上海欧易生物医学科技有限公司
上海浦东张江高科技园区祖冲之路1505弄138号6层
电　　话：021-38762096　021-38760130
免费热线：4006400826　4006400860

lncRNA 芯片数据查看说明

数据文件夹中包含内容

文件	描述
1 原始数据	芯片杂交后获得的原始数据。
2 标准化数据	样本的原始信号值，标准化信号值以及检出情况。
3 差异筛选	根据芯片确认单中提供的分析要求进行的比较数据，在设生物学重复下，利用火山图展示差异表达情况
4 GO 分析	对差异基因进行 GO 富集分析，统计每个 GO term 中所包括的差异基因个数，并用统计检验的方法计算每个 GO term 中差异基因富集的显著性。
5 Pathway 分析	结合 KEGG 数据库，对差异基因进行 Pathway 显著性分析。分析每个 Pathway 中所包含的差异基因个数，用统计检验计算每个 Pathway 中差异基因富集的显著性。
6 聚类分析	对差异表达基因进行非监督层次聚类，用热图展示。要求设生物学或者技术重复。

1、原始数据(该文件用 excel 打开)

利用 Feature Extraction 软件 (version10.7.1.1, Agilent Technologies)从扫描图片上提取得到的数据。具体的表头解释见“原始数据”文件夹中“原始数据表头说明”文件。

2、标准化数据(该文件用 excel 打开)

将原始数据导入 Genespring 软件(version 12.5,Agilent)，利用 quantile 的方法进行标准化后得到的结果，包含原始信号值，标准化信号值，检出情况以及详细的注释信息。

表头	示例
ProbeName：探针号	A_23_P326296
样本名+.txt(raw)：原始信号值	5000

样本名+.txt (normalized)：用 quantile 法标准化的信号值，并且经过 log2 转换	12
gIsWellAboveBG_Call：样本的旗标 ✧ P 代表信号与背景差异显著 ✧ A 代表信号与背景差异不显著	P
others：见芯片注释信息(下表)	

芯片注释信息

表头	含义	示例
TargetID	检测的 non-codingRNA 的 ID 号	TCONS_00011455
Probe_Sequence	探针序列	AGAGTTAGAGAAAGCAGCAAAGACTACAAGACTTCAGAGTCACAGCTGGCTACAAACCCA
lincID	LincRNA ID 号	TCONS_00011455
UCSCID	USUC 数据库登录号	uc002xip.2
Ensembl_Transcript_ID	Ensembl 数据库登录号	ENST00000455973
RefACC	Refseq 登录号	NR_026903
UniGeneID	UniGene ID 号	Hs.79110
GenbankID	Genbank 登录号	NR_026903
TIGRID	TIGR 登录号	THC2700111
Description.From.NCBI	来源于 NCBI 描述	uncharacterized LOC441204
Description.from.other.source	其他来源描述	Homo sapiens uncharacterized locus LOC441204 (LOC441204), non-coding RNA [NR_015364]
havana_transcript	Havana 登录号	OTTHUMT00000412363.1
Description.from.paper	文献中描述	
EntrezGeneID	EntrezGene 登录号	5473
GeneSymbol	基因缩写	PPBP
GeneName	基因名称	pro-platelet basic protein (chemokine (C-X-C motif) ligand 7)
DESCRIPTION	描述	Homo sapiens pro-platelet basic protein (chemokine (C-X-C motif) ligand 7) (PPBP), mRNA [NM_002704]
GO_ID	GO 分类号	GO:0002576(platelet degranulation)|GO:0005355 (glucose transmembrane transporter

		activity)|GO:0005576(extracellular region)
chrom	所在染色体	9
Strand	正负链	+
TxStart	转录起始	85068667
TxEnd	转录终止	85070166
GenomicCoordinates	染色体位置	
Cytoband	Cytoband 登录号	hs|9q21.32
Accessions	登录号	ens|ENST00000422010|tc|THC2700111|linc|TCONS_00015701|linc|TCONS_00016624
Probe.type	探针类型	non-coding RNA
Sequence.source	序列来源	agilent G3 V2

3、差异筛选

- **差异筛选（该文件用 excel 打开）**

在筛选差异基因之前，先进行探针过滤，用于比较的每组样本中至少有一组75%标记为“P”的探针留下进行后续分析。对于有生物学重复的分析，利用 T 检验得到的差异显著性 P 值和标准化信号值的差异倍数 Fold change 值进行筛选，标准为 Fold change 值>= 2.0 且 P 值<= 0.05。

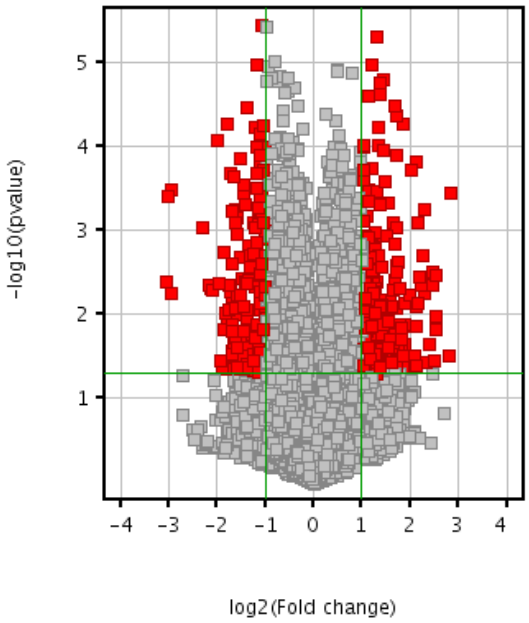
表头	示例
ProbeName：探针号	A_23_P326296
p：TTest 统计分析后基因的显著性 P 值，p 值越小显著性差异越大	0.01
FC: 组间差异倍数(负值表示下调)	-3
Log FC: 组间差异倍数绝对值，经 log2 转换(负值表示下调)	1
FC(abs): 组间差异倍数绝对值，非 log2 转换	2.5
Regulation：组间差异趋势，up 为上调，down 为下调	UP
样本名+.txt(normalized)：用 quantile 法标准化的信号值，且经过 log2 转换	12
gIsWellAboveBG_Call：样本的旗标 ✧ P 代表信号与背景差异显著 ✧ A 代表信号与背景差异不显著	P
others：芯片注释信息	
差异分析筛选标准：	

- 对于生物学重复够做统计分析的组间比较，FC(abs) 大于或等于 2 且 p 小于等于 0.05，且用于比较的两组数据至少有一组数据 75%样本标记为 P。
- 对于生物学重复不够作统计分析的组间比较，FC(abs) 大于或等于 2，且用于比较的两组数据至少有一组数据 100%样本标记为 P。

注：以上筛选阈值仅供参考，您可根据实验需要调整筛选标准，以得到合适的差异表达基因。

- **火山图：**

将比较所产生的差异表达情况反映到火山图中，X 轴是差异倍数以 2 为底取对数的值，Y 轴是 p-value 值以 10 为底取负对数的值。在设生物学重复的情况下（即使用统计分析法分析差异表达）将给出火山图。



灰色点：Fold change 绝对值＜2，P 值＞0.05 的基因。

红色点：Fold change 绝对值≥2，P 值≤0.05 的基因，这些为显著性差异基因。

4、GO 分析

对差异基因进行 GO 分析，从而对这个基因的功能进行描述。GO 包括三大板块，Biological Process，Cellular Component 和 Molecular Function，所以有三类结果。统计每个 GO 条目中所包括的差异基因个数，并用统计检验的方法计算每个 GO 条目中差异基因富集的显著性。计算的结果会返回一个富集显著性的 P 值，小的 p 值表示差异基因在该 GO 条目中出现了富集。可以根据 GO 分析的

结果结合生物学意义从而挑选用于后续研究的基因。**结果包含网页和条形图图片。**

网页结果展示：

红色标注的是需要重点关注的！

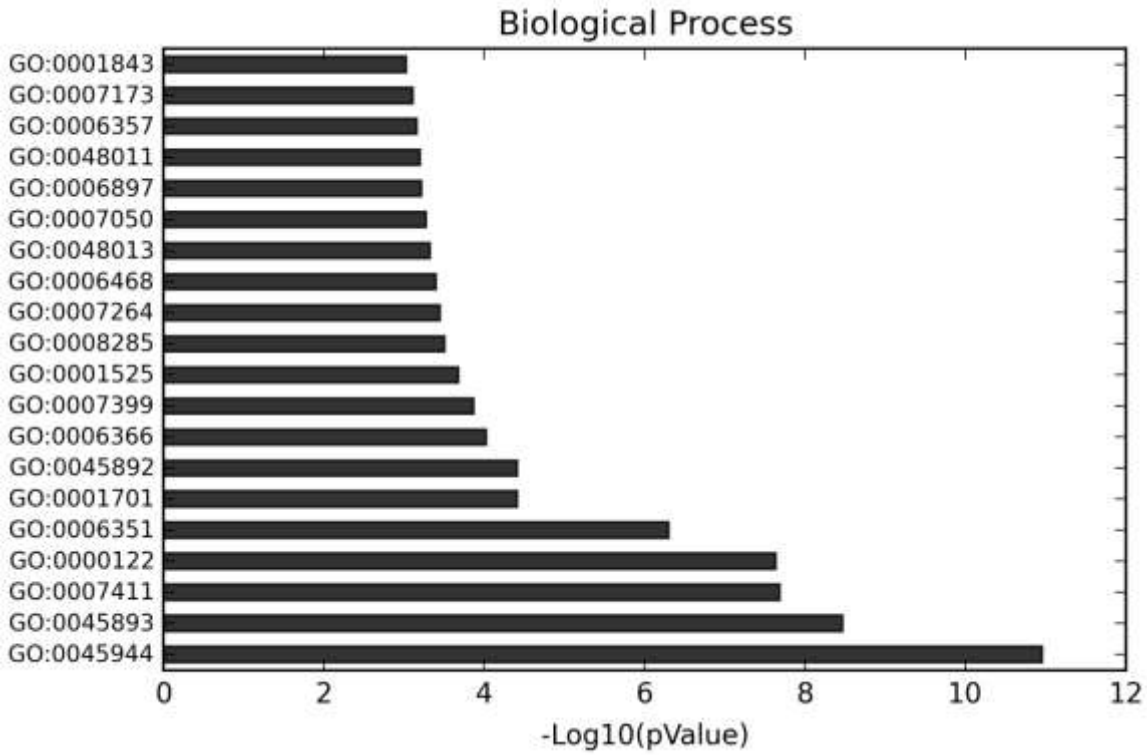
Terms(Sorted by pvalue)

Rank	TermID	TermDescription	List Hits	List Total	Population Hits	Population Total	FoldEnrichment	P-value	FDR_bh
0	GO:0045944	positive regulation of transcription from RNA polymerase II promoter	632	8403	871	16329	1.44	1.1E-11	4.4E-08
1	GO:0045893	positive regulation of transcription, DNA-templated	377	8403	493	16329	1.51	3.4E-09	6.6E-06
2	GO:0007411	axon guidance	298	8403	378	16329	1.55	2.1E-08	2.3E-05

表头	示例
TermID：条目在 Gene Ontology 的登录号	GO:0009612
TermDescription：该条目的描述	**cell growth**
List Hits：该 GO 条目中靶基因的个数	**11**
List Total：注释到 GO 的总靶基因数	84
Population Hits：整张芯片注释到该条目中的基因个数	99
Population Total：整张芯片注释到 GO 的基因数	5085
FoldEnrichment：(list hits/List total)/(population hits/poplation total)	1.55
P-value：富集显著性 P 值，P≤0.05 表示显著富集	**3.71E-06**
FDR_bh:采用 Benjamini-Hochberg 方法矫正后的 P 值	**4.4E-08**

图片结果展示：

按照 pValue 排序，对 rank 在前 20 位的条目(若总条目不大于 20，则全部展示)利用条形图进行展示。



5、Pathway 分析

利用 KEGG 数据库对差异基因进行 Pathway 分析，并且用统计检验的方法计算每个 Pathway 条目中差异基因富集的显著性。计算的结果会返回一个富集显著性的 P 值，小的 P 值表示差异基因在该 pathway 中出现了富集。Pathway 分析对实验结果有提示的作用，通过差异基因的 Pathway 分析，可以找到富集差异基因的 Pathway 条目，寻找不同样品的差异基因可能和哪些细胞通路的改变有关。**结果包含网页、csv 文件、条形图图片和通路图。**

网页结果展示：

红色标注的是需要重点关注的！

Terms(Sorted by pvalue)

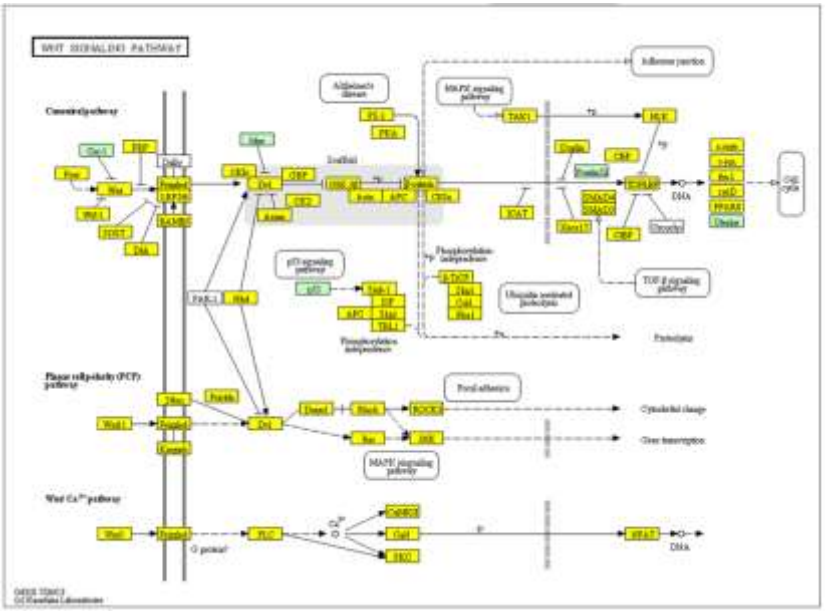
Rank	TermID	TermDescription	List Hits	List Total	Population Hits	Population Total	FoldEnrichment	P-value	FDR_bh
0	path:hsa05200	Pathways in cancer	292	3434	398	6899	1.52	1.6E-07	4.5E-05
1	path:hsa05205	Proteoglycans in cancer	159	3434	204	6899	1.59	1.3E-05	0.0018
2	path:hsa04360	Axon guidance	107	3434	127	6899	1.71	4E-05	0.0038
3	path:hsa04310	Wnt signaling pathway	113	3434	140	6899	1.64	8.4E-05	0.006

表头	示例
TermID：条目在 KEGG 数据库的登录号	path：hsa04514
TermDescription：该条目的描述	**Endocytosis**
List Hits：该条目中靶基因的个数	**37**

List Total：注释到 KEGG 的总靶基因数	546
Population Hits：整张芯片注释到该条目中的基因个数	203
Population Total：整张芯片注释到 KEGG 的基因数	5981
FoldEnrichment：(list hits/List total)/(population hits/poplation total)	1.59
P-value：富集显著性 P 值，P≤0.05 表示显著富集	**3.71E-06**
FDR_bh:采用 Benjamini-Hochberg 方法矫正后的 P 值	**4.4E-08**

点击通路条目的名称，则超链接进通路图，结果展示如下：

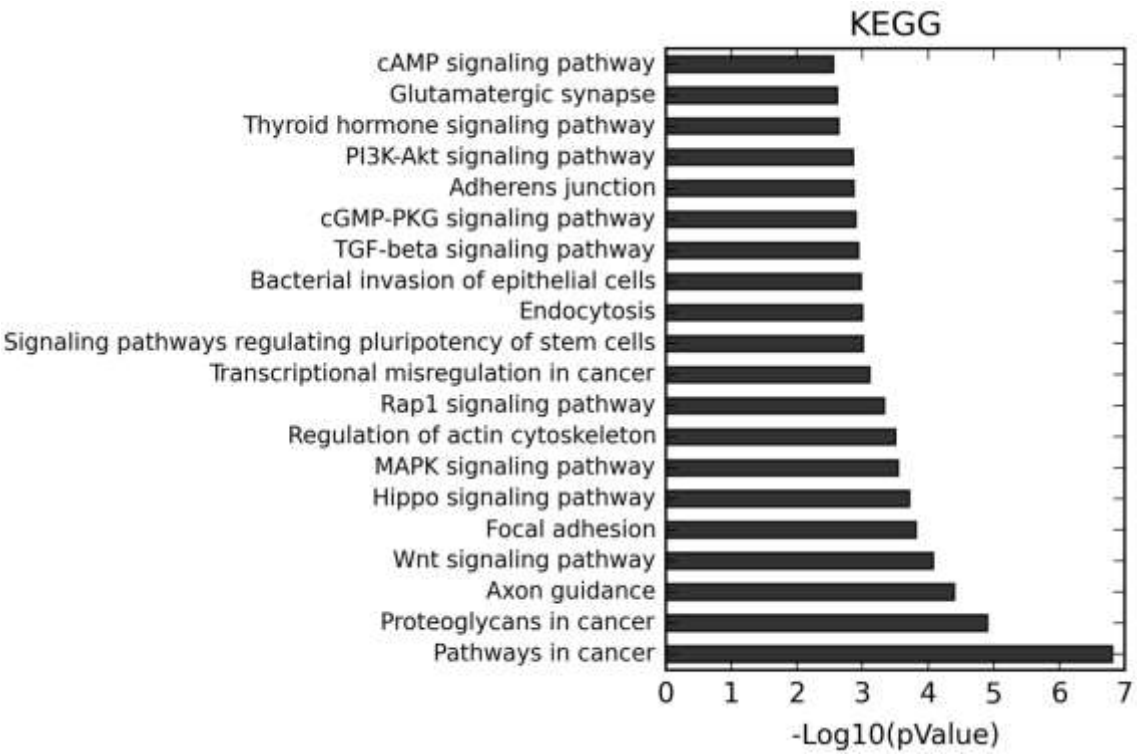
path:hsa04310 Wnt signaling pathway　点击此处，则可以得到如下通路图



黄色表示该靶基因在该通路中的位置

条形图图片结果展示：

按照 pValue 排序，对 rank 在前 20 位的条目(若总条目不大于 20，则全部展示)利用条形图进行展示。

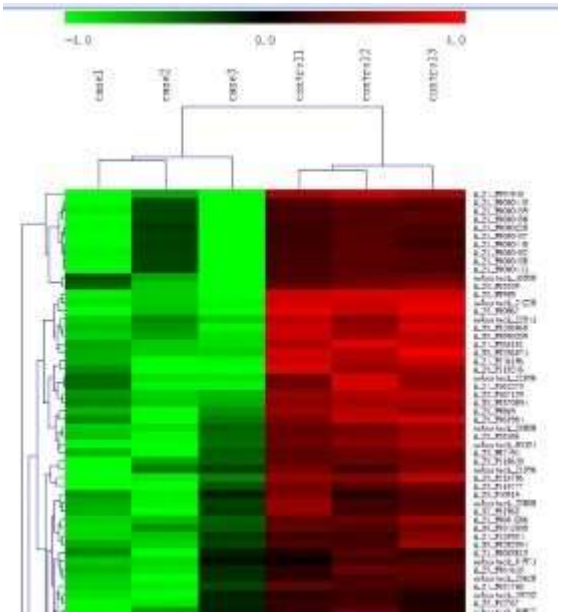


6、聚类分析

对差异表达基因进行非监督层次聚类。计算多个样品两两之间的距离，构成距离矩阵，合并距离最近的两类为一新类，计算新类与当前各类的距离，再合并、计算，直至只有一类为止，用挑选的差异基因的表达情况来计算样品直接的相关性，一般来说，同一类样品能通过聚类出现在同一个簇中，聚在同一个簇中的基因可能具有相似的生物学功能。用 Heatmap 展示。

通常初步分析后的差异基因比较多，初步的聚类结果只是针对样本进行聚类。如果老师想对样本和基因进行双向聚类，则需要您指定用于聚类分析的样本和基因，每一次聚类至少包含 3 个以上样本，基因数量尽量不要超过 1000。我们建议老师也可以尝试利用“报告\实验文件\3 常用软件\1 MEV 软件”进行聚类，该文件夹中有对应程序以及 PPT。

上海欧易生物医学科技有限公司
上海浦东张江高科技园区祖冲之路1505弄138号6层
电　　话：021-38762096　021-38760130
免费热线：4006400826　4006400860



色阶表示基因表达量从相对低（绿）到相对高（红）变化。行名代表探针名称，列名代表样本名称。