UCLA
UCLA Previously Published Works

Title

Transcriptome sequencing of the human pathogen Corynebacterium diphtheriae NCTC 13129 provides detailed insights into its transcriptional landscape and into DtxR-mediated transcriptional regulation

Permalink

https://escholarship.org/uc/item/4t14f30c

Journal

BMC Genomics, 19(1)

ISSN

1471-2164

Authors

Wittchen, Manuel

Busche, Tobias

Gaspar, Andrew H

et al.

Publication Date

2018-12-01

DOI

10.1186/s12864-018-4481-8

Copyright Information

Peer reviewed

BMC Genomics

RESEARCH ARTICLE

Open Access

CrossMark

Transcriptome sequencing of the human pathogen *Corynebacterium diphtheriae* NCTC 13129 provides detailed insights into its transcriptional landscape and into DtxR-mediated transcriptional regulation

Manuel Wittchen[1], Tobias Busche[1,2], Andrew H. Gaspar[3], Ju Huck Lee[4,5], Hung Ton-That[4], Jörn Kalinowski[1*] and Andreas Tauch[1]

Abstract

Background: The human pathogen *Corynebacterium diphtheriae* is the causative agent of diphtheria. In the 1990s a large diphtheria outbreak in Eastern Europe was caused by the strain *C. diphtheriae* NCTC 13129. Although the genome was sequenced more than a decade ago, not much is known about its transcriptome. Our aim was to use transcriptome sequencing (RNA-Seq) to close this knowledge gap and gain insights into the transcriptional landscape of a *C. diphtheriae tox*+ strain.

Results: We applied two different RNA-Seq techniques, one to retrieve 5′-ends of primary transcripts and the other to characterize the whole transcriptional landscape in order to gain insights into various features of the *C. diphtheriae* NCTC 13129 transcriptome. By examining the data we identified 1656 transcription start sites (TSS), of which 1202 were assigned to genes and 454 to putative novel transcripts. By using the TSS data promoter regions recognized by the housekeeping sigma factor σ^A and its motifs were analyzed in detail, revealing a well conserved −10 but an only weakly conserved −35 motif, respectively. Furthermore, with the TSS data 5′-UTR lengths were explored. The observed 5′-UTRs range from zero length (leaderless transcripts), which make up 20% of all genes, up to over 450 nt long leaders, which may harbor regulatory functions. The *C. diphtheriae* transcriptome consists of 471 operons which are further divided into 167 sub-operon structures. In a differential expression analysis approach, we discovered that genetic disruption of the iron-sensing transcription regulator DtxR, which controls expression of diphtheria toxin (DT), causes a strong influence on general gene expression. Nearly 15% of the genome is differentially transcribed, indicating that DtxR might have other regulatory functions in addition to regulation of iron metabolism and DT. Furthermore, our findings shed light on the transcriptional landscape of the DT encoding gene *tox* and present evidence for two *tox* antisense RNAs, which point to a new way of transcriptional regulation of toxin production.

Conclusions: This study presents extensive insights into the transcriptome of *C. diphtheriae* and provides a basis for future studies regarding gene characterization, transcriptional regulatory networks, and regulation of the *tox* gene in particular.

Keywords: *Corynebacterium diphtheriae*, Transcriptome sequencing, RNA-Seq, Transcription start site, Promoter, DtxR, Diphtheria toxin

---

* Correspondence: joern@cebitec.uni-bielefeld.de
[1]Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany
Full list of author information is available at the end of the article

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 2 of 17

Background

Corynebacterium diphtheriae is a Gram-positive bacterium causing the communicable disease diphtheria in humans by colonizing the upper respiratory tract or skin. Although a vaccine was introduced more than 100 years ago by von Behring, outbreaks still occur worldwide. A clinical isolate from a severe diphtheria outbreak in Eastern Europe in the 1990s, named *C. diphtheriae* NCTC 13129, was subjected to genomic sequencing in 2003 [1]. The genome of this *tox*⁺ strain has a size of about 2.5 Mbp with a G + C content of about 53% [1]. The published genomic information is a sound basis for further studies particularly concerning the pathogenicity of this bacterium [2–5].

Considered as the most important virulence factor of *C. diphtheriae*, diphtheria toxin (DT), encoded by the corynephage *tox* gene, has been studied extensively, with its structure [6–9] and mode of action [10, 11] now well characterized. Strain NCTC 13129 also harbors three pilus gene clusters, which encode three distinct adhesive pilus types that are assembled by sortase enzymes and critical for bacterial virulence [12, 13]. These pilus gene clusters and their variations are also identified in many pathogenic isolates from cases of diphtheria, endocarditis and pneumonia [3]. It is important to note here that *C. diphtheriae* mutants devoid of pili or DT are highly attenuated in the *Caenorhabditis elegans* and rodent models of infection [14, 15], supporting that DT and pili are the major virulence factors in *C. diphtheriae*.

Quite early the effect of increased DT is produced when *C. diphtheriae* is grown under iron-limiting conditions, and the basis of this modulation has been well studied and attributed to the transcriptional regulator DtxR [16–18]. In the presence of iron, DtxR is activated, forming a dimer that binds a 19 bp-palindromic sequence within the *tox* promoter, hence repressing the expression of the *tox* gene; in the iron-limiting conditions, DtxR is deactivated, leading to expression of DT [18]. Further research showed that DtxR is a dual regulator of iron homeostasis in many bacteria and its binding site was found upstream of several iron uptake related genes, like siderophores and heme oxygenases [3, 19–23]. It is not clear, however, if DtxR regulates genes coding for pili and the sortase machines.

Although the regulator and the genetic origin of the *tox* gene was identified many years ago, only a few studies focus on other genetic properties (e.g. promoter) of the DT encoding gene [24, 25]. In all only a few studies regarding the transcriptional organization of *C. diphtheriae* were performed [19, 22, 26, 27]. Promoters and operon structures are not known for the majority of genes.

For the analysis of transcriptomes a broad range of techniques exists. Northern blotting [28], Reverse transcription PCR [29], RACE (Rapid Amplification of cDNA Ends) [30] or Microarrays [31] are suitable for the analysis of transcripts and / or transcript abundance. The major drawback of most of these techniques is the fact that they only allow the analysis of a few to several targets in parallel, rendering the analysis of whole transcriptomes difficult and time consuming. Microarrays are designed for high-throughput screening of transcripts and their abundances but give little information about the transcript's sequence [32]. Transcriptome sequencing (RNA-Seq) solves a lot of these problems and delivers some unique features such as de novo gene discovery. The technique provides both, the characterization of transcripts with single nucleotide resolution and transcript quantification with a high dynamic range. It is therefore considered ideal for the analysis of complete transcriptomes [33]. RNA-Seq revealed an unexpected complexity of bacterial transcriptomes and shed light on novel transcripts like small and antisense RNAs [34, 35]. Furthermore, a deep view into the transcriptome can be used to improve genomic annotations by identifying novel transcripts and correcting translation start sites (TLS) [36, 37]. Next to the regular sequencing of full length mRNAs (also called whole transcriptome sequencing), which is used for transcription profiling and differential gene expression analysis [33, 34], new RNA-Seq protocols emerged, which allow the analysis of very specific RNA features. Specific RNA-Seq protocols were invented to exactly map transcript ends for the identification of transcription start sites (TSS) [38] or terminator structures [39], by keeping the benefits of high dynamic range and resolution. This data can be used to identify promoter regions, analyze 5′ or 3′ untranslated regions (UTRs) and their inherent regulatory elements such as riboswitches [36, 40, 41].

Another feature of RNA-Seq is the ability to identify operon structures [38, 39]. Operon detection based on genomic data relies on function prediction of genes, the proximity of genes to each other and the encoding strand [42, 43]. The disadvantage of this approach is the requirement of a good genome annotation, since unknown genes cannot be assigned to operons. In addition, the combination of whole transcriptome and TSS data enables the identification of sub-operons, which are shorter transcripts originating from the same primary operon with an internal TSS [38, 39, 41].

This study aims to use RNA-Seq to gain insights into the transcription start sites (TSS) by enrichment of native 5′-ends of RNA and the transcriptional profile of *C. diphtheriae* wild type and Δ*dtxR* mutant strains by using whole transcriptome libraries. The obtained TSS data was further analyzed to get information about promoters and 5′-UTRs, the shorter of which only contain ribosomal binding sites and the longer ones may contain complex regulatory structure such as riboswitches, RNA

thermometers or attenuators. Furthermore the wild type whole transcriptome data was used to detect operon structures. By combining the primary 5′-end and the whole transcriptome data of the wild type possible sub-operon structures can be characterized.

The whole transcriptome of the *C. diphtheriae* wild type and the Δ*dtxR* mutant was sequenced and compared to identify differentially expressed genes. The positions of DtxR binding sites relative to detected TSS are investigated. As the DT encoding gene *tox* is essential for the pathogenicity of *C. diphtheriae* we give detailed insights on the transcriptional landscape of this important gene. To the best of the authors' knowledge the data presented here is the first RNA-Seq analysis of *C. diphtheriae*.

Methods
Bacterial strains and culture conditions
C. diphtheriae strains were grown in heart infusion broth (HIB) or on heart infusion agar (HIA), whereas *Escherichia coli* strains were grown on Luria broth (LB).

In-frame deletion of the *C. diphtheriae dtxR* gene was performed according to a published protocol [12]. Briefly, primer sets dtxR-A/B and dtxR-C/D were used to amplify 600 bp fragments upstream and downstream of *dtxR*, respectively, from the chromosomal DNA of strain NCTC 13129. The products were used for cross-over PCR with primers A and D to generate a 1.2 kbp fragment, while appending BamHI sites. The 1.2 kbp product was cloned into the BamHI sites of the conjugative vector pK18*mobsacB* [44]. DNA sequencing was employed to verify the insertion and the resulting plasmid was introduced into *E. coli* S17−1. The *E. coli* S17−1 strain harboring the resulting conjugative vector was used for deletion of *dtxR* in the parental strain NCTC 13129 via homologous recombination as previously described [12]. Confirmation of the defined *dtxR* deletion in the generated mutant was performed by PCR using the primers A and D (Additional file 1: Table S1).

Cell sampling and RNA isolation
Overnight cultures of *C. diphtheriae* NCTC 13129 and its isogenic Δ*dtxR* mutant were used to inoculate fresh cultures in HIB at 37 °C. Total RNA was isolated from cells grown at exponential phase ($OD_{600} = 0.5$) using Trizol reagent (Invitrogen). The bacterial pellet obtained from 4 mL culture was resuspended in 1 mL Trizol, transferred into FastPrep Lysis Beads & Tube (MP Biomedicals) and mechanically lysed using beadbeater at a maximum speed for 20 s six times. After adding 200 μL chloroform to the lysed cells followed by centrifugation at 12,000×g for 15 min at 4 °C, the aqueous supernatant was taken and then precipitated using 500 μL isopropanol. Afterwards, crude RNA samples

were treated with DNase I (Roche Diagnostics). After purification using phenol/chloroform/isoamyl alcohol (ratio 25:24:1), RNA was precipitated with ethanol. Purified total RNA pellets were dissolved in 50 μL RNase-free water. The purified RNA was quantified with a NanoDrop (Peqlab) and by Agilent RNA Nano 6000 kit on Agilent 2100 Bioanalyzer (Agilent Technologies). PCR was performed to assure no DNA remained in the samples.

cDNA library preparations and sequencing
For whole transcriptome cDNA library preparations 2 μg total RNA from *C. diphtheriae* NCTC 13129 and the isogenic Δ*dtxR* mutant were used. Stable RNAs were depleted with the Ribo-Zero rRNA Removal Kit (Bacteria) according to manufacturer's instructions (Epicentre). Afterwards the remaining mRNA was purified using RNA MinElute columns (Qiagen) and checked for quality with the Agilent RNA Pico 6000 kit and the Agilent 2100 Bioanalyzer (Agilent Technologies). Fragmentation of mRNA, reverse transcription to cDNA, adenylation of 3′-ends, adapter ligation and PCR amplification were performed according to TrueSeq Stranded mRNA library instructions (Illumina). Prior to paired-end sequencing of the cDNA libraries on an Illumina MiSeq, their quality and concentration were checked using the Agilent High Sensitivity DNA kit and the Agilent 2100 Bioanalyzer (Agilent Technologies).

For the primary 5′-end cDNA library 2× 5 μg RNA from the wild type *C. diphtheriae* NCTC 13129 was used. The preparation protocol has been described previously in detail [39]. In the present study, the experimental work-flow was slightly modified at three steps. Non-primary transcripts were digested with terminator exonuclease (Epicentre) at 30 °C for 60 min and subsequently at 42 °C for 30 min. The 5′ adapter ligation was performed for 60 min at 30 °C with 1 μL 60 μM adapter (Additional file 1: Table S1). After cDNA amplification the two libraries were purified and size-selected by gel electrophoresis for fragment sizes between 100 and 1000 bp. The cutoff of 100 bp was chosen to reduce the presence of adapter dimers in the library. Due to the fact that the library preparation involves the use of two adapters, which together have a length of 66 nt, only transcripts smaller than 40 nt are not present in the final RNA-Seq data. Sequencing was performed in single-read mode with 75 nt read length for the 5′-enriched cDNA library on an Illumina MiSeq.

Bioinformatics data analysis
Read mapping and visualization
The paired-end reads of the whole transcriptome cDNA libraries from the wild-type and the Δ*dtxR* mutant were trimmed for low quality bases from both ends and a sliding window trimming (removing bases when the average

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 4 of 17

quality per base in a window of 4 nt decreases below 15) using trimmomatic v0.36 [45]. Reads which were trimmed to a length shorter than 39 nt were discarded. The remaining paired-end reads were mapped with bowtie2 v2.2.7 [46] to the *C. diphtheriae* NCTC 13129 genome (RefSeq NC_002935.1) with default settings for paired-end read mapping. The single-end reads of the primary 5′-end cDNA library were trimmed from the 3′ end only with trimmomatic. The remaining reads with a minimal length of 39 nt were mapped with bowtie2 using default settings for single-end read mapping. All mapped sequence data was converted from SAM to BAM format to decrease usage of disk space with SAMtools v1.3 [47] and visualized and inspected with ReadXplorer v.2.2 [48].

Identification of transcription start sites

To automatically detect transcription start sites (TSS), the TSS detection mode of the 'Transcription Analyses' function of ReadXplorer was applied on the primary 5′-end cDNA library data. After empirical testing and inspection of various parameter sets based on the automatic parameter estimation by ReadXplorer the criteria for the automatic detection of putative TSS were a minimal coverage increase of 100% with at least 28 read starts at a particular genomic position. These values were found to result in an appropriate signal to noise ratio after manual inspection of the predicted TSS. The resulting list of predicted TSS was manually checked for false-positives. A putative TSS was considered as false-positive if no clear accumulation of read starts was observed at the particular genomic position and the putative TSS was found inside an uneven gradient of accumulated read starts, as it was often the case for putative TSS detected within a highly transcribed coding region.

Identification of novel transcripts

The TSS data from the primary 5′-end cDNA library were used to identify novel transcripts. The predicted TSS were associated with an annotated gene if they are located up to 500 bp upstream of the respective coding region. All TSS not associated with a known gene were assigned as novel transcripts. To further characterize this class of TSS it was divided into three groups: (a) TSS which are located between two annotated genes are considered as intergenic TSS; (b) TSS which are positioned inside an annotated gene are denoted as intragenic TSS; and (c) TSS which are located on the opposite strand of an annotated gene were classified as antisense TSS. To find novel transcripts which encode putative proteins the transcript length was estimated by the coverage from the wild type whole transcriptome cDNA library. The covered genomic region was searched for open reading frames (ORFs) by using UGENE [49]

with AUG, GUG, UUG and CTG as start codon settings. Additionally, the corresponding stop codon had to be located within the covered genomic region. The predicted ORFs were checked for potential homologous proteins using NCBI BLAST [50, 51]. In case no ORF or protein homologue was detected, the sequence downstream of the TSS was analyzed for potential ncRNAs and RNA motifs using RFAM [52]. Newly identified genes were assigned with a locus tag containing a unique identifier.

Analysis of sequence motifs

To find conserved DNA sequence motifs in the *C. diphtheriae* NCTC 13129 genome, the motif-finding software Improbizer [53] and the visualization program WebLogo 3 [54] were used. Depending on the assumed motif location different input sequences were used. For the identification of σ^A promoter motifs the DNA sequence 40 bp upstream of the TSS assigned to a gene were taken as input. Improbizer reported an extended −10 region consisting of four unpreserved leading bases, the conserved core hexamer and two unpreserved tailing bases. For simplification the identified region was truncated to the conserved core hexamer and used as −10 motif in this work.

As the −35 motif of σ^A promoters is more frequent in presence of a −10 motif with low similarity to consensus [55], only sequences upstream of non-optimal −10 regions were used to identify possible −35 motifs. In addition to that the maximal allowed distance between the −35 and the −10 motif was set to 23 bp.

For the determination of ribosomal binding sites all genes, including predicted proteins encoded by novel transcripts, with an 5′-UTR longer than 5 nt were considered and the DNA sequence 20 bp upstream of the translation start site (TLS) was used as input for Improbizer.

The identified motifs are represented in the text in upper case letters if the frequency for the particular base is > 80% and in lower case letters if the frequency is below 80% of all analyzed sequences.

Identification of operon structures

Two or more genes are transcriptionally connected and part of an operon if they are transcribed from a single promoter. The detection of operon structures in *C. diphtheriae* was performed with ReadXplorer [48]. For this purpose the read pairs (with a minimal mapping quality ≥30) from the wild type whole track cDNA libraries spanning two neighboring genes on the same strand were counted. If more than five spanning reads were found the two genes were assigned to an operon structure (primary operon). The process was continued consecutively for the next genes until no more genes could be assigned to that operon. All three wild type cDNA libraries from the replicates were combined for

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 5 of 17

the operon detection to increase the number of reads in low coverage regions. Genes represented by novel transcripts encoding proteins or ncRNAs (i.e. tmRNA) were manually checked for operon structures, as ReadXplorer can only allocate already annotated genes to operon structures. The list of primary operons was further divided into two groups: experimentally validated operons which have an assigned TSS and predicted operons which do not have an assigned TSS to their first gene. Sub-operons were identified in case a TSS was detected for a posterior gene in a primary operon. All genes which could not be connected into an operon were considered as monocistronic transcripts.

Differential gene expression analysis

Prior to differential expression analysis the whole transcriptome data from the *C. diphtheriae* wild type and the Δ*dtxR* mutant cDNA libraries were processed as described under 'Read mapping and visualization'. For differential expression analysis the reads belonging to genes of three replicates per condition were counted with ReadXplorer and tested for differential expression with DESeq2 [56] using default settings. Genes with a false discovery rate corrected *p*-value below 0.05 and a $\log_2$(fold change) above +1.0 or below -1.0, respectively, were considered to be differentially transcribed under the examined conditions.

Results and discussion

cDNA sequencing and mapping to the *C. diphtheriae* NCTC 13129 genome

Although the genome of *C. diphtheriae* NCTC 13129 was published in 2003 [1] only a few examples about the transcriptional organization, promoter motifs and non-coding RNAs are known [18, 19, 24, 27, 57]. To perform qualitative and quantitative analyses of the *C. diphtheriae* transcriptional landscape, a technology capable of single-nucleotide resolution with great accuracy is required. Furthermore a technology with high dynamic range is needed for the analysis of the quantitative transcriptome. The recent technology of cDNA sequencing or RNA-Sequencing (RNA-Seq) fulfills all requirements and allows qualitative and quantitative transcriptome analyses in parallel [34, 58, 59]. Therefore we constructed two different types of cDNA libraries: a primary 5′-end-specific cDNA library of the wild type and whole transcriptome cDNA libraries of the wild type and of an isogenic Δ*dtxR* mutant. Triplicates of the wild type strain and the Δ*dtxR* mutant were cultivated to exponential growth in heart infusion broth medium, resulting in six individual whole transcriptome strand-specific cDNA libraries. All libraries were sequenced using a strand-specific protocol and an Illumina MiSeq machine with a read length of 75 nt or 2×75 nt for single-end and paired-end reads, respectively. The

reads were quality-trimmed with trimmomatic [45] and mapped with bowtie2 [46] to the *C. diphtheriae* NCTC 13129 genome, using default parameters. Between 96 to 99% of the reads were mapped to the genomic reference (Additional file 2: Table S2). For visualization and further analysis, the mapped reads were imported into ReadXplorer [48].

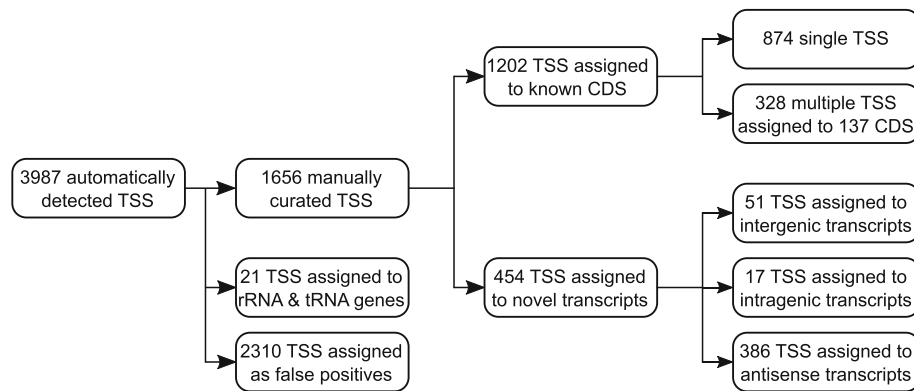### Identification of transcription start sites of primary transcripts

The analysis of the primary 5′-end cDNA data with the software ReadXplorer [48] resulted in the automatic detection of 3987 putative transcription start sites (TSS) in the *C. diphtheriae* NCTC 13129 genome. After manual inspection of the automatically detected TSS, 2310 false-positive TSS and 21 TSS assigned to rRNA and tRNA were discarded, leaving a list of 1656 manually curated TSS. These TSS were divided into two main categories: TSS that can be associated with annotated genes of the reference genome and TSS that probably belong to novel, not yet annotated transcripts. TSS assigned to annotated genes were further split into two categories: genes with a single TSS (874 genes) and genes with multiple TSS (137 genes), the latter case containing a total of 328 TSS. The TSS belonging to novel transcripts were classified into three groups: (a) intergenic TSS where the novel transcript is located between two annotated genes, (b) intragenic TSS where the novel transcript is located in the annotated coding region and (c) antisense TSS where the novel transcript is located on the opposite strand of an annotated gene.

All in all, 1202 TSS were assigned to protein-coding genes. In addition, 454 TSS associated with novel transcripts were detected which are not assigned to previously annotated genes: 51 TSS belong to novel intergenic transcripts, 17 TSS to intragenic and 386 TSS to antisense transcripts (Fig. 1; Additional file 4: Table S4 and Additional file 6: Table S6).

In rare cases a RNA might be rapidly dephosphorylated but somehow stabilized from degradation in the cell (for example by translating ribosomes or secondary structure). During library preparation this kind of RNA will be degraded leading to a loss of the TSS signal. For the 4.5S RNA and the 6C RNA no TSS could be detected, but both were abundant in the whole transcriptome data set. Nevertheless, the vast majority of all transcripts were covered.

Identification of the house-keeping sigma factor σ^A promoter motif

The identified TSS were used to analyze the upstream promoter regions for conserved motifs, representing DNA signals for the corynebacterial housekeeping sigma factor σ^A. For this search the software Improbizer [53] was used

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 6 of 17

Fig. 1 Classification of detected transcription start sites (TSS). The identification, curation and classification of TSS is shown. From the automatically detected 3987 TSS 2310 false positive TSS and 21 TSS belonging to rRNA and tRNA genes were removed, resulting in 1656 putative TSS assigned to different transcript types

to scan the upstream sequence of the detected TSS for conserved −10 motifs. The identification of the −35 motif was performed by searching the DNA sequence 23 bp upstream of non-optimal −10 motifs as the −35 motif of σ^A promoters is more frequent in presence of a −10 motif with low similarity to the consensus [55]. The −10 motif TAgaaT was identified upstream of 1190 (98.9%) TSS (Fig. 2). The recognized −35 motif (ttgcaa) is not well conserved, but it was found within a distance of 16–20 bp upstream of 1031 TSS that also possess a non-optimal −10 motif. The spacer length between the −10 motif and the detected TSS is 6 to 9 bp with a mean of 6.9 bp. The TSS itself is mainly (91%) a purine (A or G). The determined −10 motif and spacer length are in good agreement with data from *C. glutamicum* [39, 60, 61], a non-pathogenic relative of *C. diphtheriae*. The −35 motif of the *C. glutamicum* σ^A promoter (ttgcaa) is identical to that of *C. diphtheriae* and also not well conserved [39]. The comprehensive promoter data presented here lays the cornerstone for an in depth analysis of promoter motifs, which has already been done for *C. glutamicum* [55].

Characteristics of 5′-untranslated regions (5′-UTRs)
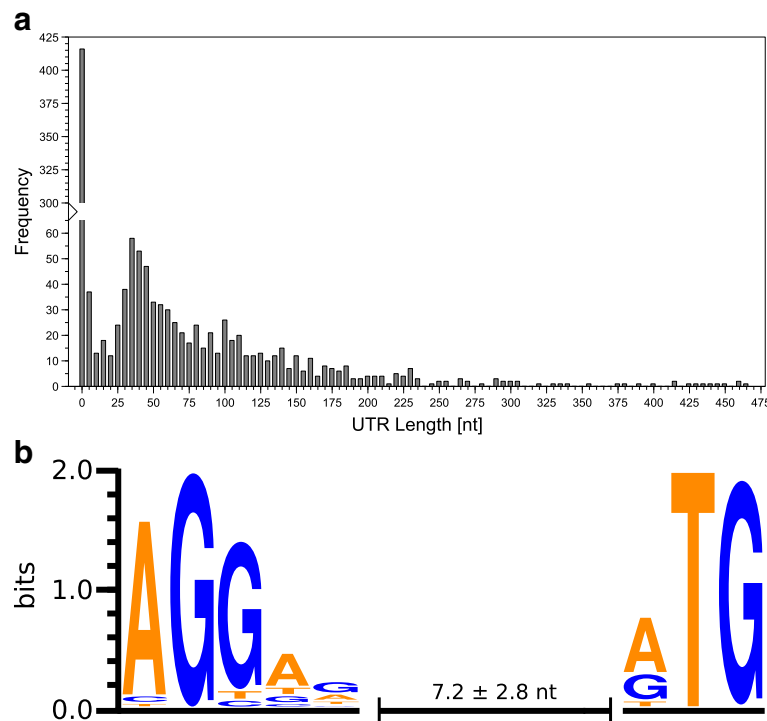### 5′-UTR length distributions
By analyzing the region between the TSS and the translation start site (TLS) in the primary 5′-end data, it was possible to obtain information on 5′-untranslated regions (5′-UTRs) in mRNA. The set of 1202 TSS assigned to known and 29 intergenic TSS assigned to novel protein-coding regions were used to characterize the 5′-UTRs. The length of the 5′-UTRs in *C. diphtheriae* mRNAs varies from 0 nt to 463 nt. The latter distance is for gene DIP1924A, a novel transcript identified in this study, encoding a hypothetical protein (Fig. 3a). Leaderless transcripts are mRNAs with 5′-UTRs that are too short for harboring a ribosomal binding site (RBS). Therefore, we categorized all genes with a 5′-UTR length from 0 to 5 nt as leaderless, as they cannot contain a canonical RBS with spacer. By using the primary 5′-end data 20% (452 of 2265) of the *C. diphtheriae* genes were found to be translated from leaderless transcripts (Additional file 3: Table S3). A large number of leaderless transcripts is a common feature of Actinobacteria [62] in general and Corynebacteria in particular, as the *C. glutamicum* transcriptome contains ~ 33% leaderless transcripts [39].

Further, the start codon usage was analyzed in both classes of transcripts, leaderless and leadered transcripts. For both classes AUG is the most frequently used start codon, followed by GUG. Around 80% of the leaderless transcripts contain an AUG, while only 62% of the leadered transcripts use this triplet as a start codon. The
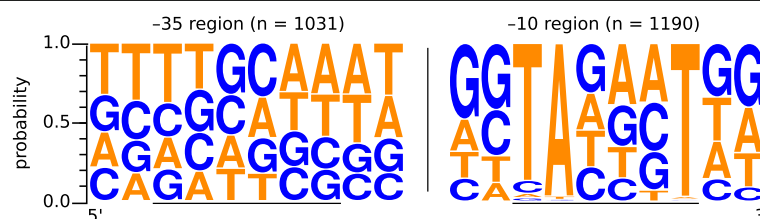
Fig. 2 Promoter motifs for the sigma factor σ^A of *C. diphtheriae*. The size of the nucleotide represents the relative abundance of the particular nucleotide at this position. The −10 motif was found upstream of 1190 TSS and the −35 motif upstream of 1031 TSS. The data was visualized with Weblogo [54]

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 7 of 17

Fig. 3 Histogram of 5'-UTR lengths and Ribosomal binding site motif in *C. diphtheriae*. **a** For the 5'-UTR analysis 1232 TSS from known genes and novel transcripts were used. The distribution of 5'-UTR lengths ranges from 0 nt (leaderless) to a maximum of 463 nt. Transcripts with a 5'-UTR of up to 5 nucleotides make up one third of all protein-coding genes. The bars represent UTR length increments of five nucleotides (1–5 nt, 6–10 nt, etc.), except for the first bar which represents UTR length of zero nucleotides. **b** Ribosomal binding site motif and mean distance to start codon. All 5'-UTRs longer than 5 nt were used for motif search. The y-axis shows the information content in bits for every nucleotide position. The diagram was created with Weblogo [54]

start codon UUG (~ 8%) is only found in transcripts having a ribosomal binding site.

As shown in Fig. 3a, a minor fraction (18%) of 5'-UTRs in *C. diphtheriae* has a length between 26 and 49 nt. These 5'-UTRs are long enough to contain a complete RBS with sufficient spacer length to the start codon, but are probably too short to harbor cis-regulatory elements.

Riboswitches and other RNA motifs

Another large group of 5'-UTRs have a length of > 100 nt. In total 264 (21%) genes are specified by a 5'-UTR of 100 nt or longer. These long 5'-UTRs might contain cis-regulatory elements which can influence transcription or translation of the mRNA by distinct sequence motifs or by folding into specific secondary structures. For various bacteria cis-regulatory elements in 5'-UTRs are known and can contain sequences of attenuators, riboswitches or binding sites for trans-regulatory elements [63, 64]. To find possible cis-regulatory elements in the 5'-UTRs, the genome sequence of *C. diphtheriae* NCTC 13129 was analyzed with the software Infernal [65] and the Rfam database [52] as search space. The results were compared with the 5'-UTR data from the primary 5′-end data set. Seven regulatory

elements were predicted in the *C. diphtheriae* NCTC 13129 genome sequence, of which five elements were found to be transcribed at the applied conditions (Table 1). In addition to riboswitches the predicted RNA motifs common in actinobacteria and named *mraW* RNA and *msiK* RNA, presumably involved in peptidoglycan synthesis and in sugar import [66, 67], respectively, were detected as transcribed.

Leader peptides are small peptides encoded upstream of some amino acid biosynthesis operons. Their translation leads to a differential folding of the attenuator RNA depending on the intracellular availability of certain amino acids [68]. In *C. diphtheriae* NCTC 13129 three putative leader peptide genes were identified: the *trp*, *ilvB* and *leu* leader peptide genes. Upstream of the first gene of the operon for tryptophan (W) biosynthesis, *trpB1*, the *trp* leader peptide (MTNMNAHN*WWW*RA*) encoded at nucleotide positions 2,456,505–2,456,545 bp was found, but no TSS was identified for the leader peptide gene. The *ilvB* leader peptide (MNIIRL*VV*I TTRRLP*) is encoded upstream of *ilvB* at nucleotide positions 1,081,747–1,081,794 bp with a TSS at the leader start position rendering it a leaderless transcript. The *ilvB* gene is involved into the biosynthesis of the amino acids isoleucine (I) and valine (V). The *leu* leader

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 8 of 17

Table 1 Predicted cis-regulatory elements in the 5′-UTRs. The predictions were obtained by using Infernal 1.1.2 with the Rfam 12.1 database and compared with the primary 5′-end data set. The list is sorted by Bit Score in descending order. Abbreviation: n.a., not applicable

Infernal prediction with Rfam						RNA-Seq detection			
Name	ID	Start	End	Bit Score	Strand	Status	Start	End	Gene
Cobalamin riboswitch	RF00174	1,066,117	1,066,317	90.8	+	observed	1,066,117	1,066,317	DIP1084
Cobalamin riboswitch	RF00174	862,214	862,414	81.8	+	not observed[a]	n.a.	n.a.	n.a.
cspA	RF01766	285,506	285,923	64.8	+	observed	285,506	285,923	DIP0320 / cspA
mraW	RF01746	1,640,095	1,639,987	70.2	−	observed[b]	1,640,095	1,639,987	DIP1606 / mraW
TPP riboswitch	RF00059	924,778	924,890	60.8	+	observed[c]	924,778	924,890	DIP0953
TPP riboswitch	RF00059	922,838	922,969	60.0	+	observed[c]	922,838	922,969	DIP0951
msiK	RF01747	509,334	509,277	48.2	−	observed	509,334	509,277	DIP0539
TPP riboswitch	RF00059	27,456	27,566	57.6	+	not observed[a]	n.a.	n.a.	n.a.

[a]No TSS detected and insufficient read coverage in that area
[b]Motif of the *mraW* region is located upstream of the TSS of DIP1606 / *mraW*
[c]Last base of TPP riboswitch is located in the respective CDS

peptide (MNRAN*LLLL*RRGGSQA*) is encoded at nucleotide positions 230,506–230,455 bp upstream of *leuA*. The *leuA* gene encodes the first step in leucine (L) biosynthesis. Neither for *leuA* nor for its leader peptide gene a TSS could be assigned due to weak transcription.

Ribosomal binding site motif
By using the 5′-UTR sequence information, a scan for ribosomal binding sites (RBS) was performed. Analysis of 779 5′-UTRs with a length larger than 5 nt by the software Improbizer [53] resulted in the conserved RBS motif AGGag in about 87% of all analyzed 5′-UTRs (Fig. 3b). The mean distance between the predicted RBS and the translation start site (TLS) of the coding region is 7.2 ± 2.8 nt. The identified RBS motif of *C. diphtheriae* is identical and the determined mean distance from RBS to TLS is very similar to that of *C. glutamicum* [39]. This was expected since the RBS-binding 3′-end of the 16S rRNA is identical in both organisms.

Re-annotation of coding sequences and detection of novel transcripts
By knowing the exact position of transcription start sites (TSS) of mRNA in the *C. diphtheriae* NCTC 13129 genome it is possible to verify, correct and re-annotate predicted coding sequences in the reference genome sequence. Furthermore, novel transcripts can be detected in the genome sequence and annotated. Accordingly two scenarios were anticipated; we corrected the translation start site (TLS) of coding sequences if the TSS is located downstream of the annotated TLS: a) In case the TSS is located at the first base of a potential start codon (ATG or GTG) that is in-frame with the annotated CDS, the TLS is shifted to the TSS position, resulting in a leaderless transcript. b) In case the TSS is not located at a start codon, the TLS is shifted to the next downstream in-frame start codon, resulting in a shortened CDS with a

5′-UTR of a length greater than 0 nt. By applying the two rules mentioned above, 120 TLS of predicted coding regions were corrected, of which 104 CDS are leaderless and 16 CDS have a 5′-UTR length greater than 5 nt (Additional file 3: Table S3 and Additional file 5: Table S5). These corrections were cross-checked by amino acid sequence similarity searches to orthologous proteins in databases and considered in the analysis of the 5′-UTR length distributions and in the motif searches.

By analyzing the intergenic, intragenic and antisense TSS, it is also possible to identify novel transcribed regions in the genome of *C. diphtheriae* NCTC 13129. As mentioned above, 454 TSS were classified as novel transcripts (Fig. 1). The intergenic TSS indicate novel not yet annotated coding regions or non-coding RNAs (ncRNAs). Only a few intergenic (51) and intragenic (17) TSS were detected. It is not clear which function these intragenic TSS have in *C. diphtheriae*, as they might lead to shortened proteins. For a range of organisms, e.g. bacteria [38], viruses [69] and eukaryotes [70], intragenic TSS and their shortened gene products have been described. These intragenic transcripts might contain regulatory regions which increase the genomic information content [71].

To find novel protein-coding transcripts, the sequences downstream of intergenic TSS were analyzed for open reading frames (ORFs) using the software UGENE [49]. In case a potential ORF was found, its amino acid sequence was analyzed with BLASTp [50, 51] to detect possible protein homologues in public databases. In case no ORF or protein homologue was found, the sequence downstream of the TSS was searched for ncRNAs or RNA motifs with RFAM [52]. For 29 of the 51 intergenic TSS a potential ORF was found and assigned with a distinct locus tag. The two ncRNAs tmRNA and RNase P M1 RNA were also identified as novel transcripts (Additional file 6: Table S6).

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 9 of 17

Around 40% of the newly detected ORFs were predicted to encode hypothetical proteins or proteins of unknown functions, but some proteins with metabolic function were also predicted. These proteins encode formate C-acetyltransferase, an ammonium transporter, magnesium chelatase, and glycine dehydrogenase. In addition, a gene encoding a putative helix-turn-helix (HTH) family transcriptional regulator (DIP1817A) was identified in the intergenic region between DIP1816 and DIP1817 (Fig. 4).

Analysis of operon structures by combining the primary 5′-end and the whole transcriptome data sets

By combining the primary 5′-end and the whole transcriptome data sets it is possible to obtain further insights into the transcriptional landscape of *C. diphtheriae* NCTC 13129, in particular into operon structures. An operon is a polycistronic transcript consisting of at least two genes transcribed from a common promoter. We defined a requirement of at least five reads spanning two adjacent genes to assign them to a primary operon. A primary operon was considered as 'experimentally validated' if a TSS was assigned to the first gene of that operon and 'experimentally validated by read pairs' if no TSS could be detected. In case an additional TSS is located inside of a deduced primary operon, a shortened transcript is generated during gene expression that defines a sub-operon, containing one or more genes. All genes not assigned to an operon were classified as monocistronic transcripts and were categorized regarding their TSS detection.

Under the studied conditions 471 primary operons containing 1417 genes were deduced from the transcriptome data. Of the 471 primary operons 337 operons (72%) are experimentally validated, as a TSS was assigned to their first genes, leaving 134 operons as experimentally validated by read pairs only. When considering internal TSS, the primary operons contain 167

sub-operons (Fig. 5). The two ncRNAs tmRNA and RNase P M1 RNA are co-transcribed in operons with protein-coding regions. The tmRNA is part of an operon with the *smpB* (DIP0750) gene encoding the SsrA-binding protein. The M1 RNA is transcribed in a primary operon consisting of DIP1679, DIP1678, M1 RNA and DIP1677. This primary operon is further divided into two sub-operons ranging from DIP1678 to DIP1677 and from M1 RNA to DIP1677 indicating a complex expression pattern of this genomic region. For the 4.5S RNA (DIP0256) and the known actinobacterial 6C RNA no TSS was detected in this study.
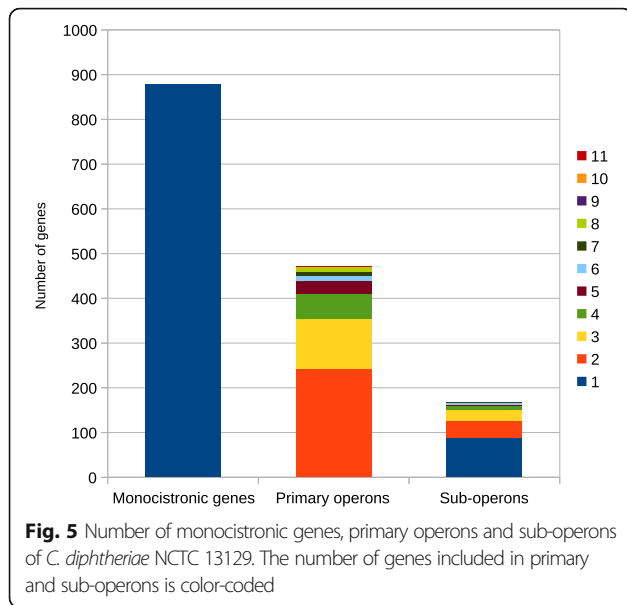
The largest primary operon covers eleven genes which code for various ribosomal proteins of the 30S and 50S subunits. Ten operons covering eight genes exist in the *C. diphtheriae* transcriptome containing genes involved in various cellular functions from replication to carbohydrate metabolism (Table 2). A list of all detected operons and sub-operons of *C. diphtheriae* NCTC 13129 is provided in the Supplemental Material (Additional file 7: Table S7).

The number of monocistrons in the *C. diphtheriae* genome accounts for 878 genes (38% of the predicted genes), of which 550 genes (63%) were associated with a TSS in this study (Fig. 5). Considering the number of genes assigned to primary operons as well as monocistrons with an assigned TSS, nearly 87% of all annotated genes of *C. diphtheriae* NCTC 13129 were detected as actively transcribed in this study.

We compared our results from the operon detection to the *in silico* operon predictions from the Database of prOkaryotic OpeRons (DOOR) [72]. Our RNA-Seq based operon detection is in agreement with the vast majority (89%) of all primary operons predicted by DOOR. The missing 11% were not evaluated in this study due to insufficient read coverage in the respective regions caused by low transcription.
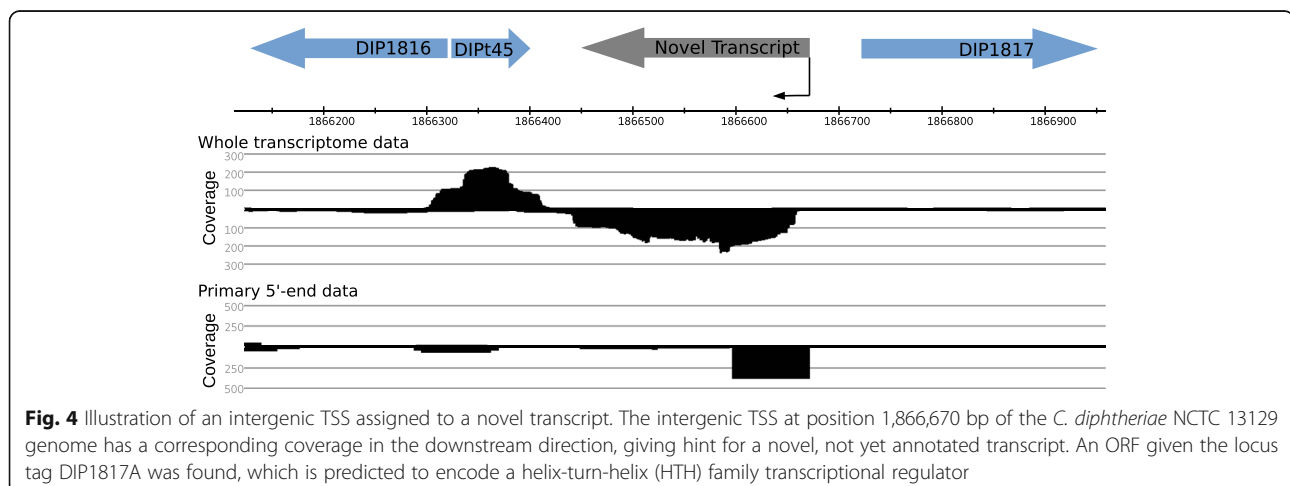
Fig. 4 Illustration of an intergenic TSS assigned to a novel transcript. The intergenic TSS at position 1,866,670 bp of the *C. diphtheriae* NCTC 13129 genome has a corresponding coverage in the downstream direction, giving hint for a novel, not yet annotated transcript. An ORF given the locus tag DIP1817A was found, which is predicted to encode a helix-turn-helix (HTH) family transcriptional regulator

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 10 of 17

Fig. 5 Number of monocistronic genes, primary operons and sub-operons of *C. diphtheriae* NCTC 13129. The number of genes included in primary and sub-operons is color-coded

Considering all detected TSS and identified primary operons as well as monocistrons, around 87% of all annotated genes are represented in this study. The remaining genes not covered are most likely due to two reasons: We analyzed transcription during exponential growth phase in complex media. Genes only active in other growth phases or conditions are not considered. Furthermore the applied method for capturing 5′-ends of transcripts relies on the fact that actively transcribed RNA is triphosphorylated at the 5′-end, which might not be the case for some transcripts. However, the large majority of transcripts was covered in this study.

Analysis of the DtxR regulon by comparing two whole transcriptome data sets

The diphtheria toxin repressor (DtxR) is the transcriptional regulator of iron homeostasis and the diphtheria toxin gene *tox* in *C. diphtheriae* and therefore important for the pathogenicity of this bacterium [73]. The iron-sensing transcription regulator DtxR binds to the DtxR

Table 2 Largest primary operons in *C. diphtheriae* NCTC 13129. The predictions from the eggNOG database (v4.5) were used to classify genes by their functions

Genes	Number of genes	Strand	TSS	Gene names and classification by eggNOG database
DIP0472 - DIP0482	11	+	detected	*rpsJ, rplC, rplD, rplW, rplB, rpsS, rplV. rpsC, rplP, rpmC, rpsQ.* Translation, ribosomal structure and biogenesis (ribosomal proteins)
DIP0203 - DIP0209	8	+	not detected	DIP0203, DIP0204, DIP0205, DIP0206, DIP0207, DIP0208, DIP0208A, DIP0209. Function unknown, phage proteins
DIP0407 - DIP0414	8	+	detected	*hemE, hemG, hemL,* DIP0410, DIP0411, DIP0412, DIP0413, DIP0414. Coenzyme metabolism; Carbohydrate metabolism and transport; Post-translational modification, protein turnover, and chaperones
DIP0438 - DIP0445	8	+	detected	DIP0438, DIP0439, DIP0440, DIP0441, DIP0442, DIP0443, DIP0444, DIP0445.Inorganic ion transport and metabolism; Function unknown
DIP0719 - DIP0726	8	+	detected	DIP0719, DIP0720, DIP0721, DIP0722, DIP0723, DIP0724, DIP0725, *uvrD.* Function unknown; Replication, recombination and repair; Inorganic ion transport and metabolism
DIP0969 - DIP0976	8	+	detected	DIP0969, DIP0970, DIP0971, DIP0972, *fdxA,* DIP0974, DIP0975, DIP0976. Signal transduction mechanisms; Amino Acid metabolism and transport; Function unknown; Energy production and conversion
DIP1248 - DIP1241	8	–	detected	DIP1248, DIP1247, DIP1246, DIP1245, DIP1244, DIP1243, *tatA, tatC.* Intracellular trafficking, secretion, and vesicular transport; Transcription; Amino Acid metabolism and transport; Post-translational modification, protein turnover, and chaperones; Function unknown
DIP1603 - DIP1596	8	–	not detected	*murE, murF, mraY, murD,* DIP1599, *murG, murC,* DIP1596. Cell wall, membrane, envelop biogenesis
DIP1720 - DIP1713	8	–	not detected	*dnaJ2,* DIP1719, DIP1718, DIP1717, DIP1716, DIP1715, *era, recO.* Replication, recombination and repair; Cell wall, membrane, envelop biogenesis; Coenzyme metabolism; Inorganic ion transport and metabolism; Function unknown; Signal transduction mechanisms; Post-translational modification, protein turnover, and chaperones
DIP1779 - DIP1772	8	–	not detected	*obgE,* DIP1778, *proB, proA, nadD,* DIP1774, DIP1773, DIP1772. Function unknown; Carbohydrate metabolism and transport; Coenzyme metabolism; Amino Acid metabolism and transport; Energy production and conversion
DIP1857 – DIP1850	8	–	detected	*clpS,* DIP1856, DIP1855, DIP1854, *murI,* DIP1852, *rph,* DIP1850. Nucleotide metabolism and transport; Translation, ribosomal structure and biogenesis; Function unknown; Cell wall, membrane, envelop biogenesis; Amino Acid metabolism and transport; Secondary metabolites biosynthesis, transport, and catabolism; Post-translational modification, protein turnover, and chaperones

motif on the DNA under iron excess conditions and thereby regulates the expression of genes coding for proteins involved in iron metabolism [3, 4]. It was shown experimentally in *C. glutamicum* that DtxR is a dual regulator which represses genes related to iron uptake but activates genes related to iron storage under iron excess conditions [23]. To analyze the dual characteristics of DtxR in *C. diphtheriae*, we compared the genome-wide transcription profile of a Δ*dtxR* mutant with that of the wild type strain. The mapped paired reads were counted with ReadXplorer [48] and the DESeq2 tool [56] was used to measure differential transcription (Additional file 9: Table S8). To assure that all three biological replicates are suitable for comparison, the normalized read counts calculated from DESeq2 were plotted against each other and the Pearson correlation coefficient R^2 was calculated. All replicates from both strains showed high R^2 values demonstrating the highly reproducible experimental set-up (Additional file 8: Figure S1).

Genes with a $\log_2$(fold change) (LFC) above +1.0 or below −1.0 and an adjusted *p*-value below 0.05 were considered as differentially transcribed. In total 224 genes showed elevated transcript levels and 113 genes decreased transcript levels in the mutant when compared with the wild type (Fig. 6 and Additional file 9: Table S8). The deletion of the dtxR gene had a remarkable influence on the transcriptome of the mutant strain affecting around 15% of all genes either directly or by indirect effects. The gene

with the largest $\log_2$(fold change) (LFC 6.28) is DIP2330 encoding a putative secreted protein of unknown function. Among the 40 genes with known or predicted DtxR binding sites, 25 (63%) were differentially transcribed in the Δ*dtxR* mutant. According to the state of differential transcription the genes with DtxR binding sites are either repressed or activated by DtxR (Table 3). The majority of the 25 differentially transcribed genes showed an enhanced transcription in the mutant strain and are therefore repressed by DtxR in the wild type under iron excess conditions. Among this group of genes are those encoding hemin receptors, iron transporters and iron siderophores. The genes *ftn*, *sdhB*, *narK* and *ycdA* are weakly transcribed in the mutant. These genes are therefore probably activated by DtxR in the wild type in an iron-rich condition. The iron storage gene *ftn* showed the lowest transcription in the *dtxR* mutant when compared with the wild type, which might underline the dual regulatory function of DtxR. Intriguingly, *srtC*, coding for the pilus-specific sortase SrtC involved in the assembly of the SpaD-type pili [2], was stronger transcribed in the absence of *dtxR* (Additional file 9: Table S8; LFC 1.27). No other sortase and pilin genes were observed as differentially transcribed in the Δ*dtxR* mutant.

The locations of binding sites and the respective TSS are in accordance with regulatory models. The DtxR binding sites of all DtxR-activated genes are located at least 37 bp upstream of the detected TSS. In contrast to
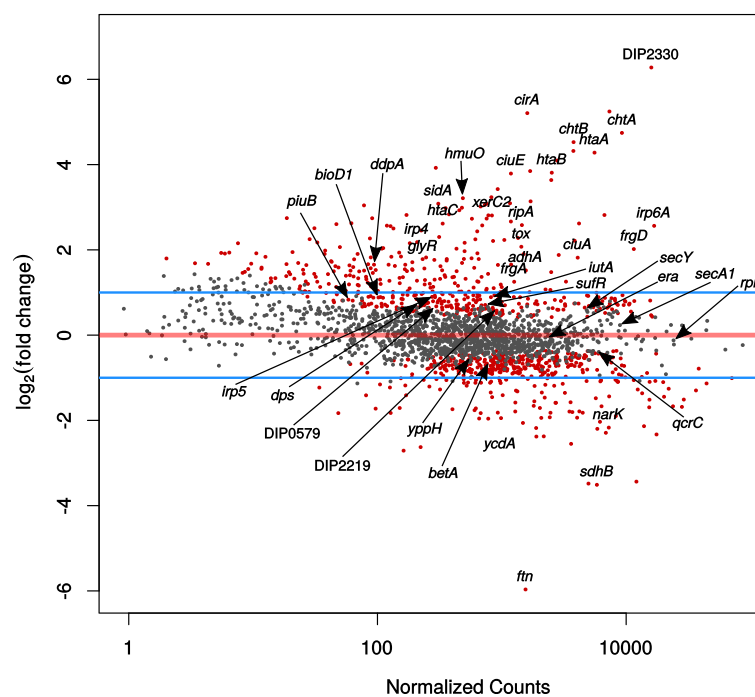
Fig. 6 M/A plot of differentially transcribed genes in the Δ*dtxR* mutant. Genes with an adjusted *p*-value below 0.05 are shown in red. The blue lines indicate the $\log_2$(fold change) threshold of +1.0 and −1.0, respectively. Genes with a known or predicted DtxR binding site are labeled

Table 3 Differential expression and σ^A promoter region of genes with known or predicted DtxR binding site. Based on the differential expression values the genes are classified into three groups: repressed, activated and not differentially expressed. In case no TSS could be assigned to the gene only the translation start codon is shown. In case a TSS could be assigned, but no −10 motif could be identified the start codon is preceded with a dotted line. Counts, normalized read counts; LFC, log$_2$(fold change); Start, start codon. The asterisks mark genes with an experimentally shown DtxR regulation

	Gene with DtxR binding site	Counts	LFC	Adjusted p-value	σ^A Promoter Region −35	−10	Start
Repressed	DIP2330	15933.8	6.28	2.7E-159	CCAGCT... N$_{18}$	...TAGGGT... N$_{386}$	...TTG
	DIP0523 / *cirA*^{a*}	1603.1	5.21	9.0E-101	....		GTG
	DIP1520 / *chtA**	9247.5	4.74	3.0E-053	TTAGGT... N$_{19}$	...TATGAT... N$_{50}$	...TTG
	DIP1519 / *chtB**	3776.1	4.53	4.7E-026	CTGGTT... N$_{18}$	...TAAGGT... N$_{13}$	...GTG
	DIP0625 / *htaA**	5566.6	4.28	9.1E-023	CTGGAC... N$_{18}$	...TAGCCT... N$_{154}$	...GTG
	DIP0629 / *htaB**	2522.4	3.81	2.9E-042	TTGAAA... N$_{17}$	...TAAGGT... N$_{41}$	...ATG
	DIP0586 / *ciuE*^{b*}	1184.3	3.79	3.6E-031			GTG
	DIP2161 / *sidA*^{a*}	308.4	3.08	4.0E-035	....		ATG
	DIP1669 / *hmuO**	478.0	2.99	2.4E-039	ATGTGA... N$_{17}$	...TATCTT... N$_{54}$	...ATG
	DIP1510 / *xerC2*	832.7	2.81	2.6E-040	TTGCGT... N$_{17}$	...TAATGT... N$_{217}$	...GTG
	DIP0624 / *htaC**	334.6	2.62	7.7E-004	................TAGGCT... N$_{53}$		...ATG
	DIP0922 / *ripA**	1450.5	2.58	5.1E-004	TTGCCT... N$_{17}$	...TAACCT... N$_{89}$	...GTG
	DIP0108 / *irp6A**	16792.9	2.56	6.6E-009	TTCTTT... N$_{17}$	...TATAGT... N$_{42}$	...GTG
	DIP2356 / *irp4**	208.5	2.18	5.6E-008	ATACGT... N$_{18}$	...TAACCT... N$_{51}$	...ATG
	DIP0222 / *tox*^{c*}	1435.8	2.07	5.3E-018	...............TAGGAT... N$_{38}$		...GTG
	DIP1062 / *frgD*^{c*}	11524.2	2.02	9.2E-008	CTAACC... N$_{19}$	...TAACCT... N$_7$	...ATG
	DIP0582 / *ciuA**	4077.3	1.82	3.8E-008	TTCCCC... N$_{18}$	...TACCTT... N$_8$	...ATG
	DIP0415 / *glyR*^{b}	230.8	1.78	5.8E-012			ATG
	DIP2162 / *ddpA*^{a*}	93.8	1.65	2.2E-006	....		ATG
	DIP2114 / *adhA*^{a*}	1561.2	1.52	1.6E-005	....		ATG
	DIP1061 / *frgA**	1257.1	1.25	5.3E-004	TTACCC... N$_{19}$	...TAGAAT... N$_6$	...ATG
	DIP1189 / *bioD1*^{a}	105.7	0.89	2.3E-002	....		GTG
	DIP0894 / *irp5*^{d*}	290.3	0.89	3.3E-002	TTCACC... N$_{18}$	...TACACT... N$_8$	...ATG
	DIP0169 / *iutA*	939.4	0.85	1.5E-002	CTAGCT... N$_{19}$	...TAAGAT... N$_{42}$	...GTG
	DIP0124 / *piuB*^{a*}	58.4	0.83	2.4E-002	....		ATG
Not differentially expressed	DIP1296 / *sufR**	772.4	0.66	1.1E-001	TTGGCG... N$_{18}$	...TAGTGT... N$_{104}$	...TTG
	DIP2303 / *dps*	211.2	0.65	7.4E-003	TTGCAC... N$_{16}$	...CACACT... N$_7$	...ATG
	DIP0579	269.8	0.56	7.0E-002	TTGCCT... N$_{18}$	...TAGGGT... N$_7$	...GTG
	DIP2219^a	906.5	0.52	3.4E-002	....		TTG
	DIP0540 / *secY**	4874.4	0.43	4.9E-002	CTAGTG... N$_{19}$	...TAGGGT... N$_{119}$	...GTG
	DIP0699 / *secA1*	8693.0	0.10	6.5E-001	TTGCTG... N$_{20}$	...TAGGCT... N$_{112}$	...GTG
	DIP0437 / *rplL*^{b}	24043.2	-0.12	6.3E-001			ATG
	DIP1714 / *era*^{d}	2424.0	-0.15	4.9E-001	...............TAACAT... N$_{48}$		...GTG
	DIP1626 / *qcrC*^{a}	6020.6	-0.39	6.7E-002	....		ATG
	DIP1473 / *yppH*^{d}	557.8	-0.61	5.6E-002	TCCGCT... N$_{18}$	...TATCCT... N$_{48}$	...ATG
	DIP2202 / *betA*	788.3	-0.81	2.9E-005	GTGGCA... N$_{18}$	...CACACT... N$_{125}$	...GTG
Activated	DIP1391 / *ycdA*^{c}	2322.6	-1.01	1.2E-007	TTGCTG... N$_{19}$	...TATATT... N$_{114}$	...TTG
	DIP0501 / *narK*	7249.6	-2.17	1.0E-006	TTAATT... N$_{16}$	...TAGCTT... N$_{94}$	...ATG
	DIP0370 / *sdhB*	5816.8	-3.51	1.1E-031	GTGGTC... N$_{19}$	...TAAAAT... N$_{21}$	...ATG
	DIP1866 / *ftn*	1551.5	-5.97	1.9E-045	CTGACA... N$_{17}$	...TAGCAT... N$_7$	...ATG

^{a}No TSS detected, hence no promoter motif predictable
^{b}Gene is part of a primary operon and lacks own TSS
^{c}Multiple TSS detected and only TSS closest to DtxR binding site shown. In case of the *tox* gene the −10 motif of the TSS (TSS 2) closest to the start codon is shown
^{d}Gene is the first one in a sub-operon and therefore has own TSS

that the DtxR binding sites of all DtxR-repressed genes overlap the −10 region of the σ^A promoter or are located downstream of the detected TSS. The mechanism of repression by DtxR most likely works by simply covering the promoter site and thereby preventing the RNA polymerase from binding or by roadblocking which forces the RNA polymerase to halt prematurely [74]. Binding of DtxR upstream of a promoter seems to have an activating effect on gene transcription but it is unclear how the mechanism of gene activation works. Nevertheless for some genes with known or predicted DtxR binding site no TSS was detected in this study presumably due to low transcription (Table 3).

Comprehensive transcriptomic view on the phage island and the tox gene encoding diphtheria toxin

C. diphtheriae NCTC 13129 is a *tox*⁺ strain as it carries a corynephage that harbors the diphtheriae toxin gene *tox*. The diphtheriae toxin (DT) is one of the strongest bacterial toxins and essential for the pathogenicity of *C. diphtheriae* [73, 75]. Here we use the primary 5′-end data and the whole transcriptome data to gain a comprehensive view on the transcriptional features of the phage island and the *tox* gene region in particular.

The phage island is located between the tRNA^Arg genes DIPt10 and DIPt11 and consists of 44 genes (from DIP0180 to DIP0222/*tox*). Only six TSS were assigned to corynephage genes. One of these genes is transcribed leaderless (DIP0180), while the others have varying 5′-UTR lengths ranging from 28 nt to 413 nt. Furthermore, six TSS assigned to putative antisense transcripts were detected. By operon analysis, the phage island is transcriptionally structured in 7 primary operons containing 34 genes and one sub-operon containing one gene (DIP0197), leaving the remaining 10 genes transcribed as monocistrons (Fig. 7a).

Interestingly, the transcription of genes in the middle of the phage island is relatively low compared to genes
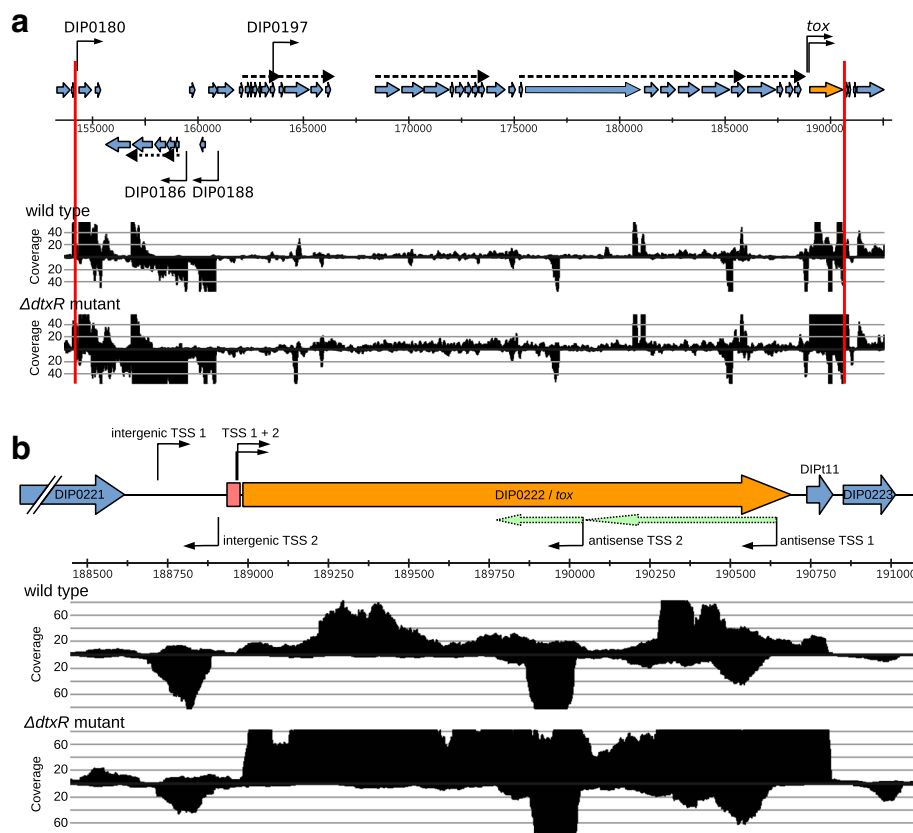
Fig. 7 Transcription profile of the phage island and the *tox* gene. **a** The transcription profile of the phage island (framed by two red lines) of the wild type and the isogenic Δ*dtxR* mutant is shown. The genes which could be assigned with a TSS are labeled with a black arrow. Primary operon structures are indicated by dashed arrows. The *tox* gene is shown in dark yellow. TSS assigned to intergenic and antisense transcripts were omitted for clarity. **b** Detailed view on the monocistronic *tox* gene. It possesses two TSS (TSS 1 and TSS 2) with 5′-UTR lengths of 43 nt and 38 nt. The DtxR binding site (red box) is located close to the start codon of the CDS and overlaps both TSS and their corresponding −10 promoter motifs. Two additional TSS for putative antisense RNAs were found (light green arrows) and two additional TSS (intergenic TSS 1 and intergenic TSS 2) on the forward and reverse strand, respectively, were identified upstream of the *tox* gene, indicating two putative novel transcripts. The image is a modified screenshot from ReadXplorer [48] showing non-normalized coverage data of one exemplary replicate per condition

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 14 of 17

in the exterior regions of the island. This is the case for both wild type and Δ*dtxR* mutant strains. However, 20 genes (45%) in the phage island are differentially transcribed in the Δ*dtxR* mutant. These genes do not cluster in a specific region, as genes in the exteriors and the middle part of the phage island are affected. Although the *tox* gene is the only gene in the phage island with a DtxR binding site, additional 19 corynephage genes are stronger transcribed in the Δ*dtxR* mutant. Various genes encoding for phage repression and capsid assembly as well as the putative phage integrase and a putative transcription regulator are among these genes.

Although the *tox* gene is part of the corynephage genome, its expression is under bacterial control. Upstream of the *tox* gene a DtxR binding site is located that is blocked under high-iron conditions by the regulator protein DtxR encoded on the corynebacterial chromosome [16–18]. Many studies focus on the structure of the DT protein [6–8, 76] or its domains [9, 77–79] but only a few studies addressed the transcriptomic characteristics of the *tox* encoding gene [24, 25, 80].

In front of the *tox* coding region two TSS (named TSS 1 and TSS 2) with predicted σ^A promoter motifs were detected (Additional file 3: Table S3), resulting in 5'-UTRs of 43 nt and 38 nt respectively (Fig. 7b). The DtxR binding site is located 32 bp upstream of the start codon and overlaps the −10 promoter motif of both TSS. As expected, the transcription of *tox* was increased > 4-fold in the Δ*dtxR* mutant when compared with the wild type. In early studies of the *tox* transcription two overlapping promoters resulting in two TSS at positions −38 and −43 relative to the GTG start codon were found [25, 81]. As shown by site-directed mutagenesis the −10 motif proximal to the start codon is more active than the other [25]. Our primary 5'-end data supports these findings as the number of read starts at the TSS closer to the start codon (TSS 2) is ~ 5-fold higher compared to the distal TSS (TSS 1) (data not shown).

In addition to the TSS of the *tox* coding region, two putative antisense RNA (asRNA) and their TSS as well as two intergenic TSS were deduced from the transcriptome data (Fig. 7b). To the best of the authors' knowledge this is the first description of *tox* related asRNA in *C. diphtheriae*. The two asRNA start close to the end of the *tox* CDS (asRNA TSS 1) and approximately in the middle of the CDS (asRNA TSS 2). The lengths estimated from overlapping reads are 580 nt for asRNA 1 and 260 nt for asRNA 2. Antisense RNAs have a broad range of functions effecting transcription, stability or translation of the sense mRNA [82]. However the function of the identified antisense RNAs is not known at this point.

The two intergenic TSS (intergenic TSS 1 and 2, Fig. 7b) indicate the presence of novel transcripts upstream of the *tox* gene. Both novel transcripts are on opposing strands facing each other. While the coverage of the transcript of the reverse strand is clearly visible, the coverage of the other transcript on the forward strand is very low. Both the BLAST search of potential ORFs and the ncRNA prediction by Infernal/RFAM gave any hints on possible functions. Although the *tox* gene regulation by DtxR is known for decades, our findings illustrate that the transcriptional landscape of this gene region is far more complex and still compelling. Further research is needed to shed light onto the complex transcriptional patterns in the *tox* gene region.

Conclusion

This study comprises the first reported whole transcriptome and transcription start site (TSS) analysis of *C. diphtheriae* NCTC 13129. We provide a comprehensive list of TSS, promoter motifs and ribosomal binding sites as well as 5'-UTRs for the majority of genes. Furthermore, we corrected several predicted coding regions based on the experimentally detected TSS data and found hundreds of putative novel transcripts. By combining the whole transcriptome with the 5´-enriched cDNA library data operon and sub-operon structures were predicted. In addition, differential gene expression analysis of a *dtxR* deletion mutant was performed to identify the global effects of DtxR regulation that includes the diphtheria toxin gene *tox*. As the *tox* gene is a major factor contributing to the pathogenicity of *C. diphtheriae* we present a detailed analysis of the transcriptional landscape of this important gene region.

The findings presented here greatly expand the understanding of transcript regulation and provide a solid foundation for further transcriptome studies of this important pathogen. In particular the cornerstone was laid for in depth analyses of promoters motifs and for transcriptional analysis of host pathogen interactions. Furthermore, future studies might be focused on small and antisene RNAs, which could harbor new regulatory elements and functions.

Additional files

Additional file 1: Table S1. Oligonucleotide sequences. (XLSX 5 kb)

Additional file 2: Table S2. Number of cDNA reads of the cDNA libraries. (XLSX 5 kb)

Additional file 3: Table S3. List of transcription start sites assigned to known CDS. (XLSX 70 kb)

Additional file 4: Table S4. List of intragenic and antisense TSS. (XLSX 15 kb)

Additional file 5: Table S5. List of corrected CDS start sites. (XLSX 9 kb)

Additional file 6: Table S6. List of intergenic TSS. (XLSX 8 kb)

Additional file 7: Table S7. List of operons, sub-operons and monocistrons. (XLSX 38 kb)

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 15 of 17

Additional file 8: Figure S1. Reproducibility of differential expression analysis with varying cDNA library replicates. (PDF 1397 kb)

Additional file 9: Table S8. List of differentially transcribed genes. (XLSX 418 kb)

Availability of data and materials

The data sets supporting the conclusions of this article are available in the NCBI Gene Expression Omnibus repository, under the accession number GSE98202, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE98202

Authors' contributions

MW created the whole transcriptome cDNA libraries, conducted all bioinformatic analyses and drafted the manuscript. TB and MW prepared the primary 5′-end cDNA library. AHG constructed the ΔdtxR mutant. AHG and JHL performed the cultivation and RNA isolation. AT, JK and HTT coordinated the study and participated in data analysis. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The two strains *Corynebacterium diphtheriae* NCTC 13129 and *Corynebacterium diphtheriae* NCTC 13129 Δ*dtxR* were used in this study. The first strain derived from the National Collection of Tissue Cultures which is operated by Public Health England (Salisbury, United Kingdom). The strain number in this culture collection is NCTC 13129. The second strain derived from the first strain via deletion of the gene *dtxR*.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

^{1}Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany. ^{2}Institute for Biology-Microbiology, Freie Universität Berlin, D-14195 Berlin, Germany. ^{3}Department of Molecular Biology and Biophysics, University of Connecticut Health Center, Farmington, CT, USA. ^{4}Department of Microbiology & Molecular Genetics, University of Texas McGovern Medical School at Houston, Houston, USA. ^{5}Present address: Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, 181 Ipsin-gil, Jeollabuk-do 56212, Republic of Korea.

References

1. Cerdeño-Tárraga AM, Efstratiou A, Dover LG, Holden MTG, Pallen M, Bentley SD, Besra GS, Churcher C, James KD, De Zoysa A, Chillingworth T, Cronin A, Dowd L, Feltwell T, Hamlin N, Holroyd S, Jagels K, Moule S, Quail MA, Rabbinowitsch E, Rutherford KM, Thomson NR, Unwin L, Whitehead S, Barrell BG, Parkhill J. The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. Nucleic Acids Res. 2003;31:6516–23.
2. Gaspar AH, Ton-That H. Assembly of distinct pilus structures on the surface of Corynebacterium diphtheriae. J Bacteriol. 2006;188:1526–33.
3. Trost E, Blom J, SdC S, Huang I-H, Al-Dilaimi A, Schröder J, Jaenicke S, Dorella FA, Rocha FS, Miyoshi A, Azevedo V, Schneider MP, Silva A, Camello TC, Sabbadini PS, Santos CS, Santos LS, Hirata R Jr, Mattos-Guaraldi AL, Efstratiou A, Schmitt MP, Ton-That H, Tauch A. Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol. 2012;194:3199–215.
4. Merchant AT, Spatafora GA. A role for the DtxR family of metalloregulators in gram-positive pathogenesis. Mol Oral Microbiol. 2014;29:1–10.
5. Allen CE, Schmitt MP. Utilization of host iron sources by Corynebacterium diphtheriae: multiple hemoglobin-binding proteins are essential for the use of iron from the hemoglobin-haptoglobin complex. J Bacteriol. 2015;197:553–62.
6. Kantardjieff K, Collier RJ, Eisenberg D. X-ray grade crystals of the enzymatic fragment of diphtheria toxin. J Biol Chem. 1989;264:10402–4.
7. Choe S, Bennett MJ, Fujii G, Curmi PM, Kantardjieff KA, Collier RJ, Eisenberg D. The crystal structure of diphtheria toxin. Nature. 1992;357:216–22.
8. Louie GV, Yang W, Bowman ME, Choe S. Crystal structure of the complex of diphtheria toxin with an extracellular fragment of its receptor. Mol Cell. 1997;1:67–78.
9. Li J, Rodnin MV, Ladokhin AS, Gross ML. Hydrogen-deuterium exchange and mass spectrometry reveal the pH-dependent conformational changes of diphtheria toxin T domain. Biochemistry. 2014;53:6849–56.
10. Hoch DH, Romero-Mira M, Ehrlich BE, Finkelstein A, DasGupta BR, Simpson LL. Channels formed by botulinum, tetanus, and diphtheria toxins in planar lipid bilayers: relevance to translocation of proteins across membranes. Proc Natl Acad Sci U S A. 1985;82:1692–6.
11. Ladokhin AS. pH-triggered conformational switching along the membrane insertion pathway of the diphtheria toxin T-domain. Toxins. 2013;5:1362–80.
12. Ton-That H, Schneewind O. Assembly of pili on the surface of Corynebacterium diphtheriae. Mol Microbiol. 2003;50:1429–38.
13. Mandlik A, Swierczynski A, Das A, Ton-That H. Corynebacterium diphtheriae employs specific minor pilins to target human pharyngeal epithelial cells. Mol Microbiol. 2007;64:111–24.
14. Broadway MM, Rogers EA, Chang C, Huang I-H, Dwivedi P, Yildirim S, Schmitt MP, Das A, Ton-That H. Pilus gene pool variation and the virulence of Corynebacterium diphtheriae clinical isolates during infection of a nematode. J Bacteriol. 2013;195:3774–83.
15. Reardon-Robinson ME, Osipiuk J, Chang C, Wu C, Jooya N, Joachimiak A, Das A, Ton-That H. A disulfide bond-forming machine is linked to the Sortase-mediated Pilus assembly pathway in the gram-positive bacterium Actinomyces Oris. J Biol Chem. 2015;290:21393–405.
16. Boyd J, Oza MN, Murphy JR. Molecular cloning and DNA sequence analysis of a diphtheria tox iron-dependent regulatory element (dtxR) from Corynebacterium diphtheriae. Proc Natl Acad Sci U S A. 1990;87:5968–72.
17. Schmitt MP, Holmes RK. Iron-dependent regulation of diphtheria toxin and siderophore expression by the cloned Corynebacterium diphtheriae repressor gene dtxR in C. Diphtheriae C7 strains. Infect Immun. 1991;59:1899–904.
18. Tao X, Schiering N, Zeng HY, Ringe D, Murphy JR. Iron, DtxR, and the regulation of diphtheria toxin expression. Mol Microbiol. 1994;14:191–7.
19. Schmitt MP. Transcription of the Corynebacterium diphtheriae hmuO gene is regulated by iron and heme. Infect Immun. 1997;65:4634–41.
20. Qian Y, Lee JH, Holmes RK. Identification of a DtxR-regulated operon that is essential for siderophore-dependent iron uptake in Corynebacterium diphtheriae. J Bacteriol. 2002;184:4846–56.
21. Kunkle CA, Schmitt MP. Analysis of the Corynebacterium diphtheriae DtxR regulon: identification of a putative siderophore synthesis and transport system that is similar to the Yersinia high-pathogenicity island-encoded yersiniabactin synthesis and uptake system. J Bacteriol. 2003;185:6826–40.

Wittchen *et al. BMC Genomics* (2018) 19:82

Page 16 of 17

22. Kunkle CA, Schmitt MP. Analysis of a DtxR-regulated iron transport and siderophore biosynthesis gene cluster in Corynebacterium diphtheriae. J Bacteriol. 2005;187:422–33.

23. Brune I, Werner H, Hüser AT, Kalinowski J, Pühler A, Tauch A. The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of Corynebacterium glutamicum. BMC Genomics. 2006;7:21.

24. Leong D, Murphy JR. Characterization of the diphtheria tox transcript in Corynebacterium diphtheriae and Escherichia Coli. J Bacteriol. 1985;163:1114–9.

25. Boyd J, Murphy JR. Analysis of the diphtheria tox promoter by site-directed mutagenesis. J Bacteriol. 1988;170:5949–52.

26. Lee JH, Wang T, Ault K, Liu J, Schmitt MP, Holmes RK. Identification and characterization of three new promoter/operators from Corynebacterium diphtheriae that are regulated by the diphtheria toxin repressor (DtxR) and iron. Infect Immun. 1997;65:4273–80.

27. Oram DM, Jacobson AD, Holmes RK. Transcription of the contiguous sigB, dtxR, and galE genes in Corynebacterium diphtheriae: evidence for multiple transcripts and regulation by environmental factors. J Bacteriol. 2006;188:2959–73.

28. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci U S A. 1977;74:5350–4.

29. Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, Lind K, Sindelka R, Sjöback R, Sjögreen B, Strömbom L, Ståhlberg A, Zoric N. The real-time polymerase chain reaction. Mol Asp Med. 2006;27:95–125.

30. Frohman MA. On beyond classic RACE (rapid amplification of cDNA ends). PCR Methods Appl. 1994;4:S40–58.

31. Taub F, DeLeo JM, Thompson EB. Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. DNA. 1983;2:309–27.

32. Chen YA, Chou C-C, Lu X, Slate EH, Peck K, Xu W, Voit EO, Almeida JS. A multivariate prediction model for microarray cross-hybridization. BMC Bioinformatics. 2006;7:101.

33. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

34. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

35. Dornenburg JE, Devita AM, Palumbo MJ and Wade JT. Widespread antisense transcription in *Escherichia coli*. mBio. 2010;1.

36. Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 2008;9:R175.

37. Raghavan R, Groisman EA, Ochman H. Genome-wide detection of novel regulatory RNAs in E. Coli. Genome Res. 2011;21:1487–97.

38. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J. The primary transcriptome of the major human pathogen helicobacter pylori. Nature. 2010;464:250–5.

39. Pfeifer-Sancar K, Mentz A, Rückert C, Kalinowski J. Comprehensive analysis of the Corynebacterium glutamicum transcriptome using an improved RNAseq technique. BMC Genomics. 2013;14:888.

40. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus salmonella typhi. PLoS Genet. 2009;5:e1000569.

41. Irla M, Neshat A, Brautaset T, Rückert C, Kalinowski J, Wendisch VF. Transcriptome analysis of thermophilic methylotrophic bacillus methanolicus MGA3 using RNA-sequencing provides detailed insights into its previously uncharted transcriptional landscape. BMC Genomics. 2015;16:73.

42. Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: a database for prokaryotic operons. Nucleic Acids Res. 2009;37:D459–63.

43. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Del Moral-Chávez V, Rinaldi F, Collado-Vides J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016;44:D133–43.

44. Schäfer A, Tauch A, Jäger W, Kalinowski J, Thierbach G, Pühler A. Small mobilizable multi-purpose cloning vectors derived from the Escherichia Coli plasmids pK18 and pK19: selection of defined deletions in the chromosome of Corynebacterium glutamicum. Gene. 1994;145:69–73.

45. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–120.

46. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.

47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000GPDP. The sequence alignment/map format and SAMtools. Bioinformatics (Oxford, England). 2009;25:2078–9.

48. Hilker R, Stadermann KB, Doppmeier D, Kalinowski J, Stoye J, Straube J, Winnebald J, Goesmann A. ReadXplorer–visualization and analysis of mapped sequences. Bioinformatics. 2014;30:2247–54.

49. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28:1166–7.

50. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.

51. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schäffer AA, Yu Y-K. Protein database searches using compositionally adjusted substitution matrices. FEBS J. 2005;272:5101–9.

52. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43:D130–7.

53. Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. Science. 2004;305:1743–6.

54. Crooks GE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14:1188–90.

55. Albersmeier A, Pfeifer-Sancar K, Rückert C and Kalinowski J. Genome-wide determination of transcription start sites reveals new insights into promoter structures in the actinomycete Corynebacterium glutamicum. J biotechnol. 2017;257:99–09.

56. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

57. Burgos JM, Schmitt MP. The ChrA response regulator in Corynebacterium diphtheriae controls hemin-regulated gene expression through binding to the hmuO and hrtAB promoter regions. J Bacteriol. 2012;194:1717–29.

58. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, Huang X, Han B. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res. 2010;20:1238–49.

59. van Vliet and Arnoud H M. Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol Lett. 2010;302:1-7.

60. Pátek M, Nešvera J, Guyonvarch A, Reyes O, Leblon G. Promoters of Corynebacterium glutamicum. J Biotechnol. 2003;104(1–3):311–23.

61. Pátek M, Nešvera J. Sigma factors and promoters in Corynebacterium glutamicum. J Biotechnol. 2011;154:101–13.

62. Zheng X, Hu G-Q, She Z-S, Zhu H. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics. 2011;12:361.

63. Winkler WC, Breaker RR. Regulation of bacterial gene expression by riboswitches. Annu Rev Microbiol. 2005;59:487–517.

64. Meyer MM. The role of mRNA structure in bacterial translational regulation. Wiley interdisciplinary reviews. RNA. 2017;8

65. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.

66. Bertram R, Schlicht M, Mahr K, Nothaft H, Saier MH, Titgemeyer F. In silico and transcriptional analysis of carbohydrate uptake systems of Streptomyces Coelicolor A3(2). J Bacteriol. 2004;186:1362–73.

67. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol. 2010;11:R31.

68. Kolter R, Yanofsky C. Attenuation in amino acid biosynthetic operons. Annu Rev Genet. 1982;16:113–34.

69. Löchelt M, Muranyi W, Flügel RM. Human foamy virus genome possesses an internal, Bel-1-dependent and functional promoter. Proc Natl Acad Sci U S A. 1993;90:7317–21.

70. Schoenfeld A, Davidowitz EJ, Burk RD. A second major native von Hippel-Lindau gene product, initiated from an internal translation start site, functions as a tumor suppressor. Proc Natl Acad Sci U S A. 1998;95:8817–22.

71. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T,

Harrow J, Guigo R, Gingeras TR, Antonarakis SE, Reymond A. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res. 2007;17:746–59.

72. Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, Mao F, Lai W, Xu Y. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res. 2014;42:D654–9.

73. Baron S. Medical Microbiology. Galveston (TX): University of Texas Medical Branch at Galveston; 1996.

74. Rodionov DA. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem Rev. 2007;107:3467–97.

75. Freeman VJ. Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. J Bacteriol. 1951;61:675–88.

76. Bennett MJ, Choe S, Eisenberg D. Refined structure of dimeric diphtheria toxin at 2.0 a resolution. Protein Sci. 1994;3:1444–63.

77. Weiss MS, Blanke SR, Collier RJ, Eisenberg D. Structure of the isolated catalytic domain of diphtheria toxin. Biochemistry. 1995;34:773–81.

78. Man P, Montagner C, Vitrac H, Kavan D, Pichard S, Gillet D, Forest E, Forge V. Accessibility changes within diphtheria toxin T domain when in the functional molten globule state, as determined using hydrogen/deuterium exchange measurements. FEBS J. 2010;277:653–62.

79. Trujillo C, Taylor-Parker J, Harrison R, Murphy JR. Essential lysine residues within transmembrane helix 1 of diphtheria toxin facilitate COPI binding and catalytic domain entry. Mol Microbiol. 2010;76:1010–9.

80. Costa JJ, Michel JL, Rappuoli R, Murphy JR. Restriction map of corynebacteriophages beta c and beta vir and physical localization of the diphtheria tox operon. J Bacteriol. 1981;148:124–30.

81. Kaczorek M, Zettlmeissl G, Delpeyroux F, Streeck RE. Diphtheria toxin promoter function in Corynebacterium diphtheriae and Escherichia Coli. Nucleic Acids Res. 1985;13:3147–59.

82. Thomason MK, Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? Annu Rev Genet. 2010;44:167–88.