



TrueBac™ ID - Genome

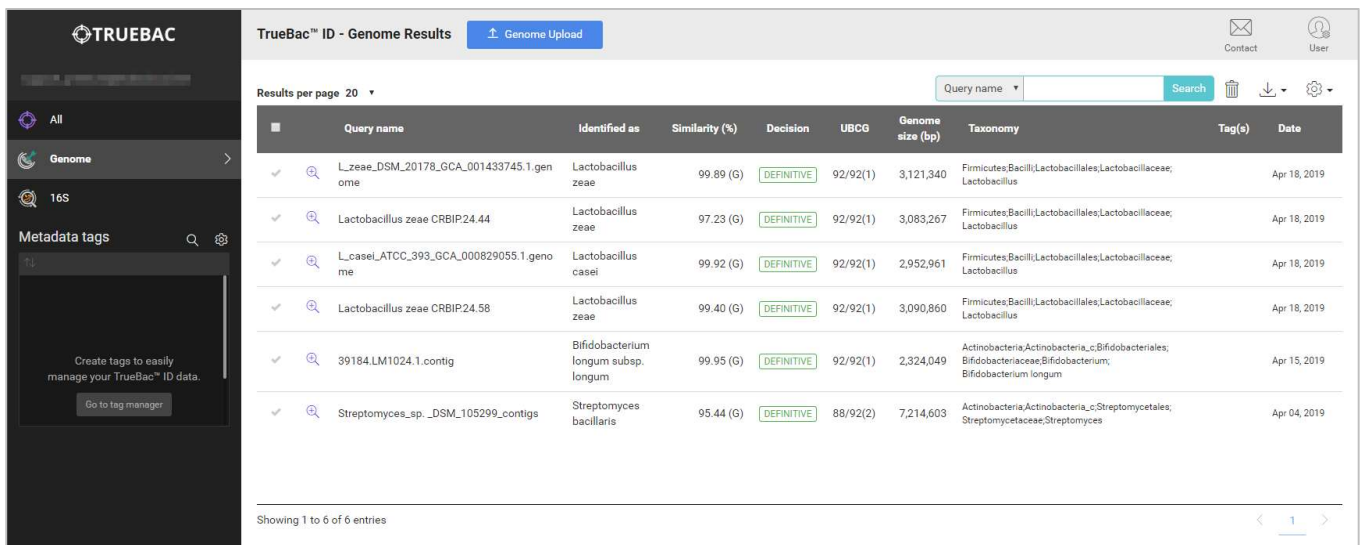
The following document contains basic instructions for the use of TrueBac™ ID – Genome

How to access TrueBac™ ID – Genome



Genome-based Identification in Action

- Go to TrueBac ID (<https://www.truebacid.com>) and login.
- On the main page, click on [Genome(s)] button under Data Center.



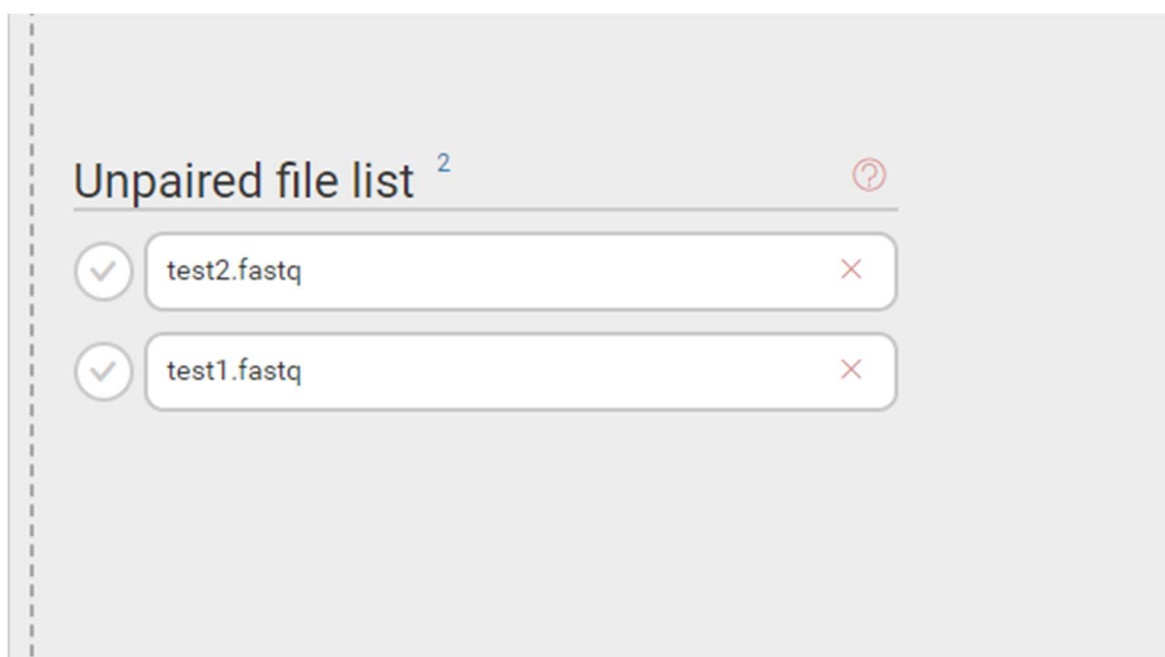
Query name	Identified as	Similarity (%)	Decision	UBCG	Genome size (bp)	Taxonomy	Tag(s)	Date
L_zeae_DSM_20178_GCA_001433745.1.genome	Lactobacillus zeae	99.89 (G)	DEFINITIVE	92/92(1)	3,121,340	Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus		Apr 18, 2019
Lactobacillus zeae CRBIP24.44	Lactobacillus zeae	97.23 (G)	DEFINITIVE	92/92(1)	3,083,267	Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus		Apr 18, 2019
L_casei_LATCC_393_GCA_000829055.1.genome	Lactobacillus casei	99.92 (G)	DEFINITIVE	92/92(1)	2,952,961	Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus		Apr 18, 2019
Lactobacillus zeae CRBIP24.58	Lactobacillus zeae	99.40 (G)	DEFINITIVE	92/92(1)	3,090,860	Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus		Apr 18, 2019
39184.LM1024.1.contig	Bifidobacterium longum subsp. longum	99.95 (G)	DEFINITIVE	92/92(1)	2,324,049	Actinobacteria; Actinobacteria_c; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; Bifidobacterium longum		Apr 15, 2019
Streptomyces_sp._DSM_105299_contigs	Streptomyces bacillaris	95.44 (G)	DEFINITIVE	88/92(2)	7,214,603	Actinobacteria; Actinobacteria_c; Streptomycetales; Streptomycetaceae; Streptomyces		Apr 04, 2019

Supported Data Types

- Currently, TrueBac™ ID's upload center supports two types of data (FASTA and FASTQ) with following extensions.
 - ❖ FASTA: '*.fasta', '*.fna', '*.fna_nt', '*.fas', '*.fsa', '*.fa'
 - ❖ FASTQ: '*.fastq', '*.fq'
- All files can be gzipped (*.gz) or zipped (*.zip) or decompressed.
- For paired-end FASTQ files, please use the following naming convention for auto paring of two fastq files.

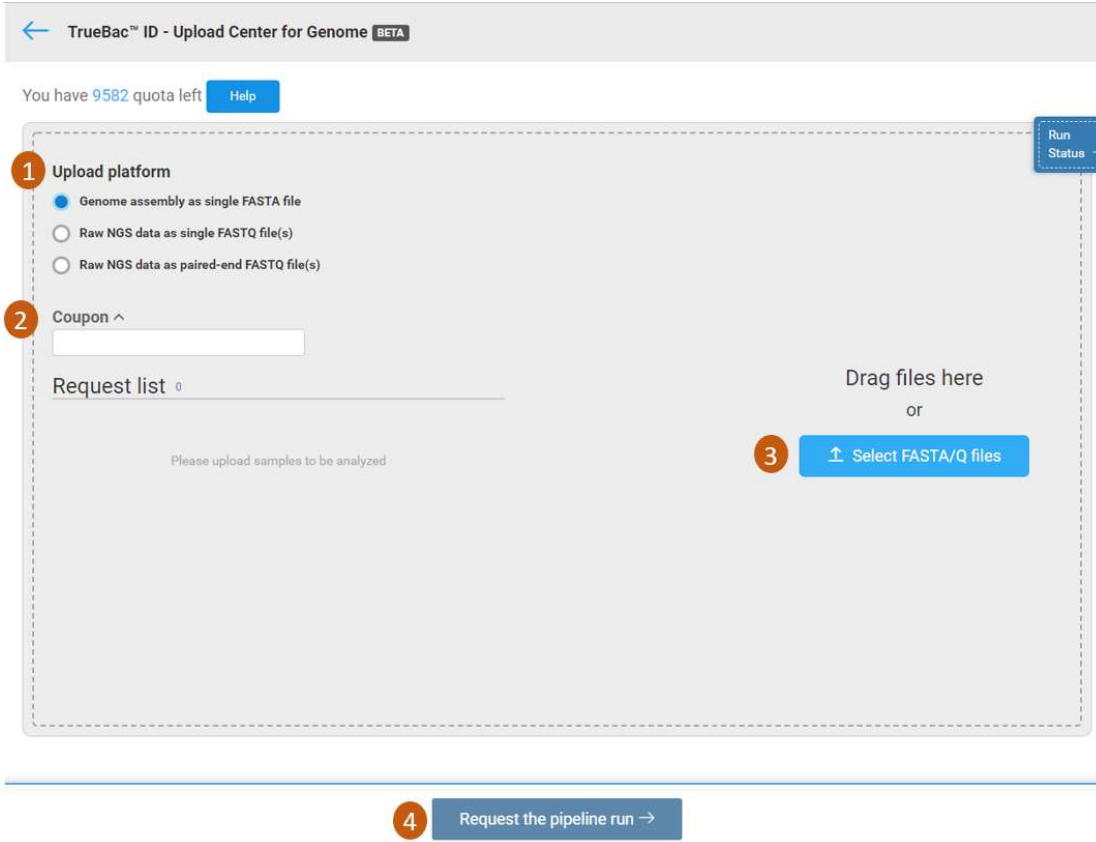
SampleName_SampleNumber_Lane_Read_FlowCellIndex.fastq.gz
(e.g., SampleName_S1_L001_R1_001.fastq.gz / SampleName_S1_L001_R2_001.fastq.gz)

- If auto paring does not work, you may also select the pairs manually to join them to one sample.



[Click on both fastq files to manually join them to a paired-end sample.]

How to upload the data



TrueBac™ ID - Upload Center for Genome **BETA**

You have 9582 quota left [Help](#)

1 Upload platform

- Genome assembly as single FASTA file
- Raw NGS data as single FASTQ file(s)
- Raw NGS data as paired-end FASTQ file(s)

2 Coupon ^

Request list ⁰

Please upload samples to be analyzed

Drag files here
or

3 [Select FASTA/Q files](#)

4 [Request the pipeline run →](#)

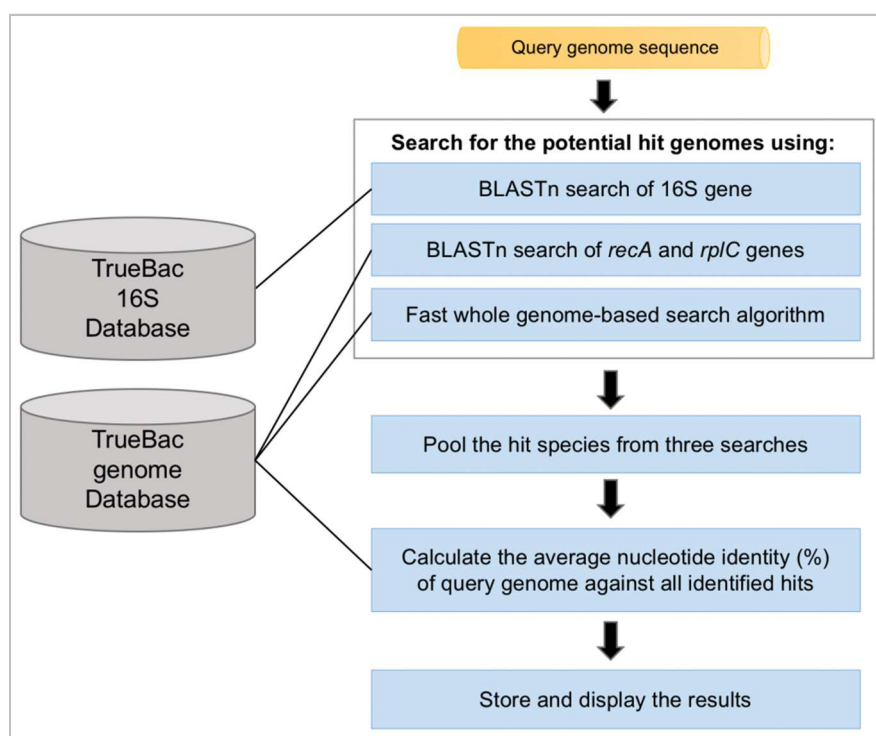
- Click [Genome Upload](#) button on the top to open upload center page for TrueBac™ ID – Genome page.
1. Three options are available under Upload platform menu.
 - **Genome assembly as single FASTA file:** Assembled genome file
 - **Raw NGS data as single FASTQ file(s):** Raw NGS data files sequenced in single read.
 - **Raw NGS data as paired-end FASTQ file(s):** Raw NGS data files sequenced in paired-end read.
 2. Coupon: Insert the coupon code if you have a free coupon.
 3. Select FASTA/Q files: You can upload the genome file(s) by locating your file(s) manually or you can simply drag and drop your file(s).
 4. Request the pipeline run: When all the data is ready, click this button to start ID analysis.

TrueBac™ ID – Genome (Algorithm)

A bacterial isolate can be confidently identified at the species level using genome sequence information ([Richter & Rosselló-Móra, 2009](#); [Chun & Rainey, 2014](#); [Chun et al., 2018](#)).

- **If strain 1 belongs to species A, the necessary conditions are:**
 1. The genome sequence of the type strain of species A must be available.
 2. The average nucleotide sequence identity (ANI) value between the genomes of the strain 1 and type strain of species A should be higher than the proposed cutoff for bacterial species definition, i.e., 95~96%.

Ideally, ANI values are calculated for the genome sequence of the isolate against the genome of type strains of all known species. This would require an enormous computing resource and is not efficient as we are only interested in the species identification. In other words, only closely related species would matter. For this reason, **TrueBac™ ID-Genome** adopted a two-step approach where, first, the potential hit species are identified using four-way searches which are then used to compute ANI values with the query genome. Sometimes, a good-quality sequence cannot be extracted from the final genome assembly. Therefore, in addition to the 16S gene, the *recA* and *rplC* genes are used for searching the potential neighboring species. The RplC is a ribosomal protein whereas RecA is not. They are members of the recently revised bacterial core gene set [[Learn more](#)].



[TrueBac™ ID-Genome algorithm]

¹ Average nucleotide identity (ANI): A widely accepted overall genome related index (OGRI) value that shows high correlation with the conventional DNA-DNA hybridization (DDH) method. A typical species boundary for ANI is $\geq 95\%$.

The TrueBac™ ID-Genome algorithm (figure above) performs the following steps:

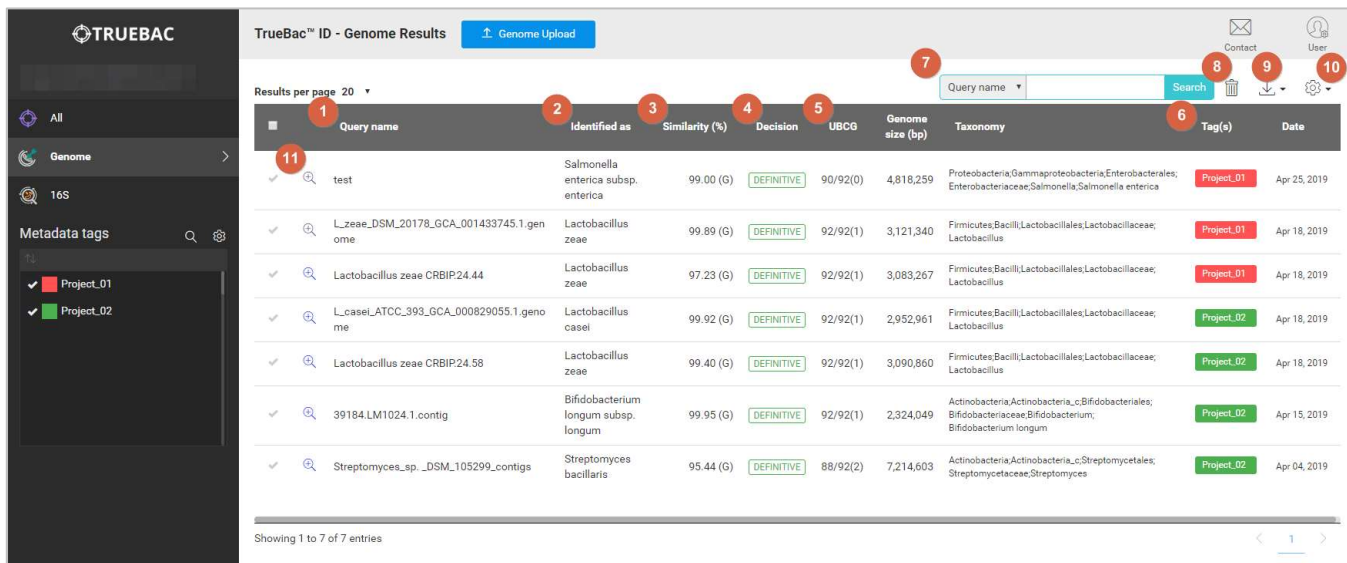
1. The query genome is used to find the potential hit species against the TrueBac™ DB using the following search engines:
 - 1.1. If the query genome contains 16S sequence, it will be used to search against the TrueBac™ DB using BLASTn program.
 - 1.2. If the query genome contains *recA* gene sequence, it will be used to search against the TrueBac™ DB using BLASTn program.
 - 1.3. If the query genome contains *rplC* gene sequence, it will be used to search against the TrueBac™ DB using BLASTn program.
 - 1.4. Additionally, at least one search algorithm that utilizes whole-genome information is used. At present, the MASH tool ([Ondov et al., 2016](#)) is included in our pipeline.
2. The potential hit species identified in the step 1 are pooled and used for computing average nucleotide identity using the MUMmer tool (ANIm; [Richter & Rosselló-Móra, 2009](#))
3. The decision for species-level identification is made by considering ANIm and 16S similarity values.

The TrueBac™ ID-Genome algorithm (figure above) decision making process:

1. If there is a known species (species with valid name or [genomospecies](#)) with ≥ 95 % ANI, the decision is made as “Identified correctly to that species.”
2. If there are no known species with ≥ 95 % ANI, the decision is made as “sp. nov.” (meaning novel species).
3. If there is an ambiguity in making an identification call, the decision is made as “sp.” (e.g., *Bacillus* sp.). It includes the following cases.
 - 3.1. The genome sequence(s) are not available for the type strains of some of the potential hit species. In this case, the result of identification can be updated later when TrueBac™ DB is updated with that missing genomes in future.
 - 3.2. The query genome shows identical or very similar ANI values to two or more species. In most cases, the latter species belong to the same species. In other words, they are synonyms but the necessary taxonomic change (i.e., combining them into a single species) has not proposed yet.

Please note that the TrueBac™ DB contains >2,000 genomospecies [[Learn more](#)]. If your query genome is identified as one of these, it means that you have a novel species. Genomospecies are included in the database and are being expanded constantly to provide users the way of tracking isolates belonging to novel species. For example, [CP015110](#) (a genome sequence deposited in NCBI) represents a novel species in the genus *Acinetobacter* so it was named the genomospecies [CP015110_s](#) in TrueBac DB. If a user isolates multiple strains belonging to this species, they will be identified as “CP015110_s” instead of “*Acinetobacter* sp. nov.” by TrueBac™ ID-Genome service. In this way, most, if not all, isolates can be properly classified and organized easily according to the species.

TrueBac™-Genome ID results [General overview]



1	2	3	4	5	6	7	8	9	10
Query name	Identified as	Similarity (%)	Decision	UBCG	Genome size (bp)	Taxonomy	Tag(s)		
test	Salmonella enterica subsp. enterica	99.00 (G)	DEFINITIVE	90/92(0)	4,818,259	Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica	Project_01		
L_zeae_DSM_20178_GCA_001433745.1.genome	Lactobacillus zeae	99.89 (G)	DEFINITIVE	92/92(1)	3,121,340	Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus	Project_01		
Lactobacillus zeae CRBIP24.44	Lactobacillus zeae	97.23 (G)	DEFINITIVE	92/92(1)	3,083,267	Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus	Project_01		
L_casei_ATCC_393_GCA_000829055.1.genome	Lactobacillus casei	99.92 (G)	DEFINITIVE	92/92(1)	2,952,961	Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus	Project_02		
Lactobacillus zeae CRBIP24.58	Lactobacillus zeae	99.40 (G)	DEFINITIVE	92/92(1)	3,090,860	Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus	Project_02		
39184.LM1024.1.contig	Bifidobacterium longum subsp. longum	99.95 (G)	DEFINITIVE	92/92(1)	2,324,049	Actinobacteria;Actinobacteria_c;Bifidobacteriales;Bifidobacteriaceae;Bifidobacterium;Bifidobacterium longum	Project_02		
Streptomyces_sp._DSM_105299_contigs	Streptomyces bacillaris	95.44 (G)	DEFINITIVE	88/92(2)	7,214,603	Actinobacteria;Actinobacteria_c;Streptomycetales;Streptomycetaceae;Streptomyces	Project_02		

Showing 1 to 7 of 7 entries

- On the TrueBac™ ID – Genome page, the summary of all genomes uploaded will be shown. In here, a user can manage overall ID results from tagging to deletion.
- Query name:** The sample name given by the user.
 - Identified as:** Taxon name identified by TrueBac™ ID system.
 - Similarity (%):** ANI (%) value or 16S rRNA similarity (%).
 - Decision:** four types of decision types are available.
 - DEFINITIVE:**
Identification has been made based on ANI value against the unique reference genome ($\geq 95\%$)
 - DEFINITIVE_16S:**
Identification has been made based on the unique reference 16S rRNA ($\geq 97\%$)
 - AMBIGUOUS:**
Multiple reference genome was found with ANI value larger than or equal to 95% threshold value.
 - UNIDENTIFIED:**
There was neither genome nor 16S rRNA sequence to make the conclusion.
 - UBCG:** Abbreviation for Up-to-date Bacterial Core Gene, which indicates how many core genes that designated genome has. The number in the bracket represent the number of paralogs found. [\[Learn more\]](#)
 - Tag(s):** Attach tag(s) (project number, sample(s), study types, etc.) to manage your genome data.
 - Search:** Search any terms related to Query name, Identified as, Taxonomy and Date
 - Delete results:** A button to delete selected genomes.
 - Export as spreadsheet:** Export selected or all genome(s) in tab-separated-values (tsv) format.
 - Display column:** Used to turn on/off columns.
 - Click of a magnify glass on each sample (🔍) will show more in-depth report of a sample (Please refer to the next page).

TrueBac™-Genome ID results [in-depth data]

A. Summary & Candidate species

TRUEBAC™ ID - Genome Result

test - *Salmonella enterica* subsp. *enterica* - genomic evidence

Candidate species

No.	Hit taxon	ANI (%)	ANI coverage (%)	16S (%)	recA (%)	rplC (%)	Taxonomy
1	<i>Salmonella enterica</i> subsp. <i>enterica</i>	98.55	91.3	98.98	99.34	99.52	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica
2	<i>Salmonella enterica</i> subsp. <i>salamae</i>	96.30	85.8	99.32	97.83	99.05	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica
3	<i>Salmonella enterica</i> subsp. <i>indica</i>	95.82	83.1	98.57	98.02	98.41	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica
4	<i>Salmonella enterica</i> subsp. <i>diarizonae</i>	95.49	79.9	99.39	97.36	99.36	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica
5	<i>Salmonella enterica</i> subsp. <i>houtenae</i>	95.21	81.5	98.98	97.46	98.57	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica
6	<i>Citrobacter koseri</i>	85.33	52.6	98.90	N/A	96.19	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Citrobacter
7	<i>Salmonella bongori</i>	90.21	79.8	98.63	92.84	97.78	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella
8	<i>Salmonella enterica</i> subsp. <i>arizonae</i>	93.56	79.9	98.50	96.52	98.89	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Salmonella;Salmonella enterica
9	<i>Citrobacter sedlakii</i>	84.81	45.1	98.29	N/A	N/A	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Citrobacter
10	PQLZ_s	84.87	42.1	98.29	91.40	N/A	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Citrobacter
11	CP011132_s	84.87	40.9	97.88	N/A	96.19	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Citrobacter
12	FCOP_s	84.49	28.8	97.74	N/A	95.87	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Enterobacter

Showing 1 to 15 of 15 entries

- 1) Sample name, Decision status, General statistics about the genome are shown.
- 2) Download: You can download the results in various formats.
 - The results such as Candidate species [Hits], antibiotic resistance genes [AMR], and virulence factors [VF] can be downloaded either excel (*.xlsx) or JSON (*.json) formats.
 - Sequence file(s): Assembled genome, 16S rRNA, and two core genes (*recA*, *rplC*) can be downloaded in FASTA format (*.fasta).
- 3) Decision statement
- 4) Candidate species: List of species that were considered as a candidate prior to Average Nucleotide Identification (ANI) calculation.

B. Antimicrobial resistance gene(s)

Antimicrobial resistance (AMR) gene(s) are found using the AMRFinderPlus, a tool that identifies AMR genes using either protein annotations or nucleotide sequence. (<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>) via National Center for Biotechnology Information.

In AMRFinderPlus, two main search method is used namely, BLASTP and the search against Hidden Markov Models (HMM).

Antibiotic resistance gene(s)

Contig	Location	Orientation	Gene symbol	Sequence name	Class	% Coverage	% Identity	Alignment length	Closest acc	HMM id
7	676..1476	+	erm(F)	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(F)	MACROLIDE	100.00	98.87	266	WP_002682030.1	NF012223.0
17	52923..53372	+	arr	NAD(+)-rifampin ADP-ribosyltransferase	RIFAMYCIN	NA	NA	NA	NA	NF033144.1
20	53465..55390	+	blaESP	ESP-1 family subclass B3 metallo-beta-lactamase	BETA-LACTAM	NA	NA	NA	NA	NF000455.2
49	6951..7814	-	aadS	aminoglycoside 6-adenyltransferase AadS	AMINOGLYCOSIDE	100.00	99.65	287	WP_003013318.1	NF033387.1
4	12096..12980	+	bla	class A beta-lactamase, subclass A2	BETA-LACTAM	NA	NA	NA	NA	NF012099.1

Showing 1 to 5 of 5 entries

- 1) Contig: Genome's contig number where the AMR gene is located
- 2) Location: Position of the gene. (c stands for complementary)
- 3) Orientation: Orientation of the gene with respect to the reference
- 4) Gene symbol: The name of the gene(s)
- 5) Sequence name: Gene family name of AMRs
- 6) Class: A relevant Drug class for AMR
- 7) % Coverage: Describes how much of the reference gene is covered by query gene *
- 8) % Identity: Identity value of amino acids sequence of query and reference genes *
- 9) Alignment length: Total length of alignment *
- 10) Closest acc: Genbank accession number *
- 11) HMM id: A relevant HMM id in AMRFINDER HMM database

* The values are only obtained when sequence(s) obtained from BLASTP search

C. Virulence factors

The famous Virulence Factors Database (VFDB, <http://www.mgc.ac.cn/VFs/>) is used to find genes encoding toxins.

Virulence factors

Contig	Location	Description	E-value	% Identity	% Coverage
6728	25..315	NP_752613 (entB) isochorismatase [Enterobactin (VF0228)] [Escherichia coli CFT073]	2.700e-140	97.59	100
6665	25..330	NP_752608 (fepD) ferrienterobactin ABC transporter permease [Enterobactin (VF0228)] [Escherichia coli CFT073]	7.660e-135	94.44	100
6230	c(22..351)	NP_757247 (fimG) FimG protein precursor [Type 1 fimbriae (VF0221)] [Escherichia coli CFT073]	7.770e-167	99.09	100
5923	c(5..319)	NP_460110 (csgG) curli production assembly/transport protein CsgG [Agf (VF0103)] [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]	3.370e-95	84.13	100
4726	15..200	NP_752613 (entB) isochorismatase [Enterobactin (VF0228)] [Escherichia coli CFT073]	1.190e-91	99.46	100

Showing 1 to 186 of 186 entries

*Unlike AMRFinder, VFDB does not provide specific cutoff values. A user may have to make an educative guess based on the value provided.

- 12) Contig: Genome's contig number where the gene is located
- 13) Location: Position of the gene. (c stands for complementary)
- 14) Description: A brief description of VF provided by VFDB
- 15) E-value: A parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases.
- 16) % Identity: Identity value of amino acids sequence of query and reference genes.
- 17) % coverage: Describes how much of the query sequence is covered by the target sequence.