

# 生物信息分析流程

## 目录

原始下机数据处理.....	2
样品组装.....	2
基因组组分分析.....	2
编码基因.....	2
重复序列.....	2
非编码RNA.....	3
基因岛.....	3
前噬菌体.....	3
CRISPR.....	3
基因功能分析.....	3
基因功能注释.....	4
效应子.....	4
分泌蛋白预测.....	4
分泌系统蛋白及T3SS 效应蛋白预测.....	4
次级代谢基因簇分析.....	4
毒力或致病性分析.....	4

## 原始下机数据处理

测序得到的原始数据(Raw Data)会存在一定比例的低质量数据, 为了保证后续信息分析结果的准确可靠, 首先要对原始数据进行过滤处理, 得到有效数据(Clean Data)。结果见“01.Cleandata.tar.gz 或 01.Cleandata”。

具体处理步骤如下:

- 1) 去除所含低质量碱基 (质量值 $\leq 20$ ) 超过一定比例 (默认设为 40%) 的 reads;
- 2) 去除 N 碱基达到一定比例的 reads (默认设为 10%);
- 3) 去除与 Adapter 之间 overlap 超过一定阈值 (默认设为 15bp), 且错配数小于 3 的 reads;
- 4) 对于小基因组等项目, 如果样品存在宿主污染, 需与宿主数据库进行比对, 过滤掉可能来源于宿主的 reads。

## 样品组装

从各样品质控后的 Clean Data 出发, 进行基因组组装, 得到能反映样品基因组基本情况的序列文件, 并对组装结果进行评价。结果见“02.Assembly.tar.gz 或 02.Assembly”。

基因组组装的具体处理步骤如下:

- 1) 经过预处理后得到 Clean Data, 使用 SOAP denovo 组装软件进行组装:  
选取不同的 K-mer (默认选取 95、107、119) 进行组装, 根据项目类型选择最优的 kmer, 选取最少 scaffold 的组装结果; 利用最优 kmer 并调节其他参数 (-d -u -R -F 等) 再次筛选得到初步组装结果;
- 2) 使用 SPAdes 软件进行组装:  
选取不同的 K-mer (默认选取 99 和 127) 进行组装, 根据项目类型选择最优的 kmer, 选取最少 scaffold 的组装结果;
- 3) 使用 Abyss 软件进行组装:  
选取 K-mer 64 进行组装, 获得组装结果;
- 4) 使用 CISA 软件整合三个软件的组装结果, 并挑选其中最少数 scaffold 的组装结果;
- 5) 采用 gapclose 软件对初步组装结果进行补洞, 并且通过过滤低测序深度 (小于平均深度的 0.35) 的 reads 去除同 lane 污染, 从而得到最终的组装结果;
- 6) 过滤掉 500bp 以下的片段, 并进行评估和统计分析以及后续基因预测。

## 基因组组分析

### 编码基因

从各样品最终的组装结果 ( $\geq 500$ bp) 出发, 采用对应的软件进行 ORF (Open Reading Frame) 预测及过滤。结果见“03.Genome\_Component.tar.gz 或 03.Genome\_Component 中 Gene”。

对于细菌、古菌等样本采用 GeneMarkS 软件 (verison 4.17) (<http://topaz.gatech.edu/>) 进行编码基因预测及过滤。

### 重复序列

根据重复的序列在基因组上的分布, 分为两大类: 散在重复序列、串联重复序列。散在重复序列是与串联重复序列的组织形式不同的另一类重复序列, 是散在方式分布于基因组内的散在重复序列。串联重复序列 (Tandem Repeat, TR), 即相邻的、重复两次或多次特定核酸序列模式的重复序列。结果见“03.Genome\_Component.tar.gz 或 03.Genome\_Component 中 Repeat”。

通过 RepeatMasker 软件 (verison 4.0.5) 进行散在重复序列预测, TRF (Tandem repeats finder, verison 4.07b) 搜寻 DNA 序列中的串联重复序列。

## 非编码RNA

非编码RNA (ncRNA) 是一类广泛存在于小基因组、细菌、古生菌和真核生物生物体内, 执行多种生物学功能的RNA 分子, 其本身并不携带翻译为蛋白质的信息, 直接在RNA 水平对生命活动发挥作用。对于微生物而言, 非编码 RNA 的主要类型包括sRNA、rRNA、tRNA、snRNA、snoRNA 及 miRNA 等等。结果见“03.Genome\_Component.tar.gz 或 03.Genome\_Component 中 ncRNA”。

常见类型 ncRNA 预测方法如下:

- 1) tRNA: 通过 tRNAscan-SE 软件可预测tRNA 区域和tRNA 的二级结构。
- 2) rRNA: 通过比对 rRNA 库查找与数据库近缘的 rRNA (identity 默认 $\geq 50\%$ ), 并利用 rRNAmmer 软件预测新出现及未被注释的rRNAs, 综合二者结果。
- 3) sRNA: 首先用Rfam 软件进行Rfam database 比对注释, 接着用cmsearch 程序 (version 1.1rc4) 确定最终的sRNA。

## 基因岛

基因岛 (Genomics Islands, GIs) 是一些细菌、噬菌体或质粒中有横向起源迹象的一部分基因组。基因岛大小通常从 10 到 500 Kbp, 其 GC 含量和密码子选择与基因组的其余部分不同。基因岛的许多亚分类基于它们所表现出来的功能, 如病原性基因岛 (Pathogenicity islands, PAIs) 其上包含有毒力基因, 与致病机理相关, 抗生素抗性基因岛包含有许多抗生素抗性基因。结果见 “03.Genome\_Component.tar.gz 或 03.Genome\_Component 中 Genomic\_Island”。

采用 IslandPath-DIOMB 软件预测基因岛, 其通过检测序列中二核苷酸偏向性 (phylogenetically bias) 和移动性基因 (mobility genes, 如转座酶或整合酶) 以判定基因岛以及潜在的水平基因转移。

## 前噬菌体

整合在宿主基因组上的温和噬菌体的核酸称之为前噬菌体。前噬菌体预测是使用 phiSpy 工具 (Nov11\_v2.3), 这是一个预测精度和准确性都很高的工具, 支持 GenBank 格式格式文件, 结果见“03.Genome\_Component.tar.gz 或 03.Genome\_Component 中 Prophage”。

## CRISPR

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat Sequences, 成簇的、规律间隔的短回文重复序列) 是在 40%已测序的细菌和 90%已测序的古细菌中都有报道过的、含有一个与噬菌体和质粒同源的短的重复序列, 通过对外来同源的 DNA 作用对噬菌体有抗性, 影响质粒的连接, 是原核生物的免疫系统一部分。

采用 CRISPRdigger 工具对样品基因组进行 CRISPR 预测, CRISPRdigger 准确性高并能预测其中的短回文重复结构。结果见 “03.Genome\_Component.tar.gz 或 03.Genome\_Component 中 CRISPR”。

## 基因功能分析

## 基因功能注释

目前提供注释的通用功能数据库主要有：

GO(Gene Ontology, <http://geneontology.org/>)、KEGG (Kyoto Encyclopedia of Genes and Genomes, <http://www.genome.jp/kegg>)、COG(Cluster of Orthologous Groups of proteins, <http://www.ncbi.nlm.nih.gov/COG/>)、NR(Non-Redundant Protein Database)、TCDB(Transporter Classification Database)、Swiss-Prot(<http://www.ebi.ac.uk/uniprot/>)、CAZy(Carbohydrate-Active enZymes Database)等。结果见“04.Genome\_Function.tar.gz 或 Genome\_Function 中 General\_Gene\_Annotation 内各数据库”。

功能注释基本步骤如下：

- 1) 将预测基因与各功能数据库进行 BLAST 比对 (blastp,  $\text{evalue} \leq 1e-5$ ) ;
- 2) BLAST 结果过滤：对于每一条序列的 BLAST 结果，选取 score 最高的比对结果 (默认  $\text{identity} \geq 40\%$ ,  $\text{coverage} \geq 40\%$ ) 进行注释。

## 效应子

### 分泌蛋白预测

分泌蛋白是指在细胞内合成后，在信号肽的引导下穿过细胞膜分泌到细胞外起作用的蛋白质。分泌蛋白中有许多是生命活动所需的重要酶类。分泌蛋白的 N 端是由 15~30 个氨基酸组成的信号肽，对分泌蛋白的分泌起主导作用。

使用信号肽预测工具 SignalP 进行预测，检测是否含有信号肽及跨膜结果，综合预测蛋白序列是否是分泌蛋白。结果见“04.Genome\_Function.tar.gz 或 04.Genome\_Function/\*/Effector/Secretory\_Protein”。

### 分泌系统蛋白及T3SS 效应蛋白预测

病原菌通过分泌系统TNSS (type N secretion systems, 目前确定的有7种, I型-VII型) 将该类蛋白分泌至胞外或是宿主细胞中，通过控制免疫应答反应以及细胞衰亡引起病理反。其中革兰氏阴性菌的T3SS通常用来从分子水平研究病原菌，感染机制，毒力作用等，是研究得比较多的分泌系统。

对于 TNSS 系统，通过蛋白序列功能数据库注释结果中，提取分泌系统相关蛋白进行注释。结果见“04.Genome\_Function.tar.gz 或 04.Genome\_Function/\*/Effector/TnSS”。

对于革兰氏阴性菌，另外采用 EffectiveT3 软件预测 T3SS 效应蛋白。结果见“04.Genome\_Function.tar.gz 或 04.Genome\_Function/\*/Effector/T3SS”。

### 次级代谢基因簇分析

次级代谢产物是微生物在一定的生长时期，以初级代谢产物为前体合成的对微生物的生命活动无明确功能，并非生长繁殖所必需的物质。采用 antiSMASH 程序 (version 2.0.2) 对基因组进行预测。

结果见“04.Genome\_Function.tar.gz 或 04.Genome\_Function/\*/Effector/Secondary\_Metabolism”。

### 毒力或致病性分析

对于细菌，目前提供注释的病原细菌致病性和耐药性数据库主要有 PHI (Pathogen Host Interactions Database)、ARDB (Antibiotic Resistance Genes Database)、CARD (Comprehensive Antibiotic Research Database)。结果见“04.Genome\_Function.tar.gz 或 04.Genome\_Function 中 Pathogenicity 内各数据库”。