GENOMIC PROFILING OF THE NEONATAL MENINGITIC *CRONOBACTER SAKAZAKII* CLONAL COMPLEX 4

NAQASH MASOOD

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of

Doctor of Philosophy

March 2015

ABSTRACT

The *Cronobacter sakazakii* clonal complex 4 (CC4) represents a dominant lineage of the genus *Cronobacter* in *Cronobacter* PubMLST. A strong association of *C. sakazakii* CC4 with neonatal infections especially neonatal meningitis has been established. The composition of *C. sakazakii* CC4 isolates in *Cronobacter* PubMLST indicates that at least 70% of these isolates were obtained either from clinical sources or infant formula. The dominant association of the *C. sakazakii* CC4 with neonatal infections especially meningitis and its predominant isolation from the infant formula and environment was intriguing which drove the rationale for the in-depth genomic analysis presented in the current PhD study. The whole genome phylogeny revealed that despite their geographical and temporal spread, the CC4 isolates cluster tightly with each other representing a clonal and a stable lineage within the *Cronobacter* genus. An exhaustive search of the sequenced genomes to identify virulence or environmental fitness associated traits indicated no significant difference between the virulence potential of *C. sakazakii* CC4 and *C. sakazakii* non-CC4. The interesting observation was the presence of two hypothetical proteins predominant in CC4 isolates, one of which was the homologue of an inner membrane protein. In addition, an hypothetical protein was noted to be largely absent from the *C. sakazakii* CC4 genomes. The O:2 was found to be the dominant serotype of CC4, however not exclusive to CC4. A giant adhesion associated gene was also noted predominantly in the *C. sakazakii* CC4 genomes. Single nucleotide polymorphism indicated low degree of sequence diversity within CC4 with average distance of 300-400 SNPs against the reference isolate. The subdivision of the low invasive CC4 isolates was intriguing, however no unique invasion associated traits were determined in a highly invasive CC4 isolate whereas one low invasive isolate indicated the presence of heavy metal resistance associated traits. The metal resistance assays could not differentiate the high and low invasive CC4 isolates. The case study of the 1994 French outbreak using genome sequenced data suggested powdered infant formula (PIF) to be the dominant, yet not the exclusive source of outbreak for the *C. sakazakii* isolates. The current PhD study was the first to explore the genomes of *C. sakazakii* CC4 revealing some interesting variations. Future studies are warranted to characterise hypothetical proteins predominant in CC4 to elucidate their significance in this clonal lineage. Furthermore, transcriptomics studies are warranted to find out any unique genes differentially expressed in the CC4 genomes under different stressful conditions.

DECLARATION

It is hereby certified that the experimental work and analysis embodied in this thesis is the original research carried out by the author, unless otherwise stated at the School of Science and Technology, Nottingham Trent University, UK. This work is the intellectual property of the author. You may copy up to 5% of this work for the private study or personal, non-commercial research. Any information used from this thesis should be fully cited.

Naqash Masood

ACKNOWLEDGEMENTS

First of all I would like to thank ALLAH ALMIGHTY for giving me the opportunity, courage, strength, determination and guidance to carry out this research.

I would like to convey a very special thanks to my Director of Studies Professor Stephen Forsythe for his continuous support and guidance throughout the course of this PhD study. He has been an excellent and a supportive supervisor, not only did he provide me research guidance but also helped develop my professional career by encouraging me to attend international conferences and training workshops.

I would also like to extend my gratefulness to the second supervisor Dr Alan McNally for his expert opinion, whenever need. His adept opinion helped me a great deal in achieving the goals of my PhD.

A special thanks to Dr Ben Dickins who assisted me on a number of occasions to help sort technical issues related to computing. I am also grateful to Professor Nadia Chuzhanova as she was always kind and supportive whenever I needed her help in statistical analysis.

I owe special thanks to my previous supervisors, Dr Gina Manning and Dr Muhammad Afzal Javed as their supervision gave me the foundation of a good researcher. They had been really kind to me and always supported me, whenever needed. I would also like to take this opportunity to thank all my school teachers, particularly Mr Shahid Qureshi at the Crescent School, who was always motivating and indeed developed my strong interest in Biology.

My heartfelt thanks to whole of my family; my sweet mum, my loving dad, my caring brothers and to my lovely sister Nadia. Their moral support, encouragement and motivation helped me achieve this important milestone of my life. A special thanks to my loving and caring wife Tania who has been understanding, supportive and patient during the course of my PhD study.

Many thanks to all my friends and colleagues at NTU and beyond who have been supportive and encouraging.

I would like to acknowledge the Vice Chancellor, Nottingham Trent University for funding my PhD study.

DEDICATION

I would like to dedicate this thesis to my beloved mother!

TABLE OF CONTENTS

PUBLICATIONS

Masood, N., Moore, K., Farbos, A., Hariri, S., Block, C., Paszkiewicz, K., Dickins, B., McNally, A. & Forsythe, S. (2015). Genomic dissection of the *Cronobacter sakazakii* outbreak in a French neonatal intensive care unit 1994, using Single Nucleotide Polymorphism. ***BMC Genomics,*** 16:750 doi 10.1186/s12864-015-1961-y.

Masood, N., Jackson, E., Moore, K., Farbos, A., Paszkiewicz, K., Dickins, B., McNally, A., & Forsythe, S. (2014). Draft genome sequence of "*Candidatus* Cronobacter colletis" NCTC 14934^T, a new species in the genus *Cronobacter*. ***ASM Genome Announcements***, 2(3) e00585-14.

Masood, N., Moore, K., Farbos, A., Hariri, S., Block, C., Paszkiewicz, K., Dickins, B., McNally, A., & Forsythe, S. (2013). Draft genome sequence of a meningitic isolate of *Cronobacter sakazakii* Clonal Complex 4, strain 8399. ***ASM Genome Announcements,*** 1(5) doi:pii: e00833-13. 10.1128/genomeA.00833-13.

Masood, N., Moore, K., Farbos, A., Hariri, S., Paszkiewicz, K., Dickins, B., McNally, A., & Forsythe, S. (2013). Draft genome sequences of *Cronobacter helveticus* LMG23732^T, *Cronobacter pulveris* LMG24059 and *Cronobacter zurichensis* LMG23730^T: three newly identified species in the genus *Cronobacter*. ***ASM Genome Announcements***, 1(5). doi:pii: e00783-13. 10.1128/genomeA.00783-13.

Masood, N., Moore, K., Farbos, A., Hariri, S., Paszkiewicz, K., Dickins, B., McNally, A., & Forsythe, S. (2013). Draft genome sequence of the earliest *Cronobacter sakazakii* sequence type 4 strain NCIMB 8272. ***ASM Genome Announcements***, 1(5). doi:pii: e00782-13. 10.1128/genomeA.00782-13.

Jackson, E., **Masood, N**., Ibrahim, K., Urvoy, N., Hariri, S., & Forsythe, S. (2015). *Siccibacter colletis* sp. nov., a new *Siccibacter* species isolated from plant material. ***International Journal of Systematic And Evolutionary Microbiology***, 65:1335-1341.

Jackson, E., Sonbol, H., **Masood, N**., & Forsythe, S. (2014). Genotypic and phenotypic characteristics of *Cronobacter* species, with particular attention to the newly reclassified species *C. helveticus*, *C. pulveris*, and *C. zurichensis*. ***Food Microbiology***, 44:226-235.

Joseph, S., Desai, P., Ji, Y., Hamby, S. E., **Masood, N**., Hariri, S., Sonbol, H., Cumming, C. A., Rico, A., Shih, S. M., Degoricja, L., Brzoska, P., Chuzhanova, N., McClelland, M., Furtado, M. R., & Forsythe, S. (2012). Comparative analysis of genome sequences covering the seven *Cronobacter* species ***PLOS ONE***, 7:e49455. 10.1128/JCM.00905-12.

Joseph, S., Hariri, S., **Masood, N**., & Forsythe, S. (2013). Sialic acid utilization by *Cronobacter sakazakii*. ***Microbial Informatics and Experimentation****,* 3:3.

Alkeskas, A. Odrodzki, P., Saad, M., **Masood, N**., Rhouma, N., Moore, K., Farbos, A., Paszkiewicz, K. & Forsythe, S. (2015). Colonisation of neonatal nasogastric feeding tubes by *Escherichia coli* K1". ***BMC Infectious Diseases,*** in print.

CONFERENCE PRESENTATIONS (POSTERS)

Naqash Masood, Alan Mcnally, Konrad Paszkiewicz, Karen Moore, Audrey Farbos, Stephen Forsythe. **2015.** Genomic analysis of *C. sakazakii* isolates from a French neonatal intensive care unit outbreak of 1994 revealed powdered infant formula to be the potential source of outbreak. Accepted for ***115th ASM General Meeting New Orleans, USA***.

Emily Jackson, Hana Sonbol, **Naqash Masood,** Stephen Forsythe. **2014**. Genotypic and phenotypic characteristics of *Cronobacter* species, with particular attention to the newly reclassified species *C. helveticus*, *C. pulveris*, and *C. zurichensis*. ***ASM Boston, USA***

Naqash Masood, Alan Mcnally, Konrad Paszkiewicz, Karen Moore, Audrey FarboS, Stephen Forsythe. **2013**. Whole genome comparative analysis revealed clustering of neonatal meningitic *C. sakazakii* ST4 strains. ***FEMS, Leipzig, Germany***

Naqash Masood and Stephen Forsythe. **2012**. Single Nucleotide Polymorphism revealed a clonal signature for the neonatal meningitic *Cronobacter sakazakii* ST4. ***5th Cold Spring Harbor Laboratory/Wellcome Trust Scientific conference on infectious Disease Genomics and Global Health, Cambridge, UK***

Naqash Masood and Stephen Forsythe. **2012**. Expanding the multilocus sequence typing of *Cronobacter sakazakii* to include virulence relatedness. ***ASM, San Francisco, USA***

Naqash Masood and Stephen Forsythe. **2012**. Multilocus sequence typing of *Cronobacter sakazakii*; ST4 and variation in *ompA* and *recN* sequences. ***SGM, Dublin, Republic of Ireland.***

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ABC	ATP binding cassette DNA Deoxyribonucleic acid
ACT	Artemis comparison tool DR Direct repeats
BBB	Blood brain barrier
BLAST	Basic local alignment search tool
BRIG	BLAST ring image generator FUF Follow up formula
CC	Clonal complex HBMEC Human brain microvascular epithelial cells
CDC	Centers for Disease Control and prevention
CHO	Chinese hamster ovary ITS Internal transcribed spacer
CNS	Central nervous system
CSF	Cerebrospinal fluid
DDH	DNA-DNA hybridization
f-AFLP	Amplified fragment length polymorphisms
FAO	Food and Agriculture Organization of the UN
GIT	Gastro intestinal tract
MLSA	Multilocus sequence analysis
MLST	Multilocus sequence typing
NCBI	National Centre for Biotechnology Information
NEC	Necrotising enterocolitis
NGS	Next generation sequencing
NICU	Neonatal intensive care unit
NTU	Nottingham Trent University
O-LPS	Oligo-lipopolysaccharide
OM	Outer membrane
ORF	Open reading frame
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
PIF	Powdered infant formula
SOLiD	Sequencing by oligonucleotide ligation and detection
ST	Sequence type
TSA	Trypticase soy agar
WHO	World Health Organization

1 CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

This section will cover the literature review of the genus *Cronobacter*. The topics covered include physiology, taxonomic reviews, reservoirs, infections caused in adults and neonates, potential virulence and environmental fitness associated traits, genome studies and significance of *Cronobacter sakazakii* Clonal Complex 4 (CC4).

1.1.1 PHYSIOLOGY

The members of genus *Cronobacter* consist of Gram negative, motile, peritrichous, non-spore forming, oxidase negative, catalase positive, methyl red negative, able to reduce nitrate to nitrite, facultative anaerobic rods belonging to the family Enterobacteriaceae and related closely to *Citrobacter* and *Enterobacter* genera (Iversen et al. 2008). They can grow over a wide temperature range of 4^{o}C to 44-47^{o}C (Iversen et al. 2004).

A number of physiological characteristics of *Cronobacter* spp. are of importance due to their association with infant formula and neonatal infections. The ability of organism to resist dry heat is of significance. *Cronobacter* are well known for their desiccation tolerance (Osaili and Forsythe, 2009). The organism can tolerate desiccation for more than 2 years and on re-constitution grow rapidly (Caubilla-Barron et al. 2007).The ability of *Cronobacter* to resist osmotic stress and therefore to survive the dried and desiccated conditions of PIF, is considered to be better than other pathogens such as *E. coli*, *Citrobacter* and *Salmonella* (Breeuwer et al. 2003). Several studies have tried to explain the osmotic tolerance of *Cronobacter*. A proteomic based study identified at least 53 proteins associated with osmotic stress in *Cronobacter* (Riedel and Lehner, 2007). Trehalose synthesis by *Cronobacter* has also been associated with osmotic stress response; trehalose is a molecular chaperon which can protect proteins and membranes from denaturation, as in case of *E. coli* (Breeuwer et al. 2003; Horlacher and Boos 1997). Several osmotolerance genes have been identified in *Cronobacter* and other Enterobacteriaceae, for example Feeney and colleagues (2014) described seven homologues of the Prop osmolyte uptake system; ESA_02131, ESA_01706, ESA_04214, ESA_pESA3p05450, ESA_01226, ESA_00673 and ESA_03328 in *C. sakazakii* BAA-894 genome (Feeney et al. 2014). Another study identified homologues of *E. coli* genes *TrkH, trkG, trkA, trkE, kdpA, kdpB, kdpC, kdpD, kdpE, phoP, phoQ, ompC, ompF, envZ* and *ompR* associated with uptake of K^{+} which act as

osmoprotectants as a primary response to osmotic stress, in *Cronobacter* (Feeney et al. 2011) .

Yellow pigment production is an important characteristic of *Cronobacter*, since approximately 80% of the strains have shown the production of non-diffusible yellow pigment on Trypticase Soy Agar (TSA) at 25^{o}C (Iversen and Forsyth, 2004). The production of this yellow carotenoid based pigment may help protect the organism against the free oxygen radicals generated by sunlight, if it is colonizing the plant materials (Osaili and Forsythe, 2009). The yellow pigment production may explain the frequent isolation of *Cronobacter* from plant ingredients and products (Osaili and Forsythe, 2009).

Cronobacter are able to produce copious capsule on milk agar plates (Caubilla-Barron et al. 2007). The production of capsular material may help the organism to form biofilms on different surfaces which can be resistant to disinfectants (Beuchat et al. 2009). The production of capsule may help to protect the organism against desiccation as indicated by Caubilla-Barron and Forsythe (2007). Their study using 27 Enterobacteriaceae strains showed that capsulated *Cronobacter* strains were recoverable from the dehydrated infant formula after 2.5 years. The capsulation of *Cronobacter* is likely to be associated with biofilm formation as the organism has been found as part of the mixed flora in enteral feeding tubes obtained from the NICU of neonates who were not getting infant formula as their feed; this colonization of the enteral feeding tubes may lead to the contamination of the subsequent feeds (Hurrell et al. 2009a, b). Moreover, *Cronobacter* are able to colonise equipment which is used to manufacture feed preparation including silicon, stainless steel, polyvinyl chloride, glass, latex and polycarbonate (Iversen et al. 2004; Lehner et al. 2005).

The properties of pigment production, desiccation resistance and polysaccharide production has been considered as the indicators of the plants are the natural habitats of *Cronobacter* (Iversen and Forsythe, 2003). The capsular polysaccharide may help in the attachment of the organism to the plant surfaces, desiccation resistance help the organism to survive severe environmental conditions while yellow pigment help protect the organism from the sunlight produced free radicals (Iversen and Forsythe, 2003).

Another important trait of *Cronobacter* is thermotolerance which has been studied in detail. Although isolates differ in their ability to tolerate heat, however all the strains were inactivated at the pasteurization temperature of 72^{o}C (Nazarowec-White and Farber 1997;

Breeuwer et al. 2003). These observations were used to review the guide lines for re-constitution of powdered infant formula (PIF) in order to reduce the risk of neonatal infections by *Cronobacter* (FAO-WHO 2004, 2006). *Cronobacter* are able to survive the reconstituted PIF in a temperature range of 6 to 47^{o}C (Iversen et al. 2004). In laboratory, the strains can grow at 37^{o}C in 16 to 18 hours on media such as TSA and MacConkey's agar. A study by Gajdosova et al. (2011) proposed an 18 kb long genetic region consisting of 22 open reading frames (*orf*) associated with thermotolerance in *C. sakazakii*. Their study showed that *orf*HIJK were experimentally linked with thermotolerance since cloning of these *orf*s into *E. coli* caused a two fold increase in its D_{58} values (Gajdosova et al. 2011).

Cronobacter isolates have shown moderate resistance to the acidic conditions of up to pH 3.0; the resistance increases if the strains are pre-exposed to sub-lethal acidic conditions (Edelson-Mammel et al. 2005). The *ompR* gene has been proposed as crucial in response of *C. sakazakii* against highly acidic conditions (Alvarez-Ordonez et al. 2014). Acids are important in food preservation and recently studies have aimed to exploit the bacteriostatic properties of acids to protect them against the *Cronobacter* contamination of baby foods and PIF. For example, a combination of lactic acid and copper at sub-lethal concentrations have shown inhibitory effects on *Cronobacter* spp. In infant formula (Holy et al. 2011).

1.1.2 TAXONOMIC REVIEWS

The genus *Cronobacter* has undergone numerous taxonomic reviews over the last 3 decades. Initially, the genus was identified as yellow pigmented *Enterobacter cloacae*. In 1980, Famer performed DNA-DNA hybridization (DDH) and phenotypic studies for the definition of "*Enterobacter sakazakii*". The results of the DDH indicated that the organism had 41% genetic relatedness with *Citrobacter freundii* and 54% to *Enterobacter cloacae*. Since, phenotypically the organism was more closely related to *E. cloacae*, therefore it was included in the genus *Enterobacter* as a new species known as *Enterobacter sakazakii*, named after the Japanese biologist Riichi Sakazaki (Farmer et al. 1980).

Nearly 2 decades later, Iversen and colleagues (2004) using 16S rDNA and *hsp60* gene sequences analysed the phylogenetic relationship of *E. sakazakii* and *Citrobacter*. Their study identified at least 4 distinct clusters within *E. sakazakii*, each cluster representing a potential novel species (Iversen et al. 2004); the relationship between this genotypic clustering and the biogroups identified by Farmer et al. (1980) was determined by Iversen

et al. (2006) in their further analysis. Follow up studies by Iversen et al. (2007 & 2008) using Amplified Fragment Length Polymorphisms (f-AFLP), 16S rDNA sequence analysis, DDH, ribotyping and phenotypic characterization resulted in taxonomic revision of *E. sakazakii* into the genus *Cronobacter* which belongs to the family Enterobacteriaceae (Iversen et al. 2007; Iversen et al. 2008). At that stage the genus *Cronobacter* consisted of 6 species, *C. sakazakii*, *C. malonaticus*, *C. turicensis*, *C. muytjensii*, *C. dublinensis* and *C. genomospecies* 1 (Iversen et al. 2008). Further studies using optical mapping and genome sequencing data validated the re-classification scheme by Iversen et al. (2007 & 2008; Kotewicz & Tall 2009; Kucerova et al. 2010). Due to high genetic similarity and clonal nature of the members of genus *Cronobacter*, it was difficult to distinguish between *C. sakazakii* and *C. malonaticus* based on the 16S sequence analysis. Baldwin et al. (2009) described a Multilocus Sequence Typing (MLST) scheme using 7 housekeeping genes (concatenated sequence length of 3036 bp) which include ATP synthase b chain (*atpD*), elongation factor G (*fusA*), glutaminyl tRNA synthetase (*glnS*), glutamate synthase large subunit (*gltB*), DNA gyrase subunit B (*gyrB*), translation initiation factor IF-2 (*infB*) and phosphoenolpyruvate synthase A (*ppsA*) to discriminate the species within *Cronobacter* genus (Baldwin et al, 2009; http://pubmlst.org/*cronobacter*/). The scheme is more discriminatory and robust than 16S sequence analysis and showed the separation of *C. sakazakii* and *C. malonaticus* (Baldwin et al. 2009; Yan et al. 2012). It has revealed a stable clonal nature of the virulent *C. sakazakii* strains and was also used in the description of the 2 new *Cronobacter* species *C. universalis* and *C. condimenti* (Joseph et al. 2012a). Joseph and colleagues (2012) applied the MLST scheme to 350 strains. The scheme has indicated certain associations of *Cronobacter* Sequence Types (STs); the ST1 has been associated with infant formula and isolated from clinical sources, the ST8 strains with clinical sources while ST4 or Clonal Complex 4 (CC4) associated dominantly with neonatal meningitis. On the other hand *C. malonaticus* ST7 has been associated with adult infections (Joseph et al. 2012a; Hariri et al. 2013). The MLST scheme has to date identified more than 300 STs in the *Cronobacter* genus. The scheme is hosted by University of Oxford and curated by Professor Stephen Forsythe of NTU, it is available online at (http://pubmlst.org/*cronobacter*/; Forsythe et al. 2014).

Brady et al. (2013) proposed that three non-pathogenic *Enterobacter* spp. (*Enterobacter pulveris*, *Enterobacter helveticus* and *Enterobacter turicensis*), which were originally excluded by Iversen et al. (2008) should be included into *Cronobacter* as 3 new species as *C.*

pulveris, C. helveticus and C. zurichensis (Brady et al. 2013). However, there were ambiguities as no virulence associated traits were attributed to these species. Stephan et al. (2014) proposed that these 3 species should constitute 2 new genera namely *Franconibacter* and *Siccibacter* (Stephan et al. 2014). Furthermore, recent 7 loci MLSA analysis by Forsythe and colleagues indicated that the 3 species do not cluster with other 7 *Cronobacter* species (Forsythe et al. 2014).

To date, *Cronobacter* consists of 7 species which include *C. sakazakii*, *C. malonaticus*, *C. turicensis*, *C. muytjensii*, *C. dublinensis* and *C. condimenti.* Amongst Enterobacteriaceae, the close relatives of the genus *Cronobacter* include *Citrobacter, Enterobacter, Franconibacter* and *Siccibacter* (Baldwin et al, 2009, Forsythe et al. 2014, Stephan et al. 2014).

1.1.3 RESERVOIRS

Cronobacter are ubiquitous in the environment. They have been isolated from a wide range of sources including food, environment, animals and humans (Holy & Forsythe, 2014). *Cronobacter* have been isolated from herbs, spices, salads, cereals, ready to eat food, vegetables, meat and confectionary (Iversen & Forsythe, 2004; Friedemann 2007; Baumgartner et al. 2009).

A wide variety of environmental sources seem to host *Cronobacter*, these include, soil, grass, silage and hot mineral water springs (Neelam et al. 1987; Mosso et al. 1994; Van Os et al. 1996). One of the earliest isolates which has been included in the present study was isolated from dried milk in 1950, the genome of which has now been sequenced (Masood et al. 2013a). *Cronobacter* have been isolated frequently from the processing and non-processing environments within the powdered infant formula (PIF) manufacturing factories which include floors, bays, roller dryers, tankers and air filters (Hein et al. 2009; Craven et al. 2010; Jacobs et al. 2011). The organism has also been isolated from house dust (Kandhai et al. 2004). Additionally, the organism was isolated from the marine environment (Agogue et al. 2006). *Cronobacter* strains also have the ability to colonise and grow on the enteral feeding tubes (Hurrell et al. 2009a, b).

The organism has also been isolated from animal sources, an example is the isolation of a *Cronobacter* strain from the nostrils of a horse (Holy et al. 2011). Similarly, flies can also be the potential vector for *Cronobacter* as a number of different isolates have been obtained

from different species of flies which include Mexican fruit flies and wild house flies (Kuzina et al. 2001; Mramba et al. 2006; Butler et al. 2010; Pava-Ripoll et al. 2012).

Among the human and clinical sources of the organism include cerebrospinal fluid (CSF), blood, sputum, wound infection, bone marrow, urine and faeces (Farmer et al. 1980; Muytjens et al. 1983; Gallagher & Ball, 1991; Iversen et al. 2006; Caubilla-Barron et al. 2007).

One of the most noteworthy sources of isolation for *Cronobacter* which has caught much attention around the world is the powdered infant formula (PIF) which is often used as a substitute for the breast milk to feed neonates and infants. It is important to mention that as a product PIF is not sterile and has been the source of isolation for *Cronobacter* and other bacterial isolates of Enterobacteriaceae family (Kucerova et al. 2011). Worldwide surveys between 1988 and 2004 have shown the frequent isolation of *Cronobacter* strains from PIF (Muytjens et al. 1988; Nazarowec-White & Farber, 1997; Iversen & Forsythe, 2004). A study by Muytjens et al. (1988) showed that *Cronobacter*, then known as *Enterobacter sakazakii* were the third most frequent coloniser of PIF after *E. agglomerans* and *E. cloacae* and was detected in 20 of 141 samples of infant formula tested (Muytjens et al. 1988). In another survey amongst 120 samples of Canadian infant formula tested, at least 6.7% were found to be contaminated with *Cronobacter* spp. (Nazarowec-White & Farber, 1997). However, it must be considered that the bacterial contamination of PIF may not be exclusively intrinsic as extrinsic factors such as contamination during the manufacturing and packaging as well as during the reconstitution by using contaminated water, utensils and hands is also possible. Temperature abuse is another factor which may contribute towards the contamination of prepared formula (Kucerova et al. 2011). Not only do *Cronobacter* infections due to consumption of contaminated PIF risk neonatal health, but they have economic consequences as well, since the product recalls had occurred as in the case of the Tennessee outbreak in 2001 (Himelright et al. 2002) and the December 2011 which claimed the life of a baby in Missouri, USA after which the supermarket chain Walmart recalled a particular batch of PIF and the manufacturers suffered a significant economic loss (CDC, 2012).

1.1.4 INFECTIONS CAUSED BY *CRONOBACTER*

Almost all *Cronobacter* species have been associated with infections, an exception being C. *condimenti* (Cruz-Cordova et al. 2012). Mainly the members of *Cronobacter* have been

associated with neonatal infections which include meningitis, necrotising enterocolitis (NEC), bacteraemia and sepsis (Healy et al. 2010). The species of *Cronobacter* which are most often associated with infantile infections are *C. sakazakii*, *C. turicensis* and *C. malonaticus* (Joseph et al. 2012a, b). *Cronobacter* are often regarded as opportunistic pathogens predominantly associated with life threatening infections in new-born babies generally less than 4 weeks old (Bar-Oz et al. 2001; Mullane et al. 2008), however there have been reports of infections in older infants (Bowen & Braden, 2006). *Cronobacter* spp. not only cause infections in neonates but have also been associated with adult infections especially in elderly or immunocompromised individuals (Caubilla-Barron et al. 2007; See et al. 2007). Although infections caused by *Cronobacter* are rare, the mortality rate in infants is as high as up to 80% (Bowen & Braden, 2006; Friedemann, 2007).

The joint United Nations, World Health Organisation (WHO) and Food and Agriculture Organisation (FAO) joint meetings in 2004 and 2006 reported the annual incidence of invasive "*E. sakazakii*" (i.e. *Cronobacter*) in infants (children below 12 months of age) and neonates; the survey was originally conducted by the United States FoodNEt in 2002 in its nine surveillance sites. According to the survey the annual incidence of the invasive *Cronobacter* strains (isolated from sterile sites only) in infants was 1 per 100, 000 infants, whereas it was 8.7 per100, 000 low-birthweight neonates (FAO/WHO, 2006).

To date, no global tracking system is in place for *Cronobacter* however the expert panel of the WHO tracked 120 cases of *Cronobacter* infection from 1961 to 2008 in infants and children under 3 years of age. *Cronobacter* infections are rarer than other epidemic outbreaks. However it is not the number of cases, but the mortality rate of up to 80%, which is of real concern (about *Cronobacter* infections). Furthermore, the majority of the infected neonates who do recover from the central nervous system (CNS) infections are likely to develop lifelong chronic mental disabilities (Lai, 2001).

1.1.5 INFECTIONS IN NEONATES

1.1.5.1 NECROTISING ENTEROCOLITIS

The general route of entry for *Cronobacter* into the human body is thought to be the gastro intestinal tract (GIT) where they may induce necrotising enterocolitis (NEC) (Liu et al. 2012). Several factors are thought to contribute towards the onset of NEC such as a susceptible neonate with an immature immune system, hypoxia, hypothermia, ingestion of the enteral

formula feed instead of breast milk which contains beneficial natural protective components, and increased bacterial colonisation (Grishin et al. 2013). All of these conditions collectively induce inflammation which in turn cause the production of host inflammatory components such as nitric oxide, cytokines and platelet activating factor thereby damaging the apical GIT (Grishen et al. 2013). It is also believed that the virulence of *Cronobacter* is dose as well as host dependent, they may just remain asymptomatic in term infants, yet causing infections in preterm infants (Hamby et al. 2011; Cetinkaya et al. 2013).

1.1.5.2 THE CNS INFECTIONS

Cronobacter has a tropism towards the central nervous system (CNS), once it enters the systemic circulation it can cause meningitis in low birth-weight babies whereas sepsis and bacteraemia in babies with higher birth-weight (Yan et al. 2012). Once the pathogen crosses the blood brain barrier (BBB), it may induce ventriculitis and form brain abscesses or cysts. The formation of the brain abscesses may in turn develop into a condition called hydrocephalus which is a condition in which cerebrospinal fluid excessively accumulates in the brain (Bowen & Braden, 2006; Chenu & Cox, 2009). The abnormal accumulation of CSF in turn induce enlargement of the brain spaces called ventricles. As a result of the formation of enlarged CSF filled ventricles the balance between the CSF absorption and production is disturbed which subsequently cause intracranial pressure on the brain tissues (Jaradat et al. 2014).

1.1.5.3 ADULT INFECTIONS

Cronobacter infections have also been detected in adults especially in the elderly patients and other immunocompromised individuals (Lai, 2001; See et al. 2007). Members of the *Cronobacter* have been associated with urinary tract (UTI) infections, aspiration pneumonia in stroke patients, diarrhoea, conjunctivitis, osteomyelitis and wound infections (Gosney et al. 2006; Friedemann, 2009; Flores et al. 2011; Yan et al. 2012; Tsai et al. 2013). It is reported that almost 50% of the adults who develop *Cronobacter* infections have underlying malignancy of some sort (Lai, 2001; See et al. 2007).

The nosocomial infections in adults such as UTI and conjunctivitis can be associated with environmental sources such as water, surfaces of medical equipment and person to person contacts as alternative routes of transmission (Friedemann, 2009; Flores et al. 2011).

1.1.6 OUTBREAKS OF *CRONOBACTER*

There have been several outbreaks across different parts of the world, some of the major *Cronobacter* outbreaks are listed here. Muytjens and colleagues (1983) reported outbreak of *Cronobacter* in Netherlands in which at least 8 infants were reported to have been infected. Of these 8 infants, 5 were from the same hospital. All of the infants developed meningitis and 6 of them died. It was the first report that linked *Cronobacter* outbreak with the contamination of the infant formula (Muytjens et al. 1983).

In 1994, a neonatal intensive care unit (NICU) in France, an outbreak of *Cronobacter* spp. occured in which 3 babies died, 1 due to meningitis and other from NEC; the outbreak was analysed by Caubilla-Barron et al. (2007). The Pulsed Field Gel Electrophoresis (PFGE) analysis by Caubilla-Barron et al. (2007) divided the isolates into 4 clusters. It was of high significance that all 3 deaths were associated with the isolates from cluster 2 which formed the largest cluster (Caubilla-Barron et al. 2007). More importantly, all of the cluster 2 isolates belonged to Clonal Complex 4 (CC4) which is at the heart of the present PhD study. Therefore, most of the isolates from this outbreak were sequenced using Illumina MiSeq and analysed further as part of the present study. The genomic analysis of these sequenced strains is presented in chapter 8.

In 2001, van Acker and colleagues reported another outbreak of *Cronobacter* in an NICU in Belgium, however this time *Cronobacter* were associated with NEC in neonates. All the infants infected by *Cronobacter* were low birth weight, preterm and were fed infant formula. Two of the infected babies suffering from NEC eventually died. This was the first report linking *Cronobacter* with NEC (van Acker et al. 2001). Two of the *C. sakazakii* isolates from this outbreak, *C. sakazakii* 2106 (ST21) and 2107 (ST12) have now been sequenced by our group and are being analysed by Emily Jackson as part of her PhD project.

In the same year, 2001, another outbreak of *Cronobacter* spp. occurred in a hospital in Tennessee in which an infant died of meningitis. The outbreak was reported by Himelright and colleagues in 2002. Due to the death of the neonate, 49 more babies were tested of which at least 8 were found to be positive for *Cronobacter*. Of these 8 babies, 6 remained asymptomatic while 2 had respiratory illnesses. Further investigations linked the outbreak with a formula milk product as a result of which a product recall was announced by the manufacturing company. Nonetheless, it must be noted that in this outbreak the formula

fed to infants was not manufactured by the company with an intention to be consumed by neonates which raised further questions about infants' care in the hospitals (Himelright et al. 2002).

Another outbreak of *Cronobacter* was reported by Block et al. (2002) in a hospital in Jerusalem. They isolated *Cronobacter* strains from 5 infants, 3 of which suffered from bacteraemia and meningitis while 2 remained asymptomatic. The meningitic isolate *C. sakazakii* 8399 (*C. sakazakii* 1587) was obtained by our group, its genome sequenced and announced; the genome revealed several virulence associated traits such as an iron acquisition system and some toxic metal resistance associated genes (Masood et al. 2013b). An important observation in this outbreak was that one of the isolate was obtained from the blender which was used to prepare infant formula, more interestingly the pulsetypes of this isolate matched with the pulsetypes of the infants when analysed using PFGE. The organism had established the blender as an environmental niche as it was persistently isolated from the blender for nearly 5 months even after repeated de-contamination of the blender (Block et al. 2002).

1.1.7 POTENTIAL VIRULENCE ASSOCIATED TRAITS

The last decade has seen an increased growth in *Cronobacter* research due to infections associated with the organism. However, as yet the mechanism by which it causes infections is not clearly understood (Jaradat et al. 2014). *Cronobacter* isolates demonstrate a variable virulence phenotype as showed by epidemiological as well as *in vitro* tissue culture studies (Caubilla-Barron et al. 2007; Townsend et al. 2007, 2008). Neonatal infections have mainly been associated with *C. sakazakii*, *C. malonaticus* and *C. turicensis* (Healy et al. 2010, Kucerova et al. 2010). Although the mechanism of *Cronobacter* infections is not clear, several potential virulence associated markers have been determined (Kucerova et al. 2010; Stephan et al. 2011).

Enterotoxin like compounds were the first to be described as putative virulence factors in *Cronobacter* (Pagotto et al. 2003). At least 4 of the 18 *Cronobacter* isolates were able to show enterotoxin production when they were tested using a suckling mice assay (Pagotto et al. 2003). The authors also showed that *Cronobacter* isolates, in addition to the production of enterotoxin, were also able to produce proteolytic enzymes that cause tissue damage to the infection site in mice (Pagotto et al. 2003). A zinc metalloprotease has been

reported by Kothary et al. (2007) which induced the rounding of the Chinese hamster ovary (CHO) cells thereby causing cell damage (Kothary et al. 2007). An enhancement in translocation of *C. sakazakii* has been observed to the blood brain barrier (BBB) from the rat gut when infant formula is supplemented both with *C. sakazakii* and endotoxin (Townsend et al. 2007). One possible mechanism is thought to be the tight junction disruption by lipopolysaccharides (LPS) which may increase the permeability of the BBB to the organism (Kim & Loessner, 2008). Additionally, enterocyte migration and epithelial restitution was also found to be impaired by the endotoxin thereby enhancing the permeability of the BBB to the bacteria (Cetin et al. 2004). Since the endotoxin was found to be heat stable at 100^{o}C, its presence in infant formula may enhance its pathogenicity for neonates (Townsend et al. 2007).

Adhesion is the first and essential step towards bacterial pathogenesis as it facilitates the colonisation and subsequent infection of host cells (Soto & Hultgren, 1999). Fimbriae are the appendages which help bacteria to adhere to the host cell surface thereby helping them with colonisation and successive infection process (Soto & Hultgren, 1999). At least ten putative fimbrial clusters have been described in *Cronobacter* (Joseph et al. 2012b). An important observation was the absence of the curli fimbriae in all of the *C. sakazakii* genomes analysed which indicates that curli fimbriae are not essential for the virulence of *C. sakazakii* unlike *E. coli* (Joseph et al. 2012b).

Mittal and colleagues (2009) have described the role of ompA and ompX proteins in bacterial penetration through the BBB (Mittal et al. 2009). The role of *ompA* in *Cronobacter*'s invasion of human brain microvascular epithelial cells (HBMEC) *in vitro* has also been described (Mohan & Venkitanarayanan, 2006; Singamsetty et al. 2008). However since genes for both *ompA* and *ompX* are present in all of the *Cronobacter* genomes, their role regarding the variability in pathogenicity of *Cronobacter* is not clear (Joseph et al. 2012b). In the present study, an attempt was made to study the variation in the sequence both at nucleotide and amino acid level to check for variability amongst *C. sakazakii* sequenced strains with special emphasis on *C. sakazakii* CC4. The analysis is presented in chapter 3.

Kucerova et al. (2010) showed that the genome of *C. sakazakii* BAA-894 encode a cation efflux system (*cusA, cusB, cusC and cusF*) along with a regulatory gene *cusR*; according to

their analysis these genes were found only in the isolates associated with infections in neonates while missing in other isolates which may in part explain the variation in the pathogenic ability of *Cronobacter* (Kucerova et al. 2010). Two of the iron acquisition systems have been found encoded on the *Cronobacter* plasmids pESA3 and pCTU1; an ATP binding cassette transport-mediated iron uptake (*eitCBAD*) and another iron acquisition system which is mediated through siderophore (*iucABCD/iutA*) (Kucerova et al. 2010; Franco et al. 2011a; Joseph et al. 2012b). Since iron acquisition systems have been detected in all *Cronobacter*, their role towards the variability in pathogenicity of *Cronobacter* is not understood (Joseph et al. 2012b).

Secretion systems are important towards the pathogenicity of bacteria as they are able to translocate proteins and nucleoprotein complexes and may contribute towards the virulence of *Cronobacter* as a plasmid borne type IV secretion system has been identified in *C. sakazakii* and *C. turicensis* encoded on the plasmids pESA2 and pCTU2 respectively (Kucerova et al. 2010; Franco et al. 2011a; Joseph et al. 2012b). A newly described type six secretion system (T6SS) which has been associated with invasion of the BBB in *E. coli* K1 (Zhou et al. 2012) was found encoded on the large plasmid pESA3 of *C. sakazakii* BAA-894, which was found to be variably present in other *Cronobacter* isolates (Kucerova et al. 2010; Joseph et al. 2012b).

Other potential virulence associated traits include a plasmid encoded plasminogen activator called *cpa* (ESA_ pESA3p05434) that helps in the activation of plasminogen and provides resistance to *C. sakazakii* against the bactericidal activity of the serum (Franco et al. 2011b) and a superoxide dismutase (*sodA*) for macrophage survival both of which were detected uniformly in all *Cronobacter* genomes (Joseph et al. 2012b).

1.1.8 OTHER VIRULENCE ASSOCIATED TRAITS

1.1.8.1 SIALIC ACID UTILISATION

The human brain is rich in sialic acid which could be utilised by *C. sakazakii* as an alternative source of carbon and nitrogen thereby allowing them to survive and persist in brain, inflicting brain damage. Recent genome analysis (Joseph et al. 2012b; Joseph et al. 2013) has shown that sialic acid catabolism genes required for the utilisation of exogenous sialic acid are found only in *C. sakazakii* species. No evidence of the *de novo* synthesis of sialic acid exists in *C. sakazakii.* The *nanAEKTR, nagA* and *nagB* genes, essential for the transport

and catabolism of sialic acid have been found in all the *C. sakazakii* strains. The gene encoding the outer membrane porin protein *nanC* had not been annotated in *C. sakazakii* BAA894, however pBLAST homology searches reveal ESA_03302 as the probable *nanC*. The *nanAET* genes are usually found clustered in bacteria (Vimr et al. 2004) as per *Edwardsiella tarda* but in *C. sakazakii* the proposed *nanE* gene, ESA_00529 is located at a different loci. The *C. sakazakii* strains have shown the ability to grow on sialic acid (Joseph et al. 2013). Certain bacteria have shown the ability to cleave sialic acid from sialoglycoconjugates using the enzyme sialidase (Corfield et al. 1992), however the presence of sialidases is not frequent in bacteria. Bacterial species lacking sialidases but still catabolising sialic acid either use sialic acid cleaved by other pathogens colonising the same niche (Shakhnovich et al. 2002) or by the host sialidases (Sohanpal et al. 2004). Extensive homology searches were carried out in order to find a homologue of a sialidase however the gene was not found in *C. sakazakii.* This study has now been published and our group was the first to show that sialic acid can be an alternative carbon and energy source for *C. sakazakii* (Joseph et al. 2013).

1.1.8.2 INOSITOL FERMENTATION

Based on the presence of the inositol monophosphate gene (*suhB*), inositol fermentation was proposed as the marker for pathogenicity in *Cronobacter* (Hamby et al. 2011), however the presence of the GR29 operon associated with inositol fermentation in the environmental isolates of *Cronobacter* and its absence in the pathogenic isolates seems to contradict its association with pathogenicity in *Cronobacter*, therefore the role of inositol fermentation in virulence of *Cronobacter* is not clear (Grim et al. 2013).

1.1.8.3 BIOFILM FORMATION

Cronobacter are able to form biofilms on different surfaces such as latex, silicon, polyvinyl chloride, stainless steel and polycarbonate (Iversen et al. 2004; Lehner et al. 2005). The ability of *Cronobacter* to form biofilms on these surfaces convey resistance against various environmental stresses and in turn makes them resilient to cleaning by sensitisers and other anti-microbial agents (Ravishankar et al. 2003; FAO/WHO, 2006; Kim et al. 2007). The ability of bacteria to form biofilms is of special concern in the food industry as these biofilms can act as reservoirs for contamination and spoilage of the food products which are being processed (Lehner et al. 2005; Hartmann et al. 2010). All *Cronobacter* spp. studied by Joseph

et al. (2012) possessed capsular polysaccharides genes *wzABCKM* (ESA_01155-01175) which may be important towards the biofilm formation in *Cronobacter* (Joseph et al. 2012b). A study has shown the role of flagella in biofilm formation by *C. sakazakii*, the authors also indicated 2 hypothetical proteins (ESA_00281-2) which may contribute towards the biofilm formation in *C. sakazakii* (Cruz-Cordova et al. 2012).

1.1.8.4 ANTIBIOTIC RESISTANCE AND SUSCEPTABILITY

There have been contradictory reports regarding the antimicrobial susceptibility and resistance of *Cronobacter*. *Cronobacter* were found to be the most susceptible amongst the Enterobacteriaceae when tested for susceptibility to 29 antimicrobial agents (Muytjens & van der Ros-van de Repe, 1986). This was validated by another study showing that *Cronobacter* isolates did not show expression of β-lactamase and were constantly susceptible to β-lactams (Stock & Widemann 2002). However, Pitout et al. (1997) reported a low level production of β-lactamase while Caubilla-Barron et al. (2007) reported 2 of the *Cronobacter* isolates showing β-lactamase activity (Pitout et al. 1997; Caubilla-Barron et al. 2007). Similarly, another report suggested that isolates differ in their ability to resist β-lactam antibiotics, the same study reported a PIF isolate for its β-lactamase activity (Zhou et al. 2011).

Before 1985, *Cronobacter* infections were treated mainly with ampicillin, gentamicin and/or chloramphenicol. However, in 1988, to treat meningitis caused by *Cronobacter*, Willis and Robinson recommended a combination of ampicillin and gentamicin (Willis & Robinson, 1988). Another study reported that all *Cronobacter* isolates used in their study were susceptible to aminoglycosides and trimethoprim/sulfamethoxazole and were able to resist extended-spectrum penicillins, cefazolin and ampicillin (Lai, 2001). A recent study has shown that antibiotics such as streptomycin, ciprofloxacin, kanamycin and gentamicin might be appropriate choices for treating the infections as all of these were effective against *C. sakazakii* in both its unstressed and stressed states (Al-Nabulsi et al. 2011).

The overall antibiotic resistance of *Cronobacter* was found to be low compared with other food-borne pathogens, even though *Cronobacter* isolates were found to harbour multiple antibiotic resistance operon (*mar*). The overuse of antibiotics could be a cause of the antibiotic resistance in *Cronobacter* (Lee et al. 2012).

1.1.9 ENVIRONMENTAL RESISTANCE

The frequent isolation of *Cronobacter* from PIF and from human hosts show the ability of this organism to withstand stresses during the processing of PIF or inside the human body. The ability to resist environmental stresses is crucial for the survival, persistence and the ability of *Cronobacter* to induce infections in susceptible host.

1.1.9.1 HEAT RESISTANCE

Like antibiotic resistance and susceptibility, there are contradictory reports about the heat tolerance of *Cronobacter*. Some of the studies suggest that *Cronobacter* isolates are thermotolerant while other studies report no heat tolerance (Nazarowec-White & Farber, 1997; Breeuwer et al. 2003; Osaili & Forsythe, 2009). The contradiction in these studies could be explained by several factors which could influence the heat resistance ability of *Cronobacter* such as the source of isolation, growth conditions and variations between the isolates (Arroyo et al. 2009; Walsh et al. 2011; Jaradat et al. 2014). Different food properties such as high fat content and low water activity can also influence the ability of *Cronobacter* to resist heat (Osaili & Forsythe, 2009). A study by Gajdosova et al. (2011) reported an 18kbp region containing 22 open reading frames (ORFs) which were up-regulated when the isolates were subjected to heat adaptation conditions. According to the authors, genes within this cluster indicated homologies against known bacterial proteins which are involved in stress response such as heat, acid stress and oxidation (Gajdosova et al. 2011; Jaradat et al. 2014).

1.1.9.2 DESICCATION RESISTANCE

The ability of *Cronobacter* to resist drying and osmotic stresses is considered to be higher than for other members of the Enterobacteriaceae (Feeney & Sleator, 2011). A study has shown that *Cronobacter* strains were recoverable from PIF after 2.5 years (Barron & Forsythe, 2007). The ability of *Cronobacter* to resist desiccant conditions could be due to the accumulation of an osmoprotectant called trehalose inside the cells where it protects the cells from dying due to low moisture content (Breeuwer et al. 2003). A number of genes have been associated with osmoprotection in *Cronobacter* such as trehalose and betaine encoding genes which were detected in all *Cronobacter* strains analysed (Joseph et al. 2012b). It was assumed that the products of these genes help *Cronobacter* to survive in the food ingredients which are used in dry food preparations. Using bioinformatics analysis,

Feeney and Sleator (2011) showed that ProP and OpuC which are the homologues of the osmoprotectants are present in *C. sakazakii* genomes and the osmotolerance response is regulated at the transcriptional, translational and post-translational levels. They further showed that RpoS is probably the global transcriptional regulator (Feeney & Sleator, 2011). Additionally several other genes were also identified whose synergistic expression might provide resistance against dry and desiccant conditions such as colanic acid exopolysaccharide and cellulose biosynthesis genes (Grim et al. 2013).

1.1.9.3 ACID RESISTANCE

In order to cause infections in neonates and other immuno-compromised individuals, *Cronobacter* strains should have the ability to survive very high acidic conditions of the stomach. Studies have shown the ability of *Cronobacter* to survive and grow under high acidic conditions. In one study it was reported that at least 79.2% of the tested *Cronobacter* strains were able to grow at a pH as low as 3.9 (Dancer et al. 2009). Another study showed that out of 12 *Cronobacter* tested strains at least 10 strains showed only <1 log decline in their growth at pH 3.5 as compared to acid sensitive strains which showed a ~3.5 log decline, over a 5 hours incubation at 37^{o}C (Edelson-Mammel et al. 2005). Similarly another study has reported the ability of *Cronobacter* not only to survive but to grow at a low pH 4.5 (Johler et al. 2009).

1.1.10 O-ANTIGEN TYPING OF *CRONOBACTER*

In Gram negative bacteria the outer surface of the lipopolysaccharide (LPS) contains O-antigen which is responsible for serological diversity of bacteria. Like other Gram negative bacteria, the LPS covers the outer membrane of *Cronobacter* along with lipid A core. The lipid A is the toxic moiety covering at least 8% of the total LPS weight (MacLean et al. 2009).

Like other Gram negative bacteria, O-antigen has also been used to characterise *Cronobacter* spp. based on their O-antigen structure (Mullane et al. 2008; Jarvis et al. 2011, 2013). In Gram negative bacteria, the *rfb* gene locus constitutes a highly variable region of the O-antigen and is located between *galF* and *gnd;* the first O-antigen serotyping method for *Cronobacter* spp. was developed by Mullane et al. (2008) based on the amplification of *rfb* locus followed by digestion using *Mbo*II. A PCR-RFLP profile can be generated with this method which can then be used to compare different *Cronobacter* isolates. The O-antigen serotypes, O:1 and O:2 of *Cronobacter* were characterised using this approach (Mullane et

al. 2008). Five more serotypes were added to *Cronobacter* serotypes by Sun et al. (2011). The original O-antigen serotyping scheme was further extended and 8 new serotypes were described by Jarvis et al. (2011). Of these 8 new serotypes at least 2 were found in more than 1 *Cronobacter* species including *C. sakazakii*, *C. muytjensii*, *C. turicensis* and *C. malonaticus* (Jarvis et al. 2011, 2013). To date there are 17 recognized serotypes in *Cronobacter* covering 7 species.

Most of the virulence associated attributes discussed above including the environmental adaption of *Cronobacter*, have been analysed in sequenced *C. sakazakii* genomes and the analysis is presented in chapter 3, in order to observe if there is any association of these traits to a particular lineage with special emphasis on *C. sakazakii* clonal complex 4 (CC4), a stable clonal lineage with sequence type 4 (ST4) as the dominant ST with single or double locus variants associated frequently with neonatal meningitis, which is the main focus of this PhD thesis.

1.1.11 PUBLISHED GENOMES

The first annotated genome of *C. sakazakii* BAA-894 was published by an NTU research group (Kucerova et al. 2010). The whole genome sequence of *C. sakazakii* BAA-894 (isolated from an NICU outbreak) revealed one chromosome (4.36837 Mb) and three plasmids of which only two have been sequenced and released; pESA2 (31 Kb) and pESA3 (131 Kb). To date, 8 genomes of *C. sakazakii* have been released in Genbank, of which 6 genomes have been sequenced and released by our group in collaboration with the University of Exeter, UK and Life Technology Inc. USA. In 2012, Shin and colleagues announced the finished genome of *C. sakazakii* strain ES15 isolated from ground whole grains consisting of a single chromosome (4.27 Mb). The first draft genome of a *C. sakazakii* CC4 (ST4) strain (701) was sequenced by our group in collaboration with Life technologies and has been made available via Genbank (Joseph et al, 2012b). Power and colleagues (2013) announced the genome of the first finished *C. sakazakii* CC4 (ST4) strain SP291 which had been isolated from a powdered infant formula (PIF) production site, consisting of 1 chromosome (4.3 Mb) and 3 plasmids (118kb, 52kb and 4.4kb). Recently our group has released the earliest *C. sakazakii* CC4 (ST4) strain NCIMB 8272 (NTU ID; 377) in international culture collection (accession: AWFW00000000). This genome was sequenced in collaboration with the University of Exeter (UK).

1.1.12 GENOME STUDIES

The association of *Cronobacter* with neonatal infections has caught the attention of researchers worldwide especially in the last 2 decades, and several genome analysis studies have been undertaken, some of the relevant studies are highlighted here. The first genome of *Cronobacter* genus *C. sakazakii* BAA-894 was published by Kucerova et al. (2010) who used a whole genome microarray to analyse the genus diversity of the genus *Cronobacter* (Kucerova et al. 2010). The genome consisted of 1 chromosome (4.4 Mb) and 3 plasmids (2 sequenced) pESA2 (31 kb) and pESA3 (131 kb) and can be accessed from Genbank (accession: NC_009778-80). The authors used the genome sequence of *C. sakazakii* BAA-894 to construct 384, 030 probe oligonucleotides to tile the DNA microarray. They used 10 of the isolates as query strains which belonged to 5 different members of *Cronobacter* genus, namely *C. sakazakii*, *C. malonaticus*, *C. turicensis*, *C. muytjensii* and C. *dublinensis*. Their analysis indicated that greater diversity within *Cronobacter* is attributed to the presence of mobile genetic elements and prophage regions. Furthermore a number of different virulence associated traits such as type 6 secretion system (T6SS), iron acquisition genes and fimbrial gene clusters were also identified (Kucerova et al. 2010), these have been discussed and analysed in chapter 3.

In 2011, Stephan and co-workers announced the genome of *C. turicensis* z3032 which was obtained from a fatal case in a children's hospital in Zurich, Switzerland (Stephan et al. 2011). The genome of *C. turicensis* z3032 (Accession No. NC_013282-85) consisted of a chromosome (4.38 Mb) and 3 plasmids of sizes 138kb, 53kb and 22kb respectively.

A comparative genomic study was conducted by Franco et al. (2011) in which they compared the large plasmid pESA3 (131 kb) of *C. sakazakii* BAA-894 with the large plasmid pCTU3 of *C. turicensis* z3032 (138 kb). Their study showed that both of these plasmids have homologous gene contents. They designated these plasmids as virulence plasmids sharing the same backbone since they harbour virulence associated genes such as T6SS, iron acquisition, filamentous haemagglutinin (*fha* locus) and plasminogen activator (*cpa*) (Franco et al. 2011a, b).

A comprehensive comparative genomics study of *Cronobacter* was undertaken by Joseph et al. (2012) in which a total of 14 *Cronobacter* genomes covering 7 species were analysed (Joseph et al. 2012b). The genomic analysis indicated a larger pangenome comprising over

6000 genes with a small core genome of *Cronobacter* comprised 2000 genes. A considerable degree of variation was observed for the presence of virulence associated traits such as adhesins, T6SS, heavy metal resistance genes associated with resistance to tellurite, copper and silver. The authors showed that *C. sakazakii* genomes were unique as compared to other members of the genus as they harboured genes essential for the utilisation of exogenous sialic acid. The comparative study also included 2 *C. sakazakii* CC4 genomes, 701 and E899, however no unique traits were observed in the CC4 genomes which could be linked with neonatal meningitis. In addition, several prophage fragments were also detected which contribute towards the genomic diversity in *Cronobacter* (Joseph et al. 2012b). The comparative genome analysis by Joseph et al. (2012b) provided the foundation for further genomic analysis and will be referred to in other sections of this thesis.

In another study, Grim et al. (2013) analysed 6 strains representing 6 of the 7 species, their genomic analysis was based on average nucleotide identity (ANI), whole genome alignment and phylogenetic re-construction. The authors suggested that the species within *Cronobacter* diverged in a bidirectional manner and this divergence is influenced by niche adaptation. According to the study, 2 of the species: *C. dublinensis* and *C. muytjensii,* have acquired genomic traits such as maltose operon, Xylose utilization genes, galactouronate, phyto-toxin and β-carotene associated genes which make them more suitable for the environmental and plants associated habitats. On the other hand, genomes of other 4 species *C. sakazakii*, *C. malonaticus*, *C. universalis* and *C. turicensis* have acquired more virulence associated traits such as T6SS, iron acquisition and serum resistance associated genes, which make them more suitable for the host environment. Their genomic study also revealed a larger core genome consisting of 3160 coding sequences (CDS) which was much larger than previously reported i.e. 1899 and 2000 genes, by Kucerova et al. (2010) and Joseph et al . (2012b) respectively. Their study also listed 84 genomic regions which were present in 2 or more *Cronobacter* genomes and 45 genomic regions which were unique in a subset of the genomes, the notable traits were the same which were previously described by Kucerova et al. (2010) and Joseph et al. (2012b) including T6SS, heavy metal resistance such as tellurium, copper/silver and prophage elements (Grim et al. 2013). Since their study only analysed 6 genomes representing 6 of 7 species of *Cronobacter*, only further genomic analysis comprising larger number of genomes representing each species in the genus can validate the hypothesis of species level bidirectional divergence within *Cronobacter*.

1.1.13 *CRONOBACTER SAKAZAKII* ST4/CC4 AND NEONATAL MENINGITIS

The *Cronobacter* MLST scheme has identified over 300 STs for *Cronobacter* (Forsythe et al. 2014). The *C. sakazakii* ST4 is found to be the predominant ST within CC4 associated with neonatal meningitis and frequently isolated from CSF (Joseph & Forsythe, 2011). A previous study by Joseph and Forsythe (2011) suggested a strong association of *C. sakazakii* ST4 with neonatal meningitis. In their study they analysed 41 *Cronobacter* strains of which almost half (20/41) were *C. sakazakii* ST4 (Table 1.1) (Joseph & Forsythe, 2011). Furthermore 9 of 12 meningitic isolates were *C. sakazakii* ST4, thus the study indicated a strong association of *C. sakazakii* ST4 and neonatal meningitis, moreover it also indicated *C. sakazakii* ST4 as a clonal lineage as the isolates were obtained from 7 different countries over 50 years period of time as shown in Table 1.1 from Joseph and Forsythe (2011).

Table 1. 1 *Cronobacter sakazakii* strains, their clinical origin and *Cronobacter* MLST sequence type (Source; Joseph & Forsythe, 2011*)

Strain	Patient type/age (EGA)†	Clinical signs/outcome	Isolation site	Year	Country	ST
1218	Neonate/<1 mo (30 wk)	Fatal meningitis	CSF	2001	USA	1
1241	Infant/7 mo	Sudden infant	Blood	2008	USA	1
12	Adult/74 y	UNK	Feces	2004	Czech Rep	1
HPB 3290	Neonate (33 wk)	Meningitis	CSF	2001	USA	1
984	Neonate/3–4 wk	UNK	Enteral	2007	UK	3
978	Neonate/<1 wk	UNK	Enteral	2007	UK	3
553	Neonate/1 d	UNK	UNK	1977	Netherlands	4
557	Neonate/5 d	UNK	UNK	1979	Netherlands	4
1242	Infant/7 wk	Fatal meningitis	Brain	2008	USA	4
1222	Infant/8 mo	Fever, recovered	Blood	2003	USA	4
1224	Infant/10 mo	Fever, severe	Blood	2004	USA	4
558	UNK	UNK	UNK	1983	Netherlands	4
20	Child/6 y	UNK	Feces	2004	Czech Rep	4
695	Neonate/15 d (32 wk)	Fatal NEC II	Trachea	1994	France	4
701	Neonate/28 d (28 wk)	Fatal NEC III	Peritoneal	1994	France	4
709	Neonate/18 d (29 wk)	Septicemia	Trachea	1994	France	4
767	Neonate/19 d (31 wk)	Fatal meningitis	Trachea	1994	France	4
721	Neonate/2 wk	Meningitis	CSF	2003	USA	4
1219	Neonate/<1 mo (36 wk)	Fatal meningitis	CSF	2002	USA	4
1221	Neonate/<1 mo	Meningitis	CSF	2003	USA	4
1225	Neonate/<1 mo (35 wk)	Fatal meningitis	Blood	2007	USA	4
1231	Neonate (33 wk)	Fatal neurologic	Feces	2004	New Zealand	4
1220	Infant/6 wk (37 wk)	Brain abscess,noFatal	CSF	2003	USA	4
1223	Infant/6 wk (31 wk)	UNK, in ICU	Blood	2004	USA	4
1240	Infant/7 wk	Fatal meningitis	CSF	2008	USA	4
HPB 2853	UNK	UNK	UNK	1990	Canada	4
HPB 2852	UNK	UNK	UNK	1990	Canada	8
CDC 407–77	UNK	UNK		1977	USA	8
CDC 996–77	UNK	UNK	Spinal	1977	USA	8
511	UNK	UNK	UNK	1983	Czech Rep	8
513	UNK	UNK	UNK	1983	Czech Rep	8
526	UNK	UNK	UNK	1983	Czech Rep	8
ATCC 29544	Child	UNK	Throat	1980	USA	8
696	Neonate/17 d (32 wk)	NEC II	Feces	1994	France	12
690	Neonate/27 d (31 wk)	Asymptomatic	Feces	1994	France	12
520	UNK	UNK	UNK	1983	Czech Rep	12
693	Neonate/13 d (41 wk)	Asymptomatic	Feces	1994	France	13
HPB 2856	Child/6 y	UNK	UNK	2002	Canada	15
NCTC 9238	UNK	UNK	Abdomen pus	1953	UK	18
1249	Neonate	Fatal infection	UNK	2009	UK	31
CDC 0743–75	UNK	Foot wound	Wound	1975	USA	41

*No copyright permission needed (http://wwwnc.cdc.gov/eid/page/copyright-and-disclaimers), d; day, wk; week, mo; month, y; year, UNK; unknown, CSF; cerebrospinal fluid, NEC; necrotising enterocolitis, the table was re-ordered according to sequence type.

These findings were strengthened by further studies in our group by Hariri et al. (2013). The authors analysed 15 *Cronobacter* isolates from the Centers for Disease Control (CDC) which showed an unequal distribution pattern for *C. sakazakii* isolates (Hariri et al. 2013). According to the authors, all 5 CSF isolates were either ST4 or within the ST4 complex (one

or two loci variants) thereby forming ST4 complex called Clonal Complex 4 (CC4). Further to Joseph and Forsythe (2011), Hariri and colleagues (2013) introduced the concept of *C. sakazakii* CC4, a clonal lineage consisting predominantly of ST4 strains and strains which differ in only 1 or 2 MLST loci and associated frequently with neonatal infections especially neonatal meningitis (Joseph & Forsythe, 2011; Hariri et al. 2013). Another study by Joseph and Forsythe (2012) also showed the frequent isolation of *C. sakazakii* CC4 from food, clinical sources as well as from powder milk as shown in figure 1.1 of the geoBURST analysis by Joseph and Forsythe, 2012.

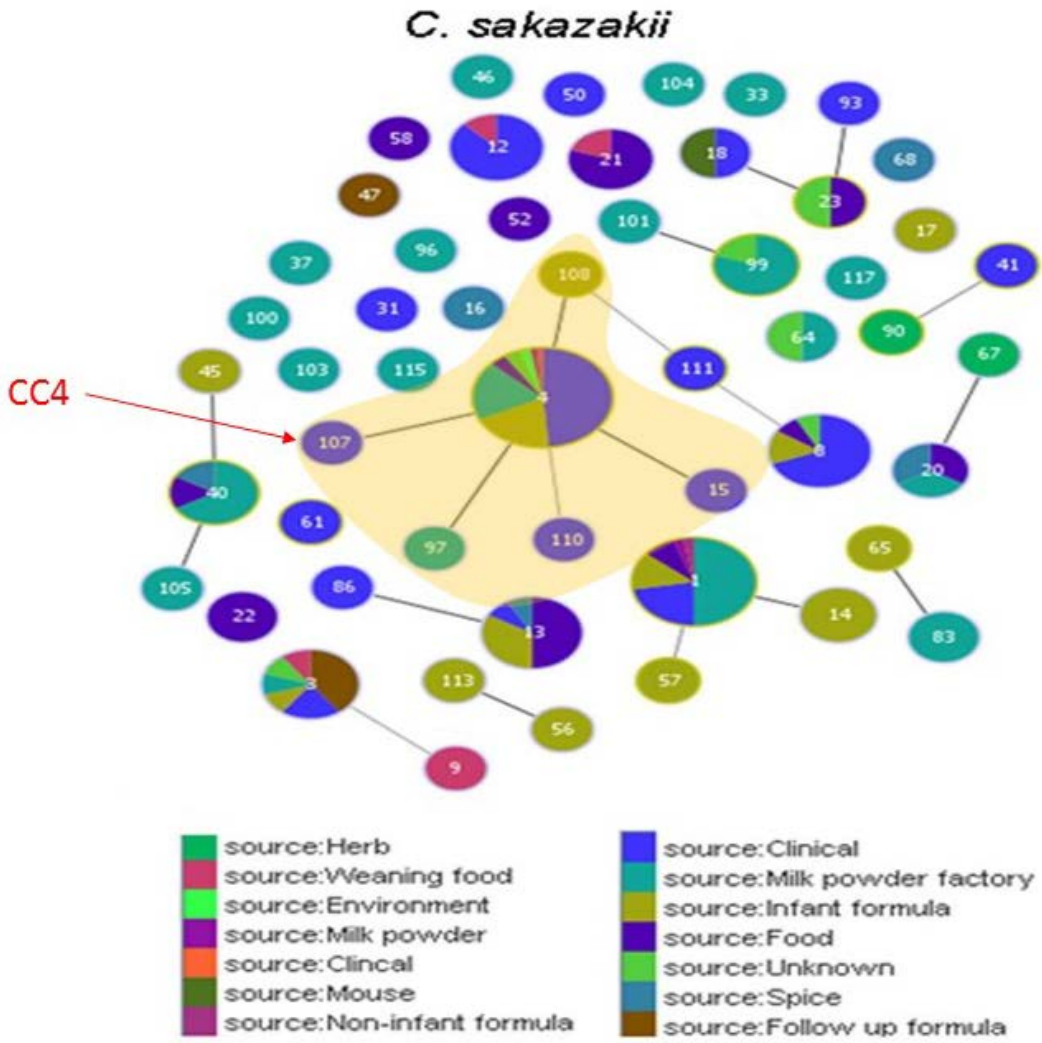


Figure 1. 1 The geoBURST image of the *C. sakazakii* clonal complexes (Source; Joseph & Forsythe, 2012*).

This cropped section of the figure from Joseph & Forsythe (2012) shows the association of *C. sakazakii* clonal complexes with different sources as published by Joseph & Forsythe (2012). The yellow highlighted region indicates the *C. sakazakii* CC4 i.e. *C. sakazakii* ST4 and its single/double loci variants. Each colour indicates a different source of origin for the isolates. The image is cropped to show *C. sakazakii* clonal complexes only. *No permission needed for the figure reuse under Frontiers Creative Commons Attribution License (CC BY).

Furthermore, it was not only the association of *C. sakazakii* CC4 with neonatal meningitis which was of significance, it was the occurrence of *C. sakazakii* CC4 in milk powder and its manufacturing environment which is of great concern. In another study by our group, Sonbol et al. (2013) described the association of *Cronobacter* STs in global surveys. Sonbol et al. (2013) analysed 85 *Cronobacter* isolates by MLST, collected between 1988 and 2009 from 14 different countries. The findings indicated that at least 24% of these isolates were *C. sakazakii* CC4 which were isolated from milk powder and dairy manufacturing environment including roller-dryers, tanker bays, evaporators, roofs and spray drying areas (Sonbol et al. 2013). These findings were further strengthened by Muller et al. (2013); the MLST analysis of the selected *Cronobacter* isolates from a Swiss infant formula production facility indicated that most of these isolates were *C. sakazakii* ST4 (Muller et al. 2013).

It is now well understood that *C. sakazakii* ST4 is a complex of STs forming a clonal complex 4 or CC4 which is dominated by *C. sakazakii* ST4 isolates. A clear link has been established between *C. sakazakii* CC4 and neonatal infections especially meningitis. Furthermore it has also been well established that this is a stable and a clonal lineage which has the ability to persist in PIF and its manufacturing environment. The mortality rate of up to 80% in neonates presses the need to study this neonatal health associated lineage at the genomic level. Therefore the present study was aimed mainly at the genomic interrogation of the sequenced *C. sakazakii* genomes with the target to identify any unique virulence associated determinants in *C. sakazakii* CC4 which help them induce meningitis in newborns or any environmental fitness associated traits which support them to persist in the environment ultimately increasing the risk of exposure to neonates. The identification of these virulence associated genes would be key to combat infections caused by this pathogenic lineage thereby helping save the precious lives of newborn babies.

1.1.14 POWDERED INFANT FORMULA

As described above, the frequent association of *Cronobacter* with PIF, particularly *C. sakazakii* CC4 is alarming. It is therefore of important to discuss the powdered infant formula manufacturing process and critical points of contamination.

Human breast milk is considered to be the best form of nutrition for the new-born babies (Vandenplas, 2002). However, there are instances when breast feeding is not possible. In these circumstances human breast milk substitutes are made available in the form of either

powdered infant formula or liquid formula, in order to maintain the normal growth of the neonate (Kent et al. 2015). However, due to their immature immune system and permeable gastro-intestinal tract, neonates are more susceptible to food borne pathogens such as *Cronobacter* and *Salmonella* which are able to survive in PIF (Iversen and Forsythe, 2003; Lonnerdal, 2012). The consumption of PIF contaminated with food borne pathogens such as *Cronobacter* therefore can cause serious illnesses in neonates (Iversen and Forsythe, 2003; Caubilla-Barron et al. 2007; Joseph and Forsythe, 2011). *Cronobacter* spp. has ability to survive the desiccated conditions of PIF for over two years and upon reconstitution grow rapidly (Caubilla-Barron et al. 2007; Osaili and Forsythe, 2009). Contamination of PIF with *Cronobacter* over the last two to three decades is of great concern as it has claimed the lives of number of neonates (Caubilla-Barron et al. 2007). The manufacture process of PIF and critical points of contamination are discussed briefly.

The United states department of Food and Drug Administration regulates the nutrients of PIF. Although components may vary depending upon the formula type, nevertheless all the formulae contain the source of proteins, fats and carbohydrates, diluents, minerals such as iron, copper, zinc, calcium, phosphate, potassium, chloride, magnesium, sulphur, sodium and iodine. Iron is considered to be one of the most important components of PIF since all babies need iron for their normal growth. Vitamins added in PIF include vitamin A, B12, C, D, E, riboflavin, thiamine, pyridoxine, pantothenate, niacin and folacin. A variety of emulsifiers or stabilizers are added so that formula remains homogenous.

An important component of bovine milk based PIF is lactoferrin (Masson and Heremans, 1971). Lactoferrin is a glycoprotein of the transferrin family of proteins found in mucosal surfaces and a variety of external secretions, its main source being milk (Sanchez et al. 1992; Harouna et al. 2015). An important characteristic of lactoferrin is that it can bind to two atoms of iron and therefore can help bacteria in iron acquisition milk (Sanchez et al. 1992; Harouna et al. 2015). Iron is an important component of PIF, however its presence along with lactoferrin may enhance the bacterial growth, in cases where contamination of PIF occurs. It has been shown that if lactoferrin is devoid of iron, it can prevent iron utilisation by some bacteria such as *Legionella pneumophila* (Orsi, 2004).

1.1.14.1 MANUFACTURE OF THE POWDERED INFANT FORMULA

Most of the infant formulae are based on bovine milk, however it contains higher level of proteins, fats and minerals, therefore it is diluted and skimmed so that it closely resemble the composition of human milk (Koletzko et al. 2005; CAC, 2007). Infant formulae based on soy are also available, however due the potential harmful effects on neonates due to the presence phytoestrogens, it is recommended that their use should be limited (Bhatia and Greer, 2008; Badger et al. 2009).

According to the industrial experts form Europe and USA, manufacturing of the powdered infant formulae involves mixing of the ingredients, processing and packaging (FAO/WHO, 2004). Each of the step is discussed briefly below.

1. **MIXING**.

The process of mixing PIF ingredients is divided into three categories.

a. **Wet Mix process**: In this process all the ingredients of the PIF are handled in the liquid state. After mixing, the mixture is heat treated through pasteurization or sterilization followed by drying.

b. **Dry-mix process:** In this process the individual ingredients of PIF are prepared followed by heat treatments, drying and dry-blending.

c. **Combined process:** Part of the ingredients are prepared using the wet mix process to make the base powder. The remaining ingredients are prepared using the dry-mix process and then added to the base powder.

2. PROCESSING

Ingredients which include milk, derivatives of milk vitamins, amino acids, fatty acids, carbohydrates, soy proteins, minerals and food additives make up PIF. These ingredients, either in the powdered or liquid form are usually mixed with water to make a liquid mix. This liquid is heat treated (pasteurized for 15 seconds at 72^{o}C or for 25 seconds at 74.4^{o}C). Following pasteurization, the liquid mix is dried to powdered form ($a_w \leq 0.3$) by spray drying. Vitamins and heat sensitive ingredients are added prior to dying. During the drying process, the liquid mix is heated to 82^{o}C and then pumped through the spray nozzles under high

pressures. The hot air causes instantaneous drying of the liquid mix. The resultant powder falls to the bottom of the drying chamber from where it is collected.

3. PACKAGING

The final product, after completion of the drying and blending steps, is transferred from the storage to the filling machinery. The final powdered product is then filled in flexible containers which are filled with inert gas, sealed, coded and labelled. The finished product is not released until it undergoes final testing for microbiological analysis, nutrient contents and uniformity.

1.1.14.2 CRITICAL POINTS OF CONTAMINATION

The primary microorganisms associated with contamination of PIF, identified by the joint FAO/WHO meetings (2004, 2006) included *Cronobacter* spp., *Salmonella enteritidis*, *Enterobacter agglomerans*, *Hafnia alvei*, *Klebsiella pneumoniae*, *Citrobacter koseri*, *Citrobacter freundii*, *Klebsiella oxytoca*, *Enterobacter cloacae*, *Escherichia coli*, *Serratia* sp., *Acinetobacter* sp., *Bacillus cereus*, *Clostridium difficile*, *Clostridium perfringens*, *Clostridium botulinum*, *Listeria monocytogenes* and *Staphylococcus* sp. Any of these organisms if survive in PIF or extrinsically contaminates the product, have great chances of growth and multiplication upon reconstitution, as milk feeds act as an excellent media (Agostoni et al. 2004).

According to the joint FAO/WHO meetings (2004), critical points of contamination of PIF with above mentioned organisms are:

Contamination through the addition of heat sensitive ingredients such as amino acids, vitamins and fatty acids, during the dry-mixing and combined-mixing processes of PIF manufacturing. In dry-mixing process, there is no heat treatment process to kill the bacteria in the finished project. Therefore if one or more ingredients are contaminated with even a low level of bacteria then their presence in the finished product is likely.

In wet-mixing process, the spray drying method is used which requires the processing equipment including spray dryer itself and the fluidized bed to be regularly wet cleaned. This frequent wet cleaning may provide the moisture contents for bacterial growth. If the bacterial growth is not controlled, it can ultimately lead to the product contamination.

In all three processes of manufacturing, the environmental contamination of PIF may occur especially after thermal processing i.e. during drying or packing of the PIF.

It should be noted that the dry-mix ingredients added to PIF are not raw as they have to be processed by the suppliers and therefore contamination of PIF could to be the result of post-heat treatment.

The likelihood of contamination of the ingredients added to PIF vary and depends on the ingredient type, for example starches have higher risk of contamination with PIF than oil.

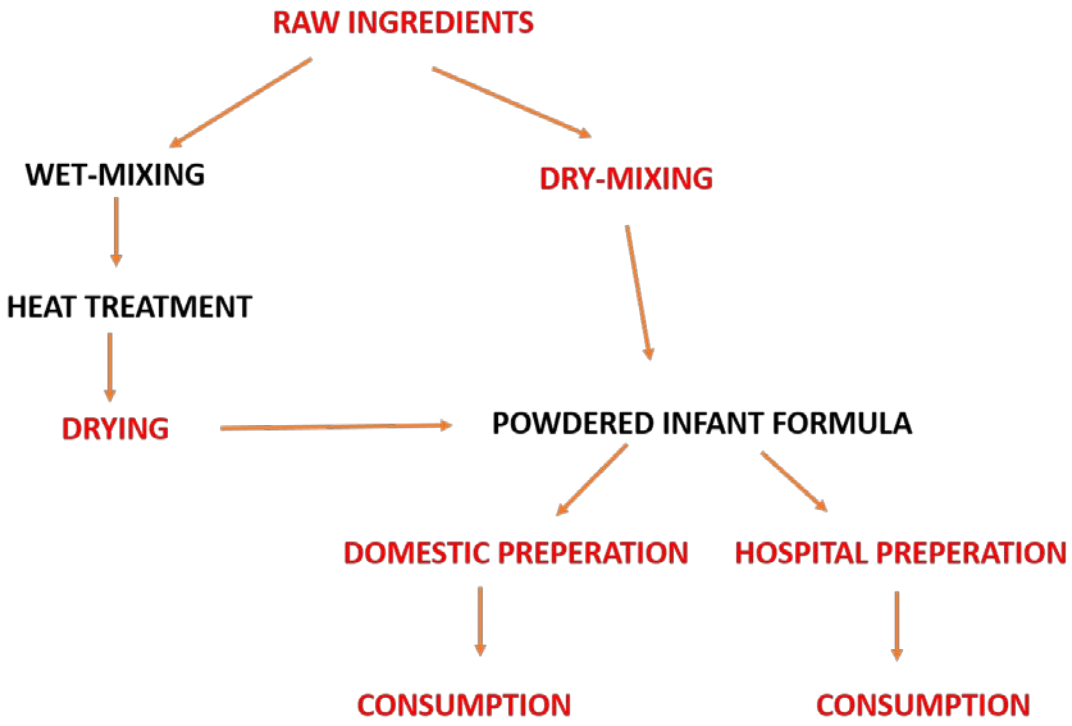


Figure 1. 2 Critical points of contamination of PIF manufacturing process (adapted from FAO/WHO, 2004).

The figure is the schematic representation of the process of PIF manufacturing highlighting critical points of contamination. The text in red font indicates potential critical points of contamination.

1.1.15 NEXT GENERATION SEQUENCING

Next generation sequencing (NGS) or massively parallel sequencing has revolutionised genetics and genomics. The availability of the high throughput instruments and the rapid development in the analysis programs has a huge impact on research in the field of genetics. The development of the NGS technology has contributed a wealth of information in recent years and continues to bring enormous changes in our understanding of the basic biological knowledge, genetics and biological research (Mardis, 2008, 2011). The constant and rapid development of the NGS has significantly dropped the cost of genome sequencing and laboratories with a moderate research budgets can afford to sequence small to large size genomes. However, although the low cost of genome sequencing has resulted in accumulation of sequencing data, it is the analysis of this sequencing data which is still challenging to the scientific community. Currently, there are 3 platforms which are in widespread use for the genome sequencing, illumina/Solexa genome analyser (Bentley, 2006) (http://www.illumina.com/), (Roche/454 FLX (Margulies et al. 2005) (http://www.454.com/)) and the Applied Biosystems SOLiD™ System (http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html). All of these 3 genome sequencing systems are based on the principal of massively parallel sequencing but differ in their methodology. Since all of the genomes in the current study were sequenced using the Illumina sequence by synthesis process, it is discussed briefly below.

1.1.15.1 ILLUMINA SEQUENCE BY SYNTHESIS, AN OVERVIEW

The Illumina sequencing technology is based on the cluster generation of the DNA to generate large sequencing data in a rapid and accurate manner. This technology has a wide range of applications not only in genomics but also in transcriptomics and epigenetics. Figure 1. 2 (adapted from Mardis, 2008) shows the schematic representation of the Illumina sequence by synthesis workflow. The genomic DNA is fragmented and the fragments are ligated with adapters to each single stranded DNA which is immobilised on the surface of the flow cell channels. Addition of the unlabelled nucleotides and DNA polymerase initiates the solid phase bridge amplification. The enzyme converts the single stranded nucleotide bridges to double stranded bridges by incorporating the unlabelled nucleotides. The denaturation of the double stranded nucleotide bridges leaves single stranded nucleotide

templates which are anchored on the solid substrate. The amplification of the single stranded templates generates millions of double stranded DNA fragments in each channel of the flow cell. The sequencing cycle begins when a labelled deoxynucleoside (dNTP) is added, the label acts as the reversible terminator therefore after the incorporation of each nucleotide, the florescent dye is imaged to identify the base and cleaved enzymatically which allows the incorporation of the next dNTP. The base calls are made directly from the signal light intensity for each single base in each sequencing cycle, therefore yielding a high quality base by base sequencing. In paired end sequencing, the sequence by synthesis cycle is repeated for both forward and reverse strands of the DNA. The forward and reverse reads are paired to create a contiguous sequence. This contiguous sequence can then be aligned against the reference genome for variant identification or can be used for any other genomic analysis based on the individual needs and the research question being answered (Mardis, 2008).

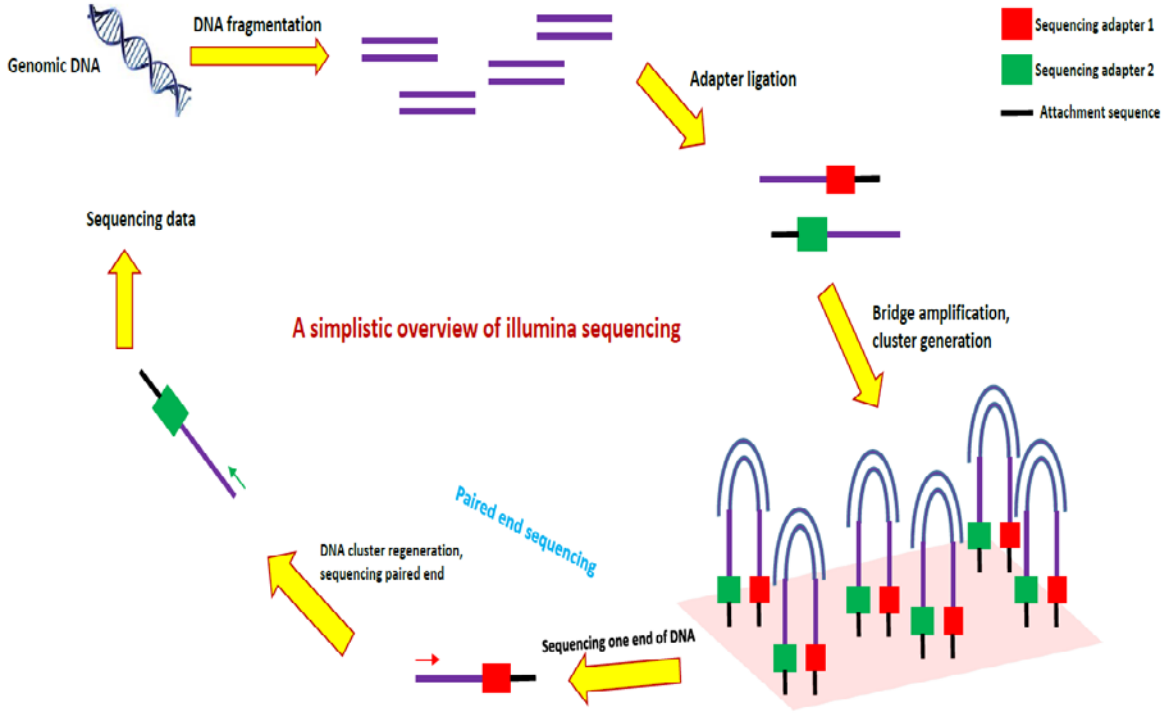


Figure 1. 3 Schematic representation of the steps involved in Illumina sequence by synthesis.

Genomic DNA is fragmented and ligated to adapters. Single stranded fragments are bound to the inside surface of the flow cell. The addition of unlabelled nucleotides and enzyme initiates the solid phase bridge amplification. The double stranded DNA are denatured to leave the single stranded fragments attached to the flow cell. First sequencing cycle begins when all 4 labelled dNTPs, primers and DNA polymerase are added. When laser light is bombarded the dye emits fluorescence which is captured for each cluster. Half way through all the cycles, the ends are switched. Paired end sequencing generates reads for both forward and reverse strand. The forward and reverse reads are paired to generate a contiguous sequence, which can then be aligned with a reference isolate for variant identification.

To date Illumina is dominating the genome sequencing industry and the technology is adapted by the institutions across the world. The sequencing technology is continually improving and different platforms are available according to individual sequencing needs including (GAIIx), Illumina MiSeq, Illumina HiSeq (1000, 2000, 2500) and Next Seq 500. The Illumina GAIIx systems is largely replaced by the recent systems like MiSeq and HiSeq.

1.2 AIMS AND OBJECTIVES OF THE STUDY

The *Cronobacter* spp. is an emerging food borne pathogen which has been associated with neonatal infections and has caught the attention of researchers especially over the last 2 decades. Although not all the members of *Cronobacter* are pathogenic, one particular lineage known as *C. sakazakii* CC4 which consists predominantly of *C. sakazakii* ST4 has been associated frequently with neonatal meningitis. The association of *C. sakazakii* CC4 with neonatal meningitis and its persistence in the environment especially PIF has been established by previous studies in our group. There have been several outbreaks of *Cronobacter* which have claimed the lives of a number of new born babies, the largest outbreak was in France in 1994 in a NICU which claimed the lives of 3 babies; all of the babies who died were infected with *C. sakazakii* CC4. Therefore it was of importance to identify any unique genes within this lineage. To achieve this, several *C. sakazakii* isolates including a number of *C. sakazakii* CC4 strains have been sequenced in collaboration with the University of Exeter UK; these isolates have been sequenced during different phases of the project. The study was divided into three parts, the first part was to identify unique genes in *C. sakazakii* CC4 using different comparative genomic approaches. The second part was to study the diversity within *C. sakazakii* CC4 and the third part was aimed at the genomic dissection of the 1994 French outbreak where majority of the isolates were *C. sakazakii* CC4, in order to precisely determine the source of outbreak.

The present study had 3 main aims:

1. **COMPARATIVE GENOMICS OF *CRONOBACTER SAKAZAKII* CC4 AND *CRONOBACTER SAKAZAKII* NON-CC4 GENOMES**

   The comparative genomics had the following objectives

   - Whole genome alignment.
   - Whole genome phylogeny determination.

- Genome screening of *C. sakazakii* CC4 and *C. sakazakii* non-CC4 genomes for the presence or absence of virulence and environmental fitness associated traits which have been described by previous studies in order to observe any CC4 specific pattern.
- Core genome extraction.
- Identification of the genes specific to the *C. sakazakii* CC4 core genome.
- Identification of the unique genes in the pan *C. sakazakii* CC4 genome based on the comparative genomics of the oldest *C. sakazakii* CC4 strain 377 against *C. sakazakii* non-CC4 strain ES15 using a gene by gene search approach.
- Fragmented genomic analysis of the *C. sakazakii* CC4 and *C. sakazakii* non-CC4 genomes using Gegenees to identify CC4 specific and CC4 missing traits.

The results of the above genomic analyses have been presented in chapters 3, 4, 5 and 6.

2. DIVERSITY WITHIN *C. SAKAZAKII* CC4

The analysis was divided into following main objectives

- Study the diversity of sequenced *C. sakazakii* CC4 isolates using Single Nucleotide Polymorphism.
- Genome comparison of high and low *C. sakazakii* CC4 genomes.
- The comparative analysis of the *C. sakazakii* CC4 isolates using *Cronobacter* BIGsDB.

The results of the above genomic analysis have been presented in chapter 7.

3. GENOMIC ANALYSIS OF *CRONOBACTER SAKAZAKII* STRAINS ISOLATED DURING OUTBREAK IN A NEONATAL INTENSIVE CARE UNIT IN FRANCE, 1994

The analysis was divided into following objectives

- Whole genome phylogeny determination to observe overall strains relatedness of all sequenced *C. sakazakii* isolates.
- SNP analysis of each cluster using the earliest isolate to observe strain relatedness in each cluster in order to identify potential source of origin.

The results of the above analysis is presented in chapter 8.

NOTE: This project was undertaken in parallel with two other laboratory based PhD projects;
1) Sumyya Hariri whose main objective was to identify virulence associated traits in *C. sakazakii* with special emphasis on *C. sakazakii* CC4 using laboratory experiments.

2) Hana Sonbol whose project was aimed at physiological profiling of *C. sakazakii* with special emphasis on *C. sakazakii* CC4 using laboratory experiments.

Therefore where appropriate references will be made for these parallel laboratory studies. In addition, references to the previous laboratory based PhD studies (Nasreddin Rhouma 2012; Faisal Almajed 2014) will also be made where needed.

Secondly, the sequence in which the results are presented in this thesis may not represent the chronological order in which the analysis was conducted due to the nature of the study.

CHAPTER 2

MATERIALS AND METHODS

2 MATERIALS AND METHODS

The current study was mainly aimed at the genomic analysis of the sequenced *Cronobacter sakazakii* isolates and there was only a minimal amount of laboratory work involved. Most of the analysis was undertaken using available online genomic tools and by following their manuals. The Material and Methods sections is divided into laboratory based methods and computer based methods.

2.1 LABORATORY BASED METHODS

2.1.1 MEDIA PREPERATION

2.1.2 TRYPTICASE SOY AGAR (TSA)

The TSA (casein soya bean digest agar) is a general purpose media that supports the growth of wide variety of organisms, both aerobic and anaerobic. The components of TSA include agar, pancreatic digest of casein, enzymatic (papain) digest of soya bean and sodium chloride. The TSA media in its powdered state was purchased from Oxoid Thermo Scientific UK. To prepare the media, 20g of the powdered media was dissolved in 500ml of sterile distilled water. The media was sterilised by autoclaving it for 15min at 121^{o}C. The sterile prepared media was allowed to cool to 50^{o}C and approximately 20ml was poured into each sterile petri dishes. The TSA media plates were stored at 4^{o}C for up-to 3 weeks. The TSA plates were always dried before culturing the bacterial strains on them.

2.1.3 TRYPTICASE SOY BROTH (TSB)

The TSB media was purchased from Oxoid Thermo Scientific UK. To prepare the TSB, 30g of the base was dissolved in 1L of the sterile distilled water and autoclaved for 15min at 121^{o}C. The prepared broth was stored at 4^{o}C.

2.1.4 STOCK SOLUTION PREPARATION FOR METAL RESISTANCE ASSAY

The 1M stock solution of copper (II) sulphate (Sigma-Aldrich®, UK), sodium arsenate (Sigma-Aldrich®, UK), nickel chloride (Fisher Scientific, UK), cobalt (II) nitrate (Sigma-Aldrich®, UK) silver nitrate (Alfa Aesar®, UK), zinc sulphate (BDH chemicals England), sodium tellurite (Sigma-Aldrich®, UK) and cadmium carbonate was prepared by dissolving 159.60g, 180.03g, 129.60g, 182.94g, 169.87g, 161.47g, 221.58g and 172.42g respectively into 1L of distilled water (molecular weight equivalent of the substance into 1L of distilled water). Further

dilutions of 0.1M (100mm), 0.01M (10mm) and 0.001M (1 mm) were prepared from 1M (1000mm) stock solution.

2.1.5 BACTERIAL CULTURING AND STORAGE

All *Cronobacter sakazakii* strains used in this study were maintained on TSA and stored at -20^{o}C for short term storage and -80^{o}C for long term storage. When needed the cultures were recovered from the frozen stock by subculturing them on TSA at 37^{o}C for 24h.

2.1.6 METAL RESISTANCE ASSAY

Selected *C. sakazakii* isolates were tested for their ability to tolerate metals (toxic to bacteria at high concentrations). Isolates were tested against 8 different metals which included copper (II) sulphate (Sigma-Aldrich®, UK), sodium arsenate (Sigma-Aldrich®, UK), nickel chloride (Fisher Scientific, UK), cobalt (II) nitrate (Sigma-Aldrich®, UK), silver nitrate (Alfa Aesar®, UK), zinc sulphate (BDH chemicals England), sodium tellurite (Sigma-Aldrich®, UK) and cadmium carbonate (Harrington, USA). Each isolate was tested for 4 different concentrations; 1M (1000 mm), 0.1M (100mm) , 0.01M (10mm) and 0.001M (1mm) of each heavy metal.

Each isolate was streaked onto TSA and incubated for overnight at 37^{o}C. At least 5 colonies were picked from the TSA for each isolate and grown in TSB at 37^{o}C for overnight in a shaking incubator. A sterile cotton swab was soaked in the TSB culture of each isolate and streaked all over the TSA. Four paper discs were placed equidistant on the TSA plate and pressed gently with the flame sterilised forceps so that each disc is in proper contact with the culture on the media. Using a pipette, 7µl of each dilution was added onto the centre of each paper disc. The inoculated TSA plates were grown for overnight at 37^{o}C and the zone of inhibition was measured. Each experiment was conducted in triplicate.

2.1.7 EXTENDED SPECTRUM BETA LACTAMASES (ESβL) SUSCEPTIBILITY TESTING

Bacterial resistance against the β-lactam antibiotics including penicillin and extended spectrum cephalosporin is mediated through the production of extended spectrum beta lactamases (ESβL) (Beceiro et al. 2004). The ESβL are the enzymes produced by members of Enterobacteriaceae family particularly *E. coli* and *Klebsiella pneumoniae* as well as *Acinetobacter baumannii* and *Pseudomonas aeruginosa*. The enzymes act by hydrolysing the β-lactam rings of penicillin and extended spectrum cephalosporin, before they can

access the penicillin binding sites at the cytoplasmic membrane. Commonly expressed ESβL include TEM, S CTX-M and SHV, which in most cases are plasmid borne and confer resistance against 3rd generation cephalosporin such as cefpodexime, cefoxitin and cefotetan (Beceiro et al. 2004). The infections caused by ESβL producing organisms are usually treated by carbapenem antibiotics. Clavalunic acid is the inhibitor of ESβL (Beceiro et al. 2004).

The AmpC beta-lactamases are bacterial enzymes which confer resistance to 3rd generation extended spectrum cephamycins and cephalosporins (Perez-Perez et al. 2002). The production of these enzymes is generally the result of hyper-production or de-repression of the chromosomally encoded AmpC when a plasmid-mediated *ampC* gene is acquired (Perez-Perez et al. 2002; Nasim et al. 2004). Like ESβL, infections caused by AmpC producing bacteria are also treated by carbapenem antibiotics. AmpC are inhibited by class C inhibitors such as cloxacillin and boronic acid (Perez-Perez et al. 2002; Nasim et al. 2004).

Additionally, bacterial resistance against carbapenem and other β-lactam antibiotics is also mediated through the production of carbapenemases (Babini et al. 2000; Coudron et al. 2000). Carbapenemases are the enzymes which hydrolyse antibiotics such as carbapenems, cephamycins and oxyiminocephalosporins. Metallo-β-lactamases (MBL) belong to class B carbapenemases which can efficiently hydrolyse carbapenems antibiotics. The MBLs need zinc as a co-factor for their catalytic activity. Besides Enterobacteriaceae, the MBLs have been detected in *A. baumannii* and *P. aeruginosa*. Chelating agents such as EDTA are the inhibitors of MBLs (Babini et al. 2000; Coudron et al. 2000)

In the present study, screening of *C. sakazakii* isolates for ESβL and/or AmpC was undertaken using D68C[1] test of Mast discs detection system. The D68C[1] is used to detect the ESβL positive strains, de-repressed or hyper-produced AmpC positive strains as well as the co-production of ESβL and AmpC. The presence of ESβL and/or AmpC can easily be determined through zone size comparsions using antibiotic (cefpodoxime) plus antibiotic inhibitor. The antibiotic inhibitors were not specified in the flier supplied by manufacturer since it was a proprietary information. However, as stated earlier, clavulanic acid is the general inhibitor of ESβL while AmpC is inhibited by cloxacillin and boronic acid.

The assay was performed following the guidance of British Society for Antimicrobial Chemotherapy (BSAC, version 12, May 2013, http://bsac.org.uk/wp-

content/uploads/2012/02/Version-12-Apr-2013_final.pdf/). The *C. sakazakii* isolates were streaked onto the TSA and incubated overnight at 37°C. At least 5 colonies were picked up from the TSA and transferred to 5ml of sterile saline and mixed thoroughly using a vortex mixer. The optical density (OD) of the suspension was then adjusted to 0.5 McFarland standard. After adjusting the OD, a cotton swab was soaked in the suspension and gently streaked over the TSA. The TSA plates were allowed to dry for approximately 5 min. The ESβL and AmpC discs which were purchased from the MAST Group limited (UK) were placed equidistant on the TSA plates and pressed gently with flame sterilised forceps to ensure complete contact with the TSA media. The TSA plates with ESβL and AmpC discs were incubated at 37°C for 24h. For each disc the diameter of the zone was measured and compared with guidance image (D68C[1], Figure 2.1) provided by the Mast group.

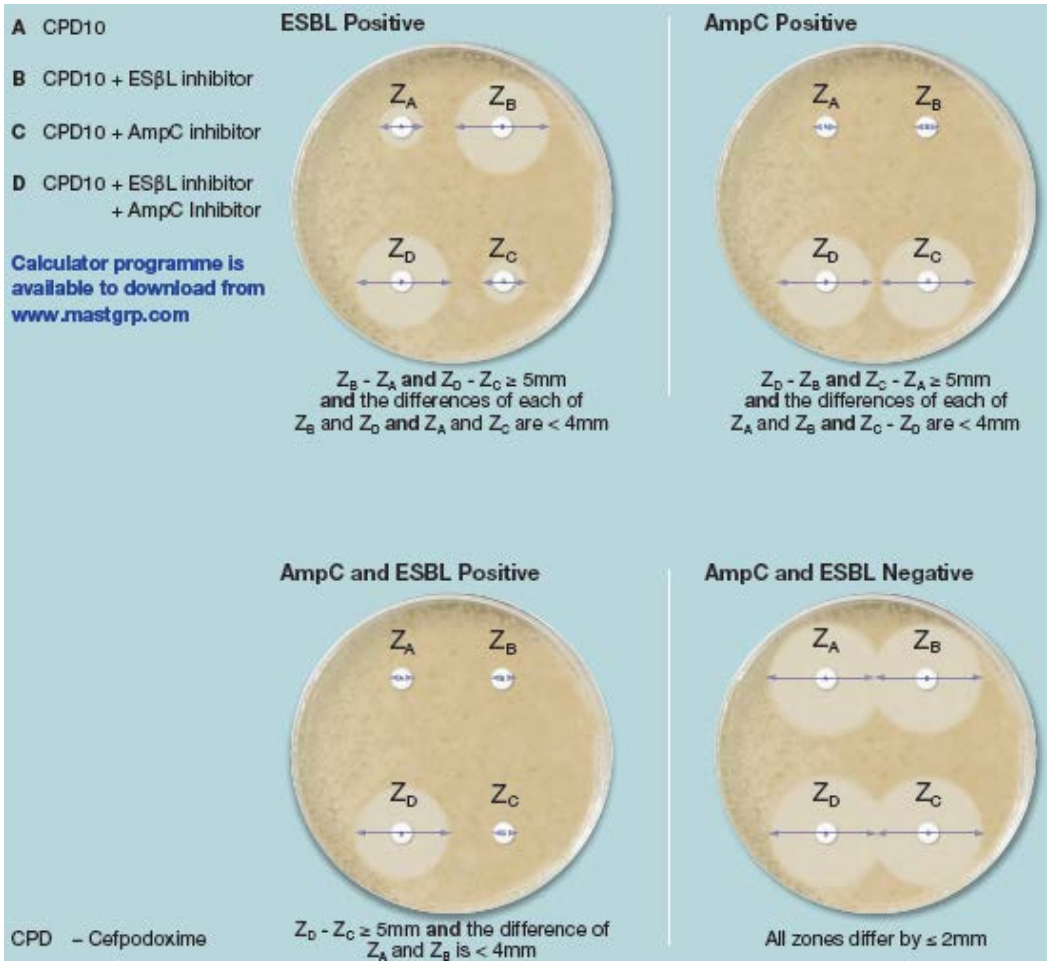


Figure 2. 1 Guide to ESβL and AmpC resistance and sensitivity (source; http://www.mastgrp.com/).

The image above is a cropped section showing guidance for the ESβL and AmpC resistance and guidance. After performing the disc diffusion assay, the guide was consulted to decide the ESβL and AmpC resistance and sensitivity characteristic of the isolate.

2.1.8 GENOMIC DNA EXTRACTION

To prepare the genomic DNA, bacterial cultures were grown in 1.5ml of TSB for overnight at 37^{o}C in the shaking incubator. Genomic DNA extraction of the overnight grown cultures was undertaken using GenElute™ Bacterial Genomic DNA Kit (Sigma Aldrich®, UK) following the manufacturer's guide. Using a Nanodrop 2000 (Thermo Scientific, UK) the purity and concentration of the genomic DNA was checked to ensure each DNA sample has a minimum 260/280nm values of 1.8 and 260/230nm values of 2.0 before sending them for sequencing.

2.1.9 GENOME SEQUENCING

Bacterial DNA was extracted from 1-day old cultures using GenElute™ bacterial genome kit (Sigma Aldrich®, USA) using the manufacturer's protocol. A total of 50 *C. sakazakii* genomes were sequenced during this project using different versions of illumina (GAIIx, HiSeq and MiSeq) sequencing technology (Table 2.1). All of the strains were sequenced by Exeter Sequencing Service, Wellcome Trust Biomedical Informatics Hub, University of Exeter, UK. All of the isolates were paired end sequenced. Paired end sequencing involves sequencing both ends of a fragment thus generating high quality alignable data. Paired end sequencing facilitates the detection of repetitive sequences, gene fusions, genetic rearrangements as well as novel transcripts. Paired end sequences are more likely to align with the reference thereby enhancing the quality of the entire data set. Sequencing both ends of a fragments allows long range positional information to achieve high quality alignment of reads. The superior alignment across repetitive DNA regions helps generate longer contigs in *de novo* assemblies by filling the gaps in consensus sequences. Longer and less number of contigs mean fewer scaffolds with less gaps thus improving the overall quality of a draft bacterial genome (http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html).

2.2 Computer based methods

The genome analysis workflow is given in figure 2.1.

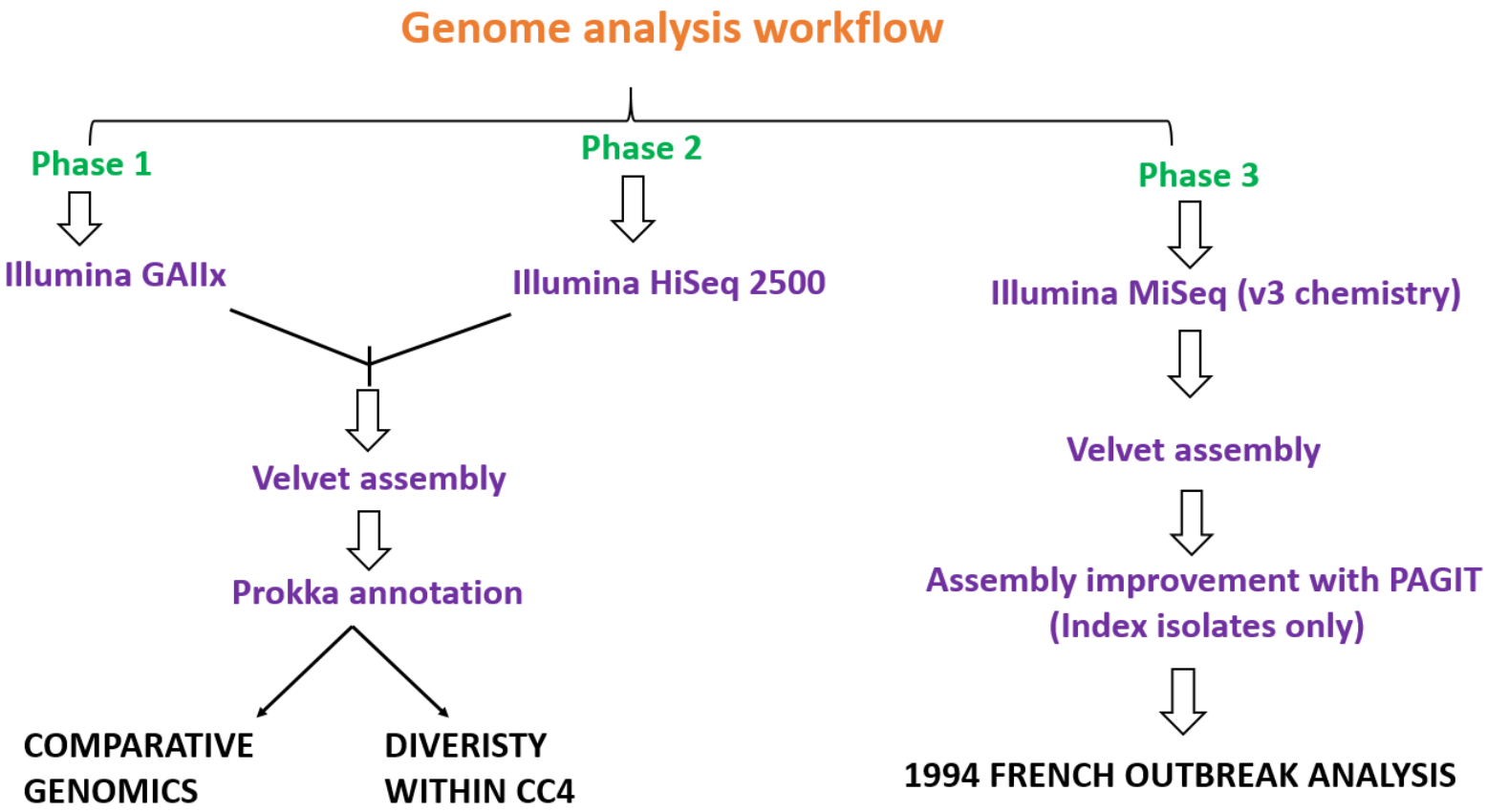


Figure 2. 2 The schematic representation of the genome analysis workflow.

The genome sequencing was undertaken in 3 phases during the course of this PhD study. Genome assembly was mainly undertaken using Velvet and improved using PAGIT for selected isolates. The genome annotation for all isolates was undertaken using Prokka.

2.2.1 SELECTION OF THE ISOLATES AND SEQUENCING PLATFORM

Due to the nature of study different genomes were sequenced at different stages. In total there were three sequencing phases. Genome sequences in phase 1 and 2 have been analysed and presented in chapters 3-7 while phase 3 genomes have mainly been used in the 1994 French outbreak analysis presented in chapter 8. Phase 1 consisted of 23 *C. sakazakii* strains (No. 1-14, 19-21, 23-26, 28; Table 2.1) which were sequenced at the beginning of this project using Illumina GAIIx (RTA version: 2.8) generating 36 bp, paired end reads.

In phase 2, *C. sakazakii* CC4 isolates 377 and 1587 were genome sequenced (No. 15, 16; Table 2.1) on Illumina HiSeq 2500 (CASAVA 1.8.2, RTA 1.13) generating paired end reads of 150 bp.

In phase 3 the genome sequences of 25 *C. sakazakii* strains were generated on Illumina MiSeq using v3 chemistry and 300bp paired end reads using dual indexed Nextera XT libraries. The mean insert size was around 250-300bp (No. 30-54; Table 2.1).

Additional genomes which included *C. sakazakii* BAA-894 (NC_009778-80), *C. sakazakii* SP291 (NC_020260-63), *C. sakazakii* ES15 (NC_017933.1), *C. sakazakii* 696 (NZ_CALF00000000.1) and *C. sakazakii* 680 (NZ_CALG00000000.1) were accessed from NCBI for comparative purposes.

Please note that *C. sakazakii* CC4 isolates 701 and non-CC4 isolate 696 (No. 30 and 45; Table 2.1) were initially accessed from Genbank for the analysis presented in chapters 3-7; however these were re-sequenced as part of the French outbreak 1994 analysis (phase 3) to obtain their raw files (FASTQ) for SNP analysis (chapter 8). Similarly, *C. sakazakii* CC4 isolate 767 was sequenced in phase 1 and was also included in the analysis presented in chapter 8 since it was originally isolated from 1994 French outbreak (No. 5; Table 2.1).

2.2.2 ACCESSION NUMBERS

The whole gnome shotgun projects for all 50 *C. sakazakii* genomes have been submitted in Genbank by following the instructions given by NCBI (http://www.ncbi.nlm.nih.gov/genbank/genomesubmit/) and the accession numbers are given in table 2.1.

Table 2. 1 Sequenced strains used in this PhD study

No	*C. sakazakii* isolate	ST	CC	Sequencing phase	Country	Genbank accession	Date	Source	Extra information	Coverage	N50
1	20	4	4	Phase 1	Czech Republic	JNBH00000000	2003	Clinical; Faeces	Adult, 74 years old	13	54736
2	553	4	4	Phase 1	Netherlands	JNBJ00000000	1977	Clinical	1 day infant	16	70973
3	557	4	4	Phase 1	Netherlands	JNBK00000000	1979	Clinical	4 day old infant	9	16011
4	558	4	4	Phase 1	Netherlands	JNBL00000000	1983	Clinical		46	217574
5	767	4	4	Phase 1	France	JNCX00000000	1994	Trachea	Fatal meningitis	16	130723
6	1219	4	4	Phase 1	USA	JNCY00000000	2009	Clinical	Fatal meningitis	12	264907
7	1220	4	4	Phase 1	USA	JNDH00000000	2003	Clinical; CSF		14	305405
8	1221	4	4	Phase 1	USA	JNDI00000000	2003	Clinical; CSF		14	126619
9	1225	4	4	Phase 1	USA	JNDF00000000	2007	Clinical; Blood	Fatal meningitis	11	54922
10	1240	4	4	Phase 1	USA	JNDD00000000	2009	Clinical; CSF		16	228799
11	721	4	4	Phase 1	USA	JNDA00000000	2003	Clinical; CSF		13	26058
12	1231	4	4	Phase 1	New Zealand	JNDG00000000	2005	Clinical; Faeces	Infant with meningitis	15	143701
13	6	4	4	Phase 1	Canada	JNBG00000000	1990	Clinical		68	42403
14	4	15	4	Phase 1	Canada	JMSR00000000	1990	Clinical		18	32494
15	377	4	4	Phase 2		AWFW00000000	1950	Milk powder		103	339411
16	1587	10 9	4	Phase 2	Israel	AWSP00000000	2000	CSF isolate	Fatal infant isolate	91	632792
17	SP291	4	4	Genbank	Ireland	CP004091-4	Unknown	PIF manufacturing plant			
18	BAA-894	1	1	Genbank	USA	NC_009778-80	2001	Non-infant formula	Fatal meningitis		
19	1218	1	1	Phase 1	USA	JNDC00000000	2001	CSF	Fatal meningitis	13	241193
21	5	8	8	Phase 1	Canada	JNBF00000000	1990	Clinical		17	63514
22	680	8	8	Genbank	USA	CALG01000001-201	1977	CSF			

Table 2.1 continued

No	*C. sakazakii* isolate	ST	CC	Sequencing phase	Country	Genbank accession	Date	Source	Extra information	Coverage	N50
23	140	40	45	Phase 1	India	JMSS00000000	2005	Spice	Ajwan cumin seed	22	85643
24	150	16	16	Phase 1	South Korea	JNDB00000000	2005	Spice	Yoojeon food spices	14	50208
25	978	3	3	Phase 1	UK	JNCZ00000000	2007	Clinical	Enteral feeding tubes	12	232916
26	520	12		Phase 1	Czech Republic	JNBI00000000	1983	Clinical		16	67964
27	1249	31		Phase 1	UK	JNDE00000000	2010	Clinical	Fatal infant isolate	18	179201
28	ES15	125	100	Genbank	South Korea	PRJNA81953	Unknown	Ground whole grain			
29	730	4	4	Phase 3	France	JOLT00000000	1994	Trachea	Meningitis (DIED)	12	105211
30	701*	4	4	Genbank/ Phase 3	France	NZ_CALE00000000	1994	stools	NEC I	14	131008
31	691	4	4	Phase 3	France	JOLQ00000000	1994	Sputum	No-details	13	210466
32	692	4	4	Phase 3	France	JOLV00000000	1994	stools	NEC II	17	164025
33	694	4	4	Phase 3	France	JOLM00000000	1994	Conjunctivae	NEC II	26	257787
34	695	4	4	Phase 3	France	JOLG00000000	1994	Trachea	NEC II (DIED)	14	127060
35	698	4	4	Phase 3	France	JOLR00000000	1994	Trachea	No-details	15	164033
36	702	4	4	Phase 3	France	JOLF00000000	1994	Trachea	NECII	13	74465
37	705	4	4	Phase 3	France	JOLI00000000	1994	Stools	NEC II	15	161660
38	706	4	4	Phase 3	France	JOLH00000000	1994	Skin	NEC II	10	75457
39	707	4	4	Phase 3	France	JOLD00000000	1994	Trachea	No-symptoms	17	178992
40	709	4	4	Phase 3	France	JOLJ00000000	1994	stools	No-symptoms	14	137377
41	711	4	4	Phase 3	France	JOLC00000000	1994	Prepared formula	NEC II	17	210291
43	690	12		Phase 3	France	JOLN00000000	1994	stools	No-symptoms	12	54751
44	696*	12		Genbank /Phase 3	France	JOLW00000000	1994	stools	NECII	15	127060
45	699	12		Phase 3	France	JOLK00000000	1994	stools	Digestive problems	14	223653
46	703	12		Phase 3	France	JOLY00000000	1994	Trachea	NEC II	12	198512

Table 2.1 continued

No	*C. sakazakii* isolate	ST	CC	Sequencing phase	Country	Genbank accession	Date	Source	Extra information	Coverage	N50
47	708	12		Phase 3	France	JOLZ00000000	1994	Trachea	Septicaemia	14	225681
48	693	13	13	Phase 3	France	JOLO00000000	1994	stools	No-symptoms	15	134695
49	713	13	13	Phase 3	France	JOLX00000000	1994	Bottle		17	166211
50	714	13	13	Phase 3	France	JOLU00000000	1994	Prepared formula		15	187343
51	715	13	13	Phase 3	France	JOLP00000000	1994	Infant formula	No-symptoms	13	116348
52	700	86	13	Phase 3	France	JOLL00000000	1994	Peritoneal fluid	NEC III (DIED)	16	161684
53	716	14	1	Phase 3	France	JOLS00000000	1994	stools	NEC I	15	195858

ST; Sequence Type 4, CC; Clonal Complex, Isolates No 1 to 30 were used in the analysis presented in chapters 3, 4, 5, 6 and 7; Isolates No 30 to 54 were used in the analysis presented in chapter 7. Date; refers to isolation date, Isolate No 5 was sequenced in phase 1 but was used in the analysis presented in chapters 3-7 as well as chapter 8, since it was isolated from French outbreak 1994; *Initially isolates 701 and 696 were accessed from Genbank and used in analysis of chapters 3-7. These were re-sequenced for chapter 8.

2.2.3 GENOME ASSEMBLY

The initial *de novo* genome assembly was performed using Velvet (versions 1.2.02 and 1.2.09) (Zerbino & Birney, 2008) by Exeter Sequencing Services, UK and the assembled genome contigs for all of the strains were provided.

2.2.4 CONTIG RE-ORDERING OF THE SELECTED GENOMES USING PAGIT

The genome assembly of the strains which were used as index for Single Nucleotide Polymorphism (SNP) analysis (chapter 8) was improved using Post Assembly Genome Improvement (PAGIT) (Swain et al. 2012). The finished genome of *C. sakazakii* SP291 (Genbank accessions CP004091-4) was used as reference for contig reordering.

2.2.4.1 PRINCIPLE OF PAGIT

PAGIT aims to improve the assembly from the draft to a high quality which can then be used either for comparative genomics or for precise variant determinants in a genome. PAGIT consists of 4 open source programs (Figure 2.2) which can be used either individually or together; in this case all 4 programs were set to run automatically. These 4 programs are:

1. **ABACAS (Algorithm-Based Automatic Contiguation of the Assembled Sequences)**; It is an orientation tool that re-orders the contigs by aligning them against a reference genome. The reference genome should be of high quality. In the present study, the finished *C. sakazakii* SP291 was used as reference for contig re-ordering of the selected draft *C. sakazakii* genomes.
2. **IMAGE (Iterative Mapping and Assembly for Gap Elimination)** is a tool that tends to close the gaps between the contigs by extending the contig ends using the paired end sequencing information.
3. **ICORN (Iterative Correction of Reference Nucleotides)** is a tool that corrects the error in the base calls within the consensus sequence including single base pairs as well as small insertions and deletions by iterative mapping of the reads to the sequence.
4. **RATT (Rapid Annotation Transfer Tool)** is a synteny based algorithm which transfers the annotation to the draft genome from the reference genome.

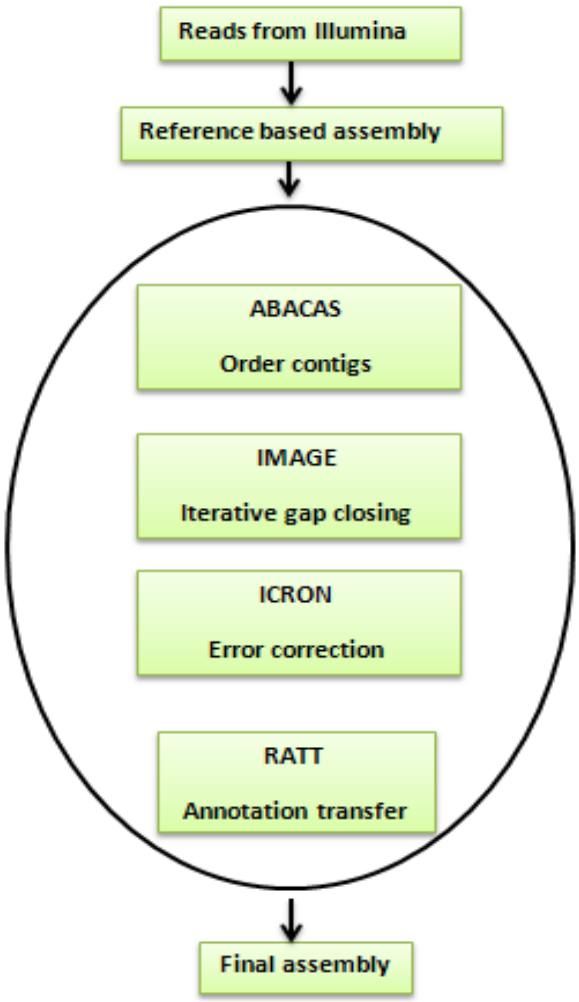


Figure 2. 3 The schematic representation of the workflow of PAGIT.

PAGIT is a combination of 4 open source computer programs; ABACAS re-orders the contigs of the draft genome by aligning them against the reference genome, IMAGE iteratively closes the gaps by extending the contig ends using information from the paired end sequences, ICORN corrects the error is base calls, insertions and deletions by iteratively mapping of the draft assemblies and finally RATT transfers the annotation from the reference onto the daft genome.

2.2.4.2 PAGIT METHODOLOGY

The program in its bundled form was downloaded from Sanger Institute website (https://www.sanger.ac.uk/resources/software/pagit/#Download/) with all the supporting scripts on a HP EliteBook (Linux X64bit) on a Unix 10.04 operating system. Once installed the program was executed from the relevant directory when running a genome assembly improvement following the instructions (Swain et al. 2012). The PAGIT yielded output data for each of the 4 components of the program in an individual folder in the sourced directory. In this case, the output was the EMBL file which was located within the RUNRATT folder.

The EMBL file was opened in Artemis (Rutherford et al. 2000) to visualise the improved genome assembly.

2.2.5 PROKKA GENOME ANNOTATION

2.2.5.1 PRINCIPLE OF PROKKA

The genome annotation refers to the identification and labelling of all the features on a sequenced genome (Richardson & Watson, 2013). There are several valuable annotation tools available such as RAST and Xbase2, however the genome annotation by Prokka has been found to be more accurate than the former 2 tools (Seemann, 2014). In addition it is faster and the output files can be easily analysed using web-based genome analysis tools like Artemis. In the present study, the genome annotation of the sequenced *C. sakazakii* isolates was undertaken using Prokka; a command line software tool which takes preassembled genomes as input and requires the BLAST+ to be preinstalled on the Unix system (Seemann, 2014). Prokka depends on the external tools for feature prediction which include Prodigal for predicting Coding sequence (CDS) (Hyatt et al. 2010), RNAmmer for Ribosomal RNA genes (rRNA) (Lagesen et al. 2007), Aragorn to predict Transfer RNA genes (Laslett & Canback, 2004), SignalP to predict Signal leader peptides (Petersen et al. 2011) and Infernal for non-coding RNA (Kolbe & Eddy, 2011).

2.2.5.2 PROKKA METHODOLOGY

The program was installed and executed using the instructions available on the Victorian Bioinformatics Consortium website (http://bioinformatics.net.au/prokka-manual.html/). The Prokka yielded outputs in FASTA (.fna, .fsa) and the annotation files (.gbk, .gff)
Both FASTA and annotation files were analysed and viewed using the online web-based tools such as Artemis and Artemis Comparison Tool (ACT).

2.2.6 ARTEMIS

2.2.6.1 PRINCIPLE OF ARTEMIS

One of the frequently used genomic tool throughout this study was Artemis, hosted at the Sanger Institute (http://www.sanger.ac.uk/resources/software/artemis/#download). Artemis is a genomic tool that can be used for the visualisation and annotation of the compact bacterial, archaeal or other lower eukaryotic genomes (Rutherford et al. 2000;

Carver et al. 2008, 2012). Artemis can read the annotated genome files either in EMBL, Genbank or FASTA format. The Artemis manual containing the guidelines for using the program was used (ftp://ftp.sanger.ac.uk/pub/resources/software/artemis/artemis.pdf).

2.2.6.2 ARTEMIS METHODOLOGY

In this study, Artemis was frequently used to visualise *C. sakazakii* genomes at particular regions of interest, to observe G+C and to save the sequence of any gene of interest in the FASTA format. BLAST searches under "Run" option were used to perform homology searches for any trait of interest. The "DNA plotter" feature within Artemis was used to mark any feature on the genome and to represent it on a circular genomic map, for example in marking of SNPs. Artemis was also used to visualise the BAM (Binary alignment/Map) in conjunction with the VCF (Variant Call Format) files to manually curate the SNP calls (section 2.2.9).

2.2.7 ACT and WebACT

2.2.7.1 PRINCIPLE OF ACT and WebACT

Another important tool which has been used frequently in analysis of the sequenced genomes in the current study was Artemis Comparison Tool (ACT) developed by the researchers at Sanger institute (Carver et al. 2005, 2008) and hosted at their website (https://www.sanger.ac.uk/resources/software/act/). A web version of the ACT called WebACT has been developed and is hosted at the Imperial College London (http://www.webact.org/WebACT/home).

2.2.7.2 METHODOLOGY OF ACT and WebACT

To compare genomes in ACT, a comparison file was needed for the two genomes which are being compared; the comparison file was created using WebACT simply by right clicking the alignment and saving the comparison file of the given two genomes to be compared. Both ACT and WebACT are essentially the same programs, the difference is that the former can be installed on a computer and does not need an active internet connection while the latter is only available online and requires an active connection to the internet. Once the genome files were chosen and submitted, the files were then uploaded to the server allowing comparison to be launched via ACT (Figure 2.3).

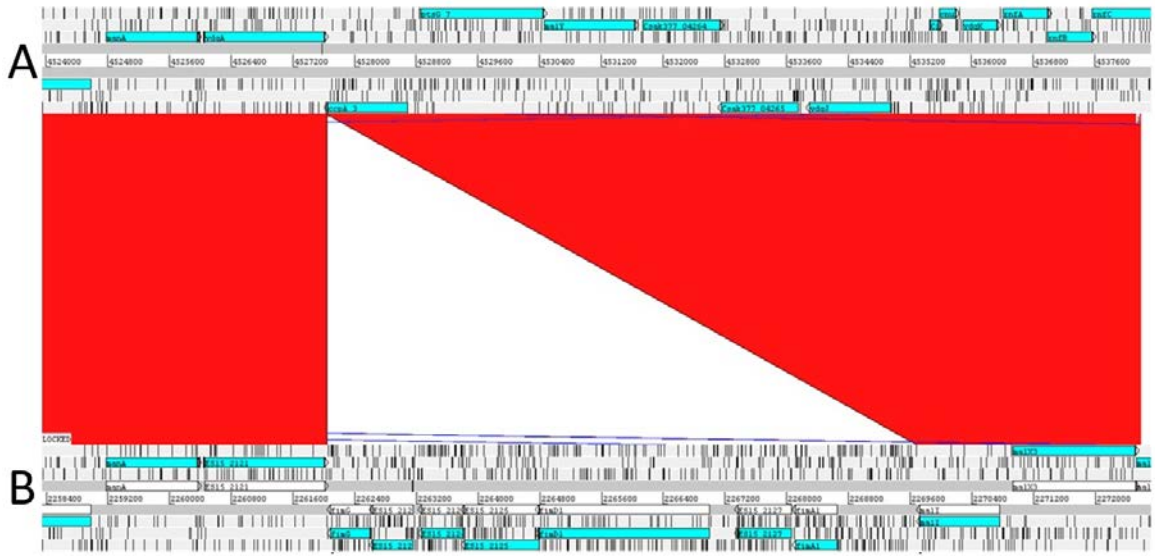


Figure 2. 4 Example screenshot of the ACT genome alignment.

The screenshot above is an example of the two genomes A and B viewed in ACT for comparative analysis. The image clearly show a genomic region absent in A.

The program can be used for several different genomic tasks such as viewing, annotation, BLAST searching for genes of interest, depending on the user requirement. However in the present study these programs were primarily used for the comparative genomic purposes which mainly involved a manual gene by gene search to identify the genes present in a particular genome compared to another genome and has been used throughout this study. Any gene of interest was searched for homology using the BLAST search feature from the "Run" menu. Any additional information about using the program was accessed online (ftp://ftp.sanger.ac.uk/pub/resources/software/act/act.pdf).

2.2.8 DETERMINATION OF STATISTICAL SIGNIFICANCE

The statistical significance for the presence/absence of selected genes within two sub-sets of *C. sakazakii* isolates i.e. CC4 and non-CC4, was determined by Fisher's exact test (one tailed chi-square analysis) using the software called IBM SPSS Statistics 22.

2.2.9 PHYLOGENETIC ANALYSIS

The phylogenetic analysis of 30 *C. sakazakii* genomes presented in chapter 3 and 26 French outbreak isolates presented in chapter 8 was undertaken in 3 steps

1. Whole genome alignment using Mugsy
2. Core genome extraction

3. Construction of maximum likelihood phylogeny using RaxML

Each of the above steps is described below.

2.2.9.1 PRINCIPLE OF WHOLE GENOME ALIGNMENT USING MUGSY

The whole genome alignments are important in studying bacterial diversity. The present study was aimed to describe the genetic diversity of different Sequence Types (ST) of *C. sakazakii* strains. In the present study, a computationally efficient, fast and accurate tool called Mugsy was used for the whole genome alignment of *C. sakazakii* genomes (Angiuoli & Salzberg, 2011). Mugsy uses Nucmer which is a whole genome pairwise aligner to identify homology amongst genomes. Mugsy accepts a set of multiple FASTA files as input and invokes through a single command line. It produces the output as a multiple alignment file (MAF). The program was used for the whole genome alignment of multiple *C. sakazakii* genomes; 30 *C. sakazakii* genomes (chapter 3) and 26 *C. sakazakii* genomes (chapter 8).

2.2.9.2 METHODOLOGY

The program was downloaded from the website (http://mugsy.sf.net) on HP EliteBook (Linux X64bit) using UNIX (version 10.04) operating system. The whole genome phylogeny was estimated using 3 steps as described below.

2.2.9.2.1 WHOLE GENOME ALIGNMENT

The genome sequences of the isolates to be aligned were copied to the appropriate folder and the alignment was executed using the command line (Appendix text 1.1). The alignment output in the MAF format was used for the core genome extraction.

2.2.9.2.2 CORE GENOME EXTRACTION

Once the multiple alignment (MAF) of *C. sakazakii* genomes was obtained, the core genome was extracted using mothur (Schloss et al. 2009) through a pipeline of scripts written in Python by Jason Sahl and adapted by Alan McNally (Sahl et al. 2011; McNally et al. 2013). The algorithms used to extract the core genome are given in the Appendix text 1.1. The core genome alignment was produced as a double filtered concatenated file, "concat_alignment.fas" which was then viewed using SeaView (http://doua.prabi.fr/software/seaview), an example is shown in Figure 2.4. The alignment was saved in the Phylip (Phylogeny Inference Package) format which was then used for phylogeny estimation using RaxML (Stamtakis et al. 2005).

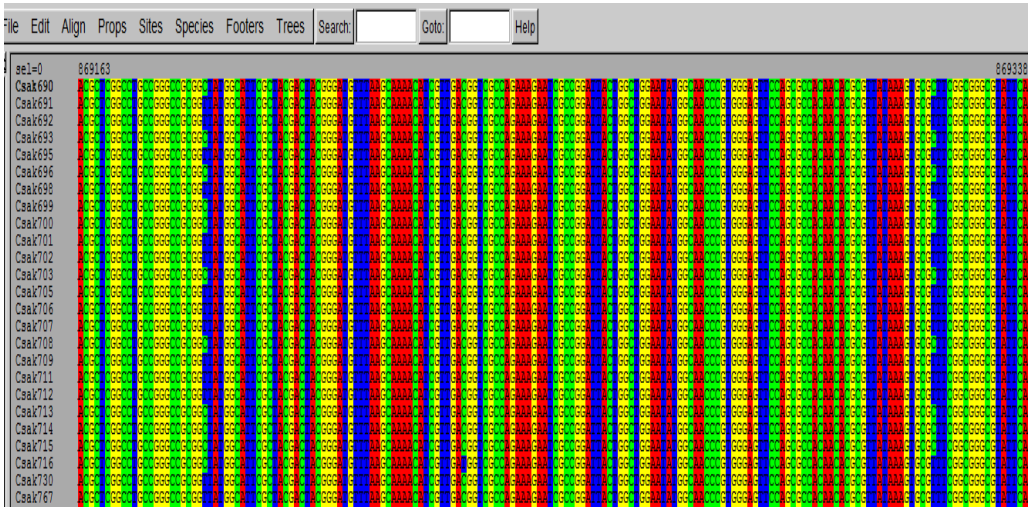


Figure 2. 5 Example image of the whole genome alignment visualisation in SeaView.

The image is an example of the screenshot showing the core genome alignment contained in concatenated alignment file "concat_alignment.fas". SeaView allows the conversion of concat_alignment.fas to Phylip (Phy) format which can then be fed to RaxML for phylogeny estimation.

2.2.9.2.3 PHYLOGENY ESTIMATION

The final alignment file in the Phylip format was then used to determine the maximum likelihood phylogeny by RaxML (Stamtakis et al. 2005) using general time reversible (GTR) model of GAMMA correction and a rapid bootstrap function performing 100 bootstraps. The resultant best tree was visualised and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

2.2.10 MAPPING OF THE SEQUENCED GENOMES FOR SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS

The selected sequenced genomes were mapped against the reference genomes (depending upon the analysis performed) for the identification of the variants in the form of Single Nucleotide Polymorphisms (SNPs). The analysis has been presented in chapters 7 and 8. To observe the diversity within *C. sakazakii* CC4 genomes, the index genome used was *C. sakazakii* SP291 (chapter 7), however when the target was to observe strain relatedness within each cluster, the earliest isolate of each cluster was used as the reference strain (chapter 8). The SNP analysis was performed on sequenced *C. sakazakii* isolates using a

program called SMALT (https://www.sanger.ac.uk/resources/software/smalt/) in combination with SAMtools and BCFtools utilities (http://samtools.sourceforge.net/) (Li et al. 2009). The mapping was executed in a stream of scripts (Appendix text 1.2). The input files for the query genome were in the FASTQ format (FASTQ format is the text based format used to store both nucleotide sequence and its corresponding quality scores by bundling both the FASTA and its quality data). The FASTA file of the reference was used to create the index file which was then mapped against the query genomes using a stream of scripts in the Linux operating system (Appendix text 1.2). The final output was the list of SNPs for the query genome in the form of VCF (Variant Call Format) text file containing information about the reference base, its position in the genome, quality and its variant in the query genome. The SNPs were then filtered in Artemis with the following parameters; minimum combined depth (DP) of 8.0, minimum quality (QUAL) score of 30.0 and minimum allele frequency (AFI) of 0.75 (0.90 when investigating the outbreak).

The remaining SNPs were further subjected to manual curation (only when there was working number of SNPs) using the BAM file in Artemis to exclude any SNPs which are not in the reliable areas of high coverage.

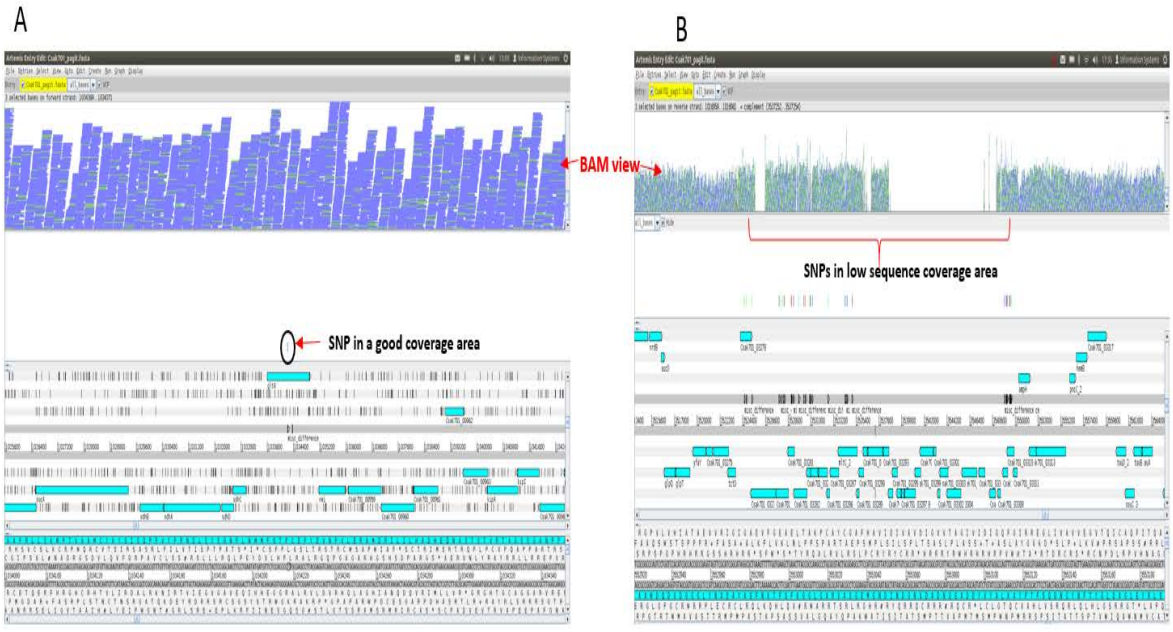


Figure 2. 6 Manual curation of the SNPs in Artemis.

The image above is an example of the manual curation of SNPs in Artemis. The reference genome was opened in Artemis and the VCF and BAM files of the query genomes were loaded on it. Using the BAM view only those SNPs which were in high coverage were selected (panel A) while the SNPs in low coverage area were omitted (panel B).

After manual curation of the SNPs, to visualise the position of these SNPs in the genome "create features from variants" option was used in Artemis and then using DNA plotter feature of the Artemis these SNPs were marked on the circular DNA map (Figure 2.6). The filtered VCF file was saved which showed the number of filtered SNPs in the query genome.

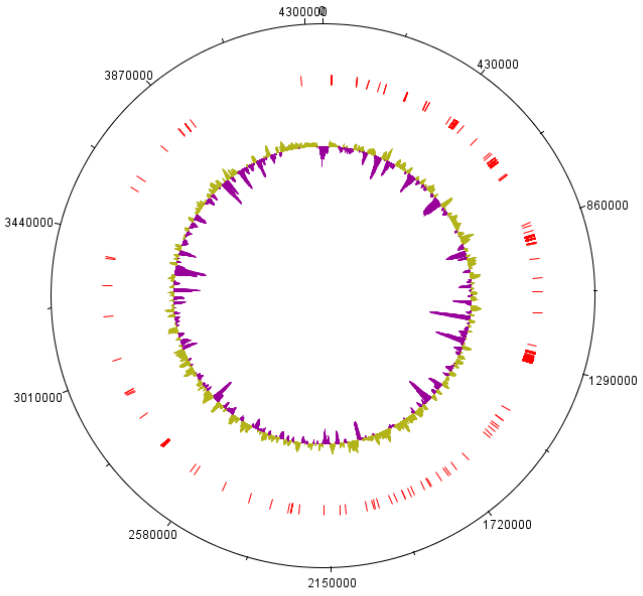


Figure 2. 7 Example image of the SNP marking on DNA plotter.

The image is an example DNA plotter image which indicates the location of SNPs on the reference genome. The SNPs were marked using the DNA plotter feature of Artemis. The outer most black circle with tick marks show the genome size and positions on the reference isolate. The inner red markings indicate the position of SNPs in the bacterial isolate. The inner most spiked circle indicates the GC contents (green = above average; purple = below average).

2.2.11 FRAGMENTED GENOMIC ANALYSIS USING GEGENEES

2.2.11.1 PRINCIPLE OF GEGENEES

Agren and colleagues (2012) recently developed a program called Gegenees which is a stand-alone computer program with a graphical user interface, the program can take hundreds of genomes to perform comparative genome analysis. The principal of Gegenees is based on the fragmentation of all the genomes to perform all against all comparison controlled by multi-threaded BLAST control engine. The most useful aspect of the Gegenees which was exploited in the current study was that the program allows the selection of a particular group of genomes called "target group" and to identify any unique genomic traits which differentiates the target group from the background group (genomes in the

comparison other than the target group genomes). Depending upon the uniqueness score within the target group, the program gives a score to the unique region called biomarker score; the biomarker score of 1.0 represents a genomic region uniquely conserved in the target group while missing uniformly from the background group. The program is written in java and was download from the website (http://www.gegenees.org/). The input files could be in either FASTA or Genbank format. Although the program can be used for several different tasks, since the present study was mainly targeted in identifying the unique regions within *C. sakazakii* CC4 genomes, Gegenees was used to identify genes unique in CC4 genomes which represented our target group.

2.2.11.2 GENERAL PARAMETERS OF GEGENEES USED IN THIS STUDY

Fragmented all against all comparison was performed using Gegenees (v2.0.3-LNX64) on an HP EliteBook (UNIX 10.04). The *C. sakazakii* genomes (FASTA format) were imported from the saved directory. Settings were manipulated by keeping the fragment size to 500bp and search method to BLASTN (Figure 2.7).

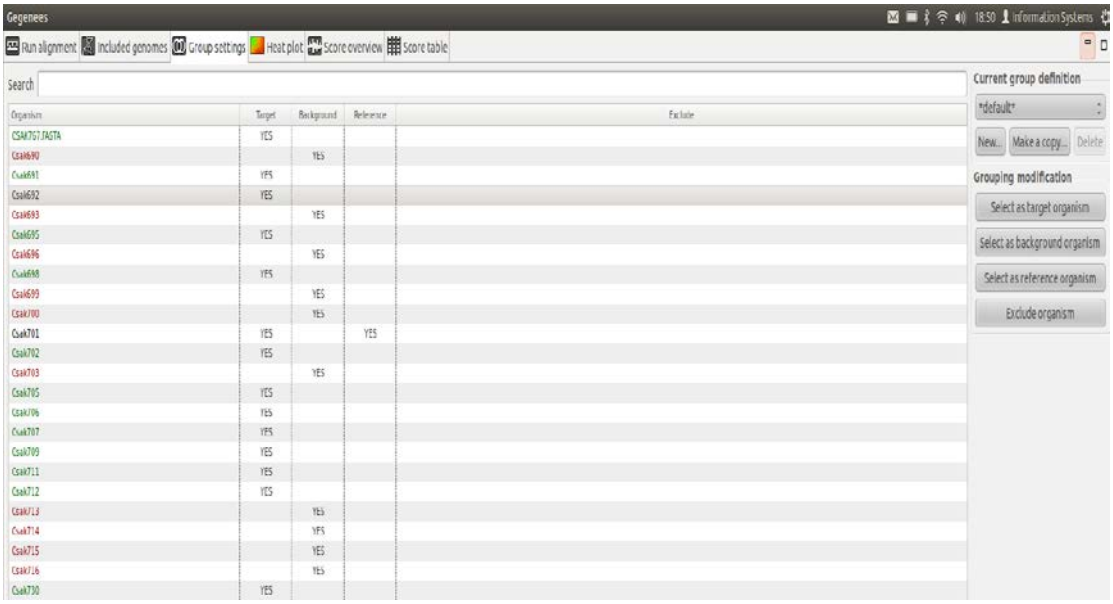


Figure 2. 8 Example image showing target group and background group settings.

The image above is a screenshot showing the settings of Gegenees, the genomes in green font colour indicate the target group while red font colour indicates the background group genomes. The black font colour shows the target group genome used as reference. The tabs on the top can be used to view and edit the included genomes, the Heat plot tab is used to view and export the heat plot generated after specifying the target and background group, Score overview tab is used to view the biomarker regions and Score table shows the biomarker score for each fragment which can then be exported and further analysed in Artemis.

After specifying the settings, the alignment was executed. The program performed the fragmentation of each genome according to the specified fragment size (in this case 500bp), BLASTed each fragment in each genome against all the fragments in other genomes. Once the fragmented alignment was completed, the program then allowed to select the target and background group (Figure 2.7).The target group was then selected and a reference genome within target group was assigned. Biomarker score settings were set to max/min and only the biomarker traits with the maximum score i.e. 1.0 were selected. Once the target and background group settings were specified, the program generated a heat plot showing the relatedness of the all the genomes which can be exported in the form an image. The biomarker score table was then exported and the coordinates were used to identify the genomic regions using the annotated Genbank file of the reference isolate in Artemis.

2.2.12 BLAST RING IMAGE GENERATOR

2.2.12.1 PRINCIPLE OF BRIG

Alikhan et al. (2012) developed a tool called BLAST ring image generator (BRIG) which can generate comparative genome images of the prokaryotic genomes. These comparative genome images are of great value as they are a quick way of visualising several genomes to observe genomic homologies and differences across certain genomes based on a single central reference genome. The program is written in Java 1.6 and uses CGView to render the images (Stothard et al. 2004) and BLAST to perform genomic comparisons.

2.2.12.2 PARAMETERS OF BRIG USED

The program was downloaded from http://brig.sourceforge.net/). The comparative genomic images were created using the BRIG tutorial (http://brig.sourceforge.net/brig-tutorial-1-whole-genome-comparisons/). The BLAST type used was BLASTn and minimum identity threshold was kept to 50% (Figure 2.8). It must be noted that in the present analysis, when performing the comparative genome analysis, the results of the BRIG analysis were validated through manual analysis using ACT.

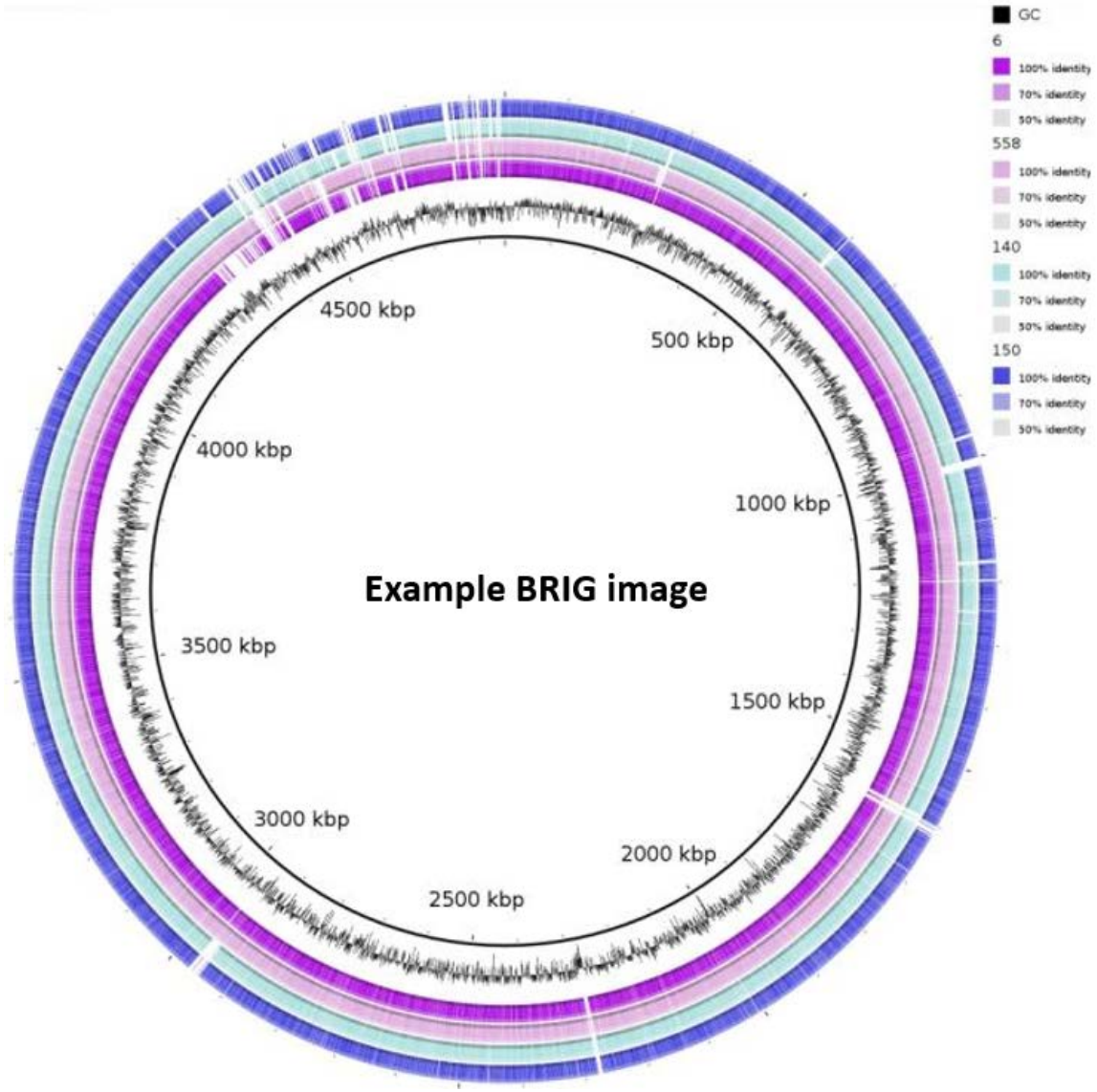


Figure 2. 9 An example comparative genome image constructed using BRIG.

The central back line indicates the reference genome. The black rings with spikes indicate the G+C content while the coloured circular rings indicate different *C. sakazakii* genomes. The white regions on the coloured rings show the genomic regions which are missing in the query genome while unique to the reference genome.

Once a comparative genomic image was created, it indicated the unique regions (white regions in example Figure 2.8) specific to the reference genome. This was further manually investigated by ACT and Artemis using the annotated genome files of the reference genome.

All of the above described genomic tools have been used in analysis throughout this study and the reference will be made to these sections when referring to any of these methods.

3 CHAPTER 3

GENOME SCREENING OF *C. SAKAZAKII* CC4 FOR VIRULENCE AND ENVIRONMENTAL FITNESS ASSOCIATED TRAITS

3.1 BACKGROUND

A multilocus sequence type (MLST) scheme has been adopted for the *Cronobacter* genus and is being extensively used for the identification of *Cronobacter* sequence types (ST) across the world (http://pubmlst.org/*cronobacter*/) (Baldwin et al. 2007).

The clonal complex 4 (CC4) refers to *C. sakazakii* ST4 and its single or double loci variants and constitutes a dominant clonal complex in the *Cronobacter* PubMLST. The association of *C. sakazakii* CC4 with neonatal meningitis has been well documented (Joseph & Forsythe, 2011; Hariri et al. 2013). Comparative analysis by Joseph & Forsythe (2011) using 41 clinical strains obtained from 7 different countries isolated during 1953-2008 indicated that almost half (20/41) of these isolates were ST4. More importantly, their study indicated the clinical significance of ST4 since amongst a total of 12 isolates from meningitis, at least 9 were *C. sakazakii* ST4 which were isolated from cerebrospinal fluid (CSF), blood and trachea. These ST4 strains analysed by Joseph and Forsythe were isolated between 1977 and 2008 from 6 different countries (the Netherlands, France, United States, New Zealand, Czech Republic, Canada); therefore suggesting that ST4 is a virulent and stable clone strongly associated with neonatal meningitis (Joseph & Forsythe, 2011).

Another study by our group reinforced the association of *C. sakazakii* CC4 with neonatal meningitis (Hariri et al. 2013). Hariri and colleagues analysed 15 *Cronobacter* strains which were provided by the Centers for Disease Control (CDC). The study showed that all of the 5 CSF isolates were either *C. sakazakii* ST4 or its one loci or two variants together forming a clonal complex known as clonal complex 4 or CC4 (Hariri et al. 2013).

The above mentioned observations are further strengthened by that fact that in the *Cronobacter* outbreak of the French neonatal intensive care unit (NICH) 1994, amongst 30 of *C. sakazakii* isolates, 16 isolates were *C. sakazakii* CC4; more importantly all of the 3 babies who died were infected by *C. sakazakii* CC4 strain (Caubilla-Barron et al. 2007). These studies suggested a close relationship of *C. sakazakii* CC4 with neonatal infections especially neonatal meningitic cases.

Not only the predominant association of *C. sakazakii* CC4 with neonatal meningitis is of serious concern, another alarming fact is its frequent isolation from PIF and its manufacturing environment such as include floors, bays, roller dryers, tankers and air fillers factories (Neelam et al. 1987; Mosso et al. 1994; Van Os et al. 1996; Hein et al. 2009; Craven

et al. 2010; Jacobs et al. 2011). The frequent isolation of *C. sakazakii* CC4 from PIF and its manufacturing factories poses a significant threat to the lives of newborn babies.

To date, CC4 is the dominant clonal complex reported in *Cronobacter* and represents approximately 20% of the total *Cronobacter* isolates deposited in the *Cronobacter* PubMLST at the time of writing current section in this thesis (http://pubmlst.org/perl/bigsdb/bigsdb.pl?db=pubmlst_*cronobacter*_isolates/, as of 22nd November 2014, BST; 16:02).

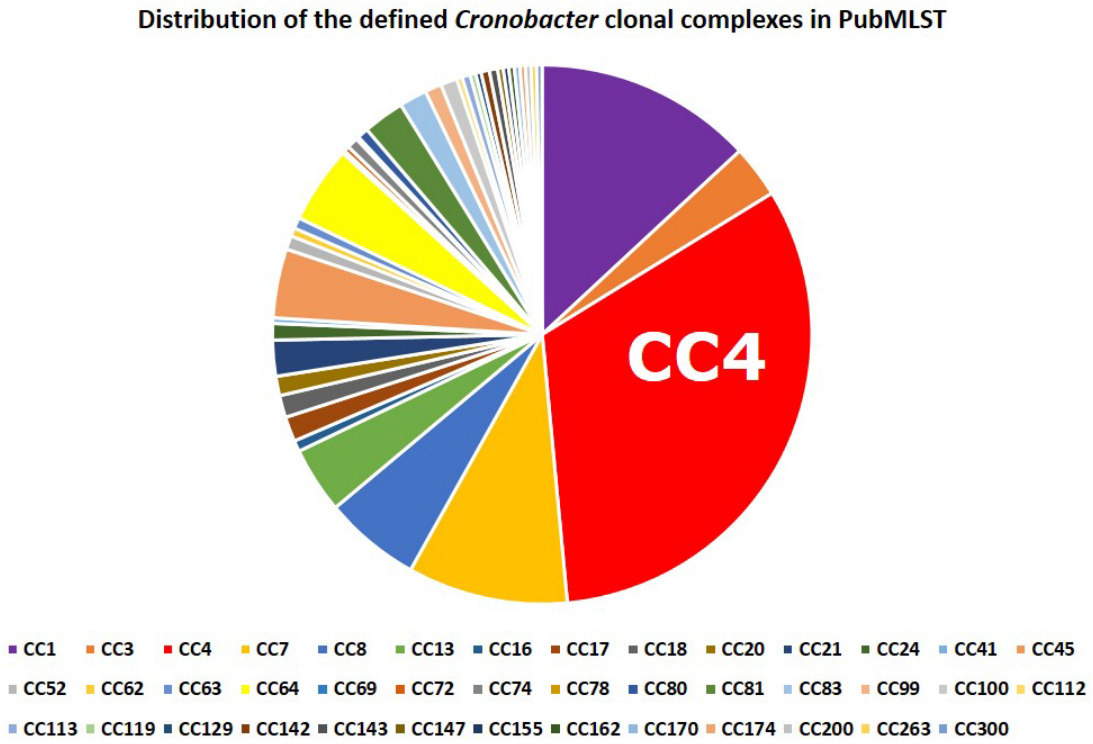


Figure 3. 1 Distribution of clonal complexes in *Cronobacter* PubMLST.

The pie-chart represents the distribution of 604 (of total 1007) *Cronobacter* isolates in clonal complexes defined to date. Please note the dominance of CC4 (red coloured area) amongst different clonal complexes. The information was true as of 22nd November 2014.

More significantly, it is not just the association of *C. sakazakii* CC4 with clinical or neonatal meningitic cases, an increasingly disturbing fact is its strong association with powdered infant formula (PIF) or infant formula (IF) which is making this clonal lineage of *Cronobacter* a high concern for neonatal health. Our group has previously reported that 24% of the *Cronobacter* strains isolated from the environment of 6 milk powder manufacturing plants in Australia and Germany (Sonbol et al. 2013), as well re-identified strains from an international survey in 1988 of *Cronobacter* in PIF were *C. sakazakii* CC4 (Muytjens et al. 1988). It is interesting to note that amongst 195 *C. sakazakii* CC4 strains (as of date, 22nd

November 2014) listed in *Cronobacter* PubMLST, at least 44% were clinical isolates, nearly 24% were infant formula isolates, about 8% were isolated from food and ingredients while approximately 16% represented environmental and about 0.5% were water isolates (Figure 3.2). Taken together clinical and IF isolates represented approximately 70% of the total *C. sakazakii* CC4 isolates signifying a threat to the neonatal health.

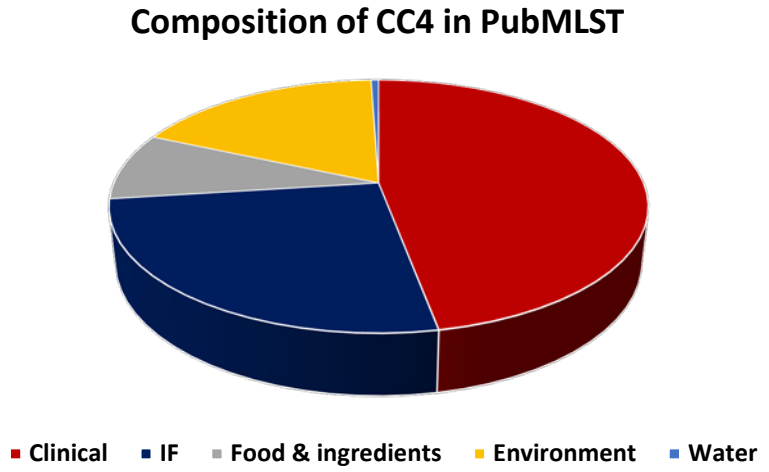


Figure 3. 2 Composition of *C. sakazakii* CC4 in *Cronobacter* PubMLST.

The pie-chart represents the distribution of 195 *C. sakazakii* CC4 isolates deposited in *Cronobacter* PubMLST. Please note the dominance of CC4 in clinical (dark red coloured area) and in infant formula (navy blue coloured area). The information was true on 22nd November 2014.

Furthermore, the relative abundance of the environmental CC4 isolates and its occurrence in PIF and PIF manufacturing plants is indicative of its ability to survive under stressful environmental conditions. The finished genome sequence of a *C. sakazakii* ST4 strain (SP291) has been published which was notable for its persistence in a PIF production facility in Ireland (Power et al. 2013).

The prevailing association of *C. sakazakii* CC4 with neonatal meningitis and its relatively increased persistence in PIF and its manufacturing environment warranted further investigation at the genomic level. As part of this genomic analysis 30 *C. sakazakii* sequenced genomes were included in this study consisting of 18 *C. sakazakii* CC4 and 12 *C. sakazakii* non-CC4 isolates (Table 3.1). The genomic analysis presented in this chapter aimed to screen these 30 *C. sakazakii* genomes for the virulence and environmental associated factors which have already been described in previous studies, in order to identify if any of these traits are specific to *C. sakazakii* CC4.

3.2 HYPOTHESIS

C. sakazakii CC4 isolates have unique virulence or environmental fitness associated traits.

3.3 AIMS OF THE CHAPTER

- Whole genome phylogeny determination of 30 *C. sakazakii* isolates.
- Screen *C. sakazakii* genomes for virulence associated traits previously described in *Cronobacter.*
- To screen *C. sakazakii* genomes for the presence of environmental fitness associated traits previously described in *Cronobacter.*
- *In silico* serotyping of *C. sakazakii* genomes against the known serotypes described previously in *C. sakazakii.*
- *In silico* plasmid analysis of *C. sakazakii* genomes by aligning them against previously described virulence associated plasmid pESA3 of *C. sakazakii.*
- To screen *C. sakazakii* genomes for the presence of neonatal meningitic associated traits previously described in *E. coli* K1.

3.4 METHODS

3.4.1 BACTERIAL STRAINS USED IN THE ANALYSIS

A total of 30 *C. sakazakii* genomes were used in the study presented in this chapter (Table 3.1); 18 *C. sakazakii* CC4 and 12 *C. sakazakii* non-CC4. These strains were primarily clinical isolates which had been isolated between 1950 and 2010 from 11 different countries. The finished genomes were already available for three strains; *C. sakazakii* BAA-894 (Genbank accessions CP000783-5) (Kucerova et al. 2010), *C. sakazakii* SP291 (Genbank accessions CP004091-4) (Power et al. 2013) and a ground whole grain isolate *C. sakazakii* ES15 (Genbank accessions CP003312) (Shin et al. 2012). Draft genomes were already available for five genomes; *C. sakazakii* 701 (Genbank accessions CALE01000001-768), *C. sakazakii* 696 (Genbank accessions CALF01000001-569), *C. sakazakii* 680 (Genbank accessions CALG01000001-201) (Joseph et al. 2012b), *C. sakazakii* 377 (AWFW00000000) (Masood et al. 2013a), *C. sakazakii* 1587 (AWSP00000000) (Masood et al. 2013b). An additional 22 genomes were sequenced as part of this study (Table 3.1). The whole genome shotgun projects for these isolates were deposited at Genbank and their accession numbers are given in Table 3.1.

Table 3. 1 Genomes sequenced *C. sakazakii* isolates analysed in chapters 3-7

C. sakazakii strain	ST	CC	Country	Date*	Source	Extra information	Genbank accession
			***C. sakazakii* CC4 strains**				
377 (NCIMB 8272)	4	4	UK	1950	Milk powder	PIF manufacturing plant	AWFW00000000
SP291	4	4	Ireland	Unkn			CP004091-4
20	4	4	Czech Republic	2003	Clinical; Faeces	Adult, 74 years old	JNBH00000000
553	4	4	Neth	1977	Clinical	1 day infant	JNBJ00000000
557	4	4	Neth	1979	Clinical	4 day old infant	JNBK00000000
558	4	4	Neth	1983	Clinical		JNBL00000000
701	4	4	France	1994	Clinical; Peritoneal fluid	Fatal NECIII	CALE01000001-768
767	4	4	France	1994	Clinical; Trachea	Fatal meningitis	JNCX00000000
1219	4	4	USA	2009	Clinical	Fatal meningitis	JNCY00000000
1220	4	4	USA	2003	Clinical; CSF		JNDH00000000
1221	4	4	USA	2003	Clinical; CSF		JNDI00000000
1225	4	4	USA	2007	Clinical; Blood	Fatal meningitis	JNDF00000000
1240	4	4	USA	2009	Clinical; CSF		JNDD00000000
721	4	4	USA	2003	Clinical; CSF		JNDA00000000
1231	4	4	New Zealand	2005	Clinical; Faeces	Infant with meningitis	JNDG00000000
6	4	4	Canada	1990	Clinical		JNBG00000000
4	15	4	Canada	1990	Clinical		JMSR00000000
1587	109	4	Israel	2000	Clinical; CSF	Severe anatomical damage to brain	AWSP00000000
			***C. sakazakii* non-CC4 strains**				
BAA-894	1	1	USA	2001	Non-infant formula	Fatal meningitis	NC_009778-80
1218	1	1	USA	2001	Clinical; CSF	Fatal meningitis	JNDC00000000
ES15	125	100	South Korea		Ground whole grain		PRJNA81953
1	8	8	USA	1980	Clinical; Throat		JMRI00000000
5	8	8	Canada	1990	Clinical		JNBF00000000
680	8	8	USA	1977	Clinical; CSF		CALG01000001-201
140	40	45	India	2005	Spice	Ajwan cumin seed	JMSS00000000
150	16	16	South Korea	2005	Spice	Yoojeon food spices	JNDB00000000
978	3	3	UK	2007	Clinical	Enteral feeding tubes	JNCZ00000000
696	12		France	1994	Clinical; Faeces	NECII infant	CALF01000001-569
520	12		Czech Republic	1983	Clinical		JNBI00000000
1249	31	31	UK	2010	Clinical	Fatal infant isolate	JNDE00000000

*; Date of isolation, ST; Sequence Type, CC; Clonal Complex, Neth; Netherlands, PIF; Powdered infant formula, NEC; Necrotising enterocolitis, CSF; Cerebrospinal fluid, Unkn; Unknown

3.4.2 GENOME SEQUENCING ASSEMBLY AND ANNOTATION

Genome sequencing of *C. sakazakii* strains was carried out by the Exeter sequencing service, University of Exeter (UK). The genome sequences were generated on different versions of Illumina sequencing platforms (section 2.2.1). The genome assembly was performed using Velvet (section 2.2.1). The genomes were annotated using prokaryotic genome annotation system (PROKKA) (Seemann et al. 2014) as described in section 2.2.5.

3.4.3 PHYLOGENETIC ANALYSIS

The genome sequences of 30 *C. sakazakii* genomes (Table 3.1) were aligned using Mugsy (Angiuoli et al. 2011) and the core genome extracted as described previously (Sahl et al. 2011; Clark et al. 2012; McNally et al. 2013) (section 2.2.8). Maximum likelihood phylogeny was then reconstructed using RaxML with the GTR-gamma model (Stamatakis et al. 2005) and the resulting trees visualised and annotated using Figtree (http://tree.bio.ed.ac.uk/software/figtree/). The method is described in more detail in section 2.2.8.

3.4.4 GENOME COMPARISONS AND GENOME SCREENING FOR THE TRAITS OF INTEREST

The genome comparison was undertaken using Artemis Comparison Tool (ACT) (Carver et al. 2005) described in detail in section 2.2.7. The genome sequences of 30 *C. sakazakii* strains analysed in the present study are now available at *Cronobacter* BIGsDB (http://pubmlst.org/perl/bigsdb/bigsdb.pl?db=pubmlst_*cronobacter*_isolates/). In addition to ACT, the BLAST feature of *Cronobacter* BLAST was also used to search for traits of interest, where appropriate. The nucleotide sequence of any gene of interest was BLASTed against 30 *C. sakazakii* genomes to determine its presence in the sequenced genomes (http://pubmlst.org/perl/bigsdb/bigsdb.pl?page=plugin&name=BLAST&db=pubmlst_*cronobacter*_isolates/). A gene was designated as present when it had ~50% nucleotide alignment with any sequenced genome i.e. when approximately half the length of a gene sequence was present.

3.5 RESULTS AND DISCUSSION

A number of virulence associated traits and environmental fitness associated traits previously described in *Cronobacter* were investigated. The aim was to find if any of these traits show specificity or predominance in *C. sakazakii* CC4.

Note. For the purpose of clarity and continuity, a brief background for each of the trait investigated is given in the result and discussion section of this chapter.

3.5.1 WHOLE GENOME PHYLOGENY

The core genome phylogeny of 30 *C. sakazakii* genomes was determined (Figure 3.3). These strains had been collected over 6 decades from 11 different countries (Table 3.1). The resulting core genome phylogeny showed a close clustering of all 18 CC4 strains despite their temporal and geographic spread. This confirms that CC4 is a genetically distinct cluster of *C. sakazakii*. The observation is suggestive of a clade with very low levels of diversity.

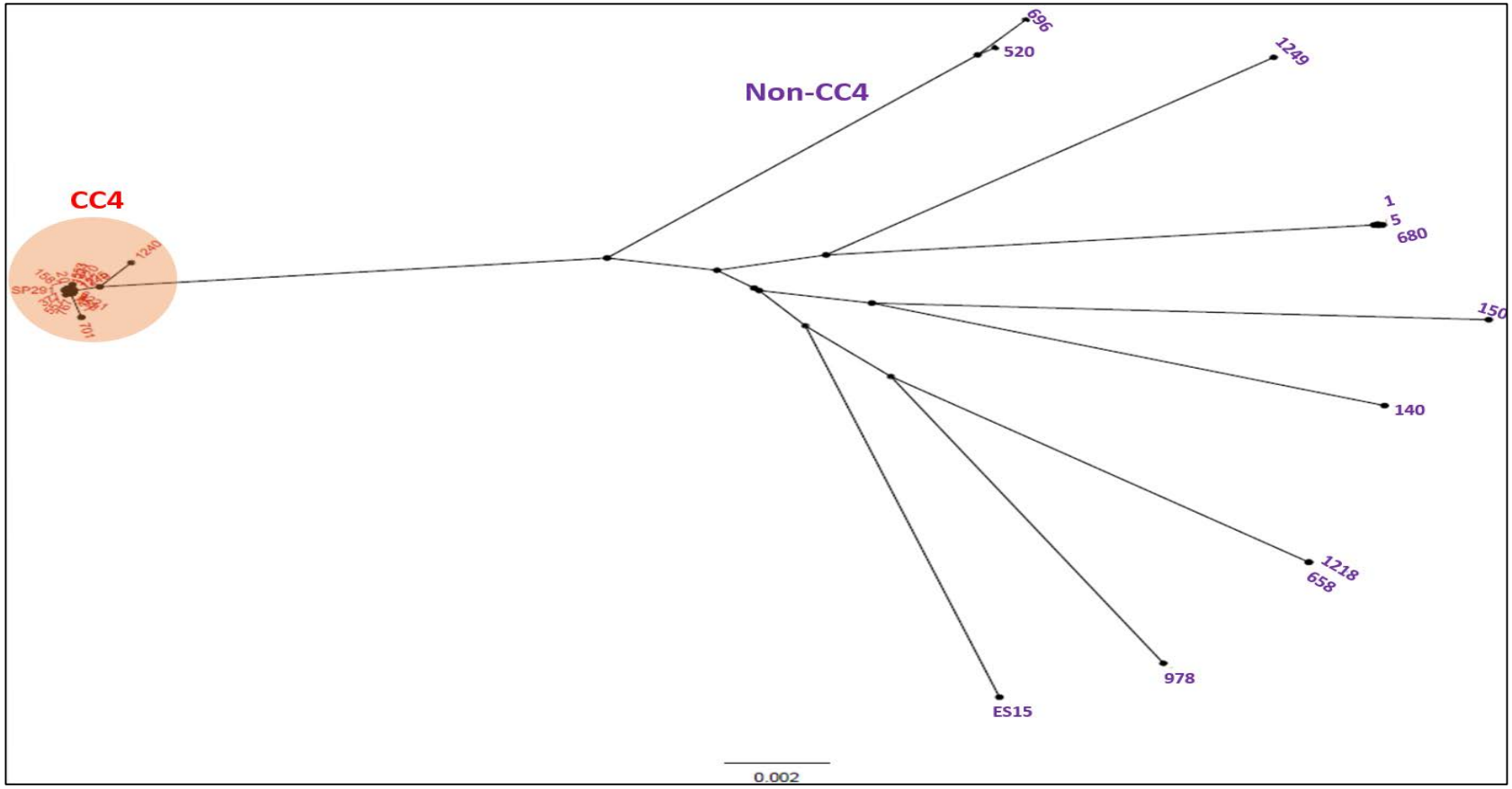


Figure 3. 3 The whole genome based maximum likelihood tree of 30 *C. sakazakii* strains.

The genomes of 30 *C. sakazakii* (n=30) strains approximately 4.5 million bp were aligned using Mugsy, core genome extracted and maximum likelihood phylogeny generated using RAxML. Phylogeny was viewed and annotated using FigTree. All 18 *C. sakazakii* isolates (red tip labels, orange encircled) clustered tightly with each other. The tip labels in purple represent non-CC4 isolates (n=12). The scale bar at the bottom shows the number of nucleotide substitutions per base.

3.5.2 GENOME SCREENING OF SEQUENCED *C. SAKAZAKII* STRAINS FOR VIRULENCE ASSOCIATED TRAITS

Generally a bacterium has to cross several barriers before it can successfully establish an infection in its host (Wilson, 2002); this would need an underlying genetic machinery. A number of virulence associated genes were investigated in sequenced *C. sakazakii* isolates and are described below.

3.5.2.1 ADHESION ASSOCIATED TRAITS

Adhesion is the first and essential step towards bacterial pathogenesis as it facilitates the colonisation and subsequent infection of host cells (Soto & Hultgren, 1999). Fimbriae are the appendages which help bacteria to adhere to the host cell surface thereby helping them with colonisation and successive infection process (Soto & Hultgren, 1999). At least ten putative fimbrial gene clusters have been described in *Cronobacter* (Joseph et al. 2012b). The *C. sakazakii* genomes were interrogated for any plausible association of these fimbrial clusters with *C. sakazakii* CC4 genomes. The genomes were screened using ACT and *Cronobacter* BLAST (Table S1). The results are presented here as percentage of CC4 (n=18) and non-CC4 strains (n=12) positive for a gene under investigation (Figure 3.4). Although no *C. sakazakii* CC4 specific pattern of fimbrial gene clusters was observed, some interesting variation was noted, for example fimbrial region 1 (ESA_01970-6) was absent in most *C. sakazakii* strains but was detected in both *C. sakazakii* ST1 strains; 1 and 1218, therefore absence of this region in clinically significant strains indicate that it may not be essential towards virulence of *C. sakazakii* (Table S1). Curli fimbriae have been associated with adhesion to the host cell in *E. coli* (Doran et al. 1993; Olsen et al. 1993). A curli fimbrial gene cluster (ctu_16160-230) has been reported in the genome of *C. turicensis* z3032 (Joseph et al. 2012b); however none of the curli fimbriae homologues was noted in any of the *C. sakazakii* genome analysed in this study. Type 1 fimbriae have also been associated with adhesion and virulence of the *E. coli* (Connel et al. 1996). The type 1 fimbrial operon (*fimABCDEFGHI*) located on genome of *E. coli str. K-12 substr.* MG1655 (accession: NC_000913.3) was used in this study to detect any of its homologue in *C. sakazakii* genomes, however none was detected in any of the *C. sakazakii* genome (data not shown).

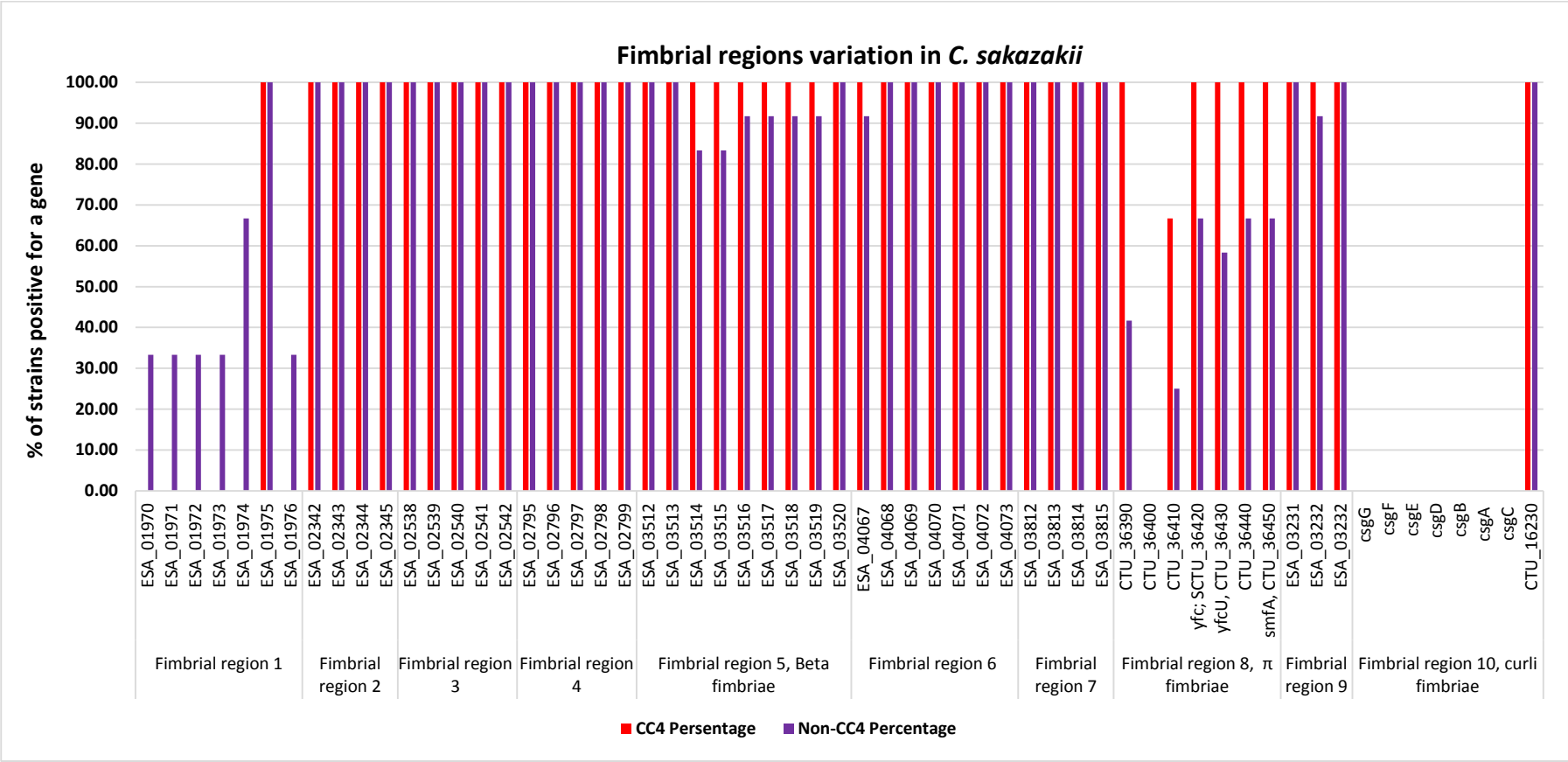


Figure 3. 4 Fimbrial regions variation in *C. sakazakii* genomes.

Genomes of 30 *C. sakazakii* were screened for the presence of 10 different fimbrial region previously described in the genomes of *C. sakazakii* BAA-894 and *C. turicensis* z3032. The data is presented in the graph as a percentage of *C. sakazakii* CC4 (n=18) and non-CC4 (n=12) genomes positive for a gene under investigation. Red bars indicate the percentage of CC4 genomes positive for a given gene while purple bars indicate the non-CC4 genomes positive for any given gene within the fimbrial regions.

3.5.2.2 INVASION ASSOCIATED TRAITS

Like adhesion, invasion is a crucial mechanism of bacterial pathogenicity. The role of outer membrane protein A (ompA) towards brain microvascular endothelial cell (HBMEC) has been studied in *E. coli* K1 (Badger et al. 2000). It has also been proposed that ompA and ompX proteins in *C. sakazakii* are important and act synergistically towards the invasion of Caco-2 and IN-407 cell lines (Kim et al. 2010). Similarly, another study by Kim et al. (2010) suggested that ompA along with ompX protein is involved in the basolateral invasion of *C. sakazakii*; their study showed that the mutants lacking *ompA* and *ompX* genes were significantly reduced in their ability to invade the human enterocyte-like epithelial Caco-2 and human intestinal epithelial INT-407 (Kim et al. 2010). A recent study has shown that in *C. sakazakii* 29544, a putative Inv (invasion protein) encoded by *Inv* gene is essential for the basolateral invasion of Caco-2 cells and acts synergistically with ompA (Chandrapala et al. 2014). Genome interrogation of 30 *C. sakazakii* strains revealed that the *ompA* and *ompX* genes were noted in 100% of the CC4 and non-CC4 isolates. However, Inv gene was absent in 6 CC4 while 7 non-CC4 isolates (Figure 3.5, Table S1). The absence of the *Inv* gene in clinically important CC4 and non-CC4 strains such as isolates 721 and 696 (CSF isolates) indicate that this gene may not be essential for the invasion of *C. sakazakii* in general and CC4 in particular.

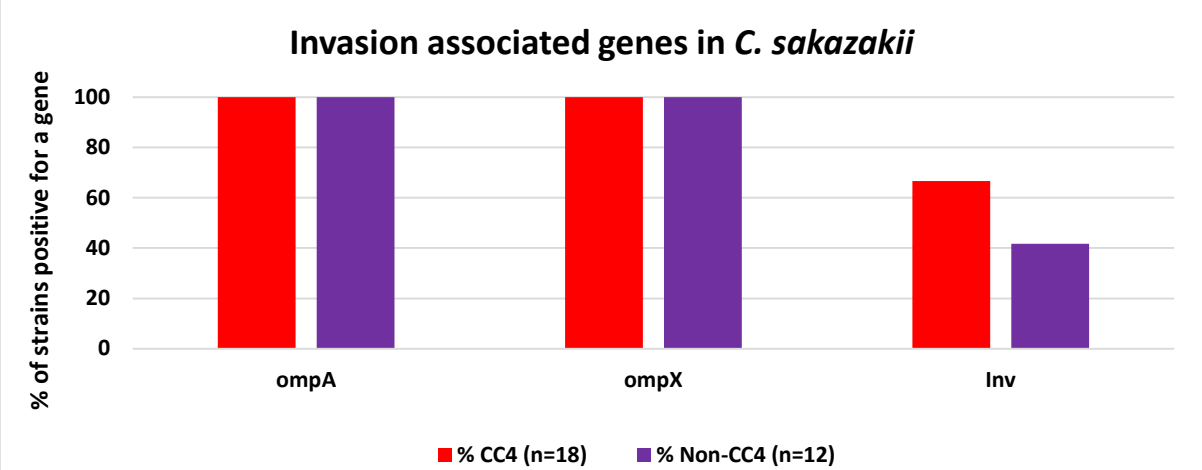


Figure 3. 5 Invasion associated genes in *C. sakazakii* genomes.

Genomes of 30 *C. sakazakii* strains were screened for the presence of invasion associated genes *ompA*, *ompX* and *Inv*. The data is presented in the graph as a percentage of *C. sakazakii* CC4 (n=18) and non-CC4 (n=12) genomes positive for a gene under investigation. Red bars indicate the percentage of CC4 genomes positive for a given gene while purple bars indicate the non-CC4 genomes positive for the invasion associated gene.

3.5.2.3 OUTER MEMBRANE PROTEIN A (ompA)

Several studies have established the role of the ompA protein towards the invasion of the neonatal meningitic *E. coli* K1 as well as *C. sakazakii* (Mohan et al. 2009; Kim et al. 2010). A study by Kim and Wang (2002) indicated that a mutant of *E. coli* K1 strain RS218 (018:K1:H7) in which the *ompA* gene was deleted, showed less invasion of Brain Microvascular Endothelial Cells (BMEC) when using *in vitro* invasion assays (Wang & Kim, 2002). Similarly Nair et al. (2009) showed that ompA is a major fibronectin binding protein in *C. sakazakii* and is essential for the invasion of the BMEC. Their study showed that the *C. sakazakii* mutants in which the *ompA* gene was knocked out exhibited significantly reduced invasion of the BMEC *in vitro*, hence indicating that *ompA* is required for the invasion and therefore the pathogenicity of *C. sakazakii* (Mohan et al. 2009). Another study by Koebnik (1999) showed that in *E. coli*, all the loops of the ompA protein were essential for its functioning (Koebnik, 1999). Since the role of *ompA* has been well established towards the invasion and therefore virulence of *C. sakazakii*, it was of high significance to further analyse this important virulence associated trait in sequenced *C. sakazakii* isolates in order to observe if there is any variation in the sequence of *ompA* which could give advantage to *C. sakazakii* CC4 isolates as compared to non-CC4. The *ompA* gene was detected uniformly across all the sequenced strains in this study (Figure 3.5, Table S1). The results were validated by the PCR amplification using the primers and reaction conditions previously defined by Nair and Venkitanarayanan (2006) (Figure S1). The aim of the present analysis was to align the *ompA* gene sequence of *C. sakazakii* CC4 and non-CC4 isolates used in this study to observe any variations at the nucleotide level such as SNPs, insertions or deletions and whether these genetic changes are translated to the protein level which could ultimately change the binding ability to the host tissues (Weissman et al. 2003). The *ompA* gene sequence of 30 *C. sakazakii* isolates was extracted from their genomes and aligned using phylogeny.fr (http://phylogeny.lirmm.fr/phylo_cgi/index.cgi/). Initial alignment indicated an incomplete gene sequence of *C. sakazakii* strain 150 which could be due to low sequence coverage at that area of the genome; therefore this strain was omitted from *ompA* analysis. The full length *ompA* gene sequence of *C. sakazakii* was 1077 bp. The gene sequences of 29 *C. sakazakii* isolates was aligned using online genomic tool phyolgeny.fr. The resulting tree was viewed and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The *ompA* phylogeny indicated close clustering of *C. sakazakii* CC4 isolates except for the two

CC4 isolates; 4 and 1219 which branched distantly and with non-CC4 strains (Figure 3.6). Similarly *C. sakazakii* CC4 isolates 553 and 557 showed longer branch length, although on the same branch as other CC4 isolates. Another interesting observation was the clustering of one *C. sakazakii* non-CC4 isolate 140 within the *C. sakazakii* CC4. Additionally, 2 *C. sakazakii* non-CC4 isolates were located on the same branch as *C. sakazakii* CC4, however with a longer branch length (Figure 3.6).

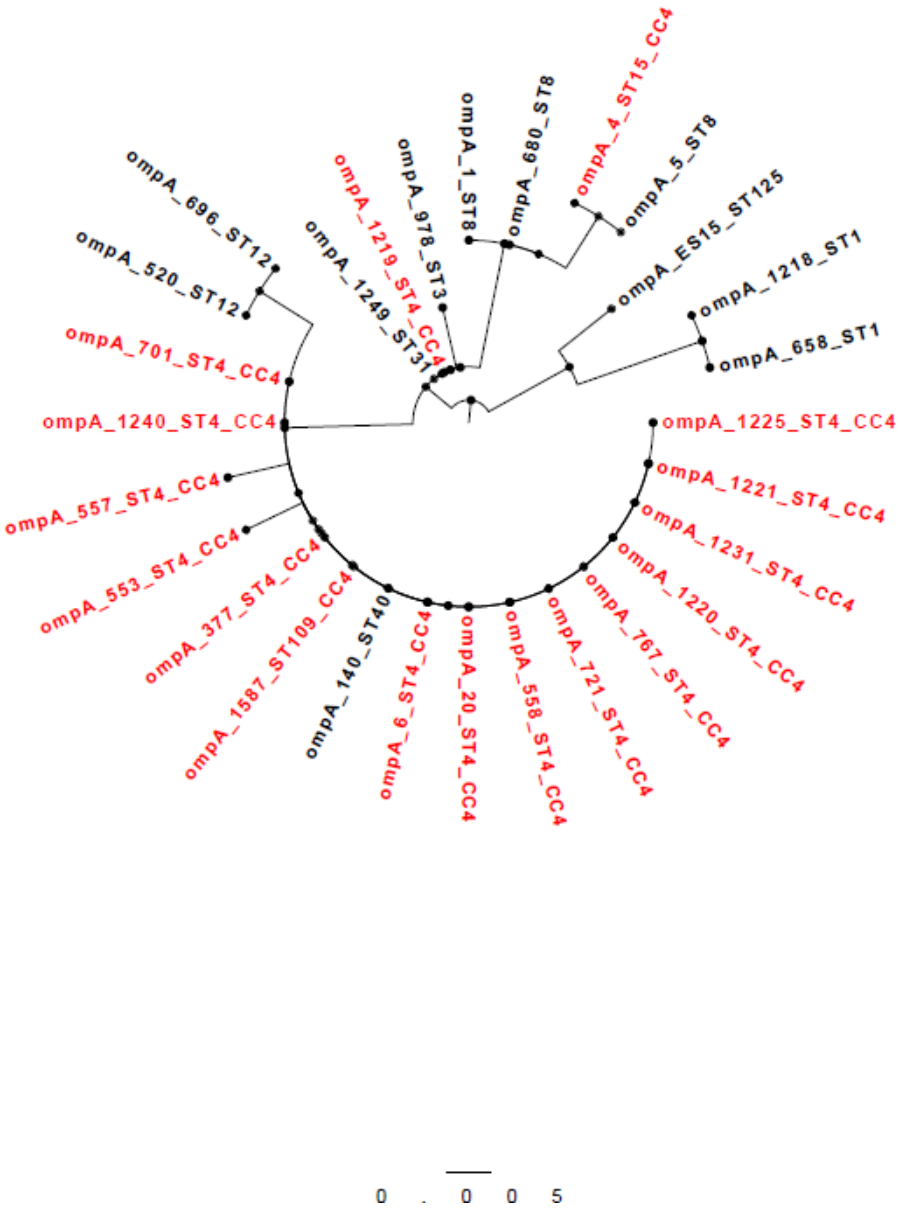


Figure 3. 6 The phylogeny of 29 *C. sakazakii* isolates based on ompA gene sequences.

The figure shows the midpoint rooted phylogenetic tree for the 29 *C. sakazakii* isolates based on the *ompA* gene sequence (1077 bp). The alignment was constructed using phylogeny.fr and the tree annotated using FigTree. The red font colour indicates *C. sakazakii* CC4 (n=18) while black font colour indicates *C. sakazakii* non-CC4 isolates (n=11). The scale bar indicates the number of nucleotide substitutions per base.

These observations were further investigated by manually analysing the gene alignment using Jalview which was integrated within phylogeny.fr (Figure 3.7).

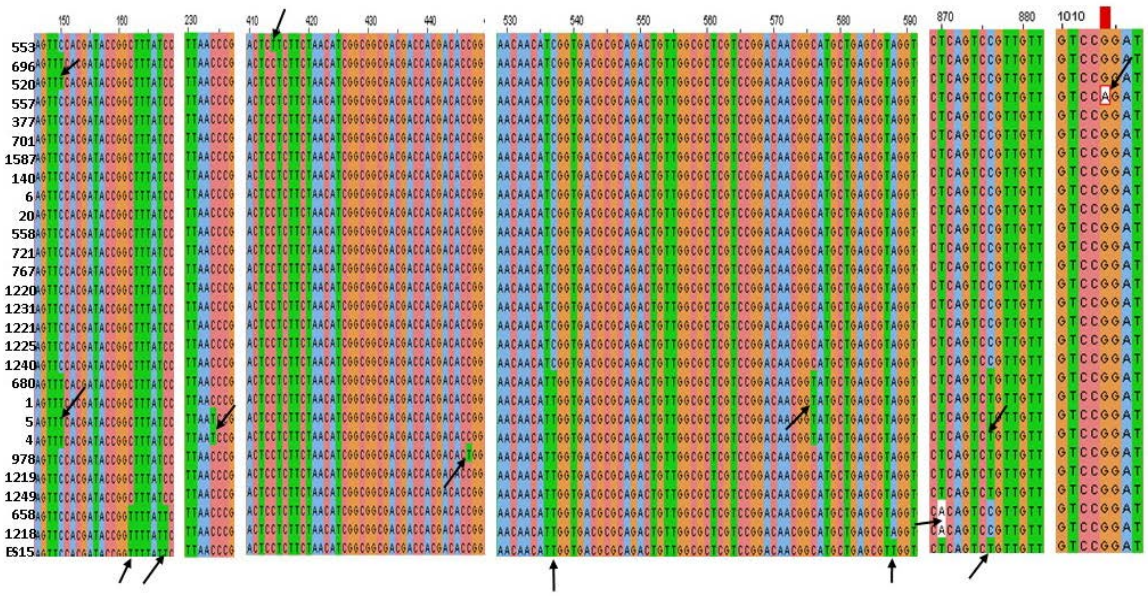


Figure 3. 7 Manual curation of the *ompA* gene sequence alignment.

The figure above represents the cropped segments of the screen shots combined as one image showing only the sections where variations in the *ompA* gene sequence was observed. The black arrows point towards the Single Nucleotide Variations (SNP) changes in *C. sakazakii* isolates.

The manual curation of the *ompA* gene sequence alignment indicated an overall great degree of sequence consensus across both *C. sakazakii* CC4 and non-CC4 with nucleotide variations at only 12 positions. The gene sequence was found to be highly conserved across most CC4 isolates while non-CC4 isolates showed some Single Nucleotide Polymorphism (SNP) changes at only fewer positions across the length of the alignment (Table 3.2). Two of the CC4 isolates; 4 and 1219 which branched with non-CC4 isolates also had fewer SNPs some of which were shared with non-CC4 isolates (Table 3.2). The SNPs can cause variation in the resulting structure of the protein which may alter their functional characteristics. In bacteria, for example these altered functional properties of a protein may enhance their virulence potential. Since, ompA is a protein which is associated with adhesion and invasion of *E. coli* K1 and *C. sakazakii*, it was therefore important to investigate whether these changes in the nucleotide sequence are translated into the amino acid sequence (i.e. non-synonymous changes). To investigate this the amino acid structure for the *ompA* gene sequences of all 29 *C. sakazakii* isolates was deduced using an online nucleotide to amino

acid translation tool available at (http://in-silico.net/tools/biology/sequence_conversion/). However, the deduced amino acid sequence for all *C. sakazakii* isolates showed no variation as 100% sequence conservation was observed across all the *C. sakazakii* isolates. Hence all the nucleotide substitutions shown in Table 3.2 were synonymous. Therefore, the present analysis of the *ompA* gene and amino acid sequence analysis suggested that although *ompA* might be essential in the virulence of *C. sakazakii*, it does not form the basis of differentiation between *C. sakazakii* CC4 and non-CC4 and hence it contributes equally towards the virulence potential of both subsets of *C. sakazakii*.

Table 3. 2 Nucleotide variations observed in *ompA* gene sequence alignment

No	SNP Position	*C. sakazakii* Isolates	Nucleotide substitution (with reference to general consensus)	Type of SNP
1	150	696, 520, 680, 1, 5, 4	C → T	Synonymous
2	162	658, 1218, ES15	C → T	Synonymous
3	168	658, 1218, ES15	C → T	Synonymous
4	234	5, 4	C → T	Synonymous
5	414	553	C → T	Synonymous
6	447	978	C → T	Synonymous
7	537	680, 1, 5, 4, 978, 1219, 1249, 658, 1218, ES15	C → T	Synonymous
8	576	680, 1, 5, 4	C → T	Synonymous
9	588	ES15	A → T	Synonymous
10	870	658, 1218,	T → A	Synonymous
11	876	680, 1, 5, 4, 978, 1219, 1249, ES15	C → T	Synonymous
12	1014	557	G → A	Synonymous

Green font colour shows the nucleotide in general consensus; Red font colour shows the substituted nucleotide

3.5.2.4 THE FILAMENTOUS HAEMAGGLUTININ, THE FHA LOCUS

Filamentous haemagglutinin (FHA) is the large, rod shaped protein (>200 kDa) which is expressed by bacteria such as *Bordetella pertussis* (Makhov et al. 1994). The FHA facilitates bacterial adhesion to epithelial cells and macrophages *in vitro* and is involved in colonisation of trachea *in vivo* (Julio & Cotter, 2005). Franco and colleagues, showed the presence of a 27 kb region known as FHA locus (ctu_1p01140 to Ctu_1p01210) on the pCTU1 plasmid (accession; NC_013283) of *C. turicensis* z3032 encoding a filamentous haemagglutinin gene (*fhaB*), its specific transporter gene (*fhaC*) and 5 allied putative adhesins (Franco et al. 2011a). The entire FHA locus as a whole was not detected in any of the *C. sakazakii* CC4 or

non-CC4 strains (Figure 3. 8, Table S1), hence indicating that it may not be essential for the virulence of *C. sakazakii* in general or CC4 in particular. Five of the hypothetical proteins (ctu_1p01170-01210) within FHA locus were noted in genomes of *C. sakazakii* CC4 strains 553 and 557 and in non-CC4 strains 658 and 978. Moreover, the *fhaB* and its specific transporter gene (*fhaC*) was noted in 3 *C. sakazakii* non-CC4 isolates 1, 5 and 680 all of which were of clinical origin and belong to *Cronobacter* sequence type 8 (ST8). Whether the presence of an incomplete FHA locus is important in these isolates remains unclear (Table S1).

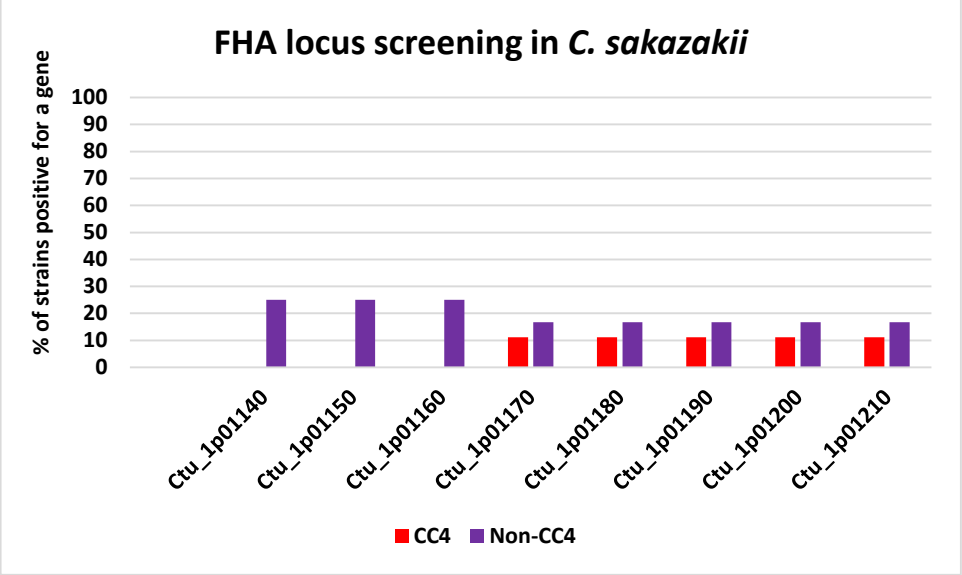


Figure 3. 8 The FHA locus genes screening in *C. sakazakii*.

The FHA locus genes were screened in 30 *C. sakazakii* strains. The figure is a graphical representation indicating the difference in the percentage of strains positive for the tested virulence associated traits. Overall the whole FHA locus was not detected in any of the CC4 (n=18) or non-CC4 strains (n=12).

3.5.2.5 MACROPHAGE SURVIVAL AND PROTEASE ACTVITY ASSOCIATED TRAITS

Like adhesion and invasion, macrophage survival is also crucial for bacterial pathogenicity; homologue of *pagC* gene (STM1246) required for *in vitro* macrophage survival of *Salmonella* Typhimurium (Pulkkinen & Miller, 1991) was not detected in any of the *C. sakazakii* genome while homologues of the *sodA* gene encoding for superoxide dismutase which has also been associated with macrophage survival was detected in all *C. sakazakii* genomes irrespective of their sequence type or clonal complex (Table S1).

Other putative virulence associated factors included zinc-containing metalloprotease (*zpx*) which has been reported to induce rounding of the Chinese Hamster Ovary (CHO) cells in tissue culture (Kothary et al. 2007). However homologues of a zinc metalloprotease gene

(CSSP291_03690) was noted in all *C. sakazakii* genomes (Table S1) except one CC4 isolate 721; hence not differentiating *C. sakazakii* CC4 from non-CC4. This observation correlates with the parallel PhD study by Hana Sonbol, in her study she performed protease activity assays by studying the expression of protease on milk agar at 20^{O}C for 72 hours. Her experiments included 23 sequenced *C. sakazakii* isolates, 15 CC4 and 8 non-CC4 (CC4 isolates; 553, 767, 6, 20, 1240, 721, 1219, 1220, 1221, 1225, 1231, 558, 557, 701 and 4. Non-CC4 isolates; 1, 5, 680, 658, 1218, 520, 696 and 1249) analysed in the present study. Her results indicated that all of the tested *C. sakazakii* isolates showed protease activity irrespective of their sequence type (Hana Sonbol, personal communications). In her assays, the CC4 isolate 721 lacking zinc metalloprotease encoded by CSSP291_03690 also showed protease activity. It must, however be noted that the bacterial genomes can have several homologues of proteases, hence whether this protease activity was caused by CSSP291_03690 or other protease encoding genes is unclear.

3.5.2.6 SERUM RESISTANCE ASSOCIATED TRAITS

Another important mechanism associated with bacterial pathogenicity is the evasion of the host bactericidal activity of serum, for instance in *Salmonella enterica* (serovar typhimurium) LT2, the *rck* gene was located on the plasmid pSLT and has been associated with serum resistance (Heffernan et al. 1992), however its homologue was not detected in any of the *C. sakazakii* genome (Figure 3.9; Table S1).

A plasmid encoded plasminogen activator called *cpa* (ESA_ pESA3p05434) helps in the activation of plasminogen and provides resistance to *C. sakazakii* against the bactericidal activity of the serum (Franco et al. 2011b). Genomic investigation indicated that like other virulence associated traits, *cpa* was also detected in most *C. sakazakii* strains (Figure 3.9; Table S1) except for *C. sakazakii* CC4 strain 6 and non CC4 strains 1, 520, 680 and ES15; all of which do not carry “full length” pESA3-like plasmid (section 3.4.5).

The serum resistance activity in *C. sakazakii* is likely to be regulated by the *cpa* gene as the present genomic analysis was validated by the serum resistance assays performed in a parallel PhD study by Sumyya Hariri. Her experiment included 22 sequenced *C. sakazakii* isolates, 16 CC4 and 6 non-CC4 (CC4 isolates; 377, 553, 767, 6, 20, 1240, 721, 1219, 1220, 1221, 1225, 1231, 558, 557 and 701, 4. Non-CC4 isolates; 5, 658, 1218, 520, 696 and 1249) analysed in the present study. Her results indicated that all of the tested isolates were able

to resist serum activity apart from the CC4 isolate 6 and non-CC4 isolate 520 which lack the plasmid pESA3 and hence *cpa*. Other isolates lacking plasmid pESA3, 1, 680 and ES15 were not part of her study (Sumyya Hariri personal communications). Hence serum resistance is not a CC4 specific phenomenon rather it is a plasmid borne characteristic in *C. sakazakii*.

3.5.2.7 HAEMOLYSIN ASSOCIATED TRAITS

Certain bacteria such as *E. coli* O157:H7 str. Sakai are able to induce haemolysis of red blood cells which is crucial towards the onset of severe illnesses like Haemolytic Uremic Syndrome (Wong et al. 2012). Six different putative haemolysin associated genes in *Cronobacter* have been listed; these include ESA_00102, ESA_00432, ESA_00643, ESA_02810, ESA_02937 and ESA_03540 encoding for haemolysin activator protein precursor, predicted membrane protein haemolysin III homologue, haemolysins and related proteins containing CBS domains, haemolysin expression modulating protein, putative haemolysin and a haemolysin precursor respectively (Joseph et al. 2012b). All of these genes were noted uniformly across CC4 and non-CC4 genomes (Figure 3.9, Table S1).

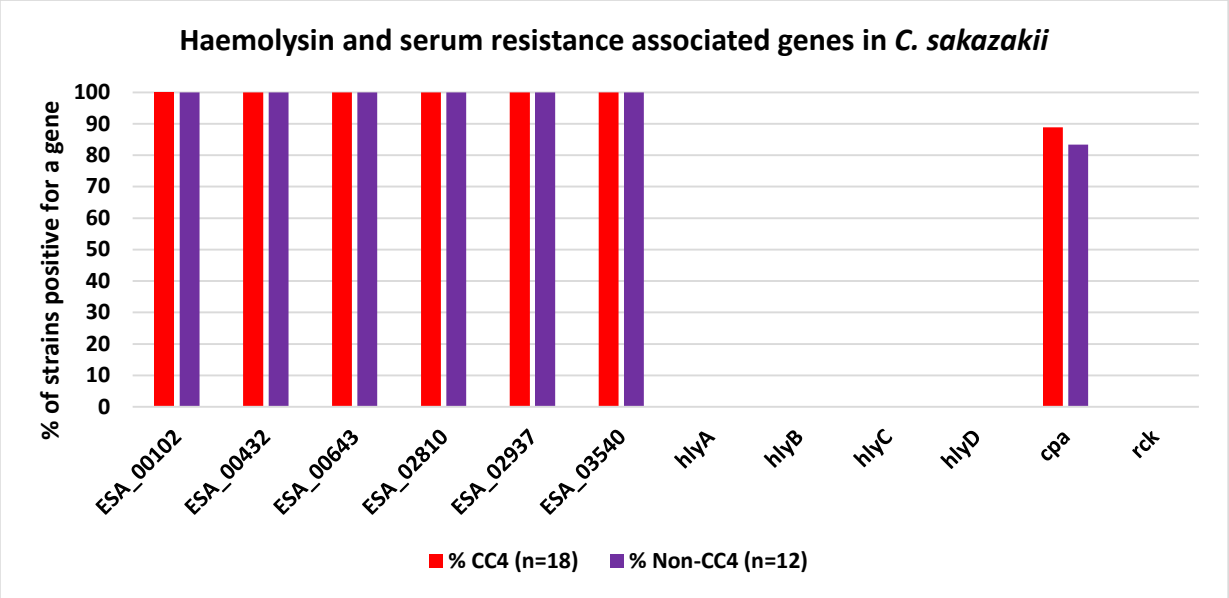


Figure 3. 9 Haemolysin associated genes screening in *C. sakazakii.*

The haemolysin associated genes were screened in 30 *C. sakazakii* strains. The figure is the graphical representation indicating the difference in the percentage of strains positive for the tested virulence associated traits. Overall the whole FHA locus was not detected in any of the CC4 (n=18) or non-CC4 strains (n=12).

Moreover, haemolysin associated genes *hlyABCD* located on the plasmid pO157 (accession: NC_002128) of *E. coli* O157:H7 str. Sakai were also searched for the presence of their homologues in 30 *C. sakazakii* genomes. The interrogation of *C. sakazakii* genomes for the

presence *hlyABCD* did not reveal any homologue in *C. sakazakii* genomes (Figure 3.9, Table S1).

These genetic features did not differentiate CC4 from non-CC4 isolates. Furthermore, 4 of the sequenced *C. sakazakii* isolates including 1 CC4 isolate 701 and 3 non-CC4 isolates 680, 696 and 658 were assayed by Hana Sonbol for the haemolysis activity on horse and sheep blood cells at 37^{o}C for 48 hours. Her study indicated that all of these 4 isolates were able to induce haemolysis on both horse (β haemolysis) and sheep blood (α haemolysis) (Hana Sonbol, personal communications).

3.5.2.8 TYPE 6 SECRETION SYSTEM

Type 6 secretion system (T6SS) is a newly described secretion system which has previously been associated with invasion of blood brain barrier in *E. coli* K1 (Zhou et al. 2012). Several T6SS associated genes have been noted on the genome of *C. sakazakii* BAA 894 (Kucerova et al. 2010; Joseph et al. 2012b). In this study, five different T6SS associated genes clusters were screened in *C. sakazakii* genomes, however all of these T6SS regions indicated a variable pattern of presence or absence and none of the clusters were found specific to *C. sakazakii* CC4 genomes (Figure 3.10, Table S1).

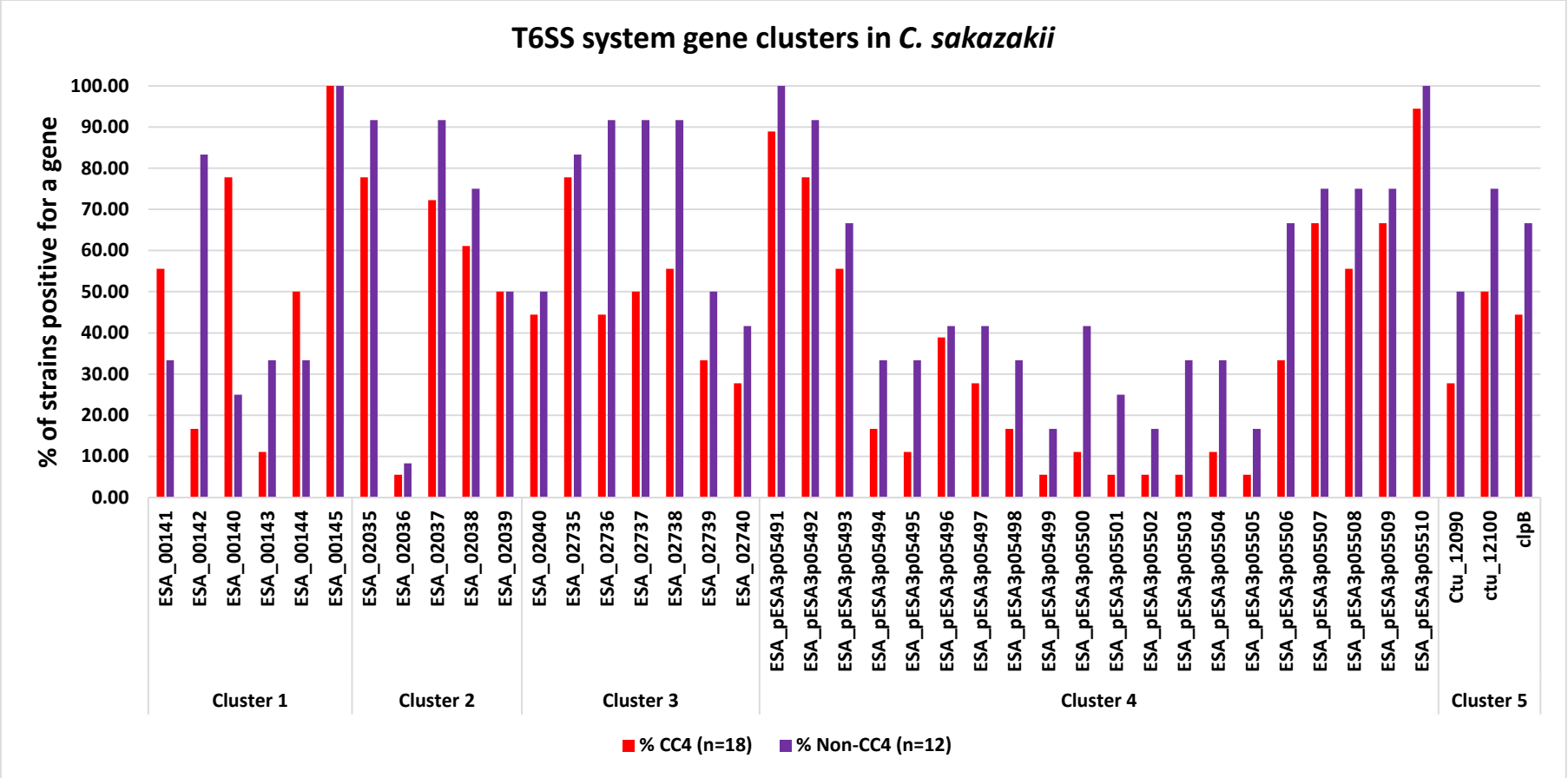


Figure 3. 10 Type 6 Secretion System associated gene clusters in *C. sakazakii.*

Five different T6SS associated gene clusters were screened in 30 *C. sakazakii* sequenced genomes. The figure is a graphical representation indicating the difference in the percentage of strains positive for any gene. A mosaic pattern for the presence of these T6SS associated genes clusters was noticed across CC4 (n=18) and non-CC4 (n=12).

3.5.2.9 COPPER AND SILVER RESISTANCE ASSOCIATED GENES

Metals such as copper and silver may be important for the bacterial survival and fitness, however, their accumulation in the bacterial cells can be toxic at higher concentrations. Copper is an important micronutrient which is required for the survival of the bacteria as it participates in the redox reactions by acting as a co-factor for the enzymes. However, due to its high chemical reactivity, the accumulation of copper can be a hazard for the bacteria. Therefore, bacteria have evolved mechanisms to control the copper traffic in and out of the cell to maintain the homoeostasis for copper synthesis and to avoid its potential toxic effects (Arguello et al. 2013). Two copper and silver resistance associated regions; ESA_04236-45 (*cusESRCFBA/silEP)* and ESA_04248–55 (*pcoABCERS*) have been reported in *C. sakazakii* (Kucerova et al. 2010; Joseph et al. 2012b). The annotation of these genes was verified using BLASTx analysis as well by aligning the genome of *C. sakazakii* 658 against pRJ1004 (accession no. X83541.1; Brown et al. 1995; Hao et al. 2015).

The genomes of 30 *C. sakazakii* isolates were screened which indicated that both of these clusters were present variably across two subsets of the genomes, indicating no CC4 specific pattern exists for the presence of these genes (Figure 3.11, Table S1).

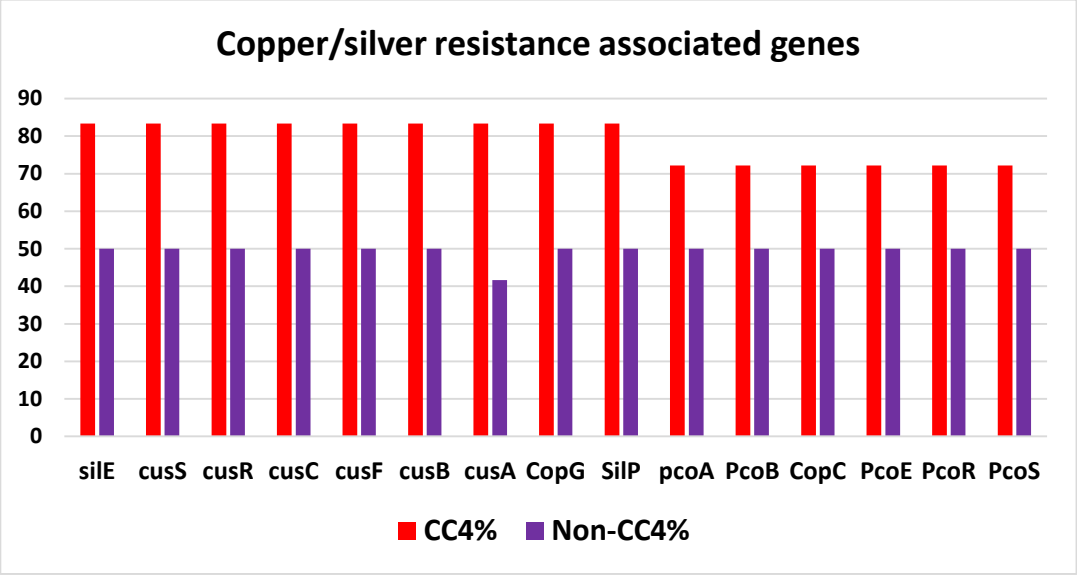


Figure 3. 11 Copper/silver resistance associated genes in *C. sakazakii.*

Two closely located copper and silver associated gene clusters; ESA_04236-45 (*cusESRCFBA/silEP)* and ESA_04248–55 (*pcoABCERS*) were screened in the 30 *C. sakazakii* genomes. The figure is a graphical representation indicating the difference in the percentage of strains positive for any gene. A mosaic pattern for the presence of these genes clusters was noticed across CC4 (n=18) and non-CC4 (n=12).

It was interesting to note that two high invasive *C. sakazakii* CC4 strains 767 and 701 lacked ESA_04248–55 (*pcoABCDR*). Additionally, both of these regions were absent in three *C. sakazakii* CC4 strains 20, 721 and 1221, the latter two being CSF isolates. Both of these regions were also missing in six of the non-CC4 strains 520, 1249, 140, 150, ES15 and 978; of these non-CC4 strains, 1249 was a fatal meningitic isolate (Table 3.1). The absence of the copper/silver resistance regions from the significant strains of clinical origin suggest that these regions may not be essential for the virulence of *C. sakazakii* in general or CC4 clade in particular.

3.5.2.10 IRON UPTAKE GENES

Iron is an essential nutrient for virtually all forms of life, its significance is based on its role in cellular processes such as energy generation, oxygen transport, DNA replication and oxidative stress response (Skaar, 2010). Bacterial pathogens have no exemption from iron as they require iron for their survival and growth, especially within the vertebrate host where they encountered iron limiting conditions. Amongst the first line of defences utilized by vertebrate hosts against bacterial infections is the process called “nutritional immunity” which refers to the withholding of nutrients from bacterial pathogens to prevent their outgrowth. The most important form of nutritional immunity is the sequestration of iron (Kehl-Fie and Skaar, 2009). In vertebrates, the vast majority of iron is intracellular, contained either in iron storage proteins called ferritin or complexed with heme (a cofactor of haemoglobin). Moreover, the neutral pH and the aerobic environment makes the extracellular iron insoluble. The strong binding of iron with vertebrate proteins transferrin/lactoferrin further enhances the vertebrate iron sequestration (Bullen and Griffiths, 1999). All of these factors significantly reduce the amount of freely available iron needed by bacterial pathogens to grow inside the vertebrate host and cause infection (Skaar, 2010).

To combat host iron sequestration, bacterial pathogens have evolved mechanism to survive and grow within limiting iron conditions. Sensing iron depletion, has evolved in bacterial pathogens as a marker for the vertebrate environment. This sensing controls the transcriptional regulation of an iron- dependent repressor called Fur (ferric uptake regulator) (Hantke, 1981). Fur operon controls the expression of iron-regulated genes, in the presence

of iron the genes are repressed while this repression is lifted in the absence of iron (Masse et al. 2007). The Fur iron regulation has been identified in bacterial pathogens such as *E. coli, Bacillus subtilis, Vibrio sp., Shigella flexneri,* and *Pseudomonas aeruginosa* (Masse et al. 2007).

Most bacterial pathogens, in order to compete with the host iron sequestration have generally evolved three iron uptake mechanisms; siderophore based iron uptake, heme acquisition system and lactoferrin/transferrin receptors (Figure 3.12).

Siderophores are low molecular weight, small chelating compounds secreted by bacterial pathogens which can scavenge iron from the host proteins such as transferrin and lactoferrin (Kehl-Fie and Skaar, 2009). Once removed from the host proteins, iron bound siderophores interact with cognate receptors on the bacterial surface and are internalized where iron is released (Kehl-Fie and Skaar, 2009).

Heme based iron acquisition system in bacterial pathogens usually involve surface receptors either recognizing heme or hemoproteins bound to heme such as hemoglobin. Heme is removed from the hemoproteins and is transferred to the bacterial cytoplasm where iron is removed and released through the action of heme oxygenases (Wilks, 2002).

In addition to iron acquisition from transferrin and lactoferrin through siderophore, some bacterial pathogens such as members of Neisseriaceae and Pasteurellaceae can directly recognise these proteins through specialised receptors (Ratledge and Dover, 2000). These proteins recognise human transferrins and subsequently transfer them to bacterial cytoplasm where iron is released and utilised as a nutrient (Skaar, 2010).

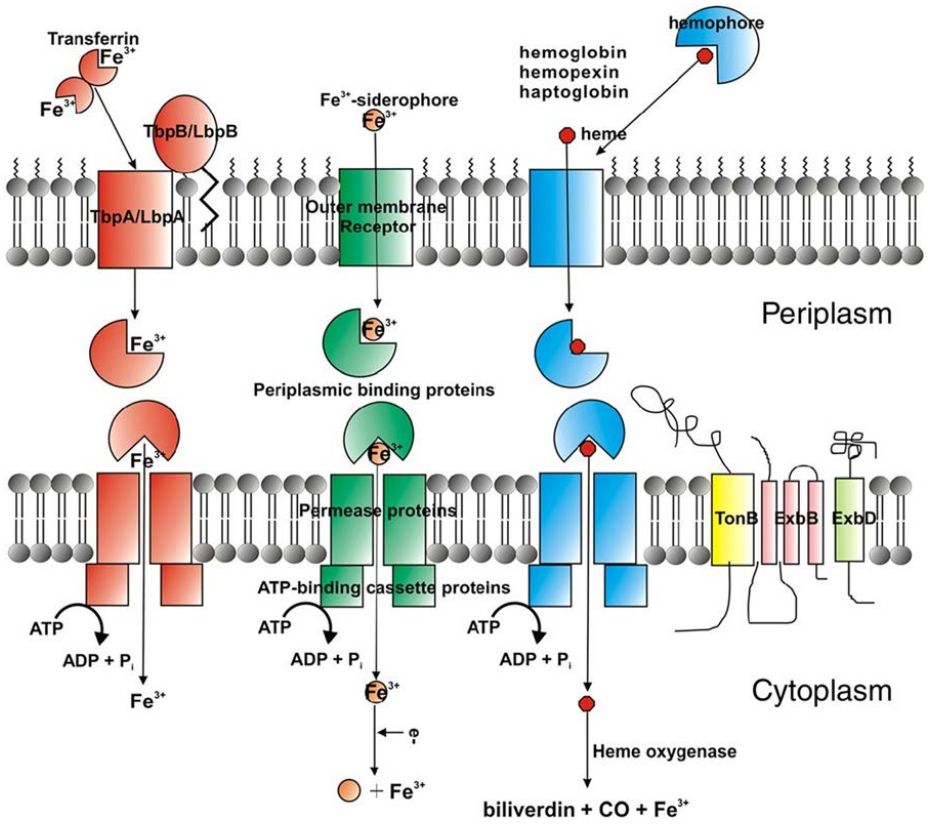


Figure 3. 12 Schematic representation of iron acquisition mechanisms in Gram negative bacteria (Krewulak and Vogel, 2007).

The figure shows the transferrin, siderophore and heme based iron acquisition systems. All of these iron uptake systems need an outer membrane receptor, a PBP, and an inner-membrane ABC transporter. Transport through the outer membrane receptors is mediated by the action of TonB system (TonB, ExbB, ExbD). All three systems are not present in all bacteria, however some pathogens have more than one iron acquisition system.

In *Cronobacter*, Several gene clusters associated with iron acquisition have been reported which include an enterobactin synthesis (*entABCDEFS*, ESA_00791-800) and transport (*fepABCDEG*; ESA_02727-31), a plasmid-borne operon for aerobactin synthesis (*iucABCD*) and its receptor *iutA* (ESA_pESA3p05547-51) and a hydroxamate type siderophore synthesis associated gene cluster (*fhuABCDE*, ESA_03187-90 & ESA_02242) (Kucerova et al. 2010; Joseph et al. 2012b). The present comparative genomic analysis did not reveal any particular association of *C. sakazakii* CC4 genomes with the above mentioned iron acquisition and transport associated genes as these genes were noted on most of the *C. sakazakii* strains analysed. The plasmid borne iron acquisition region *iutA-iucABCD* was found in all except for plasmid-less *C. sakazakii* CC4 strain 6 and two *C. sakazakii* non-CC4 strains 520 and ES15 (section 3.4.5). Hence, none of the iron acquisition genes showed any particular association with *C. sakazakii* CC4 genomes (Figure 3.13, Table S1).

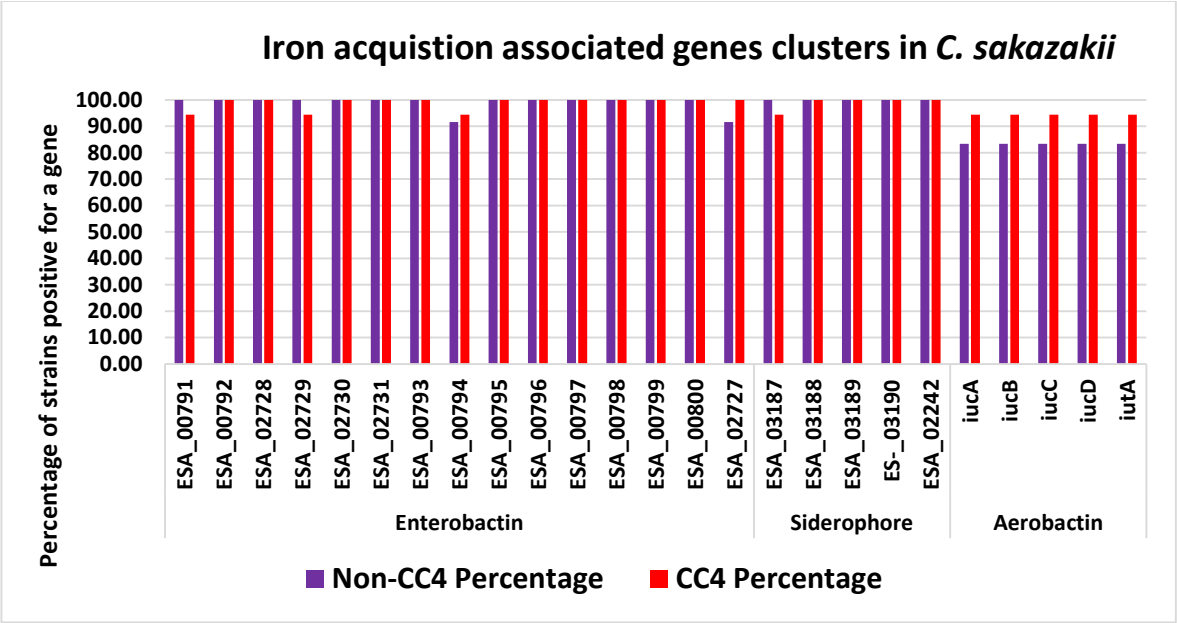


Figure 3. 13 Iron acquisition associated genes in *C. sakazakii*

Three different iron acquisition associated gene clusters; enterobactin, siderophore and aerobactin were screened in *C. sakazakii* sequenced genomes (n=30). The figure is a graphical representation indicating the difference in the percentage of CC4 strains (n=18) and non-CC4 strains (n=12) positive for any gene. All of these genes were detected predominantly across both subsets of *C. sakazakii.*

3.5.2.11 SIALIC ACID UTILISATION GENES

Another important virulence associated attribute in Gram negative bacteria is the utilisation of sialic acid, for example sialic acid utilisation by neonatal meningitic *E. coli* K1 has been reported (Vimr et al. 2004; Severi et al. 2007). Sialic acid is present in breast milk intestinal mucin, human milk and gangliosides in brain (Wang, 2009). Additionally, sialic acid is also a supplement in some infant formula (Joseph et al. 2013). More recently, our group has shown the ability of *C. sakazakii* to utilise sialic acid (Joseph et al. 2013). Previous studies by our group have described the presence of *yhcH* encoding a putative sugar isomerase and sialic acid utilisation genes *nanAKTR* (encoding N-acetylneuraminate and N-acetylmannosamine degradation), *nanC* (outer membrane porin) and *nanE* (ManNAc-6-P epimerase) uniquely in the genomes of *C. sakazakii* (Joseph et al. 2012b). The *nanE* gene is usually found clustered with *nanAKTR*, however in *C. sakazakii* this gene is located distantly as in case of *Citrobacter freundii* and *Edwardsiella tarda* (Joseph et al. 2013). Additionally *nagA* (encoding for GlcNAc-6-P deacetylase) and *nagB* (encoding for glucosamine-6-P

deaminase) which contributes to the synthesis of fructose-6-phosphate were also noted in *C. sakazakii* genomes which support the role of sialic acid as carbon source in *C. sakazakii* (Joseph et al. 2013). In brief, *nanC* encodes for the outer membrane porin protein. The *nanT* gene encodes a major facilitator superfamily (MFS) protein which transports the sialic acid into the bacterial cell. After the transport of the sialic acid into the cell, Neu5Ac lyase encoded by *nanA* converts it into N-acetylmannosamine (ManNAc) and phosphoenolpyruvate (PEP). The *NanK* enocodes for ATP-dependant kinase which is specific for generating N-acetylmannosamine-6-phosphate (ManNAc-6-P) from ManNAc. The *nanE* encodes for ManNAc-6-P epimerase which then converts ManNAc-6-P into N-acetylglucosamine-6-phosphate (GlcNAc-6-P). The products of *nagA* and *nagB* then catalyse the conversion of GlcNAc-6-P into fructose-6-phosphate which is a substrate for the glycolytic pathway. The *nanR* is the repressor involved in the regulation of the sialic acid genes activity.

The ability of *C. sakazakii* to grow on sialic acid and on ganglioside GM1 media has been demonstrated in a parallel laboratory based study by Sumyya Hariri (Joseph et al. 2013). An additional TRAP transporter *siaPQM* was also noted on all *Cronobacter* genomes, however since only *C. sakazakii* showed the growth on sialic acid, therefore it indicates that *nanT* is the only active transporter in *C. sakazakii* (Joseph et al. 2013). Presence of the sialic acid utilisation genes could be a crucial factor for the predominance of *C. sakazakii* in neonatal infections and in infant formula. Since amongst *C. sakazakii*, the clonal lineage *C. sakazakii* CC4 is the most dominant in neonatal infections and in infant formula, it was important to investigate whether it is linked with sialic acid utilisation. As previous study (Joseph et al. 2013) only investigated 5 of the sequenced *C. sakazakii* genomes, therefore in the present study the genomic investigation was expanded to include 30 *C. sakazakii* sequenced isolates in order to observe if there is any variation for the presence of sialic acid genes in *C. sakazakii* CC4 and non-CC4. The *nanAKTR*, *nanC*, *nanE, nagA, nagB* and *siaPQM* were all screened for their presence in 30 *C. sakazakii* genomes analysed in this study. In agreement with previous study by our group, all of these genes were noted in 100% of the *C. sakazakii* CC4 and non-CC4 genomes and hence not indicating a variation between CC4 and non-CC4 (Table S1). It also emphasises the fact all *C. sakazakii* are able to utilise exogenous sialic acid as a carbon source.

3.5.3 ENVIRONMENTAL FITNESS ASSOCIATED TRAITS

3.5.3.1 STRESS RESPONSE GENES

Members of *Cronobacter*, especially *C. sakazakii* are known to be better able to resist environmental stresses such as heat, acidic conditions and osmotic shock compared with other Enterobacteriaceae (Nazarowec-White & Farber, 1997; Breeuwer et al. 2003; Edelson-Mammel et al. 2005; Dancer et al. 2009). The interrogation of the sequenced *C. sakazakii* CC4 and non-CC4 isolates indicated that all of these isolates harboured a number of genes annotated as stress response associated genes.

3.5.3.2 POLYMORPHISM IN *rpoS* GENE IN SEQUENCED *C. SAKAZAKII* ISOLATES

The alternative sigma factor *rpoS* is well recognised in modulating the general stress response in Gram-negative bacteria (Dodd and Aldsworth, 2002). A recent study by Ordonez et al. (2012) indicated that the *rpoS* gene encoding for the alternative sigma factor σ contributes to stress resistance in *C. sakazakii* since according to their study a positive correlation between the *rpoS* activity and resistance against acid, alkaline and osmotic stresses was observed (Ordonez et al. 2012). Their study suggested that variability in the gene sequence is associated with the stress response and the full length *rpoS* gene sequence is important for the *rpoS* functionality as large disruptions in the gene sequence can hamper its activity thereby altering the response of bacterial isolate against stress (Ordonez et al. 2012). Therefore, the *rpoS* gene sequence was analysed in detail in the sequenced *C. sakazakii* isolates in the present study to investigate if there are any CC4 specific variations within the *rpoS* gene. The homologue of the *rpoS* gene (CSSP291_02780) was detected uniformly across CC4 and non-CC4 genomes analysed in the present study, which was expected as the gene is distributed widely across the Gram negative bacteria. The *rpoS* gene sequence of the each isolate was extracted by aligning it against the *rpoS* gene (CSSP291_02780) of *C. sakazakii* SP291. In order to check the variation in the gene sequence, the *rpoS* gene sequences of 18 CC4 and 12 non-CC4 were aligned using phylogeny.fr (http://phylogeny.lirmm.fr/phylo_cgi/index.cgi/) and phylogeny determined (Figure 3.14). As seen in the figure 3.14, variations were observed within CC4 and non-CC4 *C. sakazakii* isolates. Apart from the CC4 isolate 6, the remaining 17 CC4 isolates were located on the same branch; however variations were noted within the CC4 cluster as 7 of the CC4 isolates; 4, 721, 1219, 1220, 1221, 1225 and 1240 were located on a sub-branch within CC4. Similarly

CC4 isolate 1231 showed a slightly longer branch length. The *C. sakazakii* CC4 isolate 6 was located on a different branch with a non-CC4 isolate 5 (ST8). Variations can also be seen within 3 of the ST8 isolates; 1, 5 and 680 (Figure 3.14). A closer look at the sequence alignment indicated an overall high degree of sequence consensus across CC4 and non-CC4 isolates used in the analysis; the only exception was a non-CC4 isolate 978 which had a deletion of 87 nucleotides at the beginning of the gene sequence which could be due to sequencing or assembly error (Figure 3.15, Table 3.3). The variations on the phylogenetic tree were attributed only to 17 SNPs observed between 30 isolates. The major difference between CC4 (except *C. sakazakii* 6) and non-CC4 isolates on the phylogenetic tree was only due to two SNPs at positions 372 and 558 (Figure 3.15, Table 3.3). The *C. sakazakii* CC4 isolate 6 which did not cluster with other CC4 strains had 6 SNPs shared with non-CC4 isolates (Figure 3.15, Table 3.3).

Further analysis was undertaken to study the type of these SNPs. *In silico* conversion of *rpoS* nucleotide sequence to amino acids (http://web.expasy.org/translate/) and their alignment using "phylogeny.fr" indicated 99% of amino acid sequence conservation. Only two of the 17 SNPs were found to be non-synonymous. In isolate 680, corresponding to the nucleotide position 209 (amino acid position 70), a Leucine (L) was replaced by (P) Proline. In isolates 5 and 6, corresponding to the nucleotide position 544 (amino acid position 182), a Threonine (T) was replaced by Serine (S). In addition, a deletion mutation was also noted; in isolates 1218 and 658, both belonging to ST1, had a deletion of Glutamine at nucleotide position 601 (amino acid position 201). Whether these non-synonymous SNPs affect the stress response of these isolates warrants further investigation.

Due to very high sequence conservation of *rpoS,* both at nucleotide and amino acid level in CC4 and non-CC4 isolates, it seems unlikely that it can may any unique stress tolerance ability to *C. sakazakii* CC4. As explained in the following sections, the stress response assays such as heat tolerance, osmotolerance and acid resistance assays in a parallel PhD study did not form the basis of differentiation between CC4 and non-CC4 isolates (Hana Sonbol, personal communications).

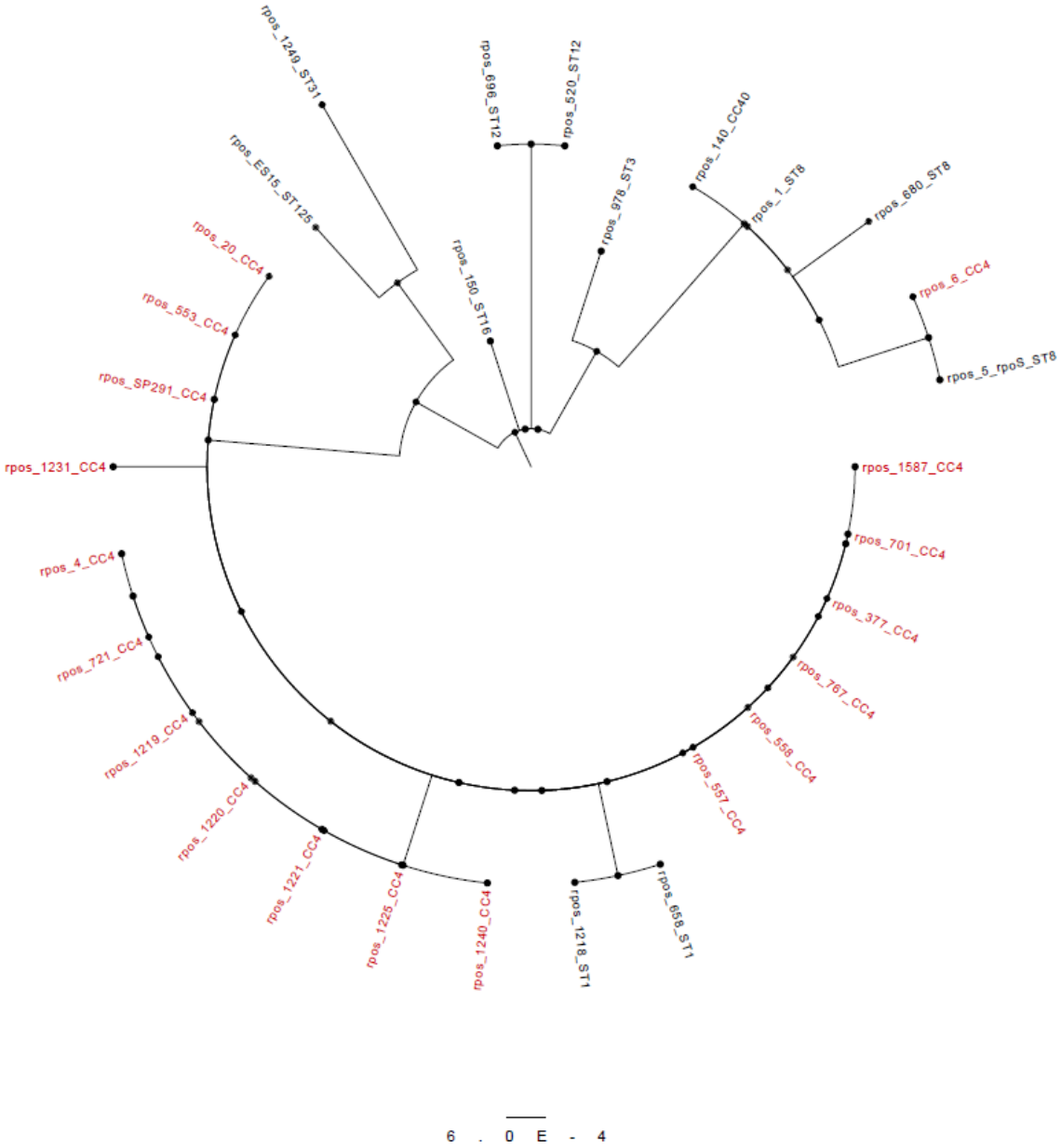


Figure 3. 14 The phylogeny of 30 *C. sakazakii* isolates based on *rpoS* gene sequence.

The figure shows the midpoint rooted phylogenetic tree for the 30 *C. sakazakii* isolates based on the *rpoS* gene sequence (993 bp). The alignment was constructed using phylogeny.fr and the tree annotated using FigTree. The red font colour indicates *C. sakazakii* CC4 (n=18) while black font colour indicates *C. sakazakii* non-CC4 isolates (n=12).

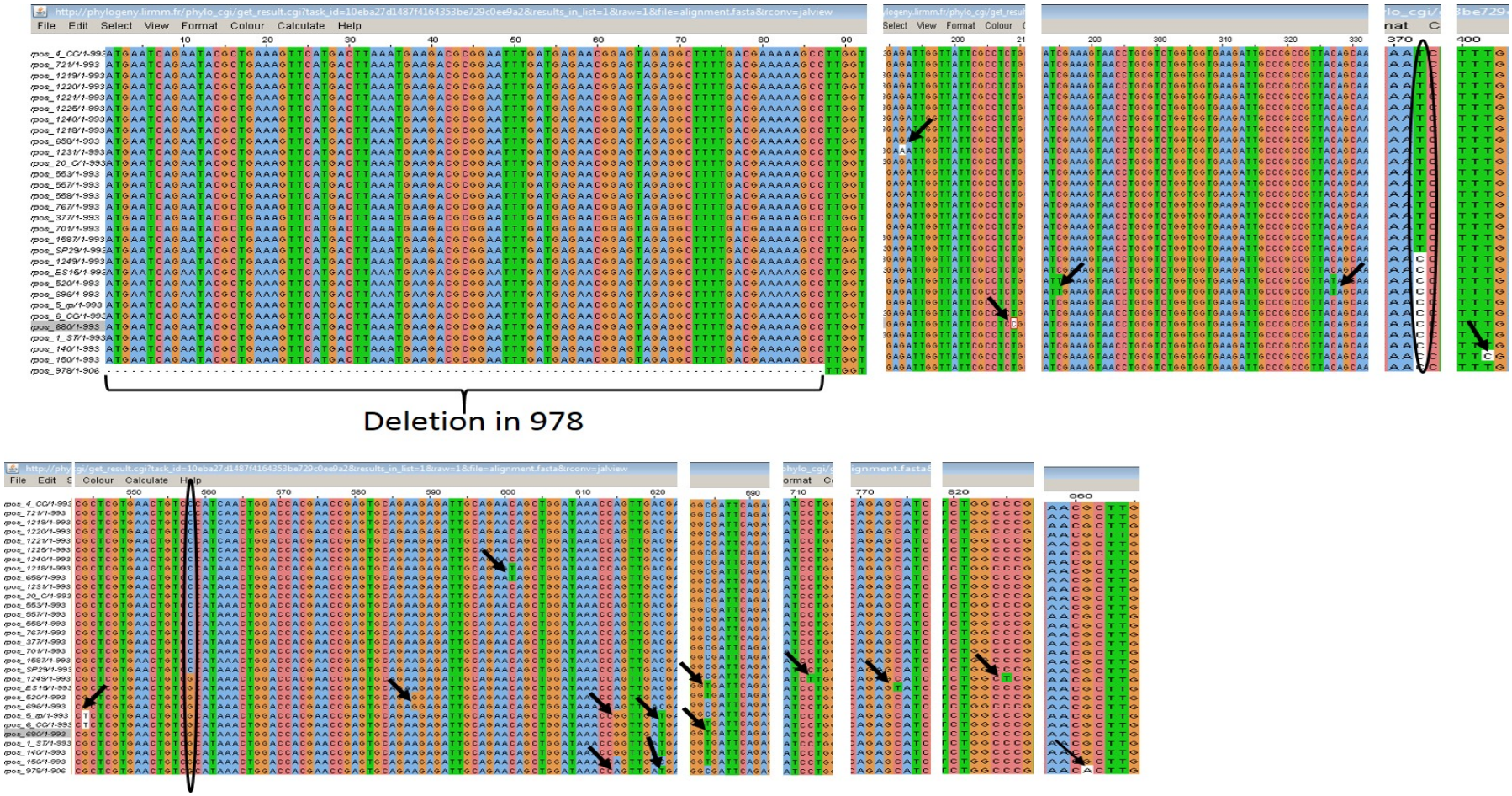


Figure 3. 15 Manual curation of the *rpoS* gene sequence alignment.

The figure above represents the cropped segments of the screen shots combined as one image showing only the sections where variations in the *rpoS* gene sequence was observed. The black arrows point towards the Single Nucleotide Variations (SNP) changes in *C. sakazakii* isolates. The black encircled areas represent areas of CC4 and non-CC4 variation.

Table 3. 3 Nucleotide variations observed in *rpoS* gene sequence alignment

No	SNP Position	*C. sakazakii* Isolates	Nucleotide substitution (with reference to general consensus)	Type of SNP
1	192	1231	G → A	Synonymous
2	209	680	T → C	Non-synonymous L → P Leucine to Proline
3	285, 327	696, 520	C → T	Synonymous
4	372	1249, ES15, 520, 696, 5, 6*, 680, 1, 140, 150, 978	C → T	Synonymous
5	402	150	T → C	Synonymous
6	544	5, 6*	G → T	Synonymous T → S Threonine to Serine
7	558	1249, ES15, 520, 696, 5, 6*, 680, 1, 140, 150, 978	C → G	Synonymous
8	587	520, 696	A → G	Synonymous
9	601	1218, 658	C → T	Deletion deletion at amino acid position 201, in isolates 1218 and 658 Q → - Glutamine deleted
10	615, 621	140, 1, 680, 6*, 5	A → G	Synonymous
11	621	140, 1, 680, 6*, 5	C → T	Synonymous
12	621	978	C → T	Synonymous
13	684	140, 1, 680, 6*, 5, ES15, 1249	C → T	Synonymous
14	712	1249	C → T	Synonymous
15	774	ES15	C → T	Synonymous
16	825	1249	C → T	Synonymous
17	861	978	G → A	Synonymous

Green font colour shows the nucleotide in general consensus, Red font colour shows the substituted nucleotide, *; CC4 isolate

3.5.3.3 GENETIC REGION ASSOCIATED WITH THERMOTOLERANCE IN *C. SAKAZAKII*

The increased stress tolerance of *Cronobacter* such as thermotolerance and desiccation can enhance its survival in the environment of PIF manufacturing factories which can increase its chances of contaminating the PIF thereby growing the risk of exposure to neonate. A number of studies have been conducted on thermoresistance of *Cronobacter*; although variations in the D-values have been observed between different studies, yet most of these studies suggest the thermotolerance of *Cronobacter* is either equivalent or higher than other Enterobacteriaceae (Nazarowec-White & Farber, 1997; Breeuwer et al. 2003; Iversen et al. 2004). A positive correlation between thermotolerance and desiccation has also been proposed for *Cronobacter* (Dancer et al. 2009). Gajdosova and colleagues (2011) identified an 18 kb long region in *C. sakazakii* ATCC 29544 associated with its thermotolerance at 58^{o}C. The region consisted of 22 open reading frames (*orf*) having homologies with stress response proteins such as heat, oxidation, stress response and acid resistance (Gajdosova et al. 2011). The region contained transposases at both ends. Gajdosova and colleagues propose its acquisition from plants, since plant pathogens such as *Burkholderia* and *Ralstonia* species also contain an homologous genetic region (Forsythe, 2004; Friedmann 2007; Gajdosova et al. 2011). Furthermore, the authors experimentally confirmed the association of *orf*HIJK with thermotolerance as the cloning of these *orf*s into *E. coli* resulted in two fold increase in its D_{58} value (Gajdosova et al. 2011). Additionally, the *orf*I was found to be a homologue of the hypothetical protein (Mfla_1165) in thermotolerant *Methylobacillus flagellatus* KT; the authors found a positive correlation with the presence of *orf*I and elevated thermotolerance in *C. sakazakii* and *C. malonaticus* at 58^{o}C (Gajdosova et al. 2011).

As discussed earlier, the predominant association of *C. sakazakii* CC4 with neonatal meningitis and its occurrence in the environment especially in the PIF manufacturing environment is of high concern as it can increase its exposure to neonate, therefore it was of great significance to investigate the genomic region identified by Gajdosova et al. (2011) in the sequenced *C. sakazakii* genomes in order to observe its variation across *C. sakazakii* CC4 and non-CC4 genomes. To investigate this, the whole 18 kb region was accessed from Genbank (accession; FR714908). Each of the 22 *orf*s including the transposases were screened for their presence in sequenced *C. sakazakii* genomes using *Cronobacter* BLAST and ACT. The findings of this analysis were quite interesting as the entire region (except for

transposases *orf*T5 and *orf*T1) varied greatly between *C. sakazakii* CC4 and *C. sakazakii* non-CC4 (Figure 3.16, Table 3.4). Most interestingly, the region *orf*HIJK which had been experimentally confirmed by Gajdosova et al. (2011) to be associated with high thermotolerance (after cloning in *E. coli*), was detected in *C. sakazakii* CC4 genomes significantly more ($P < 0.05$) in CC4 isolates (50% of CC4 isolates) than *C. sakazakii* non-CC4 isolates (less than 10% of non-CC4 isolates).

Some interesting observations were also noted. The *C. sakazakii* CC4 genomes 701 and 767 both of which were fatal isolates from the French outbreak of 1994 and isolated only 26 days apart varied for the presence *orfHIJK*. The *orfHIJK* was detected in the *C. sakazakii* 767 while absent from the *C. sakazakii* CC4 isolate 701.

An important observation was that the *C. sakazakii* CC4 isolate SP291 which was originally isolated from the PIF manufacturing factory lacked the *orfHIJK* region (Power et al. 2013). The CC4 isolate *C. sakazakii* SP291 has been regarded as the "persistent thermotolerant isolate" by Power and colleague (Power et al. 2013). The absence of the *orf*HIJK from *C. sakazakii* SP291 questions the association of this thermotolerant region with thermotolerance.

Furthermore the genomic analysis did not correlate with the laboratory based dry heat tolerance assays performed by Hana Sonbol. In her study, Hana Sonbol, after desiccation treatment subjected 7 CC4 strains (1221, 701, 721, 1587, 1542, 1537, 1533) and 6 non-CC4 strains (658, 1536, 520, 696, 680 and 1) to determine the effect of dry heat on the survival of these isolates at 60^{o}C, 80^{o}C and 100^{o}C. The results of Hana Sonbol suggested that CC4 strains were more heat tolerant than non-CC4 strains at 100^{o}C while at 60^{o}C, 80^{o}C no significant difference in the viability was observed (Hana Sonbol, personal communications). It must be noted that the CC4 isolates used by Hana Sonbol included isolates 1221, 701, 1587 and 721, all of which lack the *orfHIJK* which have been proposed as key thermotolerant genes in *C. sakazakii* by Gajdosova et al. (2011). Hence, the laboratory experiments did not correlate with the genomic analysis presented here.

The analysis of the genes indicated a significant variation for their presence across CC4 and non-CC4 genomes. However the absence of the key thermotolerant genes proposed by Gajdosova et al. (2011) in a thermotolerant CC4 isolate SP291 and its absence in the fatal meningitic CC4 isolates 701 and 1587 and CSF isolates 721 and 1221 weakens the likelihood

that these are the key genes in regulating the thermotolerance of *C. sakazakii*. Furthermore, these observations did not correlate with the laboratory studies performed in a parallel PhD study by Hana Sonbol as discussed above. Hence, although the association of these thermotolerant genes with thermotolerance cannot be negated, the observations in the present study suggest that these may not be the only genes responsible for thermotolerance in *C. sakazakii*.

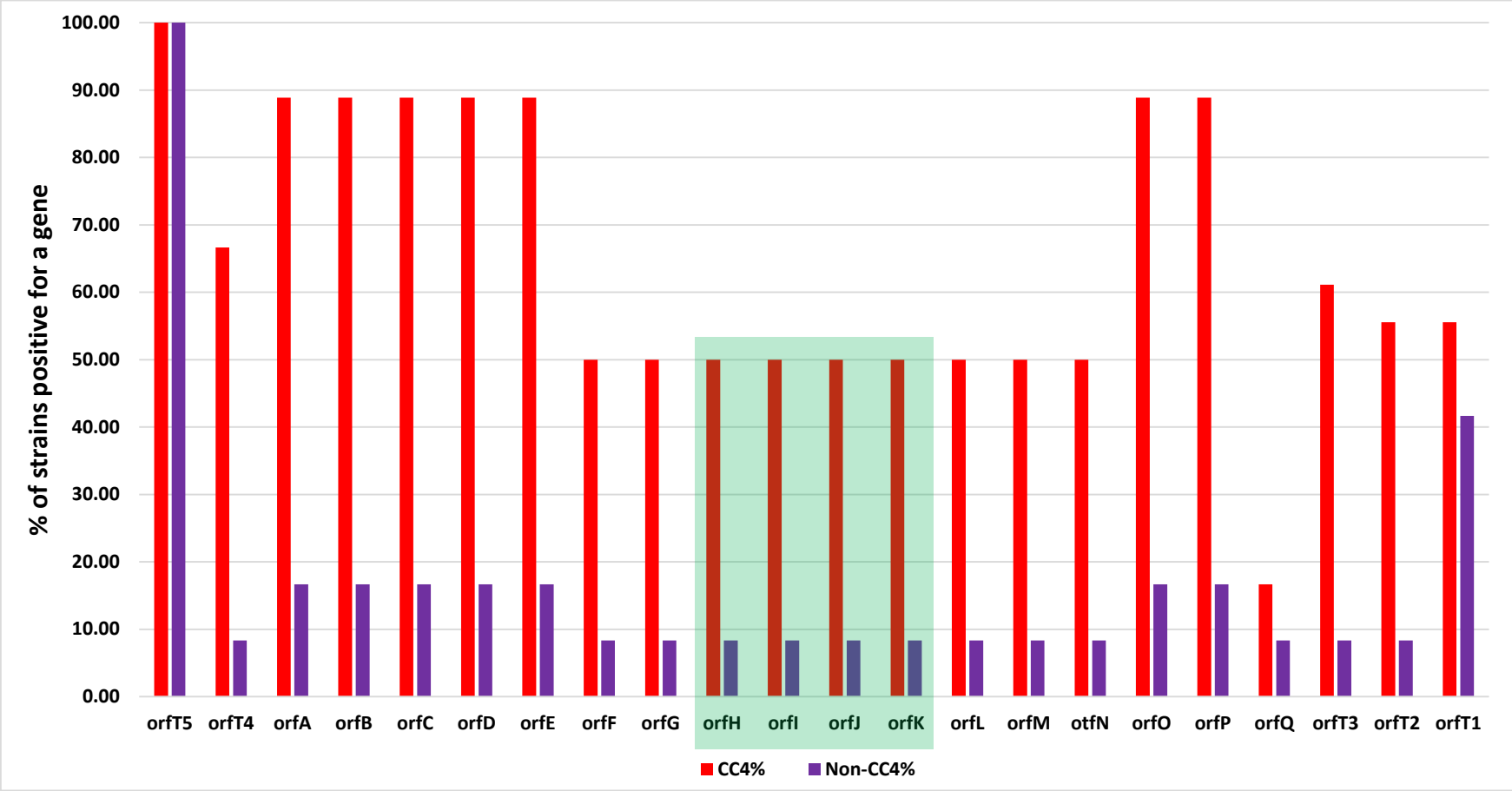


Figure 3. 16 Thermotolerance associated genetic region in *C. sakazakii* genomes.

The genomes of 30 *C. sakazakii* isolates (n=30) were screened for the presence of homologues of 22 *orfs* of thermotolerance region of *C. sakazakii* ATTCC 29544. The data is presented in the graph as a percentage of *C. sakazakii* CC4 (n=18) and non-CC4 genomes (n=12) positive for an *orf* under investigation. Red bars indicate the percentage of CC4 genomes positive for an *orf* while purple bars indicate the non-CC4 genomes positive for an *orf*. The green shaded area indicate the genes orfHIJK experimentally linked with thermotolerance.

Table 3. 4 Genome screening of *C. sakazakii* genomes for the presence of thermotolerance region homologue of *C. sakazakii* ATCC 29544

Strain	ST/CC	*orfT5* Put. transposase	*orfT4* Put. transposase	*orfA* Hypothetical protein	*orfB* Put. phage transcriptional regulator/Put. excisionase,	*orfC* Put. small heat shock protein Hsp-20	*orfD* Put. ATPase with chaperone activity, clpB	*orfE* Hypothetical protein	*orfF* ypothetical protein	*orfG* Hypothetical protein	*orfH* Put. small heat shock protein Hsp-20	*orfI* Put. YfdX family	*orfJ* Put. YfdX family	*orfK* Put. conservative transmembrane protein, HdeD	*orfL* Conserved Hypothetical protein	*orfM* Put. thioredoxin	*orfN* Put. sodium/hydrogen exchanger	*orfO* Put. conserved hypotetical protein	*orfP* Put. peptidase M48	*orfQ* Put. DegP2 peptidase	*orfT3* Put. transposase	*orfT2* Put. transposase IS4 family	*orfT1* Put. transposase fragment
1221	4/4	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-
721	4/4	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
701	4/4	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+
767	4/4	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
377	4/4	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
1587	109/4	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
SP291	4/4	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
553	4/4	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
1231	4/4	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
4	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
6	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
20	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
557	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
558	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1219	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
1220	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+

Table 3. 4 continued

Strain	ST/CC	*orfT5* Put. transposase	*orfT4* Put. transposase	*orfA* Hypothetical protein	*orfB* Put. phage transcriptional regulator/Put. excisionase,	*orfC* Put. small heat shock protein Hsp-20	*orfD* Put. ATPase with chaperone activity, clpB	*orfE* Hypothetical protein	*orfF* Hypothetical protein	*orfG* Hypothetical protein	*orfH* Put. small heat shock protein Hsp-20	*orfI* Put. YfdX family	*orfJ* Put. YfdX family	*orfK* Put. conservative transmembrane protein, HdeD	*orfL* Conserved Hypothetical protein	*orfM* Put. thioredoxin	*orfN* Put. sodium/hydrogen exchanger	*orfO* Put. conserved hypotetical protein	*orfP* Put. peptidase M48	*orfQ* Put. DegP2 peptidase	*orfT3* Put. transposase	*orfT2* Put. transposase IS4 family	*orfT1* Put. transposase fragment
1225	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
1240	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
1	8/8	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
696	12	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+
5	8/8	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
680	8/8	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
140	40/140	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
150	16/16	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
520	12	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
658	1/1	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1218	1/1	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1249	31	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ES15	125/100	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
978	3/3	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

orf = open reading frame, **+** = isolate containing an *orf*, **-** = isolate not containing an *orf*, **%** = refers to the percentage of CC4 or non-CC4 isolates positive for an *orf* , nd = not defined, put. = putative, the green shadowed are indicates *orfhijk* experimentally confirmed to be linked with thermotolerance after cloning in *E. coli* by (Gajdosova et al. 2011).

3.5.3.4 ACID RESISTANCE GENE

In order to cause infection in neonates, bacterial cells should have the ability to grow in the acidified environment of the neonate's gut. The extremely acidic environment of human stomach (pH 1-3) not only facilitates digestion of food but also act as a strong barrier against food borne-pathogens (Smith, 2003). Certain food borne-pathogens such as *E. coli* and *Helicobacter pylori* have the ability to survive this strong host defence and cause infection (Gordon and Small, 1993).

A recent study by Alvarez-Ordonez et al. (2014) identified the *ompR* gene as a key player in response of *C. sakazakii* against highly acidic conditions (Alvarez-Ordonez et al. 2014). The *ompR* gene is a homologue of the gene ESA_04334 encoding for osmolarity response regulator (*ompR*). In this study, the genomes of 30 *C. sakazakii* strains were screened for the presence of *ompR*. The genomic analysis revealed that *ompR* was present in all *C. sakazakii* strain. The presence of *ompR* in all CC4 and non-CC4 strains suggested that it is not specific to CC4 (Table S1). In a parallel PhD study by Hana Sonbol, 7 CC4 strains (1221, 701, 721, 1587, 1542, 1537, 1533) and 6 non-CC4 strains (658, 1536, 520, 696, 680 and 1) were assayed for their ability to resist exposure to pH3.5 for 2 hours. No significant difference between CC4 and non-CC4 isolates was observed as most of the isolates showed only a 2 log reduction after 2 hours exposure to pH 3.5 (Hana Sonbol, personal communications).

3.5.3.5 BIOFILM AND CAPSULE FORMATION

Biofilm formation together with stress response is crucial for bacteria to survive under stressful conditions. The present genome analysis detected a biofilm associated operon *bcsBEFZ* (encoding for cellulose biosynthesis) in *C. sakazakii* 377 genome (Masood et al. 2013a), however no specific association of biofilm forming genes was attributed to *C. sakazakii* CC4 since *bcsBEFZ* homologues were noted in all *C. sakazakii* CC4 and non-CC4 genomes. Similarly, Hartmann et al. (2010) showed that two hypothetical proteins (ESA_00281 and ESA_00282) and at least three flagellar genes *flhE* (ESA_01356), *fliD* (ESA_01287) and *flgJ* (ESA_02266) contribute to adhesion of *C. sakazakii* ES5; but present genomic analysis could not detect the association of these genes with any particular group of *C. sakazakii* as all of these genes were detected uniformly in both *C. sakazakii* CC4 and non-CC4 genomes (Table S1). A previous PhD study in our group showed that both tested

C. sakazakii CC4 (4, 721, 1219, 1220, 1221, 1225, 1231, 1240, 767, 557, 558) and *C. sakazakii* non-CC4 isolates (1, 5, 658, 520, 1249) irrespective of their sequence type were able to form biofilms on soya based, casein and whey based formula, although variation in the degree of biofilm was observed at different temperatures and different milk formula (Nassrudin Rhouma 2012, unpublished data).

The capsular polysaccharides on the surface of the bacterial cells can be important for their pathogenicity as well as for their interaction with environment (Joseph et al. 2012b; Caubilla-Barron et al. 2007). The capsule production and biofilm formation has been linked in *Cronobacter* (Caubilla-Barron et al. 2007). Approximately 25 kb long cluster of genes (ESA_01155-01175; *wzABCKM*) has been reported in *Cronobacter* encoding for colanic acid (Joseph et al. 2012b). The genomic interrogation in the present study revealed that this region as a whole was detected in majority of the CC4 and non-CC4 isolates; the exceptions were *C. sakazakii* CC4 isolates 1240 and 557 and non-CC4 isolates 140 and 696 where the region was partially present (Table S1). Moreover, in a previous PhD study, Nassrudin Rhouma tested 18 *C. sakazakii* CC4 isolates (767, 557, 558, 6, 4, 20, 1225, 721, 1220, 1222, 1223, 1231, 1224, 1219, 553, 1221, 1240, 1242) and 11 *C. sakazakii* non-CC4 isolates (658, 680, 5, 1, 520, 580, 1249, 1241, 1019, 555, 12) for capsule production on the milk agar. Of these strains, only 4 CC4 isolates (20, 1219, 1220, 1225) while 5 non-CC4 isolates (1, 5, 680, 520, 1249) were negative for capsule production on milk agar; all of the remaining CC4 and non-CC4 isolates were able to produce capsule on the milk agar (Nassrudin Rhouma 2012, unpublished data). Hence, biofilm formation and capsule production cannot be regarded as a *C. sakazakii* CC4 specific phenomenon.

3.5.3.6 OSMOTOLERANCE ASSOCIATED TRAITS

The predominance of *C. sakazakii* CC4 in PIF and PIF manufacturing environment suggests that this clonal lineage of *C. sakazakii* may have unique stress tolerance genes which favour its growth in PIF (a_w of ~0.2) and the surrounding desiccated environment. A number of different osmotolerance genes have been described previously in *Cronobacter* and other Enterobacteriaceae*,* for example Feeney and colleagues (2014) described seven homologues of the Prop osmolyte uptake system; ESA_02131, ESA_01706, ESA_04214, ESA_pESA3p05450, ESA_01226, ESA_00673 and ESA_03328 in *C. sakazakii* BAA-894 genome (Feeney et al. 2014). Another study by the same group identified homologues of *E. coli* genes *TrkH, trkG, trkA, trkE, kdpA, kdpB, kdpC, kdpD, kdpE, phoP, phoQ, ompC, ompF, envZ* and *ompR* associated with uptake of K^+ which act as osmoprotectants as a primary response to osmotic stress (Feeney et al. 2011) . The present genomic analysis showed that most of these genes were uniformly present in *C. sakazakii* genomes proposing that osmotolerance may not be unique to CC4 but it is a common stress survival mechanism adapted by *C. sakazakii* (Figure 3.19; Table S1).

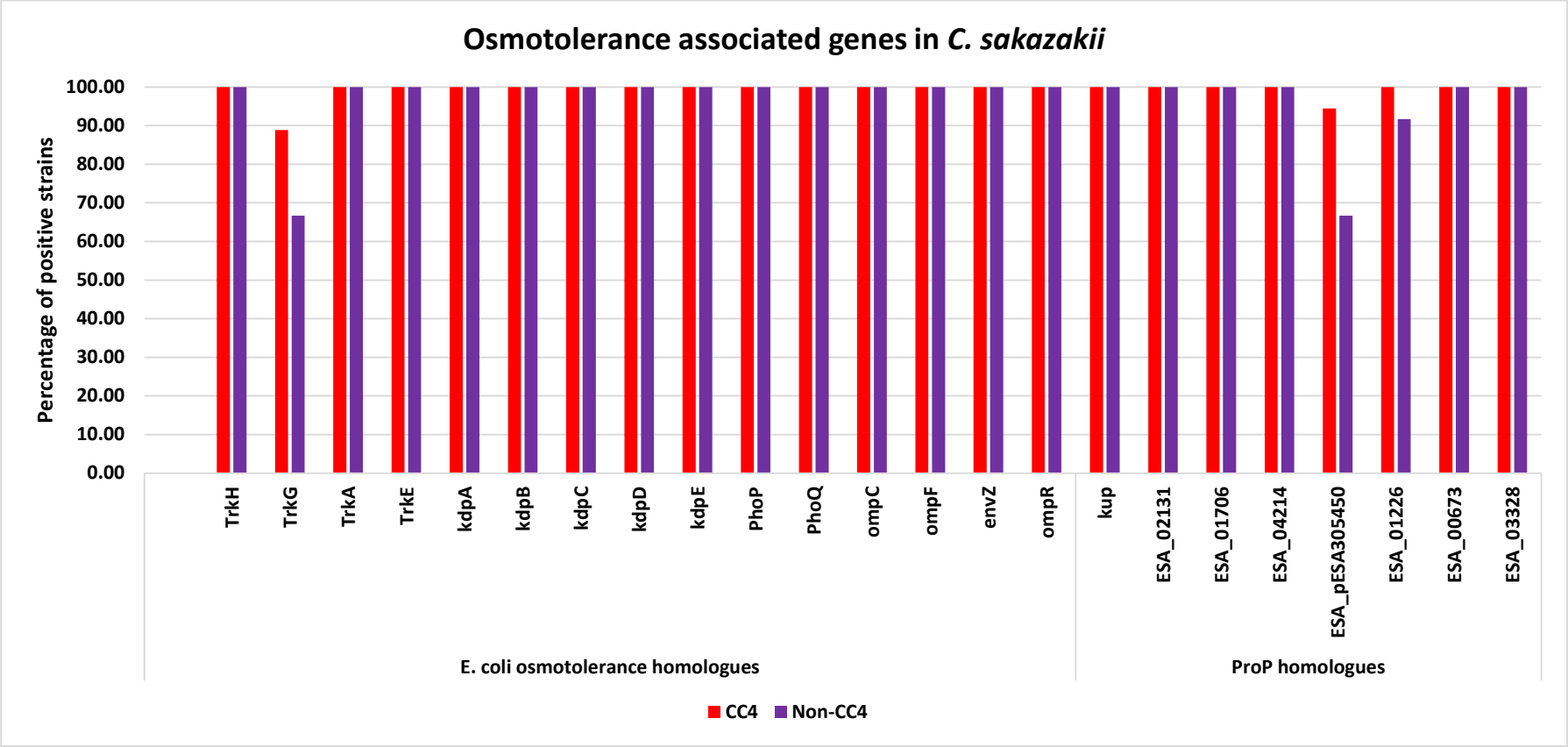


Figure 3. 17 Osmotolerance associated genes screening in *C. sakazakii.*

The osmotolerance associated genes previously described in *E. coli* and *C. sakazakii* were screened in 30 *C. sakazakii* strains for the presence of their homologues. The figure is a graphical representation indicating the difference in the percentage of CC4 (n=18) and non-CC4 (n=12) strains positive for the tested genes. Most of the genes were predominantly present across both sus-sets of *C. sakazakii*.

3.5.4 *IN SILICO* SEROTYPING OF *C. SAKAZAKII* STRAINS

In Gram-negative bacteria, the O-antigen is a highly divergent part of the lipopolysaccharide (LPS). The sugar residues present within the O-antigen region vary in their number and order which constitute a characteristic serotype (Mullane et al. 2008; Jarvis et al. 2011; Sun et al. 2012). The O-antigen region in *Cronobacter* consists of two conserved genes UTP glucose-1-phosphate uridylyltransferase subunit (*galF*) and dTDP-D-glucose-4, 6-dehydratase (*rfbB*) while the rest of the genes are divergent in this genus (Kucerova et al. 2010; Joseph et al. 2012b). To date at least 17 different serotypes have described in *Cronobacter* of which seven O-antigen types (O:1-O:7) have been described in *C. sakazakii* (Mullane et al. 2008; Jarvis et al. 2011, 2013; Sun et al. 2012) (Table 3.5).

Table 3. 5 Defined serotypes in *Cronobacter sakazakii*

No	O-antigen	Isolate	Accession	Length (bp)	Reference
1	O:1	*C. sakazakii* strains NCTC 8155	EU076545	12875	Mullane et al. 2008
2	O:2	*C. sakazakii* strains NCTC 11468	EU076546	12868	Mullane et al. 2008
3	O:3	*C. sakazakii* 2156	HQ646168	13701	Jarvis et al. 2011
4	O:4	*C. sakazakii* strain G2594	JQ674747	11887	Sun et al. 2012
5	O:5	*C. sakazakii* strain G2706	JQ674748	12303	Sun et al. 2012
6	O:6	*C. sakazakii* strain G2704	JQ674749	7388	Sun et al. 2012
7	O:7	*C. sakazakii* strain G2592	JQ674750	8932	Sun et al. 2012

Given the high clinical significance of neonatal meningitis, it was of interest whether the O-antigen region could provide a unique clonal signature for *C. sakazakii* CC4 strains. In order to compare the O-antigen region of *C. sakazakii* strains used in this study, the genomic regions corresponding to 7 defined serotypes were obtained from Genbank (Table 3.5) and compared against all 30 *C. sakazakii* genomes used in this study using *Cronobacter* BLAST and ACT. The results of the genomic comparison indicated that 30 *C. sakazakii* isolates belonged only to 4 serotypes i.e. O:1-O:4. The most significant observation in this analysis was that except for one CC4 isolate 1240, all of the remaining 17 CC4 isolates were found to be O:2 serotype. The CC4 isolate 1240 showed complete alignment with O:4. Nevertheless, the O:2 serotype was not unique to CC4 as two of the non-CC4 isolates 978 (ST3) and 1249 (ST31) were also O:2. Hence, although O:2 was found to be the dominant serotype in CC4, yet it was not exclusive to CC4. Moreover, the O:2 cannot be termed as

unique signature for neonatal meningitis, as two of the fatal meningitic non-CC4 isolates 658 and 1218 belonged to O:1 serotype (Table 3.5). Furthermore, the data analysis by an independent genomic data analysis of 107 *Cronobacter* isolates in *Cronobacter* PubMLST has indicated that the O:2 is the dominant serotype in *C. sakazakii* isolates belonging to at least 19 different sequence types (STs; 4, 15 ,109, 218, 64, 268, 261, 267, 22, 282, 31, 287, 12,17, 50, 262, 13, 86, 3) (Khaled Ibrahim, personal communications; unpublished data).

Table 3. 6 Serotypes designation to *C. sakazakii* CC4 strains based on the genomic comparison

Isolate	% identity	Alignment length	Reference	Reference Length (bp)	Designated serotype	ST	CC
1	100*	12875	EU076545	12875	O:1	8	8
5	100*	12875	EU076545	12875	O:1	8	8
680	100*	12875	EU076545	12875	O:1	8	8
150	100*	12875	EU076545	12875	O:1	16	16
1218	97.54	13018	EU076545	12875	O:1	1	1
658	99.05	12876	EU076545	12875	O:1	1	1
ES15	99.05	12876	EU076545	12875	O:1	125	100
1220	98.65	12956	EU076546	12868	O:2	4	4
767	99.33	12914	EU076546	12868	O:2	4	4
377	100	12868	EU076546	12868	O:2	4	4
558	99.99	12868	EU076546	12868	O:2	4	4
1587	99.98	12868	EU076546	12868	O:2	109	4
553	100*	12868	EU076546	12868	O:2	4	4
SP291	100	12868	EU076546	12868	O:2	4	4
1219	100*	12868	EU076546	12868	O:2	4	4
1220	100*	12868	EU076546	12868	O:2	4	4
1221	100*	12868	EU076546	12868	O:2	4	4
1225	100*	12868	EU076546	12868	O:2	4	4
1231	100*	12868	EU076546	12868	O:2	4	4
6	100*	12868	EU076546	12868	O:2	4	4
20	100*	12868	EU076546	12868	O:2	4	4
4	100*	12868	EU076546	12868	O:2	15	4
701	100*	12868	EU076546	12868	O:2	4	4
721	100*	12868	EU076546	12868	O:2	4	4
978	**100***	**12868**	**EU076546**	**12868**	**O:2**	**3**	**3**
1249	**100***	**12868**	**EU076546**	**12868**	**O:2**	**31**	**31**
520	100*	13701	HQ646168	13701*	O:3	12	
696	99.85	11886	JQ674747	11887	O:4	12	
1240	**99.02**	**11979**	**JQ674747**	**11887**	**O:4**	**4**	**4**
140	99.76	11900	JQ674747	11887	O:4	40	45

ST; Sequence type, CC; Clonal complex, bp; base pair, *; Based on ACT comparison

3.5.5 *IN SILICO* PLASMID PROFILING

Whole genome sequencing of *C. sakazakii* BAA-894 and *C. turicensis* revealed that their genomes possess repFIB like plasmids, pESA3 (131 kb) and pCTU1 (138 kb) which are almost similar in size. These plasmids have been regarded as virulence plasmids (Kucerova et al. 2010; Franco et al. 2011a). Both of these plasmid share a common backbone and differ only in a few traits. Both plasmids possess a repFIB replication origin gene (*repA*), iron acquisition system *eitCBAD* and *iucABCDE/iutA* (section 3.4.2.10) and a *Cronobacter* plasminogen activator (*cpa*) (section 3.4.2.6). In addition pESA3 contains a 17 kb T6SS (section 3.4.2.8) while pCTU1 contains a 24 kb FHA locus (section 3.4.2.4) (Franco et al. 2011a). These traits have already been investigated independently in this chapter. No antibiotic resistance associated traits were found on these plasmids. It was important to screen *C. sakazakii* genomes analysed in this study for the presence of full length plasmids and to observe if there is any variation between *C. sakazakii* CC4 and non-CC4 genomes regarding the presence of full length plasmid.

3.5.5.1 ALIGNMENT OF pESA3 PLASMID WITH *C. SAKAZAKII* GENOMES

The 131 kb plasmid pESA3 was aligned against 18 *C. sakazakii* CC4 and 11 *C. sakazakii* non-CC4 genomes (please note that the *C. sakazakii* BAA-894 was excluded since the reference plasmid belongs to this strain) using Blast Ring Image Generator (BRIG) (Figure 3.20). The BRIG alignment indicated that the plasmid was completely absent only in one *C. sakazakii* CC4 strain 6 while it was completely absent from two of the *C. sakazakii* non-CC4 isolates; ES15 and 520. This finding is consistent with the laboratory based plasmid profiling in a parallel PhD study by Sumyya Hariri. The plasmid profiling experiment performed by Sumyya Hariri included 23 sequenced isolates (analysed in the present study) of which only strains 6 and 520 appear to lack plasmid, (please note that isolate ES15 was accessed from Genbank and is not present in our culture collection) (Sumyya Hariri; personal communications, unpublished data). In addition one *C. sakazakii* CC4 isolate 557 while two non-CC4 isolates 1 and 680 contained only partial length pESA3 plasmid; the CC4 isolate 557 showed homology with only ~90 kb while non-CC4 isolates 1 and 680 showed homology with only ~70 kb region of pESA3 plasmid. The *C. sakazakii* non-CC4 isolate 696 showed interspersed homology across the length of the pESA3 plasmid which is likely to be the sequencing or genome assembly artefact; please note this genome was accessed from the

Genbank and was not sequenced as part of the present study (Table 3.1). Approximately 10 kb region consisting of 10 hypothetical proteins (ESA_pESA3p05493 to ESA_pESA3p05504) was found to be largely missing in a majority of the *C. sakazakii* CC4 and non-CC4 isolates. This region has also been investigated previously as one of the T6SS system clusters in section (section 3.4.2.8). It remains unclear as to why this region has been lost from most of the *C. sakazakii* plasmids. An important finding in this analysis was that *C. sakazakii* CC4 strain 6 completely lacked the plasmid pESA3; interestingly this strain has been found to be low invasive on BMEC using *in vitro* tissue culture studies by Faisal Almajed as part of his PhD study (personal communications; unpublished data). The detailed comparative genomic analysis of high and low invasive *C. sakazakii* genomes is presented in chapter 7. In addition, a starvation sensing protein encoded by the *rspA* gene (ESA_pESA3p05434) was absent only in the isolates which either completely or partially lack the pESA3 plasmid i.e. 6, 557, 520, 680, 1 and ES15 (Table S1).

In brief, the pESA3 plasmid showed variable presence across CC4 and non-CC4 isolates; the plasmid was missing completely in one CC4 isolate 6 and two non-CC4 isolates 520 and ES15. The CC4 isolate 557 and non-CC4 isolates 1 and 680 only contained partial length pESA3 plasmid.

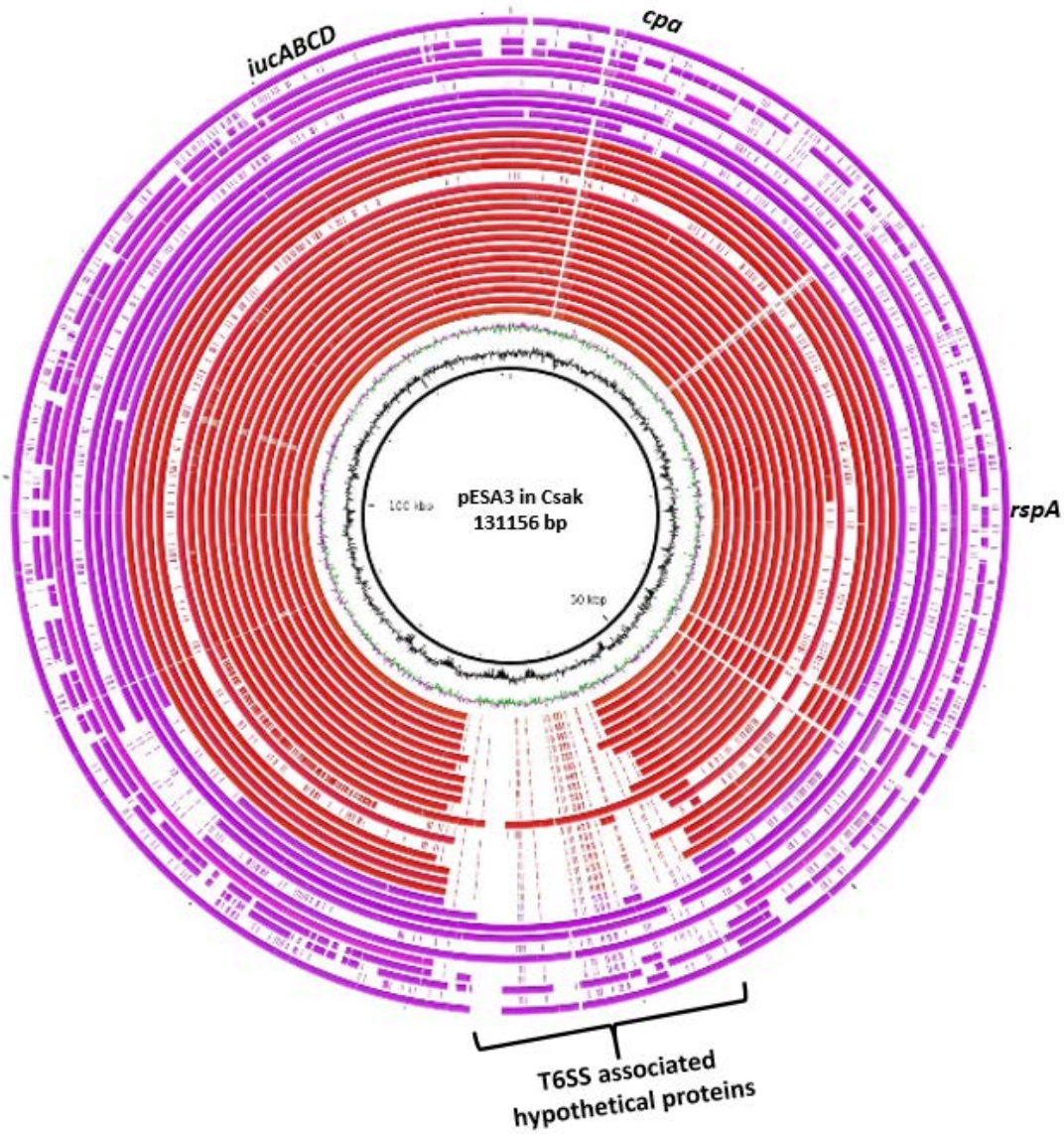


Figure 3. 18 BRIG alignment of the pESA3 plasmid with *C. sakazakii* isolates.

The image above represents the alignment of the pESA3 plasmid (131 kb) with 18 *C. sakazakii* CC4 (red rings) and 11 *C. sakazakii* non-CC4 (purple rings). The legend on the left hand side of the figure indicates ring for each isolate. The central solid black circle represent the reference plasmid backbone, the second and third rings from inside out represent the GC contents and GC skew respectively. The red coloured rings No. 4-21 from inside out represent the CC4 isolates 1219, 1220, 1221, 1225, 1231, 1240, 1587, 20, 377, 4, 553, 557, 558, 6, 721, 767, 701 and SP21 respectively. A large T6SS associated region can be seen missing in most *C. sakazakii* CC4 (except isolate 553).The purple coloured rings No. 22-32 from inside out represent isolates 1249, 1, 1218, 140, 520, 978, 150, 680, 696, ES15 and 5 respectively.

The BRIG alignment with plasmid pCTUI (138 kb) also revealed a similar pattern as with pESA3 (Figure S2), reassuring that both of these plasmids share a similar backbone. Additional 27 kb FHA region encoded on the plasmid pCTU1 has been investigated in section 3.4.2.4.

3.5.6 SCREENING OF THE NEONATAL MENINGITIC *ESCHERICHIA COLI* (NMEC) VIRULENCE DETERMINANTS IN *C. SAKAZAKII*

Amongst the bacterial pathogens causing neonatal meningitis, *E. coli* remains the second most common cause after group B streptococci (GBS) (Bonacorsi & Bingen, 2005; May, 2005). Majority of the NMEC strains belong to the *E. coli* phylogenetic group B2 and are distributed in a few clones (Bonacorsi & Bingen, 2005). Amongst the NMEC *E. coli,* strains having K1 capsule are predominant and represent approximately 80% of the NMEC associated cases (Glode et al. 1977; Kim et al. 1992). The O18:K1:H7 is considered to be the most common clone of NMEC and is distributed worldwide while other clones such as O83:K1 and O45:K1 are restricted to only specific parts of the world (Bonacorsi & Bingen, 2005). Over the years, considerable progress has been made to understand the pathogenic mechanism of *E. coli* K1, but the virulence associated determinants described in *E. coli* K1 are not sufficient to understand their virulence potential in the global context. The *E. coli* K1 is associated predominantly with neonatal meningitis which is induced by a complicated multistage process (Kim, 2001, 2002). This involves the colonisation of the gastrointestinal tract (GIT), mucous membrane translocation, intravascular space invasion, survival and multiplication. Once a threshold level for bacteraemia is reached then the bacteria penetrates the blood brain barrier (BBB) and invades the central nervous system (CNS) (Dietzman, 1974; Kim, 2001). The entry of the bacteria into CNS causes an increased movement of the white blood cells across BBB (pleocytosis) as well as release of toxic and pro-inflammatory compounds ultimately leading to meningitis; a swelling of meninges which provides protective covering to brain and spinal cord (Dietzman, 1974; Kim, 2001, 2002). A number of virulence associated determinates have been described in *E. coli* K1 of which ompA and ibe proteins, *Asla*, *TraJ* and *cnf1* genes are considered essential for its virulence mechanism which include transversal of BBB and *in vitro* invasion of HMBEC (Bonacorsi & Bingen, 2005). In addition, more recently, the plasmid of *E. coli* K1 S88 strain has been associated with high level of bacteraemia in neonatal meningitic rat model (Peigne et al. 2009). Since the present study was focused on screening the virulence associated traits in neonatal meningitic *C. sakazakii* CC4, it was therefore important to analyse whether these virulence associated traits which have been described in *E. coli* K1, are also present in the genomes of *C. sakazakii* and to determine if any of these traits are specific to *C. sakazakii* CC4.

The variation of *C. sakazakii* CC4 for *ompA* at the nucleotide and amino acid level has been covered in section 3.4.2.3. The *E. coli* K1 traits including *aslA* gene, the ibe proteins, *TraJ* and *cnf1* were screened in *C. sakazakii* genomes and have been discussed here.

3.5.6.1 The *aslA* gene

A study by Hoffman et al. (2000) elucidated the role of *aslA* (encoding for the arylsulfatase enzymes which contain highly conserved sulfatase motifs). The mutants of the *aslA* gene constructed in their study by targeted gene disruption and gene deletion showed reduced invasion of the HBMEC *in vitro* and *in vivo*. The genomes of 30 *C. sakazakii* genomes in this study were screened for the presence of this gene, however the gene was detected in none of the *C. sakazakii* isolates (Table 3.7, Figure 3.21).

3.5.6.2 The ibes proteins

A number of studies based on the transposon mutagenesis, gene deletion and complementation have revealed the ibe proteins which include ibeA, ibeB and ibeC are involved in the invasion of HMBMEC in *E. coli* K1, of these *ibeA* gene was reported specific to *E. coli* K1 while *ibeB* and *ibeC* have homologues *p77211, yijP* in *E. coli* K12 (Huang et al. 1999; Wang et al. 1999; Huang et al 2001, Wang & Kim, 2002). The *ibeA, ibeB* and *ibeC* gene screening in *C. sakazakii* genomes revealed that *ibeA* and *ibeC* were not detected in any of the *C. sakazakii* isolates while *ibeB* was detected in 83% and 50% of *C. sakazakii* CC4 and non-CC4 isolates respectively (Figure 3.21). However, the BLASTx analysis of the *ibeB* gene against *Cronobacter* revealed homology against the copper/silver efflux system outer membrane protein CusC (ESA_04239) in *C. sakazakii* 658, and has already been covered in section 3.4.2.9. This could be due to the sequence similarities between these two genes. Furthermore, the absence of the *ibeB* homologue in *C. sakazakii* i.e. CusC in clinically significant *C. sakazakii* CC4 and non-CC4 isolates such as strains 20, 721 and 1249 indicates that it may not be essential towards the virulence of *C. sakazakii* in general and *C. sakazakii* CC4 in particular.

3.5.6.3 The *TraJ* gene

A study by Hill and colleagues (2003) demonstrated the role of *Traj* in the *E. coli* K1 virulence. Their study indicated that the *TraJ* mutants were less virulent than the wild type and showed reduced ability to disseminate from the mesenteric lymph nodes tissue to the

deeper tissue of the spleen and liver. Moreover, their study also indicated that *TraJ* contributes towards the macrophage survival and therefore to the intracellular nature of the *E. coli* K1 as *TraJ* mutants were less frequently internalized by macrophages as compared to the wild type (Hill et al. 2003). The *TraJ* of *E. coli* K1 was screened in *C. sakazakii* genomes for the presence of its homologue, however it was not detected in any of the *C. sakazakii* genomes (Table 3.7, Figure 3.21).

3.5.6.4 The cytotoxic necrotising factor 1

The cytotoxic necrotising factor 1 (*cnf1*) has been reported to play a role towards the invasion of BMECs in *E. coli* K1. A study by Khan et al. (2002) has shown that *cnf1* contributed towards the invasion of BMECs *in vitro* through modulation of the cytoskeletal rearrangements by activating RhoA and penetration of the central nervous system *in vivo* (Khan et al. 2002). The gene sequence of *cnf1* was screened in *C. sakazakii* genomes using *Cronobacter* BLAST for the presence of its homologue, however none of the isolates indicated its presence (Table 3.7, Figure 3.21).

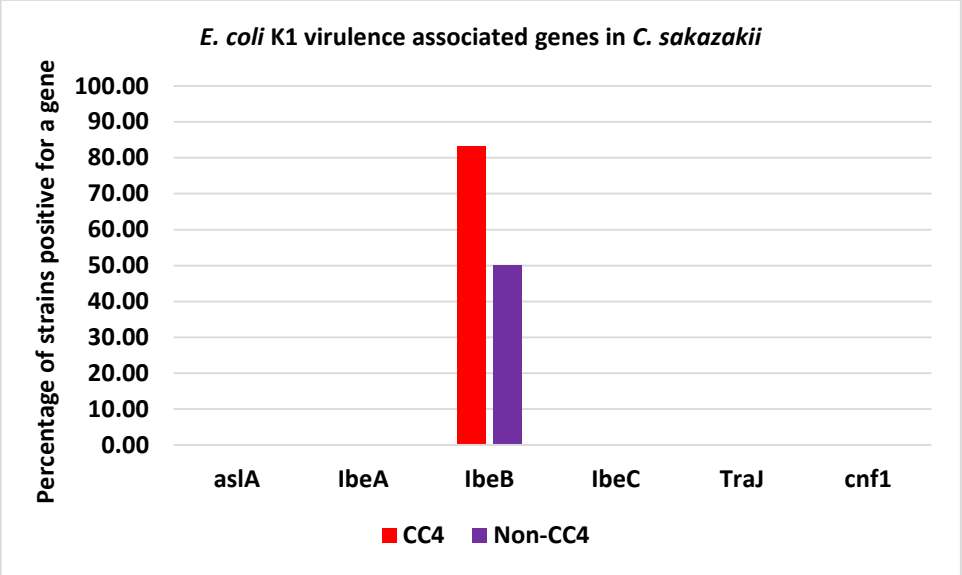


Figure 3. 19 The *E. coli* K1 virulence associated gene screening in *C. sakazakii* strains.

The homologues of the *aslA, IbeA, IbeB, IbeC*, *TraJ* and *cnf1*. The chart represents the percentage of the CC4 (n=18) and non-CC4 (n=12) positive *C. sakazakii* isolates for the presence of the genes. Except for *IbeB*, none of the homologue was not detected in any of the *C. sakazakii* genomes.

Table 3. 7 *E. coli* K1 virulence associated genes screening in *C. sakazakii*.

		***E. coli* K1 gene**	***aslA***	***IbeA***	***IbeB****	***IbeC***	***TraJ***	***cnf1***
		Access.	AAG10151	AAF98391	AAD30205	AAD28716	AAA92657	CAA50007
		Putative Function	Invasion	Invasion	Invasion	Invasion	Macrophage survival	Invasion
		Ref.	Hoffman et al. 2000	Huang et al. 2001	Huang et al. 2001	Wang et al. 1999	Hill et al. 2003	Falbo et al. 1993, Khan et al. 2002
***C. sakazakii* isolate**	**ST**	**CC**						
6	**4**	**4**	-	-	+	-	-	-
20	**4**	**4**	-	-	-	-	-	-
377	**4**	**4**	-	-	+	-	-	-
553	**4**	**4**	-	-	+	-	-	-
557	**4**	**4**	-	-	+	-	-	-
558	**4**	**4**	-	-	+	-	-	-
701	**4**	**4**	-	-	+	-	-	-
721	**4**	**4**	-	-	-	-	-	-
767	**4**	**4**	-	-	+	-	-	-
1219	**4**	**4**	-	-	+	-	-	-
1220	**4**	**4**	-	-	+	-	-	-
1221	**4**	**4**	-	-	-	-	-	-
1225	**4**	**4**	-	-	+	-	-	-
1231	**4**	**4**	-	-	+	-	-	-
1240	**4**	**4**	-	-	+	-	-	-
1587	**109**	**4**	-	-	+	-	-	-
SP291	**4**	**4**	-	-	+	-	-	-
4	**15**	**4**	-	-	+	-	-	-
1	**8**	**8**	-	-	+	-	-	-
5	**8**	**8**	-	-	+	-	-	-
680	**8**	**8**	-	-	+	-	-	-
140	**40**	**45**	-	-	-	-	-	-
150	**16**	**16**	-	-	-	-	-	-
658	**1**	**1**	-	-	+	-	-	-
1218	**1**	**1**	-	-	+	-	-	-
520	**12**		-	-	-	-	-	-
696	**12**		-	-	+	-	-	-
978	**3**	**3**	-	-	-	-	-	-
1249	**31**	**nd**	-	-	-	-	-	-
ES15	**125**	**100**	-	-	-	-	-	-

ST; Sequence Type, CC; Clonal Complex; * CusC homologue in *C. sakazakii,* Access.; Accession, Ref.; Reference

3.5.6.5 NEONATAL MENINGITIC *E. COLI* K1 PLASMID PECOS88 IN *C. SAKAZAKII* GENOMES

Similarly, the plasmid PECOS88 (accession. CU928146), a homologue of the pAPEC-O2-R plasmid in *Escherichia coli* strain S88 which according to Peigne and colleagues (2009) was linked with high level of bacteraemia in a neonatal meningitic rat model was also aligned against all *C. sakazakii* genomes, however the whole length homologue was not detected in any of the isolate (Figure 3.22). The putative virulence genes located on the PECOS88 included three different iron uptake systems; salmochelin (*iroBCDEN*), aerobactin (*iucABCD* and *iutA*) and the *sitABCD* genes (Johnson et al. 2006; Sabri et al. 2006). Additional virulence associated traits included an *iss* gene involved in compliment resistance by providing increased serum survival (Chuba et al. 1989). Type 1 secretion system (T1SS) genes *etsABC* (Johnson et al. 2006), an outer membrane protease *ompT* (Stumpe et al. 1998) and a haemolysin encoding gene *hlyF* (Morales et al. 2004). However, the BRIG alignment (Figure 3.22) showed that the whole length homologue of the plasmid PECOS88 was not present in any of the *C. sakazakii* genome being analysed in this study. Further investigation by WebACT and *Cronobacter* indicated that except for the homologues of iron uptake aerobactin (*iucABCD* and *iutA*) which was detected in most of the *C. sakazakii* genomes; iron uptake system associated genes have been analysed in more detail (section 3.4.2.1), none of the homologue of the above mentioned virulence associated trait was detected in *C. sakazakii* genomes. Additionally a partial homologue of the *IroN*, salmochelin siderophore receptor and several homologues of transposes were noted in most of the *C. sakazakii* genomes (Figure 3.22).

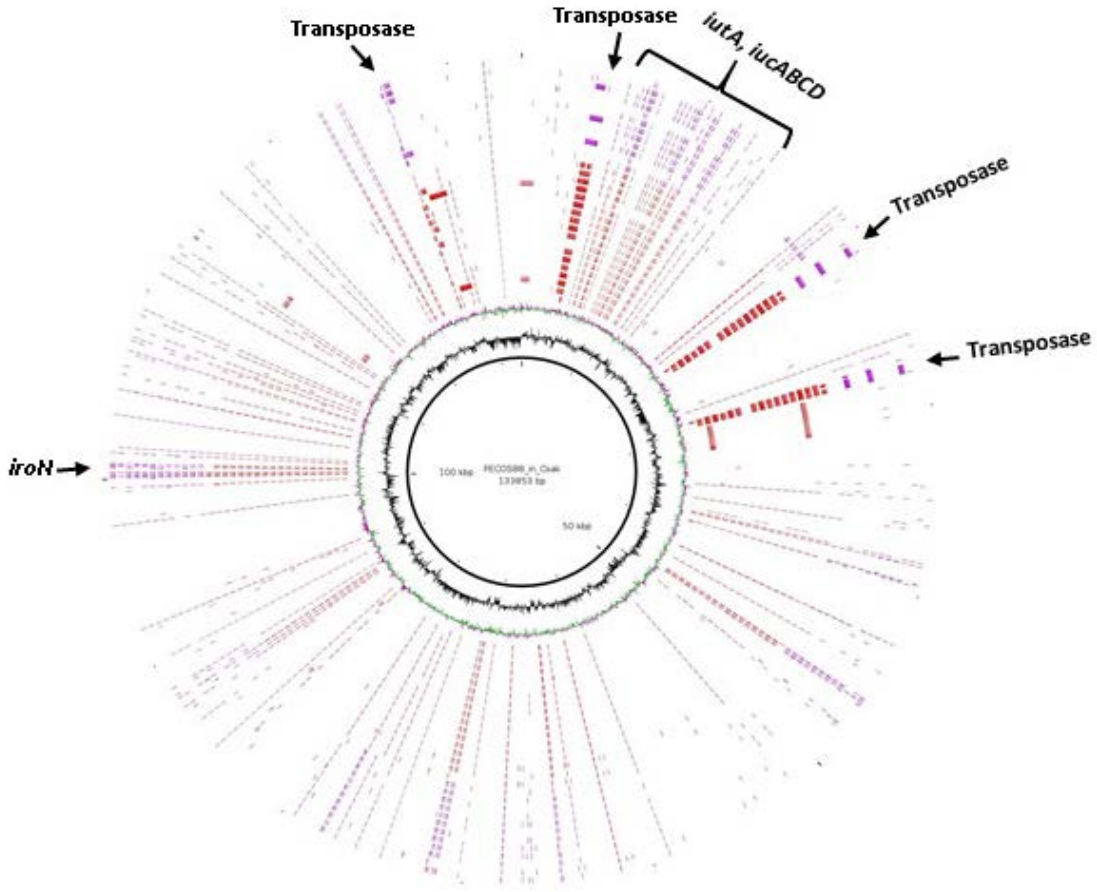


Figure 3. 20 BRIG alignment of the PECOS88 plasmid with *C. sakazakii* isolates.

The image above represents the alignment of the PECOS88 plasmid with 18 *C. sakazakii* CC4 (red rings) and 12 *C. sakazakii* non-CC4 (purple rings). The legend on the left hand side of the figure indicates ring for each isolate. The central solid black circle represent the reference plasmid backbone, the second and third rings from inside out represent the GC contents and GC skew respectively.

In brief, majority of the *E. coli* K1 virulence associated traits investigated here were not detected in most of the *C. sakazakii* CC4 and non-CC4 genomes.

3.6 GENERAL DISCUSSION AND CONCLUSION

The predominant association of *C. sakazakii* CC4 with neonatal meningitis has been established strongly. In addition to its association with neonatal meningitis, this clone has been frequently isolated from powdered infant formula (Joseph & Forsythe, 2011; Hariri et al. 2013; Sonbol et al. 2013). An attempt was made in the study presented in chapter 3 to exhaustively screen previously described virulence and environmental fitness associated traits in the sequenced *C. sakazakii* isolates.

The whole genome alignment and phylogeny of 30 *C. sakazakii* genomes (Table 3.1) revealed a close clustering of 18 *C. sakazakii* CC4 genomes (Figure 3.3). This indicated a strong clonal nature of this lineage, as 18 CC4 genomes were isolated from 9 different countries over 6 decades. This observation fabricated a strong base for the investigation at the genomic level to reveal the traits responsible for the uniqueness of this lineage.

In order to cause infection such as meningitis in neonates and infants, it is a prerequisite for a bacterial strain to cross a number of barriers before reaching the brain. After ingestion, it would need to colonise the neonate's stomach and grow in highly acidic environment (~pH 3.5). To enter blood circulation, the bacteria needs to attack the intestinal epithelial linings and therefore it must contain the necessary genetic information to encode invasion proteins. In order to reach to the brain, the organism would need to enter the systemic blood circulation where again it has to overcome the host defences such as the ability to survive within macrophages. Finally to get access to the brain it needs to cross the blood brain barrier (BBB) and once in the brain it should have the ability to acquire limited nutrients such as sialic acid in order to survive and grow (Wilson, 2002). Similarly, the frequent isolation of this important lineage and its increased isolation from powdered infant formula and its manufacturing environment warranted the investigation of the environmental fitness associated traits such as environmental stress resistance, thermotolerance, acid resistance, osmotolerance and biofilm formation.

All of the above factors were taken into consideration and an exhaustive genome screening of the virulence associated genes previously identified in *Cronobacter* and neonatal meningitic *E. coli* K1 as well as environmental fitness associated traits, was undertaken in the sequenced *C. sakazakii* genomes. The aim was to identify unique CC4 specific trait. However none of traits investigated was found to be strictly CC4 specific. Two of the traits

which showed significant variation between CC4 and non-CC4 isolates was the dominant O:2 antigen serotype of *C. sakazakii* CC4 and the predominance of thermotolerance associated genes cluster (Gajdosova et al. 2011) in *C. sakazakii* CC4 genomes, however these traits could not be termed as "unique" to CC4 as non-CC4 isolates also have O:2 serotype and no significant difference in the heat tolerance was observed between CC4 and non-CC4 in the laboratory assays performed in an independent study by Hana Sonbol.

The analysis presented in this chapter on the virulence and environmental fitness associated traits suggested nearly equal spread of these traits across both groups. With few exceptions, the analysis at this stage suggests that both *C. sakazakii* CC4 and non-CC4 may have equal virulence potential as many of the traits were noted in both groups in almost equal proportion. Further genomic analysis using different approaches was undertaken to identify strictly *C. sakazakii* CC4 specific traits and is presented in chapters 4, 5 and 6.

4 CHAPTER 4

COMPARATIVE GENOMICS TO IDENTIFY UNIQUE GENES IN *C. SAKAZAKII* CC4 CORE GENOME

4.1 BACKGROUND

As discussed earlier, the predominant association of *C. sakazakii* CC4 with neonatal meningitis and its prevalence in PIF and manufacturing environment is of great concern to neonatal health (Joseph & Forsythe, 2011; Hariri et al. 2013; Muller et al. 2013; Power et al. 2013; Sonbol et al. 2013; Forsythe et al. 2014). This warranted investigation at the genomic level which forms the basis for the current PhD thesis. The whole genome phylogeny indicated that despite the geographical and temporal spread, *C. sakazakii* CC4 constitute a strong clonal group (Figure 3.3). Although some interesting variations were noted, yet the genome screening of 30 *C. sakazakii* isolates including *C. sakazakii* 18 CC4 and 12 *C. sakazakii* non-CC4 genomes for the virulence and environmental fitness associated traits which have been mentioned in previous studies, could not identify traits which could directly be linked specifically with virulence of *C. sakazakii* CC4 (chapter 3). The clonal nature of *C. sakazakii* CC4 and their dominance with neonatal infections warranted that there might be some traits unique to this lineage. Therefore, in order to take the investigation further, different comparative genome approaches were used in order to determine the genomic features responsible for the uniqueness of *C. sakazakii* CC4. The genomic analysis presented in this chapter was aimed primarily to identify the unique traits within *C. sakazakii* CC4 core genome. This involved construction of *C. sakazakii* CC4 and *C. sakazakii* non-CC4 core genomes and comparative genomics using online comparative genomic tool called Artemis Comparison Tool (ACT) (Carver et al. 2005).

4.2 HYPOTHESIS

C. sakazakii CC4 isolates have unique virulence or environmental fitness associated traits.

4.3 AIMS OF THE CHAPTER

The main aim of the study presented in this chapter was to find out unique traits in *C. sakazakii* CC4 core genome; the analysis had following objectives.

- Construction of *C. sakazakii* CC4 core genome based on 18 *C. sakazakii* CC4 genomes.
- Construction of *C. sakazakii* non-CC4 core genome based on 12 *C. sakazakii* non-CC4 genomes.
- Compare the genome of *C. sakazakii* CC4 isolate 557 against non-CC4 isolate 658 using a gene by gene search approach to list unique traits in CC4 isolate 557.
- Using a three way ACT, align *C. sakazakii* CC4 genome 557 with *C. sakazakii* CC4 core genome and *C. sakazakii* non-CC4 core genome.
- Check the presence or absence of the genes identified unique to *C. sakazakii* CC4 isolate 557 in CC4 core genome and non-CC4 core genome.
- List unique genes in *C. sakazakii* CC4 core genome.
- Check the variation of the genes identified unique to *C. sakazakii* core genome in *C. sakazakii* pangenome.
- Screen only the traits of interest in *Cronobacter* BLAST to observe their variation in 107 *Cronobacter* genomes representing 7 *Cronobacter* species.

4.4 METHODOLOGY

4.4.1 CORE GENOME CONSTRUCTION

The core genome construction involved aligning the genomes using Mugsy (Angiuoli & Salzberg, 2011). The aligned genomes were subjected to series of algorithms written in Python to extract the core genome (section 2.2.8.1, Appendix text 1.1). For *C. sakazakii* CC4 core genome construction, the genomes of 18 *C. sakazakii* CC4 isolates were used while for non-CC4 core genome construction, the genomes of 12 *C. sakazakii* non-CC4 isolates were used (Table 3.1). The core genomes were annotated using Prokka (section 2.2.5).

4.4.2 COMPARATIVE GENOME ANALYSIS

The comparative genome analysis presented in this chapter was undertaken in 3 steps.

1. Initially, the genome of *C. sakazakii* CC4 isolate 557 was chosen to compare against a *C. sakazakii* non-CC4 isolate 658. The rational for choosing *C. sakazakii* CC4 isolate 557 amongst other CC4 isolates was that it was the earliest clinical isolate (1979) at the time this study was initiated. The *C. sakazakii* non-CC4 658 was chosen as representative *C. sakazakii* non-CC4 genome since it was the first finished *C. sakazakii* genome available publically (Kucerova et al. 2010). The comparison of isolate 557 against 658 was undertaken using ACT using a gene by gene search approach. The unique genes in *C. sakazakii* CC4 isolate 557 were listed.

2. Once the unique genes in *C. sakazakii* CC4 isolate 557 were identified. Then using a three way ACT comparison, the genome of *C. sakazakii* CC4 isolate 557 was aligned against *C. sakazakii* CC4 core genome and *C. sakazakii* non-CC4 core genome, any trait present only in *C. sakazakii* CC4 isolate 557 and *C. sakazakii* core genome while absent from *C. sakazakii* non-CC4 core genome was designated as “unique in *C. sakazakii* CC4 core genome”.

3. Once the traits were determined as unique to “*C. sakazakii* CC4 core genome”; these traits were then checked for their variation in pangenome of *C. sakazakii* isolates (n=30). For this purpose, the genome of each *C. sakazakii* isolate (Table 3.1) was checked manually using ACT for the presence or absence of these genes.

4.5 RESULTS AND DISCUSSION

4.5.1 THE *C. SAKAZAKII* CC4 CORE GENOME

The *C. sakazakii* CC4 core genome was constructed using 18 *C. sakazakii* CC4 genomes (section 2.2.8.1). The core *C. sakazakii* CC4 genome consisted of 4497 coding sequences (CDS) or genes and was ~3.22 Mb in length (data not shown), representing a very large core genome size as compared to the average genome size of a CC4 genome (~4.5Mbp). The larger core genome is indicative of very low sequence diversity within this lineage. At least 17% of the core CC4 genome consisted of hypothetical proteins of unknown function. Other genes encoded diverse range of functions which included but are not limited to starvation sensing proteins, multiple drug resistance proteins, penicillin binding proteins, fimbrial and flagellar proteins, general stress response proteins, glycine betaine/L-proline transport proteins, heat shock proteins, haemolysins and heme transport proteins, iron, zinc, nickel transport, ABC transporters, magnesium transport, methyl accepting chemotaxis proteins, transcriptional regulators and a large number of other housekeeping traits.

4.5.2 THE *C. SAKAZAKII* NON-CC4 CORE GENOME

The *C. sakazakii* non-CC4 core genome was constructed using 12 non-CC4 genomes (Table 3.1) using Mugsy whole genome alignment followed by the series of scripts to extract the core genome (section 2.2.8.1, Appendix text 1.1). The *C. sakazakii* non-CC4 core genome was much smaller in size compared to *C. sakazakii* CC4 core genome consisting of ~1.59 Mb and 1807 CDS or genes (data not shown). The smaller core genome size is indicative of greater sequence diversity within *C. sakazakii* non-CC4 genomes, it was expected since 12 non-CC4 genomes represented 8 different sequence types and at least 6 different clonal complexes in contrast to CC4 core genome which represented only 3 different sequence types and only one clonal complex (Table 3.1). The composition of the *C. sakazakii* non-CC4 core genome indicated at least 50% of hypothetical proteins of unknown function. The remaining traits included a large number of housekeeping traits in addition to other traits belonging to diverse functional categories such as transcriptional regulators, multi drug resistance proteins, methyl accepting chemotaxis proteins, iron transport, fimbrial proteins, general stress proteins, nitrate reductases and ABC transporters, majority of which were also part of the *C. sakazakii* CC4 core genome.

Since the core genome analysis was based on the comparison of two genomes, It should be noted however that there are likely to be genes which are present in both of these isolates i.e. 557 and 658, however absent from the non-CC4 core genome.

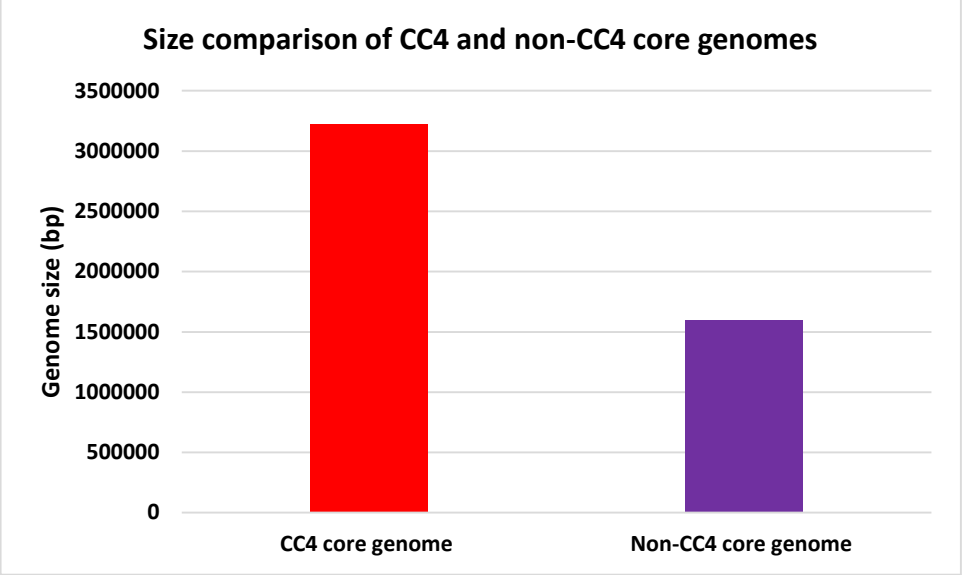


Figure 4. 1 Size comparison of *C. sakazakii* CC4 and *C. sakazakii* non-CC4 core genomes.

C. sakazakii core CC4 and *C. sakazakii* non-CC4 genomes were constructed by Mugsy and post Mugsy scripts using 18 and 12 genomes respectively. The Y axis represents the genome sizes in base pairs (bp) for CC4 core genome (3224000 bp) and non-CC4 core genome (1597600 bp).

4.5.3 COMPARATIVE GENOME ANALYSIS

The comparative genome analysis was undertaken as explained above (section 4.3.2). The initial genomic comparison of *C. sakazakii* CC4 isolate 557 against *C. sakazakii* non-CC4 isolate 658 identified a total of 442 genes unique to *C. sakazakii* CC4 isolate 557; the large number of these traits included hypothetical proteins (234/442) of unknown function and phage associated traits (22/442) (Table S2). All of these 442 genes were screened using three way ACT comparison to shortlist only those genes which are present only in *C. sakazakii* CC4 isolate 557 and *C. sakazakii* CC4 core genome while absent from *C. sakazakii* non-CC4 core genome; hence designated as unique CC4 core genes (Figure 4.2).

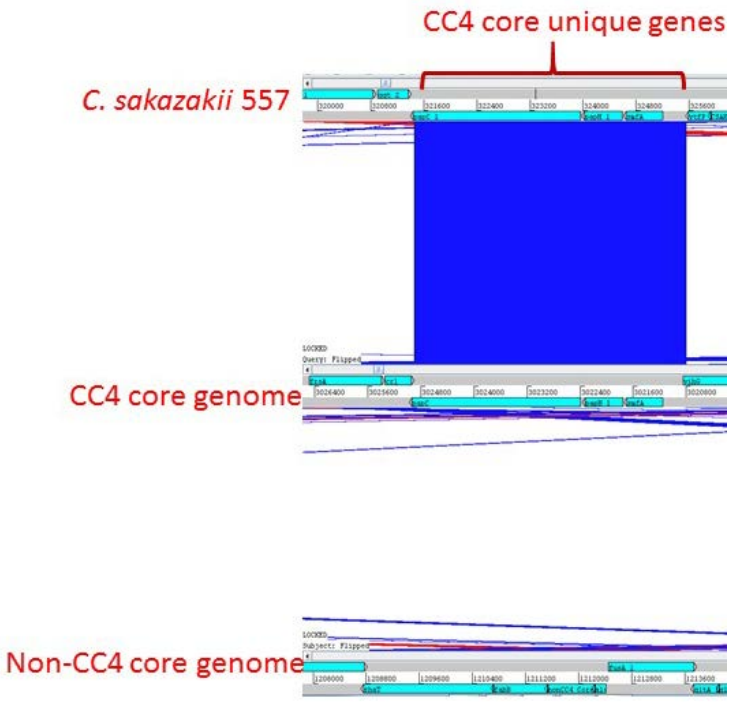


Figure 4. 2 Three way ACT comparison.

The figure is an example showing the selection of genes unique to CC4 core genome. Only those traits were considered as unique to CC4 core genome which were present only in CC4 isolate 557 (top) and core CC4 genome (middle) while absent from non-CC4 core genome (bottom).

Three way ACT genomic comparison indicated that out of 442 genes which were unique to *C. sakazakii* CC4 isolate 557 in comparison with *C. sakazakii* non-CC4 isolate 658, only 39 genes were unique to *C. sakazakii* CC4 core genome (present in *C. sakazakii* CC4 isolate 557 and CC4 core genome but absent in *C. sakazakii* non-CC4 isolate 658) (Table 4.2). These 39 genes included 12 hypothetical proteins. The BLASTx analysis for most hypothetical proteins

could not identify any functional relevance as hits with significant degree of homology also represented hypothetical proteins.

In order to validate the above results, each of the 39 genes was checked manually in 30 *C. sakazakii* genomes to check their variation in the pan *C. sakazakii* genome (n=30; Table 3.1). Apart from their conservation in *C. sakazakii* CC4, most of the genes were also noticeable in *C. sakazakii* non-CC4 genomes with variable degree of presence and none of the genes was found absent in all *C. sakazakii* non-CC4 genomes (n=12) (Figure 4.3, Table 4.1). Of particular interest were 5 genes which were noted to be absent in majority of the *C. sakazakii* non-CC4 genomes (n=12). Four of these genes with locus_tag CSAK557_01621, CSAK557_01380, CSAK557_03949 and CSAK557_03870 encoded for the hypothetical proteins of unknown function while one gene with locus_tag CSAK557_02008 encoded for a metallo-beta-lactamase family protein (Table 4.1).

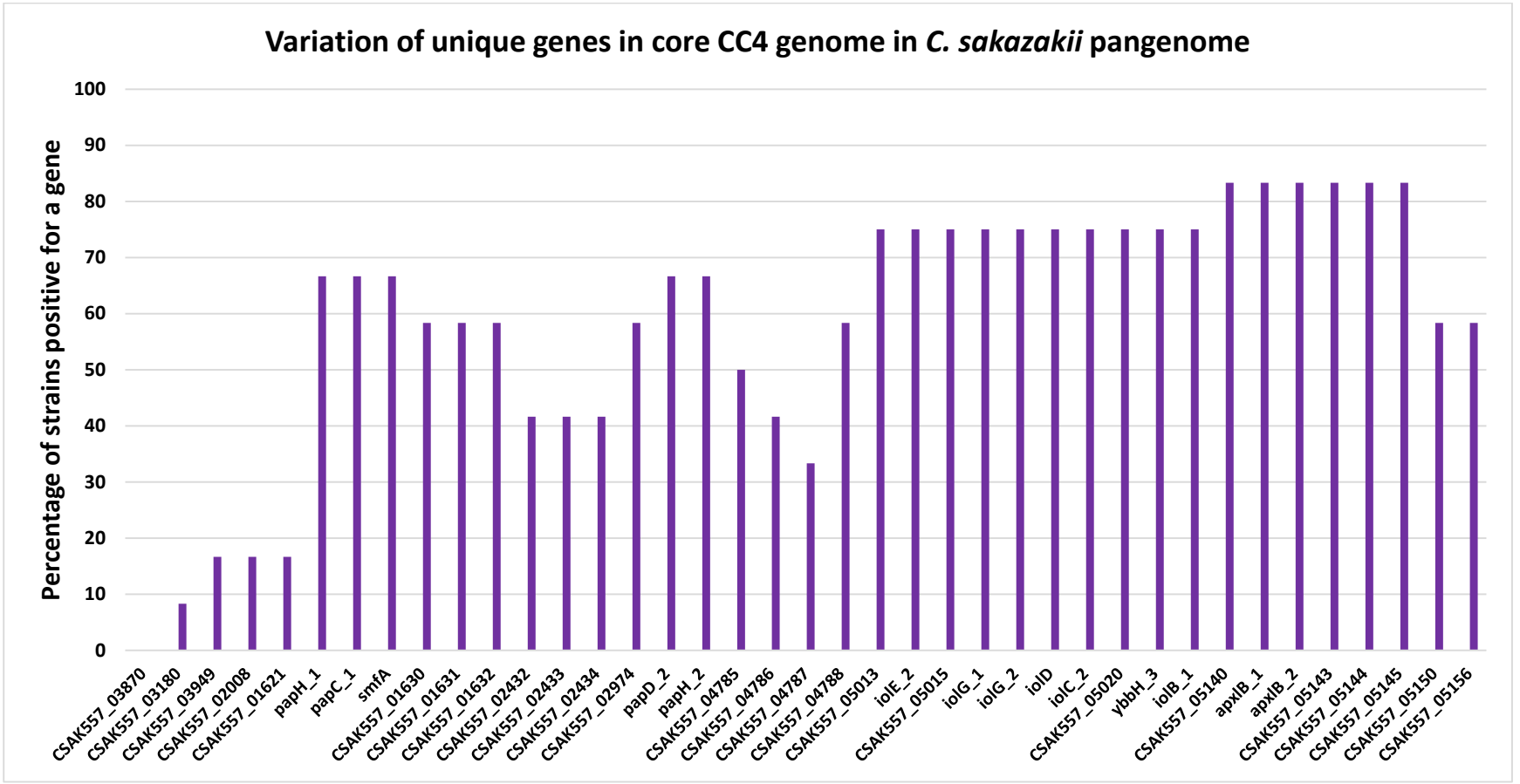


Figure 4. 3 Variation of *C. sakazakii* CC4 core unique genes in 30 *C. sakazakii* isolates.

The graph represents the variation of 39 unique genes in the core CC4 genome, across the pangenome of 30 *C. sakazakii* isolates. The X axis shows the genes while Y axis represents the percentage of non-CC4 isolates (n=12) positive for any gene.

Table 4. 1 Distribution of 39 *C. sakazakii* CC4 core unique genes in pan *C. sakazakii* genome (n=30)

Genes	Annotation	Str.																												
CSAK557 01621	H	6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_03949	H	20	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_03870*	H	377	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_03180	H	553	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_02008	Metallo-beta-lactamase superfamily protein	557	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557 01621	H	558	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
papH_1	PAP fimbrial minor pilin protein precursor	701	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
papC_1	OM usher protein papC precursor	721	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
smfA	Fimbria A protein precursor	767	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_01630	Hothetical protein	1219	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_01631	H	1220	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CSAK557_01632	H																													
CSAK557_02432	H																													
CSAK557_02433	H																													
CSAK557_02434	H																													
CSAK557_02974	H																													
papD_2	Chaperone protein papD precursor																													
papH_2	PAP fimbrial minor pilin protein precursor																													
CSAK557_04785	putative fimbrial subunit SteE																													
CSAK557_04786	putative minor fimbrial subunit StfF																													
CSAK557_04787	Major MR/P fimbria protein precursor																													
CSAK557_04788	O-antigen ligase RfaL																													
CSAK557_05013	H																													
iolE_2	Inosose dehydratase																													
CSAK557_05015	Hydroxypyruvate isomerase																													
iolG_1	Inositol 2-dehydrogenase																													
iolG_2	Inositol 2-dehydrogenase																													
iolD	trihydroxycyclohexane-1,2-dione hydrolase																													
iolC_2	5-dehydro-2-deoxygluconokinase																													
CSAK557_05020	Hothetical protein																													
ybbH_3	putative HTH-type transcriptional regulator																													
iolB_1	5-deoxy-glucuronate isomerase																													
CSAK557_05140	T1SS membrane fusion protein, HlyD																													
apxIB_1	RTXI toxin determinant B																													
apxIB_2	RTX-I toxin determinant B																													
CSAK557_05143	Hothetical protein																													
CSAK557_05144	OMchannel protein																													
CSAK557_05145	TISS OM protein, TolC family																													
CSAK557_05150	H																													
CSAK557 05156	H																													

Table 4.1 continued

	CSAK557_01621	CSAK557_02008	CSAK557_03180	CSAK557_03870*	CSAK557_03949	papH_1	papC_1	smfA	CSAK557_01630	CSAK557_01631	CSAK557_01632	CSAK557_02432	CSAK557_02433	CSAK557_02434	CSAK557_02974	papD_2	papH_2	CSAK557_04785	CSAK557_04786	CSAK557_04787	CSAK557_04788	CSAK557_05013	iolE_2	CSAK557_05015	iolG_1	iolG_2	iolD	iolC_2	CSAK557_05020	ybbH_3	iolB_1	CSAK557_05140	apxIB_1	apxIB_2	CSAK557_05143	CSAK557_05144	CSAK557_05145	CSAK557_05150	CSAK557_05156
	H	Metallo-beta-lactamase superfamily protein	H	H	H	PAP fimbrial minor pilin protein precursor	OM usher protein papC precursor	Fimbria A protein precursor	Hothetical protein	H	H	H	H	H	H	Chaperone protein papD precursor	PAP fimbrial minor pilin protein precursor	putative fimbrial subunit SteE	putative minor fimbrial subunit StfF	Major MR/P fimbria protein precursor	O-antigen ligase RfaL	H	Inosose dehydratase	Hydroxypyruvate isomerase	Inositol 2-dehydrogenase	Inositol 2-dehydrogenase	trihydroxycyclohexane-1,2-dione hydrolase	5-dehydro-2-deoxygluconokinase	Hothetical protein	putative HTH-type transcriptional regulator	5-deoxy-glucuronate isomerase	T1SS membrane fusion protein, HlyD	RTXI toxin determinant B	RTX-I toxin determinant B	Hothetical protein	OMchannel protein	TISS OM protein, TolC family	H	H
1221	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1225	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1231	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1240	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1587	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
SP291	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
658	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1218	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
978	+	-	+	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
1	-	-	-	-	-	+	+	+	+	+	+	-	-	-	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-

Table 4.1 continued

		CSAK557_01621	CSAK557_02008	CSAK557_03180	CSAK557_03870*	CSAK557_03949	papH_1	papC_1	smfA	CSAK557_01630	CSAK557_01631	CSAK557_01632	CSAK557_02432	CSAK557_02433	CSAK557_02434	CSAK557_02974	papD_2	papH_2	CSAK557_04785	CSAK557_04786	CSAK557_04787	CSAK557_04788	CSAK557_05013	iolE_2	CSAK557_05015	iolG_1	iolG_2	iolD	iolC_2	CSAK557_05020	ybbH_3	iolB_1	CSAK557_05140	apxIB_1	apxIB_2	CSAK557_05143	CSAK557_05144	CSAK557_05145	CSAK557_05150	CSAK557_05156
		H	Metallo-beta-lactamase superfamily protein	H	H	H	PAP fimbrial minor pilin protein precursor	OM usher protein papC precursor	Fimbria A protein precursor	Hothetical protein	H	H	H	H	H	H	Chaperone protein papD precursor	PAP fimbrial minor pilin protein precursor	putative fimbrial subunit SteE	putative minor fimbrial subunit StfF	Major MR/P fimbria protein precursor	O-antigen ligase RfaL	H	Inosose dehydratase	Hydroxypyruvate isomerase	Inositol 2-dehydrogenase	Inositol 2-dehydrogenase	trihydroxycyclohexane-1,2-dione hydrolase	5-dehydro-2-deoxygluconokinase	Hothetical protein	putative HTH-type transcriptional regulator	5-deoxy-glucuronate isomerase	T1SS membrane fusion protein, HlyD	RTXI toxin determinant B	RTX-I toxin determinant B	Hothetical protein	OMchannel protein	TISS OM protein, TolC family	H	H
5		-	-	-	-	-	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
680		-	-	-	-	-	+	+	+	+	+	+	-	-	-	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
520		-	+	-	-	-	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
696		-	+	-	-	-	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
140		-	-	-	-	-	+	+	+	-	-	-	+	+	+	+	+	+	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
150		+	-	-	-	+	+	+	+	-	-	-	+	+	+	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1249		-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ES15		-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

*CSAK557_03870 is an homologue of the CSSP291_18210, H; Hypothetical protein, Str.; Strain, Red font colour indicate CC4 isolates while purple font colour show non-CC4 isolates.

4.5.3.1 THE HYPOTHETICAL PROTEIN CSAK557_01621

The hypothetical protein CSAK557_01621 despite its presence in 100% of the *C. sakazakii* CC4 genomes was also noted in two non-CC4 isolates 978 and 150 (Table 4.1). No functional relevance could be designated to this hypothetical as the BLASTx hits also revealed hypothetical protein of unknown function. The analysis of the flanking region of the gene showed that it was flanked by hypothetical proteins and some housekeeping traits such as proline porter II and cytochrome b-561 (involved in vitamin c regeneration) (Figure 4.4). The presence of CSAK557_01621 next to *prop_1* encoding for the proline porter II could be of significance as proline transport in bacteria has been associated with osmoregulation in bacteria such as *E. coli* K12 (Grothe et al. 1986; Milner & Wood, 1989). It has been well established that *C. sakazakii* are able to survive under very low water activity (a_w of ~0.2) environments such as PIF (Barron & Forsythe, 2007). Therefore, one possibility could be the regulation of proline porter II by the hypothetical protein CSAK557_01621, however only future studies aimed at its characterisation could elucidate its precise role and its significance in *C. sakazakii*.

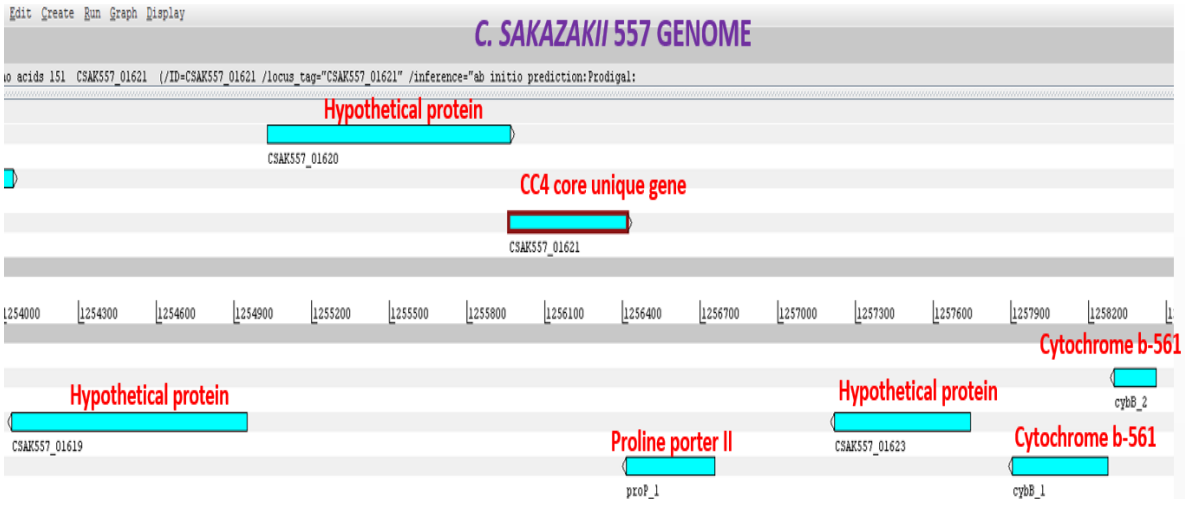


Figure 4. 4 The flanking region of the hypothetical protein CSAK557_01621.

The figure above represents the cropped section of the screen shot showing flanking region of the hypothetical protein CSAK557_01621 on the genome of *C. sakazakii* CC4 isolate 557 as viewed in Artemis using annotated EMBL file.

4.5.3.2 THE HYPOTHETICAL PROTEIN CSAK557_03180

The hypothetical protein CSAK557_03180 which was conserved in 100% of the CC4 genome was also noted in only one non-CC4 genome 978. The BLASTx analysis indicated a hit against prophage protein in *C. turicensis* 564 (accession; CCJ90293), hence indicating that it is a phage acquired trait which has been carried by CC4 isolates.

4.5.3.3 THE HYPOTHETICAL PROTEIN CSAK557_03949

Besides 100% of CC4 genomes, the hypothetical protein CSAK557_03949 was also noted in two of the non-CC4 isolates 150 and 1249. The BLASTx analysis could not reveal any functional relevance for this gene. Moreover, the gene was also flanked by hypothetical proteins both upstream and downstream when the flanking region was examined on the genome of *C. sakazakii* isolate 557. Further characterisation may clarify the role of this hypothetical protein and its predominance in CC4.

4.5.3.4 METALLO-BETA-LACTAMASE SUPERFAMILY PROTEIN (CSAK557_02008)

Another gene unique in the core CC4 genome was CSAK557_02008 encoding for metallo-beta lactamase superfamily protein. This was an important observation as β lactamase family proteins are important towards antibiotic resistance in bacteria. The gene was also noted in two of the non-CC4 isolates 520 and 680 both of which belong to ST12. Moreover, both of these ST12 strains have clinical origin (Table 3.1), therefore the initial observation was suggesting beta lactam resistance could be restricted to CC4 and ST12 lineages.

4.5.3.5 THE HYPOTHETICAL PROTEIN CSAK557_03870 or CSSP291_18210

One of the most important observation in the present analysis was the hypothetical protein CSAK557_03870 which was conserved in 100% of the *C. sakazakii* CC4 genomes (n=18) analysed in the present study while absent entirely from the *C. sakazakii* non-CC4 genomes (n=12). This was an important observation, since the aim of the present analysis was to detect such a trait which was uniformly present in CC4 while uniformly absent from non-CC4 genomes.

It must be noted that the hypothetical protein CSAK557_03870 in CC4 isolate 557 is an homologue of the hypothetical protein CSSP291_18210 in *C. sakazakii* CC4 isolate SP291, therefore from this point onwards, in order to avoid any confusion which may arise later in

this thesis (chapter 6), the hypothetical protein CSAK557_03870 will be called as "hypothetical protein CSSP291_18210".

In order to predict any functional relevance, the BLASTx analysis of CSSP291_18210 having nucleotide sequence length of 270bp and translated amino acid sequence length of 89 amino acids was undertaken. One of the top BLASTx hit with significant degree of homology (identities = 73% (65/89) and 85% positives (76/89) was against a putative inner membrane protein of *Enterobacter aerogenes* (accession: YP_007387546).

Another hit with far less homology (identities = 46% (57/87) and 62% positives (54/87) was against the "chain A, crystal structure of the type VI effector-immunity complex Ssp1-rap1a from *Serratia marcescens*" (accession; 4BI8_A). The flanking region of the CSSP291_18210 was also analysed on the genome of *C. sakazakii* CC4 isolate 557 (Figure 4.5), however it was flanked both upstream and downstream by hypothetical proteins of unknown function.

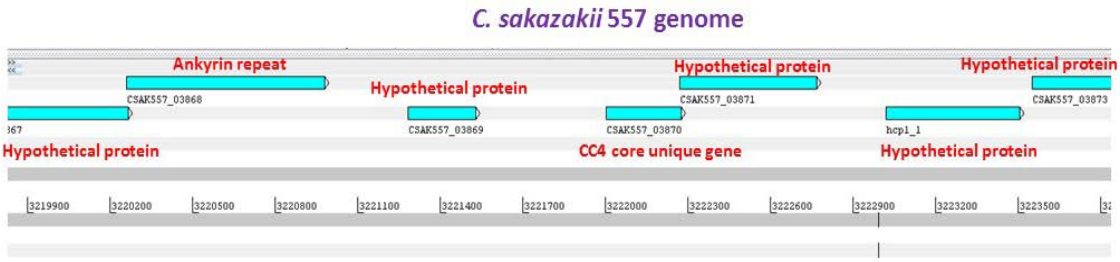


Figure 4. 5 The flanking region of the hypothetical protein CSSP291_18210
The figure above represents the cropped section of the screen shot showing flanking region of the hypothetical protein CSSP291_18210 on the genome of *C. sakazakii* CC4 isolate 557 as viewed in Artemis using annotated EMBL file.

Furthermore, an attempt was also made to obtain any additional functional and structural relevance for this hypothetical protein using an online protein prediction tool available at http://ppopen.informatik.tu-muenchen.de/ (Bigelow et al. 2004; Ofran & Rost, 2007; Hamp et al. 2013). Although, the translated amino acid sequence of the hypothetical protein CSSP291_18210 (89 amino acid) could not assign any functional category, its predicted secondary structure indicated at least 10 protein binding sites and one polynucleotide binding region (Ofran & Rost, 2007) (Figure 4.6). These binding sites could be important for its interaction with other proteins thereby modulating either virulence or other important physiological function. Although from the present analysis, the homology of CSSP291_18210 did not indicate any direct association with virulence, yet the BLASTx homology against the inner membrane protein and the presence of the protein binding

sites on its predicted secondary structure could be of significance. However, these are just predictions and only the full characterisation of the hypothetical protein CSSP291_18210 could describe its precise role and its importance in *C. sakazakii* CC4.

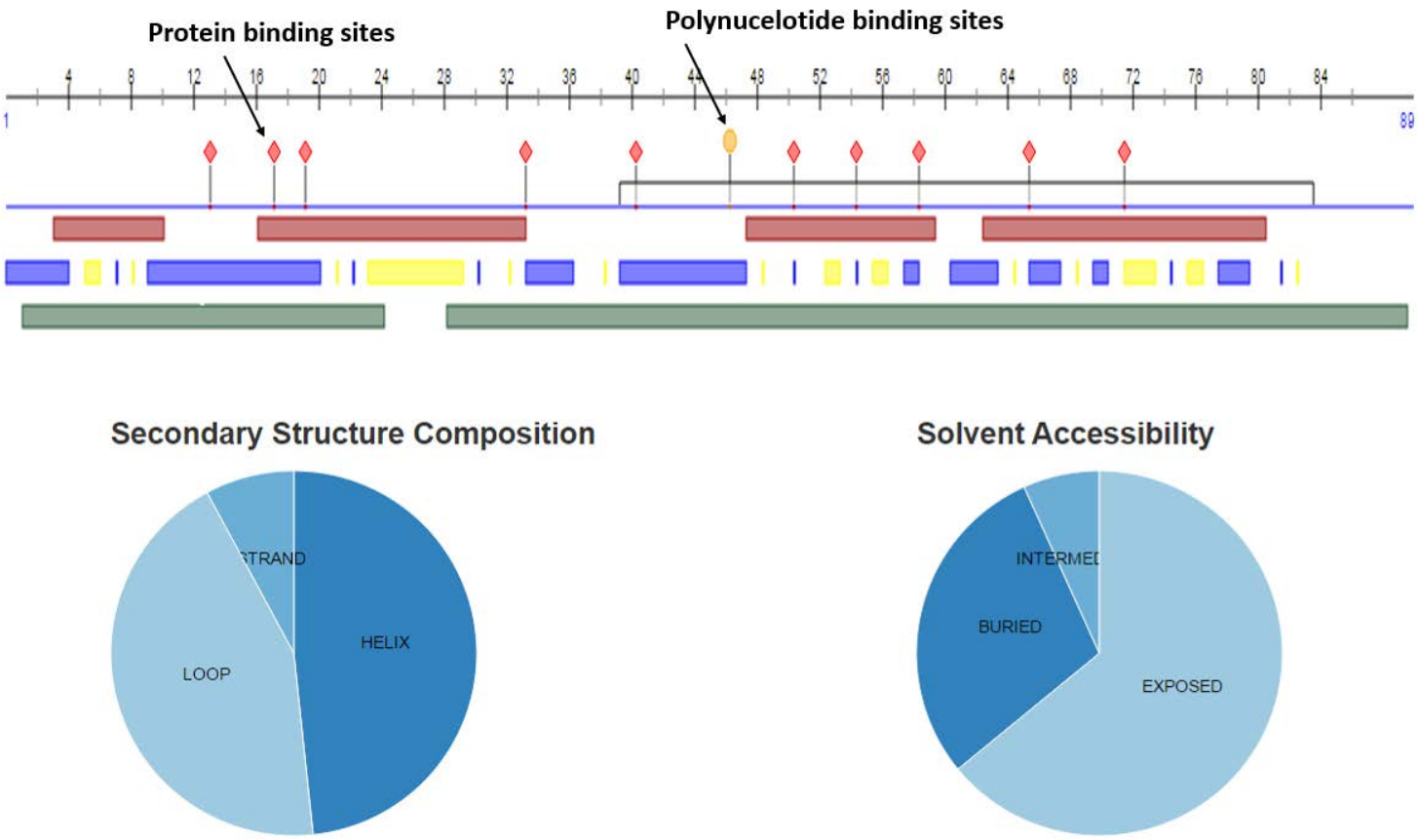


Figure 4. 6 The *In silico* protein prediction of hypothetical protein CSSP291_18210.

The figure above represents the secondary structure prediction of CSSP291_18210 (89 amino acid long) using online protein prediction tool http://ppopen.informatik.tu-muenchen.de/. The red squares with stalks represent the protein binding sites while the yellow circle with a stalk shows a polynucleotide binding region. The red, blue and yellow rectangles represent the helices, exposed and buried parts of the protein. The composition of the secondary structure and its solvent accessibility is represented by the pie-chart.

4.5.4 SCREENING OF THE SELECTED GENES IN 107 *CRONOBACTER* GENOMES USING *CRONOBACTER* BLAST

During the later stages of the current PhD study (by the end of 3rd year), 107 *Cronobacter* genomes which also included 30 *C. sakazakii* genomes analysed in the present study (Table 3.1) were made available at *Cronobacter* BIGsDB which has a BLAST facility. These 107 genomes included 37 CC4 genomes (all of which were *C. sakazakii*) and 70 non-CC4 genomes belonging to 7 different species of *Cronobacter*. (http://pubmlst.org/perl/bigsdb/bigsdb.pl?db=pubmlst_*cronobacter*_isolates&page=plugin&name=BLAST) (last accessed; 20th Feb 2015). Hence it provided an opportunity to screen any trait of interest against 107 *Cronobacter* genomes.

This opportunity was utilised in the present study and 4 hypothetical proteins CSAK557_01621, CSAK557_03949, CSAK557_02008 and CSSP291_18210 which showed predominant presence in *C. sakazakii* CC4 (Table 4.1) were screened in 107 *Cronobacter* genomes which represent the entire *Cronobacter* genus. Please note that the 37 CC4 genomes in *Cronobacter* BLAST were all *C. sakazakii* while the remaining non-CC4 genomes represented 7 different species of *Cronobacter*. The hypothetical protein CSAK557_03180 was excluded since it was found to be the homologue of a phage associated trait.

The *Cronobacter* BLAST analysis of the hypothetical protein CSAK557_01621 indicated that the gene was present in 100% of the CC4 isolates (n=37) while only 12 of the 70 *Cronobacter* spp. non-CC4 isolates (including isolates belonging to *C. sakazakii, C. malonaticus, C. turicensis* and *C. dublinensis*), showing significant variation ($P < 0.05$) (Table S3).

The *Cronobacter* BLAST analysis of the hypothetical protein CSAK557_03949 also indicated that the gene was present in 100% of the CC4 isolates (n=37) while it was also noted in 13 of 70 *Cronobacter* spp. non-CC4 genomes including *C. sakazakii*, *C. dublinensis*, *C. turicensis*, *C. malonaticus* and *C. condimenti*. The hypothetical protein CSAK557_03949 varied significantly between CC4 and non-CC4 isolates ($P < 0.05$). This observation indicates that although the CSAK557_03949 might be an essential part of the *C. sakazakii* core CC4 genome yet it is also noticeable in the pangenome of *Cronobacter* spp. non-CC4 isolates.

The *Cronobacter* BLAST analysis of the hypothetical protein CSAK557_02008 also indicated significant variation ($P < 0.05$) between CC4 and non-CC4 isolates. The gene encoding for

the metallo-β-lactamase indicated that despite its presence in 100% of the CC4 genomes (n=37) the gene was also noted in majority i.e. 42 of 70 *Cronobacter* spp. non-CC4 genomes representing all 7 species of *Cronobacter*. Hence, although an integral part of the core CC4 genome, yet it is not exclusive to CC4 and its distribution across non-CC4 genomes does not signify its uniqueness in *C. sakazakii* CC4 core genome.

For the hypothetical protein CSSP291_18210, interestingly the BLAST results indicated that 100% of the CC4 genomes (n=37) present in the *Cronobacter* PubMLST were positive. On the other hand, it was detected only in 2 of 70 non-CC4 *Cronobacter* spp. genomes present in *Cronobacter* PubMLST, indicating significant variation ($P < 0.05$) between CC4 and non-CC4 isolates. Two of the non-CC4 genomes which were positive for the hypothetical protein CSSP291_18210 included *C. sakazakii* 2051 (CC64) and *C. malonaticus* 685 (CC129).

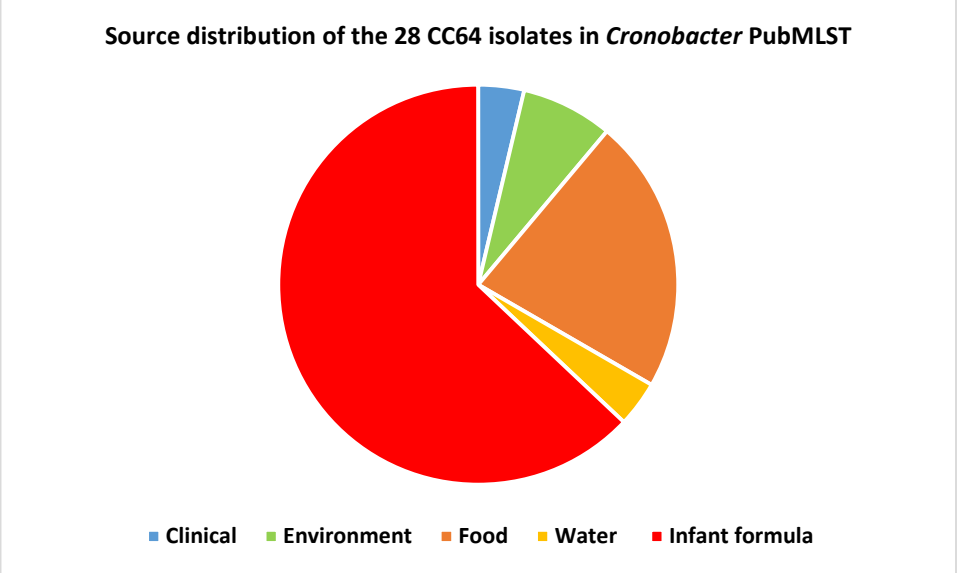


Figure 4. 7 Distribution of Clonal Complex 64 isolates in *Cronobacter* PubMLST.

The analysis of the *Cronobacter* spp. isolates deposited in *Cronobacter* PubMLST indicates that a total of 28 isolates belong to CC64, all of which were *C. sakazakii*. More importantly 82 % (23/28) of the CC64 isolates were obtained from food including infant formula ingredients while 37 % (16/28) represented isolates obtained from infant formula (the information was true on 3rd Feb, 2015). Therefore, the increased association of CC64 with food, especially infant formula is alarming as isolates from this clonal complex may also pose a potential risk to neonatal health and immuno-compromised individuals due to their association with infant formula and food such as ready to eat food, respectively. Whether

the hypothetical protein CSSP291_18210 is also predominant in CC64 is not clear at the moment, since *C. sakazakii* 2051 represented the only sequenced CC64 isolate in the *Cronobacter* PubMLST. Hence, it is warranted that future studies also include isolates from CC64.

On the other hand, *C. malonaticus* is another *Cronobacter* spp. which has been associated with adult infections (Joseph & Forsythe, 2012; Joseph et al. 2012a, b; Hariri et al. 2013). However, the hypothetical protein CSSP291_18210 was only found in 1 sequenced isolate of *C. malonaticus* 685 (CC129) amongst a total of 14 sequenced *C. malonaticus* isolates present in the *Cronobacter* PubMLST.

From the present analysis, it is suggested that the hypothetical protein CSSP291_18210 is predominant in CC4 and 2 of the non-CC4 isolates positive for this hypothetical protein may represent the rare variants which could have acquired this protein through horizontal gene transfer. Full characterisation of this hypothetical protein is warranted in future studies which may reveal important insights towards its significance in CC4.

4.6 GENERAL DISCUSSION AND CONLCUSION

The analysis presented in this chapter was aimed primarily at the identification of unique traits in *C. sakazakii* CC4 with the hope to find a possible link with its virulence or environmental persistence. In order to achieve this, a comparative genomics approach based on the core genomes was used in the analysis presented in this chapter. The pangenome of a bacterial lineage is the combination of a core genome which is conventionally defined as those genes present in all strains, and an accessory genome which includes the genes absent from one or more strains (Tettelin et al. 2005; van Tonder et al. 2014).

The aim was to first shortlist only those traits which are unique to CC4 core genome and then validate their presence in the pangenome of *C. sakazakii* isolates. The core genomes of both CC4 and non-CC4 genomes were constructed by using Mugsy and post-Mugsy scripts (section 2.2.8.2, Appendix text 1.1). As compared to non-CC4 core genome, the core CC4 genome was much larger and was almost double its size (Figure 4.1). It must however be noted that defining a core genome is arbitrary and depends on the number and nature of the isolates used to construct a core genome; conventionally larger the number of isolates, smaller would be the core genome size (van Tonder et al. 2014). However, since the number of isolates used for the construction of CC4 core genome (n=18) were higher than non-CC4 (n=12), the large size of *C. sakazakii* CC4 genomes indicated a very low sequence diversity within this lineage. The observation was interesting as the *C. sakazakii* isolates used in the present analysis were geographically and temporally diverse; the 18 CC4 isolates were obtained during the last 6 decades from at least 9 different countries of the world (Table 3.1). Yet their close homology to each other is of great significance. The larger core genome size further validates the clonality of CC4 isolates observed by whole genome phylogeny of 30 *C. sakazakii* isolates (Figure 3.3). On the other hand, the smaller core genome size for *C. sakazakii* non-CC4 could be explained by the variation in the isolates used to construct it; the 12 non-CC4 isolates represented at least 8 different STs and 6 different clonal complexes (CCs) contrary to CC4 core genome which represented only one CC and three STs.

The results of the current analysis were based on the comparative genomics between one CC4 isolate 557 and one non-CC4 isolate 658 which was undertaken using ACT. A total of 442 genes were identified unique to CC4 isolate 557 (Table S2). To validate how many of these traits are unique to CC4 core genome; a three way ACT comparison was undertaken using CC4 isolate 557, *C. sakazakii* core CC4 genome and *C. sakazakii* non-CC4 core genome. Any gene detected only in *C. sakazakii* CC4 isolate 557 and *C. sakazakii* core-CC4 genome while absent from *C. sakazakii* non-CC4 core genome was termed as unique to "CC4 core genome" (Figure 4.2). This way only 39 genes were detected unique to *C. sakazakii* CC4-core genome.

The 39 unique genes in *C. sakazakii* CC4-core genome included hypothetical proteins and housekeeping traits such as hydrolases, dehydrogenases; a number of traits of interest were also detected as unique to *C. sakazakii* CC4-core genome. These included 8 fimbrial associated genes found at different locations and a Type 1 Secretion System (T1SS) associated genes. Bacterial fimbriae are the appendages which act as adhesins and facilitate the bacterial attachment with the host surface and could play an important role towards the virulence of a bacterial pathogen (Connell et al. 1996). For example in uropathogenic *E. coli*, fimbrial mediated adhesion is an important virulence mechanism (Delepelaire, 2004). Similarly, bacterial secretion systems are important in establishing a successful host-pathogen interaction, a number of protein secretion systems have been described in Gram negative bacteria (Type 1-VI) (Tseng et al. 2009). The T1SS associated genes found unique in the core-CC4 genome encoded for important TISS associated components such as HlyD family protein, RTX-I toxin determinant B and a TolC protein (Table 4.1) all of which are important parts of the TISS (Tseng et al. 2009).

Although, conserved in *C. sakazakii* CC4 core genome and absent from the non-CC4 core genome, it was likely that these 39 genes might also be present in the pangenome of *C. sakazakii* non-CC4 isolates, therefore it was important to screen the pangenome of non-CC4 isolates before any gene could be designated as CC4 specific. Therefore, in order to validate whether these 39 are strictly specific to CC4 genomes or also found in the pangenome of non-CC4 isolates, all of these genes were checked for their presence in 18 *C. sakazakii* CC4 genomes and 12 *C. sakazakii* non-CC4 genomes. The results of the analysis partly validated the current analysis as all 39 genes were detected in 100% of the CC4 genomes (n=18). However, 38 of 39 genes were also detected in *C. sakazakii* non-CC4 genomes (pangenome)

with variable degree of presence indicating that while these traits are not part of the *C. sakazakii* core non-CC4 genome, yet their pangenome harbour these traits. The results were quantified to observe the degree of variability between *C. sakazakii* CC4 and *C. sakazakii* non-CC4 isolates for all 39 genes (Figure 4.3). Amongst these 39 genes five of the genes were of particular interest as they were noted with low frequency in non-CC4 isolates and analysed further. The CSAK557_01380 hypothetical protein is believed to be a phage acquired trait as indicated by the BLASTx hit and seem to be carried stably by CC4 isolates.

The *Cronobacter* PubMLST contains 107 *Cronobacter* genomes which include 37 *C. sakazakii* CC4 genomes and 70 non-CC4 genomes which belong to 7 different species of the genus *Cronobacter*. The *Cronobacter* BLAST facility was used to check the variability of the 4 hypothetical proteins CSAK557_01621, CSAK557_03949, CSAK557_02008 and CSSP291_18210 (CSAK557_03870) which were noticed predominantly in *C. sakazakii* CC4 genomes (Table 4.1). Initial analysis with 30 *C. sakazakii* genomes indicated that the hypothetical protein CSAK557_01621 was found to be conserved across all *C. sakazakii* CC4 and present only in 2 *C. sakazakii* non-CC4 isolates 140 and 978 (Table 4.1). However the gene was noted with very low frequency in non-CC4 isolates even when the analysis was expanded to 107 *Cronobacter* genomes; it was only noted in 18% of *Cronobacter* spp. non-CC4 isolates (12/70) (section 4.4.4). The BLASTx analysis could not indicate a functional relevance for this protein. The presence of the proline porter II next to this gene provided some hint that this gene might be involved in the regulation of osmosis. Previous studies have described the role of a number of genes in the osmotolerance of *C. sakazakii* (Feeney and Sleator, 2011; Feeney et al. 2014), however these genes did not form the basis of differentiation between *C. sakazakii* CC4 and *C. sakazakii* non-CC4 (section 3.4.3.6). Therefore, only further characterisation of CSAK557_01621 gene can justify its predominance in *C. sakazakii* CC4.

Similarly, when the analysis for the hypothetical protein CSAK557_03949 was expanded to 107 genomes, it was noted predominantly in CC4 isolates (37/37) while only 7 % (13/70) of the *Cronobacter* spp. non-CC4 isolates including *C. turicensis* and *C. dublinensis* which have not previously been associated with infections. The BLASTx could not identify any functional relevance while the flanking regions also consisted of hypothetical proteins. The characterisation of CSAK557_03949 may also provide insight into its predominance in CC4.

The CSAK557_02008 encoding for metallo-beta lactamase, besides its uniform presence in all CC4 isolates was also noted in two of the *C. sakazakii* non-CC4 isolates (520 and 680) of clinical origin, both of which belong to ST12. The antibiotic resistance can enhance the virulence potential of bacterial pathogens as they become difficult to treat with conventional antibiotics. Antibiotic resistance has previously been reported in *C. sakazakii* against penicillin, tetracycline, ciprofloxacin and nalidixic acid (Caubilla-Barron et al. 2007; Kilonzo-Nthenge et al. 2012). Although resistance against ESβL (extended spectrum beta lactamases) has been reported in two of the CC4 isolates 701 and 767 (Caubilla-Barron et al. 2007), however the laboratory ESβL testing of 15 CC4 isolates indicated none of them were ESβL resistant (chapter 8). Furthermore, when the *Cronobacter* BLAST analysis was expanded to 107 genomes, the results indicated that the CSAK557_02008 encoding for metallo-beta lactamase was also present in 60% of the *Cronobacter* spp. non-CC isolates belonging to 7 different species of *Cronobacter*. Hence, the trait could not be regarded as specific to CC4 and or ST12.

The most important observation in the initial analysis was the hypothetical protein CSSP291_18210 which was conserved across all *C. sakazakii* CC4 (n=18) while missing entirely form *C. sakazakii* non-CC4 isolates (n=12). Even when the analysis was expanded to 107 *Cronobacter* spp. genomes, the gene was noted only in 2 of 70 *Cronobacter* spp. non-CC4 isolates, *C. sakazakii* 2051 and *C. malonaticus* 685; these two isolates may represent rare variants within non-CC4 isolates. The *C. sakazakii* 2051 belongs to CC64 which represents an important clonal complex in *Cronobacter* PubMLST, as at least 37% of the strains were isolated from the infant formula (Figure 4.7). However, since there was one sequenced isolate *C. sakazakii* 2051 present in the *Cronobacter* PubMLST, the association of this hypothetical protein with CC64 cannot be determined. The important observation was its presence uniformly in all 37 CC4 isolates. The BLASTx analysis indicated hit against an inner membrane protein in *Enterobacter aerogenes* with high homology and to less homology with a T6SS effector protein (section 4..4.3.5). The flanking region of this hypothetical protein was investigated on the genome of *C. sakazakii* 557 however it was also flanked by uncharacterised hypothetical proteins (Figure 4.5). Using an online protein prediction tool (http://ppopen.informatik.tu-muenchen.de/), an attempt was made to obtain any further functional relevance for this protein using its amino acid sequence. Although no functional category was assigned, it was important to note the presence of

several protein binding sites from the deduced structure of this protein (Figure 4.6). Therefore, these observations suggested that this inner membrane like protein "unique" to CC4 protein could play some significant role towards virulence or fitness of CC4. Hence, further characterisation of this gene is warranted in order to precisely elucidate its role towards virulence or fitness of *C. sakazakii* CC4. The gene is now being analysed in a laboratory based independent study by Abdlrhman Alsonosi as part of his PhD study.

An obvious observation in the present analysis was that all of the 39 genes which were initially found to be unique to the *C. sakazakii* CC4 core genome (Table 4.1) were uniformly absent from the two non-CC4 isolates 1218 and 658 both of which belong to CC1 (ST1); more importantly both of these isolates have clinical origin and were isolated from the fatal meningitic cases (Table 3.1). Nevertheless their conservation in CC4 genomes is important and warrants further investigation.

The core genome analysis presented in this chapter was aimed to identify unique genes which could explain the clonality of *C. sakazakii* CC4 and their predominance in neonatal meningitis or environmental fitness. Relatively large core genome of CC4 indicated less sequence diversity within this lineage. Although none of the trait identified could directly be linked with the virulence of *C. sakazakii* CC4, yet the presence of the hypothetical protein CSSP291_18210 predominantly in CC4 isolates was significant and warrants future studies.

The analysis presented in this chapter indicated that 34 of 39 genes which were unique to core-CC4 genome were present largely in the pangenome of *C. sakazakii* non-CC4 isolates. Therefore, from this point onwards, it was decided to screen the pangenome of *C. sakazakii* CC4 and *C. sakazakii* non-CC4 by comparative genomics to identify unique genes in *C. sakazakii* CC4 genomes, the analysis is presented in chapters 5 and 6.

5 CHAPTER 5

IDENTIFICATION OF *C. SAKAZAKII* CC4 SPECIFIC TRAITS USING GENE BY GENE SEARCH APPROACH

5.1 BACKGROUND

To date, over 300 sequence types (ST) in the *Cronobacter* genus have been defined. Genome analysis has revealed clonality within the genus with over 40 clonal complexes (CC) identified so far (Forsythe et al. 2014). Amongst these clonal complexes, the CC4 isolates dominates the *Cronobacter* isolates in PubMLST (Figure 3.1). The ST4 constitutes the dominant sequence type in CC4 representing about 91% of all CC4 isolates in *Cronobacter* PubMLST (as of 2nd December 2014). A retrospective review of *Cronobacter* isolates from cerebral spinal fluid collected over a 60 year period across 6 countries revealed that majority were *C. sakazakii* ST4 (Joseph & Forsythe, 2011). This was substantiated by more recent analysis of a number of highly publicised cases in the US where *C. sakazakii* meningitis associated cases were all ST4 or its single/double loci variants constituting CC4 (CDC 2012; Hariri et al. 2013). It is notable that *C. sakazakii* CC4 (ST4) has been reported to be frequently isolated from milk powder factories, powdered infant formula (PIF) processing plants and from PIF in Ireland, Switzerland, Germany and Australia (Muller et al. 2013; Power et al. 2013; Sonbol et al. 2013; Forsythe et al. 2014). At least 24% of *Cronobacter* strains isolated from the environment of 6 milk powder manufacturing plants in Australia and Germany were reported to be CC4 (Sonbol et al. 2013). The finished genome sequence of a *C. sakazakii* CC4 strain (SP291) was notable for its persistence in a PIF production facility in Ireland (Power et al. 2013). Therefore an improved understanding of *C. sakazakii* CC4 strains at genomic level was warranted. It was of high significance to identify virulence associated determinants which are specific to this lineage. The main aim of the present PhD study was to analyse the genomes of *C. sakazakii* with the aim to identify any virulence associated traits which are specific only to *C. sakazakii* CC4. Different comparative genomic approaches were used in this study to identify any CC4 specific traits as described in chapters 3 and 4. In the present comparative genomic analysis presented in this chapter a gene by gene search approach was used to list all the *C. sakazakii* CC4 specific traits which could be linked with virulence or its persistence in the environment especially in PIF or its production environment.

5.2 HYPOTHESIS

C. sakazakii CC4 isolates have unique virulence or environmental fitness associated traits.

5.3 AIMS OF THE CHAPTER

The genomic analysis presented in this chapter is the continuation of comparative genomic analysis from chapters 3 and 4. The main aim of the genomic analysis was to find out any *C. sakazakii* CC4 specific trait which could explain its predominant association with neonatal meningitis or its persistence in PIF. The objectives of the chapter 5 were;

- Undertake a comparative genome analysis of *C. sakazakii* CC4 isolate 377 against a *C. sakazakii* non-CC4 isolate *C. sakazakii* ES15 through a gene by gene search approach using Artemis Comparison Tool (ACT).
- List all the unique traits in *C. sakazakii* CC4 isolate 377.
- Check for the presence and absence of the genes unique to *C. sakazakii* CC4 isolate 377 in the genomes of remaining *C. sakazakii* CC4 and *C. sakazakii* non-CC4 isolates to determine *C. sakazakii* CC4 specific traits.

5.4 METHODS

5.4.1 STRAINS, SEQUENCING, GENOME ASSEMBLY AND ANNOTATION

A total of 30 *C. sakazakii* strains were used in this study (Table 3.1); 18 *C. sakazakii* CC4 and 12 *C. sakazakii* non-CC4. These strains were primarily clinical isolates which had been isolated between 1950 and 2010 from 11 countries (Table 3.1). Genome sequencing of 22 *C. sakazakii* strains was carried out by the University of Exeter (UK) using Illumina Hiseq2500 generating paired end reads. The genome assembly was with Velvet (Zerbino et al. 2008). The genomes were annotated using the prokaryotic genome annotation system (PROKKA) (Seemann, 2014) (section 2.2.5).

5.4.2 COMPARATIVE GENOME ANALYSIS USING ACT

Initially one *C. sakazakii* CC4 strain 377, was compared against a non-CC4 genome *C. sakazakii* strain ES15 using ACT. The genome of *C. sakazakii* 377 represents the earliest *C. sakazakii* CC4 (ST4) isolate in the *Cronobacter* PubMLST and in our culture collection and its genome has recently been announced by our group (Masood et al. 2013a). The *C. sakazakii* ES15 representing non-CC4 genomes was isolated from ground whole grains and belongs to CC100 (ST125) (Shin et al. 2012). The comparison was conducted manually using ACT by walking along the entire length of the genome by checking each and every single gene; this involved checking presence of each gene in *C. sakazakii* 377 against *C. sakazakii* ES15 using ACT. It should be noted that coverage filter was “by eye“ (section 2.2.7). All of the genes identified as unique to *C. sakazakii* CC4 strain 377 against *C. sakazakii* ES15 were listed. The comparison was then extended to the remaining *C. sakazakii* CC4 and non-CC4 genomes. From initially identified unique genes in *C. sakazakii* 377, each of these genes was checked manually in the genomes of the remaining *C. sakazakii* CC4 and *C. sakazakii* non-CC4 using ACT and the presence/absence of each gene was recorded for each genome. The final spreadsheet was then checked to determine if any pattern exists for the presence of certain genes in *C. sakazakii* CC4 or *C. sakazakii* non-CC4 genomes. For the ease of explaining the results the unique genes in *C. sakazakii* CC4 strain 377 were divided into the genomic regions (GRs) based on either their location or functionality. To represent the presence of the genes within a GR, a heat map of each GR was generated using SPSS representing the percentage of presence of each gene in a group by the intensity of colour i.e. higher the

intensity of the colour, higher the percentage of presence of a gene in a group (CC4 or non-CC4) and *vice versa*.

5.5 RESULTS AND DISCUSSION

5.5.1 WHOLE GENOME COMPARISON OF *C. SAKAZAKII* CC4 AND NON-CC4 GENOMES USING ACT

The genome of the oldest CC4 strain in our culture collection, *C. sakazakii* 377 isolated in 1950 was chosen to compare against *C. sakazakii* ES15, the latter being a non-clinical isolate from ground whole grain. A total of 465 genes were identified unique to *C. sakazakii* CC4 strain 377 (Table S4). Each of these unique genes was checked for its presence in remaining 17 CC4 and 11 non-CC4 genomes included in this study (Table 3.1). The plasmids of *C. sakazakii* BAA-894 (also known as 658) and *C. sakazakii* SP291 were also checked for the presence of these genes. A large proportion (295/465) representing at least 63% of these genes were hypothetical proteins of unknown function and traits associated with phages. To simplify the genomic comparison at this stage, the hypothetical proteins and cryptic phages were omitted and the remaining 168 genes which included a wide range of traits were divided into seven genomic regions (GR1-7) based on their function and location on the genome (Table S4). A heat map for each GR (Figures 5.1-7) shows the presence of a gene in both CC4 and non-CC4 group. Each of the GR is discussed briefly below.

5.5.1.1 THE GR1

The GR1 consisted of 27 genes located between Csak377_00005-00053. Of significance were the two copper and silver homeostasis (*cus/sil* and *pco* regions). Both of these regions have already been investigated in chapter 3 (section 3.4.2.9). None of these genomic regions within GR1 were detected as *C. sakazakii* CC4 specific. Both of these regions were present variably across the 2 groups (Figure 5.1, Table 5.1). The variation has been discussed in more detail in section 3.4.2.9.

In addition, the GR1 also contained two toxin/anti-toxin system associated proteins vapC, vagC (annotated as a virulence associated protein). Another toxin/antitoxin system proteins parD1/parE4 were found adjacent to the first copper/silver homeostasis region. Both the toxin/anti-toxin system associated genes were present variably across the genomes of CC4 and non-CC4 strains (Figure 5.1, Table 5.1).

Other traits in GR1 found variably across two sets of genomes included a disulfide isomerase, a putative restriction endonuclease, a transcriptional repressor pifC, a RepFIB replication protein called repB_1, a C protein motif parG and a parF protein (annotated as VirC1 protein). None of these traits were found to be specific to CC4 genomes.

The absence of the GR1 in the clinically significant *C. sakazakii* CC4 isolates such as strains 721 and 1221 both of which are the CSF isolate suggest that these genes might not play a significant role in the virulence potential of CC4 (Table 5.1). Moreover the entire contents of the GR1 were also absent from the clinically significant non-CC4 isolates including 520 and 1249 which further suggest that the genes in the GR1 might also not be crucial for the *C. sakazakii* virulence in general.

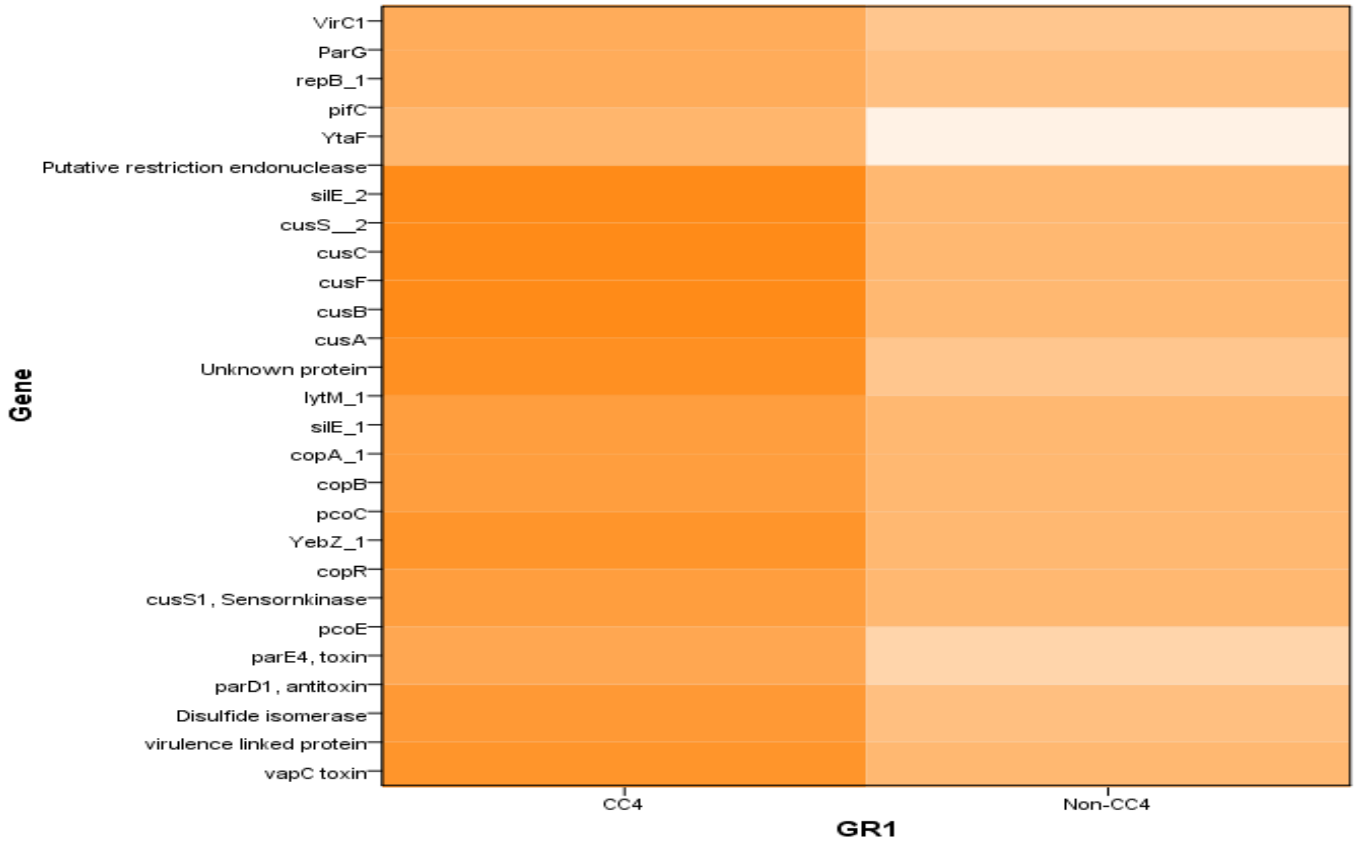


Figure 5. 1 Heat maps of variable regions GR1.

The image indicates heat map of variable region GR1 representing visually the percentage of presence of any trait within GR1 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 1 Gene contents of GR1 and their variation across *C. sakazakii* CC4 and non-CC4 genomes

Strain	Csak377_00005	Csak377_00006	Csak377_00009	Csak377_00018	Csak377_00019	Csak377_00023	Csak377_00028	Csak377_00031	Csak377_00041	Csak377_00044	Csak377_00047	Csak377_00051	Csak377_00052	Csak377_00053
	vapC toxin	Virulence associated protein	Disulfide isomerase	parD1, antitoxin protein	parE4, toxin protein	YebZ_1, inner membrane protein	lytM_1, Glycyl-glycine endopeptidase precursor	Protein of unknown function	Putative restriction endonuclease	YtaF, putative sporulation protein	pifC, Transcriptional repressor	repB_1, RepFIB replication protein A	ParG, C protein motif	VirC1 protein
1587	+	+	+	+	+	+	+	+	+	+	+	+	+	+
6	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1225	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1220	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1219	+	+	+	+	+	+	+	+	+	+	+	+	+	+
557	+	+	+	+	+	+	+	+	+	-	+	+	+	+
1231	+	+	-	+	+	+	+	+	+	-	+	-	-	-
1240	+	+	+	+	+	+	+	+	+	+	+	+	+	+
553	+	+	+	+	+	+	+	+	+	-	+	+	+	+
SP291	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4	+	+	+	+	+	+	+	+	+	+	+	+	+	+
767	+	+	+	-	-	-	+	+	-	-	-	-	-	-
701	+	+	+	-	-	-	-	+	-	-	-	-	-	+
558	-	-	-	-	-	+	+	+	-	-	-	-	-	-
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
721	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1221	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 5.1 continued

Strain	Csak377_00005	Csak377_00006	Csak377_00009	Csak377_00018	Csak377_00019	Csak377_00023	Csak377_00028	Csak377_00031	Csak377_00041	Csak377_00044	Csak377_00047	Csak377_00051	Csak377_00052	Csak377_00053
	vapC toxin	Virulence associated protein	Disulfide isomerase	parD1, antitoxin protein	parE4, toxin protein	YebZ_1, inner membrane protein	lytM_1, Glycyl-glycine endopeptidase precursor	Protein of unknown function	Putative restriction endonuclease	YtaF, putative sporulation protein	pifC, Transcriptional repressor	repB_1, RepFIB replication protein A	ParG, C protein motif	VirC1 protein
520	-	-	-	-	-	-	-	-	-	-	-	-	-	-
150	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1249	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1218	-	-	-	-	-	+	+	+	-	-	-	-	-	-
658	-	-	-	-	-	+	+	+	-	-	+	-	-	-
978	+	+	+	-	-	-	-	-	-	-	-	+	+	+
140	+	+	+	-	-	-	-	-	-	-	+	+	+	+
696	+	+	-	+	+	+	+	+	+	-	+	+	-	+
1	+	+	+	+	+	+	+	+	+	-	+	+	+	+
5	+	+	+	+	+	+	+	+	+	-	+	+	+	+
680	+	+	+	+	+	+	-	-	+	-	+	-	-	+

Red font colour indicates CC4 while purple indicates no-CC4 isolates.

5.5.1.2 The GR2

The GR2 region consisted of 37 genes located between Csak377_00100-00243 on the genome of the *C. sakazakii* CC4 isolate 377. The region was dominantly present in only two of the CC4 strains 1587 and 558 besides the reference CC4 strain 377. A tellurite resistance gene (Csak377_00166) was present in three of the CC4 strains as well as three non-CC4 strains. Three genes (Csak377_00169-00171) encoding type 1 restriction enzymes were present only in four of the CC4 strains of which two strains 1587 and 1231 were fatal meningitic isolates. A putative DNA double-strand break repair Rad50 ATPase (Csak377_00147) was present only in three CC4 and one non-CC4 strain. A RepFIB replication protein repB_2 (Csak377_00142) was present only in four CC4 and three non-CC4 strains whereas a thymidylate synthase thyA_1 (Csak377_00153) was detected in three CC4 and five non-CC4 strains. A XerD, site-specific recombinase encoding gene (Csak377_00241) was noted only in genomes of two CC4 and two non-CC4 strains, all of which were clinical isolates.

An RNA one modulator protein rop (Csak377_00238) was uniformly missing in all the CC4 and non-CC4 strains apart from the reference CC4 strain 377.

Other traits which were uniformly present in only three of the CC4 strains and missing from the rest of the CC4 and all non-CC4 strains included two aerobic cobaltochelatase subunit proteins cobT and cobS (Csak377_00125-6), a DNA polymerase III subunit alpha called dnaE_1 (Csak377_00127), a DNA polymerase thermostable protein polA_1 (Csak377_00134), a recombinase A protein recA_1 (Csak377_00136), calcineurin-like phosphoesterase superfamily domain protein (Csak377_001450, a DNA polymerase III subunit epsilon dnaQ_1 (Csak377_00158), a ribonuclease HI rnhA_1 (Csak377_00164), type 1 restriction enzyme hsdR_1 (Csak377_001690, a repressor of phase-1 flagellin protein (Csak377_001850, excinuclease ABC subunit B uvrB_1 (Csak377_00188), a putative chromosome-partitioning protein parB (Csak377_00190), a RepB family partition protein (Csak377_00190), an alpha-haemolysin translocation ATP-binding protein HlyB (Csak377_00191), a Flp pilus assembly protein CpaE (Csak377_00221), a replicative DNA helicase (Csak377_00225) a DNA primase (Csak377_00225) and a DNA ligase (Csak377_00230).

Four of the conjugation associated protein coding genes Csak377_00100-5 were noted only in five of the CC4 isolates 1587, SP291, 558, 1220, 1240 and two non-CC4 isolates 1 and 8 both of which belong to ST8. The presence of a transposase in the close proximity may indicate the acquisition of these genes through horizontal gene transfer.

In brief, the GR2 did not show any CC4 specific pattern since the genes within this region were present variably across the two groups.

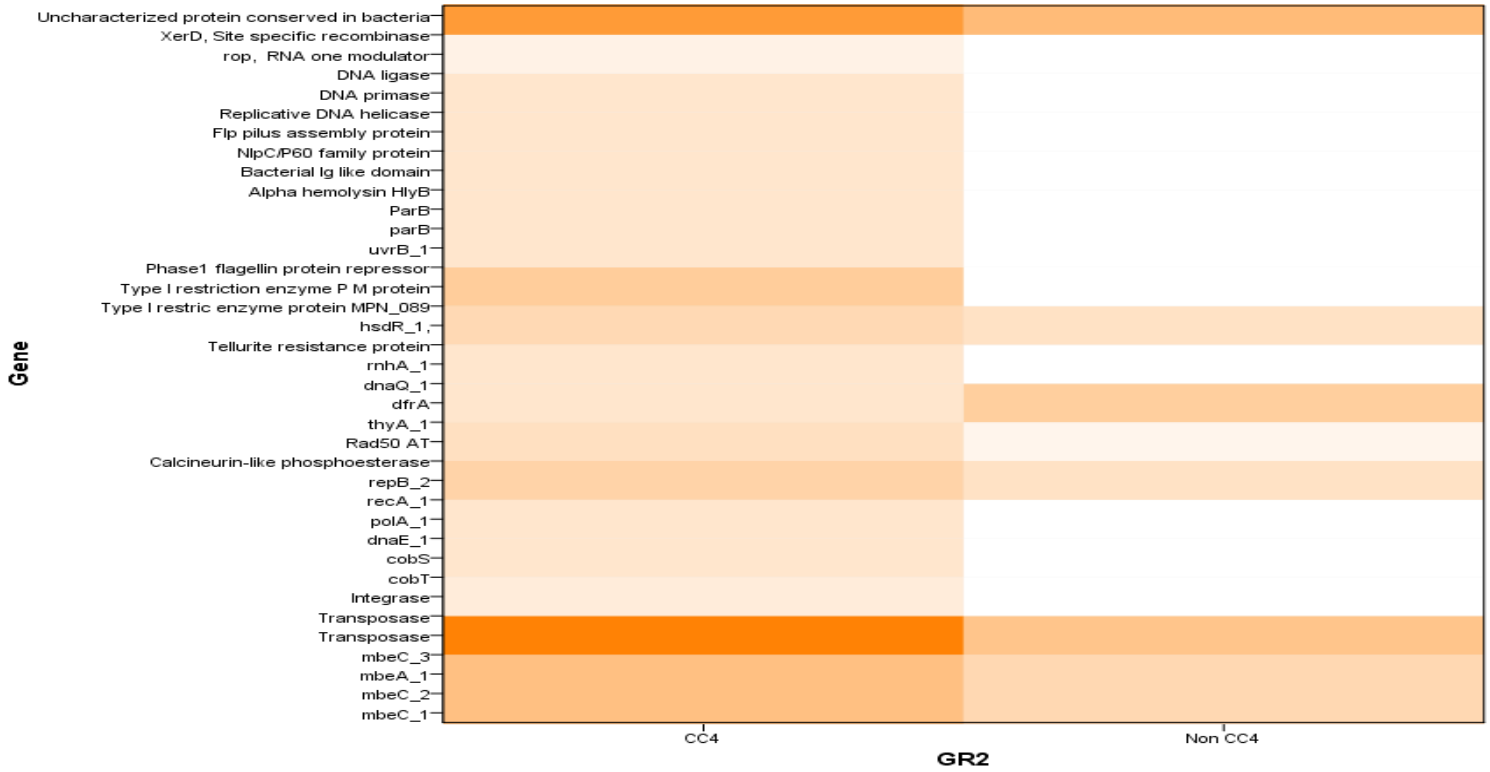


Figure 5. 2 Heat maps of variable regions GR2.

The image indicates heat map of variable region GR2 representing visually the percentage of presence of any trait within GR2 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 2 Gene contents of GR2 and their variation in *C. sakazakii* CC4 and non-CC4 strains

C. sakazakii isolate	Csak377_00100	Csak377_00101	Csak377_00102	Csak377_00105	Csak377_00117	Csak377_00121	Csak377_00122	Csak377_00125	Csak377_00126	Csak377_00127	Csak377_00134	Csak377_00136	Csak377_00142	Csak377_00145	Csak377_00147	Csak377_00153	Csak377_00155	Csak377_00158	Csak377_00164	Csak377_00166	Csak377_00169	Csak377_00170	Csak377_00171	Csak377_00185	Csak377_00188	Csak377_00189	Csak377_00190	Csak377_00191	Csak377_00205	Csak377_00211	Csak377_00221	Csak377_00225	Csak377_00226	Csak377_00230	Csak377_00238	Csak377_00241	Csak377_00243
	mbeC_1, Conjugative accessory protein	mbeC 2. Coniugative accessory protein	mbeA_1, DNA relaxase	mbeC_3, Conjugative accessory protein	Transposase	Transposase	Integrase	cobT, Aerobic cobaltochelatase subunit	cobS, Aerobic cobaltochelatase subunit	dnaE_1, DNA polymerase III subunit alpha	polA_1, DNA polymerase thermostable protein	recA_1, Recombinase A	repB_2, RepFIB replication protein A	Calcineurin-like phosphoesterase	Putative DNA double-strand break repair Rad50 ATPase	thyA_1, Thymidylate synthase	dfrA, Dihydrofolate reductase	dnaQ_1, DNA polymerase III subunit epsilon	rnhA_1, Ribonuclease HI	Tellurite resistance protein	hsdR_1, Type 1 restriction enzyme EcoR124II R protein	Type I restriction enzyme specificity protein MPN_089	Putative type I restriction enzymeP M protein	Repressor of phase-1 flagellin protein	uvrB_1, Excinuclease ABC subunit B,	parB, putativechromosome-partitioning protein	ParB/RepB/SpoOJ family partition protein	hlyB, Alpha-hemolysin	Bacterial Ig-like domain	NlpC/P60 family protein	Flp pilus assembly protein 2C ATPase CpaE	Replicative DNA helicase	DNA primase	DNA ligase	rop, RNA one modulator	XerD, Site-specific recombinase	Uncharacterized protein conserved in bacteria
6	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
701	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
767	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1225	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
20	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+
721	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+
1221	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+
1219	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
557	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1231	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+
4	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
553	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+

Table 5.2 continued

C. sakazakii isolate	Csak377_00100 mbeC_1, Conjugative accessory protein	Csak377_00101 mbeC 2. Coniugative accessory protein	Csak377_00102 mbeA_1, DNA relaxase	Csak377_00105 mbeC_3, Conjugative accessory protein	Csak377_00117 Transposase	Csak377_00121 Transposase	Csak377_00122 Integrase	Csak377_00125 cobT, Aerobic cobaltochelatase subunit	Csak377_00126 cobS, Aerobic cobaltochelatase subunit	Csak377_00127 dnaE_1, DNA polymerase III subunit alpha	Csak377_00134 polA_1, DNA polymerase thermostable protein	Csak377_00136 recA_1, Recombinase A	Csak377_00142 repB_2, RepFIB replication protein A	Csak377_00145 Calcineurin-like phosphoesterase	Csak377_00147 Putative DNA double-strand break repair Rad50 ATPase	Csak377_00153 thyA_1, Thymidylate synthase	Csak377_00155 dfrA, Dihydrofolate reductase	Csak377_00158 dnaQ_1, DNA polymerase III subunit epsilon	Csak377_00164 rnhA_1, Ribonuclease HI	Csak377_00166 Tellurite resistance protein	Csak377_00169 hsdR_1, Type 1 restriction enzyme EcoR124II R protein	Csak377_00170 Type I restriction enzyme specificity protein MPN_089	Csak377_00171 Putative type I restriction enzymeP M protein	Csak377_00185 Repressor of phase-1 flagellin protein	Csak377_00188 uvrB_1, Excinuclease ABC subunit B,	Csak377_00189 parB, putativechromosome-partitioning protein	Csak377_00190 ParB/RepB/SpoOJ family partition protein	Csak377_00191 hlyB, Alpha-hemolysin	Csak377_00205 Bacterial Ig-like domain	Csak377_00211 NlpC/P60 family protein	Csak377_00221 Flp pilus assembly protein 2C ATPase CpaE	Csak377_00225 Replicative DNA helicase	Csak377_00226 DNA primase	Csak377_00230 DNA ligase	Csak377_00238 rop, RNA one modulator	Csak377_00241 XerD, Site-specific recombinase	Csak377_00243 Uncharacterized protein conserved in bacteria
1220	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1240	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1587	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+
558	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-
SP291	+	+	+	+	+	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
680	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
1218	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
658	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-

Table 5.2 continued

C. sakazakii isolate	Csak377_00100	Csak377_00101	Csak377_00102	Csak377_00105	Csak377_00117	Csak377_00121	Csak377_00122	Csak377_00125	Csak377_00126	Csak377_00127	Csak377_00134	Csak377_00136	Csak377_00142	Csak377_00145	Csak377_00147	Csak377_00153	Csak377_00155	Csak377_00158	Csak377_00164	Csak377_00166	Csak377_00169	Csak377_00170	Csak377_00171	Csak377_00185	Csak377_00188	Csak377_00189	Csak377_00190	Csak377_00191	Csak377_00205	Csak377_00211	Csak377_00221	Csak377_00225	Csak377_00226	Csak377_00230	Csak377_00238	Csak377_00241	Csak377_00243
	mbeC_1, Conjugative accessory protein	mbeC 2. Coniugative accessory protein	mbeA_1, DNA relaxase	mbeC_3, Conjugative accessory protein	Transposase	Transposase	Integrase	cobT, Aerobic cobaltochelatase subunit	cobS, Aerobic cobaltochelatase subunit	dnaE_1, DNA polymerase III subunit alpha	polA_1, DNA polymerase thermostable protein	recA_1, Recombinase A	repB_2, RepFIB replication protein A	Calcineurin-like phosphoesterase superfamily domain protein	Putative DNA double-strand break repair Rad50 ATPase	thyA_1, Thymidylate synthase	dfrA, Dihydrofolate reductase	dnaQ_1, DNA polymerase III subunit epsilon	rnhA_1, Ribonuclease HI	Tellurite resistance protein	hsdR_1, Type 1 restriction enzyme EcoR124II R protein	Type I restriction enzyme specificity protein MPN_089	Putative type I restriction enzymeP M protein	Repressor of phase-1 flagellin protein	uvrB_1, Excinuclease ABC subunit B,	parB, putativechromosome-partitioning protein	ParB/RepB/SpoOJ family partition protein	hlyB, Alpha-hemolysin translocation+ATP-binding+protein HlyB	Bacterial Ig-like domain	NlpC/P60 family protein	Flp pilus assembly protein 2C ATPase CpaE	Replicative DNA helicase	DNA primase	DNA ligase	rop, RNA one modulator	XerD, Site-specific recombinase	Uncharacterized protein conserved in bacteria
696	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
520	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
140	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
150	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1249	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
978	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Red font colour indicates CC4 while purple indicates no-CC4 isolates.

5.5.1.3 THE GR3

The GR3 region consisted of 51 genes located between Csak377_00325-00480. The region was present in most CC4 and non-CC4 strains and included wide range of traits associated with diverse range of functions. Of significance were arsenical efflux proteins, an iron acquisition region *iutAiucABCD,* multidrug efflux associated proteins and a starvation sensing protein, discussed below (Figure 5.3, Table 5.3). Some of the traits in GR3 seem to be plasmid borne such as *iutAiucABCD*. It is important to refer to the *in silico* plasmid profiling (section 3.4.5) which suggested that plasmid pESA3 was completely missing in isolates 6, ES15 and 520 while it was partially detected in isolates 557, 1 and 680 (section 3.4.5).

5.5.1.3.1 ARSENIC RESISTANCE

Three genes present in majority of *C. sakazakii* genomes were associated with arsenic resistance. A homologue of the regulatory gene *arsR* (Csak377_00333) annotated as *aseR*, was found missing in three CC4 strains 701, 6 and 557 as well as three non-CC4 strains 1, 520 and 680; all of these isolates, except for CC4 isolate 701 either lack full or partial length plasmid pESA3 (section 3.4.5). The absence of this trait from CC4 isolate 701 remains unclear as this isolate contains the full length plasmid pESA3. Other two genes *arsB* (Csak377_00334) encoding for arsenic efflux pump protein and *arsC* (Csak377_00335) encoding an arsenate reductase were found to be uniformly present in all CC4 strains but missing in one non-CC4 strain 520. Heavy metal resistance is an important characteristic adaptable by bacteria which enables them to survive in the environment containing toxic metals such as Arsenic. Since the arsenic resistance associated genes were observed variably across both groups and did not show a dominant presence in any of the *C. sakazakii* CC4 or *C. sakazakii* non-CC4 group, the arsenic resistance genes may not form the basis of differentiation between CC4 and non-CC4 (Figure 5.3, Table 5.3), however they may represent strain specific variations. Moreover the plasmid less CC4 isolate 6 contains a number of homologue of the arsenic resistance associated genes (section 7.4.2.2.1) indicating that arsenic resistance in *C. sakazakii* is not strictly associated with plasmid profile.

5.5.1.3.2 POTASSIUM UPTAKE, PILUS ASSEMBLY PROTEIN AND STARVATION-SENSING PROTEIN

A potassium uptake gene annotated as *trkG_1* (Csak377_00336) was noted in all genomes except two CC4 strains 6 and 557 and three non-CC4 strains 1, 680 and 520, all of which lack either full length or partial length plasmid pESA3 (section 3.4.5) indicating that this is a plasmid borne trait. In addition, these isolates also lacked a pilus assembly associated gene *pilA* (Csak377_00342) encoding Tfp pilus assembly and a starvation sensing gene *rspA* (Csak377_00344) encoding for the starvation-sensing protein. The absence of these traits from the clinically significant CC4 and non-CC4 isolates may reflect that these genes might not be essential for the virulence of *C. sakazakii* (Figure 5.3, Table 5.3). Moreover their absence mainly in the plasmid less strains further indicates the plasmid borne nature of these traits.

5.5.1.3.3 THE IRON ACQUISITION REGION

The iron acquisition region *iutA*, *iucABCD* (Csak377_00390-00394) has previously been reported by Joseph et al (2012b) and investigated in section 3.4.2.10. The region was found in all except one CC4 strain 6 (low invasive strain) and one non-CC4 strain 520 in addition to the ES15 isolate. As mentioned earlier (section 3.4.5.1) that isolates ES15, 6 and 520 lack the full length plasmid indicating that this iron acquisition region is plasmid borne in *C. sakazakii* and is not linked with CC4 or non-CC4.

5.5.1.3.4 MULTIDRUG EFFLUX AND ANTIMICROBIAL PROTEINS

Two multidrug efflux associated genes were noted in GR3, a gene *mexB* (Csak377_00370) encoding for the multidrug-efflux transporter was present in all CC4 strains except strain 6 while absent from three non-CC4 strains 1, 680 and 520 all of which lack either full or partial length plasmid pESA3. Another multidrug efflux gene (Csak377_00381) encoding for multidrug efflux system protein MdtL was only found missing in CC4 strain 6 and non-CC4 strain 520, these strains also lacked two additional genes (Csak377_00375) annotated as antimicrobial resistance protein MdtL and a Metallo-beta-lactamase superfamily protein (Csak377_00413). A fibrinolysin associated gene *pla* (Csak377_00380) annotated as a coagulase/fibrinolysin precursor was noted in all the genomes except CC4 strain 6 and three non-CC4 strains 1, 5 and 520. The GR3 also contained two genes encoding for a putative

transcriptional regulator (Csak377_00473) and a recombination protein F (Csak377_00480) both of which were present only in six CC4 genomes and missing uniformly from the non-CC4 genomes. However, since these genes were missing in clinically significant *C. sakazakii* CC4 isolates such as 1225 and 701 (CSF isolates from fatal cases), therefore these traits may not be crucial in virulence of *C. sakazakii* CC4 (Figure 5.3, Table 5.3).

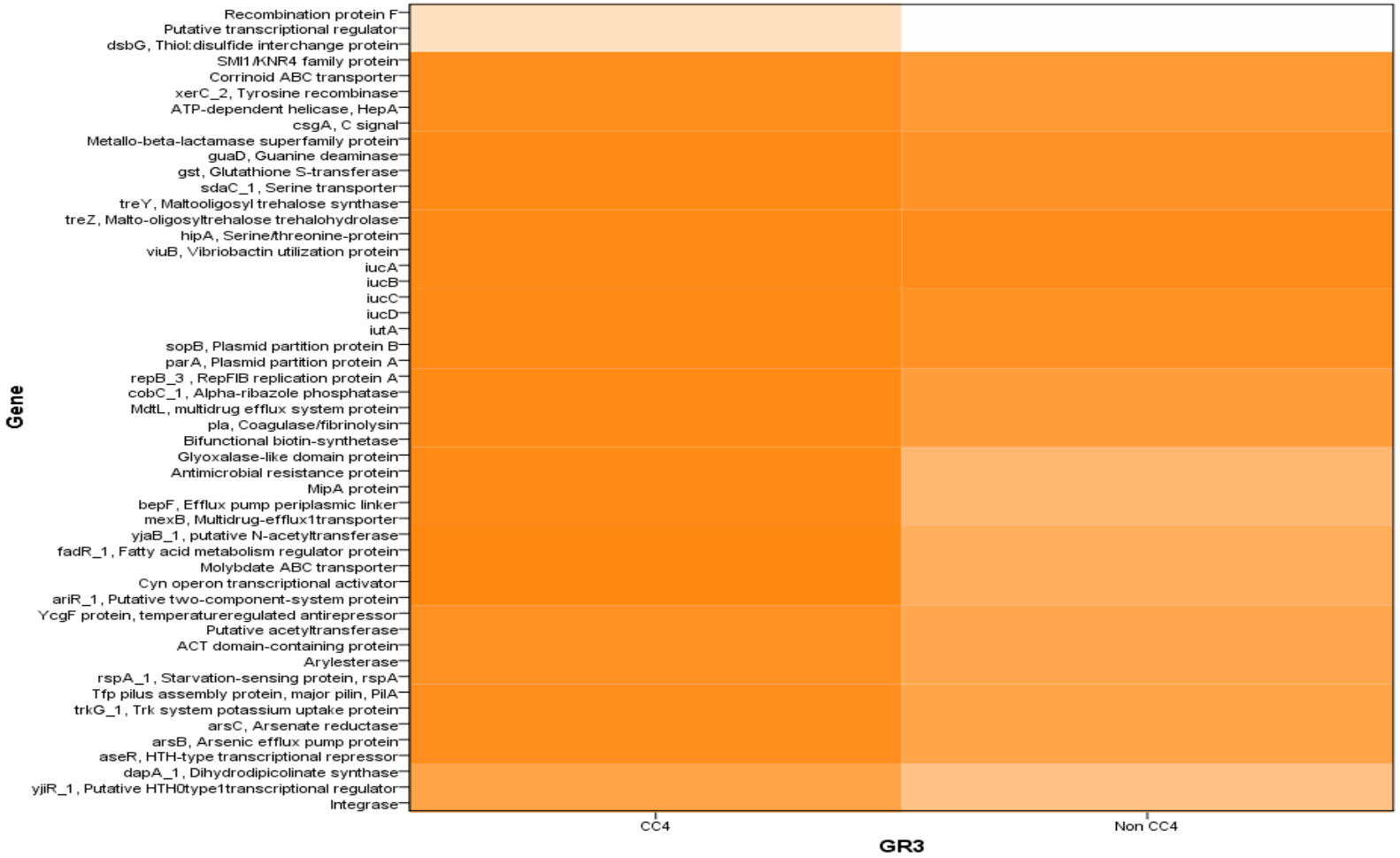


Figure 5. 3 Heat maps of variable regions GR3.

The image indicates heat map of variable region GR3 representing visually the percentage of presence of any trait within GR3 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 3 Gene contents of GR3 and their variation in *C. sakazakii* CC4 and non-CC4 strains

C. sakazakii isolate	Csak377_00325	Csak377 00330	Csak377 00332	Csak377 00333	Csak377 00334	Csak377 00335	Csak377 00336	Csak377 00342	Csak377 00344	Csak377 00346	Csak377 00348	Csak377 00350	Csak377 00354	Csak377 00357	Csak377 00362	Csak377 00363	Csak377 00365	Csak377 00367	Csak377 00370	Csak377 00371	Csak377 00374	Csak377 00375	Csak377 00376	Csak377 00377	Csak377 00380	Csak377 00381	Csak377 00385	Csak377 00386	Csak377 00388	Csak377 00389	Csak377 00390	Csak377 00391	Csak377 00392	Csak377 00393	Csak377 00394	Csak377 00396	Csak377 00398	Csak377 00399	Csak377 00400	Csak377 00406	Csak377 00411	Csak377 00412	Csak377 00413	Csak377 00417	Csak377 00418	Csak377 00421	Csak377 00423	Csak377 00427	Csak377 00432	Csak377 00473
	Integrase	viiR 1. transcriptional regulator viiR	Dihydrodipicolinate synthase dapA 1.	aseR. HTH-type transcriptional repressor AseR	arsB. Arsenic efflux pump protein	arsC. Arsenate reductase	trkG 1. Trk system potassium uptake protein trkG	Tfp pilus assembly protein, major pilin PilA	rspA 1. Starvation-sensing protein rspA	Arylesterase	ACT domain-containing protein	Putative acetyltransferase	antirepressor YcgF protein	ariR 1. Putative two-component-system connector	Cvn operon transcriptional activator	molybdate ABC transporter	fadR 1. Fatty acid metabolism regulator protein	viaB 1. putative N-acetyltransferase YiaB	mexB. Multidrug-efflux+transporter	bepF. Efflux pump periplasmic linker BepF protein	MltA-interacting protein MipA protein	Antimicrobial resistance protein Mig-14	Glyoxalase-like domain protein	Bifunctional biotin-(acetyl-CoA-carboxylase)	pla. Coagulase/fibrinolysin precursor	MdtL. multidrug efflux system protein	cobC 1. Alpha-ribazole phosphatase	repB 3 . RepFIB replication protein A	parA. Plasmid partition protein B	sopB. Plasmid partition protein B	iutA . Iron-regulated outer membrane proteins	iucD . L-lysine 6-monooxygenase	iucC. Aerobactin synthase IucC	iucB. N(6)-hydroxylysine O-acetyltransferase	iucA. Aerobactin synthase IucA	viuB. Vibriobactin utilisation protein ViuB	hipA. Serine/threonine-protein kinase HipA	treZ. Malto-oligosyltrehalose trehalohydrolase	treY. Maltooligosyl trehalose synthase	sdaC 1. Serine transporter	gst. Glutathione S-transferase GST-4.5	guaD. Guanine deaminase	Metallo-beta-lactamase superfamily protein	csgA. C signal	ATP-dependent helicase. HepA	xerC 2. Tyrosine recombinase XerC	Corrinoid ABC transporter	c SM11 / KNR4 family protein	dsbG. Thiol:disulfide	Putative transcriptional regulator
1587	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
558	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
701	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-
767	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1225	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
20	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
721	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
1221	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
1220	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
1219	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1240	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
1231	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
553	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-

Table 5.3 continued

C. sakazakii isolate	Csak377_00325 Integrase	Csak377_00330 yjiR_1. transcriptional regulator yjiR	Csak377_00332 dapA_1. Dihydrodipicolinate synthase	Csak377_00333 aseR. HTH-type transcriptional repressor AseR	Csak377_00334 arsB. Arsenic efflux pump protein	Csak377_00335 arsC. Arsenate reductase	Csak377_00336 trkG_1. Trk system potassium uptake protein trkG	Csak377_00342 Tfp pilus assembly protein, major pilin PilA	Csak377_00344 rspA_1. Starvation-sensing protein rspA	Csak377_00346 Arylesterase	Csak377_00348 ACT domain-containing protein	Csak377_00350 Putative acetyltransferase	Csak377_00354 antirepressor YcgF protein
6	-	-	-	-	+	+	-	-	-	-	-	-	+
557	+	-	-	-	+	+	-	-	-	-	-	-	+
520	-	-	-	-	-	-	-	-	-	-	-	-	+
1	+	-	-	-	+	+	-	-	-	+	-	-	-
680	-	-	+	+	+	+	-	-	-	+	-	-	-
SP291	+	+	+	+	+	+	+	+	+	+	+	+	+
1218	-	+	+	+	+	+	+	+	+	+	+	+	+
658	-	+	+	+	+	+	+	+	+	+	+	+	+
696	+	+	+	-	+	+	+	+	+	+	+	+	+
140	+	+	+	+	+	+	+	+	+	+	+	+	+
150	-	+	+	+	+	+	+	+	+	+	+	+	+
1249	-	+	+	+	+	+	+	+	+	+	+	+	+
978	+	+	+	+	+	+	+	+	+	+	+	+	+

C. sakazakii isolate	Csak377_00357 ariR_1. Putative two-component-system connector	Csak377_00362 Cyn operon transcriptional activator	Csak377_00363 molybdate ABC transporter	Csak377_00365 fadR_1. Fatty acid metabolism regulator protein	Csak377_00367 yiaB_1. putative N-acetyltransferase YiaB	Csak377_00370 mexB. Multidrug-efflux+transporter	Csak377_00371 bepF. Efflux pump periplasmic linker BepF protein	Csak377_00374 MltA-interacting protein MipA protein	Csak377_00375 Antimicrobial resistance protein Mig-14	Csak377_00376 Glyoxalase-like domain protein	Csak377_00377 Bifunctional biotin-(acetyl-CoA-carboxylase)	Csak377_00380 pla. Coagulase/fibrinolysin precursor	Csak377_00381 MdtL. multidrug efflux system protein
6	-	-	+	-	-	-	-	-	-	-	-	-	-
557	+	+	+	+	+	+	+	+	+	+	+	+	+
520	-	-	+	-	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-	-	-	-	-
680	-	-	+	-	-	-	-	-	-	-	-	-	-
SP291	+	+	+	+	+	+	+	+	+	+	+	+	+
1218	+	+	+	+	-	+	+	+	+	+	+	+	+
658	+	+	+	+	+	+	+	+	+	+	+	+	+
696	+	+	+	+	+	+	+	+	+	+	+	+	+
140	+	+	+	+	+	+	+	+	+	+	+	+	+
150	+	+	+	+	+	+	-	-	+	+	+	+	+
1249	+	+	+	+	+	+	+	+	+	+	+	+	+
978	+	+	+	+	+	+	+	+	+	+	+	+	+

C. sakazakii isolate	Csak377_00385 cobC_1. Alpha-ribazole phosphatase	Csak377_00386 repB_3. RepFIB replication protein A	Csak377_00388 parA. Plasmid partition protein A	Csak377_00389 sopB. Plasmid partition protein B	Csak377_00390 iutA. Iron-regulated outer membrane proteins	Csak377_00391 iucD. L-lysine 6-monooxygenase	Csak377_00392 iucC. Aerobactin synthase IucC	Csak377_00393 iucB. N(6)-hydroxylysine O-acetyltransferase	Csak377_00394 iucA. Aerobactin synthase IucA	Csak377_00396 viuB. Vibriobactin utilisation protein ViuB	Csak377_00398 hipA. Serine/threonine-protein kinase HipA	Csak377_00399 treZ. Malto-oligosyltrehalose trehalohydrolase
6	-	-	-	-	-	-	-	-	-	-	-	-
557	+	+	+	+	+	+	+	+	+	+	+	+
520	-	-	-	-	-	-	-	-	-	-	+	-
1	+	+	+	+	+	+	+	+	+	+	+	+
680	+	+	+	+	+	+	+	+	+	+	+	+
SP291	+	+	+	+	+	+	+	+	+	+	+	+
1218	+	+	+	+	+	+	+	+	+	+	+	+
658	+	+	+	+	+	+	+	+	+	+	+	+
696	+	+	+	+	+	+	+	+	+	+	+	+
140	+	+	+	+	+	+	+	+	+	+	+	+
150	+	+	+	+	+	+	+	+	+	+	+	+
1249	+	+	+	+	+	+	+	+	+	+	+	+
978	+	+	+	+	+	+	+	+	+	+	+	+

C. sakazakii isolate	Csak377_00400 treY. Maltooligosyl trehalose synthase	Csak377_00406 sdaC_1. Serine transporter	Csak377_00411 gst. Glutathione S-transferase GST-4.5	Csak377_00412 guaD. Guanine deaminase	Csak377_00413 Metallo-beta-lactamase superfamily protein	Csak377_00417 csgA. C signal	Csak377_00418 ATP-dependent helicase. HepA	Csak377_00421 xerC_2. Tyrosine recombinase XerC	Csak377_00423 Corrinoid ABC transporter	Csak377_00427 c SMI1 / KNR4 family protein	Csak377_00432 dsbG. Thiol:disulfide	Csak377_00473 Putative transcriptional regulator
6	-	-	-	-	-	-	-	-	-	-	-	-
557	+	+	+	+	+	+	+	+	+	-	-	+
520	-	-	-	-	-	-	-	-	-	-	-	-
1	+	+	+	+	+	+	+	+	+	+	+	-
680	+	+	+	+	+	+	+	+	+	+	+	-
SP291	+	+	+	+	+	+	+	+	+	+	+	+
1218	+	+	+	+	+	+	+	+	+	+	+	-
658	+	+	+	+	+	+	+	+	+	+	+	-
696	+	+	-	+	+	+	+	+	+	+	+	-
140	+	+	+	+	+	+	-	+	+	+	+	-
150	+	+	+	+	+	+	+	+	+	+	+	-
1249	+	+	+	+	+	+	-	+	+	+	+	-
978	+	+	+	+	+	+	-	+	+	+	+	-

Red font colour indicates CC4 while purple indicates non-CC4 isolates

5.5.1.4 The GR4

The GR4 region consisted of eight genes encoding fimbrial associated proteins which include major fimbrial protein precursor mrpA, a putative minor fimbrial subunit StfF, a putative fimbrial subunit SteE, a PAP fimbrial minor pilin protein precursor papH_1, a chaperone protein precursor papD, an outer membrane usher protein precursor papC_1, a PAP fimbrial minor pilin protein precursor papH_2 and fimbrial A protein precursor smfA (Csak377_00977-00984). The latter four fimbrial associated genes (Csak377_00981-984) were the homologues of π fimbriae (CTU_36420-450) in *C. turicensis* which were reported missing from the genome of *C. sakazakii* BAA-894 (Joseph et al. 2012b). The whole of the GR4 was uniformly present across all CC4 genomes however the region was variably present in non-CC4 genomes (Figure 5.4, Table 5.4). It is interesting to note that besides CC4 isolates, these fimbrial genes were also uniformly present in two ST12 isolates 696 (NEC II isolate) and 520 both of which although not CC4, yet represent clinically significant isolates. Three of the non-CC4 strains 1218 (CSF isolate), 978 and *C. sakazakii* 658 (CSF isolate) having a clinical origin, completely lacked this region while it was missing partially from three other non-CC4 strains 680, 140 and 150. Fimbriae are the appendage which bacteria use to adhere to and colonise the hosts surface such as intestinal epithelial cell lines (Baumler et al. 1996). The presence of this fimbrial region may give advantage to *C. sakazakii* CC4, however it seems unlikely since the region does not give a CC4 specific pattern as it is completely present in about half of the non-CC4 isolates investigated in this study. Moreover, non-CC4 isolates also possess other fimbrial gene clusters as discussed in (section 3.4.2.1).

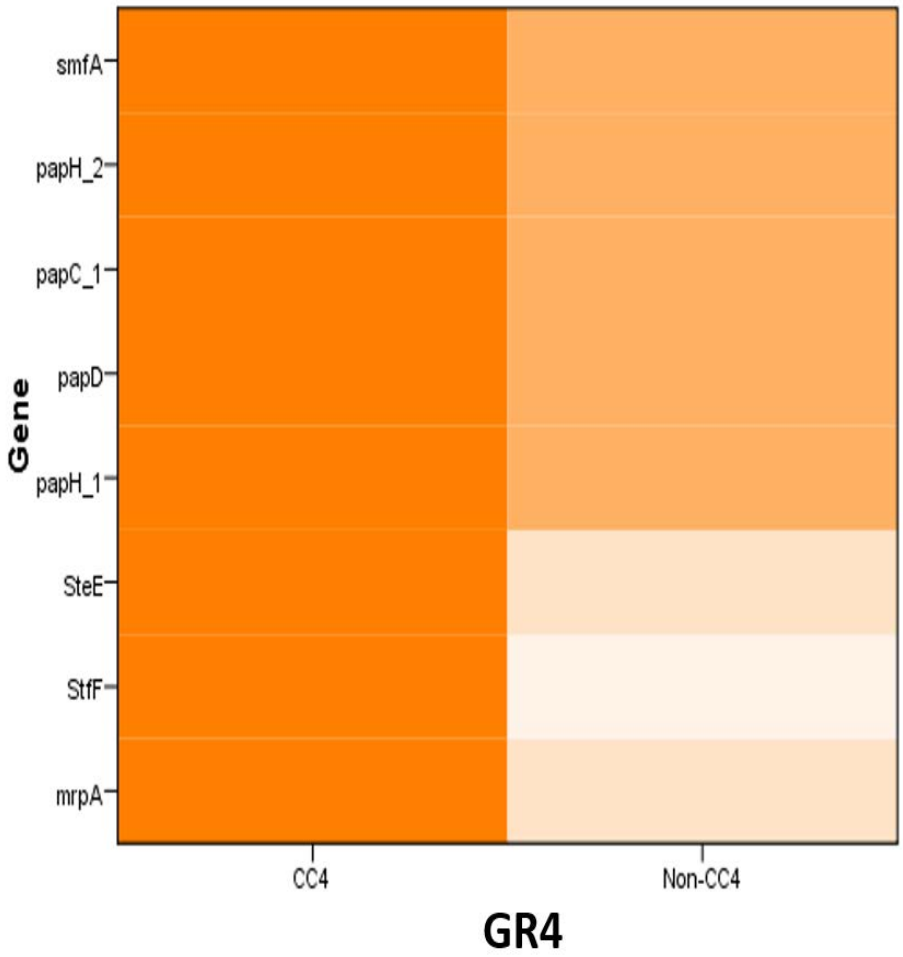


Figure 5. 4 Heat maps of variable regions GR4.

The image indicates heat map of variable region GR4 representing visually the percentage of presence of any trait within GR4 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 4 Gene contents of GR4 and their variation in *C. sakazakii* CC4 and non-CC4 strains

	Louc_tag	Csak377_00977	Csak377_00978	Csak377_00979	Csak377_00980	Csak377_00981	Csak377_00982	Csak377_00983	Csak377_00984
C. sakazakii isolate	ST/CC	Major MR/P fimbria protein precursor, mrpA	Putative minor fimbrial subunit StfF	Putative fimbrial subunit SteE	PAP fimbrial minor pilin protein precursor, papH_1	Chaperone protein papD precursor, papD	Outer membrane usher protein papC precursor, papC_1	PAP fimbrial minor pilin protein precursor, papH_2	Fimbria A protein precursor, smfA
1587	109/4	+	+	+	+	+	+	+	+
558	4/4	+	+	+	+	+	+	+	+
6	4/4	+	+	+	+	+	+	+	+
767	4/4	+	+	+	+	+	+	+	+
1225	4/4	+	+	+	+	+	+	+	+
20	4/4	+	+	+	+	+	+	+	+
721	4/4	+	+	+	+	+	+	+	+
1221	4/4	+	+	+	+	+	+	+	+
1220	4/4	+	+	+	+	+	+	+	+
1219	4/4	+	+	+	+	+	+	+	+
701	4/4	+	+	+	+	+	+	+	+
557	4/4	+	+	+	+	+	+	+	+
1231	4/4	+	+	+	+	+	+	+	+
1240	4/4	+	+	+	+	+	+	+	+
553	4/4	+	+	+	+	+	+	+	+
4	15/4	+	+	+	+	+	+	+	+
SP291	4/4	+	+	+	+	+	+	+	+
1	8/8	+	-	+	+	+	+	+	+
5	8/8	+	+	+	+	+	+	+	+
680	8/8	-	-	-	+	+	+	+	+
1218	1/1	-	-	-	-	-	-	-	-
658	1/1	-	-	-	-	-	-	-	-
978	3/3	-	-	-	-	-	-	-	-
140	40/45	-	-	-	+	+	+	+	+
150	16/16	-	-	-	+	+	+	+	+
696	12	+	+	+	+	+	+	+	+
520	12	+	+	+	+	+	+	+	+
1249	31/31	+	+	+	+	+	+	+	+

5.5.1.5 The GR5

The GR5 consisted of 15 genes encoding for diverse range of functions located between Csak377_01070 and Csak377_01731. The genes in GR5 encoded for the antitermination protein, KilA-N domain protein, anaerobic benzoate catabolism transcriptional regulator, lactose operon repressor laci_1, putative transcriptional regulator, a spore protein hspA (hypothetical protein homologue), a putative membrane protein, SMI1/KNR4 family protein, ribulose-5-phosphate 4-epimerase and aldolases, site-specific recombinase XerD, ATP-dependent helicase HepA, type-1 restriction enzyme R protein hsdR_2, putative toxin YpjF and antitoxin YeeU. Although the above mentioned traits were noted more in CC4, yet all of these traits show a mosaic pattern for the presence across two groups (Figure 5.5, Table 5.5) and hence difficult to link with the virulence potential of *C. sakazakii* CC4.

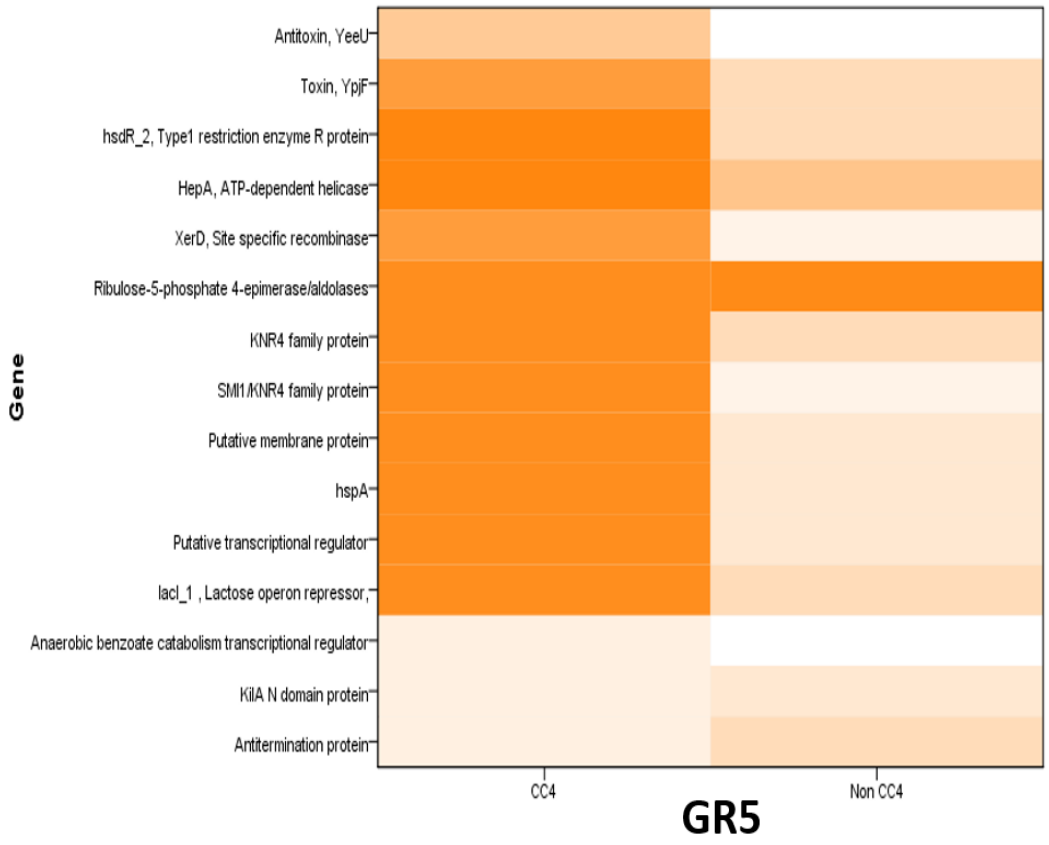


Figure 5. 5 Heat maps of variable regions GR5.

The image indicates heat map of variable region GR5 representing visually the percentage of presence of any trait within GR5 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 5 Gene contents of GR5 and their variation in *C. sakazakii* CC4 and non-CC4 strains

		Csak377_01070	Csak377_01072	Csak377_01091	Csak377_01213	Csak377_01218	Csak377_01219	Csak377_01222	Csak377_01316	Csak377_01318	Csak377_01337	Csak377_01587	Csak377_01727	Csak377_01729	Csak377_01730	Csak377_01731
***C. sakazakii* isolate**	**ST/CC**	**Antitermination protein**	**KilA-N domain protein**	**Anaerobic benzoate catabolism transcriptional regulator**	**Lactose operon repressor, lacI_1**	**Putative transcriptional regulator**	**Spore protein SP21, hspA**	**Putative membrane protein**	**SMI1 / KNR4 family protein**	**SMI1 / KNR4 family protein**	**Ribulose-5-phosphate 4-epimerase and aldolases**	**Site-specific recombinase XerD**	**ATP-dependent helicase HepA**	**Type-1 restriction enzyme R protein, hsdR_2**	**Putative toxin YpjF, ypjF**	**Antitoxin YeeU**
1225	**4/4**	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1231	**4/4**	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
1587	**109/4**	-	-	-	+	+	+	+	-	+	+	+	+	+	+	-
558	**4/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
6	**4/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
767	**4/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-
20	**4/4**	-	-	-	-	+	+	+	+	+	+	-	+	+	+	-
721	**4/4**	-	-	-	+	-	-	-	+	+	-	-	+	+	-	-
1221	**4/4**	-	-	-	+	-	-	-	+	+	-	-	+	+	-	-
1220	**4/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-
1219	**4/4**	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-
701	**4/4**	-	-	-	+	+	+	+	+	-	+	+	+	+	-	-
557	**4/4**	-	-	-	+	+	+	+	-	-	+	-	+	+	+	+
1240	**4/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
553	**4/4**	-	-	-	+	+	+	+	+	+	+	+	-	-	+	-
4	**15/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
SP291	**4/4**	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
1	**8/8**	+	-	-	+	+	+	+	-	-	+	-	-	-	+	-
696	**12**	-	-	-	+	+	+	+	+	-	+	+	+	+	+	-
5	**8/8**	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
680	**8/8**	+	-	-	+	-	-	-	-	-	+	-	-	-	+	-
1218	**1/1**	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-
658	**1/1**	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-
150	**16/16**	-	-	-	-	-	-	-	-	+	+	-	+	-	-	-
1249	**31/31**	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
978	**3/3**	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-
520	**12**	+	+	-	-	-	-	-	-	-	-	-	+	-	-	-
140	**40/45**	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-

5.5.1.6 The GR6

The GR6 contained 7 genes (located between Csak377_02402 to Csak377_02980). A leucine efflux protein leuE (Csak377_02402) was detected in only six of the CC4 and only one non-CC4 strain 978 (Figure 5.6, Table 5.6). A site-specific recombinase XerD (Csak377_02405) was detected in all except one CC4 strain 6, however only in four non-CC4 strains. Other five genes in GR6, Csak377_02976-80 encoding for a phenylacetate-CoA ligase, phenyloxazoline synthase mbtB, 1-deoxy-D-xylulose-5-phosphate synthase dxs_2, transketolase 2 tktB_1 and 3-oxoacyl-(acyl-carrier-protein) reductase fabG_5 respectively; were present uniformly across the CC4 group and six of the non-CC4 strains while variably present in remaining five of the non-CC4 strains. Hence, none of these traits show a CC4 specific pattern and therefore could not be linked with the virulence potential of CC4 (Figure 5.6, Table 5.6).

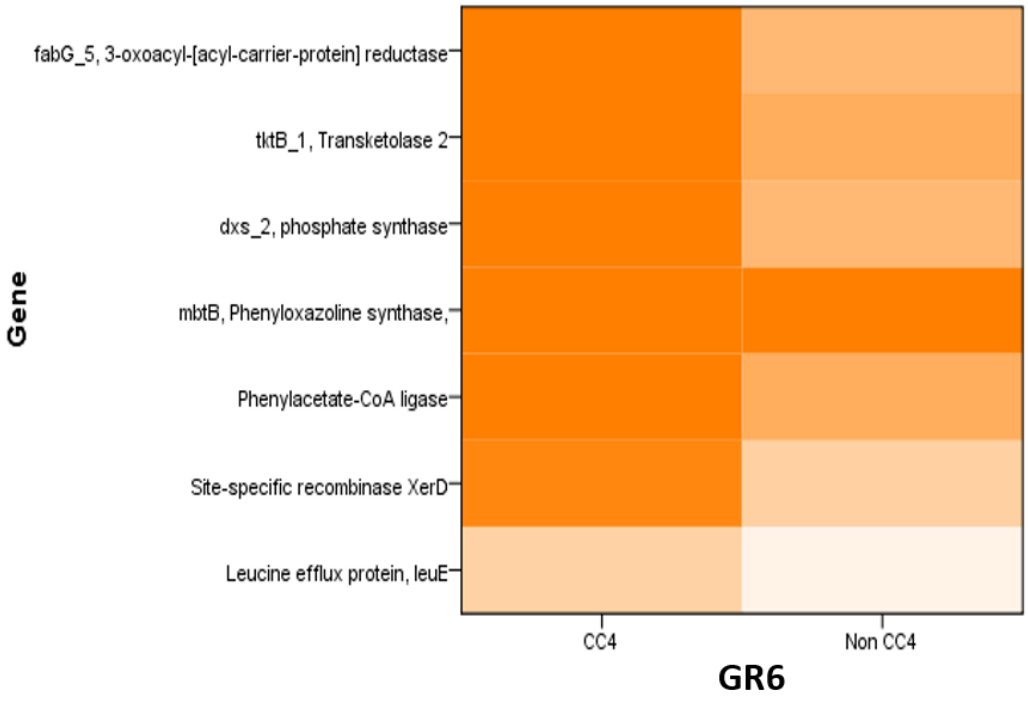


Figure 5. 6 Heat maps of variable regions GR6.

The image indicates heat map of variable region GR6 representing visually the percentage of presence of any trait within GR6 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 6 Gene contents of GR6 and their variation in *C. sakazakii* CC4 and non-CC4 strains

C. sakazakii isolate	ST /CC	Csak377_02402 Leucine efflux protein, leuE	Csak377_02405 Site-specific recombinase XerD	Csak377_02976 Phenylacetate-CoA ligase	Csak377_02977 Phenyloxazoline synthase MbtB, mbtB	Csak377_02978 1-deoxy-D-xylulose-5-phosphate synthase, dxs_2,	Csak377_02979 Transketolase 2, tktB_1	Csak377_02980 3-oxoacyl-(acyl-carrier-protein) reductase FabG, fabG_5
1587	109/4	+	+	+	+	+	+	+
558	4/4	-	+	+	+	+	+	+
6	4/4	-	+	+	+	+	+	+
767	4/4	+	-	+	+	+	+	+
1225	4/4	-	+	+	+	+	+	+
20	4/4	+	+	+	+	+	+	+
721	4/4	-	+	+	+	+	+	+
1221	4/4	-	+	+	+	+	+	+
1220	4/4	-	+	+	+	+	+	+
1219	4/4	-	+	+	+	+	+	+
701	4/4	-	+	+	+	+	+	+
557	4/4	+	+	+	+	+	+	+
1231	4/4	-	+	+	+	+	+	+
1240	4/4	-	+	+	+	+	+	+
553	4/4	+	+	+	+	+	+	+
4	15/4	-	+	+	+	+	+	+
SP291	4/4	+	+	+	+	+	+	+
1218	1/1	-	+	+	+	+	+	+
658	1/1	-	+	+	+	+	+	+
1	8/8	-	+	-	+	-	-	-
5	8/8	-	-	-	+	-	-	-
680	8/8	-	-	-	+	-	-	-
696	12	-	-	-	+	-	-	-
520	12	-	-	+	+	-	+	-
140	40/45	-	-	+	+	+	+	+
150	16/16	-	-	+	+	+	+	+
1249	31/31	-	-	+	+	+	+	+
978	3/3	+	+	+	+	+	+	+

5.5.1.7 The GR7

The region GR7 contained 18 genes (located between Csak377_03205- 03963) (Figure 5.7, Table 5.7). The genes located between Csak377_03841-03850 were the O-antigen associated genes. All of these genes are part of the O:2 serotype of *C. sakazakii* described by Mullane et al. (2008). The genes were present uniformly across the CC4 group except strain 1240. The observation further strengthened the findings of the chapter 3 (section 3.4.4) where all of the CC4 isolates aligned with O:2 serotype except for 1240 which belongs to a different O-antigen group (O:4). The variation of strain 1240 in GR7 from other CC4 strains further justify its longer branch length on the SNP phylogeny (Figure 7.1). The O-antigen variation between CC4 and non-CC4 has been discussed in more detail in (section 3.4.4). In brief, the variation indicates that O:2 is the dominant serotype in CC4, however it is not exclusive to CC4 as some non-CC4 isolates such as 1249 and 978 also belong to O:2 (section 3.4.4).

Additionally in GR7, an acyltransferase family protein (Csak377_03205) was present in all CC4 isolates except strain 1587 while present only in one *C. sakazakii* non-CC4 isolate 978. Similarly, a putative toxin/anti-toxin Ykf/YfjZ (Csak377_03206-7) was also detected mainly in CC4 strains (present in 14/18 strains) while missing in majority of non-CC4 strains (present in 3/11 strains). Other traits in GR7 included an antirestriction protein (Csak377_03210) were present mostly in CC4 group (13/18) while detected only in three non-CC4 strains. Antirestriction proteins such as ArdA are typically plasmid encoded or found in bacteriophages; these antirestriction proteins protect the host DNA from degradation by inhibiting the restriction modification enzymes of the invading foreign DNA (Zavilgelsky and Rastorguev, 2008). The enrichment of antirestriction protein (Csak377_03210) in CC4 could be important towards genomic stability of *C. sakazakii* CC4 and hence warrants further investigation.

A GTPase era_2 (Csak377_03216) was also detected in most of the CC4 strains (12/18) while only in three of the non-CC4 strains. GTPases are hydrolase enzymes universally distributed in bacteria; these enzymes are involved in functions such as protein synthesis, cell division and protein translocation (Caldon and March, 2003). Whether the enrichment of Csak377_03216 encoding a GTPase in CC4 genomes, is of significance, requires further investigation.

A site-specific tyrosine recombinase XerC (Csak377_03917) was detected in three CC4 and one non-CC4 strain, a putative HTH-type transcriptional regulator (Csak377_03919) was detected in only two CC4 and two non-CC4 strains. A transcriptional repressor DicA (Csak377_03920) was detected in only two of the CC4 strains 1587 and SP291 and was not noted in any of the non-CC4 strains. Likewise, a putative two-component system connector protein ariR_2 (Csak377_03963) was also detected in only three of the CC4 strains while missing uniformly across the non-CC4 group.

In brief, the GR7 was of significance as it contained some of the O-antigen associated genetic traits. From the previous section 3.4.4, it has been shown the O:2 is the dominant serotype for *C. sakazakii* CC4, however it is not specific to CC4 as non-CC4 isolates 1249 and 978 were also O:2 serotype. The remaining traits in GR7 showed a mosaic pattern for their presence across two groups and hence none of the trait was CC4 specific.

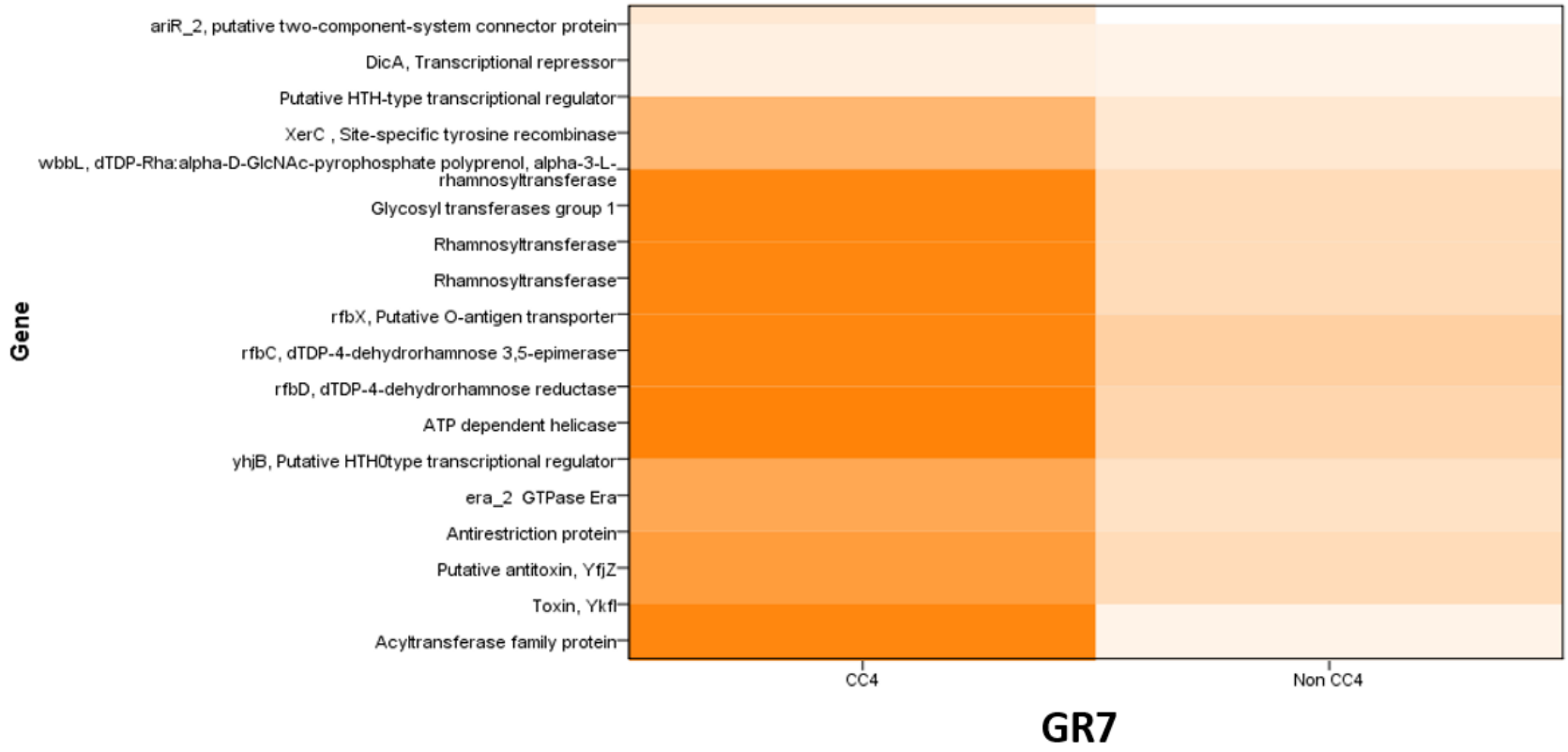


Figure 5. 7 Heat maps of variable regions in GR7.

The image indicates heat map of variable region GR7 representing visually the percentage of presence of any trait within GR7 across CC4 and non-CC4 groups. The intensity of the colour increases as the percentage of presence of a gene in a group increases and vice versa. The heat maps were generated in SPSS.

Table 5. 7 Gene contents of GR7 and their variation in *C. sakazakii* CC4 and non-CC4 strains

	Locus_tag	Csak377_03205	Csak377_03206	Csak377_03207	Csak377_03210	Csak377_03216	Csak377_03419	Csak377_03431	Csak377_03841	Csak377_03843	Csak377_03844	Csak377_03845	Csak377_03848	Csak377_03849	Csak377_03850	Csak377_03917	Csak377_03919	Csak377_03920	Csak377_03963
	Product	Acyltransferase family protein	Toxin Ykfl	Putative antitoxin YfjZ,	Antirestriction protein	GTPase Era	Putative HTH-type transcriptional regulator yhjB	ATP-dependent helicase	dTDP-4-dehydrorhamnose reductase, rfbD	dTDP-4-dehydrorhamnose 3,5-epimerase, rfbC	Putative O-antigen transporter, rfbX	Rhamnosyltransferase	rhamnosyltransferase	Glycosyl transferases group 1	dTDP-Rha:alpha-D-GlcNAc-pyrophosphate polyprenol, alpha-3-L-rhamnosyltransferase, wbbL	Site-specific tyrosine recombinase XerC	Putative HTH-type transcriptional regulator	Transcriptional repressor DicA	ariR_2putative two-component-system connector protein AriR
Strain	**ST/CC**																		
1240	4/4	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
1587	109/4	-	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+
558	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
6	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
767	4/4	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-
1225	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
20	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+
721	4/4	+	-	-	-	-	+	+	+	+	+	+	+	+	+	-	-	-	-
1221	4/4	+	-	-	-	-	+	+	+	+	+	+	+	+	+	-	-	-	-
1220	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
1219	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
701	4/4	+	-	-	-	-	+	+	+	+	+	+	+	+	+	-	-	-	-
557	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
1231	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-

Table 5.7 continued

	Locus_tag	Csak377_03205	Csak377_03206	Csak377_03207	Csak377_03210	Csak377_03216	Csak377_03419	Csak377_03431	Csak377_03841	Csak377_03843	Csak377_03844	Csak377_03845	Csak377_03848	Csak377_03849	Csak377_03850	Csak377_03917	Csak377_03919	Csak377_03920	Csak377_03963
	Product	Acyltransferase family protein	Toxin YkfI	Putative antitoxin YfjZ,	Antirestriction protein	GTPase Era	Putative HTH-type transcriptional regulator yhjB	ATP-dependent helicase	dTDP-4-dehydrorhamnose reductase, rfbD	dTDP-4-dehydrorhamnose 3,5-epimerase, rfbC	Putative O-antigen transporter, rfbX	Rhamnosyltransferase	rhamnosyltransferase	Glycosyl transferases group 1	dTDP-Rha:alpha-D-GlcNAc-pyrophosphate polyprenol, alpha-3-L-rhamnosyltransferase, wbbL	Site-specific tyrosine recombinase XerC	Putative HTH-type transcriptional regulator	Transcriptional repressor DicA	ariR_2putative two-component-system connector protein AriR
553	4/4	+	+	+	+	-	+	-	+	+	+	+	+	+	+	-	-	-	-
4	15/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
SP291	4/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1249	31/31	-	-	-	-	-	+	+	+	+	+	+	+	+	+	-	-	-	-
978	3/3	+	-	-	-	-	-	-	+	+	+	+	+	+	+	-	+	-	-
1	8/8	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	-	-
5	8/8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
680	8/8	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
1218	1/1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
658	1/1	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
696	12	-	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-
520	12	-	-	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-
140	40/45	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
150	16/16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-

The genomic comparison using the whole genome based comparative approach identified a number of unique traits in *C. sakazakii* 377, however none of the trait could form the basis of clear differentiation between CC4 and non-CC4 strains.

5.5.1.8 LARGE ADHESION ASSOCIATED GENE IN *C. SAKAZAKII* GENOMES

An interesting observation in this study was the presence of a large gene (~15kb) CSAK377_00585 (or CSSP291_03155 in CC4 isolate SP291) annotated as hypothetical protein which was present uniformly in CC4 genomes (18/18) yet variable in non-CC4 genomes (6/12) (Table 5.8). Previously, a large (600 kDa) non-fimbrial adhesin encoding gene *siiE* (CAE11797) has been reported in *Salmonella* Typhimurium (Gerlach et al. 2007). The CSSP291_03155 was suspected to be a homologue of *siiE*, however the ACT comparison of CSSP291_03155 and *siiE* did not show any homology. The whole length nucleotide and protein sequences of CSSP291_03155 were searched for homology using BLAST in NCBI. It showed homology only in *Cronobacter* with partial gene homologues being identified only in non-CC4 *C. sakazakii* strain ES15 (ES15_0912) and *C. turicensis* z3032 (CTU_32080). Further investigation was undertaken and all the possible open reading frames ORFs within this gene were predicted using Artemis for ease of BLAST searching. The Artemis predicted a number of ORFs which were searched for homology using BLASTx in NCBI (Table 5.9). Majority of the ORFs were the homologues of adhesins or T1SS proteins; ten of these ORFs were the partial homologues of adhesin proteins, five of them were the homologues of type 1 secretion system (T1SS) while remaining two were the homologues of large repetitive proteins. Bacterial adhesins are important in mediating the attachment and colonisation of bacteria to the host cells. Bacterial adhesins can recognize and bind to the receptors on the host cells which in turn may trigger cellular pathways which can facilitate the bacterial colonisation or invasion. The attachment of bacterial cells through adhesion to the host cells sometime alters bacterial gene expression which may facilitate the bacterial pathogenesis (Soto & Hultgren, 1999). The presence of this large adhesion associated gene may be important in virulence of *C. sakazakii* in general, however it is not a CC4 specific trait since it was also found to be variably present in half of the non-CC4 *C. sakazakii* genomes (6/12). Further investigation is warranted to characterise this genomic island to elucidate its precise role in *C. sakazakii.*

Table 5. 8 Presence of the full length adhesion associated gene CSSP291_03155 in *C. sakazakii* strains

C. sakazakii isolate	ST	CC	CSSP291_03155
6	4	4	+
20	4	4	+
377	4	4	+
553	4	4	+
558	4	4	+
701	4	4	+
721	4	4	+
767	4	4	+
1219	4	4	+
1220	4	4	+
1221	4	4	+
1225	4	4	+
1231	4	4	+
1240	4	4	+
SP291	4	4	+
557	4	4	+
1587	109	4	+
4	15	4	+
978	3	3	+
520	12		+
696	12		+
150	16	16	+
1249	31	31	+
140	40	45	+
658	1	1	-
1218	1	1	-
ES15	125	100	-
1	8	8	-
5	8	8	-
680	8	8	-

Table 5. 9 BLASTx analysis of the giant adhesion associated gene in *C. sakazakii* SP291*

		CSSP291_03155 analysis after division into ORFs	
ORF No	**Coordinates**	**BLASTx homology**	**Accession**
1	654177:655082	Adhesin, AT-2 family transporter (*Pantoea sp. Sc1*)	WP_009089074
2	654577:655104	Hemagglutinin/hemolysin/adhesin-like protein (*Erwinia billingiae* Eb661)	EbC_30340
3	654698:655690	Type 1 secretion target domain-containng protein, partial (*Enterobacter mori*)	WP_010426201.1
4	655192:656457	Adhesin for cattle intestine colonisation (*Pectobacterium wasabiae*)	EbC_30340
5	655908:656390	Autotransporter adhesin (*Yersinia mollaretii* ATCC 43969)	EEQ10543
6	656054:657160	Adhesin for cattle intestine colonisation (*Escherichia coli*)	CE10_0459; AEQ11302
7	656607:657062	Autotransporter adhesin (*Yersinia mollaretii* ATCC 43969)	WP_004875316.1
8	656698:657705	Type 1 secretion target domain-containing protein (*Enterobacter cloacae*)	WP_020884206
9	657204:657641	Adhesin for cattle intestine colonisation (*Escherichia coli*)	WP_001454914
10	657332:665359	Adhesin for cattle intestine colonisation (*Pectobacterium wasabiae* CFBP 3304)	EJS92462
11	658528:659079	Type 1 secretion target domain-containing protein, partial (*Enterobacter mori*)	WP_010430919
12	659406:659918	Hemagglutinin/hemolysin/adhesin-like protein (*Erwinia billingiae* Eb661)	CAX60565
13	660046:660459	Large repititive protein, CTU_32110 (*Cronobacter turicensis* z3032)	CTU_32110
14	661789:662382	Type 1 secretion target domain-containing protein (*Enterobacter sp.* SST3)	WP_008502299
15	661875:662246	Large repetitive protein (*Cronobacter turicensis* 564)	WP_007763466.1
16	663088:663549	Autotransporter adhesin (*Yersinia mollaretii*)	WP_004875316
17	665360:665869	Putative TolC family type I secretion outer membrane protein (*Cronobacter sakazakii* ES15)	ES15_0913; AFJ98486

***** Not all the ORFs within CSSP291_03155 are listed here

5.6 GENERAL DISCUSSION AND CONCLUSION

Whole genome based phylogeny of 30 *C. sakazakii* strains indicated close clustering of 18 CC4 strains collected from 9 different countries for over 6 decades; indicating a clonal genome signature for this clade (Figure 3.3). Although some interesting variations were observed, previous analysis presented in this thesis based on the screening of the potential virulence associated traits (chapter 3) and core genome comparative analysis (chapter 4) in *C. sakazakii* genomes could not identify any CC4 specific associated gene which could clearly explain its predominance in neonatal meningitis or its persistence in the environment. Therefore, it was decided to take the comparative genomic analysis further and use a different approach in the view that it may help identify any trait which is CC4 specific. Hence a gene by gene search approach was used by comparing one *C. sakazakii* CC4 (377) and *C. sakazakii* non-CC4 genome (ES15) to identify unique traits in *C. sakazakii* CC4 genome and expand the analysis to the remaining *C. sakazakii* CC4 and non-CC4 genomes to determine genes specific to CC4. The genomic comparison was undertaken using ACT. The gene by gene search identified 465 genes which were unique in the genome of *C. sakazakii* 377 (Table S4). However, 63% of these gene represented hypothetical proteins of unknown function and phage associated traits. To make the genomic comparison simpler, the hypothetical and phage associated traits were omitted and the remaining genes were divided into 7 genomic regions based on their location and functional categories. It must however be noted that the all genes within a GR do not necessarily located next to each other, the GRs were assigned just to present the genomic comparison in a simplified manner based on functional similarities and close location.

The traits in the GR1 were the homologues of copper and silver homoeostasis associated genes which have previously been described (Kucerova et al. 2010; Joseph et al. 2012b). The genes within GR1 were detected in most of the *C. sakazakii* CC4 and non-CC4 genomes, however no CC4 specific pattern was observed for any genes within GR1. Furthermore, since the region was absent in clinically significant *C. sakazakii* CC4 and non-CC4 isolates such as isolates 20, 1221 and 1249 all of which were CSF isolates, it suggested that the region may not be essential for the virulence potential of *C. sakazakii* in general or CC4 in particular.

The GR2 consisted of genes with diverse range of functions. Overall the genes within GR2 were missing largely in both CC4 and non-CC4 strains and no CC4 specific pattern was observed for any of the traits within this region.

The GR3 consisted of a number of important virulence associated traits such as arsenic resistance, iron acquisition, potassium uptake, pilus assembly, starvation sensing and multi drug efflux associated genes. However, the region did not indicate the basis of differentiation between CC4 and non-CC4 as most of these genes were detected in majority of the *C. sakazakii* isolates across both groups. Iron acquisition, heavy metal resistance, multi-drug efflux are the characteristics adapted by several bacterial pathogens which help them to survive within hostile conditions and cause infections in their susceptible hosts. The presence of these traits in most CC4 and non-CC4 isolates indicate that these traits contribute equally towards the virulence potential of *C. sakazakii* and cannot be linked specifically to any of the subgroup. The iron acquisition associated genes in GR3 have been also have been investigated in chapter 3 (3.4.2.10).

The GR4 consisted of 8 genes encoding for the fimbrial adhesion genes. The fimbriae are the important bacterial appendages which help them adhere to the host surface (Proft & Baker, 2008). A number of fimbrial genes have been described in *Cronobacter* (Joseph et al. 2012b) and investigated in chapter 3 (section 3.4.2.1). The GR4 represents an additional fimbrial region. This region was detected uniformly in all *C. sakazakii* CC4 isolates, however only 3 of *C. sakazakii* non-CC4 isolates completely lacked this region. Since the region was detected in more than half of the non-CC4 isolates (n=12) it may not form the basis of differentiation between the two groups, moreover its absence in clinically significant *C. sakazakii* isolates 658 and 1218 (fatal meningitic isolates) further suggested that it may not be essential for the virulence potential of *C. sakazakii* CC4 in particular and *C. sakazakii* in general.

Similarly, the GR5 consisted of 15 genes which encoded for diverse range of proteins including transcriptional regulators, toxin/anti toxin system, however all of them showed a mosaic pattern of presence across the two groups and none of them indicated any *C. sakazakii* CC4 pattern.

Similarly, the GR6 consisted of nine genes encoding for the house keeping enzymes such as reductases, synthases, ligases, transketolase and a leucine efflux protein, however none of the gene was CC4 specific.

Of particular significance was the GR7, other than the toxin/anti-toxin, actetyltransferases, anti-restriction proteins, transcriptional regulators and recombinases none of which showed a CC4 specific pattern, the genes located between Csak377_03841-03850 were part the O-antigen specific genes (Figure 5.7, Table 5.7). It was interesting to note that all of the O-antigen associated genes were uniformly present in all CC4 isolates except isolate 1240. In addition, non-CC4 isolates 978 and 1249 also contained these O-antigen associated genes. The O-antigen region has been investigated in detail in chapter 3 (section 3.4.5) where it was observed that all of the CC4 isolates belonged to O:2 except isolate 1240, while non-CC4 isolates 978 and 1249 also had O:2 serotype. Thus O:2 is not the unique signature for *C. sakazakii* CC4 (section 3.4.5).

An interesting observation in the present comparative genomic analysis was the presence of a giant ~15 kb long gene present in all of *C. sakazakii* CC4 isolates (n=18), however present as a whole in only half (6 of 12) non-CC4 genomes. Further investigation of this after dividing into smaller ORFs revealed that most were the homologues of either T1SS or adhesion associated genes. Adhesins are important towards the bacterial pathogenicity as they help a bacterium colonise the host surface (Soto and Hultgren, 1999). The gene was also noted in half of the non-CC4 isolates (6/12). However, this gene could be important towards the virulence of *C. sakazakii* in general and hence further investigation at the laboratory level is warranted to elucidate its importance in *C. sakazakii*.

The analysis presented in this chapter was aimed for the identification of any potential virulence or environmental fitness associated trait in *C. sakazakii* CC4 using a gene by gene comparison of *C. sakazakii* CC4 and non-CC4 isolates. A total of 465 genes were identified unique in the genome of *C. sakazakii* CC4 isolate 377 against comparison with *C. sakazakii* non-CC4 isolate ES15. After omitting the phages and hypothetical traits, the remaining functional genes divided into 7 groups, could not form the basis of differentiation between CC4 and non-CC4. The GR7 which contained the O-antigen associated genes further strengthened the *in silico* serotyping presented in section 3.4.5. In addition a giant adhesion associated gene was noted in all *C. sakazakii* CC4 isolates, however due to its presence in

50% of the *C. sakazakii* non-CC4 isolates, it cannot be classed as CC4 specific trait. Although some interesting variations were observed, yet the present genomic analysis could not find any gene with functional annotation uniquely specific to *C. sakazakii* CC4 which could clearly explain its predominance in neonatal meningitis and its persistence in the environment especially in PIF and its associated production environment. The genomic analysis was carried further by fragmented genomic comparison using a program called Gegenees which can identify any unique genes in a target group. The analysis is presented in the following chapter 6.

It is important to point out that although the hypothetical proteins and phage associated traits were omitted from this analysis, yet there were two hypothetical proteins CSAK377_02657 (homologue of CSAK557_03870 in CC4 isolate 557 and CSSP291_18210 in CC4 isolate SP291) and Csak377_03526 (homologue of CSSP291_10445 in *C. sakazakii* CC4 isolate SP291) which were found specific to *C. sakazakii* CC4 (Table S4). These hypothetical proteins have been discussed in more detail in chapter 6.

6 CHAPTER 6

FRAGMENTED COMPARATIVE GENOME ANALYSIS OF *C. SAKAZAKII* STRAINS USING GEGENEES TO IDENTIFY UNIQUE TRAITS IN *C. SAKAZAKII* CC4

6.1 BACKGROUND

Due to the predominant association of *C. sakazakii* CC4 with neonatal meningitis and its relatively frequent isolation from PIF and the environment (Joseph & Forsythe, 2011; Hariri et al. 2013; Muller et al. 2013; Sonbol et al. 2013; Forsythe et al. 2014); it was of significance to identify any unique traits in this lineage. The analysis presented in chapter 3 aimed to identify any potentially virulence and environmental fitness associated traits previously described in *Cronobacter* or closely related organisms revealed some interesting variations, however could not identify any trait strictly specific to *C. sakazakii* CC4. The analysis was carried forward in chapter 4 by investigating the core genome of *C. sakazakii* CC4, however except for one hypothetical protein (CSSP291_18210), none of the traits could directly be linked with CC4 clonality and its virulence potential or environmental fitness. Whole genomic comparison using Artemis Comparison Tool (ACT) through gene by gene search approach identified genomic regions which showed variable degree of presence across *C. sakazakii* CC4 and non-CC4 genomes; although some interesting variations such as the giant adhesion associated gene was observed, nevertheless no other trait could explain the clonality, virulence potential or environmental fitness of the CC4 (chapter 5).

The genomic analysis presented in this chapter was aimed to identify the unique traits in *C. sakazakii* CC4 using a program called Gegenees (Agren et al. 2012). Gegenees is a program which fragments the genomes and compares each fragment against all the fragmented genomes used in comparison. It also allows a target group to be selected and to identify genomic fragments specific to that target group. The program has been used for the identification of unique traits in strains from the 2011 *Escherichia coli* O104:H4 outbreak, in the *Bacillus cereus* group and in the foot and mouth disease viruses (Agren et al. 2012). In this analysis, this program was used to analyse 30 *C. sakazakii* genomes keeping 18 CC4 genomes as our target group.

6.2 HYPOTHESIS

C. sakazakii CC4 isolates have unique virulence or environmental fitness associated traits.

6.3 AIMS OF THE CHAPTER

The aims of the analysis presented in this chapter were;

- To undertake fragmented genomic analysis by Gegenees using 18 *C. sakazakii* CC4 genomes as the target group to identify CC4 specific traits.

- To undertake fragmented analysis by Gegenees keeping 12 non-CC4 genomes as the target group to identify any traits missing uniformly in CC4 genomes.

- Validate the findings of Gegenees by *Cronobacter* BLAST and ACT.

6.4 METHODS

6.4.1 STRAINS, SEQUENCING, GENOME ASSEMBLY AND ANNOTATION

A total of 30 *C. sakazakii* isolates which included 18 CC4 and 12 non-CC4 were used in this genomic analysis (Table 3.1). The 22 genomes were sequenced at University of Exeter using Illumina Hiseq2500 while additional genomes were accessed from Genbank (Table 3.1). The genome assembly was using Velvet (Zerbino et al. 2008). The genome annotation was undertaken using PROKKA (Seemann, 2014) (section 2.2.5).

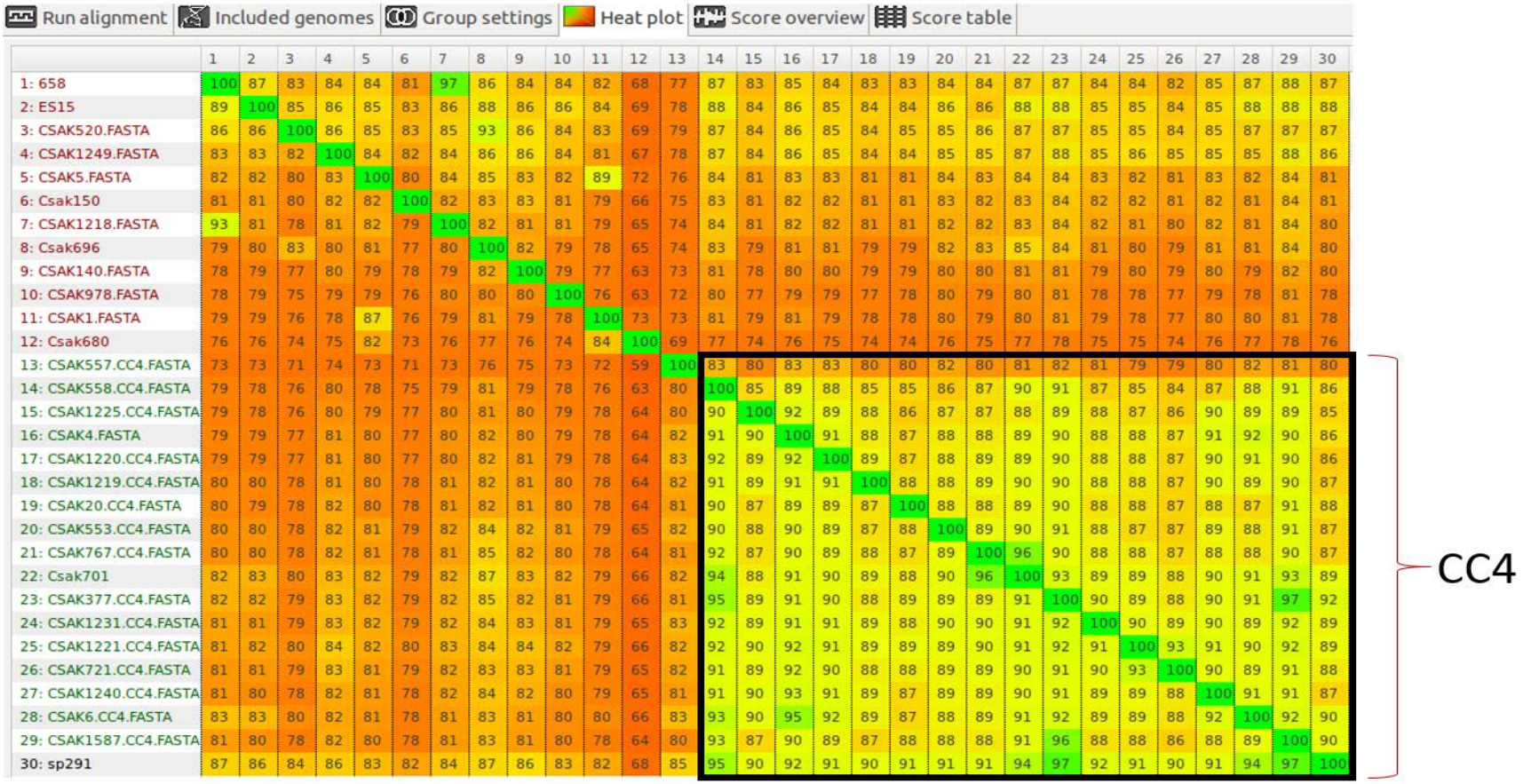
6.4.2 FRAGMENTED GENOMIC ANALYSIS USING GEGENEES

The finished *C. sakazakii* CC4 genome *C. sakazakii* SP291 was used as the reference strain. Fragmented all against all comparison was performed by setting fragment size to 500bp and sliding step size as 500bp to ensure there are no overlaps. The search method used was BLASTN (section 2.2.10). The input files for each genome were in FASTA format. Automated fragmentation by Gegenees fragmented each genome into fragments of 500bp (section 2.2.10). Initially 18 *C. sakazakii* CC4 genomes were used as the target group to identify the biomarker or unique traits in CC4, in this case *C. sakazakii* CC4 genome SP291 was used as the reference, hence all the genomic coordinates (or annotation) of the unique traits identified would refer to *C. sakazakii* SP291. The unique traits in *C. sakazakii* non-CC4 isolates refer to the missing traits in *C. sakazakii* CC4. When identifying missing genomic fragments in *C. sakazakii* CC4, the genomes of 12 non-CC4 isolates were used as the target group, in this case the genome of *C. sakazakii* 658 was used as the reference strain, hence the genomic coordinates (or gene annotation) of all the missing genomic fragments in CC4 (or unique traits in non-CC4) would refer to the genome of *C. sakazakii* 658.

6.5 RESULTS AND DISCUSSION

6.5.1 FRAGMENTED GEGENEES ANALYSIS TO IDENTIFY UNIQUE TRAITS IN *C. SAKAZAKII* CC4

The fragmented all against all genomic analysis of 30 *C. sakazakii* genomes was undertaken using Gegenees. To identify unique genomic fragments (genes), 18 *C. sakazakii* CC4 genomes were used as the target group (section 2.2.10). Interestingly, as with the whole genome alignment (Figure 3.3) the Gegenees alignment of 30 *C. sakazakii* genomes also revealed close clustering of the CC4 genomes as shown by the heat map (Figure 6.1) further validating the clonal nature of the CC4 genomes. The fragmented all against all genomic comparison of 30 *C. sakazakii* genomes by Gegenees revealed a unique signature for the CC4 genomes (Figure 6.1).

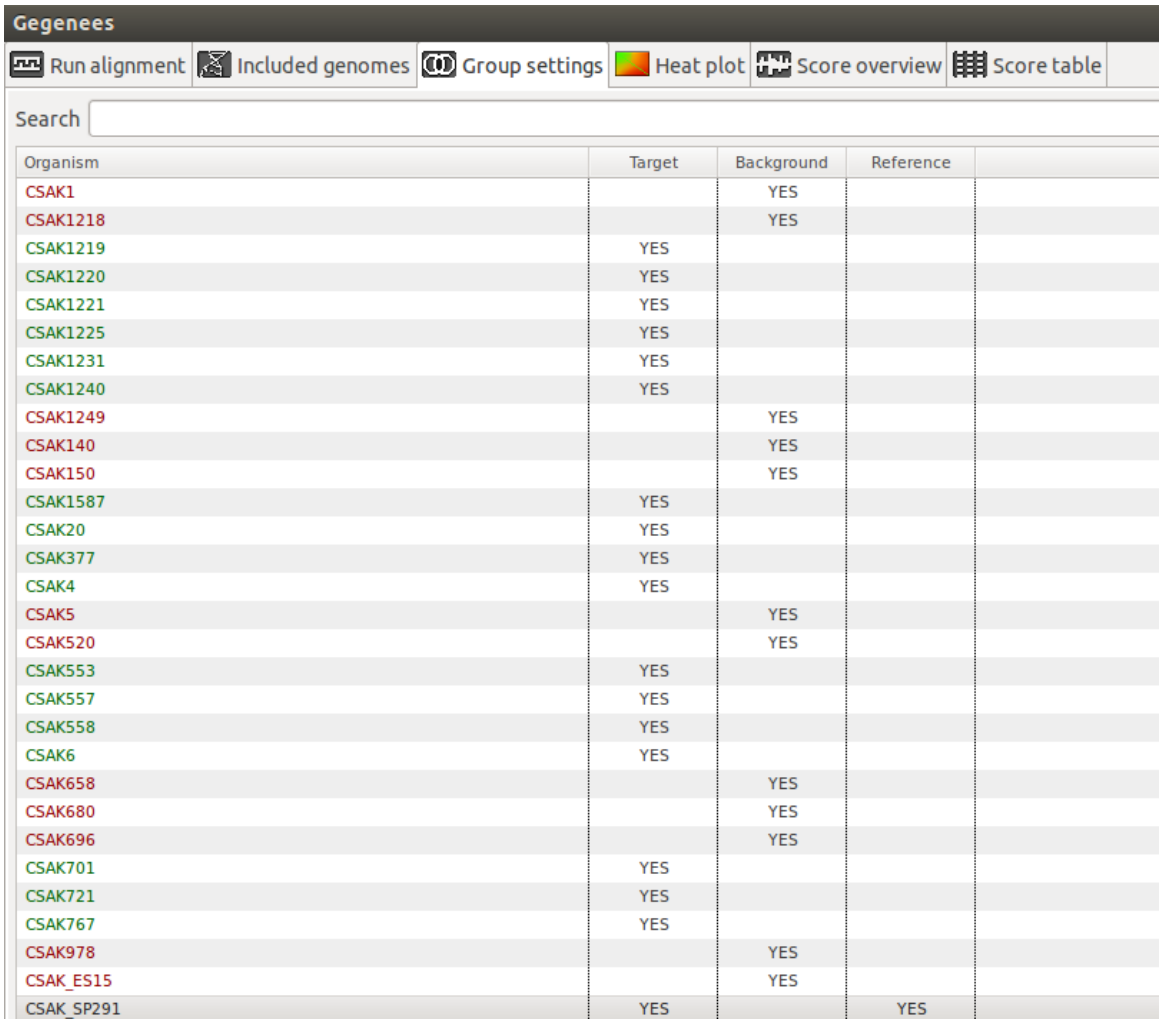


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1: 658	100	87	83	84	84	81	97	86	84	84	82	68	77	87	83	85	84	83	83	84	84	87	87	84	84	82	85	87	88	87
2: ES15	89	100	85	86	85	83	86	88	86	86	84	69	78	88	84	86	85	84	84	86	86	88	88	85	85	84	85	88	88	88
3: CSAK520.FASTA	86	86	100	86	85	83	85	93	86	84	83	69	79	87	84	86	85	84	85	85	86	87	87	85	85	84	85	87	87	87
4: CSAK1249.FASTA	83	83	82	100	84	82	84	86	86	84	81	67	78	87	84	86	85	84	84	85	85	87	88	85	86	85	85	85	88	86
5: CSAK5.FASTA	82	82	80	83	100	80	84	85	83	82	89	72	76	84	81	83	83	81	81	84	83	84	84	83	82	81	83	82	84	81
6: Csak150	81	81	80	82	82	100	82	83	83	81	79	66	75	83	81	82	82	81	81	83	82	83	84	82	82	81	82	81	84	81
7: CSAK1218.FASTA	93	81	78	81	82	79	100	82	81	81	79	65	74	84	81	82	82	81	81	82	82	83	84	82	81	80	82	81	84	80
8: Csak696	79	80	83	80	81	77	80	100	82	79	78	65	74	83	79	81	81	79	79	82	83	85	84	81	80	79	81	81	84	80
9: CSAK140.FASTA	78	79	77	80	79	78	79	82	100	79	77	63	73	81	78	80	80	79	79	80	80	81	81	79	80	79	80	79	82	80
10: CSAK978.FASTA	78	79	75	79	79	76	80	80	80	100	76	63	72	80	77	79	79	77	78	80	79	80	81	78	78	77	79	78	81	78
11: CSAK1.FASTA	79	79	76	78	87	76	79	81	79	78	100	73	73	81	79	81	79	78	78	80	79	80	81	79	78	77	80	80	81	78
12: Csak680	76	76	74	75	82	73	76	77	76	74	84	100	69	77	74	76	75	74	74	76	75	77	78	75	75	74	76	77	78	76
13: CSAK557.CC4.FASTA	73	73	71	74	73	71	73	76	75	73	72	59	100	83	80	83	83	80	80	82	80	81	82	81	79	79	80	82	81	80
14: CSAK558.CC4.FASTA	79	78	76	80	78	75	79	81	79	78	76	63	80	100	85	89	88	85	85	86	87	90	91	87	85	84	87	88	91	86
15: CSAK1225.CC4.FASTA	79	78	76	80	79	77	80	81	80	79	78	64	80	90	100	92	89	88	86	87	87	88	89	88	87	86	90	89	89	85
16: CSAK4.FASTA	79	79	77	81	80	77	80	82	80	79	78	64	82	91	90	100	91	88	87	88	88	89	90	88	88	87	91	92	90	86
17: CSAK1220.CC4.FASTA	79	79	77	81	80	77	80	82	81	79	78	64	83	92	89	92	100	89	87	88	89	89	90	88	88	87	90	91	90	86
18: CSAK1219.CC4.FASTA	80	80	78	81	80	78	81	82	81	80	78	64	82	91	89	91	91	100	88	88	89	90	90	88	88	87	90	89	90	87
19: CSAK20.CC4.FASTA	80	79	78	82	80	78	81	82	81	80	78	64	81	90	87	89	89	87	100	88	88	89	90	88	88	87	88	87	91	88
20: CSAK553.CC4.FASTA	80	80	78	82	81	79	82	84	82	81	79	65	82	90	88	90	89	87	88	100	89	90	91	88	87	87	89	88	91	87
21: CSAK767.CC4.FASTA	80	80	78	82	81	78	81	85	82	80	78	64	81	92	87	90	89	88	87	89	100	96	90	88	88	87	88	88	90	87
22: Csak701	82	83	80	83	82	79	82	87	83	82	79	66	82	94	88	91	90	89	88	90	96	100	93	89	89	88	90	91	93	89
23: CSAK377.CC4.FASTA	82	82	79	83	82	79	82	85	82	81	79	66	81	95	89	91	90	88	89	89	89	91	100	90	89	88	90	91	97	92
24: CSAK1231.CC4.FASTA	81	81	79	83	82	79	82	84	83	81	79	65	83	92	89	91	91	89	88	90	90	91	92	100	90	89	90	89	92	89
25: CSAK1221.CC4.FASTA	81	82	80	84	82	80	83	84	84	82	79	66	82	92	90	92	91	89	89	89	90	91	92	91	100	93	91	90	92	89
26: CSAK721.CC4.FASTA	81	81	79	83	81	79	82	83	83	81	79	65	82	91	89	92	90	88	88	89	89	90	91	90	93	100	90	89	91	88
27: CSAK1240.CC4.FASTA	81	80	78	82	81	78	82	84	82	80	79	65	81	91	90	93	91	89	87	89	89	90	91	89	89	88	100	91	91	87
28: CSAK6.CC4.FASTA	83	83	80	82	81	78	81	83	81	80	80	66	83	93	90	95	92	89	87	88	89	91	92	89	89	88	92	100	92	90
29: CSAK1587.CC4.FASTA	81	80	78	82	80	78	81	83	81	80	78	64	80	93	87	90	89	87	88	88	88	91	96	88	88	86	88	89	100	90
30: sp291	87	86	84	86	83	82	84	87	86	83	82	68	85	95	90	92	91	90	91	91	91	94	97	92	91	90	91	94	97	100

Figure 6. 1 Heat map of similarity indices of 30 *C. sakazakii* genomes.

The heat plot of 30 *C. sakazakii* genomes based on BLASTN alignment of 500/500 is shown. 18 CC4 genomes (green font) were used as the target group, all the remaining genomes represented background group (red font). No threshold was used in order to screen both conserved and non-conserved genomic fragments for the CC4 unique genomic fragments.

In Gegenees analysis, a biomarker (max/min) score value of 1.0 in the target group refers to the unique genomic fragment or gene in the that target group (Agren et al. 2012). In the present analysis, 18 *C. sakazakii* CC4 genomes were used as the target group while the genome of *C. sakazakii* SP291 was used as the reference genome (Figure 6.2).



Organism	Target	Background	Reference
CSAK1		YES	
CSAK1218		YES	
CSAK1219	YES		
CSAK1220	YES		
CSAK1221	YES		
CSAK1225	YES		
CSAK1231	YES		
CSAK1240	YES		
CSAK1249		YES	
CSAK140		YES	
CSAK150		YES	
CSAK1587	YES		
CSAK20	YES		
CSAK377	YES		
CSAK4	YES		
CSAK5		YES	
CSAK520		YES	
CSAK553	YES		
CSAK557	YES		
CSAK558	YES		
CSAK6	YES		
CSAK658		YES	
CSAK680		YES	
CSAK696		YES	
CSAK701	YES		
CSAK721	YES		
CSAK767	YES		
CSAK978		YES	
CSAK_ES15		YES	
CSAK_SP291	YES		YES

Figure 6. 2 The target group settings in Gegenees.

The image above is the cropped section of the screen shot indicating the total number of genomes included in the fragmented Gegenees analysis and the target group (green font colour) and background group (red font colour) settings. The black font colour shows the reference genome in the target group.

As a result of the fragmented all against all genomic comparison, regions with highest biomarker score i.e. 1.0 were found on the reference *C. sakazakii* SP291 genome demonstrating that these regions are unique to the target CC4 group (Figure 6.3).

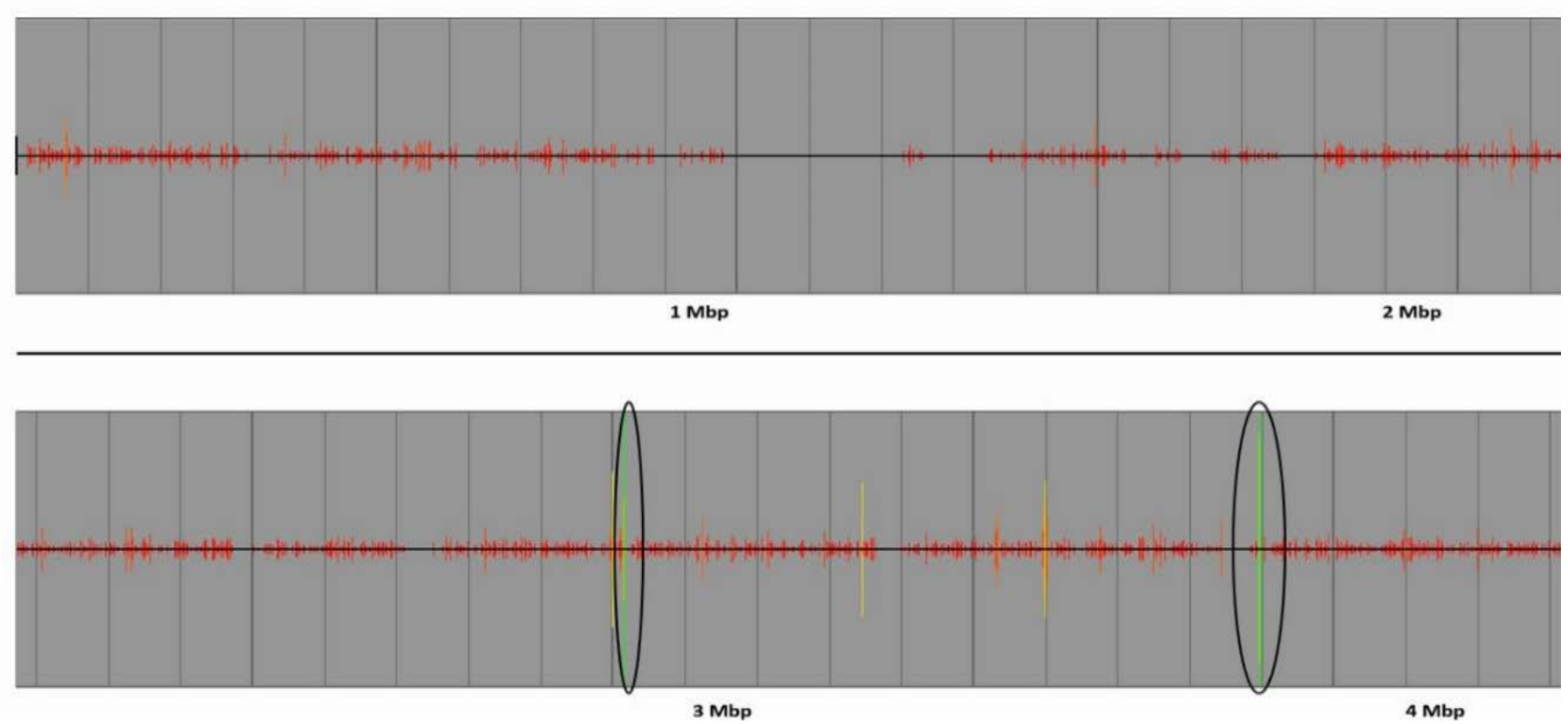


Figure 6. 3 Unique regions in the *C. sakazakii* CC4 genomes.

The figure shows CC4 unique regions (green, black encircled) identified by BLASTN alignment of 30 *C. sakazakii* genomes. The target group consisted of 18 CC4 genomes while remaining 12 non-CC4 genomes represented the background group. The genomic positions are with reference to the *C. sakazakii* SP291.

The investigation of these genomic regions was undertaken using their coordinates in the reference *C. sakazakii* SP291. These genomic regions represented 2 genes with locus_tag CSSP291_18210 and CSSP291_10445, both of which were annotated as hypothetical proteins (Table 6.1).

Table 6. 1 Unique genes identified in *C. sakazakii* CC4 genomes (n=18) by Gegenees fragmented genomic analysis

Locus_tag	Annotation	Blastx
CSSP291_18210	Hypothetical protein	Inner membrane protein of *Enterobacter aerogenes* (accession: YP_007387546) 73% (65/89)
CSSP291_10445	Hypothetical protein	Hypothetical protein COCC4DRAFT_56662 [*Bipolaris maydis* ATCC 48331 30/91(33%)

In order to investigate further and to predict the potential function of these hypothetical proteins, homology searches were performed in NCBI using BLASTx. The hypothetical protein CSSP291_18210 showed partial homology (identities = 73% (65/89) and 85% positives (76/89) was against a putative inner membrane protein of *Enterobacter aerogenes* (accession: YP_007387546). Another hit with less homology (identities = 46% (57/87) and 62% positives (54/87) was against the "chain A, crystal structure of the type VI effector-immunity complex Ssp1-rap1a from *Serratia marcescens*" (accession; 4BI8_A) (46% identity). This observation was quite interesting as the hypothetical protein CSSP291_18210 turned out to be the same as CSAK557_03870 which has been indicated as enriched in CC4 in chapter 4 (section 4.4.3.5). The hypothetical protein has been discussed in detail in section 4.4.3.5. In brief, this hypothetical protein was initially noticed only in *C. sakazakii* CC4 isolates (n=18) while absent from all *C. sakazakii* non-CC4 isolates (n=12). The flanking regions revealed only hypothetical proteins while the secondary structure *in silico* protein indicated several protein binding sites (Figures 4.5-6). For more detail on this hypothetical protein please see section 4.4.3.5.

An additional finding with Gegenees analysis was the identification of the hypothetical protein CSSP291_10445 as enriched in CC4. The BLASTx could not predict any functional relevance for this hypothetical protein as most of the hits with considerable degree of homology indicated only hypothetical proteins of unknown function. The location of the hypothetical protein CSSP291_10445 was also taken into consideration. Using Artemis the flanking regions of these CC4 unique hypothetical proteins was investigated (Figure 6.4).

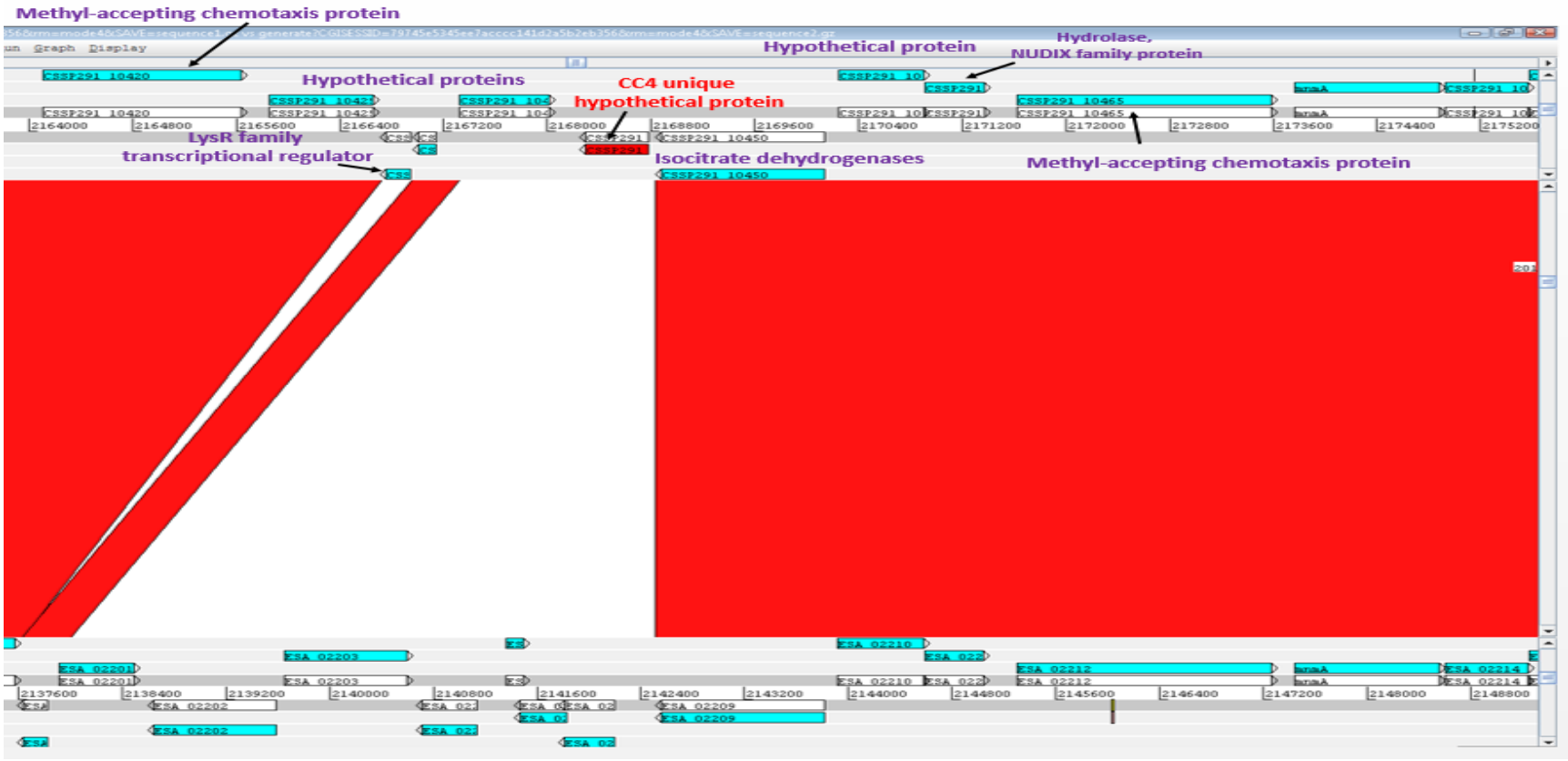


Figure 6. 4 ACT comparison of the flanking region of the hypothetical protein CSSP291_10445 in one *C. sakazakii* CC4 SP291 and *C. sakazakii* non-CC4 658.

The image above indicates the ACT comparison of the flanking region of the hypothetical protein CSSP291_10445 unique in 18 *C. sakazakii* CC4 genomes as identified by Gegenees fragmented genome comparison analysis. The flanking region was noticed in non-CC4 genomes as well.

The hypothetical protein CSSP291_10445 was flanked by housekeeping traits which included such as isocitrate dehydrognese, LysR family transcriptional regulator protein, hydrolase enzyme and other hypothetical proteins; methyl accepting chemotaxis proteins. However the flanking region was also noticed in non-CC4 genome 658 (Figure 6.4), it was therefore difficult to predict the any functional relevance for the unique hypothetical protein CSSP291_10445.

An attempt was also made to predict the functional and structural properties of the hypothetical protein CSSP291_10445 using online protein prediction tool http://ppopen.informatik.tu-muenchen.de/ (Bigelow et al. 2004; Goldberg et al. 2012; Ofran & Rost, 2007; Hamp et al. 2013). No functional relevance could be predicted, however the location of this hypothetical protein was predicted to be the inner membrane protein. Furthermore, a secondary structure was also predicted with several protein binding sites on the surface of hypothetical protein CSSP291_10445 (Figure 6.5). The structural characteristic such as binding sites is an important feature which may mediate some virulence or fitness associated mechanism, however only full characterisation of this hypothetical protein can unravel its functional attributes and its significance in CC4.

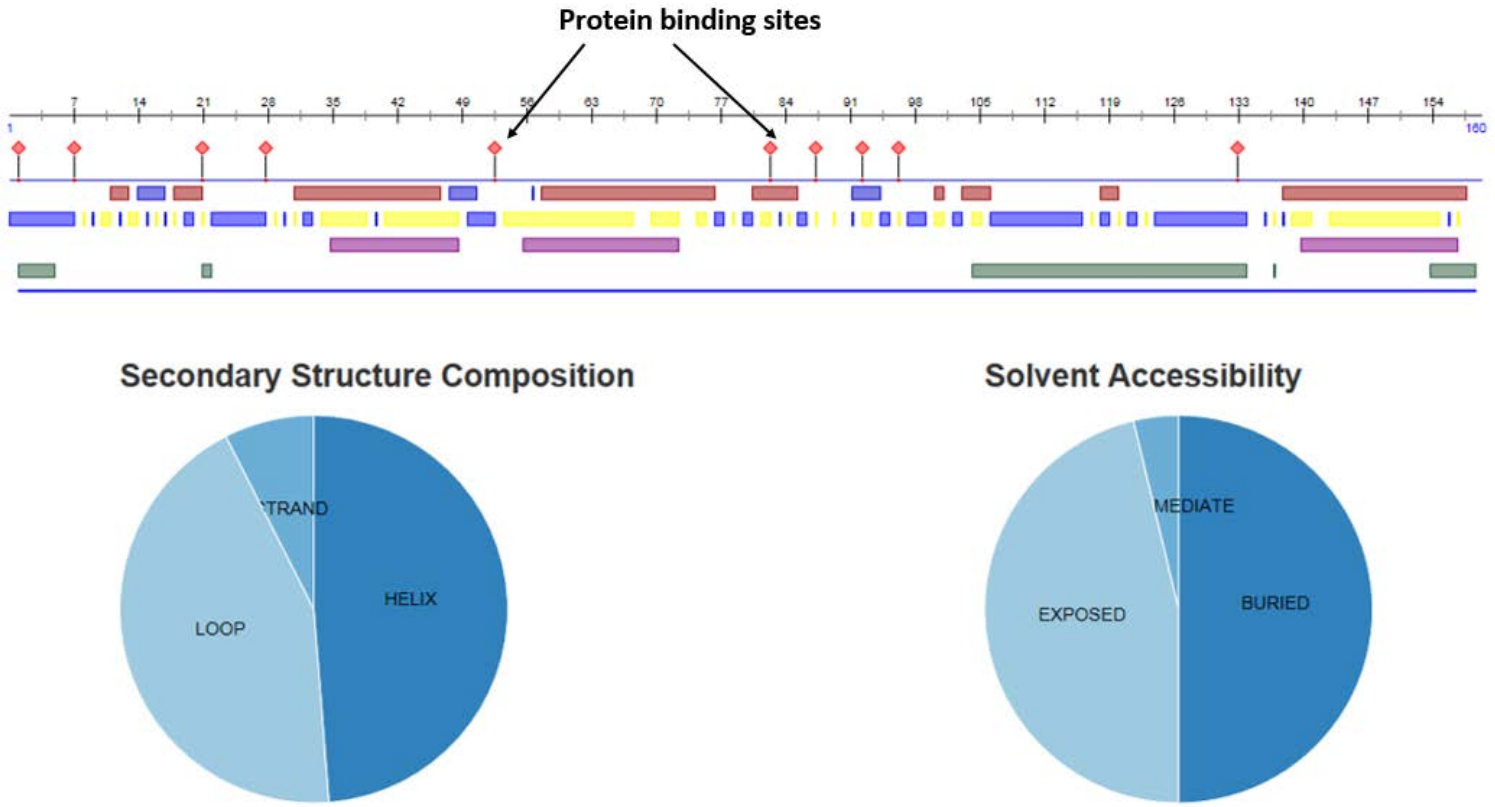


Figure 6. 5 Secondary structure prediction for the hypothetical protein CSSP291_10445.

The figure above represents the secondary structure prediction of CSSP291_10445 using online protein prediction tool http://ppopen.informatik.tu-muenchen.de/. The red squares with stalks represent the protein binding sites while the yellow circle with a stalk shows a polynucleotide binding region. The red, blue and yellow rectangles represent the helices, exposed and buried parts of the protein regarding solvent accessibility. The composition of the secondary structure and its solvent accessibility is represented by the pie-chart.

As mentioned elsewhere in this thesis, the PubMLST website for *Cronobacter* also facilitates the genomic comparison of all publically available *Cronobacter* genomes using BLAST facility. A total of 107 *Cronobacter* spp. genomes (last accessed; 20th February 2015) are available publically in the BIGSdb facility in *Cronobacter* PubMLST. Of these, 37 were CC4 genomes, all of which were *C. sakazakii* strains. Additional 70 genomes represent non-CC4 isolates representing all 7 species of the genus *Cronobacter* (http://pubmlst.org/perl/bigsdb/bigsdb.pl?db=pubmlst_*cronobacter*_isolates&page=plugin&name=BLAST/). The opportunity was exploited in order to observe the variation of the hypothetical protein CSSP291_10445 at a larger scale using *Cronobacter* BLAST. The BLAST results against 107 sequenced *Cronobacter* spp. genomes indicated that the hypothetical protein CSSP291_10445 was noted in 100% of the CC4 genomes (n=37) while it was detected in merely 5 out of 70 non-CC4 isolates which included 693, 700, 713, 714 and 715 all of which were *C. sakazakii* belonging to CC13. Hence, the hypothetical protein CSSP291_10445 varied significantly between CC4 and non-CC4 isolates ($P < 0.05$). Please note that these genomes represent the 1994 French outbreak isolates which have been investigated in chapter 8. The observation is also of significance as all of the CC13 isolates in *Cronobacter* PubMLST were *C. sakazakii*; moreover almost half of the CC13 strains (12/26) were food isolates while ~36% (9/26) represented infant formula isolates. Whether, the hypothetical protein CSSP291_10445 is also a dominant trait in CC13 cannot be predicted with certainty as the genome sequences of only 5 of the 26 CC13 isolates in *Cronobacter* PubMLST are available at the time of writing this section of the thesis. Future studies are warranted to find the association of the hypothetical protein CSSP291_10445 with CC13.

6.5.2 FRAGMENTED GEGENEES ANALYSIS TO IDENTIFY MISSING GENES IN *C. SAKAZAKII* CC4

Bacterial genomes constantly undergo gene gain and loss, a process called genome flux (Lawrence 1999; den Bakker et al. 2010). The loss of genes can sometimes enhance the virulence of a bacterium, for example the loss of *ompT* and *cadA* from *Shigella* where the presence of these genes can attenuate the virulence (Nakata et al. 1993; Maurelli et al. 1998). Moreover the loss of function through deletion of certain genes can enhance bacterial adaptation to a certain habitat (Hottes et al. 2013). Since, *C. sakazakii* CC4 have frequently been associated with neonatal infections and isolated frequently from PIF, therefore it was important to investigate if any of the genes have been lost specifically from *C. sakazakii* CC4 genomes. Thus, the missing genes in *C. sakazakii* CC4 were also identified using Gegenees. In this case the target group settings were changed and the 12 non-CC4 genomes (Table 3.1) were used as the target group while the non-CC4 isolate *C. sakazakii* 658 was used as the reference strain. The initial Gegenees analysis coupled with *Cronobacter* BLAST and manual genomic analysis by ACT identified only one hypothetical protein (ESA_02200) as specific to 12 *C. sakazakii* non-CC4 genomes while missing uniformly in 18 *C. sakazakii* CC4 genomes used in the present genomic analysis. The analysis was expanded by *Cronobacter* BLAST to 107 *Cronobacter* genomes including 37 CC4 and 70 *Cronobacter* spp. non-CC4 genomes. The results of the BLAST analysis indicated uniform absence of this hypothetical protein in 36 CC4 isolates while present only in one *C. sakazakii* CC4 isolate ES713; however the results also indicated the hypothetical protein ESA_02200 was not uniformly present in all 70 *Cronobacter* spp. non-CC4 isolates as it was found absent in 18 other *Cronobacter* spp. isolates as well (Table S5). Hence, although there was significant variation between CC4 and non-CC4 isolates ($P < 0.05$), the loss of ESA_02200 is not strictly specific to CC4. The BLASTx was performed to predict any functional relevance, however most of the hits with significant degree of homology only indicated uncharacterised hypothetical proteins. Furthermore, the analysis of the flanking region in the reference *C. sakazakii* BAA-894 could not indicate any functional relevance for this hypothetical protein as it was flanked by uncharacterised hypothetical, proteins both upstream and downstream (Figure 6.6). An attempt was also made to predict the function of this hypothetical protein using the *in silico* protein prediction tool http://ppopen.informatik.tu-muenchen.de/ (Bigelow et al. 2004; Goldberg et al. 2012;

Ofran & Rost, 2007; Hamp et al. 2013). However, no functional category and location of hypothetical protein ESA_02200 could be predicted. In future studies, characterisation of the hypothetical protein ESA_02200 may clarify its loss, predominantly from the CC4 isolates.

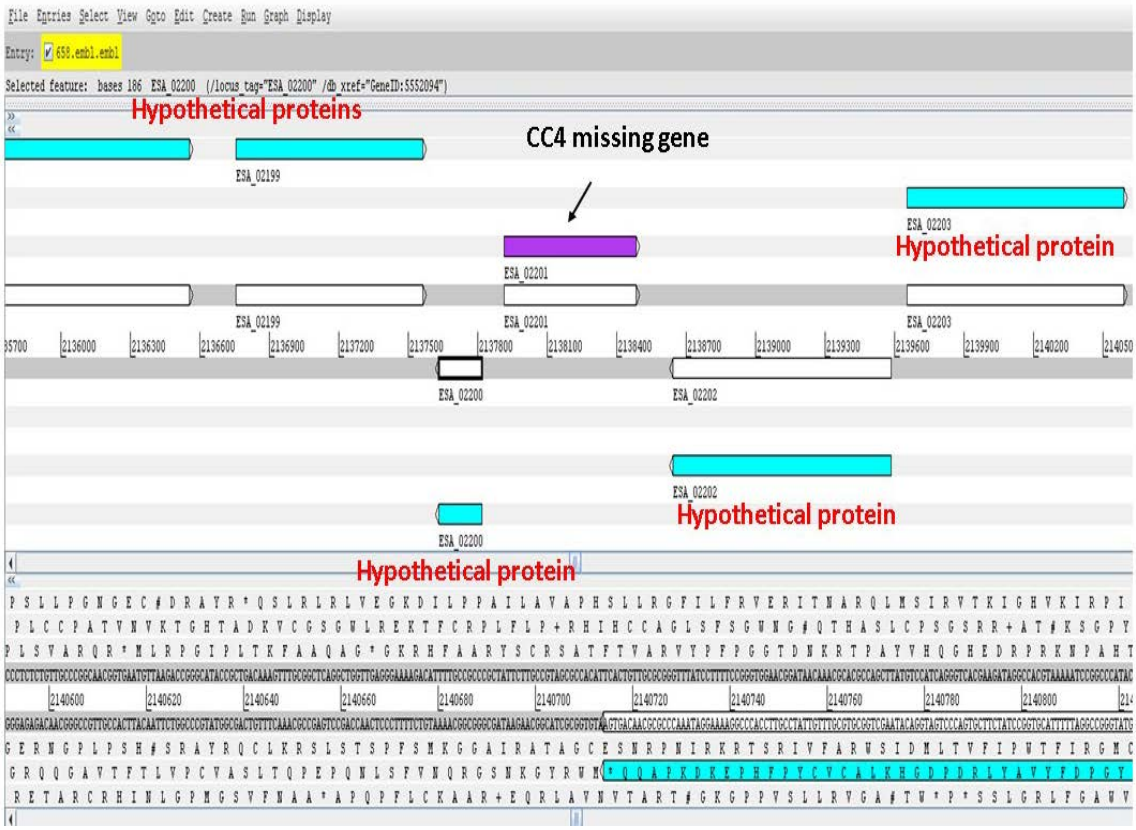


Figure 6. 6 Location of the CC4 missing gene

The figure indicates the location the gene encoding for the hypothetical protein (purple colour). found missing in *C. sakazakii* CC4 genomes (n=18) while present uniformly in non-CC4 genomes (n=12).

6.6 GENERAL DISCUSSION AND CONCLUSION

The predominance of *C. sakazakii* CC4 in neonatal meningitis and its frequent isolation from the PIF is of high concern to neonatal health. Although, interesting variations were observed in chapters 3-5, however no key virulence associated gene which could explain the association of CC4 with virulence and environmental fitness, could be identified. Recently a new genomic analysis program has been developed which can help identify unique genes within a target group, the program is called Gegenees and has been used for the identification of unique genes in the *Bacillus cereus group* and in the foot and mouth disease viruses (Agren et al. 2012). Hence, it was of particular interest to utilise this program with the aim to identify unique genes within *C. sakazakii* CC4. As with the previous analysis, the same set of 30 *C. sakazakii* genomes consisting of 18 CC4 and 12 non-CC4 genomes were used in the Gegenees analysis (Table 3.1). The aim was twofold, first to identify any unique genes within CC4 and secondly to identify any genes specifically lost from the CC4 genomes.

The fragmented analysis by Gegenees identified two hypothetical proteins CSSP291_18210 and CSSP291_10445 enriched in CC4 isolates. The hypothetical protein CSSP291_18210 was also identified as enriched in CC4, in genome analysis presented in chapter 4. The present analysis further strengthened the finding of the chapter 4. The hypothetical protein CSSP291_18210 has been discussed in detail in chapter 4; in brief it was an homologue of the inner membrane protein as revealed by BLASTx and was only detected in 2 of the non-CC4 isolates *C. sakazakii* 2051 and *C. malonaticus* (CC129). The *C. sakazakii* 2051 belongs to CC64 which has also been isolated frequently from the infant formula (section 4.4.4).

An additional finding in the present analysis was the hypothetical protein CSSP291_10445 which was found to be enriched in CC4 genomes. The expansion of the analysis to 107 *Cronobacter* spp. genomes indicated that besides 100% of the CC4 genomes, the hypothetical protein was also detected in five of the CC13 isolates; an important lineage associated with food and infant formula (section 6.4.1). The BLASTx analysis could not predict homologies against a protein with known function as most hits indicated hypothetical proteins. Further, characterisation using the *in silico* protein prediction (section 6.4.1) also could not assign any functional relevance however, the secondary structure predicted the presence of some protein binding sites on its surface. These binding sites may be crucial for the interaction of this hypothetical protein with other molecules.

However, only further characterisation of this hypothetical protein can elucidate its importance in CC4.

The loss of genes can also be important towards the virulence of bacterium or towards the adaptation of a particular habitat (Lawrence & Roth, 1999). Therefore, the Gegenees was also used to find out any genes which could have been lost specifically from CC4 genomes. In this case the 12 non-CC4 genomes were used as the target group where *C. sakazakii* 658 was used as the reference genome. Only one hypothetical protein ESA_02200 was identified as unique to non-CC4 genomes while absent uniformly from the CC4 genomes (except for the isolate ES713). The analysis was taken further to suggest any functional relevance of this hypothetical protein using BLASTx and analysing the flanking region of this hypothetical protein, however no clue was found since the BLASTx results returned only hypothetical proteins while the flanking regions also was consisted of hypothetical proteins (Figure 6.6). Furthermore, the *Cronobacter* BLAST indicated that the loss of this hypothetical protein is not exclusive to CC4 as 18 other *Cronobacter spp*. non-CC4 genomes also lack this hypothetical proteins in their genomes (Table S5). The detailed laboratory based studies dedicated at the characterisation of these hypothetical proteins and to assay their effect on the virulence or fitness of *C. sakazakii* CC4 may elucidate their precise role.

Although no key virulence associated genes could be noted, the detailed comparative genomic analysis presented in chapters 3-6 aimed to identify *C. sakazakii* CC4 specific traits revealed interesting variations including two hypothetical proteins (CSSP291_18210, CSSP291_10445) as predominant in CC4 isolates, the O:2 to be the dominant serotype for *C. sakazakii* CC4 and a giant adhesion associated gene present dominantly in the *C. sakazakii* CC4 isolates. In addition, the hypothetical protein ESA_02200 was noted to be predominantly missing from the CC4 genomes.

It is important to mention that while undertaking fragmented genome analysis, contrasts other than CC4 against non-CC4 were also considered. Fragmented genome analysis was also performed by trying clinical, CSF and meningitic isolates as target groups, however no unique genes were detected in any of the target group (data not shown).

The focus of the analysis was now moved to the second and third main aims of this thesis which were to study the diversity within *C. sakazakii* CC4 isolates (presented in the chapter 7) and the genomic analysis with primary aim of source identification of *C. sakazakii* strains

isolated from the 1994 French outbreak of neonatal intensive care unit (NICU) which led to the deaths of 3 babies (presented in chapter 8).

7 CHAPTER 7

DIVERSITY WITHIN *CRONOBACTER SAKAZAKII* CLONAL COMPLEX 4

7.1 BACKGROUND

Previous and current PhD studies at NTU have elaborated that there are differences amongst CC4 isolates in their clinical presentations, for example variation in the ability of CC4 isolates regarding adhesion and invasion was observed *in vitro* (Nasreddin Rhouma, 2012; Faisal Almajed 2014, unpublished data). Hence, it was of significance to analyse the diversity within *C. sakazakii* CC4 isolates using genome sequence data, which was the focus of the analysis presented in this chapter. With the revolution in whole genome sequencing technologies, the detection of Single Nucleotide Polymorphism (SNP) amongst different bacterial isolates of the same species has become relatively easier and can therefore be used as a genomic tool to study intra-species variation or in other words to analyse the diversity within closely related bacterial isolates. This approach was exploited in the analysis presented in this study. The analysis presented in this chapter was based primarily on the SNP analysis and further comparative genomic studies within *C. sakazakii* CC4 isolates were undertaken to answer the questions arising from the SNP analysis. The sequenced *C. sakazakii* genomes analysed in this chapter are also available on *Cronobacter* BIGSdb (Jolley & Maiden, 2010). The genome comparator facility in the *Cronobacter* BIGSdb facilitates the comparative genomics of these isolates using the COG-cgMLST (Forsythe et al. 2014), therefore this platform provided an opportunity to correlate the findings with SNP analysis. The COG-cgMLST analysis of *C. sakazakii* CC4 isolates was based on the sequence backbone of *C. sakazakii* SP291 (Power et al. 2013).

7.2 HYPOTHESIS

C. sakazakii CC4 is genetically a clonal lineage.

7.3 AIMS OF THE CHAPTER

The main aim of the analysis presented in this chapter was to study the diversity within *C. sakazakii* CC4 isolates. The analysis was undertaken with the following objectives.

- SNP calling of *C. sakazakii* CC4 isolates using finished *C. sakazakii* CC4 isolate SP291 as an index strain.
- Manual filtration of the SNPs to retain only high quality SNPs.
- Phylogeny estimation of the CC4 isolates using the filtered variant call format (VCF) files.
- Variation analysis of the isolates as indicated by the SNP phylogeny.
- Comparative genomics of high and low invasive *C. sakazakii* CC4 strains.
- Heavy metal resistance assay of the selected *C. sakazakii* high and low invasive isolates.
- The COG-cgMLST analysis of *C. sakazakii* CC4 isolates using *C. sakazakii* SP291 as the reference backbone.

7.4 METHODOLOGY

7.4.1 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS OF *C. SAKAZAKII* CC4 GENOMES

Single Nucleotide Polymorphism (SNP) analysis was performed on the *C. sakazakii* CC4 isolates (Table 7.1) using SMALT and SAMtools (Li et al. 2009). The FASTA file of the published genome of the CC4 strain *C. sakazakii* SP291 was used as the reference or index. The resulting VCF files were filtered using VCFTools to include only SNPs with minimum quality score of 30, minimum depth of 8, and minimum allele frequency of 0.75 (Sahl et al. 2011; Clark et al. 2012; McNally et al. 2013). A consensus sequence file was then created for each genome and used to create a maximum likelihood phylogeny by Dr Alan McNally using his scripts. The method is described in more detail in section 2.2.9. The resulting tree was visualised and annotated using the FigTree (http://beast.bio.ed.ac.uk/figtree).

7.4.2 COMPARATIVE GENOME ANALYSIS

The comparative genome analysis of *C. sakazakii* CC4 isolates was undertaken using Artemis Comparison Tool (ACT). The methods are described in more detail in sections 2.2.6 and 2.2.7. The genome comparator facility of *Cronobacter* PubMLST was also used.

7.4.3 METAL RESISTANCE ASSAY

Selected *C. sakazakii* isolates were tested for their ability to tolerate metals toxic at higher concentrations. Isolates were tested for 8 different metals which included copper (II) sulphate (Sigma-Aldrich®, UK), sodium arsenate (Sigma-Aldrich®, UK), nickel chloride (Fisher Scientific, UK), silver nitrate (Alfa Aesar®, UK), zinc sulphate (BDH chemicals England), sodium tellurite (Sigma-Aldrich®, UK), cadmium carbonate (Harrington, USA) and cobalt (II) nitrate (Sigma-Aldrich®, UK). Each of the isolate was tested for 4 different concentrations of 1M, 0.1M, 0.01M and 0.001M of each heavy metal. The method is described in detail in section 2.1.6.

7.5 RESULTS AND DISCUSSION

7.5.1 SNP ANALYSIS OF *C. SAKAZAKII* CC4 STRAINS

The SNP analysis of *C. sakazakii* CC4 genomes was undertaken using SMALT and filtered manually (section 2.2.9). The reference genome used was *C. sakazakii* SP291. The SNP phylogenetic tree was viewed and annotated using FigTree. The SNP phylogenetic tree of *C. sakazakii* CC4 (Figure 7.1) indicated that most of the *C. sakazakii* CC4 (n=13) strains were equidistant to the reference genome, however four of the CC4 strains; 6, 558, 1587 and 1240 revealed variations on the SNP tree; these isolates showed greater number of SNPs compared to other isolates against the reference isolate SP291 (Table 7.1). The strain 1240 is notable as it is serotype O:4 unlike other CC4 isolates which are O:2 (section 3.4.4).

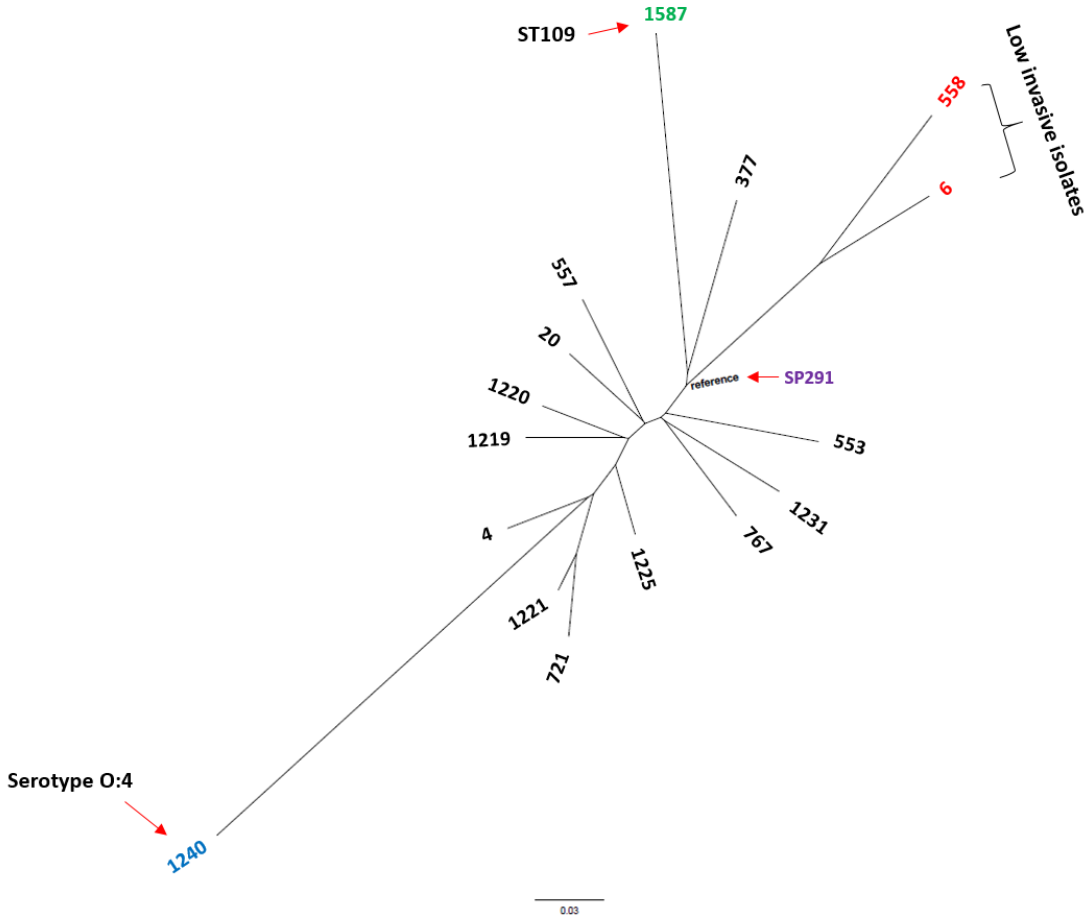


Figure 7. 1 The SNP based phylogeny of 18 CC4 strains.

The SNPS were called using SMALT and SAMtools to generate the VCF files which were then merged and the maximum likelihood phylogeny was estimated. Colour coding: purple = reference isolate; red = low invasive strains; blue = O:4; green = ST109. The scale bar shows the number of nucleotide substitutions. (Comment; Please note that at the time the SNP analysis was initiated the raw sequencing files (FASTQ) of the CC4 isolate 701 were not available and therefore could not be included in the SNP phylogeny).

The number of SNPs against the reference strain varied from as low as only 16 to as high as 1245 SNPs. Most of the CC4 isolate (n=10) had SNPs in the range of 300 to 400 (Table 7.1). Considering the temporal and geographic diversity, this observation further supports the genomic stability of the CC4 isolates.

Table 7. 1 SNP calls in *C. sakazakii* CC4 isolates against the reference *C. sakazakii* CC4 isolate SP291

***C. sakazakii* CC4 isolate**	**No of SNPs**	**Year of isolation**	**Country**	**Source**
377	**16**	1950	UK	Clinical
553	**335**	1977	Netherlands	Clinical
557	**307**	1979	Netherlands	Clinical; Faeces
6	**668**	1990	Canada	Clinical
558	**785**	1983	Netherlands	Milk powder
4	**403**	1990	Canada	Clinical
767	**294**	1994	France	Clinical; Peritoneal fluid
1587	**554**	2000	Israel	Clinical; CSF
20	**239**	2003	Czech Republic	Clinical; Trachea
721	**414**	2003	USA	Clinical
1220	**280**	2003	USA	Clinical; CSF
1221	**360**	2003	USA	Clinical; CSF
1231	**343**	2005	New Zealand	Clinical; Blood
1225	**312**	2007	USA	Clinical; Faeces
1219	**306**	2009	USA	Clinical; CSF
1240	**1245**	2009	USA	Clinical; CSF

CSF; Cerebrospinal fluid

It was interesting to note that the earliest CC4 isolate 377 which was isolated in 1950 had only 16 SNPs against *C. sakazakii* CC4 isolate SP291. Unfortunately, the date of isolation for the reference strain SP291 is not known. The analysis of these 16 SNPs indicated, 9 of these SNPs were located in a single gene CSSP291_02330 encoding for a putative integrase, of which only 1 SNP was non-synonymous while the remaining 8 SNPs were synonymous (Table S6). Integrases are the enzymes that catalyse the integration of the viruses such as bacteriophages into the host chromosomal DNA, found ubiquitously in bacteria and are considered important for the horizontal gene transfer (Zaneveld et al. 2008). The presence of SNPs in a potentially phage acquired are unlikely to be significant, moreover these genomic areas also represent low sequence coverage areas. Three of the SNPs were located in a gene CSSP291_t20316 potentially encoding for a transfer RNA, however both of these SNPs were synonymous. The remaining 4 SNPs were all intergenic and synonymous. Hence,

the SNP difference between the reference isolate SP291 and the earliest isolate 377 indicate the close relatedness of these 2 isolates. It is important to consider the source of isolation for both these isolates as strain 377 was isolated from the milk powder in the UK while SP291 was isolated from a PIF manufacturing plant in Republic of Ireland, nevertheless both of these isolates were obtained from two neighbouring countries: UK and Ireland. Low number of SNP differences between these isolates suggest a possible common origin.

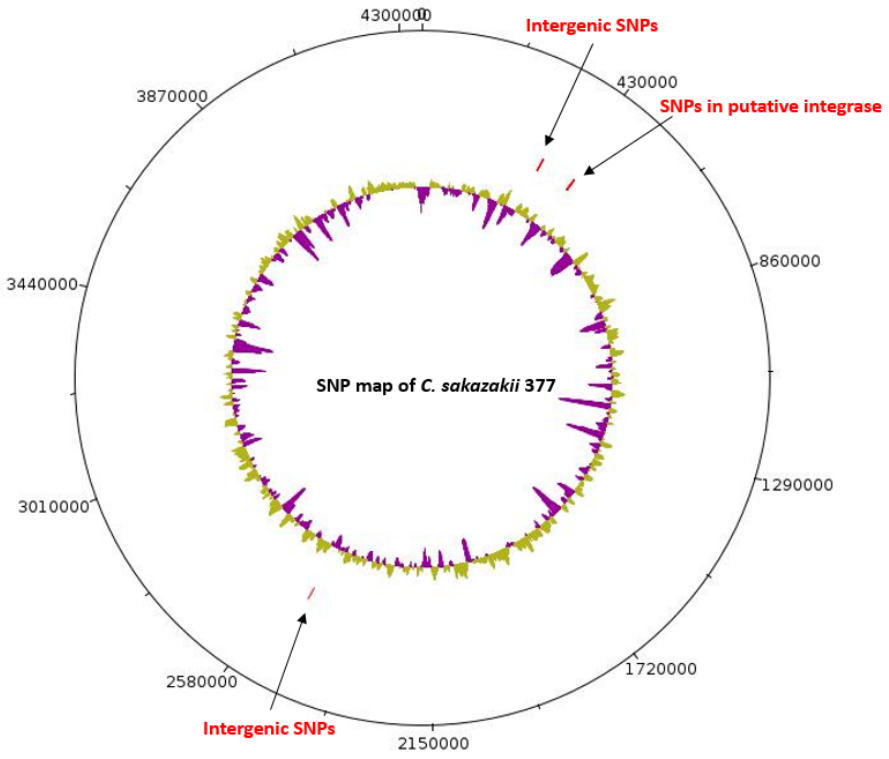


Figure 7. 2 SNP map of *C. sakazakii* 377.

The figure represents the location of SNPs on *C. sakazakii* 377. The location of the SNPs is with reference to *C. sakazakii* genome SP291. The outermost black ring with tick marks represents the *C. sakazakii* SP291 genome. The inner circle with red marks shows the location of SNPs while innermost circle with purple (above average) and green (below average) spikes indicate GC contents in a particular location of the genome. The image was created using the DNA plotter feature of Artemis. The arrows indicate the location and annotation (red fonts) of the SNPs in isolate 377.

The SNP phylogeny showed that although most of the CC4 isolates clustered close to each other, some differences were observed. Four of the CC4 strains; 6, 558, 1587 and 1240 revealed variations on the SNP tree (Figure 7.1). Of these strains, *C. sakazakii* 6 and 558 clustering together but relatively distant to other CC4 isolates, have previously been identified as low invasive on human brain microvascular endothelial cell lines (HBMEC) *in*

vitro (Faisal Almajed 2014; unpublished data). Further analysis was undertaken to study the variation in these 4 isolates.

7.5.1.1 SNP VARIATION IN *C. SAKAZAKII* 1240

A relatively large number of SNPs were detected in *C. sakazakii* CC4 isolate 1240 against the reference strain SP291 compared to other CC4 isolates. A total of 1245 SNPs were detected in CC4 isolate 1240; the SNP map indicated a high concentration of SNPs located between region 860000bp and 1290000bp (Figure 7.3).

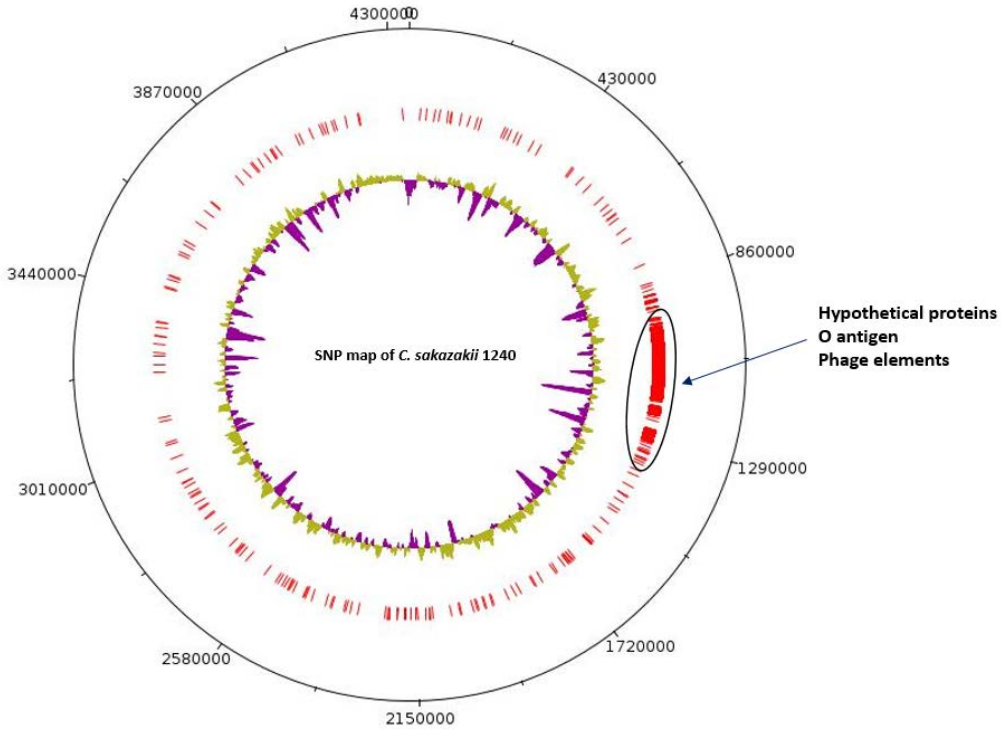


Figure 7. 3 SNP map of *C. sakazakii* 1240.

The figure represents the location of SNPs on *C. sakazakii* 1240. The location of the SNPs is with reference to *C. sakazakii* genome SP291. The outermost black ring with tick marks represents the *C. sakazakii* SP291 genome. The inner circle with red marks shows the location of SNPs while innermost circle with purple (above average) and green (below average) spikes indicate GC contents in a particular location of the genome. The image was created using the DNA plotter feature of Artemis. The black encircled is the genomic region having high concentration of SNPs (>1000) containing hypothetical proteins, some phage and O-antigen associated traits.

Whether it is due to recombination, warrants further investigation. The analysis of the filtered VCF file indicated over 1000 SNPs in this region alone. The analysis of the genomic region indicated over 400 genes in this region, the majority of which were hypothetical proteins of unknown function and some phage associated traits. A notable observation was

the O-antigen associated genes located between CSSP291_05550 and CSSP291_05665. Previous *in silico* analysis indicated that 1240 is different from other CC4 isolates that it has an O:4 serotype, whereas the remaining strains are O:2 (section 3.4.4). The analysis indicates that O-antigen as well as hypothetical and page associated traits are the basis of longer branch length and hence variation of 1240 on the SNP tree. Further investigation of the hypothetical proteins may provide better insight into their significance in isolate 1240.

7.5.1.2 SNP VARIATION IN *C. SAKAZAKII* 1587

The *C. sakazakii* CC4 isolate 1587 had a total of 554 SNPs against the reference *C. sakazakii* CC4 isolate SP291, however like *C. sakazakii* 1240, this isolate also had a greater branch length on the SNP tree (Figure 7.1). The analysis of the VCF files after applying the filtering parameters revealed that the genomic region between region 320012bp to 351260bp had high density of SNPs in isolate 1587; this region alone contained 397 SNPs out of total 554 SNPs. A SNP map of the isolate 1587 was also drawn using DNA plotter (Figure 7.4).

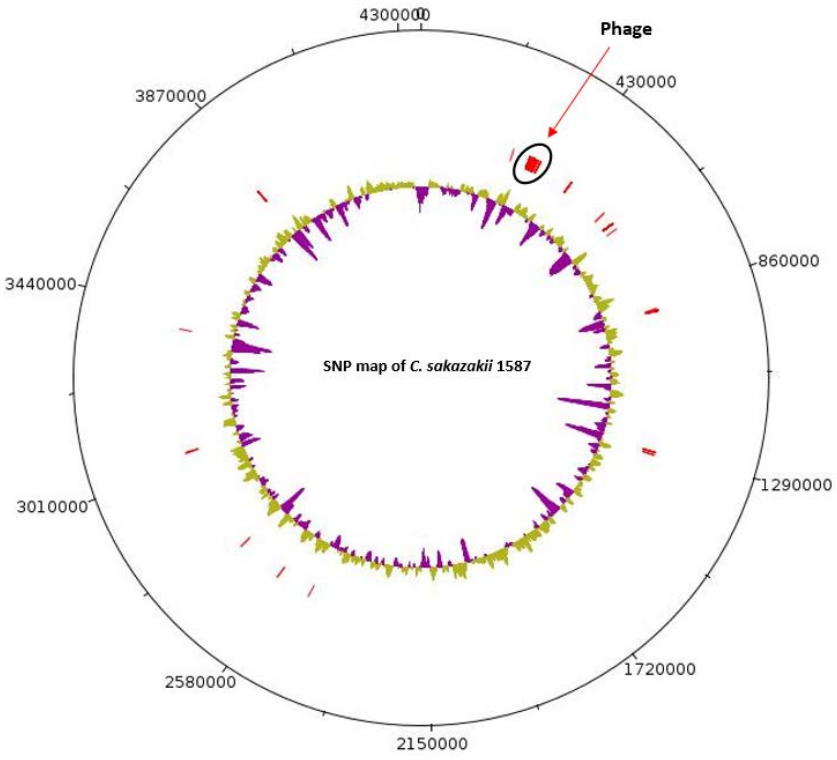


Figure 7. 4 SNP map of *C. sakazakii* 1587.

The figure represents the location of SNPs on *C. sakazakii* 1587. The location of the SNPs is with reference to *C. sakazakii* genome SP291. The outermost black ring with tick marks represents the *C. sakazakii* SP291 genome. The inner circle with red marks shows the location of SNPs while innermost circle with purple (above average) and green (below average) spikes indicate GC contents in a particular location of the genome. The image was created using the DNA plotter feature of Artemis. The black encircled is the genomic region having high concentration of SNPs (397) containing only phage associated traits.

The black encircled region indicates the high density SNP region in 1587. The analysis of this genomic region revealed only phage associated traits. Hence, the major variation between CC4 isolate 1587 from other CC4 isolates was likely due to this strain specific phage acquisition which explains its longer branch length on the SNP tree (Figure 7.1).

7.5.1.3 SNP VARIATION IN LOW INVASIVE *C. SAKAZAKII* ISOLATES 6 AND 558

Low invasive *C. sakazakii* isolates 6 and 558 clustered closely on the SNP tree, having 668 and 785 SNPs respectively (Table 7.1, Figure 7.1). The analysis of the SNPs in low invasive *C. sakazakii* isolates indicated that both of these isolates had high SNP density between genomic positions 430000 and 1290000 on the reference genome (Figure 7.5). This region had approximately 600 SNPs distributed in both isolates, the region consisted of wide variety of genomic traits such as large number of hypothetical proteins, housekeeping genes, transcriptional regulators and some phage associated traits. Due to time limitation, studying each of these SNPs and gene affected by that SNP was beyond the scope of the present study. However, the analysis of the SNP calls indicated that both of these isolates shared a common SNP (G to C; CAG to CAC) at position 83510 (with reference to *C. sakazakii* SP291 genome). When the reference genome was viewed in Artemis, the position referred to a hypothetical protein CSSP291_04255. This was a non-synonymous SNP which changed the amino acid asparagine (CAG) to histidine (CAC). The flanking region of CSSP291_04255 consisted of hypothetical proteins of unknown function. The BLASTx analysis of CSSP291_04255 indicated a phage protein in *Escherichia coli* UMEA 3108-1 (accession; EQW65158). Therefore, the “common” SNP in isolates 6 and 558 did not reveal any functional relevance.

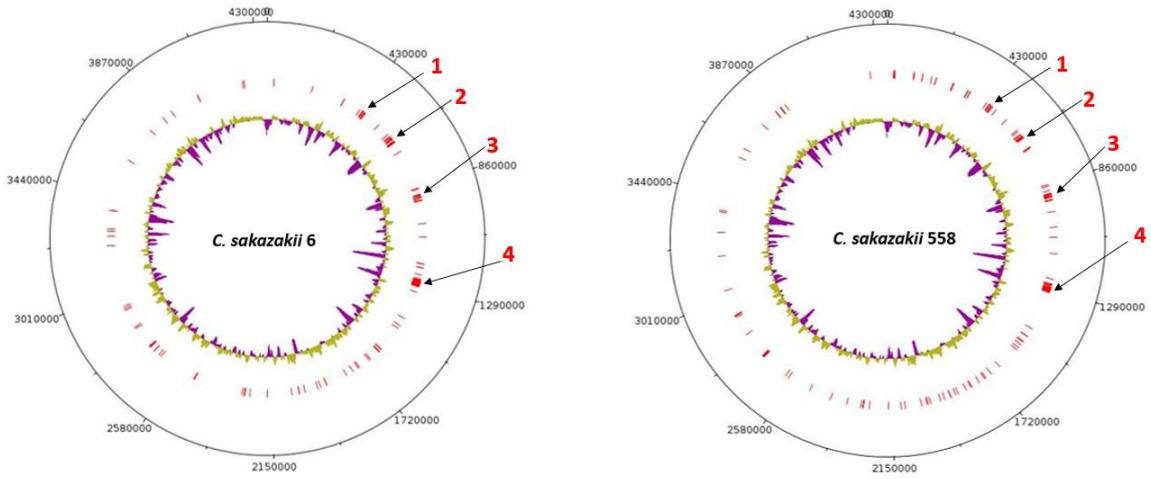


Figure 7. 5 SNP mapping of low invasive *C. sakazakii* isolates.

The figure represents the position of SNPs on *C. sakazakii* isolates 6 and 558. The location of the SNPs is with reference to *C. sakazakii* genome SP291. The outermost black ring with tick marks represents the *C. sakazakii* SP291 genome. The arrows marked as 1, 2, 3 and 4 represent the SNP dense areas common between isolates 6 and 558. The inner circle with red marks shows the location of SNPs while innermost circle with purple (above average) and green (below average) spikes indicate GC contents in a particular location of the genome. The image was created using the DNA plotter feature of Artemis.

However, it was important to know if there are any differences in the genomic contents of high and low invasive *C. sakazakii* isolates. It was therefore decided to undertake comparative genomic study by choosing the genomes of representative high and low invasive CC4 isolates with the hope to identify potential invasion associated traits in high invasive isolates as well as unique traits in low invasive isolates.

7.5.2 COMPARATIVE GENOMICS OF HIGH AND LOW INVASIVE *C. SAKAZAKII* CC4 ISOLATES

The relatively distant location of the low invasive isolates indicated that there might be some unique regions within these isolates making them distant from other CC4 isolates. Fragmented genomic analysis was undertaken using two CC4 isolates 6 and 558 as the target group, however no traits were identified which were unique in these two low invasive CC4 isolates (data not shown). It was therefore decided to carry the genomic analysis further by comparing a high invasive *C. sakazakii* CC4 isolate against a low invasive CC4 isolate; the comparison was aimed to identify any potential invasion associated trait in high invasive CC4 isolate as well as to identify any traits unique to low invasive CC4 isolate. For

this purpose high invasive *C. sakazakii* CC4 isolate 767 was chosen to compare against the low invasive CC4 isolate 6 by a manual gene by gene search approach using ACT. Using a gene by gene search approach, the genes unique in high invasive isolate 767 as well as low invasive isolate 6 were listed (Table S7).

7.5.2.1 UNIQUE TRAITS IN HIGH INVASIVE *C. SAKAZAKII* ISOLATE 767

A total of 369 genes were found as unique to *C. sakazakii* isolate 767 in comparison with the low invasive isolate 6 (Table S7). Majority of the unique traits in high invasive *C. sakazakii* isolate 767 were dominated by hypothetical proteins and phage associated traits (241/369). The 128 remaining genes included a wide variety of traits such as multidrug-efflux transporter MexB, a number of conjugal transfer genes, transcriptional regulators, serine transporters, aerobactin and iron acquisition genes (Table S7). Of particular interest was CSAK767_00886 annotated as invasion protein IagB, however the BLASTx analysis of this genes suggested it to be a homologue of the lytic transglycosylase in *E. coli* (accession; WP_024186589). The lytic transglycosylases are found ubiquitously in Gram negative bacteria and have been associated with the transport of macromolecules such as type 3 secretion system and type IV pilus synthesis (Koraimann, 2003; Scheurwater et al. 2008). None of the traits identified unique to the high invasive strain 767 could directly be linked with its invasion associated capability.

7.5.2.2 UNIQUE TRAITS IN LOW INVASIVE *C. SAKAZAKII* ISOLATE 6

More traits were identified to be unique to the low invasive strain 6 compared to high invasive strain 767. A total of 324 genes were identified to be unique to strain 6 compared to 767 as shown in the supplementary table (Table S7). Again, a large number of these genes (168/324) were hypothetical proteins and phage associated traits. Transposases and integrases were found to be interspersed among these regions. Other traits included membrane proteins, a number of conjugation elements, transcriptional regulators, toxin, antitoxin system, type IV biogenesis and secretory proteins. Of particular interest were traits associated with the “heavy metal” resistance and transport which included arsenic, copper, manganese, chromate and nickel. Metals such as nickel, arsenic, chromium and manganese at higher concentrations are generally toxic to the bacterial cells although certain metals are essential for the activities of enzymes (Nies, 1999). It has been suggested that the term “heavy metal” is both meaningless and misleading, as it is a general tendency

to regard all the heavy metals as toxic or ecotoxic, however it is not supported by the facts (Duffus, 2002). Therefore, the heavy metals, in this thesis are referred either as metals or toxic metals (since these metals are toxic to bacteria at higher concentrations).

Each of the metal resistance region in strain 6 is discussed below.

7.5.2.2.1 ARSENIC RESISTANCE

Arsenic is a toxic element which is released into the environment either by anthropogenic activities or through natural activities such as weathering and volcanic activities (Cullen and Reimer, 1989). Although arsenic is very toxic to microorganism, however due to its abundance in environment, certain bacterial species have developed resistance mechanism against arsenic compounds (Mukhopadhyay et al. 2002; Hobman, J. L., & Crossman, L. (2014). Several arsenic resistance associated, *ars* genes were detected as unique in low invasive *C. sakazakii* isolate 6 containing eight arsenic resistance associated genes with locus_tag CSAK6_00140-47 encoding for arsenical resistance operon trans-acting repressor *arsD_1*, arsenical pump-driving ATPase *arsA_1*, phosphinothricin N-acetyltransferase, arsenate reductase *arsC_1*, arsenic efflux pump protein *arsB_1*, arsenical pump-driving ATPase *arsA*_2, arsenical resistance operon trans-acting repressor *arsD*_2 and HTH-type transcriptional repressor *aseR*. However, it must be noted that the homologues of phosphinothricin N-acetyltransferase, arsenate reductase *arsC_1*, arsenic efflux pump protein *arsB_1* were also detected in high invasive CC4 isolate 767.

The arsenical compounds may contaminate several environments such as fresh water, ground water and sea water. However organisms such as bacteria, fungi and algae have developed mechanisms to survive in environments contaminated with arsenic. These mechanisms can include arsenite oxidation which is mediated by *aox* genes, respiratory arsenate reduction which is conferred by *arr* genes or resistance to arsenic through arsenic extrusion for cytoplasm defence mediated by *ars* genes (Branco et al. 2008). The arsenic resistance (*ars*) genes are found both in Gram negative and Gram positive bacteria such as *E. coli*, *Staphylococcus* species and *Pseudomonas aeruginosa* (Silver et al. 1981; Rosenstein et al. 1992; Carlin et al. 1995; Cai et al. 1998). The *ArsA* is part of the arsenic resistance operon which acts by forming a transmembrane complex with *ArsB* (Dey, 1997). The *ArsR* is a regulatory protein which controls the expression of the *ars* operon (Xu et al. 1996). The

function of the *ArsH* has not yet been fully described, however the *ArsH* was not found to be essential for the resistance against arsenic in *E. coli* (Branco, et al. 2008).

In order to test whether there is any difference between the high and low invasive CC4 isolates in their ability to resist arsenic, 10 different *C. sakazakii* isolates were tested including low invasive isolates 6 and 558 as well as high invasive isolates 767 and 20 (Faisal Almajed 2014, unpublished data). The strains were subjected to 4 different concentrations; 1M, 0.1M, 0.01M and 0.001M of sodium arsenate (section 2.1.6). The results indicated all isolates were sensitive to 1M sodium arsenate while all of the isolates were able to resist up to 0.01M concentrations of this heavy metal. The interesting observation, however was the low invasive isolate 6 which was able to resist 0.1M concentration of sodium arsenate. This laboratory observation is in agreement with the genomic analysis. However it must be noted that the other low invasive isolate 558 was not able to tolerate 0.1M concentration of sodium arsenate. Moreover, the results indicated the ability of all isolates irrespective of their invasion capability or their clonal complex to tolerate relatively low concentrations of sodium arsenate up to 0.01M, indicating that arsenic resistance may not be a unique mechanism to low invasive *C. sakazakii* CC4 isolates. However further investigation may clarify the link between invasion and heavy metal resistance such as arsenic in *C. sakazakii.*

Table 7. 2 Disc diffusion assay to compare toxic metals sensitivity in *C. sakazakii*

C. sakazakii strain	ST/CC	Invasion	Sodium Arsenate				Copper (II) sulphate				Silver nitrate				Sodium tellurite			
			1M	0.1	0.01	0.001	1M	0.1	0.01	0.001	1M	0.1	0.01	0.001	1M	0.1	0.01	0.001
767	4/4	High invasive*	35	19	0	0	12	8	0	0	10	13	7	0	45	35	24	8
20	4/4	High invasive	40	26	0	0	13	5	0	0	11	14	9	0	45	38	18	0
6	4/4	Non-invasive	20	0	0	0	10	6	0	0	12	15	8	0	35	25	9	0
558	4/4	Low invasive	31	15	0	0	13	0	0	0	8	11	8	0	43	40	30	22
377	4/4		35	15	0	0	11	6	0	0	13	12	15	0	45	40	32	14
140	40/45	Low invasive	35	23	0	0	10	5	0	0	7	13	14	10	35	25	9	0
150	16/16	Low invasive	26	11	0	0	13	5	0	0	12	10	14	0	45	32	30	15
978	3/3		31	14	0	0	10	0	0	0	10	11	10	0	50	35	10	8
658	1/1	Low invasive	28	10	0	0	13	6	0	0	10	11	10	0	42	35	15	0
520	12	Invasive	40	28	0	0	10	0	0	0	10	15	7	0	40	32	14	12

Table 7. 3 continued; Disc diffusion assay to compare toxic metals sensitivity in *C. sakazakii*

C. sakazakii strain	ST/CC	Invasion	Cobalt (II) nitrate				Nickel chloride				Zinc sulphate				Cadmium carbonate			
			1M	0.1	0.01	0.001	1M	0.1	0.01	0.001	1M	0.1	0.01	0.001	1M	0.1	0.01	0.001
767	4/4	High invasive*	21	8	0	0	14	0	0	0	22	9	0	0	0	0	0	0
20	4/4	High invasive	22	10	0	0	17	0	0	0	15	9	0	0	0	0	0	0
6	4/4	Non-invasive	20	0	0	0	9	9	0	0	15	0	0	0	0	0	0	0
558	4/4	Low invasive	17	7	0	0	18	0	0	0	15	7	0	0	0	0	0	0
377	4/4		22	13	0	0	15	0	0	0	14	7	0	0	0	0	0	0
140	40/45	Low invasive	23	7	0	0	15	0	0	0	14	0	0	0	0	0	0	0
150	16/16	Low invasive	18	7	0	0	14	0	0	0	15	7	0	0	0	0	0	0
978	3/3		23	8	0	0	18	0	0	0	15	7	0	0	0	0	0	0
658	1/1	Low invasive	20	8	0	0	17	0	0	0	15	7	0	0	0	0	0	0
520	12	Invasive	20	12	0	0	17	0	0	0	15	8	0	0	0	0	0	0

* The invasion assays were conducted by Faisal Almajed, 2014. The diameter for the zone of clearance was measured in mm.

7.5.2.2.2 NICKEL AND COBALT EFFLUX

A nickel/cobalt efflux system encoding gene, *rcnA* was also found to be unique in the genome of low invasive *C. sakazakii* isolate 6 (Table S7). Metals such as nickel and cobalt are required by bacteria in a variety of metabolic functions though in trace amounts; high concentrations of these metals can be toxic to bacteria. To overcome the high concentration of nickel and cobalt, one mechanism that bacteria have acquired is the efflux of these toxic metals from the cell (Mergeay et al. 2003; Nies, 2003). Rodrigue et al. (2005) showed that inactivation of the *yohM* gene induces sensitivity to nickel or cobalt to *E. coli* cells, hence they proposed the new de-nomination for the *yohM* gene that is *rcnA* (resistance to cobalt and nickel) (Rodrigue et al. 2005).

The laboratory assay was conducted to test the variability in nickel/cobalt resistance between the low and high invasive *C. sakazakii* isolates (Table 7.2). The results indicated that all of the isolates were able to resist nickel chloride at low concentrations of up to 0.1M. Similarly all the isolates were resistant to cobalt (II) nitrate at concentrations of up to 0.01M. At 0.1M, all the isolates were sensitive to cobalt (II) nitrate except for low invasive isolate 6. Interestingly, the low invasive isolate 6 showed variation at concentration of 0.1M; it was sensitive to nickel chloride while resistant to cobalt (II) nitrate at this concentration. The results here indicated that *rcnA* gene may not be the only gene responsible for nickel/cobalt resistance or efflux in *C. sakazakii* in general and low invasive isolates in particular as all the isolates showed some degree of resistance against these toxic metals .

7.5.2.2.3 COPPER AND SILVER RESISTANCE

Similarly, some copper and silver resistance and transport associated genes which included *copA_1* (copper resistance protein A precursor), *copB* (copper resistance protein B precursor), *pcoC* (copper resistance protein C precursor), *csoR* (copper-sensitive operon repressor) and *pcoE* (copper resistant protein PcoE) were detected unique to low invasive *C. sakazakii* isolate 6 (Table S7). The copper and silver resistance and transport genes have been investigated in section 3.4.2.9; the observation in this analysis is in agreement with previous analysis.

The laboratory assays indicated that all of the tested isolates were able to resist copper (II) sulphate at concentrations of up to 0.01M whereas at concentration of 0.1M, only three of

the isolates 558, 978 and 520 were resistant, it should be noted that the isolate 558 is a low invasive CC4 isolate. However, isolate 520 which was moderately invasive on HBMEC was also resistant to copper (II) sulphate at concentration of 0.1M. In addition, the isolate 978 was also resistant to 0.1M concentration of copper (II) sulphate. Similarly all the isolates were found to be sensitive to the silver nitrate at concentrations of 1M, 0.1M and 0.01M. At concentration of 0.001M all the strains were able to resist silver nitrate, the only exception was *C. sakazakii* isolate 140 which showed some sensitivity to silver nitrate at this concentration. Overall, all of the isolates showed at least some degree of sensitivity to copper (II) sulphate and silver nitrate and no clear pattern was observed for the copper and silver resistance which could form the basis of subdivision within *C. sakazakii* isolates.

7.5.2.2.4 IRON TRANSPORT

At least three genes associated with iron transport *FecA* (Iron(III) dicitrate transport protein), *fecC* (Iron(III) dicitrate transport system permease protein) and *fecD* (Iron(III) dicitrate transport system permease protein) were also noted as unique in the low invasive *C. sakazakii* isolate 6 (Table S7). A number of iron acquisition and siderophore associated genes have been analysed and discussed in section 3.4.2.10.

The comparative genomic analysis of high invasive *C. sakazakii* isolate 767 with low invasive *C. sakazakii* CC4 isolate 6 did not identify a key invasion associated gene which could directly be linked with invasion in high invasive isolates. However, a number of metal resistance and transport genes associated with copper, iron, nickel, arsenic, chromium were identified as unique to the low invasive *C. sakazakii* CC4 strain 6. Although laboratory test did not indicate a clear differentiation between high and low invasive isolates based on their metal tolerance ability (for tested metals) yet, the association of heavy metal resistance with invasion in *C. sakazakii* warrants further investigation.

7.5.3 The BIGsDB ANALYSIS OF *C. SAKAZAKII* CC4 ISOLATES

The genome sequences of *C. sakazakii* isolates analysed in the present study are now available at the *Cronobacter* PubMLST which facilitates the genome comparison using the *Cronobacter* BIGsDB (Bacterial Isolate Genome Sequence Database) (http://pubmlst.org/perl/bigsdb/bigsdb.pl?page=plugin&name=GenomeComparator&db=pubmlst_*cronobacter*_isolates) (Forsythe et al. 2014). Since the SNP analysis revealed genomic diversity within CC4 isolates based on the invasion phenotype, it was an opportunity to use this facility to study the diversity of CC4 genomes and to observe if there is any correlation between the findings of SNP analysis and BIGsDB comparative analysis of *Cronobacter* PubMLST. The comparative analysis was performed on 18 *C. sakazakii* CC4 isolates (n=18) using the genome of *C. sakazakii* CC4 isolate SP291 as the reference backbone (NC_020260.1). A total of 3917 coding sequences were analysed of which of 545 loci were exactly the same in all 18 CC4 isolates while 2682 loci showed allelic differences in 18 CC4 isolates. None of these 2682 variable loci had a unique pattern in low invasive isolates 6 and 558 as well as for the high invasive isolate 767 (data not shown). The neighbour-net joining tree did not indicate the clustering of low invasive isolates 6 and 558 based on the alignment of 3917 coding sequences (Figure 7.6). However, it was interesting to note the clustering of the reference *C. sakazakii* SP291 with the oldest CC4 isolate 377 and a fatal meningitic CC4 isolate 1587, despite their temporal as well as geographical diversity (Table 3.1). This is in agreement with the SNP phylogeny (Figure 7.1). Moreover, it further reinforces the close relatedness of the reference *C. sakazakii* CC4 isolate SP291 with the oldest CC4 isolate 377 (SNPs=16) as identified by SNP analysis (Table 7.1). The distance matrix of the 18 *C. sakazakii* CC4 isolates indicated that based on 3917 loci, the loci variation ranged from as low as 219 (1587 against 377) and as high as 2230 (701 against 557) (Figure 7.7). Again low loci variation was observed between the reference isolate SP291 and the oldest CC4 isolate 377, both of which differ from each other only by 263 loci. Please note that the loci variation refers to the allelic variation between the isolates and is based on the reference isolate SP291. Additionally a total of 33 paralogues were also identified across 18 CC4 isolates.

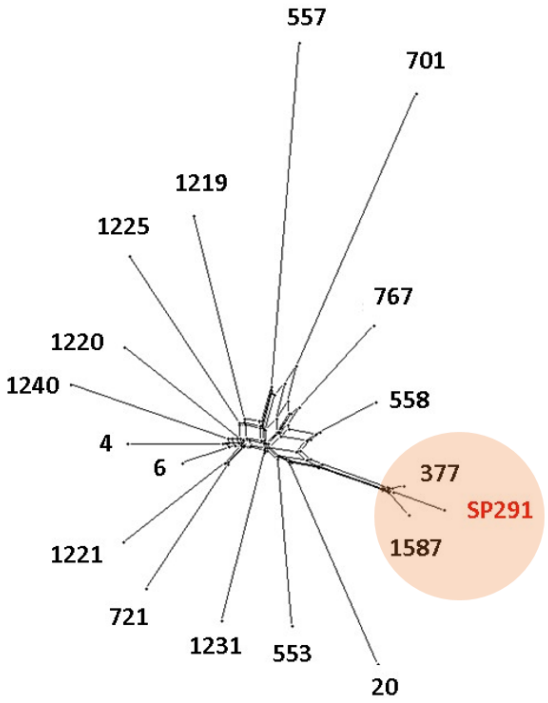


Figure 7. 6 The neighbour-net joining tree of *C. sakazakii* CC4 isolates (n = 18).

The tree indicates the variation amongst CC4 isolates based on 3917 coding sequences of the reference CC4 isolate SP291. The red font colour indicates the reference isolate. The orange highlighted circle indicates the clustering of the reference isolate with the oldest CC4 isolate 377 and 1587.

	ref	553	557	558	701	767	6	20	377	4	721	1219	1220	1221	1231	1240	1225	1587
ref	0																	
553	1250	0																
557	1939	1801	0															
558	864	1169	1819	0														
701	1989	1978	2230	1698	0													
767	1309	1271	1873	943	1719	0												
6	800	1105	1806	538	1746	1050	0											
20	1360	1430	1872	1403	2086	1482	1361	0										
377	263	1092	1828	701	1870	1182	630	1216	0									
4	1140	1152	1778	906	1856	1091	572	1354	986	0								
721	1320	1307	1822	1126	1905	1232	897	1474	1192	958	0							
1219	1649	1544	1957	1449	2219	1527	1259	1677	1535	1230	1385	0						
1220	1385	1327	1888	1163	2040	1260	945	1510	1258	985	1147	1420	0					
1221	1355	1354	1874	1157	2036	1284	912	1556	1223	942	946	1479	1191	0				
1231	1370	1316	1888	1186	2051	1300	1125	1542	1214	1125	1268	1575	1363	1351	0			
1240	1511	1443	1935	1261	2132	1411	948	1633	1348	968	1208	1532	1312	1266	1413	0		
1225	1576	1508	1926	1345	2057	1452	1098	1648	1426	1102	1301	1526	1350	1348	1453	1376	0	
1587	323	1168	1888	784	1912	1222	706	1268	219	1053	1213	1596	1301	1249	1309	1422	1491	0

Figure 7. 7 The distance matrix of the *C. sakazakii* CC4 isolates (n = 18) produced by the *Cronobacter* BIGsDB comparative genome analysis.

The distance matrix was produced by the alignment of 3917 coding sequences of the reference CC4 isolate SP291 and shows the allelic variability across 18 CC4 isolates. High homology of the reference isolate SP291 with 377 and 1587 is notable.

7.6 GENERAL DISCUSSION AND CONCLUSION

The close association of the *C. sakazakii* CC4 with neonatal meningitis was at the heart of the present study. The genomic analysis in the present study has indicated that *C. sakazakii* CC4 is a stable and a clonal lineage as revealed by the whole alignment of 30 *C. sakazakii* isolates (Figure 3.3). This finding is further strengthened through a recent study by our group using whole genome-MLST (1865 loci) and ribosomal-MLST (51-loci) in *Cronobacter* PubMLST validated the clonal nature of *C. sakazakii* CC4, as CC4 isolates formed a distinct cluster on the phylogenetic trees (Forsythe et al. 2014). The comparative genomic analysis using different comparative genomic strategies could only identify hypothetical proteins as unique in the *C. sakazakii* CC4 genomes without a clear link with its virulence potential (chapters 3-6). Previous and a current parallel PhD study has indicated that *C. sakazakii* CC4 isolates do differ in their clinical presentation such as their adhesion and invasion capability *in vitro* (Nasreddin Rhouma 2012; Faisal Almajed, 2014, unpublished data). Hence, it was of significance to study the genomic diversity within *C. sakazakii* CC4 using whole genome sequencing data. Single Nucleotide Polymorphism (SNP) has emerged as an important comparative genomic tool to study the genomic diversity within bacterial isolates of the same species. The SNP identification can help in phylogenetic characterisation of bacterial isolates, tracing strains during an epidemic, forensic investigations and comparing phenotype to genotype (Gardner & Slezak, 2010). The whole genome alignment of the CC4 isolates used in the present study indicated a very large core genome which suggested a low sequence diversity within this lineage (section 4.4.1). However, SNPs represent the finest resolution of a DNA sequence (Lorenc et al. 2012). Therefore, in the present study, the SNP analysis was used to investigate if further sub-division occurs within this important clinical lineage. The finished genome of *C. sakazakii* SP291 was used as an index for SNP calling in the CC4 isolates. The SNPs were filtered to the high quality parameters previously defined to retain only good quality SNPs (section 2.2.9). The tree topology and the number of SNPs indicated that most of the isolates clustered closely to each other and had SNPs in the range of 300 to 400 (Table 7.1, Figure 7.1). Nevertheless, subdivisions were also observed. The larger branch lengths of *C. sakazakii* 1240 and 1587 were attributed to the increased density of SNPs in genomic areas containing phage acquired traits (Figures 7.3, 7.4). An interesting observation was the close relatedness of the oldest CC4 isolate 377 with the index isolate SP291 as both of these isolates merely differ by 16 SNPs of which most

SNPs were located in a putative integrase in a low sequence coverage area. The possibility of both these isolates share a common origin cannot be excluded as the isolate 377 was obtained from the milk powder while SP291 was obtained from PIF manufacturing environment both from neighbouring countries (UK and Ireland) having free trade with each other. The only ambiguity is the unknown date of isolation for the CC4 isolate SP291. Given the very low sequence diversity between the CC4 isolates 377 and SP291 and close relatedness of their source of isolation, it is proposed that both of these isolates most likely share a common origin.

Another interesting observation was the clustering of low invasive *C. sakazakii* isolates 6 and 558 with each other on the SNP tree (Figure 7.1). Both of these CC4 isolates were temporally and geographically diverse. Isolate 6 was obtained from a clinical source from Canada in 1990 while isolate 558 was obtained from the Netherlands in 1983 from a clinical source (Table 7.1). This was significant as both of these isolates were suggested as low invasive on the HBMEC cell lines *in vitro* in a parallel PhD study (Faisal Almajed 2014, unpublished data). Although only 1 SNP was shared between these 2 isolates, yet both isolates had SNP dense areas; however located within phage associated traits (Figure 7.5). Nevertheless the observation was of significance and hence prompted comparative genomic analysis between high and low invasive *C. sakazakii* CC4 isolates. An attempt was made to identify unique traits in low invasive isolates 6 and 558 by keeping them as a target group in Gegenees (Agren et al. 2012), however none of the traits could be identified as specific to both of these isolates compared to other CC4 isolates (data not shown). Manual comparative genome analysis was undertaken using *C. sakazakii* 767 as high invasive strain against the low invasive *C. sakazakii* isolate 6. A number of unique traits were identified in both high and low invasive isolates (Table S7). No unique invasion associated trait could be identified unique to the high invasive isolate 767. The analysis of the invasion associated genes including *ompA* and *ompX* presented in chapter 3 did not differentiate the low and high invasive CC4 isolates based on their presence as these genes were uniformly detected in all of the *C. sakazakii* isolates (section 3.4.2.2). The interesting observation in this genomic comparison was the presence of a number of metal resistance genes associated with resistance against arsenic nickel, cobalt, chromate, copper and silver which were identified as unique in the genome of low invasive CC4 isolate 6. The metal resistance in bacteria is important in their survival in environments containing high metal concentrations. However,

the association of metal resistance with bacterial invasion is not clear. To test this observation, a total of 10 *C. sakazakii* isolates with different invasion phenotypes were chosen to compare their ability to resist 8 different toxic metals at 4 different concentrations (1M (1000 mm), 0.1M (100 mm), 0.1M (10 mm) and 0.001M (1 mm) (Table 7.2). The choice of the isolates for heavy metal resistance assays was made in a way that they not only represent high and low invasive isolates but also represent equal proportion of CC4 and non-CC4 isolates. With few exceptions, all of the isolates were able to tolerate at least 6 different metals at concentrations of up to 0.1M. The results indicated that metal resistance in *C. sakazakii* is not limited to a certain group as the behaviour of the isolates was independent of their invasion phenotype (Table 7.2).

It should be considered that the metal resistance assays in this study were conducted only to obtain a general overview of the level of difference in resistance of *C. sakazakii* high and low invasive isolates against selected metals. Hence, no control isolates were used. Another limitation is that the dilutions used were not suited for a general purpose metal toxicity assays. Moreover, the composition of the media may also affect the metal resistance behaviour of a bacterial isolate. Therefore, in future, in order to obtain a true picture of the metal toxicity in *C. sakazakii*, the experiments may be repeated using Minimum Inhibitory Concentration (MIC) assays. Appropriate control strains should be introduced in MIC assays while giving consideration to the choice of media for bacterial growth and dilutions used.

The availability of the *C. sakazakii* CC4 isolates analysed in the present study in the *Cronobacter* PubMLST database along with additional genomes provided an opportunity to substantiate the findings of SNP analysis. The BIGsDB comparative analysis was performed using 18 *C. sakazakii* CC4 isolates with isolate SP291 as the reference backbone. The analysis of 3917 genes partly reinforced the findings of SNP analysis as indicated by the close relatedness of *C. sakazakii* isolates SP291 and 377 (Figures 7.6, 7.7). It further indicated that based on the reference 3917 genes, the diversity within CC4 isolates ranged from as low 219 to as high as 2230 loci (allelic variation). The close relatedness of the CC4 isolate 377 with the reference isolate SP291 was obvious further strengthening the proposal that both of these isolates share a common origin.

The present analysis aimed to study the diversity within *C. sakazakii* CC4 isolates indicated the close relatedness of these isolates by SNP analysis. Although subdivisions were

observed within CC4 isolates, yet in most of the cases these differences were attributed to phage associated traits. The presence of heavy metal resistance associated genes in low invasive isolates was of interest, yet the laboratory data did not support that the hypothesis that low invasive *C. sakazakii* are heavy metal resistant. Further studies are warranted to precisely describe the association of metal resistance with invasion of *C. sakazakii*.

8 CHAPTER 8

GENOMIC ANALYSIS OF *CRONOBACTER SAKAZAKII* STRAINS ISOLATED DURING OUTBREAKS IN A NEONATAL INTENSIVE CARE UNIT IN FRANCE, 1994

8.1 BACKGROUND

There have been several outbreaks of *C. sakazakii* during the recent years (van Acker et al. 2001; Himelright et al. 2002) but one of the largest outbreak of *C. sakazakii* which caught the attention of the researchers hit a NICU in France claiming the lives of at least 3 babies (Caubilla-Barron et al. 2007). The outbreak lasted for about 3 months (111 days; 5th May 1994 to 11th July 1994). A total of 18 neonates were infected or colonised with *C. sakazakii* of which 3 neonates died (Table 8.1); neonate H died of meningitis while neonates J and F died of NECII and NECIII respectively. Most of the neonates were underweight with average weight of 1461g. Besides neonate D, all of the remaining neonates were delivered preterm. An autopsy of the neonate H who died of meningitis revealed cerebral lesions. All of the infected neonates developed clinical symptoms within 28 days of birth, exception was neonate K who developed symptoms of NEC1 after 78 days of birth. All neonates (F, H, J) who died were low weight with weights of 1000g, 1500g and 1560g, respectively. Four of the neonates (C, E, O, Q) were asymptomatically colonised while only 2 of the neonates N and P developed digestive problems of moderate nature. A detailed genotypic and phenotypic analysis of the *C. sakazakii* strains isolated from this outbreak has previously been undertaken (Caubilla-Barron et al. 2007). The PFGE analysis by Caubilla Barron and colleagues (2007) divided the strains into 4 distant pulsetypes (PFGE 1-4).

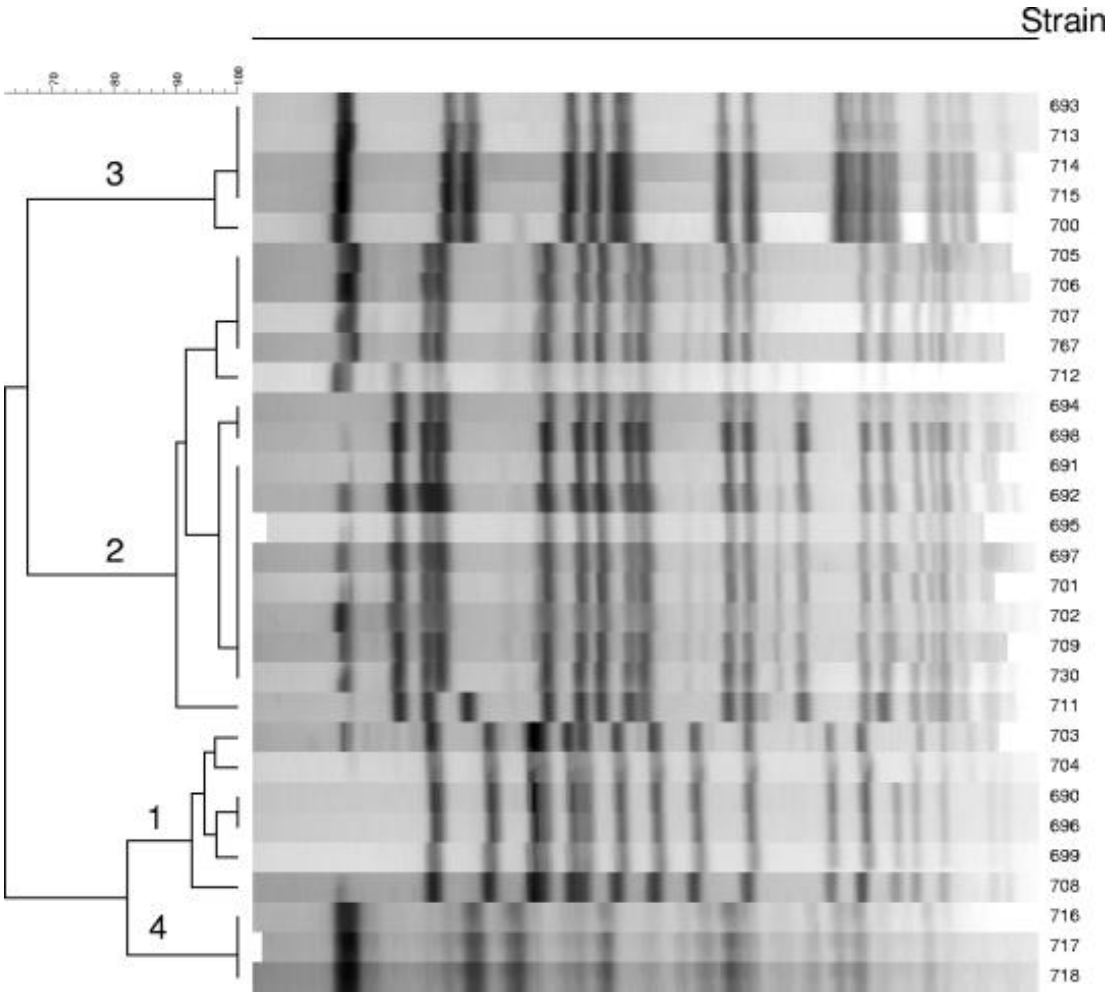


Figure 8. 1 PFGE profile of *C. sakazakii* isolates of French outbreak 1994, generated by Caubilla-Barron et al. 2007*

Dendrogram created from PFGE profiles of *C. sakazakii* isolates (Bionumerics software, version 3.5). Clustering was performed with UPGMA by using the Dice coefficient. Pulsotypes are identified on the left. The tolerance in the band was 1.5%, with an optimisation of 1.5%. The figure has been reused from Caubilla-Barron et al. 2007.
*No copyright permission needed (http://journals.asm.org/site/misc/reprints.xhtml).

It was important to note that all of the 3 fatalities during this outbreak were attributed to the PFGE 2 and more importantly all the PFGE 2 strains belong to clonal complex 4 (CC4) of *Cronobacter* MLST scheme (http://www.pubmlst.org/cronobacter/, Baldwin et al. 2009). Previous studies by our group have indicated strong association of *C. sakazakii* CC4 with neonatal meningitis (Joseph & Forsythe, 2011; Hariri et al. 2013). The advent of next generation sequencing has greatly facilitated the epidemiological investigation at the genomic level. The present study was focused on the reassessment as well as detailed genomic analysis of *C. sakazakii* strains isolated from this outbreak with the main aim of tracing the precise source of outbreak. Nearly all *C. sakazakii* strains previously analysed by

Caubilla-Barron et al. (2007) were sequenced using Illumina MiSeq v3 chemistry (Table 8.1). Unfortunately, due to the long gap between the outbreak and present study, 2 of the isolates (704 of PFGE 1 and 697 of PFGE 2) have been lost from the culture collection, however the absence of these strains did not significantly affect the in depth analysis of this outbreak at the genomic level presented in this study.

8.2 HYPOTHESIS

Powdered Infant Formula (PIF) is the potential source of outbreak**.**

8.3 AIMS OF THE CHAPTER

- Whole genome alignment and phylogeny determination of the 26 sequenced *C. sakazakii* isolates of the 1994 French outbreak.
- Single Nucleotide Polymorphism analysis of the sequenced *C. sakazakii* isolates to trace the source of outbreak.
- Extended spectrum β lactamase (ESβL) profiling of the CC4 isolates of the 1994 French outbreak.
- Virulence associated plasmid alignments of plasmids pAPEC-O2-R and PECOS88 against sequenced isolates.

8.4 METHODS

8.4.1 BACTERIAL STRAINS

A total of 26 *C. sakazakii* genomes were analysed in this study (Table 8.1). The strains were divided into 4 pulsetypes by Caubilla-Barron (2007). Since 3 strains 716, 717 and 718 forming PFGE 4 in the PFGE profile (Figure 8.1) were isolated from an unopened can of the infant formula on the same day therefore, it was assumed that these are the multiple isolates of the same strain and hence only one isolate 716 was sequenced from this cluster.

Table 8. 1 Sequenced strains isolated from *C. sakazakii* the 1994 French NICU outbreak and SNP typing

No	*C. sakazakii* strain	Pulsetype	ST/CC	Baby	Symptoms	Isolation Date (1994)	Day-No	Isolation site	SNPs
1	699	PFGE1	12	A	No-details	23-Mar	1	Trachea	**Reference strain***
2	703	PFGE1	12	B	NECII	25-Apr	34	Trachea	63
3	708	PFGE1	12	C	No symp.	09-May	48	Trachea	52
4	696	PFGE1	12	D	NECII	08-Jun	78	Stools	14
5	690	PFGE1	12	E	No symp.	19-Jun	89	Stools	64
6	701	PFGE2	4	F	NEC III (DIED)	07-Apr	16	Peritoneal fluid	**Reference strain***
7	691	PFGE2	4	G	No-details	19-Apr	28	Sputum	9
8	767	PFGE3	4	H	Meningitis (DIED)	11-May	50	Trachea	38
9	709	PFGE2	4	C	Septicaemia	12-May	51	Trachea	13
10	705	PFGE2	4	B	NEC II	24-May	63	Trachea	339
11	695	PFGE2	4	J	NEC II (DIED)	07-Jun	77	Trachea	11
13	692	PFGE2	4	L	NEC II	13-Jun	83	Stools	15
14	702	PFGE2	4	K	NEC I	13-Jun	83	Stools	10
15	694	PFGE2	4	M	NEC II	14-Jun	84	Conjunctivae	8
16	712	PFGE2	4		NEC II	17-Jun	87	Prepared formula	6
17	707	PFGE2	4	B	NEC II	26-Jun	96	Skin	337
18	711	PFGE2	4	O	No-symptoms	27-Jun	97	Stools	14
19	730	PFGE2	4	K	NEC I	27-Jun	97	Stools	10
20	698	PFGE2	4	D	NEC II	01-Jul	101	Stools	10

Table 7.1 continued

No	*C. sakazakii* strain	Pulsetype	ST/CC	Baby	Symptoms	Isolation Date (1994)	No	Isolation site	SNPs
21	700	PFGE3	13	P	Digestive problems	15-Jun	85	Stools	**Reference strain***
22	693	PFGE3	13	Q	No symp.	18-Jun	88	Stools	4
23	713	PFGE3	13			20-Jun	90	End of bottle	4
24	714	PFGE3	13			27-Jun	97	End of bottle	3
25	715	PFGE3	13			27-Jun	97	Prepared formula	3
26	716	PFGE4	14/1			11-Jul	111	Infant formula	

ST; Sequence Type, CC; Clonal Complex, NEC; Necrotising enterocolitis; Symp.; symptoms, reference strain for SNP calling.

8.4.2 GENOME SEQUENCING, ASSEMBLY, ANNOTATION

Bacterial DNA was extracted from 1-day old cultures using GenElute™ bacterial genome kit (Sigma Aldrich®, USA). The genome sequences of 26 *C. sakazakii* strains were generated on Illumina MiSeq generating 300bp paired end reads (section 2.2.1). The *de novo* assembly was performed using Velvet (version 1.2.09, Zerbino et al. 2007). The genome assembly of the strains which were used as index for SNP calling was improved using Post Assembly Genome Improvement (PAGIT) suite of programmes (Swain et al. 2012). The finished genome of *C. sakazakii* SP291 (Genbank accessions CP004091-4) was used as references for contig reordering. The genomes were annotated using the prokaryotic genome annotation system (PROKKA) (Seemann, 2014). The methods are described in more detail in sections 2.2.3-5.

8.4.3 PHYLOGENETIC ANALYSIS

The genome sequences of 26 *C. sakazakii* genomes (Table 8.1) were aligned using Mugsy (Angiuoli et al. 2012) and the core genome extracted as described previously (Sahl et al. 2011; Clark et al. 2012; McNally et al. 2013). Maximum likelihood phylogeny was then reconstructed using RAxML with the GTR-gamma model (Stamatakis, 2005) and the resulting trees visualised and annotated using Figtree (http://tree.bio.ed.ac.uk/software/figtree/). The method is discussed in detail in section 2.2.8.

8.4.4 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS

Single Nucleotide Polymorphism (SNP) analysis was performed on sequenced *C. sakazakii* strains belonging to cluster 1, 2 and 3 (Table 8.1) using SMALT and SAMtools (Li et al. 2009). The SNP calling was done independently for each cluster using the earliest isolate in each cluster as the reference or index genome. The resulting VCF files were filtered in Artemis to include only SNPs with minimum quality score of 30, minimum depth of 8, and minimum allele frequency of 0.90 (Sahl et al. 2011; Clark et al. 2012; McNally et al. 2013). A consensus sequence file was then created for each genome and used to create a maximum likelihood phylogeny by Dr Alan McNally (section 2.2.9).

8.4.5 EXTENDED SPECTRUM B LACTAMASE DETECTION

The presence of Extended Spectrum β Lactamases (ESβL) and/or AmpC was determined using disc diffusion system of MASTDISC^{TM} by zone size comparison which involved simultaneously testing the antibiotic and antibiotic plus inhibitor combinations. The method is described in more detail in section 2.1.7.

8.5 RESULTS AND DISCUSSION

8.5.1 WHOLE GENOME PHYLOGENY

Whole genome phylogeny of the 26 *C. sakazakii* strains was constructed from the core genome using Mugsy (Angiuoli et al. 2011). The topology of the whole genome based phylogeny was in agreement with the PFGE profiles of Caubilla-Barron et al. (2007). The phylogeny indicated 4 clusters within these strains, cluster 2 being the largest group (Figure 8.2). Three of the clusters (1, 2, 3) were isolated from neonates while the fourth "cluster" consisted of lone strain isolated from unopened can of infant formula. The result of the whole genome phylogeny was in complete agreement with previous study by Caubilla-Barron et al. (2007). Since the grouping of the strains was identical to the PFGE profile, for clarity the clusters were given the same numbers as PFGE i.e. PFGE 1, 2, 3 and 4 represent the same strains as clusters 1, 2, 3 and 4 respectively in the present study. Please note that although isolate 716 was a lone strain, it was still designated it as "cluster 4" to avoid confusion since it was a part of PFGE 4 (Figure 8.1) by Caubilla-Barron et al. (2007). It was interesting to note that all the strains within the cluster 2 belong to *Cronobacter* MLST sequence type 4 or clonal complex 4 (CC4) which has previously been associated predominantly with neonatal meningitis (Joseph & Forsythe, 2011; Hariri et al. 2013) and all three neonates who died were infected by cluster 2 strains (Caubilla-Barron et al. 2007).

Figure 8. 2 Midpoint rooted whole genome maximum likelihood tree of 26 *C. sakazakii* strains isolated from French NICU outbreak, 1994.

The genomes of 26 *C. sakazakii* strains approximately 4 million bp were aligned using Mugsy and maximum likelihood phylogeny generated using RAxML using GTRGAMMA substitution model. The tree was viewed and annotated using FigTree. The phylogeny indicated 4 distinct clusters. Cluster 1; ST12 and PFGE1 (Blue shaded), Cluster 2; CC4 and PFGE2 (Orange shaded), Cluster 3; ST13 and PFGE 3 (Purple shaded), Cluster 4; ST1 and PFGE4 (Green shaded). The tree was rooted to midpoint. The scale bar indicates the number of nucleotide substitutions.

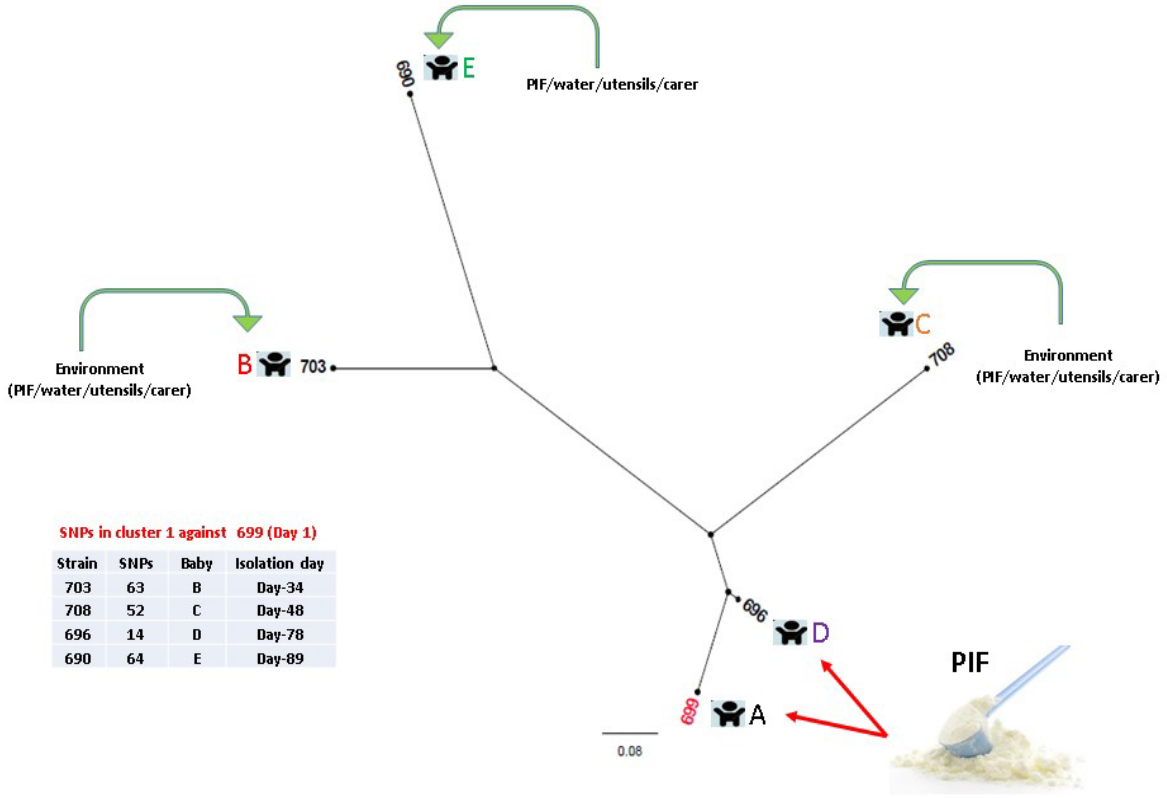
8.5.2 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS

Pathogen typing is essential to identify food-borne infections which have common origins and to help remove the contaminated food which is still in trade, it also helps in tracing back the contaminants to its source. Although PFGE is still widely used in outbreak analysis, its limitations are emerging, for instance co-migration of similar sized bands can obscure their discrimination (Turabelidze et al. 2013). The Single Nucleotide Polymorphism (SNPs) can cause significant changes in phylogenetic distances but they may not change the PFGE pattern. Therefore the isolates may appear identical on PFGE when they are not (Turabelidze et al. 2013). The variation in PFGE pattern in epidemiological data can hinder undertaking informed action against any bacterial outbreak. SNP analysis has proven to be more discriminatory and is being used to differentiate outbreak isolates which often show very low sequence diversity (Turabelidze et al. 2013; Leekitcharoenphon et al. 2014). SNP analysis has been used for the outbreak analysis in a number of studies including the outbreak of *E. coli* O157:H7 attributed to salad bar and romaine lettuce (Turabelidze et al. 2013).

In the present study, 26 *C. sakazakii* isolates were divided into four clusters according to whole genome phylogeny (Figure 8.2) and PFGE (Caubilla-Barron et al. 2007). The SNP analysis was performed only on cluster 1, 2 and 3 as the genome sequence of only a single strain from cluster 4 was available (Table 8.1). Since in the this study the SNP calls were based on the reference genome, the reference genome has to be closely related to the strains being investigated. Using an un-related reference genome can significantly increase the number of SNPs which ultimately can affect the SNP phylogeny (Leekitcharoenphon et al. 2014). Therefore in the present study each of 3 clusters was analysed independently of the isolates from other clusters. The earliest isolate in each cluster was used as the reference or the index strain to identify SNPs in the subsequently isolated strains. To improve the quality of the identified SNPs, assembly of the reference genome was improved using PAGIT. The SNP calls were determined using SMALT and SAMtools as described previously.

8.5.2.1 SNP TYPING OF CLUSTER 1

Five of the isolates formed cluster 1 which were isolated between 23rd March and 19th June from babies A, B, C, D and E (Table 8.1). Two of these neonates (B and D) developed symptoms of NECII while 2 (C and E) remained asymptomatic, for the remaining 1 neonate (A) no clinical details were available. All of these isolates belong to ST12 of *Cronobacter* MLST scheme (Baldwin et al. 2009). SNP typing was undertaken in order to observe the strain relatedness within this cluster; the earliest isolate of this cluster i.e. *C. sakazakii* 699 isolated on 23rd March 1994 was used as the reference for SNP typing and SNP phylogeny determined (Figure 8.3).



SNPs in cluster 1 against 699 (Day 1)

Strain	SNPs	Baby	Isolation day
703	63	B	Day-34
708	52	C	Day-48
696	14	D	Day-78
690	64	E	Day-89

Figure 8. 3 SNP phylogeny of the cluster 1 strains.

The SNPs were called using SMALT and SAMtools to generate the VCF files which were filtered using VCFTools to include only SNPs with minimum quality score of 30, minimum depth of 8, and minimum allele frequency of 0.90. A consensus sequence file was then created for each genome and used to create a maximum likelihood phylogeny. The red font colour indicates the reference strain. The scale bar indicates the number of nucleotide substitutions.

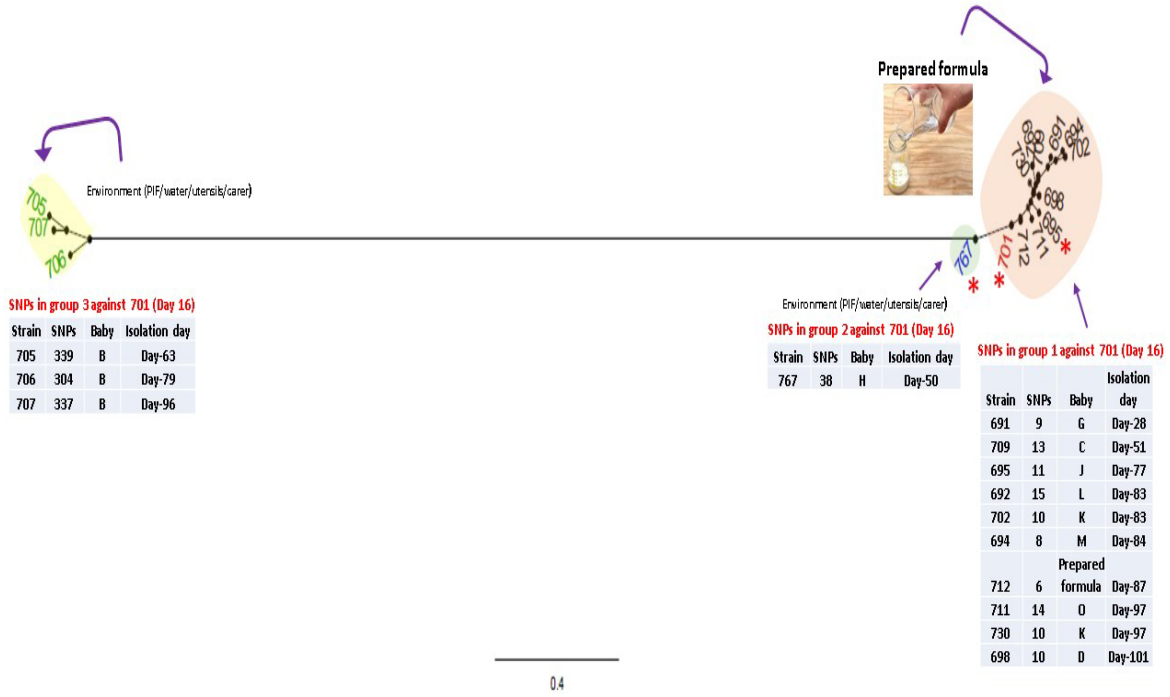
The number of SNPs and tree topology indicated that the cluster 1 strains were relatively distant to the reference strain 699. The reference strain differed from the strain 696 by only 14 SNPs indicating a possible common origin. One possibility could be the transfer of the strain from baby A to D but it seems unlikely as both of these isolates were obtained 78 days apart. Since both of these strains were isolated over 2 months apart from each other therefore from the very low number of SNPs it is reasonable to speculate that these isolates share a common static origin where their growth was halted. It is unlikely that environmental sources in a hospital environment such as water or carer's hands could be the origin for these 2 isolates. Therefore from these observations it is proposed that the origin for the isolates 699 and 696 is highly likely to be the PIF and both of the babies A and D could have been fed the same batch of PIF.

The remaining three strains; 703, 708 and 690 were distant to the reference strain by over 50 SNPs each; therefore for the cluster 1 strains, it is assumed that apart from isolates 699 and 696 which are likely to be originated from PIF, the other 3 isolates were acquired independently from different sources, it could be PIF, water used to re-constitute the formula, the utensils used to handle the PIF or carer's hand. It is interesting to note that like strain 703, three of the cluster 2 isolates (705, 706 and 707) were also isolated from baby B; however the strains in the two clusters were entirely different (~70,000 SNPs) indicating that baby B was co-infected with different unrelated strains of *C. sakazakii*.

8.5.2.2 SNPS IN CLUSTER 2

A total of 15 sequenced strains belonged to cluster 2 and had been isolated between 7th April and 1st July, 1994. These stains were isolated from different sites including peritoneal fluid, sputum, trachea, stools, skin and conjunctivae. The period of isolation of these strains was in parallel with cluster 1 strains. Of these strains, 11 were isolated from NEC cases at different stages, 1 from a septicaemia and 1 from a meningitis case, the remaining 2 isolates were obtained from babies O and G for whom no clinical details were provided. In the present analysis, the earliest isolate of this cluster *C. sakazakii* 701 isolated on 7th April 1994 was used as a reference strain for SNP calls in order to observe the strain relatedness and to determine the possible route of transmission. The SNPs in each of the strain isolated according to the date order were determined using SAMTools (Table 8.1) and SNP phylogeny (Figure 8.4) was constructed as described earlier.

The tree topology indicated at least three subdivisions within cluster 2 strains; a group of 11 isolates (group 1) joined by a short arm to isolate 767 (group 2) and through a large arm with three isolates 705, 706 and 707 (group 3). All of the isolates in group 1 (Figure 8.4, orange encircled) had remarkable degree of similarity with the reference (index) strain as shown by the number of SNP differences; maximum of 15 SNP differences to the index strain.



SNPs in group 3 against 701 (Day 16)

Strain	SNPs	Baby	Isolation day
705	339	B	Day-63
706	304	B	Day-79
707	337	B	Day-96

SNPs in group 2 against 701 (Day 16)

Strain	SNPs	Baby	Isolation day
767	38	H	Day-50

SNPs in group 1 against 701 (Day 16)

Strain	SNPs	Baby	Isolation day
691	9	G	Day-28
709	13	C	Day-51
695	11	J	Day-77
692	15	L	Day-83
702	10	K	Day-83
694	8	M	Day-84
712	6	Prepared formula	Day-87
711	14	O	Day-97
730	10	K	Day-97
698	10	D	Day-101

Figure 8. 4 SNP phylogeny of the cluster 2 strains.

The SNPs calls were obtained using SMALT and SAMtools to generate the VCF files which were filtered using VCFTools to include only SNPs with minimum quality score of 30, minimum depth of 8, and minimum allele frequency of 0.90. A consensus sequence file was then created for each genome and used to create a maximum likelihood phylogeny. The tip labels shaded orange, green and yellow indicates groups 1, 2 and 3 respectively within cluster 2. The red tip label shows the reference strain. The red asteric represents the strains isolated from the fatal cases. The scale bar indicates the number of nucleotide substitutions.

An interesting observation was the close relatedness of the reference isolate 701 with the prepared formula isolate 712 which differed from the reference isolate 701 by only 6 SNPs. Although the reference strain 701 was isolated at least two months earlier than the prepared formula isolate 712, the tree topology and small number of SNP differences suggest a common source of origin. Since *C. sakazakii* has the ability to survive in a wide variety of environments, it is unclear how the prepared formula was contaminated; in this case the potential source of contamination could either be PIF itself, water used to prepare formula, colonisation of bacterium on the utensils used for the feed preparation or the carer's hands. However, since all of these 11 strains were isolated over 73 days period (25th April–19th June, 1994), it suggested that these isolates are likely to have a common origin. It seems unlikely that the same water was used over 73 days period to prepare PIF or the organism survived in the utensils or on the carer's hands for that long, therefore it seems plausible that these isolates originated from a particular batch of the PIF which was used to

prepare baby formula. The analysis therefore suggest PIF to be the potential source of transmission of cluster 2 strains to at least 10 babies; F, H, C, O, J, D, L, G, M and K of which babies F, H and J died. Two of the isolates; 702 and 730 were obtained from the same baby K, 14 days apart; both of these strains differ from each other by 24 SNPs which could be the result of microevolution.

The second "group" within cluster 2 was formed of a single isolate *C. sakazakii* 767 isolated from baby H, who died of meningitis (Figure 8.4, shaded green). This strain differed from the reference strain by 38 SNP, therefore from the SNP typing and tree topology, it seems likely that this strain is relatively different from all other isolates in cluster 2 and could have been acquired from a different source. The SNP mapping on the DNA plotter indicated a localisation of SNPs in this strain (Figure 8.5).

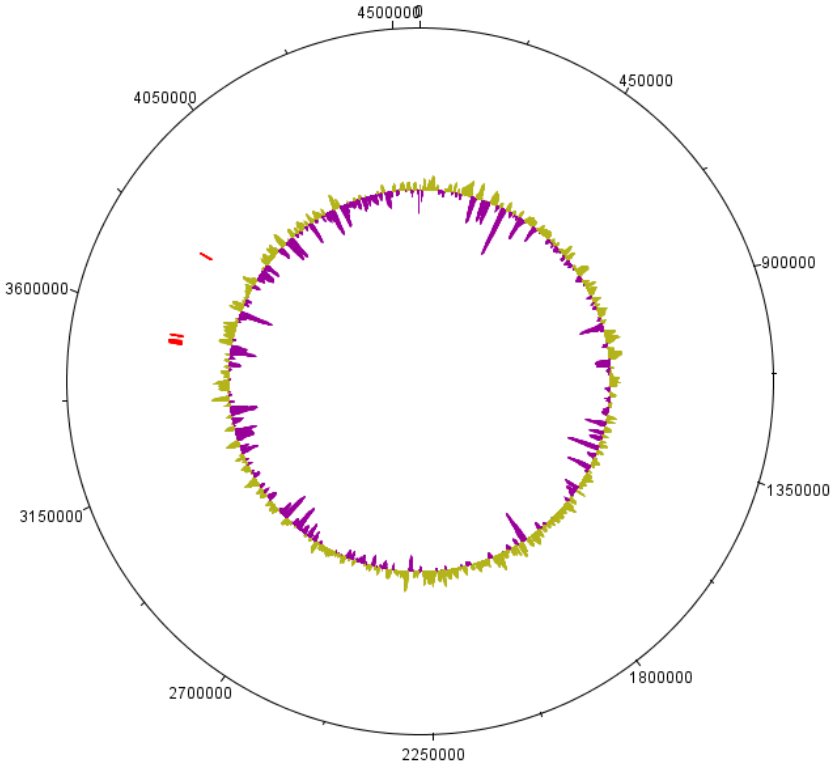


Figure 8. 5 SNPs in *C. sakazakii* 767 against index isolate 701.

The DNA plotter image indicating the localization of the SNPs (red mrkings) in isolate 767 in reference to isolate 701. The outer most black circle with tick marks show the genome size and positions. The red markings indicate the location of the SNPs in isolate 767 while the inner most spiked circle indicates the GC content (green: above average, purple: below average, average: 50%).

It was of importance to analyse these SNPs in isolate 767 in more detail. Therefore, further manual curation of these SNPs was performed by observing the precise location of these SNPs, gene annotation and visualising the sequence coverage at that area using the BAM file which was generetaed as an output of SAMTools (Figure 8.6).

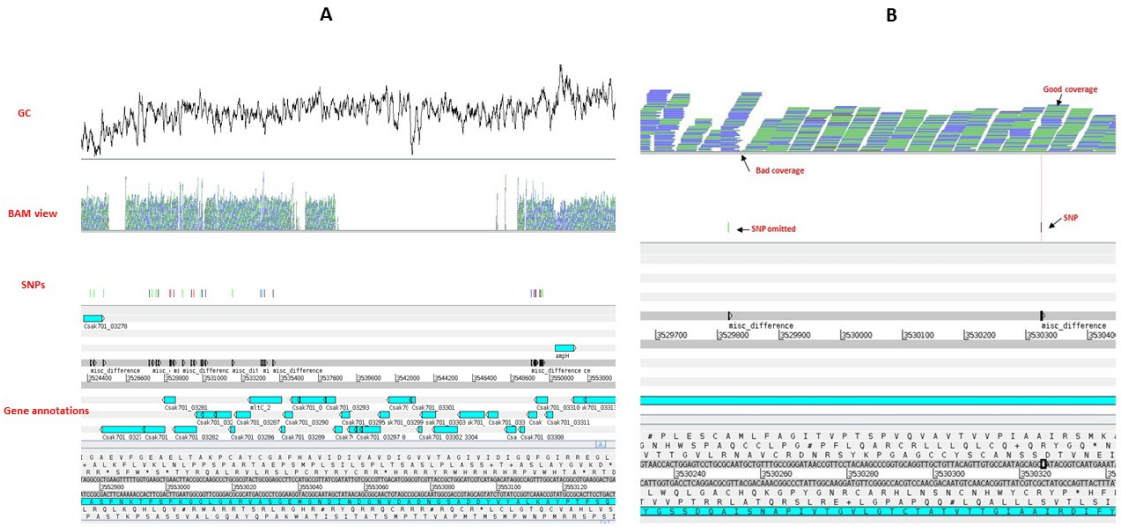


Figure 8. 6 Manual curation of the SNPs in *C. sakazakii* isolate 767.

Panel A; All the SNPs in *C. sakazakii* 767 in reference to *C. sakazakii* 701, the bottom section shows the genomic annotation of the region containing the SNPs, above the bottom section the coloured markings indicate SNPs, below the top section of the panel is the BAM view showing the sequence coverage of the region while the top section with black spikes indicate the variation in GC content. **Panel B**; The zoomed in view showing an example of the details for the manual curation, the SNPs in the good coverage area were included while SNPs in bad coverage area were omitted.

The in depth analysis indicated that out of 38 SNPs in *C. sakazakii* 767, at least 29 SNPs were located within phage associated traits (Table 8.2). The remaining 9 SNPs were located in uncharacterised hypothetical proteins, whether these hypothetical proteins also belong to the acquired traits remains unclear, however the lower GC content at that region as seen on DNA plotter shows it is highly likey. Therefore, these differences could not be linked with any functionality in *C. sakazakii* 767 which only seem to differ from the 11 strains of group 1 within cluster 2 by potentially acquired genetic traits. Therefore from these observations, whether the source of origin for isolate 767 was the same as isolates in group 1 remains ambiguous.

Table 8. 2 Genomic annotations containing SNPs in *C. sakazakii* 767 against the index isolate 701

No	POS (701)	REF	ALT	QUAL	GENE/LOCUS_TAG	ANNOTATION
1	3524559	C	A	53	Csak701_03278	Acyltransferase family protein
2	3524760	G	A	157	Csak701_03278	Acyltransferase family protein
3	3525295	G	A	113	Csak701_03279	Hypothetical protein
4	3527922	G	T	222	Csak701_03280	Phage Tail Collar Domain protein
5	3528092	C	A	50	Csak701_03280	Phage Tail Collar Domain protein
6	3528306	G	A	139	Csak701_03280	Phage Tail Collar Domain protein
7	3528444	C	T	99	Csak701_03280	Phage Tail Collar Domain protein
8	3529084	T	C	110	Csak701_03281	Phage Tail Collar Domain protein
9	3529146	A	C	222	Csak701_03281	Phage Tail Collar Domain protein
10	3529315	A	G	117	Csak701_03281	Phage Tail Collar Domain protein
11	3529818	C	A	102	Csak701_03282	Hypothetical protein
12	3530325	G	T	81	Csak701_03282	Hypothetical protein
13	3530469	T	C	53.1	Csak701_03282	Hypothetical protein
14	3530891	C	A	48.1	Csak701_03283	Hypothetical protein
15	3530960	A	T	39	Csak701_03283	Hypothetical protein
16	3531134	T	G	72	Csak701_03284	phage baseplate assembly protein V
17	3532674	G	A	86	Csak701_03286	Hypothetical protein
18	3534298	A	G	66	mltc_2, Csak701_03288	Membrane-bound lytic murein transglycosylase precursor C
19	3534409	G	A	106	mltc_2, Csak701_03288	Membrane-bound lytic murein transglycosylase precursor C
20	3534514	A	G	110	mltc_2, Csak701_03288	Membrane-bound lytic murein transglycosylase precursor C
21	3534984	C	T	191	mltc_2, Csak701_03288	Membrane-bound lytic murein transglycosylase precursor C
22	3549746	A	G	108	Csak701_03309	Hypothetical protein
23	3549914	A	G	115	Csak701_03309	Hypothetical protein
24	3549980	G	T	222	Csak701_03309	Hypothetical protein
25	3550007	T	C	222	Csak701_03309	Hypothetical protein
26	3550237	C	G	154	Csak701_03310	putative chitinase
27	3550291	G	C	164	Csak701_03310	putative chitinase
28	3550321	G	A	156	Csak701_03310	putative chitinase
29	3550402	T	A	126	Csak701_03310	putative chitinase
30	3797427	A	G	38.3	intA_3, Csak701_03534	Prophage CP4-57 integrase
31	3797700	G	T	151	intA_3, Csak701_03534	Prophage CP4-57 integrase
32	3797793	T	G	63	intA_3, Csak701_03534	Prophage CP4-57 integrase
33	3797858	C	A	123	intA_3, Csak701_03534	Prophage CP4-57 integrase
34	3797868	G	A	47.3	intA_3, Csak701_03534	Prophage CP4-57 integrase
35	3797919	C	A	91	intA_3, Csak701_03534	Prophage CP4-57 integrase
36	3797949	G	T	114	intA_3, Csak701_03534	Prophage CP4-57 integrase
37	3798204	C	T	153	intA_3, Csak701_03534	Prophage CP4-57 integrase
38	3798339	C	A	119	Un-annotated	

POS; Position, REF; reference nucleotide, ALT; altered nucleotide

The third group within cluster 2 consisted of 3 strains (Figure 8.4, shaded yellow) was distant to the reference strain as indicated by tree topology and SNP typing; all of the isolates in this group were obtained from the same infant i.e. baby B. Each of these isolates differ from the reference strain by more than 300 SNPs. This observation is in agreement with Caubilla-Barron et al. (2007) where these strains formed a subcluster within the PFGE 2 (Figure 8.1). Further manual analysis showed that all of these 3 strains had a concentration of SNPs between 3.5-4.5 MB region loci which likely was the difference between these isolates and others in cluster 2 (Figure 8.7).

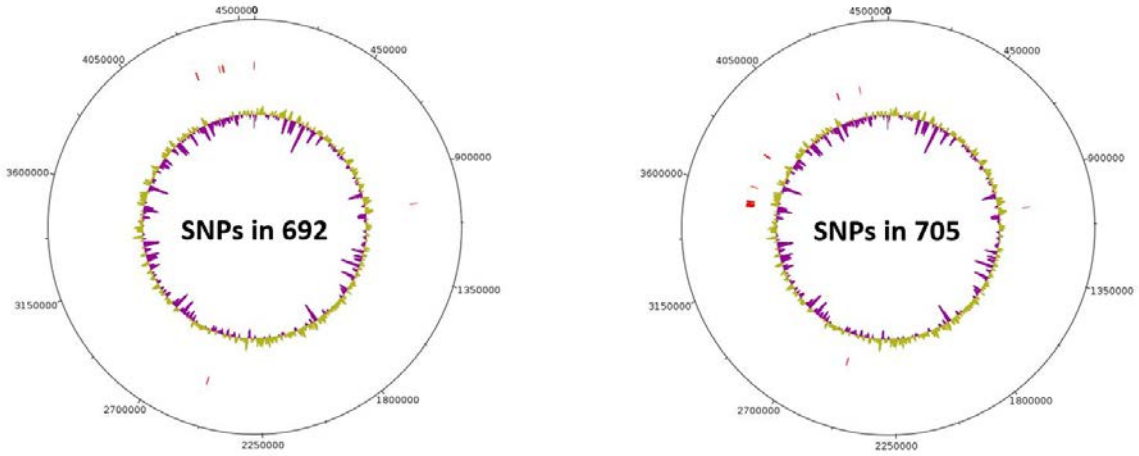


Figure 8. 7 The SNP differences between group 1 and group 3 within cluster 3.

The figure above shows the comparison between the strain 692 (group 1 of cluster 2) and 705 (group 3 in cluster 2). The outer most, black circle with tick marks shows the genome size and positions. The red markings indicate the location of the SNPs while the innermost circle with green (above average) and purple spikes (below average) indicate GC contents. SNP image of one isolate from each group was chosen to show the differences, the concentrated SNPs as shown in 705 were missing from all the group 1 isolates within cluster 2.

It was beyond the scope of the present study to curate each of these single SNP and its annotation, however the genomic region in general contained a wide variety of traits including large number of hypothetical proteins. It was interesting to note that location of these SNPs was correlated with those of isolate 767, however the SNP density and spread (3.5-4.5MB) for the group 3 isolates was much higher. Whether this SNP concentration indicates a recent recombination event within a subset of cluster 2 or *C. sakazakii* CC4 requires further investigation, yet it shows the common origin of these 3 isolates which differs from other isolates in cluster 2. Further analysis of these group 3 isolates within

cluster 2 was undertaken by SNP typing to observe homology amongst themselves. The SNP typing suggested that all 3 isolates showed a remarkable degree of homology amongst themselves and differed from each other by a maximum of 16 SNPs (data not shown); since these three isolates were isolated from baby B at different time points, the SNP differences between them could be the result of a microevolution. As distant on the phylogenetic tree from the reference and other isolates in cluster 2 (Figure 8.4), it is assumed that these strains might have infected the baby B through a different route of transmission, therefore the same batch of the PIF is unlikely to be the source of infection for baby B.

8.5.2.3 SNPS IN CLUSTER 3

Five of the isolates belonging to cluster 3 were isolated between 15th to 27th June (Table 8.1). Two of the isolates (700, 693) were obtained from babies P and Q, the former had moderate digestive problems while the later was asymptomatic. Of the remaining 3 isolates, 2 isolates (713, 714) were obtained from the bottle while 1 isolate (715) was obtained from unused prepared formula. The date of isolation of these isolates overlapped with cluster 1 and 2 isolates. The MLST profile of these isolates indicated all of these isolates were ST13. All of the 5 strains were obtained between days 85 to 97 of the outbreak. The earliest isolate within this cluster, *C. sakazakii* 700 was used as the index to call SNPs in the remaining 4 isolates. The SNPs were typed as described before and a phylogeny obtained (Figure 8.8). The SNP typing and tree topology of the isolates obtained from the babies P and Q, isolates obtained from end of the bottle and prepared formula indicates that all of these strains were highly identical to the index strain as a maximum of 4 SNPs were obtained. This observation led to the assumption that the potential source of the baby infection or colonisation was prepared formula. Whether the prepared formula was contaminated because of the survival of *C. sakazakii* in PIF or was due to contamination from a different source such as water, utensils used to prepare formula feed, temperature abuse or carer's hands remains unclear. However, like group 1 of the cluster 2, it seems likely that PIF was also the source of contamination for the cluster 3 strains as these isolates seem to have a common origin and they were isolated over a 12 day period where it seems unlikely that contamination could be from any other source such as water.

SNPs in group 3 against 700

Strain	SNPs	Baby	Isolation day
693	4	Q	Day-88
713	3	End of bottle	Day-90
714	3	End of bottle	Day-97
715	3	Prepared formula	Day-97

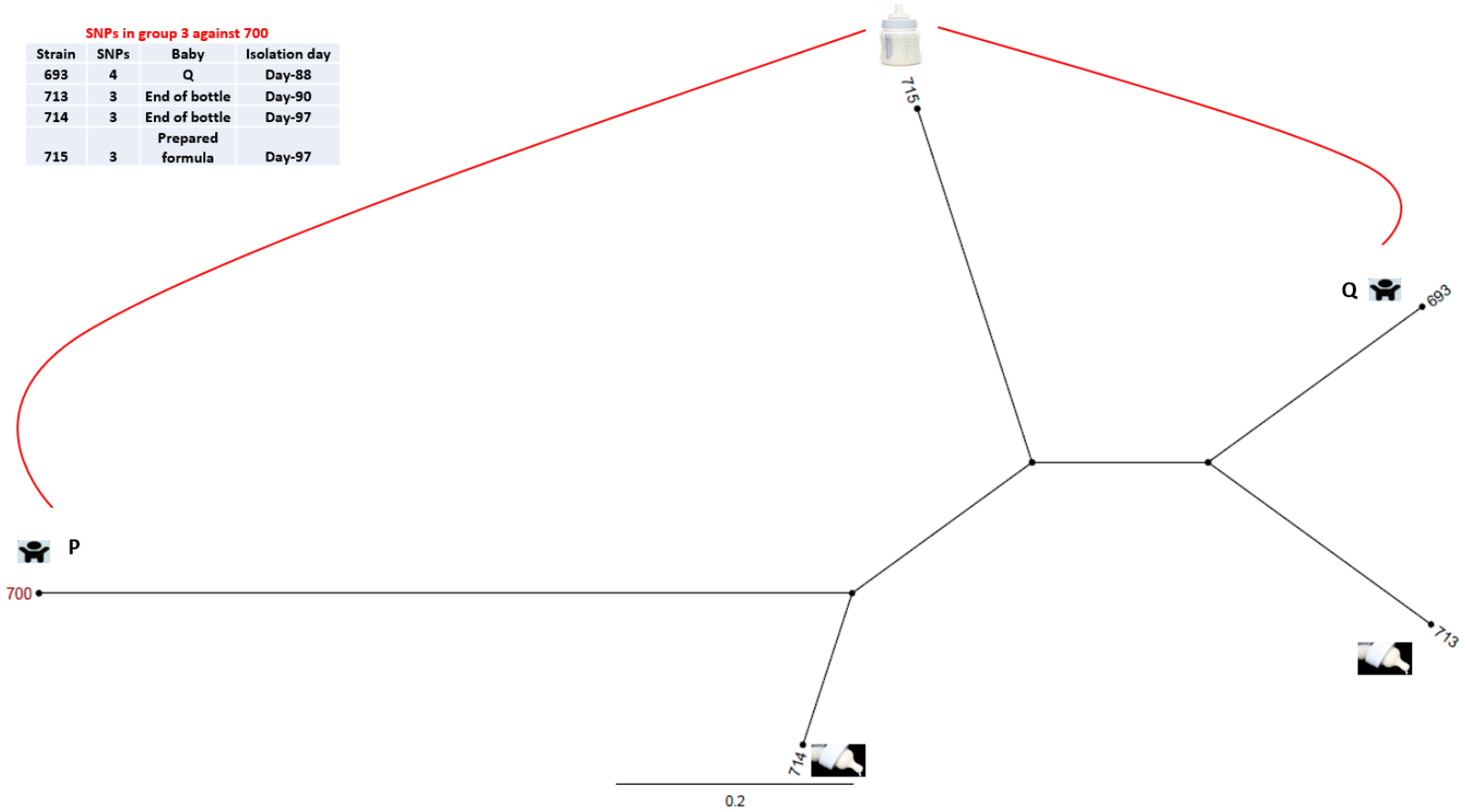


Figure 8. 8 SNP phylogeny of the cluster 3 strains.

The SNPS were called using SMALT and SAMtools to generate the VCF files which were were filtered using VCFTools to include only SNPs with minimum quality score of 30, minimum depth of 8, and minimum allele frequency of 0.90. A consensus sequence file was then created for each genome and used to create a maximum likelihood phylogeny. The purple font colour indicates the reference strain. The scale bar indicates the number of nucleotide substitutions.

8.5.2.4 SUMMARY

In summary the SNP level investigation revealed that there were at least 3 outbreaks of diverse *C. sakazakii* isolates in the French NICU and there was a likelihood of a 4th outbreak. Contaminated batch (or batches) of PIF was the likely source of dissemination for at least two of the cluster 1, eleven of the cluster 2 and all five isolates of cluster 3 while origin of three cluster 1 and four of the cluster 2 isolates remains unclear as *C. sakazakii* has the ability to survive in wide variety of environment. There was a likelihood of a 4th outbreak as the cluster 4 isolates were obtained from the unopened can of formula; it is important to note that cluster 4 strain *C. sakazakii* 716 belongs to clonal complex 1 (CC1) which has frequently been associated with infant formula. At least 35% (as of date 20th Sep 2014) of the CC1 isolates in *Cronobacter* PubMLST database have either been isolated from infant formula, infant formula ingredients or the milk powder manufacturing factories while at least 12% of the CC1 isolates belong to clinical cases which also include *C. sakazakii* BAA-894 (Kucerova et al. 2010). Therefore, the risk of a 4th outbreak was quite high if the formula was fed to a neonate, it is plausible due to heightened risk that the hygienic practices were improved which may have avoided the dissemination of this clinically important lineage of *C. sakazakii* into neonates.

8.5.3 ESβL profiling of the cluster 2 isolates

Extended spectrum beta-lactamases (ESβL) are the enzymes harboured by certain bacterial species which provide resistance to bacteria against penicillin and cephalosporin antibiotics. The emergence of antibiotic resistance presents a serious challenge as it limits the treatment options available against the bacterial infections. Since in the French outbreak, only cluster 2 strains were associated with the deaths of 3 neonates, it was important to screen these strains for the ability to resist antibiotic treatments. All the strains in cluster 2 were tested for their ability to resist ESβL and/or AmpC using disc diffusion method of MASTDIC^{TM}. All of the strains were found to be sensitive to ESβL and/or AmpC (Table 8.3).

In addition to the laboratory ESβL profiling, several known resistance genes which confer resistance against beta lactam drugs such as *blaCTX-M3* (*Citrobacter freundii*), *blaCTX-M-1 (Escherichia coli* O25b:H4-ST131 str. EC958), *ampC* (*Escherichia coli str.* K-12 substr. MG1655) and *blaSHV-5* (*Klebsiella pneumoniae*) were also searched for the homologues in all of the *C. sakazakii* genomes, however no homologue was detected in any of the *C.*

sakazakii genomes analysed in this chapter. In future studies, it is also recommended that Antibiotic Resistance Genes Database (ARDB) (http://ardb.cbcb.umd.edu/) should also be searched for antibiotic resistance associated genes using genome sequenced *C. sakazakii* isolates.

Table 8. 3 ESβL and/or AmpC profile of the French outbreak cluster 2 strains using disc diffusion system of MASTDISC^{TM*}

Strain	A (CPD10)	B (CPD10 + ESβL inhibitor)	C (CPD10 + AmpC inhibitor)	D (CPD10 + ESβL inhibitor + AmpC Inhibitor)						
	Dia.	Dia.	Dia.	Dia.	B-A	D-C	B-D	A-C	ESβL	AmpC
707	28	30	29	30	2	1	0	-1	Neg.	Neg.
691	27	28	27	27	1	0	1	0	Neg.	Neg.
695	23	24	24	23	1	-1	1	-1	Neg.	Neg.
701	28	27	28	29	-1	1	-2	0	Neg.	Neg.
767	22	22	21	22	0	1	0	1	Neg.	Neg.
730	27	28	27	27	1	0	1	0	Neg.	Neg.
711	30	29	29	29	-1	0	0	1	Neg.	Neg.
706	28	27	30	28	-1	-2	-1	-2	Neg.	Neg.
712	30	30	28	28	0	0	2	2	Neg.	Neg.
702	30	28	28	27	-2	-1	1	2	Neg.	Neg.
692	29	30	30	30	1	0	0	-1	Neg.	Neg.
698	28	30	30	28	2	-2	2	-2	Neg.	Neg.
694	29	30	29	27	1	-2	3	0	Neg.	Neg.
705	28	28	28	28	0	0	0	0	Neg.	Neg.
709	30	30	30	29	0	-1	1	0	Neg.	Neg.

The ESβL and AmpC discs (MAST Group limited, UK) were placed equidistant on the TSA plates and pressed gently with flame sterilised forceps to ensure complete contact with the TSA media. The TSA plates with ESβL and AmpC discs were incubated at 37^{o}C for 24h. For each disc the diameter of the zone was measured and compared with guidance image (D68C[1], Figure 2.1) provided by the Mast group. *refer to Figure 2.1 (section 2.1.7), Dia.; Diameter in mm, Neg.; Negative.

8.5.4 Alignment of *C. sakazakii* genomes against virulence associated plasmids

The development of resistance against antibiotics and toxic metals in bacteria pose an increasing challenge to the treatment of infectious disease. Antibiotic and metal resistance was thought to be linked with the presence of resistance genes both against antibiotics and toxic metals, located in close proximity on the bacterial plasmids (Samanta et al. 2012). The presence of the R plasmids such as pAPEC_O1-R and pAPEC-O2-R plasmids in extra intestinal pathogenic *E. coli* which are transmissible across bacteria suggest that these plasmids may act as the reservoir for the antimicrobial agents which may ultimately affect human health.

A homologue of the pAPEC-O2-R plasmid called PECOS88 has been found to be encoded on the genome of the neonatal meningitic *E. coli* strain S88 (O45: K1: H7) where it has been linked with increased level of bacteraemia in neonatal rat model of meningitis (Peigne et al. 2009). Resistance genes encoded on pAPEC-O2-R in addition to the *silCERS*, *silAB* and *silP* which provide resistance against silver and copper, also include *folA* providing resistance to trimethoprim, *aadA5* which encodes resistance to aminoglycoside, *catB3* which contributes to resistance against chloramphenicol, and a beta lactamase gene (*bla*) contributing to resistance against beta lactam drugs (Johnson et al. 2005). *C. sakazakii* is ubiquitous in the environment and a frequent coloniser of humans and animals; the sharing of habitats with other commensal and pathogenic bacteria such as *E. coli* increase the chances of gene transfers through genetic exchange or through horizontal gene transfer. It was therefore of interest to find out whether a complete or partial homologue of pAPEC-O2-R or any of its resistance genes have homologues in *C. sakazakii* genomes being investigated in the present study and if present do these traits show variability across the 4 *C. sakazakii* clusters. To answer these questions plasmid pAPEC-O2-R was aligned against all 26 *C. sakazakii* genomes using BLAST Ring Image Generator (BRIG) and a comparative genome image obtained (Figure 8.9).

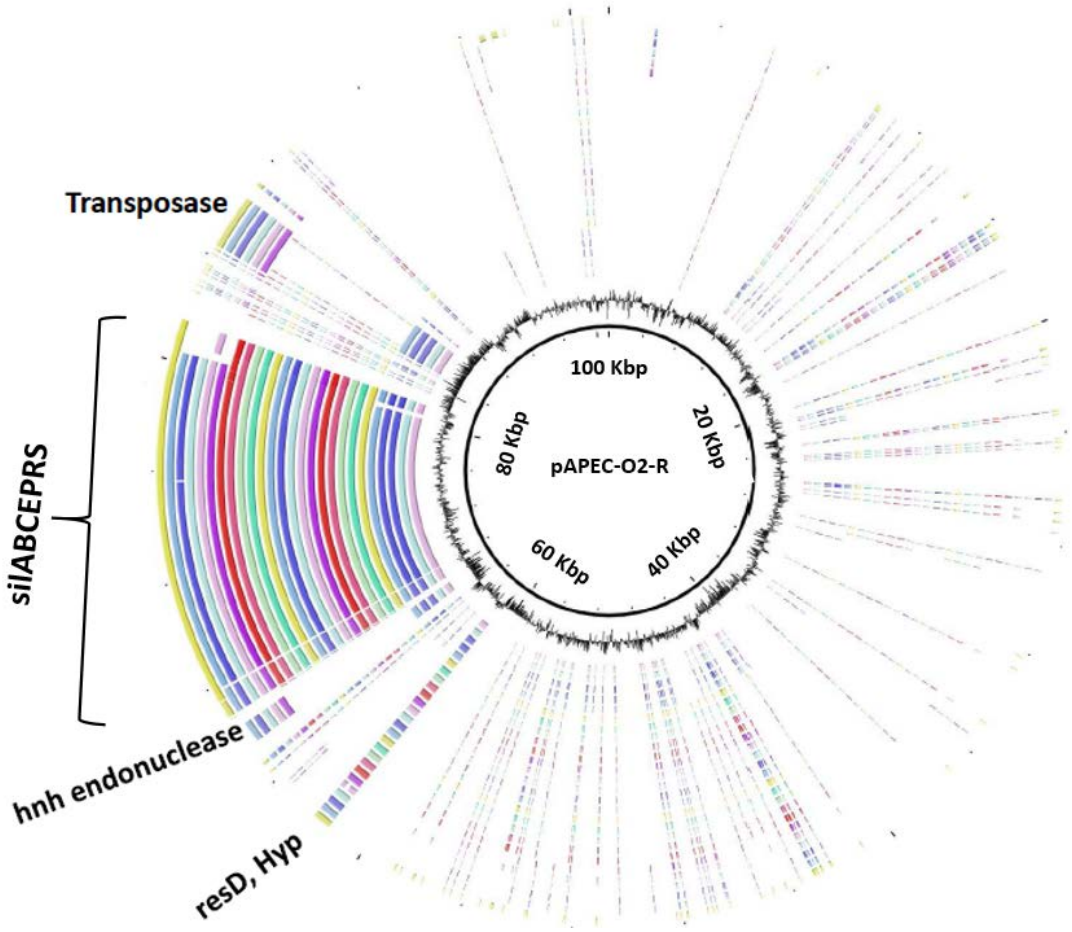


Figure 8. 9 Comparative genome image showing the alignment of pAPEC-O2-R plasmid against *C. sakazakii* genomes.

The image was generated using BRIG. The central black solid circle indicates the plasmid pAPEC-O2-R (~100 kb). The distance between each dot is 4 kb. Second black spiked circle from inside out represents GC contents. Each of the coloured rings represent a different *C. sakazakii* genome, the colour coding is shown on the colour legend on the right of the figure. The solid colour rings are indicative of high BLAST homology with the reference while fade and broken rings represent low degree of homology. Rings from inside out represent isolates 690, 696, 699, 703, 708, 691, 692, 694, 695, 698, 701, 702, 705, 706, 707, 709, 711, 712, 730, 767, 693, 700, 713, 714, 715 and 716 respectively. (Hyp: hypothetical protein).

The result in Figure 8.9 showed that complete homologue of the pAPEC-O2-R was not detected in any of the *C. sakazakii* genome. However, it was interesting to note that a region of pAPEC-O2-R located between ~61-88 kb showed homology to most of the *C. sakazakii* genomes, the region was designated as homologous region. Further analysis by Artemis and WebACT indicated that the region consisted of 25 genes (Locus_tag: O2R_88 to O2R_111) (Table 8.4). This similarity between pAPEC-O2-R of avian *E. coli* and *C. sakazakii* was interesting and could be the result of horizontal gene transfer between *E. coli* and *C. sakazakii*, however to date there has been no report of avian isolation of *C. sakazakii*.

This region was selected and further investigated by *Cronobacter* BLAST (http://pubmlst.org/perl/bigsdb/bigsdb.pl?db=pubmlst_*cronobacter*_isolates&page=plugin&name=BLAST/). Each of the genes was BLAST searched to precisely check its presence in 26 *C. sakazakii* genomes. The region consisted of a gene cluster *silABCEPRS* encoding for silver and copper efflux and binding proteins, this gene cluster was uniformly detected in all 26 *C. sakazakii* genomes irrespective of the cluster grouping; the region has been previously described by our group (Kucerova et al. 2010; Joseph et al. 2012b) and has been investigated previously in this thesis (section 3.2.4.9). Since the region was present in all 4 clusters it indicates that although the region might be essential for the fitness of *C. sakazakii* in general. Another set of genes encoded on the plasmid pAPEC-O2-R was *tetA* and *tetR* encoding for tetracycline efflux protein and tetracycline repressor protein respectively, both of these genes were uniformly absent in all the of *C. sakazakii* genomes. The gene *O2R_88* encoding for resolvases (*resD*) was present in all while another resolvase encoded by *O2R_110* was absent from all *C. sakazakii* genomes. The hypothetical proteins encoding genes *O2R_89-90* were missing in all while other hypothetical proteins encoding genes *O2R_92, O2R_94, O2R_99, O2R_102, O2R_104.2* and *O2R_104* were detected in all of the *C. sakazakii* genomes. The gene O2R_93 encoding for hnh endonuclease was present in all the *C. sakazakii* genomes of cluster 1 and 3 while absent in all cluster 2 strains except *C. sakazakii* 701, the gene was also absent from the cluster 4 strain *C. sakazakii* 716. Genes *O2R_105* and *O2R_108* encoding a hypothetical protein and a relaxase respectively, were uniformly absent from all genomes whereas *O2R_109* encoding for a transposase was missing only in cluster 2 strains. Hence, except for a transposase which was absent specifically from the cluster 2 genomes, no pattern for the presence of the traits could be observed in any particular cluster of the *C. sakazakii* strains analysed in the present study.

Table 8. 4 Investigation of the pAPEC_O2-R homologous region using *Cronobacter* BLAST

	Gene	resD	O2R_89	O2R_90	O2R_91	O2R_92	O2R_93, hnh	O2R_94	silE	silS	silR	silC	O2R_99	silB	silA	O2R_102	silP	O2R_104	O2R_104.2	O2R_105	tetA	tetR	O2R_108	O2R_109	O2R_110	TnpM
	Annotation	Resolvase	Hyp	Hyp	Hyp	Hyp	Hnh endonuclease	Hyp	Silver-binding protein	silS	silR	Outer membrane efflux protein	Hyp	Efflux transporter, RND family, MFP s	Heavy metal efflux pump	Hyp	Copper-translocating P-type ATPase	Hyp	Transposase	Hyp	Tetracycline efflux protein	Tetracycline repressor protein	Relaxase	Tn1721 transposase	Resolvase	TnpM
C. sakazakii Isolate																										
690	Cluster 1	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-
699		+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-
703		+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-
708		+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-
696		+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-
691	Cluster 2	+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
692		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
694		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
695		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
698		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
701		+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
702		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
705		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
706		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
707		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
709		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
711		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
712		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
730		+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-

Table 8.4 continued

	Gene	resD	O2R_89	O2R_90	O2R_91	O2R_92	hnh	O2R_93,	O2R_94	silE	silS	silR	silC	O2R_99	silB	silA	O2R_102	silP	O2R_104	O2R_104.2	O2R_105	tetA	tetR	O2R_108	O2R_109	O2R_110	TnpM
	Annotation	Resolvase	Hyp	Hyp	Hyp	Hyp		Hnh endonuclease	Hyp	Silver-binding protein	silS	silR	Outer membrane efflux protein	Hyp	Efflux transporter, RND family, MFP s	Heavy metal efflux pump	Hyp	Copper-translocating P-type ATPase	Hyp	Transposase	Hyp	Tetracycline efflux protein	Tetracycline repressor protein	Relaxase	Tn1721 transposase	Resolvase	TnpM
767		+	-	-	+	+	-		+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
693	Cluster 3	+	-	-	+	+	+		+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-
700		+	-	-	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-
713		+	-	-	+	+	+		+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-
714		+	-	-	+	+	+		+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-
715		+	-	-	+	+	+		+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-
716	Cluster 4	+	-	-	+	+	-		+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-

Hyp; Hypothetical protein

Other antibiotic resistance genes including *folA*, *aad5* and *bla* were not found in any of the *C. sakazakii* genomes analysed in the present study.

Similarly, the plasmid PECOS88, a homologue of the pAPEC-O2-R plasmid in *Escherichia coli* strain S88 which according to Peigne and colleagues (2009) was linked with high level of bacteraemia in a neonatal meningitic rat model was also aligned against all *C. sakazakii* genomes, however whole length homologue was not detected in any of the genome (Figure 8.10).

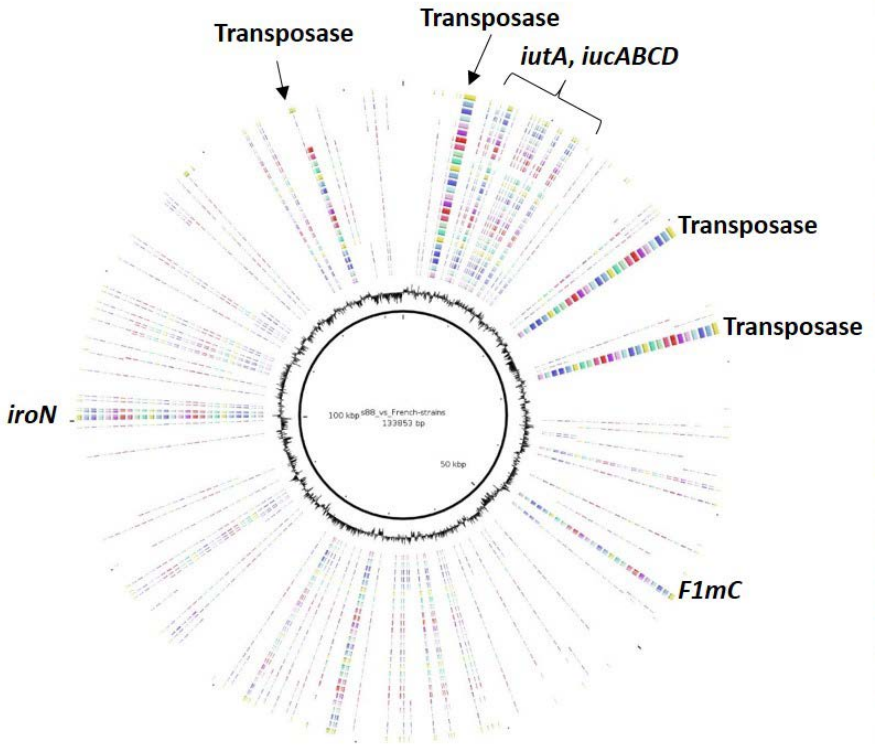


Figure 8. 10 Comparative genome image showing the alignment of PECOS88 plasmid against *C. sakazakii* genomes

The image was generated through BRIG. The central black solid circle indicates the plasmid PECOS88 (~133 kb). The distance between each dot is 10 kb. Second black spiked circle from inside represents GC contents. Each of the coloured rings represent a different *C. sakazakii* genome, the colour coding is shown on the colour legend on the right of the figure. The solid colour rings are indicative of high BLAST homology with the reference while fade and broken rings represent low degree of homology. Rings from inside out represent isolates 690, 696, 699, 703, 708, 691, 692, 694, 695, 698, 701, 702, 705, 706, 707, 709, 711, 712, 730, 767, 693, 700, 713, 714, 715 and 716 respectively.

The putative virulence genes located on the PECOS88 included three different iron uptake systems; salmochelin (*iroBCDEN*), aerobactin (*iucABCD* and *iutA*) and the *sitABCD* genes (Johnson et al. 2006; Sabri et al. 2006). Additional virulence associated traits included an *iss* gene involved in compliment resistance by providing increased serum survival (Chuba et al. 1989). Type 1 secretion system (T1SS) genes *etsABC* (Johnson et al. 2006), an outer membrane protease *ompT* (Stumpe et al. 1998) and a hemolysin encoding gene *hlyF* (Morales et al. 2004). However, the BRIG alignment showed that the whole length homologue of the plasmid PECOS88 was not present in any of the *C. sakazakii* genome analysed in this chapter. Further investigation by WebACT and *Cronobacter* indicated that except for the homologues of iron uptake aerobactin (*iucABCD* and *iutA*) which was detected in most of the *C. sakazakii* genomes, none of the other homologues were detected in *C. sakazakii* genomes; additionally, several homologues of transposes were found in most of the *C. sakazakii* genomes (Table 8.5).

Table 8. 5 BLAST investigation of the homologues of the PECOS88 genes in *C. sakazakii* isolates of the French outbreak 1994

	Gene	pECS88_0035 InsA	flmC C	pECS88_0025 insB	pECS88_0036 insB	iutA	iucA	iucB	iucC	iucD	pECS88_0007 InsB	pECS88_0008 InsA	iroN	PECS88_0149
	Annotation	Transposase	putative F-plasmid maintenance protein	fragment of Transposase	Transposase	Ferric aerobactin receptor precursor	Aerobactin siderophore biosynthesis protein	N(6)-hydroxylysine acetylase	Aerobactin siderophore biosynthesis protein	L-lysine 6-monooxigenase	Transposase	Transposase	salmochelin siderophore receptor	Putative transposase
***C. sakazakii* Isolate**														
690	**Cluster 1 (CC12)**	+	+	+	+	+	+	+	+	+	+	+	+	+
699		+	+	+	+	+	+	+	+	+	+	+	+	+
703		+	+	+	+	+	+	+	+	+	+	+	+	+
708		+	+	+	+	+	+	+	+	+	+	+	+	+
696		+	+	-	-	+	+	+	+	+	-	-	+	+
691	**Cluster 2 (CC4)**	+	+	+	+	+	-	+	+	+	+	+	+	+
692		+	+	+	+	+	+	+	+	+	+	+	+	+
694		+	+	+	+	+	+	+	+	+	+	+	+	+
695		+	+	+	+	+	+	+	+	+	+	+	+	+
698		+	+	+	+	+	+	+	+	+	+	+	+	+
701		+	+	+	+	+	-	+	+	+	+	+	-	+
702		+	+	+	+	+	+	+	+	+	+	+	-	+
705		+	+	+	+	+	+	+	+	+	+	+	-	+
706		+	+	+	+	+	+	+	+	+	+	+	+	+
707		+	+	+	+	+	+	+	+	+	+	+	+	+
709		+	+	+	+	+	+	+	+	+	+	+	+	+

Table 8.5 continued

	Gene	pECS88_0035 InsA	flmC C	pECS88_0025 InsB	pECS88_0036 InsB	iutA	iucA	iucB	iucC	iucD	pECS88_0007 InsB	pECS88_0008 InsA	iroN	PECS88_0149
	Annotation	Transposase	putative F-plasmid maintenance protein	fragment of Transposase	Transposase	Ferric aerobactin receptor precusor	Aerobactin siderophore biosynthesis protein	N(6)-hydroxylysine acetylase	Aerobactin siderophore biosynthesis protein	L-lysine 6-monooxigenase	Transposase	Transposase	salmochelin siderophore receptor	Putative transposase
711		+	+	+	+	+	+	+	+	+	+	+	+	+
712		+	+	+	+	-	-	-	-	-	+	+	+	+
730		+	+	+	+	+	+	+	+	+	+	+	+	+
767		+	+	-	-	+	+	+	+	+	-	+	+	+
693	Cluster 3 (CC13)	+	+	+	+	+	+	+	+	+	+	+	+	-
700		+	+	+	+	+	+	+	+	+	+	+	+	-
713		+	+	+	+	+	+	+	+	+	+	+	+	-
714		+	+	+	+	+	+	+	+	+	+	+	+	-
715		+	+	+	+	+	+	+	+	+	+	+	+	-
716	Cluster 4 (CC14)	+	+	+	+	+	+	+	+	+	+	+	+	-

Thus the alignment of the plasmid PECOS88 with *C. sakazakii* genomes revealed that except for a putative transposase (pECS88_0149) which was absent specifically from the cluster 3 and 4 strains, no specific pattern for the presence or absence of the genes was observed which could be related with a particular cluster of *C. sakazakii* strain being analysed in the current study.

In summary, the comparative genome analysis by BRIG alignment against virulence associated plasmids could not identify any particular virulence associated trait which could form the basis of differentiation between different clusters.

8.6 GENERAL DISCUSSION AND CONCLUSION

The present study was aimed at the genomic dissection of the 1994 French *C. sakazakii* outbreak which lasted over 3 months. Whole genome phylogeny indicated at least 4 distinct clusters within the sequenced *C. sakazakii* strains. Although, Single Nucleotide Polymorphism has proven to be more discriminatory than the current epidemiological investigation methods like PFGE, it does re-emphasise the usefulness of PFGE. Genomic examination at the SNP level across different clusters authenticated that the strains in different clusters were completely un-related. Moreover, SNP analysis was undertaken to analyse the relatedness within each cluster independently of the other clusters. SNP analysis of the cluster 1, 2 and 3 revealed genetic diversity within each cluster; although some of the strains were highly related and likely to share a common ancestor, there were diverse strains within each cluster. SNP typing in combination with SNP phylogeny suggested PIF to be the possible origin for at least 40% of the cluster 1, 73% of the cluster 2 and 100% of the cluster 3 sequenced strains. The possible origin for the remaining strains in these clusters remains unclear. An attempt was made to link the remaining isolates of cluster 1 PIF isolates from cluster 2 and 3, however SNP typing revealed that these strains were highly divergent from PIF isolates of cluster 2 and 3. Therefore present genomic level analysis of the isolates from the 1994 French NICU outbreak suggested that there were at least 3 independent but parallel outbreaks of *C. sakazakii* which lasted for 3 months with the possibility of a 4^{th} outbreak as the cluster 4 strains belong to an unopened can of PIF; the cluster 4 isolate 716 belongs to CC1 which has frequently been isolated from the infant formula and also from the clinical sources. Furthermore, SNP analysis in this study suggested that all of these 3 outbreaks (clusters 1, 2 and 3) to some extent were associated with the consumption of contaminated PIF; however not all of the isolates within cluster 1 and 2 could be linked with PIF. The origin of the remaining isolates in cluster 1 and 2 remains unclear as the ability of *C. sakazakii* to resist dry and desiccated conditions allows its survival not only in the PIF but wide range of environmental sources including water (Kucerova et al. 2011).

Another important observation in this study was that different strains from different cluster groups of *C. sakazakii* colonised the same baby indicating the ability of *C. sakazakii* strains to co-exist. For instance, 4 isolates were recovered from baby B at different time points during the outbreaks; 3 of these isolates belonged to cluster 2 (CC4) while one isolate

belonged to cluster 1 (ST12). These isolates from two different clusters were highly divergent as indicated by SNP differences (>70,000 SNPs) (data not shown) suggesting that they were acquired independently of each other. This was also the case with babies C and D which were colonised and infected by highly divergent (>70,000 SNPs) clusters 1 and 2 of *C. sakazakii*. The isolation of distant *C. sakazakii* strains suggests the ability of *C. sakazakii* to co-exist, whether co-existence of *C. sakazakii* is linked with neonatal infections warrants further investigation; however the present observation did not support this as baby C who was colonised by isolates 708 and 709 of cluster 1 and 2 respectively, was asymptomatic.

In the study presented in this chapter, the day of isolation of a *C. sakazakii* strain with the number of SNPs within the same outbreak could not be linked, which is contrary to other observations where the time of isolation and number of SNPs were correlated such as in the case of methicillin-resistant *Staphylococcus aureus* (MRSA) (Harris et al. 2010). This could be due to the differences in the epidemiology between these organisms, MRSA spreads between humans while *C. sakazakii* is ubiquitous in the environment and can survive and persist in non-human environments such as PIF, water and on solid surfaces (Kucerova et al. 2011). The observations in the present study were similar to the observations by Leekitcharoenphon et al. (2014) who studied different *Salmonella* outbreaks using whole genome sequencing data (Leekitcharoenphon et al. 2014).

Another important observation was the association of the cluster 2 strains with all three casualties during the outbreak. More importantly all of the cluster 2 strains belong to clonal complex 4 (CC4) of the *Cronobacter* MLST scheme. The investigation was taken further to profile all the cluster 2 or CC4 strains for the presence of any extended spectrum beta lactamases using the disc diffusion method (MASTDISC^{TM}), however all of the strains were found to be ESβL sensitive (Table 8.3). Similarly BLAST searching the genomes also did not indicate any ESβL associated trait in *C. sakazakii* genomes analysed in this study.

Plasmids are important sources for the dissemination of the virulence associated traits amongst bacterial species. Homologue of the pAPEC-O2-R plasmid called PECOS88 carrying antibiotic and metal resistance genes has been detected in the *E. coli* strain s88 (O45:K1:H7) which was linked with high level of bacteraemia in neonatal meningitis rat model (Peigne et al. 2009). *E. coli* strain s88 (O45:K1:H7) is a representative of the French clone (O45:K1:H7) which is associated with one third of the *E. coli* neonatal meningitis (ECNM), the strain was

isolated from the CSF of a new-born baby in Robert Debré Hospital, Paris, France (Bonacorsi et al. 2003). Each of the plasmids pAPEC-O2-R and PECOS88 were aligned independently with all the sequenced *C. sakazakii* isolates. However, no complete homologue for either of these plasmids was detected in any of the *C. sakazakii* genomes. An approximately ~12 kb region of the pAPEC-O2-R was found in all of the isolates consisting mainly of *silABCEPRS* associated with silver and copper efflux and binding protein; this region has previously been illustrated in *Cronobacter* (Kucerova et al. 2010; Joseph et al. 2012b) and analysed in this thesis (section 3.2.4.9). On the other hand PECOS88 alignment only indicated the presence of some transposases and partial homologues of aerobactin associated traits (*iucABCD* and *iutA*) which also have been described previously in *Cronobacter* (Kucerova et al. 2010; Joseph et al. 2012) (section 3.4.2.10). None of the virulence associated traits encoded on these plasmids such as *tetA*, *tetR*, *folA* (trimethoprim resistance), *aad5* (aminoglycoside resistance) and bla (beta lactamase) encoded on pAPEC-O2-R and T1SS, *iss* (compliment resistance), *hlyF* (haemolysin) encoded on PECOS88 plasmid was detected in any of the *C. sakazakii* isolate from the 1994 French outbreak.

In conclusion, the present study indicated PIF to be a major but not the exclusive source for the transmission of *C. sakazakii* to neonates in the 1994 French NICU outbreak. Improve practices in the manufacturing and preparation in addition to improved personal hygiene are essential to prevent the growth and transmission of this neonatal health associated pathogen.

9 CHAPTER 9

OVERVIEW, CONCLUSIONS AND FUTURE DIRECTIONS

9.1 OVERVIEW

Cronobacter sakazakii is a member of the genus *Cronobacter* and came to prominence for its association with neonatal infections which include bacteraemia, septicaemia and neonatal meningitis (Bar-Oz et al., 2001; Mullane et al., 2008; Joseph & Forsythe, 2011; Joseph et al. 2012a). However, it is one lineage of *C. sakazakii* known as Clonal Complex 4 (CC4) which is of deep concern as it has been predominantly associated with neonatal meningitis (Joseph & Forsythe, 2011; Hariri et al. 2013; Forsythe et al. 2014). Not only does *C. sakazakii* CC4 dominantly associated with neonatal infections, it has been isolated frequently from the environment, more importantly from powdered infant formula (PIF) and associated production environments such as floors, bays, roller dryers, tankers and air filters (Neelam et al. 1987; Mosso et al. 1994; Van Os et al. 1996; Hein et al. 2009; Craven et al. 2010; Jacobs et al. 2011; Muller et al. 2013; Power et al. 2013; Sonbol et al. 2013).

C. sakazakii CC4 represents the dominant clonal complex of the *Cronobacter* and represents the single or double loci variants of ST4 (Figure 3.1). ST4 is the dominant sequence type representing at least 91% of the total CC4 isolates in *Cronobacter* PubMLST (last checked; 2nd December 2014). Large number of CC4 isolates have clinical origin (Figure 3.2). A number of neonatal fatalities have been associated with the infection of *C. sakazakii* CC4, for example in the French outbreak of 1994, each of 3 babies who died were infected by *C. sakazakii* CC4 (Caubilla-Barron et al. 2007). The predominant association of *C. sakazakii* CC4 with neonatal infections and its predominant occurrence in the environment warranted in depth investigation. Although a number of studies have described virulence associated genes in *C. sakazakii*, none of the studies was aimed specifically to explain the association of *C. sakazakii* CC4 with neonatal infections and to explain the factors behind its uniqueness. Therefore, it was of great significance and it raised certain questions;

Is *C. sakazakii* CC4 a clonal lineage?

What makes *C. sakazakii* CC4 unique compared to other lineages of *C. sakazakii*?

Are there any virulence associated genes specific to *C. sakazakii* CC4?

Are there any environmental fitness associated genes specific to *C. sakazakii* CC4?

How diverse is *C. sakazakii* CC4?

To answer all these question was important due to the direct association of *C. sakazakii* CC4 with neonatal health.

Next generation sequencing (NGS) or massively parallel sequencing has revolutionised genetics and genomics. The availability of high throughput instruments and rapid developments in analysis programs has a huge impact on research in the field of genomics. The development of the NGS technology has contributed a wealth of information in recent years and continues to bring enormous change in our understanding of the basic biological knowledge, genetics and biological research (Mardis, 2008, 2011). The constant and rapid development of NGS has exponentially dropped the cost of genome sequencing and nowadays the laboratories with modest research budget can afford to sequence small to large size genomes. The present study exploited this opportunity and a total of 50 genomes were sequenced in three different phases of the current PhD study (Table 2.1). The study was initiated with 30 *C. sakazakii* strains while 26 *C. sakazakii* isolates were sequenced only for the 1994 French outbreak analysis.

The three main aims of the present study were;

- Comparative genomic analysis for the identification of the CC4 specific traits.
- Analyse the diversity within CC4.
- Case study: Genome analysis of the isolates obtained from the French outbreak (1994) with the main aim of source identification.

9.2 COMPARATIVE GENOMICS TO IDENTIFY CC4 SPECIFIC TRAITS

The strategy used for the identification of unique traits within *C. sakazakii* CC4 can be divided into 4 parts. The genome screening for the traits which previously have been described in literature as associated with either virulence or environmental fitness of *Cronobacter*, core CC4 genome analysis, gene by gene search using ACT and fragmented comparative genomic analysis using Gegenees.

9.2.1 GENOME SCREENING FOR VIRULENCE AND ENVIRONMENTAL FITNESS ASSOCIATED TRAITS

9.2.1.1 CLONALITY OF *C. SAKAZAKII* CC4

Using genome sequences of 30 *C. sakazakii* isolates, whole genome phylogeny was constructed (Figure 3.3). The results of the phylogenetic analysis were quite interesting as 18 CC4 isolates formed a distinct cluster on the phylogenetic tree. It was interesting to note that 18 CC4 isolates were temporally and geographically diverse, isolated over 60 years from 9 different countries across the globe. This unique clustering of the CC4 isolates was quite intriguing and prompted detailed genomic analysis to reveal which genomic traits are specifically found in CC4 conferring a clonal signature to this neonatal meningitic associated clade. The genome interrogation was undertaken using different strategies for the identification of unique traits within CC4. The first part of the study was to screen the traits which have previously been indicated in the literature to be associated with either virulence or environmental fitness in *C. sakazakii* or other *Cronobacter* species. The aim was to observe how these traits vary for their presence across CC4 and non-CC4 groups and to check if there is any trait which is specific to CC4 which could explain its clonal nature, its predominant association with neonatal infections especially meningitis and its frequent isolation from the environment especially PIF. For this purpose both virulence associated traits as well as environmental fitness associated traits were screened in 30 *C. sakazakii* genomes (Table 3.1).

9.2.1.2 VIRULENCE ASSOCIATED GENE SCREENING

Generally a bacterium has to cross several barriers before it can successfully establish an infection in its host (Wilson, 2002). A large number of *C. sakazakii* strains, especially CC4 have been isolated from CSF of neonates with meningitis (Joseph & Forsythe, 2011; Hariri

et al. 2013). To survive within the host, the bacteria should have the ability to adopt to the survival conditions within the host environment and should have underlying genetic machinery to cause infection. In order to reach a neonates' brain, *C. sakazakii* would have to overcome several host defence mechanisms which include the acidic environment of stomach, attachment and invasion of the intestinal epithelium to reach blood circulation, to resist the complement attack, survive within macrophages and cross the blood brain barrier (BBB) to get access to the brain (Wilson, 2002). The increased association of *C. sakazakii* CC4 with neonatal meningitis and its frequent isolation from clinical sources especially CSF warranted the investigation at the genomic level to reveal the virulence associated genes found uniquely in this clade.

The first step for a bacterial pathogen to establish an infection is their ability to successfully colonise their host; it is considered an important mechanism for bacterial pathogens to attach to and colonise the surfaces such as the epithelial lining of the human stomach (Soto & Hultgren, 1999). Fimbriae are present predominantly in Gram negative bacteria helping them to adhere to the host surfaces which could ultimately lead to the infection. At least 10 different fimbriae associated gene clusters which have previously been described in *Cronobacter* (Kucerova et al. 2010; Joseph et al. 2012b) were screened in 30 *C. sakazakii* genomes. The analysis indicated that clusters 2-7 and 9 (homologues of fimbrial genes in *C. sakazakii* BAA-894) were noted in most of the CC4 and non-CC4 isolates while the remaining 3 clusters, fimbrial region 1, fimbrial region 8 (π fimbriae homologues of *C. turicensis* z3032) and fimbrial region 10 (curli fimbriae homologues of *C. turicensis* z3032) showed greater variation for their presence across two subsets. None of these showed specific association for its presence either in CC4 or non-CC4 group. In addition, the homologues of the type 1 fimbrial operon (*fimABCDEFGHI*) of *E. coli str. K-12 substr. MG1655* were not detected in any of the *C. sakazakii* genome (section 3.4.2.1).

Invasion, like adhesion, is also considered crucial for bacterial pathogens for the establishment of successful infection (Badger et al. 1999). Several studies have established the role of outer membrane protein A (ompA) towards the invasion of the neonatal meningitic *E. coli* K1 as well as *C. sakazakii* (Badger et al. 2000; Mohan et al. 2009; Kim et al. 2010). A study by Kim and Wang (2002) indicated that a mutant of *E. coli* K1 strain RS218 (018:K1:H7) in which the *ompA* gene was deleted, showed less invasion on Brain Microvascular Endothelial Cells (BMEC) when using *in vitro* invasion assays (Wang & Kim,

2002). Similarly Nair et al. (2009) showed that ompA is a major fibronectin binding protein in *C. sakazakii* and is essential for the invasion of the BMEC. Their study showed that the *C. sakazakii* mutants in which the *ompA* gene was knocked out exhibited significantly reduced invasion of the BMEC *in vitro*, indicating that *ompA* is required for the invasion and therefore the pathogenicity of *C. sakazakii* (Mohan et al. 2009). Another study by Koebnik (1999) showed that all the loops of the ompA protein were essential for its functioning (Koebnik, 1999). It has also been proposed that ompA and ompX proteins in *C. sakazakii* are important and act synergistically towards the invasion of Human Brain Microvascular Cells (HBMEC) (Kim et al. 2010). It has been suggested that ompA along with ompX protein is involved in the basolateral invasion of *C. sakazakii*; the study showed that the mutants lacking *ompA* and *ompX* genes were significantly reduced in their ability to invade the human enterocyte-like epithelial Caco-2 and human intestinal epithelial INT-407 cell lines (Kim et al. 2010).

The genome screening of sequenced *C. sakazakii* isolates indicated that *ompA* and *ompX* were noted in 100% of CC4 and non-CC4 genomes. Due to the significance of *ompA* for the invasion of *C. sakazakii* which has been established in the previous studies, it was important to analyse the *ompA* gene in greater detail in order to observe any variations at the nucleotide and amino acid level which are unique to CC4. The *ompA* gene sequences of CC4 and non-CC4 isolates were aligned and phylogeny determined; the resulting phylogeny was interesting as all except 2 CC4 isolates formed a unique cluster on the phylogenetic tree (Figure 3.6; section 3.4.2.3). The analysis of the gene sequence alignment revealed a very low sequence diversity (12 SNPs) across CC4 and non-CC4 isolates for the *ompA* (Table 3.2). The conversion of the nucleotide sequence into amino acid sequence *in silico* indicated that all of these SNPs were synonymous and hence were unlikely to confer any structural or functional changes to ompA. Hence, it is unlikely that variation in the ompA gene sequence may give a competitive advantage to CC4 isolates. The results indicated the *ompA* gene is equally important for both CC4 and non-CC4 isolates.

Another gene associated with invasion is *Inv* encoding for invasin protein and has been linked with pathogenesis of *Salmonella* and *Yersinia* (Pepe & Miller, 1990). It has been revealed that in *C. sakazakii* 29544, a putative Inv (invasion protein) encoded by *Inv* gene is essential for the basolateral invasion of Caco-2 cells and acts synergistically with ompA (Chandrapala et al. 2014). The genome screening of sequenced *C. sakazakii* isolates showed that *Inv* gene was absent from 6 CC4 and 7 non-CC4 isolates (section 3.4.2.2). The absence

of *Inv* gene from clinically important strain like a CSF isolate 721 weakens the probability that this gene is essential for the invasion of *C. sakazakii* in general and CC4 in particular (section 3.4.2.2).

Filamentous haemagglutinin (FHA) is the large, rod shaped protein (>200 kDa) which is expressed by bacteria such as *Bordetella pertussis* (Makhov et al. 1994). The FHA facilitates bacterial adhesion to epithelial cells and macrophages *in vitro* and is involved in colonisation of trachea *in vivo* (Julio & Cotter, 2005). A 27 kb FHA locus, reported by Franco et al. (Franco et al. 2011a) located on the pCTU1 plasmid of *C. turicensis* z3032, was interrogated in the sequenced genomes (section 3.4.2.4), however none of the CC4 or non-CC4 isolates showed the presence of the complete locus indicating that this trait may not be crucial for the virulence of *C. sakazakii* in general or CC4 in particular.

Macrophages are the immune cells which provide the first line of defence against bacterial pathogens to invade the internal cells. Once engulfed by the macrophages, the bacterial cells are enclosed in a vacuole called phagosome which then fuses with lysosome and bacteria get digested (Ernst et al. 1999). Bacteria have adapted mechanisms to avoid or survive within the hostile environment of macrophages (Pizarro-Cerda et al. 1996). A macrophage survival associated gene *sodA* encodes for superoxide dismutase in *C. sakazakii* 658 (Kucerova et al. 2010). The *sodA* gene was noted invariably across two subsets of the sequenced genomes (section 3.4.2.5). Similarly, zinc metalloprotease *zpx* associated with rounding of the CHO cells (Kothary et al. 2007) was noted in most of the isolates across CC4 and non-CC4. The results of the protease activity assays by Hana Sonbol indicated that all of the tested isolates (15 CC4 and 8 non-CC4) showed protease activity irrespective of their clonal complex (section 3.4.2.5).

Evasion of the bactericidal effects of serum is a common characteristic feature of the invasive isolates of several species such as *Streptococcus* (Hol et al. 1995; Hoe et al. 1999; Williams et al. 2001). A plasmid encoded plasminogen activator (cpa) which has been linked with resistance against bactericidal activity of serum in *C. sakazakii* was noted only in the isolates which showed the presence of the full length pESA3 plasmid. It was not detected in isolates which lack either full length or partial length plasmid including CC4 isolate 6 and non-CC4 isolates 1, 680, 520 and ES15. Although it is an important trait, it does not form the basis of differentiation between two groups. The parallel laboratory study by Sumyya Hariri

using 16 CC4 and 6 non-CC4 sequenced isolates indicated that in *C. sakazakii* serum resistance is strongly related with the presence of plasmid pESA3 as plasmid less isolates 6 and 520 were sensitive to serum (section 3.4.2.6: Sumyya Hariri, personal communications).

Lysis of red blood cells known as haemolysis is an important pathogenic mechanism employed by certain bacteria such as *E. coli* (Wong et al. 2012). A number of haemolysin associated traits (ESA_00102, ESA_00432, ESA_00643, ESA_02810, ESA_02937 and ESA_03540) which have been reported previously (Joseph et al. 2012b) were searched for homology in the sequenced isolates and were noted uniformly across all of the isolates. The parallel laboratory study by Hana Sonbol showed that all 4 isolates (701, 680, 696 and 658) tested for haemolysis activity were positive on both horse and sheep blood (section 3.4.2.7).

In *E. coli* K1, Type 6 Secretion System (T6SS), which is a newly described bacterial secretion system has been linked with the invasion of blood brain barrier (Zhou et al. 2012). At least 10 different T6SS associated gene clusters were indicated previously by Joseph et al. (2012b). The genome interrogation of the sequenced isolates indicated that none of isolates contained any of these cluster as a complete unit, rather a mosaic pattern for the presence of the genes within these clusters was noticed (Table S1). The analysis suggested that T6SS may not be essential for the virulence of *C. sakazakii* (section 3.4.2.8).

In trace amounts, toxic metals such as copper are essential for the bacterial cells for the metabolic and growth associated cellular processes. However, their presence in excess can be toxic to bacterial cells. For example, copper is an important micronutrient which is required for the survival of the bacteria as it participates in the redox reactions by acting as a co-factor for the enzymes. However, due to its high chemical reactivity, the accumulation of copper can be a hazard for the bacteria. Hence bacteria have evolved mechanisms to efflux toxic metals when exposed to high concentrations (Arguello et al. 2013). The genomes of 30 *C. sakazakii* isolates were screened for the presence of two copper and silver resistance associated regions ESA_04236-45 (*cusESRCFBA/silEP)* and ESA_04248–55 (*pcoABCERS*) (section 3.4.2.9). Genome analysis of CC4 isolate 377 indicated that *cus/sil* and *pco* genes were located together on its chromosome (Figure S3), however no CC4 specific pattern was observed as both regions indicated a variable pattern of presence across two groups. However, it was important to note that CC4 and non-CC4 isolates of clinically significant origin such as 701, 767, 721, 1221 and 1249 all of which were CSF isolates lacked

either one or both of these copper and silver resistance associated regions which questions their significance towards the virulence of *C. sakazakii* in general and CC4 in particular (section 3.4.2.9).

Iron is essential for all forms of life including bacterial pathogens. It is involved in in such crucial biological activities as DNA synthesis and energy generation. Vertebrate hosts use a mechanism called nutritional immunity to sequester iron from invading bacterial pathogens. Once within the vertebrate host, bacterial pathogens need iron to maintain their growth and survival, where they encounter iron limiting conditions. In order to combat host iron sequestration, most bacterial pathogens have generally developed three iron uptake mechanisms; siderophore based iron uptake, heme acquisition system and lactoferrin/transferrin receptors (Kehl-Fie and Skaar, 2009; Skaar, 2010) (section 3.5.2.10).

The interrogation of the sequenced isolates for the enterobactin synthesis and transport system, plasmid-borne operon for aerobactin synthesis (*iucABCD*) and its receptor *iutA* a siderophore synthesis which have previously been reported (Joseph et al. 2012b) indicated that all of these regions were present predominantly across both CC4 and non-CC4 groups. The exception was plasmid less isolates 6, 520 and ES15 which lacked the plasmid-borne aerobactin synthesis and its receptor (*iucABCD*, *iutA*) (section 3.4.2.10). The analysis indicated that iron acquisition associated genes do not give any competitive advantage to CC4 over non-CC4 and that iron acquisition machinery has evolved efficiently and commonly across *C. sakazakii*.

Another important virulence associated mechanism is the ability of the bacteria to exploit alternative carbon and energy sources such as sialic acid. Gram negative bacteria such as neonatal meningitic *E. coli* K1 have evolved the ability to utilise sialic acid (Wang, 2009). Sialic acid is present in its conjugated form in the human brain, therefore it may provide bacterial pathogens such as *C. sakazakii* with an alternative carbon and energy source (Wang, 2009). Since *C. sakazakii* has been associated with neonatal meningitis and frequently been isolated from the CSF, it was important to screen the sequenced isolates for the presence of any genes which have been associated with sialic acid utilisation (Joseph et al. 2012b; Joseph et al. 2013). Previous studies by our group has shown that amongst *Cronobacter*, *C. sakazakii* has the necessary genes and the ability to utilise sialic acid as an energy source (Joseph et al. 2012b; Joseph et al. 2013). Parallel laboratory studies by

Sumyya Hariri has shown that all of the tested *C. sakazakii* isolates were able to grow on ganglioside GM1 media (Joseph et al. 2013). The sialic acid utilisation genes *nanAKTR*, *nanC, nagAB, neuC* and *siaPQM* were screened in the sequenced isolates in order to observe any variation cross CC4 and non-CC4. However, as expected, all of the sialic acid genes were noted uniformly across the two groups indicating that all of the *C. sakazakii* isolates irrespective of which clonal complex or sequence type they belong, have acquired the ability to utilise sialic acid (section 3.4.2.11). Recent research in our group indicates that the sialic acid utilisation and the sialic acid genes may not be exclusive to *C. sakazakii* as isolates of other *Cronobacter* species are giving a hint for the presence of sialic acid genes (Sumyya Hariri; personal communications, unpublished data).

In summary, the genome screening for the above mentioned potentially virulence associated genes did not indicate a clear CC4 specific pattern for any of the trait. Most of the traits analysed indicated a variable pattern for the presence of these genes across both groups. Hence, based on these traits, no significant variation in the virulence potential of CC4 and non-CC4 could be deduced.

9.2.1.3 GENOME SCREENING FOR THE ENVIRONMENTAL FITNESS ASSOCIATED TRAITS

The analysis was taken further for screening the genetic traits which have been proposed to be associated with environmental fitness of *C. sakazakii* in previous studies. It is important to mention that all of the sequenced isolates contained genes associated with stress response (data not shown). Some of the environmental fitness associated traits which were screened in the sequenced isolates included *rpoS*, thermotolerance, acid, osmotic stress, biofilm and capsule formation (colanic acid synthesis) associated genes.

In Gram-negative bacteria the *rpoS* gene encoding for the alternative sigma factor σ^S regulates a large regulon which modulates the expression of ~10% of the genome, including genes associated with general stress response (Weber et al. 2005). In *E. coli* and *Salmonella* spp. the *rpoS* gene has been regarded as highly mutable (Robbe-Saule et al. 2003; Bhagwat et al. 2006; Dong et al. 2009). Moreover in *E. coli*, the level of expression of the *rpoS* gene has been linked with variation in stress response (Robey et al. 2001; Chiang et al. 2011). A study by Ordonez et al. (2012) elucidated the role of the *rpoS* gene in *C. sakazakii* indicating that full length gene sequence is important for the bacterial response against acidic, alkaline and osmotic stress (Ordonez et al. 2012). The presence and

variability of the *rpoS* gene in sequenced *C. sakazakii* isolates was analysed (3.4.3.2). The gene was detected uniformly in all of the isolates. Although, CC4 isolates did indicate distinct clustering on the phylogenetic tree (Figure 3.14), the alignment of *rpoS* indicated a very low sequence diversity with 17 SNPs distributed across 30 sequenced isolates (Figure 3.15, Table 3.3). This high sequence conservation was unlikely to cause significant changes in the stress response behaviour of the CC4 compared to non-CC4 isolates. This observation was supported by the stress response assays such as dry heat stress response, osmotolerance and acid resistance assays where no significant variation was observed between CC4 and non-CC4 (Hana Sonbol; personal communications). The analysis indicated that although *rpoS* might be important for the general stress response in *C. sakazakii*, it does not vary significantly between CC4 and non-CC4 and therefore unlikely to affect the stress response behaviour between the 2 groups.

The ability to resist the acidic conditions of the stomach is important for the bacterial pathogen for its growth and survival. The *ompR* gene which has been associated with acid resistance in *C. sakazakii* (Alvarez-Ordonez et al. 2014) was detected in all of the sequenced isolates, both CC4 and non-CC4. Furthermore, the acid resistance assays performed by Hana Sonbol indicated no significant variation in the ability of *C. sakazakii* CC4 and non-CC4 isolates to grow at pH 3.5 (section 3.4.3.4).

Another important characteristic of the bacterial cells is their ability to form biofilms on the biotic and abiotic surfaces. The biofilm can provide competitive advantage to bacterial cells and can make them resistant to desiccation, disinfectants, antibiotics, extreme temperatures as well as help them survive in environments with low pH such as stomach (Scher et al., 2005; Wood et al., 2006; Dancer et al., 2009; Osaili & Forsyhte 2009). However, the genome interrogation of the sequenced isolates did not indicate any variation for the presence of the biofilms associated genes. All of the sequenced isolates irrespective of the clonal complex or sequence type, harboured biofilm formation genes *bcsBEFZ* homologues of the earliest *C. sakazakii* CC4 isolate 377 (Masood et al. 2013a) as well as the homologues of hypothetical proteins ESA_00281 and ESA_00282 suggested by Hartmann et al. (2010) to be involved in biofilm formation (section 3.4.3.5). Previous study using majority of the sequenced isolates analysed in the present study, indicated that all of these isolates were able to form biofilms on the soya, whey and casein based formulas (Nassrudin Rhouma 2012, unpublished data) (section 3.4.3.5). The colanic acid synthesis associated genes cluster

(ESA_01155-01175; *wzABCKM*) was noted in most of the CC4 and non-CC4 isolates. Moreover, as per previous PhD study, the capsule production was not found to be specific to CC4 as most tested *C. sakazakii* CC4 and non-CC4 isolates were able to produce capsule on the milk agar (Nassrudin Rhouma 2012, unpublished data).

The frequent isolation of *C. sakazakii* CC4 from the PIF and from the dry and desiccated environment indicates its ability to survive in conditions with very low water activity. A number of genes associated with osmotolerance in *C. sakazakii* were proposed by Feeney et al. (2014). The Prop osmolyte uptake system homologues indicated by ESA_02131, ESA_01706, ESA_04214, ESA_pESA3p05450, ESA_01226, ESA_00673 and ESA_03328 in *C. sakazakii* BAA-894 genome (Feeney et al. 2014) were detected dominantly across both CC4 and non-CC4 groups showing no significant variation between the 2 groups except for the plasmid borne gene ESA_pESA3p05450. Similarly the *E. coli* homologues of the osmo-protectants *TrkH, trkG, trkA, trkE, kdpA, kdpB, kdpC, kdpD, kdpE, phoP, phoQ, ompC, ompF, envZ* and *ompR* (Feeney et al. 2011) were also noted uniformly across both groups except for *trkG* which showed only slight variation (section 3.4.3.6).

The increased stress tolerance of *C. sakazakii* such as thermotolerance and desiccation tolerance can enhance its survival in the environment of PIF manufacturing factories which can increase its chances of contaminating the PIF, thereby growing the risk of exposure to the neonate. Thermotolerant bacteria having the ability to adapt to sub-boiling temperatures may ultimately contaminate the pasteurized food product (Hsieh & Ren, 2001; Fyre & Donnelly, 2005; Prejit et al. 2007). A number of studies suggest the thermotolerance of *Cronobacter* either equivalent or higher than other Enterobacteriaceae (Nazarowec-White & Farber, 1997; Breeuwer et al. 2003; Iversen et al. 2004; Dancer et al. 2009). Gajdosova and colleagues (2011) identified an 18kb long region in *C. sakazakii* ATCC 29544 associated with its thermotolerance at 58^{o}C. The region consisted of 22 open reading frames (*orf*) of which *orf*HIJK was experimentally linked with thermotolerance as the cloning of these *orf*s into *E. coli* resulted in two fold increase in its D_{58} value (Gajdosova et al. 2011). The genomes of 30 *C. sakazakii* isolates were screened for the presence of 22 *orfs* suggested by Gajdosova et al. 2011. Interestingly, all of these 22 *orfs* showed high degree of variation between CC4 and non-CC4 isolates. More importantly, the *orfHIJK* were noted predominantly in the CC4 isolates (>50% isolates) than non-CC4 (<10%), showing significant variation. Nevertheless, the absence of these regions in clinically significant isolate 701 and

a “persistent thermotolerant” CC4 isolate SP291 (Power et al. 2013) questioned its significance in CC4. Additionally, the laboratory heat tolerance assays performed by Hana Sonbol on 7 CC4 and 6 non-CC4 sequenced isolates (section 3.4.3.3) did not indicate a significant difference between the 2 groups for their ability to tolerate dry heat. Hence, the analysis here suggested that although the association of *orfHIJK* with thermotolerance cannot be negated, yet it may not be the only region responsible for thermotolerance in *C. sakazakii* (section 3.4.3.3).

9.2.1.4 *IN SILICO* SEROTYPING OF SEQUENCED ISOLATES

The O-antigen is the hypervariable and the outermost region of the lipopolysaccharides (LPS) in Gram negative bacteria. The O-antigen is important for its role in the innate immunity and its variation has been used to define serotyping schemes in Gram negative bacteria (Wang et al. 2010). To date, up to 7 different O-antigen types have been described in *C. sakazakii* (Table 3.5). It was therefore of high interest to align the genetic regions associated with these serotypes to observe any variation between CC4 and non-CC4 based on O-antigen. The analysis indicated a highly interesting pattern as 17/18 CC4 isolates showed complete alignment with O:2 antigen, the only exception being isolate 1240 which aligned with O:4 (Table 3.6). However it was noticeable that O:2 antigen was not limited to CC4 as 2 non-CC4 isolates 978 and 1249 also aligned with O:2. The remaining 11 non-CC4 isolates were divided into three O-antigen types i.e. O:1, O:3 and O:4 indicating sequence type independent pattern of O-antigen in *C. sakazakii* (Table 3.6). More recent analysis of genomes in *Cronobacter* PubMLST has indicated that O:2 antigen is a dominant antigen type in *C. sakazakii* and is distributed amongst 19 different sequence types of which 16 sequence types are non-CC4 (Khaled Ibrahim, personal communications; unpublished data). Hence, although O:2 antigen was noted dominantly in CC4, yet it cannot be considered a unique signature for CC4 (section 3.4.4). Furthermore, it is distributed widely amongst *C. sakazakii* and cannot either be considered as the signature for the neonatal meningitis as a number of different isolates of different STs with non-clinical origin also belong to O:2 serotype.

9.2.1.5 *IN SILICO* PLASMID PROFILING OF THE SEQUENCED ISOLATES

The large plasmid pESA3 of *C. sakazakii* BAA-894 and pCTU1 of *C. turicensis* z3032 have been regarded as virulence plasmids (Franco et al. 2012). The virulence associated genes located on these plasmids such as *eitCBAD* and *iucABCDE/iutA* (section 3.4.2.10), *Cronobacter*

plasminogen activator (*cpa*) (section 3.4.2.6), 17 kb T6SS (section 3.4.2.8) and a filamentous haemagglutinin (FHA locus) (section 3.4.2.4) have been covered separately. In order to investigate the presence of the full length plasmids in the sequenced isolates, pESA3 plasmid was aligned with 29 sequenced isolates (excluding *C. sakazakii* BAA-894) using Blast Ring Image Generator (BRIG). The results did not show a CC4 specific variation. The plasmid was missing completely in one CC4 isolate 6 while only 90kb region of the plasmid was found in the CC4 isolate 557. The plasmid was completely absent in two non-CC4 isolates 520, ES15 and was partially present in isolates 1 and 680 (70kb). Hence the *in silico* plasmid profiling indicated only strain specific variations across CC4 and non-CC4 (section 3.4.5, Figure 3.20). A similar pattern was observed for the pCTU1 plasmid (Figure S2).

9.2.1.6 SCREENING OF THE NEONATAL MENINGITIC *ESCHERICHIA COLI* (NMEC) VIRULENCE DETERMINANTS IN *C. SAKAZAKII*

Amongst the NMEC, *E. coli* strains having K1 capsular polysaccharide are predominant and represent almost 80% of the NMEC associated cases (Glode et al. 1977). The *E. coli* K1 is associated predominantly with neonatal meningitis which is induced by a complicated multistage process (Kim, 2001, 2002). It was therefore of significance to screen the sequenced *C. sakazakii* isolates for potential virulence associated determinants of *E. coli* K1 for the presence of the homologues. The traits which were screened for in sequenced *C. sakazakii* isolates included ibe proteins, *Asla*, *TraJ* and *cnf1* all of which have been associated with virulence in *E. coli* K1 (Bonacorsi & Bingen, 2005).

The *aslA* (encoding for the arylsulfatase enzymes which contain highly conserved sulfatase motifs) has been linked with invasion of HBMEC (Hoffman et al. 2000). Similarly the *TraJ* (conjugal transfer transcriptional regulator) has been associated with macrophage survival and invasion of *E. coli* K1 (Hill et al. 2003). Moreover, the *cnf*1 (cytotoxic necrotising factor 1) is linked with the invasion of BMEC and penetration of *E. coli* K1 to the central nervous system (Khan et al. 2002). The genome analysis indicated that homologue of none of these genes was present in any of the 30 sequenced *C. sakazakii* isolates (Figure 3.21, Table 3.7).

Similarly, the ibe proteins including ibeA, ibeB and IbeC have also been associated with invasion of HBMEC (Huang et al. 1999; Wang et al. 1999; Huang et al 2001, Wang & Kim, 2002). The genome screening of the sequenced *C. sakazakii* isolates indicated that except for *ibeB*, none of the homologue was detected in any of the isolates. The *ibeB* gene was

noted more in CC4 isolates (15/18) than non-CC4 (6/12). However, the BLASTx analysis of *ibeB* was indicating homology against the *cusC* gene in *C. sakazakii* which was part of the copper/silver efflux system (section 3.4.2.9). Furthermore, the trait was absent in clinically significant isolates such as 20, 721 and 1249 (Figure 3.21, Table 3.7).

The plasmid PECOS88, a homologue of the pAPEC-O2-R plasmid in *Escherichia coli* strain S88 which according to Peigne and colleagues (2009) was linked to high level of bacteraemia in a neonatal meningitic rat model was also aligned against all *C. sakazakii* genomes, however the whole length homologue was not detected in any of the genome (Figure 3.22). The putative virulence genes located on the PECOS88 included three different iron uptake systems; salmochelin (*iroBCDEN*), aerobactin (*iucABCD* and *iutA*) and the *sitABCD* genes (Johnson et al. 2006; Sabri et al. 2006). Additional virulence associated traits included an *iss* gene involved in compliment resistance by providing increased serum survival (Chuba et al. 1989), type 1 secretion system (T1SS) genes *etsABC* (Johnson et al. 2006), an outer membrane protease *ompT* (Stumpe et al. 1998) and a haemolysin encoding gene *hlyF* (Morales et al. 2004). It was curious to investigate whether any of these sequenced isolates contain the homologue of the plasmid PECOS88. The Blast Ring Image Generator (BRIG) was used to align 30 sequenced isolates against the plasmid PECOS88, however none of the isolates contained full length plasmid (Figure 3.22). The regions of homology included *iucABCD* and *iutA* which has already been covered in section 3.4.2.10 and partial homologue of the *iroN* and several transposases.

The analysis until this stage aimed to screen the virulence and environmental fitness associated traits of *Cronobacter* and *E. coli* K1 could not indicate any trait which can define clear differentiation between CC4 and non-CC4 isolates. The analysis was taken further using three different strategies i.e. core genome analysis, a gene by gene search approach and fragmented genomic analysis using Gegenees; aimed to identify traits unique to *C. sakazakii* CC4.

9.2.2 UNIQUE GENES IN *C. SAKAZAKII* CC4 CORE GENOME

The genome screening of the *C. sakazakii* isolates for the previously described virulence and environmental fitness associated traits revealed interesting variations such as O:2 antigen and thermotolerance associated genes; however none of the traits was found completely specific to CC4 which could clearly explain its predominance in neonatal infections and in the environment (chapter 3). The clonal nature of the of the *C. sakazakii* CC4 genomes (Figure 3.3) was quite intriguing and it was hypothesised that there might be some unique traits within its core genome which might explain its unique nature associated with its virulence or environmental persistence. For this purpose *C. sakazakii* CC4 core genome and *C. sakazakii* non-CC4 core genome were constructed using 18 and 12 genomes respectively (section 2.2.8.2). A relatively larger core genome was revealed for *C. sakazakii* CC4 compared to non-CC4 core genome (Figure 4.1). The large size of the CC4 core genome further strengthened its clonal nature and low sequence diversity as was revealed earlier by whole genome phylogeny (Figure 3.3).

The comparative genomic analysis was undertaken in 3 steps. Initially one *C. sakazakii* CC4 isolate 557 was compared against one *C. sakazakii* non-CC4 isolate 658 (CC1/ST1). The unique genes (total of 442) in the *C. sakazakii* CC4 isolate 557 were listed which consisted of variety of different traits including large number of hypothetical proteins and some phage associated traits (Table S2). Using a 3 way ACT, the CC4 genome 557 was aligned against *C. sakazakii* CC4 core genome and *C. sakazakii* non-CC4 core genome (Figure 4.2). Three way ACT indicated only 39 genes unique to *C. sakazakii* CC4 core genome (Table 4.1). Since the aim of the present analysis was to look for the traits which are CC4 specific, hence the pangenome of 30 *C. sakazakii* isolates (18 CC4 and 12 non-CC4, Table 3.1) was interrogated for the variation of 39 genes. This was done by manually checking each of the 39 genes. The analysis indicated that 34 of 39 genes were also noticeable in *C. sakazakii* non-CC4 isolates suggesting that although these genes were not part of the *C. sakazakii* non-CC4 core genome, yet they were present in its pangenome. Five of these 39 genes with locus_tag CSAK557_01621, CSAK557_01380, CSAK557_03949, CSAK557_03870 and CSAK557_02008 were of particular interest as they were noted predominantly in *C. sakazakii* CC4 genomes (Table 4.1). All of these genes were annotated as hypothetical protein of unknown function, the exception being CSAK557_02008 which was annotated as metallo-β-lactamase family protein.

The BLASTx analysis of the 5 hypothetical proteins also could not reveal any trait with functional importance which could be linked with virulence or environmental fitness of *C. sakazakii* CC4. Amongst these hypothetical proteins, of particular significance was CSAK557_03870 or CSSP291_18210 which was detected only in 18 *C. sakazakii* CC4 genomes while absent from all 12 *C. sakazakii* non-CC4 genomes (Table 4.1). The predominance of the hypothetical protein CSSP291_18210 in CC4 was further strengthened when the analysis was expanded to include 107 *Cronobacter* genomes in *Cronobacter* PubMLST; these genomes contain 37 *C. sakazakii* CC4 genomes while 70 non-CC4 *Cronobacter* spp. genomes which belong to 7 different species of the genus *Cronobacter* (section 4.4.4). The *Cronobacter* BLAST indicated that the hypothetical protein CSSP291_18210 was conserved across 37 *C. sakazakii* CC4 genomes while only detected in 2 of 70 *Cronobacter* spp. non-CC4 isolates; *C. sakazakii* 2051 and *C. malonaticus* 685. These 2 isolates likely represent the rare variants amongst non-CC4 which might have acquired these genes through horizontal gene transfer. Hence, the hypothetical protein CSSP291_18210 seemed likely to be CC4 specific trait as it was not only found to be conserved in the *C. sakazakii* core CC4 genome but also predominant in the CC4 pangenome when the analysis was expanded (section 4.4.4). It is also important to point out that *C. sakazakii* 2051 belongs to CC64; the source analysis of the CC64 isolates in *Cronobacter* PubMLST indicated large proportion of the isolates from food (23/28) including infant formula (Figure 4.7). Whether hypothetical protein CSSP291_18210 is also present dominantly in CC64 remains unclear since *C. sakazakii* 2051 represented the only sequenced isolate of CC64.

The BLASTx analysis of the CSSP291_18210 revealed homologies against a membrane protein in *Enterobacter aerogenes* (section 4.4.3.5). Furthermore an attempt was also made for *in silico* prediction of this hypothetical protein using an online protein prediction tool (section 4.4.3.5). However, no functional category was assigned to this hypothetical protein although its secondary structure indicated some protein binding sites which could be important for its interaction with other proteins (Figure 4.6). Nevertheless, characterisation of this hypothetical protein in future studies is warranted in order to illustrate its importance in *C. sakazakii* CC4.

9.2.3 GENE BY GENE SEARCH OF THE PAN *C. SAKAZAKII* GENOME FOR THE IDENTIFICATION OF CC4 SPECIFIC TRAITS

The core genome analysis, initially, could only identify hypothetical protein CSSP291_18210 specific to *C. sakazakii* CC4. The genomic analysis was taken further to screen the pan *C. sakazakii* genome (n=30) using a gene by gene search approach. The aim of the analysis was to compare to the earliest *C. sakazakii* CC4 isolates 377 against a non-CC4 isolate ES15 using a gene by search in ACT. All the genes unique in CC4 isolate 377 were listed and then checked for their presence in the remaining 28 *C. sakazakii* isolates (17 CC4 and 11 non-CC4). At this stage hypothetical proteins and phage associated traits were omitted from the genomic comparison in order to make the comparison simpler.

The unique genes in CC4 isolate 377 were divided into 7 regions based on their function or location. Wide variety of traits were found in these 7 regions such as copper and silver homeostasis (*cus/sil* and *pco*) (which have already been studied in detail section 3.4.2.9), arsenic resistance genes, potassium uptake, pilus assembly protein and starvation-sensing protein, iron acquisition genes *iutA*, *iucABCD* (already investigated in detail in section 3.4.2.10), multidrug efflux and antimicrobial proteins, fimbrial associated genes and housekeeping genes such as relaxes, endonuclease, transcription, translation, toxin/anti-toxin and conjugation associated traits. Although some strains specific variations were observed, however no CC4 specific genes were noticed in these regions.

Of considerable interest were 7 genes located between Csak377_03841 and Csak377_03850 associated with O-antigen synthesis which were found uniformly in CC4 isolates but missing in all non-CC4 isolates except isolates 1249 and 978. These genes represented only part of the O-antigen and not the entire O:2 locus. The O-antigen region have been investigated in section 3.4.4 and the findings are largely in agreement as O:2 was predominant serotype in CC4, however not exclusive as two of the non-CC4 isolates 1249 and 978 were also O:2 (Table 5.7).

The most interesting observation in the present analysis was the giant adhesion associated gene CSSP291_03155 which was noted predominantly in CC4 isolates. The full length genes was detected in all of the CC4 isolates (18/18), however only half of the non-CC4 isolates (6/12) possessed the full length gene. For the ease of BLAST searching, the CSSP291_03155 gene was divided into smaller open reading frames (*orfs*) and searched for homology using

BLASTx. Interestingly majority of the hits indicated adhesion associated traits or type 1 secretion system associated traits. However, due to its presence in 50% of the non-CC4 isolates, the trait cannot be regarded as a CC4 specific trait. Yet, its significance cannot be neglected as it might play an important role towards virulence of *C. sakazakii* in general. Future studies aimed at characterising this gene may reveal important insights into its significance in *C. sakazakii*.

Another significant observation in this study was the presence of 2 hypothetical proteins CSAK377_02657 (homologue of CSSP291_18210 in *C. sakazakii* CC4 isolate SP291) and Csak377_03526 (homologue of CSSP291_10445 in *C. sakazakii* CC4 isolate SP291) which were specific to 18 CC4 isolates against 12 non-CC4 isolates (Table S4). However, since this analysis omitted the hypothetical protein and phage associate traits, therefore these hypothetical proteins were not discussed in detail here. Both of these hypothetical proteins were highlighted in the other parts of the study.

The search for the CC4 specific traits was taken one step ahead by fragmented genomic analysis using Gegenees.

9.2.4 FRAGMENTED GENOMIC ANALYSIS FOR THE IDENTFICATION OF CC4 SPECIFIC TRAITS

With the lowering cost of genome sequencing, the need for the platforms for genome analysis is becoming more popular. The advanced genomic analysis tools in this area are continuously being developed. One such tool which has recently been developed is called Gegenees (Agren et al. 2012). The principle of the program relies on the genomic fragmentation into sizeable fragments and then BLAST searching. The program allows to choose a target group within a set of genomes in order to find biomarker traits (unique genes) within the target group. Gegenees has been used for the identification of the biomarker traits within the *Bacillus cereus* group and in the foot and mouth disease viruses (Agren et al. 2012). Hence, it was another opportunity for the identification of the CC4 specific traits.

The fragmented genomic analysis was performed on 30 *C. sakazakii* isolates (Table 3.1) by setting the fragment size to 500 bp. The heat map for the similarity indices generated by Gegenees further reinforced the clonality of the 18 *C. sakazakii* CC4 isolates (Figure 6.1).

The biomarker score settings were set to the max of 1.0 which indicates only those genes which are specific to the target group. This way two of the genomic fragments were identified as specific to 18 CC4 isolates against 12 non-CC4 isolates, in the initial fragmented analysis (Figure 6.3). The interrogation of the genome annotation for the reference *C. sakazakii* CC4 isolate indicated these genomic fragments correspond to two hypothetical proteins. These hypothetical proteins were CSSP291_18210 and CSSP291_10445. Interestingly, the CSSP291_18210 was the homologue of CSAK557_03870 in CC4 isolate 557 which was identified as CC4 enriched in chapter 4 and a homologue of CSAK377_02657 in CC4 isolate 377 which was identified as CC4 enriched in chapter 5 (Table S4). Hence, this hypothetical protein was indicated as CC4 enriched by three different types of genomic analysis presented in this thesis. To avoid the locus_tag confusion, this hypothetical protein is denoted only by CSSP291_18210.

As discussed previously, the hypothetical protein CSSP291_18210 is the homologue of the inner membrane protein with some predicted protein binding sites and was also noted in only 2 of the non-CC4 isolates *C. sakazakii* 2051 and *C. malonaticus* 685 of which *C. sakazakii* 2051 belongs to CC64; an important clonal complex related to food especially infant formula (section 4.4.3.5).

An additional finding by Gegenees fragmented analysis was the hypothetical protein CSSP291_10445 which was found to be CC4 enriched. The analysis of the 107 *Cronobacter* spp. revealed 100% conservation of this hypothetical protein in CC4 genomes while it was noted only in 5 of the *C. sakazakii* non-CC4 isolates all of which belong to CC13 including 693, 700, 713, 714 and 715. The CC13 is also an important clonal complex as all of the isolates in this clonal complex are *C. sakazakii*. Moreover, a large proportion of the CC13 in the *Cronobacter* PubMLST were isolated from the infant formula. Hence, the CC13 also warrants attention in the future studies (section 6.4.1).

The BLASTx as well as *in silico* protein prediction could not reveal any functional relevance for hypothetical protein CSSP291_10445. However, it was predicted to be located in the inner membrane of the bacterial cell and also contain some protein binding sites on its

surface. These characteristics are important which may mediate some important virulence or environmental fitness associated mechanism in *C. sakazakii* CC4. Future studies are warranted to characterise this hypothetical protein to describe its significance in CC4. It is also suggested that future studies may also include some CC13 isolates as it seems to be an important clonal complex due to its association with food especially infant formula (section 6.4.1).

The gene loss can also enhance the virulence potential of bacteria (Nakata et al. 1993; Maurelli et al. 1998). An additional finding through fragmented genomic analysis was an hypothetical protein ESA_02200 missing from the CC4 strains. No functional relevance could be identified for this hypothetical protein. The analysis of the 107 *Cronobacter* genomes indicated that this gene was also missing from the 18 *Cronobacter* spp. non-CC4 isolates. Future studies are warranted to clarify the absence of ESA_02200 predominantly in CC4 (section 6.4.2).

Until this stage, as a result of the comparative genomics (chapter 3-6), two hypothetical proteins were detected as predominant in the *C. sakazakii* CC4 isolates. From this point onwards, the focus of the analysis was moved to study the diversity within CC4 and French outbreak (1994) investigation using the genome sequence data.

9.3 DIVERSITY WITHIN *C. SAKAZAKII* CC4

The strong clustering of the CC4 isolates on whole genome phylogeny (Figure 3.3) and large core genome size of the CC4 isolates (Figure 4.1) suggested very low sequence diversity within this clonal complex. However laboratory experiments in previous and current parallel PhD studies suggested differences in the CC4 isolates regarding their clinical presentations such adhesion and invasion assays (Nasreddin Rhouma 2012; Faisal Almajed 2014). Hence, it was of significance to study the diversity within CC4. For this purpose two different strategies were used, SNP analysis and COG-cgMLST.

The SNP analysis was undertaken using SMALT and SAMtools and filtered to retain high quality SNPs (section 2.2.9). The finished *C. sakazakii* CC4 isolate SP291 was used as the reference for SNP calling in remaining 17 CC4 isolates (Table 7.1). With exceptions, the

average number of SNPs between most CC4 isolates and the index strain SP291 was ~300-400 SNPs (Table 7.1). This is a relatively low number of SNPs considering the geographical and temporal spread of the CC4 isolates (Table 3.1). The SNP phylogeny showed some interesting variations as 2 of the isolates 1240 and 1587 showed longer branch; the longer branch length of the CC4 isolate 1240 was not a surprise as the *in silico* serotyping indicated that this isolate, unlike other 17 CC4 isolates, belong to *C. sakazakii* O:4 serotype (section 3.4.4). The lowest number of SNP differences was between the reference isolate SP291 and the earliest CC4 isolate 377 which differ from each other only by 16 SNPs. The high genetic similarity as well as relatedness of the source of isolation between these 2 isolates suggested that both of these isolates may share a common origin.

The isolate 1240 had the most number of SNPs differing from the index isolate SP291 by 1245 SNPs (Table 7.1). The analysis of these SNPs indicated most of the SNPs in a region containing genetic traits associated with O-antigen, hypothetical proteins and some phage associated traits (Figure 7.3).

The CC4 isolate 1587 with longer branch length had a concentration of SNPs between region 320012bp to 351260bp, the analysis of this genetic region indicated only phage associated traits (Figure 7.4).

An interesting observation on the SNP phylogeny was the clustering together of the 2 CC4 isolates 6 and 558, both of which have been categorised as low invasive on HBMEC, *in vitro* in a parallel PhD study (Faisal Almajed 2014, unpublished data). The analysis of the SNPs indicated a high concentration of SNPs in both of these isolates in the phage associated genomic regions (Figure 7.5). It was important to know whether there are genomic differences between the high and low invasive CC4 isolates in terms of gene contents. This observation formed the basis of comparative genomics of high and low invasive isolates.

Comparative genomics of high invasive CC4 isolate 767 was undertaken against the low invasive CC4 isolate 6. None of the unique traits in the high invasive CC4 isolate 767 could directly be related to its invasiveness. On the other hand, a number of metal resistance associated genes including arsenic, copper, manganese, chromate, nickel and copper were noted in low invasive isolate 6 (Table S7). Although, in trace amounts the metals are required for the metabolic functions, yet there increased concentrations are harmful for the bacterial cells. Hence, the acquisition of metal resistance genes in bacteria is important as

it helps them survive in a wide variety of environments containing higher concentrations of these toxic metals .

The relationship between the bacterial invasion and metal resistance is not clear. In order to verify if there is any subdivision between the *C. sakazakii* isolates based on metal resistance, a total of 10 *C. sakazakii* isolates were chosen to test for 4 different concentrations (1M, 0.1M, 0. 01M and 0.001M) of 8 different toxic metals which included copper (II) sulphate, sodium arsenate, nickel chloride, cobalt (II) nitrate, silver nitrate, zinc sulphate, cadmium carbonate and sodium tellurite. The choice of the isolates was made in such a way that these isolates not only represent high invasive (767, 20) and low invasive (6, 558) but also equal proportion of the CC4 and non-CC4 isolates (5 isolate from each group). The results of the metals resistance assays indicated that most of the isolates were able to resist toxic metals at concentrations of up to 0.01M irrespective of their invasion phenotype or clonal complex (section 7.3.3). Further investigation may clarify the association of metal resistance with invasion in *C. sakazakii*.

The genome sequences of the isolates analysed in the present study are also available at the *Cronobacter* PubMLST which host a genome comparator facility. The opportunity was utilised to perform BIGsDB analysis for 18 of the CC4 isolates keeping the CC4 isolate SP291 as the reference isolate. The analysis was based on the 3917 genes of the reference isolate SP291. Of total 3917 loci, at least 2682 loci showed allelic variations amongst 18 CC4 isolates. The allelic variation ranged from as low as 219 (1587 against 377) and as high as 2230 (701 vs 557). The results, in part verified the findings of the SNP analysis as three of the CC4 isolates SP291, 377 and 1587 clustered together on the neighbour net joining tree (Figure 7.6). These isolates also clustered closely on the SNP phylogeny, although with a longer branch length of isolate 1587 (Figure 7.1). The low number of SNP difference, low frequency of the loci variations as well as the relatedness in the source of isolation all support the proposal that isolates 377 and SP291 share a common origin. The fact that that SNP analysis and BIGsDB analysis may not be in complete agreement can be explained as the SNP calls in the CC4 isolates were subjected to rigorous filtering parameters (section 2.2.9).

The SNP analysis, overall, revealed a low sequence diversity within CC4. The analysis of the sequenced CC4 isolates by SNP analysis revealed interesting pattern as clustering together of the low invasive isolates. The genomic comparison between high and low invasive strains

revealed a number of metal resistance associated genes in the low invasive isolate 6. However, the metal resistance assays did not correlate with invasion phenotype and metal resistance ability of *C. sakazakii*. The BIGsDB comparative analysis of 18 CC4 isolates in part verified the findings of the SNP analysis and supported the proposal that CC4 isolates SP291 and 377 are likely to have a common source of origin.

9.4 GENOMIC ANALYSIS OF THE *C. SAKAZAKII* STRAINS ISOLATED DURING OUTBREAKS IN A NEONATAL INTENSIVE CARE UNIT IN FRANCE, 1994

As discussed earlier, there have been several outbreaks associated with *C. sakazakii*, however the largest outbreak hit a French NICU, 1994 lasting for approximately 3 months. A total of 18 neonates were infected or asymptomatically colonised during the length of the outbreak. The most significant observation in this outbreak was that all 3 babies which died during this outbreak were infected by the *C. sakazakii* isolates belonging to clonal complex 4. Hence, it was of significance to analyse this outbreak in greater detail using the genome sequence data. The previous study by Caubilla-Barron et al. (2007) focused only on the genotypic and phenotypic characteristics of the *C. sakazakii* isolates obtained from this outbreak and divided the isolates into 4 clusters using PFGE (Figure 8.1). In the present study, 26 of the *C. sakazakii* isolates from this outbreak were sequenced and their genomic data used for the detailed investigation. The aim was to construct the whole genome phylogeny to overview the larger subdivisions and then trace the outbreak source for these isolates using single nucleotide polymorphism.

The whole genome phylogeny (Figure 8.2) was largely in agreement with the PFGE clustering of Caubilla-Barron et al. (2007). The whole genome phylogeny indicated 4 divisions which were all in agreement with Caubilla-Barron et al. (2007). The largest group was cluster 2 containing 15 isolates all of which were CC4. The cluster 1, 3 and 4 contained 5, 5 and 1 isolates respectively belonging to CC12, CC13 and CC1 respectively. The whole genome phylogeny (Figure 8.2) in the present study favours the PFGE for observing the larger subdivision.

The SNP analysis was used to analyse the strain relatedness and to predict the potential source of outbreak. The SNP level analysis is advantageous over the PFGE analysis as it provides more resolution in an outbreak investigation. Although, PFGE is still widely used for the outbreak investigations, however it has limitations such as co-migration of the similar sized bands which may hinder the precise outbreak source (Turabelidze et al. 2013). With the low sequencing costs, the SNP analysis is becoming more affordable by the smaller laboratories where it can be used for the bacterial outbreak investigation. The SNP analysis is particularly significant in the cases where the bacterial isolates show very low sequence diversity. The SNP analysis has been used for the outbreak investigation of the *E. coli* O157:H7 associated with salad bar and romaine lettuce (Turabelidze et al. 2013; Leekitcharoenphon et al. 2014).

In the present study, the strategy was to use the earliest isolate in each cluster as index strain to call SNPs in the remaining isolates of that cluster. The number of SNPs were used as a measure of the strain relatedness i.e. low the number of SNPs between two isolates, the more related these isolates are and *vice versa*.

The SNP analysis of the cluster 1 showed a low number of SNPs (14 SNPs) between the index isolate 699 and the isolate 696 indicating a possible common source of origin. Both of these isolates were isolated approximately 2 months apart. The *C. sakazakii* are known for their ability survive the dry and desiccated environment of the PIF (Caubilla-Barron et al. 2007). In this environment there growth would be halted and the genome evolution would reduce to minimum. It is therefore likely that both of these isolates 699 and 696 were transmitted to babies through same batch of PIF where they survived and their growth was halted. The remaining 3 isolates could have been acquired from the independent sources such as water, carer's hands, utensils used to prepare PIF as these isolates showed relatively greater number of SNPs against the index isolate (section 8.4.2.1, Figure 8.3, Table 8.1).

The cluster 2 represented the most important cluster as all 3 deaths during this outbreak were attributed to this cluster, more importantly, all of these isolates were CC4 which was the rationale for initiating this outbreak investigation. From the total of 15 CC4 isolates, 11 were isolated from NEC cases, 1 from septicaemia and 1 from a meningitis case, the remaining 2 isolates were obtained from the asymptomatic babies. The earliest isolate of this cluster, *C. sakazakii* 701 was used as the index strain. The SNP typing and the tree

topology for the cluster 2 or CC4 isolates indicated further subdivisions within this cluster dividing it into the groups of 11, 1 and 3 isolates. The index isolate 701 clustered with the largest group and differ by only 6 SNPs from the prepared formula isolate 712. From this observation, it was suggested that the possible source of origin for 11 of these isolates was PIF. The isolate 767 showed relatively more SNPs against the index isolate 701, most of which were located in the phage area making it difficult to predict its source of origin. The remaining 3 isolates (705, 706, 707) which were obtained from the same baby showed more than 300 SNPs against the index isolate 701 indicating an independent acquisition for these 3 isolates. Moreover, these 3 isolates were highly identical to each other.

The cluster 3 contained 5 isolates obtained within 10 days from each other. The earliest isolate 693 was used as the index isolate to call SNPs. The SNP calls and tree topology indicated a maximum of 4 SNP differences between the index and the remaining isolates. The close resemblance of the index isolate with the prepared formula isolate 715, suggested that all of these 5 isolates were potentially originated from PIF.

Hence, SNP level investigation suggested PIF to be the potential source of outbreak for at least 40% of the cluster 1, 73% of the cluster 2 and 100% of the cluster 3 *C. sakazakii* isolates. There was likelihood of a 4th outbreak as the cluster 4 isolate 716 was obtained for the unopened can of the PIF. It is important to note that isolate 716 belongs to CC1 and the *Cronobacter* PubMLST database showed that CC1 has also been isolated frequently from clinical sources as well as from the infant formula.

Given the clinical significance, the ESβL and AmpC profiling was also undertaken only for the cluster 2 or CC4 isolates of the French outbreak, however none of these isolates was found to be ESβL or AmpC resistant (section 8.4.3).

Additionally, all of the sequenced isolates were also aligned against some known virulence associated plasmids which included pAPEC-O2-R and PECOS88 (Peigne et al. 2009). A homologue of the pAPEC-O2-R plasmid called PECOS88 has been found to be encoded on the genome of the neonatal meningitic *E. coli* strain S88 (O45: K1: H7) where it has been linked with an increased level of bacteraemia in a neonatal rat model of meningitis (Peigne et al. 2009). Both of these plasmid were aligned against the 26 sequenced *C. sakazakii* isolates (Table 8.1). The pAPEC-O2-R contain genes such as *folA* encoding resistance to trimethoprim *aadA5* which encodes resistance to aminoglycoside, *catB3* which contributes

to resistance against chloramphenicol and a beta-lactamase gene (*bla*) contributing to resistance against beta-lactam drugs; in addition *silCERS*, *silAB* and *silP* which provide resistance against silver and copper (Johnson et al. 2005). The complete homologue of the pAPEC-O2-R was not detected in any of the sequenced isolate, the only homology was against the copper and silver resistance region which has already been described in previous studies (Kucerova et al. 2010; Joseph et al. 2012b) and has been investigated previously in this study (section 3.4.2.9).

Similarly the PECOS88 plasmid of the *E. coli* S88 (O45: K1: H7) which has been associated with high level bacteraemia in neonatal meningitic rat model was also aligned against the 26 sequenced isolates using BRIG. Again, the complete homologue was not detected in any of the isolate, the only homology was against the iron acquisition region *iutA, iucABCD* which has been has been discussed previously in this thesis (section 3.4.2.10).

9.5 CONCLUSION

The association of *C. sakazakii* CC4 with neonatal infections especially neonatal meningitis, its persistence in the environment and its frequent isolation from the infant formula has been well established by the previous studies. However, the genetic factors behind the virulence potential as well as environmental fitness of this important neonatal health associated lineage of *Cronobacter* are not known. The present study was the first of its kind aimed primarily for the identification of the traits unique in the *C. sakazakii* CC4. Whole genome phylogeny revealed close clustering of the CC4 isolates despite their geographic as well as temporal diversity. Comparative genomics could only reveal two hypothetical proteins CSSP291_18210 and CSSP291_10445 predominant in *C. sakazakii* CC4 genomes. In addition, an hypothetical protein ESA_02200 was found to be dominantly absent from the CC4 isolates. The screening of the virulence and environmental fitness associated genes indicated a nearly equal spread for most of the traits investigated. The O:2 serotype was found to be the dominant serotype for CC4, however not exclusive to CC4. A giant adhesion associated gene CSSP291_03155 was noted in all *C. sakazakii* CC4 isolates yet variably in non-CC4 isolates. These results of the comparative genomics suggested that *C. sakazakii* CC4 might not be more virulent than non-CC4. The dominance of *C. sakazakii* CC4 in neonatal infections could be due to its predominance in the environment and PIF, ultimately increasing its exposure to neonates. Whether the hypothetical proteins identified as CC4 enriched in the present study have a role towards the virulence or fitness of *C. sakazakii* CC4 remains unclear at this stage. These hypothetical proteins warrant further investigation in future studies in order to elucidate their significance in CC4.

The low sequence diversity of the CC4 was further strengthened by the SNP analysis as average of ~300-400 SNP differences were noted between the CC4 isolates and the index strain SP291. Two of the low invasive CC4 isolates clustered together. Comparative genomics revealed metal resistance associated genes in the low invasive isolate. The laboratory studies could not reveal any difference in the metal resistance between high and low invasive isolates.

The investigation of the 1994 French outbreak isolates using whole genome sequencing data revealed 4 subdivisions within these isolates indicating three parallel outbreaks with the possibility of a 4th outbreak. It further revealed PIF to be the potential outbreak source

not only for the cluster 2 or CC4 isolates but for majority of the isolates in three clusters. The analysis also revealed the ability of *C. sakazakii* to co-infect its host. The SNP analysis revealed that multiple isolates from the same baby belonging to different clonal complexes or sequence types were completely unrelated to each other, indicating the ability of *C. sakazakii* isolates from different clonal complexes to co-exist in its human host which might ultimately increase the risk of infection.

The present PhD study has revealed important insights into the genomics of the neonatal meningitic *C. sakazakii* CC4 and has paved the way for further analysis to help better understand the pathogenicity of this important neonatal health associated bacterial lineage.

9.6 KEY NOVEL FINDINGS OF THE CURRENT PhD STUDY

- Whole genome phylogeny indicated close clustering of the 18 CC4 isolates.
- The *C. sakazakii* CC4 core genome was larger in size indicating low sequence diversity within this clonal complex.
- Comparative genomics revealed two hypothetical proteins; CSSP291_18210 and CSSP291_10445 predominant in CC4.
- Comparative genomics revealed hypothetical protein ESA_02200 predominantly absent from the CC4 genomes.
- The O:2 was the dominant serotype for *C. sakazakii* CC4 isolates.
- The thermotolerant region *orfHIJK* was dominantly detected in CC4 isolates.
- *C. sakazakii* isolates 6, ES15, 520 completely lacked the plasmid PESA3 while *C. sakazakii* isolates 1, 680 and 557 had partial length plasmid PESA3.
- Comparative genomics revealed a large adhesion associated gene CSSP291_03155 predominantly in *C. sakazakii* CC4 isolates.
- The SNP phylogeny revealed subdivisions within CC4 isolates as low invasive CC4 isolates 6 and 558 clustered together.
- Comparative genomics of high and low invasive *C. sakazakii* isolates revealed metal resistance associated genes in low invasive isolate.
- Metal resistance assays revealed no significant difference in metal sensitivity between high and low invasive isolates.
- The SNP analysis indicated that the earliest CC4 isolate 377 and CC4 isolate SP291 might have originated from a common source.
- The outbreak investigation using genome sequence data for the 1994 French outbreak revealed 3 parallel outbreaks with the potential of a 4^{th} outbreak and PIF to be the major potential outbreak source.
- The outbreak investigation revealed that *C. sakazakii* isolates belonging to different sequence types and clonal complexes have the ability to co-infect and hence co-exist in a human host.

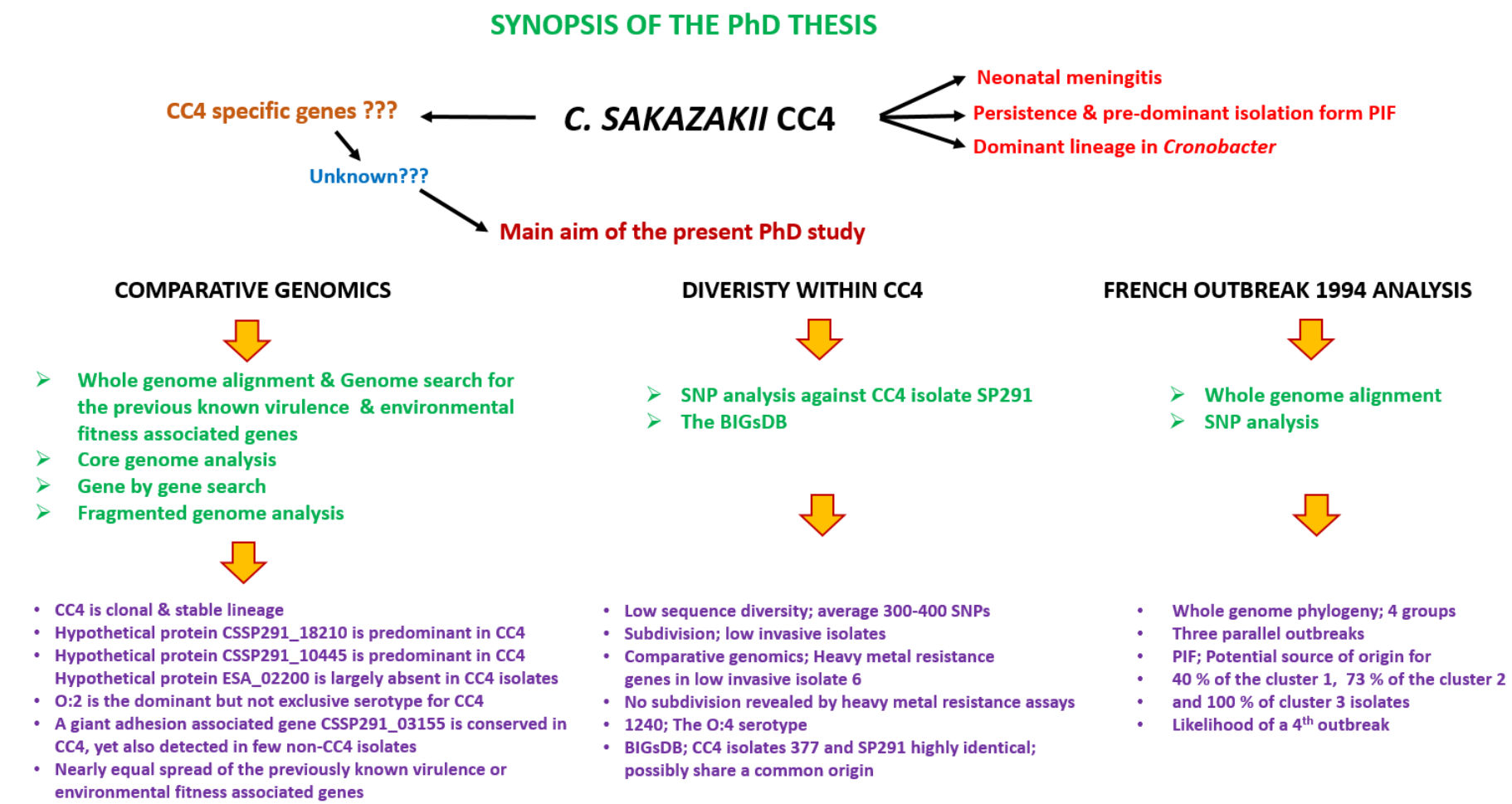


Figure 9. 1 The schematic representation of the work flow undertaken in the current PhD study.

The work was divided into three main sections (Block capitals). The green font colour shows the methodology while purple colour fonts highlight the key findings of the each section.

9.7 FUTURE DIRECTIONS

9.7.1 CHARACTERISATION OF THE CC4 ENRICHED HYPOTHETICAL PROTEINS

Future studies are warranted to characterise the two CC4 enriched hypothetical proteins CSSP291_18210 and CSSP291_10445. Initially, the studies may involve site directed mutagenesis and tissue culture studies by creating knock down or knock out gene mutants for these hypothetical proteins. The mutants and the wild type strains can then be compared for their ability to invade and attach the human intestinal and brain cell lines and can also be tested for any variation in the physiological traits such as desiccation resistance, heat tolerance and osmotolerance. It is also suggested that the *C. sakazakii* isolates from CC13 may also be included as this lineage has also been isolated frequently from PIF and the hypothetical protein CSSP291_10445 was also detected in isolates belonging to this lineage.

Furthermore, the difference in the expression of these hypothetical proteins encoding genes can also be studied after subjecting the selected isolates to hostile conditions such as desiccation, heat and osmotic shock.

Once the significance of these hypothetical proteins in *C. sakazakii* CC4 is determined additional proteomics based approaches such as mass spectrometry, MALDI-TOF, Western blot etc. may be used further characterisation of these hypothetical proteins.

Although not completely specific to CC4, due to the predominant presence in the CC4 isolates, the characterisation of the giant adhesion associated gene CSSP291_03155 using above strategies may also clarify its significance towards virulence of *C. sakazakii*.

9.7.2 CHARACTERISATION OF THE CC4 MISSING HYPOTHETICAL PROTEIN ESA_02200

Future studies are also warranted to characterise the hypothetical protein ESA_02200 which was found to be missing predominantly in the CC4 genomes. The characterisation of this hypothetical protein using same strategy as above (section 9.10.1), may provide important insights towards its absence in nearly all CC4 isolates analysed in the present study.

9.7.3 TRANSCRIPTOMICS FOR THE DIFFERENTIAL GENE EXPRESSION ANALYSIS

In the present PhD study, an exhaustive genome search was undertaken to find out the genes which are specific to *C. sakazakii* CC4. Apart from two hypothetical proteins enriched in CC4, no other genes could be found specific to CC4 genomes which can clearly explain its predominance in neonatal infections and its prevalence in the environment. In future studies, it is suggested that the gene expression studies are undertaken to investigate the gene contents which are differentially expressed in *C. sakazakii* CC4 after exposing them to stress conditions such as desiccation, heat and osmotic shock. The gene expression studies may result in the identification of important genes expressed exclusively in CC4 or overexpressed in the CC4 genomes under different hostile environmental conditions.

9.7.4 RELATIONSHIP BETWEEN SEROTYPE AND VIRULENCE POTNETIAL OF *C. SAKAZAKII*

Since the present study indicated that O:2 antigen is the dominant serotype in *C. sakazakii* CC4, it is recommended that future studies should also be undertaken to clarify if any relationship exists between *C. sakazakii* serotypes and their virulence potential.

9.7.5 RECOMBINATION DETECTION IN *C. SAKAZAKII* CC4

Recombination is a key driver and an important evolutionary force through which DNA segments are exchanged between closely related organisms. It is an important evolutionary factor which enables the acquisition of resistance elements and virulence traits. It is therefore recommended to detect the recombination events in *C. sakazakii* CC4 using Bayesian approaches. The recombination detection may provide important insights in to the genome evolution of *C. sakazakii* CC4.

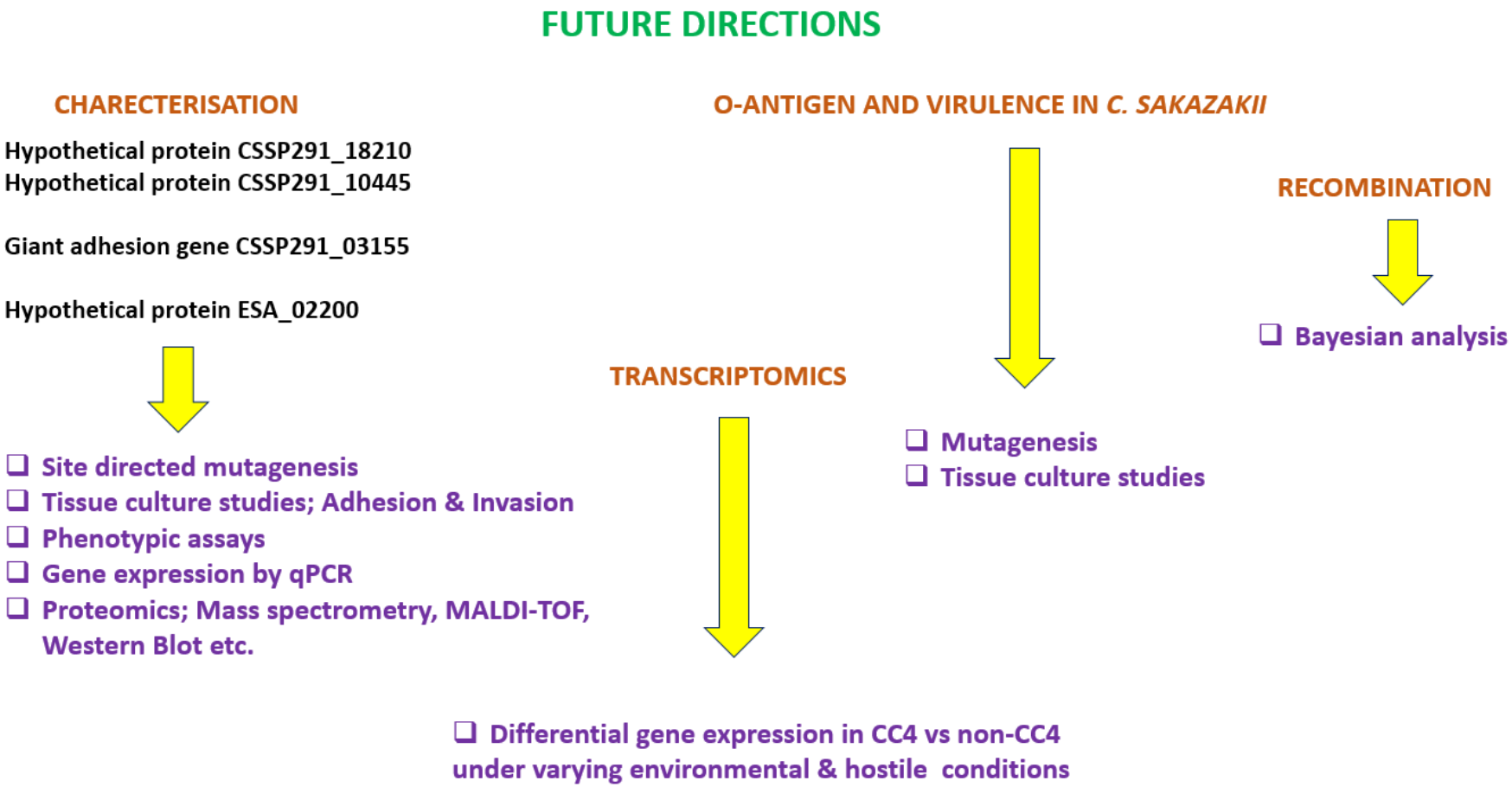


Figure 9. 2 Schematic representation of the future work emerging from the present PhD study.

Three different types of studies are suggested to better understand the virulence potential of *C. sakazakii* CC4. The text in the block capitals (dark orange) indicates the objectives for the future work while the text in the purple colour suggest the strategies which could be used.

REFERENCES

Agogue, H., Casamayor, E., Bourrain, M., Obernosterer, I., Joux, F., Herndl, G. & Lebaron, P. (2005). A survey on bacteria inhabiting the sea surface microlayer of coastal ecosystems. *FEMS Microbiology Ecology* **54**, 269-280.

Agostoni, C., Axelsson, I., Goulet, O., Koletzko, B., Michaelsen, K.F., Puntis, J.W.L., Rigo, J., Shamir, R., Szajewska, H., Turck, D. (2004). Preparation and handling of powdered infant formula: A commentary by the espghan committee on nutrition. *Journal of Pediatric Gastroenterology and Nutrition* **39**, 320.

Agren, J., Sundstrom, A., Hafstrom, T. & Segerman, B. (2012). Gegenees: Fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS ONE* **7**, e. 39107.

Alikhan, N., Petty, N., Ben Zakour, N. & Beatson, S. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402.

Al-Nabulsi, A., Osaili, T., Elabedeen, N., Jaradat, Z., Shaker, R., Kheirallah, K., Tarazi, Y. & Holley, R. (2011). Impact of environmental stress desiccation, acidity, alkalinity, heat or cold on antibiotic susceptibility of *Cronobacter sakazakii*. *International Journal of Food Microbiology* **146**, 137-143.

Alvarez-Ordonez, A., Begley, M. & Hill, C. (2012). Polymorphisms in *rpoS* and stress tolerance heterogeneity in natural isolates of *Cronobacter sakazakii*. *Applied and Environmental Microbiology* **78**, 3975-3984.

Alvarez-Ordonez, A., Begley, M., Clifford, T., Deasy, T., Collins, B. & Hill, C. (2014). Transposon mutagenesis reveals genes involved in osmotic stress and drying in *Cronobacter sakazakii*. *Food Research International* **55**, 45-54.

Angiuoli, S. & Salzberg, S. (2010). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334-342.

Arguello, J., Raimunda, D. & Padilla-Benavides, T. (2013). Mechanisms of copper homeostasis in bacteria. *Frontiers in Cellular and Infection Microbiology* **3**.

Arroyo, C., Condon, S. & Pagan, R. (2009). Thermobacteriological characterization of *Enterobacter sakazakii*. *International Journal of Food Microbiology* **136**, 110-118.

Babini, G. S. & Livermore, D. M. (2000). Antimicrobial resistance amongst *Klebsiella* spp. collected from intensive care units Southern and Western Europe in 1997–1998. *Journal of Antimicrobial Chemotherapy* **45**, 183–9.

Badger, J., Stins, M., & Kim, K. (1999). *Citrobacter freundii* invades and replicates in human brain microvascular endothelial cells. *Infection and immunity*, **67**, 4208-4215.

Badger, J., Wass, C. & Kim, K. (2000). Identification of *Escherichia coli* K1 genes contributing to human brain microvascular endothelial cell invasion by differential fluorescence induction. *Molecular Microbiology* **36**, 174-182.

Badger, T. M., Gilchrist, J. M., Pivik, R. T., Andres, A., Shankar, K., Chen, J. R., Ronis, M. J. (2009). The health implications of soy infant formula. *The American Journal of Clinical Nutrition* **89**, 1668S–1672S.

Baldwin, A., Loughlin, M., Caubilla-Barron, J., Kucerova, E., Manning, G., Dowson, C. & Forsythe, S. (2009). Multilocus sequence typing of *Cronobacter sakazakii* and *Cronobacter malonaticus* reveals stable clonal structures with clinical significance which do not correlate with biotypes. *BMC Microbiology* **9**, 223.

Bar-Oz, B., Preminger, A., Peleg, O., Block, C. & Arad, I. (2001). *Enterobacter sakazakii* infection in the newborn. *Acta Paediatrica* **90**, 356-358.

Barron, J. C., & Forsythe, S. J. (2007). Dry stress and survival time of *Enterobacter sakazakii* and other Enterobacteriaceae in dehydrated powdered infant formula. *Journal of food protection*, **70**, 2111-2117.

Baumgartner, A., Grand, M., Liniger, M. & Iversen, C. (2009). Detection and frequency of *Cronobacter* spp. (*Enterobacter sakazakii*) in different categories of ready-to-eat foods other than infant formula. *International Journal of Food Microbiology* **136**, 189-192.

Baumler, A. J., Tsolis, R. M. & Heffron, F. (1996). Contribution of fimbrial operons to attachment to and invasion of epithelial cell lines by *Salmonella typhimurium*. *Infection and Immunity* **64**, 1862-1865.

Beceiro, A., Bou, G. (2004). Class C b-lactamases: an increasing problem worldwide. *Reviews in Medical Microbiology* **15**, 141–52.

Bentley, D. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics & Development* **16**, 545-552.

Beuchat, L., Kim, H., Gurtler, J., Lin, L., Ryu, J. & Richards, G. (2009). *Cronobacter sakazakii* in foods and factors affecting its survival, growth, and inactivation. *International Journal of Food Microbiology* **136**, 204-213.

Bhagwat, A., Tan, J., Sharma, M., Kothary, M., Low, S., Tall, B. & Bhagwat, M. (2006). functional heterogeneity of RpoS in stress tolerance of enterohemorrhagic *Escherichia coli* strains. *Applied and Environmental Microbiology* **72**, 4978-4986.

Bhatia, J., Greer, F. (2008). Use of soy protein-based formulas in infant feeding. *Pediatrics* **121,** 1062–1068.

Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D. & Rost, B. (2004). Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Research* **32**, 2566-2577.

Block, C., Peleg, O., Minster, N., Bar-Oz, B., Simhon, A., Arad, I. & Shapiro, M. (2002). Cluster of neonatal infections in Jerusalem due to unusual biochemical variant of *Enterobacter sakazakii*. *European Journal of Clinical Microbiology and Infectious Diseases* **21**, 613-616.

Bonacorsi, S., Clermont, O., Houdouin, V., Cordevant, C., Brahimi, N., Marecat, A., Tinsley, C., Nassif, X., Lange, M. & Bingen, E. (2003). Molecular analysis and experimental virulence of French and north American *Escherichia coli* neonatal meningitis isolates: Identification of a new virulent clone. *The Journal of Infectious Diseases* **187**, 1895-1906.

Bowen, A. & Braden, C. (2006). Invasive *Enterobacter sakazakii* disease in infants. *Emerging Infectious Diseases* **12**, 1185-1189.

Brady, C., Cleenwerck, I., Venter, S., Coutinho, T. & De Vos P (2013). Taxonomic evaluation of the genus *Enterobacter* based on multilocus sequence analysis (MLSA): proposal to reclassify *E. nimipressuralis* and *E. amnigenus* into *Lelliottia* gen. nov. as *Lelliottia nimipressuralis* comb. nov. and *Lelliottia amnigena* comb. nov., respectively, *E. gergoviae* and *E. pyrinus* into *Pluralibacter* gen. nov. as *Pluralibacter gergoviae* comb. nov. and *Pluralibacter pyrinus* comb. nov., respectively, *E. cowanii*, *E. radicincitans, E. oryzae* and *E. arachidis* into *Kosakonia* gen. nov. as *Kosakonia cowanii* comb. nov., *Kosakonia radicincitans* comb. nov., Kosakonia oryzae comb. nov. and *Kosakonia arachidis* comb. nov., respectively, and *E. turicensis*, *E. helveticus* and *E. pulveris* into *Cronobacter* as *Cronobacter zurichensis* nom. nov., *Cronobacter helveticus* comb. nov. and *Cronobacter pulveris* comb. nov., respectively, and emended description of the genera *Enterobacter* and *Cronobacter*. *Systematic and Applied Microbiology* **36**, 309-19.

Branco, R., Chung, A. & Morais, P. (2008). Sequencing and expression of two arsenic resistance operons with different functions in the highly arsenic-resistant strain *Ochrobactrum tritici* SCII24T. *BMC Microbiology* **8**, 95.

Breeuwer, P., Lardeau, A., Peterz, M. & Joosten, H. (2003). Desiccation and heat tolerance of *Enterobacter sakazakii*. *Journal of Applied Microbiology* **95**, 967-973.

Brown, N. L., Barrett, S. R., Camakaris, J., Lee, B. T., Rouch, D. A. (1995). Molecular genetics and transport analysis of the copper-resistance determinant (pco) from *Escherichia coli* plasmid pRJ1004. *Molecular Microbiology* 1**7**, 1153–1166.

Bullen, J. J., Griffiths, E. (1999). Iron and infection: molecular, physiological and clinical aspects. New York: John Wiley and Sons.

Butler, J., Garcia-Maruniak, A., Meek, F. & Maruniak, J. (2010). Wild Florida house flies (*Musca domestica*) as carriers of pathogenic bacteria. *Florida Entomologist* **93**, 218-223.

CAC (2007). *Standards for Infant Formula and Formulas for Special Medical Purposes Intended for Infants*; CODEX STAN 72-108. Joint FAO/WHO Food Standards Programme, Food and Agriculture Organization of the United Nations: Rome, Italy.

Cai, J., Salmon, K. & DuBow, M. (1998). A chromosomal *ars* operon homologue of *Pseudomonas aeruginosa* confers increased resistance to arsenic and antimony in *Escherichia coli*. *Microbiology* **144**, 2705-2729.

Caldon, C. E., & March, P. E. (2003). Function of the universally conserved bacterial GTPases. *Current Opinion in Microbiology* 6, 135-139.

Carlin, A., Shi, W., Dey, S. & Rosen, B. P. (1995). The *ars* operon of *Escherichia coli* confers arsenical and antimonial resistance. *Journal of Bacteriology* **177,** 981-986.

Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B., Parkhill, J. & Rajandream, M. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672-2676.

Carver, T., Harris, S., Berriman, M., Parkhill, J. & McQuillan, J. (2011). Artemis: an integrated platform for visualisation and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464-469.

Carver, T., Rutherford, K., Berriman, M., Rajandream, M., Barrell, B. & Parkhill, J. (2005). ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422-3423.

Caubilla-Barron, J., Hurrell, E., Townsend, S., Cheetham, P., Loc-Carrillo, C., Fayet, O., Prere, M. & Forsythe, S. (2007). Genotypic and phenotypic analysis of *Enterobacter sakazakii* strains from an outbreak resulting in fatalities in a neonatal intensive care unit in France. *Journal of Clinical Microbiology* **45**, 3979-3985.

Centers for Disease Control and Prevention (2012). CDC Update: Investigation of *Cronobacter* infections among infants in the United States. January 13, 2012. http://www.cdc.gov/foodsafety/diseases/*cronobacter*/investigation.html.

Cetin, S., Ford, H., Sysko, L., Agarwal, C., Wang, J., Neal, M., Baty, C., Apodaca, G. & Hackam, D. (2004). Endotoxin Inhibits intestinal epithelial restitution through activation of Rho-GTPase and increased focal adhesions. *Journal of Biological Chemistry* **279**, 24592-24600.

Cetinkaya, E., Joseph, S., Ayhan, K. & Forsythe, S. (2013). Comparison of methods for the microbiological identification and profiling of *Cronobacter* species from ingredients used in the preparation of infant formula. *Molecular and Cellular Probes* **27**, 60-64.

Chandrapala, D., Kim, K., Choi, Y., Senevirathne, A., Kang, D., Ryu, S. & Kim, K. (2014). Putative Inv is essential for basolateral invasion of Caco-2 cells and acts synergistically with OmpA to affect *in vitro* and *in vivo* virulence of *Cronobacter sakazakii* ATCC 29544. *Infection and Immunity* **82**, 1755-1765.

Chenu, J. & Cox, J. (2009). *Cronobacter* (*Enterobacter sakazakii*): current status and future prospects. *Letters in Applied Microbiology* **49**, 153-159.

Chiang, S., Dong, T., Edge, T. & Schellhorn, H. (2011). Phenotypic diversity caused by differential RpoS activity among environmental *Escherichia coli* isolates. *Applied and Environmental Microbiology* **77**, 7915-7923.

Chuba, P., Leon, M., Banerjee, A. & Palchaudhuri, S. (1989). Cloning and DNA sequence of plasmid determinant iss, coding for increased serum survival and surface exclusion, which has homology with lambda DNA. *MGG Molecular & General Genetics* **216**, 287-292.

Clark, G., Paszkiewicz, K., Hale, J., Weston, V., Constantinidou, C., Penn, C., Achtman, M. & McNally, A. (2012). Genomic analysis uncovers a phenotypically diverse but genetically homogeneous *Escherichia coli* ST131 clone circulating in unrelated urinary tract infections. *Journal of Antimicrobial Chemotherapy* **67**, 868-877.

Connell, I., Agace, W., Klemm, P., Schembri, M., Marild, S. & Svanborg, C. (1996). Type 1 fimbrial expression enhances *Escherichia coli* virulence for the urinary tract. *Proceedings of the National Academy of Sciences* **93**, 9827-9832.

Coudron, P. E. & Moland, E. S. (2000). Thomson KS. Occurrence and detection of AmpC beta-lactamases among *Escherichia coli*, *Klebsiella pneumoniae* and *Proteus mirabilis* isolates at a veterans medical center. *Antimicrobial Agents and Chemotherapy* **38**, 1791–6.

Craven, H., McAuley, C., Duffy, L. & Fegan, N. (2010). Distribution, prevalence and persistence of *Cronobacter* (*Enterobacter sakazakii*) in the nonprocessing and processing environments of five milk powder factories. *Journal of Applied Microbiology* **109**, 1044-1052.

Cruz-Cordova, A., Rocha-Ramirez, L., Ochoa, S., Gonzalez-Pedrajo, B., Espinosa, N., Eslava, C., Hernandez-Chinas, U., Mendoza-Hernandez, G. & Rodriguez-Leviz, A. et al. (2012). Flagella from five *Cronobacter* species induce pro-inflammatory cytokines in macrophage derivatives from human monocytes. *PloS one* **7**, 52091.

Cullen, W.R. & Reimer, K. J. (1989). Arsenic speciation in the environment. *Chemical Reviews* **89**, 713–764.

Dancer, G., Mah, J., Rhee, M., Hwang, I. & Kang, D. (2009). Resistance of *Enterobacter sakazakii* (*Cronobacter* spp.) to environmental stresses. *Journal of Applied Microbiology* **107**, 1606-1614.

Delepelaire, P. (2004). Type I secretion in gram-negative bacteria. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1694**, 149-161.

den Bakker, H., Cummings, C., Ferreira, V., Vatta, P., Orsi, R., Degoricija, L., Barker, M., Petrauskene, O., Furtado, M. & Wiedmann, M. (2010). Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics* **11**, 688.

Dey, S. (1997). Alternate Energy Coupling of ArsB, the Membrane Subunit of the Ars Anion-translocating ATPase. *Journal of Biological Chemistry* **272**, 326-331.

Dietzman, D., Fischer, G. & Schoenknecht, F. (1974). Neonatal *Escherichia coli* septicemia bacterial counts in blood. *The Journal of Pediatrics* **85**, 128-130.

Dodd, C. & Aldsworth, T. (2002). The importance of RpoS in the survival of bacteria through food processing. *International Journal of Food Microbiology* **74**, 189-194.

Dong, T., Chiang, S., Joyce, C., Yu, R. & Schellhorn, H. (2009). Polymorphism and selection of rpoS in pathogenic *Escherichia coli*. *BMC Microbiology* **9**, 118.

Doran, J. L., Collinson, K., Burian, J., Sarlos, G., Todd, E. C. D., Munro, C. K., Kay, C. M., Banser, P. A., Peterkin, P. I. & Kay, W. W. (1993). DNA-based diagnostic tests for *Salmonella* species targeting *agfA*, the structural gene for thin, aggregative fimbriae. *Journal of Clinical Microbiology* **31**, 2263–2273.

Duffus, J. H. (2002). "Heavy metals": a meaningless term?. *Pure and Applied Chemistry* **74**, 793-807.

Edelson-Mammel, S., Porteous, M. & Buchanan, R. (2005). Survival of *Enterobacter sakazakii* in a dehydrated powdered infant formula. *Journal of Food Protection* **68**, 1900-1902.

Ernst, R., Guina, T. & Miller, S. (1999). How intracellular bacteria survive: surface modifications that promote resistance to host innate immune responses. *The Journal of Infectious Diseases* **179**, 326-330.

FAO/WHO (2004). *Enterobacter sakazakii* and other microorganisms in powdered infant formula: meeting report. *Microbiological Risk Assessment Series* **6***.* WHO-FAO of the United Nations, Geneva and Rome. WHO Press, Geneva, Switzerland.

FAO/WHO (2006). *Enterobacter sakazakii* and *Salmonella* in powdered infant formula (meeting report). *Microbiological Risk Assessment Series* **10**. Rome: Food and Agriculture Organization of the United Nations/World Health Organization.

Farmer, J., Asbury, M., Hickman, F. & Brenner, D. (1980). *Enterobacter sakazakii*: A new species of "Enterobacteriaceae" isolated from clinical specimens. *International Journal of Systematic Bacteriology* **30**, 569-584.

Feeney, A. & Sleator, R. (2011). An *in silico* analysis of osmotolerance in the emerging gastrointestinal pathogen *Cronobacter sakazakii*. *Bioengineered Bugs* **2**, 260-270.

Feeney, A., Johnston, C., Govender, R., O'Mahony, J., Coffey, A. & Sleator, R. (2014). Analysis of the role of the *Cronobacter sakazakii* ProP homologues in osmotolerance. *Gut Pathogens* **6**, 15.

Flores, J., Medrano, S., S'anchez, J. & Fernandez-Escartin, E. (2011). Two cases of hemorrhagic diarrhea caused by *Cronobacter sakazakii* in hospitalized nursing infants associated with the consumption of powdered infant formula. *Journal of Food Protection* **74**, 2177-2181.

Forsythe, S.J., Dickins, B. & Jolley, K. A. (2014). *Cronobacter*, the emergent bacterial pathogen *Enterobacter sakazakii* comes of age; MLST and whole genome sequence analysis. *BMC Genom*ics **15**, 1121.

Franco, A., Hu, L., Grim, C., Gopinath, G., Sathyamoorthy, V., Jarvis, K., Lee, C., Sadowski, J. & Kim, J. et al. (2011a). Characterization of putative virulence genes on the related RepFIB plasmids harbored by *Cronobacter* spp. *Applied and Environmental Microbiology* **77**, 3255-3267.

Franco, A., Kothary, M., Gopinath, G., Jarvis, K., Grim, C., Hu, L., Datta, A., McCardell, B. & Tall, B. (2011b). Cpa, the outer membrane protease of *Cronobacter sakazakii*, activates plasminogen and mediates resistance to serum bactericidal activity. *Infection and Immunity* **79**, 1578-1587.

Friedemann, M. (2007). *Enterobacter sakazakii* in food and beverages (other than infant formula and milk powder). *International Journal of Food Microbiology* **116**, 1-10.

Friedemann, M. (2009). Epidemiology of invasive neonatal *Cronobacter* (*Enterobacter sakazakii*) infections. *European Journal of Clinical Microbiology & Infectious Diseases* **28,** 1297-1304.

Fyre, C. & Donnelly, C. W. (2005). Comprehensive survey of pasteurised fluid milk produced in the United States reveals a low prevalence of *Listeria monocytogenes*. *Journal of Food Protection* **68**, 973- 979.

Gajdosova, J., Benedikovicova, K., Kamodyova, N., Tothova, L., Kaclikova, E., Stuchlik, S., Turna, J. & Drahovska, H. (2011). Analysis of the DNA region mediating increased thermotolerance at 58°C in *Cronobacter* sp. and other *Enterobacterial* strains. *Antonie van Leeuwenhoek* **100**, 279–289.

Gallagher, P. & Ball, W. (1991). Cerebral infarctions due to CNS infection with *Enterobacter sakazakii*. *Pediatric Radiology* **21**, 135-136.

Gardner, S. & Slezak, T. (2010). Scalable SNP analysis of 100+ bacterial or viral genomes. *Journal of Forensic Research* **01**. doi: 10.4172/2157-7145.1000107

Gerlach, R., Jackel, D., Stecher, B., Wagner, C., Lupas, A., Hardt, W. & Hensel, M. (2007). *Salmonella* pathogenicity island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. *Cellular Microbiology* **9**, 1834-1850.

Glode, M., Sutton, A., Robbins, J., McCracken, G., Gotschlich, E., Kaijser, B. & Hanson, L. (1977). Neonatal meningitis due to *Escherichia coli* K1. *Journal of Infectious Diseases* **136**, 93-97.

Goldberg, T., Hamp, T. & Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics* **28**, 458-465.

Gorden, J. and Small, P. (1993). Acid resistance in enteric bacteria. *Infection and Immunity* **61**, 364–367.

Gosney, M., Martin, M., Wright, A. & Gallagher, M. (2006). *Enterobacter sakazakii* in the mouths of stroke patients and its association with aspiration pneumonia. *European Journal of Internal Medicine* **17**, 185-188.

Grim, C., Kotewicz, M., Power, K., Gopinath, G., Franco, A., Jarvis, K., Yan, Q., Jackson, S. & Sathyamoorthy, V. et al. (2013). Pangenome analysis of the emerging foodborne pathogen *Cronobacter* spp. suggests a species-level bidirectional divergence driven by niche adaptation. *BMC Genomics* **14**, 366.

Grishin, A., Papillon, S., Bell, B., Wang, J. & Ford, H. (2013). The role of the intestinal microbiota in the pathogenesis of necrotising enterocolitis. *Seminars in Pediatric Surgery* **22,** 69-75.

Grothe, S., Krogsrud, R. L., McClellan, D. J., Milner, J. L., & Wood, J. M. (1986). Proline transport and osmotic stress response in *Escherichia coli* K-12. *Journal of Bacteriology* **166**, 253-259.

Hamby, S., Joseph, S., Forsythe, S. & Chuzhanova, N. (2011). *In silico* identification of pathogenic strains of *Cronobacter* from Biochemical data reveals association of inositol fermentation with pathogenicity. *BMC Microbiol*ogy **11**, 204.

Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A. & Braun, T. et al. (2013). Homology-based inference sets the bar high for protein function prediction. *BMC Bioinformatics* **14**, 7.

Hantke, K. (1981). Regulation of ferric iron transport in Escherichia coli K12: isolation of a constitutive mutant. *Molecular Genetics and Genomics* **182**, 288–292.

Hao, X., Lüthje, F. L., Qin, Y., McDevitt, S. F., Lutay, N., Hobman, J. L., Karishma Asiani, K., Fernando C. Soncini, F. C., German, N., Zhang, S., Zhu, Y. & Rensing, C. (2015). Survival in amoeba—a major selection pressure on the presence of bacterial copper and zinc resistance

determinants? Identification of a “copper pathogenicity island”. *Applied Microbiology and Biotechnology* **99**, 5817-5824.

Hariri, S., Joseph, S. & Forsythe, S. (2013). *Cronobacter sakazakii* ST4 strains and neonatal meningitis, United States. *Emerging Infectious Diseases* **19**, 175-177.

Harouna, S., Carraminana, J. J., Navarro, F., Perez, M. D., Calvo, M., & Sanchez, L. (2015). Antibacterial activity of bovine milk lactoferrin on the emerging foodborne pathogen *Cronobacter sakazakii*: Effect of media and heat treatment. *Food Control* **47**, 520-525.

Harris, S., Feil, E., Holden, M., Quail, M., Nickerson, E., Chantratita, N., Gardete, S., Tavares, A. & Day, N. et al. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-474.

Hartmann, I., Carranza, P., Lehner, A., Stephan, R., Eberl, L. & Riedel, K. (2010). Genes involved in *Cronobacter sakazakii* biofilm formation. *Applied and Environmental Microbiology* **76**, 2251-2261.

Healy, B., Cooney, S., O'Brien, S., Iversen, C., Whyte, P., Nally, J., Callanan, J. & Fanning, S. (2010). *Cronobacter* (*Enterobacter sakazakii*): An opportunistic foodborne pathogen. *Foodborne Pathogens and Disease* **7**, 339-350.

Heffernan, E. J., Harwood, J., Fierer, J. & Guiney, D. (1992). The *Salmonella typhimurium* virulence plasmid complement resistance gene *rck* is homologous to a family of virulence-related outer membrane protein genes, including *pagC* and ail. *Journal of Bacteriology*, **174**, 84-91.

Hein, I., Gadzov, B., Schoder, D., Foissy, H., Malorny, B. & Wagner, M. (2009). Temporal and spatial distribution of *Cronobacter* isolates in a milk powder processing plant determined by Pulsed-Field Gel electrophoresis. *Foodborne Pathogens and Disease* **6**, 225-233.

Hill, V., Townsend, S., Arias, R., Jenabi, J., Gomez-Gonzalez, I., Shimada, H. & Badger, J. (2003). TraJ-Dependent *Escherichia coli* K1 Interactions with professional phagocytes are important for early systemic dissemination of infection in the neonatal rat. *Infection and Immunity* **72**, 478-488.

Himelright, I., Harris, E., Lorch, V. & Anderson, M. (2002). *Enterobacter sakazakii* infections associated with the use of powdered infant formula -Tennessee, 2001. *Journal of the American Medical Association* **287**, 2204–2205.

Hobman, J. L., & Crossman, L. (2014). Bacterial antimicrobial metal ion resistance. *Journal of Medical Microbiology* **64**, 471-497.

Hoe, N. P., Nakashima, K., Lukomski, S., Grigsby, D., Liu, M. & Kordari, P. et al. (1999). Rapid selection of complement-inhibiting protein variants in group A *Streptococcus* epidemic waves. *Nature Medicine* **5***,* 924-929.

Hoffman, J., Badger, J., Zhang, Y., Huang, S. & Kim, K. (2000). *Escherichia coli* K1 aslA contributes to invasion of brain microvascular endothelial cells *in vitro* and *in vivo*. *Infection and Immunity* **68**, 5062-5067.

Hol, C., Verduin, C., Dijke, E., Verhoef, J., Fleer, A. & Dijk, H. (1995). Complement resistance is a virulence factor of *Branhamella (Moraxella) catarrhalis*. *FEMS Immunology & Medical Microbiology* **11,** 207-211.

Holy, O. & Forsythe, S. (2014). *Cronobacter* spp. as emerging causes of healthcare-associated infection. *Journal of Hospital Infection* **86**, 169-177.

Holy, O., Matouskova, I., Holy, V., Koukalova, D. & Chmelar, D. (2011). Isolation of *Cronobacter* spp.(formerly *Enterobacter sakazakii*) from nostrils of healthy stable horse-short communication. *Epidemiologie, mikrobiologie, imunologie: casopis Spolecnosti pro epidemiologii a mikrobiologii Ceske lekarske spolecnosti JE Purkyne* **60**, 167-169.

Horlacher, R. & Boos, W. 1997. Characterization of TreR, the major regulator of the *Escherichia coli* trehalose system. *Journal of Biological Chemistry* **272**, 13026–13032.

Hottes, A., Freddolino, P., Khare, A., Donnell, Z., Liu, J. & Tavazoie, S. (2013). Bacterial adaptation through loss of function. *PLoS Genetics* **9**, e1003617.

Hsieh, Y. H. P. & Ren, Z. (2001). Determination of microbial quality in pasteurized milk using a non-instrumental microrespirometer. *Food Microbiology* **88**, 24- 28.

Huang, S. H., Chen, Y.H., Fu, Q., Wang, Y., Stins, M., Wass, C. & Kim, K. S. (1999). Identification and characterization of an *E. coli* invasion gene locus *ibeB* required for penetration of brain microvascular endothelial cells. *Infection and Immun*ity **67**, 2103–2109.

Huang, S., Wan, Z., Chen, Y., Jong, A. & Kim, K. (2001). Further characterization of *Escherichia coli* brain microvascular endothelial cell invasion gene ibeA by deletion, complementation, and protein expression. *The Journal of Infectious Diseases* **183,** 1071-1078.

Hurrell, E., Kucerova, E., Loughlin, M., Caubilla-Barron, J. & Forsythe, S. (2009a). Biofilm formation on enteral feeding tubes by *Cronobacter sakazakii*, *Salmonella* serovars and other Enterobacteriaceae. *International Journal of Food Microbiology* **136**, 227-231.

Hurrell, E., Kucerova, E., Loughlin, M., Caubilla-Barron, J., Hilton, A., Armstrong, R., Smith, C., Grant, J., Shoo, S. & Forsythe, S. (2009b). Neonatal enteral feeding tubes as loci for colonisation by members of the Enterobacteriaceae*. BMC Infectious Diseases* **9**, 146.

Hyatt, D., Chen, G., LoCascio, P., Land, M., Larimer, F. & Hauser, L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Iversen, C. & Forsythe, S. (2003). Risk profile of *Enterobacter sakazakii*, an emergent pathogen associated with infant milk formula. *Trends in Food Science and Technology* **14**, 443-454.

Iversen, C. & Forsythe, S. (2004). Isolation of *Enterobacter sakazakii* and other Enterobacteriaceae from powdered infant formula milk and related products. *Food Microbiology* **21**, 771-777.

Iversen, C., Lane, M. & Forsythe, S. (2004). The growth profile, thermotolerance and biofilm formation of *Enterobacter sakazakii* grown in infant formula milk. *Letters in Applied Microbiology*, **38**, 378-382.

Iversen, C., Lehner, A., Mullane, N., Bidlas, E., Cleenwerck, I., Marugg, J., Fanning, S., Stephan, R. & Joosten, H. (2007). The taxonomy of *Enterobacter sakazakii*: proposal of a new genus *Cronobacter* gen. nov. and descriptions of *Cronobacter sakazakii* comb. nov. *Cronobacter sakazakii* subsp. *sakazakii*, comb. nov., *Cronobacter sakazakii* subsp. malonaticus subsp. nov., *Cronobacter turicensis* sp. nov., *Cronobacter* muytjensii sp. nov., *Cronobacter* dublinensis sp. nov. and *Cronobacter* genomospecies 1. *BMC Evolutionary Biol*ogy **7**, 64.

Iversen, C., Mullane, N., McCardell, B., Tall, B., Lehner, A., Fanning, S., Stephan, R. & Joosten, H. (2008). *Cronobacter* gen. nov., a new genus to accommodate the biogroups of *Enterobacter sakazakii*, and proposal of *Cronobacter sakazakii* gen. nov., comb. nov., *Cronobacter malonaticus* sp. nov., *Cronobacter turicensis* sp. nov., *Cronobacter muytjensii* sp. nov., *Cronobacter* dublinensis sp. nov., *Cronobacter* genomospecies 1, and of three subspecies, *Cronobacter dublinensis* subsp. dublinensis subsp. nov., *Cronobacter dublinensis* subsp. *lausannensis* subsp. nov. and *Cronobacter dublinensis* subsp. *lactaridi* subsp. nov. *International Journal of Systematic And Evolutionary Microbiology* **58**, 1442-1447.

Iversen, C., Waddington, M., Farmer, J. & Forsythe, S. (2006). The biochemical differentiation of *Enterobacter sakazakii* genotypes. *BMC Microbiology* **6**, 94.

Jacobs, C., Braun, P. & Hammer, P. (2011). Reservoir and routes of transmission of *Enterobacter sakazakii* (*Cronobacter* spp.) in a milk powder-producing plant. *Journal of Dairy Science* **94**, 3801-3810.

Jaradat, Z., Al Mousa, W., Elbetieha, A., Al Nabulsi, A. & Tall, B. (2014). *Cronobacter* spp. - opportunistic food-borne pathogens. A review of their virulence and environmental-adaptive traits. *Journal of Medical Microbiology* **63**, 1023-1037.

Jarvis, K., Grim, C., Franco, A., Gopinath, G., Sathyamoorthy, V., Hu, L., Sadowski, J., Lee, C. & Tall, B. (2011). Molecular characterization of *Cronobacter* lipopolysaccharide O-Antigen gene clusters and development of serotype-specific PCR assays. *Applied and Environmental Microbiology* **77**, 4017-4026.

Jarvis, K., Yan, Q., Grim, C., Power, K., Franco, A., Hu, L., Gopinath, G., Sathyamoorthy, V. & Kotewicz, M. et al. (2013). Identification and characterization of five new molecular serogroups of *Cronobacter* spp. *Foodborne Pathogens and Disease* **10**, 343-352.

Johler, S., Stephan, R., Hartmann, I., Kuehner, K. & Lehner, A. (2009). Genes involved in yellow pigmentation of *Cronobacter sakazakii* ES5 and influence of pigmentation on persistence and growth under environmental stress. *Applied and Environmental Microbiology* **76**, 1053-1061.

Johnson, T., Johnson, S. & Nolan, L. (2006). Complete DNA Sequence of a ColBM plasmid from avian pathogenic *Escherichia coli* suggests that it evolved from closely related ColV virulence plasmids. *Journal of Bacteriology* **188**, 5975-5983.

Johnson, T., Siek, K., Johnson, S. & Nolan, L. (2005). DNA Sequence and Comparative Genomics of pAPEC-O2-R, an Avian Pathogenic *Escherichia coli* Transmissible R Plasmid. *Antimicrobial Agents and Chemotherapy* **49**, 4681-4688.

Johnson, T., Siek, K., Johnson, S. & Nolan, L. (2005). DNA Sequence and Comparative Genomics of pAPEC-O2-R, an Avian Pathogenic *Escherichia coli* Transmissible R Plasmid. *Antimicrobial Agents and Chemotherapy* **49**, 4681-4688.

Jolley, K. & Maiden, M. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595.

Joseph, S. & Forsythe, S. (2011). *Cronobacter sakazakii* sequence type 4 in neonatal infections. *Emerging Infectious Diseases* **17**, 1713-1715.

Joseph, S. & Forsythe, S. (2012). Insights into the emergent bacterial pathogen *Cronobacter* spp., generated by multilocus sequence typing and analysis. *Frontiers in Microbiology* **3**.

Joseph, S., Desai, P., Ji, Y., Hamby, S. E., Masood, N., Hariri, S., Sonbol, H., Cumming, C. A., Rico, A., Shih, S. M., Degoricja, L., Brzoska, P., Chuzhanova, N., McClelland, M., Furtado, M. R., & Forsythe, S. (2012b). Comparative analysis of genome sequences covering the seven *Cronobacter* species. *PLOS ONE* **7**, e49455.

Joseph, S., Hariri, S., Masood, N. & Forsythe, S. (2013). Sialic acid utilisation by *Cronobacter sakazakii*. *Microbial Informatics and Experimentation* **3**, 3.

Joseph, S., Sonbol, H., Hariri, S., Desai, P., McClelland, M. & Forsythe, S. (2012a). Diversity of the *Cronobacter* genus as revealed by multilocus sequence typing. *Journal of Clinical Microbiology* **50**, 3031-3039.

Julio, S. & Cotter, P. (2005). Characterization of the Filamentous Hemagglutinin-like protein FhaS in *Bordetella bronchiseptica*. *Infection and Immunity* **73**, 4960-4971.

Kandhai, M., Reij, M., Gorris, L., Guillaume-Gentil, O. & van Schothorst, M. (2004). Occurrence of *Enterobacter sakazakii* in food production environments and households. *The Lancet* **363**, 39-40.

Kehl-Fie T. E., Skaar, E. P. (2009). Nutritional immunity beyond iron: a role for manganese and zinc. *Current Opinion in Chemical Biology* **14**, 218–224.

Kent, R. M., Fitzgerald, G. F., Hill, C., Stanton, C., & Ross, R. P. (2015). Novel Approaches to Improve the Intrinsic Microbiological Safety of Powdered Infant Milk Formula. *Nutrients* **7**, 1217-1244.

Khan, N. A., Wang, Y., Kim, K. J., Chung, J. W., Wass, C. A. & Kim, K. S. (2002). Cytotoxic necrotising factor 1 contributes to *Escherichia coli* K1 invasion of the central nervous system. *Journal of Biological Chemistry* **277**, 15607–15612.

Kilonzo-Nthenge, A., Rotich, E., Godwin, S., Nahashon, S. & Chen, F. (2012). Prevalence and antimicrobial resistance of *Cronobacter sakazakii* isolated from domestic kitchens in middle Tennessee, United States. *The Journal of Food Protection* **75,** 1512-1517.

Kim, H., Ryu, J. & Beuchat, L. (2007). Effectiveness of disinfectants in killing *Enterobacter sakazakii* in suspension, dried on the surface of stainless steel, and in a biofilm. *Applied and Environmental Microbiology* **73**, 1256-1265.

Kim, K. (2001). *Escherichia coli* translocation at the blood-brain barrier. *Infection and Immunity* **69**, 5217-5222.

Kim, K. (2002). Strategy of *Escherichia coli* for crossing the blood brain barrier. *The Journal of Infectious Diseases* **186**, 220-224.

Kim, K., Itabashi, H., Gemski, P., Sadoff, J., Warren, R. & Cross, A. (1992). The K1 capsule is the critical determinant in the development of *Escherichia coli* meningitis in the rat. *Journal of Clinical Investigation* **90**, 897-905.

Kim, K., Kim, K., Choi, J., Lim, J., Lee, J., Hwang, S. & Ryu, S. (2010). Outer Membrane Proteins A (OmpA) and X (OmpX) are essential for basolateral invasion of *Cronobacter sakazakii*. *Applied and Environmental Microbiology* **76**, 5188-5198.

Koebnik, R. (1999). Structural and functional roles of the surface-exposed loops of the barrel membrane protein OmpA from *Escherichia coli. Journal of Bacteriology* **181***,* 3688-3694.

Kolbe, D. & Eddy, S. (2011). Fast filtering for RNA homology search. *Bioinformatics* **27**, 3102-3109.

Koletzko, B., Baker, S., Cleghorn, G., Neto, U.F., Gopalan, S., Hernell, O., Hock, Q.S., Jirapinyo, P., Lonnerdal, B., Pencharz, P. (2005). Global standard for the composition of infant formula: Recommendations of an espghan coordinated international expert group. *Journal of Pediatric Gastroenterology and Nutrition* **41**, 584–599.

Koraimann, G. (2003). Lytic transglycosylases in macromolecular transport systems of Gram-negative bacteria. *Cellular and Molecular Life Sciences (CMLS)* **60**, 2371-2388.

Kotewicz, M. & Tall B. (2009). Optical maps of subgroups of *Cronobacter* (formerly *Enterobacter sakazakii*) show large chromosomal regions of homology and differences among isolates. *1st International Conference on Cronobacter*, Poster Abstract **60**, Dublin, Ireland.

Kothary, M., McCardell, B., Frazar, C., Deer, D. & Tall, B. (2007). Characterization of the zinc-containing metalloprotease encoded by zpx and development of a species-specific detection method for *Enterobacter sakazakii*. *Applied and Environmental Microbiology* **73**, 4142-4151.

Kucerova, E., Clifton, S. W., Xia, X. Q., Long, F., Porwollik, S., Fulton, L., Fronick, C., Minx, P., Kyung, K., Warren, W., Futon, R., Feng, D., Wollam, A., Shah, N., Bhonagiri, V., Nash, W. E., Hallsworth-Pepin, K., Wilson, R. K., McClelland, M. & Forsythe, S. J. (2010). Genome sequence of *Cronobacter sakazakii* BAA-894 and comparative genomic hybridization analysis with other *Cronobacter* species. *PLoS ONE* **5**: e9556. doi:10.1371/Journal.pone.0009556.

Kucerova, E., Joseph, S. & Forsythe, S. (2011). The *Cronobacter* genus: ubiquity and diversity. *Quality Assurance and Safety of Crops & Foods* **3**, 104-122.

Kuzina, L., Peloquin, J., Vacek, D. & Miller, T. (2001). Isolation and identification of bacteria associated with adult laboratory Mexican fruit flies, *Anastrepha ludens* (Diptera: Tephritidae). *Current Microbiology* **42**, 290-294.

Lagesen, K., Hallin, P., Rodland, E., Staerfeldt, H., Rognes, T. & Ussery, D. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100-3108.

Lai, K. (2001). *Enterobacter sakazakii* infections among neonates, infants, children, and adults: case reports and a review of the literature. *Medicine* **80**, 113-122.

Laslett, D. & Canback, B. (2004). ARAGORN, a program to detect tRNA genesand tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**, 11–16.

Lawrence, J. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinions in Microbiology* **2**, 519-523.

Lawrence, J. G. & Roth, J. R. (1999). Genomic flux: genome evolution by gene loss and acquisition. *Organization of the prokaryotic genome* 263-289.

Lee, Y., Park, J. & Chang, H. (2012). Detection, antibiotic susceptibility and biofilm formation of *Cronobacter* spp. from various foods in Korea. *Food Control* **24**, 225-230.

Leekitcharoenphon, P., Nielsen, E., Kaas, R., Lund, O. & Aarestrup, F. (2014). Evaluation of Whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLoS ONE* **9**, e87991.

Lehner, A., Riedel, K., Eberl, L., Breeuwer, P., Diep, B., & Stephan, R. (2005). Biofilm formation, EPS production and cell-to-cell signalling in various *Enterobacter sakazakii* strains: aspects promoting environmental persistence. *Journal of Food Protection* **68**, 2287-2294.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Liebana, E., Gibbs, M., Clouting, C., Barker, L., Clifton-Hadley, F. A., Pleydell, E., Abdalhamid, B., Hanson, N. D., Martin, L., Poppe, C. & Davies, R. H. (2004). Characterization of β-lactamases responsible for resistance to extended-spectrum cephalosporins in *Escherichia coli* and *Salmonella enterica* strains from food-producing animals in the United Kingdom. *Microbial Drug Resistance* **10** 1-9.

Liu, D., Zhao, W., Fang, W. & Chen, Y. (2012). cPLA2α-mediated actin rearrangements downstream of the Akt signaling is required for *Cronobacter sakazakii* invasion into brain endothelial cells. *Biochemical and Biophysical Research Communications* **417**, 925-930.

Lonnerdal B. **(2012).** Preclinical assessment of infant formula. *Annals of Nutrition and Metabolism* **60**, 196–199.

Lorenc, M., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., Visendi, P., Berkman, P. & Lai, K. et al. (2012). Discovery of Single Nucleotide Polymorphisms in complex genomes using SGSautoSNP. *Biology* **1**, 370-382.

MacLean, L., Pagotto, F., Farber, J. & Perry, M. (2009). Structure of the antigenic repeating pentasaccharide unit of the LPS O-polysaccharide of *Cronobacter sakazakii* implicated in the Tennessee outbreak. *Biochemistry and Cell Biology* **87**, 459-465.

Makhov, A., Hannah, J., Brennan, M., Trus, B., Kocsis, E., Conway, J., Wingfield, P., Simon, M. & Steven, A. (1994). Filamentous Hemagglutinin of *Bordetella pertussis*. *Journal of Molecular Biology*, **241**, 110-124.

Mardis, E. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics* **9**, 387-402.

Mardis, E. (2011). A decade's perspective on DNA sequencing technology. *Nature* **470,** 198-203.

Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M. & Chen, Y. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.

Masood, N., Moore, K., Farbos, A., Hariri, S., Block, C., Paszkiewicz, K., Dickins, B., McNally, A. & Forsythe, S. (2013b). Draft genome sequence of a meningitic isolate of *Cronobacter sakazakii* clonal complex 4, strain 8399. *Genome announcements* **1**, 00833-13.

Masood, N., Moore, K., Farbos, A., Hariri, S., Paszkiewicz, K., Dickins, B., McNally, A. & Forsythe, S. (2013a). Draft genome sequence of the earliest *Cronobacter sakazakii* sequence type 4 strain, NCIMB 8272. *Genome announcements* **1**, 00782-13.

Masse, E., Salvail, H., Desnoyers, G., Arguin, M. (2007). Small RNAs controlling iron metabolism. *Current Opinion in Microbiology* **10**, 140–145.

Masson, P. L., & Heremans, J. F. (1971). Lactoferrin in milk from different species. *Comparative Biochemistry and Physiology part B: Comparative Biochemistry* **39**, 119-113.

Maurelli, A., Fernandez, R., Bloch, C., Rode, C. & Fasano, A. (1998). "Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proceedings of the National Academy of Sciences* **95**, 3943-3948.

May, M. (2005). Early onset neonatal meningitis in Australia and New Zealand, 1992-2002. *Archives of Disease in Childhood - Fetal and Neonatal Edition* **90**, 324-327.

McNally, A., Alhashash, F., Collins, M., Alqasim, A., Paszckiewicz, K., Weston, V. & Diggle, M. (2013). Genomic analysis of extra-intestinal pathogenic *Escherichia coli* urosepsis. *Clinical Microbiology and Infection* **19,** 328-334.

Mergeay, M., Monchy, S., Vallaeys, T., Auquier, V., Benotmane, A., Bertin, P., Taghavi, S., Dunn, J., van der Lelie, D. & Wattiez, R. (2003). *Ralstonia metallidurans*, a bacterium specifically adapted to toxic metals: towards a catalogue of metal-responsive genes. *FEMS Microbiol*ogy **27,** 385–410.

Milner, J. L. & Wood, J. M. (1989). Insertion proQ220: Tn5 alters regulation of proline porter II, a transporter of proline and glycine betaine in *Escherichia coli*. *Journal of Bacteriology* **171**, 947-951.

Mittal, R., Wang, Y., Hunter, C. J., Gonzalez-Gomez, I. & Prasadarao, N. V. (2009). Brain damage in newborn rat model of meningitis by *Enterobacter sakazakii*: a role for outer membrane protein A. *Laboratory Investigation* **89**, 263-277.

Mohan Nair, M. & Venkitanarayanan, K. (2006). Cloning and sequencing of the *ompA* gene of *Enterobacter sakazakii* and development of an *ompA*-targeted PCR for rapid detection of *Enterobacter sakazakii* in infant formula. *Applied and Environmental Microbiology* **72**, 2539-2546.

Mohan Nair, M., Venkitanarayanan, K., Silbart, L. & Kim, K. (2009). Outer Membrane Protein A (OmpA) of *Cronobacter sakazakii* binds fibronectin and contributes to invasion of Human Brain Microvascular Endothelial Cells. *Foodborne Pathogens and Disease* **6**, 495-501.

Morales, C., Lee, M., Hofacre, C. & Maurer, J. (2004). Detection of a novel virulence gene and a *Salmonella* virulence homologue among *Escherichia coli* isolated from broiler chickens. *Foodborne Pathogens and Disease* **1**, 160-165.

Mosso, M., de la Rosa, M., Vivar, C. & Medina, M. (1994). Heterotrophic bacterial populations in the mineral waters of thermal springs in Spain. *Journal of Applied Bacteriology* **77**, 370-381.

Mramba, F., Broce, A. & Zurek, L. (2006). Isolation of *Enterobacter sakazakii* from stable flies, Stomoxys calcitrans L.(Diptera: Muscidae). *Journal of Food Protection* **69**, 671-673.

Mukhopadhyay, R., Rosen, B. P., Phung, L. T., Silver, S. (2002). Microbial arsenic: from geocycles to genes and enzymes. *FEMS Microbiology Reviews* **26**, 311–325.

Mullane, N., O'Gaora, P., Nally, J., Iversen, C., Whyte, P., Wall, P. & Fanning, S. (2008). Molecular analysis of the *Enterobacter sakazakii* O-antigen gene locus. *Applied and Environmental Microbiology* **74**, 3783-3794.

Muller, A., Stephan, R., Fricker-Feer, C. & Lehner, A. (2013). Genetic diversity of *Cronobacter sakazakii* isolates collected from a Swiss infant formula production facility. *Journal of Food Protection* **76**, 883-887.

Muytjens, H. L. & van der Ros-van de Repe, J. (1986). Comparative *in vitro* susceptibilities of eight *Enterobacter* species, with special reference to *Enterobacter sakazakii*. *Antimicrobial Agents and Chemotherapy* **29**, 367–370.

Muytjens, H., Roelofs-Willemse, H. & Jaspar, G. (1988). Quality of powdered substitutes for breast milk with regard to members of the family Enterobacteriaceae. *Journal of clinical Microbiology* **26**, 743-746.

Muytjens, H., Zanen, H., Sonderkamp, H., Kollee, L., Wachsmuth, I. & Farmer, J. (1983). Analysis of eight cases of neonatal meningitis and sepsis due to *Enterobacter sakazakii*. *Journal of Clinical Microbiology* **18**, 115-120.

Nakata, N., Tobe, T., Fukuda, I., Suzuki, T., Komatsu, K., Yoshikawa, M. & Sasakawa, C. (1993). The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the *ompT* and *kcpA* loci. *Molecular Microbiology* **9**, 459-468.

Nazarowec-White, M. & Farber, J. (1997). Incidence, survival, and growth of *Enterobacter sakazakii* in infant formula. *Journal of Food Protection* **60**, 226-230.

Neelam, M., Nawaz, Z. & Riazuddin, S. (1987). Hydrocarbon biodegradation: Biochemical characterization of bacteria isolated from local soils. *Pakistan Journal of Scientific and Industrial Research*, **30**, 382-385.

Nies, D. (2003). Efflux-mediated heavy metal resistance in prokaryotes. *FEMS Microbiology Reviews* **27**, 313-339.

Ofran Y. & Rost B (2007). ISIS: “interaction sites identified from sequence” *Bioinformatics.* **23**, 13-16.

Olsen, A., Arnqvist, A., Hammar, M. & Normark, S. (1993). Environmental regulation of curli production in *Escherichia coli*. *Infectious Agents And Disease* **2**, 272–274.

Orsi, N. (2004). The antimicrobial activity of lactoferrin: current status and perspectives. *Biometals* **17**, 189-196.

Osaili, T. & Forsythe, S. (2009). Desiccation resistance and persistence of *Cronobacter* species in infant formula. *International Journal of Food Microbiology* **136**, 214-220.

Pagotto, F., Nazarowec-White, M., Bidawid, S. & Farber, J. (2003). *Enterobacter sakazakii*: infectivity and enterotoxin production *in vitro* and *in vivo*. *Journal of Food Protection* **66**, 370-375.

Pava-Ripoll, M., Pearson, R., Miller, A. & Ziobro, G. (2012). Prevalence and relative risk of *Cronobacter* spp., *Salmonella* spp., and *Listeria monocytogenes* associated with the body surfaces and guts of individual filth flies. *Applied and Environmental Microbiology* **78**, 7891-7902.

Peigne, C., Bidet, P., Mahjoub-Messai, F., Plainvert, C., Barbe, V., Medigue, C., Frapy, E., Nassif, X. & Denamur, E. et al. (2009). The plasmid of *Escherichia coli* strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic *E. coli* plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. *Infection and Immunity* **77**, 2272-2284.

Pepe, J. C., & Miller, V. L. (1990). The *Yersinia enterocolitica inv* gene product is an outer membrane protein that shares epitopes with *Yersinia pseudotuberculosis* invasin. *Journal Of Bacteriology* **172**, 3780-3789.

Perez-Perez, F. J., Hanson, N. D. (2002). Detection of plasmid-mediated AmpC b-lactamase genes in clinical isolates by using multiplex PCR. *Journal of Clinical Microbiology* **40**, 2153–62.

Petersen, T., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**, 785-786.

Pitout, J., Moland, E., Sanders, C., Thomson, K. & Fitzsimmons, S. (1997). Beta-lactamases and detection of beta-lactam resistance in *Enterobacter* spp. *Antimicrobial Agents and Chemotherapy* **41**, 35-39.

Pizarro-Cerda, J., Moreno, E., Desjardins, M., & Gorvel, J. P. (1997). When intracellular pathogens invade the frontiers of cell biology and immunology. *Histology and histopathology* **12**, 1027-1038.

Power, K., Yan, Q., Fox, E., Cooney, S. & Fanning, S. (2013). Genome sequence of *Cronobacter sakazakii* SP291, a persistent thermotolerant isolate derived from a factory producing powdered infant formula. *Genome announcements* **1**, 00082-13.

Prejit Nanu, E. & Latha, C. (2007). Safety and keeping quality of pasteurized milk under refrigeration. *Journal of Food Science and Technology* **44**, 363- 366.

Proft, T. & Baker, E. (2008). Pili in Gram-negative and Gram-positive bacteria structure, assembly and their role in disease. *Cellular and Molecular Life Sciences* **66**, 613-635.

Pulkkinen, W. S. & Miller, S. I. (1991). A *Salmonella typhimurium* virulence protein is similar to a *Yersinia enterocolitica* invasion protein and a bacteriophage lambda outer membrane protein. *Journal of Bacteriology* **173,** 86-93.

Ratledge, C., Dover, L. G. (2000). Iron metabolism in pathogenic bacteria. *Annual Review of Microbiology* **54**, 881–941.

Ravishankar, S., Juneja, V., Yousef, A. & Juneja, V. (2003). Adaptation or resistance responses of microorganisms to stresses in the food processing environment. *Microbial Stress Adaptation and Food Safety* 105-158.

Richardson, E. & Watson, M. (2013). The automatic annotation of bacterial genomes. *Briefings in Bioinformatics* **14**, 1-12.

Robbe-Saule, V., Algorta, G., Rouilhac, I. & Norel, F. (2003). Characterization of the *RpoS* status of clinical isolates of *Salmonella enterica*. *Applied and Environmental Microbiology* **69**, 4352-4358.

Robey, M., Benito, A., Hutson, R., Pascual, C., Park, S. & Mackey, B. (2001). Variation in resistance to high hydrostatic pressure and *rpoS* heterogeneity in natural isolates of *Escherichia coli* O157:H7. *Applied and Environmental Microbiology* **67**, 4901-4907.

Rodrigue, A., Effantin, G. & Mandrand-Berthelot, M. (2005). Identification of *rcnA* (*yohM*), a nickel and cobalt resistance gene in *Escherichia coli*. *Journal of Bacteriology* **187**, 2912-2916.

Rosenstein, R., Peschel, A., Wieland, B. & Gotz, F. (1992). Expression and regulation of the antimonite, arsenite, and arsenate resistance operon of *Staphylococcus xylosus* plasmid pSX267. *Journal of Bacteriology* **174**, 3676-3683.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. & Barrell, B. (2000). Artemis: sequence visualisation and annotation. *Bioinformatics* **16**, 944-945.

Sabri, M., Leveille, S. & Dozois, C. M. (2006). A *SitABCD* homologue from an avian pathogenic *Escherichia coli* strain mediates transport of iron and manganese and resistance to hydrogen peroxide. *Microbiol*ogy **152**, 745-758.

Sahl, J., Johnson, J., Harris, A., Phillippy, A., Hsiao, W., Thom, K. & Rasko, D. (2011). Genomic comparison of multi-drug resistant invasive and colonising *Acinetobacter baumannii* isolated from diverse human body sites reveals genomic plasticity**.** *BMC Genomics* **12,** 291**.**

Samanta, A., Bera, P., Khatun, M., Sinha, C., Pal, P., Lalee, A. & Mandal, A. (2012). An investigation on heavy metal tolerance and antibiotic resistance properties of bacterial strain *Bacillus* sp. isolated from municipal waste. *Journal of Microbiology and Biotechnology Research* **2**, 178-189.

Sanchez, L., Calvo, M., & Brock, J. H. (1992). Biological role of lactoferrin. *Archives of disease in childhood* **67**, 657.

Scher, K., Romling, U. & Yaron, S. (2005). Effect of heat, acidification, and chlorination on *Salmonella Enterica serovar Typhimurium* cells in a biofilm formed at the air-liquid interface. *Applied and Environmental Microbiology* **71**, 1163-1168.

Scheurwater, E., Reid, C. & Clarke, A. (2008). Lytic transglycosylases: Bacterial space-making autolysins. *The International Journal of Biochemistry & Cell Biology* **40**, 586-591.

Schloss, P., Westcott, S., Ryabin, T., Hall, J., Hartmann, M., Hollister, E., Lesniewski, R., Oakley, B. & Parks, D. et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537-7541.

See, K., Than, H. & Tang, T. (2007). *Enterobacter sakazakii* bacteraemia with multiple splenic abscesses in a 75-year-old woman: a case report. *Age and Ageing* **36**, 595-596.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069.

Severi, E., Hood, D. & Thomas, G. (2007). Sialic acid utilisation by bacterial pathogens. *Microbiology* **153**, 2817-2822.

Shakhnovich, E., King, S. & Weiser, J. (2002). Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of *Neisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract. *Infection and Immunity* **70**, 7161-7164.

Shin, H., Lee, J., Choi, Y. & Ryu, S. (2012). Complete genome sequence of the opportunistic food-borne pathogen *Cronobacter sakazakii* ES15. *Journal of Bacteriology* **194**, 4438-4439.

Silver, S., Budd, K., Leahy, K. M., Shaw, W. V., Hammond, D., Novick, R. P., Willsky, G. R., Malamy, M. H. & Rosenberg, H. (1981). Inducible plasmid-determined resistance to arsenate, arsenite, and antimony(III) in *Escherichia coli* and *Staphylococcus aureus*. *Journal of Bacteriology* **146,** 983-996.

Singamsetty, V., Wang, Y., Shimada, H. & Prasadarao, N. (2008). Outer membrane protein A expression in *Enterobacter sakazakii* is required to induce microtubule condensation in human brain microvascular endothelial cells for invasion. *Microbial Pathogenesis* **45**, 181-191.

Singh, B. (2009). Thermotolerance and multidrug resistance in bacteria isolated from equids and their environment. *Veterinary Record* **164**, 746-750.

Skaar, E. (2010). The battle for iron between bacterial pathogens and their vertebrate hosts. *PLoS Pathogens* **6**, e1000949.

Smith, J. L. (2003). The role of gastric acid in preventing foodborne disease and how bacteria overcome acid conditions. *Journal of Food Protection* **66**, 1292-1303.

Sohanpal, B., El-Labany, S., Lahooti, M., Plumbridge, J. & Blomfield, I. (2004). Integrated regulatory responses of fimB to N-acetylneuraminic (sialic) acid and GlcNAc in *Escherichia coli* K-12. *Proceedings of the National Academy of Sciences* **101**, 16322-16327.

Sonbol, H., Joseph, S., McAuley, C., Craven, H. & Forsythe, S. (2013). Multilocus sequence typing of *Cronobacter* spp. from powdered infant formula and milk powder production factories. *International Dairy Journal* **30**, 1-7.

Soto, G. & Hultgren, S. (1999). Bacterial adhesins: common themes and variations in architecture and assembly. *Journal of Bacteriology* **181**, 1059-1071.

Stamatakis, A., Ludwig, T. & Meier H. (2005). RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees**.** *Bioinformatics***, 21,** 456-463**.**

Stephan, R., Grim, C., Gopinath, G., Mammel, M., Sathyamoorthy, V., Trach, L., Chase, H., Fanning, S. & Tall, B. (2014). Re-examination of the taxonomic status of *Enterobacter helveticus, Enterobacter pulveris* and *Enterobacter turicensis* as members of the genus *Cronobacter* and their reclassification in the genera *Franconibacter* gen. nov. and *Siccibacter* gen. nov. as *Franconibacter helveticus* comb. nov., *Franconibacter pulveris* comb. nov. and *Siccibacter turicensis* comb. nov., respectively. *International Journal of Systematic and Evolutionary Microbiology* **64**, 3402-3410.

Stephan, R., Lehner, A., Tischler, P. & Rattei, T. (2011). Complete genome sequence of *Cronobacter turicensis* LMG 23827, a food-borne pathogen causing deaths in neonates. *Journal of Bacteriology* **193**, 309-310.

Stock, I. & Wiedemann, B. (2002). Natural antibiotic susceptibility of *Enterobacter amnigenus, Enterobacter cancerogenus, Enterobacter gergoviae* and *Enterobacter sakazakii* strains. *Clinical Microbiology and Infection* **8**, 564-578.

Stothard, P. & Wishart, D. (2005). Circular genome visualisation and exploration using CGView. *Bioinformatics* **21**, 537-539.

Stumpe, S., Schmid, R., Stephens, D. L., Georgiou, G., & Bakker, E. P. (1998). Identification of OmpT as the protease that hydrolyzes the antimicrobial peptide protamine before It enters growing cells of *Escherichia coli*. *Journal of Bacteriology*, **180**, 4002-4006.

Sun, Y., Wang, M., Liu, H., Wang, J., He, X., Zeng, J., Guo, X., Li, K., Cao, B. & Wang, L. (2011). Development of an O-antigen serotyping scheme for *Cronobacter sakazakii*. *Applied and Environmental Microbiology* **77**, 2209-2214.

Swain, M., Tsai, I., Assefa, S., Newbold, C., Berriman, M. & Otto, T. (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nature protocols* **7**, 1260-1284.

Tettelin, H., Masignani, V., Cieslewicz, M., Donati, C., Medini, D., Ward, N., Angiuoli, S., Crabtree, J. & Jones, A. et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* **102**, 13950-13955.

Townsend, S., Hurrell, E. & Forsythe, S. (2008). Virulence studies of *Enterobacter sakazakii* isolates associated with a neonatal intensive care unit outbreak. *BMC Microbiology* **8**, 64.

Townsend, S., Hurrell, E., Gonzalez-Gomez, I., Lowe, J., Frye, J., Forsythe, S. & Badger, J. (2007). *Enterobacter sakazakii* invades brain capillary endothelial cells, persists in human macrophages influencing cytokine secretion and induces severe brain pathology in the neonatal rat. *Microbiology* **153**, 3538-3547.

Tsai, H., Liao, C., Huang, Y., Lee, P. & Hsueh, P. (2013). *Cronobacter* infections not from infant formula, Taiwan. *Emerging Infectious Diseases* **19**, 167.

Tseng, T., Tyler, B. & Setubal, J. (2009). Protein secretion systems in bacterial-host associations, and their description in the gene ontology. *BMC Microbiology* **9**, 2.

Turabelidze, G., Lawrence, S., Gao, H., Sodergren, E., Weinstock, G., Abubucker, S., Wylie, T., Mitreva, M. & Shaikh, N. et al. (2013). Precise dissection of an *Escherichia coli* O157:H7 outbreak by Single Nucleotide Polymorphism analysis. *Journal of Clinical Microbiology* **51**, 3950-3954.

van Acker, J., de Smet, F., Muyldermans, G., Bougatef, A., Naessens, A. & Lauwers, S. (2001). Outbreak of necrotising enterocolitis associated with *Enterobacter sakazakii* in powdered milk formula. *Journal of Clinical Microbiology* **39**, 293-297.

Van Os, M., Van Wikselaar, P. & Spoelstra, S. (1996). Formation of biogenic amines in well fermented grass silages. *The Journal of Agricultural Science* **127**, 97-107.

van Tonder, A., Mistry, S., Bray, J., Hill, D., Cody, A., Farmer, C., Klugman, K., von Gottberg, A. & Bentley, S. et al. (2014). Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Computational Biology* **10**, e1003788.

Vandenplas Y. (2002). Oligosaccharides in infant formula. *British Journal of Nutrition* **87**, 293–296.

Vimr, E., Kalivoda, K., Deszo, E. & Steenbergen, S. (2004). Diversity of microbial sialic acid metabolism. *Microbiology and Molecular Biology Reviews* **68**, 132-153.

Walsh, D., Molloy, C., Iversen, C., Carroll, J., Cagney, C., Fanning, S. & Duffy, G. (2011). Survival characteristics of environmental and clinically derived strains of *Cronobacter sakazakii* in infant milk formula (IMF) and ingredients. *Journal of Applied Microbiology* **110**, 697-703.

Wang, B. (2009). Sialic acid is an essential nutrient for brain development and cognition. *Annual Review of Nutrition* **29**, 177-222.

Wang, Y. & Kim, K. (2002). Role of *OmpA* and *IbeB* in *Escherichia coli* K1 invasion of brain microvascular endothelial cells *in vitro* and *in vivo*. *Pediatric Research* **51**, 559-563.

Weber, H., Polen, T., Heuveling, J., Wendisch, V. & Hengge, R. (2005). Genome-wide analysis of the general stress response network in *Escherichia coli*: sigmaS-dependent genes, promoters, and sigma factor selectivity. *Journal of Bacteriology* **187**, 1591-1603.

Weissman, S., Moseley, S., Dykhuizen, D. & Sokurenko, E. (2003). *Enterobacterial* adhesins and the case for studying SNPs in bacteria. *Trends in Microbiology* **11**, 115-117.

Wilks, A. (2002). Heme oxygenase: evolution, structure, and mechanism. *Antioxidants and Redox Signaling* **4**, 603-614.

Williams, B., Morlin, G., Valentine, N. & Smith, A. (2001). Serum resistance in an invasive, nontypeable *Haemophilus influenzae* strain. *Infection and Immunity* **69**, 695-705.

Willis, J. & Robinson, J. (1988). *Enterobacter sakazakii* meningitis in neonates. *The Pediatric Infectious Disease Journal* **7**, 196-199.

Wilson, J. (2002). Mechanisms of bacterial pathogenicity. *Postgraduate Medical Journal* **78**, 216-224.

Wong, C., Mooney, J., Brandt, J., Staples, A., Jelacic, S., Boster, D., Watkins, S. & Tarr, P. (2012). Risk factors for the hemolytic uremic syndrome in children infected with *Escherichia coli* O157:H7: a multivariable analysis. *Clinical Infectious Diseases* **55**, 33-41.

Wood, T., Gonzalez Barrios, A., Herzberg, M. & Lee, J. (2006). Motility influences biofilm architecture in *Escherichia coli*. *Applied Microbiology and Biotechnology* **72**, 361-367.

Xu, C., Shi, W. & Rosen, B. (1996). The chromosomal *arsR* gene of *Escherichia coli* encodes a trans-acting metalloregulatory protein. *Journal of Biological Chemistry* **271**, 2427-2432.

Yan, Q., Condell, O., Power, K., Butler, F., Tall, B. & Fanning, S. (2012). *Cronobacter* species (formerly known as *Enterobacter sakazakii*) in powdered infant formula: a review of our current understanding of the biology of this bacterium. *Journal of Applied Microbiology* **113**, 1-15.

Zaneveld, J., Nemergut, D. & Knight, R. (2008). Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology* **154**, 1-15.

Zavilgelsky, G. B., & Rastorguev, S. M. (2009). Antirestriction proteins ArdA and Ocr as efficient inhibitors of type I restriction-modification enzymes. *Molecular biology* **43**, 241-248.

Zerbino, D. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**, 821-829.

Zhou, Y., Tao, J., Yu, H., Ni, J., Zeng, L., Teng, Q., Kim, K., Zhao, G., Guo, X. & Yao, Y. (2012). Hcp family proteins secreted via the type VI secretion system coordinately regulate *Escherichia coli* K1 interaction with human brain microvascular endothelial cells. *Infection and Immunity* **80**, 1243-1251.

APPENDICES

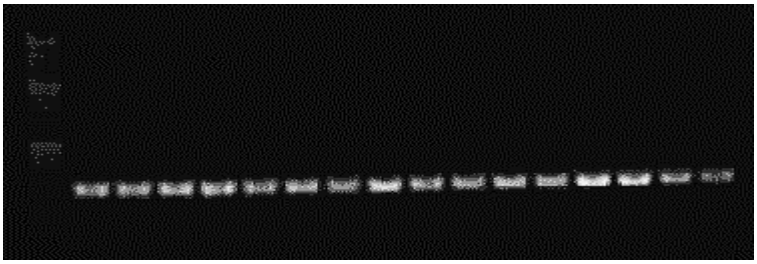


Figure S1. PCR amplification of the *ompA* in selected *C. sakazakii* CC4 and non-CC4 genomes

Five microlitre of the PCR product was loaded into each well. Lane 1 contains 1 KB DNA ladder (promega). Lane 2 contains PCR amplified DNA from *C. sakazakii* 658 as a positive control. Lane 3-15 contain the PCR products of *C. sakazakii* isolates 1, 2, 4, 5. 6, 12, 20, 520, 553, 555, 557, 558, 580 and 680 respectively. As seen in the figure all the strains were positive for ompA gene as 500 bp DNA product was yielded for each strain. A negative control was also run containing only the PCR reaction mixture without any template DNA.

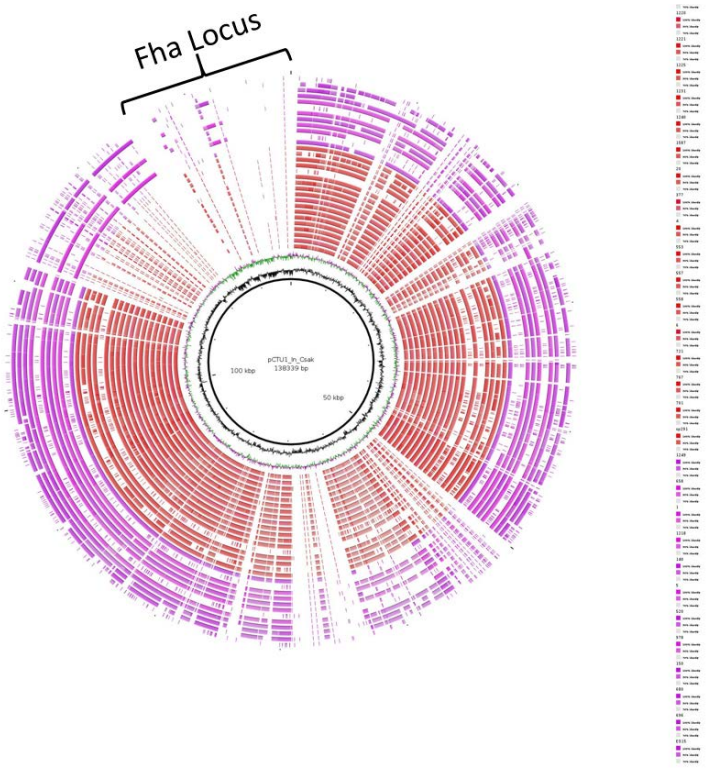


Figure S2. BRIG alignment of the pCTU1 plasmid with *C. sakazakii* isolates

The image above represents the alignment of the pCTU1 plasmid (138 kb) with 18 *C. sakazakii* CC4 (red rings) and 11 *C. sakazakii* non-CC4 (purple rings). The legend on the left hand side of the figure indicates ring for each isolate. The central solid black circle represent the reference plasmid backbone, the second and third rings from inside out represent the GC contents and GC skew respectively.

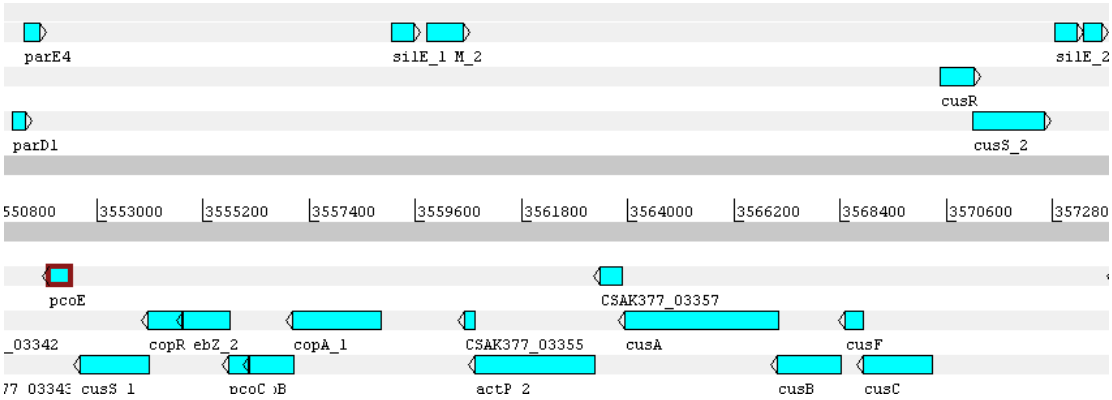


Figure S3 . Clustering of *cus/sil* and *pco* genes.

The genome of *C. sakazakii* isolate 377 was analysed using Artemis. The screen shot indicating the clustering of cus/sil and pco genes potentially linked with copper and silver resistance or efflux.

Table S1. Virulence and environmental fitness associated traits in sequenced *C. sakazakii* isolates

		Isolate	6	20	377	553	557	558	701	721	767	1219	1220	1221	1225	1231	1240	SP291	4	1587	658	1218	978	1	5	680	520	696	150	1249	140	ES15
		ST	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	15	109	1	1	3	8	8	8	12	12	16	31	40	125
		CC	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1	1	3	8	8	8			16	31	45	100
Category	**Gene/Locus_tag**																															
	ESA_01970		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+
	ESA_01971		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+
	ESA_01972		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+
	ESA_01973		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+
	ESA_01974		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	+	-	-	+
	ESA_01975		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_01976		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+
Fimbrial region 1 [ESA_01970-6]	ESA_02342		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02343		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02344		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02345		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Fimbrial region 2 [ESA_02342-5]	ESA_02538		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02539		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02540		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02541		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02542		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Fimbrial region 3 [ESA_02538-42]	ESA_02795		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02796		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02797		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02798		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02799		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Fimbrial region 4 [ESA_02795-99]	ESA_03512		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

	ESA_03513		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03514		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+
	ESA_03515		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+
	ESA_03516		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+
	ESA_03517		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+
	ESA_03518		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+
	ESA_03519		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
	ESA_03520		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04067		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
Fimbrial region 5 [ESA_03512-20], Beta fimbriae	ESA_04068		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04069		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04070		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04071		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04072		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04073		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Fimbrial region 6 [ESA_04067-73]	ESA_03812		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03813		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03814		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03815		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	CTU_36390		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	+	+	+	+	-	-
Fimbrial region 7 [ESA_03812-15]	CTU_36400		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CTU_36410		-	+	+	-	-	+	+	-	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-	-	+	+	-	+	-	-
	yfc; SCTU_36420		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	-
	yfcU, CTU_36430		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+	-	+	+	+	-
	CTU_36440		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	-
Fimbrial region 8 [ctu_36390-36450], ? fimbriae	smfA, CTU_36450		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	-
	ESA_03231 putative major pilin subunit		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03232		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
	ESA_03232		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Fimbrial region 9 [ESA_03231-33	csgG, CTU_16160		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

	csgF, CTU_16170		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	csgE, CTU_16180		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	csgD, CTU_16190		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	csgB, CTU_16200		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	csgA, CTU_16210		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	csgC, CTU_16220		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CTU_16230		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	fimA		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fimbrial region 10 [CTU-16160-230], curli fimbriae	fimB		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	fimC		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	fimD		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	fimE		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	fimF		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	fimG		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	fimH		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Type 1 fimbrial operon, fimABCDEFGHI, E. coli str. K-12 substr. MG1655	fimI		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ompA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ompX		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Invasion	Inv		+	-	+	-	-	+	+	-	+	+	+	+	-	+	+	+	-	+	+	+	-	-	+	-	-	-	-	+	-	+
	Ctu_1p01140		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-
	Ctu_1p01150		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-
	Ctu_1p01160		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-
	Ctu_1p01170		-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-
	Ctu_1p01180		-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-
	Ctu_1p01190		-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-
	Ctu_1p01200		-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-
FHA locus	Ctu_1p01210		-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-
	ESA_00141		-	+	-	-	+	+	-	+	-	+	+	+	+	-	+	-	+	-	-	+	-	-	+	-	-	+	+	-	-	-
	ESA_00142		+	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	+	-	-	+	+	+	-	+	-	+	+	+	-	+
T6SS	ESA_00140		-	+	-	-	+	+	-	+	-	+	+	+	+	+	+	-	+	+	-	+	-	-	+	-	+	+	-	-	+	-

	ESA_00143		-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	+	-	+	-	-
	ESA_00144		-	+	+	-	+	+	+	-	+	-	-	-	-	-	-	-	+	-	+	-	+	-	+	+	+	-	-	-	+	-
	ESA_00145		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02035		+	+	+	+	+	+	+	+	+	+	-	-	-	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
	ESA_02036		-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	ESA_02037		+	+	+	+	+	+	+	+	+	+	-	-	-	+	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+
	ESA_02038		+	+	+	+	+	+	-	+	+	+	-	-	-	+	-	-	+	-	+	-	+	+	+	+	+	+	-	-	+	+
	ESA_02039		+	-	-	-	+	-	+	+	+	+	-	+	-	+	-	-	+	-	+	-	-	+	-	+	+	+	+	-	-	-
	ESA_02040		+	-	-	-	+	-	+	+	+	+	-	+	-	+	-	-	+	-	+	-	-	+	-	+	-	+	+	-	-	-
	ESA_02735		+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	-	+	+	+	+	+	+	+	+	-	+	+	-
	ESA_02736		+	+	+	+	+	+	-	+	-	+	-	-	-	-	-	-	+	-	+	+	+	+	+	+	-	+	-	+	+	+
	ESA_02737		+	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	+	-	+	+	+	+	+	+	-	+	-	+	+	+
	ESA_02738		+	+	+	+	+	+	+	+	-	+	-	-	-	+	-	-	+	-	+	+	+	+	+	+	-	+	-	+	+	+
	ESA_02739		+	-	-	-	+	-	+	-	-	+	+	-	-	+	-	-	+	-	+	+	-	+	-	+	-	-	-	+	-	-
	ESA_02740		-	-	+	-	+	+	-	+	-	+	-	-	-	-	-	-	+	-	+	+	-	-	-	-	-	+	-	-	-	+
	ESA_pESA3p05491		+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_pESA3p05492		+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	+	+	+	+	+	+	-	-	+	+	+	+	+
	ESA_pESA3p05493		+	+	+	-	+	+	+	-	-	+	-	+	-	-	-	-	+	-	+	-	+	+	+	+	+	-	+	-	+	+
	ESA_pESA3p05494		+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	+	-	-	+	-	+	-	-	-
	ESA_pESA3p05495		+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	+	-	+	-	-	-	-	-	-
	ESA_pESA3p05496		+	-	-	-	+	-	-	-	+	+	+	+	-	+	-	-	+	-	+	-	-	+	-	+	-	-	+	-	-	-
	ESA_pESA3p05497		+	-	-	-	+	-	-	-	+	-	+	+	-	-	-	-	+	-	+	-	-	+	-	+	-	-	+	-	-	-
	ESA_pESA3p05498		-	-	-	-	+	-	-	-	-	-	+	-	-	+	-	-	+	-	-	-	-	+	-	+	-	-	+	-	-	-
	ESA_pESA3p05499		-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-
	ESA_pESA3p05500		+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	+	-	+	-	-	+	-	-	-
	ESA_pESA3p05501		-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	+	-	-	-	-	-	-
	ESA_pESA3p05502		-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-
	ESA_pESA3p05503		+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	+	-	-	-	-	-	-
	ESA_pESA3p05504		+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	+	-	+	-	-	-	-	-	-
	ESA_pESA3p05505		-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-

	ESA_pESA3p05506		+	-	-	-	+	-	+	+	-	-	-	+	-	-	-	+	+	-	-	+	-	+	-	+	+	+	+	+	-	-
	ESA_pESA3p05507		+	-	-	+	+	-	+	+	+	-	+	+	+	+	+	-	+	+	+	+	-	+	-	+	+	+	+	+	-	-
	ESA_pESA3p05508		-	+	+	+	+	+	+	+	-	-	+	-	-	+	-	-	+	-	-	+	+	-	+	+	+	+	-	+	+	+
	ESA_pESA3p05509		-	+	+	+	+	+	+	+	+	-	+	-	-	+	-	-	+	+	-	+	+	-	+	+	+	+	-	+	+	+
	ESA_pESA3p05510		+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Ctu_12090		+	-	-	-	-	-	-	-	-	+	+	-	-	+	-	+	+	-	+	+	-	+	-	+	-	-	-	+	-	-
	ctu_12100		+	+	+	-	-	+	-	-	-	+	+	-	-	+	-	-	+	-	+	+	+	+	+	+	+	-	-	+	+	+
	clpB		+	+	+	-	-	+	-	+	-	+	-	-	-	-	-	-	+	-	+	-	+	+	+	+	+	+	-	-	+	+
	Yhch		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nanK		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nanT		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nanA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nanR		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nanC		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nanE		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nagA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	nagB		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	neuC		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	siaP		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	siaQ		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sialic acid	siaM		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00791		+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00792		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02728		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02729		+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02730		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02731		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00793		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00794		+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
Enterobaction	ESA_00795		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

	ESA_00796		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00797		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00798		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00799		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00800		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02727		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
	ESA_03187		+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+
	ESA_03188		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03189		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ES-_03190		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Siderophore	ESA_02242		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	iucA		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
	iucB		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
	iucC		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
	iucD		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
Aerobaction	iutA		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-
	ESA_00281		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00282		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	flhE (ESA_01356)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	fliD (ESA_01287)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	flgJ, ESA_02266		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	bcsB		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	bcsE		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	bcsF		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Biofilm	bcsZ		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Colanic acid	ESA_01155-01175; wzABCKM		+	+	+	+	+ (12 kb)	+	+	+	+	+	+	+	+	+	+	+ (17 kb)	+	+	+	+	+	+	+	+	+	+ (17 kb)	+	+	+ (17 kb)	+
Stress response	rpoS		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	TrkH		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
E. coli osmotolerance homologues	TrkG		+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	-	+	+	+	-	+	+	+	+	+	+	+	-	+

	TrkA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	TrkE		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	kdpA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	kdpB		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	kdpC		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	kdpD		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	kdpE		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	PhoP		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	PhoQ		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ompC		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ompF		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	envZ		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ompR		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	kup		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02131		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_01706		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_04214		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_pESA305450		+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	-	+	+	+	-	+	+	+	+	+	+	+	+	+
	ESA_01226		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00673		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Prop homologues	ESA_03328		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00102 (Hemolysin activator protein precursor)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00432 (Predicted membrane protein hemolysin III homologue)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_00643 (Hemolysins and related proteins containing CBS domains)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02810 (Hemolysin expression modulating protein)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_02937 (Putative hemolysin (smaller - ~1790 bp))		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ESA_03540 (21 kDa hemolysin precursor)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	hlyA		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	hlyB		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Haemolysin genes	hlyC		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

	hlyD		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ESA_04236 (silE)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04237 (cusS)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04238 (cusR)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04239 (cusC)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04240 (cusF)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04241 (cusB)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04242 (cusA)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	-	-	+	-	-	-	-	-
	ESA_04243 (CopG)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04244 (silP_1)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04245 (silP_2)		+	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04248 (pcoE)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04249 (pcoA)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04250 (pcoB)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04251 (CopC)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04252 (PcoE_1)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04253 (PcoR)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	ESA_04254 (PcoS)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
Copper & silver resistance	ESA_04255 (PcoE_2)		+	-	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	-	-	-
	sodA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Hha		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Zn metallo		+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Pla		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	rspA (ESA_pESA3p05473)		-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	+	+	+	+	+	-
	cpa (ESA_pESA3p05434)		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	+	+	+	+	+	-
	hlyA		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	hlyB		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	hlyC		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Miscellaneous	hlyD		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

	rcK		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	pagC		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S2. Unique genes in *C. sakazakii* CC4 isolate 557 against *C. sakazakii* non-CC4 isolate 658

No	Locus_tag/gene	start	end	Annotation
1	CSAK557_00247	183951	184301	Phage-related protein
2	papC_1	321444	323951	Outer membrane usher protein papC precursor
3	papH_1	324024	324584	PAP fimbrial minor pilin protein precursor
4	smfA	324656	325189	Fimbria A protein precursor
5	CSAK557_00627	492999	493751	hypothetical protein
6	intS_1	762314	763513	Putative prophage CPS-53 integrase
7	rfbD	895112	896008	dTDP-4-dehydrorhamnose reductase
8	CSAK557_01138	897724	898731	Glycosyl transferasesgroup 1
9	wbbL	898721	899485	dTDP-Rha:alpha-D-GlcNAc-pyrophosphate polyprenol, alpha-3-L-rhamnosyltransferase
10	leuE	961085	961729	Leucine efflux protein
11	CSAK557_01226	961957	962727	hypothetical protein
12	CSAK557_01620	1255030	1255965	hypothetical protein
13	CSAK557_01621	1255962	1256417	hypothetical protein
14	CSAK557_01630	1262993	1263475	hypothetical protein
15	CSAK557_01631	1263475	1263702	hypothetical protein
16	CSAK557_01632	1263832	1264056	hypothetical protein
17	CSAK557_01683	1298645	1299625	Transposase DDE domain protein
18	ykfI	1302939	1303280	Toxin YkfI
19	yfjZ	1303301	1303618	Putative antitoxin YfjZ
20	CSAK557_01690	1303639	1303860	hypothetical protein
21	ykfG_2	1303869	1304345	hypothetical protein
22	CSAK557_01692	1304361	1304819	Antirestriction protein
23	ykfF	1304921	1305160	hypothetical protein
24	CSAK557_01694	1305237	1305704	hypothetical protein
25	yafY	1305728	1306171	putative lipoproteinyafY precursor
26	CSAK557_01696	1306171	1306407	hypothetical protein
27	CSAK557_01697	1306444	1307145	hypothetical protein
28	CSAK557_01698	1307362	1308183	hypothetical protein
29	era_2	1308275	1309138	GTPase Era

30	CSAK557_01884	1453553	1453924	hypothetical protein
31	CSAK557_01941	1506048	1506770	hypothetical protein
32	CSAK557_02001	1559194	1559880	hypothetical protein
33	CSAK557_02002	1559877	1561226	hypothetical protein
34	CSAK557_02008	1565465	1566094	Metallo-beta-lactamase superfamily protein
35	**CSAK557_02176**	**1705806**	**1706288**	**hypothetical protein homologue of cssp291_10445**
36	ugl	1936512	1937699	Unsaturated glucuronyl hydrolase
37	lacY_1	1937716	1938975	Lactose-proton symport
38	CSAK557_02434	1938993	1940837	hypothetical protein
39	CSAK557_02444	1949015	1949917	hypothetical protein
40	CSAK557_02445	1950088	1950906	hypothetical protein
41	CSAK557_02447	1952469	1953518	site-specific tyrosine recombinase XerC
42	CSAK557_02756	2235475	2236521	hypothetical protein
43	CSAK557_02757	2236518	2237129	hypothetical protein
44	fimH	2427497	2428483	hypothetical protein
45	CSAK557_03180	2610127	2611212	hypothetical protein
46	CSAK557_03287	2705906	2707072	hypothetical protein
47	CSAK557_03838	3193036	3193263	hypothetical protein
48	CSAK557_03839	3193315	3194415	hypothetical protein
49	CSAK557_03840	3194559	3195047	hypothetical protein
50	CSAK557_03841	3195605	3195718	hypothetical protein
51	CSAK557_03844	3197550	3197831	hypothetical protein
52	**CSAK557_03870**	**3222005**	**3222274**	**hypothetical protein homologue of cssp291_18210**
53	clpV1_3	3228066	3230822	hypothetical protein
54	elfA_2	3230886	3231416	hypothetical protein
55	papD_1	3231503	3232261	Chaperone protein papD precursor
56	htrE_2	3232349	3234949	Heat shock protein E
57	CSAK557_03883	3234934	3235530	fimbrial-like adhesinprotein SfmF
58	hcp1_2	3237828	3238313	hypothetical protein
59	CSAK557_03887	3238448	3239026	hypothetical protein
60	CSAK557_03888	3239013	3240359	hypothetical protein
61	CSAK557_03943	3284696	3285562	hypothetical protein

62	intA_3	3285737	3286288	Prophage CP4-57 integrase
63	CSAK557_03949	3289374	3289808	hypothetical protein
64	CSAK557_03951	3290767	3290916	hypothetical protein
65	CSAK557_03952	3291632	3292111	hypothetical protein
66	CSAK557_03953	3292111	3292386	hypothetical protein
67	CSAK557_04237	3539876	3540067	hypothetical protein
68	arsA_1	3540427	3540780	Arsenical pump-driving ATPase
69	ywnH	3540982	3541509	Putative phosphinothricin acetyltransferase YwnH
70	aqpZ2	3541559	3541801	Aquaporin Z 2
71	CSAK557_04241	3541913	3542011	hypothetical protein
72	CSAK557_04558	3807335	3808522	FRG domain protein
73	CSAK557_04559	3808928	3809152	hypothetical protein
74	CSAK557_04560	3809364	3810098	hypothetical protein
75	CSAK557_04561	3810340	3811185	hypothetical protein
76	CSAK557_04562	3811175	3811330	hypothetical protein
77	CSAK557_04563	3811296	3811775	conjugal transfer protein TrbJ
78	CSAK557_04564	3811963	3812943	conjugal transfer protein TrbL
79	arsD	3812940	3813302	Arsenical resistanceoperon trans-acting repressor ArsD
80	arsA_2	3813320	3815071	Arsenical pump-driving ATPase
81	arsB	3815119	3816408	Arsenic efflux pump protein
82	arsC	3816421	3816756	Arsenate reductase
83	kfoC_2	3817039	3819135	Chondroitin polymerase
84	CSAK557_04570	3819214	3820236	phage major capsid protein, P2 family
85	CSAK557_04571	3820297	3821121	Phage capsid scaffolding protein (GPO) serine peptidase
86	CSAK557_04572	3821163	3822323	hypothetical protein
87	CSAK557_04573	3822399	3822875	hypothetical protein
88	CSAK557_04574	3823165	3823608	T5orf172 domain protein
89	CSAK557_04610	3849651	3850274	hypothetical protein
90	CSAK557_04611	3850274	3851302	hypothetical protein
91	CSAK557_04612	3851357	3852403	hypothetical protein
92	CSAK557_04613	3852408	3853403	P pilus assembly protein, pilin FimA
93	CSAK557_04614	3853415	3854752	Glycosyltransferase (GlcNAc)

94	CSAK557_04615	3854746	3855345	hypothetical protein
95	oprF_2	3855349	3856968	Outer membrane porinF precursor
96	CSAK557_04617	3856961	3857446	hypothetical protein
97	CSAK557_04618	3857510	3858463	ribosomal-protein-alanine acetyltransferase
98	ygaP	3858530	3859066	Inner membrane protein ygaP
99	ygaV	3859068	3859373	putative HTH-type transcriptional regulator ygaV
100	CSAK557_04621	3859579	3860772	galactoside permease
101	CSAK557_04622	3860839	3861192	putative membrane protein
102	CSAK557_04623	3861197	3861676	hypothetical protein
103	ytcD	3861727	3862110	putative HTH-type transcriptional regulator ytcD
104	cpnA	3862211	3862939	Cyclopentanol dehydrogenase
105	ripA	3863123	3863836	HTH-type transcriptional repressor of iron proteins A
106	CSAK557_04627	3863989	3864381	hypothetical protein
107	CSAK557_04628	3864454	3865140	hypothetical protein
108	CSAK557_04629	3865340	3866158	transcriptional activator FtrA
109	CSAK557_04630	3866233	3866595	Cupin domain protein
110	CSAK557_04631	3866592	3867053	hypothetical protein
111	CSAK557_04632	3867590	3867808	hypothetical protein
112	CSAK557_04633	3867844	3868392	Restriction endonuclease
113	CSAK557_04634	3868398	3868625	hypothetical protein
114	CSAK557_04635	3868622	3870733	Type IV secretory pathway, VirD4 components
115	CSAK557_04636	3870792	3871481	integrating conjugative element membrane protein, family
116	CSAK557_04637	3871674	3871991	integrative conjugative element protein, RAQPRD family
117	CSAK557_04638	3872576	3872875	site-specific tyrosine recombinase XerC
118	CSAK557_04639	3873718	3874416	hypothetical protein
119	CSAK557_04640	3874419	3874832	hypothetical protein
120	CSAK557_04641	3874832	3875131	hypothetical protein
121	CSAK557_04642	3875697	3875960	Transposase
122	CSAK557_04643	3875987	3876817	insertion element IS2transposase InsD
123	CSAK557_04644	3877011	3877748	hypothetical protein
124	CSAK557_04645	3877876	3878091	putative transcriptional regulator
125	CSAK557_04646	3878519	3879577	hypothetical protein

126	CSAK557_04647	3880198	3880776	hypothetical protein
127	CSAK557_04648	3880918	3881655	hypothetical protein
128	CSAK557_04649	3881676	3882362	hypothetical protein
129	CSAK557_04650	3882507	3882812	hypothetical protein
130	CSAK557_04651	3882848	3883210	hypothetical protein
131	CSAK557_04652	3883456	3884904	hypothetical protein
132	CSAK557_04653	3884931	3885347	conjugative transferregion lipoprotein
133	CSAK557_04654	3885347	3888163	Type IV secretory pathway, VirB4 components
134	CSAK557_04655	3888160	3888567	hypothetical protein
135	CSAK557_04656	3888834	3889241	integrating conjugative element protein, family
136	CSAK557_04657	3889238	3890209	integrating conjugative element protein, family
137	CSAK557_04658	3891152	3892066	putative restrictionendonuclease
138	repA_1	3892261	3892719	Replication-associated protein
139	CSAK557_04660	3893412	3893519	hypothetical protein
140	CSAK557_04661	3893605	3893769	type IV conjugative transfer system pilin TraA
141	CSAK557_04662	3893766	3894251	conjugal transfer protein TraM
142	CSAK557_04663	3895213	3895827	site-specific tyrosine recombinase XerC
143	CSAK557_04664	3896708	3897472	hypothetical protein
144	CSAK557_04665	3897555	3897731	hypothetical protein
145	CSAK557_04666	3897934	3898554	hypothetical protein
146	CSAK557_04667	3898909	3899151	hypothetical protein
147	CSAK557_04668	3899271	3899573	Transposase
148	CSAK557_04669	3899542	3900663	Transposase
149	CSAK557_04670	3901014	3901676	hypothetical protein
150	CSAK557_04676	3908810	3910477	hypothetical protein
151	CSAK557_04677	3910480	3911028	hypothetical protein
152	CSAK557_04678	3911000	3911725	hypothetical protein
153	CSAK557_04679	3911715	3912254	Bacteriophage tail assembly protein
154	CSAK557_04680	3912233	3912670	hypothetical protein
155	CSAK557_04681	3912934	3913620	hypothetical protein
156	CSAK557_04682	3913613	3914200	hypothetical protein
157	CSAK557_04683	3914209	3915564	Tail fiber protein

158	intA_4	3915620	3916861	Prophage CP4-57 integrase
159	CSAK557_04685	3916845	3919382	hypothetical protein
160	CSAK557_04686	3919363	3920418	hypothetical protein
161	CSAK557_04706	3939959	3940309	hypothetical protein
162	CSAK557_04707	3940306	3942291	hypothetical protein
163	CSAK557_04708	3942325	3942600	Transposase
164	CSAK557_04709	3942645	3943022	IS1 transposase
165	CSAK557_04710	3943201	3945396	putative P-loop ATPase
166	CSAK557_04711	3945463	3946023	hypothetical protein
167	CSAK557_04712	3946109	3946321	hypothetical protein
168	CSAK557_04713	3946501	3946731	Virulence-associatedprotein
169	vapC	3946728	3947144	tRNA(fMet)-specific endonuclease VapC
170	CSAK557_04715	3947193	3947690	hypothetical protein
171	CSAK557_04716	3947694	3948392	hypothetical protein
172	CSAK557_04717	3948896	3949012	hypothetical protein
173	CSAK557_04718	3949009	3950403	integrating conjugative element protein, family
174	CSAK557_04719	3950405	3950743	hypothetical protein
175	CSAK557_04720	3950754	3952109	hypothetical protein
176	traC	3952525	3953106	DNA primase TraC
177	CSAK557_04722	3953198	3954151	hypothetical protein
178	CSAK557_04723	3954282	3954965	Type I restriction-modification system methyltransferase subunit
179	CSAK557_04724	3955049	3955327	hypothetical protein
180	CSAK557_04725	3955324	3955971	hypothetical protein
181	CSAK557_04726	3956342	3956965	Integrase
182	CSAK557_04727	3957097	3958176	type II secretion system protein F
183	CSAK557_04728	3958186	3958782	hypothetical protein
184	CSAK557_04729	3958787	3959266	invasion protein IagB
185	CSAK557_04730	3959263	3959931	Type IV leader peptidase family protein
186	CSAK557_04731	3959931	3961082	hypothetical protein
187	CSAK557_04732	3960990	3961547	hypothetical protein
188	CSAK557_04733	3961504	3961758	conjugal transfer protein TraR
189	CSAK557_04734	3961751	3963466	integrating conjugative element, PFGI_1 class, ParB family protein

190	dnaB_2	3963463	3963834	Replicative DNA helicase
191	CSAK557_04736	3963834	3964454	hypothetical protein
192	CSAK557_04737	3964826	3965560	hypothetical protein
193	CSAK557_04738	3965635	3966354	hypothetical protein
194	CSAK557_04739	3966347	3968173	hypothetical protein
195	pifC_1	3968136	3968504	Transcriptional repressor pifC
196	dnaB_3	3968851	3969303	Replicative DNA helicase
197	CSAK557_04742	3969290	3969847	hypothetical protein
198	CSAK557_04743	3969844	3970170	hypothetical protein
199	CSAK557_04744	3970170	3972230	Phage-related minor tail protein
200	CSAK557_04745	3972418	3972675	hypothetical protein
201	CSAK557_04746	3972822	3973154	hypothetical protein
202	CSAK557_04747	3973154	3973495	hypothetical protein
203	CSAK557_04748	3973496	3973789	Phage holin family 2
204	CSAK557_04749	3973799	3974254	hypothetical protein
205	CSAK557_04750	3974251	3974922	hypothetical protein
206	CSAK557_04751	3975119	3975610	hypothetical protein
207	CSAK557_04752	3975683	3975886	hypothetical protein
208	ssb_2	3975904	3976452	Helix-destabilizing protein
209	CSAK557_04754	3976531	3976773	hypothetical protein
210	CSAK557_04755	3976757	3977005	hypothetical protein
211	CSAK557_04756	3977115	3977396	hypothetical protein
212	CSAK557_04757	3977594	3979132	integrating conjugative element relaxase, PFGI-1 class
213	xerD_1	3979173	3980126	Tyrosine recombinaseXerD
214	CSAK557_04759	3980752	3981075	hypothetical protein
215	CSAK557_04760	3981053	3981856	hypothetical protein
216	topB_3	3981908	3983584	DNA topoisomerase 3
217	CSAK557_04762	3983598	3984320	integrating conjugative element protein, family
218	CSAK557_04763	3984802	3985110	hypothetical protein
219	CSAK557_04764	3985111	3985722	site-specific tyrosine recombinase XerC
220	CSAK557_04765	3987607	3987798	hypothetical protein
221	CSAK557_04766	3987816	3987983	hypothetical protein

222	CSAK557_04767	3988085	3988882	hypothetical protein
223	CSAK557_04768	3989401	3990417	hypothetical protein
224	CSAK557_04769	3990505	3991722	hypothetical protein
225	CSAK557_04770	3991732	3992262	hypothetical protein
226	CSAK557_04771	3992446	3993360	hypothetical protein
227	CSAK557_04772	3994120	3995460	hypothetical protein
228	CSAK557_04773	3995904	3996902	Plasmid encoded RepAprotein
229	traJ	3997755	3998108	Relaxosome protein
230	CSAK557_04775	3998190	3998546	conjugal transfer protein TrbJ
231	CSAK557_04776	3999188	3999619	Bacteriophage CI repressor helix-turn-helix domain protein
232	CSAK557_04777	4000198	4000512	hypothetical protein
233	CSAK557_04778	4000730	4002592	Bacteriophage replication gene A protein (GPA)
234	CSAK557_04779	4002711	4002935	hypothetical protein
235	CSAK557_04780	4002909	4003178	Ogr/Delta-like zinc finger
236	CSAK557_04781	4003229	4003561	phage portal protein,PBSX family
237	ompF_3	4003584	4003829	Porin OmpF
238	papD_2	4003856	4004614	Chaperone protein papD precursor
239	papH_2	4004658	4005212	PAP fimbrial minor pilin protein precursor
240	CSAK557_04785	4005242	4006165	putative fimbrial subunit SteE
241	CSAK557_04786	4006240	4006653	putative minor fimbrial subunit StfF
242	mrpA	4006662	4007189	Major MR/P fimbria protein precursor
243	CSAK557_04788	4007432	4008715	O-antigen ligase RfaL
244	ygiW_3	4008791	4009219	hypothetical protein
245	rcnA	4009391	4010533	Nickel/cobalt effluxsystem rcnA
246	csoR	4010546	4010815	Copper-sensitive operon repressor
247	CSAK557_04792	4010921	4012219	enterobactin exporterEntS
248	CSAK557_04793	4012493	4013446	hypothetical protein
249	CSAK557_04794	4013486	4014919	chromate transporter,chromate ion transporter (CHR) family
250	CSAK557_04795	4015148	4015906	hypothetical protein
251	CSAK557_04796	4015960	4016880	hypothetical protein
252	CSAK557_04797	4016944	4017315	hypothetical protein
253	CSAK557_04798	4017456	4018145	hypothetical protein

254	CSAK557_04799	4018091	4018918	hypothetical protein
255	xerC_3	4019092	4019685	Tyrosine recombinaseXerC
256	repB_2	4020065	4020937	RepFIB replication protein A
257	CSAK557_04802	4021402	4021641	ParG
258	CSAK557_04803	4021694	4022311	putative crown gall tumor protein VirC1
259	CSAK557_04804	4022862	4023575	hypothetical protein
260	ybaQ	4023831	4024265	putative HTH-type transcriptional regulator ybaQ
261	CSAK557_04806	4024347	4024820	Type IV secretory pathway, TrbL components
262	CSAK557_04807	4024865	4025581	hypothetical protein
263	rmlA	4025639	4027222	Glucose-1-phosphate thymidylyltransferase
264	rfbC	4027226	4027768	dTDP-4-dehydrorhamnose 3,5-epimerase
265	rfbX	4027833	4029089	Putative O-antigen transporter
266	CSAK557_04811	4029086	4030063	rhamnosyltransferase
267	CSAK557_04812	4030075	4030830	hypothetical protein
268	CSAK557_04813	4031115	4033940	hypothetical protein
269	CSAK557_04814	4034378	4034980	hypothetical protein
270	CSAK557_04815	4035925	4036689	hypothetical protein
271	CSAK557_04816	4036709	4037539	rhamnosyltransferase
272	xerC_4	4037565	4037894	Tyrosine recombinaseXerC
273	umuC_7	4037922	4039196	hypothetical protein
274	CSAK557_04819	4039153	4039482	hypothetical protein
275	CSAK557_04820	4039546	4039932	hypothetical protein
276	CSAK557_04821	4040054	4040779	integrating conjugative element protein, family
277	CSAK557_04822	4040792	4041658	hypothetical protein
278	CSAK557_04823	4041865	4042986	site-specific tyrosine recombinase XerC
279	CSAK557_04824	4043427	4043711	hypothetical protein
280	CSAK557_04825	4043794	4044612	hypothetical protein
281	CSAK557_04826	4044612	4044851	hypothetical protein
282	dam_2	4044752	4045255	DNA adenine methylase
283	CSAK557_04828	4045485	4045718	hypothetical protein
284	CSAK557_04829	4045785	4046186	hypothetical protein
285	CSAK557_04830	4046190	4046606	hypothetical protein

286	CSAK557_04831	4046599	4046790	hypothetical protein
287	CSAK557_04832	4046801	4047304	Phage regulatory protein CII (CP76)
288	CSAK557_04833	4047521	4050640	hypothetical protein
289	CSAK557_04834	4051088	4052917	hypothetical protein
290	CSAK557_04835	4053261	4053884	hypothetical protein
291	CSAK557_04836	4053966	4054274	hypothetical protein
292	CSAK557_04837	4054429	4054794	hypothetical protein
293	CSAK557_04838	4054883	4055299	hypothetical protein
294	CSAK557_04839	4055523	4056113	integrating conjugative element protein, family
295	CSAK557_04840	4056127	4056549	integrating conjugative element protein, family
296	CSAK557_04841	4056546	4057247	integrating conjugative element protein, family
297	CSAK557_04842	4057461	4058120	integrating conjugative element protein, family
298	CSAK557_04843	4058117	4058323	conjugative transferregion protein
299	CSAK557_04844	4058657	4058920	hypothetical protein
300	CSAK557_04845	4058923	4060074	hypothetical protein
301	dnaB_4	4060165	4060692	Replicative DNA helicase
302	cpsD	4060689	4061603	Tyrosine-protein kinase CpsD
303	CSAK557_04848	4061766	4062548	hypothetical protein
304	CSAK557_04849	4063092	4063724	hypothetical protein
305	CSAK557_04850	4063833	4064693	hypothetical protein
306	CSAK557_04851	4064707	4066089	Phage-related lysozyme (muraminidase)
307	CSAK557_04852	4065901	4066371	hypothetical protein
308	CSAK557_04853	4066364	4067500	hypothetical protein
309	gcvA_3	4075036	4075473	Gcv operon activator
310	CSAK557_04862	4075936	4076916	Transposase DDE domain protein
311	CSAK557_04863	4079051	4079614	hypothetical protein
312	CSAK557_04864	4079706	4079870	hypothetical protein
313	topB_4	4080153	4080638	DNA topoisomerase 3
314	CSAK557_04866	4081069	4081329	hypothetical protein
315	CSAK557_04867	4081668	4081814	hypothetical protein
316	CSAK557_04868	4082042	4082710	hypothetical protein
317	CSAK557_04869	4082735	4083085	hypothetical protein

318	CSAK557_04872	4085014	4086318	Site-specific recombinase XerD
319	CSAK557_04873	4086427	4086582	hypothetical protein
320	CSAK557_04874	4086579	4086824	hypothetical protein
321	CSAK557_04875	4087470	4087781	anaerobic benzoate catabolism transcriptional regulator
322	CSAK557_04876	4087990	4088280	DNA-binding transcriptional regulator Nlp
323	CSAK557_04939	4135125	4135781	hypothetical protein
324	CSAK557_04942	4138720	4139952	hypothetical protein
325	CSAK557_04943	4140031	4140261	hypothetical protein
326	CSAK557_04944	4140261	4140857	hypothetical protein
327	CSAK557_04945	4140942	4141412	hypothetical protein
328	CSAK557_04946	4141569	4142447	putative periplasmiciron-binding protein precursor
329	mntB	4142450	4143322	Manganese transport system membrane protein mntB
330	znuC_1	4143319	4144011	Zinc import ATP-binding protein ZnuC
331	CSAK557_04949	4144373	4144618	hypothetical protein
332	CSAK557_04986	4170733	4171539	Phage small terminasesubunit
333	CSAK557_04987	4171532	4171984	Phage head completionprotein (GPL)
334	CSAK557_04988	4171981	4172481	P2 phage tail completion protein R (GpR)
335	CSAK557_04989	4172481	4173188	Phage virion morphogenesis family protein
336	CSAK557_04990	4173185	4173751	hypothetical protein
337	CSAK557_04991	4173822	4174097	hypothetical protein
338	CSAK557_04992	4174265	4174480	hypothetical protein
339	epsE_2	4174846	4176168	Type II traffic warden ATPase
340	CSAK557_04994	4176176	4176673	type IV pilus biogenesis protein PilP
341	CSAK557_04995	4176660	4177940	Pilin accessory protein (PilO)
342	bfpB	4177944	4179617	Bundle-forming pilusB
343	CSAK557_04997	4179637	4179864	PilM
344	CSAK557_04998	4180082	4181437	integrating conjugative element protein PilL, PFGI-1 class
345	dam_3	4181908	4182510	DNA adenine methylase
346	CSAK557_05000	4182507	4182815	hypothetical protein
347	CSAK557_05001	4182792	4183058	hypothetical protein
348	CSAK557_05002	4183285	4186248	putative ATPase
349	CSAK557_05003	4186306	4186974	integrating conjugative element protein, family

350	CSAK557_05004	4186986	4187387	hypothetical protein
351	repA_2	4187921	4188439	Replication-associated protein
352	pifC_2	4189131	4189514	Transcriptional repressor pifC
353	aseR	4189554	4189937	HTH-type transcriptional repressor AseR
354	CSAK557_05008	4190784	4191119	hypothetical protein
355	CSAK557_05009	4191010	4191360	integrating conjugative element membrane protein, family
356	CSAK557_05010	4191373	4191741	integrating conjugative element protein, family
357	CSAK557_05011	4191731	4192150	Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, ATPase PilB
358	CSAK557_05012	4192099	4192326	hypothetical protein
359	csbC	4192360	4193856	putative metabolite transport protein CsbC
360	iolE_2	4194287	4195186	Inosose dehydratase
361	CSAK557_05015	4195216	4196100	Hydroxypyruvate isomerase
362	iolG_1	4196136	4196279	Inositol 2-dehydrogenase/D-chiro-inositol 3-dehydrogenase
363	iolG_2	4196342	4197166	Inositol 2-dehydrogenase/D-chiro-inositol 3-dehydrogenase
364	iolD	4197223	4199163	3D-(3,5/4)-trihydroxycyclohexane-1,2-dione hydrolase
365	iolC_2	4199656	4200831	5-dehydro-2-deoxygluconokinase
366	CSAK557_05020	4201076	4201591	hypothetical protein
367	ybbH_3	4201738	4202592	putative HTH-type transcriptional regulator ybbH
368	iolB_1	4202652	4202801	5-deoxy-glucuronate isomerase
369	lacI_3	4206953	4207369	Lactose operon repressor
370	lacI_4	4207332	4208120	Lactose operon repressor
371	CSAK557_05027	4208590	4208874	Transposase
372	CSAK557_05028	4208922	4209440	hypothetical protein
373	htpX_2	4229851	4230606	hypothetical protein
374	CSAK557_05053	4230609	4230842	hypothetical protein
375	CSAK557_05054	4230820	4231317	putative membrane protein
376	kefB_2	4231320	4232042	NEM-activable K( )/H() antiporter
377	kefC_3	4232045	4233082	K( )/H( ) antiporter
378	trxC_2	4233086	4233526	Thioredoxin-2
379	CSAK557_05058	4233516	4234661	hypothetical protein
380	CSAK557_05059	4234741	4235352	hypothetical protein
381	CSAK557_05060	4235442	4236329	YfdX protein

382	CSAK557_05061	4236432	4236830	YfdX protein
383	CSAK557_05062	4236980	4237393	hypothetical protein
384	hspA_1	4237386	4237631	Spore protein SP21
385	CSAK557_05064	4237826	4238320	ATP-dependent metallopeptidase HflB
386	ftsH_2	4238287	4238841	ATP-dependent zinc metalloprotease FtsH
387	CSAK557_05066	4238902	4239093	hypothetical protein
388	clpC	4239093	4240892	ATP-dependent Clp protease ATP-binding subunit ClpC
389	clpB_1	4240877	4241266	Chaperone protein ClpB
390	CSAK557_05069	4241359	4241985	type VI secretion ATPase, ClpV1 family
391	hspA_2	4242152	4242721	Spore protein SP21
392	CSAK557_05071	4242861	4243037	DNA binding domain, excisionase family
393	CSAK557_05072	4243344	4243856	hypothetical protein
394	CSAK557_05097	4268093	4268773	Acyltransferase family protein
395	CSAK557_05098	4268809	4269639	hypothetical protein
396	CSAK557_05099	4271471	4272112	hypothetical protein
397	CSAK557_05100	4272127	4273242	hypothetical protein
398	CSAK557_05101	4273218	4275215	hypothetical protein
399	CSAK557_05139	4302834	4303052	hypothetical protein
400	CSAK557_05140	4303078	4303461	type I secretion membrane fusion protein, HlyD family
401	apxIB_1	4303471	4304118	RTX-I toxin determinant B
402	apxIB_2	4304073	4305677	RTX-I toxin determinant B
403	CSAK557_05143	4305701	4306621	outer membrane channel protein
404	CSAK557_05144	4306720	4307100	type I secretion outer membrane protein, TolC family
405	CSAK557_05145	4307185	4309131	hypothetical protein
406	CSAK557_05146	4309110	4309436	hypothetical protein
407	CSAK557_05147	4310040	4310600	hypothetical protein
408	CSAK557_05148	4310866	4311693	hypothetical protein
409	CSAK557_05149	4311740	4313281	hypothetical protein
410	CSAK557_05150	4313320	4313697	hypothetical protein
411	CSAK557_05151	4313828	4314067	hypothetical protein
412	CSAK557_05152	4314129	4314662	hypothetical protein
413	CSAK557_05153	4314601	4316088	hypothetical protein

414	CSAK557_05154	4316484	4317164	hypothetical protein
415	CSAK557_05155	4318304	4318492	hypothetical protein
416	CSAK557_05156	4318676	4319083	hypothetical protein
417	CSAK557_05242	4381149	4381523	hypothetical protein
418	dxs_2	4381793	4382158	1-deoxy-D-xylulose-5-phosphate synthase
419	CSAK557_05391	4487245	4488321	hypothetical protein
420	CSAK557_05392	4488276	4488794	hypothetical protein
421	intS_2	4581144	4581905	Putative prophage CPS-53 integrase
422	intA_5	4581953	4582366	Prophage CP4-57 integrase
423	CSAK557_05527	4582577	4582774	hypothetical protein
424	CSAK557_05530	4586884	4587210	Recombinational DNA repair protein (RecE pathway)
425	CSAK557_05534	4587842	4587973	Lambda Phage CIII
426	CSAK557_05535	4588072	4588221	hypothetical protein
427	CSAK557_05537	4590119	4590241	hypothetical protein
428	CSAK557_05538	4590291	4590644	hypothetical protein
429	CSAK557_05540	4592380	4593096	T5orf172 domain protein
430	CSAK557_05541	4594013	4594192	hypothetical protein
431	CSAK557_05546	4601214	4601405	hypothetical protein
432	CSAK557_05547	4602777	4603004	hypothetical protein
433	CSAK557_05548	4604484	4604684	Head fiber protein
434	CSAK557_05549	4604794	4605003	hypothetical protein
435	CSAK557_05550	4605191	4605685	hypothetical protein
436	CSAK557_05551	4608155	4608439	hypothetical protein
437	CSAK557_05552	4608491	4608853	hypothetical protein
438	CSAK557_05553	4610363	4610521	hypothetical protein
439	CSAK557_05554	4610935	4611195	hypothetical protein
440	CSAK557_05555	4611197	4611613	hypothetical protein
441	CSAK557_05556	4611628	4611825	hypothetical protein
442	CSAK557_05557	4611825	4612130	hypothetical protein

Table S3. Variation of five CC4 core genome unique genes in 107 *Cronobacter* genomes

No	Isolate	species	CC4	CSAK557_03870	CSAK557_01621	CSAK557_03180	CSAK557_03949	CSAK557_02008
1	4	*C. sakazakii*	CC4	+	+	+	+	+
2	6	*C. sakazakii*	CC4	+	+	+	+	+
3	20	*C. sakazakii*	CC4	+	+	+	+	+
4	377	*C. sakazakii*	CC4	+	+	+	+	+
5	553	*C. sakazakii*	CC4	+	+	+	+	+
6	557	*C. sakazakii*	CC4	+	+	+	+	+
7	558	*C. sakazakii*	CC4	+	+	+	+	+
8	691	*C. sakazakii*	CC4	+	+	+	+	+
9	692	*C. sakazakii*	CC4	+	+	+	+	+
10	694	*C. sakazakii*	CC4	+	+	+	+	+
11	695	*C. sakazakii*	CC4	+	+	+	+	+
12	698	*C. sakazakii*	CC4	+	+	+	+	+
13	701	*C. sakazakii*	CC4	+	+	+	+	+
14	702	*C. sakazakii*	CC4	+	+	+	+	+
15	705	*C. sakazakii*	CC4	+	+	+	+	+
16	706	*C. sakazakii*	CC4	+	+	+	+	+
17	707	*C. sakazakii*	CC4	+	+	+	+	+
18	709	*C. sakazakii*	CC4	+	+	+	+	+
19	711	*C. sakazakii*	CC4	+	+	+	+	+
20	712	*C. sakazakii*	CC4	+	+	+	+	+
21	721	*C. sakazakii*	CC4	+	+	+	+	+
22	730	*C. sakazakii*	CC4	+	+	+	+	+
23	767	*C. sakazakii*	CC4	+	+	+	+	+
24	1105	*C. sakazakii*	CC4	+	+	+	+	+
25	1219	*C. sakazakii*	CC4	+	+	+	+	+

26	1220	*C. sakazakii*	CC4	+	+	+	+	+
27	1221	*C. sakazakii*	CC4	+	+	+	+	+
28	1225	*C. sakazakii*	CC4	+	+	+	+	+
29	1231	*C. sakazakii*	CC4	+	+	+	+	+
30	1240	*C. sakazakii*	CC4	+	+	+	+	+
31	1533	*C. sakazakii*	CC4	+	+	+	+	+
32	1537	*C. sakazakii*	CC4	+	+	+	+	+
33	1542	*C. sakazakii*	CC4	+	+	+	+	+
34	1587	*C. sakazakii*	CC4	+	+	+	+	+
35	ES713	*C. sakazakii*	CC4	+	+	+	+	+
36	G-2151	*C. sakazakii*	CC4	+	+	+	+	+
37	SP291	*C. sakazakii*	CC4	+	+	+	+	+
38	1	*C. sakazakii*	Non-CC4	-	-	-	-	-
39	5	*C. sakazakii*	Non-CC4	-	-	-	-	-
40	92	*C. turicensis*	Non-CC4	-	+	-	-	+
41	140	*C. sakazakii*	Non-CC4	-	-	-	-	-
42	150	*C. sakazakii*	Non-CC4	-	+	-	+	-
43	507	*C. malonaticus*	Non-CC4	-	-	-	-	+
44	510	*C. malonaticus*	Non-CC4	-	-	-	-	+
45	520	*C. sakazakii*	Non-CC4	-	-	-	-	+
46	530	*C. muytjensii*	Non-CC4	-	-	+	-	+
47	564	*C. turicensis*	Non-CC4	-	-	+	-	+
48	581	*C. universalis*	Non-CC4	-	-	-	+	+
49	582	*C. dublinensis*	Non-CC4	-	+	-	+	+
50	583	*C. dublinensis*	Non-CC4	-	-	-	+	+
51	658	*C. sakazakii*	Non-CC4	-	-	-	-	-
52	680	*C. sakazakii*	Non-CC4	-	-	-	-	-
53	681	*C. malonaticus*	Non-CC4	-	-	-	-	+
54	685	*C. malonaticus*	Non-CC4	+	-	-	+	+
55	687	*C. malonaticus*	Non-CC4	-	-	-	+	+
56	690	*C. sakazakii*	Non-CC4	-	-	-	-	+

57	693	*C. sakazakii*	Non-CC4	-	-	-	-	-
58	696	*C. sakazakii*	Non-CC4	-	-	-	-	+
59	699	*C. sakazakii*	Non-CC4	-	-	-	-	+
60	700	*C. sakazakii*	Non-CC4	-	-	-	-	-
61	703	*C. sakazakii*	Non-CC4	-	-	-	-	+
62	708	*C. sakazakii*	Non-CC4	-	-	-	-	+
63	713	*C. sakazakii*	Non-CC4	-	-	-	-	-
64	714	*C. sakazakii*	Non-CC4	-	-	-	-	-
65	715	*C. sakazakii*	Non-CC4	-	-	-	-	-
66	716	*C. sakazakii*	Non-CC4	-	-	-	-	-
67	978	*C. sakazakii*	Non-CC4	-	+	+	-	-
68	984	*C. sakazakii*	Non-CC4	-	+	+	-	-
69	1210	*C. dublinensis*	Non-CC4	-	-	-	+	+
70	1211	*C. turicensis*	Non-CC4	-	-	-	-	+
71	1218	*C. sakazakii*	Non-CC4	-	-	-	-	-
72	1249	*C. sakazakii*	Non-CC4	-	-	-	+	-
73	1330	*C. condimenti*	Non-CC4	-	-	-	-	+
74	1536	*C. sakazakii*	Non-CC4	-	-	-	-	-
75	1545	*C. malonaticus*	Non-CC4	-	-	-	-	+
76	1553	*C. turicensis*	Non-CC4	-	-	-	-	+
77	1554	*C. turicensis*	Non-CC4	-	+	-	-	+
78	1556	*C. dublinensis*	Non-CC4	-	+	-	+	+
79	1558	*C. malonaticus*	Non-CC4	-	-	-	-	+
80	1560	*C. dublinensis*	Non-CC4	-	+	+	-	+
81	1569	*C. malonaticus*	Non-CC4	-	-	-	-	+
82	1846	*C. malonaticus*	Non-CC4	-	-	-	+	+
83	1880	*C. turicensis*	Non-CC4	-	+	-	-	+
84	2030	*C. dublinensis*	Non-CC4	-	+	-	-	+
85	2045	*C. malonaticus*	Non-CC4	-	-	-	-	+
86	2046	*C. malonaticus*	Non-CC4	-	-	-	-	+
87	2048	*C. sakazakii*	Non-CC4	-	-	-	-	-

88	2051	*C. sakazakii*	Non-CC4	+	+	-	-	-
89	2064	*C. sakazakii*	Non-CC4	-	-	-	-	-
90	2087	*C. sakazakii*	Non-CC4	-	-	-	-	-
91	2089	*C. sakazakii*	Non-CC4	-	-	-	-	-
92	2106	*C. sakazakii*	Non-CC4	-	-	-	-	-
93	2107	*C. sakazakii*	Non-CC4	-	-	-	-	+
94	2109	*C. malonaticus*	Non-CC4	-	-	-	-	+
95	2161	*C. sakazakii*	Non-CC4	-	-	-	-	-
96	ATCC 51329	*C. muytjensii*	Non-CC4	-	-	-	-	+
97	CMCC 45402	*C. malonaticus*	Non-CC4	-	-	-	-	+
98	E764	*C. sakazakii*	Non-CC4	-	-	-	-	+
99	ES15	*C. sakazakii*	Non-CC4	-	-	-	-	-
100	ES35	*C. sakazakii*	Non-CC4	-	-	-	-	-
101	HPB5174	*C. sakazakii*	Non-CC4	-	-	-	-	-
102	LMG 23823	*C. dublinensis*	Non-CC4	-	-	-	+	+
103	LMG 23824	*C. dublinensis*	Non-CC4	-	+	-	+	+
104	LMG 23825	*C. dublinensis*	Non-CC4	-	+	-	-	+
105	LMG23826	*C. malonaticus*	Non-CC4	-	-	-	-	+
106	NBRC 102416T	*C. sakazakii*	Non-CC4	-	-	-	-	-
107	NCTC9529T	*C. universalis*	Non-CC4	-	-	-	+	+

Table S4. Presence of the unique genes identified in *C. sakazakii* CC4 isolate 377 against non-CC4 isolate ES15, in the remaining 17 CC4 and 11 non-CC4 isolates

			Strain	1587	558	6	767	1225	20	721	1221	1220	1219	701	557	1231	1240	553	4	SP291	1	5	680	1218	658	696	520	140	150	1249	978
			ST	109/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST4/CC4	ST15/CC4	ST109/CC4	ST8/CC8	ST8 /CC8	ST8/CC8	ST1/CC1	ST1/CC1	ST12	ST12	ST40/45	ST16/CC16	ST31	ST3/CC3
				Fatal men	Clin	Clin	Fatal men	Fatal men	Clin	CSF	CSF	CSF	Fatal men	Fatal men	Clin	Fatal men	Fatal men	CSF	Clin	Clin	Prod. Site	Clin	Clin	Clin	CSF	CSF	Fatal men	Clin	Clin	Spice	Spice	Fatal men	EFT
No	Locus_tag	Product																														
1	Csak377_00002	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+
2	Csak377_00003	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+
3	Csak377_00004	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+
4	Csak377_00005	vapC toxin		+	-	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	-	-	+
5	Csak377_00006	Virulence associated protein		+	-	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	-	-	+
6	Csak377_00008	Hypothetical protein		+	-	+	+	+	-	-	-	+	+	+	+	-	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	+
7	Csak377_00009	Disulfide isomerase		+	-	+	+	+	-	-	-	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	+
8	Csak377_00010	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
9	Csak377_00011	Hypothetical protein		+	-	+	+	+	-	-	-	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	+	-	+	-	-	-
10	Csak377_00012	Hypothetical protein		+	-	+	+	+	-	-	-	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	+	-	+	-	-	-
11	Csak377_00014	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+
12	Csak377_00015	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+
13	Csak377_00016	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+
14	Csak377_00017	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+
15	Csak377_00018	parD1, antitoxin protein		+	-	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-	-	-	-
16	Csak377_00019	parE4, toxin protein		+	-	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-	-	-	-
17	Csak377_00020	pcoE, putative copper-binding protein precursor		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
18	Csak377_00021	CusS1, Sensornkinase		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-

19	Csak377_00022	copR, Transcriptional activator protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
20	Csak377_00023	YebZ_1, inner membrane protein		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
21	Csak377_00024	pcoC, Copper resistance proteinC precursor		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
22	Csak377_00025	copB, Copper resistance protein B precursor		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
23	Csak377_00026	copA_1, Copper resistance protein A precursor		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
24	Csak377_00027	silE_1, Silver-binding protein silE precursor BLASTx; copper resistant protein PcoE		+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
25	Csak377_00028	lytM_1, Glycyl-glycine endopeptidase precursor		+	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-
26	Csak377_00029	Hypothetical protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-
27	Csak377_00031	Protein of unknown function		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-
28	Csak377_00032	cusA, Cation efflux system protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
29	Csak377_00033	cusB, Cation efflux system protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
30	Csak377_00034	cusF, Cation efflux system protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
31	Csak377_00035	cusC, Cation efflux system protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
32	Csak377_00037	cusS__2, Sensor kinase		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
33	Csak377_00038	silE_2, Silver-binding protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
34	Csak377_00039	Hypothetical protein		+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
35	Csak377_00040	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-
36	Csak377_00041	Putative restriction endonuclease		+	-	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	-	-	+	-	-	-	-	-
37	Csak377_00043	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
38	Csak377_00044	YtaF, putative sporulation protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
39	Csak377_00045	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
40	Csak377_00046	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
41	Csak377_00047	pifC, Transcriptional repressor		+	-	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	-	+	+	-	+	-	-	-
42	Csak377_00048	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
43	Csak377_00049	Hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	+	-	+	+	+	+	+	+	+	-	-	+	-	+	-	-	-
44	Csak377_00051	repB_1, RepFIB replication protein A		+	-	+	-	+	-	-	-	+	+	-	+	-	+	+	+	+	+	+	-	-	-	+	-	+	-	-	+
45	Csak377_00052	ParG, C protein motif		+	-	+	-	+	-	-	-	+	+	-	+	-	+	+	+	+	+	+	-	-	-	-	-	+	-	-	+
46	Csak377_00053	VirC1 protein		+	-	+	-	+	-	-	-	+	+	+	+	-	+	+	+	+	+	+	+	-	-	+	-	+	-	-	+
47	Csak377_00054	Hypothetical protein		+	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-

48	Csak377_00100	mbeC_1, Conjugative accessory protein		+	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-
49	Csak377_00101	mbeC_2, Conjugative accessory protein		+	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-
50	Csak377_00102	mbeA_1, DNA relaxase		+	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-
51	Csak377_00103	Hypothetical protein		-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-
52	Csak377_00104	Hypothetical protein		+	+	-	-	-	+	+	+	+	-	-	-	+	+	-	+	+	+	-	+	-	-	-	-	-	-	-	-
53	Csak377_00105	mbeC_3, Conjugative accessory protein		+	+	-	-	-	+	-	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-
54	Csak377_00116	Hypothetical protein		+	+	-	-	-	+	-	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-
55	Csak377_00117	Transposase		+	+	-	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	-	-	-	-	+	-	-	+
56	Csak377_00118	Hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
57	Csak377_00121	Transposase		-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
58	Csak377_00122	Integrase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
59	Csak377_00123	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
60	Csak377_00124	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
61	Csak377_00125	cobT, Aerobic cobaltochelatase subunit		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
62	Csak377_00126	cobS, Aerobic cobaltochelatase subunit		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
63	Csak377_00127	dnaE_1, DNA polymerase III subunit alpha		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
64	Csak377_00128	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
65	Csak377_00129	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
66	Csak377_00130	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
67	Csak377_00131	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
68	Csak377_00132	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
69	Csak377_00133	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
70	Csak377_00134	polA_1, DNA polymerase thermostable protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
71	Csak377_00135	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
72	Csak377_00136	recA_1, Recombinase A		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
73	Csak377_00137	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
74	Csak377_00138	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
75	Csak377_00139	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
76	Csak377_00140	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
77	Csak377_00141	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

78	Csak377_00142	repB_2, RepFIB replication protein A		+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	+	+	-
79	Csak377_00143	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
80	Csak377_00144	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
81	Csak377_00145	Calcineurin-like phosphoesterase superfamily domain protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
82	Csak377_00146	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
83	Csak377_00147	Putative DNA double-strand break repair Rad50 ATPase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-
84	Csak377_00148	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
85	Csak377_00149	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
86	Csak377_00152	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
87	Csak377_00153	thyA_1, Thymidylate synthase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	+
88	Csak377_00154	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
89	Csak377_00155	dfrA, Dihydrofolate reductase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
90	Csak377_00156	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
91	Csak377_00157	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
92	Csak377_00158	dnaQ_1, DNA polymerase III subunit epsilon		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
93	Csak377_00159	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
94	Csak377_00160	Hypothetical protein		+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
95	Csak377_00161	Hypothetical protein		+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
96	Csak377_00162	phage N-6-adenine-methyltransferase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
97	Csak377_00163	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
98	Csak377_00164	rnhA_1, Ribonuclease HI		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
99	Csak377_00165	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
100	Csak377_00166	Tellurite resistance protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-
101	Csak377_00167	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
102	Csak377_00168	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
103	Csak377_00169	hsdR_1, Type 1 restriction enzyme EcoR124II R protein		+	+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
104	Csak377_00170	Type I restriction enzyme specificity protein MPN_089		+	+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
105	Csak377_00171	Putative type I restriction enzymeP M protein		+	+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
106	Csak377_00172	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
107	Csak377_00173	Hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

108	Csak377_00174	Hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
109	Csak377_00176	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
110	Csak377_00177	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
111	Csak377_00178	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
112	Csak377_00179	Hypothetical protein		-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
113	Csak377_00180	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
114	Csak377_00181	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-
115	Csak377_00182	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
116	Csak377_00183	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
117	Csak377_00184	Hypothetical protein		-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
118	Csak377_00185	Repressor of phase-1 flagellin protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
119	Csak377_00186	Hypothetical protein		-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
120	Csak377_00187	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
121	Csak377_00188	uvrB_1, Excinuclease ABC subunit B,		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
122	Csak377_00189	parB, putativechromosome-partitioning protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
123	Csak377_00190	ParB/RepB/Spo0J family partition protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
124	Csak377_00191	hlyB, Alpha-hemolysin translocation+ATP-binding+protein HlyB		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
125	Csak377_00192	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
126	Csak377_00193	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
127	Csak377_00194	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
128	Csak377_00195	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
129	Csak377_00196	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
130	Csak377_00197	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
131	Csak377_00198	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
132	Csak377_00199	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
133	Csak377_00200	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
134	Csak377_00201	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
135	Csak377_00202	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
136	Csak377_00203	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

137	Csak377_00204	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
138	Csak377_00205	Bacterial Ig-like domain		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
139	Csak377_00206	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
140	Csak377_00207	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
141	Csak377_00208	Phage tape measure domain protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
142	Csak377_00209	Phage related protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
143	Csak377_00210	Phage minor tail protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
144	Csak377_00211	NlpC/P60 family protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
145	Csak377_00212	Phage-related protein tail component		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
146	Csak377_00213	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
147	Csak377_00214	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
148	Csak377_00216	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-
149	Csak377_00217	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
150	Csak377_00218	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
151	Csak377_00219	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
152	Csak377_00220	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
153	Csak377_00221	Flp pilus assembly protein 2C+ATPase+CpaE		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
154	Csak377_00222	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
155	Csak377_00223	Hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
156	Csak377_00224	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
157	Csak377_00225	Replicative DNA helicase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
158	Csak377_00226	DNA primase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
159	Csak377_00227	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
160	Csak377_00228	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
161	Csak377_00229	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
162	Csak377_00230	DNA ligase		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
163	Csak377_00231	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
164	Csak377_00232	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
165	Csak377_00233	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
166	Csak377_00234	Hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

167	Csak377_00236	Hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
168	Csak377_00237	Hypothetical protein		+	+	-	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+
169	Csak377_00238	rop, RNA one modulator		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
170	Csak377_00239	Hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
171	Csak377_00240	Hypothetical protein		-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+	-	-	-	-	-	-
172	Csak377_00241	XerD, Site-specific recombinase		-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-
173	Csak377_00243	Uncharacterized protein conserved in bacteria		+	-	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	+	+	+	-	+	+	-	-
174	Csak377_00325	Integrase		+	+	-	+	-	+	+	+	-	-	+	+	+	-	+	+	+	+	-	-	-	-	+	-	+	-	-	+
175	Csak377_00327	Hypothetical protein		+	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	+	+	-	-	+	+	-	-	-	-	-	+
176	Csak377_00330	yjiR_1, Putative HTH-type+transcriptional regulator yjiR		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
177	Csak377_00332	dapA_1, Dihydrodipicolinate synthase		+	+	-	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	+	+	+	+	+	-	+	+	+	+
178	Csak377_00333	aseR, HTH-type transcriptional repressor AseR		+	+	-	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	+	+	+	+	-	-	+	+	+	+
179	Csak377_00334	arsB, Arsenic efflux pump protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
180	Csak377_00335	arsC, Arsenate reductase		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
181	Csak377_00336	trkG_1, Trk system potassium uptake protein trkG		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
182	Csak377_00337	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	-	-	+	+	+	-	+	+	+	+
183	Csak377_00339	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
184	Csak377_00341	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
185	Csak377_00342	Tfp pilus assembly protein, major pilin PilA		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
186	Csak377_00343	mngR_1, Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
187	Csak377_00344	rspA_1, Starvation-sensing protein rspA		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
188	Csak377_00346	Arylesterase		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
189	Csak377_00347	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
190	Csak377_00348	ACT domain-containing protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
191	Csak377_00349	yqjZ, Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
192	Csak377_00350	Putative acetyltransferase		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
193	Csak377_00351	yycE, Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
194	Csak377_00354	Blue light and temperature-regulated antirepressor YcgF protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+
195	Csak377_00355	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
196	Csak377_00356	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+

197	Csak377_00357	ariR_1, Putative two-component-system connector protein AriR protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
198	Csak377_00359	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
199	Csak377_00361	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
200	Csak377_00362	Cyn operon transcriptional activator		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
201	Csak377_00363	molybdate ABC transporter periplasmic molybdate-binding protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+
202	Csak377_00365	fadR_1, Fatty acid metabolism regulator protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
203	Csak377_00366	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
204	Csak377_00367	yjaB_1, putative N-acetyltransferase YjaB		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	+	-	+	+	+	+
205	Csak377_00368	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
206	Csak377_00369	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
207	Csak377_00370	mexB, Multidrug-efflux+transporter		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
208	Csak377_00371	bepF, Efflux pump periplasmic linker BepF protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
209	Csak377_00374	MltA-interacting protein MipA protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
210	Csak377_00375	Antimicrobial resistance+protein Mig-14		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
211	Csak377_00376	Glyoxalase-like domain protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
212	Csak377_00377	Bifunctional biotin-[acetyl-CoA-carboxylase] synthetase/biotin+operon repressor		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
213	Csak377_00378	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	-	-	+	+
214	Csak377_00379	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
215	Csak377_00380	pla, Coagulase/fibrinolysin precursor		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+
216	Csak377_00381	MdtL, multidrug efflux system protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
217	Csak377_00382	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
218	Csak377_00383	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
219	Csak377_00384	Hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+
220	Csak377_00385	cobC_1, Alpha-ribazole phosphatase		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
221	Csak377_00386	repB_3 , RepFIB replication protein A		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
222	Csak377_00387	hypothetical protein		+	+	-	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	+	+	-	-	+	+	+	+
223	Csak377_00388	parA, Plasmid partition protein A		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
224	Csak377_00389	sopB, Plasmid partition protein B		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
225	Csak377_00390	iutA , Iron-regulated outer membrane proteins		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+

226	Csak377_00391	iucD , L-lysine 6-monooxygenase		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
227	Csak377_00392	iucC, Aerobactin synthase IucC		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
228	Csak377_00393	iucB, N(6)-hydroxylysine O-acetyltransferase		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
229	Csak377_00394	iucA, Aerobactin synthase IucA		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
230	Csak377_00396	viuB, Vibriobactin utilisation protein ViuB		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
231	Csak377_00398	hipA, Serine/threonine-protein kinase HipA		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
232	Csak377_00399	treZ, Malto-oligosyltrehalose trehalohydrolase		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
233	Csak377_00400	treY, Maltooligosyl trehalose synthase		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
234	Csak377_00403	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
235	Csak377_00406	sdaC_1, Serine transporter		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
236	Csak377_00409	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
237	Csak377_00410	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+
238	Csak377_00411	gst, Glutathione S-transferase GST-4.5		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+
239	Csak377_00412	guaD, Guanine deaminase		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
240	Csak377_00413	Metallo-beta-lactamase superfamily protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
241	Csak377_00415	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
242	Csak377_00416	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
243	Csak377_00417	csgA, C signal		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
244	Csak377_00418	ATP-dependent helicase, HepA		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-
245	Csak377_00419	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-
246	Csak377_00420	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-
247	Csak377_00421	xerC_2, Tyrosine recombinase XerC		+	+	-	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
248	Csak377_00423	Corrinoid ABC transporter substrate-binding protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
249	Csak377_00426	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
250	Csak377_00427	c SMI1 / KNR4 family protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
251	Csak377_00428	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
252	Csak377_00429	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
253	Csak377_00430	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
254	Csak377_00432	dsbG, Thiol:disulfide interchange protein DsbG precursor		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
255	Csak377_00433	hypothetical protein		+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+

256	Csak377_00434	hypothetical protein		+	+	-	+	-	+	-	-	-	-	+	-	+	-	+	-	+	+	+	+	-	+	+	-	+	+	+	-
257	Csak377_00470	intA_1, Prophage CP4-57 integrase		-	+	-	+	-	-	-	-	-	-	-	+	+	-	-	-	+	-	-	-	+	+	+	-	-	-	-	-
258	Csak377_00471	Transposase		+	+	+	+	+	+	-	-	+	+	-	+	-	+	-	+	+	+	-	-	-	-	-	-	-	-	-	-
259	Csak377_00472	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
260	Csak377_00473	putative transcriptional regulator		-	+	-	+	-	-	-	-	-	+	-	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
261	Csak377_00474	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
262	Csak377_00475	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
263	Csak377_00476	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
264	Csak377_00477	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
265	Csak377_00478	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
266	Csak377_00479	hypothetical protein		-	+	-	+	-	-	-	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
267	Csak377_00480	Recombination protein F		-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
268	Csak377_00660	celllulose biosynthesis operon protein BcsF/YhjT		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
269	Csak377_00661	cellulose biosynthesis protein BcsE		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
270	Csak377_00662	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+
271	Csak377_00663	cell division protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+
272	Csak377_00665	bcsB, Cellulose synthase regulatory subunit		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
273	Csak377_00666	bcsZ, Endoglucanase precursor		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
274	Csak377_00821	Microvirus H protein (pilot protein)		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
275	Csak377_00822	Bacteriophage replication gene A protein (GPA)		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
276	Csak377_00823	Bacteriophage scaffolding protein D		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
277	Csak377_00824	Capsid protein (F protein)		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
278	Csak377_00825	Major spike protein (G protein)		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
279	Csak377_00910	hypothetical protein		+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	-	-	+	+	-
280	Csak377_00911	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	+	+	-	+	-	+	-	-	-	+	+	-	-	-	+	-	-
281	Csak377_00912	hypothetical protein		+	-	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	+	+	-	-	-	+	-	-
282	Csak377_00913	hypothetical protein, dltB		+	-	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	+	+	-	-	-	+	-	-
283	Csak377_00917	hypothetical protein		-	-	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	-	-
284	Csak377_00918	hypothetical protein		-	-	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	-	-
285	Csak377_00977	Major MR/P fimbria protein precursor, mrpA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	-

286	Csak377_00978	Putative minor fimbrial subunit StfF		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-	+	+	-	-	+	-
287	Csak377_00979	Putative fimbrial subunit SteE		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	-
288	Csak377_00980	PAP fimbrial minor pilin protein precursor, papH_1		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-
289	Csak377_00981	Chaperone protein papD precursor, papD		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-
290	Csak377_00982	Outer membrane usher protein papC precursor, papC_1		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-
291	Csak377_00983	PAP fimbrial minor pilin protein precursor, papH_2		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-
292	Csak377_00984	Fimbria A protein precursor, smfA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-
293	Csak377_01070	Antitermination protein		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	+	-	+	-	-	-	+	-	-	-	-
294	Csak377_01072	KilA-N domain protein		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
295	Csak377_01076	hypothetical protein		+	-	+	-	+	-	+	+	-	-	-	-	+	+	-	+	+	+	+	+	+	+	-	+	+	-	+	+
296	Csak377_01081	hypothetical protein		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-	+	+	-	-	-	-	-	-
297	Csak377_01082	hypothetical protein		-	-	-	-	+	-	-	-	-	-	-	_	+	-	-	-	-	+	+	-	+	+	-	-	-	-	-	-
298	Csak377_01085	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	-	-	+
299	Csak377_01086	hypothetical protein		-	-	+	-	+	-	+	+	-	-	-	-	+	+	-	+	-	+	+	-	+	+	-	-	+	-	-	+
300	Csak377_01089	hypothetical protein		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
301	Csak377_01090	hypothetical protein		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
302	Csak377_01091	Anaerobic benzoate catabolism transcriptional regulator		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
303	Csak377_01212	Transposase		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	-
304	Csak377_01213	Lactose operon repressor, lacI_1		+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-
305	Csak377_01216	IS1 transposase		+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-	-	-	-
306	Csak377_01217	Transposase		+	+	+	+	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-
307	Csak377_01218	Putative transcriptional regulator		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-
308	Csak377_01219	Spore protein SP21, hspA		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-
309	Csak377_01221	hypothetical protein		+	+	+	+	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-
310	Csak377_01222	Putative membrane protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-
311	Csak377_01223	hypothetical protein, htpX_1		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-
312	Csak377_01224	hypothetical protein		+	-	-	+	-	-	-	-	-	-	+	-	+	-	+	-	+	+	-	-	-	-	+	-	-	-	-	-
313	Csak377_01225	Integrase core domain protein		+	-	-	+	-	-	-	-	-	-	+	-	+	-	+	-	+	+	-	-	-	-	+	-	-	-	-	-
314	Csak377_01316	SMI1 / KNR4 family protein		-	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	-	-	-	-	+	-	-	-	-	-
315	Csak377_01318	SMI1 / KNR4 family protein		+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	-	-	-	+	+	-	+

316	Csak377_01319	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	-	-	-	+	+	-	+
317	Csak377_01320	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-	+	-	+
318	Csak377_01324	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	+	-	-	-	-	-	-	+	-
319	Csak377_01330	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
320	Csak377_01331	hypothetical protein		+	+	+	+	+	+	-	-	+	+	-	+	+	+	+	+	+	-	-	-	-	-	-	-	+	-	+	-
321	Csak377_01332	hypothetical protein		+	+	+	+	+	+	-	-	+	+	-	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	+	-
322	Csak377_01333	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	-
323	Csak377_01334	hypothetical protein		+	+	+	+	+	+	-	-	+	+	-	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	-	-
324	Csak377_01335	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
325	Csak377_01337	Ribulose-5-phosphate 4-epimerase and aldolases		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
326	Csak377_01338	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+
327	Csak377_01339	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+
328	Csak377_01352	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	+	+
329	Csak377_01353	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	+	+
330	Csak377_01357	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	-	-	-	-	-
331	Csak377_01397	Prophage CP4-57 integrase, intA_2		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	+	-
332	Csak377_01399	hypothetical protein		+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
333	Csak377_01400	hypothetical protein		+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	+	-
334	Csak377_01401	hypothetical protein		+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	+	-
335	Csak377_01402	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	+	-	+	+	-
336	Csak377_01403	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	+	+	-
337	Csak377_01574	hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
338	Csak377_01575	hypothetical protein		+	+	+	-	+	-	-	-	+	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-
339	Csak377_01576	hypothetical protein		+	+	+	-	+	-	-	-	+	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-
340	Csak377_01577	hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
341	Csak377_01578	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
342	Csak377_01579	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
343	Csak377_01580	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
344	Csak377_01581	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
345	Csak377_01582	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

346	Csak377_01583	hypothetical protein		+	-	+	-	+	-	-	-	+	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
347	Csak377_01584	hypothetical protein		+	+	+	-	+	-	-	-	+	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
348	Csak377_01585	hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
349	Csak377_01586	hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
350	Csak377_01587	Site-specific recombinase XerD		+	+	+	+	+	-	-	-	+	+	+	-	+	+	+	+	+	-	-	-	-	-	+	-	-	-	-	-	-
351	Csak377_01727	ATP-dependent helicase HepA		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	+	+	+	+	-	+	-	-	-
352	Csak377_01728	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	+	+	+	+	-	+	-	-	-
353	Csak377_01729	Type-1 restriction enzyme R protein, hsdR_2		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	+	+	+	-	-	-	-	-	-
354	Csak377_01730	Putative toxin YpjF, ypjF		+	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-
355	Csak377_01731	Antitoxin YeeU		-	+	+	-	+	-	-	-	-	-	-	+	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-
356	Csak377_01732	hypothetical protein		-	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-
357	Csak377_01733	hypothetical protein, ykfG_1		+	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
358	Csak377_01734	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-
359	Csak377_01735	hypothetical protein		-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
360	Csak377_01736	hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
361	Csak377_01737	Retron-type reverse transcriptase		+	+	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	-	-
362	Csak377_01740	hypothetical protein		-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
363	Csak377_01909	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
364	Csak377_02398	hypothetical protein		+	-	-	-	-	+	+	-	-	-	-	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+
365	Csak377_02399	hypothetical protein		+	-	-	-	-	+	+	-	-	-	-	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+
366	Csak377_02400	hypothetical protein		+	-	-	-	-	+	+	-	-	-	-	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+
367	Csak377_02402	Leucine efflux protein, leuE		+	-	-	+	-	+	-	-	-	-	-	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+
368	Csak377_02403	hypothetical protein		+	+	-	-	-	+	-	-	-	-	-	+	-	-	+	-	+	+	-	+	-	-	-	-	-	+	-	-	+
369	Csak377_02405	Site-specific recombinase XerD		+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	-	-	-	-	-	-	+
370	Csak377_02475	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+
371	Csak377_02476	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+
372	Csak377_02635	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-	-	-	-	-
373	Csak377_02652	hypothetical protein		+	+	-	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
374	Csak377_02653	hypothetical protein		+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
375	CSAK377_02657	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-

376	Csak377_02668	hypothetical protein, fimH		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	+	-
377	Csak377_02976	Phenylacetate-CoA ligase		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+
378	Csak377_02977	Phenyloxazoline synthase MbtB, mbtB		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
379	Csak377_02978	1-deoxy-D-xylulose-5-phosphate synthase, dxs_2,		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	+	+
380	Csak377_02979	Transketolase 2, tktB_1		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	+	+	+	+	+
381	Csak377_02980	3-oxoacyl-[acyl-carrier-protein] reductase FabG, fabG_5		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	+	+
382	Csak377_02982	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	+	+	+	+
383	Csak377_03006	hypothetical protein		+	-	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
384	Csak377_03007	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
385	Csak377_03008	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
386	Csak377_03009	Retron-type reverse transcriptase		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
387	Csak377_03011	hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
388	Csak377_03205	Acyltransferase family protein		-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+
389	Csak377_03206	Toxin YkfI, ykfI		+	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-
390	Csak377_03207	Putative antitoxin YfjZ, yfjZ		+	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-
391	Csak377_03208	hypothetical protein		+	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-
392	Csak377_03209	hypothetical protein, ykfG_2		-	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-
393	Csak377_03210	Antirestriction protein		-	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-
394	Csak377_03211	hypothetical protein, ykfF		-	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-
395	Csak377_03212	hypothetical protein		+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	+	-	-	+	-	-	-	-	-
396	Csak377_03213	hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	+	+	-	+	-	-	+	-	-	-	-	-
397	Csak377_03214	hypothetical protein		-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	+	+	-	-	-	-	+	-	-	-	-	-
398	Csak377_03215	hypothetical protein		-	+	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-
399	Csak377_03216	GTPase Era, era_2		-	+	+	-	+	+	-	-	+	+	-	+	+	+	-	+	+	+	-	+	-	-	+	-	-	-	-	-
400	Csak377_03217	hypothetical protein		-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	+	-	+	-	-	+	-	-	-	-	-
401	Csak377_03218	hypothetical protein		-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	+	-	+	-	-	+	-	-	-	-	-
402	Csak377_03412	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	-	+	-
403	Csak377_03413	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-
404	Csak377_03414	Phage-related lysozyme (muraminidase)		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-
405	Csak377_03415	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-

406	Csak377_03416	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-
407	Csak377_03417	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-
408	Csak377_03418	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-
409	Csak377_03419	Putative HTH-type transcriptional regulator yhjB, yhjB		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	-	+	-
410	Csak377_03428	hypothetical protein		+	+	+	-	+	+	+	+	+	+	-	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	+	-
411	Csak377_03429	hypothetical protein		+	+	-	-	+	+	+	+	+	+	-	+	+	+	-	+	+	+	-	-	+	+	+	+	-	+	+	+
412	Csak377_03431	ATP-dependent helicase		+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	+	+	-	-	-	+	-
413	Csak377_03432	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	+	-	+	-
414	Csak377_03526	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
415	Csak377_03841	dTDP-4-dehydrorhamnose reductase, rfbD		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
416	Csak377_03843	dTDP-4-dehydrorhamnose 3,5-epimerase, rfbC		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	+	-	-	+	+
417	Csak377_03844	Putative O-antigen transporter, rfbX		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
418	Csak377_03845	Rhamnosyltransferase		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
419	Csak377_03846	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
420	Csak377_03847	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
421	Csak377_03848	rhamnosyltransferase		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
422	Csak377_03849	Glycosyl transferases group 1		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
423	Csak377_03850	dTDP-Rha:alpha-D-GlcNAc-pyrophosphate polyprenol, alpha-3-L-rhamnosyltransferase, wbbL		+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+
424	Csak377_03917	Site-specific tyrosine recombinase XerC		+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-
425	Csak377_03918	hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-
426	Csak377_03919	Putative HTH-type transcriptional regulator		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	+
427	Csak377_03920	Transcriptional repressor DicA		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
428	Csak377_03921	hypothetical protein		+	+	+	-	-	-	+	+	-	-	-	-	-	+	-	+	+	-	-	-	-	-	-	-	+	+	+	+
429	Csak377_03927	hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	+
430	Csak377_03932	hypothetical protein		+	+	+	-	-	-	+	+	-	-	-	-	-	+	-	+	+	+	-	-	-	-	-	-	-	+	-	-
431	Csak377_03943	hypothetical protein		+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
432	Csak377_03956	hypothetical protein		+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	+	-	-
433	Csak377_03962	hypothetical protein		+	-	+	-	-	+	+	+	-	-	-	-	-	+	-	+	+	-	-	-	+	+	-	-	-	-	+	-
434	Csak377_03963	ariR_2putative two-component-system connector protein AriR		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-

435	Csak377_03964	Phage terminase, small subunit		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-
436	Csak377_03965	Phage terminase-like protein, large subunit		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-
437	Csak377_03966	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-
438	Csak377_03967	phage portal protein, HK97 family		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-
439	Csak377_03968	phage prohead protease, HK97 family		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
440	Csak377_03969	putative phage phi-C31 gp36 major capsid-like protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
441	Csak377_03970	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
442	Csak377_03971	Bacteriophage head-tail adaptor		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
443	Csak377_03972	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
444	Csak377_03973	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
445	Csak377_03974	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
446	Csak377_03975	Mu-like prophage tail sheath protein gpL		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
447	Csak377_03976	Phage tail tube protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
448	Csak377_03977	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
449	Csak377_03978	phage tail tape measure protein, TP901 family, core region		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	+	-	-	+	+	-
450	Csak377_03979	Mu-like prophage DNA circulation protein		+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	-	-	-	-	-	+	+	-	+
451	Csak377_03980	Mu-like prophage tail protein gpP		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
452	Csak377_03981	Mu-like prophage protein gp45		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
453	Csak377_03982	Phage protein GP46		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
454	Csak377_03983	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
455	Csak377_03984	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-
456	Csak377_03985	hypothetical protein		+	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	+	-	+	-	-	+	-
457	Csak377_03986	hypothetical protein		+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-
458	Csak377_03987	hypothetical protein		+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	-	-	-	-	+	-
459	Csak377_03988	hypothetical protein		+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	-	-	-	+	+	-
460	Csak377_04031	Transposase		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+
461	Csak377_04039	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	+	-	+
462	Csak377_04040	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	+	+	+	+
463	Csak377_04047	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	+	+
464	Csak377_04048	hypothetical protein		+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	+	+

465	Csak377_04220	hypothetical protein		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	-	-	-	+	+	+

Table S5. Investigation of the hypothetical proteins identified by Gegenees as unique in 18 *C. sakazakii* CC4 genomes (against 12 non-CC4 genomes) and missing in CC4, using *Cronobacter* BLAST

Isolate	Species		CSSP291_10445	CSSP291_18210	ESA_02200
4	*C. sakazakii*	CC4	+	+	-
6	*C. sakazakii*	CC4	+	+	-
20	*C. sakazakii*	CC4	+	+	-
377	*C. sakazakii*	CC4	+	+	-
553	*C. sakazakii*	CC4	+	+	-
557	*C. sakazakii*	CC4	+	+	-
558	*C. sakazakii*	CC4	+	+	-
691	*C. sakazakii*	CC4	+	+	-
692	*C. sakazakii*	CC4	+	+	-
694	*C. sakazakii*	CC4	+	+	-
695	*C. sakazakii*	CC4	+	+	-
698	*C. sakazakii*	CC4	+	+	-
701	*C. sakazakii*	CC4	+	+	-
702	*C. sakazakii*	CC4	+	+	-
705	*C. sakazakii*	CC4	+	+	-
706	*C. sakazakii*	CC4	+	+	-
707	*C. sakazakii*	CC4	+	+	-
709	*C. sakazakii*	CC4	+	+	-
711	*C. sakazakii*	CC4	+	+	-
712	*C. sakazakii*	CC4	+	+	-
721	*C. sakazakii*	CC4	+	+	-
730	*C. sakazakii*	CC4	+	+	-
767	*C. sakazakii*	CC4	+	+	-
1105	*C. sakazakii*	CC4	+	+	-
1219	*C. sakazakii*	CC4	+	+	-
1220	*C. sakazakii*	CC4	+	+	-
1221	*C. sakazakii*	CC4	+	+	-
1225	*C. sakazakii*	CC4	+	+	-
1231	*C. sakazakii*	CC4	+	+	-
1240	*C. sakazakii*	CC4	+	+	-
1533	*C. sakazakii*	CC4	+	+	-
1537	*C. sakazakii*	CC4	+	+	-
1542	*C. sakazakii*	CC4	+	+	-

1587	*C. sakazakii*	CC4	+	+	-
ES713	*C. sakazakii*	CC4	+	+	+
G-2151	*C. sakazakii*	CC4	+	+	-
SP291	*C. sakazakii*	CC4	+	+	-
1	*C. sakazakii*	Non-CC4	-	-	+
5	*C. sakazakii*	Non-CC4	-	-	+
92	*C. turicensis*	Non-CC4	-	-	-
140	*C. sakazakii*	Non-CC4	-	-	+
150	*C. sakazakii*	Non-CC4	-	-	+
507	*C. malonaticus*	Non-CC4	-	-	+
510	*C. malonaticus*	Non-CC4	-	-	+
520	*C. sakazakii*	Non-CC4	-	-	+
530	*C. muytjensii*	Non-CC4	-	-	-
564	*C. turicensis*	Non-CC4	-	-	-
581	*C. universalis*	Non-CC4	-	-	+
582	*C. dublinensis*	Non-CC4	-	-	-
583	*C. dublinensis*	Non-CC4	-	-	-
658	*C. sakazakii*	Non-CC4	-	-	+
680	*C. sakazakii*	Non-CC4	-	-	+
681	*C. malonaticus*	Non-CC4	-	-	+
685	*C. malonaticus*	Non-CC4	-	+	+
687	*C. malonaticus*	Non-CC4	-	-	+
690	*C. sakazakii*	Non-CC4	-	-	+
693	*C. sakazakii*	Non-CC4	+	-	+
696	*C. sakazakii*	Non-CC4	-	-	+
699	*C. sakazakii*	Non-CC4	-	-	+
700	*C. sakazakii*	Non-CC4	+	-	+
703	*C. sakazakii*	Non-CC4	-	-	+
708	*C. sakazakii*	Non-CC4	-	-	+
713	*C. sakazakii*	Non-CC4	+	-	+
714	*C. sakazakii*	Non-CC4	+	-	+
715	*C. sakazakii*	Non-CC4	+	-	+
716	*C. sakazakii*	Non-CC4	-	-	+
978	*C. sakazakii*	Non-CC4	-	-	+
984	*C. sakazakii*	Non-CC4	-	-	+
1210	*C. dublinensis*	Non-CC4	-	-	-

1211	*C. turicensis*	Non-CC4	-	-	-
1218	*C. sakazakii*	Non-CC4	-	-	+
1249	*C. sakazakii*	Non-CC4	-	-	+
1330	*C. condimenti*	Non-CC4	-	-	-
1536	*C. sakazakii*	Non-CC4	-	-	+
1545	*C. malonaticus*	Non-CC4	-	-	+
1553	*C. turicensis*	Non-CC4	-	-	-
1554	*C. turicensis*	Non-CC4	-	-	-
1556	*C. dublinensis*	Non-CC4	-	-	-
1558	*C. malonaticus*	Non-CC4	-	-	+
1560	*C. dublinensis*	Non-CC4	-	-	-
1569	*C. malonaticus*	Non-CC4	-	-	+
1846	*C. malonaticus*	Non-CC4	-	-	+
1880	*C. turicensis*	Non-CC4	-	-	-
2030	*C. dublinensis*	Non-CC4	-	-	-
2045	*C. malonaticus*	Non-CC4	-	-	+
2046	*C. malonaticus*	Non-CC4	-	-	+
2048	*C. sakazakii*	Non-CC4	-	-	+
2051	*C. sakazakii*	Non-CC4	-	+	+
2064	*C. sakazakii*	Non-CC4	-	-	+
2087	*C. sakazakii*	Non-CC4	-	-	+
2089	*C. sakazakii*	Non-CC4	-	-	+
2106	*C. sakazakii*	Non-CC4	-	-	+
2107	*C. sakazakii*	Non-CC4	-	-	+
2109	*C. malonaticus*	Non-CC4	-	-	+
2161	*C. sakazakii*	Non-CC4	-	-	+
ATCC 51329	*C. muytjensii*	Non-CC4	-	-	-
CMCC 45402	*C. malonaticus*	Non-CC4	-	-	+
E764	*C. sakazakii*	Non-CC4	-	-	+
ES15	*C. sakazakii*	Non-CC4	-	-	+
ES35	*C. sakazakii*	Non-CC4	-	-	+
HPB5174	*C. sakazakii*	Non-CC4	-	-	+
LMG 23823	*C. dublinensis*	Non-CC4	-	-	-
LMG 23824	*C. dublinensis*	Non-CC4	-	-	-
LMG 23825	*C. dublinensis*	Non-CC4	-	-	-
LMG23826	*C. malonaticus*	Non-CC4	-	-	+

NBRC 102416T	*C. sakazakii*	Non-CC4	-	-	+
NCTC9529T	*C. universalis*	Non-CC4	-	-	+

Table S6. SNP calls in the earliest *C. sakazakii* CC4 isolate 377 with SP291 as the reference strain

POS	REF	ref codon	amino acid	ALT	alt codon	amino acid	Nature	Gene	Annotation
351213	T	TAT	Tyrosine	C	TAC	Tyrosine	Synonymous	Intergenic	
351220	C	TTC	Phenylalanine	T	TTT	Phenylalanine	Synonymous	Intergenic	
351246	C	TGC	Cysteine	T	TGT	Cysteine	Synonymous	CSSP291_t20316	
351250	A	GGA	Glycine	G	GGG	Glycine	Synonymous	CSSP291_t20316	
351251	C	GAC	Aspartic acid	T	GAT	Aspartic acid	Synonymous	CSSP291_t20316	
455855	A	GGA	Glycine	C	GGC	Glycine	Synonymous	CSSP291_02330	Putative integrase
455858	G	TCG	Serine	A	TCA	Serine	Synonymous	CSSP291_02330	Putative integrase
455879	C	ATC	Isoleucine	T	ATT	Isoleucine	Synonymous	CSSP291_02330	Putative integrase
455886	C	CCC	Proline	G	CCG	Proline	Synonymous	CSSP291_02330	Putative integrase
455957	G	TCG	Serine	A	TCA	Serine	Synonymous	CSSP291_02330	Putative integrase
455963	G	AGG	Arginine	A	AGA	Arginine	Synonymous	CSSP291_02330	Putative integrase
455980	G	TAG	Stop codon	A	TAA	Stop codon	Synonymous	CSSP291_02330	Putative integrase
455996	T	GTT	Valine	C	GTC	Valine	Synonymous	CSSP291_02330	Putative integrase
456389	G	AAG	Lysine	T	AAT	Asparagine or Aspartic acid	Non-synonymous	CSSP291_02330	Putative integrase
456419	C	ATC	Isoleucine	T	ATT	Isoleucine	Synonymous	Intergenic	
2499701	T	CTT	Leucine	A	CTA	Leucine	Synonymous	Intergenic	

Table S7. Comparative genomics of high CC4 isolate 767 and low invasive CC4 isolate 6

Locus_tag	Unique genes in *C. sakazakii* 767		Locus_tag	Unique genes in *C. sakazakii* 6
CSAK767_03653	Glyoxalase-like domain protein			
iucA_2	Aerobatin synthase IuA		**Gene/Locus Tag**	**Product**
iucA_1	Aerobatin synthase IuA		kduI	4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase
ahpF_1	Alkyl hydroperoxide redutase subunit F		CSAK6_02171	anaerobic benzoate catabolism transcriptional regulator
kgtP_2	Alpha-ketoglutarate permease		CSAK6_03882	Antirestriction protein
cobC_2	Alpha-ribazole phosphatase		CSAK6_03709	antitoxin HipB
CSAK767_03680	Antibioti biosynthesis monooxygenase		parD1	Antitoxin ParD1
CSAK767_03684	Arylesterase		yeeU	Antitoxin YeeU
CSAK767_03682	AT domain-ontaining protein		arsC_1	Arsenate reductase
CSAK767_03620	Autotransporter protein or domain, integral membrane beta-barrel involved in protein seretion		arsC_2	Arsenate reductase
cph2_6	Bateriophytohrome ph2		arsB_1	Arsenic efflux pump protein
CSAK767_03652	bifuntional biotin--[aetyl-oA-arboxylase] synthetase/biotin operon repressor		arsA_1	Arsenical pump-driving ATPase
ycgF_1	Blue light- and temperature-regulated antirepressor YgF		arsA_2	Arsenical pump-driving ATPase
CSAK767_03669	division inhibitor protein		arsR	Arsenical resistance operon repressor
dmlR_10	D-malate degradation protein R		arsD_1	Arsenical resistance operon trans-acting repressor ArsD
bepF	Efflux pump periplasmi linker BepF		arsD_2	Arsenical resistance operon trans-acting repressor ArsD
gst	Glutathione S-transferase GST-.		clpP1	ATP-dependent Clp protease proteolytic subunit 1
CSAK767_04220	Host-nulease inhibitor protein Gam		pcrA	ATP-dependent DNA helicase pcrA
ycgE_3	HTH-type transriptional repressor YgE		CSAK6_03864	ATP-dependent metallopeptidase HflB
iutA_2	Iron-regulated outer membrane proteins		bfpB	Bundle-forming pilus B
iutA_3	loain reeptor		fdhA_2	Glutathione-independent formaldehyde dehydrogenase
CSAK767_03608	Metallo-beta-latamase superfamily protein		ydhP_2	Inner membrane transport protein ydhP

map_2	Methionine aminopeptidase		CSAK6_03873	IS2 transposase Tnp
CSAK767_03666	molybdate AB transporter, periplasmi molybdate-binding protein		CSAK6_04024	putative P-loop ATPase
CSAK767_03648	multidrug efflux system protein MdtL		CSAK6_02025	site-specific tyrosine recombinase XerC
mdtH_2	Multidrug resistane protein MdtH		CSAK6_01955	Transposase DDE domain protein
iucB	N()-hydroxylysine O-aetyltransferase		CSAK6_01248	chaperone protein DnaJ
pla	oagulase/fibrinolysin preursor		CSAK6_00941	chromate transporter, chromate ion transporter (CHR) family
CSAK767_03552	conjugal transfer mating pair stabilization protein TraN		CSAK6_03644	chromate transporter, chromate ion transporter (CHR) family
CSAK767_03580	plasmid partitioning protein		CSAK6_03567	conjugal transfer protein TraR
CSAK767_03582	plasmid SOS inhibition protein A		CSAK6_03610	conjugative transfer region lipoprotein
yjiR_2	putative HTH-type transriptional regulator yjiR		CSAK6_03605	conjugative transfer region protein
CSAK767_04120	putative membrane protein		copA_1	Copper resistance protein A precursor
csgA	signal		copB	Copper resistance protein B precursor
baeS_1	Signal transdution histidine-protein kinase BaeS		pcoC	Copper resistance protein C precursor
rspA_3	Starvation-sensing protein rspA		csoR	Copper-sensitive operon repressor
trkG_2	Trk system potassium uptake protein trkG		rusA_1	Crossover junction endodeoxyribonuclease rusA
xerC_2	Tyrosine reombinase Xer		dmlR_12	D-malate degradation protein R
iucC	Aerobatin synthase Iu		CSAK6_03776	DNA methylase
phoP_2	Alkaline phosphatase synthesis transriptional regulatory protein phoP		traC	DNA primase TraC
CSAK767_03700	arboxylate/amino aid/amine transporter		topB_2	DNA topoisomerase 3
CSAK767_03614	ATP-dependent lp protease ATP-binding subunit lpA		CSAK6_02170	DNA-binding transcriptional regulator Nlp
CSAK767_04202	Bateriophage lysis protein		CSAK6_03646	enterobactin exporter EntS
fadR_2	Fatty aid metabolism regulator protein		fbpC2	Fe(3) ions import ATP-binding protein FbpC 2
guaD	Guanine deaminase		CSAK6_00928	fec operon regulator FecR
aseR	HTH-type transriptional repressor AseR		CSAK6_03626	FRG domain protein
CSAK767_04130	III protein family protein		era_2	GTPase Era

CSAK767_03571	Ketosteroid isomerase-related protein		ssb_3	Helix-destabilizing protein
iucD	L-lysine -monooxygenase		CSAK6_03982	hemin importer ATP-binding subunit
CSAK767_03629	Major Failitator Superfamily protein		CSAK6_01257	HNH endonuclease
treY	Maltooligosyl trehalose synthase		aseR	HTH-type transcriptional repressor AseR
treZ_1	Malto-oligosyltrehalose trehalohydrolase		CSAK6_01998	**hypothetical protein**
CSAK767_03655	MltA-interating protein MipA		CSAK6_02022	**hypothetical protein**
mexB	Multidrug-efflux transporter MexB		CSAK6_02023	**hypothetical protein**
CSAK767_04181	Nlp/P family protein		CSAK6_02024	**hypothetical protein**
CSAK767_03539	conjugal transfer pilus assembly protein TraW		CSAK6_02035	**hypothetical protein**
CSAK767_03540	conjugal transfer protein TrbI		CSAK6_02036	**hypothetical protein**
CSAK767_03597	orrinoid AB transporter substrate-binding protein		CSAK6_02040	**hypothetical protein**
parA	Plasmid partition protein A		CSAK6_03927	**hypothetical protein**
ybcO	prophage-derived unharaterized protein ybO		CSAK6_03948	**hypothetical protein**
CSAK767_04225	reombination protein F		CSAK6_03950	**hypothetical protein**
repB	RepFIB repliation protein A		CSAK6_03954	**hypothetical protein**
rusA_2	rossover juntion endodeoxyribonulease rusA		CSAK6_04023	**hypothetical protein**
tsr_8	Serine hemoreeptor protein		CSAK6_00129	**hypothetical protein**
sdaC_2	Serine transporter		ykfG_1	**hypothetical protein**
CSAK767_03593	SMI / KNR family protein		CSAK6_00131	**hypothetical protein**
CSAK767_04200	Terminase small subunit		CSAK6_00132	**hypothetical protein**
dsbG	Thiol:disulfide interhange protein DsbG preursor		CSAK6_00133	**hypothetical protein**
btuC_2	Vitamin B import system permease protein Btu		CSAK6_00134	**hypothetical protein**
cpdA_2	3',-yli adenosine monophosphate phosphodiesterase pdA		CSAK6_00136	**hypothetical protein**
CSAK767_01960	AIPR protein		CSAK6_00137	**hypothetical protein**
CSAK767_02305	APdomain protein		CSAK6_00346	**hypothetical protein**
arsC	Arsenate redutase		CSAK6_00348	**hypothetical protein**
CSAK767_04192	Baterial Ig-like domain (group 2)		CSAK6_00351	**hypothetical protein**
CSAK767_01936	Bateriophage I repressor helix-turn-helix domain protein		CSAK6_00943	**hypothetical protein**

dapA_2	Dihydrodipiolinate synthase		CSAK6_01066	**hypothetical protein**
yhdJ	DNA adenine methyltransferase YhdJ		CSAK6_01067	**hypothetical protein**
bglF_4	EIIBA-Bgl		CSAK6_01068	**hypothetical protein**
glgX_2	Glyogen debranhing enzyme		CSAK6_01070	**hypothetical protein**
mgtB	Magnesium-transporting ATPase, P-type		CSAK6_01071	**hypothetical protein**
treZ_2	Malto-oligosyltrehalose trehalohydrolase		CSAK6_01072	**hypothetical protein**
nodD2_5	Nodulation protein D 2		CSAK6_01073	**hypothetical protein**
CSAK767_03546	cconjugal transfer pilus assembly protein TraE		CSAK6_01074	**hypothetical protein**
CSAK767_03547	cconjugal transfer pilus assembly protein TraL		CSAK6_01178	**hypothetical protein**
CSAK767_03534	cconjugal transfer pilus assembly protein TraU		CSAK6_01235	**hypothetical protein**
CSAK767_03533	cconjugal transfer pilus assembly protein Trb		CSAK6_01236	**hypothetical protein**
CSAK767_03545	cconjugal transfer protein TraK		CSAK6_01238	**hypothetical protein**
sopB	Plasmid partition protein B		CSAK6_01239	**hypothetical protein**
yjaB_2	putative N-aetyltransferase YjaB		CSAK6_01240	**hypothetical protein**
CSAK767_03246	reombination protein F		CSAK6_01241	**hypothetical protein**
CSAK767_04214	Repliation protein P		CSAK6_01242	**hypothetical protein**
repA	Repliation-assoiated protein		CSAK6_01247	**hypothetical protein**
trg_9	Ribose and galatose hemoreeptor protein		CSAK6_01249	**hypothetical protein**
hipA	Serine/threonine-protein kinase HipA		CSAK6_01250	**hypothetical protein**
CSAK767_01932	site-speifi tyrosine reombinase Xer		CSAK6_01251	**hypothetical protein**
slt_1	Soluble lyti murein transglyosylase preursor		CSAK6_01253	**hypothetical protein**
CSAK767_02302	Terminase small subunit		CSAK6_01254	**hypothetical protein**
CSAK767_03688	Tfp pilus assembly protein, major pilin PilA		CSAK6_01255	**hypothetical protein**
CSAK767_02301	Transposase		CSAK6_01256	**hypothetical protein**
CSAK767_03869	Transposase		CSAK6_01258	**hypothetical protein**
CSAK767_03551	Ureidoglyolate lyase		CSAK6_01264	**hypothetical protein**
viuB	Vibriobatin utilisation protein ViuB		CSAK6_01268	**hypothetical protein**
cynR_1	yn operon transriptional ativator		CSAK6_01533	**hypothetical protein**

CSAK767_03654	antimirobial resistane protein Mig-		CSAK6_01534	**hypothetical protein**
hipB	Antitoxin HipB		CSAK6_01536	**hypothetical protein**
CSAK767_04129	Bateriophage lambda Kil protein		CSAK6_01537	**hypothetical protein**
CSAK767_04208	phage antitermination protein Q		CSAK6_01542	**hypothetical protein**
CSAK767_04182	phage minor tail protein L		CSAK6_01543	**hypothetical protein**
CSAK767_04187	phage tail protein		CSAK6_01544	**hypothetical protein**
CSAK767_04203	phage-related lysozyme (muraminidase)		CSAK6_01545	**hypothetical protein**
CSAK767_04185	phage-related minor tail protein		CSAK6_01791	**hypothetical protein**
CSAK767_04183	phage-related protein		CSAK6_01792	**hypothetical protein**
CSAK767_04180	phage-related protein, tail omponent		CSAK6_01793	**hypothetical protein**
CSAK767_04179	phage-related protein, tail omponent		CSAK6_01794	**hypothetical protein**
CSAK767_03679	putative aetyltransferase		CSAK6_02026	**hypothetical protein**
ywnA	putative HTH-type transriptional regulator ywnA		CSAK6_02033	**hypothetical protein**
pknH	putative serine/threonine-protein kinase pknH		CSAK6_02034	**hypothetical protein**
yusV_2	putative siderophore transport system ATP-binding protein YusV		CSAK6_02037	**hypothetical protein**
ariR	putative two-omponent-system onnetor protein AriR		CSAK6_02038	**hypothetical protein**
CSAK767_01933	S ribosomal protein L22/unknown domain fusion protein		CSAK6_02039	**hypothetical protein**
dsbD_2	Thiol:disulfide interhange protein DsbD preursor		CSAK6_02133	**hypothetical protein**
CSAK767_04198	3hypothetial protein		CSAK6_02137	**hypothetical protein**
CSAK767_03863	Abi-like protein		CSAK6_02138	**hypothetical protein**
relB_2	Antitoxin RelB		CSAK6_02139	**hypothetical protein**
CSAK767_00888	Ar-like DNA binding domain protein		CSAK6_02140	**hypothetical protein**
CSAK767_01258	Ar-like DNA binding domain protein		CSAK6_02172	**hypothetical protein**
CSAK767_02270	Bateriophage II protein		CSAK6_02173	**hypothetical protein**
CSAK767_02234	Bateriophage lysis protein		CSAK6_02631	**hypothetical protein**
CSAK767_01964	Bateriophage P2-related tail formation protein		CSAK6_02633	**hypothetical protein**
CSAK767_01943	Bateriophage repliation gene A protein (GPA)		CSAK6_02634	**hypothetical protein**

traD	DNA transport protein TraD		CSAK6_02635	**hypothetical protein**
CSAK767_01974	DNA-binding transriptional regulator		CSAK6_02636	**hypothetical protein**
fieF_1	Ferrous-iron efflux pump FieF		CSAK6_02637	**hypothetical protein**
CSAK767_01962	Gene 2-like lysozyme		CSAK6_02638	**hypothetical protein**
CSAK767_03640	hypothetial protein		CSAK6_02664	**hypothetical protein**
CSAK767_04122	hypothetial protein		CSAK6_02895	**hypothetical protein**
CSAK767_03681	hypothetial protein		CSAK6_02895	**hypothetical protein**
CSAK767_03592	hypothetial protein		CSAK6_02896	**hypothetical protein**
CSAK767_03651	hypothetial protein		dltB	**hypothetical protein**
CSAK767_03690	hypothetial protein		clpV1_1	**hypothetical protein**
CSAK767_03586	hypothetial protein		clpV1_2	**hypothetical protein**
CSAK767_03670	hypothetial protein		CSAK6_03562	**hypothetical protein**
CSAK767_03601	hypothetial protein		CSAK6_03563	**hypothetical protein**
CSAK767_03645	hypothetial protein		CSAK6_03568	**hypothetical protein**
CSAK767_03668	hypothetial protein		CSAK6_03569	**hypothetical protein**
CSAK767_03559	hypothetial protein		CSAK6_03570	**hypothetical protein**
CSAK767_03701	hypothetial protein		CSAK6_03571	**hypothetical protein**
CSAK767_03619	hypothetial protein		CSAK6_03574	**hypothetical protein**
CSAK767_03615	hypothetial protein		CSAK6_03575	**hypothetical protein**
mngR_1	hypothetial protein		CSAK6_03577	**hypothetical protein**
CSAK767_04133	hypothetial protein		CSAK6_03578	**hypothetical protein**
CSAK767_03673	hypothetial protein		CSAK6_03582	**hypothetical protein**
recT	hypothetial protein		CSAK6_03590	**hypothetical protein**
CSAK767_03600	hypothetial protein		CSAK6_03594	**hypothetical protein**
CSAK767_03603	hypothetial protein		CSAK6_03599	**hypothetical protein**
CSAK767_04222	hypothetial protein		CSAK6_03609	**hypothetical protein**
yycE	hypothetial protein		CSAK6_03612	**hypothetical protein**
CSAK767_03683	hypothetial protein		CSAK6_03616	**hypothetical protein**

CSAK767_03594	hypothetial protein		CSAK6_03617	**hypothetical protein**
CSAK767_04038	hypothetial protein		CSAK6_03618	**hypothetical protein**
psiB	hypothetial protein		CSAK6_03620	**hypothetical protein**
CSAK767_03587	hypothetial protein		CSAK6_03621	**hypothetical protein**
CSAK767_03602	hypothetial protein		CSAK6_03622	**hypothetical protein**
CSAK767_03612	hypothetial protein		CSAK6_03623	**hypothetical protein**
CSAK767_03584	hypothetial protein		umuC_1	**hypothetical protein**
CSAK767_03577	hypothetial protein		CSAK6_03627	**hypothetical protein**
CSAK767_03694	hypothetial protein		ygiW_3	**hypothetical protein**
CSAK767_03644	hypothetial protein		CSAK6_03634	**hypothetical protein**
CSAK767_03578	hypothetial protein		CSAK6_03635	**hypothetical protein**
CSAK767_04094	hypothetial protein		CSAK6_03636	**hypothetical protein**
CSAK767_03585	hypothetial protein		CSAK6_03637	**hypothetical protein**
CSAK767_03579	hypothetial protein		CSAK6_03638	**hypothetical protein**
CSAK767_03572	hypothetial protein		CSAK6_03639	**hypothetical protein**
CSAK767_03646	hypothetial protein		CSAK6_03640	**hypothetical protein**
CSAK767_03650	hypothetial protein		CSAK6_03641	**hypothetical protein**
traV	hypothetial protein		CSAK6_03642	**hypothetical protein**
CSAK767_03561	hypothetial protein		CSAK6_03643	**hypothetical protein**
CSAK767_04190	hypothetial protein		CSAK6_03645	**hypothetical protein**
CSAK767_04215	hypothetial protein		CSAK6_03762	**hypothetical protein**
CSAK767_03660	hypothetial protein		CSAK6_03763	**hypothetical protein**
CSAK767_03591	hypothetial protein		htpX_2	**hypothetical protein**
CSAK767_03641	hypothetial protein		CSAK6_03769	**hypothetical protein**
CSAK767_04124	hypothetial protein		CSAK6_03774	**hypothetical protein**
CSAK767_04201	hypothetial protein		CSAK6_03775	**hypothetical protein**
CSAK767_04135	hypothetial protein		CSAK6_03778	**hypothetical protein**
CSAK767_04204	hypothetial protein		CSAK6_03852	**hypothetical protein**

CSAK767_04205	hypothetial protein		CSAK6_03853	**hypothetical protein**
CSAK767_03538	hypothetial protein		CSAK6_03854	**hypothetical protein**
CSAK767_03560	hypothetial protein		CSAK6_03876	**hypothetical protein**
CSAK767_03647	hypothetial protein		CSAK6_03877	**hypothetical protein**
CSAK767_04194	hypothetial protein		CSAK6_03878	**hypothetical protein**
CSAK767_03542	hypothetial protein		CSAK6_03880	**hypothetical protein**
CSAK767_03692	hypothetial protein		ykfF	**hypothetical protein**
CSAK767_04132	hypothetial protein		ykfG_2	**hypothetical protein**
CSAK767_04212	hypothetial protein		CSAK6_03884	**hypothetical protein**
CSAK767_03568	hypothetial protein		yfdG	**hypothetical protein**
CSAK767_03535	hypothetial protein		CSAK6_04013	**hypothetical protein**
CSAK767_04193	hypothetial protein		CSAK6_04014	**hypothetical protein**
CSAK767_04211	hypothetial protein		CSAK6_04015	**hypothetical protein**
CSAK767_04184	hypothetial protein		CSAK6_04242	**hypothetical protein**
CSAK767_03537	hypothetial protein		CSAK6_04243	**hypothetical protein**
CSAK767_04213	hypothetial protein		CSAK6_04244	**hypothetical protein**
CSAK767_03606	hypothetial protein		CSAK6_04245	**hypothetical protein**
CSAK767_04186	hypothetial protein		CSAK6_04247	**hypothetical protein**
CSAK767_04191	hypothetial protein		CSAK6_04248	**hypothetical protein**
traC	hypothetial protein		CSAK6_04249	**hypothetical protein**
CSAK767_03605	hypothetial protein		CSAK6_04250	**hypothetical protein**
CSAK767_04128	hypothetial protein		yebZ_2	Inner membrane protein YebZ
CSAK767_03611	hypothetial protein		yjeO	Inner membrane protein yjeO
CSAK767_04127	hypothetial protein		CSAK6_03770	Integrase core domain protein
CSAK767_03583	hypothetial protein		CSAK6_03601	integrating conjugative element membrane protein, family
CSAK767_03689	hypothetial protein		CSAK6_03604	integrating conjugative element membrane protein, family

CSAK767_04178	hypothetial protein		CSAK6_03583	integrating conjugative element protein PilL, PFGI-1 class
CSAK767_03567	hypothetial protein		CSAK6_03572	integrating conjugative element protein, family
CSAK767_03590	hypothetial protein		CSAK6_03595	integrating conjugative element protein, family
CSAK767_03661	hypothetial protein		CSAK6_03597	integrating conjugative element protein, family
CSAK767_03663	hypothetial protein		CSAK6_03603	integrating conjugative element protein, family
CSAK767_04134	hypothetial protein		CSAK6_03606	integrating conjugative element protein, family
CSAK767_04197	hypothetial protein		CSAK6_03607	integrating conjugative element protein, family
CSAK767_03536	hypothetial protein		CSAK6_03608	integrating conjugative element protein, family
CSAK767_04188	hypothetial protein		CSAK6_03613	integrating conjugative element protein, family
CSAK767_03574	hypothetial protein		CSAK6_03614	integrating conjugative element protein, family
CSAK767_04189	hypothetial protein		CSAK6_03615	integrating conjugative element protein, family
CSAK767_04131	hypothetial protein		CSAK6_02639	integrating conjugative element relaxase, PFGI-1 class
CSAK767_01934	hypothetial protein		CSAK6_03566	integrating conjugative element, PFGI_1 class, ParB family protein
CSAK767_01935	hypothetial protein		CSAK6_03602	integrative conjugative element protein, RAQPRD family
CSAK767_01948	hypothetial protein		CSAK6_03591	invasion protein IagB
CSAK767_01949	hypothetial protein		fecA_2	Iron(III) dicitrate transport protein FecA
CSAK767_01975	hypothetial protein		fecC	Iron(III) dicitrate transport system permease protein fecC
yfdG	hypothetial protein		fecD	Iron(III) dicitrate transport system permease protein fecD
CSAK767_02238	hypothetial protein		CSAK6_03633	**IS1 transposase**
CSAK767_02262	hypothetial protein		CSAK6_03708	**IS1 transposase**
CSAK767_02263	hypothetial protein		kefC_2	K( )/H( ) antiporter
CSAK767_02267	hypothetial protein		lexA_1	LexA repressor
CSAK767_02273	hypothetial protein		mntB	Manganese transport system membrane protein mntB
CSAK767_02276	hypothetial protein		azr_1	NADPH azoreductase

CSAK767_02278	hypothetial protein		rcnA	Nickel/cobalt efflux system rcnA
CSAK767_02279	hypothetial protein		CSAK6_01540	ParG
CSAK767_02280	hypothetial protein		CSAK6_01265	**Phage gp6-like head-tail connector protein**
CSAK767_02281	hypothetial protein		CSAK6_01266	**Phage head-tail joining protein**
CSAK767_02283	hypothetial protein		CSAK6_01243	**phage N-6-adenine-methyltransferase**
CSAK767_02284	hypothetial protein		CSAK6_01261	**phage portal protein, HK97 family**
CSAK767_02285	hypothetial protein		CSAK6_01267	**phage protein, HK97 gp10 family**
CSAK767_02287	hypothetial protein		CSAK6_00935	**phage resistance protein**
CSAK767_02288	hypothetial protein		CSAK6_01259	**Phage terminase, small subunit**
CSAK767_02289	hypothetial protein		CSAK6_01260	**Phage terminase-like protein, large subunit**
CSAK767_02290	hypothetial protein		bar	Phosphinothricin N-acetyltransferase
CSAK767_02291	hypothetial protein		CSAK6_03586	Pilin accessory protein (PilO)
CSAK767_02292	hypothetial protein		CSAK6_03584	PilM
CSAK767_02293	hypothetial protein		intA_2	**Prophage CP4-57 integrase**
CSAK767_02294	hypothetial protein		yhfQ	Putative ABC transporter substrate-binding lipoprotein yhfQ precursor
CSAK767_02295	hypothetial protein		yfjZ	Putative antitoxin YfjZ
CSAK767_02296	hypothetial protein		CSAK6_00148	putative ATPase
CSAK767_02297	hypothetial protein		CSAK6_01179	putative ATPase
CSAK767_02298	hypothetial protein		CSAK6_03619	putative ATPase
CSAK767_02303	hypothetial protein		CSAK6_01252	putative chitinase
CSAK767_03805	hypothetial protein		pcoE_1	putative copper-binding protein pcoE precursor
repE	hypothetial protein		CSAK6_01541	putative crown gall tumor protein VirC1
umuD_2	hypothetial protein		pinR	Putative DNA-invertase from lambdoid prophage Rac
CSAK767_03822	hypothetial protein		CSAK6_03705	putative enzyme related to lactoylglutathione lyase
CSAK767_03836	hypothetial protein		CSAK6_00939	putative flavoprotein
CSAK767_03837	hypothetial protein		yghU	putative GST-like protein yghU
CSAK767_03842	hypothetial protein		yafY	putative lipoprotein yafY precursor

CSAK767_03844	hypothetial protein		CSAK6_03766	putative membrane protein
CSAK767_04424	hypothetial protein		CSAK6_03951	Putative NADH-flavin reductase
CSAK767_04425	hypothetial protein		CSAK6_03952	putative outer membrane protein
CSAK767_04426	hypothetial protein		CSAK6_03579	putative periplasmic iron-binding protein precursor
hcp1_3	hypothetial protein		CSAK6_01263	putative phage phi-C31 gp36 major capsid-like protein
CSAK767_00337	hypothetial protein		CSAK6_01246	putative phage-encoded protein
CSAK767_00338	hypothetial protein		intS_1	Putative prophage CPS-53 integraseintS_2e
CSAK767_00885	hypothetial protein		fecI	putative RNA polymerase sigma factor fecI
CSAK767_01257	hypothetial protein		yusV_2	putative siderophore transport system ATP-binding protein YusV
CSAK767_01937	hypothetial protein		CSAK6_01532	putative sporulation protein YtaF
CSAK767_01939	hypothetial protein		ypjF	putative toxin YpjF
CSAK767_01940	hypothetial protein		CSAK6_00135	putative transcriptional regulator
CSAK767_01941	hypothetial protein		CSAK6_00347	putative transcriptional regulator
ybiI_2	hypothetial protein		bin3	Putative transposon Tn552 DNA-invertase bin3
CSAK767_01944	hypothetial protein		CSAK6_04246	Relaxase/Mobilisation nuclease domain protein
smf_2	hypothetial protein		repB	RepFIB replication protein A
CSAK767_01955	hypothetial protein		dnaB_2	Replicative DNA helicase
CSAK767_01963	hypothetial protein		CSAK6_03598	Restriction endonuclease
CSAK767_01965	hypothetial protein		sigM	RNA polymerase sigma factor sigM
CSAK767_01966	hypothetial protein		frmA_2	S-(hydroxymethyl)glutathione dehydrogenase
CSAK767_02128	hypothetial protein		cusS_1	Sensor kinase CusS
CSAK767_02129	hypothetial protein		CSAK6_01237	Serine dehydrogenase proteinase
CSAK767_02130	hypothetial protein		frmB	S-formylglutathione hydrolase frmB
CSAK767_02131	hypothetial protein		silE_1	Silver-binding protein silE precursor
CSAK767_02232	hypothetial protein		CSAK6_00350	Site-specific recombinase XerD
CSAK767_02240	hypothetial protein		CSAK6_01234	Site-specific recombinase XerD
CSAK767_02254	hypothetial protein		CSAK6_00349	site-specific tyrosine recombinase XerC

CSAK767_02257	hypothetial protein		CSAK6_01180	site-specific tyrosine recombinase XerC
CSAK767_02272	hypothetial protein		CSAK6_01928	site-specific tyrosine recombinase XerC
CSAK767_02274	hypothetial protein		CSAK6_03593	site-specific tyrosine recombinase XerC
CSAK767_02275	hypothetial protein		hspA_2	Spore protein SP21
CSAK767_03127	hypothetial protein		trxC_1	Thioredoxin-2
CSAK767_03235	hypothetial protein		parE4	Toxin ParE4
CSAK767_03806	hypothetial protein		ykfI	Toxin YkfI
CSAK767_03807	hypothetial protein		copR	Transcriptional activator protein CopR
CSAK767_03808	hypothetial protein		CSAK6_00940	transcriptional regulator, Acidobacterial, PadR-family
parM	hypothetial protein		frmR_2	Transcriptional repressor frmR
CSAK767_03817	hypothetial protein		pifC	Transcriptional repressor pifC
CSAK767_03818	hypothetial protein		CSAK6_03596	Transglycosylase SLT domain protein
CSAK767_03821	hypothetial protein		CSAK6_00934	**Transposase**
CSAK767_03823	hypothetial protein		CSAK6_03560	**Transposase**
CSAK767_03824	hypothetial protein		CSAK6_03561	**Transposase**
CSAK767_03831	hypothetial protein		CSAK6_03698	**Transposase**
CSAK767_03832	hypothetial protein		CSAK6_03699	**Transposase**
CSAK767_03833	hypothetial protein		CSAK6_03700	**Transposase**
CSAK767_03834	hypothetial protein		CSAK6_03768	**Transposase DDE domain protein**
CSAK767_03846	hypothetial protein		CSAK6_03703	**Transposase IS116/IS110/IS902 family protein**
CSAK767_03849	hypothetial protein		CSAK6_00944	**Transposase, TnpA family**
CSAK767_03861	hypothetial protein		CSAK6_02632	Type I restriction-modification system methyltransferase subunit
CSAK767_03944	hypothetial protein		CSAK6_03589	type II secretion system protein F
CSAK767_04423	hypothetial protein		epsE_3	Type II traffic warden ATPase
CSAK767_03827	Integrase ore domain protein		CSAK6_03592	Type IV leader peptidase family protein
CSAK767_03868	Integrase ore domain protein		CSAK6_03587	type IV pilus biogenesis protein PilP
CSAK767_03862	Integrase ore domain protein		CSAK6_03611	Type IV secretory pathway, VirB4 components

CSAK767_00886	invasion protein IagB		CSAK6_03600	Type IV secretory pathway, VirD4 components
iutA_1	Iron-regulated outer membrane proteins		CSAK6_03628	tyrosine recombinase XerC
treZ_3	Malto-oligosyltrehalose trehalohydrolase		xerD_3	Tyrosine recombinase XerD
CSAK767_03126	modulator of post-segregation killing protein		cpsD	Tyrosine-protein kinase CpsD
CSAK767_02282	Mu-like prophage protein gp		CSAK6_03866	YfdX protein
CSAK767_01959	Mu-like prophage protein gpG		CSAK6_03867	YfdX protein
traI_1	Multifuntional onjugation protein TraI		CSAK6_00945	Zinc-type alcohol dehydrogenase-like protein
CSAK767_02255	NinB protein			
cpo	Non-heme hloroperoxidase			
CSAK767_03848	conjugal transfer mating pair stabilization protein TraG			
CSAK767_03532	conjugal transfer mating pair stabilization protein TraN			
CSAK767_03945	conjugal transfer pilus aetylation protein TraX			
CSAK767_03544	conjugal transfer pilus assembly protein TraB			
CSAK767_03847	conjugal transfer pilus assembly protein TraH			
CSAK767_00887	conjugal transfer protein TraM			
CSAK767_03850	conjugal transfer surfae exlusion protein TraT			
CSAK767_01950	Phage apsid saffolding protein (GPO) serine peptidase			
CSAK767_02299	phage head morphogenesis protein, SPP gp family			
CSAK767_01953	Phage head ompletion protein (GPL)			
CSAK767_01957	phage lysis regulatory protein, LysB family			
CSAK767_01951	phage major apsid protein, Pfamily			
CSAK767_01968	phage major tail tube protein			
CSAK767_01961	Phage Pbaseplate assembly protein gpV			
CSAK767_01947	phage portal protein, PBSX family			
CSAK767_01973	Phage protein D			
CSAK767_01972	Phage protein U			
CSAK767_01938	Phage regulatory protein II (P)			
CSAK767_01256	Phage regulatory protein Rha (Phage_pRha)			

CSAK767_01952	Phage small terminase subunit			
CSAK767_02277	phage Tail ollar Domain protein			
CSAK767_01969	Phage tail protein E			
CSAK767_01954	Phage Tail Protein X			
CSAK767_01967	Phage tail sheath protein			
CSAK767_01970	phage tail tape measure protein, TP family, ore region			
CSAK767_02300	phage-assoiated protein, family			
CSAK767_01956	Phage-related lysozyme (muraminidase)			
CSAK767_02233	Phage-related lysozyme (muraminidase)			
CSAK767_01971	Phage-related minor tail protein			
CSAK767_03816	Plasmid stability protein			
pstP	PP2-family Ser/Thr phosphatase			
CSAK767_01958	Pphage tail ompletion protein R (GpR)			
intA_2	Prophage P- integrase			
intA_3	Prophage P- integrase			
parB	putative hromosome-partitioning protein parB			
CSAK767_02264	putative HTH-type transriptional regulator			
CSAK767_02235	putative phage-enoded protein			
CSAK767_03838	putative rown gall tumor protein Vir			
CSAK767_01945	ribose-phosphate pyrophosphokinase			
CSAK767_04125	Siphovirus Gp			
CSAK767_02239	Terminase small subunit			
CSAK767_02371	Transposase			
CSAK767_03128	Transposase			

Appendix text 1. 1 Mugsy whole genome alignment and algorithms used to extract the core genome

(The comment in red font within curly brackets describes each command)

Whole genome alignment using Mugsy

To run Mugsy, all the files must be in the fasta format (e.g. 1.fasta) and in the Mugsy folder

Navigate to the Mugsy folder from the terminal

To run mugsy use the following command

./mugsy 1.fasta 2.fasta 3.fasta {if we have 3 genomes}

Output file will be a multiple alignment file in the form of tmp file rename it (*.maf)and run through the following scripts before running the raxML ($1 is the *.maf so replace it)

Core genome extraction

python maf_to_fasta.py < $1 > temp.fasta {generates a temporary fasta file}

python get_record_ids.py temp.fasta > tmp {generates a temporary file used to make record}

sort -u -k 1,1 tmp > record_ids.txt {generates a text file containing the records}

python maf_thread_for_species_list.py record_ids.txt < $1 > new.maf {creates a new maf file}

python maf_to_concat_fasta_list.py record_ids.txt < new.maf > concat.fasta {generates a concatenated fasta file from new maf file}

mothur "#filter.seqs(fasta=concat.fasta, vertical=F, trump=-)" {filters the concatenated fasta file}

mothur "#filter.seqs(fasta=concat.filter.fasta, vertical=F, trump=.)" {again filters the concatenated fasta file}

mv concat.filter.filter.fasta concat_alignment.fas {creates a double filtered concatenated alignment file}

Visualise the concat_alignment.fas file using the seaview and save as phylip file (*.phy) as raxML only accepts the phy files

Phylogeny estimation by RAxML

To run RaxML move the *.phy file to the Mugsy folder

Navigate to the Mugsy folder from the terminal and run the following command line

ramlHPC –f a –m GTRGAMMA –x12345 -# 100 –s *.phy –n *tree {where –s is the name of the phylip file generated before and the –n is the name of the output file o raxML}

The RAxML best tree can then be viewed and annotated using FigTree.

Appendix text 1. 2 Commands used for SNP calls using SMALT, SAMtools and BCFtools

(The comment in red font within curly brackets describes each command)

./smalt_x86_64 index -k 17 -s 2 Refindex Csak.fasta {builds a hash table for the reference genome in file Csak.fasta}

./smalt_x86_64 map -f sam -o 1.sam Refindex 1_read1.fastq 1_read2.fastq {loads the hash table created by the previous step into memory and maps paired-end reads in the files1_read1.fastq 1_read2.fastq}

samtools faidx Csak.fasta {Index reference sequence Csak.fasta in the FASTA format or extract subsequence from indexed reference sequence}

samtools view -bS -t Csak.fasta.fai -o 1.bam 1.sam {Extract all or sub alignments in SAM or BAM format. Since, no regions were specified it extracted all the alignments}

samtools sort 1.bam 1.sort {sorts the alignment by the leftmost coordinates}

samtools rmdup 1.sort.bam 1.out.bam {remove potential duplicates to retain the highest quality variant calls}

samtools mpileup -ugf Csak.fasta 1.out.bam | bcftools view -bvcg - > var.raw.bcf {Produce a BCF for the BAM file. The records of the alignment are grouped by sample identifiers in @RG header lines}

bcftools view var.raw.bcf | perl Vcfutils.pl varFilter -D 100 > 1var.vcf {converts BCF file to VCF file}

bgzip 1var.vcf {produce and compress the TAB-delimited VCF file}

tabix -p vcf 1var.vcf.gz {indexes the TBA-delimited genome position file from the compressed and sorted VCF file}
samtools index 1.out.bam {index the sorted alignment for the fast and random access, index file 1.out.bam.bai will also be created}