从NCBI的dataset搜索数据集下载分析

以GSE283056为例

目录

基本流程

查看数据

清洗后操作

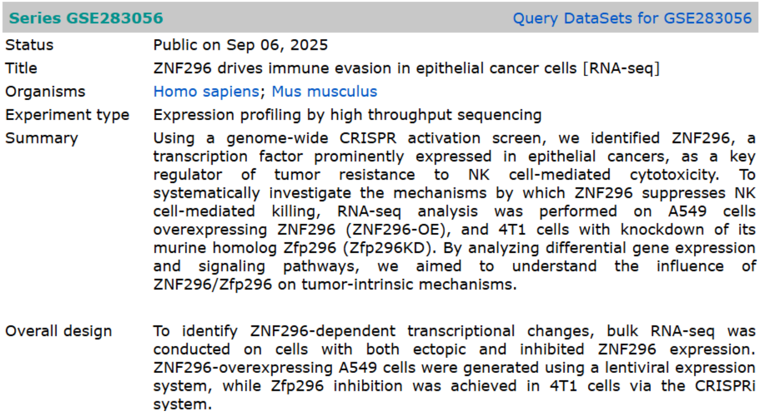
不足

基本流程

基本流程： → 选定数据集或者自行测序，得到数据集 → 解读数据集 → 清洗数据表 → 代码分析 → 绘制统计图

NCBI搜索→breast cancer→查看数据集283056

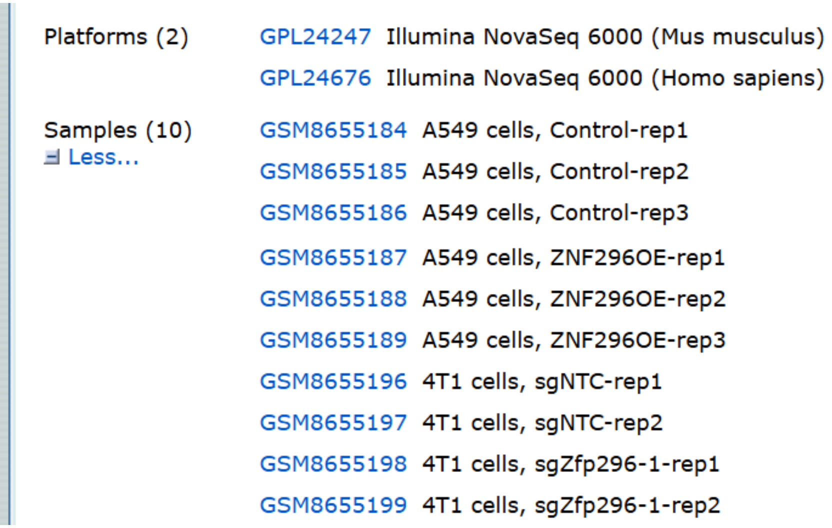
283056的界面截图



Series GSE283056	Query DataSets for GSE283056
Status	Public on Sep 06, 2025
Title	ZNF296 drives immune evasion in epithelial cancer cells [RNA-seq]
Organisms	Homo sapiens; Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	Using a genome-wide CRISPR activation screen, we identified ZNF296, a transcription factor prominently expressed in epithelial cancers, as a key regulator of tumor resistance to NK cell-mediated cytotoxicity. To systematically investigate the mechanisms by which ZNF296 suppresses NK cell-mediated killing, RNA-seq analysis was performed on A549 cells overexpressing ZNF296 (ZNF296-OE), and 4T1 cells with knockdown of its murine homolog Zfp296 (Zfp296KD). By analyzing differential gene expression and signaling pathways, we aimed to understand the influence of ZNF296/Zfp296 on tumor-intrinsic mechanisms.
Overall design	To identify ZNF296-dependent transcriptional changes, bulk RNA-seq was conducted on cells with both ectopic and inhibited ZNF296 expression. ZNF296-overexpressing A549 cells were generated using a lentiviral expression system, while Zfp296 inhibition was achieved in 4T1 cells via the CRISPRi system.

→

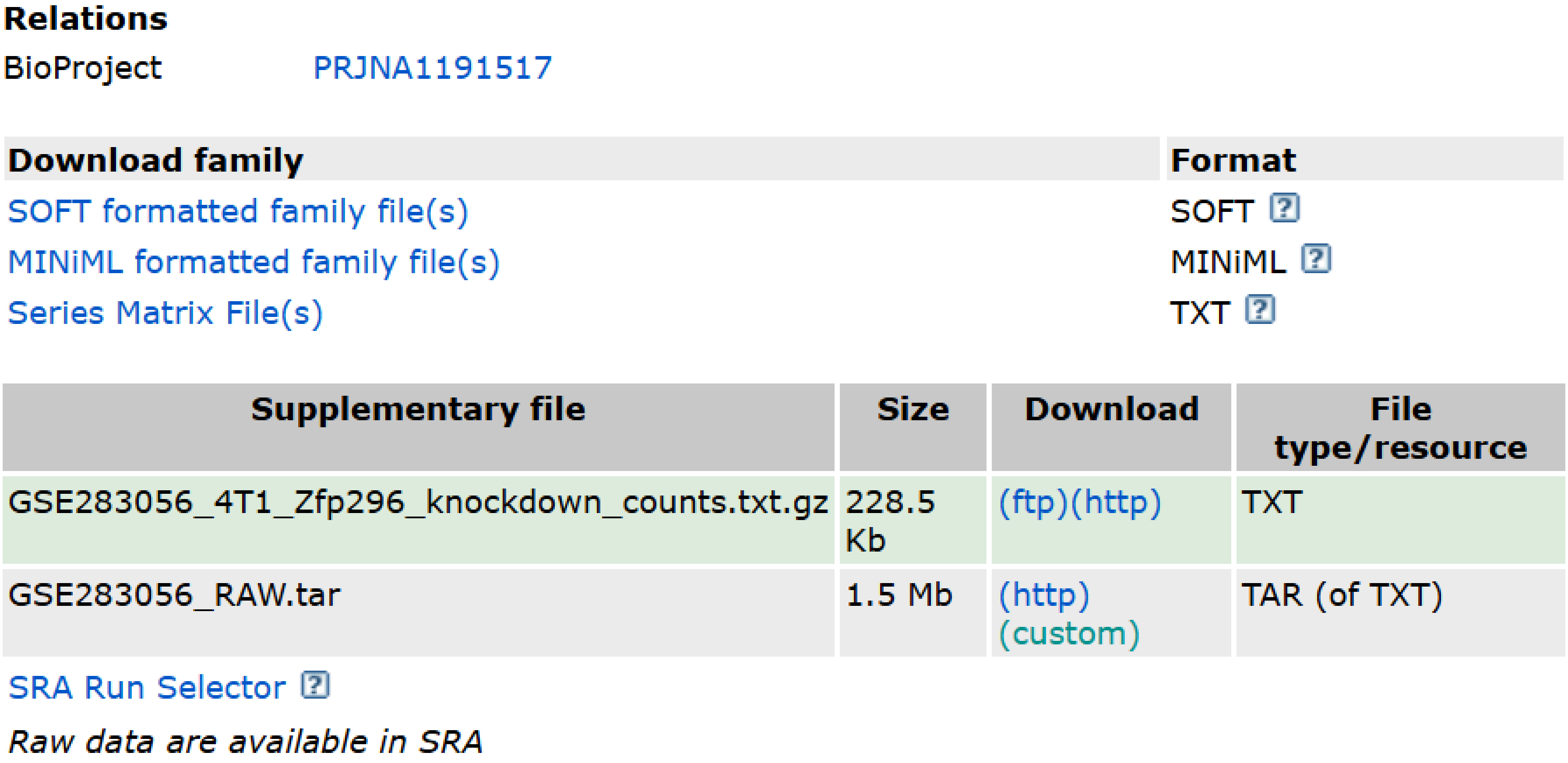
可以看出测序平台和样本分组情况



Platforms (2)	GPL24247	Illumina NovaSeq 6000 (Mus musculus)
	GPL24676	Illumina NovaSeq 6000 (Homo sapiens)
Samples (10) ⊟ Less...	GSM8655184	A549 cells, Control-rep1
	GSM8655185	A549 cells, Control-rep2
	GSM8655186	A549 cells, Control-rep3
	GSM8655187	A549 cells, ZNF296OE-rep1
	GSM8655188	A549 cells, ZNF296OE-rep2
	GSM8655189	A549 cells, ZNF296OE-rep3
	GSM8655196	4T1 cells, sgNTC-rep1
	GSM8655197	4T1 cells, sgNTC-rep2
	GSM8655198	4T1 cells, sgZfp296-1-rep1
	GSM8655199	4T1 cells, sgZfp296-1-rep2

查看数据

查看`soft`,`matrix`,`soft`,`supplementary file`



Relations

BioProject	PRJNA1191517

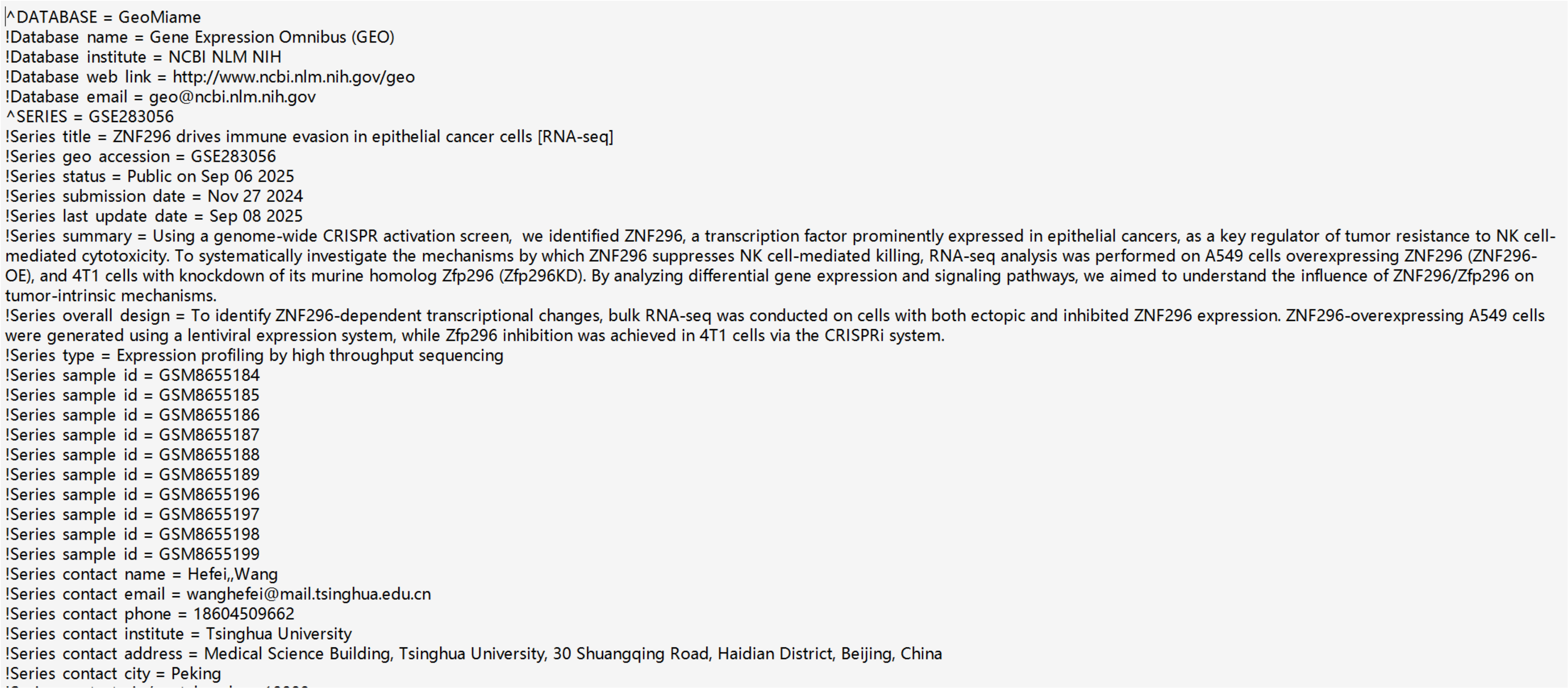
Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE283056_4T1_Zfp296_knockdown_counts.txt.gz	228.5 Kb	(ftp)(http)	TXT
GSE283056_RAW.tar	1.5 Mb	(http) (custom)	TAR (of TXT)

SRA Run Selector ?

Raw data are available in SRA

soft文件详情



```
^DATABASE = GeoMiame
!Database name = Gene Expression Omnibus (GEO)
!Database institute = NCBI NLM NIH
!Database web link = http://www.ncbi.nlm.nih.gov/geo
!Database email = geo@ncbi.nlm.nih.gov
^SERIES = GSE283056
!Series title = ZNF296 drives immune evasion in epithelial cancer cells [RNA-seq]
!Series geo accession = GSE283056
!Series status = Public on Sep 06 2025
!Series submission date = Nov 27 2024
!Series last update date = Sep 08 2025
!Series summary = Using a genome-wide CRISPR activation screen,  we identified ZNF296, a transcription factor prominently expressed in epithelial cancers, as a key regulator of tumor resistance to NK cell-mediated cytotoxicity. To systematically investigate the mechanisms by which ZNF296 suppresses NK cell-mediated killing, RNA-seq analysis was performed on A549 cells overexpressing ZNF296 (ZNF296-OE), and 4T1 cells with knockdown of its murine homolog Zfp296 (Zfp296KD). By analyzing differential gene expression and signaling pathways, we aimed to understand the influence of ZNF296/Zfp296 on tumor-intrinsic mechanisms.
!Series overall design = To identify ZNF296-dependent transcriptional changes, bulk RNA-seq was conducted on cells with both ectopic and inhibited ZNF296 expression. ZNF296-overexpressing A549 cells were generated using a lentiviral expression system, while Zfp296 inhibition was achieved in 4T1 cells via the CRISPRi system.
!Series type = Expression profiling by high throughput sequencing
!Series sample id = GSM8655184
!Series sample id = GSM8655185
!Series sample id = GSM8655186
!Series sample id = GSM8655187
!Series sample id = GSM8655188
!Series sample id = GSM8655189
!Series sample id = GSM8655196
!Series sample id = GSM8655197
!Series sample id = GSM8655198
!Series sample id = GSM8655199
!Series contact name = Hefei,,Wang
!Series contact email = wanghefei@mail.tsinghua.edu.cn
!Series contact phone = 18604509662
!Series contact institute = Tsinghua University
!Series contact address = Medical Science Building, Tsinghua University, 30 Shuangqing Road, Haidian District, Beijing, China
!Series contact city = Peking
```

soft文件解读的关键

^SAMPLE = GSM8655198下列对应的

!Sample_taxid_ch1 = 10090

!Sample_characteristics_ch1 = genotype: Zfp296 knockdown的描述

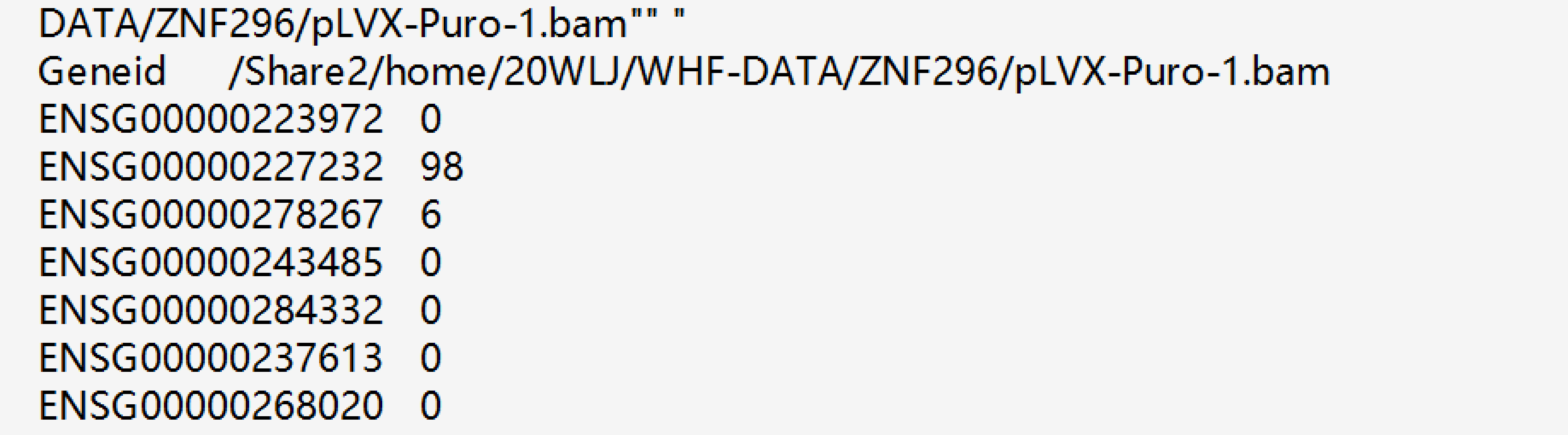
这些可以在以后分析时确定分组做后续检验

matrix文件解读的关键

!Sample_title "A549 cells, Control-rep1" "A549 cells, Control-rep2" "A549 cells, Control-rep3" "A549 cells, ZNF296OE-rep1" "A549 cells, ZNF296OE-rep2" "A549 cells, ZNF296OE-rep3" `这些描述可以确定分组，可以对比soft文件，互相检验分组是否设置正确`

"ID_REF" "GSM8655184" "GSM8655185" "GSM8655186" "GSM8655187" "GSM8655188" "GSM8655189" `这些描述是和sample_title分组对应的样本名称`

Supplementary file的解读



```
DATA/ZNF296/pLVX-Puro-1.bam"" "
Geneid     /Share2/home/20WLJ/WHF-DATA/ZNF296/pLVX-Puro-1.bam
ENSG00000223972   0
ENSG00000227232   98
ENSG00000278267   6
ENSG00000243485   0
ENSG00000284332   0
ENSG00000237613   0
ENSG00000268020   0
```

bam文件是专门的测序格式，可以用Linux或者R的bioconductor包解开，python需要从源文件入手，逐行阅读，用pandas打开为csv或者tsv格式，再来清洗表头

清洗后操作

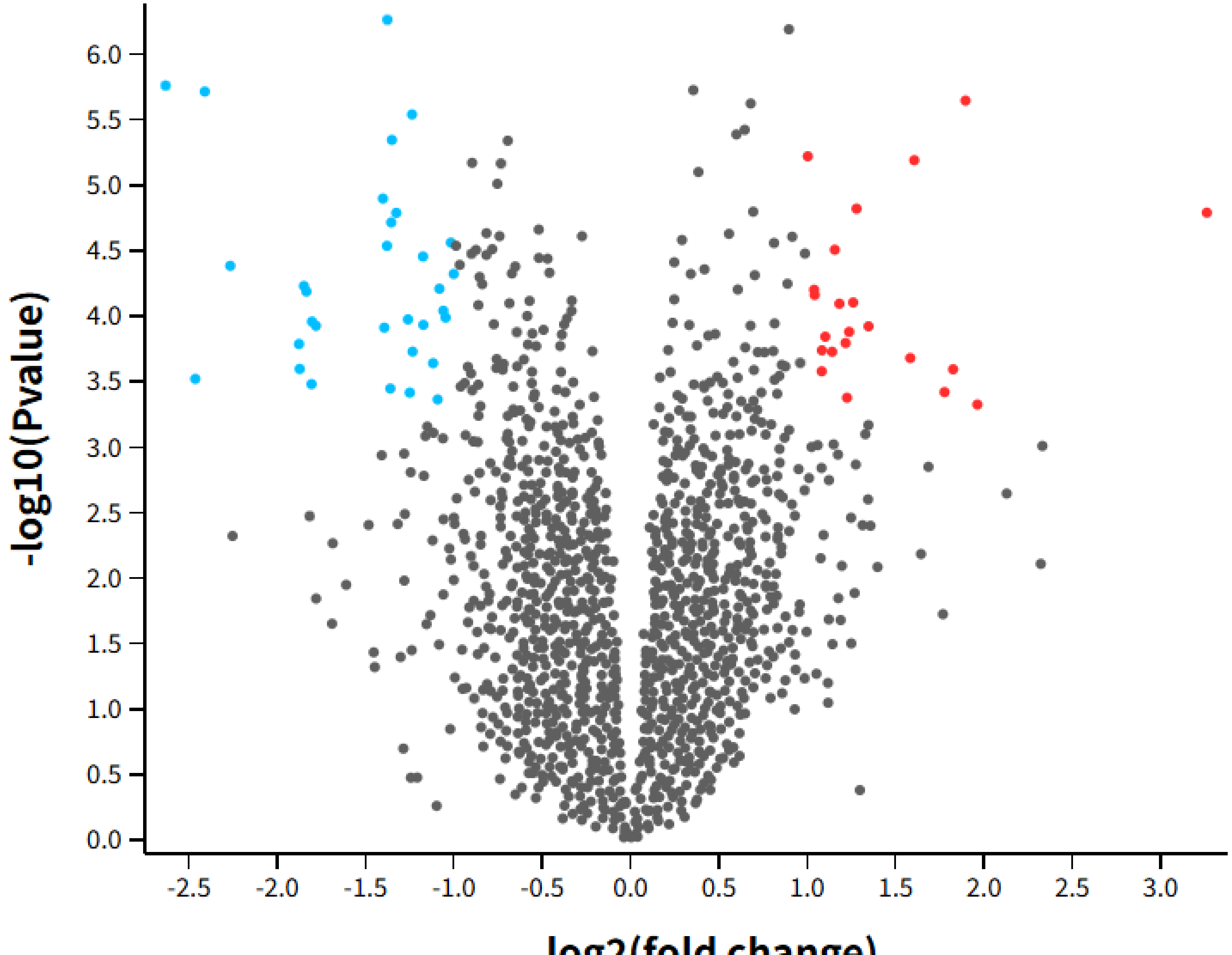
清洗后的数据格式

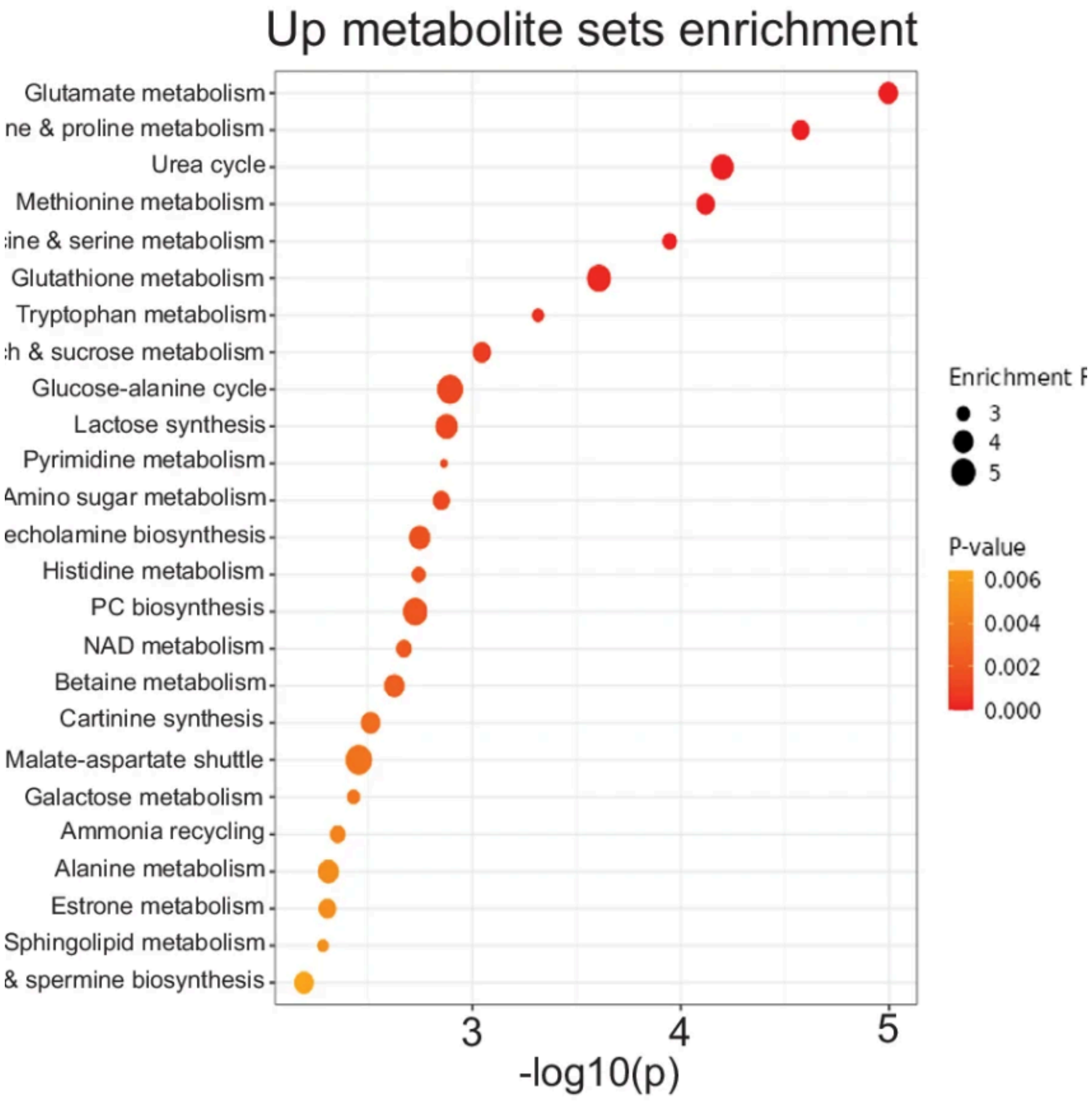
清洗好的数据表格应该是包含了基因id，表达值，样本名，分组的表格，类似于

	sample1/group1	simple2/group2
基因id	expr1	expr2

然后利用统计软件或者专门的分析软件来对其进行分析，画出分析图像

例如使用ggplot2绘制出的基因表达火山图





利用火山图分析出来的差异基因绘制的富集分析图

不足

现阶段问题

代码自动化处理问题

使用pandas模块，清洗数据表，例如去除无关的元数据表述，塑造成一个标准的数据框格式，方便后续统计

统计学问题

统计学检验需要进一步学习，例如， `control` 组和 `disease` 组各有3个样本，是 `control1` 对比 `disease1` 来检验p值，还是 `control1` 分别对比 `disease1` ， `disease2` ， `disease3` 来检验3个p值

更多的分析问题

仅靠差异基因分析和富集分析不足以支撑一个完整的研究，后续需要学习更多的分析来丰富