医口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

Agilent 2100 bioanalyzer：精确检测RNA完整性和总量。

2、文库构建与质检

建库起始RNA为total RNA，总量>=1ug。通过Oligo(dT)磁珠富集带有polyA 尾的mRNA，随后在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、加A 尾并连接测序接头，用AMPure XP beads 筛选370~420 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、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。

4、差异表达分析

使用DESeq2 软件（1.20.0）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（3.22.5）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。

5、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。Reactome数据库汇集了人类等模式物种各项反应及生物学通路。

Reactome数据库以小于0.05 的校正的P 值作为显著性富集的阈值。DO(Disease Ontology)是描述人类基因功能与疾病相关的数据库。DO富集以小于0.05 的校正的P 值作为显著性富集的阈值。DisGeNET数据库整合了人类疾病相关基因。DisGeNET富集以小于0.05 的校正的P 值作为显著性富集的阈值。我们使用clusterProfiler（3.4.4）软件分析Reactome通路，DO通路以及DisGeNET通路中差异表达基因的统计富集。

6、基因集富集分析

基因集富集分析(Gene Set Enrichment Analysis, GSEA)不需要指定明确的差异基因阈值，使用预先定义的基因集（通常来自功能注释或先前实验的结果），将基因按照在两类样本中的差异表达程度排序，然后检验预先设定的基因集合是否在这个排序表的顶端或者底端富集。基因集富集分析检测基因集合而不是单个基因的表达变化，因此可以包含这些细微的表达变化，预期得到更为理想的结果。我们使用本地版GSEA分析工具http://www.broadinstitute.org/gsea/index.jsp，分别对该物种的GO、KEGG、Reactome、DO、DisGeNET数据集进行GSEA分析。

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。

10、融合基因分析

融合基因是指两个基因的全部或部分序列融合而成的嵌合基因，一般由染色体易位、缺失等原因所致。我们使用STAR-Fusion 软件进行融合基因的检测，STAR-Fusion（1.2.0）是利用STAR 比对的融合输出结果来检测融合转录本的软件包，分为SATR比对，STARFusion. predict，STAR-Fusion.filter 等步骤，并利用校验工具FusionInspector对STAR-Fusion的预测结果进行校正，以保证融合基因结果的准确性。

11、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。