生物信息分析流程

目录

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≤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；

2）首先选取不同的 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（细菌项目选取最少 scaffold，真菌项目选取最大的 N50）；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；

3）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；

4）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。

基因组组分分析

编码基因

从各样品最终的组装结果（>=500bp）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”。

对于真菌样本基因预测的方法如下：（1）若提供转录组数据，进行基于转录组数据 Transdecoder/Glimmer/Snap 的从头 PASA 预测、基于转录组数据的 Cufflinks 预测、从头 Augustus（version 2.7）预测和以近缘序列（若有提供）做参考的同源 Genewise（version 2.4.1）预测，然后将多种方法的结果进行 EVM 整合和 PASA 第二轮验证。（2）若未提供转录组数 据，但提供近缘参考序列，则进行同源 Genewise 预测。（3）若未提供转录组数据和近缘参 考序列，则进行从头 Augustus 预测。

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、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>=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/ ）、KOG、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 04.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，evalue ≤ 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identity>=40%，coverage>=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”。

细胞色素 P450 数据库注释

细胞色素 P450（cytochromeP450 或 CYP450,简称 CYP450）为一类亚铁血红素—硫醇盐蛋白的超家族，它参与内源性物质和包括药物、环境化合物在内的外源性物质的代谢。

结果见 “04.Genome_Function.tar.gz 或 04.Genome_Function/*/Effector/P450”。

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”。

致病性分析

对于真菌，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）、DFVF（database of fungal virulence）。结果见“04.Genome_Function.tar.gz或04.Genome_Function 中 Pathogenicity 内各数据库”。