**生信分析报告**

**项目标题： 高级别浆液性卵巢癌 ;**

**单 号： BSXN240936 ;**

**分析人员： 黄礼闯 ;**

**分析类型： 生信协助 ;**

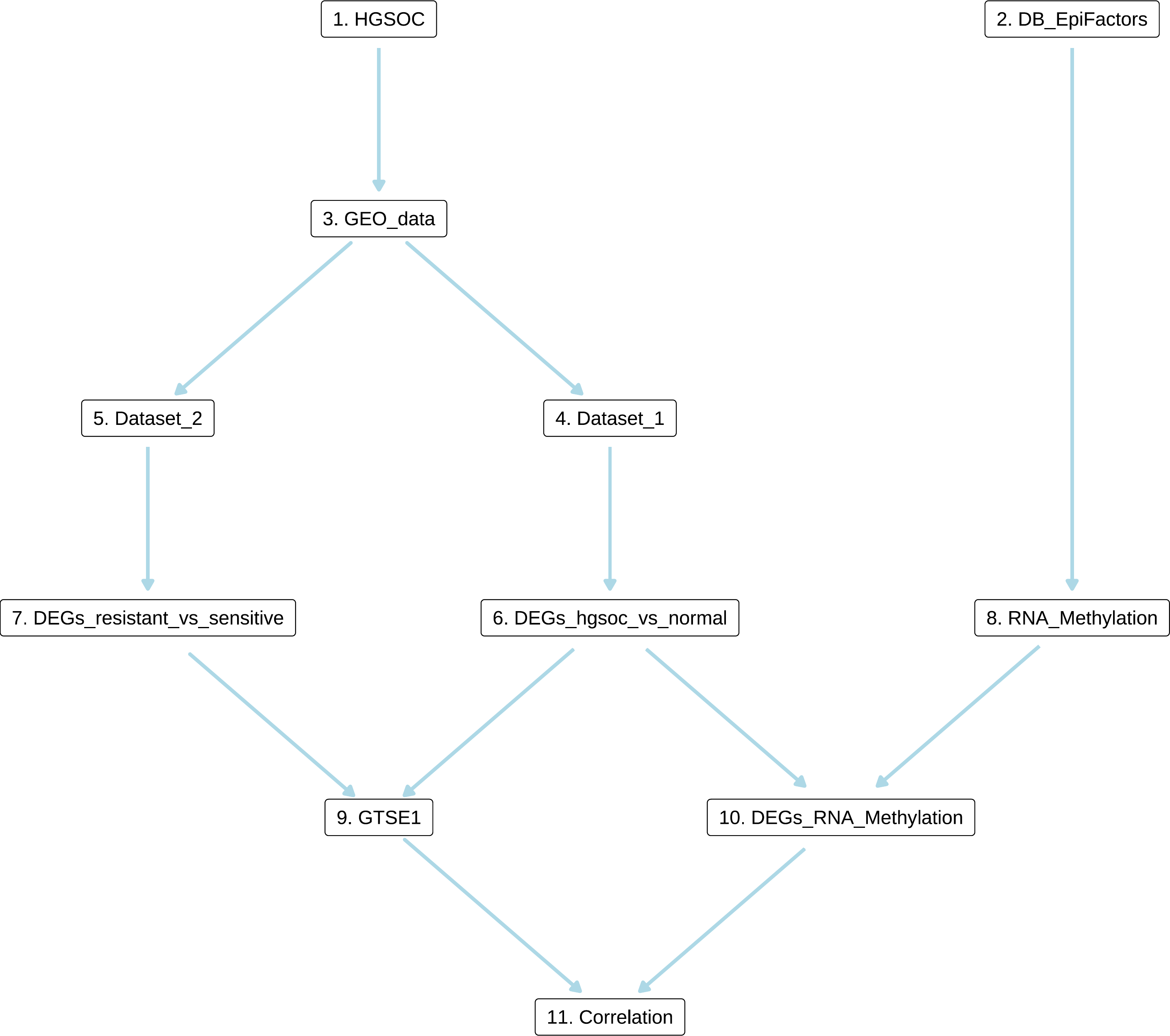
**委 托 人： 李扬 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程

## 1.1 需求

* GTSE1 在 HGSOC 中高表达促进肿瘤增值与迁移。 GTSE1 与化疗耐药有关。在卵巢癌 (HGSOC) 中的表达，在化疗耐药卵巢癌数据集中的表达。
* GTSE1 的 mRNA 序列上的 m6A 甲基化，GTSE1可能主要被ALKBH5、IGF2BP1、METTL14、METTL3、YTHDF1、YTHDF2。 GTSE1与甲基化相关酶的表达进行相关性分析



**Fig.** **1** Route

**(File path: Figure+Table/1.1.0\_需求/Route.pdf)**

# 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: OV)

以 R 包 TCGAbiolinks (2.35.1) (2015, **IF:16.6**, Q1, Nucleic Acids Research)1 获取 TCGA-OV 数据集。

## 2.3 GSE 数据搜索 (Dataset: HGSOC)

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: ((high-grade serous ovarian cancer[Description] AND (normal[Description]) AND ((6:1000[Number of Samples]) AND (GSE[Entry Type]) AND (Homo Sapiens[Organism]))。 以正则匹配，滤除包含 ‘single cell’ 或 ‘scRNA’ 的数据例。仅查询临床样本信息，因此滤除匹配到 ‘cells’, ‘cell type’ 或 ‘cell line’ 的实验数据例。 此外，去除了以特定 Marker 细胞类型为研究对象的数据例 (CD4、CD8 T 细胞等)。仅获取类型包含 ‘Expression profiling by high throughput sequencing’ 或 ‘Expression profiling by array’ 的数据例。

## 2.4 GEO 数据获取 (Dataset: HGSOC)

以 R 包 GEOquery (2.74.0) 获取 GSE146553 数据集。

## 2.5 Limma 差异分析 (Dataset: HGSOC)

使用 log2 和 limma::normalizeBetweenArrays 对数据标准化。 以 limma (3.62.1) (2005)2 差异分析。创建设计矩阵，对比矩阵，差异分析：hgsoc vs normal。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 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

## 2.6 GSE 数据搜索

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: ((ovarian cancer[Description] AND (Drug resistance[Description]) AND ((6:1000[Number of Samples]) AND (GSE[Entry Type]) AND (Homo Sapiens[Organism]))。

## 2.7 GEO 数据获取 (Dataset: OV\_GSE173579)

以 R 包 GEOquery (2.74.0) 获取 GSE173579 数据集。 以 GEOquery:::getRNASeqQuantResults 获取 RNA count 数据 (NCBI-generated data, 参考 <https://www.ncbi.nlm.nih.gov/geo/info/rnaseqcounts.html>) 以及基因注释。

## 2.8 Limma 差异分析 (Dataset: OV\_GSE173579)

以 R 包 edgeR (4.4.0) ()3 对数据预处理。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)。 以 limma (3.62.1) (2005)2 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：Carboplatin\_high\_Resistant\_SKOV3 vs Non\_resistance\_SKOV3。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 P.Value 小于 0.05，|Log2(FC)| 大于 0.3 的统计结果。

## 2.9 GEO 数据获取 (Dataset: OC\_GSE140996)

以 R 包 GEOquery (2.74.0) 获取 GSE140996 数据集。

## 2.10 Limma 差异分析 (Dataset: OC\_GSE140996)

以 limma (3.62.1) (2005)2 差异分析。创建设计矩阵，对比矩阵，差异分析：W1CR\_none vs W1\_none。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 P.Value 小于 0.05，|Log2(FC)| 大于 0.3 的统计结果。

## 2.11 GEO 数据获取 (Dataset: OC\_GSE143152)

以 R 包 GEOquery (2.74.0) 获取 GSE143152 数据集。 以 GEOquery:::getRNASeqQuantResults 获取 RNA count 数据 (NCBI-generated data, 参考 <https://www.ncbi.nlm.nih.gov/geo/info/rnaseqcounts.html>) 缺失样本: GSM4251001 (‘NCBI-generated data’ 缺失样本计数数据的原因包括运行未通过 50% 的对齐率或由于技术原因处理失败) 以及基因注释。

## 2.12 Limma 差异分析 (Dataset: OC\_GSE143152)

以 R 包 edgeR (4.4.0) ()3 对数据预处理。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)。 以 limma (3.62.1) (2005)2 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：prexasertib\_resistant\_Kuramochi vs parental\_Kuramochi, prexasertib\_resistant\_OV vs parental\_OV, prexasertib\_resistant\_EFO vs parental\_EFO, prexasertib\_resistant\_OVSAHO vs parental\_OVSAHO。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 P.Value 小于 0.05，|Log2(FC)| 大于 0.3 的统计结果。

## 2.13 GEO 数据获取 (Dataset: OC\_GSE285737)

以 R 包 GEOquery (2.74.0) 获取 GSE285737 数据集。

## 2.14 Limma 差异分析 (Dataset: OC\_GSE285737)

以 R 包 edgeR (4.4.0) ()3 对数据预处理。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)。 以 limma (3.62.1) (2005)2 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：Carboplatin\_resistant\_OVCAR\_3 vs OVCAR\_3。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 P.Value 小于 0.05，|Log2(FC)| 大于 0.3 的统计结果。

## 2.15 GEO 数据获取 (Dataset: HGSOC\_GSE141630)

以 R 包 GEOquery (2.74.0) 获取 GSE141630 数据集。

## 2.16 Limma 差异分析 (Dataset: HGSOC\_GSE141630)

以 limma (3.62.1) (2005)2 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：carboplatin\_resistant vs sensitive。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 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

## 2.17 EpiFactors 表观遗传调控因子数据获取 (Dataset: M6A)

从数据库 EpiFactors (2023, **IF:16.6**, Q1, Nucleic acids research)4 获取表观遗传调控蛋白的数据。

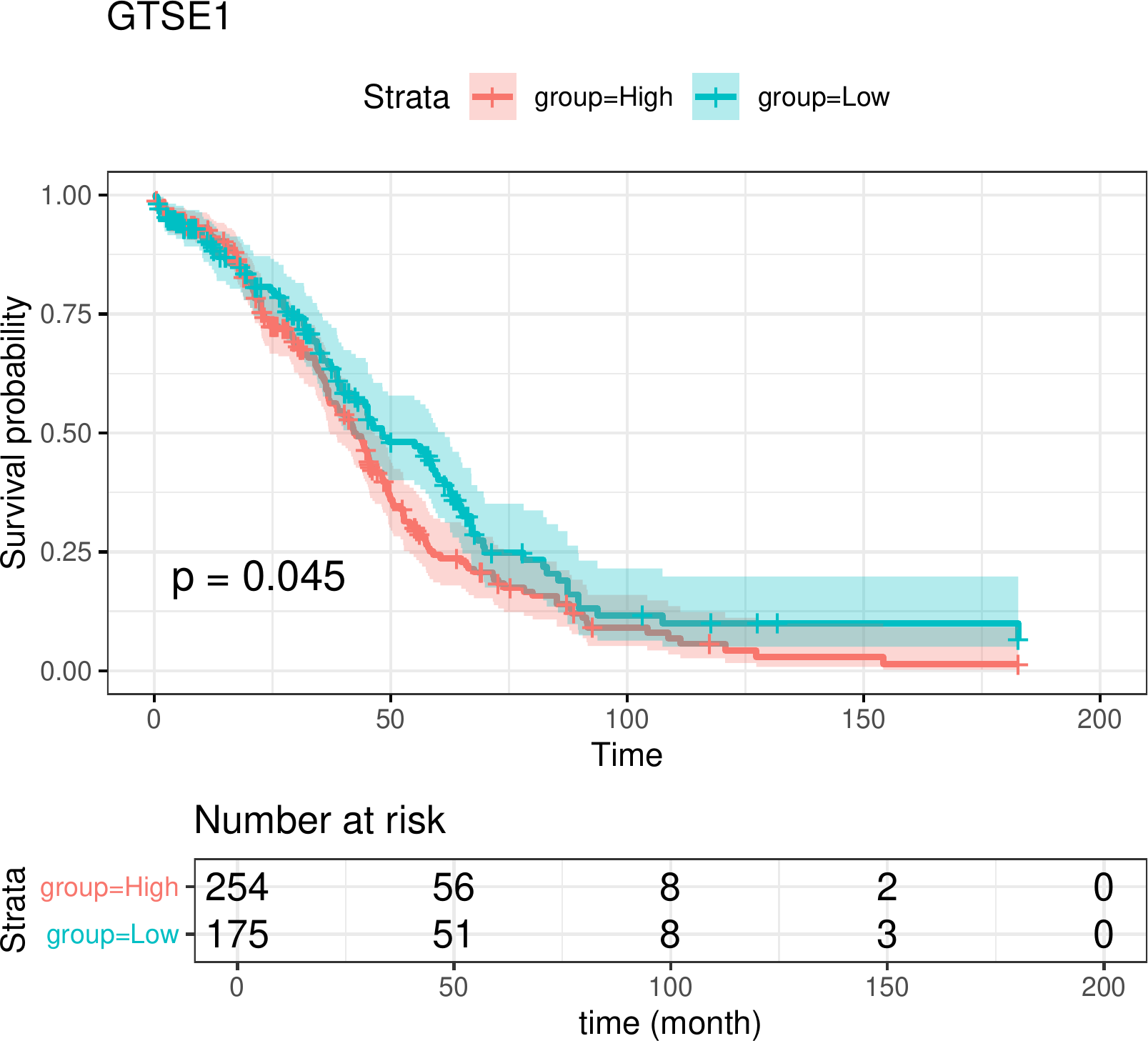
# 3 分析结果

## 3.1 TCGA 数据获取 (OV)

获取 TCGA-OV 数据。

## 3.2 Survival 生存分析 (OV)

Fig. **[2](#OV-survival-curve-of-GTSE1)** 为 GTSE1 生存曲线。



**Fig.** **2** OV survival curve of GTSE1

**(File path: Figure+Table/3.2.0\_Survival\_生存分析\_(OV)/OV-survival-curve-of-GTSE1.pdf)**

## 3.3 GSE 数据搜索 (HGSOC)

以 Entrez Direct (EDirect) 搜索 GEO 数据库 (检索条件见方法章节) 。

**Tab.** **1** High grade serous ovarian cancer EDirect query

| GSE | Title | Summary | Taxon | GdsType |
| --- | --- | --- | --- | --- |
| GSE209964 | Gene count level ... | Transcriptomic pr... | Homo sapiens | Expression profil... |
| GSE181955 | Gene expression o... | We report RNA seq... | Homo sapiens | Expression profil... |
| GSE146556 | Global DNA hypome... | This SuperSeries ... | Homo sapiens | Methylation profi... |
| GSE146553 | Global DNA hypome... | Comparison of DNA... | Homo sapiens | Expression profil... |
| GSE135886 | LncRNA and mRNA e... | Ovarian cancer is... | Homo sapiens | Expression profil... |
| ... | ... | ... | ... | ... |

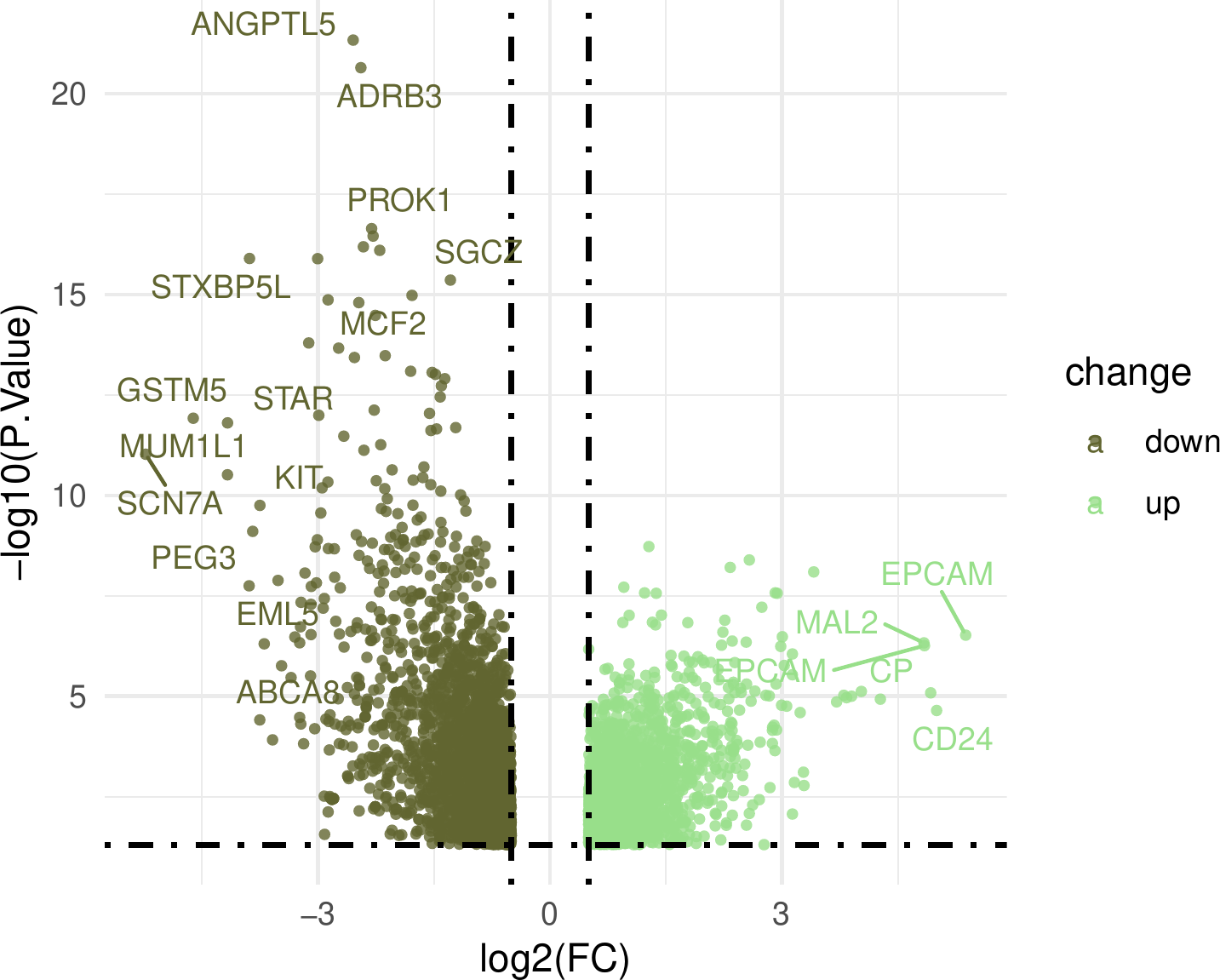
**(File path: Figure+Table/3.3.0\_GSE\_数据搜索\_(HGSOC)/High-grade-serous-ovarian-cancer-EDirect-query.xlsx)**

## 3.4 GEO 数据获取 (HGSOC)

## 3.5 Limma 差异分析 (HGSOC)

筛选 group 为 “normal”, “hgsoc”，最终得到 43 例数据。样本分组：hgsoc (n=40) , normal (n=3) , Others (n=0) 。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：hgsoc vs normal。(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 DEGs 统计：up (n=1992) , down (n=2347)

Fig. **[4](#HGSOC-Box-Plot-Of-DEGs)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。

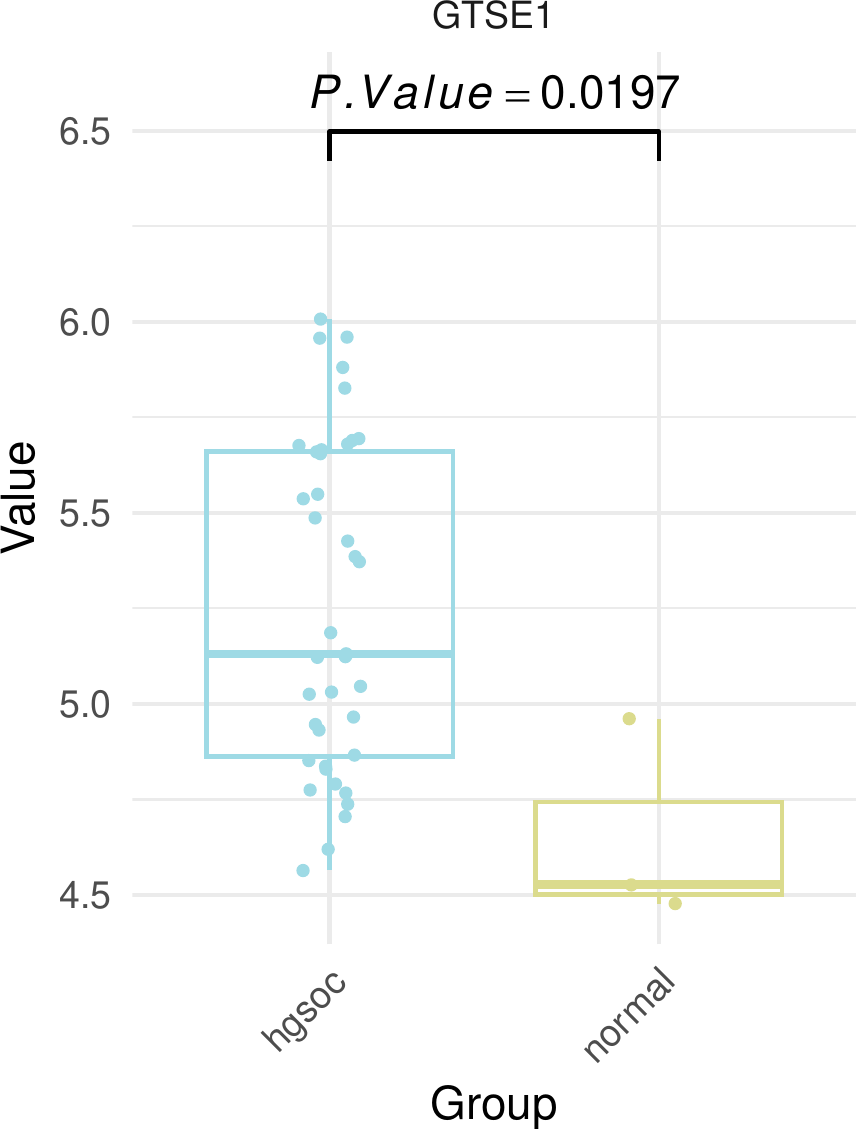


**Fig.** **3** HGSOC hgsoc vs normal

**(File path: Figure+Table/3.5.0\_Limma\_差异分析\_(HGSOC)/HGSOC-hgsoc-vs-normal.pdf)**

* P.Value cut-off: 0.05
* Log2(FC) cut-off: 0.5

**(See: Figure+Table/3.5.0\_Limma\_差异分析\_(HGSOC)/HGSOC-hgsoc-vs-normal-content)**



**Fig.** **4** HGSOC Box Plot Of DEGs

**(File path: Figure+Table/3.5.0\_Limma\_差异分析\_(HGSOC)/HGSOC-Box-Plot-Of-DEGs.pdf)**

## 3.6 Drug resistant

### 3.6.1 GSE 数据搜索 (HGSOC\_RES)

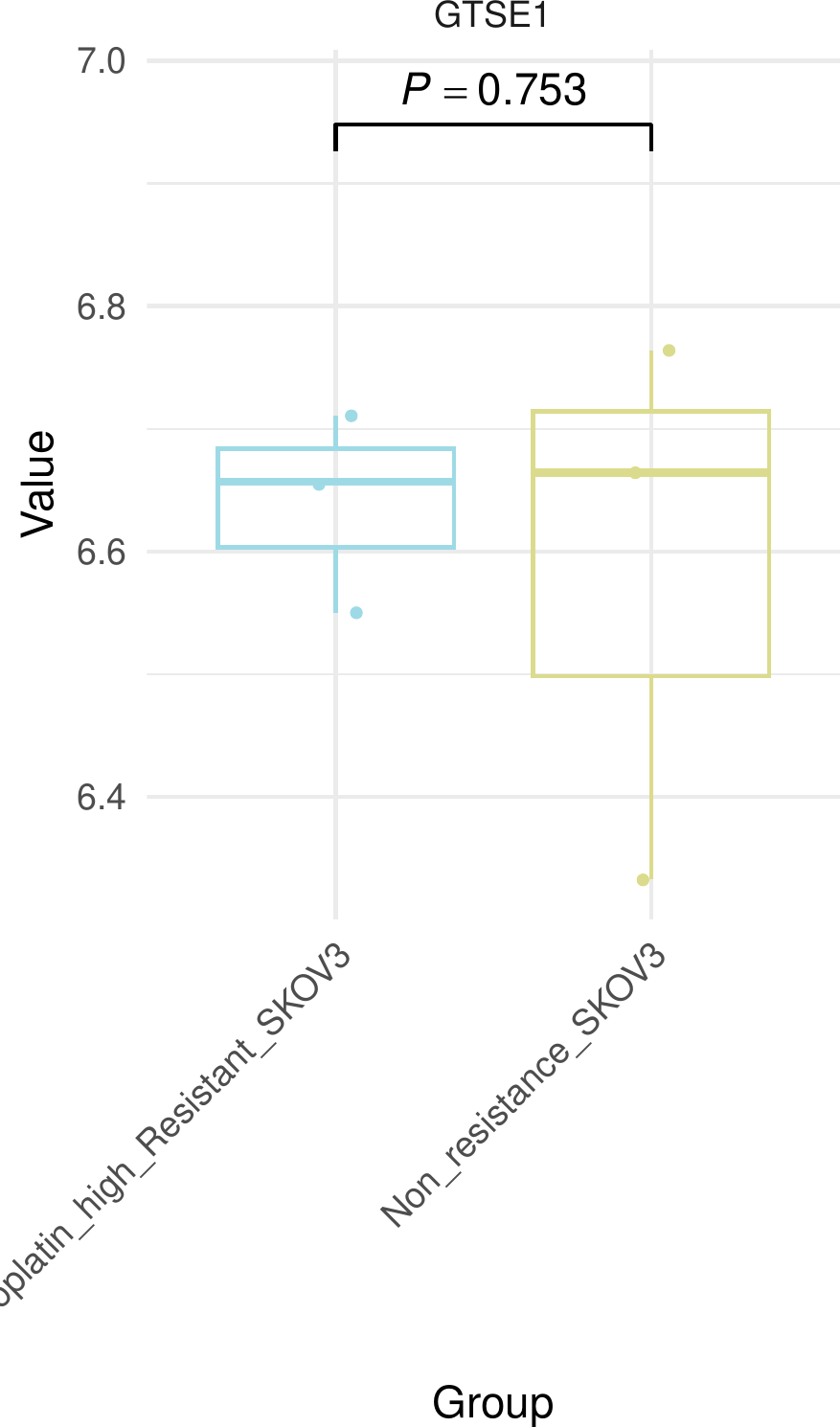
### 3.6.2 GEO 数据获取 (OV\_GSE173579)

以 GEOquery 获取 GSE173579 的数据信息。

### 3.6.3 Limma 差异分析 (OV\_GSE173579)

以 edgeR 将GSE173579 RNA-seq 数据标准化 (详见方法章节)。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Carboplatin\_high\_Resistant\_SKOV3 vs Non\_resistance\_SKOV3。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。

Fig. **[5](#OV-GSE173579-Box-Plot-Of-DEGs)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。



**Fig.** **5** OV GSE173579 Box Plot Of DEGs

**(File path: Figure+Table/3.6.3\_Limma\_差异分析\_(OV\_GSE173579)/OV-GSE173579-Box-Plot-Of-DEGs.pdf)**

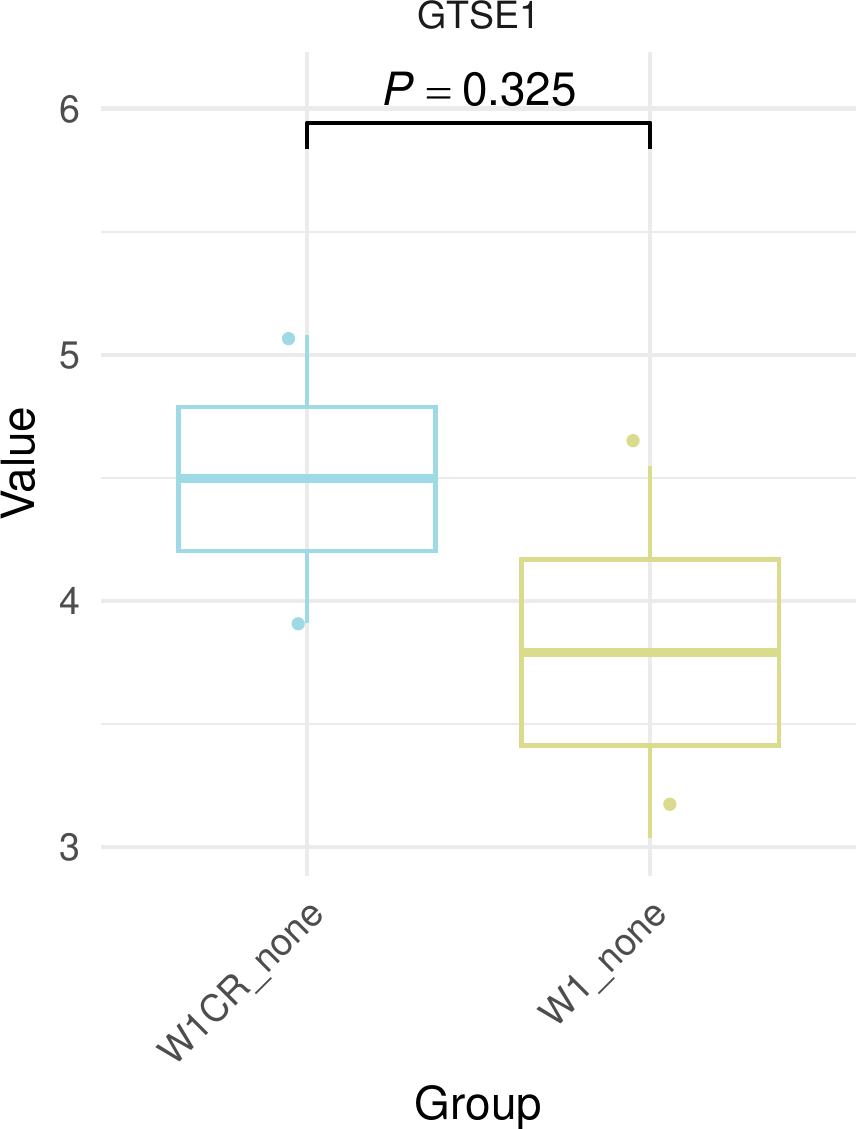
### 3.6.4 GEO 数据获取 (OC\_GSE140996)

以 GEOquery 获取 GSE140996 的数据信息。

### 3.6.5 Limma 差异分析 (OC\_GSE140996)

匹配 group 中包含“none”的描述，最终得到 4 例数据。样本分组：W1\_none (n=2) , W1CR\_none (n=2) 。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：W1CR\_none vs W1\_none。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。

Fig. **[6](#OC-GSE140996-Box-Plot-Of-DEGs)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。



**Fig.** **6** OC GSE140996 Box Plot Of DEGs

**(File path: Figure+Table/3.6.5\_Limma\_差异分析\_(OC\_GSE140996)/OC-GSE140996-Box-Plot-Of-DEGs.pdf)**

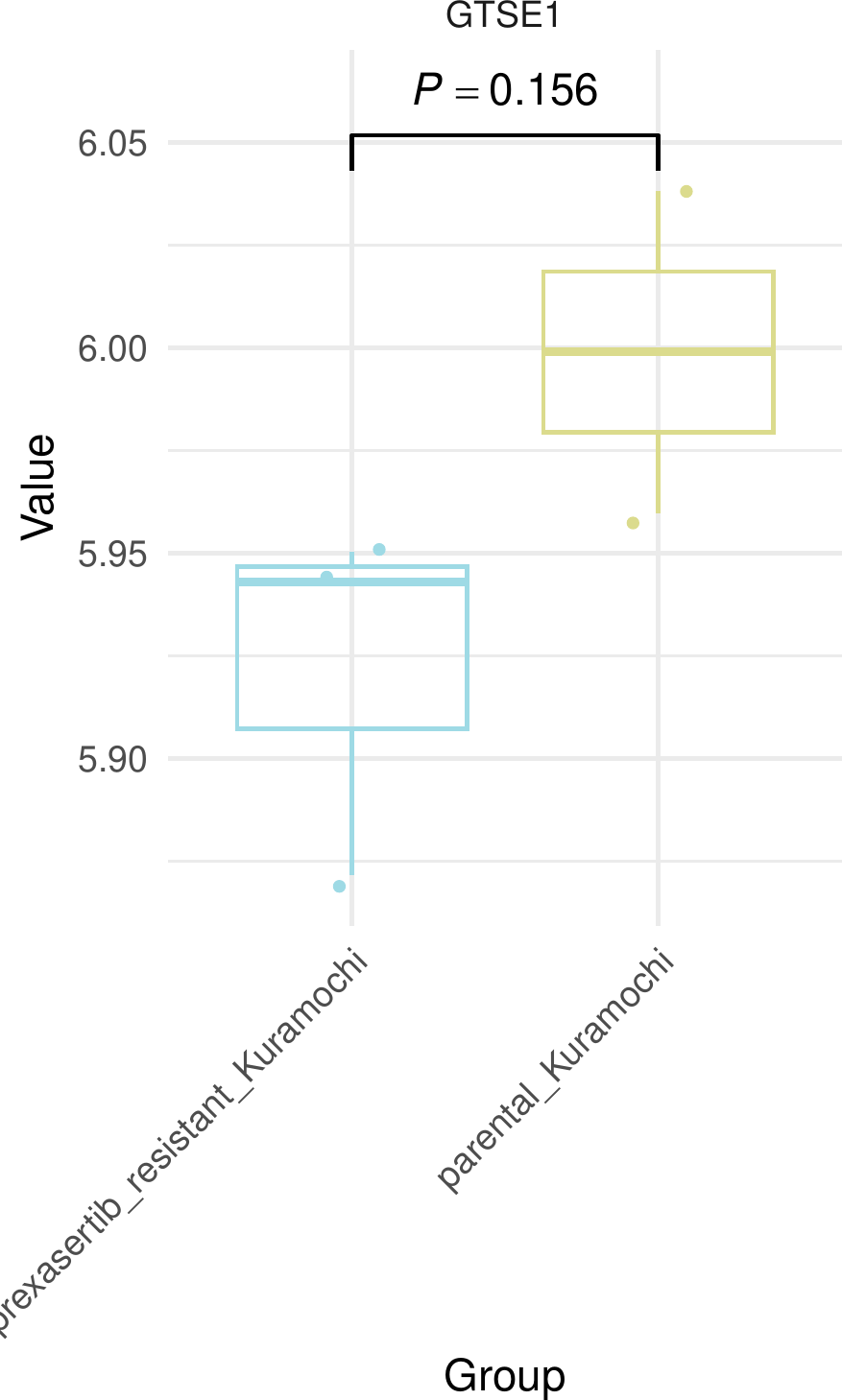
### 3.6.6 GEO 数据获取 (OC\_GSE143152)

以 GEOquery 获取 GSE143152 的数据信息。

### 3.6.7 Limma 差异分析 (OC\_GSE143152)

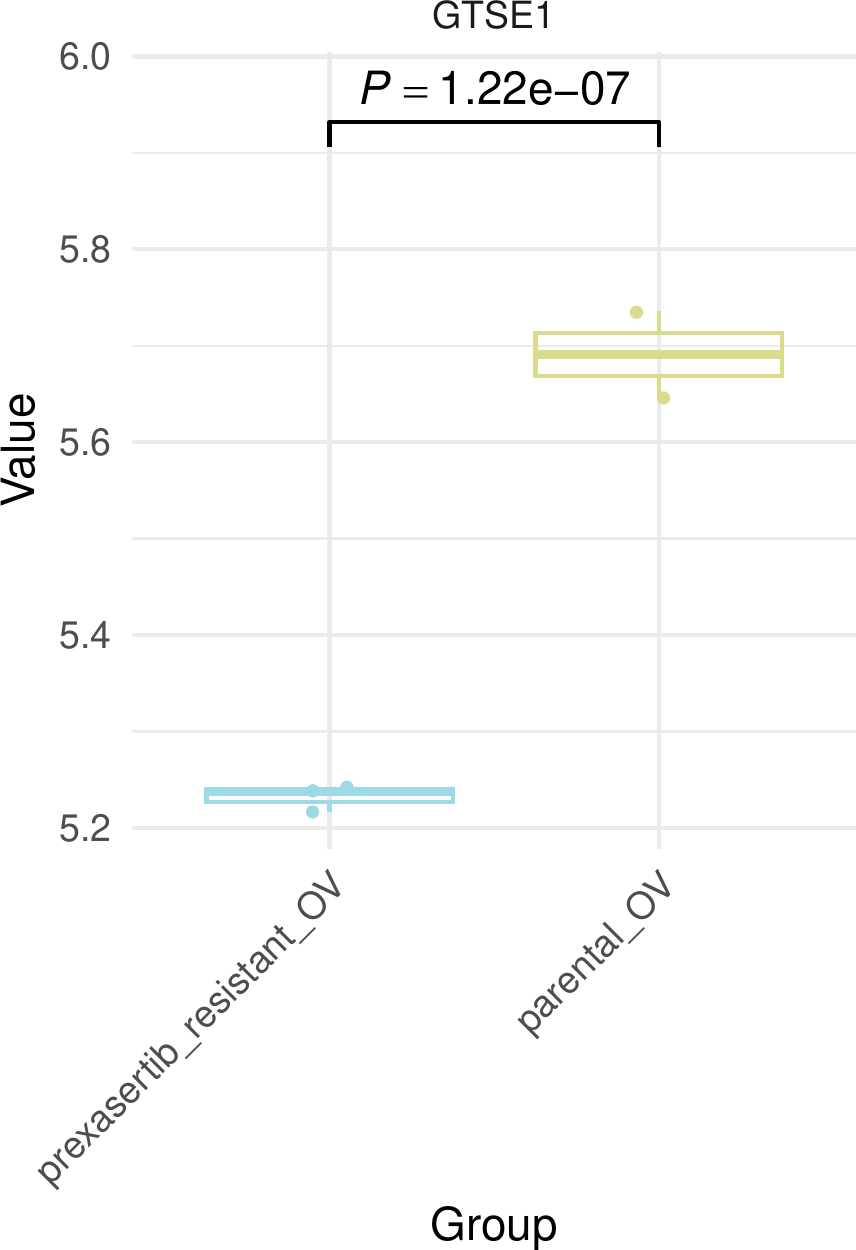
以 edgeR 将GSE143152 RNA-seq 数据标准化 (详见方法章节)。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：prexasertib\_resistant\_Kuramochi vs parental\_Kuramochi, prexasertib\_resistant\_OV vs parental\_OV, prexasertib\_resistant\_EFO vs parental\_EFO, prexasertib\_resistant\_OVSAHO vs parental\_OVSAHO。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。

Fig. **[7](#OC-GSE143152-Box-Plot-Of-DEGs-1)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。 Fig. **[8](#OC-GSE143152-Box-Plot-Of-DEGs-2)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。 Fig. **[9](#OC-GSE143152-Box-Plot-Of-DEGs-3)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。 Fig. **[10](#OC-GSE143152-Box-Plot-Of-DEGs-4)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。 Tab. **[2](#OC-GSE143152-Statistic-of-Focused-genes-extra)** 为聚焦分析的基因的统计附表。



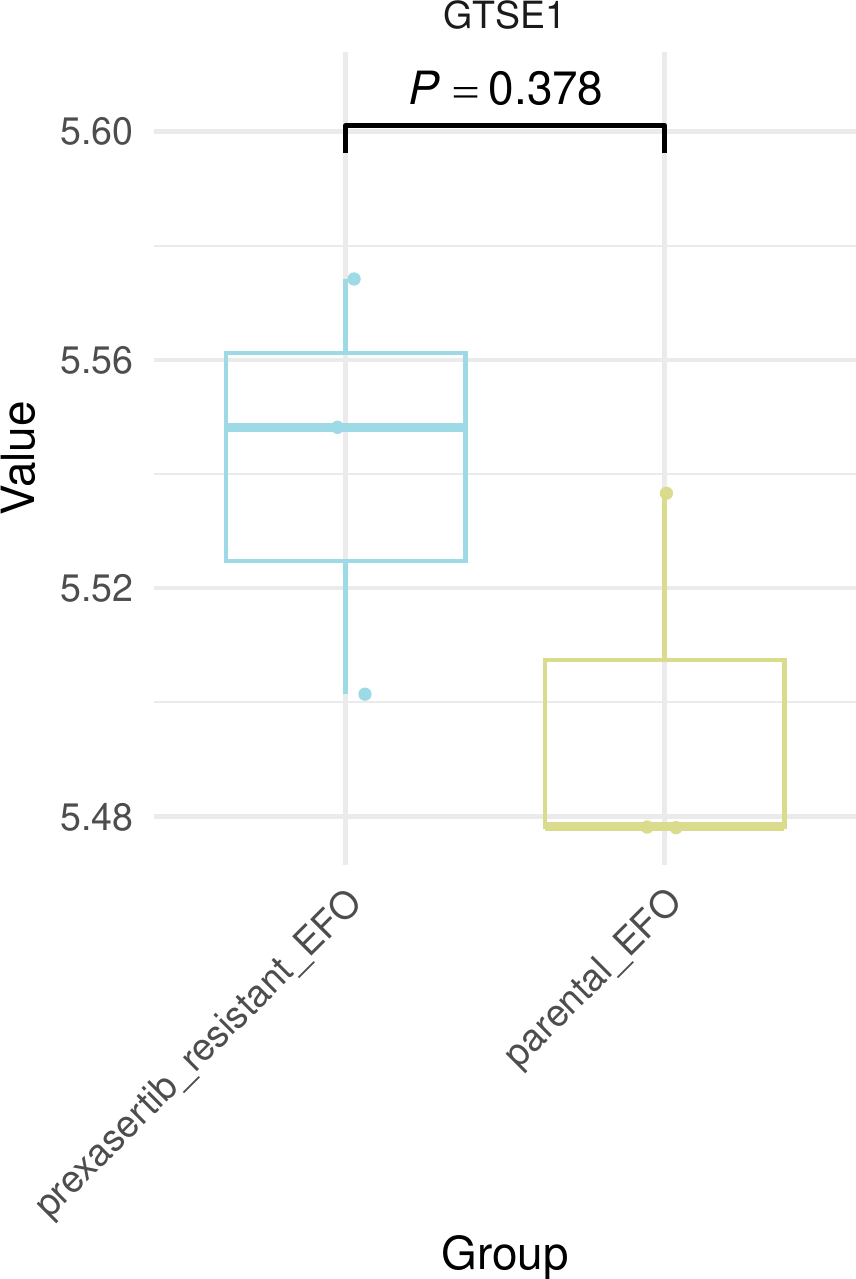
**Fig.** **7** OC GSE143152 Box Plot Of DEGs 1

**(File path: Figure+Table/3.6.7\_Limma\_差异分析\_(OC\_GSE143152)/OC-GSE143152-Box-Plot-Of-DEGs-1.pdf)**



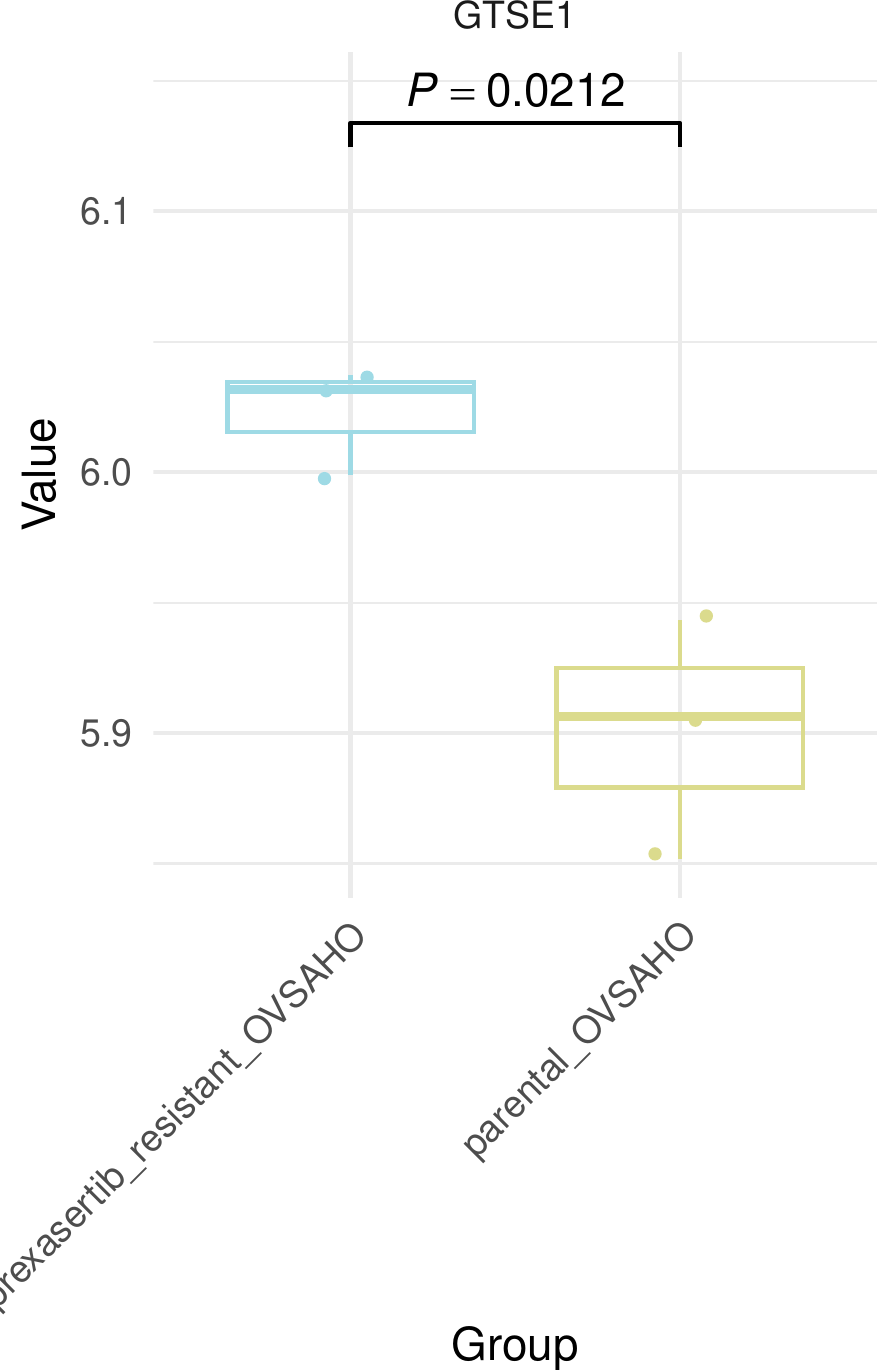
**Fig.** **8** OC GSE143152 Box Plot Of DEGs 2

**(File path: Figure+Table/3.6.7\_Limma\_差异分析\_(OC\_GSE143152)/OC-GSE143152-Box-Plot-Of-DEGs-2.pdf)**



**Fig.** **9** OC GSE143152 Box Plot Of DEGs 3

**(File path: Figure+Table/3.6.7\_Limma\_差异分析\_(OC\_GSE143152)/OC-GSE143152-Box-Plot-Of-DEGs-3.pdf)**



**Fig.** **10** OC GSE143152 Box Plot Of DEGs 4

**(File path: Figure+Table/3.6.7\_Limma\_差异分析\_(OC\_GSE143152)/OC-GSE143152-Box-Plot-Of-DEGs-4.pdf)**

**Tab.** **2** OC GSE143152 Statistic of Focused genes extra

| Hgnc symbol | LogFC | Adj.P.Val | P.Value | Rownames |
| --- | --- | --- | --- | --- |
| GTSE1 | 0.1212 | 0.05954 | 0.02116 | 51512 |

**(File path: Figure+Table/3.6.7\_Limma\_差异分析\_(OC\_GSE143152)/OC-GSE143152-Statistic-of-Focused-genes-extra.xlsx)**

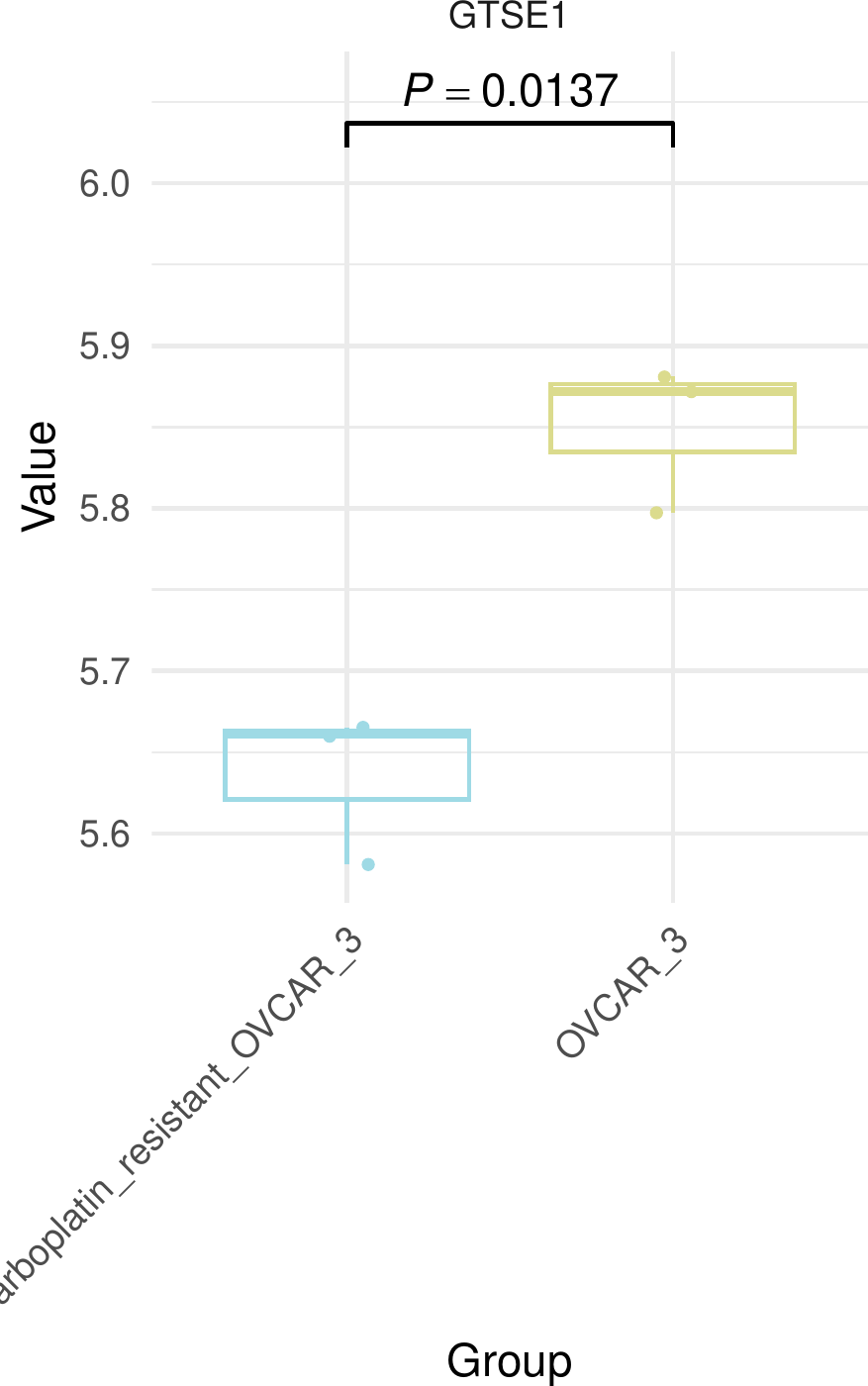
### 3.6.8 GEO 数据获取 (OC\_GSE285737)

以 GEOquery 获取 GSE285737 的数据信息。

### 3.6.9 Limma 差异分析 (OC\_GSE285737)

以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Carboplatin\_resistant\_OVCAR\_3 vs OVCAR\_3。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。

Fig. **[11](#OC-GSE285737-Box-Plot-Of-DEGs)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。



**Fig.** **11** OC GSE285737 Box Plot Of DEGs

**(File path: Figure+Table/3.6.9\_Limma\_差异分析\_(OC\_GSE285737)/OC-GSE285737-Box-Plot-Of-DEGs.pdf)**

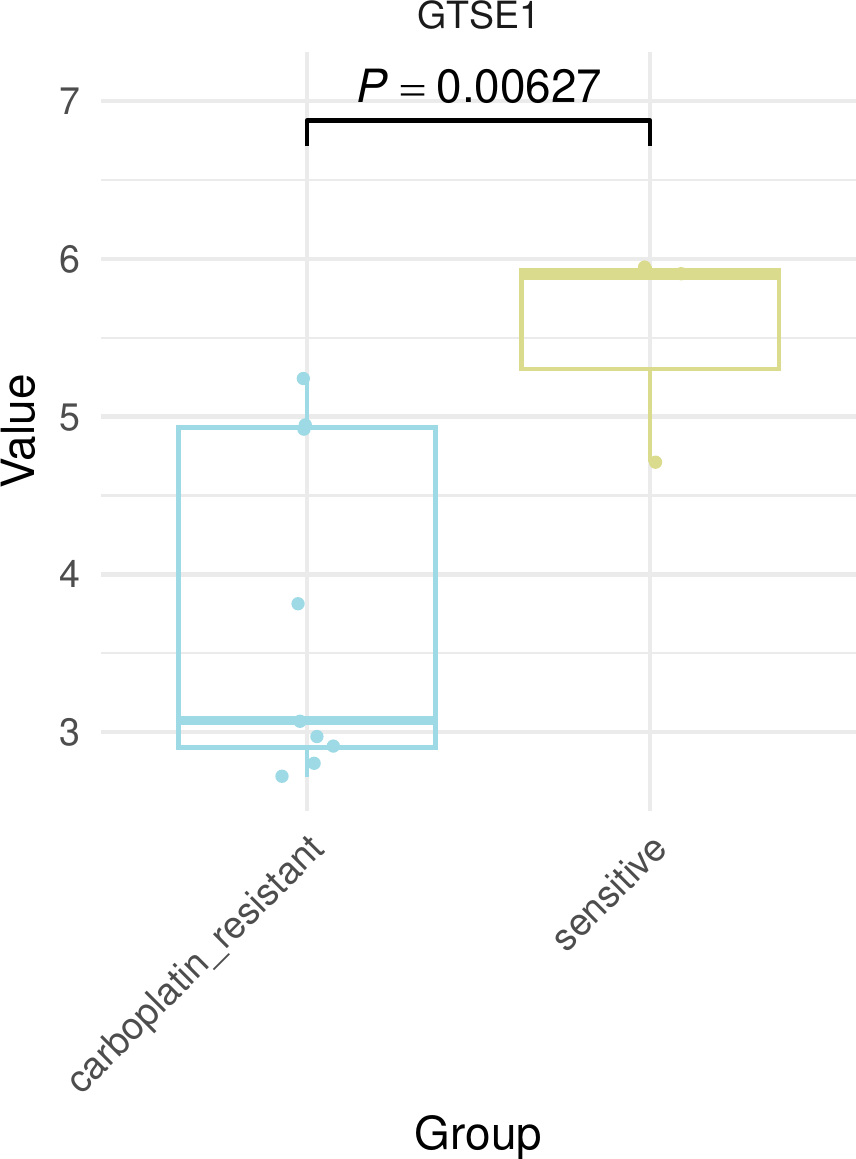
### 3.6.10 GEO 数据获取 (HGSOC\_GSE141630)

以 GEOquery 获取 GSE141630 的数据信息。

### 3.6.11 Limma 差异分析 (HGSOC\_GSE141630)

以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：carboplatin\_resistant vs sensitive。(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 DEGs 统计：up (n=1608) , down (n=1331)

Fig. **[12](#HGSOC-GSE141630-Box-Plot-Of-DEGs)** 基因 GTSE1 表达水平，以及对应的 limma 差异分析显著水平。

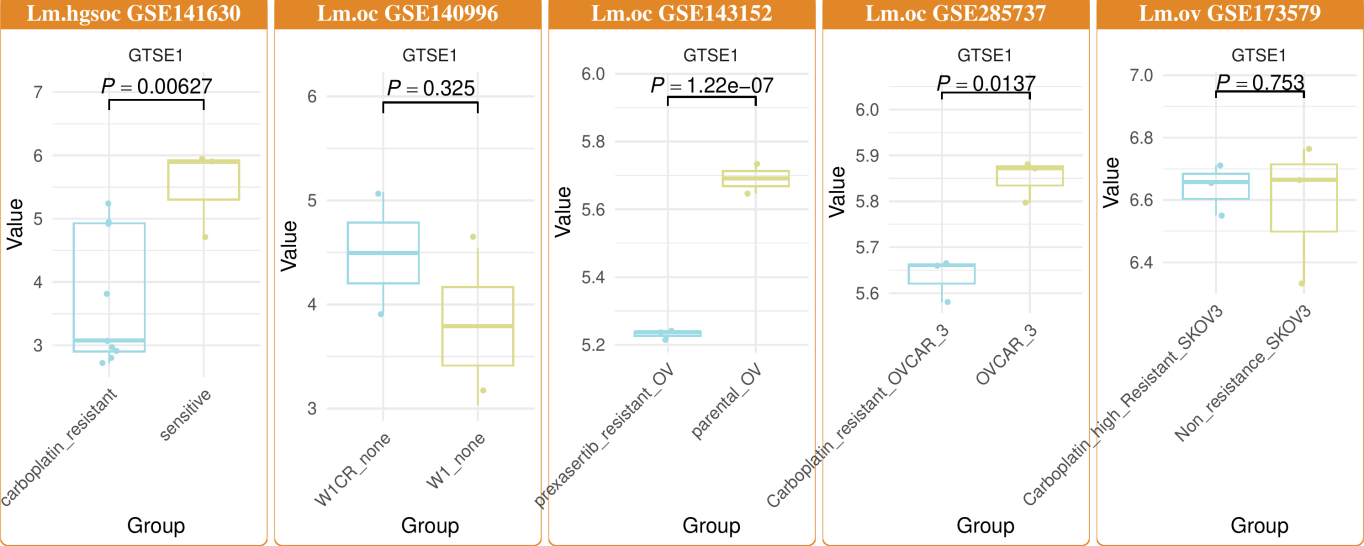


**Fig.** **12** HGSOC GSE141630 Box Plot Of DEGs

**(File path: Figure+Table/3.6.11\_Limma\_差异分析\_(HGSOC\_GSE141630)/HGSOC-GSE141630-Box-Plot-Of-DEGs.pdf)**

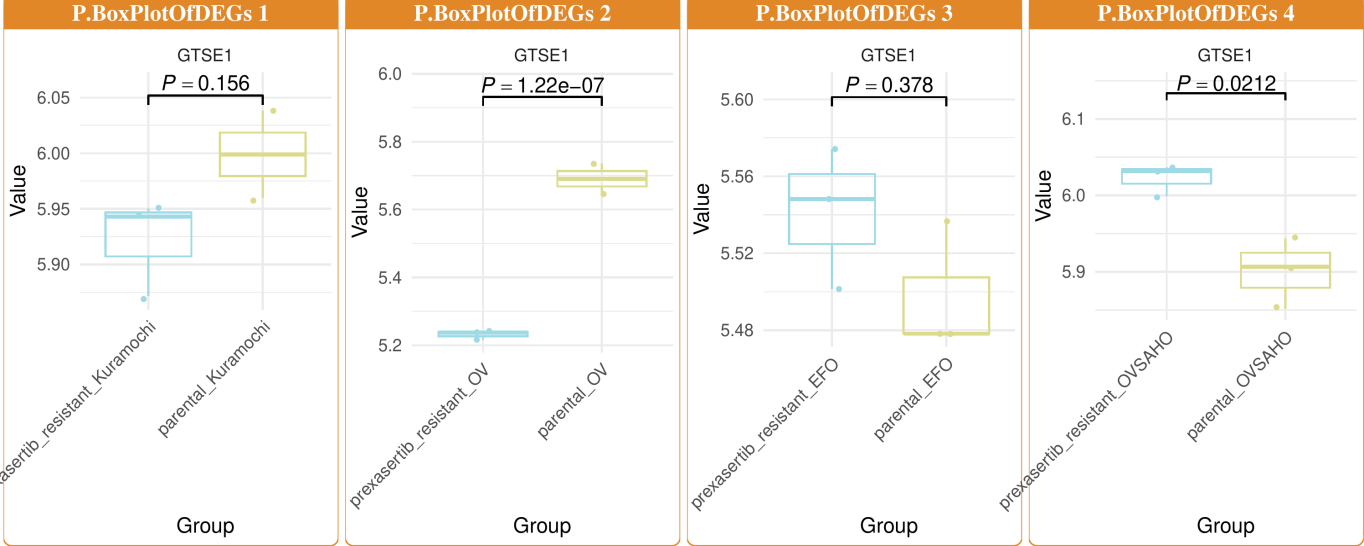
### 3.6.12 汇总

Fig. **[13](#Multiple-Datasets-GTSE1)** Multiple Datasets GTSE1 Fig. **[14](#GSE143152-differential-cell-lines-GTSE1)** GSE143152 differential cell lines GTSE1



**Fig.** **13** Multiple Datasets GTSE1

**(File path: Figure+Table/3.6.12\_汇总/Multiple-Datasets-GTSE1.pdf)**



**Fig.** **14** GSE143152 differential cell lines GTSE1

**(File path: Figure+Table/3.6.12\_汇总/GSE143152-differential-cell-lines-GTSE1.pdf)**

## 3.7 EpiFactors 表观遗传调控因子数据获取 (M6A)

获取 EpiFactors 中的表观遗传调控蛋白。匹配 Modification 中包含“RNA methylation”的描述，最终得到 12 例数据。

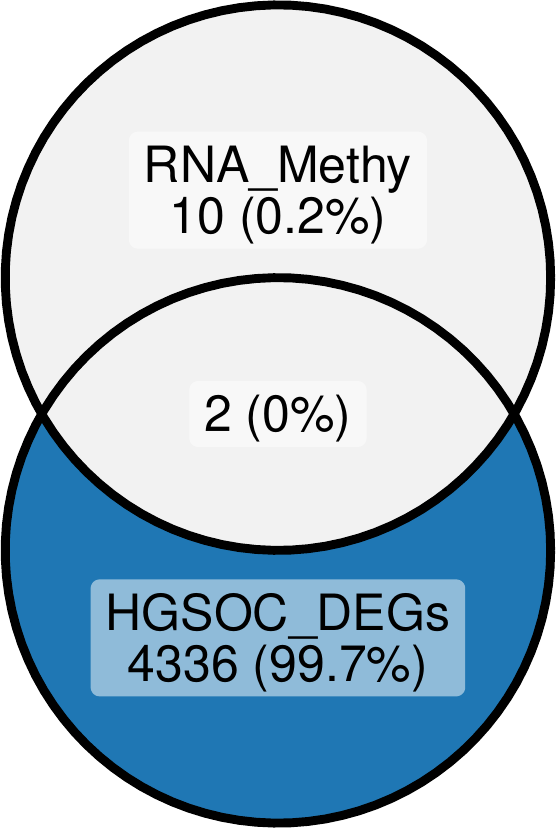
**Tab.** **3** M6A all protein of epigenetic regulators

| Id | HGNC symbol | Status | HGNC ID | HGNC name |
| --- | --- | --- | --- | --- |
| 96 | CBLL1 | New | 21220 | E3 ubiquitin-prot... |
| 387 | METTL14 | New | 29330 | Methyltransferase... |
| 388 | METTL16 | New | 28480 | Methyltransferase... |
| 390 | METTL3 | New | 17560 | Methyltransferase... |
| 391 | METTL4 | New | 24730 | Methyltransferase... |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.7.0\_EpiFactors\_表观遗传调控因子数据获取\_(M6A)/M6A-all-protein-of-epigenetic-regulators.xlsx)**

## 3.8 交集: HGSOC\_DEGs + RNA\_Methy (SIG)

Fig. **[15](#Intersection-of-HGSOC-DEGs-with-RNA-Methy)** 以下取交集： - 基因集 (hgsoc - normal, 来自于Limma 差异分析(Section: HGSOC)) - 基因集 (CBLL1, METTL14, etc., 来自于EpiFactors 表观遗传调控因子数据获取(Section: M6A))



**Fig.** **15** Intersection of HGSOC DEGs with RNA Methy

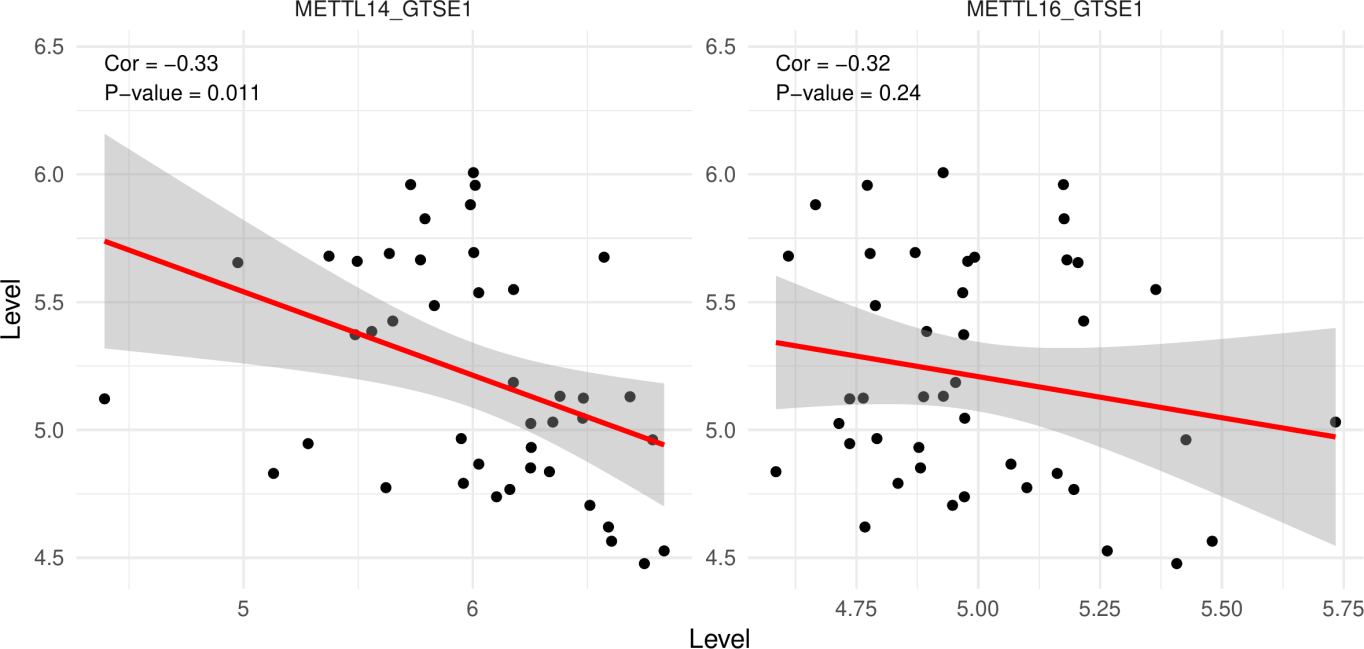
**(File path: Figure+Table/3.8.0\_交集:\_HGSOC\_DEGs\_+\_RNA\_Methy\_(SIG)/Intersection-of-HGSOC-DEGs-with-RNA-Methy.pdf)**

* All\_intersection: METTL14, METTL16

**(See: Figure+Table/3.8.0\_交集:\_HGSOC\_DEGs\_+\_RNA\_Methy\_(SIG)/Intersection-of-HGSOC-DEGs-with-RNA-Methy-content)**

## 3.9 关联分析 (HGSOC)

将基因 (METTL14, METTL16 -> GTSE1) 关联分析。



**Fig.** **16** Linear regression

**(File path: Figure+Table/3.9.0\_关联分析\_(HGSOC)/Linear-regression.pdf)**

# 4 总结

Fig. **[4](#HGSOC-Box-Plot-Of-DEGs)** ， 结果表明 GTSE1 癌症高表达， 耐药数据集相对以于癌症低表达 (多数为低表达，见 Fig. **[13](#Multiple-Datasets-GTSE1)** ) (Fig. **[10](#OC-GSE143152-Box-Plot-Of-DEGs-4)** 为耐药高表达)。

随后，筛选出的差异表达基因 METTL14，与 GTSE1 显著关联，见 Fig. **[16](#Linear-regression)** 。

# 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](https://doi.org/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.

4. Marakulina, D. *et al.* EpiFactors 2022: Expansion and enhancement of a curated database of human epigenetic factors and complexes. *Nucleic acids research* **51**, D564–D570 (2023).