

生信分析报告

项目标题: 骨肉瘤

单号: BSZD231122

分析人员: 黄礼闯

分析类型: 分析优化

委托人: 杨立宇

受托人: 杭州铂赛生物科技有限公司





Contents

1 分析流程	1
2 材料和方法	1
2.1 数据分析平台	1
2.2 Seurat 集成单细胞数据分析 (Dataset: OS)	1
2.3 CopyKAT 癌细胞鉴定 (Dataset: OS)	2
2.4 scFEA 单细胞数据的代谢通量预测 (Dataset: OS_SAMPLE)	2
2.5 Seurat 细胞亚群分析 (Dataset: OS_CANCER)	2
2.6 Limma 代谢通量差异分析 (Dataset: OS_CANCER_FLUX)	2
2.7 TCGA 数据获取 (Dataset: OS)	2
2.8 COX 回归 (Dataset: TCGA_OS)	2
2.9 Survival 生存分析 (Dataset: TCGA_OS)	2
2.10 GSE 数据搜索 (Dataset: OS)	3
2.11 GEO 数据获取 (Dataset: OS_GSE39057)	3
2.12 GEO 数据获取 (Dataset: OS_GSE39055)	3
2.13 GEO 数据获取 (Dataset: OS_GSE16091)	3
2.14 GEO 数据获取 (Dataset: OS_GSE21257)	3
2.15 Survival 生存分析 (Dataset: OS_OUTER)	3
2.16 ClusterProfiler 富集分析 (Dataset: PROG)	3
3 分析结果	4
3.1 Seurat 集成单细胞数据分析 (OS)	4
3.2 CopyKAT 癌细胞鉴定 (OS)	11
3.3 scFEA 单细胞数据的代谢通量预测 (OS_SAMPLE)	13
3.4 Seurat 细胞亚群分析 (OS_CANCER)	15
3.4.1 Seurat-copyKAT 癌细胞注释 (OS_CANCER)	16
3.4.2 Limma 代谢通量差异分析 (OS_CANCER_FLUX)	17
3.5 TCGA 数据获取 (OS)	20
3.6 COX 回归 (TCGA_OS)	20
3.7 Survival 生存分析 (TCGA_OS)	25
3.8 外部数据集验证	26
3.8.1 GSE 数据搜索 (OS)	26
3.8.2 GEO 数据获取 (OS_GSE39057)	26
3.8.3 GEO 数据获取 (OS_GSE39055)	26
3.8.4 GEO 数据获取 (OS_GSE16091)	26
3.8.5 GEO 数据获取 (OS_GSE21257)	26

3.8.6 Survival 生存分析 (OS_OUTER)	26
3.9 ClusterProfiler 富集分析 (PROG)	28
4 总结	29
Reference	29



List of Figures

1 Route	1
2 Pre Quality control	4
3 OS After Quality control	5
4 OS Standard deviations of PCs	6
5 OS UMAP Unintegrated	7
6 OS UMAP Integrated	7
7 OS Marker Validation	9
8 OS SCSA Cell type annotation	10
9 OS SCSA Cell Proportions in each sample	11
10 OS proportions of aneuploid and diploid	12
11 OS SAMPLE Convergency of the loss terms during training	13
12 OS SAMPLE cells metabolic flux	15
13 OS CANCER The scsa cell	16
14 OS CANCER Cancer Cell type annotation	17
15 OS CANCER cancer cell proportions	17
16 OS CANCER FLUX Malignant cell BC vs Benign cell BC	18
17 OS SAMPLE Malignant cell Benign cell Cell flux ridge plot	20
18 TCGA OS lasso COX model	22
19 TCGA OS lasso COX coeffients lambda min	23
20 TCGA OS lasso COX coeffients lambda 1se	23
21 TCGA OS lasso COX ROC lambda min	24
22 TCGA OS lasso COX ROC lambda 1se	25
23 OS OUTER all datasets survival plot	27
24 OS OUTER all datasets ROC validation	27
25 PROG KEGG enrichment	28
26 PROG GO enrichment	29

List of Tables

1	OS significant markers of cell clusters	8
2	BC10 copyKAT prediction data	12
3	OS SAMPLE annotation of metabolic flux	14
4	OS SAMPLE metabolic flux matrix	14
5	OS CANCER FLUX data Malignant cell BC vs Benign cell BC	19
6	TCGA OS sig Univariate Cox Coefficients	21

1 分析流程

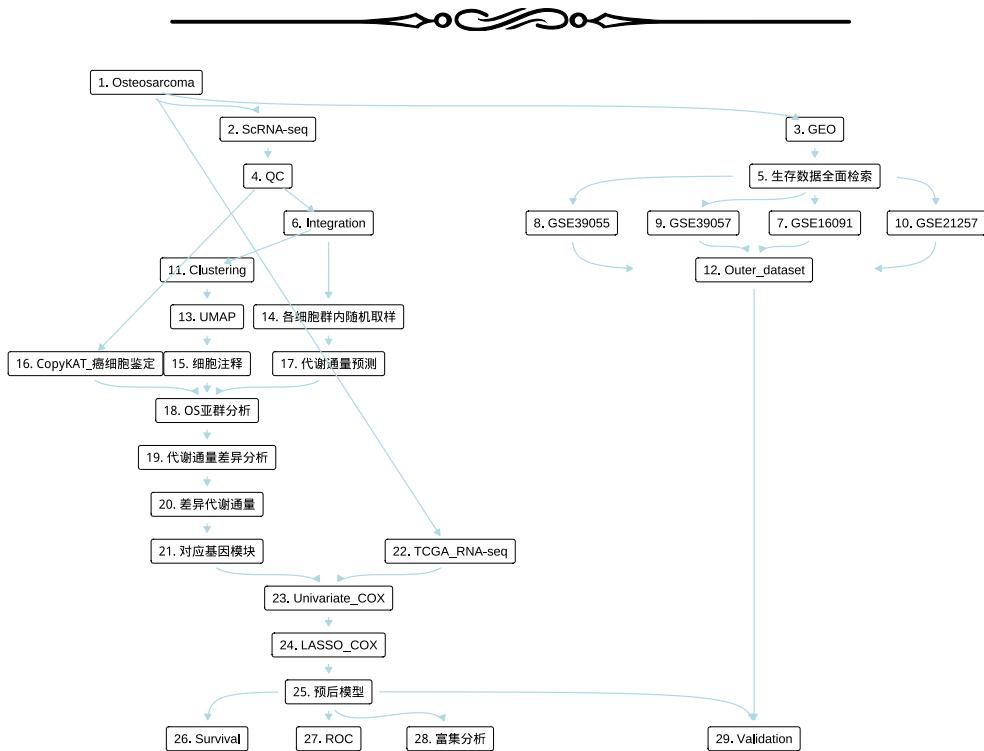


Figure 1: Route

2 材料和方法

2.1 数据分析平台

在 Linux pop-os x86_64 (6.9.3-76060903-generic) 上，使用 R version 4.4.2 (2024-10-31) (<https://www.r-project.org/>) 对数据统计分析与整合分析。

2.2 Seurat 集成单细胞数据分析 (Dataset: OS)

使用 Seurat R 包 (5.1.0) 进行单细胞数据质量控制 (QC) 和下游分析。依据 https://satijalab.org/seurat/articles/integration_introduction 为指导对单细胞数据预处理。一个细胞至少应有 1000 个基因，并且基因数量小于 7000。线粒体基因的比例小于 10%。根据上述条件，获得用于下游分析的高质量

细胞。执行标准 Seurat 分析工作流 (NormalizeData, FindVariableFeatures, ScaleData, RunPCA)。以 ElbowPlot 判断后续分析的 PC 维度。以 Seurat::IntegrateLayers 集成数据，去除批次效应 (使用 HarmonyIntegration 方法)。在 1-10 PC 维度下，以 Seurat::FindNeighbors 构建 Nearest-neighbor Graph。随后在 1.2 分辨率下，以 Seurat::FindClusters 函数识别细胞群并以 Seurat::RunUMAP 进行 UMAP 聚类。以 Seurat::FindAllMarkers (LogFC 阈值 0.25; 最小检出率 0.01) 为所有细胞群寻找 Markers。以 Python 工具 SCSA ((2020, IF:2.8, Q2, Frontiers in genetics)¹) (<https://github.com/bioinfo-ibms-pumc/SCSA>) 对细胞群注释。

2.3 CopyKAT 癌细胞鉴定 (Dataset: OS)

R 包 CopyKAT 用于鉴定恶性细胞 (2021, IF:33.1, Q1, Nature Biotechnology)²。CopyKAT 可以区分整倍体与非整倍体，其中非整倍体被认为是肿瘤细胞，而整倍体是正常细胞 (2012, IF:39.1, Q1, Nature Reviews Genetics)³。由于 CopyKAT 不适用于多样本数据 (批次效应的存在)，因此，对各个样本独立鉴定。

2.4 scFEA 单细胞数据的代谢通量预测 (Dataset: OS_SAMPLE)

将 Seurat 的 RNA Assay ('counts') 作为输入数据，以 scFEA 预测细胞的代谢通量 (2021, IF:6.2, Q1, Genome research)⁴。参考 https://github.com/changwn/scFEA/blob/master/scFEA_tutorial1.ipynb 和 https://github.com/changwn/scFEA/blob/master/scFEA_tutorial2.ipynb。

2.5 Seurat 细胞亚群分析 (Dataset: OS_CANCER)

执行标准 Seurat 分析工作流 (NormalizeData, FindVariableFeatures, ScaleData, RunPCA)。以 ElbowPlot 判断后续分析的 PC 维度。在 1-15 PC 维度下，以 Seurat::FindNeighbors 构建 Nearest-neighbor Graph。随后在 1.2 分辨率下，以 Seurat::FindClusters 函数识别细胞群并以 Seurat::RunUMAP 进行 UMAP 聚类。

2.6 Limma 代谢通量差异分析 (Dataset: OS_CANCER_FLUX)

以 limma (3.62.2) (2005)⁵ 差异分析。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。创建设计矩阵，对比矩阵，差异分析：Malignant_cell_BC vs Benign_cell_BC。使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 拟合线形模型。以 limma::topTable 提取所有结果，并过滤得到 adj.P.Val 小于 0.05, |Log2(FC)| 大于 0.5 的统计结果。

2.7 TCGA 数据获取 (Dataset: OS)

以 R 包 TCGAbiolinks (2.35.1) (2015, IF:16.6, Q1, Nucleic Acids Research)⁶ 获取 TARGET-OS 数据集。

2.8 COX 回归 (Dataset: TCGA_OS)

以 R 包 survival (3.8.3) 进行单因素 COX 回归 (survival::coxph)。筛选 $\text{Pr}(>|z|) < .05$ 的基因。以 R 包 glmnet(4.1.8) 作 lasso 处罚的 cox 回归，以 cv.glmnet‘ 函数作 5 交叉验证获得模型。

2.9 Survival 生存分析 (Dataset: TCGA_OS)

以 R 包 survival (3.8.3) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

2.10 GSE 数据搜索 (Dataset: OS)

使用 Entrez Direct (EDirect) <https://www.ncbi.nlm.nih.gov/books/NBK3837/> 搜索 GEO 数据库 (`esearch -db gds`)，查询信息为: ((Osteosarcoma[Description]) AND ((6:300[Number of Samples]) AND (GSE[Entry Type]) AND (Homo sapiens[Organism])))，转化为数据表格。以正则匹配，滤除‘summary’或‘title’中包含‘single cell’或‘scRNA’的数据例。仅查询临床数据，因此滤除匹配到关键词 *in vitro*, *cell line*, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。(注：以上仅为查找合适的 GEO 数据所做的数据筛选，与实际分析无关)。仅获取类型包含‘Expression profiling by high throughput sequencing’或‘Expression profiling by array’的数据例。此外，排除 summary 或 title 中匹配到字符集 EX (KO, WT, *WT*, KO, wildtype, mutant, knock, deficien, absen, SuperSeries, transgenic, CD[0-9]+) 的数据例。上述得到 147 个 GSE 数据集。以 R 抓取网页 (例如, `RCurl::getURL` 抓取)，解析‘Overall design’和‘Samples’模块，匹配字符集 EX，排除匹配到的数据例。排除‘Overall design’中包含 *in vitro*, *cell line*, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。仅获取包含‘protein coding’测序的数据集，排除‘Samples’和‘Overall design’中包含 siRNA, miRNA, miR, lncRNA 字符的数据例。余下共 73 个。以 `GEOquery` 获取 GSE 数据集 (n=73)。从元数据中匹配包含关键词的数据：‘Survival|Event|Dead|Alive|Status|Day|Time’，共得到 17 个数据集。

2.11 GEO 数据获取 (Dataset: OS_GSE39057)

以 R 包 `GEOquery` (2.74.0) 获取 GSE39057 数据集。

2.12 GEO 数据获取 (Dataset: OS_GSE39055)

以 R 包 `GEOquery` (2.74.0) 获取 GSE39055 数据集。

2.13 GEO 数据获取 (Dataset: OS_GSE16091)

以 R 包 `GEOquery` (2.74.0) 获取 GSE16091 数据集。

2.14 GEO 数据获取 (Dataset: OS_GSE21257)

以 R 包 `GEOquery` (2.74.0) 获取 GSE21257 数据集。

2.15 Survival 生存分析 (Dataset: OS_OUTER)

以 R 包 `survival` (3.8.3) 生存分析，以 R 包 `survminer` (0.5.0) 绘制生存曲线。

2.16 ClusterProfiler 富集分析 (Dataset: PROG)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)⁷ 进行 KEGG 和 GO 富集分析。以 `p.adjust` 表示显著水平。

3 分析结果

3.1 Seurat 集成单细胞数据分析 (OS)

读取 BC10, BC11, BC16, BC17, BC2, BC20, BC21, BC22, BC3, BC5, BC6 样本的数据集。前期质量控制，一个细胞至少应有 1000 个基因，并且基因数量小于 7000。线粒体基因的比例小于 10%。数据归一化，PCA 聚类 (Seurat 标准工作流，见方法章节) 后，绘制 PC standard deviations 图。去除批次效应后 (详见方法章节)，在 1-10 PC 维度，1.2 分辨率下，对细胞群 UMAP 聚类。计算所有细胞群的 Marker。使用特异性 Marker，以 SCSA 对细胞群注释。

(鉴定所用 Markers 来源于原作文献 PMID:33303760 (2020, IF:14.7, Q1, Nature communications)⁸)

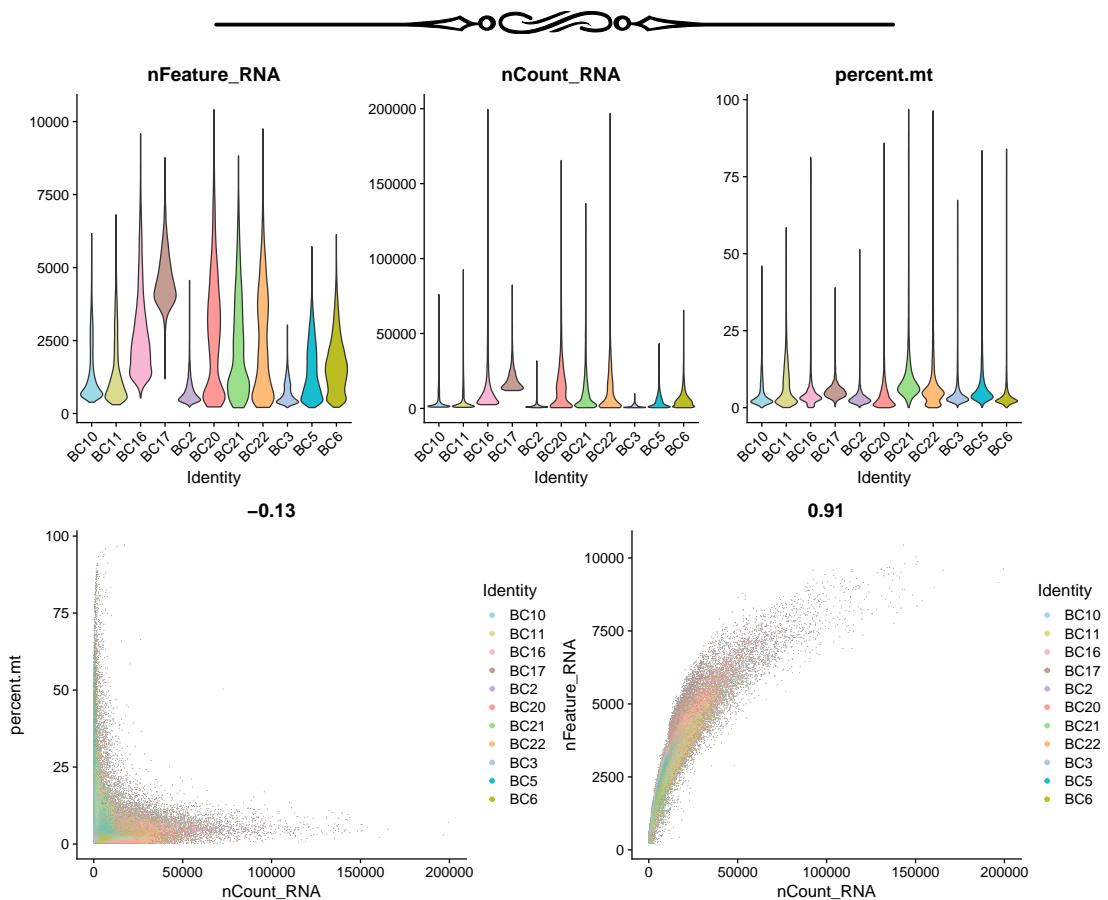


Figure 2: Pre Quality control

Fig. 2

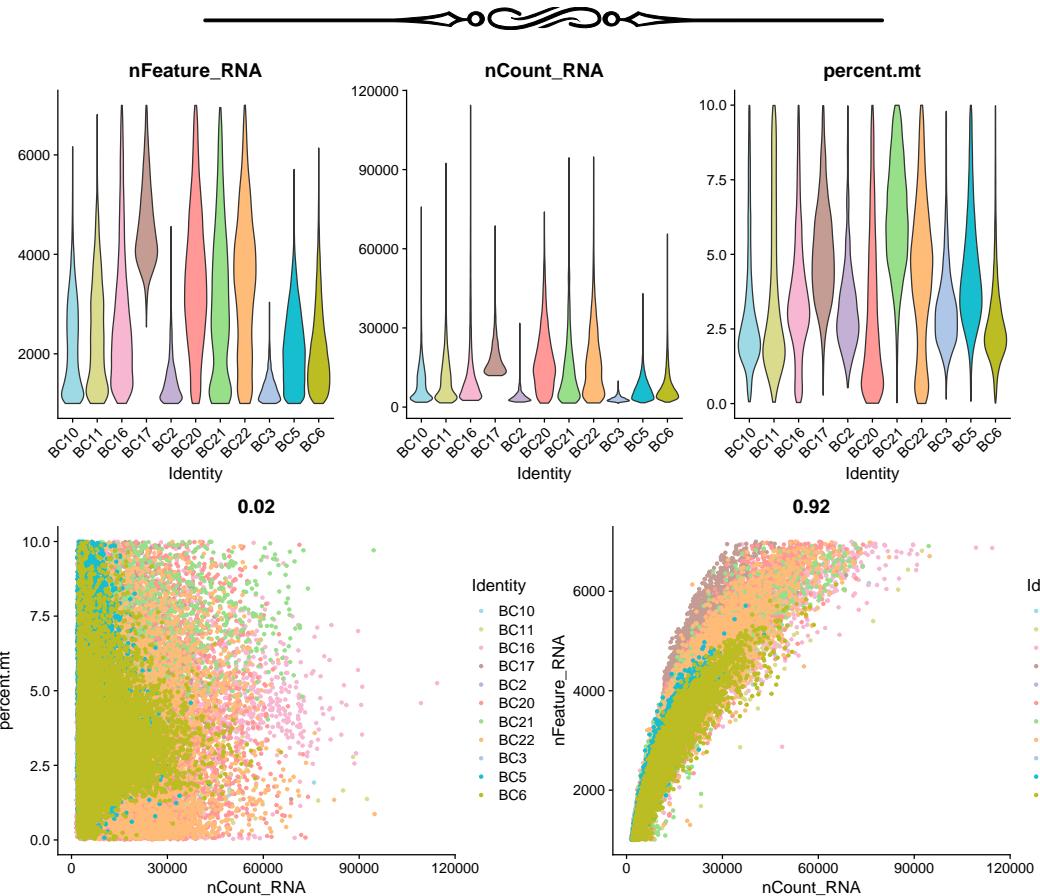


Figure 3: OS After Quality control

Fig. 3

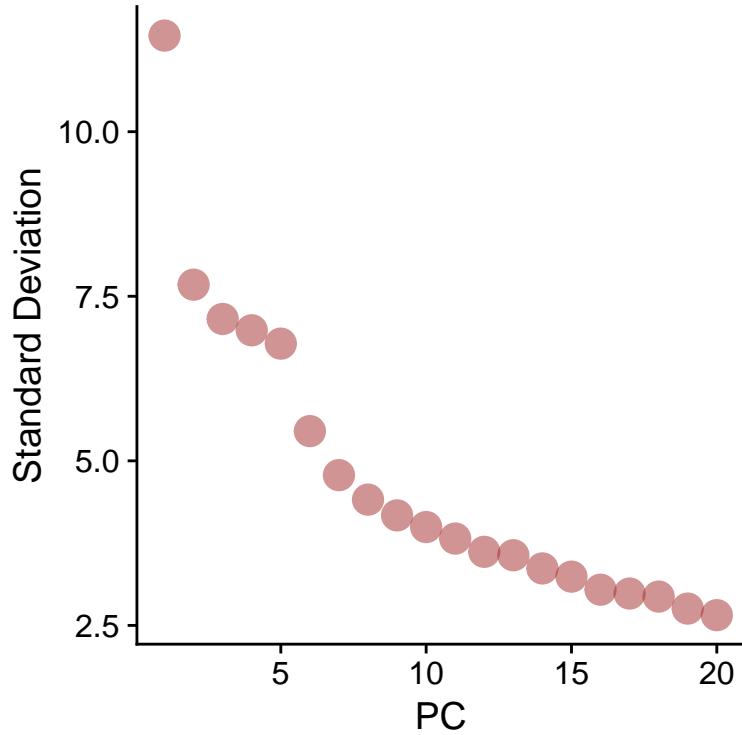


Figure 4: OS Standard deviations of PCs



Fig. 4



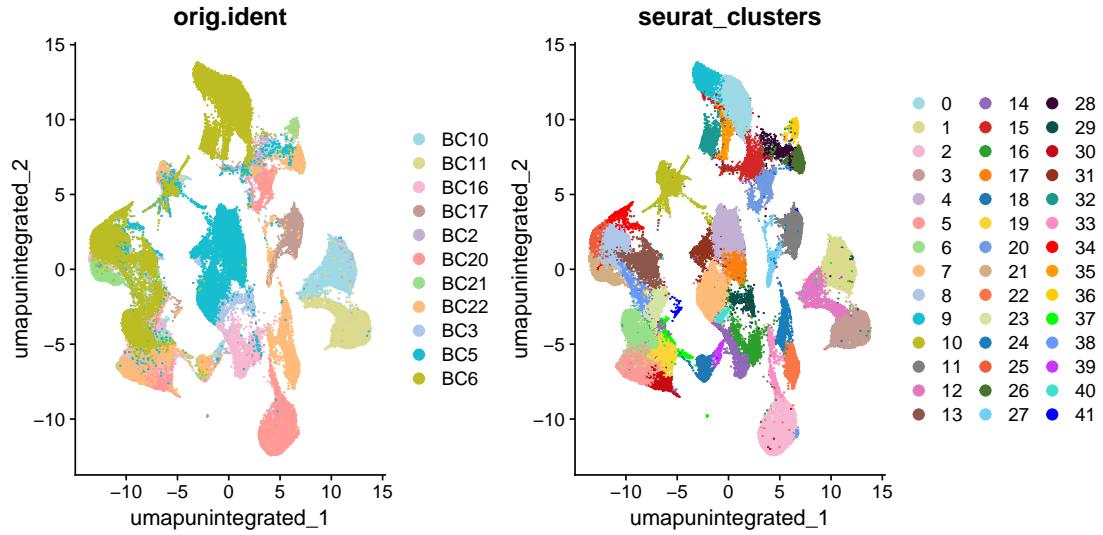


Figure 5: OS UMAP Unintegrated



Fig. 5

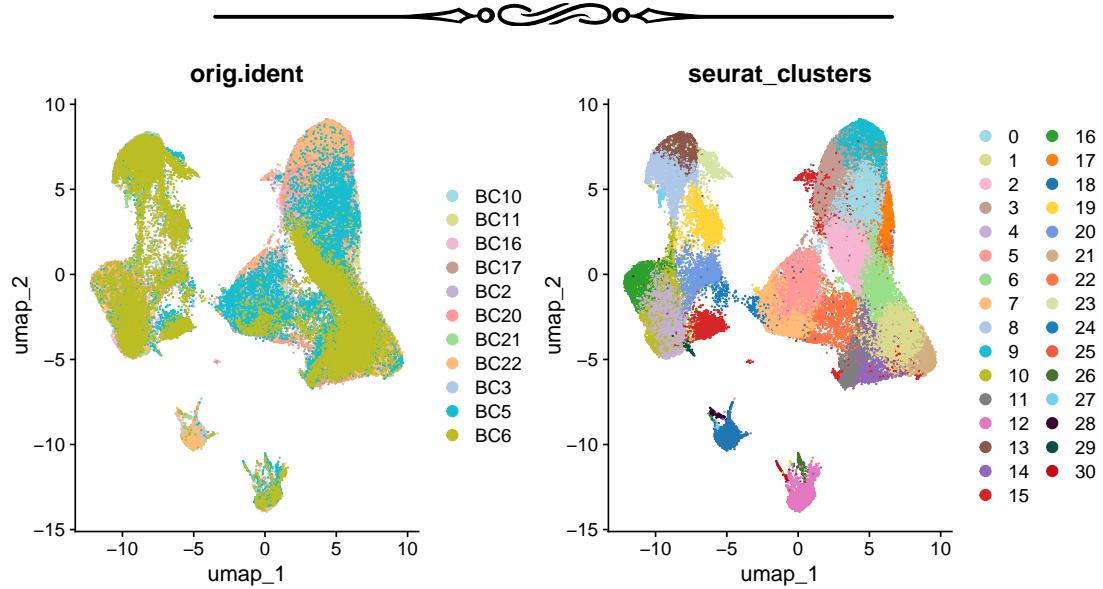


Figure 6: OS UMAP Integrated



Fig. 6



Table 1: OS significant markers of cell clusters

rownames	p_val	avg_log2FC	pct.1	pct.2
ALPL	0	1.494	0.915	0.412
PANX3	0	2.181	0.645	0.156
IFITM5	0	2.458	0.802	0.323
LY6K	0	1.191	0.77	0.314
RHBDL2	0	1.733	0.644	0.21
...



Tab. 1



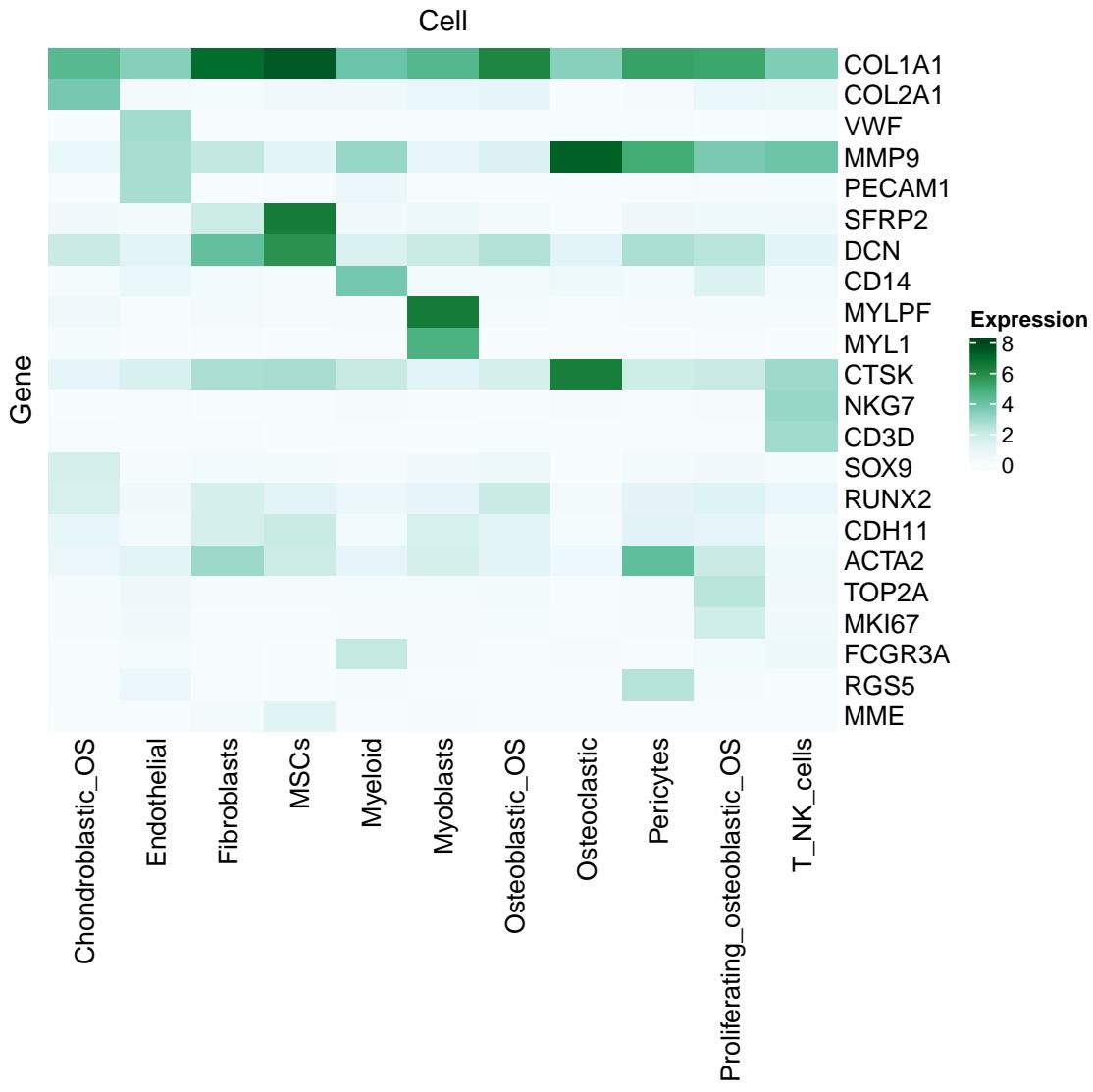


Figure 7: OS Marker Validation



Fig. 7



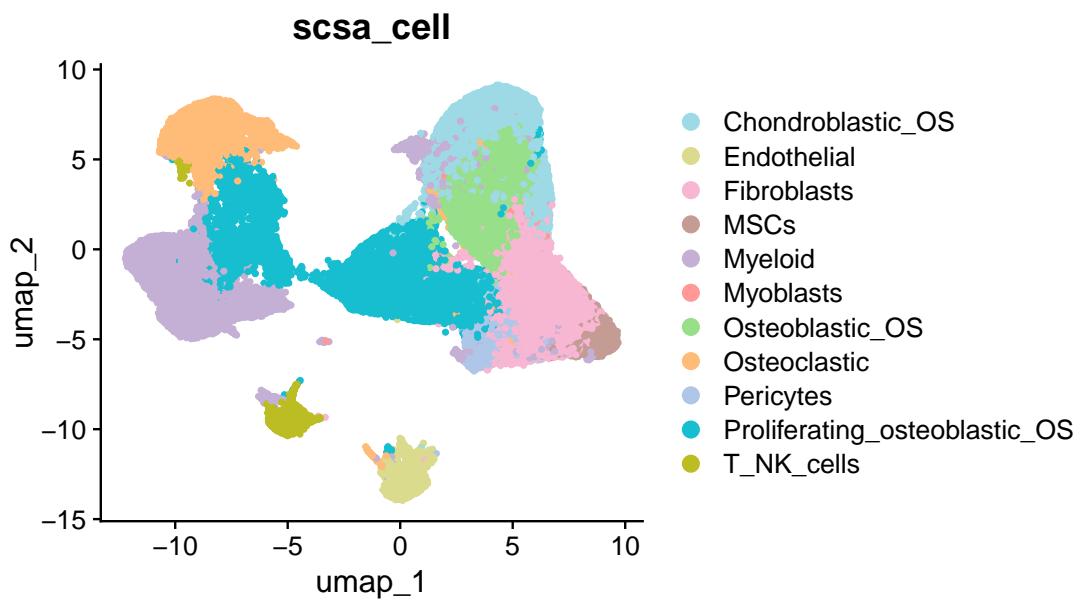


Figure 8: OS SCSA Cell type annotation

Fig. 8

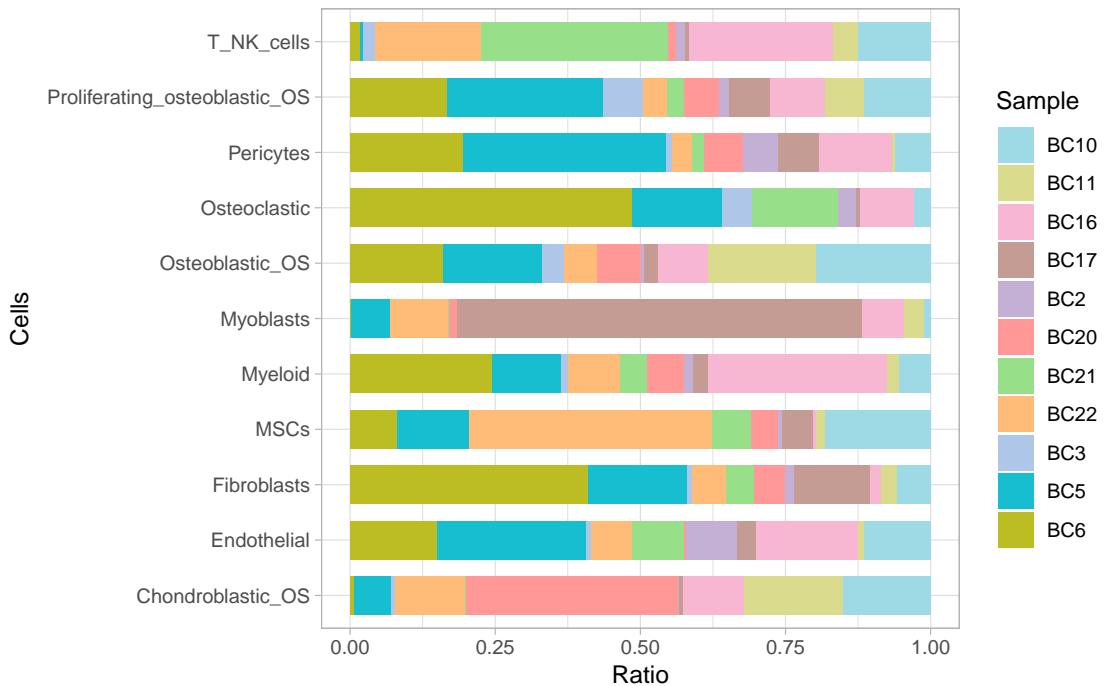


Figure 9: OS SCSA Cell Proportions in each sample

Fig. 9

3.2 CopyKAT 癌细胞鉴定 (OS)

以 CopyKAT 鉴定恶质细胞。



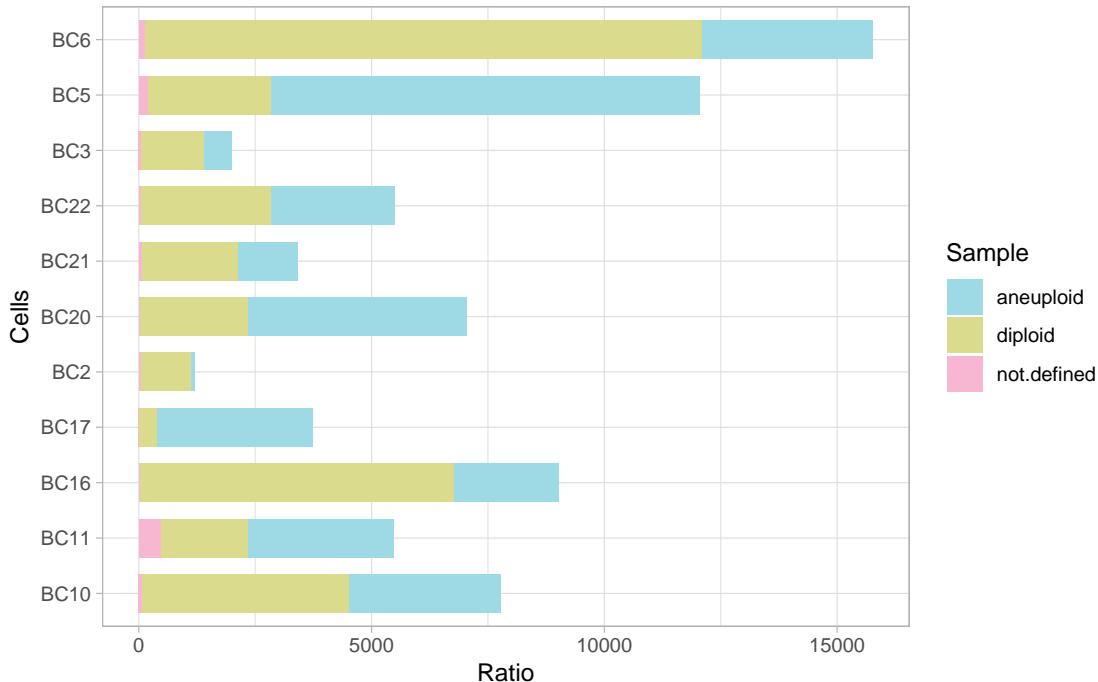


Figure 10: OS proportions of aneuploid and diploid



Fig. 10

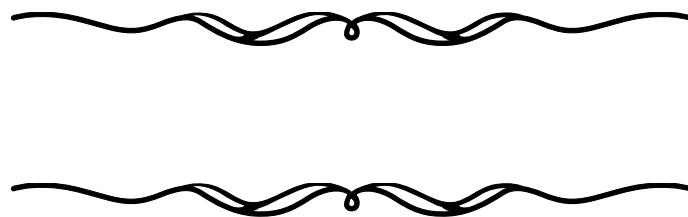


Table 2: BC10 copyKAT prediction data

orig.ident	cell.names	copykat.pred	copykat_cell
BC10	AAACCTGAGACTCGGA-1_1	aneuploid	Cancer cell
BC10	AAACCTGAGGAACTGC-1_1	diploid	Normal cell
BC10	AAACCTGAGGATGGAA-1_1	diploid	Normal cell
BC10	AAACCTGAGGTGCTT-1_1	diploid	Normal cell
BC10	AAACCTGAGTAGCGGT-1_1	diploid	Normal cell
...



Tab. 2



3.3 scFEA 单细胞数据的代谢通量预测 (OS_SAMPLE)

根据样本和细胞类型分组，将细胞随机抽样 (各组比例为: 0.5) (细胞数量较多，通过随机抽样的方式减少计算负担) (随机种子: 987456)。将 Seurat (所有细胞) 以 scFEA 预测代谢通量。

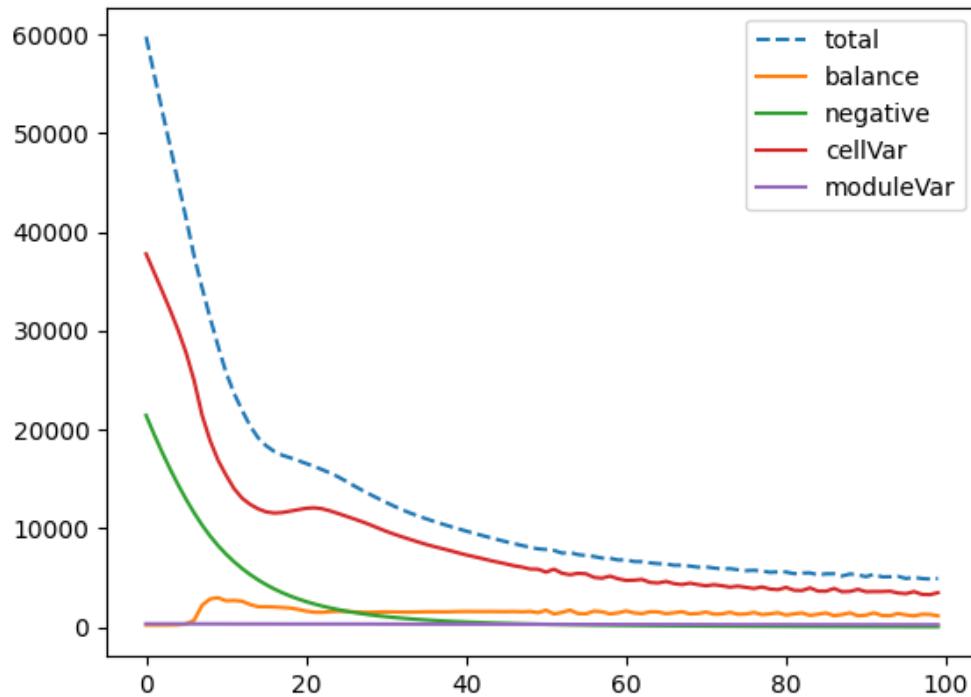


Figure 11: OS SAMPLE Convergence of the loss terms during training



Fig. 11



Table 3: OS SAMPLE annotation of metabolic flux

V1	Module_id	Compound_IN_name	Compound_IN_ID	Compound_OUT_name
M_1	1	Glucose	C00267	G6P
M_2	2	G6P	C00668	G3P
M_3	3	G3P	C00118	3PD
M_4	4	3PD	C00197	Pyruvate
M_5	5	Pyruvate	C00022	Acetyl-Coa
...



Tab. 3



Table 4: OS SAMPLE metabolic flux matrix

V1	M_1	M_2	M_3	M_4
AAACCTGAGGATGGAA-1_1	0.01224	0.01727	0.0649	0.1126
AAACCTGCACAACGCC-1_1	0.01224	0.02175	0.06084	0.07629
AAACCTGTCATCATTC-1_1	0.01618	0.02072	0.02948	0.03468
AAACCTGTCGTCCGTT-1_1	0.01832	0.07996	0.1203	0.2113
AAACCTGTCTTGCATT-1_1	0.01683	0.06143	0.09314	0.1331
...



Tab. 4

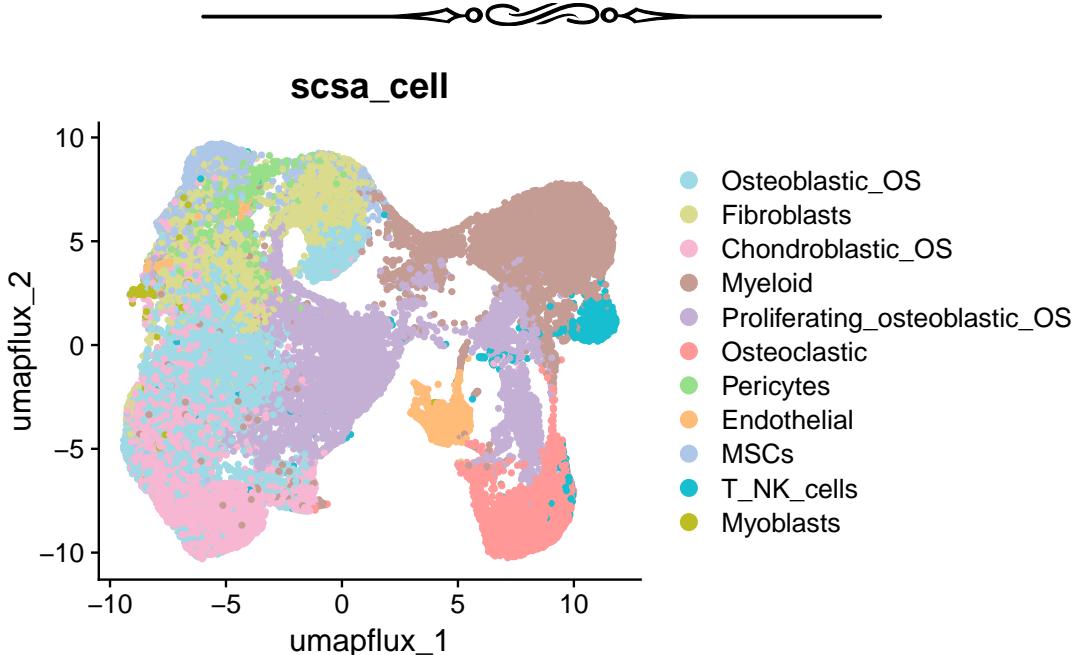


Figure 12: OS SAMPLE cells metabolic flux

Fig. 12

3.4 Seurat 细胞亚群分析 (OS_CANCER)

成骨细胞和软骨细胞骨肉瘤是临幊上常见的两种主要骨肉瘤类型 (2020, IF:14.7, Q1, Nature communications)⁸。在这里，聚焦于注释结果中的 Proliferating_osteoblastic_OS, Chondroblastic_OS, Osteoblastic_OS 细胞，重新聚类分析。匹配 scsa_cell 中包含”_OS\$” 的描述，最终得到 34230 例数据。分析其亚群。数据归一化，PCA 聚类 (Seurat 标准工作流，见方法章节) 后，绘制 PC standard deviations 图。在 1-15 PC 维度，1.2 分辨率下，对细胞群 UMAP 聚类。

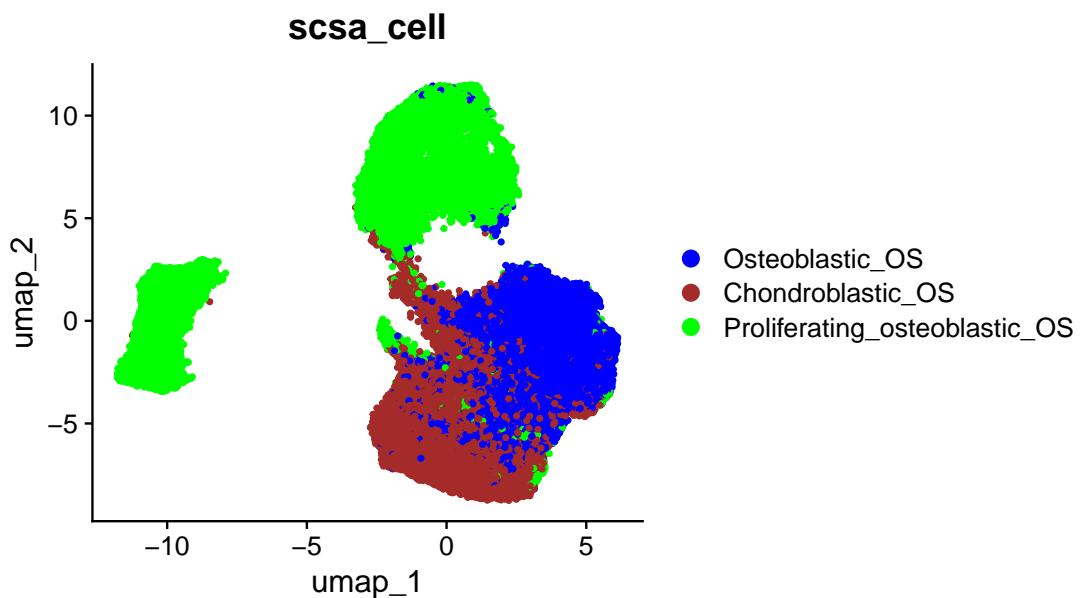


Figure 13: OS CANCER The scsa cell



Fig. 13

3.4.1 Seurat-copyKAT 癌细胞注释 (OS_CANCER)

将 CopyKAT 的预测结果映射细胞注释中。



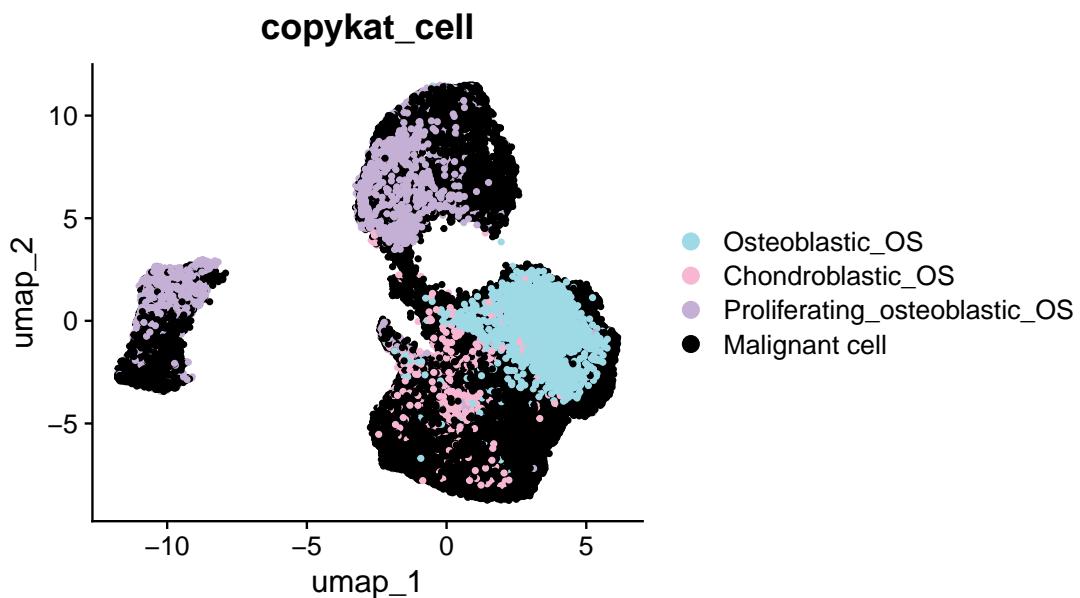


Figure 14: OS CANCER Cancer Cell type annotation

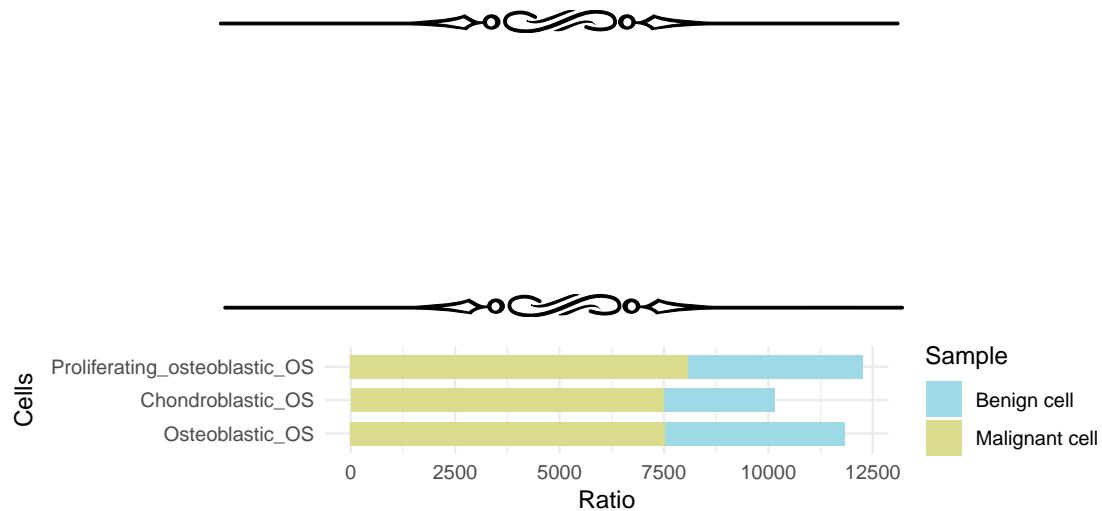


Figure 15: OS CANCER cancer cell proportions

Fig. 15

3.4.2 Limma 代谢通量差异分析 (OS_CANCER_FLUX)

匹配 scsa_cell 中包含”_OS\$” 的描述，最终得到 17122 例数据。以公式 $\sim 0 + \text{group}$ 创建设计矩阵 (design matrix)。差异分析：Malignant_cell_BC vs Benign_cell_BC。(若 A vs B，则为前者比后者，LogFC 大于

0 时, A 表达量高于 B)。上调或下调 DMFs 统计: up (n=29) , down (n=4)

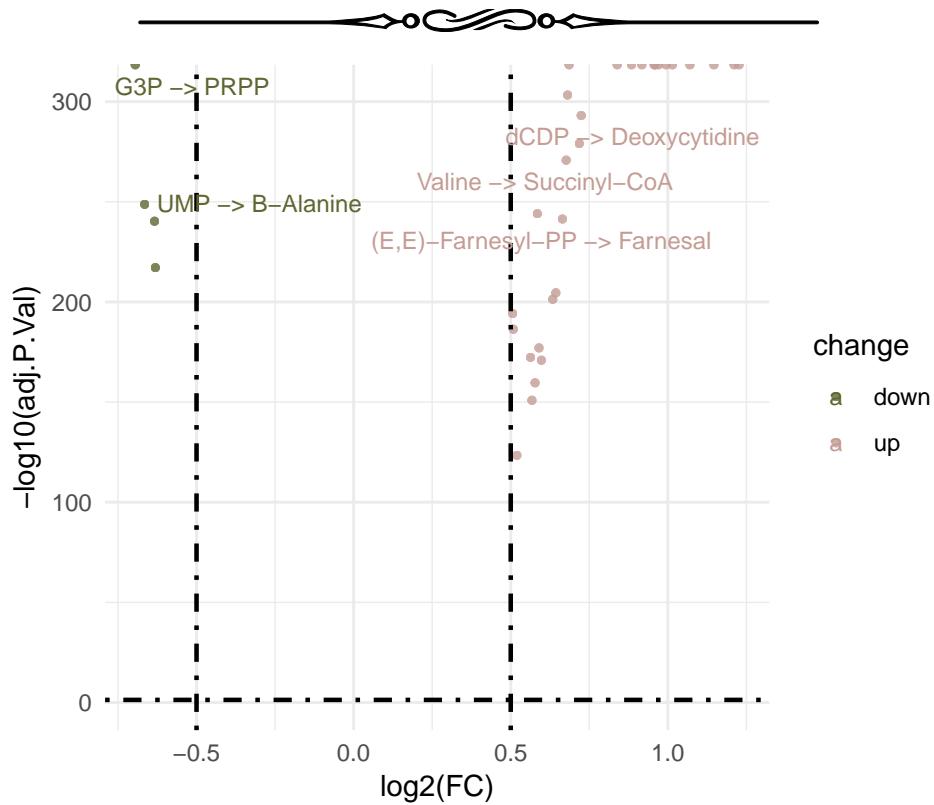


Figure 16: OS CANCER FLUX Malignant cell BC vs Benign cell BC

adj.P.Val cut-off :

0.05

Log2(FC) cut-off :

0.5

(See: Figure+Table/3.4.2_Limma_ 代谢通量差异分析_(OS_CANCER_FLUX)/OS-CANCER-FLUX-Malignant-cell-BC-vs-Benign-cell-BC

Fig. 16

Table 5: OS CANCER FLUX data Malignant cell BC vs Benign cell BC

name	logFC	adj.P.Val	rownames	Module_id
3PD -> Pyruvate	1.227	0	M_4	4
G3P -> 3PD	1.211	0	M_3	3
ADP -> Deoxyadeno...	1.015	0	M_140	140
lysine -> Acetyl-CoA	1.07	0	M_60	60
Pyruvate -> Lactate	1.147	0	M_6	6
...

Tab. 5



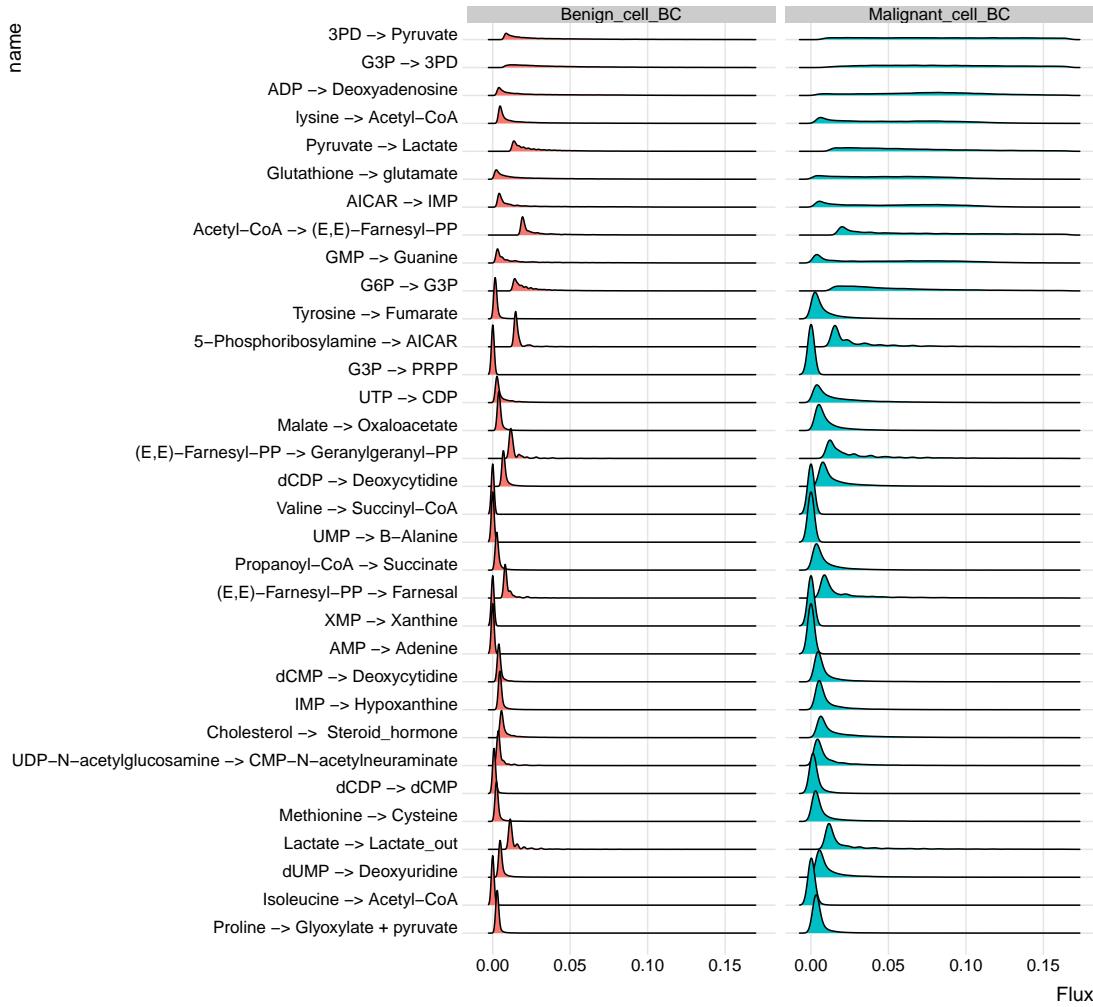


Figure 17: OS SAMPLE Malignant cell Benign cell Cell flux ridge plot



Fig. 17

3.5 TCGA 数据获取 (OS)

获取 TARGET-OS 数据。

3.6 COX 回归 (TCGA_OS)

将基因集 (Malignant_cell_Benign_cell, 来自于 scFEA 单细胞数据的代谢通量预测 [Section: OS_SAMPLE]) 用于模型建立。共 298 个基因在数据集 TARGET-OS 中找到 (根据基因名匹配)。所有数据生存状态 (去除生存状态未知的数据), (Alive (n=57), Dead (n=29))。执行单因素 COX 回归, 筛选 P 值 < 0.05, 共筛选到 25 个基因。在单因素回归得到的基因 (P < 0.01) 的基础上, 使用

`glmnet::cv.glmnet` 作 5 倍交叉验证 (评估方式为 C-index), 筛选 lambda 值。lambda.min, lambda.1se 值分别为 0.006, 0.07 (R 随机种子为 987456)。对应的特征数 (基因数) 分别为 10, 10。

Table 6: TCGA OS sig Univariate Cox Coefficients

feature	coef	exp(coef)	se(coef)	z
ACAT1	0.4221	1.525	0.2025	2.084
UPRT	-0.6027	0.5473	0.208	-2.897
UGT2B10	0.3296	1.39	0.1409	2.339
PCCB	0.4687	1.598	0.175	2.679
PGLS	-0.3914	0.6761	0.1773	-2.207
...

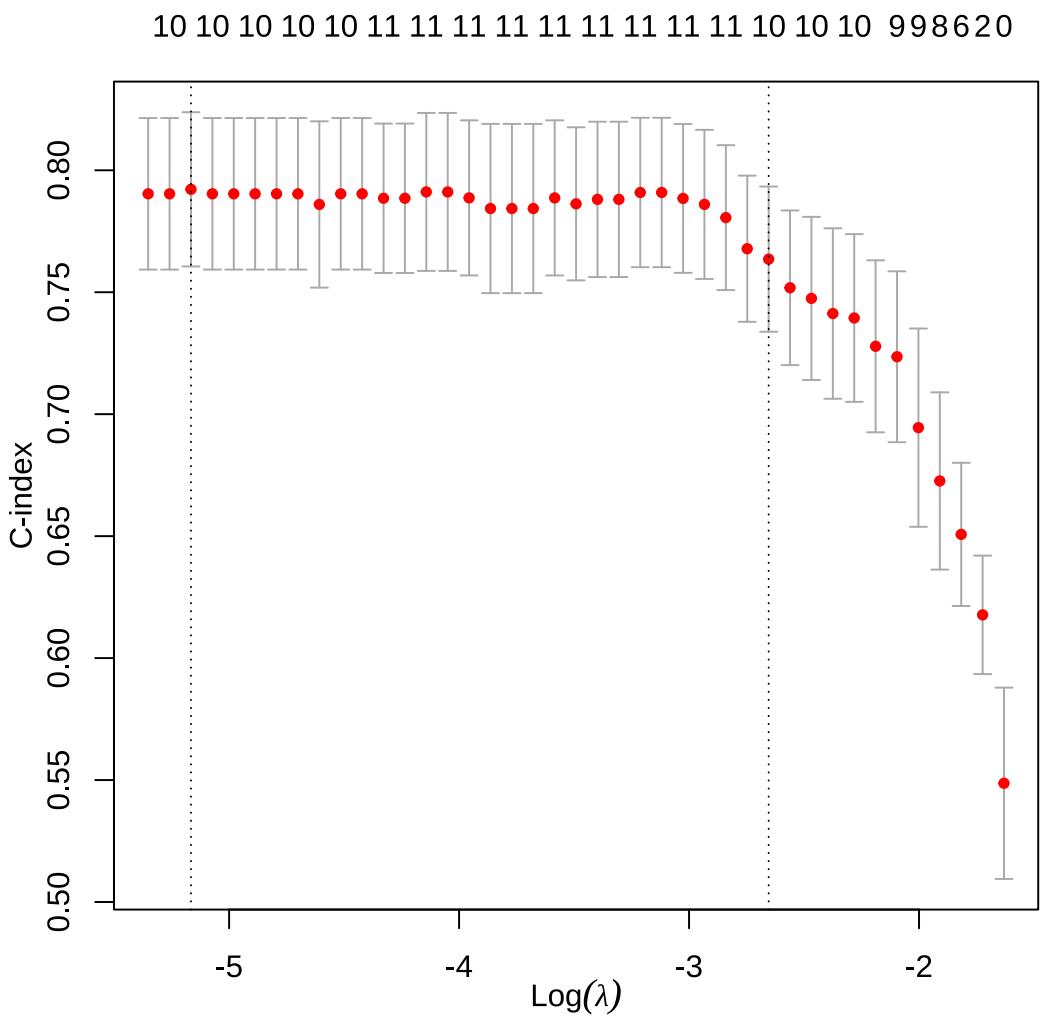
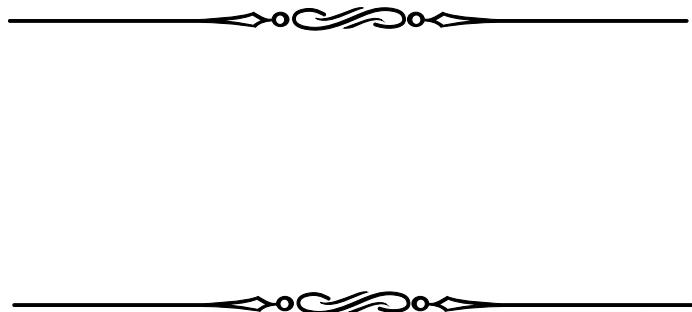


Figure 18: TCGA OS lasso COX model



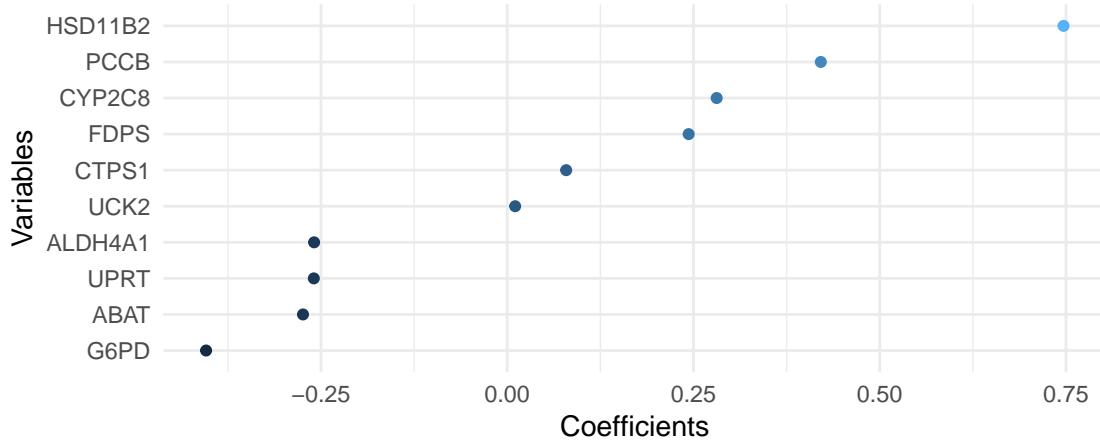


Figure 19: TCGA OS lasso COX coefficients lambda min

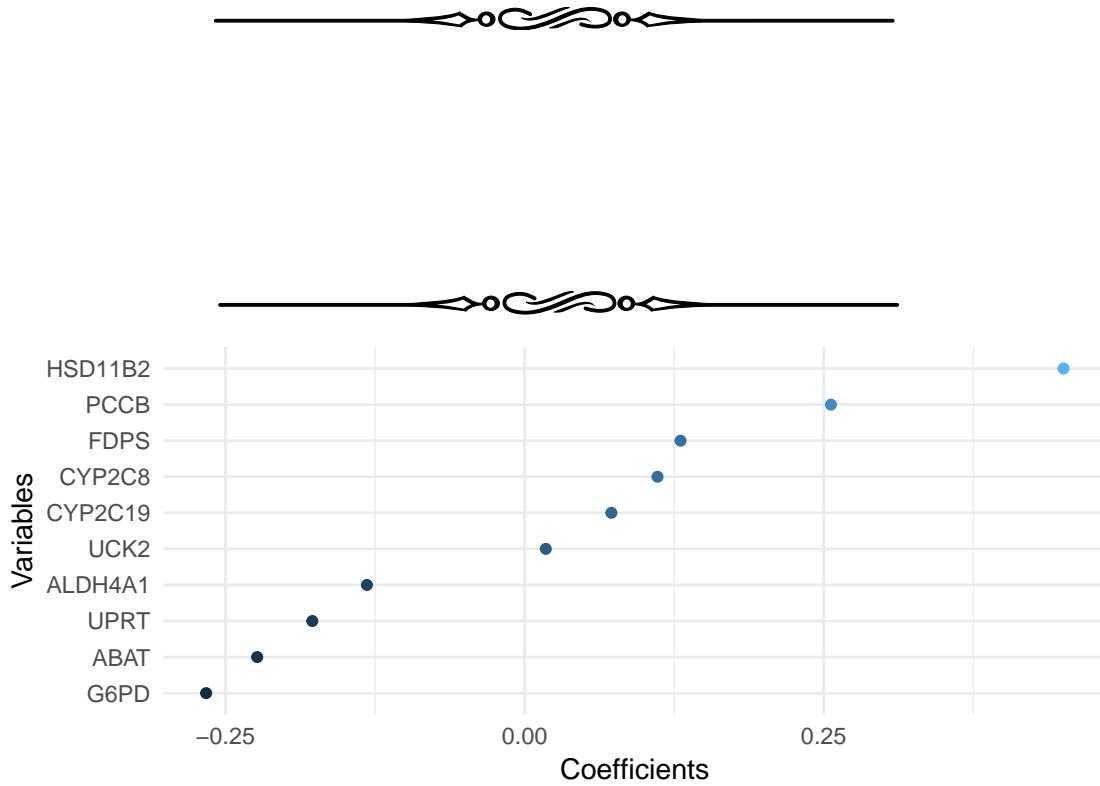


Figure 20: TCGA OS lasso COX coefficients lambda 1se



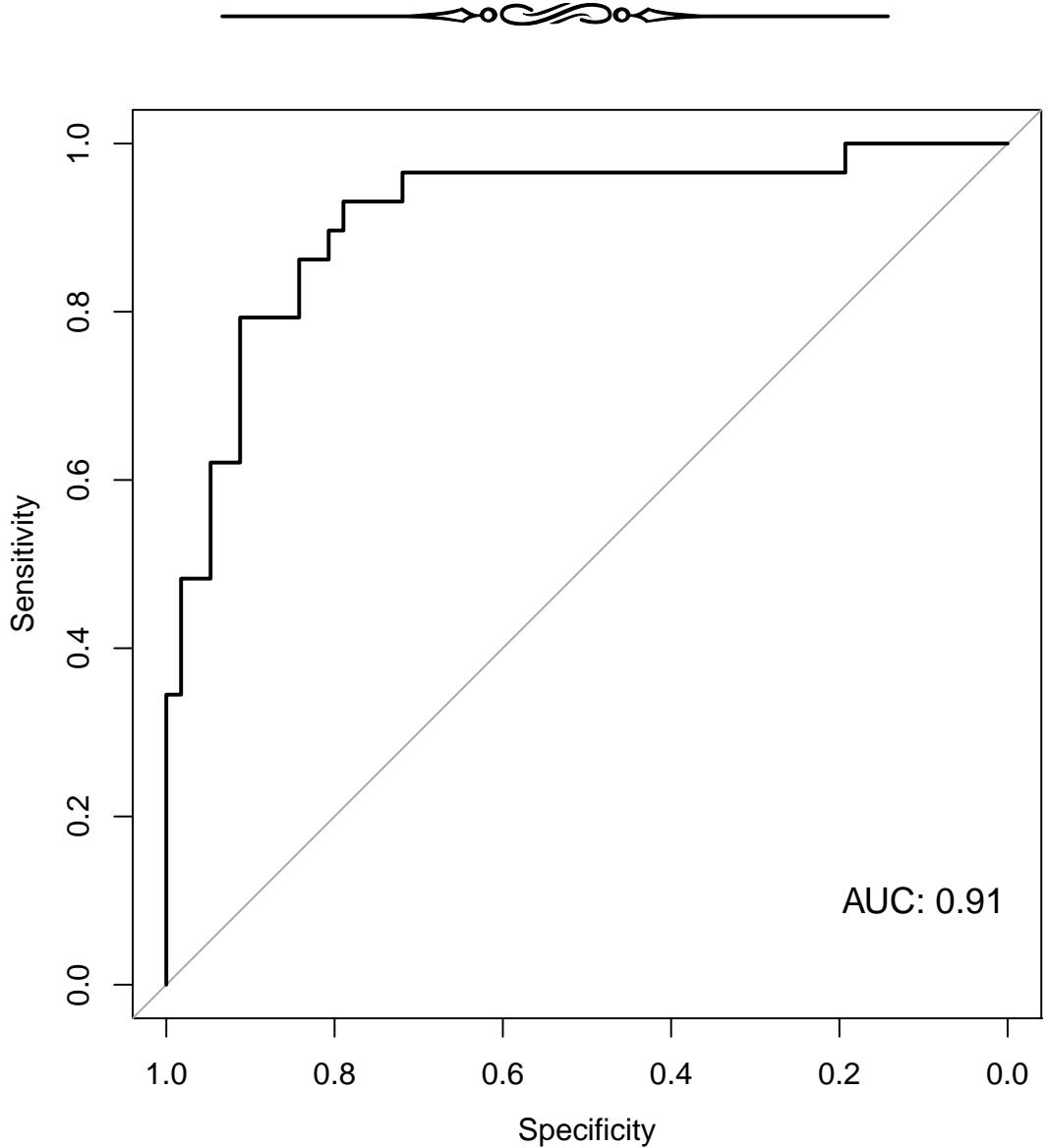
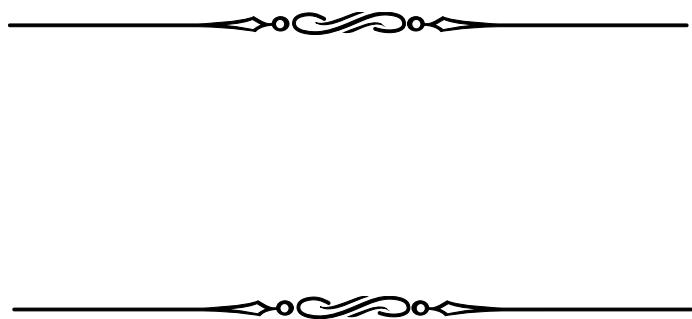


Figure 21: TCGA OS lasso COX ROC lambda min



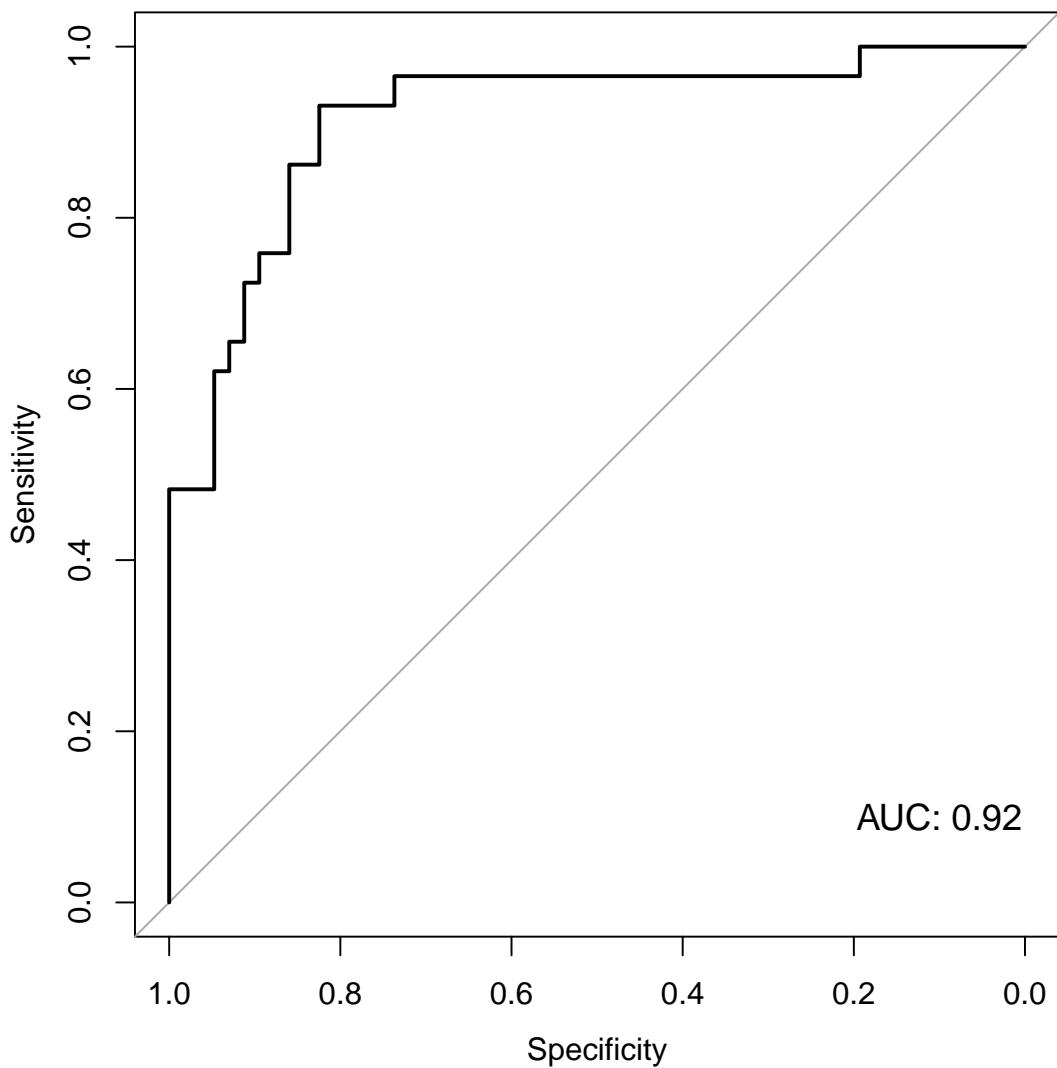


Figure 22: TCGA OS lasso COX ROC lambda 1se

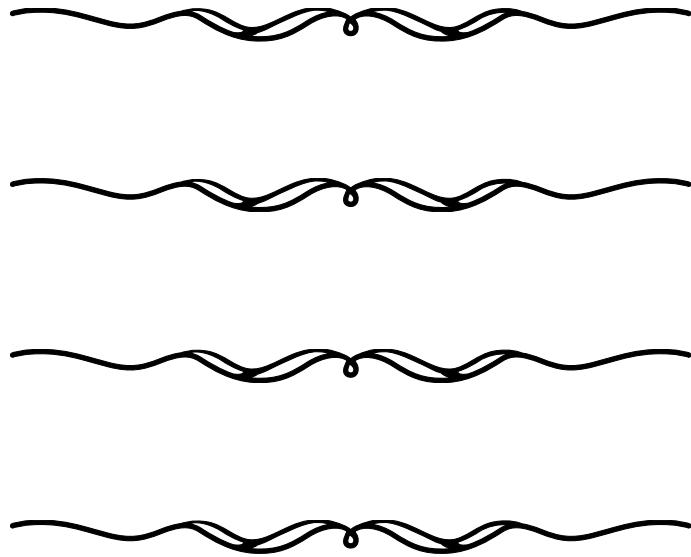


3.7 Survival 生存分析 (TCGA_OS)

选择 lambda.min 时得到的特征集，包含 10 个基因，分别为：UPRT, PCCB, CYP2C8, UCK2, ALDH4A1, G6PD, FDPS, CTPS1, HSD11B2, ABAT。以回归系数构建风险评分模型。

$$Score = \sum(expr(Gene) \times coef)$$

按 `survminer::surv_cutpoint` 计算的 cutoff，将样本分为 Low 和 High 风险组 (cutoff: 0.55168695257864) (High (n=28), Low (n=58))，随后进行生存分析。



3.8 外部数据集验证

3.8.1 GSE 数据搜索 (OS)

以 Entrez Direct (EDirect) 搜索 GEO 数据库 (检索条件见方法章节)。在检索匹配后，经人工确认，全部带有生存数据的 Osteosarcoma 为：GSE16091, GSE39055, GSE39057, GSE21257

3.8.2 GEO 数据获取 (OS_GSE39057)

以 GEOquery 获取 GSE39057 的数据信息。

3.8.3 GEO 数据获取 (OS_GSE39055)

以 GEOquery 获取 GSE39055 的数据信息。

3.8.4 GEO 数据获取 (OS_GSE16091)

以 GEOquery 获取 GSE16091 的数据信息。

3.8.5 GEO 数据获取 (OS_GSE21257)

以 GEOquery 获取 GSE21257 的数据信息。

3.8.6 Survival 生存分析 (OS_OUTER)

对在 GEO 找到的所有具备生存信息的 Osteosarcoma 数据集做了外部验证。合并数据集 (GSE16091, GSE39055, GSE39057, GSE21257)。对于不同注释来源的基因名，以 `org.Hs.eg.db::org.Hs.eg.db` 获取基因的别名 (ALIAS)，根据 (ALIAS) 的一致性合并。查找预后模型中基因的 ALIAS，在未找到对应基因的情况下，使用该基因的 ALIAS 查找。(原模型基因：UPRT, PCCB, CYP2C8, UCK2, ALDH4A1, G6PD, FDPS,

CTPS1, HSD11B2, ABAT; 以 ALIAS 匹配后, 基因为: UPRT, PCCB, CYP2C8, UCK2, ALDH4A1, G6PD, FDPS, CTPS1, HSD11B2, ABAT)。随后, 将基因表达数据归一化 (Z-score)。按 `survminer::surv_cutpoint` 计算的 cutoff, 将样本分为 Low 和 High 风险组 (cutoff: 1.12270604442518) (High (n=15), Low (n=94)), 随后进行生存分析。此外, 对于未合并前的各个数据集, 以相同的方式生存分析。

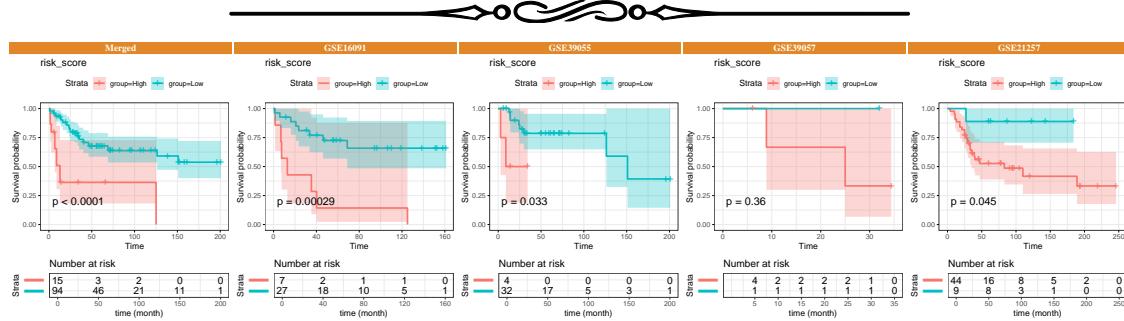


Figure 23: OS OUTER all datasets survival plot

Fig. 23

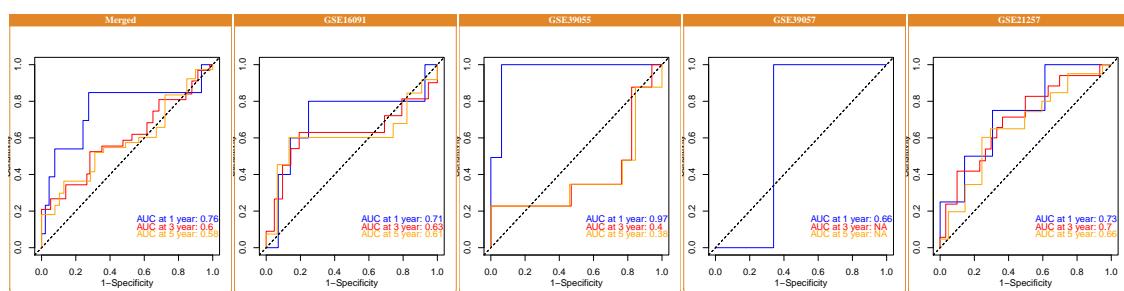


Figure 24: OS OUTER all datasets ROC validation

Fig. 24

3.9 ClusterProfiler 富集分析 (PROG)

对基因集 (UPRT, PCCB, CYP2C8, ...[n = 10], 来自于 Survival 生存分析 [Section: TCGA_OS]) 进行 ClusterProfiler 富集分析。

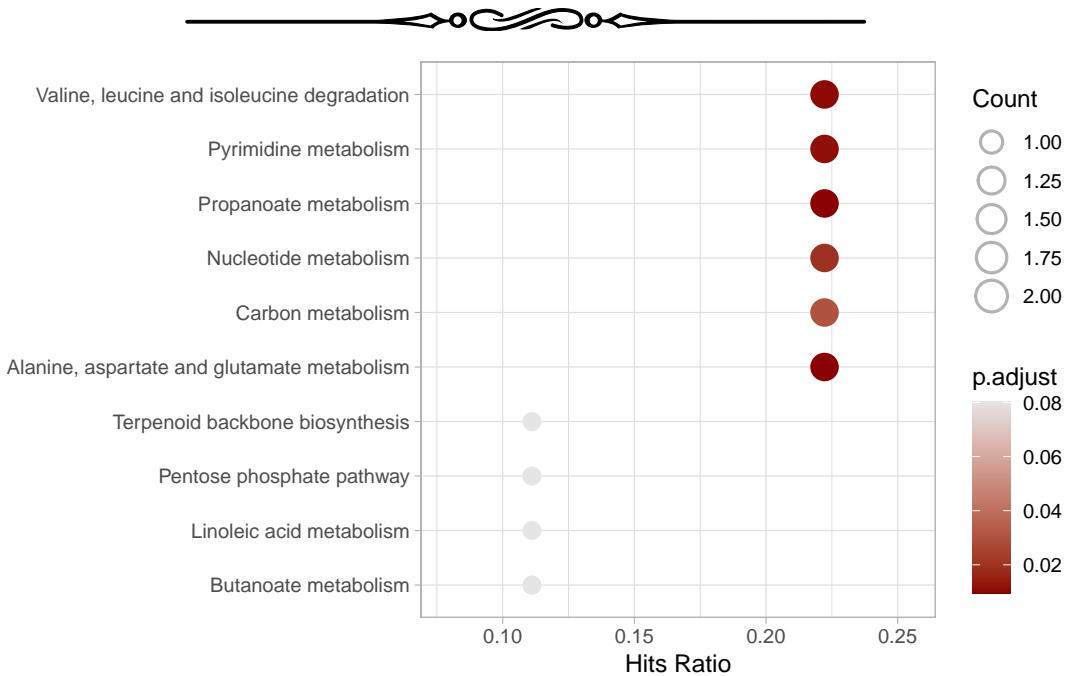


Figure 25: PROG KEGG enrichment

Fig. 25

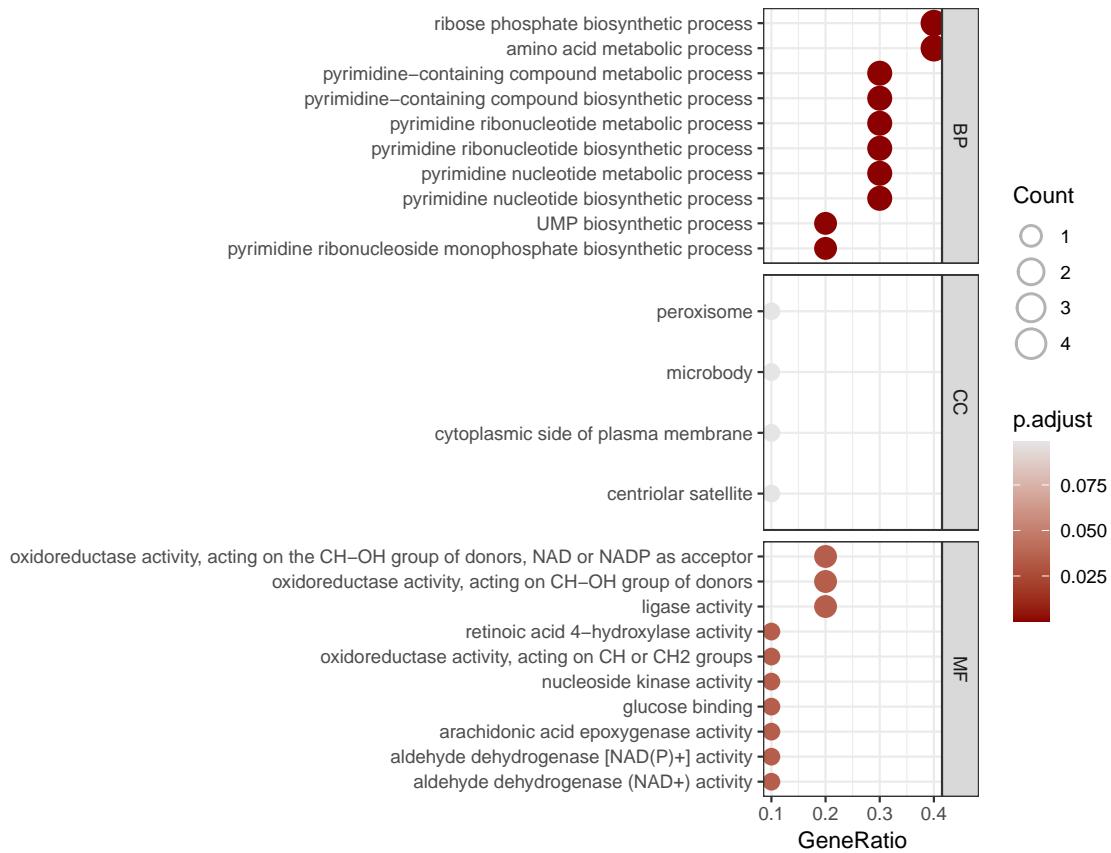


Figure 26: PROG GO enrichment



Fig. 26

4 总结

癌症的本质特征与癌细胞自身代谢的改变息息相关 (2008, **IF:48.8**, Q1, Cancer cell)⁹。对于骨肉瘤，目前仍缺少研究从单细胞水平探究癌症的代谢变化。本分析从单细胞水平鉴定的恶质细胞 (肿瘤细胞) 出发，分析正常细胞与癌症细胞之间的代谢通量差异，进而获取对应代谢模块的基因，建立预后模型，以代谢改变的角度，预测疾病的进展。模型以 TARGET-OS 数据集建立，进而在 GEO 数据库搜索了所有可用的带有生存信息的基因表达数据集，用以验证预后模型的可靠性。

Reference

1. Cao, Y., Wang, X. & Peng, G. SCSA: A cell type annotation tool for single-cell rna-seq data. *Frontiers in genetics* **11**, (2020).

2. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature Biotechnology* **39**, 599–608 (2021).
3. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics* **13**, 189–203 (2012).
4. Alghamdi, N. *et al.* A graph neural network model to estimate cell-wise metabolic flux using single-cell rna-seq data. *Genome research* **31**, 1867–1884 (2021).
5. Smyth, G. K. Limma: Linear models for microarray data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer-Verlag, 2005). doi:10.1007/0-387-29362-0_23.
6. Colaprico, A. *et al.* TCGAbiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research* **44**, (2015).
7. Wu, T. *et al.* ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).
8. Zhou, Y. *et al.* Single-cell rna landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nature communications* **11**, (2020).
9. Kroemer, G. & Pouyssegur, J. Tumor cell metabolism: Cancers achillesheel. *Cancer cell* **13**, 472–482 (2008).