二代宏基因组分析流程详解

二代宏基因组分析流程详解：

01：data

测试数据：肠道微生物宏基因组

SAL-10.R1.fq.gz/SAL-10.R1.fq.gz，2GB

宿主基因组数据：人类基因组

GCF_000001405.39_GRCh38.p13_genomic.fna.gz，921MB

02：质控

```
''' 

    /export/personal/software/software/fastp/v0.20.0/fastp --thread 10 -i /export/personal1/liupb/ngs_meta/raw_data/SAL-10.R1.fq.gz -I /export/personal1/liupb/ngs_meta/raw_data/SAL-10.R2.fq.gz -o SAL-10.R1.clean.fq.gz -O SAL-10.R2.clean.fq.gz -h ASL-10.html

''' 
```

10个线程，质控时长约5分钟

03：去宿主

bowtie2:

index构建：

```
''' 

    /export/personal/software/software/bowtie2/v2.4.2/bowtie2-build ../host_genomic/GCF_000001405.39_GRCh38.p13_genomic.fna.gz bowtie2_index/index

''' 
```

1个线程，构建时长约为30分钟

比对：

```
''' 

    /export/personal/software/software/bowtie2/v2.4.2/bowtie2 -p 10 -x bowtie2_index/index -1 ../filter/SAL-10.R1.clean.fq.gz -2 ../filter/SAL-10.R2.clean.fq.gz --un-conc unpaired -S bowtie2.sam

''' 
```

10个线程，比对时长约为30分钟

minimap2:

index构建：

```
''' 

    /export/personal/software/software/minimap2/v2.17/minimap2 -d index ../host_genomic/GCF_000001405.39_GRCh38.p13_genomic.fna.gz -t 4
```

’ ”

4个线程，构建时长约10分钟

比对：

“ ’

```
/export/personal/software/software/minimap2/v2.17/minimap2 -ax sr index ../filter/SAL-10.R1.clean.fq.gz ../filter/SAL-10.R2.clean.fq.gz -t 10 > mini.sam
```

’ ”

10个线程，比对时长约22分钟

综合比较：minimap2整体比对速度比bowtie2快1/3左右，比对到宿主中的数量比bowtie2多，但质量相对较低。bowtie2比minimap2比对质量高，且能够生成未比对上的结果文件，minimap2需要脚本从头读取sam文件提取未必对文件，相对较为繁琐。综合比较选择bowtie2作为流程去宿主比对软件

04：组装

megahit:

’’’

```
/export/personal/software/software/megahit/v1.2.9/bin/megahit -1 ../bowtie/X2X_R1.fasta -2 ../bowtie/X2X_R2.fasta -o assmble -t 10
```

“ ’

10个线程，组装时长约31分钟

’ ”

```
碱基总数： 4970976
reads总数： 305
reads平均长度： 16298.28
reads最长长度： 443489
N50： 154170
GC含量： 0.52
N含量： 0.00
```

’ ”

metaSPAdes:

’’’

```
/export/personal1/liupb/test/metaSPAdes/SPAdes-3.15.3-Linux/bin/metaspades.py --only-assembler -t 10 -1 ../bowtie/X2X_R1.fastq -2 ../bowtie/X2X_R2.fastq -o result_spades
```

’’’

10个线程，耗时8小时32分钟

“ ’

```
碱基总数： 5156829
reads总数： 1319
reads平均长度： 3909.65
reads最长长度： 526401
N50： 206371
GC含量： 0.52
N含量： 0.00
```

' ''

综合比较：megahit耗时和内存使用比metaSPAdes有明显的差距，虽然相对与metaSPAdes在组装质量上稍有欠缺。综合考虑默认软件使用megahit，但流程中依旧保留metaSPAdes供用户选择

05：基因预测：

'' '

```
/export/personal/software/software/prokka/v1.14.6/bin/prokka /export/personal1/liupb/ngs_meta/Megahit/assmble/final.contigs.fa --metagenome --kingdom Bacteria --cpus 10 --prefix X2X --outdir X2X_prokka
```

' ''

10个线程，预测时长约5分钟

06：去冗余

'' '

```
/export/personal/software/software/cdhit/v4.8.1/cd-hit -i ../prokaa/X2X_prokka/X2X.faa -o X2X.protein.fasta -c 0.8 -aS 0.8 -d 0 -T 8
```

' ''

8个线程，运行时长约4秒

07：基因丰度统计

构建Index:

'' '

```
/export/personal1/liupb/test/salmon/Salmon-0.7.2_linux_x86_64/bin/salmon index -t /export/personal1/liupb/ngs_meta/Megahit/assmble/final.contigs.fa -i transcript_index --type quasi -k 31
```

' ''

1个线程，耗时约20秒

基因丰度统计：

' ''

```
/export/personal1/liupb/test/salmon/Salmon-0.7.2_linux_x86_64/bin/salmon quant -i transcript_index --libType IU -1 ../filter/SAL-10.R1.clean.fq.gz -2 ../filter/SAL-10.R2.clean.fq.gz -o X2X.quant -p 10
```

'' '

10个线程，耗时20秒（#改软件reads文件类型识别靠文件后缀）

08：功能注释

CAZY数据库注释：

' ''

> */export/personal/software/software/diamond/v0.9.26/diamond blastp --query /export/personal1/liupb/ngs_meta/cd-hit/X2X.protein.fasta --db /export/personal/software/database/CAZy/v9/CAZy.dmnd --outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen stitle --max-target-seqs 5 --evalue 1e-05 --threads 8 --out X2X.fasta.cazy.m6*

‘’ ’

8个线程，耗时约5分20秒

KEGG数据库注释：

’ ‘’

> */export/personal/software/software/diamond/v0.9.26/diamond blastp --query /export/personal1/liupb/ngs_meta/cd-hit/X2X.protein.fasta --db /export/personal/software/database/KEGG/20180701/prokaryotes.kegg.dmnd --outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen stitle --max-target-seqs 5 --evalue 1e-05 --threads 8 --out X2X.fasta.kegg.m6*

‘’ ’

8个线程，耗时约5分20秒

NOG数据库注释：

’ ‘’

> */export/personal/software/software/diamond/v0.9.26/diamond blastp --query /export/personal1/liupb/ngs_meta/cd-hit/X2X.protein.fasta --db /export/personal/software/database/NOG/20190302/microbe.nog.dmnd --outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen stitle --max-target-seqs 5 --evalue 1e-05 --threads 8 --out X2X.fasta.nog.m6*

‘’ ’

8个线程，耗时约5分40秒