Chapter 3
Comparative Genomics and Pathogenicity Islands of *Corynebacterium diphtheriae*, *Corynebacterium ulcerans*, and *Corynebacterium pseudotuberculosis*

Eva Trost and Andreas Tauch

Abstract The systematic application of next-generation DNA sequencing technologies has provided detailed insights into the genomics of corynebacteria. The genomes of 13 *Corynebacterium diphtheriae* strains isolated from cases of classical diphtheria, endocarditis and pneumonia were completely sequenced and annotated, providing first insights into the pan-genome of this species. Comparative gene content analyses revealed an enormous collection of variable pilus gene clusters relevant for adhesion properties of *C. diphtheriae*. Variation in the distributed genome is apparently a common strategy of *C. diphtheriae* to establish differences in host-pathogen interactions. Molecular data deduced from the complete genome sequences of two *Corynebacterium ulcerans* strains provided considerable knowledge of candidate virulence factors, including a novel type of ribosome-binding protein with striking structural similarity to Shiga-like toxins. Likewise, functional data deduced from the complete genome sequences of six *Corynebacterium pseudotuberculosis* isolates from various sources greatly extended the knowledge of virulence factors and indicated that this species is equipped with a distinct gene set promoting its survival under unfavorable environmental conditions encountered in the mammalian host.

Keywords Genome sequence · Core genome · Pan-genome · Pathogenicity island · Virulence factor

A. Tauch (✉) · E. Trost
Institut für Genomforschung und Systembiologie, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 27, 33615, Bielefeld, Germany
e-mail: tauch@cebitec.uni-bielefeld.de

E. Trost
e-mail: etrost@cebitec.uni-bielefeld.de

A. Burkovski (ed.), *Corynebacterium diphtheriae and Related Toxigenic Species*,
DOI 10.1007/978-94-007-7624-1_3, © Springer Science+Business Media Dordrecht
(outside the USA) 2014

3.1 Introduction

A new generation of DNA sequencing approaches, collectively called next-generation DNA sequencing technologies, has provided unprecedented opportunities for high-throughput functional genome research (Mardis 2008; Shendure and Ji 2008). Since first introduced to the market in 2005, these technologies have been used for standard sequencing applications, such as whole-genome sequencing and genome resequencing, and for novel applications previously unexplored by the 'classical' Sanger sequencing strategy (Morozova and Marra 2008). Despite the many advances in chemistries and the robust performance of modern Sanger sequencers, the application of this relatively expensive method to large genome sequencing projects has remained beyond the means of the typical grant-funded investigator. An inherent limitation of Sanger sequencing is the requirement of *in vivo* amplification of DNA fragments that are to be sequenced, which is usually achieved by cloning into bacterial hosts (Morozova and Marra 2008; Shendure and Ji 2008). The Roche/454 technology, the first next-generation DNA sequencing technology released to the market, circumvents the cloning requirement by taking advantage of a highly efficient *in vitro* DNA amplification method known as emulsion PCR (Rothberg and Leamon 2008; Droege and Hill 2008). Moreover, the use of the picotiter plate system in the Roche/454 approach allows hundreds of thousands of pyrosequencing reactions to be carried out in parallel, massively increasing the sequencing throughput (Droege and Hill 2008).

Recent scientific discoveries in the field of corynebacterial genomics resulted from the systematic application of next-generation DNA sequencing technologies; i.e. the Roche/454 Genome Sequencer FLX System and the Life Technologies SOLiD System (Tauch et al. 2008a; Cerdeira et al. 2011a). Not surprisingly, the first next-generation sequencing studies have focused on the genomes of corynebacterial pathogens because of their importance in human disease, including *Corynebacterium urealyticum*, *Corynebacterium kroppenstedtii*, *Corynebacterium aurimucosum* and *Corynebacterium resistens* (Soriano and Tauch 2008; Tauch et al. 2008a, b; Trost et al. 2010a; Schröder et al. 2012). The Genomes OnLine Database GOLD (Pagani et al. 2012) lists additional corynebacterial species, whose genomes have been sequenced to high-quality draft status in the course of the human microbiome project (Lewis et al. 2012).

This chapter describes the current status of genome sequencing projects of the three closely related pathogenic species *Corynebacterium diphtheriae*, *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis*, the so-called 'diphtheria' group, and summarizes the major findings of comparative genomic studies, thereby focussing on the detection of pathogenicity islands and virulence factors. To date, complete genome sequences of 21 strains have been published and sequencing of more clinical isolates is currently ongoing (Pagani et al. 2012). General features of the completely sequenced genomes of *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis* are listed in Table 3.1. The deduced genomic data considerably improve our understanding of the architecture and evolution of corynebacterial genomes, species-specific traits and potential factors contributing to pathogenicity in humans and animals.

Table 3.1 Overview of completely sequenced corynebacterial isolates and general features of the genome sequences

Strain	Genome size (bp)	G+C content (%)	No. of genes	No. of tRNAs	No. of rRNA operons (16S-23S-5S)	No. of unique genes	Types of CRISPRs (number of repeats)	GenBank Accession No.	Reference
Corynebacterium diphtheriae									
NCTC 13129	2,488,635	53.5	2,368	54	5	124	I (7); II (26)	BX248353	Cerdeño-Tarrága et al. 2003
C7(β)^{tox}+	2,499,189	53.5	2,350	56	5	126	I (6)	CP003210	Trost et al. 2012
PW8	2,530,683	53.7	2,361	53	5	101	III (15)	CP003216	Trost et al. 2012
CDC-E8392	2,433,326	54.6	2,270	54	5	52	III (12)	CP003211	Trost et al. 2012
31A	2,535,346	53.6	2,402	51	5	104	I (28)	CP003206	Trost et al. 2012
241	2,426,551	53.4	2,260	50	5	6	I (15); II (4)	CP003207	Trost et al. 2012
VA01	2,395,441	53.4	2,196	50	5	27	I (7)	CP003217	Trost et al. 2012
HC01	2,427,149	53.4	2,260	53	5	7	I (15); II (4)	CP003212	Trost et al. 2012
HC02	2,468,612	53.7	2,244	53	5	69	I (5)	CP003213	Trost et al. 2012
HC03	2,478,364	53.5	2,268	50	5	35	III (42)	CP003214	Trost et al. 2012
HC04	2,484,332	53.5	2,280	50	5	13	III (15)	CP003215	Trost et al. 2012
INCA 402	2,449,071	53.6	2,235	50	5	44	III (17)	CP003208	Trost et al. 2012
BH8	2,485,519	53.6	2,375	53	5	85	I (1)	CP003209	Trost et al. 2012
Corynebacterium ulcerans									
809	2,502,095	53.3	2,182	52	4	90	IV (28); V (12); VI (67)	CP002790	Trost et al. 2011
BR-AD22	2,606,374	53.4	2,338	52	4	132	IV (38); V (10); VI (32)	CP002791	Trost et al. 2011
Corynebacterium pseudotuberculosis									
FRC41	2,337,914	52.2	2,110	49	4	49	IV (1)	CP002097	Trost et al. 2010b
I19	2,337,730	52.2	2,124	49	4	1	IV (1)	CP002251	Silva et al. 2011
1002	2,335,112	52.2	2,111	48	4	1	IV (1)	CP001809	Ruiz et al. 2011
C231	2,328,208	52.2	2,103	48	4	4	IV (1)	CP001829	Ruiz et al. 2011
PAT10	2,335,323	52.2	2,079	48	4	4	IV (1)	CP002924	Cerdeira et al. 2011b
CIP 52.97	2,320,595	52.1	2,057	47	4	86	–	CP003061	Cerdeira et al. 2011c

3.2 The Pan-Genome of *C. diphtheriae* and Deduced Pathogenicity Islands

3.2.1 The Reference Genome of C. diphtheriae NCTC 13129

The first genome of the 'diphtheria' group to be sequenced was that of *C. diphtheriae* NCTC 13129, which was initially isolated from a pharyngeal membrane of a patient with clinical diphtheria (Cerdeño-Tarrága et al. 2003). This toxigenic strain is a representative of the clone responsible for an outbreak of diphtheria in the states of the former Soviet Union in the 1990s (Dittmann et al. 2000). The whole-genome shotgun method with Sanger technology has been applied to determine the genome sequence of *C. diphtheriae* NCTC 13129 (Cerdeño-Tarrága et al. 2003). The complete genome sequence derived from two genomic shotgun libraries and terminal sequences from a large-insert bacterial artificial chromosome (BAC) library that was used for generating a scaffold. The genome of *C. diphtheriae* NCTC 13129 has a size of 2,488,635 bp with a G+C content of 53.5% and contains 2,320 predicted coding regions, of which 45 were annotated as pseudogenes (Cerdeño-Tarrága et al. 2003). Very recently, a comprehensive re-annotation of this genome sequence has been performed as a new approach to make the *C. diphtheriae* NCTC 13129 reference genome more descriptive and current with relevant features regarding the organism's lifestyle (Salzberg 2007; D'Afonseca et al. 2012). This *in silico* strategy is facilitated by the massive amounts of publicly available data linked to sequenced genomes of other species of the genus *Corynebacterium* (Pagani et al. 2012). With respect to structural genomics of *C. diphtheriae* NCTC 13129, 23 protein-coding regions were deleted and 71 new genes were added to the initial genome annotation (D'Afonseca et al. 2012). Nevertheless, all gene regions previously assigned as pseudogenes were validated and ten new pseudogenes were created. In relation to functional genomics, about 57% of the initial genome annotation was updated to become functionally more informative, as the product descriptions of 973 predicted proteins were updated. Among them, 370 gene products previously annotated as 'hypothetical proteins' now have more informative descriptions (D'Afonseca et al. 2012). The re-annotation resulted in the discovery of new genes in the *C. diphtheriae* NCTC 13129 genome sequence, correction of coding strands and the significant improvement of functional description of protein-coding regions, including classical virulence genes (D'Afonseca et al. 2012). The re-annotated archives of *C. diphtheriae* NCTC 13129 are available at: http://lgcm.icb.ufmg.br/pub/C_diphtheriae_reannotation.embl.

Genomic islands in the genome of *C. diphtheriae* NCTC 13129 were detected by examining local anomalies in the nucleotide composition of the DNA, such as G+C content, GC skew and/or dinucleotide frequency deviations that can be indicative of the recent acquisition of DNA regions by horizontal gene transfer (Cerdeño-Tarrága et al. 2003). The most prominent genomic island of *C. diphtheriae* NCTC 13129 comprises the complete genome of a *tox*$^+$ corynephage encoding diphtheria toxin. This prophage has a size of 36,566 bp with a G+C content of 52.2% and encodes

43 predicted proteins. Sequence similarities on the amino acid level were detected to proteins of phage BFK20 from *Brevibacterium flavum* (Bukovska et al. 2006). The diphtheria toxin gene *tox* is located at the right end of the prophage genome, adjacent to the attachment site and within a DNA region of low G+C content. This specific location of *tox* is indicative of a bacterial gene that was acquired from a previous host and is dispensable for the life cycle of the phage, but may affect the phenotype or fitness of the lysogenic bacterium (Brüssow et al. 2004). In addition to the *tox*$^{+}$ corynephage (PICD 1), twelve genomic regions (PICD 2–13) with local anomalies in the nucleotide composition were detected in *C. diphtheriae* NCTC 13129 (Table 3.2). Several genes potentially involved in pathogenicity of *C. diphtheriae* NCTC 13129 are located on the detected genomic islands. These putative pathogenicity islands encode, for instance, a siderophore biosynthesis and export system, a putative lantibiotic biosynthesis system, and three types of sortase-related adhesive pili (Table 3.2). It is therefore likely that *C. diphtheriae* NCTC 13129 has recently acquired by horizontal transfer specialized genes that may be involved in the pathogenic lifestyle by encoding variable pilus structures for the adherence of the bacterium to host cell surfaces (Cerdeño-Tarrága et al. 2003).

3.2.2 The Pan-Genome of the Species C. diphtheriae

Very recently, the knowledge of the gene content of *C. diphtheriae* isolates was considerably extended, as the genomes of twelve clinical strains initially recovered from cases of classical diphtheria, endocarditis, and pneumonia were completely sequenced and annotated (Trost et al. 2012). The selected collection of *C. diphtheriae* strains (Table 3.1) includes the prominent ancestor of many toxoid vaccine producers *C. diphtheriae* PW8 (Park and Williams 1896) and the laboratory strain *C. diphtheriae* C7(β)$^{tox+}$ (Freeman 1951; Barksdale and Pappenheimer 1954). Including the genome sequence of the reference strain *C. diphtheriae* NCTC 13129, a comparative analysis of these genomes allowed the first characterization of the pan-genome of the species *C. diphtheriae* (Trost et al. 2012). The microbial pan-genome is defined as the total gene repertoire in a bacterial species and comprises the ‘core genome’, which is shared by all individuals, the ‘dispensable genome’ containing genes present only in a subset of individuals, and the ‘unique genome’, which is unique to an individual (Medini et al. 2005; Tettelin et al. 2008).

The twelve *C. diphtheriae* genomes were sequenced by pyrosequencing with the Roche/454 Genome Sequencer FLX System and sequencing depths ranging from 29× to 55× (Trost et al. 2012). All genomic sequences were assembled to circular chromosomes with 2.395 Mb to 2.535 Mb in size (Table 3.1). The average G+C content of each genome is in the range of 53 %, which is consistent with the G+C content of the reference genome of *C. diphtheriae* NCTC 13129 (Cerdeño-Tarrága et al. 2003). The annotation of the twelve *C. diphtheriae* genomes and re-annotation of the *C. diphtheriae* NCTC 13129 genome sequence (D’Afonseca et al. 2012) revealed a median number of 2,294 protein-coding genes for each strain, with the lowest number of 2,196 genes annotated in the genome of *C. diphtheriae* VA01 and the highest number of 2,402 genes in *C. diphtheriae* 31A (Table 3.1). A comparative gene content analysis showed that the mean number of genes shared by two strains comprises 1,903 ± 54 orthologous genes, while the mean number of genes not shared by a distinct pair of strains comprises 644 ± 134 genes, indicating the large variability of the gene repertoire in the sequenced *C. diphtheriae* isolates (Trost et al. 2012).

The number of core genes of *C. diphtheriae* was determined with the software EDGAR using bidirectional best BLASTP hits for genome comparisons (Blom et al. 2009). Based on a series of calculations using all *C. diphtheriae* genomes individually as a reference, the core genome of the sequenced *C. diphtheriae* strains comprises 1,632 genes that can therefore be regarded as highly conserved in this species (Trost et al. 2012). To deduce the development of the core genome in dependence on the number of sequenced *C. diphtheriae* strains, the median number of core genes in each genome was calculated based on the permutation of all possible genome comparisons. According to this approach, the number of core genes present in *C. diphtheriae* will comprise about 1,611 protein-coding genes when adding further genome sequences to the current data set. This value revealed a genetic backbone of the *C. diphtheriae* genome, which includes approximately 70 % of the gene repertoire of the sequenced strains, with about 30 % of the gene content being variable to some extent and therefore belonging to the dispensable portion of the *C. diphtheriae* genome. The full complement of protein-coding regions that are part of the dispensable genome of *C. diphtheriae* was determined as 2,361 distributed genes (Trost et al. 2012).

The bioinformatic characterization of the unique genome (Table 3.1) revealed the average number of 61 ± 43 strain-specific genes per sequenced *C. diphtheriae* isolate (Trost et al. 2012). To deduce the development of the number of unique genes in dependence on the number of sequenced *C. diphtheriae* genomes, the median number of strain-specific coding regions was determined using the permutation of all possible genome comparisons. The respective calculation indicated that the median number of unique genes estimated to occur in additionally sequenced *C. diphtheriae* genomes comprises about 65 genes. Accordingly, the sum total of protein-coding regions representing the pan-genome of *C. diphtheriae* currently comprises 4,786 genes, which is about three times the size of the deduced core genome (Trost et al. 2012). This calculation was corroborated by applying Heaps' law: $n = \kappa \times N^{\gamma}$, with N being the number of sequenced genomes (Tettelin et al. 2008). Hence, the number of protein-coding regions added to the pan-genome of *C. diphtheriae* will increase by 69 genes per newly sequenced genome, indicating an open pan-genome for the species *C. diphtheriae* (Trost et al. 2012). In general, a microbial pan-genome can be classified as 'closed' or 'open' (Tettelin et al. 2008). A pan-genome is considered to be closed, if the number of new genes added per newly sequenced genome converges to zero. Therefore, a closed microbial pan-genome indicates a static gene content of a bacterial species that is no longer expendable by genome sequencing. On the other hand, a pan-genome is considered open when each newly sequenced strain can be expected to reveal some genes unique within the species, regardless of the number of already analyzed genomes. An open pan-genome is therefore associated with a dynamic gene content of a bacterial species (Halachev et al. 2011).

3.2.3 Genetic Variability of CRISPR/cas Regions in **C. diphtheriae**

Due to the genetic diversity of *C. diphtheriae* isolates, a number of typing methods have been established for inter-strain differentiation, such as amplified fragment length polymorphism analysis, multilocus enzyme electrophoresis, pulsed-field gel electrophoresis, ribotyping, randomly amplified polymorphic DNA analysis (Mokrousov 2009), and multi-locus sequence typing (Jolley et al. 2004). These methods allow the identification of clonal groups of closely related *C. diphtheriae* strains with different sensitivities. A newer method to determine the phylogenetic relationship of *C. diphtheriae* strains is the so-called spoligotyping (spacer oligonucleotide typing), which is based on the presence of arrays of clustered regularly interspaced short palindromic repeats (CRISPRs) in the genome sequence (Mokrousov et al. 2005). These arrays are composed of direct repeats that are separated by non-repetitive, similar-sized spacer sequences (Deveau et al. 2010). CRISPRs and associated *cas* genes represent a widespread genetic system across bacteria that causes RNA interference against foreign nucleic acids, for instance resistance to bacteriophages (Deveau et al. 2010; Marraffini and Sontheimer 2010). The CRISPR/*cas* system of *C. diphtheriae* therefore participates in a constant evolutionary battle between the bacterium and corynephages through the addition or deletion of spacer sequences in the bacterial genome and mutations or deletion in phage genomes. Targets for spoligotyping are the spacer regions between the direct repeats, as variations in the number or nucleotide sequence of spacers provide patterns for the differentiation between clonal groups of *C. diphtheriae* isolates (Mokrousov et al. 2005). In a macroarray-based approach of spoligotyping, 154 clinical *C. diphtheriae* strains were subdivided into 34 spoligotypes (Mokrousov et al. 2005).

Three types of CRISPR/*cas* systems were detected in the genomes of the sequenced *C. diphtheriae* strains (Fig. 3.1). A detailed classification of the CRISPR/*cas* regions is listed in Table 3.1. CRISPR/*cas* type I was detected in the genomes of eight strains and is composed of three *cas* genes (*cas1* to *cas3*). The number of associated spacer sequences ranges from one to 28. CRISPR/*cas* type II is additionally present in three *C. diphtheriae* genomes and contains eight *cas* genes (*cas4* to *cas11*). The number of repeats in these arrays ranged from four to 26. CRISPR/*cas* type III is present in five genomes, with varying numbers of repeats ranging from 12 to 42. The type III CRISPR/*cas* region is flanked by eight *cas* genes (*cas12* to *cas19*). A nucleotide sequence comparison of the identified spacer sequences revealed that only 48 out of the 219 spacers are shared by two or three *C. diphtheriae* strains, supporting the view that CRISPR/*cas* regions provide an attractive target for the solid discrimination between different *C. diphtheriae* isolates (Mokrousov et al. 2005; Trost et al. 2012).

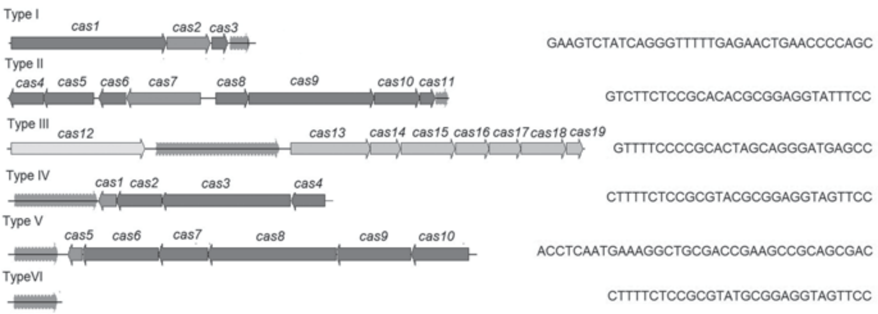
Type I
cas1
cas2
cas3
GAAGTCTATCAGGGTTTTTGAGAACTGAACCCCAGC
Type II
cas4
cas5
cas6
cas7
cas8
cas9
cas10
cas11
GTCTTCTCCGCACACGCGGAGGTATTTCC
Type III
cas12
cas13
cas14
cas15
cas16
cas17
cas18
cas19
GTTTTCCCCGCACTAGCAGGGATGAGCC
Type IV
cas1
cas2
cas3
cas4
CTTTTCTCCGCGTACGCGGAGGTAGTTCC
Type V
cas5
cas6
cas7
cas8
cas9
cas10
ACCTCAATGAAAGGCTGCGACCGAAGCCGCAGCGAC
TypeVI
CTTTTCTCCGCGTATGCGGAGGTAGTTCC

Fig. 3.1 Schematic representation of CRISPR/*cas* regions detected in the genomes of *C. diphtheriae*, *C. ulcerans*, and *C. pseudotuberculosis*. The CRISPR/*cas* regions show different numbers and arrangements of *cas* genes (*labeled arrows*). The position of the CRISPR locus is also variable. The nucleotide sequences of the conserved repeats are shown. CRISPR types I–III were detected in the sequenced *C. diphtheriae* genomes. CRISPR types IV–VI were found in the genomes of *C. ulcerans* isolates. CRISPR type VI of *C. ulcerans* is lacking associated *cas* genes. *C. pseudotuberculosis* isolates contain only CRISPR type IV. See also Table 3.1 for a detailed classification of CRISPR/*cas* regions

3.2.4 Genetic Variability of tox^+ Corynephages in C. diphtheriae

The pan-genome project also provided more detailed information about the genetic variability of corynephages harboring the diphtheria toxin gene *tox* that was identified in *C. diphtheriae* NCTC 13129 (Cerdeño-Tarrága et al. 2003) and in *C. diphtheriae* strains C7(β)$^{tox+}$, CDC-E8392, PW8 and 31A (Trost et al. 2012). In the case of *C. diphtheriae* PW8, two non-tandem copies of the corynephage ω^{tox+} were detected in the complete genome sequence, as suggested previously from restriction endonuclease maps of phage DNA (Rappuoli et al. 1983). The 36-kb genome sequences of both ω^{tox+} corynephages are almost identical, as they show only five nucleotide mismatches. Both copies of the prophage are separated by a 2-kb genomic region encoding a putative membrane protein that is flanked by two copies of a tRNA^{Arg} gene representing the known attachment site of corynephages in *C. diphtheriae* (Ratti et al. 1997). Nucleotide sequence comparisons of the tox^+ corynephages revealed that the ω^{tox+} phage of *C. diphtheriae* PW8 is similar to the β^{tox+} phage present in *C. diphtheriae* C7(β)$^{tox+}$, which is consistent with an early report demonstrating that both corynephages differ in only three genomic regions (Rappuoli et al. 1983). A highly different tox^+ prophage was detected in the genome sequence of *C. diphtheriae* 31A (Trost et al. 2012). Significant nucleotide sequence similarity to β-like corynephages was observed only at the right-hand end of the prophage genome, which harbors the *tox* gene region. Other regions of the prophage genome revealed homology at the amino acid level to proteins of the prophage ΦCULC22IV, which is present in the tox^- strain *C. ulcerans* BR-AD22 (Trost et al. 2011). It has been proposed previously that the diphtheria toxin gene *tox* was acquired by corynephage β due to the terminal location of this gene in the prophage and the significantly decreased G + C content of this region in the phage genome (Cerdeño-Tarrága et al. 2003; Brüssow et al. 2004). The detection of identical *tox* genes in different prophages now indicates that the acquisition of the diphtheria toxin gene *tox* occurred independently in two different corynephages or that gene shuffling is frequently found in this group of phages (Trost et al. 2012).

3.2.5 Pathogenicity Islands and Pilus Gene Clusters of C. diphtheriae

The plasticity of the *C. diphtheriae* genome was analyzed by two comparative approaches designed to detect differences in the repertoire of pathogenicity islands that were initially assigned in the reference genome of *C. diphtheriae* NCTC 13129 (Cerdeño-Tarrága et al. 2003). The distribution of pathogenicity islands PICD 3 and PICD 8 was investigated by a PCR-based approach in eleven *C. diphtheriae* strains (Soares et al. 2011). The pathogenicity island PICD 8 was detected in only one strain, *C. diphtheriae* HC01 (Table 3.1), whereas PICD 3 was more widely distributed and present in six *C. diphtheriae* strains. This data indicated that the pathogenicity islands of *C. diphtheriae* strains can be differentiated by their variable genomic stability, thereby contributing to genome evolution and the lifestyle of this bacterium (Soares et al. 2011). Another study analyzed the global genome organization of *C. diphtheriae* C7(−) and *C. diphtheriae* PW8 by comparative genomic hybridization including probes representing the 13 pathogenicity islands of *C. diphtheriae* NCTC 13129 (Iwaki et al. 2010). Remarkably, eleven of the 13 pathogenicity islands were considered to be absent in the genome of *C. diphtheriae* C7(−), although this strain retained clear signs of pathogenicity, including adhesion to Detroit 562 cells and the formation of abscesses in animal skin. In contrast, the genome of *C. diphtheriae* PW8 was considered to lack only three pathogenicity islands, but exhibited more reduced signs of pathogenicity in a model system. These results already suggested a great genomic heterogeneity of the species *C. diphtheriae*, not only in genome organization, but also in pathogenicity (Iwaki et al. 2010).

This data were recently confirmed on a much broader scale during the pan-genome project of *C. diphtheriae* (Trost et al. 2012). Genomic islands and candidate pathogenicity islands of the sequenced *C. diphtheriae* strains were identified with the new pathogenicity island prediction software PIPS, which performs a combined analysis of DNA sequences based on typical features of genomic islands (Soares et al. 2012). In total, 57 genomic islands were identified in the sequenced *C. diphtheriae* genomes, including 24 islands (PICD 1–24) in the reference genome of *C. diphtheriae* NCTC 13129 (Table 3.2). Additionally detected genomic islands in *C. diphtheriae* NCTC 13129 carry, for instance, the *irp6ABC* operon encoding a siderophore-dependent iron uptake system (PICD 15) (Qian et al. 2002) and the siderophore biosynthesis and transport gene cluster *ciuABCDEFG* (PICD 19) (Kunkle

Table 3.2 Overview of predicted pathogenicity islands of *C. diphtheriae* NCTC 13129

Reference	Name	Begin CDS	End CDS	Begin position	End position	Length (bp)	Prominent function of island gene(s)
NCTC 13129	PICD 1	DIP0179	DIP0222	152426	190661	38236	Diphtheria toxin encoding corynephage
NCTC 13129	PICD 2	DIP0223	DIP0247	190816	210235	19420	Adhesive pilus
NCTC 13129	PICD 3	DIP0282	DIP0290	249276	256671	7396	Iron transport system
NCTC 13129	PICD 4	DIP0334	DIP0359	305695	326445	20751	Secreted proteins, including polysaccharide degradation enzyme
NCTC 13129	PICD 5	DIP0438	DIP0445	400793	409559	8767	Metal transport system and secreted proteins
NCTC 13129	PICD 6	DIP0750	DIP0766	726536	742372	15837	Lantibiotic biosynthesis proteins
NCTC 13129	PICD 7	DIP0794	DIP0823	776261	796859	20599	Phage-related proteins
NCTC 13129	PICD 8	DIP1645	DIP1664	1680222	1700446	20225	Secreted proteins, including extracellular matrix-binding protein
NCTC 13129	PICD 9	DIP1817	DIP1843	1866720	1883310	16591	Phage-related proteins
NCTC 13129	PICD 10	DIP2010	DIP2015	2065290	2070753	5464	Adhesive pilus
NCTC 13129	PICD 11	DIP2064	DIP2093	2114764	2144615	29852	Fimbrial-associated protein and surface-anchored protein
NCTC 13129	PICD 12	DIP2143	DIP2170	2210329	2244077	33749	Siderophore biosynthesis and transport proteins
NCTC 13129	PICD 13	DIP2208	DIP2234	2297805	2322967	25163	CRISPR locus
NCTC 13129	PICD 14	DIP0028	DIP0051	27025	43284	16260	CRISPR locus
NCTC 13129	PICD 15	DIP0071	DIP0115	63353	94809	31457	Iron transport system
NCTC 13129	PICD 16	DIP0267	DIP0275	230464	239860	9397	Antibiotic resistance protein
NCTC 13129	PICD 17	DIP0320	DIP0326	285652	291817	6166	Transport system with unknown function
NCTC 13129	PICD 18	DIP0448	DIP0466	418336	436609	18274	Two-component system and transport system with unknown function
NCTC 13129	PICD 19	DIP0582	DIP0607	550390	573656	23267	Siderophore biosynthese and transport proteins
NCTC 13129	PICD 20	DIP1891	DIP1901	1941600	1956347	14748	Transport system with unknown function
NCTC 13129	PICD 21	DIP1944	DIP1971	1998093	2018900	20808	Diverse functions and proteins with unknown function
NCTC 13129	PICD 22	DIP2021	DIP2049	2076036	2096446	20411	Secreted proteins, including secretory lipases
NCTC 13129	PICD 23	DIP2123	DIP2135	2176625	2198998	22374	Transport system with unknown function
NCTC 13129	PICD 24	DIP2302	DIP2345	2398584	2451928	53345	Two-component system and transport system with unknown function

and Schmitt 2005). Therefore, the extended search for genomic islands in the sequenced *C. diphtheriae* strains revealed additional gene clusters with characteristics of horizontal gene transfer, which are probably involved in iron acquisition. Comparative *in silico* analysis of the predicted genomic islands revealed that some are strain-specific, whereas others are partially or completely conserved in more than one strain (Trost et al. 2012). Only eight genomic islands can be regarded as highly conserved in all *C. diphtheriae* genomes, demonstrating the great genomic plasticity of *C. diphtheriae* (Trost et al. 2012). Many genomic islands encode typical phage products and the respective genomic regions of the *C. diphtheriae* genomes can be regarded as remnants of prophages. Some genomic islands encode proteins involved in specific metabolic pathways and were assigned as metabolic islands of the *C. diphtheriae* genome, whereas others encode proteins involved in antibiotic resistance or heavy metal ion resistance, such as cadmium, copper, mercury, and arsenic resistance.

The plasticity of the *C. diphtheriae* genome is also obvious when visualizing the gene content of the sequenced isolates with the BRIG software (Alikhan et al. 2011) and using the genome of *C. diphtheriae* NCTC 13129 as a reference (Fig. 3.2). Variations in the gene repertoire of the *C. diphtheriae* isolates seem to cluster in genomic regions assigned as pathogenicity islands, indicating that horizontal gene transfer is a major force in shaping the gene content and physiological traits of *C. diphtheriae* strains.

The search for pathogenicity islands in the genomes of the sequenced *C. diphtheriae* strains led to the detection of several islands harboring gene clusters for adhesive pili (Trost et al. 2012), which play important roles in bacterial colonization and pathogenesis (Ton-That and Schneewind 2003). Pilus assembly has been studied extensively in *C. diphtheriae* and occurs by a two-step mechanism, whereby pilin subunits are first polymerized and then covalently anchored to cell wall peptidoglycan. A pilin-specific sortase catalyzes the polymerization of the pilus, consisting of the shaft protein, the tip pilin and the base pilin (Rogers et al. 2011). Based on amino acid sequence homology searches using the pilin motif and cell wall sorting signal as queries, at least two pilus gene clusters were identified in each of the sequenced *C. diphtheriae* isolates, with *C. diphtheriae* HC04 haboring four pilus gene clusters (Fig. 3.3). Six different types of pilus gene clusters were detected according to the arrangement of genes encoding pilus subunits (*spa*) or pilin-specific sortases (*srt*). It is noteworthy to mention that the genome of *C. diphtheriae* PW8 contains a highly degenerated SpaD gene cluster with multiple intact and disrupted genes encoding SpaD, SpaE, SpaF pilins and sortases SrtB and SrtE, in addition to a SpaA gene cluster with a disrupted *spaC* gene (Fig. 3.3). Mobile DNA elements were also detected in the SpaD locus of *C. diphtheriae* PW8, suggesting horizontal gene transfer for gene duplication. Phylogenetic trees reconstructed with the neighbor-joining algorithm revealed that the protein components of the pilus, i.e. shaft protein, tip pilin and base pilin, and the cognate pilin-specific sortases display a great diversity in their amino acid sequences (Trost et al. 2012). Therefore, most *spa* and *srt* genes present on the predicted pathogenicity islands of the sequenced *C. diphtheriae* strains were

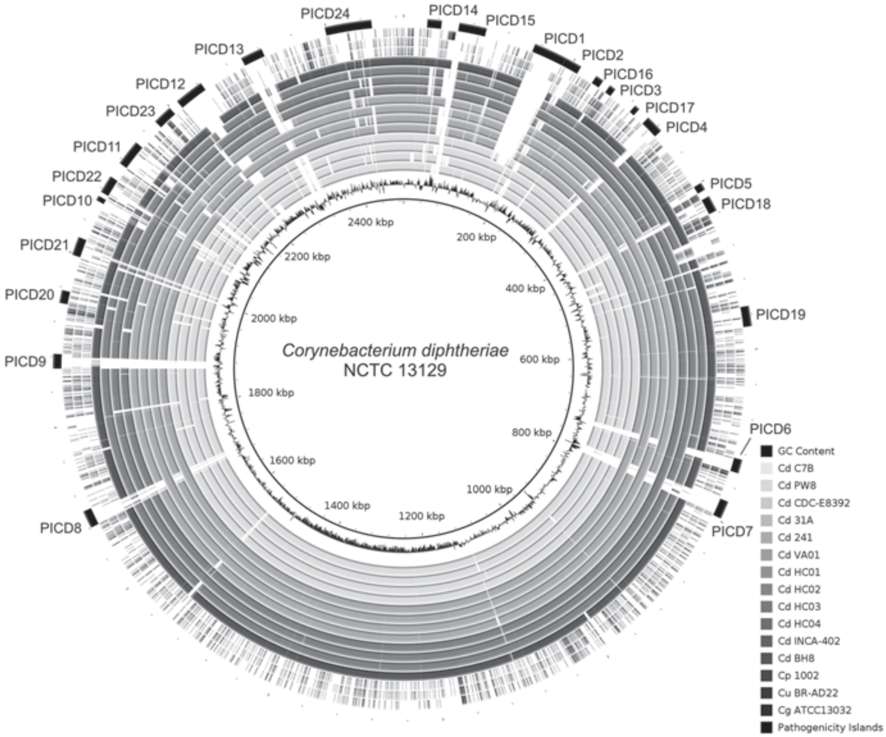

Fig. 3.2 Circular genome comparison between *C. diphtheriae* strains using *C. diphtheriae* NCTC 13129 as a reference. The genome comparison was generated with the BLAST Ring Image Generator BRIG. It shows the positions of candidate pathogenicity islands in the genome of *C. diphtheriae* NCTC 13129 and the presence/absence of these islands in other *C. diphtheriae* strains or species of the genus *Corynebacterium*. Abbreviations: GC Content, G+C profile of a genome region; Cd, *C. diphtheriae*; Cp, *C. pseudotuberculosis*; Cu, *C. ulcerans*; Cg, *C. glutamicum*; PICD, putative pathogenicity island of *C. diphtheriae*

assigned as unique genes during the pan-genome analysis. This result strongly implies that important variations exist on the cell surface of *C. diphtheriae* strains, which might be relevant for the initial step of an infection (Trost et al. 2012). Previous studies demonstrated different degrees of attachment of *C. diphtheriae* to HEp-2 cell monolayers (Hirata et al. 2004), differences in adhesion of *C. diphtheriae* C7(−) and *C. diphtheriae* PW8 to Detroit 562 cells (Iwaki et al. 2010) and strain-specific differences of *C. diphtheriae* in adhesion, invasion and intracellular survival (Ott et al. 2010). Moreover, mutations in the base pilin SpaB and the tip pilin SpaC of the SpaA-type pilus reduced the adhesive activity of *C. diphtheriae* (Mandlik et al. 2007).

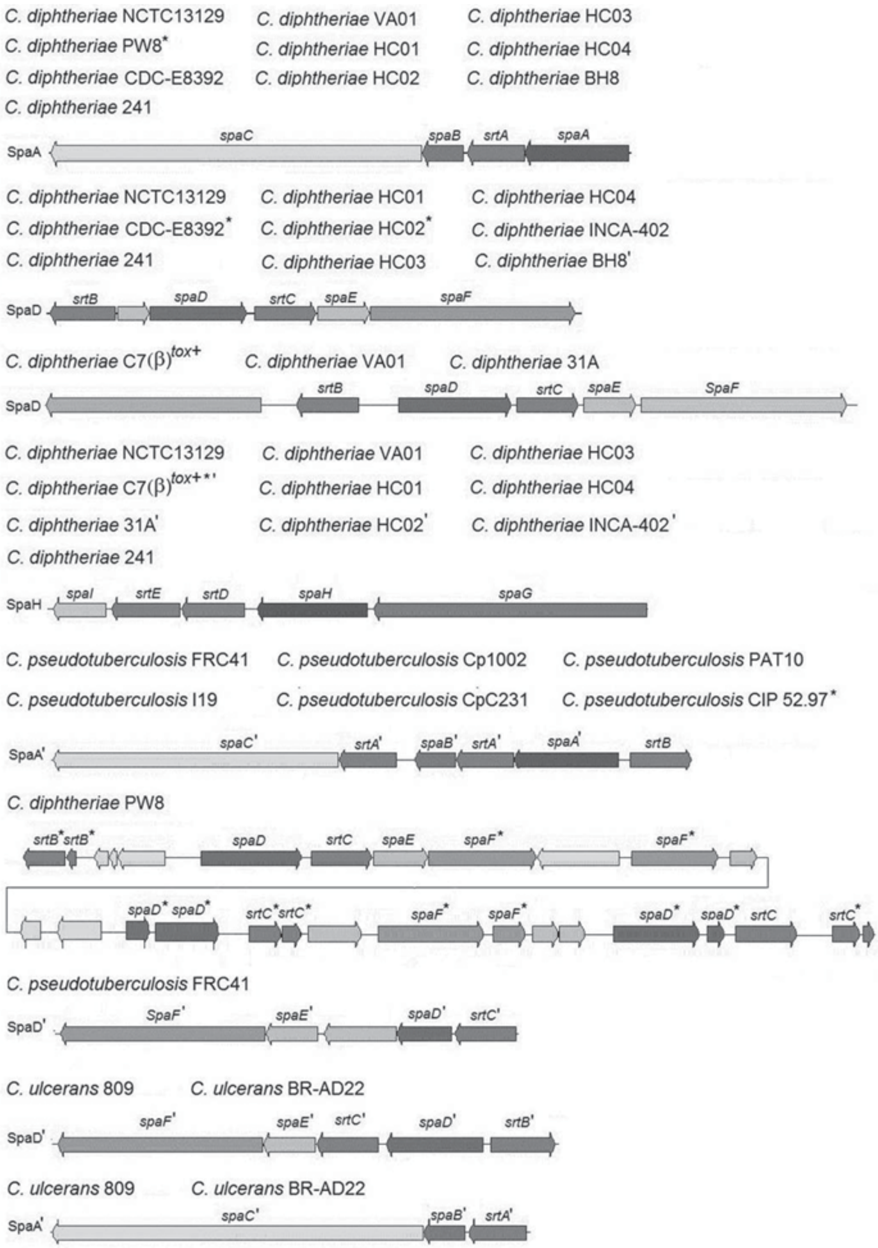

Fig. 3.3 Schematic representation of pilus gene clusters in the genomes of *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis*. The detected pilus gene clusters revealed different arrangements of genes encoding subunits of adhesive pili (*spa*) or pilin-specific sortases (*srt*). Assigned strains are listed above the gene clusters. Each *C. diphtheriae* genome contains at least two pilus gene clusters. *C. diphtheriae* PW8 contains a degenerated gene cluster with multiple intact and disrupted genes. Symbols: *asterisk* (*), clusters of the labeled strains contain a fragmented gene; *prime* (‘), denotes variants of the respective pilus gene cluster

3.3 Comparative Genomics of *C. ulcerans* and Candidate Virulence Factors

3.3.1 Reference Genomes of C. ulcerans *from Human and Animal Sources*

C. ulcerans has been detected as a commensal in domestic and wild animals that may serve as reservoirs for zoonotic infections (Hogg et al. 2009). As the knowledge of the bacterium's lifestyle and additional virulence factors besides the diphtheria toxin was very limited, the complete genome sequences of two *C. ulcerans* strains from human and animal sources were recently determined and characterized by comparative genomics (Trost et al. 2011). *C. ulcerans* 809 was isolated from an elderly woman with rapidly fatal pulmonary infection and a history of chronic bilateral limb ulcers. The woman lived in the metropolitan area of Rio de Janeiro and was hospitalized in coma, with shock and acute respiratory failure. The patient died 23 days after hospitalization despite an intensive medical treatment (Mattos-Guaraldi et al. 2008). *C. ulcerans* BR-AD22 was recovered from a nasal sample of a young asymptomatic dog that was kept in an animal shelter in Rio de Janeiro (Dias et al. 2010). The complete genome sequences of *C. ulcerans* 809 and *C. ulcerans* BR-AD22 were determined by pyrosequencing with the Roche/454 Genome Sequencer FLX System. This approach revealed sequence coverages of 42.8× and 22.9×, respectively (Trost et al. 2011). The chromosome of *C. ulcerans* 809 has a size of 2,502,095 bp and encodes 2,182 proteins, whereas the genome of *C. ulcerans* BR-AD22 is 104,279 bp larger and comprises 2,338 protein-coding regions (Table 3.1). The difference in size of the genomes is mainly caused by prophage-like elements that are present only in the genome of *C. ulcerans* BR-AD22. Both genomes show a highly similar order of orthologous genes and share a common set of 2,076 protein-coding regions, which demonstrates the very close phylogenetic relationship of both isolates. The pan-genome of the species *C. ulcerans* currently comprises 2,445 protein-coding genes. Obviously, more genome sequences of *C. ulcerans* isolates are necessary to determine the development of core genes, unique genes and the pan-genome of this species precisely.

*3.3.2 Genetic Variability of CRISPR/*cas *Regions and Prophages in* C. ulcerans

A screening of the genome sequences of *C. ulcerans* 809 and *C. ulcerans* BR-AD22 with the CRISPRFinder revealed the presence of three CRISPR/*cas* regions, herein named CRISPR types IV–VI (Fig. 3.1). CRISPR type IV is present in both *C. ulcerans* genomes and flanked by four *cas* genes. The direct repeats of this locus are 29 bp in length and separated by spacers with variable nucleotide sequences that are completely different in both *C. ulcerans* strains. Similar features were observed for the second CRISPR/*cas* region in the *C. ulcerans* genomes (Fig. 3.1). CRISPR type V is characterized by six *cas* genes and repeats of 36 bp. The spacer sequences present in *C. ulcerans* 809 are also different to those present in the corresponding CRISPR/*cas* region of the *C. ulcerans* BR-AD22 genome. CRISPR type VI of *C. ulcerans* is lacking associated *cas* genes in the direct proximity (Fig. 3.1). The spacer sequences of this CRISPR type have a length of 29 bp and show the largest variation between both strains, with 67 spacers present in the genome of *C. ulcerans* 809 and 32 spacers in *C. ulcerans* BR-AD22 (Table 3.1). The detection of CRISPR/*cas* regions in the genome of *C. ulcerans* strains and the sequence variations of the CRISPR loci suggests the use of these markers for a precise typing of clonal groups of *C. ulcerans* isolates from human and animal sources (Trost et al. 2011).

In accordance with the *tox*$^{-}$ phenotype of *C. ulcerans* 809 and *C. ulcerans* BR-AD22, both genomes were devoid of nucleotide sequences of a *tox*$^{+}$ corynephage encoding diphtheria toxin. However, the genomes of *C. ulcerans* 809 and *C. ulcerans* BR-AD22 harbor the highly similar prophages ΦCULC809I and ΦCULC22I with sizes of about 42 kb (Trost et al. 2011). Both prophages were detected at the same genomic position and apparently integrated at slightly different sites into a coding region for a hypothetical protein that might represent the integration site of these phages in the *C. ulcerans* chromosome. Minor differences between the prophages were detected in the number of genes, as ΦCULC809I comprises 45 genes, whereas 42 genes were assigned to the ΦCULC22I genome. According to global amino acid sequence alignments, both prophages share 36 genes that code for gene products with at least 98 % amino acid sequence identity, indicating the very close relationship of both prophages from different *C. ulcerans* isolates (Trost et al. 2011). The genome sequence of *C. ulcerans* BR-AD22 contains the additional prophages ΦCULC22II, ΦCULC22III and ΦCULC22IV, of which ΦCULC22III is incomplete and probably a defective remnant of a formerly active corynephage (Trost et al. 2011). Most strain-specific genes of the animal isolate *C. ulcerans* BR-AD22 were assigned to the additional prophage-like regions ΦCULC22II, ΦCULC22III and ΦCULC22IV. Therefore, only 92 protein-coding regions of this strain were regarded as unique genes, of which 13 genes were annotated with putative physiological functions (Trost et al. 2011).

3.3.3 Pathogenicity Islands and Virulence Factors of **C. ulcerans**

The search for unique genes by reciprocal best BLASTP matches revealed 90 strain-specific genes for the human isolate *C. ulcerans* 809, of which ten were annotated with putative physiological functions (Trost et al. 2011). This set of gene regions includes the *vsp2* gene coding for a secreted serine protease and the *rbp* gene encoding a putative ribosome-binding protein. Both gene products represent candidate virulence factors of *C. ulcerans* 809. The *rbp* gene is located between a gene coding for a putative phage integrase and a transposase gene and is moreover specified by the low G+C content of 45.1 %, suggesting the lateral transfer of *rbp* to *C. ulcerans* 809. The respective tyrosine recombinase detected in *C. ulcerans* 809 shares 92 % amino acid sequence identity with the integrase of the β-type corynephage present in the reference genome of *C. diphtheriae* NCTC 13129 and is also encoded directly adjacent to a tRNA^{Arg} gene. This gene annotation supports the assumption that a lysogenic β-type corynephage had been integrated downstream of the *rbp* gene in the *C. ulcerans* genome in former times (Trost et al. 2011). The protein product of the *rbp* gene showed weak similarity to the A chains of the Shiga-like toxins SLT-1 and SLT-2 from *Escherichia coli*, but contains all highly conserved amino acid residues relevant for the catalytic N-glycosidase activity (O'Loughlin et al. 2001; LaPointe et al. 2005). In contrast, the amino acid sequence of the Rbp protein lacks the typical ER-targeting sequence at the C-terminal end, which is necessary for the retranslocation of the catalytic domain of SLT-1 from the endoplasmatic reticulum (ER) into the cytosol of the host cell (O'Loughlin et al. 2001). As *C. ulcerans* can probably persist as a facultative intracellular pathogen in mammalian host cells, a retranslocation of the Rbp protein into the cytosol is nonessential for activity. The secretion of the putative toxin into the cytosol of host cells is supported instead by a typical signal sequence at the amino-terminal end of the protein (Trost et al. 2011). The enzymatic activity of the ribosome-binding protein Rbp probably leads to inhibition of protein biosynthesis by depurination of a single adenosine residue in the 28S rRNA of the eukaryotic ribosome (O'Loughlin et al. 2001). A genome screening for further virulence factors revealed the presence of endoglycosidase E (see below), neuraminidase H and adhesive pili of the SpaA' and SpaD' type that are encoded in both *C. ulcerans* genomes (Fig. 3.3). The *C. ulcerans* genome is apparently equipped with genes for a broad spectrum of virulence factors, including a novel ribosome-binding protein that is encoded only in the human isolate *C. ulcerans* 809.

Putative pathogenicity islands of *C. ulcerans* were detected the in larger genome of the animal isolate *C. ulcerans* BR-AD22 (Table 3.3). The *C. ulcerans* BR-AD22 genome contains 14 putative pathogenicity islands, including a phospholipase D gene region, an operon encoding urease and genes for iron uptake systems, which are generally associated with virulence. Most genes assigned to the pathogenicity islands of *C. ulcerans* have diverse roles in cellular metabolism or even hitherto unknown functions. However, all candidate virulence factors in the detected pathogenicity islands have characteristics that are indicative of horizontal gene transfer.

3.4 Towards the Pan-Genome of *C. pseudotuberculosis*

3.4.1 The Reference Genome of C. pseudotuberculosis *1002*

C. pseudotuberculosis is an important animal pathogen and the causative agent of a disease that is commonly called caseous lymphadenitis (Dorella et al. 2006a). This

Table 3.3 Overview of predicted pathogenicity islands of *C. ulcerans* BR-AD-22

Reference	Name	Begin CDS	End CDS	Begin position	End position	Length (bp)	Prominent function of island gene(s)
BR-AD22	PICU1	CULC22_00019	CULC22_00042	19833	44743	24911	CRISPR locus, phospholipase D and iron acquisition genes
BR-AD22	PICU2	CULC22_00051	CULC22_00112	54498	118292	63795	CRISPR locus, ABC transport systems, two-component systems and transcriptional regulators
BR-AD22	PICU3	CULC22_00166	CULC22_00182	175545	192468	16924	Diverse functions and proteins with unknown function
BR-AD22	PICU4	CULC22_00224	CULC22_00236	248273	264854	16582	Diverse functions and proteins with unknown function
BR-AD22	PICU5	CULC22_00667	CULC22_00683	725593	742611	17019	Putrescine synthesis and ABC transport protein systems
BR-AD22	PICU6	CULC22_01155	CULC22_01200	1276769	1325503	48735	Phage-related proteins
BR-AD22	PICU7	CULC22_01654	CULC22_01724	1835091	1891202	56112	Diverse functions and proteins with unknown function
BR-AD22	PICU8	CULC22_01773	CULC22_01788	1944778	1967893	23116	Secreted proteins and proteins with unknown function
BR-AD22	PICU9	CULC22_01794	CULC22_01816	1972423	1989721	17299	Diverse functions and proteins with unknown function
BR-AD22	PICU10	CULC22_01921	CULC22_01985	2108976	2164364	55389	Diverse functions and proteins with unknown function
BR-AD22	PICU11	CULC22_02033	CULC22_02044	2214650	2227049	12400	Chaperone and proteins with unknown function
BR-AD22	PICU12	CULC22_02071	CULC22_02085	2254055	2265204	11150	Cytochrome C biosynthesis and proteins with unknown function
BR-AD22	PICU13	CULC22_02134	CULC22_02168	2333661	2374767	41107	Iron and oligopeptide transport system, urease operon and diverse functions
BR-AD22	PICU14	CULC22_02307	CULC22_02325	2550850	2574374	23525	Diverse functions and proteins with unknown function

disease is found in the major sheep and goat production areas worldwide and causes significant economic losses. The strain selected for the first genome sequencing project was *C. pseudotuberculosis* 1002, which was isolated from goat caseous granulomas in Bahia state (Brazil). This strain has been licensed as a live attenuated vaccine strain in Brazil (Dorella et al. 2006a). The genome of *C. pseudotuberculosis* 1002 was sequenced using both 'classical' Sanger and pyrosequencing technologies (Ruiz et al. 2011). The genome sequencing project initially started with only 215 genomic survey sequences (GSSs) obtained from random samples of a BAC library of *C. pseudotuberculosis* 1002 (Dorella et al. 2006b). This representative library contained about 18,000 clones with inserts ranging in size from 25 to 120 kb and provided a 390-fold coverage of the *C. pseudotuberculosis* genome. Many GSSs (80.4 %) revealed significant similarity to the genome sequence of *C. diphtheriae* NCTC 13129 confirming the very close phylogenetic relationship of both species (Dorella et al. 2006b). Pyrosequencing was carried out with the Roche/454 Genome Sequencer FLX System and a sequencing depth finally resulting in 31 × coverage of the *C. pseudotuberculosis* 1002 genome (Ruiz et al. 2011). The chromosome of *C. pseudotuberculosis* 1002 has a size of 2,335,112 bp with a G + C content of 52.19 % and contains 2,111 predicted protein-coding regions, of which 53 were annotated as pseudogenes (Ruiz et al. 2011).

Meanwhile, five additional genome sequences of *C. pseudotuberculosis* isolates have been determined and published (Table 3.1), including *C. pseudotuberculosis* C231 from a sheep in Australia (Ruiz et al. 2011), *C. pseudotuberculosis* I19 from a cow with mastitis in Israel (Silva et al. 2011), *C. pseudotuberculosis* PAT10 from a sheep with lung abscess in Argentina (Cerdeira et al. 2011b), *C. pseudotuberculosis* CIP 52.97 from a horse with ulcerative lymphangitis in Kenya (Cerdeira et al. 2011c) and *C. pseudotuberculosis* FRC41 from a young French girl with necrotizing lymphadenitis (Trost et al. 2010b), which was the first genome sequence publicly available for this species. The complete genome sequences of *C. pseudotuberculosis* C231 and *C. pseudotuberculosis* FRC41 were both determined with the Roche/454 Genome Sequencer FLX System, whereas the Life Technologies SOLiD System was used for the remaining three genome projects. To address the problem of short reads in the case of the latter next-generation sequencing technology, a new hybrid *de novo* assembly strategy was developed combining De Bruijn graph and Overlap-Layout-Consensus methods (Cerdeira et al. 2011a). This *in silico* approach was used in a case study to assemble the complete genome sequence of *C. pseudotuberculosis* I19 from short reads (Cerdeira et al. 2011a). Briefly, contigs were assembled *de novo* from the short reads and were oriented using the complete genome sequence of *C. pseudotuberculosis* FRC41 as a reference for anchoring. Remaining gaps in the genome sequence of *C. pseudotuberculosis* I19 were closed using an iterative anchoring of additional short reads adjacent to sequence gaps (Cerdeira et al. 2011a). This new assembly strategy is feasible as the sequenced *C. pseudotuberculosis* genomes show a highly similar architecture and a highly conserved order of orthologous coding regions (Ruiz et al. 2011).

3.4.2 Comparative Genomics and the Pan-Genome of **C. pseudotuberculosis**

The number of core genes of *C. pseudotuberculosis* was calculated with the software EDGAR using bidirectional best BLASTP hits for genome comparisons (Blom et al. 2009). Based on a set of calculations using all *C. pseudotuberculosis* genomes individually as references, the core genome of the hitherto sequenced *C. pseudotuberculosis* isolates comprises 1,810 genes that can be regarded as highly conserved in this species. The bioinformatic characterization of the unique genome of *C. pseudotuberculosis* revealed a very low number of strain-specific genes in the four *C. pseudotuberculosis* biovar *ovis* isolates 1002, C231, I19 and PAT10, whereas 86 unique genes were detected in the genome of the *C. pseudotuberculosis* biovar *equi* isolate CIP 52.7 and 49 in *C. pseudotuberculosis* FRC41 from a human clinical source (Table 3.1). This preliminary data indicate that the genomes of *C. pseudotuberculosis* biovar *ovis* isolates have very similar gene contents. Accordingly, the sum total of protein-coding regions representing the pan-genome of *C. pseudotuberculosis* currently comprises only 2,630 genes, which is just about 1.5 times the size of the predicted core genome.

The close similarity between the sequenced *C. pseudotuberculosis* strains is also evident when comparing the structure of the CRISPR/*cas* regions. All *C. pseudotuberculosis* isolates share a CRISPR type IV with only one repeat sequence, with the exception of *C. pseudotuberculosis* CIP 52.97 that completely lacks a CRISPR/*cas* region (Table 3.1). It is therefore unlikely that spoligotyping is a suitable approach to analyze the genetic diversity of *C. pseudotuberculosis* isolates.

3.4.3 Pathogenicity Islands and Virulence Factors of **C. pseudotuberculosis**

Pathogenicity islands of *C. pseudotuberculosis* were detected and annotated in the reference genome of *C. pseudotuberculosis* 1002 by the means of the recently developed software PIPS (Ruiz et al. 2011; Soares et al. 2012). The *C. pseudotuberculosis* 1002 genome includes eleven putative pathogenicity islands (Table 3.4), which contain several classical virulence factors, including genes for pilus subunits, adhesion factors, iron uptake systems and secreted toxins. All of the candidate virulence factors in the islands have characteristics that indicate horizontal transfer. Comparative *in silico* analysis of the predicted pathogenicity islands with the BRIG software (Alikhan et al. 2011) revealed that most of the respective genes belong to the distributed genome of *C. pseudotuberculosis* and are only present in genome sequences of biovar *ovis* isolates, whereas others are partially or completely conserved in almost all strains (Fig. 3.4). This data indicates that prominent differences exist in the genetic repertoires of isolates belonging to the *C. pseudotuberculosis* biovars *ovis* or *equi*.

Table 3.4 Overview of predicted pathogenicity islands of *C. pseudotuberculosis* 1002

Reference	Name	Begin CDS	End CDS	Begin position	End position	Length (bp)	Prominent function of island gene(s)
1002	PICP1	Cp1002_0022	Cp1002_0031	19903	29136	9234	Phospholipase D and iron acquisition genes
1002	PICP2	Cp1002_0040	Cp1002_0067	38609	68350	29742	Iron and choline transport system and transcriptional regulators
1002	PICP8	Cp1002_0159	Cp1002_0167	163700	176472	12773	Purine nucleoside phosphorylase and deoxyribonucleoside regulator
1002	PICP3	Cp1002_0174	Cp1002_0185	186099	196852	10754	Iron transport system
1002	PICP9	Cp1002_0553	Cp1002_0573	575444	604590	29147	Diverse functions and proteins with unknown function
1002	PICP4	Cp1002_0980	Cp1002_0992	1057400	1076800	19401	Iron transport system
1002	PICP5	Cp1002_1445	Cp1002_1472	1588030	1617021	28992	Iron transport system and transcriptional regulators
1002	PICP6	Cp1002_1553	Cp1002_1565	1701830	1713424	11595	Transport system with unknown function
1002	PICP10	Cp1002_1617	Cp1002_1633	1766713	1788178	21466	Diverse functions
1002	PICP7	Cp1002_1903	Cp1002_1932	2089913	2123600	33688	Iron and oligopeptide transport system, urease operon and diverse functions
1002	PICP11	Cp1002_2069	Cp1002_2080	2290665	2301762	11098	Diverse functions and proteins with unknown function

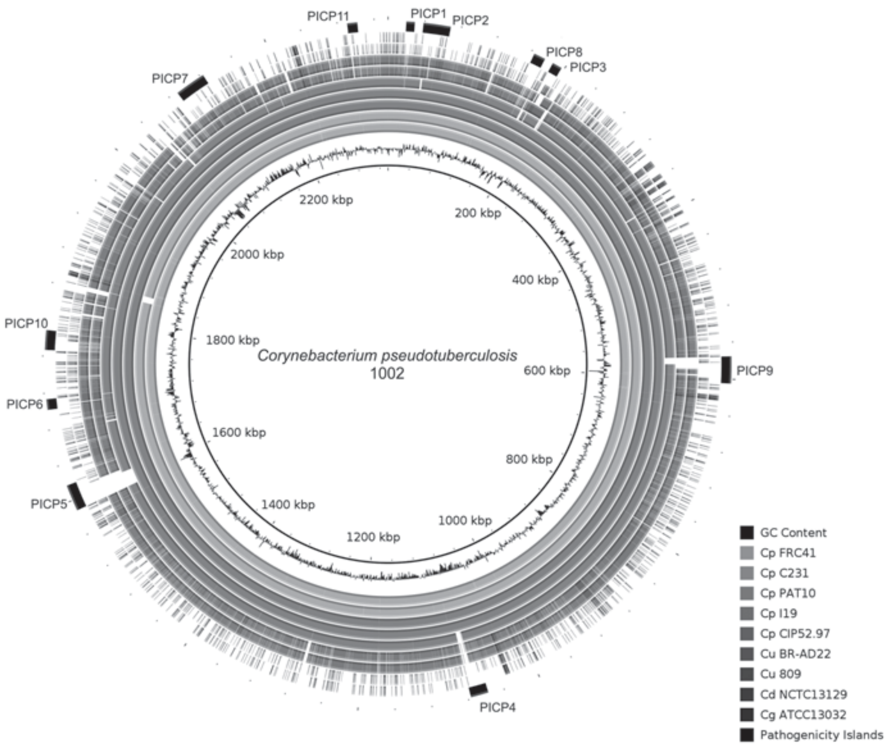

Fig. 3.4 Circular genome comparison between *C. pseudotuberculosis* strains using *C. pseudotuberculosis* 1002 as a reference. The circular genome comparison shows the positions of putative pathogenicity islands in the genome of the reference strain *C. pseudotuberculosis* 1002 (biovar *ovis*) and the presence/absence of these islands in other *C. pseudotuberculosis* biovar *ovis* strains (C231, PAT10 and I19), a *C. pseudotuberculosis* biovar *equi* strain (CIP 52.97), an isolate from a human clinical source (FRC41), and in other corynebacterial species. Abbreviations: *GC Content* G+C profile of a genome region; *Cp C. pseudotuberculosis*; *Cu C. ulcerans*; *Cd C. diphtheriae*; *Cg C. glutamicum*; *PICP* putative pathogenicity island of *C. pseudotuberculosis*

Despite the importance of *C. pseudotuberculosis* for animal health, there is little information about the pathogenesis and the facultative intracellular lifestyle of this bacterium. Only few virulence factors were identified previously in *C. pseudotuberculosis* (Dorella et al. 2006a), of which the most prominent one is phospholipase D (Pld), a sphingomyelin-degrading exotoxin (McKean et al. 2007). The annotation of the complete *C. pseudotuberculosis* FRC41 genome sequence provided additional knowledge of candidate virulence factors in this species (Trost et al. 2010b). In addition to the virulence factor phospholipase D, the endoglycosidase EndoE (misleadingly described as corynebacterial protease CP40 in previous studies) is encoded in the genome of this human isolate. The *ndoE* gene product of *C. pseudotuberculosis* FRC41 revealed sequence similarity to the α-domain of the secreted endoglycosidase EndoE from *Enterococcus faecalis* (Collin and Fischetti 2004.).

EndoE from *E. faecalis* is a two-domain protein that is characterized by two distinct activities involved in the degradation of N-linked glycans from ribonuclease B and the hydrolysis of the conserved glycans on IgG. The latter activity of the enzyme was assigned exclusively to the β-domain of EndoE, suggesting that the homologous protein from *C. pseudotuberculosis* has only endoglycosidase activity. In this way, *C. pseudotuberculosis* is probably able to interact directly with the mammalian host by glycolytic modulation of host glycoproteins (Trost et al. 2010b). The genome annotation of *C. pseudotuberculosis* FRC41 revealed serine proteases, neuraminidase H, nitric oxide reductase, an invasion-associated protein and acyl-CoA carboxylase subunits involved in mycolic acid biosynthesis as additional candidate virulence factors. Moreover, a gene-regulatory network analysis suggested that the cAMP-sensing transcriptional regulator GlxR plays a key role in controlling the expression of several genes contributing to virulence of *C. pseudotuberculosis* (Trost et al. 2010b). The human isolate *C. pseudotuberculosis* FRC41 is furthermore equipped with SpaA‘ and SpaD‘ gene clusters encoding protein subunits involved in the sortase-mediated polymerization of adhesive pili (Fig. 3.3). The pilus gene cluster of the SpaA‘-type is present in all sequenced genomes of *C. pseudotuberculosis* (Fig. 3.3).

3.5 Future Perspectives

The development of ultra-fast next-generation sequencing technologies has opened a new era of microbial genomics, enabling the rapid and detailed characterization of bacterial genomes and associated bacterial lifestyles. This progress in microbial genomics is obviously helping to shape our understanding of bacterial evolution. In particular, comparative genomics, or on a broader scale pan-genomics, affords the opportunity to detect species-specific features of a genome or strain-specific traits such as virulence factors contributing to the pathogenicity of bacteria. The systematic application of next-generation sequencing technologies also provides the possibility to generate bacterial DNA sequence data of extraordinary resolution, making it possible to identify single nucleotide changes within entire genomes and to map genome-wide single-nucleotide polymorphisms (SNPs). This type of studies provides data of very fine-scale resolution and enables the detection of the evolutionary history of multiple isolates within a clonal bacterial lineage. Distinguishing clonal groups within a pathogenic bacterial species was initially performed by phenotypic and subsequently by genotypic typing techniques and has been the cornerstone of infectious disease epidemiology, allowing the identification and tracking of clones responsible for infection and disease. Sequence-based typing approaches, such as multilocus sequence typing, have relied on the variation within a few selected marker genes. Although this technique is highly informative, it has a limited resolution when applied to closely related isolates. Approaches based on next-generation sequencing are also suitable for identifying subtle evolutionary events or for distinguishing clonal strains within a recent epidemic when applied on bacterial collections of known origin. Studies of the phylogeny of a bacterial species or of a clonal lineage within a species are highly dependent on the quantity and diversity of sampled isolates. However, the recent pan-genomic study of *C. diphtheriae* demonstrated that it has become possible to fully sequence significant numbers of isolates in a strain collection in reasonable time, thereby revealing new information on the plasticity of the *C. diphtheriae* genome. In principle, the size and the composition of the investigated strain collection is crucial for subsequent biological interpretations. This is particularly relevant for bacterial pathogens that reside in multiple niches and is therefore considered in the ongoing genome sequencing of *C. pseudotuberculosis* isolates from various geographical regions and sources, including sheep, goats, cows, horses, and humans. This strategy prevents bias in the data sets and provides a more complete picture of the true diversity of the bacterial species. A pan-genomic approach is also feasible for the characterization of *C. ulcerans*, which has been recognized in a broad spectrum of animal hosts. Whole-genome sequencing also facilitates the identification of gene losses and gene gains that can play a significant role in the evolution or pathogenicity of a bacterial species, as indicated by the detection of the candidate virulence factor *rbp* in *C. ulcerans* 809. Therefore, next-generation sequencing technologies provide a means of rapidly detecting associations between phenotype and genotype. The next few years will see an increase in the biological interpretation of such data using either high-throughput *in vitro* assays or the selected testing by targeted genetic experiments. This should further improve our understanding of the various forces that are important in the evolution of bacterial pathogens and enable the development of appropriate interventions. The next few years also promise an enhanced understanding of how and why epidemic clones emerge or disappear, and ultimately the better management and treatment of infectious diseases.

Acknowledgement The authors thank Siomar de Castro Soares for bioinformatic support, running the PIPS pipeline and visualizing comparative genomic data with the BRIG software.

References

Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402

Barksdale WL, Pappenheimer AM Jr (1954) Phage-host relationships in nontoxigenic and toxigenic diphtheria bacilli. J Bacteriol 67:220–232

Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter FJ, Zakrzewski M, Goesmann A (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10:154

Brüssow H, Canchaya C, Hardt WD (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68:560–602

Bukovska G, Klucar L, Vlcek C, Adamovic J, Turna J, Timko J (2006) Complete nucleotide sequence and genome analysis of bacteriophage BFK20—a lytic phage of the industrial producer *Brevibacterium flavum*. Virology 348:57–71

Cerdeira LT, Carneiro AR, Ramos RT, Almeida SS de, D'Afonseca V, Schneider MP, Baumbach J, Tauch A, McCulloch JA, Azevedo VA, Silva A (2011a) Rapid hybrid *de novo* assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. J Microbiol Methods 86:218–223

Cerdeira LT, Pinto AC, Schneider MP, Almeida SS de, Santos AR dos, Barbosa EG, Ali A, Barbosa MS, Carneiro AR, Ramos RT, Oliveira RS de, Barh D, Barve N, Zambare V, Belchior SE, Guimarães LC, Castro SS de, Dorella FA, Rocha FS, Abreu VA de, Tauch A, Trost E, Miyoshi A, Azevedo V, Silva A (2011b) Whole-genome sequence of *Corynebacterium pseudotuberculosis* PAT10 strain isolated from sheep in Patagonia, Argentina. J Bacteriol 193:6420–6421

Cerdeira LT, Schneider MP, Pinto AC, Almeida SS de, Santos AR dos, Barbosa EG, Ali A, Aburjaile FF, Abreu VA de, Guimarães LC, Castro SS de, Dorella FA, Rocha FS, Bol E, Gomes deSPH, Lopes TS, Barbosa MS, Carneiro AR, Jucá Ramos RT, Coimbra NA, Lima AR, Barh D, Jain N, Tiwari S, Raja R, Zambare V, Ghosh P, Trost E, Tauch A, Miyoshi A, Azevedo V, Silva A (2011c) Complete genome sequence of *Corynebacterium pseudotuberculosis* strain CIP 52.97, isolated from a horse in Kenya. J Bacteriol 193:7025–7026

Cerdeño-Tarrága AM, Efstratiou A, Dover LG, Holden MT, Pallen M, Bentley SD, Besra GS, Churcher C, James KD, De Zoysa A, Chillingworth T, Cronin A, Dowd L, Feltwell T, Hamlin N, Holroyd S, Jagels K, Moule S, Quail MA, Rabbinowitsch E, Rutherford KM, Thomson NR, Unwin L, Whitehead S, Barrell BG, Parkhill J (2003) The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res 31:6516–6523

Collin M, Fischetti VA (2004) A novel secreted endoglycosidase from *Enterococcus faecalis* with activity on human immunoglobulin G and ribonuclease B. J Biol Chem 279:22558–22570

D'Afonseca V, Soares SC, Ali A, Santos AR, Pinto AC, Magalhães AA, Faria CJ, Barbosa E, Guimarães LC, Eslabão M, Almeida SS, Abreu VA, Zerlotini A, Carneiro AR, Cerdeira LT, Ramos RT, Hirata R Jr, Mattos-Guaraldi AL, Trost E, Tauch A, Silva A, Schneider MP, Miyoshi A, Azevedo V (2012) Reannotation of the *Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria. Open Access Bioinformatics 4:1–13

Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64:475–93

Dias AA, Silva FC Jr, Pereira GA, Souza MC, Camello TC, Damasceno JA, Pacheco LG, Miyoshi A, Azevedo VA, Hirata R Jr, Bôas MH, Mattos-Guaraldi AL (2010) *Corynebacterium ulcerans* isolated from an asymptomatic dog kept in an animal shelter in the metropolitan area of Rio de Janeiro, Brazil. Vector Borne Zoonotic Dis 10:743–748

Dittmann S, Wharton M, Vitek C, Ciotti M, Galazka A, Guichard S, Hardy I, Kartoglu U, Koyama S, Kreysler J, Martin B, Mercer D, Ronne T, Roure C, Steinglass R, Strebel P, Sutter R, Trostle M. 2000. Successful control of epidemic diphtheria in the states of the Former Union of Soviet Socialist Republics: lessons learned. J Infect Dis. 181:10–22

Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V (2006a) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet. Res 37:201–218

Dorella FA, Fachin MS, Billault A, Dias Neto E, Soravito C, Oliveira SC, Meyer R, Miyoshi A, Azevedo V (2006b) Construction and partial characterization of a *Corynebacterium pseudotuberculosis* bacterial artificial chromosome library through genomic survey sequencing. Genet Mol Res 5:653–663

Droege M, Hill B (2008) The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete datasets. J. Biotechnol 136:3–10

Freeman VJ (1951) Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. J Bacteriol 61:675–688

Halachev MR, Loman NJ, Pallen MJ (2011) Calculating orthologs in bacteria and archaea: a divide and conquer approach. PLoS One 6:e28388

Hirata R Jr, Souza SM, Rocha-de-Souza CM, Andrade AF, Monteiro-Leal LH, Formiga LC, Mattos-Guaraldi AL (2004) Patterns of adherence to HEp-2 cells and actin polymerisation by toxigenic *Corynebacterium diphtheriae* strains. Microb Pathog 36:125–130

Hogg RA, Wessels J, Hart J, Efstratiou A, De Zoysa A, Mann G, Allen T, Pritchard GC (2009) Possible zoonotic transmission of toxigenic *Corynebacterium ulcerans* from companion animals in a human case of fatal diphtheria. Vet Rec 165:691–692

Iwaki M, Komiya T, Yamamoto A, Ishiwa A, Nagata N, Arakawa Y, Takahashi M (2010) Genome organization and pathogenicity of *Corynebacterium diphtheriae* C7(–) and PW8 strains. Infect Immun 78:3791–3800

Jolley KA, Chan MS, Maiden MC (2004) mlstdbNet – distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics 5:86

Kunkle CA, Schmitt MP (2005) Analysis of a DtxR-regulated iron transport and siderophore biosynthesis gene cluster in *Corynebacterium diphtheriae*. J Bacteriol 187:422–433

LaPointe P, Wei X, Gariepy J (2005) A role for the protease-sensitive loop region of Shiga-like toxin 1 in the retrotranslocation of its A1 domain from the endoplasmic reticulum lumen. J Biol Chem 280:23310–23318

Lewis CM Jr, Obregón-Tito A, Tito RY, Foster MW, Spicer PG (2012) The Human Microbiome Project: lessons from human genomics. Trends Microbiol 20:1–4

Mandlik A, Swierczynski A, Das A, Ton-That H (2007) *Corynebacterium diphtheriae* employs specific minor pilins to target human pharyngeal epithelial cells. Mol Microbiol 64:111–124

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genom Hum Genet 9:387–402

Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11:181–190

Mattos-Guaraldi AL, Sampaio JL, Santos CS, Pimenta FP, Pereira GA, Pacheco LG, Miyoshi A, Azevedo V, Moreira LO, Gutierrez FL, Costa JL, Costa-Filho R, Damasco PV, Camello TC, Hirata R Jr (2008) First detection of *Corynebacterium ulcerans* producing a diphtheria-like toxin in a case of human with pulmonary infection in the Rio de Janeiro metropolitan area, Brazil. Mem Inst Oswaldo Cruz 103:396–400

McKean SC, Davies JK, Moore RJ (2007) Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death. Microbiology 153:2203–2211

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Mokrousov I (2009) *Corynebacterium diphtheriae*: genome diversity, population structure and genotyping perspectives. Infect Genet Evol 9:1–15

Mokrousov I, Narvskaya O, Limeschenko E, Vyazovaya A (2005) Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. J Clin Microbiol 43:1662–1668

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–264

O'Loughlin EV, Robins-Browne RM (2001) Effect of Shiga toxin and Shiga-like toxins on eukaryotic cells. Microbes Infect 3:493–507

Ott L, Höller M, Rheinlaender J, Schäffer TE, Hensel M, Burkovski A (2010) Strain-specific differences in pili formation and the interaction of *Corynebacterium diphtheriae* with host cells. BMC Microbiol 10:257

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40:D571–579.

Park WH, Williams AW (1896) The production of diphtheria toxin. J Exp Med 1:164–185

Qian Y, Lee JH, Holmes RK (2002) Identification of a DtxR-regulated operon that is essential for siderophore-dependent iron uptake in *Corynebacterium diphtheriae*. J Bacteriol 184:4846–4856

Rappuoli R, Michel JL, Murphy JR (1983) Restriction endonuclease map of corynebacteriophage γ_c^{tox+} isolated from the Park-Williams no. 8 strain of *Corynebacterium diphtheriae*. J Virol 45:524–530

Ratti G, Covacci A, Rappuoli R (1997) A tRNA$_2^{Arg}$ gene of *Corynebacterium diphtheriae* is the chromosomal integration site for toxinogenic bacteriophages. Mol Microbiol 25:1179–1181

Rogers EA, Das A, Ton-That H (2011) Adhesion by pathogenic corynebacteria. Adv Exp Med Biol 715:91–103

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26:1117–1124

Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AA, Lopes DO, Dorella FA, Pacheco LG, Costa MP, Turk MZ, Seyffert N, Moraes PM, Soares SC, Almeida SS, Castro TL, Abreu VA, Trost E, Baumbach J, Tauch A, Schneider MP, McCulloch J, Cerdeira LT, Ramos RT, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, Pedrosa AL, Vieira CU, Guimarães CT, Bartholomeu DC, Oliveira DM, Santos FR, Rabelo ÉM, Lobo FP, Franco GR, Costa AF, Castro IM, Dias SR, Ferro JA, Ortega JM, Paiva LV, Goulart LR, Almeida JF, Ferro MI, Carneiro NP, Falcão PR, Grynberg P, Teixeira SM, Brommonschenkel S, Oliveira SC, Meyer R, Moore RJ, Miyoshi A, Oliveira GC, Azevedo V (2011) Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. PLoS One 6:e18551

Salzberg SL (2007) Genome re-annotation: a wiki solution? Genome Biol 8:102

Schröder J, Maus I, Meyer K, Wördemann S, Blom J, Jaenicke S, Schneider J, Trost E, Tauch A (2012) Complete genome sequence, lifestyle, and multi-drug resistance of the human pathogen *Corynebacterium resistens* DSM 45100 isolated from blood samples of a leukemia patient. BMC Genomics 13:141

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Silva A, Schneider MP, Cerdeira L, Barbosa MS, Ramos RT, Carneiro AR, Santos R, Lima M, D'Afonseca V, Almeida SS, Santos AR, Soares SC, Pinto AC, Ali A, Dorella FA, Rocha F, Abreu VA de, Trost E, Tauch A, Shpigel N, Miyoshi A, Azevedo V (2011) Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis. J Bacteriol 193:323–324

Soares SC, Dorella FA, Pacheco LG, R. Hirata R Jr, Mattos-Guaraldi AL, Azevedo V, Miyoshi A (2011) Plasticity of *Corynebacterium diphtheriae* pathogenicity islands revealed by PCR. Genet Mol Res 10:1290–1294

Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A, Baumbach J, Trost E, Tauch A, Hirata R Jr, Mattos-Guaraldi AL, Miyoshi A, Azevedo V (2012) PIPS: Pathogenicity Island Prediction Software. PLoS One 7:e30848

Soriano F, Tauch A (2008) Microbiological and clinical features of *Corynebacterium urealyticum*: urinary tract stones and genomics as the Rosetta Stone. Clin Microbiol Infect 14:632–643

Tauch A, Trost E, Tilker A, Ludewig U, Schneiker S, Goesmann A, Arnold W, Bekel T, Brinkrolf K, Brune I, Götker S, Kalinowski J, Kamp PB, Lobo FP, Viehoever P, Weisshaar B, Soriano F, Dröge M, Pühler A (2008a) The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. J Biotechnol 136:11–21

Tauch A, Schneider J, Szczepanowski R, Tilker A, Viehoever P, Gartemann KH, Arnold W, Blom J, Brinkrolf K, Brune I, Götker S, Weisshaar B, Goesmann A, Dröge M, Pühler A (2008b) Ultrafast pyrosequencing of *Corynebacterium kroppenstedtii* DSM44385 revealed insights into the physiology of a lipophilic corynebacterium that lacks mycolic acids. J Biotechnol 136:22–30

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477

Ton-That H, Schneewind O (2003) Assembly of pili on the surface of *Corynebacterium diphtheriae*. Mol Microbiol 50:1429–1438

Trost E, Götker S, Schneider J, Schneiker-Bekel S, Szczepanowski R, Tilker A, Viehoever P, Arnold W, Bekel T, Blom J, Gartemann KH, Linke B, Goesmann A, Pühler A, Shukla SK, Tauch A (2010a) Complete genome sequence and lifestyle of black-pigmented *Corynebacterium aurimucosum* ATCC 700975 (formerly *C. nigricans* CN-1) isolated from a vaginal swab of a women with spontaneous abortion. BMC Genomics 11:91

Trost E, L O, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Castro SS de, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A (2010b) The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genomics 11:728

Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoever P, Burkovski A, Soares SC, Almeida SS, Dorella FA, Miyoshi A, Azevedo V, Schneider MP, Silva A, Santos CS, Santos LS, Sabbadini P, Dias AA, Hirata R Jr, Mattos-Guaraldi AL, Tauch A (2011) Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. BMC Genomics 12:383

Trost E, Blom J, Castro SS de, Huang IH, Al-Dilaimi A, Schröder J, Jaenicke S, Dorella FA, Rocha FS, Miyoshi A, Azevedo V, Schneider MP, Silva A, Camello TC, Sabbadini PS, Santos CS, Santos LS, Hirata R Jr, Mattos-Guaraldi AL, Efstratiou A, Schmitt MP, Ton-That H, Tauch A (2012) Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol 194:3199–3215