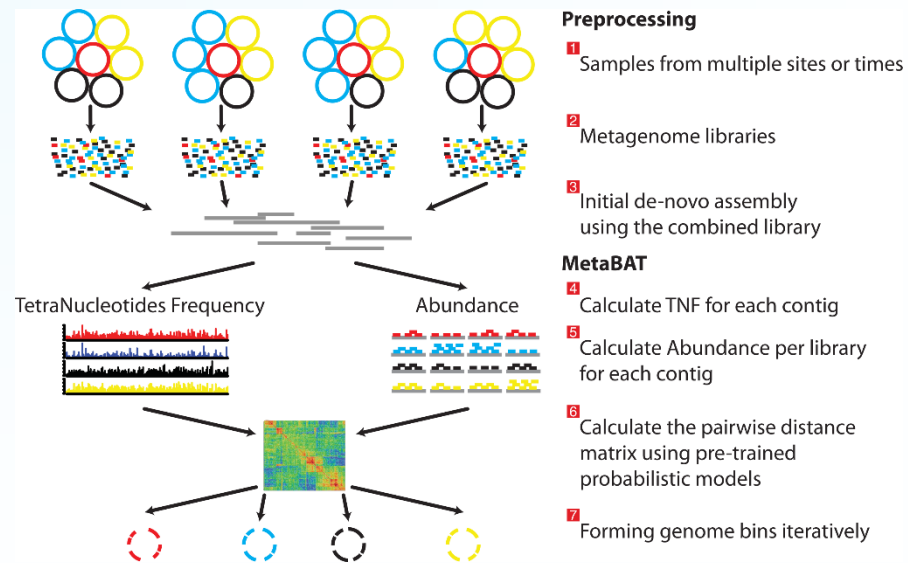


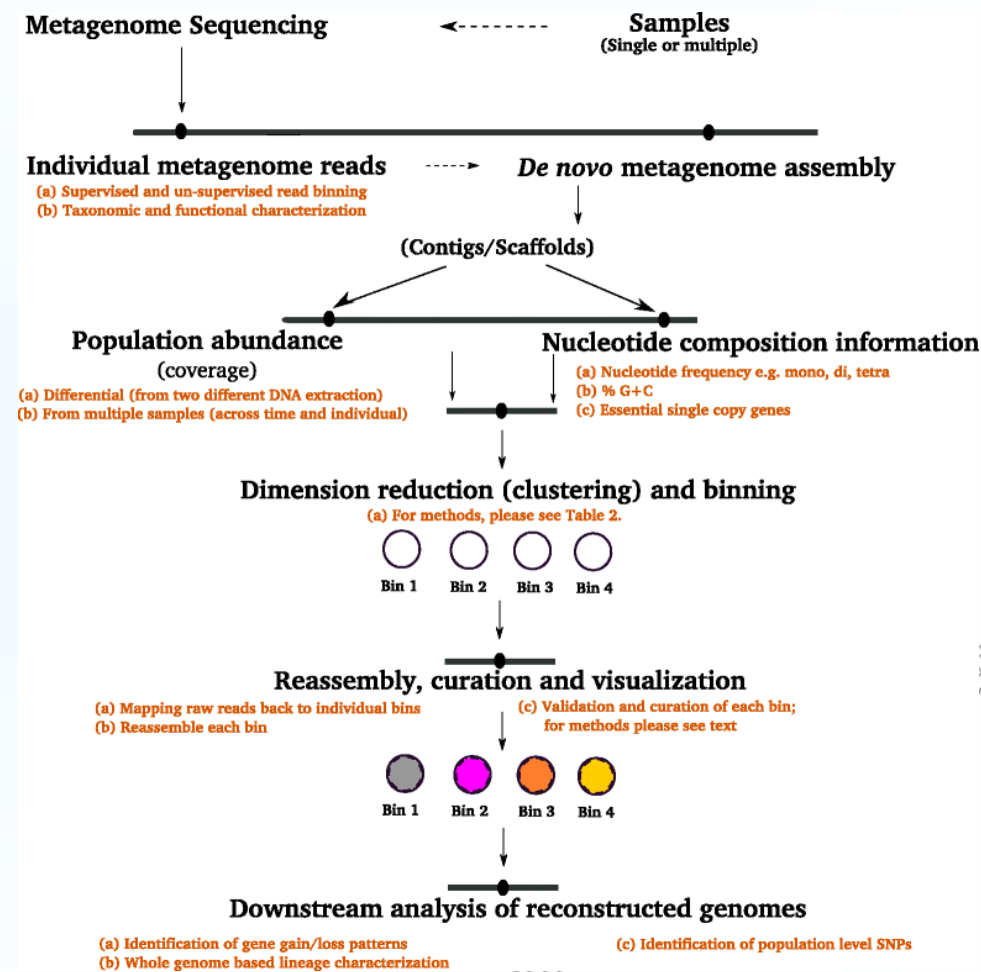


34分箱Binning

易生信
2019年6月23日



- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)



Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4, 8, doi:10.1186/s40168-016-0154-5 (2016).

- 一. **Binning简介**
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)

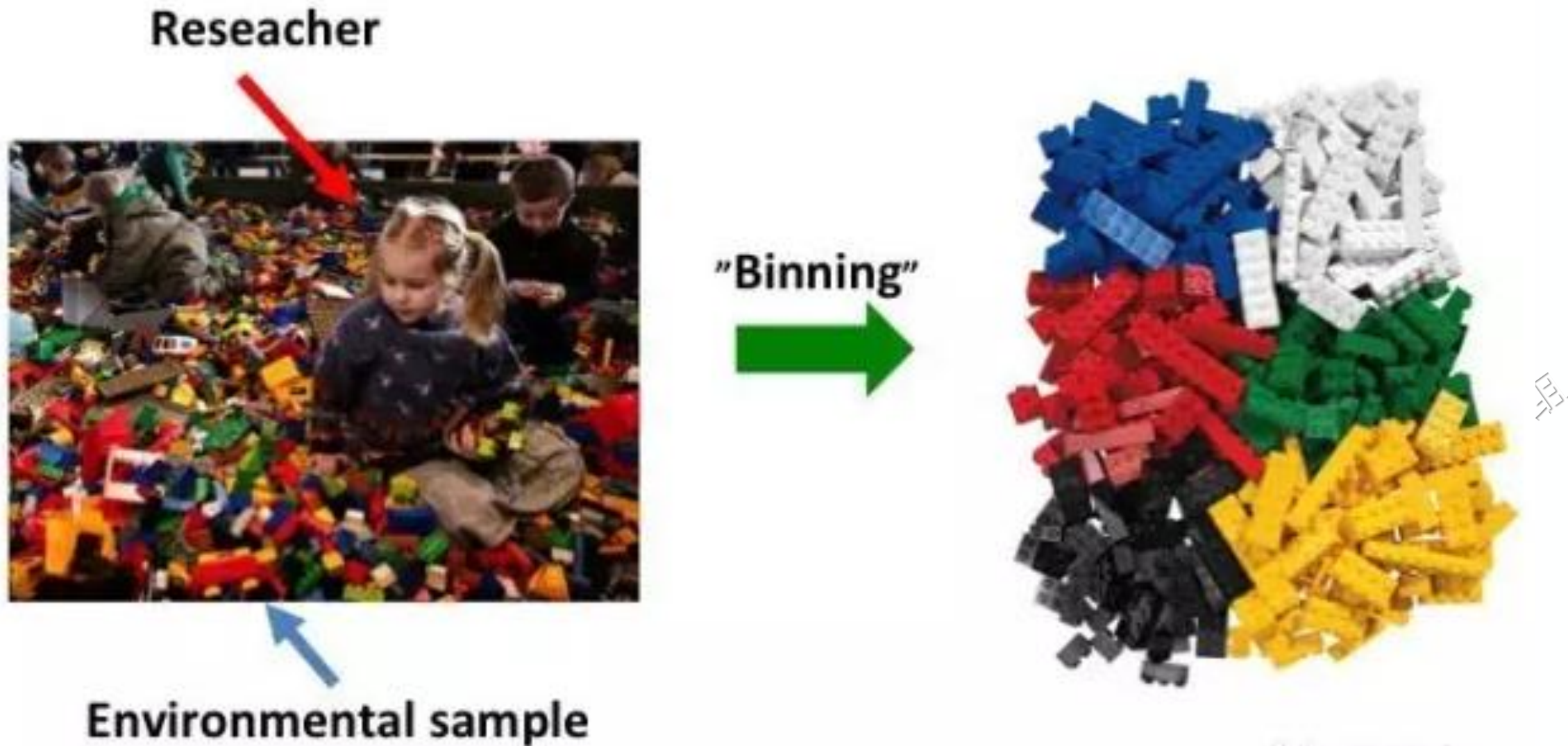


什么是分箱(Binning)?

- 宏基因组研究中，你想不想知道哪些序列来自哪些菌，想不想拼出未培养菌或未知菌的基因组呢？
- 其实这些可以有，很多高水平文章中都有，这个过程就叫Binning(我习惯将其翻译为分箱)，是宏基因组分析提升档次的关键步骤。
- 想了解Bining的背景知识，请阅读 [《一文读懂宏基因组binning》](#)。想了解哪些软件可以Binning，请看 [《精选30余款宏基因组分析软件》](#)。想知道更全面的 Bin 软件及评估，可以阅读 Nature Method(<https://www.nature.com/articles/nmeth.4458>)，或阅读中文导读 [《Nature Method: 史上最权威宏基因组软件评估—人工重组宏基因组基准数据集》](#)，其中有9款Bin软件的简介和比较。



Binning的原理



Binning有两方面的重要应用

○ 关联分析

即通过binning得到的bins (strain-level clusters / strain-level taxonomic units) 可以进行宏基因组关联分析以及多组学联合分析，将特定功能代谢产物与特定物种、特定基因进行关联研究，推动其因果机制的探究，为疾病监控、环境监测提供了菌株水平的生物靶标。

[Nature综述: 宏基因组关联分析-深入研究微生物组\(王俊&贾慧钰\)](#)

[Nature Protocols: 整合宏基因组、代谢组和表型分析的的计算框架](#)

○ 单菌组装

通过对binning得到的bins进行后续组装，可以得到很多不能在实验室里培养的细菌、古菌或病毒的基因组草图，然后根据单菌组装结果进行菌株水平的基因和功能注释、比较基因组分析、进化分析等，使我们得以洞察菌株的生态适应、营养互作和新陈代谢等，研究在生态环境和复杂疾病中作用的菌种以及致病菌和宿主的互作机制及其微进化机制。

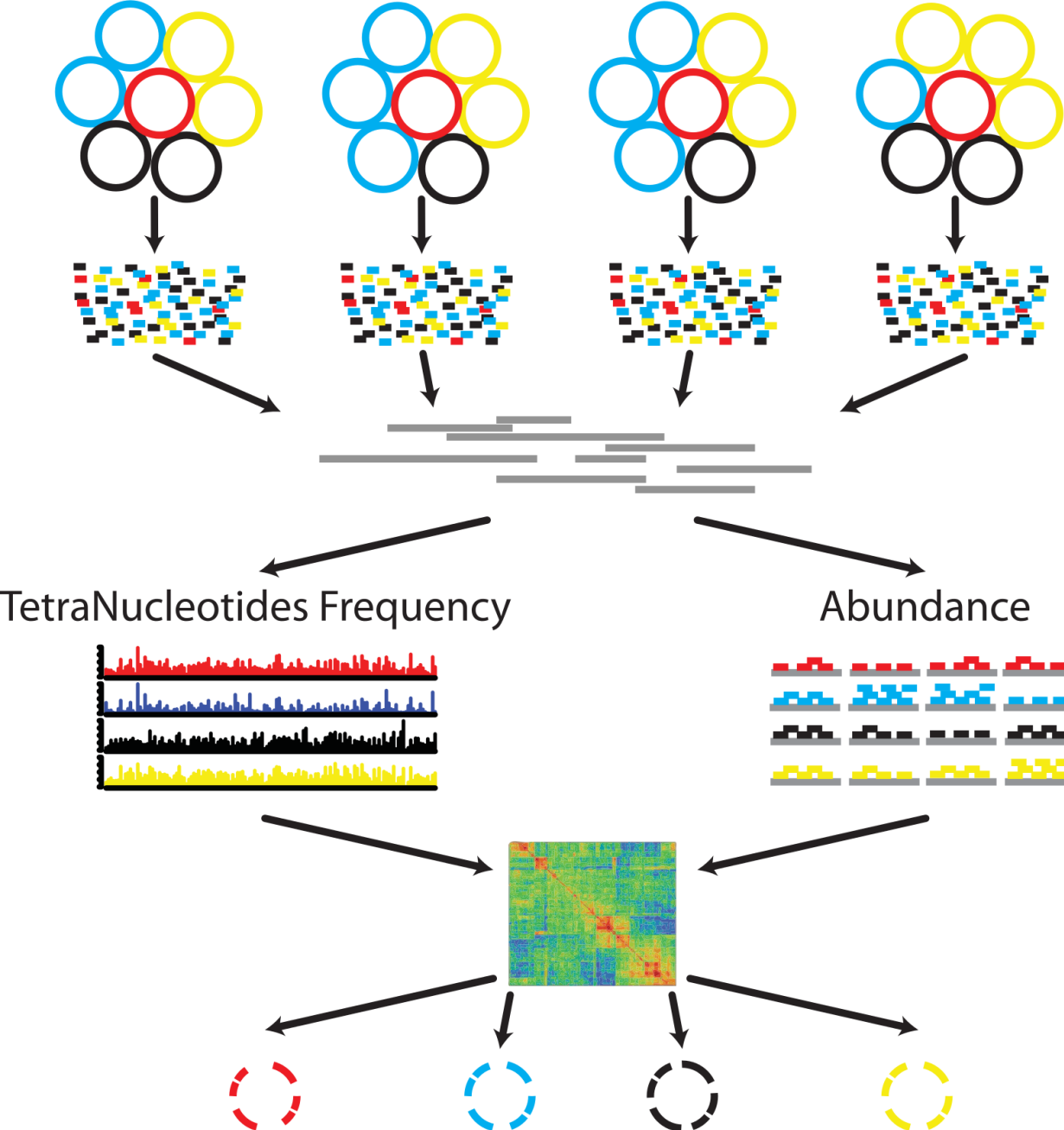
[Nature Communications: 宏基因组学提示曙古菌门的代谢和进化](#)

Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology* **14**, 508, doi:10.1038/nrmicro.2016.83 (2016).

Pedersen, H. K. *et al.* A computational framework to integrate high-throughput ‘-omics’ datasets for the identification of potential mechanistic links. *Nature Protocols* **13**, 2781-2800, doi:10.1038/s41596-018-0064-z (2018).

Hua, Z.-S. *et al.* Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. *Nature Communications* **9**, 2832, doi:10.1038/s41467-018-05284-4 (2018).





Preprocessing

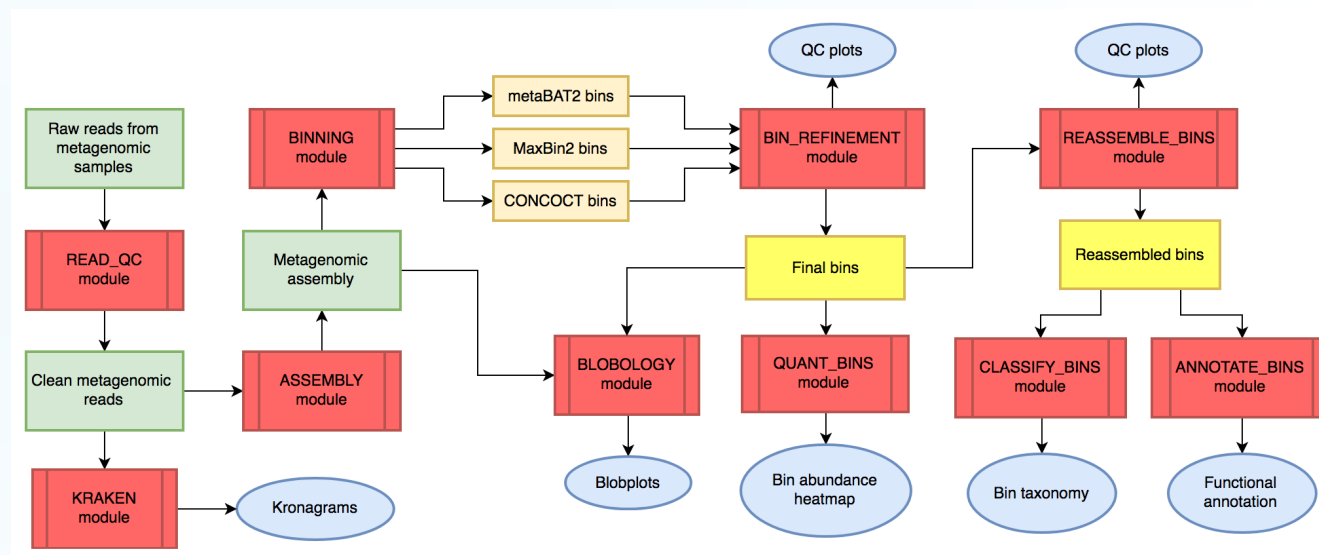
- 1 Samples from multiple sites or times
- 2 Metagenome libraries
- 3 Initial de-novo assembly using the combined library

MetaBAT

- 4 Calculate TNF for each contig
- 5 Calculate Abundance per library for each contig
- 6 Calculate the pairwise distance matrix using pre-trained probabilistic models
- 7 Forming genome bins iteratively

1. 不同时间地点的样品
2. 宏基因组测序
3. 组装为重叠群
4. 计算重叠群4核苷酸频率
5. 计算样本中重叠群丰度
6. 计算成对距离矩阵
7. 迭代获得分箱

- 一. Binning简介
- 二. **MetaWRAP**流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)



易生信


SOFTWARE

Open Access



MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis

Gherman V. Uritskiy, Jocelyne DiRuggiero*  and James Taylor*

[\[HTML\] MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis](#) 

[GV Uritskiy, J DiRuggiero, J Taylor - Microbiome, 2018 - microbiomejournal.biomedcentral ...](#)

Abstract

Background: The study of microbiomes using whole-metagenome shotgun sequencing enables the analysis of uncultivated microbial populations that may have important roles in their environments. Extracting individual draft genomes (bins) facilitates metagenomic analysis at the single genome level. Software and pipelines for such analysis have become diverse and sophisticated, resulting in a significant burden for biologists to access and use them. Furthermore, while bin extraction algorithms are rapidly improving, there is still a lack of tools ...

★  Cited by 9 Related articles All 11 versions 

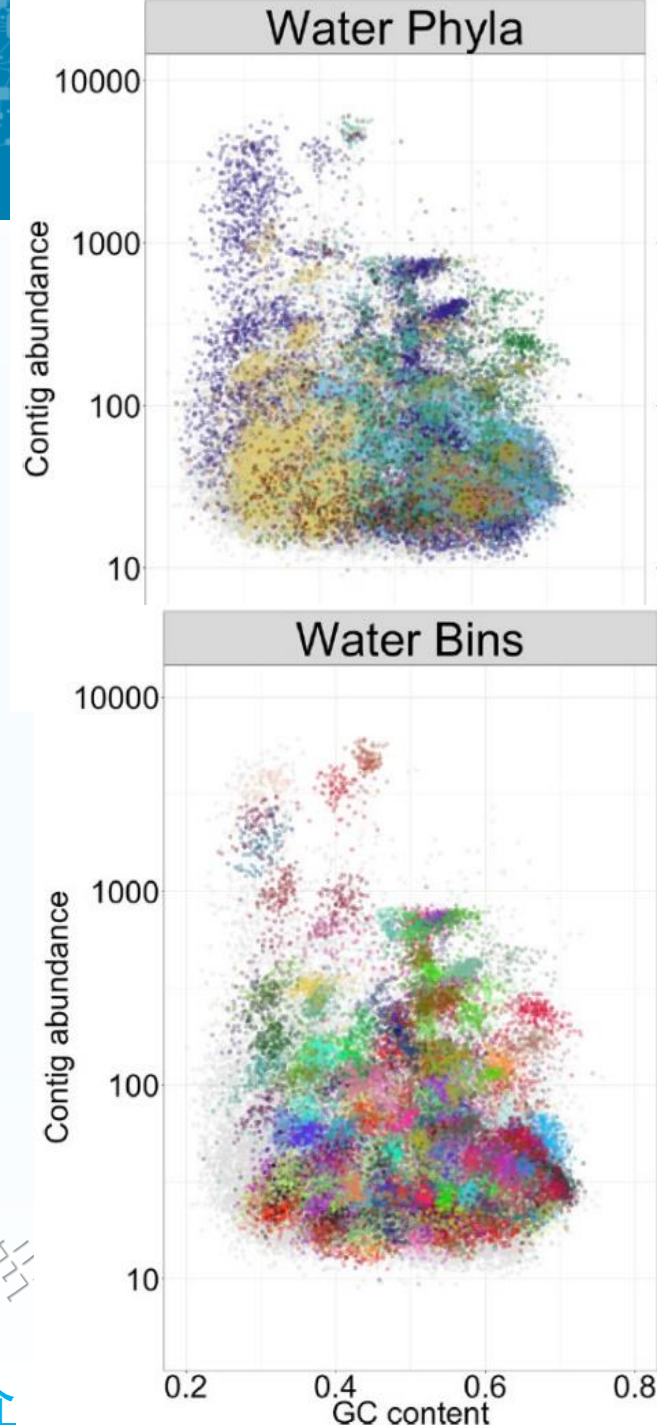
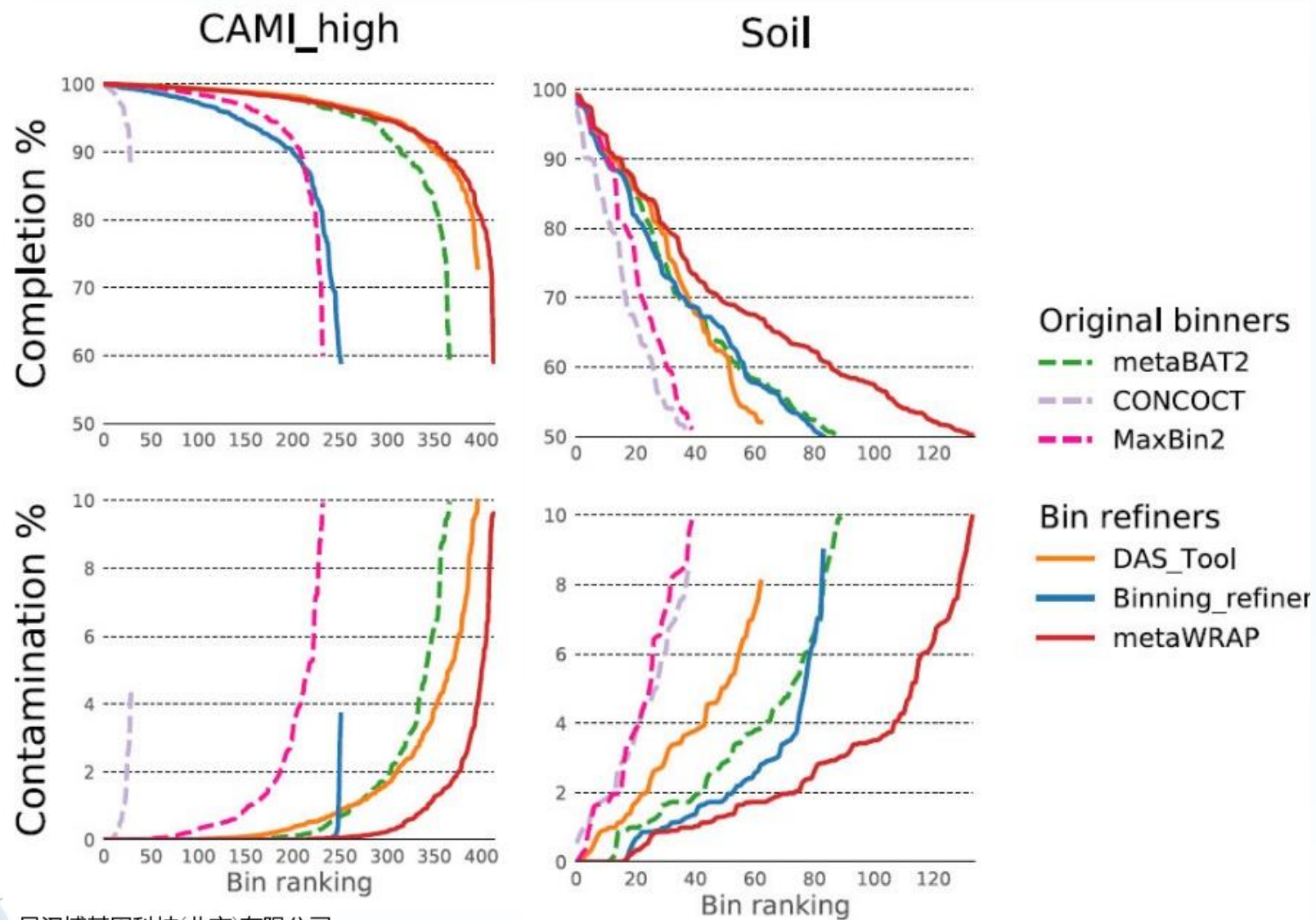


灵活的宏基因组数据挖掘单菌基因组分析流程

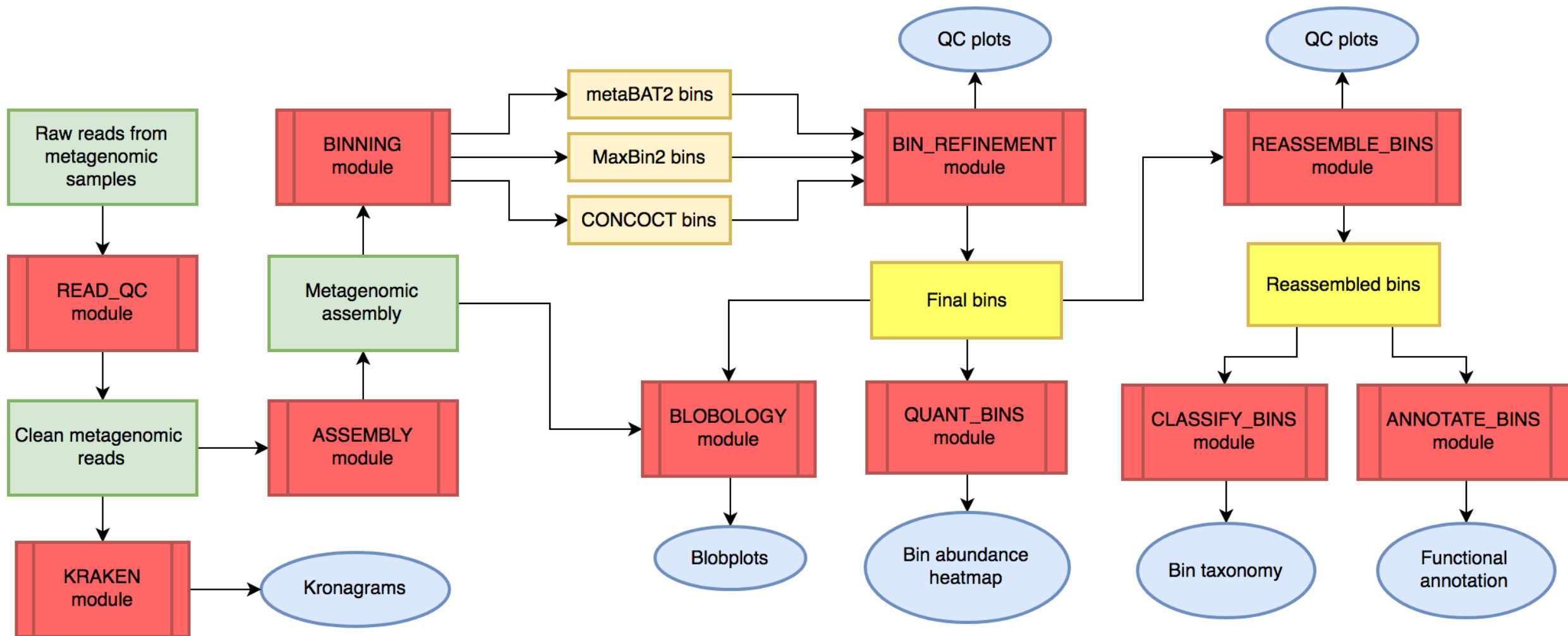
- ① MetaWRAP是一款整合了质控、拼接、分箱、提纯、评估、物种注释、丰度估计、功能注释和可视化的分析流程，纳入超140个工具软件，可一键安装；
- ② 流程整合了CONCOCT、MaxBin、 metaBAT等三款分箱工具以及提纯和重组装算法；
- ③ 与以上三种分析工具单独使用，以及与使用分箱提纯工具DAS_tool、Binning-refiner相比，分箱结果更佳；
- ④ MetaWRAP还可实现宏基因组分析从原始数据到结果可视化的全部流程，同时也可灵活使用各个模块独立分析，弹性多变。



MetaWRAP评估结果



分析流程——全流程+可视化



MetaWRAP的功能模块

宏基因组数据预处理模块

- 1) 质控Read_QC: read质控剪切和移除人类宿主
- 2) 组装Assembly: 质控、使用megahit或metaSPAdes拼接
- 3) 物种注释Kraken: 对reads和contigs层面进行可视化

分箱Bin处理模块

- 1) 分箱Binning: 利用MaxBin2, metaBAT2, 和CONCOCT三个软件分别分箱;
- 2) 提纯Bin_refinement: 对多种Bin结果评估和综合分析, 获得更好的结果;
- 3) 重组装Reassemble_bins: 利用原始序列和评估软件二次组装, 改善Bin的N50、完整度
- 4) 定量Quant_bins: 估计样品中每个bin的丰度并热图展示
- 5) 气泡图Blobology: blobplots可视化群体的contigs的物种和Bin分布
- 6) 物种注释Classify_bins: 对Bin物种注释
- 7) 基因注释Annotate_bins: 预测Bin中的基因



- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. **MetaWRAP安装**
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)

易生信 生信宝典 宏基因组



- 系统要求是由处理的数据量决定的。
- 其中一些软件，如KRAKEN、metaSPAdes对内存需求较高，推荐服务器配置**48+核，512+GB内存(但不是必须)**，至少服务器**24核，64GB内存**，仅支持64位Linux系统。
- 对于300 GB以上数据用户，推荐配置48核，512GB内存或更高。
- 软件原作者的教程中参数使用了96线程和900G内存，可以推断软件开发和测试所用服务器至少为96线程和1TB内存
- 常见胖结点配置24核64GB内存(1-10万)、48核1T内存(10-30万)、96核2T内存(30-70万)

```
conda create -n metawrap python=2.7 # 创建虚拟环境
conda activate metawrap # 启动虚拟环境，防冲突
conda config --add channels defaults # 添加依赖关系的源
conda config --add channels conda-forge
conda config --add channels bioconda
conda config --add channels ursky
conda install -c ursky metawrap-mg # 安装流程，超140个软件
```



常用数据库大小和功能

Database	Size	Used in module
Checkm	1.7GB	binning, bin_refinement, reassemble_bins
NCBI_nt	111GB	blobology, classify_bins
NCBI_tax	156MB	blobology, classify_bins
Indexed hg38	34GB	read_qc
KRAKEN	233GB	kraken

按官网、或公众号教程安装每个数据库，下载时间由网速决定，1-30天不等。

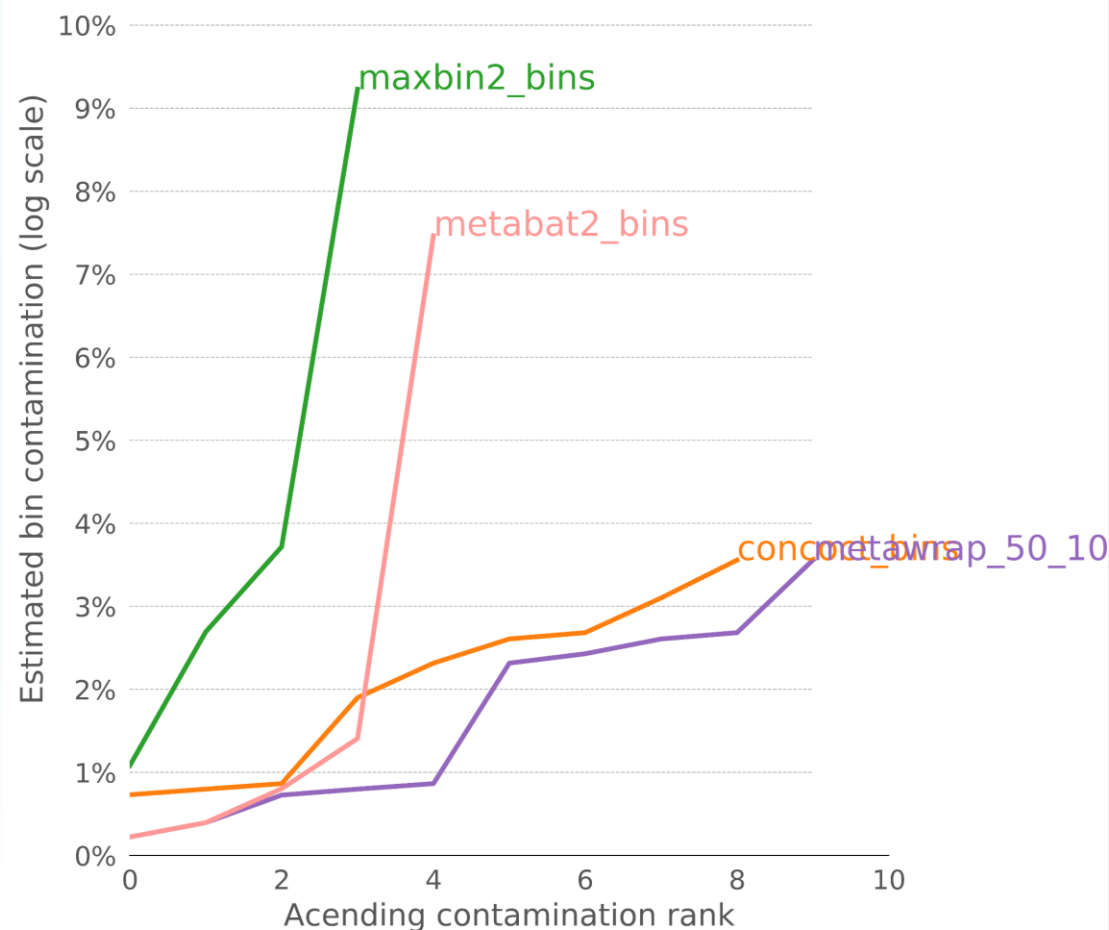
[MetaWRAP安装和数据库部署](https://github.com/bxlab/metaWRAP/blob/master/installation/database_installation.md)

https://github.com/bxlab/metaWRAP/blob/master/installation/database_installation.md



- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. **Binning**组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)

Bin contamination ranking



4.1 运行三种bin软件

输入文件为contig和clean reads

调用三大主流binning程序cococt, maxbin2, metabat2

8线程耗时0.5 - 2 小时

nohup 和 &bg 任务转后台不中断，记录输出内容到 nohup.out(可选)

nohup metawrap binning -o **temp/binning** -t 8 \

-a temp/megahit/**final.contigs.fa** \

--metabat2 --maxbin2 --concoct temp/qc/**ERR*.fastq** &bg

用自己的文件，替换输出文件名为 ***1_kneaddata_paired*.fastq**

输出文件夹 **temp/binning** 包括3种软件结果和中间文件



4.2 Bin提纯

一般要求c完整度70，x污染率5，数据少降低阈值保证有结果可演示

8线程耗时1 - 2 小时

```
metawrap bin_refinement -o temp/bin_refinement -t 8 \
```

```
-A temp/binning/metabat2_bins/ \
```

```
-B temp/binning/maxbin2_bins/ \
```

```
-C temp/binning/concoct_bins/ -c 50 -x 10
```

查看高质量Bin的数量

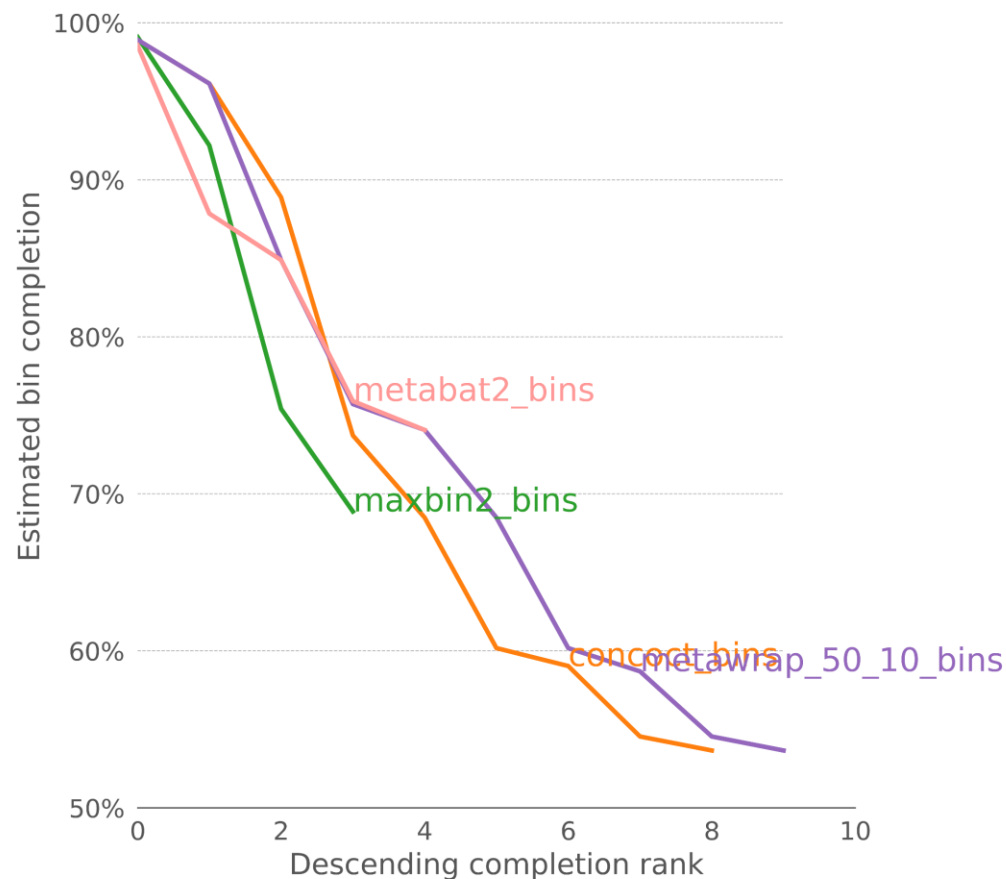
```
cat temp/bin_refinement/metawrap_bins.stats | awk '$2>50 && $3<10' |  
wc -l
```

结果改进程度见temp/bin_refinement/figures/目录

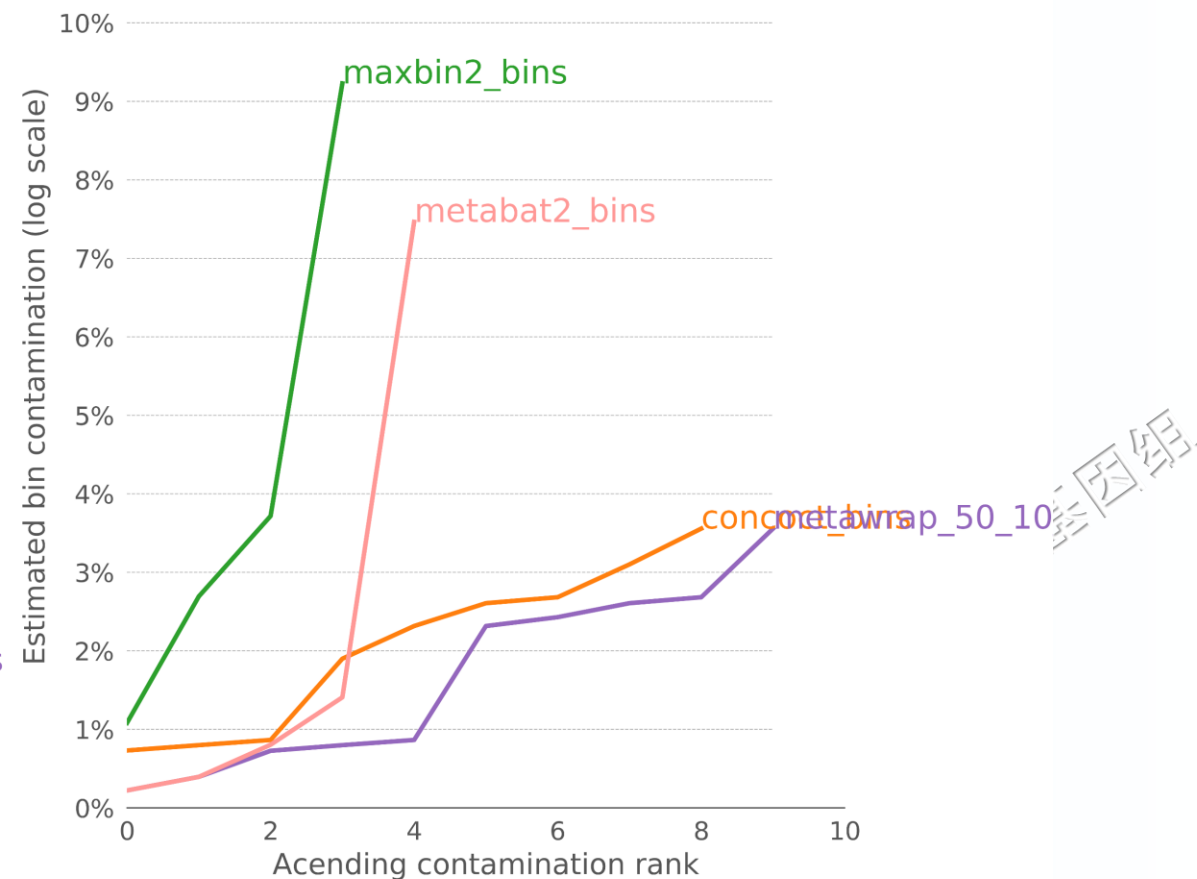


提纯后与前三款软件结果比较

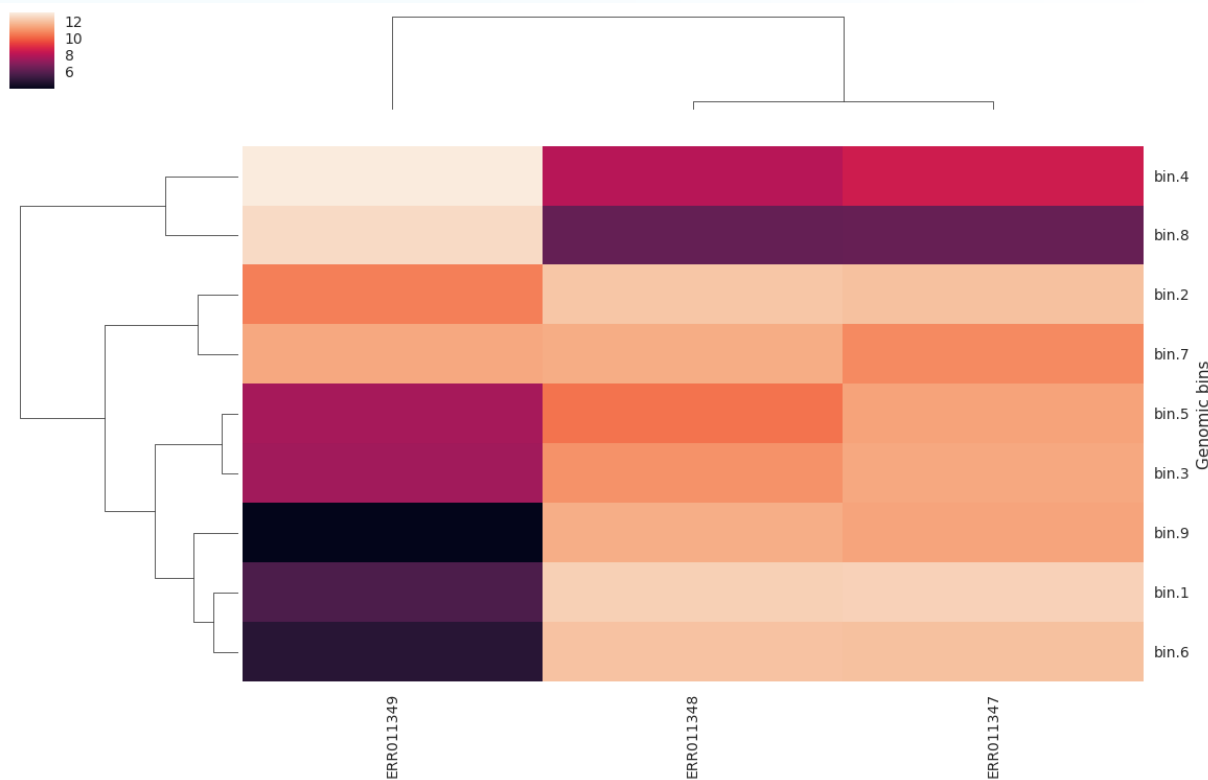
Bin completion ranking



Bin contamination ranking



- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. **Bin定量、物种和基因注释**
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)



4.3 Bin定量

使用salmon计算每个bin在样本中相对丰度

耗时3m，系统用时10m，此处可设置线程，但salmon仍调用全部资源

需要指定输出文件夹，包括4.3中的参数的输出目录

```
nohup metawrap quant_bins -b temp/bin_refinement/metawrap_50_10_bins -t 8 \
```

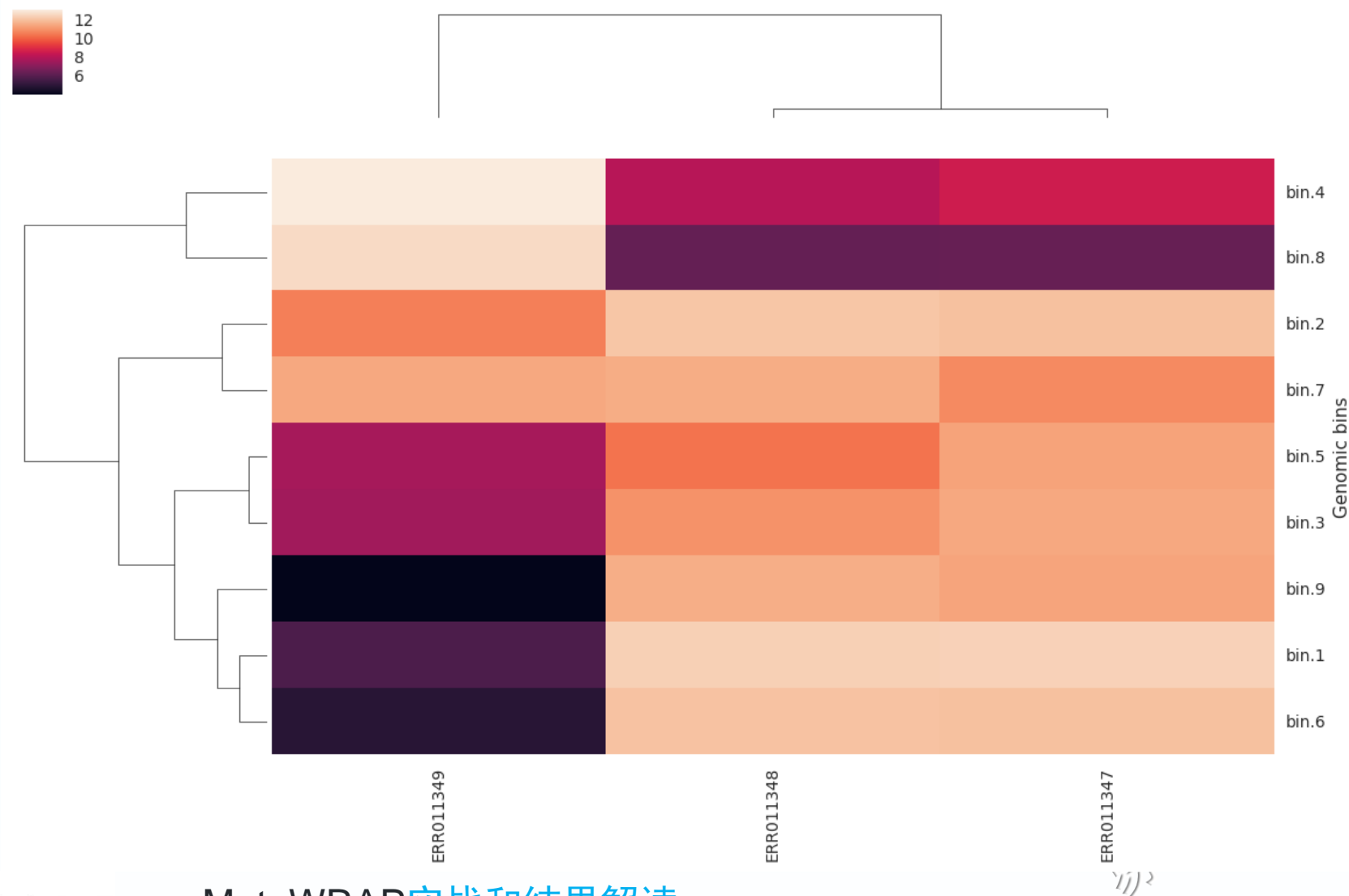
```
-o temp/bin_quant -a temp/megahit/final.contigs.fa temp/qc/ERR*.fastq &bg
```

结果包括bin丰度热图`temp/quant_bins/genome_abundance_heatmap.png`

如果想自己画图，原始数据位于`temp/quant_bins/abundance_table.tab`



Bin定量热图



宏基因组

富集

4.4 Bin注释

Taxator-tk对每条contig物种注释，再估计bin整体的物种，11m

```
nohup metawrap classify_bins -b temp/bin_refinement/metawrap_50_10_bins \  
-o temp/bin_classify -t 8 &
```

注释结果见`temp/classify_bins/bin_taxonomy.tab`

基于prokka基因注释，4m

```
metaWRAP annotate_bins -o temp/bin_annotate -t 8 \  
-b temp/bin_refinement/metawrap_50_10_bins
```

每个 bin 基因注释的 gff 文件 bin_funct_annotations, 核酸 ffn 文件 bin_untranslated_genes, 蛋白 faa 文件 bin_translated_genes



目录

- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. **Krona物种组成可视化(可选)**
- 七. Bin可视化(可选)
- 八. 重组装(可选)



S4.1 物种组成预测(kraken+krona, 可选)

- # kraken物种注释, -o输出目录, -t线程数, -s抽样1M减少计算量, 质控数据, contig文件
- # 需要有256GB以上可用内存, 7G演示数据, 8线程耗时15m, 测试服务器暂不开放, 防多人运行内存耗尽死机
- `metawrap kraken -o temp/kraken -t 8 -s 1000000 \`
- `temp/final.contigs.fa temp/ERR*.fastq`
- # 结果文件夹中有注释结果文件*.kraken(每个reads的注释结果)、*.krona(注释结果分类汇总)
- # 可视化的Krona网页图表kronagram.html





易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. **Bin可视化(可选)**
- 八. 重组装(可选)

易生信 生信宝典 宏基因组



S4.2 Bin可视化

计算每个contig的GC含量和在每个样本中的丰度

8线程耗时1-6h

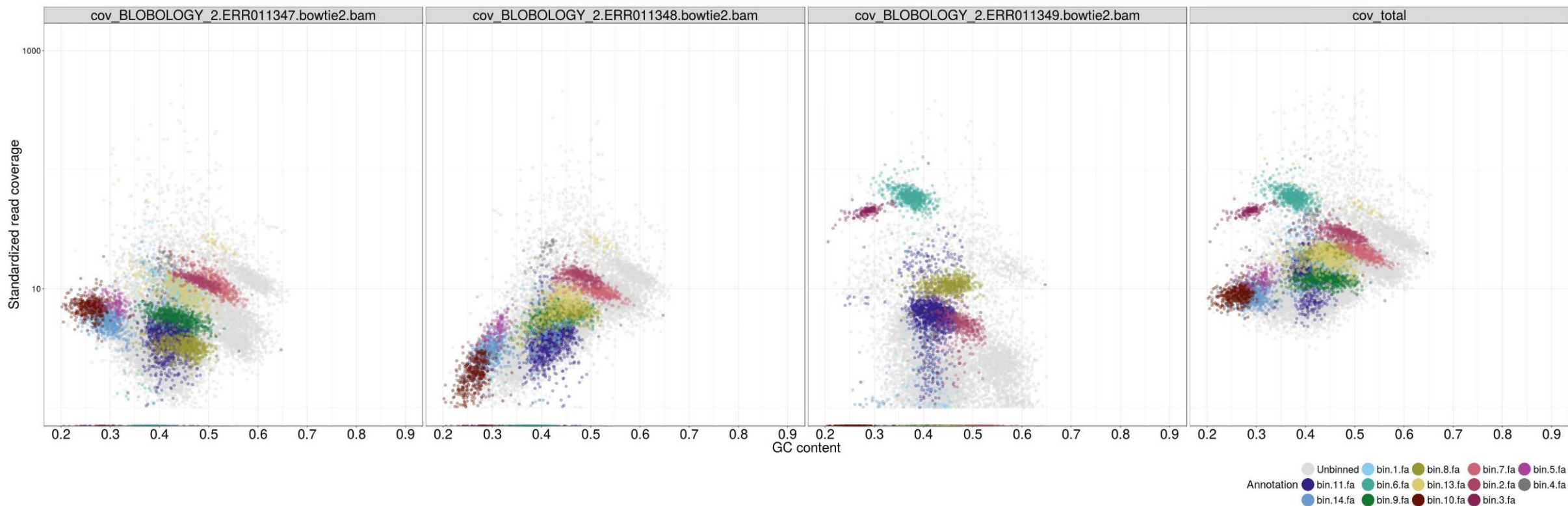
```
metawrap blobology -a temp/megahit/final.contigs.fa -t 8 \  
-o temp/blobology --bins temp/bin_refinement/metawrap_50_10_bins \  
temp/qc/ERR*.fastq
```

结果为final.contigs.binned.blobplot, 方便使用ggplot2可视化

参考脚本 `${soft}/envs/metawrap/bin/metawrap-scripts/blobology/makeblobplot_with_bins.R`(脚本需修改)

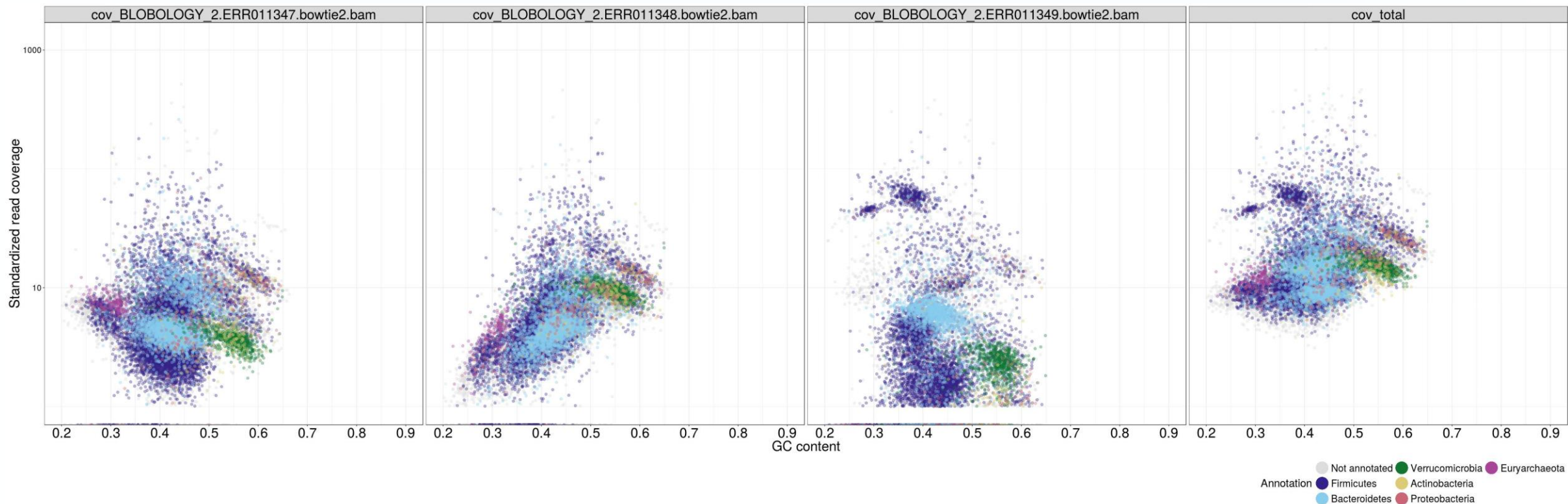


官方示例分箱结果展示blobology图片(我没运行成功)



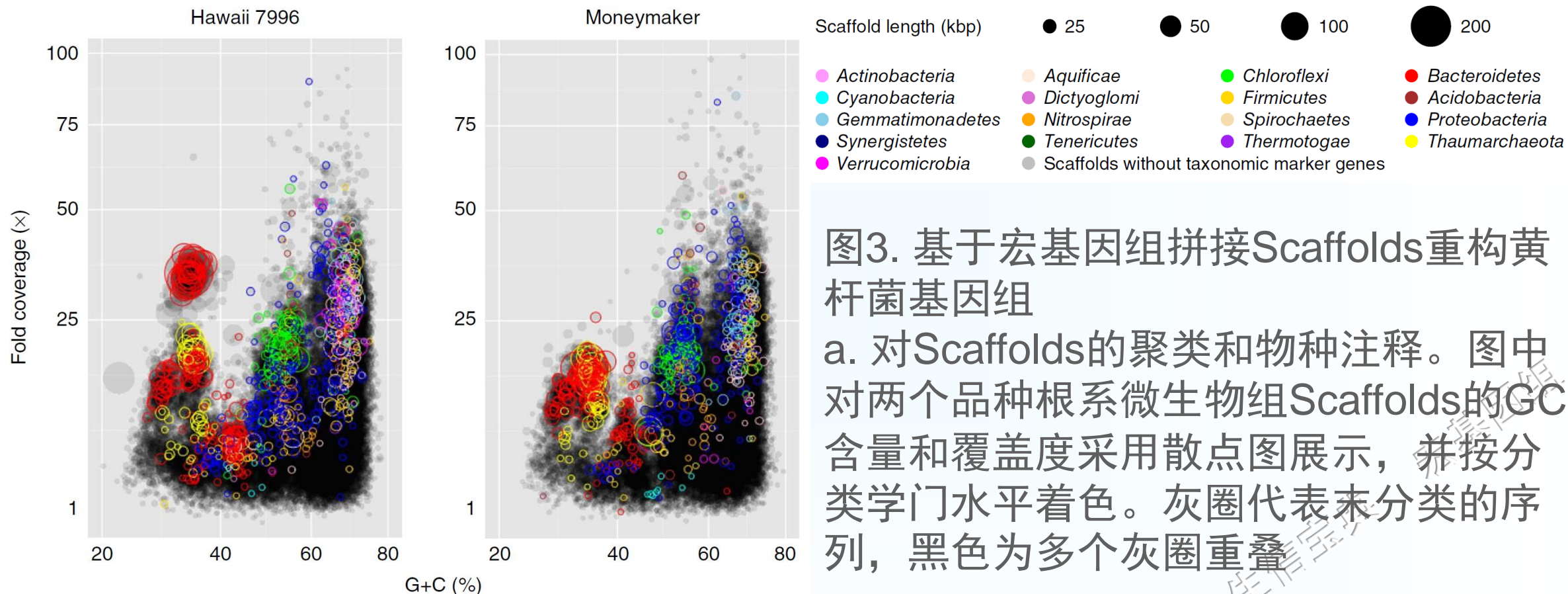
基于Contigs的GC含量和标准化的Reads覆盖深度散点图(blobology图)
四个图分别为3个样品和总体，按Bin编号着色

官方示例分箱结果展示blobology图片(我没运行成功)



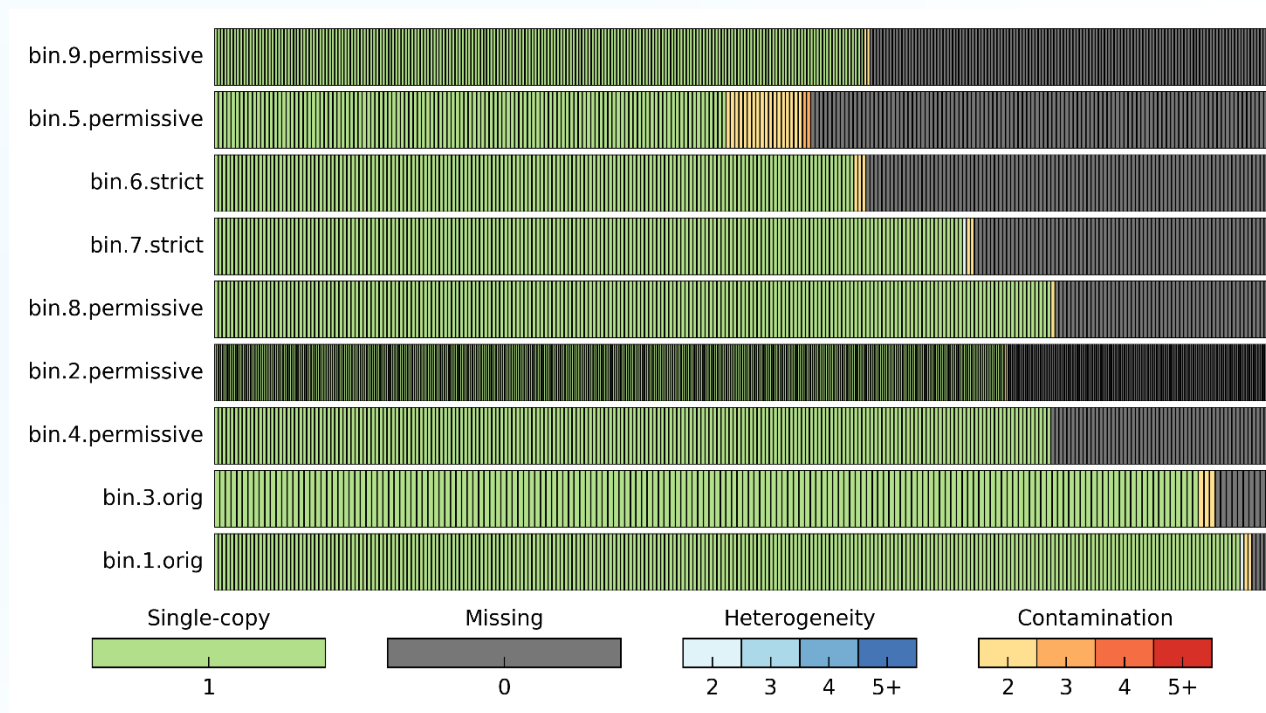
基于Contigs的GC含量和标准化的Reads覆盖深度散点图(blobology图)
四个图分别为3个样品和总体，按分类学门水平着色

NBT-番茄根系宏基因组分箱 *Flavobacteriaceae* 基因组



Kwak, M.-J. *et al.* Rhizosphere microbiome structure alters to enable wilt resistance in tomato. *Nature Biotechnology* **36**, 1100, doi:10.1038/nbt.4232 (2018). [NBT: 根际微生物组抗番茄枯萎病 PPT思路梳理](#)

- 一. Binning简介
- 二. MetaWRAP流程简介
- 三. MetaWRAP安装
- 四. Binning组装和提纯
- 五. Bin定量、物种和基因注释
- 六. Krona物种组成可视化(可选)
- 七. Bin可视化(可选)
- 八. 重组装(可选)



S4.3 重组装

需合并所有样本作为此步输入, 6s

```
cat temp/qc/ERR*_1.fastq > temp/qc/all_1.fq
```

```
cat temp/qc/ERR*_2.fastq > temp/qc/all_2.fq
```

提纯的bin还可以通过再组装进一步改善结果, 8核, 100G内存, 用时2小时

```
nohup metawrap reassemble_bins -o temp/bin_reassemble \
```

```
-1 temp/qc/all_1.fq -2 temp/qc/all_2.fq -t 8 -m 100 \
```

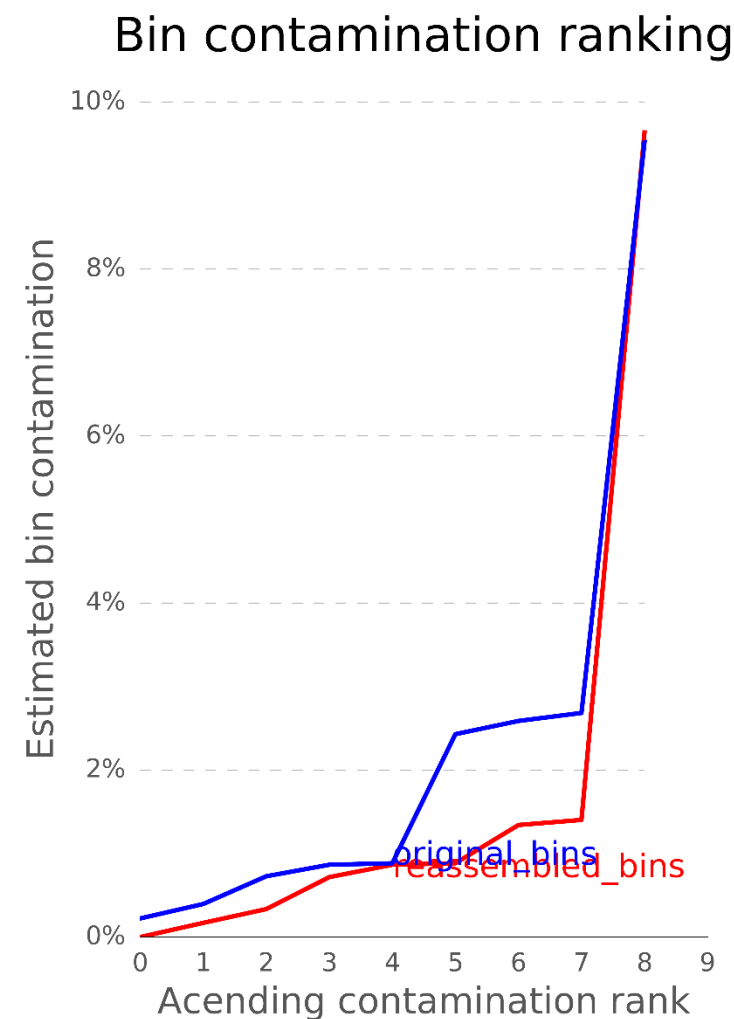
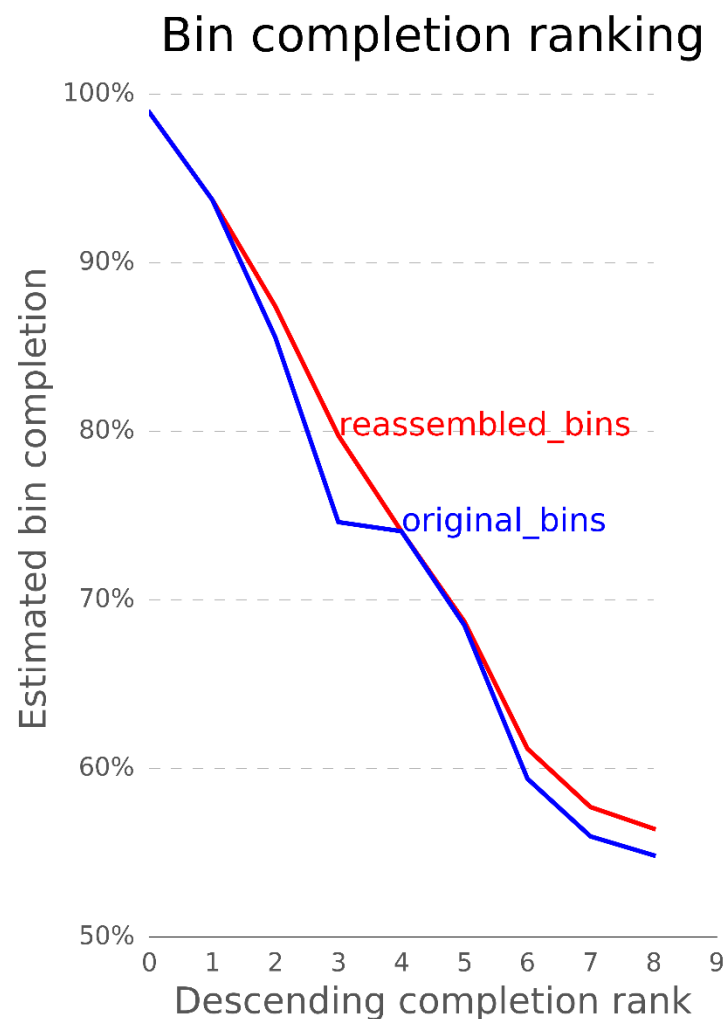
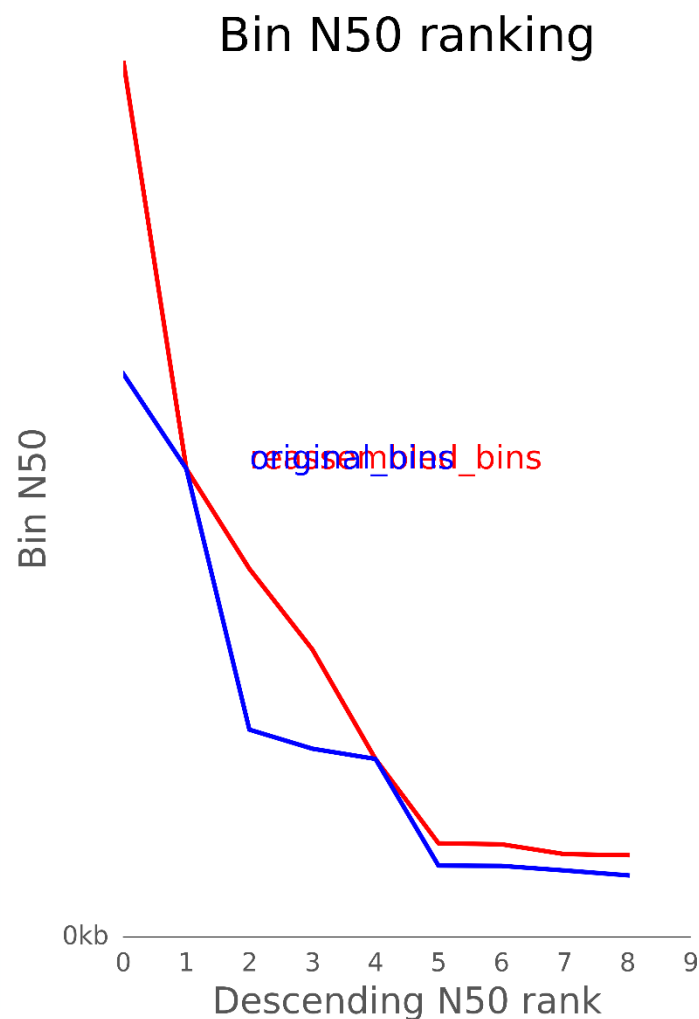
```
-c 50 -x 10 -b temp/bin_refinement/metawrap_50_10_bins &bg
```

结果统计见`temp/bin_reassemble/reassembled_bins.stats`,
`reassembly_results.png`, 比对重组装前后的变化, N50, 这完整度和污染率均有改进。

`reassembled_bins.png`展示CheckM对bin评估结果的可视化。



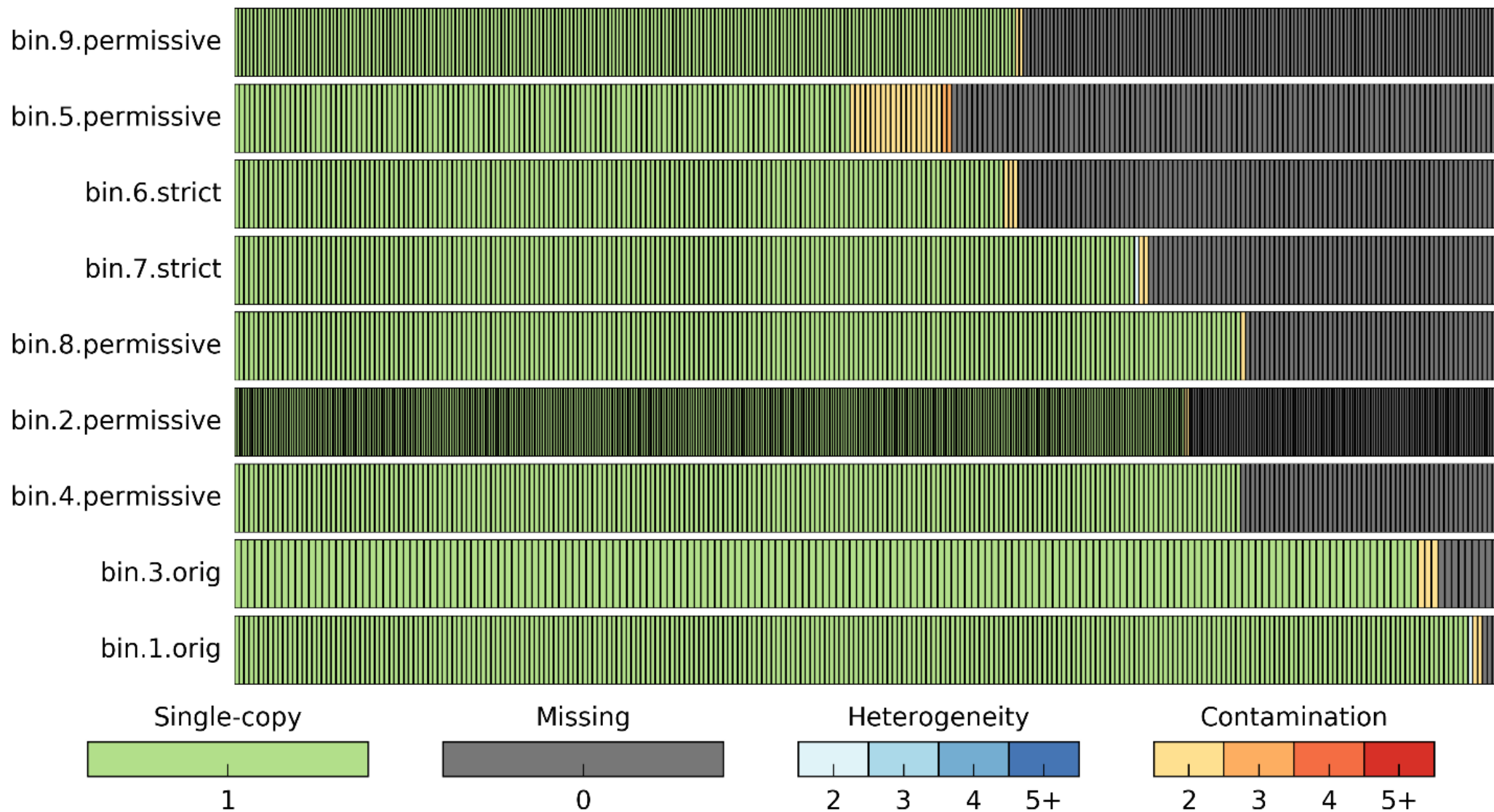
重组装前后对比



temp/reassemble_bins/reassembly_results.png



重组装后评估完整度和污染率可视化

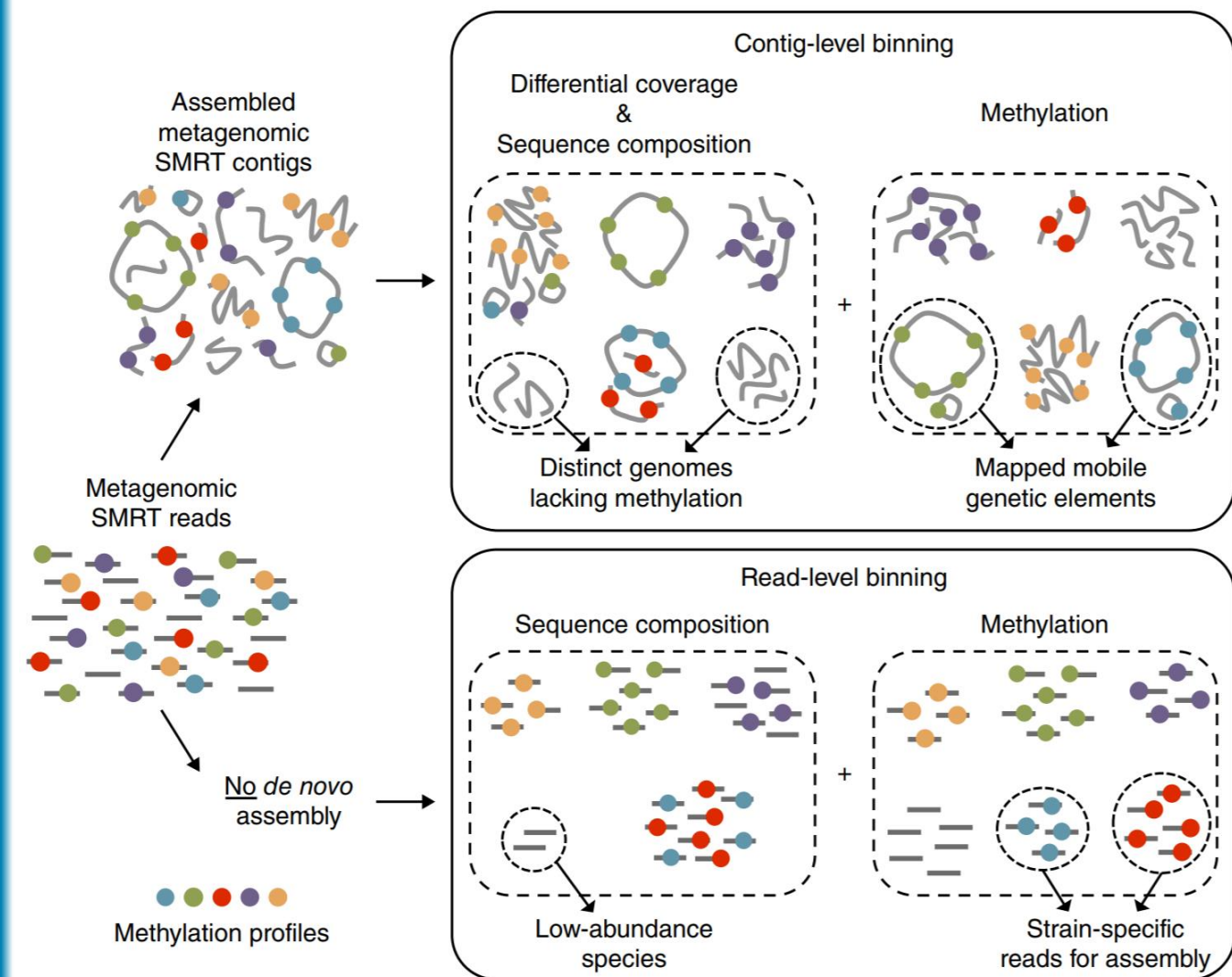


基因组



展望：新技术改进宏基因组Binning

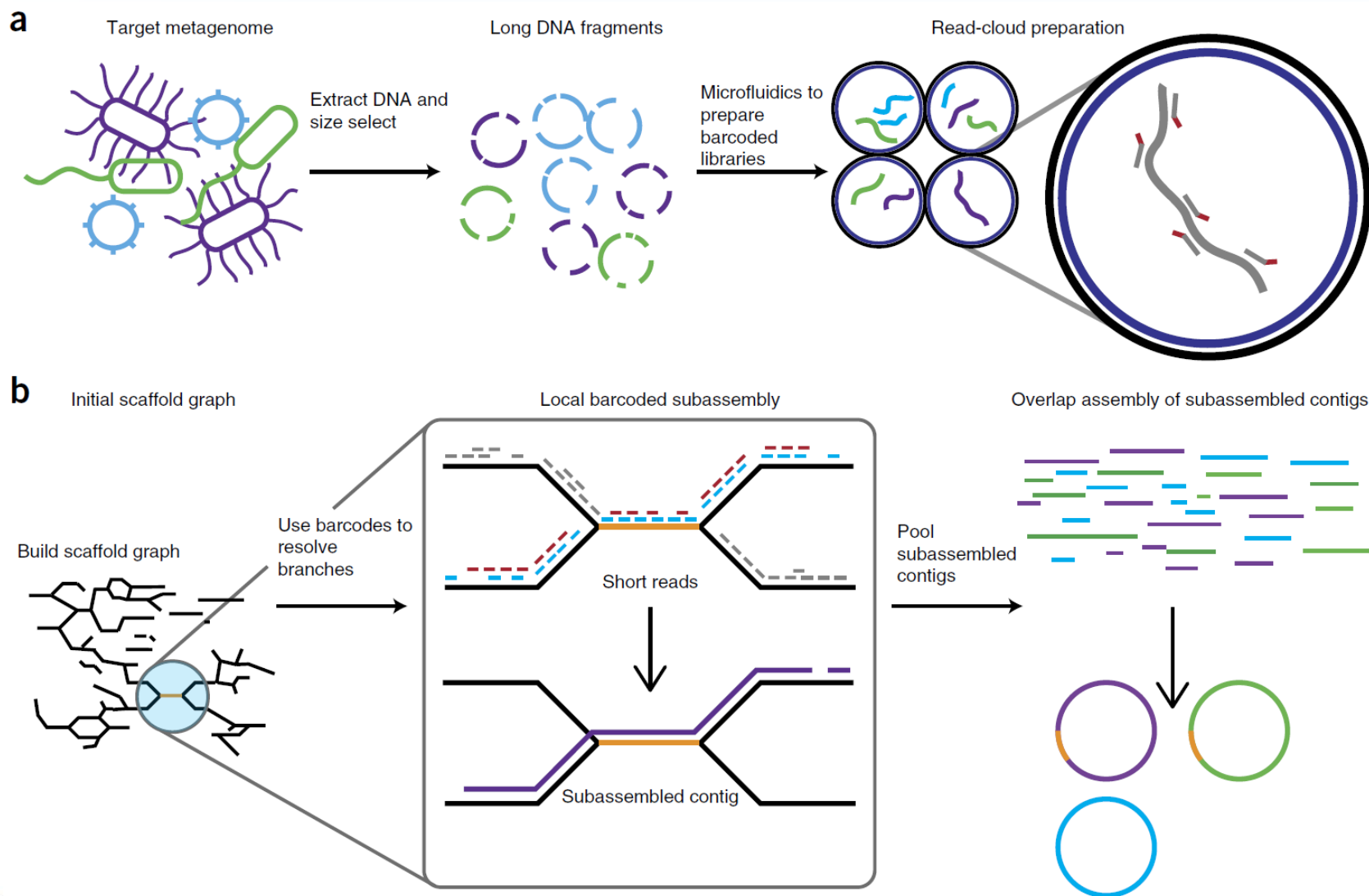
NBT: 宏表观组—DNA甲基化辅助宏基因组binning



基于三代PacBio测序的甲基化结果， 辅助分箱

Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nature Biotechnology* **36**, 61, doi:10.1038/nbt.4037 (2017).

NBT: "读云"建库+雅典娜组装获得高质量基因组



基于10X建库技术测序宏基因组。每个大片段来源序列有相同的Barcode，降低目标片段复杂度，提高组装准确率和长度。

Bishara, A. *et al.* High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nature Biotechnology* **36**, 1067, doi:10.1038/nbt.4266 (2018).

NBT: 超高速细菌基因组检索技术

2019年2月 *Nature Biotechnology* 封面文章

新的位片式索引技术，使检索全球44万个可用细菌基因组资源成为可能

分箱前比对数据库，排除已知菌基因组，降低复杂度可提高分箱质量

Bradley, P., den Bakker, H. C., Rocha, E. P. C., McVean, G. & Iqbal, Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology* **37**, 152-159, doi:10.1038/s41587-018-0010-1 (2019).



宏基因组

易生信



- Binning(分箱)就是拼乐高积木，按每个物种片段的特征进行分类；
- MetaWRAP依赖140多个软件的分箱流程，可采用Conda方式一键安装；它自带了质控、拼接和物种注释流程，可以从原始数据起始Binning；
- Binning提纯是综合了MaxBin2、MetaBat2和CONCOCT的结果，选出最优解；
- 包括Bin定量、物种和功能注释，方便下游差异比较或与表型关联；
- 重组装可进一步优化结果，但较耗时，根据实际情况选择；
- Binning领域还有很多路要走，如综合考虑更多算法和参考数据库等；
- 分箱的计算资源消耗巨大，本质上降低复杂度才是关键，结合新技术如三代测序、10X方式建库测序、高速检索等是未来的趋势。



- 宏基因组公众号 https://mp.weixin.qq.com/s/5jQspEvH5_4Xmart22gjMA
- 生信宝典公众号 <https://mp.weixin.qq.com/s/i71OMaUu6QtcY0pt1njHQA>
- 加拿大生信网 <https://bioinformatics.ca/>
- 美国高通量开源课程 <https://github.com/ngs-docs>
- [一文读懂宏基因组binning](#)
- 30余款宏基因组分析软件经验总结[上](#) [中](#) [下](#)
- [微生物组学数据分析工具综述 | 16S+宏基因组+宏病毒组+宏转录组](#)





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

