筛选主动脉-下腔静脉瘘 ACF 模型 DEGs 并功能分析

2024-02-02

LiChuang Huang

@ 立效研究院

Contents

List of Figures

List of Tables

1 摘要

需求：

生物信息学分析筛选对照组动物和 ACF 动物之间有差异表达的 XXX mRNA（若缺少动物数据库，可以筛选血液透析患者的血管差异基因）。GO 和 KEGG 分析与内皮-间质转化相关的显著富集的通路 YYY。

结果：

- 筛选的差异表达基因见 Fig. 4
- 富集结果见 Fig. 5 和 Fig. 6
  - GO:0048771 'tissue remodeling' 为显著富集并与 ET 相关的通路。

2 前言

3 材料和方法

3.1 材料

All used GEO expression data and their design:

- **GSE232594**: Comparative gene expression profiling analysis of RNA-seq data for right atrium free wall myocardium in volume overload and sham-operated C57/BL6 mice on postnatal day21.

3.2 方法

Mainly used method:

- R package `biomaRt` used for gene annotation[1].
- The `biomart` was used for mapping genes between organism (e.g., mgi_symbol to hgnc_symbol)[1].
- R package `ClusterProfiler` used for gene enrichment analysis[2].
- `Fastp` used for Fastq data preprocessing[3].
- GEO https://www.ncbi.nlm.nih.gov/geo/ used for expression dataset aquisition.
- The Human Gene Database `GeneCards` used for disease related genes prediction[4].
- `Kallisto` used for RNA-seq mapping and quantification[5].
- R package `Limma` and `edgeR` used for differential expression analysis[6,7].
- 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 数据来源GSE232594

由于该数据集(以及相似的其它数据集)的原作者没有导出Count数据(适应于差异分析)，因此这里下载了SRA (PRJNA972912)原始数据从头开始分析该RNA-seq数据集。

Table 1 (下方表格)为表格GSE metadata概览。

(对应文件为 `Figure+Table/GSE-metadata.xlsx`)

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

Table 1: GSE metadata

rownames	title	genotype.ch1	strain.ch1	tissue.ch1	treatment.ch1
GSM7359743	RA, sham-o...	WT	C57BL/6	Right atrium	sham-operated
GSM7359744	RA, sham-o...	WT	C57BL/6	Right atrium	sham-operated
GSM7359745	RA, sham-o...	WT	C57BL/6	Right atrium	sham-operated
GSM7359746	RA, Volume...	WT	C57BL/6	Right atrium	volume ove...
GSM7359747	RA, Volume...	WT	C57BL/6	Right atrium	volume ove...
GSM7359748	RA, Volume...	WT	C57BL/6	Right atrium	volume ove...

6.1.1 SRA

Table 2 (下方表格)为表格SRA metadata概览。

(对应文件为 `Figure+Table/SRA-metadata.xlsx`)

> 注：表格共有6行45列，以下预览的表格可能省略部分数据；表格含有6个唯一‘Run’。

Table 2: SRA metadata

Run	spots	bases	spots_...	avgLength	size_MB	Assemb...	downlo...	Experi...	Librar...
SRR245...	23554439	706633...	23554439	300	2164	NA	https:...	SRX203...	GSM735...
SRR245...	23066894	692006...	23066894	300	2125	NA	https:...	SRX203...	GSM735...
SRR245...	22691185	680735...	22691185	300	2136	NA	https:...	SRX203...	GSM735...
SRR245...	23061459	691843...	23061459	300	2141	NA	https:...	SRX203...	GSM735...
SRR245...	21413791	642413...	21413791	300	2006	NA	https:...	SRX203...	GSM735...

Run	spots	bases	spots_...	avgLength	size_MB	Assemb...	downlo...	Experi...	Librar...
SRR245...	21966609	658998...	21966609	300	2050	NA	https:...	SRX203...	GSM735...

6.2 RNA-seq 前处理

6.2.1 QC

‘QC report’ 数据已全部提供。

(对应文件为 `./fastp_report/`)

> 注：文件夹./fastp_report/共包含 6 个文件。
> 1. SRR24578639.html
> 2. SRR24578640.html
> 3. SRR24578641.html
> 4. SRR24578642.html
> 5. SRR24578643.html
> 6. ...

6.2.2 定量

cDNA 参考基因注释 (使用的是 mus musculus 的参考基因) 。https://ftp.ensembl.org/pub/release-110/fasta/mus_musculus/

Table 3 (下方表格) 为表格 Quantification 概览。

(对应文件为 `Figure+Table/Quantification.csv`)

> 注：表格共有 116873 行 7 列，以下预览的表格可能省略部分数据；表格含有 116873 个唯一 ‘target_id’。

Table 3: Quantification

target_id	SRR245786391	SRR245786401	SRR245786411	SRR245786421	SRR245786431	SRR245786441...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...
ENSMUST0000..	0	0	0	0	0	...

target_id	SRR245786391	SRR245786401	SRR245786411	SRR245786421	SRR245786431	SRR245786441	...
ENSMUST000...	0	0	0	0	0	0	...
ENSMUST000...	0	0	0	0	0	0	...
ENSMUST000...	0	0	0	0	0	0	...
ENSMUST000...	0	0	0	0	0	0	...
ENSMUST000...	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...

6.3 差异分析

6.3.1 QC

Figure 1 (下方图) 为图 Filtered 概览。

(对应文件为 `Figure+Table/Filtered.pdf`)

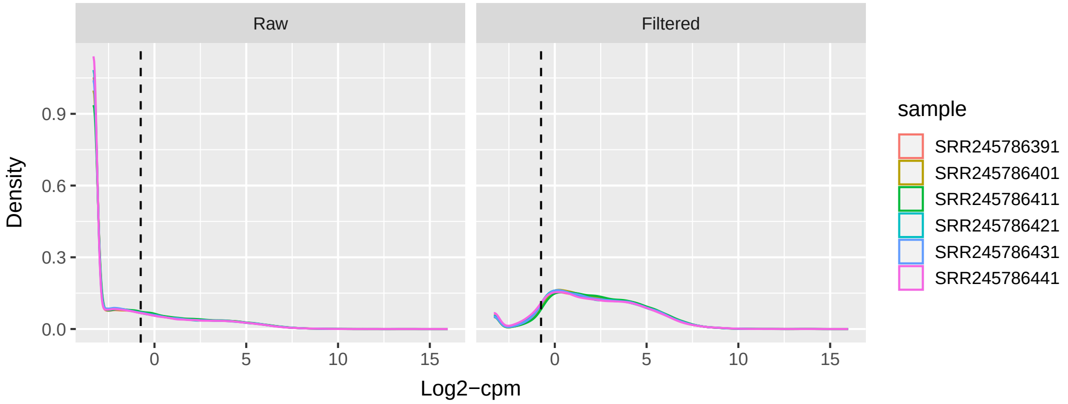

Figure 1: Filtered

Figure 2 (下方图) 为图 Normalization 概览。

(对应文件为 `Figure+Table/Normalization.pdf`)

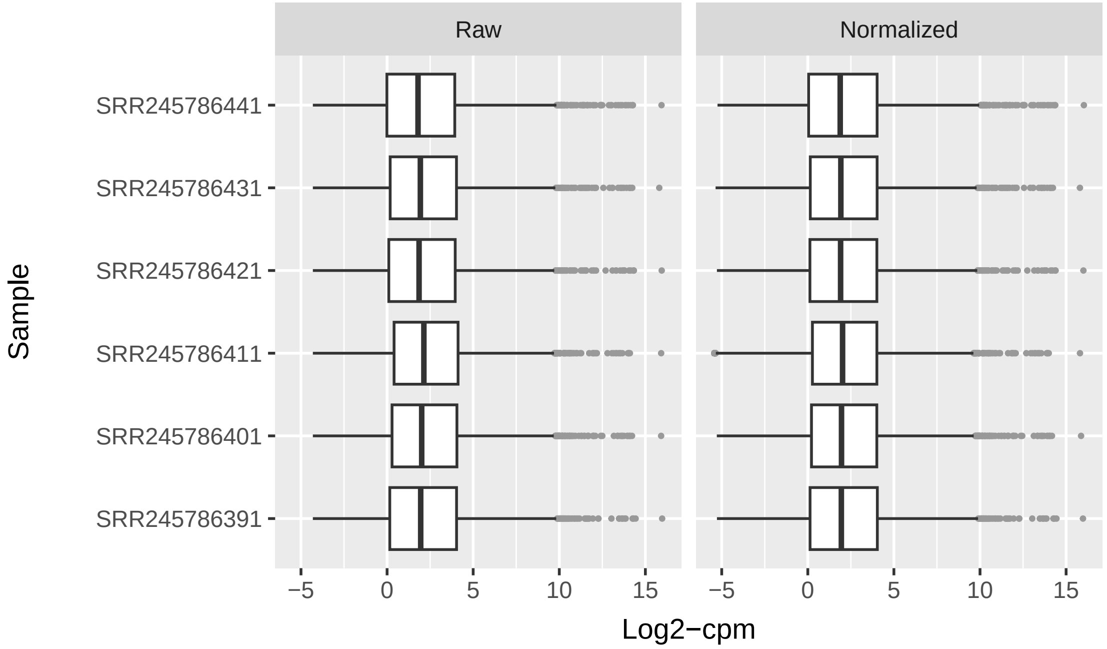

Figure 2: Normalization

6.3.2 结果

Figure 3 (下方图) 为图 Model vs control DEGs 概览。

(对应文件为 `Figure+Table/Model-vs-control-DEGs.pdf`)

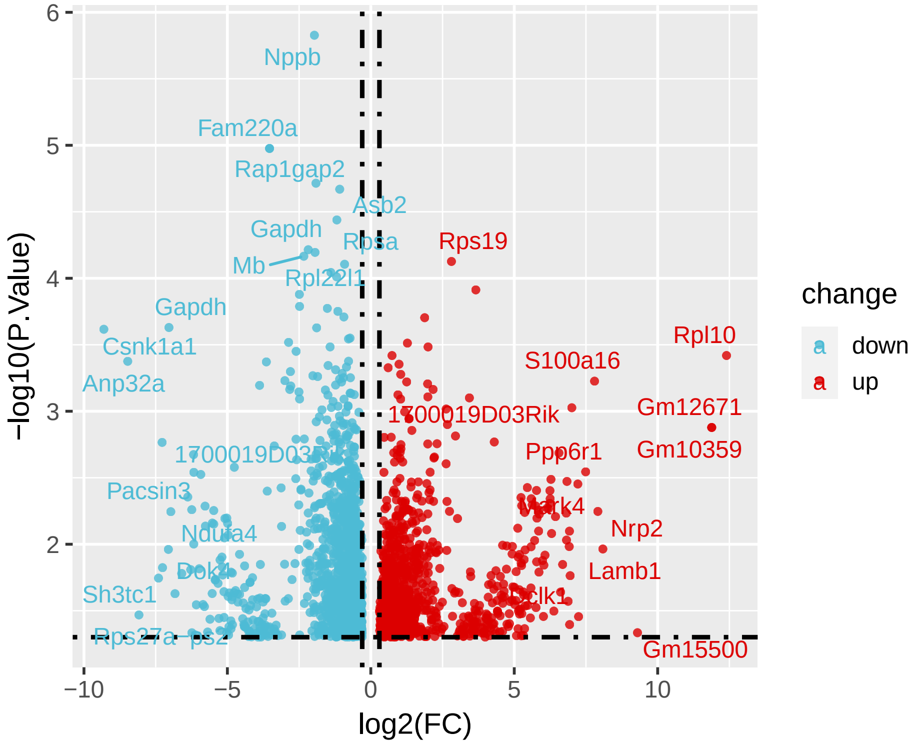

Figure 3: Model vs control DEGs

Table 4 (下方表格) 为表格 Data model vs control DEGs 概览。

(对应文件为 `Figure+Table/Data-model-vs-control-DEGs.xlsx`)

注：表格共有 2522 行 13 列，以下预览的表格可能省略部分数据；表格含有 2108 个唯一 ‘mgi_symbol’。

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

Table 4: Data model vs control DEGs

rownames	ensemb......2	mgi_sy...	ensemb......4	entrez...	hgnc_s...	descri...	logFC	AveExpr	...
7133	ENSMUS...	Nppb	ENSMUS...	18158	NA	natriu...	-1.967...	7.6029...	...
31330	ENSMUS...	Asb2	ENSMUS...	65256		ankyri...	-1.084...	6.5994...	...
89274	ENSMUS...	Rap1gap2	ENSMUS...	380711	NA	RAP1 G...	-1.912...	3.8988...	...
78044	ENSMUS...	Rpsa	ENSMUS...	16785	NA	riboso...	-1.184...	8.2501...	...
85206	ENSMUS...	Gapdh	ENSMUS...	14433	NA	glycer...	-2.180...	8.8656...	...
36842	ENSMUS...	Mb	ENSMUS...	17189	NA	myoglo...	-2.336...	6.9739...	...
96280	ENSMUS...	Fam220a	ENSMUS...	67238	NA	family...	-3.529...	2.4402...	...
96286	ENSMUS...	Fam220a	ENSMUS...	67238	NA	family...	-3.529...	2.4402...	...
70694	ENSMUS...	Fbxl22	ENSMUS...	74165	NA	F-box ...	-0.913...	6.5089...	...
60553	ENSMUS...	Rpl22l1	ENSMUS...	68028	NA	riboso...	-1.944...	3.7509...	...
82742	ENSMUS...	Copa	ENSMUS...	12847	NA	coatom...	-1.395...	5.2027...	...
13717	ENSMUS...	Ankrd1	ENSMUS...	107765	NA	ankyri...	-1.193...	10.790...	...
63231	ENSMUS...	Rps19	ENSMUS...	20085	NA	riboso...	2.8106...	3.0550...	...
51535	ENSMUS...	Fau	ENSMUS...	14109	NA	Finkel...	-2.483...	4.8874...	...
58716	ENSMUS...	Frmd5	ENSMUS...	228564	NA	FERM d...	-1.151...	4.5589...	...
...	...	...	...	...	...	...	...	...	...

6.3.3 基因名映射到人类的基因

Table 5 (下方表格) 为表格 Mapped Data model vs control DEGs 概览。

(对应文件为 `Figure+Table/Mapped-Data-model-vs-control-DEGs.xlsx`)

注：表格共有 2108 行 13 列，以下预览的表格可能省略部分数据；表格含有 1980 个唯一 ‘hgnc_symbol’。

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

Table 5: Mapped Data model vs control DEGs

hgnc_s...	mgi_sy...	logFC	P.Value	rownames	ensemb......6	ensemb......7	entrez...	descri...	...
NPPB	Nppb	-1.967...	1.4872...	7133	ENSMUS...	ENSMUS...	18158	natriu...	...
ASB2	Asb2	-1.084...	2.1399...	31330	ENSMUS...	ENSMUS...	65256	ankyri...	...
RAP1GAP2	Rap1gap2	-1.912...	1.9290...	89274	ENSMUS...	ENSMUS...	380711	RAP1 G...	...
RPSAP58	Rpsa	-1.184...	3.6418...	78044	ENSMUS...	ENSMUS...	16785	riboso...	...
GAPDH	Gapdh	-2.180...	6.0987...	85206	ENSMUS...	ENSMUS...	14433	glycer...	...
MB	Mb	-2.336...	6.8296...	36842	ENSMUS...	ENSMUS...	17189	myoglo...	...
FAM220A	Fam220a	-3.529...	1.0560...	96280	ENSMUS...	ENSMUS...	67238	family...	...
FBXL22	Fbxl22	-0.913...	7.8460...	70694	ENSMUS...	ENSMUS...	74165	F-box ...	...
RPL22L1	Rpl22l1	-1.944...	6.3883...	60553	ENSMUS...	ENSMUS...	68028	riboso...	...
COPA	Copa	-1.395...	9.0391...	82742	ENSMUS...	ENSMUS...	12847	coatom...	...
ANKRD1	Ankrd1	-1.193...	9.7473...	13717	ENSMUS...	ENSMUS...	107765	ankyri...	...
RPS19	Rps19	2.8106...	7.4734...	63231	ENSMUS...	ENSMUS...	20085	riboso...	...
FAU	Fau	-2.483...	0.0001...	51535	ENSMUS...	ENSMUS...	14109	Finkel...	...
FRMD5	Frmd5	-1.151...	0.0001...	58716	ENSMUS...	ENSMUS...	228564	FERM d...	...
NA	Rpl11	-0.940...	0.0001...	54544	ENSMUS...	ENSMUS...	67025	riboso...	...
...	...	...	...	...	...	...	...	...	...

6.4 内皮-间质转化 (endothelial-to-mesenchymal transition, ET)

6.4.1 ET 来源

从 GeneCards 获取相关的基因集。

Table 6 (下方表格) 为表格 ET related targets from GeneCards 概览。

(对应文件为 `Figure+Table/ET-related-targets-from-GeneCards.xlsx`)

注：表格共有 96 行 7 列，以下预览的表格可能省略部分数据；表格含有 96 个唯一‘Symbol’。

Table 6: ET related targets from GeneCards

Symbol	Description	Category	UniProt_ID	GIFtS	GC_id	Score
TGFB1	Transformi...	Protein Co...	P01137	61	GC19M041301	5.71
H19	H19 Imprin...	RNA Gene		34	GC11M001995	4.52
MIR21	MicroRNA 21	RNA Gene		31	GC17P102034	4.40
BMP7	Bone Morph...	Protein Co...	P18075	55	GC20M057168	3.85
MIR126	MicroRNA 126	RNA Gene		29	GC09P136670	3.49
MIRLET7C	MicroRNA L...	RNA Gene		28	GC21P017033	3.49
CTNNB1	Catenin Be...	Protein Co...	P35222	62	GC03P041194	3.41
TGFB2	Transformi...	Protein Co...	P61812	60	GC01P218345	3.41
TMX2-CTNND1	TMX2-CTNND...	RNA Gene		23	GC11P057712	2.96
BMPR2	Bone Morph...	Protein Co...	Q13873	59	GC02P202376	2.88
ROCK1	Rho Associ...	Protein Co...	Q13464	57	GC18M032996	2.88
SNAI1	Snail Fami...	Protein Co...	O95863	52	GC20P049982	2.88
MALAT1	Metastasis...	RNA Gene		31	GC11P084571	2.88
MIR532	MicroRNA 532	RNA Gene		23	GC0XP056752	2.88
RUNX3	RUNX Famil...	Protein Co...	Q13761	51	GC01M024899	2.78
...	...	...	...	...	...	...

6.4.2 与 DEG 交集

Figure 4 (下方图) 为图 Intersection of ET with DEGs 概览。

(对应文件为 `Figure+Table/Intersection-of-ET-with-DEGs.pdf`)

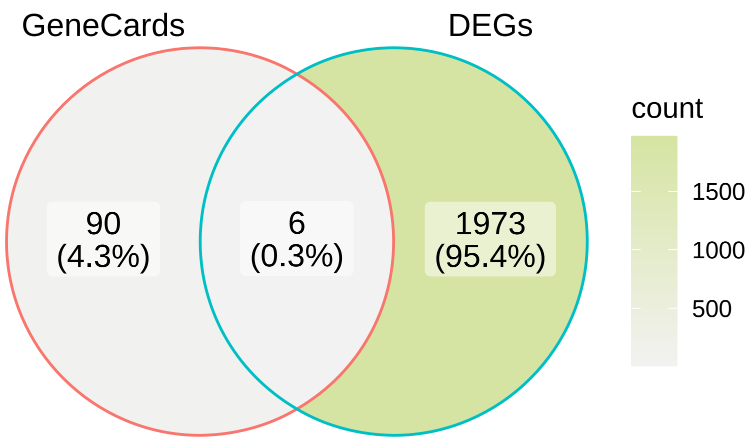

Figure 4: Intersection of ET with DEGs

Intersection :

CTNNB1, NFKB1, HSPB1, ACVRL1, ACTA2, FOXO1

(上述信息框内容已保存至 `Figure+Table/Intersection-of-ET-with-DEGs-content`)

6.5 富集分析

Figure 5 (下方图) 为图 Ids KEGG enrichment 概览。

(对应文件为 `Figure+Table/Ids-KEGG-enrichment.pdf`)

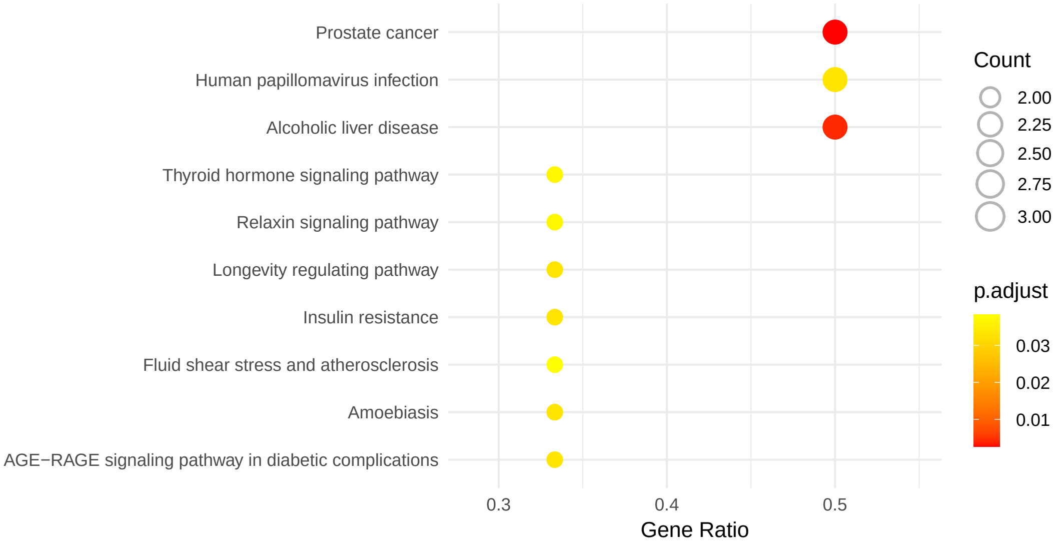

Figure 5: Ids KEGG enrichment

Figure 6 (下方图) 为图 Ids GO enrichment 概览。

(对应文件为 `Figure+Table/Ids-GO-enrichment.pdf`)

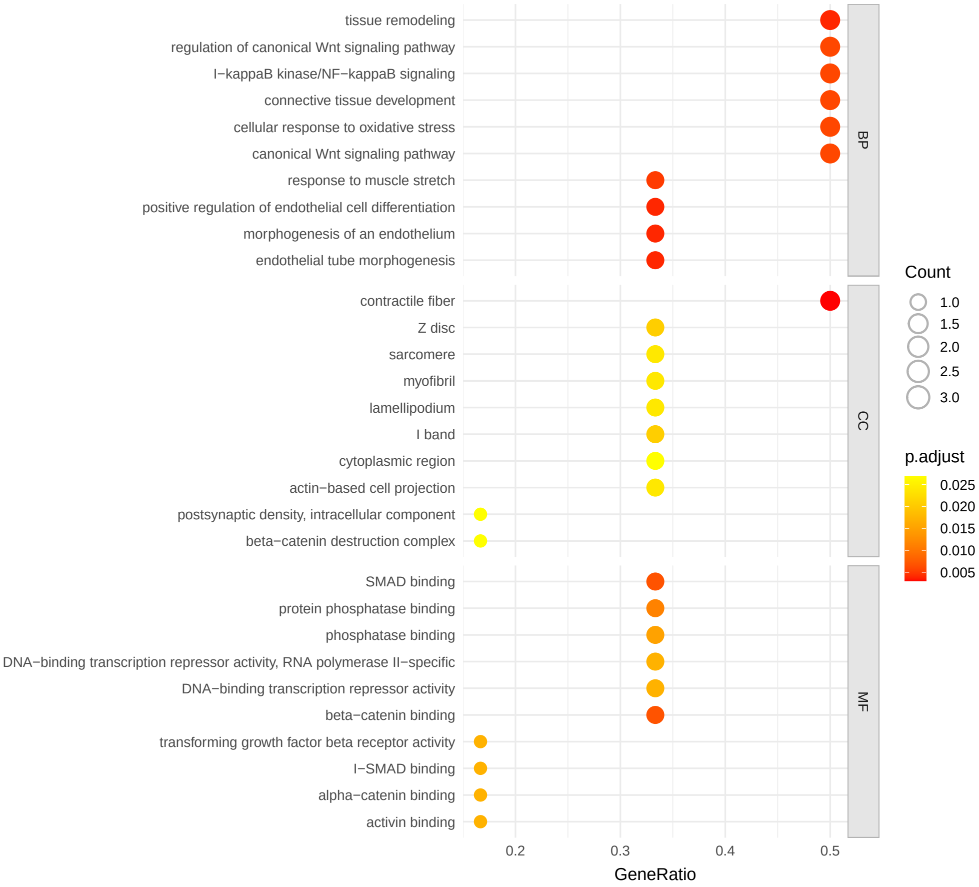

Figure 6: Ids GO enrichment

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. Wu, T. *et al.* ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).

3. Chen, S. Ultrafast one-pass fastq data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, (2023).

4. Stelzer, G. *et al.* The genecards suite: From gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **54**, 1.30.1–1.30.33 (2016).

5. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology* **34**, (2016).

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

7. Chen, Y., McCarthy, D., Ritchie, M., Robinson, M. & Smyth, G. EdgeR: Differential analysis of sequence

read count data user's guide. 119.