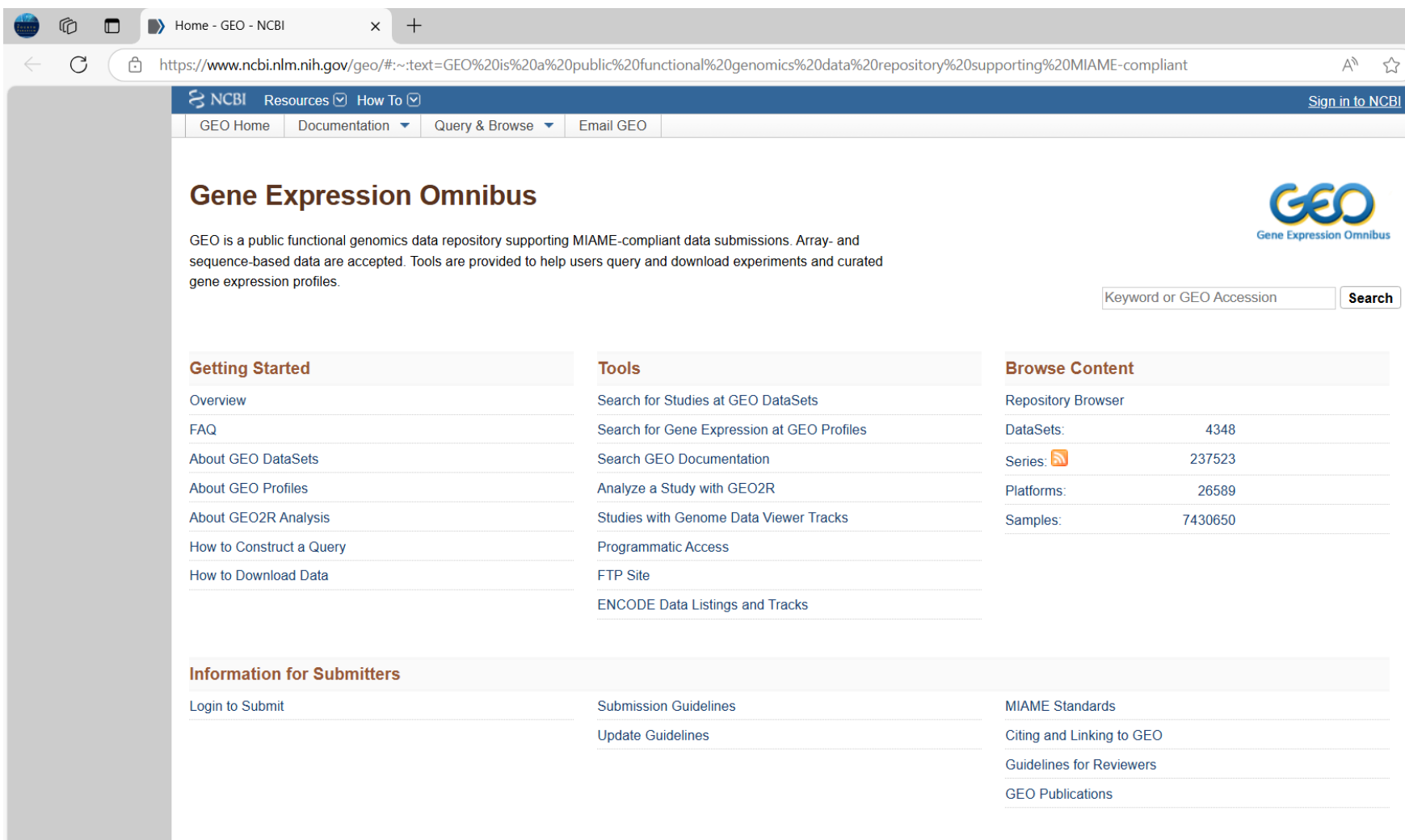


GEO2R

陈明杰

202410

Geo数据库简介



The screenshot shows the NCBI Gene Expression Omnibus (GEO) homepage. The browser address bar displays the URL: <https://www.ncbi.nlm.nih.gov/geo/#:~:text=GEO%20is%20a%20public%20functional%20genomics%20data%20repository%20supporting%20MIAME-compliant>. The page features a navigation bar with links to NCBI, Resources, and How To. Below the navigation bar, the main heading is "Gene Expression Omnibus". A description states: "GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles." A search bar is located on the right side of the main content area. The page is organized into three main columns: "Getting Started" (Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, How to Download Data), "Tools" (Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, Studies with Genome Data Viewer Tracks, Programmatic Access, FTP Site, ENCODE Data Listings and Tracks), and "Browse Content" (Repository Browser, DataSets: 4348, Series: 237523, Platforms: 26589, Samples: 7430650). At the bottom, there is a section for "Information for Submitters" (Login to Submit, Submission Guidelines, Update Guidelines, MIAME Standards, Citing and Linking to GEO, Guidelines for Reviewers, GEO Publications).

1, 芯片, 测序原始数据仓库
基因组, 转录组, 修饰等
不存蛋白质谱和代谢数据

2, 发文章一般都要上传原始数据
可重复性, 真实性

3, 大量数据共享, 可挖掘发文章
没有数据, 挖掘GEO
数据不够, GEO来凑

4, 数据质量参差不齐

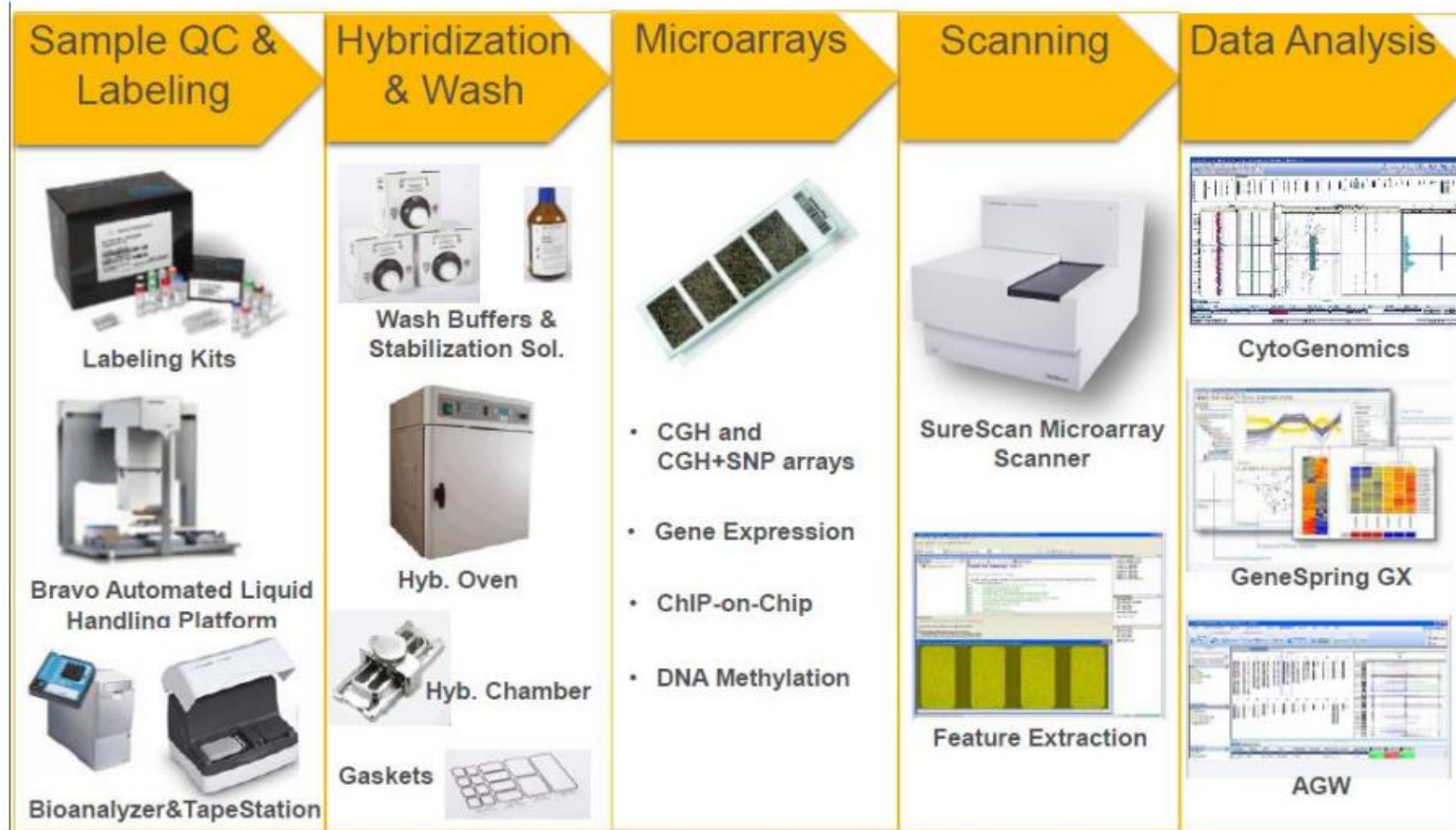
GEO2R简介

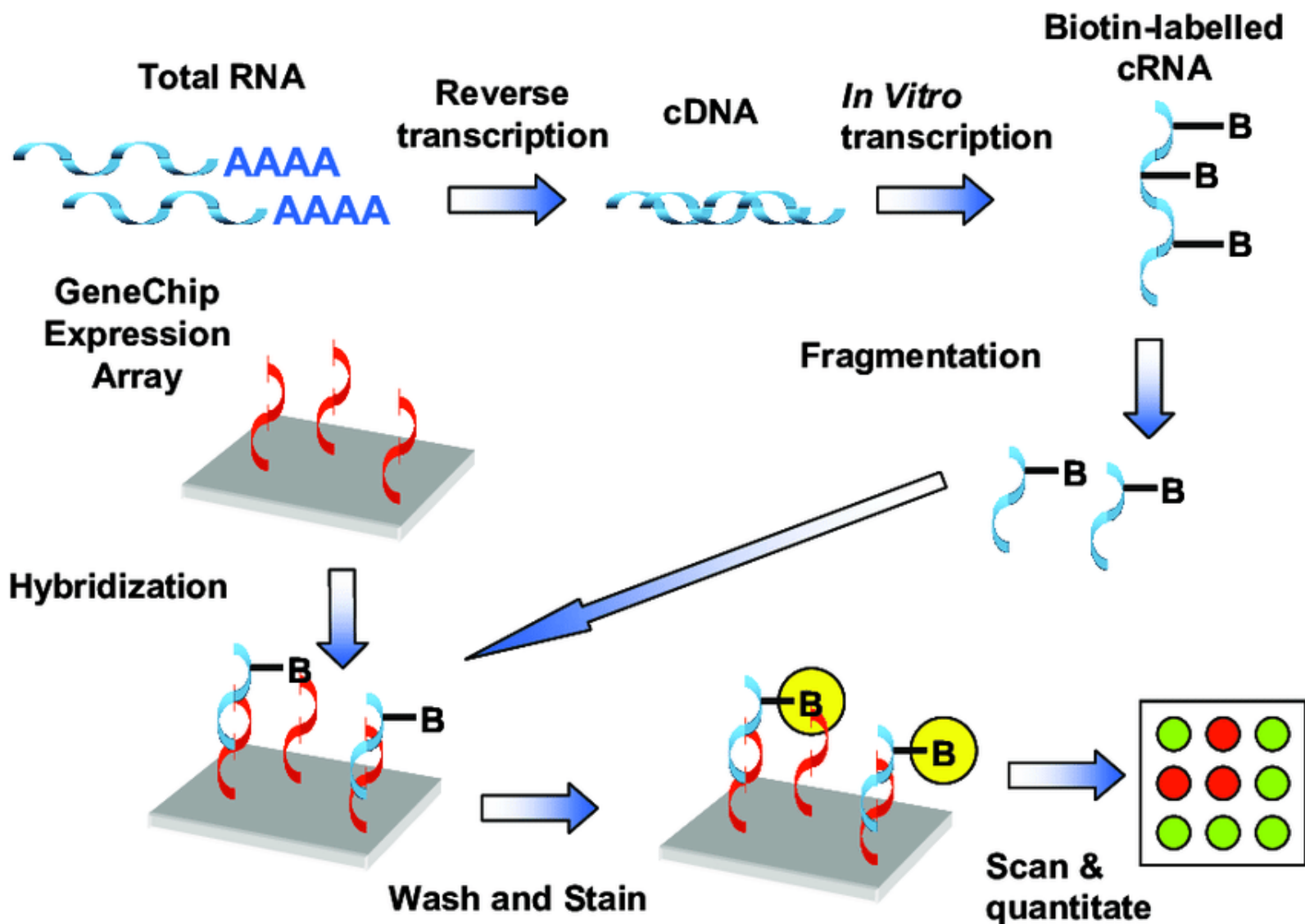
- GEO2R是一个交互式网络工具，允许用户比较GEO系列中的两个或更多样本组，以识别在实验条件下差异表达的基因。结果以按P值排序的基因表形式呈现，并以一系列图表的形式收集，以帮助可视化差异表达的基因并评估数据集质量。
- GEO2R使用GEOquery和limma来执行差异表达分析，使用原始提交者提供的处理后的数据表作为输入。GEOquery将GEO数据解析成R数据结构，这些数据结构可以被其他R包使用。limma进行差异分析，并在P值上应用多重测试校正，以帮助校正假阳性的发生。

GEO2R局限

- 结果跟文章可能不一致
 - 仅支持分组比较，不支持1 vs 1
 - 部分样品可能不可用
 - 仅一个series内的，不能跨series
-
- 建议： 下载原始数据分析，然后用geo2r验证

芯片（标记，杂交，扫描，数据分析）





芯片上探针是预先设计好的，固定的，长度~60 nt，探针荧光强度表示表达高低（可定制）

测序flowcell上“探针”是“长”出来的，长度25-150 nt，长出来的reads经过比对后获得每个基因来源的reads个数表示强度

简化：
150张芯片叠起来
从平面到立体

实操

GEO上传

1, Metadata

Series信息，实验设计，目的，物种，数据类型，贡献者等

用geo2r进行分析

2, Processed data

3, raw data

每个样品的信息

表达矩阵

原始数据cel文件

探针集

13个样品

61	ID_REF	GSM631755	GSM631756	GSM631757	GSM631758	GSM631759	GSM631760	GSM631761	GSM631762	GSM631763	GSM631764	GSM631765	GSM631766	GSM631767
62	1007_s_at	9.37362	9.27457	9.31466	8.81297	9.22501	8.6195	9.79252	8.71202	8.8878	9.41581	8.86084	10.1835	8.7555
63	1053_at	4.88624	4.57343	4.66537	4.41698	4.53738	4.89242	4.57879	4.3637	4.33044	4.33198	4.17537	4.39334	4.65933
64	117_at	4.96282	4.91762	4.9978	5.68013	7.17241	5.41695	5.21297	5.26965	5.54095	6.35407	5.62223	5.3208	5.87183
65	121_at	7.80585	7.25204	7.10025	7.98722	7.83563	8.09589	7.66055	7.46696	8.01953	8.5485	8.89975	8.2442	8.10341
66	1255_g_at	2.99568	3.13716	3.31115	3.39302	3.22608	3.4484	3.53151	3.37321	3.58914	3.60277	3.8355	3.60085	3.84845
67	1294_at	5.86496	5.908	5.5454	7.08601	5.85381	6.22509	6.05949	6.24108	6.61087	7.30959	6.8064	7.06536	6.51999
68	1316_at	4.714	4.76368	4.73688	5.13795	4.77754	4.99426	5.0032	5.1503	5.38395	5.40588	5.67363	5.2636	5.56291
69	1320_at	3.81331	3.62043	3.76574	3.96046	3.9283	4.00209	3.83565	4.01251	4.10997	4.09588	4.35531	3.95444	4.05609
70	1405_i_at	3.74983	3.52672	3.48854	3.90788	3.31866	3.65133	3.53986	3.80711	3.99118	3.91932	4.43341	3.59144	5.00821
71	1431_at	3.09368	3.12216	3.02732	4.02962	3.03975	3.34048	3.03704	3.23148	3.22122	3.529	3.58041	3.29748	3.51201
72	1438_at	5.31156	5.15311	5.3069	6.14691	5.57046	6.28229	5.37652	5.54494	6.31894	6.21323	6.41962	5.85699	6.10687
73	1487_at	7.27482	7.31441	6.92615	7.20505	7.57457	7.05009	7.26466	7.32946	6.93762	6.9453	7.18271	7.34664	6.91876

NCBI

NCBI » GEO » GEO2R » GSE25724

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Rest instructions [YouTube](#)

GEO accession Set Expression data from type 2 diabetic and non-diabetic isolated human islets

选择具体数据

SamplesDefine groups

Enter a group name: List

Group	Accession	Source name	Tissue	Disease state	
-	GSM631755	Non-diabetic islets, rep1	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631756	Non-diabetic islets, rep2	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631757	Non-diabetic islets, rep3	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631758	Non-diabetic islets, rep4	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631759	Non-diabetic islets, rep5	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631760	Non-diabetic islets, rep6	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631761	Non-diabetic islets, rep7	human islets, non-diabetic	pancreatic islets	non-diabetic
-	GSM631762	Type 2 diabetic islets, rep1	human islets, diabetic	pancreatic islets	type 2 diabetes
-	GSM631763	Type 2 diabetic islets, rep2	human islets, diabetic	pancreatic islets	type 2 diabetes
-	GSM631764	Type 2 diabetic islets, rep3	human islets, diabetic	pancreatic islets	type 2 diabetes
-	GSM631765	Type 2 diabetic islets, rep4	human islets, diabetic	pancreatic islets	type 2 diabetes
-	GSM631766	Type 2 diabetic islets, rep5	human islets, diabetic	pancreatic islets	type 2 diabetes
-	GSM631767	Type 2 diabetic islets, rep6	human islets, diabetic	pancreatic islets	type 2 diabetes

设置分组

SamplesDefine groups

Enter a group name: List

Cancel selection

Type2-diabetic (6 samples)

Non-diabetic (7 samples)

Group	Accession	
Non-diabetic	GSM631755	
Non-diabetic	GSM631756	
Non-diabetic	GSM631757	
Non-diabetic	GSM631758	Non-diabetic islets, rep4
Non-diabetic	GSM631759	Non-diabetic islets, rep5
Non-diabetic	GSM631760	Non-diabetic islets, rep6
Non-diabetic	GSM631761	Non-diabetic islets, rep7
Type2-diabetic	GSM631762	Type 2 diabetic islets, rep1
Type2-diabetic	GSM631763	Type 2 diabetic islets, rep2
Type2-diabetic	GSM631764	Type 2 diabetic islets, rep3
Type2-diabetic	GSM631765	Type 2 diabetic islets, rep4
Type2-diabetic	GSM631766	Type 2 diabetic islets, rep5
Type2-diabetic	GSM631767	Type 2 diabetic islets, rep6

R代码

GEO2ROptionsProfile graphR script

Apply adjustment to the P-values. More...

☒ Benjamini & Hochberg (False discovery rate)

☐ Benjamini & Yekutieli

☐ Bonferroni

☐ Holm

Apply log transformation to the data. More...

☒ Auto-detect

☐ Yes

☐ No

Apply limma precision weights (vooma). More...

☐ Yes

☒ No

Force normalization. More...

☐ Yes

☒ No

Category of Platform annotation to display on results.

☐ Submitter supplied

☒ NCBI generated

Plot displays. More...

Significance level cut-off (enter number between 0 and 1)

Log 2 fold change threshold

Volcano and Mean-difference plot contrasts (select up to 5)

0 selected (clear)

☐ Type2-diabetic vs Non-diabetic

If you edit Options after performing an analysis, click Reanalyze to apply the edits.

Reanalyze

参数选择

比较方式

分析

注：一次不成功，多试几次

GEO2R

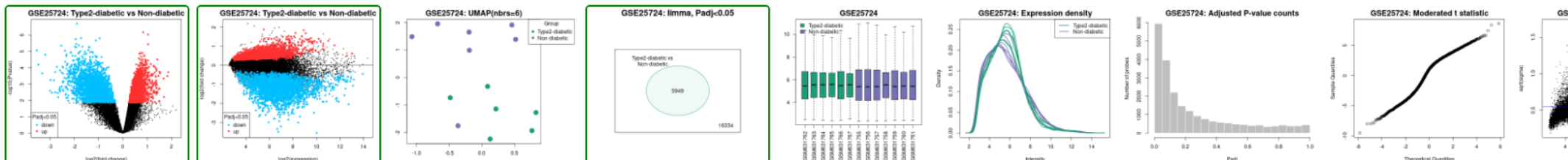
Options

Profile graph

R script

Reanalyze if you changed any options.

Visualization ?



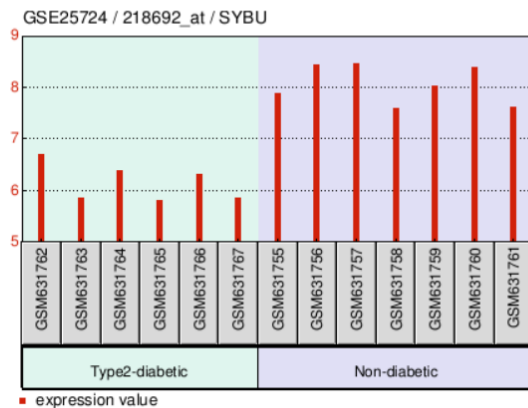
Top differentially expressed genes ?

下载全表 [Download full table](#) [Select columns](#) 选择列 (若无基因名, 可以试试)

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
▸ 218692_at	0.0043	1.93e-07	-9.52	7.19	-1.894	SYBU	syntabulin
▸ 220296_at	0.00641	6.64e-07	8.58	6.14	0.837		
▸ 211465_x_at	0.00641	9.18e-07	8.35	5.86	1.04	FUT6	fucosyltransferase 6

ID	adj.P.Val	P.Value	t
▸ 218692_at	0.0043	1.93e-07	-9.52

Sample values



Log2倍数变化

基因名

基因描述信息

正: up
负: down

功能富集分析

结果合并

61	ID_REF	GSM631755	GSM631756	GSM631757	GSM631758	GSM631759	GSM631760	GSM631761	GSM631762	GSM631763	GSM631764	GSM631765	GSM631766	GSM631767
62	1007_s_at	9.37362	9.27457	9.31466	8.81297	9.22501	8.6195	9.79252	8.71202	8.8878	9.41581	8.86084	10.1835	8.7555
63	1053_at	4.88624	4.57343	4.66537	4.41698	4.53738	4.89242	4.57879	4.3637	4.33044	4.33198	4.17537	4.39334	4.65933
64	117_at	4.96282	4.91762	4.9978	5.68013	7.17241	5.41695	5.21297	5.26965	5.54095	6.35407	5.62223	5.3208	5.87183
65	121_at	7.80585	7.25204	7.10025	7.98722	7.83563	8.09589	7.66055	7.46696	8.01953	8.5485	8.89975	8.2442	8.10341
66	1255_g_at	2.99568	3.13716	3.31115	3.39302	3.22608	3.4484	3.53151	3.37321	3.58914	3.60277	3.8355	3.60085	3.84845
67	1294_at	5.86496	5.908	5.5454	7.08601	5.85381	6.22509	6.05949	6.24108	6.61087	7.30959	6.8064	7.06536	6.51999
68	1316_at	4.714	4.76368	4.73688	5.13795	4.77754	4.99426	5.0032	5.1503	5.38395	5.40588	5.67363	5.2636	5.56291
69	1320_at	3.81331	3.62043	3.76574	3.96046	3.9283	4.00209	3.83565	4.01251	4.10997	4.09588	4.35531	3.95444	4.05609
70	1405_i_at	3.74983	3.52672	3.48854	3.90788	3.31866	3.65133	3.53986	3.80711	3.99118	3.91932	4.43341	3.59144	5.00821
71	1431_at	3.09368	3.12216	3.02732	4.02962	3.03975	3.34048	3.03704	3.23148	3.22122	3.529	3.58041	3.29748	3.51201
72	1438_at	5.31156	5.15311	5.3069	6.14691	5.57046	6.28229	5.37652	5.54494	6.31894	6.21323	6.41962	5.85699	6.10687
73	1487_at	7.27482	7.31441	6.92615	7.20505	7.57457	7.05009	7.26466	7.32946	6.93762	6.9453	7.18271	7.34664	6.91876

GSE25724.top.table.tsv									
1	ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title	
2	218692_at	0.0043	1.93e-07		-9.515988	7.190784	-1.8943593	SYBU	syntabulin
3	220296_at	0.00641	6.64e-07		8.580331	6.139189	0.8369024		
4	211465_x_at	0.00641	9.18e-07		8.345812	5.85724	-1.0402888	FUT6	fucosyltransferase,6
5	217724_at	0.00641	1.50e-06		-8.001034	5.428612	-1.721919	SERBP1	SERPINE1,mRN.
6	201145_at	0.00641	1.54e-06		-7.981476	5.403784	-0.9596948	HAX1	HCLS1,associat
7	207098_s_at	0.00641	1.72e-06		-7.903182	5.30383	-1.9506498	MFN1	mitofusin,1

表达

差异

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	ID_REF	GSM631755	GSM631756	GSM631757	GSM631758	GSM631759	GSM631760	GSM631761	GSM631762	GSM631763	GSM631764	GSM631765	GSM631766	GSM631767	ID	adj.P.Val	P.Value	t	B	logFC	Gene.sym	Gene.title
2	1007_s_at	9.37362	9.27457	9.31466	8.81297	9.22501	8.6195	9.79252	8.71202	8.8878	9.41581	8.86084	10.1835	8.7555	1007_s_at	0.8428	7.95E-01	-0.2644	-6.24393	-0.06592	MIR4640/	microRNA
3	1053_at	4.88624	4.57343	4.66537	4.41698	4.53738	4.89242	4.57879	4.3637	4.33044	4.33198	4.17537	4.39334	4.65933	1053_at	0.07705	2.91E-02	-2.43459	-3.74412	-0.27439	RFC2	replication
4	117_at	4.96282	4.91762	4.9978	5.68013	7.17241	5.41695	5.21297	5.26965	5.54095	6.35407	5.62223	5.3208	5.87183	117_at	0.66524	5.88E-01	0.555272	-6.12221	0.183155	HSPA6	heat shock
5	121_at	7.80585	7.25204	7.10025	7.98722	7.83563	8.09589	7.66055	7.46696	8.01953	8.5485	8.89975	8.2442	8.10341	121_at	0.08172	3.22E-02	2.382024	-3.83512	0.536949	PAX8	paired box
6	1255_g_at	2.99568	3.13716	3.31115	3.39302	3.22608	3.4484	3.53151	3.37321	3.58914	3.60277	3.8355	3.60085	3.84845	1255_g_at	0.04359	1.01E-02	2.977222	-2.77017	0.349796	GUCA1A	guanylate
7	1294_at	5.86496	5.908	5.5454	7.08601	5.85381	6.22509	6.05949	6.24108	6.61087	7.30959	6.8064	7.06536	6.51999	1294_at	0.04666	1.15E-02	2.912473	-2.88889	0.681345	MIR5193/	microRNA
8	1316_at	4.714	4.76368	4.73688	5.13795	4.77754	4.99426	5.0032	5.1503	5.38395	5.40588	5.67363	5.2636	5.56291	1316_at	0.0155	4.36E-04	4.589245	0.200464	0.531353	THRA	thyroid ho
9	1320_at	3.81331	3.62043	3.76574	3.96046	3.9283	4.00209	3.83565	4.01251	4.10997	4.09588	4.35531	3.95444	4.05609	1320_at	0.07133	2.55E-02	2.503398	-3.62389	0.250798	PTPN21	protein tyr
10	1405_i_at	3.74983	3.52672	3.48854	3.90788	3.31866	3.65133	3.53986	3.80711	3.99118	3.91932	4.43341	3.59144	5.00821	1405_i_at	0.06188	1.98E-02	2.633528	-3.39348	0.527566	CCL5	C-C motif

excel先排序，再合并，合并后可以挑选感兴趣基因绘制热图
根据表达值，验证下log2FC，上下调
挑选差异基因做热图等

GEO2R视频

- <https://www.bilibili.com/video/BV12m4y1E7ti/>

limma差异分析

```
05_limma_ok.r
1 # 载入limma包
2 library(limma)
3
4 # 读取表达矩阵
5 df = read.table(file="GSE25724_series_matrix.txt", header=TRUE, row.names=1, sep="\t", check.names=FALSE,quote = "")
6
7 # 数据是否log2, 根据表达值来看: 数据范围50以下的, 一般不用。1000以上的需要log2
8 #df1 <- log2(df+0.1)
9
10 # 绘制box图, 查看样品表达情况
11 pdf("box.before.pdf")
12 boxplot(df)
13 dev.off()
14
15 # 按列取样品 (对照组放前面, 实验组在后)
16 df1 = subset(df, select = c("GSM631755", "GSM631756", "GSM631757", "GSM631758", "GSM631759", "GSM631760", "GSM631761",
17 "GSM631762", "GSM631763", "GSM631764", "GSM631765", "GSM631766", "GSM631767"))
18
19 # 若标准化过, 则不用再标准化 (看页面描述信息)
20 #df1 = normalizeBetweenArrays(df)
21
22 #pdf("box.after.pdf")
23 #boxplot(df1)
24 #dev.off()
25
26 # 标准化数据导出, 用于绘制热图等
27 write.table(df1, file='normdata.txt',quote=FALSE, sep = "\t")
28
29 # 设置分组信息 (对照组放前面, 实验组在后, 样品一一对应)
30 group_list <- factor(c("Non-diabetic", "Non-diabetic", "Non-diabetic", "Non-diabetic", "Non-diabetic", "Non-diabetic", "Non-diabetic",
31 "Type2-diabetic", "Type2-diabetic", "Type2-diabetic", "Type2-diabetic", "Type2-diabetic", "Type2-diabetic"),
32 levels=c("Non-diabetic", "Type2-diabetic"))
33
34 design <- model.matrix(~group_list)
35
36 # 建模计算差异
37 fit <- lmFit(df1, design)
38 fit <- eBayes(fit, 0.01)
39
40 # 差异结果
41 result_limma <- topTable(fit, adjust="fdr", sort.by="B", coef=2, number=Inf)
42
43 # 导出差异表格
44 write.table(result_limma, file="results.txt", sep="\t", quote=F)
45
46 ###合并结果
47 # 使用merge函数根据新列合并两个数据框
48 result_limma$probe <- rownames(result_limma)
49 df1$probe <- rownames(df1)
50
51 df1_sorted <- df1[match(result_limma$probe, df1$probe), ]
52 merged_df <- merge(result_limma, df1_sorted, by = "probe")
53 write.table(merged_df, file="results2.txt", sep="\t", quote=F, row.names=F)
54
```

表达矩阵 (从Series Matrix处下载)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID_REF	GSM631755	GSM631756	GSM631757	GSM631758	GSM631759	GSM631760	GSM631761	GSM631762	GSM631763	GSM631764	GSM631765	GSM631766	GSM631767
2	1007_s_at	9.37362	9.27457	9.31466	8.81297	9.22501	8.6195	9.79252	8.71202	8.8878	9.41581	8.86084	10.1835	8.7555
3	1053_at	4.88624	4.57343	4.66537	4.41698	4.53738	4.89242	4.57879	4.3637	4.33044	4.33198	4.17537	4.39334	4.65933
4	117_at	4.96282	4.91762	4.9978	5.68013	7.17241	5.41695	5.21297	5.26965	5.54095	6.35407	5.62223	5.3208	5.87183
5	121_at	7.80585	7.25204	7.10025	7.98722	7.83563	8.09589	7.66055	7.46696	8.01953	8.5485	8.89975	8.2442	8.10341
6	1255_g_at	2.99568	3.13716	3.31115	3.39302	3.22608	3.4484	3.53151	3.37321	3.58914	3.60277	3.8355	3.60085	3.84845
7	1294_at	5.86496	5.908	5.5454	7.08601	5.85381	6.22509	6.05949	6.24108	6.61087	7.30959	6.8064	7.06536	6.51999
8	1316_at	4.714	4.76368	4.73688	5.13795	4.77754	4.99426	5.0032	5.1503	5.38395	5.40588	5.67363	5.2636	5.56291
9	1320_at	3.81331	3.62043	3.76574	3.96046	3.9283	4.00209	3.83565	4.01251	4.10997	4.09588	4.35531	3.95444	4.05609

提取待分析样品 (对照在前, 实验在后)

设置样品对应的组信息 (对照在前, 实验在后)

	A	B	C	D	E	F	G
1	probe	logFC	AveExpr	t	P.Value	adj. P.Val	B
2	218692_at	-1.89436	7.197765	-9.51599	1.93E-07	0.0043	7.190784
3	220296_at	0.836902	5.030437	8.580331	6.64E-07	0.006405	6.139189
4	211465_x	1.040289	7.876416	8.345812	9.18E-07	0.006405	5.85724

probeid

差异

Annotations