Integrated repeat analysis and metadata report for *Bacillus thuringiensis* CTC

Table of contents

- General information
- Genome assembly information
- Species and taxonomy
- Annotation information
- Link to ProGenomes
- Information from StrainSelect
- Potential QC issues
- Results of repeat analyses
- 16S rRNA gene analysis
- Predicted biosynthetic gene clusters (antiSMASH)
- Proteogenomics database (basis to improve genome annotation)

General information

Several of the information fields provided below were parsed from the respective Genbank file. We also include information derived from additional analyses that we conducted.

NCBI database:	RefSeq	**Used annotation date:**	06-02-2023
Assembly name:	ASM145534v1	**Assembly accession:**	GCF_001455345.1
Assembly date:	2015-12-02	**Biosample accession:**	SAMN04278951

Sequencing technology: short read only
Reason for RefSeq exclusion: -

Genome assembly information

Total genome size (bp):	5,352,926
Number of genomic elements:	2
Number of plasmids:	1
Number of linear elements:	0

Overview of genomic elements (ordered as they appear in the GBK or FASTA file):
Genome coverage: 150x

Accession	**NZ_CP013274.1**	**NZ_CP013273.1**
Length (bp)	5,327,397	25,529
Circular/linear	circular	circular
Chr/plasmid	strain CTC chromosome	strain CTC plasmid unnamed
Reference assembly	GCA_000190515.1	GCA_000190515.1

Genome Compendium

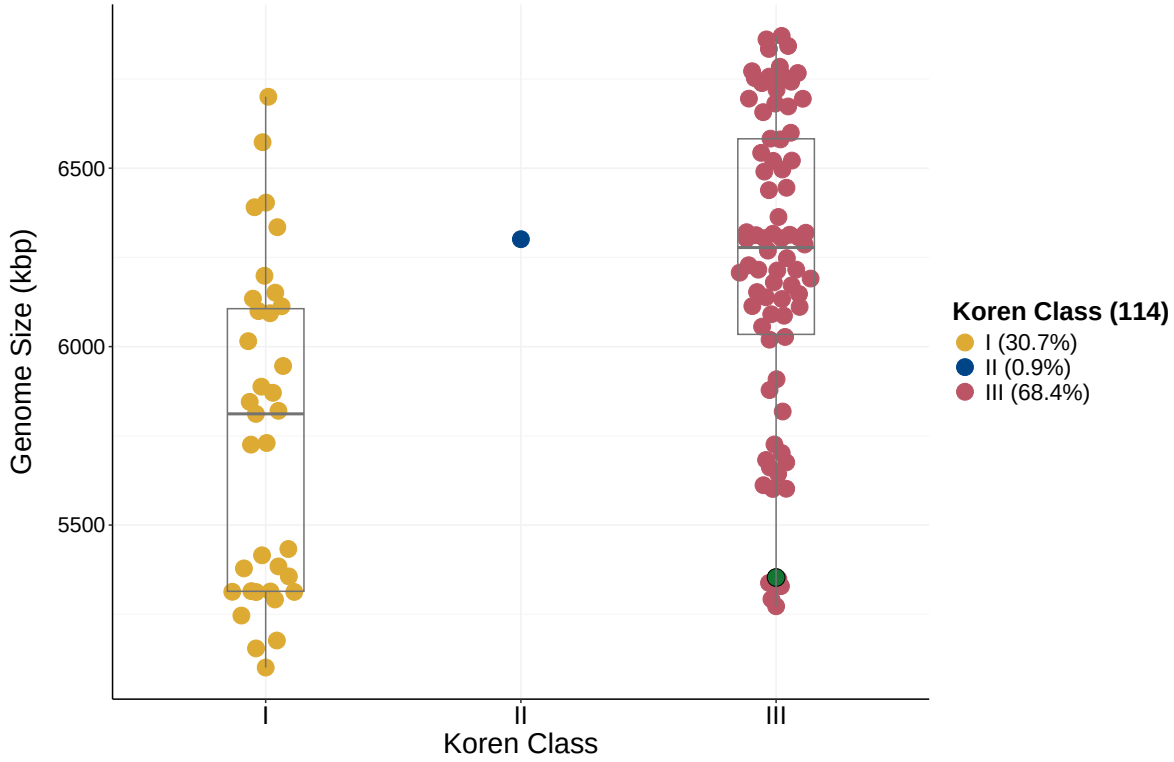

Figure 1: Overview of genome size distributions for the species *Bacillus thuringiensis* with respect to three classes of genome assembly complexity (Koren et al., 2013). The genome size of strain CTC is shown in green.

Species and taxonomy

Species:	*Bacillus thuringiensis*
Strain:	CTC
NCBI taxon ID:	1428

Taxonomy and number of total assemblies:

Taxonomic rank	Refseq	Genbank
Bacteria	55,436	19,664
Bacillota	12,876	3,520
Bacilli	11,447	2,995
Bacillales	6,743	1,479
Bacillaceae	2,698	379
Bacillus	2,168	334
Bacillus thuringiensis	105	9

Annotation information

Number of genes:	5,527	**Total number of CDS:**	5,396
Number of coding CDS:	5,260	**Number of pseudogenes:**	136
Number of rRNAs:	11, 11, 11 (5S, 16S, 23S)	**Number of tRNAs:**	93
Number of ncRNAs:	5		

Link to ProGenomes

ProGenomes is a very useful resource that provides consistent annotations for both complete and fragmented genome assemblies of prokaryotes. As annotations are sometimes not available in summary format from selected Refseq/Genbank assemblies, we also provide the link to the identical assembly from ProGenomes (if available).

https://progenomes.genome-compendium.com/GCF_001455345.1

Information from StrainSelect

StrainSelect is a database providing taxonomic, functional, and procurement information for bacterial strains (if available). It includes genome annotations and identifies strains available from biological resource centers for experimental use.

Not available.

Potential QC issues

We advise researchers interested in assemblies with a reported potential quality flag to assess the provided information in the respective Genbank and/or assembly information file. Genomes that are based on elements of a reference-guided assembly strategy may well miss relevant genomic information that a de novo assembly strategy could provide. Genomes assembled solely from short reads but containing long repeats may harbor assembly errors.

Reference-based assembly:	**Yes**
Short-read only assembly, long repeats:	**Yes**
Directional terminal repeats on linear contig:	**No**

Results of repeat analyses

Assembly complexity class:	3
Number of repeat regions:	44
Total repeat region content:	1.88%
Longest repeat region:	8,247

Repeat type	Number of repeat regions	Longest repeat	Median length
Interspersed	28	5,404	1,955
Tandem	16	8,247	708
Terminal inverted	0	0	0
Terminal directional	0	0	0

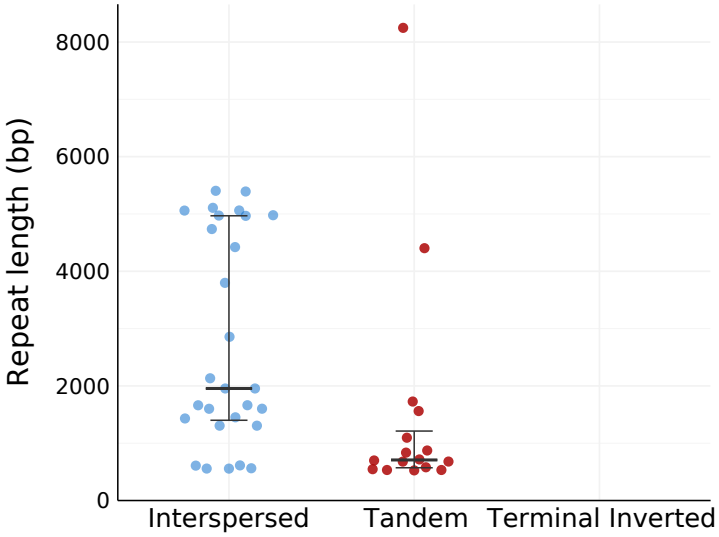

Figure 2: Swarm plot showing the length distribution for three repeat classes. Terminal directional repeats are not shown (QC aspect).

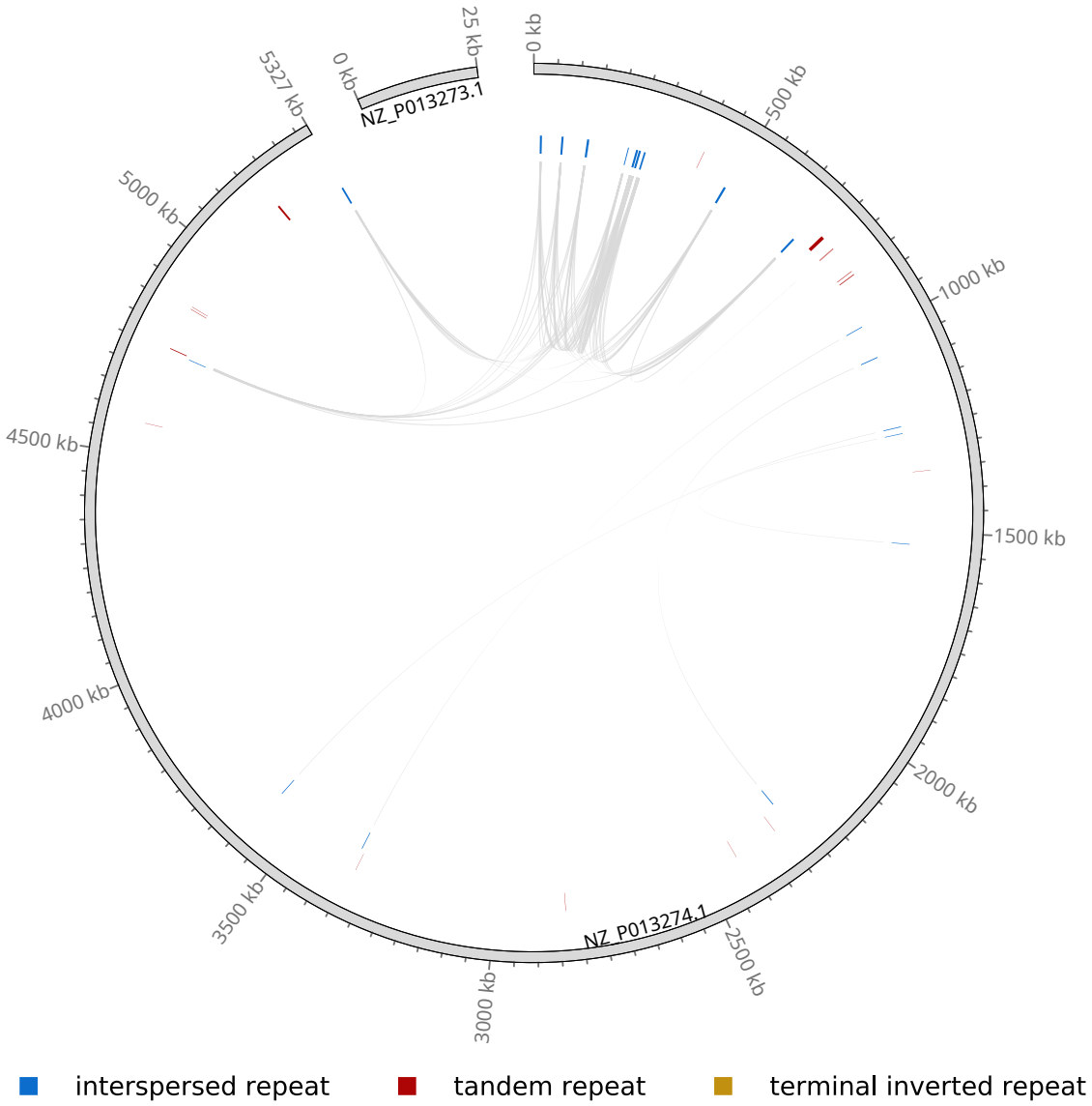

Figure 3: Circos plot visualizing three repeat classes and their genomic location for *Bacillus thuringiensis* CTC. Interspersed repeats are connected by gray bands.

16S rRNA gene analysis

Accession	Start Coordinate	Stop Coordinate	Strand	Length	Contig Type
NZ_CP013274.1	15,006	16,558	1	1,552	chrom
NZ_CP013274.1	68,289	69,841	1	1,552	chrom
NZ_CP013274.1	131,468	133,020	1	1,552	chrom
NZ_CP013274.1	234,043	235,595	1	1,552	chrom
NZ_CP013274.1	254,726	256,278	1	1,552	chrom
NZ_CP013274.1	261,926	263,478	1	1,552	chrom
NZ_CP013274.1	274,950	276,490	1	1,540	chrom
NZ_CP013274.1	489,167	490,719	1	1,552	chrom
NZ_CP013274.1	700,745	702,317	1	1,572	chrom
NZ_CP013274.1	4,754,512	4,756,065	-1	1,553	chrom
NZ_CP013274.1	5,322,791	5,324,343	1	1,552	chrom

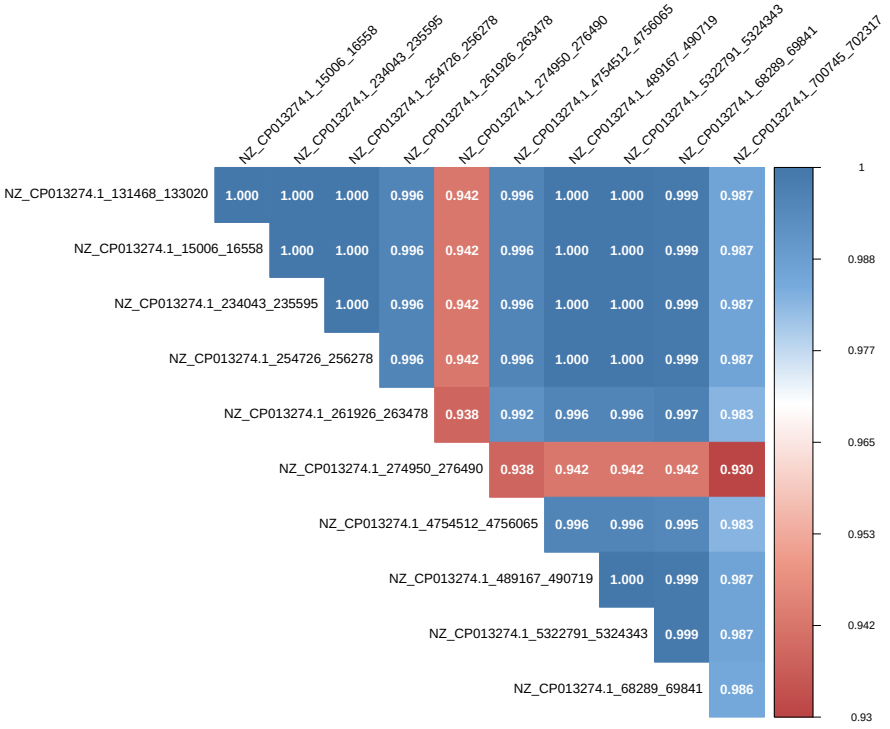

Figure 4: Overview of percent sequence identities for all global optimal pairwise sequence alignments (VSEARCH), of the encoded 16S rRNA genes shown in a heatmap for *Bacillus thuringiensis* CTC (self-alignments are excluded).

Predicted biosynthetic gene clusters (antiSMASH)

Cluster Type	Count
NRPS	2
RiPP-like	2
azole-containing-RiPP	1
betalactone	1
NI-siderophore	1
NRP-metallophore	1
terpene	1
terpene-precursor	1
Total	**10**

BGC ID	Count
BGC0000032.5	1
BGC0000300.5	1
BGC0000309.5	1
BGC0000343.5	1
BGC0000359.5	1
BGC0000366.5	1
BGC0000368.5	1
BGC0000371.5	1
BGC0000392.5	1
BGC0000401.5	1
BGC0000407.5	1
BGC0000454.5	1
BGC0000459.5	1
BGC0000460.5	1
BGC0000607.5	1
...	...
Total	**56**

Proteogenomics database

This integrated proteogenomics database (iPtgxDB, https://iptgxdb.expasy.org/index/) enables the identification of genes missed in the NCBI RefSeq annotation with the help of proteomics data. Mass spectrometry data can be directly searched against this database using search engines such as MSFragger, identifying expressed pseudogenes, corrected start sites for annotated genes and completely novel genes missing from the current annotation. Help on identifying novelty from the search results can be found here.

The iPtgxDB was created by using the NCBI RefSeq annotation as a basis and integrating gene predictions from Prodigal 2.6.3 and GMSCMapper 0.1.0. The GenBank files of each annotation were converted to GFF format using iPtgxDB_convert.jar (https://iptgxdb.expasy.org/software/), requiring the GMSCMapper results to be converted to GenBank format using a custom Python script beforehand. The steps outlined here were then followed to obtain an iPtgxDB from these GFF files. You can download the full results from https://genome-compendium.com/download_iptgxdb/GCF_001455345.1

annotation	clusters	entries	clusters +	entries +	reductions	extensions
RefSeq (refseq)	5396	5396	5261	5261	0	0
Prodigal (prod)	5389	5389	185	352	145	67
GMSC (ref)	811	811	30	32	0	3
TOTAL	-	-	5476	5645	145	70

Summary table: Composition of iPtgxDB. The number of clusters (proteins grouped by same stop codon location), entries (total number of proteins), and how many are new compared to the annotations in the rows above, as well as the number of novel start sites that reduce or extend a previous annotation are detailed in the other columns.

Reason	RefSeq	Prodigal	GMSC	TOTAL
too small	0	65	4	69
all-pseudo	135	0	0	135
indistinguishable start	0	97	15	112

Excluded table: Overview table of predictions that were not included. Reasons for the exclusion include peptides that are too small to be detected using a mass spectrometer (<6aa), pseudogene-only entries that could lead to excess ambiguous peptides, and internal start sites that cannot be distinguished from the RefSeq-annotated proteoform.

The annotations included in this iPtgxDB were selected to enable researchers to identify novelty based on a small set of predictions. These small custom iPtgxDBs are similar in size to a RefSeq protein search database and have specific advantages (better search statistics). Larger iPtgxDBs (20–40× larger), containing more annotations and in silico predictions that may identify additional novelty, can be generated using the webserver. (https://iptgxdb.expasy.org/iptgxdb/submit/).