RNA-seq 分析转录因子 mEmBP-1a 过表达对水稻的影响

刘宇杰(171240517)，高志伟(171240503)

(南京大学匡亚明学院　江苏 南京　210023)

摘要：本文利用 GEO 数据库中过表达玉米转录因子 mEmBP-1a 的水稻株系的转录组测序数据，利用 RNA-seq 数据处理的 limma、edgeR 等 R 包和 DAVID、KOBAS 等在线工具进行分析，对差异表达基因进行 GO 功能注释和 KEGG 富集通路分析。分析结果表明，与物质运输、氧气反应及相关代谢通路有关的基因表达显著上调，初步解释了光合作用效率提高的机制。

关键词：RNA-seq；差异表达基因；GO 和 KEGG；光合作用

一、研究背景

RNA-seq 技术又称转录组测序技术，指用高通量测序技术对全部或部分转录产物进行测序分析。其实验的主要流程为：提取样品总 RNA，富集 mRNA 并打断为小片段，反转录为 cDNA，再进行 PCR 扩增，最后上机测序。得到高通量测序结果后，需对原始数据进行处理和分析。首先通过质量控制、序列比对和表达定量进行数据预处理，再进行差异基因分析，最后根据需求进行功能分析、通路富集、共表达网络构建等后续处理。

对于 RNA-seq 数据，来自 Bioconductor 项目的 edgeR 和 limma 等 R 包提供了一套完善的统计学处理方法[1]。edgeR 中的归一化算法考虑了可能出现的异常高表达值的情况，其核心思想是表达量居中的基因或转录本在所有样本中表达量都应该是相似的。edgeR 采用 TMM 校正方法，在去除高表达和高差异基因后计算加权系数，使剩余基因在校正后差异倍数尽可能小。此类算法的校正结果较为稳定，可使差异表达分析结果更为可靠。但缺点是没有校正基因长度的影响，且选取不同样本比较时会得到不同的表达值，不利于共表达等整合分析[2]。

对于作物而言，提高光合作用效率是提高作物单产能力的主要途径。迄今为止，大多数作物的增产方式是操纵一个或几个与光合作用相关的基因。目前的研究发现，一种来源于玉米的转录因子 mEmBP-1 的过表达会导致光合作用途径中几乎所有基因的表达同时增加，如叶绿素 a 和 b 的结合蛋白，Rubisco，GAPDH，

FBA，TK 和 PRK 等[3]。这种光合作用相关基因表达的增加将导致在温室和田间植物的光合效率均有增加，谷物产量提高。

本文分析所用数据的提供者利用 DNA-蛋白质结合实验，发现 mEmBP-1a 蛋白可以直接与光合作用基因的启动子区域结合，从而上调这些基因的表达[3]。本文利用其所提供的 RNA-seq 原始数据，通过多种分析手段，对比四个过表达 mEmBP-1a 的水稻样品和四个野生型水稻的转录组测序结果，揭示该转录因子对水稻基因表达的影响，从而解释光合作用效率提高的原因。

二、方法与数据

1. 数据来源

分析所用数据下载自 NCBI 网站 Gene Expression Omnibus (GEO)数据库，序列号为 GSE143259，具体网址为 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143259。

数据上传者为中国科学院上海生命科学研究院的 Faming Chen 课题组。为进行 RNA-seq 分析，该组在田间生长的抽穗初期的植物中，于上午 10:00 左右收集 mEmBP-OE-4 系和野生型的顶部第二片叶子。分别收集了 mEmBP-OE-4 系(mEmBP-OE-4_rep1-4)和 WT(WT_rep1-4)的四个生物重复样本。按流程使用 Life Technologies Corporation 的 PureLink RNA Mini Kit 提取 RNA。构建利用 Illumina 测序的 RNA 文库，并使用 Illumina X Ten 平台以双末端 150 模式测序[3]。

笔者下载的数据形式为八个样品 mRNA 对应基因的计数原始数据。

2. 数据分析方法

数据分析主要分为两部分。

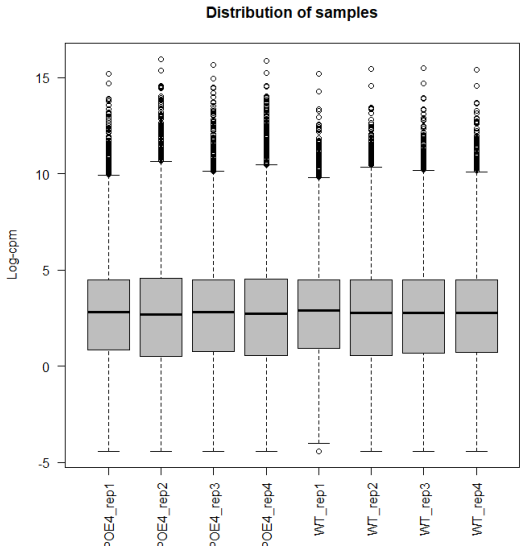
第一部分主要源自 WEHI workshop 的 RNA-seq 教学内容及其具体处理流程的参考文献[1]，修改后的代码见附录部分。

第二部分参考陈家辉、任学义等[4]的分析方法和相关教程[5]，利用在线 DAVID 软件(http://david.abcc.ncifcrf.gov/home.jsp)[6]进行显著差异表达基因的 GO 功能注释和 Kyoto Encyclopedia of Genes and Genomes(KEGG)富集通路分析，并参考相关教程[7]利用 KOBAS(http://kobas.cbi.pku.edu.cn/kobas3/?t=1)[8]再次进行 KEGG 分析以验证结果。

三、结果与讨论

1. 第一部分：WEHI RNA-seq workshop 方法演练

八个样品测序结果原始数据共 36850 个基因片段，去除八个样本中均为零表达的基因 7687 个，低表达的基因 8473 个，剩余 20690 个基因。对剩余基因进行归一化处理，使所有样品的 log-CPM 分布类似，如图 1 所示。



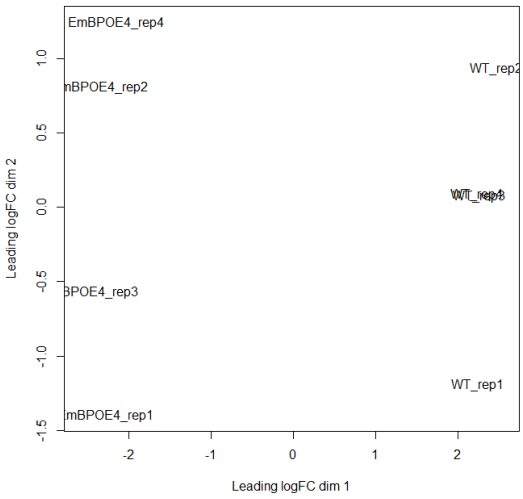
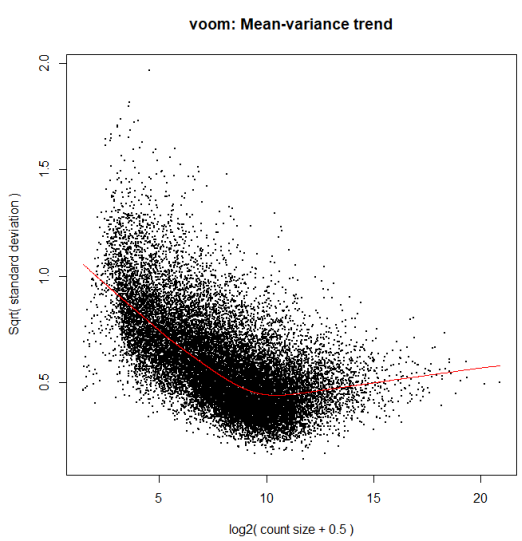


图 1. 八个样品归一化处理后的 log-CPM 图像　　图 2. 八个样品无监督聚类处理结果

归一化完成后，对样本进行无监督聚类，结果如图 2 所示。可看出在维度 1 上，野生型样品和 mEmBP-1a 过表达样品可以很好地按照实验分组聚类，但在维度 2 上聚类效果较差。其中，野生型样品 3 号和 4 号在维度 1 和维度 2 上都十分接近，批次效应很小。

接下来，创建设计矩阵 design matrix 进行对比。由于本文所分析数据仅分为两组，故此过程较为简略。

之后，从表达计数数据中删除异方差。由于文献中显示，当使用原始计数或者其 log-CPM 转化值时，方差与均值并不独立，但使用负二项分布模拟计数的方法有均值方差为二次关系的假设。在 limma 中，假设 log-CPM 符合正态分布，并使用由 voom 函数计算得到的精确权重来调整均值和方差的关系，便可对 log-CPM 进行线性建模[1]。用 voom 函数调整的前后结果分别如图 3(a)(b)所示，变化明显，处理前方差有变化趋势，处理后方差为恒定值。



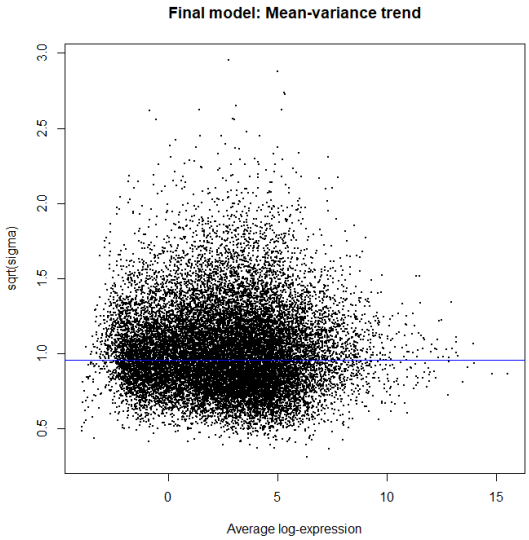
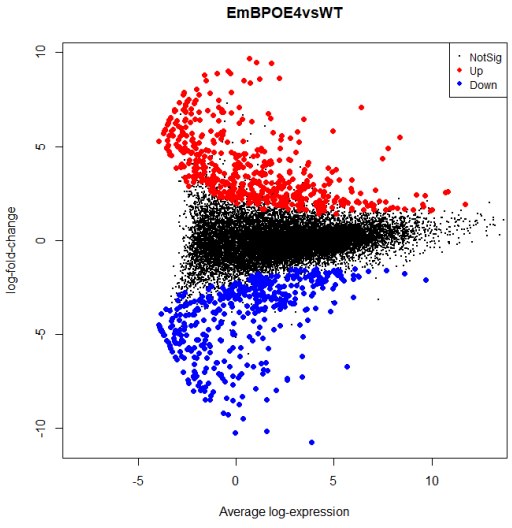


图 3. (a) 处理前均值-方差变化趋势图　　(b)处理后均值-方差变化趋势图

检验差异表达基因数目，得到实验组相较于对照组上调的基因共 4108 个，下调的基因共 4067 个，数目较多。为将研究范围缩小至显著差异基因，此处设置阈值 logFC≥1 (FC = fold change)，最终得到显著上调基因共 509 个，显著下调基因共 406 个。

显著差异表达基因的 Mean-Difference 图像可视化结果如图 4(a)所示。

也可利用 Glimma 包绘制交互式的 MD 图像，如图 4(b)所示。同时可以按照 P 值排序得到差异最显著的 10 个基因，如图 5 所示。



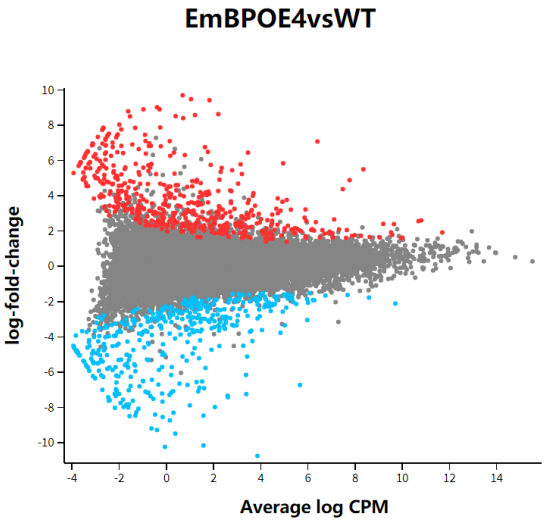


图 4. (a) limma 包绘制的 MD Plot　　(b) Glimma 包绘制的 MD Plot

GeneID	anno	logCPM	logFC	Adj.PValue
Os04g0612500	Os04g0612500	8.351	5.498	0.00001917
Os12g0601800	Os12g0601800	2.791	5.444	0.00001917
Os03g0782200	Os03g0782200	6.407	7.074	0.0000194
Os05g0227600	Os05g0227600	4.722	3.852	0.0000194
Os11g0606400	Os11g0606400	1.405	-6.889	0.0000194
Os11g0639400	Os11g0639400	-2.401	-7.592	0.0000194
Os12g0528801	Os12g0528801	1.297	4.9	0.0000194
Os12g0554100	Os12g0554100	2.346	5.581	0.0000194
Os11g0636050	Os11g0636050	-1.541	8.497	0.00002104
Os02g0151800	Os02g0151800	3.379	-6.159	0.00002952

图 5. 表达差异最显著的 10 个基因

最后，绘制热图展示差异表达最显著的 100 个基因，如图 6 所示。

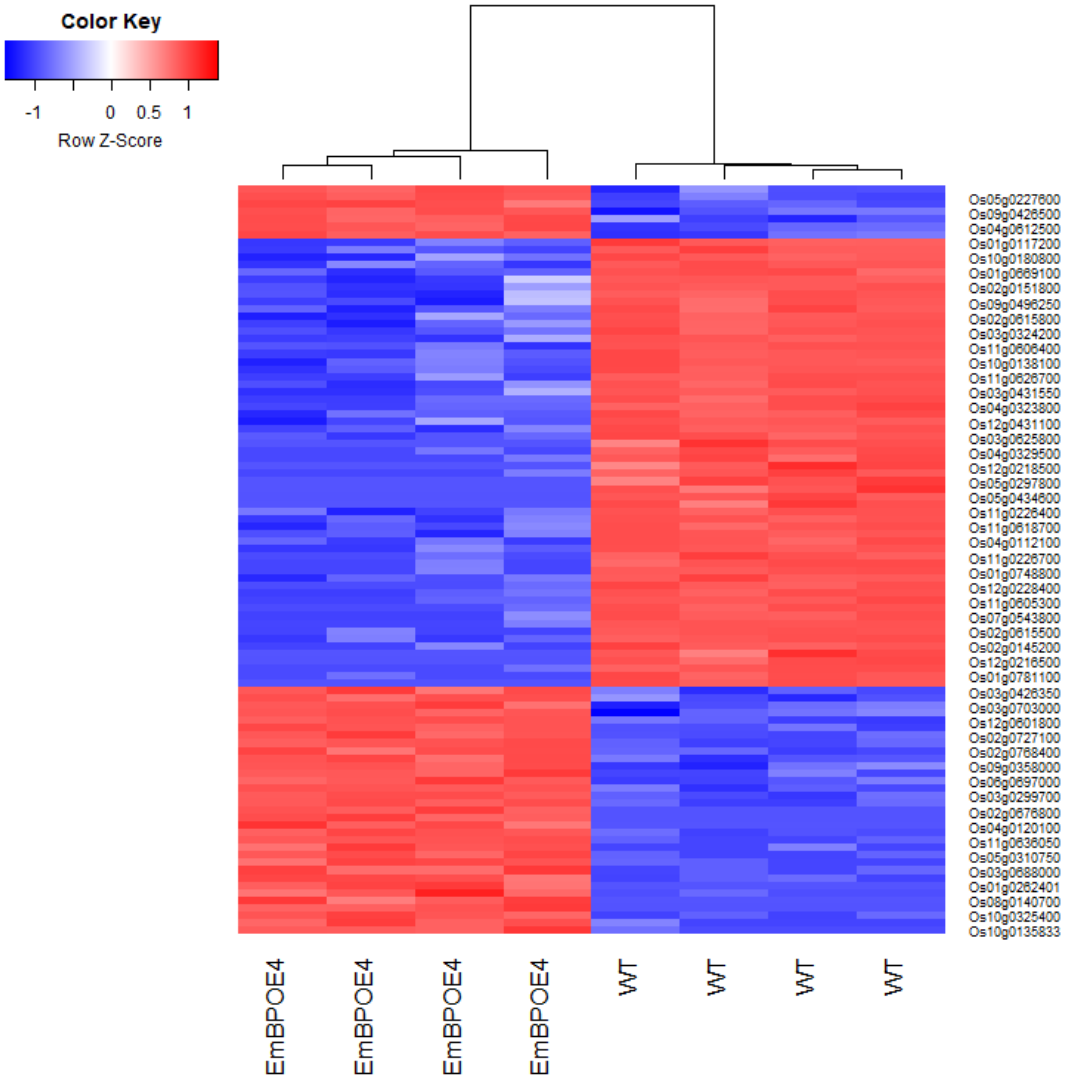


图 6. 用热图展示差异表达最显著的基因

2. 第二部分：GO 功能注释和 KEGG 富集通路分析

利用 DAVID 在线工具进行显著差异表达基因的 GO 功能注释。在 DAVID 中，显著上调的 509 个基因共找到 434 个；显著下调的 406 个基因共找到 263 个。对这些基因的注释结果利用 OriginPro 2020 软件[9]作图，如图 7、8 所示。

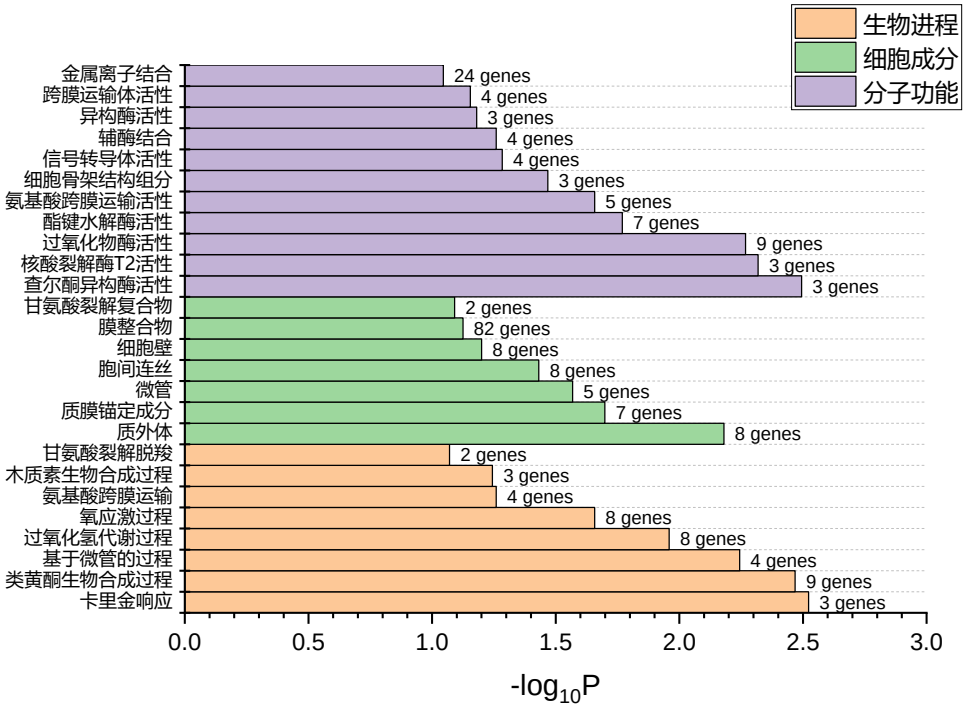


图 7. 显著上调基因的 GO 功能注释

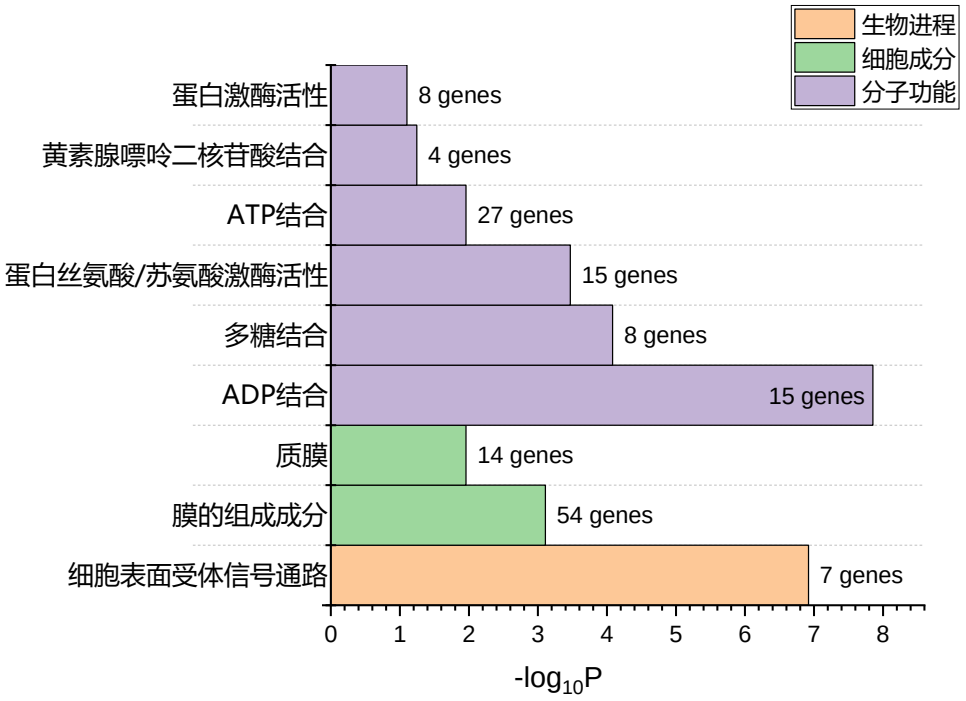


图 8. 显著下调基因的 GO 功能注释

由图可知，GO 功能注释分为生物进程、细胞成分和分子功能三类。

显著上调基因可以很好地解释过表达 mEmBP-1a 对光合作用的促进：金属离子结合、辅酶结合可能有利于叶绿素的合成和稳定；过氧化物酶活性、氧应激过程、过氧化氢代谢过程可以解释细胞对于光合作用产氧增加、活性氧产生量增加的自我保护；与膜整合物相关的基因多达 82 个，与胞间连丝、微管、基于微管的过程等均可解释光合作用增加所对应的物质运输量增加。

显著下调基因 P 值较小且相关基因较多的是 ADP 结合、膜组成成分和细胞表面受体信号通路等。

DAVID 分析结果还提供了显著差异基因中的关键词统计，如图 9、10 所示。显然，mEmBP-1a 的过表达对于细胞信号通路和跨膜蛋白部分的影响最为明显。

接下来进行 KEGG 富集通路分析。由于显著下调基因在 DAVID 所提供的数据库中匹配数目较少，故对其进行 KEGG 分析时，未得到结果。显著上调基因的 KEGG 分析结果如图 11 所示，共有 8 组通路的基因有显著的富集现象，分别是：苯丙素生物合成、代谢通路、次生代谢产物生物合成、氮代谢、氰氨基酸代谢、类黄酮生物合成、蔗糖和淀粉代谢、乙醛酸和二羧酸酯代谢。

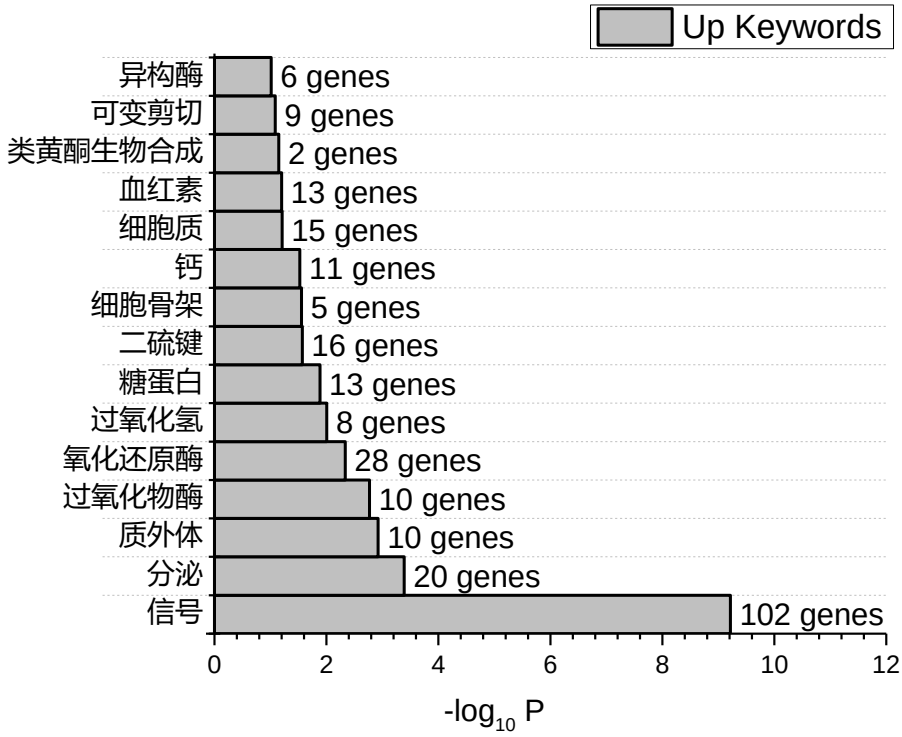


图 9. 显著上调基因 Up Keywords 的 P 值和基因数目

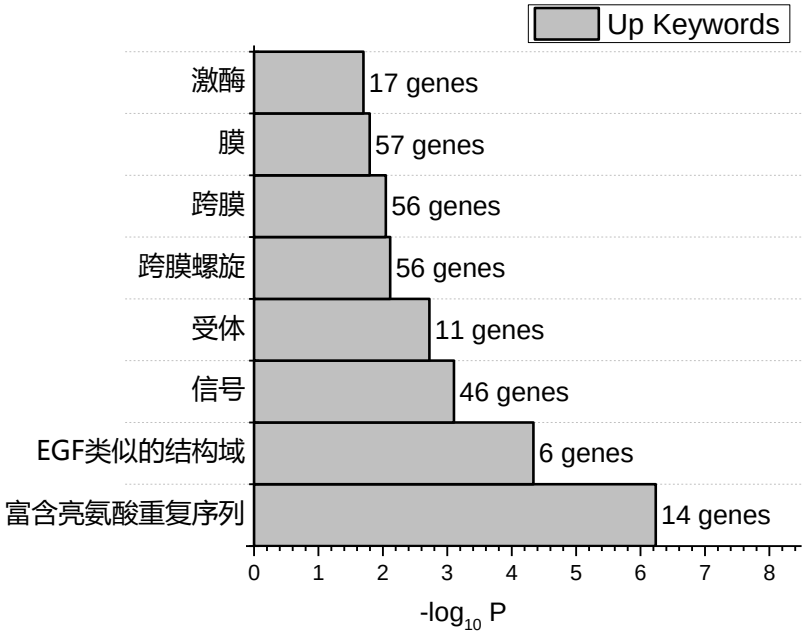


图 10. 显著下调基因 Up Keywords 的 P 值和基因数目

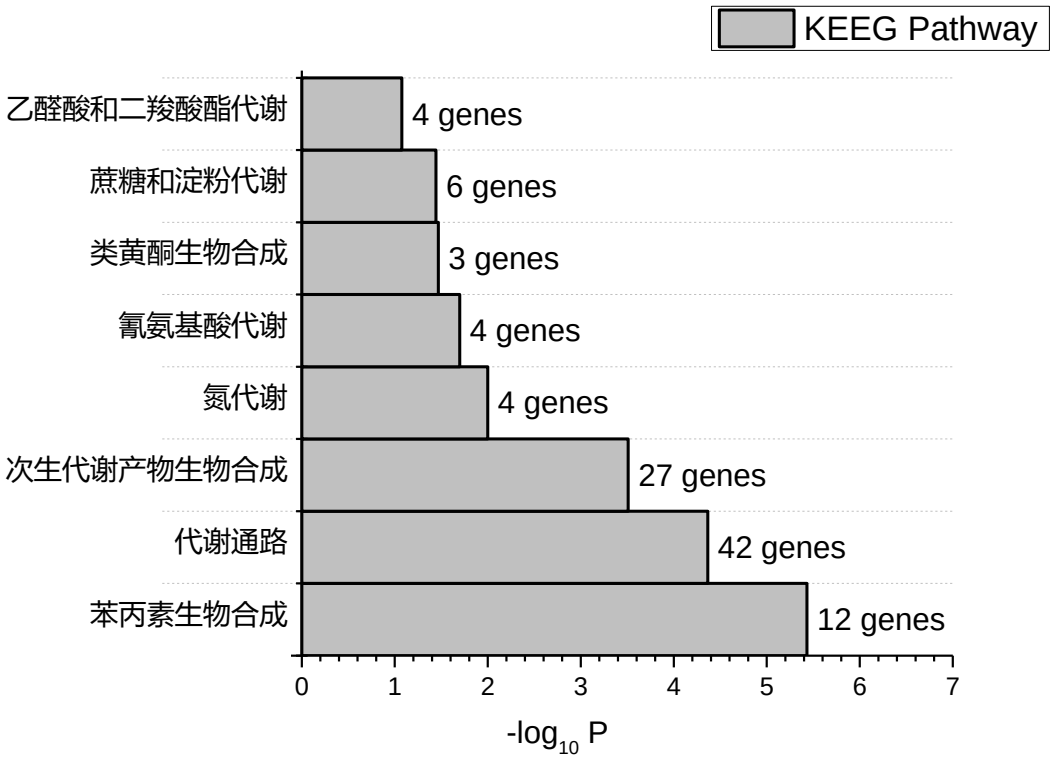


图 11. 显著上调基因的 KEGG 富集通路分析结果

再利用 KOBAS 进行 KEGG 分析以验证结果。显著下调基因仍然没有 P 值小于 0.1 的显著结果，与 DAVID 分析的结果基本吻合。显著上调基因如下图 12 所示，与 DAVID 分析得到的八个结果完全一致，仅校正后的 P 值有大小差别。其中淀粉和蔗糖代谢通路 P 值约为 0.08，稍大于 0.05。

#Term	Database	ID	Input number	Background number	P-Value	Corrected P-Value	Input
Metabolic ...	KEGG PATHW...	osa01100	59	2290	5.43259499...	0.00030574...	9272308|43...
Phenylprop...	KEGG PATHW...	osa00940	14	233	9.55452195...	0.00030574...	4333841|43...
Biosynthes...	KEGG PATHW...	osa01110	35	1177	3.44278105...	0.00073445...	4348690|43...
Flavonoid ...	KEGG PATHW...	osa00941	5	26	6.08021815...	0.00097283...	4334588|43...
Cyanoamino...	KEGG PATHW...	osa00460	5	54	0.00130740...	0.01673477...	4336033|43...
Nitrogen m...	KEGG PATHW...	osa00910	4	34	0.00179871...	0.01918625...	4324249|43...
Starch and...	KEGG PATHW...	osa00500	7	161	0.00885043...	0.08091823...	4333841|43...

图 12. KOBAS 中显著上调基因 KEGG 富集通路分析结果

从 KEGG 分析结果可见，在转录因子 mEmBP-1a 过表达的影响下，差异表达的基因大量富集于与代谢相关的通路中。推测可能由于光合作用合成产物转化为糖类、氨基酸、类黄酮等次生代谢产物及其他物质，上调下游代谢反应水平。

四、总结

本文以 WEHI RNA-seq Workshop 所学知识为出发点，利用 GEO 数据库中最新上传(2020.1.8)的转录组测序原始数据，按课上所学流程进行 RNA-seq 的初步分析，并拓展利用 DAVID 和 KOBAS 在线工具实现了水稻在转录因子 mEmBP-1a 过表达条件下的差异基因鉴定和通路富集分析，初步解释了光合作用效率提高的机制，与数据提供者所描述的实验结果可以很好印证。

本文由匡亚明学院大三本科生刘宇杰和高志伟共同合作完成，两人均参与数据寻找、结果分析、问题处理、报告撰写的整个过程，参与程度一致。在完成该项目的过程中，进一步熟悉了相关生物信息学数据库和工具的使用及结果分析，复习了利用 R 进行 RNA-seq 数据处理的原理和步骤，对相关生物学问题也有了更为深入的理解。

参考文献

[1] Law C W , Alhamdoosh M , Su S , et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR[J]. F1000 Research, 2016, 5:1408.

[2] 陈铭主编. 生物信息学(第三版)[M]. 科学出版社, 2018.

[3] Over-expression of a transcription factor mEmBP-1a increases rice photosynthesis, biomass and yield by regulating the expression of multiple key photosynthesis genes [EB/OL].[2020-01-08].https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143259.

[4] 陈家辉,任学义,李丽敏,卢诗意,程湉,谭量天,梁少东,何丹林,罗庆斌,聂庆华,张细权,罗文.转录组测序揭示细胞周期通路参与鸡腹脂沉积[J].遗传,2019,41(10):962-973.

[5] DAVID——让注释简单起来！[EB/OL].[2017-01-10].http://www.360doc.com/content/17/0110/23/19913717_621651875.shtml.

[6] (a) Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57. (b) Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.

[7] GO、KEGG 富集分析——DAVID 与 KOBAS 在线分析工具[EB/OL]. [2018-10-16]. https://www.jianshu.com/p/85ff36fae727.

[8] Xie, C. et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research 39, W316–W322 (2011).

[9] OriginPro, Version 2020. OriginLab Corporation, Northampton, MA, USA.

附录：数据分析第一部分代码

```
#Load packages
library(limma)
library(edgeR)
library(RColorBrewer)
```

```
#Load data
files=list.files(pattern='txt$')

#Read and Merge a Set of Files Containing Count Data
x=readDGE(files,columns=c(1,2))
class(x)
dim(x)
x

#Add group labels
group=as.factor(c('EmBPOE4','EmBPOE4','EmBPOE4','EmBPOE4','WT','WT','WT','WT'))
x$samples$group=group

#Data pre-processing
#Filterx$genes=rownames(x)ing
table(rowSums(x$counts==0)==8) #Genes with zero counts in all samples
keep.exprs=filterByExpr(x,group=group) #Remove genes with very low counts across all groups
x=x[keep.exprs,,keep.lib.sizes=F]
dim(x)

#Add gene ID
x$genes=rownames(x)

#Normalisation: Remove systematic bias due to technical effects
x=calcNormFactors(x,method='TMM')
x$samples$norm.factors

#Data visualisation
```

```
#Boxplots
lcpm=cpm(x,log=T) #log Counts per Million
boxplot(lcpm,las=2,col='grey',main="Distribution of samples",ylab="Log-cpm")

#MDS plots: Multidimensional scaling plot
plotMDS(lcpm)
plotMDS(lcpm,labels=group)



#Differential expression analysis by limma
#Design matrix
design=model.matrix(~0+group)
colnames(design)=gsub('group','',colnames(design))
design #Effects are estimated for parameters specified by the design matrix

#Contrasts
contr.matrix=makeContrasts(EmBPOE4vsWT=EmBPOE4-WT,levels=colnames(design))
contr.matrix

#Estimate mean-variance relationship
v=voom(x,design,plot=T) #voom: Use mean-var trend to assign weights to each observation
v #Weights got from voom used in linear modelling removes mean-variance trend

#Fit linear model
vfit=lmFit(v,design) #Fits a linear model for each gene
vfit=contrasts.fit(vfit,contrast=contr.matrix) #Computes statistics and fold changes for comparisons
of interest
efit=eBayes(vfit) #Computes more precise estimates by sharing gene information using empirical
Bayes moderation
```

```
plotSA(efit,main='Final model: Mean-variance trend') #Plots residual standard deviation versus average log expression for fitted model

#Test for Differential Expression(DE)
summary(decideTests(efit)) #adjusted p-value < 0.05

#Top 10 genes from DE analysis
topTable(efit,coef='EmBPOE4vsWT')
  #LogFC: Log fold change, Fold change = ratio between two values
  #AveExpr: Average gene expression across all samples


##Test for DE relative to a threshold: when number of DE genes is large
tfit=treat(vfit,lfc=1) #threshold: logFC >= 1
dt=decideTests(tfit)
summary(dt)

vennDiagram(dt[,1],circle.col=c("turquoise","salmon"))
write.fit(tfit,dt,file="results.txt")

##Top genes from DE analysis
EmBPOE4.vs.WT=topTreat(tfit,coef='EmBPOE4vsWT',n=Inf)
head(EmBPOE4.vs.WT)

#Mean-difference plot
plotMD(tfit,column=1,status=dt[,1],main=colnames(tfit)[1],xlim=c(-8,13))

#Interactive version of MD plot using the Glimma package
library(Glimma)
```

```
glMDSPlot(lcpm,labels=group,groups=x$samples[,c(2,4)],launch=T)
glMDPlot(tfit,coef=1,status=dt,main=colnames(tfit)[1],counts=lcpm,groups=group)

#Heatmap
library(gplots)
EmBPOE4.vs.WT.topgenes=EmBPOE4.vs.WT$ID[1:100]
i=which(v$genes %in% EmBPOE4.vs.WT.topgenes)
mycol=colorpanel(1000,"blue","white","red")
heatmap.2(lcpm[i,],scale="row",labRow=v$genes[i],labCol=group,col=mycol,trace="none",densit
y.info="none",margin=c(8,6),lhei=c(2,10),dendrogram="column")
```