生信分析报告

项目标题：骨肉瘤分析 ZDHHC 家族成员

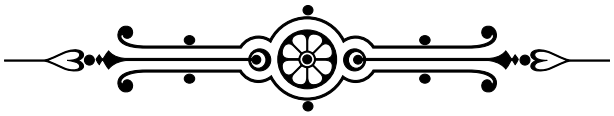
单　　号：BSHQ240303

分析人员：黄礼闯

分析类型：生信分析

委 托 人：张永旭

受 托 人：杭州铂赛生物科技有限公司



Contents

List of Figures

List of Tables

1 分析流程

1.1 需求

根据方案 2 中的设计，完成第一部分生信分析（骨肉瘤）：

1. GEPIA 等数据库，分析 ZDHHC 家族成员的差异表达
2. TCGA、TIMER、GSE 等数据集，分析 ZDHHC 家族成员的预后情况
3. 通过预后、表达的相关趋势，利用韦恩图，筛选明显上调的棕榈酰化酶 ZDHHC
4. 验证集分析：更换其他的数据库、GEO 数据集，证明 ZDHHC 明显高表达、预后较差。
5. 通过 HPA 数据库验证以上差异蛋白的 IHC 表达结果。

1.2 实际分析

1. GEO 数据库获取 Osteosarcoma 数据集，差异分析 Tumor vs Normal (GEPIA 使用的是 TCGA 数据，不包含 Osteosarcoma)
2. 使用 TARGET-OS 数据集，分析 ZDHHC 家族预后。
3. 筛选差异表达和预后显著的 ZDHHC 基因。
4. 基因较少，未能通过多个数据集的验证。
5. HPA 不包含筛选的 ZDHHC 的 Osteosarcoma 的数据。

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 TCGA 数据获取 (Dataset: OS)

以 R 包 `TCGAbiolinks` (2.34.0) (2015, **IF:16.6**, Q1, Nucleic Acids Research)[1] 获取 TCGA 数据集。

2.3 Survival 生存分析 (Dataset: OS)

去除了生存状态未知的数据。以 R 包 `survival` (3.7.0) 生存分析，以 R 包 `survminer` (0.5.0) 绘制生存曲线。以 R 包 `timeROC` (0.4) 绘制 1, 3, 5 年生存曲线。

2.4 GEO 数据获取 (Dataset: GEOOS2)

以 R 包 `GEOquery` (2.74.0) 获取 GSE99671 数据集。

2.5 Limma 差异分析 (Dataset: GEOOS2)

以 R 包 `limma` (3.62.1) (2005, **IF:**, ,)[2] `edgeR` (4.4.0) (, **IF:**, ,)[3] 进行差异分析。以 `edgeR::filterByExpr` 过滤 count 数量小于 10 的基因。以 `edgeR::calcNormFactors`，`limma::voom` 转化 count 数据为 log2 counts-

per-million (logCPM)。分析方法参考 https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html。随后，以公式 ~ 0 + group + pairs 创建设计矩阵 (design matrix) 用于线性分析。使用 `limma::lmFit`, `limma::contrasts.fit`, `limma::eBayes` 差异分析对比组：TUMOR vs NORMAL。以 `limma::topTable` 提取所有结果，并过滤得到 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

2.6 GEO 数据获取 (Dataset: GEOOS4)

以 R 包 `GEOquery` (2.74.0) 获取 GSE253548 数据集。

2.7 Biomart 基因注释 (Dataset: GEOOS4)

以 R 包 `biomaRt` (2.62.0) 对基因进行注释，获取各数据库 ID 或注释信息，以备后续分析。

2.8 Limma 差异分析 (Dataset: GEOOS4)

使用 `limma::lmFit`, `limma::contrasts.fit`, `limma::eBayes` 差异分析对比组：TUMOUR vs NORMAL。以 `limma::topTable` 提取所有结果，并过滤得到 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

3 分析结果

3.1 TARGET 数据获取 (OS)

获取 TARGET-OS 数据集，用于生存分析。

3.2 Survival 生存分析 (OS)

生存分析的统计结果见 Tab. 1

'OS Survival plots' 数据已全部提供。

(File path: `Figure+Table/OS-Survival-plots`)

Note: The directory 'Figure+Table/OS-Survival-plots' contains 30 files.

1. 1_ZDHHC6.pdf
2. 10_ZDHHC12.pdf
3. 11_ZDHHC3.pdf
4. 12_ZDHHC19.pdf
5. 13_ZDHHC16.pdf
6. ...

Table 1: OS Significant Survival PValue

name	pvalue
ZDHHC15	0.0123699184476175
ZDHHC7	0.0487669724778526
ZDHHC3	0.00170983043763898
ZDHHC23	0.0298228620445287

Table 1 (下方表格) 为表格 OS Significant Survival PValue 概览。

(File path: `Figure+Table/OS-Significant-Survival-PValue.csv`)

注：表格共有 4 行 2 列，以下预览的表格可能省略部分数据；含有 4 个唯一 'name'。

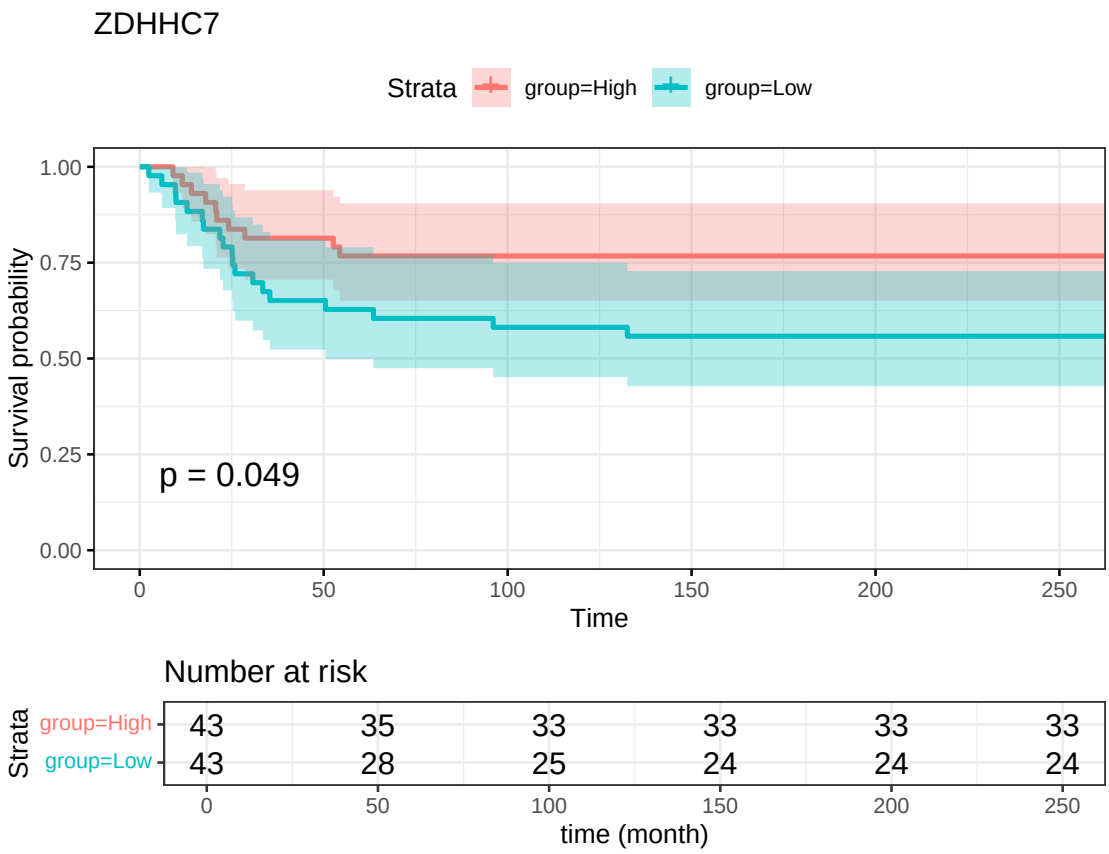


Figure 1: OS survival curve of ZDHHC7

Figure 1 (下方图) 为图 OS survival curve of ZDHHC7 概览。

(File path: `Figure+Table/OS-survival-curve-of-ZDHHC7.pdf`)

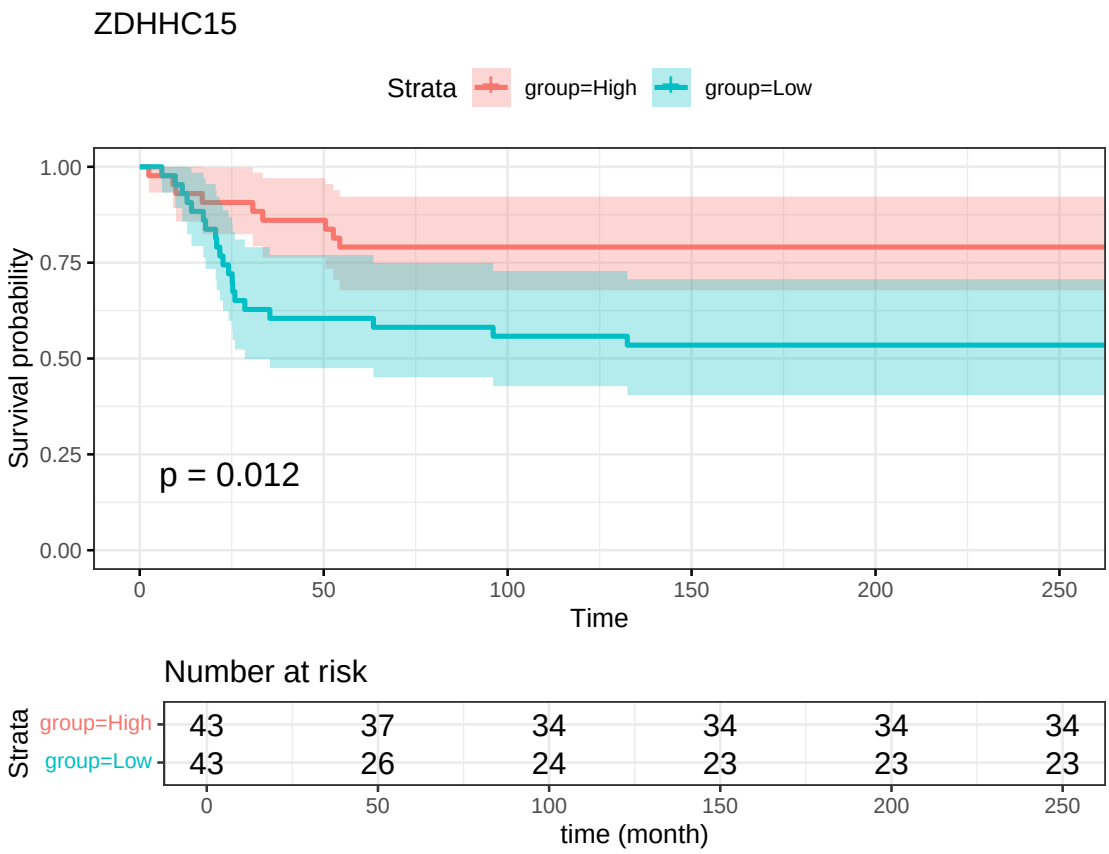


Figure 2: OS survival curve of ZDHHC15

Figure 2 (下方图) 为图 OS survival curve of ZDHHC15 概览。

(File path: `Figure+Table/OS-survival-curve-of-ZDHHC15.pdf`)

3.3 GEO 数据获取 (GEOOS2)

获取 GEO 数据，用于差异分析。

Data Source ID :

GSE99671

data_processing :

Color-space base calling

data_processing.1 :

Mapping, alignment with Lifescope

data_processing.2 :

Lifescope transcriptome workflow

data_processing.3 :

Genome_build: hg19

(Others) :

...

(见 `Figure+Table/GEOOS2-GSE99671-content`)

3.4 Limma 差异分析 (GEOOS2)

用到的样本见 Tab. 3，差异分析结果见 Fig. 3

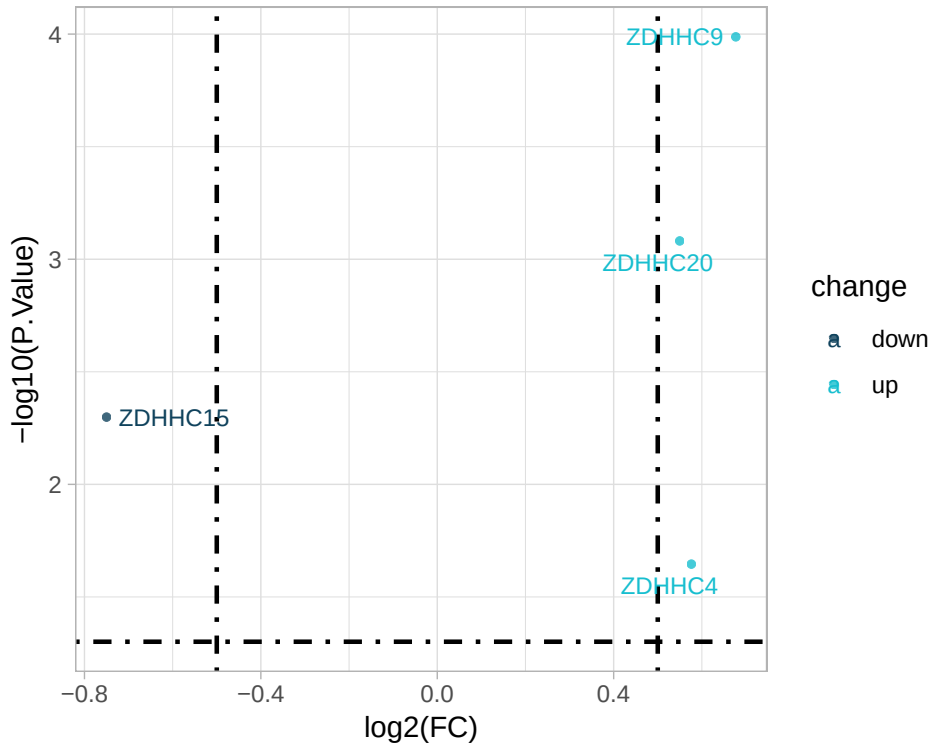


Figure 3: GEOOS2 TUMOR vs NORMAL

Figure 3 (下方图) 为图 GEOOS2 TUMOR vs NORMAL 概览。

(File path: `Figure+Table/GEOOS2-TUMOR-vs-NORMAL.pdf`)

P.Value cut-off :

0.05

Log2(FC) cut-off :

0.5

(See: `Figure+Table/GEOOS2-TUMOR-vs-NORMAL-content`)

Table 2: GEOOS2 data TUMOR vs NORMAL

rownames	V1	symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
23851	23851	ZDHHC9	0.6769...	17.023...	4.3671...	0.0001...	0.0019...	1.1699...
14111	14111	ZDHHC20	0.5497...	16.715...	3.6502...	0.0008...	0.0078...	-0.782...
23525	23525	ZDHHC15	-0.749...	13.234...	-2.989...	0.0050...	0.0318...	-2.249...
7957	7957	ZDHHC4	0.5761...	13.974...	2.3829...	0.0226...	0.0715...	-3.438...

Table 2 (下方表格) 为表格 GEOOS2 data TUMOR vs NORMAL 概览。

(File path: `Figure+Table/GEOOS2-data-TUMOR-vs-NORMAL.csv`)

注：表格共有 4 行 9 列，以下预览的表格可能省略部分数据；含有 4 个唯一 'rownames；含有 4 个唯一 'symbol'。

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

Table 3: GEOOS2 metadata of used sample

sample	group	lib.size	norm.f...	pairs	batch	rownames	title	barcod...	chemot...
OSVN001T	TUMOR	950659	1	BC1	B	GSM264...	OSVN00...	BC1	NA
OSVN001N	NORMAL	1962162	1	BC2	L	GSM264...	OSVN00...	BC2	NA
OSDN001N	NORMAL	3398664	1	BC3	B	GSM264...	OSDN00...	BC3	NA
OSDN001T	TUMOR	4601178	1	BC4	M	GSM264...	OSDN00...	BC4	NA
OSVN003N	NORMAL	4462111	1	BC5	L	GSM264...	OSVN00...	BC5	NA
...	...	...	...	...	...	...	...	...	...

Table 3 (下方表格) 为表格 GEOOS2 metadata of used sample 概览。

(File path: `Figure+Table/GEOOS2-metadata-of-used-sample.csv`)

> 注：表格共有 36 行 15 列，以下预览的表格可能省略部分数据；含有 36 个唯一 'sample'。

3.5 GEO 数据获取 (GEOOS4)

> **Data Source ID :**
>
> GSE253548
>
> **data_processing :**
>
> Illumina DRAGEN BCL, then fastq files were analysed with salmon to get counts data. The counts were imported to DESeq2.
>
> **data_processing.1 :**
>
> Assembly: GRCh38
>
> **data_processing.2 :**
>
> Supplementary files format and content: DESeq2 normalised counts

(见 `Figure+Table/GEOOS4-GSE253548-content`)

3.6 Biomart 基因注释 (GEOOS4)

由于该数据集不包含 Symbol 等基因注释信息，因此，使用 biomaRt 对其注释。

3.7 Limma 差异分析 (GEOOS4)

用到样本见 Tab. 5，差异分析结果见 Fig. 4。

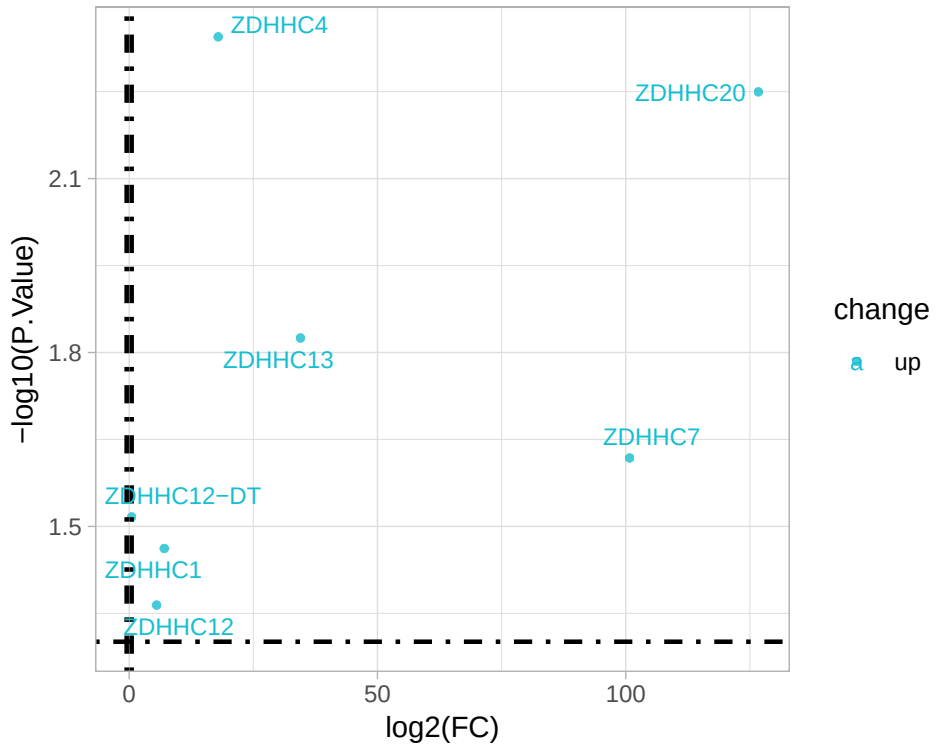


Figure 4: GEOOS4 TUMOUR vs NORMAL

Figure 4 (下方图) 为图 GEOOS4 TUMOUR vs NORMAL 概览。

(File path: `Figure+Table/GEOOS4-TUMOUR-vs-NORMAL.pdf`)

P.Value cut-off :

0.05

Log2(FC) cut-off :

0.5

(See: `Figure+Table/GEOOS4-TUMOUR-vs-NORMAL-content`)

Table 4: GEOOS4 data TUMOUR vs NORMAL

rownames	logFC	AveExpr	t	P.Value	adj.P.Val	B	hgnc_s…
ENSG00…	17.939…	12.858…	2.9137…	0.0045…	0.0928…	-2.073…	ZDHHC4
ENSG00…	126.70…	98.071…	2.8381…	0.0056…	0.0928…	-2.252…	ZDHHC20
ENSG00…	34.500…	25.344…	2.4818…	0.0149…	0.1645…	-3.042…	ZDHHC13
ENSG00…	100.79…	83.675…	2.2951…	0.0240…	0.1898…	-3.420…	ZDHHC7
ENSG00…	0.5141…	1.2910…	2.1994…	0.0304…	0.1898…	-3.604…	ZDHHC1…
…	…	…	…	…	…	…	…

Table 4 (下方表格) 为表格 GEOOS4 data TUMOUR vs NORMAL 概览。

(File path: `Figure+Table/GEOOS4-data-TUMOUR-vs-NORMAL.csv`)

注：表格共有 7 行 8 列，以下预览的表格可能省略部分数据；含有 7 个唯一 'rownames'；含有 7 个唯一 'hgnc_symbol'。

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

Table 5: GEOOS4 metadata of used sample

sample	group	rownames	title	ageatd…	diseas…	Sex.ch1	status…	tissue…	treatm…
Q01B03…	TUMOUR	GSM802…	Q01B03…	16	TUMOUR	F	deceased	bone	chemo
Q02B03…	NORMAL	GSM802…	Q02B03…	14	NORMAL	F	deceased	bone	chemo
Q02B03…	TUMOUR	GSM802…	Q02B03…	14	TUMOUR	F	deceased	bone	chemo
Q04B02…	NORMAL	GSM802…	Q04B02…	16	NORMAL	M	alive	bone	chemo
Q04B02…	TUMOUR	GSM802…	Q04B02…	16	TUMOUR	M	alive	bone	chemo

sample	group	rownames	title	ageatd...	diseas...	Sex.ch1	status...	tissue...	treatm...
...	...	...	...	...	...	...	...	...	...

Table 5 (下方表格) 为表格 GEOOS4 metadata of used sample 概览。

(File path: `Figure+Table/GEOOS4-metadata-of-used-sample.csv`)

注：表格共有 90 行 10 列，以下预览的表格可能省略部分数据；含有 90 个唯一 'sample'。

3.8 预后显著且差异表达的 ZDHHC

3.8.1 预后分析 + GEO2 (GSE99671)

以生存分析显著的基因 Tab. 1 ，与差异分析结果 Tab. 2 取交集，见 Fig. 5 。交集基因生存分析见 Fig. 2。

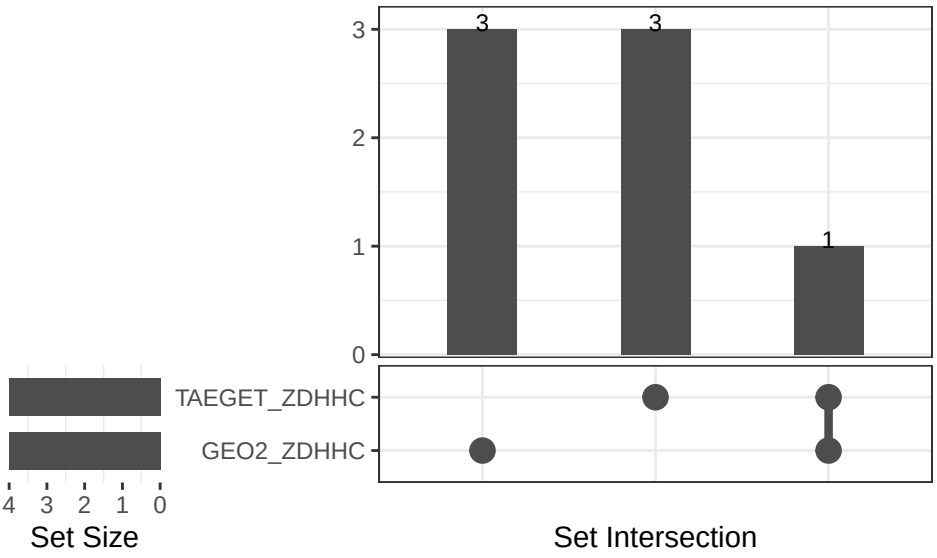


Figure 5: Intersection of GEO2 ZDHHC with TAEGET ZDHHC

Figure 5 (下方图) 为图 Intersection of GEO2 ZDHHC with TAEGET ZDHHC 概览。

(File path: `Figure+Table/Intersection-of-GEO2-ZDHHC-with-TAEGET-ZDHHC.pdf`)

All_intersection :

ZDHHC15

(See: `Figure+Table/Intersection-of-GEO2-ZDHHC-with-TAEGET-ZDHHC-content`)

3.8.2 预后分析 + GEO4 (GSE253548)

以生存分析结果 Tab. 1，与差异分析结果 Tab. 4 取交集，结果见 Fig. 6。交集基因生存分析图见 Fig. 1。

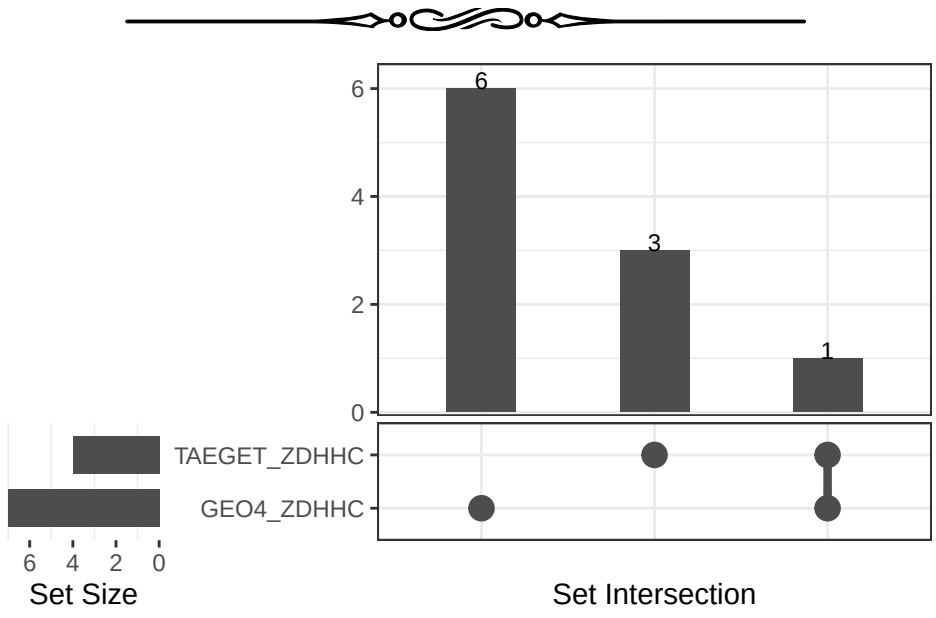


Figure 6: Intersection of GEO4 ZDHHC with TAEGET ZDHHC

Figure 6 (下方图) 为图 Intersection of GEO4 ZDHHC with TAEGET ZDHHC 概览。

(File path: `Figure+Table/Intersection-of-GEO4-ZDHHC-with-TAEGET-ZDHHC.pdf`)

All_intersection :

ZDHHC7

(See: `Figure+Table/Intersection-of-GEO4-ZDHHC-with-TAEGET-ZDHHC-content`)

3.9 HPA 数据库

HPA 数据库不包含上述基因的 Osteosarcoma 数据。

4 总结

按实际分析的结果，筛选的两个基因见 Fig. 5，Fig. 6

Reference

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

2. 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.

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