**生信分析报告**

**项目标题： 预测甲基化调控因子 ;**

**单 号： BSCL240914 ;**

**分析人员： 黄礼闯 ;**

**分析类型： 生信分析 ;**

**委 托 人： 邱美婷 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程

## 1.1 需求

通过软件预测甲基化调控因子（如METTL14）的靶基因，并通过数据库筛选于PCOS患者中表达水平具有显著差异性的基因，合并交集，并对该交集中的基因进行功能富集和KEGG通路富集分析，筛选PCOS患者中可能的METTL14甲基化调控基因及其相关通路；

## 1.2 实际流程

从 EpiFactors 获取表观遗传调控因子，筛出甲基化相关调控因子 (A 集合) 。 获取 PCOS GEO 数据，差异分析得到 DEGs，发现可能存在甲基化修饰位点的基因 B 集合。 在 PCOS 中筛选出差异表达的甲基化调控因子 (C集合) ，与 B 集合关联分析，随后富集分析。

补充了显著关联的基因的 PTMs 预测 (组蛋白修饰位点预测，包括 Methylarginine，Methyllysine 类型。

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

从数据库 EpiFactors (2023, **IF:16.6**, Q1, Nucleic acids research)1 获取表观遗传调控蛋白的数据。

## 2.3 GEO 数据获取 (Dataset: PCOS)

以 R 包 GEOquery (2.74.0) 获取 GSE277906 数据集。

## 2.4 Limma 差异分析 (Dataset: PCOS)

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 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：pcos vs control。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。 对 GSE277906 的 mRNA 数据 (protein\_coding) 差异分析

## 2.5 富集分析 (Dataset: SIGCOR\_05)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)4进行 KEGG 和 GO 富集分析。 以 pathview R 包 (1.46.0) 对选择的 KEGG 通路可视化。

## 2.6 MusiteDeep 蛋白质转录后修饰位点预测 (Dataset: SIGCOR\_05)

以 biomaRt (2009, **IF:13.1**, Q1, Nature protocols)5 获取蛋白质 (hsa) 的序列 (biomaRt::getSequence 获取 peptide)。 以 Python 工具 MusiteDeep (2020, **IF:16.6**, Q1, Nucleic Acids Research)6 预测 Methylarginine, Methyllysine 修饰位点，设定 PTM 得分截断为 0.5。

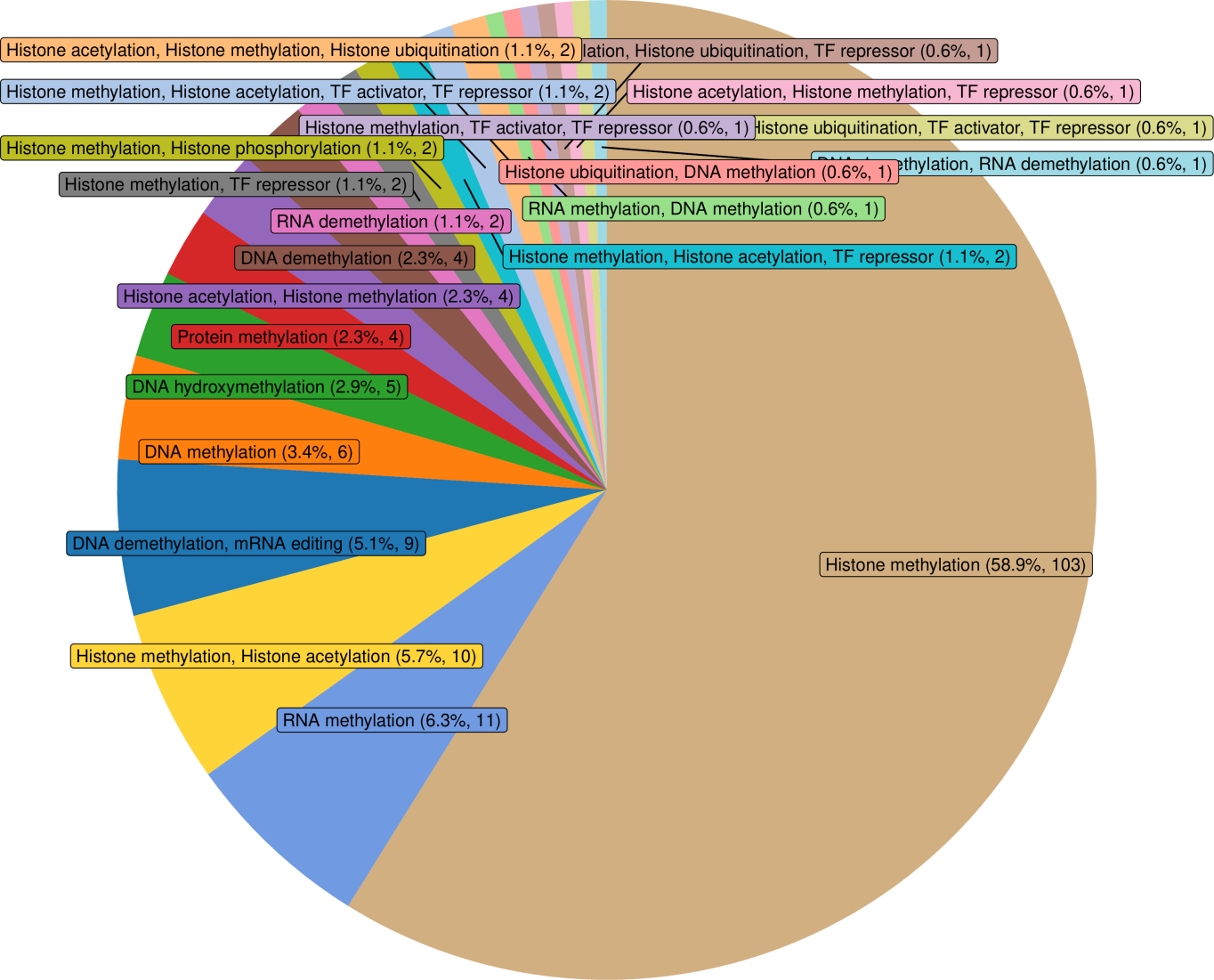
## 2.7 富集分析 (Dataset: PTMS)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)4进行 KEGG 和 GO 富集分析。 以 pathview R 包 (1.46.0) 对选择的 KEGG 通路可视化。

# 3 分析结果

## 3.1 EpiFactors 表观遗传调控因子数据获取 (METHY)

从所有 表观调控因子 Fig. 中筛选出甲基化修饰调控因子，见 Tab.



**Fig.** **1** Distribution all protein of epigenetic regulators

**(File path: Figure+Table/Distribution-all-protein-of-epigenetic-regulators.pdf)**

**Tab.** **1** METHY regulators

| Id | HGNC s... | Status | HGNC ID | HGNC name | GeneID | UniPro......7 | UniPro......8 | Domain | MGI sy... |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 11 | AEBP2 | # | 24051 | AE bin... | 121536 | Q6ZN18 | AEBP2 ... | Pfam-B... | Aebp2 |
| 12 | AICDA | # | 13203 | Activa... | 57379 | Q9GZX7 | AICDA ... | APOBEC... | Aicda |
| 15 | ALKBH1 | New | 17911 | Nuclei... | 8846 | Q13686 | ALKB1 ... | PF13532 | Alkbh1 |
| 16 | ALKBH4 | New | 21900 | Alpha-... | 54784 | Q9NXW9 | ALKB4 ... | PF13532 | Alkbh4 |
| 17 | ALKBH5 | New | 25996 | AlkB h... | 54890 | Q6P6C2 | ALKB5 ... | PF13532 | Alkbh5 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/METHY-regulators.xlsx)**

## 3.2 GEO 数据获取 (PCOS)

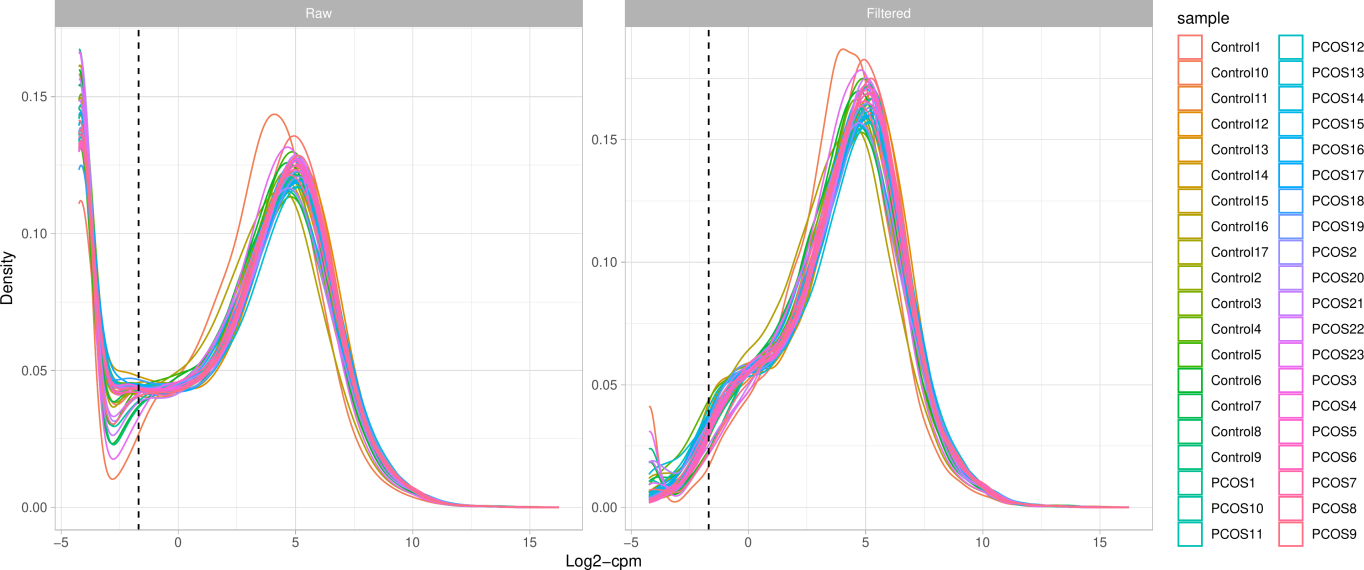
获取 GEO PCOS 数据，用于筛选差异表达基因。

* Data Source ID: GSE277906
* data\_processing: Illumina Casava1.7 software used for basecalling.
* data\_processing.1: Raw reads of fastq format were firstly processed using fastp and the low quality reads were removed to obtain the clean reads.
* data\_processing.2: The clean reads were mapped to the reference genome using HISAT2. FPKM of each gene was calculated and the read counts of each gene were obtained by HTSeq-count
* data\_processing.3: Assembly: GRCh38
* (Others): …

**(见Figure+Table/PCOS-GSE277906-content)**

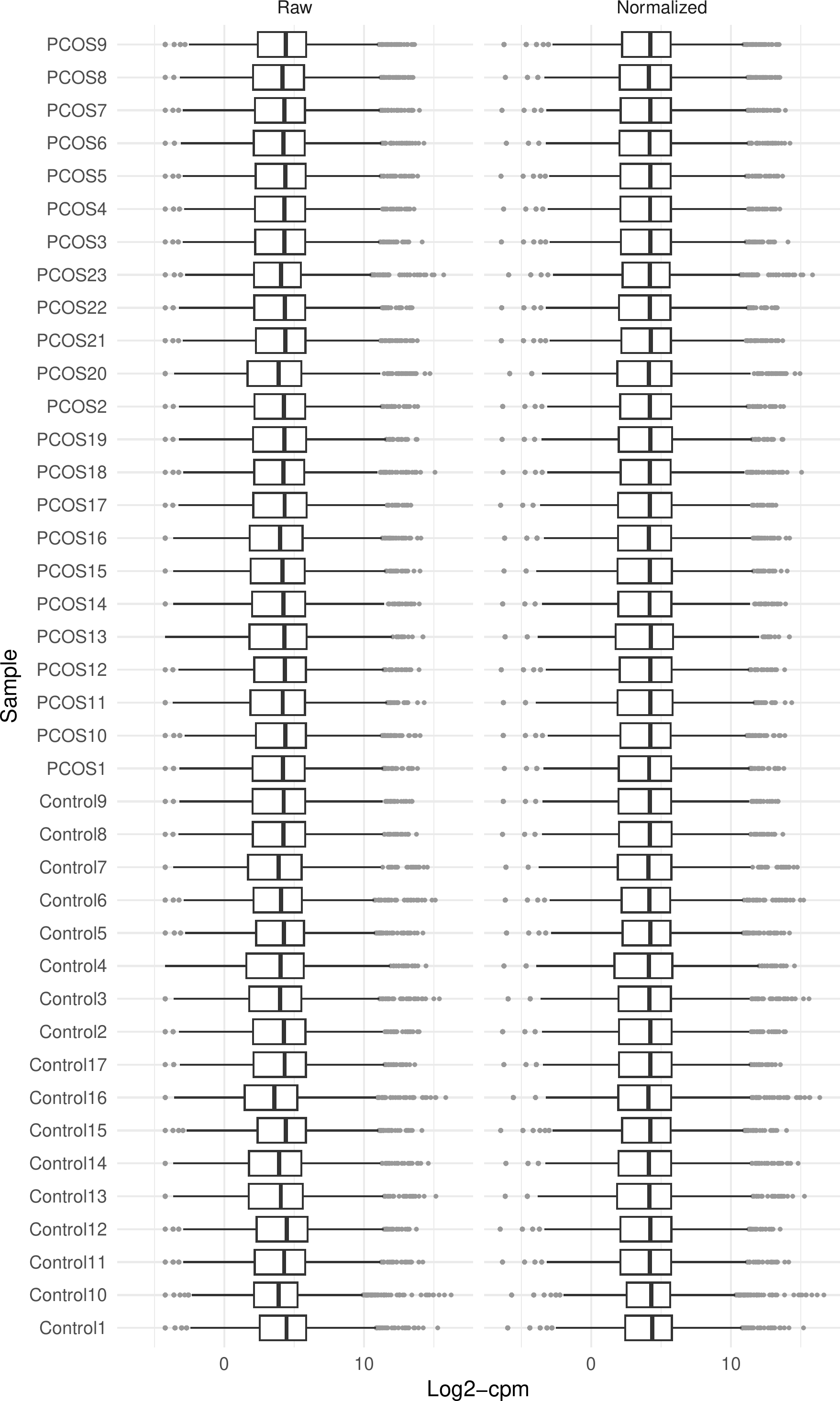
## 3.3 Limma 差异分析 (PCOS)

差异分析，得到 DEGs 见 Fig.



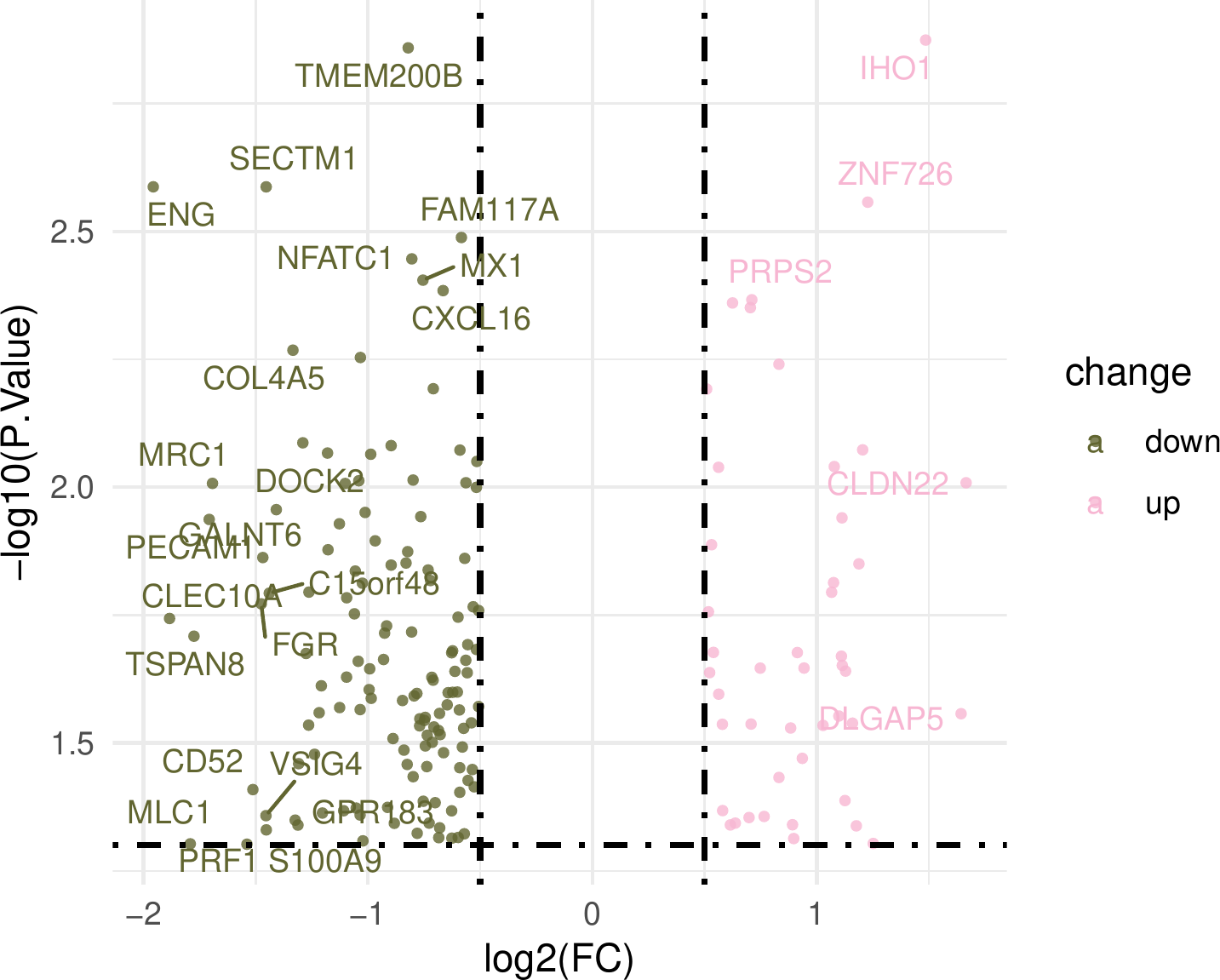
**Fig.** **2** PCOS Filter low counts

**(File path: Figure+Table/PCOS-Filter-low-counts.pdf)**



**Fig.** **3** PCOS Normalization

**(File path: Figure+Table/PCOS-Normalization.pdf)**



**Fig.** **4** PCOS pcos vs control

**(File path: Figure+Table/PCOS-pcos-vs-control.pdf)**

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

**(See: Figure+Table/PCOS-pcos-vs-control-content)**

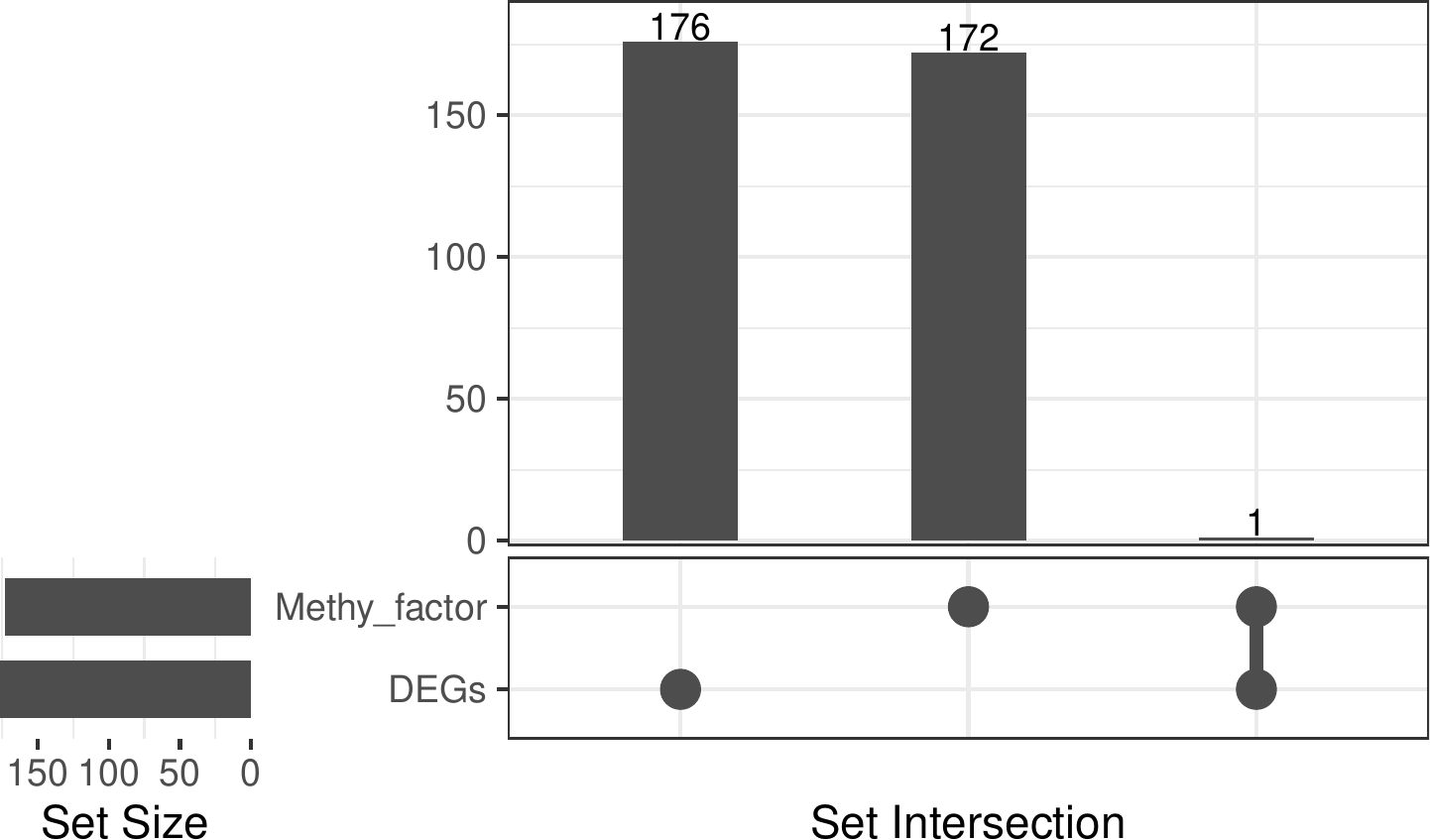
**Tab.** **2** PCOS data pcos vs control

| Rownames | Id | Gene D... | Coding... | Descri... | Pathway | Pathwa... | GO ID | GO term | Wiki ID |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PRPS2 | PRPS2 | 5634 | Protei... | Phosph... | Hsa000... | Pentos... | GO:000... | Magnes... |  |
| FXYD6 | FXYD6 | 53826 | Protei... | FXYD d... |  |  | GO:000... | Molecu... |  |
| MMP15 | MMP15 | 4324 | Protei... | Matrix... | Hsa04928 | Parath... | GO:000... | Metall... | WP5283... |
| CXCL16 | CXCL16 | 58191 | Protei... | C-X-C ... | Hsa040... | Cytoki... | GO:000... | Low-de... | WP5115... |
| MX1 | MX1 | 4599 | Protei... | MX dyn... | Hsa032... | Viral ... | GO:000... | GTPase... | WP5115... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/PCOS-data-pcos-vs-control.xlsx)**

## 3.4 差异表达的 Methylation Factors

将差异表达基因与 Tab. **[1](#METHY-regulators)** 中的因子取交集， 见 Fig. 。



**Fig.** **5** Intersection of Methy factor with DEGs

**(File path: Figure+Table/Intersection-of-Methy-factor-with-DEGs.pdf)**

* All\_intersection: PRDM6

**(See: Figure+Table/Intersection-of-Methy-factor-with-DEGs-content)**

**Tab.** **3** Intersection METHY epigenetic regulators

| Id | HGNC s... | Status | HGNC ID | HGNC name | GeneID | UniPro... | UniPro...1 | Domain | MGI sy... |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 510 | PRDM6 | # | 9350 | PR dom... | 93166 | Q9NQX0 | PRDM6 ... | SET PF... | Prdm6 |

**(File path: Figure+Table/Intersection-METHY-epigenetic-regulators.xlsx)**

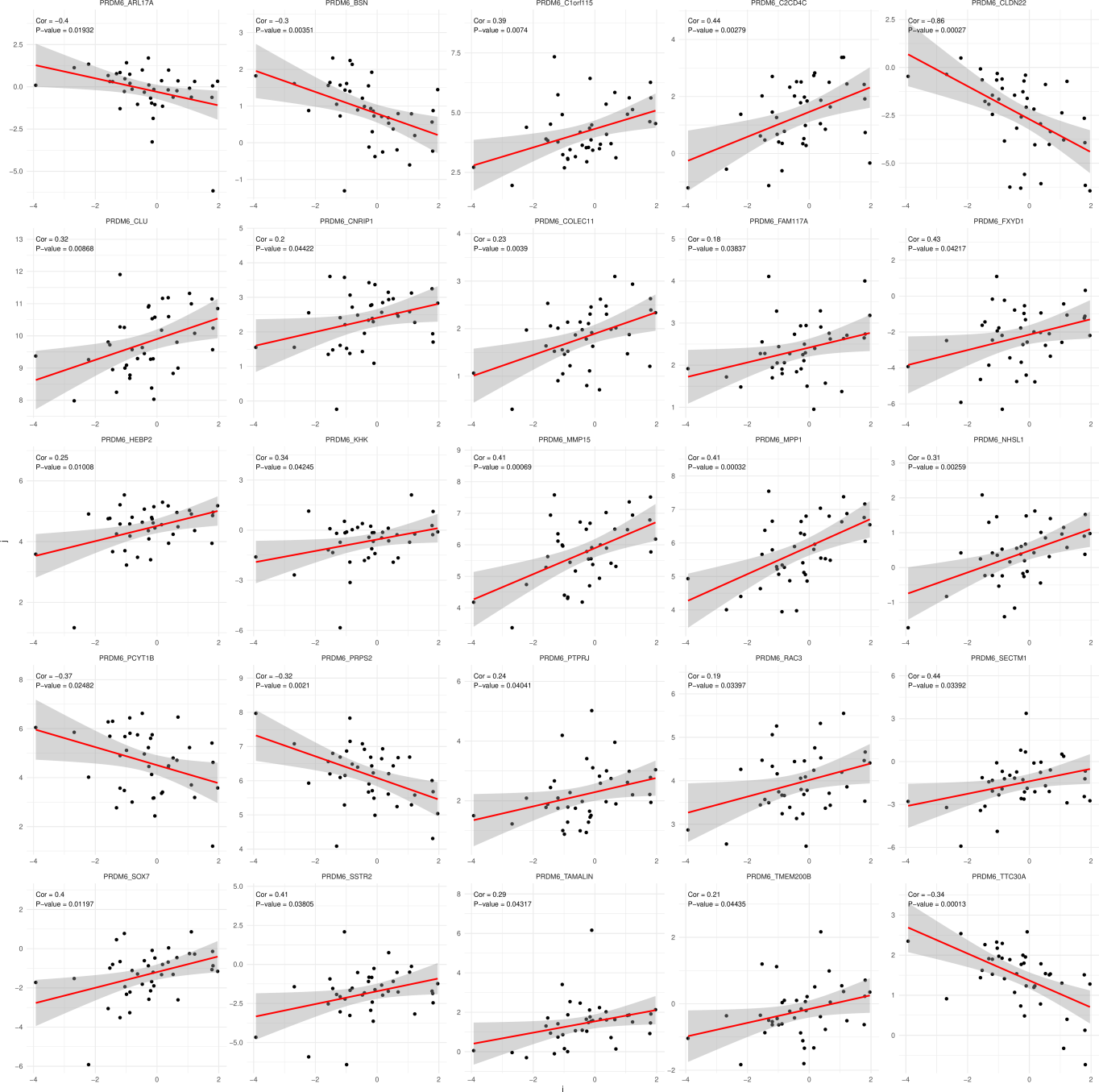
## 3.5 Methylation Factors 与 DEGs 关联分析

为了寻找 Fig. **[5](#Intersection-of-Methy-factor-with-DEGs)** 中发现的差异表达的 Methylation Factors 可能调控的 DEGs 修饰，将两个数据集作关联分析，结果见 Tab. 。 以 pvalue < 0.05 为条件筛选，见 Tab. ， Fig. 。 其中，pvalue < 0.001 的见Fig. 。

**Tab.** **4** All correlation results

| From | To | Cor | Pvalue | Model | -log2(... | Signif... | Sign |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PRDM6 | A2M | 0.1286... | 0.49596 | C(Cont... | 1.0117... | > 0.05 | - |
| PRDM6 | AARD | -0.195... | 0.35308 | C(Cont... | 1.5019... | > 0.05 | - |
| PRDM6 | AATK | -0.035... | 0.74928 | C(Cont... | 0.4164... | > 0.05 | - |
| PRDM6 | ABCC9 | -0.146... | 0.45304 | C(Cont... | 1.1422... | > 0.05 | - |
| PRDM6 | ADAMTSL2 | 0.1576... | 0.30011 | C(Cont... | 1.7364... | > 0.05 | - |
| ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/All-correlation-results.xlsx)**



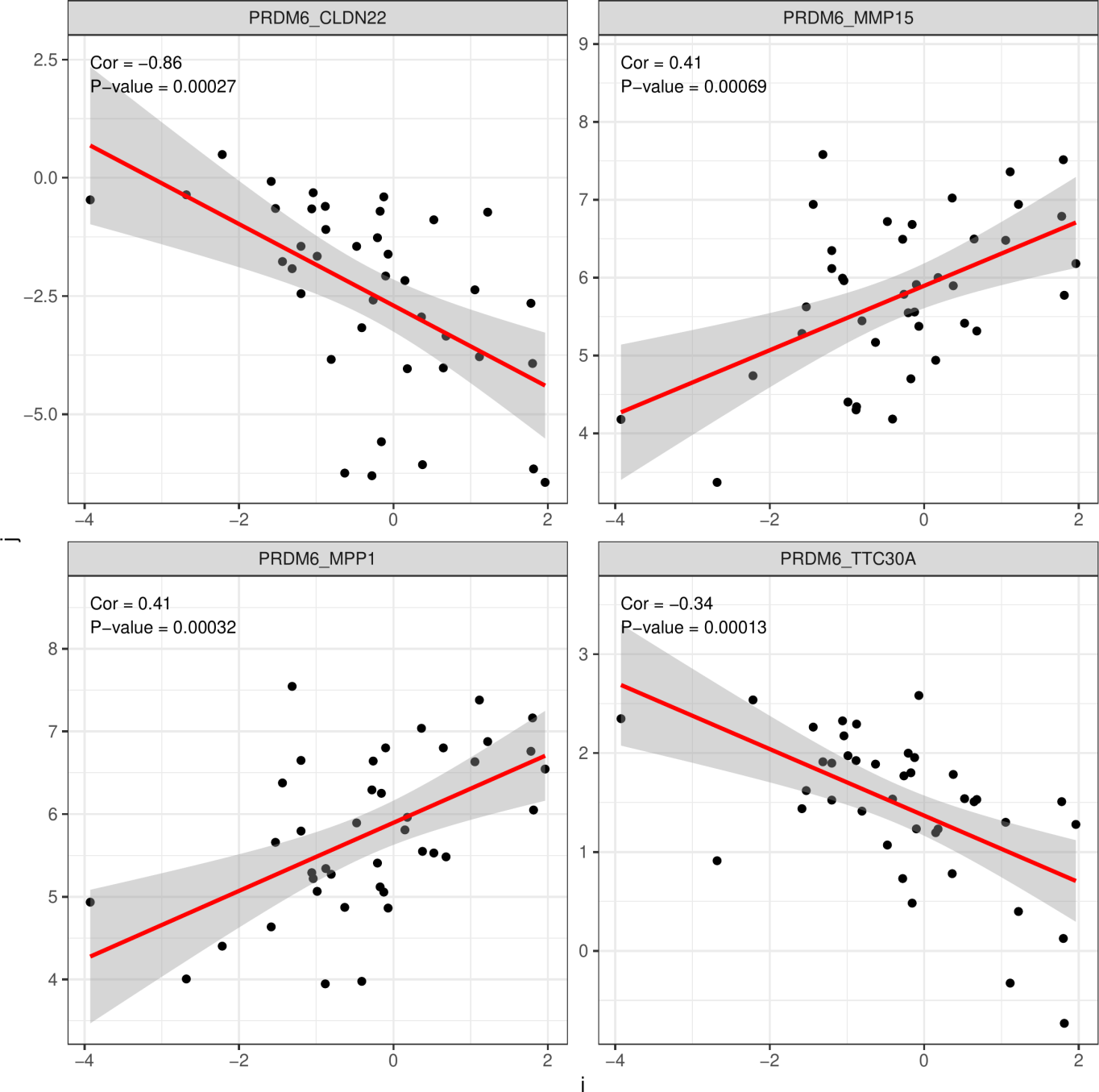
**Fig.** **6** Significant correlation

**(File path: Figure+Table/Significant-correlation.pdf)**

**Tab.** **5** Correlation results 05

| From | To | Cor | Pvalue | Model | -log2(... | Signif... | Sign |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PRDM6 | ARL17A | -0.401... | 0.019322 | C(Cont... | 5.6936... | < 0.05 | \* |
| PRDM6 | BSN | -0.295... | 0.0035095 | C(Cont... | 8.1545... | < 0.05 | \* |
| PRDM6 | C1orf115 | 0.3889... | 0.0074019 | C(Cont... | 7.0778... | < 0.05 | \* |
| PRDM6 | C2CD4C | 0.4359... | 0.0027875 | C(Cont... | 8.4868... | < 0.05 | \* |
| PRDM6 | CLDN22 | -0.862... | 0.0002... | C(Cont... | 11.862... | < 0.001 | \*\* |
| ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/correlation-results-05.xlsx)**



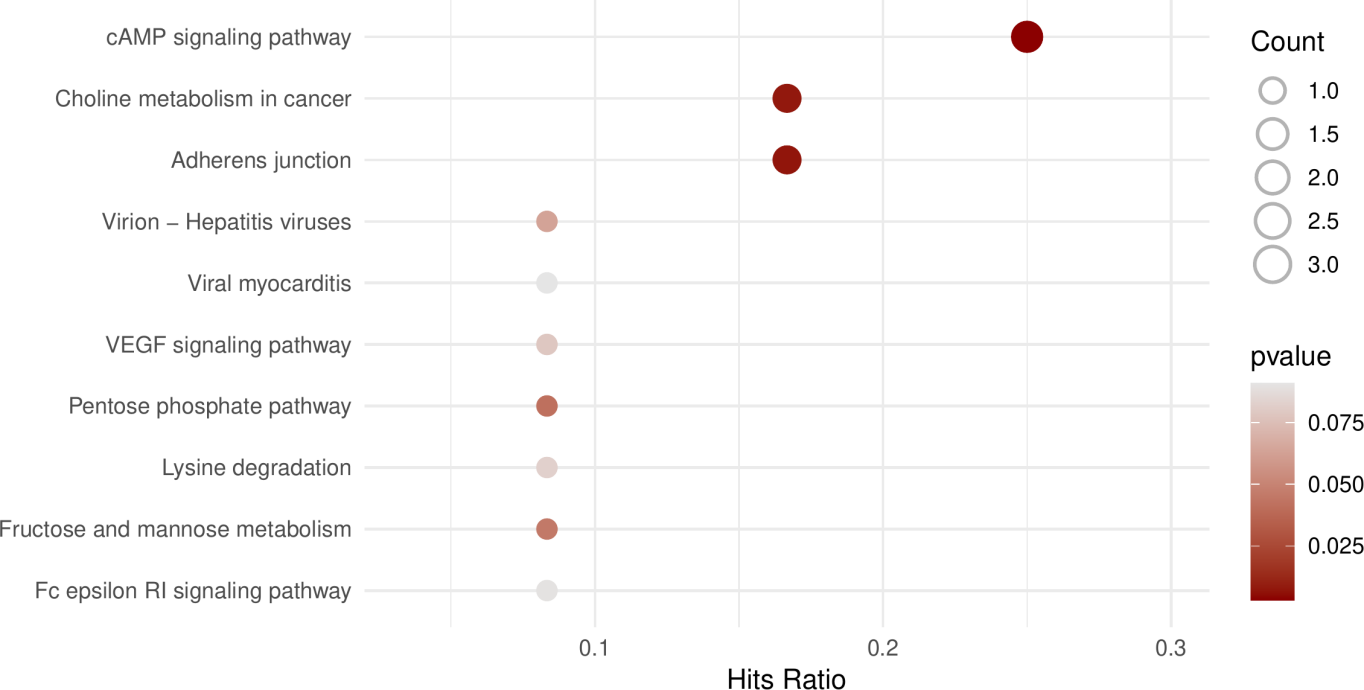
**Fig.** **7** Correlation results 001

**(File path: Figure+Table/correlation-results-001.pdf)**

## 3.6 富集分析 (SIGCOR\_05)

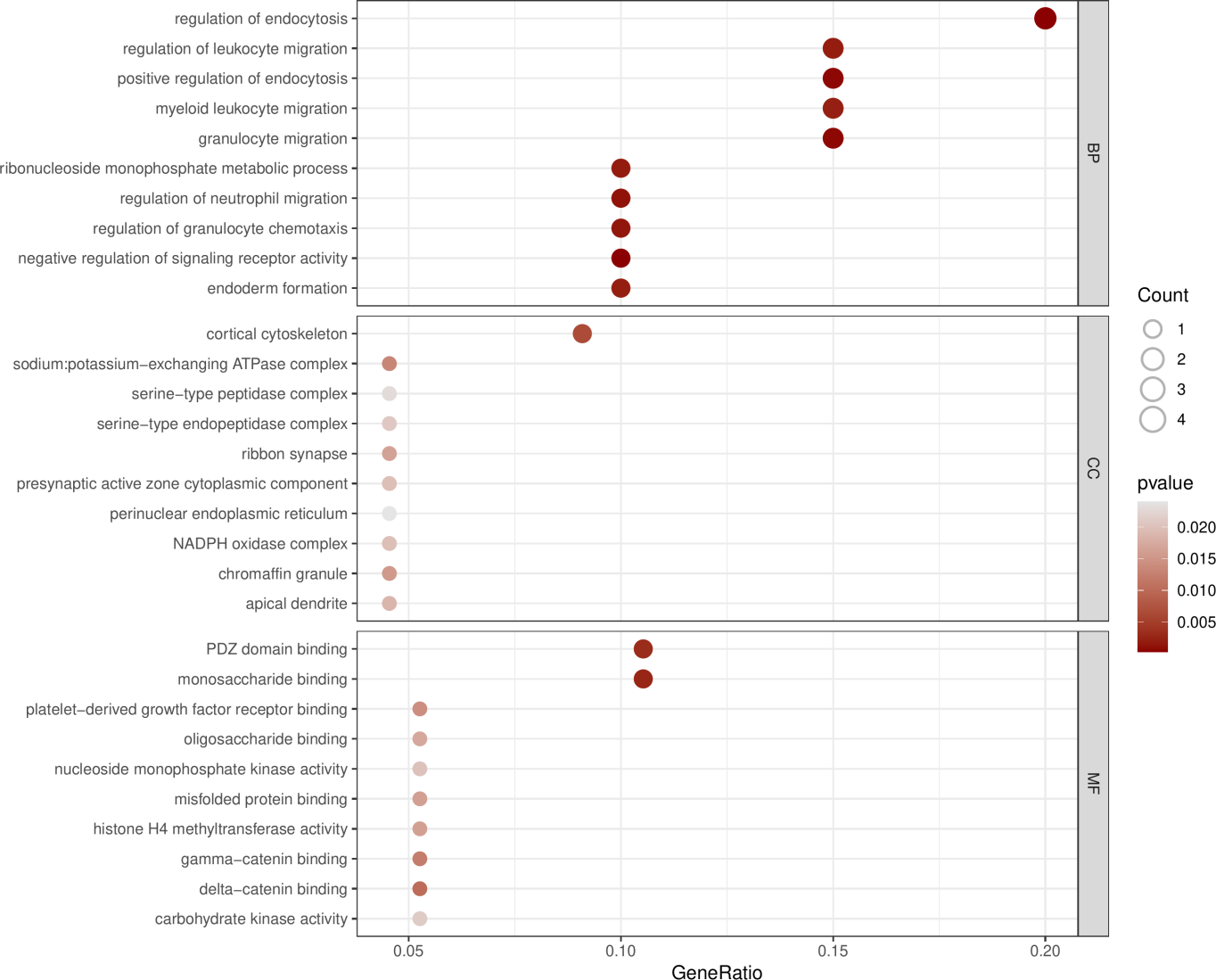
将 Tab. **[5](#correlation-results-05)** 中的基因富集分析 (包含 PRDM6)，

KEGG，GO 结果见 Fig. ， Fig. 。 Fig. 为 KEGG 中最为显著的 cAMP 通路，可能与 PCOS 中甲基化调控相关。 富集分析的数据表格见 Tab. ， Tab. 。



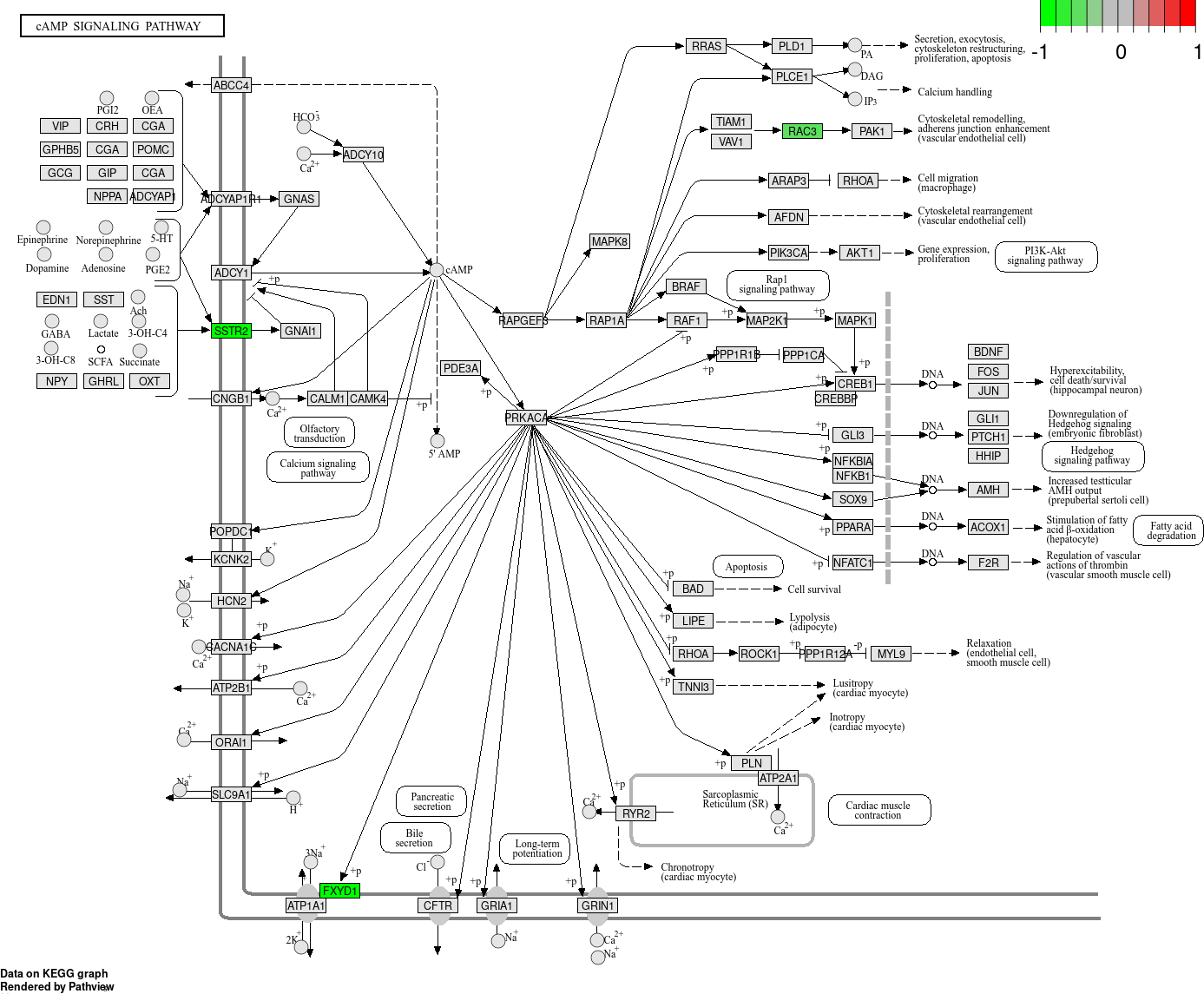
**Fig.** **8** SIGCOR 05 KEGG enrichment

**(File path: Figure+Table/SIGCOR-05-KEGG-enrichment.pdf)**



**Fig.** **9** SIGCOR 05 GO enrichment

**(File path: Figure+Table/SIGCOR-05-GO-enrichment.pdf)**



**Fig.** **10** SIGCOR 05 hsa04024 visualization

**(File path: Figure+Table/SIGCOR-05-hsa04024-visualization.png)**

* Interactive figure:
* Enriched genes: SSTR2, FXYD1, RAC3

**Tab.** **6** SIGCOR 05 KEGG enrichment data

| Category | Subcat... | ID | Descri... | GeneRatio | BgRatio | Pvalue | P.adjust | Qvalue | GeneID |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Enviro... | Signal... | Hsa04024 | CAMP s... | 3/12 | 226/8868 | 0.0030... | 0.1085... | 0.1009... | 5348/5... |
| Cellul... | Cellul... | Hsa04520 | Adhere... | 2/12 | 93/8868 | 0.0067... | 0.1085... | 0.1009... | 5795/5881 |
| Human ... | Cancer... | Hsa05231 | Cholin... | 2/12 | 99/8868 | 0.0075... | 0.1085... | 0.1009... | 9468/5881 |
| Metabo... | Carboh... | Hsa00030 | Pentos... | 1/12 | 31/8868 | 0.0411... | 0.2895... | 0.2693... | 5634 |
| Metabo... | Carboh... | Hsa00051 | Fructo... | 1/12 | 34/8868 | 0.0450... | 0.2895... | 0.2693... | 3795 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/SIGCOR-05-KEGG-enrichment-data.xlsx)**

**Tab.** **7** SIGCOR 05 GO enrichment data

| Ont | ID | Descri... | GeneRatio | BgRatio | Pvalue | P.adjust | Qvalue | GeneID | Count |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BP | GO:003... | Regula... | 4/20 | 307/18986 | 0.0002... | 0.0821... | 0.0605... | 1191/7... | 4 |
| BP | GO:200... | Negati... | 2/20 | 25/18986 | 0.0003... | 0.0821... | 0.0605... | 25927/... | 2 |
| BP | GO:009... | Granul... | 3/20 | 156/18986 | 0.0005... | 0.0821... | 0.0605... | 4354/5... | 3 |
| BP | GO:004... | Positi... | 3/20 | 159/18986 | 0.0005... | 0.0821... | 0.0605... | 1191/7... | 3 |
| BP | GO:190... | Regula... | 2/20 | 47/18986 | 0.0011... | 0.1115... | 0.0822... | 4354/5881 | 2 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

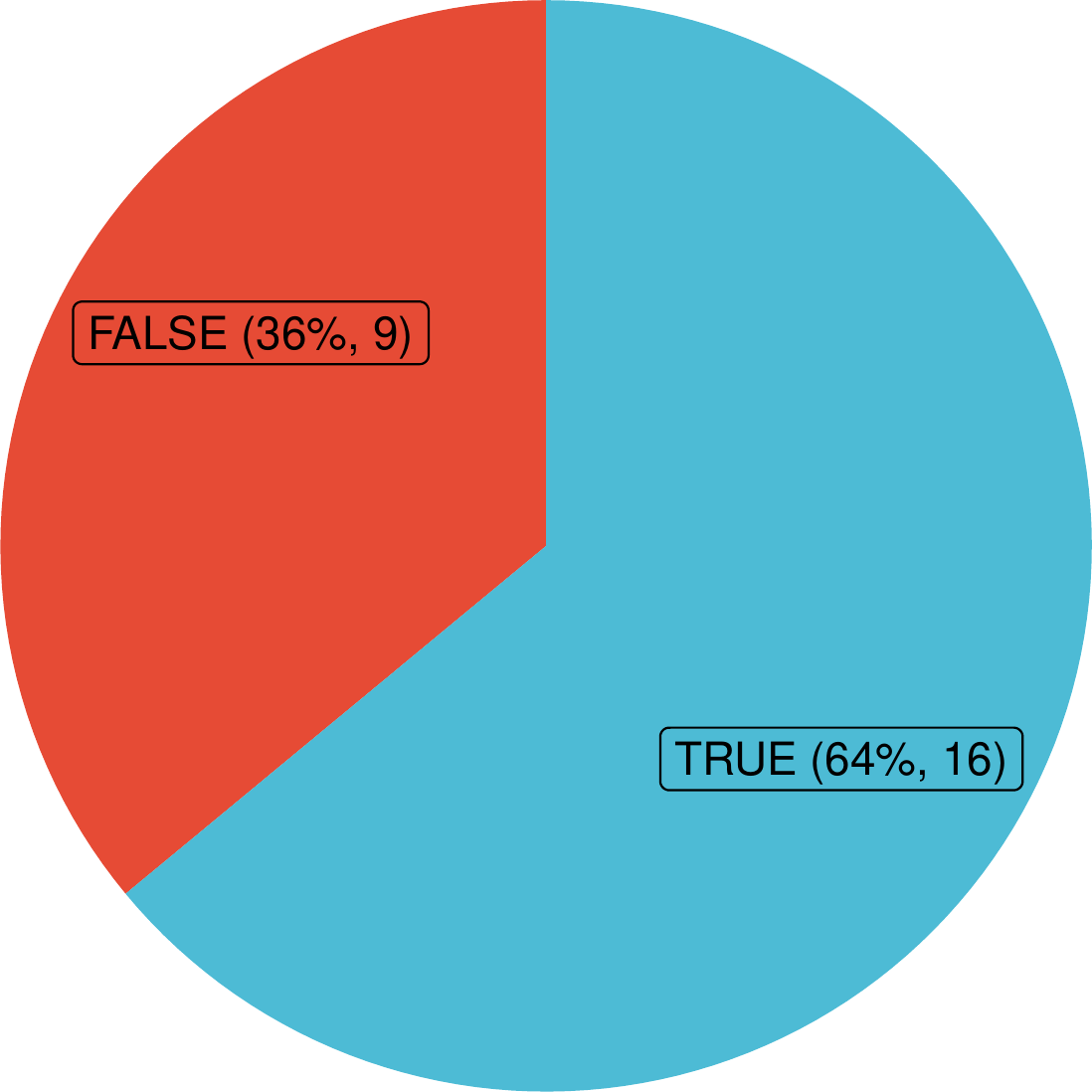
**(File path: Figure+Table/SIGCOR-05-GO-enrichment-data.xlsx)**

## 3.7 MusiteDeep 蛋白质转录后修饰位点预测 (SIGCOR\_05)

对 Tab. **[5](#correlation-results-05)** 中的基因预测了组蛋白修饰位点 (“Methylarginine”, “Methyllysine”) 。

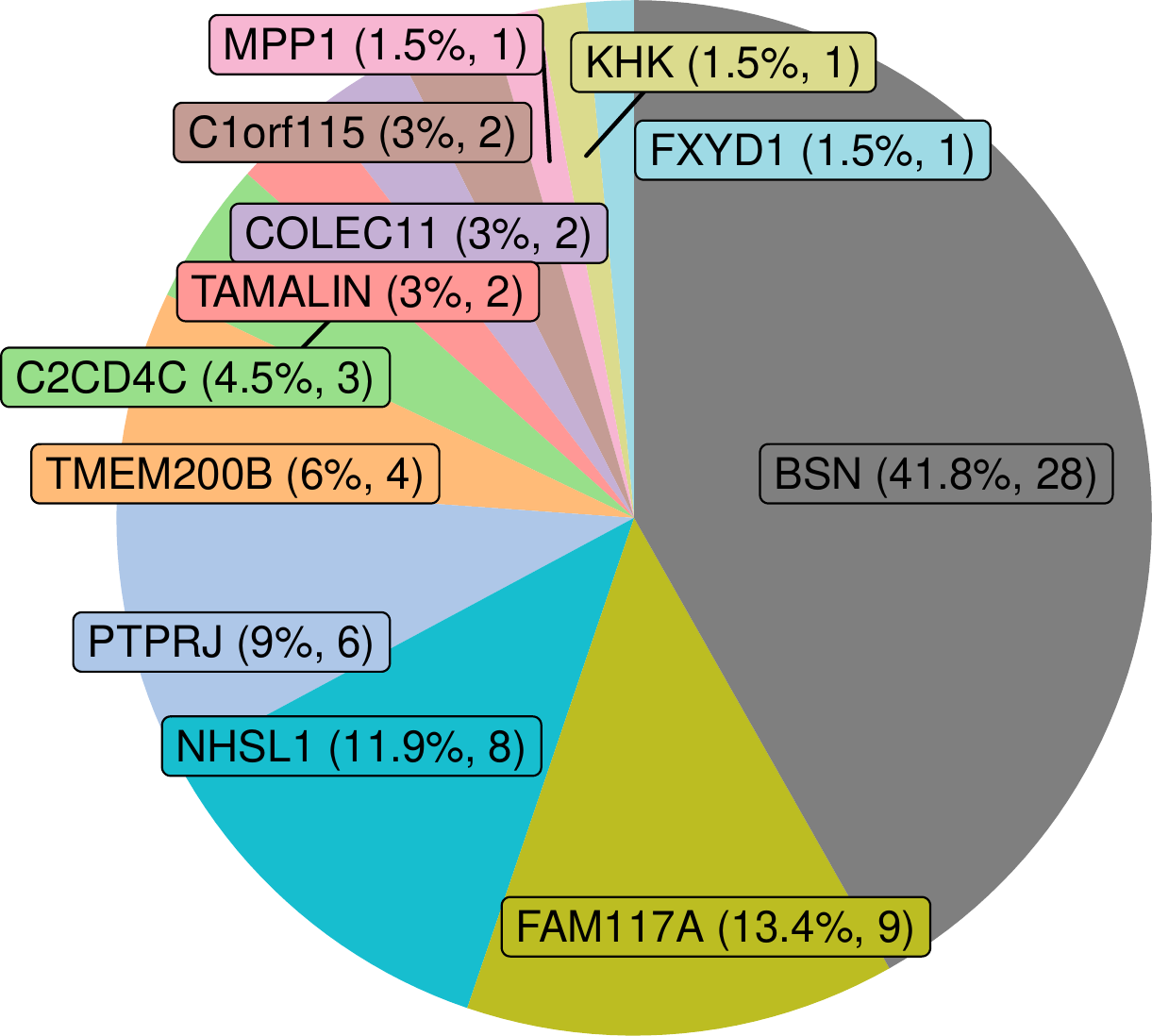
Fig. 统计了含有修饰位点的基因 (设置了 PTM score cutoff，见 **[Error! Bookmark not defined.](#Intersection)**)。 Fig. ，Fig. 为各类型统计。

Fig. ， Fig. ， Fig. ，为所有 cAMP 通路的三个基因的修饰位点，以 Fig. 为最佳。



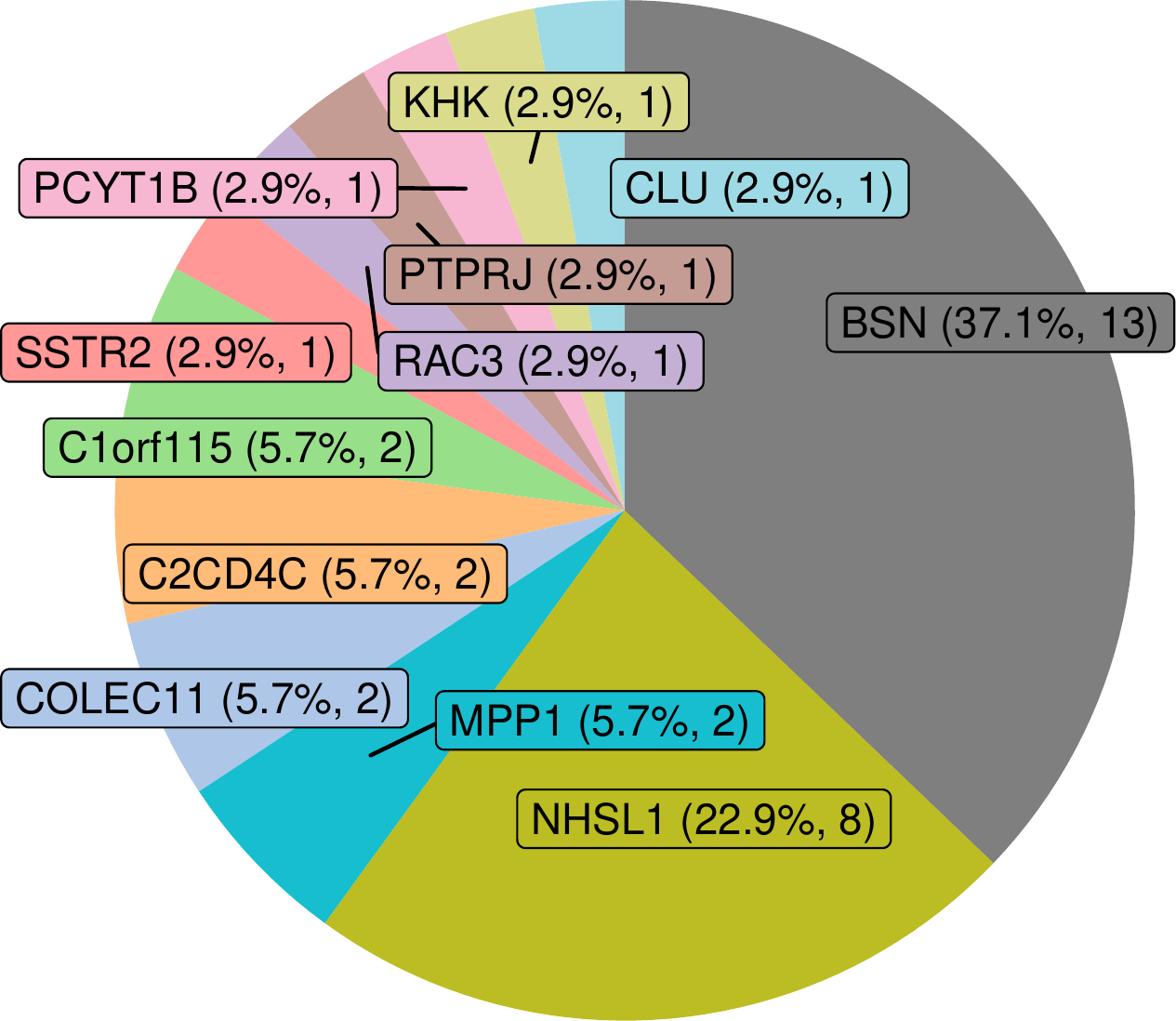
**Fig.** **11** SIGCOR 05 has PTMs

**(File path: Figure+Table/SIGCOR-05-has-PTMs.pdf)**



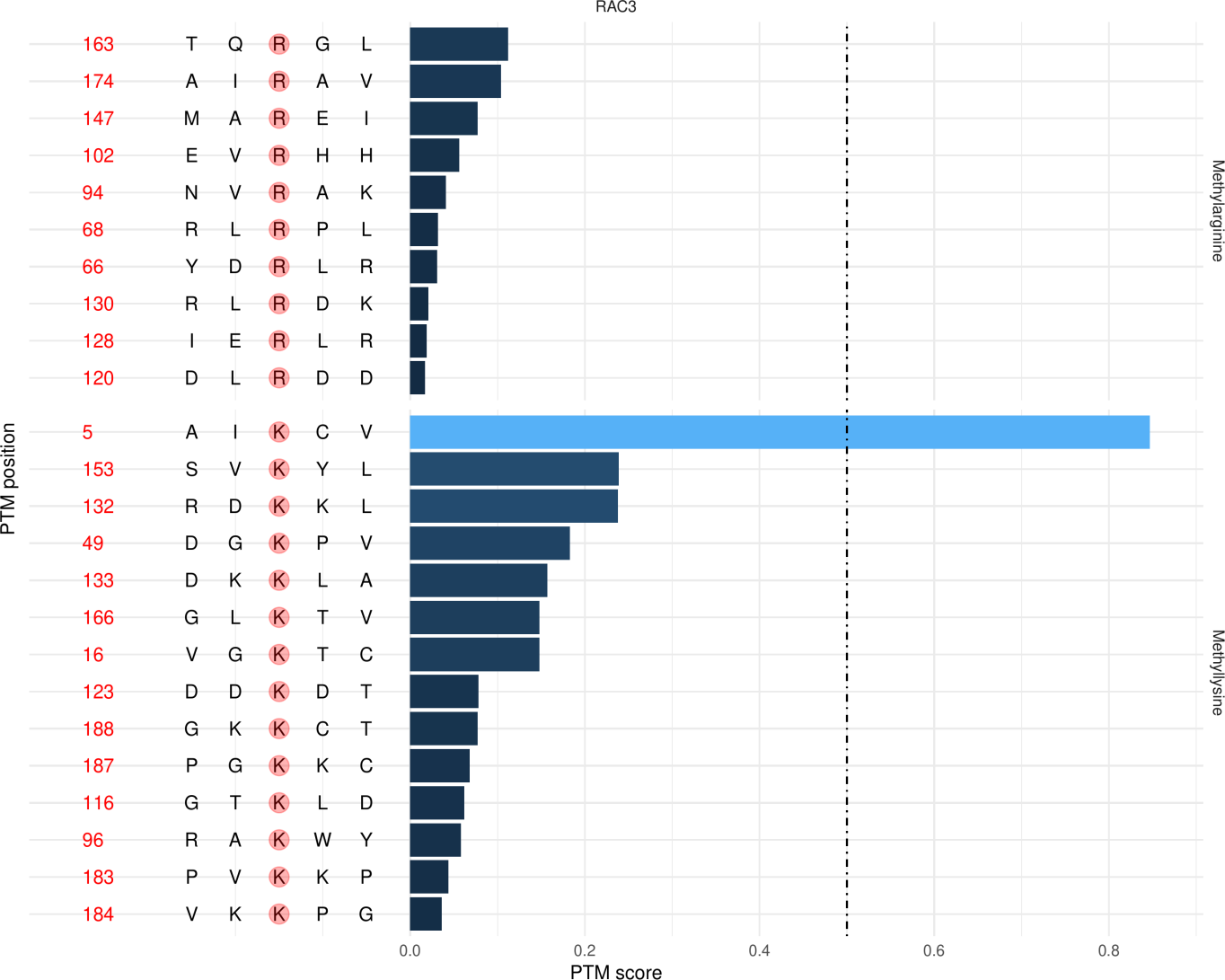
**Fig.** **12** SIGCOR 05 Methylarginine PTM numbers

**(File path: Figure+Table/SIGCOR-05-Methylarginine-PTM-numbers.pdf)**



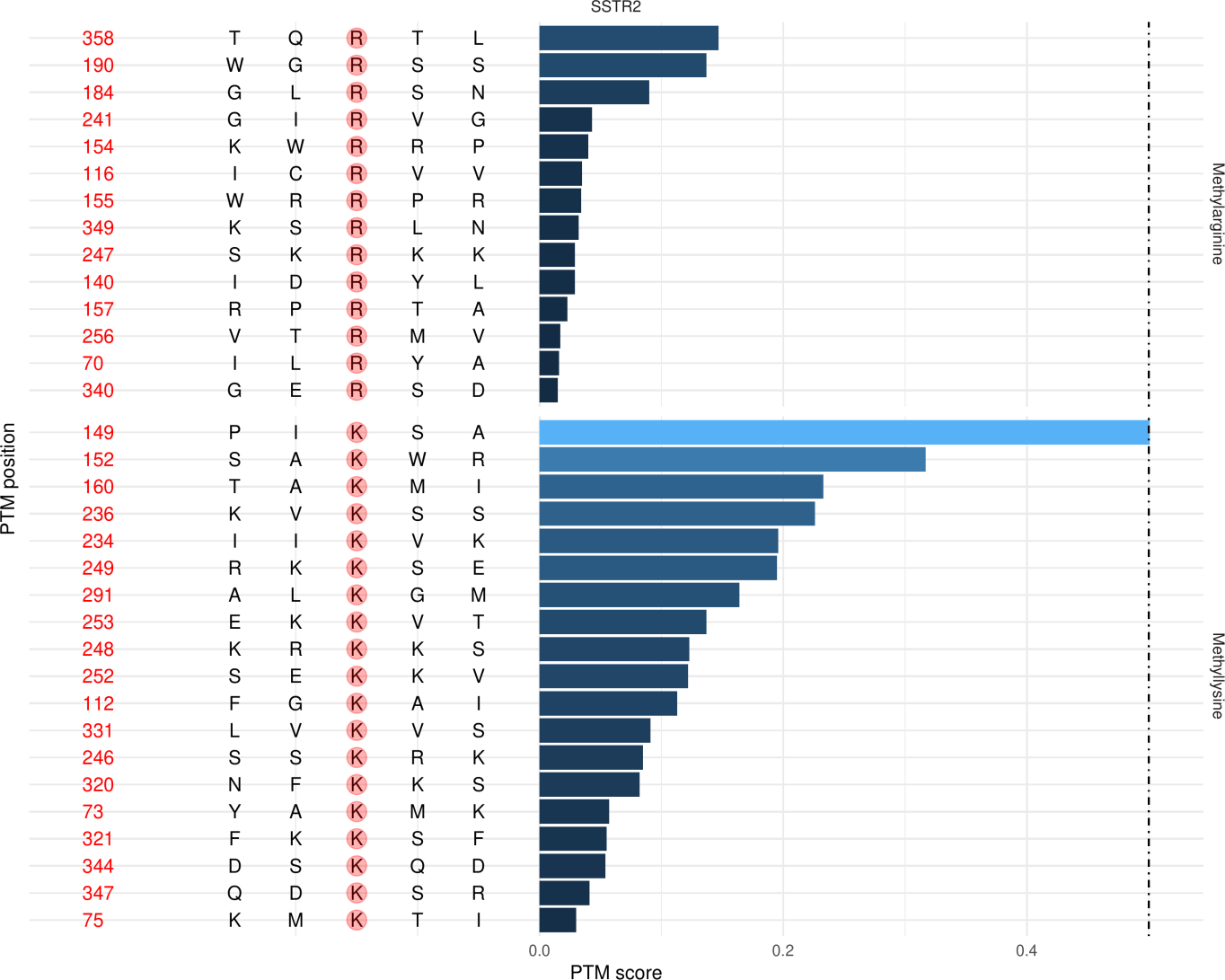
**Fig.** **13** SIGCOR 05 Methyllysine PTM numbers

**(File path: Figure+Table/SIGCOR-05-Methyllysine-PTM-numbers.pdf)**



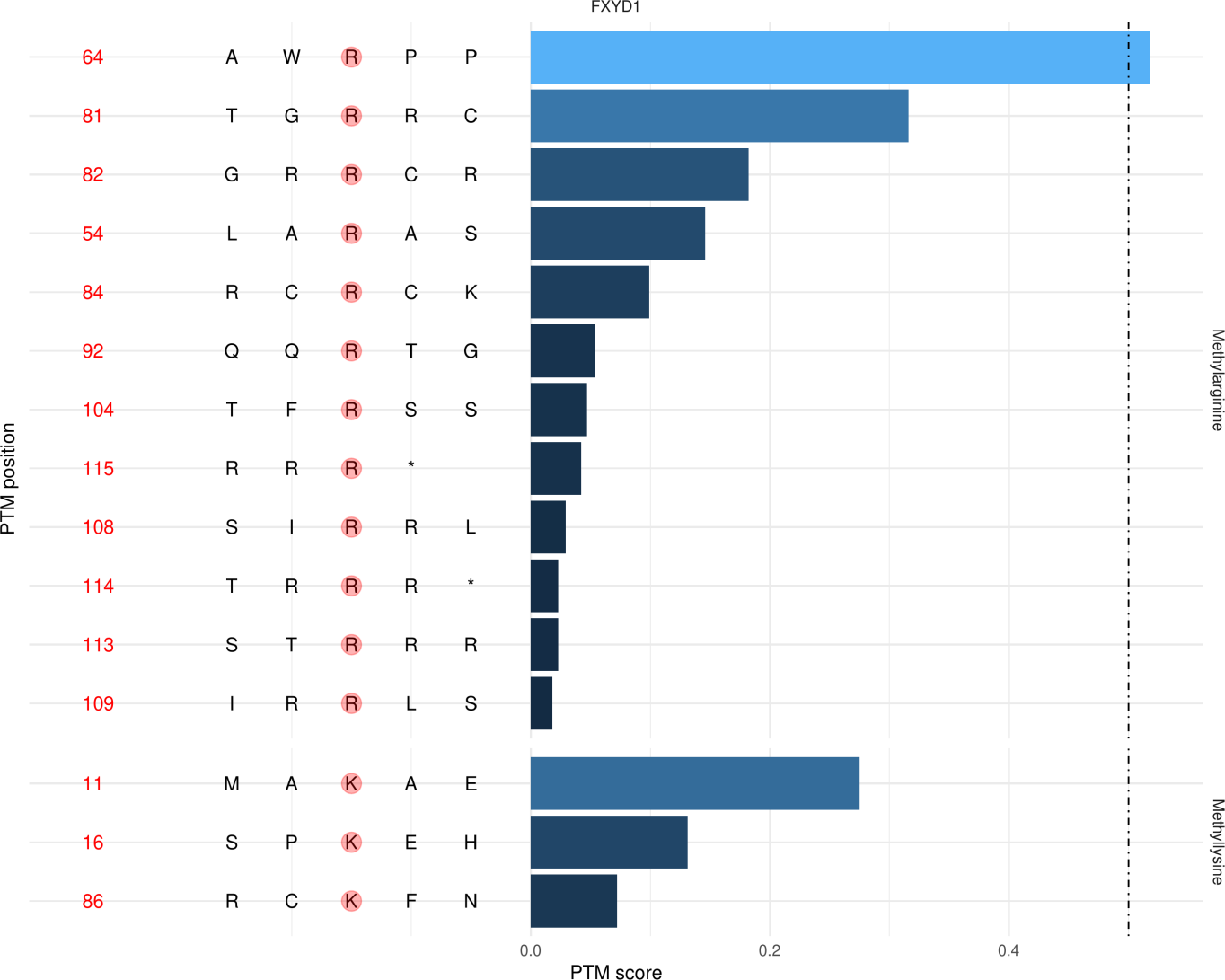
**Fig.** **14** SIGCOR 05 RAC3 PTM score

**(File path: Figure+Table/SIGCOR-05-RAC3-PTM-score.pdf)**



**Fig.** **15** SIGCOR 05 SSTR2 PTM score

**(File path: Figure+Table/SIGCOR-05-SSTR2-PTM-score.pdf)**



**Fig.** **16** SIGCOR 05 FXYD1 PTM score

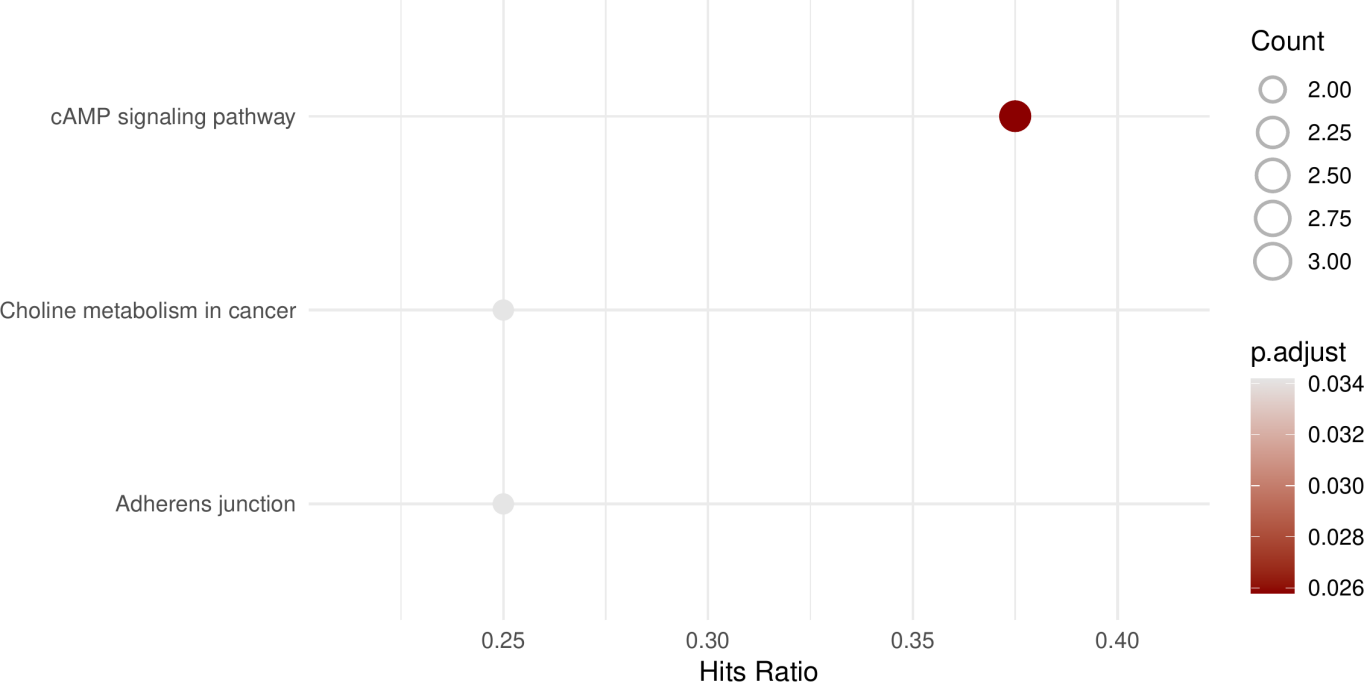
**(File path: Figure+Table/SIGCOR-05-FXYD1-PTM-score.pdf)**

Note: The directory 'Figure+Table/SIGCOR-05-All-PTM-score' contains 24 files.  
  
1 1\_ARL17A.pdf  
2 10\_FXYD1.pdf  
3 11\_HEBP2.pdf  
4 12\_KHK.pdf  
5 13\_MMP15.pdf  
6 ...

**(File path: Figure+Table/SIGCOR-05-All-PTM-score)**

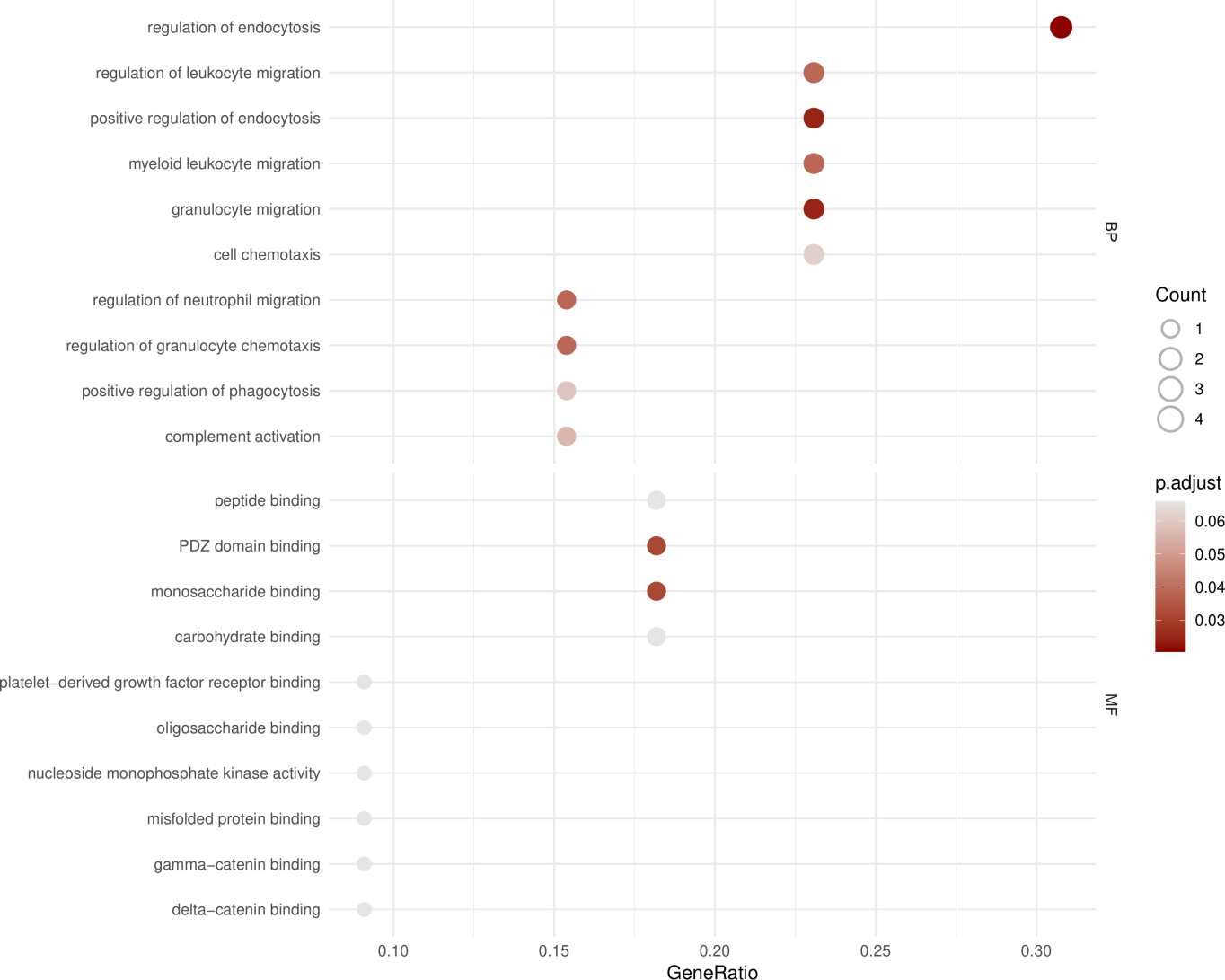
## 3.8 富集分析 (PTMS)

补充了上述有 PTMs 的基因的富集分析，发现同样聚焦于 cAMP 通路，见Fig. 。



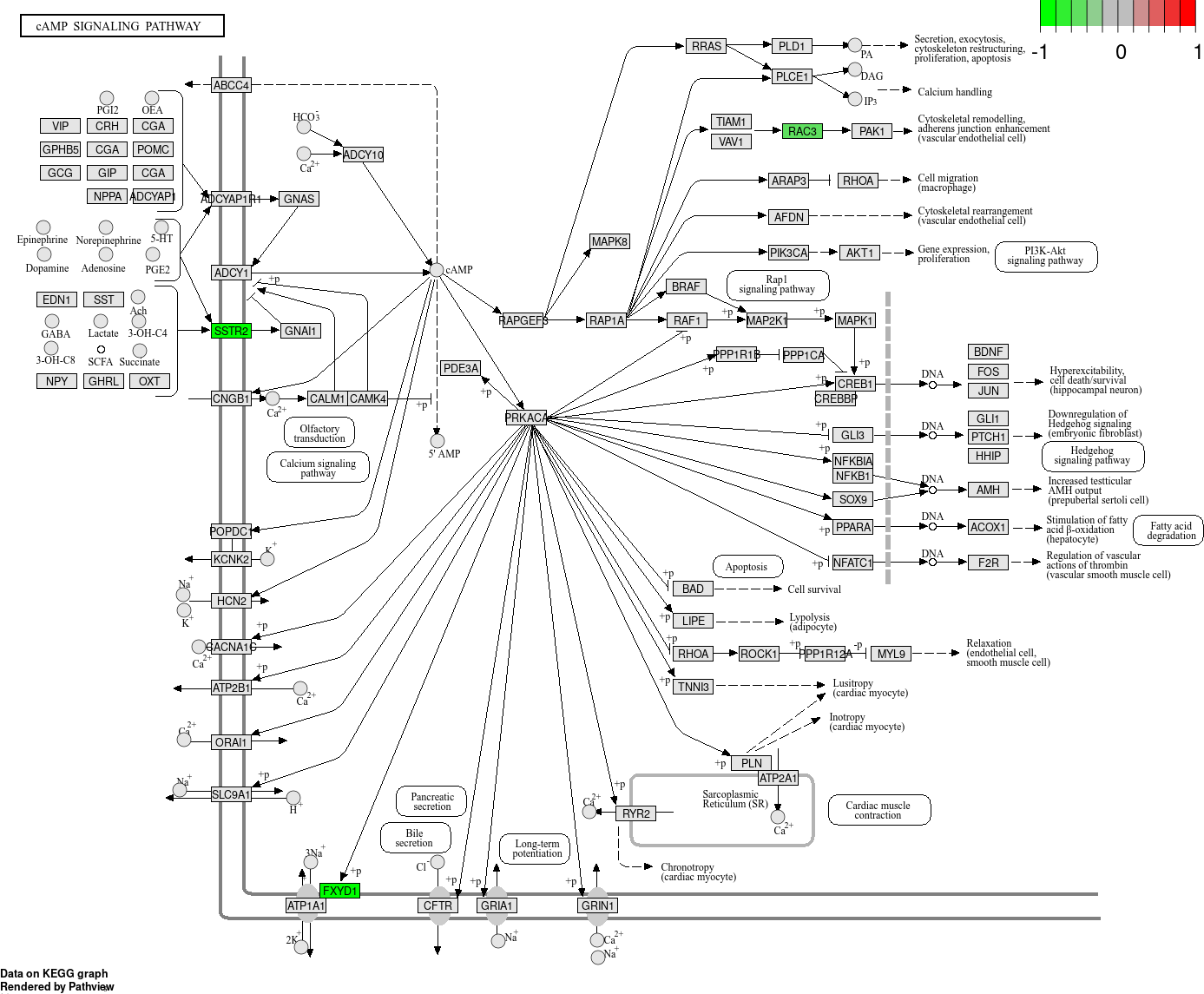
**Fig.** **17** PTMS KEGG enrichment

**(File path: Figure+Table/PTMS-KEGG-enrichment.pdf)**



**Fig.** **18** PTMS GO enrichment

**(File path: Figure+Table/PTMS-GO-enrichment.pdf)**



**Fig.** **19** PTMS hsa04024 visualization

**(File path: Figure+Table/PTMS-hsa04024-visualization.png)**

* Interactive figure:
* Enriched genes: SSTR2, FXYD1, RAC3

# 4 总结

筛选的甲基化调控因子为 PRDM6，可能调控的基因见 Tab. **[5](#correlation-results-05)** ， 富集分析结果中，cAMP 通路最为显著，Fig. **[10](#SIGCOR-05-hsa04024-visualization)** 。

cAMP 通路的三个基因的修饰位点见 Fig. **[14](#SIGCOR-05-RAC3-PTM-score)**, Fig. **[16](#SIGCOR-05-FXYD1-PTM-score)**, Fig. **[15](#SIGCOR-05-SSTR2-PTM-score)**。

# Reference

1. 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).

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

5. 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).

6. Wang, D. *et al.* MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research* **48**, W140–W146 (2020).