pCY360 encodes a significant number of proteins predicted to span the membrane or that are secreted (35, 9%), along with a number of proteins which potentially affect the organism"s membrane topology, it was reasoned that the replicon may enhance the ability of *B. proteoclasticus* to interact with its environment. Due to the microbiologically-rich nature of the rumen it is likely *B. proteoclasticus* interacts with a diverse range of microorganisms. Therefore *B. proteoclasticus* was grown in co-culture with the rumen methanogen, *Methanobrevibacter ruminantium* strain M1 (DSM 1093). *B. proteoclasticus* was grown on xylan as a sole carbon source while *M. ruminantium* relied on H_2 produced as an end product of *B. proteoclasticus* growth. As early as 120 min after co-inoculation, microscopic examination of the co-culture showed co-aggregation of *B. proteoclasticus* and *M. ruminantium* cells (Fig. 4.11). This interaction does not appear to influence the growth rate of *B. proteoclasticus*, as its generation time within the co-culture was not significantly different to when it was grown in mono-culture. To examine the contribution of *B. proteoclasticus* to this interspecies interaction, total RNA was extracted at late exponential phase from *B. proteoclasticus* and *M. ruminantium* each grown in mono-culture, and from a co-culture of the two organisms, (described in Section 2.2.13). A large, dense, white pellet formed in each extract, which was attributed to co-purification of polysaccharide, (most likely xylan) from the growth medium. Contaminating polysaccharide was removed using an RNeasy Midi kit (Qiagen, Hilden, Germany; described in Section 2.2.13.4). Spectrophotometric analysis revealed the resulting RNA to be pure (Appendix III, Table A3).

To analyse the integrity of the extracted RNA, a Bioanalyzer RNA 6000 nano-assay was conducted (Agilent Technologies, Santa Clara, CA, USA). Aliquots of each sample were combined and the quality of each total RNA sample was analysed using an RNA 6000 Nano LabChip® kit with an Agilent 2100 Bioanalyzer. The resulting data is displayed as both an electropherogram and an electrophoretic gel-image translation of that information (Fig. 4.12).

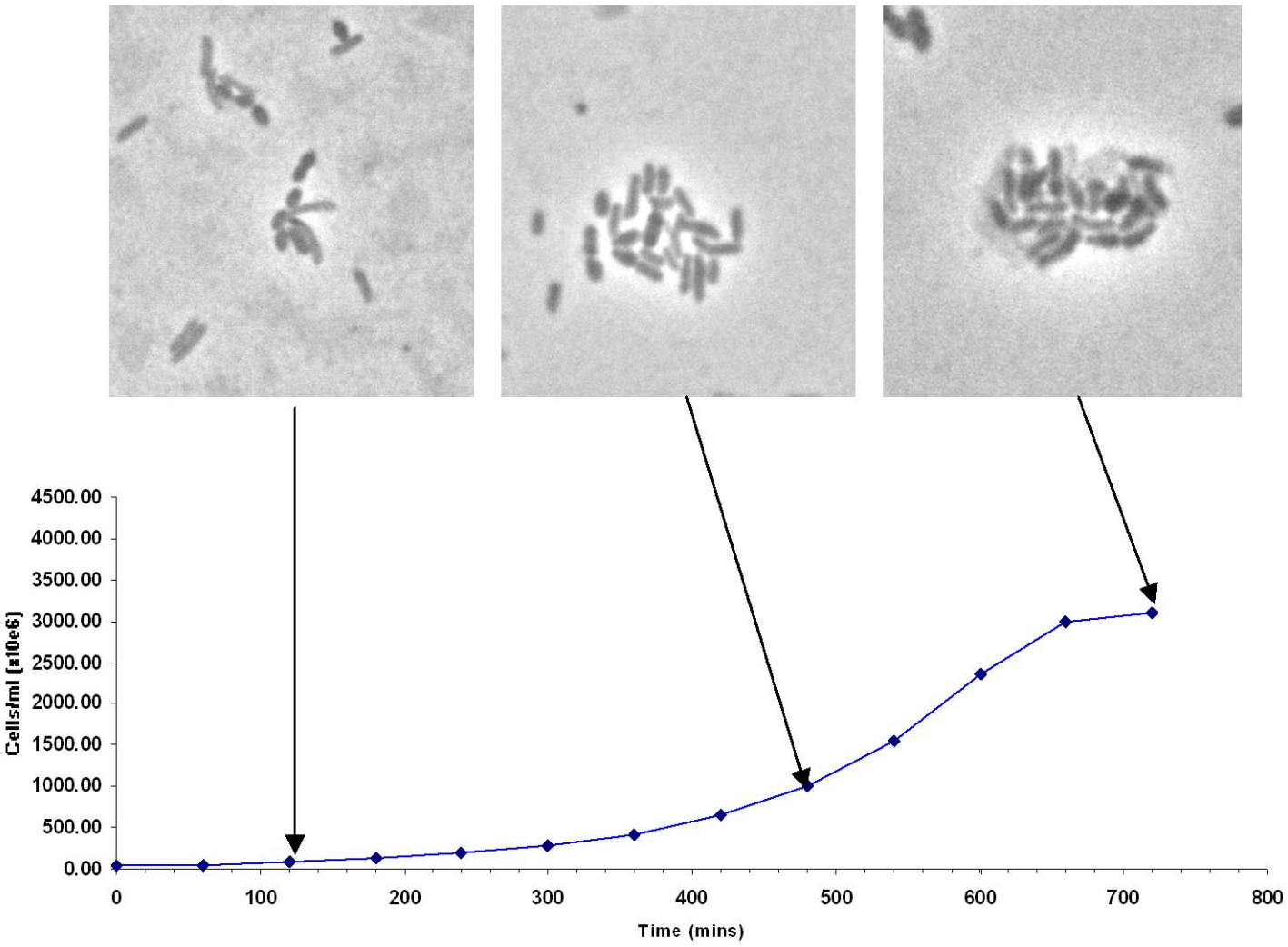


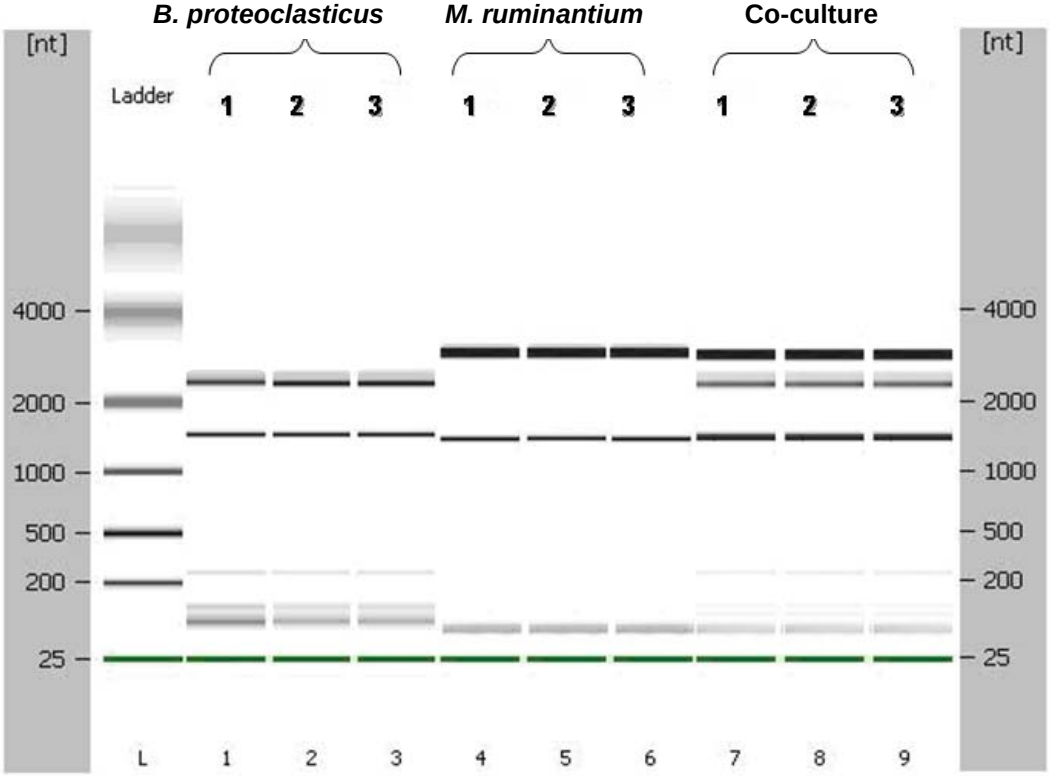
Figure 4.11. **Observation of interspecies interactions between *B. proteoclasticus* and *Methanobrevibacter ruminantium***. Microscope images taken at 2, 8 and 12 h post-inoculation of *B. proteoclasticus* (lighter rod-shaped organism) into BY+ (+ 0.2% xylan) media containing a mid-exponential *M. ruminantium* culture (darker, short ovoid rod-shaped organism). Growth, as determined by Thoma slide enumeration, is shown along with sampling time.

The electrophoretic gel-image translations show ribosomal RNA (rRNA) bands including the 16*S* (~1500 nt) and 23*S* rRNAs (~2600 nt for *B. proteoclasticus* and ~3000nt for *M. ruminantium*) present in all total RNA samples. Analysis of the electropherograms reveals the 23*S* and 16*S* rRNAs to be present in both mono-culture samples at ratio of 1 or greater indicating little or no RNA degradation has occurred (Ambion TechNotes 11(2) 2004). A 23*S*/16*S* ratio could not be determined for the co-culture RNA extracts due to the similar mobility of the 16*S* of each organism. Analysis of the electropherograms shows the baseline between 16*S* and 23*S* rRNA peaks, as well as immediately below the 16*S* rRNA peak to be relatively flat in all samples consistent with little or no degradation of the 23*S* rRNA (Ambion TechNotes 11(2) 2004).

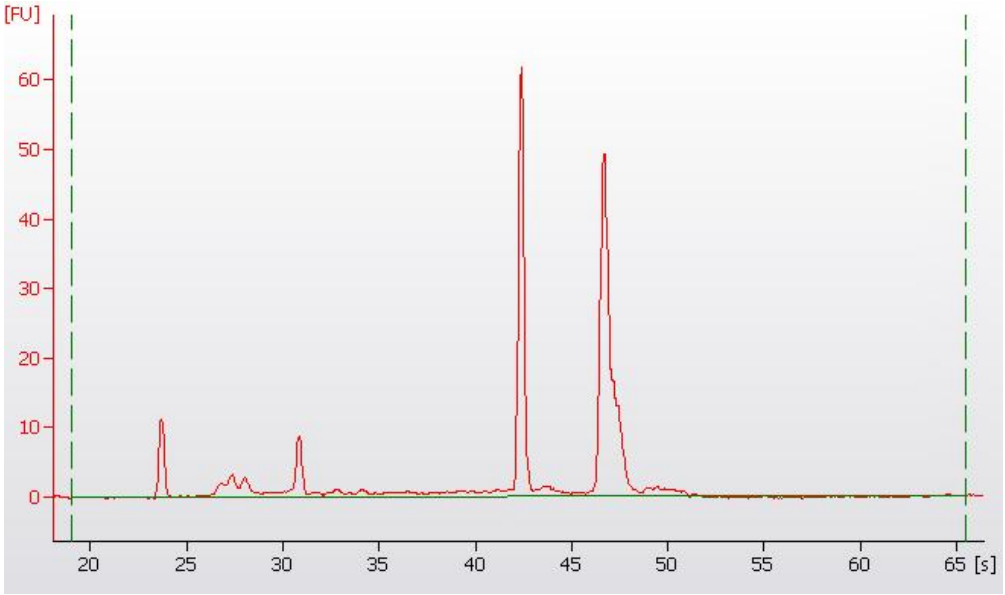
4.13.1 Microarray hybridisation and scanning

RNA was reverse-transcribed using Superscript^{TM} III Reverse Transcriptase and purified (described in Section 2.2.11.7 and 2.2.11.8 respectively). The proportion of cDNA derived from *B. proteoclasticus* and *M. ruminantium* in the co-culture was estimated by qPCR, (described in Section 2.2.13.10) using primers specific for the gene encoding butyryl-CoA dehydrogenase (*bcd*; primers bcdqfp and bcdqrp) from *B. proteoclasticus* and the gene encoding N^5,N^{10}-methenyl-H_4MPT cyclohydrolase, (*mch*; primers mchqfp and mchqrp) from *M. ruminantium*. Homologues of both genes have previously been shown to be constitutively expressed in closely related species (Reeve *et al*., 1997, Asanuma *et al*., 2005). Both qPCR reactions gave strong amplification efficiencies (0.97 and 0.98 respectively) and melting curve and gel-electrophoretic analysis revealed the reactions to specifically produce a PCR product of the expected size (266 bp for *bcd* and 419 bp for *mch*, Fig. 4.13).

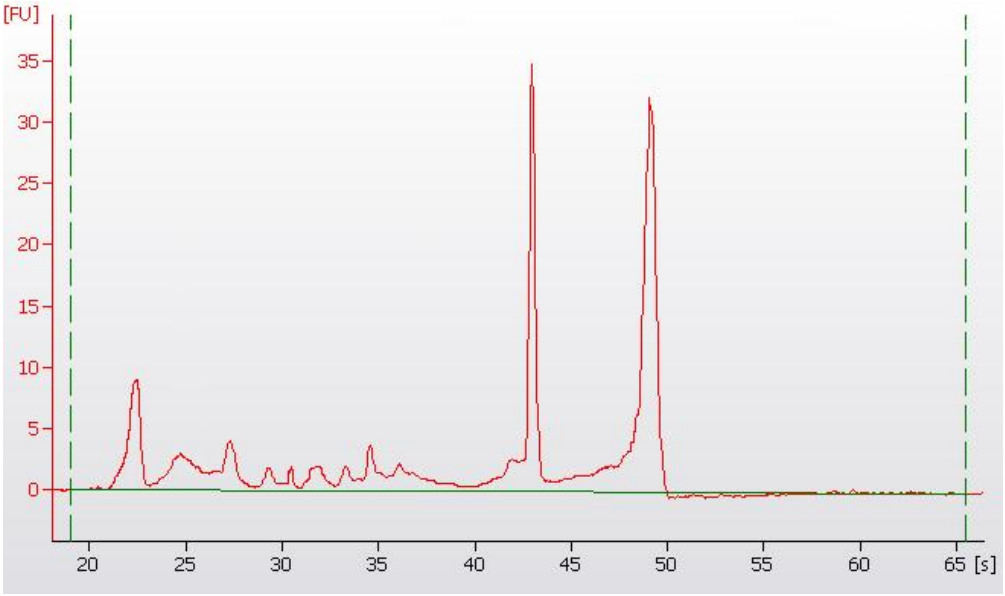
A


B. proteoclasticus
M. ruminantium
Co-culture
[nt]
Ladder
1 2 3 1 2 3 1 2 3
4000
2000
1000
500
200
25
L 1 2 3 4 5 6 7 8 9

B


B. proteoclasticus
[FU]
60
50
40
30
20
10
0
20 25 30 35 40 45 50 55 60 65 [s]

M. ruminantium



Co-culture

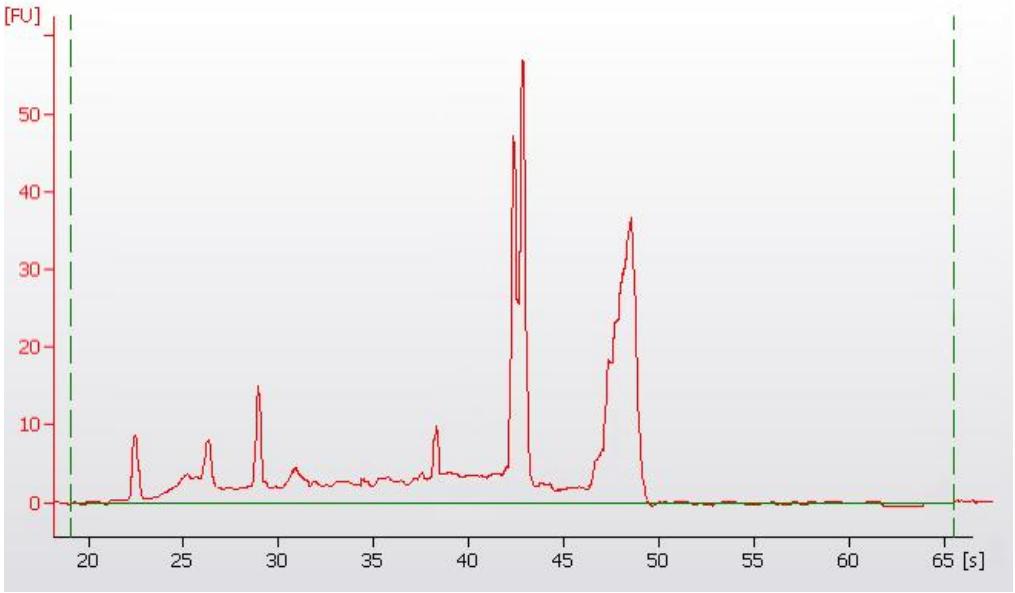


Figure 4.12. Electropherograms and electrophoretic gel-translation of total RNA extracts from *B. proteoclasticus*, *M. ruminantium* and the co-culture of both organisms. The integrity of RNA extracts were examined by a Bioanalyzer RNA 6000 nano-assay and the resulting images are shown as electropherograms (b) and the translated gel image (a). The two dominant peaks occurring at ~42.5 and 47.5 s correspond to the 16*S* and 23*S* rRNAs respectively.

No amplification was detected from the blank or the RNA controls. qPCR analysis indicated that cDNA derived from *B. proteoclasticus* comprised approximately 45% of the total RNA extracted from each co-culture (Table 4.4). Mono-cultures were combined at the qPCR-determined ratio and each cDNA sample was labeled, as described (2.2.13.10), with both Cy3 and Cy5 dyes. The purpose of labelling divided cDNA samples with different dyes was to enable dye swaps to be performed, thereby eliminating dye bias.

Labeled cDNAs were hybridised, (described in Section 2.2.13.12) to a specially constructed microarray containing oligonucleotide probes against the almost complete genomes of both *B. proteoclasticus* and *M. ruminantium*. Each microarray was scanned at a high, medium and low PMT gain value, (described in Section 2.2.13.13), with two exceptions; replicates 2a and 3b were not scanned at a low level due to excessive photo-bleaching of the Cy5 fluorophore prior to the successful completion of this scan. PMT gain values and red to green ratios are shown with microarray scan images in Appendix III (Table. A4 and Fig. A4).

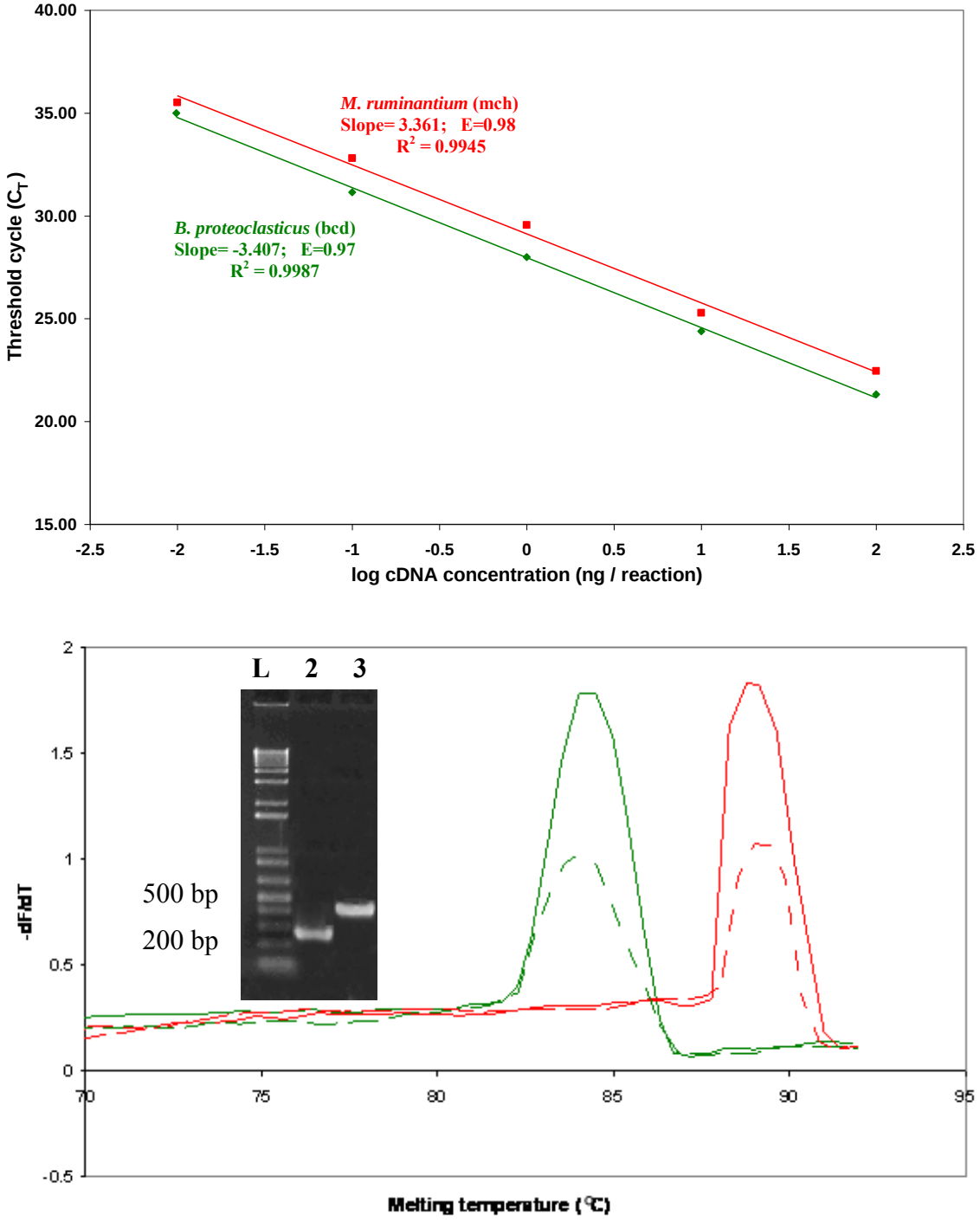


Figure 4.13 RT-qPCR amplification efficiencies (A) and specificities (B). (A) The slope of each threshold cycle vs. $\log_{10}$ of the cDNA concentration of the monocultures (solid line) of *B. proteoclasticus* (green) and *M. ruminantium* (red). The slope, efficiency (E), and correlation coefficient (R^2) are shown. (B) Melting peaks were examined for each PCR product using both monocultures (solid line) and co-cultures (broken line). Inset: Gel-electrophoresis of the RT-qPCR products obtained using primer sets *bcd* to enumerate the *B. proteoclasticus* cDNA contribution (lane 2), and *mch* to enumerate *M. ruminantium* cDNA contribution (lane 3). Band sizes (266 bp, *bcd* and 412 bp *mch*) were determined using the 1Kb+ ladder (L)

Table 4.4. Percentage contribution of cDNAs from either *B. proteoclasticus* or *M. ruminantium* to co-culture RNA extracts.

Co-culture sample	% cDNA contribution*	
	B. proteoclasticus	***M. ruminantium***
Replicate 1	45.50 +/- 3.22%	55.48 +/- 3.04%
Replicate 2	44.83 +/- 4.15%	54.77 +/- 6.40%
Replicate 3	47.49 +/- 4.11%	51.67 +/- 8.02%

*Average percentage of the five 10-fold dilutions +/- 99 percentile confidence interval. Results derived from reverse transcription qPCR analysis of cDNA contribution from each organism to each co-culture extract.

4.13.2 Microarray analysis

The microarray data were subsequently analysed to test its integrity and determine what normalisation procedures were required, (described in Section 2.2.13.14). Visual inspection of the scanned slide image identified two small regions (Fig. 4.14) where the background had clearly corrupted several oligonucleotide spots. These spots were „bad-flagged" and removed from further analysis. Box plots of the $\log_2$ intensity distribution for each scan (Fig. 4.15) show the background intensity to be approximately 1 fold lower than the foreground for the red scan in all but the low scans of 1a, 1b and 2b. The background intensities for all green scans were consistently 1-3 fold lower than the foreground.

.

The high scan level was selected to provide the base for subsequent analysis as few features fell outside the useable range (described in Section 2.2.13.14) for analysis compared to the other scan levels. The average intensity span was 600 – Saturation (65535). Plots of the foreground and background signal intensities (Fig. 4.16) show little or no spatial bias in either the Cy5 (red) or Cy3 (green) foreground. However, greater spatial bias was evident in the background of both Cy5 and Cy3 plots. As this spatial variation did not correspond to that observed with the foreground, it was decided that background normalisation would unfairly bias the results and so was not performed. Four small regions that lacked any signal were observed in the foreground plots of all slides from both Cy5 and Cy3 scans. These blank regions are most likely

attributable to improper slide printing and were removed from further analysis. MA plots (described in Section 2.2.13.14; Fig. 4.17) suggested that in all slide scans, there was no bias to detection of either fluorophore through the increasing signal intensities. The „print-tip loess" method was selected for within-slide normalisation. This method removes any spatial trends within the slide by fitting a loess smoother to each of the 32 print-tip blocks on each slide and was selected to remove the bias introduced from the different pins of the spotting robot. To determine the efficacy of normalisation, density plots were produced for the data prior to and following normalisation. Density plots show a bias toward the Cy5 dye in the raw data (Fig. 4.18) which is largely mitigated following normalisation (Fig. 4.19). Box plots of the intensity distribution of each feature by category (Fig. 4.20) show blanks to have a very narrow range of intensities and in most instances show little bias to either dye, particularly for the high intensity scans. Negative controls tended to have a wider range of intensities particularly in the dye-swap scans of replicates 2 (2b) and 3 (3b). Constitutively expressed genes *frhB* and *dnaK* show little bias to either fluorophore suggesting the proportion of cDNA from each organism is approximately equivalent in the co-culture and mono-culture conditions. Conversely the intensity distribution of *B. proteoclasticus* and *M. ruminantium* transcripts has a greater span suggesting genes are differentially regulated between the two conditions.

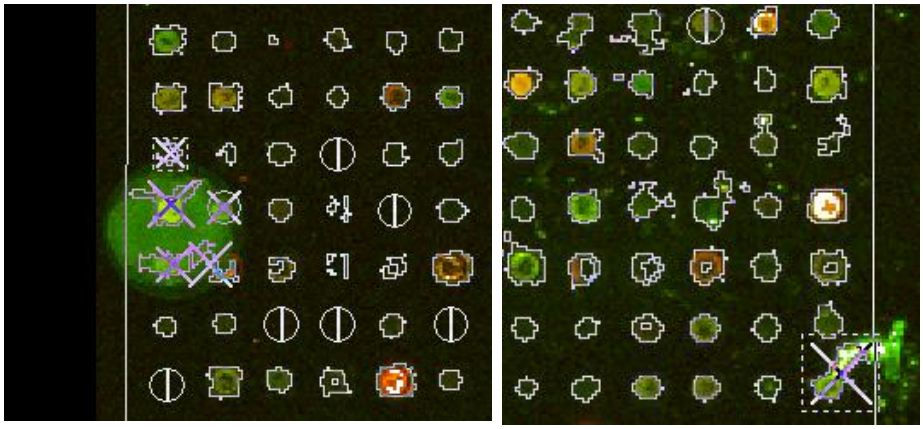


Figure 4.14. **Low quality regions of microarray slides 'bad-flagged' prior to analysis.** Two regions identified by visual inspection were flagged, using Gene-Pix 6.0 software, as being of poor quality and were excluded from further analysis.

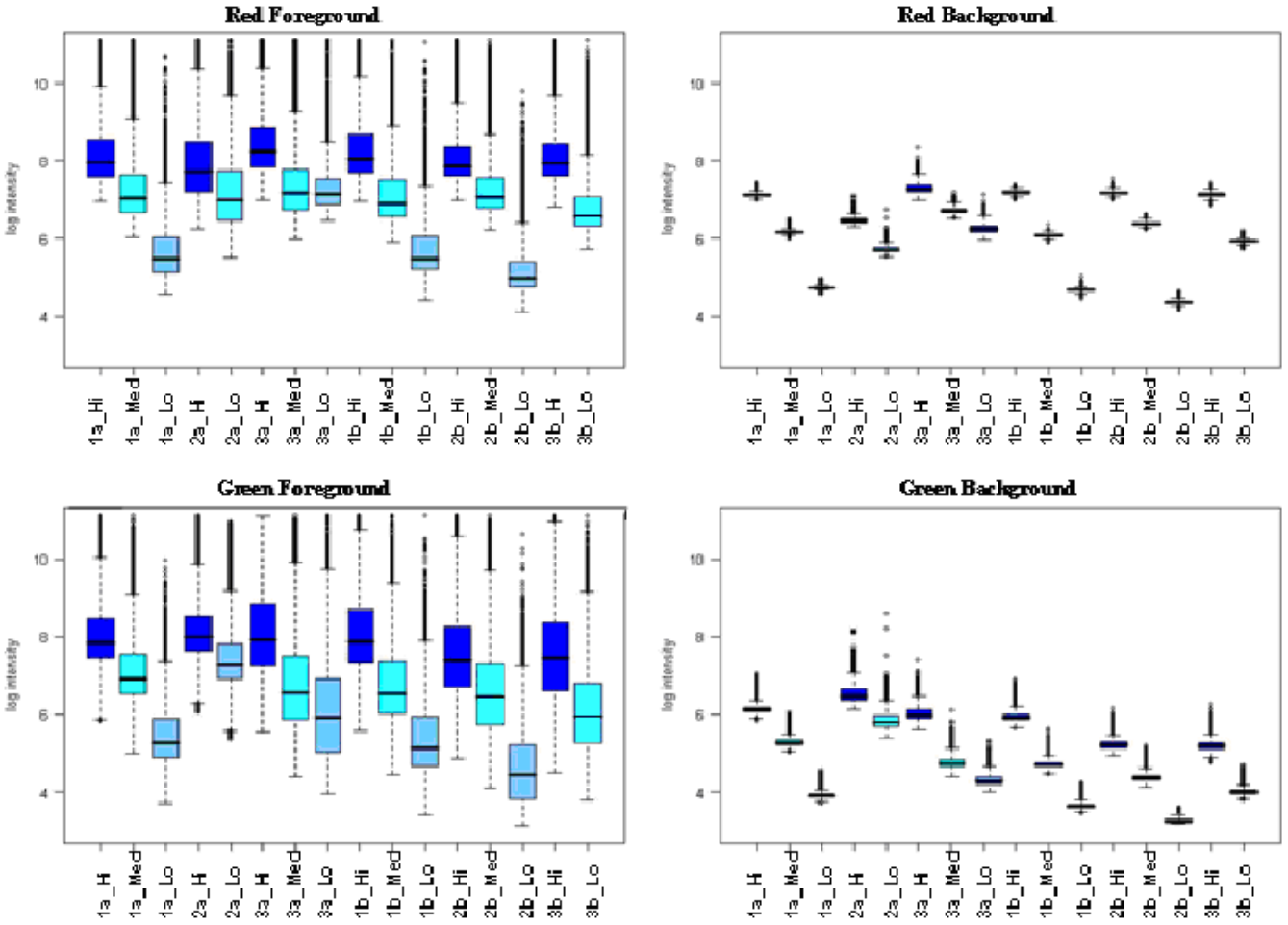
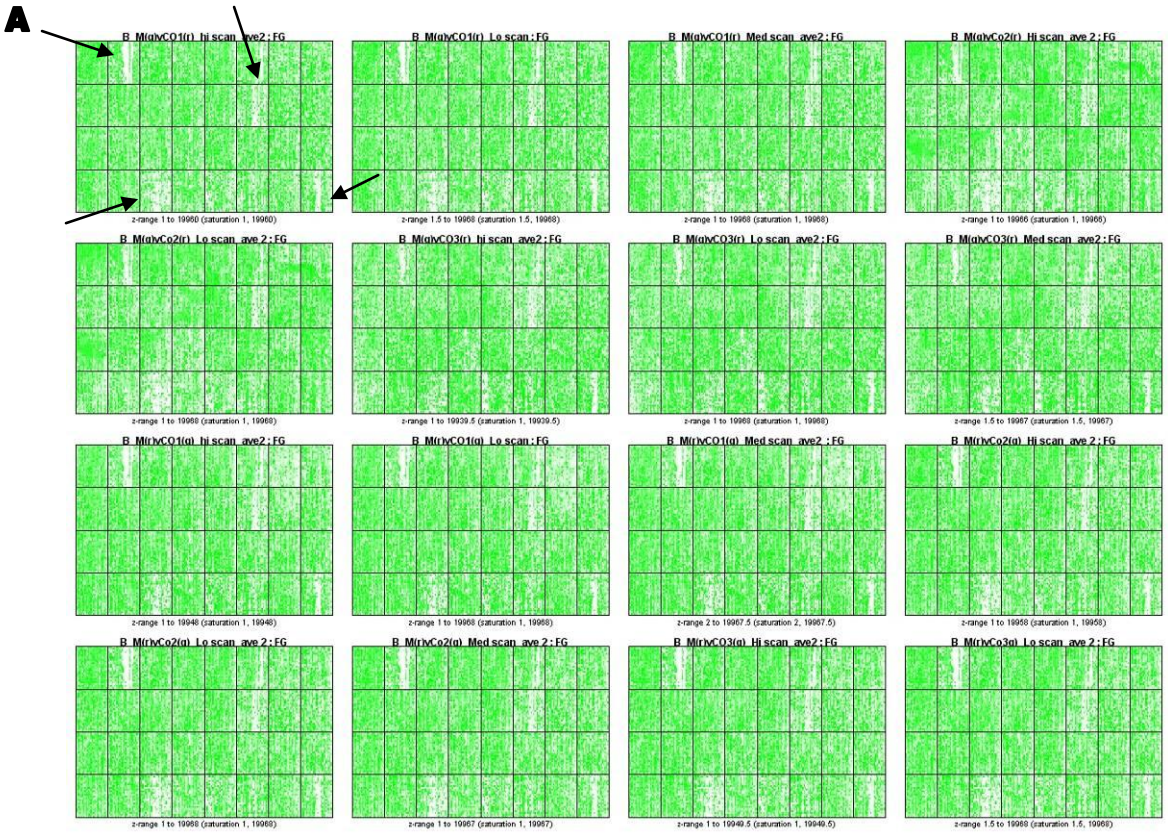
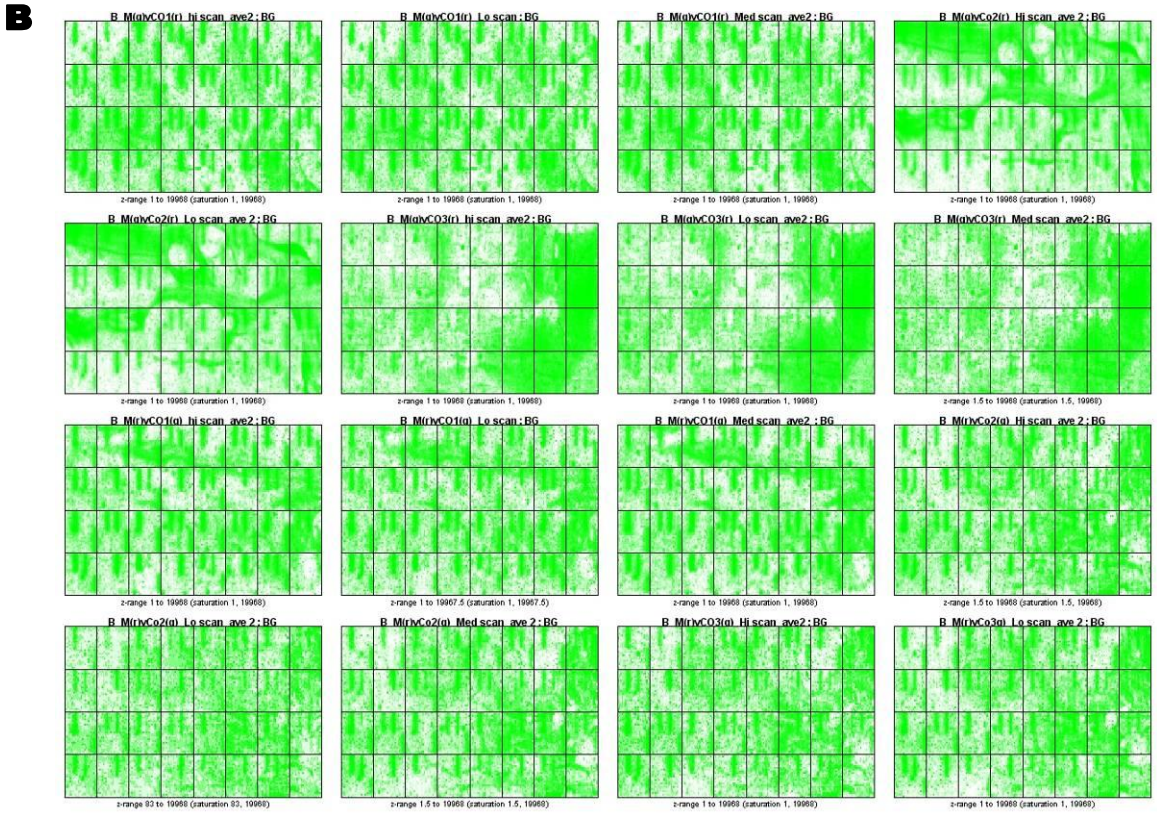


Figure 4.15. **Distribution of background and foreground intensities for each feature**. Box plots show the distribution of signal intensities detected in the foreground and background of each feature on each slide scan. High intensity scans are shown in dark blue, medium intensity scans are shown in turquoise and low intensity scans are in light blue. This figure was created in Bioconductor, using Limma and generated using the graphics module of R.


A
B_M(g)yCO1(r)_hi scan_ave2 : FG
B_M(g)yCO1(r)_Lo scan : FG
B_M(g)yCO1(r)_Med scan_ave2 : FG
B_M(g)yCo2(r)_Hi scan_ave 2 : FG
B_M(g)yCo2(r)_Lo scan_ave 2 : FG
B_M(g)yCO3(r)_hi scan_ave2 : FG
B_M(g)yCO3(r)_Lo scan_ave2 : FG
B_M(g)yCO3(r)_Med scan_ave2 : FG
B_M(r)yCO1(g)_hi scan_ave2 : FG
B_M(r)yCO1(g)_Lo scan : FG
B_M(r)yCO1(g)_Med scan_ave2 : FG
B_M(r)yCo2(g)_Hi scan_ave 2 : FG
B_M(r)yCo2(g)_Lo scan_ave 2 : FG
B_M(r)yCo2(g)_Med scan_ave 2 : FG
B_M(r)yCO3(g)_Hi scan_ave2 : FG
B_M(r)yCo3(g)_Lo scan_ave 2 : FG


B
B_M(g)yCO1(r)_hi scan_ave2 : BG
B_M(g)yCO1(r)_Lo scan : BG
B_M(g)yCO1(r)_Med scan_ave2 : BG
B_M(g)yCo2(r)_Hi scan_ave 2 : BG
B_M(g)yCo2(r)_Lo scan_ave 2 : BG
B_M(g)yCO3(r)_hi scan_ave2 : BG
B_M(g)yCO3(r)_Lo scan_ave2 : BG
B_M(g)yCO3(r)_Med scan_ave2 : BG
B_M(r)yCO1(g)_hi scan_ave2 : BG
B_M(r)yCO1(g)_Lo scan : BG
B_M(r)yCO1(g)_Med scan_ave2 : BG
B_M(r)yCo2(g)_Hi scan_ave 2 : BG
B_M(r)yCo2(g)_Lo scan_ave 2 : BG
B_M(r)yCo2(g)_Med scan_ave 2 : BG
B_M(r)yCO3(g)_Hi scan_ave2 : BG
B_M(r)yCo3(g)_Lo scan_ave 2 : BG

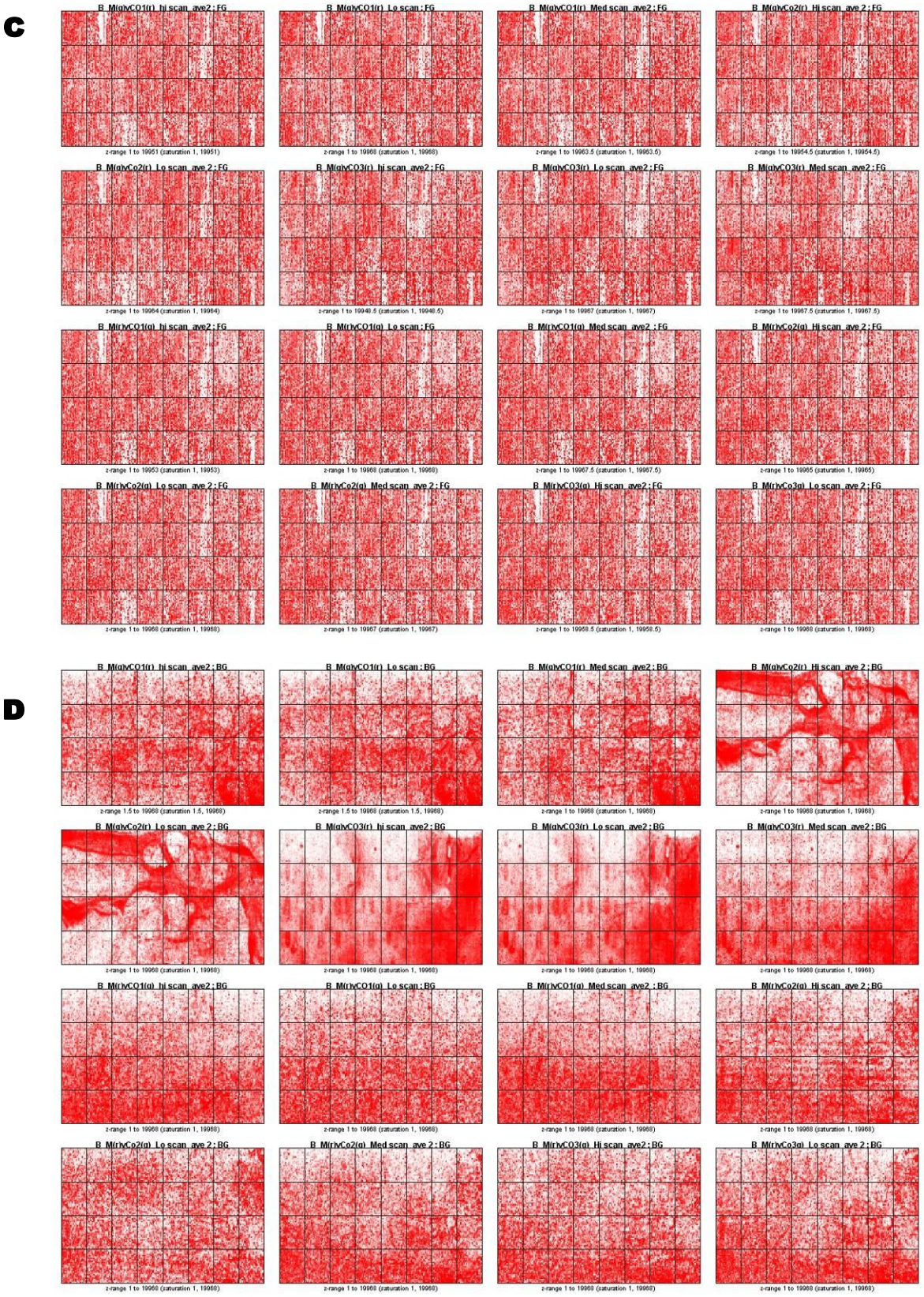
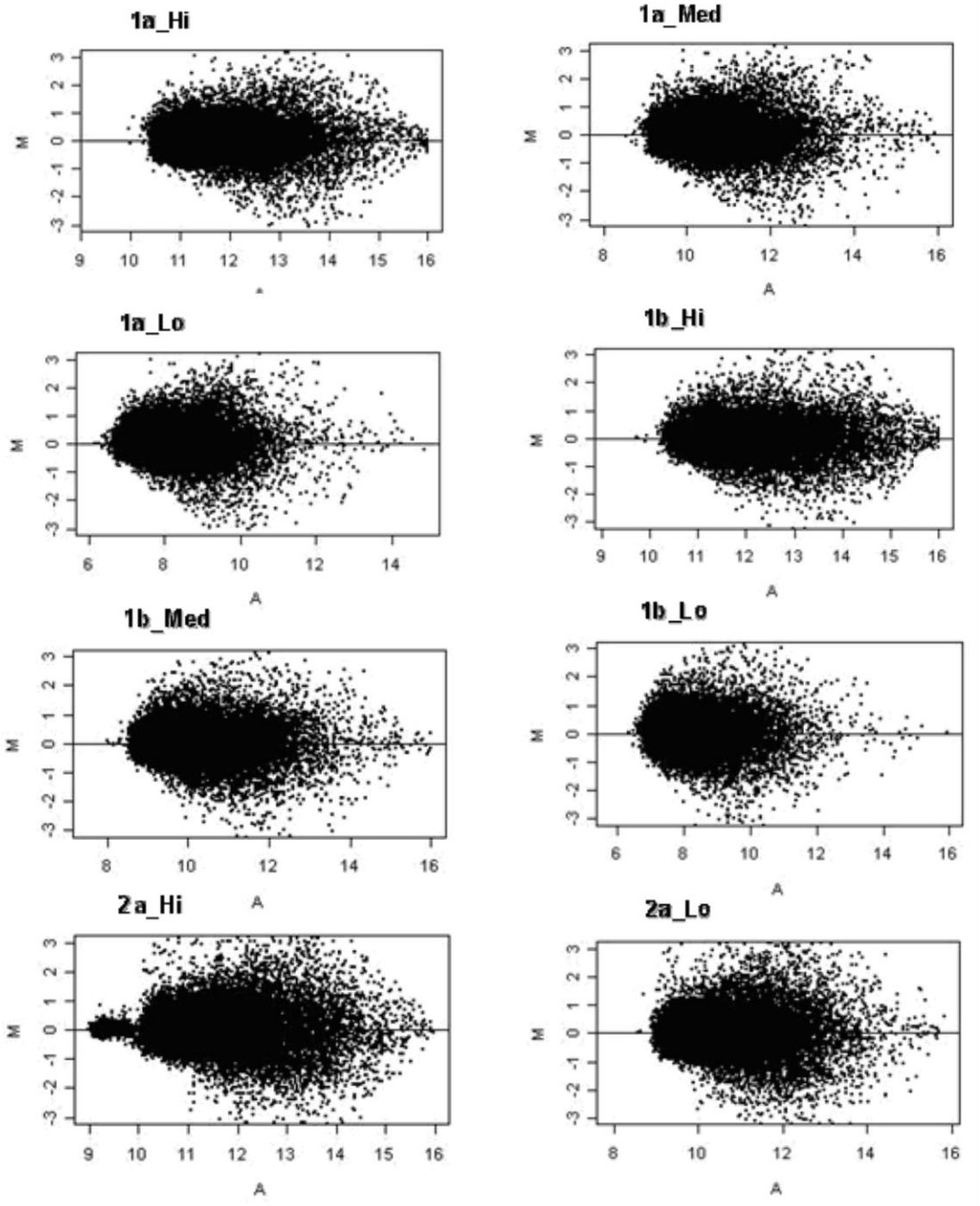


Figure 4.16. **Spatial distribution plots of background and foreground intensities.** Plots of the signal intensity distribution observed in the foreground (A and C) and background (B and D) of both the Cy5 (A and B) and Cy3 (C and D) scans. Slides are, in order left to right, top to bottom, 1a_Hi, Lo, Med, 2a_Hi, Lo, 3a_Hi, Lo, Med, 1b_Hi, Lo, Med, 2b_Hi, Lo, Med, 3b_Hi, Lo. The identified printing errors are indicated in the first scan of the Cy5 foreground by four arrows, and can be observed in each foreground plot. The figures were created in Bioconductor, using Limma and generated using the graphics module of R.


1a_Hi
1a_Med
1a_Lo
1b_Hi
1b_Med
1b_Lo
2a_Hi
2a_Lo
M
A

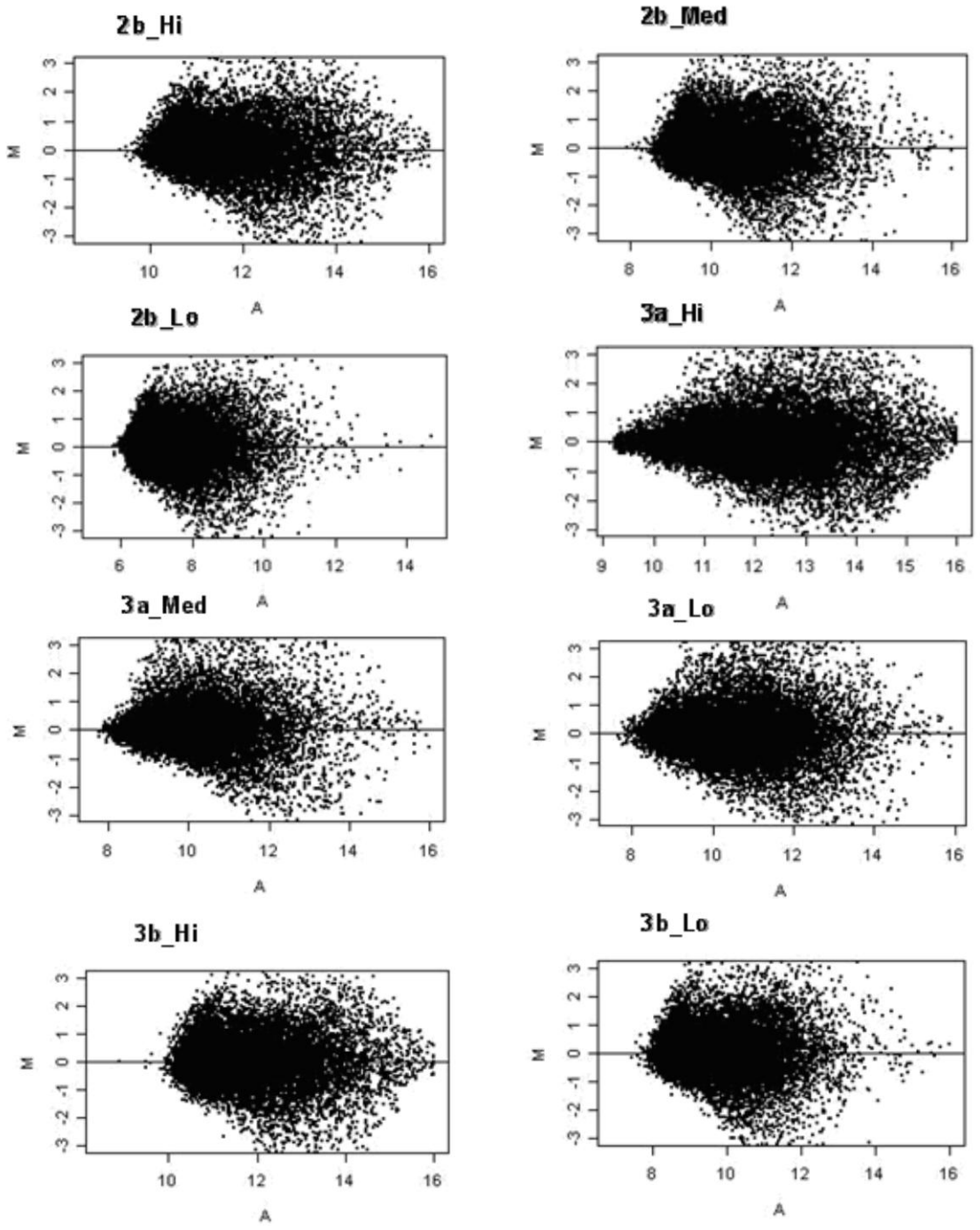


Figure 4.17. MA plots for all analysed features. M (log2(Cy5) - log2(Cy3)) vs. A (1/2(log2Cy5+Cy3)) plots are shown. Plots maintain symmetry along the A axis, indicating no bias toward either fluorophore throughout the signal intensities measured. The figures were created in Bioconductor, using Limma and generated using the graphics module of R.

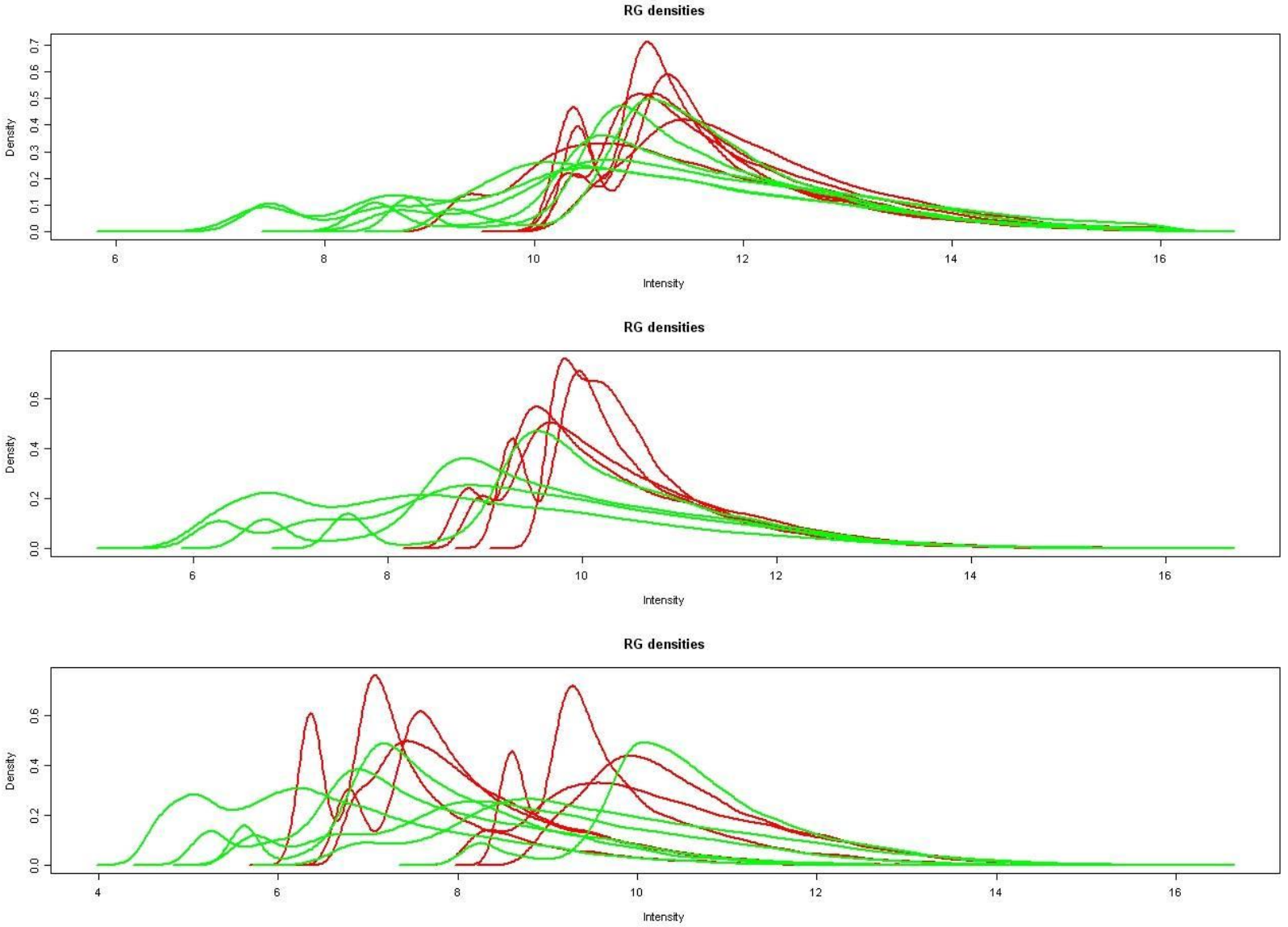


Figure 4.18. **Density plots (raw data).** The proportion of spots at each intensity, before normalisation is shown for each of the three scan levels (high, medium and low). All appear to show a bias to the Cy5 (red) fluorophore at higher intensity and the Cy3 (green) fluorophore at lower intensities.

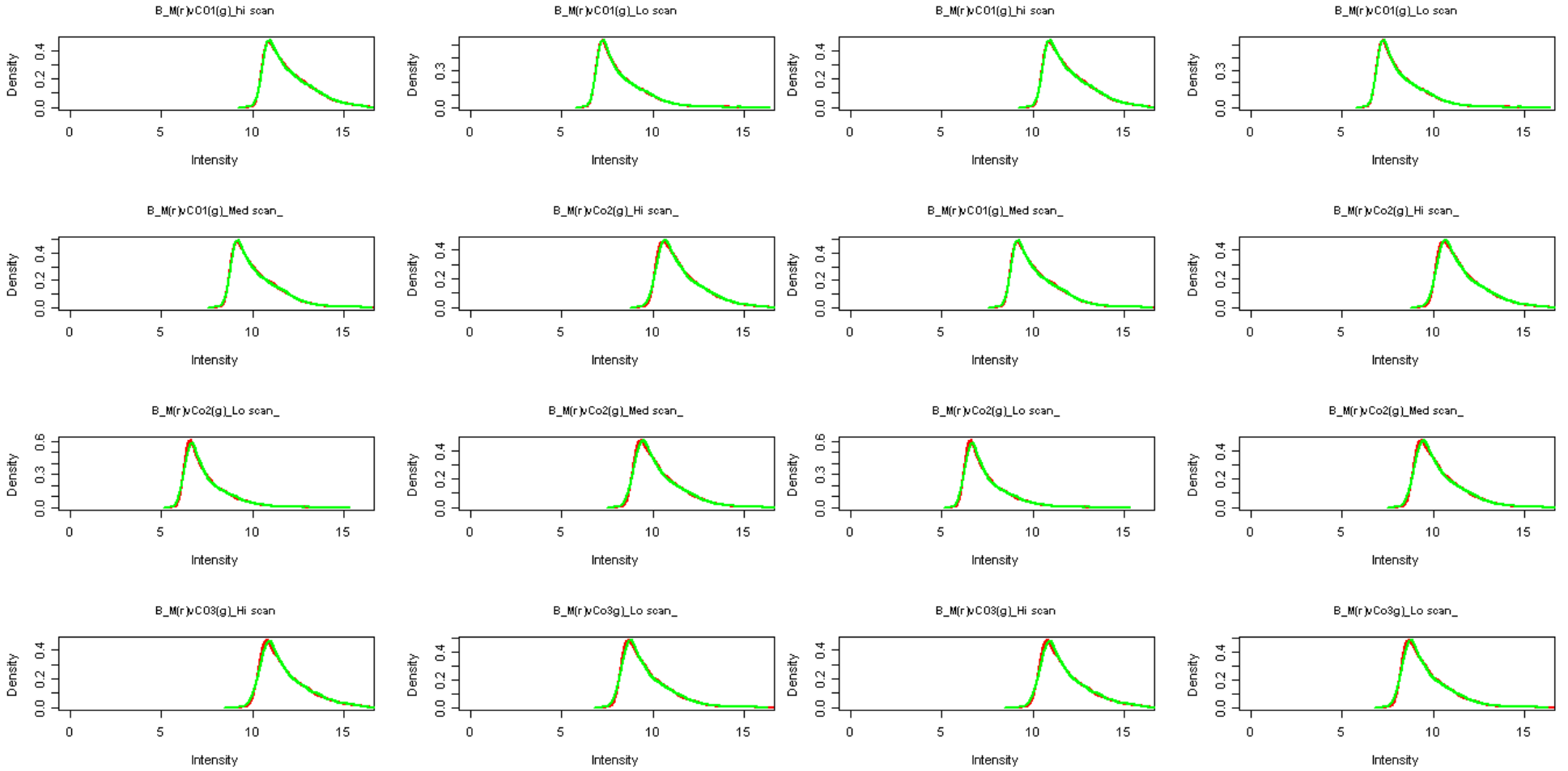
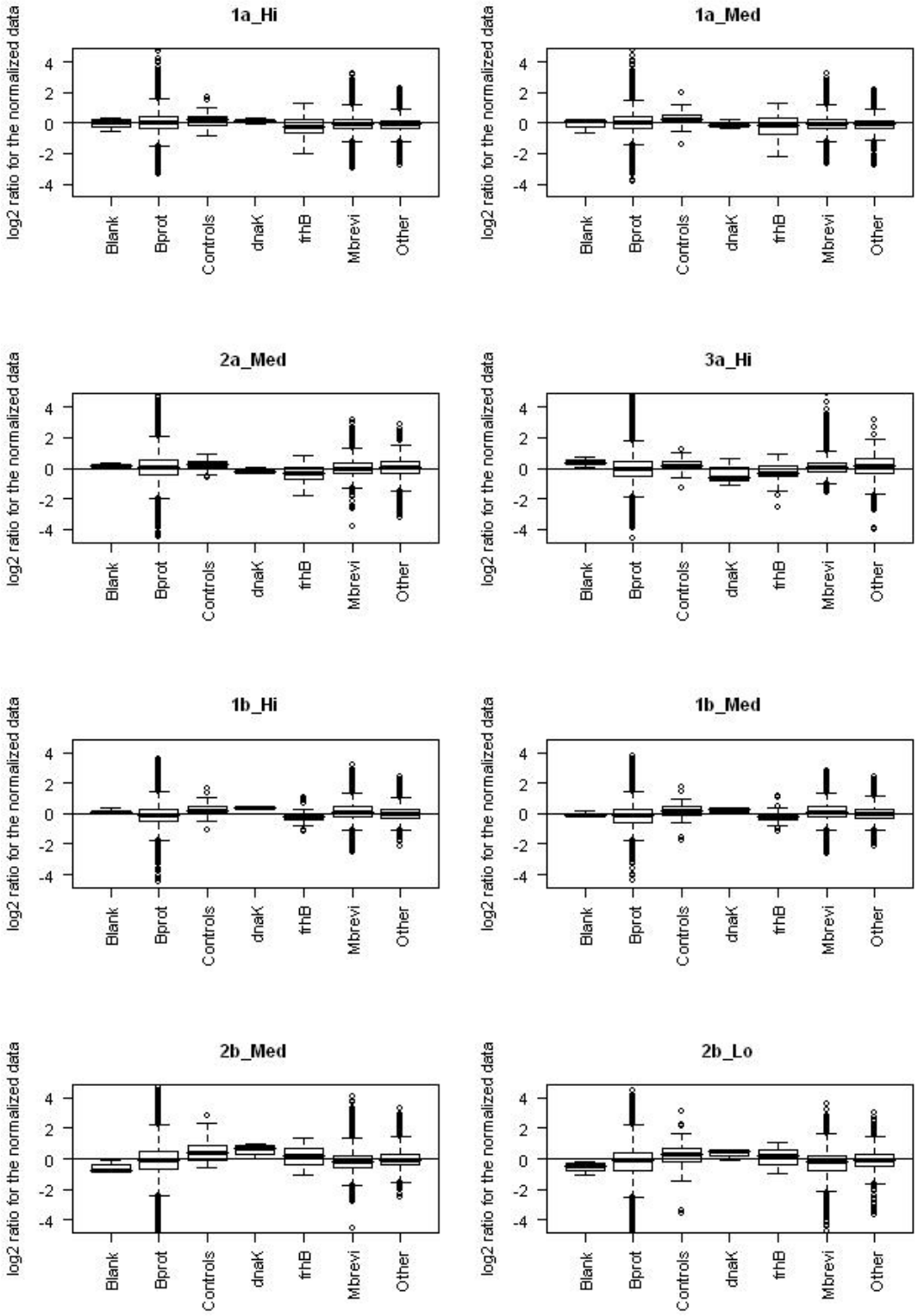


Figure 4.19. **Density plots by slide (normalised data).** The proportion of spots at each intensity, following within-slide normalisation. The Cy5 (red) and Cy3 (green) curves are nearly identical in all slide scans indicating little or no bias to either fluorophore following normalisation.


1a_Hi
1a_Med
2a_Med
3a_Hi
1b_Hi
1b_Med
2b_Med
2b_Lo
log2 ratio for the normalized data
4
2
0
-2
-4
Blank
Bprot
Controls
dnaK
frhB
Mbrevi
Other

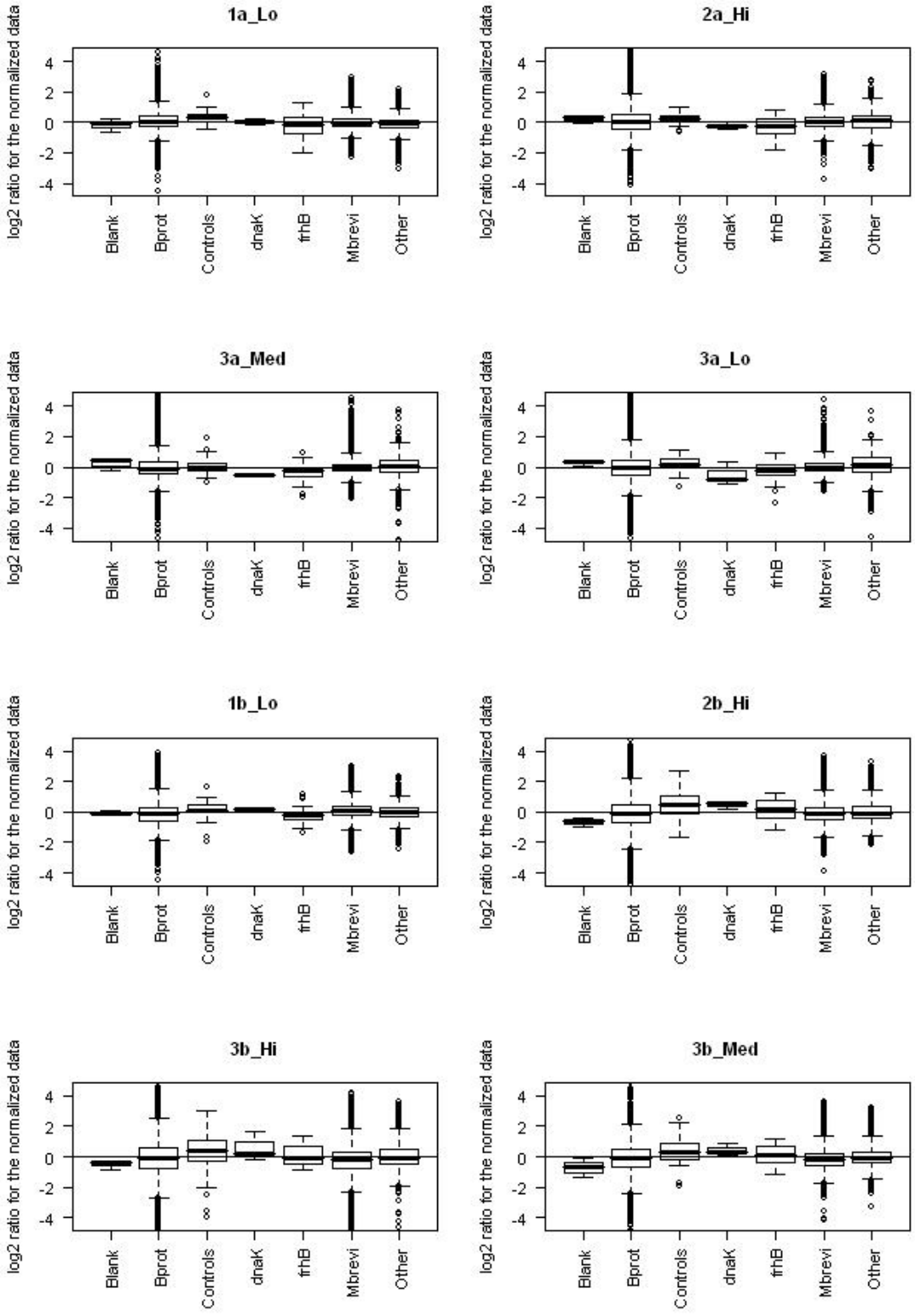


Figure 4.20. **Distribution of feature intensities by category (normalised data)**. Bprot=*B. proteoclasticus* ORFs, Mbrevi = *M. ruminantium* ORFs, Controls = Negative controls, Bprot *dnaK* and Mbrevi *frhB* = constitutively expressed genes Heat shock protein 70 and coenzyme F420 reducing hydrogenase of *B. proteoclasticus* and *M. ruminantium* respectively, Unrecognised = spots which have not been assigned to a category. This figure was created in Bioconductor, using Limma and generated using the graphics module of R.

Analysis of the mRNA transcript levels of the co-culture versus the pooled mono-cultures revealed several processes to be up-regulated in each of the organisms (Table 4.5). Increased expression was seen in many *B. proteoclasticus* ORFs whose products are predicted to have roles in exopolysaccharide biosynthesis and export, flagellar formation, the phosphoenolpyruvate sugar import system (PTS), particularly as it related to fructose, glycogen biosynthesis and storage, RNA catabolism and the conversion of glutamate to NAD^+. Decreased expression was seen in many of *B. proteoclasticus'* ORFs whose products are predicted to be glycoside hydrolases, except those predicted to be involved in the degradation of xylan; non-PTS sugar transporters; and fatty acid biosynthesis, except those predicted to have a role in butyrate production. Increased expression was seen in many of *M. ruminantium* ORFs whose products are predicted to have roles in adherence, glutamate production and methanogenesis. 364 (94%) of the 389 putative pCY360 ORFs were detected at a significant pixel intensity (9,000 > 65,000). Of these, 12 ORFs (4.5%) were found to be up-regulated in the co-culture condition greater than 2-fold with a false discovery rate (FDR) less than 0.05, while 3 (1%) were found to be down-regulated greater than 2-fold (FDR <0.05; Table 4.6). Those genes that were significantly up-regulated are mostly predicted to encode hypothetical proteins (8; ORFs 41, 44, 142, 204, 363, 365, 370 and 378). Four ORFs were assigned more informative annotations, these include: ORF 23, a putative lipoprotein; ORF 42, a ribonuclease H protein; ORF 151, a putative IS605-family transposase; and ORF 208, a putative DNA-binding protein. Those genes significantly down-regulated include: ORF 13, a PilT N-terminus domain protein; ORF 20, a hypothetical protein; and ORF 347, a hypothetical transmembrane protein. Analysis of the distribution of differential regulation around the megaplasmid reveals two operonic structures each encompassing 5 ORFs (Fig. 4.21). The first operon (ORFs 39 to 43), are up-regulated by an average of 2.37 fold $^+/_-$ 0.69 (99% confidence) and encode 4 hypothetical proteins and a ribonuclease H protein. The ORFs of the second operon (ORFs 363 to 367), are collectively up-regulated by an average of 2.26 $^+/_-$ 0.43 (99% confidence), and encode 5 hypothetical proteins that all show ~30% amino acid identity to ORF8 of the bacteriophage f237 isolated from *Vibrio parahaemolyticus*. Adjacent to this potential operon is the most significantly up-regulated gene, ORF370 which is up-regulated 4.56 fold. The constitutively expressed genes butyryl-CoA dehydrogenase (*bcd*; *B. proteoclasticus*), and the gene encoding *M. ruminantium*'s N^5, *N10*-methenyl-H_4MPT cyclohydrolase

(*mch*; *M. ruminantium*) were found to have biological fold changes of 1.07 (FDR = 0.01) and 1.12 (FDR = 0.047), respectively. Blank controls had a biological fold change of 1.27 (FDR = 0.15), and the eight *Arabidopsis* controls all had a low biological fold change (1.00 – 1.08) and high FDR (0.67 – 0.98).

Table 4.5. Observed transcriptional differences in co-culture in comparison to monoculture

B. proteoclasticus

Number	Encoded function*	Fold change*	FDR
	Cellular Aggregation		
Regulation			
5	Diguanylate cyclase (GGDEF-domain) protein	↑ 2.3 – 3.3x	0.5 – 7 x 10^{-3}
1	Sensory box / GGDEF / EAL domain fusion protein	↑ 2.6x	2 x 10^{-3}
Exopolysaccharide production			
3	Polysaccharide biosynthesis protein	↑ 2.1 - 2.8x	1 – 10 x 10^{-3}
5	Glycosyltransferase	↑ 2.1 - 3.8	0.07 – 2 x 10^{-2}
1	Polysaccharide export protein	↑ 2.2x	2 x 10^{-3}
Flagella formation			
2	Flagellar motor apparatus (components A & B)	↑ 2.8 - 5.4x	0.02 - 4 x 10^{-2}
2	Flagellar motor switch protein	↑ 2.6x	3 x 10^{-3}
1	Flagellar basal-body rod protein	↑ 2.0x	3 x 10^{-3}
8	Flagellar biosynthesis proteins	↑ 2.1 - 5.1x	0.07 - 4 x 10^{-2}
1	Flagellin	↑ 3.8x	3 x 10^{-3}
1	Flagellar assembly protein	↑ 2.2x	1 x 10^{-2}
1	Flagellar associated GTP binding protein	↑ 2.6x	4 x 10^{-3}
1	Flagellum-specific ATP synthase	↑ 2.2x	7 x 10^{-3}
	Carbohydrate metabolism		
Regulation			
1	Sugar fermentation stimulation protein, *sfsA*	↑ 6.5x	2 x 10^{-4}
1	PTS-associated DeoR-family transcriptional regulator	↑ 6.0x	5 x 10^{-4}
1	Carbon storage regulator	↑ 2.2x	3 x 10^{-3}
Glycosyl hydrolase			
	Endo-1,4-beta-glucanase, *cel5C*	↓ 6.2x	6 x 10^{-4}
	Alpha-D-glucuronidase, *agu67A*	↓ 4.9x	3 x 10^{-4}
	Alpha-L-rhamnosidase, *ram78A*	↓ 2.8x	6 x 10^{-3}
	Feruloyl esterase, *est1A*	↓ 2.4x	2 x 10^{-2}
	Pectate lyase, *pel1A*	↓ 2.4x	1 x 10^{-3}
	Alpha-L-fucosidase, *fuc29A*	↓ 2.2x	4 x 10^{-3}
	Beta-glucosidase, *bgl3D*	↓ 2.2x	1 x 10^{-3}
Transport			
6	Sugar ABC transport components	↓ 2.3 - 8.3x	0.02 - 3 x 10^{-2}
1	Sugar (glycoside-pentoside-hexuronide) transport protein	↑ 2.0	2 x 10^{-2}

	Feruloyl esterase, *est1A*	↓	2.4x	2×10^{-2}
	Pectate lyase, *pel1A*	↓	2.4x	1×10^{-3}
	Alpha-L-fucosidase, *fuc29A*	↓	2.2x	4×10^{-3}
	Beta-glucosidase, *bgl3D*	↓	2.2x	1×10^{-3}
Transport				
6	Sugar ABC transport components	↓	2.3 - 8.3x	$0.02 - 3 \times 10^{-2}$
1	Sugar (glycoside-pentoside-hexuronide) transport protein	↑	2.0	2×10^{-2}
2	PTS system (fructose components)	↑	3.8 – 4.4 x	$2 – 3 \times 10^{-3}$
1	Fructuronic acid transporter, *gntP*	↑	4.2	4×10^{-3}
Fructose metabolism				
1	Fructose-1,6-bisphosphatase	↑	2.1	4×10^{-3}
1	1-phosphofructokinase	↑	4.2	3×10^{-4}
1	6-phosphofructokinase	↑	4.5	4×10^{-4}
1	Transketolase	↑	2.0	7×10^{-3}
1	Transaldolase	↓	2.6	4×10^{-3}
Glycogen synthesis				
3	Glucose-1-phosphate adenylyltransferase (*glgC* & *glgD*)	↑	2.0 – 3.3x	$0.4 - 2 \times 10^{-2}$
	Nitrogen metabolism			
Protease				
1	Serine protease	↑	3.6	1×10^{-2}
Transport				
1	Amino acid ABC transport system	↑	2.1 - 2.9x	$1 – 2 \times 10^{-3}$
Aspartate metabolism				
1	Aspartate-semialdehyde dehydrogenase	↑	3.3x	8×10^{-3}
1	Asparagine synthase (glutamine hydrolysing)	↓	2.0x	1×10^{-2}
	RNA catabolism			
2	RNA-binding protein	↑	2.0 – 2.7x	$0.9 – 2 \times 10^{-3}$
1	Ribonuclease H	↑	3.0x	5×10^{-4}
2	hicA/B proteins	↑	2.3 – 3.2x	$0.5 – 1 \times 10^{-3}$
1	Uracil permease	↑	2.6x	3×10^{-4}
	Fatty acid synthesis			
2	3-oxoacyl- synthase (I & II)	↓	2.2 - 2.3	3×10^{-2}
1	3-oxoacyl-(acyl-carrier-protein) reductase	↓	2.1x	4×10^{-2}
1	Enoyl-(acyl-carrier-protein) reductase II	↓	2.1x	4×10^{-2}
2	Acyl carrier protein	↓	2.2 - 3.1x	$3 – 4 \times 10^{-3}$
1	Acyl carrier protein phosphodiesterase	↓	2.2x	8×10^{-3}
1	Phosphate butyryltransferase	↑	2.3x	8×10^{-4}
1	Butyrate kinase	↑	2.2x	1×10^{-2}
	NAD^+ synthesis			
1	NH_3-dependent NAD+ synthase	↑	2.8x	1×10^{-2}

M. ruminantium

Number	Encoded function*	Fold change	FDR
	Cellular Aggregation		
5	Adhesin protein	↑ 2.0 – 4.1x	$0.5 – 1 \times 10^{-3}$
	Nitrogen metabolism		
1	hydroxylamine reductase, hcp	↑ 5.9x	2×10^{-4}
1	Glu/Leu/Phe/Val dehydrogenase	↑ 2.0x	1×10^{-2}
	Methanogenesis		
2	Formate dehydrogenase subunits, (FdhA, FdhB)	↑ 2.3 - 3.5x	$0.8 – 3 \times 10^{-3}$
1	Tungsten formylmethanofuran dehydrogenase subunit A FwdA	↑ 2.1x	2×10^{-3}
4	Methyl-coenzyme M reductase subunits (McrB, McrC,McrD, McrG	↑ 2.3 - 3.4x	$0.01 – 2 \times 10^{-2}$
1	Methyl viologen-reducing hydrogenase gamma subunit MvhG	↑ 2.2	4×10^{-3}
4	Tetrahydromethanopterin S-methyltransferase subunits (MtrA, MtrB, MtrC, MtrH)	↑ 2.1 - 3.2x	$0.5 – 1 \times 10^{-2}$

*Results derived from microarray analysis of mRNA transcript levels of *B. proteoclasticus* and *M. ruminantium* grown as a co-culture compared to mono-cultures. The results show the annotated function, number of genes with that annotation affected, biological fold change and false discovery rate (FDR) value. Arrows indicate if ORFs were up or down regulated in co-culture.

Table 4.6 pCY360 ORFs up-regulated in co-culture of *B. proteoclasticus* and *M. ruminantium.*

UP-REGULATED IN CO-CULTURE

	Putative function	Fold change	FDR
370	Hypothetical protein	4.56	$1.7x10^{-3}$
208	Putative DNA-binding protein, HU	3.73	$8.1 x10^{-4}$
42	RibonucleaseH, RnaHA	3.01	$5.2 x10^{-4}$
44	Hypothetical protein	3.00	$1.9 x10^{-3}$
363	Hypothetical protein	2.71	$1.8 x10^{-3}$
365	Hypothetical protein	2.56	$5.8 x10^{-3}$
23	Putative lipoprotein	2.44	$1.2 x10^{-3}$
378	Hypothetical protein	2.4	$8.6 x10^{-3}$
41	Hypothetical protein	2.22	$3.2 x10^{-3}$
204	Hypothetical protein	2.22	$3.3 x10^{-2}$
142	Hypothetical protein	2.12	$3.5 x10^{-2}$
151	Putative IS605-family transposase, TnpBC	2.12	$1.2 x10^{-2}$

DOWN-REGULATED IN CO-CULTURE

ORF	Putative function	Fold change	FDR
347	Hypothetical transmembrane protein	2.79	$1 x10^{-2}$
20	Hypothetical protein	2.16	$3 x10^{-3}$
13	PilT N-terminus domain protein, PinA	2.06	$2 x10^{-2}$

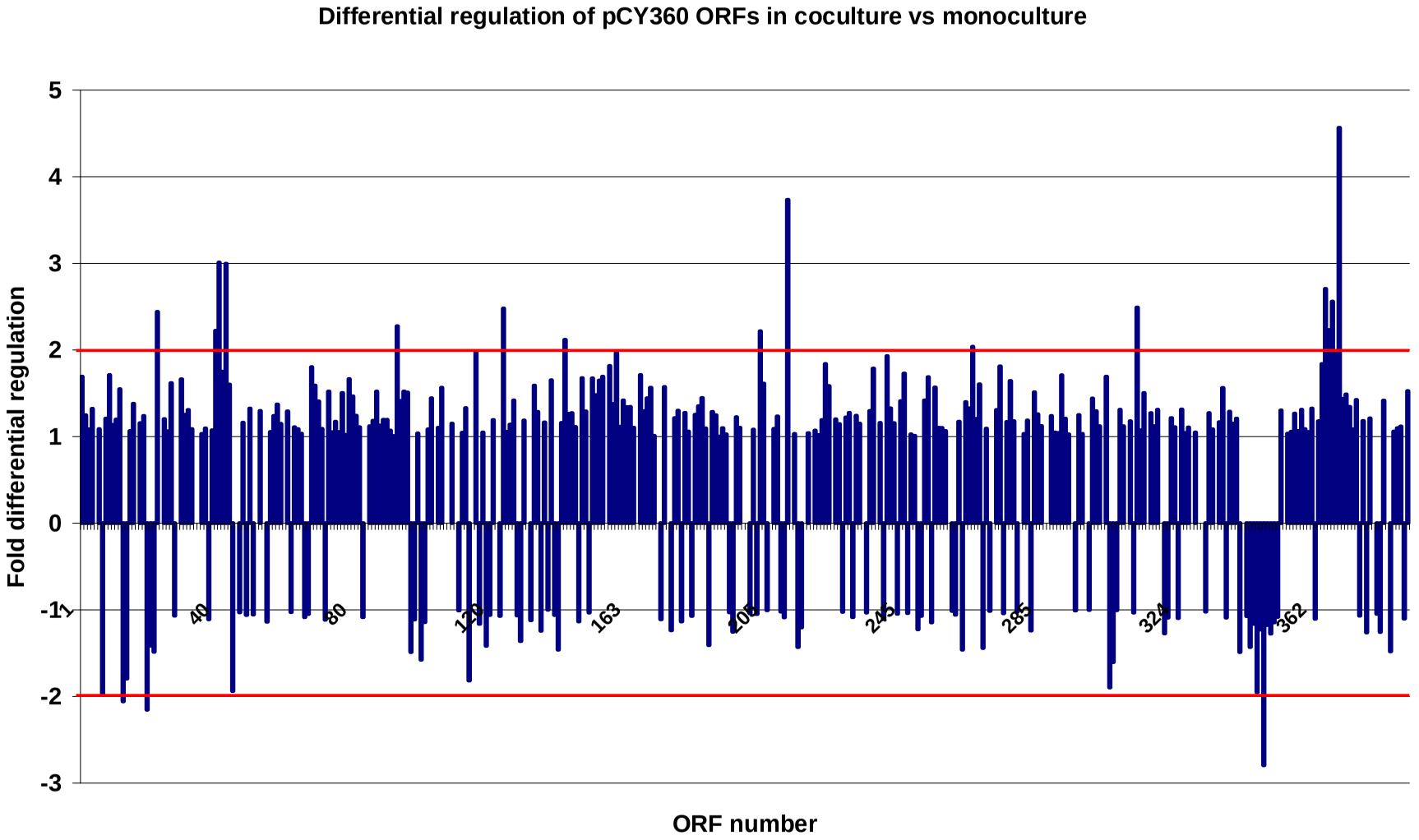


Figure 4.21. **Differential regulation of all 389 predicted ORFs of pCY360. T**he differential regulation of gene transcripts, are arranged in numerical order, for all 389 predicted ORFs of pCY360 as determined by microarray analysis. Red bars show the point of significance (a 2-fold regulatory difference). Two predicted operonic structures with similar levels of up-regulation are evident (ORFs 39 to 43 and ORFs 363 to 367).

4.14 The distribution of large replicons in other *Butyrivibrio* / *Pseudobutyrivibrio* species

The occurrence of auxiliary replicons in bacteria closely related to *B. proteoclasticus* B316^T was investigated using PFGE (Fig. 4.22). Large extra-chromosomal DNA bands were observed in a number of *Butyrivibrio* and *Pseudobutyrivibrio* species, including a 145 Kb band in the *P. ruminis* strain DSM9787, 160 and 240 Kb bands in *B. fibrisolvens* strain D1 and a 360 Kb band in *B. fibrisolvens* C211, 100, 360 and 500 Kb bands in *B. hungatei* strains JK615, C219a and A38 respectively a 145Kb band in *P. xylanivorans* strain Ce52 and 340Kb bands in both *Butyrivibrio sp*. JK619 and *B. proteoclasticus* UC142. Of all strains tested *B. proteoclasticus* B316^T is the only organism possessing three auxiliary replicons. There was no phylogenetic correlation with episomal DNA content with different strains of *B. fibrisolvens* and *B. hungatei* possessing auxiliary replicons of different sizes, while *P. ruminis* DSM9787 is the only one of the three *P. ruminis* strains tested to contain an auxiliary replicon.

4.14.1 Relatedness of pCY360 to auxiliary replicons from other *Butyrivibrio* species

The relationship of the auxiliary replicons detected in other *Butyrivibrio* and *Pseudobutyrivibrio* strains to pCY360 was examined using Southern blots probed with pCY360-derived *repB* amplicons (Fig. 4.22b). The pCY360 *repB* probe detected the pCY360 band as expected, and weakly to the co-resident replicons. The *repB* probe also hybridised strongly with the 145 Kb DNA band in *P. ruminis* but not to any other episomal band. The probes hybridised with DNAs retained in the wells and the irresolvable fraction (containing linear chromosomal as well as open-circular DNAs of each replicon) from all *Butyrivibrio* and *Pseudobutyrivibrio* strains except *B. fibrisolvens* strain C211.

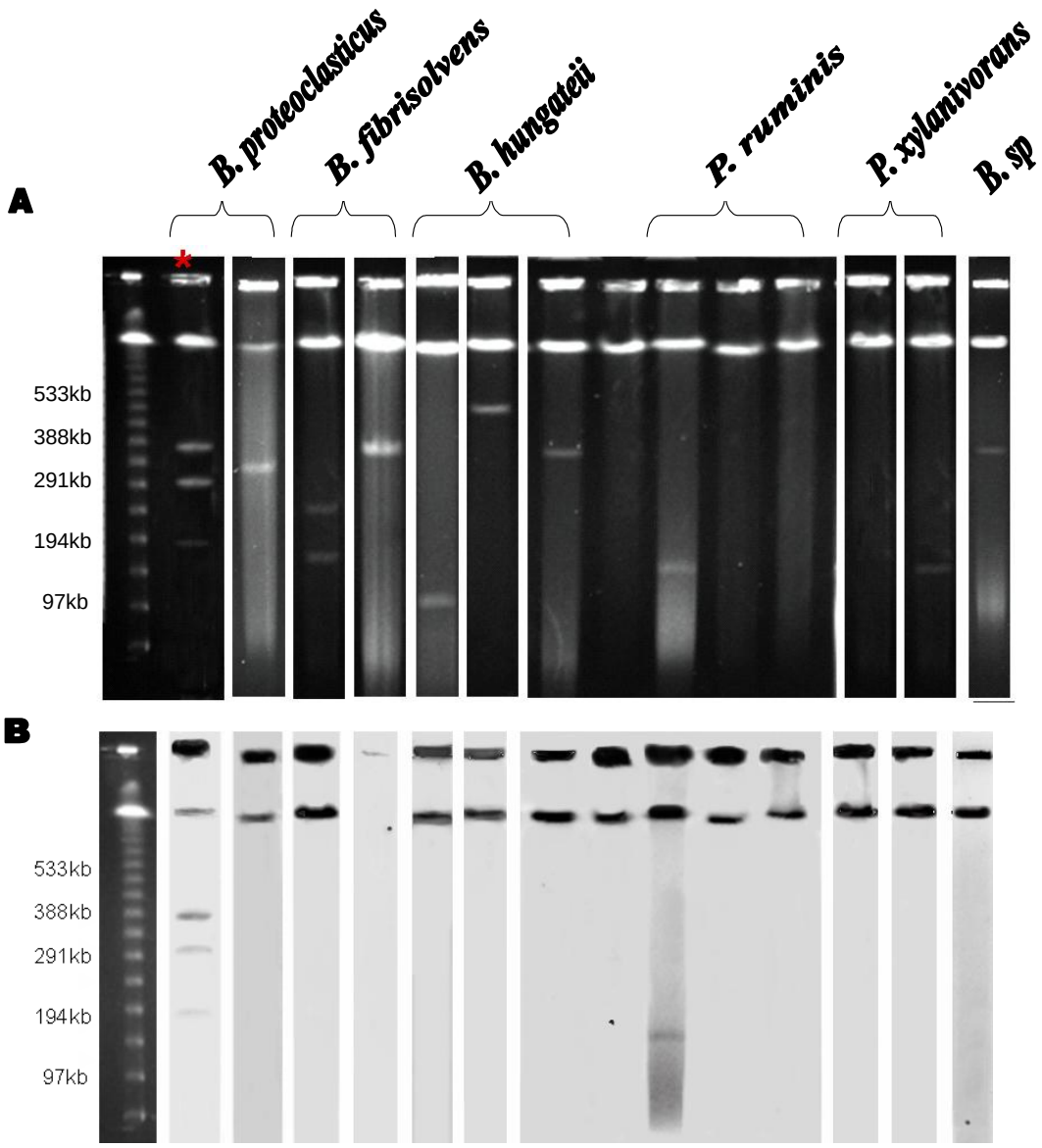


Fig. 4.22 Distribution and relatedness of large extra-chromosomal replicons in *Butyrivibrio* / *Pseudobutyrivibrio* spp. (A) Collated PFGE images of uncut whole genomic DNA extracts from strains representing species of the *Butyrivibrio* / P*seudobutyrivibrio* assemblage. These include, from left to right, *B. proteoclasticus* strains B316^{T} (red asterisk) and UC142, *B. fibrisolvens* strains D1 and C211, *B. hungatei* strains JK615, A38 and C219, *B. crossotus* DSM2876, *P. ruminis* strains DSM9787, CF3 and CF1b, *P. xylanivorans* strains Mz5^{T} and Ce52 and *Butyrivibrio* sp. JK619 (B) Southern blots of the above gels using as a probe an amplicon derived from pCY360s *repB* gene.

4.15 Discussion

4.15.1 Replication of pCY360

The site identified as likely containing the pCY360 origin of replication includes ORFs that align to RepB and ParA proteins. This site also contains two large inverted repeats (IR1 and IR2), an 18bp direct repeat (DR1) and 8 potential *dnaA* boxes. The divergent nature of the ORFs surrounding IR1, along with iterons in the form of DR1 and several *dnaA* box candidates suggests that this region is likely to contain the *oriR*. The structure of this region is consistent with that found in plasmids that replicate via the theta mode of replication (del Solar *et al*., 1998). Several additional findings support this: 1) a 366 bp region aligns strongly with the putative *oriR* of the theta-replicating plasmid pRJF1, 2) The RepB protein aligns best with replication proteins of theta-replicating plasmids and more distantly to those that replicate via strand displacement or rolling-circle, 3) rolling-circle replication has not previously been seen in autonomously replicating DNAs the size of pCY360. However, third-position GC skew analysis suggests a unidirectional mode of replication. Additionally, unlike previously described large extra-chromosomal replicons, or other theta-replicating episomes, most ORFs (89%) are located on one strand, also supporting a unidirectional mode of replication. Several examples exist of unidirectional theta replicating plasmids (Bruand *et al.*, 1991, Le Chatelier *et al.*, 1993, Sun *et al.*, 2006a) and this is proposed to be the most likely style of replication employed by pCY360. Along with RepB and ParA, several other gene products are likely involved in the replication and maintenance of pCY360 including a XerC-family integrase/recombinase and a putative PIN domain toxin-antitoxin (TA) operon. RepB and ParA are thought to enhance plasmid retention through conferring autonomous replication (Moscoso *et al.*, 1995, Moscoso *et al.*, 1997) and equi-partitioning of the replicon during cell division (Meacock & Cohen, 1980), respectively. The XerC-family integrase/recombinase is thought to enhance plasmid stability through the resolution of plasmid multimers (Cornet *et al*., 1994) and the PIN domain toxin-antitoxin (TA) operon is thought to be involved in post-segregational killing (Arcus *et al*., 2005).

4.15.2 Minimal Gene Set ORFs

The presence of the anaerobic ribonucleotide reductase genes *nrdD* and *nrdG* in the pCY360 sequence is unusual for an extra-chromosomal element. the only previous

example being on the conjugative 450Kb megaplasmid pHG1 from *Alcaligenes eutrophus* (Siedow *et al*., 1999). NrdD and NrdG collectively comprise a class III ribonucleotide reductase system that catalyzes the anaerobic reduction of ribonucleotide triphosphates to their corresponding deoxyribonucleotide triphosphates, a process essential for DNA synthesis and repair. In *A. eutrophus* they are essential for anaerobic respiration when the organism uses nitrate as the terminal electron acceptor. This is also the first report of *ftsHA*, *gmkA* and *udgA* genes being encoded on a megaplasmid. In *E. coli,* the *ftsH* gene is chromosomally located and encodes an essential ATP-dependant metalloprotease that is anchored in the cytoplasmic membrane. It is implicated in the degradation of misassembled membrane proteins (Kihara *et al.*, 1995, Kihara *et al.*, 1998, Akiyama *et al.*, 1996) as well as certain cytoplasmic regulatory proteins such as σ32 and λCII. FtsH has also been shown in *E. coli* to control the cellular level of UDP-3-*O*-(R-3-hydroxymyristoyl)–*N*-acetylglucosamine deacetylase (LpxC) (Tomoyasu *et al.*, 1995, Shotland *et al.*, 1997, Ogura *et al.*, 1999). This post-translational control of LpxC, which catalyses the first committed step in lipopolysaccharide biosynthesis, ensures balanced formation of the inner and outer membranes (Ogura *et al*., 1999). In *ftsH* mutants lipopolysaccharide accumulates, altering membrane composition and interfering with plasmid partitioning (Inagawa *et al*., 2001). The *gmkA* gene encodes guanylate kinase, an enzyme that catalyzes the ATP-dependent phosphorylation of guanosine monophosphate (GMP), or deoxy-GMP (dGMP), to guanosine diphosphate (GDP) or dGDP respectively (Oeschger & Bessman, 1966). This function is essential for recycling (d)GMP, and is critical to the biosynthesis of the RNA and DNA purine nucleotides guanosine 5′-triphosphate (GTP) and dGTP respectively. Consequently it is also indirectly involved in the recycling of the bacterial secondary messenger cyclic di-GMP (Hall & Kuhn, 1986, Jenal, 2004), which has been implicated in the regulation of cellular differentiation (Jenal, 2004), cellulose synthesis (Amikam & Benziman, 1989, Ross *et al.*, 1990), and cell surface adhesiveness (Kim & McCarter, 2007, Gjermansen *et al.*, 2006, Simm *et al.*, 2007). The *udgA* gene encodes uracil-DNA glycosylase, a DNA repair enzyme that excises uracil residues from DNA by cleaving the N-glycosidic bond (Duncan *et al*., 1978). Uracil in DNA arises through the deamination of cytosine (Lindahl & Nyberg, 1974) or as a result of the misincorporation of dUTP in place of dTTP, a consequence of the formation of large amounts of dUTP as an intermediate in the biosynthesis of thymidylate (O'Donovan

& Neuhard, 1970). The removal of uracil under these circumstances is required for maintenance of DNA integrity. Interestingly, three of these genes (*nrdGA, nrdDA* and *ssbA*) are found clustered near the predicted *oriR* of pCY360 (Fig. 2). In addition, a thymidylate synthase pseudogene is found within this cluster. The thymidylate synthase pseudogene appears to be truncated, matching reasonably well to the N-terminal ~ 60 amino acids of characterised thymidylate synthases from *E. coli* (Belfort *et al*., 1983) and *Bacillus subtilis* (Tam & Borriss, 1995). The presence of several genes from the Bacterial Minimal Gene Set (Koonin, 2000) in the pCY360 sequence is unusual. Each appears to supply a redundant copy of a gene encoded on the *B. proteoclasticus* B316^T genome. Several of these proteins have roles in DNA metabolism, particularly nucleotide metabolism, and may collectively act to enhance pCY360 retention by reducing the metabolic burden imposed through the titration of such proteins, or their products, away from the chromosome. The FtsH orthologue may act to ensure effective partitioning of pCY360, as well as regulating the large assortment of membrane proteins encoded by this replicon.

4.15.3 Unique enzymatic contributions

The pCY360 replicon appears to encode the only copies of the *B. proteoclasticus* enzymes poly(ADP-ribose) polymerase (PARP; putative) and thermonuclease. PARP, a protein typically encoded by eukaryotic organisms, comprises a regulatory enzyme induced by DNA damage (Oliver *et al*., 1999). It catalyses the covalent attachment of ADP-ribose units from NAD^+ to itself and a limited number of other DNA binding proteins, decreasing their affinity for DNA. The presence of PARP on pCY360 is curious given the typical eukaryotic lineage of the enzyme. Thermonuclease, a remarkably heat-stable endonuclease (Chen *et al.*, 2000, Chesbro & Auborn, 1967), is implicated in the non-specific degradation of DNA and RNA oligomers. Rumen fluid has previously been shown to possess nuclease activity (Ruiz *et al*., 2000), and non-specific nucleases have been shown to be produced by a number of rumen bacteria, including species of the genra *Bacteroidetes*, *Prevotella*, and *Fibrobacter* (Accetto & Avgustin, 2001, MacLellan & Forsberg, 2001, Flint & Thomson, 1990). The nucleotides that would ultimately result from this degradation are known to be scavenged by rumen bacteria and recycled as various cell precursors or, in some cases, as an alternative energy source (Cotta, 1990). The requirement of *B. proteoclasticus* for a heat stable nuclease cannot be explained but may be shared by

several other *Butyrivibrio* strains as discussed (4.15.10).

4.15.4 *oriT* and Mob proteins

The current understanding of bacterial conjugative transfer is based on the system found in Gram-negative organisms. This process is mediated by two multi-protein complexes, the relaxosome and the mating pair formation (mpf) complex, which are coupled via a coupling protein such as TraG. While much of the process appears to be the same in Gram-positive organisms, differences lie in the mpf mechanism used to establish donor:recipient contact, the details of which remain to be fully elucidated. The relaxosome is a multi-protein complex commonly composed of both plasmid- and chromosomally-encoded proteins. The relaxosome complex associates with plasmid DNA at its origin of transfer (*oriT*). In Gram-negative bacteria the mpf complex is a subset of type IV secretion systems, composed of several proteins that establish intimate contact between the donor and recipient, as well as mediating the transfer of the DNA. The pCY360 replicon possesses several ORFs related to components of this mpf complex, the relaxosome, as well as an ORF that aligns strongly to a TraG coupling protein. Both ORF172 (a putative MobA) and ORF292 (a putative RecD/TraA helicase) are potential relaxases, which forms the central component of the relaxosome (Grohmann *et al*., 2003). Relaxases of Gram-positive bacteria have previously been reported to contain two of three conserved motifs (Motifs I and III), characterised by a conserved tyrosine residue and two histidine residues, respectively (Grohmann *et al*., 2003). Alignment of the pCY360 MobA and RecD/TraA proteins with relaxases from the Gram-positive bacterial plasmids pMRC01, pSK41 and pRE25, (Dougherty *et al.*, 1998, Firth *et al.*, 1993, Kurenbach *et al.*, 2003) shows conservation of Motif I in the putative RecD/TraA helicase protein but not in the putative MobA. However, neither proteins show conservation of Motif III (Fig. 4.3).

After forming at the *oriT*, the DNA relaxase acts to cleave a single strand at a specific di-nucleotide site, known as *nic*. Analysis of Gram-positive bacterial plasmid transfer has identified a core *oriT* consensus sequence that contains this *nic* site (Grohmann *et al*., 2003) A total of 21 sequences, all containing absolutely conserved nucleotides of the *nic* site consensus sequence, are present on pCY360. Five of these sequences are present in the putative transfer locus of pCY360 (bp 238026 – 238044) including one

(Candidate 14) that sits within the loop structure of the 19 bp inverted repeat, IR10. Sixteen further sites occur outside the transfer region including two candidates that reside upstream of each of the relaxase candidates (MobA and RecD/TraA; Candidates 12 and 19) and one that also resides within the loop structure of an inverted repeat (IR7; Candidate 13). It is not clear which of these *oriT* candidates might be utilised if pCY360 were conjugatively active.

Type IV secretion systems can transport DNA-protein complexes between cells. These multi-protein systems arrange themselves to form a translocation channel through surface associated structures. This translocation channel mediates the transfer of DNA-protein substrates from cell to cell (Christie *et al*., 2005). There are numerous protein components of this system including structural pilus proteins, muramidases for facilitating access across cell walls, membrane-located proteins that presumably help translocate material across the membrane and lipoproteins which appear to help stabilise other components of the system. Two lipoproteins are found in close association to the *oriT* of pCY360 consistent with VirB7-like Type IV pilus components (Christie *et al*., 2005). The region also contains six novel proteins predicted to have two or more transmembrane spanning regions, consistent with the several membrane-located components involved in a typical Type IV system. Four novel proteins also carry signal peptidase I signal sites (Bendtsen *et al*., 2004). One of these signal peptide-containing proteins matches over a short region to a lytic transglycosylase from the *Enterobacter* phage, P1. Proteins showing weak matches to lytic transglycosylases have previously been found in conjugative transfer systems of other Gram-positive bacteria (Grohmann *et al*., 2003). It has been postulated that lytic transglycosylases may allow DNA and/or proteins to cross the cell envelope by opening local regions of the peptidoglycan (Grohmann *et al*., 2003, Abajy *et al.*, 2007). Type IV secretion systems are also implicated in protein secretion (Lu *et al*., 1997). The pCY360 replicon contains 2 signal peptidases (TraF, ORF232 and SipB, ORF279) that reside either side of the 35 Kb region containing the majority of conjugative transfer-related proteins. This locus may either enable conjugative transfer of the pCY360 replicon, utilising a novel relaxase, or it may mediate Type IV protein secretion. Additionally this site may lead to the formation of a Type IV pilus involved in bacterial adhesion as previously seen in Gram-positive rumen bacteria (Rakotoarivonina *et al*., 2002).

4.15.5 pCY360 can potentially influence the cell envelope composition of *B. proteoclasticus*

The pCY360 replicon encodes 31 proteins that are predicted to span the membrane two or more times, most of which encode novel (or hypothetical) transmembrane proteins. Several other pCY360-encoded proteins have the potential to influence cell membrane topology, including an ATP-dependent zinc metalloproteinase, FtsH, and two signal peptidase I proteins (both described above); a putative membrane-bound GGDEF protein; and two putative class B sortase proteins (Dramsi *et al*., 2005). GGDEF, a regulatory protein that catalyses cyclic di-GMP synthesis (Simm *et al*., 2004), acts as a secondary messenger regulating cell surface adhesiveness (D'Argenio & Miller, 2004). As discussed above, pCY360 also encodes *gmkA* which may act indirectly in the recycling of cyclic di-GMP (Hall & Kuhn, 1986, Jenal, 2004). Recent work suggests that GGDEF can also modulate this process independently of cyclic di-GMP (Holland *et al*., 2008). Sortase B proteins are transamidases that covalently link proteins that possess a C-terminal NXZTN or NPKTG motif (Maresso *et al*., 2006) to the peptidoglycan of Gram-positive bacteria (Bierne *et al*., 2004). Collectively these pCY360-encoded proteins have the potential to significantly impact the composition and functions of the *B. proteoclasticus* cell envelope.

4.15.6 Transposases

The presence of transposase genes within the pCY360 sequence with sequence similarity to transposases found on both the chromosome and BPc2, suggests that transposon-mediated gene shuttling between pCY360 and other *B. proteoclasticus* replicons may have occurred in the past. However, given the low sequence similarity between each it is unlikely that such gene shuttling still occurs. Broker and colleagues (2004), in their analysis of the 101-kilobase-pair megaplasmid pKB1, isolated from *Gordonia westfalica* Kb1 found evidence to suggest that transposons may flank mobile "metabolic regions". Analysis of pCY360 found only one similar region of 10 Kb flanked by IS605-family transposases. However, the 14 genes within this region encoded a conserved hypothetical protein, a DNA ligase, a protein resembling an archaeal histone-like protein, and 11 hypothetical proteins. Therefore no obvious metabolic function was discernable within this region.

4.15.7 RepB phylogeny and codon usage

Phylogenetic analysis of the pCY360 replicon, shows it forms a distinct branch from other plasmids suggesting this replicon diverged from its closest related plasmid, currently described, some time ago. Following its acquisition several mechanisms appear to have been available or were subsequently acquired enhancing this replicons retention. These include autonomous replication, plasmid partitioning, multimer resolution, post-segregational killing (all described above), and possibly restriction modification (ORF106, a modification methylase and possibly ORF143, which shows weak similarity to an N6 adenine-specific DNA methyltransferase). Three pieces of evidence suggest pCY360 has co-evolved with the major chromosome for a long evolutionary time:

1) The %G + C content, which is known to be relatively constant within a species but vary between species (Forsdyke & Mortimer, 2000), is similar between the two replicons.

2) The replicons not only utilise the same major codons, but have near identical codon usage frequencies. It is known that in most organisms selection acts to bias codon usage frequencies towards a subset of all potential codons. The codons used typically reflect the abundance of the corresponding tRNAs available within the organism (Kanaya *et al*., 1999). Consequently, genes which use the preferred codons are thought to maximize their speed of elongation, to minimize the costs of their proofreading, and maximize the accuracy of their translation (Stoletzki & Eyre-Walker, 2007). It has been shown in several organisms that translationally optimized genes have higher expression levels (Gouy & Gautier, 1982, Gupta & Ghosh, 2001).

3) Phylogenetic analysis of the pCY360 RepB protein shows it to be significantly diverged from other plasmid replication initiation proteins. However, this apparent divergence could be confounded by the limited amount of sequence information derived from plasmids or species within the rumen environment and available in public databases.

4.15.8 Is pCY360 an essential part of the *B. proteoclasticus* genome?

The inability to cure *B. proteoclasticus* of the pCY360 replicon suggests it may be required for the survival of *B. proteoclasticus*. If this is true it is not clear which component(s) of pCY360 constitute the essential element(s). The genes described in the Bacterial Minimal Gene Set are likely to be important, however, they are not

unique to the genome and therefore it is more likely they alleviate the metabolic cost of the megaplasmid and potentially its co-residing auxiliary replicons. The two unique enzymes contributed to the genome (PARP and thermonuclease) are unlikely to be essential to *B. proteoclasticus*. One possibility is post-segregational killing via the PIN-domain protein, which putatively acts as the toxin in a toxin-antitoxin (TA) operon. However, the retention of pCY360 over the time it appears to have resided within the genome of *B. proteoclasticus*, as indicated by GC content and codon usage, is unlikely to be purely attributable to this TA operon. This is because the only portion of the plasmid protected from mutation over time by the TA operon would be the antitoxin component. Therefore an inactivating mutation within the toxin component or deleterious mutation to the remainder of pCY360 could freely occur, provided the replicon encoded nothing to enhance the fitness of *B. proteoclasticus*.

4.15.9 Microarray analysis

The presence of a large number of extracellular and membrane proteins encoded by pCY360, as well as the presence of proteins involved in the regulation of membrane topology led to the hypothesis that pCY360 may contribute to interspecies interactions through signalling and/or cell to cell adhesion. Previous studies have shown that many interspecies interactions occur within the rumen environment. For example, the transfer of succinate for conversion to propionate (Scheifinger & Wolin, 1973), the transfer of hydrogen for conversion to methane (Ushida *et al.*, 1997, Wolin, 1976) and during the degradation of plant biomass with respect to cellulose (Fondevila & Dehority, 1996), hemicellulose (Miron *et al*., 1994) and plant cell wall protein (Debroas & Blanchart, 1993) hydrolysis. *M. ruminantium* was selected to investigate the potential contribution of the megaplasmid to intraspecies interactions because it is known to interact with hydrogen-producing bacterial species from the rumen (Vogels & Stumm, 1980, Ushida *et al*., 1997) and is therefore likely to participate in interspecies hydrogen transfer with *Butyrivibrio*, and an almost complete genome microarray was available to analyse changes in genome-wide gene expression. This analysis was unique in respect of the fact that preceding studies had focused on this interaction using only cellulytic bacteria as the hydrogen-donor. Microscopic examination of *B. proteoclasticus* in co-culture with *M. ruminantium* demonstrated the formation of cell to cell aggregates within 120 min of *B. proteoclasticus* inoculation. Both organisms appeared to up-regulate genes whose

products are predicted to be involved in co-aggregation. *B. proteoclasticus* appears to mediate aggregation through EPS and flagellar biosynthesis, while *M. ruminantium* expresses genes encoding large adhesin proteins. EPS production is a typical bacterial tool to mediate biofilm formation (Crawford *et al.*, 2008, Danese *et al.*, 2000). Although tempting to speculate that *B. proteoclasticus*‟ EPS proteins may be allosterically-regulated through the secondary messenger cyclic di-GMP, stimulated by the five di-guanylate cyclase-encoding genes and the GGDEF/EAL-dual domain protein found to be up-regulated in co-culture, as previously demonstrated (D'Argenio & Miller, 2004); cyclic di-GMP has additionally been implicated in the regulation of several other physiological responses (Amikam & Benziman, 1989, Amikam *et al.*, 1995). Flagella production is typically inversely correlated with cyclic di-GMP levels and EPS production, likely due to its opposing role in motility, however, the ability to form a flagellum has been shown to be essential to biofilm formation in several organisms (Gavin *et al.*, 2002, O'Toole & Kolter, 1998). The up-regulation of 17 genes related to flagellar biosynthesis in *B. proteoclasticus* suggests co-aggregate formation in this organism is also likely to involve its flagellum. Further, flagellar biosynthesis does not appear to be negatively regulated by di-guanylate cyclases in *B. proteoclasticus*. *B. proteoclasticus* up-regulates genes whose products are thought to be involved in PTS transport of sugars. The expression of genes related to EPS biosynthesis has recently been found to be co-regulated with components of the PTS system in *Vibrio cholera* (Houot & Watnick, 2008). *B. proteoclasticus* also appears to up-regulate genes that interact in the conversion of fructose to glycogen as well as its subsequent storage within the cell. Collectively this suggests *B. proteoclasticus* in co-culture attempts to scavenge PTS sugars, in-particular fructose, for conversion to glycogen as an energy storage mechanism. This may be a general biofilm response, where *in-vivo* the organism would engage in a biofilm involving a consortia of bacteria, including cellulolytic bacteria that would supply sugars compatible with the *B. proteoclasticus*‟ PTS transport system. It is likely that a catabolite repression response invoked by the PTS system is responsible for the down-regulation of most genes involved in polysaccharide degradation as well as non-PTS sugar uptake systems. Yet, this response does not appear to extend to those involved in xylan-degradation. Analysis of the *B. proteoclasticus* genome suggests it lacks two enzymes critical to *de-novo* biosynthesis of NAD^+, and instead relies on scavenging exogenous nicotinamide riboside using enzymes encoded by BPc2 (discussed later). The final

step in this altered biosynthetic pathway is the transfer of an ammonium ion from glutamate to deamino-NAD^+ forming NAD^+. This step is catalysed by the glutamate-dependent NAD-synthase, NadE. *M. ruminantium* appears to up-regulate genes involved in the production of glutamate, while *B. proteoclasticus* up-regulates *nadE* along with the ATPase and substrate-binding protein cognate to an amino acid ABC-transporter present on BPc2. The upregulated ABC-transporter has strong sequence identity (~70% aa identity) to a glutamine-specific ABC-transporter from *Bacillus sterothermophilus* (Wu & Welker, 1991). Collectively this suggests cooperative production of this essential cofactor. *M. ruminantium* up-regulates many critical components of the methanogenesis pathway suggesting it can utilise H_2 and formate more efficiently when supplied by *B. proteoclasticus*. *B. proteoclasticus* is known to produce formate as an end product of fermentation (Attwood *et al*., 1996). With the exception of two genes involved in butyrate synthesis, many genes involved in fatty acid metabolism are down-regulated. Previous studies have found that some fatty acids inhibit biofilm formation (Inoue *et al*., 2008). Consequently, fatty acids have been proposed to have a regulatory function.

The vast majority of ORFs (364 ORFs, 94%) assigned to the pCY360 replicon after manual annotation were detected by the microarray at intensities significantly greater than the background, suggesting they were transcribed and consequently supporting their status as coding sequences. Of these ORFs, 15 were found to be differentially regulated between the mono- and co-culture conditions. Twelve ORFs were up-regulated, including genes predicted to encode a lipoprotein, a putative DNA-binding protein, a ribonuclease H protein, a putative IS605-family transposase and two hypothetical proteins that show approximately 30% identity to ORF 8 of the bacteriophage f237. Ribonuclease H is a non-specific endonuclease that catalyzes the hydrolysis of the 3'-O-P-bond of RNA in DNA/RNA duplexes. Ribonuclease H has a role in DNA replication, where it is responsible for removing the RNA primer to allow completion of the newly synthesized DNA. However, it is unlikely that the rate of *B. proteoclasticus* DNA replication was enhanced by the interaction in co-culture as no other gene directly related to DNA replication was found to be significantly up-regulated and the growth rate of *B. proteoclasticus* was not significantly different between the two conditions. Ribonuclease H is also important in retrovirus and retrotransposon replication, where it performs three related functions. It mediates the

degradation of the original RNA template following reverse transcription, it generates a polypurine tract (the primer for plus-strand DNA synthesis), and it also carries out the final removal of RNA primers from newly synthesized double stranded DNA. While it is tempting to speculate the up-regulated ORFs that bear sequence similarity to the bacteriophage f237, may encode a retrovirus that has been activated during co-culture, there is no reverse transcriptase gene encoded nearby, or at all, on the pCY360 replicon. A reverse transcriptase is encoded by the *B. proteoclasticus* main chromosome, but it was not found to be differentially regulated in this experiment. Therefore the significance of ribonuclease H up-regulation in the co-culture condition remains unclear at this stage.

Contrary to the experimental hypothesis, only one of the proteins predicted to be extracellular or associated with the cell membrane, (a putative lipoprotein), was found to be significantly up-regulated. Lipoproteins are implicated in transportation and adhesion, both of which would be potentially valuable to the interspecies interaction. The up-regulated putative DNA-binding protein may act as a novel transcriptional regulator. Two potential operonic structures were identified in pCY360, each consisting of five genes. The first encodes four hypothetical proteins and the ribonuclease H, while the second encodes the five hypothetical proteins that resemble ORF 8 of bacteriophage f237. The significance of ribonuclease H up-regulation has been discussed above and the co-regulated hypothetical proteins within this operon provide no further evidence of a potential function. However, recent microarray analysis has shown this operon is also up-regulated when *B. proteoclasticus* is grown in xylose as opposed to xylan (Kong, 2007), the significance of this is unclear. ORF 8 of bacteriophage f237 encodes a protein of unknown function. Its presence has previously been found to correlate with virulence of the phage (Nasu *et al*., 2000). ORF 8 is unique to the bacteriophage of *V. parahaemolyticus* strain O3:K6, a highly infectious strain (Nasu *et al*., 2000). Its presence upon pCY360 and within the *B. proteoclasticus* genome, and its potential role in interspecies interactions with *M. ruminantium* are unclear.

Three ORFs were found to be significantly down-regulated, these include ORFs predicted to encode a protein with a PilT N-terminal domain (PIN domain) and a hypothetical transmembrane protein. The PIN domain containing protein, as

previously mentioned, potentially encodes the toxic component of a toxin-antitoxin operon (Arcus *et al*., 2005). The ORF predicted to encode the antitoxin component (ORF 12) was found to be down-regulated to a similar extent (1.79 fold; FDR = 0.002) consistent with the hypothesis that these proteins are polycistronic. While the down-regulation of the PIN-domain protein is on the border of significance and falls below this threshold when collectively analysed alongside its antitoxin component, it is conceivable that the toxic component could, in co-culture, affect *M. ruminantium*. Therefore its down-regulation could be to safeguard the methanogenic symbiont.

4.15.10 Distribution and relatedness of auxiliary replicons in other *Butyrivibrio* /*Pseudobutyrivibrio* spp

The genomic architecture of 14 species representing all genera of the *Butyrivibrio – Pseudobutyrivibrio* assemblage was investigated. Several further strains were not able to be examined due to a non-specific nuclease activity that was unable to be neutralised by heating or treatment with an alkaline buffer. Large molecular weight DNAs, consistent with megaplasmid-like elements, were found to be common to this assemblage of bacteria, with 10 of the 14 strains tested possessing at least 1 megaplasmid-like element. No phylogenetic correlation was observed with regard to their size or distribution within the assemblage. Phylogenetic discordance was also observed in other bacterial assemblages investigated for megaplasmid distribution (Rosenberg *et al*., 1981, Li *et al.*, 2007). Southern blot analysis using the *repB* gene from pCY360 further showed a lack of phylogenetic relationship with only a 145 Kb band from *P. ruminis* DSM9787 hybridising to the probe. 16*S* rRNA gene phylogeny (Fig. 1.1; Moon *et al*. 2008) shows *P. ruminis* to be one of the most distantly related species to *B. proteoclasticus* of the entire assemblage.

4.16 Summary

The pCY360 megaplasmid accounts for nearly 10% of the entire *B. proteoclasticus* B316^T genome and constitutes the fourth largest extra-chromosomal replicon described to date in a Gram-positive bacterium. Despite contributing a significant part of the genome, there is no clear function attributable to pCY360. The majority of ORFs have no significant hits to any previously described genes, nor matches to genes of unknown function. The metabolic cost on the organism of retaining such a large replicon at approximately four copies per chromosome is likely to be large. Despite this, pCY360 was not able to be cured from *B. proteoclasticus* by any of the methods tested. The megaplasmid contains many genes that are predicted to encode proteins capable of affecting the topology of the *B. proteoclasticus* membrane, yet these genes show no up-regulation in co-culture with the rumen methanogenic archaeon *M. ruminantium*. Several genes are differentially-regulated during this interaction, and although their functions are not yet clear, their identification provides a direction for future analysis. Similar large DNAs are observed in other species of the *Butyrivibrio* / *Pseudobutyrivibrio* assemblage, however their distribution and size shows no phylogenetic correlation. Furthermore all except a megaplasmid-like replicon in a comparatively distantly-related *Pseudobutyrivibrio ruminis* species appear to utilise distinct replication machinery.

5 The secondary chromosome of *B. proteoclasticus*

5.1 Introduction

Bacterial genomes are classically described as being composed of a single circular chromosome. This paradigm was initially supported by analysis of genomic architectures of the archetypical Gram-negative and Gram-positive bacteria *Escherichia coli* and *Bacillus subtilis*, respectively. It was later discovered that bacterial genomes could additionally possess one or more plasmids of various sizes. However, the autonomous replication and dispensability of plasmids under certain conditions suggested they were transient or parasitic entities. Over time it has become apparent that many bacterial genomes are not confined to a simple single chromosome structure and approximately 30 species have been described as possessing more than one chromosome. These bacterial species are almost exclusively of the α, β and γ classes of the Proteobacterial phylum. To date the only described Gram-positive organism possessing two chromosomes is *Deinococcus radiodurans* R1. In *D. radiodurans* the replicon is designated as a secondary chromosome, based on the presence of a tRNA, and proteins involved in amino acid utilisation and cell envelope formation (White *et al*., 1999).

This chapter reports the detailed analysis of the third largest *B. proteoclasticus* replicon, BPc2, which satisfies all current definitions of a secondary chromosome. BPc2 was completely sequenced and finished to a Q40 (equivalent to Phred 40) quality standard. It also examines the distribution of rRNA operons among other large replicons of the *Butyrivibrio* / *Pseudobutyrivibrio* assemblage.

5.2 Sequence analysis of BPc2

The complete sequence of the BPc2 replicon is 302,355 bp in length. The replicon has a %G + C content of 40.04% (40.67% in coding regions, 36.80% in non-coding regions), which is nearly identical to that of the major chromosome (40.2%). GLIMMER (Salzberg *et al*., 1998) analysis identified 260 potential ORFs, however 19 were eliminated for reasons described previously (see Section 4.2). Additionally, 14 ORFs were identified manually that appeared to have been missed by GLIMMER analysis, giving BPc2 a total of 255 ORFs likely to be protein coding regions. BPc2 encodes two complete ribosomal RNA operons (16*S*, 5*S*, 23*S* rRNAs) and two tRNAs (aspartic acid and threonine). The predicted BPc2 ORFs have an average size of 1,020 bp, this is biased by a small number of much larger ORFs as the median ORF size is 873 bp (Fig. 5.1). BPc2 has a gene density of 0.84 genes / kb and gene coverage of 86%. Unlike pCY360, BPc2 ORFs are evenly spread between the Watson (41%) and Crick (59%) strands and analysis shows the replicon to have a classical symmetric third-position GC skew (Fig. 5.2). The distribution of start codons was found to be 85% ATG, 8% GTG and 7% TTG. Proteins encoded by BPc2 ORFs have an average isoelectric point of 6.23 (median 5.18; Fig. 5.1) and unlike pCY360, the majority of these predicted proteins can be ascribed a putative function (167, 65%) or described as conserved hypothetical proteins (36; 14.12%). BPc2 ORFs encode 50 (20%) proteins predicted to span the membrane more than once and a further 44 ORFs (18%) predicted to be secreted proteins. Some (21) of the predicted proteins have during the tenure of this thesis been identified by tandem mass spectrophotometry analysis of proteins separated on 1-dimensional gels (Dunne, in preparation). These protein identifications resulted in two hypothetical proteins being reclassified as an uncharacterised protein, and an uncharacterised secreted protein and one conserved hypothetical trans-membrane protein being reclassified as an uncharacterised conserved trans-membrane protein. The ORFs that are able to be assigned functions appear to form several clusters (Fig. 5.2). These clusters appear to be largely dedicated to replication, cofactor uptake and metabolism, chemotaxis, energy metabolism and detoxification. The energy metabolism cluster of ORFs is the largest, forming two distinct groups, one which is interrupted by the detoxification cluster.

5.3 The BPc2 origin of replication

A single site spanning nucleotides (298075 - 1164) was identified as likely possessing the origin of replication (*oriR*) (Fig. 5.2). The region encodes a putative replication initiation protein, RepB (ORF1) and a plasmid partitioning protein, ParA (ORF255). These proteins show 53% and 23% amino acid identity to RepB and ParA encoded by pCY360. These ORFs are divergently transcribed and separated by a large (3,457 bp) intergenic region. The divergent nature and size of this intergenic region, suggested that it may contain the BPc2 *oriR*. Analysis of the intergenic region identified 21 direct repeats, three inverted repeats and a 33 bp inverse repeat (IV1). Aside from the inverse repeat, all repeats fell into four families: a 16 bp inverted repeat, IR1; a 17 bp direct repeat DR1 (two copies); and a direct repeat DR2 that has ten copies that share significant nucleotide identity but vary in length from 12 to 30 bp. IR1 appeared three times therefore can additionally form a series of three direct repeats on each strand. Eleven potential *dnaA* boxes were identified as well as a 366 bp AT rich (17% GC) region that shows 47% nucleotide identity to the putative *oriR* of *B. fibrisolvens* plasmid, pRJF1 and 53% identity to the predicted *oriR* of pCY360 (Fig. 5.2). This region spans two dnaA box candidates and a copy of DR1 and DR2.

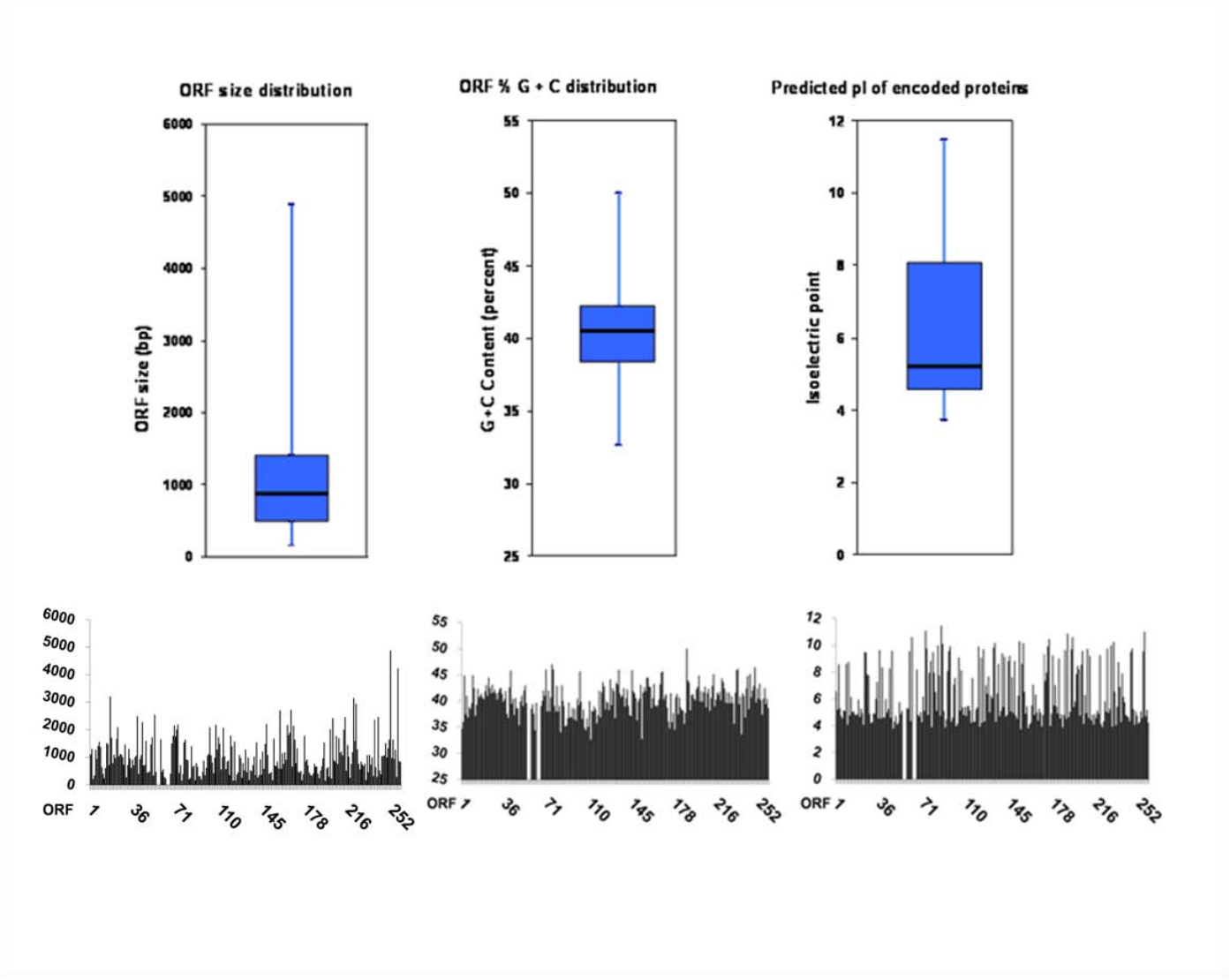
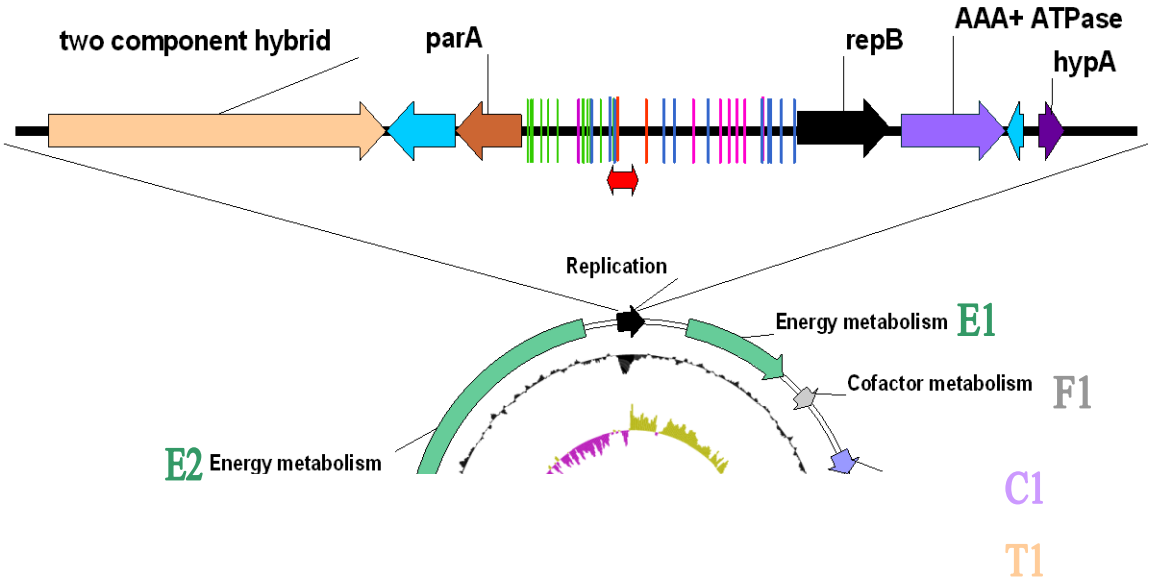
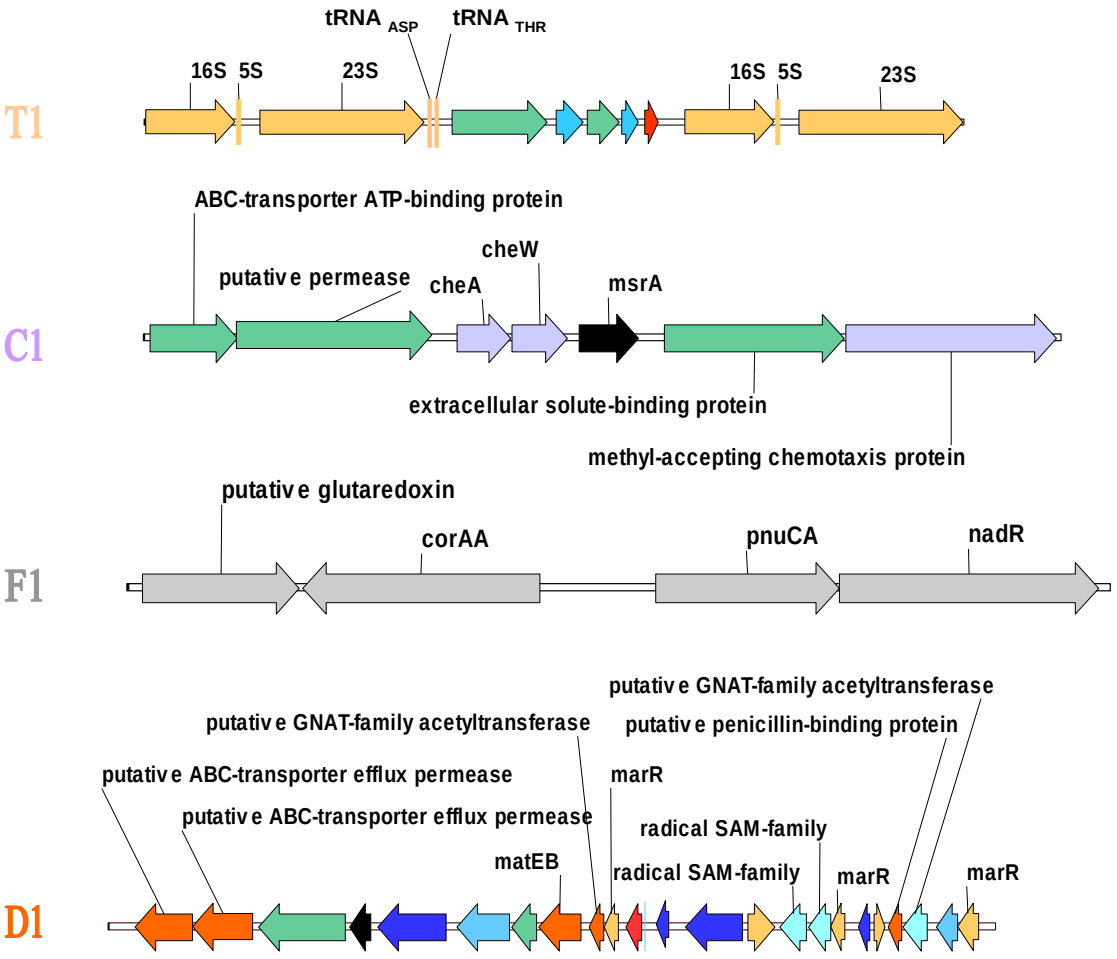


Fig.5.1. Predicted ORFs and encoded protein composition of BPc2 . Box plots show size, %G + C and isometric point distribution for each quartile of the data. Column charts show the distribution of each ORF from 1 – 255.


two component hybrid
parA
repB
AAA+ ATPase
hypA
Replication
Energy metabolism
E1
Cofactor metabolism
F1
E2
Energy metabolism
C1
T1

D1

E3


T1
tRNA ASP
tRNA THR
16S
5S
23S
16S
5S
23S
C1
ABC-transporter ATP-binding protein
putative permease
cheA
cheW
msrA
extracellular solute-binding protein
methyl-accepting chemotaxis protein
F1
putative glutaredoxin
corAA
pnuCA
nadR
D1
putative GNAT-family acetyltransferase
putative GNAT-family acetyltransferase
putative penicillin-binding protein
putative ABC-transporter efflux permease
marR
putative ABC-transporter efflux permease
radical SAM-family
matEB
radical SAM-family
marR
marR

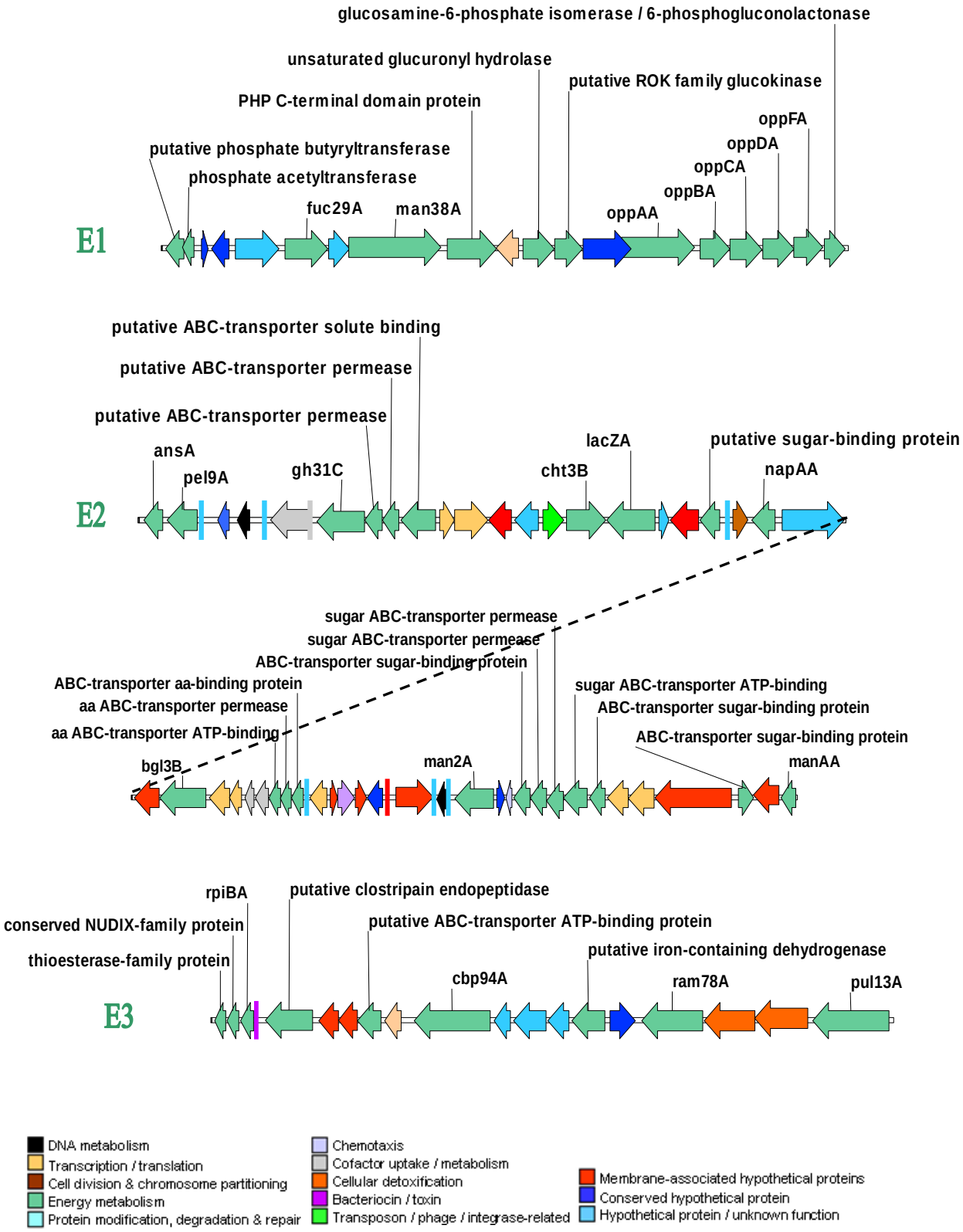


Fig. 5.2. *In-silico* map of BPc2. The main circular map shows, from outside to inside, the major functional clusters, the deviation from the average %G + C, and the third-position GC skew. The putative *oriR*-containing locus is enlarged above. The *dnaA* boxes (blue), the direct repeats (DR1, red; and DR2, green), inverse repeat (IVR1, purple) and inverted repeats (IR1, pink) are shown as well as the region aligning to the putative origins of the *B. fibrisolvens* plasmid pRJF1 and the co-resident pCY360 megaplasmid (red double ended arrow). Expanded views of each functional cluster are also shown, with ORFs colour coded by function according to the key.

5.4 Energy and Nitrogen metabolism

BPc2 encodes a large number of proteins that are predicted to contribute directly, or indirectly, to energy or nitrogen metabolism. This includes proteins involved in the uptake or breakdown of peptides, amino acids, simple sugars and an assortment of complex polysaccharides. It also includes proteins involved in the breakdown of nucleotides, and the reduction of fumarate and aldehydes (listed in Table 5.1 and many shown in Fig. 5.2). Eight of the proteins involved in energy metabolism are uniquely encoded within the *B. proteoclasticus* genome by BPc2. This includes: a phosphate acetyltransferase (ORF 11); a putative α-L-fucosidase, Fuc29A (ORF 15); a putative α-mannosidase, Man38A (ORF 17); a rhomboid-family serine peptidase (ORF 57); a fumarate reductase, FccA (ORF 64); an *L*-asparaginase, AnsA (ORF190); and a putative β-mannosidase, Man2A (ORF 237). These enzymes catalyse crucial steps in several different biological pathways including pyruvate and propionate metabolism (phosphate acetyltransferase); the entire (N-) glycan degradation pathway (Fuc29A, Man38A and Man2A, Fig. 5.3); anaerobic respiration (FccA); and nitrogen and aspartate metabolism (AnsA). BPc2 also encodes four ATP-binding cassette (ABC) uptake systems. One of these (ORFs 201-203) is unable to be assigned a specific substrate but is found in a cluster of genes dedicated to polysaccharide degradation. The other uptake systems are dedicated to the ATP-dependent transport of oligopeptides (ORFs 23-27), amino acids (ORFs 224-226) and sugars (ORFs 240-244 and 248). Each ABC-transport system varies in structure, but typically contains 1 or 2 permease components, 1 or 2 ATP-binding components and a substrate binding protein. The exceptions are the ABC transport system of unknown specificity that appears to lack an ATP-binding component, and the sugar ABC transport system that has three substrate-binding proteins. Additionally, several transcriptional regulators are found adjacent to these ABC-transport systems suggesting they contribute to the control of energy metabolism. A GntR-family transcriptional regulator (ORF 228) and a LytR-family two component system (ORF 220, sensor kinase and ORF 221, response regulator) are found just upstream and downstream, respectively, of the amino acid ABC-transport system. An AraC-family, two component system (ORFs 245, a sensor kinase and ORF 246, a response regulator) is found between two of the sugar-substrate binding proteins of the sugar ABC-transport system. Another AraC-

family, two component system (ORFs 204, a response regulator and ORF 205, a sensor kinase) also occurs just upstream of the ABC-transport system of unknown substrate-specificity (Fig. 5.2 E2, E2 and E1, respectively).

Table 5.1 BPc2-encoded proteins involved in energy and nitrogen metabolism

ORF	Function	EC number
	Nitrogen metabolism	
23	ABC-transporter oligopeptide-binding protein, OppAA	
24	Oligopeptide ABC transporter permease protein, OppBA	
25	Oligopeptide ABC transporter permease protein, OppCA	
26	Oligopeptide ABC transporter ATP-binding protein, OppDA	
27	Oligopeptide ABC transporter ATP-binding protein, OppFA	
55	Putative 5‟-nucleotidase	3.1.3.5
57	Rhomboid-family serine peptidase	3.4.21.105
71	Putative amino acid permease protein	
75	Putative amino acid permease protein	
91	Hypothetical NUDIX-family protein	
101	Putative ATP-dependent metalloprotease	
145	Conserved NUDIX-family protein	
148	Putative clostripain endopeptidase	3.4.22.8
190	*L*-asparaginase, AnsA*	3.5.1.1
224	Amino acid ABC-transporter ATP-binding protein	
225	Amino acid ABC-transporter permease protein	
226	ABC-transporter amino acid binding protein	
	Carbohydrate metabolism	
10	Phosphate butyryltransferase	2.3.1.19
11	Phosphate acetyltransferase	2.3.1.8
15	Putative α-L-fucosidase, Fuc29A	3.2.1.51
17	Putative α-mannosidase, Man38A	3.2.1.24
20	Unsaturated glucuronyl hydrolase, Ugl88A	3.2.1.88
21	Putative ROK-family glucokinase	2.7.1.2
28	glucosamine-6-phosphate isomerase	3.5.99.6
61	Oligosaccharide-specific β-xylosidase, Xyl39A	3.2.1.37
66	α-galactosidase, Aga36A	3.2.1.22
92	Putative sugar-phosphate isomerase	5.3.1.-
95	Putative endo-1,4-β-galactanase, Bgn53A	3.2.1.89
146	Ribose-5-phosphate isomerise, RpiB	5.3.1.6
153	Putative cellobiose phosphorylase, Cbp94A	2.4.1.20

159	Putative α-L-rhamnosidase, Ram78A	3.2.1.40
162	Pullulanase, Pul13A	3.2.1.41
166	Phosphoglycerate / bisphosphoglycerate mutase	5.4.2.-
191	Pectate lyase, Pel9A	4.2.2.2
200	Glycoside hydrolase-family 31 protein, Gh31C	3.2.1.-
209	Putative β-hexosaminidase, Bhx3B	3.2.1.52
210	β-galactosidase, LacZA	3.2.1.23
213	Putative sugar-binding protein	
219	β-glucosidase, Bgl3B	3.2.1.21
237	Putative β-mannosidase, man2A	3.2.1.25
240	ABC-transporter sugar-binding protein	
241	Sugar ABC-transporter permease protein	
242	Sugar ABC-transporter permease protein	
243	Sugar ABC-transporter ATP-binding protein	
244	ABC-transporter sugar-binding protein	
248	ABC-transporter sugar-binding protein	
250	Mannose-6-phosphate isomerise, ManA	5.3.1.8
	Other	
64	Fumarate reductase, FccA	1.3.99.1
144	Thioesterase-family protein	
157	Iron-containing alcohol dehydrogenase	1.1.1.1
201	ABC transporter permease protein	
202	ABC transporter permease protein	
203	Solute-binding protein	

*Proteins uniquely encoded by BPc2 are shown in red.

5.5 Putative detoxification functions encoded by BPc2

BPc2 encodes nine proteins that are annotated as being involved in the efflux or degradation of an assortment of toxic compounds, including the terpenoid camphor, β-lactam antibiotics and heavy metals (Table 5.2). Two ABC-type efflux permeases and three multidrug and toxin extrusion proteins (MatEA, MatEB and MatEC) are also encoded. Two of the MatE proteins are found either side of, and close to, the replication origin while many of the remaining detoxification proteins reside in a 28 Kb locus (Fig. 5.2 D1). This locus also encodes two putative GNAT-family acetyltransferases, one of which (ORF 168) is very similar to a phosphinothricin *N*-acetyltransferase from *Bacillus halodurans* (50% amino acid identity, E= $2e^{-39}$). The locus also contains three multiple-antibiotic resistance (MarR)-family transcriptional regulators, one of which is found directly upstream of the ORFs encoding the putative phosphinothricin *N*-acetyltransferase and MatEB.

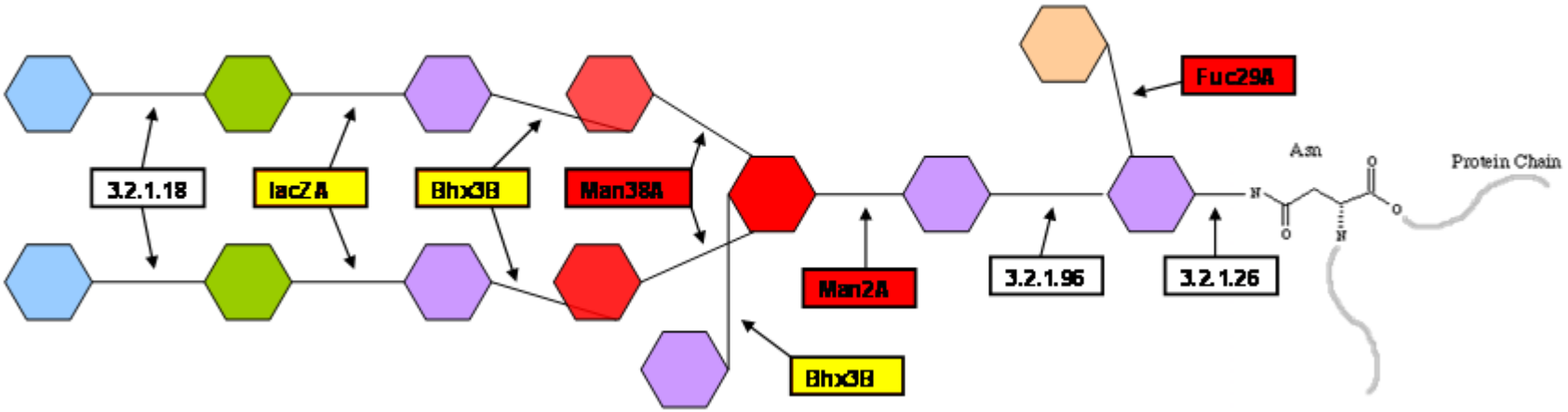


Figure 5.3. *N*-Glycan degradation capacity of *B. proteoclasticus*. A typical N-glycan structure is shown along with the enzymes involved in its breakdown. Monosaccharide subunits neuraminic acid (blue), galactose (green), N-acetyl galactosamine (purple), mannose (red) and fucose (tan) are shown. Those enzymes that are uniquely encoded by BPc2 are shown in red, those also encoded on the major chromosome are shown in yellow, and those not found in the *B. proteoclasticus* genome are not shaded.

Table 5.2 BPc2-encoded proteins involved in detoxification

ORF	Function	EC number
9	Multidrug and toxin extrusion protein, MatEA	
29	Putative β-lactamase	3.5.2.6
53	Heavy metal translocating P-type ATPase	
138	CrcB protein*	
160	ABC-transporter efflux permease	
161	ABC-transporter efflux permease	
167	Multidrug and toxin extrusion protein, MatEB	
168	Putative GNAT-family acetyltransferase	2.3.1.-
181	Putative penicillin-binding protein	
251	Multidrug and toxin extrusion protein, MatEC	

*Proteins uniquely encoded by BPc2 are shown in red.

5.6 Chemotaxis and flagellar formation

Despite being non-motile, *B. proteoclasticus* has a complete set of flagellar and chemotaxis-related genes. BPc2 has five proteins implicated in flagellar formation and chemotaxis including a putative flagellar protein, the chemotaxis proteins CheA, CheW and CheY, and the methyl-accepting chemotaxis protein, McpD (Table 5.3). Most of the chemotaxis-related genes are found in a small cluster spanning less than 5 Kb (Fig. 5.2 C1), while *cheY* falls outside this region. *cheY* is found immediately upstream of the sugar ABC-transport system (described above; seen in Fig. 5.2 E2). A putative permease and an extracellular solute-binding protein, both of unknown specificity, are located adjacent to the 5 Kb chemotaxis cluster. BPc2 also encodes a single protein (ORF 60) putatively involved in flagellar formation situated directly downstream of the second rRNA operon. No chemotaxis or flagellar formation proteins appear to be uniquely encoded by BPc2.

Table 5.3 BPc2-encoded proteins involved in chemotaxis

ORF	Function
47	Chemotaxis protein, CheA
48	Chemotaxis protein, CheW
51	Methyl-accepting chemotaxis protein, McpD
60	Putative flagellar protein
239	Chemotaxis protein, CheY

5.7 Cofactor and vitamin uptake and metabolism

BPc2 encodes nine proteins implicated in the uptake or metabolism of several important cofactors, including magnesium, cobalt, Fe^{2+}, nicotinamide adenine di-nucleotide (NAD), coenzyme A, and a glutaredoxin, along with the vitamins biotin (Vitamin H or B_7) and pantothenate (Vitamin B_5) (Table 5.4). BPc2 also contains a 6,7-dimethyl-8-ribityllumazine synthase pseudogene. This truncated gene (51 amino acids) shows significant sequence identity (63%) to the C-terminus of a characterised 6,7-dimethyl-8-ribityllumazine synthase from *Bacillus subtilis* (Fischer *et al*., 2003). Six of the proteins implicated in the uptake or metabolism of these cofactors appear to be uniquely encoded within the *B. proteoclasticus* genome by the BPc2 replicon, including a putative glutaredoxin, a putative magnesium and cobalt transport protein (CorAA), a nicotinamide riboside transporter (PnuC), a nicotinamide riboside kinase, an adenylyltransferase (NadR), an acyl-carrier protein phosphodiesterase and a biotin biosynthesis protein (BioY). While *B. proteoclasticus* encodes alternate systems for the acquisition of cobalt and magnesium, the biosynthesis of both NAD and biotin appears to be dependent on the BPc2-encoded proteins PnuC, NadR and BioY. Analysis of the *B. proteoclasticus* genome reveals it lacks homologues to the essential NAD-biosynthetic enzymes quinolinate synthase (NadA), nicotinate-nucleotide pyrophosphorylase (NadC) and the nicotinamidase, PncA, along with the biotin biosynthetic enzymes 8-amino-7-oxopelargonate synthase (BioF), 7,8-diaminopelargonic acid aminotransferase (BioA), dethiobiotin synthetase (BioD) and biotin synthase (BioB).

Table 5.4 BPc2-encoded proteins involved in cofactor and vitamin metabolism

ORF	Function	EC number
34	Putative glutaredoxin*	
35	Putative magnesium and cobalt transport protein, CorA	
36	Nicotinamide mononucleotide transporter, PnuC	
37	Nicotinamide-nucleotide adenylyltransferase, NadR	2.7.7.1
54	Acyl-carrier protein phosphodiesterase	1.7.-.-
69	*6,7-dimethyl-8-ribityllumazine synthase pseudogene*	
130	Putative flavodoxin	
198	Ferrous iron transport protein, FeoB	
199	Putative ferrous iron transport protein, FeoA	
222	Biotin transport protein, BioY	
223	Biotin acetyl-CoA-carboxylase-ligase, BirA	6.3.4.15

*Proteins uniquely encoded by BPc2 are shown in red.

5.8 Ribosomal RNAs and Transfer RNAs

BPc2 encodes two complete ribosomal RNA (rRNA) operons giving *B. proteoclasticus* a total of six rRNA operons (Fig. 5.2 T1). All six rRNA operons are identical in structure (16*S* – 5*S* – 23*S*) and share greater than 99% nucleotide sequence identity. BPc2 also encodes an aspartic acid (Asp-GTC) and a threonine (Thr-TGT) tRNA. An orthologue of each of these tRNAs is found on the major chromosome.

5.9 BPc2 contains genes described in the minimal gene set

BPc2 has 13 genes described in the Bacterial Minimal Gene Set (Koonin, 2000) listed in Table 5.5. Two of these genes, a putative acyl-carrier protein phosphodiesterase and a peptide methionine sulfoxide reductase, are not found elsewhere in the *B. proteoclasticus* genome. Aside from the unique enzymes described above, BPc2 also encodes a unique copy of the RNA 2'-phosphotransferase enzyme, KptA.

5.10 Transposases

Four transposase genes were identified within the BPc2 sequence, and each can be assigned to one of three transposase families; IS200 (ORF 89), IS605 (ORF 80 and ORF 127) and IS110 (ORF 94). The IS605-family transposases show 94% amino acid identity to each other, and 29 to 32% identity to transposases found in pCY360 (Section 4.8). The IS110 and IS200-family transposases show 100% identity at both the amino acid and nucleotide level to four IS110 and two IS200 transposases, respectively, upon the major chromosome. The IS110-family transposase also shows 100% identity to a transposase found on pCY186. Several genes indicative of a retro-transposon were identified within a 5.3 Kb locus. These include an RNA-dependent DNA polymerase (ORF 114), a putative integrase (ORF 116) and an IstB-family ATP-binding protein (ORF118). The intervening ORFs (ORFs 115 and 117) are hypothetical proteins.

Table 5.5. Genes of the Bacterial Minimal Gene Set found on BPc2.

ORF	Size (aa)	Putative function	Source of Best Blast match	E-value	% ID	GenBank Accession
24	344	Oligopeptide ABC-transporter, permease	*Paenibacillus sp.*	8 e^{-72}	43%	ZP_02849269
25	372	Oligopeptide ABC-transporter, permease	*Paenibacillus sp.*	1 e^{-111}	56%	ZP_02849268
26	363	Oligopeptide ABC-transporter, ATP-binding protein	*Paenibacillus sp.*	1 e^{-112}	59%	ZP_02849267
27	333	Oligopeptide ABC-transporter, ATP-binding protein	*Thermotoga maritima*	1 e^{-100}	55%	NP_227873
31	238	Anaerobic ribonucleotide triphosphate reductase activating protein, NrdGB	*Ruminococcus gnavus*	1 e^{-71}	54%	ZP_02041952
32	434	Putative anaerobic ribonucleotide triphosphate reductase, NrdDB	*Clostridium bolteae*	1 e^{-168}	73%	ZP_02089280
33	318	Anaerobic ribonucleotide triphosphate reductase, NrdDC	*Clostridium scindens*	1 e^{-110}	68%	ZP_02431419
49	61	Peptide methionine sulfoxide reductase, MsrAA *	*Clostridium thermocellum*	1 e^{-34}	47%	YP_001039379
54	154	Acyl carrier protein phosphodiesterase	*Clostridium acetobutylicum*	2 e^{-20}	38%	NP_350011
131	237	HAD-superfamily hydrolase	*Dorea formicigenerans*	1 e^{-42}	58%	ZP_02234927
163	216	Exodeoxyribonuclease III	*Streptococcus sanguinis*	1 e^{-89}	71%	YP_001035639
166	260	Phosphoglycerate / bisphosphoglycerate mutase	*Clostridium phytofermentans*	6 e^{-43}	38%	YP_001560430
196	212	Putative S-adenosylmethionine-dependent methyltransferase	*Dorea formicigenerans*	1 e^{-69}	60%	ZP_02234927

*Genes of the Bacterial Minimal Gene Set unique to BPc2 are shown in red.

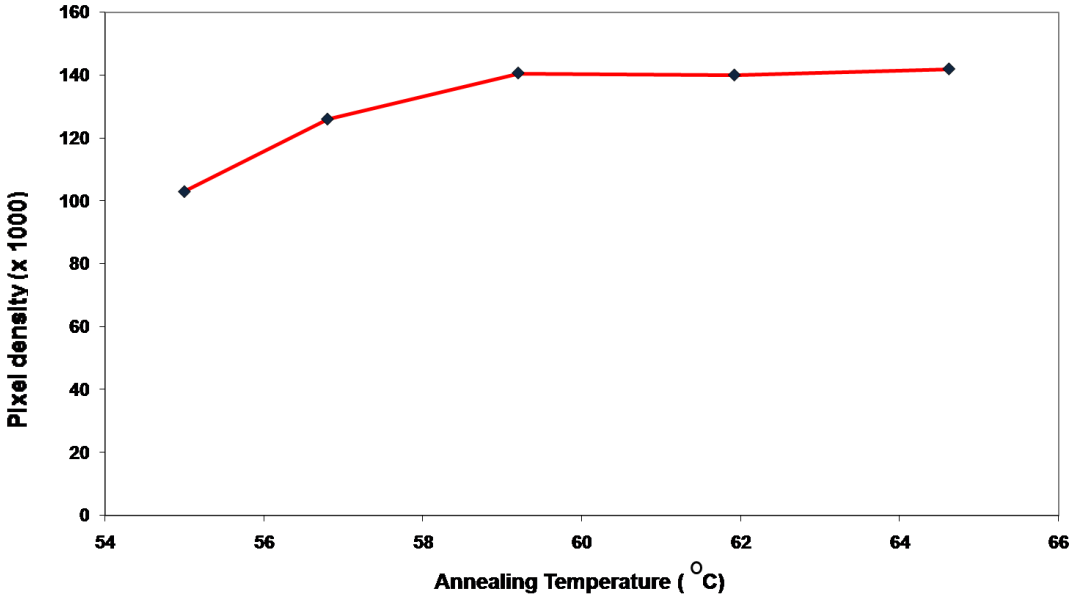
5.11 Maintenance of BPc2 under curing conditions

As previously described (Section 4.11), *B. proteoclasticus* was grown at 45 °C or treated with sub-lethal levels of novobiocin, acriflavine, ethidium bromide, or acridine orange and a total of 5,688 colonies were transferred to nylon membranes to screen for the loss of BPc2 by colony hybridisation. The probe used was a 762 bp, largely intergenic, region of BPc2 amplified by PCR using the p300_probe2 primer set. All colonies examined were detected by the probe, indicating that none of the *B. proteoclasticus* colonies screened were cured of the BPc2 replicon.

5.12 Copy number of BPc2

The copy number of BPc2 was derived by comparison of its absolute copy number to that of the major chromosome (as described in Section 4.12), using qPCR. The T_m of the BPc2 primer set, (p300_probe2) was optimised by conventional PCR (as described in Section 4.12.1). The amount of detected product was nearly identical between the annealing temperatures of 59 to 65 °C (Fig. 5.4a). An annealing temperature of 60 °C was selected as this matched the annealing temperature used for qPCR of the major chromosome. The optimum $MgCl_2$ concentration, (Section 4.12.1) at an annealing Tm of 60 ºC, was determined to be 3 mM (Fig. 5.4b). The specificity of the qPCR reaction was determined by melting curve analysis and gel-electrophoresis. Melting curve analysis revealed a single peak for the p300_probe2 reaction occurring at 85.55 ºC. Subsequent gel-electrophoresis showed a band of the expected size (762 bp, Fig 5.5). The standard curve of the Threshold Cycle (C_T) vs. the $\log_{10}$ of the replicon copy numbers for the Topo-recombinant plasmid carrying the p300_probe2 sequence (Fig 5.6a) had a slope of -3.357. This equates to amplification efficiency for the reaction of 0.99. The standard curve was linear in the range tested ($1 \times 10^5 - 1 \times 10^9$ copies / µl; $R^2 > 0.99$). A standard curve of the C_T vs. the $\log_{10}$ of the fold dilution for the target (Fig. 5.6b) was also determined to ensure they were approximately equal. The slope was 3.409 for the p300p2 target, which equates to an amplification efficiency of 0.96 for the reaction. The standard curve was linear in the range tested ($1 \times 10^{-1} - 1 \times 10^{-5}$ DNA extract / reaction; $R^2 > 0.99$). Quantitative analysis of BPc2 DNA revealed the replicon to be present in $1.24 \times 10^{10} \pm 4.24 \times 10^9$ copies in the tested sample (99% confidence) per $1.3 \times 10^{10} \pm 1.75 \times 10^9$ copies of the chromosome (99% confidence). This is equivalent to a BPc2 copy number of 0.95 ± 0.37 copies per chromosome (99% confidence).

A



B

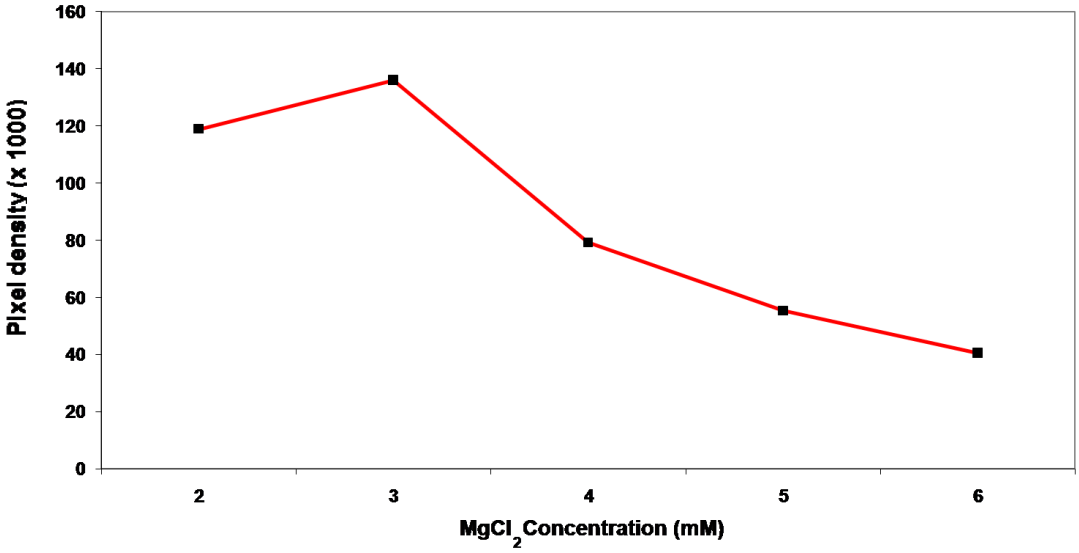


Figure 5.4. Optimisation of qPCR reactions. Quantitative PCR reactions were optimised for temperature (A) and $MgCl_2$ concentration (B). Optimal conditions were determined as those producing the greatest amount of the desired qPCR product which was determined by measuring the sum of the pixel density from the region of interest (ROI) using KODAK 1D Image analysis software.

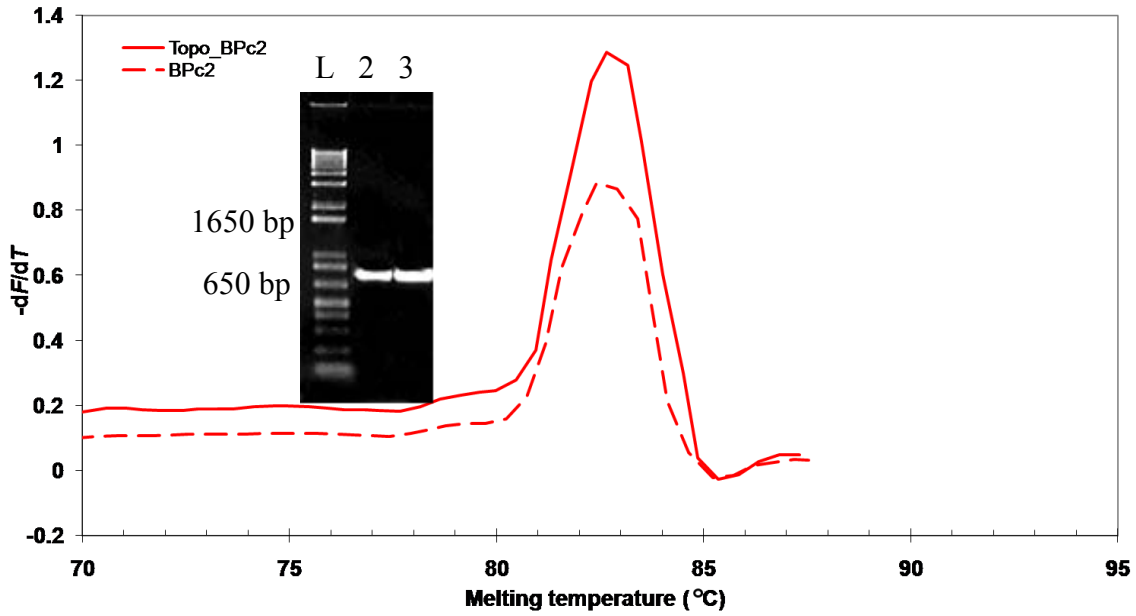


Figure 5.5. Confirmation of qPCR amplification specificities of BPc2 primer set. Melting peaks were examined for each set using as a template either the *B. proteoclasticus* total DNA preparation (broken line) or the quantified Topo-recombinant plasmids carrying the target sequence (solid line). Inset: Agarose gel-electrophoresis of the qPCR products. Lanes: L, the 1Kb+ ladder, lanes 2 and 3, 762 bp PCR-products obtained using as templates either (2) quantified Topo-recombinant plasmids carrying the target sequence or (3) the total DNA preparation.

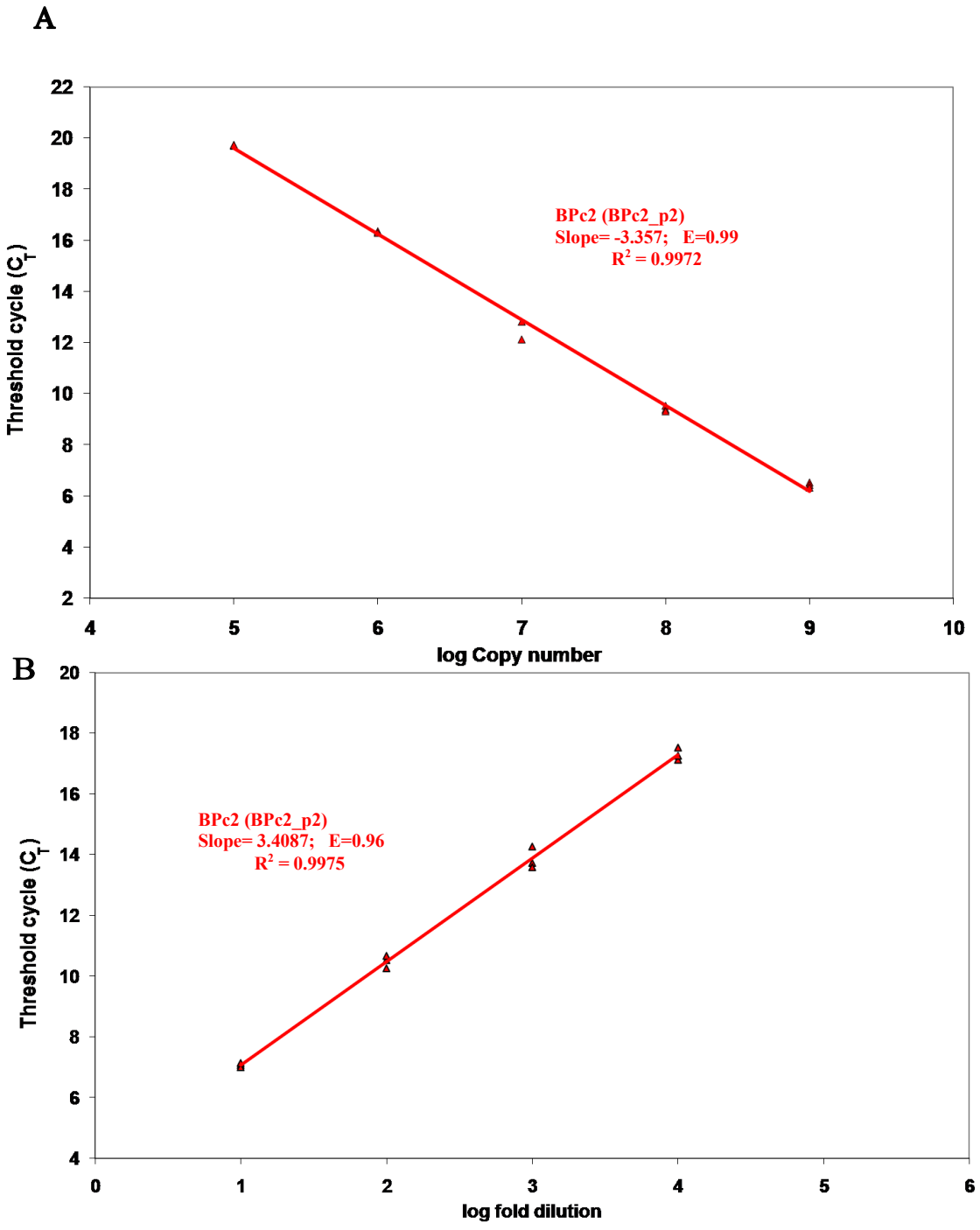


Figure 5.6. Standard curves of qPCR reactions. The slope of each Threshold cycle vs. $\log_{10}$ plot of the copy number in each standard (A) or Threshold Cycle vs. Log_{10} fold dilution (B) is shown. The slope, efficiency (E), and correlation (R^2) are indicated next to their respective slopes.

5.13 Microarray analysis of BPc2 gene expression in mono-culture versus in co-culture with *M. ruminantium*

BPc2 encodes a significant number of proteins annotated as having functions that would affect the cell membrane or extracellular environment and consequently BPc2 may have a role in interspecies interactions. BPc2 encodes 50 membrane-spanning proteins, 44 secreted proteins, four di-guanylate cyclase proteins (ORFs 7, 8, 65 and 249), an EAL-domain protein (207), a putative sortase B protein (ORF 247), a putative band 7 protein (ORF 14) and a conserved hypothetical protein that contains a cupin domain (ORF 178). BPc2 also possesses four two-component sensor kinase / response regulator systems (ORFs 204 and 205, 220 and 221, 245 and 246, and 253). All of these two-component systems have a predicted membrane spanning module in the ORF encoding the histidine kinase component. Analysis of the microarray data from the *B. proteoclasticus* mono-culture versus the *B. proteoclasticus-M. ruminantium* co-culture, (Section 4.13), revealed 95% (224 of the 235 putative BPc2 ORFs, identified upon the BPc2 replicon at the time the microarray slides were printed,18 ORFs were not included in the microarray analysis), were detected at a significant pixel intensity (9,000 > 65,000). Of the genes analysed 26 ORFs (10%) were found to be up-regulated greater than 2-fold with a false discovery rate (FDR) less than 0.05 in the co-culture condition, while a further 12 (5%) were down-regulated greater than 2-fold (FDR <0.05) (Table 5.6). Several genes, whose encoded proteins could potentially influence the membrane or extracellular environment, were found to be significantly up-regulated including six membrane-spanning proteins (ORFs 44, 46, 59, 60, 76 and 234), four secreted proteins (ORFs 154, 156, 217 and 226), two di-guanylate cyclases (ORFs 7 and 8), and the band 7 protein (ORF 14). The putative flagellar protein (ORF 60) is up-regulated during co-culture, as are many of the flagellar proteins located on the main chromosome, as discussed previously. The most significantly up-regulated gene are a putative ABC-transporter permease protein (ORF 46, 11.5 fold) and it‟s associated ATP-binding protein (ORF 45, 7.5 fold). A hypothetical membrane spanning-protein, contiguous with these genes (ORF 44), is also significantly up-regulated (3.5 fold). A putative extracellular solute-binding protein found downstream of these ORFs was not found to be differentially regulated (ORF 50; 1.01 fold). The substrate specificity of this up-regulated ABC-transport system is not clear. Another ABC-transport system, with amino acid-specificity, is also up-regulated. The extracellular amino acid-specific binding protein

encoded by ORF 226 is up-regulated 2.88 fold, while its ATP-binding protein component (ORF 224) falls narrowly below the significance threshold with 1.97 fold up-regulation. Many of the genes that were down-regulated are involved in carbohydrate metabolism, including an ABC-type sugar transport system (ORFs 240, 241, 242 and 244), Fuc29A (ORF 15), Ram78A (ORF 159) and Pel9A (ORF 191). Analysis of the distribution of differential gene regulation around the replicon reveals four clusters of genes that appear to be co-regulated (Fig. 5.7). The ABC-type transport systems with sugar (down-regulated) or unknown (up-regulated) specificity described above, are seen, along with a cluster of ORFs consisting of a hypothetical protein upstream of two GNAT-family acetyltransferases (ORFs 123 to 124, down-regulated), and another cluster consisting of a hypothetical protein that resides between two hypothetical secreted protein (ORFs 154 to 156, up-regulated). The extent of up- or down- regulation of ORFs within each cluster appears to differ markedly suggesting they are not likely to be polycistronic.

Table 5.6. Differential regulation of BPc2 ORFs during co-culture

ORF	Putative function	Fold change	FDR
	Up-regulated in co-culture		
46	Putative ABC-transporter permease protein	11.48	3.2×10^{-4}
45	ABC-transporter ATP-binding protein	7.51	1.2×10^{-3}
156	Hypothetical secreted protein	6.80	5.2×10^{-4}
14	Putative band-7 protein	6.09	1.8×10^{-3}
234	Hypothetical membrane protein	5.09	3.5×10^{-4}
32	Putative anaerobic ribonucleotide-triphosphate reductase	4.53	1.7×10^{-3}
155	Hypothetical protein	3.73	1.9×10^{-3}
44	Conserved transmembrane protein	3.48	1.9×10^{-2}
7	Di-guanylate cyclase, GGDEF protein	3.26	4.3×10^{-3}
136	HicB protein	3.25	5.2×10^{-4}
40	Hypothetical protein containing UBA/TS-N domain	2.93	2×10^{-2}
8	Di-guanylate cyclase, GGDEF protein	2.88	1.3×10^{-3}
226	ABC-transporter amino acid binding protein	2.88	3.0×10^{-3}
59	Conserved transmembrane protein	2.84	3.1×10^{-2}
154	Hypothetical secreted protein	2.58	2.4×10^{-2}
217	Hypothetical secreted protein	2.47	8.6×10^{-4}
60	Putative flagellar protein	2.42	1.1×10^{-3}
33	Anaerobic ribonucleotide-triphosphate reductase, NrdDC	2.41	1.5×10^{-2}
29	Putative β-lactamase	2.31	7.6×10^{-4}
252	DNA-binding protein, HU	2.28	6.3×10^{-3}
135	HicA protein	2.26	1.9×10^{-3}
38	Hypothetical protein	2.12	3.0×10^{-3}
76	Hypothetical transmembrane protein	2.07	4.4×10^{-3}
2	Putative AAA+ ATPase	2.07	1.4×10^{-2}
5	Conserved hypothetical protein	2.03	2.2×10^{-3}

Down-regulated in co-culture

240	ABC-transporter sugar-binding protein	8.28	6.3×10^{-4}
241	Sugar ABC-transporter permease protein	4.95	3.2×10^{-4}
122	Hypothetical protein	3.75	3.1×10^{-3}
244	ABC-transporter sugar-binding protein	3.32	1.4×10^{-2}
242	Sugar ABC-transporter permease protein	2.83	4.0×10^{-3}
159	Putative α-L-rhamnosidase, Ram78A	2.78	7.0×10^{-3}
124	Putative GNAT-family acetyltransferase	2.74	6.4×10^{-4}
52	Conserved hypothetical protein	2.63	1.7×10^{-2}
89	IS200-family transposase, TnpCB	2.54	9.7×10^{-3}
191	Pectate lyase, Pel9A	2.38	1.4×10^{-4}
54	Acyl-carrier protein phosphodiesterase	2.25	8.5×10^{-3}
15	α-L-fucosidase, Fuc29A	2.21	4.4×10^{-3}

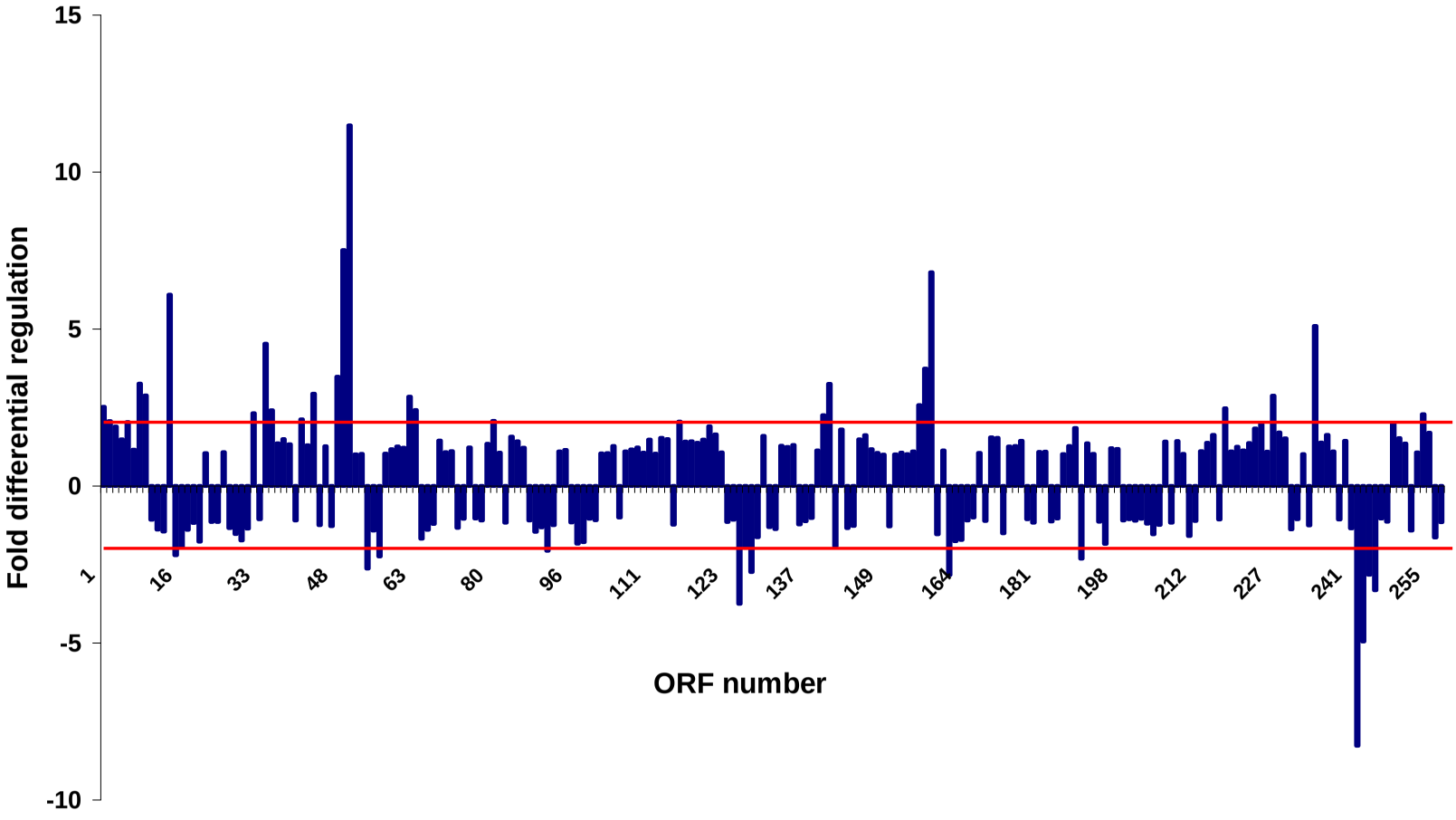


Figure 5.7. Differential regulation of BPc2-located genes during co-culture of *B.. proteoclasticus* with *M. ruminantium*. The differential regulation of gene transcripts, for ORFs of BPc2 as determined by microarray analysis. Red bars show the significance threshold (a 2-fold expression difference).

5.14 Distribution of rRNA operons in the *Butyrivibrio* assemblage

The 10 strains from the *Butyrivibrio* / *Pseudobutyrivibrio* assemblage that were found to have multi-replicon genomes were analysed by Southern blot analysis using a 209 bp amplicon corresponding to nucleotides 510 – 719 of the 16*S* rRNA gene from *B. proteoclasticus*. The resulting image revealed no hybridisation to any of the megaplasmid-like replicons other than BPc2 (Fig. 5.8).

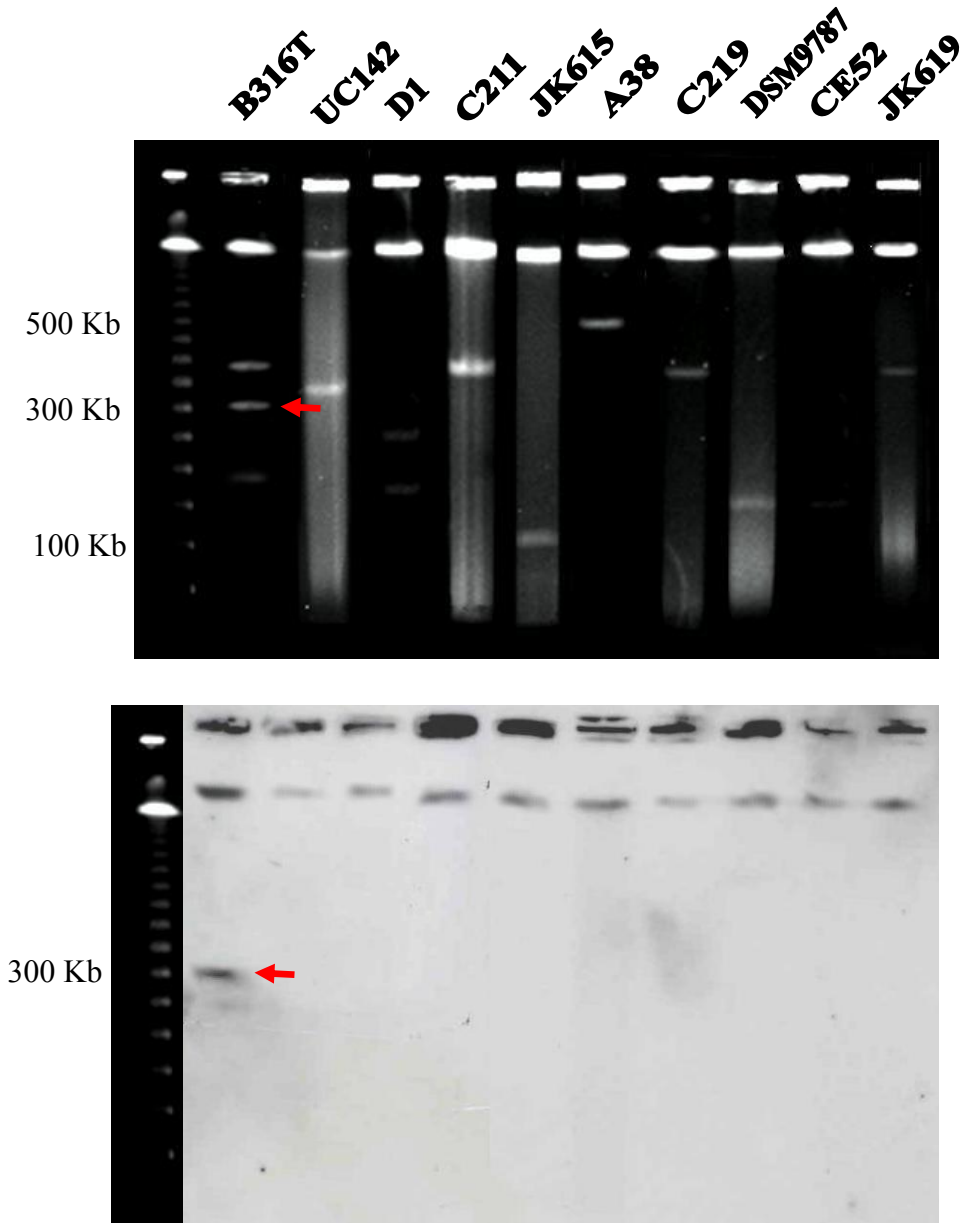


Fig. 5.8. Distribution of rRNA operons among megaplasmid-like replicons in *Butyrivibrio* / *Pseudobutyrivibrio* spp. (a) PFGE images of uncut whole genomic DNA extracts from strains of the *Butyrivibrio* / *Pseudobutyrivibrio* assemblage found to possess megaplasmid-like replicons. These include from left to right *B. proteoclasticus* strains B316^{T} and UC142, *B. fibrisolvens* strains D1 and C211, *B. hungatei* strains JK615, A38 and C219, *P. ruminis* strain DSM9787, *P. xylanivorans* strain Ce52 and *Butyrivibrio* sp. JK619 (B) Southern blots of the above gel was probed using an amplicon derived from the BPc2 16*S* rRNA gene.

5.15 Discussion

5.15.1 BPc2 replication

The locus identified as most likely to contain the BPc2 *oriR* encompasses ORFs encoding the RepB and ParA proteins. Unlike pCY360, where the *oriR* is predicted to reside upstream of both *repB* and *parA*, the BPc2 *oriR* is predicted to reside in a large intergenic region between these two divergently transcribed ORFs. This intergenic region contains a large number of elements, including direct (DR1 and DR2), inverted (IR1) and inverse (IVR1) repeats, that are capable of forming various secondary structures along with eleven potential *dnaA* boxes. The large size of this intergenic region, coupled with the divergent nature of the ORFs surrounding it initially suggested it likely contained the BPc2 *oriR*. The secondary structures, *dnaA* box candidates and the AT-rich nature of the region further support this assertion. The region contains a 366 bp sequence that exhibits significant nucleotide identity to the putative *oriR* of both *B. fibrisolvens* plasmid, pRJF1 and the co-resident pCY360. This, along with the significant amino acid identity between the BPc2 and pCY360 RepB protein, suggests BPc2 also replicates via a theta mode (del Solar *et al*., 1998). However, third-position GC skew analysis, and the even distribution of ORFs on both DNA strands suggests, that BPc2 utilises a bidirectional replication mode.

5.15.2 BPc2 is a secondary chromosome

Several features of the BPc2 replicon distinguish it from the two megaplasmids, pCY360 and pCY186 (described in Chapters 4 and 6 respectively). Firstly, BPc2 encodes two complete ribosomal RNA operons and two transfer RNAs. Excluding the major chromosomes of bacteria, relatively few replicons have been described as possessing non-coding translational machinery. A number of examples are found in Gram-negative bacteria, almost exclusively from the α, β and γ classes of the Proteobacterial phyla (Chain *et al*., 2006; Pohlmann *et al*., 2006; Suwanto and Kaplan, 1989). The only other Gram-positive example is *D. radiodurans*, which encodes a secondary chromosome that lacks an rRNA operon but possesses a single methionine tRNA (White *et al*., 1999). In most instances, replicons possessing rRNA operons are designated as secondary chromosomes, although these replicons often

possess chromosomal characteristics other than the translational machinery (see Choudhary *et al*., 1997 and Komatsu *et al*., 2003 for examples).

Secondly, the vast majority of the ORFs assigned to BPc2 show sufficient amino acid sequence similarity to proteins of known or predicted function, enabling a putative function to be assigned. The proportion of ORFs unable to be assigned a function is almost identical to that seen on major chromosome.

Thirdly, BPc2 exists as a single copy per chromosome. This is consistent with the observations of Egan *et al*. (2004) who found the replication of the *Vibrio cholerae* secondary chromosome was co-ordinated with that of the major chromosome, and consequently to the cell-cycle. Thus secondary chromosome replication appears distinct from the replication of megaplasmids which are believed to replicate autonomously and independent from the chromosome (del Solar *et al*., 1998).

Fourthly, unlike pCY186, BPc2 is unable to be cured from *B. proteoclasticus*, suggesting a function or functions encoded by BPc2 are indispensible. Also unlike pCY360, no TA operon is evident in BPc2.

Several BPc2 genes may be essential for *B. proteoclasticus* viability including two genes described in the Bacterial Minimal Gene Set (*acpD* and the peptide methionine sulfoxide reductase genes), genes essential for the uptake and biosynthesis of NAD (*pnuC* and *nadR*), or biotin (*bioY*), the only copy of fumarate reductase and a raft of enzymes that enhance the metabolic capabilities of *B. proteoclasticus*. Collectively these observations satisfy all current definitions of a secondary chromosome (Bartosik *et al.*, 2002, Ng *et al.*, 1998, Jumas-Bilak *et al.*, 1998, Ochman, 2002) with the exception of size (Schwartz & Friedrich, 2001). No other auxiliary replicon of the *Butyrivibrio* assemblage was found to possess a 16*S* rRNA suggesting that this feature is unique to *B. proteoclasticus* B316^{T}.

5.15.3 The role of BPc2 in energy metabolism

Six clusters of BPc2 ORFs are dedicated to energy metabolism, chemotaxis, cellular detoxification and cofactor uptake and metabolism. Energy metabolism appears to be the predominant role of BPc2, with 21% of its ORFs having roles in the reduction,

oxidation, uptake or breakdown of either peptides, amino acids, complex polysaccharides, simple sugars, nucleotides, fumarate or aldehydes. These ORFs form three distinct clusters, although each cluster does not conform to a single type of energy metabolism. Other megaplasmid-sized replicons of Gram-positive bacteria have previously been described as having significant roles in energy metabolism including pREL1 and pREC1 of *Rhodococcus erythropolis* PR4 (alcohol and fatty acid metabolism, respectively; Sekine *et al*., 2006).

The breakdown of glycosidic bonds within polysaccharides is mediated by enzymes known as glycoside hydrolases (GHs). There are numerous types of GHs, necessitated by the extensive variety in stereochemistry of carbohydrates (Geremia *et al.*, 1996, Henrissat, 1991, Henrissat & Bairoch, 1993). They are grouped into approximately 100 families based on amino acid sequence identity (CAZY website, www.cazy.org). BPc2 encodes 13 GH enzymes each from a different GH-family. Three of these GH activities, Fuc29A, Man38A, and Man2A, are not encoded elsewhere in the *B. proteoclasticus* genome. Fuc29A, a GH29-family enzyme, is found in plants, vertebrates and bacteria, and particularly bacteria that colonise a mammalian host. It is an exoglycosidase capable of cleaving α-linked *L*-fucose residues from glycol-conjugates, in which the most common linkages are to galactose or *N*-acetylglucosamine residues. Man38A, a GH38-family enzyme, hydrolyses terminal, non-reducing α-*D*-mannose residues in α-*D*-mannosides. Man2A, is a GH2-family enzyme, which hydrolyses β-mannose-1-4-*N*-acetylglucosamine linkages releasing mannose-oligosaccharides and chitobiose. All of these enzymes, along with LacZA and Bhx3B (homologues of which are also encoded on the major chromosome) are required for the catabolism of *N*-glycans released during the degradation of glycoproteins. In organisms that inhabit the mammalian intestinal tract it is speculated that the presence of an *N*-glycan-degrading pathway may allow the bacterium to degrade glycosylated proteins secreted from the epithelium (Zwierz *et al.*, 1981, Ruas-Madiedo *et al.*, 2008, Argueso & Gipson, 2006). While *B. proteoclasticus* may be capable of performing a similar role in the gastrointestinal tract of the ruminant host, the rumen itself is a non-secreting organ. Therefore it is more likely, in this environment, that these enzymes are involved in the degradation of plant polysaccharides where mannose and fucose are abundant (de La Torre *et al*., 2002). *B.*

proteoclasticus also possesses fucose and mannose up-take systems and it is believed the organism may import and utilise these sugars in exopolysaccharide production, as previously observed in other bacterial species (Sutherland & Wilkinson, 1968, Bramhachari *et al.*, 2007, Kosenko & Mal'tseva, 1984). Therefore Fuc29A, Man38A and Man2A may be involved in providing and recycling exopolysaccharide material.

Aside from Pul13A and Bgl3B, all BPc2-encoded GH enzymes lack secretory signals or transmembrane helices, suggesting they act within the cell. *B. proteoclasticus* encodes more than 20 ABC-transport systems identified as being involved in the uptake of sugars or polysaccharides. ABC-transport systems couple ATP-hydrolysis to the importation (or export) of a specific substrate. Substrates are recognised by the substrate binding-protein, which acts on the outside of the cell to capture the substrate with high affinity and deliver it to the cognate transport system. BPc2 encodes a single sugar ABC-transporter system, but encompasses three substrate-binding proteins which is unusual and may indicate the system is capable of transporting multiple sugars. Once simple six-carbon sugars are transported into the cell, they are phosphorylated to prevent their escape, by hexokinase-like enzymes. BPc2 encodes a ROK-family glucokinase, which catalyses the phosphorylation of glucose to give glucose-6-phosphate. Glucose-6-phosphate can be further metabolised by the pentose-phosphate or glycolytic pathways. The pentose-phosphate pathway provides reducing power in the form of reduced nicotinamide adenine di-nucleotide phosphate (NADPH) and ribose-5-phosphate, an intermediate for the formation of amino acids, vitamins, nucleotides, and cell wall constituents (Sprenger, 1995). BPc2 participates in the pentose-phosphate pathway by contributing one of three RpiB proteins encoded by *B. proteoclasticus*. RpiB reversibly catalyses the conversion of ribose-5-phosphate to ribulose-5-phosphate, an essential reaction in the oxidative phase of the pentose-phosphate pathway. The forward reaction of RpiB additionally provides intermediates for glycolysis. The reverse reaction provides the final step in the transformation of glucose 6-phosphate to ribose 5-phosphate, which is essential to the synthesis of nucleotides and several cofactors (Zhang *et al*., 2003).

Glucose-6-phosphate molecules can alternatively be catabolised, via glycolysis, to pyruvate, producing ATP either directly or through anaerobic respiration (Stryer, 1995; Unden and Bongaerts, 1997). BPc2 contributes directly to glycolysis by

encoding one of eight phosphoglycerate mutase enzymes produced by *B. proteoclasticus*. This enzyme catalyses step eight of the glycolytic pathway converting 3-phosphoglycerate to 2-phosphoglycerate. The enzymes ManA and glucosamine-6-phosphate isomerase, encoded by BPc2, enable carbon-6-phosphorylated intermediates of mannose and glucosamine monosaccharides, respectively, to enter the glycolytic pathway. ManA and the glucosamine-6-phosphate isomerase convert mannose-6-phosphate or glucosamine-6-phosphate, respectively, to fructose-6-phosphate, the second intermediate in the glycolysis pathway. A third sugar-phosphate isomerase is encoded by BPc2, however, its sugar-specificity is unclear. Another BPc2-encoded enzyme, Cbp94A, degrades cellobiose to α-D-glucose-1-phosphate and *D*-glucose, both of which can readily be metabolised to glucose-6-phosphate (Alexander, 1968).

Following glycolysis, the uniquely, BPc2-encoded phosphate acetyltransferase, catalyses the transfer of the acetyl-group from acetyl-CoA to a phosphate molecule forming acetyl-phosphate and CoA. This enzyme plays an important role in the catabolism of pyruvate through to acetate (Fedorov *et al.*, 2006, Bock & Schonheit, 1995). Acetyl-phosphate is subsequently converted, by an acylphosphatase, to acetate in the final ATP-yielding step. Phosphate acetyltransferase also plays a role in the catabolism of propionate, where it is thought to convert propionyl-CoA to propionyl-phosphate (Palacios *et al*., 2003). Therefore, the activity of this enzyme may influence the ratio of volatile fatty acids (VFA) produced in *B. proteoclasticus*. This is consistent with VFA products released by *B. proteoclasticus* in Complete Carbohydrate medium which were 1.4 mM acetate : 1.2 mM propionate (Attwood *et al*., 1996).

5.15.4 The contribution of BPc2 to nitrogen metabolism

The greatest source of nitrogen available to rumen bacteria is in the form of plant protein. The metabolism of protein nitrogen initially involves the degradation of polypeptides, typically by membrane-bound proteases (Brock *et al*., 1982) releasing shorter oligopeptides and amino acids. BPc2 uniquely encodes a Rhomboid-family protein that forms a membrane-located serine protease. Rhomboid-family proteases cleave peptide bonds through nucleophilic attack of a serine hydroxyl group on the

scissile carbonyl bond, releasing peptides and amino acids. These peptides and amino acids can then be transported into the cell by specific transporters and BPc2 encodes two complete ABC-transporter systems involved in this process. One of these ABC-transporters is dedicated to the uptake of oligopeptides, while the other is involved in the importation of amino acids. BPc2 also encodes two proteins that resemble ABC-type permease components involved in amino acid uptake. However, this system lacks an identifiable substrate-binding component and these ORFs reside more than 10 Kb from the nearest potentially cognate ATP-binding component. Once inside the cell peptides may be further degraded to amino acids by various peptidases; numerous examples exist upon the major chromosome. Amino acids can then be incorporated into freshly synthesised protein or further degraded to volatile fatty acids, CO_2, and ammonia (Bach *et al*., 2005). BPc2 encodes enzymes involved in both the catabolic degradation of oligopeptides (a putative clostripain endopeptidase), and the hydrolysis of amino acids (AnsA). AnsA is present only on BPc2, and catalyses the hydrolysis of *L*-asparagine to *L*-aspartate with the release of ammonia. *B. proteoclasticus* encodes the enzymatic systems that are able to metabolise *L*-aspartate to form *L*-lysine or *L*-methionine, or the organic compounds oxaloacetate or fumarate, which are important to the processes of gluconeogenesis and anaerobic respiration respectively.

Nucleotides can also provide an alternate source of nitrogen for bacteria (Thwaites *et al.*, 1979, Wang & Lampen, 1952, Campbell, 1957). BPc2 encodes a putative 5"-nucleotidase, and two NUDIX-family proteins, that potentially contribute to nitrogen metabolism through the degradation of nucleic acids. 5'-nucleotidases are membrane-bound enzymes that catalyze the extracellular hydrolysis of the phosphate esterified at carbon 5' of the ribose and deoxyribose portions of nucleotide molecules producing their nucleoside derivatives (Zimmermann, 1992). The resulting nucleosides are able to be transported into the cell where they can be further degraded, firstly by nucleoside phosphorylases, releasing their nitrogenous bases, then by nucleobase deaminases, releasing ammonia. Several nucleoside phosphorylases and nucleobase deaminases are encoded upon the *B. proteoclasticus* major chromosome along with an uptake system for uracil and xanthine. The function of the NUDIX- (nucleoside diphosphate linked to another moiety, X)-family proteins is unclear. These proteins include a diverse range of enzymes that are grouped by their ability to hydrolyse nucleoside diphosphate derivatives, including deoxynucleoside di- or tri- phosphates

(Buchko *et al*., 2008), nucleotide sugars, and dinucleoside polyphosphates (Bessman *et al*., 1996). At this time, no evidence exists to demonstrate the ability of *B. proteoclasticus* to catabolise nucleic acids for the production of nitrogen, however the systems appear to be present and nucleic acids are known to be abundant within the rumen (Arambel *et al*., 1982). These enzymes may also have an anabolic role in the scavenging and / or recycling of nucleic acids.

5.15.5 BPc2 encoded genes involved in other forms of energy metabolism

BPc2 encodes an iron-containing alcohol dehydrogenase and fumarate reductase, FccA, both of which have potential roles in energy metabolism. The iron-containing alcohol dehydrogenase is one of three alcohol-dehydrogenases encoded by *B. proteoclasticus*. Despite the name, these enzymes appear to preferentially catalyse the reverse reaction in bacteria, reducing aldehydes to alcohols and thereby recycling NAD^+ (Mackenzie *et al.*, 1989, Ying *et al.*, 2007, Ma & Adams, 1999, Chen, 1995). The production of primary and/or secondary alcohols as a by-product of fermentation by *B. proteoclasticus* has not been tested but the production of small amounts of ethanol have been observed in other *Butyrivibrio* species under carbohydrate-limiting conditions (Strobel, 1994). FccA, catalyses the reduction of fumarate to succinate (Maier *et al*., 2003). The enzyme is occasionally found to also catalyse the reverse reaction and in such cases is thought to play a role in the degradation of benzoate (Ampe *et al*., 1997). Fumarate reductase is essential to anaerobic respiration when utilising fumarate as the terminal electron acceptor (Iverson *et al*., 1999). *Butyrivibrio* species are known to obtain ATP via electron-transport-mediated processes (Dawson *et al*., 1979) and all essential components appear to be accounted for within the *B. proteoclasticus* genome (Dawson *et al.*, 1979). However, fumarate respiration probably plays a limited role as fumarate has the lowest electrochemical potential of all possible terminal electron acceptors, and is therefore typically only utilised when other potential anoxic terminal electron acceptors, (e.g. nitrate, sulphate, dimethyl sulphoxide), are in limited supply (Unden & Bongaerts, 1997). Also, electron-transport-mediated ATP-generation is believed to account for less than half of all ATP generation in *Butyrivibrio* species (Dawson *et al*., 1979). Nevertheless, numerous rumen bacteria, including *Fibrobacter succinogenes*, *Selenomonas ruminantium*, *Veillonella parvula*, *Ruminococcus albus*, *Prevotella ruminicola*, and *Anaerovibrio lipolytica* have previously been shown to utilise fumarate (Asanuma &

Hino, 2000). This process in some, but not all, rumen-bacteria is coupled to the oxidation of H_2, suggesting its use as the terminal electron acceptor. The oxidation of H_2 during fumarate respiration has been proposed to be important in reducing H_2 partial pressures, which can be inhibitory to anaerobic fibre-degradation by rumen bacteria (Russell, 1987, Chung, 1976).

5.15.6 The role of BPc2 in cellular homeostasis

BPc2 appears to have a significant role in cellular homeostasis, encoding nine proteins that potentially extrude or degrade a wide assortment of potentially toxic compounds (Table 5.2). BPc2 encodes two ABC-type efflux permeases of unknown specificity, and three multidrug and toxin extrusion-family proteins (MatEA, MatEB and MatEC). ABC-efflux permeases are typically involved in the removal of specific toxic compounds from the cell (Schluter *et al*., 2007; Schouten *et al*., 2008). MatE proteins are membrane-bound transporters that utilise either sodium (Long *et al*., 2008) or proton (He *et al*., 2004) gradients to extrude a broad range of antimicrobial substrates (He *et al*., 2004, Morita *et al.*, 1998) including intercalating agents such as ethidium bromide, antiseptics such as acriflavin,; antibiotics such as novobiocin, doxorubicin and a number of aminoglycosides, monovalent and divalent biocides, fluoroquinolones, the isoquinoline berberine; and lipophilic quaternary compounds such as tetraphenylphosphonium (Kaatz *et al.*, 2005, Long *et al*., 2008, Begum *et al.*, 2005, He *et al*., 2004). Two of the MatE proteins are found close to the replication origin, suggesting their functions are of significant importance to *B. proteoclasticus* (Mira & Ochman, 2002, Horimoto *et al.*, 2001, Liu & Sanderson, 1996). The CrcB protein is implicated along with CrcA and a cold shock protein homologue, in resistance to the chromosome-decondensation effects elicited by the terpene, camphor (Hu *et al*., 1996). While cold shock protein-homologues are present on all replicons except BPc2, no CrcA homologues are evident within the genome, therefore the role of CrcB is uncertain and its placement within the detoxification category is tentative. Likewise, the penicillin-binding protein and the β-lactamase encoded by BPc2 may have roles in the breakdown of β-lactam antibiotics. Homologues of each of these proteins are known to contribute to the biosynthesis and recycling of the peptidoglycan component of bacterial cell walls (Goffin & Ghuysen, 1998, Park & Uehara,

2008), although peptidoglycan recycling is currently thought to only occur in Gram-negative species (Park, 1995). Many of the detoxification genes reside in a 28 Kb locus that also encodes three multiple-antibiotic resistance (MarR)-family transcriptional regulators and two putative GNAT-family acetyltransferases (Fig. 5.2 D1). MarR-family transcriptional regulators are typically ligand-sensing transcriptional repressors, which, in the absence of ligand, bind to the -35, -10 and/or ribosome binding sites of regulated genes. The *marR* gene is typically located within the cluster of genes it regulates and is predominantly autoregulatory. MarR regulators are not confined to the regulation of genes involved in antimicrobial resistance and have been implicated in the control of virulence factor production, response to oxidative stresses and the catabolism of aromatic compounds. One of the MarR-family transcriptional regulators is found just upstream of *matEB* and the putative GNAT-family acetyltransferase. A MarR homologue has previously been shown to regulate *matE* in *Staphylococcus aureus* (Kaatz *et al*., 2005), while various other homologues are known to regulate multi-substrate efflux systems implicated in multiple antimicrobial resistances in a wide assortment of bacteria (Poole *et al.*, 1996, Srikumar *et al.*, 2000, Lomovskaya *et al.*, 1995, Lee *et al.*, 2003). GNAT- (Gcn5-related N)-family acetyltransferases are a large and diverse-family that catalyse the transfer of an acetyl group from acetyl-coA to different substrates (Vetting *et al*., 2005). One of the GNAT-family acetyltransferase proteins encoded within the central detoxification locus, and mentioned above, shows significant sequence identity to a phosphinothricin *N*-acetyltransferase from *Bacillus halodurans* which is implicated in the inactivation of the glutamine-synthase inhibitor, phosphinothricin. Phosphinothricin is derived from the degradation of phosphinothricyl-alanyl-alanine, a tripeptide produced by aerobic soil-colonizing bacteria of the genus *Streptomyces* (Bayer *et al.*, 1972, Seto *et al.*, 1982) and is used as the active component in several broad-spectrum herbicides. Collectively these genes are likely to enhance the tolerance of *B. proteoclasticus* toward a broad range of antimicrobial molecules.

5.15.7 Chemotaxis and flagella-related proteins

B. proteoclasticus possesses a single sub-terminal flagellum, yet it is non-motile in broth cultures (Attwood *et al*., 1996). The reason for this apparent anomaly is believed to be due the presence of a feruloyl esterase gene that disrupts the flagellar biosynthetic gene cluster resulting in a non-functional flagellum. *B. proteoclasticus* also encodes a large number of proteins involved in chemotaxis. While most of the proteins dedicated to flagellar biosynthesis and chemotactic motility are encoded by the major chromosome, BPc2 encodes a putative flagellar protein and the chemotaxis proteins CheA, CheW, CheY and McpD, all of which have homologues on the major chromosome. CheA, CheY and CheW act as a signal transduction system analogous to the two component regulatory systems common to bacteria. CheA, the sensor-histidine kinase, associates with CheW and trans-membrane-located chemoreceptors, such as McpD. In the presence of chemoreceptor-specific substrate, CheA autophosphorylates then transfers its phosphate to the response regulator CheY (Stewart, 1997, Kentner & Sourjik, 2006, Lux *et al.*, 1995). Phosphorylated CheY then diffuses to the flagellar motor where it acts as an allosteric regulator to promote rotation (Sourjik & Berg, 2002, Alon *et al.*, 1998). This process requires the hydrolysis of ATP and is therefore an energy-consuming process, which is apparently futile if it does not result in cell motility. It therefore seems likely that the loss of motility in *B. proteoclasticus* is a recent event. If this is correct, BPc2 may have contributed to cell chemotaxis in the past and may still play a minor role in flagellar biosynthesis.

5.15.8 Vitamin and cofactor uptake and metabolism

BPc2 encodes ten proteins annotated to be associated with the uptake and biosynthesis of various cofactors, including CorAA, PnuC, NadR BioY, an acyl-carrier protein phosphodiesterase and a putative glutaredoxin. While other systems can substitute for CorAA in the acquisition of cobalt and magnesium, the ability to scavenge exogenous biotin or nicotinamide riboside, (an essential precursor for NAD), appear to be conferred only by BPc2. PnuC acts as an uptake permease specific for nicotinamide riboside (Grose *et al.*, 2005, Sauer *et al.*, 2004), while *nadR* encodes a nicotinamide riboside kinase and adenylyltransferase, which converts imported nicotinamide riboside to nicotinamide mononucleotide, thereby preventing its escape from the cell. NadR can convert nicotinamide mononucleotide to NAD, although this is also performed by NadD and NadE (Kurnasov *et al.*, 2002, Singh *et al.*, 2002, Grose *et al.*, 2005), two enzymes encoded by the major chromosome. The cofactor NAD^+ and it‟s reduced- (NADH) and phosphorylated- (NADP and NADPH) derivatives are essential electron carriers implicated in numerous catabolic and biosynthetic redox reactions (Magalhaes *et al.*, 2008; Joyner and Baldwin, 1966). Collectively, PnuC and NadR are essential for the biosynthesis of NAD via the pyridine recycling pathway (Kurnasov *et al.*, 2002). The *B. proteoclasticus* genome lacks homologues of *nadA, nadC* and *pncA*. NadA and NadC are each critical to the de-novo biosynthesis of NAD, while PncA is critical to another known, but poorly elucidated, scavenging and recycling pathway (Foster *et al.*, 1979, Grose *et al.*, 2005). The NadR-PnuC pathway appears to be one of only two pathways available to *B. proteoclasticus* for NAD biosynthesis, the other involving the conversion of glutamate.

B. proteoclasticus appears to be dependent upon the BPc2-encoded BioY for the provision of biotin. BioY appears to encode a high capacity biotin importer (Hebbeln *et al.*, 2007) and can presumably satisfy the biotin requirements of *B. proteoclasticus* through the uptake of exogenous biotin. Analysis of the *B. proteoclasticus* genome sequence reveals an absence of *bioA, bioB* and *bioD* homologues and shows the presence of an α/β hydrolase with only weak similarity to *bioF*. All four enzymes (BioA, B, D, and F) are required for the *de-novo* biosynthesis of biotin (Krell & Eisenberg, 1970, Ploux & Marquet, 1992, Eisenberg & Krell, 1969, Ohsawa *et al.*, 1989). The requirement of exogenous biotin has been reported previously for a

Butyrivibrio species (Wachenheim & Patterson, 1992). A gene that encodes the bifunctional protein, biotin-acetyl-CoA-carboxylase ligase / biotin operon regulator, *birAA*, resides directly upstream of *bioY*. In characterised systems, BirA-homologues negatively regulate the biotin biosynthesis operon and are responsible for the biotinylation of various biotin-dependent carboxylase enzymes including acetyl-CoA-carboxylase and pyruvate carboxylase, both of which are present on the *B. proteoclasticus* major chromosome (Bower *et al.*, 1995, Campbell *et al.*, 1980). These carboxylase enzymes utilize biotin as a cofactor in a number of important biological processes including gluconeogenesis, the metabolism of fatty acids, amino-acids and carbon dioxide fixation (Gavin & Umbreit, 1965, Hartman, 1970, Renner & Bernlohr, 1972). Another BPc2-encoded enzyme potentially implicated in fatty acid metabolism is the putative acyl carrier protein phosphodiesterase. Acyl carrier protein phosphodiesterases catalyze the hydrolytic cleavage of the 4'-phosphopantetheine residue from the prosthetic group of an acyl carrier protein, a cofactor in fatty acid biosynthesis. The gene encoding the acyl carrier protein phosphodiesterase is listed in the Bacterial Minimal Gene Set. Glutaredoxin is a small redox enzyme, best characterised as a hydrogen donor to aerobic ribonucleotide reductases (Holmgren, 1976). The entire glutaredoxin system, which additionally consists of glutathione and glutathione reductase, is not common to Gram positive organisms and is less common in obligate anaerobes (Fahey *et al*., 1978). In these organisms the function of the glutaredoxin system is able to be replaced by the thioredoxin system (Rocha *et al*., 2007). Analysis shows the *B. proteoclasticus* genome to be auxotrophic for glutathione reductase, yet encode all required components of a thioredoxin system. Therefore the presence of a glutaredoxin-like gene upon BPc2 is unusual and its encoded protein is unlikely to fulfil the archetypical glutaredoxin role. BPc2 also encodes a 6,7-dimethyl-8-ribityllumazine synthase pseudogene. In its native form 6,7-dimethyl-8-ribityllumazine synthase condenses 5-amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione with 3,4-dihydroxy-2-butanone 4-phosphate to give 6,7-dimethyl-8-ribityllumazine, a precursor in the biosynthesis of riboflavin (vitamin B_2).

5.15.9 Other genes unique to BPc2

BPc2 also encodes a peptide methionine sulfoxide reductase, MsrA and an RNA 2′-phosphotransferase, KptA. MsrA is described in the bacterial minimal gene set and catalyses the reduction of free or protein-bound methionine sulfoxides, which occur due to exposure to reactive oxygen or nitrogen species during aerobic metabolism. The MsrA enzyme is highly conserved among obligate anaerobes and in endosymbiotic bacteria, and in these systems it may reduce exogenous methionine that has been oxidized by host-released reactive oxygen species (Brot *et al.*, 1981, Ezraty *et al.*, 2005, St John *et al.*, 2001, Alamuri & Maier, 2004). MsrA also has activity against several other sulphoxide compounds and *msrA* expression in *Ochrobactrum anthropi* is up-regulated during growth in the presence of phenolic compounds (Tamburro *et al*., 2001) indicating it may have an additional role in cellular detoxification.

KptA is homologous to Tpt1 of *Saccharomyces cerevisiae* (Spinelli *et al*., 1998) which transfers the 2′-phosphate from tRNA to NAD in the final step of tRNA splicing (Culver *et al*., 1997). However, as tRNA splicing is not known to occur in bacteria, the function of this enzyme in *B. proteoclasticus* is unclear. Phylogenetic analysis of the enzyme shows no evidence of a lateral acquisition event by bacteria, suggesting the enzyme has an analogous role in bacteria (Spinelli *et al*., 1998). Mutational inactivation of *kptA* in *E. coli* is not lethal, unlike in yeast, therefore the importance of this process is uncertain (Spinelli *et al*., 1998, Culver *et al*., 1997). KptA may interact with the pCY360-encoded APPR-1-P processing enzyme (ORF 22) which catalyzes the conversion of ADP-ribose-1′′′monophosphate, a hydrolysed derivative of ADP-ribose-1′′′-2′′′′cyclic phosphate which is produced in the final step of tRNA splicing, to ADP-ribose (Kumaran *et al*., 2005).

5.15.10 Microarray analysis of BPc2 gene expression during co-culture of *B. proteoclasticus* with a rumen methanogen

B. proteoclasticus was grown as a mono-culture or in co-culture with the methanogenic archaeon *M. ruminantium* and changes in gene transcript levels between the two conditions were analysed by microarrays. The microarray data supported the transcription of 224 (95%) of the BPc2 ORFs present on the array. Of these, 38 were found to be regulated significantly differently between the two conditions. Most of the up-regulated ORFs were unable to be ascribed a function, suggesting that one or more biological processes occurring during this interaction are either poorly described or undertaken by unique enzymes acting in known pathways in *B. proteoclasticus*. The majority of the up-regulated hypothetical or conserved hypothetical proteins are predicted to span the membrane or to be secreted from the cell, suggesting they may interact directly with *M. ruminantium*. Three further significantly up-regulated ORFs encode proteins that are likely to have a significant impact on the *B. proteoclasticus* membrane, including a band-7 protein homologue and two di-guanylate cyclase (GGDEF) proteins. The band-7 protein is homologous to an integral membrane protein, common within eukaryotes, that is thought to regulate cation conductance (Lande *et al*., 1982), while a homologue in *Bacillus halodurans* was up-regulated under alkaline conditions (Zhang *et al*., 2005). Di-guanylate cyclise enzymes synthesise the secondary messenger, 3‘-5‘-cyclic di-guanylate monophosphate (di-GMP). Di-GMP acts antagonistically with phosphodiesterase A, (a member of the EAL-domain-family), to regulate intracellular di-GMP levels (Simm *et al*., 2004). Increased intracellular levels of di-GMP has been implicated in the allosteric activation of enzymes involved in exopolysaccharide biosynthesis leading to increased biofilm formation (D'Argenio & Miller, 2004, Gjermansen *et al*., 2006, Kim & McCarter, 2007). Consistent with this finding, exopolysaccharide biosynthesis genes appear to be up-regulated by *B. proteoclasticus* during co-culture. Genes implicated in flagellar formation are also up-regulated during co-culture, including the putative flagellar protein-encoding ORF of BPc2. Flagellar are known to play a significant, and occasionally essential, role in biofilm formation (Pratt & Kolter, 1998, Gavin *et al*., 2002, O'Toole & Kolter, 1998). Together with exopolysaccharides, flagellar are likely to contribute to the formation of the *B. proteoclasticus – M. ruminantium* co-aggregates observed during co-culture (Fig. 4.11). Three genes, *hicA*, *hicB* and the gene encoding the DNA-binding protein,

HU were found to be up-regulated on BPc2. HicA has a double-stranded RNA-binding domain and HicB possesses a partial RNase H fold (Makarova *et al*., 2006) which suggests they may have roles in RNA metabolism. The DNA-binding protein HU, is a histone-like non-specific DNA-binding protein that affects nucleoid condensation (Endo *et al*., 2002) and consequently transcriptional capacity. An increase in the amount of HU would theoretically result in a more condensed, and consequently less transcriptionally active, DNA molecule. HU orthologues are also present upon the major chromosome and pCY186, however, neither were identified by GLIMMER analysis of the draft genome sequence and therefore probes to these genes were not included on the microarray slides. Most of the BPc2-encoded ORFs found to be significantly down-regulated have roles in carbohydrate metabolism. This includes ORFs encoding the enzymes Ram78A, Pel9A and Fuc29A along with most components of the sugar-specific ABC-transport system. This is consistent with the overall *B. proteoclasticus* transcriptome profile that shows transcriptional repression of genes encoding polysaccharide-degrading enzymes and sugar-specific ABC-transport systems (Section 4.13.2). Their repression may be mediated by catabolite repression invoked by the up-regulation of several components of the phosphoenolpyruvate-phosphotransferase (PTS) system, encoded by the major chromosome. The PTS system is a multi-component transport system, specific for simple sugars, that is implicated in various regulatory roles, including catabolite repression. The carbon substrate supplied to *B. proteoclasticus* in both mono- and co-cultures was xylan, which is degraded mainly to the monosaccharides, xylose and arabinose, neither of which are typical PTS substrates. However, genes encoding PTS system components have previously been shown to be co-regulated with genes involved in exopolysaccharide biosynthesis and mutagenic disruption of several PTS genes has been shown to inhibit biofilm formation (Houot & Watnick, 2008, Moorthy & Watnick, 2005, Knobloch *et al.*, 2003). Furthermore, the induction of the PTS system and concomitant down-regulation of sugar-importation and polysaccharide-degrading systems may be a general biofilm response. In the rumen, *B. proteoclasticus* is likely to exist in a biofilm, associated with other rumen microorganisms. Under these circumstances one would expect a variety of sugars, principally, glucose, xylose and arabinose would be released through the hydrolysis of plant celluloses and hemicelluloses (Matulova *et al.*, 2008, Thurston *et al.*, 1993). Therefore, it would be an advantage, energetically, to use an available substrate such

as glucose first rather than synthesize a range of polysaccharide-degrading enzymes to release fermentable sugars from plant material. The down-regulated genes that encode Ram78A, Pel9A and Fuc29A enzymes, are all involved in the degradation of common constituents of bacterial biofilms (Ratto *et al*., 2005). This may be to prevent the futile-degradation of exopolysaccharides produced by *B. proteoclasticus* in the co-culture, which appear to function to enhance interspecies aggregation.

The BPc2 gene encoding the acyl-carrier protein phosphodiesterase is down-regulated in the co-culture, as are the chromosomally-encoded acyl carrier protein and many of the enzymes involved in fatty acid synthesis. The transcriptional repression of genes, or their encoded enzymes, involved in fatty acid synthesis has been observed previously and is thought to result in an altered membrane composition and consequential reduction in hydrophobicity. It has also been shown that biofilm-associated cells tend to be more hydrophilic than their planktonic equivalents (Allison *et al*., 1990). Previous studies have found cells can detach from a biofilm surface through the increase in hydrophobicity (Nikolaev Iu *et al*., 2001). The attenuation of cell surface hydrophobicity is believed to contribute to the adhesion process (Gianotti *et al*., 2008).

5.16 Summary

Analysis of the BPc2 sequence reveals it is likely to contribute significantly to a variety of important biological processes in *B. proteoclasticus*. BPc2 encodes enzymes that appear essential to the uptake of biotin, important to the uptake and biosynthesis of nicotinamide adenine mononucleotide and the ability to utilise fumarate as the terminal electron acceptor during anaerobic respiration. The replicon also makes extensive contributions to carbohydrate and nitrogen metabolism as well as cellular detoxification. These features, along with the presence of two ribosomal RNA operons, two transfer RNAs and two uniquely-encoded enzymes described in the Bacterial Minimal Gene Set, suggest BPc2 is a secondary chromosome. It appears to make significant contributions to the interspecies interaction with *M. ruminantium* contributing to cellular aggregation and central metabolism in *B. proteoclasticus*. Many of these processes require enzymes encoded by both the major chromosome and BPc2, suggesting that dynamic regulatory processes must operate in order to coordinate expression from the different replicons of *B. proteoclasticus.* The presence of a secondary chromosome appears to be unique within the *Butyrivibrio* species assemblage to *B. proteoclasticus* B316^{T}.

6 pCY186

6.1 Introduction

The presence of multiple megaplasmid-sized replicons has previously been observed in a number of bacteria, particularly among Gram-negative bacteria. However examples of megaplasmids in Gram-positive bacteria also exist including *Arthrobacter aurescens* TC1, *Arthrobacter sp.* FB24, *D. radiodurans* R1, *Mycobacterium sp*. KMS, *Rhodococcus erythropolis* PR4 and *Rhodococcus sp*. RHA1. Of these organisms, only *Rhodococcus sp.* RHA1 and *D. radiodurans* R1 possess more than two megaplasmid-sized replicons, each, having three. In all described Gram-positive examples the co-resident replicons largely encode unique functions (Mongodin *et al.*, 2006, White *et al.*, 1999, Sekine *et al.*, 2006) despite occasional locus duplication being observed between the co-residing megaplasmids (McLeod *et al*., 2006) or between the megaplasmid and the hosting major chromosome (Mongodin *et al*., 2006). Nevertheless, examples do exist where co-residing replicons appear to cooperatively contribute to pathways, such as the mineralisation of alkanes (Sekine *et al*., 2006), resistance to heavy metals (Mongodin *et al*., 2006), DNA-damage repair (White *et al*., 1999) or plasmid partitioning (McLeod *et al*., 2006, Jerke *et al.*, 2008).

Analysis of the largest two auxiliary replicons from *B. proteoclasticus* revealed them to be significantly different with regard to their gene complements. The pCY186 replicon proved technically difficult to sequence completely with high throughput techniques resulting in a significant under-representation in particular regions of the replicon. However, following array-based pyrosequencing, the sequence was able to be finished to a Q40 quality standard.

This chapter describes the detailed analysis of the pCY186 replicon, the attempted elucidation of its function(s), origin and contribution to the biology of *B. proteoclasticus*.

6.2 Sequence analysis of pCY186

The complete sequence of the pCY186 replicon was found to be 186,328 bp in length, consistent to within 2% of the size estimated from pulsed-field gel analysis. The replicon has the lowest overall %G + C content of all the *B. proteoclasticus* replicons at 38.10%, and is lower in both coding (38.93%) and non-coding (34.46%) regions than the major chromosome. A total of 205 potential ORFs were identified by GLIMMER (Salzberg *et al*., 1998) analysis. Seven of these were eliminated for reasons described in Section 4.2, giving pCY186 a total of 198 ORFs considered to be genuine protein coding regions. Additionally, pCY186 encodes a single alanine tRNA. The predicted ORFs have an average size of 807 bp, and this again appears to be exaggerated by a small portion of much larger ORFs, the median size being just 602 bp (Fig. 6.1). This gives pCY186 a gene density of 1.06 genes per kb and gene coverage of 81%. The ORFs are evenly spread between the Watson (38%) and Crick (62%) strands and analysis shows the replicon to have a classical third-position GC skew, similar to the two *B. proteoclasticus* chromosomes (Fig. 6.2). Start codon distribution is typical, with 95% ATG, 3.5% GTG and 1.5% TTG. The proteins encoded by pCY186 ORFs have an average isoelectric point (pI) of 6.45 (median 5.72; Fig. 6.1). As with the larger megaplasmid, pCY360, the majority of the predicted pCY186 proteins are unable to be ascribed a putative function (63% hypothetical proteins, 13% conserved hypothetical proteins). pCY186 encodes 28 (14%) proteins predicted to span the membrane more than once and a further 19 (10%) proteins predicted to be secreted. Five of the predicted proteins have been identified by tandem mass spectrophotometry analysis of 1-dimensional protein gels (Dunne, in preparation). These identifications resulted in two hypothetical proteins being reclassified as uncharacterised proteins (ORFs 151 and 177, respectively).

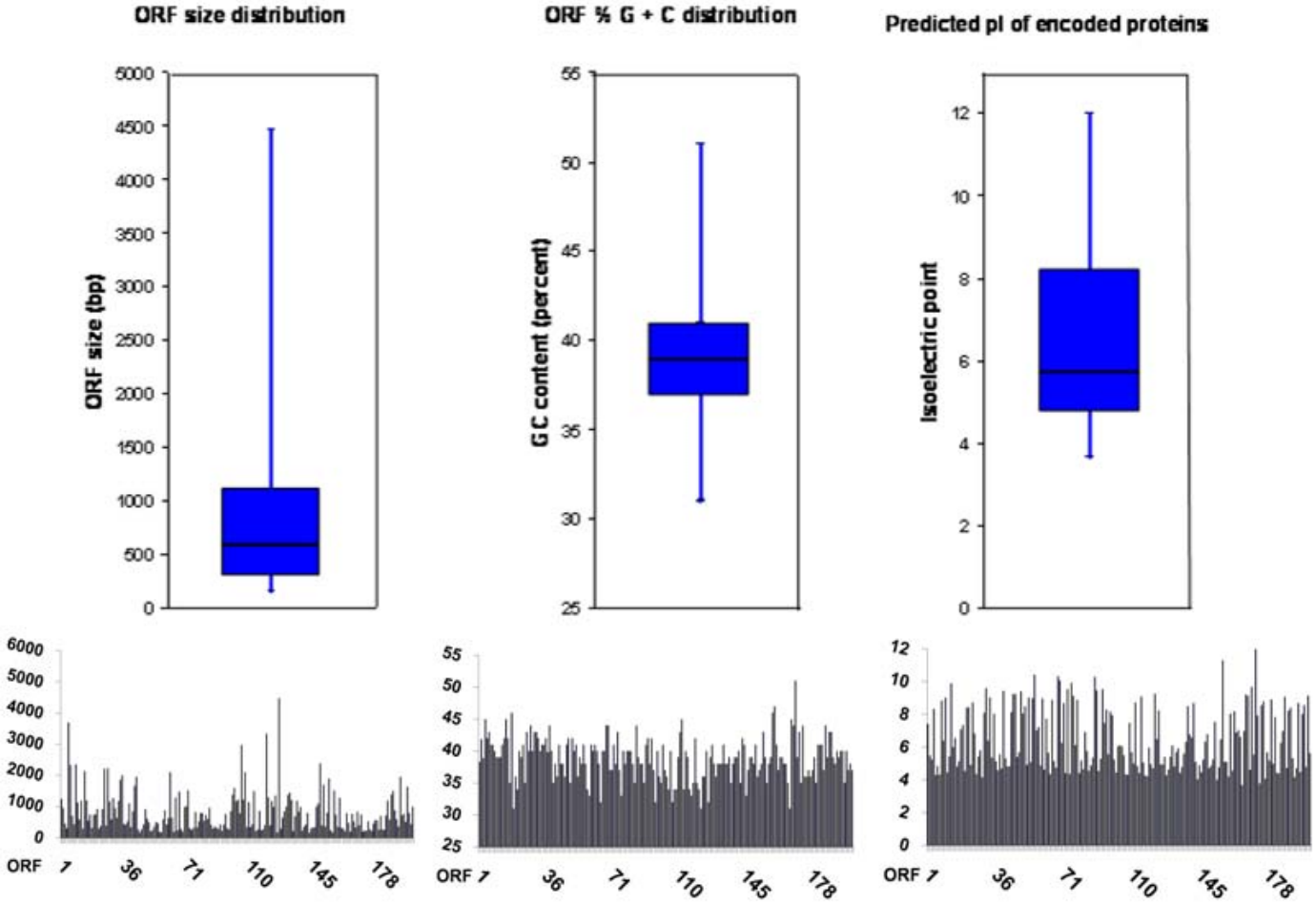


Fig. 6.1. ORFs of pCY186 and encoded protein composition . Box plots show size, %G + C and isometric point distribution for each quartile of the data. Column charts show the distribution of each feature from ORF 1 – 198.

6.3 The replication origin of pCY186

The pCY186 replicon encodes three RepB-family replication initiation proteins, two of which flank a locus that encodes proteins resembling chromosomal-type replication machinery. The proteins encoded by this locus show significant amino acid sequence identity to the chromosomal replication initiation protein DnaA, the replicative DNA-helicase, DnaB, a complete DNA polymerase III (PolIII), PolCA, along with two additional proteins encoding PolIII α-subunits and a protein encoding the PolIII γ and τ subunits. Despite this wealth of replicative machinery a second locus was favoured as possessing the replicative origin of pCY186. As with pCY360 and BPc2, this locus, (nt 183123 - 1309), encodes a putative replication initiation protein, RepB (ORF1) and a plasmid partitioning protein, ParA (ORF196). However unlike the other two replicons this locus also encodes the partitioning protein, ParB (ORF198). The 1359 bp intergenic region found upstream of *parA* was found to contain a 439 bp A+T rich (18% GC) region with significant nucleotide identity (53%) to the putative *oriR*s of *B. fibrisolvens* plasmid, pRJF1 pCY360 (61%) and BPc2 (55%). Analysis of the pCY186 intergenic region identified two direct repeat families (DR1, 21 bp and DR2, 23 bp), one-family of inverted repeats (IR1, 28 bp), two inverse repeats (IV1, 14 bp and IV2 13 bp) and ten *dnaA* box candidates. The IR1 occurs nine times, therefore they can additionally form a series direct repeats on each strand. DR1 is repeated four times, DR2 twice, while IV1 and IV2 both occur once. Four *dnaA* box candidates, one copy of DR2, three copies of DR1 and both IR1 and IV2 are encompassed by the sequence that aligns to the putative *oriR*s of pRJF1, pCY360 and BPc2. This locus is also supported as containing the true *oriR* by the third-position GC skew (Fig. 6.2) which suggests replication initiates, or terminates within this intergenic region.

6.4 DNA metabolism

In addition to the proteins encoded by pCY186 that resemble chromosomal replicative machinery, the replicon encodes three proteins with potential roles in DNA metabolism. This includes a putative UmuC-like DNA repair protein (ORF 157), and two hypothetical proteins that show weak similarity to *yolD* (ORF 158) and DNA primase (ORF 182).

6.5 Restriction modification

A cluster of genes residing between nt 88130 - 97336 have functions dedicated to restriction modification. This includes four components of a Type I Restriction Modification system (ORFs 98, *hsdM*; 99, *hsdSA*; 100, *hsdSB*; and 102 *hsdR*) and two Type II Restriction Endonucleases (ORFs 97 and 101).

6.6 pCY186 contains genes described in the Bacterial Minimal Gene Set

As with pCY360 and BPc2, pCY186 encodes ORFs described in the Bacterial Minimal Gene Set (Koonin, 2000). These include a DNA replication-initiation protein (DnaA), a replicative DNA helicase (DnaB), two DNA polymerase III α-subunits, a single stranded DNA-binding protein, an AAA+ ATPase and a zinc metalloprotease, FtsH (Table 6.2). DNA primase is also described in the Bacterial Minimal Gene Set, consequently ORF 182 may also belong in this group. Most of these ORFs have functions related to DNA metabolism, however, as with pCY360, none of these proteins are unique to the *B. proteoclasticus* genome.

6.7 Transposases

Two transposase-encoding ORFs were identified within the pCY186 sequence. This includes an IS110-family transposase (ORF10) and an IS1182-family transposase (ORF195). The IS110-family transposase shows 100% identity at both the amino acid and nucleotide level to four transposases on the major chromosome and one found on BPc2, (previously discussed in Section 5.10). The IS1182-family transposase shows strong identity (98 - 100%) to the carboxyl-termini of two transposases found on the major chromosome.

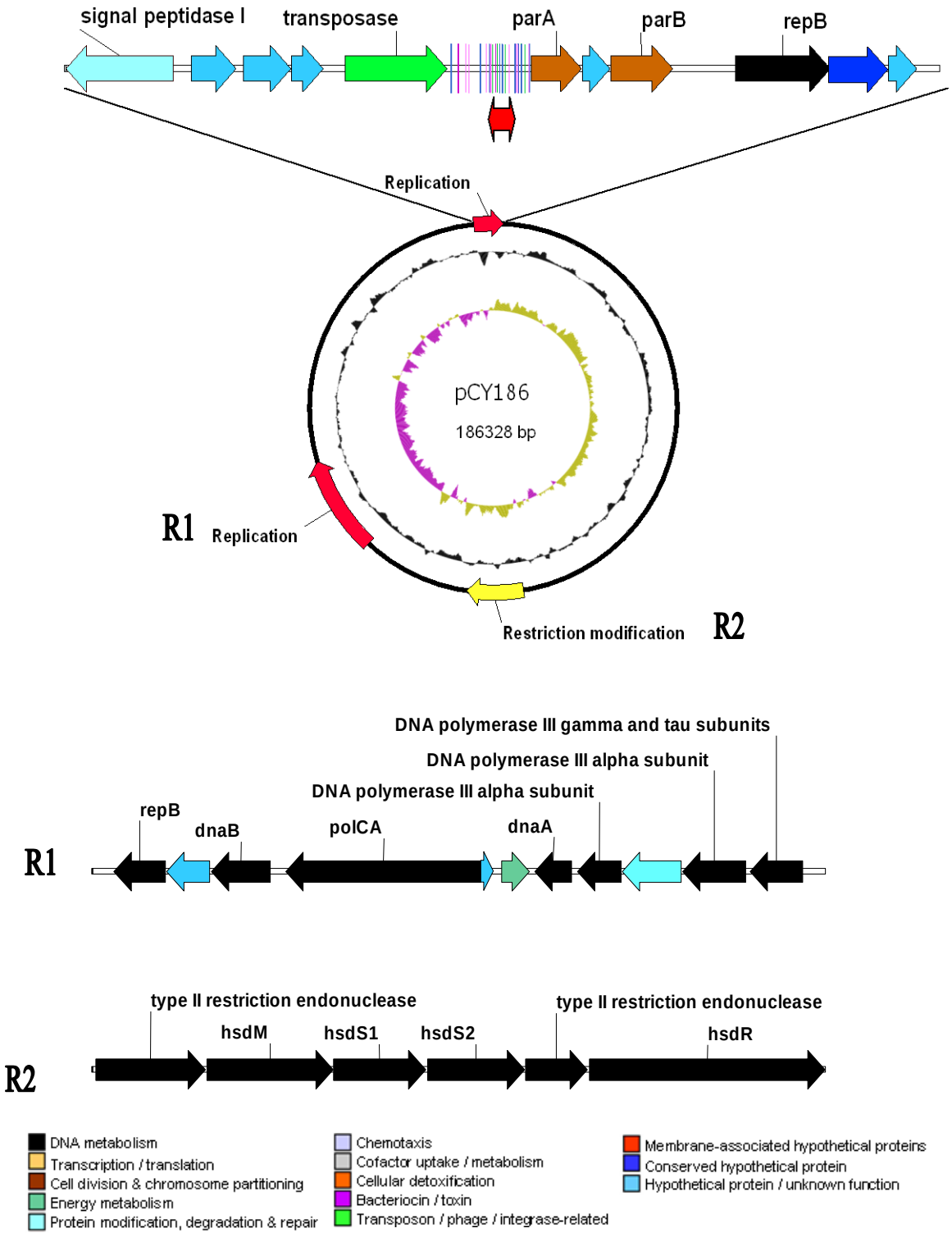


Fig. 6.2. Compositional map of pCY186. The main feature shows, from outside to inside, gene clusters with an identifiable function, the deviation from the average %G + C, and the third-position GC skew. The putative *oriR*-containing locus (R1) is enlarged above. *dnaA* boxes (blue), direct repeats (DR1 and DR2, green), inverse repeats (IV1 and IV2, purple) and inverted repeats (IR1, pink) are shown as well as the predicted *oriR*. Expanded views of the replication and RM clusters are shown, with ORFs colour coded by function according to the key. Maps were generated with Vector NTI (InforMax, 2001)

Table 6.2 pCY186 ORFs described in the bacterial minimal gene set

ORF	Size (aa)	Putative function	Organism with best BLAST match	E-value	% ID	GenBank Accession
47	139	Single stranded DNA-binding protein	*Anaerostipes caccae*	6 e^{-31}	49%	ZP_02419566
80	260	AAA+ ATPase	*Paenibacillus larvae*	$5e^{-21}$	30%	ZP_02327396
121	449	Replicative DNA helicase, DnaB	*Ruminococcus obeum*	$1e^{-142}$	57%	ZP_01963276
126	281	Chromosomal replication initiation protein, DnaA	*Thermotoga petrophila*	$9e^{-28}$	42%	YP_001243606
127	332	DNA polymerase III α-subunit	*Moorella thermoacetica*	$2e^{-16}$	38%	YP_429902
128	458	ATP-dependent zinc metalloprotease, FtsHB	*Faecalibacterium prausnitzii*	$2e^{-50}$	33%	ZP_02091431
129	480	DNA polymerase III α-subunit	*Clostridium scindens*	$5e^{-40}$	34%	ZP_02432860

6.8 Attempts to cure *B. proteoclasticus* of pCY186

Attempts were made to create strains devoid of the pCY186 replicon (as described in Section 4.11). A total of 5,688 colonies (1,107 from treatment with novobiocin; 1,016 from acriflavine treatment, 1,198 from ethidium bromide treatment; 1,156 from acridine orange treatment; 1,211 from growth at 45 ºC) were screened by colony hybridisation using the pCY186 specific probe (p190_probe2), designed to target a 745 bp, region of the megaplasmid spanning part of two hypothetical proteins (ORFs 4 and 5) unique within the *B. proteoclasticus* genome. A total of 142 colonies 48 (4.3%) from the novobiocin treatment, 33 (3.2%) from the acriflavine treatment, 19 (1.6%) from the ethidium bromide treatment, 22 (1.9%) from the acridine orange treatment, and 20 (1.7%) from growth at 45 °C were not detected by the Southern probe, suggesting these *B. proteoclasticus* cells lacked the pCY186 megaplasmid. Analysis of the positive control (*B. proteoclasticus* grown in M704 at 39 °C under non-stressful conditions), revealed the pCY186 megaplasmid to be absent in 2 of the 148 colonies examined. Twenty seven of the potentially cured strains (5 from each condition and the 2 cured positive controls) were examined by pulsed-field gel-electrophoresis and all were found to lack the 186 Kb band associated with the pCY186 megaplasmid (Fig 6.3a).

6.8.1 Morphological differences

The morphology of *B. proteoclasticus* devoid of the pCY186 megaplasmid (ΔpCY186) was examined. Microscopic analysis revealed cells occurring in long chains with a significantly greater frequency to the wild type *B. proteoclasticus* (Fig. 6.3a). This feature was evident throughout the complete growth cycle of the organism. The ΔpCY186 cultures grown on solid media were found to form white colonies, quite distinct from the grey colony typically observed in the wild type. However, the white colony phenotype was occasionally observed in wild type cultures.

6.8.2 Growth rate

The growth rate of ΔpCY186 was compared to the wild type to determine if the presence of pCY186 had a positive or negative impact on the generation time of *B. proteoclasticus*. Growth rate was determined by spectrophotometry (described in Section 2.2.3). The generation time of ΔpCY186, (105.4 ± 14.1 mins; 99% confidence), was not significantly different from the wild type (99 ± 13.7 mins; 99% confidence) (Fig. 6.3b).

6.9 Copy number of pCY186

The copy number of pCY186 was determined as previously described (Section 4.12). The T_m of the pCY186 primer set, p190_probe2, was optimised by conventional PCR as described (Section 4.12.1). The optimal annealing temperature for the pCY186 primer set was determined to be 60 °C (Fig. 6.4a). The reaction was then optimised for $MgCl_2$ concentration, as described (Section 4.12.1). From this analysis the optimal $MgCl_2$ concentration using an annealing Tm of 60ºC, ensuring both specificity and optimal amplification efficiency, was determined to be 3 mM (Fig. 6.4b). The specificity of the qPCR reaction was determined by melting curve analysis and gel-electrophoresis. Melting curve analysis revealed a single peak for the p190_probe2 reaction occurring at 85.34 ºC and subsequent gel-electrophoresis revealed it to be consistent with the expected 745 bp size of the p190_probe2 product (Fig 6.5). The standard curve of the threshold cycles vs. the $\log_{10}$ of the replicon copy numbers for the Topo-recombinant plasmid carrying the p190_probe2 sequence (Fig 6.6a) had a slope of -3.448. This equates to amplification efficiency for the reaction of 0.95. The standard curve was linear in the range tested ($1 \times 10^5 – 1 \times 10^9$ copies / µl; R^2 >0.99). A standard curve of the threshold cycles vs. the $\log_{10}$ of the fold dilution for the target (Fig. 6.6b) was also determined to ensure they were approximately equal. The slope was determined to be 3.413 for the p190_probe1 target. This equates to an amplification efficiency of 0.96 for the reaction. The standard curve was linear in the range tested ($1 \times 10^{-1} – 1 \times 10^{-5}$ DNA extract / reaction; R^2 >0.99). Analysis revealed the replicon to be present at $1.14 \times 10^{10} \pm 1.03 \times 10^9$ copies in the tested sample (99% confidence) per $1.3 \times 10^{10} \pm 1.75 \times 10^9$ copies of the chromosome (99% confidence). This equates to a copy number of 0.88 $^{+}$/. 0.16 copies per chromosome (99% confidence).

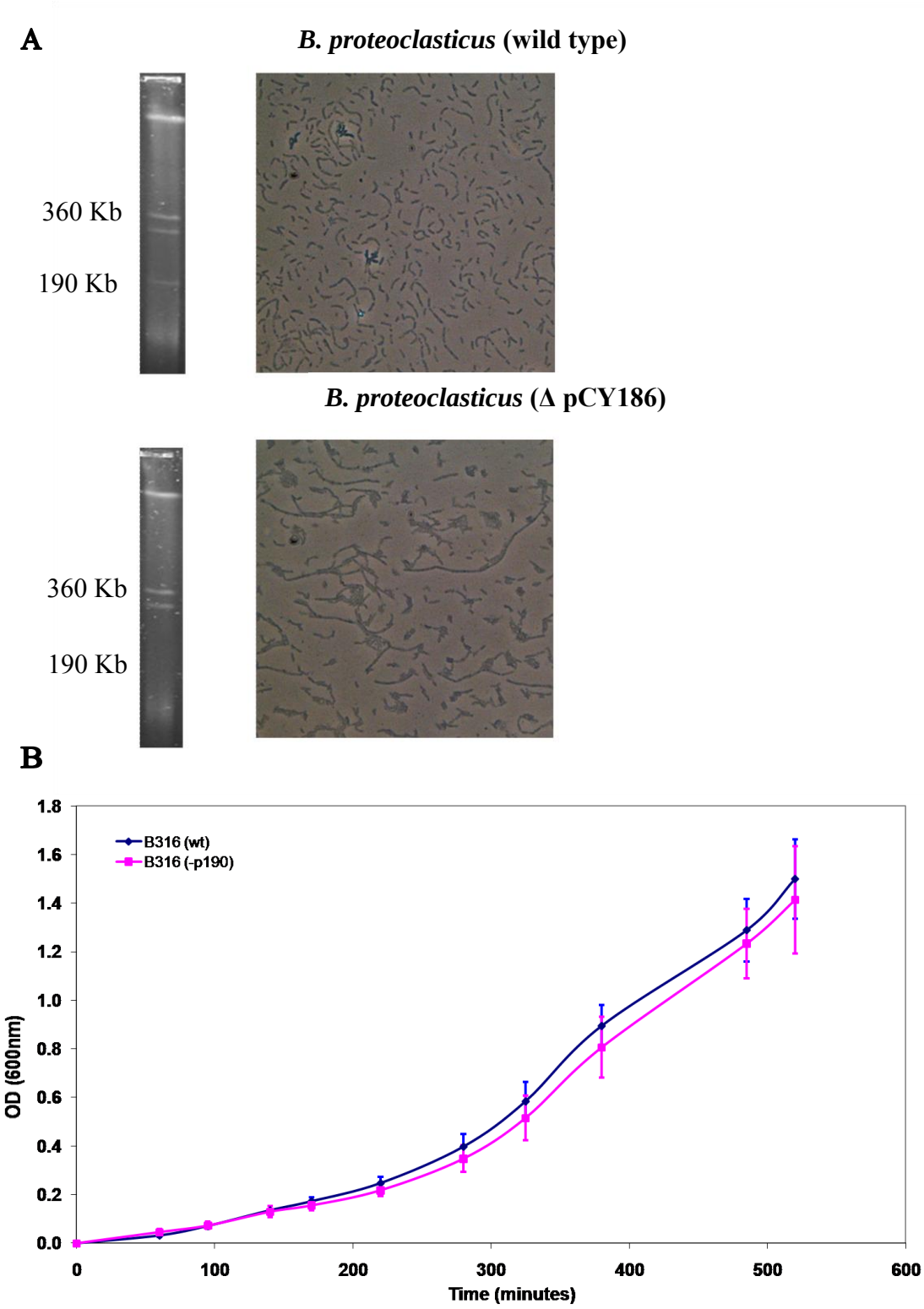


Fig. 6.3. Analysis of ΔpCY186. (A) Comparison of the auxiliary replicon composition by PFGE (left panels) and cell morphology by phase contrast microscopy using a 100x oil immersion objective (right panels) of the wild-type (top) and pCY186-cured (ΔpCY186; bottom) *B. proteoclasticus*. (B) The effect of the absence of pCY186 on culture growth as measured by optical density increase at 600 nm.

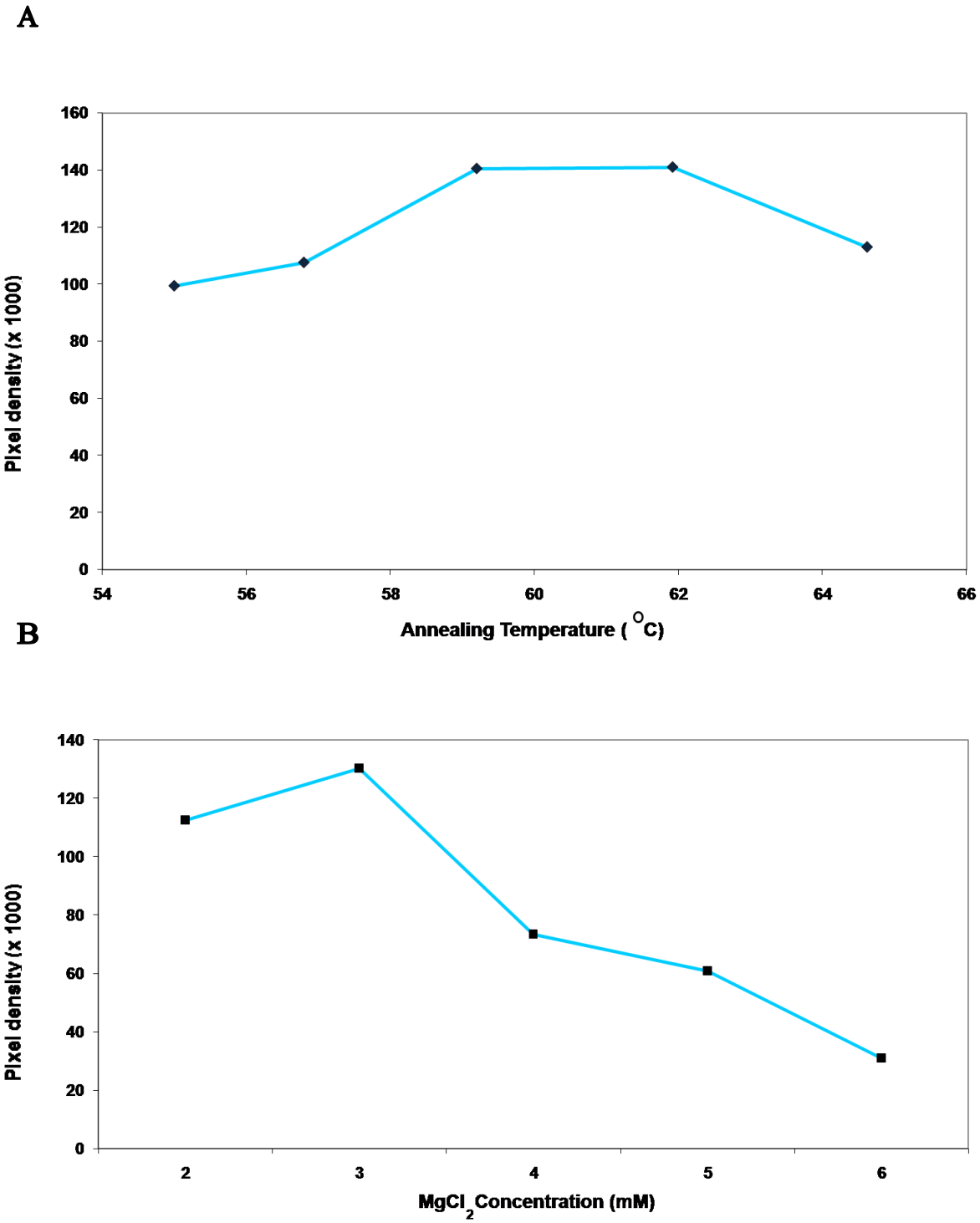


Fig. 6.4 Optimisation of qPCR reactions. The qPCR reactions were optimised for temperature (A) and $MgCl_2$ concentration (B) for amplification of p190_probe2. Optimal conditions were determined as those producing the greatest amount of the desired qPCR product determined by measuring the sum of the pixel density from the region of interest (ROI) using KODAK 1D Image analysis software.

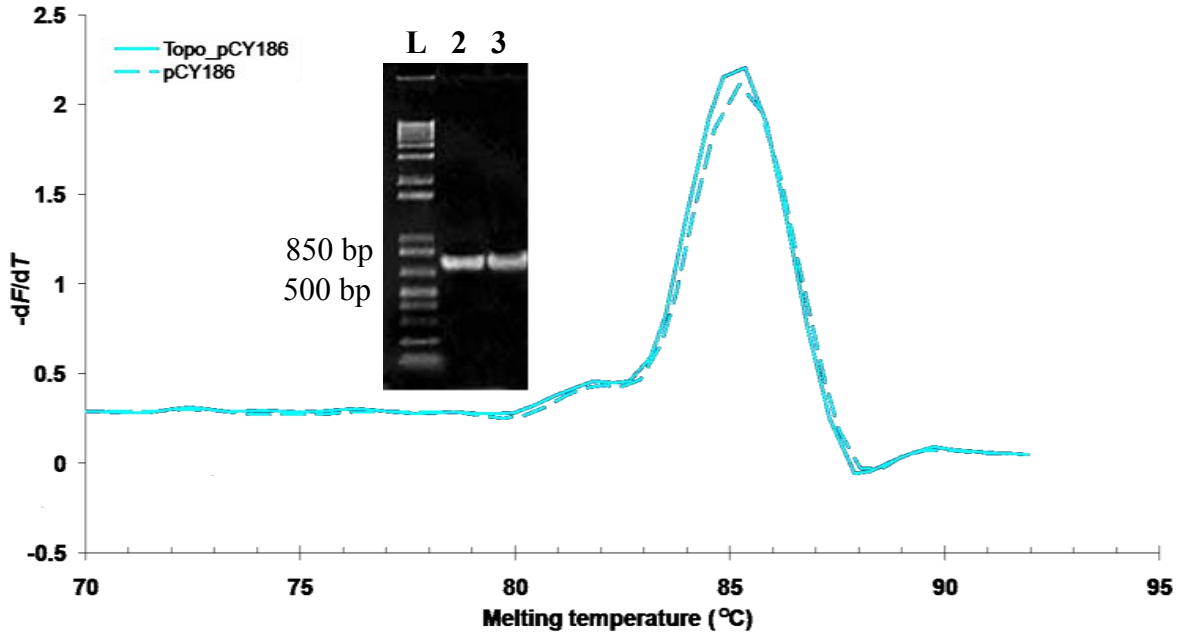


Fig. 6.5. Confirmation of qPCR amplification specificity. Melting peaks were examined for each qPCR product from both the *B. proteoclasticus* total DNA preparation (broken line) and the quantified Topo-recombinant plasmids carrying the target sequence (solid line). Inset: Lanes: L, the 1Kb+ ladder, lanes 2 and 3, 745 bp PCR-products obtained using as templates either (2) quantified Topo-recombinant plasmids carrying the target sequence or (3) the total DNA preparation.

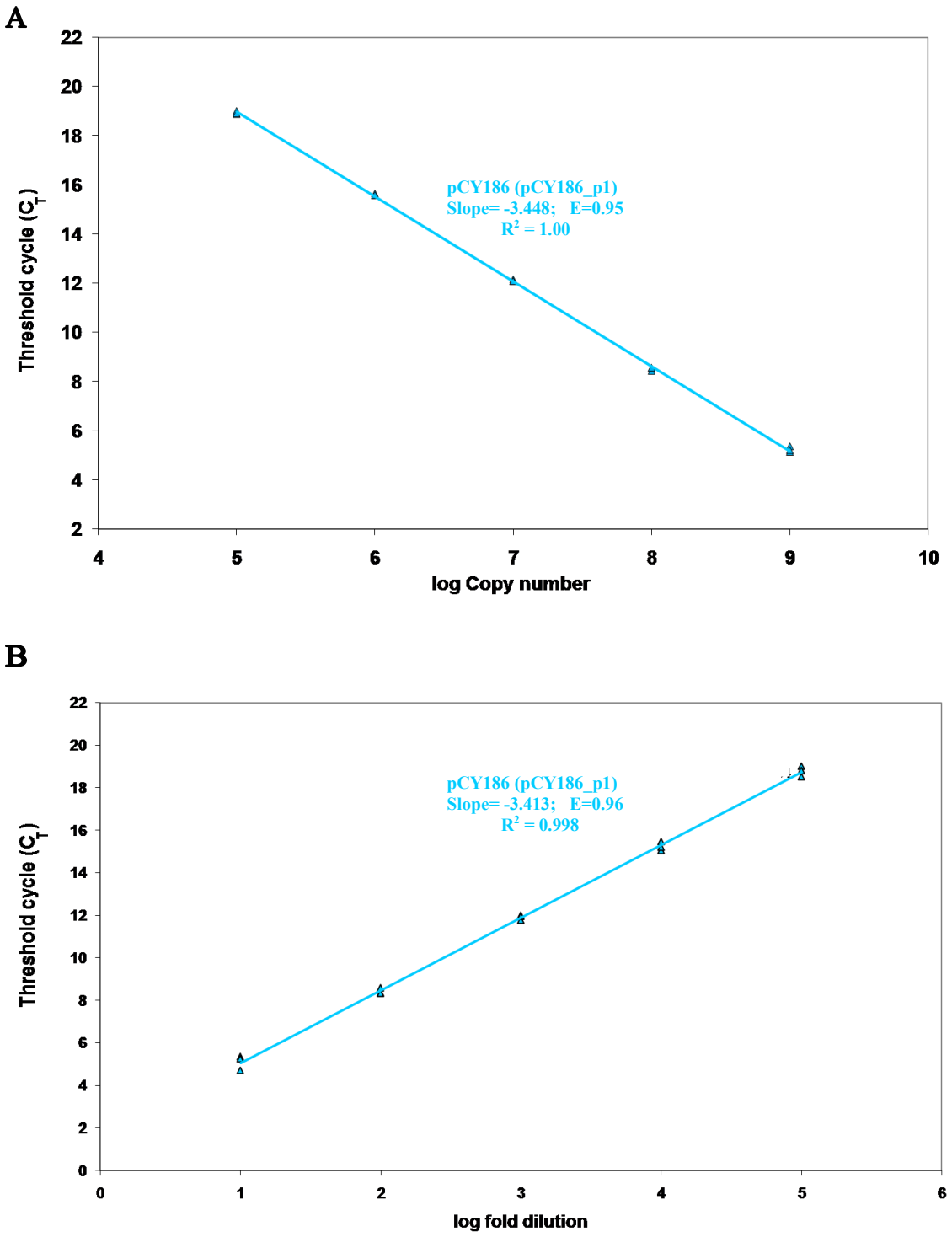


Fig. 6.6. Standard curves for the qPCR reactions. The slope of each threshold cycle vs. $\log_{10}$ of the copy number in each standard (A) or threshold cycle vs. $\log_{10}$ fold dilution are shown (B). The slope, efficiency (E), and correlation (R^2) are also shown.

6.10 Microarray analysis of pCY186 gene expression in mono-culture versus co-culture with *M. ruminantium*

Analysis of the microarray data of the mono-culture versus the co-culture, (described previously in Section 4.13), revealed 127 (85%) of the 149 putative pCY186 ORFs (identified at the time of microarray slide printing), were detected at a significant pixel intensity (9,000 > 65,000). As the microarrays were printed prior to complete closure of this replicon 49 ORFs were not included in the microarray analysis. Of the genes analysed, 12 ORFs (6%) were found to be up-regulated greater than 2-fold with a false discovery rate (FDR) less than 0.05 in the co-culture condition, while only a single hypothetical protein (ORF152) was found to be down-regulated greater than 2-fold (FDR <0.05; Table 6.2). Almost all of the ORFs found to be up-regulated in the co-culture were of unknown function (9 hypothetical proteins and 2 conserved hypothetical proteins). The only ORF that was able to be ascribed a function was ORF 15, a putative glycosyl transferase of the GT28-family.

Analysis of the distribution of differentially regulated genes around the pCY186 replicon revealed four clusters of genes that appear to be co-regulated (Fig. 6.7). The first is a cluster of seven ORFs encompassing ORFs 15 to 22 (ORF 19 being absent from the array). All seven ORFs are up-regulated above, or just below, the threshold for significance. This cluster consists of six hypothetical proteins, two that are predicted to span the membrane, and the putative GT28-family glycosyl transferase. The second cluster, encompassing ORFs 90 to 93, similarly encodes four hypothetical proteins, one of which is predicted to span the membrane. A third cluster exists that includes ORFs 187 and 190 (both hypothetical proteins), both of which show strong up-regulation in the co-culture (10.7 and 7.98 fold respectively). Unfortunately ORFs 188 and 189, two hypothetical proteins, were not identified at the time the microarray was printed and were unable to be analysed. An operon-like structure is also seen encompassing two hypothetical proteins (ORFs 168 and 169), both being up-regulated by 3 fold in the co-culture.

Table 6.2 pCY186 ORFs significantly up-regulated in co-culture

ORF	Putative function	Fold change	FDR
	UP-REGULATED IN CO-CULTURE		
187	Hypothetical protein	10.70	8.8×10^{-5}
92	Hypothetical protein	8.56	2.8×10^{-4}
190	Hypothetical protein	7.97	3.3×10^{-4}
82	Hypothetical protein	5.7	2.5×10^{-4}
18	Hypothetical transmembrane protein	4.61	1.1×10^{-3}
15	Putative glycosyl transferase GT28-family protein	3.76	2.1×10^{-2}
90	Hypothetical protein	3.46	1.3×10^{-3}
16	Conserved hypothetical protein	3.36	2.5×10^{-3}
169	Conserved hypothetical protein	3.07	2.6×10^{-2}
22	Hypothetical protein	3.02	1.2×10^{-2}
91	Hypothetical protein	2.73	6.3×10^{-4}
93	Hypothetical protein	2.35	3.6×10^{-2}
	DOWN-REGULATED IN CO-CULTURE		
152	Hypothetical protein	3.85	2.5×10^{-2}

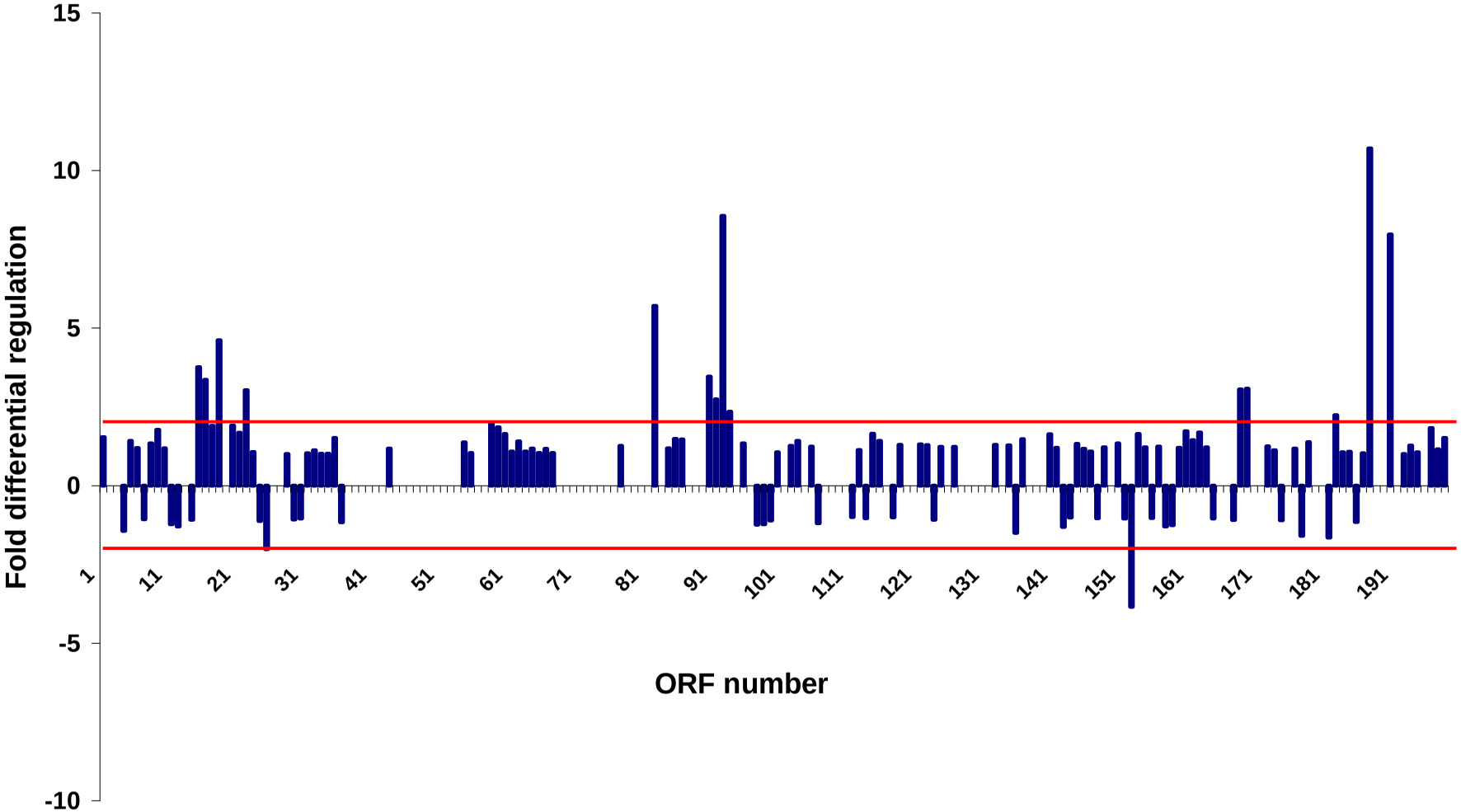


Fig. 6.7. Differential pCY186 gene regulation during co-culture with *M. ruminantium*. The histogram shows the differential regulation of gene transcripts, in order, for all 198 predicted ORFs of pCY186 as determined by microarray analysis. Red bars show the point of significance (2-fold regulatory difference). Four potential operon-like structures are evident (ORFs 15 to 22, ORFs 90 to 93, ORFs

6.11 Discussion

6.11.1 Replication of pCY186

Initially two loci were identified as potentially possessing the pCY186 *oriR*. The first locus encoded several proteins with significant similarity to chromosomal-type replication machinery. These included the chromosomal replication initiation protein, DnaA, the replicative DNA-helicase, DnaB, a complete DNA polymerase III (PolIII), PolCA, along with two additional proteins encoding PolIII α-subunits and a protein encoding the PolIII γ and τ subunits. Despite the presence of various chromosomal replication components being reported from large plasmids, to date, no plasmid or secondary chromosome has been shown to replicate via chromosomal-type replication machinery (Ng *et al*., 1998). Flanking this locus are two *rep* genes. A third *rep* gene is located in a second locus that has a similar arrangement to the described replication loci of pCY360 and BPc2. In addition to *repB* (ORF1), this locus encodes *parA* (ORF196), along with *parB* (ORF198). While both loci contained sufficient replicative machinery to facilitate the replication of pCY186, the latter was favoured based on several findings. Firstly, as with pCY360 and BPc2, an A+T rich tract was identified within the intergenic region upstream of *parA* that showed substantial sequence identity to the *oriR*s of pRJF1, pCY360 and BPc2. Secondly, pCY186 had a near perfect third position GC skew that suggested replication was initiated or terminated within this AT-rich region. Analysis of this intergenic region encompassing the pCY186 putative *oriR* identified a large number of potential secondary structure-forming DNA sequences including inverted, inverse and direct repeats along with ten *dnaA* box candidates. Copies of both direct repeat families were found within the putative *oriR* and provide potential iterons. Based on the phylogeny of its Rep protein, GC skew and the even spread of genes on both DNA strands, pCY186, as with BPc2, is predicted to replicate using a bi-directional theta mechanism.

6.12.2 DNA metabolism

Of those genes encoded by pCY186 that are able to be assigned a function, several are found to have functions in DNA metabolism, particularly in the process of DNA replication. This includes the cluster of genes described above that encode chromosomal-type replication apparatus, DnaA, DnaB, PolCA and the polymerase III α, γ and τ subunits. In addition, a hypothetical protein that shows weak similarity to a DNA primase was observed. Evidence, discussed above, suggests replication is initiated at a locus distinct to that encoding these chromosomal-type replication proteins. It also suggests the initiation of replication is mediated by a RepB protein, as is common for plasmids. However, theta-replicating plasmids, including those of *B. proteoclasticus*, also commonly encode binding sites for DnaA, suggesting a role for the chromosomal initiator in plasmid replication (Singh & Banerjee, 2007, Krasowiak *et al.*, 2006). Following RepB-mediated plasmid replication initiation, chromosomal-type replicative apparatus, like those found on pCY186, have been shown to assemble and contribute to replication (Krasowiak *et al*., 2006, Titok *et al.*, 2006, Park *et al.*, 1998, MacAllister *et al.*, 1991). The presence of these replication proteins encoded by pCY186 may alleviate the metabolic cost caused through the titration of paralogous proteins away from the major chromosome. The pCY186 replicon also uniquely encodes a putative UmuC-like DNA repair protein. UmuC, is the catalytic subunit of DNA polymerase V (Reuven *et al*., 1999). It is implicated in the replication of damaged DNA and thereby prevents the lethal effects of stalled replication. In a native setting it would be induced by the DNA recombination protein, RecA, as part of the „SOS" response. A RecA protein is encoded by the *B. proteoclasticus* major chromosome. While UmuC alone has been shown to have a limited DNA synthesis capability on undamaged DNA, its ability to complete synthesis across damaged DNA, also requires the second DNA polymerase V component, UmuD (Reuven *et al*., 1999). An adjacent ORF shows weak similarity to YolD, an uncharacterised protein-family that is believed to be functionally equivalent to UmuD (Permina *et al*., 2002). Although the componentry for the type V DNA polymerase appear to be uniquely encoded by pCY186 within the *B. proteoclasticus* genome, a second SOS response system, DNA polymerase IV (DinB) is also encoded by the major chromosome (Napolitano *et al*., 2000).

6.12.3 Restriction modification

A second group of genes appear to have roles in restriction modification. Restriction modification (RM) systems have previously been observed in several *Butyrivibrio* species (Mrazek *et al*., 2005). The majority of the RM genes found on pCY186, including *hsdMA, hsdSA, hsdSB*, and *hsdRA* and two Type II restriction endonucleases are located within a distinct cluster. Two additional genes, encoding an McrB-like restriction enzyme and a McrC-like restriction enzyme modulator protein, also have roles in RM and are found outside of this cluster. RM systems are widespread throughout prokaryotic species and recognise and prevent invasion of foreign DNA. The genes *hsdMA, hsdSA, hsdSB,* and *hsdRA* encode all the components of a Type I RM system, functioning in cognate DNA modification, foreign sequence recognition (x2) and restriction endonucleolytic digestion, respectively (reviewed Murray, 2000). The Type II restriction endonucleases appear to lack the corresponding methylation systems. While the target sequence specificity of both Type II restriction endonucleases is unclear, the potential lethality of DNA breakage makes it likely that this sequence is not, or is no longer, found within *B. proteoclasticus* DNA. Two further proteins resemble the methyl-cytosine restriction (Mcr) system that was first observed in *E. coli*. The Mcr system degrades all DNAs possessing 5-methyl-cytosine (Raleigh & Wilson, 1986). The RM complement of pCY186 accounts for the vast majority of the identifiable restriction systems in *B. proteoclasticus*. This RM system is therefore likely to fulfil an important role *in-vivo* in preventing the invasion of foreign DNA, such as from phage, which are known to be abundant in the rumen (Klieve & Swain, 1993).

6.11.4 ORFs from the Bacterial Minimal Gene Set

As with pCY360 and BPc2, pCY186 encodes ORFs described within the Bacterial Minimal Gene Set (Koonin, 2000). These include several genes of the chromosomal-type replication apparatus described above (*dnaA, dnaB*, and two DNA polIII α) along with a single-stranded DNA-binding protein, *ssb,* an AAA+ ATPase and a zinc metalloprotease, *ftsH* (Table 6.2). It also includes a DNA primase which shows only weak similarity to known DNA primases. Genes encoding Ssb and FtsH are also located on pCY360. AAA+ ATPases are a large, functionally diverse protein-family that exerts their activity through the energy-dependent unfolding of macromolecules.

As with pCY360, none of the encoded proteins of pCY186 genes described in the Bacterial Minimal Gene Set are unique to the *B. proteoclasticus* genome.

6.11.5 Transposases

The strong sequence identity observed between the transposase genes of pCY186 and those found upon other *B. proteoclasticus* replicons suggests they may still be active. While low sequence similarity indicates that genes have not moved recently, or are no longer being transferred to pCY360 by transposon-mediated gene shuttling, the strong sequence identities seen between transposases located upon pCY186 and the two chromosomes suggests that such transposon-mediated gene shuttling may still occur in these replicons (Amabile-Cuevas & Chicurel, 1992).

6.11.6 Dispensability

The finding that pCY186 was easily lost *in-vitro*, even in the absence of a curing agent suggests it was either transiently present in *B. proteoclasticus* at the time of its isolation or it provides a function or functions required *in-vivo* that aren't required *in-vitro*. The preceding hypothesis is unlikely given the significant identity of its replicative machinery to BPc2 and pCY360, along with its %G+C and codon usage which match the major chromosome almost perfectly. Collectively these findings suggest pCY186 has resided with *B. proteoclasticus* for a significant period of time. The pCY186-endoded RM systems may provide protection of genomic DNA from invasion by foreign nucleic acid in the rumen, but absent *in-vitro*. Morphological observations of *B. proteoclasticus* ΔpCY186 show it almost exclusively forms long chains. The loss of pCY186 suggests that, *in-vitro,* the replicon is a burden to *B. proteoclasticus*. However, no statistically significant difference in the growth rate of *B. proteoclasticus* was observed in the presence or absence of pCY186.

6.11.7 Microarray analysis

Due to the sequencing difficulties that resulted in the under-representation of various regions of pCY186, only 75% (149) of the 198 ORFs ultimately assigned to pCY186 were utilised in the construction of the microarray. Of these, 85% (127) were detected at intensities significantly greater than the background suggesting they were being transcribed and consequently supporting their status as genuine coding sequences. Of these, 13 were found to be differentially regulated between the mono-and co-culture conditions. The twelve ORFs found to be up-regulated were, with the exception of a gene encoding a-family 28 glycosyl transferase, all of unknown function. This included two conserved hypothetical proteins (ORFs 16 and 169) that have been identified in several other species of the phylum Firmicutes, mostly of the orders *Bacillales* and *Clostridiales*. Glycosyl Transferase-family 28 (GT28) proteins catalyse various di-, oligo-, or poly-saccharide biosynthetic reactions, and characterised GT28 proteins include a 1,2-diacylglycerol 3-β-galactosyltransferase, 1,2-diacylglycerol 3-β-glucosyltransferase and β-N-acetylglucosamine transferase. Likewise, the only down-regulated ORF has an unknown function. Four clusters of co-regulated ORFs were evident, the first encompassing seven ORFs including the GT28 protein and six hypothetical proteins, two that are predicted to span the membrane. The co-regulation of these ORFs with the GT28-family protein may suggest a role in exopolysaccharide production and transport across the membrane. The other three gene clusters all encompass genes of unknown function.

6.12 Summary

The pCY186 replicon is one of two, largely cryptic, megaplasmids found in *B. proteoclasticus*. pCY186 is the only replicon of the four found in *B. proteoclasticus* that can be displaced *in-vitro* without the loss of cell viability. Several pieces of evidence suggest this replicon is not transient but rather has co-resided with the host chromosome for a significant period of time. This would imply its function is specific to, and important within its natural environment. One potential explanation for this is its possession of RM systems that may be important to *B. proteoclasticus* in the defence against invasion by ruminal phage.

7 General discussion, conclusions and future directions

7.1 Introduction

The term „megaplasmid" was first coined in 1981 by Rosenburg *et al*. to describe autonomously replicating plasmid molecules greater than 100 Kb in size. Since then more than 200 megaplasmids have been described in over 30 genera of mostly Gram-negative bacteria. In some cases the term megaplasmid has been replaced with miniature, secondary, or alternate-chromosome due to the presence of indispensable or distinctly chromosomal elements within these replicons (Bartosik *et al*., 2002, Ng *et al*., 1998, Suwanto & Kaplan, 1989). The genome of the Gram-positive, rumen-inhabiting bacterium *B. proteoclasticus* B316^T consists of 4 replicons that were found to consist of a 3.5 Mb major chromosome, a secondary chromosome of 302 Kb and two megaplasmids of 186 Kb and 361 Kb. The presence of large auxiliary replicons had previously been observed in a number of *Butyrivibrio* strains (Teather, 1982), however, at the time, their large sizes precluded detailed analysis of their structures or functions. The sequencing of the *B. proteoclasticus* genome provided an ideal platform to investigate the respective gene complements, replication mechanisms, copy numbers and dispensability of three examples of these large *Butyrivibrio* plasmids. Through various characterisation techniques this work has attempted to determine the contribution of each replicon to the biological capacity of *B. proteoclasticus*. This work was able to build on the first observation of these large replicons, providing a more comprehensive investigation into the distribution of large auxiliary replicons in *Butyrivibrio* and *Pseudobutyrivibrio* species. Further, by probing PFGE gels for the pCY360 replication initiation protein or ribosomal RNA operons we were able to determine the relationship of the *B. proteoclasticus* replicons to auxiliary replicons from other *Butyrivibrio* species.

7.2 Comparing replication machinery

The three auxiliary replicons of *B. proteoclasticus* were found to encode similar replication machinery, regardless of their status as a megaplasmid or secondary chromosome. Comparison of the composition of their replicative machineries identified four elements that appear to be cognate to the replicative apparatus of *B. proteoclasticus* replicons (Figure 7.1). This included the *oriR* (physical origin of replication), ORFs encoding replication and plasmid partitioning proteins RepB and ParA, respectively and a hypothetical protein conserved in all three auxiliary replicons. In pCY186 this protein contained a predicted AT hook-like motif, in pCY360 this motif was found to be below the trusted threshold and it was absent from the BPc2 homologue. Nevertheless, this hypothetical protein showed 22 and 32% amino acid sequence identity to those found associated with pCY360 and pCY186 respectively. AT hook motifs are thought to promote binding to DNA, preferentially at AT-rich regions, such as those found within the *oriR* (Aravind & Landsman, 1998). The deduced amino acid sequence of the RepB protein was well conserved (36 to 54% aa identity; Table 7.1), as were the *oriR*s, which shared 55 to 61% nucleotide identity (Table 7.1). The ParA sequences were less conserved (26 to 38% amino acid identity; Table 7.1). The *oriR* and the adjacent ORFs showed partial synteny. The gene encoding the hypothetical protein was always found directly downstream of *parA*, while the *oriR* was always located directly upstream of *parA*. The ORFs surrounding the *oriR*s of pCY360 and BPc2 were found to be divergently transcribed, while those of pCY186 were not. In all cases the intergenic region encompassing the *oriR* contained *dnaA* box candidates and sequences capable of forming secondary structures. However, their number, type and location relevant to the *oriR* and *repB* varied. Additionally the nucleotide sequences of repeat units were not conserved between replicons. These differences were expected given that plasmid incompatibility precludes two replicons with near-identical replicative machinery from co-existing in the same cell, particularly with low copy number replicons (reviewed in Section 1.12).

Previous studies have found that the cognate Rep protein cleaves a distinct TA di-nucleotide pair within the *oriR*. Secondary structure analysis of the region surrounding this di-nucleotide suggests it occurs within an unpaired region of a terminal, or internal loop of a DNA hairpin stem-loop structure (Koepsel *et al*., 1985, Thomas *et al*., 1990, Grohmann *et al*., 1998, Noirot-Gros *et al*., 1994, Puyet *et al*., 1988). The *oriR*s of the three auxiliary replicons from *B. proteoclasticus* along with other theta replicating *Butyrivibrio* plasmids, pRJF1 and pRJF2, were aligned using Dialign to identify potential *nic* sites. Two potential *nic* sites, each encoding a TA di-nucleotide pair, were found to be conserved throughout all plasmids tested (pCY360 nt 98 – 99 and nt 326 - 327; BPc2 nt 111 – 112 and nt 352 - 353; pCY186 nt 143 – 144 and nt 383 - 384; pRJF1 nt 44 – 45 and nt 284 -285; pRJF2 nt 43 – 44 and nt 286 - 287; Figure 7.2). Secondary structure predictions of all *oriR*s found both potential *nic* sites to occur in unpaired loop regions of each plasmid except pCY360. However structural analysis of each replicon suggests two alternate *nic* sites for pCY360 occurring within similarly-located unpaired stem-loop regions (Figure 7.2). These differences in the pCY360 *oriR* coincide with differences observed or predicted in its replication mechanism. The pCY360 replicon is predicted to replicate in a unidirectional theta mechanism and it was determined to exist at a copy number of approximately 5 per chromosome. On the other hand, BPc2 and pCY186 are both predicted to replicate in a bi-directional theta mechanism, and both are present at approximately 1 copy per chromosome.

Given the low copy number of each *B. proteoclasticus* replicon, an effective partitioning system would be essential for equal segregation of replicons to daughter cells. Aside from the major chromosome, pCY186 is the only replicon to encode the ParB component of the partitioning complex. Similar observations have been made in other genomes containing multiple replicons (McLeod *et al*., 2006, Jerke *et al*., 2008). The mechanism through which ParA and ParB confer equal partitioning to mother and daughter cells is not fully elucidated. ParA is an ATPase that has been shown to regulate expression of the *par* operon in characterised systems (Bouet and Funnell, 1999). ParA interacts with the other components of the partition complex, where it is thought the protein acts to move plasmid copies from the centre of the cell, where the division septum forms, toward the termini through an oscillating motion between mid-cell and quarter-cell positions (Bouet

and Funnell, 1999; Davis *et al*., 1996; Li *et al*., 2004; Youngren and Austin, 1997). ParB is known to bind to DNA at a centromere-like region known as *parS* (Austin & Abeles, 1983, Funnell & Gagnier, 1993). Both partitioning components have been found to be essential to the partitioning process in bacterial systems that have been tested (Friedman & Austin, 1988, Ogura & Hiraga, 1983). If ParB is essential to the equal partitioning of *B. proteoclasticus* replicons, deficient replicons may be complemented in trans by the pCY186-encoded ParB, as has been previously proposed in other genomes containing multiple replicons (Jerke *et al*., 2008, McLeod *et al*., 2006). The capacity of *B. proteoclasticus* to replicate pCY360 and BPc2 following curing of the pCY186 replicon may be due to a compensatory effect elicited by the chromosomally-encoded ParB. The chromosomally-encoded ParB may be sufficient to allow effective partitioning of two of the three auxiliary replicons, while a third replicon may require a second *parB* homologue. Alternatively the predominantly filamentous growth phenotype observed in strains cured of pCY186 may be due to an impaired segregational ability of the organism caused by the absence of the pCY186-encoded ParB. Filamentous grown has previously been observed in *Caulobacter crescentus* following the depletion of ParB (Mohl *et al*., 2001).

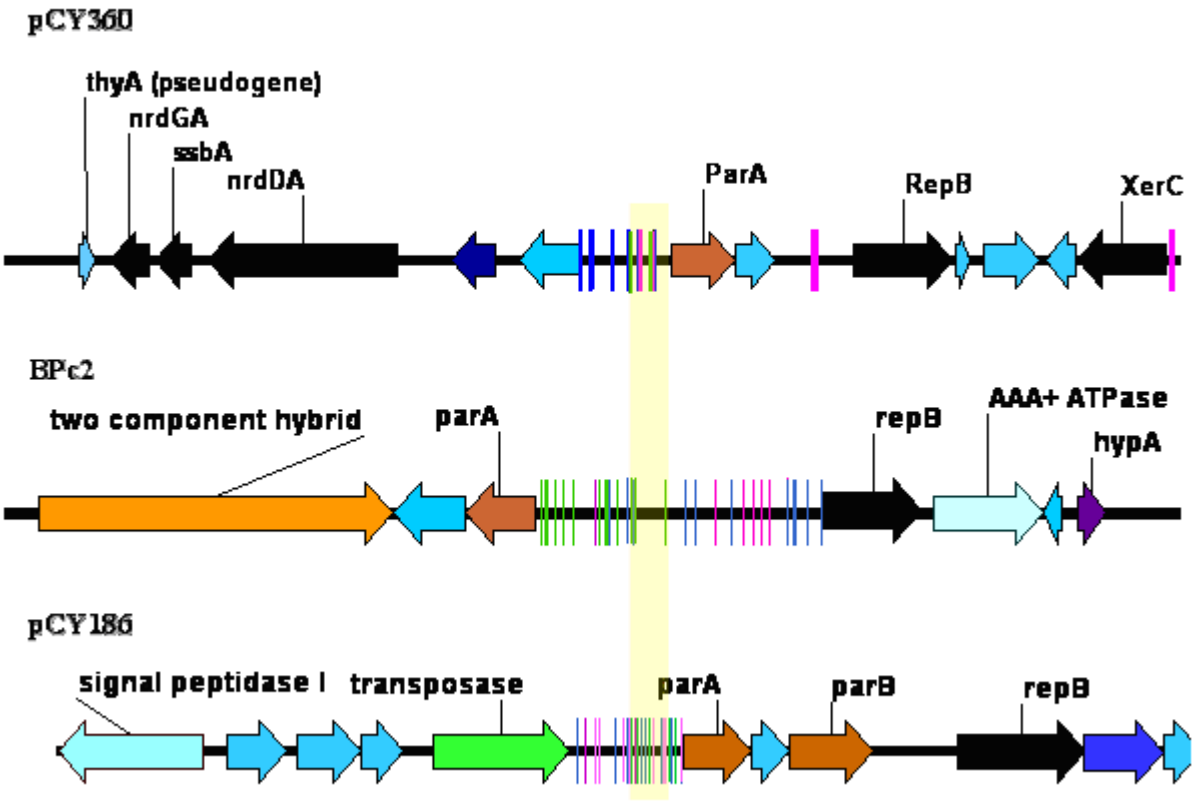


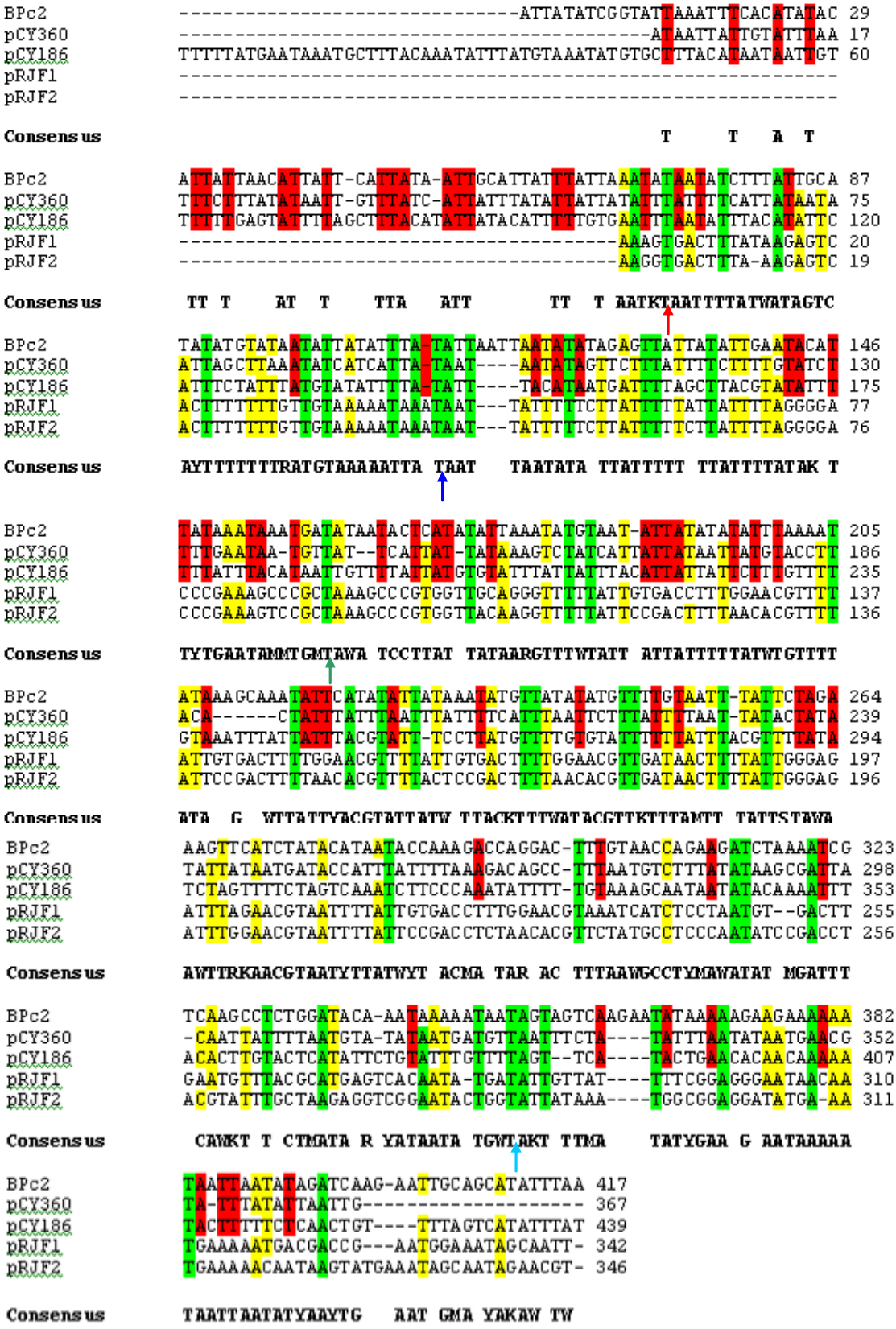
Figure 7.1. Comparison of the replication loci of the *B. proteoclasticus* auxiliary replicons. The replication loci of each of the *B. proteoclasticus*" auxiliary replicons are shown with their cognate *oriR*s in the centre (highlighted yellow). ORFs located within 7 Kb up- or down- stream of the *oriR* are colour coded by function and shown along with *dnaA* boxes (blue), direct repeats (green), inverse repeats (purple) and inverted repeats (pink). Maps were generated with Vector NTI (InforMax, 2001)

Table 7.1 Sequence identity of replicative machinery.

Orthologue in:		
pCY360	**BPc2**	**pCY186**
RepB	54%*	36%
ParA	26%	38%
Hypothetical protein	22%	29%
oriR	56%	61%
	RepB	43%
	ParA	28%
	Hypothetical protein	32%
	oriR	55%

*Sequence identity is based on amino acid identity (except the *oriR* which is nucleotide identity) as determined by Clustal alignments.

A



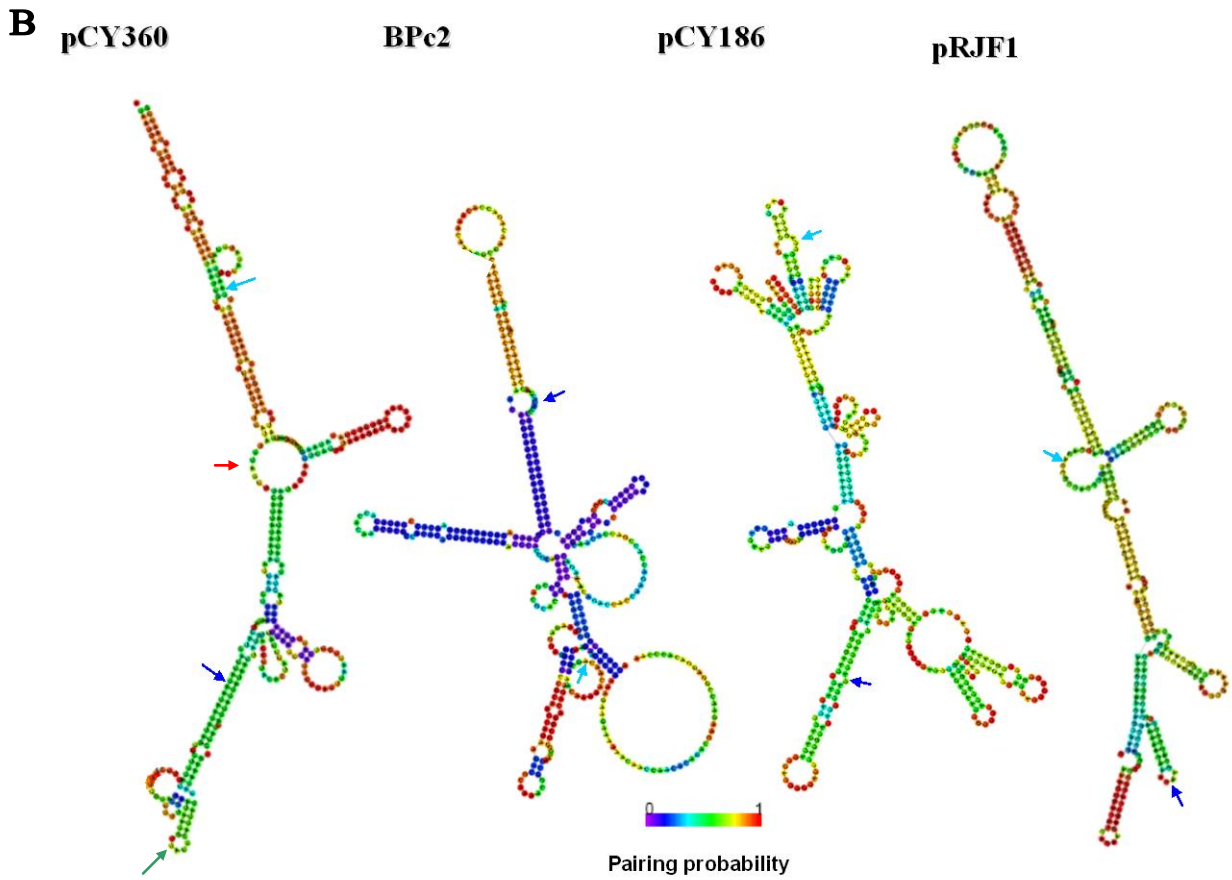


Figure 7.2. Comparison of replication origins. The oriR of pCY360, BPc2, pCY186, pRJF1 and pRJF2 were aligned using ClustalX (A). Nucleotides were colour-coded as follows: absolutely conserved (green); strongly conserved (yellow); conserved in all *B. proteoclasticus* replicons but not pRJF1 or pRJF2 (red). Secondary structures of the *oriR*s of each replicon were predicted using RNAfold (pRJF2 was near identical to pRJF1 and is not shown; B). Paired-nucleotides are colour coded according to the legend by the probability that the nucleotides pair as predicted. Potential *nic* sites are indicated by a dark blue or light blue arrow in both A and B. Structurally predicted *nic* sites for pCY360 are shown as a red or green arrow.

7.3 Evolutionary aspects

There are four conceivable pathways leading to the formation or acquisition of an auxiliary replicon the size of those observed in *B. proteoclasticus* B316^{T}: i) The replicon could have been acquired horizontally in its current form; ii) through the concatenation of several smaller plasmids; iii) derived from a smaller plasmid that grew through the acquisition of genes from the host chromosome, or from the extracellular environment; or iv) split from a larger ancestral chromosome following the acquisition of apparatus allowing autonomous replication. There is evidence in support of several of these mechanisms seen in the composition of each of the three auxiliary replicons.

The potential conjugative ability of pCY360 suggests it may have been acquired as a megaplasmid. The rumen is known to be a favourable environment for conjugative transfer of plasmids (Mizan *et al*., 2002). Conjugative elements, including plasmids, larger than pCY360 have been shown to be transferable (Hogrefe *et al*., 1984, Schwartz *et al*., 2003, Sullivan *et al*., 2002). In pCY186 the three ORFs that best resemble plasmid replication initiation proteins may be remnants of the replicative apparatus of several concatenated plasmids. Other evidence points to the *B. proteoclasticus* replicons originating as smaller entities. The predicted secondary structure of the pCY360 *oriR* markedly resembles the *oriR*s of the smaller *Butyrivibrio* plasmids pRJF1 and pRJF2. Further, the nucleotide sequence of the *oriR*s of each of the *B. proteoclasticus* auxiliary replicons, along with their cognate Rep proteins, show a strong to moderate degree of nucleotide or amino acid sequence identity respectively, arguing in favour of a similar evolutionary origin. As previously discussed transposon transposition is a common method for gene shuttling within or between replicons (Amabile-Cuevas & Chicurel, 1992). The potential for acquisition of genetic material from the chromosome is found in the presence of IS110-family and IS1182-family transposases that have identical or near identical sequences on the major and secondary chromosome as well as on pCY186. IS4-family, IS200-family and IS605-family transposases with low sequence identity are also found distributed among the major and secondary chromosomes and pCY360 suggesting gene shuttling may have also occurred between these replicons in the past. DNA may also have been acquired directly from the external environment via natural transformation. The high bacterial population density in the rumen and the tendency for rumen bacteria to

congregate on surfaces creates conditions favourable for gene transfer events. The presence of a chromosomal-type replication apparatus upon pCY186, rRNA operons and other chromosomal-features of BPc2 and members of the Bacterial Minimal Gene Set upon all replicons could be rationalised either by the splitting of a larger ancestral chromosome to form these replicons, or by gene shuttling from the chromosome. In one of the Southern blot experiments (Figure 4.22) the *repB* probe was seen to hybridise to the chromosomal component of most species within the *Butyrivibrio-Pseudobutyrivibrio* assemblage. This suggests a gene encoding a RepB protein is found upon many *Butyrivibrio* and *Pseudobutyrivibrio* chromosomes and provides a potential mechanism for autonomous replication of DNA split from the major chromosome. Collectively, there is insufficient evidence to propose a single mechanism for the formation of these auxiliary replicons.

Regardless of the mechanism leading to their formation the *B. proteoclasticus* replicons appear to have diverged from one another some time ago as evidenced by the small number of significant similarities seen between the gene complements of each replicon. Further, those genes that do show any level of similarity (with exception to the proposed replicative machinery surrounding the *oriR*) do not appear contiguous within the replicons (Figure 7.3).

7.4 Contributions to the biology of *B. proteoclasticus*

Bioinformatic analysis of the BPc2 DNA sequence show that this replicon has evolved chromosomal qualities including rRNA operons, unique copies of a number of enzymes involved in the metabolism of carbohydrate, nitrogen, fatty acids and various co-factors. This includes two enzymes encoded by genes of the Bacterial Minimal Gene Set. Essential metabolic functions provided by BPc2 are the uptake of biotin, the uptake and biosynthesis of nicotinamide adenine mononucleotide and the potential to utilise fumarate as the terminal electron acceptor during anaerobic respiration. Additionally, BPc2 appears to encode significant detoxification functions. These significant contributions made by BPc2 would account for the inability to derive *B. proteoclasticus* mutants devoid of this replicon.

The two other auxiliary replicons, pCY360 and pCY186 have both evolved as megaplasmids. The larger pCY360 could not be cured from *B. proteoclasticus*, suggesting that it contributes an essential function to the organism. However, this function could not be associated with any of the ORFs identified by bioinformatic analysis of this plasmid, particularly because no chromosomal-like features were observed. In contrast to pCY360, the smaller megaplasmid pCY186 was easily cured. Its presence in rumen isolates of *B. proteoclasticus* B316^{T} and apparent long association with *B. proteoclasticus* (as suggested by codon usage analysis) suggest pCY186 provides a function advantageous in the rumen environment. This could include protection against invading nucleic acid but also effective partitioning of the nucleoid. Both pCY360 and pCY186 are largely cryptic, with 76% and 63%, respectively, of their gene complements (a total of 421 genes) appearing novel within the sequence databases. These results are consistent with the findings from other characterised plasmids derived from *Butyrivibrio* species, which are cryptic with regard to function (Hefford *et al*., 1997, Hefford *et al*., 1993, Kobayashi *et al*., 1995). The compositional differences between the megaplasmids and chromosomes of *B. proteoclasticus* become apparent when comparing COG-defined functional distributions (Figure 7.4). An attempt was made to infer the function of the hypothetical proteins identified in this work. A hypothesis was formed that the proteins encoded by the auxiliary replicons of *B. proteoclasticus* contributed to interspecies interactions. This hypothesis was based on the large number of ORFs encoding proteins predicted to impact the bacterial membrane, being membrane-associated or secreted into the extracellular environment of *B. proteoclasticus*. This hypothesis led to the investigation of interspecies interactions between *B. proteoclasticus* and the methanogen *M. ruminantium* using microarray analysis. The analysis of microarray data for the *B. proteoclasticus* genes located on auxiliary replicons identified 30 up-regulated genes encoding proteins of unknown function during co-culture with *M. ruminantium* (25 hypothetical proteins, 5 conserved hypothetical proteins). The broad functional categories to which the up-regulated genes belong indicate that the encoded proteins may have roles in the following processes: i) formation of cell co-aggregates (through exopolysaccharide production or flagellar formation); ii) the uptake of glutamate or its subsequent conversion to NAD; iii) the uptake of PTS-sugars or their conversion to, or storage as, glycogen; iv) RNA catabolism. Some of the up-

regulated genes may also have roles in poorly defined or novel processes, such as the interspecies transfer of hydrogen, formate or CO_2. Five genes from the auxiliary replicons encoding proteins of unknown function (4 hypothetical proteins and 1 conserved hypothetical protein) were found to be down-regulated. These similarly may have roles in polysaccharide degradation, sugar uptake or fatty acid metabolism, consistent with down-regulated genes that have an annotated function.

7.5 Distribution of auxiliary replicons among species within the *Butyrivibrio/ Pseudobutyrivibrio* assemblage

Using PFGE, the genomic architecture of 14 strains representing each species of the *Butyrivibrio – Pseudobutyrivibrio* assemblage was evaluated. Large extrachromosomal elements were found to be common among this group of bacteria, with representatives observed in each genus. However there appears to be no phylogenetic correlation with regard to their size or the extent of their distribution. Teather (1982) concluded, through the analysis of restriction enzyme digests of the large replicons in *B. fibrisolvens* strains ATCC 19171, OR383, OR391 and OR399, that they were likely to be of similar, though not identical composition. In contrast, the work presented in this thesis demonstrated that the three auxiliary replicons of *B. proteoclasticus* show very distinct gene complements. Further, the replication initiation protein used by the auxiliary replicons of *B. proteoclasticus* appear to be distinct from those utilised by 9 of the 10 large replicons identified in other species of the *Butyrivibrio / Pseudobutyrivibrio* assemblage, despite the co-resident auxiliary replicons of *B. proteoclasticus* sharing reasonable sequence identity. A 145 Kb DNA band from *P. ruminis* DSM9787 was the only large replicon that shares sufficient identity to the pCY360 *repB* gene, as determined by Southern blot analysis. Therefore it seems likely the three auxiliary replicons of *B. proteoclasticus* and the 145 Kb replicon from *P. ruminis* comprise a single related group of *Butyrivibrio* plasmids distinct from the others observed.

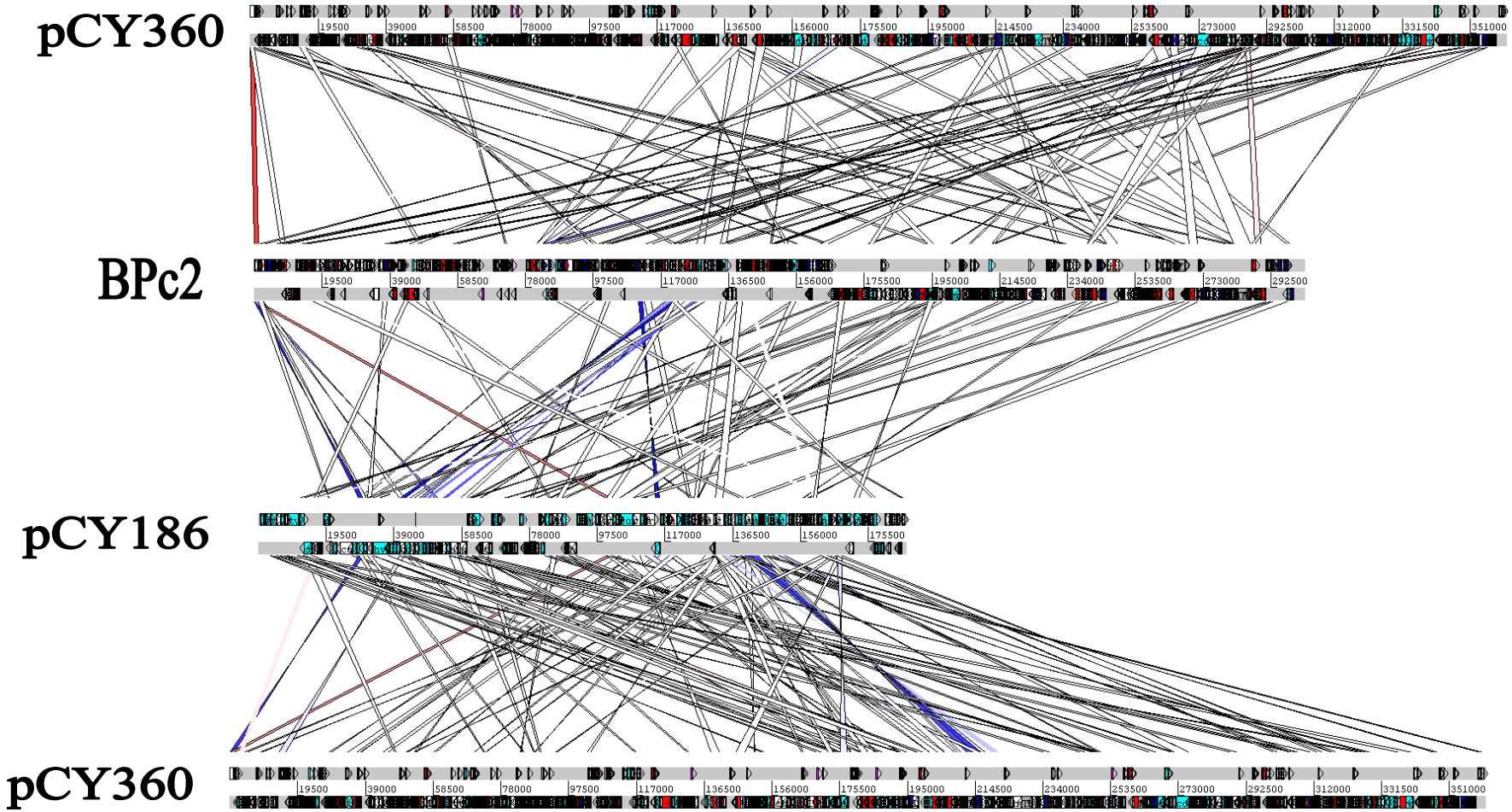


Figure 7.3. ACT comparison of gene synteny of *B. proteoclasticus*' auxiliary replicons. Lines show the position of matches in each replicon. Matches are colour coded to indicate the strength of match: red = strong same orientation, blue = strong reverse orientation. The intensity of the colours ranges from white (bordered by black lines; very weak match) to red or blue (strong match). Comparisons were made using the software compACTor and displayed using ACT. Note the pCY186 replicon is not orientated in this analysis and the rep gene is located in the centre of the replicon.

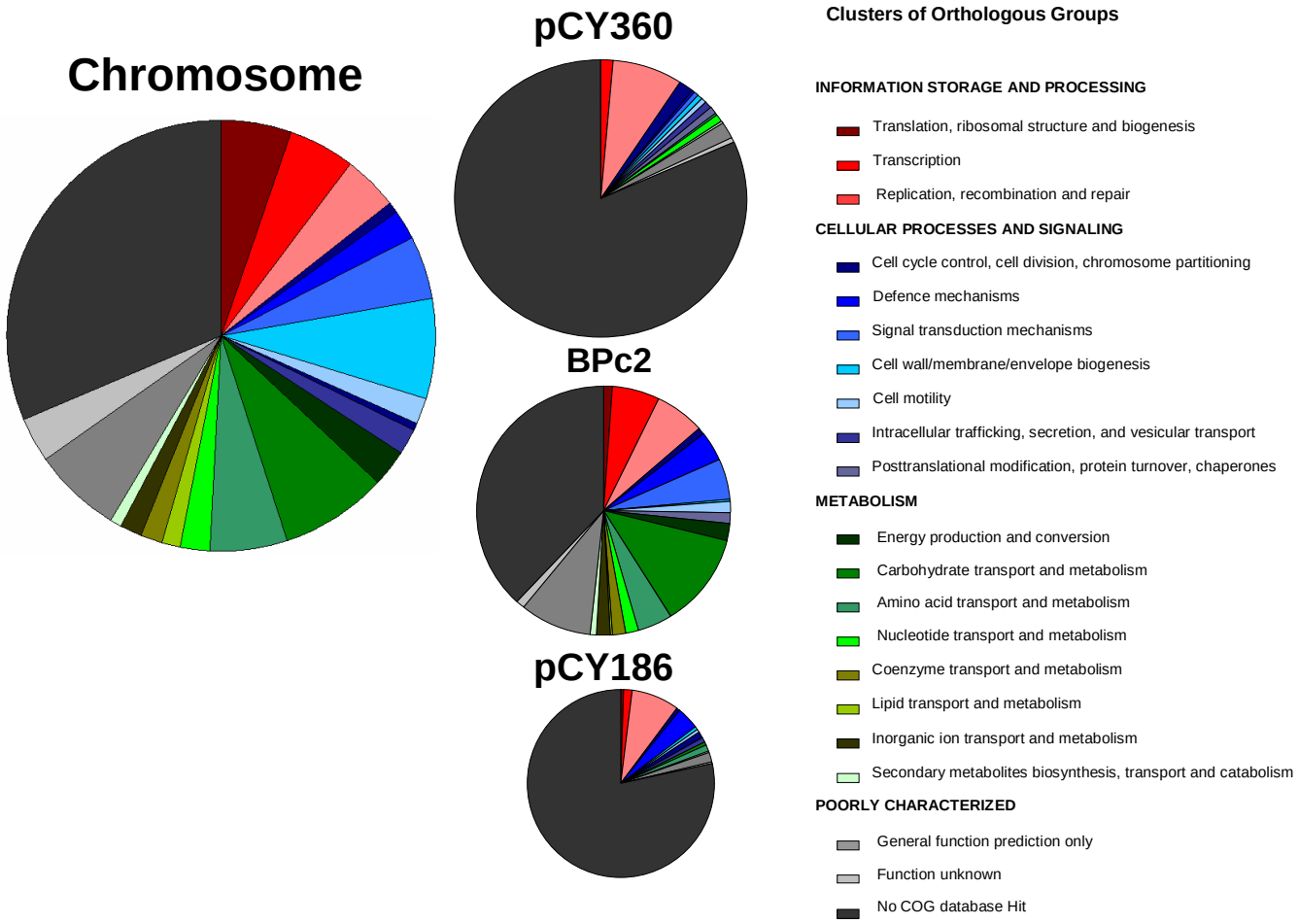


Figure 7.4. COG category distribution. The gene content of each replicon was assigned to COG functional categories. The proportion of COG matches in each functional category is presented as the pie-charts.

7.6 Looking Forward

The logical extension of the work reported in this thesis is the development of an effective genetic system for gene transfer into, and mutagenesis of, *B. proteoclasticus*. Attempts to develop systems for the genetic manipulation of ruminal bacteria have been ongoing for a little over two decades following the recognised potential for improvement of rumen and consequently animal performance through manipulation of rumen microbes (Mackie & White, 1990). The pCY186 replicon appears to be the best candidate on which to base the construction of a suitable shuttle vector for two reasons: i) it can be eliminated from the cell *in-vitro* without a reduction in fitness; ii) it possesses the majority of the RM systems encoded by *B. proteoclasticus* therefore its absence would remove a significant transformation barrier. A brief attempt to construct such a vector was unsuccessful. This is consistent with other attempts to introduce described *Butyrivibrio*- (Beard *et al.*, 1995, Hefford *et al.*, 1997) or *Clostridium*- (Lyras & Rood, 1998) shuttle vectors into *B. proteoclasticus* which have been unsuccessful. Difficulties have also been reported in the transformation of other *Butyrivibrio* species with *Butyrivibrio-E. coli* shuttle vectors (Mann *et al*., 1986). Collectively this suggests significant barriers are present that prevent the establishment of foreign DNA in *Butyrivibrio* species. RM systems are typically regarded as the primary defence against invading or introduced DNAs. Aside from RM systems, non-specific nucleases are also likely to be involved. Non-specific nucleases have previously been reported to inhibit transformation in *Vibrio cholerae* (Focareta & Manning, 1991). The *B. proteoclasticus* genome has more than 20 potential endo- and exo- nucleases, including the thermostable nuclease encoded by pCY360, which are predicted to operate both intra- and extra-cellularly. Recently the use of *in-vitro* methylation of vectors prior to their transformation has been shown to successfully overcome these barriers (Nierop Groot *et al*., 2008, Chen *et al*., 2008, Accetto *et al*., 2005). Future efforts may look to use an alternative second host such as *Clostridium perfringens* which is more closely related than *E. coli* and capable of anaerobic growth. A future vector may also require the incorporation of genetic elements other than *repB* and the *oriR* to successfully replicate and be retained by *B. proteoclasticus*, such as the *parA-parB* machinery to overcome potential partitioning problems associated with the vectors low copy number. Future efforts to develop a genetic transformation system for *B. proteoclasticus* would be of significant value and

would allow a much wider range of functional genomic analyses to be carried out using the completed genome sequence.

A number of interesting observations were made during the course of this thesis that warrant further attention.

Analysis of the enzymatic pathways of *B. proteoclasticus* suggests the organism is incapable of de-novo biosynthesis of biotin and NAD and instead relies on the BPc2-encoded systems to import exogenous biotin and nicotinamide riboside along with the systems to metabolise the latter through to the essential cofactor NAD. Establishing these requirements would be useful to eliminate the dependence on products of poorly defined composition (such as rumen fluid or yeast extract) from media for future experimentation where a defined media is required.

BPc2 may also allow *B. proteoclasticus* to utilise fumarate as the terminal electron acceptor for anaerobic respiration when more electronegative terminal electron acceptors, such as nitrate, sulphate or dimethyl sulphoxide, are in limited supply.

During attempts to cure *B. proteoclasticus* of each of the auxiliary replicons it was observed that the organism was capable of growth in media containing 5% (w/v) SDS. Recently the ability to resist high concentrations of biocides, including SDS, was observed in a number bacteria isolated from a dental unit water line (Liaqat & Sabri, 2008). Their resistance mechanism appeared to be related to an increase in diaminopimelic acid in the cell wall. It would be interesting to determine if a similar resistance mechanism operates in *B. proteoclasticus*, and the functional relevance of this type of resistance to life in the rumen

Following curing of the pCY186 megaplasmid, morphological analysis of *B. proteoclasticus* (ΔpCY186) has shown that it occurs almost exclusively in long chains. This is indicative of problems with cell division or cell separation. Possible reasons include a defective nucleoid partitioning system. The investigation of the exact effects of the loss of pCY186 is warranted as this may determine the role of some (or all) of the genes with unknown function and help identify the overall contribution of this replicon to the biology of *B. proteoclasticus*.

Finally, bacterial genome sequencing is unveiling a large number of genes that appear unique or are of unknown function. The sequencing and analysis of the auxiliary replicons of *B. proteoclasticus* has resulted in a further 503 novel genes along with 83 previously observed genes of unknown function. Transcriptional and occasionally proteomic evidence suggests most of these genes are expressed. One challenge that has yet to be overcome in this, and other bacterial genome sequencing projects, is ascribing function to such genes. A complete understanding of *B. proteoclasticus* physiology and metabolic capability depends on unravelling the functions encoded by its functionally unknown gene complement. Several bioinformatic methods, such as predicting protein function from high-resolution structural analysis (Hermann *et al*., 2007), are being developed to help address this problem, while gene knockouts can be used in organisms with developed genetic systems to look for phenotypic differences.

However, like all information derived from the *B. proteoclasticus* genome sequence, it needs to be linked to proteomic data and put into an environmental context. Microarray investigation into the effects on genetic pathways during the interspecies interaction between *B. proteoclasticus* and *M. ruminantium* revealed a potentially symbiotic interaction between the two species. It also revealed significant changes in *B. proteoclasticus* metabolism, through the down-regulation of many GH-family enzymes and sugar uptake systems and up-regulation of PTS transport systems. This suggests that the manner in which *B. proteoclasticus* acts in a pure culture is different to how it acts in its natural setting, where it is likely to interact with a large consortia of other bacteria, archaea, fungi, phage and protozoa. Similarly, other studies of interspecies interactions illustrate differences in *in-vitro* experimentation using isolated cultures compared to what is occurring *in vivo* (Johnson *et al*., 2006). Therefore the challenge for the future is to link the *B. proteoclasticus* genome sequence to functional data derived from *in-vivo* experiments or experiments that mimic conditions that are relevant to the rumen.

Appendix I Gram stain

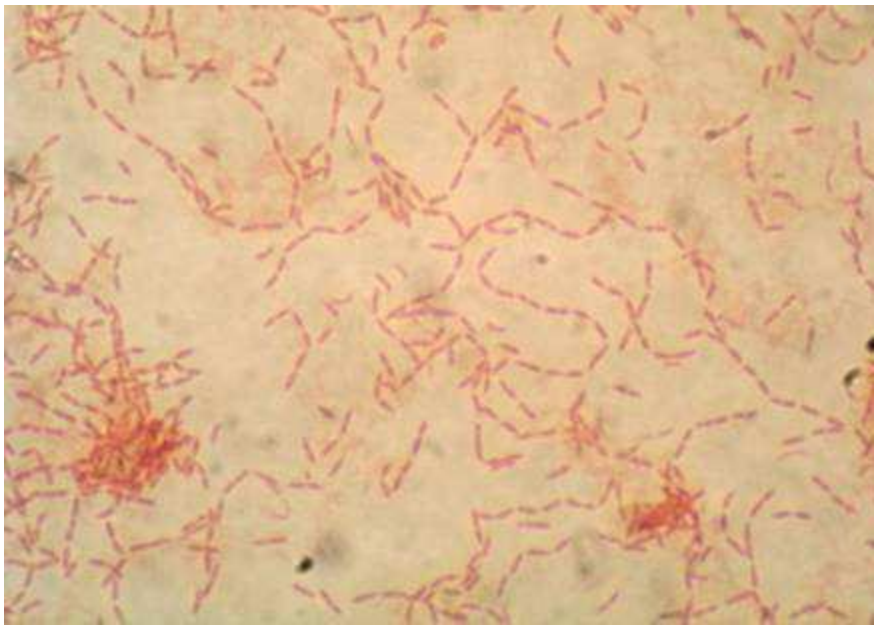


Figure A1 Typical Gram-staining of wild-type *B. proteoclasticus* grown in M704 broth media.

Appendix II Primers

Table A1 Primer details

Primer Name	Sequence	Primer Name	Sequence
pCY360			
p360_196fp	gatatctgccttctggtacgct	p360_SsbFp	aggaagttcctcatcgat
p360_196rp	gcatatataagggtatcgtcggat	p360_SsbRp	cttatgggacgtctcact
p360_341fp	cttctgtacaaagcatctgccatat	p360_Thyfp	atatagacacaagttacgcag
p360_343rp	tcattgccggaactatacaaa	p360_Thyrp	ttgagcaaaagacattcc
p360_398fp	tcatacaaataaaaatagcgccagc	p360_UNPPFp	ggcattaaaactccggttat
p360_398rp	aaaaccaacgaccttgcaca	p360_UNPPRp	gcataaagagcacatgcg
p360_ChypOrifp	tttcttaaaactccgggcat	p360_Morifp	tgtccttctttcttctgtca
p360_ChypOriRp	tgcggagcattaacgatt	p360_Morirp	tcttccaggatgagctca
p360_IG3Probe	ttaatagaaaaaggactaaaaata	p360OriT_P1 fp	aatcatcaagtgcaagtatg
p360_Int/RecFp	tggctcccgaaagattaat	p360OriT_P1 rp	gcagaccatatcatgcta
p360_Int/RecRp	gtcaatttacggaggagc	p360OriT_P2 fp	gtttcaatgggcttataatc
p360_IR10Probe	ggttccctttttttaccacctataa	p360OriT_P2 rp	cttttgtctttgcatgatag
p360_IR11Probe	ttttttatattataataaagtagttcgctaaa	p360OriT_P3 fp	ctatcatgcaaagacaaaagc
p360_IR2rp	aggatgaacaagaagctg	p360OriT_P3 rp	agaaaaagggctttaagg
p360_NrdDfp	tcgactataagaactttacgtggg	p360OriT_P4 fp	ttcttctggaactctctg
p360_NrdDrp	tggcacagaagttgagtt	p360OriT_P4 rp	cagagcattgatacttct
p360_NrdGfp	ctaagcgttctctgtggt	p360OriT_P5 fp	cagcaaaagaagtatcaatgc
p360_NrdGrp	ataaagtccccaattaat	p360OriT_P5 rp	cgaatcatttggagcata
p360_OriRaFp	aattgttcgtcatctactgtcac	p360OriT_P6 fp	ttacgctcatatgtcctcc
p360_OriRaRp	gagtgatacatcatacttgagaagt	p360OriT_P6 rp	aatttttggaagtcgagc
p360_ParAFp	tacaatctcaacgattaacc	p360OriT_P7 fp	gacttccaaaaattctgcacag
p360_ParARp	cctccaacttttctacca	p360OriT_P7 rp	aagcaggtaaatggcgac
p360_RepBFp	aaagtagtgaactttcactaataa	p360OriT_P8 fp	tactgttcgataagagacatga
p360_RepBRp	tttcaagaagctcgaaat	p360OriT_ rp	ttgtttcttccatgaaaa
BPc2			
c48probefp	aaccaaaacagttgaggatag	p300c49a	taaactttcacaggccctcc
c48proberp	atactcgctaccaattactcc	p300c49b	atagcggaaacgcagaaataga
c49a'	ccaggtttcaactataccatt	p300probe1fp	tcaacagccttattacctgtaaag
c49a6	ttcaagtaagccctacgaaa	p300probe1rp	ttatgccctggaacaagaaa
c49b6	caatctgacgggatttaagt	p300probe2fp	caggtaaaagagcgtgataact
p2_1.p1ca	gtgagcctcttctgtactcgt	p300probe2rp	ctactgttccaagagtcgga

p2_1.p1cb	aagtggctctcgtataatgc
p2_1.q1ca	ctgaccaacacgttatcatc
p300_parAfp	cacttatcctcctctaagaa
p300_parArp	tacaatagcatttagcaacc
p300_repB2fp	gtgggggacttatgccagat
p300_REPb2rp	tagcctgcagcttcagcaat
p300_repBrp	tatctactgtaatatgctgc
p300108B2	gaggaacggtatcattgattca
p300109A2	actacagccgccatgataact
p300c107a	cacaaactactttatcaact
p300c107a4	aagcatatgcgtctcgattg
p300c107b	tttagatgatgcatataagaaat
p300c107b	cacaatgctcacacatcattc
p300c107b'	gcccattgccgcagctccaa
p300c107b	tttagatgatgcatataagaaat
p300c107b4	ctcagccgtgggcttgcttt
p300c107b4confp	tgacggtacctgactaagaa
p300c107b4conrp	atctaatcctgtttgctccc
p300c107b5	gagactgccagggataacct
p300c107b6	accttaccagatcttgagatcc
p300c107d	cacaatgctcacacatcattc
p300c108a'	attagggcttgcgctcactg
p300c108a	tcgataaattcaaggaaaca
p300c108a4	cgcgggtacagcatctcact
p300c108a4confp	gcggagtcgctgttgggata
p300c108a4conrp	cgtcgatgtgaactcttggg
p300c108a5	cttcggtgtcgtgtttcagc
p300c108a6	tactcattccggcattctct
p300c108b	tagagctttatgccaaggatcc
p300c108b'	gtctggagctgtatagtgaa
p300c108b	tagagctttatgccaaggatcc
p300c108b4	gtataagtacggcaatagacat
p300c108b6	gcagaagtatgcccagatca
p300c108c	attagggcttgcgctcactg
p300c108d	gtctggagctgtatagtgaa
p300c109a'	gagtgccttaacatgctcctt
p300c109a	cctttatcgtcttcctgtctttat
p300c109a3	actacaggcgcttctgtact
p300c109a4	tttcatcaccactggatt
p300c109a6	gtgagttgccaatctgacgg
p300repB2fp	gtgggggacttatgccagat
p300repB2rp	tagcctgcagcttcagcaat
pcr116.bwp	tgaggatgataagcaggttg
pcr116.gwp	ttctggttctcctggatatg
pcr117.bwp	tctgtggtagtggagtttgag
pcr117.bwp10	tgagtgtagcggtaatgctc
pcr117.bwp5	ctctggagctgacatagacatag
pcr117.bwp7	cgcacttacaggtacaagtc
pcr117.bwp8	ggttacaaggtctgacatgc
pcr117.gwp	tgaagatcctgtggcagatg
pcr117.gwp3	cttgaaattgatagtgataacag
pcr117.gwp3	tcctccatcctgaattattg
pcr117.gwp3	cttgaaattgatagtgataacag
pcr117.gwp4	tcctccatcctgaattattg
pcr117.gwp5	cctgaagtctcttgacgaag
pcr117.gwp6	ccggattatcagagcttcac
pcr117.gwp7	aggcaggtcagttagttgtg
pcr117.gwp8	aagtgcatccttatcctcac
pcr117.gwp9	gctgagcagattagtgaagc
pcr118.bwp	gtattcagcagctctatgacac
pcr118.gwp	ctacatgatcttcggaccag
RNA3a.1	tgtttcactatcggtcacca
RNA3a.10	tcgaattaaaccacatgctc
RNA3a.11	tcgggtctatgtgaagtgac
RNA3a.12	taacaaggtagccgtaggag
RNA3a.13	ttctatacactccgctgctc
RNA3a.14	cgcagtaagtattccacctg
RNA3a.15	tcctacggctaccttgttac
RNA3a.16	tggagcatgtcgtttaattc
RNA3a.2	acgtgttaaccgtcgtgttc
RNA3a.3	tggattattgtggcatagaag
RNA3a.4	cgacacagaacctttgacaa
RNA3a.5	ttgtcaaggttctgtgtcg
RNA3a.6	tggattattgtggcatagaag
RNA3a.7	ctacagatcgttgctttggtg
RNA3a.8	ccacattgggactgagacac
RNA3a.9	ctagtgtagcggtgaaatgc
RNA3afp	atccaatgctgttaatctagg
RNA3arp	ggacttatacttggagctatcg
RNA3b.1	ttacctagagcgatgagc

p300c109a7	gaaacaaaagagttctccgt	RNA3b.10	gatgtgaactcttgggagtg
p300c109b'	ctttcaggatgagaatggcga	RNA3b.11	acttagtgatccggtggtatg
p300c109b	caaatccggcaaacacactt	RNA3b.12	ctaggatgttccctcagaag
p300c109B2	gagcgtaactccctcatctgct	RNA3b.13	gtaaccagcatcttcactgg
p300c109c	gagtgccttaacatgctcctt	RNA3b.2	ggtcgctaataacgtatg
p300c109d	ctttcaggatgagaatggcga	RNA3b.3	tggtgaccgatagtgaaac
p300c110/111a	atgacagcatctggtggagc	RNA3b.4	gataggcacaaggtgtaagc
p300c110/111A2	gctttggtggaacactcatt	RNA3b.5	gaaggcatctaagcgtgaag
p300c110/111b	accctaaggtttgtgaatcttg	RNA3b.6	atacgttattagcgacccgaag
p300c110/111B2	atgtcgagatattcccgcca	RNA3b.7	acttcacatagacccgaaac
p300c110a'	accatagtctattcatatttcc	RNA3b.8	gattatcaaactccgaatgc
p300c110a4	taaaaggttacaggcagttg	RNA3b.9	tcggctagtgagctattacg
p300c110a5	tgccaaggaatactacatg	RNA3bfp	tggagttccctgtgcataatc
p300c110c	accatagtctattcatatttcc	RNA3brp	ttgttggcgcagatgtattatc
p300c112a	gccgtatgcttgatgtcaat		
p300c112A2	atgcatctgacgcaaatggg		
p300c112b	ttacttattgactaggcatatat		
p300c112b4	actgctctcctcaagctcaa		
p300c145+a'	cattcaggagaccaaacttc		
p300c145+a2	gaacctgtctgcttaatctg		
p300c145+b'	atccggcgctgagcagatta		
p300c145+c	cattcaggagaccaaacttc		
p300c145+d	atccggcgctgagcagatta		
p300c145A	gcctcttcctgtttggcatt		
p300c145B	acatgctggattcaggtgct		
p300c145b4	acctgcatttggtgtatata		
		pCY190	
p190_parAfp	cttctgacaatttcatcaac	p190_parA2fp	ctgacaatttcatcaacaatacgtc
p190_parArp	taacaatagcaatggcaaac	p190_parA2rp	gcaacaatacttacaaagctaggta
p190_repBfp	taataatatcttttccagctctac	p190_parB2fp	ctttttttctagtttctcaagcg
p190_repBrp	gaaaaaagatgaatcttatgat	p190_parB2rp	ctttttttccatgaggagcgt
p190_5a5	cggaaaacaggttgttcatg	p190_probe1fp	ggtttaggttgagatactatcattg
p190_c105acomp	ggcaacatccgataagaata	p190_probe1rp	tcaaaccaagtgcaaattga
p190_c105a5	gaaccttcatctgaataaatcc	p190_probe2fp	ttcactttggtgaagacgaa
p190_c210a5	gcggataatagatcagattagcaa	p190_probe2rp	ctgtagggagtaaggcataca
p190_c210b5	tctccgaaactcctatatcaca	p190_repB2fp	taataatatgttttccagctctag
p190_c211a5	gccatccttaacaagtttctt	p190_repB2rp	gaaaaaagatgaatcttatgat
p190_c240A	caattacgacaccaagacatca	p3.bwp	tatgcggattgctaagatg
p190_c240a'	catctgctctctgttcatact	p3.bwp1	atgaagaactcgctgctatc
p190_c240B	tgggtaaatatcaaagaatgacaa	p3.bwp2	gcgatattcatcttcattgg

p190_c240b'	ctgctttacatattctcttaac
p190_c240b5	aatcggtgcattgttgcctt
p190_c240c	catctgctctctgttcatact
p190_c240d	ctgctttacatattctcttaac
p190_c3A	attcttctgctccaccaacc
p190_c3a'	gacgattgtaaccggctcgta
p190_c3B	tagttaatgatctccaaataattat
p190_c3b'	tgcaacaggctgcttgaact
p190_c3b5	tgcgactaaacaaaatcttacaagg
p190_c3c	gacgattgtaaccggctcgta
p190_c3d	tgcaacaggctgcttgaact
p190_c49c	ccaggtttcaactataccatt
p190_c4A	atttcgcatcaacaaatact
p190_c4a'	agagcacagcagcaaaggcatc
p190_c4a_cfm	catagaaaggacgacagttt
p190_c4B	tgggaataagaagcaactatggga
p190_c4b'	tatcaacgaaacggattacc
p190_c4b‟	tatcaacgaaacggattacc
p190_c4b5	gaaaggggaattcaatatgc
p190_c4c	agagcacagcagcaaaggcatc
p190_c5A	taagtggatagcgcatgcgt
p190_c5a'	gctggccttctgtgatgtgc
p190_c5a5	cggaaaacaggttgttcatg
p190_c5B	gatgccttctaactgcatcaat
p190_c5b'	ccattcatatacgcagttatg
p190_c5c	gctggccttctgtgatgtgc
p190_c5d	ccattcatatacgcagttatg
p190_OriAfp	ggatccgtgtggtggaattc
p190_OriArp	atcatctagagcgacagccc
p190_OriBfp	ccgcttattttagcacaatattgac
p190_OriBrp	aatagcaatggcaaaccaaa
p190_OriCbrp	gacaatgctccagcaagaaa
p190_OriDrp	gagcatgaagaaaagcaagaagtaa
p190_OriErp	ccatattcttctcttcgcca
p190_oriGfp	ttaagataaggcactgatcattctg
p190_oriGrp	gttggtaaaacaacctcaactatag
p190_oriHfp	cgtaaaaaaaggagaggtaac
p190_oriHrp	gagcagattatcaacggaaa
p190_oriIfp	acttcttaaatcatgctccaaatgc
p190_oriIrp	ccagcaagaaatattcctctaaaca
p3.gwp	tgttcattactatatgccatgc
p3.gwp1	atccgctacattcctgatag
p3.gwp2	gcatttatctcagtctcagc
p3_1.p1ca	ttcacgtatccaacgtctaac
p3_1.q1ca	catggagtggaagagaagag
C5probefp	ccgtgacaacaacctatgtaca
C5proberp	acgctgctccaccaactaaa
pcr119.bwp	ggtcgaagtagttgagaagaag
pcr119.bwpc	gacgtaggcaattacgacac
pcr119.gwp	tacatcttcaacaaggctctc
pcr119.gwpc	cacatctgttgttccaccac
pcr120.bwp	tgtggttacagacaatggaac
pcr120.bwp3	acatcttctccaagcatcac
pcr120.bwp5	gacttcaaccgcttctgtag
pcr120.bwp6	gcttggaacaaaccatacag
pcr120.bwp7	ttcaagtaccactctaagagc
pcr120.bwp9	aagagcagtcagtcaagg
pcr120.gwp	gaagtcagggatgtaatcagg
pcr120.gwp3	tctgtacgtggagtcttatgg
pcr120.gwp5	cctaaagtagcggatgaagtc
pcr120.gwp6	acacgctaaagttgccatatc
pcr120.gwp8	tccaggatgtactgatctcc
pcr204.bwp	ggagtaatctgcgtcatagc
pcr204.gwp	tcactccatgaatcagtacg
pcr204.gwpa	agcctgcttgattattcttg
pcr204.gwpb	tggattacactaccagcattg
pcr204.gwpc	gacggctatctcgtaagaag
CpP190f	tggaagacaaagcactagctggga
CpP190r	cgcttcgtgtcgtccagaaa
13666.bwp1	tgacaattatacagtcctgcag
13666.gwp1	ttatctgcatcttagcaatcc
3026.bwp1	cgattgtaaccggctcgtat
3026.bwp2	taagccaagtggacagattg
3026.gwp	tatacgagccggttagaatc
3026.gwp2	ttatggtcaggttgcattag

Chromosome		Other	
BPp1.fp	caccttcaacgaactccatc	bcdqfp	tgagaagggaacacctggat
BPp1.rp	ccctaattcacagctaatcc	bcdrfp	ttgctcttccgaactgctt
		mchqfp	gtattgcctggtgaagatgt
		mchqrp	gtcgatttggtagaagtca

Appendix III Supporting microarray data

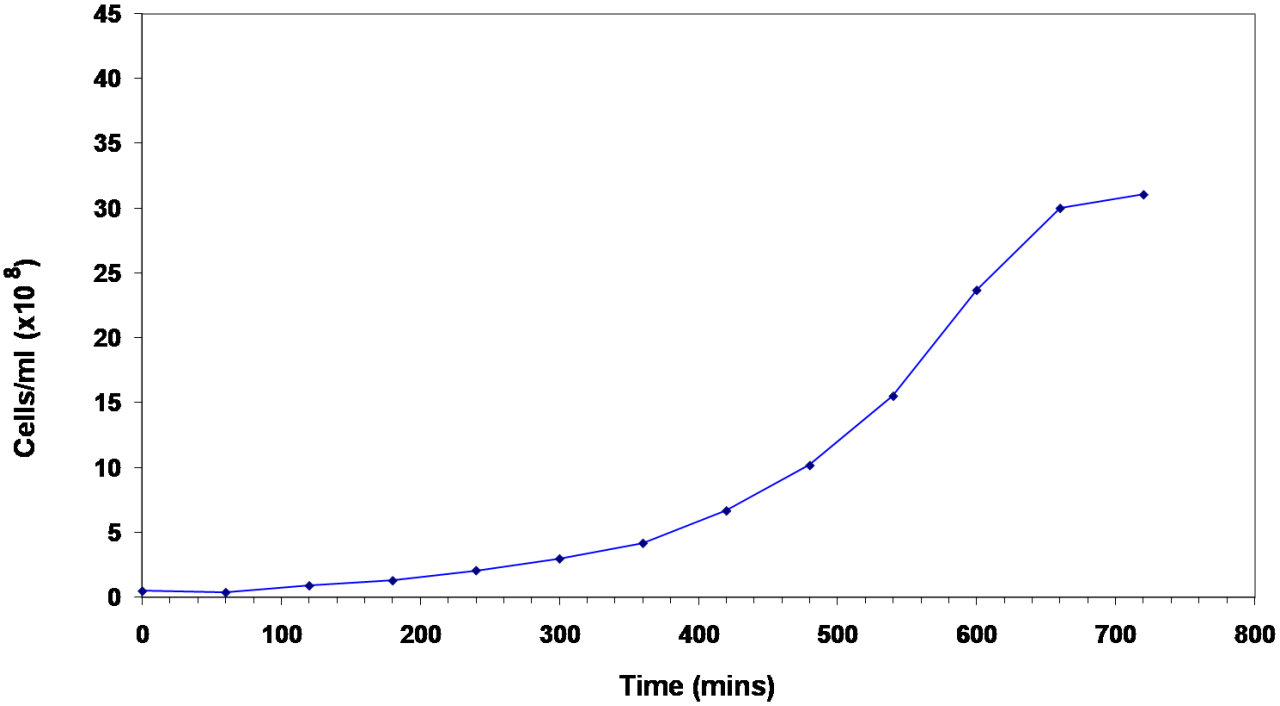


Figure A2 **Growth curve of *B. proteoclasticus* in BY+ media supplemented with 0.2% Xylan**

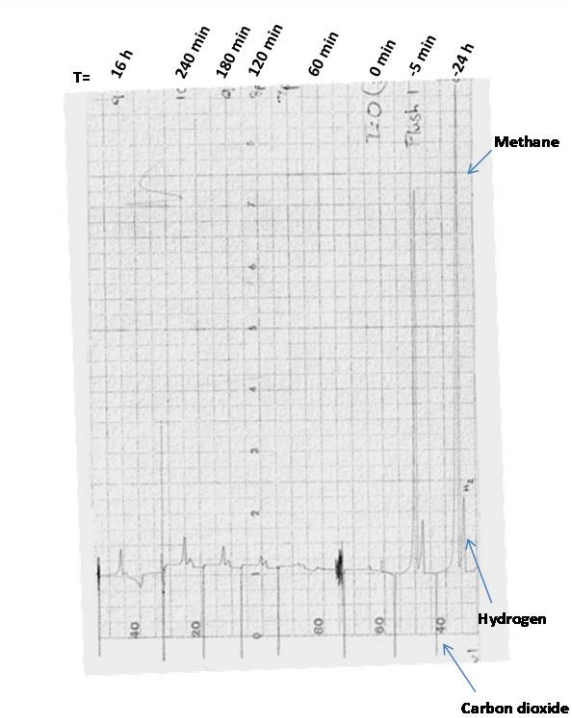


Figure A3 **GC analysis of H_2 and CH_4 in co-culture.** GC analysis of gas composition in co-culture before, up to, and after the point of extraction (510 minutes). Peaks run right to left and correspond to, in order, H_2, CH_4 and CO_2 as indicated. Cultures were flushed with CO_2 immediately prior to the addition of *B. proteoclasticus* (T=0).

Table A2 Spectrophotometric analysis of total RNA samples to determine purity and concentration.

Replicate	Sample	A_{260}/A_{280}	A_{260}/A_{230}	ng/ul	Total RNA (µg)
	Blank	-0.32	-0.08	-0.004	n/a
	B. proteoclasticus	2.00	2.35	108.83	6.53*
1	*M. ruminantium*	2.10	2.42	121.09	7.27
	Co-culture	2.11	2.38	177.83	10.67
	B. proteoclasticus	1.97	2.31	102.17	6.13
2	*M. ruminantium*	2.09	2.36	107.84	6.47
	Co-culture	2.09	2.41	193.01	11.58
	B. proteoclasticus	2.08	2.40	85.17	5.11
3	*M. ruminantium*	2.04	2.41	83.83	5.03
	Co-culture	2.08	2.32	172.39	10.34

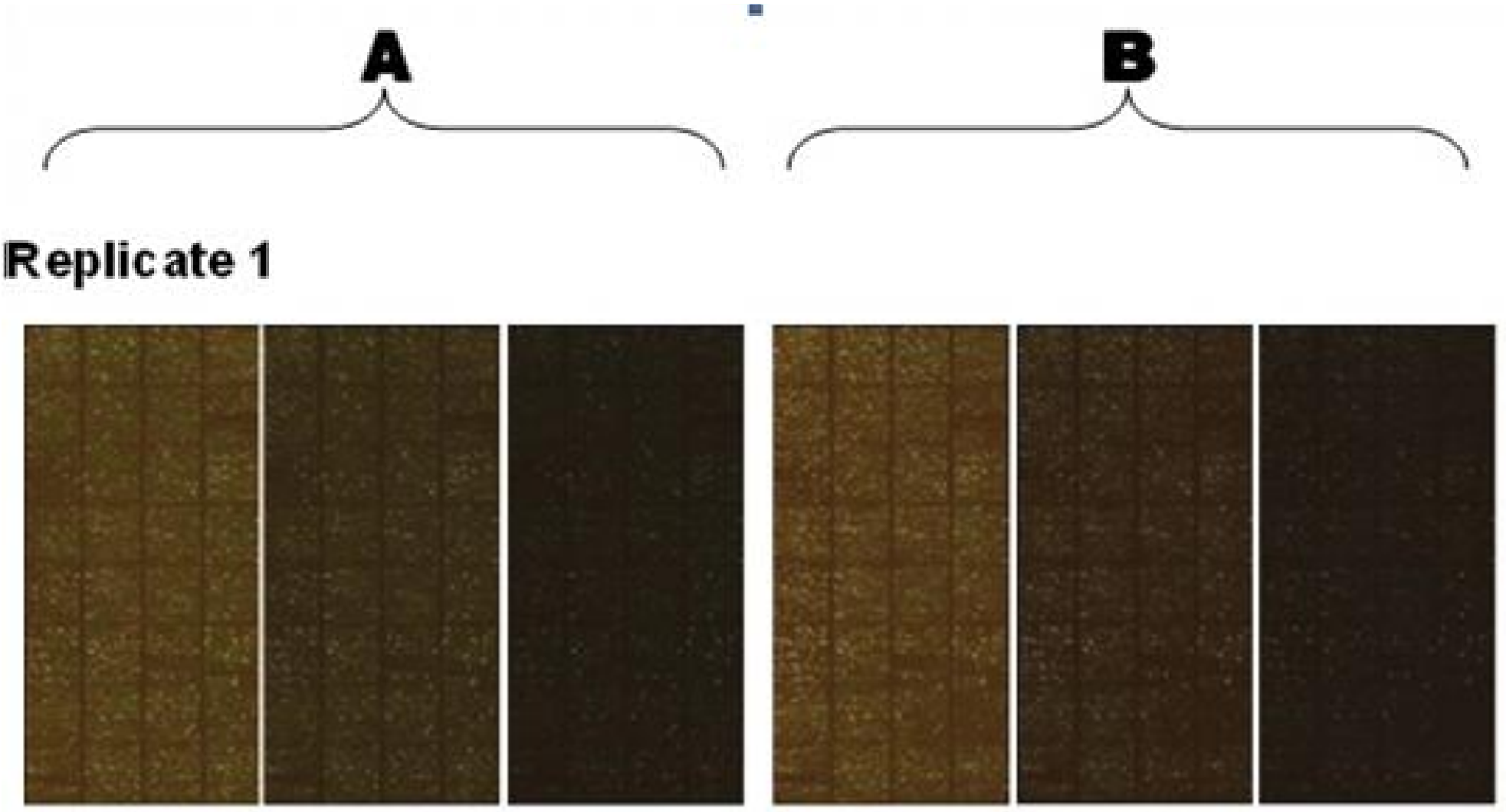
*Each total RNA sample was analysed by spectrophotometry. Absorbance readings were taken at wavelengths (λ) 230 nm, 260 nm, and 280 nm. The absorbance ratio between λ = 260 nm (A_{260}) and λ= 280 nm (A_{280}) as well as the A_{260} / A_{230} ratio are shown. RNA concentrations were determined by the Nanodrop software using an adaptation of Beers law (Concentration = (Absorbance x Extinction coefficient) / path length of the transmitted light) where the extinction coefficient for RNA is 40 ng-cm / ml. Total RNA is the total mass of RNA in the remaining 60 µl sample.

Table A3 Mixing schedule for mono-cultures

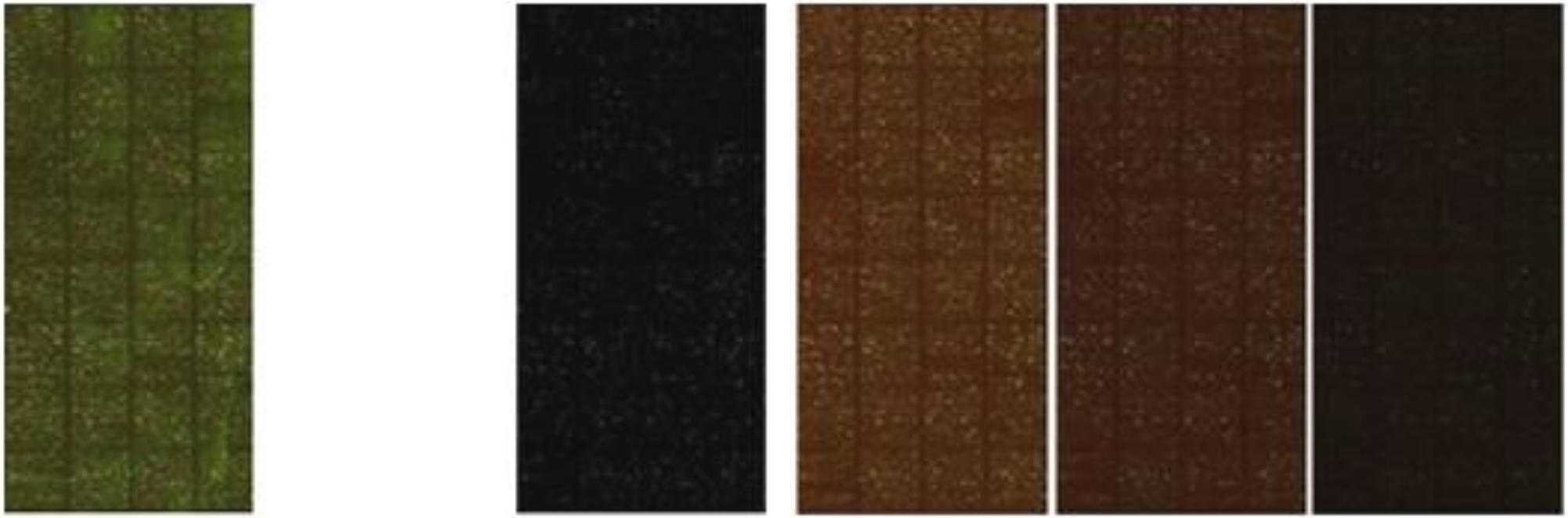
Sample	***B. proteoclasticus*** Vol (µl)	Concentration (Total RNA (%))	***M. ruminantium*** Vol (µl)	Concentration (Total RNA (%))
1	68.9	4.5 µg (45%)	75.7	5.5 µg (55%)
2	73.4	4.5 µg (45%)	85	5.5 µg (55%)
3	97.8	5 µg (50%)	99.5	5µg (50%)

Table A4 Microarray scan levels

Sample	Scan level	PMT Gain		Ratio
		λ = 532 nm (Cy3)	λ = 635 nm (Cy5)	
1A	High	537	580	0.97
1A	Med	445	490	1.00
1A	Low	362	400	1.02
1B	High	550	565	1.03
1B	Med	485	500	0.99
1B	Low	385	400	1.01
2A	High	580	510	1.00
2A	Med	525	460	0.99
2B	High	489	580	0.96
2B	Med	425	520	1.04
2B	Low	395	486	1.03
3A	High	540	580	0.99
3A	Med	444	500	0.95
3A	Low	380	478	0.99
3B	High	425	540	1.01
3B	Med	310	385	1.00


A
B
Replicate 1

Replicate 2



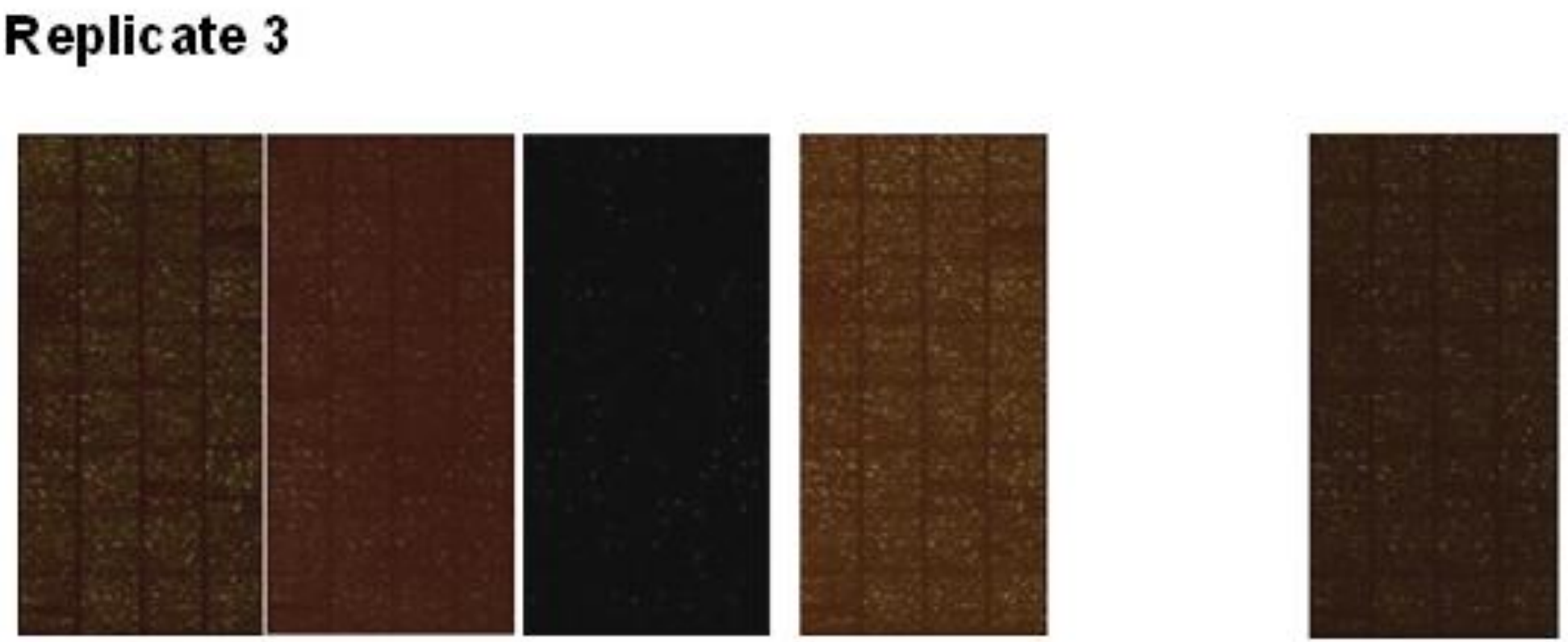


Figure A4 **Microarray scans.** Microarray scan pictures of replicates 1 to 3 are shown top to bottom and dye-swaps of each replicate are shown left to right (A and B).

Table A5 Microarray grid settings

Print	**121 CPR/MR combined**
Slides	50 epoxy
Slide rows	5
Slide columns	10
Blocks	32
rows	8
columns	4
Block spacing	Column 4550 Row 4499.5
384 well plates	17 x 3 repeats + 1 extra =52
# spots	19968
# rows X	26
# columns Y	24
spacing row	170
spacing column	170
Spot diameter	100

Table A6 Feature weightings

Slides	-100 flag*	- 50 flag	No flag
1A – High	0	1564	18404
1A – Med	0	1558	18410
1A – Low	0	1532	18436
1B – High	0	1774	18194
1B – Med	0	1909	18059
1B – Low	0	2479	17489
2A – High	0	0	19968
2A – Med	126	1191	18651
2B – High	1	2874	17093
2B – Med	1	3240	16727
2B – Low	2	4191	15775
3A – High	0	0	19968
3A – Med	0	1751	15235
3A – Low	0	4733	18217
3B – High	2	4118	15848
3B - Med	2	3269	16697

*Shows weightings applied to features of each slide. The weighting of features flagged as bad by visual inspection were given a weighting of -100 essentially removing them from further analysis, features bad flagged by Gene pix (typically low intensity or very irregular outlines) were given a weighting of -50 and features accepted by both Gene pix and visual inspection were not flagged.

Appendix IV R programming code

```
# ------------------------------------------------------------------ #
# Set up Bioconductor.
# REMEMBER: need to run this setup code at the beginning of all
sessions!!!
# ------------------------------------------------------------------ #

Sys.setenv("http_proxy"="http://webgate.agresearch.co.nz:8080")
install.packages(c("limma","iplots"), dependencies = TRUE)

# Note that after the first line of code comes up, you'll need to
enter your log-in details (as R is going outside of AgR to get the
files).
# To log-in you need to type:
# Username: Agresearch\your_AgR_username, Password:
your_AgR_password.
# You'll then be asked to choose a mirror - use NZ.  After doing
this, the packages should then be loaded into R.

library(limma)

memory.limit(size=4000)                              # Changes memory to
4000 Mb
memory.limit(size=NA)                                # Tells you how
much the current memory capacity is

workingDir <- "M:/Coculture microarray"               # Change the file
pathways to what you want
setwd(workingDir)                                    # Sets the working
directory to workingDir
dataDir <- "M:/PhD/microarray/Results"


# ------------------------------------------------------------------ #
# Create a 'target' file
# ------------------------------------------------------------------ #

# Obtaining all .gpr filenames
write.table(dir(dataDir,pattern=".gpr"), "6  microarrays of B316 and
MBB in Coculture versus not", quote = FALSE)


# ------------------------------------------------------------------ #
# Read in the files defined in the FileName column of the Target
file.
# ------------------------------------------------------------------ #

targets <- readTargets("Coculture array list.txt")         # Reads in
the targets file
targets                          # Prints out what is contained in
targets so you can doublecheck that it is correct


wt.flags <- function(x) {        # Function which gives spots with
flag values > = 0 a weighting of 1 and 0 otherwise.
  as.numeric(x$Flags >= 0)       # A weight of 0 means that the BAD
spot is multiplied by 0 and so 'disappears'.
}                                # A weight of 1 = a GOOD spot.
```

```
x <- list()                       # Checking that the weight function
is working (1 = good spots, 0 = bad spots)
x$Flags <- c(100, 0, -50, -100)
wt.flags(x)                       # Should give 1 1 0 0

RG <- read.maimages(targets$FileName, source="genepix.median",
path=dataDir,
                    wt.fun=wt.flags, other.columns="Flags")    #
Reads in the files in dataDir using the median pixel values
                                                               # and
weighting the data as defined in the weight function

# Help documentation
limmaUsersGuide()


# ------------------------------------------------------------------ #
# Saving your data in R
# ------------------------------------------------------------------ #

save.image("M:/Coculture microarray/Coculture analysis.RData")

# ------------------------------------------------------------
# To Reload

load("M:/Coculture microarray/Coculture analysis.RData")

# ------------------------------------------------------------------ #
# Quality control: types of spots on the chip...
# ------------------------------------------------------------------ #

# ----------------------------------
# Create a list which shows how many copies there are of each gene on
the chip...

# Spots with "NA" gene IDs as Limma can not deal with these
sum(is.na(RG$genes$ID))

# note: there are 0 genes!!


# Creating the list...
copies <- stack(table(RG$genes$ID))
o <- order(copies$values, decreasing = T)
copies.ordered <- copies[o,]
copies.ordered[1:10,]

# Writing a dataset of gene IDs and copy numbers to Excel for easy
viewing...
getwd()
write.table(copies.ordered, file = "Copy numbers of genes,
ordered.xls", sep = "\t")

# Summary of how many genes there are with each count...
c(table(copies.ordered$values), "total genes" = length(RG$genes$ID) )

# ----------------------------------
# Plot the expression values for each gene type...
```

```
# Create a variable identifying the gene types...

RG$genes$control.IDs <- RG$genes$ID                                #
Create a copy of the gene Name
RG$genes$control.IDs <- "Unrecognised"                                  #
Then rename the gene IDs to be one of the 6 classes of spot found on
the chip
RG$genes$control.IDs[grep("CPROT70",RG$genes$ID)] <- "Bprot"
RG$genes$control.IDs[grep("MRUM70",RG$genes$ID)] <- "Mbrevi"
RG$genes$control.IDs[grep("CPROT50 rand",RG$genes$ID)] <- "Controls"
RG$genes$control.IDs[grep("CPROT50 arabid",RG$genes$ID)] <-
"Controls"
RG$genes$control.IDs[grep("CPROT50 dnaK",RG$genes$ID)] <- "Bprot
dnaK"
RG$genes$control.IDs[grep("CPROT50 frhB",RG$genes$ID)] <- "Mbrevi
frhB"
RG$genes$control.IDs[grep("blank",RG$genes$ID)] <- "Blank"


# Plot the RAW values for each gene type...

slideNames <- rownames(RG$targets)

targets

oldpar <- par()

par(mfrow=c(4,4),mar=c(7,4,4,2)+0.1,mgp=c(3,1,0))

for( i in c(1:16)) {

    plot(as.data.frame(RG$genes$control.IDs),
log2(RG$R[,i]/RG$G[,i]),
main=slideNames[i],cex.main=1.1, ylab = "log2 ratio for the raw
data", las = 2, cex = 0.8,
ylim=c(-4.5,4.5))

abline(h=0)
}

par(oldpar)

# ------------------------------------------------------------------#
#  Quality control: summary statistics...
# ------------------------------------------------------------------#

# Summary of Red and Green foreground and background intensities:
# Tables of summary statistics

summary(RG$R)
summary(RG$Rb)
summary(RG$G)
summary(RG$Gb)

# Boxplots

targets                                                            #
Check slide order

oldpar <- par()
```

```
par(mfrow=c(1,1),mar=c(10,4,4,2)+0.1)                          # Set
margins so that the x-axis labels can be entirely seen.

colors()

colour <- rep(c("blue","skyblue2", "light blue"),4)

colour <- ("blue","skyblue2","light blue","blue","light
blue","blue","skyblue2","light blue","blue","skyblue2","light
blue","blue","skyblue2","light blue","blue","light blue")
# Set boxplot colours

boxplot(data.frame(log(RG$R)), las = 2, main = "Red foreground", ylab
= "log intensity", col = colour)  # Plots images in order given in
targets

# Plotting slides in the order high, medium, low

par(mfrow=c(2,2))                                                  # Plot
2x2 graphs per page

boxplot(data.frame(log(RG$R[,c(1,3,2,4,6,5,7,9,8,10,12,11)])), las =
2, main = "Red foreground", ylab = "log intensity", col = colour)
boxplot(data.frame(log(RG$Rb[,c(1,3,2,4,6,5,7,9,8,10,12,11)])), las =
2, main = "Red background", ylab = "log intensity", col = colour)
boxplot(data.frame(log(RG$G[,c(1,3,2,4,6,5,7,9,8,10,12,11)])), las =
2, main = "Green foreground", ylab = "log intensity", col = colour)
boxplot(data.frame(log(RG$Gb[,c(1,3,2,4,6,5,7,9,8,10,12,11)])), las =
2, main = "Green background", ylab = "log intensity", col = colour)

# Plotting slides in the order high, medium, low and with the same y-
axis range for all plots.

boxplot(data.frame(log(RG$R[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16
)])), las = 2, main = "Red foreground", ylab = "log intensity", col =
colour, ylim = c(3,11))
boxplot(data.frame(log(RG$Rb[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,1
6)])), las = 2, main = "Red background", ylab = "log intensity", col
= colour, ylim = c(3,11))
boxplot(data.frame(log(RG$G[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16
)])), las = 2, main = "Green foreground", ylab = "log intensity", col
= colour, ylim = c(3,11))
boxplot(data.frame(log(RG$Gb[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,1
6)])), las = 2, main = "Green background", ylab = "log intensity",
col = colour, ylim = c(3,11))

par(oldpar)

# ------------------------------------------------------------------ #
# Quality control: slide images...
# ------------------------------------------------------------------ #

par(mfrow=c(4,4))

for( i in c(1:16) ) {
imageplot(rank(RG$R[,i]), layout=RG$printer, low="white", high="red",
main=paste(slideNames[i], ": FG"), cex.main=1.1)
}

for( i in 1:16 ) {
```

```
imageplot(rank(RG$Rb[,i]), layout=RG$printer, low="white",
high="red",
main=paste(slideNames[i], ": BG"), cex.main=1.1)
}

for( i in 1:16 ) {
imageplot(rank(RG$G[,i]), layout=RG$printer, low="white",
high="green",
main=paste(slideNames[i], ": FG"), cex.main=1.1)
}

for( i in 1:16 ) {
imageplot(rank(RG$Gb[,i]), layout=RG$printer, low="white",
high="green",
main=paste(slideNames[i], ": BG"), cex.main=1.1)
}

# ------------------------------------------------------------------ #
# Quality control: flag diagnostics...
# ------------------------------------------------------------------ #

# ----------------------------------------
# Plotting the scans done at different intensities versus each
other...

# Splitting up the slides into high, medium and low scan types

targets
# Check slide order

high.slides <- (grep("Hi",targets$FileName, ignore.case = TRUE))
medium.slides <- (grep("Med",targets$FileName, ignore.case = TRUE))
low.slides <- (grep("Lo",targets$FileName, ignore.case = TRUE))

high.slides
medium.slides
low.slides

# High scans vs low scans...

dye <- RG$R                    # Change to RG$R or RG$G as you wish
dyecolour <- "Red"             # "Red" or "Green" to match the RG
variable above.

par(oldpar)
# Colouring the plots by their flag status...

Flag.colour <- (-1*RG$other$Flags[,high.slides[slide]])+1
# Can use the high.slides, medium.slides or low.slides flags to
colour the plots (change to what you want)
unique(Flag.colour)
iset.col(Flag.colour)
# May take a few seconds for the points to appear!

iset.col(c(0,0))
# Sets the colour back to all black

# Identifying selected points...

iset.selected()
```

```
print(cbind(RG$genes[iset.selected(),1:4], Flag.high =
RG$other$Flags[iset.selected(),high.slides[slide]], Flag.medium =
RG$other$Flags[iset.selected(),medium.slides[slide]],
Flag.low = RG$other$Flags[iset.selected(),low.slides[slide]]), quote
= F)

# Setting the weights for any extra bad-flagged spots to 0.  Wise to
make a copy of the original values first!

RG$old.weights <- RG$weights
RG$other$old.flags <- RG$other$Flags

spot.1 <- iset.selected()
spot.2 <- iset.selected()
# Create as many 'spot' variables as are needed

RG$weights[c(spot.1, spot.2),high.slides[slide]] <- 0
# Change the scanning level to that on which the spot is bad.
RG$other$Flags[c(spot.1, spot.2),high.slides[slide]] <- 0


# ------------------------------------------------------------
# Create a table showing the numbers of each type of Flag

Flags.stacked <- c(RG$other$Flags)
Slides.stacked <-
rep(colnames(RG$other$Flags),each=length(RG$other$Flags[,1]))
Flag.matrix <- as.matrix(table(Slides.stacked,Flags.stacked))

Flag.matrix


# ------------------------------------------------------------------- #
# NORMALIZATION
# ------------------------------------------------------------------- #


# ------------------------------------------------
# Removing the control spots.  i.e. setting the blank, random,
arabidopsis and spike in control values to 0...

RG$old.weights1 <- RG$weights
RG$weights[RG$genes$control.IDs=="CPROT50 rand"] <- 0
RG$weights[RG$genes$control.IDs=="CPROT50 arabid"] <- 0
RG$weights[RG$genes$control.IDs=="CPROT50 ackscat"] <- 0
RG$weights[RG$genes$control.IDs=="CPROT50 ack5het"] <- 0

RG$weights[RG$genes$ID=="blank"] <- 0

# ------------------------------------------------
# Within slide normalization...

#?backgroundCorrect -for help file
#?normalizeWithinArrays -for help file

MA <- normalizeWithinArrays(RG, method="printtiploess",
bc.method="none")


# ------------------------------------------------------------------- #
# MA plots
```

```
# ------------------------------------------------------------------ #

# Spots with zero weight NOT shown...

par(mfrow=c(4,4))
for(i in 1:16) {
    plotMA(MA, array=i, main=slideNames[i], cex.main=1.1, ylim=c(-
3,3))
abline(h=0)
}

# Show spots with zero weight  (note: y-axis range now set to (-4,4)
so that all of the bad spots can be seen)

par(mfrow=c(4,4))
for(i in 1:16) {
    plotMA(MA, array=i, main=slideNames[i], cex.main=1.1,
zero.weights = TRUE, status = RG$weights[,i], xlim = c(7,16),
ylim=c(-4,4))
abline(h=0)
}

# Show control and quality assurance spots

par(mfrow=c(4,4))
for(i in 1:16) {
    plotMA(MA, array=i, main=slideNames[i], cex.main=1.1,
zero.weights = TRUE, status = RG$genes$control.IDs, xlim = c(7,16),
ylim=c(-4,4))
abline(h=0)
}


# ------------------------------------------------------------------ #
# Boxplots of the normalized data by different spot type
# ------------------------------------------------------------------ #

oldpar <- par()

par(mfrow=c(4,4),mar=c(7,4,4,2)+0.1,mgp=c(3,1,0))

for( i in c(1:16)) {
    plot(as.data.frame(RG$genes$control.IDs), MA$M[,i],
main=slideNames[i],cex.main=1.1, ylab = "log2 ratio for the
normalized data", las = 2, cex = 0.8, ylim=c(-4.5,4.5))
abline(h=0)
}

par(oldpar)


# ------------------------------------------------------------------ #
#  Density plots...
# ------------------------------------------------------------------ #

# Comparing results when background corrected vs not

par(mfrow=c(2,1))
plotDensities(backgroundCorrect(RG, method="none"))
plotDensities(backgroundCorrect(RG, method="subtract"))
```

```
# Comparing results for high, medium and low scans (Raw data)

par(mfrow=c(3,1))
plotDensities(backgroundCorrect(RG[,high.slides], method="none"))
plotDensities(backgroundCorrect(RG[,medium.slides], method="none"))
plotDensities(backgroundCorrect(RG[,low.slides], method="none"))

# Comparing results for high, medium and low scans (Normalized data)

par(mfrow=c(3,1))
plotDensities(MA[,high.slides])
plotDensities(MA[,medium.slides])
plotDensities(MA[,low.slides])

# ------------------------------------------------------------------- #
# STEP 13 - t-tests...
# ------------------------------------------------------------------- #

library(limma)

#-------------------------------------------------------------------- #
# Divide the normalized data and targets file into subsets of high,
medium and low scan data:

MA_high <- MA[,high.slides]
MA_medium <- MA[,medium.slides]
MA_low <- MA[,low.slides]

targets_high <- targets[high.slides,]
targets_medium <- targets[medium.slides,]
targets_low <- targets[low.slides,]


#-----------------------------------------------------
# Creating the design matrix...  (Note can't remove a dye effect as
the Ref mRNA is always on Green)

design.1 <- cbind("CoculturevsMBB_B316" = c(1,1,1,0,0,0),
"MBB_B316vsCoculture" = c(0,0,0,1,1,1))

#-----------------------------------------------------
# Creating the contrast...

cont.matrix.1 <- makeContrasts("CoculturevsMBB_B316" =
CoculturevsMBB_B316 - MBB_B316vsCoculture, levels = design.1)
cont.matrix.1

#-----------------------------------------------------
# Medium scan analysis

# Fitting the linear model to the design matrix, and then to the
contrast [Coculture - Monoculture]:

fit_medium.1 <- lmFit(MA_medium, design.1)
fit_medium.1 <- contrasts.fit(fit_medium.1, cont.matrix.1)

names(fit_medium.1)
dim(fit_medium.1$coefficients)  # 44290 * 1

# Calculating simple t-statistics (NOT recommended!)
```

```
simple.t.pvals.1 <- (2 * pt(-
abs(fit_medium.1$coefficients[,1]/fit_medium.1$stdev.unscaled[,1]/fit
_medium.1$sigma), df = fit_medium.1$df.residual))

# Calculating empirical Bayes moderated t-statistics:

fit_medium.1 <- eBayes(fit_medium.1)

# Calculating FDR's:

FDR.1 <- p.adjust(as.matrix(fit_medium.1$p.value)[,1], method="BH")

#-------------------------------------------------------
# Low scan analysis - FDR's not calculated here as will be later
using resultsOut

fit_low.1 <- lmFit(MA_low, design.1)
fit_low.1 <- contrasts.fit(fit_low.1, cont.matrix.1)
fit_low.1 <- eBayes(fit_low.1)

#-------------------------------------------------------
# High scan analysis

fit_high.1 <- lmFit(MA_high, design.1)
fit_high.1 <- contrasts.fit(fit_high.1, cont.matrix.1)
fit_high.1 <- eBayes(fit_high.1)


# ------------------------------------------------------------------- #
#  STEP 14 - VOLCANO PLOTS...
# ------------------------------------------------------------------- #

par(mfrow=c(1,1))
volcanoplot(fit_medium.1, coef=1,
highlight=sum(fit_medium.1$p.value<0.0005, na.rm=T),
names=fit_medium.1$genes$ID)  # Plots log odds versus log ratio

# Can easily modify the code to plot the log(mod p) versus log ratio
instead, or change the plotting colour or plotting symbol etc...

volcanoplot

# This should output the following code in the R console...

function (fit, coef = 1, highlight = 0, names = fit$genes$ID,
    ...)
{
    if (!is(fit, "MArrayLM"))
        stop("fit must be an MArrayLM")
    if (is.null(fit$lods))
        stop("No B-statistics found, perhaps eBayes() not yet run")
    x <- as.matrix(fit$coef)[, coef]
    y <- as.matrix(fit$lods)[, coef]
    plot(x, y, xlab = "Log Fold Change", ylab = "Log Odds", pch = 16,
        cex = 0.2, ...)
    if (highlight > 0) {
        if (is.null(names))
            names <- 1:length(x)
        names <- as.character(names)
        o <- order(y, decreasing = TRUE)
        i <- o[1:highlight]
```

```
        text(x[i], y[i], labels = substring(names[i], 1, 8),
            cex = 0.8, col = "blue")
    }
    invisible()
}

# ------------------------------------------------------------------ #
#  SAVING THE RESULTS FROM THE T-TESTS...
# ------------------------------------------------------------------ #

getwd()

# Option 1:

resultsOut(Fit.detail = fit_medium.1, Fit.rank = list(low=fit_low.1,
high=fit_high.1),
coef = 1, cDNA = TRUE,
write.file = TRUE, filename = "Coculture vs monoculture, using method
1.txt",
Raw.data = RG, High.Slides = high.slides, test.gene=2012)

resultsOut(Fit.detail = fit_medium.1, Fit.rank = NULL,
coef = 1, cDNA = TRUE,
write.file = TRUE, filename = "Probiotic vs Control, using method 1,
TEST.txt",
Raw.data = RG, High.Slides = high.slides, test.gene=2012)

        getwd()
# Option 2:

resultsOut(Fit.detail = fit_medium.2, Fit.rank = list(low=fit_low.2,
high=fit_high.2), coef = 2, cDNA = TRUE,
write.file = TRUE, filename = "Coculture, using method 2.txt",
Raw.data = RG, High.Slides = high.slides, test.gene=1)
```

Appendix V Gene list

Table A7 Gene list

ORF	Gene Symbol	Protein Name	EC number	GO*		
				Function	Process	Component
pCY360						
1	*repB*	Replication initiation protein		0003917	0006270	
2		Hypothetical protein		0005554	0000004	
3		Hypothetical protein		0005554	0000004	
4		Hypothetical protein		0005554	0000004	
5	*xerCA*	Putative integrase/recombinase		0009009	0006310	
6		Hypothetical protein		0005554	0000004	
7		Hypothetical protein		0005554	0000004	
8		Hypothetical protein		0005554	0000004	
9		Hypothetical transmembrane intracellular protein		0005554	0000004	0016020
10		Membrane protein		0005554	0000004	0016020
11		Hypothetical protein		0005554	0000004	
12		Uncharacterised protein		0005554	0000004	
13	*pinA*	PilT N-terminus domain protein		0015070 / 006276		
14		Hypothetical protein		0005554	0000004	
15		Hypothetical protein		0005554	0000004	
16		PilZ domain protein		0005554	0000004	
17		Hypothetical protein		0005554	0000004	
18	*ftsHA*	ATP-dependent Zinc metalloproteinase	3.4.24.-	0004176 / 008237		0016020
19		Hypothetical transmembrane intracellular protein		0005554	0000004	0016020
20		Hypothetical protein		0005554	0000004	
21		Hypothetical protein		0005554	0000004	
22		Putative Appr-1-p processing protein		0004721		
23		Putative lipoprotein		0005554	0000004	0016020
24		Hypothetical transmembrane intracellular protein		0005554	0000004	0016020
25		Hypothetical protein		0005554	0000004	
26		Hypothetical protein		0005554	0000004	
27		Hypothetical protein		0005554	0000004	
28		Hypothetical protein		0005554	0000004	
29		Hypothetical protein		0005554	0000004	
30		Hypothetical protein		0005554	0000004	
31		Hypothetical protein		0005554	0000004	
32		Hypothetical protein		0005554	0000004	
33		Hypothetical protein		0005554	0000004	
34		Hypothetical protein		0005554	0000004	
35		Hypothetical protein		0005554	0000004	
36		Hypothetical protein		0005554	0000004	
37		Hypothetical protein		0005554	0000004	
38		Conserved hypothetical protein DUF1016		0005554	0000004	
39		Hypothetical protein		0005554	0000004	
40		Hypothetical protein		0005554	0000004	
41		Hypothetical protein		0005554	0000004	
42	*rnaHA*	Ribonuclease H	3.1.26.4	0004523	0006401	
43		Hypothetical protein		0005554	0000004	

44		Hypothetical protein		0005554	0000004	
45		Hypothetical protein		0005554	0000004	
46		Hypothetical protein		0005554	0000004	
47		Putative helicase		0004386	0000004	
48		Hypothetical protein		0005554	0000004	
49		Hypothetical protein		0005554	0000004	
50		Hypothetical protein		0005554	0000004	
51		Hypothetical protein		0005554	0000004	
52		Hypothetical protein		0005554	0000004	
53		Hypothetical protein		0005554	0000004	
54		Hypothetical protein		0005554	0000004	
55		Hypothetical transmembrane protein		0005554	0000004	0016020
56		Hypothetical protein		0005554	0000004	
57		Hypothetical protein		0005554	0000004	
58		Hypothetical protein		0005554	0000004	
59		Hypothetical protein		0005554	0000004	
60		Hypothetical protein		0005554	0000004	
61		Hypothetical protein		0005554	0000004	
62		Hypothetical protein		0005554	0000004	
63		Uncharacterised protein		0005554	0000004	
64		Hypothetical protein		0005554	0000004	
65		Hypothetical protein		0005554	0000004	
66		Hypothetical protein		0005554	0000004	
67		Hypothetical protein		0005554	0000004	
68		Hypothetical protein		0005554	0000004	
69		Hypothetical transmembrane protein		0005554	0000004	0016020
70		Hypothetical transmembrane protein		0005554	0000004	0016020
71		Hypothetical protein		0005554	0000004	
72		Hypothetical protein		0005554	0000004	
73		Hypothetical protein		0005554	0000004	
74		Hypothetical protein		0005554	0000004	
75		Hypothetical protein		0005554	0000004	
76		Hypothetical protein		0005554	0000004	
77		Hypothetical protein		0005554	0000004	
78		Hypothetical protein		0005554	0000004	
79		Conserved hypothetical protein		0005554	0000004	
80		Site-specific recombinase/Phage integrase-family protein		0009009	0006310	
81		Uncharacterised conserved protein		0005554	0000004	
82		Hypothetical transmembrane protein		0005554	0000004	0016020
83		Conserved hypothetical protein		0005554	0000004	
84		Hypothetical protein		0005554	0000004	
85		Hypothetical protein		0005554	0000004	
86		Hypothetical protein		0005554	0000004	
87		Hypothetical protein		0005554	0000004	
88		Hypothetical protein		0005554	0000004	
89		Hypothetical protein		0005554	0000004	
90	*tnpAA*	Putative IS4-family transposase		0004803	0006313	
91	*tnpAB*	Putative IS4-family transposase		0004803	0006313	
92		Hypothetical protein		0005554	0000004	
93		Conserved hypothetical protein		0005554	0000004	
94		Hypothetical protein		0005554	0000004	
95	*tnpBA*	IS605-family transposase		0004803	0006313	
96		Hypothetical protein		0005554	0000004	
97		Hypothetical protein		0005554	0000004	
98		Hypothetical protein		0005554	0000004	
99		Hypothetical protein		0005554	0000004	
100		Hypothetical protein		0005554	0000004	

101		Hypothetical protein		0005554	0000004	
102		Hypothetical protein		0005554	0000004	
103		Hypothetical protein		0005554	0000004	
104		Hypothetical protein		0005554	0000004	
105		Hypothetical protein		0005554	0000004	
106		DNA modification methylase		0009008	0006306	
107		Putative DNA-binding protein		0003677	0000004	
108		Hypothetical protein		0005554	0000004	
109		Hypothetical protein		0005554	0000004	
110		Hypothetical protein		0005554	0000004	
111		Hypothetical protein		0005554	0000004	
112		Hypothetical protein		0005554	0000004	
113		Hypothetical protein		0005554	0000004	
114		Hypothetical protein		0005554	0000004	
115		Uncharacterised protein		0005554	0000004	
116		Hypothetical protein		0005554	0000004	
117		Conserved hypothetical protein		0005554	0000004	
118		Conserved hypothetical protein		0005554	0000004	
119		Hypothetical protein		0005554	0000004	
120		Hypothetical protein		0005554	0000004	
121		Hypothetical protein		0005554	0000004	
122		Hypothetical protein		0005554	0000004	
123		Hypothetical protein		0005554	0000004	
124		Hypothetical protein		0005554	0000004	
125		Hypothetical protein		0005554	0000004	
126		Hypothetical protein		0005554	0000004	
127		Hypothetical protein		0005554	0000004	
128		Hypothetical protein		0005554	0000004	
129		Hypothetical protein		0005554	0000004	
130		Hypothetical protein		0005554	0000004	
131		Hypothetical protein		0005554	0000004	
132		Hypothetical protein		0005554	0000004	
133		Hypothetical protein		0005554	0000004	
134		Hypothetical protein		0005554	0000004	
135		Hypothetical protein		0005554	0000004	
136	*tnpBB*	IS605-family transposase,		0004803	0006313	
137		Hypothetical protein		0005554	0000004	
138		Hypothetical protein		0005554	0000004	
139		Hypothetical protein		0005554	0000004	
140		Hypothetical protein		0005554	0000004	
141		Archaeal histone-like protein		0005554	0000004	
142		Hypothetical protein		0005554	0000004	
143		Hypothetical protein		0005554	0000004	
144		Hypothetical protein		0005554	0000004	
145		Hypothetical protein		0005554	0000004	
146		Hypothetical protein		0005554	0000004	
147		Hypothetical protein		0005554	0000004	
148		Hypothetical protein		0005554	0000004	
149		Conserved hypothetical protein		0005554	0000004	
150	*ligA*	NAD-dependent DNA ligase	6.5.1.2	0003911	0006260 / 006281 / 006310	
151	*tnpBC*	IS605-family transposase		0004803	0006313	
152		Hypothetical protein		0005554	0000004	
153		Conserved hypothetical protein		0005554	0000004	
154		Hypothetical exported protein		0005554	0000004	0005576
155		Hypothetical protein		0005554	0000004	
156		Hypothetical transmembrane protein		0005554	0000004	0016020

157		Hypothetical protein		0005554	0000004	
158		Hypothetical transmembrane protein		0005554	0000004	0016020
159		Conserved hypothetical protein		0005554	0000004	
160		Hypothetical protein		0005554	0000004	
161		Hypothetical protein		0005554	0000004	
162		Serine/Threonine protein phosphatase	3.1.3.16	0004722	0006470	
163		Hypothetical protein		0005554	0000004	
164		Hypothetical protein		0005554	0000004	
165		Hypothetical protein		0005554	0000004	
166		Hypothetical protein		0005554	0000004	
167		Hypothetical protein		0005554	0000004	
168		Hypothetical protein		0005554	0000004	
169		Hypothetical protein		0005554	0000004	
170		Hypothetical protein		0005554	0000004	
171		Hypothetical protein		0005554	0000004	
172	*mobA*	Putative mobilisation protein		0003916		
173		Hypothetical membrane protein		0005554	0000004	0016020
174		Hypothetical secreted protein		0005554	0000004	
175		Uncharacterised conserved protein		0005554	0000004	
176		Conserved hypothetical protein		0005554	0000004	
177		Conserved hypothetical protein		0005554	0000004	
178		Conserved hypothetical protein		0005554	0000004	
179		Conserved hypothetical protein		0005554	0000004	
180		Hypothetical protein		0005554	0000004	
181		Hypothetical protein		0005554	0000004	
182		Hypothetical protein		0005554	0000004	
183		Hypothetical protein		0005554	0000004	
184	*resA*	Putative DNA recombinase/resolvase		0000150		
185	*resB*	DNA recombinase/resolvase		0000150		
186		Hypothetical protein		0005554	0000004	
187		Hypothetical protein		0005554	0000004	
188	*resC*	Site-specific recombinase/resolvase		0009009	0006310	
189		Hypothetical protein		0005554	0000004	
190		Hypothetical protein		0005554	0000004	
191		Hypothetical protein		0005554	0000004	
192		Hypothetical protein		0005554	0000004	
193		Hypothetical protein		0005554	0000004	
194		Hypothetical protein		0005554	0000004	
195		Hypothetical protein		0005554	0000004	
196		Hypothetical protein		0005554	0000004	
197		Hypothetical protein		0005554	0000004	
198		Hypothetical protein		0005554	0000004	
199		Conserved hypothetical protein		0005554	0000004	
200		Hypothetical protein		0005554	0000004	
201		Hypothetical protein		0005554	0000004	
202		Hypothetical protein		0005554	0000004	
203		Hypothetical protein		0005554	0000004	
204		Hypothetical protein		0005554	0000004	
205		Hypothetical protein		0005554	0000004	
206		Uncharacterised protein		0005554	0000004	
207		Hypothetical protein		0005554	0000004	
208	*hupA*	Putative DNA-binding protein HU		0006310		
209		Hypothetical protein		0005554	0000004	
210	*uvrD*	ATP-dependent DNA helicase	3.6.1.-	0004003		
211	*parpA*	Putative poly(ADP-ribose) polymerase	2.4.2.30		0006974	
212		Hypothetical protein		0005554	0000004	
213		Conserved hypothetical protein		0005554	0000004	
214		Hypothetical secreted protein		0005554	0000004	0005576

215		Hypothetical secreted protein		0005554	0000004	0005576
216		Hypothetical secreted protein		0005554	0000004	0005576
217		Hypothetical secreted protein		0005554	0000004	0005576
218		Hypothetical transmembrane protein		0005554	0000004	0016020
219		Hypothetical protein		0005554	0000004	
220		HD domain protein		0016818	0000004	
221		Hypothetical protein		0005554	0000004	
222		Hypothetical secreted protein		0005554	0000004	0005576
223		Hypothetical protein		0005554	0000004	
224		Hypothetical protein		0005554	0000004	
225		Hypothetical protein		0005554	0000004	
226		Putative DNA polIII ε subunit		0006260		0009360
227		Hypothetical protein		0005554	0000004	
228		Hypothetical protein		0005554	0000004	
229		Hypothetical protein		0005554	0000004	
230		Hypothetical protein		0005554	0000004	
231		Hypothetical protein		0005554	0000004	
232	*traF*	Signal peptidase I	3.4.21.89	0009004	0009306	
233		Hypothetical protein		0005554	0000004	
234		Hypothetical protein		0005554	0000004	
235	*dnaEA*	DNA polymerase III, α subunit	2.7.7.7	0003887	0006260	0009360
236		Phosphoesterase, putatively DNA Pol III associated	3.1.3.15	0008081	0006260	0009360
237		Hypothetical protein		0005554	0000004	
238	*traE*	Putative TraE protein			0030255 / 009291	
239		Hypothetical transmembrane protein		0005554	0000004	0016020
240		Hypothetical membrane protein		0005554	0000004	0016020
241		Hypothetical transmembrane protein		0005554	0000004	0016020
242	*traG*	TraG-family protein			0030255 / 009291	0016020
243	*cspD*	Cold-shock protein-1		0003677 / 003723	0009409	
244		Hypothetical transmembrane protein		0005554	0000004	0016020
245		Hypothetical secreted protein		0005554	0000004	0005576
246		Hypothetical secreted protein		0005554	0000004	0005576
247		Hypothetical secreted protein		0005554	0000004	0005576
248		Hypothetical membrane protein		0005554	0000004	0016020
249		Putative type IV pilus subunit		0016887	0030255 / 009291	
250		Hypothetical protein		0005554	0000004	
251		Hypothetical protein		0005554	0000004	
252		Conserved hypothetical secreted protein		0005554	0000004	0005576
253	*nlpA*	NLP/P60-family lipoprotein		0005554	0000004	0016020
254	*nuA*	Thermonuclease	3.1.31.1	0004518		
255		Hypothetical protein		0005554	0000004	
256		Hypothetical protein		0005554	0000004	
257		Putative Lipoprotein		0005554	0000004	0016020
258	*udgA*	Uracil-DNA glycosylase	3.2.2.-	0004844	0006281	
259		Hypothetical transmembrane protein		0005554	0000004	0016020
260		Conserved hypothetical protein		0005554	0000004	
261		Hypothetical protein		0005554	0000004	
262		Hypothetical protein		0005554	0000004	
263		Hypothetical protein		0005554	0000004	
264	*trbB*	Putative ATPase involved in pili biogenesis		0016887	0030255 /0009291	
265	*pphA*	Serine/Threonine protein phosphatase	3.1.3.16	0004722	0043085 / 043086	

266		Hypothetical secreted protein		0005554	0000004	0005576
267		Hypothetical secreted protein		0005554	0000004	0005576
268		Hypothetical protein		0005554	0000004	
269	*topBA*	DNA topoisomerase III		0003917	0006265	
270		Hypothetical protein		0005554	0000004	
271		Hypothetical transmembrane protein		0005554	0000004	0016020
272		Hypothetical protein		0005554	0000004	
273		Hypothetical protein		0005554	0000004	
274		Hypothetical protein		0005554	0000004	
275	*recJA*	Putative single-stranded-DNA-specific exonuclease		0008297	0006310	
276		Hypothetical protein		0005554	0000004	
277		Hypothetical protein		0005554	0000004	
278		Hypothetical protein		0005554	0000004	
279	*sipB*	Signal peptidase I	3.4.21.89	0009004	0009306	
280		Hypothetical protein		0005554	0000004	
281		Hypothetical secreted protein		0005554	0000004	0005576
282	*recG*	ATP-dependent DNA helicase	3.6.1.-	0004003		
283		Hypothetical protein		0005554	0000004	
284		Hypothetical transmembrane protein		0005554	0000004	0016020
285		Hypothetical protein		0005554	0000004	
286		Hypothetical transmembrane protein		0005554	0000004	0016020
287		Hypothetical protein		0005554	0000004	
288	*tnpAB*	IS4-family transposase		0004803	0006313	
289		Cell surface CnaB-domain containing protein		0005554	0000004	0016020
290	*rpoD*	DNA-directed RNA polymerase sigma70 factor		0016986	0006350	0005665
291		Hypothetical protein		0005554	0000004	
292		Putative RecD/TraA-family helicase		0004386		
293		Hypothetical protein		0005554	0000004	
294		Hypothetical protein		0005554	0000004	
295		Hypothetical transmembrane protein		0005554	0000004	0016020
296		Hypothetical protein		0005554	0000004	
297		Hypothetical transmembrane protein		0005554	0000004	0016020
298		Hypothetical secreted protein		0005554	0000004	0005576
299		Conserved hypothetical protein		0005554	0000004	
300		Hypothetical protein		0005554	0000004	
301		Hypothetical protein		0005554	0000004	
302		Hypothetical secreted protein		0005554	0000004	0005576
303		Hypothetical protein		0005554	0000004	
304		Hypothetical secreted protein		0005554	0000004	0005576
305		Hypothetical secreted protein		0005554	0000004	0005576
306		Hypothetical protein		0005554	0000004	
307		Hypothetical transmembrane protein		0005554	0000004	0016020
308		Hypothetical secreted protein		0005554	0000004	0005576
309		Hypothetical protein		0005554	0000004	
310		Hypothetical protein		0005554	0000004	
311		Putative membrane-bound di-guanylate cyclase		0043789	0019934	0016020
312		Hypothetical protein		0005554	0000004	
313		Hypothetical protein		0005554	0000004	
314	*pcrA*	ATP-dependent DNA helicase	3.6.1.-	0004003		
315		Hypothetical protein		0005554	0000004	
316		Hypothetical protein		0005554	0000004	
317		Hypothetical protein		0005554	0000004	
318		Conserved hypothetical protein DUF600		0005554	0000004	
319		Hypothetical protein		0005554	0000004	

320		Hypothetical protein		0005554	0000004	
321		Hypothetical protein		0005554	0000004	
322		Hypothetical protein		0005554	0000004	
323		Hypothetical protein		0005554	0000004	
324		Hypothetical protein		0005554	0000004	
325		Hypothetical protein		0005554	0000004	
326		Hypothetical protein		0005554	0000004	
327	*tnpAC*	IS4-family transposase,		0004803	0006313	
328		Hypothetical protein		0005554	0000004	
329		Hypothetical protein		0005554	0000004	
330		Putative SCP-like extracellular protein		0005554	0000004	0005576
331	*tnpBD*	Putative IS605-family transposase		0004803	0006313	
332		Hypothetical protein		0005554	0000004	
333		Hypothetical protein		0005554	0000004	
334		Hypothetical protein		0005554	0000004	
335	*srtBA*	Putative Sortase B-family protein			0016044	
336		Hypothetical protein		0005554	0000004	
337		Hypothetical protein		0005554	0000004	
338		Putative Trigger factor protein,		0003754		
339	*srtBB*	Putative Sortase B-family protein			0016044	
340		Cell surface CnaB-domain containing protein		0005554	0000004	0016020
341		Hypothetical protein		0005554	0000004	
342		Putative SCP-like extracellular protein		0005554	0000004	0005576
343	*tnpBE*	Putative IS605-family transposase		0004803	0006313	
344		Hypothetical protein		0005554	0000004	
345		Hypothetical protein		0005554	0000004	
346	*gmkAA*	Guanylate kinase	2.7.4.8	0004385		
347		Hypothetical transmembrane protein		0005554	0000004	0016020
348		Conserved hypothetical protein		0005554	0000004	
349		Uncharacterised conserved protein DUF1064		0005554	0000004	
350		Hypothetical secreted protein		0005554	0000004	0005576
351		Hypothetical protein		0005554	0000004	
352		Hypothetical protein		0005554	0000004	
353		Hypothetical protein		0005554	0000004	
354		Hypothetical secreted protein		0005554	0000004	0005576
355		Hypothetical protein		0005554	0000004	
356		Hypothetical protein		0005554	0000004	
357		Hypothetical protein		0005554	0000004	
358		Hypothetical protein		0005554	0000004	
359		Putative superfamily II helicase		0004386		
360		Hypothetical protein, similar to phage proteins		0005554	0000004	
361	*tnpBF*	Putative IS605-family transposase		0004803	0006313	
362	*tnpCA*	IS200-family transposase,		0004803	0006313	
363		Hypothetical protein, similar to bacteriophage f237 ORF8		0005554	0000004	
364		Hypothetical protein, similar to phage proteins		0005554	0000004	
365		Hypothetical protein, similar to bacteriophage f237 ORF8		0005554	0000004	
366		Hypothetical protein, similar to bacteriophage f237 ORF8		0005554	0000004	
367		Hypothetical protein, similar to phage proteins		0005554	0000004	
368		Conserved hypothetical protein		0005554	0000004	
369		Conserved hypothetical protein		0005554	0000004	

370		Hypothetical protein		0005554	0000004	
371		Hypothetical protein		0005554	0000004	
372		Hypothetical protein		0005554	0000004	
373		Hypothetical protein		0005554	0000004	
374		Hypothetical protein		0005554	0000004	
375		Hypothetical transmembrane protein		0005554	0000004	0016020
376		Putative CAAX amino terminal protease-family protein		0008487		0016020
377		Hypothetical protein		0005554	0000004	
378		Hypothetical protein		0005554	0000004	
379		Hypothetical protein		0005554	0000004	
380		Hypothetical protein		0005554	0000004	
381		Hypothetical transmembrane protein		0005554	0000004	0016020
382		Hypothetical protein (putative truncated ThyA)		0005554	0000004	
383	*nrdGA*	Anaerobic ribonucleoside triphosphate reductase activating protein	1.97.1.4	0009391		
384	*ssbA*	Single stranded DNA binding protein		0003697	0006260 / 006281 / 006310	
385	*nrdDA*	Anaerobic ribonucleoside triphosphate reductase	1.17.4.2	0016961		
386		Conserved hypothetical protein		0005554	0000004	
387		Hypothetical protein with lipase/acylhydrolase domain		0005554	0000004	
388	*parAA*	ParA-family protein		0016887	0030542	
389		Hypothetical protein		0005554	0000004	
BPc2						
1	*repB*	Replication initiation protein		0003917	0006270	
2		Putative AAA+ ATPase		0016887	0000004	
3		Hypothetical protein		0005554	0000004	
4	*hypA*	Putative hydrogenase expression/formation protein		0016151		
5		Conserved hypothetical protein DUF111		0005554	0000004	
6	*hypB*	Hydrogenase expression/formation protein HypB		0016151		
7		Di-guanylate cyclase		0043789	0019934	
8		Di-guanylate cyclase		0043789	0019934	
9	*matEA*	Multi antimicrobial extrusion protein		0015297	0006855	0016020
10		Phosphate butyryltransferase	2.3.1.19		0019605	
11		Phosphate acetyltransferase	2.3.1.8	0008959		
12		Conserved hypothetical protein		0005554	0000004	
13		Conserved hypothetical protein		0005554	0000004	
14		Band 7 protein (Flotillin)		0005554	0000004	0005576
15	*fuc29A*	α-L-Fucosidase	3.2.1.51	0004560	0005975	
16		Hypothetical protein		0005554	0000004	
17	*man38A*	Putative α -mannosidase	3.2.1.24	0004559	0005975	
18		Putative PHP-family phosphoesterase		0005554	0000004	
19		Putative AraC-family transcriptional regulator		0003700 / 0003677	0006355	
20		Unsaturated glucuronyl hydrolase	3.2.1.-	0016798		
21		Putative ROK-family glucokinase		0004340		
22		Conserved hypothetical protein		0005554	0000004	
23	*oppAA*	Oligopeptide ABC transporter oligopeptide-binding protein		0005215 / 0015198	0006857	0005576
24	*oppBA*	Oligopeptide ABC transporter, permease protein		0005215 / 0015198	0006857	0016020

				/ 0015406 / 0015637		
25	*oppCA*	Oligopeptide ABC transporter, permease protein		0005215 / 015198 / 0015406 / 0015637	0006857	0016020
26	*oppDA*	Oligopeptide ABC transporter, ATP-binding protein		0005215 / 015198 / 0042626	0006857	
27	*oppFA*	Oligopeptide ABC transporter, ATP-binding protein		0005215 / 015198 / 0042626	0006857	
28		Glucosamine-6-phosphate isomerase / 6-phosphogluconolactonase bifunctional protein	3.5.99.6 / 3.1.1.31	0004342 / 0017057	0006046	
29		Putative β-lactamase		0008800		
30		Hypothetical protein		0005554	0000004	
31	*nrdGB*	Putative anaerobic ribonucleotide-triphosphate reductase activating protein	1.97.1.4	0009391		
32	*nrdDB*	Putative anaerobic ribonucleoside-triphosphate reductase	1.17.4.2	0016961		
33	*nrdDC*	Anaerobic ribonucleotide-triphosphate reductase	1.17.4.2	0016961		
34		Putative glutaredoxin		0009487		
35	*corAA*	Putative magnesium and cobalt transporter		0015095 / 0015087	0015693 / 006824	0016020
36	*pnuC*	Nicotinamide mononucleotide transporter		0015663	0015890	0016020
37	*nadR*	Nicotinamide-nucleotide adenylyltransferase	2.7.7.1	0015663	0009435	
38		Hypothetical protein		0005554	0000004	
39		Hypothetical protein		0005554	0000004	
40		Hypothetical protein containing UBA/TS-N domain		0005554	0000004	
41		MoxR-like AAA+ ATPase		0016887	0000004	
42		Conserved hypothetical protein, DUF58		0005554	0000004	
43		Putative protease		0008233		0016020
44		Conserved transmembrane protein		0005554	0000004	0016020
45		ABC transporter ATP binding protein		0043190 / 0005524		
46		Putative permease		0015646		0016020
47	*cheA*	Chemotaxis protein CheA			0006935 / 007165	
48	*cheW*	Chemotaxis protein CheW			0006935	
49	*msrAA*	Peptide methionine sulfoxide reductase	1.8.4.11	0016491	0030091	
50		Putative extracellular solute-binding protein		0005215	0000004	0005576
51	*mcpD*	Methyl-accepting chemotaxis protein			0006935 / 007165	0016020
52		Conserved hypothtical protein DUF156		0005554	0000004	
53		Heavy metal (copper) translocating P-type ATPase		0015076 / 0015076		0016020
54	*acpD*	Acyl carrier protein phosphodiesterase	3.1.4.14	0000036 / 0008081		
	rrs	16*S* rRNA				
	rrf	5*S* rRNA				
	rrl	23*S* rRNA				
55		Putative 5‟-nucleotidase		0008252		0016020
56		C_gcaxxg_C_C-family protein		0005554	0000004	
57		Rhomboid-family protein		0008236		0016020

58		Hypothetical protein		0005554	0000004	
59		Conserved transmembrane protein		0005554	0000004	0016020
	rrs	16*S* rRNA				
	rrf	5*S* rRNA				
	rrl	23*S* rRNA				
60		Putative flagellar protein			0009296	0016020
61	*xyl39A*	Oligosaccharide-specific β-xylosidase,	3.2.1.37	0009044	0005976	
62		ABC-transporter ATP-binding protein		0043190 / 0005524		
63		ATP-dependent DNA helicase, UvrD/Rep-family	3.6.1.-	0004003	0006268	
64	*fccA*	Fumarate reductase	1.3.99.1	0009055	0009061 / 019645	0016020
65		Di-guanylate cyclase		0043789	0019934	0016020
66	*aga36a*	α -Galactosidase	3.2.1.22	0004557		
67		Conserved hypothetical protein DUF81		0005554	0000004	
68		Conserved membrane protein		0005554	0000004	0016020
69		6,7-dimethyl-8-ribityllumazine synthase pseudogene	6.3.3.-		0009231	
70		Hypothetical protein		0005554	0000004	
71		Putative amino acid permease		0015359	0006865	0016020
72		Hypothetical secreted protein		0005554	0000004	0005576
73		Hypothetical protein		0005554	0000004	
75		Putative amino acid permease		0015359	0006865	0016020
76		Hypothetical transmembrane protein		0005554	0000004	0016020
77		Hypothetical protein		0005554	0000004	
78		Hypothetical protein		0005554	0000004	
79		Hypothetical protein		0005554	0000004	
80	*tnpBG*	Putative IS605-family transposase		0004803	0006313	
81		Membrane protein		0005554	0000004	0016020
82		Hypothetical protein		0005554	0000004	
83		Hypothetical protein		0005554	0000004	
84		Conserved membrane spanning protein with CAAX amino terminal protease-family domain		0005554	0000004	0016020
85		Putative α/β-family hydrolase		0016787	0000004	
86		Conserved hypothetical protein		0005554	0000004	
87		Hypothetical protein		0005554	0000004	
88		Conserved hypothetical protein		0005554	0000004	
89	*tnpCB*	IS200-family transposase		0004803	0006313	
90		Conserved membrane spanning protein with CAAX amino terminal protease-family domain		0005554	0000004	0016020
91		Putative NUDIX-family protein,		0016787	0009132	
92		Putative sugar-phosphate isomerase	5.3.1.6	0016853	0007487	
93		α/β-family hydrolase,		0016787	0000004	0005576
94	*tnpDA*	Putative IS110-family transposase		0004803	0006313	
95	*bgn53A*	Putative endo-1,4-β-galactanase	3.2.1.89	0004553	0005975	
96		Hypothetical secreted protein		0005554	0000004	0005576
97		Phage integrase-family domain protein		0005554	0000004	
98		Hypothetical protein		0005554	0000004	
99		Conserved hypothetical protein		0005554	0000004	
100		RNA-metabolising metallo-β-lactamase		0005554	0016070	
101		Putative ATP-dependent metalloprotease		0008237 / 0004176	0030163	
102		Hypothetical protein		0005554	0000004	
103		Conserved hypothetical protein		0005554	0000004	
104		Hypothetical protein		0005554	0000004	

105		Hypothetical protein		0005554	0000004	
106		Hypothetical protein		0005554	0000004	
107		AAA+ ATPase		0016887	0000004	
108		Hypothetical protein		0005554	0000004	
109		Hypothetical protein		0005554	0000004	
110		Hypothetical protein		0005554	0000004	
111		DnaC-like DNA helicase		0004003	0006260	
112		Two-component response regulator		0000156	0000004	
113		HNH endonuclease domain protein		0004519	0000004	0005576
114		RNA-directed DNA-polymerase		0003964	0006410	
115		Hypothetical protein		0005554	0000004	
116		Putative integrase		0008907	0015074	
117		Hypothetical protein		0005554	0000004	
118	*istB*	IS21-like ATP-binding protein		0005524	0000004	
119		Hypothetical protein		0005554	0000004	
120		Putative transcriptional regulator		0003700 / 0003677	0006355	
121		Conserved hypothetical protein (UPF0027)		0005554	0000004	
122		Hypothetical protein		0005554	0000004	
123		Putative GNAT-family acetyltransferase		0016407	0000004	
124		Putative GNAT-family acetyltransferase		0016407	0000004	
125		Conserved hypothetical protein		0005554	0000004	
126		Hypothetical protein		0005554	0000004	
127	*tnpBH*	Putative IS605-family transposase		0004803	0006313	
128		Hypothetical protein		0005554	0000004	
129	*kptA*	RNA 2'-phosphotransferase	2.7.-.-	0008665	0009451	
130		Putative flavodoxin		0009457	0006118	
131		HAD-superfamily hydrolase (subfamily 1a)		0016787	0000004	
132		Putative NAD-dependent epimerase/dehydratase			0009225	
133		Conserved transmembrane protein		0005554	0000004	0016020
134		Conserved hypothetical protein		0005554	0000004	
135		HicA protein		0005554	0000004	
136		HicB protein		0005554	0000004	
137		Putative GNAT-family acetyltransferase		0016407	0000004	
138	*crcB*	CrcB protein		0005554	0000004	0016020
139		Major facilitator superfamily transporter		0015646	0005215	0016020
140		Putative TetR-family transcriptional regulator		0003700 / 0003677	0006355	
141		Radical SAM-family protein		0005554	0000004	
142		Sensory box protein		0005057	0007165	
143		Site-specific recombinase, phage integrase-family		0009009	0006310	
144		Thioesterase-family protein		0016788	0000004	
145		Conserved NUDIX-family protein		0016787	0009132	
146	*rpiB*	Ribose 5-phosphate isomerase	5.3.1.6	0004751	0005975	
147		Putative bacteriocin		0015470	0000004	
148		Putative clostripain endopeptidase	3.4.22.8	0004197		0005576
149		Conserved membrane protein		0005554	0000004	0016020
150		Conserved membrane protein		0005554	0000004	0016020
151		ABC transporter, ATP-binding protein		0043190 / 0042626		
152		TetR-family transcription regulator		0003700 / 0003677	0006355	
153	*cbp94A*	Putative cellobiose phosphorylase	2.4.1.20	0004645	0000271	
154		Hypothetical secreted protein		0005554	0000004	0005576

155		Hypothetical protein		0005554	0000004	
156		Hypothetical secreted protein		0005554	0000004	0005576
157		Iron-containing alcohol dehydrogenase	1.1.1.1	0004025	0006066	0005576
158		Conserved secreted protein		0005554	0000004	0005576
159	*ram78A*	Putative α -L-rhamnosidase	3.2.1.40	0004553	0005976	
160		ABC transporter efflux permease		0015435 / 0005524	0000004	0016020
161		ABC transporter efflux permease		0015435 / 0005524	0000004	0016020
162	*pul13A*	Pullulanase	3.2.1.41	0004553	0005976	0016020
163		Exodeoxyribonuclease III	3.1.11.2	0008853	0006284	
164		Conserved hypothetical protein		0005554	0000004	
165		Uncharacterised secreted protein		0005554	0000004	0005576
166		Phosphoglycerate / bisphosphoglycerate mutase	5.4.2.-	0016868	0006110	
167	*mateB*	Multi antimicrobial extrusion protein		0015297	0006855	0016020
168		Putative GNAT-family acetyltransferase		0016407	0000004	
169		MarR-family transcriptional regulator		0003700 / 0003677	0006355	
170		Hypothetical transmembrane protein		0005554	0000004	0016020
171		Hypothetical protein		0005554	0000004	
172		Conserved hypothetical protein		0005554	0000004	
173		Conserved hypothetical protein		0005554	0000004	
174		AraC-family trancriptional regulator		0003700 /0003677	0006355	
175		Radical SAM domain protein		0005554	0000004	
176		Radical SAM domain protein		0005554	0000004	
177		MarR-family transcriptional regulator		0003700 / 0003677	0006355	
178		Conserved hypothetical (cupin domain) protein		0005554	0000004	
179		Hypothetical protein		0005554	0000004	
180		Putative transcriptional regulator		0003700 /0003677	0006355	
181		Putative penicillin-binding protein		0008658	0000004	
182		Putative GNAT-family acetyltransferase		0016407	0000004	
183		Hypothetical secreted protein		0005554	0000004	0005576
184		MarR-family transcriptional regulator		0003700 / 0003677	0006355	
185		Conserved hypothetical protein DUF1801		0005554	0000004	
186		Conserved hypothetical protein COG4898		0005554	0000004	
187		Hypothetical protein		0005554	0000004	
188		Hypothetical secreted protein		0005554	0000004	0005576
189		Hypothetical protein		0005554	0000004	
190	*ansA*	L-Asparaginase	3.5.1.1	0004067	0006530	
191	*pelL9A*	Pectate lyase	4.2.2.2	0030570	0000272	0005576
192		Hypothetical protein		0005554	0000004	
193		Conserved hypothetical protein		0005554	0000004	
194		Hypothetical protein		0005554	0000004	
195		Hypothetical protein		0005554	0000004	
196		Putative S-adenosylmethionine-dependent methyltransferase		0008757	0000004	
197		Hypothetical protein		0005554	0000004	
198	*feoB*	Ferrous iron transport protein B		0015639	0015684	
199	*feoA*	Putative ferrous iron transport protein A		0005554	0015684	
200	*gh31C*	Glycoside hydrolase-family 31	3.2.1.-	0004553	0005976	
201		ABC transporter permease protein		0015406	0006810	0016020
202		ABC transporter permease protein		0015406	0006810	0016020

203		ABC transporter, solute-binding protein		0005215	0006810	0005576
204		AraC-family two component response regulator		0000156	0000160	
205		Putative sensory histidine kinase		0000155	0000160	0016020
206		Uncharacterised conserved protein		0005554	0000004	
207		EAL domain protein			0019934	
208		Putative XerD-family site-specific recombinase/ phage integrase,		0008907	0015074	
209	*cht3B*	Putative β-hexosaminidase	3.2.1.52	0004563	0006032	0005576
210	*lacZA*	β-Galactosidase	3.2.1.-	0004565	0005975	
211		Hypothetical protein		0005554	0000004	
212		Di-guanylate cyclase		0043789	0019934	
213		Putative sugar-binding protein,		0005529	0005975	
214		Hypothetical protein		0005554	0000004	
215	*ParA*	ParA-family protein		0016887	0030542	
216	*napAA*	Na+/H+ exchanger-family protein (antiporter)		0015385	0006818 / 006814	0016020
217		Hypothetical secreted protein		0005554	0000004	0005576
218		Hypothetical transmembrane protein		0005554	0000004	0016020
219	*bgl3B*	β-Glucosidase,	3.2.1.21	0008422	0005975	0016020
220		Sensor kinase		0000155	0000160	
221		LytR-family response regulator		0003700 /0003677	0006355	
222	*bioY*	BioY-family protein		0005554	0019351	
223	*birA*	BirA bifunctional protein	6.3.4.15	0016564 / 0004077	0006768	
224		Amino acid ABC transporter, ATP-binding protein (putative glutamine specific)		0015171 / 0042626	0006865	
225		Amino acid ABC transporter permease protein		0015171 / 0015359	0006865	0016020
226		Amino acid ABC transporter amino acid-binding protein		0015171 / 0015597	0006865	0005576
227		Hypothetical protein		0005554	0000004	
228		GntR-family transcriptional regulator		0003700 /0003677	0006355	
229		Putative Lipoprotein,		0005554	0000004	
230		Radical SAM superfamily protein		0005554	0000004	
231		Conserved membrane protein DUF1113		0005554	0000004	0016020
232		Conserved secreted protein		0005554	0000004	0005576
233		Putative Lipoprotein		0005554	0000004	0016020
234		Hypothetical membrane protein		0005554	0000004	0016020
235	*pyrE*	Putative orotate phosphoribosyltransferase	2.4.2.10	0004588	0006221	
236		Hypothetical protein		0005554	0000004	
237	*man2A*	Putative β-mannosidase	3.2.1.25	0004567	0005975	
238		Conserved hypothetical protein with HTH domain		0005554 / 0003677	0000004	
239	*cheY*	Chemotaxis protein		0040012 / 0000156	0006935	
240		Sugar ABC transporter, sugar-binding protein		0005215 / 0015145 / 0005529	0015749	0005576
241		Sugar ABC transporter, permease protein		0005215 / 0015145 / 0015406	0015749	0016020
242		Sugar ABC transporter, permease protein		0005215 / 0015145 / 0015406	0015749	0016020

243		Sugar ABC transporter, ATP binding protein		0005215 / 0015145 / 0042626	0015749	
244		Sugar ABC transporter, sugar-binding protein		0005215 / 0015145 / 0005529	0015749	0005576
245		Sensor histidine kinase		0000155	0000160	0016020
246		AraC-family response regulator		0003700 / 0003677	0006355	
247		Putative sortase B protein		0005554	0000004	0016020
248		Sugar ABC transporter Sugar-binding protein		0005215 / 0015145 / 0005529	0015749	0005576
249		Conserved di-guanylate cyclase protein		0043789	0019934	
250	*manA*	Mannose-6-phosphate isomerase	5.3.1.8	0004476	0005975	
251	*matEC*	Multi antimicrobial extrusion protein		0015297	0006855	0016020
252	*hbsB*	DNA binding protein HU		0003677 / 0008301	0030261	
253		2-component system sensor kinase/response regulator hybrid protein		0000155 / 0000156	0000160	0016020
254		Uncharacterised protein		0005554	0000004	
255	*parA*	ParA-family ATPase		0016887	0030542	
pCY186						
1	*repBA*	RepB-family replication initiation protein		0003917	0006270	
2		Hypothetical protein		0005554	0000004	
3		Hypothetical protein		0005554	0000004	
4		Hypothetical transmembrane protein		0005554	0000004	0016020
5		Hypothetical transmembrane protein		0005554	0000004	0016020
6		Hypothetical transmembrane protein		0005554	0000004	0016020
7		Hypothetical secreted protein		0005554	0000004	0005576
8		Hypothetical transmembrane protein		0005554	0000004	0016020
9		Hypothetical protein		0005554	0000004	
10	*tnpDB*	IS110-family transposase		0004803	0006313	
11		Hypothetical protein		0005554	0000004	
12		Hypothetical protein		0005554	0000004	
13		Hypothetical secreted protein		0005554	0000004	0005576
14	*traG*	TraG protein			0030255 / 009291	0016020
15		GT28-family glycosyl transferase	2.4.1.-	0005554	0005976	
16		Conserved hypothetical protein		0005554	0000004	
17		Hypothetical protein		0005554	0000004	
18		Hypothetical transmembrane protein		0005554	0000004	0016020
19		Hypothetical protein		0005554	0000004	
20		Hypothetical protein		0005554	0000004	
21		Hypothetical transmembrane protein		0005554	0000004	0016020
22		Hypothetical protein		0005554	0000004	
23		Hypothetical transmembrane protein		0005554	0000004	0016020
24		Hypothetical protein		0005554	0000004	
25		Hypothetical protein		0005554	0000004	
26		Hypothetical protein		0005554	0000004	
27		Hypothetical secreted protein		0005554	0000004	0005576
28		Hypothetical secreted protein		0005554	0000004	0005576
29		Hypothetical protein		0005554	0000004	
30		RNA-binding protein		0019843	0000004	
31		Hypothetical secreted protein		0005554	0000004	0005576
32		Hypothetical protein		0005554	0000004	
33		Hypothetical protein		0005554	0000004	

34		Hypothetical secreted protein		0005554	0000004	0005576
35		Hypothetical transmembrane protein		0005554	0000004	0016020
36		Hypothetical protein		0005554	0000004	
37		Hypothetical protein		0005554	0000004	
38		Hypothetical secreted protein		0005554	0000004	0005576
39		Hypothetical secreted protein		0005554	0000004	0005576
40		Hypothetical transmembrane protein		0005554	0000004	0016020
41		Hypothetical secreted protein		0005554	0000004	0005576
42		Hypothetical secreted protein		0005554	0000004	0005576
43		Hypothetical secreted protein		0005554	0000004	0005576
44		Hypothetical protein		0005554	0000004	
45		Hypothetical protein		0005554	0000004	
46		Hypothetical transmembrane protein		0005554	0000004	0016020
47		Single stranded DNA-binding protein		0003697	0006260 / 006281 / 006310	
48		Hypothetical protein		0005554	0000004	
49		Hypothetical transmembrane protein		0005554	0000004	0016020
50		Hypothetical protein		0005554	0000004	
51		Hypothetical protein		0005554	0000004	
52		Hypothetical protein		0005554	0000004	
53		Hypothetical transmembrane protein		0005554	0000004	0016020
54		Hypothetical protein		0005554	0000004	
55		Hypothetical protein		0005554	0000004	
56		Hypothetical protein		0005554	0000004	
57		Hypothetical protein		0005554	0000004	
58		Hypothetical protein		0005554	0000004	
59		Hypothetical protein		0005554	0000004	
60		Hypothetical protein		0005554	0000004	
61		Hypothetical secreted protein		0005554	0000004	0005576
62		Helicase domain-containing protein		0005554	0000004	
63		GDSL-family lipase/acylhydrolase		0016298 / 0004064	0000004	
64		Hypothetical protein		0005554	0000004	
65		Hypothetical transmembrane protein		0005554	0000004	0016020
66		Hypothetical protein		0005554	0000004	
67		Hypothetical protein		0005554	0000004	
68	*hupB*	DNA-binding protein		0003677	0006275 / 006355	
69		Hypothetical protein		0005554	0000004	
70		Hypothetical protein		0005554	0000004	
71		Hypothetical transmembrane protein		0005554	0000004	0016020
72		Hypothetical secreted protein		0005554	0000004	0005576
73		Hypothetical protein		0005554	0000004	
74		Hypothetical protein		0005554	0000004	
75		Hypothetical transmembrane protein		0005554	0000004	0016020
76		Hypothetical transmembrane protein		0005554	0000004	0016020
77		Hypothetical secreted protein		0005554	0000004	0005576
78		Hypothetical protein		0005554	0000004	
79		Hypothetical secreted protein		0005554	0000004	0005576
80		AAA+ ATPase		0016887	0000004	
81		Hypothetical protein		0005554	0000004	
82		Hypothetical transmembrane protein		0005554	0000004	0016020
83		Hypothetical protein		0005554	0000004	
84		Hypothetical protein		0005554	0000004	
85		Hypothetical protein		0005554	0000004	
86		Hypothetical protein		0005554	0000004	
87		Hypothetical transmembrane protein		0005554	0000004	0016020

88		Hypothetical transmembrane protein		0005554	0000004	0016020
89		Hypothetical protein		0005554	0000004	
90		Hypothetical protein		0005554	0000004	
91		Hypothetical protein		0005554	0000004	
92		Hypothetical protein		0005554	0000004	
93		Hypothetical transmembrane protein		0005554	0000004	0016020
94		Hypothetical protein		0005554	0000004	
95		Hypothetical protein		0005554	0000004	
96		Hypothetical transmembrane protein		0005554	0000004	0016020
97		Type II restriction endonuclease		0004519	0009307	
98	*hsdM*	Type I restriction modification system M subunit	2.1.1.72	0009008	0006306	
99	*hsdSA*	Type I restriction modification system S subunit		0009307	0006304	
100	*hsdSB*	Type I restriction modification system S subunit		0009307	0006304	
101		Type II restriction endonuclease		0004519	0009307	
102	*hsdR*	Type I restriction modification system R subunit		0009307	0015666	
103		Lipoprotein		0005554	0000004	
104		PASTA domain-containing protein		0005554	0000004	
105		Hypothetical protein		0005554	0000004	
106		Site-specific recombinase/resolvase		0009009	0006310	
107		Hypothetical protein		0005554	0000004	
108		Hypothetical protein		0005554	0000004	
109		Peptidase c14		0004201	0000004	
110		Hypothetical protein		0005554	0000004	
111		Hypothetical protein		0005554	0000004	
112	*cspB*	Cold shock protein		0003677 / 003723	0009409	
113		Hypothetical protein		0005554	0000004	
114		Hypothetical protein		0005554	0000004	
115		Hypothetical secreted protein		0005554	0000004	0005576
116		Protein kinase		0004672	0000004	
117		Hypothetical protein		0005554	0000004	
118		Hypothetical protein		0005554	0000004	
119	*repBB*	RepB-family replication initiation protein		0003917	0006270	
120		Hypothetical protein		0005554	0000004	
121	*dnaB*	Replicative DNA helicase		0004003	0006260	
122		Hypothetical protein		0005554	0000004	
123	*polCA*	DNA polymerase III	2.7.7.7	0003887	0006260	0009360
124		Hypothetical protein		0005554	0000004	
125		GDSL-family lipase/acylhydrolase		0016298 / 0004064	0000004	
126	*dnaA*	Chromosomal replication initiator protein		0003688	0006270	
127		DNA polymerase III, α-subunit	2.7.7.7	0003887	0006260	0009360
128	*ftsHB*	ATP-dependent Zinc Metalloprotease		0004176	0008237	
129		DNA polymerase III, α-subunit	2.7.7.7	0003887	0006260	0009360
130	*dnaX*	DNA polymerase III γ and τ subunits		0006461	0006260	0009360
131		Hypothetical protein		0005554	0000004	
132		Hypothetical protein		0005554	0000004	
133		Metallo-β-lactamase-family protein		0016787	0000004	
134		Hypothetical protein		0005554	0000004	
135		Hypothetical protein		0005554	0000004	
136		Hypothetical protein		0005554	0000004	
137		Hypothetical protein		0005554	0000004	
138		Hypothetical protein		0005554	0000004	
139	*repBC*	RepB-family replication initiation protein		0003917	0006270	

140		Hypothetical protein		0005554	0000004	
141		Hypothetical protein		0005554	0000004	
142		Hypothetical protein		0005554	0000004	
143		Hypothetical protein		0005554	0000004	
144		Hypothetical protein		0005554	0000004	
145		MerC-like restriction enzyme modulator protein	3.1.21.-		0032072	
146		McrB-like restriction enzyme			0009307	
147		Hypothetical transmembrane protein		0005554	0000004	
148		KAP-family NTPase		0017111	0000004	
149		Hypothetical protein		0005554	0000004	
150		Hypothetical secreted protein		0005554	0000004	0005576
151		Uncharacterised secreted protein		0005554	0000004	0005576
152		Hypothetical protein		0005554	0000004	
153		Hypothetical protein		0005554	0000004	
154		Hypothetical protein		0005554	0000004	
155		Hypothetical protein		0005554	0000004	
156		Hypothetical protein		0005554	0000004	
157		UmuC-like DNA repair protein		0003685	0006281	
158		Hypothetical protein		0005554	0000004	
159		Hypothetical protein		0005554	0000004	
160		Hypothetical protein		0005554	0000004	
161		Hypothetical protein		0005554	0000004	
162		Hypothetical protein		0005554	0000004	
163		Hypothetical protein		0005554	0000004	
164		Hypothetical protein		0005554	0000004	
165		Hypothetical protein		0005554	0000004	
166		Hypothetical protein		0005554	0000004	
167		Hypothetical protein		0005554	0000004	
168		Hypothetical protein		0005554	0000004	
169		Conserved hypothetical protein		0005554	0000004	
170		Hypothetical protein		0005554	0000004	
171		Hypothetical protein		0005554	0000004	
172		Hypothetical protein		0005554	0000004	
173		Hypothetical transmembrane protein		0005554	0000004	0016020
174		Hypothetical protein		0005554	0000004	
175		Hypothetical protein		0005554	0000004	
176		Hypothetical protein		0005554	0000004	
177		Uncharacterised protein		0005554	0000004	
178		Hypothetical transmembrane protein		0005554	0000004	0016020
179		Hypothetical protein		0005554	0000004	
180		Metallophosphoesterase		0008081	0000004	
181		Hypothetical protein		0005554	0000004	
182		Hypothetical protein		0005554	0000004	
183		Hypothetical protein		0005554	0000004	
184		Hypothetical protein		0005554	0000004	
185		Hypothetical transmembrane protein		0005554	0000004	0016020
186		Hypothetical protein		0005554	0000004	
187		Hypothetical protein		0004386	0000004	
188		Hypothetical protein		0005554	0000004	
189		Hypothetical protein		0005554	0000004	
190		Hypothetical protein		0005554	0000004	
191		Signal peptidase I	3.4.21.89	0009004	0009306	
192		Hypothetical protein		0005554	0000004	
193		Hypothetical protein		0005554	0000004	
194		Hypothetical protein		0005554	0000004	
195		Transposase IS1182-family		0004803	0006313	
196	*parA*	ParA-family partitioning protein		0016887	0030542	

197		Hypothetical protein		0005554	0000004	
198	*parB*	ParB-family partitioning protein		0003677	0030542	

*All proteins were additionally described with the gene ontology (GO): 0003694

References

Abajy, M. Y., J. Kopec, K. Schiwon, M. Burzynski, M. Doring, C. Bohn & E. Grohmann, (2007) A type IV-secretion-like system is required for conjugative DNA transport of broad-host-range plasmid pIP501 in Gram-positive bacteria. J Bacteriol **189**: 2487-2496.

Abeles, A. L., L. D. Reaves, B. Youngren-Grimes & S. J. Austin, (1995) Control of P1 plasmid replication by iterons. Mol Microbiol **18**: 903-912.

Accetto, T. & G. Avgustin, (2001) Non-specific DNAases from the rumen bacterium *Prevotella bryantii*. Folia Microbiol **46**: 33-35.

Accetto, T., M. Peterka & G. Avgustin, (2005) Type II restriction modification systems of *Prevotella bryantii* TC1-1 and *Prevotella ruminicola* 23 strains and their effect on the efficiency of DNA introduction via electroporation. FEMS Microbiol Lett **247**: 177-183.

Akiyama, Y., A. Kihara & K. Ito, (1996) Subunit a of proton ATPase F0 sector is a substrate of the FtsH protease in *Escherichia coli*. FEBS letters **399**: 26-28.

Alamuri, P. & R. J. Maier, (2004) Methionine sulphoxide reductase is an important antioxidant enzyme in the gastric pathogen *Helicobacter pylori*. Mol Microbiol **53**: 1397-1406.

Alexander, J. K., (1968) Purification and specificity of cellobiose phosphorylase from *Clostridium thermocellum*. J Biol Chem **243**: 2899-2904.

Allignet, J., N. Liassine & N. el Solh, (1998) Characterization of a staphylococcal plasmid related to pUB110 and carrying two novel genes, vatC and vgbB, encoding resistance to streptogramins A and B and similar antibiotics. Antimicrob Agents Chemother. **42**: 1794-1798.

Allison, D. G., M. R. Brown, D. E. Evans & P. Gilbert, (1990) Surface hydrophobicity and dispersal of *Pseudomonas aeruginosa* from biofilms. FEMS Microbiol Lett **59**: 101-104.

Alon, U., L. Camarena, M. G. Surette, B. Aguera y Arcas, Y. Liu, S. Leibler & J. B. Stock, (1998) Response regulator output in bacterial chemotaxis. Embo J **17**: 4238-4248.

Altermann, E. & T. R. Klaenhammer, (2003) GAMOLA: a new local solution for sequence annotation and analyzing draft and finished prokaryotic genomes. OMICS. **7**: 161-169.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller & D. J. Lipman, (1997) Basic local alignment search tool. Available on Proweb: http://www.proweb.org/proweb/Tools/selfblast.html

Amabile-Cuevas, C. F. & M. E. Chicurel, (1992) Bacterial plasmids and gene flux. Cell **70**: 189-199.

Ambrozic, J., D. Ferme, M. Grabnar, M. Ravnikar & G. Avgustin, (2001) The bacteriophages of ruminal prevotellas. Folia Microbiologica **46**: 37-39.

Amikam, D. & M. Benziman, (1989) Cyclic diguanylic acid and cellulose synthesis in *Agrobacterium tumefaciens*. J Bacteriol **171**: 6649-6655.

Amikam, D., O. Steinberger, T. Shkolnik & Z. Ben-Ishai, (1995) The novel cyclic dinucleotide 3'-5' cyclic diguanylic acid binds to p21ras and enhances DNA synthesis but not cell replication in the Molt 4 cell line. Biochem J **311** (Pt 3): 921-927.

Amils, R., N. Irazabal, D. Moreira, J. P. Abad & I. Marin, (1998) Genomic organization analysis of acidophilic chemolithotrophic bacteria using pulsed field gel electrophoretic techniques. Biochimie **80**: 911-921.

Ampe, F., J. L. Uribelarrea, G. M. Aragao & N. D. Lindley, (1997) Benzoate degradation via the ortho pathway in *Alcaligenes eutrophus* is perturbed by succinate. Appl Environ Microbiol **63**: 2765-2770.

Anderson, S., (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res **9**: 3015-3027.

Andrup, L., (1998) Conjugation in Gram-positive bacteria and kinetics of plasmid transfer. APMIS Suppl **84**: 47-55.

Andrup, L., J. Damgaard & K. Wassermann, (1993) Mobilization of small plasmids in *Bacillus thuringiensis* subsp. *israelensis* is accompanied by specific aggregation. J Bacteriol **175**: 6530-6536.

Anton, J., R. Amils, C. Smith & P. LopezGarcia, (1995) Comparative restriction maps of the archaeal megaplasmid pHM300 in different *Haloferax mediterranei* strains. Sys Appl Microbiol **18**: 439-447.

Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist & E. M. Zdobnov, (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res **29**: 37-40.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi & L. S. Yeh, (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res **32**: D115-119.

Arai, N., K. Arai & A. Kornberg, (1981) Complexes of Rep protein with ATP and DNA as a basis for helicase action. J Biol Chem **256**: 5287-5293.

Arambel, M. J., E. E. Bartley, G. S. Dufva, T. G. Nagaraja & A. D. Dayton, (1982) Effect of diet on amino and nucleic acids of rumen bacteria and protozoa. J Dairy Sci **65**: 2095-2101.

Aravind, L. & D. Landsman, (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res **26**: 4413-4421.

Arcus, V. L., P. B. Rainey & S. J. Turner, (2005) The PIN-domain toxin-antitoxin array in mycobacteria. Trends Microbiol **13**: 360-365.

Argueso, P. & I. K. Gipson, (2006) Quantitative analysis of mucins in mucosal secretions using indirect enzyme-linked immunosorbent assay. Meth Molec Biol **347**: 277-288.

Arp, D. J., P. S. Chain & M. G. Klotz, (2007) The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. Annu Rev Microbiol **61**: 503-528.

Asanuma, N. & T. Hino, (2000) Activity and properties of fumarate reductase in ruminal bacteria. J Gen Appl Microbiol **46**: 119-125.

Asanuma, N., M. Ishiwata, T. Yoshii, M. Kikuchi, Y. Nishina & T. Hino, (2005) Characterization and transcription of the genes involved in butyrate production in *Butyrivibrio fibrisolvens* type I and II strains. Curr Microbiol **51**: 91-94.

Attwood, G. T., K. Reilly & B. K. Patel, (1996) *Clostridium proteoclasticum* sp. nov., a novel proteolytic bacterium from the bovine rumen. Int J Syst Bacteriol **46**: 753-758.

Attwood, T. K., M. D. Croning, D. R. Flower, A. P. Lewis, J. E. Mabey, P. Scordis, J. N. Selley & W. Wright, (2000) PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res **28**: 225-227.

Austin, S. & A. Abeles, (1983) Partition of unit-copy miniplasmids to daughter cells. I. P1 and F miniplasmids contain discrete, interchangeable sequences sufficient to promote equipartition. J Mol Biol **169**: 353-372.

Axon Instruments, (2004) GenePix Pro 6.0 users guide.

Bach, A., S. Calsamiglia & M. D. Stern, (2005) Nitrogen metabolism in the rumen. J Dairy Sci **88**: E9-21.

Backhed, F., R. E. Ley, J. L. Sonnenburg, D. A. Peterson & J. I. Gordon, (2005) Host-bacterial mutualism in the human intestine. Science **307**: 1915-1920.

Bairoch, A., (1991) PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res **19**: Suppl:2241-2245.

Bandelt, H. J. & A. W. Dress, (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol **1**: 242-252.

Barcroft, J., R. A. McAnally & A. T. Phillipson, (1944) Absorption of volatile acids from the alimentary tract of sheep and other animals. J. Exp. Biol **20**: 120-129.

Barker, H. A., M. D. Kamen & B. T. Bornstein, (1945) The synthesis of butyric and caproic acids from ethanol and acetic acid by *Clostridium kluyveri*. Proc Natl Acad Sci USA **31**: 373-381.

Barton, B. M., G. P. Harding & A. J. Zuccarelli, (1995) A general-method for detecting and sizing large plasmids. Anal Biochem **226**: 235-240.

Bartosik, D., J. Baj, B. A.A. & M. Wlodarczyk, (2002) Characterization of the replicator region of megaplasmid pTAV3 of *Paracoccus versutus* and search for plasmid-encoded traits. Microbiol **148**: 871-881.

Bascand, G., (2007) Agricultural production statistics (provisional): June 2007 In.: Ministry of Agriculture and Forestry. Available: http://www.stats.govt.nz

Bayer, E., K. H. Gugel, K. Hagele, H. Hagenmaier, S. Jessipow, W. A. Konig & H. Zahner, (1972) Metabolic products of microorganisms: Phosphinothricin and phosphinothricyl-alanyl-analine. Helvetica chimica acta **55**: 224-239.

Beard, C. E., M. A. Hefford, R. J. Forster, S. Sontakke, R. M. Teather & K. Gregg, (1995) A stable and efficient transformation system for *Butyrivibrio fibrisolvens* OB156. Curr Microbiol **30**: 105-109.

Begum, A., M. M. Rahman, W. Ogawa, T. Mizushima, T. Kuroda & T. Tsuchiya, (2005) Gene cloning and characterization of four MATE family multidrug efflux pumps from *Vibrio cholerae* non-O1. Microb Immun **49**: 949-957.

Belfort, M., G. F. Maley & F. Maley, (1983) Characterization of the *Escherichia coli thyA* gene and its amplified thymidylate synthetase product. Proc Natl Acad Sci USA **80**: 1858-1861.

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

Benjamini, Y. & D. Yekutieli, (2005) Quantitative trait loci analysis using the false discovery rate. Genetics **171**: 783-790.

Benkel, B. F. & Y. Fong, (1996) Long range-inverse PCR (LR-IPCR): extending the useful range of inverse PCR. Genet Anal **13**: 123-127.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell & D. L. Wheeler, (2008) GenBank. Nucleic Acids Res **36**: D25-30.

Bentley, S. D., S. Brown, L. D. Murphy, D. E. Harris, M. A. Quail, J. Parkhill, B. G. Barrell, J. R. McCormick, R. I. Santamaria, R. Losick, M. Yamasaki, H. Kinashi, C. W. Chen, G. Chandra, D. Jakimowicz, H. M. Kieser, T. Kieser & K. F. Chater, (2004)

SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). Mol Microbiol **51**: 1615-1628.
Bergquist, P. L., S. Saadi & W. K. Maas, (1986) Distribution of basic replicons having homology with *repFIA*, *repFIB*, and *repFIC* among IncF group plasmids. Plasmid **15**: 19-34.

Bernal, A., U. Ear & N. Kyrpides, (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. NAR **29**: 126-127.

Bertani, G., (1951) Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. J Bacteriol **62**: 293-300.

Bertani, G., (2004) Lysogeny at mid-twentieth century: P1, P2, and other experimental systems. J Bacteriol 186: 595-600.

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

Bessman, M. J., D. N. Frick & S. F. O'Handley, (1996) The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. J Biol Chem **271**: 25059-25062.

Beveridge, T. J., (1990) Mechanism of Gram-variability in select bacteria. J Bacteriol **172**: 1609-1620.

Bierne, H., C. Garandeau, M. G. Pucciarelli, C. Sabet, S. Newton, F. Garcia-del Portillo, P. Cossart & A. Charbit, (2004) Sortase B, a new class of sortase in *Listeria monocytogenes*. J Bacteriol **186**: 1972-1982.

Birch, P. & S. A. Khan, (1992) Replication of single-stranded plasmid pT181 DNA *in-vitro*. Proc Natl Acad Sci USA **89**: 290-294.

Blackburn, T. H. & P. N. Hobson, (1962) Further studies on the isolation of proteolytic bacteria from the sheep rumen. J Gen Microbiol **29**: 69-81.

Blakesley, R. W., N. F. Hansen, J. C. Mullikin, P. J. Thomas, J. C. McDowell, B. Maskeri, A. C. Young, B. Benjamin, S. Y. Brooks, B. I. Coleman, J. Gupta, S. L. Ho, E. M. Karlins, Q. L. Maduro, S. Stantripop, C. Tsurgeon, J. L. Vogt, M. A. Walker, C. A. Masiello, X. Guan, G. G. Bouffard & E. D. Green, (2004) An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res **14**: 2235-2244.

Bock, A. K. & P. Schonheit, (1995) Growth of *Methanosarcina barkeri* (Fusaro) under nonmethanogenic conditions by the fermentation of pyruvate to acetate: ATP synthesis via the mechanism of substrate level phosphorylation. J Bacteriol **177**: 2002-2007.

Boekhorst, J., M. W. de Been, M. Kleerebezem & R. J. Siezen, (2005) Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. J Bacteriol **187**: 4928-4934.

Borodovsky, M. & J. McIninch, (1993) GeneMark: parallel gene recognition for both DNA strands. Comput Chem **17**: 123-133.

Bower, S., J. Perkins, R. R. Yocum, P. Serror, A. Sorokin, P. Rahaim, C. L. Howitt, N. Prasad, S. D. Ehrlich & J. Pero, (1995) Cloning and characterization of the *Bacillus subtilis birA* gene encoding a repressor of the biotin operon. J Bacteriol **177**: 2572-2575.

Boye, E. & K. Nordstrom, (2003) Coupling the cell cycle to cell growth. EMBO reports **4**: 757-760.

Brady, G., J. Frey, H. Danbara & K. N. Timmis, (1983) Replication control mutations of plasmid R6-5 and their effects on interactions of the RNA-I control element with its target. J Bacteriol. **154**: 429-436.

Bramhachari, P. V., P. B. Kishor, R. Ramadevi, R. Kumar, B. R. Rao & S. K. Dubey, (2007) Isolation and characterization of mucous exopolysaccharide (EPS) produced by *Vibrio furnissii* strain VB0S3. J Microbiol Biotech **17**: 44-51.

Brink, G. E. & T. E. Fairbrother, (1994) Cell-wall composition of diverse clovers during primary spring growth. Crop Sci **34**: 1666-1671.

Brock, F. M., C. W. Forsberg & J. G. Buchanan-Smith, (1982) Proteolytic activity of rumen microorganisms and effects of proteinase inhibitors. Appl Environ Microbiol **44**: 561-569.

Broker, D., M. Arenskotter, A. Legatzki, D. H. Nies & A. Steinbuchel, (2004) Characterization of the 101-kilobase-pair megaplasmid pKB1, isolated from the rubber-degrading bacterium *Gordonia westfalica* Kb1. J Bacteriol **186**: 212-225.

Brom, S., A. Garcia-de los Santos, L. Cervantes, R. Palacios & D. Romero, (2000) In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons. Plasmid **44**: 34-43.

Brot, N., L. Weissbach, J. Werth & H. Weissbach, (1981) Enzymatic reduction of protein-bound methionine sulfoxide. Proc Natl Acad Sci USA **78**: 2155-2158.

Bruand, C., S. D. Ehrlich & L. Janniere, (1991) Unidirectional theta replication of the structurally stable *Enterococcus faecalis* plasmid pAM beta 1. Embo J **10**: 2171-2177.

Bruand, C., E. Le Chatelier, S. D. Ehrlich & L. Janniere, (1993) A fourth class of theta-replicating plasmids: the pAM beta 1 family from Gram-positive bacteria. Proc Natl Acad Sci USA **90**: 11668-11672.

Brudno, M., A. Poliakov, S. Minovitsky, I. Ratnere & I. Dubchak, (2007) Multiple whole genome alignments and novel biomedical applications at the VISTA portal. Nucleic Acids Res **35**: W669-674.

Bryant, M. P. & N. Small, (1956) The anaerobic monotrichous butyric acid-producing curved rod-shaped bacteria of the rumen. J Bacteriol **72**: 16-21.
Buchan, D. W., S. C. Rison, J. E. Bray, D. Lee, F. Pearl, J. M. Thornton & C. A. Orengo, (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. Nucleic Acids Res **31**: 469-473.

Buchko, G. W., O. Litvinova, H. Robinson, A. F. Yakunin & M. A. Kennedy, (2008) Functional and structural characterization of DR_0079 from *Deinococcus radiodurans*, a novel Nudix hydrolase with a preference for cytosine (deoxy)ribonucleoside 5'-di- and triphosphates. Biochem **47**: 6571-6582

Bulach, D. M., R. L. Zuerner, P. Wilson, T. Seemann, A. McGrath, P. A. Cullen, J. Davis, M. Johnson, E. Kuczek, D. P. Alt, B. Peterson-Burch, R. L. Coppel, J. I. Rood, J. K. Davies & B. Adler, (2006) Genome reduction in *Leptospira borgpetersenii* reflects limited transmission potential. Proc Natl Acad Sci USA **103**: 14560-14565.

Butler, J. E., Q. He, K. P. Nevin, Z. He, J. Zhou & D. R. Lovley, (2007) Genomic and microarray analysis of aromatics degradation in *Geobacter metallireducens* and comparison to a *Geobacter* isolate from a contaminated field site. BMC Genomics **8**: 180.

Campbell, A., R. Chang, D. Barker & G. Ketner, (1980) Biotin regulatory (bir) mutations of *Escherichia coli*. J Bacteriol **142**: 1025-1028.

Campbell, L. L., Jr., (1957) Reductive degradation of pyrimidines. II. Mechanism of uracil degradation by *Clostridium uracilicum*. J Bacteriol **73**: 225-229.

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard & B. Henrissat, (2008) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res **37**: D233-D238

Carraro, D. M., A. A. Camargo, A. C. Salim, L. Gonzaga, G. C. Costa, A. T. Vasconcelos & A. J. Simpson, (2004) Closure of rRNA related gaps in the *Chromobacterium violaceum* genome with the PCR-assisted contig extension (PACE) protocol. Genet Mol Res **3**: 53-63.

Casjens, S., N. Palmer, R. van Vugt, W. M. Huang, B. Stevenson, P. Rosa, R. Lathigra, G. Sutton, J. Peterson, R. J. Dodson, D. Haft, E. Hickey, M. Gwinn, O. White & C. M. Fraser, (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the lyme disease spirochete *Borrelia burgdorferi*. Mol Microbiol **35**: 490-516.

Cavicchioli, R., (2002) Extremophiles and the search for extraterrestrial life. Astrobiol **2**: 281-292.

Chagan, I., M. Tokura, J. P. Jouany & K. Ushida, (1999) Detection of methanogenic archaea associated with rumen ciliate protozoa. J Gen Appl Microbiol. **45**: 305-308.

Chain, P. S., V. J. Denef, K. T. Konstantinidis, L. M. Vergez, L. Agullo, V. L. Reyes, L. Hauser, M. Cordova, L. Gomez, M. Gonzalez, M. Land, V. Lao, F. Larimer, J. J. LiPuma, E. Mahenthiralingam, S. A. Malfatti, C. J. Marx, J. J. Parnell, A. Ramette, P. Richardson, M. Seeger, D. Smith, T. Spilker, W. J. Sul, T. V. Tsoi, L. E. Ulrich, I. B. Zhulin & J. M. Tiedje, (2006) *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci USA **103**: 15280-15287.

Chan, C. K., A. L. Hsu, S. L. Tang & S. K. Halgamuge, (2008) Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. J Biomed Biotechnol **2008**: 513701.

Chargaff, E., (1951) Structure and function of nucleic acids as cell constituents. Fed. Proc **10**: 654-659.

Chen, J., Z. Lu, J. Sakon & W. E. Stites, (2000) Increasing the thermostability of Staphylococcal nuclease: implications for the origin of protein thermostability. J Mol Biol **303**: 125-130.

Chen, J. S., (1995) Alcohol dehydrogenase: multiplicity and relatedness in the solvent-producing clostridia. FEMS Microbiol Rev **17**: 263-273.

Chen, Q., J. R. Fischer, V. M. Benoit, N. P. Dufour, P. Youderian & J. M. Leong, (2008) *In-vitro* CpG methylation increases the transformation efficiency of *Borrelia burgdorferi* strains harboring the endogenous linear plasmid lp56. J Bacteriol **24**: 7885-7891

Cheng, K. J. & J. W. Costerton, (1977) Ultrastructure of *Butyrivibrio fibrisolvens*: a Gram-positive bacterium. J Bacteriol **129**: 1506-1512.

Chesbro, W. R. & K. Auborn, (1967) Enzymatic detection of the growth of *Staphylococcus aureus* in foods. Appl Microbiol **15**: 1150-1159.

Chesson, A., C. S. Stewart, K. Dalgarno & T. P. King, (1986) Degradation of isolated grass mesophyll, epidermis and fiber cell-walls in the rumen and by cellulolytic rumen bacteria in axenic culture. J Appl Bacteriol **60**: 327-336.

Choudhary, M., C. Mackenzie, K. Nereng, E. Sodergren, G. M. Weinstock & S. Kaplan, (1997) Low-resolution sequencing of *Rhodobacter sphaeroides* 2.4.1^{T}: chromosome II is a true chromosome. Microbiol **143**: 3085-3099.

Christie, P. J., (2001) Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. Mol Microbiol **40**: 294-305.

Christie, P. J., K. Atmakuri, V. Krishnamoorthy, S. Jakubowski & E. Cascales, (2005) Biogenesis, architecture, and function of bacterial type IV secretion systems. Annu Rev Microbiol **59**: 451-485.

Chung, K. T., (1976) Inhibitory effects of H_2 on growth of *Clostridium cellobioparum*. Appl Environ Microbiol **31**: 342-348.

Claesson, M. J., Y. Li, S. Leahy, C. Canchaya, J. P. van Pijkeren, A. M. Cerdeno-Tarraga, J. Parkhill, S. Flynn, G. C. O'Sullivan, J. K. Collins, D. Higgins, F. Shanahan, G. F. Fitzgerald, D. van Sinderen & P. W. O'Toole, (2006) Multireplicon genome architecture of *Lactobacillus salivarius*. Proc Natl Acad Sci USA **103**: 6718-6723.

Clark, B. F. & K. A. Marcker, (1966) The role of N-formyl-methionyl-sRNA in protein biosynthesis. J Mol Biol **17**: 394-406.

Clark, R. G., K. J. Cheng, L. B. Selinger & M. F. Hynes, (1994) A conjugative transfer system for the rumen bacterium, *Butyrivibrio fibrisolvens*, based on Tn916-mediated transfer of the *Staphylococcus aureus* plasmid pUB110. Plasmid **32**: 295-305.

Clarke, R. T. J., (1968) The microbiology of "pre-gastric" fermentation. Aust J Sci **31**: 141-146.

Cocconcelli, P. S., E. Triban, M. Basso & V. Bottazzi, (1991) Use of DNA probes in the study of silage colonization by *Lactobacillus* and *Pediococcus* strains. J Appl Bacteriol **71**: 296-301.

Coenye, T. & P. Vandamme, (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol Lett **228**: 45-49.

Cohen, S. N., (1976) Transposable genetic elements and plasmid evolution. Nature **263**: 731-738.

Cole, S. T. & I. Saint Girons, (1994) Bacterial genomics. FEMS Microbiol Rev **14**: 139-160.

Copeland, A., S. Lucas, A. Lapidus, K. Barry, J. C. Detter, T. Glavina, N. Hammon, S. Israni, S. Pitluck, P. Chain, S. Malfatti, M. Shin, L. Vergez, J. Schmutz, F. Larimer, M. Land, N. Kyrpides, A. Lykidis & P. Richardson, (2007) Complete sequence of chromosome 2 of *Burkholderia sp*. 383. Unpublished Data.

Cornet, F., I. Mortier, J. Patte & J. M. Louarn, (1994) Plasmid pSC101 harbors a recombination site, *psi*, which is able to resolve plasmid multimers and to substitute for the analogous chromosomal *Escherichia coli* site dif. J Bacteriol. **176**: 3188-3195.

Corpet, F., J. Gouzy & D. Kahn, (1998) The ProDom database of protein domain families. Nucleic Acids Res. **26**: 323-326.

Cotta, M. A., (1988) Amylolytic activity of selected species of ruminal bacteria. Appl Environ Microbiol **54**: 772-776.

Cotta, M. A., (1990) Utilization of nucleic acids by *Selenomonas ruminantium* and other ruminal bacteria. Appl Environ Microbiol **56**: 3867-3870.

Crawford, R. W., D. L. Gibson, W. W. Kay & J. S. Gunn, (2008) Identification of a bile-induced exopolysaccharide required for *Salmonella* biofilm formation on gallstone surfaces. Infect Immun **76**: 5341-5349.

Crespi, M., E. Messens, A. B. Caplan, M. van Montagu & J. Desomer, (1992) Fasciation induction by the phytopathogen *Rhodococcus fascians* depends upon a linear plasmid encoding a cytokinin synthase gene. EMBO J **11**: 795-804.

Culver, G. M., S. M. McCraith, S. A. Consaul, D. R. Stanford & E. M. Phizicky, (1997) A 2'-phosphotransferase implicated in tRNA splicing is essential in *Saccharomyces cerevisiae*. J Biol Chem **272**: 13203-13210.

Currier, T. C. & E. W. Nester, (1976) Isolation of covalently closed circular DNA of high-molecular weight from Bacteria. Anal Biochem **76**: 431-441.

D'Argenio, D. A. & S. I. Miller, (2004) Cyclic di-GMP as a bacterial second messenger. Microbiol **150**: 2497-2502.

Dahlberg, C. & L. Chao, (2003) Amelioration of the cost of conjugative plasmid carriage in *Eschericha coli* K12. Genetics **165**: 1641-1649.

Dale, H. E., R. E. Stewart & S. Brody, (1954) Rumen temperature. I. Temperature gradients during feeding and fasting. Cornell Vet **44**: 368-374.

Danese, P. N., L. A. Pratt & R. Kolter, (2000) Exopolysaccharide production is required for development of *Escherichia coli* K-12 biofilm architecture. J Bacteriol **182**: 3593-3596.

Dawson, K. A., M. C. Preziosi & D. R. Caldwell, (1979) Some effects of uncouplers and inhibitors on growth and electron transport in rumen bacteria. J Bacteriol **139**: 384-392.

de La Torre, F., J. Sampedro, I. Zarra & G. Revilla, (2002) AtFXG1, an *Arabidopsis* gene encoding alpha-L-fucosidase active against fucosylated xyloglucan oligosaccharides. Plant Physiol **128**: 247-255.

De Wilde, M., J. E. Davies & F. J. Schmidt, (1978) Low molecular weight RNA species encoded by a multiple drug resistance plasmid. Proc Natl Acad Sci USA **75**: 3673-3677.

Debroas, D. & G. Blanchart, (1993) Interactions between proteolytic and cellulolytic rumen bacteria during hydrolysis of plant cell wall protein. Repr Nutr Dev **33**: 283-288.

Dehority, B. A. & H. W. Scott, (1967) Extent of cellulose and hemicellulose digestion in various forages by pure cultures of rumen bacteria. J. Dairy Sci **50**: 1136-1141.

del Solar, G. & M. Espinosa, (2000) Plasmid copy number control: an ever-growing story. Mol Microbiol **37**: 492-500.
del Solar, G., R. Giraldo, M. J. Ruiz-Echevarria, M. Espinosa & R. Diaz-Orejas, (1998) Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev **62**: 434-464.

Delcher, A. L., D. Harmon, S. Kasif, O. White & S. L. Salzberg, (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res **27**: 4636-4641.

Diaz-Lopez, T., M. Lages-Gonzalo, A. Serrano-Lopez, C. Alfonso, G. Rivas, R. Diaz-Orejas & R. Giraldo, (2003) Structural changes in RepA, a plasmid replication initiator, upon binding to origin DNA. J Biol Chem **278**: 18606-18616.

Donachie, W. D., (1968) Relationship between cell size and time of initiation of DNA replication. Nature **219**: 1077-&.

Dougherty, B. A., C. Hill, J. F. Weidman, D. R. Richardson, J. C. Venter & R. P. Ross, (1998) Sequence and analysis of the 60 kb conjugative, bacteriocin-producing plasmid pMRC01 from *Lactococcus lactis* DPC3147. Mol Microbiol. **29**: 1029-1038.

Dramsi, S., P. Trieu-Cuot & H. Bierne, (2005) Sorting sortases: a nomenclature proposal for the various sortases of Gram-positive bacteria. Res Microbiol **156**: 289-297.

Dryselius, R., K. Kurokawa & T. Iida, (2007) *Vibrionaceae*, a versatile bacterial family with evolutionarily conserved variability. Res Microbiol **158**: 479-486.

DSMZ, (1993) DSMZ 704 medium. Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH. Available: http://www.dsmz.de/media/med704.htm.

Duenas-Gonzalez, A., M. Candelaria, C. Perez-Plascencia, E. Perez-Cardenas, E. de la Cruz-Hernandez & L. A. Herrera, (2008) Valproic acid as epigenetic cancer drug: Preclinical, clinical and transcriptional effects on solid tumors. Cancer Treat Rev. **34**: 206-222

Duncan, B. K., P. A. Rockstroh & H. R. Warner, (1978) *Escherichia coli* K-12 mutants deficient in uracil-DNA glycosylase. J Bacteriol **134**: 1039-1045.

Dunne, J., (in preparation) A proteomic analysis of fibre degradation by *Butyrivibrio proteoclasticus*. In: School of Biological Sciences. Victoria University, Wellington, New Zealand.

Durre, P., (2007) Biobutanol: an attractive biofuel. Biotechnol J **2**: 1525-1534.

Dutta, A., S. K. Singh, P. Ghosh, R. Mukherjee, S. Mitter & D. Bandyopadhyay, (2006) *In-silico* identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. In Silico Biol **6**: 43-47.

Eckhardt, T., (1978) A rapid method for the identification of plasmid desoxyribonucleic acid in bacteria. Plasmid **1**: 584-588.

Edwards, R. A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr. & F. Rohwer, (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics **7**: 57.

Egan, E. S., M. A. Fogel & M. K. Waldor, (2005) Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol **56**: 1129-1138.

Egan, E. S., A. Lobner-Olesen & M. K. Waldor, (2004) Synchronous replication initiation of the two *Vibrio cholerae* chromosomes. Curr Biol **14**: R501-502.

Eisenberg, M. A. & K. Krell, (1969) Synthesis of dethiobiotin from 7,8-diaminopelargonic acid in biotin auxotrophs of *Escherichia coli* K-12. J Bacteriol **98**: 1227-1231.

El-Mansi, M., K. J. Anderson, C. A. Inche, L. K. Knowles & D. J. Platt, (2000) Isolation and curing of the *Klebsiella pneumoniae* large indigenous plasmid using sodium dodecyl sulphate. Res Microbiol **151**: 201-208.

Elsden, S. R., (1946) The application of silica gel partition chromotography to the estimation of volatile fatty acids. Biochem. J. **40**: 252-256.

Endo, T., N. Sasaki, I. Tanaka & M. Nakata, (2002) Compact form of DNA induced by DNA-binding protein HU. Biochem Biophys Res Comm **290**: 546-551.

Eppendorf, (2000) TripleMaster PCR system users manual.

Ewing, B. & P. Green, (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8**: 186-194.

Ewing, B., L. Hillier, M. C. Wendl & P. Green, (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8**: 175-185.

Ezraty, B., J. Bos, F. Barras & L. Aussel, (2005) Methionine sulfoxide reduction and assimilation in *Escherichia coli*: new role for the biotin sulfoxide reductase BisC. J Bacteriol **187**: 231-237.

Fahey, R. C., W. C. Brown, W. B. Adams & M. B. Worsham, (1978) Occurrence of glutathione in bacteria. J Bacteriol **133**: 1126-1129.

Faria, J. C. & P. Grosjean, (2005) Tinn-R. In. The R Foundation for Statistical Computing. Vienna, Austria. pp. 12

Fedorov, D. V., D. A. Podkopaeva, M. L. Miroshnichenko, E. A. Bonch-Osmolovskaia, A. V. Lebedinskii & M. Grabovich, (2006) Investigation of the catabolism of acetate and peptides in the new anaerobic thermophilic bacterium *Caldithrix abyssi*. Mikrobiologiia **75**: 154-159.

Ferrer, C., F. J. M. Mojica, G. Juez & F. RodriguezValera, (1996) Differentially transcribed regions of *Haloferax volcanii* genome depending on the medium salinity. J Bacteriol **178**: 309-313.
Firth, N., K. P. Ridgway, M. E. Byrne, P. D. Fink, L. Johnson, I. T. Paulsen & R. A. Skurray, (1993) Analysis of a transfer region from the *Staphylococcal* conjugative plasmid pSK41. Gene **136**: 13-25.

Fischer, M., I. Haase, K. Kis, W. Meining, R. Ladenstein, M. Cushman, N. Schramek, R. Huber & A. Bacher, (2003) Enzyme catalysis via control of activation entropy: site-directed mutagenesis of 6,7-dimethyl-8-ribityllumazine synthase. J Mol Biol **326**: 783-793.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty & J. M. Merrick, (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**: 496-512.

Flemming, A., (1922) On a remarkable bacteriolytic element found in tissues and secretions. Proc Roy Soc Ser B **93**: 306-317.

Flint, H. J. & A. M. Thomson, (1990) Deoxyribonuclease activity in rumen bacteria. Lett Appl Microbiol **11**: 18-21.

Focareta, T. & P. A. Manning, (1991) Distinguishing between the extracellular DNases of *Vibrio cholerae* and development of a transformation system. Mol Microbiol **5**: 2547-2555.

Fondevila, M. & B. A. Dehority, (1996) Interactions between *Fibrobacter succinogenes*, *Prevotella ruminicola*, and *Ruminococcus flavefaciens* in the digestion of cellulose from forages. J Anim Sci **74**: 678-684.

Forns, X., J. Bukh, R. H. Purcell & S. U. Emerson, (1997) How *Escherichia coli* can bias the results of molecular cloning: preferential selection of defective genomes of hepatitis C virus during the cloning procedure. Proc Natl Acad Sci USA **94**: 13909-13914.

Forsdyke, D. R. & J. R. Mortimer, (2000) Chargaffs legacy. Gene **261**: 127-137.

Foster, J. W., D. M. Kinney & A. G. Moat, (1979) Pyridine nucleotide cycle of *Salmonella typhimurium*: isolation and characterization of *pncA*, *pncB*, and *pncC* mutants and utilization of exogenous nicotinamide adenine dinucleotide. J Bacteriol **137**: 1165-1175.

Francia, M. V., A. Varsaki, M. P. Garcillan-Barcia, A. Latorre, C. Drainas & F. de la Cruz, (2004) A classification scheme for mobilization regions of bacterial plasmids. FEMS Microbiol Rev **28**: 79-100.

Frangeul, L., K. E. Nelson, C. Buchrieser, A. Danchin, P. Glaser & F. Kunst, (1999) Cloning and assembly strategies in microbial genome projects. Microbiol **145**: 2625-2634.

Fraser, C. M., J. A. Eisen, K. E. Nelson, I. T. Paulsen & S. L. Salzberg, (2002) The value of complete microbial genome sequencing (you get what you pay for). J Bacteriol **184**: 6403-6405.
Frazer, A., (2005) Meat and Wool New Zealand Scholars Meeting. Wellington: Personal communication.

Frere, J., M. Novel & G. Novel, (1993) Molecular analysis of the *Lactococcus lactis* subspecies *lactis* CNRZ270 bidirectional theta replicating lactose plasmid pUCL22. Mol Microbiol **10**: 1113-1124.

Friedman, S. A. & S. J. Austin, (1988) The P1 plasmid-partition system synthesizes two essential proteins from an autoregulated operon. Plasmid **19**: 103-112.

Fuhrmann, S., M. Ferner, T. Jeffke, A. Henne, G. Gottschalk & O. Meyer, (2003) Complete nucleotide sequence of the circular megaplasmid pHCG3 of *Oligotropha carboxidovorans*: function in the chemolithoautotrophic utilization of CO, H_2 and CO_2. Gene **322**: 67-75.

Fukuchi, S. & K. Nishikawa, (2004) Estimation of the number of authentic orphan genes in bacterial genomes. DNA Res **11**: 219-231, 311-313.

Funnell, B. E. & L. Gagnier, (1993) The P1 plasmid partition complex at *parS*. II. Analysis of ParB protein binding activity and specificity. J Biol Chem **268**: 3616-3624.

Furste, J. P., W. Pansegrau, G. Ziegelin, M. Kroger & E. Lanka, (1989) Conjugative transfer of promiscuous IncP plasmids: interaction of plasmid-encoded products with the transfer origin. Proc Natl Acad Sci USA **86**: 1771-1775.

Gamon, M. R., E. C. Moreira, S. S. de Oliveira, L. M. Teixeira & C. Bastos Mdo, (1999) Characterization of a novel bacteriocin-encoding plasmid found in clinical isolates of *Staphylococcus aureus*. Ant Van Leeu **75**: 233-243.

Garzon, A., D. A. Cano & J. Casadesus, (1995) Role of Erf recombinase in P22-mediated plasmid transduction. Genetics **140**: 427-434.

Gavin, J. J. & W. W. Umbreit, (1965) Effect of biotin on fatty acid distribution in *Escherichia coli*. J Bacteriol **89**: 437-443.

Gavin, R., A. A. Rabaan, S. Merino, J. M. Tomas, I. Gryllos & J. G. Shaw, (2002) Lateral flagella of *Aeromonas* species are essential for epithelial cell adherence and biofilm formation. Mol Microbiol **43**: 383-397.

Gentleman, R., V. Carey, W. Huber, R. Irazarry & S. Dudoit, (2005) Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York.

Gerdes, K., J. Moller-Jensen & R. Bugge Jensen, (2000) Plasmid and chromosome partitioning: surprises from phylogeny. Mol Microbiol **37**: 455-466.

Geremia, R. A., E. A. Petroni, L. Ielpi & B. Henrissat, (1996) Towards a classification of glycosyltransferases based on amino acid sequence similarities: prokaryotic alpha-mannosyltransferases. Biochem J **318**: 133-138.
Gianotti, A., D. Serrazanetti, S. Sado Kamdem & M. E. Guerzoni, (2008) Involvement of cell fatty acid composition and lipid metabolism in adhesion mechanism of *Listeria monocytogenes*. Int J Food Microbiol **123**: 9-17.

Gil, R., F. J. Silva, J. Pereto & A. Moya, (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev **68**: 518-537.

Gilmour, M. W., J. E. Gunton, T. D. Lawley & D. E. Taylor, (2003) Interaction between the IncHI1 plasmid R27 coupling protein and type IV secretion system: TraG associates with the coiled-coil mating pair formation protein TrhB. Mol Microbiol. **49**: 105-116.

Gioia, J., X. Qin, H. Jiang, K. Clinkenbeard, R. Lo, Y. Liu, G. E. Fox, S. Yerrapragada, M. P. McLeod, T. Z. McNeill, L. Hemphill, E. Sodergren, Q. Wang, D. M. Muzny, F. J. Homsi, G. M. Weinstock & S. K. Highlander, (2006) The genome sequence of *Mannheimia haemolytica* A1: insights into virulence, natural competence, and *Pasteurellaceae* phylogeny. J Bacteriol **188**: 7257-7266.

Gjermansen, M., P. Ragas & T. Tolker-Nielsen, (2006) Proteins with GGDEF and EAL domains regulate *Pseudomonas putida* biofilm formation and dispersal. FEMS Microbiol Lett **265**: 215-224.

Goering, R. V. & M. A. Winters, (1992) Rapid method for epidemiologic evaluation of Gram-positive cocci by field inversion gel-electrophoresis. J Clin Microbiol. **30**: 577-580.

Goffin, C. & J. M. Ghuysen, (1998) Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. Microbiol Mol Biol Rev **62**: 1079-1093.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier & J. C. Venter, (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA **103**: 11240-11245.

Golovin, A., T. J. Oldfield, J. G. Tate, S. Velankar, G. J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, J. M. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, A. Pajon, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, G. J. Swaminathan, M. Tagari, S. Tromm, W. Vranken & K. Henrick, (2004) E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res. **32**: D211-216.

Gough, J., K. Karplus, R. Hughey & C. Chothia, (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. **313**: 903-919.

Gouy, M. & C. Gautier, (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res **10**: 7055-7074.

Gram, H. C., (1884) Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. Fortschritte der Medizin **2**: 185-189.

Grantham, R., C. Gautier & M. Gouy, (1980a) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res **8**: 1893-1912.

Grantham, R., C. Gautier, M. Gouy, R. Mercier & A. Pave, (1980b) Codon catalog usage and the genome hypothesis. Nucleic Acids Res **8**: r49-r62.

Grindley, N. D., G. O. Humphreys & E. S. Anderson, (1973) Molecular studies of R factor compatibility groups. J Bacteriol **115**: 387-398.

Grohmann, E., M. Moscoso, E. L. Zechner, G. del Solar & M. Espinosa, (1998) *In-vivo* definition of the functional origin of leading strand replication on the *Lactococcal* plasmid pFX2. Mol Gen Genet **260**: 38-47.

Grohmann, E., G. Muth & M. Espinosa, (2003) Conjugative plasmid transfer in Gram-positive bacteria. Microbiol Mol Biol Rev **67**: 277-301.

Grose, J. H., U. Bergthorsson, Y. Xu, J. Sterneckert, B. Khodaverdian & J. R. Roth, (2005) Assimilation of nicotinamide mononucleotide requires periplasmic AphA phosphatase in *Salmonella enterica*. J Bacteriol **187**: 4521-4530.

Gudgeon, J., (2007) New Zealand external trade statistics. Available: www.stats.govt.nz/externaltrade.

Gupta, S. K. & T. C. Ghosh, (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene **273**: 63-70.

Habel, R. E., (1975) Ruminant digestive system. In: The anatomy of domestic animals. S. Sisson, J. D. H. Grossman & R. Getty (eds). London: W. B. Saunders Company, pp. 881-897.

Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen & O. White, (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res **29**: 41-43.

Halami, P. M., A. Ramesh & A. Chandrashekar, (2000) Megaplasmid encoding novel sugar utilizing phenotypes, pediocin production and immunity in *Pediococcus acidilactici* C20. Food Microbiol **17**: 475-483.

Hall, S. W. & H. Kuhn, (1986) Purification and properties of guanylate kinase from bovine retinas and rod outer segments. Eur J Biochem **161**: 551-556.

Hanahan, D., J. Jessee & F. R. Bloom, (1991) Plasmid transformation of *Escherichia coli* and other bacteria. Methods Enzymol. **204**: 63-113.

Hartman, R. E., (1970) Carbon dioxide fixation by extracts of *Streptococcus faecalis* var. *liquefaciens*. J Bacteriol **102**: 341-346.

He, G. X., T. Kuroda, T. Mima, Y. Morita, T. Mizushima & T. Tsuchiya, (2004) An H(+)-coupled multidrug efflux pump, PmpM, a member of the MATE family of transporters, from *Pseudomonas aeruginosa*. J Bacteriol **186**: 262-265.

Hebbeln, P., D. A. Rodionov, A. Alfandega & T. Eitinger, (2007) Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. Proc Natl Acad Sci USA **104**: 2909-2914.

Hefford, M. A., Y. Kobayashi, S. E. Allard, R. J. Forster & R. M. Teather, (1997) Sequence analysis and characterization of pOM1, a small cryptic plasmid from *Butyrivibrio fibrisolvens*, and its use in construction of a new family of cloning vectors for butyrivibrios. Appl Environ Microbiol **63**: 1701-1711.

Hefford, M. A., R. M. Teather & R. J. Forster, (1993) The complete nucleotide sequence of a small cryptic plasmid from a rumen bacterium of the genus butyrivibrio. Plasmid **29**: 63-69.

Heidelberg, J. F., J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, S. R. Gill, K. E. Nelson, T. D. Read, H. Tettelin, D. Richardson, M. D. Ermolaeva, J. Vamathevan, S. Bass, H. Qin, I. Dragoi, P. Sellers, L. McDonald, T. Utterback, R. D. Fleishmann, W. C. Nierman, O. White, S. L. Salzberg, H. O. Smith, R. R. Colwell, J. J. Mekalanos, J. C. Venter & C. M. Fraser, (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature **406**: 477-483.

Henrissat, B., (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J **280**: 309-316.

Henrissat, B. & A. Bairoch, (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J **293**: 781-788.

Hermann, J. C., R. Marti-Arbona, A. A. Fedorov, E. Fedorov, S. C. Almo, B. K. Shoichet & F. M. Raushel, (2007) Structure-based activity prediction for an enzyme of unknown function. Nature **448**: 775-779.

Hespell, R. B. & P. J. O'Bryan-Shah, (1988) Esterase activities in *Butyrivibrio fibrisolvens* strains. Appl Environ Microbiol **54**: 1917-1922.

Hespell, R. B., R. Wolf & R. J. Bothast, (1987) Fermentation of xylans by *Butyrivibrio fibrisolvens* and other ruminal bacteria. Appl Environ Microbiol **53**: 2849-2853.

Heuer, H., R. E. Fox & E. M. Top, (2007) Frequent conjugative transfer accelerates adaptation of a broad-host-range plasmid to an unfavorable *Pseudomonas putida* host. FEMS Microbiol Ecol **59**: 738-748.
Hirochika, H., K. Nakamura & K. Sakaguchi, (1984) A linear DNA plasmid from *Streptomyces rochei* with an inverted terminal repetition of 614 base pairs. Embo J **3**: 761-766.

Hitoshi, U., M. Fujihiko & W. Chieko, (1999) Regulation of DNA replication by iterons: an interaction between the ori2 and incC regions mediated by RepE-bound iterons inhibits DNA replication of mini-F plasmid in *Escherichia coli*. EMBO J **18**: 3856-3867.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker & P. Schuster, (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem **125**: 167-188.

Hogrefe, C., D. Romermann & B. Friedrich, (1984) *Alcaligenes eutrophus* hydrogenase genes (hox). J Bacteriol **158**: 43-48.

Holland, L. M., S. T. O'Donnell, D. A. Ryjenkov, L. Gomelsky, S. R. Slater, P. D. Fey, M. Gomelsky & J. P. O'Gara, (2008) A staphylococcal GGDEF domain protein regulates biofilm formation independently of c-di-GMP. J Bacteriol **15**: 5178-5189

Holmgren, A., (1976) Hydrogen donor system for *Escherichia coli* ribonucleoside-diphosphate reductase dependent upon glutathione. Proc Natl Acad Sci USA **73**: 2275-2279.

Honda, Y., T. Nakamura, K. Tanaka, A. Higashi, H. Sakai, T. Komano & M. Bagdasarian, (1992) DnaG-dependent priming signals can substitute for the two essential DNA initiation signals in *oriV* of the broad host-range plasmid RSF1010. Nucleic Acids Res **20**: 1733-1737.

Horimoto, K., S. Fukuchi & K. Mori, (2001) Comprehensive comparison between locations of orthologous genes on archaeal and bacterial genomes. Bioinformatics (Oxford, England) **17**: 791-802.

Houot, L. & P. I. Watnick, (2008) A novel role for enzyme I of the *Vibrio cholerae* phosphoenolpyruvate phosphotransferase system in regulation of growth in a biofilm. J Bacteriol **190**: 311-320.

Hu, K. H., E. Liu, K. Dean, M. Gingras, W. DeGraff & N. J. Trun, (1996) Overproduction of three genes leads to camphor resistance and chromosome condensation in *Escherichia coli*. Genetics **143**: 1521-1532.

Hungate, R. E., (1966) The rumen and its microbes. Academic Press, New York.

Hungate, R. E., (1988) Introduction: The ruminant and the rumen. In: The Rumen Microbial Ecosystem. P. N. Hobson (ed). London and New York: Elsevier Applied Science, pp. 1-20.

Hunter, W. J., F. C. Baker, I. S. Rosenfeld, J. B. Keyser & S. B. Tove, (1976) Biohydrogenation of unsaturated fatty acids. Hydrogenation by cell-free preparations of *Butyrivibrio fibrisolvens*. J Biol Chem **251**: 2241-2247.
Huson, D. H., (1998a) SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**: 68-73.

Huynen, M. A. & B. Snel, (2000) Gene and context: integrative approaches to genome analysis. Adv Protein Chem **54**: 345-379.

Igloi, G. L. & R. Brandsch, (2003) Sequence of the 165-kilobase catabolic plasmid pAO1 from *Arthrobacter nicotinovorans* and identification of a pAO1-dependent nicotine uptake system. J Bacteriol. **185**: 1976-1986.

Ihaka, R. & R. Gentleman, (1996) R: a language for data analysis and graphics. J Comput Graph Statist **5**: 299-314.

Ilyina, T. V. & E. V. Koonin, (1992) Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res **20**: 3279-3285.

Inagawa, T., J. Kato, H. Niki, K. Karata & T. Ogura, (2001) Defective plasmid partition in *ftsH* mutants of *Escherichia coli*. Mol Genet Genomics **265**: 755-762.

InforMax Inc., (2001) Vector NTI. In. Bethesda: InforMax, Inc.

Innes, R. W., M. A. Hirose & P. L. Kuempel, (1988) Induction of nitrogen-fixing nodules on clover requires only 32 kilobase pairs of DNA from the *Rhizobium trifolii* symbiosis plasmid. J Bacteriol **170**: 3793-3802.

Inoue, T., R. Shingaki & K. Fukui, (2008) Inhibition of swarming motility of *Pseudomonas aeruginosa* by branched-chain fatty acids. FEMS Microbiol Lett **281**: 81-86.

Inselburg, J. & A. Oka, (1975) Discontinuous replication of colicin E1 plasmid deoxyribonucleic acid. J Bacteriol **123**: 739-742.

Ishikawa, T., Y. Hayashida, K. Hirayasu, K. Ozawa, N. Yamamoto, T. Tanaka & S. Matsuura, (2003) Use of transcriptional sequencing in difficult to read areas of the genome. Anal Biochem **316**: 202-207.

Iverson, T. M., C. Luna-Chavez, G. Cecchini & D. C. Rees, (1999) Structure of the *Escherichia coli* fumarate reductase respiratory complex. Science **284**: 1961-1966.

Jacq, C., J. R. Miller & G. G. Brownlee, (1977) A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell **12**: 109-120.

Jenal, U., (2004) Cyclic di-guanosine-monophosphate comes of age: a novel secondary messenger involved in modulating cell surface structures in bacteria? Curr Opin Microbiol **7**: 185-191.

Jerke, K., C. H. Nakatsu, F. Beasley & A. Konopka, (2008) Comparative analysis of eight *Arthrobacter* plasmids. Plasmid **59**: 73-85.
Johnson, M. R., S. B. Conners, C. I. Montero, C. J. Chou, K. R. Shockley & R. M. Kelly, (2006) The *Thermotoga maritima* phenotype is impacted by syntrophic interaction with *Methanococcus jannaschii* in hyperthermophilic coculture. Appl Environ Microbiol **72**: 811-818.

Joyner, A. E., Jr. & R. L. Baldwin, (1966) Enzymatic studies of pure cultures of rumen microorganisms. J Bacteriol **92**: 1321-1330.

Jumas-Bilak, E., S. Michaux-Charachon, G. Bourg, M. Ramuz & A. Allardet-Servent, (1998) Unconventional genomic organization in the alpha subgroup of the proteobacteria. J Bacteriol. **180**: 2749-2755.

Juncker, A. S., H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen & A. Krogh, (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. **12**: 1652-1662.

Kaatz, G. W., F. McAleese & S. M. Seo, (2005) Multidrug resistance in *Staphylococcus aureus* due to overexpression of a novel multidrug and toxin extrusion (MATE) transport protein. Antimicro Agen Chemo **49**: 1857-1864.

Kalkus, J., M. Reh & H. G. Schlegel, (1990) Hydrogen autotrophy of *Nocardia opaca* strains is encoded by linear megaplasmids. J Gen Microbiol **136**: 1145-1151.

Kalmokoff, M. L. & R. M. Teather, (1997) Isolation and characterization of a bacteriocin (Butyrivibriocin AR10) from the ruminal anaerobe *Butyrivibrio fibrisolvens* AR10: evidence in support of the widespread occurrence of bacteriocin-like activity among ruminal isolates of *B. fibrisolvens*. Appl Environ Microbiol **63**: 394-402.

Kanaya, S., Y. Yamada, Y. Kudo & T. Ikemura, (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**: 143-155.

Karlin, S., J. Mrazek & A. Campbell, (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol **179**: 3899-3913.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren & E. S. Lander, (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423**: 241-254.

Kentner, D. & V. Sourjik, (2006) Spatial organization of the bacterial chemotaxis system. Curr Opin Microbiol **9**: 619-624.

Kepler, C. R., K. P. Hirons, J. J. McNeill & S. B. Tove, (1966) Intermediates and products of the biohydrogenation of linoleic acid by *Butyrivibrio fibrisolvens*. J Biol Chem. **241**: 1350-1354.

Khan, S. A., R. W. Murray & R. R. Koepsel, (1988) Mechanism of plasmid pT181 DNA replication. Bioch Biophy Acta **951**: 375-381.
Kihara, A., Y. Akiyama & K. Ito, (1995) FtsH is required for proteolytic elimination of uncomplexed forms of SecY, an essential protein translocase subunit. Proc Natl Acad Sci USA **92**: 4532-4536.

Kihara, A., Y. Akiyama & K. Ito, (1998) Different pathways for protein degradation by the FtsH/HflKC membrane-embedded protease complex: an implication from the interference by a mutant form of a new substrate protein, YccA. J Mol Biol **279**: 175-188.

Kim, U. J., H. Shizuya, P. J. de Jong, B. Birren & M. I. Simon, (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res **20**: 1083-1085.

Kim, Y. K. & L. L. McCarter, (2007) ScrG, a GGDEF-EAL protein, participates in regulating swarming and sticking in *Vibrio parahaemolyticus*. J Bacteriol **189**: 4094-4107.

Kleckner, N., (1981) Transposable elements in prokaryotes. Annu Rev Genet **15**: 341-404.

Klemm, D., B. Heublein, H. P. Fink & A. Bohn, (2005) Cellulose: fascinating biopolymer and sustainable raw material. Angew Chem Int Ed Engl **44**: 3358-3393.

Klieve, A. V. & R. A. Swain, (1993) Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. Appl Environ Microbiol **59**: 2299-2303.

Kneisel, J. J., (1968) A clinical trial of an enzyme inhibitor in pancreatitis. Arch Surg **96**: 442-449.

Knobloch, J. K., M. Nedelmann, K. Kiel, K. Bartscht, M. A. Horstkotte, S. Dobinsky, H. Rohde & D. Mack, (2003) Establishment of an arbitrary PCR for rapid identification of Tn917 insertion sites in *Staphylococcus epidermidis*: characterization of biofilm-negative and nonmucoid mutants. Appl Environ Microbiol **69**: 5812-5818.

Kobayashi, Y., R. J. Forster, M. A. Hefford, R. M. Teather, M. Wakita, K. Ohmiya & S. Hoshino, (1995) Analysis of the sequence of a new cryptic plasmid, pRJF2, from a rumen bacterium of the genus *Butyrivibrio* - comparison with other *Butyrivibrio* plasmids and application in the development of a cloning vector. Fems Microbiology Letters **130**: 137-143.

Kobayashi, Y., N. Okuda, M. Matsumoto, K. Inoue, M. Wakita & S. Hoshino, (1998) Constitutive expression of a heterologous *Eubacterium ruminantium* xylanase gene (xynA) in *Butyrivibrio fibrisolvens*. FEMS Microbiol Lett **163**: 11-17.

Koepsel, R. R., R. W. Murray, W. D. Rosenblum & S. A. Khan, (1985) The replication initiator protein of plasmid pT181 has sequence-specific endonuclease and topoisomerase-like activities. Proc Natl Acad Sci USA **82**: 6845-6849.

Kohli, J. & H. Grosjean, (1981) Usage of the three termination codons: compilation and analysis of the known eukaryotic and prokaryotic translation termination sequences. Mol Gen Genet **182**: 430-439.

Kojic, M., I. Strahinic & L. Topisirovic, (2005) Proteinase PI and lactococcin A genes are located on the largest plasmid in *Lactococcus lactis* subsp. *lactis* bv. *diacetylactis* S50. Can J Microbiol **51**: 305-314.

Kolsto, A. B., (1997) Dynamic bacterial genome organization. Mol Microbiol **24**: 241-248.

Komatsu, H., Y. Imura, A. Ohori, Y. Nagata & M. Tsuda, (2003) Distribution and organization of auxotrophic genes on the multichromosomal genome of *Burkholderia multivorans* ATCC 17616. J Bacteriol **185**: 3333-3343.

Konforti, B., (2007) Name that gene! Nat Struct Molec Biol **14**: 681.

Kong, Z., (2007) Microarray analysis of *Clostridium proteoclasticum* genes involved in hemicellulose degradation. In: School of biological sciences. Auckland: University of Auckland.

Koonin, E. V., (2000) How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet **1**: 99-116.

Koonin, E. V., (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol **1**: 127-136.

Kopecny, J., M. Zorec, J. Mrazek, Y. Kobayashi & R. Marinsek-Logar, (2003) *Butyrivibrio hungatei* sp nov and *Pseudobutyrivibrio xylanivorans* sp nov., butyrate-producing bacteria from the rumen. Int J Syst Evol Microbiol. **53**: 201-209.

Kornberg, A., J. F. Scott & L. L. Bertsch, (1978) ATP utilization by rep protein in the catalytic separation of DNA strands at a replicating fork. J Biol Chem **253**: 3298-3304.

Kosenko, L. V. & N. N. Mal'tseva, (1984) Exopolysaccharide of *Mycobacterium flavum*. Mikrobiologiia **53**: 547-550.

Kozak, M., (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev **47**: 1-45.

Krasowiak, R., Y. Sevastsyanovich, I. Konieczny, L. E. Bingle & C. M. Thomas, (2006) IncP-9 replication initiator protein binds to multiple DNA sequences in *oriV* and recruits host DnaA protein. Plasmid **56**: 187-201.

Krause, D. O., R. J. Bunch, B. D. Dalrymple, K. S. Gobius, W. J. Smith, G. P. Xue & C. S. McSweeney, (2001) Expression of a modified *Neocallimastix patriciarum* xylanase in *Butyrivibrio fibrisolvens* digests more fibre but cannot effectively compete with highly fibrolytic bacteria in the rumen. J Appl Microbiol **90**: 388-396.

Krause, D. O., S. E. Denman, R. I. Mackie, M. Morrison, A. L. Rae, G. T. Attwood & C. S. McSweeney, (2003) Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics. FEMS Microbiol Rev. 27: 663-693.

Krawiec, S. R., M., (1990) Organization of the bacterial chromosome. Microbiol Rev. 54: 502-539.

Krell, K. & M. A. Eisenberg, (1970) The purification and properties of dethiobiotin synthetase. J Biol Chem **245**: 6558-6566.

Krogh, A., B. Larsson, G. von Heijne & E. L. Sonnhammer, (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol **305**: 567-580.

Krum, J. G. & S. A. Ensign, (2001) Evidence that a linear megaplasmid encodes enzymes of aliphatic alkene and epoxide metabolism and coenzyme M (2-mercaptoethanesulfonate) biosynthesis in *Xanthobacter* strain Py2. J Bacteriol **183**: 2172-2177.

Kumaran, D., S. Eswaramoorthy, F. W. Studier & S. Swaminathan, (2005) Structure and mechanism of ADP-ribose-1"-monophosphatase (Appr-1"-pase), a ubiquitous cellular processing enzyme. Protein Sci **14**: 719-726.

Kundig, C., C. Beck, H. Hennecke & M. Gottfert, (1995) A single rRNA gene region in *Bradyrhizobium japonicum*. J Bacteriol **177**: 5151-5154.

Kurenbach, B., C. Bohn, J. Prabhu, M. Abudukerim, U. Szewzyk & E. Grohmann, (2003) Intergeneric transfer of the *Enterococcus faecalis* plasmid pIP501 to *Escherichia coli* and *Streptomyces lividans* and sequence analysis of its *tra* region. Plasmid **50**: 86-93.

Kurnasov, O. V., B. M. Polanuyer, S. Ananta, R. Sloutsky, A. Tam, S. Y. Gerdes & A. L. Osterman, (2002) Ribosylnicotinamide kinase domain of NadR protein: identification and implications in NAD biosynthesis. J Bacteriol **184**: 6906-6917.

Kyrpides, N., (1999) Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide. Bioinformatics **15**: 773-774.

Lambrecht, F. L., (1966) Notes on the growth curve of *Trypanosoma cruzi chagas* 1909 as determined by optical density. Rev Inst Med Trop Sao Paulo **8**: 249-254.

Lande, W. M., P. V. Thiemann & W. C. Mentzer, Jr., (1982) Missing band 7 membrane protein in two patients with high Na, low K erythrocytes. J Clin Invest **70**: 1273-1280.

Lanka, E. & B. M. Wilkins, (1995) DNA processing reactions in bacterial conjugation. Annu Rev Biochem **64**: 141-169.

Laukova, A., M. Kuncova & V. Kmet, (1990) Isolation of several conjugative plasmids of the rumen bacteria *Enterococcus faecium*. Biologia **45**: 533-538.

Le Chatelier, E., S. D. Ehrlich & L. Janniere, (1993) Biochemical and genetic analysis of the unidirectional theta replication of the *S. agalactiae* plasmid pIP501. Plasmid **29**: 50-56.

Lechner, R. L., M. J. Engler & C. C. Richardson, (1983) Characterization of strand displacement synthesis catalyzed by bacteriophage T7 DNA polymerase. J Biol Chem **258**: 11174-11184.

Lechner, R. L. & C. C. Richardson, (1983) A preformed, topologically stable replication fork. Characterization of leading strand DNA synthesis catalyzed by T7 DNA polymerase and T7 gene 4 protein. J Biol Chem **258**: 11185-11196.

Lee, C., J. Kim, S. G. Shin & S. Hwang, (2006) Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. J Biotech **123**: 273-280.

Lee, E. H., C. Rouquette-Loughlin, J. P. Folster & W. M. Shafer, (2003) FarR regulates the *farAB*-encoded efflux pump of *Neisseria gonorrhoeae* via an MtrR regulatory mechanism. J Bacteriol **185**: 7145-7152.

Lessl, M., D. Balzer, K. Weyrauch & E. Lanka, (1993) The mating pair formation system of plasmid RP4 defined by RSF1010 mobilization and donor-specific phage propagation. J Bacteriol **175**: 6415-6425.

Li, Y., C. Canchaya, F. Fang, E. Raftis, K. A. Ryan, J. P. van Pijkeren, D. van Sinderen & P. W. O'Toole, (2007) Distribution of megaplasmids in *Lactobacillus salivarius* and other lactobacilli. J Bacteriol **189**: 6128-6139.

Liaqat, I. & A. N. Sabri, (2008) Analysis of cell wall constituents of biocide-resistant isolates from dental-unit water line biofilms. Curr Microbiol **57**: 340-347.

Lin, L. L. & J. A. Thomson, (1991) An analysis of the extracellular xylanases and cellulases of *Butyrivibrio fibrisolvens* H17c. FEMS Microbiol Lett **68**: 197-203.

Lindahl, T. & B. Nyberg, (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry **13**: 3405-3410.

Liolios, K., N. Tavernarakis, P. Hugenholtz & N. Kyrpides, C., (2006) The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. NAR **34**: D332-334

Liu, S. L., A. Hessel & K. E. Sanderson, (1993) Genomic mapping with *I-Ceu I*, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella spp*., *Escherichia coli*, and other bacteria. Proc Natl Acad Sci USA **90**: 6874-6878.

Liu, S. L. & K. E. Sanderson, (1996) Highly plastic chromosomal organization in *Salmonella typhi*. Proc Natl Acad Sci USA **93**: 10303-10308.

Liu, Y. & Y. L. Fan, (1991) Location of 43 kD mosquito-larvicidal toxin gene of highly toxic *Bacillus sphaericus* 10. Sci China B **34**: 593-598.

Lomovskaya, O., K. Lewis & A. Matin, (1995) EmrR is a negative regulator of the *Escherichia coli* multidrug resistance pump EmrAB. J Bacteriol **177**: 2328-2334.

Long, F., C. Rouquette-Loughlin, W. M. Shafer & E. W. Yu, (2008) Functional cloning and characterization of the multidrug efflux pumps NorM from *Neisseria gonorrhoeae* and YdhE from *Escherichia coli*. Antimicrob Agen Chemo **52**: 3052-3060

Lopezgarcia, P., J. Anton, J. P. Abad & R. Amils, (1994) Halobacterial megaplasmids are negatively supercoiled. Mol Microbiol **11**: 421-427.

Lowe, T. M. & S. R. Eddy, (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25**: 955-964.

Lu, H. M., S. T. Motley & S. Lory, (1997) Interactions of the components of the general secretion pathway: role of *Pseudomonas aeruginosa* type IV pilin subunits in complex formation and extracellular protein secretion. Mol Microbiol **25**: 247-259.

Lukashin, A. V. & M. Borodovsky, (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26**: 1107-1115.

Lux, R., K. Jahreis, K. Bettenbrock, J. S. Parkinson & J. W. Lengeler, (1995) Coupling the phosphotransferase system and the methyl-accepting chemotaxis protein-dependent chemotaxis signaling pathways of *Escherichia coli*. Proc Natl Acad Sci USA **92**: 11583-11587.

Lyras, D. & J. I. Rood, (1998) Conjugative transfer of RP4-oriT shuttle vectors from *Escherichia coli* to *Clostridium perfringens*. Plasmid **39**: 160-164.

Ma, K. & M. W. Adams, (1999) An unusual oxygen-sensitive, iron- and zinc-containing alcohol dehydrogenase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol **181**: 1163-1170.

MacAllister, T. W., W. L. Kelley, A. Miron, T. T. Stenzel & D. Bastia, (1991) Replication of plasmid R6K origin gamma *in-vitro*. Dependence on dual initiator proteins and inhibition by transcription. J Biol Chem **266**: 16056-16062.

Mackenzie, K. F., C. K. Eddy & L. O. Ingram, (1989) Modulation of alcohol dehydrogenase isoenzyme levels in *Zymomonas mobilis* by iron and zinc. J Bacteriol **171**: 1063-1067.

Mackie, R. I. & B. A. White, (1990) Recent advances in rumen microbial ecology and metabolism: potential impact on nutrient output. J Dairy Sci **73**: 2971-2995.

MacLellan, S. R. & C. W. Forsberg, (2001) Properties of the major non-specific endonuclease from the strict anaerobe *Fibrobacter succinogenes* and evidence for disulfide bond formation *in-vivo*. Microbiol **147**: 315-323.

Magalhaes, M. L., A. Argyrou, S. M. Cahill & J. S. Blanchard, (2008) Kinetic and mechanistic analysis of the *Escherichia coli* ribD-encoded bi-functional deaminase-reductase involved in riboflavin biosynthesis. Biochem **47**: 6499-6507.

Maier, T. M., J. M. Myers & C. R. Myers, (2003) Identification of the gene encoding the sole physiological fumarate reductase in *Shewanella oneidensis* MR-1. J Bas Microbiol 43: 312-327.

Makarova, K. S., N. V. Grishin & E. V. Koonin, (2006) The HicAB cassette, a putative novel, RNA-targeting toxin-antitoxin system in archaea and bacteria. Bioinformatics **22**: 2581-2584.

Mann, S. P., G. P. Hazlewood & C. G. Orpin, (1986) Characterization of a cryptic plasmid (p0M1) in *Butyrivibrio fibrisolvens* by restriction endonuclease analysis and its cloning in *Escherichia coli*. Curr Microbiol **13**: 17.

Maresso, A. W., T. J. Chapa & O. Schneewind, (2006) Surface protein IsdC and Sortase B are required for heme-iron scavenging of *Bacillus anthracis*. J Bacteriol **188**: 8145-8152.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley & J. M. Rothberg, (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**: 376-380.

Marsin, S., E. Marguet & P. Forterre, (2000) Topoisomerase activity of the hyperthermophilic replication initiator protein Rep75. Nucleic Acids Res **28**: 2251-2255.

Martin, S. A. & R. G. Dean, (1989) Characterization of a plasmid from the ruminal bacterium *Selenomonas ruminantium*. Appl Environ Microbiol **55**: 3035-3038.

Martinez-Vaz, B. M., Y. Xie, W. Pan & A. B. Khodursky, (2005) Genome-wide localization of mobile elements: experimental, statistical and biological considerations. BMC Genomics **6**: 81.

Mathew, M. K., C. L. Smith & C. R. Cantor, (1988) High-resolution seperation and accurate size determination in pulsed-field gel-electrophoresis of DNA. 1. DNA size standards and the effect of agarose and temperature. Biochem **27**: 9204-9210.

Mattarelli, P., B. Biavati, A. Alessandrini, F. Crociani & V. Scardovi, (1994) Characterization of the plasmid pVS809 from *Bifidobacterium globosum*. New Microbiol **17**: 327-331.

Matulova, M., R. Nouaille, P. Capek, M. Pean, A. M. Delort & E. Forano, (2008) NMR study of cellulose and wheat straw degradation by *Ruminococcus albus* 20. FEBS J **275**: 3503-3511.

McCarthy, C., (2003) Chromas. In. School of Health Sciences, Griffith University Southport, Australia.

McDowell, D. G., N. A. Burns & H. C. Parkes, (1998) Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR. Nucleic Acids Res **26**: 3340-3347.

McInerney, P. & M. O'Donnell, (2004) Functional uncoupling of twin polymerases: mechanism of polymerase dissociation from a lagging-strand block. J Biol Chem **279**: 21543-21551.

McLeod, M. P., R. L. Warren, W. W. Hsiao, N. Araki, M. Myhre, C. Fernandes, D. Miyazawa, W. Wong, A. L. Lillquist, D. Wang, M. Dosanjh, H. Hara, A. Petrescu, R. D. Morin, G. Yang, J. M. Stott, J. E. Schein, H. Shin, D. Smailus, A. S. Siddiqui, M. A. Marra, S. J. Jones, R. Holt, F. S. Brinkman, K. Miyauchi, M. Fukuda, J. E. Davies, W. W. Mohn & L. D. Eltis, (2006) The complete genome of *Rhodococcus sp.* RHA1 provides insights into a catabolic powerhouse. Proc Natl Acad Sci USA **103**: 15582-15587.

McMurray, A. A., J. E. Sulston & M. A. Quail, (1998) Short-insert libraries as a method of problem solving in genome sequencing. Genome Res **8**: 562-566.

Meacock, P. A. & S. N. Cohen, (1980) Partitioning of bacterial plasmids during cell division: a cis-acting locus that accomplishes stable plasmid inheritance. Cell. **20**: 529-542.

Mercer, D. K., S. Patel & H. J. Flint, (2001) Sequence analysis of the plasmid pRR12 from the rumen bacterium *Prevotella ruminicola* 223/M2/7 and the use of pRR12 in *Prevotella*/*Bacteroides* shuttle vectors. Plasmid **45**: 227-232.

Mesas, J. M., M. C. Rodriguez & M. T. Alegre, (2004) Plasmid curing of *Oenococcus oeni*. Plasmid **51**: 37-40.

Miller, T. L. & S. E. Jenesel, (1979) Enzymology of butyrate formation by *Butyrivibrio fibrisolvens*. J Bacteriol **138**: 99-104.

Mira, A. & H. Ochman, (2002) Gene location and bacterial sequence divergence. Mol Biol Evol **19**: 1350-1358.

Miron, J., S. H. Duncan & C. S. Stewart, (1994) Interactions between rumen bacterial strains during the degradation and utilization of the monosaccharides of barley straw cell-walls. J Appl Bacteriol **76**: 282-287.

Mitchell, J. A., A. T. McCray & O. Bodenreider, (2003) From phenotype to genotype: issues in navigating the available information resources. Methods Inf Med **42**: 557-563.

Mizan, S., M. D. Lee, B. G. Harmon, S. Tkalcic & J. J. Maurer, (2002) Acquisition of antibiotic resistance plasmids by enterohemorrhagic *Escherichia coli* O157:H7 within rumen fluid. J Food Prot **65**: 1038-1040.

Mochizuki, S., K. Hiratsu, M. Suwa, T. Ishii, F. Sugino, K. Yamada & H. Kinashi, (2003) The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. Mol Microbiol **48**: 1501-1510.

Mohl, D. A., J. Easter, Jr. & J. W. Gober, (2001) The chromosome partitioning protein, ParB, is required for cytokinesis in *Caulobacter crescentus*. Mol Microbiol **42**: 741-755.

Mongodin, E. F., N. Shapir, S. C. Daugherty, R. T. DeBoy, J. B. Emerson, A. Shvartzbeyn, D. Radune, J. Vamathevan, F. Riggs, V. Grinberg, H. Khouri, L. P. Wackett, K. E. Nelson & M. J. Sadowsky, (2006) Secrets of soil survival revealed by the genome sequence of *Arthrobacter aurescens* TC1. PLoS Gen **2**: e214.

Moon, C. D., D. M. Pacheco, W. J. Kelly, S. Leahy, D. Li, J. Kopečný & G. T. Attwood, (2008) Reclassification of *Clostridium proteoclasticum* as *Butyrivibrio proteoclasticus* comb. nov., a butyrate-producing ruminal bacterium. Int J Sys Evol Microbiol **58**: 2041-2045.

Moorthy, S. & P. I. Watnick, (2005) Identification of novel stage-specific genetic requirements through whole genome transcription profiling of *Vibrio cholerae* biofilm development. Mol Microbiol **57**: 1623-1635.

Moreno, E., (1998) Genome evolution within the alpha Proteobacteria: why do some bacteria not possess plasmids and others exhibit more than one different chromosome? Fems Microbiol Rev **22**: 255-275.

Morgenstern, B., (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res **32**: W33-36.

Morita, Y., K. Kodama, S. Shiota, T. Mine, A. Kataoka, T. Mizushima & T. Tsuchiya, (1998) NorM, a putative multidrug efflux protein, of *Vibrio parahaemolyticus* and its homolog in *Escherichia coli*. Antimic Agen Chemo **42**: 1778-1782.

Morris, E. J. & N. P. Van Gylswyk, (1980) Comparison of the action of rumen bacteria on cell walls of Eragrostis tef. J Agric Sci **95**: 313-323.

Morrison, N. A., C. Y. Hau, M. J. Trinick, J. Shine & B. G. Rolfe, (1983) Heat curing of a sym plasmid in a fast-growing *Rhizobium sp*. that is able to nodulate legumes and the nonlegume *Parasponia sp*. J Bacteriol **153**: 527-531.

Morton, R. K., (1950) Separation and purification of enzymes associated with insoluble particles. Nature **166**: 1092-1095.

Moscoso, M., G. del Solar & M. Espinosa, (1995) Specific nicking-closing activity of the initiator of replication protein RepB of plasmid pMV158 on supercoiled or single-stranded DNA. J Biol Chem **270**: 3772-3779.

Moscoso, M., R. Eritja & M. Espinosa, (1997) Initiation of replication of plasmid pMV158: mechanisms of DNA strand-transfer reactions mediated by the initiator RepB protein. J Mol Biol **268**: 840-856.

Mrazek, J., M. Piknova, P. Pristas & J. Kopecny, (2005) Occurrence of restriction-modification systems in ruminal butyrate-producing bacteria. Anaerobe **11**: 280-284.

Muesing, M., J. Tamm, H. M. Shepard & B. Polisky, (1981) A single base-pair alteration is responsible for the DNA overproduction phenotype of a plasmid copy-number mutant. Cell **24**: 235-242.

Muller, E., K. Fahlbusch, R. Walther & G. Gottschalk, (1981) Formation of N,N-dimethylglycine, acetic acid, and butyric acid from betaine by *Eubacterium limosum*. Appl Environ Microbiol **42**: 439-445.

Mural, R. J., (2000) ARTEMIS: a tool for displaying and annotating DNA sequence. Brief Bioinform **1**: 199-200.

Murray, N. E., (2000) Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev **64**: 412-434.

Murzin, A. G., S. E. Brenner, T. Hubbard & C. Chothia, (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247**: 536-540.

Nakabachi, A., A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran & M. Hattori, (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science **314**: 267.

Nallapareddy, S. R., K. V. Singh & B. E. Murray, (2006) Construction of improved temperature-sensitive and mobilizable vectors and their use for constructing mutations in the adhesin-encoding acm gene of poorly transformable clinical *Enterococcus faecium* strains. Appl Environ Microbiol **72**: 334-345.

Napolitano, R., R. Janel-Bintz, J. Wagner & R. P. Fuchs, (2000) All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. Embo J **19**: 6259-6265.

Nasu, H., T. Iida, T. Sugahara, Y. Yamaichi, K. S. Park, K. Yokoyama, K. Makino, H. Shinagawa & T. Honda, (2000) A filamentous phage associated with recent pandemic *Vibrio parahaemolyticus* O3:K6 strains. J Clin Microbiol **38**: 2156-2161.

Navarre, W. W. & O. Schneewind, (1994) Proteolytic cleavage and cell wall anchoring at the LPXTG motif of surface proteins in Gram-positive bacteria. Mol Microbiol **14**: 115-121.

Naylor, G. E., (1998) The isolation and characterisation of ruminal mycoplasmas and their interactions with ruminal cellulolytic microorganisms. In: Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand. pp. 101.

NCBI, (2008) Entrez: Genomes bacterial plasmid taxonomy. Available: http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=2&type=2&name=Bacteria%20Plasmids

Neimark, H. & B. C. Kirkpatrick, (1993) Isolation and characterization of full-length chromosomes from non-culturable plant-pathogenic *Mycoplasma*-like organisms. Mol Microbiol **7**: 21-28.

Nesbo, C. L., M. Dlutek & W. F. Doolittle, (2006) Recombination in *Thermotoga*: Implications for species concepts and biogeography. Genetics **172**: 759-769.

Ng, W. V., S. A. Ciufo, T. M. Smith, R. E. Bumgarner, D. Baskin, J. Faust, B. Hall, C. Loretz, J. Seto, J. Slagel, L. Hood & S. DasSarma, (1998) Snapshot of a large dynamic replicon in a halophilic archaeon: Megaplasmid or minichromosome? Gen Res **8**: 1131-1141.

Niazi, J. H., D. T. Prasad & T. B. Karegoudar, (2001) Initial degradation of dimethylphthalate by esterases from *Bacillus* species. FEMS Microbiol Lett **196**: 201-205.

Nierop Groot, M., F. Nieboer & T. Abee, (2008) Enhanced transformation efficiency of recalcitrant *Bacillus cereus* and *Bacillus weihenstephanensis* isolates upon *in-vitro* methylation of plasmid DNA. Appl Environ Microbiol.

Nikolaev Iu, A., N. S. Panikov, S. M. Lukin & G. A. Osipov, (2001) Saturated C21-C33 hydrocarbons are involved in the self-regulation of *Pseudomonas fluorescens* adhesion to a glass surface. Mikrobiologiia **70**: 174-181.

Noirot-Gros, M. F., V. Bidnenko & S. D. Ehrlich, (1994) Active site of the replication protein of the rolling circle plasmid pC194. Embo J **13**: 4412-4420.

Nolling, J., G. Breton, M. Omelchenko, K. S. Makarova, Q. D. Zeng, R. Gibson, H. M. Lee, J. Dubois, D. Y. Qiu, J. Hitti, Y. I. Wolf, R. L. Tatusov, F. Sabathe, L. Doucette-Stamm, P. Soucaille, M. J. Daly, G. N. Bennett, E. V. Koonin & D. R. Smith, (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol **183**: 4823-4838.

Novick, R. P., (1987) Plasmid incompatibility. Microbiol Rev **51**: 381-395.

Novick, R. P. & F. C. Hoppensteadt, (1978) On plasmid incompatibility. Plasmid **1**: 421-434.

O'Donovan, G. A. & J. Neuhard, (1970) Pyrimidine metabolism in microorganisms. Bacteriol Rev **34**: 278-343.

O'Toole, G. A. & R. Kolter, (1998) Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. Mol Microbiol **30**: 295-304.

Ochman, H., (2002) Bacterial evolution: chromosome arithmetic and geometry. Curr Biol **12**: R427-428.

Ochman, H. & L. M. Davalos, (2006) The nature and dynamics of bacterial genomes. Science **311**: 1730-1733.

Ochman, H., A. S. Gerber & D. L. Hartl, (1988) Genetic applications of an inverse polymerase chain reaction. Genetics **120**: 621-623.

Oeschger, M. P. & M. J. Bessman, (1966) Purification and properties of guanylate kinase from *Escherichia coli*. J Biol Chem **241**: 5452-5460.

Ogata, K., T. Sekizaki, R. I. Aminov, K. Tajima, M. Nakamura, T. Nagamine, H. Matsui & Y. Benno, (1999) A small cryptic plasmid from *Ruminobacter amylophilus* NIAH-3 possesses functional mobilization properties. Fems Microbiol Let **181**: 41-48.

Ogura, T. & S. Hiraga, (1983) Partition mechanism of F plasmid: two plasmid gene-encoded products and a cis-acting region are involved in partition. Cell **32**: 351-360.

Ogura, T., K. Inoue, T. Tatsuta, T. Suzaki, K. Karata, K. Young, L. H. Su, C. A. Fierke, J. E. Jackman, C. R. Raetz, J. Coleman, T. Tomoyasu & H. Matsuzawa, (1999) Balanced biosynthesis of major membrane components through regulated degradation of the committed enzyme of lipid A biosynthesis by the AAA protease FtsH (HflB) in *Escherichia coli*. Mol Microbiol **31**: 833-844.

Ohsawa, I., D. Speck, T. Kisou, K. Hayakawa, M. Zinsius, R. Gloeckler, Y. Lemoine & K. Kamogawa, (1989) Cloning of the biotin synthetase gene from *Bacillus sphaericus* and expression in *Escherichia coli* and *Bacilli*. Gene **80**: 39-48.

Okinaka, R. T., K. Cloud, O. Hampton, A. R. Hoffmaster, K. K. Hill, P. Keim, T. M. Koehler, G. Lamke, S. Kumano, J. Mahillon, D. Manter, Y. Martinez, D. Ricke, R. Svensson & P. J. Jackson, (1999) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. J Bacteriol **181**: 6509-6515.

Oliver, F. J., J. Menissier-de Murcia & G. de Murcia, (1999) Poly(ADP-ribose) polymerase in the cellular response to DNA damage, apoptosis, and disease. Am J Hum Genet **64**: 1282-1288.

Olson, S. A., (2002) EMBOSS opens up sequence analysis. Brief Bioinform **3**: 87-91.

Osborn, A. M., F. M. da Silva Tatley, L. M. Steyn, R. W. Pickup & J. R. Saunders, (2000) Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. Microbiol **146**: 2267-2275.

Paillard, D., N. McKain, M. T. Rincon, K. J. Shingfield, D. I. Givens & R. J. Wallace, (2007) Quantification of ruminal *Clostridium proteoclasticum* by real-time PCR using a molecular beacon approach. J Appl Microbiol **103**: 1251-1261.

Palacios, S., V. J. Starai & J. C. Escalante-Semerena, (2003) Propionyl coenzyme A is a common intermediate in the 1,2-propanediol and propionate catabolic pathways needed for expression of the prpBCDE operon during growth of *Salmonella enterica* on 1,2-propanediol. J Bacteriol **185**: 2802-2810.

Pansegrau, W., D. Balzer, V. Kruft, R. Lurz & E. Lanka, (1990) *In-vitro* assembly of relaxosomes at the transfer origin of plasmid RP4. Proc Natl Acad Sci USA **87**: 6555-6559.

Pansegrau, W., G. Ziegelin & E. Lanka, (1988) The origin of conjugative IncP plasmid transfer: interaction with plasmid-encoded products and the nucleotide sequence at the relaxation site. Biochim Biophys Acta **951**: 365-374.

Papadopoulos, D., D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski & M. Blot, (1999) Genomic evolution during a 10,000-generation experiment with bacteria. Proc Natl Acad Sci USA **96**: 3807-3812.

Park, J. T., (1995) Why does *Escherichia coli* recycle its cell wall peptides? Mol Microbiol **17**: 421-426.

Park, J. T. & T. Uehara, (2008) How bacteria consume their own exoskeletons (turnover and recycling of cell wall peptidoglycan). Microbiol Mol Biol Rev **72**: 211-227.

Park, K. & D. K. Chattoraj, (2001) DnaA boxes in the P1 plasmid origin: The effect of their position on the directionality of replication and plasmid copy number. J Mol Biol **310**: 69-81.

Park, K., S. Mukhopadhyay & D. K. Chattoraj, (1998) Requirements for and regulation of origin opening of plasmid P1. J Biol Chem **273**: 24906-24911.

Pearl, F. M., C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton & C. A. Orengo, (2003) The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res **31**: 452-455.

Peitsch, M. C., (1996) ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochem Soc Trans **24**: 274-279.

Permina, E. A., A. A. Mironov & M. S. Gelfand, (2002) Damage-repair error-prone polymerases of eubacteria: association with mobile genome elements. Gene **293**: 133-140.

Pfennig, N. & S. Wagener, (1986) An improved method of preparing wet mounts for photomicrographs of microorganisms. J Microbiological Met 4: 303-306.

Pieper, U., N. Eswar, F. P. Davis, H. Braberg, M. S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B. M. Webb, D. Eramian, M. Y. Shen, L. Kelly, F. Melo & A. Sali, (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res **34**: D291-295.

Pierce, K. E., J. A. Sanchez, J. E. Rice & L. J. Wangh, (2005) Linear-After-The-Exponential (LATE)-PCR: primer design criteria for high yields of specific single-stranded DNA and improved real-time detection. Proc Natl Acad Sci USA **102**: 8609-8614.

Pink, B., (2005) Agricultural Production Survey. In. Wellington: Ministry of Agriculture and Forestry.

Pizzonia, J., (2001) Electrophoresis gel image processing and analysis using the KODAK 1D software. Biotechniques **30**: 1316-1320.

Ploux, O. & A. Marquet, (1992) The 8-amino-7-oxopelargonate synthase from *Bacillus sphaericus*. Purification and preliminary characterization of the cloned enzyme overproduced in *Escherichia coli*. Biochem J **283**: 327-331.

Podar, M. & A. L. Reysenbach, (2006) New opportunities revealed by biotechnological explorations of extremophiles. Curr Opin Biotechnol **17**: 250-255.

Pohlmann, A., W. F. Fricke, F. Reinecke, B. Kusian, H. Liesegang, R. Cramm, T. Eitinger, C. Ewering, M. Potter, E. Schwartz, A. Strittmatter, I. Voss, G. Gottschalk, A. Steinbuchel, B. Friedrich & B. Bowien, (2006) Genome sequence of the bioplastic-producing "Knallgas" bacterium *Ralstonia eutropha* H16. Nat Biotech **24**: 1257-1262.

Polan, C. E., J. J. McNeill & S. B. Tove, (1964) Biohydrogenation of unsaturated fatty acids by rumen bacteria. J Bacteriol **88**: 1056-1064.

Ponting, C. P., J. Schultz, F. Milpetz & P. Bork, (1999) SMART: Identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res **27**: 229-232.

Poole, K., K. Tetro, Q. Zhao, S. Neshat, D. E. Heinrichs & N. Bianco, (1996) Expression of the multidrug resistance operon *mexA-mexB-oprM* in *Pseudomonas aeruginosa*: *mexR* encodes a regulator of operon expression. Antimic Agen Chemo **40**: 2021-2028.

Pratt, L. A. & R. Kolter, (1998) Genetic analysis of *Escherichia coli* biofilm formation: roles of flagella, motility, chemotaxis and type I pili. Mol Microbiol **30**: 285-293.

Puyet, A., G. H. del Solar & M. Espinosa, (1988) Identification of the origin and direction of replication of the broad-host-range plasmid pLS1. Nucleic Acids Res **16**: 115-133.

Rakotoarivonina, H., G. Jubelin, M. Hebraud, B. Gaillard-Martinie, E. Forano & P. Mosoni, (2002) Adhesion to cellulose of the Gram-positive bacterium *Ruminococcus albus* involves type IV pili. Microbiol **148**: 1871-1880.
Raleigh, E. A. & G. Wilson, (1986) *Escherichia coli* K-12 restricts DNA containing 5-methylcytosine. Proc Natl Acad Sci USA **83**: 9070-9074.

Rasmussen, R., (2001). Quantification on the LightCycler. S. Meuer, C. Wittwer & K. Nakagawara (eds). Springer-Verlag, pp. 21-34.

Ratto, M., M. L. Suihko & M. Siika-aho, (2005) Polysaccharide-producing bacteria isolated from paper machine slime deposits. J Indust Microbiol Biotech **32**: 109-114.

Reeve, J. N., J. Nolling, R. M. Morgan & D. R. Smith, (1997) Methanogenesis: genes, genomes, and who's on first? J Bacteriol **179**: 5975-5986.

Renner, E. D. & R. W. Bernlohr, (1972) Characterization and regulation of pyruvate carboxylase of *Bacillus licheniformis*. J Bacteriol **109**: 764-772.

Reuven, N. B., G. Arad, A. Maor-Shoshani & Z. Livneh, (1999) The mutagenesis protein UmuC is a DNA polymerase activated by UmuD', RecA, and SSB and is specialized for translesion replication. J Biol Chem **274**: 31763-31766.

Reva, O. & B. Tummler, (2008) Think big--giant genes in bacteria. Env Microbiol **10**: 768-777.

Reymond, N., H. Charles, L. Duret, F. Calevro, G. Beslon & J. M. Fayard, (2004) ROSO: optimizing oligonucleotide probes for microarrays. Bioinformatics **20**: 271-273.

Rice, P., I. Longden & A. Bleasby, (2000) EMBOSS: the European molecular biology open software suite. Trends Genet **16**: 276-277.

Robbins, H., (1956) An Empirical Bayes approach to statistics. In: Proceeding of the third Berkeley symposium on mathematical statistics. Berkeley: University of California Press, pp. 157-163.

Rocha, E. R., A. O. Tzianabos & C. J. Smith, (2007) Thioredoxin reductase is essential for thiol/disulfide redox control and oxidative stress survival of the anaerobe *Bacteroides fragilis*. J Bacteriol **189**: 8015-8023.

Roche, (2003) LightCycler fastStart DNA master SYBR green I instruction manual. Basel, Switzerland.

Roelcke, D. & G. Uhlenbruck, (1969) Proteinase K: a new serological effective protease from fungi. Z Med Mikrobiol Immunol **155**: 156-170.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen & P. Nyren, (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem **242**: 84-89.

Rosenberg, C., P. Boistard, J. Denarie & F. Cassedelbart, (1981) Genes controlling early and late functions in symbiosis are located on a megaplasmid in *Rhizobium meliloti*. Mol Gen Genet **184**: 326-333.
Roses, A. D., P. L. St Jean & M. G. Ehm, (2007) Use of whole-genome association scans in disease gene identification, drug discovery and development. IDrugs **10**: 797-804.

Ross, P., R. Mayer, H. Weinhouse, D. Amikam, Y. Huggirat, M. Benziman, E. de Vroom, A. Fidder, P. de Paus, L. A. Sliedregt, G. A. van der Marel & J. H. van Boom, (1990) The cyclic diguanylic acid regulatory system of cellulose synthesis in *Acetobacter xylinum*. Chemical synthesis and biological activity of cyclic nucleotide dimer, trimer, and phosphothioate derivatives. J Biol Chem **265**: 18933-18943.

Ruas-Madiedo, P., M. Gueimonde, M. Fernandez-Garcia, C. G. de los Reyes-Gavilan & A. Margolles, (2008) Mucin degradation by *Bifidobacterium* strains isolated from the human intestinal microbiota. Appl Environ Microbiol **74**: 1936-1940.

Ruiz, T. R., S. Andrews & G. B. Smith, (2000) Identification and characterization of nuclease activities in anaerobic environmental samples. Can J Microbiol **46**: 736-740.

Russell, J. B., (1985) Fermentation of cellodextrins by cellulolytic and noncellulolytic rumen bacteria. Appl Environ Microbiol **49**: 572-576.

Russell, J. B., (1987) Effect of extracellular pH on growth and proton motive force of *Bacteroides succinogenes*, a cellulolytic ruminal bacterium. Appl Environ Microbiol **53**: 2379-2383.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream & B. Barrell, (2000) Artemis: sequence visualization and annotation. Bioinformatics **16**: 944-945.

Sadowsky, M. J. & B. B. Bohlool, (1983) Possible involvement of a megaplasmid in nodulation of soybeans by fast-growing Rhizobia from china. Appl Environ Microbiol **46**: 906-911.

Saeki, H., M. Akira, K. Furuhashi, B. Averhoff & G. Gottschalk, (1999) Degradation of trichloroethene by a linear-plasmid-encoded alkene monooxygenase in *Rhodococcus corallinus* (*Nocardia corallina*) B-276. Microbiol UK **145**: 1721-1730.

Saito, H. & K. I. Miura, (1963) Preparation of transforming deoxyribonucleic acid by phenol treatment. Biochim Biophys Act **72**: 619-629.

Salanoubat, M., S. Genin, F. Artiguenave, J. Gouzy, S. Mangenot, M. Arlat, A. Billault, P. Brottier, J. C. Camus, L. Cattolico, M. Chandler, N. Choisne, C. Claudel-Renard, S. Cunnac, N. Demange, C. Gaspin, M. Lavie, A. Moisan, C. Robert, W. Saurin, T. Schiex, P. Siguier, P. Thebault, M. Whalen, P. Wincker, M. Levy, J. Weissenbach & C. A. Boucher, (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature **415**: 497-502.

Salmond, G. P. C., (1994) Secretion of extracellular virulence factors by plant-pathogenic bacteria. Annu Rev Phytopathol **32**: 181-200.

Salzberg, S. L., A. L. Delcher, S. Kasif & O. White, (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res **26**: 544-548.

Sambrook, J. & D. W. Russell, (2001) Molecular cloning: A laboratory manual. Cold Spring Harbor, New York.

Sandberg, R., C. I. Branden, I. Ernberg & J. Coster, (2003) Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene **311**: 35-42.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe & M. Smith, (1977a) Nucleotide sequence of bacteriophage *phi X174* DNA. Nature **265**: 687-695.

Sanger, F. & A. R. Coulson, (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol **94**: 441-448.

Sanger, F., S. Nicklen & A. R. Coulson, (1977b) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA **74**: 5463-5467.

Sargent, K. L., P. Ng, C. Evelegh, F. L. Graham & R. J. Parks, (2004) Development of a size-restricted pIX-deleted helper virus for amplification of helper-dependent adenovirus vectors. Gene Ther **11**: 504-511.

Sauer, E., M. Merdanovic, A. P. Mortimer, G. Bringmann & J. Reidl, (2004) PnuC and the utilization of the nicotinamide riboside analog 3-aminopyridine in *Haemophilus influenzae*. Anti Agen Chemo **48**: 4532-4541.

Scaife, J. & J. D. Gross, (1962) Inhibition of multiplication of an Flac factor in Hfr cells of *Escherichia coli* K-12. Biochem Biophys Res Commun **7**: 403-407.

Schaefer, D. M., C. L. Davis & M. P. Bryant, (1980) Ammonia saturation constants for predominant species of rumen bacteria. J Dairy Sci **63**: 1248-1263.

Scheifinger, C. C. & M. J. Wolin, (1973) Propionate formation from cellulose and soluble sugars by combined cultures of *Bacteroides succinogenes* and *Selenomonas ruminantium*. Appl Microbiol **26**: 789-795.

Schneiker, S., O. Perlova, O. Kaiser, K. Gerth, A. Alici, M. O. Altmeyer, D. Bartels, T. Bekel, S. Beyer, E. Bode, H. B. Bode, C. J. Bolten, J. V. Choudhuri, S. Doss, Y. A. Elnakady, B. Frank, L. Gaigalat, A. Goesmann, C. Groeger, F. Gross, L. Jelsbak, L. Jelsbak, J. Kalinowski, C. Kegler, T. Knauber, S. Konietzny, M. Kopp, L. Krause, D. Krug, B. Linke, T. Mahmud, R. Martinez-Arias, A. C. McHardy, M. Merai, F. Meyer, S. Mormann, J. Munoz-Dorado, J. Perez, S. Pradella, S. Rachid, G. Raddatz, F. Rosenau, C. Ruckert, F. Sasse, M. Scharfe, S. C. Schuster, G. Suen, A. Treuner-Lange, G. J. Velicer, F. J. Vorholter, K. J. Weissman, R. D. Welch, S. C. Wenzel, D. E. Whitworth, S. Wilhelm, C. Wittmann, H. Blocker, A. Puhler & R. Muller, (2007) Complete genome sequence of the myxobacterium *Sorangium cellulosum*. Nat Biotech **25**: 1281-1289.

Schofield, D. A., C. Westwater, B. D. Hoel, P. A. Werner, J. S. Norris & M. G. Schmidt, (2003) Development of a thermally regulated broad-spectrum promoter system for use in pathogenic Gram-positive species. Appl Environ Microbiol **69**: 3385-3392.

Schroeter, A., S. Riethdorf & M. Hecker, (1988) Amplification of different ColE1 plasmids in an *Escherichia coli relA* strain. J Basic Microbiol **28**: 553-555.

Schultz, J., R. R. Copley, T. Doerks, C. P. Ponting & P. Bork, (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res **28**: 231-234.

Schultz, J., F. Milpetz, P. Bork & C. P. Ponting, (1998) SMART, a simple modular architecture research tool: Identification of signaling domains.Proc Natl Acad Sci USA **95**: 5857-5864.

Schwartz, D. C. & C. R. Cantor, (1984) Seperation of yeast chromosome sized DNAs by pulsed-field gradient gel-electrophoresis. Cell **37**: 67-75.

Schwartz, E. & B. Friedrich, (2001) A physical map of the megaplasmid pHG1, one of three genomic replicons in *Ralstonia eutropha* H16. Fems Microbiol Let **201**: 213-219.

Schwartz, E., A. Henne, R. Cramm, T. Eitinger, B. Friedrich & G. Gottschalk, (2003) Complete nucleotide sequence of pHG1: A *Ralstonia eutropha* H16 megaplasmid encoding key enzymes of H_2-based lithoautotrophy and anaerobiosis. J Mol Biol **332**: 369-383.

Sekine, M., S. Tanikawa, S. Omata, M. Saito, T. Fujisawa, N. Tsukatani, T. Tajima, T. Sekigawa, H. Kosugi, Y. Matsuo, R. Nishiko, K. Imamura, M. Ito, H. Narita, S. Tago, N. Fujita & S. Harayama, (2006) Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4. Env Microbiol **8**: 334-346.

Seto, H., S. Imai, T. Tsuruoka, A. Satoh, M. Kojima, S. Inouye, T. Sasaki & N. Otake, (1982) Studies on the biosynthesis of bialaphos (SF-1293). 1. Incorporation of ^{13}C- and ^{2}H-labeled precursors into bialaphos. J Antibio **35**: 1719-1721.

Sewell, G. W., H. C. Aldrich, D. Williams, B. Mannarelli, A. Wilkie, R. B. Hespell, P. H. Smith & L. O. Ingram, (1988) Isolation and characterization of xylan-degrading strains of *Butyrivibrio fibrisolvens* from a Napier grass-fed anaerobic digester. Appl Environ Microbiol **54**: 1085-1090.

Shane, B. S., L. Gouws & A. Kistner, (1969) Cellulolytic bacteria occurring in the rumen of sheep conditioned to low-protein teff hay. J Gen Microbiol **55**: 445-457.

Sherman, L. A. & M. L. Gefter, (1976) Studies on the mechanism of enzymatic DNA elongation by *Escherichia coli* DNA polymerase II. J Mol Biol **103**: 61-76.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki & H. Ishikawa, (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp*. APS. Nature **407**: 81-86.

Shimizu, S., H. Kobayashi, E. Masai & M. Fukuda, (2001) Characterization of the 450-kb linear plasmid in a polychlorinated biphenyl degrader, *Rhodococcus sp*. strain RHA1. Appl Environ Microbiol **67**: 2021-2028.

Shine, J. & L. Dalgarno, (1975) Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16*S* RNA and translational specificity of the ribosome. Eur J Biochem **57**: 221-230.

Shoemaker, N. B., G. R. Wang & A. A. Salyers, (1992) Evidence for natural transfer of a tetracycline resistance gene between bacteria from the human colon and bacteria from the bovine rumen. Appl Environ Microbiol **58**: 1313-1320.

Shotland, Y., S. Koby, D. Teff, N. Mansur, D. A. Oren, K. Tatematsu, T. Tomoyasu, M. Kessel, B. Bukau, T. Ogura & A. B. Oppenheim, (1997) Proteolysis of the phage lambda CII regulatory protein by FtsH (HflB) of *Escherichia coli*. Mol Microbiol **24**: 1303-1310.

Shuvaev, A. N. & A. V. Brilkov, (2007) A model of bacterial cell cycle duration based on DnaA dynamics and estimation of the population cost of bacterial plasmids. Dokl Biochem Biophys **416**: 233-236.

Siedow, A., R. Cramm, R. A. Siddiqui & B. Friedrich, (1999) A megaplasmid-borne anaerobic ribonucleotide reductase in *Alcaligenes eutrophus* H16. J Bacteriol **181**: 4919-4928.

Simakov, O., (2006) MB DNA Analysis. In. Germany: Molbiosoft, Available: http://www.molbiosoft.de/html/downloads.htm.

Simm, R., A. Lusch, A. Kader, M. Andersson & U. Romling, (2007) Role of EAL-containing proteins in multicellular behavior of *Salmonella enterica* sv. *typhimurium*. J Bacteriol **189**: 3613-3623.

Simm, R., M. Morr, A. Kader, M. Nimtz & U. Romling, (2004) GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. Mol Microbiol **53**: 1123-1134.

Singh, S. K. & P. C. Banerjee, (2007) Nucleotide sequence analysis of cryptic plasmid pAM5 from *Acidiphilium multivorum*. Plasmid **58**: 101-114.

Singh, S. K., O. V. Kurnasov, B. Chen, H. Robinson, N. V. Grishin, A. L. Osterman & H. Zhang, (2002) Crystal structure of *Haemophilus influenzae* NadR protein. A bifunctional enzyme endowed with NMN adenyltransferase and ribosylnicotinimide kinase activities. J Biol Chem **277**: 33291-33299.

Skovgaard, M., L. J. Jensen, S. Brunak, D. Ussery & A. Krogh, (2001) On the total number of genes and their length distribution in complete microbial genomes. Trends Genet **17**: 425-428.

Smith, B., W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector & C. Rosse, (2005) Relations in biomedical ontologies. Genome Biol **6**: R46.

Smyth, G. K., (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat App Gene Mol Biol **3**: Article 3.

Smyth, G. K., (2005) Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor. R. Gentleman, V. Carey, W. Huber, R. Irazarry & S. Dudoit (eds). New York: Springer, pp. 397-420.

Sobral, B. W. S., R. J. Honeycutt, A. G. Atherly & M. McClelland, (1991) Electrophoretic seperation of the 3 *Rhizobium meliloti* replicons. J Bacteriol **173**: 5173-5180.

Sonnhammer, E. L., S. R. Eddy & R. Durbin, (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins **28**: 405-420.

Sourjik, V. & H. C. Berg, (2002) Binding of the *Escherichia coli* response regulator CheY to its target measured *in-vivo* by fluorescence resonance energy transfer. Proc Natl Acad Sci USA **99**: 12669-12674.

Spinelli, S. L., H. S. Malik, S. A. Consaul & E. M. Phizicky, (1998) A functional homolog of a yeast tRNA splicing enzyme is conserved in higher eukaryotes and in Escherichia coli. Proc Natl Acad Sci USA **95**: 14136-14141.

Sprenger, G. A., (1995) Genetics of pentose-phosphate pathway enzymes of *Escherichia coli* K-12. Arch Microbiol **164**: 324-330.

Srikumar, R., C. J. Paul & K. Poole, (2000) Influence of mutations in the *mexR* repressor gene on expression of the MexA-MexB-oprM multidrug efflux system of *Pseudomonas aeruginosa*. J Bacteriol **182**: 1410-1414.

St John, G., N. Brot, J. Ruan, H. Erdjument-Bromage, P. Tempst, H. Weissbach & C. Nathan, (2001) Peptide methionine sulfoxide reductase from *Escherichia coli* and *Mycobacterium tuberculosis* protects bacteria against oxidative damage from reactive nitrogen intermediates. Proc Natl Acad Sci USA **98**: 9901-9906.

Staden, R., (1996) The Staden sequence analysis package. Mol Biotech **5**: 233-241.

Staden, R., K. F. Beal & J. K. Bonfield, (2000) The Staden package, 1998. Meth Molec Biol **132**: 115-130.

Stecker, C., A. Johann, C. Herzberg, B. Averhoff & G. Gottschalk, (2003) Complete nucleotide sequence and genetic organization of the 210-kilobase linear plasmid of *Rhodococcus erythropolis* BD2. J Bacteriol **185**: 5269-5274.

Stewart, C. S., H. J. Flint & M. P. Bryant, (1997) The rumen bacteria. In: The Rumen Microbial Ecosystem. P. N. Hobson & C. S. Stewart (eds). London: Chapman and Hall, pp. 10-72.

Stewart, R. C., (1997) Kinetic characterization of phosphotransfer between CheA and CheY in the bacterial chemotaxis signal transduction pathway. Biochemistry **36**: 2030-2040.

Stoletzki, N. & A. Eyre-Walker, (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol **24**: 374-381.

Stouthamer, A. H. & S. Kooijman, (1993) Why it pays for bacteria to delete disused DNA and to maintain megaplasmids. Int J Gener Mol Microbiol **63**: 39-43.

Strobel, H. J., (1994) Pentose transport by the ruminal bacterium *Butyrivibrio fibrisolvens*. FEMS Microbiol Lett **122**: 217-222.

Stryer, L., (1995) Biochemistry. W.H. Freeman and company, New York.

Sueoka, N., (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc Natl Acad Sci USA **47**: 1141-1149.

Suhas, P. J. Carrott & M. M. Ribeiro Carrott, (2007) Lignin--from natural adsorbent to activated carbon: A review. Bioresour Technol **98**: 2301-2312.

Sullivan, J. T., J. R. Trzebiatowski, R. W. Cruickshank, J. Gouzy, S. D. Brown, R. M. Elliot, D. J. Fleetwood, N. G. McCallum, U. Rossbach, G. S. Stuart, J. E. Weaver, R. J. Webby, F. J. De Bruijn & C. W. Ronson, (2002) Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. J Bacteriol **184**: 3086-3095.

Sun, C., M. Zhou, Y. Li & H. Xiang, (2006a) Molecular characterization of the minimal replicon and the unidirectional theta replication of pSCM201 in extremely halophilic archaea. J Bacteriol **188**: 8136-8144.

Sun, X. M., Y. P. Tang, X. Z. Meng, W. W. Zhang, S. Li, Z. R. Deng, Z. K. Xu & R. T. Song, (2006b) Sequencing and analysis of a genomic fragment provide an insight into the *Dunaliella viridis* genomic sequence. Acta Biochim Biophys Sin **38**: 812-820.

Sutherland, I. W. & J. F. Wilkinson, (1968) The exopolysaccharide of *Klebsiella aerogens* A3 (S1) (type 54). The isolation of O-acetylated octasaccharide, tetrasaccharide and trisaccharide. Biochem J **110**: 749-754.

Suwanto, A. & S. Kaplan, (1989) Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. J Bacteriol **171**: 5850-5859.

Suzek, B. E., M. D. Ermolaeva, M. Schreiber & S. L. Salzberg, (2001) A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics **17**: 1123-1130.

Taghavi, S., M. Mergeay & D. vanderLelie, (1997) Genetic and physical maps of the *Alcaligenes eutrophus* CH34 megaplasmid pMOL28 and its derivative pMOL50 obtained after temperature-induced mutagenesis and mortality. Plasmid **37**: 22-34.

Tam, N. H. & R. Borriss, (1995) The *thyA* gene from *Bacillus subtilis* exhibits similarity with the phage phi 3T thymidylate synthase gene. Microbiol **141**: 291-297.

Tamura, K., J. Dudley, M. Nei & S. Kumar, (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**: 1596-1599.

Tatusov, R. L., M. Y. Galperin, D. A. Natale & E. V. Koonin, (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28**: 33-36.

Tatusov, R. L., E. V. Koonin & D. J. Lipman, (1997) A genomic perspective on protein families. Science **278**: 631-637.

Teather, R. M., (1982) Isolation of plasmid DNA from *Butyrivibrio fibrisolvens*. Appl Environ Microbiol **43**: 298-302.

Thein, S. L. & R. B. Wallace, (1993) The use of synthetic oligonucleotides as specific hybridisation probes in the diagnosis of genetic disorders. In: Human genetic disease analysis: a practical approach. K. E. Davis (ed). New York IRL Press.

Thomas, C. D., D. F. Balson & W. V. Shaw, (1990) *In-vitro* studies of the initiation of *Staphylococcal* plasmid replication. Specificity of RepD for its origin (oriD) and characterization of the Rep-*ori* tyrosyl ester intermediate. J Biol Chem **265**: 5519-5530.

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan & A. Narechania, (2003) PANTHER: A library of protein families and subfamilies indexed by function. Genome Res **13**: 2129-2141.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin & D. G. Higgins, (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25**: 4876-4882.

Thompson, J. D., D. G. Higgins & T. J. Gibson, (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**: 4673-4680.

Thurston, B., K. A. Dawson & H. J. Strobel, (1993) Cellobiose versus glucose utilization by the ruminal bacterium *Ruminococcus albus*. Appl Environ Microbiol **59**: 2631-2637.

Thwaites, W. M., C. H. Davis, N. Wallis-Biggart, L. M. Wondrack & M. T. Abbott, (1979) Urea: obligate intermediate of pyrimidine-ring catabolism in *Rhodosporidium toruloides*. J Bacteriol **137**: 1145-1150.

Tichopad, A., A. Dzidic & M. W. Pfaffl, (2002) Improving quantitative real-time RT-PCR reproducibility by boosting primer-linked amplification efficiency. Biotech Lett **24**: 2053-2056.

TIGR, (2001) MANATEE. Available: http://www.tigr.org/software/

Titok, M., C. Suski, B. Dalmais, S. D. Ehrlich & L. Janniere, (2006) The replicative polymerases PolC and DnaE are required for theta replication of the *Bacillus subtilis* plasmid pBS72. Microbiol **152**: 1471-1478.

Titok, M. A., J. Chapuis, Y. V. Selezneva, A. V. Lagodich, V. A. Prokulevich, S. D. Ehrlich & L. Janniere, (2003) *Bacillus subtilis* soil isolates: Plasmid replicon analysis and construction of a new theta-replicating vector. Plasmid **49**: 53-62.

Tomoyasu, T., J. Gamer, B. Bukau, M. Kanemori, H. Mori, A. J. Rutman, A. B. Oppenheim, T. Yura, K. Yamanaka, H. Niki & et al., (1995) *Escherichia coli* FtsH is a membrane-bound, ATP-dependent protease which degrades the heat-shock transcription factor sigma 32. Embo J **14**: 2551-2560.

Ton-That, H., G. Liu, S. K. Mazmanian, K. F. Faull & O. Schneewind, (1999) Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of *Staphylococcus aureus* at the LPXTG motif. Proc Natl Acad Sci USA **96**: 12424-12429.

Touchman, J. W., D. M. Wagner, J. Hao, S. D. Mastrian, M. K. Shah, A. J. Vogler, C. J. Allender, E. A. Clark, D. S. Benitez, D. J. Youngkin, J. M. Girard, R. K. Auerbach, S. M. Beckstrom-Sternberg & P. Keim, (2007) A North American *Yersinia pestis* draft genome sequence: SNPs and phylogenetic analysis. PLoS One **2**: e220.

Trinci, A. P., S. E. Lowe, A. Milne & M. K. Theodorou, (1988) Growth and survival of rumen fungi. Biosystems **21**: 357-363.

Tung, W. L. & K. C. Chow, (1995) A modified medium for efficient electrotransformation of *E. coli*. Trends Genet **11**: 128-129.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis & J. I. Gordon, (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444**: 1027-1031.

Unden, G. & J. Bongaerts, (1997) Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. Biochim Biophys Act **1320**: 217-234.

Urbanek, S., (2007) rJAVA: Low-level R to Java interface. Available: http://CRAN.R-project.org/package=rjava

Ushida, K., C. J. Newbold & J. P. Jouany, (1997) Interspecies hydrogen transfer between the rumen ciliate *Polyplastron multivesiculatum* and *Methanosarcina barkeri*. J Gen Appl Microbiol **43**: 129-131.

Van Larebeke, N., G. Engler, M. Holsters, S. Van den Elsacker, I. Zaenen, R. A. Schilperoort & J. Schell, (1974) Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. Nature **252**: 169-170.

Vanin, E. F., (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet **19**: 253-272.

Varga, G. A. & E. S. Kolver, (1997) Microbial and animal limitations to fiber digestion and utilization. J Nutr **127**: 819S-823S.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers & H. O. Smith, (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science **304**: 66-74.

Vetting, M. W., S. d. C. LP, M. Yu, S. S. Hegde, S. Magnet, S. L. Roderick & J. S. Blanchard, (2005) Structure and functions of the GNAT superfamily of acetyltransferases. Arch Biochem Biophys **433**: 212-226.

Vogels, G. D. & C. Stumm, (1980) Interactions between methanogenic bacteria and hydrogenic ciliates in the rumen. Ant Van Leeuwen **46**: 108.

Wachenheim, D. E. & J. A. Patterson, (1992) Anaerobic production of extracellular polysaccharide by *Butyrivibrio fibrisolvens* nyx. Appl Environ Microbiol **58**: 385-391.

Wallace, R. J., L. C. Chaudhary, N. McKain, N. R. McEwan, A. J. Richardson, P. E. Vercoe, N. D. Walker & D. Paillard, (2006) *Clostridium proteoclasticum*: A ruminal bacterium that forms stearic acid from linoleic acid. FEMS Microbiol Lett **265**: 195-201.

Wang, T. P. & J. O. Lampen, (1952) Metabolism of pyrimidines by a soil bacterium. J Biol Chem **194**: 775-783.

Wang, X., R. Gorlitsky & J. S. Almeida, (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol **23**: 1099-1103.

Ware, C. E., T. Bauchop, J. F. Hudman & K. Gregg, (1992) Cryptic plasmid Pbf1 from *Butyrivibrio fibrisolvens* Ar10 - its use as a replicon for recombinant plasmids. Curr Microbiol **24**: 193-197.

Weaver, D. T. & M. L. DePamphilis, (1984) The role of palindromic and non-palindromic sequences in arresting DNA synthesis *in-vitro* and *in-vivo*. J Mol Biol **180**: 961-986.

Wensvoort, M., (2002) Agricultural production statistics (final) 2002 In.: Ministry of Agriculture and Forestry. Available: http://www.stats.govt.nz/

Whelan, J. A., N. B. Russell & M. A. Whelan, (2003) A method for the absolute quantification of cDNA using real-time PCR. J Immunol Meth **278**: 261-269.

White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft, M. L. Gwinn, W. C. Nelson, D. L. Richardson, K. S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J. J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K. S. Makarova, L. Aravind, M. J. Daly & C. M. Fraser, (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science **286**: 1571-1577.

Willems, A., M. Amat-Marco & M. D. Collins, (1996) Phylogenetic analysis of *Butyrivibrio* strains reveals three distinct groups of species within the *Clostridium* subphylum of the Gram-positive bacteria. Int J Syst Bacteriol **46**: 195-199.

Wolin, M. J., (1976) Interactions between H_2-producing and methane producing species. In: Microbial formation and utilization of gases. H. G. Schlegel, G. Gottschalk & N. Pfennig (eds). Gottingen: Goltze Press pp. 141-150.

Wolin, M. J., (1981) Fermentation in the rumen and human large intestine. Science **213**: 1463-1468.

Woodsmall, R. M. & D. A. Benson, (1993) Information resources at the National Center for Biotechnology Information. Bull Med Libr Assoc 81: 282-284.
Wu, C. H., A. Nikolskaya, H. Huang, L. S. Yeh, D. A. Natale, C. R. Vinayaka, Z. Z. Hu, R. Mazumder, S. Kumar, P. Kourtesis, R. S. Ledley, B. E. Suzek, L. Arminski, Y. Chen, J. Zhang, J. L. Cardenas, S. Chung, J. Castro-Alvear, G. Dinkov & W. C. Barker, (2004) PIRSF: Family classification system at the protein information resource. Nucleic Acids Res **32**: D112-114.

Wu, L. & N. E. Welker, (1991) Cloning and characterization of a glutamine transport operon of *Bacillus stearothermophilus* NUB36: effect of temperature on regulation of transcription. J Bacteriol **173**: 4877-4888.

Wyman, S. K., R. K. Jansen & J. L. Boore, (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20**: 3252-3255.

Xue, G. P., J. S. Johnson, K. L. Bransgrove, K. Gregg, C. E. Beard, B. P. Dalrymple, K. S. Gobius & J. H. Aylward, (1997) Improvement of expression and secretion of a fungal xylanase in the rumen bacterium *Butyrivibrio fibrisolvens* OB156 by manipulation of promoter and signal sequences. J Biotech **54**: 139-148.

Yeats, C., M. Maibaum, R. Marsden, M. Dibley, D. Lee, S. Addou & C. A. Orengo, (2006) Gene3D: modelling protein structure, function and evolution. Nucleic Acids Res **34**: D281-284.

Ying, X., Y. Wang, H. R. Badiei, V. Karanassios & K. Ma, (2007) Purification and characterization of an iron-containing alcohol dehydrogenase in extremely thermophilic bacterium *Thermotoga hypogea*. Arch Microbiol **187**: 499-510.

Yu, J., T. E. Cleveland, W. C. Nierman & J. W. Bennett, (2005) *Aspergillus flavus* genomics: gateway to human and animal health, food safety, and crop resistance to diseases. Rev Iberoam Micol **22**: 194-202.

Zacher, A. N. r., C. A. Stock, J. W. n. Golden & G. P. Smith, (1980) A new filamentous phage cloning vector: fd-tet. Gene **9**: 127-140.

Zhang, H. M., Z. Li, M. Tsudome, S. Ito, H. Takami & K. Horikoshi, (2005) An alkali-inducible flotillin-like protein from *Bacillus halodurans* C-125. Prot J **24**: 125-131.

Zhang, R. G., C. E. Andersson, T. Skarina, E. Evdokimova, A. M. Edwards, A. Joachimiak, A. Savchenko & S. L. Mowbray, (2003) The 2.2 A resolution structure of RpiB/AlsB from *Escherichia coli* illustrates a new approach to the ribose-5-phosphate isomerase reaction. J Mol Biol **332**: 1083-1094.

Zimmermann, H., (1992) 5'-Nucleotidase: molecular structure and functional aspects. Biochem J **285**: 345-365.

Zund, P. & G. Lebek, (1980) Generation time-prolonging R-plasmids - correlation between increases in the generation time of *Escherichia coli* caused by R-plasmids and their molecular-size. Plasmid **3**: 65-69.

Zwierz, K., A. Gindzienski, D. Glowacka & T. Porowski, (1981) The degradation of glycoconjugates in the human gastric mucous membrane. Acta medi Academ Scienti Hungari **38**: 145-152.