# 病原微生物高通量测序数据分析本地云平台实操

广东省疾病预防控制中心
张昌
2025年6月27日

# 实操内容

- 基因组组装
  - 从头组装（De Novo assembly)
  - 有参组装（reference-based assembly)
  - 二代+三代混合组装（略）
- 基因特征分析
  - 基因组注释、耐药/毒力基因分析、MLST分型
- 溯源分析
  - 病毒全基因组进化树构建
  - 细菌核心基因组SNP（cgSNP）溯源分析
- 宏基因组测序病原体鉴定

- 连接WIFI，名称：training，密码：training。
- 打开浏览器，登录网址：http://192.168.31.*:8080
- 用户名：姓名全拼；密码：20250627

# 从头组装（De Novo assembly）


**参考代码：**

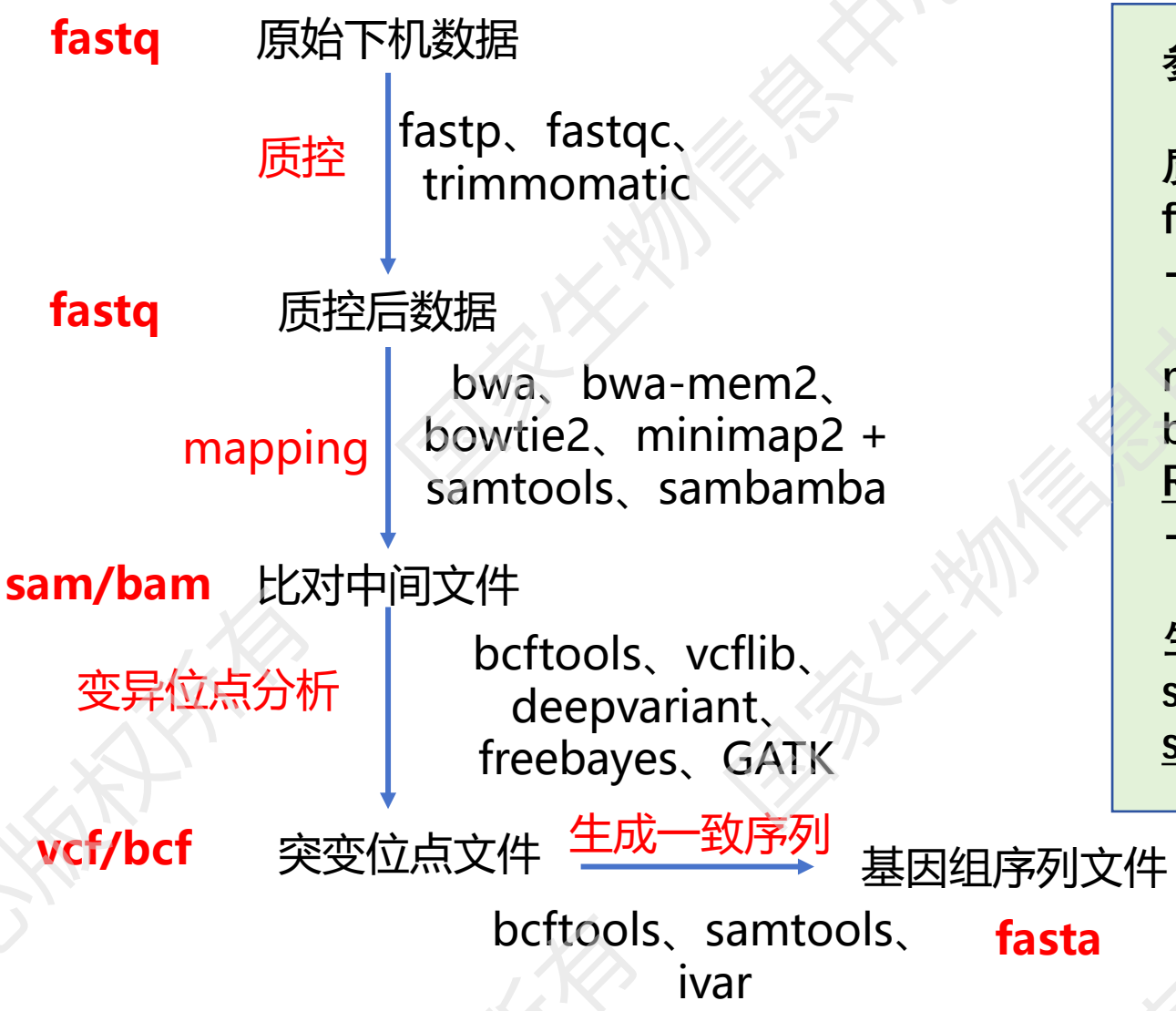
**二代：**

```
fastp -i R1.fastq.gz -I R2.fastq.gz -o R1_clean.fastq.gz
-O R2_clean.fastq.gz -h report.html
spades.py -1 R1_clean.fastq.gz -2 R2_clean.fastq.gz -o
result_out
```

**三代：**

```
flye --nano-raw barcode.fastq.gz --out-dir result_out
--threads 8
```

# 有参组装（reference-based assembly）




**参考代码：**

**质控：**

```
fastp -i R1.fastq.gz -I R2.fastq.gz -o R1_clean.fastq.gz
-O R2_clean.fastq.gz -h report.html
```

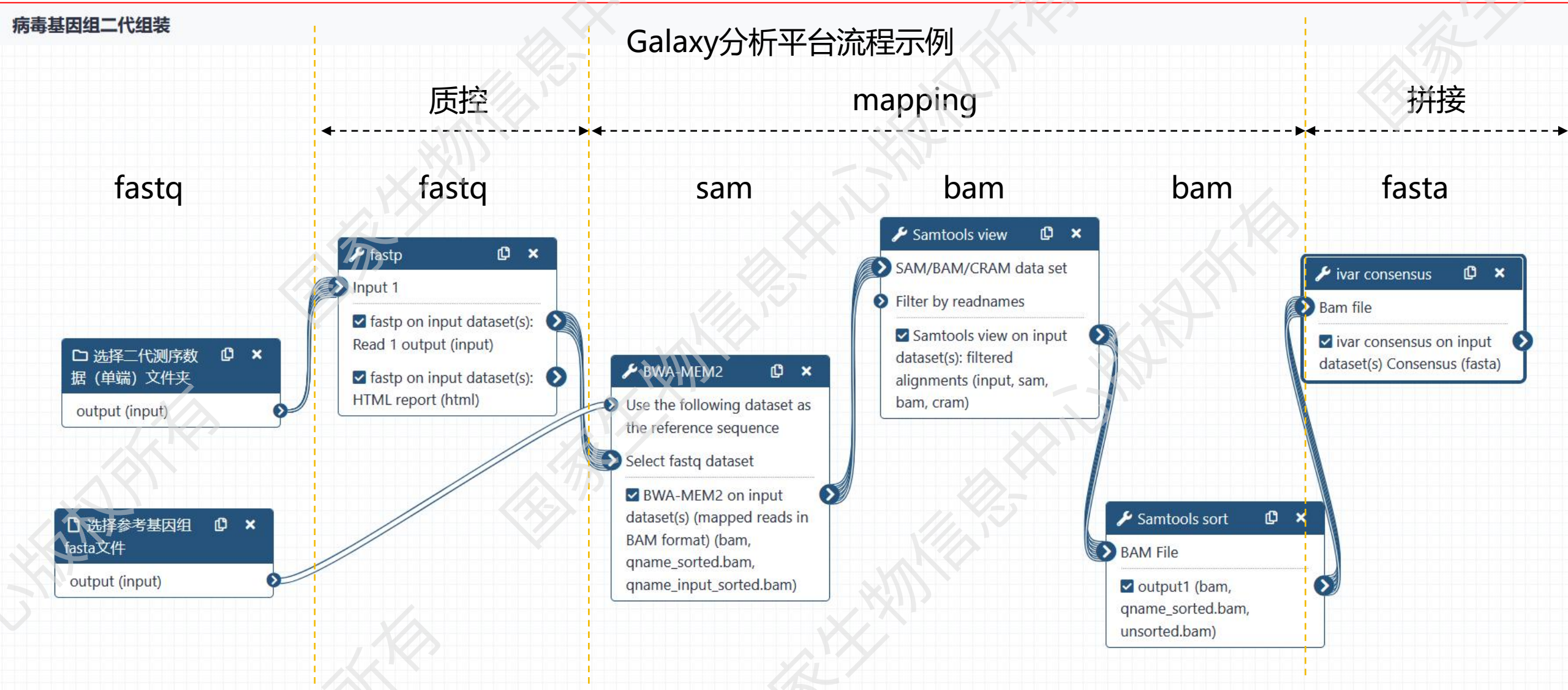
**mapping：**

```
bwa mem reference.fasta R1_clean.fastq.gz
R2_clean.fastq.gz | samtools sort | samtools view -F 4
-o sample.sorted.bam
```

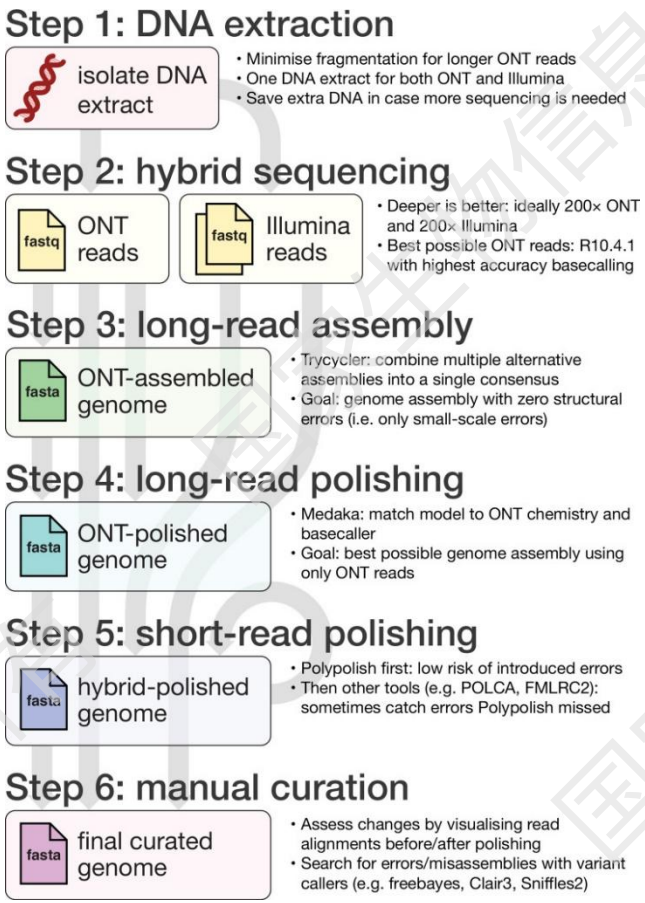
**生成一致序列：**

```
samtools consensus -m simple -d 10
sample.sorted.bam -c 0.5 > assembly.fasta
```

# 有参组装流程图


病毒基因组二代组装
Galaxy分析平台流程示例
质控
mapping
拼接
fastq
fastq
sam
bam
bam
fasta
选择二代测序数据（单端）文件夹
output (input)
选择参考基因组fasta文件
output (input)
fastp
Input 1
fastp on input dataset(s): Read 1 output (input)
fastp on input dataset(s): HTML report (html)
BWA-MEM2
Use the following dataset as the reference sequence
Select fastq dataset
BWA-MEM2 on input dataset(s) (mapped reads in BAM format) (bam, qname_sorted.bam, qname_input_sorted.bam)
Samtools view
SAM/BAM/CRAM data set
Filter by readnames
Samtools view on input dataset(s): filtered alignments (input, sam, bam, cram)
Samtools sort
BAM File
output1 (bam, qname_sorted.bam, unsorted.bam)
ivar consensus
Bam file
ivar consensus on input dataset(s) Consensus (fasta)

# 混合组装




**参考代码：**

1.三代序列组装

```
flye --nano-hq barcode.fastq.gz  -o ONT_assembly
```

2.三代组装结果校正

```
medaka_consensus -d ONT_assembly/assembly.fasta -i
barcode.fastq.gz -o ONT_assembly -m
r1041_e82_400bps_sup_g615
```

3.使用二代序列进一步校正

```
cd ONT_assembly
bwa index consensus.fasta
bwa mem -a consensus.fasta R1.fastq.gz > R1.sam
bwa mem -a consensus.fasta R2.fastq.gz > R2.sam
polypolish filter --in1 R1.sam --in2 R2.sam --out1
R1.filtered.sam --out2 R2.filtered.sam
polypolish polish consensus.fasta R1.filtered.sam
R2.filtered.sam > consensus.polished.fasta
```

# 基因特征分析

- 使用prokka进行编码基因注释（基于参考基因组）
  - prokka --proteins reference.gbk contigs.fasta
- 耐药性分析
  - arm数据库
    - amrfinder (-p <protein_fasta> | -n <nucleotide_fasta)
  - CARD数据库
    - rgi main -i contigs.fasta -t contig -o output.out --clean
- 血清型分析（部分细菌）
  - 肺炎克雷伯（Kleborate）、肺炎链球菌（PneumoKITy）、流感嗜血杆菌（hicap)、脑膜炎耐瑟氏球菌（meningotype）、酿脓链球菌（emm_typing)、产气荚膜梭菌（TOXIper）、肠沙门氏菌（sistr_cmd）、单核增生李斯特氏菌（LisSero）、副溶血弧菌（VPsero)、金黄色葡萄球菌（SaLTy)、志贺氏菌属（shigatyper)、致病性大肠埃希菌（ecoli_serotyping）、无乳链球菌（GBS_SBG)

# 基因特征分析

- 使用pubMLST数据库进行MLST分型
  - https://github.com/tseemann/mlst
  - **mlst genomes/***

| | | | |
|---|---|---|---|
| genomes/6008.fna | saureus | 239 | arcc(2) aroe(3) glpf(1) gmk_(1) pta_(4) tpi_(4) yqil(3) |
| genomes/strep.fasta.gz | ssuis | 1 | aroA(1) cpn60(1) dpr(1) gki(1) mutS(1) recA(1) thrA(1) |
| genomes/NC_002973.gbk | lmonocytogenes | 1 | abcZ(3) bglA(1) cat(1) dapE(1) dat(3) ldh(1) lhkA(3) |
| genomes/L550.gbk.bz2 | leptospira | 152 | glmU(26) pntA(30) sucA(28) tpiA(35) pfkB(39) mreA(29) caiB(29) |

# 病毒全基因组进化树构建


**参考代码：**

**多序列比对：**

```
mafft --auto seq.fasta > seq_aligned.fasta
```

**进化树构建：**

```
iqtree2 -s seq_aligned.fasta -T AUTO -B 1000
```

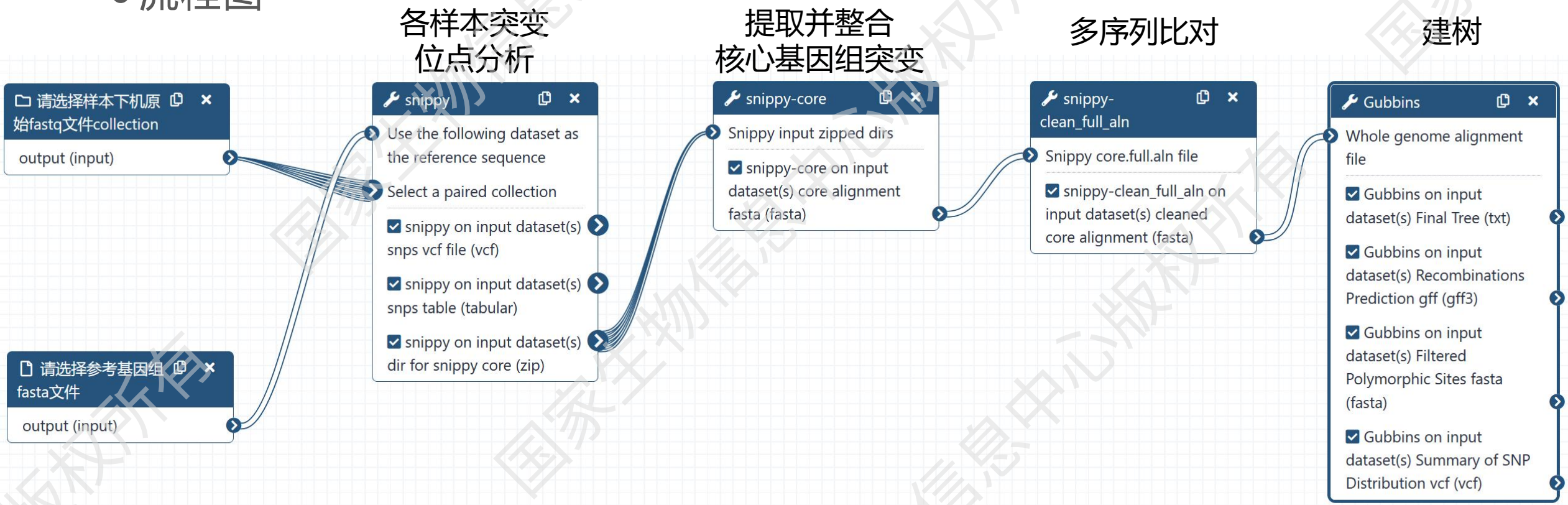
# 细菌核心基因组SNP（cgSNP）溯源分析

- 使用snippy进行cgSNP分析和建树
  - https://github.com/tseemann/snippy
  - 示例代码
  - 注意事项
    - 组装质量评估
    - 参考基因组

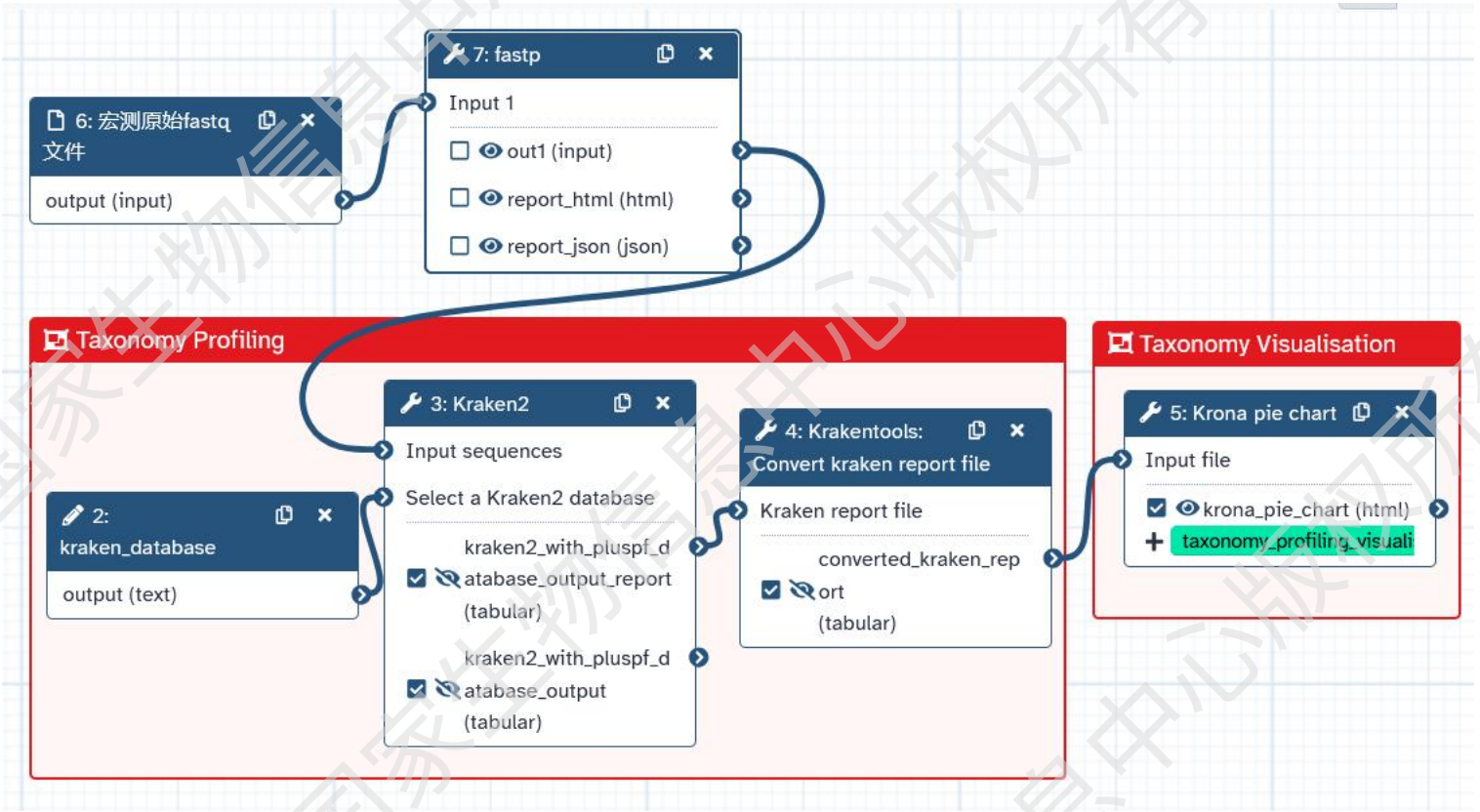
```
#!/bin/bash
echo "The PID of current script is: $$"
export PATH=/home/ngs_analysis/tool_packages/snippy/bin/:$PATH
ls Vibrio* > seqlist.txt
ls GD* >> seqlist.txt
IFS=$'\n'
for line in $(< ./seqlist.txt)
do
    outdir=`echo "$line" | tr -s ' ' "_" | cut -d "." -f 1`
    echo $outdir >> dirlist.txt
    snippy --outdir $outdir --ref ref.fasta --cpus 16 --ctgs "$line" &
done
wait
unset IFS
corepara=`cat dirlist.txt | tr -s ['\t','\n'] " "`
snippy-core --prefix core $corepara --ref ref.fasta
snippy-clean_full_aln core.full.aln > clean.full.aln
source /home/ngs_analysis/anaconda3/bin/activate gubbins_git
run_gubbins.py -p gubbins clean.full.aln
source /home/ngs_analysis/anaconda3/bin/deactivate
~/tool_packages/snippy/binaries/linux/snp-sites -c gubbins.filtered_polymorphic_sites.fasta > clean.core.aln
FastTreeMP -gtr -nt clean.core.aln > clean.core.tree
```

# 细菌核心基因组SNP进化树的构建

- 流程图

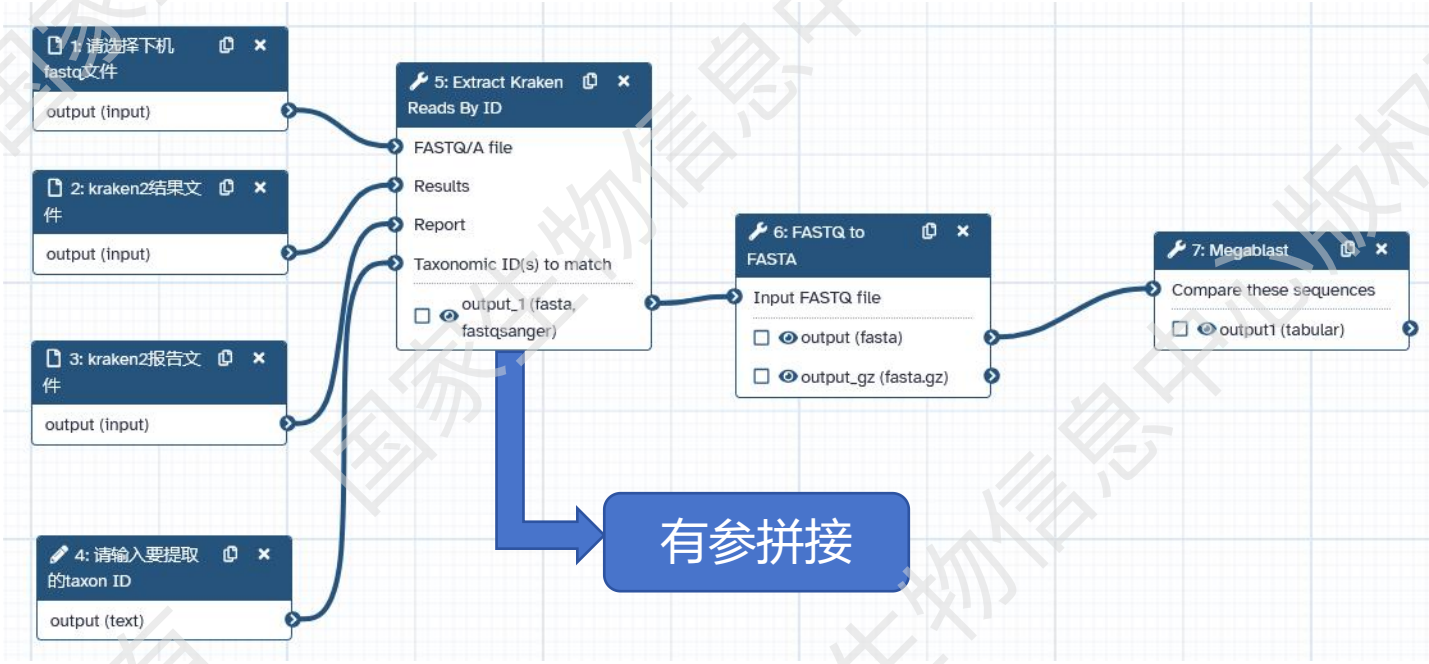


# 宏基因组测序病原体鉴定



# 如何进行宏基因组测序鉴定到的阳性病原体的复核?

- 根据病原体鉴定结果，从原始下机fastq文件中提取病原体特异性reads，并进行blast比对
- blast比对确认无误后，将提取到的reads进行有参拼接，观察reads在参考基因组上的分布情况




宏基因鉴定结果复核流程图

# 谢谢！

# Thank you!