MAFGEN Manual
Ver 1.0

Collaborative Genome Database for Clinical Isolates
That Could Not Be Identified
By MALDI-TOF MS ID Systems

TRUEBAC™ ID

Table of contents

About the MAFGEN project

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight mass spectrometry (MALDI-TOF MS) is a technique that has been widely used for diagnosis of infectious diseases. The word MAFGEN is an abbreviation for 'MALDI-TOF MS FAILED GENOME'. It is a collaborative research project that aims to collect reference grade bacterial genome sequence data from clinical isolates which could not be identified by current MALDI-TOF MS based systems.

Participating laboratories can obtain genome-based bacterial species identification with unprecedented accuracy along with antibiotic resistance, pathogenicity and virulence related gene profiles. We believe that this project can benefit diagnostic laboratories at all levels, from local, national, even to global scale.

Requirement to join the network:

You must represent your medical institution where clinical bacterial isolates are routinely identified by a MALDI-TOF MS systems for diagnostic purposes.

Participants will provide:

- The database will only accommodate bacterial isolates that could not be identified by an up-to-date MALDI-TOF ID system. Fungal isolates will not be considered at this stage.
- Metadata from your isolate (source/patient information, geographical location, isolation date, etc.)
- Detailed ID result from your MALDI-TOF MS system
- High-quality genomic DNA from your bacterial isolates (up to 10 strains/year)

Participants will get:

- Fully assembled genome sequence data (contigs) and, if requested, raw FASTQ data. (Data will be stored for 1 year). We will use Illumina MiSeq or iSeq 100 systems for genome sequencing. If you already have FASTQ sequence files, you can submit those instead of sending DNA samples to us.
- Profile of antimicrobial resistance (AMR) gene(s).
- Profile of potential virulence factor(s).
- Genome-based identification result (including novel species).
- Access to all data from the MAFGEN network except for raw (=FASTQ) and assembled sequence data.

Participants will have an access to:

- MAFGEN website: https://www.truebacid.com/genome/mafgen
- Browse shared MAFGEN data: https://www.truebacid.com/genome/browse_mafgen (you need a registered account to browse data; if you want to try before joining the MAFGEN network, please send your request to Dr. Daniel Ha at sungmin.ha@chunlab.com.

Data ownerships & rights

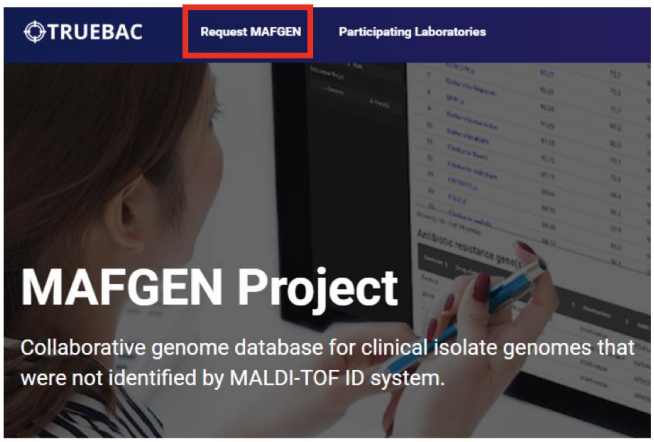
- By submitting the data, you and your organization agree that both ChunLab and your organization share the ownership of the data.
- You will have the right to use and publish your own data for academic purposes.
- ChunLab will have the right to use the data for commercial purposes.
- The submission quota for each organization is ten strains/year.
- For more information please contact our project manager, Dr. Daniel Ha at sungmin.ha@chunlab.com.

How to become a participant

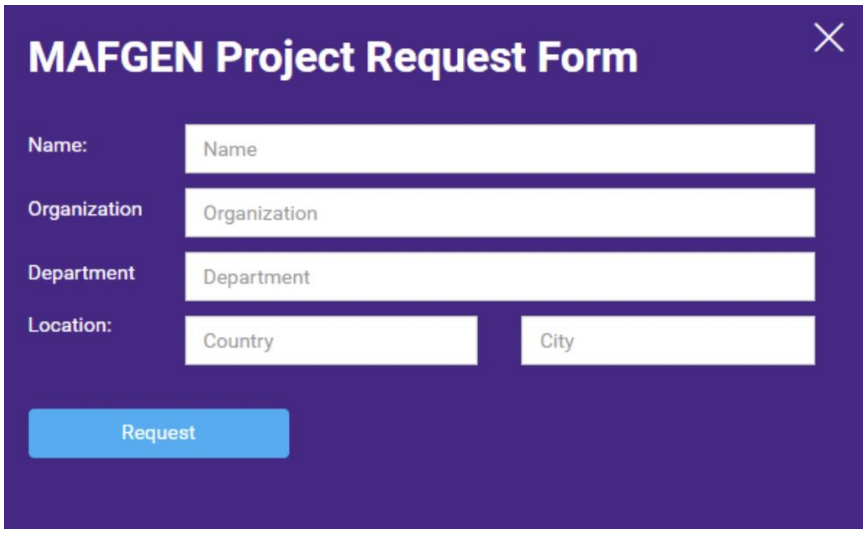
To become a participant,

1. Go to https://www.truebacid.com/genome/mafgen

2. Click "Request MAFGEN" on the top (If you haven't log-in yet the system will redirect you to log-in)



3. Fill in the below form and click the "Request" button.



4. You will get a notification email from us within a day or two.

How to extract genomic DNA

DNA QC Guideline

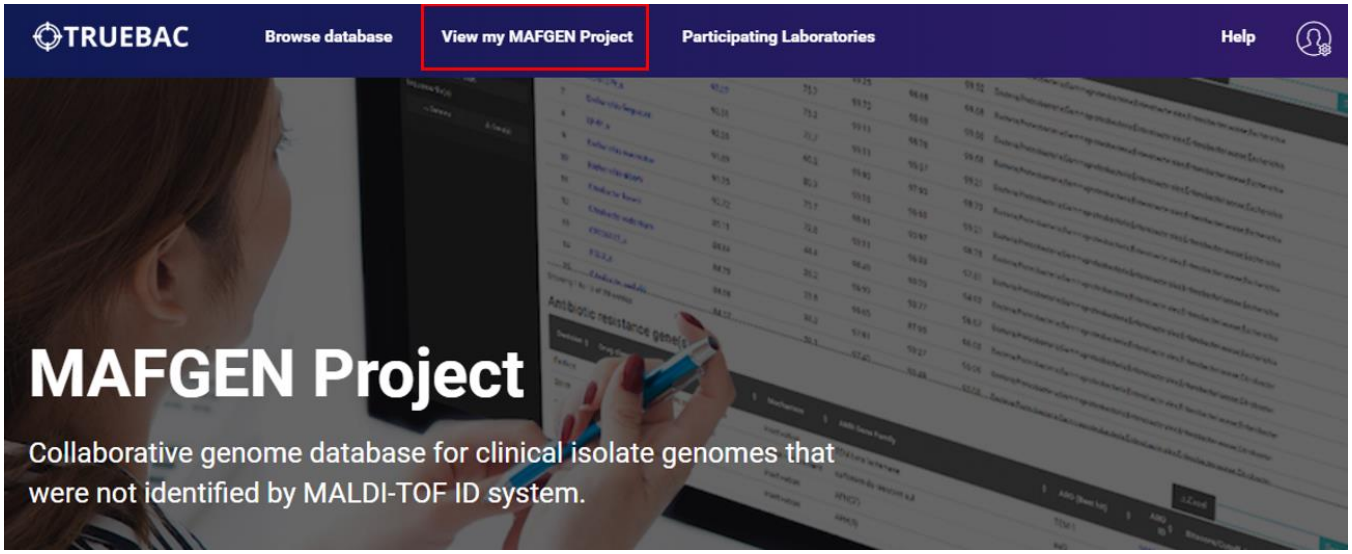
Concentration (ng/ul)	Volume (ul)	A_{260}/A_{280} Ratio
≥ 20 ng/ul	≥ 20	≥ 1.8

You can use any bacterial genomic DNA extraction kit, if it can meet the above QC guideline.

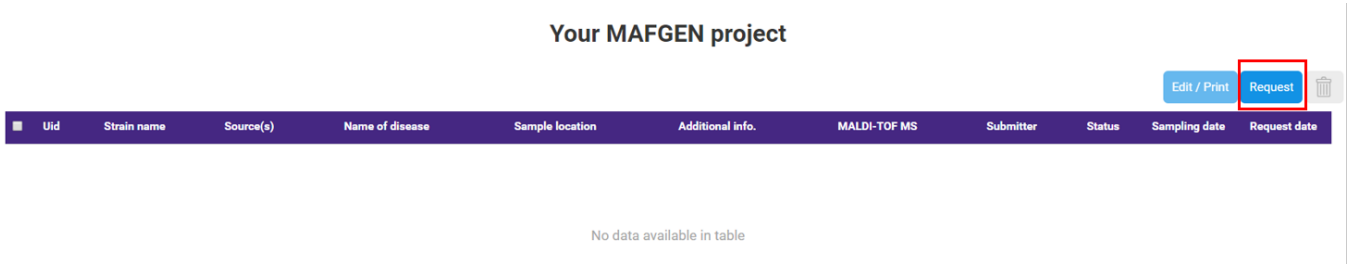
How to write a request from

Once you have become a participant and the genomic DNA is ready for shipping, you need to fill in the request form.

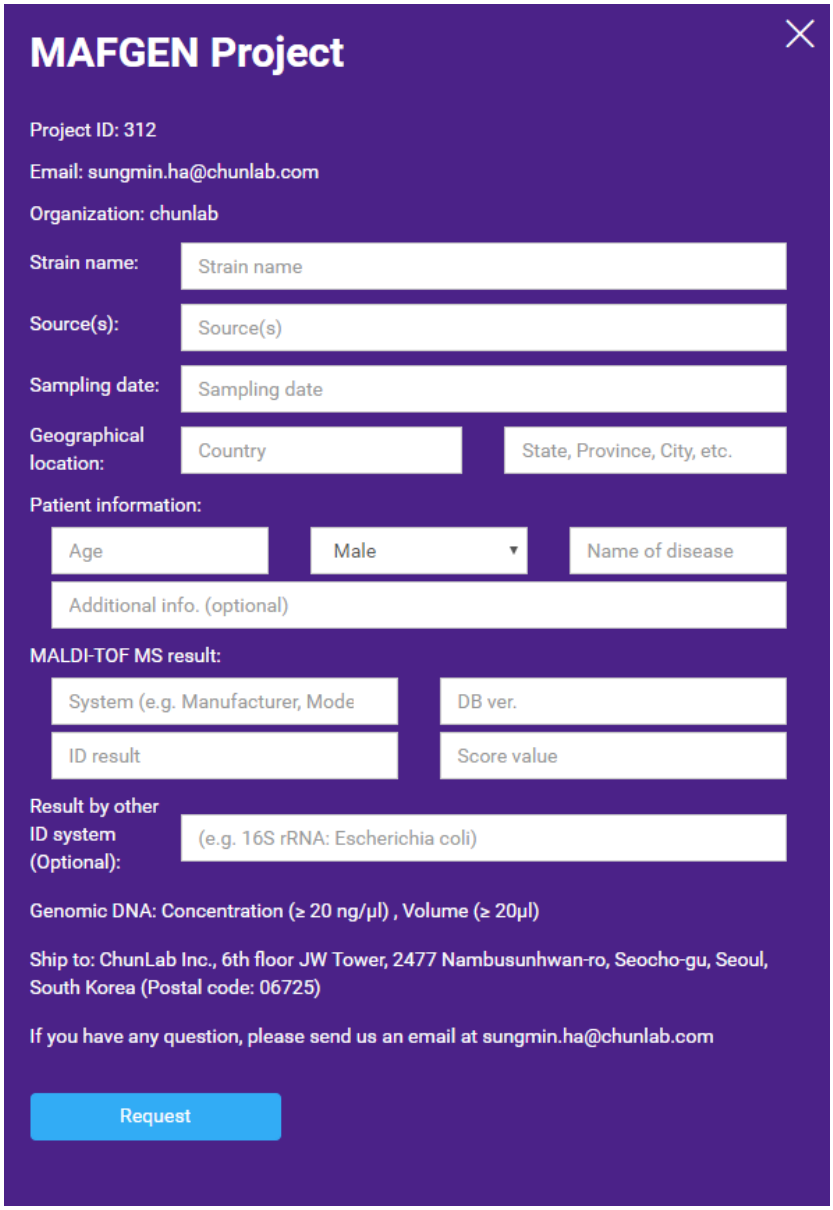
1. On https://www.truebacid.com/genome/mafgen click on “View my MAFGEN Project”



2. Your project manager table will appear. Click on the “Request” button at the top right corner of the table.



3. Complete the following form and click the “Request” button.



Strain name: Name of isolate strain

Source(s): Source of your isolate (e.g., Blood, stool, rectal swap, etc.)

Sampling date: Date when sampling was done

Geographical location: Country and state/province

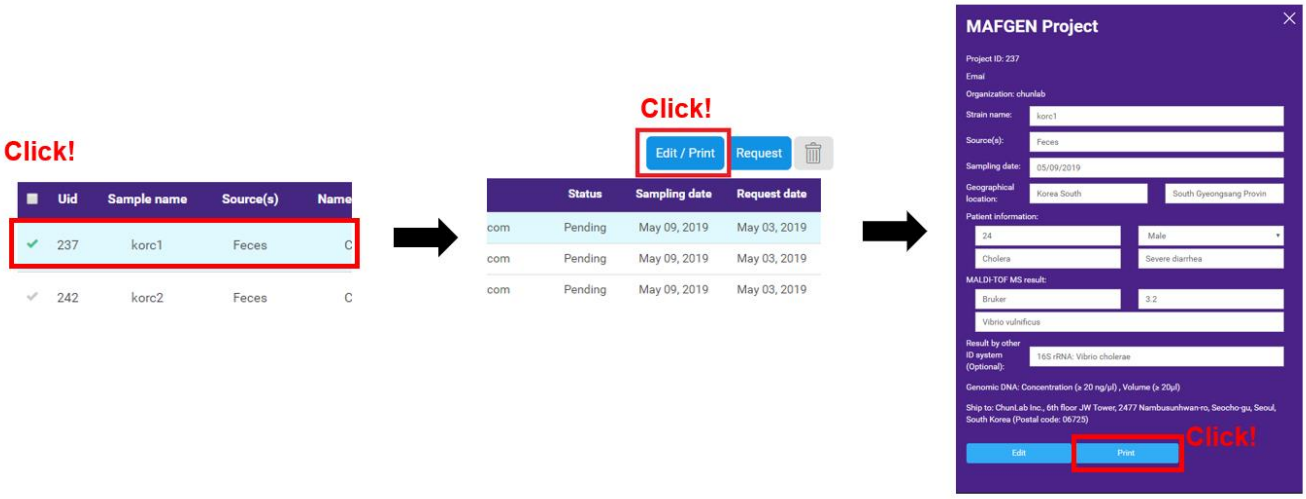
Patient information: Age, gender, and the disease name that the patient has been diagnosed with.

MALDI-TOF MS result: Manufacturer of the system (e.g., BRUKER, VITEK MS, ASTA), Database version, ID result (e.g., *Vibrio cholerae*), and the score value given by the MALDI-TOF system.

Result by other ID system: If you have used another ID method (e.g., 16S rRNA) please provide it here.

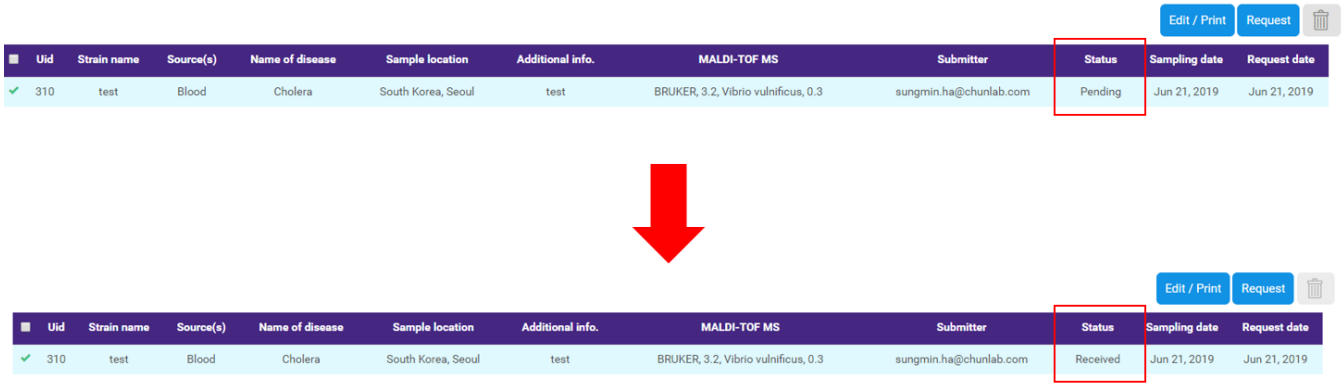
4. Print the request form and send it to us along with the extracted genomic DNA.

***Shipping address**: MAFGEN project, ChunLab Inc., 6th floor JW Tower, 2477 Nambusunhwan-ro, Seocho-gu, Seoul, South Korea (Postal code: 06725)



Monitoring the sequencing process

Once your sample arrives at ChunLab, your project status will be changed from "Pending" to "Received" and the following procedures are performed for sequencing.



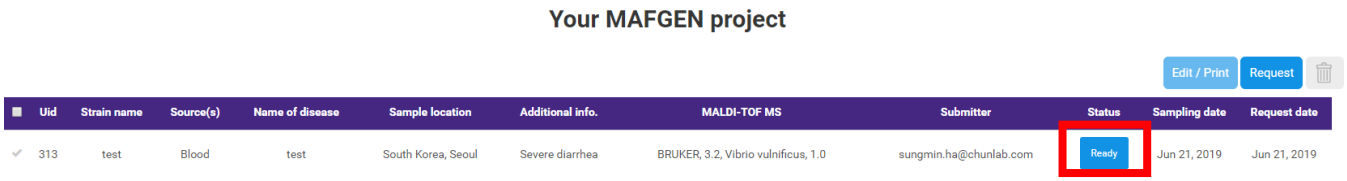
You may use the method stated below for your own research paper:

> The library is constructed using the TruSeq Nano DNA LT Library Prep kit (Illumina) according to the manufacturer's protocols. The library is quantified using the Bioanalyzer 2100 (Agilent) with the DNA 7500 kit. Whole-genome sequencing is performed on the Illumina platform.

Getting data and results

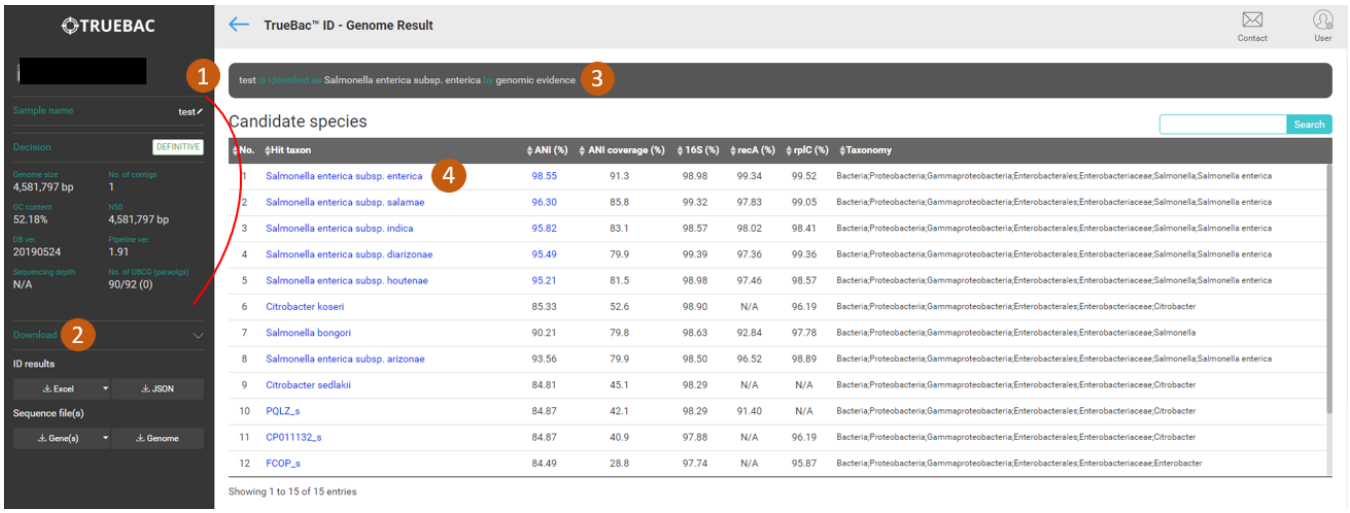
After the sequencing is done, your genome data will be assembled and identified using the TrueBac™ ID cloud platform. Once the identification is ready, you will get a notification email that your project is ready for your review.

1. Double click "Ready" button in the status column in your MAFGEN project table. This button will redirect you to the ID result page.



2. You will be redirected to Genome ID result page.

A. Summary & Candidate species



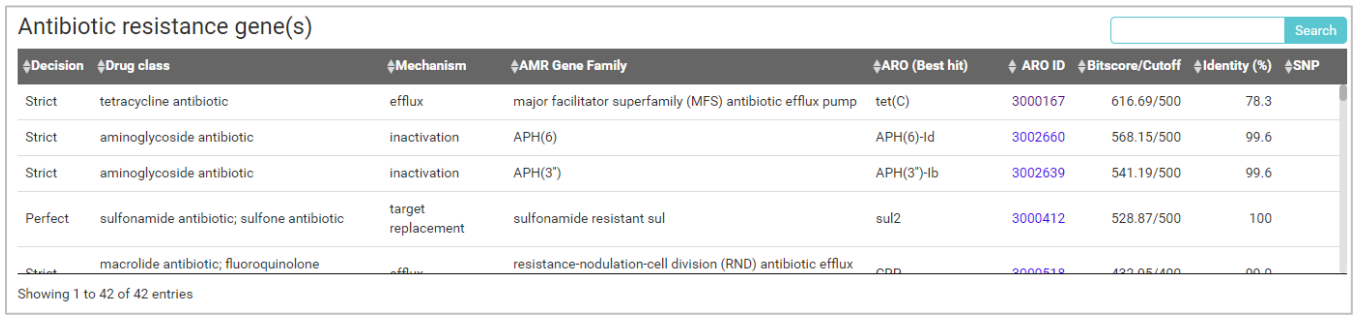
1) Sample name, Decision status, General statistics about the genome are shown.
2) Download: You can download the results in various formats.

- The results such as Candidate species [Hits], antibiotic resistance genes [AMR], and virulence factors [VF] can be downloaded in 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 identified using the Comprehensive Antibiotic Resistance Database (CARD, https://card.mcmaster.ca/) developed by McMaster University, Canada. The database only contains AMRs that have undergone rigorous experiments for their authenticity. AMR search is made with the Resistance Gene Identifier (RGI), which applies different cutoff value for each AMRs, provided by CARD.


Antibiotic resistance gene(s)

Search

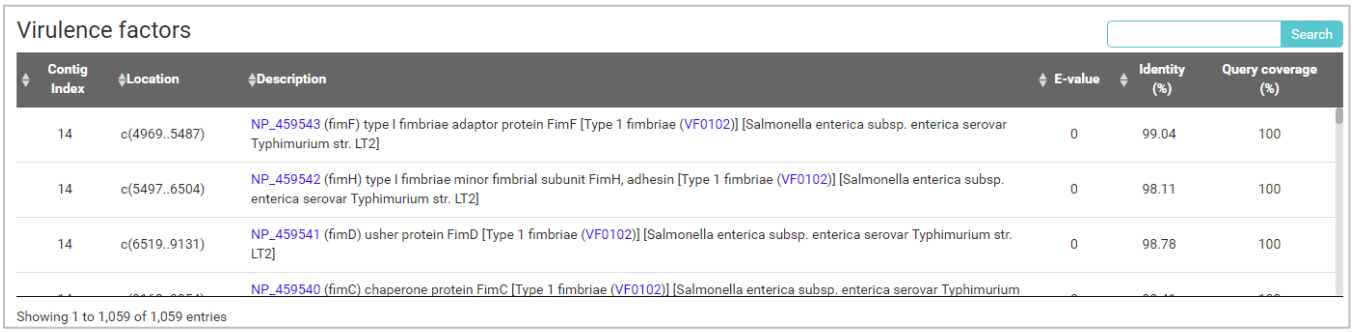
Decision	Drug class	Mechanism	AMR Gene Family	ARO (Best hit)	ARO ID	Bitscore/Cutoff	Identity (%)	SNP
Strict	tetracycline antibiotic	efflux	major facilitator superfamily (MFS) antibiotic efflux pump	tet(C)	3000167	616.69/500	78.3	
Strict	aminoglycoside antibiotic	inactivation	APH(6)	APH(6)-Id	3002660	568.15/500	99.6	
Strict	aminoglycoside antibiotic	inactivation	APH(3")	APH(3")-Ib	3002639	541.19/500	99.6	
Perfect	sulfonamide antibiotic; sulfone antibiotic	target replacement	sulfonamide resistant sul	sul2	3000412	528.87/500	100	
[illegible]	macrolide antibiotic; fluoroquinolone	[illegible]	resistance-nodulation-cell division (RND) antibiotic efflux	[illegible]	[illegible]	[illegible]	[illegible]	

Showing 1 to 42 of 42 entries

- Decision: If the bit-score and identity are above the threshold, the decision will be stated as Perfect. If only one of the values satisfies the threshold, it will be stated as Strict. If none of the values are significant, it is stated as Loose
- Drug class: Drug class that the AMR is coding for.
- Mechanism: A brief description of how AMR blocks antibiotics
- AMR gene Family: Gene family name of AMRs
- ARO (Best hit): Antibiotic Resistance Ontology provided by CARD
- ARO ID: Each ARO's Unique ID (You will be redirected to CARD if clicked)
- Bit-score/Cutoff: Bit-score value from the search result and the cutoff value provided by CARD
- Identify (%): Identity of amino acids sequence of query and reference genes
- SNP: Known variants in CARD

C. Virulence factors

The Virulence Factors Database (VFDB, http://www.mgc.ac.cn/VFs/) is used to find genes encoding toxins and other virulence factors.


Virulence factors

Search

Contig Index	Location	Description	E-value	Identity (%)	Query coverage (%)
14	c(4969..5487)	NP_459543 (fimF) type I fimbriae adaptor protein FimF [Type 1 fimbriae (VF0102)] [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]	0	99.04	100
14	c(5497..6504)	NP_459542 (fimH) type I fimbriae minor fimbrial subunit FimH, adhesin [Type 1 fimbriae (VF0102)] [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]	0	98.11	100
14	c(6519..9131)	NP_459541 (fimD) usher protein FimD [Type 1 fimbriae (VF0102)] [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]	0	98.78	100
		NP_459540 (fimC) chaperone protein FimC [Type 1 fimbriae (VF0102)] [Salmonella enterica subsp. enterica serovar Typhimurium			

Showing 1 to 1,059 of 1,059 entries

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

- Contig Index: Genome's contig number where the gene is located
- Location: Position of the gene. (c stands for complementary)
- Description: A brief description of the VF provided by VFDB
- 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.
- Identity: Identity value of amino acids sequence of query and reference genes.
- Query coverage: Describes how much of the query sequence is covered by the target sequence.