



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

PART

1

Overview

PART

2

Platform Support

PART

3

Analysis Workflow

PART

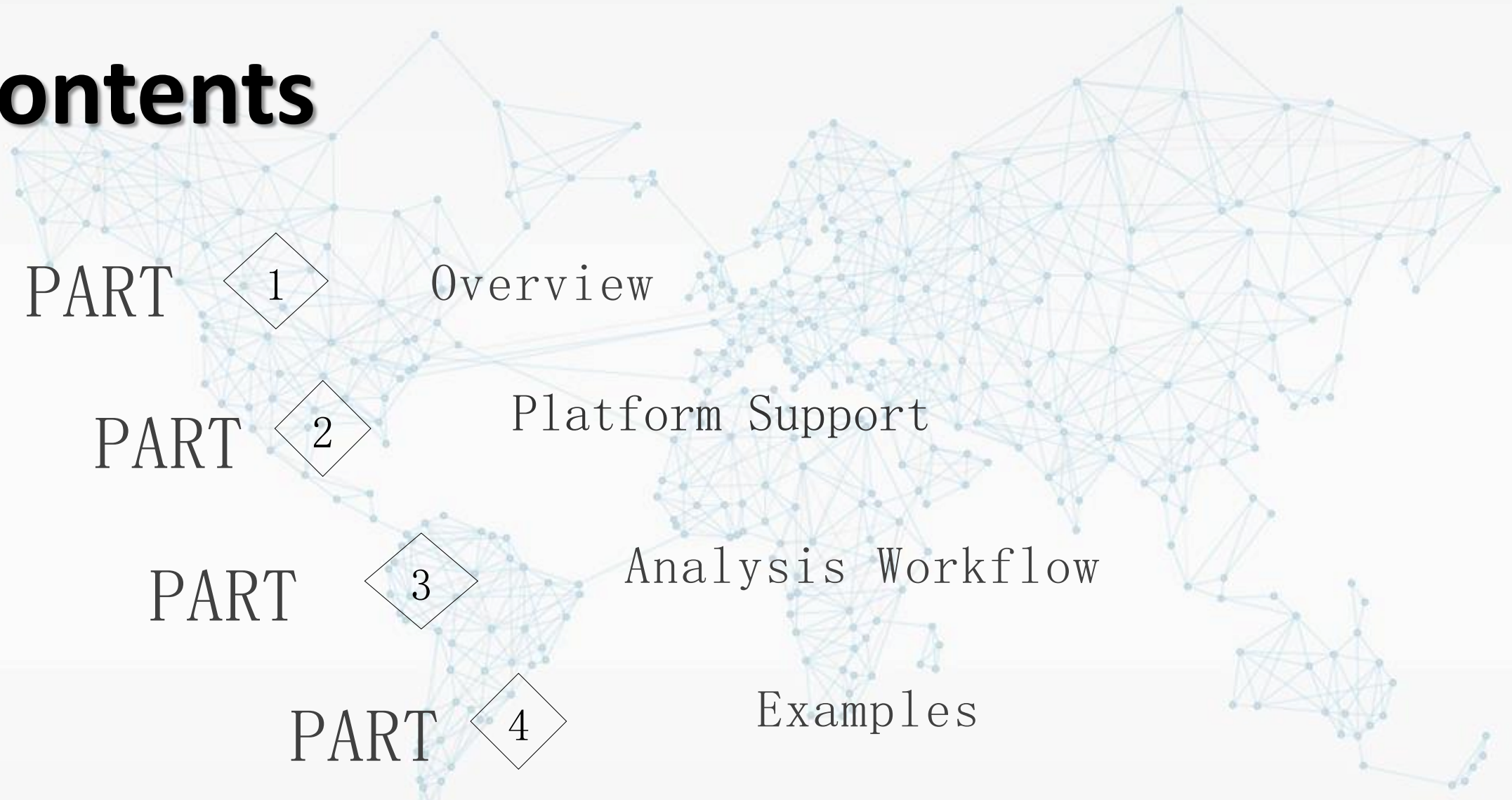
4

Examples

PART

5

Analysis Tools





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.

data share
Data sharing

点对点共享方式
—— 确保数据分享机制的安全性和可控性

The graphic depicts a network of server racks. At the top, there are four server racks. Below them, a series of blue lines connect to five circular nodes on the ground, representing a distributed network or data sharing mechanism.

2. Microbial Data Analysis Cloud Platform

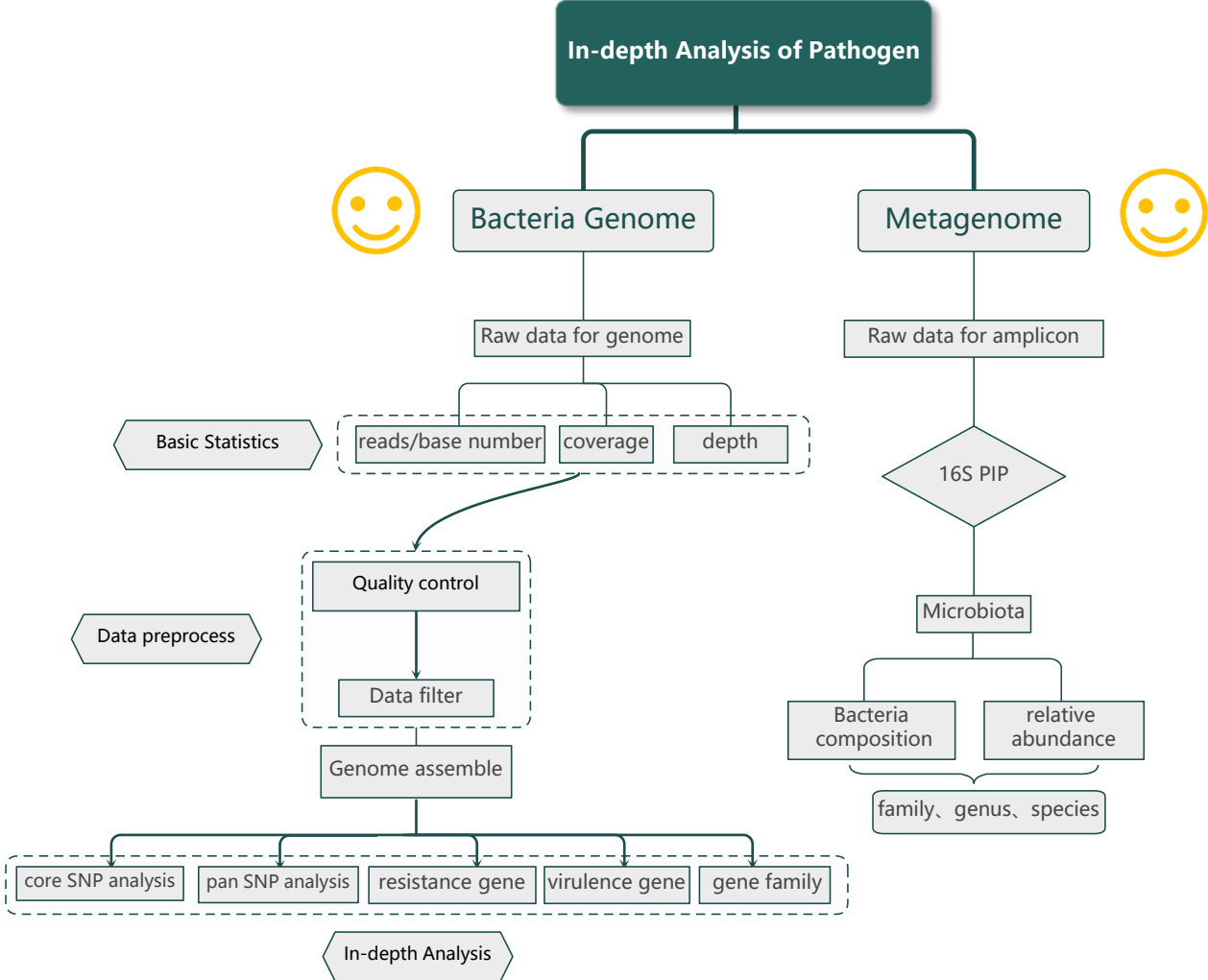
<https://analysis.mypathogen.org/>





Analysis Workflow

Analysis Workflow





4

Example

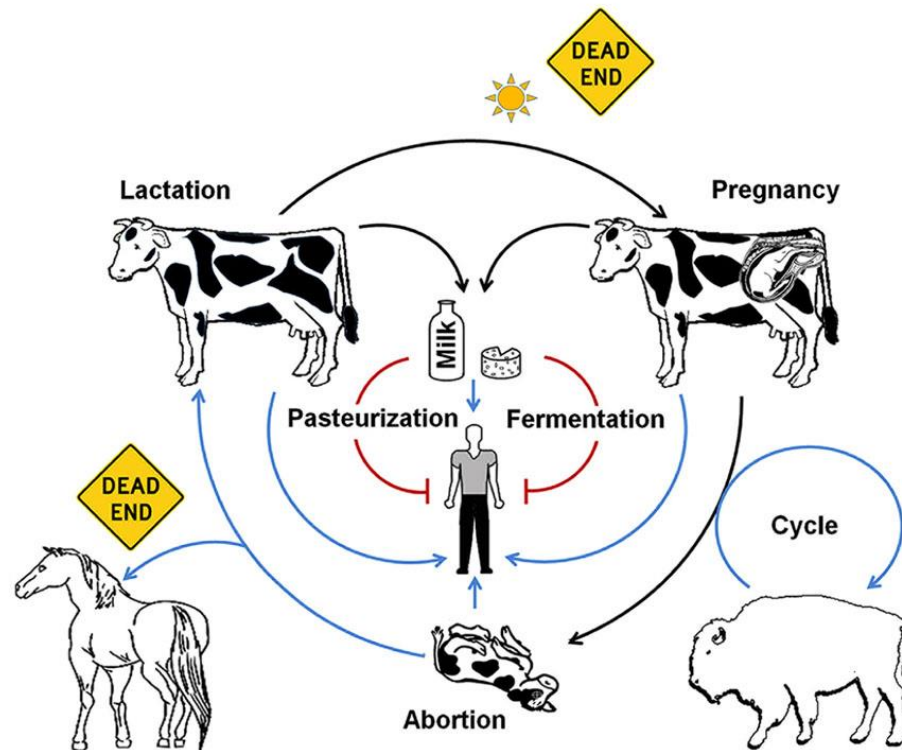
Example



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

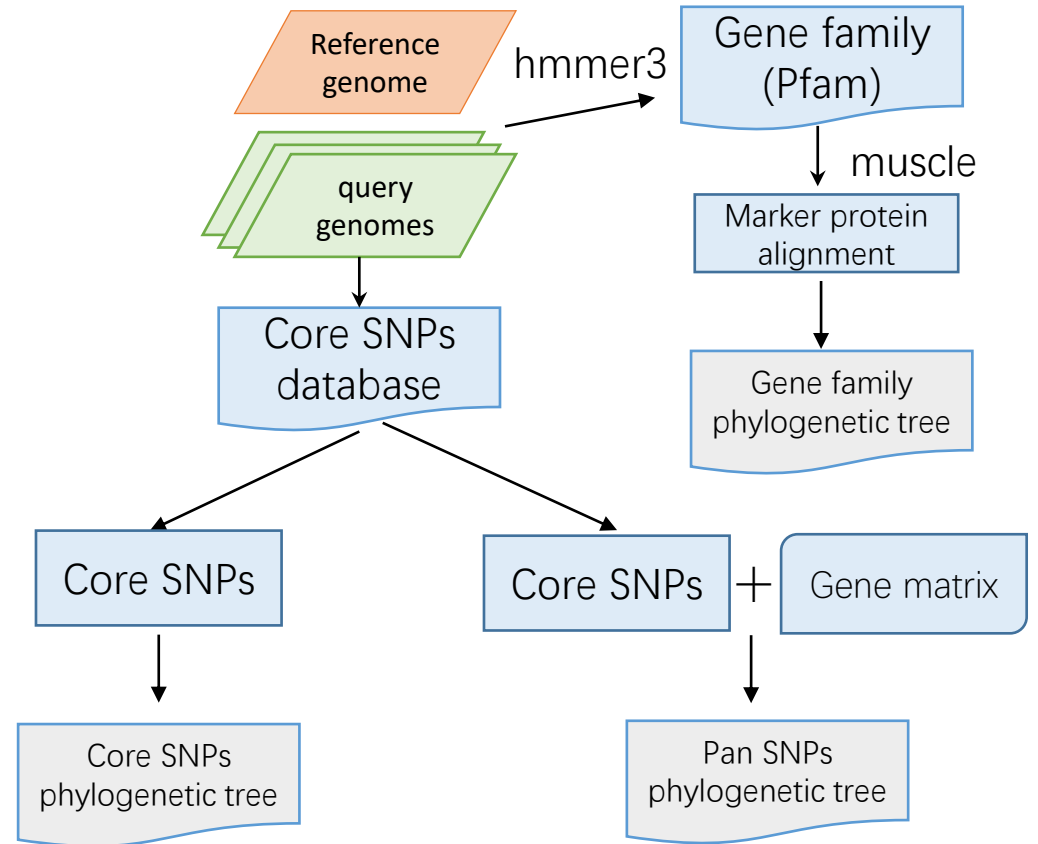


Fig 1. Three methods based on whole genome analysis

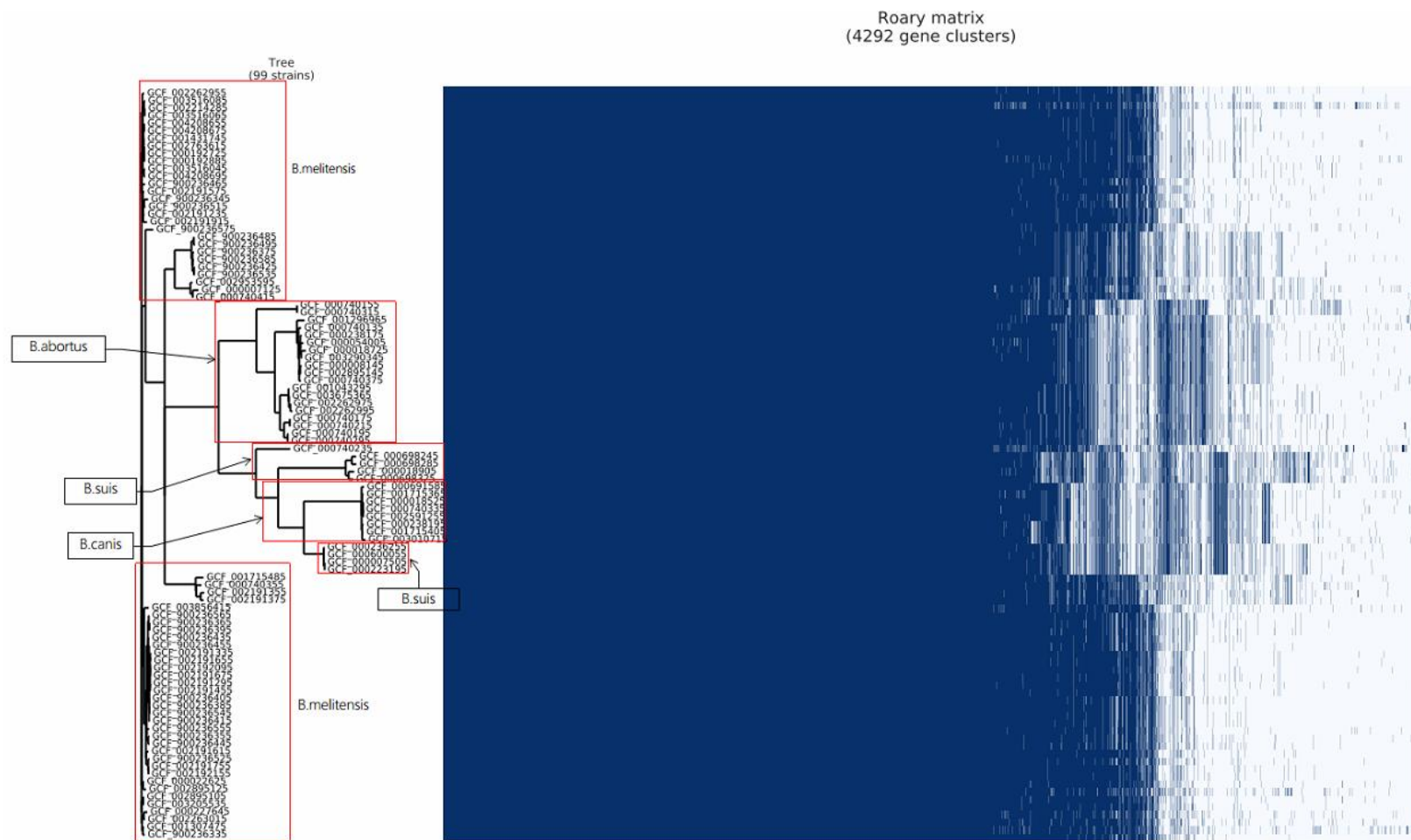


Figure 1. Phylogenetic tree of pan-SNP based on Brucella genus

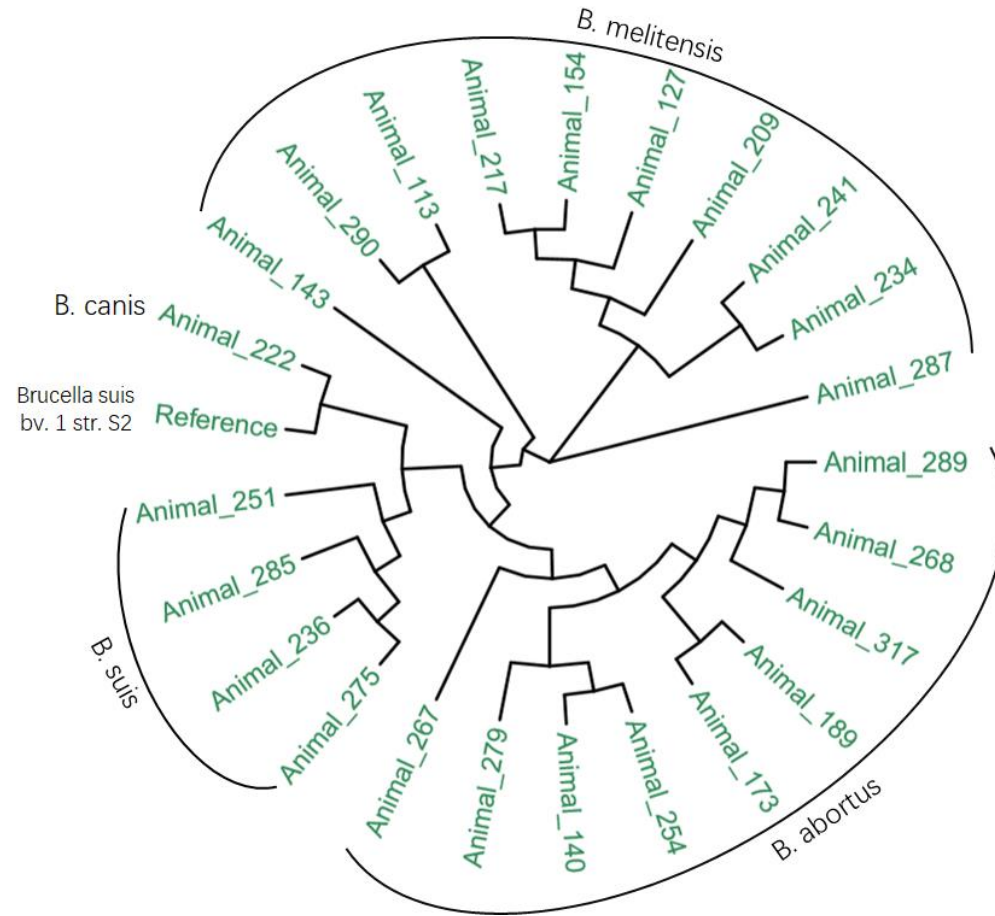


Figure 2. Phylogenetic tree of core SNP based on Brucella genus

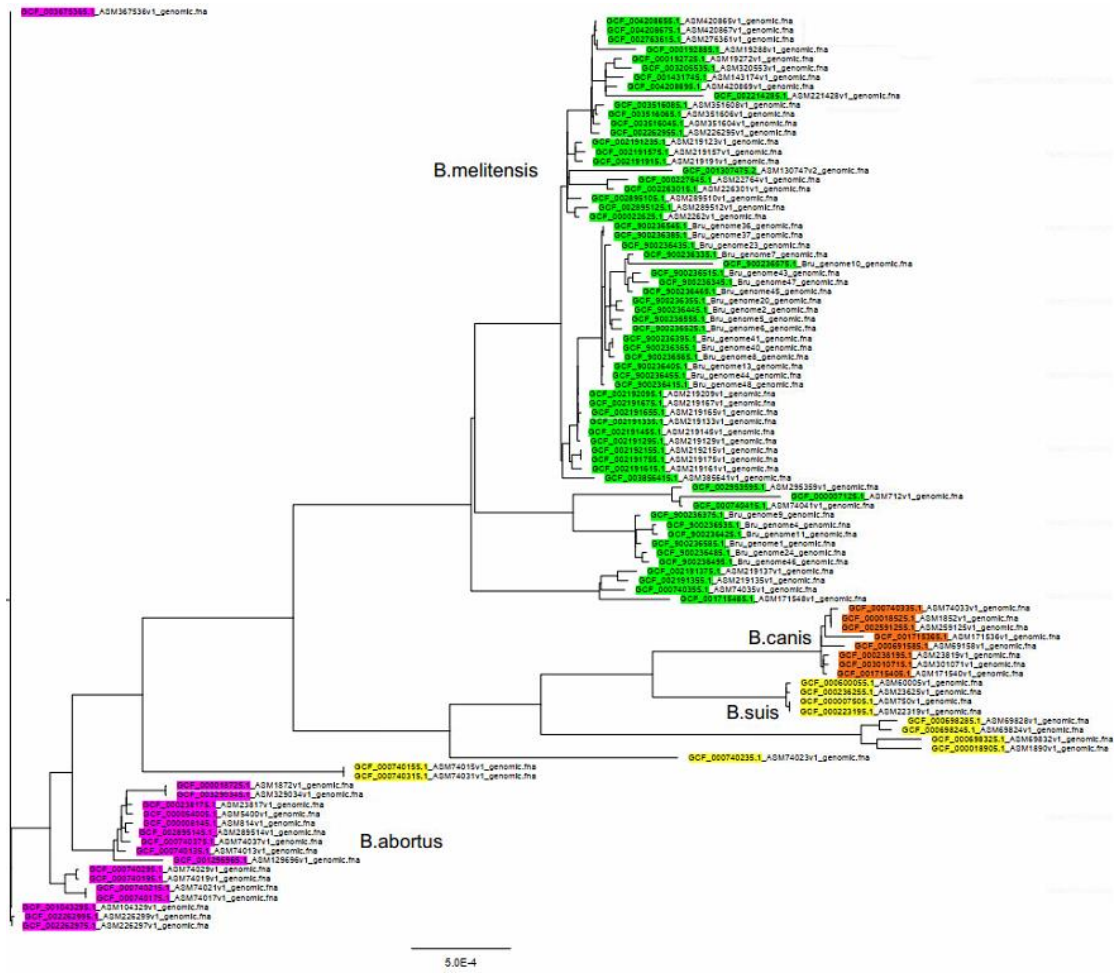
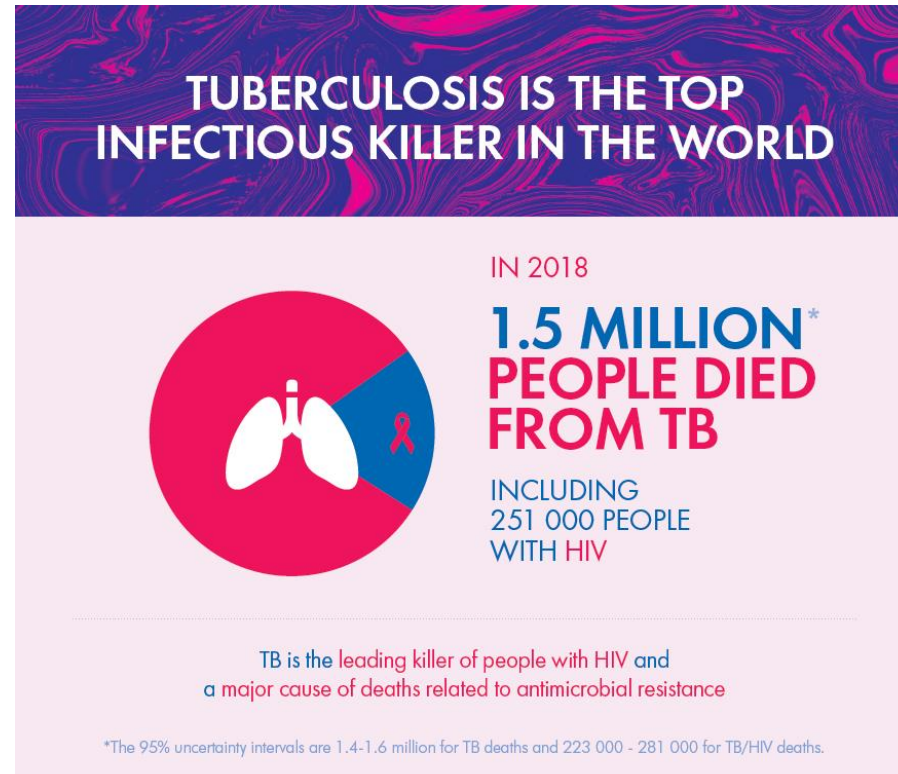


Figure3. Phylogenetic tree of gene family based on Brucella genus

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)

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_00275 5.2	31442 4	31496 9	-	AAC(2')- lc	1-546/546	=====	0/0	100	100	card	AL123456.3:314309-314855	AAC(2')- lc	aminoglyco side
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	76185 5	76527 4	+	rpoB2	64- 3462/3489	=====	19/4 1	97.13	79.56	card	AP006618.1:4835200-4838689	rpoB2	rifamycin
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	22290 13	22295 52	+	Erm(37)	1-540/540	=====	0/0	100	100	card	AL123456:2231680-2232220	Erm(37)	streptogra min/ macrolide/l incosamide
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	23101 54	23104 67	+	RbpA	1-314/345	=====	0/0	91.01	87.58	card	HQ203032:1-346	RbpA	rifamycin
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	31472 51	31488 43	-	efpA	1- 1593/1593	=====	0/0	100	100	card	AL123456.3:3153039-3154632	efpA	phenicol
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	36215 63	36222 49	-	mtrA	1-687/687	=====	0/0	100	100	card	AL123456.3:3626663-3627350	mtrA	antibacteri al_free_ fatty_acids
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	37354 28	37356 18	+	Erm(38)	777- 967/1161=====	0/0	16.45	76.44	card	AY154657.2:63-1224	Erm(38)	streptogra min/ macrolide/l incosamide
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	37652 40	37657 91	-	mfpA	1-552/552	=====	0/0	100	100	card	AL123456:3773016-3773568	mfpA	fluoroquino lone

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:

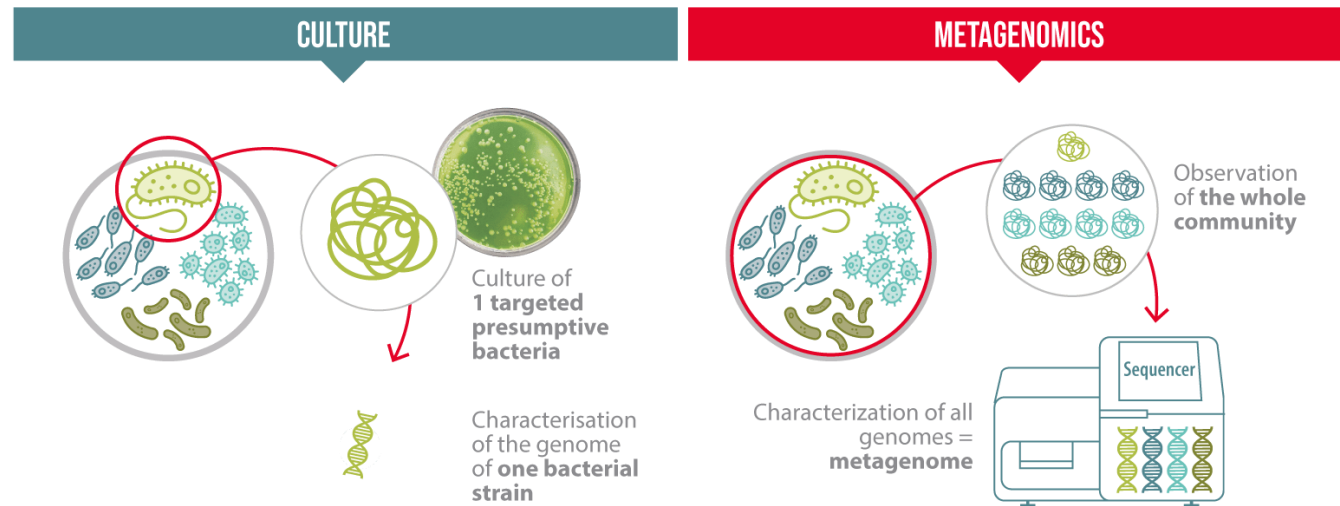
- Whole genome
- Virulence gene

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

#FILE	GCA_00006745.1_ 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
lpaA	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



Analysis Tools

1. Basic statistics of sequencing data (SeqStat)

Basic statistics of sequencing data (SeqStat)
Basic statistics of sequencing data (Seq_Stat) workflow can perform statistics and calculations on a series of reads number, base number, sequencing coverage, sequencing depth, and other basic information based on sequencing raw data.

Input file:
Query: Raw sequencing data, supports fastq, fq format.
File size can not exceed 100MB Browse

Add files

Reference: Reference genome sequence, supports fasta, fna, fa format.
File size can not exceed 100MB Browse

RUN

Output File :
Fq_Stat: Statistics of reads number and base number based on raw data, txt format.
Coverage: Sequencing coverage calculation, txt format.
Depth: Sequencing depth calculation, gz format.

Task: Seq_Stat_2020_10_09_11_6_56

File Name	File Status	Action
Depth_Stat.csv	available	下载
Coverage_Stat.csv	available	下载
Fq_Stat.csv	available	下载

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
------------------	---------------------

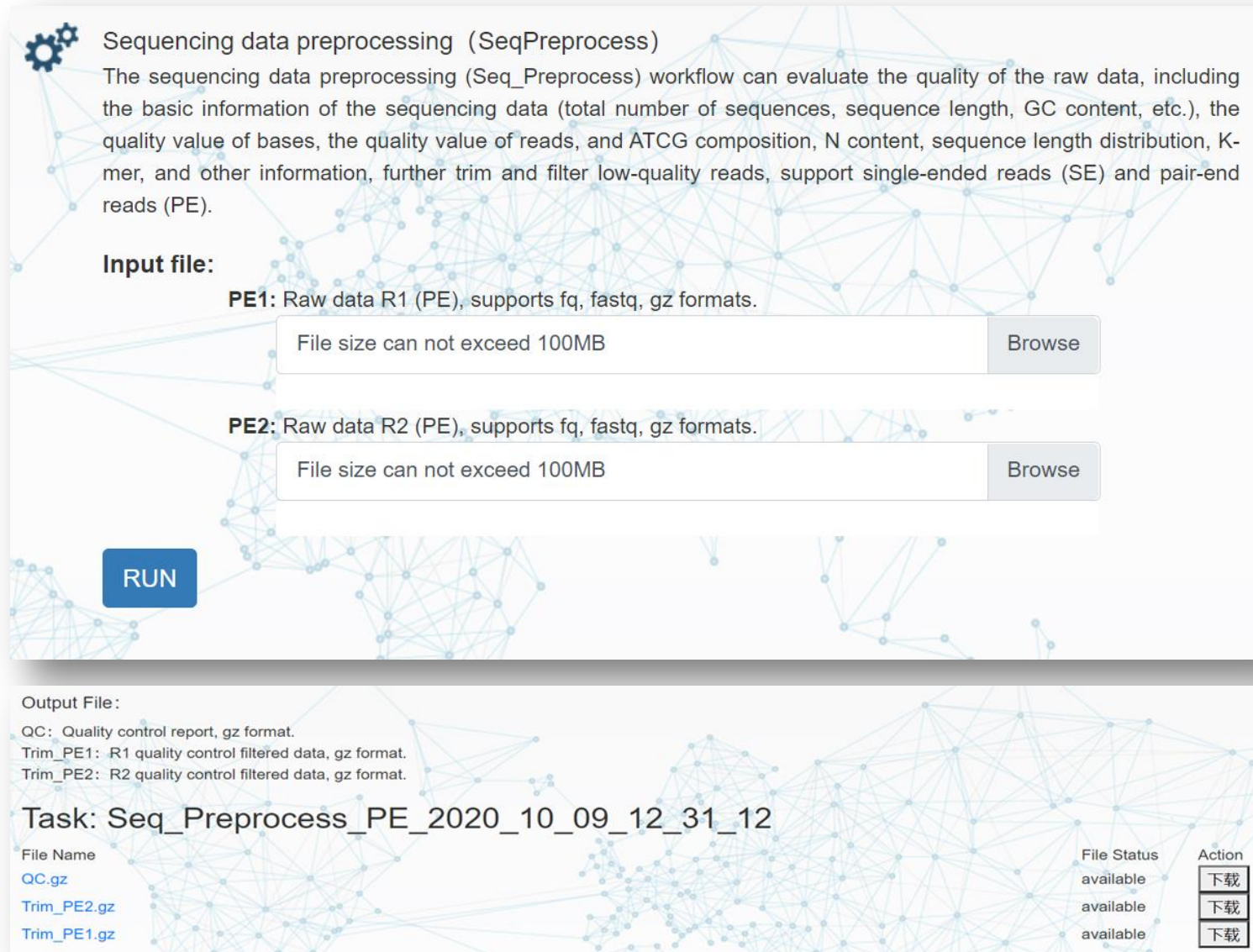
Depth_Stat.csv

#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)



Sequencing data preprocessing (SeqPreprocess)

The sequencing data preprocessing (Seq_Preprocess) workflow can evaluate the quality of the raw data, including the basic information of the sequencing data (total number of sequences, sequence length, GC content, etc.), the quality value of bases, the quality value of reads, and ATCG composition, N content, sequence length distribution, K-mer, and other information, further trim and filter low-quality reads, support single-ended reads (SE) and pair-end reads (PE).

Input file:

PE1: Raw data R1 (PE), supports fq, fastq, gz formats.

File size can not exceed 100MB

PE2: Raw data R2 (PE), supports fq, fastq, gz formats.

File size can not exceed 100MB





Output File:

QC: Quality control report, gz format.
Trim_PE1: R1 quality control filtered data, gz format.
Trim_PE2: R2 quality control filtered data, gz format.

Task: Seq_Preprocess_PE_2020_10_09_12_31_12

File Name	File Status	Action
QC.gz	available	<input type="button" value="下载"/>
Trim_PE2.gz	available	<input type="button" value="下载"/>
Trim_PE1.gz	available	<input type="button" value="下载"/>

Example:

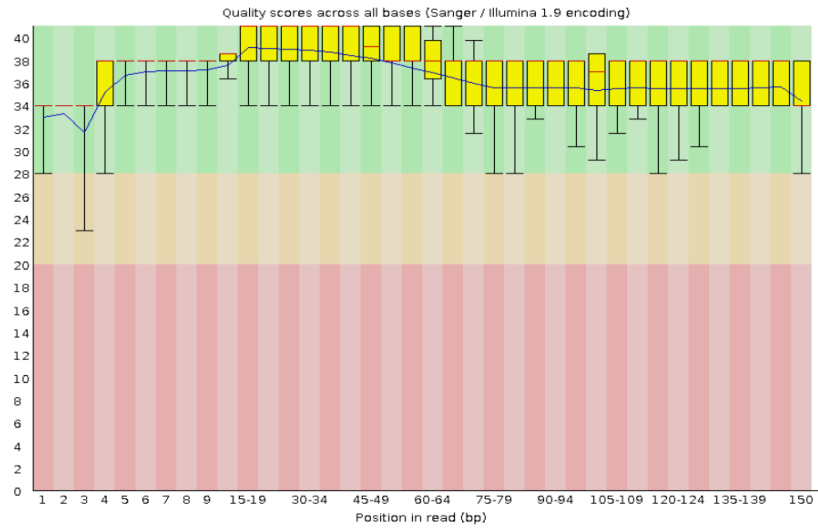
 PE1_fastqc.html →
 PE1_fastqc.zip
 PE2_fastqc.html
 PE2_fastqc.zip

FastQC Report

Summary

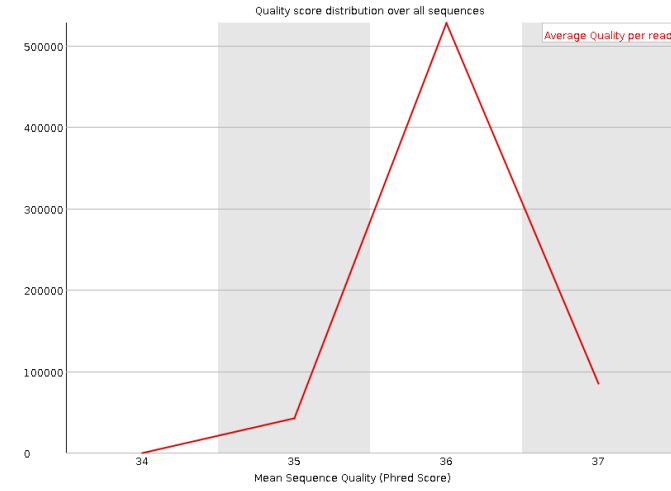
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter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

✔ Per base sequence quality



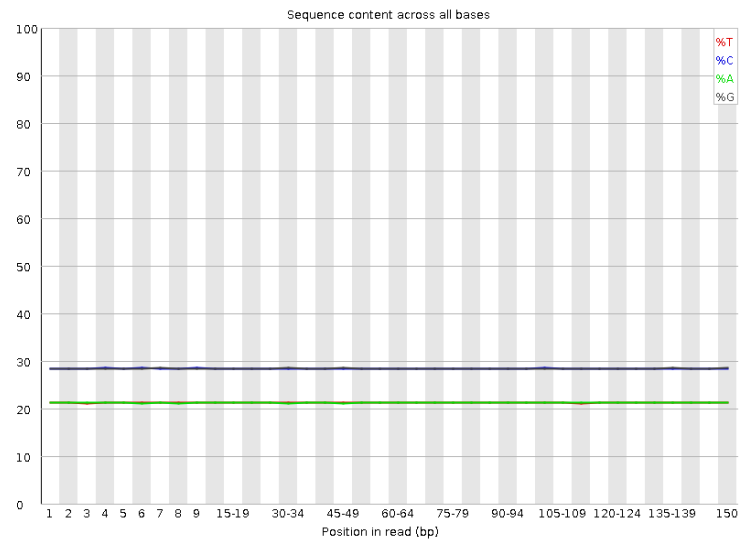
A

✔ Per sequence quality scores



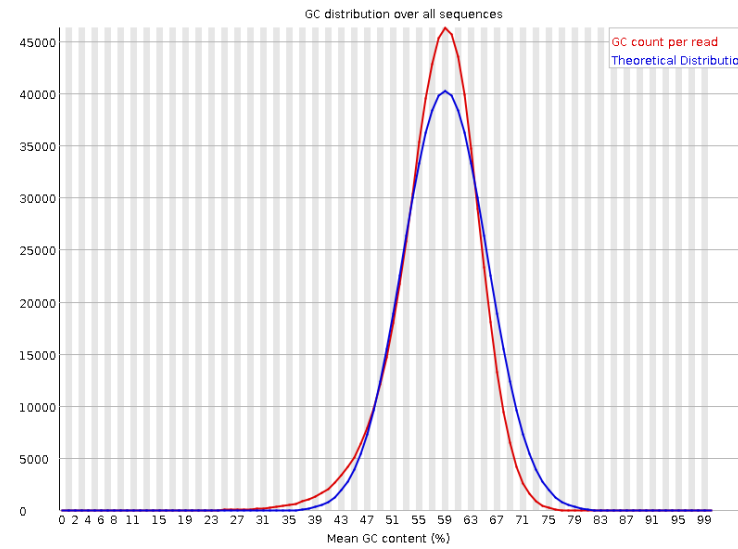
B

✔ Per base sequence content



C

! Per sequence GC content



D

3. Genome assembly and splicing (AssemblyUnicycler)

Genome assembly and splicing (AssemblyUnicycler)
Genome assembly and splicing (Assembly_Unicycler) is a tool for bacterial genome assembly. Not only can you assemble Illumina data (SPAdes), but you can also assemble PacBio or Nanopore long sequence sets (miniasm + Racon). You can also mix and assemble Illumina and third-generation sequencing data at the same time.

Input file:

longreads: nanopore or PacBio sequencing file (optional), fastq or fasta formats.
File size can not exceed 100MB

forward: Illumina forward R1 sequence file, fastq formats.
File size can not exceed 100MB

unpaired: Illumina unpaired sequencing sequence (optional), fastq formats.
File size can not exceed 100MB

reverse: Illumina forward R2 sequence file, fastq formats.
File size can not exceed 100MB

Parameters:

threads: the number of threads, the larger the value, the faster the analysis speed, the value is 1-16, the default: 2.









mode: splicing assembly mode, including conservative (very low splicing error rate, can not produce complete assembly), normal, bold (easy to produce complete assembly, high error rate), default: normal.

Output File:
assemble: tar.gz result compression package..

Task: Assembly_Unicycler_2020_10_09_14_5_15

File Name	File Status	Action
result.tar.gz	available	<input type="button" value="下载"/>

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 x
0 10 20 30 40 50 60 70
1 >1 length=358774 depth=1.00x
2 AGCGGTTCCCTGTTTTAACAGAATCGCCGGAACCGCTCTAACTATTTGTTTTGTCGCATTATCCAACGCAA
3 AACCGTTTCACACTTTTGCTGGAAATGCTCTATGCCGTTTTGGCCCGATGCACAAGCAGGAGTTCGCGCGC
4 TCATGACCTGAAAGGGGCTCAAACAAAAACGCCCTCTCCGGATGGAAAGAGCGTTTAAACTATACCCGGC
5 TTGCGCCCGATCAGAGATCGACGTCGAGAATGGCCATCGAGAAATTATACGACAATTCGCCTTCATCCTC
6 ATCCTTGTAGATAAGGCCGAGAAATTCGTCGCGCAGGTAAAGCTCGGCGGAATCATTCTTGCGCGGACGC
7 GCCTTCACCTGAAGATCAGGATTGTTGAAGGTCCGTTTGAAATAGGCGTCCAGTTTTTTGAGTTCCTCGG
8 GTTCAACCGGATTCTCCTGACATTGATTTGAAGCCGTTTTATTGCACCCGGCATTTCCTTATGTAAACC
9 CCTGACATCGCCTATCGCAGGCTGTAGCGACGAATTAACCTTCGGTTTGGCATGGGCAAAGTGCGAAGC
10 GAGGTCAGAAATCAGTCCATTGAACTGATTTCCCCGCATGGACGTTTCACATTTTGGGCCCGAAGACGCT
11 CCAAACAAAGCTTTCACCACATATAGGCCGTCTGCGGGATAAAACCAGCAGACAACCCAACCTTATGGCA
12 TGAGCCGTTATCAATATTGGAATTAGATGCCGTCAACCAGCATCTGGTCCATCACGCGGGATGGCTCCGT
13 GCAGCCCGTTTCGCCGATGATGCGCGCAGGCACGCCCGCCACGGTTACATTATGCGGCACGGACTTCAGC
14 ACAACGGAACCAGCCGCGATCTTGGAGCACTGCCCTACCTGGATGTTGCCGAGGATTTTTGCGCCTGCC
15 CAATCAGCACGCCCTGACGGATTTTCGGATGGCGATCACCGCTCGATTTGCCGGTTCGCCCCAGCGTAAC
16 CCCATGCAGGATCGAGACATTATCTTCTACAACGGCCGTTTCACCCACCACAAGCCCTGTAGCATGGTCG
17 AGAAACAGCCCGCTGCCAAGCCTGGCCGCCGGTGAATATCGGTCTGAAAGATCGATGACGAGCGGCTCT
18 GAAGATAATAGGCGAAATCCTTGCACCTTGCTTGTAAAGCCAGTGCGCCAGACGGTGTGTCTGGATCGC
19 ATGGAAGCCCTCAGGTAAGAACCAGGATCCATAAAGCGGCTATAGGCCGGATCGCGGTCATAAACAGCC
20 TGAATATCCACACGCAGGACGTGCGACCATTCCGGTTTTGCCTCAAGCATCGTATCGAAGGTCTGGCGCA
21 GGATATCCGCGCTCACATCGGGATGGCCAAGGCCTTCGGCGATACGATGCATCACCGCTTCTTCAAGGCT
```


4. Phylogenetic analysis of core genome SNPs (CoreSNP_Phylo)

Phylogenetic analysis of core genome SNPs (CoreSNP_Phylo)
The core genome SNPs phylogenetic analysis (CoreSNP_Phylo) workflow is used to find SNP sites between the sequencing data and the reference genome. As a result, a phylogenetic tree will be built based on the core genome SNP sites.

Input file:
PE: Illumina forward R1 sequence file and reverse R2 sequence file, supports fq, fastq, gz formats.
File size can not exceed 100MB

Reference: Reference genome, supports fna, fasta, fa formats.
File size can not exceed 100MB

Output File:
CoreSNP_Phylo: tgz compressed package, containing phylogenetic tree in newick and pdf.

Task: CoreSNP_Phylo_2020_10_09_16_14_35

File Name	File Status	Action
treefile.tgz	available	<input type="button" value="下载"/>

Example:

treefile.tgz
↓
core.newick
Rplots.pdf

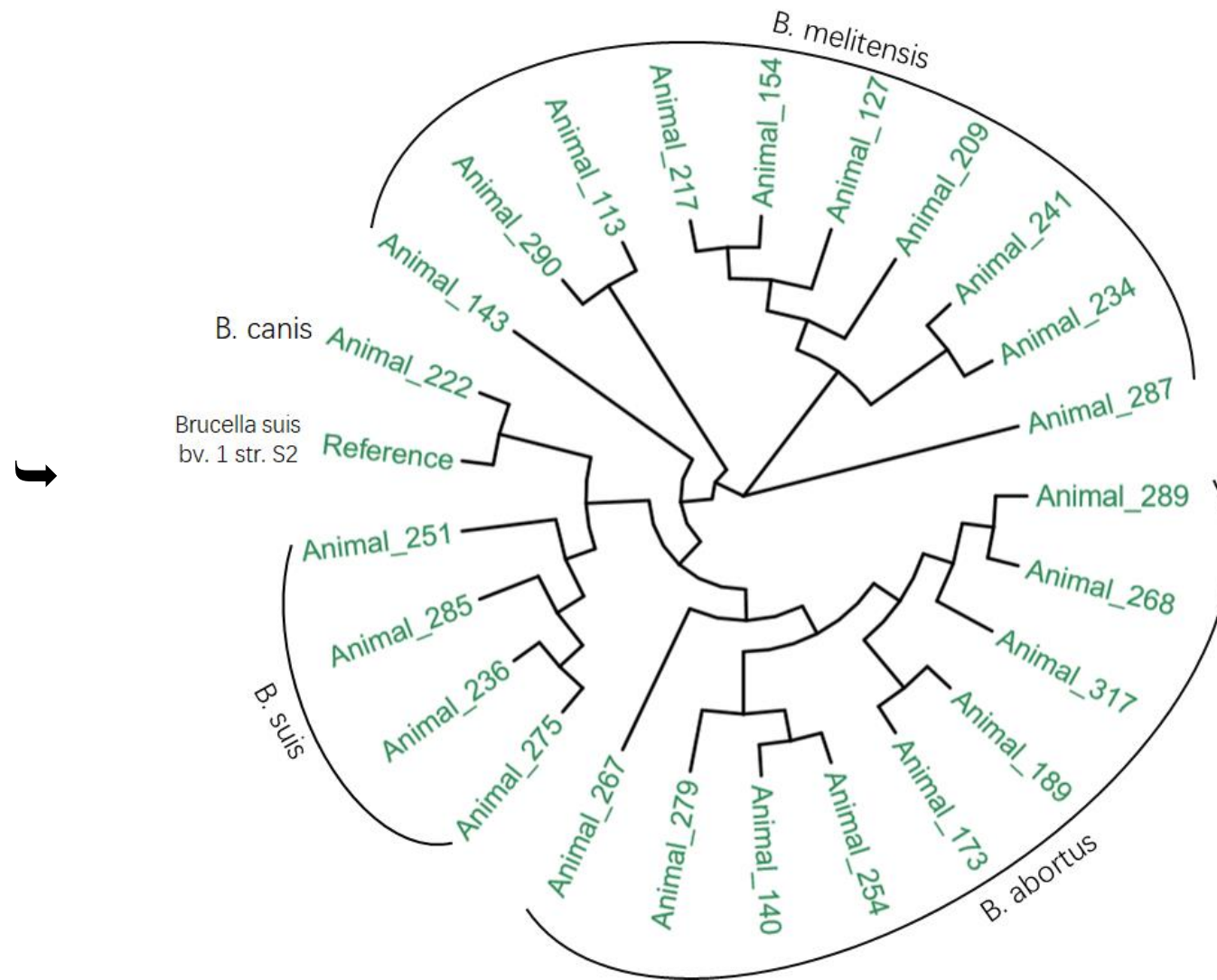
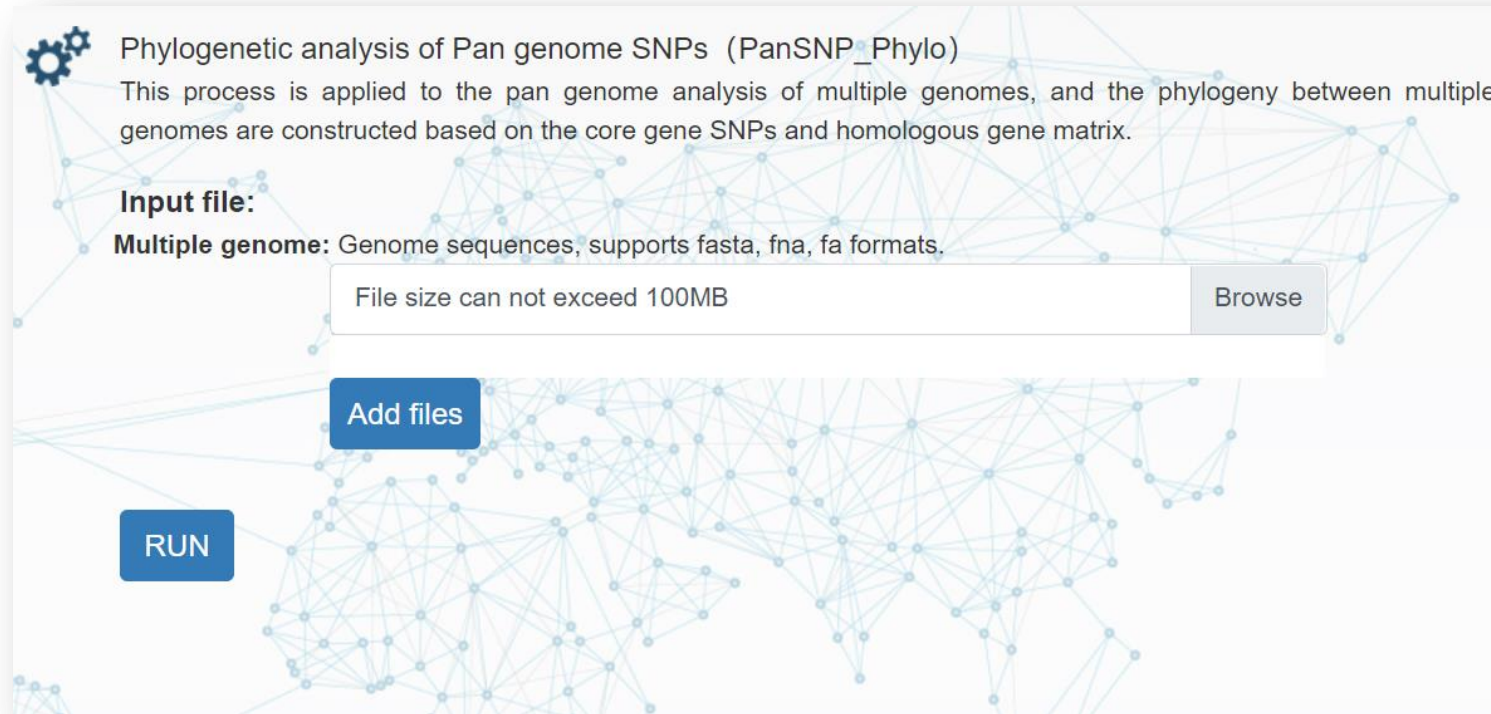



Figure 1. Core SNP-based phylogenetic tree of 24 Brucella strains

5. Phylogenetic analysis of Pan genome SNPs (PanSNP_Phylo)



 Phylogenetic analysis of Pan genome SNPs (PanSNP_Phylo)
This process is applied to the pan genome analysis of multiple genomes, and the phylogeny between multiple genomes are constructed based on the core gene SNPs and homologous gene matrix.

Input file:
Multiple genome: Genome sequences, supports fasta, fna, fa formats.

File size can not exceed 100MB

Output File:

Pangenome_matrix: tgz compressed package, which provides 3 figures, showing the tree compared to a matrix with the presence and absence of core and accessory genes. The next is an pie chart of the breakdown of genes and the number of isolate they are present in. And finally there is a graph with the frequency of genes versus the number of genomes.

Task: PanSNP_Phylo_2020_10_09_16_56_52

File Name	File Status	Action
Pangenome_matrix.tgz	available	<input type="button" value="下载"/>

Example:

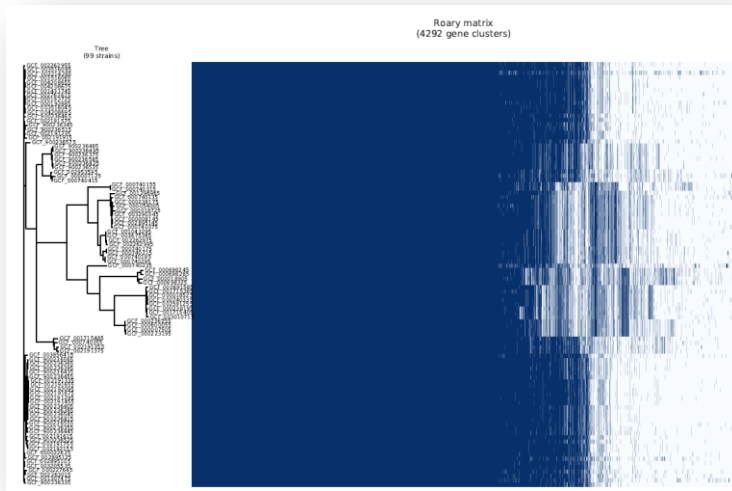


Fig 1. pangenome_matrix

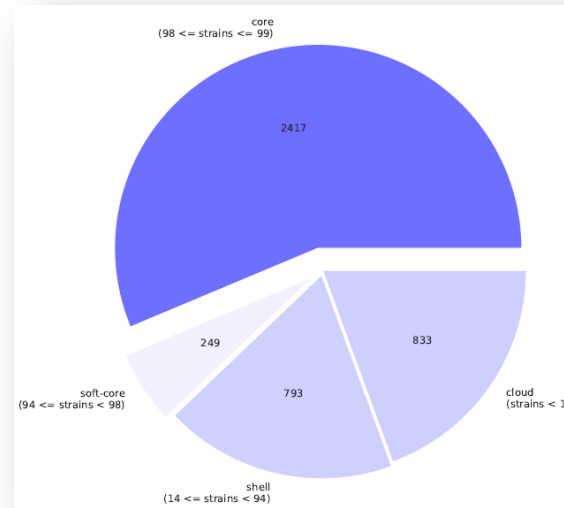


Fig 2. pangenome_pie

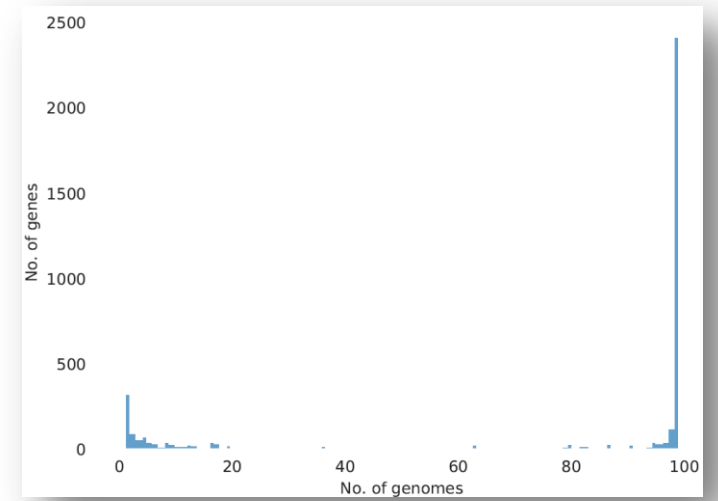
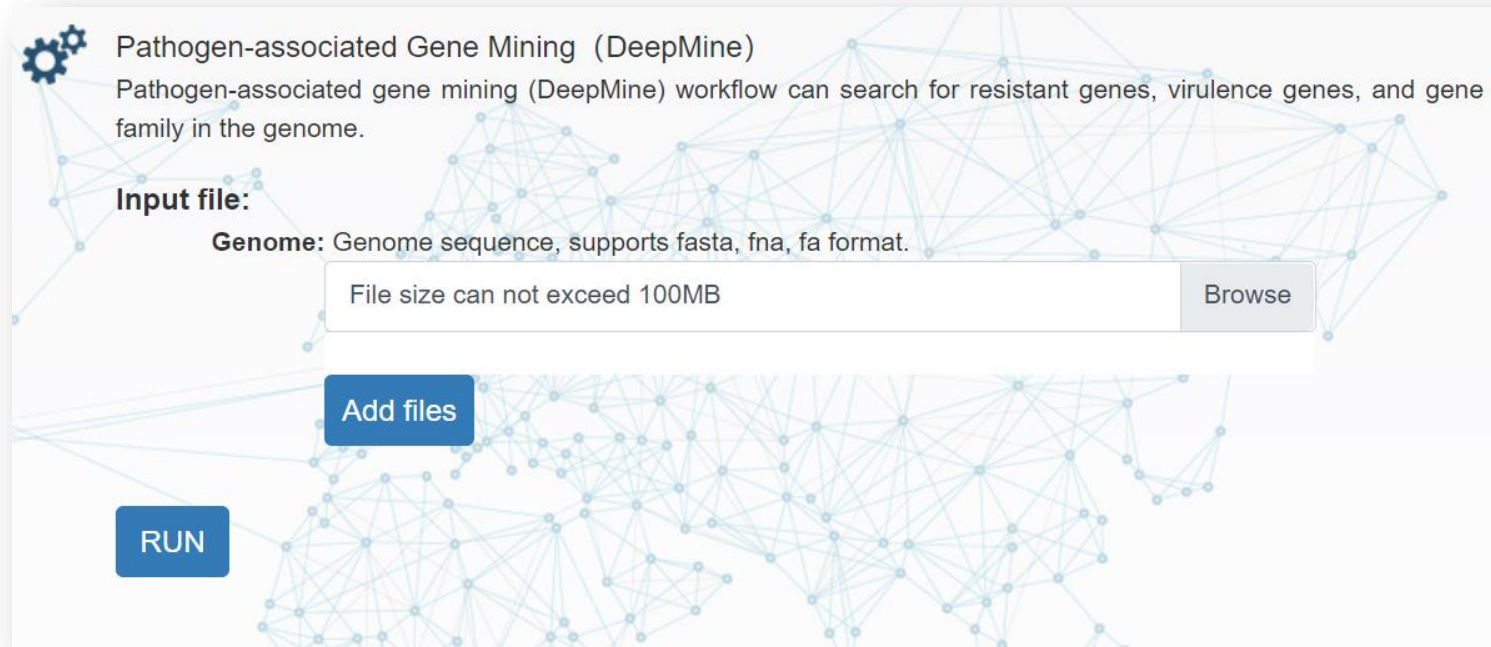


Fig 3. pangenome_frequency

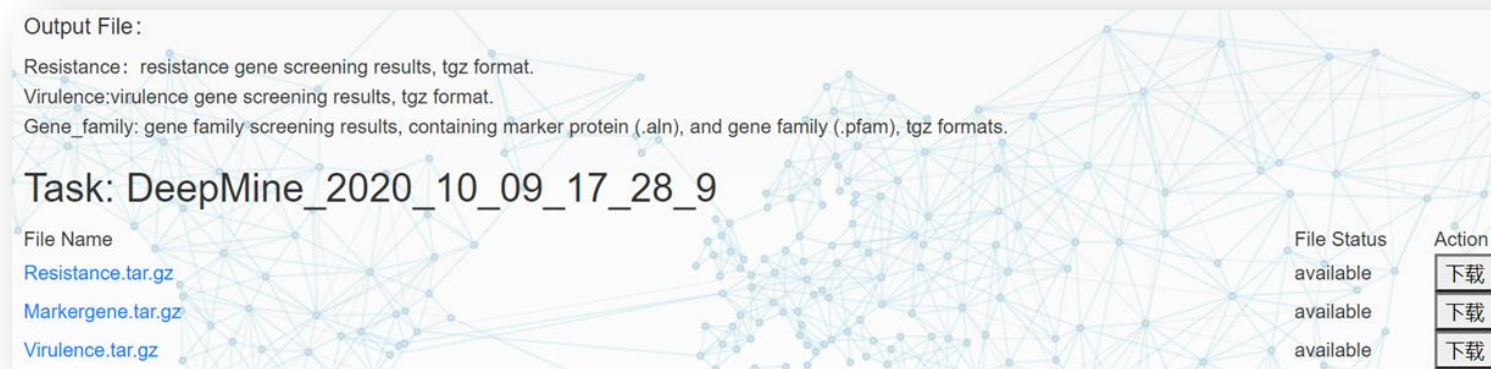
6. Pathogen-associated Gene Mining (DeepMine)



Pathogen-associated Gene Mining (DeepMine)
Pathogen-associated gene mining (DeepMine) workflow can search for resistant genes, virulence genes, and gene family in the genome.

Input file:
Genome: Genome sequence, supports fasta, fna, fa format.

File size can not exceed 100MB



Output File:
Resistance: resistance gene screening results, tgz format.
Virulence: virulence gene screening results, tgz format.
Gene_family: gene family screening results, containing marker protein (.aln), and gene family (.pfam), tgz formats.

Task: DeepMine_2020_10_09_17_28_9

File Name	File Status	Action
Resistance.tar.gz	available	<input type="button" value="下载"/>
Markergene.tar.gz	available	<input type="button" value="下载"/>
Virulence.tar.gz	available	<input type="button" value="下载"/>

Example:








-  GCF_000008585.1_ASM858v1_genomic.fna.tab
-  GCF_000009445.1_ASM944v1_genomic.fna.tab
-  GCF_000010685.1_ASM1068v1_genomic.fna.tab
-  GCF_000016145.1_ASM1614v1_genomic.fna.tab
-  GCF_000153685.2_ASM15368v2_genomic.fna.tab
-  GCF_000193185.2_ASM19318v2_genomic.fna.tab
-  resistance.summary.csv

Fig 1. Resistance







-  GCF_000008585.1_ASM858v1_genomic.fna.tab
-  GCF_000009445.1_ASM944v1_genomic.fna.tab
-  GCF_000010685.1_ASM1068v1_genomic.fna.tab
-  GCF_000016145.1_ASM1614v1_genomic.fna.tab
-  GCF_000153685.2_ASM15368v2_genomic.fna.tab
-  GCF_000193185.2_ASM19318v2_genomic.fna.tab

Fig 2. Virulence



-  genome.out.aln
-  genome.out.pfam

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_00275 5.2	31442 4	31496 9	-	AAC(2')-lc	1-546/546	=====	0/0	100	100	card	AL123456.3:314309-314855	AAC(2')-lc	aminoglycoside
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	76185 5	76527 4	+	rpoB2	64- 3462/3489	=====	19/4 1	97.13	79.56	card	AP006618.1:4835200-4838689	rpoB2	rifamycin
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	22290 13	22295 52	+	Erm(37)	1-540/540	=====	0/0	100	100	card	AL123456:2231680-2232220	Erm(37)	streptomycin/ macrolide/ lincosamide
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	23101 54	23104 67	+	RbpA	1-314/345	=====	0/0	91.01	87.58	card	HQ203032:1-346	RbpA	rifamycin
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	31472 51	31488 43	-	efpA	1- 1593/1593	=====	0/0	100	100	card	AL123456.3:3153039-3154632	efpA	phenicol
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	36215 63	36222 49	-	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_00275 5.2	37354 28	37356 18	+	Erm(38)	777- 967/1161=====	0/0	16.45	76.44	card	AY154657.2:63-1224	Erm(38)	streptomycin/ macrolide/ lincosamide
GCF_000008585.1_ASM858v1_genomic.fna	NC_00275 5.2	37652 40	37657 91	-	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
lpa	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	OTCase_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, Cytidyltransferase
PF04392.11	ABC_sub_bind, ABC
PF01121.19	CoaE, Dephospho-CoA
PF04279.14	IspA, Intracellular
PF03461.14	TRCF, TRCF
PF10984.7	DUF2794, Protein

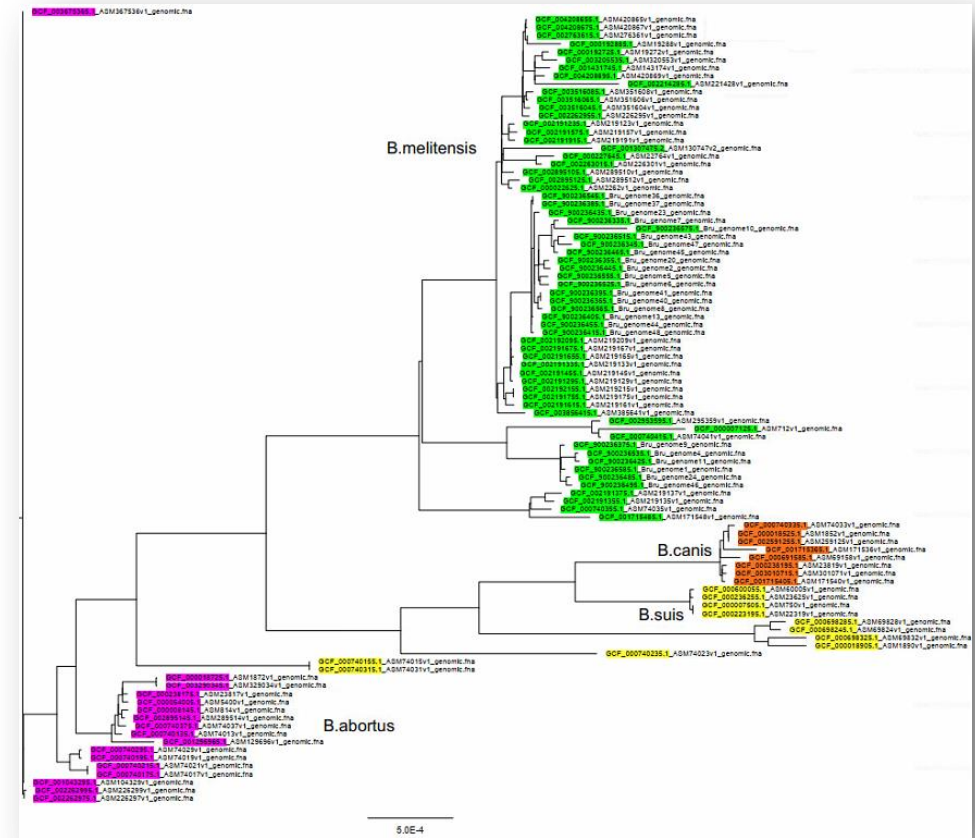



Figure3. Phylogenetic tree based on Brucella "genus" level gene-related protein sequence

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

 Pathogen screening based on 16s amplicon sequencing(Pathogen_16sPIP)

Function introduction: Pathogen_16sPIP is a comprehensive analysis process with multiple integrated components, including data format conversion, quality control, sequence screening, fast alignment, report generation and other processes, and provides two modes of fast and sensitive, which can be used for single-ended or dual 16S metagenomic sequencing data for pathogenic bacteria screening; in emergencies, fast mode can be preferentially selected, and 346 kinds of human health-related pathogens can be quickly screened in combination with clinical symptoms. If the user wants to identify the presence of other species and study the population diversity of the microbial community, the sensitive mode is a better choice

Input file:

forward: Forward sequence (R1) of pair-end Illumina sequencing data in fastq formats, or Single-end data in fasta formats, and 454 sequencing data in sff, sam, bam formats.

File size can not exceed 100MB

reverse: reverse sequence (R2) of pair-end Illumina sequencing data in fastq format. (Optional)

File size can not exceed 100MB

Parameters:

mode:Use analysis mode, the value is fast or sensitive, default fast.

thread:Number of threads used for analysis, value 1~16.

format:To analyze the NGS data formats, fastq, fasta, sam, bam or sff.

Output File:

Task: Pathogen_16sPIP_2020_10_09_18_7_13

File Name	File Status	Action
reulst.tar.gz	available	<input type="button" value="下载"/>
pdf.pdf	available	<input type="button" value="下载"/>

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/>

Pathogen in-depth analysis cloud platform | Introduction | Platform | Workflow | Example | Tools | 中文 | English

- Basic statistics of sequencing data
- Sequencing data preprocessing
- Genome assembly and splicing
- Phylogenetic analysis of core genome SNPs
- Phylogenetic analysis of Pan genome SNPs
- Pathogen-associated Gene Mining
- Pathogen screening based on 16s amplicon sequencing**

Pathogen screening based on 16s amplicon sequencing(Pathogen_16sPIP)

Function introduction: Pathogen_16sPIP is a comprehensive analysis process with multiple integrated components, including data format conversion, quality control, sequence screening, fast alignment, report generation and other processes, and provides two modes of fast and sensitive, which can be used for single-ended or dual 16S metagenomic sequencing data for pathogenic bacteria screening; in emergencies, fast mode can be preferentially selected, and 346 kinds of human health-related pathogens can be quickly screened in combination with clinical symptoms. If the user wants to identify the presence of other species and study the population diversity of the microbial community, the sensitive mode is a better choice

Input file:

forward: Forward sequence (R1) of pair-end Illumina sequencing data in fastq, fastq.gz, fasta formats, and 454 sequencing data in sff, sff.gz formats. File size can not exceed 100MB

reverse: reverse sequence (R2) of pair-end Illumina sequencing data in fastq, fastq.gz, fasta formats, and 454 sequencing data in sff, sff.gz formats. File size can not exceed 100MB

Parameters:

mode: Use analysis mode, the value is fast or sensitive.
fast

thread: Number of threads used for analysis, value 1-16
8

format: To analyze the NGS data formats, fastq, fasta, sff, sff.gz, and 454 sequencing data in sff, sff.gz formats.
fastq

RUN

Pathogen in-depth analysis cloud platform | Introduction | Platform | Workflow | Example | Tools | 中文 | English

布鲁氏菌

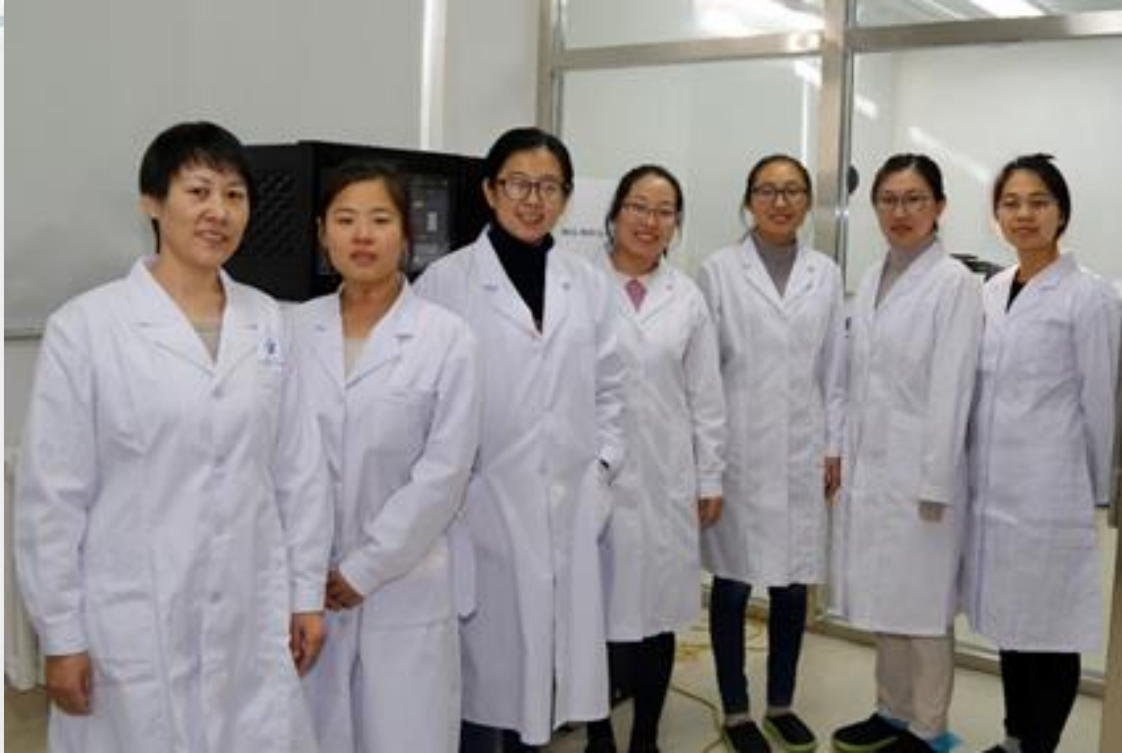
霍乱弧菌

结核分枝杆菌

临床样本

The visualization shows a stylized brain composed of colored segments (green, orange, red, grey) with circular callouts pointing to specific pathogens: 布鲁氏菌 (Brucella), 霍乱弧菌 (Vibrio cholerae), and 结核分枝杆菌 (Mycobacterium tuberculosis). A laptop with a magnifying glass icon is positioned below the brain, and the text 临床样本 (Clinical sample) is on the right. The background features a network graph pattern.

Our team



Dr. Zhang Wen

Dr. Hanna

Zhang Tingting

Qiang Yujun

Li Xiuwen



**THANK YOU FOR
WATCHING !**

pengxianhui@icdc.cn