

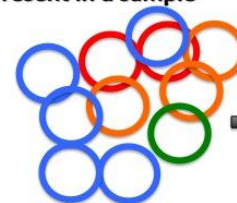


## 31 拼接/组装Assembly

易生信

2023年11月26日

Bacterial genomes  
present in a sample



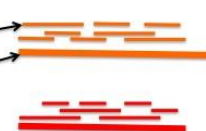
Genomes cut into small  
fragments



Sequencing of many random  
fragments from pool of  
fragments



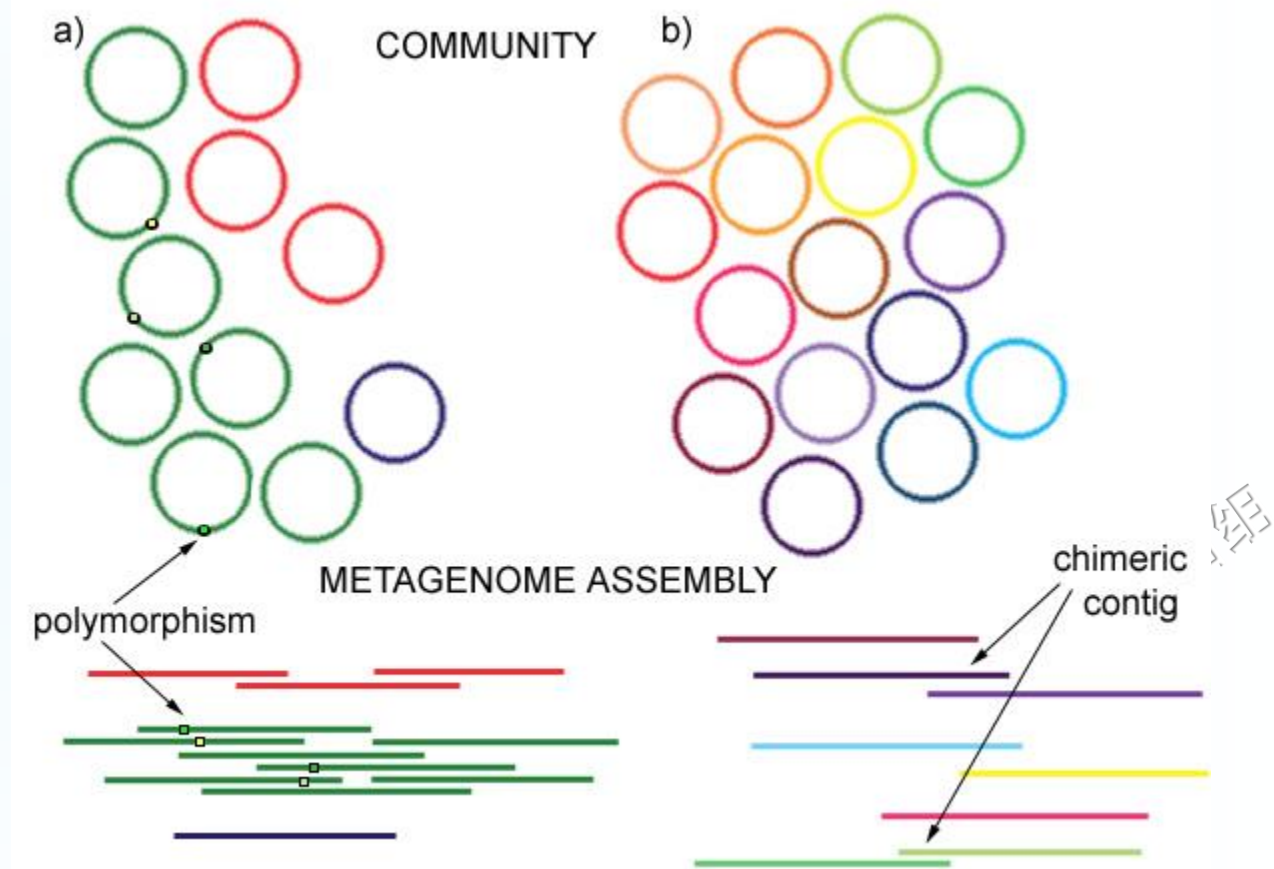
DNA  
sequences  
Computer-  
assembled  
consensus  
sequence



Alignment of DNA sequences with a computer  
program to create a larger consensus sequence

# 目录

- 一. 质控KneadData
- 二. 物种分类kraken 2
- 三. **序列组装/拼接**
- 四. 基因预测/注释
- 五. 基因聚类cd-hit
- 六. 基因定量salmon
- 七. 基因功能注释



# 组装/拼接 (Assemble)的基本原理

Bacterial genomes present in a sample



Genomes cut into small fragments



Sequencing of many random fragments from pool of fragments

DNA sequences  
Computer-assembled consensus sequence



Alignment of DNA sequences with a computer program to create a larger consensus sequence

宏基因组  
生信宝典





# 拼接中常见名词

- Read: 读长，高通量测序平台产生的序列
- Contig: 重叠群，基于读长之间的重叠区关系拼接获得的更长序列
- Scaffold: 支架，双端测序时，同一条序列的两端读长分布于不同的重叠群上，可确定两个重叠群的方向和距离时，将重叠群中间用N连接后的更长序列
- N50: 将重叠群或支架按长度由大到小排列，累加总长度50%时，所在序列长度，用于表示拼接质量的重要参数
- Depth: 测序深度，即测序总碱基与基因组大小的比值，如人类30x，即90G数据，宏基因组中要求较完整获得相对丰度1%的细菌基因组，测序量为： $5 \text{ MB} \times 30x \div 1\% = 15\text{GB}$
- Coverage: 覆盖度，测序获得的序列占整个基因组的比例，如97%即3%没测到。



# 拼接软件时间和内存比较

## (1) IDBA-UD

Running Time	33h 54m
Memory Utilization (GB)	123.84

## (2) SPAdes

Running Time	67h 02m
Memory Utilization (GB)	381.79

## (3) MEGAHIT

Running Time	1h 53m
Memory Utilization (GB)	33.41

Yu Peng, Henry C. M. Leung, S. M. Yiu, Francis Y. L. Chin. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420-1428. [10.1093/bioinformatics/bts174](https://doi.org/10.1093/bioinformatics/bts174)  
Cited by 2622

[IDBA-UD: 组装非均匀覆盖度的宏基因组和单细胞数据](#)

Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, Pavel A. Pevzner. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* 27: 824-834.

[10.1101/gr.213959.116](https://doi.org/10.1101/gr.213959.116) Cited by 2192

[metaSPAdes: 新型多功能宏基因组拼接工具](#)

Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiro Sadakane, Tak-Wah Lam. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31: 1674-1676.

[10.1093/bioinformatics/btv033](https://doi.org/10.1093/bioinformatics/btv033) Cited by 3607

[MEGAHIT: 复杂宏基因组拼接的超快速解决方案](#)

[综述metaSPAdes、IDBA-UD、MetaQuast、Prokka、metaProdigal](#)

# MEGAHIT——多快好省的组装神器

- 最快，最省内存，且在宏基因组拼接中质量可接受的软件
- -h显示参数详细
- -1/2左或右端文件，支持多文件；--12双端交替(interleave)的单文件；-r单端
- -t设置线程数，默认全用
- --use-gpu 支持GPU运算
- --continue 支持中断继续运行
- --k-min 27 --k-max 191 --k-step 20 手动设置kmer，**调整速度&精度**

[组装拼接MEGAHIT\(多快好省\)和评估quast](#)  
[MEGAHIT文章解读](#)

Li, Dinghua, et al. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph." *Bioinformatics* 31.10 (2015): 1674-1676.

Li, Dinghua, et al. "MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices." *Methods* 102 (2016): 3-11.



## 3.1.1 MEGAHIT拼接

### ○ 方法1. 混合组装(少量样本的推荐)

优点：简单快速获得一套参考序列，基因冗余度低，混合增加低丰度菌测序深度并且提高拼接长度和完整度

缺点：需要更大内存，混样提高错误拼接、嵌合体风险，高丰度区域碎片化

### ○ 方法2. 单样本组装(大量样本推荐)

优点：内存资源消耗少，防止样本间污染和嵌合体组装，高丰度菌重叠群更长

缺点：低丰度菌难组装较完整，样品间基因大量冗余，去冗余计算时间长

### ○ 方法3. 混合+单样本组装(样本量可完成计算下推荐)

优点：混合提高低丰度覆盖度，单样本防止样品间混淆，基因最完整

缺点：计算资源和时间消耗大，下游基因注释、去冗余时间长

[组装拼接MEGAHIT\(多快好省\)和评估quast](#)

[MEGAHIT文章解读](#)



# MEGAHIT拼接，混合快，单样本累计慢

# 组装，10~30m，TB级数据需几天至几周

time megahit -t 6 \

-1 `tail -n+2 result/metadata.txt|cut -f 1|sed 's/^/tempVqcV/;s\$/\_1.fastq/'| tr '\n' ','|sed 's/,,\$/'` \

-2 `tail -n+2 result/metadata.txt|cut -f 1|sed 's/^/tempVqcV/;s\$/\_2.fastq/'| tr '\n' ','|sed 's/,,\$/'` \

-o temp/megahit

# -t设置线程数量，默认使用所有线程，可能会影响其他人工作

# -1/2输入文件：反引号(`)使用shell命令基于元数据获得输入文件列表

# -o 输出目录，必须不存在，否则需要删除再运行

# 超过300GB，k-mer尽量调大，如29+，否则会超软件上限

增加参数加速：--k-min 29 --k-max 141 --k-step 20



## 3.1.2 metaSPAdes精细拼接

- 主页: <http://cab.spbu.ru/software/spades/>
- `conda install spades` # 安装软件
- `metaspades.py -h` # 查看帮助
- Meta帮助: <http://cab.spbu.ru/files/release3.12.0/manual.html#meta>
- `metaspades.py --test` # 运行测试数据
- 此软件 `--iontorrent` 支持PGM数据, 甚至支持`--pacbio`和`--nanopore`三代测序数据
- 原文简介: [metaSPAdes: 新型多功能宏基因组拼接工具](#)



# (可选) Metaspades组装，混合慢，单样本更快

# 混合组装：6线程 15分钟，内存100G

```
time metaspades.py -t 6 -m 100 \
```

```
`tail -n+2 result/metadata.txt|cut -f 1|sed 's/^/temp\qcV/;s/$/_1.fastq/'|sed 's/^/-1 /'|tr '\n' ' ' \
```

```
`tail -n+2 result/metadata.txt|cut -f 1|sed 's/^/temp\qcV/;s/$/_2.fastq/'|sed 's/^/-2 /'|tr '\n' ' ' \
```

```
-o temp/metaspades
```

# t控制线程，m控制内存上限，反引号(`)使用shell命令基于元数据获得输入文件

```
-1 temp/qc/C1_1.fastq -1 temp/qc/C2_1.fastq .....
```

# 23M，contigs体积更大，megahit仅为8.3M

```
ls -sh temp/metaspades/contigs.fasta
```

90G土壤样本，2T内存，1个月没完成。相同数据量，不同数据复杂度消耗时间可差数十至数百倍。



# Metaspades二、三代混合组装(提高片段长度)

- 以Illumina和Nanopore数据为例
- # 3G数据，耗时3h

i=SampleA

```
time metaspades.py -t 48 -m 500 \  
  -1 seq/${i}_1.fastq -2 seq/${i}_2.fastq \  
  --nanopore seq/${i}.fastq \  
  -o temp/metaspades_${i}
```

易生信 生信宝典 宏基因组



# OPERA-MS二、三代混合拼接

- OPERA-MS是发表于Nature Biotechnology的专业二、三代混合组装工具，基于对短读长megahit/metaspades的组装结果，再进行组装以提高片段长度。

```
perl ../OPERA-MS.pl \  
  --short-read1 R1.fastq.gz \  
  --short-read2 R2.fastq.gz \  
  --long-read long_read.fastq \  
  --no-ref-clustering --num-processors 24 \  
  --out-dir RESULTS
```

易生信  
宏基因组





# OPERA-MS二代组装+三代优化

- 结果卡在第9步polishing，可添加--no-polishing参数跳过此步；短序列只支持成对文件，多个文件需要cat合并

```
perl ~/soft/OPERA-MS/OPERA-MS.pl \  
  --contig-file temp/megahit/final.contigs.fa \  
  --short-read1 R1.fastq.gz \  
  --short-read2 R2.fastq.gz \  
  --long-read long_read.fastq \  
  --num-processors 32 \  
  --no-ref-clustering \  
  --no-strain-clustering \  
  --no-polishing \  
  --out-dir temp/opera
```

易生信 毕生缘 宏基因组



## 3.1.3 QUAST评估

### QUAST: quality assessment tool for genome assemblies

[A Gurevich](#), [V Saveliev](#), [N Vyahhi](#), [G Tesler](#) - Bioinformatics, 2013 - [academic.oup.com](#)

Limitations of genome sequencing techniques have led to dozens of assembly algorithms, none of which is perfect. A number of methods for comparing assemblers have been developed, but none is yet a recognized benchmark. Further, most existing methods for comparing assemblies are only applicable to new assemblies of finished genomes; the problem of evaluating assemblies of previously unsequenced species has not been adequately considered. Here, we present QUAST—a quality assessment tool for evaluating ...

☆ 77 Cited by 3049 Related articles All 18 versions

`quast.py -h` # 显示帮助，评估单个组装结果，生成网页报告

`quast.py temp/megahit/final.contigs.fa -o result/megahit/quast`

# 评估多种组装结果

`quast.py --label "megahit,metapasdes" temp/megahit/final.contigs.fa \`

`temp/metaspades/contigs.fasta -o temp/quast`

- ☐ basic\_stats
- ☐ icarus.html
- ☐ icarus\_viewers
- ☐ quast.log
- ☐ report.html
- ☐ report.pdf
- ☐ report.tex
- ☐ report.tsv
- ☐ report.txt
- ☐ transposed\_report.tex
- ☐ transposed\_report.tsv
- ☐ transposed\_report.txt



# 评估结果: megahit vs metaspades

Worst Median Best

☒ Show heatmap

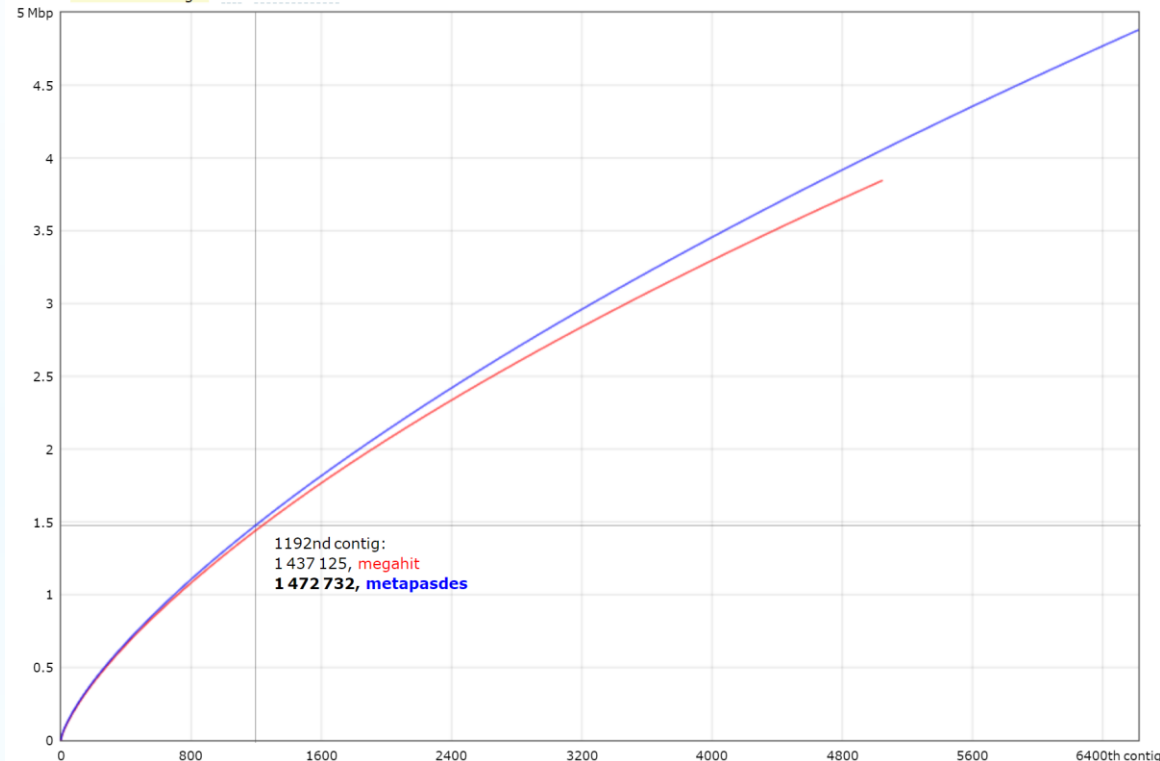
## Statistics without reference

	megahit	metaspades
# contigs	5048	6623
# contigs ( $\geq 0$ bp)	15 862	63 700
# contigs ( $\geq 1000$ bp)	724	800
# contigs ( $\geq 5000$ bp)	2	3
# contigs ( $\geq 10000$ bp)	0	1
# contigs ( $\geq 25000$ bp)	0	0
# contigs ( $\geq 50000$ bp)	0	0
Total length	3 846 110	4 879 999
Total length ( $\geq 0$ bp)	7 921 207	21 026 917
Total length ( $\geq 1000$ bp)	1 006 830	1 105 965
Total length ( $\geq 5000$ bp)	11 400	23 293
Total length ( $\geq 10000$ bp)	0	11 863
Total length ( $\geq 25000$ bp)	0	0
Total length ( $\geq 50000$ bp)	0	0
N50	736	707
N75	592	581
L50	1807	2429
L75	3277	4349
GC (%)	41.73	41.93

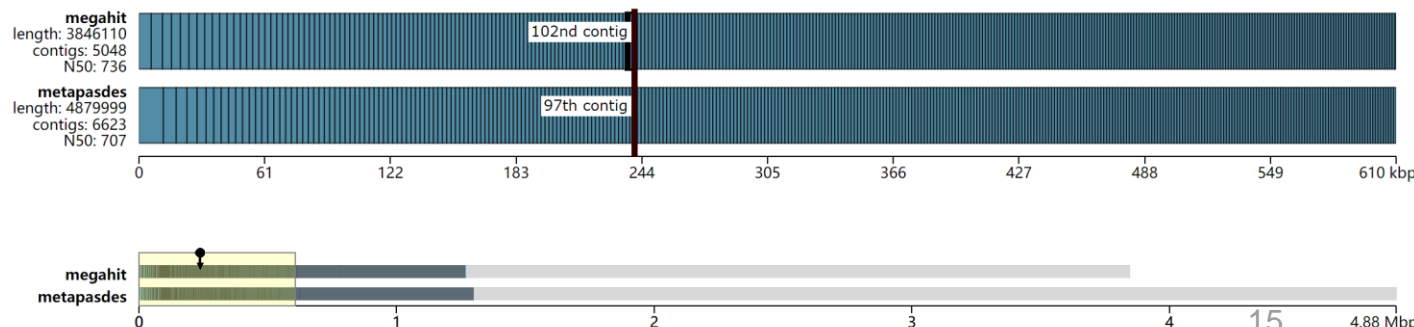
## Mismatches

# N's	0	0
# N's per 100 kbp	0	0

Plots: Cumulative length Nx GC content



Contig size viewer. For better performance, only largest 1000 contigs of each assembly were loaded



# (可选) MetaQUAST评估基因组完整度

# 依赖数据库更全面评估，下载SILVA数据库确定细菌种类；然后在NCBI下载最高丰度的50个株的基因组，分析覆盖度(数据下载受网络限制，可能需很久，我测试下载极慢)

```
metaquast.py result/megahit/final.contigs.fa -o result/megahit/metaquast
```

## MetaQUAST: evaluation of metagenome assemblies

[A Mikheenko](#), [V Saveliev](#), [A Gurevich](#) - Bioinformatics, 2016 - [academic.oup.com](#)

During the past years we have witnessed the rapid development of new metagenome assembly methods. Although there are many benchmark utilities designed for single-genome assemblies, there is no well-recognized evaluation and comparison tool for metagenomic-specific analogues. In this article, we present MetaQUAST, a modification of QUAST, the state-of-the-art tool for genome assembly evaluation based on alignment of contigs to a reference. MetaQUAST addresses such metagenome datasets features as (i) ...

☆ 77 Cited by 205 Related articles All 13 versions

[MetaQuast: 评估宏基因组拼接](#)



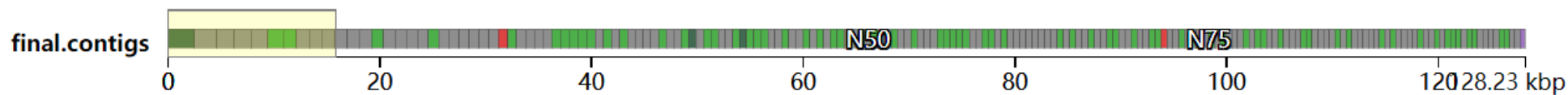
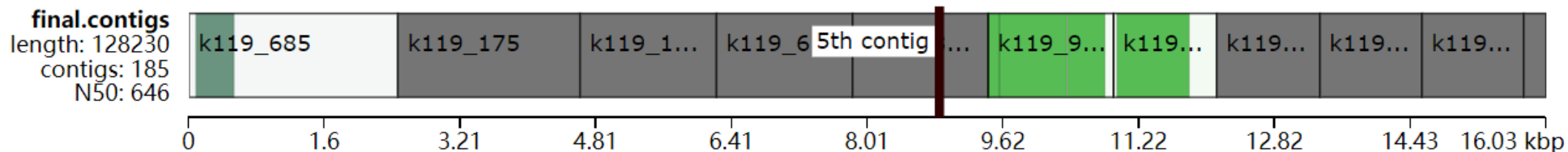


# MetaQuast结果：评估错误组装、错配和插入缺失

## Contig size viewer

结果见：[result/megahit/metaquast/report.html](http://result/megahit/metaquast/report.html)

[View in Icarus contig browser](#) — Contig size viewer

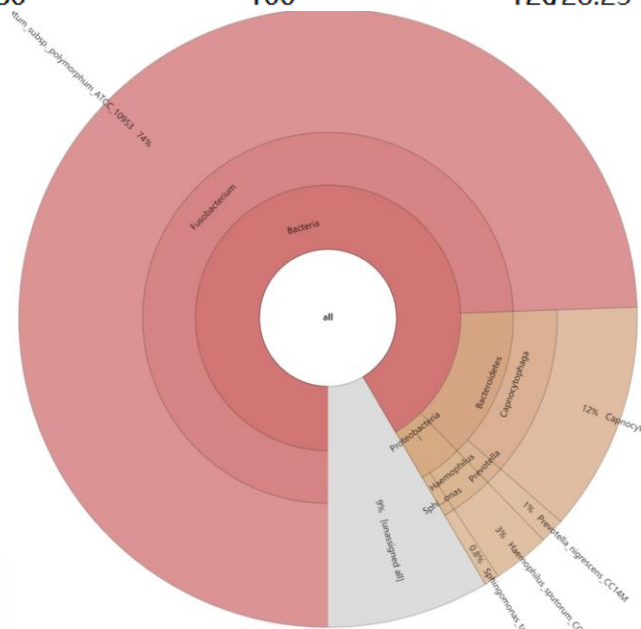


## Misassemblies

# misassemblies	2
Capnocytophaga_ochracea_str._Holt_25	0
Capnocytophaga_sputigena_ATCC_33612	0
Fusobacterium_nucleatum_subsp._polymorphum_ATCC_10953	2
Haemophilus_sputorum_CCUG_13788	0
Prevotella_nigrescens_CC14M	0
Sphingomonas_taxi	0
Misassembled contigs length	1426

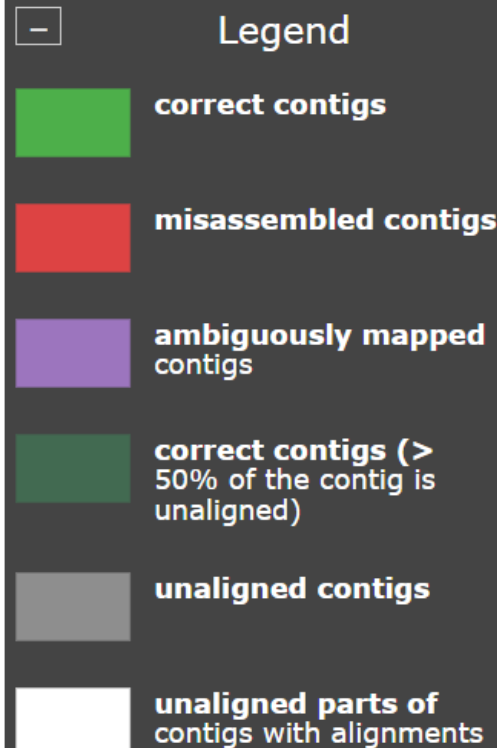
## Mismatches

# mismatches per 100 kbp	2715.06
# indels per 100 kbp	153.15
# N's per 100 kbp	0



## Contig info

<click on a contig to get details>



评估错误组装、错配和插入缺失



- MEGAHIT快速组装，适合30G~300G范围多样本混合组装，节省计算和内存资源；默认按95%相似度种水平聚类，无法拼接株水平序列。
- metaSPAdes精细组装，但内存和时间消耗极大，适合单样本分别组装，可以拼接株水平重叠群，30G组装需上百线程1周，90G无法完成；
- 拼接长度和错误率也成正比，N50提高也伴随时嵌合体升高风险；
- 二、三代测序数据混合组装，首选metaSPAdes安装方便，显著提高片段长度；
- 二、三代测序数据混合组装OPERA-MS无Conda安装麻烦，但速度较快；
- QUAST快速评估常用组装指标，提供html/pdf报告，支持多个组装结果共同评估和比较；
- metaQUAST基于参考数据库进行更细致的评估，但下载成功率不高。



- [宏基因组公众号文章目录](#)      [生信宝典公众号文章目录](#)
- [科学出版社《微生物组数据分析》——50+篇](#)
- [Bio-protocol《微生物组实验手册》——153篇](#)
- [Protein Cell: 扩增子和宏基因组数据分析实用指南](#)
- [CMJ: 人类微生物组研究设计、样本采集和生物信息分析指南](#)
- [加拿大生信网 <https://bioinformatics.ca/> 宏基因组课程中文版](#)
- [美国高通量开源课程 <https://github.com/ngs-docs>](#)
- [Curtis Huttenhower <http://huttenhower.sph.harvard.edu/>](#)
- [Nicola Segata <http://segatalab.cibio.unitn.it/>](#)





扫码关注生信宝典，学习更多生信知识



扫码关注宏基因组，获取专业学习资料

# 易生信，没有难学的生信知识

