Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Reannotation of the *Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria

This article was published in the following Dove Press journal:
Open Access Bioinformatics
24 February 2012
Number of times this article has been viewed

Vívian D'Afonseca[1,*]
Siomar C Soares[1,*]
Amjad Ali[1]
Anderson R Santos[1]
Anne C Pinto[1]
Aryane AC Magalhães[1]
Cássio de Jesus Faria[1]
Eudes Barbosa[1]
Luis C Guimarães[1]
Marcus Eslabão[2]
Sintia S Almeida[1]
Vinicius AC Abreu[1]
Adhemar Zerlotini[3,4]
Adriana R Carneiro[5]
Louise T Cerdeira[5]
Rommel TJ Ramos[5]
Raphael Hirata Jr[6]
Ana L Mattos-Guaraldi[6]
Eva Trost[7]
Andreas Tauch[7]
Artur Silva[5]
Maria P Schneider[5]
Anderson Miyoshi[1]
Vasco Azevedo[1]

[1]Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [2]Federal University of Pelotas, Pelotas, Rio Grande do Sul, Brazil; [3]FIOCRUZ - CEBIO, Belo Horizonte, Minas Gerais, Brazil; [4]EMBRAPA - CNPTIA, Campinas, São Paulo, Brazil; [5]Federal University of Pará, Belém, Pará, Brazil; [6]Rio de Janeiro State University, Rio de Janeiro, Brazil; [7]Center for Biotechnology, Bielefeld University, Bielefeld, Germany

*These authors contributed equally to this work

Correspondence: Vasco Azevedo
Federal University of Minas Gerais,
Belo Horizonte, 31907-270,
Minas Gerais, Brazil
Tel +55 31 3409 2610
Fax +55 31 3409 2610
Email vasco@icb.ufmg.br

Background: The reannotation of genomes already on file is a new approach to discovering new genetic elements and to make the genomes more descriptive and current with relevant features regarding the organism's lifestyle. Within this approach, the present study aimed to reannotate the genome of the Gram-positive human pathogen *Corynebacterium diphtheriae*, which causes diphtheria. The deposit of massive amounts of information linked to other species of the genus *Corynebacterium* has facilitated the updating of the genomic interpretation of this microorganism. Additionally, the emergence of invasive disease by nontoxigenic strains of *C. diphtheriae* and the reemergence of diphtheria in partially immunized populations have given impetus to new studies in relation to its structural and functional genome.

Results: In relation to structural genomics, 23 coding regions (coding sequences) were deleted and 71 new genes were added to the genome annotation. Nevertheless, all the pseudogenes were validated and ten new pseudogenes were created. In relation to functional genomics, about 57% of the genome annotation was updated and became functionally more informative. The product descriptions of 41% (973 proteins) were updated. Among them, 370 that were previously annotated as "hypothetical proteins," now have more informative descriptions. With the new annotation, the plasticity of the genome became evident, which shows improvements in the annotation of 13 pathogenicity islands already described in the literature. In addition, the large number of transposases and the presence of structural genes of bacteriophages make their genomic versatility evident. Contrasting with this reality, it also allowed the clarification of some aspects concerned with mechanisms used by *C. diphtheriae* to stop the invasion of the genome by bacteriophages, mediated by the clustered regularly interspaced short palindromic repeats region.

Conclusion: The reannotation of the *C. diphtheriae* genome provided an improvement in annotation of the *C. diphtheriae* genome in several aspects, such as virulence characteristics and plasticity events. Moreover, the protocol used here can be extended to various other pathogens in order to improve the genomic information already on file in public databases and to minimize propagating errors. The reannotated archive and updated archive are available at: http://lgcm.icb.ufmg.br/pub/C_diphtheriae_reannotation.embl.

Keywords: *Corynebacterium diphtheriae*, diphtheria, reannotation, CRISPR, pathogenicity islands, genome

Background

In recent years, genomics has regained its foothold in the areas of science that are in full development. With the advent of new sequencing platforms, known as the next generation, the amount of genomic data available in public databases has increased exponentially.[1]

This is due to the fact that, currently, the acquisition of genomic data happens in a rapid, efficient, accurate, and low-cost manner. Research groups may then begin projects with their favorite organisms.[2] The reflection of this expansion may be seen in the database Genomes OnLine Database v 3 (http://www.genomesonline.org). There are currently around 10,420 genome projects in progress, and approximately 1700 genomes have been completed and published.

Meanwhile, on the one side, the massive generation of genomic data is good for science on the whole; on the other side, it has brought about the propagation of errors from the annotation where genomes, annotated automatically and without physical oversight, are deposited daily in public domain databases. Connected to this, many genome annotations deposited years ago were not updated, thus worsening this scenario.[3,4] As one approach to improving this panorama and minimizing the propagation of errors, some groups are already undertaking the process called reannotation, in which a genome already deposited passes through a new process of the prediction of genes and other structural elements of the genome; afterwards, they are reviewed manually by a specialist on the organism, and every open reading frame (ORF) has its product reannotated with the aim of improving its description.[2]

Few organisms were reannotated until today.[4–8] However, the reported improvement and the description of new genomic elements have motivated the practice of this new approach. In *Escherichia coli* CFT073, the update allowed the identification of 299 new ORFs, including various classical elements of virulence present in pathogenicity islands (PAIs), which were not predicted in the first version of genome annotation.[5] In the *Campylobacter jejuni* NCTC11168 pathogen, various pseudogenes have been identified, and around 97.8% of the previous genome experienced some type of update, including a change in the description of the gene product, the gene symbol, and new description for hypothetical proteins.[6] Both of the latter cited studies undertook the updating of the genome annotations 7 years after they were published for the first time.

Based on this approach, the present study attempted to reannotate the genome of *Corynebacterium diphtheriae* biotype *gravis*, strain NCTC13129. The sequencing and subsequent availability of the *C. diphtheriae* NCTC13129 genome in public domain databases occurred in 2003, under accession number NC_002935, contributing to the understanding of the pathogenicity, virulence, and lifestyle of this pathogen.[9]

C. diphtheriae is a Gram-positive human pathogen and has a high guanine-cytosine content.[9] This pathogen has the ability to colonize the human respiratory tract and, through the action of its exotoxin, diphtheria toxin, forms a membranous exudate over the tonsils, pharynx, and/or nasal cavity.[10] Diphtheria was under control for decades, but it is currently among the reemerging diseases. More than 150,000 cases and 5000 deaths were reported from diphtheria from 1990–1999 in the Eastern European region. This number is in great contrast to reports from the previous decade, which did not surpass 600 cases, according to the World Health Organization. In addition to the European continent, other continents such as Asia, Africa, and South America have also reported significant numbers of diphtheria cases. After a huge effort from the health organizations to contain the disease, the immunization coverage has reached approximately 82% of the world population in 2009, decreasing the number of reported cases to 857 in the same year (World Health Organization). However, the identification of nontoxigenic *C. diphtheriae* strains, ie, strains unable to produce the diphtheria toxin, which have caused invasive diseases, such as endocarditis,[11,12] has to be treated as a new potential problem in public health.

Therefore, it has become highly relevant to improve and update the already existing data about this pathogen, with the goal of increasing genetic knowledge linked to its genome and to propose new approaches and precise diagnostics that will prevent or minimize the effects of its resurgence. The reannotated file can be downloaded freely through the authors' server, at: http://lgcm.icb.ufmg.br/pub/C_diphtheriae_reannotation.embl. Alternatively, it can also be downloaded through a public server: http://www.bioinformatics.org/groups/?group_id=1103.

Methods

Genome reannotation

The reannotation procedures involved the use of several algorithms, in a multi-step process. Structural annotation was performed using the following software: FGENESB: bacterial operon and gene predictor (http://www.softberry.com; Softberry, Inc, Mount Kisco, NY); RNAmmer: ribosomal ribonucleic acid (RNA) predictor (Center for Biological Sequence Analysis, Lyngby, Denmark);[13] tRNAscan-SE: transfer RNA predictor (Lowe Lab, Biomolecular Engineering, University of California, Santa Cruz, CA);[14] and Tandem Repeat Finder: repetitive deoxyribonucleic acid (DNA) predictor (Boston University, Boston, MA).[15]

Functional annotation was performed by similarity analyses, using Basic Local Alignment Search Tool (protein) – National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov; Bethesda, MD) with

a cutoff of 10^{-6} against a nonredundant database of proteins, InterProScan (European Bioinformatics Institute, Hinxton, Cambridgeshire, UK) and SignalP (Center for Biological Sequence Analysis) analysis.[16] Manual annotation was performed using Artemis (Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK).[17]

Criteria for manual curation

To improve the annotation, all coding sequences (CDSs) were manually curated. The correction of the initial methionine was guided by the presence of a signal peptide, and matched with homologous proteins of related organisms. The hits generated in a similarities search, with a minimum identity of 60% and the presence of the same results in almost all hits, were used to update the predicted products.

To improve the annotation of either hypothetical proteins or proteins without available product descriptions, large predicted conserved domains linked to CDS were used, when available.

Subcellular location of predicted proteins and gene targets for vaccine development

Predictions of the cellular locations of *Corynebacterium* proteins were made using the "subcellular localization" option of the software Vaxign. Classification of predicted proteins was done using "Dynamic Vaxign Analysis" of the Vaxign software (University of Michigan Medical School, Ann Arbor, MI),[18] in secreted and cell wall categories of subcellular location. Additionally, the software searched for MHC classes I and II binding proteins, transmembrane helices, and adhesion probability.

In silico identification of PAIs

In order to identify and classify accurately the putative PAIs in the corynebacterial genomes, the authors developed a combined computational approach using several in-house scripts to integrate the prediction of diverse algorithms and databases (http://www.genoma.ufpa.br/lgcm/pips). The algorithms and databases were: Colombo – SIGI-HMM (Institute of Computer Science, University of Göttingen, Göttingen, Germany),[19] Artemis,[17] tRNAscan-SE,[14] HMMER (v 3.0; Howard Hughes Medical Institute, Chevy Chase, MD),[20] Artemis Comparison Tool (Wellcome Trust Sanger Institute),[21] and mVIRdb (Lawrence Livermore National Laboratory, Livermore, CA).[22]

In silico metabolic pathway construction

The two main data sources used for reconstructing the *C. diphtheriae* metabolic pathways were the genome sequence file in FASTA format, and the genome annotation file in GenBank format. Metabolic pathway databases for strains 1002 and C231 were created using the Pathway Tools 13 software, developed by SRI International (Menlo Park, CA).[23] The Pathway Tools software contains algorithms that predict the metabolic pathways of an organism from its genome, by comparison to a reference pathways database known as MetaCyc.[24]

Construction of a metabolic pathways database was done, using BioCyc,[25] in order to compare the pathways of the bacteria *C. pseudotuberculosis* I19, *C. efficiens* YS-314, *C. glutamicum* ATCC 13032, and *C. jeikeium* K411 to the deduced *C. diphtheriae* pathways.

Results

Improving of the *C. diphtheriae* NCTC13129 genome annotation

C. diphtheriae NCTC13129 genome remained annotated without changes for 7 years. Today, the vast genomic information present in the databases allows this scenario to be altered.

Presently, the complete reannotation and updating of the *C. diphtheriae* genome annotation allowed its modification, in its various structural and functional aspects, of which approximately 57% of the prior genome annotation has undergone alteration, making it more descriptive. Additionally, this process assisted in the discovery of new genetic elements, which can provide us with new understanding about the virulence and plasticity of the microorganism. Based on this information, the updated genome annotation shows 2368 genes in contrast to the previous version which had 2320. Within this new tally, 23 CDSs of *C. diphtheriae* were deleted and 71 new CDSs were predicted and validated, as shown in Table 1. The new CDSs and pseudogenes, along with all the predicted PAIs, are represented in Figure 1. For more information about the similarity between the new CDSs and pseudogenes with proteins of the nonredundant database of proteins from NCBI, please refer to Additional File 1.

The criterion for the deletion of the CDSs was their use in the formation of new pseudogenes or the absence of biological evidence. In addition, in three cases (DIP0404, DIP0700, and DIP1975), the new prediction of CDS was done in the DNA strand opposite the predicted CDS in the genome deposited in NCBI Reference Sequence. These new predictions presented biological evidence with strong similarity to other species of the genus *Corynebacterium*, in contrast to the three deleted CDSs that were ORFans or showed meaningless matches in the protein databases. Figure 2 illustrates the reannotated

Table 1 Coding sequences deleted and/or modified from the previous version of the genome annotation (Below are the new coding sequences of the updated *Corynebacterium diphtheriae* genome)

Gene ID Cd* RefSeq	Product	New prediction	Begin	End	Strand	Product	Status
DIP0017	Hypothetical protein	DIP0016A	19709	20107	R	Hypothetical membrane protein	Match with *Corynebacterium* species
DIP0018/DIP0019/DIP0020	Hypothetical protein	DIP0020	20341	21185	R	Putative glycosylase (pseudogene)	Match with *Corynebacterium* species
DIP0039	Hypothetical protein	–	–	–	–	–	Overlap with CRISPR region
DIP0040	Hypothetical protein	–	–	–	–	–	Overlap with CRISPR region
DIP0142/DIP0143	Hypothetical protein	DIP0143	118306	119988	F	Transposase (pseudogene)	Pseudogene increased
DIP0239/DIP0240/DIP0241	Hypothetical protein	DIP0239	205490	206485	R	Putative surface-anchored protein (pseudogene)	Pseudogene increased
DIP0404	Hypothetical protein	DIP0403a	369731	370213	F	Putative membrane protein	Match with *Corynebacterium* species
DIP0700	Hypothetical protein	DIP0699a	676852	677295	R	Conserved hypothetical protein	Match with *Corynebacterium* species
DIP0734/DIP0735	Putative membrane protein	DIP0734	711318	712185	R	Putative sodium/glutamate symporter	Pseudogene increased
DIP0757/DIP0757A	IS element transposase (partial)	DIP0757	736001	736487	R	IS element transposase (partial)	Pseudogene increased
DIP0898/DIP0899	Hypothetical protein	DIP0899	868809	869972	F	HNH endonuclease domain protein	Pseudogene increased
DIP1106/DIP1110	Conserved hypothetical protein (pseudogene)	DIP1106	1093117	1094859	F	Putative signal-transduction protein containing cAMP-binding	Pseudogene increased
DIP1654/DIP1655	Conserved hypothetical protein	DIP1654	1689384	1690688	R	LGFP repeat superfamily protein	Pseudogene increased
DIP1820	Putative membrane protein	DIP1819A	1867528	1867842	R	Hypothetical protein	Match with *Corynebacterium* species
DIP1975	Hypothetical protein	DIP1974A	2022366	2022947	F	Lipoprotein LpqE	Match with *Corynebacterium* species
DIP2023/DIP2024	Hypothetical protein	DIP2023	2076614	2077463	F	Filamentation induced by cAMP protein	Match with *Corynebacterium* species
DIP2033/DIP2034	Putative transposase	DIP2034	2085350	2086593	R	Transposase, mutator family	Pseudogene increased
DIP2149/DIP2152	Putative transposase	DIP2149	2212812	2214416	R	Transposase for insertion sequence	Pseudogene increased
DIP2222	Putative exported protein	DIP2220A/ DIP2220B	2310566/ 2310715	2310790/ 2310927	F	Hypothetical protein	Match with *Corynebacterium* species
DIP2309/DIP2310	Putative DNA-binding protein	DIP2309	2405722	2407315	R	Divergent AAA domain protein	Create new pseudogene with CDS

New CDS with function

Gene ID Cd* RefSeq	Product	Status	Begin	End	Strand
DIP0020	glycosylase	New pseudogene	20341	21185	R
DIP0201	gp1, terminase	Pseudogene unmerged	166488	166832	F
DIP0201A	gp2, terminase	Pseudogene unmerged	166822	168423	F
DIP0493A	Putative molybdopterin converting factor	New CDS	459868	460140	F
DIP1267	Putative short chain dehydrogenase	New CDS	1275251	1275829	F
DIP1974A	Lipoprotein LpqE	New CDS	2022366	2022947	F
DIP2156A	Possible amidohydrolase	New CDS	2217532	2217855	F

New CDSs with unknown functions

Amount	Product
50	Hypothetical protein
9	Hypothetical secreted protein
4	Transposases
1	Hypothetical membrane protein

Notes: **C. diphtheriae*. RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviations: cAMP, cyclic adenosine monophosphate; CDS, coding sequence; CRISPR, clustered regularly interspaced palindromic repeats; DNA, deoxyribonucleic acid; ID, identification; IS, insertion sequence.

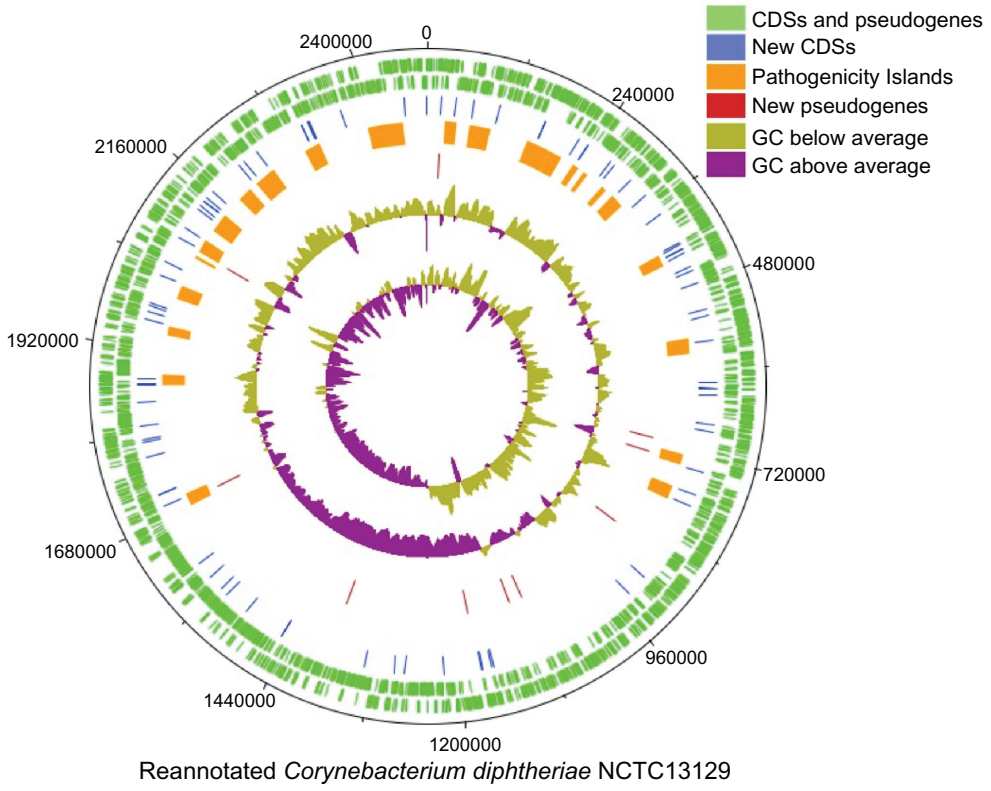

Figure 1 Genomic map showing the new coding sequences (CDSs) and pseudogenes along with all pathogenicity islands.
Notes: Rings from outside to inside: first and second rings (green), CDSs and pseudogenes which have not underpassed through modifications in length; third ring (blue), new CDSs; fourth ring (orange), pathogenicity islands; fifth ring (red), new pseudogenes; sixth ring (purple and brown), guanine-cytosine (GC) plot; seventh ring (purple and brown), GC skew.

region, in which the CDS DIP1975 was previously predicted in the NCBI Reference Sequence.

The product of the new CDS (DIP1974a) had strong similarity with the "lipoprotein – LpqE" protein, with various matches to homologous proteins from other species of *Corynebacterium*. In contrast, CDS DIP1975 did not show any hits with any phylogenetically related organism and for this reason was removed.

In spite of the sparse description of Gram-positive pathogens, it is known that lipoproteins, structural components of the membrane of various bacteria, are important factors linked to the stimulation of an immune response, especially in humans.[26] Hence, the new gene DIP1975 may be intimately related to the virulence of the *C. diphtheriae* and may be the target of studies for the development of new therapies.

Functional reannotation: new annotation reveals genetic elements of *C. diphtheriae* acting against foreign DNA

Various fields may be altered, based on searches for similarity and protein domains conserved in the new annotation. As shown in Figure 3, 43% of the genome annotation remains unaltered,

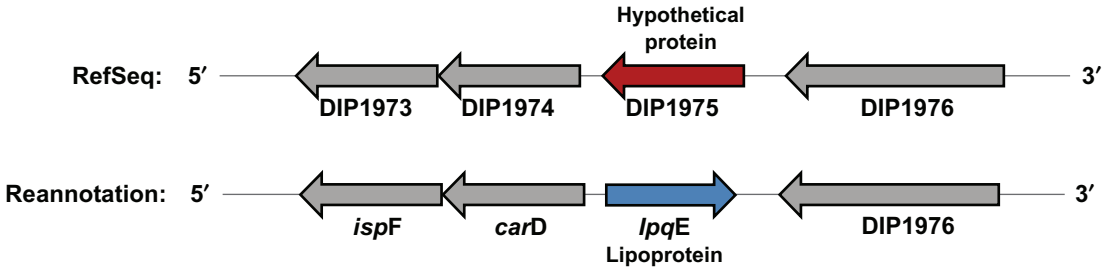

Figure 2 Illustrative schematic of the correction of open reading frames, for the correct orientation of the genome, based on protein similarity.
Notes: Open reading frame DIP1975 is shown in red, in the wrong orientation from the first annotation of the *Corynebacterium diphtheriae* NCTC13129 genome. The corrected open reading frame (DIP_1976) is illustrated in blue with its probable genetic product, which was predicted based on searches for protein similarity (Basic Local Alignment Search Tool [protein]) against the nonredundant protein database (cutoff: 10^{-6}). RefSeq is the National Center for Biotechnology Information Reference Sequence database.

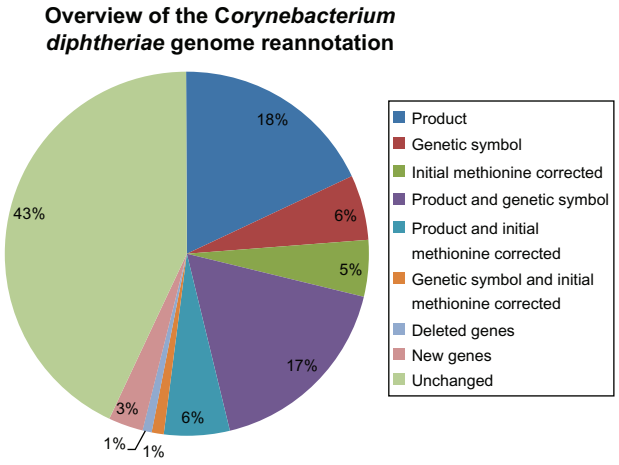

Figure 3 Overview of the changes that occurred in the *Corynebacterium diphtheriae* NCTC13129 genome after the process of reannotation, in the principal categories of change.

while 57% underwent alterations in various aspects. The field most frequently altered was the "product" of the genes (41%): that is, 973 proteins. Among them, 370 proteins ceased to be hypothetical proteins, conserved hypothetical proteins, and/or hypothetical membrane proteins. This alteration provided significant genetic knowledge of the various genes previously annotated as hypothetical proteins, and today, their functions are known, including specific genes encoding virulence factors and pathogenicity. Several acetyltransferases, receptors for iron-binding, fimbrial subunits, transposases, and proteins possibly linked to bacterial virulence were identified.

A common feature of those virulence factors is their location on PAIs, large regions acquired through horizontal gene transfer, which play important roles in the evolution of pathogenic bacteria. *C. diphtheriae* is known to harbor 13 putative PAIs (PiCds 1–13) with genes coding for putative iron transport genes, exported proteins, two component system proteins, insertion sequence transposases, and the diphtheria toxin coding gene (*tox*), which is located in a corynephage-acquired region.[9] Through the use of the software PIPS (http://www.genoma.ufpa.br/lgcm/pips), 11 additional PAIs were identified in the genome sequence of *C. diphtheriae* (Additional File 2) in the original genome annotation (PiCds 14–24). After the reannotation of the genome sequence, the same PiCds 1–24 were identified. However, this new finding deserves special attention.

Two clustered regularly interspaced short palindromic repeats (CRISPR) elements which were initially annotated as "hypothetical proteins" were found to be located in two regions identified as the fourteenth and thirteenth putative PAIs of *C. diphtheriae* (PiCd 14 and PiCd 13), respectively. Table 2 shows these ORFs with their new description. The existence of these regions and of gene families (*cas*) associated with CRISPR regions is known to play an important role against infection by bacteriophages and other mobile genetic elements.[27,28] Such sites showed various repetitions and, in these or some studies, were interpreted as an immune response mechanism against bacterial invasion.[29]

Proteins associated with these regions recognize foreign DNA and use it in a mechanism to silence DNA, similar to RNA interference.[28,30] The DNA is fragmented by *cas*-type proteins in segments of approximately 30 base pairs; the fragments are then inserted into the repetitive regions of CRISPR, which are expressed constitutively.[31,32] These expressed RNAs become guides for other *cas* proteins to process the foreign DNA, as occurs in the RNA interference mechanism.[28,30]

The function of CRISPR in immunity against mobile elements is clearly shown in *Enterococcus faecalis*, where the antibiotic-sensitive strain OG1RF, which possesses two CRISPR arrays, lacks most of the antibiotic resistance genes that are harbored by the hospital-adapted strain V583.[33] Interestingly, Palmer and Gilmore showed that five hybrid

Table 2 Clustered regularly interspaced short palindromic repeats (CRISPR) elements and associated genes described in the new annotation of the *Corynebacterium diphtheriae* genome NCTC13129

Gene ID RefSeq	New ID	Previous annotation RefSeq	Reannotation of *C. diphtheriae*	CRISPR Region
DIP0036	DIP0036	Conserved hypothetical protein	CRISPR-associated protein, Csn1 family	1
DIP0037	DIP0037	Conserved hypothetical protein	CRISPR-associated protein, Cas1 family	1
DIP0038	DIP0038	Conserved hypothetical protein	CRISPR-associated protein, Cas2 family	1
DIP2208	DIP2208	Conserved hypothetical protein	CRISPR-associated protein, Cas5 family	2
DIP2209	DIP2209	Conserved hypothetical protein	CRISPR-associated protein, Cas4 family	2
DIP2210	DIP2210	Conserved hypothetical protein	CRISPR-associated protein, Cse2 family	2
DIP2212	DIP2212	Conserved hypothetical protein	CRISPR-associated protein, Cse3 family	2
DIP2213	DIP2213	Putative helicase	CRISPR-associated helicase Cas3 family	2
DIP2214	DIP2214	Conserved hypothetical protein	CRISPR-associated protein, Cas1 family	2
DIP2215	DIP2215	Conserved hypothetical protein	CRISPR-associated protein, Cas2 family	2

Note: RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviation: ID, identification.

strains, originating by the acquisition of a resistance island of the strain V583 by OG1RF, are deficient in one CRISPR array possibly due to displacement of this region during DNA incorporation.[33] Moreover, they speculated that modern antibiotic therapy may facilitate the increase in plasticity through the disruption of the balance between the two opposing forces, the acquisition of foreign DNA and degradation of this DNA by self-defense mechanisms.[33]

Following the reannotation of the *C. diphtheriae* genome, the existence of two major operons became clear, denoted as CRISPR 1 region and CRISPR 2 region, which could assume that role in this pathogen. This information has already been presented in studies performed by Mokrousov,[34] but the name of genes and their products remain unchanged in the currently available *C. diphtheriae* genome file.

As shown in Table 2, the CRISPR 1 region is composed of three genes (DIP0036, DIP0037, and DIP0038), *cns1*, *cas1*, and *cas2*, respectively, which participate in the cascade of recognition and silencing of foreign DNA. The CRISPR 2 site of *C. diphtheriae* has seven genes (DIP2208–DIP2210 and DIP2212–DIP2215), containing the *casD*, *casC*, *casB*, *casF*, DIP2211, *casG*, and *casF* genes. Genes superimposed in these two regions were deleted from the annotation.

In spite of a vast genetic repertoire, containing genes that resist the invasion of bacteriophages and mobile genetic elements, the *C. diphtheriae* genome shows another reality. The new annotation showed a large number of transposases and structural proteins originating from bacteriophages. The presence of the CRISPR regions may be an indication that the genome could be even more plastic in their absence.

Pseudogenes, transposases, and other phase-variable elements

The reannotation validated all the pseudogenes of the *C. diphtheriae* genome. In the previous version there were 48 pseudogenes, and 51 pseudogenes were included in the new version. Of the existing pseudogenes, 31 remained the same between the two annotations, and five were no longer pseudogenes and led to eight normal CDSs. These are: DIP0201 (DIP0201 and DIP0201A), DIP0269 (DIP0269), DIP1267 (DIP1267), DIP1523 (DIP1522A and DIP1522B), and DIP2026 (DIP2026). The descriptions of their products are found in Table 3. Furthermore, ten new pseudogenes were detected, which are also shown in Table 3.

It is worth noting that a large part of the features called pseudogenes are probably transposases. In all, there are 56 transposases encoded along the genome. This number, in fact, is a characteristic seen in many species of the *Corynebacterium* genus. The locations of many annotated transposases were in areas flanked by probable PAIs,

Table 3 Coding sequences that ceased to be pseudogenes and new pseudogenes not described in the previous version of the *C. diphtheriae* NCTC13129 genome

New ID	RefSeq ID	Begin	End	Strand	CDS that are no longer pseudogenes	
					Product	**Status**
DIP0201	DIP0201	166488	166832	F	gp1, terminase	No longer a pseudogene
DIP0201A	DIP0201	166822	168423	F	gp2, terminase	No longer a pseudogene
DIP0269	DIP0269	233365	233853	R	Hypothetical protein	No longer a pseudogene
DIP1267	DIP1267	1275251	1275829	F	Putative short chain dehydrogenase	No longer a pseudogene
DIP1522A	DIP1523	1546421	1546561	F	Hypothetical protein	No longer a pseudogene
DIP1522B	DIP1523	1546837	1546971	F	Hypothetical protein	No longer a pseudogene
DIP2026	DIP2026	2079750	2079992	R	Putative transposase for insertion sequence element	No longer a pseudogene
					CDS – new pseudogenes	
DIP0020	DIP0018/DIP0019/DIP0020	20341	21185	R	Glycosidases	New pseudogene
DIP0734	DIP0734/DIP0735	711318	712185	R	Putative sodium/glutamate symporter	New pseudogene
DIP0757	DIP0757/DIP0757A	736001	736487	R	IS element transposase (partial)	New pseudogene
DIP0899	DIP0898/DIP0899	868809	869972	F	HNH endonuclease domain protein	New pseudogene
DIP1095	DIP1095	1077807	1078934	F	Conserved hypothetical integral membrane protein	New pseudogene
DIP1118	DIP1118	1101135	1103803	F	Integral membrane protein, MmpL family	New pseudogene
DIP1177	DIP1177	1175409	1176715	F	Conserved hypothetical protein	New pseudogene
DIP1367	DIP1367	1384058	1384607	R	Transposase of insertion sequence	New pseudogene
DIP1654	DIP1654/DIP1655	1689384	1690688	R	LGFP repeat superfamily protein	New pseudogene
DIP2023	DIP2023/DIP2024	2076614	2077463	F	Filamentation induced by cAMP protein	New pseudogene

Note: RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviations: cAMP, cyclic adenosine monophosphate; CDS, coding sequence; ID, identification; IS, insertion sequence.

reinforcing the idea of the acquisition of islands by lateral transfer. This is because most of the transposases seen in prokaryotic organisms are of exogenous origin.[35] Nevertheless, the transposases perform an important role in the diversification and evolution of bacterial genomes.[36] In the reannotation of *C. diphtheriae*, the insertion of transposases into genes may be noted, interrupting their reading phases, as in the *hsdM* gene, shown in Figure 4. Today, the high plasticity of the *C. diphtheriae* genome is known;[34] perhaps the large number of transposases present have an important role in diversification and could also confirm the increase of such atypical and more virulent strains. Furthermore, these genes are identified inside PAIs of *C. diphtheriae*.

The interruption of the gene cited above is an interesting finding. The *hsd*(*R*, *S*, and *M*) gene encodes type I restriction enzymes, generally with three subunits, involved in the methylation of adenine residues. The subunit encoded by the *hsdS* gene identifies the DNA region to be methylated, and the *hsdM* gene performs the methyltransferase activity. Finally, the *hsdR* gene translocates the *hsdS-M* complex to the target region, even though they are kilobases apart. This mechanism is seen as a preventive action taken by the cell against invasion by foreign DNA, principally by bacteriophages.[37,38]

In *C. diphtheriae*, there is a complete operon with the three genes present: *hsdR-S-M* (DIP2312, DIP2313, and DIP2314, respectively), and the interrupted gene *hsdM* (DIP2081) in another region of the genome. In many populations of *Mycoplasma pulmonis*, the presence of these enzymes was not detected, and even with intact genes, the bacterium is susceptible to infection by bacteriophages. Additionally, analyzing the operon structurally in *C. diphtheriae*, it appeared functional. However, one of its genes, although extra, was interrupted by the insertion of a transposase (Figure 4).

The presence of CRISPR arrays and restriction enzymes inside PAIs raises the question about what extent genes with functions related to immunity against mobile elements may be incorporated from infecting phages or acquired plasmids to avoid coinfection and/or cotransformation by other incoming DNAs, maintaining the DNA balance.

Discussion

Improvement of annotations of specific genes encoding virulence factors

An important approach currently used in prokaryote genomes is data mining to search for genes that may be linked to virulence and pathogenicity pathways and activities.[39] Many characteristics are taken into consideration in this search, such as immunogenicity of the likely products, proteins with adhesive properties, host-pathogen interaction, bacterial dissemination through the host tissues, and proteins without homology to the host. Therefore, following the reannotation of the *C. diphtheriae* genome, a search was made for these gene targets using the Vaxign software.[40] This program is currently used mainly in the search for new candidate genes in the development of vaccines, but the purpose of the present work was to identify the gene targets connected to virulence, pathogenicity, and immunogenic properties. Furthermore, these results can help us understand the reemergence of diphtheria in the world.

The cell wall and extracellular as well as subcellular locations of the proteins were used. The choice was based on the characteristics of these proteins, as they are the first factors to come in contact with the host to promote the dissemination of the microorganism and are frequently highly immunogenic adhesion molecules.[18] The focus of this study was directed at those proteins whose description showed little detail, or even lacked information in the previous version of the genome. Now, such reannotated proteins provide a better description of their function or their activity. A search through the entire genome revealed 23 good extracellular gene targets, as shown in Table 4. Among the secreted proteins, nine were formerly considered as hypothetical proteins, and they now have a better description. Moreover, this mining of the genome for cell wall proteins resulted in 14 suggested proteins and their probable functions, shown in Table 4.

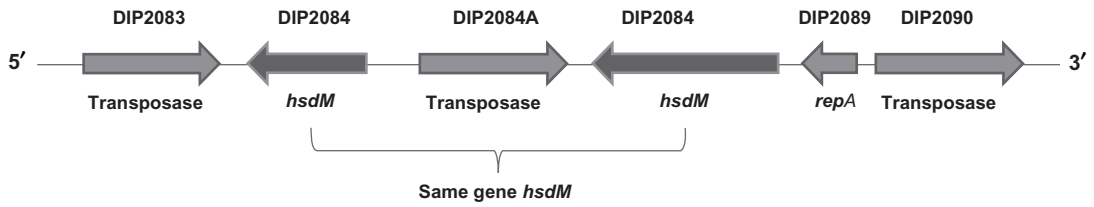

Figure 4 Illustrative schematic of the *hsdM* (DIP_2081) gene, interrupted by the insertion of a transposase.
Notes: Highlighted in dark gray is the *hsdM* gene. The gene was interrupted by the insertion of a transposase (light gray). In addition, the gene is flanked by two more probable transposases (DIP_2080 and DIP_2084). It is possibly a "hotspot" region for the insertion of mobile genetic elements. The interruption of this gene occurs by the addition of the DIP_2082 transposase.

Table 4 New description for *Corynebacterium diphtheriae* genes which have immunological properties and virulence activity (University of Michigan Medical School, Ann Arbor, MI)[27]

Gene ID RefSeq	Product RefSeq	New product	Adhesion probability
DIP0225	Putative secreted polysaccharide deacetylase	Polysaccharide deacetylase	0.125
DIP0298	Putative penicillin-binding secreted protein	Penicillin-binding protein 1B, secreted protein – Pbp1B	0.560
DIP0365	Surface layer protein A	Surface layer protein A	0.577
DIP0543	Putative sialidase precursor	Neuraminidase (sialidase) – NanH	0.200
DIP0554	Putative subtilisin-like cell wall associated serine protease (mycosin)	Subtilisin-like serine protease (mycosin)	0.375
DIP0559	ESAT-6-like protein	ESAT-6-like protein – EsxT	0.521
DIP0640	Hypothetical protein DIP0640	NPL/P60-family secreted protein	0.631
DIP0793	Hypothetical protein DIP0793	Putative twin-arginine translocation pathway signal protein	0.513
DIP0836	Hypothetical protein DIP0836	Putative secreted metalloendopeptidase – MepA	0.825
DIP1097	Putative low molecular weight protein antigen 6	Putative low molecular weight protein antigen 6	0.146
DIP1281	Putative invasion protein	Resuscitation-promoting factor interacting protein – RpfI	0.527
DIP1621	Hypothetical protein DIP1621	NlpC/P60 family protein	0.465
DIP1622	Hypothetical protein DIP1622	NlpC/P60 family protein	0.556
DIP1701	Putative ribonuclease	Guanyl-specific ribonuclease Sa3	0.577
DIP2034	Putative transposase	Transposase, mutator family	0.000
DIP2193	Putative secreted antigen	Trehalose corynomycolyl transferase C – CmtC	0.398
DIP2194	Putative secreted antigen	Trehalose corynomycolyl transferase B – CmtB	0.424
DIP2294	Putative penicillin-binding protein	Penicillin-binding protein C – PbpC	0.741
DIP2339	Putative major secreted protein	Trehalose corynomycolyl transferase A – CmtA	0.346
Cell wall proteins: genetics target of *C. diphtheriae* vaccine			
DIP0139	Hypothetical protein DIP0139	Conserved hypothetical protein	0.669
DIP0235	Putative fimbrial subunit	Putative fimbrial protein	0.739
DIP0237	Putative surface-anchored protein	Surface-anchored protein (fimbrial subunit) – SpaE	0.517
DIP0238	Putative surface-anchored fimbrial subunit	Surface-anchored protein (fimbrial subunit) – SpaF	0.809
DIP0515	Putative transport system secreted protein	ABC-type dipeptide/oligopeptide/nickel transport system – OppA	0.549
DIP0956	Putative peptide transport system secreted protein	ABC transporter solute-binding protein – OppA1	0.502
DIP1740	ABC transporter solute-binding protein	ABC transporter solute-binding protein	0.405
DIP2010	Putative surface-anchored membrane protein	Possible surface-anchored membrane protein	0.565
DIP2013	Putative surface-anchored fimbrial subunit	Putative surface-anchored fimbrial subunit	0.559
DIP2066	Putative surface-anchored fimbrial associated protein	Putative surface-anchored fimbrial associated protein	0.843
DIP2093	Sdr family related adhesin	Putative Sdr-family related adhesin	0.818
DIP2162	ABC transporter solute-binding protein	Periplasmic binding protein-like II	0.221
DIP2226	Surface-anchored fimbrial subunit	Surface-anchored fimbrial subunit – SpaH	0.741
DIP2227	Surface-anchored fimbrial subunit	Surface-anchored fimbrial subunit – SpaG	0.816

Note: RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviation: ID, identification.

An important characteristic noted was the presence of classical virulence factors such as adhesins (DIP2093), fimbrial subunits (DIP0235, DIP0237, DIP0238, DIP2013, DIP2066, DIP2226, and DIP2227), and adenosine triphosphate-binding cassette transporter proteins connected to the transport of solutes (DIP0515, DIP0956, and DIP1740). These proteins, which are located in the cell wall fraction, were analyzed in silico (Table 4). The knowledge of these elements that promote interaction with the host is vast.[40] As they are some of the first elements to make contact with the host cell, and show no counterpart in the host, they generally are the targets that are used in the development of vaccines. Dealing with the classical elements, which are already well described in the literature and in databases, the reannotation did not result in significant changes in these proteins.

However, the extracellular portion underwent a significant change in the functional annotation. There is a large number of proteins connected to the membrane, such as polysaccharide deacetylase (DIP0225), NlpC/60 protein families (DIP1621 and DIP1622), penicillin-binding

proteins (DIP0298 and DIP2294), and unique proteins of actinobacteria, for example, subtilisin-like serine protease (mycosin, DIP0554) with high proteolytic capacity and structural proteins, including trehalose corynomycolyl transferase (DIP2193, DIP2194, and DIP2339).[41] Most of these proteins, connected to the extracellular part of the membrane, show enormous capacity to promote recognition and immune response in the host. It is because of this characteristic that they are good candidates for the development of more effective therapies.

Additionally, one protein on the list shown in Table 4 deserves highlighting. In dealing with a human pathogen such as *C. diphtheriae*, having the ability to colonize the mucosa, the presence of the neuraminidase (sialidase) gene (DIP0543) confers an extra ability to use solutes present only in animal host cells, such as sialic acid. Some pathogens have the capacity to use this sugar as a source of carbon and thereby possess an extra mechanism for surviving within the cell, in a hostile environment.[42]

In addition, the use of this compound can interfere with the defense system of the host by diminishing the viscosity of the mucus and diminishing the activity of inflammatory cells. A rapid and sensitive assay for neuraminidase using peanut lectin hemagglutination was used to study the prevalence of neuraminidase activity among sucrose-fermenting and nonsucrose-fermenting toxigenic *C. diphtheriae* strains. Neuraminidase activity was found in all isolates regardless of biotype, hemagglutinating activity, and site of isolation of bacteria. Besides expressing neuraminidase activity that hydrolyzes sialic acid from glycoconjugates, *C. diphtheriae* was also capable of transferring sialic acid residues from a sialyl-lactose donor. A single molecule probably expresses both neuraminidase and trans-sialidase activity. The trans-sialidase activity was documented by observations of the interactions of bacterial cells with wheat germ agglutinin and peanut lectins. *C. diphtheriae* expressed a trans-sialidase activity located on the cell surface that produced asialoglycoconjugates from a sialyl donor substrate and at the same time generated bacterial sialyl derivatives of beta-galactosidase acceptors.[43] Therefore, the action of this protein can be a strong indication of the ability of *C. diphtheriae* to colonize the mucosa of human airways, escaping from the human immune system and causing disease.

Aside from proposing various genes that may be broadly used as targets in therapy studies, the findings presented here show a bit of the virulence and the pathogenicity of this reemerging and diversifying pathogen, now in a more detailed way.

Description of other operons in *C. diphtheriae* PAIs

Additionally, other genes linked to virulence could be described. Several operons inside the PAIs of *C. diphtheriae* had been assigned a gene name on the genome information, such as *cyd*, *dha*, and *pdx* operons. The *cydABCD* operon codes for an oxygen-scavenging enzyme (cytochrome d) which is reported to be elevated in situations where oxygen is a limiting factor for bacterial growth. Besides, cytochrome d has several different roles in bacteria, including scavenging oxygen that could inactivate oxygen-sensitive nitrogenases, contributing to energy conservation under microaerobiosis and protecting bacteria from oxidative stress.[44,45]

The *dha* operon (PiCd 24) pertains to a family of enzymes that utilize phosphate donors, such as adenosine triphosphate or phosphoproteins, to phosphorylate a toxic compound formed during glycerol metabolism (dihydroxyacetone) into a nontoxic compound (dihydroxyacetone phosphate).[46,47]

Finally, the *pdx* operon (PiCd2) is composed of the genes *pdxS* and *pdxT,* which code for pyridoxal 59-phosphate synthase. Pyridoxal 59-phosphate synthase is regulated by another gene of the *pdx* operon, *pdxR*, and plays an important role in the de novo biosynthesis of vitamin B6, which, in turn, is an essential cofactor for several enzymes catalyzing a variety of biochemical reactions.[48]

Conclusion

C. diphtheriae genome annotation underwent alteration in 57% of its contents, after reannotation. The entire approach was manual, following protocols already established in the literature for the functional annotation of genomes. Updating genomes already available in databases, in addition to supporting research groups performing experiments using the genomic data, also helps in the minimization of annotation errors.

The reannotation resulted in the discovery of new genes in the *C. diphtheriae* genome sequence, correction of ORF strands, and improvement of the functional description of the genome, including classical virulence genes. In addition, it assisted in the search for new gene targets for the development of more effective therapies, information hitherto unpublished in the literature. Nevertheless, the improvement in the description of the proteins linked to the different bacterial defense mechanisms present in the genome, besides providing knowledge of how *C. diphtheriae* may respond to invasion by mobile genetic material, provides indications about its plasticity and the modulation of the genome.

Finally, the protocol used in the present genome can be applied to any genome, whether already on file or not,

aimed at the improvement, accuracy of the annotation, and the search for virulence and pathogenicity genes of microorganisms.

Authors' contributions

SCS, VD, AA, ARS, ACP, AACM, CJF, EB, LCG, ME, SSA, VACA, AZN, ARC, LTC, RTJR were involved in all of: predictions of genes, transfer RNA, ribosomal RNA, and conserved domains of proteins; in similarity searches of *C. diphtheriae* genomes against several databases; and, in the functional reannotation. SCS performed the PAIs analysis. ARS located the subcellular proteins in the genome. VD was responsible for searching new genetic targets for the development of vaccines. VA and AM coordinated and participated in the conception, design, and supervision of the whole project. VA, AM, ALMG, RHJ, SCS, VD, ET, AT, AS, and MPS were involved in writing the manuscript.

Acknowledgments

Funding was provided by CNPq (Conselho Nacional de Desenvolvimento Científico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Médigue C, Moszer I. Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol*. 2007;158(10):724–736.
2. Salzberg SL. Genome re-annotation: a wiki solution? *Genome Biol*. 2007;8(1):102.
3. Petty NK. Genome annotation: man versus machine. *Nature*. 2010; 8(11):762.
4. Boneca IG, de Reuse H, Epinat J, Pupin M, Labigne A, Moszer I. A revised annotation and comparative analysis of Helicobacter pylori genomes. *Nucleic Acids Res*. 2003;31(6):1704–1714.
5. Luo C, Hu GQ, Zhu H. Genome reannotation of Escherichia coli CFT073 with news insights into virulence. *BMC Genomics*. 2009;10:522.
6. Gundogdu O, Bentley SD, Holden MT, Parkhill J, Dorrell N, Wren BW. Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics*. 2007;8:162.
7. Camus JC, Pryor MJ, Médigue C, Cole ST. Re-annotation of genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*. 2002; 148(Pt 10):2967–2973.
8. Dandekar T, Huynen M, Regula JT, et al. Re-annotating the Mycoplasma pneumonia genome sequence: adding value, function and reading frames. *Nucleic Acids Res*. 2000;28(17):3278–3288.
9. Cerdeño-Tárraga AM, Efstratiou A, Dover LG, et al. The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. *Nucleic Acids Res*. 2003;31(22):6516–6523.
10. Rappuoli R, Podda A, Giovannoni F, Nencioni F, Peragallo L, Francolini P. Absence of protective immunity against diphtheria in a large proportion of young adults. *Vaccine*. 1993;11(5):576–577.
11. Pimenta FP, Hirata R Jr, Rosa AC, Milagres LG, Mattos-Guaraldi AL. A multiplex PCR assay for simultaneous detection of Corynebacterium diphtheriae and differentiation between non-toxigenic and toxigenic isolates. *J Med Microbiol*. 2008;57(Pt 11):1438–1439.
12. Gomes DL, Martins CA, Faria LM, et al. Corynebacterium diphtheriae as an emerging pathogen in nephrostomy catheter-related infection: evaluation of traits associated with bacterial virulence. *J Med Microbiol*. 2009;58(Pt 11):1419–1427.
13. Lagesen K, Hallin P, Rødland EA, Staerfeldt H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007;35(9):3100–3108.
14. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25(5):955–964.
15. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573–580.
16. Zdobnov EM, Apweiler R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001;17(9):847–848.
17. Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. *Bioinformatics*. 2000;16(10):944–945.
18. He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and an application for vaccine development. *J Biomed Biotechnol*. 2010;2010:297505.
19. Waack S, Keller O, Asper R, et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*. 2006;7:142.
20. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39:W29–W37.
21. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. *Bioinformatics*. 2005;21(16):3422–3423.
22. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T. MvirDB – a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res*. 2007;35:D391–D394.
23. Karp PD, Paley S, Romero P. The Pathway Tools software. *Bioinformatics*. 2002;18 Suppl 1:S225–S232.
24. Caspi R, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2008;36: D623–D631.
25. Caspi R, Karp PD. Using the MetaCyc pathway database and the BioCyc database collection. *Curr Protoc Bioinformatics*. 2007;Chapter 1 (Unit 1):17.
26. Bubeck Wardenburg J, Williams WA, Missiakas D. Host defenses against Staphylococcus aureus infection require recognition of bacterial lipoproteins. *Proc Natl Acad Sci U S A*. 2006;103(37): 13831–13836.
27. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*. 2007;8:172.
28. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*. 2010;11(3): 181–190.
29. Zegans ME, Wagner JC, Cady KC, Murphy DM, Hammond JH, O'Toole GA. Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa. *J Bacteriol*. 2009;191(1):210–219.
30. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*. 2002;43(6):1565–1575.
31. Mojica FJ, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*. 2005;60(2):174–182.
32. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*. 2005;151(Pt 3):653–663.
33. Palmer KL, Gilmore MS. Multidrug-resistant enterococci lack CRISPR-cas. *MBio*. 2010;1(4):e00227–e00310.

34. Mokrousov I. Corynebacterium diphtheriae: genome diversity, population structure and genotyping perspectives. *Infect Genet Evol*. 2009;91(1):–15.
35. Mes TH, Doeleman M. Positive selection on transposase genes of insertion sequences in the Crocosphaera watsonii genome. *J Bacteriol*. 2006;188(20):7176–7185.
36. Ooka T, Ogura Y, Asadulghani MD, et al. Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in Escherichia coli O157 genomes. *Genome Res*. 2009;19(10):1809–1816.
37. Dybvig K, Sitaraman R, French CT. A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc Natl Acad Sci U S A*. 1998;95(23): 13923–13928.
38. Obarska-Kosinska A, Taylor JE, Callow P, Orlowski J, Bujnicki JM, Kneale GG. HsdR subunit of the type I restriction-modification enzyme EcoR124I: biophysical characterisation and structural modelling. *J Mol Biol*. 2008;376(2):438–452.
39. Rappuoli R. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*. 2001;19(17–19):2688–2691.
40. Webb SA, Kahler C. Bench-to-bedside review: bacterial virulence and subversion of host defences. *Crit Care*. 2008;12(6):234–242.
41. Ramulu HG, Adindla S, Guruprasad L. Analysis and modeling of mycolyl-transferases in the CMN group. *Bioinformation*. 2006;1(5): 161–169.
42. Lichtensteiger CA, Vimr ER. Neuraminidase (sialidase) activity of Haemophilus parasuis. *FEMS Microbiol Lett*. 1997;152(2):269–274.
43. Mattos-Guaraldi AL, Formiga LC, Andrade AF. Trans-sialidase activity for sialic acid incorporation on Corynebacterium diphtheriae. *FEMS Microbiol Lett*. 1998;168(2):167–172.
44. Iuchi S, Chepuri V, Fu HA, Gennis RB, Lin EC. Requirement for terminal cytochromes in generation of the aerobic signal for the arc regulatory system in Escherichia coli: study utilizing deletions and lac fusions of cyo and cyd. *J Bacteriol*. 1990;172(10):6020–6025.
45. Winstedt L, Yoshida K, Fujita Y, von Wachenfeldt C. Cytochrome bd biosynthesis in Bacillus subtilis: characterization of the cydABCD operon. *J Bacteriol*. 1998;180(24):6571–6580.
46. Bächler C, Schneider P, Bähler P, Lustig A, Erni B. Escherichia coli dihydroxyacetone kinase controls gene expression by binding to transcription factor DhaR. *EMBO J*. 2005;24(2):283–293.
47. Bizzini A, Zhao C, Budin-Verneuil A, et al. Glycerol is metabolized in a complex and strain-dependent manner in Enterococcus faecalis. *J Bacteriol*. 2010;192(3):779–785.
48. Jochmann N, Götker S, Tauch A. Positive transcriptional control of the pyridoxal phosphate biosynthesis genes pdxST by the MocR-type regulator PdxR of Corynebacterium glutamicum ATCC 13032. *Microbiology*. 2011;157(Pt 1):77–88.

Additional File 1 Similarity analyses between new coding sequences and pseudogenes against the nonredundant database of proteins from National Center for Biotechnology Information

Abbreviations: ABC, adenosine triphosphate-binding cassette; ACP, acyl carrier protein; ATP, adenosine triphosphate; CoA, Coenzyme A; CRISPR, clustered regularly interspaced palindromic repeats; DNA, deoxyribonucleic acid; GMP, guanosine monophosphate; GTP, guanosine triphosphate; IS, insertion sequence; MFS, major facilitator superfamily; NAD, nicotinamide adenine dinucleotide, NADH, reduced form of NAD; NAD(P)H, NAD phosphate; PAI, pathogenicity island; RNA, ribonucleic acid; rRNA, ribosomal RNA; SNARE, soluble N-ethylmaleimide-sensitive factor attachment protein receptor; tRNA, transfer RNA.

Additional File 2 Coding sequences of pathogenicity islands predicted by the software PIPS (http://www.genoma.ufpa.br/lgcm/pips) in the reannotated Corynebacterium diphtheriae NCTC13129 genome

Abbreviation: ABC, adenosine triphosphate-binding cassette.