转录组材料方法中文版

一、样品收集和准备

1、RNA 提取与检测

①琼脂糖凝胶电泳：分析样品 RNA 完整性及是否存在 DNA 污染；

②NanoPhotometer spectrophotometer：检测 RNA 纯度（OD260/280 及 OD260/230 比值）；

③Agilent 2100 bioanalyzer：精确检测 RNA 完整性。

2、文库构建与质检

建库起始 RNA 为 total RNA，总量>=1ug。建库中使用的建库试剂盒为 Illumina 的 NEBNext® UltraTM RNA Library Prep Kit。通过 Oligo(dT)磁珠富集带有 polyA 尾的 mRNA，随后在 NEB Fragmentation Buffer 中用二价阳离子将得到的 mRNA 随机打断。以片段化的 mRNA 为模版，随机寡核苷酸为引物，在 M-MuLV 逆转录酶体系中合成 cDNA 第一条链，随后用 RNaseH 降解 RNA 链，并在 DNA polymerase I 体系下,以dNTPs 为原料合成 cDNA 第二条链。纯化后的双链 cDNA 经过末端修复，经过末端修复、加 A 尾并连接测序接头，用 AMPure XP beads 筛选 250~300 bp 左右的 cDNA，进行 PCR 扩增并再次使用 AMPure XP beads 纯化 PCR 产物，最终获得文库。

文库构建完成后，先使用 Qubit2.0 Fluorometer 进行初步定量，稀释文库至 1.5ng/ul，随后使用 Agilent 2100 bioanalyzer 对文库的 insert size 进行检测，insert size 符合预期后，qRT-PCR 对文库有效浓度进行准确定量（文库有效浓度高于 2nM），以保证文库质量。

3、上机测序

库检合格后，把不同文库按照有效浓度及目标下机数据量的需求 pooling 后进行 Illumina 测序, 并产生150 bp 配对末端读数。测序的基本原理是边合成边测序（Sequencing by Synthesis）。在测序的 flow cell 中加入四种荧光标记的 dNTP、DNA 聚合酶以及接头引物进行扩增，在每一个测序簇延伸互补链时，每加入一个被荧光标记的 dNTP 就能释放出相对应的荧光，测序仪通过捕获荧光信号，并通过计算机软件将光信号转化为测序峰，从而获得待

测片段的序列信息。

二、数据分析

1、数据质控

测序片段被高通量测序仪测得的图像数据经 CASAVA 碱基识别转化为序列数据（reads），文件为 fastq格式，其中主要包含测序片段的序列信息以及其对应的测序质量信息。测序获得的原始数据中包含少量带 有测序接头或测序质量较低的 reads。为了保证数据分析的质量及可靠性，需要对原始数据进行过滤。主要包括去除带接头(adapter)的 reads、去除含 N(N 表示无法确定碱基信息)的 reads、去除低质量 reads(Qphred <=20 的碱基数占整个 read 长度的 50％以上的 reads)。同时，对 clean data 进行 Q20，Q30 和 GC 含量计算。后续所有分析均是基于 clean data进行的高质量分析 。

2、序列比对到参考基因组

直接从基因组网站下载参考基因组和基因模型注释文件。使用HISAT2 v2.0.5 构建参考基因组的索引，并使用 HISAT2 v2.0.5 将配对末端 clean reads 与参照基因组比对。我们选择 HISAT2 作为比对工具，因为HISAT2可以基于基因模型注释文件生成拼接连接的数据库，因此比其他非拼接比对工具有更好的比对效果。

3、新转录本预测

采用 StringTie（1.3.3b）（Mihaela Pertea.et al. 2015）进行新基因预测。StringTie 应用网络流算法以及可选的从头组装来拼接转录本。相对于 cufflinks 等软件，stringtie 有以下优势：(1)拼接出更完整的转录本;(2)拼接处更准确的转录本；(3)更好的估计转录本的表达水平;(4)拼接速度更快。

4、基因表达水平定量

featureCounts（1.5.0-p3）用于计算映射到每个基因的读数。然后根据基因的长度计算每个基因的 FPKM，并计算映射到该基因的读数。 FPKM 指每百万碱基对测序的转录本序列片段的每千碱基片段的预期数量。同时考虑了测序深度和基因长度对读数计数的影响，并且是当前最常用于估计基因表达水平的方法。

5、差异表达分析

使用DESeq2 软件（1.16.1）进行两个比较组合之间的差异表达分析（每个组两个生物学重复）。DESeq2 提供了统计程序，用于使用基于负二项式分布的模型来确定数字基因表达数据中的差异表达。使用 Benjamini 和 Hochberg 的方法来调整所得 P 值以控制错误发现率。通过 DESeq2 发现调整的 P 值<0.05 的基因被分配为差异表达的。（对于没有生物学重复的采用 edgeR ）在进行差异基因表达分析之前，对于每个测序文库，通过一个比例归一化因子通过 edgeR 程序包调整读取计数。两个条件的差异表达分析使用 edgeR 软件包（3.18.1）进行。使用 Benjamini＆Hochberg 方法调整 P 值。校正后的P 值以及|log2foldchange|作为显著差异表达的阈值。

6、差异基因富集分析

通过 clusterProfiler （3.4.4）软件实现差异表达基因的 GO 富集分析，其中修正了基因长度偏差。考虑具有小于 0.05 的校正的 P 值的 GO term 通过差异表达基因显著富集。KEGG 是一个数据库资源，用于从分子水平的信息，特别是基因组测序产生的大规模分子数据集和其他高通量数据库中了解生物系统的高级功能和效用，如细胞，生物体和生态系统等。我们使用 clusterProfiler（3.4.4）软件分析 KEGG 通路中差异表达基因的统计富集。

7、差异基因蛋白网络互作分析

差异表达基因的 PPI 分析基于已知和预测蛋白质-蛋白质相互作用的 STRING 数据库。对于数据库中存在的物种，我们通过从数据库中提取目标基因列表来构建网络;否则，使用diamond（0.9.13）将目标基因序列与选择的参考蛋白序列进行比对，然后根据所选参考物种的已知相互作用建立网络。

8、可变剪切分析

选择性剪接是调控基因表达和蛋白质变量的重要机制。使用 rMATS（3.2.5）软件分析 AS 事件，主要包括 SE、RI、MXE、A5SS、A3SS 五种可变剪接事件。

9、SNP 分析

使用 GATK（3.7）软件对样本数据进行变异位点分析，并用 SnpEff（4.3q）软件对变异位点进行注释。

10、WGCNA分析

加权基因共表达网络分析 (WGCNA, Weighted correlation network analysis)是用

来描述不同样品之间基因关联模式的系统生物学方法，可以用来鉴定高度协同变化的基因集,并根据基因集的内连性和基因集与表型之间的关联鉴定候补生物标记基因或治疗靶点。R包WGCNA是用于计算各种加权关联分析的功能集合，可用于网络构建，基因筛选，基因簇鉴定，拓扑特征计算，数据模拟和可视化等。WGCNA适用于多样品数据模式，一般要求样本数多于15个，样本数多于20时效果更好，样本越多，结果越稳定。输入数据文件一个是样本信息的文件，即描述样本性状特征的矩阵：用于关联分析的性状必须是数值型特征；如果是区域或分类变量，需要转换为0-1矩阵的形式。另外一个是基因表达数据的文件，对于转录组测序即可以用基因表达的FPKM数据。