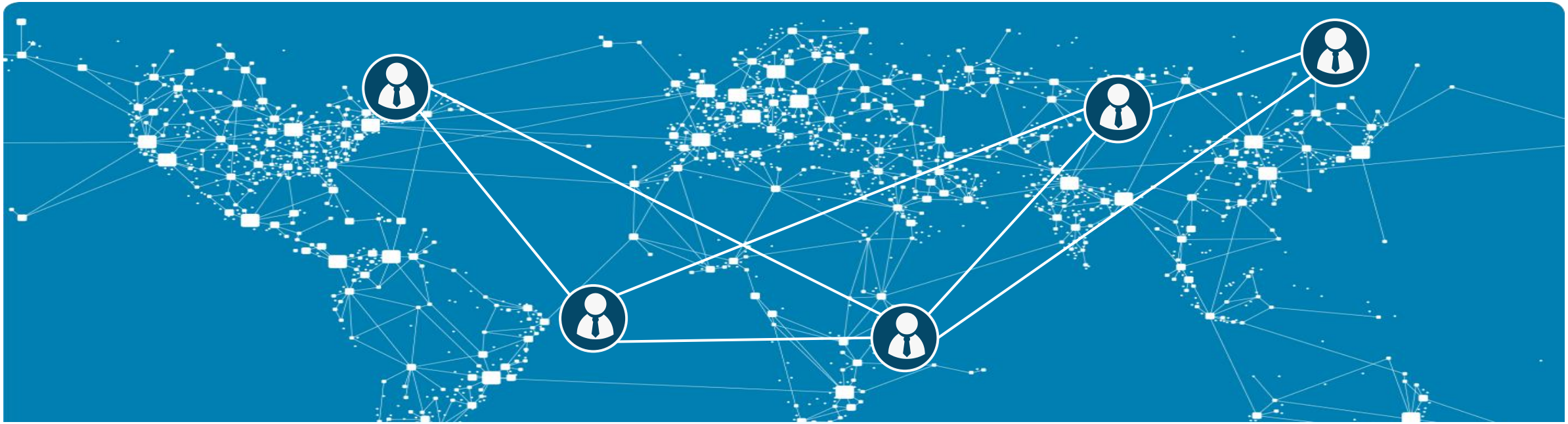
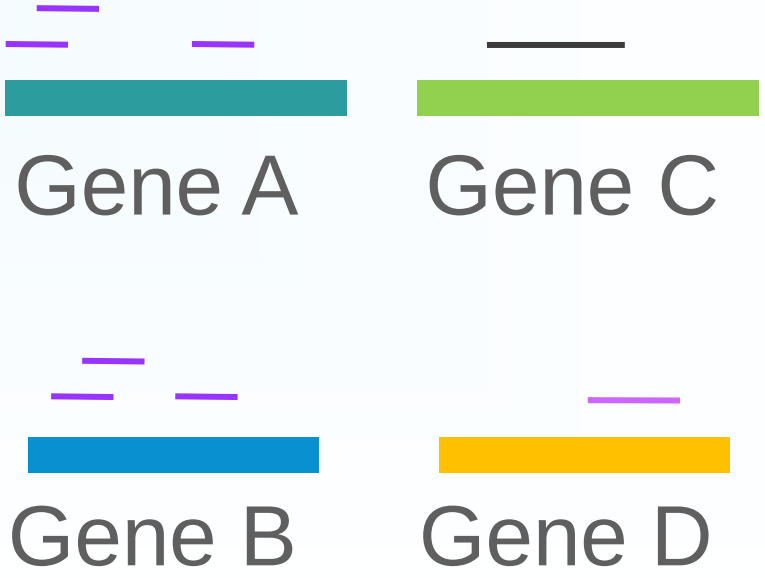
微生物组—宏基因组分析专题研讨会第14期



# 32 基因预测和定量Genes

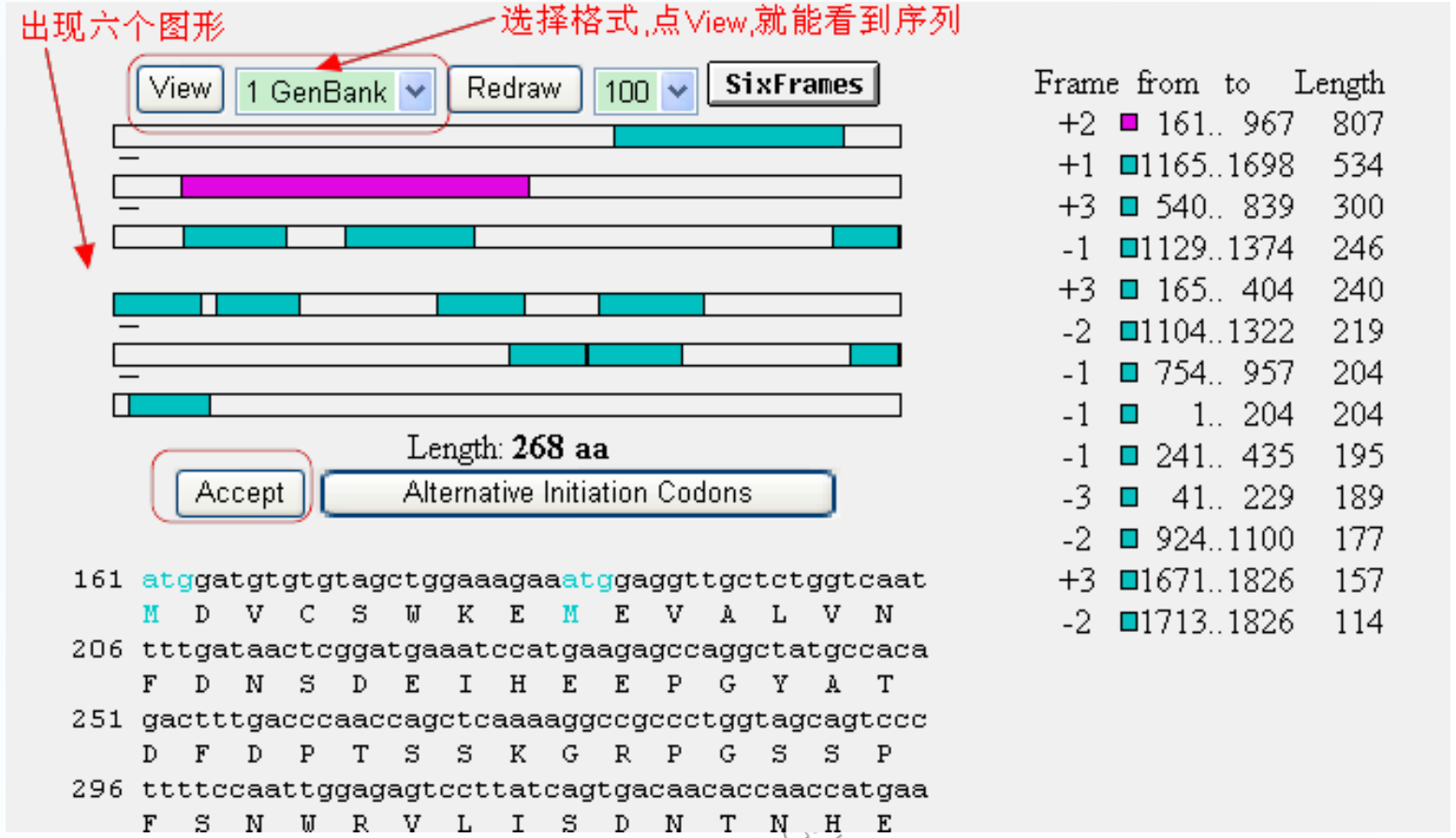



易生信

2022年3月27日

# 目录






https://www.ncbi.nlm.nih.gov/orffinder

# Prokka基因注释

Victorian Bioinformatics Consortium

## VICTORIAN BIOINFORMATICS CONSORTIUM

http://www.vicbioinformatics.com/software.prokka.shtml

ABOUT

STAFF

SOFTWARE

WEB TOOLS

### PROKKA

#### Description

Prokka is a software tool for the rapid annotation of prokaryotic genomes. A typical 4 Mbp genome can be fully annotated in less than 10 minutes on a quad-core computer, and scales well to 32 core SMP systems. It produces GFF3, GBK and SQN files that are ready for editing in Sequin and ultimately submitted to Genbank/DDJB/ENA.

#### Download

Prokka v1.12 — 14 March 2017 — Download (360MB) — MD5 — Changes — Docs — Paper — GitHub

Prokka: **rapid prokaryotic genome annotation**

T Seemann - Bioinformatics, 2014 - academic.oup.com

The multiplex capability and high yield of current day DNA-sequencing instruments has made bacterial whole **genome** sequencing a routine affair. The subsequent de novo assembly of reads into contigs has been well addressed. The final step of **annotating** all

☆ 99 Cited by 6825 Related articles All 8 versions

使用方法和常见问题附录

# Prodigal基因注释

[HTML] Prodigal: prokaryotic gene recognition and translation initiation site identification

D Hyatt, GL Chen, PF LoCascio… - BMC …, 2010 - bmcbioinformatics.biomedcentral …

The quality of automated gene prediction in microbial organisms has improved steadily over the past decade, but there is still room for improvement. Increasing the number of correct identifications, both of genes and of the translation initiation sites for each gene, and reducing the overall number of false positives, are all desirable goals. With our years of experience in manually curating genomes for the Joint Genome Institute, we developed a new gene prediction algorithm called Prodigal (PROkaryotic DYnamic programming Gene …

☆ 99 Cited by 5038 Related articles All 29 versions »» Prodigal：原核基因识别

Gene and translation initiation site prediction in metagenomic sequences

D Hyatt, PF LoCascio, LJ Hauser… - …, 2012 - academic.oup.com

Motivation: Gene prediction in metagenomic sequences remains a difficult problem. Current sequencing technologies do not achieve sufficient coverage to assemble the individual genomes in a typical sample; consequently, sequencing runs produce a large number of short sequences whose exact origin is unknown. Since these sequences are usually smaller than the average length of a gene, algorithms must make predictions based on very little data. Results: We present MetaProdigal, a metagenomic version of the gene prediction …

☆ 99 Cited by 348 Related articles All 15 versions metaProdigal：宏基因组中基因预测

# metaProdigal基因预测

```
mkdir -p temp/prodigal

# prodigal的meta模式预测基因，7s，>和2>&1记录分析过程至gene.log
time prodigal -i  result/megahit/final.contigs.fa \
    -d temp/prodigal/gene.fa \
    -o temp/prodigal/gene.gff \
    -p meta -f gff > temp/prodigal/gene.log 2>&1

# 查看日志是否运行完成，有无错误
tail temp/prodigal/gene.log
# 统计基因数量
grep -c '>' temp/prodigal/gene.fa
```

# 统计和提取完整基因

- # 统计完整基因数量，数据量大可只用完整基因部分
- 原核起始密码子通常为ATG, GTG或TTG，终止为TAA, TGA或TAG
- 00完整，01缺少终止密码子，10缺少起始密码子，11代表双端均缺失

```
grep -c 'partial=00' temp/prodigal/gene.fa
```

- # 提取完整基因

```
grep 'partial=00' temp/prodigal/gene.fa | cut -f1 -d ' '| sed 's/>//' >
temp/prodigal/full_length.id
seqkit grep -f temp/prodigal/full_length.id temp/prodigal/gene.fa >
temp/prodigal/full_length.fa
seqkit stat temp/prodigal/full_length.fa
```





详见：https://github.com/hyattpd/prodigal/wiki/understanding-the-prodigal-output

# 其它基因注释软件

- MetaGeneAnnotator conda install metagene_annotator
  http://metagene.nig.ac.jp/ 2006 NAR，2008 DNA Res，引用516和504次
- FragGeneScan conda install fraggenescan
  https://sourceforge.net/projects/fraggenescan/ 2010年发表于NAR，引用580次
- MetaGeneMark 未被conda收录，有在线工具
  http://exon.gatech.edu/GeneMark/ 2010年发表于NAR，引用796次
- GeneMarkS-2 未被conda收录，有在线工具
  http://exon.gatech.edu/GeneMark/genemarks2.cgi 2018年发表于GR，引用51次


Noguchi, Hideki, Jungho Park, and Toshihisa Takagi. "MetaGene: prokaryotic gene finding from environmental genome shotgun sequences." *Nucleic acids research* 34.19 (2006): 5623-5630.
Noguchi, Hideki, Takeaki Taniguchi, and Takehiko Itoh. "MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes." DNA research 15.6 (2008): 387-396.
Rho, Mina, Haixu Tang, and Yuzhen Ye. "FragGeneScan: predicting genes in short and error-prone reads." *Nucleic acids research* 38.20 (2010): e191-e191.
Zhu, Wenhan, Alexandre Lomsadze, and Mark Borodovsky. "Ab initio gene identification in metagenomic sequences." *Nucleic acids research* 38.12 (2010): e132-e132.
Lomsadze, A., Gemayel, K., Tang, S. & Borodovsky, M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Research 28, 1079-1089, doi:10.1101/gr.230615.117 (2018).

# 可选的最新预测方法GeneMarkS-2

**Input sequence**

**Enter sequence (FASTA or multi FASTA format)**

http://exon.gatech.edu/GeneMark/genemarks2.cgi

**or, upload file:** 选择文件 未选择任何文件

**Action** Start GeneMarkS-2 重置

**Options**

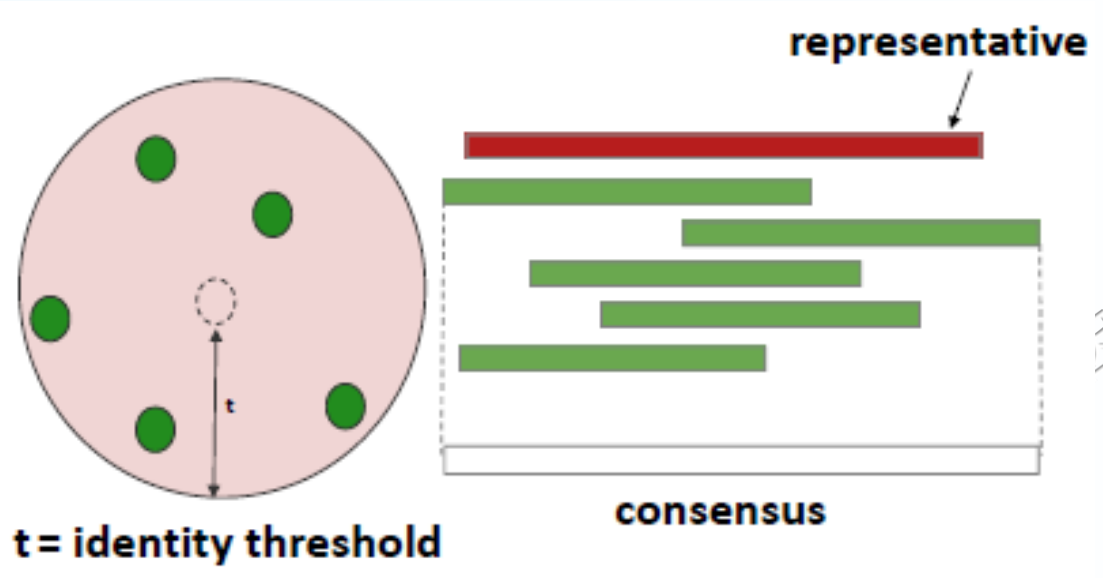
| Sequence type | Output format for gene prediction | Output options | Optional: results by E-mail |
|---|---|---|---|
| ◉ Prokaryotic<br>○ Bacteria<br>○ Archaea | ○ LST<br>◉ GFF3<br>○ GTF | ☑ Protein sequence<br>☑ Gene nucleotide sequence | E-mail<br><br>Subject<br>GeneMarkS-2<br>☐ Compress files |

**Advanced options**

◉ Genetic code 11.
○ Genetic code 4. "TGA" codon as Tryptophan (not a stop codon).
○ Genetic code 25. "TGA" codon as Glycine (not a stop codon).

# 目录





# 构建非冗余基因集

- 基因聚类，实现去除冗余基因、降低基因数量级
- 多样本、多批次或公共数据合并为一致参考序列(Reference)
- 通过CD-HIT将所有样本的基因序列根据序列相似性进行聚类，去除冗余序列(宏基因组常用阈值：coverage > 90%, identity > 95%）。

cd-hit

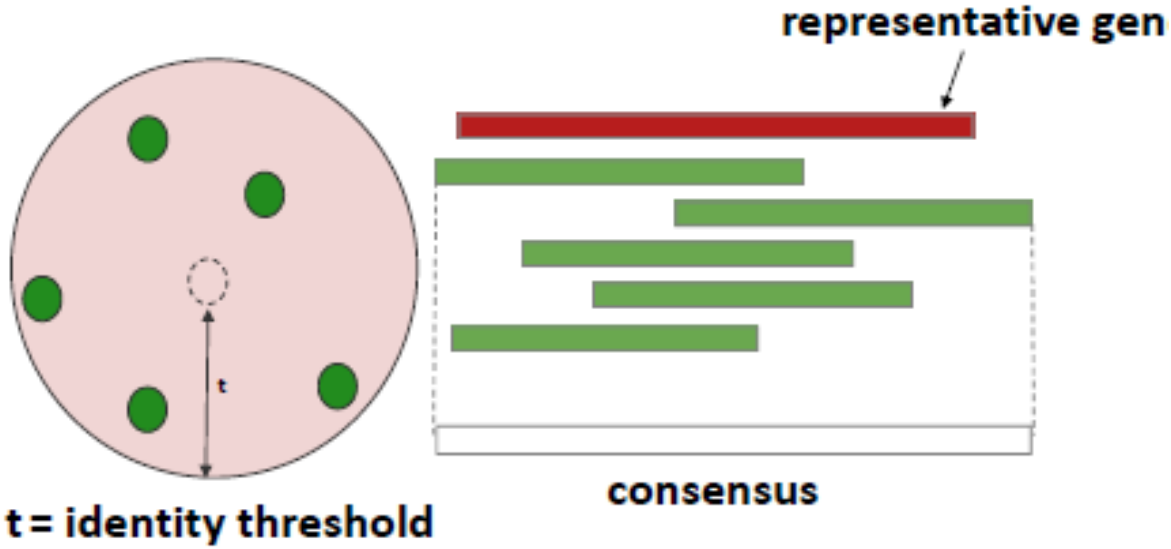
About 14,200 results (**0.03** sec)

**Cd-hit**: a fast program for clustering and comparing large sets of protein nucleotide sequences

W Li, A Godzik - Bioinformatics, 2006 - academic.oup.com

Abstract Motivation: In 2001 and 2002, we published two papers (Bioinformatics, 17, 282–283, Bioinformatics, 18, 77–82) describing an ultrafast protein sequence clustering program called **cd-hit**. This program can efficiently cluster a huge protein database with millions of …

☆ Cited by 6061 Related articles All 15 versions



# CD-HIT家族

| 小工具 | 功能 | 应用 |
| --- | --- | --- |
| cd-hit | 按指定相似度聚类蛋白质序列 | 非冗余蛋白集构建，如UniRef |
| **cd-hit-est** | 按指定相似度聚类核酸序列 | 非冗余基因集构建、重复序列家族分析、聚类OTUs |
| **cd-hit(-est)-2d** | 两个数据库比对 | 多批量、来源宏基因组构建非冗余基因集 |
| cd-hit-otu | 16S序列聚类 | 早期OTU鉴定方法 |

# 3.2.2 cd-hit构建非冗余基因集

```
# aS覆盖度，c相似度，G局部比对，g最优解，M内存0不限制，T多线程
time cd-hit-est -i temp/prodigal/gene.fa \
    -o result/NR/nucleotide.fa \
    -aS 0.9 -c 0.95 -G 0 -g 0 -T 0 -M 0
# 统计非冗余基因数量19146，目前研究经常达千万级
grep -c '>' temp/prodigal/nucleotide.fa

# 翻译核酸为对应蛋白序列, --trim去除结尾的*
seqkit translate --trim result/NR/nucleotide.fa > result/NR/protein.fa
```

# cd-hit-est-2d 两批次构建非冗余基因集

- A和B基因集，分别有M和N个非冗余基因
- 两批数据合并后用cd-hit-est去冗余，计算量是(M + N) X (M + N -1)
- cd-hit-est-2d比较，只有M X N的计算量

```
# 计算B中特有的基因
cd-hit-est-2d -i A.fa -i2 B.fa -o B.uni.fa \
    -aS 0.9 -c 0.95 -G 0 -g 0 \
    -T 96 -M 0 -d 0
# 合并为非冗余基因集
cat A.fa B.uni.fa > NR.fa
```

# 目录








http://wiki.biomine.skelleftea.se/wiki/index.php/Metagenomics

# Salmon非比对定量

- Salmon(三文鱼)是一款新的、极快的转录组计数软件。它与Kallisto(熊神星)和Sailfish(旗鱼)类似，可以不通过mapping而获得基因的counts值。Salmon的结果可由edgeR / DESeq2等进行counts值的下游分析。

Salmon provides fast and bias-aware quantifi

R Patro, G Duggal, MI Love, RA Irizarry, C Kingsford - Natu

We introduce Salmon, a lightweight method for quantifying
seq reads. Salmon combines a new dual-phase parallel infe
bias models with an ultra-fast read mapping procedure. It is
quantifier to correct for fragment GC-content bias, which, as
substantially improves the accuracy of abundance estimate
subsequent differential expression analysis.

☆ Cited by 3097 Related articles All 10 versior

nature methods

Brief Communication | Published: 06 March 2017

Salmon provides fast and bias-aware quantification of transcript expression

Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry & Carl Kingsford

*Nature Methods* **14**, 417–419 (2017) | Download Citation

不比对快速估计基因丰度Salmon

# Salmon安装和基本功能

- conda install salmon # 安装
- samlon -h # 查看帮助
- salmon v1.4，主要提供以下5类功能

**index Create a salmon index # 建索引**

**quant Quantify a sample # 样本定量**

alevin single cell analysis # 单细胞分析

swim Perform super-secret operation

**quantmerge Merge multiple quantifications into a single file # 合并样本结果**

# 3.2.3 基因定量——建索引

```
# 建索引,-t转录本，-p线程数， -i 索引
salmon index -t result/NR/nucleotide.fa \
    -p 9 -i temp/salmon/index
# 直接运行salmon找不到lib库，可重定义程序完整路径 (error while loading shared libraries: liblzma.so.0)
alias salmon="/conda2/envs/metagenome_env/share/salmon/bin/salmon"
```

解决思路和另一种使用方法见附录

# 3.2.3 基因定量——定量

```
# 定量，l文库类型自动选择，p线程 ，--meta宏基因组模式
tail -n+2 result/metadata.txt|cut -f1|rush -j 2 \
  'salmon quant -i temp/salmon/index -l A -p 3 --meta \
  -1 temp/qc/{1}_1_kneaddata_paired_1.fastq \
  -2 temp/qc/{1}_1_kneaddata_paired_2.fastq \
  -o temp/salmon/{1}.quant'
```

# 3.2.3 基因定量——合并为丰度矩阵

```
mkdir -p result/salmon
# 合并百万分比TPM
salmon quantmerge --quants temp/salmon/*.quant \
  -o result/salmon/gene.TPM
# 合并原始reads count值
salmon quantmerge --quants temp/salmon/*.quant \
  --column NumReads -o result/salmon/gene.count
sed -i '1 s/.quant//g' result/salmon/gene.*
```

# 基因定量完成

- 基因丰度矩阵(TPM)，可下游STAMP、LEfSe、limma、t.test等统计

| Name | C1 | C2 | C3 | C4 | C5 | C6 | N1 | N2 | N3 | N4 | N5 | N6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k6_1 | 0 | 0 | 44.1 | 0 | 0 | 175.4 | 0 | 67.6 | 0 | 0 | 0 | 0 |
| k4_1 | 0 | 36.4 | 62.6 | 33.6 | 0 | 0 | 98.9 | 0 | 110 | 0 | 0 | 0 |
| k3_1 | 0 | 12.1 | 10.4 | 54.0 | 0 | 0 | 0 | 0 | 0 | 43.1 | 20.3 | 0 |
| k2_1 | 0 | 0 | 0 | 27.8 | 0 | 0 | 0 | 0 | 45.1 | 44.4 | 170.2 | 0 |

- 基因定量原始计数，下游计算多样性或edgeR、DESeq2统计差异

| Name | C1 | C2 | C3 | C4 | C5 | C6 | N1 | N2 | N3 | N4 | N5 | N6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k6_1 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| k4_1 | 0 | 1 | 2 | 1 | 0 | 0 | 2 | 0 | 3 | 0 | 0 | 0 |
| k3_1 | 0 | 1 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 5 | 2 | 0 |
| k2_1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 | 4 | 13 | 0 |

# 总结

- Prokka提供了基因预测、注释流程，但依赖关系多容易报错；
- Prodigal的meta模式是Prokka的核心，只使用这部分可提速100倍；
- Cd-hit可建立非冗余基因集，多基因集合并等，方便开展多样品定量、比较；
- Cd-hit-2d可进行两批非冗余基因的冗余，显著减少计算时间；
- Salmon基于k-mer的非比对定量方法：快速，准确，节约空间(非比对方法没有序列比对中间文件)；主要分为建索引，定量和合并3步；
- 软件更新快，使用出问题时，需要正对照。

# 参考资源

- 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

# 附录：基因注释Prokka

- Prokka: rapid prokaryotic genome annotation
- Prokka是一个命令行软件工具，可以在一台典型台式机上在约10分钟内充分注释一个细菌基因组草图。它产生标准兼容的输出文件以进行进一步分析或者在基因组浏览器中查看。
- 2014年发表于Bioinformatics，最新版本1.12于2017年3月14日更新，大小360MB。因为它是一个复杂的分析流程，依赖关系众多。
- 安装：conda install prokka





Seemann, Torsten. "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30.14 (2014): 2068-2069.
http://www.vicbioinformatics.com/software.prokka.shtml

# 3.2.1 prokka基因注释

```
# 查看文件大小，预估时间
ls -sh temp/megahit/final.contigs.fa # 31 Mb
time prokka temp/megahit/final.contigs.fa --outdir temp/prokka \
  --prefix mg --metagenome --kingdom Archaea,Bacteria,Mitochondria,Viruses \
  --force --cpus 3
# 以mg开头，注释宏基因组，多界，强制覆盖输出
# 3线程，耗时2m
```

# Prokka运行常见问题1：Java版本过高

Error: A JNI error has occurred, please check your installation and try again
Exception in thread "main" java.lang.UnsupportedClassVersionError: minced has been compiled by a more recent version of the Java Runtime (class file version 55.0), this version of the Java Runtime only recognizes class file versions up to 52.0

- 错误提示：系列中为Java55，而软件最高支持Java52
- 解决办法：使用另一个conda虚拟环境，如metawrap
- source ${soft}/bin/activate metawrap
- 结束后 conda deactivate 退出

# Prokka运行常见问题2：找不到Perl模块

- Can't locate XML/Simple.pm
- 错误描述：找不到*.pm，即Perl模块
- 解决思路：手动查找位置，并添加环境变量
- locate XML/Simple.pm # locate或find找模块位置
- # export设置PERL5LIB变量包括找到的包位置即可
- export PERL5LIB=$PERL5LIB:${soft}/envs/metawrap/lib/perl5/site_perl/5.22.0

# Prokka结果说明

**.gff: 基因注释文件，包括gff和序列，可用igv直接查看**

.gbk: Genebank格式，来自gff

.fna: 输入contig核酸文件

**.faa: 翻译CDS的AA序列**

**.ffn: 所有转录本核酸序列**

.sqn: 用于提交的序列

.fsa: 输入序列，但有sqn的描述，用于tbl2asn生成sqn文件

.tbl: 特征表，用于tbl2asn生成sqn文件

.err: 错误报告

.log: 日志

.txt: 统计结果

.tsv: 所有注释基因特征表格