<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Genome mining for radical SAM protein determinants reveals multiple sactibiotic-like gene clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Murphy, Kiera; O'Sullivan, Orla; Rea, Mary C.; Cotter, Paul D.; Ross, R. Paul; Hill, Colin</td>
</tr>
<tr>
<td><strong>Editor(s)</strong></td>
<td>Read, Timothy D.</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2011-07-08</td>
</tr>
<tr>
<td><strong>Type of publication</strong></td>
<td>Article (peer-reviewed)</td>
</tr>
</tbody>
</table>
| **Link to publisher's version** | http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0020852  
http://dx.doi.org/10.1371/journal.pone.0020852  
Access to the full text of the published version may require a subscription. |
| **Rights** | © 2011 Murphy et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. http://creativecommons.org/licenses/by/2.5/ |
| **Item downloaded from** | http://hdl.handle.net/10468/824 |

Downloaded on 2022-06-14T11:49:53Z
Genome Mining for Radical SAM Protein Determinants Reveals Multiple Sactibiotic-Like Gene Clusters

Kiera Murphy1,2, Orla O’Sullivan2*, Mary C. Rea2,3, Paul D. Cotter2,3, R. Paul Ross2,3, Colin Hill1,3

1 Microbiology Department, University College Cork, Cork, Ireland, 2 Teagasc Food Research Centre, Moorepark, Fermoy, County Cork, Ireland, 3 Alimentary Pharmabiotic Centre, Cork, Ireland

Abstract
Thuricin CD is a two-component bacteriocin produced by Bacillus thuringiensis that kills a wide range of clinically significant Clostridium difficile. This bacteriocin has recently been characterized and consists of two distinct peptides, Trnβ and Trnα, which both possess 3 intrapeptide sulphur to α-carbon bridges and act synergistically. Indeed, thuricin CD and subtilosin A are the only antimicrobials known to possess these unusual structures and are known as the sactibiotics (sulphur to alpha carbon-containing antibiotics). Analysis of the thuricin CD-associated gene cluster revealed the presence of genes encoding two highly unusual SAM proteins (TrnC and TrnD) which are proposed to be responsible for these unusual post-translational modifications. On the basis of the frequently high conservation among enzymes responsible for the post-translational modification of specific antimicrobials, we performed an in silico screen for novel thuricin CD–like gene clusters using the TrnC and TrnD radical SAM proteins as driver sequences to perform an initial homology search against the complete non-redundant database. Fifteen novel thuricin CD–like gene clusters were identified, based on the presence of TrnC and TrnD homologues in the context of neighbouring genes encoding potential bacteriocin structural peptides. Moreover, metagenomic analysis revealed that TrnC or TrnD homologs are present in a variety of metagenomic environments, suggesting a widespread distribution of thuricin-like operons in a variety of environments. In-silico analysis of radical SAM proteins is sufficient to identify novel putative sactibiotic clusters.

Introduction
The ever-increasing issue of bacterial resistance to conventional antibiotics and consumer demands for safe minimally processed foods have stimulated research interest in natural antimicrobial agents such as bacteriocins [1,2,3]. A variety of approaches can be employed to identify novel bacteriocins with the most traditional being the isolation of strains and a culture-based assessment of their ability to produce novel antimicrobials which have a broad antimicrobial spectra. In contrast, the discovery of thuricin CD was the result of the specific mining of the gut microbiota for potentially therapeutic bacteriocins with a narrow spectrum of activity [4]. This alternative approach stems from the fact that in the clinical setting the use of broad spectrum antimicrobials has lead to the emergence of multi-drug resistant pathogens and the observation that the disruption of the colonic microbiota through administration of broad spectrum antibiotic treatment predisposes susceptible individuals to infection by the nosocomial pathogen Clostridium difficile [5,6]. Thuricin CD, produced by the human faecal isolate Bacillus thuringiensis DPC 6431, is a two peptide antimicrobial, composed of Trnβ and Trnα. The two thuricin CD peptides act synergistically, exhibiting a narrow spectrum of activity targeting mainly the spore forming genera - Bacillus and Clostridium. Interestingly thuricin CD kills a wide range of clinical C. difficile ribotypes commonly associated with Clostridium difficile associated diarrhoea (CDAD) including the hypervirulent strain B1/NAP1/027 [7]. In a distal colon model thuricin CD, while killing C. difficile, had no significant impact on the composition of the microbiota unlike other broad spectrum antimicrobials [8]. However to be used as an oral therapeutic for CDAD thuricin CD, because it is sensitive to proteolytic enzymes, will require the use of encapsulation technologies to ensure the delivery of biologically active peptides to the colon.

Both Trnα and Trnβ are post-translationally modified peptides which contain unusual sulphur to α-carbon linkages, and as a result thuricin CD has been classified as a sactibiotic (sulphur to alpha carbon antibiotic) [9]. The thuricin CD-associated gene cluster has been identified and contains genes, trnC and trnD, predicted to encode two radical S’-adenosylmethionine (SAM) proteins. Radical SAM proteins are characteristically associated with unusual posttranslational modifications and are likely to be involved in the formation of the aforementioned cysteine to α-carbon linkages [10]. Radical SAM-encoding genes are rare in bacteriocin-associated clusters, with subtilosin A and propionin F being notable exceptions [11,12]. Propionin F is not a modified peptide and in this instance it is thought that the associated radical SAM enzyme contributes to N-terminal cleavage [12]. However, subtilosin A is a cyclic peptide which undergoes posttranslational modification involving a thiol linkage of sulphur to α-carbon and can therefore be classified as a sactibiotic. The mechanism
associated with the formation of this unusual posttranslational modification has not been established but is thought to be due to the associated radical SAM enzyme, AlbA.

An alternative approach to the identification of novel posttranslationally modified peptides has involved in silico screening to identify the most highly conserved component of the associated gene clusters i.e. the genes encoding the modification enzymes. This approach has been employed, for example, to identify novel type I and type II lantibiotics [13,14,15,16] and thiazole/oxazole-modified microcin (OMICS; [17,18,19,20,21]). Here this approach has been adapted to identify novel thuricin CD-like gene clusters. More specifically, an in silico screen of databases was undertaken, using TrnC and TrnD as driver sequences, to identify similar gene clusters. The search revealed 100 TrnC homologs and 53 TrnD homologs, which upon further investigation led to the identification of 15 novel putative thuricin CD-like clusters, i.e. clusters containing genes encoding TrnC and TrnD homologs as well as at least one structural peptide, across three phyla.

Results

In silico screen for TrnC and TrnD Protein

The sequences of TrnC and TrnD, radical SAM proteins encoded within the thuricin CD gene cluster [4], were used as driver sequences to perform a homology search against the complete non-redundant database (November 2010). The search revealed 100 TrnC homologs and 54 TrnD homologs (homology being defined as BLASTP E-value < 10^-6), in 112 unique genomes. Tables S1 and S2 provide a full list of TrnC and TrnD homologs, which were found in the genomes of sequenced microorganisms. In each case the corresponding bacterial genomes were viewed and the regions containing the genes of interest were located. None of the strains identified have previously been associated with a bacteriocin producing phenotype.

A more detailed examination of these gene clusters involved an analysis of the regions flanking the trnC and trnD determinants for other open reading frames (ORFs) potentially involved in the biosynthesis of, or immunity to, bacteriocins. The presence of a second nearby Radical SAM gene (including those encoding another TrnC or TrnD homolog) was screened for in each case. Putative Radical SAM proteins were examined for the presence of signature motifs. Enzymes of the Radical SAM superfamily generate radicals by combining a 4Fe-4S cluster and S-adenosylmethionine (SAM) in close proximity. The Radical SAM signature motif, C-X_a-C-X_b-C, which coordinates the iron-sulphur and SAM in close proximity was screened for, as were the putative SAM binding sequence motifs, GGEPLL and TNG.

Particular focus was placed on screening the aforementioned clusters for genes potentially encoding structural peptides which, given the variable nature of the structural peptides encoded within modified bacteriocin associated gene clusters, may or may not resemble the very similar TrnA and TrnB peptides. One feature that was screened for was the presence of a conserved double-glycine (GG) motif, which is a common feature among bacteriocins that are synthesized as biologically inactive precursor peptides (prepeptides) and plays a key role in secretion and activation [22,23]. Interestingly, in thuricin CD cleavage occurs between the GG residues rather than after which is the norm. The TrnA and TrnB peptides consist of a cysteine-less leader region characterised by a high frequency of Glu and Asp residues [4]. Both genes are preceded by perfect ribosomal binding sequences (RBS) identical to that of the Shine-Dalgarno consensus that are embedded within identical stretches of nucleotides (AAAAA-TAAGGGAGAATCT). Also screened for was the presence of conserved cysteine residues such as those that occur at positions +5, +9 and +13 in TrnA and TrnB, forming a CX3CX3C formation. Cys +5 is flanked by small hydrophobic amino acids Ala and Val, and those at positions +9 and +13 are coupled with variant small hydrophobic residues. A conserved Ser residue is present at position +13. Additionally, Gly residues occur at positions +17 and +19, which are followed by variant hydrophobic and hydrophilic residues, respectively [4].

Sequencing had revealed that TrnA and TrnB are posttranslationally modified at their respective 21st, 25th and 29th residues [4]. It is believed that each peptide has three sulphur to α-carbon thioether linkages between the cysteine side chains and the α-carbons of the modified residues. These unusual sulphur to α-carbon linkages have also been reported in the antimicrobial subtilosin A [24]. Notably, a SAM-containing enzyme is thought to be responsible for forming the thioether linkages in subtilosin A [24]. This protein bears some resemblance to the SAM-containing enzymes encoded within the thuricin CD operon [4]. Other features that were specifically screened for were the presence of other genes predicted to encode bacteriocin associated proteins, including those involved in transport and processing.

Following this screen, 15 gene clusters from 15 different bacterial strains were deemed worthy of closer analysis. Of these, 10 were gene clusters that contained both TrnC and TrnD encoding genes with putative prepeptide-encoding genes. The remaining 5 contained either a TrnC or a TrnD encoding gene but were deemed noteworthy as the majority of the other thuricin CD operon-associated features were present. Thirteen of these 15 total gene clusters were in the genomes of strains within the phylum Firmicutes. The other 2 strains were Bacteroides sp. 3_1_19 of the phylum Bacteroidetes and Petrotoga mobilis SJ95 of the phylum Thermotogae. These 15 clusters are described below and are grouped according to the phylum and genus of the associated strain. Arrow diagrams representing each of these clusters are depicted in Figure 1.

Gene clusters in Firmicutes

Identification of novel Bacillus-associated bacteriocin gene clusters. Bacillus is a large and diverse genus which contains both free living and pathogenic species. Members are Gram-positive, rod-shaped, sporulating and are known to produce a wide range of antimicrobial substances including bacteriocins. Many of the Bacillus bacteriocins are lantibiotics, a category of post-translationally modified peptides which are grouped as a consequence of the presence of [B-methyl] lanthionine structures. Examples include subtilin [25], ericin S and ericin A [26], haloderacin [14,16] and lichenicidin [13]. Our in silico screen highlighted the presence of thuricin CD-like gene clusters in the genomes of three Bacillus strains, i.e. B. cereus 95/8201, B. thuringiensis serovar pulsans BGSC 4CC1 and B. thuringiensis serovar huazhangensis BGSC 4BD1.

B. cereus 95/8201. Strains of B. cereus are frequently isolated from soil but also found in foods of plant and animal origin. B. cereus is an opportunistic pathogen and has frequently been associated with human food poisoning [27].

B. cereus strain 95/8201 (GenBank accession ACHF00000000) was isolated from a human patient with endocarditis (Bacillus cereus group Taurasse-Helgason MLST Database) and sequenced by the Naval Medical Research Centre in 2009. A gene cluster was identified in this strain which is very similar to the thuricin CD gene cluster in B. thuringiensis DPC 6431. Bcere0016_53380 (499 residues) resembles TrnC (506 amino acids). The C-X_a-C- X_b-C motif occurs as CNLRCYEYC at Cys107 in Bcere0016_53380 and...
is identical to the corresponding region in TrnC. Putative SAM binding sequence motifs, GGEPLL (identified at Gly165) and TNG (identified at Thr195) are also conserved between Bcere0016_53390 and TrnC. Bcere0016_53390 (408 amino acids) resembles TrnD (414 residues). The signature motif occurs as CMNMCKHC (at Cys103 in Bcere0016_53390) in both. The GGEPLL (Gly150) and TNG (Thr175) motifs in Bcere0016_53390 are identical to those in TrnD. Bcere0016_53360 and Bcere0016_53370 are the predicted structural genes and are 100% identical to B. thuringiensis DPC 6431 structural genes, triß and trn respectively. Other genes of note in this cluster are those encoding a putative ABC transporter (Bcere0016_53350) and a C-terminal protease (Bcere0016_53400). The ABC transporter may be responsible for the export of the bacteriocin from the cytoplasm and the peptide may have a role in immunity or leader cleavage.

**B. thuringiensis serovar pulsiensis** BGSC 4CC1. *B. thuringiensis* are noted for their ability to produce a parasporal crystal upon sporulation. Due to its insecticidal properties it has been used commercially as a biological insecticide [28]. *B. thuringiensis serovar pulsiensis* BGSC 4CC1 (Accession NZ_ACN00000000) was sequenced by the Naval Medical Research Centre in 2009. Within this genome, putative TrnC (Bthur0012_58010) and TrnD (Bthur0012_58000) encoding genes were identified which are surrounded by other genes associated with bacteriocin synthesis. The TrnC homolog (e-value 3e-30) encoded by Bthur0012_58010 contains the GGEPLL motif at Gly35 while the TrnD homolog (e-value 8e-23) encoded by Bthur0012_58000 has a GGDPLL at Gly35 which is surrounded by other genes associated with bacteriocin synthesis. The TrnD operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recognise or alternatively cleavage in this operon we have failed to recogni...
C-X$_3$-C-X$_2$-C motif, namely CNLRCKYC, is identified at Cys104. Interestingly, the CNLRCKYC sequence was also observed in Calrk_0222, a TrnC homolog in *C. brisbanensis* 177R1B. Cthe_0907 is believed to encode a structural gene. It is 46 amino acids long, contains a GG motif at Gly19 and a C-X$_3$-C-X$_2$-C motif (CQTSCQSC) at Cys24. Within this cluster other genes potentially of relevance include *secD*, a gene involved in protein export (Cthe_0904) and also a signal peptidase-encoding gene (Cthe_0909).

**C. difficile QCD-63q42.** *C. difficile* is the leading cause of nosocomial infections associated with antibiotic use and is responsible for substantial morbidity and mortality worldwide [34]. *C. difficile* QCD-63q42 is one of three *C. difficile* strains in which thuricin CD-like gene clusters were identified in this study. *C. difficile* QCD-63q42 (Accession NZ_ABHD00000000) was isolated from a 67 year old male patient with severe *C. difficile* Associated Disease (CDAD) and was sequenced in 2007 by McGill University. Like the aforementioned *C. thermocellum* ATCC 27405 gene cluster, the QCD-63q42 gene cluster is lacking a TrnD homolog and only has one structural gene but was included for the same reasons. The TrnC homolog encoded by CdiQCD-6_020200015426 (e-value 4e-21) contains 475 amino acids, a GGEPLL motif at Gly148 and the C-X$_3$-C-X$_2$-C motif (CNLHCDYC) at Cys97. The proposed structural peptide, encoded by CdiQCD-6_020200015431, is larger than those previously observed (79 amino acids) but does contain two GG motifs, at Gly 21 and Gly 39, and also a C-X$_3$-C-X$_2$-C motif, CGALCANLC, at Cys25. Nearby genes include those encoding a hydrolase (CdiQCD-6_020200015426) and an ABC transporter (CdiQCD-6_020200015451).

**C. difficile 630.** The complete genome sequence of *C. difficile* strain 630 (Accession NC_009069) was sequenced in 2006 by the Sanger Institute [34]. The strain was isolated from a patient with severe pseudomembranous colitis in Switzerland and is a virulent, highly transmissible, drug resistant strain. CD0162 is a homolog of both TrnC and TrnD (e-values of 5e-30 and 2e-0, respectively). While the adjacent CD0163 was annotated as hypothetical, alignment of its sequence with TrnD confirmed it is a TrnD homolog (e-value 1e-12). Both CD0162 (473 amino acids) and CD0163 (356 amino acids) lack the signature C-X$_3$-C-X$_2$-C motif but both possess a GG motif.

CD0163B is thought to encode a potential structural gene of 67 amino acids containing a GG motif at Gly47 and a C-X$_3$-C-X$_2$-C motif (CTIMCPYTC) at Cys20. This gene cluster also contains genes that code for an ABC transporter (CD0161) and a two-component response regulator (CD0159 and CD0160).

**C. difficile QCD-32g58.** *C. difficile* QCD-32g58 (Accession NZ_AAML00000000) was sequenced in 2007 by Washington Figure 2. Protein sequence alignment of TrnD and all the TrnD homologs identified in this study.

doi:10.1371/journal.pone.0020852.g002

Genome Mining for Sactibiotic-Like Gene Clusters
University. This hyper-virulent strain was responsible for a multi-institutional outbreak of CDAD in Quebec, Canada [35]. Analysis of this genome revealed the presence of a TrnC homolog-encoding gene, CdiQ_040500000917 (e-value 6e-30). The corresponding protein contains a GGEPLL motif at Gly153 and a C-X3-C-X3-C motif, CNLRCDYC. As is the case with some of the other Clostridia-associated clusters strains examined, an adjacent gene, CdiQ_040500000922, was found to encode a protein (356 amino acids) that does not significantly resemble TrnD (e-value 4.5). A screen for potential structural genes revealed the presence of three adjacent genes of relevant size. Upon further analysis, the 67 amino acid CdiQ_040500000932 was determined to be most likely to be the structural peptide-encoding gene as its product is predicted to contain a GG motif at Gly47 and a C-X3-C-X3-C motif, CTIMCPYTC at Cys20. Surrounding genes of note include those encoding an ABC transporter (CdiQ_040500000912) and a response regulator (CdiQ_040500000907).

**Anaerococcus tetradus ATCC 35098.** *A. tetradus* is a strictly anaerobic bacterium, usually found in water and hot springs. Lantibiotics have previously been associated with related microbes, such as ruminococcin A produced by *Ruminococcus gnavus* [36]. *A. tetradus* ATCC 35098 (Accession ACGC000000000) was sequenced in 2009 by Baylor College of Medicine. The gene cluster of interest within this genome may contain as many as 11 genes, including a number of apparently bacteriocin-associated motifs. BLAST searches revealed that the proteins encoded by HMPREF0077_1416 (e value 3e-17) and HMPREF0077_1424 (e value 3e-23) are TrnC homologs, HMPREF0077_1416 contains 728 amino acids, a GGEPLL motif at Gly79, a TNG at Thr117 and a C-X3-C-X3-C motif, CNFSCLHC, at Cys18 but lacked GGEPLL and TNG motifs. A sequence alignment with TrnD showed a good agreement and gave an e value of 4e-4. While this is outside the threshold set at e-value <10^-6, its presence remains noteworthy. In addition to the predicted radical SAM proteins, a hypothetical protein associated with this cluster shares homology with a methylase (HMPREF0077_1423). Methylases contribute to the production of a number of non-ribosomal antibiotics [37]. HMPREF0077_1413 and HMPREF0077_1414 encode homologs of the SaliRK two component signal transduction system associated with the production of the lanthionine salvicin A2, produced by *Streptococcus salivarius* K12 [38]. In close proximity are genes encoding a response regulator (HMPREF0077_1414) and two ABC transporters (HMPREF0077_1418 and HMPREF0077_1426).

**A. prevotii DSM 20548.** *A. prevotii* is an anaerobic, mesophilic, non-motile, non-sporulating bacterium. It is an opportunistic pathogen found in the microflora of the mouth, skin or vagina of the host [39]. Strain DSM 20548 (Accession CP001709) was isolated from a human plasma sample and was sequenced in 2009 by the US DOE Joint Genome Institute and contains a putative bacteriocin-gene cluster. The *T. mathranii A3* genome contains two radical SAM-encoding genes, the TrnC homolog-encoding Tmath_1971 (e value 3e-10) and the TrnD homolog-encoding Tmath_1976 (e value 3e-6). Tmath_1971 contains 461 residues, a GGEPLL at Gly153 and interestingly a C-X3-C-X3-C motif, namely CNLRCKYC, at Cys104, which has been observed in other TrnC homologs previously. Tmath_1976 is 453 amino acids long and contains a GGEPLL motif at Gly151, a TNG at Thr185 and a C-X3-C-X3-C motif, CNLKCCKYC, at Cys101. A putative structural gene, Tmath_1977, was also identified. The corresponding peptide is predicted to contain 32 residues, a GG motif at Gly14 and two C-terminal cysteine residues. Other surrounding genes of note include those encoding two ABC transporters (Tmath_1974 and Tmath_1975) and *secD*, a gene involved in protein export (Tmath_1973).

**Caldicellulosiruptor kristjanssonii 177R1B.** *Caldicellulosiruptor* is a genus of anaerobic, extreme thermophilic bacteria [41]. *C. kristjanssonii* 177R1B (Accession NC_014721) was isolated from a hot spring biomat in Iceland and sequenced in 2010 by the

<table>
<thead>
<tr>
<th>From Alignment</th>
<th>Alignment position</th>
<th>% Conservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motifs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YD</td>
<td>25–26</td>
<td>55</td>
</tr>
<tr>
<td>DE</td>
<td>70–71</td>
<td>55</td>
</tr>
<tr>
<td>CK6CX3C</td>
<td>123–131</td>
<td>100</td>
</tr>
<tr>
<td>GGXPLL</td>
<td>184–189</td>
<td>91</td>
</tr>
<tr>
<td>TNG</td>
<td>218–220</td>
<td>64</td>
</tr>
<tr>
<td>CK6CX3C</td>
<td>483–492</td>
<td>100</td>
</tr>
<tr>
<td>PC</td>
<td>420–421</td>
<td>64</td>
</tr>
<tr>
<td>Single Residues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>25; 85</td>
<td>64; 73</td>
</tr>
<tr>
<td>G</td>
<td>82; 144; 425; 435; 450</td>
<td>82; 73; 82; 64</td>
</tr>
<tr>
<td>D</td>
<td>192; 340; 393; 434</td>
<td>73; 73; 64; 73</td>
</tr>
<tr>
<td>K</td>
<td>195; 273; 309; 477</td>
<td>73; 64; 64; 64</td>
</tr>
<tr>
<td>F</td>
<td>231; 257</td>
<td>82; 73</td>
</tr>
<tr>
<td>N</td>
<td>235; 271</td>
<td>73; 64</td>
</tr>
<tr>
<td>I</td>
<td>237</td>
<td>73</td>
</tr>
<tr>
<td>S</td>
<td>264</td>
<td>64</td>
</tr>
<tr>
<td>P</td>
<td>382; 439</td>
<td>82; 73</td>
</tr>
<tr>
<td>C</td>
<td>440; 495;530</td>
<td>82; 73; 64</td>
</tr>
</tbody>
</table>

Residues are numbered according to their position in the multiple alignments. 

doi:10.1371/journal.pone.0020852.001
Figure 3. Protein sequence alignment of TrnC and all the TrnC homologs identified in this study.
doi:10.1371/journal.pone.0020852.g003
Meta genomic analysis

A search for putative TrnD and TrnC-encoding genes was also performed against metagenomic DNA. The search revealed 365 TrnC and 151 TrnD homologs in metagenomic databases (homology being defined as BLASTP e-value < 10^{-6}) (Tables S3 and S4). This analysis revealed the presence of potential TrnC and TrnD-encoding genes in such diverse environments as the Indian Ocean [44], hypersaline lagoons from the Galapagos Islands, coastal sea water from the Gulf of Mexico [45,46], farm soil from Waseca County, USA [47] and a coral reef from French Polynesia [44]. This search provides further insight into the diverse distribution of microorganisms potentially capable of producing thuracin CD-like products.

Alignment of TrnC and TrnD-like Radical SAM proteins

The availability of a significant number of TrnC- and TrnD-like radical SAM protein sequences enabled further in silico analysis to identify conserved motifs and residues. Alignment of TrnD and 9 homologs (Fig. 2) reveals a number of completely conserved motifs and residues, which are summarized in Table 1. A radical SAM superfamily signature motif CX3CX2C was identified in 8 sequences. B. thuringiensis serovar huazhongensis BGSC 4BD1 Bthur0011_60830 and C. difficile CD0163 did not contain this motif. Radical SAM enzymes generate catalytic radicals by

Table 2. Highly conserved residues shared by trnC and 18 homologs.

<table>
<thead>
<tr>
<th>Residues</th>
<th>Alignment position</th>
<th>% Conservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C 535; 553; 629</td>
<td>100; 100; 100</td>
<td></td>
</tr>
<tr>
<td>R 352</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>E 324; 346; 650</td>
<td>74; 74; 63</td>
<td></td>
</tr>
<tr>
<td>N 333</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>K 293; 300</td>
<td>100; 58</td>
<td></td>
</tr>
</tbody>
</table>

Single Residues

| D | 44;243;349;401 | 58; 95; 79; 74 |
| G | 137;538;566; 573 | 58; 84; 100; 84 |
| T | 199 | 74 |
| Y | 222; 281 | 63; 68 |
| M | 233 | 89 |
| S | 254 | 68 |
| F | 280; 289 | 84; 74 |
| K | 293; 300 | 100; 58 |
| N | 333 | 74 |
| E | 324; 346; 650 | 74; 74; 63 |
| R | 352 | 100 |
| C | 535; 553; 629 | 100; 100; 100 |

Residues are numbered according to their position in the multiple alignments. doi:10.1371/journal.pone.0020852.t002
combining a 4Fe-4S cluster and S-adenosylmethionine (SAM) in close proximity. The reaction requires the input of one electron, which is supplied by the 4Fe-4S cluster that is bound to the protein via cysteine ligands that reside in the CX₃-CX₂-C motif [40]. Kaminska et al [10] demonstrated that mutation of this motif abolishes activity of Cfr, a member of radical SAM superfamily which encodes a methyltransferase. Conserved cysteine residues are also located in a CX₂-CX₃-C conformation starting at alignment position 456. Another feature common to all radical SAM proteins is a glycine-rich sequence motif proposed to be the SAM-binding site. Both GGE/DP and TNG/A motifs were identified in 60% and 50% of radical SAM proteins respectively. These putative SAM binding sequence motifs have also been observed in AtsB, a Fe-S oxidoreductase arylsulfatase regulatory protein [49]. In addition to these, other motifs of note included a YD/N motif present in 60% of the aligned proteins, a CN motif also found in 60% of the alignments and a PC/S motif present in 50%. A number of conserved single residues were identified including a number of aromatic residues. Alignment of the TrnC homologs (Fig. 3) also revealed several conserved regions summarized in Table 2. As with the TrnD homologs, the heterogeneity of TrnC homologs is most pronounced within the C-terminal regions which are believed to be responsible for binding of substrates and auxiliary cofactors, whereas the sequences of the N-termini are well conserved. The characteristic CX₃-CX₂-C motif was found in 12 TrnC homologs; B. thuringiensis serovar pulsilens BGSC 4CC1 Bthur0012_58010, Thermoanaeobacter mathronii A3 Tmath_1971 and Clostridium difficile QCD-63q42 CdifQCD-6_020200015426 did not contain the motif. Tmath_1971 and CdifQCD-6_020200015426 were also lacking the GGE/DP and CX₂-CX₅-CX₄-C motif. Bthur0012_58010 did contain the GGEPLL and the CX₂-CX₅-CX₄-C motif. Other motifs of note include a SXDG motif located starting at alignment position 340 and conserved in 89% of the strains. A CEK/R and a MX₂-D/E motif were conserved in 68.75% of the proteins. Highly conserved single residues included, again, a number of highly conserved aspartic acid residues. One of four conserved aspartic acid residues was found to be adjacent to a glycine in TrnC and 7 other homologs. These are likely to be important for the maintenance of the SAM binding site [50].

Figure 4. Rectangular cladogram of TrnD and all the TrnD homologs encountered during the screen. Cladogram was generated using the R package and visualised in DENDROSCOPE.

doi:10.1371/journal.pone.0020852.g004
The conserved nature of these TrnC and Trn D homologs facilitated a phylogenetic analysis of their relatedness. A neighbour-joining tree generated from aligned conserved regions on the basis of percent identity was constructed for both sets of homologs. The unrooted cladogram of the Trn D homologs showed that B. thuringiensis DPC 6431 TrnD is most closely related to its homolog in B. cereus 95/8201, clustering together in the same branch (Fig. 4). This supports the findings from the in-silico screen. The cladogram of Trn C homologs (Fig. 5) is quite similar to that of the Trn D homologs. Once again B. thuringiensis DPC 6431 is shown to be most closely related to B. cereus 95/8201.

**Discussion**

Thuricin CD, a bacteriocin with a narrow spectrum of activity which includes C. difficile, was recently identified as providing a potential therapeutic solution to this important pathogen [4]. C. difficile infection (CDI) is the leading cause of nosocomial infections associated with antibiotic use and is responsible for substantial morbidity and mortality worldwide [51]. The posttranslational modifications found in thuricin CD are unusual, and had not previously been associated with a two-peptide bacteriocin. In addition to potentially providing a treatment for CDAD, it, together with subtilosin A, may also represent a previously undefined structural class of bacteriocins, the sactibiotics. Bearing this in mind an in-silico approach was used to identify other thuricin CD-like gene clusters. An in-silico approach to identify strains that may be capable of bacteriocin production has recently become a viable alternative to culture based approaches as a consequence of the increasing generation and availability of genomic and metagenomic sequence data. While in silico identification of gene clusters in a strain will not always lead to the detection of an associated bacteriocin, previous studies have shown that there is a high correlation [13,52].

The initial in silico strategy adopted here resulted in the initial identification of 99 TrnC and 53 TrnD homologs. Further analysis

![Figure 5. Rectangular cladogram of TrnC and all the TrnC homologs encountered during the screen. Cladogram was generated using the R package and visualised in DENDROSCOPE. doi:10.1371/journal.pone.0020852.g005](https://www.plosone.org/doi/10.1371/journal.pone.0020852.g005)
led to the identification of 15 novel clusters which prior to this study had not been identified as potential bacteriocin-like clusters. These gene clusters are predominantly within the Firmicutes, with all but two gene clusters from species within this phylum. It was interesting to also find gene clusters within the Bacteroidetes phylum, namely Bacteroides sp 3.1.193, and also within the Thermotogae phylum, namely Petrotoga mobilis SJ95, because bacteriocin production within these phyla is quite rare. The availability in the future of additional species corresponding to these genera will reveal whether these gene clusters are exceptional or not. The analysis of metagenomic sequence data further highlighted the distribution of gene clusters containing TrnC and TrnD-like genes. While at present it is not possible to predict definitively if these metagenomic homologues are encoded with bacteriocin-like gene clusters, the possibility that at least a portion of these encode novel sacitibiotics is a fascinating prospect. Indeed, the harnessing of such novel antimicrobials from genomic/metagenomic material through heterologous expression or the harnessing of such novel antimicrobials from genomic/metagenomic datasets. The criterion for homolog detection was a threshold of 1e-6 and greater than 20% identity. Genes surrounding the TrnC and TrnD homologs were examined using the genome viewer on NCBI and also using the DNA sequence viewer ARTEMIS [54].

Phylogenetic analysis
Protein alignments were generated by MUSCLE [55]. A phylogenetic tree was constructed using the TrnC and TrnD homologs identified in the screen. Protein trees were built using the R statistical package (http://cran.r-project.org) and the resultant phylogenetic trees were visualised using the DENDROSCOPE package [56].

Metagenomic Analysis
The TrnC and TrnD protein sequences were used as driver sequences to search for homologs in metagenomic datasets. The metagenomic portal CAMERA (http://web.camera.calit2.net) [44] was used to BLAST against all publicly available metagenomic datasets. The criterion for homolog detection was a threshold of 1e-6 and greater than 20% identity.

Supporting Information
Table S1 Bacterial genomes in which TrnC homologs were identified. (DOC)
Table S2 Bacterial genomes in which TrnD homologs were identified. (DOC)
Table S3 TrnC homologs in metagenomic databases. (DOC)
Table S4 TrnD homologs in metagenomic databases. (DOC)

Author Contributions
Conceived and designed the experiments: OOS PC. Performed the experiments: KM. Analyzed the data: KM OOS. Wrote the paper: KM OOS PC MR. Principal investigator and intellectual input: RPR CH.

Materials and Methods
Sequence Analysis
The protein sequences of TrnC and TrnD from Bacillus thuringiensis DPC 6431 were used as driver sequences to mine bacterial genomes using the BLASTP web server on NCBI (www.ncbi.nlm.nih/BLAST) using default parameters [53]. Homology was defined as an e-value greater than 1e-6 and greater than 20% identity. Genes surrounding the TrnC and TrnD homologs were examined using the genome viewer on NCBI and also using the DNA sequence viewer ARTEMIS [54].

References


58. Genome Mining for Sactibiotic-Like Gene Clusters

59. PathFinder for Sactibiotic-Like Gene Clusters

60. Genome Mining for Sactibiotic-Like Gene Clusters