广东省公共卫生研究院
GUANGDONG PROVINCIAL INSTITUTE OF PUBLIC HEALTH

宏基因组测序在疾控工作中的应用

刘哲

2025年6月28日

广东省公共卫生研究院
GUANGDONG PROVINCIAL INSTITUTE OF PUBLIC HEALTH

第一章	宏基因组测序样品制备
第二章	宏基因组测序数据分析
第三章	细菌宏基因组样本制备

第一章
宏基因组测序样品制备

1. 原理应用范围
2. 样品富集浓缩
3. 文库构建方法比较
4. 探针捕获测序
5. 测序平台与测序数据量
6. 细菌基因组样品制备

1 原理与应用范围

- 大部分病毒基因组为RNA，而高通量测序上机样品为双链DNA。因此需要逆转录、二链合成等额外步骤。
- 病毒为全寄生生长，培养需要借助其它生物（细胞），因此测序时极易被其它生物（宿主细胞、细菌）的基因组污染，这增加了高通量测序文库构建的复杂性
- 病毒基因组很小，使得样本中的病毒核酸量一般不足1ng，而高通量测序文库构建对核酸浓度有要求，因此PCR不可避免。

1 技术优势与劣势

优势：

- 不受病毒序列变异的影响
- 建库方法的通用性：一套试剂通吃所有病毒测序

劣势：

- 灵敏度很低，仅适用于病毒载量较高的样本
- 需要较高的测序数据量，对测序仪提出了更高的要求

1 适合的应用场景

- 培养病毒株的全基因组测序
- 病毒载量较高的临床样本病毒全基因组测序
- 背景较为“干净”的临床标本的病毒全基因组测序，如血清、血浆、脑脊液、肺泡灌洗液等等
- 多病原检测阴性样本的病原鉴定
- 不推荐直接使用宏基因组测序方法分析污水等环境标本中的病毒构成

1 利用宏基因组测序毒株

Name	Number of raw reads	Classified reads	Chordate reads	Artificial reads	Unclassified reads	Microbial reads	Bacterial reads	Viral reads	Fungal reads	Protozoan reads
XH3_S5	276,680	41.4%	30.5%	0%	58.6%	10.7%	0.0195%	5.26%	0.0173%	0.144%
XH4_S6	346,466	39.4%	31.4%	0%	60.6%	7.93%	0.017%	5.93%	0.00317%	0.0687%
XH5_S7	302,446	25.9%	17.6%	0%	74.1%	8.35%	0.0155%	6.95%	0.0086%	0.0456%
XH6_S8	286,364	55.1%	15.1%	0%	44.9%	39.9%	34.3%	3.88%	0.00349%	0.0367%
XH7_S9	326,340	62.3%	20%	0%	37.7%	42.2%	37.8%	3.41%	0.00521%	0.0371%
XH8_S10	283,864	73.9%	28.2%	0%	26.1%	45.7%	41%	2.31%	0.00528%	0.0447%

MiSeq V2 Nano芯片 300cycle试剂盒测序10份标本

2 环境污水的病毒浓缩富集

- 化学沉淀法： 加入聚乙二醇（PEG）或硫酸铵等化学试剂，使病毒或微生物絮凝沉淀 。

优势：低成本、仅需要离心机

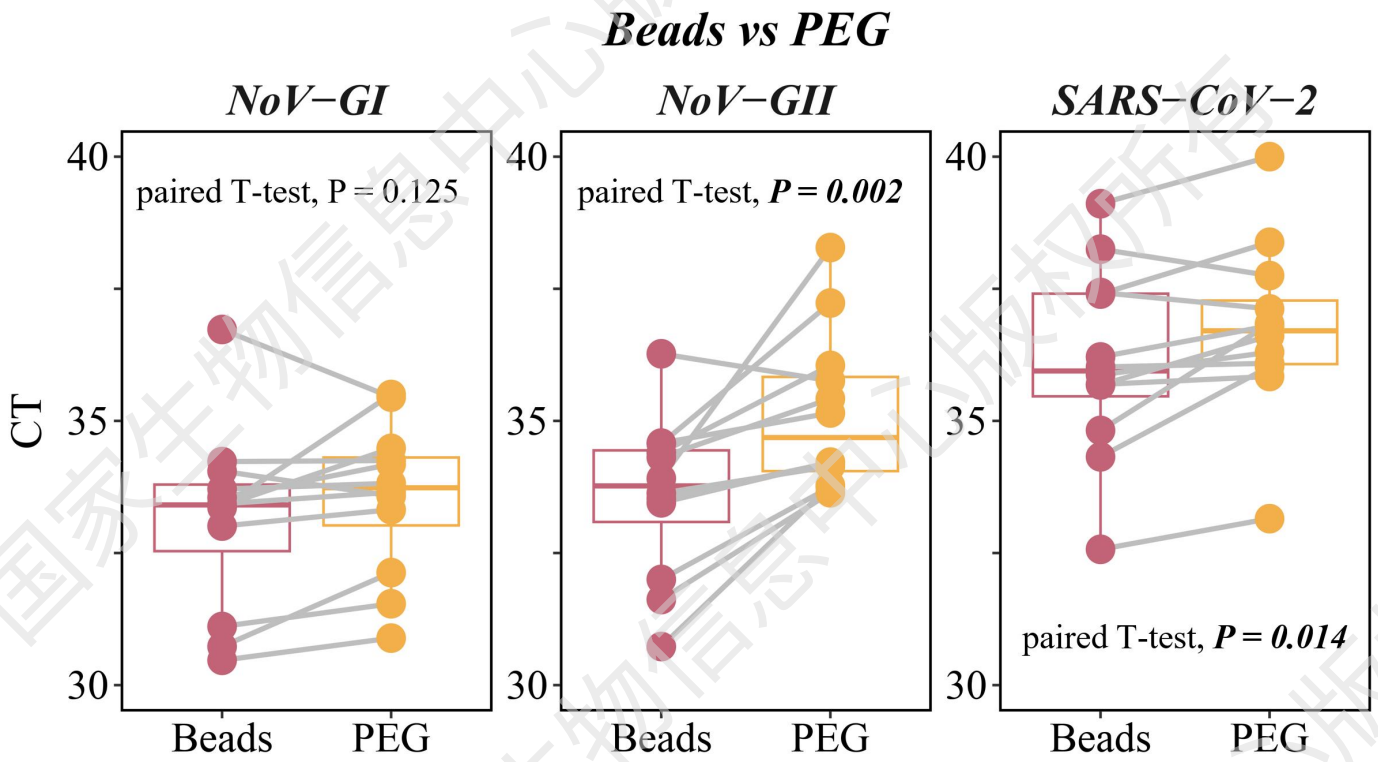
劣势：操作繁琐、难以自动化，不稳定，受操作人员的影响很大

- 磁珠法：利用了磁珠表面与病毒颗粒之间的特异性或非特异性结合，并通过磁力将结合了病毒的磁珠从大量污水中分离出来，从而达到富集病毒的目的

优势：可以自动化批量操作，稳定性好

劣势：依赖仪器、磁珠价格高于PEG

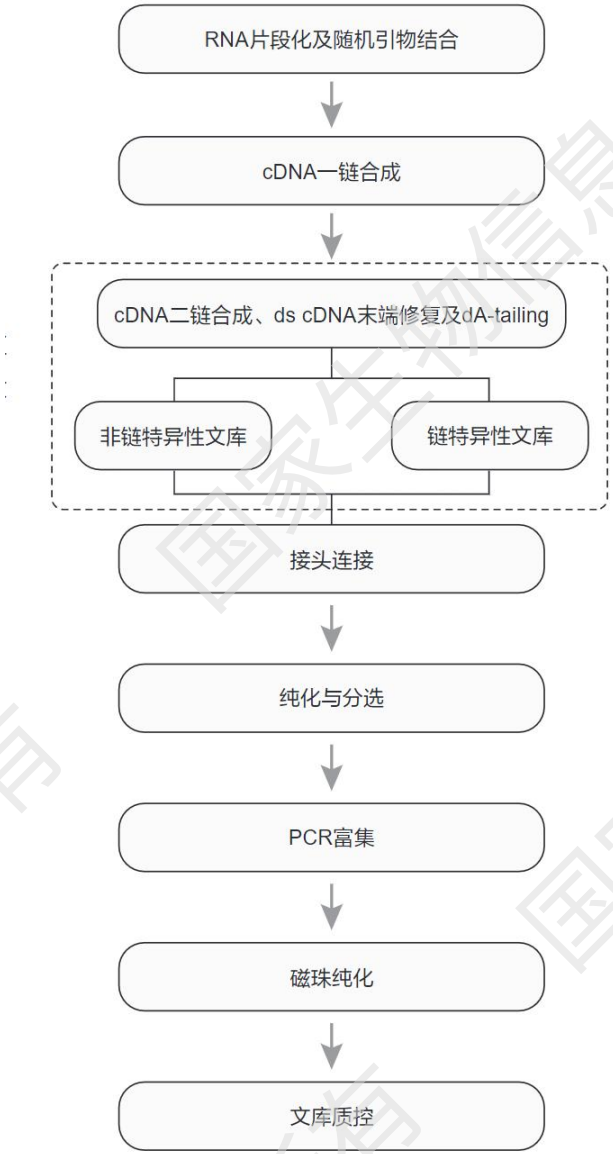
2 PEG vs 磁珠法



肠道病毒、诺如病毒、新冠病毒

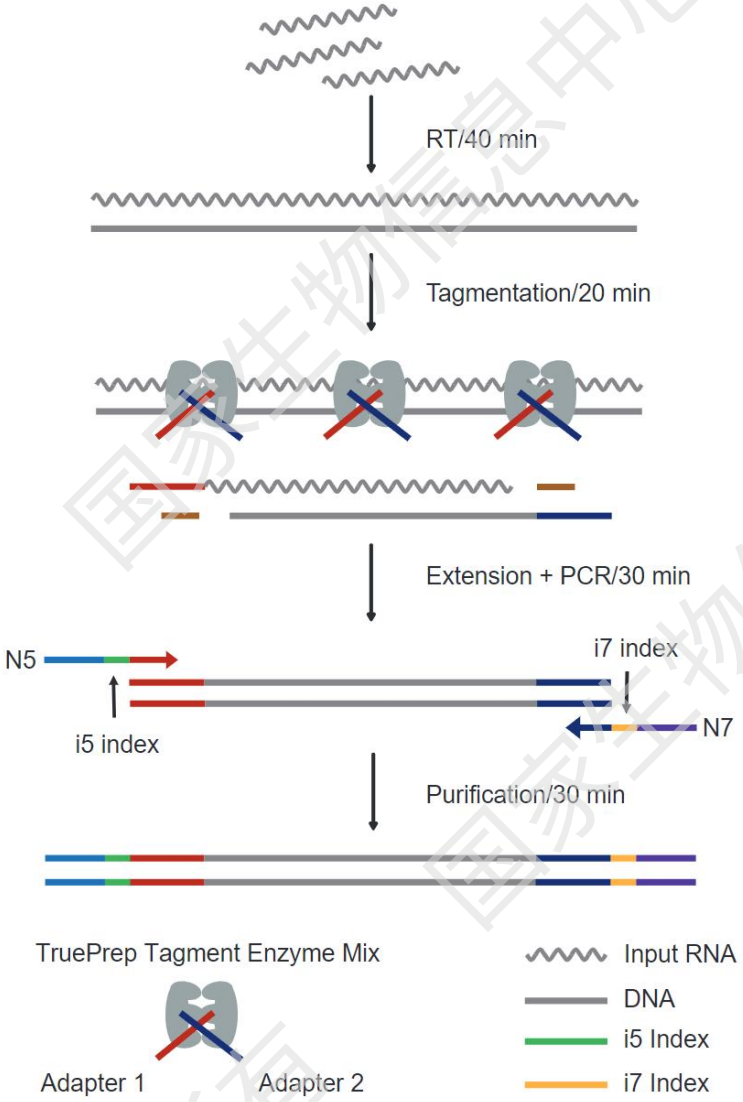
- Ct值比较：核酸病毒载量10mL磁珠富集后>40mL PEG富集
- 富集效率：~40%（磁珠）＞10%（PEG）

3 RNA病毒测序文库构建方法（酶连接法）

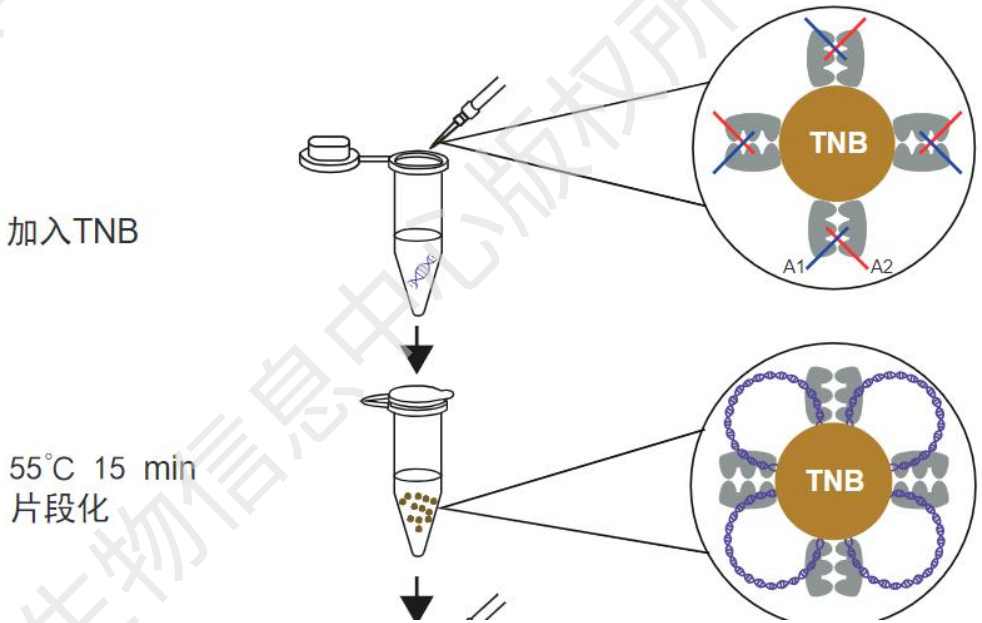


- 优点：成本较低，不挑核酸
- 缺点：效率较低、操作步骤多

3 RNA病毒测序文库构建方法（转座酶法）



- 优点：效率高、步骤少
- 缺点：价格较贵、对投入量敏感、不适用于严重降解的核酸样本

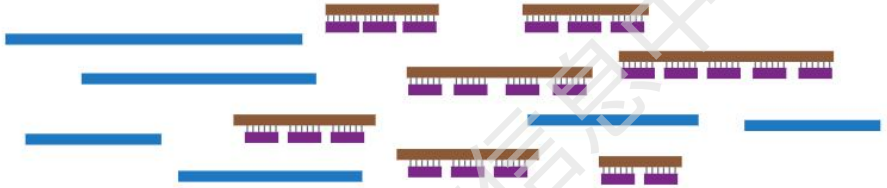


3 基因组消化与rRNA去除

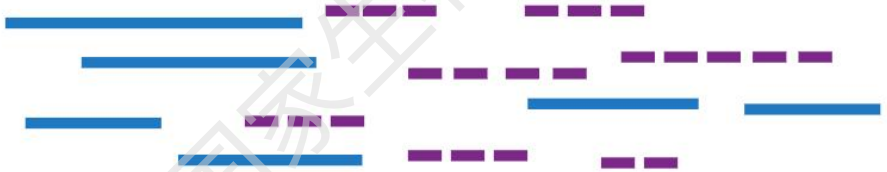
- 对于核酸浓度较低的标本（咽拭子、血清、血浆）推荐使用DNase进行消化后逆转录建库。建议选择温度敏感型Dnase（比如EzDNase），避免使用EDTA和RNA磁珠纯化
- 对于核酸浓度较高的标本（粪便、组织、全血）推荐使用rRNA去除
- 环境污水建议使用探针捕获

3 rRNA去除

1. rRNA与探针杂交（rRNA probe hybridization）



2. RNase H消化（RNase H depletion）



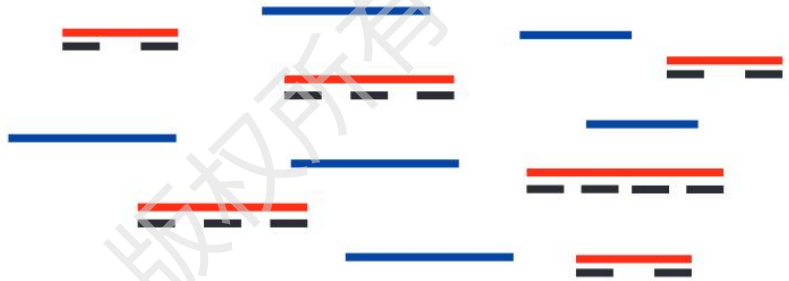
3. DNase I消化（DNase I depletion）

rRNA

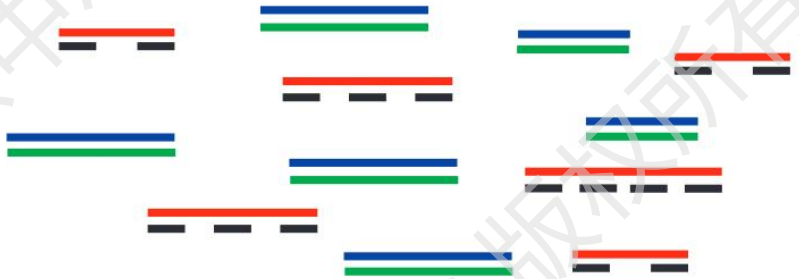
mRNA或nc-RNA

rRNA探针

1. rRNA与rRNA antigens结合



2. cDNA一链合成



3. cDNA二链合成

rRNA

mRNA或ncRNA

rRNA antigens
(Fast rRNA
Removal Mix包含
rRNA antigens)

广东省公共卫生研究院
GUANGDONG PROVINCIAL INSTITUTE OF PUBLIC HEALTH

4 探针捕获测序

- 通过预先设计的寡核苷酸探针从宏基因组文库中富集目标序列
- 对比多重PCR，有更好的序列兼容性和抗干扰能力
- 对比宏基因组测序，有很高的灵敏度，能够从低浓度样本中获得全基因组
- 缺点：操作繁琐、防污染、价格昂贵

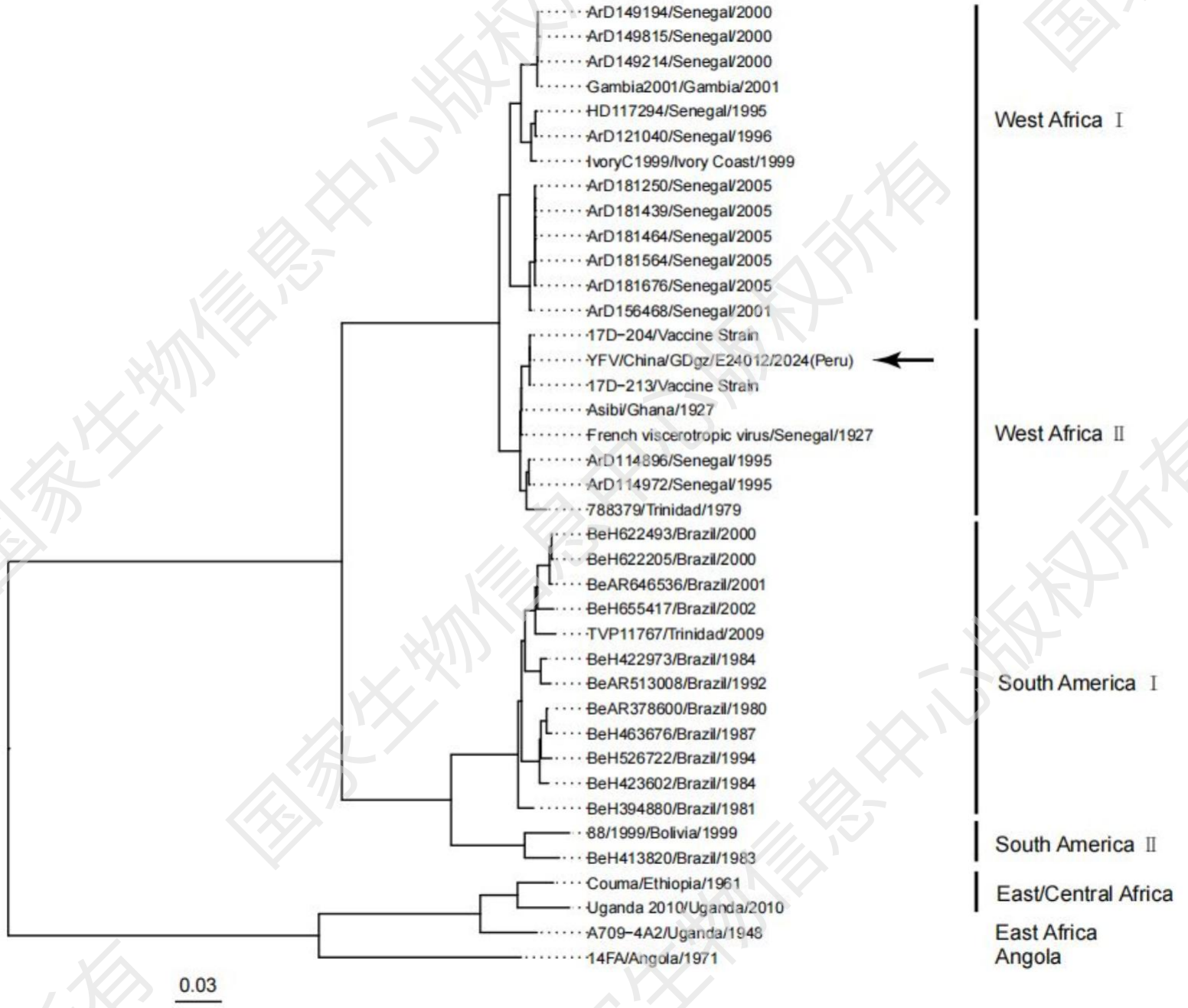
4 探针捕获测序

- 通过预先设计的寡核苷酸探针从宏基因组文库中富集目标序列
- 对比多重PCR，有更好的序列兼容性和抗干扰能力
- 对比宏基因组测序，有很高的灵敏度，能够从低浓度样本中获得全基因组
- 适合的场景：疫情标本、环境污水宏病毒组

- 缺点：操作繁琐、防污染、价格昂贵

利用探针捕获试剂处理黄热病疫情

- 2024年10月份，广交会期间，一名秘鲁籍男性在入境时被检测到黄热病病毒核酸阳性
- 患者在入境中国前接种过黄热病疫苗，并且没有非洲旅行史。
- 广东省CDC立即启动应急响应机制，对患者进行采样检测，结果同样为黄热病病毒核酸阳性，但CT值仅为33
- 最终，广东CDC使用探针捕获方案对样本中的病毒核酸进行浓缩富集，最终获得一条覆盖度94%的病毒基因组序列。
- 经过序列比对，该序列与黄热病疫苗株序列高度相似

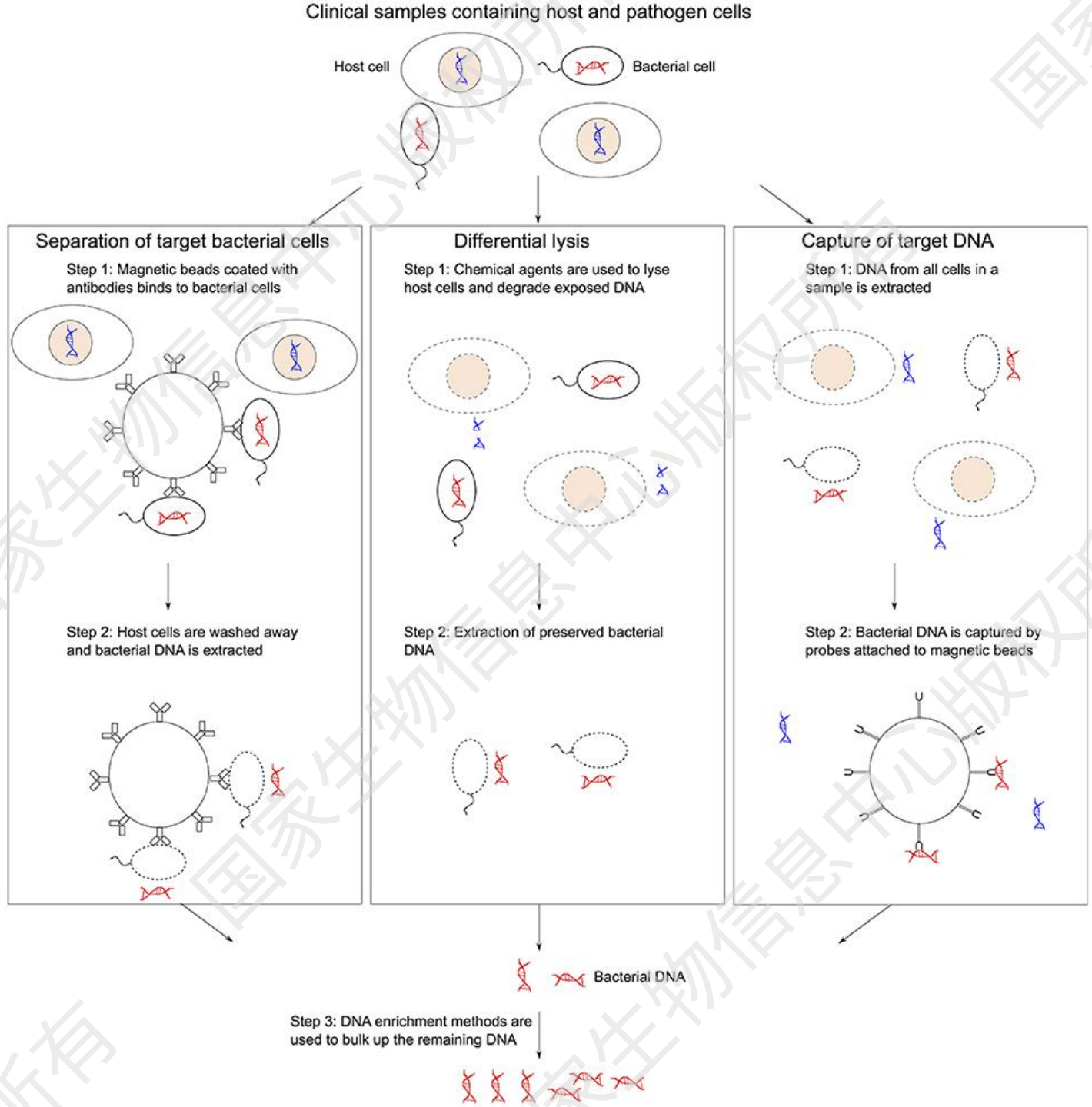

ArD149194/Senegal/2000
ArD149815/Senegal/2000
ArD149214/Senegal/2000
Gambia2001/Gambia/2001
HD117294/Senegal/1995
ArD121040/Senegal/1996
IvoryC1999/Ivory Coast/1999
ArD181250/Senegal/2005
ArD181439/Senegal/2005
ArD181464/Senegal/2005
ArD181564/Senegal/2005
ArD181676/Senegal/2005
ArD156468/Senegal/2001
17D-204/Vaccine Strain
YFV/China/GDgz/E24012/2024(Peru)
17D-213/Vaccine Strain
Asibi/Ghana/1927
French viscerotropic virus/Senegal/1927
ArD114896/Senegal/1995
ArD114972/Senegal/1995
788379/Trinidad/1979
BeH622493/Brazil/2000
BeH622205/Brazil/2000
BeAR646536/Brazil/2001
BeH655417/Brazil/2002
TVP11767/Trinidad/2009
BeH422973/Brazil/1984
BeAR513008/Brazil/1992
BeAR378600/Brazil/1980
BeH463676/Brazil/1987
BeH526722/Brazil/1994
BeH423602/Brazil/1984
BeH394880/Brazil/1981
88/1999/Bolivia/1999
BeH413820/Brazil/1983
Couma/Ethiopia/1961
Uganda 2010/Uganda/2010
A709-4A2/Uganda/1948
14FA/Angola/1971
0.03
West Africa Ⅰ
West Africa Ⅱ
South America Ⅰ
South America Ⅱ
East/Central Africa
East Africa
Angola

5 测序平台及数据量

- 三代测序平台因为reads数不足以及较高的测序成本，不适合做病毒宏基因组测序
- 小通量测序平台（MiSeq、MiniSeq）开机成本过高，只适合进行应急测序（疫情样本的测序）
- 推荐每个样本数据量不少于6Gb，使用PE150或者PE100进行宏基因组测序。

1 样品制备

- 环境污水需要使用超滤或者离心提高样本中的细菌浓度
- 临床标本需要一些宿主DNA去除策略
- **选择性裂解 (Selective Lysis):** 哺乳动物细胞膜相比于细菌细胞壁更为脆弱。利用这一特性，可以使用温和的去垢剂（如皂苷、Triton X-100）或渗透压冲击（低渗溶液）来优先裂解宿主细胞。裂解后，加入DNase I消化释放出的宿主DNA，而完整的细菌细胞则通过离心回收，再进行后续的核酸提取 。


Clinical samples containing host and pathogen cells
Host cell
Bacterial cell
Separation of target bacterial cells
Step 1: Magnetic beads coated with antibodies binds to bacterial cells
Step 2: Host cells are washed away and bacterial DNA is extracted
Differential lysis
Step 1: Chemical agents are used to lyse host cells and degrade exposed DNA
Step 2: Extraction of preserved bacterial DNA
Capture of target DNA
Step 1: DNA from all cells in a sample is extracted
Step 2: Bacterial DNA is captured by probes attached to magnetic beads
Bacterial DNA
Step 3: DNA enrichment methods are used to bulk up the remaining DNA

第二章
宏基因组数据分析

1 样品/数据质控

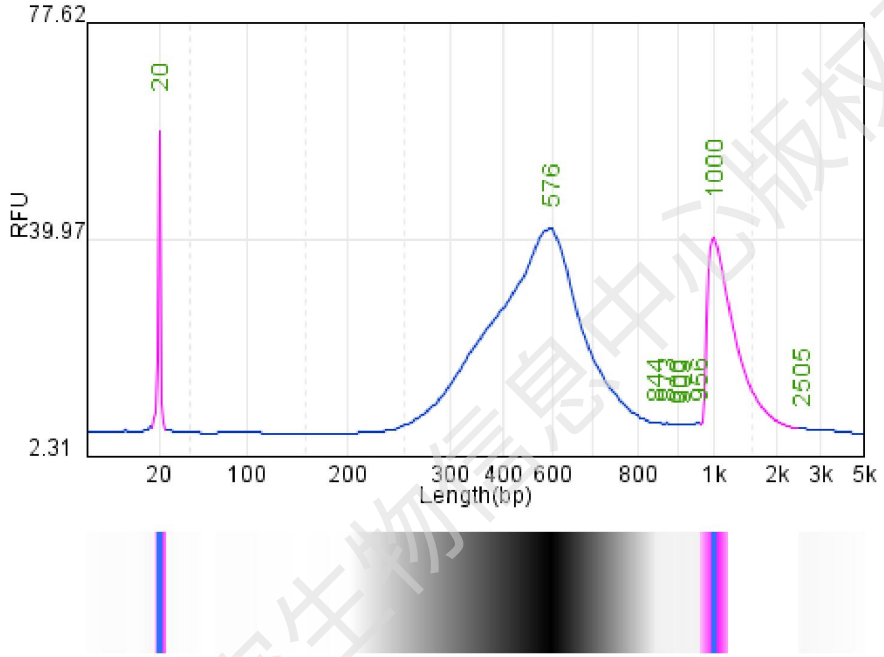
2 病毒的数据分析

3 细菌的数据分析

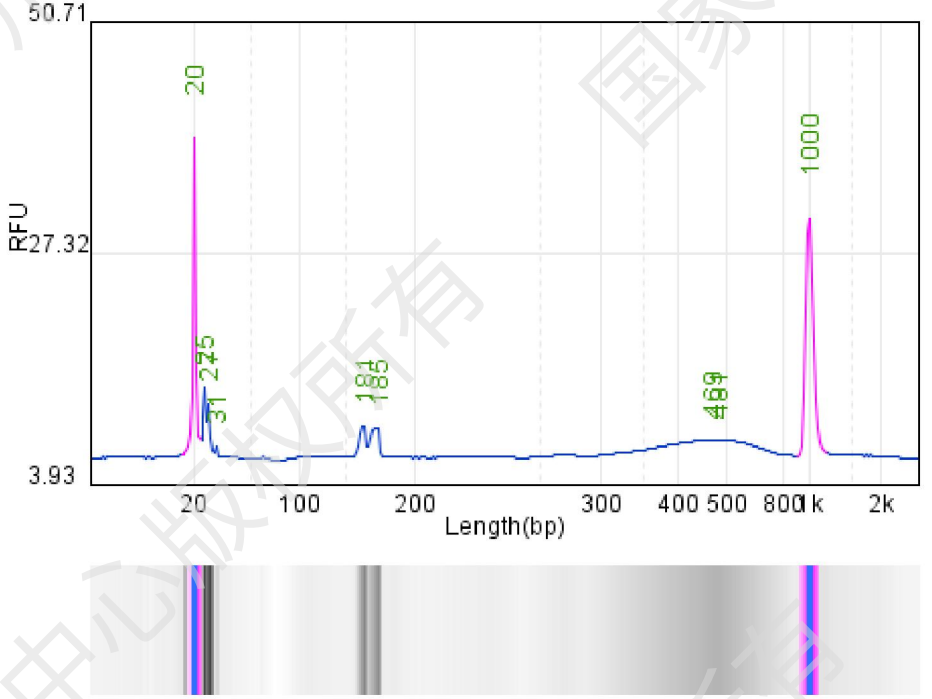
1 上机前的质量控制

- Qubit：用于文库的精确定量。Nanodrop不适合此项工作
- 毛细管电泳：Qseq或者安捷伦2100，用于确定文库插入片段的大小，以及是否有引物二聚体等问题

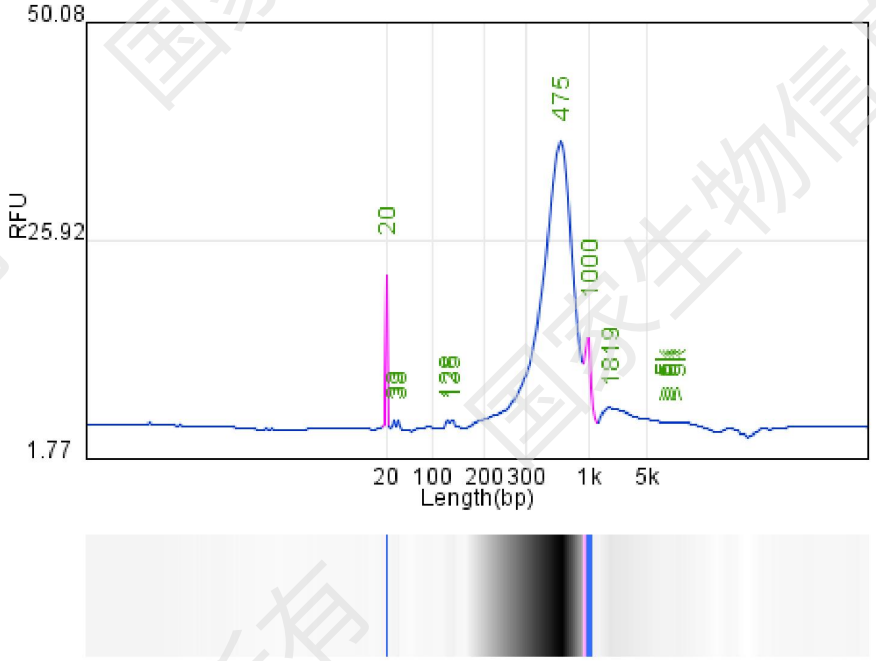
正常文库



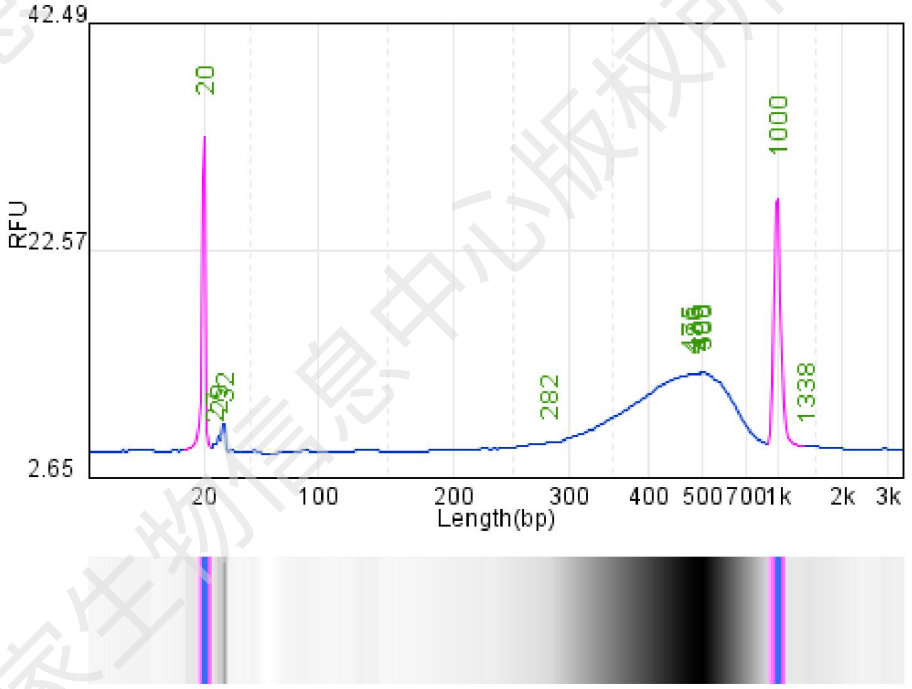
浓度过低
引物二聚体



浓度过高



小片段偏多



1 Fastp质控测序数据

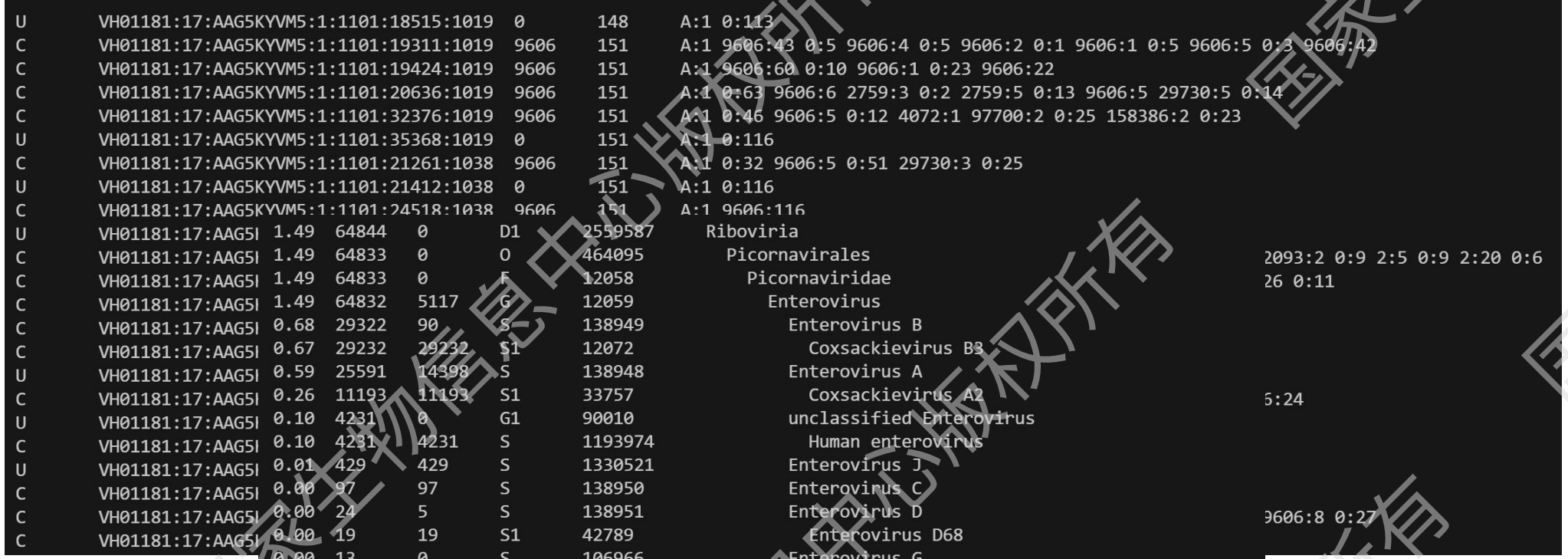
- 确定测序数据量（碱基数、reads数）
- 确定插入片段长度
- GC含量
- 数据质量
- fastp -i <输入R1文件> -o <输出R1文件> -I <输入R2文件> -O <输出R2文件> -f 10 –F 10 --t 5 –T 5 –h report.html –j report.json

2 病毒序列分析：Kraken2

- 直接对样本数据中的序列进行物种分类，生成样本中物种构成的基本分布
- 优点：速度快、结果直观
- 缺点：分类不准确，尤其对于细菌，要严格鉴定，部分结果需要二次确认
- kraken2 --db <数据库路径> --threads <线程数> --paired --output <分类结果输出文件> --report <报告文件> <R1序列文件> <R2序列文件>

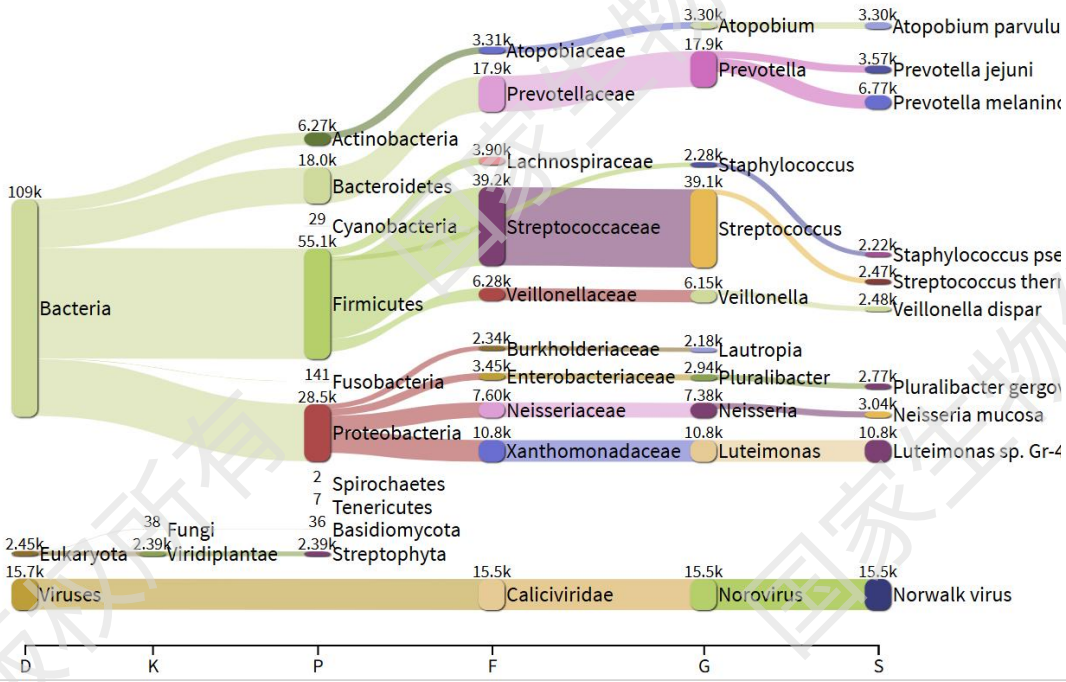
KrakenTools

- KrakenTools：分析**Kraken**、**Kraken2**、**Bracken**和**KrakenUniq**等宏基因组分类工具的输出文件。这些脚本使用户能够进行**下游分析**，例如提取特定分类单元的读数、合并报告、将数据转换为Krona或MetaPhlAn兼容格式
- python extract_kraken_reads.py -k myfile.kraken -s SEQUENCE.FILE -o OUTPUT.FASTA -t TID TID2
- https://github.com/jenniferlu717/KrakenTools



```
U       VH01181:17:AAG5KYVM5:1:1101:18515:1019  0       148     A:1 0:113
C       VH01181:17:AAG5KYVM5:1:1101:19311:1019  9606    151     A:1 9606:43 0:5 9606:4 0:5 9606:2 0:1 9606:1 0:5 9606:5 0:3 9606:42
C       VH01181:17:AAG5KYVM5:1:1101:19424:1019  9606    151     A:1 9606:60 0:10 9606:1 0:23 9606:22
C       VH01181:17:AAG5KYVM5:1:1101:20636:1019  9606    151     A:1 0:63 9606:6 2759:3 0:2 2759:5 0:13 9606:5 29730:5 0:14
C       VH01181:17:AAG5KYVM5:1:1101:32376:1019  9606    151     A:1 0:46 9606:5 0:12 4072:1 97700:2 0:25 158386:2 0:23
U       VH01181:17:AAG5KYVM5:1:1101:35368:1019  0       151     A:1 0:116
C       VH01181:17:AAG5KYVM5:1:1101:21261:1038  9606    151     A:1 0:32 9606:5 0:51 29730:3 0:25
U       VH01181:17:AAG5KYVM5:1:1101:21412:1038  0       151     A:1 0:116
C       VH01181:17:AAG5KYVM5:1:1101:24518:1038  9606    151     A:1 9606:116
U       VH01181:17:AAG5I 1.49   64844   0       D1      2559587     Riboviria
C       VH01181:17:AAG5I 1.49   64833   0       O       464095        Picornavirales                                    2093:2 0:9 2:5 0:9 2:20 0:6
C       VH01181:17:AAG5I 1.49   64833   0       F       12058           Picornaviridae                                  26 0:11
C       VH01181:17:AAG5I 1.49   64832   5117    G       12059             Enterovirus
C       VH01181:17:AAG5I 0.68   29322   90      S       138949              Enterovirus B
C       VH01181:17:AAG5I 0.67   29232   29232   S1      12072                 Coxsackievirus B3
U       VH01181:17:AAG5I 0.59   25591   14398   S       138948              Enterovirus A
C       VH01181:17:AAG5I 0.26   11193   11193   S1      33757                 Coxsackievirus A2                         5:24
U       VH01181:17:AAG5I 0.10   4231    0       G1      90010               unclassified Enterovirus
C       VH01181:17:AAG5I 0.10   4231    4231    S       1193974               Human enterovirus
U       VH01181:17:AAG5I 0.01   429     429     S       1330521             Enterovirus J
C       VH01181:17:AAG5I 0.00   97      97      S       138950              Enterovirus C
C       VH01181:17:AAG5I 0.00   24      5       S       138951              Enterovirus D                             9606:8 0:27
C       VH01181:17:AAG5I 0.00   19      19      S1      42789                 Enterovirus D68
        0.00    13      0       S       106966              Enterovirus G
        0.00    13      13      S1      64141                 Porcine enterovirus 9
        0.00    7       0       S       2169885             Enterovirus L
        0.00    7       7       S1      1826059               Enterovirus SEV-gx
        0.00    1       1       S       463676              Rhinovirus C
        0.00    1       0       F1      35296           unclassified Picornaviridae
        0.00    1       1       S       928289                Pigeon picornavirus B
        0.00    6       0       P       2497569     Negarnaviricota
        0.00    4       0       P1      2497570       Haploviricotina
        0.00    4       0       C       2497574         Monjiviricetes
        0.00    4       0       O       11157             Mononegavirales
        0.00    4       0       F       11266               Filoviridae
        0.00    4       0       G       186536                Ebolavirus
        0.00    4       4       S       186538                  Zaire ebolavirus
        0.00    2       0       P1      2497571       Polyploviricotina
        0.00    2       0       C       2497577         Insthoviricetes
        0.00    2       0       O       2499411           Articulavirales
        0.00    2       0       F       11308               Orthomyxoviridae
        0.00    2       0       G       197912                Betainfluenzavirus
        0.00    2       0       S       11520                   Influenza B virus
        0.00    2       2       S1      518987                    Influenza B virus (B/Lee/1940)
        0.00    5       0       O       76804           Nidovirales
```

2 使用Pavian可视化kraken2的结果



Click a row to see the abundance of the taxon in other samples.

Show 15 rows | CSV | Print | Copy | Column visibility

Search: Riboviria

Percentage	CladeReads	TaxonReads	TaxRank	TaxID	Name	TaxLineage
12.22%	15,503	0	-	2559587	Riboviria	Viruses
12.22%	15,503	0	F	11974	Caliciviridae	Viruses>Riboviria
12.22%	15,503	0	G	142786	Norovirus	Viruses>Riboviria>Caliciviridae
12.22%	15,503	12	S	11983	Norwalk virus	Viruses>Riboviria>Caliciviridae>Norovirus
12.2%	15,476	6	-	122929	Norovirus GII	Viruses>Riboviria>Caliciviridae>Norovirus>No
12.2%	15,470	15,470	-	552592	Norovirus GII.17	Viruses>Riboviria>Caliciviridae>Norovirus>No
0.01104%	14	14	-	95340	Norwalk-like virus	Viruses>Riboviria>Caliciviridae>Norovirus>No
0.0007885%	1	1	-	122928	Norovirus GI	Viruses>Riboviria>Caliciviridae>Norovirus>No
All	All	All	All	All	All	All

2 从头拼接、序列注释

- Megahit:快速的从头拼接工具，将fastq文件拼接为包含contig的fasta文件
  megahit -1 <Read1_序列文件> -2 <Read2_序列文件> -o <输出目录> -t 8 --min-contig-len 500
- BLASTn：最为基础和相对准确的序列注释工具。适用于对已知病毒序列的注释。缺点是速度比较慢。
  blastn -query contigs.fasta –out result.tsv -db path/to/nt -num_threads 32 -max_target_seqs 1 -outfmt "6 qseqid sacc pident length evalue bitscore scomname sskingdom stitle “
- 从数据库中提取序列：
  blastdbcmd -db /path/to/nt -entry qacc –out result.fasta

- BLASTx：将核苷酸序列翻译为氨基酸序列，可以注释未报道的病毒序列。缺点是非常慢
  blastx -query contigs.fasta –out result.tsv -db path/to/nr -num_threads 32 -max_target_seqs 1 -outfmt "6 qseqid sacc pident length evalue bitscore scomname sskingdom stitle “

- Diamond：快速的序列注释工具，替代BLASTx
  diamond blastx -d <参考数据库路径> -q <查询序列文件路径> -o <输出文件路径> --outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore staxids sscinames sskingdoms

2 基于参考序列的拼接：call consensus

- 基于一条参考序列，使用bwa index构建索引
  bwa index ref.fasta
- 使用bwa mem结合samtools将fastq的序列回帖到参考序列，生成bam文件
  bwa mem –a –t 16 ref.fasta read1 reads2 | samtoools view –bS –F 4|samtools sort –o result.bam
- 使用samtools或者iVar生成consensus序列
  samtools index result.bam
  samtools mpileup result.bam | ivar consensus –m 5 –p samplename

2 未知病毒的发现

- 基于Megahit的contig序列
- 新病毒的发现方法：

基于比对：BLAST、Diamond

基于HMM：Virsorter

基于AI：Lucaprot

3 细菌基因组分析

1 质控与组装

清洗原始数据，使用MEGAHIT或metaSPAdes等工具将短序列拼接成长片段(contigs)。

2 基因组分箱

将来自同一物种的contigs聚类，重构出宏基因组组装基因组(MAGs)。

3 物种分类

利用Kraken2、MetaPhlAn或GTDB-Tk确定MAGs的物种身份，回答“它们是谁？”

4 功能注释

比对CARD、VFDB等数据库，鉴定耐药(ARGs)和毒力(VFs)基因，回答“它们能做什么？”

5 高级分析

重建质粒、溯源ARGs的宿主、并发现全新的功能基因。

6 AI赋能

利用深度学习和大型模型优化流程，预测表型，加速新药发现。

3 基因组分箱（Binning）重构物种基因组

- 宏基因组组装的产物是数以万计的contigs混合物，它们来自样本中成百上千个不同的物种。基因组分箱（Binning）的目标，就是将这些混杂的contigs进行聚类，把来自同一个物种的contigs归拢到同一个“箱子”（bin）里，从而重建出该物种的基因组草图，即MAG。

3 主流的分箱工具

- MetaBAT 2: 结合了四核苷酸频率和跨样本的差异覆盖度信息，通过计算contig间的相似性构建一个图，然后使用一种改进的标签传播算法（label propagation algorithm）对图进行分割，实现分箱。
- MaxBin 2: 同样利用四核苷酸频率和contig覆盖度，但它采用的是一种统计模型——期望最大化（Expectation-Maximization, EM）算法，来迭代地计算每个contig属于每个bin的概率，并最终完成分配。
- CONCOCT: 也是利用序列组成和差异覆盖度，它先用主成分分析（PCA）对特征进行降维，然后使用高斯混合模型（Gaussian Mixture Models, GMM）进行聚类。

3 物种鉴定

- GTDB-Tk: 这是目前对MAGs进行分类的金标准工具。它不依赖于简单的序列比对，而是基于系统发育学的原理。它首先从待分析的MAG中鉴定出一套保守的单拷贝蛋白（细菌有120个，古菌有122个），然后将这些蛋白序列与一个巨大的、涵盖了整个细菌和古菌生命树的参考蛋白集进行比对，构建一个系统发育树，通过计算MAG在这个“生命之树”上的精确位置来确定其分类学归属。

  gtdbtk classify_wf --genome_dir ./fasta/ --out_dir gtdb/ --cpus 32 -x fasta --skip_ani_screen

3 物种鉴定

- GTDB-Tk: 这是目前对MAGs进行分类的金标准工具。它不依赖于简单的序列比对，而是基于系统发育学的原理。它首先从待分析的MAG中鉴定出一套保守的单拷贝蛋白（细菌有120个，古菌有122个），然后将这些蛋白序列与一个巨大的、涵盖了整个细菌和古菌生命树的参考蛋白集进行比对，构建一个系统发育树，通过计算MAG在这个“生命之树”上的精确位置来确定其分类学归属。

  gtdbtk classify_wf --genome_dir ./fasta/ --out_dir gtdb/ --cpus 32 -x fasta --skip_ani_screen

3 长读长测序的技术优势

- 代表： ONT, PacBio，今是科技，华大序风
- 特点： 产生数千至数万碱基的超长读段 (Reads)。
- 基因组组装： 轻松跨越重复序列，获得更完整、连续的宏基因组组装基因组 (MAGs)。
- 追踪耐药性： 可完整测序质粒等移动元件，直接关联耐药基因与宿主，为追踪其传播提供关键证据。
- 挑战： 传统上错误率和成本较高。

3 二三代混合策略

- 该策略利用长读长构建基因组的骨架，解决复杂的结构和重复区域问题，然后利用高精度的短读长对长读长组装的结果进行纠错和补洞，以获得既完整又准确的基因组。这被认为是当前获得高质量MAGs和闭合质粒的最佳途径之一 。

总结与展望

AI革命：深度学习与大型模型的兴起

人工智能正在从根本上改变宏基因组分析。模型从“特征工程”转向“表示学习”，极大地提升了分析的精度和预测能力。

序列分类与分箱

CNN和VAE（如VAMB）等模型能自动学习序列特征，实现更精准的物种分类和更高质量的MAGs恢复。

AMR预测与新药发现

AI模型能够学习复杂的“基因型-表型”关系，高精度预测耐药性，并生成全新的抗菌药物分子。

蛋白质/基因组语言模型

将生物序列视为语言，ESM-2等模型无需实验即可预测新蛋白的功能，快速评估新发病原体的风险。

广东省公共卫生研究院
GUANGDONG PROVINCIAL INSTITUTE OF PUBLIC HEALTH

感谢您的聆听