

iSeq 平台分析简介

--北大生物信息平台

介绍:

iSeq 是在北大生物信息平台开发的 RNA-seq 在线分析工具，此工具主要用于分析得到基因表达谱后的作图，差异表达基因检测，功能富集等。可以方便一键式的完成常用的 RNA-seq 分析和作图。

网址: http://202.205.131.32:3838/RNA-seq_V2/



Table of Contents

1. Upload : 上传数据	3
1.1 Expression File of Genes(.csv) -- 表达谱文件	3
1.2 Condition File of Samples(.csv) - 样本信息	3
2. Quality Control: 数据质量控制	4
2.1 summary	4
2.2 表达分布图	4
2.3 层次聚类	4
2.4 相关性聚类	5
3. Normalization : 表达谱矫正	5
3.1 boxplot	5
3.2 MA-plot	6
4. DEG calling : 差异表达基因检测	6
4.1 DESeq (需要等待 2 分钟)	6
5. Function : 功能富集分析	8
6. Plots : RNA-seq 常用作图	9
6.1 barplot	9
6.2 heatmap	9
6.3 PCA 主成分分析	10

Upload

Quality Check

Normalization

DEG calling

Function

Plots

1.1 Expression File of Genes(.csv) -- 表达谱文件

<行为基因, 列为样本>

	A	B	C	D	E	F	G	H
1		AD1	AD2	AD3	EM1	EM2	EM3	EM4
2	0610005C13Rik	9	14	21	17	30	19	23
3	0610007P14Rik	428	508	404	374	527	539	501
4	0610009B22Rik	255	257	253	169	280	230	224
5	0610009L18Rik	49	45	60	48	57	60	61
6	0610009O20Rik	676	786	691	617	755	709	712
7	0610010B08Rik	14	14	7	7	5	10	5
8	0610010F05Rik	1204	1099	950	2514	2752	2394	3376
9	0610010K14Rik	390	485	432	727	1167	1001	804
10	0610011F06Rik	347	444	426	185	296	261	232
11	0610012G03Rik	631	794	745	488	628	672	581

.cs 格式的文件，即以逗号分隔的文本文件。

示例文件路径: <http://202.205.131.32:3838/NAT.Condition.csv>

实例数据来自 <http://www.nature.com/neuro/journal/v16/n4/full/nn.3332.html>

上传成功后:

iSeq Upload Quality Check Normalization DEG calling Function Plots About

***Choose A Expression File of Genes(csv)**

选择文件 NAT.rawReads.csv

Upload complete

Choose a Condition File of Samples(csv)

选择文件 NAT.Condition.csv

Upload complete

Filter the Data

	AD_1	AD_0	AD_2	EM_1	EM_0	EM_2	EM_3
Condition:	AD	AD	AD	EM	EM	EM	EM

Show 10 entries

Search:

	AD1	AD2	AD3	EM1	EM2	EM3	EM4
0610005C13Rik	9	14	21	17	30	19	23
0610007P14Rik	428	508	404	374	527	539	501
0610009B22Rik	255	257	253	169	280	230	224
0610009L18Rik	49	45	60	48	57	60	61
0610009O20Rik	676	786	691	617	755	709	712
0610010B08Rik	14	14	7	7	5	10	5
0610010F05Rik	1204	1099	950	2514	2752	2394	3376
0610010K14Rik	390	485	432	727	1167	1001	804
0610011F06Rik	347	444	426	185	296	261	232
0610012G03Rik	631	794	745	488	628	672	581

Show 1 to 10 of 24,015 entries

Previous 1 2 3 4 5 ... 2402 Next

Download the data.

2. Quality Control: 数据质量控制

2.1 summary

对于表达谱数据的总结，给出了每个样本所有基因的统计数据。

Min. 最小值

1st Qu. 25%分位数

Median 中位数

3rd qu. 75%分位数

Max. 最大值

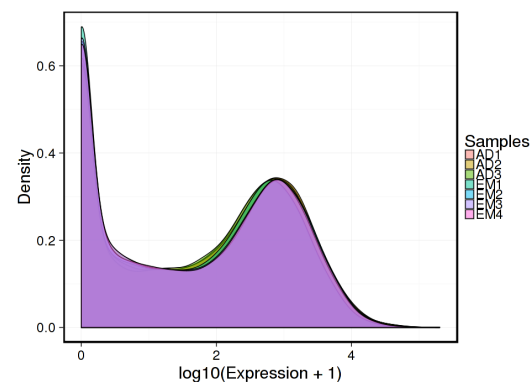
Summary:

AD1		AD2		AD3		EM1	
Min.	: 0.0	Min.	: 0.0	Min.	: 0.0	Min.	: 0.0
1st Qu.	: 1.0	1st Qu.	: 1.0	1st Qu.	: 1.0	1st Qu.	: 0.0
Median	: 99.0	Median	: 100.0	Median	: 96.0	Median	: 54.0
Mean	: 966.9	Mean	: 988.2	Mean	: 957.2	Mean	: 739.6
3rd Qu.	: 797.0	3rd Qu.	: 807.5	3rd Qu.	: 780.0	3rd Qu.	: 633.5
Max.	: 202161.0	Max.	: 105202.0	Max.	: 109825.0	Max.	: 116731.0
EM2		EM3		EM4			
Min.	: 0.0	Min.	: 0.0	Min.	: 0.0		
1st Qu.	: 0.0	1st Qu.	: 0.0	1st Qu.	: 0.0		
Median	: 68.0	Median	: 64.0	Median	: 67.0		
Mean	: 915.4	Mean	: 861.3	Mean	: 885.6		
3rd Qu.	: 787.0	3rd Qu.	: 737.0	3rd Qu.	: 766.0		
Max.	: 150246.0	Max.	: 147053.0	Max.	: 139764.0		

2.2 表达分布图

每个样本所有基因的表达量分布。在 RNA-seq 中我们的假设是所有的样本表达分布相似，所有如果有个别样本表达谱与其他样本差别很大，需要考虑是否去除此样本。

Expression Distribution:

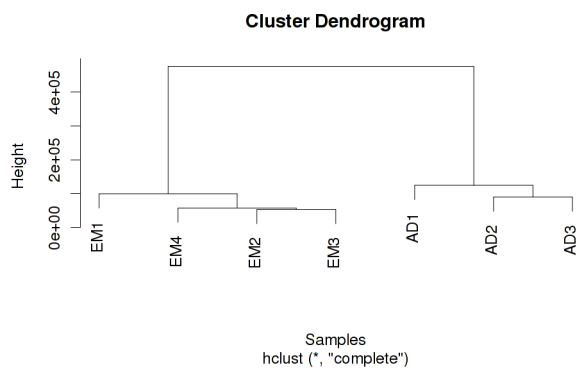


横坐标为 $\log_{10}(\text{表达量}+1)$ ，纵坐标为有此表达量的基因的比例，每种颜色代表一个样本。

2.3 层次聚类

层次聚类可以评估样本间表达谱的像似程度，越像似的样本他们在层次聚类中距离越近。此例中所属 EM 的样本聚在一起，AD 样本聚在一起。

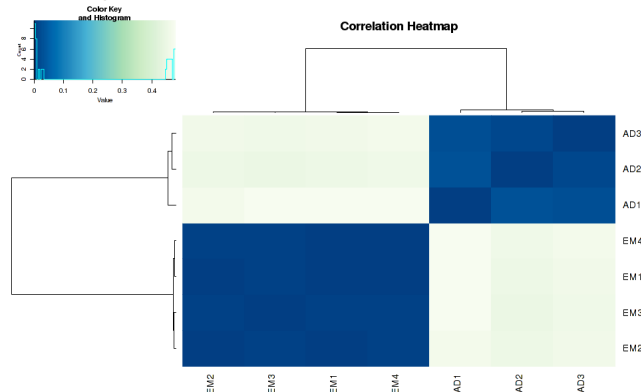
hclust



2.4 相关性聚类

类似于层次聚类，也是评估样本间像似程度的。越相近的两个样本它们所在的方块颜色越深。

Clust heatmap

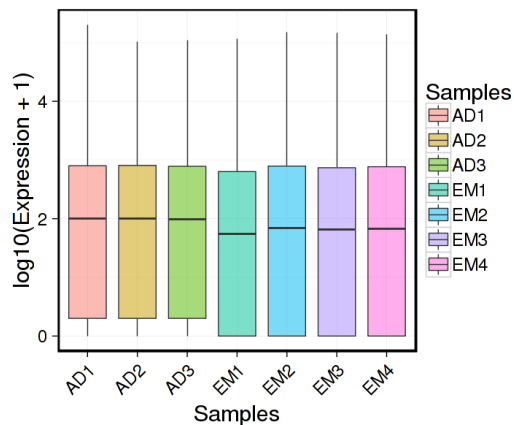


3.Normalization：表达谱矫正

3.1 boxplot

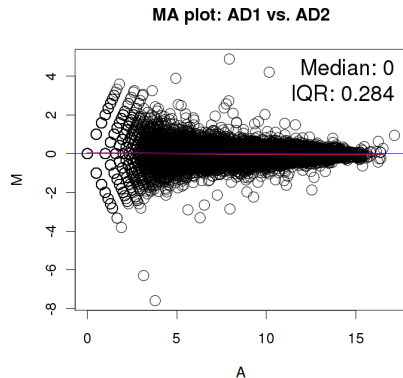
基因表达分布箱线图，每个 box 代表一个样本，box 中间的粗线为中位数，box 上下边为 25%和 75%分位数，线的顶端和低端代表最大值和最小值。一批样本矫正较好的情况是表达分布相似，及各个样本箱线较为统一。

Box plot:



3.2 MA-plot

另一个评价矫正好坏的图，具体可见：https://en.wikipedia.org/wiki/MA_plot
横坐标为两个样本 log 的表达量相加除以 2，纵坐标为 log 的表达量相减，每个点为一个基因，红线是所有点拟合出来的。其中红线与 $y=0$ 的线重合度越好代表矫正的越好。



4.DEG calling : 差异表达基因检测

4.1 DESeq (需要等待 2 分钟)

DESeq 为一个差异基因表达检测的软件。选定好需要比较的两个生物背景 (condition1 和 condition2) 就可以进行差异基因表达的检测。

参数设置：

Choose condition1:

AD

Choose condition2:

EM

Optional filter:
padj cutoff

0.05

foldChange cutoff

less than 0.25 or greater than 4

baseMean cutoff

10

Run to detect

Download Table

padj cutoff：矫正后的 p value，越小出来的差异基因统计学越显著

foldChange cutoff：两个生物背景的基因表达差异倍数，fold change 定义为 condition2 中的基因表达量除以 condition1 中的基因表达量，如果没变此值为 1，偏离 1 越远（越大或越小）代表差异越大。此值通常为设置为小于 0.5 或大于 2，当要找差异很大的基因时，此值可以设置成小于 0.25 或者大于 4。

baseMean cutoff：baseMean 表示两个 condition 下所有样本基因表达量的平均值，baseMean 主要是想要过滤掉一些表达量非常低的基因，因为这些基因的表达数值较小，往往会有非常大的 fold change（计算时由于分母小的导致 fold change 非常大，而实际很可能是随机结果）。此值对于上传为 raw reads 的数据通常设置为 10。

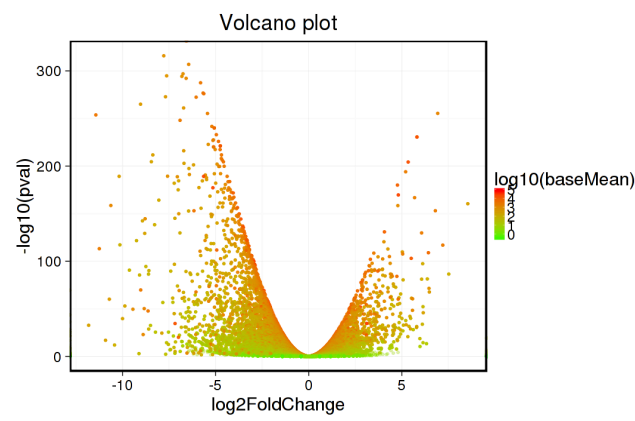
点击“Run to detect”即可获得差异表达的列表如下图，点击“Download Table”即可下载此表。

	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
Cntnap1	3176.143	7310.333	75.50	0.010327846	-6.597317	0.000000e+00	0.000000e+00
Ddn	13368.286	30755.667	327.75	0.010656573	-6.552113	0.000000e+00	0.000000e+00
S100b	1264.429	2932.667	13.25	0.004518072	-7.790077	1.242254e-316	8.344221e-313
Plekhb1	2703.571	6213.667	71.00	0.011426426	-6.451482	1.261489e-307	6.355069e-304
Pcp4l1	1618.429	3730.333	34.50	0.009248503	-6.756564	1.044693e-297	4.210322e-294
Ankrd33b	1028.571	2384.000	12.00	0.005033557	-7.634206	1.405367e-295	4.719926e-292
Itfpa	1480.286	3413.667	30.25	0.008861439	-6.818243	5.468982e-295	1.574364e-291
Gfap	2561.429	5895.000	61.25	0.010390161	-6.588638	6.106919e-293	1.538256e-289
Zfp365	5700.429	12992.667	231.25	0.017798502	-5.812100	2.797389e-288	6.263354e-285
Rasgrp1	5168.143	11752.667	229.75	0.019548755	-5.676779	1.719610e-277	3.465187e-274

Showing 1 to 10 of 2,829 entries

Previous 1 2 3 4 5 ... 283 Next

此表为差异表达的列表，每行代表一个差异表达基因，baseMean 为平均表达量，baseMeanA 为 condition1 中样本的平均表达量，baseMeanB 为 condition2 中样本的平均表达量，foldChange 为 baseMeanB/baseMeanA，log2FoldChange 为 foldChange 取 log2 后的值，pval 为差异检测显著性 p 值，padj 为 p 值经过矫正后的显著性值。



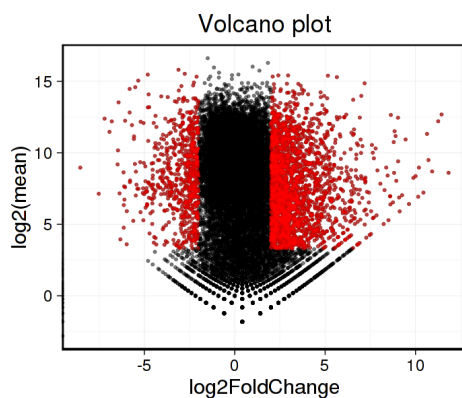
4.2 Fold Change （适用于无生物重复）

用于简单的差异表达检测，只采用 Fold change 来判断基因是否在两种 condition 下表达有差异。

	mean	C1	C2	foldChange	log2FoldChange
0610040J01Rik	54.42857	104.000000	17.25	6.02898551	2.591915
1110008P14Rik	891.42857	1757.333333	242.00	7.26170799	2.860309
1110015O18Rik	53.85714	8.333333	88.00	0.09469697	-3.400538
1110038B12Rik	372.14286	88.333333	585.00	0.15099715	-2.727407
1110046J04Rik	27.28571	49.333333	10.75	4.58914729	2.198226
1190002F15Rik	84.85714	5.000000	144.75	0.03454231	-4.855491
1300002K09Rik	37.85714	1.333333	65.25	0.02043423	-5.612868
1500015O10Rik	104.14286	182.333333	45.50	4.00732601	2.002640
1600016N20Rik	14.14286	1.666667	23.50	0.07092199	-3.817623
1600029O15Rik	137.42857	39.666667	210.75	0.18821669	-2.409534

Showing 1 to 10 of 2,830 entries

Previous 1 2 3 4 5 ... 283 Next













5.Function：功能富集分析

功能富集是对一组基因的功能进行探究，通常将差异表达的基因分为上调和下调分别进行功能富集分析。其中 Gene Ontology 与 KEGG 富集为大家经常使用的数据库。

以 david (GO:BP 富集)结果示例：

EM > AD

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle	RT		111	2.0	1.5E-50	2.9E-47
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle process	RT		82	1.5	2.4E-41	2.4E-38
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle phase	RT		73	1.3	1.2E-38	8.2E-36
<input type="checkbox"/>	GOTERM_BP_FAT	M phase	RT		68	1.2	3.8E-38	1.9E-35
<input type="checkbox"/>	GOTERM_BP_FAT	cell division	RT		63	1.1	2.1E-33	8.4E-31
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic cell cycle	RT		58	1.1	3.8E-32	1.3E-29
<input type="checkbox"/>	GOTERM_BP_FAT	M phase of mitotic cell cycle	RT		51	0.9	1.9E-30	5.5E-28
<input type="checkbox"/>	GOTERM_BP_FAT	mitosis	RT		49	0.9	7.9E-29	2.0E-26
<input type="checkbox"/>	GOTERM_BP_FAT	nuclear division	RT		49	0.9	7.9E-29	2.0E-26
<input type="checkbox"/>	GOTERM_BP_FAT	organelle fission	RT		49	0.9	4.7E-28	1.0E-25

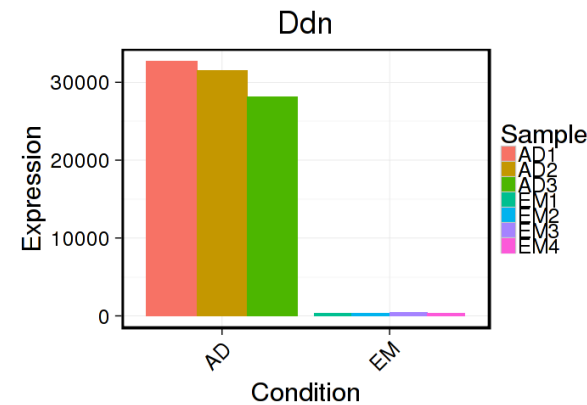
AD > EM

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamin
<input type="checkbox"/>	GOTERM_BP_FAT	metal ion transport	RT	115	0.6	1.6E-22	5.7E-19	
<input type="checkbox"/>	GOTERM_BP_FAT	ion transport	RT	156	0.8	4.2E-22	7.6E-19	
<input type="checkbox"/>	GOTERM_BP_FAT	transmission of nerve impulse	RT	76	0.4	4.3E-22	5.2E-19	
<input type="checkbox"/>	GOTERM_BP_FAT	cell-cell signaling	RT	85	0.5	2.8E-20	2.5E-17	
<input type="checkbox"/>	GOTERM_BP_FAT	cation transport	RT	122	0.6	3.3E-20	2.4E-17	
<input type="checkbox"/>	GOTERM_BP_FAT	synaptic transmission	RT	63	0.3	1.1E-19	6.5E-17	
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of system process	RT	65	0.3	5.1E-18	2.6E-15	
<input type="checkbox"/>	GOTERM_BP_FAT	potassium ion transport	RT	53	0.3	3.3E-15	1.5E-12	
<input type="checkbox"/>	GOTERM_BP_FAT	ion homeostasis	RT	76	0.4	6.9E-15	2.8E-12	
<input type="checkbox"/>	GOTERM_BP_FAT	chemical homeostasis	RT	86	0.5	3.9E-14	1.4E-11	

6.Plots : RNA-seq 常用作图

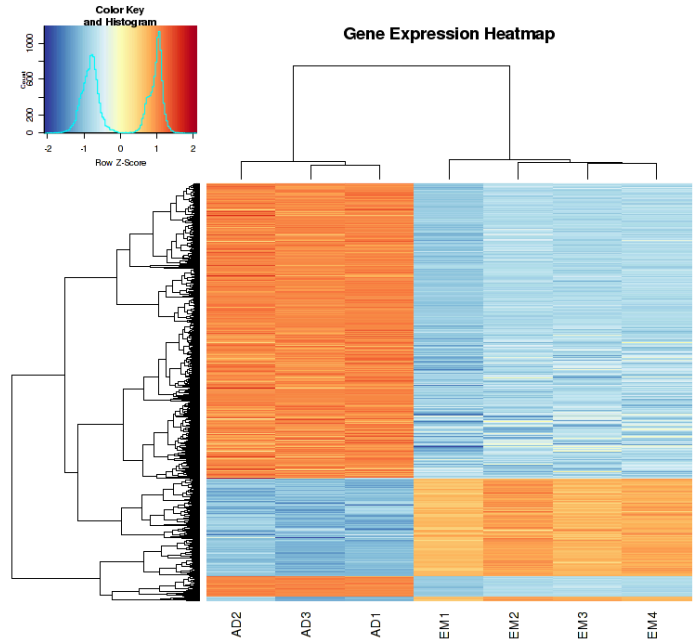
6.1 barplot

对差异表达的基因进行画图，横坐标为各个 condition 下的样本，纵坐标为表达量。此例可以看出基因 Ddn 在 AD 中的 3 个生物学重复表达远远高于 EM 中的 4 个生物学重复。



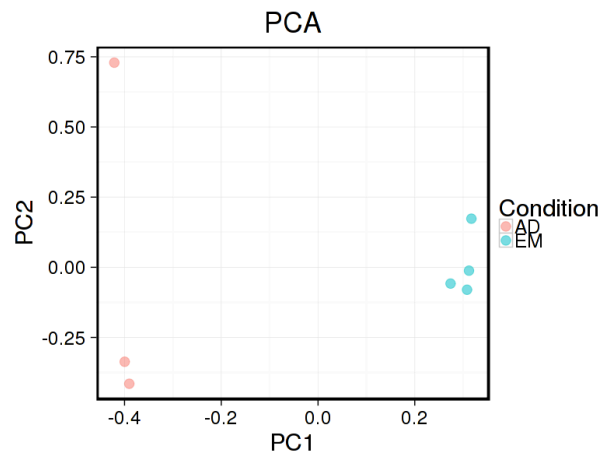
6.2 heatmap

基因表达热图，热图中每一行为一个基因，每一列为一个样本，染色代表表达量，越红表达量越高。热图一般用于整体上差异的观测，往往能发现一些特殊表达模式的基因簇。



6.3 PCA 主成分分析

将多为数据进行降维，已在低维空间上观测样本的相似性。



联系：

使用中如果发现任何 bug 或使用过程中的建议，请联系张超：

zhangchao3@hotmail.com