

Checkm介绍及云平台使用

1. checkM是用于评估分离出的微生物、单细胞和宏基因组的质量工具。
2. 使用有谱系世系关系的特有和独有基因数据集来大致估计基因组的完整度和污染程度。
3. <https://github.com/Ecogenomics/CheckM>
4. <https://github.com/Ecogenomics/CheckM/wiki>

1. pip安装的方法(pip install checkm-genome), 需要root权限 (推荐)

- > pip3 install numpy

- > pip3 install matplotlib

- > pip3 install pysam

#如果已安装可以忽略以上步骤

- > pip3 install checkm-genome

2. 通过Conda安装

- > conda install -c bioconda checkm-genome
- > conda install -c bioconda/label/cf201901 checkm-genome

3. 手动安装:

1. 需要的软件: HMMER($\geq 3.1b1$)、prodigal(2.60 or $\geq 2.6.1$)、pplacer(≥ 1.1)

> export PATH=\$PATH: /homeppplacer-Linux-v1.1.alpha17

2. 依赖的python模块: python ≥ 2.7 and < 3.0 、numpy $\geq 1.8.0$ 、scipy $\geq 0.9.0$ 、matplotlib $\geq 1.3.1$ 、pysam $\geq 0.8.3$ 、dendropy $\geq 4.0.0$ 、ScreamingBackpack $\geq 0.2.3$

1. CheckM假定基因组由contig组成，并且要处理的文件默认以.fna为后缀名;
2. -x可以指定其他后缀名（例如：-x fa）
3. *bin id*: 基因组bin的唯一标识符（来自输入的fasta文件）

工作流程

1. lineage-specific (世系特异性) 【推荐方法】

checkm lineage_wf <bin folder> <output folder>

2. taxonomic-specific (物种分类特异性)

checkm taxonomy_wf <rank> <taxon> <bin folder> <output folder>

<rank>: phylum; <taxon> : Cyanobacteria

3. custom marker genes (自行指定基因maker)

> checkm analyze <custom HMM file> <bin folder> <output folder>

> checkm qa <custom HMM file> <output folder>

下载数据库并设置数据库路径



国家微生物科学数据中心
National Microbiology Data Center

- `wget -c https://data.ace.uq.edu.au/public/CheckM_databases/checkm_data_2015_01_16.tar.gz`
- `tar -zxvf checkm_data_2015_01_16.tar.gz`
- `checkm data setRoot $PATH/checkm_data`

lineage-specific (世系特异性)



国家微生物科学数据中心
National Microbiology Data Center

1. (M) > checkm tree <bin folder> <output folder> 将基因组加入到参考基因组树中
2. (R) > checkm tree_qa <output folder> (可选) 检查树
3. (M) > checkm lineage_set <output folder> <marker file> 创建一个Marker文件，这个文件包含用于评估基因组的lineage-specific标记位点
4. (M) > checkm analyze <marker file> <bin folder> <output folder> 鉴定marker基因和评估基因组完整度和污染
5. (M) > checkm qa <marker file> <output folder> 对基因组质量进行总结

标准流程



国家微生物科学数据中心
National Microbiology Data Center

1. > checkm lineage_wf <bin folder> <output folder>

<bin folder>为基因组fna所在的输入目录

	Marker Lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination
d	root (UID1)	5656	56	24	0	0	0	0	0	56	100.00	2154.89
3	root (UID1)	5656	56	24	0	0	0	0	0	56	100.00	1064.59
7	k__Bacteria (UID203)	5449	104	58	0	10	24	18	23	29	100.00	304.81
3	k__Bacteria (UID1453)	901	171	117	4	163	4	0	0	0	97.01	2.99
7	c__Gammaproteobacteria (UID4444)	263	498	228	68	166	211	51	1	1	90.61	68.87
1	c__Gammaproteobacteria (UID4443)	356	451	270	35	402	14	0	0	0	90.19	4.07
1	k__Bacteria (UID2566)	525	208	136	19	187	2	0	0	0	89.20	1.10
5	k__Bacteria (UID2565)	2921	152	93	45	35	28	23	11	10	83.46	136.85
8	f__Flavobacteriaceae (UID2817)	81	511	283	74	406	31	0	0	0	81.81	6.66
7	c__Alphaproteobacteria (UID3305)	564	347	229	106	238	3	0	0	0	78.22	0.61
0	c__Gammaproteobacteria (UID4443)	356	451	270	91	302	56	2	0	0	76.97	12.89

3. checkm lineage_wf -h查看全部参数及用法

4. 例如: checkm lineage_wf -t 20 -x fa --nt --tab_table -f
bins_qa.txt metabat_bins bins_qa_result

1. unbinned 识别没有被分装 (unbinned) 的序列
2. coverage 计算序列的coverage
3. tetra 计算每条序列的四核苷酸频率
4. profile 计算map到每个bin的reads的百分率, 可用比较bins丰度
5. join_tables 将tab分割的不同bin信息表文件整合
6. ssu_finder 识别序列中的核糖体小亚基RNA (SSU rRNAs), 也即16S/18S

1. bin_qa_plot: 绘制bin完整度、污染度和异质性条形图
2. gc_plot: 绘制每个bin的不同序列GC含量分布直方图及误差图
3. coding_plot: 绘制每个bin序列的编码密度 (coding density, CD) 直方图及误差图
4. tetra_plot : 绘制 bin 每条序列与 bin 平均四核苷酸频率的距离 (tetranucleotide distance, TD) 直方图及误差图
5. dist_plot: 将以上三个图形绘制在一起

bins质量评估图像



国家微生物科学数据中心
National Microbiology Data Center

➤ checkm dist_plot [Options] out_folder bin_folder
plot_folder tetra_profile dist_value

➤ 例如:

```
checkm dist_plot --image_type pdf -x fa  
bins_qa_result metabat_bins checkm_plots  
../checkm_tetra.out 95
```

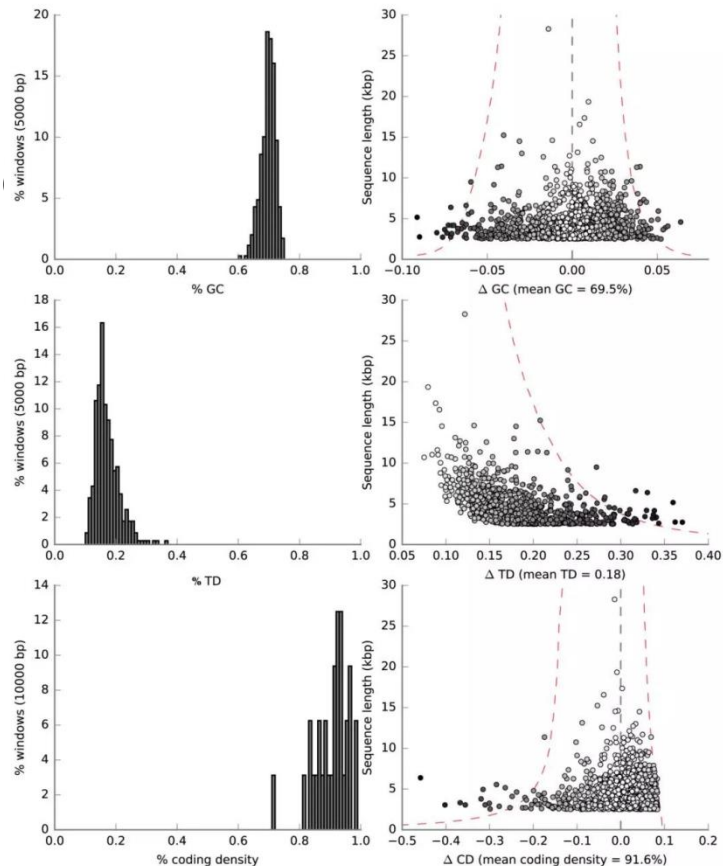


图1. bins质量评估图像

绘制bin完整度、污染度和异质性条形图



国家微生物科学数据中心
National Microbiology Data Center

```
> checkm bin_qa_plot --image_type pdf -x fa bins_qa_result metabat_bins
```

checkm_qa_plots

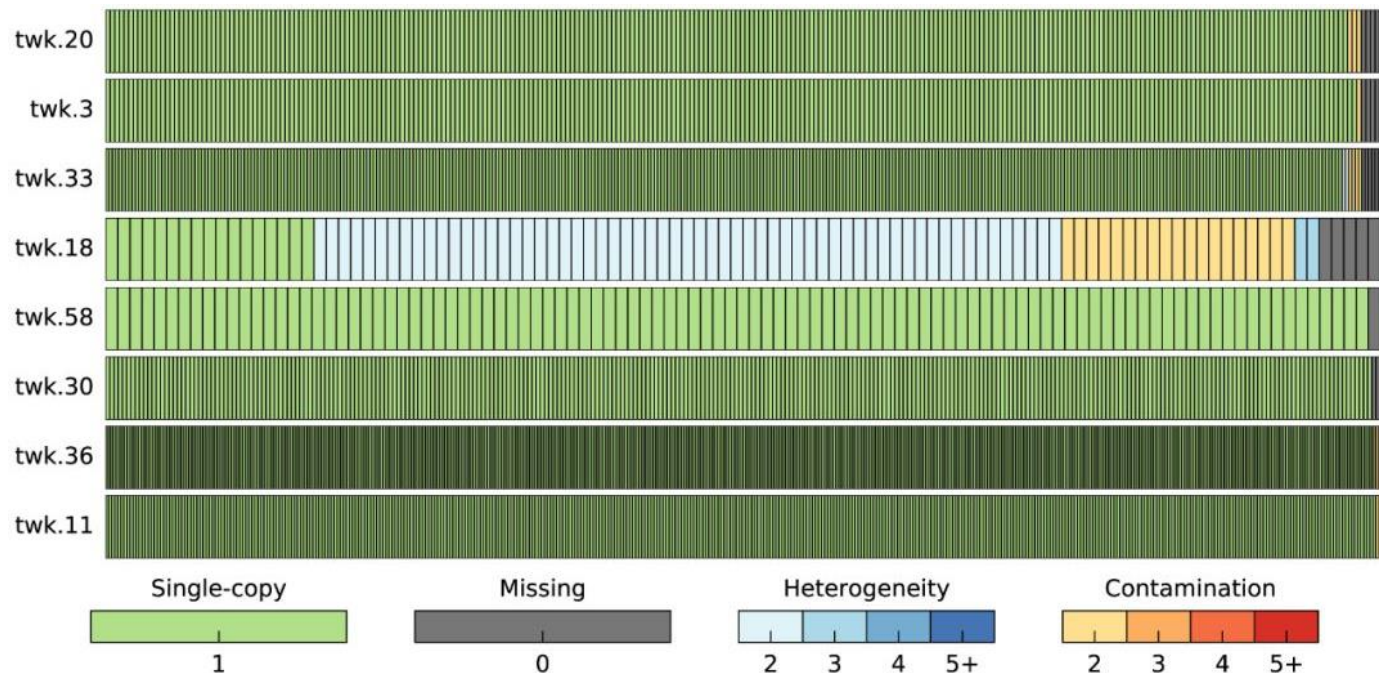


图2. bin完整度、污染度和异质性条形图

绘制Nx图

```
1. checkm nx_plot --image_type  
pdf -x fa --font_size 12  
metabat_bins checkm_Nx_plots
```

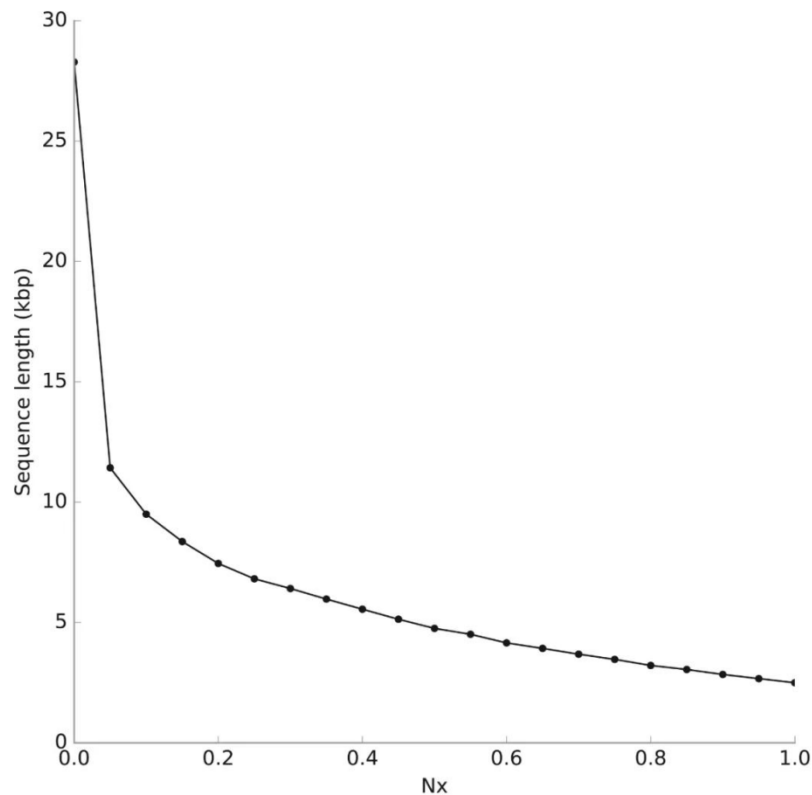


图3. bin的Nx图

首页

数据资源

元数据

数据下载

分析工具

数据汇交

服务案例

关于我们

当前位置: 首页 > 分析工具

请输入关键词



搜索: SOAPdenovo2 / CANU / PfamScan / LEfSe /

宏基因组分析流程 (3)

基因组拼接工具 (25)

基因组结构分析 (11)

基因组注释分析 (4)

元基因组分析 (20)

GenomeAssemblyAnnotation

原始链接: <https://gcmeta.wdcm...>

工具介绍: 宏基因组拼接注释是一款
针对于宏基因组测序数...

关键字: Genome Assemble An...

文章: Wenyu Shi, Heyuan Qi,
Oinolun Sun, Guomei...

ngsMetaAssembly

原始链接: <https://gcmeta.wdcm...>

工具介绍: 宏基因组拼接注释是一款
针对于宏基因组测序数...

关键字: Metagenome Assemb...

文章: Wenyu Shi, Heyuan Qi,
Oinolun Sun, Guomei...

simpleMetagenomeAnalysis

原始链接: <https://gcmeta.wdcm...>

工具介绍: 宏基因组直接注释是一款
针对于宏基因组测序数...

关键字: Metagenome Annotat...

文章: Wenyu Shi, Heyuan Qi,
Oinolun Sun, Guomei...

1. 网址: <http://biocloud.nmdc.cn/>
2. 分析软件: <http://www.nmdc.cn/analyze/>