合同编号：M-GSGC0174590

**PrimeView Human Gene Expression Array**

**人表达谱芯片**

**生物信息分析报告**

**吉凯基因****[[1]](#footnote-1)**

**目 录**

[1 原始数据及信息分析流程 3](#_Toc535928564)

[**1.1** **样本及原始数据信息** 3](#_Toc535928565)

[**1.2** **信息分析流程** 3](#_Toc535928566)

[2 芯片数据的质量评估 3](#_Toc535928567)

[**2.1** **信号强度分布统计（Signal Histogram）** 4](#_Toc535928568)

[**2.2** **相对对数信号强度统计（Relative Signal Box Plot）** 4](#_Toc535928569)

[**2.3** **样本相关性分析（Correlation Analysis）** 5](#_Toc535928570)

[**2.4** **主成分分析（Principal Component Analysis）** 5](#_Toc535928571)

[3 数据过滤 6](#_Toc535928572)

[4 显著性差异分析 7](#_Toc535928573)

[**4.1** **显著性差异分析** 7](#_Toc535928574)

[**4.2** **散点图（Scatter Plot）** 7](#_Toc535928575)

[**4.3** **层次聚类分析（Hierarchical Clustering）** 7](#_Toc535928576)

[5 显著性差异基因的生物信息分析 8](#_Toc535928577)

[**5.1** **Gene Ontology (GO) 注释及富集分析** 8](#_Toc535928578)

[**5.2** **基于KEGG & BioCarta的通路注释及富集分析** 9](#_Toc535928579)

[6 结果文件及保存位置 10](#_Toc535928580)

[7 参考文献 10](#_Toc535928581)

1. **原始数据及信息分析流程**
   1. **样本及原始数据信息**

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 原始数据文件名 | 样本名称 | 分析分组标记 |
| 1 | E8140-1\_(PrimeView).CEL | 过表达对照组 | A |
| 2 | E8140-2\_(PrimeView).CEL | 过表达对照组 | A |
| 3 | E8140-3\_(PrimeView).CEL | 过表达对照组 | A |
| 4 | E8140-4\_(PrimeView).CEL | 过表达对照组 | A |
| 5 | E8141-1\_(PrimeView).CEL | TLX2过表达组 | C |
| 6 | E8141-2\_(PrimeView).CEL | TLX2过表达组 | C |
| 7 | E8141-3\_(PrimeView).CEL | TLX2过表达组 | C |
| 8 | E8141-4\_(PrimeView).CEL | TLX2过表达组 | C |
| 9 | E8142-1\_(PrimeView).CEL | ZNF132过表达组 | B |
| 10 | E8142-2\_(PrimeView).CEL | ZNF132过表达组 | B |
| 11 | E8142-3\_(PrimeView).CEL | ZNF132过表达组 | B |
| 12 | E8142-4\_(PrimeView).CEL | ZNF132过表达组 | B |

* 1. **信息分析流程**

GO注释及富集分析

Pathway注释及富集分析

调控网络分析

差异分析及统计检验

火山图

散点图

聚类分析

信号强度分布

信号强度归一化

相关性分析

主成分分析

不合格

合格

芯片原始数据

质量控制

数据预处理

显著性差异分析

差异基因功能分析

图1-2 常规芯片信息分析流程图

在芯片信息分析过程中，通常首先对芯片原始数据进行质量评价；对于质量控制合格的数据，经过数据过滤，对剩余的符合过滤标准的数据进行后续信息分析，包括显著性差异分析及差异基因的功能分析等，从而辅助肿瘤的发生、发展、转移等分子机制的研究、靶点筛选和药物设计等。

1. **芯片数据的质量评估**
   1. **信号强度分布统计（Signal Histogram）**

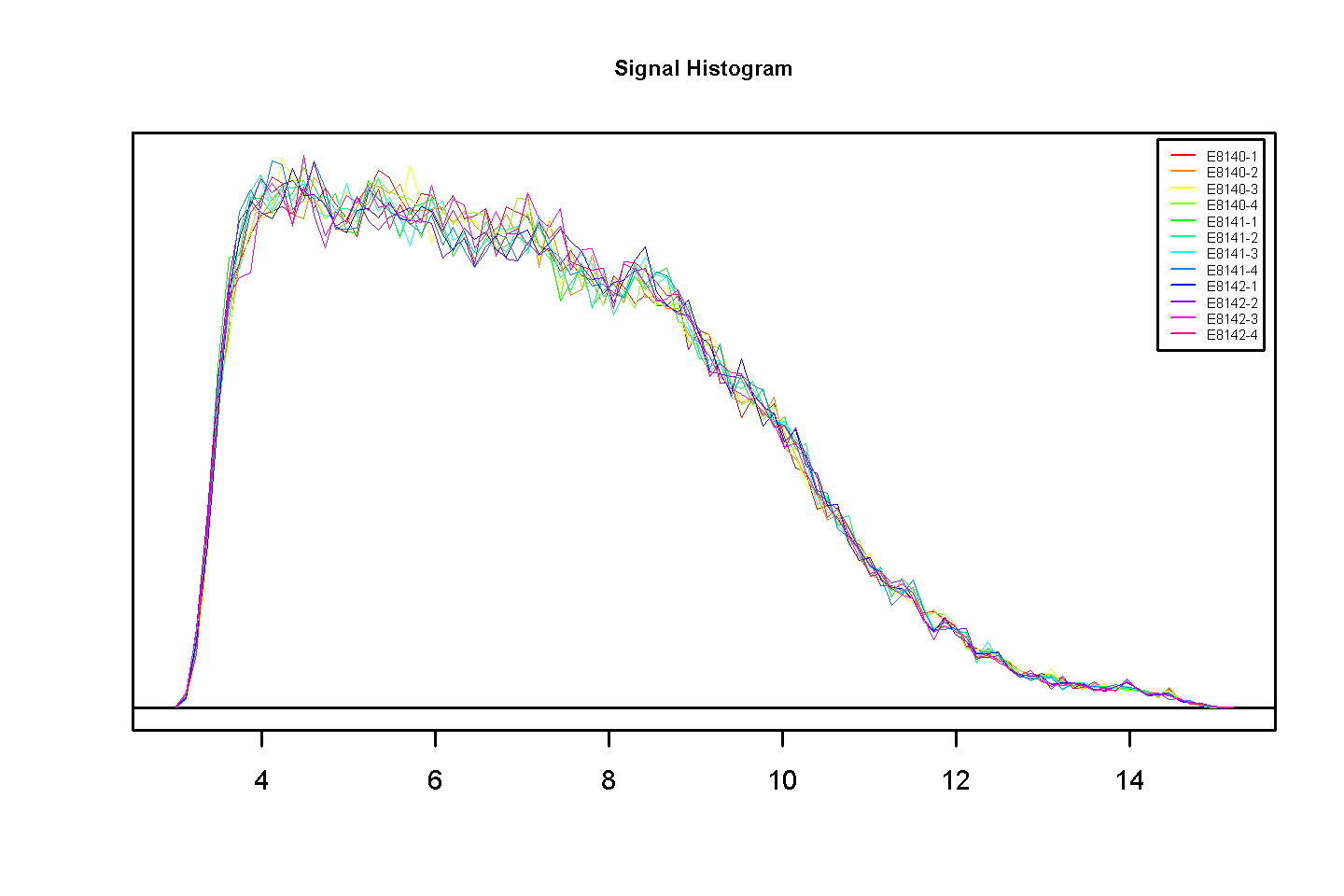


图2-1 信号强度分布曲线图

信号强度分布曲线图展示了所有芯片探针的信号强度分布情况。横坐标表示探针信号强度区间，纵坐标表示在信号强度区间内的探针集数目。不同样本的信号强度分布曲线重合度越好，表示芯片实验的可靠性越高。该项目所有样本在一个信号强度区间内的样本数的平均Z-score值均小于2，表示所有芯片结果的可靠性均符合继续分析标准。

* 1. **相对对数信号强度统计（Relative Signal Box Plot）**

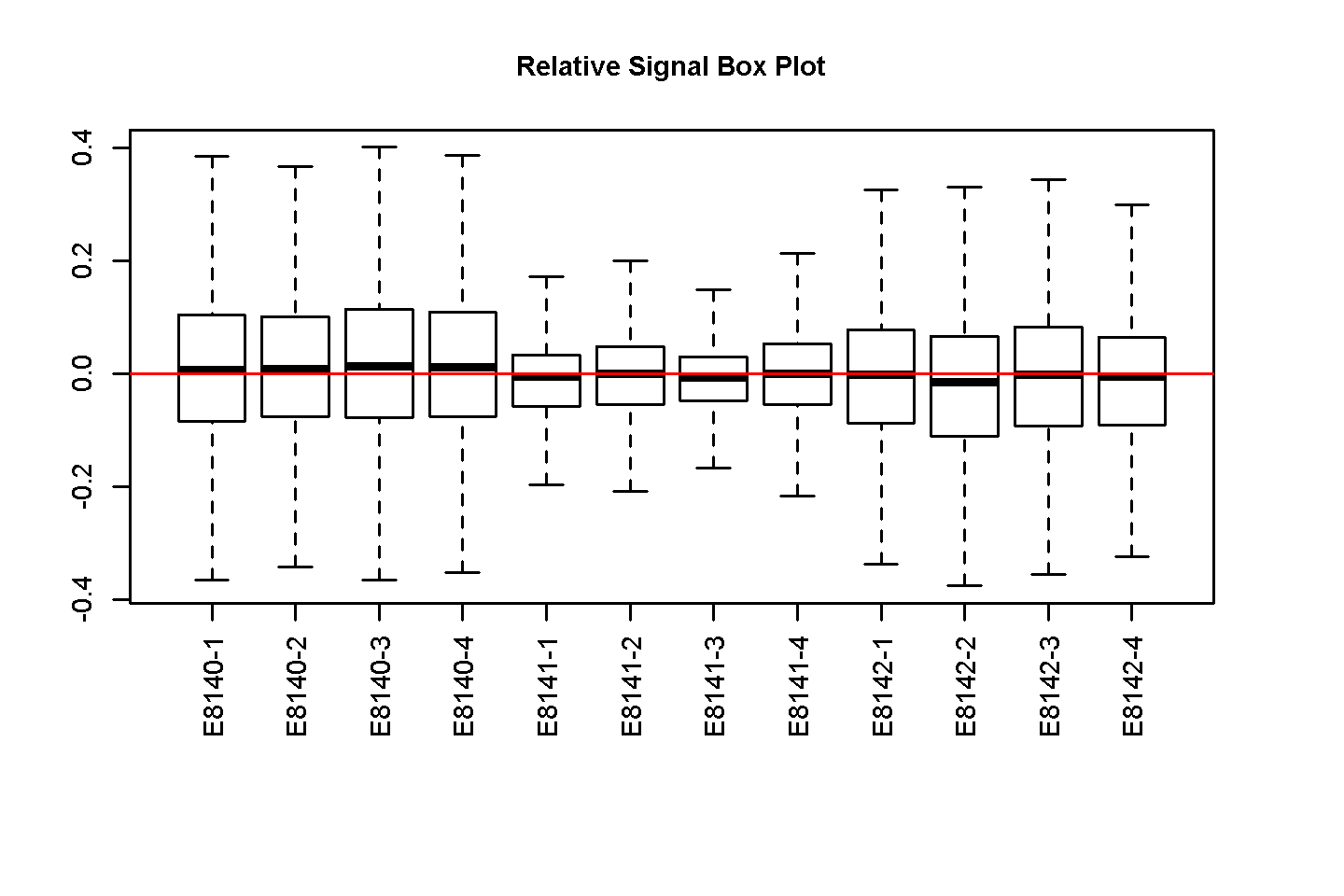


图2-2 相对对数信号强度箱线图

相对对数信号强度箱线图展示了所有芯片的相对对数信号强度的分布情况[1]。横坐标表示样本名称；纵坐标表示相对对数信号强度；中间的红线代表所有样本的相对对数信号强度的平均值；每个盒状上下部横线分别代表上下90%置信区间，盒状上下边缘表示上下四分位点，中间的黑线表示中位数。相对对数信号强度箱线图的分布越相近，表示数据的重复性越好。此项目中所有样本的中位数的Z-score值均小于2，表示该项目中的芯片实验的重复性较好，符合继续分析标准。

* 1. **样本相关性分析（Correlation Analysis）**

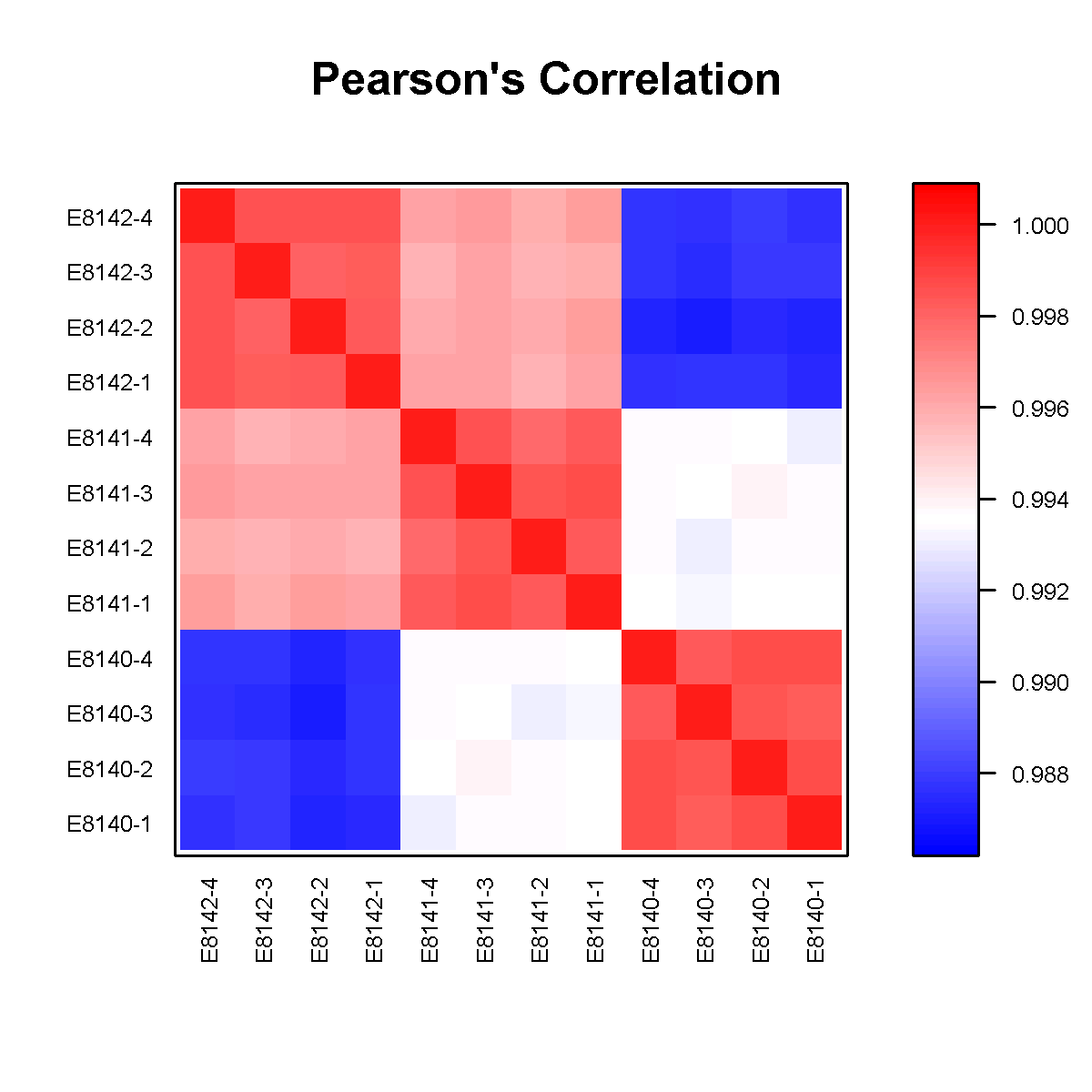


图2-3 样本间皮尔森相关系数分布图

皮尔森相关系数分布图[2]展示了所有芯片之间信号强度的相关性水平，每一个方格表示纵坐标和横坐标对应的两个样本之间的相关程度。皮尔森相关系数的范围为【-1.0,1.0】；相关系数为正值提示两样本中基因的表达模式呈正相关，为负值提示两样本中基因的表达模式呈负相关；相关系数的绝对值越接近1.0，相关程度越高。通常情况下，组内样本的基因表达模式相似，相关系数高，在相关系数分布图中呈红色；组间的基因表达模式差异大，相关系数低，在相关系数分布图中呈蓝色。本项目中，B和A组内相关系数均大于0.99，提示同组内样本的基因表达趋势相似度高；组间相关系数均显著低于组内相关系数，提示组间差异大，符合继续分析标准。

* 1. **主成分分析（Principal Component Analysis）**

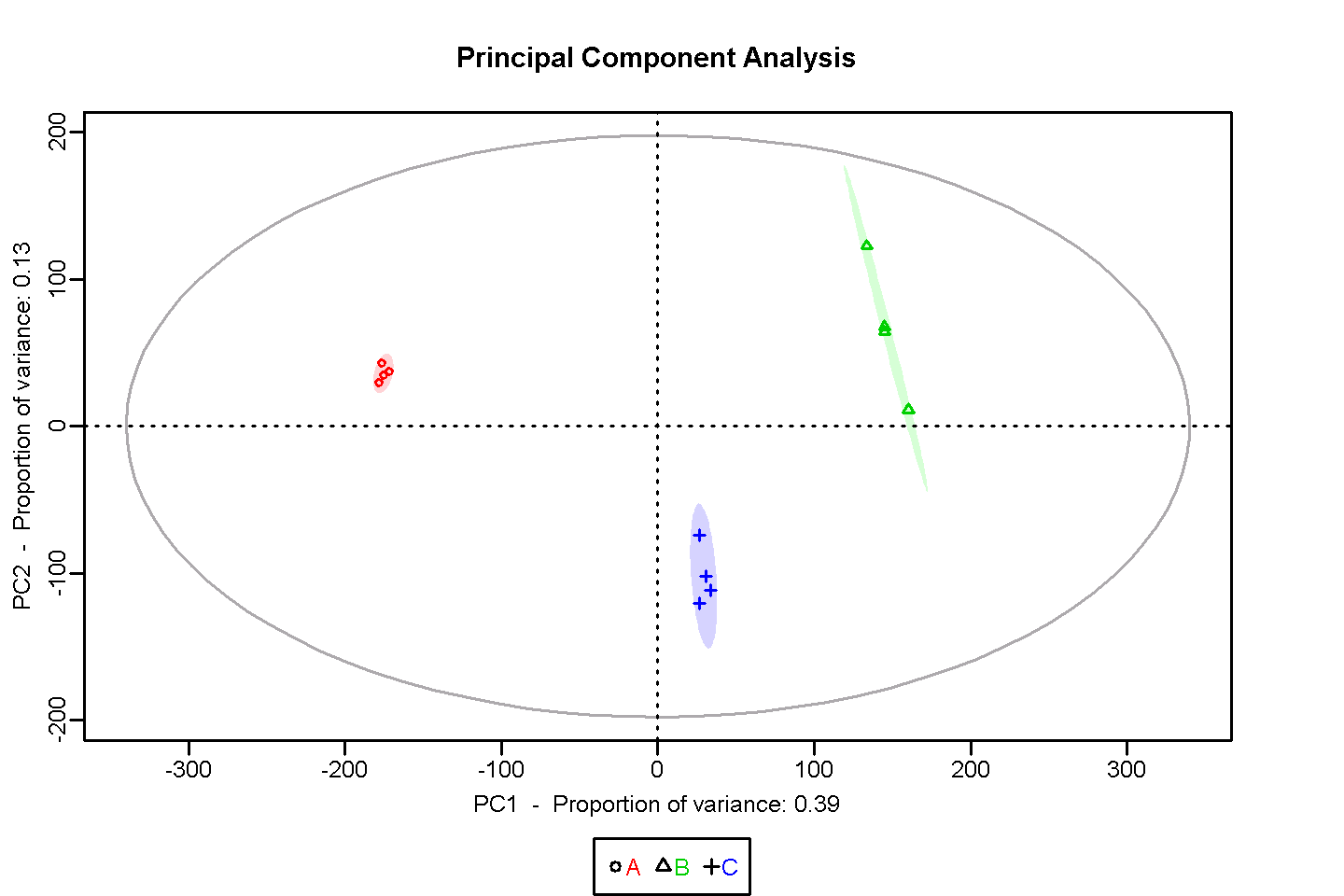


图2-4 主成分分析（PCA）得分图[3]

主成分分析（Principal Component Analysis, PCA）[4]是将芯片上的探针组重新线性组合，形成一组新的综合变量的过程。根据所分析的问题，从中选取几个（通常2-3个）综合变量，使它们尽可能多地反应原有变量的信息，从而达到降维的目的。同时，进行主成分分析还能从总体上反映组内和组间的变异度。在图2-4中，红色点代表A样本，绿色点代表B样本，蓝色点代表C样本。由PCA得分图可见，在PC1和PC2维度上，B和A组内样本聚集趋势明显，组间分离趋势明显，提示组内样本相似，组间差异显著，符合继续分析标准。

1. **数据过滤**

数据的有效性和准确性是后续获得具有统计学和生物学意义的分析结果的必要条件。由于芯片本身的设计原理以及实验过程中的人为或其他不可避免的因素，芯片原始数据中存在大量不合格的或无效的检测点。因此，对芯片数据合理的预处理对于获取准确的基因表达谱以及基于这些表达谱数据进行的后续各项分析的合理性和准确性均具有重要的意义。对芯片数据的预处理包括数据归一化[5-7]、背景噪音过滤和数据清洗（data cleaning）等过程。

我们将两个样本组别中处于所有探针组信号强度排序的最低20%范围内的探针组作为背景噪音，予以滤除。其次，利用变异系数法，计算同一探针组在同一样本组内的变异系数（Eq 3-1），对在两个组别中变异系数均大于25%的探针组予以滤除。预处理后的探针数目统计见表3-1。【注：组织样本数据不进行变异系数过滤】。

Eq 3-1

表3-1 数据过滤前后探针（组）统计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **比较组别** | **处理序号** | **数据过滤方式** | **探针数** | **探针组数** |
|  | 0 | 原始未处理 | 535,824 | 49,395 |
| B\_vs\_A | 1 | 背景噪音过滤 | - | 39,338 |
| 2 | 数据清洗 | - | 39,338 |
| C\_vs\_A | 1 | 背景噪音过滤 |  | 39377 |
| 2 | 数据清洗 |  | 39377 |

1. **显著性差异分析**
   1. **显著性差异分析**

|  |  |  |  |
| --- | --- | --- | --- |
| **编号** | **比较组别** | **上调基因数目** | **下调基因数目** |
| 1 | B vs A | 1383 | 932 |
| 2 | C\_vs\_A | 766 | 133 |
| 备注 | - | | |

在本项目中，我们使用基于经验贝叶斯分布的线性模型[8]计算显著差异性水平P-value，并使用Benjamini-Hochberg法对显著差异水平进行校正（FDR）。显著性差异基因的筛选标准为：|Fold Change|≥1.5且FDR<0.05。

* 1. **散点图（Scatter Plot）**

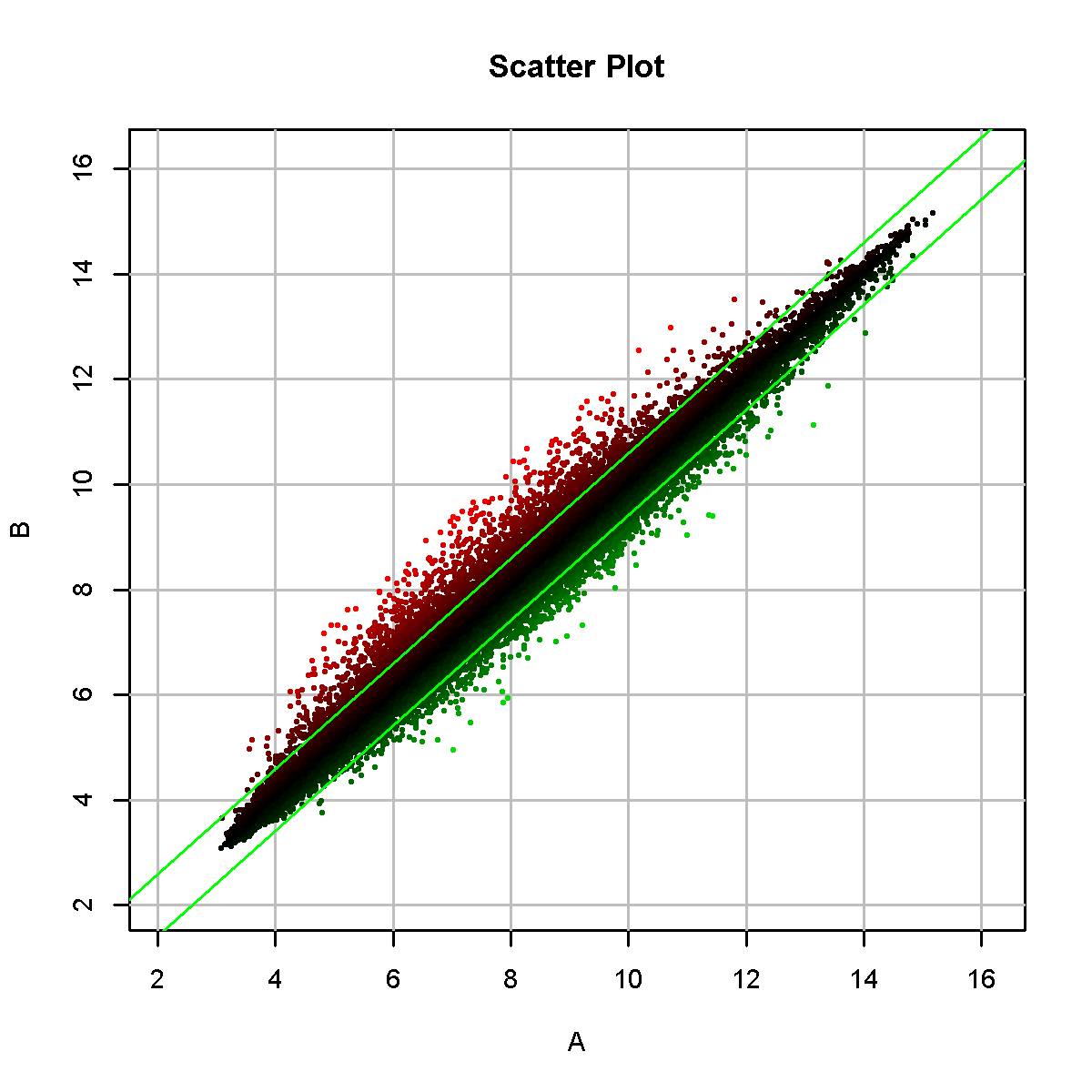
****

图4-2 散点图\_B\_vs\_A

散点图展示了实验组与对照组之间的信号强度在直角坐标系平面上的分布情况。图4-2 中，横坐标表示A组，纵坐标表示B组；图中每一个点的横纵坐标代表一个探针组在A组和B组中的信号强度；平行的绿色实线为差异参考线，参考线内区间的点代表无显著变化的探针组，区间外的红色点代表在B组中相对上调的探针组，绿色点代表在A组中相对上调的探针组。

* 1. **层次聚类分析（Hierarchical Clustering）**

聚类分析是一种常用的探索性数据分析方法，其目的是在相似性的基础上对数据进行分组、归类。聚类分组的结果中，组内的数据模式相似性较高，而组间的数据模式相似性较低。

在聚类分析过程中，聚类算法会对样本（Sample）和变量（Variable，在芯片数据中通常指基因的定量信息）两个维度进行分类。对样本的聚类结果可以检验所筛选的目标基因的合理性，即这些目标基因表达量的变化可否代表生物学处理对样本造成的显著影响；目标基因的聚类结果可以帮助我们从基因集合中区分具有不同表达模式的基因子集合，具有相近表达模式的基因可能具有相似的功能或者参与相同的生物学途径，或者在通路中处于临近的调控位置（图4-3）。

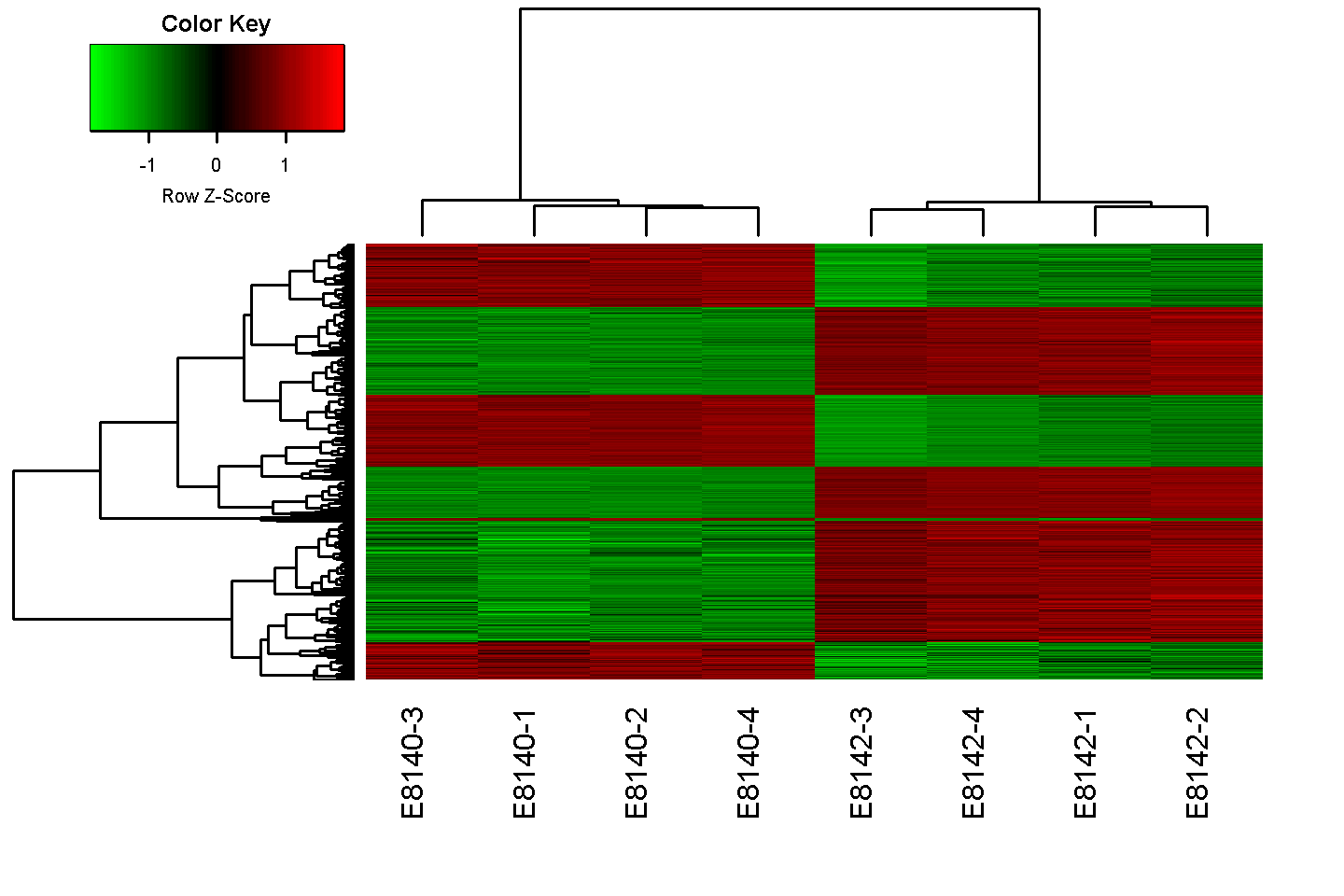
****

图4-3 聚类分析结果\_B\_vs\_A

图4-3为利用以|Fold Change|≥1.5且FDR<0.05为标准筛选的差异基因的表达谱对B和A两组样本进行层次聚类的热图[9]。在聚类分析热图中，每一列代表一个样本，每一行代表一个差异基因；上部树状结构是根据差异基因的表达谱，所有样本的聚集或分类情况；左侧树状结构表示差异基因的表达模式聚集情况；红色表示基因的表达程度相对上调，绿色表示基因的表达程度相对下调，黑色表示基因的表达没有显著变化，灰色表示基因的信号强度未检出。

1. **显著性差异基因的生物信息分析**

在基因组学中，通过基因芯片等技术产生的海量数据代表了生物体内发生的全部过程及其变化。从这些庞大而复杂的实验数据中寻找生物体的改变以及引起这些改变的源头和机制，是基因组生物信息学的主要任务。

在芯片数据分析中，常用的生物信息学分析方法包括（但不局限于）：

* GO注释及富集分析
* KEGG & BioCarta通路注释及富集分析
  1. **Gene Ontology (GO) 注释及富集分析**

基因本体（Gene Ontology）[10]是一个标准化的功能分类体系，提供了一套动态更新的标准化词汇表，并以此从三个方面描述生物体中基因和基因产物的属性：参与的生物过程（Biological Process），分子功能（Molecular Function）和细胞组分（Cellular Component）。

对目标基因集合的GO注释可以从参与的生物学过程、具有的分子功能和所处的细胞组分三个方面对这些基因进行分类。各个分类比例虽然可以在一定程度上反映实验设计中生物学处理对各个分类的影响程度大小，但是单纯依据该比例来评价各个分类受影响的显著程度是不准确的，还需要同时考虑各个分类在总体基因集合（例如：芯片上的全部基因，该物种所有已知的基因等）中的分布情况。

通常情况下，GO注释的显著性富集分析通过Fisher精确检验（Fisher’s Exact Test）来评价某个GO term基因富集度的显著性水平（图5-1）[11]。

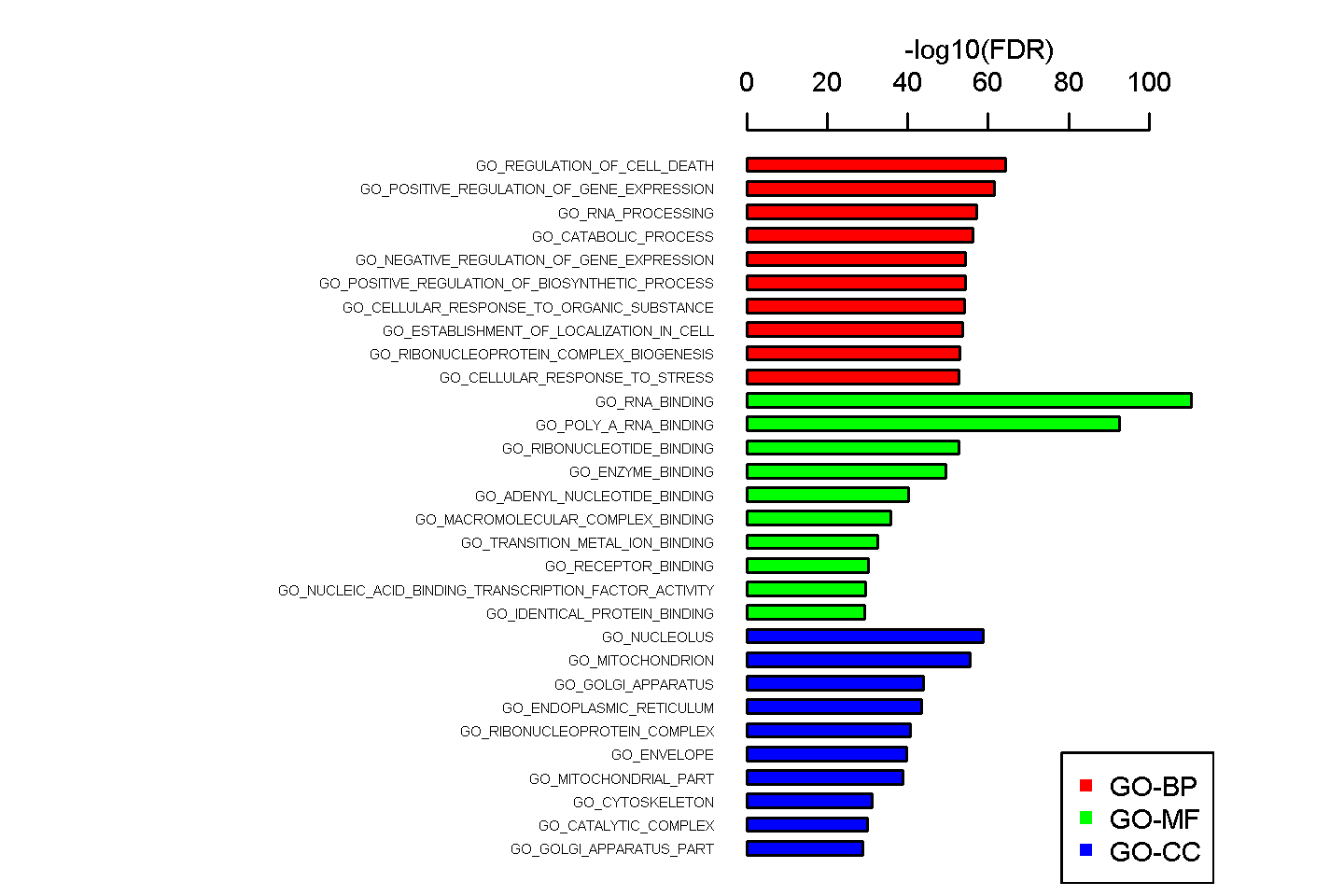


图5-1 显著富集的GO term统计\_B\_vs\_A（Top 10）

图5-1 中列举了在差异基因中富集程度最显著的GO term。在图5-1中，横坐标为富集的显著性水平FDR（以10为底的负对数变换），纵坐标为显著富集的GO term；在同类别中排名越靠前，代表在统计学上该GO term的影响或变化越显著。

* 1. **基于KEGG & BioCarta的通路注释及富集分析**

在生物体中，基因及其产物并不独立行使其功能，而是不同基因及其产物相互协调完成一系列生化反应以行使其生物学功能。因此，通路分析是更系统、全面地了解细胞的生物学过程、性状或疾病的发生机理、药物作用机制等最直接和必要的途径[12]。

KEGG[[2]](#footnote-2)和BioCarta[[3]](#footnote-3)是目前用于通路分析最常用、最权威的数据库。通路富集分析方法与GO富集分析相似，即以通路为单位，以芯片上所有基因或物种已知的全部基因为背景，通过Fisher精确检验（Fisher’s Exact Test），来分析计算各个通路基因富集度的显著性水平，从而确定受到显著影响的代谢和信号转导途径（图5-2）[11]。

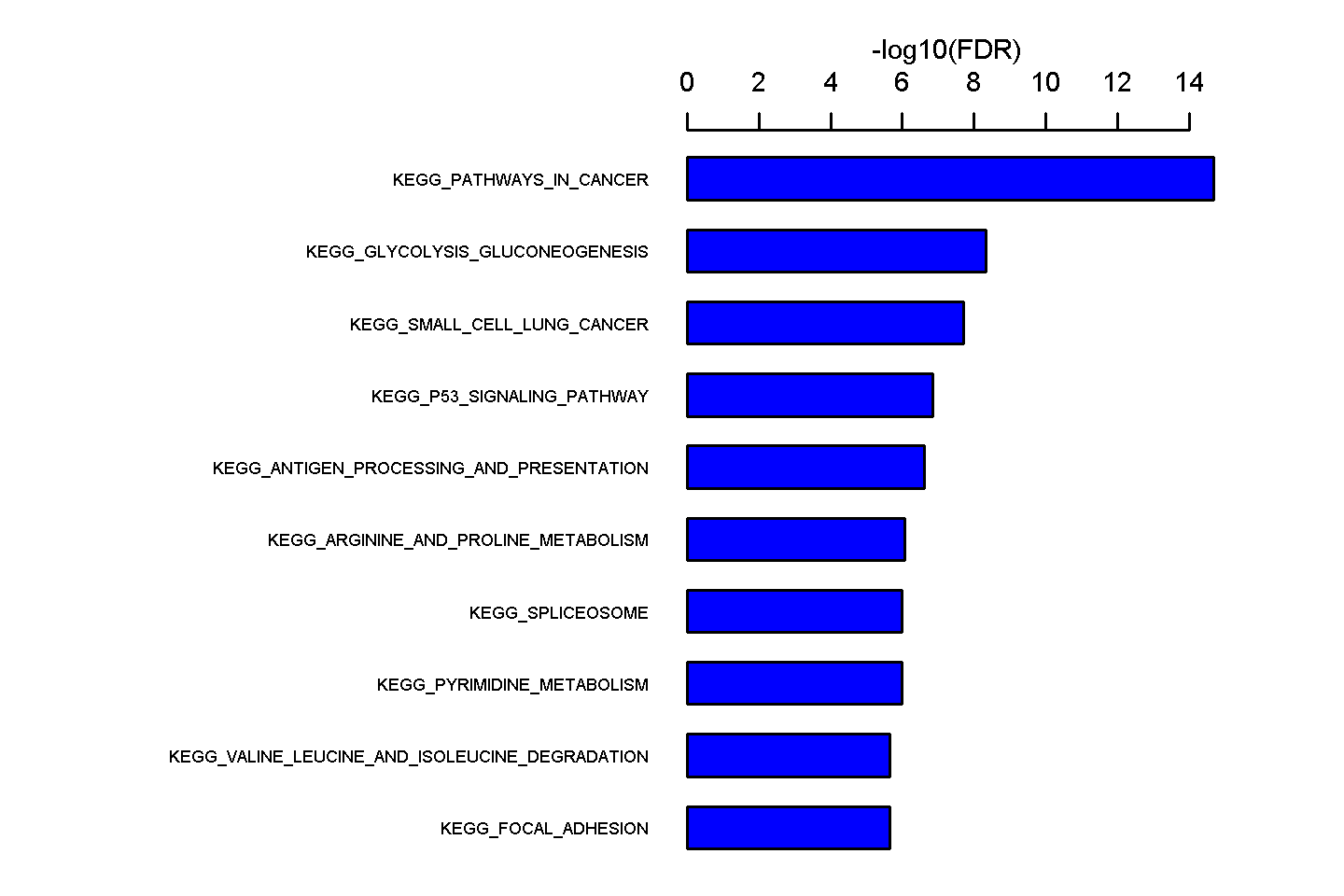


图5-2 显著富集的Pathway统计\_B\_vs\_A（Top 10）

图5-2中列举了在差异基因中富集程度最显著的通路。在图5-2中，横坐标为富集的显著性水平FDR（以10为底的负对数变换），纵坐标为显著富集的通路；排名越靠前，代表在统计学上该通路的影响或变化越显著。

1. **结果文件及保存位置**

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 文件名称 | 文件内容 | 保存位置 |
| 1 | RawData.xlsx | 原始芯片读数数据，已经RMA归一化处理 | 报告与附件\\原始及分析数据 |
| 2 | FilteredData.xlsx | 经背景噪音过滤和变异系数过滤后的数据 |
| 3 | DiffGene.xlsx | 筛选的差异基因及相对表达量列表 |
| 4 | Diffall.xlsx | 差异基因列表 |
| 5 | Signal\_Histogram.pdf | 质量控制：信号强度分布曲线图 | 报告与附件\\统计图（高清600dpi）\\质量控制 |
| 6 | Relative\_Signal\_Box\_Plot.pdf | 质量控制：相对对数信号强度箱线图 |
| 7 | Pearson\_s\_Correlation.pdf | 质量控制：样本间皮尔森相关系数分布图 |
| 8 | Principal\_Component\_Analysis.pdf | 质量控制：主成分分析（PCA）得分图 |
| 9 | Volcano\_Plot.pdf | 差异分析：火山图 | 报告与附件\\统计图（高清600dpi）\\差异分析 |
| 10 | Scatter\_Plot.pdf | 差异分析：散点图 |
| 11 | Heatmap.pdf | 差异分析：聚类分析结果 |
| 12 | Enrichment\_analysis.xlsx | GO&Pathway注释与富集分析结果 | 报告与附件\\功能及通路注释 |
| 13 | GO.pdf | 显著富集的GO term统计（Top 10） |
| 14 | Pathway.pdf | 显著富集的Pathway统计（Top 10） |
| 15 | 说明文档.pdf | - | 报告与附件 |
| 16 | M-GSGC0174590-生信分析报告.docx |

1. **参考文献**

[1] Gautier, L., et al., affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 2004. 20(3): p. 307-15.

[2] Sarkar, D., Lattice : multivariate data visualization with R. Use R! xvii, 265 pages.

[3] Hansen, M., et al., pcaGoPromoter--an R package for biological and regulatory interpretation of principal components in genome-wide gene expression data. PLoS One, 2012. 7(2): p. e32394.

[4] Jolliffe, I.T., Principal component analysis. 2nd edition. ed. Springer series in statistics. xxix, 487 pages.

[5] Bolstad, B.M., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 2003. 19(2): p. 185-93.

[6] Irizarry, R.A., et al., Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res, 2003. 31(4): p. e15.

[7] Irizarry, R.A., et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003. 4(2): p. 249-64.

[8] Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015. 43(7): p. e47.

[9] Warnes, G.R., et al., gplots: Various R Programming Tools for Plotting Data. 2016.

[10] Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. 25(1): p. 25-9.

[11] Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

[12] Kanehisa, M., et al., KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012. 40(Database issue): p. D109-14.

[](http://gcloud.taogene.com)

1. Gene Matrix云平台致力于打造一个服务医疗工作者的专业生物信息分析平台。专业团队精心打造多个生物信息分析应用模块，采用可视化交互界面，操作简易流畅，提供多种标准生物信息分析服务，个性化展示分析结果，满足客户日常科研需求。 [↑](#footnote-ref-1)
2. Kyoto Encyclopedia of Genes and Genomes (KEGG): <http://www.kegg.jp/> [↑](#footnote-ref-2)
3. BioCarta Pathways: <http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways> [↑](#footnote-ref-3)