1 测序数据质控

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 guppy v3.2.6 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 进行组装。最后采用 Pilon 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。将组装 contig 序列同 NT 数据库进行比对，确定染色体类型。

3 基因组组分分析

3.1 编码基因预测

采用 Prodiga v2.6.3 软件对组装基因组进行编码基因预测。

3.2 重复序列预测

原核生物基因组中重复序列含量极少。采用 RepeatMasker v4.0.5 软件对细菌基因组进行重复序列的预测。

3.3 非编码 RNA 预测

非编码 RNA 即不编码蛋白质的 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 v2.0 预测基因组中的 tRNA，使用软件 Infernal v1.1.3 基于 Rfam 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 CRISPR 序列预测

CRISPR 是一串包含多个短而重复的序列的碱基序列，重复序列之间是一些长度约 30 bp 的"spacer DNA"。在原核生物中，CRISPR 起到免疫系统的作用，对外来的质粒和噬菌体序列具有抵抗作用。CRISPR 能识别并使入侵的功能元件沉默。我们使用 CRT v1.2 软件对基因组进行 CRISPR 预测。

3.5 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 v1.0.4 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 v2.2.0 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。

3.6 基因岛预测

基因岛可与多种生物功能相关，如共生关系和发病机理，生物体的适应性等。基因岛基于其功能的不同可以划分为不同的子类，如病原性基因岛(pathogenicity island (PAIs))与发病机理相关，抗生素抗性岛包含许多抗生素抗性基因。相同的基因岛能在近缘物种上发生各种的水平基因转移。可通过比较分析来识别，例如系统发育分析。在细菌中，很多三型分泌系统和四型分泌系统都位于基因岛区域中。这些基因岛通常都在 10 kb 大小以上，与 tRNA 编码基因相关，GC 含量也与基因组其它序列有所差异。很多基因岛两边存在重复序列结构，包含一些其它的例如噬菌体或质粒的小元件。一些基因岛可以自主从染色体上脱离并转移到其它的序列上。我们使用软件 IslandPath-DIMOB v2.0 对细菌基因组进行基因岛预测。

3.7 前噬菌体预测

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（prophage）。基因组上带有前噬菌体的菌称为溶源菌，它们具有无需由外部感染而可产生噬菌体的遗传能力，并且这种能力可传递给后代。如果提供适当条件打破保持前噬菌体状态的机制，噬菌体基因组即变为

可增殖型而进行自主增殖，并使细胞裂解。前噬菌体序列的存在可能也会允许一些细菌获取抗生素抗性，增强对环境的适应性，提高粘附力或使细菌成为致病菌。同时，通过前噬菌体的研究可能找到特异的抗生素甚至是先进的癌症治疗方法。通过软件 PhiSpy v2.3 预测前噬菌体。

3.8 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 v5.0.0 鉴定和分析细菌和真菌基因组序列中生物合成基因簇（BGC）。

3.9 启动子预测

启动子区域是关键的调控区域，它可以使基因被转录或抑制，但是很难通过实验确定。因此，启动子的计算机识别对于引导实验工作和确定控制基因转录起始的关键区域至关重要。在此分析中，虽然启动子区域通常不如侧翼区域稳定，但它们的平均自由能随侧翼基因组序列的 GC 组成而变化。因此获得了一组自由能阈值，使用工具 PromPredict v1 对具有不同 GC 含量的基因组 DNA 进行分析，预测微生物基因组中启动子区域。

3.10 旁系同源基因预测

旁系同源在定义上对功能上没有严格要求，可能相似，但也可能并不相似(尽管结构上具一定程度的相似)，甚至于没有功能(如基因家族中的假基因)。旁系同源的功能变异可能是横向加倍后的重排变异或进化上获得了另一功能，其功能相似也许只是机械式的相关，或非直系同源基因取代新产生的非亲缘或远缘蛋白在不同物种具有相似的功能。使用 BLASTP v2.2.29 筛选出旁系同源基因。

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 Nr、KEGG、eggNOG、Swiss-Prot、TrEMBL 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 数据库的功能注释。利用软件 hmmer v3.0 基于 Pfam 数据库进行 Pfam 功能注释。

Nr 数据库的全称是 Non-Redundant Protein Database,是一个非冗余的蛋白质数据库,该数据库含有全面的蛋白序列和注释信息。与使用其他数据库相比，使用 Nr 数据库一般能使基因组中更多的基因具有注释信息，即具有最高的注释比例。但该数据库中很多蛋白序列和注释信息未经过验证,可靠性有待提高。

GO (gene ontology)是基因本体联合会(Gene Onotology Consortium)所建立的数据库，旨在建立一个适用于各种物种的，对基因和蛋白质功能进行限定和描述的，并能随着研究不断深入而更新的语义词汇标准。目前基因的相同功能在不同的功能数据库可能会使用不同的术语，这样导致使用多种数据库注释时的分歧，不利于功能注释的长期发展和使用。Gene Ontology 就是为了解决上述问题而建立的，使各种数据库中基因产物功能描述相一致。GO 数据库对于生物功能逐渐深入的倒树根形结构，最高级别功能节点包括三个：（1）细胞组分（Cellular Component）：用于描述亚细胞结构、位置和大分子复合物；（2）分子功能（Molecular Function）：用于描述基因、基因产物个体的功能；（3）生物过程（ Biological Process）：用来描述基 因编码的产物所参与的生物过程。

KEGG (Kyoto Encyclopedia of Genes and Genomes)是收集了生物的基因组、通路和化合物信息的综合性的数据库。在 kegg 中有一个“专有名词”KO（kegg Orthology），它是蛋白质（酶）的一个分类体系，将序列高度相似并且在同一条通路上有相似功能的蛋白质序列归为一组，然后打上 KO 标签，对应每个 KO 标签添加相应的功能注释信息。而这些 KO 标签对应到基因组在 kegg 中对于许多代谢通路或者细胞过程。

eggNOG 是收录生物直系同源基因簇的数据库，是在 COG 数据库的基础上的持续更新。构成每个直系同源基因簇(Cluster of Orthologous Groups)的蛋白都是被假定为来自于一个祖先蛋白，具有相同的功能。该数据库是通过比较完整基因组的蛋白序列而产生的。该数据库常被用来对新测序基因组的基因进行分类和注释。

Pfam 数据库是一种包含注释信息和多序列比对信息的蛋白家族数据库，其中的多序列比对信息是由隐马尔科夫模型产生。该数据库提供了较为完整和精确的蛋白家族和功能域的分类信息。不像通常进行的 BLAST 搜索，Pfam 使用基于隐马尔科夫模型的软件 hmmer 进行未知功能序列和 Pfam 数据库的比较，这种方法赋予了在保守位点更高的权重（给予更高的比对得分），这样可以更好地检测到较远的同源蛋白，有益于对较少注释过的生物的基因组进行注释。

SwissProt 数据库是一个人工注释的非冗余高质量蛋白序列数据库，其特点是注释结果有相应实验验证，可靠性较高。

TrEMBL 数据库是对 SwissProt 数据库的扩充，在 SwissProt 数据库的基础上增加了通过计算得到注释信息的大量蛋白序列。

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）、抗生素抗性基因数据库（CARD）、毒力因子数据库（VFDB）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（CAZyme）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 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 CARD 数据库注释

CARD(Comprehensive Antibiotic Research Database)是一个持续更新的抗生素抗性相关信息的数据库。CARD 包含描述抗生素和它们的靶标，涉及抗生素抗性基因、相关蛋白、抗生素抗性机制信息。在 CARD 的核心是一个高度开发的抗生素抗性本体（Antibiotic Resistance Ontology (ARO)），用于抗生素抗性基因数据的分类。

4.2.4 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。另外，还收录了抗真菌化合物及其靶基因。

4.2.5 VFDB 数据库注释

VFDB（virulence factor database）毒力因子数据库，用于识别细菌中含有的毒力因子。

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 v4.0 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。

5 基因组图谱

5.1 基因组圈图

利用已预测得到的基因组信息，如重复序列、GC 含量等，应用软件 Circos v0.66 绘制基因组圈图。Circos 可视化基因组，可以更清晰的探索基因组组件或位置之间的关系。

5.2 基因组图谱展示工具

将基因组中的基因、重复序列、非编码 RNA、基因岛、前噬菌体等元件线性展示在网页版文件中，可以使用 Chrome 浏览器打开，方便地查询各基因组元件的序列、类型、功能等。

5.3 基因组图谱

将基因组、编码基因、非编码 RNA 及功能注释信息整合到 gbk 格式文件。