# Introduction to the Cloud Platform for Pathogen In-depth Analysis

Xianhui Peng,Ph.D.

National Institute for Communicable Disease Control and Prevention,
Chinese Center for Disease Control and Prevention.

October 20,2020

# Contents

# 1

# Overview

# Introduction

1 Platform Support:
1.Microbial Data Analysis Cloud Platform
2. Microbial Genome Database

2 Analysis Workflow:
1.based on bacterial genomes
2.based on microbiota

3 Example:
1.Brucella
2.M. tuberculosis
3.Vibrio cholerae
4.Clinical sample

4 Analysis Tools:
1. Statistics and Preprocessing of sequencing data
2. In-Depth Analysis of Pathogen
3. Pathogen screening based on 16s amplicon sequencing

Target:

✓ Efficient

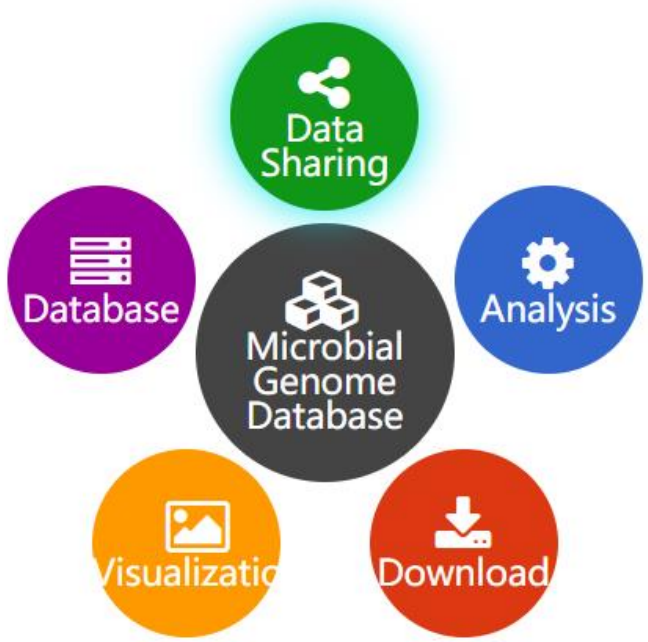
✓ Simple

# 2

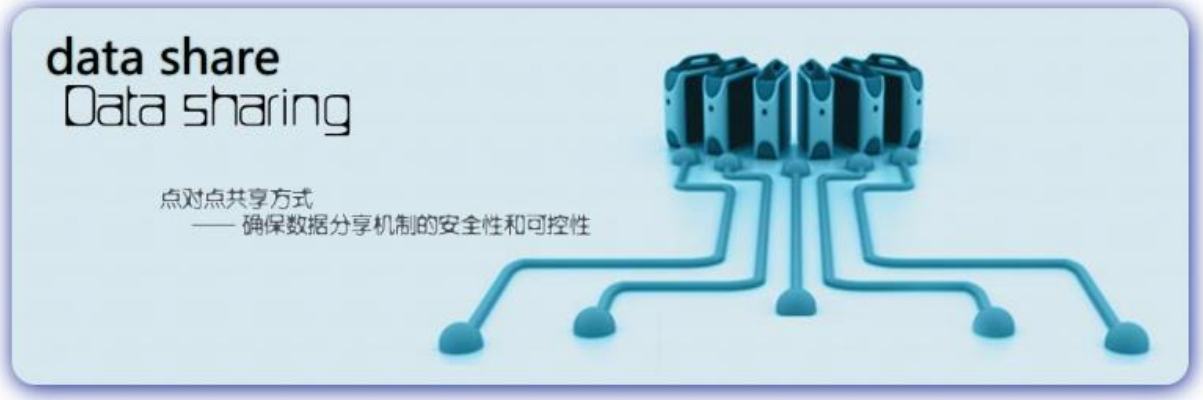
# Platform Support

# 1. Microbial Genome Database

http://data.mypathogen.org/index

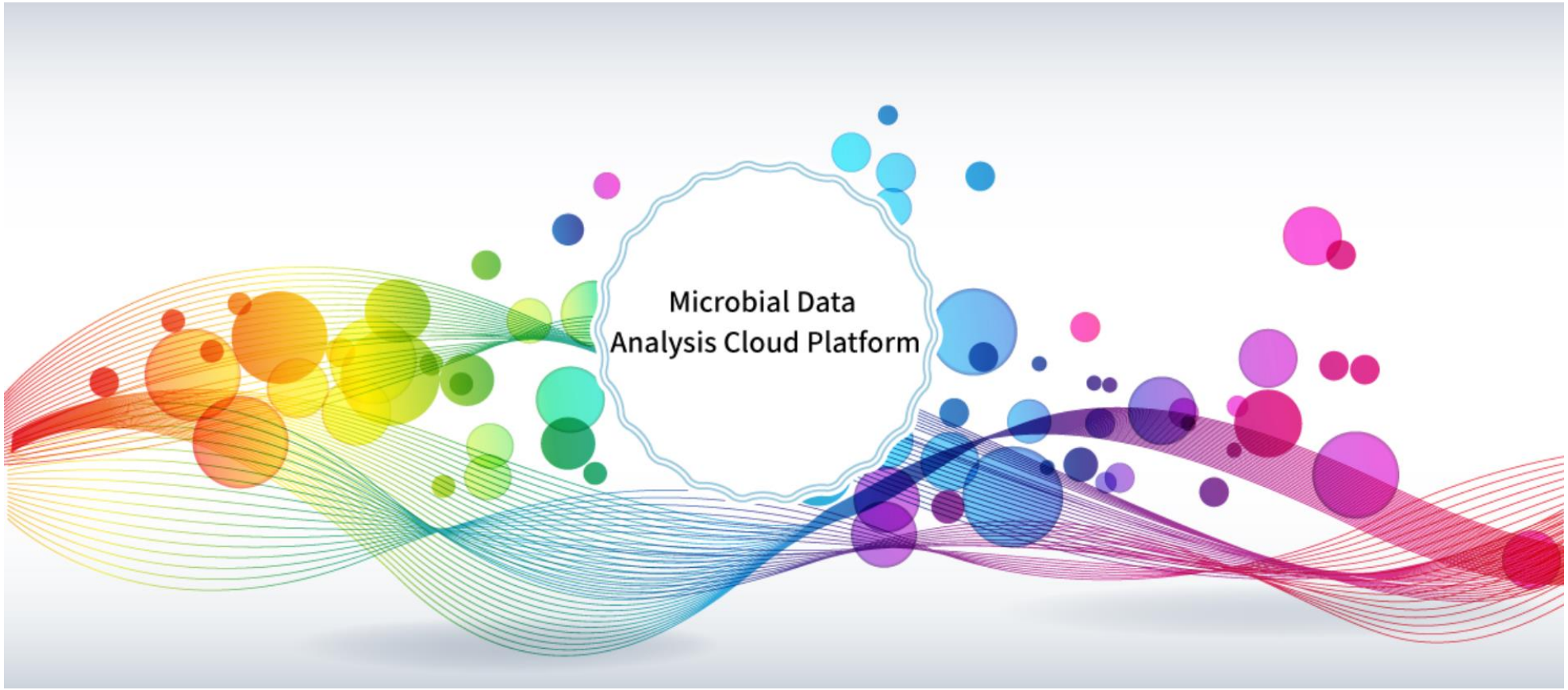



The microbial genome database is a professional database system designed to host a range of pathogenic microbial genomes and to provide users access to searching, downloading and sharing genomics data. It has included comprehensive publicly available bacterial genomic and metagenomic data.



# 2. Microbial Data Analysis Cloud Platform

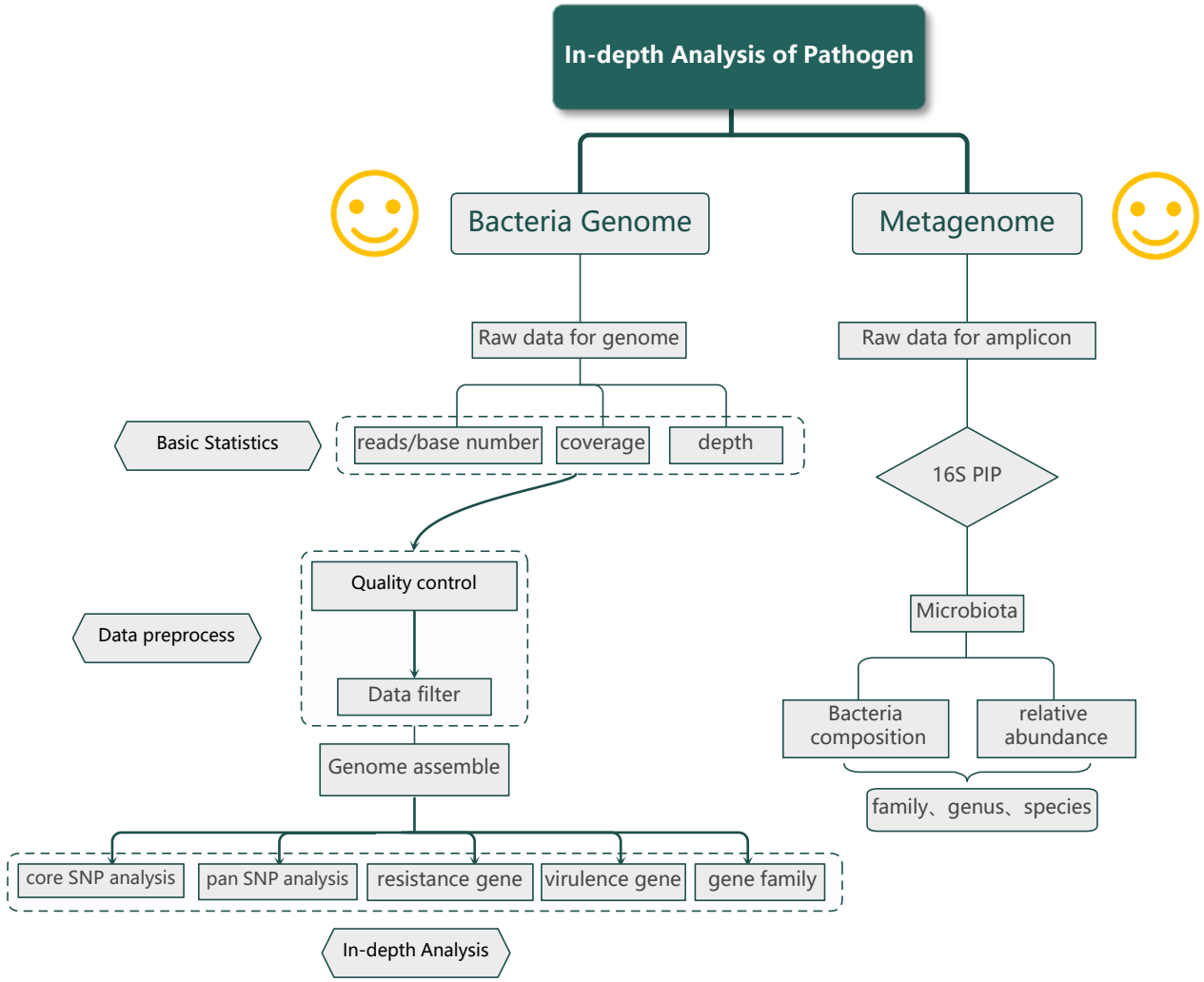
https://analysis.mypathogen.org/



# 3

# Analysis Workflow

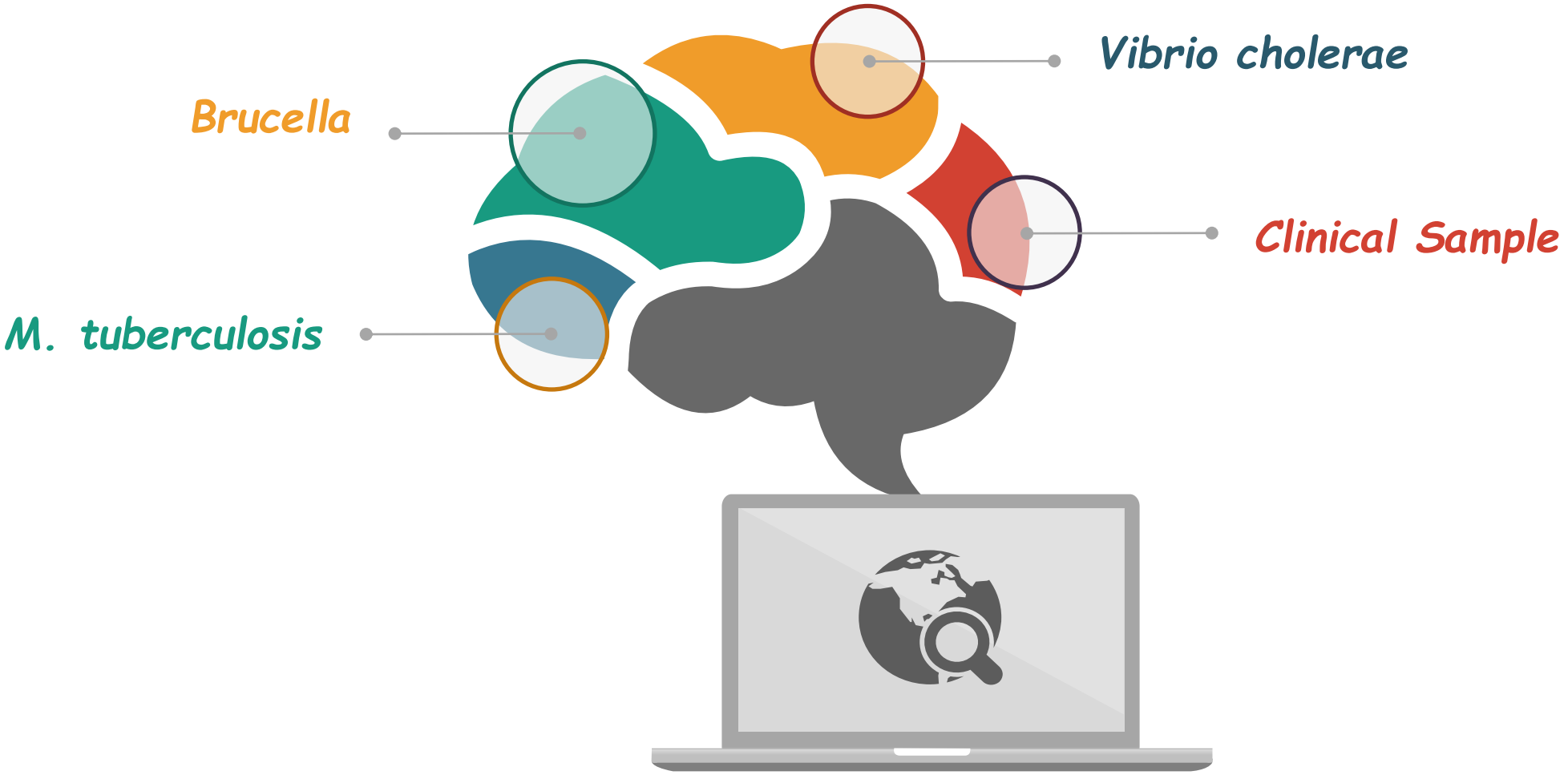
# Analysis Workflow



# 4

# Example

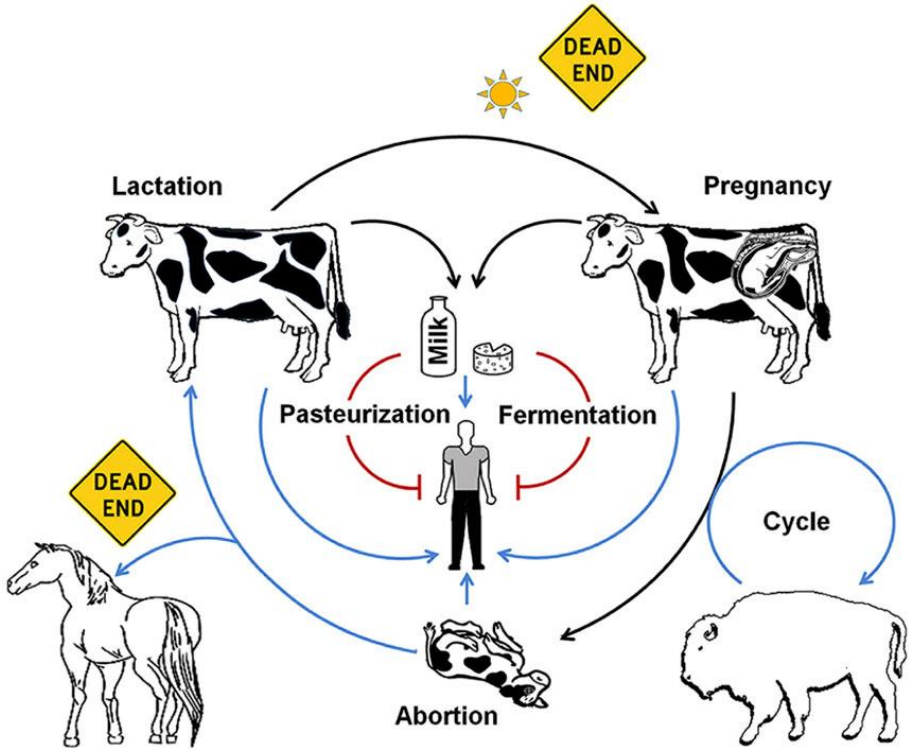
# Example



# 1. *Brucella*

## Key facts

- Brucellosis is found globally and is a reportable disease in most countries
- The disease causes flu-like symptoms, including fever, weakness, malaise and weight loss




- Person-to-person transmission is rare.
- Brucellosis is a bacterial disease caused by various Brucella species, which mainly infect cattle, swine, goats, sheep and dogs.

**Challenge:**

**Identification and molecular typing**

- 10 species and 19+ biovar
- DNA homology of ~95%
- Method: PCR\PFGE\MLST

**Solution:**

**Whole genome analysis**

- Core SNP
- Pan SNP
- Gene family analysis

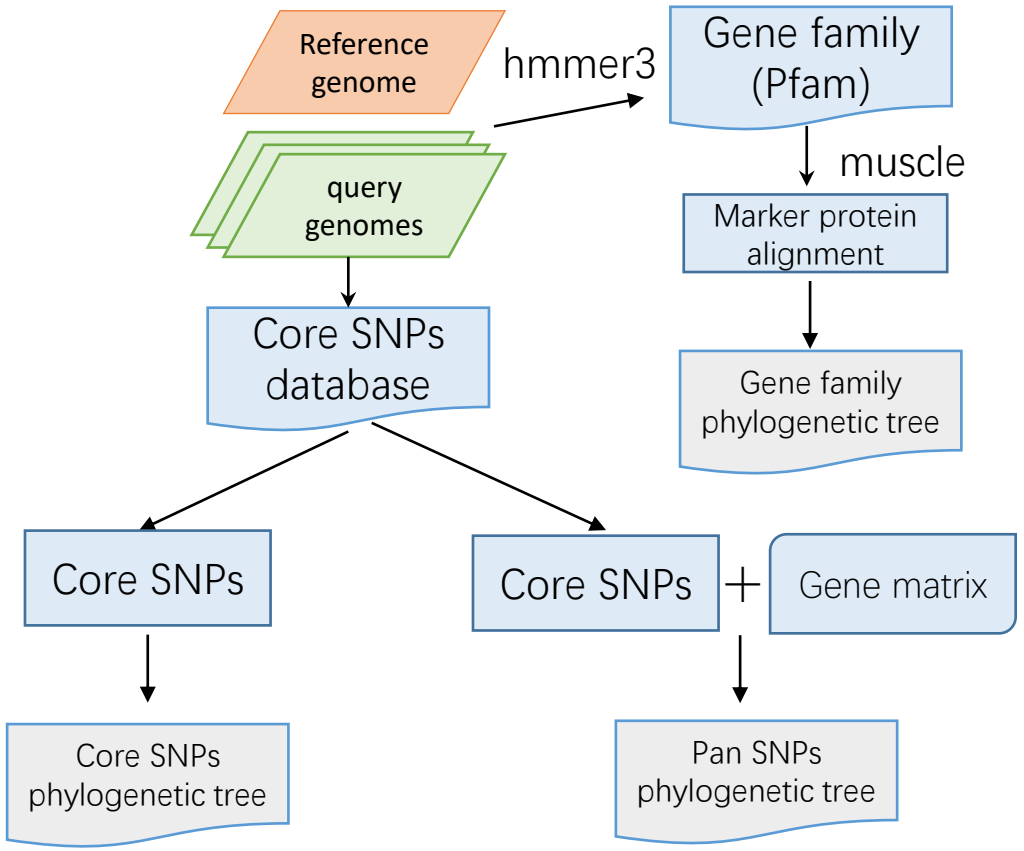



Fig 1. Three methods based on whole genome analysis

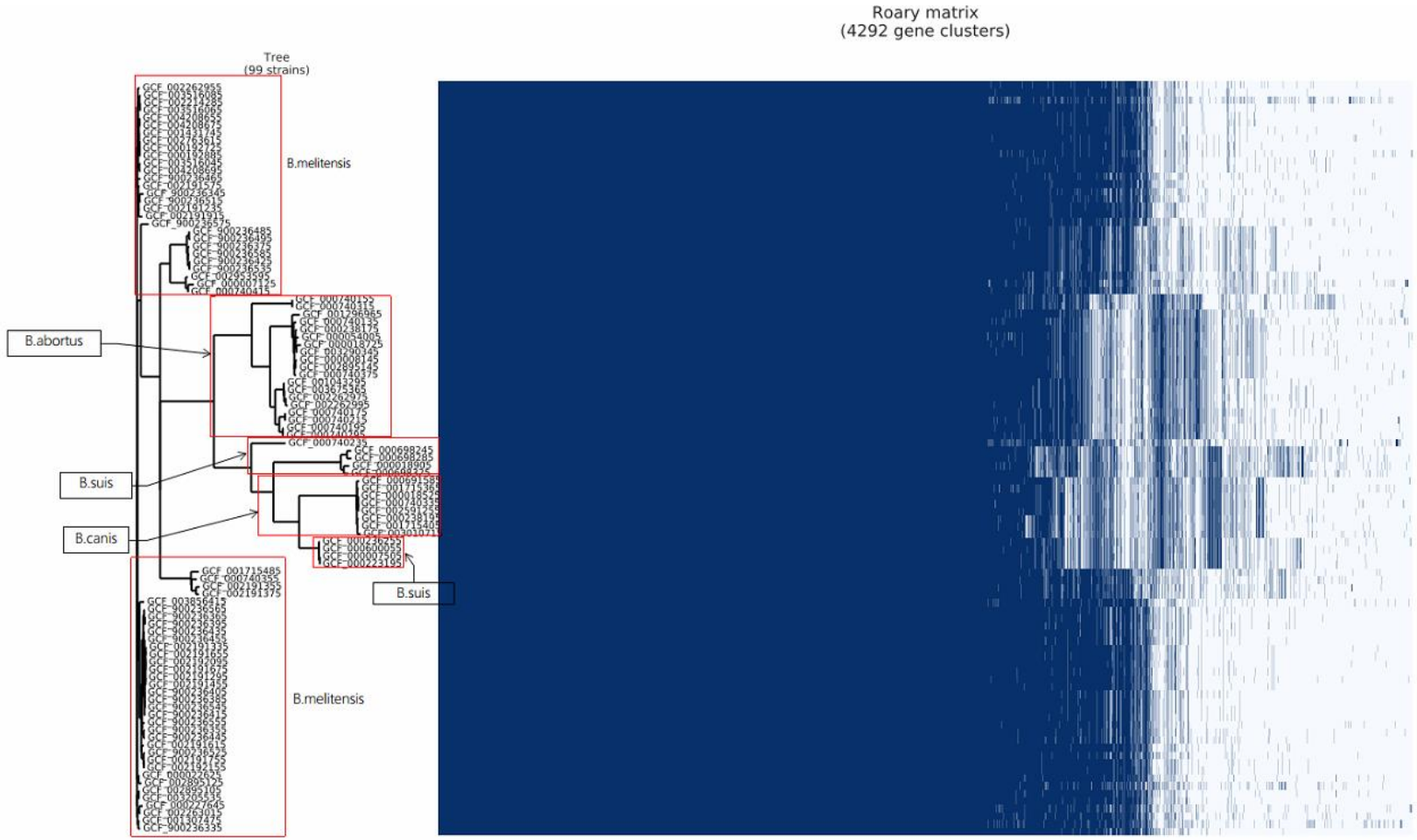



Figure 1. Phylogenetic tree of pan-SNP based on Brucella genus

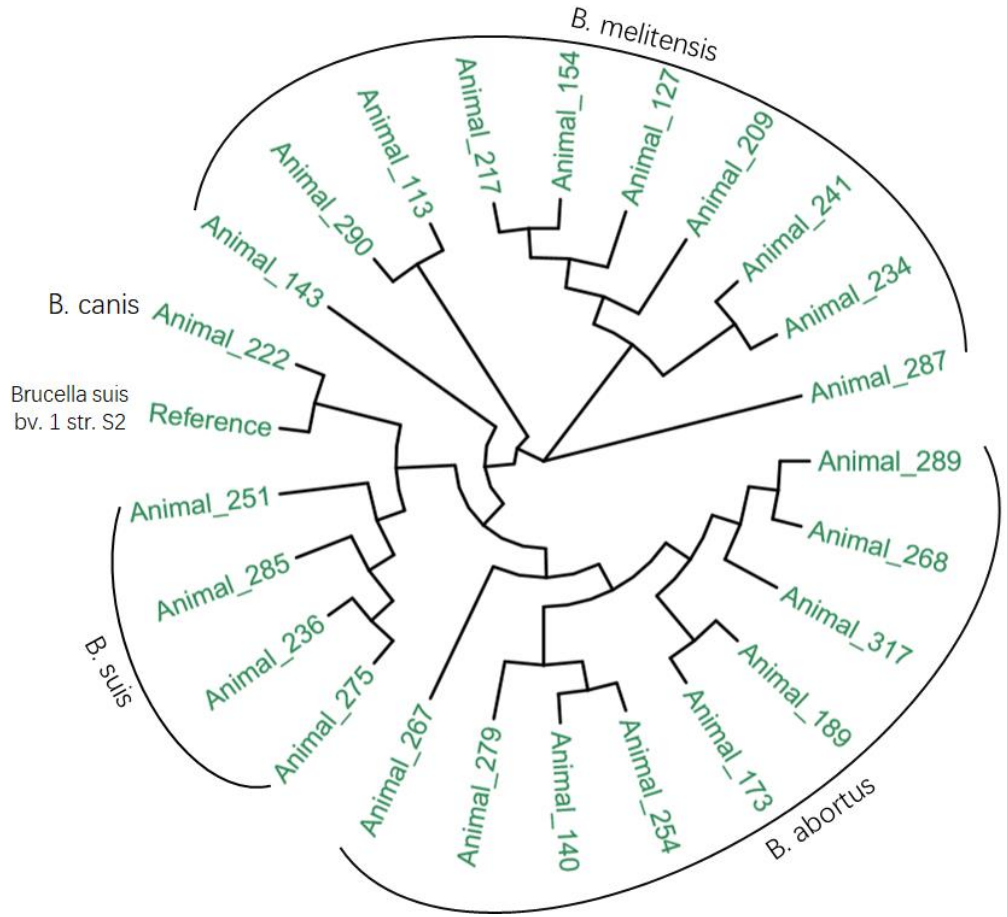



Figure 2. Phylogenetic tree of core SNP based on Brucella genus

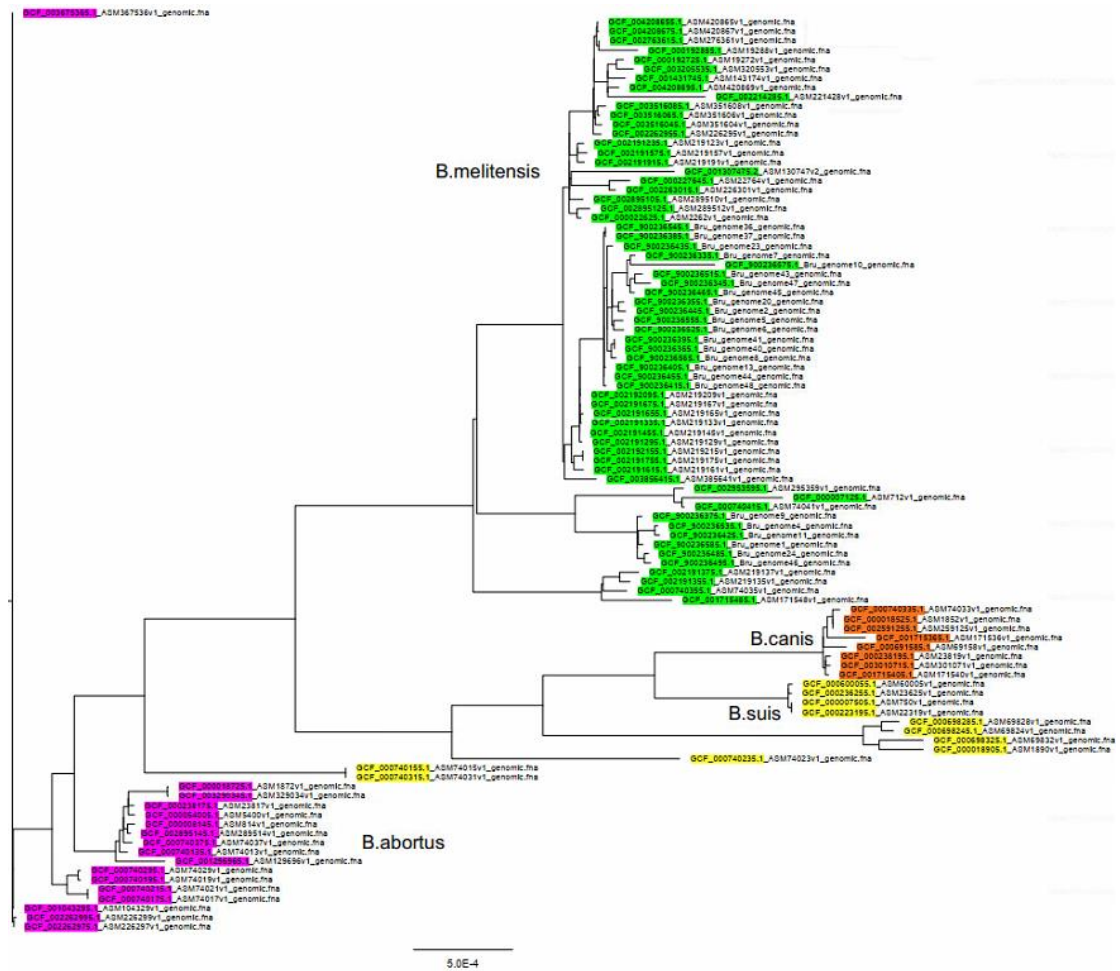
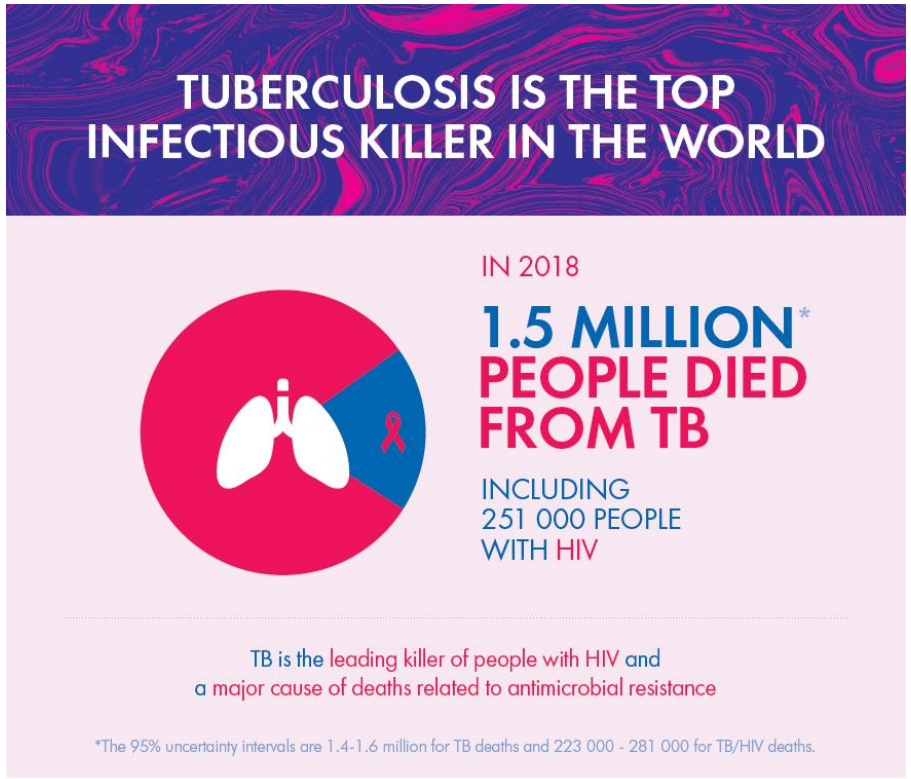



Figure3. Phylogenetic tree of gene family based on Brucella genus

# 2. *M. tuberculosis*

## Key facts

- The top infectious killer in the world.
- In 2018, 10 million, 1.5 million.
- TB is the leading killer of people with HIV.
- Multidrug-resistant TB (MDR-TB) remains a public health crisis and a health security threat.




- Globally, TB incidence is falling at about 2% per year.
- TB treatment saved about 58 million lives globally between 2000 and 2018.
- Ending the TB epidemic by 2030 is among the health targets of the Sustainable Development Goals.

**Challenge:**

- *M. tuberculosis* drug susceptibility testing (DST) is time-consuming and laborious.
- The spread of multidrug-resistant strains for first-line tuberculosis (TB) treatment.
- A growing need for data on second-line drugs, including the fluoroquinolones and aminoglycosides.

**Solution:**

- Whole genome
- resistance gene(not point mutations )★
- related SNPs (next goal)

J Clin Microbiol(2017) doi: 10.1128/JCM.02483-16.

Table 1.1 Screening results of genome-wide Mycobacterium tuberculosis resistance genes

| #FILE | NUM_FOUND | AAC(2')-Ic | Erm(37) | Erm(38) | rpoB2 | RbpA | efpA | mfpA | mtrA |
|---|---|---|---|---|---|---|---|---|---|
| GCF_000008585.1_ASM858v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000009445.1_ASM944v1_genomic.fna | 8 | 100 | 69.26 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000010685.1_ASM1068v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000016145.1_ASM1614v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000016925.1_ASM1692v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000023625.1_ASM2362v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000153685.2_ASM15368v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000154585.2_ASM15458v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000154605.2_ASM15460v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000193185.2_ASM19318v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |

Note: NUM_FOUND: the number of drug-resistant genes screened from the genome; row name: drug-resistant gene name; column name: strain name; table value:% COVERAGE

Table 1.2 Mycobacterium tuberculosis (GCF_000008585.1_ ASM858v1_genomic.fna) resistance gene screening results

| #FILE | SEQUENCE | START | END | STRAND | GENE | COVERAGE | COVERAGE_MAP | GAPS | %COVERAGE | %IDENTITY | DATABASE | ACCESSION | PRODUCT | RESISTANCE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 314424 | 314969 | - | AAC(2')-Ic | 1-546/546 | ========== | 0/0 | 100 | 100 | card | AL123456.3:314309-314855 | AAC(2')-Ic | aminoglycoside |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 761855 | 765274 | + | rpoB2 | 64-3462/3489 | ======/==== | 19/41 | 97.13 | 79.56 | card | AP006618.1:4835200-4838689 | rpoB2 | rifamycin |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 2229013 | 2229552 | + | Erm(37) | 1-540/540 | ========== | 0/0 | 100 | 100 | card | AL123456:2231680-2232220 | Erm(37) | streptogramin/macrolide/lincosamide |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 2310154 | 2310467 | + | RbpA | 1-314/345 | ========== | 0/0 | 91.01 | 87.58 | card | HQ203032:1-346 | RbpA | rifamycin |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3147251 | 3148843 | - | efpA | 1-1593/1593 | ========== | 0/0 | 100 | 100 | card | AL123456.3:3153039-3154632 | efpA | phenicol |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3621563 | 3622249 | - | mtrA | 1-687/687 | ========== | 0/0 | 100 | 100 | card | AL123456.3:3626663-3627350 | mtrA | antibacterial_free_fatty_acids |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3735428 | 3735618 | + | Erm(38) | 777-967/1161 | .....===.. | 0/0 | 16.45 | 76.44 | card | AY154657.2:63-1224 | Erm(38) | streptogramin/macrolide/lincosamide |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3765240 | 3765791 | - | mfpA | 1-552/552 | ========== | 0/0 | 100 | 100 | card | AL123456:3773016-3773568 | mfpA | fluoroquinolone |

# 3. *Vibrio cholerae*

## Key facts

- Cholera is an acute diarrhoeal disease that can kill within hours if left untreated.
- 1.3 million to 4.0 million cases of cholera, and 21 000 to 143 000 deaths.
- Up to 80% of cases treated.
- Severe cases will need rapid treatment.


- Provision of safe water and sanitation is critical to control the transmission of cholera and other waterborne diseases.
- A global strategy on cholera control with a target to reduce cholera deaths by 90% was launched in 2017.

**Challenge:**

- The study found that non-toxic strains of the O1 and O139 groups, as well as other serogroups of Vibrio cholerae, caused occasional sporadic cases or led to very small cholera outbreaks.
- *ctxAB gene* is not the only virulence gene for Vibrio cholerae.
- Vibrio cholerae may have pathogenic mechanisms other than cholera toxin(CT).

**Solution:**

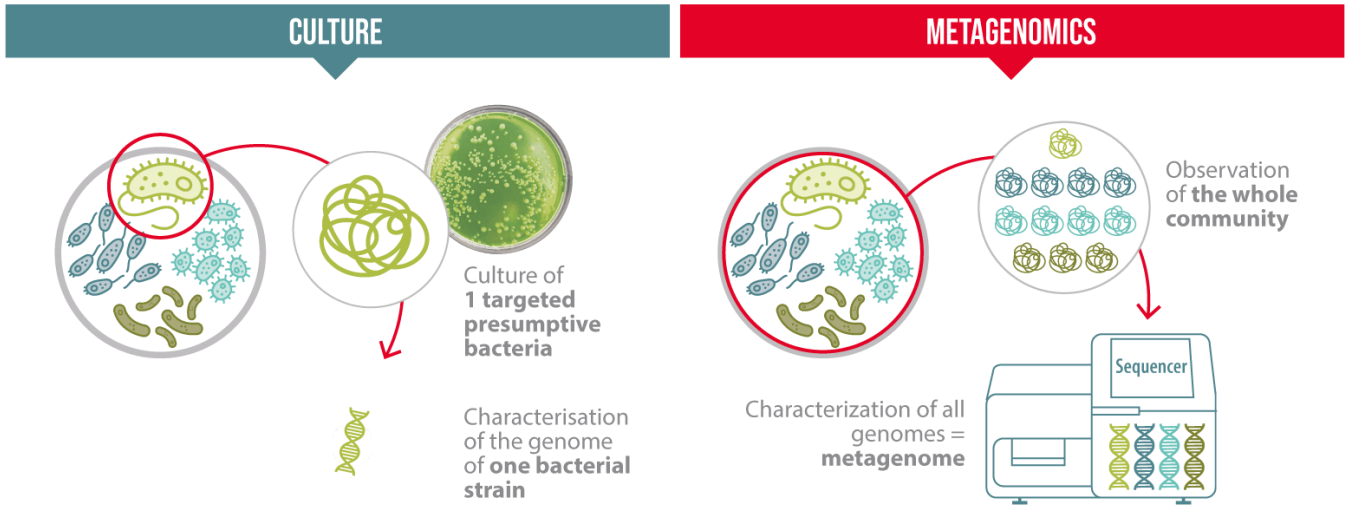
- Whole genome
- Virulence gene

Table 2. Genome-based screening results of Vibrio cholerae virulence genes

| #FILE | GCA_000006745.1_ ASM674v1_genomic.fna | GCA_000016245.1_ ASM1624v1_genomic.fna | GCA_000021605.1_ ASM2160v1_genomic.fna | GCA_00002c1625.1_ ASM2162v1_genomic.fna |
|---|---|---|---|---|
| **NUM_ FOUND** | 56 | 55 | 52 | 55 |
| **IlpA** | 100 | 100 | 100 | 100 |
| **VCA0109** | 100 | 97.26 | 100 | 97.26 |
| **VCA0122** | 100 | 100 | 100 | 100 |
| **ace** | 100 | 100.00;100.00 | . | 100.00;100.00 |
| **acfA** | 100 | 100 | 100 | 100 |
| **acfB** | 100.00;7.18 | 100.00;7.18 | 100.00;7.18 | 100.00;7.18 |
| **acfC** | 100 | 100 | 100 | 100 |
| **acfD** | 100 | 100 | 100 | 100 |
| **cheD** | 6.03;9.14;11.17 | 11.17;9.14;6.03 | 6.03;9.14;8.72;11.17 | 6.03;9.14;11.17 |
| **clpB/ vasG** | 100 | 100 | 100 | 100 |
| **cqsA** | 100 | 100 | 100 | 100 |
| **ctxA** | 100 | 100.00;100.00 | . | 100.00;100.00 |
| **ctxB** | 100 | 100.00;100.00 | . | 100.00;100.00 |
| **farA** | 11 | 11 | 11 | 11 |
| **fliN** | 48.31 | 48.31 | 48.31 | 48.31 |
| **fliP** | 25.73 | 25.73 | 25.73 | 25.73 |
| **gmhA/ lpcA** | 77.26 | 77.26 | 77.26 | 77.26 |
| **hcp-2** | 100.00;100.00 | 100.00;100.00 | 100.00;100.00 | 100.00;100.00 |
| **zot** | 100 | 100.00;100.00 | . | 100.00;100.00 |

Note: NUM_FOUND: the number of virulence genes screened from the genome; row name: strain name; column name: virulence gene name; table value:% COVERAGE

# 4. *Clinical sample*




Microbial informatics and experimentation(2012). doi.org/10.1186/2042-5783-2-3

- They allow scientists to detect rare bacteria species and ones that cannot be cultivated.
- They enable the analysis of a large number of microbial samples at the same time.
- They give us a snapshot of population diversity within a sample.

Pathogen_16sPIP:

1. microbiota structure
2. 346 kinds of pathogens related to human health

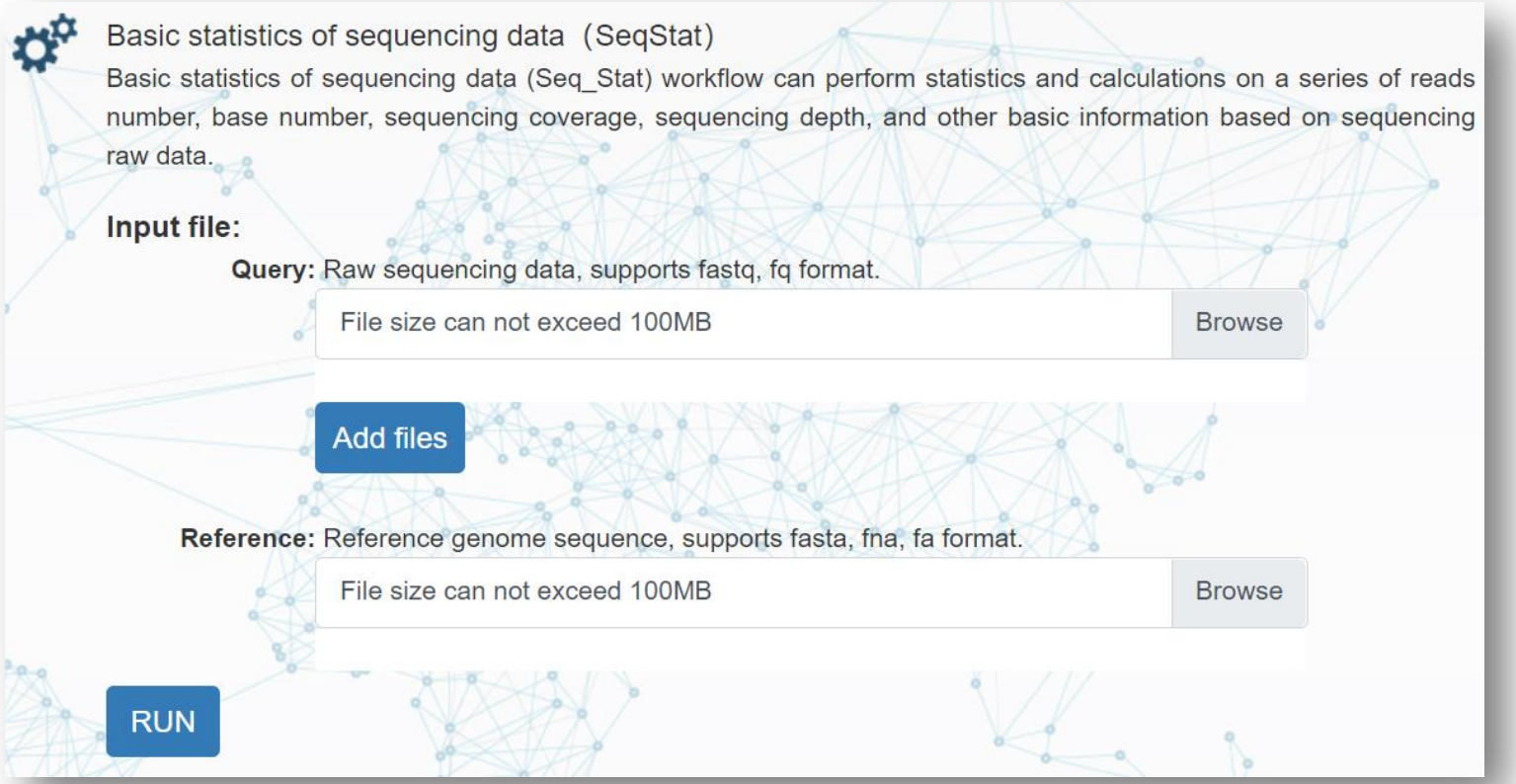
```
#=====================================================================#
#                              Summary                                #
#=====================================================================#

SampleFile: /var/data/5e748556f1e3f4001a52d2e8.fastq_trimmed_filter
Sum Data: 121.059M
Read Num: 132638
Read Len: 441
GC: 52.811
SampleSamFile: /var/data/5e748556f1e3f4001a52d2e8.fastq.sam
Match Num: 132567
Unmatch Num: 71


#=====================================================================#
#                           Test results                              #
#                The sample contains comparable strains.              #
#=====================================================================#

Species Match Num       Percentage
Pathogenic Escherichia coli     409     0.308
Prevotella copri        199     0.150
Shigella flexneri       66      0.050
Clostridium perfringens 41      0.031
Enterococcus faecium    14      0.011
Shigella sonnei 9       0.007
Streptococcus vestibularis      8       0.006
Bacteroides fragilis    4       0.003
Klebsiella pneumoniae   4       0.003
Citrobacter freundii    3       0.002
Shigella dysenteriae    3       0.002
Enterobacter asburiae   2       0.002
Streptococcus lutetiensis       2       0.002
Klebsiella oxytoca      1       0.001
Enterobacter hormaechei 1       0.001
Shigella boydii 1       0.001
Enterobacter Sakazakii  1       0.001
Edwardsiella tarda      1       0.001
Streptococcus australis 1       0.001
Streptococcus gallolyticus subsp. gallolyticus  1       0.001
Enterobacter cloacae    1       0.001
Prevotella stercorea    1       0.001
```

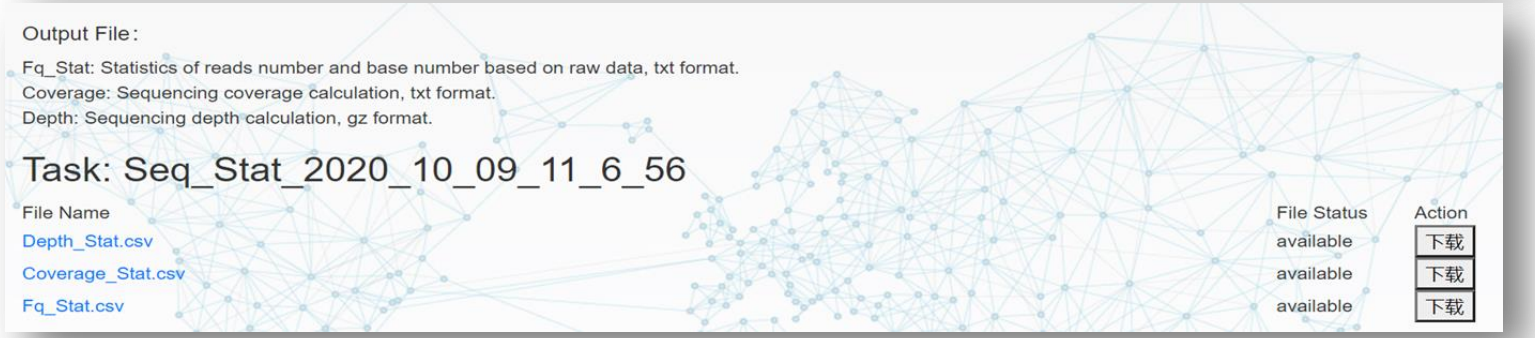
Fig 4 The composition of the flora and relative abundance of species in clinical sample

# 5

# Analysis Tools

# 1. Basic statistics of sequencing data（SeqStat）



# SeqStat : answer three questions


1. How many reads in raw data?
2. Depth in each Position?
3. The length of overlapped regions account for reference genome?

# Example:

Fq_Stat.csv

| Num reads:657120 | Num Bases: 98568000 |
|---|---|

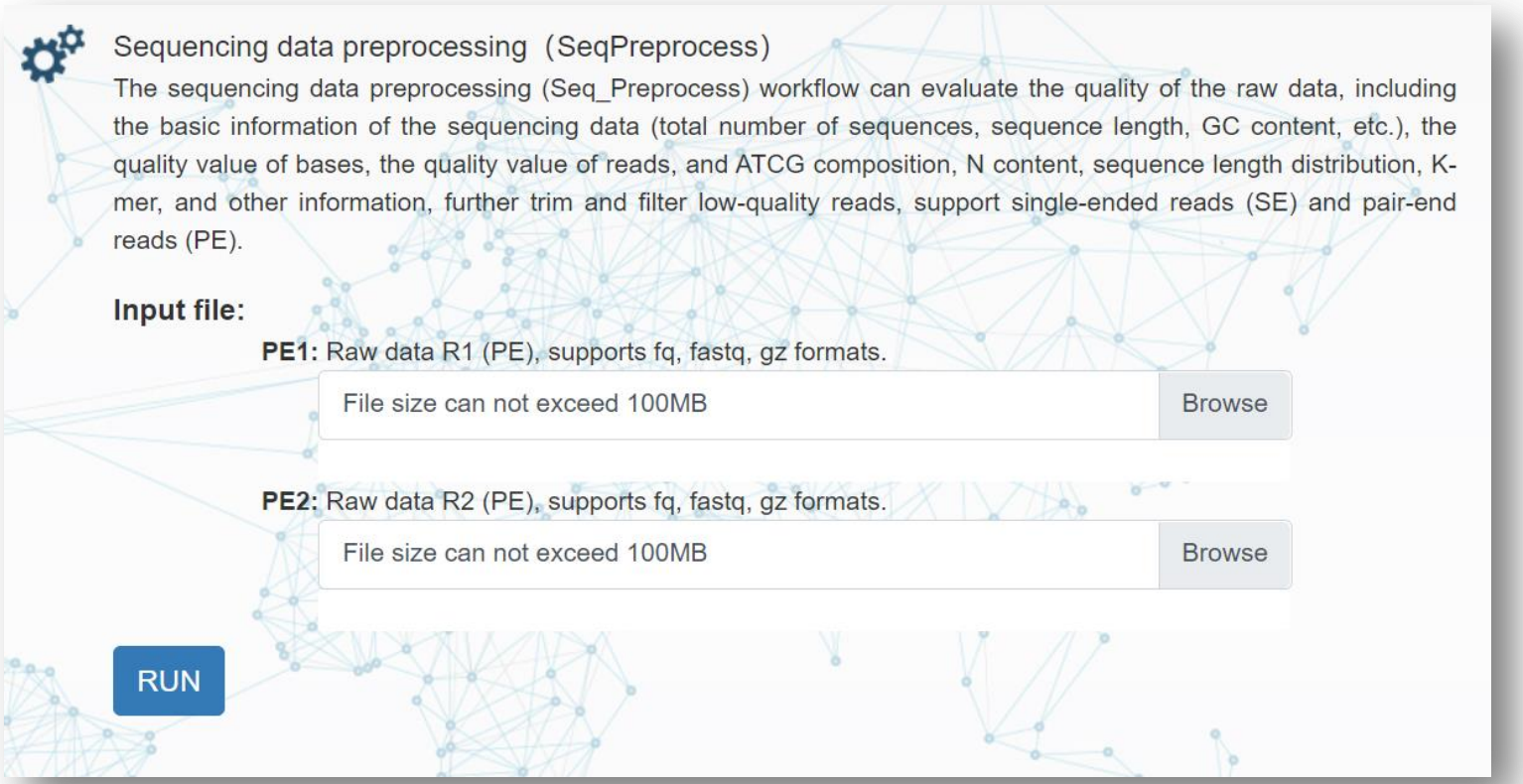
Depth_Stat.csv

| #Chr | Pos | Raw Depth | Rmdup depth | Cover depth |
|---|---|---|---|---|
| NC_004310.3 | 2 | 2 | 1 | 2 |
| NC_004310.3 | 3 | 3 | 2 | 3 |
| NC_004310.3 | 4 | 3 | 2 | 3 |
| NC_004310.3 | 5 | 3 | 2 | 3 |
| NC_004310.3 | 6 | 3 | 2 | 3 |
| …… | | | | |

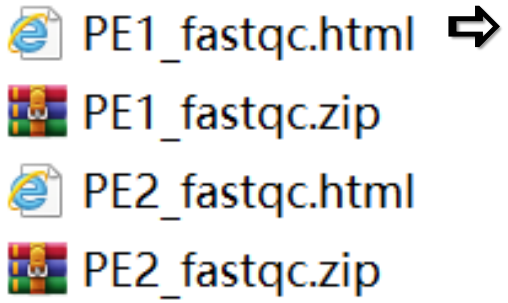
Coverage_Stat.csv

| #rname | startpos | endpos | numreads | covbases | coverage | meandepth | meanbaseq | meanmapq |
|---|---|---|---|---|---|---|---|---|
| NC_004310.3 | 1 | 2107794 | 416072 | 2102494 | 99.7486 | 29.603 | 36.6 | 41.6 |
| NC_004311.2 | 1 | 1207381 | 235061 | 1185432 | 98.1821 | 29.1965 | 36.6 | 41.7 |

# 2. Sequencing data preprocessing（SeqPreprocess）




Example:



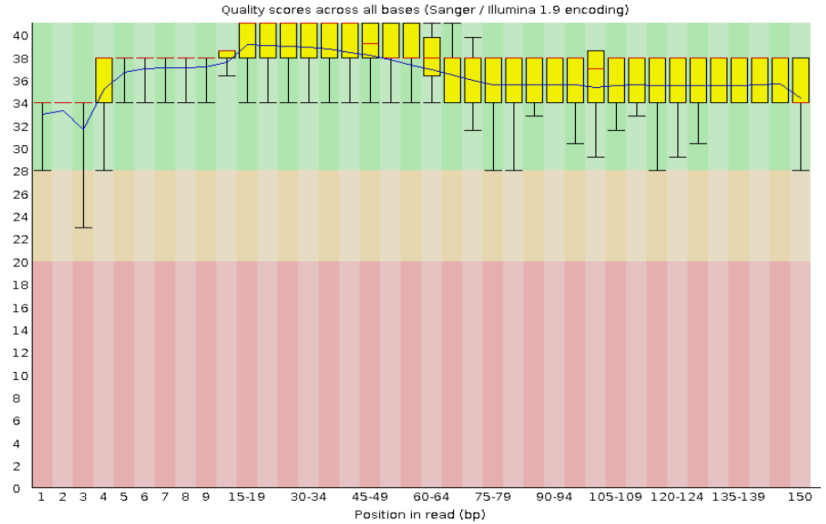

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

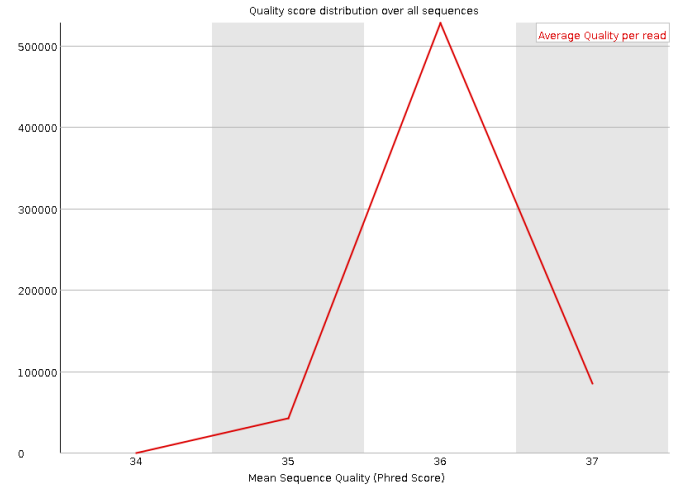
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

## Per base sequence quality



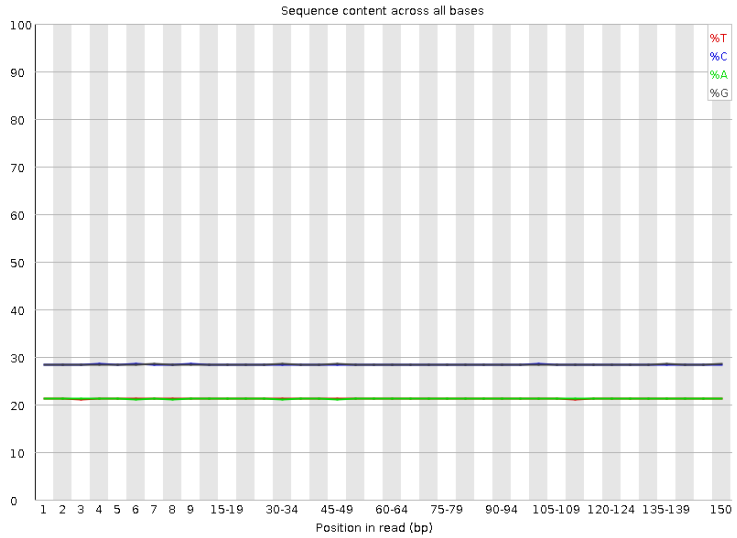

A

## Per sequence quality scores



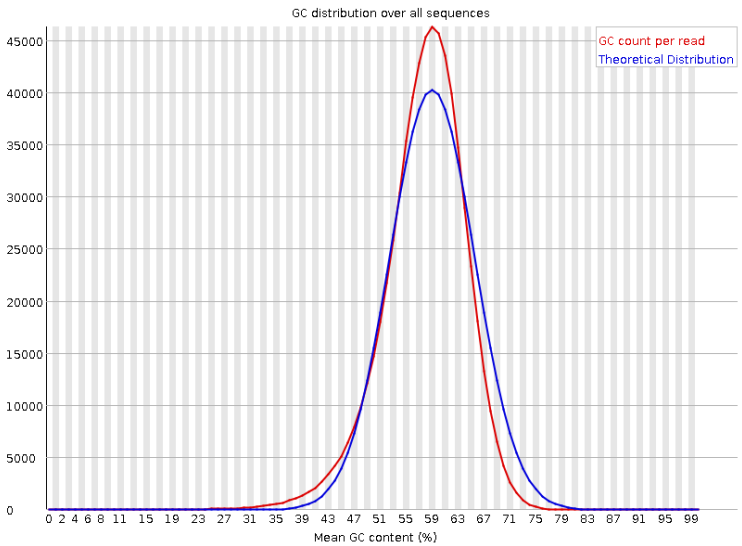

B

## Per base sequence content



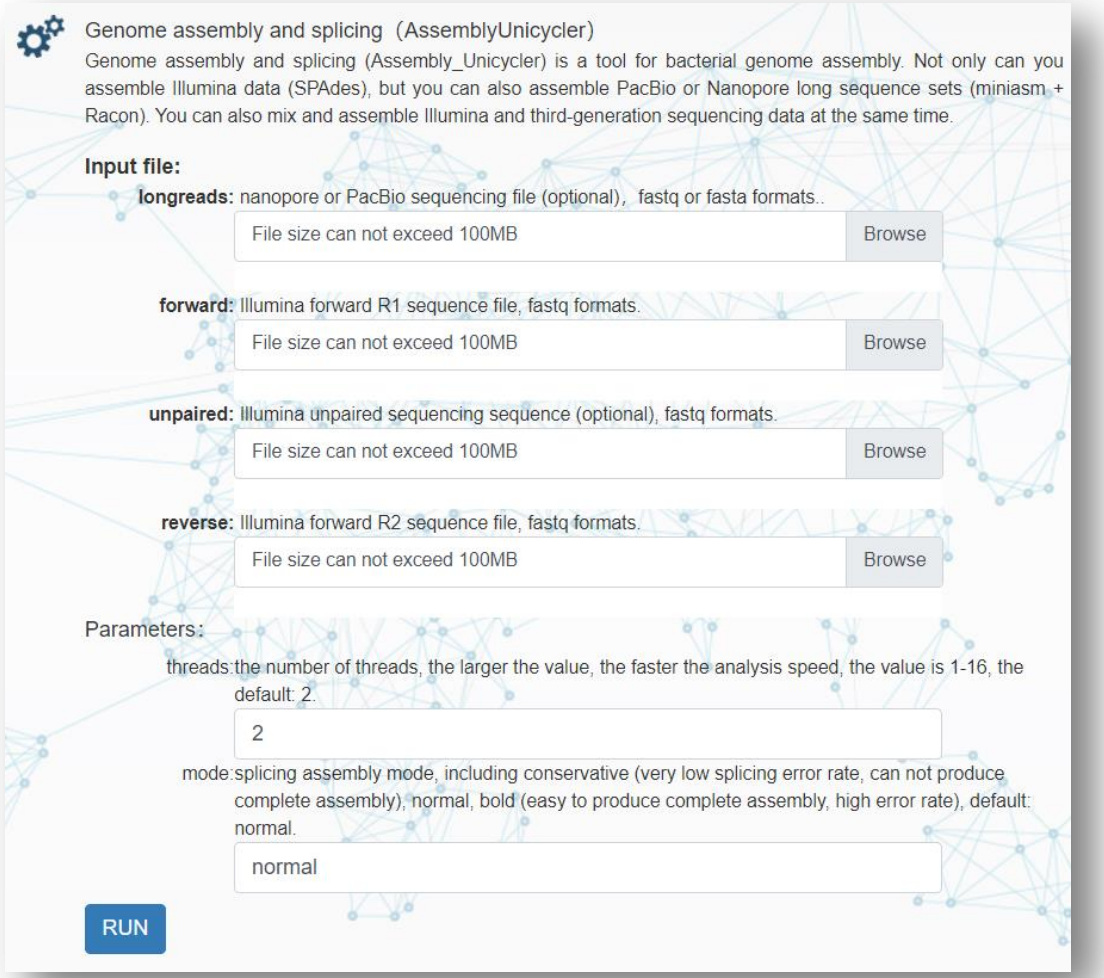

C

## Per sequence GC content




D

# 3. Genome assembly and splicing（AssemblyUnicycler）



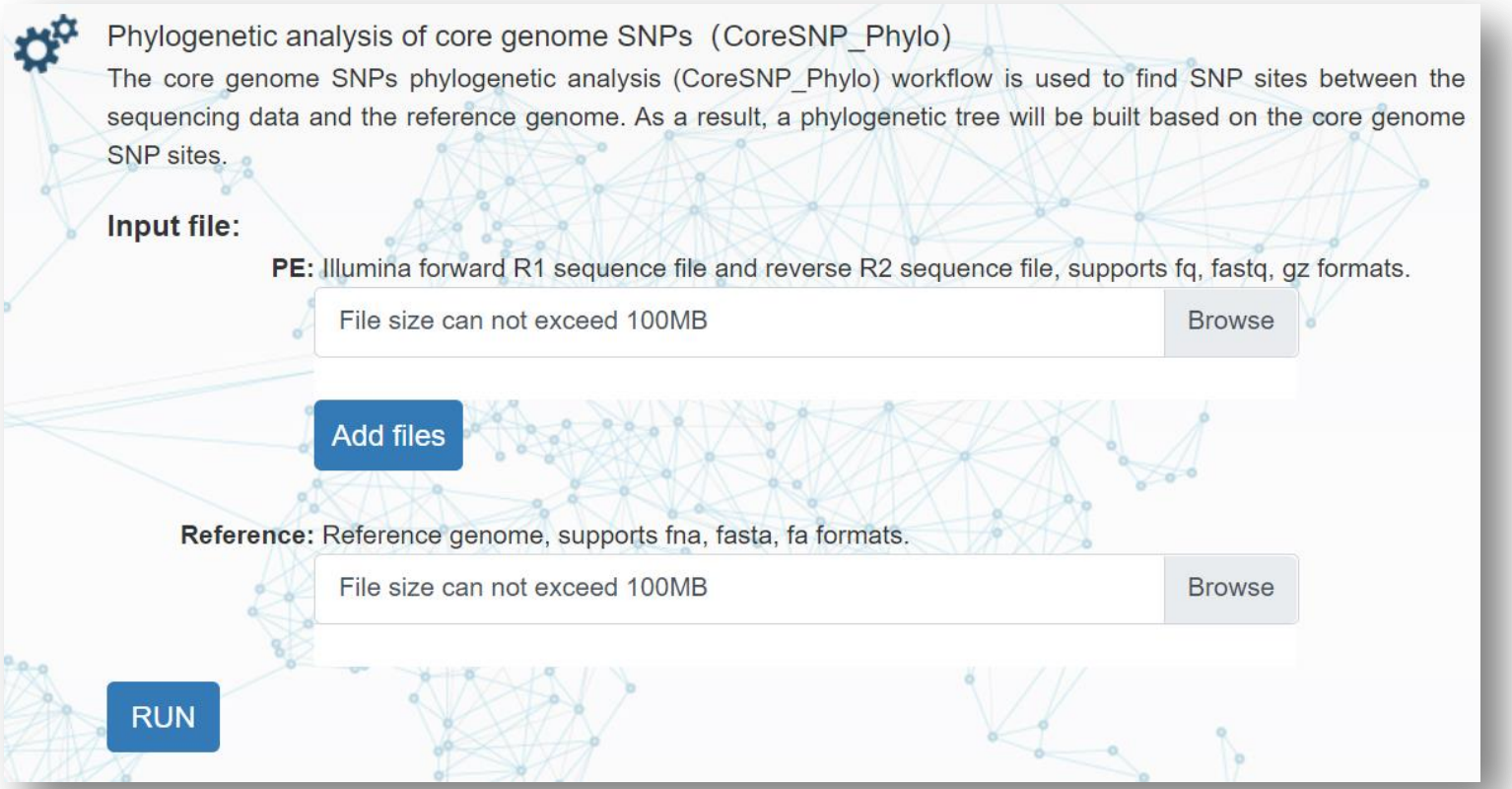

# Example:

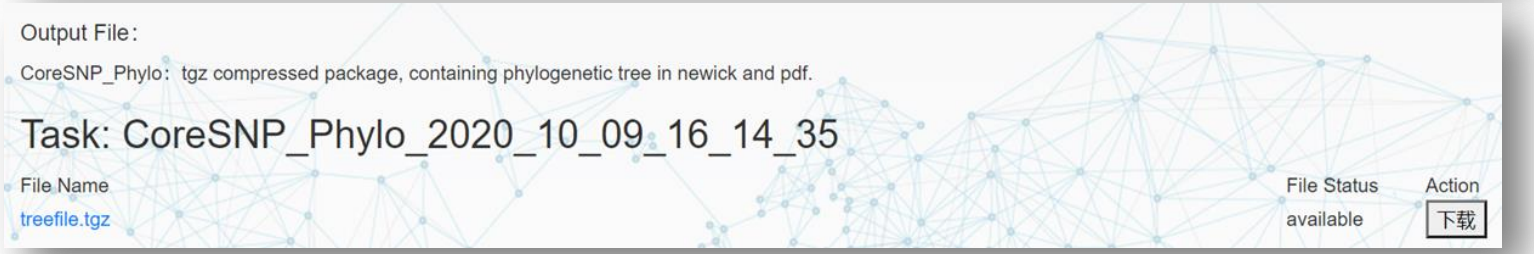
- 001_best_spades_graph.gfa
- 002_overlaps_removed.gfa
- 003_bridges_applied.gfa
- 004_final_clean.gfa
- 005_polished.gfa
- assembly.fasta ⇨
- assembly.gfa
- unicycler.log

**assembly.fasta**

```
>1 length=358774 depth=1.00x
AGCGGTTCCTGTTTTAACAGAATCGCCGGAACCGCTCTAACTATTTGTTTTGTCGCATTATCCAACGCAA
AACCGTTTCACACTTTTGCTGGAAATGCTCTATGCCGTTTTGGCCCGATGCACAAGCAGGAGTTCGCGCG
TCATGACCTGAAAGGGGCTCAAACAAAAACGCCCTCTCCGGATGGAAAGAGCGTTTAAACTATACCCGGC
TTGCGCCCGATCAGAGATCGACGTCGAGAATGGCCATCGAGAAATTATACGACAATTCGCCTTCATCCTC
ATCCTTGTAGATAAGGCCGAGAAATTCGTCGGCGAGGTAAAGCTCGGCGGAATCATTCTTGCGCGGACGC
GCCTTCACCTGAAGATCAGGATTGTTGAAGGTCCGTTTGAAATAGGCGTCCAGTTTTTTGAGTTCTTCGG
GTTTCAACCGGATTCTCCTGACATTGATTTGAAGCCGGTTTATTGCACCCGGCATTTTCTTATGTAAACC
CCTGACATCGCCTATCGCAGGCTGTAGCGACGAATTAACTTTCGGTTTGGCATGGGCAAAAGTGCGAAGC
GAGGTCAGAAATCAGTCCATTGAACTGATTTCCCCGCATGGACGTTTCACATTTTGGGCCCGAAGACGCT
CCAAACAAAGCTTTCACCACATATAGGCCGTCTGCGGGATAAAACCAGCAGACAACCCAACTTTATGGCA
TGAGCCGTTATCAATATTGGAATTAGATGCCGTCACCCAGCATCTGGTCCATCACGCGGGATGGCTCCGT
GCAGCCCGTTTCGCCGATGATGCGCGCAGGCACGCCCGCCACGGTTACATTATGCGGCACGGACTTCAGC
ACAACGGAACCAGCCGCGATCTTGGAGCACTGCCCTACCTGGATGTTGCCGAGGATTTTTGCGCCTGCCC
CAATCAGCACGCCCTGACGGATTTTCGGATGGCGATCACCGCTCGATTTGCCGGTTCCGCCCAGCGTAAC
CCCATGCAGGATCGAGACATTATCTTCTACAACGGCCGTTTCACCCACCACAAGCCCTGTAGCATGGTCG
AGAAACAGCCCGCTGCCAAGCCTGGCCGCCGGGTGAATATCGGTCTGAAAGATCGATGACGAGCGGCTCT
GAAGATAATAGGCGAAATCCTTGCGACCTTGCTTGTAAAGCCAGTGCGCCAGACGGTGTGTCTGGATCGC
ATGGAAGCCCTTCAGGTAAAGAACCGGATCCATAAAGCGGCTATAGGCCGGATCGCGGTCATAAACAGCC
TGAATATCCACACGCAGGACGTGCGACCATTCCGGGTTTGCCTCAAGCATCGTATCGAAGGTCTGGCGCA
GGATATCCGCGCTCACATCGGGATGGCCAAGGCGTTCGGCGATACGATGCATCACCGCTTCTTCAAGGCT
```

# 4. Phylogenetic analysis of core genome SNPs（CoreSNP_Phylo）



# Example:

treefile.tgz

⤵

core.newick

Rplots.pdf

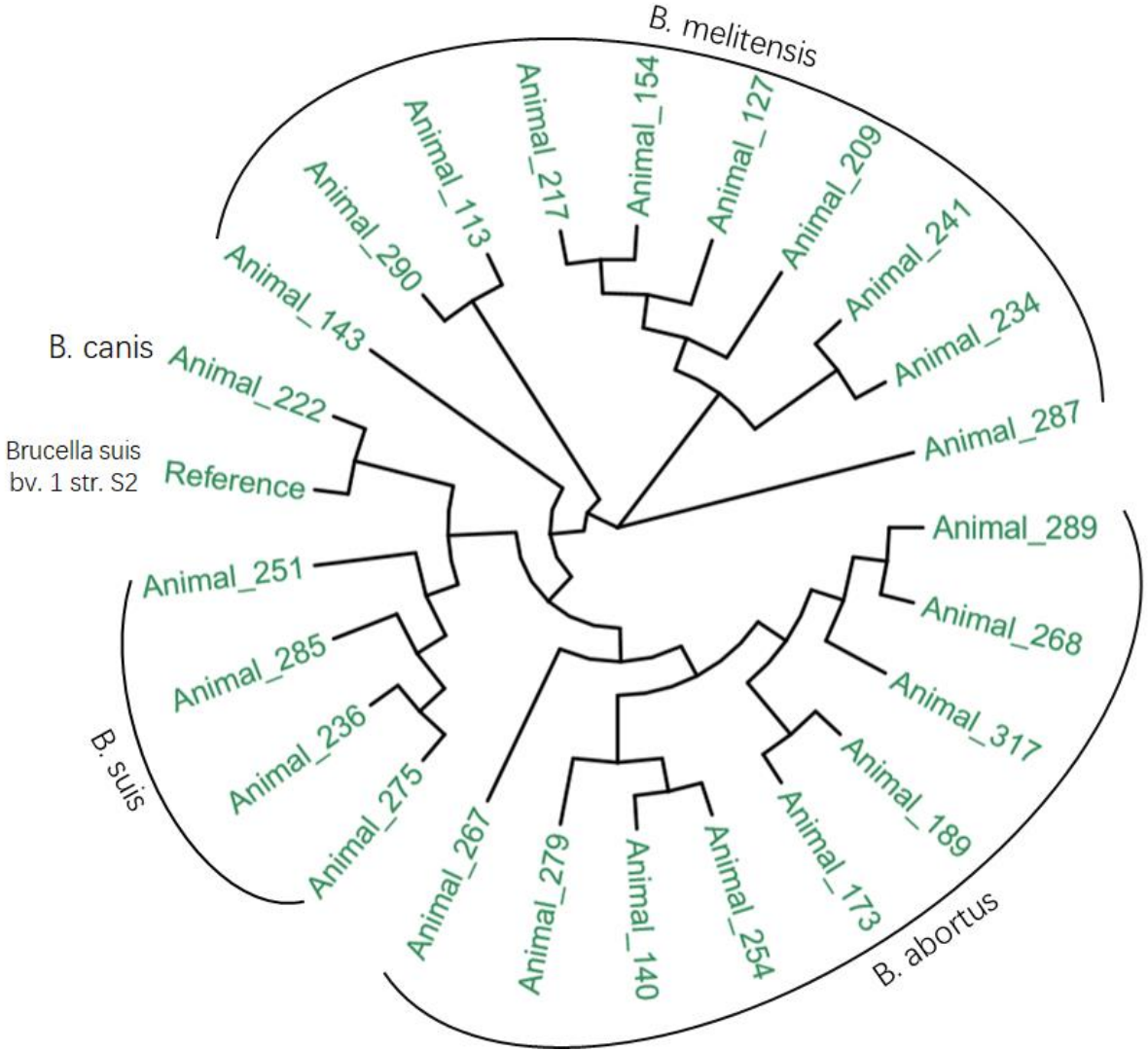
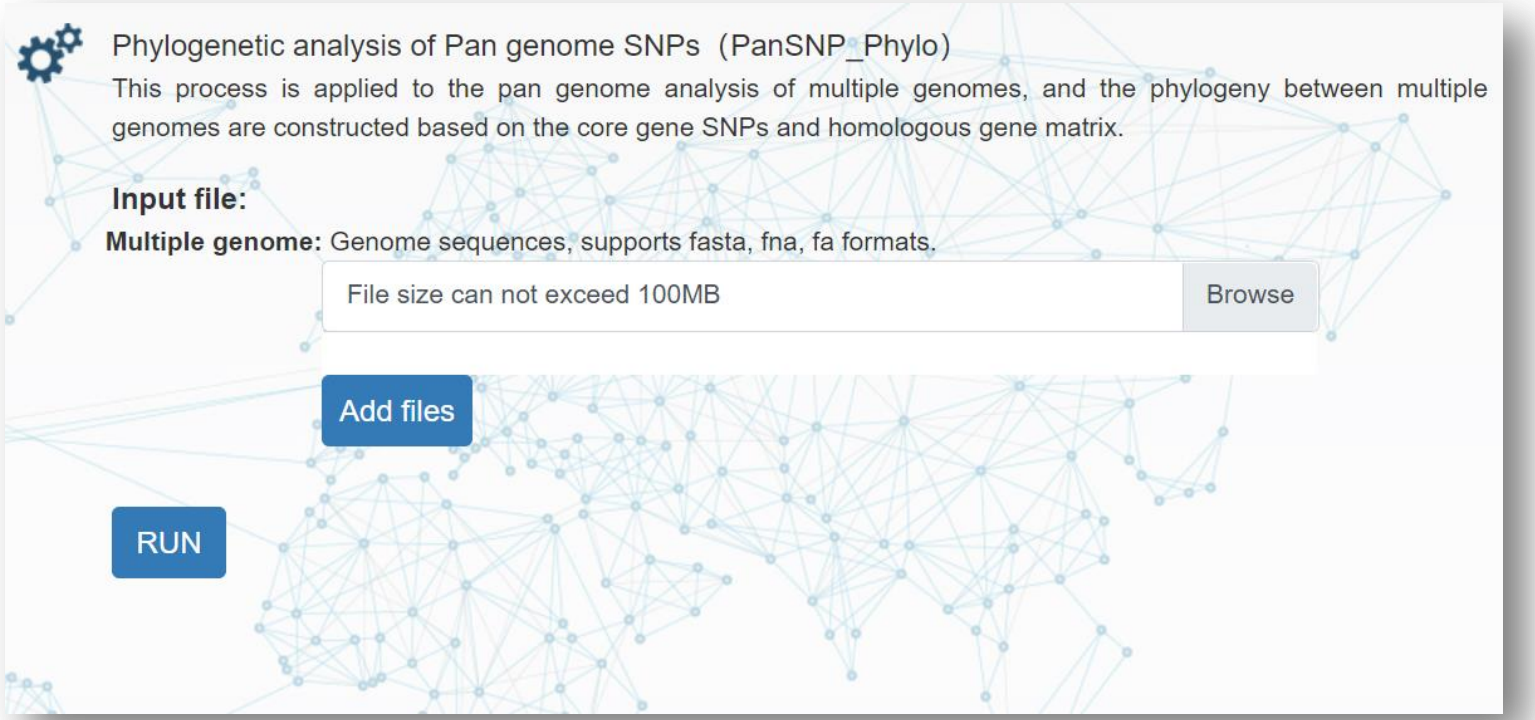
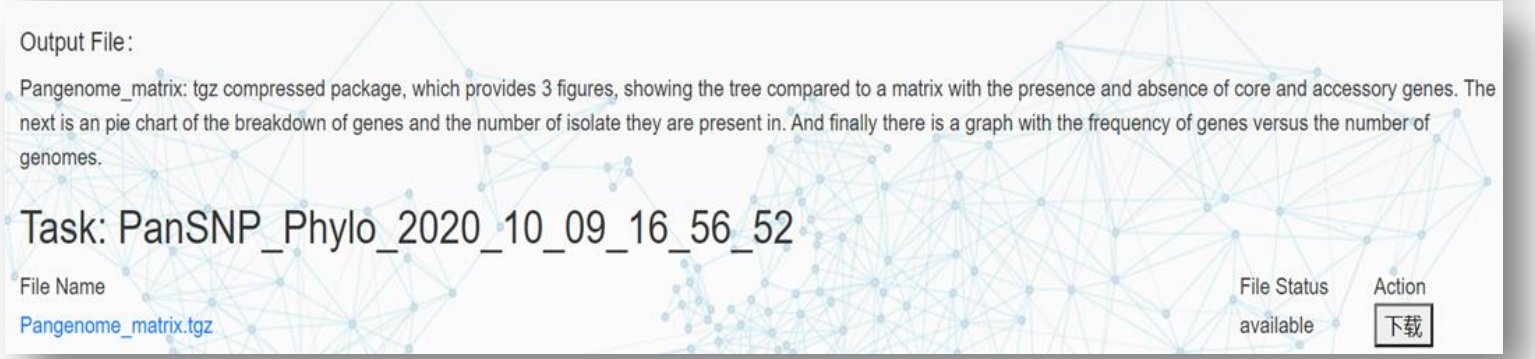



Figure 1. Core SNP-based phylogenetic tree of 24 Brucella strains

# 5. Phylogenetic analysis of Pan genome SNPs（PanSNP_Phylo）



# Example:

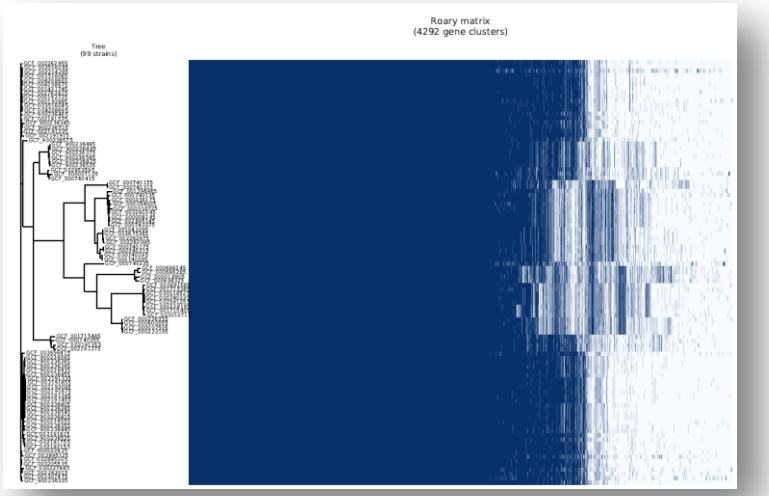



Fig 1. pangenome_matrix

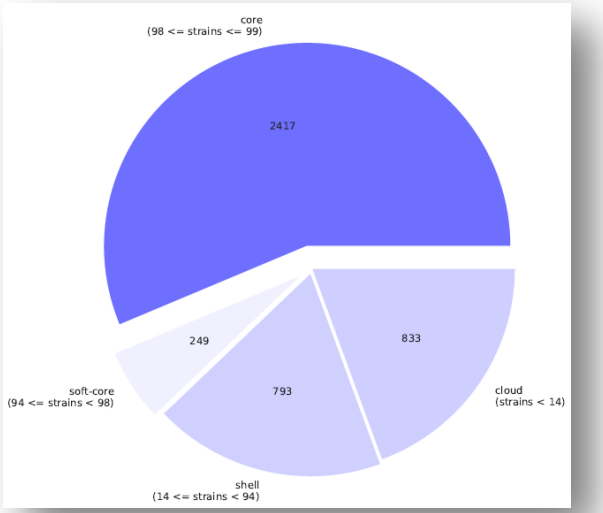



Fig 2. pangenome_pie

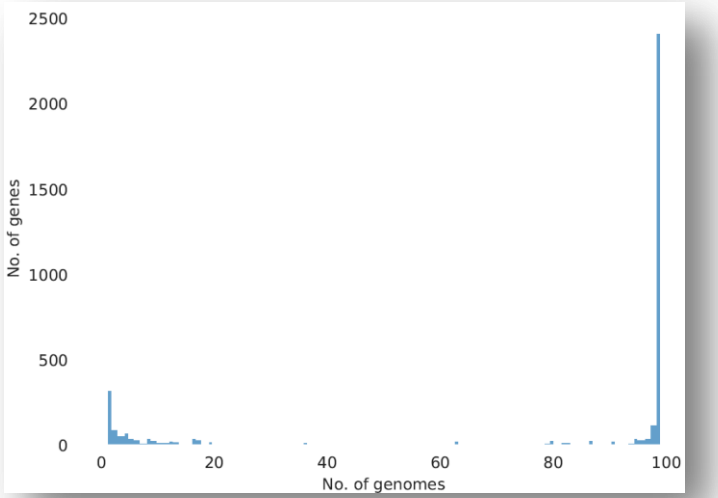
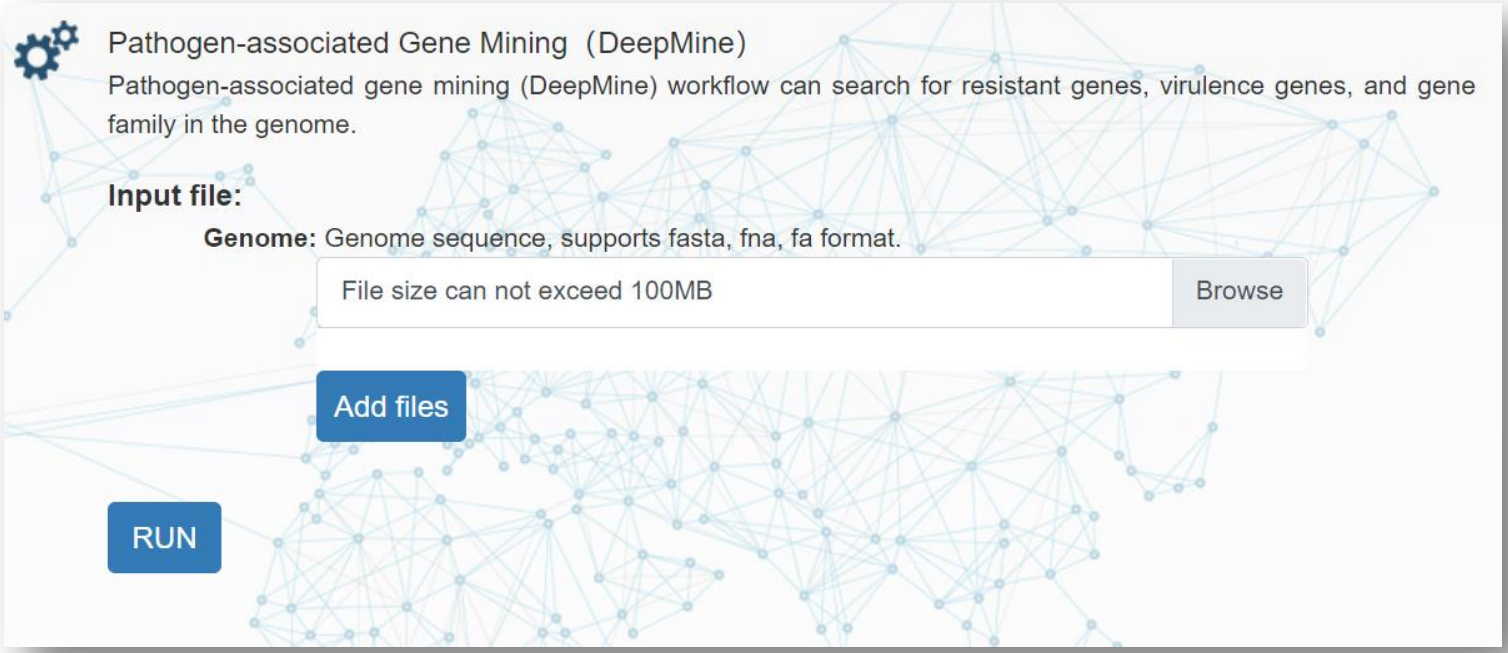
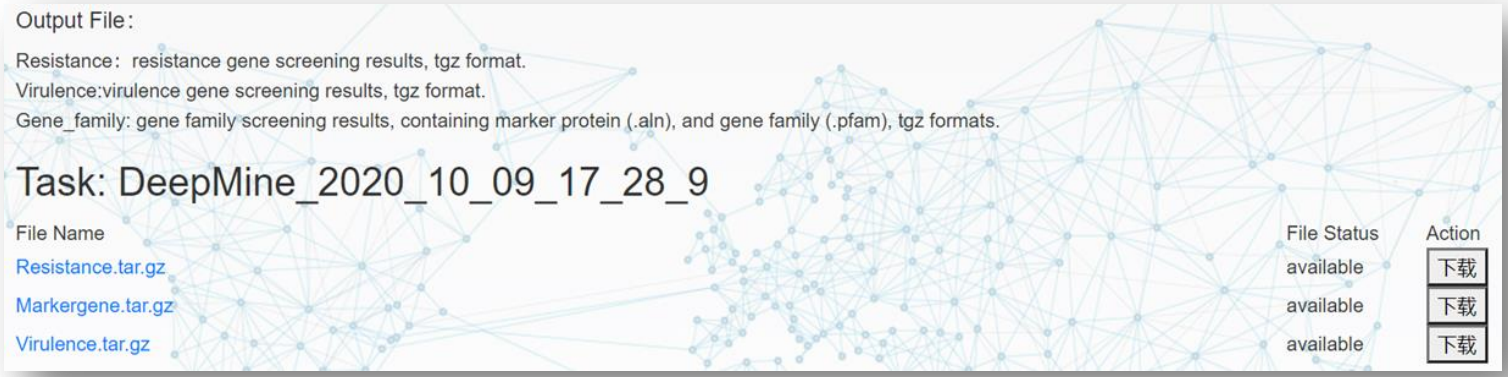



Fig 3. pangenome_frequency

# 6. Pathogen-associated Gene Mining（DeepMine）



# Example:

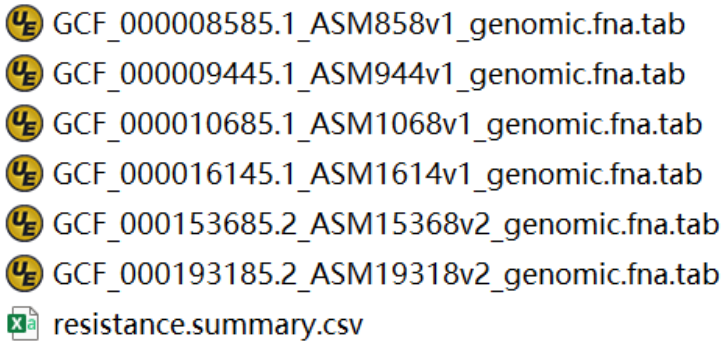



Fig 1. Resistance

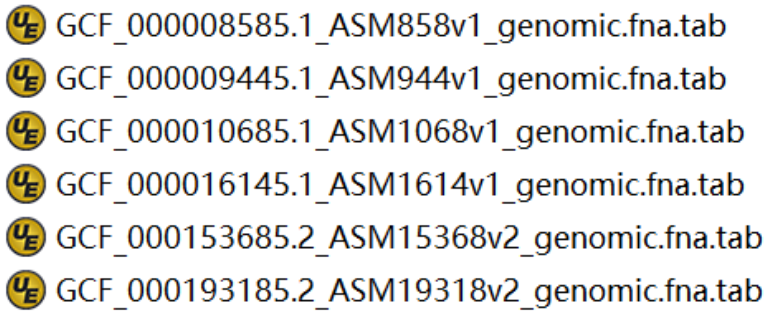



Fig 2. Virulence

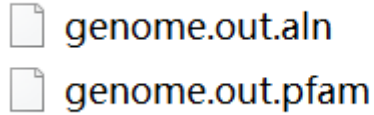



Fig 3. Markergene

Example: Resistance

Table 1.1 Screening results of genome-wide Mycobacterium tuberculosis resistance genes

| #FILE | NUM_FOUND | AAC(2')-Ic | Erm(37) | Erm(38) | rpoB2 | RbpA | efpA | mfpA | mtrA |
|---|---|---|---|---|---|---|---|---|---|
| GCF_000008585.1_ASM858v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000009445.1_ASM944v1_genomic.fna | 8 | 100 | 69.26 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000010685.1_ASM1068v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000016145.1_ASM1614v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000016925.1_ASM1692v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000023625.1_ASM2362v1_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000153685.2_ASM15368v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000154585.2_ASM15458v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000154605.2_ASM15460v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |
| GCF_000193185.2_ASM19318v2_genomic.fna | 8 | 100 | 100 | 16.45 | 97.13 | 91.01 | 100 | 100 | 100 |

Note: NUM_FOUND: the number of drug-resistant genes screened from the genome; row name: drug-resistant gene name; column name: strain name; table value:% COVERAGE

Table 1.2 Mycobacterium tuberculosis (GCF_000008585.1_ ASM858v1_genomic.fna) resistance gene screening results

| #FILE | SEQUENCE | START | END | STRAND | GENE | COVERAGE | COVERAGE_MAP | GAPS | %COVERAGE | %IDENTITY | DATABASE | ACCESSION | PRODUCT | RESISTANCE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 314424 | 314969 | - | AAC(2')-Ic | 1-546/546 | ========== | 0/0 | 100 | 100 | card | AL123456.3:314309-314855 | AAC(2')-Ic | aminoglycoside |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 761855 | 765274 | + | rpoB2 | 64-3462/3489 | ======/==== | 19/41 | 97.13 | 79.56 | card | AP006618.1:4835200-4838689 | rpoB2 | rifamycin |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 2229013 | 2229552 | + | Erm(37) | 1-540/540 | ========== | 0/0 | 100 | 100 | card | AL123456:2231680-2232220 | Erm(37) | streptogramin/macrolide/lincosamide |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 2310154 | 2310467 | + | RbpA | 1-314/345 | ========== | 0/0 | 91.01 | 87.58 | card | HQ203032:1-346 | RbpA | rifamycin |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3147251 | 3148843 | - | efpA | 1-1593/1593 | ========== | 0/0 | 100 | 100 | card | AL123456.3:3153039-3154632 | efpA | phenicol |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3621563 | 3622249 | - | mtrA | 1-687/687 | ========== | 0/0 | 100 | 100 | card | AL123456.3:3626663-3627350 | mtrA | antibacterial_free_fatty_acids |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3735428 | 3735618 | + | Erm(38) | 777-967/1161 | .....===.. | 0/0 | 16.45 | 76.44 | card | AY154657.2:63-1224 | Erm(38) | streptogramin/macrolide/lincosamide |
| **GCF_000008585.1_ASM858v1_genomic.fna** | NC_002755.2 | 3765240 | 3765791 | - | mfpA | 1-552/552 | ========== | 0/0 | 100 | 100 | card | AL123456:3773016-3773568 | mfpA | fluoroquinolone |

# Example: Virulence

Table 2. Genome-based screening results of Vibrio cholerae virulence genes

| #FILE | GCA_000006745.1_ ASM674v1_genomic.fna | GCA_000016245.1_ ASM1624v1_genomic.fna | GCA_000021605.1_ ASM2160v1_genomic.fna | GCA_00002c1625.1_ ASM2162v1_genomic.fna |
|---|---|---|---|---|
| **NUM_ FOUND** | 56 | 55 | 52 | 55 |
| **IlpA** | 100 | 100 | 100 | 100 |
| **VCA0109** | 100 | 97.26 | 100 | 97.26 |
| **VCA0122** | 100 | 100 | 100 | 100 |
| **ace** | 100 | 100.00;100.00 | . | 100.00;100.00 |
| **acfA** | 100 | 100 | 100 | 100 |
| **acfB** | 100.00;7.18 | 100.00;7.18 | 100.00;7.18 | 100.00;7.18 |
| **acfC** | 100 | 100 | 100 | 100 |
| **acfD** | 100 | 100 | 100 | 100 |
| **cheD** | 6.03;9.14;11.17 | 11.17;9.14;6.03 | 6.03;9.14;8.72;11.17 | 6.03;9.14;11.17 |
| **clpB/ vasG** | 100 | 100 | 100 | 100 |
| **cqsA** | 100 | 100 | 100 | 100 |
| **ctxA** | 100 | 100.00;100.00 | . | 100.00;100.00 |
| **ctxB** | 100 | 100.00;100.00 | . | 100.00;100.00 |
| **farA** | 11 | 11 | 11 | 11 |
| **fliN** | 48.31 | 48.31 | 48.31 | 48.31 |
| **fliP** | 25.73 | 25.73 | 25.73 | 25.73 |
| **gmhA/ lpcA** | 77.26 | 77.26 | 77.26 | 77.26 |
| **hcp-2** | 100.00;100.00 | 100.00;100.00 | 100.00;100.00 | 100.00;100.00 |
| **zot** | 100 | 100.00;100.00 | . | 100.00;100.00 |

# Example: Markergene

Table 3 List of gene families related to Brucella “genus” level gene-related protein family(partial)

| PFAM ID | family information |
|---|---|
| **PF06577.11** | DUF1134, Protein |
| **PF13779.5** | DUF4175, Domain |
| **PF02729.20** | OTCace_N, Aspartate/ornithine |
| **PF02823.15** | ATP-synt_DE_N, ATP |
| **PF03695.12** | UPF0149, Uncharacterised |
| **PF13478.5** | XdhC_C, XdhC |
| **PF00697.21** | PRAI, N-(5'phosphoribosyl)anthranilate |
| **PF00933.20** | Glyco_hydro_3, Glycosyl |
| **PF00067.21** | p450, Cytochrome |
| **PF00902.17** | TatC, Sec-independent |
| **PF01967.20** | MoaC, MoaC |
| **PF03352.12** | Adenine_glyco, Methyladenine |
| **PF03947.17** | Ribosomal_L2_C, Ribosomal |
| **PF07310.12** | PAS_5, PAS |
| **PF00366.19** | Ribosomal_S17, Ribosomal |
| **PF03239.13** | FTR1, Iron |
| **PF01148.19** | CTP_transf_1, Cytidylyltransferase |
| **PF04392.11** | ABC_sub_bind, ABC |
| **PF01121.19** | CoaE, Dephospho-CoA |
| **PF04279.14** | IspA, Intracellular |
| **PF03461.14** | TRCF, TRCF |
| **PF10984.7** | DUF2794, Protein |

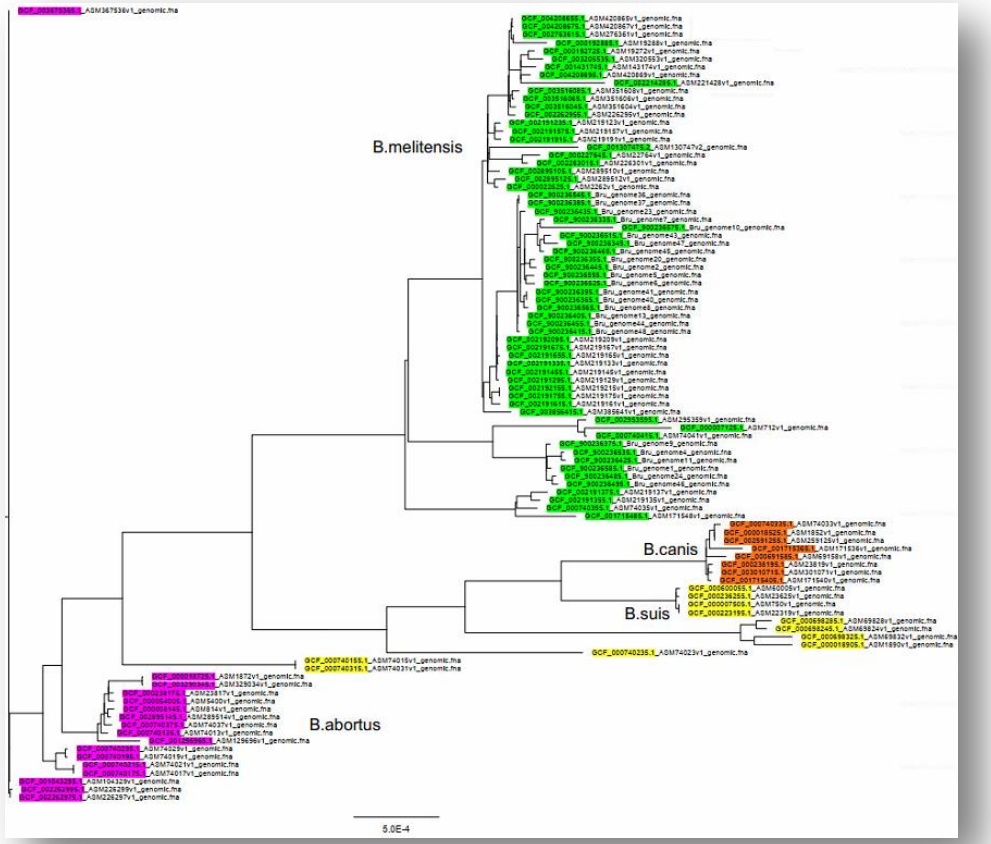
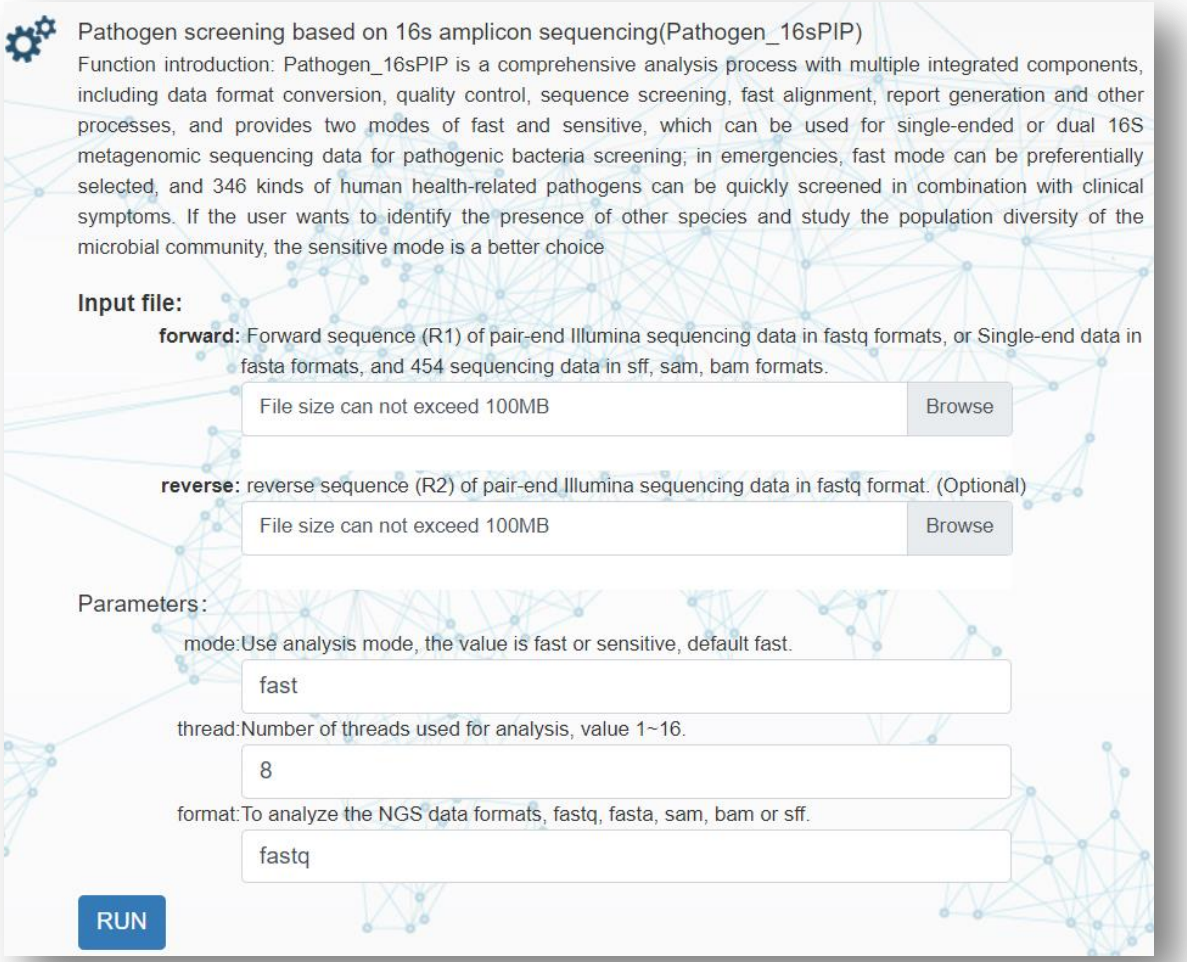



Figure3. Phylogenetic tree based on Brucella "genus" level gene-related protein sequence

# 7. Pathogen screening based on 16s amplicon sequencing(Pathogen_16sPIP)




# Example:

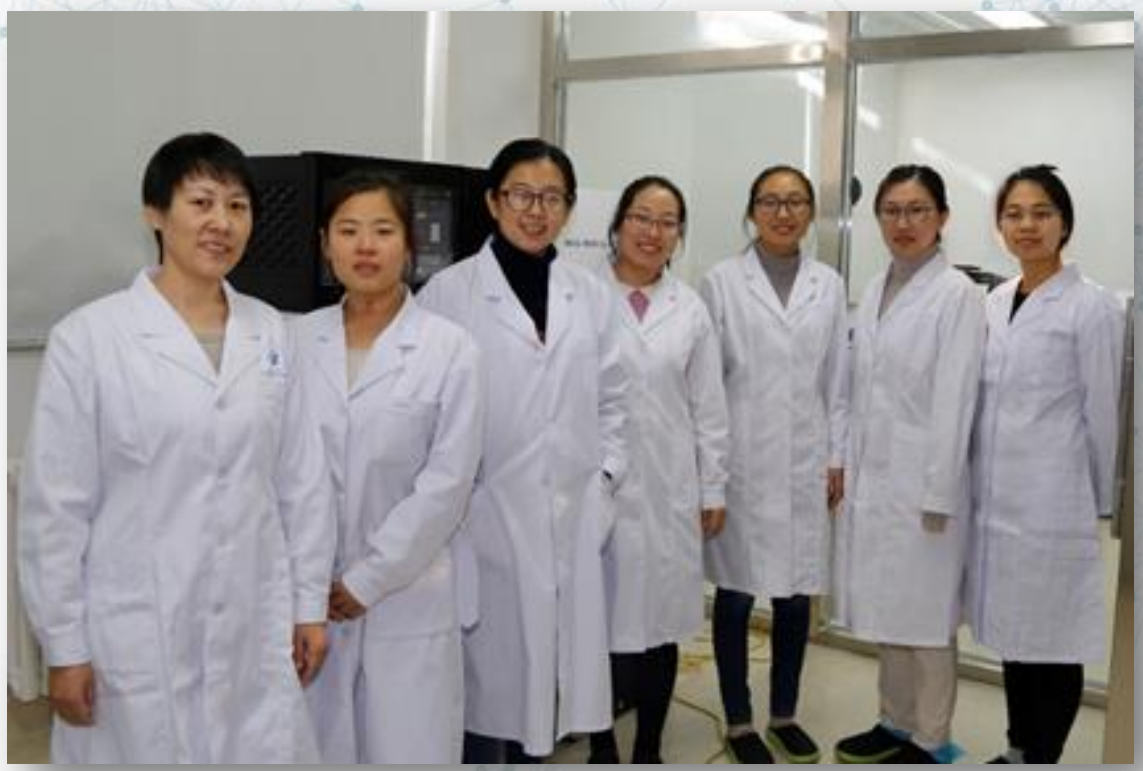
```
#=========================================================================#
#                                 Summary                                 #
#=========================================================================#

SampleFile: /var/data/5e748556f1e3f4001a52d2e8.fastq_trimmed_filter
Sum Data: 121.059M
Read Num: 132638
Read Len: 441
GC: 52.811
SampleSamFile: /var/data/5e748556f1e3f4001a52d2e8.fastq.sam
Match Num: 132567
Unmatch Num: 71


#=========================================================================#
#                              Test results                               #
#                  The sample contains comparable strains.                #
#=========================================================================#

Species Match Num       Percentage
Pathogenic Escherichia coli     409     0.308
Prevotella copri        199     0.150
Shigella flexneri       66      0.050
Clostridium perfringens 41      0.031
Enterococcus faecium    14      0.011
Shigella sonnei 9       0.007
Streptococcus vestibularis      8       0.006
Bacteroides fragilis    4       0.003
Klebsiella pneumoniae   4       0.003
Citrobacter freundii    3       0.002
Shigella dysenteriae    3       0.002
Enterobacter asburiae   2       0.002
Streptococcus lutetiensis       2       0.002
Klebsiella oxytoca      1       0.001
Enterobacter hormaechei 1       0.001
Shigella boydii 1       0.001
Enterobacter Sakazakii  1       0.001
Edwardsiella tarda      1       0.001
Streptococcus australis 1       0.001
Streptococcus gallolyticus subsp. gallolyticus  1       0.001
Enterobacter cloacae    1       0.001
Prevotella stercorea    1       0.001
```

Fig 4 The composition of the flora and relative abundance of species in clinical samples

# Pathogen In-depth Analysis Cloud Platform

http://beltroad.bio-it.cn/

# Our team



Dr. Zhang Wen

Dr. Hanna

Zhang Tingting

Qiang Yujun

Li Xiuwen

# THANK YOU FOR WATCHING !

pengxianhui@icdc.cn