宫颈癌检测研究需要补充的生物信息学数据

一、PAX1和SOX1基因研究

研究目的：

找出PAX1和SOX1基因与宫颈癌的关系，揭示以PAX1和SOX1为靶标检测宫颈癌的理论依据。

**研究内容**：

1. PAX1和SOX1基因的染色体定位，拷贝数，表达部位和功能注释信息 ，输出以下表格：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 拷贝数 | 染色体定位 | 表达部位 | 功能注释 |
| SOX1 |  |  |  |  |
| PAX1 |  |  |  |  |

1. 不同等级宫颈癌与非癌甲基化水平比较（给出平均水平数据和差异显著性统计数据）

Response:

我们从以下两篇文献获取k450甲基化数据，另外加上TCGA的数据。

[1] Farkas S A, Milutin-Gašperov N, Grce M, et al. Genome-wide DNA methylation assay reveals novel candidate biomarker genes in cervical cancer[J]. Epigenetics, 2013, 8(11): 1213-1225.

[2] Verlaat W, Snoek B C, Heideman D A M, et al. Identification and validation of a 3-gene methylation classifier for HPV-based cervical screening on self-samples[J]. Clinical Cancer Research, 2018, 24(14): 3456-3464.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Reference | platform |  |  |  |  |
| normal | Pre-Cancer | cancer | Total |
| Ref01 | 450K | 20 | 18(CIN3) | 6 | 48 |
| Ref02 | 450K | 28 | 36(CIN3) | 4(SCC) | 68 |
| TCGA | 450K | 3 | 299 | | 302 |

接着，我们鉴别出PAX1和SOX1上的CpG岛，并在这些岛上采用滑动窗口的方式，检验岛上连续探针组合的区分能力，滑动窗口大小从2个探针到岛上全部探针数目。我们找到的DMR标记如下：

PAX1：normal vs cin3

Chromosome：chr20

Start: 21705635 End:21706741 width:1106

Sensitivity：0.62 Specificity：0.635 AUC：0.758

Probes：cg19054524; cg08448701; cg01783070; cg19079845; cg07684169; cg16767801; cg20552282; cg19122389

PAX1：normal vs cancer

Chromosome：chr20

Start: 21706090 End: 21706741 width:651

Sensitivity：0.875 Specificity：0.578 AUC：0.96

Probes：cg20552282; cg19122389

PAX1：cin3 vs cancer

Chromosome：chr20

Start: 21706090 End: 21706741 width:651

Sensitivity：0.834 Specificity：0.562 AUC：0.902

Probes：cg20552282; cg19122389

SOX1：normal vs cin3

Chromosome：chr13

Start: 112067459 End: 112068020 width:561

Sensitivity：0.608 Specificity：0.621 AUC：0.731

Probes：cg11199713; cg19407095; cg06675478; cg02547394; cg16705627; cg15466862

SOX1：normal vs cancer

Chromosome：chr13

Start: 112067472 End: 112068020 width:548

Sensitivity：0.878 Specificity：0.579 AUC：0.965

Probes：cg06675478; cg02547394; cg16705627; cg15466862

SOX1：cin3 vs cancer

Chromosome：chr13

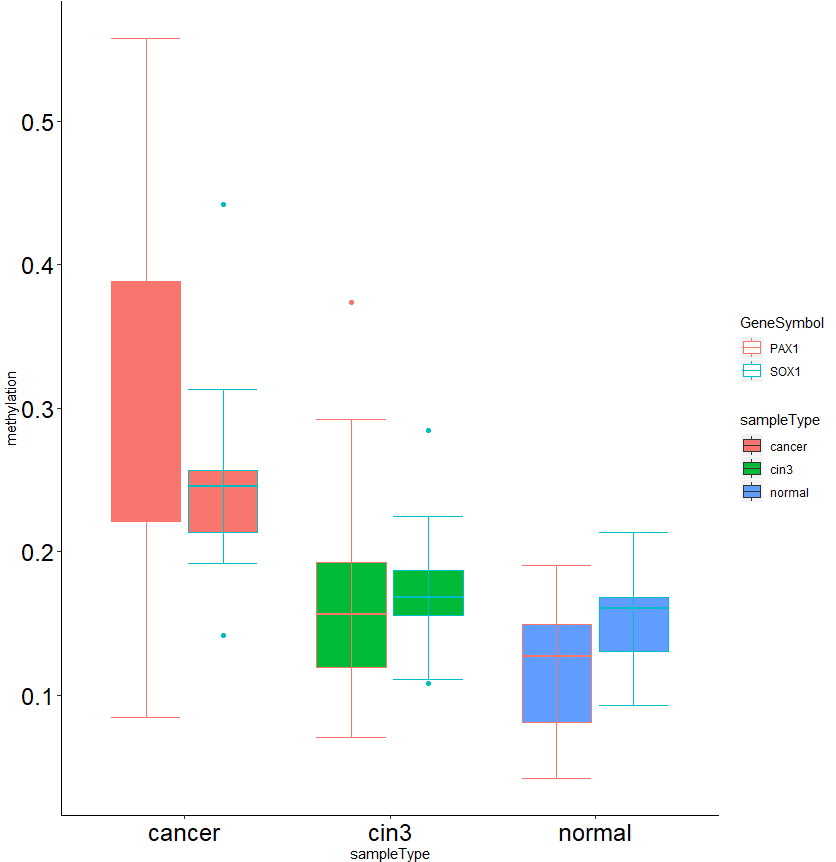
Start: 112067459 End: 112068020 width:561

Sensitivity：0.835 Specificity：0.562 AUC：0.904

Probes：cg06675478; cg02547394; cg16705627; cg15466862

用获得的DMR做特征，进行t-test，结果如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PAX1 | | SOX1 | |
|  | normal vs cin3 | normal vs cancer | normal vs cin3 | normal vs cancer |
| 样本数量 | 48 vs 54 | 48 vs 10 | 48 vs 54 | 48 vs 10 |
| 非癌均值 | 0.0765 | 0.13873 | 0.19133 | 0.18987 |
| 癌症均值 | 0.1174 | 0.32652 | 0.23740 | 0.41021 |
| p-value（t-test） | 1.5e-05 | 0.00035 | 1.5e-05 | 0.00018 |
| 95%置信区间 | -0.0585 -0.023372 | -0.26513 -0.11046 | -0.066095 -0.026052 | -0.30405 -0.41021 |



1. 不同等级宫颈癌与非癌转录水平比较（给出平均水平数据和差异显著性统计数据）。

Response:

宫颈癌与非癌在转录组上的比较，需要合适量的样本。TCGA上CESC的癌旁组织只有3个样本，统计结果不可靠。为此，我们综合TCGA和GTEx两个数据库的转录组数据集，GTEx上有11个正常宫颈的RNA-Seq样本。鉴于TCGA与GTEx两个Studies在做RNA-Seq实验的差异（样本个体、reads比对方式、Normalization 方法等），这两个数据集不能直接组合分析。需要经过集成处理，才能进行统计分析。以下文献已经做了这方面的工作，并且将处理后的数据放在figshare上共享：

Wang, Qingguo, Joshua Armenia, Chao Zhang, Alexander V. Penson, Ed Reznik, Liguo Zhang, Thais Minet et al. "Unifying cancer and normal RNA sequencing data from different sources." Scientific data 5 (2018): 180061.

t-test

|  |  |  |
| --- | --- | --- |
|  | PAX1 | SOX1 |
| 样本数量 | 252 cancer vs 14 normal | 252 cancer vs 14 normal |
| 非癌均值 | 1.202471 | 0.9481369 |
| 癌症均值 | 1.224325 | 0.9773810 |
| p-value（t-test） | 0.9813 | 0.9578 |
| 95%置信区间 | - 1.850615  1.806907 | - 1.114897  1.056409 |

1. PAX1和SOX1高甲基化群体与低甲基化群体生存分析图。

Response:

1） 我们用TCGA上宫颈癌的450k数据，该数据集包含307个病人的312个样本。我们保留单样本病人，获得302个样本；然后根据这302个病人的临床信息，删除那些没有随访信息的病人，最终保留246个病人的样本.

查询TCGA的探针信息，找到PAX1和SOX1对应的探针,去除全部为NA值的探针，剩下的有效探针分别为：

PAX1（18个）：

cg01783070; cg02155658; cg03140968; cg06654901; cg07213060;

cg07684169; cg08156066; cg08448701; cg12544951; cg15556943;

cg16767801; cg19054524; cg19079845; cg19122389; cg20432507;

cg20552282; cg22837767; cg22906273

SOX1（17个）:

cg00663972; cg01236132; cg01757312; cg02547394; cg04047221;

cg04865691; cg06675478; cg11199713; cg11750165; cg15466862;

cg16705627; cg19407095; cg19802138; cg20829347; cg24604013;

cg25463470; cg27301032

2）低甲基化样本与高甲基化样本的确定

首先，我们求出每一探针在所有样本上的甲基化beta值的均值（PAX1获得18个beta均值，SOX1获得17个均值）；然后对这些均值进行二次求平均。综合考虑样本在PAX1在探针上的均值分布，我们分别选两个cutoff值来区分样本属于低甲基化或高甲基化。

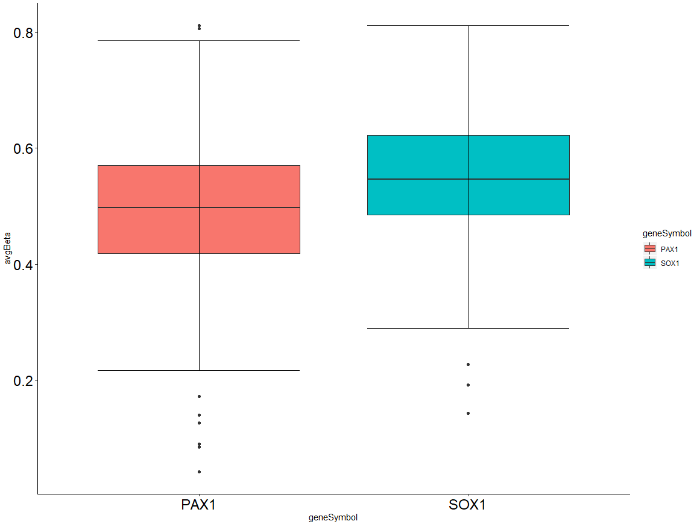
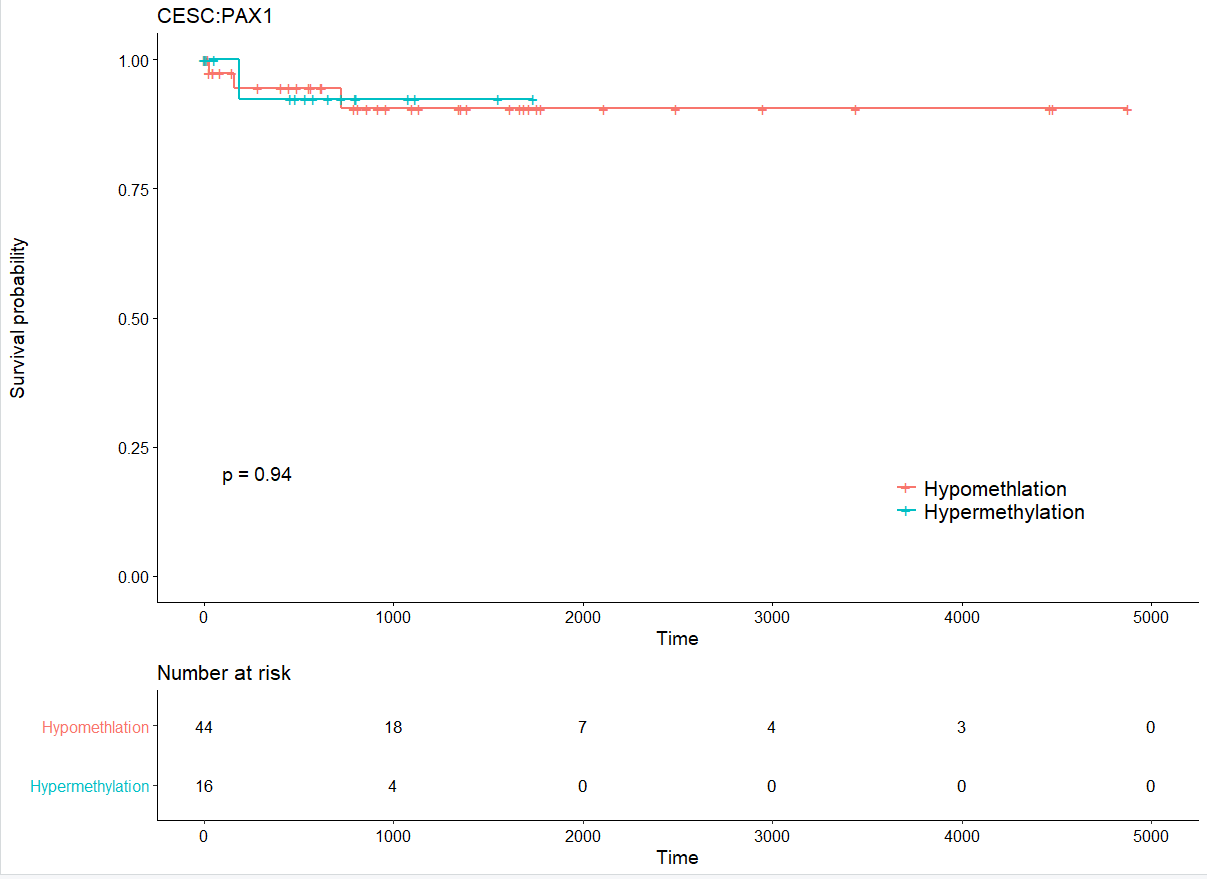
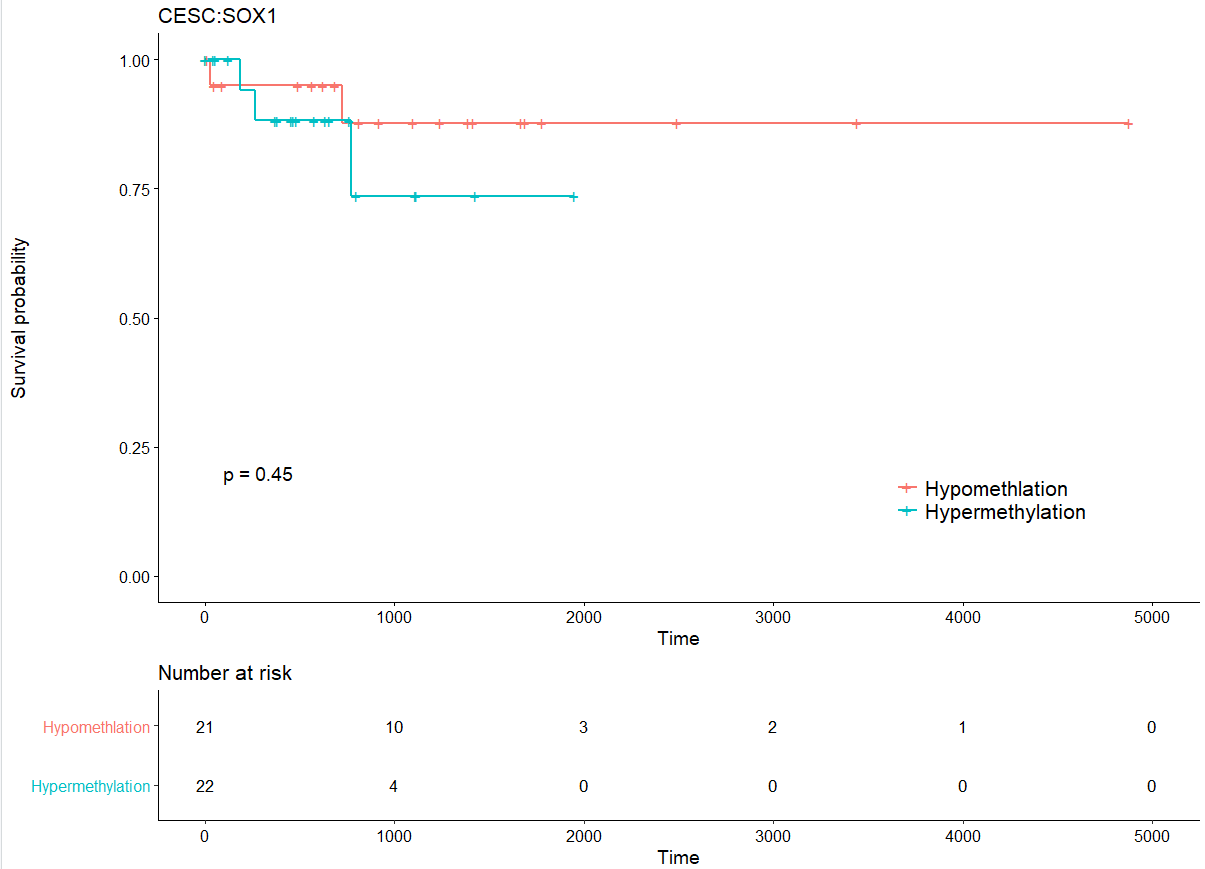


图-2：样本在PAX1和SOX1的探针上的均值分布

针对PAX1，我们样本的beta均值位于区间[0.3,0.6] 左右两侧的样本分别为低甲基化和高甲基化样本；针对SOX1，beta均值位于区间[0.4,0.7]左右两侧的样本分别为高、低甲基化样本

3）分别做SAX1和SOX1的生存分析图





1. PAX1和SOX1基因代谢通路分析，输出代谢通路图。
2. PAX1和SOX1基因以及双位点甲基化检测的对宫颈癌ROC曲线图分析，输出ROC曲线图及其相关数据。

Response：

我们采用问题2中参考文献1和参考文献2的GEO数据，构建binary 分类器(normal vs cin3, normal vs cancer, cin3 vs cancer)

