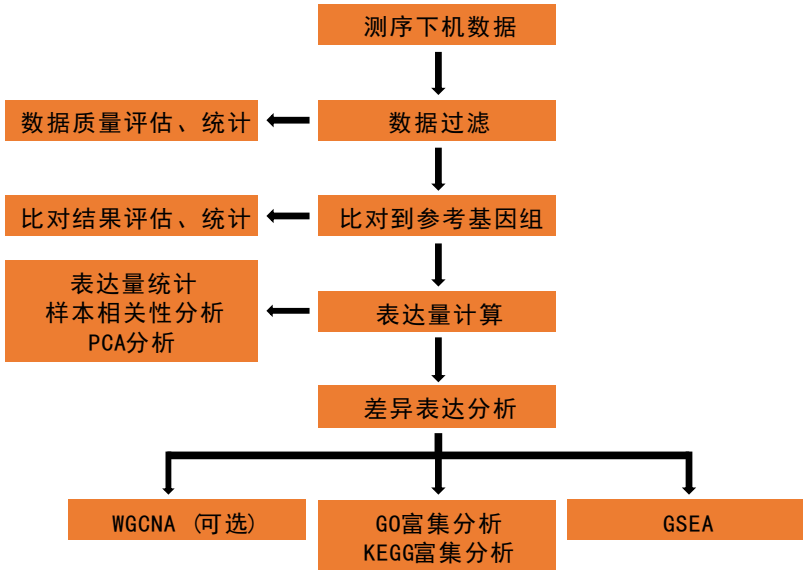


转录组分析流程

Written by 王鹏飞

Email: wangpf0608@126.com

一 分析流程图



二 分析流程及结果

1 取样、mRNA 提取、建库及测序

参考公司报告。

2 数据过滤

使用 `fastp`^[1] (version: 0.20.0) 对 raw data 进行过滤得到 clean data。统计过滤前后 total bases、total reads、Q30、Q20、GC content 以及有效数据比率 (data_stat.csv/txt)，同时使用 FastQC (version: 0.11.9) 对过滤前后的数据进行质量评估 (QC/sample_fastqc.html)。

3 比对到参考基因组

使用 HISAT2^[2] (version: 2.1.0) 将 clean reads 比对到甘蓝型油菜 ZS11 参考基因组^[3]上, 得到 SAM (Sequence Alignment/Map) 格式文件, 然后使用 SAMtools (version: 1.9) 对比对结果 (SAM 文件) 按照染色体和位置进行排序并转换为 BAM (Binary Alignment/Map) 格式文件^[4], 可以将 BAM 文件导入 IGV (Integrative Genomics Viewer)^[5] 对比对结果进行可视化。HISAT2 可以使用更少资源的同时具有更快的速度, HISAT2 比对时, 对于非链特异性文库使用默认参数, 链特异性文库需要指定文库类型 (first 使用 --rna-strandness RF, second 使用 --rna-strandness FR)。比对完成后, 我们对比对结果进行评估, 统计比对率和唯一比对率。

4 表达量计算

根据比对结果 (BAM 文件), 我们使用 R (version: 4.0.2) 软件的扩展包 Rsubread^[6] (version: 2.2.6) 中的 featureCounts 函数计算每个基因的表达量 (read count) 并进行归一化处理 (normalization), 得到 TPM (Transcripts Per Kilobase of exon model per Million mapped reads) 和 TMM (trimmed mean of M value) 表达矩阵。

5 差异表达分析

根据基因表达矩阵 (read count) 文件, 在有生物学重复的情况下, 使用 R (version: 4.0.2) 软件的扩展包 DESeq2^[7] (version: 1.28.1) 进行差异表达分析, 在没有生物学重复的情况下, 则使用 R 扩展包 edgeR^[8] (version: 3.30.3) 进行差异表达分析, 并推荐测生物学重复, 也不算太贵。对于 log2FoldChange 绝对值大于 1, 并且 padj 小于 0.05 的基因则认为是差异表达基因 (阈值需根据实际情况做出调整)。

6 功能富集分析

根据筛选出的差异表达基因, 我们使用 R (version: 4.0.2) 软件的扩展包 clusterProfiler^[9] (version: 3.16.1) 依据超几何分布检验来完成 GO 和 KEGG 富集分析 (Over-representation analysis), 设置参数 pvalueCutoff 和 qvalueCutoff 为 0.05 筛选显著富集的 GO/KEGG term。在绘图时, 如果 GO 或 KEGG 的 term 太多 (一般是 GO), 建议取前 10 或 15 个 term (GO 中 CC、BP 和 MF 各选 10 或 15 各)

进行绘图，如果相关 term 不在前 10 或 15 个内，也可以手动添加。

同时，我们使用 clusterProfiler^[9]进行 GSEA (Gene Set Enrichment Analysis)，其基本思想是将基因按照两组样本中差异表达程度排序，使用预先定义的基因集 (GO 或 KEGG)，检验预先定义的基因集是否排列在顶端或底端。GSEA 可以充分利用基因差异表达程度的信息，排除人为筛选差异表达基因的主观性，是一种更先进富集方法，与 over-representation 富集方法互为补充。

7 WGCNA

加权基因共表达网络分析 (WGCNA, Weighted correlation network analysis) 可以用来鉴定样本间高度协同变化的基因集 (模块)，同时可以根据模块特征值 (eigengene) 将模块与外部性状信息相关联，以此鉴定与性状相关的模块并进一步挖掘关键基因^[10,11]。进行 WGCNA 至少需要 15 个样本，最好是 20 个及以上。在这里我们筛选差异表达基因使用 R (version: 4.0.2) 软件的扩展包 WGCNA^[11] (version: 1.69) 进行基因模块的构建以及模块-样本、模块-形状的关联 (其中各步骤参数均需根据实际情况决定)。

参考文献

- 1 Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor[J]. Bioinformatics, 2018, 34(17).
- 2 Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett, Steven L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype[J]. Nature Biotechnology: The Science and Business of Biotechnology, 2019, 37(8).
- 3 Jia-Ming Song, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, Bo Wang, Shaoping Lu, Run Zhou, Wen-Zhao Xie, Yuanfang Cheng, Yuting Zhang, Kede Liu, Qing-Yong Yang, Ling-Ling Chen, Liang Guo. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*[J]. Nature Plants, 2020, 6(1).
- 4 Li Heng, Handsaker Bob, Wysoker Alec, Fennell Tim, Ruan Yue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard. The Sequence Alignment/Map format and SAMtools.[J]. Bioinformatics (Oxford, England), 2009, 25(16).
- 5 Robinson James T, Thorvaldsdóttir Helga, Winckler Wendy, Guttman Mitchell, Lander Eric S, Getz Gad, Mesirov Jill P. Integrative genomics viewer.[J]. Nature biotechnology, 2011, 29(1).
- 6 Liao Yang, Smyth Gordon K, Shi Wei. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads[J]. Narnia, 2019, 47(8).
- 7 Michael I Love, Wolfgang Huber, Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. Genome Biology, 2014, 15(12).
- 8 Mark D. Robinson, Davis J. McCarthy, Gordon K. Smyth. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics, 2010, 26(1).

- 9 Yu Guangchuang, Wang Li-Gen, Han Yanyan, He Qing-Yu. clusterProfiler: an R package for comparing biological themes among gene clusters.[J]. Omics : a journal of integrative biology, 2012, 16(5).
- 10 Bin Zhang, Steve Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis[J]. Statistical Applications in Genetics and Molecular Biology, 2005,4(1).
- 11 Peter Langfelder, Steve Horvath. WGCNA: an R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9(2).