Comparative genome analysis of *Campylobacter jejuni* using whole genome DNA microarrays☆

B.M. Pearson[a], C. Pin[a], J. Wright[a], K. I'Anson[a], T. Humphrey[b], J.M. Wells[a,*]

[a] *BBSRC Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK*
[b] *School of Clinical Veterinary Science, University of Bristol, The Churchill Building, Langford, Bristol BS40 5DT, UK*

Received 3 July 2003; revised 30 September 2003; accepted 2 October 2003

First published online 16 October 2003

Edited by Takashi Gojobori

Abstract Whole genome DNA microarrays were constructed and used to investigate genomic diversity in 18 *Campylobacter jejuni* strains from diverse sources. New algorithms were developed that dynamically determine the boundary between the conserved and variable genes. Seven hypervariable plasticity regions (PR) were identified in the genome (PR1 to PR7) containing 136 genes (50%) of the variable gene pool. When comparisons were made with the sequenced strain NCTC11168, the number of absent or divergent genes ranged from 2.6% (40 genes) to 10.2% (163) and in total 16.3% (269) of the genes were variable. PR1 contains genes important in the utilisation of alternative electron acceptors for respiration and may confer a selective advantage to strains in restricted oxygen environments. PR2, 3 and 7 contain many outer membrane and periplasmic proteins and hypothetical proteins of unknown function that might be linked to phenotypic variation and adaptation to different ecological niches. PR4, 5 and 6 contain genes involved in the production and modification of antigenic surface structures.

Key words: Campylobacter; Plasticity; Diversity; Genomics; Virulence

1. Introduction

Gastrointestinal infection with Campylobacter is the leading cause of bacterial diarrhoeal disease worldwide and the most common antecedent to the peripheral neuropathies Guillain Barre syndrome (GBS) and Miller Fisher syndrome (MFS) [1,2]. Campylobacter is a zoonotic pathogen of humans and livestock animals such as cows, pigs, sheep, and farmed poultry are reservoirs for the organism [2]. The majority of Campylobacter infections are sporadic in nature and are caused by two species *Campylobacter jejuni* and *Campylobacter coli*. Symptoms from infection range from very mild watery diarrhoea to profuse bloody diarrhoea with mucosal damage and inflammation especially in the ileum and jejunum [3,4]. Striking differences exist among *C. jejuni* strains in respect of phenotypic properties such as cell invasiveness, rates of translocation across cell monolayers, toxin production, and colonisation in chickens [5]. It is not known to what extent these observed differences might reflect differences in genetic factors.

The development of serotyping and other genetic typing methods for Campylobacter has contributed greatly to surveillance and epidemiological studies particularly with the identification of serotypes associated with the development of post-infection neuropathies. However, current typing methods have failed to distinguish strains with phenotypic characteristics associated with pathogenicity, virulence and different ecological habitats with one exception, a recent report linking detection of a polymerase chain reaction (PCR) amplified DNA fragment with clinically invasive strains of Campylobacter [6]. Genotyping and multilocus sequencing typing data indicate that Campylobacter strains are genetically highly diverse with only a few clonal lineages and that the population exhibits genetic plasticity [7,8].

The sequencing of microbial genomes has made it possible to construct whole genome DNA microarrays that can be used to give useful information about the genetic composition of bacterial strains by hybridisation of fluorescently labelled genomic DNA [9,10]. Limitations to this approach are that point mutations, small deletions, gene rearrangements, and novel genes not present on the microarray will not be detected. Additionally, intergenic regions containing promoter elements and non-translated RNAs are typically not present on DNA microarrays and thus are not included in the analysis. Nevertheless, comparative genome analysis using microarrays is providing new insights into microbial evolution and genetic diversity in microbial populations [11–14]. In a recent study the use of microarray genotyping was compared to random amplified polymeric DNA and the Penner serotyping system as an epidemiological tool for the analysis of *C. jejuni* clinical isolates [15].

We have constructed and used a whole genome microarray comprising unique probes for each open reading frame (ORF) in the genome to examine the genetic diversity among 18 different isolates of *C. jejuni*. To analyse the microarray data we have developed a new algorithm that dynamically determines the boundary between the conserved and variable genes. By comparing the ratios of the fluorescence signals obtained one can predict whether or not a gene is variable. A variable gene could be absent from the genome, partially deleted or divergent to the extent that it gives a significantly lower hybridisation signal than the control strain. Many published studies use a constant ratio as a cut-off for assignment of variable genes that is often empirically determined by comparison of the reference strain to a similar strain with known deletions. However, in this situation the degree of sequence

*Corresponding author.
E-mail address: jerry.wells@bbsrc.ac.uk (J.M. Wells).

☆ Supplementary data associated with this article can be found at doi: 10.1016/S0014-5793(03)01164-5

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)01164-5

divergence resulting in a ratio that falls outside the cut-off value is often undetermined. A further limitation of the constant cut-off approach is that it demands high reproducibility between microarrays because experimental variation in the width of the ratio distribution for conserved genes would result in the misclassification of genes. In this paper we have applied a novel algorithm to determine an independent cut-off for each hybridisation dataset and on this basis assigned genes as being variable (i.e. absent or divergent) or present. The algorithm is focused on finding the conserved genes and consequently the variable genes are also identified. One advantage of this approach is that once the conserved genes are found and described by a set of parameters, this can be used for the normalisation of datasets from different arrays. A similar concept lies behind the approach described recently by Kim et al. [16]. This method assumes normality for the log of the ratio of the fluorescence intensities and identifies the conserved genes by fitting the normal density function to the values with the higher frequency; the variable genes are identified in the tails of the density function. This approach could become problematic when the assumption of normality is incorrect or when the number of conserved genes is equal or smaller than the number of missing genes. In this situation it could be difficult to identify the conserved genes from the histogram of the frequency of the fluorescence ratio. In contrast the new algorithm described here does not rely on any assumptions about the distribution of fluorescence intensities and it identifies the conserved genes even if they only represent a quarter of the hybridisation dataset. Another distinguishing feature of our approach is that the boundary between the conserved and variable genes is not constant within an array because variable genes are classified according to the result of an *F* test. Application of this algorithm to a microarray based study of genetic diversity revealed new insights into the nature of genetic diversity in natural populations of *C. jejuni* and expands our knowledge of the genetic and phenotypic variation associated with variable loci.

2. Materials and methods

2.1. Construction of the C. jejuni *DNA microarray*

DNA fragments of individual ORFs were amplified using ORF specific primers for those present in strain NCTC11168 (Sigma Genosys ORFmer set). All PCR was carried out using HotStart Taq (Qiagen, UK) and an Eppendorf Mastercycler with the following parameters: 95°C for 15 min, followed by 30 cycles of 95°C 30 s, 50°C 30 s, 72°C 120 s, with a final 72°C for 600 s incubation. Genomic NCTC11168 DNA (5 ng) was used as template and approximately 80 pmol of each primer in a reaction volume of 50 μl. Successful DNA amplification was confirmed by electrophoresis in agarose gels for assessment of band size and intensity. Before microarraying, PCR products were cleaned up on a Qiagen 9600 robot using a Qiaquick 96-well Biorobot kit (Qiagen, UK), dried and resuspended in a half volume of 3× saline sodium citrate (SSC) containing 0.01% sarkosyl.

PCR probes were then spotted on poly-L-lysine coated slides using an in-house Stanford designed arrayer, see http://cmgm.stanford.edu/pbrown/mguide/index.htm for associated software and protocols. The final array contained 2304 features including controls.

2.2. Bacterial strains and growth

The origin and serotype of the Campylobacter strains used in this study is shown in Table 1. Strains were grown at 42°C under microaerophilic conditions (10% CO_2, 5% O_2, 85% N_2; relative humidity 80%) on Skirrow agar plates or in Mueller Hinton broth using a MACS-MG-1000 controlled atmosphere workstation (DW Scientific, UK).

2.3. Fluorescent labelling of genomic DNA

Genomic DNA was purified from bacteria using the Qiagen Dneasy™ method (Qiagen, UK) Approximately 10 μg of DNA mixed with random sequence hexamer oligonucleotides (Amersham Pharmacia) in 1× reaction buffer (50 mM Tris pH 7.2, 10 mM $MgSO_4$, and 0.1 mM dithiothreitol) and heated to 100°C for 5 min and then stored for approximately 10 min on ice. DNA was then labelled using Cy3 (green) or Cy5 (red) fluorescent dyes and purified for hybridisation essentially as described elsewhere (http://cmgm.stanford.edu/pbrown/protocols/4_genomic.html).

2.4. Microarray hybridisation

For each test hybridisation Cy5 labelled control genomic DNA from the sequenced strain NCTC11168 was mixed together with Cy3 labelled genomic DNA from the test strain in 50 μl of Glasshyb hybridisation buffer (Sigma, UK) and boiled for 5 min. Then 40μl of

Table 1
Summary of comparative genomics data for 18 strains of *C. jejuni*

Strain No.	Source	Serotype DA and Penner	Variable genes % (variable, conserved)	Number of runs (observed, expected, *P* value)	χ^2 distance (*P* value)
11322	NCTC, human	HS44, PEN 1	3.0 (44, 1417)	(38, 85, < 0.0001)	(3.972, 0.0463)
11351	NCTC	HS4, PEN 23	3.6 (55, 1472)	(42, 106, < 0.0001)	(9.563, 0.0020)
11392	NCTC, human	HS6, PEN 6	9.0 (141, 1427)	(68, 257, < 0.0001)	(14.467, 0.0007)
11827	NCTC, human Fla-	HS6	9.3 (137, 1331)	(68, 249, < 0.0001)	(13.425, 0.0012)
12502	NCTC	HS3, PEN 3	5.5 (84, 1453)	(64, 159, < 0.0001)	(7.816, 0.0200)
12507	NCTC	HS8, PEN 8	2.6 (40, 1484)	(30, 78, < 0.0001)	(4.279, 0.0386)
12541	NCTC	HS40 PEN 40	8.5 (133, 1431)	(68, 244, < 0.0001)	(12.522, 0.0019)
12547	NCTC	ND, PEN 36	10.2 (163, 1431)	(94, 293, < 0.0001)	(18.046, 0.0001)
12744	NCTC, milk	UT	8.8 (135, 1417)	(66, 247, < 0.0001)	(13.514, 0.0012)
1887	Puddle on farm	HS27	8.1 (127, 1426)	(94, 233, < 0.0001)	(14.390, 0.0007)
3874	Puddle on farm	ND	6.3 (100, 1452)	(56, 187, < 0.0001)	(9.776, 0.0075)
4872	Poultry faeces	ND	5.2 (85, 1419)	(56, 160, < 0.0001)	(9.828, 0.0073)
29	T2 culture (T.H.)	ND	8.8 (141, 1454)	(68, 257, < 0.0001)	(14.484, 0.0007)
30	T4 culture (T.H.)	UT	6.1 (98, 1500)	(64, 184, < 0.0001)	(12.767, 0.0017)
1090	Kitchen towel	HS13	6.3 (94, 1405)	(56, 176, < 0.0001)	(9.788, 0.0075)
249	Poultry faeces	UT	4.0 (59, 1409)	(56, 113, < 0.0001)	(4.914, 0.0857)
22-547504	Human	HS31	6.5 (97, 1401)	(52, 182, < 0.0001)	(8.807, 0.0122)
40-555943	Human	ND	3.2 (45, 1351)	(22, 87, < 0.0001)	(8.319, 0.0039)

Serotyping data are based on detection of heat stable antigens (HS) by the direct agglutination (DA) technique [36] and by the Penner method (PEN) in the case of the reference strains from the National Collection of Type Cultures (NCTC). UT: unable to be typed; ND: not done. The number of runs is equivalent to the number of groups of variable or present genes that are consecutive in the genome of the sequenced reference strain NCTC11168. The average coverage of the microarray was 92.2% of the annotated genes. The χ^2 collection statistic expresses the difference between the theoretical and observed distributions of the consecutive variable genes.

the hybridisation mixture was put onto the microarray slide and sealed with a coverslip in a GeneMachine hybridisation chamber (Anachem, UK) and incubated at 60°C for 18 h. This method, known as differential labelling, allows the hybridisation of fluorescently labelled control (Cy5) and test (Cy3) DNA to be measured for each probe on the microarray. As the genetic composition of the control strain is known from the genome sequence it serves as a control fluorescence signal for each probe and is used for comparison with the test DNA signal during the statistical analysis (see below). Following hybridisation, microarray slides were washed briefly in prewarmed 1× SSC, 0.03% sodium dodecyl sulphate (SDS) to remove the coverslip and then washed for 5 min in each of the following buffers: (a) 1× SSC, 0.03% SDS, (b) 0.2× SSC, and finally (c) 0.05× SSC. Microarray slides were dried by centrifugation at 300×*g* for 15 min before scanning.

2.5. Microarray data analysis

DNA microarrays were scanned using an Axon GenePix 4000A microarray laser scanner (Axon Instruments, CA, USA) and the data from detected features initially processed using the GenePix 3.0 software. Poor features were excluded from analysis if they contained abnormalities or were within regions of high background. A new Visual Basic program was developed to analyse the hybridisation data. A dynamic boundary between conserved and variable (i.e. absent or divergent) genes was established for each hybridisation dataset by taking into account the global variability of the whole set of conserved genes as well as the individual variability of the fluorescence measurement for each gene, based on the replicates.

Let R_p and G_p denote the Cy5 and Cy3 fluorescence signals, respectively for the conserved (p = present) genes. The relationship between the natural logarithms of the two signals is described as $\ln R_p = \alpha_p + \beta_p \ln G_p$, where α_p and β_p are estimated by regression using the average fluorescence intensities for each of the conserved genes; σ_p is the standard error of the fit. The conserved and variable genes are found simultaneously by an iterative algorithm involving a series of steps. In the first step, the parameters α_p, β_p and σ_p are assigned with initial values. Use of initial values of 0, 1 and 0.05 for α_p, β_p and σ_p, respectively, ensures that none of the variable genes are classified as conserved in the first iteration. In the second step genes are classified as being conserved or variable. The potentially variable genes are identified as having average intensities that lie outside the $3\sigma_p$ boundary estimated for the genes currently classified as conserved, i.e. $\ln R_t < (\alpha_p + \beta_p \ln G_t) - 3\sigma_p$ or $\ln R_t > (\alpha_p + \beta_p \ln G_t) + 3\sigma_p$, where R_t and G_t are the average values of the intensities for the test gene in the red and green channels. Potentially variable genes are tested using the regression models $\ln R_t = \alpha_t + \beta_p \ln G_t$ and $\ln R_t = \alpha_p + \beta_t \ln G_t$ that are fitted independently to the replicates for that gene. In these models β_p and α_p are made equal to the regression coefficients of the conserved genes. *F*-tests are carried out on the hypothesis $\alpha_p = \alpha_t$ and $\beta_p = \beta_t$; if one of these is rejected, the tested gene is classified as variable otherwise it is classified as conserved. In the third step, the values for α_p, β_p and σ_p are re-calculated by regression according to the genes classified as conserved in the second step. The second and third steps are iterated until the three parameters, α_p, β_p and σ_p, converge.

To estimate the error rate in these experiments and to check the performance of the algorithm we performed control hybridisations with the same genomic DNA samples labelled with Cy3 and Cy5. Genomic DNA from strain 12547 was digested with *Sau*3A and *Rsa*I and labelled with Cy3 and Cy5, respectively. The samples were then combined, hybridised to the microarray and the data analysed with the algorithm described above. In total 0.1% of the features were identified as variable when in fact they should have been scored as present in both samples. A similar control was subsequently performed with undigested DNA from strain 11168 on a different set of microarrays using slightly modified hybridisation conditions and the error rate was similar (0.2% of genes misclassified as variable).

2.6. Randomness of deletions test

The randomness of the location of the variable genes in the genome was tested by applying a 'Runs test' [17] and by comparing the observed and expected distributions of the number of consecutive variable genes for a circular genome (see table in supplementary material).

3. Results and discussion

Genetic diversity among 18 strains of *C. jejuni* was analysed by comparative genomic DNA hybridisation to a DNA microarray (Table 1). To analyse the microarray data we have developed a new algorithm that dynamically determines the boundary between the conserved and variable genes. Control hybridisations showed that this algorithm gives a low error rate of misclassification (0.1–0.2% of genes on the array). Sequencing data on gene loci identified as variable using the new algorithm revealed that genes that have average fluorescence intensities with small deviations from the 3σ (i.e. 3 standard deviations) boundary of the conserved genes (see data analysis in Materials and Methods) have small deletions or variable stretches of sequence whereas large deviations from 3σ boundary usually indicate complete absence of the gene in the test strain. For example, deletion of 22% of the *leuA* gene (1536 bp) sequences hybridising to the probe resulted in that gene being scored as variable in strain NCTC12547. Genes that were indeed shown to be absent from a test strain were consistently classified as variable in that strain using our statistical approach. Furthermore, comparable results were obtained with the approach described here and that of Kim et al. [16] when analysing datasets in which the conserved genes in both samples were in the majority and the log ratio of the fluorescence intensities followed a normal distribution (e.g. one or two genes differently classified; data not shown). In contrast to the method of Kim et al. [16] the new algorithm describe here does not rely on any assumptions about the normality of the distribution of fluorescence intensities and has provided reliable results, even when the logarithm of the ratios between intensities did not follow a normal distribution or when the present genes represented only 20% of the hybridisation dataset (Pin et al., unpublished). Thus this new algorithm might be especially useful for the analysis of interspecies hybridisations.

The 1654 annotated genes present in the genome of strain NCTC11168 were included in the analysis and of these 1385 (83.7%) were common to all strains tested and are referred to in this study as the core gene set. The number of variable genes for each strain ranged from 2.6% (40 genes) to 10.2% (163 genes) and in total 269 (16.3%) of the NCTC11168 genes were variable among the 18 strains tested. A previous study by Dorell et al. on 11 *C. jejuni* strains [18] reported that at least 21% of the genes present in the genome of strain NCTC11168 were dispensable and classified as absent or highly divergent using a constant cut-off method and a low cost microarray based on a tiled set of pUC18 sequencing clones covering the whole genome. Only 34.5% of the PCR probes were gene specific, making the data potentially difficult to interpret and subsequently Kim et al. [16] were unable to precisely reproduce the results of this study using their own constant cut-off re-analysis. The choice of strains and larger number of isolates included in our study may also account for differences to the previously reported estimate of the number of variable genes [11].

The genes scored as missing in our test strains were often grouped in clusters according to the genome sequence of NCTC11168, with the largest group comprising 24 consecutive genes. The theoretical probability of such a cluster of 24 consecutive variable genes occurring by random addition or loss of genes is less than 10^{-20}. The exact gene order for each

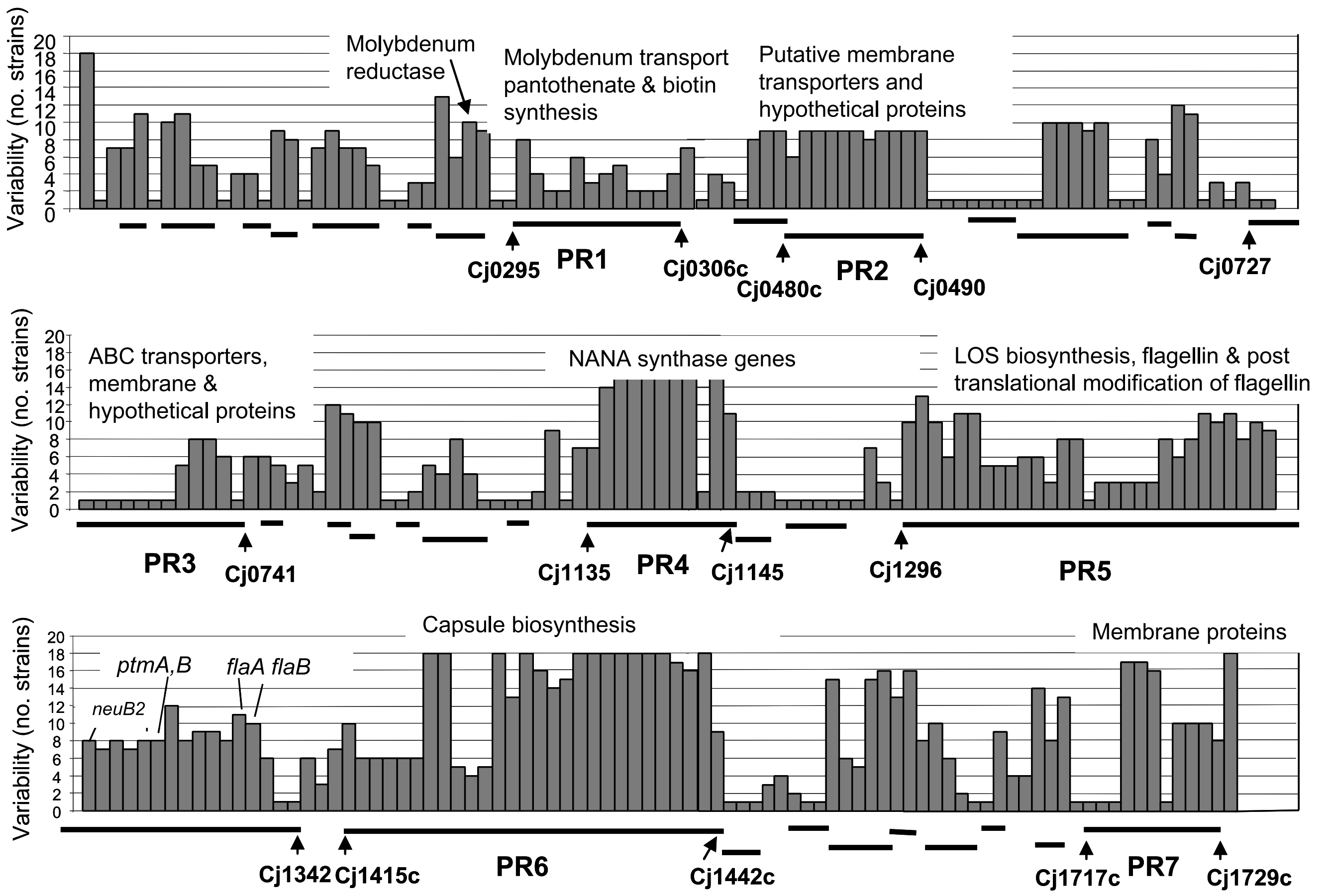

Fig. 1. Each variable gene is represented by a column, the height of which indicates the number of strains in which the gene was variable. The length of the horizontal bars underneath the columns indicates how many of the variable genes are consecutive in the genome of strain NCTC11168. The major PR (PR1 to PR7) and selected genes are indicated.

test strain is unknown but preliminary analysis of genome sequence data for another sequenced strain RM1221 (www.tigr.org) indicated that the bulk of the genome is likely to be co-linear in *C. jejuni* strains of different origin. The randomness of the location of the variable genes was also rejected for all the strains by a 'Runs test' analysis ($P<0.05$), indicating that the location of the variable genes was non-random. Thus, as expected, these genes appear to have been added or lost from the genome in groups during evolution (Table 1).

Many genes from the core set of genes found in all strains are predicted to be involved in vital functions such as energy metabolism, cell division, protein and peptide secretion, and synthesis of macromolecules such as DNA, RNA, and proteins. Known virulence genes including those encoding the cytolethal distending toxin (Cj0077c, Cj0078c, Cj0079c), putative adhesion factor *jlpA* (Cj0983), fibronectin binding protein *cadF* (Cj1478c) and *CiaB* (Cj0914c), a gene that is expressed during cell invasion, [19–21] were present in all strains tested. Two surface proteins (PEB3 Cj0289c and CgpA Cj1670c) recently shown to be N-glycosylated in *Campylobacter* [22,23] were also conserved among all strains tested. The recently described glycosylation pathway genes (i.e. Cj1120c to Cj1126c *pglB to pglG*) involved in the N-linked glycosylation of surface structures [24,22] were all conserved except *pglE* a hypothetical protein that was variable in seven strains.

In other genotyping studies the genomes of strains are often compared as a series of aligned bar charts with a line representing each gene and the colour of the line indicating its presence or absence. While this method does help to indicate hypervariable loci we found it more useful to plot the total number of variable genes versus their frequency of variability among the strains tested (Fig. 1). As seen in Fig. 1 this analysis revealed that many of the variable genes were clustered in seven large distinct regions of the genome referred to as plasticity regions (PR) containing between 11 and 45 genes; many smaller variable loci also exist. Four of the seven large PR are newly identified whereas the capsule, lipooligosacharide (LOS) and flagellin modification loci of *C. jejuni* have been shown to be highly variable in a previous microarray comparison and genetic study [11]. PR1 contains genes encoding the molybdenum transport apparatus (*modC* Cj0300c, *B* Cj0301c and *A* Cj0302), the pantothenate biosynthesis genes (*panD* Cj0296c, Cj0297c and *B* Cj0298c) and hypothetical proteins of putative or unknown function. Pantothenic acid is a precursor of coenzyme A that carries acyl groups for enzyme reactions involved in fatty acid synthesis and fatty acid oxidation and pyruvate oxidation in higher organisms. Its precise role in microbes is less well understood although a recent study showed that a pantothenate auxotroph of *Mycobacterium tuberculosis* is limited for growth in infected mice [25]. In some organisms molybdenum is involved in the reduction of nitrate by nitrate reductase, a flavoprotein enzyme containing both molybdenum and cytochrome *b*. The use of nitrate as a terminal electron acceptor in place of oxygen is important for growth under oxygen restricted conditions in *Campylobacter* [26]. Just upstream of PR1 is the molybdoenzyme reductase gene (Cj0264c) responsible for respiration of trimethylamine-*N*-oxide and dimethyl sulfoxide under oxygen limiting conditions [26]. This gene is also absent or highly divergent in 10 strains indicating that the capacity of *Campylobacter* to utilise alternative electron receptors for respiration under highly restricted oxygen conditions may be strain specific. It is possible that the acquisition of these genes has provided some strains with a selective advantage in specific ecological niches.

PR2 is flanked by truncated genes for a putative altronate hydrolase and an aldehyde dehydrogenase. The locus also contains genes encoding a putative oxidoreductase, sugar transporter and two hypothetical proteins of unknown function.

Several of the genes in PR3 are deleted in only one of the strains tested, NCTC12547, the strain most highly variable compared to the sequenced one. The genes in this region comprise components of one or more ABC transporters (all absent in strain NCTC12547) and several hypothetical proteins are also absent in some other strains.

PR4 starts downstream of the recently described protein glycosylation locus (pgl: Cj1120c to Cj1126c) involved in the N-linked glysolylation of surface proteins in *Campylobacter* [22–24]. This locus was previously thought to encode genes involved in the biosynthesis of lipopolysaccharides (LPS) (wla genes [27]). The seven *pgl* pathway genes were conserved among the strains tested with the exception of *pglE* (Cj1120c) a hypothetical protein that was absent or divergent in seven of the 18 strains. Variable genes in PR4 include three putative galactosyltransferases (Cj1136, Cj1138 and Cj1139c) and a putative glycosyltransferase (Cj1135) the exact function of which is unknown. PR4 also contains the *N*-acetyl neuraminic acid synthase genes (*neuB1*, *neuC1 and neuA1* Cj1141, Cj1142 and Cj1143) involved in the sialylation of LOS [23]. The genome of NCTC11168 also contains two other *neuB*-like genes (*neuB2* Cj1327 and *neuB3* Cj1317) but independent mutation of the three *neuB* genes identified *neuB1* as the synthase involved in sialylation of LOS [23]. No phenotype was evident for *neuB2* mutants in strain NCTC11168 but *neuB3* mutants were aflagellate and non-motile. *C. jejuni* sialylated LOS is thought to lead to the generation of cross-reactive antibodies to human sialylated gangliosides, resulting in the neuropathies associated with GBS and MFS [28].

The sequenced strain NCTC11168 contains both *neuB1* and *neuC1* and is serotype heat stable antigen HS2. NCTC11168 was previously shown to produce sialylated LOS and to be associated with the development of GBS [29,30]. The majority of the strains tested in our study (16 out of 18) lacked both *neuB1* and *neuC1*, suggesting that they would not express a sialylated LOS. This supports other evidence that only a limited number of strains and serotypes are potentially able to elicit antibody responses that cross-react with human gangliosides. Four strains also lacked both *neuB2* and *neuC2*, genes that have been previously implicated in post-translational sialylation of the flagellum [30].

The longest stretches of variable genes occur in PR5 between Cj1296 and C1342c (approximately 48 kb) and in PR6 Cj1415c to Cj1442c (approximately 34 kb) (Fig. 1). PR5 contains a large number of hypothetical proteins of unknown function belonging to protein families 1318 and 617 (see http://www.sanger.ac.uk/Projects/C_jejuni/) and variable genes encoding putative acyl carrier proteins (acpP2, P3, P4) and a β-ketoacyl-acyl carrier protein synthase III (fabH2) that is essential for fatty acid biosynthesis in other organisms [31,32]. However, the fabH2 gene is similar to fabH (Cj0328c) that was conserved among the strains tested. Thus fabH is probably the essential enzyme initiating fatty acid biosynthesis in *C. jejuni* and fabH2 may be an alternative one, which is dispensable. PR5 also contains genes involved in LOS biosynthesis, and post-translational modification of flagellin, i.e. *pmtA* (Cj1331) *pmtB* (Cj1332) and Cj1316c [33], the flagellin structural genes *flaA* (Cj1339c), *flaB* (Cj1338c) and two possible flagella proteins (Cj1312, Cj1313).

The *fla* genes are independently transcribed but *flaA* is transcribed at higher levels than *flaB*. Mutational analysis of the *flaA*/*B* genes suggests that under the conditions tested *flaA* is the major flagellum structural protein but *flaB* is incorporated in small amounts into the whole filament. The *flaA* and *flaB* genes of strain NCTC11168 have a high level of base sequence identity (approximately 93%) with the heterogeneity being confined to a small central region of the gene and the 5′ and 3′ terminal sequences. Thus the *flaA* and *flaB* probes on the microarray are not gene specific. The flagellum enables *Campylobacter* to swim through viscous environments such as the mucus found in the gastrointestinal tract and exhibit chemotaxis. Several studies have shown that the flagellum was needed for colonisation in a range of animals [34]. The flagellin genes were scored as variable in 10 out of the 18 strains tested but given the importance of the flagellum in *Campylobacter* survival and virulence it seems most likely that this result is a reflection of sequence divergence and not the absence of flagellin genes. This conclusion is supported by the fact that the absence of flaA derived PCR products is rare in studies that utilise this locus for PCR typing purposes.

PR6 is highly divergent and contains the capsule biosynthesis locus which is flanked by capsule transport genes in a similar arrangement to that found in *Escherichia coli*. Originally, the HS used for serotyping of *C. jejuni* were assumed to be LPS as in the case of *E. coli* and *Salmonella* [35]. In a study of 32 strains, it was shown that none of these expressed long chain LPS species characteristic of the O-antigen oligosaccharide repeat units and that the Penner antibodies detected a heat stable surface antigen that could be a capsular polysaccharide [36]. The sequencing of the genome of NCTC11168 led to the discovery of a 42.6 kb capsular polysaccharide biosynthesis locus (Cj1415c to Cj1442c) that is flanked by genes with similarity to capsule transport genes found in *E. coli* (i.e. *kpsS, C* (Cj1413c and Cj1414c) and *F, D, E, T and M* Cj1443c to Cj1448c). Further evidence for the role of this capsule locus in production of the serotyping antigen was demonstrated [37] by finding that four out of six different serotypes lacking genes for *kpsM*, could not be typed using Penner serotyping. Mutants in genes for *kpsS* and *kpsC* that are thought to be involved in ligation of the polysaccharide to KDO-DAGP also resulted in strain NCTC11168 (HS2) becoming unable to be typed but this was not the case with the serotypes HS1 and HS10 [37]. Interestingly the capsule transport gene *kpsM* was variable in the Penner serotype 8 reference strain NCTC12507 (HS8) and subsequently shown to be absent in this strain by Southern blotting (data not shown). In *E. coli* *kpsM* is essential for transport of the capsule so it seems likely that the capsule is not the serotypic determinant of the HS8 and Penner 8 serotype and that another HS is involved. Three strains in our collection were unable to be typed but this was not associated with variable capsule transport genes.

In the capsule biosynthesis locus (PR6) of NCTC11168 there is a cluster of five genes (Cj1416c–Cj1420c) that are either present (12 strains) or absent (six strains) as a whole cluster. This cluster of five genes encodes a phosphoenolpyruvate synthase, GMP synthase and GIP cytidyl transferase, a putative methyl transferase (Cj1419c) and one 'contingency' gene of unknown function Cj1420c that contains a homopolymeric nucleotide tract that has been shown to undergo variation in the number of repeat nucleotides, resulting in phenotypic variation [38]. The reason for the organisation of these genes with the capsule locus and any possible role they might have in capsule biosynthesis genes is unclear. There are three other contingency genes in PR6 (Cj1421c, Cj1422c, Cj1426c) that are specific to the sequenced strain NCTC11168 and one contingency gene (Cj1429c) that is found in only two other strains. Four of the seven putative glucosyl transferases (GT) present in strain NCTC11168 (HS2) were not detected in any other strain and the three other GT were rarely detected in other strains.

PR7 contains a cluster of putative outer membrane, periplasmic proteins and hypothetical proteins and several genes involved in leucine biosynthesis that were absent from only one of the 18 strains tested. PR7 also contains gene Cj1729c (*flgE2*) encoding a probable flagella hook protein that was also divergent in eight strains. This gene may be an alternative flagella hook protein utilised by *Campylobacter* as it is similar to the flagella hook protein gene *flgE* (Cj0043) that was conserved in all but one of the strains tested.

4. Conclusions

In total 16.3% (269) of the genes present in the sequenced strain NCTC11168 were either absent or highly variable in sequence among the strains of *C. jejuni* examined in this study. In each strain the variable genes were often present in large clusters, suggesting that they were acquired or lost from the genome in groups during evolution. Seven major PR (designated PR1 to PR7) were identified in the genome and these comprise 136 (50%) of the variable gene pool. Unlike the variable gene clusters (i.e. islands and pathogenicity islands) found in the genomes of *E. coli* and *Salmonella* the *C. jejuni* PR do not have a markedly different G+C content to the bulk of the genome and they are not associated with mobile elements important in horizontal DNA transfer. However, many strains of *C. jejuni* are naturally competent for DNA uptake and transformation and this is likely to have played a major role in generating genetic variability. PR4, 5 and 6 contain genes involved in the production of surface structures including LOS, flagellum, and capsule as well as the pathway enzymes for sialylation of LOS, flagellum and post-translational glycosylation of the flagellum. These findings indicate that genome diversity is linked to production of variant surface structures that might play a role in the avoidance of innate and adaptive immune responses in the host. PR1 contains genes important in the utilisation of alternative electron acceptors for respiration and may confer a selective advantage to strains in restricted oxygen environments. PR2, 3 and 7 contain many outer membrane and periplasmic proteins and hypothetical ones of unknown function. These genes warrant further investigation as they might be associated with adaptation of *Campylobacter* to different ecological niches.

The variable regions of the genome identified in this study highlight genetic factors that might be linked to phenotypic variation and adaptation to different ecological niches. The results of this work are also likely to have an impact on the design of future genetic typing schemes and microarray based epidemiological studies.

Acknowledgements: We gratefully acknowledge funding from the BBSRC core strategic grant. We thank Karen Mattick and Jenny Slader at the Food Microbiology Collaborating Laboratory, Exeter, UK, for the supply of strains and Jenny Frost and Richard Thwaites at PHLS Colindale, London, UK, for serotyping the strains and for helpful discussion.

References

[1] Wheeler Jr., J.S., Siroky, M.B., Pavlakis, A. and Krane, R.J. (1984) J. Urol. 131, 917–919.

[2] Solomon, E.B. and Hoover, D.G. (1999) J. Food Safety 19, 121–136.

[3] Wassenaar, T.M., Bleumink-Pluym, N.M.C. and van der Zeijst, B.A.M. (1991) EMBO J. 10, 2055–2061.

[4] Hickey, T.E., McVeigh, A.L., Scott, D.A., Michielutti, R.E., Bixby, A., Carroll, S.A., Bourgeois, A.L. and Guerry, P. (2000) Infect. Immun. 68, 6535–6541.

[5] Hu, L. and Kopecko, D.J. (1999) Infect. Immun. 67, 4171–4182.

[6] Carvalho, A.C., Ruiz-Palacios, G.M., Ramos-Cervantes, P., Cervantes, L.E., Jiang, X. and Pickering, L.K. (2001) J. Clin. Microbiol. 39, 1353–1359.

[7] Dingle, K.E., Colles, F.M., Wareing, D.R., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J., Urwin, R. and Maiden, M.C. (2001) J. Clin. Microbiol. 39, 14–23.

[8] Suerbaum, S., Lohrengel, M., Sonnevend, A., Ruberg, F. and Kist, M. (2001) J. Bacteriol. 183, 2553–2559.

[9] Wells, J.M. and Bennik, M.H.J. (2003) Nutr. Res. Rev. 16, 21–35.

[10] Bowtell, D.D. (1999) Nat. Genet. 21, 25–32.

[11] Israel, D.A., Salama, N., Krishna, U., Rieger, U.M., Atherton, J.C., Falkow, S. and Peek Jr., R.M. (2001) Proc. Natl. Acad. Sci. USA 98, 14625–14630.

[12] Call, D.R., Borucki, M.K. and Besser, T.E. (2003) J. Clin. Microbiol. 41, 632–639.

[13] Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L. and Falkow, S. (2000) Proc. Natl. Acad. Sci. USA 97, 14668–14673.

[14] Porwollik, S., Wong, R.M. and McClelland, M. (2002) Proc. Natl. Acad. Sci. USA 99, 8956–8961.

[15] Leonard II, E.E., Takata, T., Blaser, M.J., Falkow, S., Tompkins, L.S. and Gaynor, E.C. (2003) J. Infect. Dis. 187, 691–694.

[16] Kim, C.C., Joyce, E.A., Chan, K. and Falkow, S. (2002) Genome Biol. 3, research0065.1-research0065.17.

[17] Zar, J.H. (1999) Biostatistical Analysis, Prentice-Hall, London.

[18] Dorrell, N., Mangan, J.A., Laing, K.G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B.G., Parkhill, J., Stoker, N.G., Karlyshev, A.V., Butcher, P.D. and Wren, B.W. (2001) Genome Res. 11, 1706–1715.

[19] Pickett, C.L., Pesci, E.C., Cottle, D.L., Russell, G., Erdem, A.N. and Zeytin, H. (1996) Infect. Immun. 64, 2070–2078.

[20] Jin, S., Joe, A., Lynett, J., Hani, E.K., Sherman, P. and Chan, V.L. (2001) Mol. Microbiol. 39, 1225–1236.

[21] Konkel, M.E., Kim, B.J., Rivera-Amill, V. and Garvis, S.G. (1999) Mol. Microbiol. 32, 691–701.

[22] Wacker, M., Linton, D., Hitchen, P.G., Nita-Lazar, M., Haslam, S.M., North, S.J., Panico, M., Morris, H.R., Dell, A., Wren, B.W. and Aebi, M. (2002) Science 298, 1790–1793.

[23] Linton, D., Allan, E., Karlyshev, A.V., Cronshaw, A.D. and Wren, B.W. (2002) Mol. Microbiol. 43, 497–508.

[24] Szymanski, C.M., Yao, R., Ewing, C.P., Trust, T.J. and Guerry, P. (1999) Mol. Microbiol. 32, 1022–1030.

[25] Sambandamurthy, V.K., Wang, X., Chen, B., Russell, R.G., Der-

rick, S., Collins, F.M., Morris, S.L. and Jacobs Jr., W.R. (2002) Nat. Med. 8, 1171–1174.

[26] Sellars, M.J., Hall, S.J. and Kelly, D.J. (2002) J. Bacteriol. 184, 4187–4196.

[27] Fry, B.N., Korolik, V., ten Brinke, J.A., Pennings, M.T., Zalm, R., Teunis, B.J., Coloe, P.J. and van der Zeijst, B.A. (1998) Microbiology 144, 2049–2061.

[28] Yuki, N. (1997) J. Infect. Dis. 176, S150–S153.

[29] Aspinall, G.O., McDonald, A.G., Raju, T.S., Pang, H., Mills, S.D., Kurjanczyk, L.A. and Penner, J.L. (1992) J. Bacteriol. 174, 1324–1332.

[30] Linton, D., Karlyshev, A.V., Hitchen, P.G., Morris, H.R., Dell, A., Gregson, N.A. and Wren, B.W. (2000) Mol. Microbiol. 35, 1120–1134.

[31] Revill, P.W., Bibb, M.J., Scheu, A-K., Kieser, H.J. and Hopwood, D.A. (2001) J. Bacteriol. 183, 3526–3530.

[32] Kiatpapan, P., Kobayashi, H., Sakaguchi, M., Ono, H., Yamashita, M., Kaneko, Y. and Murooka, Y. (2001) Appl. Environ. Microbiol. 67, 426–433.

[33] Thibault, P., Logan, S.M., Kelly, J.F., Brisson, J-R., Ewing, C.P., Trust, T.J. and Guerry, P. (2001) J. Biol. Chem. 37, 34862–34870.

[34] Ketley, J.M. (1997) Microbiology 143, 5–21.

[35] Penner, J.L., Hennessy, J.N. and Congi, R.V. (1983) Infect. Immun. 55, 1806–1812.

[36] Chart, H., Frost, J.A., Oza, A., Thwaites, R., Gillanders, S. and Rowe, B. (1996) J. Appl. Bacteriol. 81, 635–640.

[37] Karlyshev, A.V., Linton, D., Gregson, N.A., Lastovica, A.J. and Wren, B.W. (2000) Mol. Microbiol. 35, 529–541.

[38] Linton, D., Gilbert, M., Hitchen, P.G., Dell, A., Morris, H.R., Wakarchuk, W.W., Gregson, N.A. and Wren, B.W. (2000) Mol. Microbiol. 37, 501–514.