筛选研究对象 AA 菌-BB 代谢产物-XX 基因

2024-06-24

LiChuang Huang

@ 立效研究院

Contents

List of Figures

List of Tables

1 摘要

1.1 要求

肠道菌-代谢物-基因关联数据为此前分析数据。选择差异最大的基因前 5，寻找关联代谢物和菌。

1.2 结果

主要思路，结合此前分析得到的数据，再以 RNA-seq (胆结石) 差异分析，根据显著性排序基因。

- 肝脏见 Tab. 5。
- 回肠见 Tab. 6。

2 前言

3 材料和方法

3.1 材料

All used GEO expression data and their design:

- **GSE66430**: RNA-seq of four female human gallbladders (3 healthy controls and 1 case with chronic gallstones) and one liver sample from the gallstone case.

3.2 方法

Mainly used method:

- R package `biomaRt` used for gene annotation[1].
- GEO https://www.ncbi.nlm.nih.gov/geo/ used for expression dataset aquisition.
- R package `Limma` and `edgeR` used for differential expression analysis[2,3].
- R version 4.4.0 (2024-04-24); Other R packages (eg., `dplyr` and `ggplot2`) used for statistic analysis or data visualization.

4 分析结果

5 结论

6 附：分析流程

6.1 GEO 数据获取 (GALLSTONE)

我们首先从 GEO 数据库中获取了与胆结石相关的数据。通过查询并筛选相关的实验，我们下载了数据集并进行了预处理。预处理步骤包括数据标准化和质量控制，以确保后续分析的准确性。

Data Source ID :

GSE66430

data_processing :

Sequencing data were demultiplexed and converted to FASTQ format.

data_processing.1 :

Paired-end reads were aligned to RefSeq (hg19) using TopHat (v2.0.9) with the parameter setting: -g 1 -N 2 -r 200.

data_processing.2 :

RNA reads-per-kilobase-per million mapped (RPKM) was calculated with RSeQC v2.3.6.

data_processing.3 :

Genome_build: GRCh37 (hg19)

(Others) :

...

(上述信息框内容已保存至 `Figure+Table/GALLSTONE-GSE66430-content`)

Table 1 (下方表格) 为表格 GALLSTONE GSE66430 metadata 概览。

(对应文件为 `Figure+Table/GALLSTONE-GSE66430-metadata.csv`)

注：表格共有 5 行 6 列，以下预览的表格可能省略部分数据；含有 5 个唯一 'rownames'。

Table 1: GALLSTONE GSE66430 metadata

rownames	title	age..years...	disease.st...	gender.ch1	tissue.ch1
GSM1622382	Non-diseas...	34	healthy	female	gall bladder
GSM1622383	Non-diseas...	46	healthy	female	gall bladder
GSM1622384	Non-diseas...	64	healthy	female	gall bladder
GSM1622385	Diseased G...	71	chronic ga...	female	gall bladder
GSM1622386	Diseased L...	71	chronic ga...	female	liver

6.2 Biomart 基因注释 (REFSEQ)

接下来，我们使用 Biomart 工具对基因进行了注释。通过链接 REFSEQ 数据库，我们将基因表达数据与基因注释信息进行匹配。

6.3 Limma 差异分析 (GALLSTONE)

使用 Limma 软件包，我们进行了胆结石相关样本的差异表达分析。通过对比正常和疾病状态下的基因表达数据，我们识别出显著差异表达的基因。

Figure 1 (下方图) 为图 GALLSTONE Disease vs Control 概览。

(对应文件为 `Figure+Table/GALLSTONE-Disease-vs-Control.pdf`)

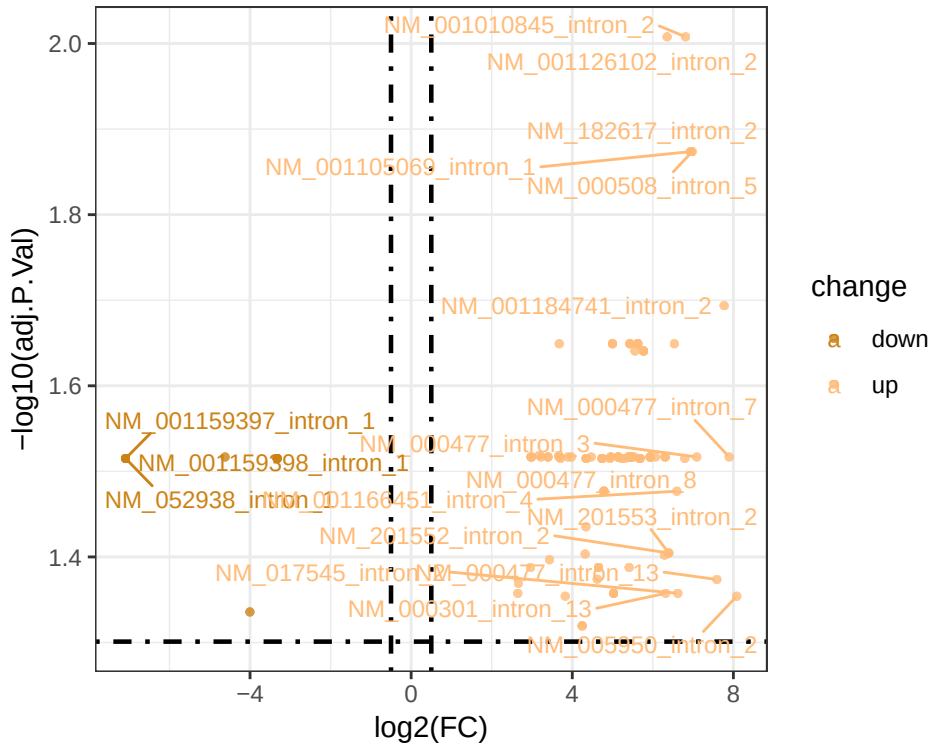

Figure 1: GALLSTONE Disease vs Control

adj.P.Val cut-off :

0.05

Log2(FC) cut-off :

0.5

(上述信息框内容已保存至 `Figure+Table/GALLSTONE-Disease-vs-Control-content`)

Table 2 (下方表格) 为表格 GALLSTONE data Disease vs Control 概览。

(对应文件为 `Figure+Table/GALLSTONE-data-Disease-vs-Control.csv`)

注：表格共有 101 行 11 列，以下预览的表格可能省略部分数据；含有 101 个唯一 'rownames'；含有 37 个唯一 'hgnc_symbol'。

1. hgnc_symbol: 基因名 (Human)
2. logFC: estimate of the log2-fold-change corresponding to the effect or contrast (for 'topTableF' there may be several columns of log-fold-changes)
3. AveExpr: average log2-expression for the probe over all arrays and channels, same as 'Amean' in the 'MarrayLM' object
4. t: moderated t-statistic (omitted for 'topTableF')
5. P.Value: raw p-value
6. B: log-odds that the gene is differentially expressed (omitted for 'topTreat')

Table 2: GALLSTONE data Disease vs Control

rownames	gene	accession	hgnc_s...	isIntron	logFC	AveExpr	t	P.Value	adj.P.Val
294507	294507	NM_001...	HP	TRUE	6.3561...	0.1962...	12.777...	5.9012...	0.0098...
298099	298099	NM_001...	NA	TRUE	6.8136...	-0.777...	12.589...	6.8856...	0.0098...
112356	112356	NM_002...	HMGCS1	TRUE	3.6763...	3.5944...	9.6243...	1.0579...	0.0224...
292286	292286	NM_001...	ACSM2B	TRUE	6.9353...	-1.200...	11.603...	1.5969...	0.0133...
292299	292299	NM_182...	ACSM2B	TRUE	6.9494...	-1.287...	11.283...	2.1265...	0.0133...
101714	101714	NM_000...	FGA	TRUE	6.9766...	-1.416...	11.176...	2.3443...	0.0133...
150259	150259	NM_017...	CYP3A4	TRUE	4.9985...	-0.306...	9.9083...	7.9150...	0.0224...
150294	150294	NM_001...	CYP3A4	TRUE	4.9985...	-0.306...	9.9083...	7.9150...	0.0224...
124063	124063	NR_033...	NA	TRUE	3.1808...	3.9078...	8.9243...	2.2292...	0.0302...
112365	112365	NM_001...	HMGCS1	TRUE	3.6700...	3.3268...	8.9506...	2.1659...	0.0302...
150260	150260	NM_017...	CYP3A4	TRUE	5.4308...	-0.822...	9.5937...	1.0919...	0.0224...
150295	150295	NM_001...	CYP3A4	TRUE	5.4308...	-0.822...	9.5937...	1.0919...	0.0224...
318508	318508	NM_002...	KRT19	TRUE	3.8891...	2.2275...	8.5802...	3.2703...	0.0304...
233713	233713	NM_005...	FGF19	TRUE	3.6670...	1.7986...	8.5826...	3.2617...	0.0304...
85809	85809	NM_001...	TF	TRUE	5.5575...	-0.804...	9.2766...	1.5233...	0.0228...
...	...	...	...	...	...	...	...	...	...

6.4 肠道菌-代谢物-基因关联数据

6.4.1 前一次的分析数据

在这部分中，我们引用了之前的分析数据。

Table 3 (下方表格) 为表格 liver data 概览。

(对应文件为 `Figure+Table/liver-data.xlsx`)

> 注：表格共有 10454 行 10 列，以下预览的表格可能省略部分数据；含有 22 个唯一 '.id'。

1. META_Rho: 关联分析结果的关联系数，绝对值越大，说明关联性越强 (源自文献的分析)
2. META_Q: 关联分析结果 P 的校正值 (源自文献的分析)
3. META_P: 关联分析结果 P 的值 (源自文献的分析)

Table 3: Liver data

.id	.id_from	Substrate	Metabo……4	Gut.Mi…	Target…	Metabo……7	META_Rho	META_Q	META_P	…
588	Metabo…		2-Imin…	Clostr…	CD59	creati…	0.4459…	6.9021…	1.8504…	…
750	Substrate	Glycine	Acetyl…	Clostr…	GHR	glycine	-0.425…	4.5068…	2.4165…	…
750	Metabo…		Glycine	Blautia	GHR	glycine	-0.425…	4.5068…	2.4165…	…
750	Metabo…		Glycine	Lactob…	GHR	glycine	-0.425…	4.5068…	2.4165…	…
5793	Substrate	D-Glucose	Acetate	Christ…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	Butyrate	Christ…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	2,3-Bu…	Escher…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	Acetoin	Escher…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	2,3-Bu…	Escher…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	2,3-Bu…	Escher…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	Acetoin	Escher…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	2,3-Bu…	Escher…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	Ethanol	Lactob…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	Acetate	Clostr…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…
5793	Substrate	D-Glucose	Butyrate	Clostr…	PLXNB2	glucose	0.3849…	1.3255…	3.5538…	…

.id	.id_from	Substrate	Metabo……4	Gut.Mi…	Target…	Metabo……7	META_Rho	META_Q	META_P	…
…	…	…	…	…	…	…	…	…	…	…

Table 4 (下方表格) 为表格 ileum data 概览。

(对应文件为 `Figure+Table/ileum-data.xlsx`)

注：表格共有 9208 行 10 列，以下预览的表格可能省略部分数据；含有 22 个唯一 '.id'。

1. META_Rho: 关联分析结果的关联系数，绝对值越大，说明关联性越强 (源自文献的分析)
2. META_Q: 关联分析结果 P 的校正值 (源自文献的分析)
3. META_P: 关联分析结果 P 的值 (源自文献的分析)

Table 4: Ileum data

.id	.id_from	Substrate	Metabo……4	Gut.Mi…	Target…	Metabo……7	META_Rho	META_Q	META_P	…
588	Metabo…		2-Imin…	Clostr…	B2M	creati…	0.5130…	0	0	…
588	Metabo…		2-Imin…	Clostr…	DSC2	creati…	0.5128…	0	0	…
588	Metabo…		2-Imin…	Clostr…	RGMB	creati…	0.4166…	5.3138…	1.4246…	…
750	Substrate	Glycine	Acetyl…	Clostr…	RET	glycine	-0.407…	9.5712…	5.1320…	…
750	Metabo…		Glycine	Blautia	RET	glycine	-0.407…	9.5712…	5.1320…	…
750	Metabo…		Glycine	Lactob…	RET	glycine	-0.407…	9.5712…	5.1320…	…
588	Metabo…		2-Imin…	Clostr…	JAM2	creati…	0.4070…	2.8618…	7.6726…	…
588	Metabo…		2-Imin…	Clostr…	CST6	creati…	0.3307…	2.1455…	5.7522…	…
588	Metabo…		2-Imin…	Clostr…	SPOCK2	creati…	-0.321…	1.3977…	1.1241…	…
588	Metabo…		2-Imin…	Clostr…	LCN2	creati…	0.3110…	2.0900…	1.1206…	…
588	Metabo…		2-Imin…	Clostr…	TNFRSF21	creati…	0.3098…	7.2094…	7.7313…	…
588	Metabo…		2-Imin…	Clostr…	SMOC1	creati…	0.3071…	2.8839…	7.7317…	…
588	Metabo…		2-Imin…	Clostr…	TNFRSF10	creati…	0.2890…	4.7330…	1.2689…	…
588	Metabo…		2-Imin…	Clostr…	COL18A1	creati…	0.2883…	3.0855…	3.3089…	…
750	Substrate	Glycine	Acetyl…	Clostr…	SLITRK5	glycine	0.2801…	1.7545…	4.7038…	…
…	…	…	…	…	…	…	…	…	…	…

6.4.2 结合 GALLSTONE RNA-seq 差异分析筛选

我们将胆结石 RNA-seq 差异分析的结果与肠道菌-代谢物-基因关联数据相结合。

Table 5 (下方表格) 为表格 Res liver 概览。

(对应文件为 `Figure+Table/Res-liver.csv`)

注: 表格共有 25 行 6 列, 以下预览的表格可能省略部分数据; 含有 5 个唯一 'hgnc_symbol'。

1. hgnc_symbol: 基因名 (Human)
2. logFC: estimate of the log2-fold-change corresponding to the effect or contrast (for 'topTableF' there may be several columns of log-fold-changes)
3. META_Q: 关联分析结果 P 的校正值 (源自文献的分析)

Table 5: Res liver

hgnc_symbol	logFC	adj.P.Val	related_me...	related_mi...	META_Q
ALB	7.09094053...	0.03042149...	Acetyl pho...	Clostridium	2.41950093...
ALB	7.09094053...	0.03042149...	Glycine	Blautia	2.41950093...
ALB	7.09094053...	0.03042149...	Glycine	Lactobacil...	2.41950093...
ALB	7.09094053...	0.03042149...	Serine	Blautia	3.22595060...
ALB	7.09094053...	0.03042149...	3-Indolepr...	Lachnospir...	7.78730300...
CYP3A4	4.99857331...	0.02242949...	2-Imino-1-...	Clostridium	1.75386285...
CYP3A4	4.99857331...	0.02242949...	Leucine	Blautia	4.16629071...
CYP3A4	4.99857331...	0.02242949...	Creatine	Akkermansia	1.00303219...
CYP3A4	4.99857331...	0.02242949...	Creatine	Lactobacillus	1.00303219...
CYP3A4	4.99857331...	0.02242949...	Creatine	Lactobacil...	1.00303219...
HP	6.35613254...	0.00982207...	Indoxyl su...	Lachnospir...	0.00525339...
HP	6.35613254...	0.00982207...	Indoxyl su...	Escherichia	0.00525339...
HP	6.35613254...	0.00982207...	Indoxyl su...	Oscillibacter	0.00525339...
HP	6.35613254...	0.00982207...	10-Keto-12...	Lactobacil...	0.01756457...
HP	6.35613254...	0.00982207...	10-Oxo-11-...	Lactobacil...	0.01756457...
...	...	...	...	...	...

Table 6 (下方表格) 为表格 Res ileum 概览。

(对应文件为 `Figure+Table/Res-ileum.csv`)

注: 表格共有 10 行 6 列, 以下预览的表格可能省略部分数据; 含有 2 个唯一 'hgnc_symbol'。

1. hgnc_symbol: 基因名 (Human)
2. logFC: estimate of the log2-fold-change corresponding to the effect or contrast (for 'topTableF' there may be several columns of log-fold-changes)
3. META_Q: 关联分析结果 P 的校正值 (源自文献的分析)

Table 6: Res ileum

hgnc_symbol	logFC	adj.P.Val	related_me…	related_mi…	META_Q
FGF19	3.66703440…	0.03042149…	Glycocholi…	Escherichia	7.27153355…
FGF19	3.66703440…	0.03042149…	Glycocholi…	Akkermansia	7.27153355…
FGF19	3.66703440…	0.03042149…	3-Phenylpr…	Clostridiu…	1.32242917…
FGF19	3.66703440…	0.03042149…	Deoxycholi…	Clostridiu…	8.88195779…
FGF19	3.66703440…	0.03042149…	Deoxycholi…	Clostridiu…	8.88195779…
TF	5.55755215…	0.02287326…	Serine	Blautia	1.40340968…
TF	5.55755215…	0.02287326…	Indole-3-l…	Clostridiu…	1.49009106…
TF	5.55755215…	0.02287326…	2-Imino-1-…	Clostridium	6.10598737…
TF	5.55755215…	0.02287326…	Leucine	Blautia	1.08359110…
TF	5.55755215…	0.02287326…	Acetyl pho…	Clostridium	1.72172512…

Reference

1. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. *Nature protocols* **4**, 1184–1191 (2009).

2. Ritchie, M. E. *et al.* Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).

3. Chen, Y., McCarthy, D., Ritchie, M., Robinson, M. & Smyth, G. EdgeR: Differential analysis of sequence read count data user's guide. 119.