

1 测序数据质控

1.1 数据产出比例统计

Nanopore 测序的下机数据的原始数据格式为包含所有原始测序信号的二进制 fast5 格式，通过 MinKNOW 软件包中的 Albacore 软件进行 base calling 后会将 fast5 格式数据转换为 fastq 格式，经进一步过滤接头、低质量及短片段(长度<2000 bp)的 reads 后，得到总的数据集。

2 基因组组装

2.1 组装结果统计

使用 Canu v1.5 软件对过滤后 subreads 进行纠错，然后采用 wtdbg v2.2 软件对纠错后 subreads 进行组装，通过 Racon 软件利用过滤后的 subreads 对组装结果进行校正，最后采用 Pilon v1.22 软件利用二代数据进一步对组装基因组进行纠错（若无二代数据，则跳过此步），得到最终准确度更高的基因组。

2.2 组装评估

分别从以下几个方面对基因组组装完整度进行评估。

2.2.1 二代测序数据回比率评估

比对率：即可以比对到基因组上的 Clean Reads 占总的 Clean Reads 数比例，比对率的高低跟基因组组装质量高低及 reads 测序质量有关，基因组组装越完整、测序 reads 质量越高，比对率越高。比对采用 bwa 软件。

Mapped (%): 比对到参考基因组的 Clean Reads 数占有所有 Clean Reads 数的百分比，使用 samtools flagstat 命令实现；

Properly mapped (%): 双端测序序列均比对到参考基因组上且距离符合测序片段的长度分布，使用 samtools flagstat 命令实现。

Coverage (%): 组装基因组的覆盖度，使用 samtools depth 命令实现。

Depth (X): 组装基因组的覆盖深度，使用 samtools depth 命令实现。

2.2.2 BUSCO 评估

BUSCO 中 fungi_odb9 数据库包含了真菌的 290 个保守的核心基因。我们使用 BUSCO v2.0 软件来评估该真菌基因组组装的完整性。

3 基因组组分析

3.1 重复序列预测

由于物种间重复序列的保守性相对较低，针对特定的物种进行重复序列的预测时需要构建特定的重复序列数据库。因此，我们借助 LTR_FINDER v1.05、MITE-Hunter v1.0.0、RepeatScout v1.0.5、PILER-DF v2.4 四个软件，基于结构预测和从头预测的原理构建该真菌基因组的重复序列数据库，用 PASTECClassifier v1.0 对数据库进行分类，然后与 Repbase (19.06)的数据库合并做为最终的重复序列数据库，然后利用 RepeatMasker v4.0.6 软件基于构建好的重复序列数据库对该真菌进行重复序列的预测。

3.1.1 重复序列预测结果描述

采用 GFF 格式对重复序列进行描述，包括重复序列元件（主要为转座子）的来源，在基因组座位，类型，特征等信息。具体可参考 GFF 格式说明，<http://www.sequenceontology.org/gff3.shtml>。

3.2 编码蛋白基因预测

3.2.1 基因预测结果统计

基因结构预测主要采用从头预测、基于同源蛋白预测和基于转录组证据预测，然后对 3

种预测结果进行整合。使用 Genscan v1.0、Augustus v2.4、GlimmerHMM v3.0.4、GeneID v1.4、SNAP (version 2006-07-28)进行从头预测；使用 GeMoMa v1.3.1 进行基于同源蛋白的预测；使用 Hisat2 v2.0.4 和 Stringtie v1.2.3 进行基于有参考转录本的组装，并利用 TransDecoder v2.0 使用 PASA v2.0.2[25]进行基于转录组组装的 Unigene 序列的预测；最后利用 EVM v1.1.1 整合上述 3 种方法得到的预测结果，并用 PASA v2.0.2 进行修饰。

3.2.2 基因基本信息统计

统计编码蛋白基因的数量、平均长度，以及内含子、外显子的数量和平均长度。

3.2.3 基因预测结果描述

GFF 格式是对基因特征（也可用于其他类似的基因组元件）进行描述的一种标准格式，描述了基因名称，来源，在基因组座位，类型，特征等信息。具体可参考 GFF 格式说明，<http://www.sequenceontology.org/gff3.shtml>。

3.3 非编码 RNA 预测

非编码 RNA 即不编码蛋白质的 RNA，包括 microRNA、rRNA 和 tRNA 等多种已知功能的 RNA，针对非编码 RNA 的结构特点，采用不同的策略预测不同的非编码 RNA。使用软件 tRNAscan-SE v1.3.1 预测基因组中的 tRNA，使用软件 Infernal v1.1 基于 Rfam v12.0 数据库预测基因组中的 rRNA 以及除了 tRNA 和 rRNA 之外的其它 ncRNA。

3.4 假基因预测

假基因(pseudogene)是具有与功能基因相似的序列，但由于插入、缺失等突变以致失去了原有的功能。我们利用已预测得到的蛋白序列与 Swiss-Prot 数据库中收录的蛋白序列，通过软件 GenBlastA v1.0.1 比对，在基因组上寻找同源的基因序列（可能的基因），然后利用软件 GeneWise wise2-2-0 寻找基因序列中的不成熟的终止密码子及移码突变，得到假基因。

3.5 基因簇预测

细菌和真菌产生的次生代谢产物是抗菌剂和其他生物活性化合物的重要来源，包括许多已经和正在被用作例如抗生素，降低胆固醇的药物或抗肿瘤药物的化学物质的生物合成途径。使用软件 antiSMASH v6.0.0 鉴定和分析细菌和真菌基因组序列中生物合成基因簇(BGC)。

4 基因组功能注释

4.1 通用数据库注释

利用预测得到的基因序列与 KOG、KEGG、Swiss-Prot (2015_01)、TrEMBL、Nr 等功能数据库做 BLAST v2.2.29 比对，得到基因功能注释结果。基于 Nr 数据库比对结果，应用软件 Blast2GO v2.5 进行 GO (releases20160907)数据库的功能注释。利用软件 hmmer v3.0 基于 Pfam (27.0)数据库进行 Pfam 功能注释。另外，对 KOG、KEGG 代谢通路富集分析、GO 功能富集分析等基因功能注释分析。

4.2 专有数据库注释

利用预测得到的基因的蛋白序列与转运蛋白分类数据库(TCDB)、病原体-宿主互作因子数据库(PHI)、细胞色素 p450 工程数据库(CYPE)、真菌毒力因子数据库(DFVF)等功能数据库做 BLAST 比对，得到相应的注释结果。另外，利用软件 hmmer 基于碳水化合物相关酶数据库(CAZy)进行碳水化合物酶类基因的功能注释。

4.2.1 CAZy 数据库注释

CAZy v4.0 全称为 Carbohydrate-Active enZymes Database，即碳水化合物活性酶数据库，参考链接 <http://www.cazy.org/>。该数据库主要包含与糖苷键降解、修饰及生成相关的酶类家族。主要包含 5 大分类：糖苷水解酶(Glycoside Hydrolases, GHs)、糖基转移酶(Glycosyl Transferases, GTs)、多糖裂解酶(Polysaccharide Lyases, PLs)、碳水化合物酯酶(Carbohydrate Esterases, CEs)、辅助活性酶(Auxiliary Activities, AAs)。此外，该数据库还包含与碳水化

合物结合相关的酶（Carbohydrate-Binding Modules, CBMs）。

4.2.2 TCDB 数据库注释

TCDB 是对膜转运蛋白进行分类的数据库，它制定了一套转运蛋白分类系统 Transporter Classification (TC) System，类似于对酶进行分类的 EC 系统，参考链接 <http://www.tcdb.org/>。TC 分类系统包含 5 个层级，因此，TC Number 包含 5 个数字或者字母，每个数字或字母实际代表某一个层级的分类。

4.2.3 PHI 数据库注释

PHI（病原宿主互作数据库），收录了经过实验验证或文献报道的能够感染动植物、真菌和昆虫的细菌、真菌等病原菌的致病基因、毒力基因和效应蛋白基因。

4.2.4 CYPED 数据库注释

CYPED 全称为 CYtochrome P450 Engineering Database，即细胞色素 p450 工程数据库，参考链接 <https://cyped.biocatnet.de/>。该数据库主要包含细胞色素 P450 的序列、结构和功能信息。CYPED v6.0 共收录了 52,674 条序列。

4.2.5 DFVF 数据库注释

DFVF 是综合的已知真菌毒力因子在线数据库，收集了 85 个属的 228 株真菌产生的 2,058 个致病基因及其蛋白产物，参考链接 <http://sysbio.unl.edu/DFVF/>。毒力因子是毒力基因的蛋白质产物，例如毒素合成酶和分泌的降解酶，使病原体能够逃避宿主的防御机制，从而导致疾病。

4.3 蛋白亚细胞定位分析

4.3.1 信号肽预测

信号肽（signal peptides），是指引导新合成的蛋白质向分泌通路转移的短肽链（长度一般为 5~30 个氨基酸）。使用软件 SignalP 4.0 对所有的预测到的基因的蛋白序列进行分析，找出含有信号肽的蛋白。

4.3.2 跨膜蛋白预测

使用软件 tmhmm v2.0 所有的预测到的基因的蛋白序列进行分析，找出含有跨膜螺旋的蛋白，即为跨膜蛋白。

4.3.3 分泌蛋白预测

在上述预测的含有信号肽的蛋白中去除含有跨膜螺旋的蛋白，剩余的蛋白即为分泌蛋白。

4.3.4 效应蛋白预测

病原真菌通过分泌效应蛋白抑制植物防御反应，从而达到侵染植物的目的。EffectorP 软件通过自我训练算法，用于识别植物-病原真菌互作的效应蛋白。我们利用 EffectorP v2.0 对分泌蛋白进一步分析，预测真菌效应蛋白。