Analysis

Huang LiChuang of Wie-Biotech

Contents

List of Figures

List of Tables

1 Title

结合 scRNA-seq 和 bulk RNA-seq 探究昼夜节律基因对于 ccRCC 的预后评估价值

（主题：昼夜节律，肿瘤，预后单细胞分析）[1]

2 Abstract

3 Introduction

生物体的行为和新陈代谢每 24 小时就会出现有节奏的波动，以预测环境的变化。这种节律时钟受到环境信息的外部和内部信号影响。最重要的外部时间是光，特别是光暗循环。所有这些信号共同改变某些基因的表达水平以及控制生理过程的代谢物或激素的产生，其改变会导致健康损害。事实上，生物钟（circadian clock，CC）可以调节：细胞周期进展、DNA 修复机制、线粒体功能障碍、代谢重编程、免疫系统[2]。CC 的改变是多种慢性疾病的危险因素，例如神经退行性疾病、糖尿病和癌症等[2,3]。事实上，CC 基因表达的改变可能导致常见恶性肿瘤的发生。虽然 CC 基因在正常生理中的功能已经被充分阐明，但 CC 基因在癌症中的改变研究仍然缺乏，例如肾细胞癌（Renal Cell Carcinoma，RCC），这使得其在肿瘤细胞中的功能的清晰度和描述存在空白[4]。CC 基因和昼夜节律在 RCC 患者中的作用是一个较新研究领域，特别是考虑到昼夜节律在抗癌治疗中的影响的新发现。在肾细胞癌中，目前治疗方法缺乏有效的疗效或耐药性生物标志物，这强烈支持探索一系列可能影响患者治疗结果的临床和行为变量的必要性[4]。

本研究通过分析肾癌单细胞测序数据集[5]，继标准的单细胞处理流程[6,7]，囊括细胞群鉴定[8,9]，拟时序分析[10,11]，加权基因共表达分析[12]，差异性分析，通路分析[13]，以多种方法筛选关键 CC 基因[14]，并筛选 top genes[14]，借助 LASSO 回归建立预后模型，表征肿瘤发展和预后效果，并在随后进行生存分析、细胞通讯分析[15] 等，较为深入的探讨了昼夜节律在肾细胞癌（主要为 Clear cell renal cell carcinoma，ccRCC）中的作用。

4 Methods

5 Results

5.1 分析 scRNA-seq 数据聚焦于 ccRCC

为了研究昼夜节律和相关基因对 ccRCC 的影响，选择 GSE207493 为分析对象[5]。该数据集共 19 个样本数据，为减少计算时间，随机选择 9 个样本，以 Seurat 整合数据集，消除批次效应。Fig. 1a 为各个样本在 PC1、PC2 上的分布。随后，根据 Fig. 1b 所示，选择 PC1-PC5 作为主要成分，进行 UMAP 降维和聚类。聚类结果如 Fig. 1c 所示。为了注释 UMAP 的聚类结果，此处获取了 Yu 等人注释的 Markder 基因（见 Tab. 2），将其重新映射到 UMAP 结果中，即 Fig. 1d。随后的分析，聚焦于 ccRCC1-4。

Figure 1为图 MAIN Preprocess GEO samples and annotate cell typle for focusing on ccRCC 概览。

(对应文件为 `./Figure+Table/fig1.pdf`)

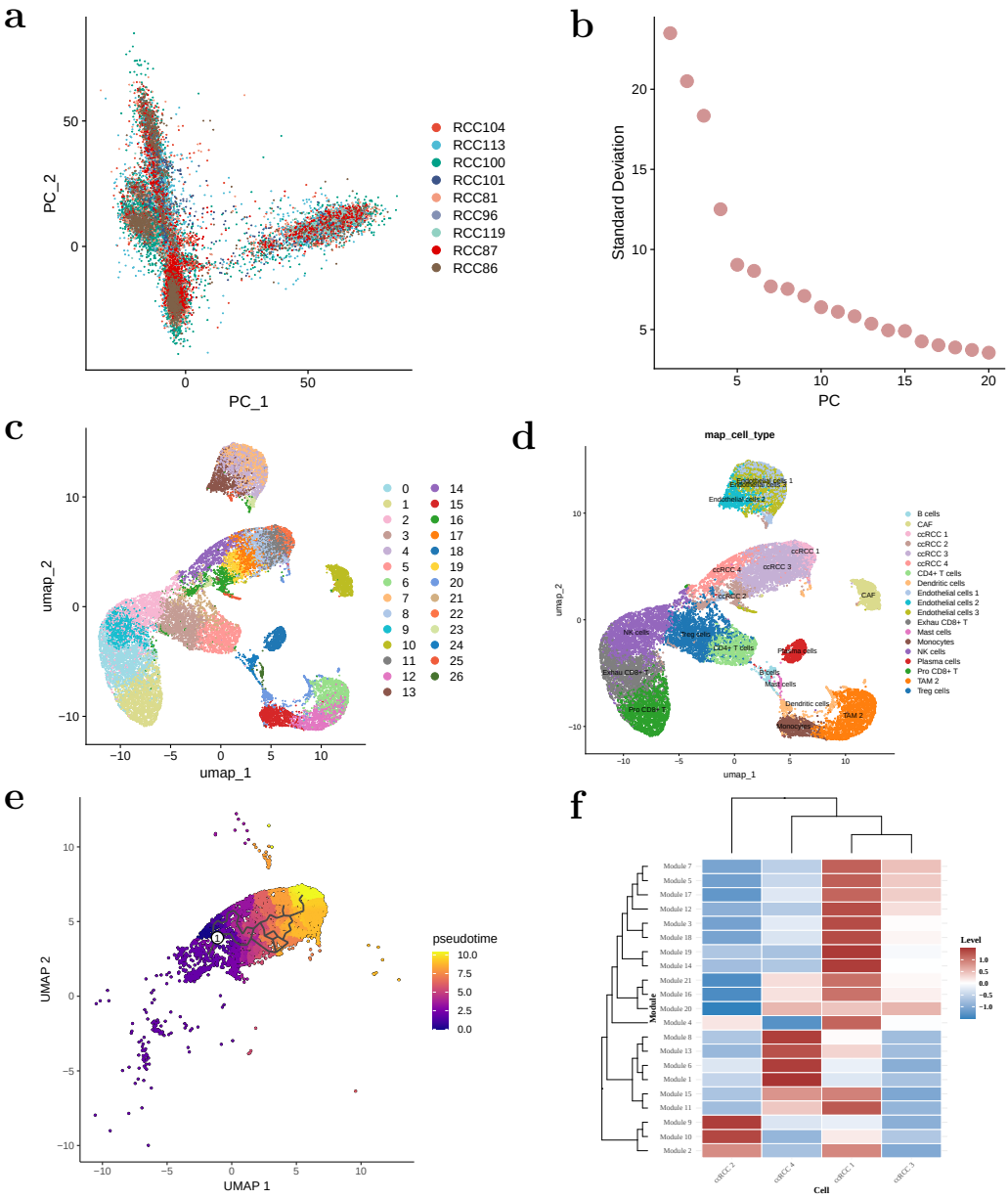


Figure 1: MAIN Preprocess GEO samples and annotate cell typle for focusing on ccRCC

5.2 WGCNA 寻找关键基因

Figure 2为图 MAIN Weighting the genes correlation network with CRDscore for finding critical genes 概览。

(对应文件为 `./Figure+Table/fig2.pdf`)

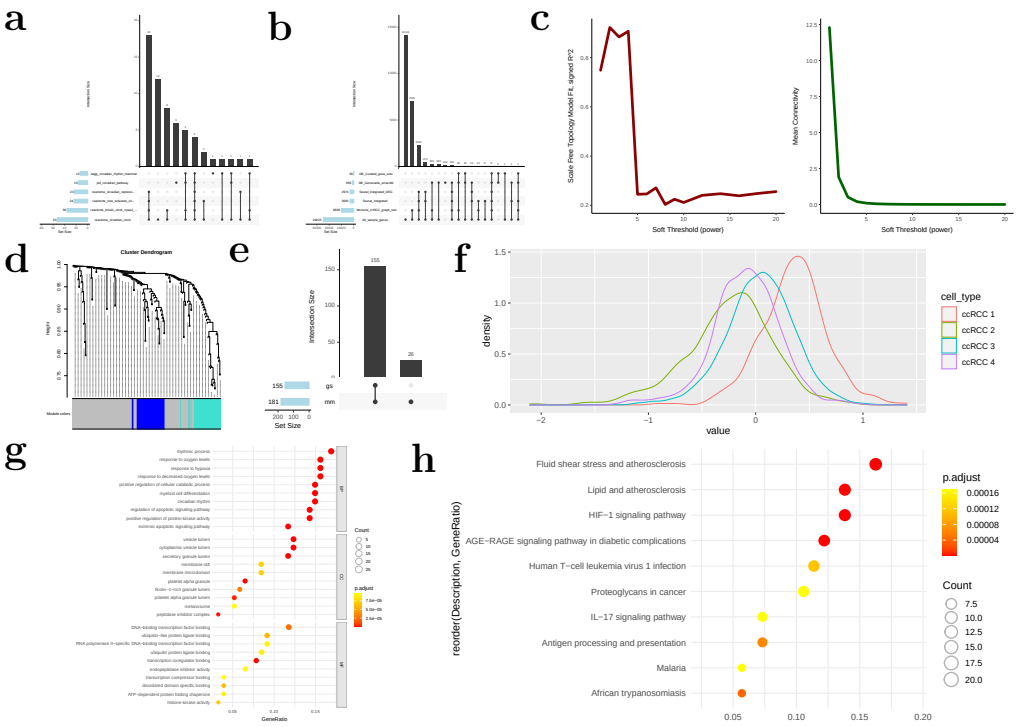


Figure 2: MAIN Weighting the genes correlation network with CRDscore for finding critical genes

5.2.1 基因数据库

5.2.2 共表达模块与显著基因

5.2.3 通路富集

5.3 CellChat 细胞通讯

Figure 3为图 MAIN Explore cell communications for finding critical cell types and signaling 概览。

(对应文件为 `./Figure+Table/fig3.pdf`)

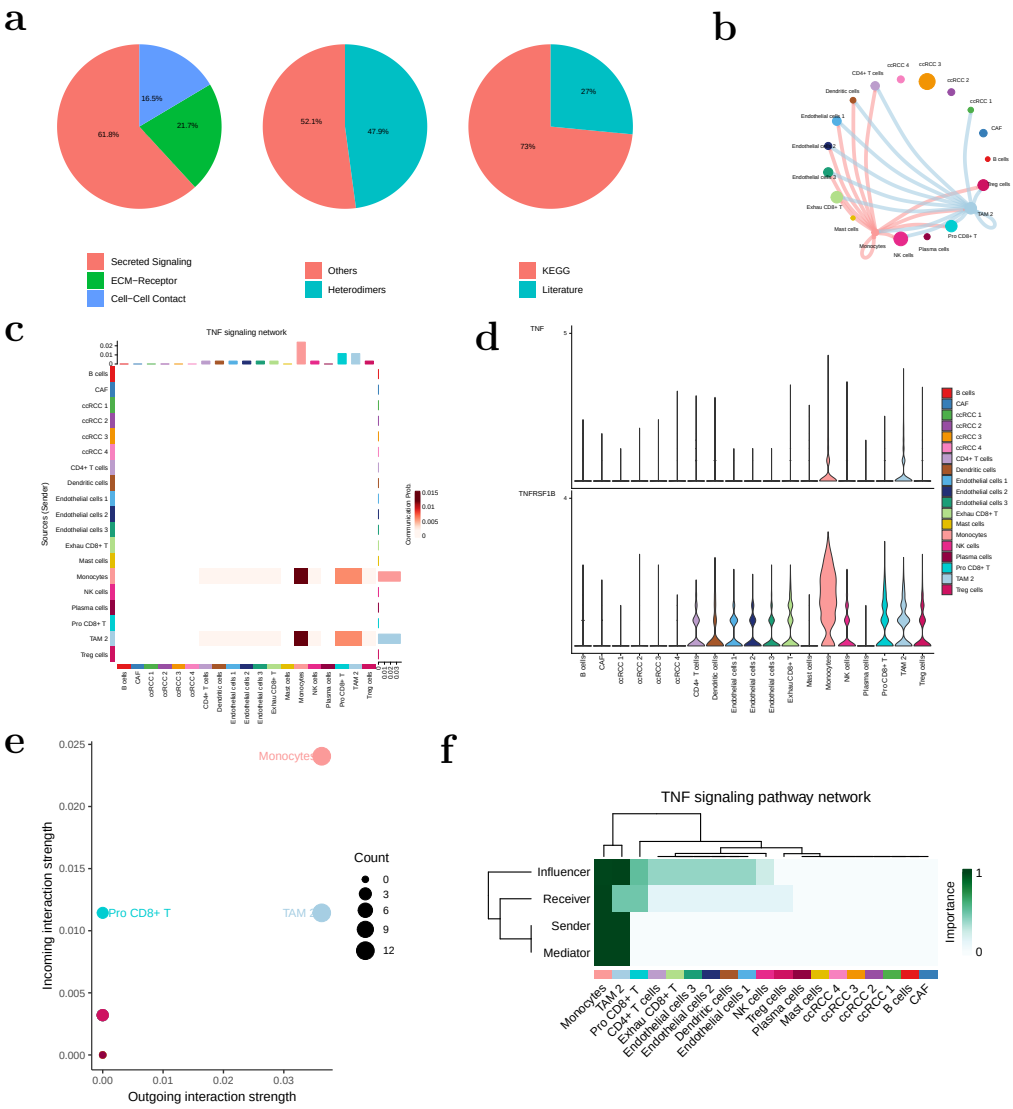


Figure 3: MAIN Explore cell communications for finding critical cell types and signaling

5.4 TCGA RNA-seq 预后模型

5.4.1 Feature selection

Figure 4为图 MAIN Perform Feature selection and LASSO regression on TCGA dataset 概览。

(对应文件为 `./Figure+Table/fig4.pdf`)

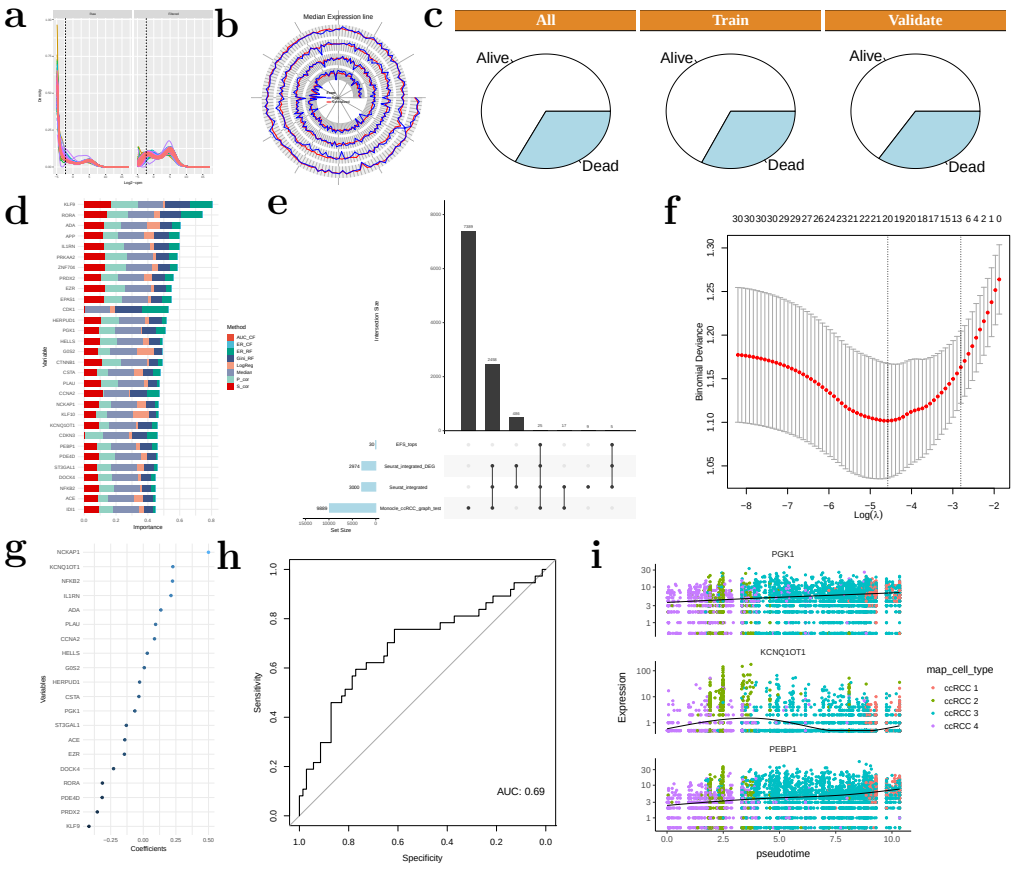


Figure 4: MAIN Perform Feature selection and LASSO regression on TCGA dataset

5.4.2 Survival

Figure 5为图 MAIN Survival analysis for validation 概览。

(对应文件为 `./Figure+Table/fig5.pdf`)

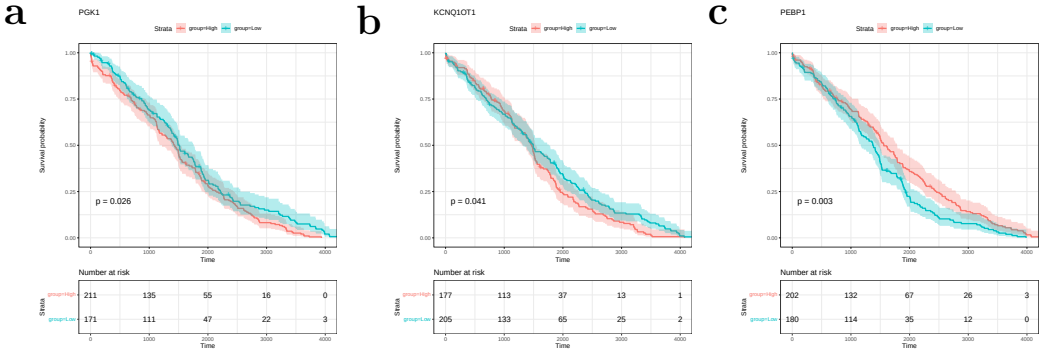


Figure 5: MAIN Survival analysis for validation

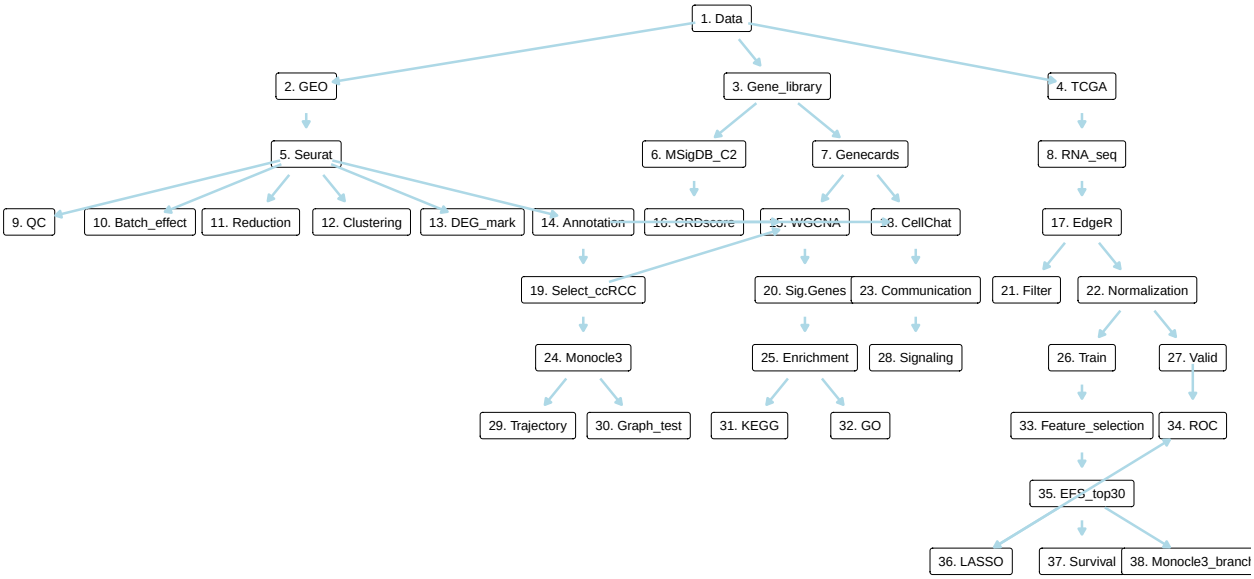
6 Discussion

7 附：分析流程

7.1 方案

- 筛选数据集
- Seurat 分析
  - 数据预处理 QC
  - 消除批次效应
  - 数据降维、聚类
  - 差异表达分析
  - 细胞群鉴定
- Monocle 分析
  - 细胞轨迹图
  - 拟时序分析
- MSigDB
  - https://bioinf.wehi.edu.au/software/MSigDB/
  - https://www.gsea-msigdb.org/gsea/msigdb
- WGCNA 分析
  - 预处理上述筛选的数据
  - 共表达基因模块
  - 结合 CRDscore 加权分析
- Clusterprofiler
  - 通路富集分析
- EFS
  - Feature selection
- LASSO
  - regression
- Cellchat 分析
  - 聚焦于上述筛选的关键基因
  - 目标基因所在细胞和其他细胞的通讯
  - 目标基因的通路（Role）
- Survival

7.2 流程设计



7.3 分析

7.3.1 数据来源

重新分析来源于 GSE207493[5] 的数据。关于 clear cell renal cell carcinoma（ccRCC）[16]。

data_processing :

Preliminary sequencing results (bcl files) were converted to FASTQ files with Cell Ranger V3.0 and CellRanger ATAC V1.2.0

data_processing.1 :

The CellRanger (10X Genomics) secondary analysis pipeline was used to generate a digital gene expression matrix

data_processing.2 :

Normalization and additional analysis by Seurat and Signac R package

data_processing.3 :

Assembly: GRCh38 for human data

data_processing.4 :

Supplementary files format and content: The barcodes, genes, expression matrix files for scRNA-seq. The peaks.bed, singlecell.csv and fragments.tsv.gz files for scATAC-seq.

7.3.2 数据预处理 Seruat

GSE207493 数据集共 19 个样本，随机选择 9 个样本用于分析。根据以下流程预处理，消除批次效应。

- https://satijalab.org/seurat/articles/sctransform_vignette
- https://satijalab.org/seurat/articles/integration_introduction

Figure 6为图 Normalization pca samples 概览。

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

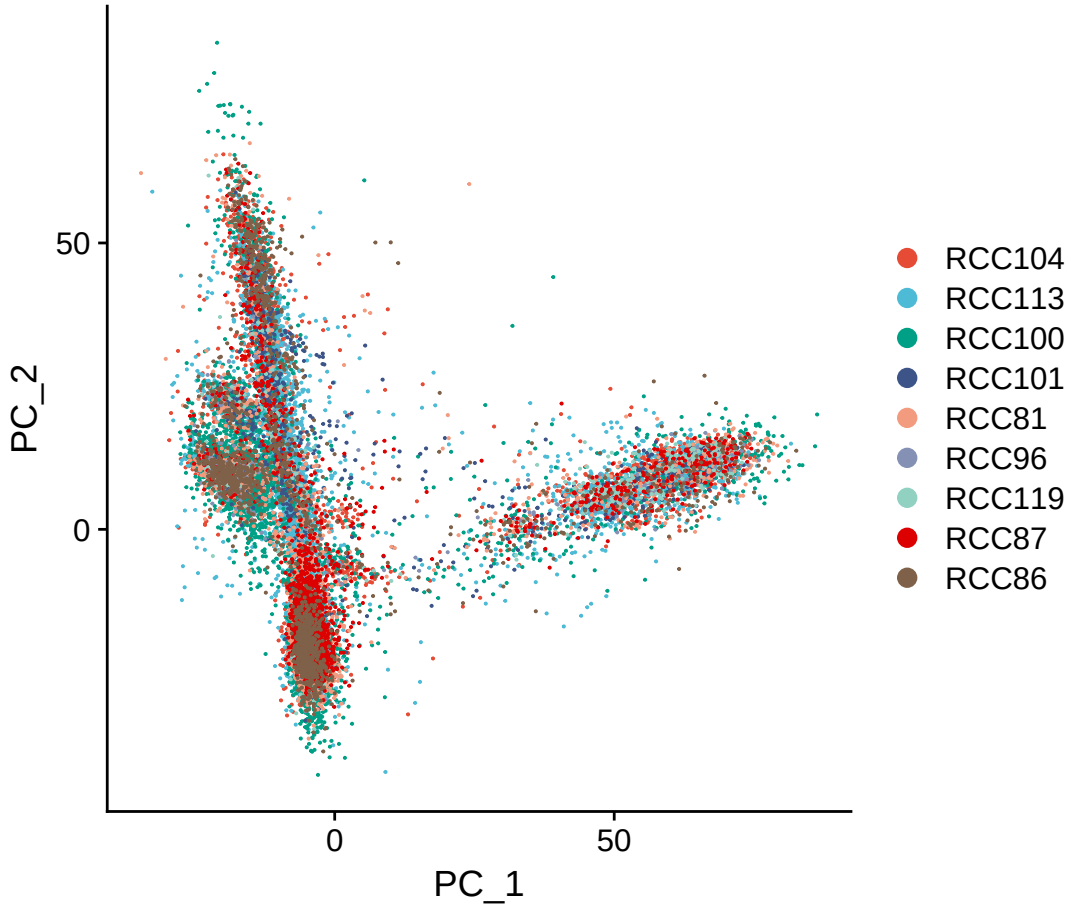


Figure 6: Normalization pca samples

Figure 7为图 pca rank 概览。根据图示，使用 1-5 个 PC 进行 UMAP 聚类。

(对应文件为 `Figure+Table/pca-rank.pdf`)

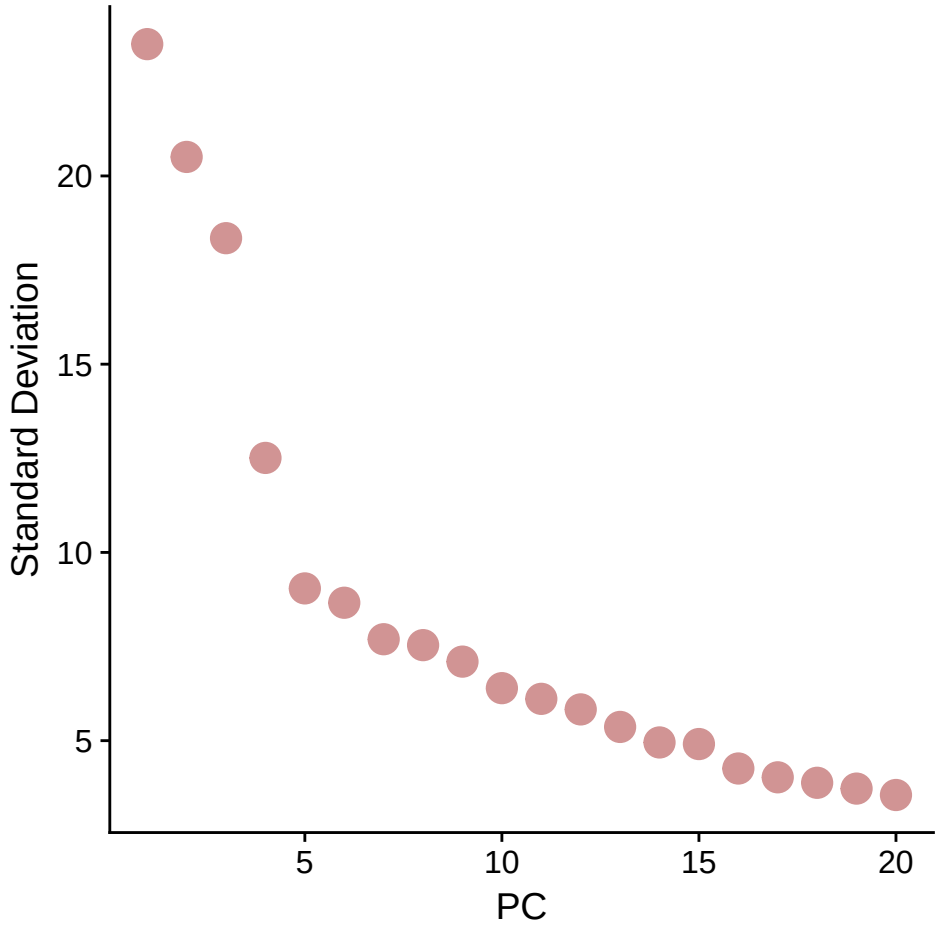


Figure 7: Pca rank

Figure 8为图 seurat cluster UMAP 概览。

(对应文件为 `Figure+Table/seurat-cluster-UMAP.pdf`)

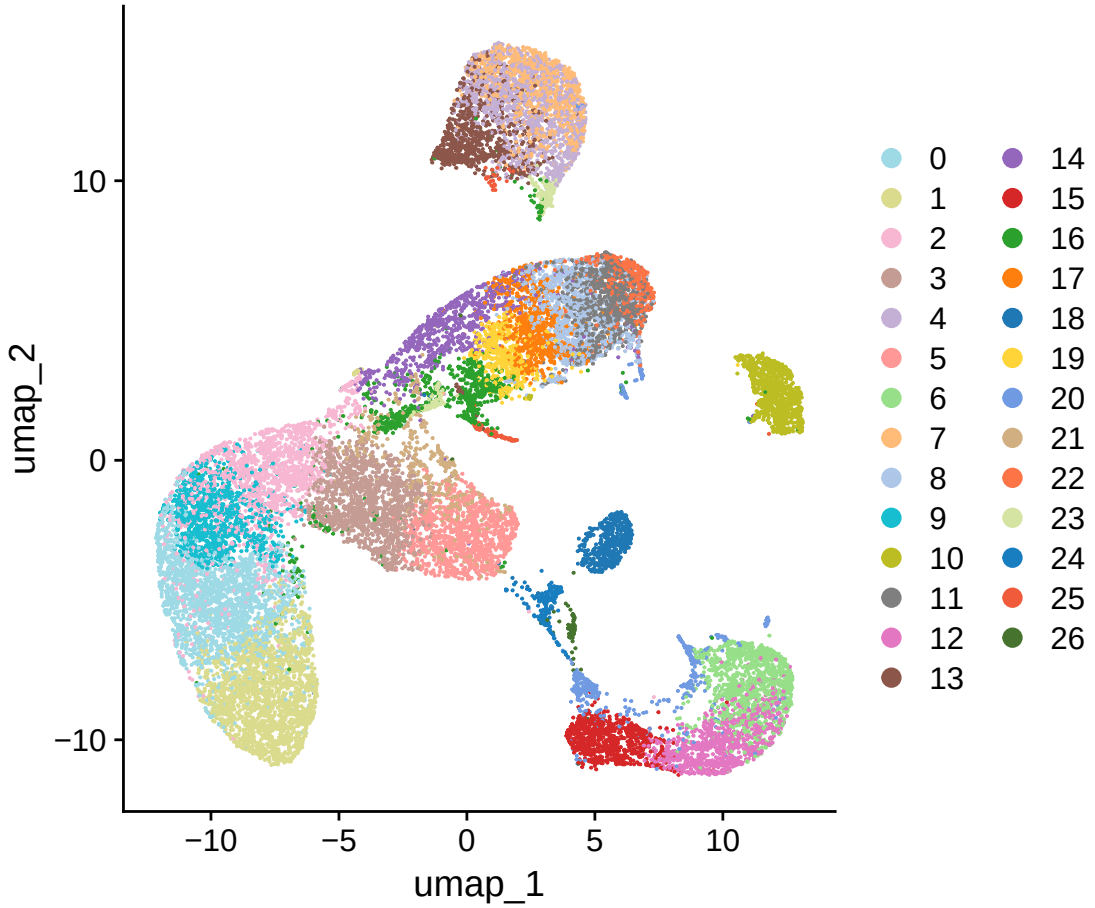


Figure 8: Seurat cluster UMAP

7.3.3 差异表达

对 UMAP 聚类结果进行差异表达分析。

Table 1为表格 Differential expressed genes 概览。

(对应文件为 `Figure+Table/Differential-expressed-genes.csv`)

> 注：表格共有 11469 行 8 列，以下预览的表格可能省略部分数据；表格含有 27 个唯一 'cluster'。

Table 1: Differential expressed genes

rownames	p_val	avg_l...	pct.1	pct.2	p_val...	cluster	gene
EFHD2	0	4.667...	0.722	0.335	0	0	EFHD2
CD247	0	4.354...	0.814	0.295	0	0	CD247
PRF1	0	4.061...	0.768	0.218	0	0	PRF1
AKNA	0	3.938...	0.774	0.329	0	0	AKNA
KLRD1	0	3.743...	0.583	0.16	0	0	KLRD1

rownames	p_val	avg_l...	pct.1	pct.2	p_val...	cluster	gene
FGFBP2	0	3.741...	0.524	0.113	0	0	FGFBP2
GZMB	0	3.594...	0.617	0.15	0	0	GZMB
GZMM	0	3.304...	0.755	0.269	0	0	GZMM
NKG7	0	3.245...	0.976	0.244	0	0	NKG7
PYHIN1	0	3.227...	0.823	0.287	0	0	PYHIN1
TXK	0	3.213...	0.638	0.196	0	0	TXK
STK17B	0	3.198...	0.855	0.401	0	0	STK17B
MYO1F	0	3.195...	0.755	0.37	0	0	MYO1F
GZMH	0	3.042...	0.775	0.211	0	0	GZMH
HCST	0	2.978...	0.924	0.409	0	0	HCST
...	...	...	...	...	...	...	...

7.3.4 细胞群注释

获取原作者的研究[5] 数据（补充材料），将差异表达基因映射到 UMAP 图中。

Table 2为表格 Original differential expressed genes DEG 概览。该表格为原作者的研究数据。

(对应文件为 `Figure+Table/Original-differential-expressed-genes-DEG.csv`)

> 注：表格共有 18602 行 7 列，以下预览的表格可能省略部分数据；表格含有 22 个唯一 'cluster'。

Table 2: Original differential expressed genes DEG

p_val	avg_l...	pct.1	pct.2	p_val...	cluster	gene
0	3.780...	0.961	0.224	0	ccRCC 1	CRYAB
0	3.216...	0.978	0.141	0	ccRCC 1	CD24
0	3.133...	0.941	0.19	0	ccRCC 1	TMEM176A
0	3.085...	0.962	0.271	0	ccRCC 1	NDUFA4L2
0	3.068...	0.895	0.152	0	ccRCC 1	SPP1
0	3.048...	0.846	0.109	0	ccRCC 1	HILPDA
0	2.887...	0.914	0.162	0	ccRCC 1	NNMT
0	2.865...	0.907	0.103	0	ccRCC 1	RARRES2
0	2.861...	0.927	0.131	0	ccRCC 1	KRT18
0	2.805...	0.949	0.232	0	ccRCC 1	ATP1B1
0	2.649...	0.836	0.104	0	ccRCC 1	ANGPTL4
0	2.625...	0.921	0.221	0	ccRCC 1	TMEM176B
0	2.617...	0.953	0.273	0	ccRCC 1	CYB5A
0	2.602...	0.93	0.201	0	ccRCC 1	VEGFA
0	2.559...	0.886	0.096	0	ccRCC 1	KRT8

p_val	avg_l...	pct.1	pct.2	p_val...	cluster	gene
...	...	...	...	...	...	...

重新映射的算法：

1. 将 UMAP 聚类结果中的每一个类群的 DEG 与原 DEG 数据比对
2. 优先保留所有细胞类型
3. 将最优比对结果注释（覆盖最多原 DEG）

Figure 9为图 map cell types in UMAP 概览。根据原作者的差异表达基因和对应的细胞类型，映射到 UMAP 图中。

(对应文件为 `Figure+Table/map-cell-types-in-UMAP.pdf`)

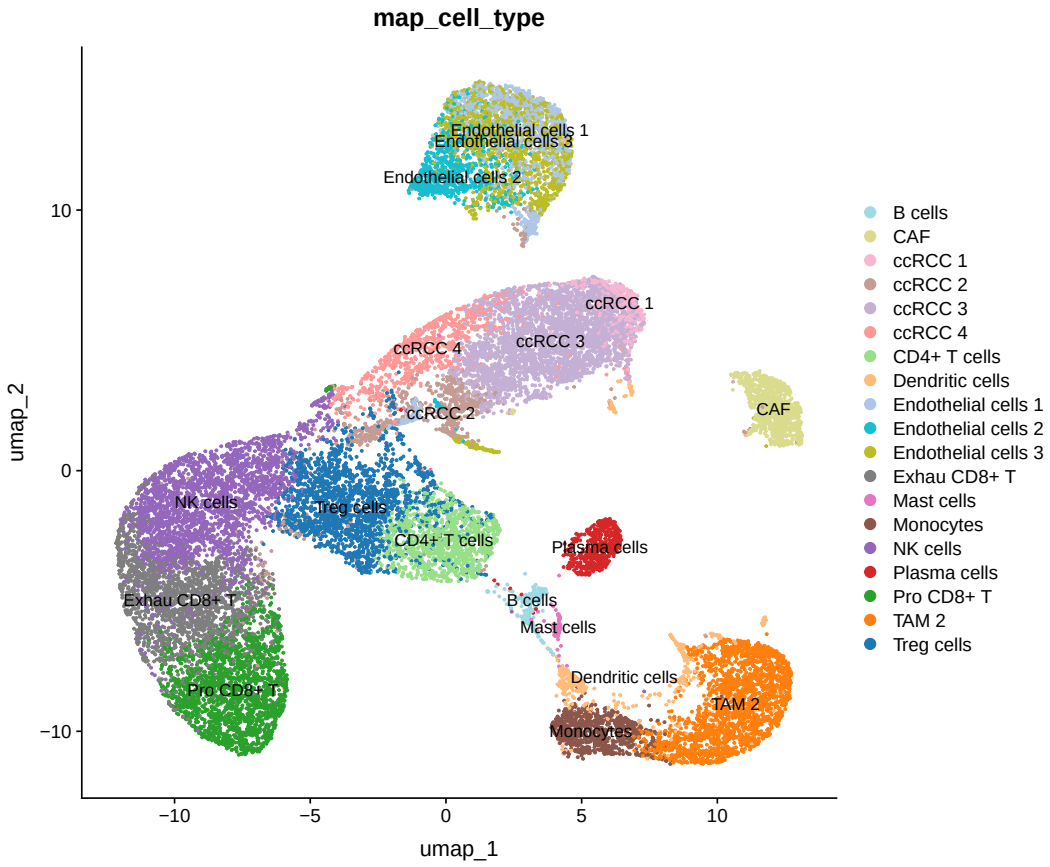


Figure 9: Map cell types in UMAP

根据原作者研究[5]，'ccRCC 4' 关联 epithelial cells 分化，为肿瘤的早期阶段。

7.3.5 拟时序分析 Monocle3

选择 ccRCC 1, ccRCC 2, ccRCC 3, ccRCC 4 作为分支数据，将 Seurat 的聚类结果以 Monocle3 做拟时序分析。

Figure 10为图 pseudotime of ccRCC cells 概览。

(对应文件为 `Figure+Table/pseudotime-of-ccRCC-cells.pdf`)

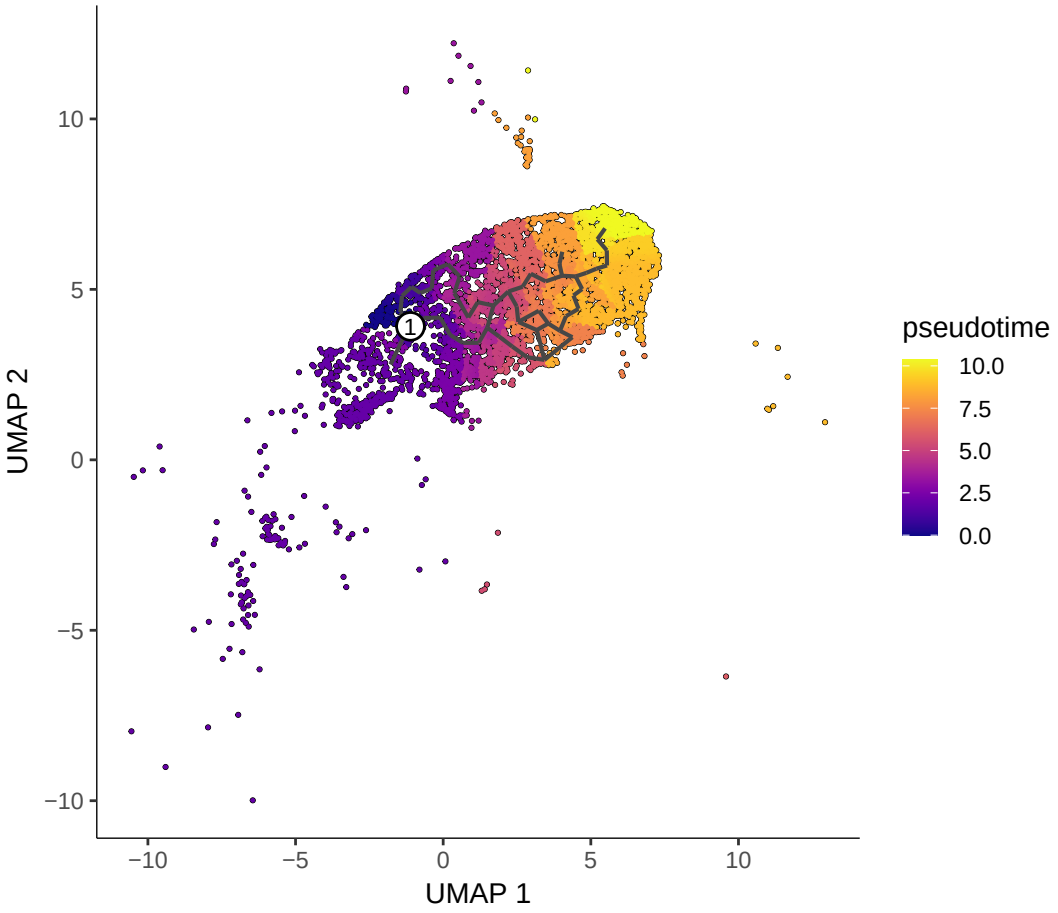


Figure 10: Pseudotime of ccRCC cells

在拟时序轨迹中寻找差异基因（`graph_test`）。

Table 3为表格 differential expressed genes DEG in trajectory 概览。

(对应文件为 `Figure+Table/differential-expressed-genes-DEG-in-trajectory.csv`)

> 注：表格共有 24823 行 6 列，以下预览的表格可能省略部分数据；表格含有 24823 个唯一 'rownames'。

Table 3: Differential expressed genes DEG in trajectory

rownames	status	p_value	moran...	morans_I	q_value
AL627...	OK	0.519...	-0.04...	-0.00...	0.699...
AL669...	OK	0.283...	0.571...	0.001...	0.449...
FAM87B	OK	0.567...	-0.17...	-0.00...	0.699...
LINC0...	OK	0.457...	0.106...	0.000...	0.656...

rownames	status	p__value	moran...	morans_I	q__value
FAM41C	OK	0.306...	0.506...	0.001...	0.476...
AL645...	OK	0.610...	-0.28...	-0.00...	0.709...
AL645...	OK	0.719...	-0.57...	-0.00...	0.769...
SAMD11	OK	0.419...	0.203...	0.000...	0.614...
NOC2L	OK	0.316...	0.476...	0.001...	0.490...
KLHL17	OK	0.350...	0.382...	0.001...	0.534...
PLEKHN1	OK	0.477...	0.056...	2.300...	0.680...
PERM1	OK	0.850...	-1.03...	-0.00...	0.868...
AL645...	OK	0.251...	0.669...	0.002...	0.406...
HES4	OK	2.821...	9.635...	0.035...	1.934...
ISG15	OK	1.840...	7.865...	0.029...	1.051...
...	...	...	...	...	...

Figure 11为图 pseudotime DEG in module 概览。

(对应文件为 `Figure+Table/pseudotime-DEG-in-module.pdf`)

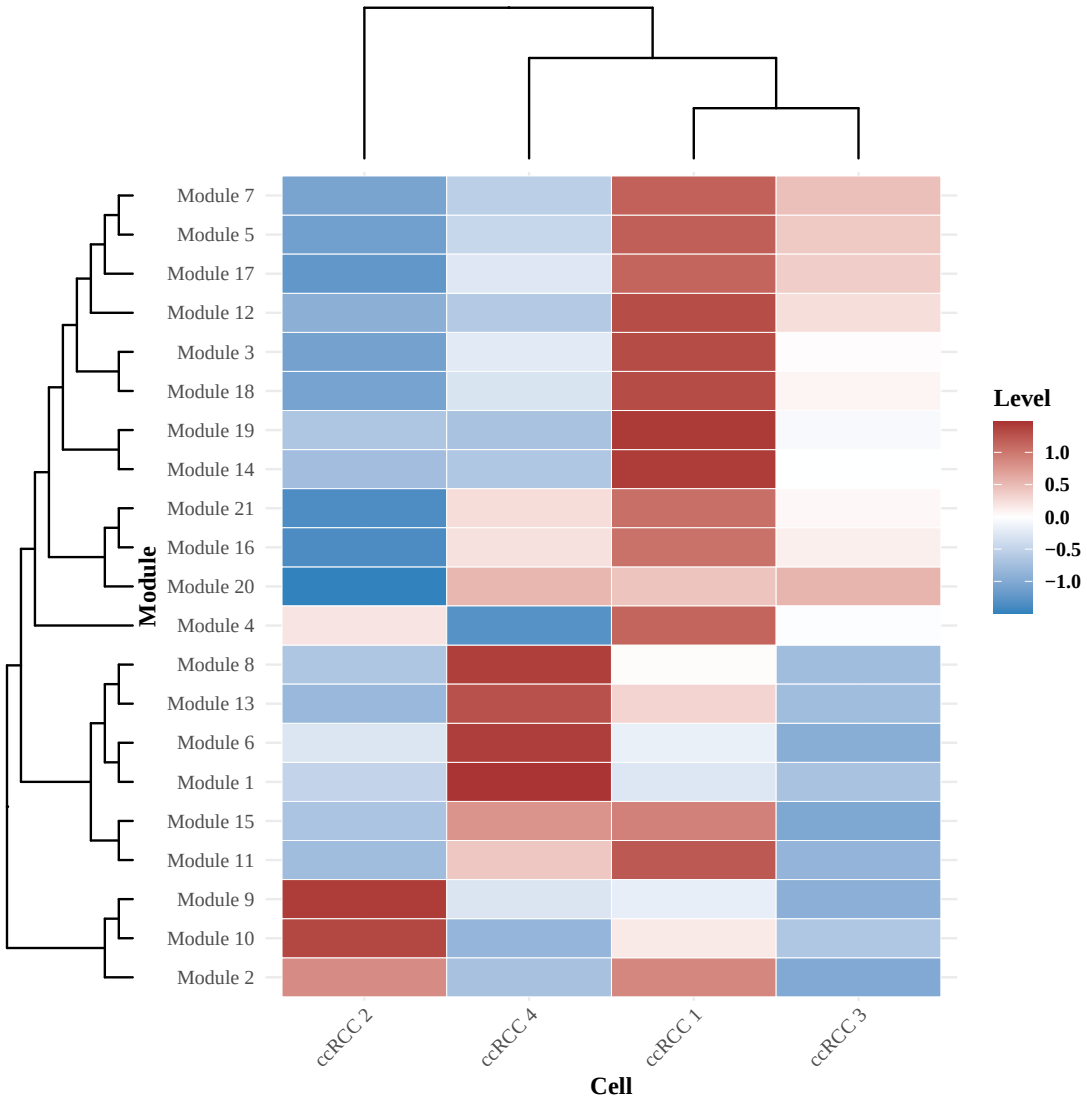


Figure 11: Pseudotime DEG in module

7.3.6 获取昼夜节律相关基因集和注释

从 https://www.genecards.org/ genecards 检索 circadian 相关基因。从 https://bioinf.wehi.edu.au/software/MSigDB/ 获取 C2 基因集合 (curated gene sets)。以 `biomaRt` 注释这些基因（无法获取注释的基因不再用于后续分析）。

检查 C2 基因集关于 circadian 的来源数据（Fig. 12）。关于 UpSet 图[17]。

Figure 12为图 intersect of sources of c2 genes 概览。

(对应文件为 `Figure+Table/intersect-of-sources-of-c2-genes.pdf`)

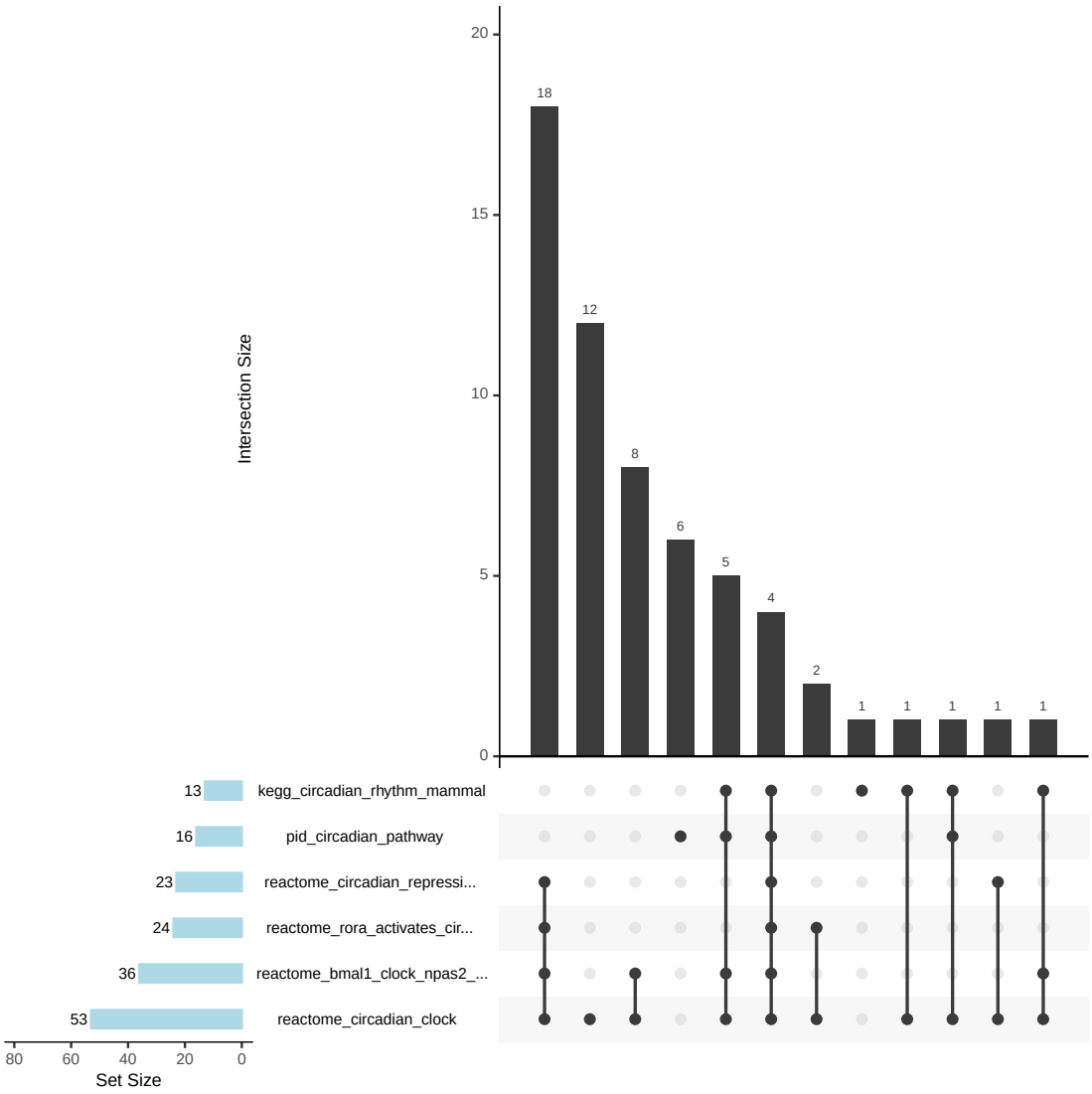


Figure 12: Intersect of sources of c2 genes

Table 4为表格 annotation of c2 gene set 概览。

(对应文件为 `Figure+Table/annotation-of-c2-gene-set.xlsx`)

> 注：表格共有 59 行 7 列，以下预览的表格可能省略部分数据；表格含有 59 个唯一 'ensembl_gene_id'。

Table 4: Annotation of c2 gene set

ensem...	entre...	hgnc_...	chrom...	start...	end_p...	descr...
ENSG0...	10135	NAMPT	7	10624...	10628...	nicot...
ENSG0...	10499	NCOA2	8	70109782	70403808	nucle...
ENSG0...	10891	PPARGC1A	4	23755041	23904089	PPARG...

ensem…	entre…	hgnc_…	chrom…	start…	end_p…	descr…
ENSG0…	11091	WDR5	9	13413…	13415…	WD re…
ENSG0…	1111	CHEK1	11	12562…	12567…	check…
ENSG0…	1374	CPT1A	11	68754620	68844410	carni…
ENSG0…	1385	CREB1	2	20752…	20760…	cAMP …
ENSG0…	1386	ATF2	2	17507…	17516…	activ…
ENSG0…	1387	CREBBP	16	3725054	3880713	CREB …
ENSG0…	1407	CRY1	12	10699…	10709…	crypt…
ENSG0…	1408	CRY2	11	45847118	45883248	crypt…
ENSG0…	1453	CSNK1D	17	82239023	82273700	casei…
ENSG0…	1454	CSNK1E	22	38290691	38318084	casei…
ENSG0…	1628	DBP	19	48630030	48637379	D-box…
ENSG0…	2033	EP300	22	41092592	41180077	E1A b…
…	…	…	…	…	…	…

Table 5为表格 annotation of genecards gene set 概览。

(对应文件为 `Figure+Table/annotation-of-genecards-gene-set.xlsx`)

> 注：表格共有 1145 行 7 列，以下预览的表格可能省略部分数据；表格含有 1084 个唯一 'ensembl_gene_id'。

Table 5: Annotation of genecards gene set

ensem…	entre…	hgnc_…	chrom…	start…	end_p…	descr…
ENSG0…	15	AANAT	17	76453351	76470117	aralk…
ENSG0…	18	ABAT	16	8674596	8784575	4-ami…
ENSG0…	5243	ABCB1	7	87503017	87713323	ATP b…
ENSG0…	22	ABCB7	X	75051048	75156732	ATP b…
ENSG0…	51099	ABHD5	3	43690108	43734371	abhyd…
ENSG0…	31	ACACA	HSCHR…	1320988	1645974	acety…
ENSG0…	31	ACACA	17	37084992	37406836	acety…
ENSG0…	35	ACADS	12	12072…	12074…	acyl-…
ENSG0…	1636	ACE	17	63477061	63498380	angio…
ENSG0…	43	ACHE	7	10088…	10089…	acety…
ENSG0…	87	ACTN1	14	68874128	68979440	actin…
ENSG0…	81	ACTN4	HG26_…	57285	141225	actin…
ENSG0…	81	ACTN4	19	38647649	38731589	actin…
ENSG0…	100	ADA	20	44584896	44652252	adeno…
ENSG0…	103	ADAR	1	15458…	15462…	adeno…
…	…	…	…	…	…	…

7.3.7 数据整备

在以上分析中，我们得到了一系列的基因数据集结果。

Figure 13为图 intersects all used gene sets 概览。

(对应文件为 `Figure+Table/intersects-all-used-gene-sets.pdf`)

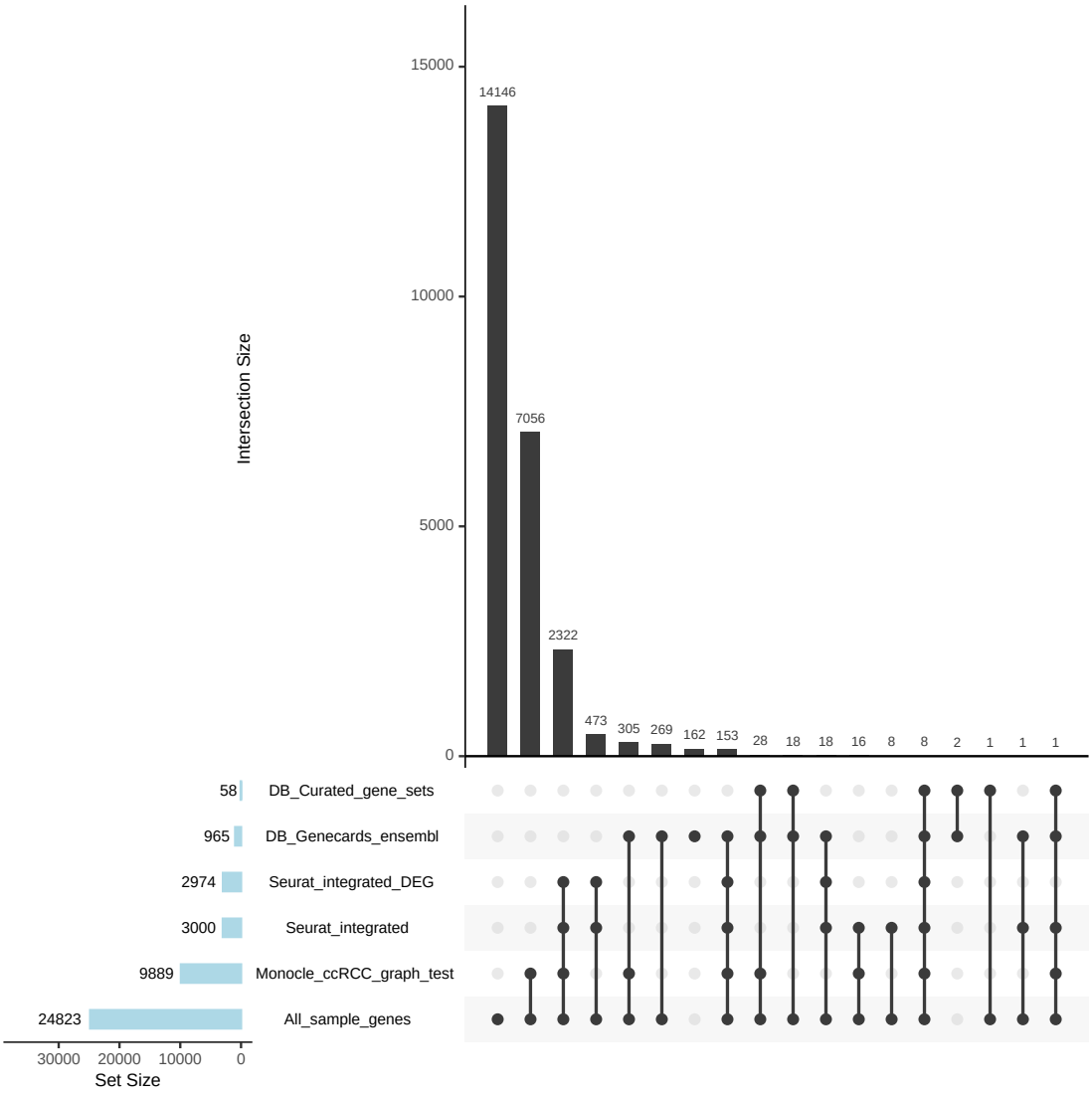


Figure 13: Intersects all used gene sets

7.3.8 加权基因共表达 WGCNA

7.3.8.1 计算基因共表达模块 使用 Fig. 13 中的 `DB_Genecards_ensembl` 基因集的基因，过滤单细胞的数据（过滤 Seurat 对象）用以 WGCNA 分析。

Figure 14为图 selection of soft threshold 概览。选择阈值为 2。

(对应文件为 `Figure+Table/selection-of-soft-threshold.pdf`)

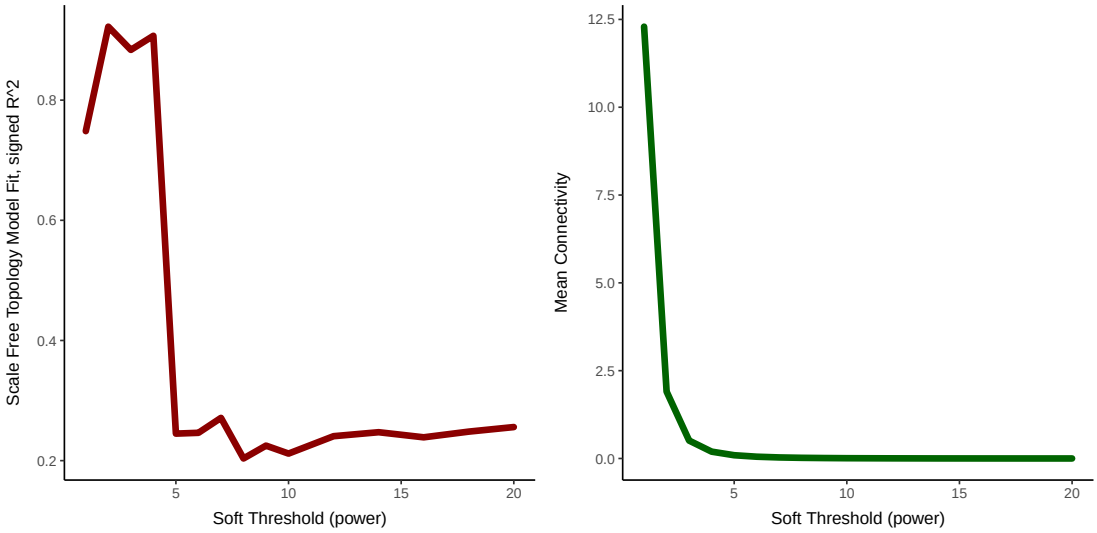


Figure 14: Selection of soft threshold

建立基因共表达模块。

Figure 15为图 gene modules 概览。

(对应文件为 `Figure+Table/gene-modules.pdf`)

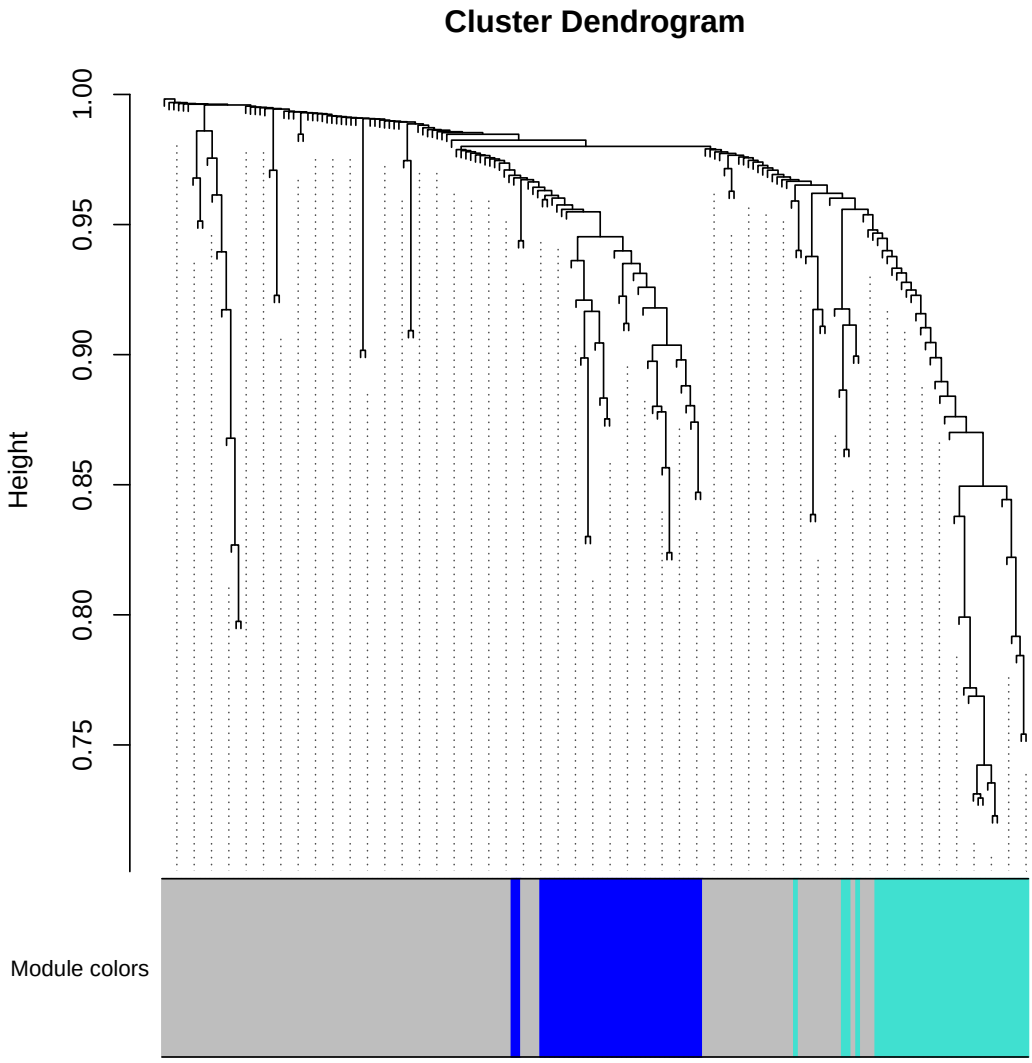


Figure 15: Gene modules

7.3.8.2 计算 CRDscore 为了对共表达的基因进行加权评估，需要 'trait' 数据。计算细胞的 CRDscore[1] 作为拟 trait 数据。该计算以 Fig. 13 中的 `DB_Curated_gene_sets` 为基准。

Figure 16为图 CRDscore distribution 概览。计算的 CRDscore 的数据分布。

(对应文件为 `Figure+Table/CRDscore-distribution.pdf`)

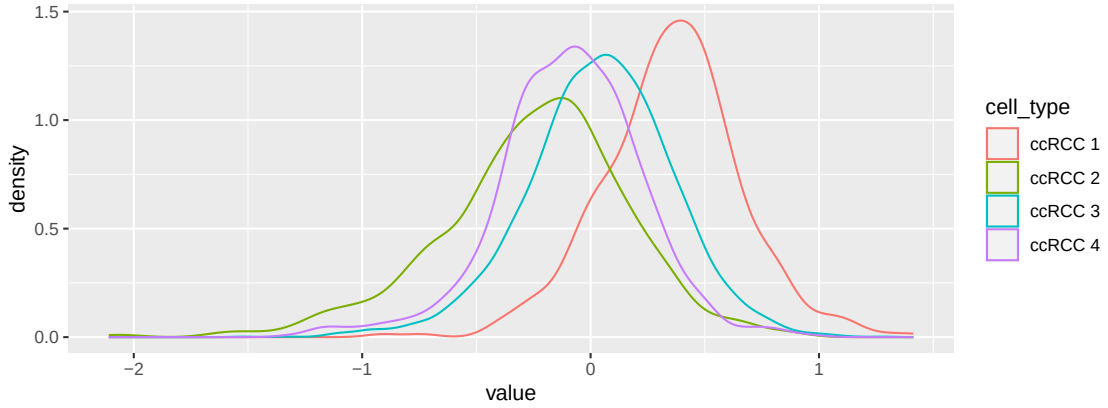


Figure 16: CRDscore distribution

7.3.8.3 筛选高权重基因

根据 CRDscore 对基因加权分析。

Table 6为表格 module membership 概览。

(对应文件为 `Figure+Table/module-membership.csv`)

> 注：表格共有 502 行 7 列，以下预览的表格可能省略部分数据；表格含有 181 个唯一 'gene'。

Table 6: Module membership

gene	module	cor	pvalue	-log2...	signi...	sign
ABCB1	ME0	-0.03	0.0142	6.137...	< 0.05	*
ABHD5	ME0	-0.04	0.003	8.380...	< 0.05	*
ACE	ME0	0.17	0	Inf	< 0.001	**
ACTN1	ME0	-0.06	0	Inf	< 0.001	**
ACTN4	ME0	0.07	0	Inf	< 0.001	**
ADA	ME0	0.29	0	Inf	< 0.001	**
AGT	ME0	-0.35	0	Inf	< 0.001	**
AHNAK	ME0	-0.04	0.0032	8.287...	< 0.05	*
AHR	ME0	0.07	0	Inf	< 0.001	**
APMAP	ME0	0.19	0	Inf	< 0.001	**
APOE	ME0	0.12	0	Inf	< 0.001	**
APP	ME0	-0.14	0	Inf	< 0.001	**
ARL4A	ME0	0.06	0	Inf	< 0.001	**
ASS1	ME0	-0.03	0.0138	6.179...	< 0.05	*
ATM	ME0	0.18	0	Inf	< 0.001	**
...	...	...	...	...	...	...

Table 7为表格 gene significant 概览。

(对应文件为 `Figure+Table/gene-significant.csv`)

> 注：表格共有 155 行 8 列，以下预览的表格可能省略部分数据；表格含有 155 个唯一 'gene'。

Table 7: Gene significant

gene	trait	cor	pvalue	-log2...	signi...	sign	p.adjust
ABHD5	CRDscore	-0.09	0	Inf	< 0.001	**	0
ACE	CRDscore	-0.07	0	Inf	< 0.001	**	0
ACTN1	CRDscore	-0.03	0.0124	6.333...	< 0.05	*	0.015...
ADA	CRDscore	-0.14	0	Inf	< 0.001	**	0
AGT	CRDscore	0.19	0	Inf	< 0.001	**	0
APMAP	CRDscore	-0.06	0	Inf	< 0.001	**	0
APOE	CRDscore	-0.06	0	Inf	< 0.001	**	0
APP	CRDscore	0.11	0	Inf	< 0.001	**	0
ARL4A	CRDscore	-0.07	0	Inf	< 0.001	**	0
ATM	CRDscore	-0.06	0	Inf	< 0.001	**	0
B2M	CRDscore	0.22	0	Inf	< 0.001	**	0
BHLHE40	CRDscore	-0.26	0	Inf	< 0.001	**	0
BHLHE41	CRDscore	0.04	0.0094	6.733...	< 0.05	*	0.011...
BTG1	CRDscore	-0.15	0	Inf	< 0.001	**	0
CAVIN3	CRDscore	0.03	0.0274	5.189...	< 0.05	*	0.032...
...	...	...	...	...	...	...	...

取 Tab. 6 和 Tab. 7 的交集。

Figure 17为图 intersect of gene significant and module membership 概览。

(对应文件为 `Figure+Table/intersect-of-gene-significant-and-module-membership.pdf`)

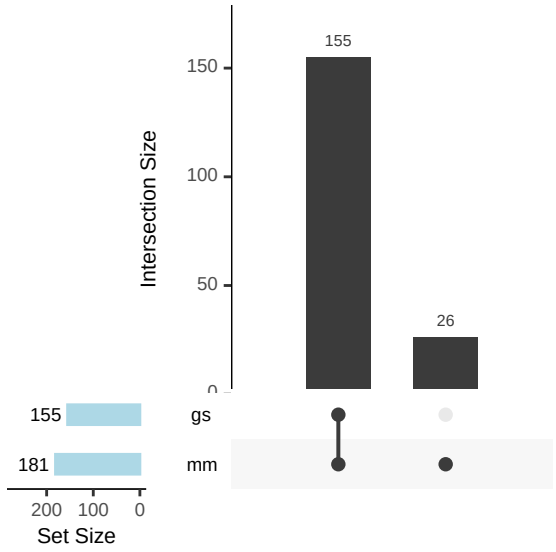


Figure 17: Intersect of gene significant and module membership

7.3.9 通路富集分析 Clusterprofiler

取 Fig. 17 的交集基因，做通路富集分析。

Figure 18为图 kegg enrich 概览。

(对应文件为 `Figure+Table/kegg-enrich.pdf`)

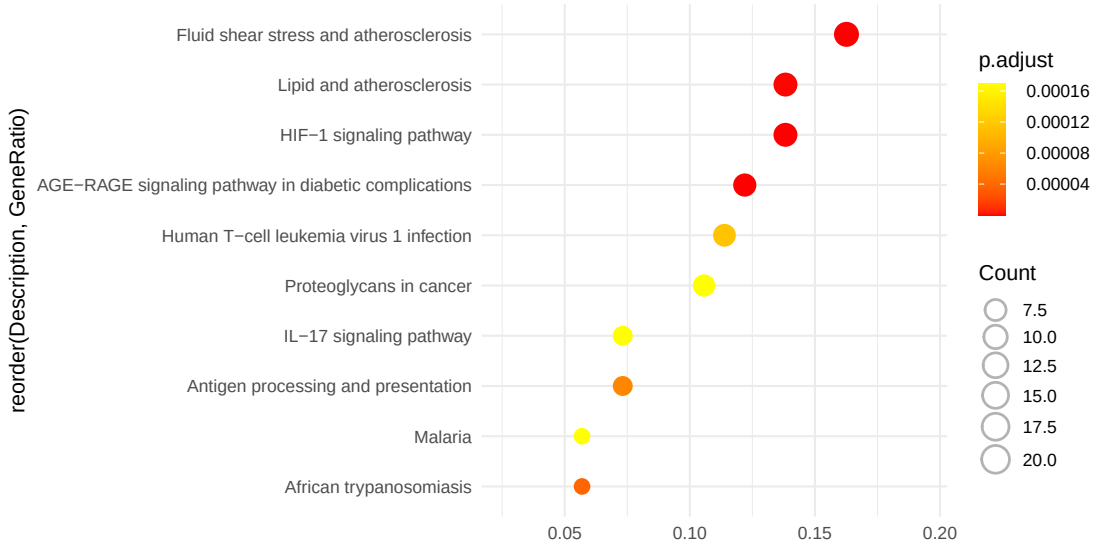


Figure 18: Kegg enrich

KEGG 富集结果显示，基因集与肿瘤（proteoglycans in cancer）相关，同时，与抗原抗体反应相关。

Figure 19为图 go enrich 概览。

(对应文件为 `Figure+Table/go-enrich.pdf`)

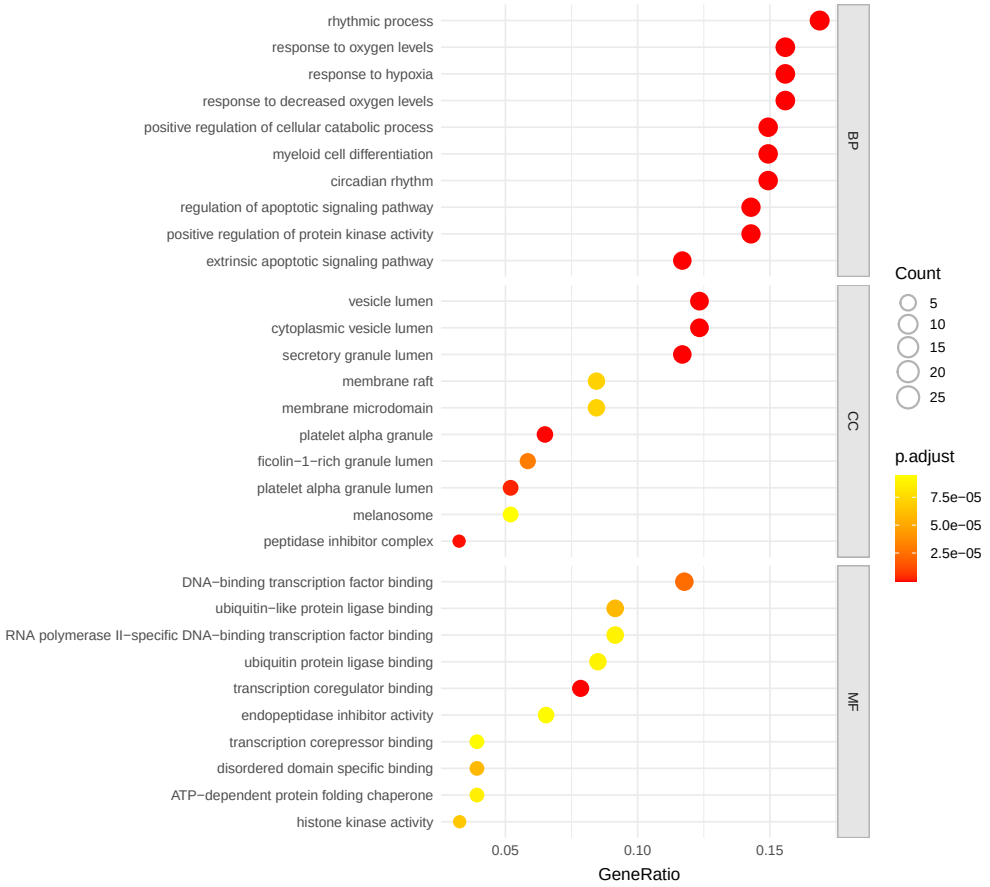


Figure 19: Go enrich

(BP, MF, CC: Biological Process, Molecular Function, and Cellular Component groups)

GO 富集结果显示，这些基因与节律过程（rhythmic process），昼夜节律（circadian rhythm）相关，并且和缺氧、氧化水平等与肿瘤微环境的通路相关联。

这些结果验证了昼夜节律对于 ccRCC 发展的影响，以及预后诊断以及治疗的可能。

7.3.10 Feature selection 和构建预后模型 (LASSO)

https://portal.gdc.cancer.gov/

7.3.10.1 TCGA 数据获取 以 `TCGAbiolinks` 下载 TCGA (RNA) 数据（KIRC）[18]。https://www.bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html

Table 8为表格 downloaded RNA seq data of ccRCC in TCGA 概览。

(对应文件为 `Figure+Table/downloaded-RNA-seq-data-of-ccRCC-in-TCGA.csv`)

注：表格共有 533 行 29 列，以下预览的表格可能省略部分数据；表格含有 533 个唯一 'id'。

Table 8: Downloaded RNA seq data of ccRCC in TCGA

id	data_……2	cases	access	file_……5	submi…	data_……7	type	file_……9	creat…	…
21702…	TSV	TCGA-…	open	c3b9c…	1907e…	Trans…	gene_…	4228535	2021-…	…
44dad…	TSV	TCGA-…	open	e4497…	d70f0…	Trans…	gene_…	4231641	2021-…	…
c1215…	TSV	TCGA-…	open	55599…	bbbda…	Trans…	gene_…	4229992	2021-…	…
b6052…	TSV	TCGA-…	open	78855…	96835…	Trans…	gene_…	4238896	2021-…	…
7b34b…	TSV	TCGA-…	open	59857…	65612…	Trans…	gene_…	4244120	2021-…	…
4845a…	TSV	TCGA-…	open	49c74…	52cd1…	Trans…	gene_…	4247550	2021-…	…
41cc6…	TSV	TCGA-…	open	87ac9…	ec8a9…	Trans…	gene_…	4242354	2021-…	…
51a84…	TSV	TCGA-…	open	6b83b…	391f3…	Trans…	gene_…	4229064	2021-…	…
2b5b6…	TSV	TCGA-…	open	0416b…	3199c…	Trans…	gene_…	4253460	2021-…	…
afe81…	TSV	TCGA-…	open	bbae6…	70241…	Trans…	gene_…	4249550	2021-…	…
99ad8…	TSV	TCGA-…	open	6a8e0…	48dfe…	Trans…	gene_…	4251551	2021-…	…
af5c0…	TSV	TCGA-…	open	4c1f1…	79a53…	Trans…	gene_…	4239264	2021-…	…
8ebdb…	TSV	TCGA-…	open	c045b…	736cd…	Trans…	gene_…	4243590	2021-…	…
5226a…	TSV	TCGA-…	open	fc526…	d0d1c…	Trans…	gene_…	4246857	2021-…	…
131cc…	TSV	TCGA-…	open	af4b5…	69aee…	Trans…	gene_…	4263556	2021-…	…
…	…	…	…	…	…	…	…	…	…	…

7.3.10.2 预处理 TCGA RNA 数据

使用 `edgeR` 预处理 RNA 数据。

- 过滤低表达的基因。
- 表达量归一化。

Figure 20为图 tcga rna data filter 概览。

(对应文件为 `Figure+Table/tcga-rna-data-filter.pdf`)

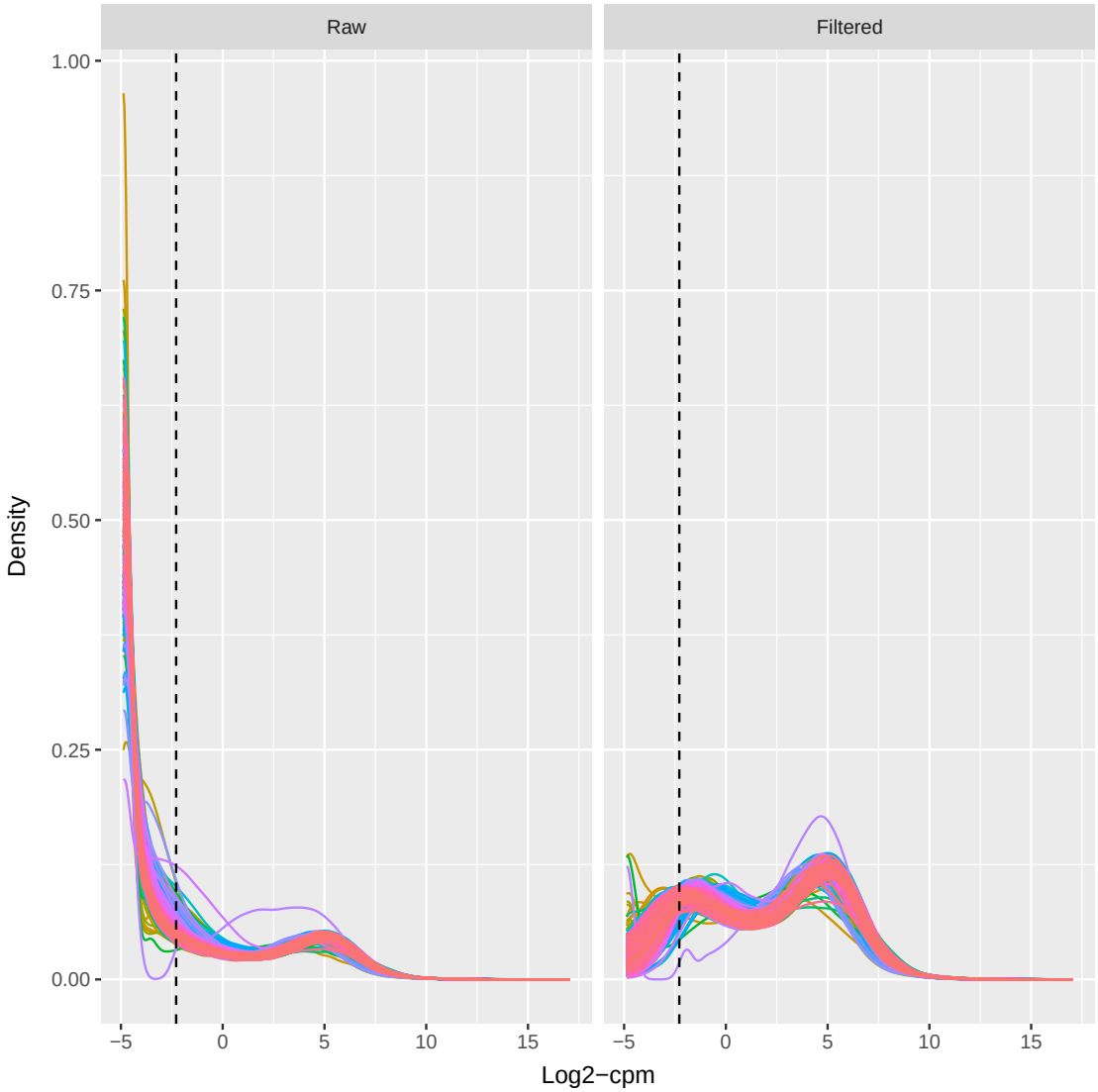


Figure 20: Tcga rna data filter

Figure 21为图 tcga rna data normalize 概览。

(对应文件为 `Figure+Table/tcga-rna-data-normalize.pdf`)

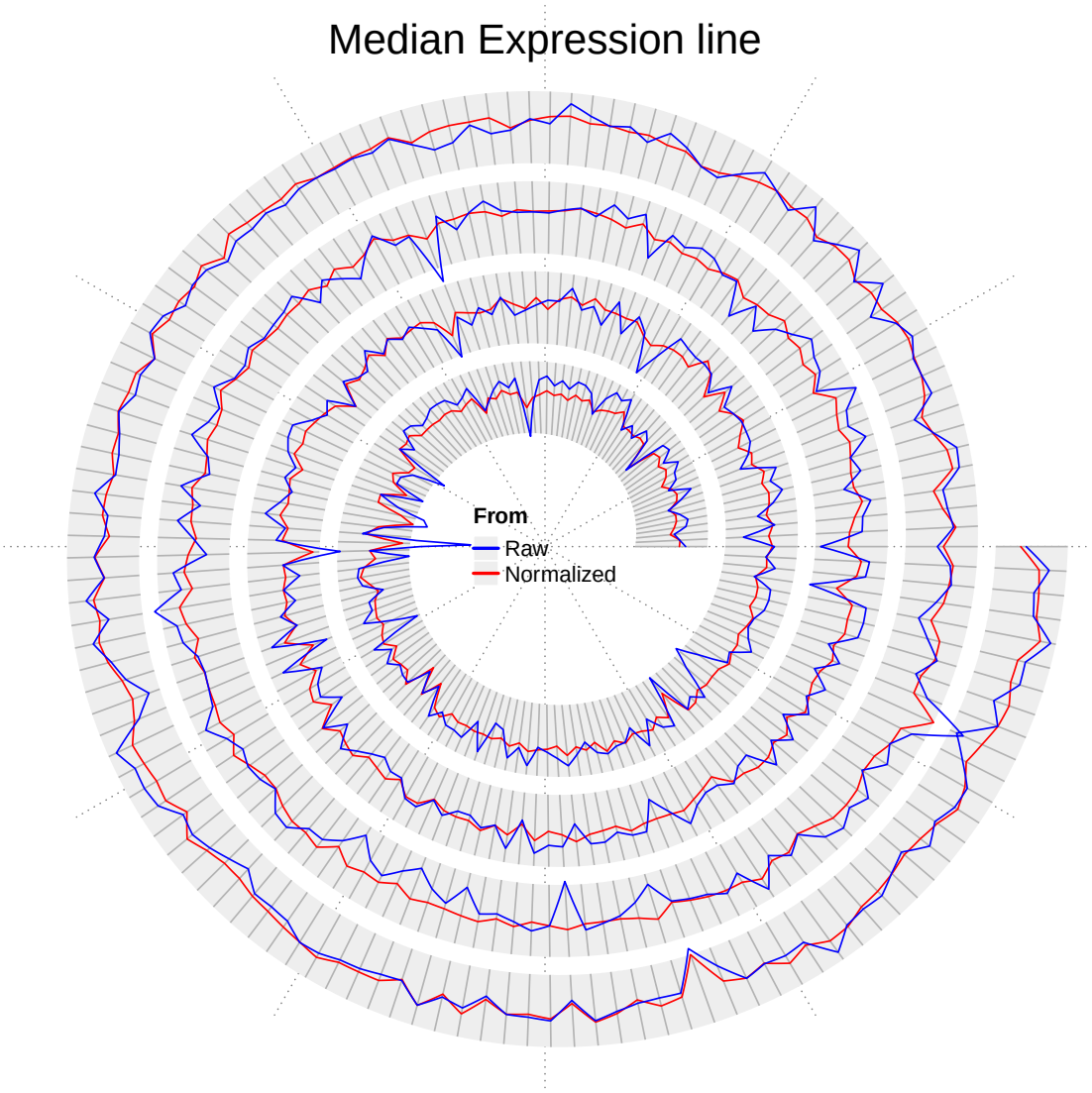


Figure 21: Tcga rna data normalize

7.3.10.3 EFS 并随机分成两组数据（4:1），以备训练和验证。

在使用 LASSO 回归之前，这里预先对基因集进行了过滤。

- 以 Fig. 17 (7.3.8.3) 中的交集的基因过滤 TCGA KIRC （即，ccRCC）的表达数据
- 使用 EFS[14] 处理训练数据集，筛选 top30 features。

Figure 22为图 random split the datasets 概览。

(对应文件为 `Figure+Table/random-split-the-datasets.pdf`)

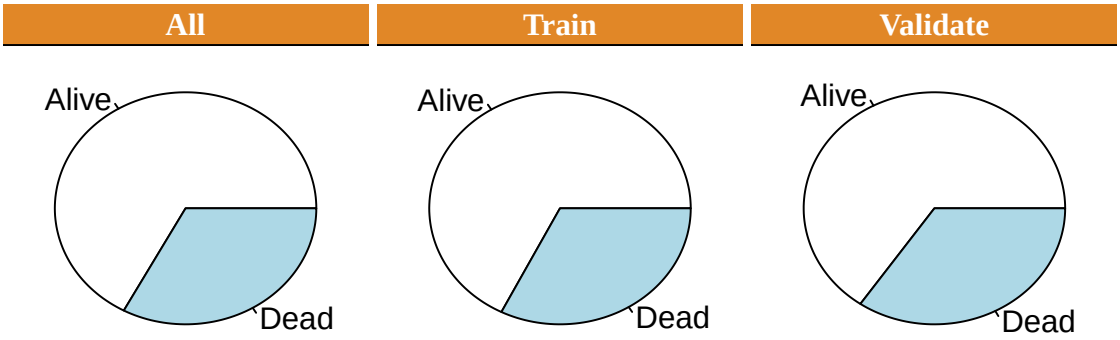


Figure 22: Random split the datasets

Figure 23为图 EFS top30 genes 概览。

(对应文件为 `Figure+Table/EFS-top30-genes.pdf`)

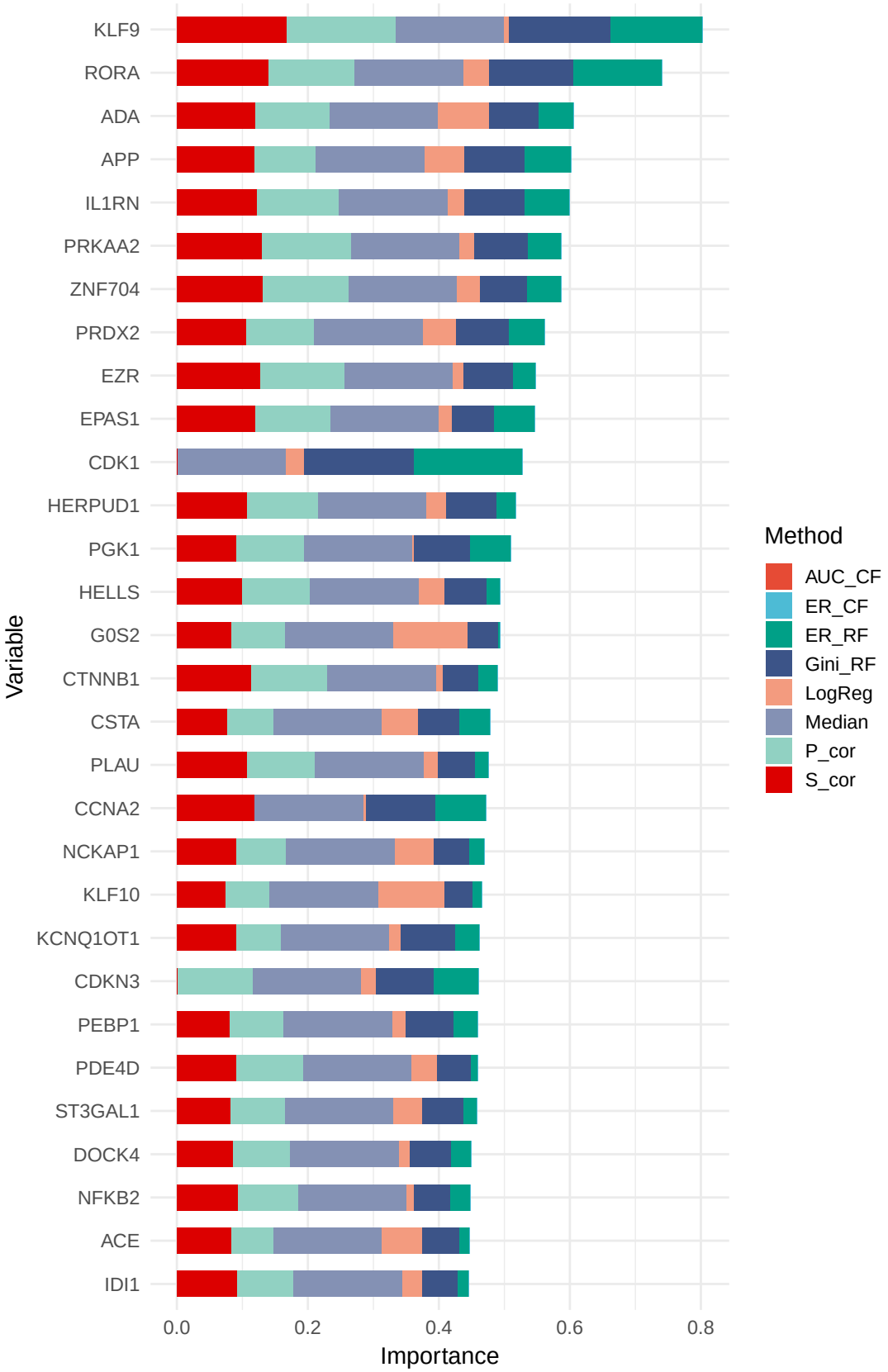

KLF9
RORA
ADA
APP
IL1RN
PRKAA2
ZNF704
PRDX2
EZR
EPAS1
CDK1
HERPUD1
PGK1
HELLS
G0S2
CTNNB1
CSTA
PLAU
CCNA2
NCKAP1
KLF10
KCNQ1OT1
CDKN3
PEBP1
PDE4D
ST3GAL1
DOCK4
NFKB2
ACE
IDI1
Variable
0.0
0.2
0.4
0.6
0.8
Importance
Method
AUC_CF
ER_CF
ER_RF
Gini_RF
LogReg
Median
P_cor
S_cor

Figure 23: EFS top30 genes

确认 top30 基因是否属于差异表达基因。

Figure 24为图 Top30 genes intersect with DEG 概览。其中 25 个属于差异表达基因。

(对应文件为 `Figure+Table/Top30-genes-intersect-with-DEG.pdf`)

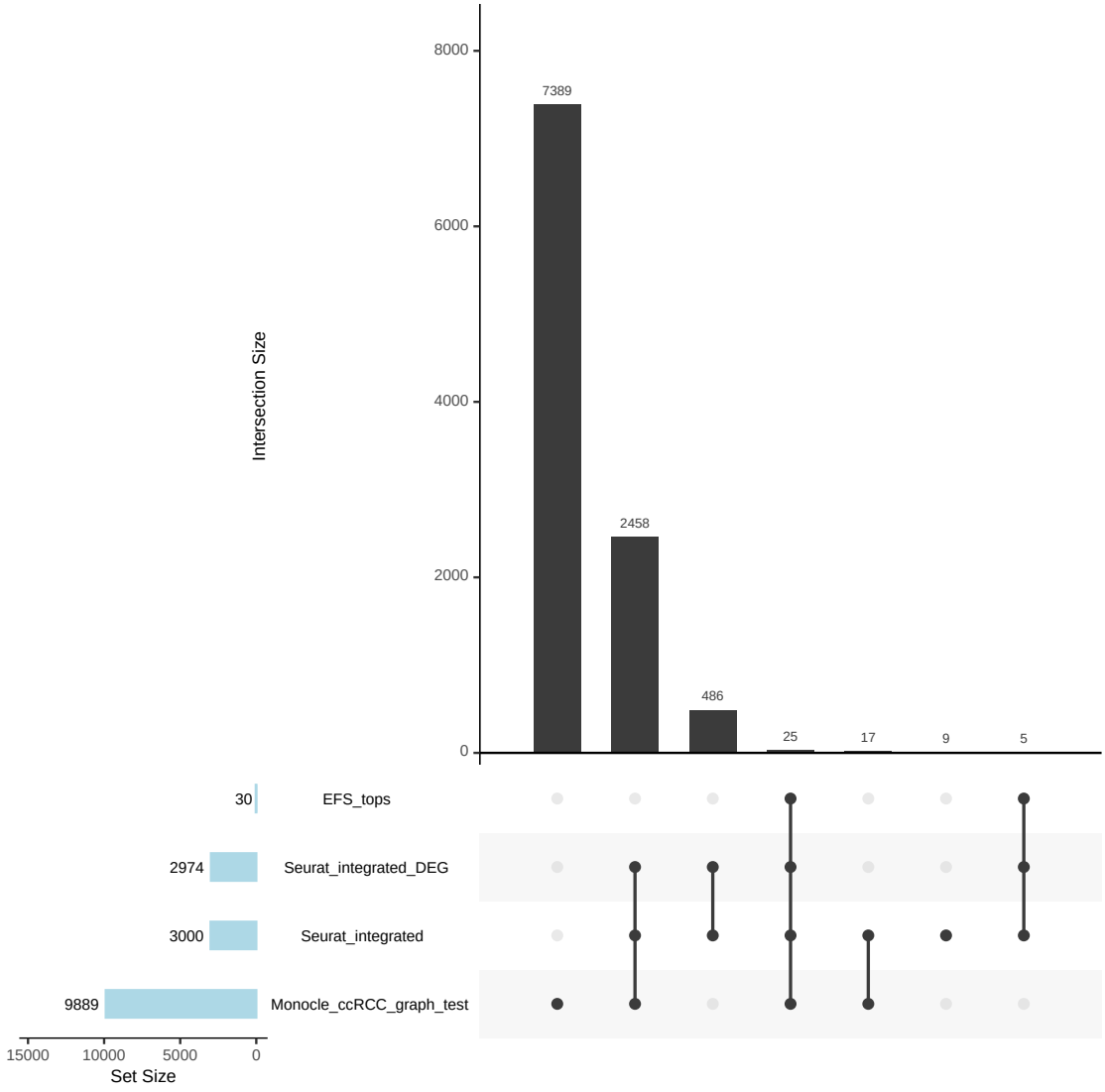


Figure 24: Top30 genes intersect with DEG

7.3.10.4 LASSO 回归 以训练数据集的 Top30 基因进行 LASSO 回归。以十倍交叉验证选择最优 Lambda 值。

Figure 25为图 LASSO model 概览。

(对应文件为 `Figure+Table/LASSO-model.pdf`)

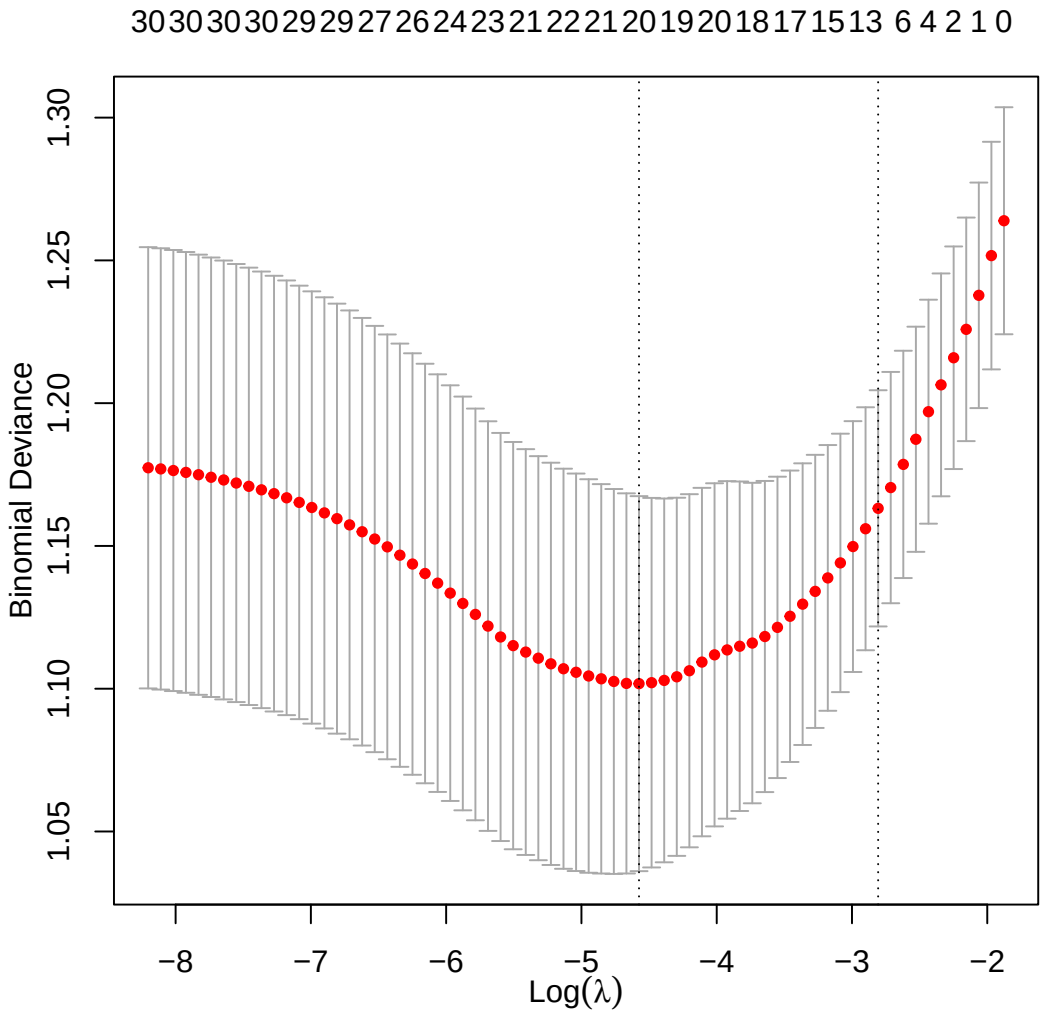


Figure 25: LASSO model

Figure 26为图 LASSO coefficents 概览。

(对应文件为 `Figure+Table/LASSO-coefficents.pdf`)

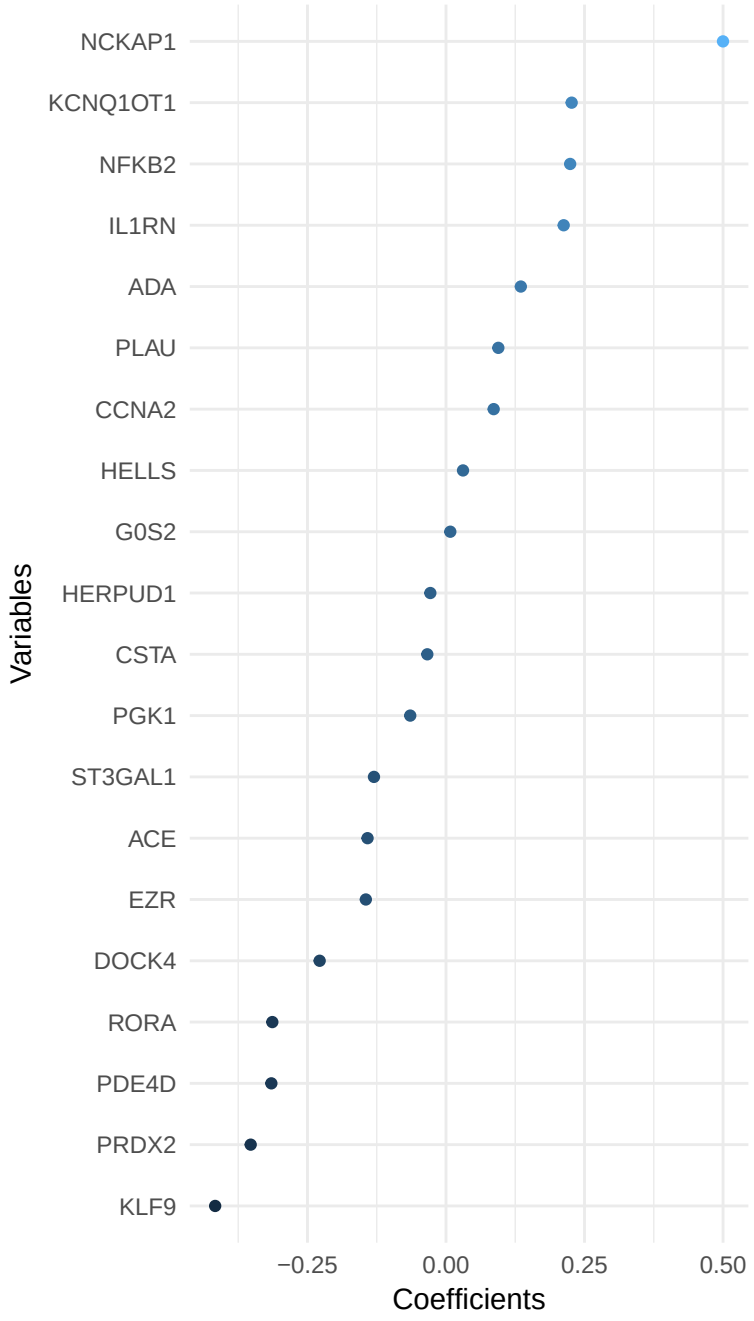


Figure 26: LASSO coefficents

ROC 结果显示，AUC 为 0.69，高于随机分类器，Fig. 26 所示基因对于预后具有一定的诊断参考价值。Figure 27为图 LASSO ROC 概览。

(对应文件为 `Figure+Table/LASSO-ROC.pdf`)

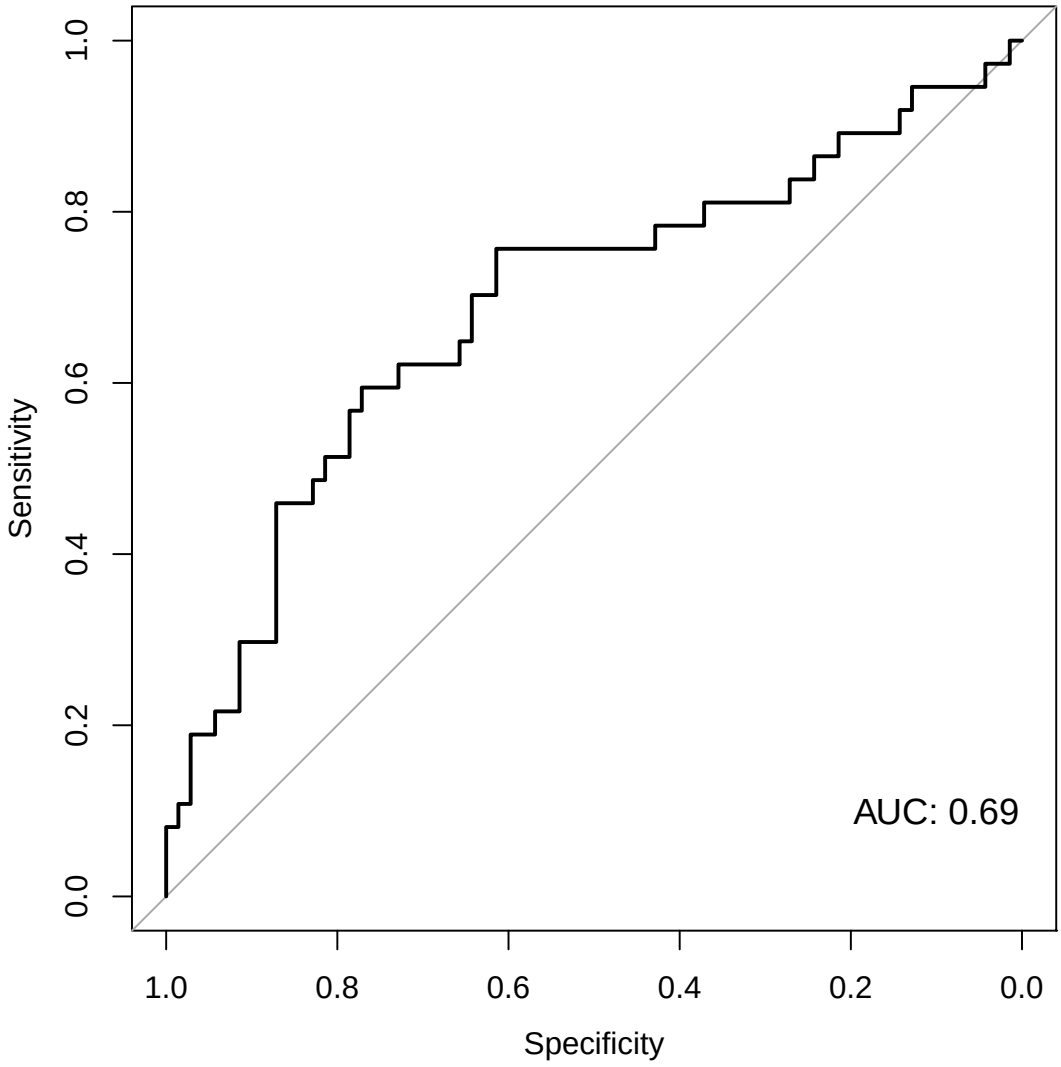


Figure 27: LASSO ROC

7.3.11 拟时序分析基因转归 Monocle3

7.3.11.1 Top 30 in pseudotime 对 Top 30 的基因在拟时序分析中追踪其表达量变化。

Figure 28为图 top 30 genes in pseudotime part 1 概览。

(对应文件为 `Figure+Table/top-30-genes-in-pseudotime-part-1.pdf`)

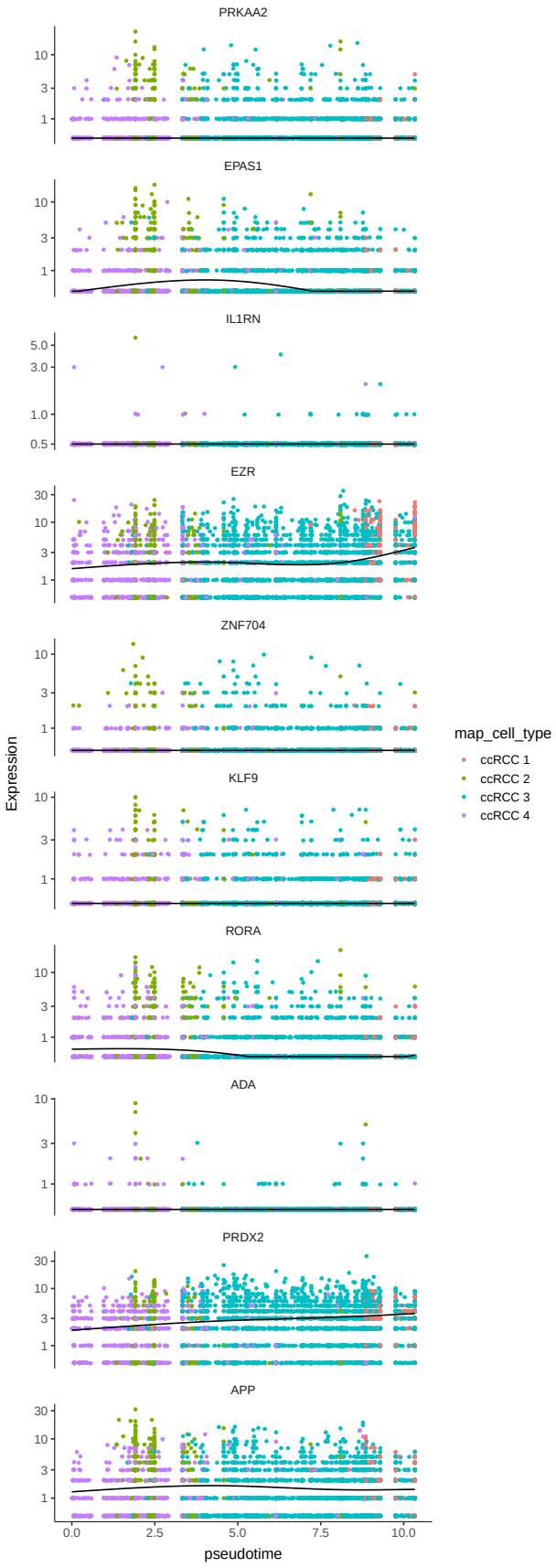

PRKAA2
EPAS1
IL1RN
EZR
ZNF704
KLF9
RORA
ADA
PRDX2
APP
Expression
pseudotime
0.0
2.5
5.0
7.5
10.0
map_cell_type
ccRCC 1
ccRCC 2
ccRCC 3
ccRCC 4

Figure 28: Top 30 genes in pseudotime part 1

Figure 29为图 top 30 genes in pseudotime part 2 概览。

(对应文件为 `Figure+Table/top-30-genes-in-pseudotime-part-2.pdf`)

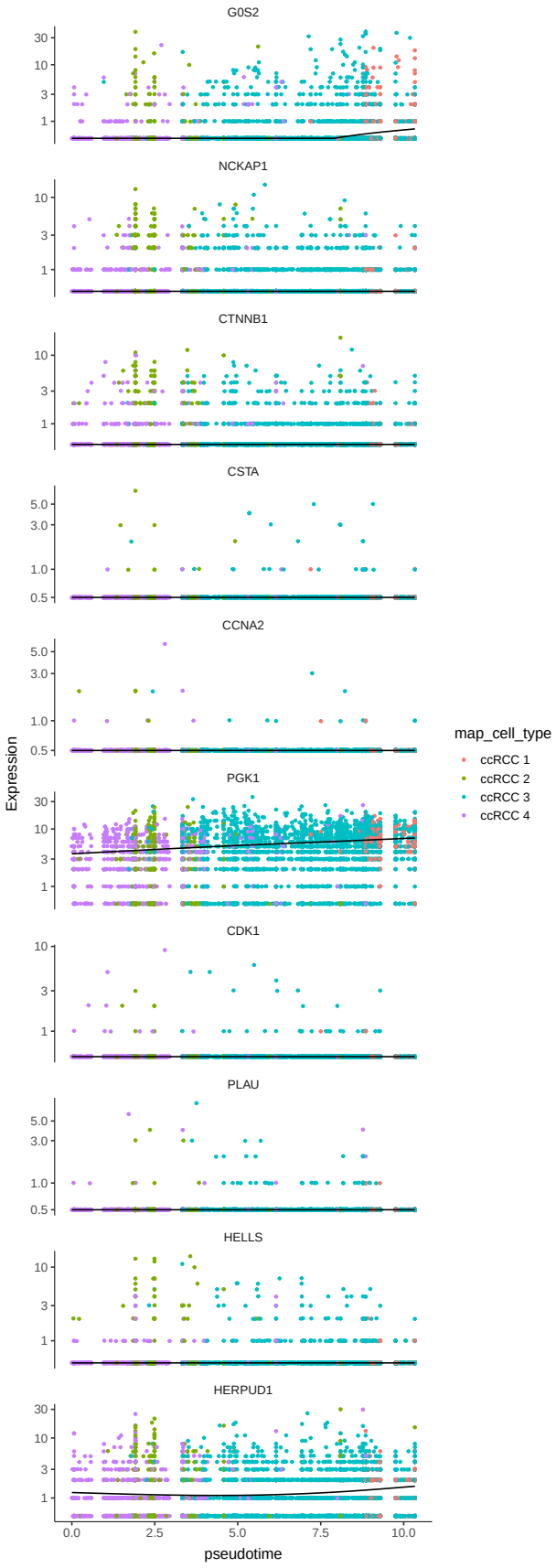

G0S2
NCKAP1
CTNNB1
CSTA
CCNA2
PGK1
CDK1
PLAU
HELLS
HERPUD1
Expression
pseudotime
map_cell_type
ccRCC 1
ccRCC 2
ccRCC 3
ccRCC 4

Figure 29: Top 30 genes in pseudotime part 2

Figure 30为图 top 30 genes in pseudotime part3 概览。

(对应文件为 `Figure+Table/top-30-genes-in-pseudotime-part3.pdf`)

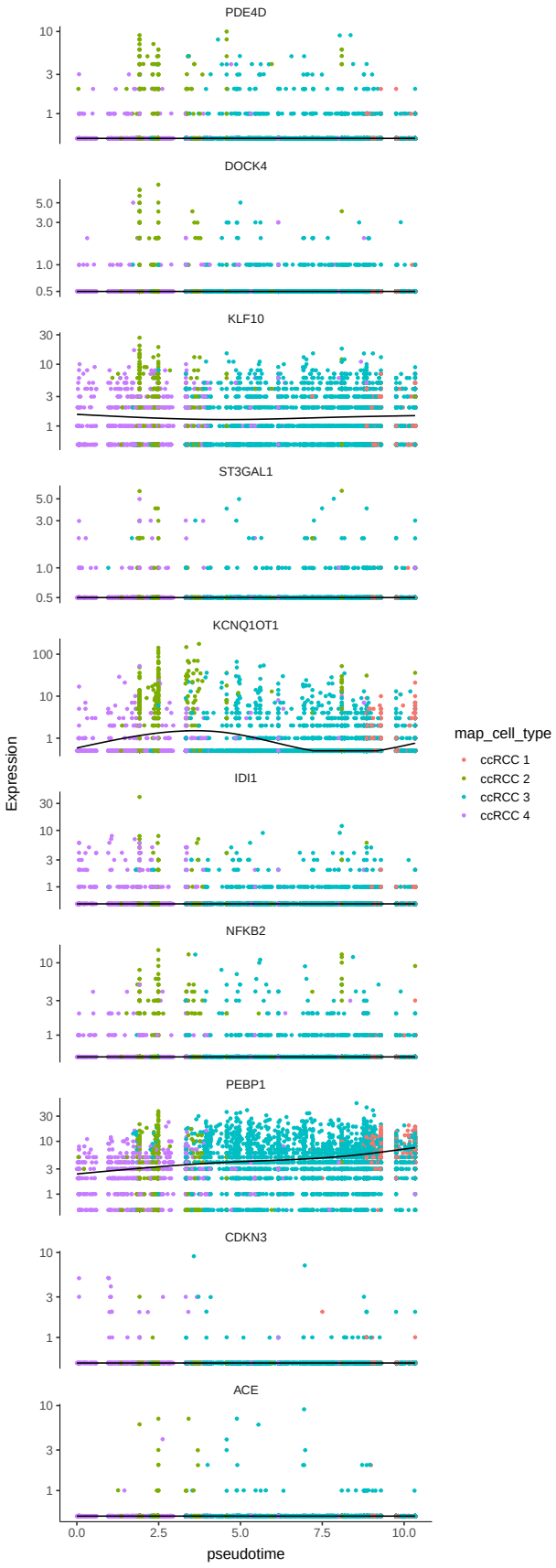

PDE4D
DOCK4
KLF10
ST3GAL1
KCNQ1OT1
IDI1
NFKB2
PEBP1
CDKN3
ACE
Expression
pseudotime
0.0
2.5
5.0
7.5
10.0
map_cell_type
ccRCC 1
ccRCC 2
ccRCC 3
ccRCC 4

Figure 30: Top 30 genes in pseudotime part3

7.3.11.2 Specific genes in Pseudotime 在 7.3.11.1 中发现，有三个基因表现出明显的曲线变化趋势。

- PGK1，呈上升曲线
- KCNQ1OT1，呈 S 变化曲线
- PEBP1，呈上升曲线

Figure 31为图 Specific genes in pseudotime 概览。

(对应文件为 `Figure+Table/Specific-genes-in-pseudotime.pdf`)

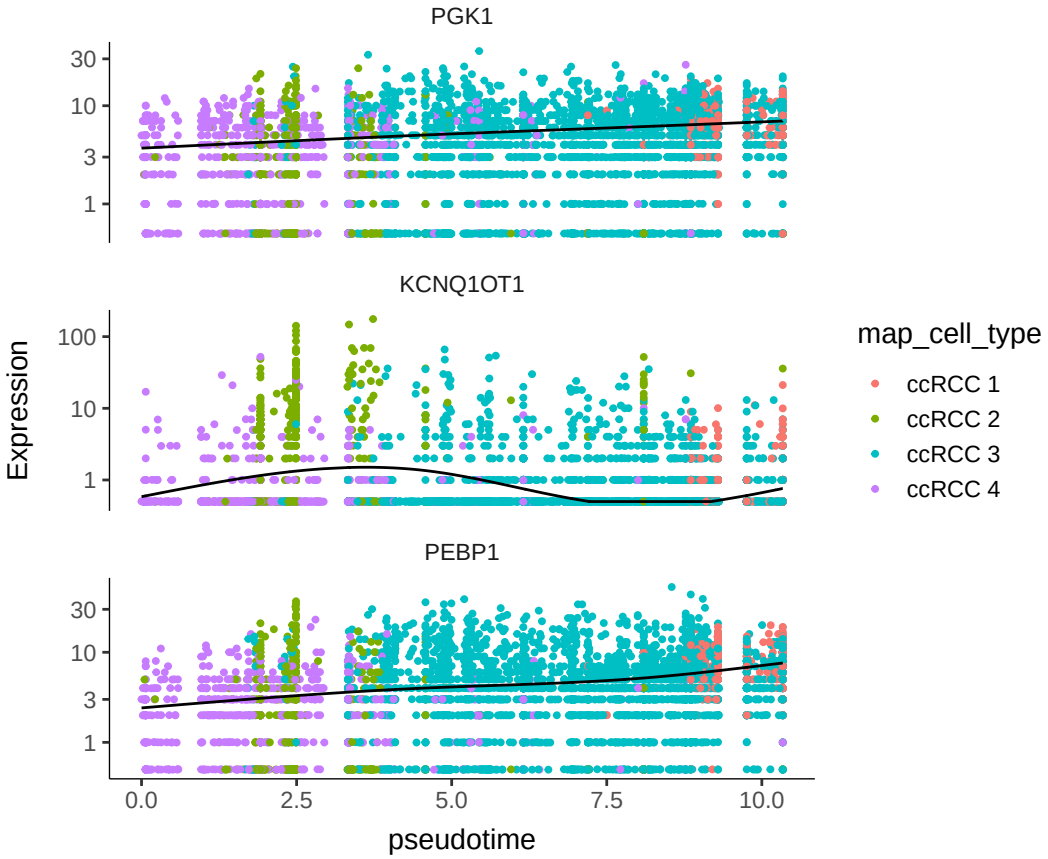


Figure 31: Specific genes in pseudotime

7.3.12 生存分析 Survival

为了进一步验证上述结果（7.3.11.2），以 TCGA 的数据做生存分析。结果显示，三者对于 ccRCC 的风险评估均有显著性意义。

Figure 32为图 Survival analysis of PGK1 概览。

(对应文件为 `Figure+Table/Survival-analysis-of-PGK1.pdf`)

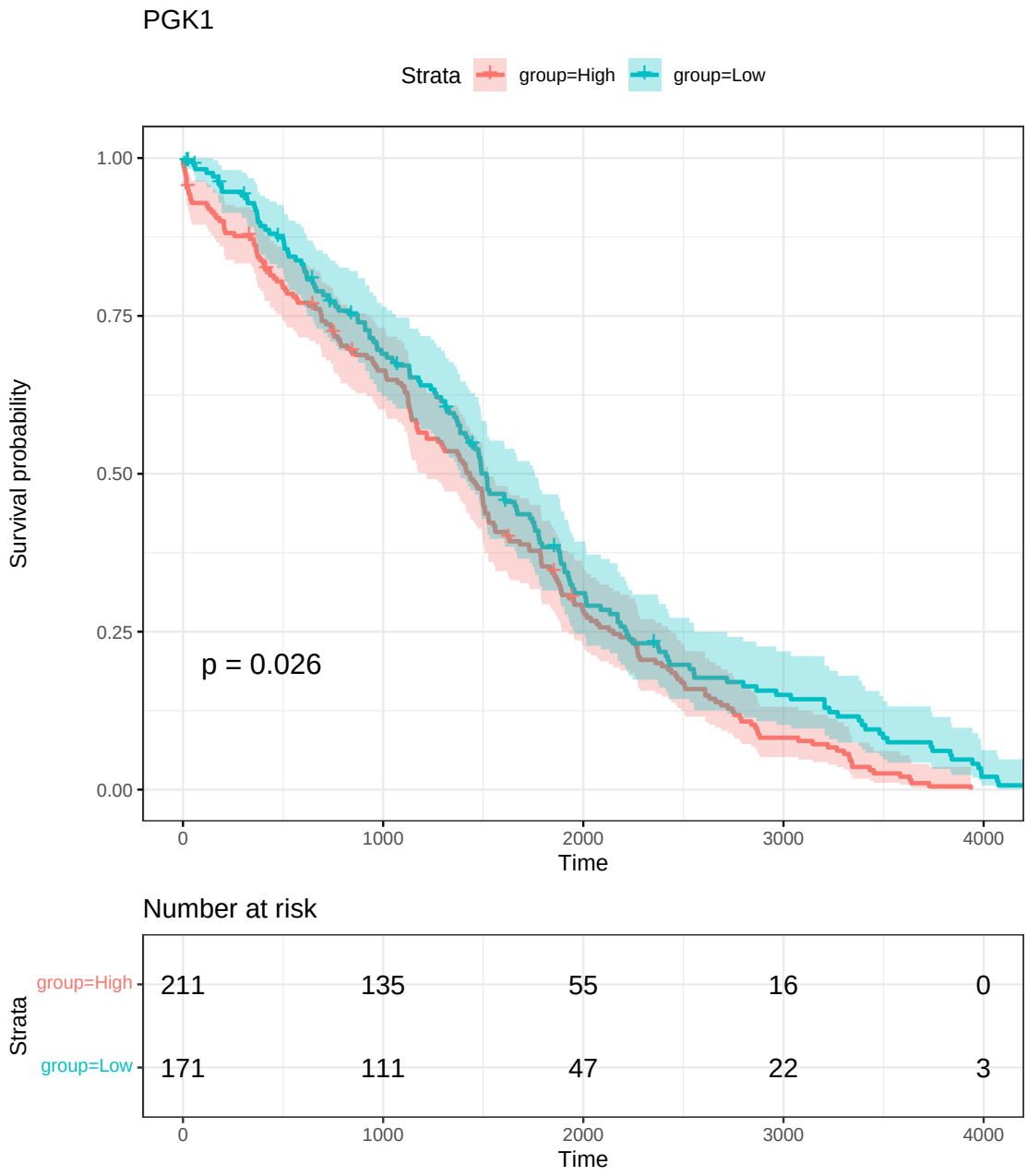


Figure 32: Survival analysis of PGK1

Figure 33为图 Survival analysis of KCNQ1OT1 概览。

(对应文件为 `Figure+Table/Survival-analysis-of-KCNQ1OT1.pdf`)

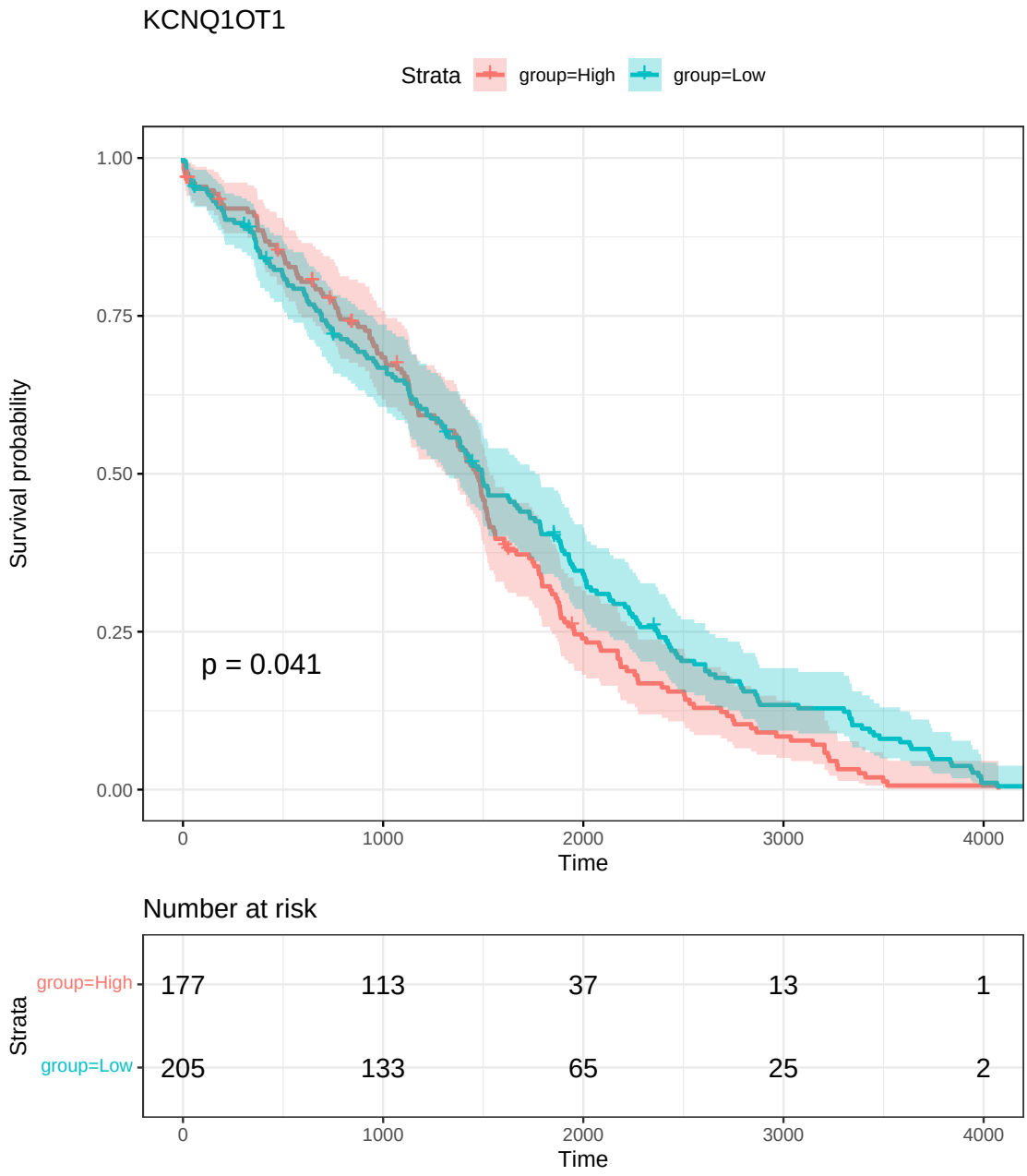


Figure 33: Survival analysis of KCNQ1OT1

Figure 34为图 Survival analysis of PEBP1 概览。

(对应文件为 `Figure+Table/Survival-analysis-of-PEBP1.pdf`)

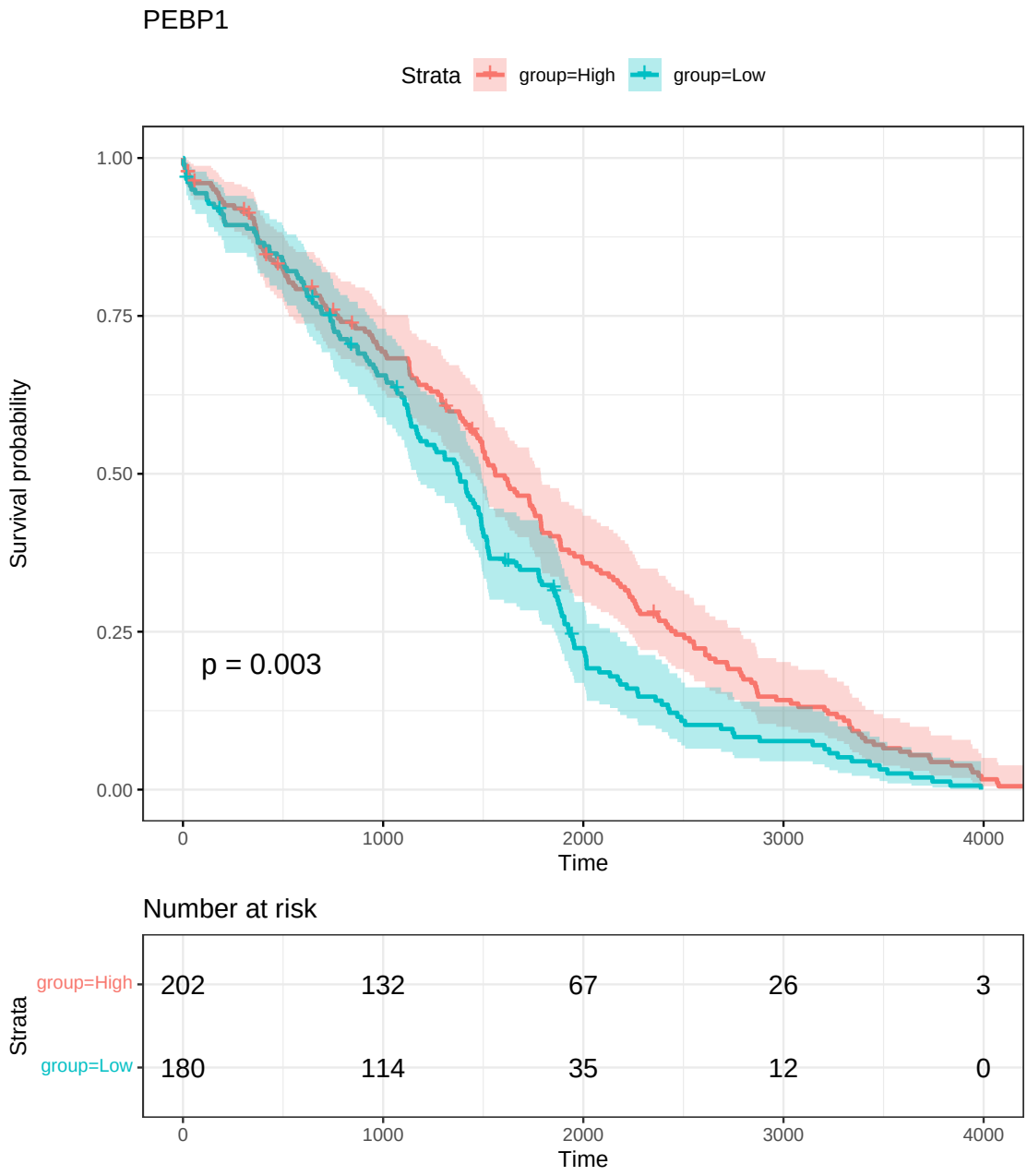


Figure 34: Survival analysis of PEBP1

7.3.13 细胞通讯 CellChat

使用 `DB_Genecards_ensembl` (Fig. 13) 基因集的基因进行细胞通讯分析。该基因集来源于 Genecards，暗示了这些基因与昼夜节律之间的关联性。

Figure 35为图 used database for cell communication 概览。

(对应文件为 `Figure+Table/used-database-for-cell-communication.pdf`)

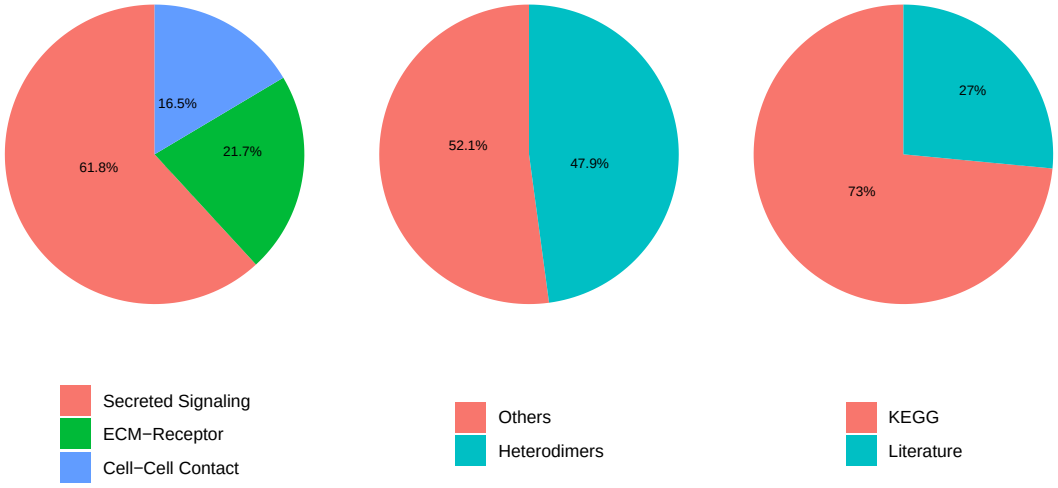


Figure 35: Used database for cell communication

Figure 36为图 cell communication count 概览。

(对应文件为 `Figure+Table/cell-communication-count.pdf`)

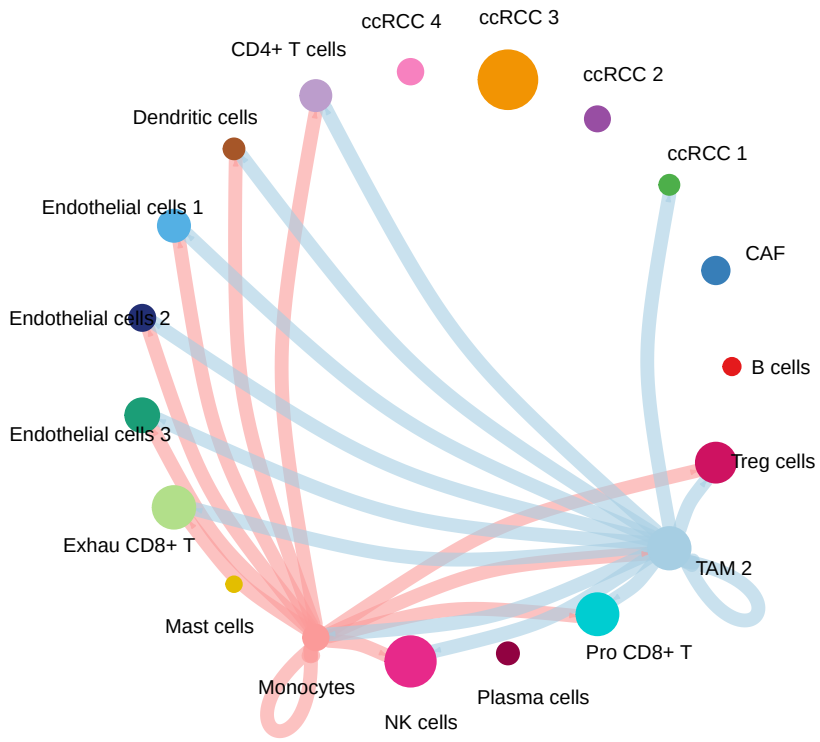


Figure 36: Cell communication count

Figure 37为图 cell communication heatmap of TNF signiling 概览。

(对应文件为 `Figure+Table/cell-communication-heatmap-of-TNF-signiling.pdf`)

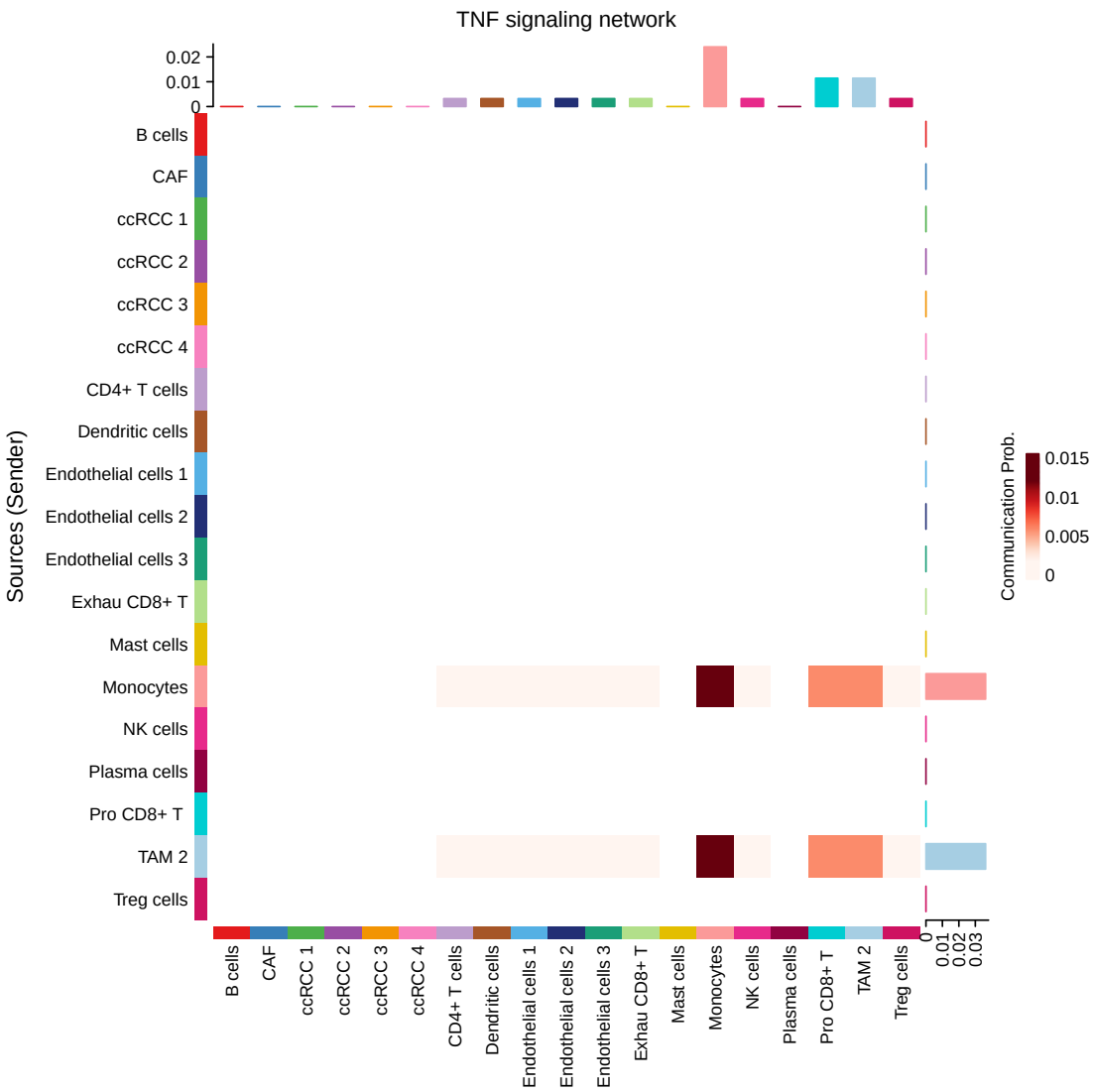


Figure 37: Cell communication heatmap of TNF signiling

Figure 38为图 gene expression of TNF signiling 概览。

(对应文件为 `Figure+Table/gene-expression-of-TNF-signiling.pdf`)

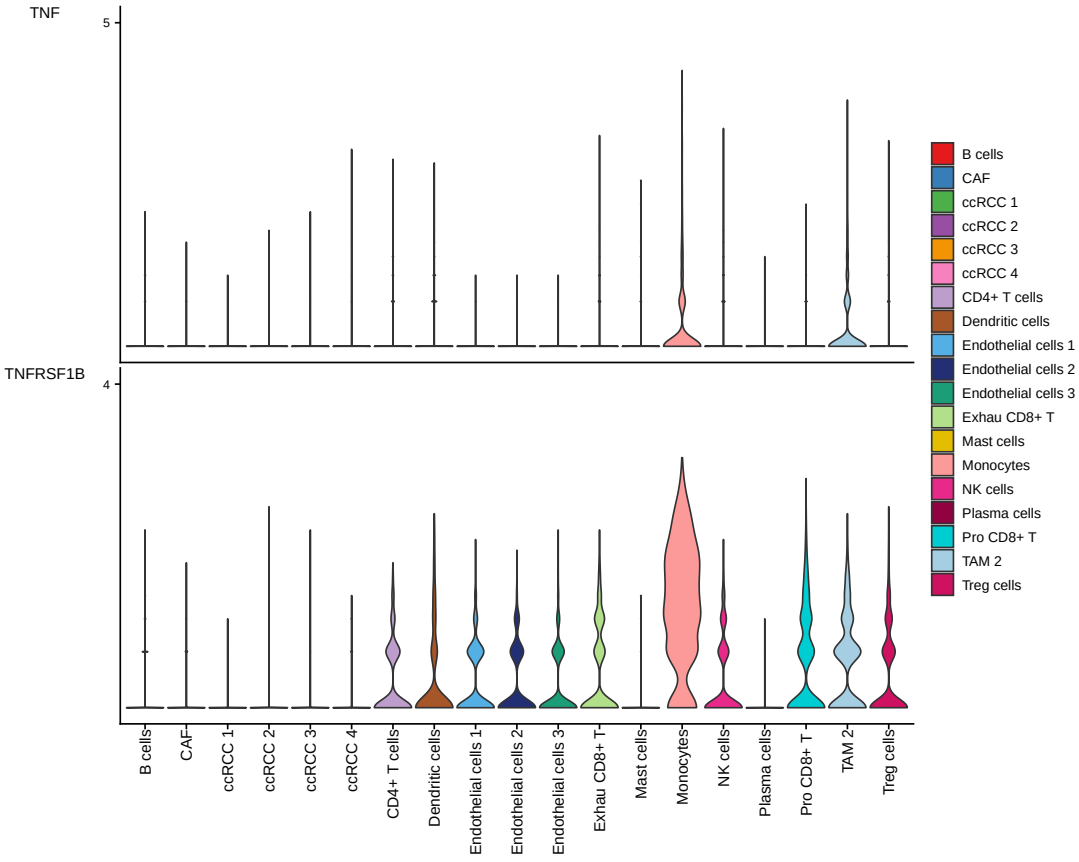


Figure 38: Gene expression of TNF signiling

Figure 39为图 role of TNF signiling 概览。

(对应文件为 `Figure+Table/role-of-TNF-signiling.pdf`)

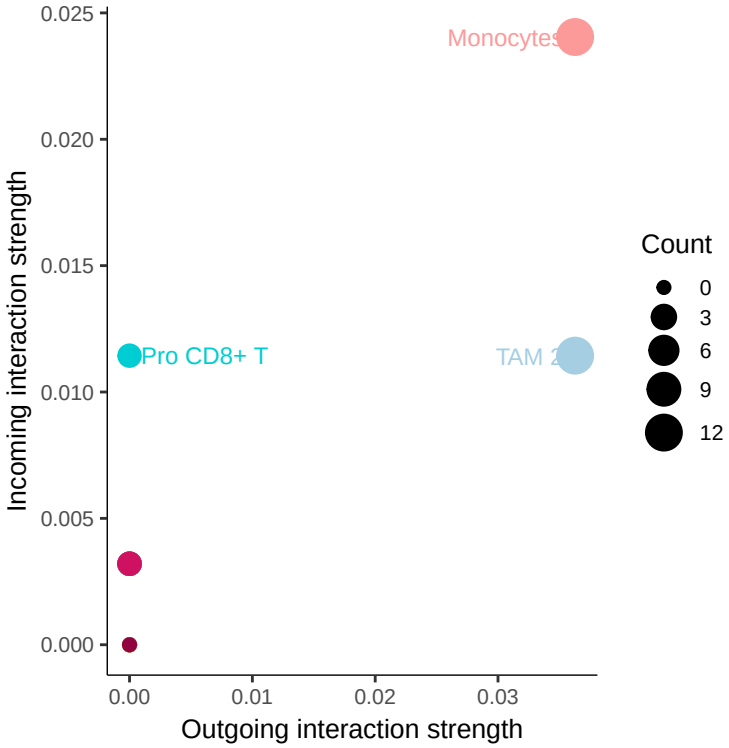


Figure 39: Role of TNF signiling

Figure 40为图 cell communication of TNF signiling 概览。

(对应文件为 `Figure+Table/cell-communication-of-TNF-signiling.pdf`)

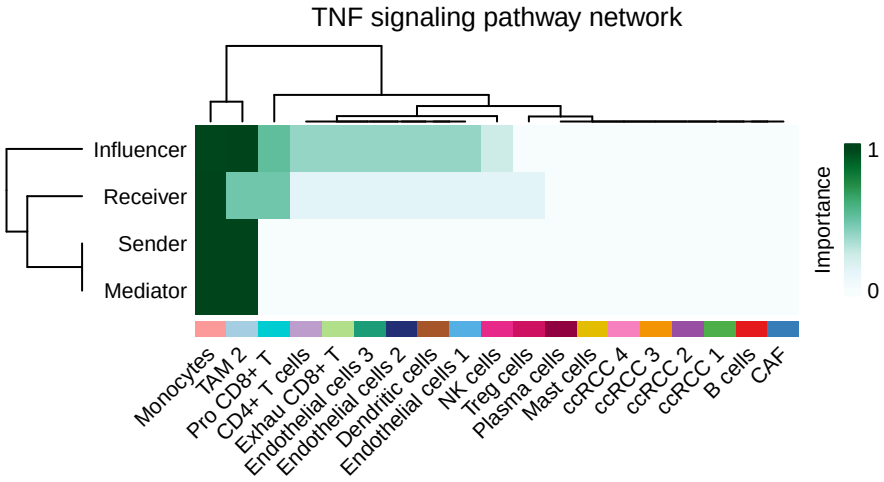


Figure 40: Cell communication of TNF signiling

细胞通讯结果显示，昼夜相关基因与 TNF （肿瘤坏死因子）通路相关。但是，细胞通讯图并不显示免疫细

胞与 ccRCC 细胞之间的交互关系，这提示昼夜节律基因并不直接作用于 ccRCC 细胞，但是通过作用于免役细胞间接生效。其中，Monocytes 可能是关键细胞（TNFRSF1B 高表达）。

7.4 其他

mda: cc review: 33384351 (science), 36291855[2], 31300477 (cancer res)[3]

Reference

1. He, L. *et al.* Single-cell transcriptomic analysis reveals circadian rhythm disruption associated with poor prognosis and drug-resistance in lung adenocarcinoma. *Journal of Pineal Research* **73**, (2022).

2. Amiama-Roig, A., Verdugo-Sivianes, E. M., Carnero, A. & Blanco, J.-R. Chronotherapy: Circadian rhythms and their influence in cancer therapy. *Cancers* **14**, (2022).

3. Shafi, A. A. & Knudsen, K. E. Cancer and the circadian clock. *Cancer Research* **79**, (2019).

4. Santoni, M. *et al.* Role of clock genes and circadian rhythm in renal cell carcinoma: Recent evidence and therapeutic consequences. *Cancers* **15**, (2023).

5. Yu, Z. *et al.* Integrative single-cell analysis reveals transcriptional and epigenetic regulatory features of clear cell renal cell carcinoma. *Cancer Research* **83**, (2023).

6. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, (2021).

7. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, (2019).

8. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nature Methods* **16**, (2019).

9. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology* **20**, (2019).

10. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, (2017).

11. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, (2014).

12. Langfelder, P. & Horvath, S. WGCNA: An r package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).

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

14. Neumann, U., Genze, N. & Heider, D. EFS: An ensemble feature selection tool implemented as r-package and web-application. *BioData Mining* **10**, 21 (2017).

15. Jin, S. *et al.* Inference and analysis of cell-cell communication using cellchat. *Nature Communications* **12**, (2021).

16. Jonasch, E., Walker, C. L. & Rathmell, W. K. Clear cell renal cell carcinoma ontogeny and mechanisms of lethality. *Nature Reviews Nephrology* **17**, (2020).

17. Lex, A. & Gehlenborg, N. Sets and intersections. *Nature Methods* **11**, (2014).

18. Colaprico, A. *et al.* TCGAbiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research* **44**, (2015).