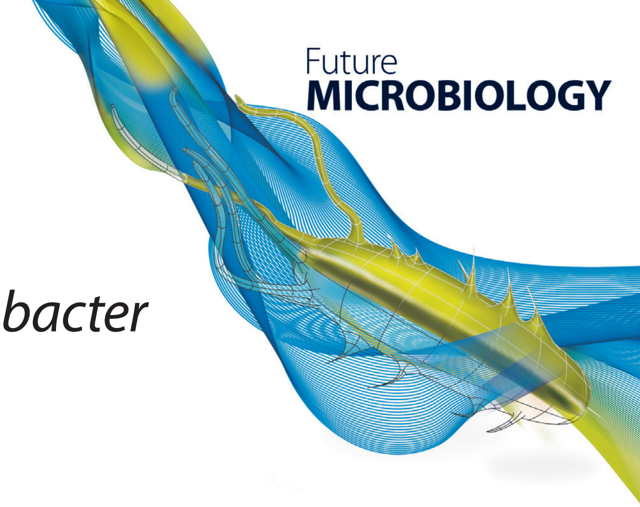
RESEARCH ARTICLE

For reprint orders, please contact: reprints@futuremedicine.com

Future
MICROBIOLOGY



CRISPR–*cas* loci profiling of *Cronobacter sakazakii* pathovars

Pauline Ogrodzki[1] & Stephen James Forsythe*,[1]

Aim: *Cronobacter sakazakii* sequence types 1, 4, 8 and 12 are associated with outbreaks of neonatal meningitis and necrotizing enterocolitis infections. However clonality results in strains which are indistinguishable using conventional methods. This study investigated the use of clustered regularly interspaced short palindromic repeats (CRISPR)–*cas* loci profiling for epidemiological investigations. **Materials & methods:** Seventy whole genomes of *C. sakazakii* strains from four clonal complexes which were widely distributed temporally, geographically and origin of source were profiled. **Results & conclusion:** All strains encoded the same type I-E subtype CRISPR–*cas* system with a total of 12 different CRISPR spacer arrays. This study demonstrated the greater discriminatory power of CRISPR spacer array profiling compared with multilocus sequence typing, which will be of use in source attribution during *Cronobacter* outbreak investigations.

First draft submitted: 13 April 2016; Accepted for publication: 5 August 2016;
Published online: 10 November 2016

Background

The bacterial pathogen *Cronobacter* has become the focus of much attention especially due to its association with neonatal meningitis [1]. A number of potential virulence traits causing cytopathogenicity of host cells have been proposed, most recent being the production of outer membrane vesicles [2,3]. The *Cronobacter* genus is composed of seven species, of which *Cronobacter sakazakii* is the species most frequently isolated from neonatal and infant cases of infection. A curated open access multilocus sequence typing (MLST) database has been established for the genus with >1400 strains and associated metadata [4–7]. This database has enabled the recognition of certain *Cronobacter* clonal lineages within the genus as pathogenic variants, whereas others are primarily commensal organisms of the environment. The major pathovars are *C. sakazakii* sequence type (ST) 4 which is more predominantly associated with neonatal meningitis, *C. sakazakii* ST12 with neonatal necrotizing enterocolitis, *C. sakazakii* clonal complex (CC) 1 strains are primarily isolated from infant formula and clinical sources, whereas *C. sakazakii* ST8 are isolated from clinical and nonformula food sources [6,8,9]. The original 7-loci MLST scheme is congruent with both 53-loci ribosomal MLST and 1865-loci core genome MLST as well as whole-genome phylogeny [6]. The reason for the predominance of certain pathovars with particular clinical presentations could be due to their greater environmental fitness resulting in increased exposure, as well as the possible encoding of virulence genes [9].

Although a number of virulence traits have been proposed in *C. sakazakii* none are unique to specific pathovars [10]. However, whole-genome analysis has now been used to generate a capsular profiling scheme which is based on the K-antigen and colanic acid encoding genes [11]. This

KEYWORDS

• CRISPR–cas • *Cronobacter sakazakii* • genotyping • MLST • sequence types

[1] Pathogen Research Group, School of Science & Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK
*Author for correspondence: Tel.: +44 115 8483529; stephen.forsythe@ntu.ac.uk

Future Medicine part of fsg

10.2217/fmb-2016-0070 © 2016 Future Medicine Ltd

ISSN 1746-0913

scheme revealed that strains of *C. sakazakii* and *C. malonaticus*, isolated from cases with the most severe neonatal clinical presentations (invasive meningitis and necrotizing enterocolitis) were capsular profile K2:CA2. Whereas the *C. sakazakii* and *C. malonaticus* strains associated with less severe clinical cases tended to be capsular profile K1:CA1 [11].

Although pulsed-field gel electrophoresis (PFGE) is commonly referred to as a 'gold standard' for genotyping, it has a number of limitations with clonal organisms which are genetically homogeneous such as *Salmonella enterica* serovar Enteritidis. For example, one pulsotype pattern using the restriction enzyme XbaI (JEGX01.0004) of *S.* Enteritidis is the most common pulsotype in the CDC database, and is given by approximately 45% of all clinical isolates [12]. Hence PFGE is unable to distinguish all unrelated *Salmonella* isolates during an epidemiological investigation. This may also account for the observation that the same pulsotype is obtained for unrelated clinical *C. sakazakii* strains [6,13].

Other genotyping methods for *Cronobacter* have been published, but none are able to distinguish strains to the same level as MLST. O-serotyping of *Cronobacter* has been investigated by several groups. Initially this was achieved by applying random fragment length polymorphism profiling across the O-antigen region and PCR assays targeting the presumed conserved serotype-specific genes *wzx* and *wzy* [14]. These methods were able to distinguish a total of seven serotypes in *C. sakazakii,* thus considerably less than the 189 defined sequence types for this species [6,15]. The O-antigen analysis has been supported in part by chemical composition determination of the frequently isolated serotypes *C. sakazakii* O:1 and O:2 [16,17]. The PCR-primer pair approach for O-serotyping has been superseded however by allele profiling of *gnd* and *galF* (encoding 6-phosphogluconate dehydrogenase and UTP-glucose-1-phosphate uridylyltransferase subunits, respectively). This DNA-sequence based method is a more reliable and expansive method for O-antigen determination of *Cronobacter* [11]. It has expanded the defined number of serotypes in the *Cronobacter* genus from 18 to 34. Nevertheless, none of the above DNA-banding pattern genotyping methods for *Cronobacter* are as discriminatory as 7-loci MLST which has >450 sequence defined STs [6].

High resolution bacterial genotyping methods can provide opportunities to improve our understanding of bacterial population genetics, evolution and epidemiology. In addition, popular DNA sequence-based methods such as MLST and PFGE have centralized curated databases (i.e., PubMLST and PulseNet) with open access to facilitate international participation. Nevertheless, the advances in next generation DNA sequencing methods have led to new genome-based typing schemes with even higher resolution than PFGE or MLST. Of particular current interest are the loci of 'clustered regularly interspaced short palindromic repeats' (CRISPRs) and CRISPR-associated genes (*cas*) protein-coding genes [18].

CRISPRs are reportedly found in approximately 80% of archeal genomes and approximately 48% of eubacterial genomes [18]. There are a number of different CRISPR–*cas* systems, often named according to their first identification organism, in other words, *Yersinia pseudotuberculosis* (YPIII), *Escherichia coli* (type I-E), *Neisseria meningitidis* (Nmeni) and *Mycobacterium tuberculosis* (Mtube) [19].

In general, CRISPR–*cas* systems have three sections: *cas* gene cluster, an AT-rich leader sequence, followed by a CRISPR spacer array composed of short (~24–48 nucleotide) direct repeat sequences separated by similarly sized, unique spacers which are usually derived from mobile genetic elements such as bacteriophages and plasmids [19,20]. *Cas* genes are found in the majority of CRISPR-containing genomes and when several CRISPRs of the same CRISPR–*cas* system are present in a single genome, then usually only a single set of *cas* genes is clustered with one of the CRISPRs.

It has been proposed that CRISPR–*cas* systems provide adaptive immunity from invasive genetic elements (phages and plasmids), regulate lysogeny and biofilm formation [21,22]. The AT leader sequence is believed to act as a promoter such that the CRISPR spacer array is transcribed and processed into small CRISPR RNAs (crRNAs). The mature crRNA and some Cas proteins target complementary nucleic acids, such as an invading phage genome, resulting in the degradation of the target DNA [23]. This leads to an acquired specific immunity against the infection by the bacteriophage [21]. Consequently, CRISPR spacer arrays may diverge between closely related strains due to recent spacer acquisition(s).

CRISPR–*cas* loci adapt by acquiring the new spacer sequences at the leader proximal end of the array. This results in polarity of the array as older spacers are at the distal end to the AT leader region. In addition, spacers are normally lost by deletion in a nonpolar manner to avoid an excessive accumulation of spacers in the CRISPR array. However there is considerable uncertainty over this issue as CRISPRs with several hundred spacers have been found, and possibly this variation is related to several factors such as the organisms' ecosystem, along with phage and plasmid exposure. Horizontal transfer of CRISPR and *cas* genes occurs between strains of the same species and even distant species and genera [24]. Subsequently not all strains within a species will necessarily possess the same sets of CRISPR–*cas* genes.

Many applications have been identified for the CRISPR–*cas* system including gene editing, evolutionary and phylogenetic studies, as well as genotyping for epidemiological investigations [19,25]. The degree of variability in the CRISPR–*cas* system is a useful marker for species diversity and evolution. This approach has been applied to *Yersinia*, *Salmonella* and *E. coli* in the Enterobacteriaceae family for phylogenetic, evolutionary and virulence-related analysis [25–27]. Since spacers are added sequentially at the leader proximal end and a given spacer is rarely acquired twice or duplicated, hierarchical relationships can be constructed between strains. Consequently, spacer arrays can be used as alternative targets for molecular subtyping and may offer higher strain resolution than MLST and PFGE. Subtyping protocols based on CRISPR–*cas* systems have been proposed for *Salmonella* and these have included combined analysis with multi-virulence-locus sequence typing and PFGE [28–30]. In contrast, CRISPR typing cannot be used for all *E. coli* strains as it is absent from the extra intestinal phylogenetic group B2 but can be used for specific identification of enterohemorrhagic and Shiga toxin producing *E. coli* serotypes [27,31,32]. This could be linked to the absence of evidence for the type I-E CRISPR–*cas* system having a role in adaptive immunity in *E. coli*, and instead is proposed to be involved in different functions such as the regulation of endogenous gene expression and possible links to virulence [33,34].

Joseph *et al.* [35] were the first to report a CRISPR spacer array in *Cronobacter* when undertaking a genus-wide whole-genome comparative study. The specific region was located in all *Cronobacter* strains except one, *C. sakazakii* 680 (ST8). The genome of the *C. dublinensis* strain 582 also showed the presence of two additional clusters of CRISPR spacer arrays. However no detailed analysis of these arrays was undertaken.

Genotyping using MLST has proven to be highly informative and led to the recognition of the clonal lineage *C. sakazakii* ST4 as the pathovar associated with fatal neonatal meningitis infections. However, neither MLST nor PFGE are able to discriminate between unrelated strains within a clonal lineage as occurs in clinical sources [13]. Therefore it is necessary to investigate more discriminatory DNA sequence-based methods, such as CRISPR–*cas* loci profiling. Such analysis may also provide additional understanding of the diversity of the species and potential virulence mechanisms.

The first objective of this study was to describe the CRISPR–*cas* loci of *C. sakazakii* and thereafter to investigate its variation within clonal groups. In total the genomes for 70 *C. sakazakii* isolates were chosen for detailed CRISPR–*cas* loci analysis. These represented the four major *C. sakazakii* pathovars: ST4, ST12, CC1 and ST8 [6]. They were also chosen to enable the comparison of variation between related and unrelated strains. These genomes are available for independent analysis using the Bacterial Isolate Genome Sequence database (BIGSdb)-supported *Cronobacter* PubMLST open access database [6,7].

Materials & methods

• Bacterial strains

For the *in silico* analysis of the CRISPR–*cas* loci, a total of 70 whole-genome-sequenced isolates were chosen as representatives of the major pathovars for the neonatal pathogen *C. sakazakii*; ST4, ST12, CC1 and ST8. They were geographically dispersed over 10 countries and temporally spread over 64 years (1950–2014) **(Supplementary Table 1)**. Additional metadata can be obtained from the open access *Cronobacter* PubMLST database [7]. In order to investigate the potential uses of *Cronobacter* CRISPR spacer array profiling for epidemiological purposes, 20 isolates from a 6-month outbreak on an NICU and 26 environmental isolates from manufacturing plants in three US states collected within a 6-week period were included in the selected isolate cohort. For genus-wide phylogenetic analysis of *cas1* and *cas3*, an additional 130 genomes were accessed from the *Cronobacter* PubMLST database.

• DNA sequences

Whole-genome DNA sequences collated at were investigated using the Cronobacter PubMLST database [7]. *In silico* analysis of the *cas* genes was carried out using search options, such as BLAST, on the *Cronobacter* PubMLST portal [36].

• DNA annotation & visualization tools

The PROKKA annotation tool uses the CRISPR recognition tool to identify CRISPR–*cas* loci [37]. The *cas* genes and CRISPR spacer arrays were extracted from the corresponding genome assemblies in the *Cronobacter* PubMLST database. These were mainly draft genomes and therefore where the full sequence of one or more *cas* genes and/or CRISPR spacer arrays could not be determined (due to beginning and end of contigs), the entire isolate was removed from analysis. Bacterial DNA sequences were investigated using the genome browser and annotation tool Artemis [38]. On occasions, spacer sequences present within a clonal lineage can differ by one single nucleotide polymorphism (SNP) or insertion/deletion (INDEL). It is unknown whether these differences have an impact on CRISPR–*cas* immunity system. These spacer sequences were therefore considered to be the same but referred to as harboring a SNP or INDEL.

• Phylogenetic analysis

The *cas1* and *cas3* genes were aligned and phylogenetically analyzed in MEGA version 5.2 using the ClustalW algorithm [39] set to default parameters settings. The phylogenetic trees were generated using the Maximum Likelihood method based on the Tamura–Nei model with the additional parameters set to default settings. All phylogenetic trees are drawn to scale with branch lengths measured in the number of substitutions per site.

Results

• Diversity of *C. sakazakii* CRISPR–*cas* loci

All *C. sakazakii* isolates harbored the CRISPR–*cas* system subtype defined as *Escherichia coli* type I-E [19]. This comprises a *cas* gene cluster composed of *cas*3, *cse1*, *cse2*, *cas*7, *cas5*, *cas6e*, *cas*1 and *cas2*, with CRISPR spacer arrays either adjacent or distant to the *cas* genes cluster **(Figure 1)**.

Notable exceptions to the possession of the type I-E CRISPR–*cas* profile were four of the eight ST8 isolates which only possessed *cas*3, which is the signature gene for type I CRISPR–Cas systems [19]. It is plausible that *cas3* is the remnant following the partial loss of a *cas* region. Alternatively, *cas3* could be the result of horizontal gene transfer (HGT). In order to investigate this, a phylogenetic analysis of the *cas*3 genes within the *C. sakazakii* ST8 and the whole *Cronobacter* genus was undertaken. This revealed that the *cas3* genes were identical within the ST8 isolates regardless of the presence of the complete *cas* gene cluster **(Figure 2)**. Since *cas1* and *cas3* are both commonly used for CRISPR–*cas* classification [19], an additional *cas1* phylogenetic tree was constructed and was found to match that of *cas3* (data not shown). The latter included all ST8 isolates from a total of 208 strains across the *Cronobacter* genus **(Figure 2)**.

The *cas3* tree conformed to the expected phylogeny across the seven species in the genus **(Figure 2)**. It showed the close relatedness of *C. sakazakii* with *C. malonaticus*, and *C. muytjensii* with *C. dublinensis*, as well as the subdivision of *C. turicensis* into two clusters. There was high sequence variation within the *C. dublinensis* species, with only half of the strains possessing the *cas3* genes. Despite this diversity within this species, further investigation of their CRISPR arrays was regarded as outside the scope of this study due to their lack of clinical significance.

The *cas3* phylogeny within *C. sakazakii* matched the clonal lineages **(Figure 2)**. The only difference was the clustering of the *C. sakazakii* CC100 strains with the *C. universalis* strains away from the general *C. sakazakii* and *C. malonaticus* clades. All *C. sakazakii* ST8 *cas3* sequences followed the species phylogeny. The *cas3* of the four major pathovars did not cluster together, but were distributed across the species **(Figure 2)**.

• CRISPR spacer array analysis of the four *C. sakazakii* pathovars

Twelve CRISPR spacer arrays were identified across *C. sakazakii* ST4, CC1, ST12 and ST8 (2, 3, 3 and 4, respectively) and contained a total of 32 different direct repeat (DR) sequences and 154 different spacer sequences (primarily 29bp and 32bp long, respectively). Where appropriate, for clarity, the sequence type (ST) is used in the CRISPR spacer array designation, in other words, ST4-CRISPR2.

• CRISPR spacer array analysis of the *C. sakazakii* ST4 lineage

Twenty-five ST4 isolates were analyzed including 14 from an NICU *Cronobacter* outbreak in France [40]. Two groups of CRISPR spacer arrays

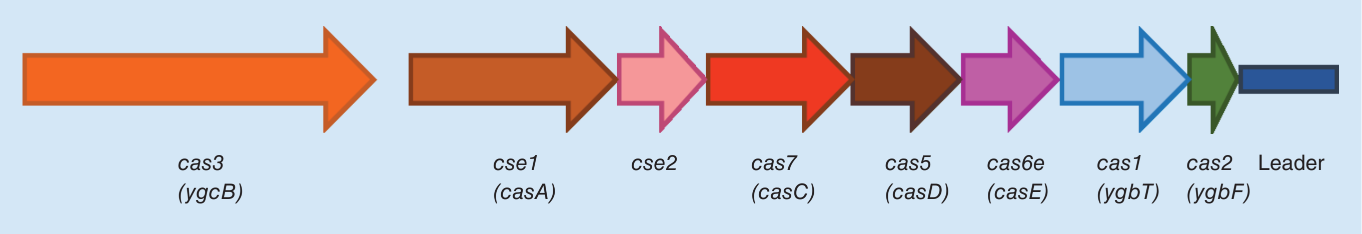


Figure 1. ***Cronobacter sakazakii*** **CRISPR–*Cas* operon architecture with former gene names are given in parenthesis.**

were found in all isolates (**Figure 3**). CRISPR1 was identical across all 25 isolates and was composed of eight DRs (four sequence variants) and seven unique spacers. The only sequence variation was in isolate 1886, which encoded a spacer containing a SNP or INDEL. CRISPR2 showed much greater variation across the 25 isolates with different combinations of spacer repeats generating a total of seven different CRISPR2 profiles. This CRISPR array contained 21–24 DRs (three sequence variants) and 23 unique spacers. Two isolates (6 and 1105) encoded a spacer containing a SNP or INDEL (different SNPs/INDELs corresponding to the same original sequence). The CRISPR2 profiles were not geographically or temporally associated as identical combinations were shared between isolates of different country, year, source and *vice versa*. For example, isolates 1537 and 1542, which were isolated from the same country and same year, show different CRISPR2 spacer combinations. In contrast, the earliest isolate 377 (from 1950, UK) had the same profile as isolate 1587 from 2000 (Israel).

With respect to the *C. sakazakii* ST4 isolates from the 1994 NICU outbreak, all had the same designated CRISPR2 profile. This profile was also found in a clinical strain (558) isolated in the Netherlands in 1983, and an isolate (1537) from a Germany milk powder factory in 2009.

• CRISPR spacer array analysis within the *C. sakazakii* CC1 lineage

Twenty-nine CC1 isolates were analyzed including 24 from environmental swabs from two different US States; two from Oregon and 22 from Wisconsin. Three CRISPR spacer arrays were found in total (**Figure 4**). CRISPR1 and CRISPR2 were found in all 29 isolates. CRISPR3 was found in seven isolates only, and was present in only two out of the 24 US environmental isolates.

CRISPR1 displayed an array of 31 DRs (11 sequence variants) and 30 unique spacers. All CRISPR1 arrays were identical across the 29 isolates apart from isolate 2064 which contained a different 29th DR sequence and isolate 716 (NICU outbreak strain), which expressed a spacer containing a SNP or INDEL. CRISPR2 displayed an array of 27–28 DRs (six sequence variants) and 26 unique and one predefined spacers (which was previously defined in ST4-CRISPR2). All isolates shared the same CRISPR2 array except for the two US manufacturing plant isolates (CFSAN022298 and CFSAN022299) from the USA Oregon state which lacked the 20th DR and spacer. The CRISPR3 of all *C. sakazakii* CC1 strains contained 11–13 DRs (two sequence variants) and 12 unique spacers. This CRISPR array was absent from the 22 US factories' isolates from Wisconsin. The only variation in CRISPR3 was found in the same two US manufacturing plant isolates as above from the US Oregon State which lacked the eighth and tenth DR and spacer sequences.

• CRISPR spacer array analysis within the *C. sakazakii* ST12 lineage

Eight ST12 isolates were analyzed including five from the French NICU outbreak (**Figure 5**). Three CRISPR arrays were identified, which were harbored by all eight isolates. The ST12 lineage showed highly conserved CRISPR array profiles compared with those within the ST4 and CC1. CRISPR1 displayed an array of 11–12 DRs (two sequence variants) and 11 unique spacers. All isolates shared the same CRISPR1 array except for two isolates (E764 and 1108) which lacked the third DR and spacer. CRISPR2 was identical across all eight isolates and displayed an array of 9 DRs (three sequence variants) and a total of eight spacers; seven unique and one predefined spacers (which was previously defined in ST4-CRISPR2). CRISPR3 displayed an array of seven to eight DRs (two sequence variants) and a total of seven spacers (five unique and two predefined spacers), which had previously defined in ST8-CRISPR3 and ST1-CRISPR3. All isolates within the *C. sakazakii* ST12 lineage shared the same

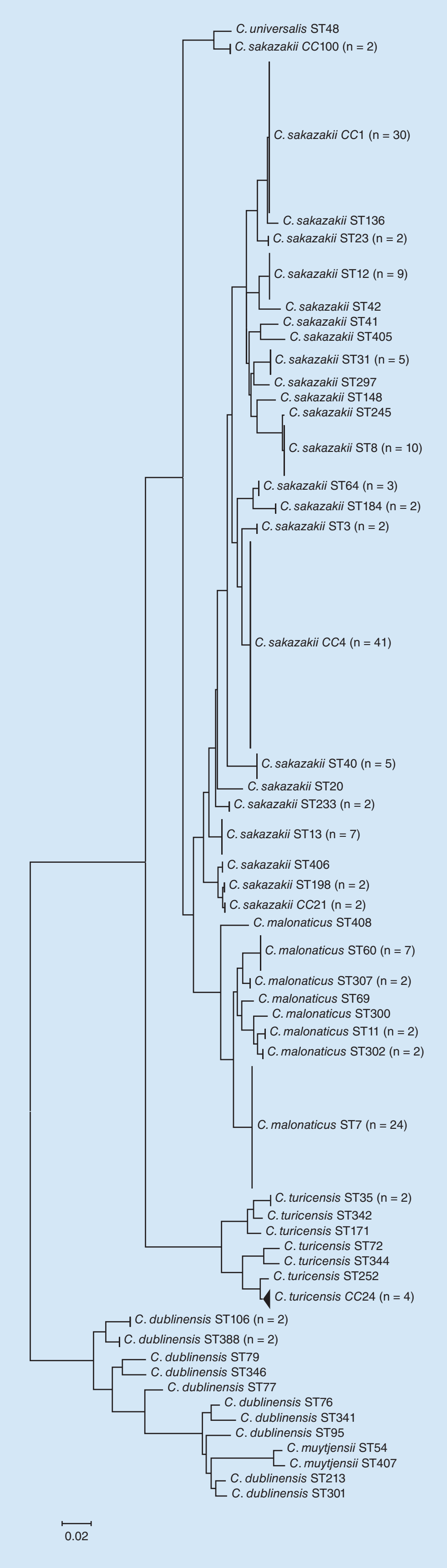

C. universalis ST48
C. sakazakii CC100 (n = 2)
C. sakazakii CC1 (n = 30)
C. sakazakii ST136
C. sakazakii ST23 (n = 2)
C. sakazakii ST12 (n = 9)
C. sakazakii ST42
C. sakazakii ST41
C. sakazakii ST405
C. sakazakii ST31 (n = 5)
C. sakazakii ST297
C. sakazakii ST148
C. sakazakii ST245
C. sakazakii ST8 (n = 10)
C. sakazakii ST64 (n = 3)
C. sakazakii ST184 (n = 2)
C. sakazakii ST3 (n = 2)
C. sakazakii CC4 (n = 41)
C. sakazakii ST40 (n = 5)
C. sakazakii ST20
C. sakazakii ST233 (n = 2)
C. sakazakii ST13 (n = 7)
C. sakazakii ST406
C. sakazakii ST198 (n = 2)
C. sakazakii CC21 (n = 2)
C. malonaticus ST408
C. malonaticus ST60 (n = 7)
C. malonaticus ST307 (n = 2)
C. malonaticus ST69
C. malonaticus ST300
C. malonaticus ST11 (n = 2)
C. malonaticus ST302 (n = 2)
C. malonaticus ST7 (n = 24)
C. turicensis ST35 (n = 2)
C. turicensis ST342
C. turicensis ST171
C. turicensis ST72
C. turicensis ST344
C. turicensis ST252
C. turicensis CC24 (n = 4)
C. dublinensis ST106 (n = 2)
C. dublinensis ST388 (n = 2)
C. dublinensis ST79
C. dublinensis ST346
C. dublinensis ST77
C. dublinensis ST76
C. dublinensis ST341
C. dublinensis ST95
C. muytjensii ST54
C. muytjensii ST407
C. dublinensis ST213
C. dublinensis ST301
0.02

Figure 2. ***Cronobacter*** **genus** ***cas3*** **phylogenetic tree.** Isolate sequences were compressed into sequence types and clonal complexes. The length of the vertical line represents the number of isolates within that sequence type or clonal complex. DNA sequences were aligned in MEGA version 5.2 using the ClustalW algorithm. The phylogenetic trees were generated using the Maximum Likelihood method. The tree was drawn to scale with branch length measured in the number of substitutions per site.

CRISPR3 array apart from two isolates (E764 and 1108) as for CRISPR1, which had an additional DR and spacer in the last position of the array.

• CRISPR spacer array analysis within the *C. sakazakii* ST8 lineage

The genomes of eight ST8 isolates were analyzed and four different CRISPR spacer arrays were identified (**Figure 6**). CRISPR1 displayed an array of 15–19 DRs (six sequence variants) and 18 unique spacers, and was only found in four out of eight isolates: 5, 513, 1888 and ES35. CRISPR1 showed variation with three different array combinations across these four isolates. Two isolates were identical (5 and 1888) and lacked three DRs and spacers at the 7th, 11th and 12th positions, and one isolate (ES35) lacked the 11th–14th DR and spacer sequences. CRISPR2 was present in all ST8 isolates and displayed an array of 8–13 DRs (five sequence variants) and 13 unique spacers (including a spacer expressing a SNP or INDEL). Variations in CRISPR2 arrays divided the isolates into two groups of four isolates; 680, NBRC102416^{T}, 2048, 1906 and 513, 5, 1888, ES35, respectively. The first group shared the same array except for one isolate (1906) which had an additional DR and spacer sequence in first position on the array. The second group shared the 2nd–5th and 7th, 8th and 13th. DRs and spacers with the first group of isolates and had an additional five to six DRs and spacers; one additional at the sixth position and four additional at position 9–12. Isolates 5 and 1888 lacked the DR and spacer at the 12th position on the array. Isolate 5 appears to lack the first few DRs and spacers of the array but this could be caused by the assembly of this genome whereby the CRISPR array was at the start of a contig. Hence the exact profile of the array cannot be determined for this isolate. CRISPR3 was identical across all eight isolates and displayed an array of five DRs (two sequence variants) and four unique spacers which includes a spacer expressing an SNP or INDEL. The only difference is the first spacer of the array showing four isolates harboring the sequence as originally defined and four isolates harboring the spacer with an SNP or INDEL, dividing the isolates into the same groups as mentioned in the CRISPR2 array. CRISPR4 is found in six out of eight isolates including isolates 513 and 1888. The CRISPR4 arrays are identical and displayed three DRs (one variant) and two unique spacers.

The presence/absence of CRISPR1 in four out of eight isolates and the variation between those

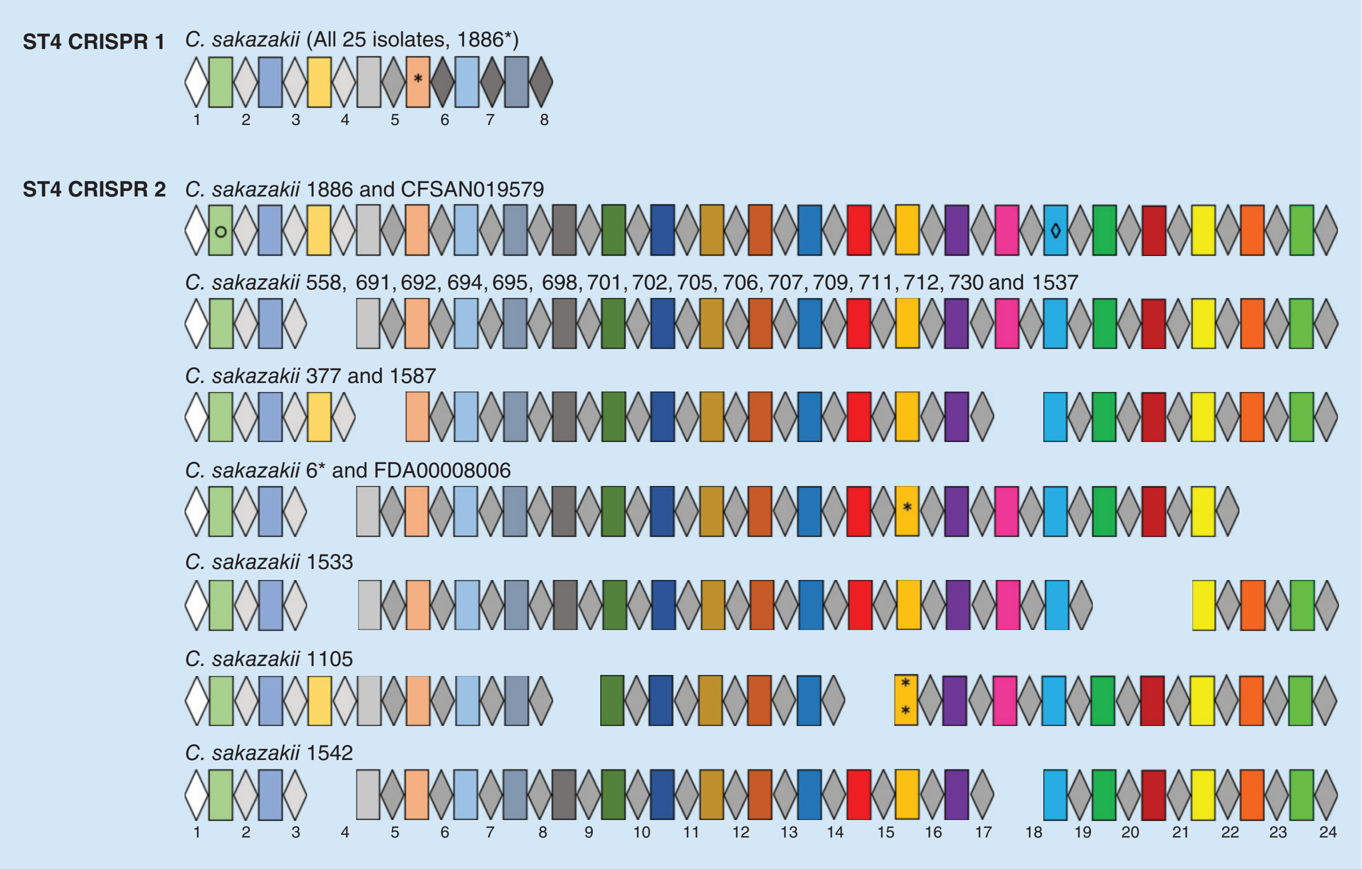

ST4 CRISPR 1
C. sakazakii (All 25 isolates, 1886*)
1
2
3
4
5
6
7
8
ST4 CRISPR 2
C. sakazakii 1886 and CFSAN019579
C. sakazakii 558, 691, 692, 694, 695, 698, 701, 702, 705, 706, 707, 709, 711, 712, 730 and 1537
C. sakazakii 377 and 1587
C. sakazakii 6* and FDA00008006
C. sakazakii 1533
C. sakazakii 1105
C. sakazakii 1542
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Figure 3. *Cronobacter sakazakii* ST4 CRISPR spacer array profiles. Grey scale lozenges correspond to direct repeats (DR) and colored boxes correspond to interspersed spacers. Lozenges with the same shading correspond to identical DR sequences within the array. Differently colored boxes correspond to unique spacers within the array. Gaps represent the absence of a spacer and its corresponding DR. Presence and number of asterisks in a spacer (*) indicate the presence and number of SNP(s) or INDEL(s) within the spacer. Spacers shared across different CRISPR arrays are indicated by different shapes; (O) ST4-CRISPR2 and CC1-CRISPR2, (◊) ST4-CRISPR and ST12-CRISPR2, (Δ) CC1-CRISPR3 and ST12-CRSPR3, and (□) ST12-CRISPR3 and ST8-CRISPR3.

same isolates in CRISPR2 and CRISPR3 correlate with presence/absence of the *cas* genes as given above; isolates 680, NBRC102416^T, 2048 and 1906 lack all of the *cas*-genes apart from *cas*3.

• Interclonal lineage similarities

Similarities between clonal lineages were identified, specifically in the DR sequences. These were used to enumerate the CRISPR spacer arrays within a lineage. CRISPR1 spacer arrays present in all four lineages mainly shared one common DR sequence, which was found in multiple copies throughout the array. Similarly, CRISPR2 arrays shared two repeating DR sequences across all four lineages. CRISPR3 arrays in lineages CC1, ST12 and ST8 harbor the same repeating pattern of DR sequences. CRISPR4 found in the ST8 lineage is unique and harbors a single DR variant.

Discussion

MLST has become a frequently used method for genotyping *Cronobacter* isolates, especially as it both speciates and indicates the pathovar or clonal lineage. However, neither MLST nor PFGE are able to discriminate strains within highly clonal groups as it occurs in *Cronobacter*. For example, the frequent isolation of unrelated *C. sakazakii* ST4 strains from clinical cases, which were also indistinguishable by PFGE [13]. Hence for epidemiological purposes, it is necessary to investigate more discriminatory methods. Given the increasing availability, lowering costs and application of NGS tools, there is an increasing trend for genome-based genotyping methods, such as CRISPR-array profiling. Due to the clinical significance of *C. sakazakii* this species was the focus of study.

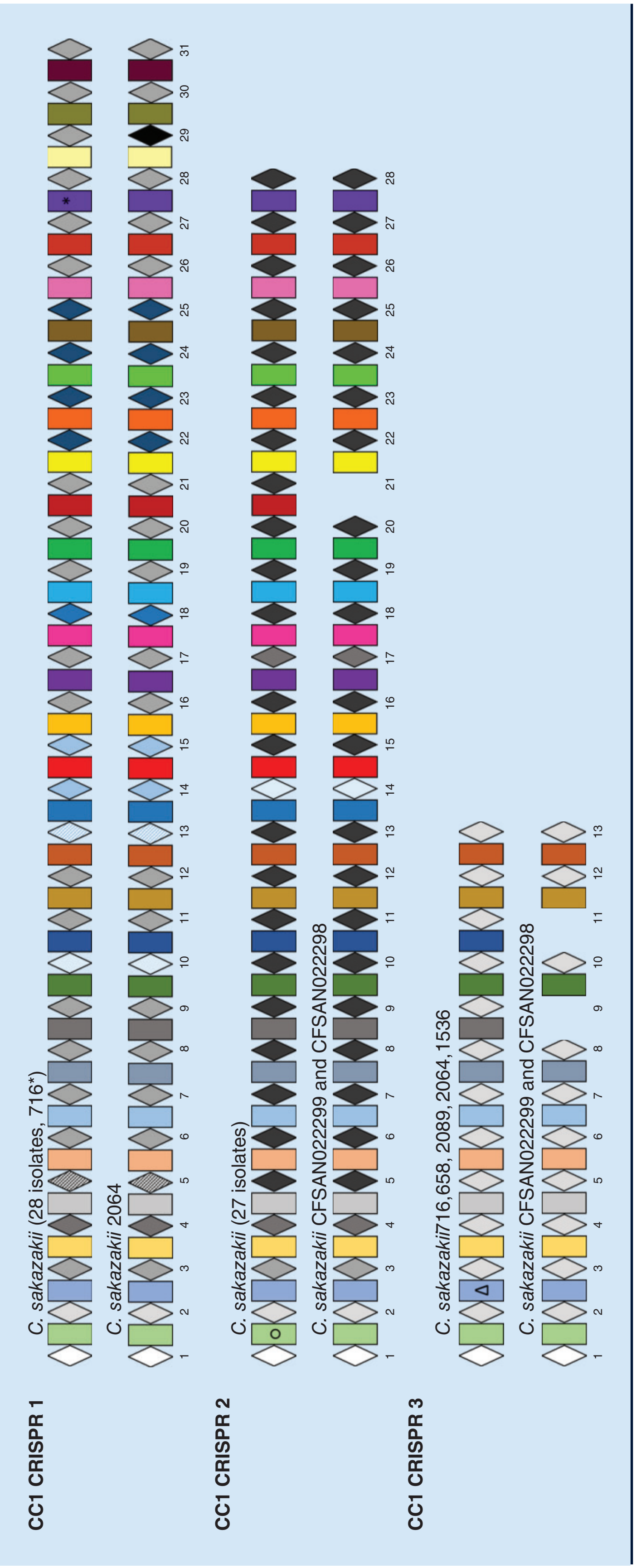

CC1 CRISPR 1
C. sakazakii (28 isolates, 716*)
C. sakazakii 2064
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
CC1 CRISPR 2
C. sakazakii (27 isolates)
C. sakazakii CFSAN022299 and CFSAN022298
CC1 CRISPR 3
C. sakazakii716,658, 2089, 2064,1536
C. sakazakii CFSAN022299 and CFSAN022298

Figure 4. *Cronobacter sakazakii* CC1 CRISPR spacer array profiles. Description and footnotes as per Figure 3.

Seventy whole-genome sequenced *C. sakazakii* isolates were selected for detailed CRISPR–*cas* loci profiling. These had been chosen as representatives of the four clinically significant *C. sakazakii* pathovars (ST4, ST12, CC1 and ST8), being also widely temporally (64 years) and globally (10 countries) distributed in their origin [6].

CRISPR–*cas* systems are a known adaptive immune mechanism in Eubacteria and Archaea [19,25]. These are generally regarded as providing immunity against foreign genetic elements such as phages and plasmids by targeting nucleic acid in a sequence specific manner. However recently it was proposed that the CRISPR–*cas* system in *E. coli* was not associated with adaptive immunity but instead with various functions such as DNA repair, regulation of gene expression and virulence [31]. Due to their conserved nature, CRISPR–*cas* loci profiling could be used for evolutionary analysis [19]. Since strains from the same environment can acquire whole CRISPR loci, such phylogenetic analysis can reveal the occurrence of HGT [41].

Previously *cas1* has been used for phylogenetic analysis of the various CRISPR–*cas* systems across a wide range of bacteria [19]. It is normally used due to its reported highly conserved distribution. However, in our analysis, four isolates of the eight ST8s lacked *cas1* and all other *cas* genes except the *cas3* gene **(Figure 1)**. This agreed with the specific lack of *cas* genes in *C. sakazakii* ST8 strain 680 as previously reported [35]. The possibility that the *cas3* in *C. sakazakii* ST8 was acquired through HGT was considered by constructing the *cas3* gene phylogeny based on 200 whole genomes from all seven *Cronobacter* species in the genus. It was found to support the previously published whole-genome phylogeny of the *Cronobacter* genus [6,35]. Similarly, the *cas1* phylogeny for all strains harboring the gene, also matched the whole-genome phylogeny (data not shown).

Bacterial strains from the same geographical and temporal region should acquire the same spacers due to localized exposure to phages and plasmids. CRISPR spacer array content has previously been strongly associated with sequence-based phylogeny and hence could be used as a rapid lineage-based detection method and as a discriminatory tool for epidemiological purposes [19,25,42].

The variability in CRISPR–*cas* loci was investigated further within *C. sakazakii,* and in particular the spacer region of strains selected for their clinical relevance and clonal lineage. The acquisition of spacers based on a CRISPR–*cas* adaptive immunity function would be expected to show geographic- and temporal-specific variations. However our analysis demonstrates that the *C. sakazakii* CRISPR spacer array profiles were neither geographically nor temporally dispersed. Instead, the identified CRISPR arrays and profiles were

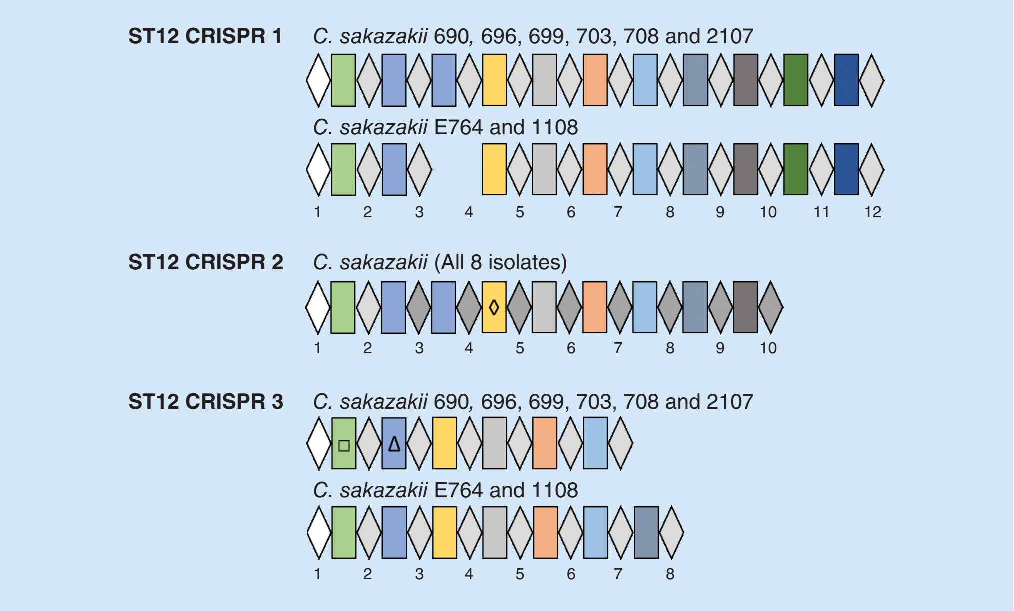

ST12 CRISPR 1
C. sakazakii 690, 696, 699, 703, 708 and 2107
C. sakazakii E764 and 1108
1 2 3 4 5 6 7 8 9 10 11 12
ST12 CRISPR 2
C. sakazakii (All 8 isolates)
1 2 3 4 5 6 7 8 9 10
ST12 CRISPR 3
C. sakazakii 690, 696, 699, 703, 708 and 2107
C. sakazakii E764 and 1108
1 2 3 4 5 6 7 8

Figure 5. *Cronobacter sakazakii* ST12 CRISPR spacer array profiles. Description and footnotes as per **Figure 3.**

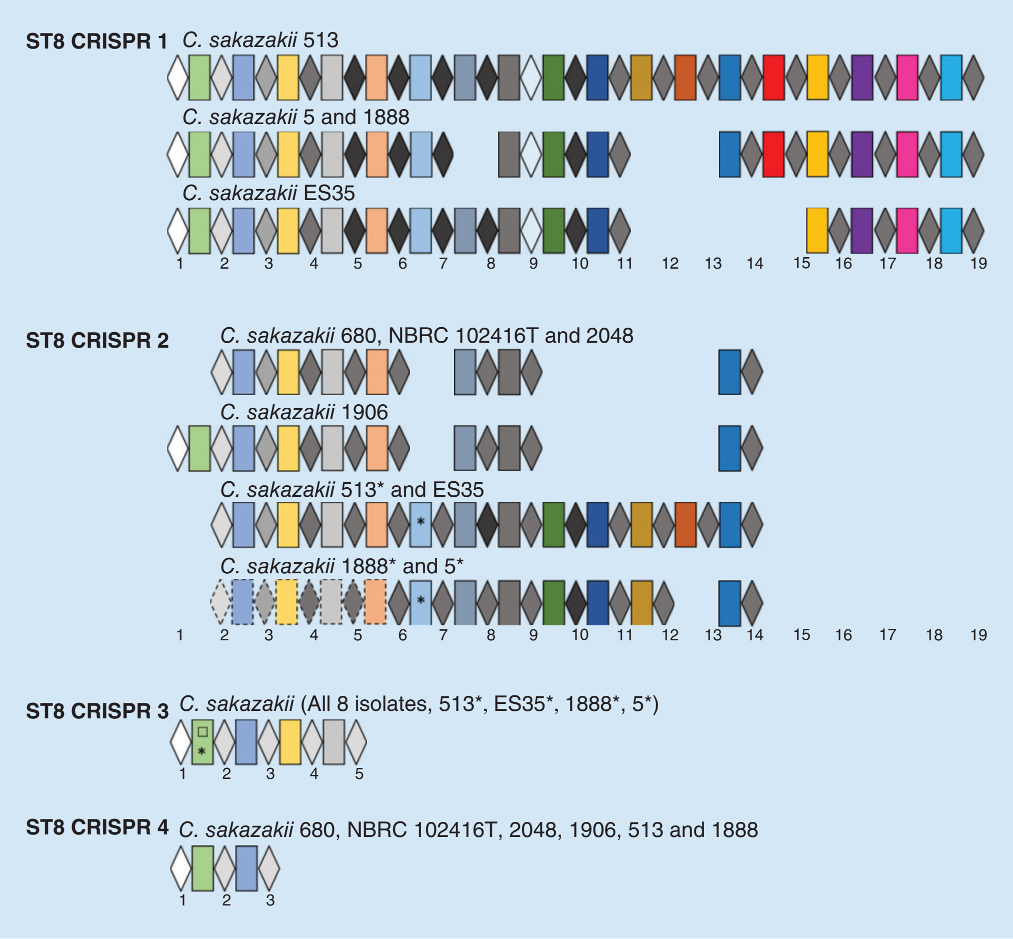


Figure 6. ***Cronobacter sakazakii*** **ST8 CRISPR spacer array profiles.** Description and footnotes as per **Figure 3**. Note: The dotted outlining of the first few DRs and spacers correspond to their absence in *C. sakazakii* isolate 5 due to the start of a contig.
DR: Direct repeat.

specific for each clonal lineage **(Figures 3 & 4–6)**. Furthermore, CRISPR array profiling was able to distinguish strains within a clonal lineage. For example, of the two CRISPR arrays identified in the ST4 lineage, CRISPR1 was highly conserved across the 25 isolates, while CRISPR2 showed seven variants based on the order and presence/absence of DR and spacer units **(Figure 3)**. Isolates of the ST8 lineage also showed high variability, especially considering the small (n = 8) number of isolates analyzed. Whereas CC1 and ST12 showed more conserved CRISPR array patterns **(Figures 3 & 4–6)**.

The four out of eight ST8 isolates lacking the majority of the *cas* genes (except *cas3*) would suggest that the CRISPR arrays harbored on the genome are no longer active and hence cannot process or utilize their transcripts. This was previously identified in *Pseudomonas aeruginosa* where certain lineages harbored degenerative systems which contained the CRISPR array but lacked the *cas* genes [42]. Since no new spacer acquisition was observed in the *Cronobacter* isolates harboring all of the *cas* genes, this could suggest that active CRISPR–*cas* systems are not necessarily linked to new spacer acquisitions but to other functions. This might explain the sudden loss of spacers in CRISPR arrays instead of a gradual CRISPR array change due to environmental exposure, as has also been proposed for *E. coli* [31].

In order to assess whether profiling of the CRISPR spacer array region could be used for epidemiological purposes, multiple isolates from the same geographic location within a short time period were included in the 70 genome cohort. The CRISPR spacer array profiles of the 14 *C. sakazakii* ST4 isolates from a NICU outbreak in France (1994) were indistinguishable from each other (**Figure 3**). However the same profile was also found in two unrelated strains (558 &1537) from different countries (Netherlands and Germany) and decades (1983 and 2009), respectively. Therefore specific *Cronobacter* CRISPR spacer array profiles within an outbreak can be used as an epidemiological tool, however such specific profiles cannot necessarily be used globally. This limitation is common with other DNA-profiling methods, for example the use of PFGE analysis for *Salmonella* serovar Enteritidis isolates [12].

Twenty-six environmental isolates from three US States had been collected within a 6-week period. *In silico* MLST analysis revealed that two isolates were belonging to the *C. sakazakii* ST4 lineage and the remaining 24 were *C. sakazakii* CC1. Other studies have also reported the common occurrence of *C. sakazakii*, and in particular ST4 and CC1 in up to 25% of isolates from milk powder-manufacturing plants [6,43,44]. It should be noted that the *C. sakazakii* ST4 pathovar is associated with neonatal meningitis through the ingestion of contaminated reconstituted infant formula and therefore its presence in manufacturing plants is undesirable [6,8,9].

Based on their CRISPR array profiles, the two *C. sakazakii* CC1 strains isolated in Oregon were distinguishable from the 22 from Wisconsin. The Oregon strains lacked DR and spacer sequences within two of the three CRISPR arrays present in this lineage (**Figure 4**). Since the *C. sakazakii* CC1 isolates within each manufacturing plant were indistinguishable from each other, this probably indicates colonization of these environments by unique strains. This supports the application of genetic profiling of *Cronobacter* strains allowing microbial source tracing of isolates from clinical cases through food source, to production site.

This is the first study to identify the widespread phylogenetic distribution of CRISPR spacer arrays within the four major pathovars of *C. sakazakii*. This analysis has demonstrated that the CRISPR array variability can provide greater power of differentiation for genotyping within clonal lineages. Future studies will expand the *in silico* CRISPR–*cas* loci profiling across the whole *Cronobacter* genus for the wider study of bacterial population diversity, in particular the adult pathovar *C. malonaticus* ST7 and the development of laboratory-based PCR typing protocols.

Conclusion & future perspective

This is the first study to identify the phylogenetic distribution of CRISPR–cas loci within the major pathovars of *C. sakazakii*; ST1, 4, 8 and 12. All strains encoded the type I-E subtype CRISPR–cas system with a total of 12 different CRISPR spacer arrays, which were not geographically or temporally associated. Each *C. sakazakii* clonal lineage could be subdivided into 2–7 CRISPR spacer array profiles, according to the direct repeat and spacer sequences. All. *C. sakazakii* neonatal meningitis pathovars (ST4) strains from a NICU outbreak had indistinguishable CRISPR2 array profiles. This study demonstrated the greater discriminatory power of CRISPR spacer array profiling compared with MLST, which will be of use in source attribution during *Cronobacter* outbreak investigations. The future use of genome-based strain differentiation will become more established over the next few years due to more affordable sequencing costs, improved availability of genomic tools and ease of use of downstream analysis. CRISPR–cas profiles are increasingly being used as global epidemiological tracking tools, and collated into curated open access databases. This trend is expected to continue as genomic profiling becomes the gold-standard for high-level differentiation of clonal organisms, replacing the current, often PCR-based, genotyping methods.

Acknowledgements

The authors gratefully thank Nottingham Trent University for their financial support. This publication made use of the Cronobacter Multi Locus Sequence Typing website (http://pubmlst.org/cronobacter/) developed by K Jolley and sited at the University of Oxford (Jolley & Maiden [45]*). The development of this site has been funded by the Wellcome Trust. The authors thank our numerous collaborators especially those involved who have submitted sequences to the Cronobacter PubMLST database.*

Financial & competing interests disclosure

This study was funded by The Nottingham Trent University. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethics statement

All clinical data are taken from previous publications associated with the sequenced bacterial strains.

EXECUTIVE SUMMARY

- This is the first study to identify the phylogenetic distribution of CRISPR–*cas* loci within the major pathovars of *C. sakazakii*; ST1, 4, 8 and 12.
- This study is needed since conventional typing methods cannot distinguish unrelated *Cronobacter* strains due to the highly level of clonality. DNA-sequence based methods such as CRISPR–*cas* loci profiling are more likely to be of use for epidemiological investigations of *Cronobacter* outbreaks and source attribution.
- This study used 70 whole-genome-sequenced *C. sakazakii* strains which were widely distributed temporally (64 years), geographically (10 countries), and origin of source (clinical, infant formula, weaning food and environmental).
- All strains encoded the type I-E subtype CRISPR–*cas* system with a total of 12 different CRISPR spacer arrays, which were not geographically or temporally associated.
- Each *C. sakazakii* clonal lineage could be subdivided into 2–7 CRISPR spacer array profiles, according to the direct repeat and spacer sequences. *C. sakazakii* ST8 contained four CRISPR arrays with up to four variants, and lacked all *cas* genes except *cas3*.
- All ST4 strains from a NICU outbreak had indistinguishable CRISPR2 array profiles. Similarly, the CRISPR arrays for 22 *C. sakazakii* CC1 strains from a US manufacturing plant were indistinguishable. This study demonstrated the greater discriminatory power of CRISPR spacer array profiling compared with multilocus sequence typing, which will be of use in source attribution during *Cronobacter* outbreak investigations.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Holy O, Forsythe SJ. *Cronobacter* species as emerging causes of healthcare-associated infection. *J. Hosp. Infect.* 86, 169–177 (2014).

2 Almajed FS, Forsythe SJ. *Cronobacter sakazakii* clinical isolates overcome host barriers and evade the immune response. *Microb. Pathog.* 90, 55–63 (2016).

3 Alzahrani H, Winter J, Boocock D *et al.* Characterisation of outer membrane vesicles from a neonatal meningitic strain of Cronobacter sakazakii. *FEMS Microbiol. Lett.* 362(12), fnv085 (2015).

4 Baldwin A, Loughlin M, Caubilla-Barron J *et al.* Multilocus sequence typing of *Cronobacter sakazakii* and *Cronobacter malonaticus* reveals stable clonal structures with clinical significance which do not correlate with biotypes. *BMC Microbiol.* 9, 223 (2009).

5 Joseph S, Sonbol H, Hariri S *et al.* Diversity of the *Cronobacter* genus as revealed by multilocus sequence typing. *J. Clin. Microbiol.* 50, 3031–3039 (2012).

• **First description of multilocus sequence typing applied across the *Cronobacter* genus.**

6 Forsythe SJ, Dickins B, Jolley KA. *Cronobacter*, the emergent bacterial pathogen *Enterobacter sakazakii* comes of age; MLST and whole genome sequence analysis. *BMC Genomics* 15, 1121 (2014).

•• **Most up-to-date coverage of *Cronobacter* diversity, pathogenicity and clinical relevance based on the DNA sequence analysis of >1000 strains and >100 whole-genome sequences.**

7 Cronobacter PubMLST Database. http://pubmlst.org/cronobacter/

8 Hariri S, Joseph S, Forsythe SJ. *Cronobacter sakazakii* ST4 strains and neonatal meningitis, US. *Emerg. Inf. Dis.* 19, 175–177 (2013).

9 Joseph S, Forsythe S. Predominance of *Cronobacter sakazakii* sequence type 4 in neonatal infections. *Emerg. Inf. Dis.* 17, 1713–1715 (2011).

•• **First recognition of *C. sakazakii* pathovar ST4.**

10 Masood N, Moore K, Farbos A *et al.* Genomic dissection of the 1994 *Cronobacter sakazakii* outbreak in a French neonatal intensive care unit. *BMC Genomics* 16, 750 (2015).

11 Ogrodzki P, Forsythe S. Capsular profiling of the *Cronobacter* genus and the association of specific *Cronobacter sakazakii* and *C. malonaticus* capsule types with neonatal meningitis and necrotizing enterocolitis. *BMC Genomics* 16, 758 (2015).

12 Centers for Disease Control. Investigation update: multistate outbreak of human *Salmonella* Enteritidis infection associated with shell eggs. www.cdc.gov/salmonella/enteritidis/

13 Alsonosi A, Hariri S, Kajsík M *et al.* The speciation and genotyping of *Cronobacter* isolates from hospitalised patients. *Eur. J. Clin. Microbiol. Infect. Dis.* 34, 1979–1988 (2015).

14 Sun Y, Wang M, Liu H *et al.* Development of an O-antigen serotyping scheme for *Cronobacter sakazakii*. *Appl. Environ. Microbiol.* 77, 2209–2214 (2011).

15 Blažková M, Javůrková B, Vlach J *et al.* Diversity of O-antigen designations within the genus *Cronobacter*: from disorder to order. *Appl. Env. Microbiol.* 81, 5574–5582 (2015).

16 Arbatsky NP, Wang M, Shashkov AS *et al.* Structure of the O-polysaccharide of *Cronobacter sakazakii* O1 containing 3-(N-acetyl-l-alanyl)amino-3,6-dideoxy-d-glucose. *Carbohydr. Res.* 345, 2095–2098 (2010).

17 Czerwicka M, Forsythe SJ, Bychowska A. Structure of the O-polysaccharide isolated from *Cronobacter sakazakii* 767. *Carbohydr. Res.* 345, 908–913 (2010).

18 Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8, 172 (2007).

19 Makarova KS, Wolf YI, Alkhnbashi OS. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* 13, 722–736 (2015).

20 Shariat N, Dudley EG. CRISPRs: molecular signatures used for pathogen subtyping. *Appl. Environ. Microbiol.* 80, 430–439 (2014).

21 Barrangou R, Fremaux C, Deveau H *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712 (2007).

22 Zegans ME, Wagner JC, Cady KC *et al.* Interaction between bacteriophage DMS3 and host CRISPR region inhibits group

behaviours of *Pseudomonas aeruginosa*. *J. Bacteriol.* 191, 210–219 (2009).

23 Garneau JE, Dupuis ME, Villion M *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468, 67–71 (2010).

24 Chakraborty S, Snijders AP, Chakravorty R *et al.* Comparative network clustering of direct repeats (DRs) and *cas* genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria. *Mol. Phylogenet. Evol.* 56, 878–887 (2010).

25 Fricke W, Mammel M, McDermott P *et al.* Comparative genomics of 28 *Salmonella enterica* isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *J. Bacteriol.* 193, 3556–3368 (2011).

26 Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* 151, 653–663 (2005).

27 Yin S, Jensen MA, Bai J *et al.* The evolutionary divergence of Shiga toxin-producing *Escherichia coli* is reflected in CRISPR spacer composition. *Appl. Environ. Microbiol.* 79, 5710–5720 (2013).

28 Fabre L, Zhang J, Guigon G *et al.* CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS ONE* 7, e36995 (2012).

29 Shariat N, DiMarzio MJ, Yin S *et al.* The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of *Salmonella enterica* subsp. *enterica* serovar Enteritidis. *Food Microbiol.* 34, 164–173 (2013).

30 Shariat N, Kirchner MK, Sandt CH *et al.* Subtyping of *Salmonella enterica* serovar Newport outbreak isolates by CRISPR-MVLST and determination of the relationship between CRISPR-MVLST and PFGE results. *J. Clin. Microbiol.* 51, 2328–2336 (2013).

31 Touchon M, Charpentier S, Clermont O *et al.* CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J. Bacteriol.* 193, 2460–2467 (2011).

32 Delannoy S, Beutin L, Fach P. Use of clustered regularly interspaced short palindromic repeat sequence polymorphisms or specific detection of enterohemorrhagic *Escherichia coli* strains of serotypes O26:H11, O45:H2, O45:H2, O132:H2, O111:H8, O121:H19, O145:H28, and O157:H7 by real-time PCR. *J. Clin. Microbiol.* 50, 4035–4040 (2012).

33 Pougach K, Semenova E, Bogdanova E *et al.* Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol. Microbiol.* 77, 1367–1379 (2010).

34 Westra ER, Buckling A, Fineran PC. CRISPR–Cas systems: beyond adaptive immunity. *Nat. Rev. Microbiol.* 12, 317–326 (2014).

35 Joseph S, Desai P, Ji Y *et al.* Comparative analysis of genome sequences covering the seven *Cronobacter* species. *PLoS ONE* 7, e49455 (2012).

- **First comparative genomic analysis of all species in the *Cronobacter* genus, including first identification of CRISPR–*cas* loci.**

36 Cronobacter PubMLST portal. http://pubmlst.org/perl/bigsdb

37 Bland C, Ramsey TL, Sabree F *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered repetitively interspaced palindromic repeats. *BMC Bioinformatics* 8, 209 (2007).

38 Carver T, Harris SR, Berriman M *et al.* Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28, 464–469 (2012).

39 Kumar S, Tamura K, Jakobsen IB *et al.* MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17, 1244–1245 (2001).

40 Caubilla-Barron J, Hurrell E, Townsend S *et al.* Genotypic and phenotypic analysis of *Enterobacter sakazakii* strains from an outbreak resulting in fatalities in a neonatal intensive care unit in France. *J. Clin. Microbiol.* 45, 3979–3985 (2007).

- **First description of neonatal outbreak indicating the variation in pathogenicity of *Cronobacter* (former name *E. sakazakii*) isolates.**

41 Yang C, Li P, Su W *et al.* Polymorphism of CRISPR shows separated natural groupings of *Shigella* subtypes and evidence of horizontal transfer of CRISPR. *RNA Biol.* 12, 1109–1120 (2015).

42 Van Belkum A, Soriaga LB, LaFave MC *et al.* Phylogenetic distribution of CRISPR–Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. *mBio* 6, e01796–e01815 (2015).

43 Sonbol H, Joseph S, McAuley C *et al.* Multilocus sequence typing of *Cronobacter* spp. from powdered infant formula and milk powder production factories. *Intl Dairy J.* 30, 1–7 (2013).

44 Fei P, Man C, Lou B *et al.* Genotyping and source tracking of the *Cronobacter sakazakii* and *C. malonaticus* isolated from powdered infant formula and an infant formula production factory in China. *Appl. Env. Microbiol.* 81, 5430–5439 (2015).

45 Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform.* 11, 595 (2010).