Friedrich-Loeffler-Institut,
Bundesforschungsinstitut für Tiergesundheit,
Institut für Bakterielle Infektionen und Zoonosen,
Jena

eingereicht über das

Institut für Mikrobiologie und Tierseuchen
des Fachbereichs Veterinärmedizin
der Freien Universität Berlin

Genome sequencing and molecular typing of *Clostridium chauvoei*

Inaugural-Dissertation
zur Erlangung des Grades eines
Doktors der Veterinärmedizin
an der
Freien Universität Berlin

vorgelegt von
Prasad Thomas
Tierarzt aus Nemmara (Indien)

Berlin 2018
Journal-Nr.: 4024

mbvberlin | mensch und buch verlag

Aus dem

**Friedrich-Loeffler-Institut,
Bundesforschungsinstitut für Tiergesundheit,
Institut für Bakterielle Infektionen und Zoonosen,
Jena**

eingereicht über das

**Institut für Mikrobiologie und Tierseuchen
des Fachbereichs Veterinärmedizin
der Freien Universität Berlin**

Genome sequencing and molecular typing of *Clostridium chauvoei*

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors der Veterinärmedizin
an der Freien Universität Berlin

vorgelegt von

Prasad Thomas
Tierarzt aus Nemmara (Indien)

Berlin 2018

Journal-Nr.: 4024

Gedruckt mit Genehmigung des Fachbereichs Veterinärmedizin
der Freien Universität Berlin

Dekan:	Univ.-Prof. Dr. Jürgen Zentek
Erster Gutachter:	Prof. Dr. Lothar H. Wieler
Zweiter Gutachter:	Prof. Dr. Heinrich Neubauer
Dritter Gutachter:	Univ.-Prof. Dr. Kerstin E. Müller

Deskriptoren (nach CAB-Thesaurus):
clostridium chauvoei*;* gangrene*;* genome analysis; comparative genomics*;* phylogeny*;* clustered regularly interspaced short palindromic repeats (CRISPR); multi-locus sequence typing (MLST); germany

Tag der Promotion: 29.01.2018

Bibliografische Information der *Deutschen Nationalbibliothek*
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

ISBN: 978-3-86387-875-7
Zugl.: Berlin, Freie Univ., Diss., 2018
Dissertation, Freie Universität Berlin
D 188

Dieses Werk ist urheberrechtlich geschützt.
Alle Rechte, auch die der Übersetzung, des Nachdruckes und der Vervielfältigung des Buches, oder Teilen daraus, vorbehalten. Kein Teil des Werkes darf ohne schriftliche Genehmigung des Verlages in irgendeiner Form reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Die Wiedergabe von Gebrauchsnamen, Warenbezeichnungen, usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürfen.

This document is protected by copyright law.
No part of this document may be reproduced in any form by any means without prior written authorization of the publisher.

Alle Rechte vorbehalten | all rights reserved
© Mensch und Buch Verlag 2018
Choriner Str. 85 - 10119 Berlin
verlag@menschundbuch.de – www.menschundbuch.de

Sponsorship
Indian Council of Agricultural Research (ICAR), Government of India.
The author is awarded with an ICAR International Fellowship.

Table of content Page no

I

Abbreviations

ATP	Adenosine Triphosphate
AFLP	Amplified Fragment Length Polymorphism
bp	Base Pair
BVSc & AH	Bachelor of Veterinary Science and Animal Husbandry
C	***Clostridium***
CARD	Comprehensive Antibiotic Research Database
CDS	Coding Sequences
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTns	Conjugative Transposons
cgMLST	Core Genome Multilocus Sequence Typing
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
dsDNA	Double Stranded DNA
ddNTP	Dideoxynucleotide Triphosphate
Dnase	Deoxyribonuclease
DSM	Deutsche Sammlung von Mikroorganismen und Zellkulturen
DUE	DNA Unwinding Element
ELISA	Enzyme Linked Immunosorbent Assay
EC	Enzyme Code
FliC	Flagellin type C
FAT	Florescent Antibody Technique
FLI	Friedrich-Loeffler-Institut
GC	Guanosine – Cytosine
GO	Gene Ontology
GAGE	Genome Assembly Gold standard Evaluation
GR	Germinant Receptor
HGAP	Hierarchical Genome Assembly Process
hqSNPs	High Quality SNPs
IVRI	Indian Veterinary Research Institute
ICEs	Integrative Elements
IBIZ	Institute of Bacterial Infections and Zoonoses
IMT	Institute of Microbiology and Epizootics
ISs	Insertion Sequences
Indels	Insertion/Deletions
In	Integrons
KAU	Kerala Agricultural University
kDa	Kilodalton
Kb	Kilobase
LCB	Locally Collinear Blocks
ML	Maximum Likelihood
MP	Maximum Parsimony
MLST	Multilocus Sequence Typing

MLVA	Multiple Locus Variable Number Tandem Repeat Analysis
Mb	Megabase
MST	Minimum Spanning Tree
MGE	Mobile Genetic Element
N-terminal	Amino-Terminal
NCBI	National Centre for Biotechnology Information
NCTC	National Collection of Type Culture
NS	Non Synonymous
OriC	Origin of replication
PacBio	Pacific Biosciences
PaLoc	Pathogenicity Locus
PCR	Polymerase Chain Reaction
PFGE	Pulse Field Gel Electrophoresis
PGM	Personal Genome Machine
rDNA	Ribosomal DNA
rRNA	Ribosomal RNA
RAPD	Randomly Amplified Polymorphic DNA
S	Synonymous
SBS	Sequencing By Synthesis
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SASP	Small, Acid-Soluble Spore Proteins
tRNA	Transfer RNA
Tns	Transposons
U	Unit
wgMLST	Whole Genome Multilocus Sequence Typing

List of tables

List of figures

Chapter 1: Introduction and review of the literature

1.1 Introduction

Blackleg caused by *Clostridium (C.) chauvoei* is a bacterial disease that affects cattle and sheep with high mortality. Blackleg can cause huge economic losses to dairy farming across the globe. Blackleg in cattle is an endogenous infection and occurs without a wound or break of the skin (Hatheway, 1990). The majority of blackleg cases occur in young cattle due to ingestion of spores. The disease is characterized by inflammation and necrosis of skeletal and cardiac muscles, toxaemia and sudden death. Sheep get infected through skin wounds after shearing, castration and tail docking (Quinn et al., 2011). The pathogenesis of blackleg in cattle is not fully known and is presumed to start with uptake of spores by the animal while grazing. It is suspected that the spores enter the body through the intestinal mucosa or lesions in the oral cavity (Useh et al., 2006, Pires et al., 2017b). Vegetative *C. chauvoei* contains flagella (H) and somatic antigen (O), along with other toxins (Chandler, 1975, Useh et al., 2003). The pathogen also harbours *Clostridium chauvoei* toxin A (CctA) which belongs to the leukocidin superfamily of bacterial toxins and is considered to be a major virulence factor and a potent protective antigen target for vaccines against blackleg (Frey et al., 2012, Frey and Falquet, 2015). The toxin was also found to be well conserved in different *C. chauvoei* strains (Rychener et al., 2017). Sialidases and hyaluronidases are other virulence factors of *C. chauvoei* (Useh et al., 2003). Sialidase (nanA) and hyalurondiases (nagH and nagI) exhibited genetic variability among the strains originating from Australia, New Zealand and United Kingdom as compared to strains originating from other regions (Rychener et al., 2017).

Currently there are no comprehensive data regarding the genomic population structure and diversity of this important animal pathogen. The species is also not described to have any genotypes so far. Since the first report of whole genome sequence data for a virulent strain from Switzerland (Falquet et al., 2013), the pathogens genome components and potential virulence factors were unravelled at the genomic level (Frey and Falquet, 2015). Genome sequencing and comparative genomics involving 20 strains form wide geographical origin have just very recently revealed a limited genetic variability for the pathogen (Rychener et al., 2017). However, a group of strains originating from Australia, New Zealand and the United Kingdom showed remarkable differences to those from Europe, Africa, and America (Rychener et al., 2017). *C. chauvoei* genomes have been reported to harbour a CRISPR element similar to the CRISPR subtype I-B system (Rychener et al., 2017) and the application of unique spacer sequences of CRISPR elements for strain differentiation has also been described (Rychener et al., 2017).

The current study was designed to gain insights in the genome content and composition of this species based on the generation of a high quality complete reference genome sequence of a recently isolated German field strain and the type strain. A comparative genomic study involved 64 *C. chauvoei* strains, mostly of European origin. Related strains, isolated from the same animal, the same outbreak and from a consecutive outbreak at the same farm (and to some extend from longitudinal time gaps) were involved to reveal microevolution of the pathogen. Genome assembly, annotation and various comparative genomic tools were applied

to discover the genome, coding sequences, virulence factors, plasmids, phages and CRISPR element composition. Pan-genome analysis was carried out to detect new gene acquisition for the species. Homologous recombination event predictions were carried out to identify the involvement of horizontal gene transfers. Pan-genome SNP based phylogeny and clustering was applied for clustering of isolates. The study also attempted to compare core genome MLST and CRISPR spacer sequence based typing options. The main goals of the study were:

- To define the genome content and composition of *C. chauvoei* by generating complete genome sequences for a field strain and the type strain.
- To define core and accessory genes, to investigate the phylogenetic structure and to evaluate genotyping options for *C. chauvoei* strains.

1.2 Review of the literature

1.2.1 The genus *Clostridium*

Clostridiae are Gram-positive anaerobes within the phylum *Firmicutes*. The genus includes more than 200 species. The majority are saprophytes, whereby around 50 species can cause clinical diseases in human and animals including birds. Pathogenic species produce various toxins and few produce highly potent neurotoxins such as *Clostridium (C.) tetani* and *C. botulinum*. Clostridia are spore formers and the endospores formed are resistant to various environmental factors such as heat, desiccation, radiation, pH variation, atmospheric pressure and chemical agents. Heat resistance can cause spoilage of canned foods. Few of the species also have industrial applications and *C. acetobutylicum* which is used for manufacturing butanol and some species are currently being explored for their potential use as therapeutic agents in tumour therapy (Kubiak and Minton, 2015). Most of the species are inhabitants of the environment and can grow in soil, water, decaying animal and plant materials and are also present in the intestinal tracts of humans and animals (Prescott et al., 2002). The type species assigned for the genus is *C. butyricum,* a normal inhabitant of the human and animal gut. The members of the genus also show phenotypical and genotypic variation with respect to Gram-staining, GC content, genome size, etc. (Kalia et al., 2011, Gupta and Gao, 2009). Intraspecies clustering and divergence are also observed in the genus and hence the genus is currently being proposed for revision with respect to inclusion and omission of species from certain families. For example recent studies have proposed a reclassification of *C. difficile* and similar species into new family-level group (Collins et al., 1994, Yutin and Galperin, 2013, Lawson et al., 2016).

The most prominent pathogenic Clostridia species are *C. botulinum*, *C. chauvoei, C. haemolyticum*, *C. novyi*, *C. perfringens*, *C. septicum* and *C. tetani.* Pathogenic clostridia including those affecting cattle and sheep can be divided into three groups i.e. neurotoxic, histotoxic and enterotoxic clostridia according to the clinical picture. Tetanus and botulism are caused by the neurotoxic species *C. tetani* and *C. botulinum,* respectively. Histotoxic *Clostridia* spp release various exotoxins that damage tissues and cause toxaemia as in blackleg and infectious necrotic hepatitis ("black disease") caused by *C. chauvoei* and *C. novyi*, respectively. Enterotoxic clostridia release toxins leading to enteritis and enterotoxaemia in diseases caused by *C. perfringens* and *C. difficile* (Quinn et al., 2011). Important clostridial

diseases affecting humans are botulism, tetanus, gas gangrene, food poisoning, pseudomembranous colitis and antibiotic associated diarrhoea. Significant clostridial diseases affecting animals, especially ruminants are black quarter or blackleg (*C. chauvoei*), malignant oedema (*C. septicum*), bacillary haemoglobinuria (*C. haemolyticum*), black disease (*C. novyi*), bradsot or braxy (*C. septicum*), pulpy kidney disease (*C. pefringens* type D), lamb dysentery (*C. pefringens* type B), big head (*C. novyi*, *C. sordellii*), struck (*C. pefringens* type C), tetanus (*C. tetani*) and botulism (*C. botulinum*) (Quinn et al., 2011, Prescott et al., 2016).

16S rRNA-based phylogenetic analysis of the genus *Clostridium* has revealed 19 phylogenetic clusters with cluster I forming the core group of the genus (Collins et al., 1994). More than half of the pathogenic species are members of this cluster including *C. chauvoei* along with other major pathogenic agents of the genus, i.e. *C. botulinum*, *C. haemolyticum*, *C. novyi*, *C. perfringens*, *C. tetani* and *C. septicum* (Stackebrandt et al., 1999). A phylogenetic tree created based on N-terminal amino acid sequences of the FliC protein showed *C. chauvoei* to be more related to *C. septicum* (Sasaki et al., 2002a). The phylogenetic positioning of *C. chauvoei* based on 16s rDNA sequence has revealed relatedness to *C. septicum* and *C. carnis* (Kuhnert et al., 1996, Kalia et al., 2011). Similar genetic relatedness has also been observed based on β-barrel pore forming toxins of the lecuocidin superfamily within the genus (*Clostridium chauvoei* toxin A (CctA), α haemolysin, NetB, β toxin). The CctA toxin showed higher homology with α haemolysin of *C. botulinum* followed by NetB and β toxin from *C. perfringens* respectively (Frey et al., 2012). Phylogenetic comparison showed that the NanA sialidase gene of *C. chauvoei* was more related to *C. septicum* (Vilei et al., 2011). A recent comparative genome sequence based study involving *C. chauvoei* and *C. perfringens* strain 13 showed limited similarity among the two species based on collinear genetic elements (Frey and Falquet, 2015).

1.2.2 Blackleg (*Clostridium chauvoei*)

Clostridium chauvoei, the causative agent of blackleg (gangrenous myositis) is classified as histotoxic along with other members of the genus such as *C. septicum*, *C. novyi*, *C. haemolyticum* and *C. sordellii* (Quinn et al., 2011). The name of the pathogen was given in honour of the French veterinarian J. B. A. Chauveau. The organism is mostly found as single cell, but is occasionally observed in pairs or short chains. The Gram staining property is instable and turns from Gram-positive to Gram-negative especially in old cultures (Abreu et al., 2016). Endospores produced by the bacteria show a typical lemon shaped appearance in microscopy and the species is motile due to peritrichous flagella (Quinn et al., 2011).

Blackleg, predominantly a disease of ruminants, is known by different names such as “black quarter”, “symptomatic anthrax”, “quarter evil”, “Rauschbrand”, “Geräusch” and “charbon symptomatique”. According to (MacLennan, 1962) “Rauschbrand” was first recognized as a distinct disease in cattle by Bollinger in 1875 and Feser in 1876 in Southern-Germany. Later in 1879 the French researchers Arloing, Cornevin and Thomas proved that the disease was caused by an anaerobic bacillus, subsequently named *Bacterium chauvoei*, then *Clostridium feseri* and now *Clostridium chauvoei*. Blackleg in cattle is an endogenous infection and occurs without a wound or break of the skin (Hatheway, 1990). The disease is characterized by

sudden death, inflammation and necrosis of skeletal and cardiac muscles, toxaemia and high mortality. The affected animal shows lameness, affected muscles show crepitation and death usually occurs within 12-24 hours preceded by signs of systemic toxaemia, recumbency and coma. The disease has also been reported from deer, mink and ostrich (Armstrong and Macnamee, 1950, Langford, 1970, Lublin et al., 1993, Nagano et al., 2008). The majority of blackleg cases occur in young cattle of less than 2 years of age due to ingestion of spores. Sheep get infected through skin wounds after shearing, castration and tail docking (Quinn et al., 2011). *C. chauvoei* infection in sheep mostly resembles malignant oedema and the animals show symptoms such as anorexia, depression, high fever, lameness and crepitating lesions followed by sudden death in many cases (Songer, 1998). Fatal human *C. chauvoei* infections were reported in 2008 (Nagano et al., 2008) and 2012 (Weatherhead and Tweardy, 2012) also pointing to the possibility of widespread under reporting of *C. chauvoei* infection in humans. Therefore, *C. chauvoei* should be regarded as a zoonotic agent.

1.2.2.1 Epidemiology

The spores of *C. chauvoei* are capable of existing for many years in soil. Pastures contaminated with spores may lead to frequent outbreaks in endemic areas. The affected animals do not directly transmit disease, but the spores which are probably released from carcasses after opening, improper carcass disposal or even with dung can contaminate pastures and can lead to outbreaks. A further reason behind sudden outbreaks may also involve earth movement by natural events or soil excavation leading to exposure of animals to buried spores.

Blackleg is a worldwide endemic disease which causes significant financial loss to cattle raisers in the United States of America, Latin America, Asia, Africa and Europe (Adams, 1998, Ramarao and Rao, 1990, Num, 2014, Chatikobo et al., 2013, Ayele et al., 2016, Tulley, 2010). In India the disease is the third important animal disease following foot and mouth disease and haemorrhagic septicaemia. The statistics generated for the year 2015 by the Department of Animal Husbandry, Dairying and Fisheries, Ministry of Agriculture and Farmers Welfare, Government of India, reported 90, 2 and 6 outbreaks in animal herds resulting in animal losses of 262, 21 and 18 in cattle, caprine/ovine and buffaloes respectively (http://aipvt.vci.nic.in/REPORT.pdf). In Iran clostridial infections are among the most important diseases of cattle and sheep and blackleg is the major disease affecting cattle (Pilehchian Langroudi, 2015). In Nigeria the losses of Zebu cattle have been estimated at US$ 4.3 million annually and the disease is categorised as a list A disease as it causes high mortality in cattle (Num, 2014, Useh et al., 2006). A recent study has concluded that blackleg is one of the major bacterial infections of cattle associated with tremendous economic losses to herders in many parts of Ethiopia. The financial costs in blackleg affected herds were estimated to be US$ 9.8 per head for local Zebu and US$ 16 per head for crossbred cattle (Ayele et al., 2016). Hence, it is considered advisable to carry out regular animal vaccination in endemic areas (Hirsh and Zee, 1999). In Germany the number of outbreaks reported from 1950 to present shows a decreasing trend averaging around 64 cases per year from 1950 to 1980 and 22 from 1980 to 2010. The decreasing trend was also observed in the subsequent years and the number of outbreaks was reported to be 13, 10, 6, 6 and 3 in years 2011, 2012, 2013, 2014 and 2015, respectively (https://www.fli.de/en/publications/annual-animal-health-

reports/). Since the report of Feser at the end of the 1870ies it is known that seasonal mountain pastures, so called "Almen", in Southern Germany, Austria and Switzerland are local and distinct endemic areas, named specifically "Rauschbrandalmen". Up to 2015, local veterinary authorities made vaccination obligate for bovines on those pastures and no outbreaks occurred. Interestingly since the obligatory vaccination program was stopped, sporadic outbreaks are observable even after decades (personal communication, Seyboldt 2017). A recent study aiming at identifying spatial and temporal clusters in the incidence of blackleg occurrence in Styria, Austria between 1986 and 2013 showed significant mean annual blackleg incidence variations between different municipalities. The study also showed blackleg cases are clustered within certain geographic areas (Wolf et al., 2017).

1.2.2.2 Pathogenesis

The pathogenesis of blackleg is not fully known and the most cited model states that the disease starts with ingestion of spores by the animal while grazing (Useh et al., 2006). The occurrence of *C. chauvoei* in 20% of spleens and livers of healthy cattle was demonstrated in a study and similar, corroborating results were found in healthy dogs livers (Barnes et al., 1975). The pathogen is believed to spread to large muscles of the foreleg, hindleg and cardiac muscles after being phagocytized in the intestine (Quinn et al., 2011). It is also suspected that these spores can remain latent in the macrophages (Pires et al., 2017b) and the germination of spores and disease occurrence are hypothesized to occur whenever the redox potential decreases in the muscles due to blunt trauma or hypoxia from excessive exercises (Abreu et al., 2016). The exact triggering factors for spore germination are still not clear. The occurrence of long lasting latency within muscles as the predominant reason for an outbreak is contradicted by reports of outbreaks occurring repeatedly where animals have been moved to new pastures, where the most probable outbreak reason is the sudden exposure of the animals to large numbers of spores. Even though the latency stage cannot be completely ruled out in such cases, involvement of additional events or a bacteraemia occurring directly from the intestine can also be further causes (Abreu et al., 2016). A recent study involving vegetative cells and spores showed that both forms could remain viable after internalization by murine and bovine macrophages. Vegetative cells and spores of *C. chauvoei* showed a pro-inflammatory and anti-inflammatory profile in the bovine macrophages, respectively. This result supports the hypothesis that macrophages may play a role in the early pathogenesis of blackleg by maintaining a latency phase (Pires et al., 2017b).

1.2.2.3 *Clostridium chauvoei* toxins and virulence factors

Clostridium chauvoei contains flagella (H) and somatic (O) antigen, along with other toxins (Chandler, 1975, Useh et al., 2003). Flagella have a pivotal role in bacterial motility, but have also been shown to be important in bacterial pathogenesis by acting as adhesins for host invasion (Haiko and Westerlund-Wikström, 2013). Immunity against *C. chauvoei* is considered to be anti-cellular rather than antitoxic and most studies have targeted somatic and flagellar antigens (Tamura et al., 1984, Chandler, 1975). Studies have shown the involvement of flagella as an important virulence factor for *C. chauvoei* (Tamura et al., 1995). Electrophoretic analysis of flagella revealed a major protein band of 46 kDa and two minor bands of 73 and 100 kDa probably representing polymorphic forms of the flagellin monomer

(Kojima et al., 1999). Sequence analysis of the *C. chauvoei* flagellin gene (*fliC*) predicted two probable glycosylation sites for the protein which were speculated to be an important post-translational modification for achieving protective immune response in mice (Kojima et al., 2000). Later studies proved the existence of a variable number of flagellin genes present among various species of the genus *Clostridium* (Tasteyre et al., 2000). For *C. chauvoei* two genes in tandem order designated as *fliC*(A) and *fliC*(B) with conserved N and C terminal ends have been reported (Sasaki et al., 2002a). Differences in length and sequence of the central variable region of the FliC genes of various species have also been applied for taxonomic grouping and the diagnosis of related clostridial species (Sasaki et al., 2002a).

Initial studies reported the production of several toxins by *C. chauvoei* and *C. septicum* including deoxyribonucleases (β-toxin), hyaluronidases (γ-toxin) and oxygen labile haemolysin, but oxygen stable haemolysins were not consistently obtained from culture filtrates (Moussa, 1958). Also deoxiribonuclease (DNase) activity was detected from 10 out of 12 *C. chauvoei* strains and from the controls used in another study (Carloni et al., 2005). Bacterial hyaluronidases are enzymes that break hyaluronate (a carbohydrate polymer that is part of the extracellular matrix) and thereby possibly help in the initiation and spread of infection. Hyaluronidases have been characterized from several Gram-positive bacteria as well as from *C. chauvoei* (Hynes and Walton, 2000, Princewill and Oakley, 1976, Frey and Falquet, 2015). An oxygen labile haemolysin was partially characterized from *C. chauvoei* which was found to cause haemolysis of sheep erythrocytes whereas horse erythrocytes were resistant (Tamura et al., 1992). *C. chauvoei* haemolysin was purified and identified as a 27 kDa protein and haemolytic experiments showed bovine, ovine and chicken erythrocytes were sensitive to haemolysin, whereas erythrocytes of rabbits, rats, mice and dogs were found to be resistant (Mudenda Hang'ombe et al., 2006).

Sialidases or neuraminidases are among the few previously characterized virulence factors of *C. chauvoei* (Useh et al., 2003). Sialidases are enzymes that hydrolyse the glycosidic linkage between sialic acid molecules and glycoconjugates and as a result degrade glycoproteins in cell junctions and mucins in infected tissue. A sialidase purified from *C. chauvoei* showed a molecular weight of 300 kDa consisting of two subunits each of 150 kDa (Heuermann et al., 1991). It was speculated that the rapid spread and brevity of blackleg pathogenesis could be an effect of sialidases causing an initial breach in cell junctions further followed by necrosis mediated by other toxins. Genetic and molecular characterization of a *C. chauvoei* NanA sialidase revealed a 150 kDa homodimer of 72 kDa monomer. Antibodies raised against a 40 kDa peptide region of the sialidase protein could neutralize the *in vitro* activity and hence was interpreted as indicative for its applicability as a potent vaccine candidate (Vilei et al., 2011).

Clostridium chauvoei toxin A (CctA) belongs to the leukocidin superfamily of bacterial toxins. This secreted toxin of 33 kDa molecular weight was highly cytotoxic to embryonic calf nasal epithelial cells and had high haemolytic activity against sheep erythrocytes. The toxin was also found to be well conserved in different *C. chauvoei* strains. The protein also displayed conservation of amino acid residues shared with other ß-barrel toxin proteins. CctA is considered to be the major virulence factor of *C. chauvoei* and a highly potent protective antigen in vaccines against blackleg (Frey et al., 2012, Frey and Falquet, 2015).

A recent study to characterize cell surface-associated proteins of *C. chauvoei* using a mass spectrometry approach identified about 150 distinct protein spots. Some of the important immune-reactive cell surface–associated proteins identified in the study include enolase, chaperonin, ribosomal protein L10, flavoprotein and glycosyl hydrolase. These proteins are also considered as potential candidates for vaccines and diagnostic assays (Jayaramaiah et al., 2016).

1.2.2.4 Diagnosis, differential diagnosis and prevention

The presumptive diagnosis of black quarter is carried out based on clinical signs, post mortem findings and case history. Post mortem lesions mainly occur in fore and hind limb muscles but may also occur in the masseter, intercostal muscles, psoas, tongue, diaphragm and heart. Pericarditis and pleurisy may also be observed. The clinical signs in sheep are similar to those in cattle. Sudden death was observed in lambs with *C. chauvoei* myocarditis. Confirmatory diagnosis is made by isolation and identification of the organism. In most cases shortly before death, septicaemia occurs and hence the organism can be cultivated not only from lesions, but also from the heart blood, liver and spleen (Hagan et al., 1988).

Other bacteria causing gas gangrene and necrotizing infections have to be considered as differential diagnosis. Gas gangrene can be caused by a variety of clostridial species, such as *C. septicum*, *C. novyi*, *C. perfringens* type A and *C. sordellii*. Among them *C. septicum,* the causative agent of malignant oedema is probably most encountered and widespread. It has been reported from ruminants, horses, pigs, elephants and avian species (Odani et al., 2009, Sasaki et al., 2001b, Murphy, 1980, Almeida e Macêdo et al., 2013, Rahman et al., 2009). The organisms involved in gas gangrene may influence the symptoms and clinical course. The disease presentation is associated with an array of potent exotoxins released by these organisms into the host tissues (Carter et al., 2014, Stevens et al., 2012). Gas gangrene usually occurs sporadically whereas disease outbreaks are mostly associated with injection of contaminated products for vaccination or other medical interventions (Morris et al., 2002a). In contrast to blackleg in bovines, gas gangrene is caused by wound contamination with spores or vegetative forms of one or more of the species (Silva et al., 2016). The initial symptoms appear after 12 to 36 hours with painful swellings at the point of infection, high fever, depression, breathing difficulties followed by convulsions and death occurring within 1 to 2 days in the majority cases.

Clostridium chauvoei is cultivated in strict anaerobic environments and characteristic colonies appear after 24-48 hours on blood agar plates. Differentiation from *C. septicum* is necessary as both pathogens are related phenotypically and are often present in clinical samples. Microscopically *C. chauvoei* cells are shorter and smaller than *C. septicum* cells. Colonies of *C. chauvoei* on blood agar are usually 2 to 4 mm in diameter, slightly raised, whitish-grey, and with a glossy surface. They are usually transparent or semi-transparent if examined at day one of incubation. Most colonies are circular and may be surrounded by a narrow zone of haemolysis. *C. chauvoei* is positive for sucrose fermentation, whereas *C. septicum* is negative. *C. chauvoei* cells spread less on blood agar plates and are more fastidious when compared to cells of *C. septicum* (Hagan et al., 1988, Hatheway, 1990).

The Fluorescent Antibody Technique (FAT) is a valuable tool for identifying and differentiating *C. chauvoei* and other clostridia such as *C. septicum* from tissue and culture smears as well as from tissue sections (Batty and Walker, 1963). FAT still has its place in diagnostics and could identify *C. chauvoei* and *C. novyi* on tissue smears taken during necropsy of a merino lamb affected with malignant oedema (Morris et al., 2002b). A case of malignant oedema caused by *C. chauvoei* was reported in a horse with a severe clinical course where first clinical signs occurred between 24 and 48 hours post infection (Almeida e Macêdo et al., 2013). A recent study carried out in order to investigate the pathogen prevalence in tissue samples with clinical and pathological evidence for clostridial myonecrosis identified *C. chauvoei* as the most frequently detected agent in both single and co-infections. The authors concluded that the high prevalence of *C. chauvoei* is attributed to its involvement in two forms of disease: blackleg and malignant oedema (Pires et al., 2017a).

Few studies have applied indirect fluorescent antibody testing and Enzyme Linked Immunosorbent Assay (ELISA) for specific detection of *C. chauvoei* and *C. septicum* antigens (Hamaoka and Terakado, 1994, Usharani et al., 2015). PCR identification has also been employed for specific detection of the pathogen as it is more reliable, fast, and easy to perform. Conventional PCR assays targeting the 16S rRNA gene, 16S-23S rRNA gene spacer region, or flagellin gene were reported by various workers (Sasaki et al., 2000a, Kojima et al., 2001, Sasaki et al., 2000b, Sasaki et al., 2002a, Sasaki et al., 2002b, Sasaki et al., 2001a). Recently developed real-time PCR protocols targeting the TPI gene, spo0A and 16s rRNA gene can also be used for pathogen detection (Garofolo et al., 2011, Lange et al., 2010, Halm et al., 2010).

The disease can be prevented in animals by vaccination. Bacterins are the commonly used vaccine candidates i.e. formalin-inactivated cultures and include both bacterial cells and culture supernatant. Polyvalent formulations are available which also include different species such as *C. novyi*, *C. septicum* and *C. sordellii*, the causative agents of gas gangrene in animals. Initial vaccination against blackleg is recommended at 2 months of age, followed by a booster after four to six weeks and later followed by half yearly or annual boosters depending on the disease prevalence (Abreu et al., 2016).

1.2.3 Bacterial genomics and phylogeny

Whole genome sequencing was completed first for *Haemophilus influenzae* (Fleischmann et al., 1995) and *Mycoplasma genitalium* (Fraser et al., 1995). Since then, sequencing of genomes of various bacterial species has been carried out and is still increasing in number and diversity with respect to genus and species. These genome sequence data provided novel insights into bacterial diversity, population structure, operons, mobile genetic elements and horizontal gene transfers (Binnewies et al., 2006). Genome sequencing is a powerful tool to get insights into bacterial evolution and pathogenesis and is more comprehensive than single gene or multigene sequence based approaches. Comparisons of bacterial genome sequence information in both interspecies and intraspecies application also helps understanding bacterial diversity and population characteristics. Today researchers are also exploiting the applicability of whole genome sequencing in clinical microbiology such as testing for

antibiotic resistance, detecting virulence determinants, outbreak detection and surveillance (Didelot et al., 2012).

1.2.3.1 Sequencing platforms

Sequencing technologies widely used for bacterial genome sequencing are pyrosequencing, reversible end sequencing and semiconductor sequencing. Their corresponding platforms are Roche GS FLX, Illumina Hiseq/Miseq and the Ion Personal Genome Machine (PGM), respectively (Bentley et al., 2008, Rothberg et al., 2011, Ronaghi et al., 1996). The general workflow for all platforms starts with a library preparation, which involves shearing of template DNA into fragments of suitable size, followed by end polishing to generate blunt end fragments and finally ligating suitable adaptor sequences to these fragments appropriate for each sequencing platform. Mechanical and enzymatic fragmentation are the widely used methods for fragmentation (Buermans and den Dunnen, 2014). Tagmentation is a novel transposase-based approach, whereby fragmentation of DNA and sequence tag incorporation occurs in one step (Adey and Shendure, 2012). All three sequencing platforms rely on sequencing by synthesis. The method employed for generation of clonally clustered amplicons of each library fragment is carried out to achieve a higher signal-to-noise ratio (Buermans and den Dunnen, 2014) by using different approaches such as emulsion PCR (Dressman et al., 2003) or bridge PCR (Adessi et al., 2000, Fedurco et al., 2006). In the PGM and FLX platforms, the clonal sequencing features are generated by emulsion PCR, whereas Illumina platforms employ bridge PCR for the same purpose (Buermans and den Dunnen, 2014).

GS FLX/454 uses pyrosequencing where the amplicon bearing beads are pre-incubated with *Bacillus stearothermophilus* (*Bst*) polymerase and single-stranded binding protein and then deposited onto a microfabricated array of picoliter-scale wells where only one bead will fit per well. The pyrophosphate ion released during nucleotide incorporation is detected using an enzyme cascade system (ATP sulfurylase and luciferase) which release a burst of light which is visualized using a charge coupled device-based signal detection system (Ronaghi et al., 1996, Shendure and Ji, 2008). The Ion Torrent platform relies on semiconductor sequencing technology. The sequence templates are generated on a bead or sphere via emulsion PCR, so that each reaction vesicle contains one library and all the reagents for amplification. The specificity of the library molecule is maintained via two complementary adaptors for the library fragments where one is present in the solution and the other bound to the sphere. The spheres containing amplified DNA are selected and deposited into a sequencing chip. The chip consists of a flow compartment and solid state pH sensor micro-arrayed wells. The release of an H+ during extension of each nucleotide is detected as a change in the pH within the sensor wells (Rothberg et al., 2011, Buermans and den Dunnen, 2014). Since both, the 454 and Ion Torrent sequencing platforms detect pyrophosphate or H+, respectively, but not the type of nucleotide, only a single kind of dNTP is added in each cycle in a predefined order (Buermans and den Dunnen, 2014).

The Illumina platform uses bridge amplification for clonal amplicon generation and employs the sequencing by synthesis (SBS) approach which typically takes place in a flow cell. Here forward and reverse oligonucleotides and one with a cleavable site (complementary to the

adapter sequences present in the library), are attached to cover the surface of the flow cell lanes. The library is loaded into the flow cell in a denatured state and hence gets hybridized to the oligos on the flow cell surface. These oligonucleotides act as primers to form an initial copy of the individual sequencing template molecule whereas the initial library molecules are removed. These flow cell-attached copied fragments are used to generate a cluster of identical template molecules using isothermal amplification. During the repeated cycles of denaturation, annealing and extension steps, the 3′ end of the copied molecules can hybridize to the complementary oligos within the flow cell resulting in the formation of a bridge structure. In the final step one strand of the dsDNA fragment is removed and all 3′ ends are blocked with ddNTP in order to prevent the open 3′ ends to act as sequencing primer sites on adjacent library molecules (Bentley et al., 2008, Buermans and den Dunnen, 2014). The Illumina sequencing platform uses a mixture of fluorescently labelled reversible terminators in each sequencing cycle. After imaging the fluorescent dye is cleaved off, the reversible terminator is deactivated and the ends are free for the next incorporation cycle (Buermans and den Dunnen, 2014).

The Single Molecule Real Time (SMRT) sequencing technology developed by Pacific Biosciences (PacBio) works differently from the above mentioned platforms. The library prepared from the template DNA does not require DNA amplification prior to sequencing and the adaptors used have a hairpin structure. The sequencing reaction takes place in zero-mode waveguide wells which are small reaction wells containing template, primer and polymerase. Here the fluorescent dye-labelled nucleotides are continuously added to a growing DNA strand and with the help of the zero-mode waveguide detector; continuous imaging of the labelled nucleotides is carried out. Here the technology differs from other technologies for its continuity as opposed to other platforms where the sequencing is carried out in interrupted cycles of extension and imaging (Buermans and den Dunnen, 2014). The Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing technology generates very long reads capable of resolving long repeat regions (Eid et al., 2009, Liao et al., 2015).

1.2.3.2 Genome assembly and annotation

Genome assembly is the process of aligning and merging short DNA sequences into longer ones in order to reconstruct the original sequence. The sequences are initially filtered according to the quality of the reads, and then overlapped to generate a contiguous sequence. Most of the assemblers employ either an overlap-layout-consensus (OLC) or a de Bruijn graph (DBG) assembly strategy (Flicek and Birney, 2009, Li et al., 2012). The OLC method generally works by finding overlaps among all the reads followed by generation of a layout of all reads and overlaps information on a graph to finally infer the consensus sequence. DBG works by chopping reads into much shorter k-mers and then using all k-mers to form a DBG to infer the genome sequence on the DBG. The OLC algorithm is optimal for low-coverage long reads, whereas the DBG algorithm is suitable for high-coverage short reads, but the assembly results can still vary with respect to genomes and sequencing technologies (Li et al., 2012). Bacterial genome assembly results are subjected to large variations, determined by several genetic factors such as the genome size, complexity, repeat elements, etc. (Hunt et al., 2013). Additional tools can also be applied to evaluate the best assembly options with or

without a high quality reference genome. These are available for several genome assemblers designed for short reads generated from bench top assemblers (Magoc et al., 2013, Hunt et al., 2013, Jünemann et al., 2014). Recently, several post assembly pipelines have been employed in bacterial genome assembly, enabling improvements such as correction of nucleotides, gap closing, joining overlapping contigs etc. (Swain et al., 2012). Hierarchical Genome Assembly Process (HGAP), a recently developed non-hybrid assembly process when applied with SMRT DNA sequencing was able to finish bacterial genomes with more than 99.999% accuracy (Liao et al., 2015, Chin et al., 2013).

Genome annotation is the process of attributing information called features to biologically important regions of the genome. The most commonly annotated features are genes including protein coding sequences (CDS), ribosomal RNA (rRNA) and transfer RNA (tRNA). Additional information on genome components such as operons, genomic islands, prophage, and clustered regularly interspaced short palindromic repeats (CRISPRs) can be added to genomes (Gary Van Domselaar, 2014). Gene Ontology (GO) based annotation follows a different approach, where every gene product is annotated according to three aspects which are based on molecular function, biological process and cellular composition (Giglio et al., 2009).

1.2.3.3 Comparative genomics and phylogeny

Bacterial species have been described within the concept of the pan-genome which comprises the core genome (genes that are present in every strain) and the dispensable genome (genes that are absent in one or more strains) (Tettelin et al., 2005). The core genome is usually represented by essential genes for basic cellular functions and the dispensable or accessory genome is believed to have an essential role in the genomic variation which may contribute to pathogenicity, drug resistance, and stress responses. Additionally, the dispensable genome may also increase the adaptability of pathogens to particular environmental conditions or hosts (Medini et al., 2005). Bacterial species can vary with respect to either having a limited or a large accessory genome. *Bacillus anthracis* is a species possessing a closed pan-genome. It is a spore forming pathogen, assumed to have very limited contact with other bacterial species in its vegetative phase and hence does not show any significant genetic variability. On the other hand, *Escherichia coli* harbour a large accessory genome correlating with the environment the species is maintained. *C. botulinum* is another example species for variable genomes capable of lateral gene movements as the Botulinum Neurotoxin (BoNT) gene has jumped between quite distantly related *Clostridium* strains (Segerman, 2012).

Evolution and population dynamics of bacterial species are new areas of research getting and requiring greater attention. Comparative genomics and phylogenetic relatedness can be explored using different approaches such as whole genome multilocus sequence typing (wgMLST), single nucleotide polymorphism (SNP) based or k-mer based approaches. Indeed every approach has its advantages and disadvantages (Klemm and Dougan, 2016). Whole-genome MLST relies on gene-by-gene analysis, using a curated database to assign an allele designation to each gene (Maiden et al., 2013). A SNP-based approach uses reads aligned to a closely related reference genome and is a helpful tool for discriminating closely related strains which differ only within few SNPs. An alignment based approach can also be employed when

a reference strain genome is unavailable or the study involves diverse strains by creating a core genome alignment (Page et al., 2015). Multiple genome alignment of bacterial genome sequences supports the understanding of genome evolution, the identification of recombination events and the ancestral genome (Darling et al., 2004, Darling et al., 2010, Angiuoli and Salzberg, 2011, Treangen et al., 2014). On the other hand k-mer based tools utilize an alignment free method relaying on relatedness among a set of sequences, based on the number or extent of k-mers (short sub-sequences of a fixed length k) which are shared among the compared genomes (Bernard et al., 2016). K-mer based tools are more robust and less biased to large genome variations, but are found to be more error prone, sensitive to sequence divergence and to the presence of incomplete sequence data (Chan et al., 2014).

Molecular phylogeny reconstruction is carried out from the sequence data using either a distance-based or character-based method. Distance matrix methods are calculating the distance between every pair of sequences. The resulting distance matrix is used for tree reconstruction such as the neighbour joining method, the most widely used distance matrix method (Saitou and Nei, 1987, Yang and Rannala, 2012). The method is a cluster algorithm and starts with a star tree joining together a pair of taxa based on the taxon distances until a fully resolved tree is obtained. The taxa to be joined are chosen in order to minimize an estimate of tree length. On the other hand, character-based methods include maximum parsimony, maximum likelihood and Bayesian inference. These approaches differ from the distance based method in as much as all sequences in an alignment are compared simultaneously while considering one character at a time to calculate a score for each tree (Yang and Rannala, 2012). The maximum parsimony method minimizes the number of changes on a phylogenetic tree by assigning character states to interior nodes on the tree where the character (or site) length is the minimum length for that site and the tree score is the sum of character length over all sites.

Maximum likelihood (ML) is a statistical method for the estimation of unknown parameters in a model. The method relies on two optimization steps, one for branch lengths to calculate the tree score for each candidate tree and the other on a search in the tree space for the maximum likelihood tree. Calculation of likelihood on a given tree can be carried out using any substitution model (Yang and Rannala, 2012, Whelan et al., 2001). Maximum likelihood is superior to distance or parsimony methods as it derives insights in the process of sequence evolution. The methods main drawback is that it is highly computationally demanding and has potentially poor statistical properties if the model is misspecified (Yang and Rannala, 2012, Whelan et al., 2001). Bayesian inference is a general methodology of statistical inference and Bayesian inference of phylogeny combines the prior probability of a phylogeny with the tree likelihood to produce a posterior probability distribution on trees. The best estimate of the phylogeny can be selected as the tree with the highest posterior probability. The method differs from maximum likelihood calculations as the topologies and branch lengths are not treated as parameters as in ML methods, but as random variables. Bayesian inference of phylogeny is computationally faster than the maximum likelihood method (Douady et al., 2003, Yang and Rannala, 2012). Recently, several tools have been developed and applied for recognizing and stripping out putative recombination sites before identifying phylogeny. Bacterial recombination events can cause different sites of the genome to have different inheritance (Croucher et al., 2009, Marttinen et al., 2012).

1.2.4 Genome composition, comparative genomics and typing options: genus *Clostridium*

1.2.4.1 Genome sequencing and composition

Genome sequencing and characterization of different *Clostridium* species have been carried out by various researchers (Nolling et al., 2001, Bruggemann et al., 2003, Shimizu et al., 2002, Sebaihia et al., 2007). The members of the genus *Clostridium* were shown to be highly heterogeneous and a limited synteny based on the genome sequence data and comparative genomic studies among *C. botulinum*, *C. acetobutylicum*, *C. perfringens* strain 13, *C. tetani*, and *C. difficile* was found. Only 568 *C. botulinum* CDSs (16%) were found to be shared with the other *Clostridium* spp. Variations were also observed with respect to genome sizes (2.7 Mb for *C. tetani* to 4.2 Mb for *C. difficle*), GC content (ranging from 28.24% to 30.93%), whereas the numbers of coding genes, rRNA operons and tRNA varied between 2368 and 3774, 6 and 11 and between 54 and 94, respectively, among the species (Sebaihia et al., 2007). With an array of diseases provoked by the members of the genus, they also exhibit variation with respect to genome sizes and genome composition within a species. Several species within the genus *Clostridium* have large plasmids and plasmid-encoded toxin genes as reported for *C. perfringens, C. sordellii,* and *C. botulinum* (Freedman et al., 2015, Couchman et al., 2015, Marshall et al., 2007). Endospore formation is unique to the phylum *Firmicutes* which includes the genus *Clostridium* (Traag et al., 2013). Spo0A, a transcriptional factor, plays a central role in the sporulation process of bacteria of the genera *Clostridium* and *Bacillus*. There is also evidence showing that Spo0A is involved in regulating various metabolic and virulence factors such as toxins in the genus *Clostridium* (Pettit et al., 2014, Paredes-Sabja et al., 2011). Bacteriophages are also identified in pathogenic bacteria which contribute to their evolution. Temperate phages can regulate toxin production and release such as the clostridial neurotoxins produced by *C. botulinum* and *C. tetani* (Fortier and Sekulovic, 2013). Other important genetic elements are the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) elements which were recognized in bacteria protected against bacteriophages. The CRISPR and CRISPR associated genes (Cas) found near the CRISPR elements together form the CRISPR-Cas system (Barrangou et al., 2007). The defence mechanism is carried out in three successive stages of adaption (insertion of a new spacer derived from the phage into the CRISPR locus), expression (protein expression of the *cas* genes and transcription of the CRISPR into a long precursor CRISPR RNA (pre-crRNA) and finally to mature crRNA by the Cas proteins and accessory factors) and interference (destruction of target nucleic acid by crRNA and Cas proteins) (Rath et al., 2015). CRISPR elements have been identified in important pathogenic species of the genus *Clostridium* e.g. a pathogenic *C. perfringens* type A strain isolated from bovines (Nowell et al., 2012) and a *C. botulinum* type III strain (BKT015925) of poultry origin. They were present on the prophage (p1) and on a plasmid (p2), respectively (Skarin et al., 2011). A variety of species within the genus carry several antibiotic resistance genes as part of their flexible genetic elements which can be transferred to other bacteria as found for *C. perfringens* (van Schaik, 2015).

1.2.4.2 Comparative genome studies and evolution

A huge genetic variability has been observed among isolates of *C. difficile*, *C. botulinum* and *C. perfringens* based on comparative genomic studies (Myers et al., 2006, Woudstra et al., 2016, Knight et al., 2015). Strain heterogeneity and evolution was studied in more detail for *C. difficile*. The major elements determining the strain variability in the case of *C. difficile* are transposons, bacteriophages, homologous recombination and natural selection (Knight et al., 2015). *C. difficile* harbours some transposons (Tns) that are mobilizable by mechanisms of the host, whilst others are self-transmissible known as conjugative transposons (CTns) and integrative elements (ICEs). These transposons can lead to heritable changes such as acquisition of possibly advantageous genes but may also lead to gene disruption. Transposons of *C. difficle* have been described as a major contributor to antimicrobial resistance unlike plasmids as observed in other pathogens (Sebaihia et al., 2006, Brouwer et al., 2012, Knight et al., 2015). Besides transposons, *C. difficle* genomes also harbour phages which play an important role in its evolution. The phages found in the genome mostly belong to the *Siphoviridae* and *Myoviridae* families. There is also increasing evidence showing the role of phages in *C. difficile* pathogenesis such as the regulation of toxin production (Hargreaves et al., 2014, Knight et al., 2015). Homologous recombination has also been shown to play an important role in the gastrointestinal adaptation and the virulence potential of *C. difficile*. The pathogenicity locus (PaLoc) of *C. difficle* which contains the major virulence factors toxin A and toxin B is found to be present in all toxigenic strains but is absent in nontoxigenic strains (Knight et al., 2015). Studies have also shown that the PaLoc is transferred between *C. difficile* strains by a conjugation-like mechanism (Brouwer et al., 2013). The other important factor which influences the divergence of *C. difficle* is the process of natural selection. A study conducted to estimate the relative ratio (*dN*/*dS*) of nonsynonymous substitutions (*dN*) and synonymous substitutions (*dS*) in the core CDSs for inferring the signatures of selection showed strong purifying selection (ratio significantly lower than 1) for deeply diverging lineages whereas the purifying selection indicated a neutral selection pressure (ratio close to 1) for recently diverged sequences. This was ascribed to a lack of time for selection to act, or the nucleotide substitutions within the species represent segregating polymorphisms rather than fixed differences (He et al., 2010, Knight et al., 2015).

A comparative genomic analysis of 40 completely sequenced clostridia genomes, including species involved in biomass degradation and disease showed a low number of core genes and a larger number of strain specific genes. The study also showed biomass degraders tend to have larger genome sizes and pan-genomes as compared to pathogenic ones (Zhou et al., 2014). A comparative study based on genome sequence data involving *C. botulinum* group I and group II genomes revealed limited genetic variabilities among group I genomes and larger variability with inversions of large genomic regions in group II genomes (Carter and Peck, 2015). A recent study of *C. botulinum* group II type E genomes based on seven closely related neuro-toxigenic subtypes (E1, E3, and E10) revealed that the strains harbour genes associated with plasmid mobility via conjugation. Few plasmid subtypes also carried a CRISPR element positioned adjacent to the neurotoxin gene cluster which may act as a hotspot for insertion of the neurotoxin gene cluster as discovered in the chromosome of *C. botulinum* (Carter et al., 2016). A comparative genomics study of four *C. perfringens* genomes showed an assignment of 90% of the genes as core genes and their high relatedness

(Ng and Lin, 2014). Studies on the *C. tetani* genome showed only few mobile elements and most of these genes were identified to be non-functional because of insertions, deletions and point mutations (Alam et al., 2010). Furthermore, the recently published genome of *C. septicum* was reported to have a 32 kb plasmid (Benamar et al., 2016). Genomic analysis of a collection of *C. sordellii* strains from diverse geographical locations revealed the presence of four clades. The four assigned clades did not reveal any significant relatedness to host, clinical presentation or geographical origin (Couchman et al., 2015).

Spore-forming bacteria can move to a stage of inactivity and persistence, and studies proved they can remain in the dormant stage for even millions of years keeping the ability to revert back to the vegetative stage (Cano and Borucki, 1995). A recent study with 200 *Firmicutes* species showed that spore-forming bacteria have longer generation times and evolve more slowly. Sporulation significantly reduces the genome-wide spontaneous DNA mutation and protein evolutionary rates (Weller and Wu, 2015).

1.2.4.3 Genotyping options: genus *Clostridium*

Various genetic typing methods have been employed for the characterization of strains belonging to the genus *Clostridium*. These studies involved various techniques such as PCR-ribotyping, Amplified Fragment Length Polymorphism (AFLP), Randomly Amplified Polymorphic DNA (RAPD), Pulse Field Gel Electrophoresis (PFGE), Multilocus Sequence Typing (MLST), Multiple Locus Variable Number Tandem Repeat Analysis (MLVA), Restriction Enzyme (RE) analysis and microarray techniques (Keto-Timonen et al., 2006, Anniballi et al., 2016, Luquez et al., 2015). Multilocus Sequence Typing (MLST) is a sequence based approach which makes comparisons across the laboratories and across time easy (Maiden et al., 2013). A recent study utilized the applicability of diversity of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR associated genes (*cas*) to understand the diversity and to create a phylogeny for *C. difficle* strains (Andersen et al., 2016).

1.2.5 Genome sequence: *Clostridium chauvoei*

Genome sequence information for the pathogen was first available for a virulent Swiss isolate (JF4335) exhibiting a 2.8Mb genome and 5 kb plasmid. The genome sequencing was carried out using PacBio and Illumina sequencing technology and the draft assembly was represented by 12 contigs for the chromosome (Falquet et al., 2013). A recent study based on genome sequence data of 20 *C. chauvoei* strains revealed the applicability of unique CRISPR spacer sequence motives for depicting genetic diversity (Rychener et al., 2017). The genome sequence analysis of strain JF4335 identified prophage elements, several primary virulence factors, proteases and antibiotic resistance genes. Potential virulence factors identified include 4 haemolysins (a haemolysin belonging to the haemolysin III-superfamily; a haemolysin A; a haemolysin of the XhlA type and the haemolytic leukocidin CctA), sialidase (NanA), hyaluronidases, potential patatine phospholipases (PPARs), two genes encoding potential collagen binding proteins and homologues to internalin A protein. Twenty flagellar biosynthesis genes were found on the chromosome of this strain. The described metabolic pathways encompass glycolysis/gluconeogenesis, sugar metabolism, purine and pyrimidine

metabolisms, as well as many amino acid metabolisms, but absence of many genes of the citric acid cycle and complete or partial lacking of few amino acid metabolisms was also noticed (Frey and Falquet, 2015).

Chapter 2: Materials and methods

2.1 Bacterial strains

Bacterial strains used in the present study were maintained in the culture collection of the Institute of Bacterial Infections and Zoonoses (IBIZ), Friedrich-Loeffler-Institut (FLI), Jena, Germany (Table 1). *Clostridium* (*C.*) *chauvoei* strains included in the study were mainly of European origin including Germany, Austria, Switzerland and Italy (only DNA), whereas few strains where of unknown origin and were obtained from miprolab GmbH, Göttingen, from the former strain collection of the Institute for Tropical Animal Health, Georg-August-University, Göttingen and from the strain collection of the Zentrales Institut des Sanitätsdienstes der Bundeswehr Kiel, Abteilung II, Veterinärmedizin. One strain was from Canada and another strain investigated was strain NCTC 08361 recovered from sheep in South Africa. The study also included the *C. chauvoei* type strain DSM 7534^T (Table 1) and other published genome sequences/reads of *C. chauvoei* and related species (Table 2). The number of strains from different geographical regions from Germany was Lower Saxony (23), Bavaria (4), Mecklenburg- Western Pomerania (3), North Rhine-Westphalia (3), Schleswig-Holstein (3) and Baden-Württemberg (2). The strains from Austria belonged to two regions namely Styria (6) and Tyrol (4). Other strains from Europe included in the study were two strains from Switzerland and one DNA from a strain isolated in Perugia, Italy. Some of the strains were isolated from the same animal (11S0315, 11S0316 and 12S0471). Few strains were also recovered from different animals involved in the same outbreak (BS79-01 and BS 80-01, 12S0467 and 12S0468, S0040-08 and S0041-08, S0021-10 and S0022-10, 13S0851 and 13S0854, BS106-02 and BS107-02, BS91-02 and BS109-02). In some cases strains coming from the same farms but with longitudinal time gaps were also included (S0121-09 (year 2009) and S0021-10, S0022-10 (year 2010)), (BS106-02, BS107-02 (year 2000) and BS91-02, BS109-02 (year 2001)), (S0260-09 (2009) and 12S0467, 12S0468 (2011)), (12S0464 (2010) and 11S0315, 11S0316, 12S0471 (2011)).

Table 1: Details of *Clostridium* strains used in this study showing number of strains, origin of strains/sample numbers, FLI strain designation, year, location and country of isolation.

Number	Origin of strain/sample number	FLI strain designation	Year (case)	Region	Country
Clostridium chauvoei					
Genome sequencing for genome completion					
1	DI201107460 St.3	12S0467	2011	Lower Saxony	Germany
2	DSM 7528^T (ATCC 100092^T, NCIMB 10665^T)	11S0147 (DSM 7528^T)	Unknown		Unknown
Genome sequencing for phylogenomics					
3	DI201110064 St.3	12S0468	2011	Lower Saxony	Germany
4	DI200103838	BS91-02	2001	Lower Saxony	Germany
5	DI200202875	BS94-02	2002	Lower Saxony	Germany
6	DI200104624	BS92-02	2001	Lower Saxony	Germany

7	DI 3055/94	BS104-02	1994	Lower Saxony	Germany
8	DI 4462/98	BS105-02	1998	Lower Saxony	Germany
9	DI200006624	BS106-02	2000	Lower Saxony	Germany
10	PA20000444	BS90-02	2000	Lower Saxony	Germany
11	DI200104694	BS109-02	2001	Lower Saxony	Germany
12	VI200928935	S0260-09	2009	Lower Saxony	Germany
13	DI200007762	BS107-02	2000	Lower Saxony	Germany
14	DI200103561	BS108-02	2001	Lower Saxony	Germany
15	DI201004872 St.2	12S0464	2010	Lower Saxony	Germany
16	DI201007522 St.3	12S0465	2010	Lower Saxony	Germany
17	DI201007544 St.4	12S0466	2010	Lower Saxony	Germany
18	DI201111103 St.2	12S0469	2011	Lower Saxony	Germany
19	DI201111234 St.3	12S0470	2011	Lower Saxony	Germany
20	DI201114061 St.4	12S0471	2011	Lower Saxony	Germany
21	DI201206789	12S0472	2012	Lower Saxony	Germany
22	none (Osterholz-Scharmbeck)	S0162-10	2010	Lower Saxony	Germany
23	DI201114061	11S0315	2011	Lower Saxony	Germany
24	DI201114061	11S0316	2011	Lower Saxony	Germany
25	Isolat 871	11S0318	2004	Mecklenburg-Western Pomerania	Germany
26	1999-RD 1268	BS80-01	1999	Mecklenburg-Western Pomerania	Germany
27	1999-RD 1267	BS79-01	1999	Mecklenburg-Western Pomerania	Germany
28	A13087497-1, xs1101/10	13S0851	2013	Baden-Württemberg	Germany
29	A13088531-1, xs1106/4	13S0854	2013	Baden-Württemberg	Germany
30	436 98 200	S0021-10	2010	North Rhine-Westphalia	Germany
31	D 115/09	S0121-09	2009	North Rhine-Westphalia	Germany
32	864 29 582	S0022-10	2010	North Rhine-Westphalia	Germany
33	P865	S0008-10	2009	Schleswig-Holstein	Germany
34	P781/1	S0040-08	2008	Schleswig-Holstein	Germany
35	P781/2	S0041-08	2008	Schleswig-Holstein	Germany
36	AZ:15-0168810-001-01	15S0088	2015	Bavaria	Germany
37	AZ:15-0219179-001-01	15S0089	2015	Bavaria	Germany

38	AZ:16-0122541-001-01	16S0579	2016	Bavaria	Germany
39	AZ:15-0168810-001-01	15S0008	2015	Bavaria	Germany
40	08064498	S0099-08	2008	Styria	Austria
41	08070153-02	S0100-08	2008	Styria	Austria
42	08075408	S0101-08	2008	Styria	Austria
43	08097017-001	S0105-08	2008	Styria	Austria
44	101/07 field isolate	S0013-08	2007	Tyrol	Austria
45	102/07 field isolate	S0014-08	2007	Tyrol	Austria
46	103/07 field isolate	S0015-08	2007	Tyrol	Austria
47	08065412-001	S0098-08	2008	Tyrol	Austria
48	08080095	S0102-08	2008	Styria	Austria
49	08084503	S0103-08	2008	Styria	Austria
50	JF 1866 (IVB 263)	BS169-00	Unknown	Unknown	Switzerland
51	JF 1869 (IVB A105)	BS171-00	Unknown	Unknown	Switzerland
52	NCTC 08361, (CN 657) Wellcome Trust Collection in 1951, Original Strain Reference: A 10	15S0023	Before 1951	Unknown	South Africa
53	Perugia	16S0574	Unknown	Perugia	Italy
54	F198	16S0578	2006	Unknown	Canada
55	628 (CN3601 Wellcome)	S0133-09	Unknown	Unknown	Unknown
56	622 (CN6299 Wellcome)	S0132-09	Unknown	Unknown	Unknown
57	620 (CN5097 Wellcome)	S0131-09	Unknown	Unknown	Unknown
58	1076 (ATCC 100092^{T}, NCIMB 10665^{T}, DSM 7528^{T})	S0136-09	Unknown	Unknown	Unknown
59	1023 (NCTC 8070, ATCC 19399; CN 690; G 1)	S0134-09	Before 1951	Unknown	Unknown
60	E8	BC93-06	Unknown	Unknown	Unknown
61	E14 (DSM 7528^{T}, ATCC 100092^{T}, NCIMB 10665^{T})	BC97-06	Unknown	Unknown	Unknown
62	H6	BC103-06	Unknown	Unknown	Unknown
63	E9	BC138-06	Unknown	Unknown	Unknown

Table 2: Details of published sequence data used for this study showing number of strains, species, strain designation, type of data and accession number.

Assembly/reads of genomes retrieved from NCBI/ENA^				
Number	Species	Strain designation	Data retrived	NCBI/ENA accession number
1	*C**. *chauvoei*	JF4335	Assembly	PRJEB3959, NZ_LT799839.1
2	*C. chauvoei*	DRS014052	Reads	DRR015944
3	*C. septicum*	CSUR P1044	Assembly	PRJEB146921
4	*C. botulinum* C/D	BKT2873	Assembly	PRJNA233460
5	*C. botulinum* C/D	BKT12695	Assembly	PRJNA233471
6	*C. botulinum* C	str.Eklund	Assembly	PRJNA20017
7	*C. botulinum* A	Hall	Assembly	NC_009698
8	*C. botulinum* E3	str. Alaska E43	Assembly	NC_010723
9	*C. botulinum* E1	str. BoNT E Beluga	Assembly	PRJNA29861
10	*C. botulinum* B	str. Eklund 17B	Assembly	NC_010674/NC_010680
11	*C. botulinum* A	ATCC 3502	Assembly	NC_009495/NC_009496
12	*C. acetobutylicum*	DSM 1731	Assembly	NC_015687
13	*C. acetobutylicum*	ATCC 824	Assembly	NC_003030
14	*C. acetobutylicum*	EA 2018	Assembly	NC_017295
15	*C. tetani*	E88	Assembly	NC_004557/NC_004565
16	*C. butyricum*	KNU-L09	Assembly	NZ_CP013252/NZ_CP013489
17	*C. butyricum*	5521	Assembly	NZ_ABDT00000000
18	*C. butyricum*	E4 BoNT E BL5262	Assembly	NZ_ACOM00000000
19	*C. haemolyticum*	NCTC 8350	Assembly	NZ_JDSA00000000
20	*C. novyi* NT	NT	Assembly	NC_8593
21	*C. novyi* B	str. ATCC	Assembly	PRJNA233467
22	*C. sporogenes*	DSM 795	Assembly	NZ_CP011663
23	*C. beijerinckii*	ATCC 35702	Assembly	NZ_CP006777
24	*C. beijerinckii*	NCIMB 8052	Assembly	NC_009617
25	*C. perfringens*	ATCC 13124	Assembly	NC_008261
26	*C. perfringens*	str. 13	Assembly	NC_003366/NC_003042
27	*C. baratii*	str. Sullivan	Assembly	NZ_CP006905/ NZ_CP006906

* *Clostridium*, ^ National Centre for Biotechnology Information/European Nucleotide Archive

2.2 Genome sequencing for genome completion

2.2.1 Bacterial strains and DNA extraction

The bacterial strains used for Pacific Biosciences (PacBio) sequencing were a virulent strain obtained from a blackleg case in cattle from northern Germany in 2011 (12S0467) and the type strain (DSM 7528^T=ATCC 10092^T). For DNA extraction, isolates were cultured in 20 ml Selzer broth (Selzer et al., 1996) (tryptone - 30g, beef extract - 20g, glucose - 4g, L-cysteine hydrochloride - 1 g in 1000 ml H_2O, pH 7.2) at 37°C for 18 to 24 hours under anaerobic conditions. Genomic DNA was extracted using the Qiagen Genomic-tip 100/Q and genomic DNA buffer set (Qiagen, Germany) with minor modification i.e., 40 U of achromopeptidase (Frey et al., 2012) was included in the enzymatic lysis buffer and the DNA was incubated at 37°C until it was dissolved. DNA quality was examined by using a Qubit 2.0 fluorometer (Life Technologies, Germany) and by agarose gel electrophoresis. The DNA quantity was measured using a Nanodrop spectrometer (Thermo Fisher Scientific, USA). Species confirmation was carried out based on published primers specific for 16-23S rDNA spacer regions to detect and identify *C. chauvoei* and *C. septicum* (Sasaki et al., 2000b).

2.2.2 Pacific Biosciences (PacBio) sequencing and assembly

Genome sequencing of 12S0467 and DSM 7528^T was carried out by Single Molecule Real Time (SMRT) DNA sequencing (Eid et al., 2009) using PacBio RSII sequencer at GATC Biotech (Germany) employing 10 kb insert libraries prepared from genomic DNA. Genome assembly was carried out using HGAP algorithm version 3 (Chin et al., 2013) which was implemented in the PacBio SMRT portal version 2.3.0 at GATC Biotech (Germany). The genome sequence of DSM 7528^T was represented by one single contig with overlapping ends. The circularization of the received contig to a bacterial chromosome was carried out at GATC Biotech (Germany) using a protocol recommended by PacBio for merging and circularization (https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Circularizing-and trimming). The method involved the Minimus 2 software (part of AMOS version 3.1 (Sommer et al., 2007)), which removes the overlapping sequence and merges the ends. The output result represents the complete genome. The error correction of the final circularized genome was carried out with the RS_Resequencing protocol in SMRT portal using the raw reads. When the genome was represented by 2 contigs, merging of the contigs was carried out using Minimus 2 (Sommer et al., 2007) and circularization was carried out using either Circlator (Hunt et al., 2015) or was achieved from the overlapping ends using Geneious 9.0.5 (Kearse et al., 2012). Circlator was used to circularize the plasmid generated by HGAP (Chin et al., 2013) to identify multiple copies of the plasmid sequence which is often generated for plasmids less than 6 kb (Hunt et al., 2015). The final circular plasmid generated in Circlator (Hunt et al., 2015) was visualized in Artemis Comparison Tool (ACT) (Carver et al., 2005).

2.3 Genome sequencing for phylogenomics

2.3.1 Bacterial strains and genome sequencing

Strains (excluding DSM 7528^T and 12S0467) were cultured on sheep blood agar plates (Yeast Extract Cysteine Medium with Sheep Blood (Beerens Formulation) Oxoid, Germany) at 37^{o}C for 24 to 48 hours under anaerobic conditions in an anaerobic chamber MACS VA500 (Don Whitley Scientific, UK). The culture material from 1 to 2 plates was used for DNA extraction. Genomic DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen, Germany) or the Qiagen Genomic-tip 100/Q and genomic DNA buffer set (Qiagen, Germany) with slight modifications, i.e. 40 U of achromopeptidase in the enzymatic lysis buffer (Frey et al., 2012). Species confirmation was carried out based on published primers specific for the 16-23S rDNA spacer region for differentiating *C. chauvoei* and *C. septicum* (Sasaki et al., 2000b). Library preparation from genomic DNA was carried out using the Nextera^{TM} (Illumina, Netherlands) library preparation method. Genome sequencing was carried out using the MiSeq^{TM} System (Illumina, USA) paired end sequencing technology (2 × 300bp) by the Institute of Microbiology and Epizootics (IMT), Freie Universität, Berlin.

2.3.2 Genome assembly evaluation

Genome assembly evaluation was carried out using the QUality ASessment Tool- 4.1 (QUAST) (Gurevich et al., 2013) run in the Genome Assembly Gold-standard Evaluation (GAGE) (Salzberg et al., 2012) mode. The completed genome sequence of the field strain from Germany (12S0467) was used as the reference for the evaluation of the three strains originating from the same region. Three assemblers were chosen for assembly evaluation (CLC Genomics Workbench version 8.5.1 (Qiagen, Aarhus, and A/S), SPAdes 3.9.0 (Bankevich et al., 2012) and MaSuRCA (Zimin et al., 2013) for important parameters described in Genome Assembly Gold-standard Evaluation for Bacteria (GAGE-B) (Magoc et al., 2013). Prior to the assembly process, the Nextera^{TM} (Illumina, USA) adapters were trimmed from the reads. Genome assembly with CLC Genomics Workbench was carried out with the following parameters for removal of low quality sequences - limit (error probability) set to 0.01, removal of ambiguous nucleotides - maximal 2 nucleotides allowed, genome assembly - with read mapping and without scaffolding. SPAdes 3.9.0 (Bankevich et al., 2012) assembly was carried out with parameters specified for long Illumina reads, with mismatch correction turned on and coverage threshold cutoff determined automatically for SPAdes assembly. MaSuRCA (Zimin et al., 2013) assembly was carried out using parameters recommended for Illumina paired read assembly which is optimal for the assembly of bacterial genomes.

2.3.3 Genome assembly and post-assembly improvements

Genome assembly for all strains was performed using SPAdes (Bankevich et al., 2012) with the settings mentioned above, but reads were initially trimmed for quality scores using Sickle version 1.33 for base quality score above 20 (Joshi NA, 2011) before error correction with BayesHammer (Nikolenko et al., 2013), which is part of SPAdes assembler. Raw reads from a *C. chauvoei* strain, sequenced with the HiSeq^{TM}system (Illumina, USA) were retrieved from

the European Nucleotide Archive submitted by Gifu University School of Medicine Pathogenic Bacterial Genetic Resource Stock Centre (GMGC) and assembled with SPAdes (Bankevich et al., 2012). Post assembly improvements of the draft genomes were carried out using the post-assembly genome-improvement toolkit (PAGIT) (Swain et al., 2012), which utilizes contig reordering by algorithm-based automatic contiguation of assembled sequences (ABACAS) (Assefa et al., 2009) (carried out only for European strains using the data of the finished field strain from Germany), gap closure using iterative mapping and assembly for gap elimination (IMAGE) (Tsai et al., 2010) for 9 iterations and error correction using iterative correction of reference nucleotides (ICORN2) (Otto et al., 2010) for 5 iterations for all strains. Assembly statistics were generated using QUAST (Gurevich et al., 2013).

2.4 Genome annotation

Protein-coding sequences were predicted by Glimmer software version 3.0 (Delcher et al., 2007), ribosomal RNA genes and transfer RNA genes were detected using RNAmmer software version 1.2 and tRNAscan-SE, respectively, for the DSM 7528^T and 12S0467 strains (Lagesen et al., 2007, Lowe and Eddy, 1997). Genome annotation for the strains was performed using automated annotation pipelines such as the Rapid Annotation using the Subsystem Technology (RAST) server (Aziz et al., 2008, Overbeek et al., 2014) and Prokka 1.11 (Seemann, 2014). Genome annotations were additionally done with Gene Ontology (GO) terms using Blast2GO PRO software in CLC genomics workbench 8.5.1 and the results were summarized in General GO slim functional categories (Conesa et al., 2005, Conesa and Gotz, 2008, Götz et al., 2008, Gotz et al., 2011). BLASTP search was carried out against the non-redundant database and the GO terms associated with each BLAST hit were used for the annotation. All draft genomes of strains used for comparative genome studies were annotated using Prokka 1.11 annotation pipeline (Seemann, 2014). Genome annotation files, gene encoding protein and nucleotide fasta files were explored and managed using Geneious 9.0.5 (Kearse et al., 2012).

2.5 Genome content and composition

2.5.1 Origin of replication (*oriC*) and antibiotic resistance genes

The origin of replication (*oriC*) was identified using Ori-Finder (Gao and Zhang, 2008). Antibiotic resistance genes were identified using Resistance Gene Identifier 3.0.9 (RGI) with the Comprehensive Antibiotic Research Database (CARD) (McArthur et al., 2013).

2.5.2 Spore resistance, sporulation and germination genes

The genetic relatedness of important genes regulating sporulation and germination of *C. chauvoei* was compared to that of *C. septicum* (CSUR P1044), *C. perfringens* (ATCC 13124) and *C. botulinum* type A (ATCC 3502) based on BLASTP searches with the respective protein coding genes.

2.5.3 Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and prophage elements

CRISPR loci were searched for using the CRISPR Recognition tool (Bland et al., 2007) in Geneious 9.0.5. Unique CRISPR spacers were identified with nucmer 3.07 (Kurtz et al., 2004) using the settings as reported earlier for *C. chauvoei* (Rychener et al., 2017). CRISPR spacer sequences unique to all strains were displayed in form of a matrix based on all CRISPR spacers numbered for strain 12S0467 followed by other spacers identified in the respective genomes as described in earlier studies (Horvath et al., 2008). Prophage elements were identified using PHAge Search Tool (PHAST) (Zhou et al., 2011) or PHAge Search Tool - Enhanced Release (PHASTER) (Arndt et al., 2016).

2.5.4 Insertion Sequences (ISs) and Genomic islands (GIs)

IS annotation was carried out with the ISsaga program (insertion sequence semi-automatic genome annotation at http://issaga.biotoul.fr/ISsaga2/issaga_index.php (Varani et al., 2011). Genomic Islands where predicted using Island Viewer tool 4 (Bertelli et al., 2017).

2.6 Comparative genome analysis and visualization

2.6.1 Circular genome plot and locally collinear blocks

The circular plot of genomes was made using Artemis and DNAPlotter (Carver et al., 2009, Rutherford et al., 2000). Locally collinear blocks (LCBs) among (DSM 7528 and 12S0467) and (DSM 7528 and CSUR P1044) were created using progressiveMauve (Darling et al., 2010).

2.6.2 Phylogenetic analysis of the genus *Clostridium*

Gegenees software_v2.2.1 was utilized for carrying out the phylogenetic analysis (Ågren et al., 2012), with TBLASTX settings of 500/500 and heat maps were generated with a cutoff threshold of 35% for non-conserved genetic material. Phylogenetic trees (Neighbour Joining method) were created in Splits Tree 4 (Huson and Bryant, 2006) using the PHYLIP file exported from the Gegenees software. The phylogenomic study involved various species within the genus such as *C. chauvoei*, *C. septicum, C. botulinum* type C/D, *C. botulinum* type C, *C. botulinum* type A, *C. botulinum* type E3, *C. botulinum* type E1, *C. botulinum type* B, *C. acetobutylicum*, *C. tetani*, *C. butyricum*, *C. haemolyticum, C. novyi* NT, *C. novyi* B, *C. sporogenes, C. beijerinckii*, *C. perfringens* and *C. baratii.*

2.6.3 Orthologous gene identification

The orthologous gene identification was carried out using the Pan-genome Ortholog Clustering Tool 3.23 (panOCT) (Fouts et al., 2012) with default parameters for three *C. chauvoei* strains (DSM 7528^{T}, 12S0467and JF4335) and the same tool was also applied to determine the shared and unique genes among the three *C. chauvoei* genomes and the *C. septicum* (CSUR P1044) genome. The venn diagram depicting the core, shared and unique genes identified using panOCT was generated using venneuler version 1.1.0 (http://www.rforge.net/venneuler/).

2.7 Comparative genomics study of *C. chauvoei* strains

2.7.1 Core, accessory and pan-genome analysis

Clostridium chauvoei draft genome sequences of this study and available genome sequences (JF4335) of earlier studies were employed in the analysis. Pan-genome analysis was carried out with Prokka (Seemann, 2014) annotated genomes using Roary 3.8.2 (Page et al., 2015). The pan-genome visualization and multidimensional scaling plot was created using Fripan (http://drpowell.github.io/FriPan/). A pie chart showing the proportion of core, softcore, shell and cloud genes were created using roary_plots.py (https://sanger-pathogens.github.io/Roary/). Core and accessory genes with large size variations or with unsplit paralog genes (considered multiple alleles) were removed from the pan-genome using roProfle tool v 1.4.5 (https://github.com/cimendes/roProfile). The curves for the core, pan-genome (S0133-09, 16S0574 and same laboratory isolates of DSM 7528 (S0136-09 and BS97-06) were excluded)) were calculated and the core vs. pan-genome and new gene discovery plot were plotted as a function of the number of genomes added sequentially using the distance guide (DG) algorithm in the PanGP software (Zhao et al., 2014). PanGP uses power-law regression ($y = Ax^{B} + C$) to model the pan-genomes generated from all permutations, where y is the total number of gene families in the pan-genome, x is the number of genomes considered, A, B and C are fitting parameters. When $0 < B < 1$, the pan-genome should be considered open because it is an unrestrained function over the number of genomes. The exponential curve fit model, $y = Ae^{Bx} + C$ was used to fit the core genome. Here, y denotes core genome size, x denotes number of genomes; A, B and C are fitting parameters. The new gene discovery plot for pan-genome analysis of *C. chauvoei* genomes was carried out where the curve represents the least-squares fit for the function $y = Ax^{B}$, where y denotes the new genes, x denotes the number of genomes and A and B are the fitting parameters. The core and accessory genes were also annotated based on gene ontology terms using Blast2GO PRO software (Conesa et al., 2005, Conesa and Gotz, 2008, Götz et al., 2008, Gotz et al., 2011) and comparative analysis of phage composition and insertional elements were carried out using progressiveMavue (Darling et al., 2010).

2.7.2 Core genome alignment

The Parsnp tool of the Harvest suite (Treangen et al., 2014) was used to make the core genome alignment using -c flag to include all the genomes. The core genome alignment FASTA file was generated using Harvest tools (Treangen et al., 2014). The visualization of alignment FASTA file was carried out using Geneious 9.0.5 (Kearse et al., 2012).

2.7.3 Recombination analysis, population structure and phylogeny

Recombination analysis of the strains within the core genome of 2.4 Mb was carried out using Genealogies Unbiased By recomBinations In Nucleotide Sequences (Gubbins) (Croucher et al., 2015) and BratNextGen (Marttinen et al., 2012). BratNextGen was run with parameters specifying hyperparameter α set to 1 and 20 iterations of the recombination estimation algorithm. The statistical significance was estimated for 100 permutations of the algorithm, setting significance at $\alpha = 0.05$. Gubbins was run with default parameters. Approximate phylogenies for all strains were inferred from the recombination free post-filtered polymorphic sites exported from Gubbins by maximum-likelihood (ML) method using RAxML 8.2.8 (Stamatakis, 2014) with the general time reversible nucleotide substitution model (GTR-GAMMA) and bootstrap support from 500 iterations. The figure for the phylogenetic tree was generated using Interactive Tree Of Life version 3.4.3 (iTOL) (Letunic and Bork, 2016) showing the bootstrap support values at the nodes.

2.7.4 Pan-genome SNP analysis, phylogeny and clustering

Pan-genome SNP analysis independent of multiple alignment and independent of a reference genome was carried out using kSNP 3.0 (Gardner et al., 2015). kSNP3 identifies SNPs based on homologous stretches of nucleotides present in all genomes with SNPs in their middle position. The optimal size of the nucleotide regions flanking the SNPs (k-mer) was identified using the program Kchooser available with the package. The SNP matrix file (SNPs_all_matrix.fasta) generated by the tool was used to produce a maximum-parsimony tree, which is a consensus of up to 100 equally parsimonious trees. The phylogenetic tree was visualized using Dendroscope 3 (Huson and Scornavacca, 2012).

2.7.5 SNP analysis and pairwise SNP differences

SNP analysis was carried out using the Snippy version 3.2 pipeline (Seemann, 2015). Snippy finds SNPs between a given reference genome and genome sequence reads to identify both substitutions (Snps) and insertions/deletions (Indels). The pipeline was used to map read-pairs of every strain to the 12S0467 strain, which was kept as reference genome as more than 90% strains were from Europe, using the Burrows–Wheeler Aligner (BWA) v0.7.12 (Li and Durbin, 2009). Average read depths were calculated with SAMtools v1.2 (Li et al., 2009) in the pipeline and finally the SNPs were identified using the variant caller FreeBayes v0.9.21 (Garrison and Marth, 2012) with a minimum depth of 10 and a minimum variant allele proportion of 0.9. Snippy was used to pool all identified SNP positions (not indels) called in at least one isolate and a multiple sequence alignment of core SNPs was generated. For the type strain for which three laboratory strains were available, read mapping was carried out to the DSM 7528^T PacBio generated reference genome sequence. Pairwise SNP differences within and between groups based on geographical and outbreak/strain levels were calculated using pairwise_snp_differences, an R script to obtain summaries of pairwise SNP differences for groups of samples (Silva, 2015).

2.7.6 SNP analysis of within-host and outbreak strain variations

To understand the microevolution, the SNPs present in strains originating from the same host or from multiple hosts in the same outbreak were studied. Two variant calling tools where used and the collinear SNPs predicted by both tools were considered only. The Snippy pipeline was called with default parameters as mentioned above (Seemann, 2015). The LYVE version of the Snp Extraction Tool (SET), a method for generating high quality SNPs (hqSNPs) for outbreak investigations was employed (Katz et al., 2017). The default Lyve-SET option was used additionally for read cleaning, done by the CG-Pipeline with read cleaner parameters (minimum quality = 15, minimum average quality = 20, bases to trim = 100, which implies each read to be trimmed from the 5′ and 3′ ends up to 100 bp, until a nucleotide has at least a Phred quality of 15). Read mapping of each genome against the *C. chauvoei* strain 12S0467 by SMALT with only identity of 95% identity or above. The Lyve-SET option VarScan v2.3.7 was used to find and detect SNPs with parameters given in brackets (any site that has <75% consensus, fewer than 10 reads, or without at least two forward and two reverse reads was masked). Lyve-SET creates an SNP matrix with the mergeVcf.sh and set_processPooledVcf.pl scripts. Finally, the matrix was also filtered to remove sites with ambiguous nucleotides, invariant sites, and/or clustered SNPs to generate hqSNPs.

2.8 Core genome MLST (cgMLST)

A core genome MLST (cgMLST) scheme was generated using Ridom SeqSphere version 3.5 (Junemann et al., 2013) for *C. chauvoei*. The core genome MLST (cgMLST) target definer analysis which uses a BLAST tool was carried out using extracted genes from the reference genome DSM 7528^T (Altschul et al., 1997) to compare these genes against multiple query genome sequences (64 assembled genomes) with default parameters (Gene identity and query coverage were set to 90% and 100%, respectively. BLAST options include match reward 1, word size 11, gap open costs 5, mismatch penalty -1, gap extension cost 2). The analysis identified the cgMLST targets and accessory targets. SeqSphere software assigns alleles for each gene and generates an allelic profile for all strains.

2.9 CRISPR spacer sequence typing

Recent studies has described the application of CRISPR spacer sequence motives for the differentiation of *C. chauvoei* strains (Rychener et al., 2017). The typing with CRISPR spacer sequence motives was applied to differentiate the strains according to (Rychener et al., 2017) and was compared to core genome MLST analysis.

Chapter 3: Results

3.1 Genome completion, genome content and comparison

3.1.1 Genome assembly and annotation summary

Hierarchical Genome Assembly Process 3 (HGAP3) assembly pipeline implemented in the PacBio's SMRT Portal version 2.3.0 generated one linear contig representing the genome of the DSM 7528^{T} strain. The circular genome was generated using Minimus 2 software. For *C. chauvoei* strain 12S0467 two contigs representing the chromosome and one contig for a plasmid were obtained. The contigs representing the chromosome were merged using Minimus 2 software and were then circularized manually in Geneious 9.0.5. The contig representing the plasmid showed multiple copies of the plasmid sequence which were resolved by the Circlator software to generate a final 4 kb plasmid sequence shown in Figure 1 (Plasmid circularization summary). The genomes were annotated with the annotation pipelines RAST and Prokka to identify open reading frames, ribosomal RNAs and also transfer RNAs. The summary of the genome assembly of the two strains and circularization summary of the complete genomes are shown in Table 3 (Genome assembly and circularization summary). The annotation features were also compared with the previous published *C. chauvoei* sequence (JF4335) (Falquet et al., 2013) are shown in Table 4 (Genome summary (DSM 7528 ^{T}, 12S0467 and JF4335).

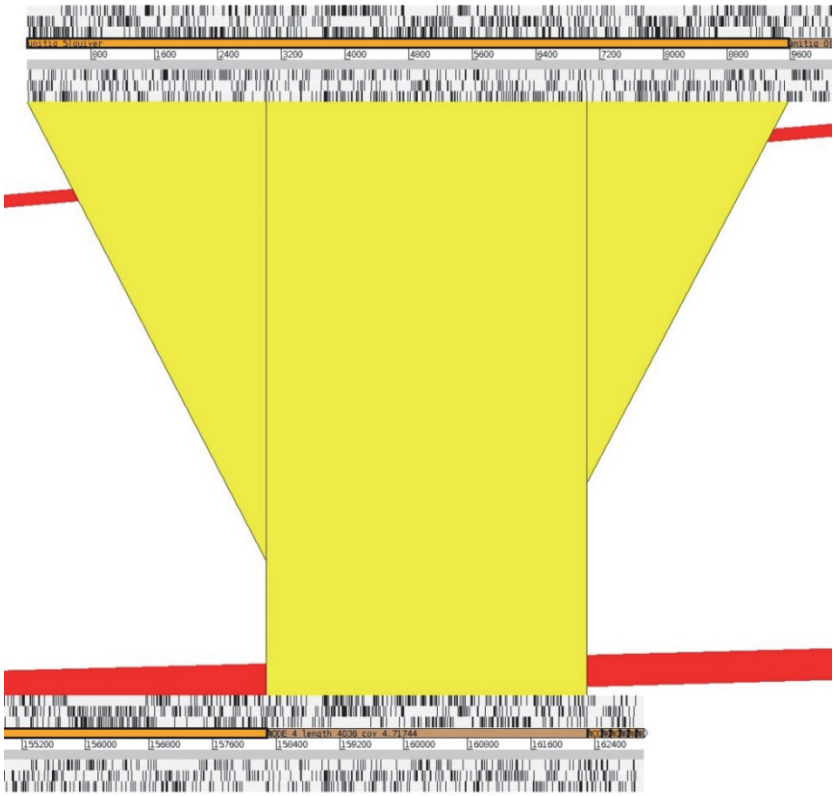


Figure 1: Artemis Comparison Tool generated view showing multiple copies of the plasmid sequence. Contig unitig_5 (shown above) generated by HGAP represents multiple copies of the plasmid sequence. Circlator reduced the multiplicity to generate contig NODE 4 (shown below) which was used to circularize and determine the 4 kb plasmid sequence.

Table 3: Genome assembly and circularization summary (*Clostridium chauvoei* strains DSM 7528[T] and 12S0467). Summary of the contig/genome size, coverage and consensus accuracy of the genome assembly (HGAP3) and subsequent circularization of the completed sequence of the two strains. Increased coverage and consensus accuracy in RS_resequening (error correction) indicate correct circularization of the genomes.

Clostridium chauvoei strain **DSM 7528 [T]**				
	Contig	Length (bp)	Coverage	Consensus accuracy
HGAP Summary				
1	unitig_0	2888463	250.2x	99.98%
Circularization Summary				
1	Genome	2872664	270x	100%
Clostridium chauvoei strain **12S0467**				
	Contigs	Length (bp)	Coverage	
HGAP Summary				
1	unitig_0	2735737	377x	99.98%
2	unitig_1	169089	394x	
3	unitig_5	9570	596x	
Circularization Summary				
1	Genome	2885630	417x	100%
2	Plasmid	3941	1218x	

Table 4: Genome summary (*Clostridium chauvoei* strains DSM 7528 [T], 12S0467 and JF4335). Genome summary table showing chromosome, plasmid, genome sizes and respective GC content. The number of rRNA operons, tRNA and coding sequences are also compared.

	DSM 7528 [T]		**12S0467**		**JF4335**	
Genome (in kb)	Chromosome	2872	Chromosome	2885	Chromosome	2825
	Plasmid	-	Plasmid	3.9	Plasmid	5.5
	Genome	2872	Genome	2889	Genome	2830
GC content	Genome	28.3%	Genome	28.3%	Genome	28.17%
	Plasmid	-	Plasmid	27%	Plasmid	25.28%
rRNA	9 operons		9 operons		5–10 operons	
tRNA	87		87		51	
CDS	RAST	2676	RAST	2632	2567	
	Prokka	2662	Prokka	2626		

rRNA: ribosomal RNA, tRNA: transfer RNA, CDS: coding sequences, RAST: Rapid Annotation using the Subsystem Technology

3.1.2 Genome features

3.1.2.1 Origin of replication of *Clostridium chauvoei*

Orifinder predicted a short and long *oriC* region of 257bp and 734bp on both sides of *dnaA* revealing five and eight DnaA box sequence motives (ttatccaca) with not more than one mismatch to the *Escherichia coli* DnaA box sequence, respectively. Marker genes commonly observed near the bacterial origin of replication were observed near the *oriC* region. Two probable DNA-unwinding elements (DUE) sites were identified within the shorter *oriC* region, based on its higher A/T composition. The predicted regions were similar to the origin of replication known for *Bacillus* (*B.*) *subtilis oriC,* displaying DnaAbox clusters in the intergenic regions both upstream and downstream of *dnaA* (Figure 2).

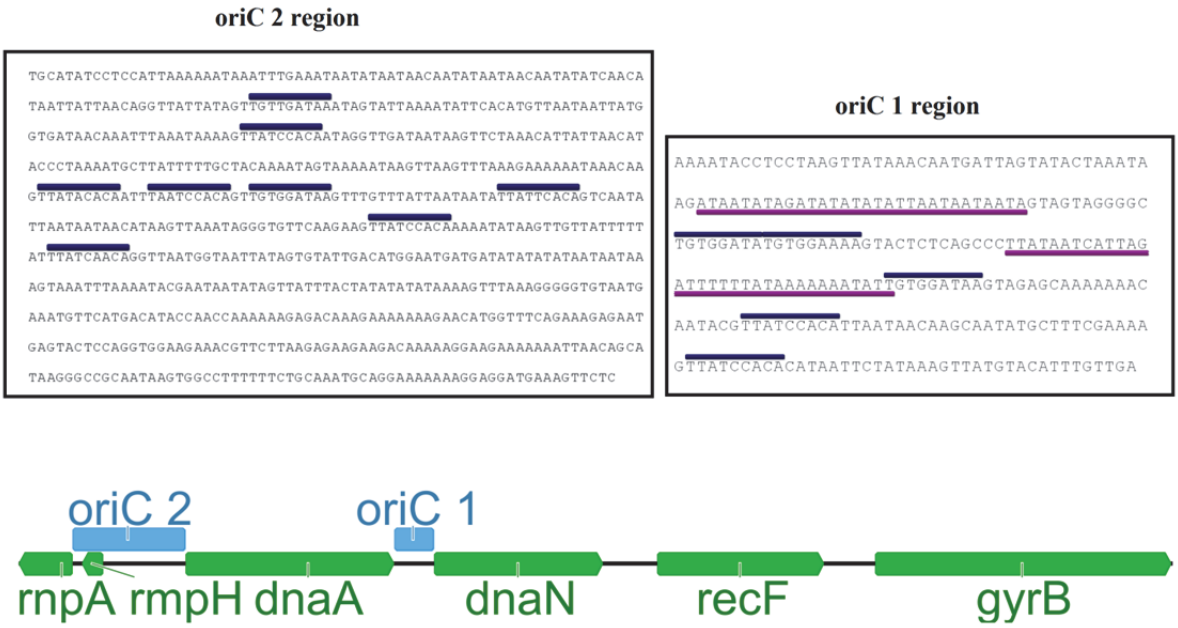


Figure 2: Schematic diagram of the origin of replication (*oriC*). The two *oriC* regions identified by orifinder are depicted in the lower part of the figure; sky blue boxes represent *oriC1* and *oriC2* with the presence of marker genes (*rnpA–rmpH–dnaA–dnaN–recF–gyrB*) near to the origin of replication. The sequence of two *oriC* regions (*oriC1* and *oriC2*) with DnaA boxes (dark blue) and two probable DNA-unwinding elements (DUE) (purple) is shown above. The schematic representation was created using Geneious 9.0.5 (http://www.geneious.com/).

3.1.2.2 Prophages and CRISPR elements

The PHAST server identified intact prophages of 53 kb and 46.3 kb for *C. chauvoei* DSM 7528^{T}and 12S0467, respectively, and an incomplete prophage of 10.2 kb for both strains with similar genetic structure. The incomplete phage element showed the presence of a small number of bacterial specific protein genes. A schematic representation of the two prophages (complete and incomplete) identified in the *C. chauvoei* DSM 7528^{T} is shown in Figure 3. A CRISPR element with two repeat sites was identified in the genomes with site lengths of around 0.6 kb (site 1) and 2 kb (site 2) separated by a distance of around 1.5 kb. There were

11 repeats for site 1 in both genomes, whereas 32 and 34 repeats were observed for site 2 in *C. chauvoei* 12S0467 and *C. chauvoei* DSM 7528^{T}, respectively. Direct repeat (DR) sequences were of the same sequence for both isolates at both sites (GATTAACATTAACATGAGATGTATTTAAAT). The spacers between the CRISPR repeat elements at site 2 were variable with 31 and 33 numbers for *C. chauvoei* 12S0467 and *C. chauvoei* DSM 7528^{T}, respectively, and a length ranging between 35 and 37 bp. Both CRISPR repeat elements were accompanied by the same genes for CRISPR associated proteins (Cas) flanking the site.

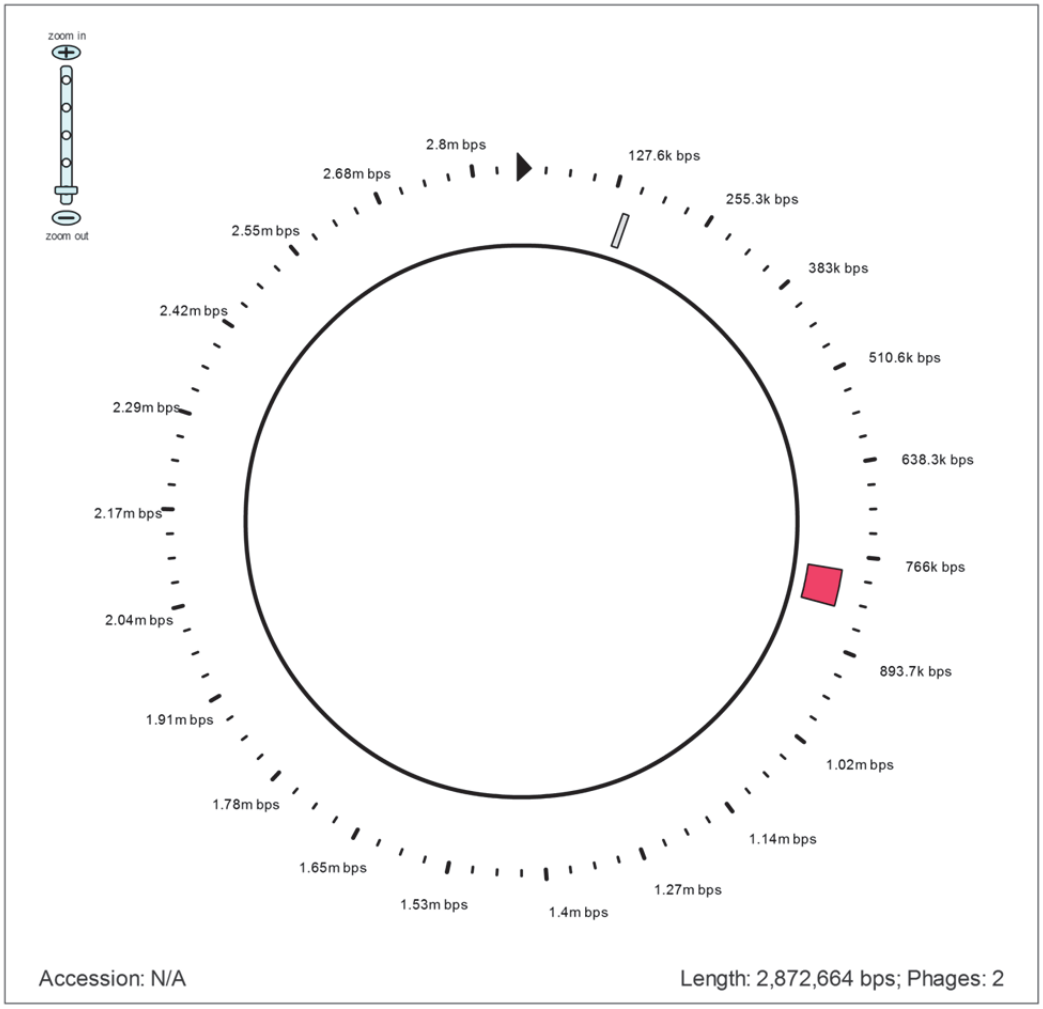


Figure 3: Schematic representation of the complete (red) and incomplete (grey) prophages in the genome of *Clostridium chauvoei* strain DSM 7528^{T}. The outer circle defines the position of the genome DSM 7528^{T}. The complete prophage is shown as red block and the incomplete prophage is shown as a grey block. The picture was generated using PHAST phage prediction tool.

3.1.2.3 Antibiotic resistance genes

Antibiotic resistance genes were predicted using the CARD database which identified eight genes potentially conferring resistance to antibiotics including aminocoumarin (two genes), beta-lactam antibiotics, rifampicin, tetracycline, macrolide, peptide antibiotic and a gene

conferring antibiotic resistance via molecular bypass for glycopeptide resistance in both *C. chauvoei* (DSM 7528^{T}and 12S0467) genomes.

3.1.2.4 Flagellin type C (fliC) genes and virulence factor genes

Flagellin is the principle component of the bacterial flagellum, an important structure in the mobility of bacterial pathogens and also involved in pathogenicity. Earlier studies carried out at the nucleotide sequence level have shown the existence of two *fliC*s which occur in tandem, designated as *fliA*(*C*) and *fliB*(*C*). They are displaying highly conserved N and C terminal regions. Sequence analysis of the two *C. chauvoei* isolates showed the presence of three *fliC* genes with a coding length of 414 amino acids. This is the first report of triplicate *fliC* genes for the species (Figure 4). Both genomes contain a gene coding for CctA, NanA sialidase, haemolysin III, haemolysin XhlA, internalin A, hyaluronidase I (NagI) and hyaluronidase H (NagH). The Blast2GO annotation identified the major functional gene classes found in the genome which includes several glycosidases and various proteins involved in metal ion binding and transportation. The genome also harbours several classes of proteolytic enzymes.

3.1.2.5 Genes for spore resistance, sporulation and germination

The characterized two *C. chauvoei* genomes have around 70 predicted genes for spore resistance, sporulation and germination which include five small, acid-soluble spore proteins (SASPs) involved in spore resistance and a set of genes involved in dipicolinic acid (DPA) synthesis similar to that identified in the Clostridia cluster I group genomes. Proteins involved in the various stages of sporulation could also be identified, including the major regulator Spo0A and key sigma factors regulating spore formation. The *C. chauvoei* genomes also harbour the *gerK* operon which encodes the Germinant Receptor (GR) and the lytic enzymes involved in the germination processes. The Spo0A amino acid sequence of all four species *C. chauvoei* (DSM 7528^T), *C. septicum* (CSUR P1044), *C. perfringens* (ATCC 13124) and *C. botulinum* type A (ATCC 3502) was compared and revealed a central region with amino acid deletions for *C. chauvoei* and *C. septicum* as shown in Figure 5. The genetic relatedness of important genes regulating sporulation and germination of *C. chauvoei* was compared to that of *C. septicum* (CSUR P1044), *C. perfringens* (ATCC 13124) and *C. botulinum* type A (ATCC 3502) based on BLASTP searches with the respective protein coding genes. Most of the *C. chauvoei* genes for sporulation and germination showed greater homology with those of *C. septicum* followed by *C. perfringens* and *C. botulinum* which suggests a certain genetic relatedness and similar mechanisms. The gene list used for comparison, length of coding sequences, pairwise similarity and coverage of query genes are shown in Table 5.

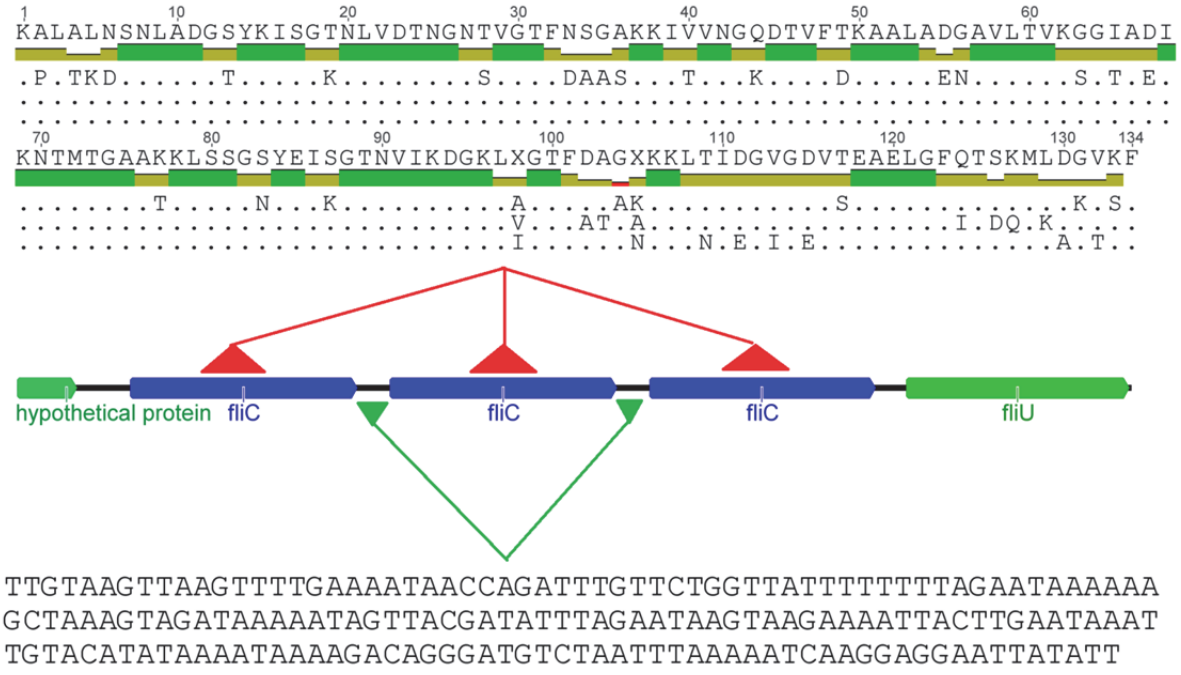


Figure 4: Schematic representation of multiple *fliC* genes. Arrangement of the three *fliC* genes in the flagellar operon. The two 181bp spacer regions between the *fliC* genes are identical and the region is indicated by the green triangle box. Alignment of the central variable region of *fliC* (amino acid position 177 to 310) is shown above with conserved amino acids depicted as dots and the region is indicated as red triangle box. Alignment and schematic representation was created using Geneious 9.0.5.

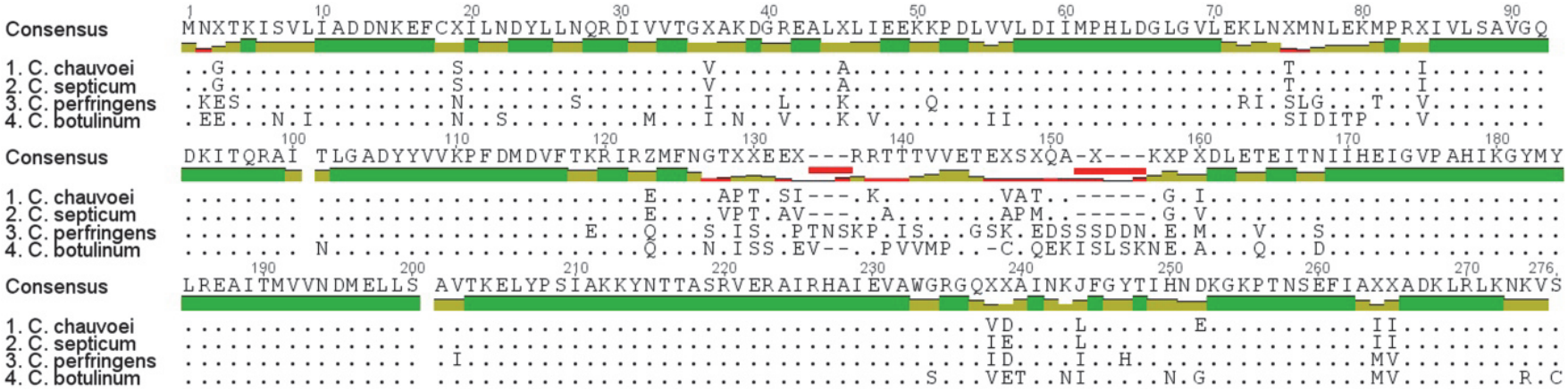


Figure 5: Alignment of the Spo0A protein sequence of four species. Alignment of the SpoA protein sequence of *Clostridium chauvoei* (DSM 7528^{T}), *C. septicum* (CSUR P1044), *C. perfringens* (ATCC 13124) and *C. botulinum* (ATCC 3502). Conserved amino acids are depicted as dots. The red line indicates the amino acid deletions for *C. chauvoei* and *C. septicum* SpoA proteins. Alignment and schematic representation was created using Geneious 9.0.5.

Table 5: Genetic relatedness of crucial genes involved in spore resistance, sporulation and germination. The genetic relatedness of crucial genes involved in spore resistance, sporulation and germination of *C. chauvoei* was compared to *C. septicum* (CSUR P1044), *C. perfringens* (ATCC 13124) and *C. botulinum* type A (ATCC 3502) proteomes based on BLASTP (Altschul et al., 1997) searches. The percentage pairwise identity and percentage query coverage of the coding sequence was computed for assessing the genetic relatedness. Most of the *C. chauvoei* genes for sporulation and germination showed greater homology to those of *C. septicum.*

Gene description	Product	CDS (L)	% PWI	% QC	% PWI	% QC	% PWI	% QC
			C. septicum		***C. perfringens***		***C. botulinum***	
4-hydroxy-tetrahydrodipicolinate synthase	DapA	293	81.2	100	65.2	98.29	72.9	100
Electron transfer flavoprotein, alpha subunit	EtfA CDS 1	397	92.2	100	76.1	99.24	72.5	99.75
Electron transfer flavoprotein, alpha subunit	EtfA CDS 2	336	96.5	100	75.7	99.40	64.2	98.81
Spore germination protein GerKA	GerKA	478	89.7	100	62.4	99.58	31.8	95.60
Spore germination protein GerKC	GerKC	373	85.8	100	48.4	100	23.4	85.75
Serine/threonine protein kinase PrkC	PrKC	661	77.3	99.24	50.6	96.22	45.7	86.54
RNA polymerase sporulation specific sigma factor SigE	SigE	236	97.4	100	83.4	100	83.8	100
RNA polymerase sporulation specific sigma factor SigF	SigF	252	98	100	72.0	98.01	70.5	96.02
RNA polymerase sporulation specific sigma factor SigG	SigG	258	99.2	100	89.5	100.0	88.3	100
RNA polymerase sporulation specific sigma factor SigH	SigH	215	93.0	100	76.0	93.46	81.1	88.79

RNA polymerase sporulation specific sigma factor SigK	SigK	232	97.8	100	70.1	98.70	68.1	99.13
Spore cortex-lytic enzyme precursor	SleB	195	79.1	100	61.2	69.07	59.3	69.07
Spore cortex-lytic enzyme SleC	SleC	445	91.0	100	60.5	98.65	40.8	16.67
Stage 0 sporulation regulator (Spo0A)	Spo0A	268	95.5	100	80.9	100	77.3	100

L - Length, PWI - Pairwise identity, QC - Query Coverage

3.1.3 Genome visualization and comparative genome analysis

3.1.3.1 Circular plot of *Clostridium chauvoei* DSM 7528^{T}

The circular genome plot with RAST predicted proteins and other elements for the strain *C. chauvoei* DSM 7528^{T}was created using DNA plotter software (Figure 6).

3.1.3.2 Phylogenetic relatedness of *Clostridium chauvoei* within the genus *Clostridium*

Phylogenomic analysis of the genus *Clostridium* was carried out with Gegenees software (Ågren et al., 2012) using genomic sequence data including species from Cluster 1 of the genus. Significant homology was observed for *C. chauvoei* and *C. septicum*, *C. botulinum* type C and D, *C. novyi* and *C. haemolyticum*, for *C. botulinum* type A and *C. sporogenes* and for *C. botulinum* type B and E*, C. beijerinckii* and *C. butyricum. C. chauvoei* showed a close relatedness to *C. septicum* (74%) in phylogenetic analysis within the genus *Clostridium* (Figure 7).

3.1.3.3 Core and accessory genomes

Comparative genome analysis carried out to identify orthologous genes among the three *C. chauvoei* genomes identified 2514 core genes and 148 accessory genes (Figure 8). This indicates the genomes harbour limited accessory genes in comparison to those observed for other related *Clostridium* spp such as *C. perfrignens* and *C. botulinum*. The number of orthologous genes shared by the three *C. chauvoei* genomes (DSM 7528^{T}, 12S0467 and JF4335) and the *C. septicum* (CSUR P1044) genome was 1559 genes. The number of species specific unique genes among the 4140 genes identified in both species together was 951 genes for *C. chauvoei* and 1481 genes for *C. septicum,* respectively.

3.1.3.4 Collinear blocks and genetic relatedness among *Clostridium chauvoei* strains (DSM 7528^T, 12S0467) and with *Clostridium septicum* (CSUR P1044)

C. chauvoei DSM 7528^T and 12S0467 genomes possessed six locally collinear blocks (LCB), with an inversion of three blocks and location change of one (Figure 9A) indicating a significant relationship among the type and field strain. DSM 7528^T and CSUR P1044 showed several identical LCB, hence *C. chauvoei* is genetically closer related to *C. septicum* than other species of *Clostridium* for which genome sequences are available (Figure 9B).

3.1.3.5 Comparative analysis of subsystem category distribution of *Clostridium chauvoei* and *Clostridium septicum*

The genome features of *C. chauvoei* (DSM 7528^T) and *C. septicum* (CSUR P1044) were compared based on subsystem category distribution features in RAST (Figure 10). Both genomes shared similar subsystem category distributions whereas the subsystem feature counts were higher for *C. septicum* for genes related to prophages, phosphorus and carbohydrate metabolism.

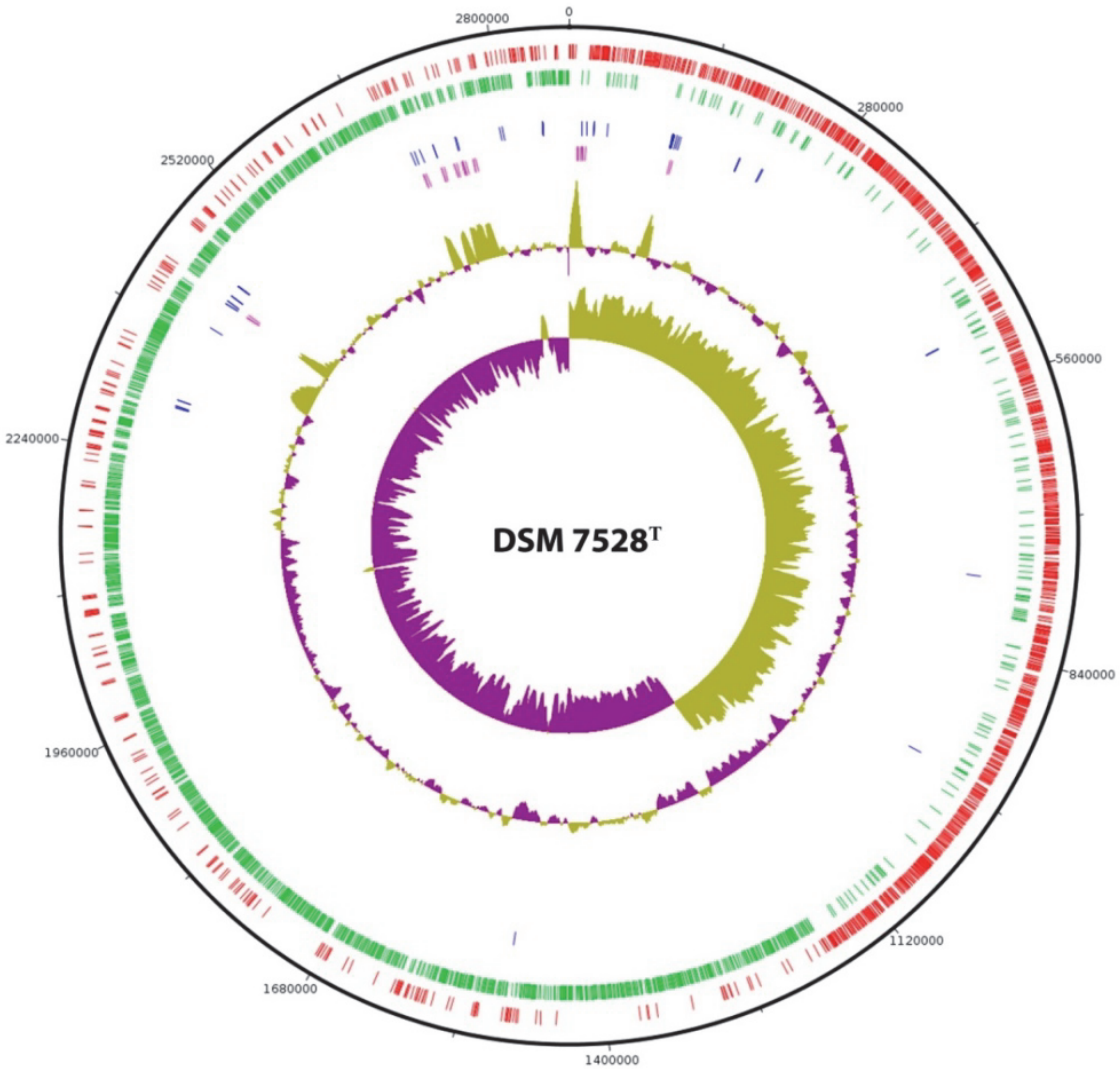


Figure 6: Circular plot of the genome of *Clostridium chauvoei* strain DSM 7528^{T}. The plot was created using DNAPlotter. Red and green indicate forward and reverse genes predicted by RAST. Blue and pink lines correspond to tRNA and rRNA genes, respectively. The inner circle displays the GC skew and the second circle from the centre displays the G+C composition.

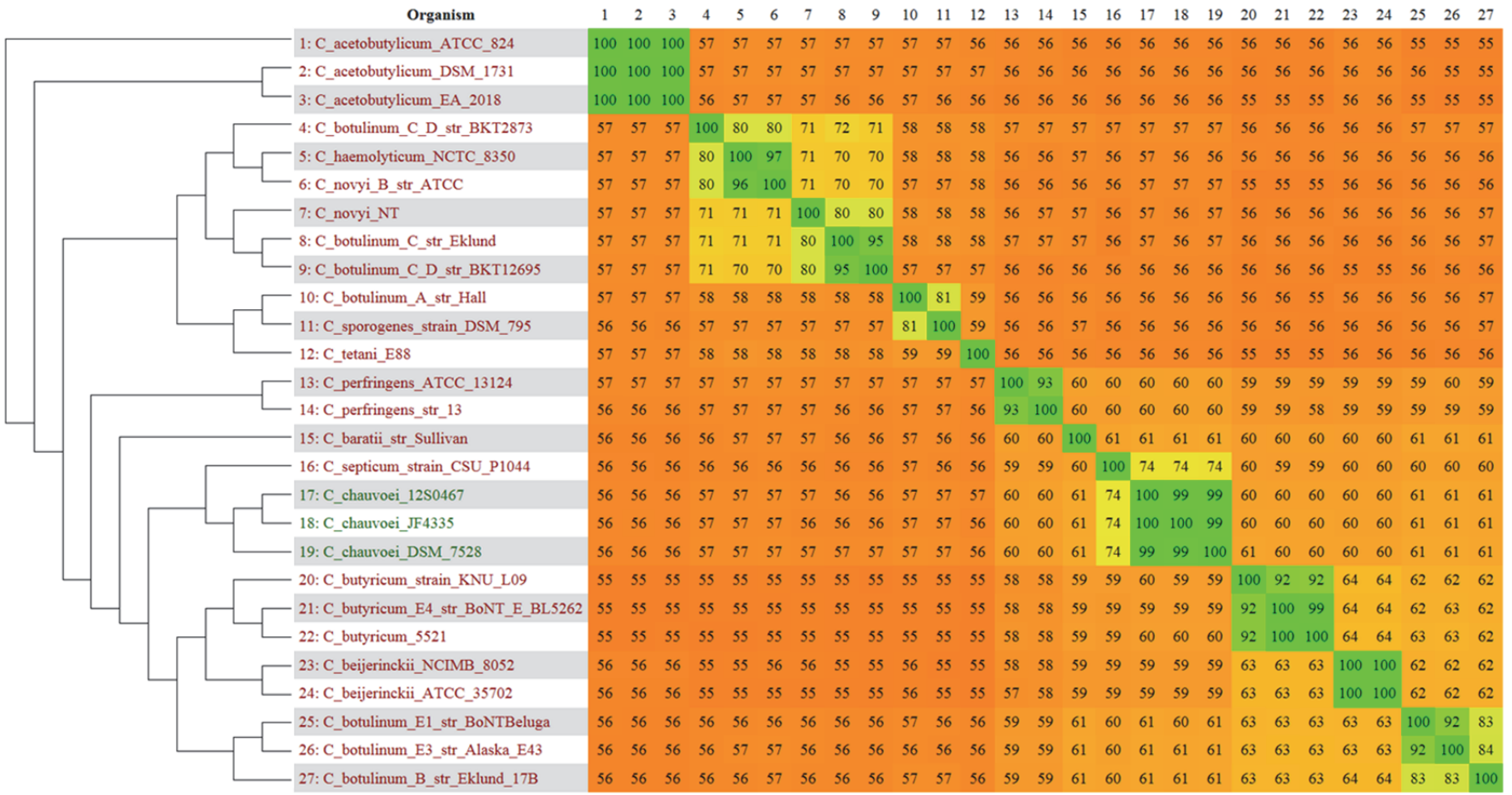


Figure 7: Phylogenetic relatedness of *Clostridium chauvoei* within the genus *Clostridium.* TBLASTX with fragment alignment settings of 500/500 were used for comparative genomics and heat maps were generated with a cutoff threshold of 35% for non-conserved genetic material. A neighbour joining method based phylogenetic tree was created in Splits Tree 4 using the PHYLIP file exported from Gegenees software.

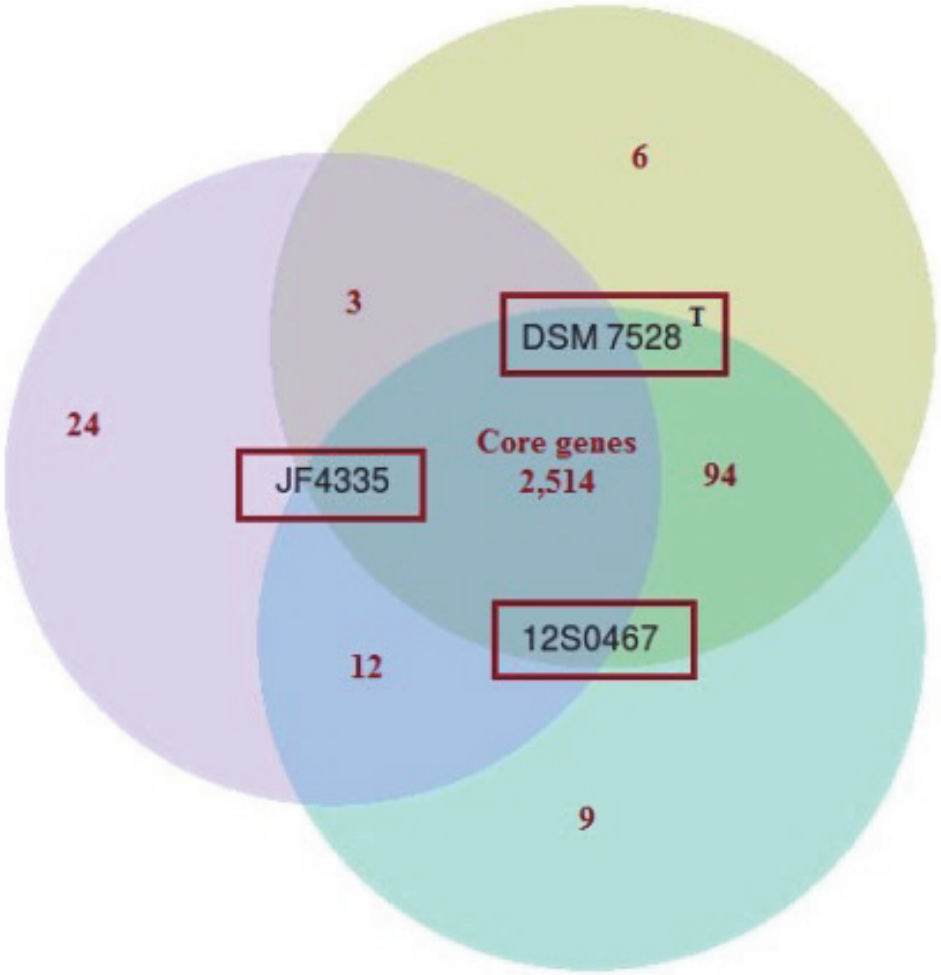


Figure 8: Venn diagram showing core genes and accessory genes (*Clostridium chauvoei* DSM 7528^T, 12S0467 and JF4335). Orthologous protein clusters identified using PanOCT shared by all three genomes were placed in the centre and considered as the core genes. The accessory genome includes the genes shared by two isolates or is unique to one isolate. The venn diagram is not to scale.

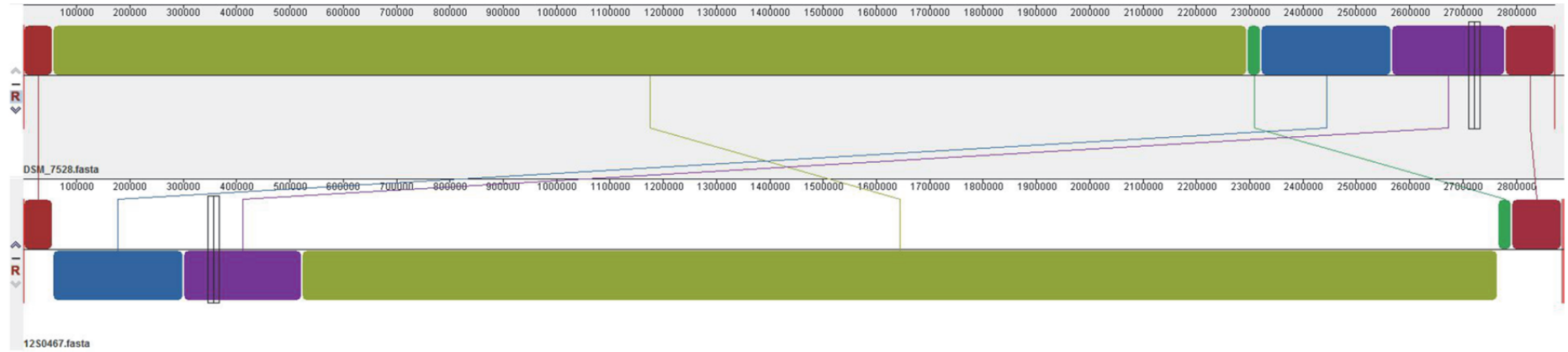


Figure 9A: *Clostridium chauvoei* DSM 7528^T (above) and *Clostridium chauvoei* 12S0467 (below)

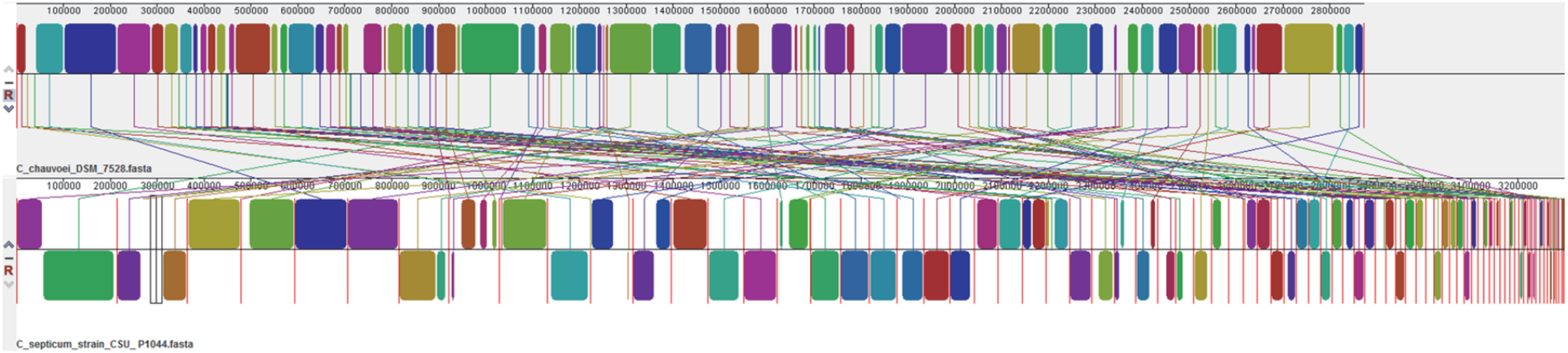


Figure 9B: *Clostridium chauvoei* DSM 7528^T (above) and *Clostridium septicum* CSUR P1044 (below)

Figure 9: Genome alignment plot created using progressiveMauve. Related Locally Collinear Blocks (LCB) between genomes are connected with a line and identified by the same colour. Blocks that are inverted are placed under the centre line of the genome. Figure 9A: *C. chauvoei* DSM 7528^T (above) and *C. chauvoei* 12S0467 (below) genomes have six LCB and there is an inversion of three and a location change of one LCB. Figure 9B: *C. chauvoei* DSM 7528^T (above) and *C. septicum* CSUR P1044 (below) showing several identical LCB and red vertical lines in the *C. septicum* strain CSUR P1044 indicate contig boundaries.

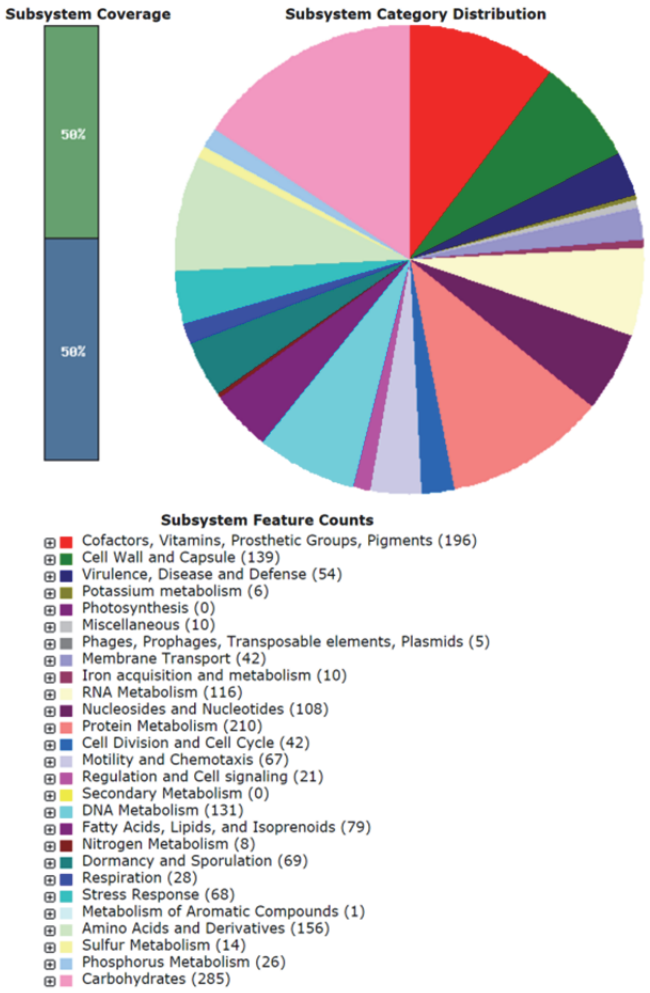


Figure 10A: ***Clostridium chauvoei*** **(DSM 7528[T])**

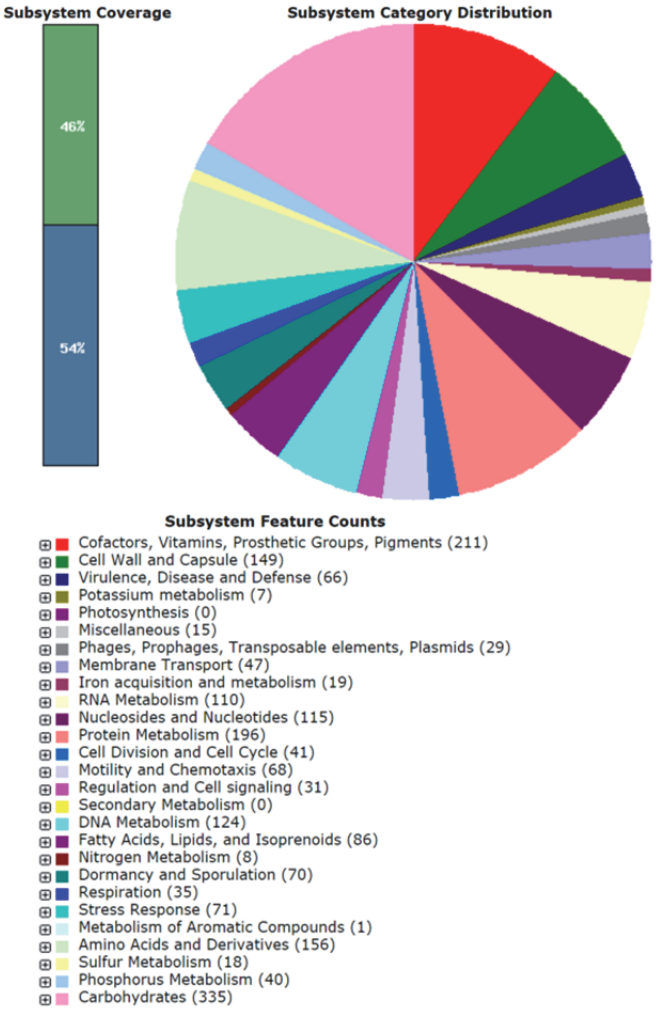


Figure 10B: ***Clostridium septicum*** **(CSUR P1044)**

Figure 10: RAST based subsystem category distribution features of *Clostridium chauvoei* (DSM 7528[T]) and *Clostridium septicum* (CSUR P1044) strains. The genome features of *C. chauvoei* (DSM 7528[T]) (Figure 10A) and *C. septicum* (CSUR P1044) (Figure 10B) were compared based on subsystem category distribution features in RAST. Both genomes share similar subsystem category distributions. Subsystem feature counts were higher for *C. septicum* for genes related to phages, prophages, plasmids, phosphorus metabolism and carbohydrates.

3.2 Comparative genomics and phylogeny

3.2.1 Genome assembly and annotation

3.2.1.1 Genome assembly evaluation

De novo assembly evaluation was carried out to determine the best assembly options for the *C. chauvoei* genome with the paired end MiSeq read data. All the assemblers generated comparable results whereby the N50 value was higher for SPAdes, misassembles (large and local), duplication ratios were comparable for SPAdes and CLC Genomic Workbench assemblers and genome fraction was better represented by MaSuRCA. The genome assembly statistics based on QUAST are depicted in Table 6.

Table 6: Genome assembly evaluation. Genome assembly evaluation was carried out using QUAST run in the Genome Assembly Gold-standard Evaluation (GAGE) mode. The assemblers providing the best results are depicted in bold for every parameter for all the three strains. The description of various parameters used in GAGE analysis is listed below the table.

	CLC (Genomic Workbench 8)			SPAdes 3.9.1			MaSuRCA		
	Str. 1	**Str. 2**	**Str. 3**	**Str. 1**	**Str. 2**	**Str. 3**	**Str. 1**	**Str. 2**	**Str. 3**
Contigs	85	85	89	**68**	**72**	**65**	111	124	98
N50	62818	59816	58237	**82671**	**82670**	**82671**	63319	69856	62627
Misassemblies	1	**0**	**1**	**0**	**0**	**1**	4	2	**6**
Local misassemblies	**0**	**1**	**2**	**0**	**1**	**2**	1	10	6
Corrected N50	61978	59816	57498	**82671**	**82671**	**82671**	62774	69856	62627
Genome fraction (%)	96.04	96.02	95.77	96.54	96.59	96.25	**97.44**	**97.53**	**97.08**
Unaligned	**0**	**0**	2	**0**	4	**0**	**0**	4	**0**
Duplication ratio	1.001	**1.000**	**1.001**	**1.000**	1.001	**1.001**	1.014	1.020	1.014

Str. 1: 12S0468, Str. 2: 12S0469, Str. 3: 12S0470

- Contigs (The number of contigs above 500 bp length).
- N50 (N50 length is defined as the shortest sequence length at 50% of the genome for given sets of contigs each with its own length in bp).
- Misassemblies (In comparison with the reference genome, the sum of the number of relocations, translocations and inversions affecting at least 1000 bp).
- Local misassemblies (Local errors, defined as misjoinining where the left and right pieces map onto the reference genome to distinct locations that are <1000 bp apart, or that overlap by <1000 bp).
- Corrected N50 (Corrected N50 size, defined as the N50 size obtained after splitting contigs at each error).
- Genome fraction (%) (The fraction of the reference genome covered by contigs).

- Unaligned contigs (The number of unaligned contigs, computed as the number of contigs that could not be aligned, even partially, to the reference genome).
- Duplication ratio (Duplication ratio, an approximation of the amount of overlaps among contigs that should have been merged).

Note: Some of the parameters defined in the analysis (Misassemblies, Local misassemblies, Unaligned) may represent true differences rather than assembly errors since the comparison involved only a related reference strain from the same region.

3.2.1.2 Genome assembly and post-assembly improvements

SPAdes was chosen as the preferred assembler for all 61 isolates (Illumina MiSeq™) for the current data sets, but the reads were initially quality trimmed using the Sickle tool before assembly. The assembly contig number ranged between 63 and 96 among the isolates. N50 values, total assembly length and GC content varied between 78 kb and 94 kb, 2.78 Mb and 2.8 Mb and 27.91% and 27.99% for the strains, respectively. This indicates that the genome variability is very limited within the species with respect to genome size and GC content. Assembly statistics of the de novo assembled genomes are shown in Table 7(A). Post-assembly improvements of the draft genomes were carried out using the post-assembly genome-improvement toolkit (PAGIT) which improved the N50 values, total assembly length and GC content of the strains. Post-assembly statistics generated for the strains are shown in Table 7(B). Exceptionally high N50 value and GC content were obtained for the strains S0133-09 and 16S0574, respectively (Shown in bold in the table).

Table 7(A) and 7 (B): Genome assembly and post assembly improvement summary. The table shows the strain designation and the number of contigs obtained, N50 value, total length, GC% summary of assembly 7(A) and post-assembly improvements 7(B).

		Table 7(A)				Table 7(B)			
	Strain designation	Genome assembly summary				Post-assembly improvement summary			
		Con-tigs	N50	Total length	GC%	Con-tigs	N50	Total length	GC%
1	12S0468	71	94539	2792907	27.94	56	116948	2822124	28.02
2	BS91-02	69	94539	2792502	27.92	51	115201	2818669	28.01
3	BS94-02	69	94539	2793530	27.94	57	95064	2816757	28.02
4	BS92-02	68	94539	2789964	27.93	50	117127	2816444	28.01
5	BS104-02	69	94539	2791614	27.92	56	95063	2818500	27.99
6	BS105-02	68	94539	2793060	27.93	50	115135	2823927	28.02
7	BS106-02	68	94539	2791111	27.92	53	115266	2823796	28
8	BS90-02	66	94539	2791927	27.93	52	115431	2826411	27.99
9	BS109-02	66	94539	2791382	27.92	53	114995	2824522	28.02
10	S0260-09	69	94539	2791226	27.93	54	114896	2820978	28.02
11	BS107-02	68	94539	2792535	27.94	55	115013	2817180	28
12	BS108-02	80	94539	2803677	27.98	63	95295	2834069	28.05
13	12S0464	66	94539	2791151	27.93	51	114972	2813664	27.99
14	12S0465	67	94539	2791480	27.93	55	114998	2815953	27.99

15	12S0466	65	94539	2791356	27.93	53	94719	2819449	28
16	12S0469	72	94539	2800342	27.94	58	95113	2828765	28.02
17	12S0470	64	87178	2782721	27.92	49	117841	2809914	27.98
18	12S0471	68	94538	2791588	27.93	56	114968	2828947	28.04
19	12S0472	71	94576	2794432	27.94	57	94945	2824613	28.03
20	S0162-10	67	94539	2793458	27.94	53	114768	2813203	27.98
21	11S0315	69	94538	2794224	27.94	55	114958	2826679	28.03
22	11S0316	69	94538	2794489	27.94	56	116494	2822724	28.01
23	BS80-01	66	94539	2803804	27.92	56	95257	2833412	27.97
24	BS79-01	68	94538	2792909	27.92	57	95124	2825159	28.04
25	13S0851	68	82671	2792608	27.93	52	95125	2825774	28.02
26	13S0854	67	82671	2792139	27.93	54	93304	2823447	28.01
27	S0021-10	67	86750	2782438	27.92	51	114144	281437	28.03
28	S0121-09	70	83756	2782631	27.92	54	95164	2819105	28.03
29	S0022-10	72	86750	2783489	27.93	54	114751	2812173	28.01
30	S0008-10	66	94539	2791851	27.93	53	116879	2823670	28.03
31	S0040-08	68	94539	2791268	27.93	59	95074	2831556	28.03
32	S0041-08	69	94561	2791553	27.93	57	94995	2823054	28.02
33	15S0088	98	63119	2801109	27.98	75	95719	2834270	28.09
34	15S0089	87	77886	2805286	27.99	63	94393	2834969	28.11
35	16S0579	89	69875	2800505	27.96	60	89230	2834215	28.1
36	15S0008	71	94539	2784669	27.93	54	94922	2813272	28
37	S0099-08	67	82671	2793800	27.93	54	95659	2821967	28.01
38	S0100-08	70	83756	2795047	27.94	57	94943	2824036	28.02
39	S0101/08	67	83755	2793533	27.93	57	95136	2820004	28.03
40	S0105-08	70	94539	2799938	27.95	57	95511	2835624	28.03
41	S0013-08	64	94576	2783773	27.93	52	116376	2810873	28.02
42	S0014-08	66	82671	2784844	27.92	53	114847	2813121	28
43	S0015-08	63	94576	2783447	27.92	54	95567	2815486	28
44	S0098-08	71	82671	2793260	27.94	58	91165	2824352	28.02
45	S0102-08	68	94539	2792191	27.93	52	116789	2823463	28.03
46	S0103-08	64	94539	2790831	27.93	53	115160	2814456	28
47	BS169-00	66	82671	2791598	27.93	53	83439	2815800	28.01
48	BS171-00	64	94539	2789906	27.92	63	115236	2831338	28.05
49	15S0023	76	68788	2782674	27.93	76	69133	2836884	28.09
50	**16S0574**	**322**	**78664**	**3158655**	**28.51**	**321**	**69775**	**3351038**	**28.74**
51	16S0578	80	68747	2794618	27.94	78	69640	2845220	28.09
52	11S0318	75	94478	2796103	27.94	61	96067	2824447	28.02
53	**S0133-09**	**69**	**152245**	**2818039**	**28.07**	**69**	**152805**	**2845421**	**28.21**
54	S0132-09	70	83740	2799076	27.98	69	84256	2839176	28.1
55	S0131-09	69	82759	2792130	27.92	68	94901	2834014	28.04
56	S0136/09	63	82671	2782334	27.91	61	95436	2821436	28.03
57	S0134-09	63	94538	2780808	27.92	62	94672	2816059	28.03
58	BC93-06	70	83756	2791595	27.93	70	84441	2829479	28

59	BC97-06	64	82671	2783212	27.92	63	95844	2831752	28.05
60	BC103-06	63	94576	2783283	27.92	63	95570	2841558	28.05
61	BC138-06	66	94539	2782085	27.93	66	95384	2829379	28.04
62	DRS014052	93	68762	2780608	27.79				

3.2.1.3 Genome annotation, plasmid, phage and CRISPR elements

Genome annotation based on the Prokka annotation pipeline revealed protein coding genes which varied from 2562 to 2608 among the 65 strains. The annotation summary (number of protein coding genes, rRNA, tRNA) for each strain is shown in Table 8 (A). A plasmid of 4 kb size already reported from the finished field strain (12S0467) was identified in all strains of the current study. All the genomes harboured two CRISPR repeat sites and showed variations for CRISPR repeat numbers among the isolates. The CRISPR repeat site 1 and 2 had 11 and 34 repeats for most of the isolates in the current study. Variation of the CRISPR repeat numbers were observed for few strains and ranged from 9 to 14 for site 1 and 21 to 65 for site 2. The genome of strain S0132-09 was denoted with only one CRISPR site of 29 repeats. The genetic structure and prophage composition was similar in all strains. CRISPR elements, repeat number and estimated sizes of the identified prophages are shown in Table 8(B). The strains also showed variations in the composition of unique CRISPR spacer sequences among the isolates. The total number of CRISPR spacer sequences was 2527 (64 strains) among all the strains, whereas 45 unique spacer sequence motives were identified. Most of the strains from Europe and also the type strain had 43 spacers. The unique spacers showed marked variations with respect to strains from Austria, Switzerland, few strains from Bavaria and Lower Saxony, 15S0023 (South Africa), strains from North Rhine-Westphalia and several strains of unknown origin. Most of the strains harboured a prophage of an average size of 31.6 kb (based on annotation with PHASTER). The incomplete phage (10 kb) predicted by PHAST for the complete genomes was not predicted by PHASTER for both, the complete (in the reanalysis with PHASTER) and draft genomes. Few strains showed variation with respect to the size of predicted prophages.

Table 8(A) and 8(B): Genome annotation, CRISPR elements and phage summary. Table 8 shows the annotation features (number of coding genes, tRNA, rRNA) in 8(A) and the CRISPR element repeat units, the corresponding repeat numbers and estimated size of prophage identified are shown 8(B).

	Table 8(A)			Table 8(B)		
Strain designation	Genome annotation Summary			Prophage	CRISPR elements	
Clostridium chauvoei strains	CDS	rRNA	tRNA	Size (kb)	Repeat no (Site 1)	Repeat no (Site 2)
S0013-08	2585	11	87	31.6	11	34
S0014-08	2595	10	87	31.6	11	34
S0015-08	2589	11	80	31.6	11	34
S0131-09	2608	12	74	31.6	11	34

S0136-09	2593	11	76	31.6	11	34
BC 97-06	2603	11	82	31.6	11	34
BC 103-06	2608	12	82	316	11	34
BC 138-06	2601	11	84	31.6	11	34
11S0316	2601	11	86	31.6	11	34
BS80-01	2628	6	86	31.6, 42.9	11	34
BS79-01	2598	11	81	31.6	11	34
BS91-02	2590	12	79	31.6	11	34
BS94-02	2584	10	84	31.6	11	34
BS92-02	2586	10	78	31.6	11	34
BS104-02	2595	11	60	31.6	11	34
BS105-02	2595	10	83	31.6	11	34
BS106-02	2595	11	78	31.6	11	34
BS109-02	2606	11	78	31.6	11	34
BS107-02	2588	11	87	31.6	11	34
BS108-02	2593	12	87	31.6	11	34
12S0464	2586	10	73	31.6	11	34
12S0465	2585	11	81	31.6	11	34
12S0466	2595	10	82	31.6	11	34
12S0469	2594	11	80	31.6	11	34
12S0471	2601	11	82	31.6	11	34
12S0472	2587	12	75	31.6	11	34
11S0315	2600	12	91	31.6	11	34
S0133-09	2599	18	95	45.6	14	65
S0008-10	2594	12	80	31.6	11	34
S0040-08	2606	11	87	31.6	11	34
S0041-08	2595	11	79	31.6	11	34
15S0008	2592	9	80	31.6	11	34
S0134-09	2587	11	81	31.6	11	34
11S0318	2599	11	80	31.6	11	34
16S0574	2999	13	93	31.6	11	34
S0098-08	2588	11	76	31.6	11	34
16S0578	2605	13	81	32.5	11	34
JF4335	2567		51	31.6	11	34
12S0468	2588	12	78	31.6	11	32
13S0851	2596	10	89	31.6	11	33
13S0854	2595	10	84	31.6	11	33
15S0088	2585	16	91	32.5	11	33
S0260-09	2595	13	83	31.6	11	32
S0162-10	2593	10	82	31,6	11	32

16S0579	2584	15	97	31.6	11	31
S0103-08	2588	10	84	31.6	11	30
DRS014052	2562	10	65	32.5, 11.9	11	26
15S0023	2597	13	82	32.5	11	24
12S0470	2596	10	84	31.6	11	21
S0021-10	2583	10	82	17. 5	11	18
S0121-09	2583	10	82	17. 5	11	18
S0022-10	2578	12	88	17. 5	11	18
S0105-08	2602	11	89	31.6	10	31
BC93-06	2598	12	78	31.6	10	31
S0100-08	2599	11	88	31.6	10	31
S0101-08	2588	13	87	31.6	10	31
BS171-00	2608	10	77	31.6	9	38
BS90-02	2596	10	83	31.6	9	36
S0102-08	2599	10	75	31.6	9	36
S0099-08	2597	11	86	31.6	9	36
BS169-00	2591	11	86	31.6	9	28
S0132-09	2604	11	81	31.6		29
15S0089	2603	17	88	32.5	Not resolved	

3.2.2 Core, accessory and pan-genome

3.2.2.1 Multidimensional scaling of pan-genome coding orthologs

Multidimensional scaling of the pan-genome orthologs using Fripan (Chapter 2, Section 2.7.1) divided strains into five major clusters (Figure 11). Most of the genomes were clustered in the centrally placed 3 clusters, the upper left cluster represents mostly the strains from Germany, Austria and Switzerland, the middle cluster represents two strains which include the completed strain from Germany and nearly complete strain from Switzerland. The placement of these two strains into separate cluster is attributed to more genes being represented by PacBio sequencing compared to draft genomes derived by Illumina sequencing. The third cluster represents the type strain, strains mostly representing other geographical areas, strains of unknown origin and one strain from Germany. Two outliers were observed, a strain form Italy (16S0574) and a strain of unknown origin (S0133-09). The review of these respective sequences indicated contaminating genetic material. Therefore strain S0133-09 was completely excluded and considered for reanalysis. The accessory genes of strain 16S0574 were not considered in assessing the pan-genome structure and composition.

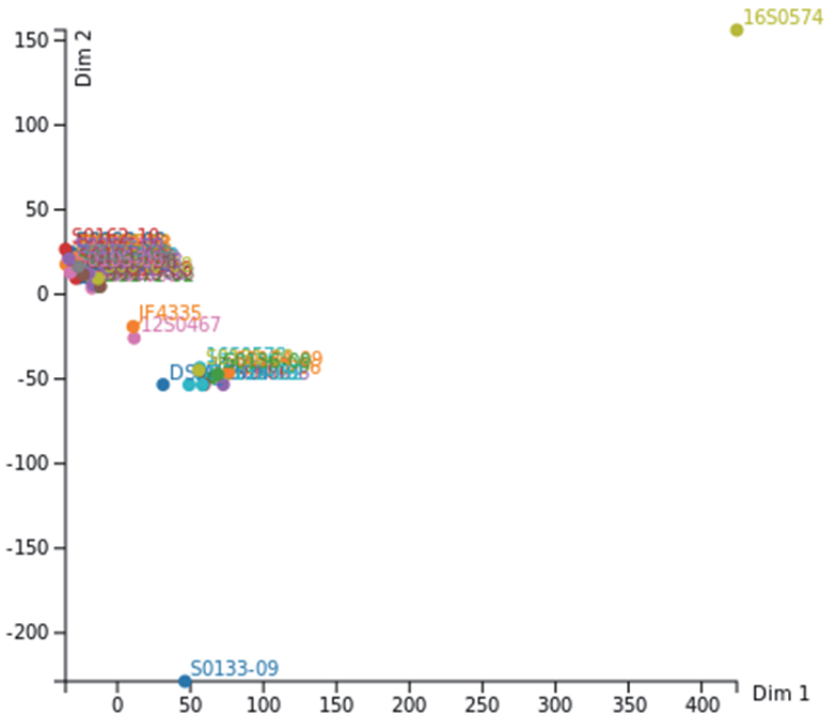


Figure 11: Multidimensional scaling of the pan-genome. Multidimensional scaling of the *Clostridium chauvoei* pan-genome divided the strains into 5 clusters with two outliers (S0133-09 and 16S0574) showing very divergent orthologous gene composition compared to the other three closely related clusters. Here the *C. chauvoei* pan-genome is represented by the 3 clusters placed centrally. The divergent outliers could be attributed to a diverse strain or contamination.

3.2.2.2. Categories of core and accessory genome and the pan-genome structure of *Clostridium chauvoei* strains

To determine the core and accessory genome as well as the pan-genome structure of the 61 *C. chauvoei* strains the Roary pan-genome pipeline was used (Chapter 2.7.1). Here, the core genome represents genes shared by 99% to 100% of the strains and the accessory genome which includes soft core genes (95% to 99% strains), shell genes (15% to 95% strains) and cloud genes (0% to 15% strains). A core genome comprising of 2352 genes was identified within a *C. chauvoei* pan-genome of 3035 genes. The accessory genome included 160 soft core genes, 104 shell genes and 419 cloud genes when all genomes (61 genomes) were analysed. The number of isolates sharing the core and category of accessory genes is depicted as a pie chart (Figure 12 A). Core and accessory genes with large size variations or with unsplit paralog genes (considered as sequence artefacts) were removed from the pan-genome using the roProfle tool. The analysis removed 9 loci identified as multiple alleles and 434 genes were removed because of large allele size variation which also involved 179 genes of the core genome. The roProfle analysis identified 2592 genes as pan-genome, 2173 genes as core genes and 419 genes as accessory genes.

For the studied genomes, the curve for the pan-genome represents the least-squares fit for the function $y = Ax^B + C$ with the best fit obtained with a correlation $r^2 = 0.999$ for A = 15.24 ± 0, B = 0.74, C = 2269.2 ± 0.02 (Figure 12 B). The extrapolated *C. chauvoei* pan-genome size for 100 genomes was found to be 2729 genes. The fitting parameter B value was 0.74, which indicates an open pan-genome. The number of core genes after addition of each new genome was plotted as a function of the number of genomes to create a core genome plot. The curve for the core genome represents the least-squares fit for the function $y = Ae^{Bx} + C$ with the best fit obtained with a correlation $r^2 = 0.996$ for A = 132.39 ± 0.15, B = -0.02, C = 2142.1 ± 0.07 (Figure 12 B). The extrapolated *C. chauvoei* core genome size for 100 genomes was found to be 2160 (Figure 12 C). The curve displaying the acquisition of new genes represents the least-squares fit for the function $y = Ax^B$ with the best fit obtained with a correlation $r^2 = 0.954$ for A = 12.38, B = -0.28. The number of new genes, extrapolated for 100 genomes, was 3 new genes per genome.

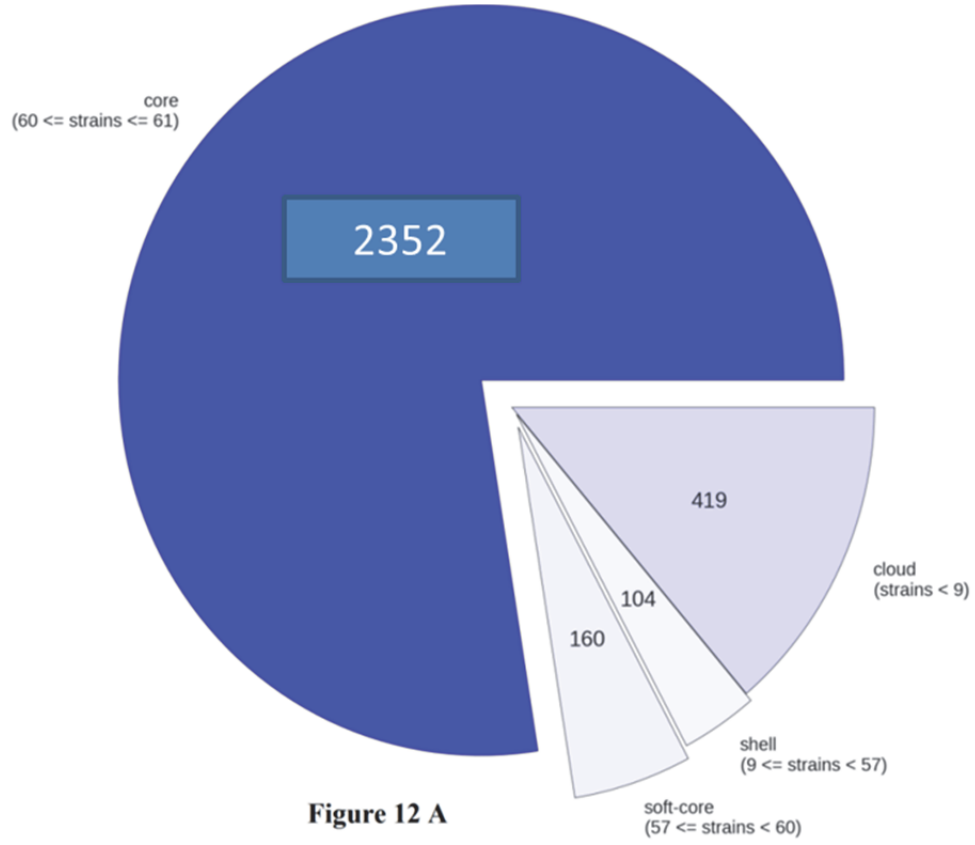


Figure 12 A

Figure 12 A: Pie-chart depicting the core and accessory genome of *Clostridium chauvoei* calculated from 61 strains.

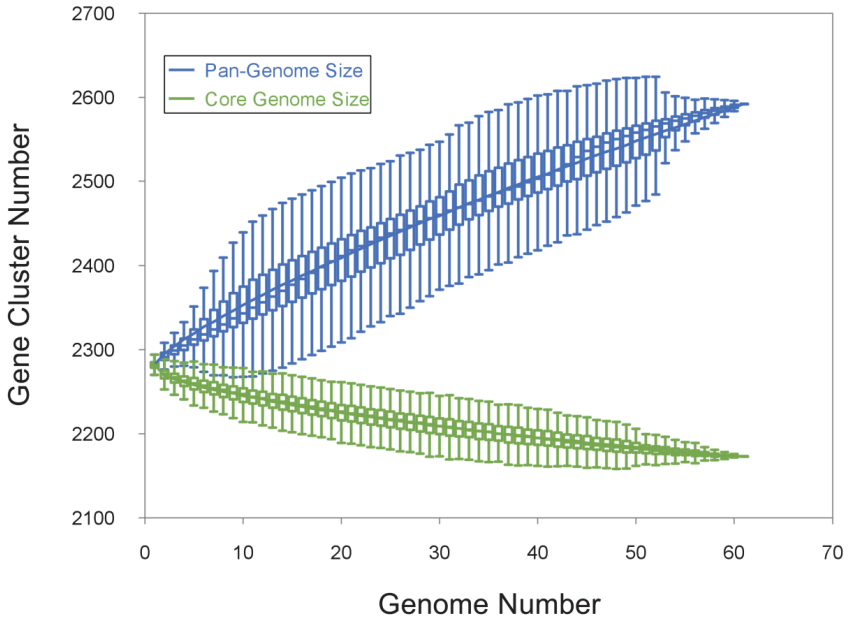
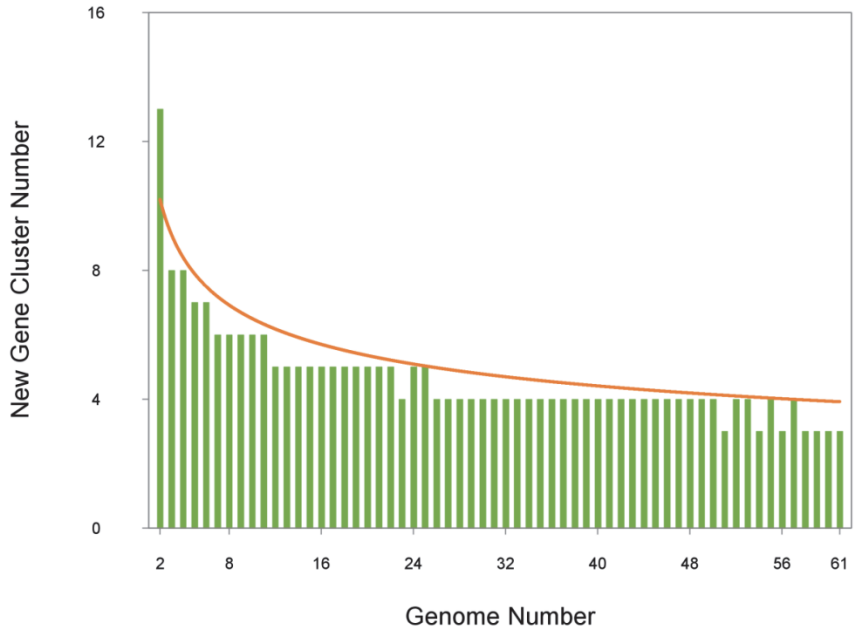


Figure 12 B: Pan-genome and core genome of *Clostridium chauvoei* genomes. For the pan-genome and the core genome analysis, the number of genes are plotted as a function of the number (n) of strains sequentially added.



.

Figure 12 C: New gene identification plot of *Clostridium chauvoei* genomes. Bars represent the number of new genes as the function of the number (n) of strains sequentially added.

3.2.2.3 Core and accessory genome composition

The core and accessory genes of 61 *C. chauvoei* strains were annotated based on gene ontology terms (Chapter 2.7.1). The composition of the core genes analysed with Blast2GO revealed similar functional classes as already predicted for the completed strain (DSM 7528T) with RAST subsystem category distribution (Figure 10).

The accessory genes were categorized according to their biological function (Figure 13). Accessory genes were related to prophages, mobile genetic elements and transposons. The variations attributed to the strains, from different geographical locations, were found to be limited. There was no clear distribution pattern of accessory genes among the strains. The genome of the analysed strains also showed limited variations with respect to phage size and number, but maintained similar composition. The strain BS80-01 harboured 2 phages, located close to each other. The three stains from North Rhine-Westphalia harboured a phage with a deletion of few genes. The variant phages were compared to the type strain phage using progressiveMauve. The alignment revealed a similar genetic structure (Figure 14). The portion of the phage region which was absent in the strains from North Rhine-Westphalia is shown in Table 9. Eight strains (DSM 7528^T, S0134-09, S0013-08, S0014-08, S0015-08, 15S0023, BC103-06 and 12S0470) also showed absence of an insertion element. The region of insertion, presence and absence are depicted based on an alignment created by progressiveMauve from a strain with insertion (12S0467) and two without insertion (DSM 7528^T and 12S0470) shown in Figure 15. 12S0470 (Lower Saxony, Germany) was the only strain form Europe lacking this insertional element. The genes of the insertional element are depicted in Table 10.

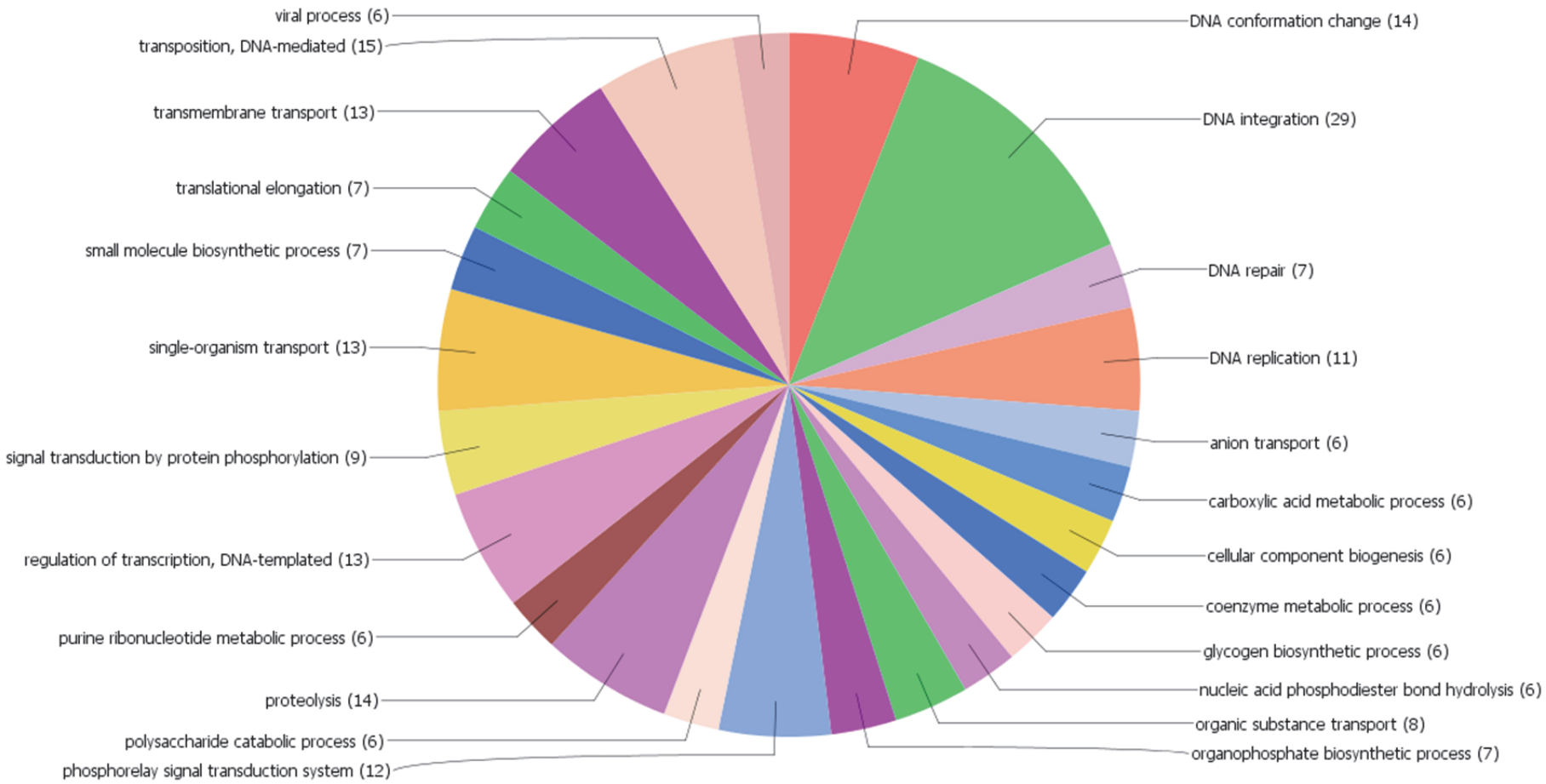


Figure 13: Categories of accessory genes grouped according to biological process. The indicated biological processes mainly involve genes related to prophages, mobile genetic elements and transposons.

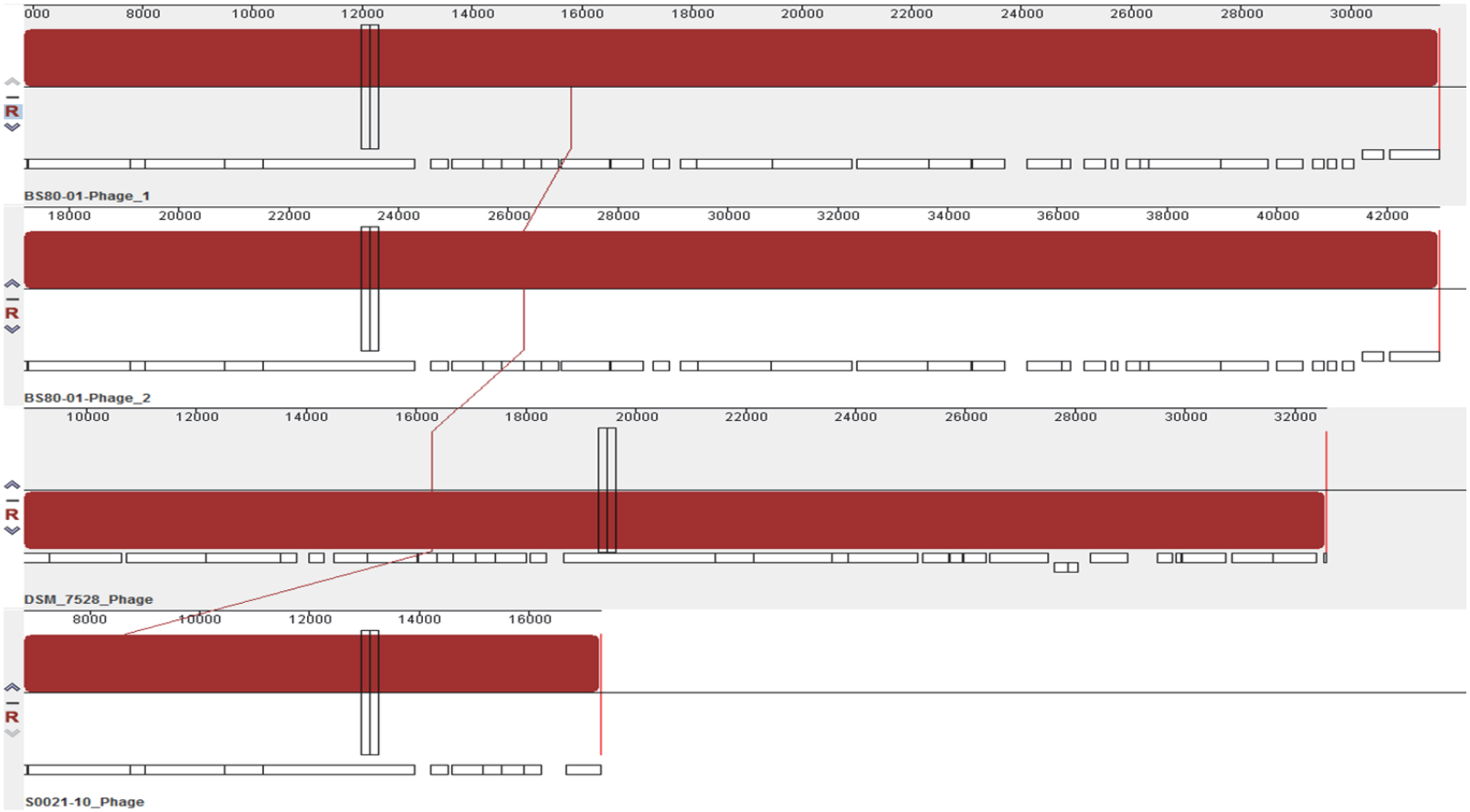


Figure 14: Structure and composition of phages. ProgressiveMauve alignment of phages present among strains which showed variation with respect to number and size of their phages. The upper two sequences represent two phages predicted from strain BS80-01 (BS80-01-Phage_1 and BS80-01-Phage_2). The third represents a phage identified from DSM 7528^{T} (DSM 7528^{T}-Phage) and the fourth represents a smaller sized phage predicted from for the three strains from North Rhine-Westphalia (S0021-10-Phage). All phages show a similar composition. The phage prediction was carried out using PHASTER.

Table 9: Phage region/genes, absent in the strains from North Rhine-Westphalia. The genes representing the region observed for DSM 7528^T strain (gene prediction by the IS saga insertion element prediction tool) which were absent for the strains from North Rhine-Westphalia (S0121-09, S0021-10, S0022-10). The table shows their location within the *C. chauvoei* DSM 7528^T genome, gene product and gene length. The table was created using Geneious 9.0.5.

Min (original sequence)	Max (original sequence)	product	Length
809,294	809,932	IS91 transposase similar to ISShvi3 element fromShewanella violacea	639
810,352	810,930	Tn3 transposase or recombinase similar to ISPa38element from Pseudomonas aeruginosa	579
810,970	811,743	Phage terminase small subunit	774
811,736	813,046	Phage terminase large subunit	1,311
813,139	814,602	Phage portal protein, SPP1 Gp6-like	1,464
814,592	815,950	NAD(+)--arginine ADP-ribosyltransferase EFV	1,359
815,943	816,254	hypothetical protein	312
816,468	816,743	hypothetical protein	276
816,931	817,524	Phage minor structural protein GP20	594
817,540	818,427	hypothetical protein	888
818,471	818,791	Phage gp6-like head-tail connector protein	321
818,788	819,102	hypothetical protein	315

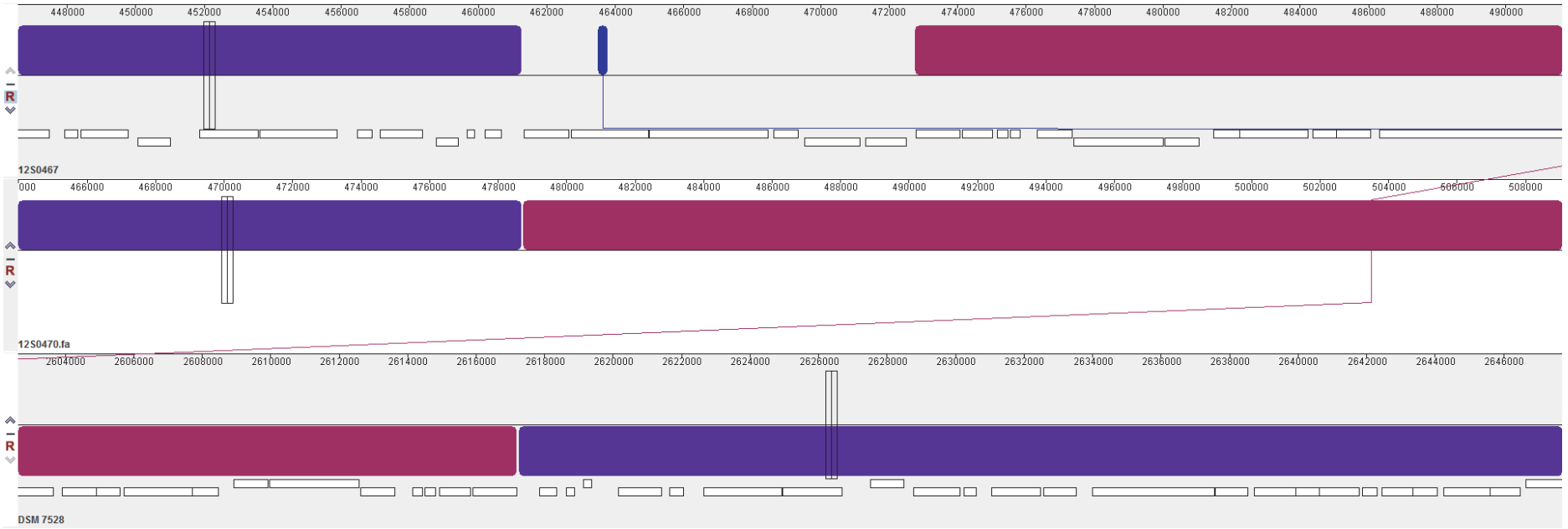


Figure 15: Insertional sequence elements in the genomes. The region of the insertion is depicted based on an alignment created by progressiveMauve using a strain (12S0467) with insertion (upper) and strains (DSM 7528^{T} and 12S0470) without insertions (middle and lower). The insertion is placed between two homologues block shared by the genomes depicted in purple and in red colour. The insertional segment is denoted as space between two blocks and the genes encoding the regions are shown as white bars below the strain 12S0467. The insertional element was located within a 19.4 kb genomic island predicted by Island Viewer tool 4 and is indicated as a blue line.

Table 10: Genes of the insertional element. Table 10 shows the genes found in the insertional element. DSM 7528^T (left) lacks the insertional element and 12S0467 (right) possess the insertional element. The insertion occurred between a duplicated coding sequence for the Lichean -permease IIC component (shown with connecting lines) in the 12S0467 strain and includes 5 genes, i.e., the β-glucanase precursor, an α-L-fucosidase, a sortase family protein, an 1S1182 transposase and a N-actetyl glucosamine repressor. The insertional element annotation was predicted using the ISsaga insertion element prediction tool.

product	Length
hypothetical protein	414
IS256 transposase similar to ISCbo4 element fromClostridium botulinum	1,239
hypothetical protein	252
Guanine deaminase	489
Lichenan permease IIC component	1,293
hypothetical protein	900
Lichenan-specific phosphotransferase enzyme IIB component	318
Lichenan-specific phosphotransferase enzyme IIA component	321
Nitrogen assimilation regulatory protein	1,389
Glutamate synthase [NADPH] small chain	2,595

product	Length
hypothetical protein	414
IS256 transposase similar to ISCbo4 element fromClostridium botulinum	1,239
hypothetical protein	636
Guanine deaminase	489
Lichenan permease IIC component	1,308
Beta-glucanase precursor	2,241
Alpha-L-fucosidase	3,477
Sortase family protein	717
IS1182 transposase similar to ISCpe5 element from Clostridium perfringens	1,638
N-acetylglucosamine repressor	1,194
Lichenan permease IIC component	1,293
hypothetical protein	900
Lichenan-specific phosphotransferase enzyme IIB component	318
Lichenan-specific phosphotransferase enzyme IIA component	321
Nitrogen assimilation regulatory protein	1,389
Glutamate synthase [NADPH] small chain	2,595

3.2.3 Variation in virulence genes

The genome sequence analysis of *C. chauvoei* strain JF4335 and previous studies have characterized primary virulence factors of the pathogen. The identified virulence factors include haemolysins (a haemolysin belonging to the haemolysin III-superfamily, a haemolysin of the XhlA type and the haemolytic leukocidin CctA), sialidase (*nanA*), hyaluronidases (*nagH* and *nagI*) and internalin A protein (Frey and Falquet, 2015, Rychener et al., 2017). The CctA gene shows conservation among the strains under study. There was also no variation observed for the haemolysin III and haemolysin XhlA genes.

The NanA sialidase gene showed lesser conservation, displaying one independent amino acid substitution for strain 15S0023 and for two strains originating from Switzerland (BS169-00, BS171-00). Bacterial hyaluronidases are enzymes that break hyaluronate (a carbohydrate polymer that is part of the extracellular matrix) and thereby help in the initiation and spread of infection. The Hyalurondiase (NagH) gene of strain S0131-09 showed multiple deletions coding for amino acids, additionally four single independent substitutions were observed for four strains (Table 11). The Hyaluronidase (NagI) gene showed two independent substitutions for two strains. N-acetylneuraminate lyase (289 aa) showed two amino acid substitutions for a Swiss strain. Internalin A (one cds) showed a single amino acid variation for one strain. For Glycosyl hydrolase family 20, one independent substitution for two strains and one substitution at the same site for five strains were identified. Similar patterns were observed for the fibronectin type III domain protein cds harbouring one to two independent substitutions in several strains and one substitution at the same site for five strains. In this study the variation with respect to primary virulence factors was studied based on protein alignments of individual genes created using all strains. Some of the protein coding genes revealed to be truncated in some strains. The amino acid substitutions/deletions identified in the virulence factors are depicted in Table 11.

Table 11: Deletion and amino acid variations for primary virulence factors. The table shows the identified virulence factors, coding sequence (cds) lengths, protein variations (either substitution or deletion), number of isolates involved and strain designation respectively.

Primary virulence factors				
Virulence factor	Length (cds)	Predicted protein variation		
		Substitution/ Deletion	Number of Isolates	Strain Designation
Clostridium chauvoei toxin A (CctA)	318	None		
Haemolysin III	223	None		
Haemolysin XhlA	78	None		
Sialidase NanA	1300	Pro to Ala	2	BS169-00, BS171-00
		Ala to Asp	1	1520023
Hyaluronidase H (NagH)	1887	Met to Iso	1	12S0467
		64 aa deletion	1	S0131-09
		Phe to Leu	1	15S0023
		Val to Gly	1	BC93-06
		Ala to Asp	1	DRS014052
Fibronectin type III domain protein	1413	Iso to Leu	2	BS106-02, BS90-02
		Trp to Arg	1	15S0089
		Ala to Glu and Gly to Ser	1	15S0023
		Thre to Iso	1	S0162-10
		Tyr to Ser	5	DSM 7528^{T}, BC103-06, S0013-08, S0014-08, S0015-08
Glycosyl hydrolase family 20	1735	Ala to Ser	1	15S0023
		Arg to Ser	1	S0098-10
		Tryp to Ser	5	DSM 7528^{T}, BC103-06, S0013-08, S0014-08, S0015-08
Hyaluronidase (NagI)	1322	Asp acid to Tyr	1	S0013-08
		Leuc to Val	1	S0041-10
N-acetylneuraminate lyase	290	Ala to Ser and Ala to Asp	1	S0169-00
Internalin A	481	None		
	447	Ser to Tyr	1	15S0023

3.2.4 Population structure and phylogeny

3.2.4.1 Recombination detection

The Parsnp software tool was used to do the core genome alignment including all 64 genomes. The Gubbins and BratNextGen recombination event prediction showed a limited number of recombination events in the 2.4 Mb core genome of the strains. Clusters predicted by BratNextGen among the strains and their respective country of origin and region are shown in Table 12. Recombination events predicted by Gubbins and BratNextGen showed strain and position wise collinearity except for some strains. Table 13 depicts the predicted recombinant regions by respective tools, regions involved and the predicted genes in the region. Most of the predicted recombination events were independent, occurring only in single strains. For two pairs of strains a similar recombination event was predicted, one pair originating from the same farm (S0021-10, S0022-10) and two Swiss strains (BS169-00, BS171-00) by BratNextGen and Gubbins, respectively (Table 13). There were also indications that some of the strains belonging to the same outbreaks/farm showed large pairwise SNP differences with the co-strain obtained from the same outbreak/farm. Some of these differences were identified to be a part of individual recombination regions encoding SNPs.

Table 12: Clusters predicted by BratNextGen among the 64 strains investigated, corresponding country of origin and region are shown.

Cluster	Strain designation	Country	Region
1	16S0579	Germany	Bavaria
	15S0089	Germany	Bavaria
	S0103-08	Austria	Styria
	S0105-08	Austria	Styria
	S0100-08	Austria	Styria
	S0101-08	Austria	Styria
	BC93-06	Unknown	Unknown
2	BS80-01	Germany	Mecklenburg-Western Pomerania
	11S0318	Germany	Mecklenburg-Western Pomerania
	BS79-01	Germany	Mecklenburg-Western Pomerania
	S0260-09	Germany	Lower Saxony
	12S0467	Germany	Lower Saxony
	12S0468	Germany	Lower Saxony
	BS104-02	Germany	Lower Saxony
	12S0466	Germany	Lower Saxony
3	BS91-02	Germany	Lower Saxony
	12S0470	Germany	Lower Saxony
	12S0469	Germany	Lower Saxony
	BS107-02	Germany	Lower Saxony

	BS92-02	Germany	Lower Saxony
	BS109-02	Germany	Lower Saxony
	12S0465	Germany	Lower Saxony
	12S0464	Germany	Lower Saxony
	BS108-02	Germany	Lower Saxony
	S0162-10	Germany	Lower Saxony
	BS94-02	Germany	Lower Saxony
	15S0008	Germany	Bavaria
	15S0088	Germany	Bavaria
	13S0854	Germany	Baden-Württemberg
	13S0851	Germany	Baden-Württemberg
	S0008-10	Germany	Schleswig-Holstein
	BS105-02	Germany	Lower Saxony
	S0098-08	Austria	Tyrol
	JF4335	Switzerland	Unknown
	S0131-09	Unknown	Unknown
	BC138-06	Unknown	Unknown
4	16S0578	Canada	Unknown
	16S0574	Italy	Perugia
	S0014-08	Austria	Tyrol
	S0015-08	Austria	Tyrol
	S0013-08	Austria	Tyrol
	S0134-09	Unknown	Unknown
	BC103-06	Unknown	Unknown
	DSM_7528	Unknown	Unknown
5	S0040-08	Germany	Schleswig-Holstein
	S0041-08	Germany	Schleswig-Holstein
6	11S0316	Germany	Lower Saxony
	11S0315	Germany	Lower Saxony
	12S0472	Germany	Lower Saxony
	12S0471	Germany	Lower Saxony
7	BS106-02	Germany	Lower Saxony
	BS90-02	Germany	Lower Saxony
	BS169-00	Switzerland	Unknown
	BS171-00	Switzerland	Unknown
	S0099-08	Austria	Styria
	S0102-08	Austria	Styria
8	S0021-10	Germany	North Rhine-Westphalia
	S0121-09	Germany	North Rhine-Westphalia
	S0022-10	Germany	North Rhine-Westphalia
9	15S0023	South Africa	Unknown
10	S0132-09	Unknown	Unknown
11	DRS014052	Unknown	Unknown

Table 13: Predicted homologous recombination events. Table 13 shows the region of the core genome (start and end nucleotide position of the core genome alignment), predicted protein coding regions by both tools (BratNextGen and Gubbins), SNP number and r/m ratio (ratio of rates at which nucleotides become substituted as a result of recombination and mutation) by Gubbins. Both tools (BratNextGen and Gubbins) showed collinearity with respect to the identified recombination region in individual strains. Several recombination event predictions were also independently predicted by both tools.

	BratNextGen			Gubbbins				
Strains	start	End	coding genes	start	end	SNP	coding genes	r/m
JF4335	289690	291658	Two-component sensor histidine kinase, Muramidase, Vancomycin resistance protein	291139	291142	4		0.44
JF4335	978140	978995	Hypothetical protein, Hypothetical protein	978411	978471	7	Hypothetical protein	
BS109-02	1132062	1133112	PolC	1132598	1132698	36	PolC	1.7143
S0008-10	513926	515023	BFD-like [2Fe-2S] binding domain protein, Hypothetical protein, Cold-shock protein	514215	514283	35	Hypothetical protein	0.897436
11S0316	1094626	1095180	Subtilase family protein	1094677	1094808	27	Subtilase family protein	1.688889
11S0316	1160987	1162026	50S ribosomal protein L32, Metal-binding protein,	1161234	1161331	36	Metal-binding protein	
11S0316	1650039	1650950	Hypothetical protein, ABC transporter ATP-binding protein	1650091	1650152	13	Hypothetical protein	
S0260-09	571807	572385		571951	572139	43		0.977273
BS104-02	193802	194701		194139	194217	32		0.969697

11S0318	1941368	1945172	ECF RNA polymerase sigma-E factor, Hypothetical protein CDS,Sensor histidine kinase YpdA, Putative response regulatory protein,	1944170	1944191	6	Sensor histidine kinase YpdA, Putative response regulatory protein,	0.375
BS169-00	1629584	1630341	Sugar ABC transporter permease CDS, Acetylneuraminate ABC transporter permease	1629928	1629999	19	Acetylneuraminate ABC transporter permease	0.703704
BS171-00				1629930	1629998	11	Acetylneuraminate ABC transporter permease	0.392857
BS169-00, BS171-00				1629930	1629998	11	Acetylneuraminate ABC transporter permease	0.186441
S0021-10	676343	679601	Hypothetical protein, Aminopeptidase, Glutaminase A					
S0022-10	676343	679601	Hypothetical protein, aminopeptidase, Glutaminase A					
15S0089	903177	904346	Hypothetical protein, Hypothetical protein					
15S0089	1021228	1021315	Hypothetical protein					

3.2.4.2 Maximum likelihood core genome phylogeny

A phylogenetic tree was deduced from the core genome post-filtered polymorphic sites involving 64 strains, exported from Gubbins, by maximum likelihood (ML) method using RAxML 8.2.8 with the general time reversible nucleotide substitution model (GTR-GAMMA). The tree generated was midpoint rooted (Figure 16). The analysis identified most of the isolates were clustering based on geographical origin and strong boot strap support values (100%) were highly prominent for strains from the same farm/outbreak/strain. The strains from geographical regions other than Europe (South Africa) and few strains of unknown origin did not form any cluster with strong bootstrap support and hence are obviously diverse (Figure 16).

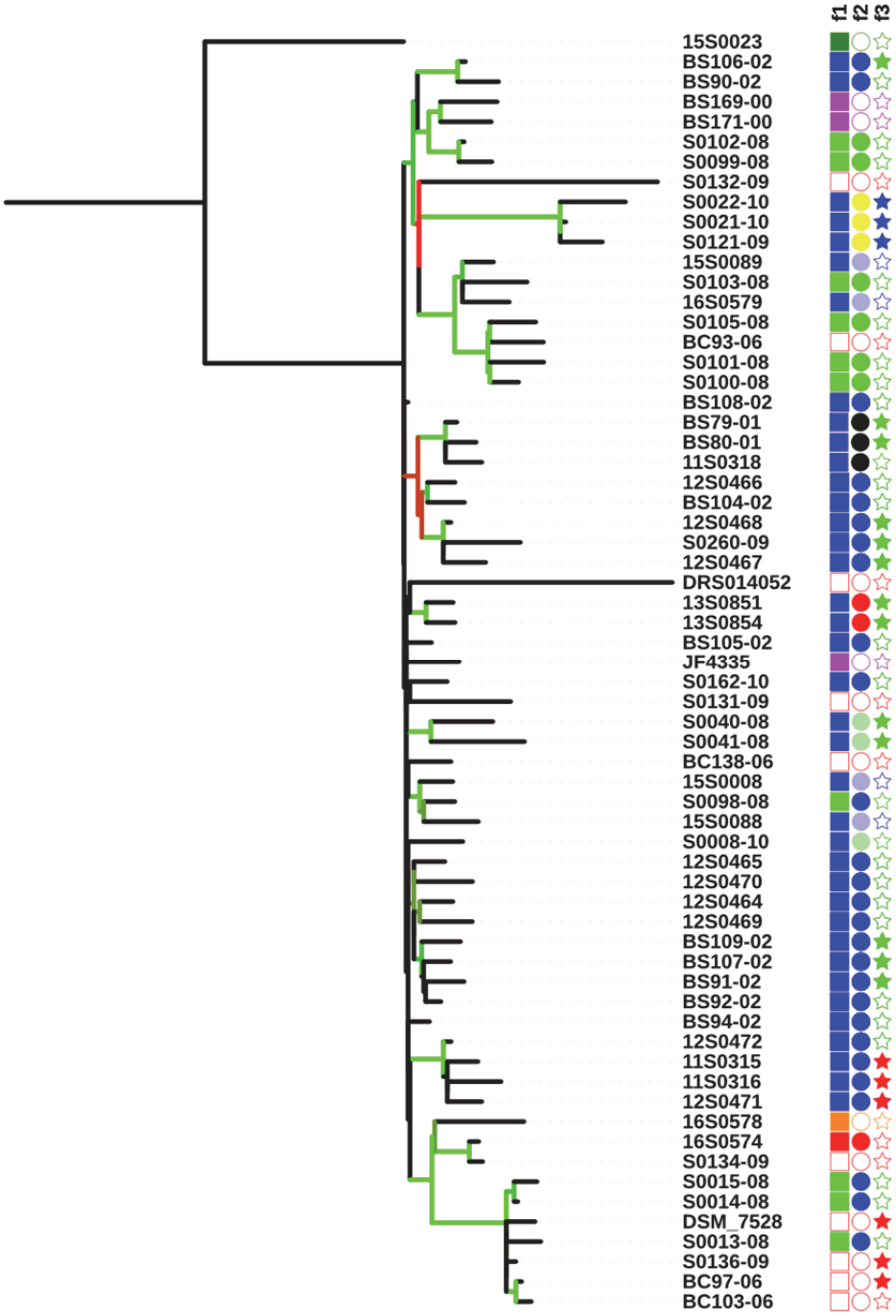


Figure 16: Phylogenetic tree based on the core genome of *Clostridium chauvoei* strains. The phylogenetic tree shows the bootstrap support values at the internal nodes ranging from 95 (red) to 100 (green). Filled square (f1), round (f2) and star (f3) shaped symbols represent the country, region and strain/outbreak/farm level relationship among the strains whereas empty boxes represent unknown information. The f1 column colours correspond to blue (Germany), green (Austria), purple (Switzerland), red (Italy), dark green (South Africa) and orange (Canada). In f2 strains from Germany are blue (Lower Saxony), black (Mecklenburg West-Pomerania), light blue (Bavaria), yellow (North Rhine-Westphalia), red (Baden-Württemberg) and light green (Schleswig-Holstein). The strains from Austria are green (Styria) and blue (Tyrol) and a single Italian strain was from Perugia (red). In the f3 column special groups are indicated: red (BC 97-06, S0136-09 and DSM 7528^T - same strain, and 11S0315, 11S0316, 12S0471 - from the same animal), green (12S0467, 12S0468 and S0260-09 - same outbreak but different animals, BS 109-02, BS107-02 and BS 90-00-same outbreak but different animals, S0040-08 and S0041-08 –same outbreak but from different animals, 13S0851 and 13S0854 – same outbreak and different animals) and blue (S0121-09 (year 2009), S0021-10 (year 2010) and S0022-10 (year 2010) - strains from same farm in two successive years)).

3.2.4.3 Pan-genome SNP analyis, phylogeny and clustering

Pan-genome analysis was carried out employing the kSNP3 package. The optimal k-mer value was identified to be 21 for the genomes under study. A total of 2294 SNPs were identified in the 62 *Clostridium chauvoei* genomes (two laboratory strains of DSM 7528^T where excluded in the analysis), of which core SNPs where 2000 and non-core SNPs where 254. Among the SNPs occurring in protein coding genes 1668 correspond to non-synonymous (NS) substitutions and 491 were synonymous (S) types resulting in high NS/S ratio of 3.397. A total of 39 SNPs were homoplastic which corresponds to SNPs shared by groups of genomes that are not clustered in the consensus maximum parsimony (MP) tree used for clustering the strains. The parsimony tree based on the pan-genome SNPs indicated 10 clusters based on the presence of at least 5 unique SNPs shared among the strains within a cluster. The strains involved in each cluster indicate genetic relatedness and have unique SNPs defining them. The clusters showed geographical relationships among each other (Figure 17). The unique SNPs shared by a cluster had a maximum value of 76 (three strains from North Rhine-Westphalia: Cluster 4). The SNPs uniquely identified in at the inner nodes within clusters ranged from 0 to 43. The maximum value of 43 was obtained for the inner node of Cluster 9 and involved 4 strains (3 strains from Austria and the type strain). Cluster 9 was identified to involve the most divergent strains, such as the type strain, 3 strains from Austria, the Canadian strain, a strain of unknown origin and the Italian strain. The strains from Germany also showed relatedness with strains from Austria (Cluster 3 and Cluster 7) (Figure 17). These strains were from Bavaria which shares geographical relatedness with Austria. Similarly, two strains from Austria clustered with two strains from Switzerland (Cluster 1) and indicate geographical relativity. Other clusters mostly involved strains from different regions of Germany, i.e. Lower Saxony (Cluster 10), Lower Saxony and Mecklenburg-Western Pomerania (Cluster 5), North Rhine-Westphalia (Cluster 4), Schleswig-Holstein (Cluster 6) and Baden-Württemberg (Cluster 9) (Figure 17). The allele specific SNPs ranged from 1 to 207. The strains harbouring allele specific SNPs above 100 are S0132-08, 1520023 and DRS104052, respectively (Figure 17). The parsimony tree indicating the clusters, the unique SNPs defining these clusters and the allele specific SNP

number are shown in the Figure 17. To get an insight into the pattern of genetic divergence of the species, the unique SNPs present in all identified clusters were annotated based on gene ontology terms. This annotation analysis showed that genes with unique SNPs were involved in various biological processes. Significantly the involved genes belonged to biological processes of transmembrane transport, carbohydrate metabolism and phosphorylation (Figure 18).

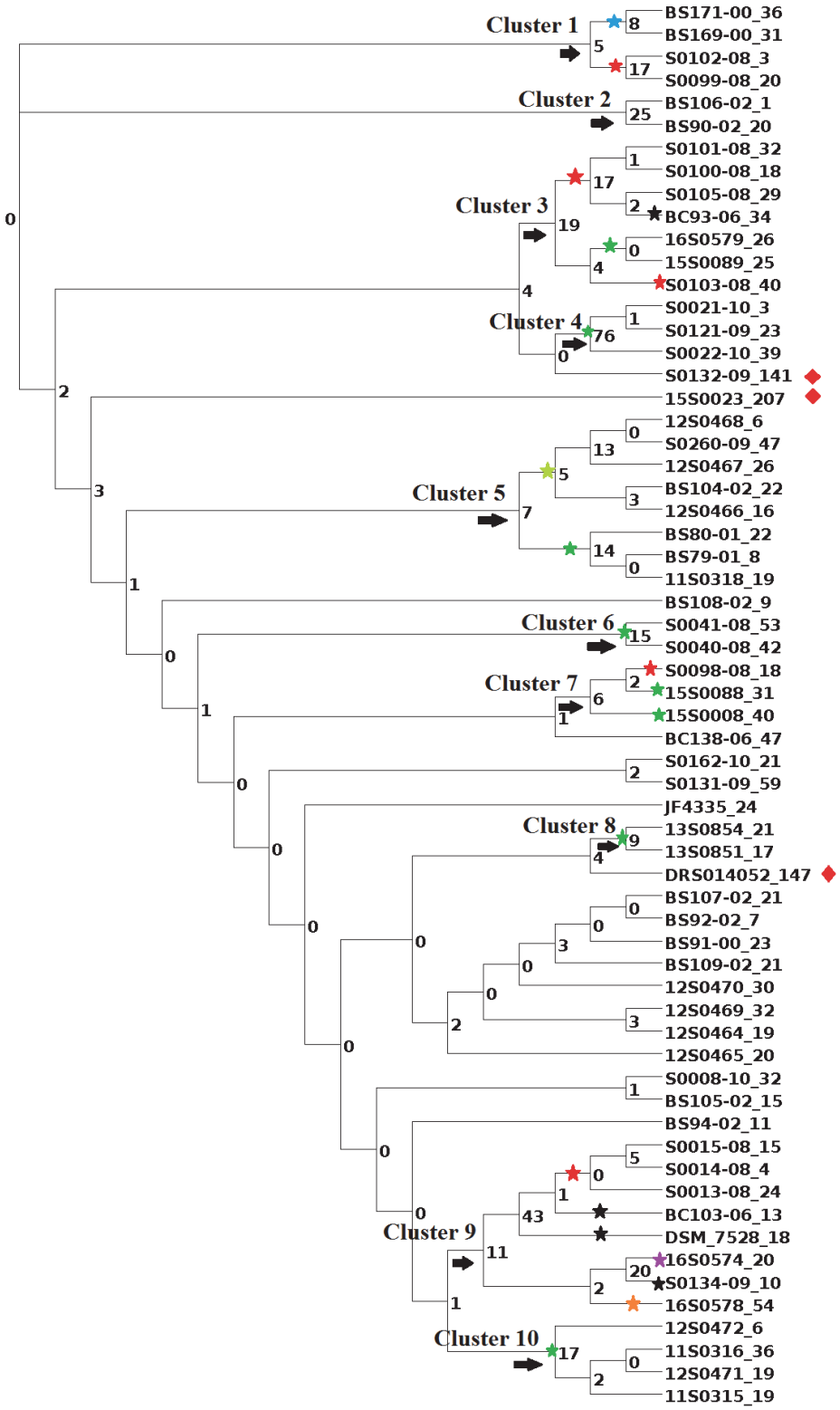

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
BS171-00_36
BS169-00_31
S0102-08_3
S0099-08_20
BS106-02_1
BS90-02_20
S0101-08_32
S0100-08_18
S0105-08_29
BC93-06_34
16S0579_26
15S0089_25
S0103-08_40
S0021-10_3
S0121-09_23
S0022-10_39
S0132-09_141
15S0023_207
12S0468_6
S0260-09_47
12S0467_26
BS104-02_22
12S0466_16
BS80-01_22
BS79-01_8
11S0318_19
BS108-02_9
S0041-08_53
S0040-08_42
S0098-08_18
15S0088_31
15S0008_40
BC138-06_47
S0162-10_21
S0131-09_59
JF4335_24
13S0854_21
13S0851_17
DRS014052_147
BS107-02_21
BS92-02_7
BS91-00_23
BS109-02_21
12S0470_30
12S0469_32
12S0464_19
12S0465_20
S0008-10_32
BS105-02_15
BS94-02_11
S0015-08_15
S0014-08_4
S0013-08_24
BC103-06_13
DSM_7528_18
16S0574_20
S0134-09_10
16S0578_54
12S0472_6
11S0316_36
12S0471_19
11S0315_19

Figure 17: Parsimony tree based on *Clostridium chauvoei* pan-genome SNPs. The node labels are showing the number of SNP alleles that are specifically present in all descendants of that node. The strain specific allele counts are indicated by an underscore symbol (_) after the strain number. The phylogenetic tree is unrooted and branch lengths are expressed in terms of changes per number of SNPs. Based on SNP alleles greater than 5 (shown in the nodes and also indicated by arrow sign), some of the strains were grouped into 10 clusters. The strains involved in these clusters and its internal nodes are indicated by star symbols in different colour to specify the geographical origin (blue: Switzerland, red: Austria, green: Germany, pink: Italy, orange: Canada, black: unknown). Cluster 5 was also identified to involve strains from two geographical regions indicated by light (Lower Saxony) and dark green (Mecklenburg-Western Pomerania) within Germany. The outer diamond shaped red symbol indicates strains with more than 100 allele specific SNPs.

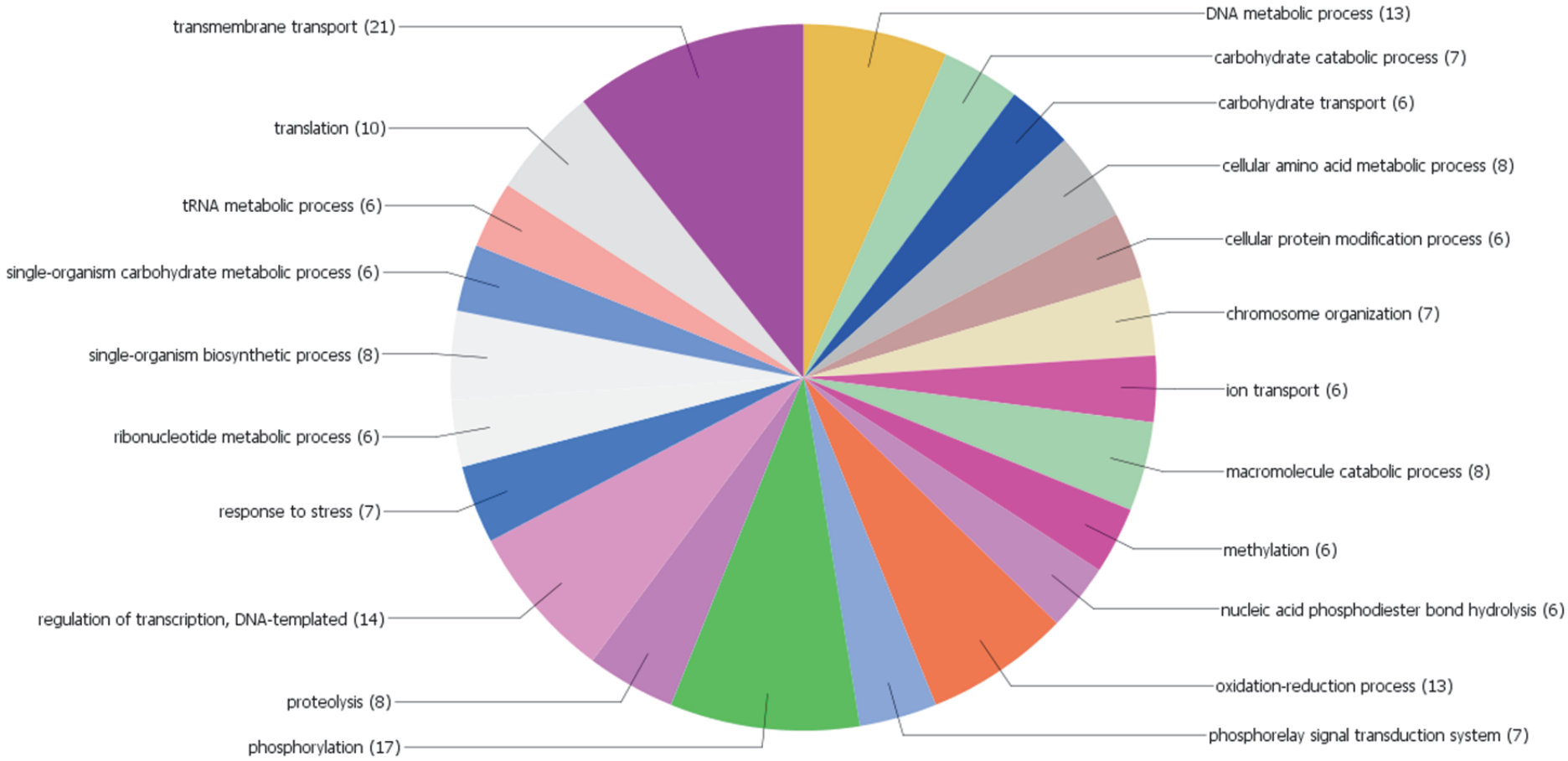


Figure 18: Pie-chart depicting the classification of genes involved in pan-genome SNP clustering. This analysis used the genes involved in the pan-genome phylogeny (harbouring SNPs). Genes were grouped based on gene ontology annotation (Blast2GO) for various biological processes. The significantly involved groups based on sequence count were genes involved in transmembrane transport and phosphorylation.

3.2.4.4 SNP analysis (Reference based mapping)

Read mapping and SNP detection was performed using the Snippy pipeline. All strains for which reads were available were mapped to *C. chauvoei* strain 12S0467 to identify SNPs. The number of identified SNPs was found to be 1755 among 61 strains. 1175 SNPs were detected among 51 strains of European origin and 835 for 38 strains from Germany. Pairwise SNP difference analysis within and among groups based on geographical origin was carried out for all strains, strains of European origin and from Germany respectively. The intra-group pairwise SNP difference median values were found to be lower for the strains from Europe and were 74 as compared to 144 for strains from unknown origin. Also the inter-group comparison shows the median pairwise SNP distance is smaller (57) for Europe vs unknown, Europe vs Canada and Canada vs unknown as compared to Europe vs South Africa and South Africa vs unknown. This indicates that the SNPs are more shared among the European, Canadian and possibly few strains from unknown origin as compared to the South African strain which is identified as the most divergent strain from a geographical point of view (Figure 19 A).

The strains originating from Europe do not show many variations with respect to inter-group comparisons of median SNP difference values ranging from 50 to slightly above 100 (Figure 19 B). But the intra-group comparison shows the 10 Austrian strains to be highly variable showing a median pairwise SNP difference value of 97 as compared to 59 occurring in 38 strains originating from Germany (Figure 19 B). The intra-group pairwise SNP median values for the strains originating from different regions of Germany were ranging from 0 to 67 Figure 19 C). High variation was observed for two other strains belonging to the same outbreak from Schleswig-Holstein which also showed SNP variation for the hyaluronidase gene, a potential virulence factor of the pathogen. Following Schleswig-Holstein the most variable strains were from Bavaria. Inter-group comparison shows higher pairwise SNP median difference values for Bavaria vs North Rhine-Westphalia, Baden-Württemberg vs. North Rhine-Westphalia, Lower Saxony vs North Rhine-Westphalia, Mecklenburg-West Pomerania vs North Rhine-Westphalia and Schleswig-Holstein vs North Rhine-Westphalia. The repeated occurrence of higher median inter-group SNP difference values for North Rhine-Westphalia strains indicates that these stains are the most divergent strains among the strains from Germany (Figure 19-C).

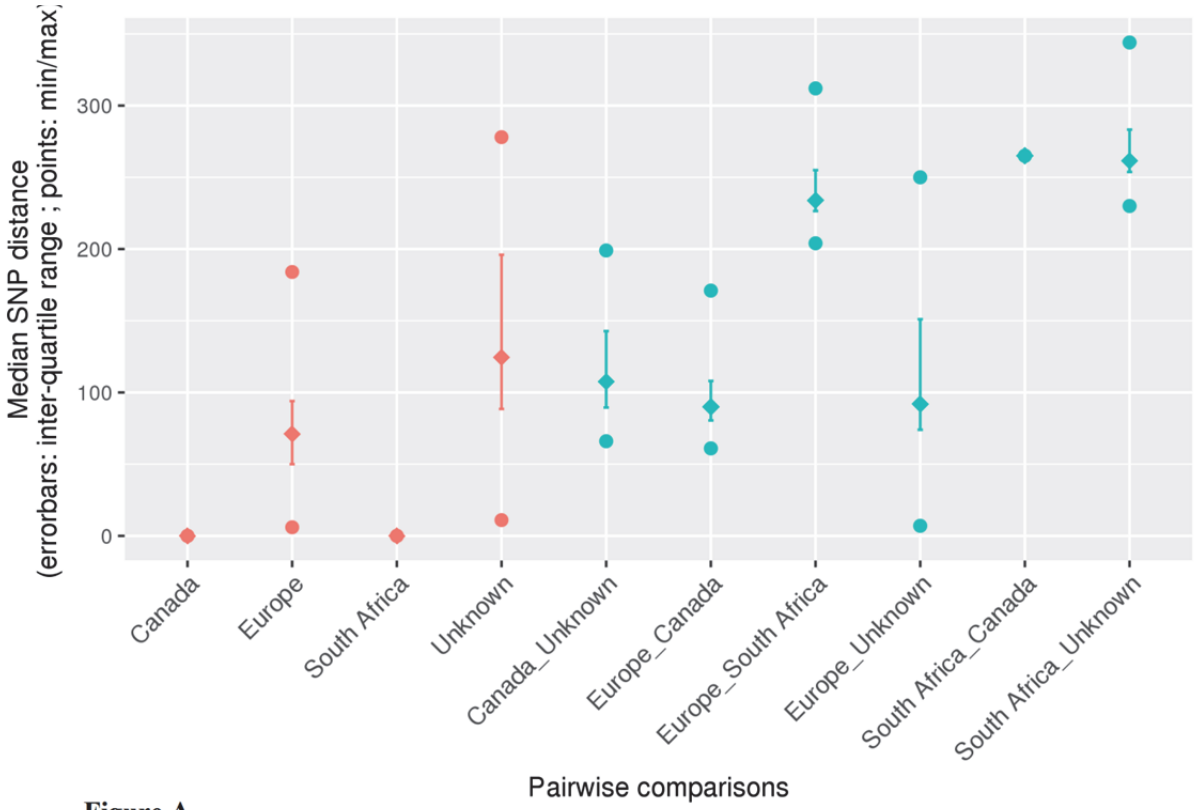


Figure A

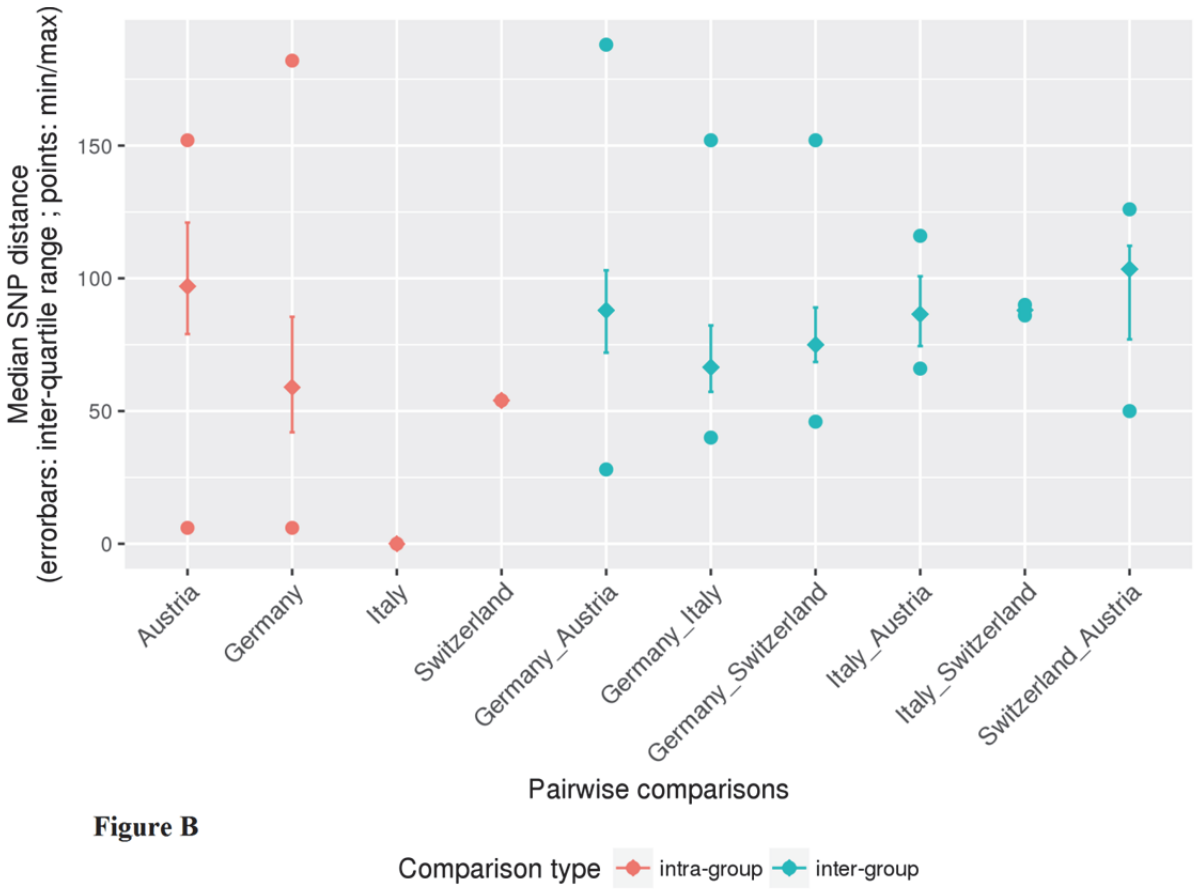


Figure B

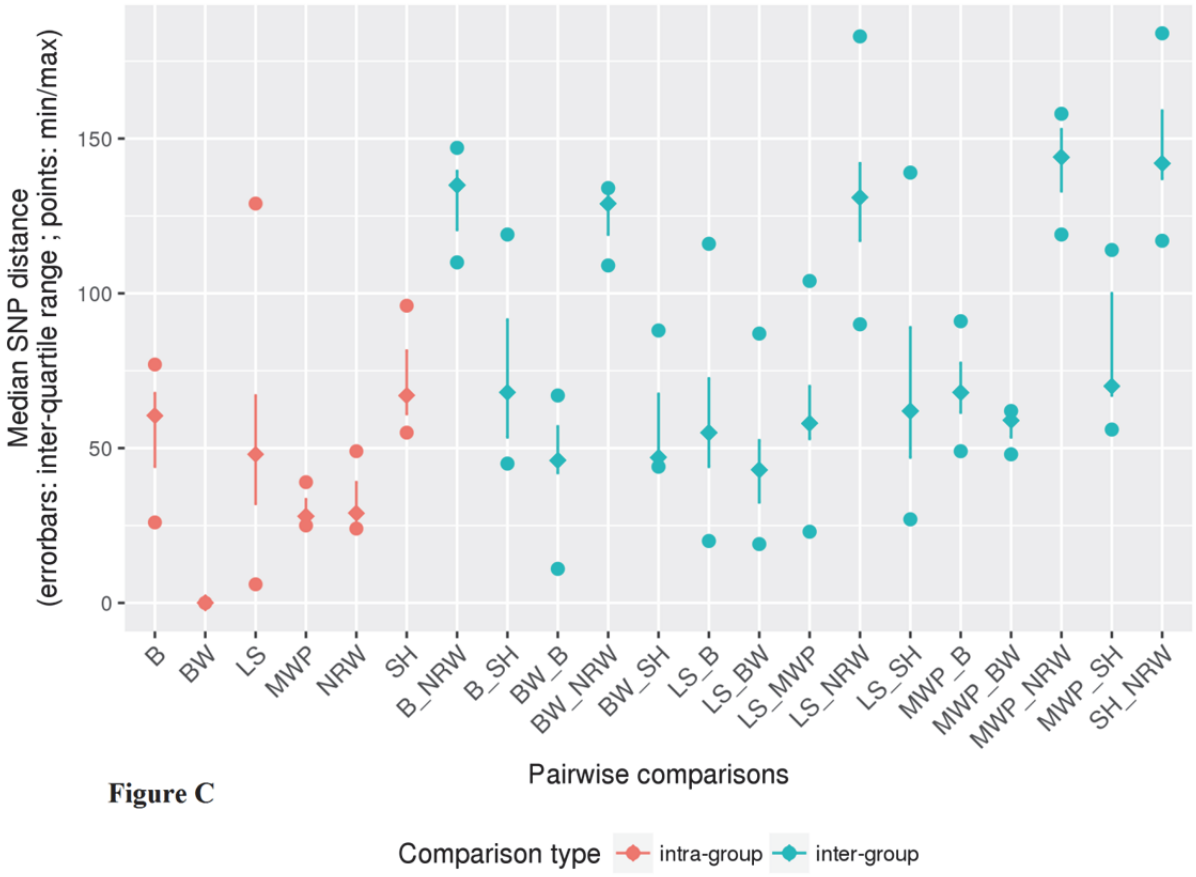


Figure 19: Pairwise SNP divergence of strains based on geographical origin. The Y-axis depicts the number of SNP differences; error bars indicate one standard deviation above and below the mean and points represent the minimum and maximum values. Figure A: Among all (61), group 1 – Canada (1), group 2 - all European strains (51), group 3- South African (1), and group 4 -unknown origin (8). Figure B: Among Europe (51), group 1- Austria (10) group 2 - Germany (38), group 3- Italy (1) and group 4 -Switzerland (2). Figure C: Among Germany (38), Group 1 – Bavaria (B) (4), Group 2- Baden-Württemberg (BW) (2), group 3- Lower Saxony (LS) (23), group 4 – Mecklenburg-Western Pomerania (MWP) (3), group 5- North Rhine-Westphalia (NRW) (3) and group 6 - Schleswig-Holstein (SH) (3).

SNP variations for the same strain with different laboratory passages and/or sources were also investigated. The three laboratory collection strains originating from the ATCC 10092^{T} strain designated as DSM 7528 ^{T}, S0136-09 and BC 97-06 were available. Of these, the DSM 7528^{T} strain was sequenced using PacBio and the other two strains using Illumina sequencing technology. The pairwise SNP difference studies show around 25 core SNPs among the strains, but the strains which were sequenced by Illumina showed only 6 SNPs differences among them (Figure 20-A). Similarly three strains originating from a single farm in North Rhine-Westphalia belong to two successive outbreaks were investigated. One strain (S0121-09) was isolated in the year 2009 and two strains (S0021-10 and S0022-10) were isolated in the year 2010. 54 core SNP differences were detected among the three strains of which strain S0121-09 had 25 SNP differences with S0021-10, whereas 52 SNP differences were observed with S0022-10. The two strains from 2010 maintained 30 SNP differences (Figure 20-B). SNP differences observed among strains from the same outbreak/animal showed variable results with lower variation of only 7 SNPs for the strains from Baden-Württemberg, but were

higher with 68 and 53 SNPs for outbreak strains from Schleswig-Holstein and Lower Saxony, respectively. During the recombination identification using Gubbins and BratNextGen of the core genome alignment file generated from Parsnp, recombination events for some strains (S0021-10, S022-10, 11S0316) were identified, but not from the co-strain recovered from the same outbreak (S0121-09, 11S0315).

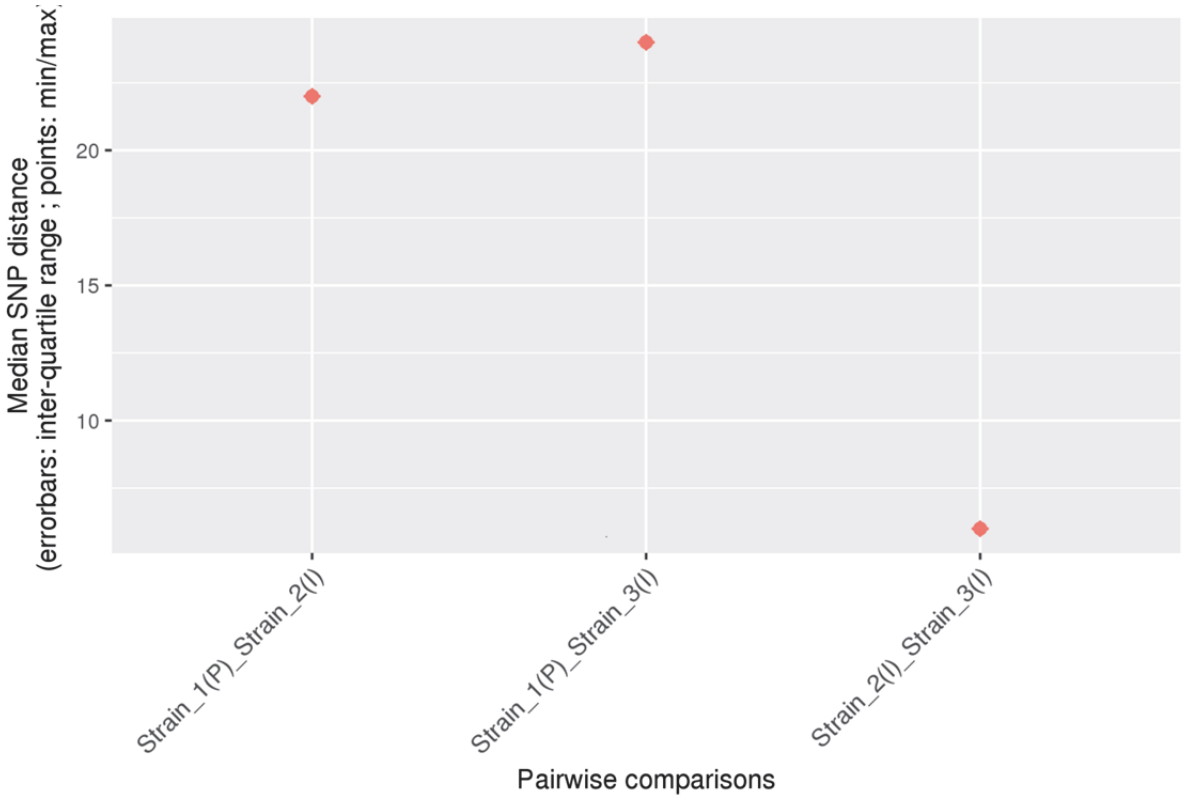


Figure A

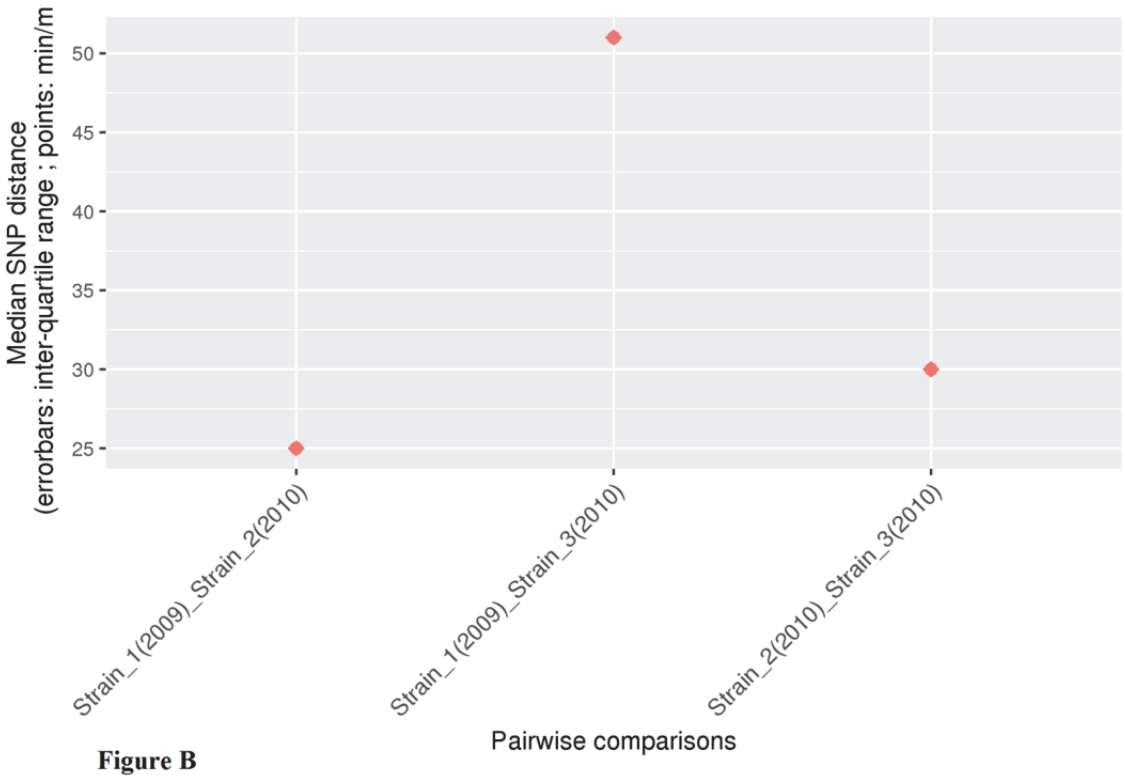


Figure 20: Pairwise SNP divergence of strains based on strains/outbreak. The Y-axis depicts the number of SNP differences and the X-axis indicates the strains. Figure A: laboratory isolates of ATCC 10092^{T}: Strain_1 (P): DSM 7528^{T} (Pac Bio sequencing), Strain_2 (I): S0136-09 (Illumina sequencing) and Strain_3 (I): BC 97-06 (Illumina sequencing). Figure B: Strains originating from the same farm in consecutive years: Strain_1 (2009): S0121-09 (isolated in the year 2009), Strain_2 (2010): S0021-10 (isolated in the year 2010), Strain_3 (2010): S0022-10 (isolated in the year 2010).

3.2.5 Within-host and outbreak strain variations

Within-host strain variations were observed in three strains isolated from single host (11S0315, 11S0316 and 12S0471). SNPs present in these strains were identified and were also compared with a possibly last common ancestor strain (12S0474) isolated from the same farm one year before using two different variant calling pipelines to confirm the SNP calling was accurate. SNPs were called by read mapping using the raw reads of the strains and the completed *C. chauvoei* field strain 12S0467 as the reference strain for SNP calling and annotation transfer. Both tools predicted the 62 SNPs and 2 SNPs differing between the tools. Some of the SNPs predicted where in noncoding genes whereas most of the SNPs where non-synonymous (NS) and hence were indicating diversifying selection occurring inside the host (Table 14). The SNPs uniquely shared by all three strains with the last common ancestor were mostly of the non-synonymous type (Table 15).

SNPs occurring in strains coming from the same outbreak, but from different animals where compared based on 5 pairs of outbreak strains (12S0467 and 12S0468, S0040-08 and S0041-08, BS79-01 and BS80-01, S0021-10 and S0022-10, 13S0851 and 13S0854). Only outbreak strains which clustered together in the kSNP3 clustering analysis were considered. The groups were identified to share none or limited SNPs between them and belonged to different

geographical regions (cluster 10,4,5,6 and 8 in Figure 17). The analysis identified high rates of non-synonymous substitutions in outbreak strains originating from the same farm (Table 16). The genes showing variation were matching to proteins identified in *Staphylococcus aureus* undergoing purifying selection within the host (Golubchik et al., 2013). The protein genes matching those of *Staphylococcus aureus* were; a putative PTS transport system IIB component, an ABC transporter ATP-binding protein, a putative cardiolipin synthetase, a sensor histidine kinases, a hyaluronidase, putative DNA binding proteins and N-acetylmuramoyl-L-alanine amidase (Similar genes are highlighted in Table 14).

Table 14: SNPs identified in three strains (11S0315, 11S0316 and 12S0471) isolated from one animal. Table 14 shows the SNPs identified in 3 strains (11S0315, 11S0316 and 12S0471) from one animal, corresponding positions in the reference strain (12S0467), the locus tag specifying the encoded protein, the product involved and the effect caused. Strains 11S0315, 11S0316 and 12S0471 are designated as 1, 2 and 3 respectively in the table. Genes highlighted in bold have been reported for within-host variations in *Staphylococcus aureus.*

Position	1	2	3	Locus tag	Product	Effect
4661	**A**	**C**	**C**	**BTM20_00030**	**DNA gyrase subunit B**	**missense_variant c.171C>A p.Phe57Leu**
97184	C	C	A	BTM20_00425	rRNA large subunit methyltransferase I	missense_variant c.1079G>T p.Arg360Ile
101623	G	T	G	BTM20_00440	isoleucine--tRNA ligase	missense_variant c.1897G>T p.Asp633Tyr
264454	C	A	C	BTM20_01335	50S ribosomal protein L13	stop_gained c.375C>A p.Tyr125*
352733	C	A	C	BTM20_01695	glucose-1-phosphate adenylyltransferase subunit GlgD	missense_variant c.218C>A p.Pro73His
398014	A	C	C	BTM20_01910	MBL fold metallo-hydrolase	stop_gained c.1372G>T p.Glu458*
449155	G	G	T	BTM20_02130	23S rRNA (uracil-5-)-methyltransferase RumA	missense_variant c.757G>T p.Val253Phe
504248	G	T	G	BTM20_02310	sodium-independent anion transporter	missense_variant c.542G>T p.Gly181Val
530835	**T**	**G**	**G**	**BTM20_02435**	**glutamine ABC transporter ATP-binding protein**	**missense_variant c.537G>T p.Met179Ile**
663865	G	G	T	BTM20_02975	mannose-1-phosphate guanylyltransferase	stop_gained c.595G>T p.Gly199*
675547	G	T	G	BTM20_03035	hypothetical protein	missense_variant c.424G>T p.Asp142Tyr
675690	G	T	G	BTM20_03035	hypothetical protein	missense_variant c.567G>T p.Lys189Asn
740123	C	A	C	BTM20_03335	hypothetical protein	missense_variant c.746G>Tp.Gly249Val
813400	C	A	C	BTM20_03640	hypothetical protein	missense_variant c.1048C>A p.His350Asn

820177	A	C	C	BTM20_03665	nucleoside-diphosphate sugar epimerase	missense_variant c.1298C>A p.Ser433Tyr
825042	G	T	G	BTM20_03690	hypothetical protein	stop_gained c.121G>T p.Glu41*
844693	T	G	G	BTM20_03785	spore germination protein	missense_variant c.517G>T p.Ala173Ser
868333	C	A	C	BTM20_03895	hypothetical protein	missense_variant c.266C>A p.Ser89Tyr
868460	C	A	C	BTM20_03895	hypothetical protein	synonymous_variant c.393C>A p.Gly131Gly
869607	**G**	**G**	**T**	**BTM20_03900**	**PTS glucose transporter subunit IIB**	**missense_variant c.584G>T p.Gly195Val**
873349	G	T	G	BTM20_03920	N-acetylneuraminate lyase	initiator_codon_variant c.3G>T p.Met1?
977308	G	G	T	BTM20_04400	hypothetical protein	stop_gained c.706G>T p.Glu236*
1045962	G	T	G	BTM20_04715	deoxyuridine 5'-triphosphate nucleotidohydrolase	missense_variant c.263G>T p.Arg88Ile
1052600	G	T	G	BTM20_04750	aminotransferase class V	stop_gained c.79G>T p.Glu27*
1123430	C	A	C	BTM20_05105	flagellar biosynthetic protein FliP	stop_gained c.63C>A p.Tyr21*
1128335	G	G	T	BTM20_05120	flagellar biosynthesis protein FlhA	stop_gained c.2044G>T p.Glu682*
1144201	C	A	C	BTM20_05195	hypothetical protein	missense_variant c.69G>T p.Met23Ile
1151894	C	A	C	BTM20_05245	deoxyribose-phosphate aldolase	missense_variant c.392C>A p.Thr131Asn
1198673	A	C	C	BTM20_05465	DNA primase	missense_variant c.1345C>A p.Pro449Thr
1320978	G	T	G	BTM20_05975	hypothetical protein	missense_variant c.956C>A p.Ala319Asp
1357085	G	G	T	BTM20_06175	MATE family efflux transporter	missense_variant c.1177G>T p.Asp393Tyr
1419022	G	T	G	BTM20_06445	sporulation sigma factor SigG	missense_variant c.95G>T p.Arg32Ile
1633161	A	C	C	BTM20_07410	hypothetical protein	missense_variant c.381C>A p.Asn127Lys
1696244	C	A	C	BTM20_07725	molybdopterin molybdenumtransferase MoeA	missense_variant c.760G>T p.Asp254Tyr
1697358	A	C	C	BTM20_07730	molybdopterin-guanine dinucleotide biosynthesis protein B	stop_gained c.376G>T p.Glu126*
1701819	C	A	C	BTM20_07750	nitrate reductase	missense_variant c.871G>T p.Val291Leu

1745933	C	A	C	BTM20_07985	SpoIVB peptidase	missense_variant c.51G>T p.Leu17Phe
1831706	A	C	C	BTM20_08425	D-alanyl-D-alanine carboxypeptidase	stop_gained c.49G>T p.Gly17*
1931131	T	G	G	BTM20_08885	tRNA guanosine(34) transglycosylase Tgt	missense_variant c.665C>A p.Ala222Glu
2031226	G	G	T	BTM20_09400	DNA topoisomerase III	missense_variant c.269C>A p.Ser90Tyr
2077127	G	G	T	BTM20_09640	6S RNA	non_coding_transcript_variant
2115213	T	G	G	BTM20_09820	ABC transporter permease	missense_variant c.274C>A p.Leu92Ile
2154708	**C**	**A**	**C**	**BTM20_09995**	**heme ABC transporter ATP-binding protein**	**missense_variant c.493G>T p.Asp165Tyr**
2235623	C	A	C	BTM20_10330	hypothetical protein	missense_variant c.132G>T p.Lys44Asn
2332138	G	T	G	BTM20_10785	cell division protein FtsH	missense_variant c.607G>T p.Val203Phe
2354690	A	C	C	BTM20_10875	DNA repair helicase	stop_gained c.682G>T p.Glu228*
2381967	G	T	G	BTM20_10975	putrescine-ornithine antiporter	missense_variant c.1025C>A p.Pro342Gln
2386328	C	A	C	BTM20_10985	hypothetical protein	missense_variant c.295G>T p.Asp99Tyr
2527890	C	A	C	BTM20_11570	cytosine deaminase	missense_variant c.763G>T p.Ala255Ser
2592148	**C**	**A**	**C**	**BTM20_11865**	**PTS glucose transporter subunit IIB**	**missense_variant c.963G>T p.Met321Ile**
2597539	**C**	**A**	**C**	**BTM20_11885**	**histidine kinase**	**stop_gained c.931G>T p.Glu311***
2818805	C	C	A	BTM20_13090	Rrf2 family transcriptional regulator	stop_gained c.376G>T p.Glu126*
2867776	T	G	G	BTM20_13365	DHH family phosphoesterase	stop_gained c.1391C>A p.Ser464*

Table 15: Unique SNPs shared by three strains of one host (11S0315, 11S0316 and 12S0471) compared to a possible ancestor strain (12S0464). The table shows the unique SNPs shared by the three strains of one host (11S0315, 11S0316 and 12S0471) to the last possible ancestor strain (12S0464) isolated one year before. The locus tag specifying the protein coding, the product involved and the effect caused are denoted.

Position	12S0464	Within-host strains	Locus Tag	Product	Effect
31672	C	G	BTM20_00160	helicase-exonuclease AddAB subunit AddB	missense_variant c.418G>C p.Glu140Gln
90060	C	T	BTM20_00385	transporter	missense_variant c.651G>A p.Met217Ile
242988	T	C	BTM20_01150	translation elongation factor Tu	synonymous_variant c.114T>C p.Asn38Asn
242994	G	A	BTM20_01150	translation elongation factor Tu	synonymous_variant c.120G>A p.Gly40Gly
493280	A	C	BTM20_02265	MFS transporter	missense_variant c.383C>A p.Ser128Tyr
504422	C	A	BTM20_02310	sodium-independent anion transporter	missense_variant c.716C>A p.Ala239Asp
541894	A	C	BTM20_02480	ABC transporter	stop_gained c.781G>T p.Gly261*
621627	G	C	BTM20_02795	PTS sugar transporter subunit IIA	missense_variant c.209G>C p.Gly70Ala
633297	T	C	BTM20_02850	hypothetical protein	missense_variant c.227T>C p.Ile76Thr
741085	G	T	BTM20_03340	hypothetical protein	missense_variant c.80C>A p.Ser27Tyr
903328	G	A	BTM20_04070	spore germination protein	missense_variant c.831G>A p.Met277Ile
1013589	C	A	BTM20_04560	anion transporter	synonymous_variant c.105C>A p.Thr35Thr

1025293	T	G	BTM20_04620	hypothetical protein	missense_variant c.602T>G p.Val201Gly
1037114	C	G	BTM20_04680	MATE family efflux transporter	missense_variant c.811G>C p.Ala271Pro
1058838	C	A	BTM20_04785	hypothetical protein	synonymous_ variant c.378C>A p.Gly126Gly
1258119	A	G	BTM20_05745	hypothetical protein	missense_variant c.352G>A p.Gly118Ser
1356681	C	T	BTM20_06175	MATE family efflux transporter	missense_variant c.773C>T p.Ser258Leu
1566309	C	A	BTM20_07105	Fe(3) ABC transporter substrate-binding protein	synonymous_ variant c.573C>A p.Thr191Thr
1576194	G	A	BTM20_07155	hypothetical protein	synonymous_ variant c.15A>G p.Thr5Thr
1660504	A	C	BTM20_07520	N-acetylmannosamine kinase	missense_variant c.624T>G p.Ser208Arg
1802770	G	A	BTM20_08300	beta-D-galactosidase	missense_variant c.68C>T p.Ala23Val
1922236	G	A	BTM20_08835	coproporphyrinogen dehydrogenase HemZ	missense_variant c.352C>T p.His118Tyr
2031858	A	C	BTM20_09405	TspO protein	missense_variant c.199G>T p.Ala67Ser
2200431	G	A	BTM20_10170	YgiQ family radical SAM protein	missense_variant c.1511C>T p.Ala504Val
2352168	T	G	BTM20_10870	DDE transposase	missense_variant c.1411T>G p.Cys471Gly
2374914	A	C	BTM20_10940	PTS transporter subunit IICB	stop_gained c.1555G>T p.Gly519*
2427299	A	C	BTM20_11180	hypothetical protein	missense_variant c.391G>T p.Asp131Tyr

2456879	T	C	BTM20_11325	glucuronyl hydrolase	synonymous_varian t c.1149A>G p.Glu383Glu
2479364	T	C	BTM20_11400	vancomycin resistance protein	missense_variant c.746G>A p.Gly249Glu
2538980	G	A	BTM20_11610	hypothetical protein	synonymous_ variant c.45C>T p.Gly15Gly

Table 16: SNPs identified from outbreak strains recovered from different animals. Table 16 shows the groups of outbreak strains (group 1 to 5), strain designation, region of outbreak, SNPs predicted by each variant calling pipeline (Snippy, Lyve-SET), co-linear SNPs and the category of SNPs identified in the protein coding genes.

	Group	Region	Snippy	Lyve-SET	Co-linear SNPs	Substitution types identified in coding genes		
						NS	S	S-G/L
1	BS79-01, BS 80-01	Mecklenburg -Western Pomerania	29	29	29	22	-	4
2	12S0467, 12S0468	Lower Saxony	32	35	21	16	-	5
3	S0041-08, S0042-08	Schleswig-Holstein	68	73	68	41	3	14
4	S0021-10, S0022-10	North Rhine-Westphalia	30	30	30	14	1	10
5	13S0851, 13S0854	Baden-Württemberg	12	13	12	6	2	1

NS - non-synonymous, S - synonymous, S-G/L -Stop gained/loosed

3.3 Core genome MLST and CRISPR spacer sequence typing

A stable core genome MLST (cgMLST) scheme was generated using SeqSphere version 3.5 for *C. chauvoei*. The core genome MLST (cgMLST) target definer analysis was carried out using extracted genes from the reference genome DSM 7528^{T}. The cgMLST scheme based neighbour joining tree was compared to the unique CRISPR spacer based matrix. Most of the clusters identified with the cgMLST scheme could be differentiated in the CRIPR spacer based array. Few of the strains which were differentiated with the cgMLST scheme showed the same unique spacer patterns in the CRISPR spacer based matrix. This was observed for strains from Mecklenburg- Western Pomerania (BS79-02, BS80-02 and 11S0318), and strains coming from Schleswig-Holstein (S0040-08 and S0041-08). These strains also formed unique clusters in pan-genome SNP based phylogeny (Figure 17). Allelic profile data were also used to generate minimum spanning trees (MST) shown in Figure 22. Figure 22-A represents the minimum spanning tree for the strains highlighting county information and 22-B highlights the region information. The tree shows that the SNP distance for strains outside Europe is greater when compared to strains within Germany. Within Germany, the strains from North

Rhine-Westphalia are the most divergent strains. The MLST scheme was also evaluated for its collinearity with the place of isolation based on the geographical map of outbreaks in Germany in the last 21 years (Figure 23). This investigation has shown the applicability of cgMLST scheme for typing the pathogen (Figure 22).

Figure 21: (Page 81) Comparison of core genome MLST and CRIPR spacer based matrix. The neighbour joining phylogenetic tree was created using cgMLST and compared to the matrix representation of CRISPR spacers identified in each strain. The spacers are numbered based on the 12S0467 strain (first line), unique spacers identified in strains are represented by blue boxes, and strains with and without gene insertion in the CRISPR array are shown in red and green respectively. The variability with respect to absence and duplication of CRISPR spacers are highlighted in orange.

Figure 22: (Page 82 and 83) Minimum spanning tree based on core genome MLST. Figure 22-A represents the minimum spanning tree for the strains highlighting country information. Figure 22-B represents the minimum spanning tree for the strains highlighting the region information. The tree shows that the SNP distance for strains outside Germany is greater when compared to strains within Germany. Within Germany, the strains from North Rhine-Westphalia are the most divergent.

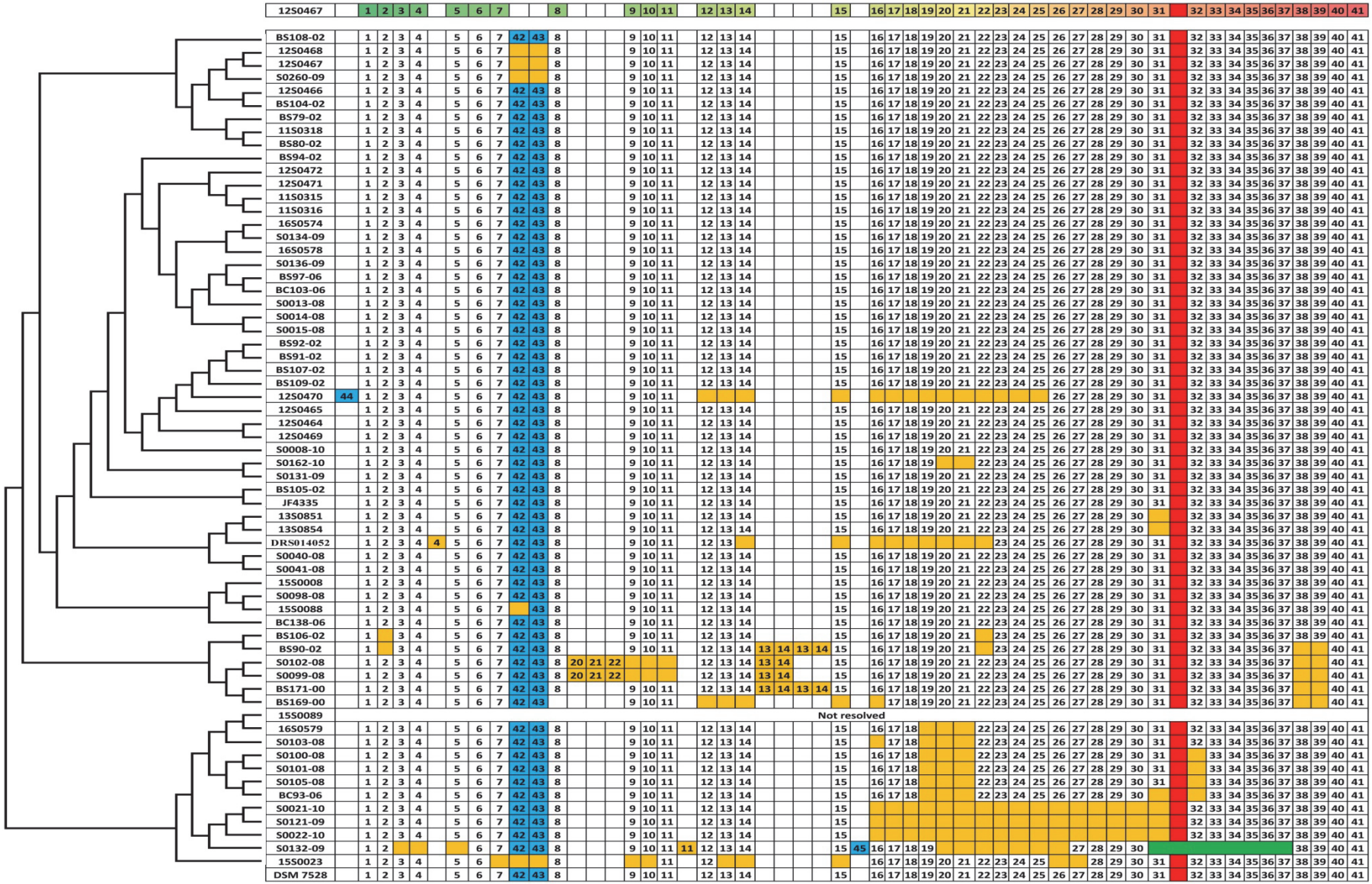

12S0467
BS108-02
12S0468
12S0467
S0260-09
12S0466
BS104-02
BS79-02
11S0318
BS80-02
BS94-02
12S0472
12S0471
11S0315
11S0316
16S0574
S0134-09
16S0578
S0136-09
BS97-06
BC103-06
S0013-08
S0014-08
S0015-08
BS92-02
BS91-02
BS107-02
BS109-02
12S0470
12S0465
12S0464
12S0469
S0008-10
S0162-10
S0131-09
BS105-02
JF4335
13S0851
13S0854
DRS014052
S0040-08
S0041-08
15S0008
S0098-08
15S0088
BC138-06
BS106-02
BS90-02
S0102-08
S0099-08
BS171-00
BS169-00
15S0089
Not resolved
16S0579
S0103-08
S0100-08
S0101-08
S0105-08
BC93-06
S0021-10
S0121-09
S0022-10
S0132-09
15S0023
DSM 7528

Figure 22-A: Minimum spanning tree for the strains highlighting country information

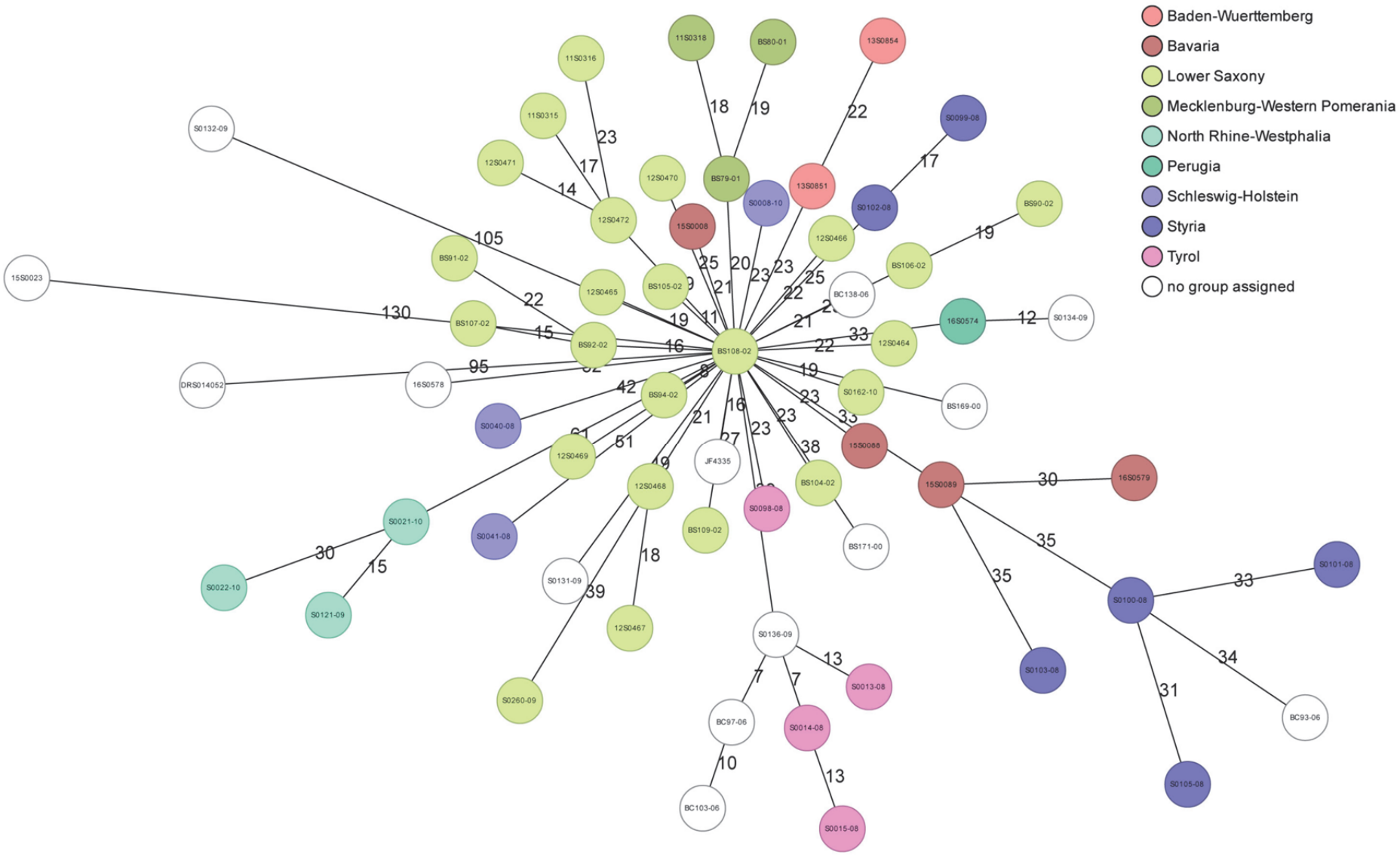


Figure 22-B: Minimum spanning tree for the strains highlighting the region information

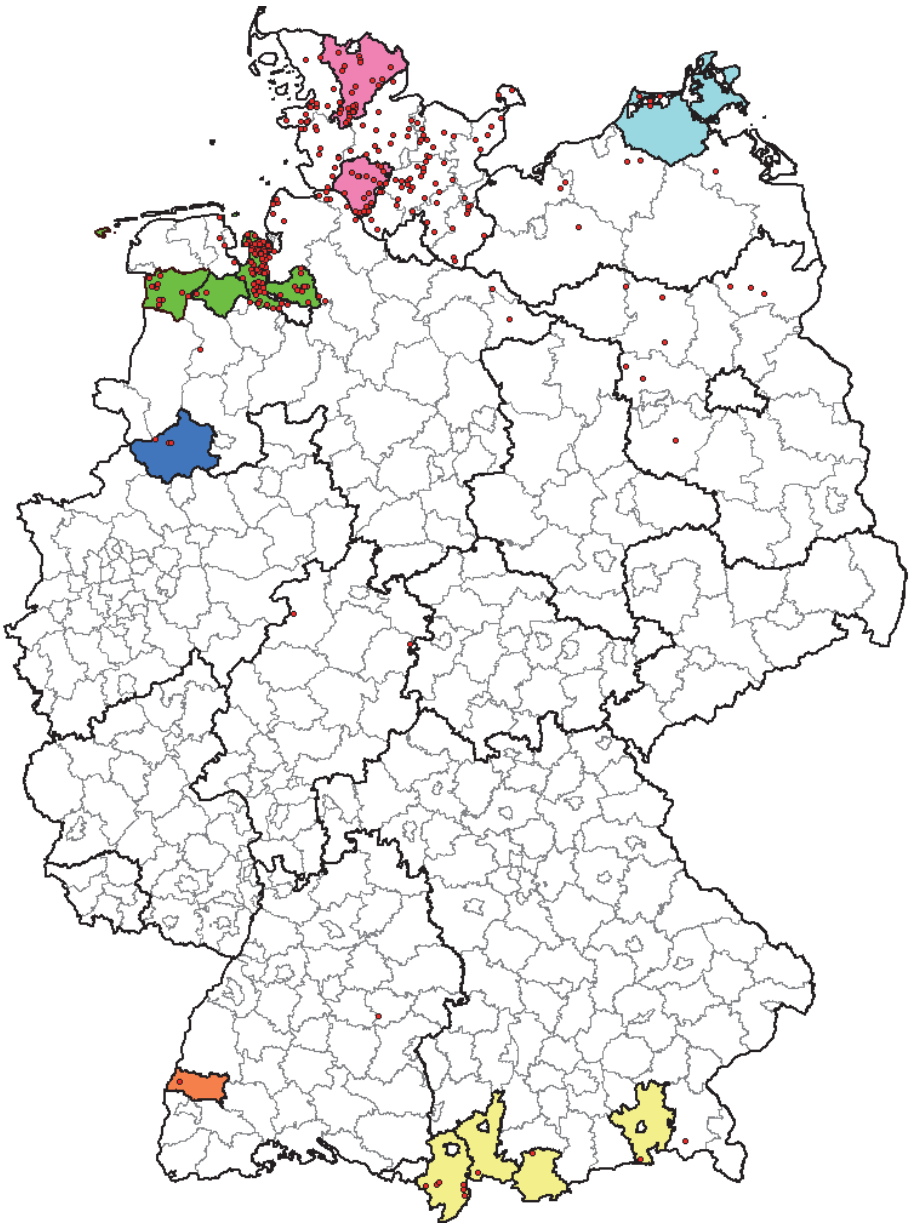


Figure 23: Geographical map of blackleg outbreaks in Germany from 1995 -2010. Map depicting the origin of the German *C. chauvoei* isolates used in this study (coloured areas) and reported outbreaks (red dots) from 01.01.1995 to 31.12.2016 (n= 334). Borders of federal states and administrative districts are delineated. Colour coding: Lower Saxony: green, North Rhine-Westphalia: blue, Schleswig-Holstein: pink, Mecklenburg-Western Pomerania: aqua, Baden-Württemberg: orange and Bavaria: yellow.

Chapter 4: Discussion

4.1 Genome completion and genome composition

4.1.1 Genome completion and origin of replication

The first complete genome sequence of two *Clostridium chauvoei* strains was achieved for the type strain (DSM 7528^T) and a strain of German origin (12S0467) isolated from a diseased animal. The genome sizes of both the type strain and the field isolate, corroborated the published draft genome sequence of the Swiss *C. chauvoei* field strain JF4335 (Falquet et al., 2013). The genome analysis showed the presence of 87 tRNA genes and 27 rRNA genes in 9 rRNA operons. The rRNA gene cluster is used as an important genetic marker to differentiate and identify *C. chauvoei* and *C. septicum* by PCR (Sasaki et al., 2000). Most of the rRNA genes are positioned close to each other, a constellation which is considered as the major hindrance for genome finishing (Koren et al., 2013). The characterization of the region for the genome seems to be difficult without long read based sequencing techniques such as the PacBio RS II system which utilizes C4 chemistry capable of generating reads with N50 close to 20kb (Rhoads and Au, 2015). The genome sequencing resulted in the generation of two high quality complete genome sequences which also included the type strain and can serve as reference for comparative genome studies. The predicted origin of replication was similar to the regions known for *Bacillus* (*B.*) *subtilis oriC* displaying DnaA box clusters in the intergenic regions both upstream and downstream of *dnaA* (Briggs et al., 2012). The number of DnaA boxes with the consensus sequence (ttatccaca) was two for both *oriC* regions (Figure 2).

4.1.2 Sporulation and germination genes

Sporulation and germination are important events in the life of spore forming bacterial pathogens. The genes involved in sporulation and germination were evaluated and compared to related species to discover species specific variations (see 3.1.2.5). Genes for small, acid-soluble spore proteins (SASPs) identified in the genome were of the α/β type. SASPs contribute significantly towards spore resistance against heat, UV radiation, and chemical agents (Setlow, 1994, Setlow, 1988, Meaney et al., 2016), *C. perfringens* strains lacking a significant number of α/β-type SASPs develop more sensitive spores (Paredes-Sabja et al., 2008a, Raju et al., 2007). Dipicolinic acid (DPA), a major component of bacterial endospores, is made from the precursor dihydro-dipicolinic acid (DHDPA) by DHDPA synthase and finally DHDPA is oxidized to DPA by the products of the *spoVF* operon in the case of *Bacillus* and many *Clostridium* species. However, many pathogens belonging to the *Clostridium* cluster I such as *C. perfringens*, *C. botulinum* and *C. tetani* have no *spoVF* orthologues in their genomes (Durre, 2014). In the case of *C. perfringens*, it has been discovered that electron transfer flavoprotein (EtfA) is involved in the production of dipicolinic acid from DHDPA (Orsburn et al., 2010). The *C. chauvoei* genome encodes a DHDPA synthase designated as DapA coding for 292 amino acids but a *spoVF* operon is absent as in other organisms of the clostridia cluster I. Two genes encoding EtfA proteins with 335 and 396 amino acids, respectively, were identified in *C. chauvoei* and were of similar length to that predicted for *C. perfringens* (Orsburn et al., 2010) (Table 5). Spo0A, a transcriptional factor plays a central role in the sporulation process of bacteria of the genera

C*lostridium* and *Bacillus*. There is also evidence showing that Spo0A is involved in regulating various metabolic and virulence factors such as toxins in the genus *Clostridium* (Pettit et al., 2014, Paredes-Sabja et al., 2011). The *C. chauvoei* genome data also revealed the presence of the *spo0A* and other genes involved in various stages of sporulation. Sporulation in *Clostridium* species is initiated by kinases known as orphan kinases as they lack the cognate response regulator and the phosphorelay mechanism found in *Bacillus* (Talukdar et al., 2015). The Spo0A gene of four species, *C. chauvoei*, *C septicum*, *C. perfingens* and *C. botulinum* type A was compared and thereby revealed a central variable region with deletions observed for *C. chauvoei* and *C. septicum* (Figure 5). The limited sequence variation of *Spo0A* has already been employed for the specific detection of *C. chauvoei* and *C. septicum* based on real-time PCR (Lange et al., 2010). The exact role of Spo0A and its interaction with orphan kinases in *C. chauvoei* waits to be explored. Furthermore also the possible role of Spo0A in regulating the metabolic pathways and virulence factors in *C. chauvoei* is unclear.

Organization of germination receptors in clostridia are found to be different from *B. subtilis* and are also strain dependent (Durre, 2014). In the case of *C. perfringens*, the *gerK* operons are organized as a monocistronic *gerKB* and a bicistronic *gerKA-gerKC* transcriptional unit (Paredes-Sabja et al., 2009). The *C. chauvoei* genome also shows the same arrangement. It can be assumed that a similar germination pattern with the necessity for L-asparagine and the possibility of KCl-induced germination as observed for *C. perfringens* (Paredes-Sabja et al., 2008b) is present. The genome also harbours a *prkC*, encoding a Ser/Thr kinase observed in *C. perfringens,* suggestive for an alternative germination pathway triggered by environmental peptidoglycan fragments as described for *B. subtilis* (Xiao et al., 2015). Most of the *C. chauvoei* genes for sporulation and germination showed close homology to those of *C. septicum,* followed by *C. perfringens* and *C. botulinum* which suggests a genetic relatedness and similar mechanisms for sporulation and germination (Table 5). The translated proteins of *C. chauvoei* DapA, GerKC, PrKC and germination-specific spore cortex-lytic enzymes SleB showed significantly lower homology to those of *C. septicum, C. perfringens* and *C. botulinum* indicating the involvement of species specific factors required for germination (Table 5).

4.1.3 Phylogenetic relatedness within the genus, comparative genomics of the completed strains

The analysis of genome sequence data to depict phylogenetic relatedness in the genus *Clostridium* revealed groups of species sharing homology. Significant homology for *C. chauvoei* was observed with *C. septicum* (Figure 7). These findings are similar to previous phylogenetic studies based on 16S rDNA (Kuhnert et al., 1996, Stackebrandt et al., 1999). Based on *rrs* sequence analysis for nucleotide signatures and unique restriction enzyme digestion patterns, *C. chauvoei* was found to be phylogenetically related to *C. septicum* and *C. carnis* (Kalia et al., 2011, Kuhnert et al., 1996). Comparison of the two completed *C. chauvoei* chromosomes showed inversion and translocation of locally collinear blocks (LCBs) (Figure 9A). Symmetric inversion around the bacterial origin of replication is a common feature for closely related species and also within species (Eisen et al., 2000). Presence of blocks switched to different positions among isolates have also been observed in strains of *C. botulinum* (Fang et al., 2010).

Based on panOCT (Fouts et al., 2012) analysis, the core and accessory genome of three *C. chauvoei* isolates were revealed (Figure 8). The genomes are highly similar in respect to gene content and differ only in a limited number of genes. The majority of the protein genes which are shared exclusively by the finished genomes DSM 7528^T and 12S0467 are genes for transposition and mobile element proteins. The genome features of *C. chauvoei* (DSM 7528^T) and *C. septicum* (CSUR P1044) were compared based on subsystem category distribution features in RAST (Aziz et al., 2008, Overbeek et al., 2014) (Figure 10). Both species share similar subsystem category distributions whereas the subsystem feature counts were higher in *C. septicum* for genes related to prophages, phosphorus metabolism and carbohydrates (Figure 10). Also the published genome for *C. septicum* has a 32 kb plasmid whereas similar large plasmids were not observed so far in characterised strains of *C. chauvoei.*

4.1.4 Antibiotic resistance genes

Antibiotic resistance genes encoded in the genome can confer resistance to bacterial pathogens. An earlier study based on genome sequence of *C. chauvoei* JF4335 reported a number of genes for antibiotic resistance (a gene for penicillin resistance (a beta-lactamase gene), an elongation factor G (EF-G) type tetracycline resistance gene (*tetM* and *tetO* analogues), and a vancomycin B-type resistance gene (*vanW*). However, that same study described the isolate (and others used in the study) sensitive to all antibiotics tested which are usually used against clostridial infections. Therefore the possibility for non-expression or expression of non-functional proteins was postulated (Frey and Falquet, 2015). Additional genes potentially conferring resistance to peptide antibiotics and macrolides were found in the present study in both genomes (see 3.1.2.3). Three genes which are protein variants in the CARD database were identified. Two encoded a *gyr*B related gene, possibly conferring resistance to aminocoumarin and one encoded *rpo*B, potentially conferring resistance to rifampicin. Even though specific SNPs can confer resistance to antibiotics in bacteria (McArthur et al., 2013), whether the protein variants identified in the *C. chauvoei* genomes play any role in antibiotic resistance is unclear. However, antimicrobial resistance was not tested in our study.

4.2 Comparative genomics

4.2.1 Genome sequencing, assembly and annotation

The study aimed also to determine the population diversity of *C. chauvoei* and was carried out with 65 *C. chauvoei* strains. The best assembly options for the species genome with the Miseq (2×300 bp) paired end reads were assessed with genome assemblers reported best in earlier studies for bacterial genome assembly (Table 6) (Magoc et al., 2013, Jünemann et al., 2014, Gurevich et al., 2013). SPAdes provided the best N50 values and less misassembles while MaSuRCA revealed the best genome fraction. The *C. chauvoei* genome has been reported to include a small cryptic plasmid of 4 to 5 kb size. The current study revealed that all the genomes contain a plasmid of 4 kb size. The genomes were annotated using the Prokka annotation pipeline, the predicted protein coding gene numbers were comparable to the numbers found in finished genomes of the species (Table 4, Table 8A). The draft genomes also showed a lower number of rRNA genes, suggestive for the inadequacy of short reads to resolve multiple rRNA gene clusters (Table 8A). One genome was completely excluded from

the subsequent analysis due to unusual genome composition identified based on assembly and annotation (S0133-09) and for one genome the accessory genome was excluded owing to the presence of unusual gene content from diverse bacterial sources (16S0574).

4.2.2 CRISPR elements

CRISPR elements identified in bacteria confer protection against bacteriophages (Barrangou et al., 2007). *C. chauvoei* genomes harbour a CRISPR element similar to the CRISPR type I-B system with *cas* genes in the order (*cas6-cas8b1-cas7-cas5-cas3-cas4-cas1-cas2*) according to the updated CRISPR type classification (Makarova et al., 2015) (3.1.2.2 and 3.2.1.3). The observed CRISPR element harbours two repeat sites. Both sites are very close to each other, separated by a single protein coding gene. More variations were observed at the CRISPR site 2 when compared to site 1, which was close to the cas proteins. There was an observable correlation between strains originating from the same region/farm showing the same CRISPR repeat numbers. Also the strains from geographical origins outside Europe showed variable repeat numbers (Table 8B). Hence it could be a possibility to type the strains based on CRISPR spacer sequences. Recently, CRISPR diversity has been applied to a CRISPR-based phylogenetic and typing analysis based on a collection of 217 *C. difficile* genomes (Andersen et al., 2016). A recent study has pointed out the significance of unique spacer sequences of *C. chauvoei* CRISPR elements to explore the genetic diversity of isolates from different geographical origin (Rychener et al., 2017). CRISPR elements have also been reported from a pathogenic *C. perfringens* type A strain isolated from a calf (Nowell et al., 2012). Interestingly no CRISPR elements were identified on the chromosome of a *C. botulinum* type III strain (BKT015925) of poultry origin but CRISPR elements were present in the prophage (p1) and on a plasmid (p2), respectively (Skarin et al., 2011). In a detailed analysis of the same strain it was found that the CRISPR type I-B system was incomplete for all *cas* genes and was predicted to be inactive (Woudstra et al., 2016). The strains involved in the current study show the presence of a complete CRISPR/Cas system and with variable repeat numbers, indicating an active CRISPR/Cas system.

4.2.3 Phages

A previous study on *C. chauvoei* genome sequence data reported the presence of prophages (Frey and Falquet, 2015). The current study showed a similar phage composition except for the size, which differed among strains (Figure 3, Table 8B). One of the haemolysin genes, previously described in the *C. chauvoei* genome belonging to the haemolysin XhlA superfamily was also found to be present in the prophage (Frey and Falquet, 2015). The predicted prophage identified in the genome was of a similar type for all strains investigated and hence indicates that the species genomes are immune to any new phage types. Limited changes were observed, such as one strain had a duplication of the same phage type and further three strains from the same farm showed a deletion of the phage region (Figure 14 and Table 9). The role of phages within the evolution of *C. chauvoei* seems to be limited in comparison to other related species.

4.2.4 *Clostridium chauvoei* pan-genome structure

Pan-genome analysis of the species, based on the available *C. chauvoei* strains, shows that the species has an open pan-genome structure (Figure 12B). The acquisition of new genes was observed to be very limited and ranged between 3 to 4 new genes per strain (Figure 12C). This shows the low potential of *C. chauvoei* genomes for acquiring or losing large unique foreign genetic elements. In contrast, other species representing the clostridia Cluster 1 group, such as *C. botulinum* and *C. perfringens* are characterised by large accessory genomes, which are mostly attributed to mobile genetic elements, phages and large plasmids that are horizontally transferred genetic elements (Skarin and Segerman, 2011, Nowell et al., 2012). The study presented here, proves that extensive genetic variability or large plasmids are absent in *C. chauvoei*. Since all the primary virulence factors previously recognised for the species were identified as part of the core genome, the species virulence is probably represented in the core genome. Pan-genome size depends on the gene gain and loss events occurring in bacteria. Gene losses are supposed to occur in new, restricted ecosystems where the non-essential genes are lost. Gene gain happens when the bacteria live in a very diverse environment and hence genetic exchange with diverse bacterial species is possible (Rouli et al., 2015). The genome size can be modified by mobile genetic elements which involve phages, integrases and transposases. The presence of phages can be regulated by active CRISPR elements present in the genome (Rouli et al., 2015). In this aspect *C. chauvoei* could be expected to encompass variable genetic elements that are present in the environment and the intestinal tracts of the hosts. On the other hand there are also species with closed pan-genomes such as *Yersinia pestis* and *Bacillus anthracis* which multiply only inside the host tissue and hence have very little chance to acquire new genes (Rouli et al., 2015). The core vs pan-genome ratio observed for *C. chauvoei* was 83% when including all the strains and was 96% when the three completed strains were compared. It can be speculated that the *C. chauvoei* pan-genome stay open, as the genome harbours several IS elements and genomic islands which could integrate new genetic elements. In contrast, the core vs pan-genome ratio was identified to be only 11% in the case of *C. botulinum.* The comparative genome sequence analysis of *C. perfringens* based on complete genomes involving various serotypes (A-E) also indicates an open pan-genome structure. Pairwise comparisons of shared orthologous protein coding genes or pseudogenes between strains ranged from 69.3% (D and E) to 85.5% (A and B) (Hassan et al., 2015). One potential reason for the high core vs pan-genome ratio in *C. chauvoei* is that the species may replicate only in isolated niches i.e. host tissues, so that the chance of acquiring foreign genetic material is limited. Another possible reason could be that the species has undergone evolution of genome reduction as observed for intracellular pathogens where a functional reduction especially of genes involved in transcription and amino acid metabolism is detected (Rouli et al., 2015). Functional reduction has been observed in an earlier study based on genome sequence data for *C. chauvoei* strain JF4335, revealing the absence of several genes involved in amino acid metabolism (Frey and Falquet, 2015). The evidence of functional reduction and intracellular existence has not yet been proven for this pathogen, even though spores and vegetative cells have recently been shown to survive in bovine macrophages (Pires et al., 2017b).

4.2.5 Genomic islands

Genomic islands (GIs) are clusters of genes which are probably transferred horizontally in bacterial genomes. GIs contribute significantly to genome evolution and often provide adaptive traits that enhance the fitness of bacteria. Sometimes they encode virulence factors, antimicrobial resistance, novel genes and provide adaptations like metal resistance, new metabolic pathways etc. (Bertelli et al., 2017). Fourteen regions were predicted to be genomic islands (GI) varying in size from 6 kb to 19 kb for the complete *Clostridium chauvoei* field strain 12S0467. A small region within one of these putative genomic islands was identified to show variability among some strains (Figure 15). Eight strains showed absence of a region within this predicted GI which contained open reading frames for potential virulence factors such as a sialidase, an alpha-L-fucosidase and a NPTQN specific sortase B gene (Table 10). Alpha-L-fucosidases are involved in the degradation of intestinal fucosyl glycans in *C. perfringens* (Fan et al., 2016) and sortase B (SrtB), a cysteine transpeptidase has been identified in *Staphylococcus aureus* that attaches a polypeptide involved in haeme-iron transport and thus helps to acquire iron during infection (Zong et al., 2004).

4.2.6 Homologous recombination

Homologous recombination in bacteria occurs as a result of transduction (mediated by an infectious agent), conjugation (direct contact between cells) or transformation (obtained directly from the environment). Recombination in bacteria involves recipient and donor cells contributing asymmetrically towards the genetic content, where the donor contributes only a small contiguous segment of DNA to the resulting genome. The recombination event prediction based on Gubbins (Croucher et al., 2015) and BratNextGen (Marttinen et al., 2012) showed that *C. chauvoei* strains in the current study undergo only limited recombination events. The regions predicted by both tools showed collinearity for most of the regions/genes. The r/m ratio (ratio of rates at which nucleotides become substituted as a result of recombination and mutation) identified at various recombination sites predicted by Gubbins, ranged from 0.37 to 1.71 indicating recombination does not contribute to the evolution of this pathogen (Table 13). Previous studies have estimated the recombination rates based on the r/m ratio for bacteria and archaea from MLST data using the recombination detection tool ClonalFrame. The r/m ratio showed variation among different species ranging from 0.1 to 63 when 2 to 8 MLST loci were analysed (Vos and Didelot, 2008).

4.2.7 Phylogeny and clustering of isolates

The current study involved only few strains from different continents, whereby the majority of the strains were from Germany. Some of them were isolated from the same outbreak or even from the same animal. Hence, different comparative genomics approaches were applied to explore the genetic relatedness of the involved strains. Core genome phylogeny and pan-genome SNP based approaches encompassed all strains and where applied to describe the genetic relatedness based on bootstrap support values and for clustering of isolates, respectively. Reference based read mapping approaches were applied to identify SNPs of closely related strains originating from Europe, different regions of Germany, outbreak strains and strains from the same host strains, respectively.

4.2.7.1 Core genome phylogeny

The maximum likelihood phylogenetic tree for the core genome (Parsnp) was midpoint rooted and 15S0023 (NCTC 08361, South Africa, sheep) was identified as the outgroup strain for the strains of the current study (Figure 16). The analysis showed that most of the isolates clustered according to their geographical origin. Strong bootstrap support values (100%) were highly prominent for strains coming from same farm/outbreak/animal (Figure 16). Few strains were also found not to cluster along with strains originating from the same region/farm, indicating that diverse strains also could be present in a population. This was mostly observed for strains from Lower Saxony. This could possibly be attributed to the specific geographical pattern of the region which has probably encountered several floods in years, decades and centuries which could have resulted in the dissemination of a diverse *C. chauvoei* population. Strains from Bavaria (15S0089 and 16S0579) and Austria (S0100-08, S S0101-08 and S0105-08) as well as Austria (S0102-08 and S099-08) and Switzerland (BS169-00 and BS171-00) formed clusters with strong bootstrap support values which indicate that these strains maintain genetic and geographical relatedness in these regions.

4.2.7.2 Pan-genome SNP analysis, phylogeny and clustering

Pan-genome analysis (see 3.2.4.3) was carried out employing the kSNP3 package. The parsimony tree was generated based on the pan-genome SNPs and indicated 10 clusters based on the presence of at least 5 unique SNPs shared among the strains within a cluster (Figure 17). A recent study has proven the usefulness of a k-mer based phylogenetic tree in generating a more informative clustering when investigating *Salmonella* strains from different serovars. The likely reason was attributed to the non-requirement of a reference genome which can cause bias if strains which are unrelated to the reference genome are involved in the study (Leekitcharoenphon et al., 2014). Similarly, a reference free k-mer based single nucleotide variant (SNV) phylogeny was applied to classify isolates into genetic lineages in *Listeria monocytogenes* (Hingston et al., 2017). Strains grouped in a cluster indicate genetic relatedness as they share unique SNPs defining them. The clusters received for *C. chauvoei* displayed geographical relationships among them (Figure 17). The functional analysis (based on Blast2GO annotation) of the genes carrying the SNPs defining the cluster or its internal node showed involvement in transmembrane transport, carbohydrate metabolism and phosphorylation (Figure 18). Genes which were commonly observed in multiple clusters or any of their internal node (indicating different SNPs in the same gene) were identified as genes coding for L-fucose isomerase and N-acetylmuramoyl-L-alanine amidase (observed in 3 different clusters or its internal node) and spore germination protein, ABC transporter, two-component sensor histidine kinase, DNA topoisomerase I, ABC transporter ATP-binding protein, glucokinase, sensor histidine kinase and three cds for hypothetical proteins (observed in 2 different clusters or its internal node). The multiple substitutions observed in these genes were mostly non synonymous. These substitutions thus may indicate fast evolving genes of the species. Since these gene associated multiple SNPs were observed beyond one cluster, one may speculate that they could have an influence on the evolution and possibly on the pathogenesis of blackleg. The unique SNPs shared by a cluster were identified to have a maximum value of 76 (Cluster 4, Figure 17), which was obtained for the three strains from North Rhine-Westphalia. This indicates that the strains form a diverse, outstanding cluster among the strains from Germany. The outbreak had a clinical history of being observed in the

animal housing rather than on the pasture in contrast to most other outbreak strains in this study. Since cotton material was used as bedding for animals, the introduction of outbreak strains from another geographical origin could also be a possibility. North Rhine-Westphalia is considered a region where no blackleg outbreaks have been reported except for this unique outbreak. Regarding the strains originating from other geographical locations, such as Lower Saxony, Bavaria and Austria (strains from different farms) there is evidence that the strains are involved in multiple clusters indicating diverse populations. One exceptional cluster (Cluster 9, Figure 17) with divergent strains was observed where the type strain (allegedly isolated before 1950) was found to be genetically related to the three strains from Austria isolated in the year 2008 (Figure 17). Strains from the same farm/outbreak showed significant relatedness. In few cases strains from the same farm/region but in longitudinal time gaps of 1 and 4 years (2 cases) showed significant relatedness (Figure 17, Cluster 4 (S0121-09, S0021-10 and S0022-10) and internal node of Cluster 5 (BS79-01, BS80-01 and 11S0318), respectively). Exceptions were observed when one strain (12S0464) isolated a year before, did not form a cluster with the three strains isolated from the same farm and one animal a year later (Cluster 10 - 11S0315, 11S0316 and 12S0471). These strains clustered with another strain (12S0472) from Lower Saxony (Figure 17). This indicates that repeated outbreaks at farm level mostly involve strains from the same lineages, although divergent strains could be observed at farm level. The sources of repeated outbreaks can be strains with evolutionary relationship, which were freshly acquired by an animal causing the new outbreak or the herd carried the bacterium in a latent stage and an outbreak was caused by one of the predisposing factors predicted for the pathogen at different time intervals. The usefulness of the limited cluster specific genes and SNPs to depict the micro evolution of this pathogen is obvious. Hence these SNPs can be applied to type the pathogen. In contrast pan-genome analysis, phylogeny and clustering is not always able to reveal geographical relatedness. A study based on eighty-three *Erysipelothrix rhusiopathiae* isolates from a range of host species and geographic origins revealed that the species is weakly clonal, but highly recombinogenic. The three distinct clades generated by comparative genomics and phylogeny showed no clear segregation based on host or geographic origin of the isolates (Forde et al., 2016). Similar results were found using comparative genomics involving 44 strains of *Clostridium sordellii* isolated from human and animal infections in the UK, the US and Australia. The study divided the species into four clades, whereby no clade was found to be specific for the country of origin, or for human or animal infections (Couchman et al., 2015).

4.3 Detection of strain variability by reference based mapping

Read mapping and SNP detection was performed using the Snippy pipeline (Seemann, 2015). All the strains for which reads were available were mapped to the 12S0467 strain to generate the SNPs. SNPs present in the with-in host and outbreak strains were detected using two different pipelines Snippy and the LYVE version of the Snp Extraction Tool (SET) (Katz et al., 2017) to confirm the SNP calling was accurate for these closely related strains.

4.3.1 Strain variability of European strains

Pairwise SNP difference analysis was carried out to infer median pairwise SNP difference values among strains belonging to various geographical regions (Figure 19A and 19 B). This value ranged from 50 to slightly above 100 for strains with European origin. The intragroup

comparison shows the 10 Austrian strains to be highly variable when compared to strains originating from Germany (Figure 19B). A recent study to identify the prevalence of blackleg in Styria, Austria, encompassed 266 confirmed cases, including 47 cases from vaccinated animals. Correlation was observed for age groups, time of year and geographical origin. The North-western parts of Styria were identified as high risk area (Wolf et al., 2017). Of the 10 strains from Austria in the current study, 6 were from Styria and significantly contributed to the large median pairwise SNP difference. Similarly higher median pairwise SNP differences were observed for strains from Bavaria when compared to strains from Lower Saxony. The larger genetic variations observed from strains for Styria and Bavaria could indicate more diverse lineages within these regions. This could also be attributed to the use of seasonal pastures and the geology of the area where animals may have a higher risk of being exposed to *C. chauvoei* spores (Wolf et al., 2017).

4.3.2 Strain variability of within-host and outbreak strains

Studies were carried out to investigate the within-host evolution of bacterial pathogens such as *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Clostridium difficile*, *Escherichia coli*, *Mycobacterium tuberculosis* etc., (Didelot et al., 2016). The differences observed within the rate of substitution per year were variable for different species. This could be a combined attribute of the genome size and the differences in the per site mutation rates which are affected by several factors e.g. the efficacy of the DNA mismatch repair systems. Another mechanism generating high rates of point mutations is hyper-mutation, which can occur when the mismatch repair system becomes disrupted. Hyper-mutation can facilitate the adaptation to new ecological niches. Within-host studies of *H. pylori* infections have identified certain genes coding for the outer membrane proteins to evolve faster than the rest of the genome (Didelot et al., 2016). The analysis involved 3 strains from the same animal and another 4 pairs of strains recovered in the same outbreaks from different animals (Table 14 and Table 15). We included only those outbreak strains which formed either a unique cluster or had unique SNPs indicative of a common ancestor (see 3.2.5). The three within-host strains showed shared SNPs at 62 sites (Table 14). We also observed a high percentage of non-synonymous substitutions of the SNPs of the within-host strains. This may indicate the possibility of emergence of polymorphic populations within one host and the major contribution of non-synonymous SNPs to microevolution.

Comparison of synonymous and non-synonymous polymorphisms within-host isolates and isolates from different hosts has revealed a purifying selection as the dominant selective force acting on *S. aureus* over short timescales based on an investigation of 131 bacterial colonies sampled from 13 singly colonized hosts. The level of genetic diversity in within-host isolates was found to be lower than isolates from different hosts whereas the higher dN/dS ratio (the ratio of substitution rates at non-synonymous and synonymous sites) indicative of a less purifying selection for within-host isolates when compared to between-host isolates (Golubchik et al., 2013). *C. chauvoei* strains isolated from the same outbreak showed SNP variations ranging from 12 to 68 and non-synonymous SNPs were dominating the variations. The genes with variations showed very limited collinearity between outbreaks. Some of the genes showing variation also matched with proteins identified in *Staphylococcus aureus* undergoing purifying selection (highlighted in Table 14) (Golubchik et al., 2013). A study conducted to elucidate the genotypic diversity of *Burkholderia dolosa* by re-sequencing 26

individual colonies from a patient sputum sample showed the presence of 6 lineages with at least 5 lineage-specific mutations in the phylogeny. The study postulated that even within a single sputum sample, the population can be so diverse that practical identity between isolates is extremely low (Lieberman et al., 2014). For *C. chauvoei* the diversity of entire populations that occur in the host during the infection is unclear. Since the *C. chauvoei* study involved very few strains which can be considered as targets for within-host evolution studies, the presence of a larger number of unique SNPs which are non-synonymous leads to the speculation that within-host evolution for the species exists. The non-synonymous SNPs and genes involved did not show much collinearity between the different outbreak groups, probably indicative of a larger diversifying population that evolves within the host. Another explanation is involvement of recombination i.e. the exchange of genetic material could be a random process depending on the population involved. A study to analyse the rate and pattern of genome evolution in *Helicobacter pylori* on two input strains identified 168 and 54 SNPs in the output strains, respectively. The study showed that the initial inflammatory immune response during the acute infection phase in the stomach epithelium causes a mutation burst in the bacteria. Outer membrane protein coding genes showed a significant high frequency of mutation and recombination events. The genes are also the most probable bacterial genes undergoing constant selection by the immune system of the host (Linz et al., 2014). The current study was not able to identify which mechanisms e.g. genetic drift, purifying selection or diversifying selection is driving the within-host evolution of *C. chauvoei*. Only for a few of the outbreak strains recombination events were predicted for corresponding proteins. The strains involved in studies of blackleg usually originate from outbreaks and the bacterium is believed to replicate within the host where the anaerobic environment prevails. Hence all strains of this study have possibly undergone within-host evolution. Also, considering the fact that there is no evidence for a contagious nature of the pathogen, few reports suggest that the bacterium could have few replication cycles in the intestine before being either absorbed to the bloodstream or excreted from it (Abreu et al., 2016, Abreu et al., 2017). Typical outbreaks of blackleg are acute in nature and the isolation of strains at longitudinal time scale from the same host may simply not be possible. The extent to which these within host evolved lineages establish themselves in the environment before causing a new outbreak is unclear.

4.4 Flagellin and virulence factors

Flagellin is the principal component of the bacterial flagellum, an important structure for the mobility of bacterial pathogens and also involved in pathogenicity. Earlier studies carried out at the nucleotide sequence level have shown the existence of two *fliCs* which occur in tandem, designated as *fliA*(*C*) and *fliB*(*C*), display highly conserved N and C terminal regions (Sasaki et al., 2002a). Flagella are considered as important vaccine and diagnostic candidate antigens of the pathogen (Tanaka et al., 1987, Tamura and Tanaka, 1984, Usharani et al., 2015). The sequence analysis of the two completed *C. chauvoei* isolates described in the current study revealed the presence of three *fliC* genes instead of two (Fig 4). This is the first report of triplicate *fliC* genes for the species, a finding of possible importance for pathogen detection and immunization. The spacer regions between the three *fliC* genes are conserved and all three genes have a central variable region (Figure 4). Flagella, toxins and lytic enzymes are primary virulence factors in *C. chauvoei* (Frey and Falquet, 2015). Recently "*Clostridium chauvoei* toxin A" (CctA) belonging to the leukocidin superfamily of bacterial

toxins was identified as a major virulence factor in *C. chauvoei*, which is not found in other clostridia (Frey et al., 2012). No genetic variability for the CctA protein at the amino acid level was found for the strains involved in the current study, even though an earlier study reported three nucleotide polymorphisms for one strain isolated 1956 in New Zealand (Frey et al., 2012). Sialidases or neuraminidases are among the few previously characterized virulence factors of *C. chauvoei*. The sialidase gene *nanA* showed variations for few strains. All genomes harbored a NagH homologue gene encoding 1887 amino acids which also showed variations at amino acid level for few strains (Table 11). Further genes potentially involved in the pathogenesis of *C. chauvoei* include phospholipases, collagen binding proteins and genes encoding homologues to internalin A (Frey and Falquet, 2015). Amino acid variations were also observable for other potential virulence factors such as fibronectin type III domain protein, glycosyl hydrolase family 20, hyaluronidase, N-acetylneuraminate lyase and internalin A for few strains (Table 11). A recent study on the membrane proteome of *C. chauvoei* has identified a putative glycosyl hydrolase to be cell-wall associated and immunogenic (Jayaramaiah et al., 2016). Proteolytic enzymes may help in the survival and spread of *C. chauvoei* in muscle tissue and thus contribute to pathogenesis. There are also many genes involved in the transportation and acquisition of iron and other metal ions which are essential for the survival of the pathogens in the host (Porcheron et al., 2013). Recent studies have also pointed out the occurrence of blackleg outbreaks in vaccinated animals in Europe (Wolf et al., 2017, Harwood et al., 2007). Outbreaks with differing pathological findings from the classical form i.e. primary involvement of the tongue and intestine without skeletal muscle or heart involvement have been reported (Harwood et al., 2007). These reports could be suggestive of variant *C. chauvoei* strains present in the pastures which can evade vaccination and cause atypical outbreaks. The strain variants occurring within one host, in different hosts and even in different outbreaks at the same farm/pasture have to be investigated intensively in the future to reveal strain microevolution and possible effects on the pathogenesis of blackleg. The protein variants observed within these strains point towards a new array of targets that could be involved in the pathogenesis and immune evasion of the pathogen and may help to develop novel vaccines. Future studies should target genome transcription to elucidate the pathogenesis of blackleg and the mechanism of how the organism evades the immune system and successfully reaches the host tissue.

4.5 Typing options for *Clostridium chauvoei*

Typing options for the *C. chauvoei* strains in the current study were evaluated based on two approaches, cgMLST and CRISPR spacer sequence matrix. A recent study based on 20 *C. chauvoei* strains from wide geographical regions such as Africa, Australia, Northern and Southern America and Europe showed 187 different CRISPR spacer sequence motives with sufficient heterogeneity to differentiate most strains (Rychener et al., 2017).The unique spacers identified in the current study are 45, indicative of less diverse strains. The CRISPR spacer matrix diversity of the studied strains was found to be inadequate to differentiate many strains of European origin; in contrast cgMLST was able to differentiate the strains even within Europe and Germany (Figure 21). This proves the applicability of cgMLST for the typing of strains from similar geographical origin and its special value for outbreak investigations. The CRISPR based typing approach is able to display genetic relatedness in a more general pattern.

Chapter 5: Summary

Genome sequencing and molecular typing of *Clostridium chauvoei*

High quality circular genome sequences were generated for the *Clostridium chauvoei* type strain DSM 7528^T (ATCC 10092^T) and a field strain 12S0467 isolated in Germany. Comparative genome analysis of the two strains revealed few inversions and translocations in local collinear blocks, indicating a conserved genome with only a small number of accessory genes. Significant homology for *C. chauvoei* was observed with *C. septicum.* The species genome shows a large number of genes, the products of which are involved in proteolysis, cleavage of glycosidic linkages (sialidases) and metal ion transportation. Triplicates of *fliC* were identified in each of the two circular genomes.

Sporulation and germination process related genes were homologous to those of the Clostridia cluster I species with highest homology to *C. septicum*, but novel variations in regulatory genes were also identified.

A comparative genomic study was carried out with a total of 64 *C. chauvoei* strains. The strain collection mostly included strains of European origin whereas few strains were of exotic origin.

Pan-genome analysis of the species based on the available *C. chauvoei* strains shows that the species has an open pan-genome structure. New gene acquisition was observed to be very limited. This probably indicates that the species is undergoing replication only in very isolated areas e.g. inside the host tissue, so that the chances of acquiring foreign genetic material are limited.

The predicted prophage identified in the genomes was similar for all strains, indicating that the genomes are immune to any new phage type. A CRISPR type I-B system was identified in all genomes which may contribute to comprehensive phage immunity. There was an obvious correlation of strains from the same region/farm i.e. displaying the same repeat numbers/spacer sequences. The strains originating from outside Europe showed a unique spacer matrix composition.

Homologous recombination plays an important role in the evolution of some bacterial pathogens. This study found only a limited number of possible recombination events contributing towards the evolution of *C. chauvoei*.

Maximum Likelihood phylogenetic analysis based on the core genome (Parsnp) identified associated isolates with strong bootstrap support values (100%). These values were highly prominent for strains originating from the same farm/outbreak/animal.

Reference genome and alignment free analysis methods (kSNP 3) based on pan-genome SNPs (SNP, Single Nucleotide Polymorphism) were found to be useful for initial clustering of isolates. The strains were clustered based on unique SNPs. The number of allele specific SNPs which are unique for each strain was higher for strains from exotic origin.

Read mapping and SNP calling (Snippy) based on a related reference genome (12S0467) was able to provide novel insights to understand the general, outbreak and within-host microevolution of a pathogen. The median pairwise SNP difference value was higher for *C. chauvoei* strains from Bavaria and Austria as compared to strains from Lower Saxony. The diversity of geographically related strains could therefore be indicative of other environmental factors or herd management influencing strain microevolution. Non-synonymous SNPs contributed significantly to the observed microevolution of the pathogen. Based on the strain variability observed for strains from the same host and strains from the same outbreak, this study also proved the possibility of genetically different *C. chauvoei* populations in one host.

A very recent study, based on genome sequence data, has pointed out the applicability of strain differentiation based on CRISPR spacer sequences for *C. chauvoei.* The current study similarly proved the applicability of typing approaches for this pathogen based on CRISPR elements and core genome MLST. The CRISPR spacer diversity was found to be inadequate to differentiate strains of European origin. However core genome MLST was able to differentiate the *C. chauvoei* strains within Europe and Germany and thus proved to be a valuable tool for strain differentiation.

The current study applied the advantage of long read based sequencing technology to generate high quality complete genome reference sequences for a field strain from Germany and the type strain of *C. chauvoei*. The phylogenetic positioning of the species within the genus was evaluated and the close relatedness to *C. septicum* was confirmed. Several genome analysis software tools provided novel insights in the genome content and composition of the pathogen. Comparative genome sequence analysis based on 64 strains showed limited horizontal gene transfer and novel gene acquisition. Strain typing based on core genome MLST analysed used the first time for this species could differentiate strains even at farm level.

Chapter 6: Zusammenfassung

Genomsequenzierung und molekulare Typisierung von *Clostridium chauvoei*

Zwei vollständige, zirkuläre Genomsequenzen von hoher Qualität wurden für den *Clostridium chauvoei* Typstamm DSM 7528^T (ATCC 10092^T) und dem in Deutschland isolierten Feldstamm 12S0467 generiert. Die vergleichende Genomanalyse der beiden Stämme zeigte wenige Inversionen und Translokationen in lokal kollinearen Blöcken. Dies weist auf ein konserviertes Genom mit nur einer kleinen Anzahl von Zusatzgenen hin. Für *C. chauvoei* wurde eine signifikante Homologie mit *C. septicum* beobachtet. Das Speziesgenom beinhaltet zudem eine große Anzahl von Genen, deren Produkte an Proteolyse, der Spaltung von glycosidischen Bindungen (Sialidasen) und Metallionen-Transport beteiligt sind. In jedem der beiden vollständigen Genome wurden drei Kopien des Gens *fliC* identifiziert.

Sporulations- und Keimungsprozeß-bezogene Gene waren homolog zu denen anderer Arten des Clostridien Clusters I, mit der höchsten Homologie zu *C. septicum*. Aber es wurden auch neue Variationen in diesen regulatorischen Genen identifiziert.

Eine vergleichende genomische Studie wurde mit insgesamt 64 *C. chauvoei*-Stämmen durchgeführt. Die Stammsammlung enthielt zum größten Teil Stämme europäischer Herkunft, während wenige Stämme exotischen Ursprungs waren.

Die Pan-Genom-Analyse der verfügbaren *C. chauvoei* Stämme zeigt, dass die Spezies eine offene Pan-Genom-Struktur aufweist. Die Akquisition neuer Gene wurde selten beobachtet. Dies deutet wahrscheinlich darauf hin, dass die Spezies nur in sehr isolierten Bereichen repliziert, z. B. innerhalb des Wirtsgewebes, sodass die Chancen, fremdes genetisches Material zu erwerben, begrenzt sind.

Der in den Genomen identifizierte Prophage war bei allen Stämme gleichartig, was darauf hinweist, dass die Genome dieser Spezies für jeden neuen Phagentyp immun sind. Ein CRISPR-Typ-I-B-System wurde in allen Genomen identifiziert, dieses trägt wahrscheinlich zur umfassenden Phagen-Immunität bei. Es gibt eine offensichtliche Übereinstimmung der Repeat-Zahlen/Spacer-Sequenzen der Stämme aus der gleichen Region/aus dem gleichen Betrieb. Stämme mit geographischem Ursprung außerhalb Europas hatten eine einzigartige Spacer-Matrix Zusammensetzung.

Homologe Rekombination spielt eine wichtige Rolle bei der Entwicklung einiger bakterieller Pathogene. Diese Studie weist lediglich auf eine eingeschränkte Anzahl von möglichen Rekombinationsereignissen hin, die zur Evolution von *C. chauvoei* beigetragen haben.

Die „Maximum-Likelihood“ phylogenetische Sequenzanalyse basierend auf dem Core-Genom (Parsnp) identifizierte zusammengehörige Isolate mit hohen Bootstrap-Test-Werten (100%). Diese waren besonders markant bei Isolaten, die aus demselben Ausbruch/Betrieb/Tier stammten.

Referenzgenom- und Alignment-freie (kSNP3) Analysemethoden auf der Basis von Pan-Genom-SNPs (SNP, engl. Single Nucleotide Polymorphism, Einzelnukleotid-Polymorphismus) waren für das initiale Clustering von Isolaten geeignet. Das Clustering der

Stämme erfolgte anhand spezifischer SNPs. Die Anzahl allel-spezifischer SNPs, die für jeden Stamm einzigartig sind, war bei Stämmen exotischen Ursprungs höher.

Read-mapping und SNP-calling (Snippy), auf der Basis eines verwandten Referenzgenoms (12S0467) können neue Einblicke zum Verständnis der generellen, ausbruchsbezogenen und im Wirt ablaufenden Mikroevolution eines Erregers ermöglichen. Der mediane paarweise SNP-Differenzwert war für Stämme aus Bayern und Österreich im Vergleich zu Stämmen aus Niedersachsen höher. Die Diversität geographisch verwandter Stämme könnte darauf hinweisen, dass Umweltfaktoren oder das Herden-Management die Mikro-Evolution der Stämme beeinflussen. Nicht-synonyme SNPs trugen wesentlich zu der beobachteten Mikro-Evolution des Erregers bei. Die beobachtete Variabilität von Stämmen aus dem gleichen Tier und Stämmen aus dem gleichen Ausbruch zeigt, dass genetisch verschiedene *C. chauvoei* Populationen in einem Wirt vorhanden sein können.

Eine kürzlich publizierte, auf Genomsequenzdaten basierende Studie hat gezeigt, dass CRISPR-Spacer-Sequenzen von *C. chauvoei* für die Stamm-Differenzierung genutzt werden können. Die aktuelle Studie beweist die Eignung von CRISPR-Typing und Core-Genom MLST. Bei den hier untersuchten Stämmen war die Diversität der CRISPR-Spacer-Sequenzen nicht ausreichend, um Stämme europäischen Ursprungs weiter zu differenzieren. Mittels Core-Genom MLST war die Differenzierung von Stämmen mit europäischem und deutschem Ursprung möglich. Sie ist deshalb ein wertvolles Werkzeug zur Unterscheidung von Stämmen.

In der vorliegenden Arbeit wurde der Vorteil von Sequenzier-Techniken mit langen Leseweiten genutzt, um zwei qualitativ hochwertige, komplette Referenz-Genomsequenzen für ein Feldisolat aus Deutschland und den Typstamm von *C. chauvoei* zu erzeugen. Die phylogenetische Position der Spezies innerhalb des Genus wurde evaluiert und eine nahe Verwandtschaft zu *C. septicum* bestätigt. Verschiedene Genom-Analyse-Programme ermöglichten neue Einblicke in die Genomzusammensetzung und in den Genominhalt des Erregers. Eine vergleichende Genomsequenz-Analyse anhand von 64 Stämmen zeigte für den Erreger einen eingeschränkten horizontalen Gentransfer und eingeschränkten Erwerb neuer Gene. Die Typisierung von Stämmen auf der Basis der Core-Genom MLST, die zum ersten Mal für die Spezies untersucht wurde, ermöglichte sogar die Differenzierung von Stämmen auf Betriebsebene.

Chapter 7: References

ABREU, C. C., EDWARDS, E. E., EDWARDS, J. F., GIBBONS, P. M., LEAL DE ARAUJO, J., RECH, R. R. & UZAL, F. A. 2017. Blackleg in cattle: A case report of fetal infection and a literature review. *J Vet Diagn Invest*, 1040638717713796.

ABREU, C. C., UZAL, F. A., UZAL, F. A., SONGER, J. G., PRESCOTT, J. F. & POPOFF, M. R. 2016. Blackleg. *Clostridial Diseases of Animals.* John Wiley & Sons, Inc.

ADAMS, L. 1998. Animal health issues in South Texas cattle. *Workshop on beef cattle production systems and natural resources conservation in semi-arid lands of South Texas and Northern Mexico.* Universidad Auto'noma dr Tamaulipas, Cd. Victoria, Tamaulipas, Mexico.

ADESSI, C., MATTON, G., AYALA, G., TURCATTI, G., MERMOD, J.-J., MAYER, P. & KAWASHIMA, E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research,* 28, e87-e87.

ADEY, A. & SHENDURE, J. 2012. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res,* 22, 1139-43.

ÅGREN, J., SUNDSTRÖM, A., HÅFSTRÖM, 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, e39107.

ALAM, S. I., DIXIT, A., TOMAR, A. & SINGH, L. 2010. Comparative genomic analysis of a neurotoxigenic *Clostridium* species using partial genome sequence: Phylogenetic analysis of a few conserved proteins involved in cellular processes and metabolism. *Anaerobe,* 16, 147-154.

ALMEIDA E MACÊDO, J. T. S., PIRES, P. S., PINHEIRO, E. E. G., OLIVEIRA, R. S. D., SILVA, R. O. S., LOBATO, F. C. F. & PEDROSO, P. M. O. 2013. Malignant edema caused by *Clostridium chauvoei* in a horse. *Acta Scientiae Veterinariae,* 41, 24.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25.

ANDERSEN, J. M., SHOUP, M., ROBINSON, C., BRITTON, R., OLSEN, K. E. P. & BARRANGOU, R. 2016. CRISPR diversity and microevolution in *Clostridium difficile*. *Genome Biology and Evolution*.

ANGIUOLI, S. V. & SALZBERG, S. L. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics,* 27, 334-342.

ANNIBALLI, F., FILLO, S., GIORDANI, F., AURICCHIO, B., TEHRAN, D. A., DI STEFANO, E., MANDARINO, G., DE MEDICI, D. & LISTA, F. 2016. Multiple-locus variable number of tandem repeat analysis as a tool for molecular epidemiology of botulism: The Italian experience. *Infect Genet Evol,* 46, 28-32.

ARMSTRONG, H. L. & MACNAMEE, J. K. 1950. Blackleg in deer. *J Am Vet Med Assoc,* 117, 212-4.

ARNDT, D., GRANT, J. R., MARCU, A., SAJED, T., PON, A., LIANG, Y. & WISHART, D. S. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*.

ASSEFA, S., KEANE, T. M., OTTO, T. D., NEWBOLD, C. & BERRIMAN, M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics,* 25.

AYELE, B., TIGRE, W. & DERESSA, B. 2016. Epidemiology and financial loss estimation of blackleg on smallholder cattle herders in Kembata Tambaro zone, Southern Ethiopia. *SpringerPlus,* 5**,** 1822.

AZIZ, R. K., BARTELS, D., BEST, A. A., DEJONGH, M., DISZ, T., EDWARDS, R. A., FORMSMA, K., GERDES, S., GLASS, E. M., KUBAL, M., MEYER, F., OLSEN, G. J., OLSON, R., OSTERMAN, A. L., OVERBEEK, R. A., MCNEIL, L. K., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G. D., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A. & ZAGNITKO, O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics,* 9**,** 75.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A., DVORKIN, M., KULIKOV, A., LESIN, V., NIKOLENKO, S., PHAM, S., PRJIBELSKI, A., PYSHKIN, A., SIROTKIN, A., VYAHHI, N., TESLER, G., ALEKSEYEV, M. & PEVZNER, P. 2012. SPAdes: a new genome assembler and its applications to single cell sequencing. *Journal of Computational Biology,* 19.

BARNES, D. M., BERGELAND, M. E. & HIGBEE, J. M. 1975. Differential diagnosis of clostridial myonecrosis. *The Canadian Veterinary Journal,* 16**,** 357-359.

BARRANGOU, R., FREMAUX, C., DEVEAU, H., RICHARDS, M., BOYAVAL, P., MOINEAU, S., ROMERO, D. A. & HORVATH, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* 315**,** 1709-12.

BATTY, I. & WALKER, P. D. 1963. Differentiation of *Clostridium septicum* and *Clostridium chauvoei* by the use of fluorescent labelled antibodies. *The Journal of Pathology and Bacteriology,* 85**,** 517-521.

BENAMAR, S., CASSIR, N., CAPUTO, A., CADORET, F. & LA SCOLA, B. 2016. Complete Genome Sequence of *Clostridium septicum* Strain CSUR P1044, Isolated from the Human Gut Microbiota. *Genome Announcements,* 4.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L. & PRATT, M. R. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456.

BERNARD, G., RAGAN, M. A. & CHAN, C. X. 2016. Recapitulating phylogenies using k-mers: from trees to networks. *F1000Research,* 5**,** 2789.

BERTELLI, C., LAIRD, M. R., WILLIAMS, K. P., LAU, B. Y., HOAD, G., WINSOR, G. L. & BRINKMAN, F. S. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res*.

BINNEWIES, T. T., MOTRO, Y., HALLIN, P. F., LUND, O., DUNN, D., LA, T., HAMPSON, D. J., BELLGARD, M., WASSENAAR, T. M. & USSERY, D. W. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics,* 6**,** 165-85.

BLAND, C., RAMSEY, T. L., SABREE, F., LOWE, M., BROWN, K., KYRPIDES, N. C. & HUGENHOLTZ, P. 2007. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics,* 8**,** 1-8.

BRIGGS, G. S., SMITS, W. K. & SOULTANAS, P. 2012. Chromosomal Replication Initiation Machinery of Low-G+C-Content Firmicutes. *Journal of Bacteriology,* 194**,** 5162-5170.

BROUWER, M. S., ROBERTS, A. P., HUSSAIN, H., WILLIAMS, R. J., ALLAN, E. & MULLANY, P. 2013. Horizontal gene transfer converts non-toxigenic *Clostridium difficile* strains into toxin producers. *Nat Commun,* 4.

BROUWER, M. S. M., ROBERTS, A. P., MULLANY, P. & ALLAN, E. 2012. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. *Mobile Genetic Elements,* 2**,** 8-12.

BRUGGEMANN, H., BAUMER, S., FRICKE, W. F., WIEZER, A., LIESEGANG, H., DECKER, I., HERZBERG, C., MARTINEZ-ARIAS, R., MERKL, R., HENNE, A. & GOTTSCHALK, G. 2003. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc Natl Acad Sci U S A,* 100**,** 1316-21.

BUERMANS, H. P. J. & DEN DUNNEN, J. T. 2014. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease,* 1842**,** 1932-1941.

CANO, R. J. & BORUCKI, M. K. 1995. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. *Science,* 268**,** 1060-4.

CARLONI, G. H., BENTANCOR, L. D. & DE TORRES, R. A. 2005. [Deoxyribonuclease activity detection in *Clostridium chauvoei* strains]. *Rev Argent Microbiol,* 37**,** 87-8.

CARTER, A. T., AUSTIN, J. W., WEEDMARK, K. A. & PECK, M. W. 2016. Evolution of Chromosomal *Clostridium botulinum* Type E Neurotoxin Gene Clusters: Evidence Provided by Their Rare Plasmid-Borne Counterparts. *Genome Biology and Evolution,* 8**,** 540-555.

CARTER, A. T. & PECK, M. W. 2015. Genomes, neurotoxins and biology of *Clostridium botulinum* Group I and Group II. *Research in Microbiology,* 166**,** 303-317.

CARTER, G. P., CHEUNG, J. K., LARCOMBE, S. & LYRAS, D. 2014. Regulation of toxin production in the pathogenic clostridia. *Molecular Microbiology,* 91**,** 221-231.

CARVER, T., THOMSON, N., BLEASBY, A., BERRIMAN, M. & PARKHILL, J. 2009. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics,* 25**,** 119-20.

CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., RAJANDREAM, M. A., BARRELL, B. G. & PARKHILL, J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics,* 21**,** 3422-3.

CHAN, C. X., BERNARD, G., POIRION, O., HOGAN, J. M. & RAGAN, M. A. 2014. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Scientific Reports,* 4**,** 6504.

CHANDLER, H. M. 1975. Rabbit immunoglobulin responses to the flagella, somatic, and protective antigens of a highly protective strain of *Clostridium chauvoei*. *Infect Immun,* 12**,** 143-7.

CHATIKOBO, P., CHOGA, T., NCUBE, C. & MUTAMBARA, J. 2013. Participatory diagnosis and prioritization of constraints to cattle production in some smallholder farming areas of Zimbabwe. *Prev Vet Med,* 109**,** 327-33.

CHIN, C. S., ALEXANDER, D. H., MARKS, P., KLAMMER, A. A., DRAKE, J., HEINER, C., CLUM, A., COPELAND, A., HUDDLESTON, J., EICHLER, E. E., TURNER, S.

W. & KORLACH, J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods,* 10**,** 563-9.

COLLINS, M. D., LAWSON, P. A., WILLEMS, A., CORDOBA, J. J., FERNANDEZ-GARAYZABAL, J., GARCIA, P., CAI, J., HIPPE, H. & FARROW, J. A. 1994. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol,* 44**,** 812-26.

CONESA, A. & GOTZ, S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics,* 2008**,** 619832.

CONESA, A., GOTZ, S., GARCIA-GOMEZ, J. M., TEROL, J., TALON, M. & ROBLES, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21**,** 3674-3676.

COUCHMAN, E. C., BROWNE, H. P., DUNN, M., LAWLEY, T. D., SONGER, J. G., HALL, V., PETROVSKA, L., VIDOR, C., AWAD, M., LYRAS, D. & FAIRWEATHER, N. F. 2015. *Clostridium sordellii* genome analysis reveals plasmid localized toxin genes encoded within pathogenicity loci. *BMC Genomics,* 16**,** 392.

CROUCHER, N. J., FOOKES, M. C., PERKINS, T. T., TURNER, D. J., MARGUERAT, S. B., KEANE, T., QUAIL, M. A., HE, M., ASSEFA, S., BAHLER, J., KINGSLEY, R. A., PARKHILL, J., BENTLEY, S. D., DOUGAN, G. & THOMSON, N. R. 2009. A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res,* 37.

CROUCHER, N. J., PAGE, A. J., CONNOR, T. R., DELANEY, A. J., KEANE, J. A., BENTLEY, S. D., PARKHILL, J. & HARRIS, S. R. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res,* 43**,** e15.

DARLING, A. C., MAU, B., BLATTNER, F. R. & PERNA, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res,* 14**,** 1394-403.

DARLING, A. E., MAU, B. & PERNA, N. T. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE,* 5**,** e11147.

DELCHER, A. L., BRATKE, K. A., POWERS, E. C. & SALZBERG, S. L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics,* 23**,** 673-9.

DIDELOT, X., BOWDEN, R., WILSON, D. J., PETO, T. E. & CROOK, D. W. 2012. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet,* 13**,** 601-12.

DIDELOT, X., WALKER, A. S., PETO, T. E., CROOK, D. W. & WILSON, D. J. 2016. Within-host evolution of bacterial pathogens. *Nat Rev Micro,* 14**,** 150-162.

DOUADY, C. J., DELSUC, F., BOUCHER, Y., DOOLITTLE, W. F. & DOUZERY, E. J. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol,* 20**,** 248-54.

DRESSMAN, D., YAN, H., TRAVERSO, G., KINZLER, K. W. & VOGELSTEIN, B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences,* 100**,** 8817-8822.

DURRE, P. 2014. Physiology and Sporulation in *Clostridium*. *Microbiol Spectr,* 2**,** Tbs-0010-2012.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P., BETTMAN, B., BIBILLO, A., BJORNSON, K., CHAUDHURI,

B., CHRISTIANS, F., CICERO, R., CLARK, S., DALAL, R., DEWINTER, A., DIXON, J., FOQUET, M., GAERTNER, A., HARDENBOL, P., HEINER, C., HESTER, K., HOLDEN, D., KEARNS, G., KONG, X., KUSE, R., LACROIX, Y., LIN, S., LUNDQUIST, P., MA, C., MARKS, P., MAXHAM, M., MURPHY, D., PARK, I., PHAM, T., PHILLIPS, M., ROY, J., SEBRA, R., SHEN, G., SORENSON, J., TOMANEY, A., TRAVERS, K., TRULSON, M., VIECELI, J., WEGENER, J., WU, D., YANG, A., ZACCARIN, D., ZHAO, P., ZHONG, F., KORLACH, J. & TURNER, S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science,* 323**,** 133-8.

EISEN, J. A., HEIDELBERG, J. F., WHITE, O. & SALZBERG, S. L. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol,* 1.

FALQUET, L., CALDERON-COPETE, S. P. & FREY, J. 2013. Draft Genome Sequence of the Virulent *Clostridium chauvoei* Reference Strain JF4335. *Genome Announc,* 1.

FAN, S., ZHANG, H., CHEN, X., LU, L., XU, L. & XIAO, M. 2016. Cloning, characterization, and production of three alpha-l-fucosidases from *Clostridium perfringens* ATCC 13124. *J Basic Microbiol,* 56**,** 347-57.

FANG, P. K., RAPHAEL, B. H., MASLANKA, S. E., CAI, S. & SINGH, B. R. 2010. Analysis of genomic differences among *Clostridium botulinum* type A1 strains. *BMC Genomics,* 11**,** 725.

FEDURCO, M., ROMIEU, A., WILLIAMS, S., LAWRENCE, I. & TURCATTI, G. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research,* 34**,** e22-e22.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A. & MERRICK, J. M. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae*Rd. *Science,* 269.

FLICEK, P. & BIRNEY, E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nat Meth,* 6**,** S6-S12.

FORDE, T., BIEK, R., ZADOKS, R., WORKENTINE, M. L., DE BUCK, J., KUTZ, S., OPRIESSNIG, T., TREWBY, H., VAN DER MEER, F. & ORSEL, K. 2016. Genomic analysis of the multi-host pathogen *Erysipelothrix rhusiopathiae* reveals extensive recombination as well as the existence of three generalist clades with wide geographic distribution. *BMC Genomics,* 17**,** 461.

FORTIER, L.-C. & SEKULOVIC, O. 2013. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence,* 4**,** 354-365.

FOUTS, D. E., BRINKAC, L., BECK, E., INMAN, J. & SUTTON, G. 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res,* 40**,** e172.

FRASER, C. M., GOCAYNE, J. D., WHITE, O., ADAMS, M. D., CLAYTON, R. A., FLEISCHMANN, R. D., BULT, C. J., KERLAVAGE, A. R., SUTTON, G. & KELLEY, J. M. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science,* 270.

FREEDMAN, J. C., THEORET, J. R., WISNIEWSKI, J. A., UZAL, F. A., ROOD, J. I. & MCCLANE, B. A. 2015. *Clostridium perfringens* type A–E toxin plasmids. *Research in microbiology,* 166**,** 264-279.

FREY, J. & FALQUET, L. 2015. Patho-genetics of *Clostridium chauvoei*. *Res Microbiol,* 166**,** 384-92.

FREY, J., JOHANSSON, A., BURKI, S., VILEI, E. M. & REDHEAD, K. 2012. Cytotoxin CctA, a major virulence factor of *Clostridium chauvoei* conferring protective immunity against myonecrosis. *Vaccine,* 30**,** 5500-5.

GAO, F. & ZHANG, C. T. 2008. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics,* 9**,** 79.

GARDNER, S. N., SLEZAK, T. & HALL, B. G. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics,* 31**,** 2877-2878.

GAROFOLO, G., GALANTE, D., SERRECCHIA, L., BUONAVOGLIA, D. & FASANELLA, A. 2011. Development of a real time PCR taqman assay based on the TPI gene for simultaneous identification of *Clostridium chauvoei* and *Clostridium septicum*. *Journal of Microbiological Methods,* 84**,** 307-311.

GARRISON, E. & MARTH, G. 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints,* 1207.

GARY VAN DOMSELAAR, M. G. A. P. S. 2014. Prokaryotic Genome Annotation. *In:* BISHOP, Ö. T. (ed.) *Bioinformatics and Data Analysis in Microbiology.* Caister Academic Press.

GIGLIO, M. G., COLLMER, C. W., LOMAX, J. & IRELAND, A. 2009. Applying the Gene Ontology in microbial annotation. *Trends Microbiol,* 17**,** 262-8.

GOLUBCHIK, T., BATTY, E. M., MILLER, R. R., FARR, H., YOUNG, B. C., LARNER-SVENSSON, H., FUNG, R., GODWIN, H., KNOX, K., VOTINTSEVA, A., EVERITT, R. G., STREET, T., CULE, M., IP, C. L. C., DIDELOT, X., PETO, T. E. A., HARDING, R. M., WILSON, D. J., CROOK, D. W. & BOWDEN, R. 2013. Within-Host Evolution of *Staphylococcus aureus* during Asymptomatic Carriage. *PLOS ONE,* 8**,** e61319.

GOTZ, S., ARNOLD, R., SEBASTIAN-LEON, P., MARTIN-RODRIGUEZ, S., TISCHLER, P., JEHL, M. A., DOPAZO, J., RATTEI, T. & CONESA, A. 2011. B2G-FAR, a species-centered GO annotation repository. *Bioinformatics,* 27**,** 919-24.

GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., WILLIAMS, T. D., NAGARAJ, S. H., NUEDA, M. J., ROBLES, M., TALÓN, M., DOPAZO, J. & CONESA, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research,* 36**,** 3420-3435.

GUPTA, R. S. & GAO, B. 2009. Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I). *Int J Syst Evol Microbiol,* 59**,** 285-94.

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29.

HAGAN, W. A., BRUNER, D. W. & TIMONEY, J. F. 1988. *Hagan and Bruner's Microbiology and Infectious Diseases of Domestic Animals: With Reference to Etiology, Epizootiology, Pathogenesis, Immunity, Diagnosis, and Antimicrobial Susceptibility*, Comstock Pub. Associates.

HAIKO, J. & WESTERLUND-WIKSTRÖM, B. 2013. The Role of the Bacterial Flagellum in Adhesion and Virulence. *Biology,* 2**,** 1242-1267.

HALM, A., WAGNER, M., KÖFER, J. & HEIN, I. 2010. Novel Real-Time PCR Assay for Simultaneous Detection and Differentiation of *Clostridium chauvoei* and *Clostridium*

septicum in Clostridial Myonecrosis. *Journal of Clinical Microbiology,* 48**,** 1093-1098.

HAMAOKA, T. & TERAKADO, N. 1994. Demonstration of common antigens on cell surface of *Clostridium chauvoei* and *C. septicum* by indirect-immunofluorescence assay. *J Vet Med Sci,* 56**,** 371-3.

HARGREAVES, K. R., KROPINSKI, A. M. & CLOKIE, M. R. J. 2014. What Does the Talking?: Quorum Sensing Signalling Genes Discovered in a Bacteriophage Genome. *PLOS ONE,* 9**,** e85131.

HARWOOD, D. G., HIGGINS, R. J. & AGGETT, D. J. 2007. Outbreak of intestinal and lingual *Clostridium chauvoei infection* in two-year-old Friesian heifers. *Vet Rec,* 161**,** 307-8.

HASSAN, K. A., ELBOURNE, L. D., TETU, S. G., MELVILLE, S. B., ROOD, J. I. & PAULSEN, I. T. 2015. Genomic analyses of *Clostridium perfringens* isolates from five toxinotypes. *Res Microbiol,* 166**,** 255-63.

HATHEWAY, C. L. 1990. Toxigenic clostridia. *Clin Microbiol Rev,* 3.

HE, M., SEBAIHIA, M., LAWLEY, T. D., STABLER, R. A., DAWSON, L. F., MARTIN, M. J., HOLT, K. E., SETH-SMITH, H. M. B., QUAIL, M. A., RANCE, R., BROOKS, K., CHURCHER, C., HARRIS, D., BENTLEY, S. D., BURROWS, C., CLARK, L., CORTON, C., MURRAY, V., ROSE, G., THURSTON, S., VAN TONDER, A., WALKER, D., WREN, B. W., DOUGAN, G. & PARKHILL, J. 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proceedings of the National Academy of Sciences,* 107**,** 7527-7532.

HEUERMANN, D., ROGGENTIN, P., KLEINEIDAM, R. G. & SCHAUER, R. 1991. Purification and characterization of a sialidase from *Clostridium chauvoei* NC08596. *Glycoconj J,* 8**,** 95-101.

HINGSTON, P., CHEN, J., DHILLON, B. K., LAING, C., BERTELLI, C., GANNON, V., TASARA, T., ALLEN, K., BRINKMAN, F. S. L., TRUELSTRUP HANSEN, L. & WANG, S. 2017. Genotypes Associated with *Listeria monocytogenes* Isolates Displaying Impaired or Enhanced Tolerances to Cold, Salt, Acid, or Desiccation Stress. *Frontiers in Microbiology,* 8**,** 369.

HIRSH, D. C. & ZEE, Y. C. 1999. *Veterinary Microbiology and Immunology*, Wiley.

HORVATH, P., ROMERO, D. A., COUTE-MONVOISIN, A. C., RICHARDS, M., DEVEAU, H., MOINEAU, S., BOYAVAL, P., FREMAUX, C. & BARRANGOU, R. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol,* 190**,** 1401-12.

HUNT, M., KIKUCHI, T., SANDERS, M., NEWBOLD, C., BERRIMAN, M. & OTTO, T. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biol,* 14.

HUNT, M., SILVA, N. D., OTTO, T. D., PARKHILL, J., KEANE, J. A. & HARRIS, S. R. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology,* 16**,** 1-10.

HUSON, D. H. & BRYANT, D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol,* 23**,** 254-67.

HUSON, D. H. & SCORNAVACCA, C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol,* 61**,** 1061-7.

HYNES, W. L. & WALTON, S. L. 2000. Hyaluronidases of Gram-positive bacteria. *FEMS Microbiology Letters,* 183**,** 201-207.

JAYARAMAIAH, U., SINGH, N., THANKAPPAN, S., MOHANTY, A. K., CHAUDHURI, P., SINGH, V. P. & NAGALEEKAR, V. K. 2016. Proteomic analysis and identification of cell surface-associated proteins of *Clostridium chauvoei*. *Anaerobe,* 39**,** 77-83.

JOSHI NA, F. J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files

JÜNEMANN, S., PRIOR, K., ALBERSMEIER, A., ALBAUM, S., KALINOWSKI, J., GOESMANN, A., STOYE, J. & HARMSEN, D. 2014. GABenchToB: A Genome Assembly Benchmark Tuned on Bacteria and Benchtop Sequencers. *PLoS ONE,* 9**,** e107014.

JUNEMANN, S., SEDLAZECK, F. J., PRIOR, K., ALBERSMEIER, A., JOHN, U., KALINOWSKI, J., MELLMANN, A., GOESMANN, A., VON HAESELER, A., STOYE, J. & HARMSEN, D. 2013. Updating benchtop sequencing performance comparison. *Nat Biotech,* 31**,** 294-296.

KALIA, V. C., MUKHERJEE, T., BHUSHAN, A., JOSHI, J., SHANKAR, P. & HUMA, N. 2011. Analysis of the unexplored features of rrs (16S rDNA) of the Genus *Clostridium*. *BMC Genomics,* 12**,** 18.

KATZ, L. S., GRISWOLD, T., WILLIAMS-NEWKIRK, A. J., WAGNER, D., PETKAU, A., SIEFFERT, C., VAN DOMSELAAR, G., DENG, X. & CARLETON, H. A. 2017. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Frontiers in Microbiology,* 8.

KEARSE, M., MOIR, R., WILSON, A., STONES-HAVAS, S., CHEUNG, M., STURROCK, S., BUXTON, S., COOPER, A., MARKOWITZ, S., DURAN, C., THIERER, T., ASHTON, B., MEINTJES, P. & DRUMMOND, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28**,** 1647-1649.

KETO-TIMONEN, R., HEIKINHEIMO, A., EEROLA, E. & KORKEALA, H. 2006. Identification of Clostridium Species and DNA Fingerprinting of *Clostridium perfringens* by Amplified Fragment Length Polymorphism Analysis. *Journal of Clinical Microbiology,* 44**,** 4057-4065.

KLEMM, E. & DOUGAN, G. 2016. Advances in Understanding Bacterial Pathogenesis Gained from Whole-Genome Sequencing and Phylogenetics. *Cell Host & Microbe,* 19**,** 599-610.

KNIGHT, D. R., ELLIOTT, B., CHANG, B. J., PERKINS, T. T. & RILEY, T. V. 2015. Diversity and Evolution in the Genome of *Clostridium difficile*. *Clinical Microbiology Reviews,* 28**,** 721-741.

KOJIMA, A., AMIMOTO, K., OHGITANI, T. & TAMURA, Y. 1999. Characterization of flagellin from *Clostridium chauvoei*. *Veterinary Microbiology,* 67**,** 231-237.

KOJIMA, A., UCHIDA, I., SEKIZAKI, T., SASAKI, Y., OGIKUBO, Y., KIJIMA, M. & TAMURA, Y. 2000. Cloning and expression of a gene encoding the flagellin of *Clostridium chauvoei*. *Vet Microbiol,* 76**,** 359-72.

KOJIMA, A., UCHIDA, I., SEKIZAKI, T., SASAKI, Y., OGIKUBO, Y. & TAMURA, Y. 2001. Rapid detection and identification of *Clostridium chauvoei* by PCR based on flagellin gene sequence. *Veterinary Microbiology,* 78**,** 363-371.

KOREN, S., HARHAY, G. P., SMITH, T. P., BONO, J. L., HARHAY, D. M., MCVEY, S. D., RADUNE, D., BERGMAN, N. H. & PHILLIPPY, A. M. 2013. Reducing

assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology,* 14**,** 1-16.

KUBIAK, A. M. & MINTON, N. P. 2015. The potential of clostridial spores as therapeutic delivery vehicles in tumour therapy. *Research in Microbiology,* 166**,** 244-254.

KUHNERT, P., CAPAUL, S. E., NICOLET, J. & FREY, J. 1996. Phylogenetic positions of *Clostridium chauvoei* and *Clostridium septicum* based on 16S rRNA gene sequences. *Int J Syst Bacteriol,* 46**,** 1174-6.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol,* 5.

LAGESEN, K., HALLIN, P., RODLAND, E. A., STAERFELDT, H. H., ROGNES, T. & USSERY, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res,* 35**,** 3100-8.

LANGE, M., NEUBAUER, H. & SEYBOLDT, C. 2010. Development and validation of a multiplex real-time PCR for detection of *Clostridium chauvoei* and *Clostridium septicum*. *Mol Cell Probes,* 24**,** 204-10.

LANGFORD, E. V. 1970. Feed-borne *Clostridium chauvoei* infection in mink. *Can Vet J,* 11**,** 170-2.

LAWSON, P. A., CITRON, D. M., TYRRELL, K. L. & FINEGOLD, S. M. 2016. Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prevot 1938. *Anaerobe,* 40**,** 95-9.

LEEKITCHAROENPHON, P., NIELSEN, E. M., KAAS, R. S., LUND, O. & AARESTRUP, F. M. 2014. Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. *PLOS ONE,* 9**,** e87991.

LETUNIC, I. & BORK, P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res,* 44**,** W242-5.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

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.

LI, Z., CHEN, Y., MU, D., YUAN, J., SHI, Y., ZHANG, H., GAN, J., LI, N., HU, X., LIU, B., YANG, B. & FAN, W. 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics,* 11**,** 25-37.

LIAO, Y. C., LIN, S. H. & LIN, H. H. 2015. Completing bacterial genome assemblies: strategy and performance comparisons. *Sci Rep,* 5**,** 8747.

LIEBERMAN, T. D., FLETT, K. B., YELIN, I., MARTIN, T. R., MCADAM, A. J., PRIEBE, G. P. & KISHONY, R. 2014. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet,* 46**,** 82-7.

LINZ, B., WINDSOR, H. M., MCGRAW, J. J., HANSEN, L. M., GAJEWSKI, J. P., TOMSHO, L. P., HAKE, C. M., SOLNICK, J. V., SCHUSTER, S. C. & MARSHALL, B. J. 2014. A mutation burst during the acute phase of *Helicobacter pylori* infection in humans and rhesus macaques. 5**,** 4165.

LOWE, T. M. & EDDY, S. R. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research,* 25**,** 0955-964.

LUBLIN, A., MECHANI, S., HOROWITZ, H. & WEISMAN, Y. 1993. A paralytic-like disease of the ostrich (Struthio camelus masaicus) associated with *Clostridium chauvoei* infection. *Veterinary Record,* 132**,** 273-275.

LUQUEZ, C., JOSEPH, L. A. & MASLANKA, S. E. 2015. Molecular subtyping of *Clostridium botulinum* by pulsed-field gel electrophoresis. *Methods Mol Biol,* 1301**,** 103-13.

MACLENNAN, J. D. 1962. THE HISTOTOXIC CLOSTRIDIAL INFECTIONS OF MAN. *Bacteriological Reviews,* 26**,** 177-274.

MAGOC, T., PABINGER, S., CANZAR, S., LIU, X., SU, Q., PUIU, D., TALLON, L. J. & SALZBERG, S. L. 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics,* 29.

MAIDEN, M. C. J., JANSEN VAN RENSBURG, M. J., BRAY, J. E., EARLE, S. G., FORD, S. A., JOLLEY, K. A. & MCCARTHY, N. D. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature reviews. Microbiology,* 11**,** 728-736.

MAKAROVA, K. S., WOLF, Y. I., ALKHNBASHI, O. S., COSTA, F., SHAH, S. A., SAUNDERS, S. J., BARRANGOU, R., BROUNS, S. J. J., CHARPENTIER, E., HAFT, D. H., HORVATH, P., MOINEAU, S., MOJICA, F. J. M., TERNS, R. M., TERNS, M. P., WHITE, M. F., YAKUNIN, A. F., GARRETT, R. A., VAN DER OOST, J., BACKOFEN, R. & KOONIN, E. V. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Micro,* 13**,** 722-736.

MARSHALL, K. M., BRADSHAW, M., PELLETT, S. & JOHNSON, E. A. 2007. Plasmid Encoded Neurotoxin Genes in *Clostridium botulinum* Serotype A Subtypes. *Biochemical and biophysical research communications,* 361**,** 49-54.

MARTTINEN, P., HANAGE, W. P., CROUCHER, N. J., CONNOR, T. R., HARRIS, S. R., BENTLEY, S. D. & CORANDER, J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research,* 40**,** e6-e6.

MCARTHUR, A. G., WAGLECHNER, N., NIZAM, F., YAN, A., AZAD, M. A., BAYLAY, A. J., BHULLAR, K., CANOVA, M. J., DE PASCALE, G., EJIM, L., KALAN, L., KING, A. M., KOTEVA, K., MORAR, M., MULVEY, M. R., O'BRIEN, J. S., PAWLOWSKI, A. C., PIDDOCK, L. J., SPANOGIANNOPOULOS, P., SUTHERLAND, A. D., TANG, I., TAYLOR, P. L., THAKER, M., WANG, W., YAN, M., YU, T. & WRIGHT, G. D. 2013. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother,* 57**,** 3348-57.

MEANEY, C. A., CARTMAN, S. T., MCCLURE, P. J. & MINTON, N. P. 2016. The role of small acid-soluble proteins (SASPs) in protection of spores of *Clostridium botulinum* against nitrous acid. *Int J Food Microbiol,* 216**,** 25-30.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Curr Opin Genet Dev,* 15**,** 589-94.

MORRIS, W. E., UZAL, F. A., FATTORINI, F. R. & TERZOLO, H. 2002a. Malignant oedema associated with blood-sampling in sheep. *Aust Vet J,* 80**,** 280-1.

MORRIS, W. E., UZAL, F. A. & PARAMIDANI, M. 2002b. Malignant oedema associated with navel infection in a Merino lamb. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia,* 54**,** 448-449.

MOUSSA, R. S. 1958. COMPLEXITY OF TOXINS FROM *CLOSTRIDIUM SEPTICUM* AND *CLOSTRIDIUM CHAUVOEI*. *Journal of Bacteriology,* 76**,** 538-545.

MUDENDA HANG'OMBE, B., KOHDA, T., MUKAMOTO, M. & KOZAKI, S. 2006. Purification and sensitivity of *Clostridium chauvoei* hemolysin to various erythrocytes. *Comp Immunol Microbiol Infect Dis,* 29**,** 263-8.

MURPHY, D. B. 1980. *Clostridium chauvoei* as the cause of malignant edema in a horse. *Vet Med Small Anim Clin,* 75**,** 1152-4.

MYERS, G. S., RASKO, D. A., CHEUNG, J. K., RAVEL, J., SESHADRI, R., DEBOY, R. T., REN, Q., VARGA, J., AWAD, M. M., BRINKAC, L. M., DAUGHERTY, S. C., HAFT, D. H., DODSON, R. J., MADUPU, R., NELSON, W. C., ROSOVITZ, M. J., SULLIVAN, S. A., KHOURI, H., DIMITROV, G. I., WATKINS, K. L., MULLIGAN, S., BENTON, J., RADUNE, D., FISHER, D. J., ATKINS, H. S., HISCOX, T., JOST, B. H., BILLINGTON, S. J., SONGER, J. G., MCCLANE, B. A., TITBALL, R. W., ROOD, J. I., MELVILLE, S. B. & PAULSEN, I. T. 2006. Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*. *Genome Res,* 16**,** 1031-40.

NAGANO, N., ISOMINE, S., KATO, H., SASAKI, Y., TAKAHASHI, M., SAKAIDA, K., NAGANO, Y. & ARAKAWA, Y. 2008. Human fulminant gas gangrene caused by *Clostridium chauvoei*. *J Clin Microbiol,* 46**,** 1545-7.

NG, V. & LIN, W.-J. 2014. Comparison of assembled *Clostridium botulinum* A1 genomes revealed their evolutionary relationship. *Genomics,* 103**,** 94-106.

NIKOLENKO, S. I., KOROBEYNIKOV, A. I. & ALEKSEYEV, M. A. 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics,* 14**,** S7.

NOLLING, J., BRETON, G., OMELCHENKO, M. V., MAKAROVA, K. S., ZENG, Q., GIBSON, R., LEE, H. M., DUBOIS, J., QIU, D., HITTI, J., WOLF, Y. I., TATUSOV, R. L., SABATHE, F., DOUCETTE-STAMM, L., SOUCAILLE, P., DALY, M. J., BENNETT, G. N., KOONIN, E. V. & SMITH, D. R. 2001. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J Bacteriol,* 183**,** 4823-38.

NOWELL, V. J., KROPINSKI, A. M., SONGER, J. G., MACINNES, J. I., PARREIRA, V. R. & PRESCOTT, J. F. 2012. Genome Sequencing and Analysis of a Type A *Clostridium perfringens* Isolate from a Case of Bovine Clostridial Abomasitis. *PLoS ONE,* 7**,** e32271.

NUM, S. M. U., NICODEMUS M. 2014. *Clostridium*: Pathogenic Roles, Industrial Uses and Medicinal Prospects of Natural Products as Ameliorative Agents against Pathogenic Species. *ordan Journal of Biological Sciences,* Vol. 7**,** 81.

ODANI, J. S., BLANCHARD, P. C., ADASKA, J. M., MOELLER, R. B. & UZAL, F. A. 2009. Malignant edema in postpartum dairy cattle. *J Vet Diagn Invest,* 21**,** 920-4.

ORSBURN, B. C., MELVILLE, S. B. & POPHAM, D. L. 2010. EtfA catalyses the formation of dipicolinic acid in *Clostridium perfringens*. *Mol Microbiol,* 75**,** 178-86.

OTTO, T. D., SANDERS, M., BERRIMAN, M. & NEWBOLD, C. 2010. Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics,* 26.

OVERBEEK, R., OLSON, R., PUSCH, G. D., OLSEN, G. J., DAVIS, J. J., DISZ, T., EDWARDS, R. A., GERDES, S., PARRELLO, B., SHUKLA, M., VONSTEIN, V., WATTAM, A. R., XIA, F. & STEVENS, R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research,* 42**,** D206-D214.

PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T. G., FOOKES, M., FALUSH, D., KEANE, J. A. & PARKHILL, J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*.

PAREDES-SABJA, D., RAJU, D., TORRES, J. A. & SARKER, M. R. 2008a. Role of small, acid-soluble spore proteins in the resistance of *Clostridium perfringens* spores to chemicals. *Int J Food Microbiol,* 122**,** 333-5.

PAREDES-SABJA, D., SARKER, N. & SARKER, M. R. 2011. *Clostridium perfringens* tpeL is expressed during sporulation. *Microb Pathog,* 51**,** 384-8.

PAREDES-SABJA, D., SETLOW, P. & SARKER, M. R. 2009. Role of GerKB in germination and outgrowth of *Clostridium perfringens* spores. *Appl Environ Microbiol,* 75**,** 3813-7.

PAREDES-SABJA, D., TORRES, J. A., SETLOW, P. & SARKER, M. R. 2008b. *Clostridium perfringens* spore germination: characterization of germinants and their receptors. *J Bacteriol,* 190**,** 1190-201.

PETTIT, L. J., BROWNE, H. P., YU, L., SMITS, W. K., FAGAN, R. P., BARQUIST, L., MARTIN, M. J., GOULDING, D., DUNCAN, S. H., FLINT, H. J., DOUGAN, G., CHOUDHARY, J. S. & LAWLEY, T. D. 2014. Functional genomics reveals that *Clostridium difficile*Spo0A coordinates sporulation, virulence and metabolism. *BMC Genomics,* 15**,** 1-15.

PILEHCHIAN LANGROUDI, R. 2015. Isolation, Specification, Molecular Biology Assessment and Vaccine Development of *Clostridium* in Iran: A Review. *Int J Enteric Pathog,* 3**,** e28979.

PIRES, P. S., ECCO, R., SILVA, R. O. S., ARAÚJO, M. R. D., SALVARANI, F. M., HENEINE, L. G. D., OLIVEIRA JÚNIOR, C. A. D. & LOBATO, F. C. F. 2017a. A retrospective study on the diagnosis of clostridial myonecrosis in ruminants in Brazil. *Ciência Rural,* 47.

PIRES, P. S., SANTOS, R. L., DA PAIXÃO, T. A., DE OLIVEIRA BERNARDES, L. C., DE MACÊDO, A. A., GONÇALVES, L. A., DE OLIVEIRA JÚNIOR, C. A., SILVA, R. O. S. & LOBATO, F. C. F. 2017b. Intracellular survival of *Clostridium chauvoei* in bovine macrophages. *Veterinary Microbiology,* 199**,** 1-7.

PORCHERON, G., GARÉNAUX, A., PROULX, J., SABRI, M. & DOZOIS, C. M. 2013. Iron, copper, zinc, and manganese transport and regulation in pathogenic Enterobacteria: correlations between strains, site of infection and the relative importance of the different metal transport systems for virulence. *Frontiers in Cellular and Infection Microbiology,* 3**,** 90.

PRESCOTT, J. F., UZAL, F. A., SONGER, J. G., PRESCOTT, J. F. & POPOFF, M. R. 2016. Brief Description of Animal Pathogenic Clostridia. *Clostridial Diseases of Animals.* John Wiley & Sons, Inc.

PRESCOTT, L. M., HARLEY, J. P. & KLEIN, D. A. 2002. *Microbiology*, McGraw-Hill.

PRINCEWILL, T. J. & OAKLEY, C. L. 1976. Deoxyribonucleases and hyaluronidases of *Clostridium septicum* and *Clostridium chauvoei*. III. Relationship between the two organisms. *Med Lab Sci,* 33**,** 10-118.

QUINN, P. J., MARKEY, B. K., LEONARD, F. C., HARTIGAN, P., FANNING, S. & FITZPATRICK, E. S. 2011. *Veterinary Microbiology and Microbial Disease*, Wiley.

RAHMAN, H., CHAKRABORTY, A., RAHMAN, T., SHARMA, R., SHOME, B. R. & SHAKUNTALA, I. 2009. Clostridial myonecrosis clinically resembling black quarter in an Indian elephant (Elephas maximus). *Rev Sci Tech,* 28**,** 1069-75.

RAJU, D., SETLOW, P. & SARKER, M. R. 2007. Antisense-RNA-mediated decreased synthesis of small, acid-soluble spore proteins leads to decreased resistance of *clostridium perfringens* spores to moist heat and UV radiation. *Appl Environ Microbiol,* 73**,** 2048-53.

RAMARAO, D. & RAO, B. U. 1990. Studies of the incidence of black quarter in Karnataka during 1979-85. *Indian Veterinary Journal,* 67**,** 795-801.

RATH, D., AMLINGER, L., RATH, A. & LUNDGREN, M. 2015. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie,* 117**,** 119-128.

RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics,* 13**,** 278-289.

RONAGHI, M., KARAMOHAMED, S., PETTERSSON, B., UHLEN, M. & NYREN, P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem,* 242**,** 84-9.

ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M., LEAMON, J. H., JOHNSON, K., MILGREW, M. J., EDWARDS, M., HOON, J., SIMONS, J. F., MARRAN, D., MYERS, J. W., DAVIDSON, J. F., BRANTING, A., NOBILE, J. R., PUC, B. P., LIGHT, D., CLARK, T. A., HUBER, M., BRANCIFORTE, J. T., STONER, I. B., CAWLEY, S. E., LYONS, M., FU, Y., HOMER, N., SEDOVA, M., MIAO, X., REED, B., SABINA, J., FEIERSTEIN, E., SCHORN, M., ALANJARY, M., DIMALANTA, E., DRESSMAN, D., KASINSKAS, R., SOKOLSKY, T., FIDANZA, J. A., NAMSARAEV, E., MCKERNAN, K. J., WILLIAMS, A., ROTH, G. T. & BUSTILLO, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475**,** 348-352.

ROULI, L., MERHEJ, V., FOURNIER, P. E. & RAOULT, D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections,* 7**,** 72-85.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M.-A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-945.

RYCHENER, L., INALBON, S., DJORDJEVIC, S. P., CHOWDHURY, P. R., ZIECH, R. E., DE VARGAS, A. C., FREY, J. & FALQUET, L. 2017. *Clostridium chauvoei*, an Evolutionary Dead-End Pathogen. *Frontiers in Microbiology,* 8.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4.

SALZBERG, S. L., PHILLIPPY, A. M., ZIMIN, A., PUIU, D., MAGOC, T. & KOREN, S. 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res,* 22.

SASAKI, Y., KOJIMA, A., AOKI, H., OGIKUBO, Y., TAKIKAWA, N. & TAMURA, Y. 2002a. Phylogenetic analysis and PCR detection of *Clostridium chauvoei, Clostridium haemolyticum*, *Clostridium novyi* types A and B, and *Clostridium septicum* based on the flagellin gene. *Veterinary Microbiology,* 86**,** 257-267.

SASAKI, Y., KOJIMA, A., KIKUCHI, E. & TAMURA, Y. 2002b. Multiplex PCR for direct detection of pathogenic clostridia in bovine clostridial infections. *Journal of Veterinary Medicine, Japan,* 55**,** 889-893.

SASAKI, Y., YAMAMOTO, K., AMIMOTO, K., KOJIMA, A., OGIKUBO, Y., NORIMATSU, M., OGATA, H. & TAMURA, Y. 2001a. Amplification of the 16S-

23S rDNA spacer region for rapid detection of *Clostridium chauvoei* and *Clostridium septicum*. *Research in Veterinary Science,* 71**,** 227-229.

SASAKI, Y., YAMAMOTO, K., KOJIMA, A., NORIMATSU, M. & TAMURA, Y. 2000a. Rapid identification and differentiation of pathogenic clostridia in gas gangrene by polymerase chain reaction based on the 16S-23S rDNA spacer region. *Research in Veterinary Science,* 69**,** 289-294.

SASAKI, Y., YAMAMOTO, K., KOJIMA, A., TETSUKA, Y., NORIMATSU, M. & TAMURA, Y. 2000b. Rapid and direct detection of *Clostridium chauvoei* by PCR of the 16S-23S rDNA spacer region and partial 23S rDNA sequences. *Journal of Veterinary Medical Science,* 62**,** 1275-1281.

SASAKI, Y., YAMAMOTO, K., TAMURA, Y. & TAKAHASHI, T. 2001b. Tetracycline-resistance genes of *Clostridium perfringens*, *Clostridium septicum* and *Clostridium sordellii* isolated from cattle affected with malignant edema. *Vet Microbiol,* 83**,** 61-9.

SEBAIHIA, M., PECK, M. W., MINTON, N. P., THOMSON, N. R., HOLDEN, M. T., MITCHELL, W. J., CARTER, A. T., BENTLEY, S. D., MASON, D. R., CROSSMAN, L., PAUL, C. J., IVENS, A., WELLS-BENNIK, M. H., DAVIS, I. J., CERDENO-TARRAGA, A. M., CHURCHER, C., QUAIL, M. A., CHILLINGWORTH, T., FELTWELL, T., FRASER, A., GOODHEAD, I., HANCE, Z., JAGELS, K., LARKE, N., MADDISON, M., MOULE, S., MUNGALL, K., NORBERTCZAK, H., RABBINOWITSCH, E., SANDERS, M., SIMMONDS, M., WHITE, B., WHITHEAD, S. & PARKHILL, J. 2007. Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes. *Genome Res,* 17**,** 1082-92.

SEBAIHIA, M., WREN, B. W., MULLANY, P., FAIRWEATHER, N. F., MINTON, N., STABLER, R., THOMSON, N. R., ROBERTS, A. P., CERDENO-TARRAGA, A. M. & WANG, H. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet,* 38.

SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.

SEEMANN, T. 2015. snippy: fast bacterial variant calling from NGS reads. *https://github.com/tseemann/snippy*.

SEGERMAN, B. 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Frontiers in Cellular and Infection Microbiology,* 2**,** 116.

SELZER, J., HOFMANN, F., REX, G., WILM, M., MANN, M., JUST, I. & AKTORIES, K. 1996. *Clostridium novyi* alpha-toxin-catalyzed incorporation of GlcNAc into Rho subfamily proteins. *Journal of Biological Chemistry,* 271**,** 25173-25177.

SETLOW, P. 1988. Small, acid-soluble spore proteins of *Bacillus* species: structure, synthesis, genetics, function, and degradation. *Annu Rev Microbiol,* 42**,** 319-38.

SETLOW, P. 1994. Mechanisms which contribute to the long-term survival of spores of Bacillus species. *Soc Appl Bacteriol Symp Ser,* 23**,** 49s-60s.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotech,* 26**,** 1135-1145.

SHIMIZU, T., OHTANI, K., HIRAKAWA, H., OHSHIMA, K., YAMASHITA, A., SHIBA, T., OGASAWARA, N., HATTORI, M., KUHARA, S. & HAYASHI, H. 2002. Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc Natl Acad Sci U S A,* 99**,** 996-1001.

SILVA, A. G. D. 2015. Pairwise_SNP_differences: an R script to summarise SNP differences among groups of samples. *github.com/MDU-PHL/pairwise_snp_differences.git*.

SILVA, R. O. S., UZAL, F. A., OLIVEIRA, C. A., LOBATO, F. C. F., UZAL, F. A., SONGER, J. G., PRESCOTT, J. F. & POPOFF, M. R. 2016. Gas Gangrene (Malignant Edema). *Clostridial Diseases of Animals.* John Wiley & Sons, Inc.

SKARIN, H., HÅFSTRÖM, T., WESTERBERG, J. & SEGERMAN, B. 2011. *Clostridium botulinum* group III: a group with dual identity shaped by plasmids, phages and mobile elements. *BMC Genomics,* 12**,** 1-13.

SKARIN, H. & SEGERMAN, B. 2011. Horizontal gene transfer of toxin genes in *Clostridium botulinum*: Involvement of mobile elements and plasmids. *Mobile Genetic Elements,* 1**,** 213-215.

SOMMER, D. D., DELCHER, A. L., SALZBERG, S. L. & POP, M. 2007. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics,* 8**,** 64.

SONGER, J. G. 1998. Clostridial diseases of small ruminants. *Vet Res,* 29**,** 219-32.

STACKEBRANDT, E., KRAMER, I., SWIDERSKI, J. & HIPPE, H. 1999. Phylogenetic basis for a taxonomic dissection of the genus *Clostridium*. *FEMS Immunol Med Microbiol,* 24**,** 253-8.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-1313.

STEVENS, D. L., ALDAPE, M. J. & BRYANT, A. E. 2012. Life-threatening clostridial infections. *Anaerobe,* 18**,** 254-9.

SWAIN, M. T., TSAI, I. J., ASSEFA, S. A., NEWBOLD, C., BERRIMAN, M. & OTTO, T. D. 2012. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc,* 7**,** 1260-84.

TALUKDAR, P. K., OLGUÍN-ARANEDA, V., ALNOMAN, M., PAREDES-SABJA, D. & SARKER, M. R. 2015. Updates on the sporulation process in *Clostridium* species. *Research in Microbiology,* 166**,** 225-235.

TAMURA, Y., KIJIMA-TANAKA, M., AOKI, A., OGIKUBO, Y. & TAKAHASHI, T. 1995. Reversible expression of motility and flagella in *Clostridium chauvoei* and their relationship to virulence. *Microbiology,* 141 (Pt 3)**,** 605-10.

TAMURA, Y., KIJIMA, M., HAMAMOTO, K. & YOSHIMURA, H. 1992. Partial characterization of the hemolysin produced by *Clostridium chauvoei*. *J Vet Med Sci,* 54**,** 777-8.

TAMURA, Y., MINAMOTO, N. & TANAKA, S. 1984. Demonstration of protective antigen carried by flagella of *Clostridium chauvoei*. *Microbiol Immunol,* 28**,** 1325-32.

TAMURA, Y. & TANAKA, S. 1984. Effect of antiflagellar serum in the protection of mice against *Clostridium chauvoei*. *Infect Immun,* 43**,** 612-6.

TANAKA, M., HIRAYAMA, N. & TAMURA, Y. 1987. Production, characterization, and protective effect of monoclonal antibodies to *Clostridium chauvoei* flagella. *Infect Immun,* 55**,** 1779-83.

TASTEYRE, A., KARJALAINEN, T., AVESANI, V., DELMEE, M., COLLIGNON, A., BOURLIOUX, P. & BARC, M. C. 2000. Phenotypic and genotypic diversity of the flagellin gene (fliC) among *Clostridium difficile* isolates from different serogroups. *J Clin Microbiol,* 38**,** 3179-86.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y

ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. & FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A,* 102**,** 13950-5.

TRAAG, B. A., PUGLIESE, A., EISEN, J. A. & LOSICK, R. 2013. Gene Conservation among Endospore-Forming Bacteria Reveals Additional Sporulation Genes in *Bacillus subtilis*. *Journal of Bacteriology,* 195**,** 253-260.

TREANGEN, T. J., ONDOV, B. D., KOREN, S. & PHILLIPPY, A. M. 2014. Rapid Core-Genome Alignment and Visualization for Thousands of Microbial Genomes. *bioRxiv*.

TSAI, I. J., OTTO, T. D. & BERRIMAN, M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology,* 11**,** 1-9.

TULLEY, W. 2010. An outbreak of blackleg (*Clostridium chauvoei*) on a beef rearing and finishing unit. *Summa, Animali da Reddito,* 5**,** 54-60.

USEH, N. M., IBRAHIM, N. D., NOK, A. J. & ESIEVO, K. A. 2006. Relationship between outbreaks of blackleg in cattle and annual rainfall in Zaria, Nigeria. *Vet Rec,* 158**,** 100-1.

USEH, N. M., NOK, A. J. & ESIEVO, K. A. 2003. Pathogenesis and pathology of blackleg in ruminants: the role of toxins and neuraminidase. A short review. *Vet Q,* 25**,** 155-9.

USHARANI, J., NAGALEEKAR, V. K., THOMAS, P., GUPTA, S. K., BHURE, S. K., DANDAPAT, P., AGARWAL, R. K. & SINGH, V. P. 2015. Development of a recombinant flagellin based ELISA for the detection of *Clostridium chauvoei*. *Anaerobe,* 33**,** 48-54.

VAN SCHAIK, W. 2015. The human gut resistome. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 370**,** 20140087.

VARANI, A. M., SIGUIER, P., GOURBEYRE, E., CHARNEAU, V. & CHANDLER, M. 2011. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology,* 12**,** R30-R30.

VILEI, E. M., JOHANSSON, A., SCHLATTER, Y., REDHEAD, K. & FREY, J. 2011. Genetic and functional characterization of the NanA sialidase from *Clostridium chauvoei*. *Veterinary Research,* 42.

VOS, M. & DIDELOT, X. 2008. A comparison of homologous recombination rates in bacteria and archaea. *ISME J,* 3**,** 199-208.

WEATHERHEAD, J. E. & TWEARDY, D. J. 2012. Lethal human neutropenic entercolitis caused by *Clostridium chauvoei* in the United States: tip of the iceberg? *J Infect,* 64**,** 225-7.

WELLER, C. & WU, M. 2015. A generation-time effect on the rate of molecular evolution in bacteria. *Evolution,* 69**,** 643-52.

WHELAN, S., LIÒ, P. & GOLDMAN, N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics,* 17**,** 262-272.

WOLF, R., HIESEL, J., KUCHLING, S., DEUTZ, A., KASTELIC, J., BARKEMA, H. W. & WAGNER, P. 2017. Spatial-temporal cluster analysis of fatal *Clostridium chauvoei*

cases among cattle in Styria, Austria between 1986 and 2013. *Prev Vet Med,* 138**,** 134-138.

WOUDSTRA, C., LE MARÉCHAL, C., SOUILLARD, R., BAYON-AUBOYER, M.-H., MERMOUD, I., DESOUTTER, D. & FACH, P. 2016. New Insights into the Genetic Diversity of *Clostridium botulinum* Group III through Extensive Genome Exploration. *Frontiers in Microbiology,* 7**,** 757.

XIAO, Y., VAN HIJUM, S. A. F. T., ABEE, T. & WELLS-BENNIK, M. H. J. 2015. Genome-Wide Transcriptional Profiling of *Clostridium perfringens* SM101 during Sporulation Extends the Core of Putative Sporulation Genes and Genes Determining Spore Properties and Germination Characteristics. *PLoS ONE,* 10**,** e0127036.

YANG, Z. & RANNALA, B. 2012. Molecular phylogenetics: principles and practice. *Nat Rev Genet,* 13**,** 303-314.

YUTIN, N. & GALPERIN, M. Y. 2013. A genomic update on clostridial phylogeny: Gram-negative spore-formers and other misplaced clostridia. *Environmental microbiology,* 15**,** 2631-2641.

ZHAO, Y., JIA, X., YANG, J., LING, Y., ZHANG, Z., YU, J., WU, J. & XIAO, J. 2014. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics,* 30**,** 1297-9.

ZHOU, C., MA, Q., MAO, X., LIU, B., YIN, Y. & XU, Y. 2014. New Insights into Clostridia Through Comparative Analyses of Their 40 Genomes. *BioEnergy Research,* 7**,** 1481-1492.

ZHOU, Y., LIANG, Y., LYNCH, K. H., DENNIS, J. J. & WISHART, D. S. 2011. PHAST: a fast phage search tool. *Nucleic Acids Res,* 39**,** W347-52.

ZIMIN, A. V., MARCAIS, G., PUIU, D., ROBERTS, M., SALZBERG, S. L. & YORKE, J. A. 2013. The MaSuRCA genome assembler. *Bioinformatics.*

ZONG, Y., MAZMANIAN, S. K., SCHNEEWIND, O. & NARAYANA, S. V. 2004. The structure of sortase B, a cysteine transpeptidase that tethers surface protein to the *Staphylococcus aureus* cell wall. *Structure,* 12**,** 105-12.

Acknowledgements

I would like to gratefully thank and acknowledge Prof. Dr. Lothar H. Wieler of Robert Koch-Institut, Berlin, Germany and feel highly privileged for getting a chance to carry out my doctoral study under his guidance. His support and guidance has given me the confidence to carry out my research at high speed and to carry out the proposed objectives.

I am highly thankful and sincerely indebted to Prof. Dr. Heinrich Neubauer for giving me an opportunity for carrying out my doctoral study under his guidance at the Institute of Bacterial Infections and Zoonoses (IBIZ), Friedrich-Loeffler-Institut, Jena, Germany. The guidance provided by Prof. Dr. Heinrich Neubauer is highly regarded and his expertise and valuable suggestions were critical in achieving the goal of the thesis. The facility and support offered by him during the entire period is highly regarded. His meticulousness suggestions and corrections have greatly improved my study, publications and thesis. I am especially indebted to the FLI president Prof. Dr. Dr. h.c. Thomas C. Mettenleiter who provided an institute position during the last months of the doctoral study.

I am highly grateful to our group leader Dr. Christian Seyboldt who has done the pivotal role in planning and execution of my thesis. He was highly generous and offered all his resources throughout my term with National Reference Laboratory (NRL) for Blackleg. His valuable suggestions and his expertise over the area of anaerobic disease investigations and epidemiology were highly useful for shaping the study. The efforts taken by him in correcting my thesis, articles and presentations were highly valuable. I would also like to acknowledge and thank the efforts taken by Prof. Dr. Heinrich Neubauer and Dr. Christian Seyboldt for getting highly valuable strains for my study.

Furthermore, I would like to thank Torsten Semmler from Robert Koch-Institut, Berlin for bioinformatics analysis of sequence data, Inga Eichhorn for carrying out the genome sequencing and Dr. Antina Lübke-Becker for giving all the research support at the Institute of Microbiology and Epizootics, Department of Veterinary Medicine, Freie Universität Berlin, Germany. I also sincerely acknowledge PD Dr. Christiane Werckenthin and other groups for providing valuable strains involved in my study.

I am highly thankful for all the help and support offered by Dr. Lisa Sprague and the effort taken by her for correcting my thesis. I am highly thankful for the help and encouragements given by PD Dr. Ulrich Methner, Dr. Gernot Schmoock, Dr. Michael Böhringer, Dr. Mandy Carolina Elschner, Dr. Falk Melzer, Dr. Anne Busch and Eric Zuchantke.

I also like to thank my doctoral colleague Mostafa Y. Abdel-Glil for all his useful help and support for my studies. I gratefully thank Dr. Dhanyalekshmi Pillai for her sincere efforts in making me confident with Linux based computational analysis.

I gratefully thank Indian Council of Agricultural Research (ICAR), Ministry of Agriculture, Government of India for funding my studies under ICAR International Fellowship 2013-14. I am highly thankful for choosing me as a candidate and sincerely acknowledge the efforts taken by the Assistant Director General (EQR), Assistant Director

General (HRD) and all the supporting staff of ICAR for helping me with the procedures. I wish to thank the Director, Indian Veterinary Research Institute (I.V.R.I) and all my colleagues and staff of I.V.R.I for providing this opportunity.

I sincerely thank our group members Frau Jutta Carmon, Frau Renate Danner, Frau Janin Rücknagel and Frau Sandra Henning for the support and care they have given me all the time. I also remember the good time and company given by Dr. Gamal Wareth, Dr. Tuan Nguyen, Dr. John Njeru, Dr. Emad Hashish, Dr. Sabrina Discher, Dr. Hosny Hasan El Adawy, Mauricio Alejandro Andino Molina and all my doctoral colleagues. I would also like to thank Marcel Trautmann and the information-technology group, the entire administration at FLI, Jena and Riems for their continuous efforts to support me.

I also remember the good time and friendship given by the Jena Malayalee's friend group and my friends from I.V.R.I and K.A.U.

Lastly I am highly indebted for the efforts taken by my wife Shalin, our son Edwin and our parents for all their support and care during all my ventures in life.

Publications

Der Inhalt der vorliegenden Arbeit wurde teilweise bereits veröffentlicht:

The content of this thesis has already been published partially:

Thomas P, Semmler T, Eichhorn I, Lübke-Becker A, Werckenthin C, Abdel-Glil MY, Wieler LH, Neubauer H, Seyboldt C. First report of two complete *Clostridium chauvoei* genome sequences and detailed in silico genome analysis. Infect Genet Evol. 2017 Jul 15;54:287-298. doi: 10.1016/j.meegid.2017.07.018.

Selbstständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbständig angefertigt habe. Ich versichere, dass ich ausschließlich die angegebenen Quellen und Hilfen in Anspruch genommen habe.

I hereby confirm that I have prepared the present work independently. I assure that I have only taken the sources and aids indicated.

Berlin, den 29.01.2018 Prasad Thomas

mbvberlin | mensch und buch verlag
49,90 Euro | ISBN: 978-3-86387-875-7