|  |
| --- |
| **甲基化分析报告** |
| **晶能生物技术（上海）有限公司** |
|  |
| **合同号 GA0123MC03-1-CH** |
| **客户名 何东仪** |
| **2015/4/13** |
|  |

目录

[1, 结果目录和分析流程 3](#_Toc419728137)

[1.1分析流程图 3](#_Toc419728138)

[1.2目录信息 3](#_Toc419728139)

[2, 数据预处理 4](#_Toc419728140)

[2.1 原始数据 4](#_Toc419728141)

[2.2 标准化方法 4](#_Toc419728142)

[2.3 标准化的数据总表 5](#_Toc419728143)

[2.4 完整注释表格说明： 6](#_Toc419728144)

[3，差异甲基化位点筛选 10](#_Toc419728145)

[4，差异甲基化位点功能分析 12](#_Toc419728146)

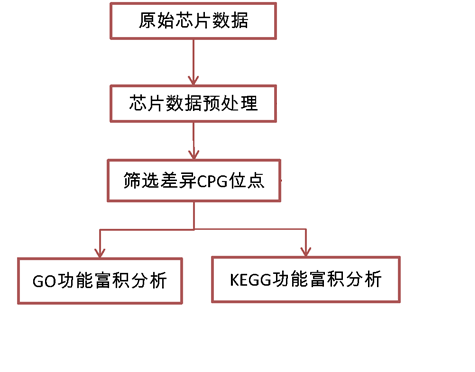
[4.1 GO功能富积分析 12](#_Toc419728147)

[4.2 KEGG功能富积分析 14](#_Toc419728148)

[5，参考文献 18](#_Toc419728149)

# 1, 结果目录和分析流程

## 1.1分析流程图

****

## 1.2目录信息

1原始数据；2均一化数据；3差异甲基化位点筛选；4功能分析：GO和KEGG功能富积分析



**说明： 附件所有的txt文件都可以用office excel文件打开**

# 2, 数据预处理

## 2.1 原始数据

1rawdata文件夹下为原始数据：

Iscan Raw Data目录包括扫描信号值IDAT文件和扫描图JPEG文件；idat文件需要illumina的genomestudio软件才能读取。

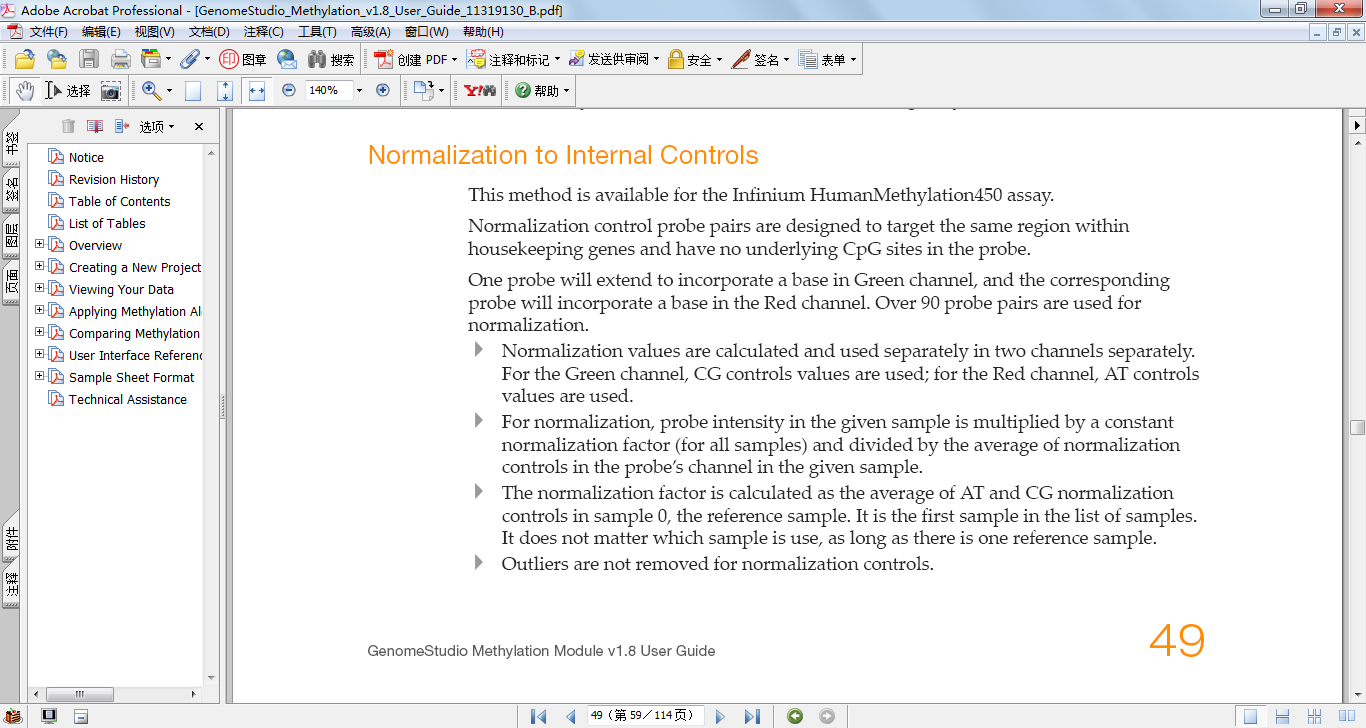
export目录包括genomestudio导出的txt格式原始数据(TableControl\_nonenorm\_nobg）是未经均一化处理，未减去背景每个样本的平均beta值avg beta，检测信号P值detection pval值

## 2.2 标准化方法

2nordata目录下：sample table是样本和芯片对应关系；TableControl\_controlnorm\_bg\_ 是均一化的信号总表

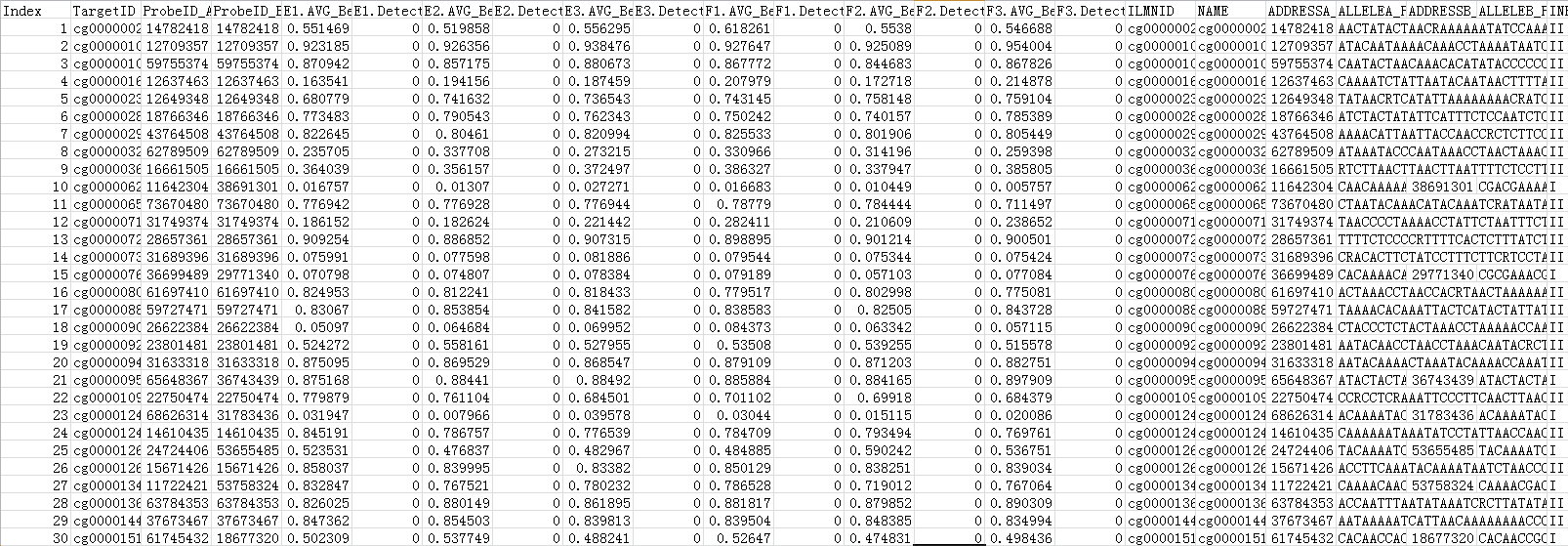
针对Illumine 450K甲基化芯片中的探针甲基化位点，多个样本的原始数据扣背景信号后进行内部对照探针平均数值标准化（Control normalization）预处理，然后计算探针水平的甲基化程度的水平（Beta score）。标准化预处理是目的消减实验可能出现的系统误差， 消减样本之间的批次效应的存在，为了使得样本之间的整体芯片信号之间可以比较， 本方法是基于内对照探针的甲基化程度没有差异的假设进行标准化， 详细英文描述信息见图2.1， 进而使得样本之间可以对Beta score指标进行差异分析，Beta score的取值范围在[0,1]，它表示甲基化探针相对非甲基化探针信号的比率，故可以作为衡量样品各CpG位点甲基化程度的指标。

图2.1 平均数值标准化英文描述信息



## 2.3 标准化的数据总表

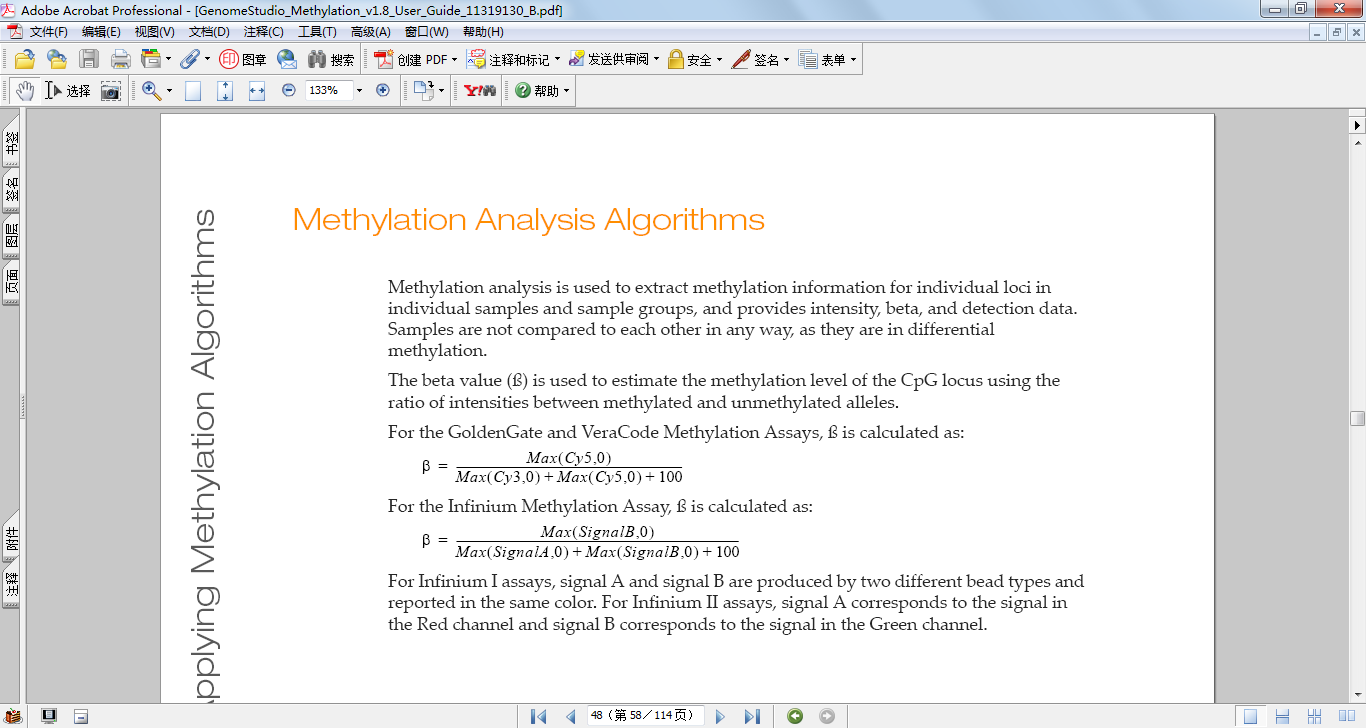
表2.2.1是标准化后的数据总表以及关键表头说明。



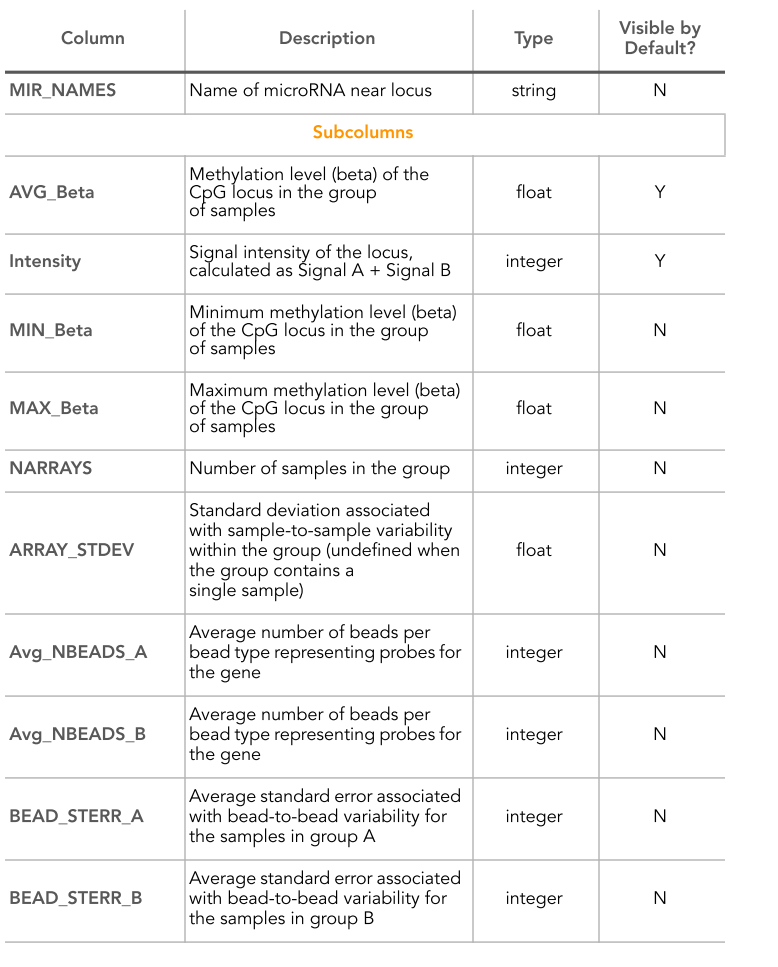
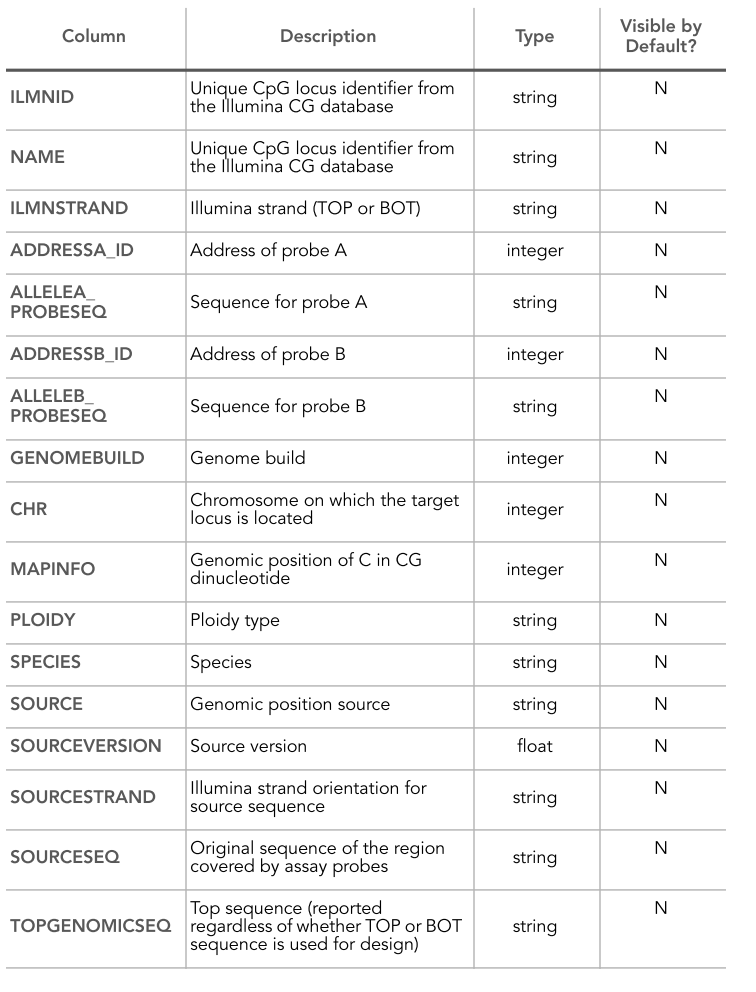
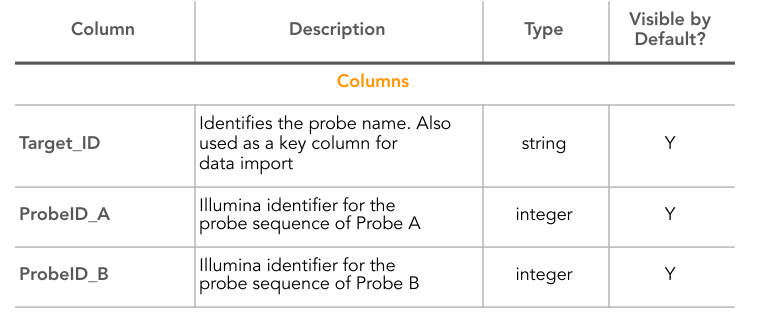
|  |  |
| --- | --- |
| 表头 | 表头说明 |
| TargetID | 探针名 |
| \*.AVG\_Beta列 | 标准化后Beta score值, 其中空值代表没有检测到信号 |
| GENOME\_BUILD | 基因组版本号 |
| CHR | 甲基化位点染色体号 |
| MAPINFO | 甲基化位点染色体位置 |
| UCSC\_REFGENE\_NAME | 甲基化位点附近的基因简称 |
| UCSC\_REFGENE\_ACCESSION | 甲基化位点附近的基因编号 |
| UCSC\_REFGENE\_GROUP | 甲基化位点在基因区域或者附近的分类， TSS200/1500是转录开始位置上游200bp和1500bp， Body是转录本的编码蛋白质区域， 5’UTR是转录本的5端UTR区域， 3’UTR是转录本的3端UTR区域，1stExon是转录本的第1个外显子区域 |
| UCSC\_CPG\_ISLANDS\_NAME | UCSC的CpG岛名称 |
| RELATION\_TO\_UCSC\_CPG\_ISLAND | 与UCSC的CpG岛位置关系， Island是岛内， N\_Shore和N\_Shelf是岛5端附近， S\_Shore和S\_Shelf是岛3端附近 |
| ENHANCER | 预测的增强子元件 |
| HMM\_ISLAND | HMM方法预测的CpG岛名称 |
| REGULATORYFEA TURE\_NAME | 调控元件名称 |
| DHS | 实验确证的DNAse 超敏感性位点 |

注：Beta score的计算方法见下图2.2.2。Detection Pval小于等于0.05表示Beta score比较可靠。

图2.2.2 Beta score的计算方法

****

## 2.4 完整注释表格说明：



# 3，差异甲基化位点筛选

1对1样本计算差异的统计方法使用illumina Custom Model



差异位点的筛选标准：实验组的deta-beta值大于0.2或小于-0.2，diffscore值大于13或小于-13，detection-p值小于0.05；且对照组的detection-p值小于0.05。

表3.1是差异甲基化的分析结果以及关键表头说明



|  |  |
| --- | --- |
| 表头 | 表头说明 |
| TargetID | CpG位点探针名 |
| AVG\_Beta | 样本Beta score |
| DiffScore | 差异统计值，体现一个探针在两种组织中的表达差异程度，绝对值越大差异越可信 |
| Delta Beta | 两个样本的Beta score差值 |

# 4，差异甲基化位点功能分析

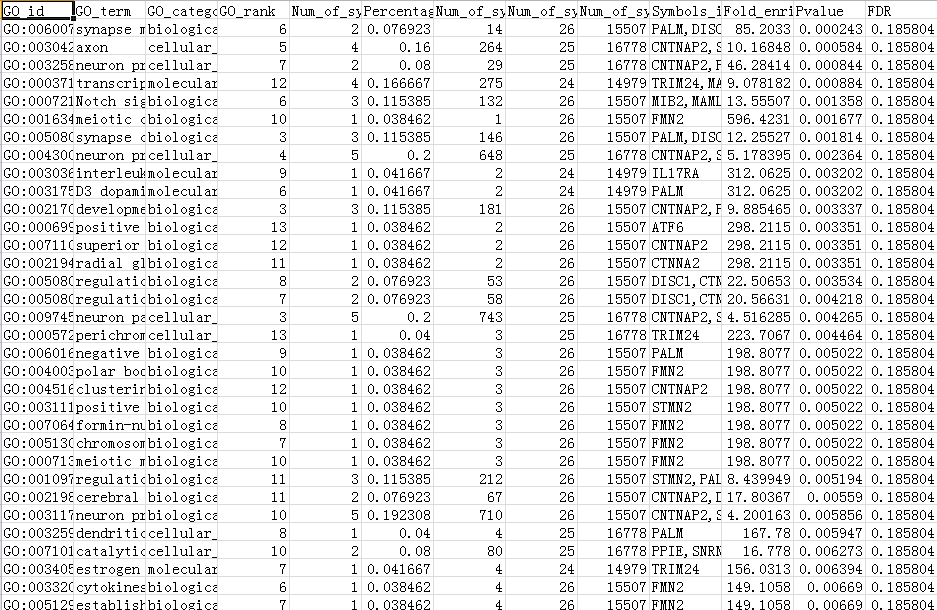
## 4.1 GO功能富积分析

针对差异化甲基化位点最近邻基因进行GO功能富积分析。 GO功能分析是针对全基因/转录本和差异基因/转录本进行功能注释和归类， 如果从全基因/转录本筛选到差异基因/转录本列表，可以对差异基因/转录本列表进行GO功能富积分析， GO功能富积分析的方法： 将全部基因/转录本作为背景列表，差异基因/转录本列表作为从背景列表中筛选出来的候选列表，利用超几何分布检验计算代表GO功能集在差异基因/转录本列表中是否显著富积的P值， 再对P值经Benjamini & Hochberg多重检验纠正后得到FDR。超几何分布检验计算P值的公式如下：



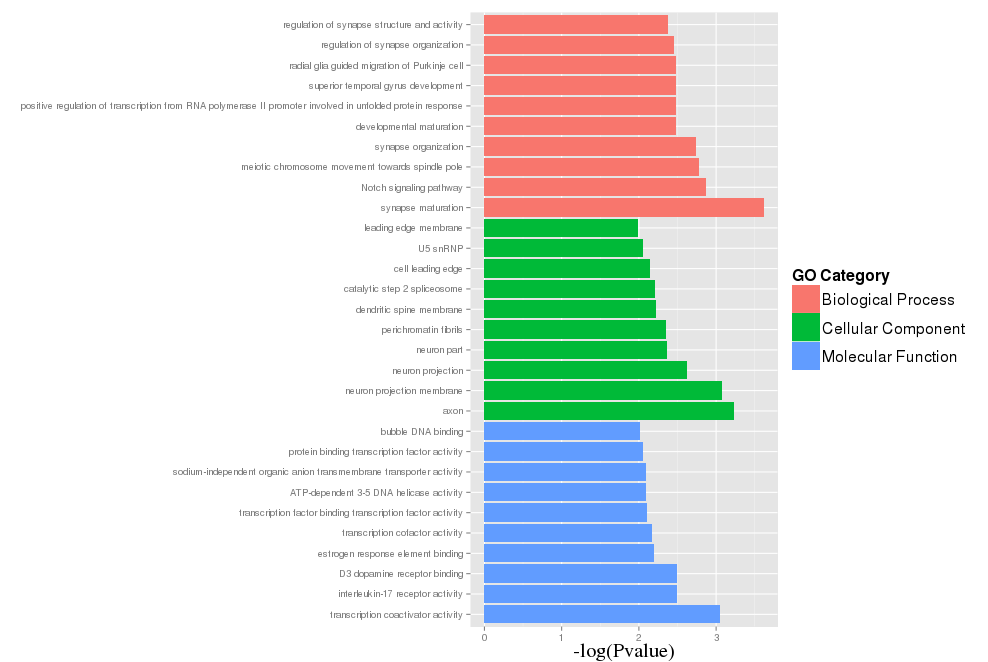
说明： N代表全基因/转录本中具有GO注释的基因数目， M代表全基因/转录本中注释为某个GO的基因/转录本数目， m代表差异基因/转录本中注释为某个GO的基因/转录本数目。GO功能富积分析结果文件见表4.1.1。显著富积GO柱状图见图4.1.1，根据pvalue小于等于0.05筛选显著富积GO， 每个GO分类下至多显示10个GO。。显著富积GO柱状图如图4.1.2，根据pvalue小于等于0.05筛选显著富积GO， 每个GO分类下至多显示10个GO。显著富积GO树状图如图9.1.2，根据pvalue小于等于0.05筛选显著富积GO， 每个GO分类下至多显示10个GO， GO树状图分为biological\_process子树、molecular\_function子树、cellular\_component子树。

表4.1.1 GO功能富积分析结果文件。



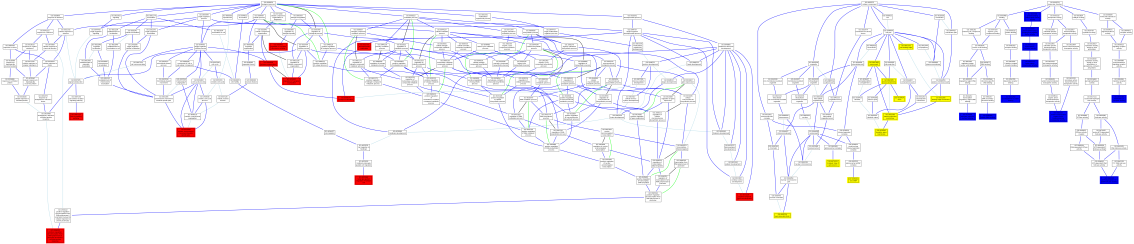
|  |  |
| --- | --- |
| 表头名 | 表头名的说明 |
| GO\_id | GO标示 |
| GO\_term | GO名称 |
| GO\_category | GO分类名 |
| GO\_rank | GO树层级 |
| Num\_of\_symbols\_in\_list\_in\_GO | 注释在这个GO且在候选基因列表中的基因总数 |
| Percentage\_of\_symbols\_in\_list\_in\_GO | 注释在这个GO且在候选基因列表中的基因总数占在候选基因列表中的基因总数的比例 |
| Num\_of\_symbols\_in\_bglist\_in\_GO | 注释在这个GO且在背景基因列表中的基因总数 |
| Num\_of\_symbols\_in\_list | 注释在这个GO顶层分类且在候选基因列表中的基因总数 |
| Num\_of\_symbols\_in\_bglist | 注释在这个GO顶层分类且在背景基因列表中的基因总数 |
| Symbols\_in\_list\_in\_GO | 注释在这个GO且在候选基因列表中的基因 |
| Fold\_enrichment | 计算方法是比例1除以比例2，比例1是Num\_of\_symbols\_in\_list\_in\_GO除以Num\_of\_symbols\_in\_list， 比例2是Num\_of\_symbols\_in\_bglist\_in\_GO除以Num\_of\_symbols\_in\_bglist |
| Pvalue | 超几何检验计算的P值 |
| FDR | 原始P值经Benjamini & Hochberg多重检验纠正后的P值 |

图4.1.1 显著富积GO柱状图。



横坐标代表-log2(Pvalue)， 纵坐标代表显著富积的GO名称

图4.1.2显著富积GO树状图

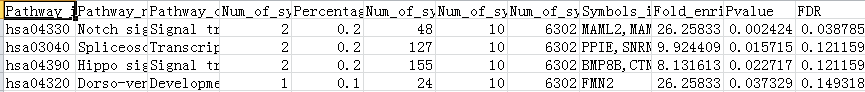


注：左边是Biological\_process子树，红色框代表显著富积GO， 中间是molecular\_function子数，黄色框代表显著富积GO，右边是cellular\_component子树，蓝色框代表显著富积GO。

## 4.2 KEGG功能富积分析

针对差异化甲基化位点最近邻基因进行KEGG Pathway功能富积分析。 KEGG pathway功能分析是针对全基因/转录本和差异基因/转录本进行KEGG数据库中Pathway的功能注释和归类， KEGG pathway功能富积分析方法与GO功能富积分析类似。KEGG Pathway功能富积分析结果文件见表4.2.1。显著富积KEGG pathway柱状图见图4.2.1。显著富积KEGG pathway散点图如图4.2.2。显著富积KEGG pathway图见图4.2.3。

表4.2.1 KEGG Pathway功能富积分析结果文件。



|  |  |
| --- | --- |
| 表头 | 表头说明 |
| Pathway\_id | Pathway标示 |
| Pathway\_name | Pathway名称 |
| Pathway\_class | Pathway分类名 |
| Num\_of\_symbols\_in\_list\_in\_path | 注释在这个pathway且在候选基因列表中的基因总数 |
| Percentage\_of\_symbols\_in\_list | 注释在这个pathway且在候选基因列表中的基因总数占在候选基因列表中的基因总数的比例 |
| Num\_of\_symbols\_in\_bglist\_in\_path | 注释在这个pathway且在背景基因列表中的基因总数 |
| Num\_of\_symbols\_in\_list | 在候选基因列表中带有KEGG pathway注释的基因总数 |
| Num\_of\_symbols\_in\_bglist | 在背景基因列表中带有KEGG pathway注释的基因总数 |
| Symbols\_in\_list\_in\_path | 注释在这个pathway且在候选基因列表中的基因列表 |
| Fold\_enrichment | 计算方法是比例1除以比例2，比例1是Num\_of\_symbols\_in\_list\_in\_path除以Num\_of\_symbols\_in\_list， 比例2是Num\_of\_symbols\_in\_bglist\_in\_path除以Num\_of\_symbols\_in\_bglist |
| Pvalue | 超几何检验计算的P值 |
| FDR | 原始P值经Benjamini & Hochberg多重检验纠正后的P值 |

图4.2.1 显著富积KEGG pathway柱状图， 横坐标代表显著富积的KEGG pathway名称， 纵坐标代表-log2(pvalue)。 纵坐标越显著表示该Pathway越富集显著， 红色柱表示显著的Pathway通路（p<=0.05）， 蓝色柱表示不显著的pathway通路

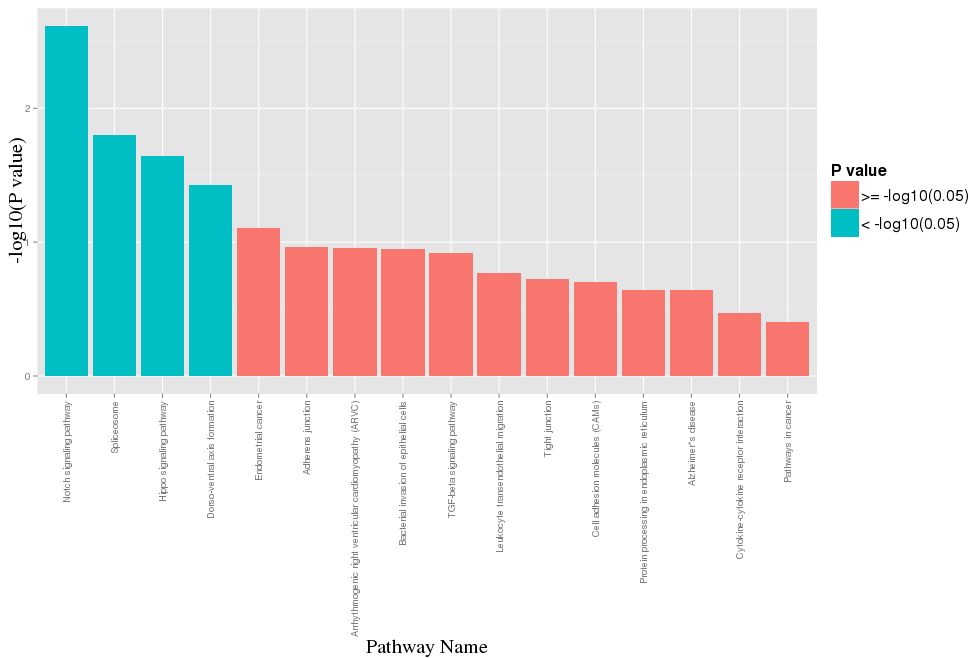


图4.2.2 显著富积KEGG pathway散点图

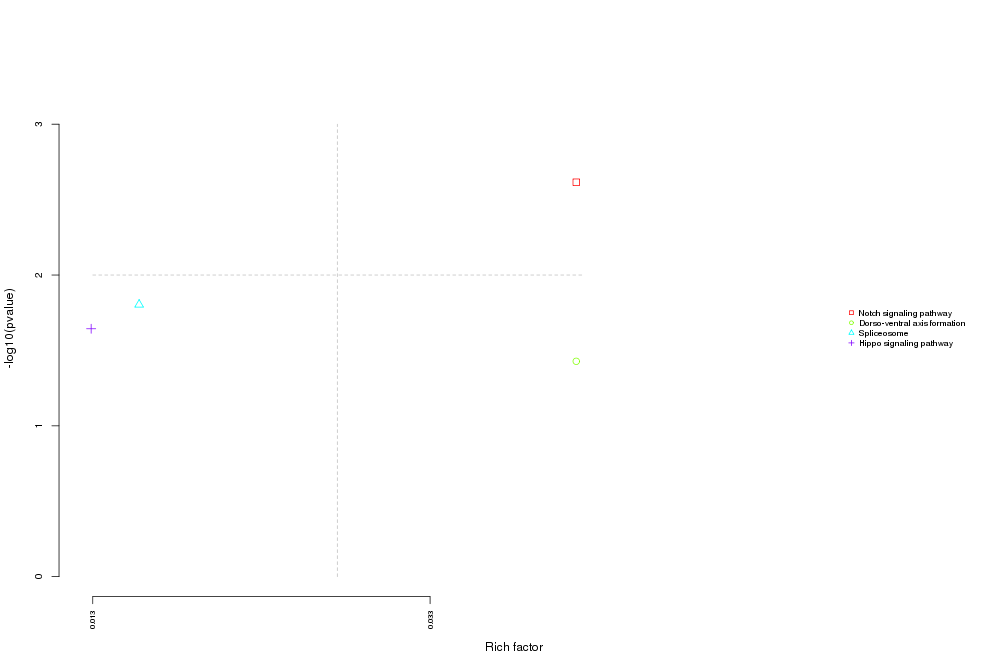
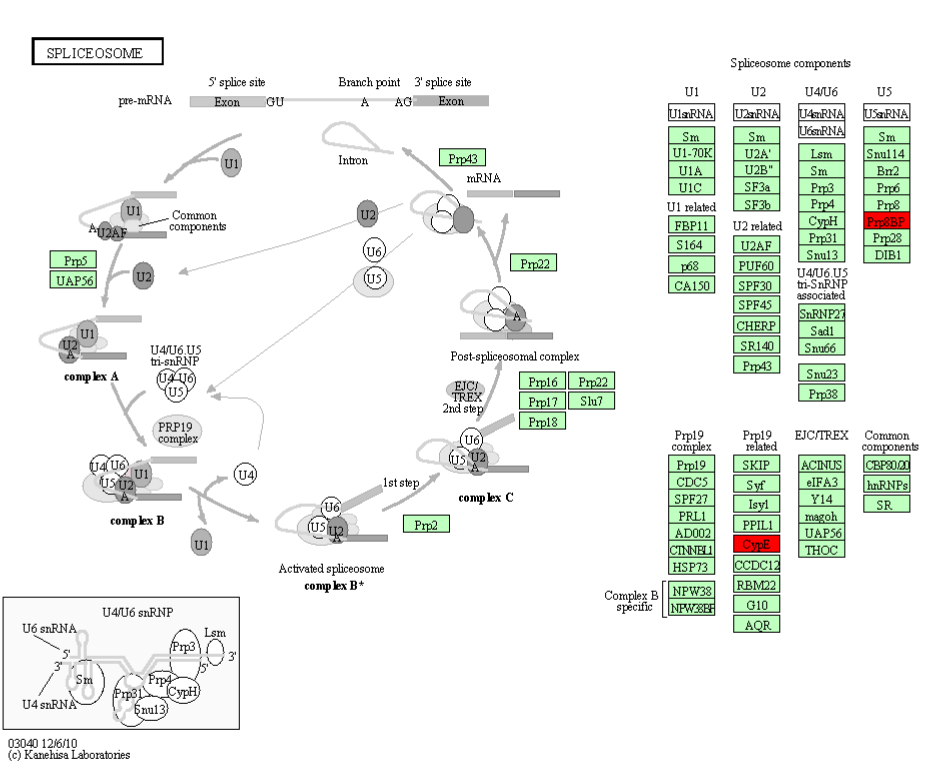


图4.2.3 显著富积KEGG pathway图, 红色方框代表差异转录本对应的基因。



# 5，参考文献

[1] Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. [Brunner AL](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Brunner%20AL%22%5BAuthor%5D), [Johnson DS](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Johnson%20DS%22%5BAuthor%5D), [Kim SW](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Kim%20SW%22%5BAuthor%5D), [Valouev A](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Valouev%20A%22%5BAuthor%5D), [Reddy TE](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Reddy%20TE%22%5BAuthor%5D), [Neff NF](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Neff%20NF%22%5BAuthor%5D), [Anton E](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Anton%20E%22%5BAuthor%5D), [Medina C](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Medina%20C%22%5BAuthor%5D), [Nguyen L](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Nguyen%20L%22%5BAuthor%5D), [Chiao E](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Chiao%20E%22%5BAuthor%5D), [Oyolu CB](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Oyolu%20CB%22%5BAuthor%5D), [Schroth GP](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Schroth%20GP%22%5BAuthor%5D), [Absher DM](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Absher%20DM%22%5BAuthor%5D), [Baker JC](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Baker%20JC%22%5BAuthor%5D), [Myers RM](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Myers%20RM%22%5BAuthor%5D). [Genome Res.](http://www.ncbi.nlm.nih.gov/pubmed/19273619) 2009 Jun.

[2] GALGO: an R package for multivariate variable selection using genetic algorithms. Victor Trevino Francesco Falciani. Bioinformatics. 2006 February.

[3] DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status. [Pike BL](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Pike%20BL%22%5BAuthor%5D), [Greiner TC](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Greiner%20TC%22%5BAuthor%5D), [Wang X](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Wang%20X%22%5BAuthor%5D), [Weisenburger DD](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Weisenburger%20DD%22%5BAuthor%5D), [Hsu YH](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Hsu%20YH%22%5BAuthor%5D), [Renaud G](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Renaud%20G%22%5BAuthor%5D), [Wolfsberg TG](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Wolfsberg%20TG%22%5BAuthor%5D), [Kim M](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Kim%20M%22%5BAuthor%5D), [Weisenberger DJ](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Weisenberger%20DJ%22%5BAuthor%5D), [Siegmund KD](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Siegmund%20KD%22%5BAuthor%5D), [Ye W](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Ye%20W%22%5BAuthor%5D), [Groshen S](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Groshen%20S%22%5BAuthor%5D), [Mehrian-Shai R](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Mehrian-Shai%20R%22%5BAuthor%5D), [Delabie J](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Delabie%20J%22%5BAuthor%5D), [Chan WC](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Chan%20WC%22%5BAuthor%5D), [Laird PW](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Laird%20PW%22%5BAuthor%5D), [Hacia JG](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Hacia%20JG%22%5BAuthor%5D). [Leukemia.](http://www.ncbi.nlm.nih.gov/pubmed/18288132) 2008 May.

[4] <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html>

[5] http://www.bioconductor.org/packages/release/bioc/html/limma.html