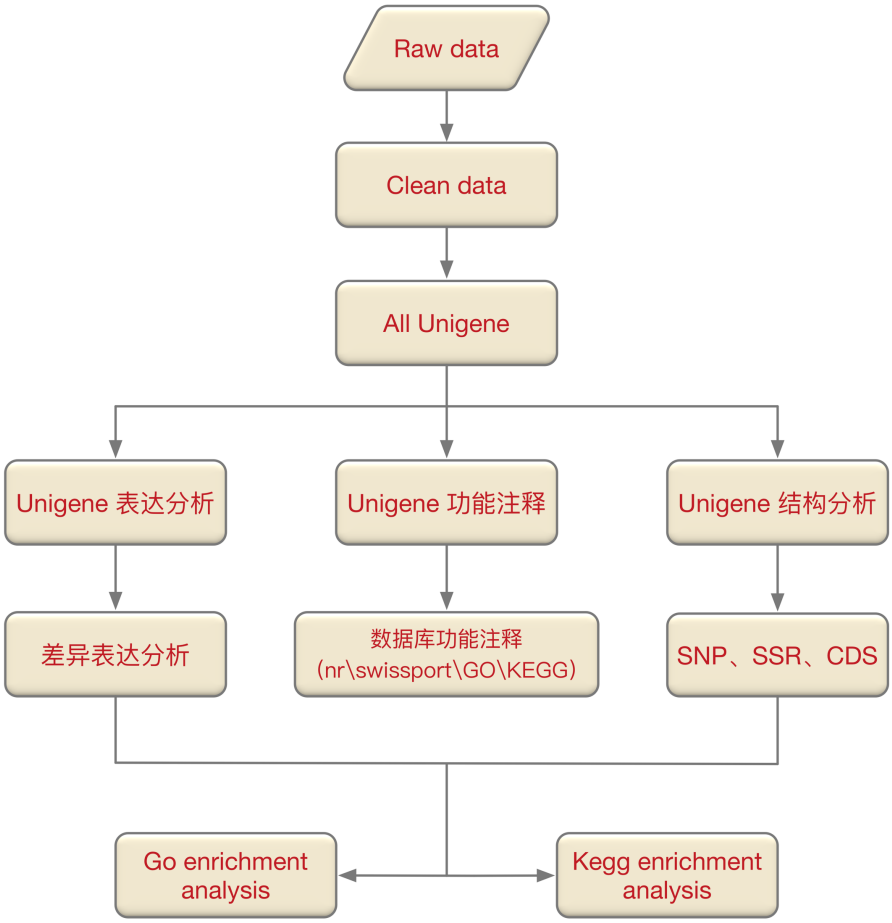


图 1 无参转录组数据分析流程图

## 数据分析流程

1) 对高通量测序原声下机数据（Raw data）进行质量过滤，去除接头序列以及低质量的 Reads 获得高质量的 Clean data。

2) 基于 Clean Data 使用 trinity 软件进行组装，组装方式分为分开组装与合并组装：

   - a:分开组装-----将不同处理间的样本进行分别组装并将分别组装合并，借助 tgicl 软件对合并的结果去除冗余，得到 Unigene 序列。随后对 Unigene 进行定量，去除(TPM<1)低表达量后得到最终的 Unigene 序列。
   - b:合并组装-----将所有样本的 clean data 的数据进行合并组装，将组装结果定量，去除(TPM<1)低表达量后得到最终的 Unigene 序列。

3) 基因定量分析：

   - 合并组装情况下：使用 bowtie 与 RSEM 将 clean data 与 Unigene 序列进行比对定量
   - 分开组装情况下：使用 kallisto 将 clean data 与 Unigene 序列进行比对比对定量

4) 差异表达分析：借助 R 软件包 edgeR 进行不同处理间差异表达分析

5) 功能注释：

   - 使用 BLAST 软件将 Unigene 序列与 NR、Swiss-Prot、GO、COG/KOG、KEGG 数据库比对，获得 Unigene 的注释信息。
   - 使用 KOBAS2.0 得到 Unigene 在 KEGG 中的 KEGG Orthology 结果。

- 使用 HMMER 软件与 Pfam 数据库比对，获得 Unigene 的注释信息。

6) 基因结构预测：

- CDS 编码区预测：TransDecoder
- SSR 预测分析：MISA
- SNP 预测分析：SAMtools 与 GATK

# 参考文献


**1: assembly[trinity]**

Grabherr M G, Haas B J, Yassour M, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data[J]. Nature biotechnology, 2011, 29(7): 644.

Haas B J, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis[J]. Nature protocols, 2013, 8(8): 1494-1512.

**2: kallisto**

Bray N L, Pimentel H, Melsted P, et al. Near-optimal probabilistic RNA-Seq quantification[J]. Nature biotechnology, 2016.

**3: RSEM**

Li B, Dewey C N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome[J]. BMC bioinformatics, 2011, 12(1): 1.

**4: bowtie**

Langmead B, Trapnell C, Pop M, et al. Bowtie: an ultrafast memory-efficient short

read aligner[J]. Genome Biol, 2009, 10(3): R25.

**5: tgicl**

Pertea G, Huang X, Liang F, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets[J]. Bioinformatics, 2003, 19(5): 651-652.

**5: bioinformatics workflow**

Yang I S, Kim S. Analysis of Whole Transcriptome Sequencing Data: Workflow and Software[J]. Genomics & informatics, 2015, 13(4): 119-125.

Conesa A, Madrigal P, Tarazona S, et al. A Survey of Best Practices for RNA-seq Data Analysis[J]. 2016.

**6: SNP calling**

http://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq

**7: DGE analysis**

Li P, Piao Y, Shon H S, et al. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data[J]. BMC bioinformatics, 2015, 16(1): 1.

Schurch N J, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? [J]. RNA, 2016, 22(6): 839-851.

Anders S, McCarthy D J, Chen Y, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor[J]. Nature protocols, 2013, 8(9):

1765-1786.

**备注：实际生物信息分析内容以项目合同为准。**