中山大学 生命科学学院

表观遗传期末论文

1 选题和简介（摘要）

本次的选题其实原本打算写的方向还不太一样，原本打算写肿瘤发生和表观遗传之间的关系。在研读了几篇综述之后，我对肿瘤的发生机制有了一些新的天真的想法。目前的主要的想法是 DNA 的突变的是肿瘤产生的原因，这诚然是对的，但是在儿童肿瘤中有相当一部分是没有常见的突变的(这也好理解，儿童肿瘤没有时间去积累突变)，或者说这些少数的突变通过对调控表观遗传的常见分子的蛋白质的变性，破坏 DNA 稳定性和表达的调控，达到常见的突变的效果（以上通过直觉得出的假设，但是我的假设一般还是挺准的。）

同时我不想抄一大段已经写好的综述和 AI 修改的内容来当作业。为了验证我以上的假设，我选取了低级别胶质瘤（Low-Grade Glioma, LGG）这种肿瘤的数据，其相对符合选题的目标肿瘤。接着我想先用非监督学习的方法，将表观遗传（准确来说是甲基化数据，目前只有甲基化数据）和预期寿命的进行预测学习，先大体看看相关性怎么样。接着如果结果可行的话，再用监督学习的方法找出权重最大的位点的哪个，我想那个位点应该就是已知的 H3.3 突变位点或者直接下游位点。

在写开头的时候我只做到大体验证可行。后面可能面临以下的问题：1，预期寿命收到很多因素的影响，可能归一因素的效果很差。2，并不是所有的低级别胶质瘤都满足我的假设模型，同一种癌症可能有着完全不同的发病机制。3，我对甲基化的数据不是很熟悉，其数据是 The Cancer Genome Atlas Program (TCGA) 数据库里面的，但我没法做任何的验证，也不知道数据是这么测出来的。但是我想反正的期末论文，体现出自己的思考就挺好的，所以这篇文章我也是边写边作分析的，希望能得出一点结论或者见解吧。

2 背景介绍（引言）

相似的和相关的表观遗传已经点明的途径：由 H3F3A 和 H3F3B 基因产生的组蛋白变体 H3.3，对于在发育和细胞分化过程中控制染色质结构及基因表达至关重要。与传统组蛋白不同，H3.3 的掺入不依赖于 DNA 复制，通常标志着活跃转录的区域。在包括弥漫性内生性脑桥胶质瘤（DIPG）和胶质母细胞瘤在内的儿科高级别胶质瘤（pHGG）中，H3.3 的体细胞突变，特别是第 27 位赖氨酸（K27M）和第 34 位甘氨酸（G34R/V）的突变，已被公认为主要的致癌驱动因素。

H3.3K27M 突变通过在第 27 位用蛋氨酸替换赖氨酸，改变了通常的翻译后修饰环境。这种突变对 PRC2 的甲基转移酶组分 EZH2 产生显性抑制，导致抑制性标记 H3K27me3 全局性下降，从而重塑表观基因组并开启致癌转录通路。尽管 H3K27me3 整体减少，但讽刺的是，特定基因的启动子可能会维持甚至获得 H3K27me3 标记，从而在特定位点产生复杂的限制性环境。

除了 K27M 之外，G34（如 G34R/V）突变会影响与染色质结合蛋白的关系，并对应特定的肿瘤部位和年龄组。G34 突变阻碍了组蛋白甲基转移酶 SETD2 的招募，导致 H3K36 三甲基化（H3K36me3）

减少，而这种改变与活跃转录和基因组稳定性有关。这些变化被认为与脑祖细胞中基因组不稳性的发展以及分化通路的改变有关。

3 初步分析

下面我会很详细的介绍我的每一步操作，说明代码的流程、思路和注意事项，原代码会上传到GitHub：https://github.com/4b4tpxh87j-sudo/epigene-study-on-LGG.git，里面会有详细的中文标注，由于时间关系，我未能写成更好看的 jupyternotebook 格式，请见谅。

3.1 TCGA 数据整合（get-embedding）

对于新的 TCGA 数据，我们做的第一件事就是查看数据的具体格式和内容。对于同一样本空间的数据的表头是一样的，可以以患者的 ID 为引导来将两个数据合并到一起，形成一个行为患者 ID，列为表观遗传位点强度值。这里的甲基化值是通过亚硫酸盐将没有甲基化的 C 变成 U，而甲基化的 C 不会改变，再通过探针杂交后比较荧光的强弱从而得到确定位点的甲基化数值。此方法可以做到高通量和较高的准确性。

注意：TCGA 的数据有一个特点：对于每一个患者 ID，不光有一行的数据，可能由于回访等情况出现一人多行的情况，对于这样的情况，可取两行的平均值或者删去一行。这里采取删去一行的操作。（我觉得这里当成两个样本来处理也是可以的）

原甲基化数据是一个 (516, 486427) 的矩阵，在去重复和加入生存时间以后后我们得到了 (511, 486429) 的矩阵。其格式是这样的：其中 NaN 是此甲基化位点没有测出，这里只是恰好展示了全是 NaN 数据的位置，OS.time 的单位是天数，OS 代表状态：1：代表死亡，而 0 代表着患者存活，时间代表着最后一次随访时间，为后面的模型预测初步失败埋下了一个坑。

```
--- 对齐后的最终数据 (merged_data) ---
形状：(511, 486429)
现在每一行是一个患者，列包含CpG探针值和临床数据。
              rs951295  rs966367  rs9839873  OS.time  OS
_PATIENT
TCGA-TQ-A7RV       NaN       NaN        NaN   1868.0   0
TCGA-DU-A5TT       NaN       NaN        NaN    743.0   0
TCGA-EZ-7264       NaN       NaN        NaN   1201.0   0
TCGA-E1-A7YN       NaN       NaN        NaN    727.0   1
TCGA-DU-6403       NaN       NaN        NaN    354.0   1
```

图 1: 连接后的数据格式

3.2 NaN 值去除和训练集测试集分割（get-embedding2）

在第二部分操作中，先去除了甲基化位点都为 NaN 的列，然后在对于单个列中少数的 NaN 位点的值又全部的列的平均值进行替换。这步是由于模型是无法处理带有 NaN 位点的矩阵的。（必须说明的是，用平均值来代替 NaN 是一个权宜之计，其并没有数据意义的支持，最好的方法一个是把 NaN 位置当作掩码，用其他行此位点和其他的关系学习得到，但这部分我还不太会。）

接着我们用方差:0.01,来筛选甲级化位点在不同样本中变化不大的位点。在筛选前一共有 421996 个位点，移除 274683 个位点，剩余 147313 个位点。这一步可以大大的减少需要考虑的维度。

再然后我们将所有的患者样本分类成训练集和测试集（留意这里，下面这步会带来很大的问题）。一共有 408 个训练集和 103 个测试集。

3.3 初步的无监督学习验证（Dimensionality-Reduction）

在甲基化位点和生存时间的关系学习中，我从简单的无监督学习中的线性关系开始：pca 分析。当我们采取用 100 个特征向量进行拟合，解释了 18.21% 的方差.

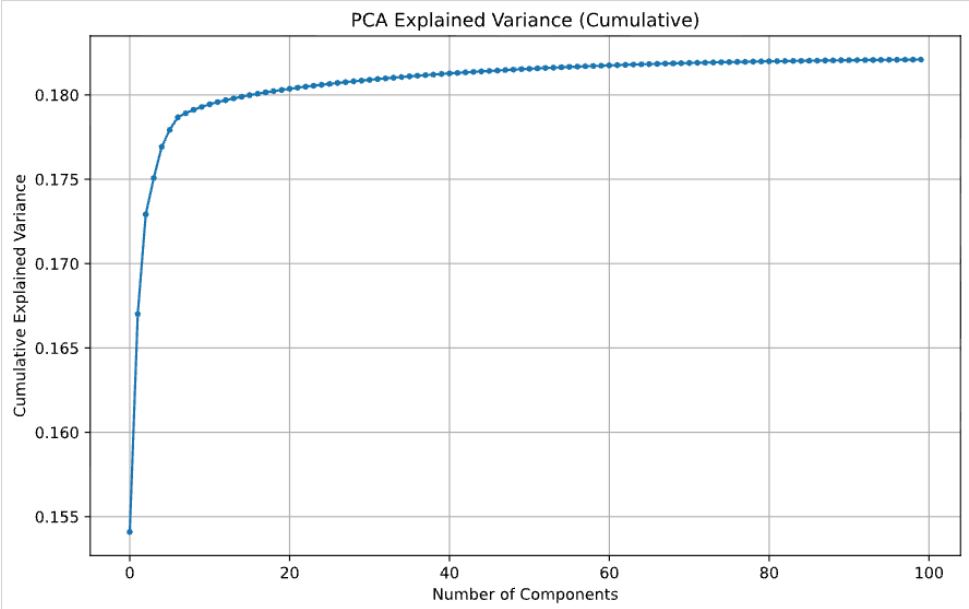

图 2: pca 拟合曲线

以此建立的随机森林回归训练的结果也可以说是惨不忍睹，平均绝对误差 (MAE): 698.78 天，均方根误差 (RMSE): 1062.20，R^2 分数 (解释性评分): -0.1061。可以说是完全没有预测的效果，误差达到了惊人的两年！

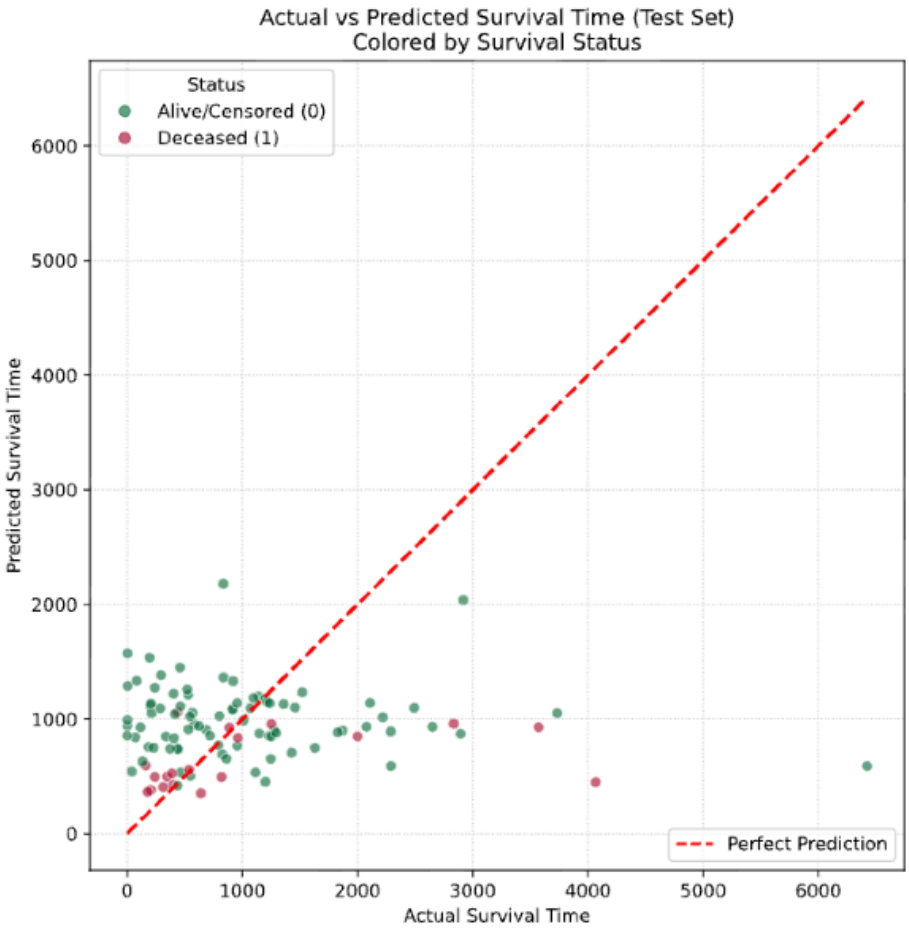

图 3: 拟合图像

在这张图片上我发现了一些问题，存活的人的时间是随访时间，并无法提供真确的生存时间信息，因此我接着只对死亡的样本进行训练和测试 (这也会带来幸存者偏差，因为这样会选中病情严重的人。)：训练样本：107 人，测试样本：19 人。前 100 个主成分总共解释了 99.26% 的方差。以此建立的随机森林回归训练的结果也可以说是惨不忍睹，平均绝对误差 (MAE): 平均绝对误差 (MAE): 641.18 天均方根误差 (RMSE): 1111.53 天,R^2 分数 (解释性评分): 0.0720.

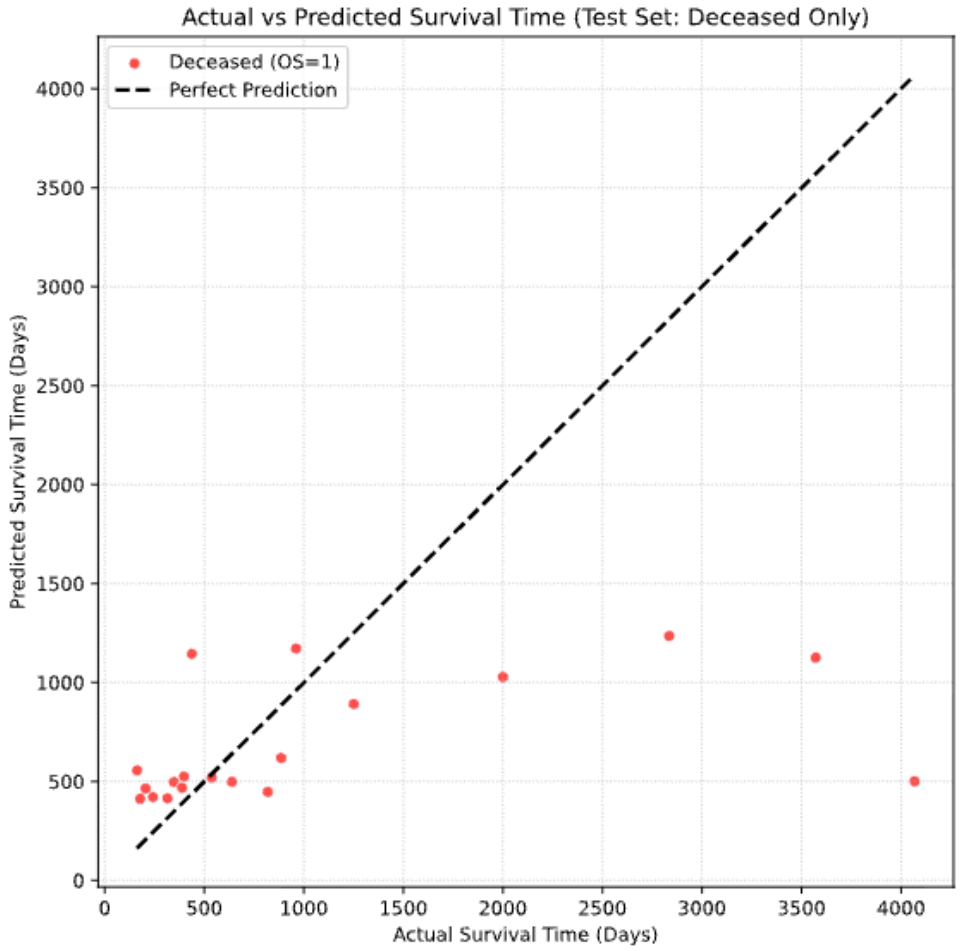

图 4: 全死亡样本的拟合图像

从以上结果中我们可以清晰的知道用简单的非监督学习是完全部分解释生存时间的。结下来我将运用监督学习的方法试一试。

4 基于监督学习的详细分析（specific）

鉴于非监督学习在处理大量的位点的过程中难以精准把握其中的奥秘所在，接下来我打算加入监督学习的部分，找出其中最关键的甲基化位点。同时说明：下面的监督学习过程都是建立在线性模型的基础上的，所以可以说这是相当简单的模型学习（关键是难得我也不会），同时在如此多维的因素中尝试将量的方差大小直接理解为作用的大小也是相当理想化的模型，再同时就算以上的位点都是正确的，影响强度也不代表着这是关键的决定性因素。

4.1 监督初筛和 Lasso 二筛

在这一部分中，我们将在第一阶段的有监督初筛中，程序利用统计学中的 F 检验（f-regression）来评估每一个自变量与目标变量（如生存时间）之间的线性相关性。这一步就像是一个“海选”过程，通过计算每个特征的 F 统计量和 p 值，将成千上万个原始位点按照相关性强弱排序，并强行保留得分最高的 2000 个特征。这种做法的主要目的是在不丢失重要信息的前提下，通过剔除大量明显的噪声特征，来显著减轻后续复杂模型的计算负担。进入第二阶段的 Lasso 二次筛选后，处理方式从单纯的统计评估转变为基于模型的正则化压缩。Lasso 回归通过在损失函数中引入 L-1 惩罚

项，能够将那些贡献较小或存在多重共线性的特征系数直接压缩为零。代码中使用的 LassoCV 带有交叉验证功能，它会自动寻找一个最优的惩罚力度参数，从而在保持模型预测能力的同时实现极致的特征稀疏化。最终，这 2000 个入围位点中绝大多数的系数会被归零，仅剩下那 25 个最具代表性且互补性最强的核心位点。

这种“先统计粗筛、后模型精选”的两步走策略，既保证了计算的效率，又利用了 Lasso 自动处理特征冗余的特性，使得最终提取出的位点在统计学和预测建模上都具有极高的价值。最终拟合结果为：平均绝对误差 (MAE): 631.57 天，R^2 分数: 0.1711。看起来效果依旧很一般，但是相比于之前的提升，我觉得已经充分的说明了监督学习为我们带来了相当有价值的信息。

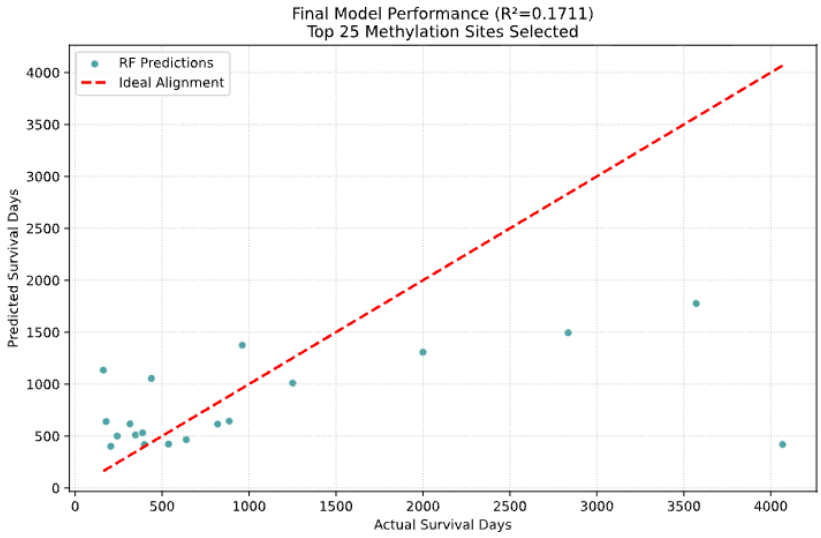

图 5: 监督学习的拟合曲线

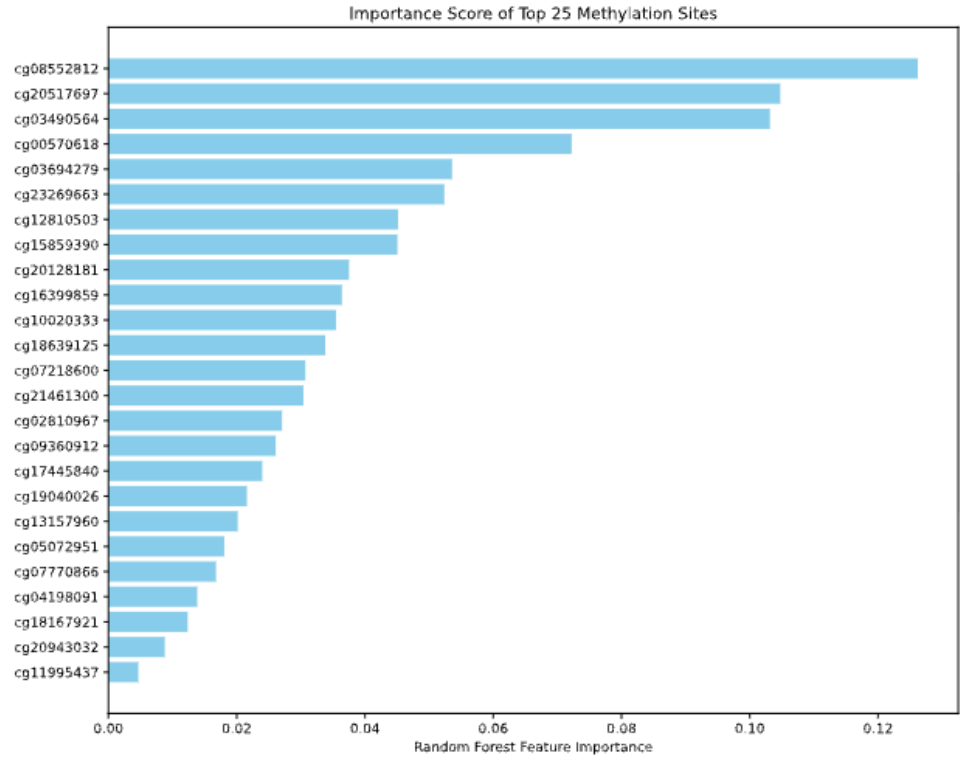

图 6: 作用位置最大的甲基化位点

4.2 作用位点的分析

说实话，我真的感谢自己本着认真负责的态度完成了本次论文的分析。原因在于：当我把第一个甲基化位点让 deepseek 分析的时候，其告诉我这是在 ZNF536 位点上，调控神经细胞的繁育和癌症的出现和发展有关。我一般不完全相信 AI 的回答。但是当我把同一个位点问 gemini 的时候，得到了完全一样的回答，一般人在这个时候就不会怀疑了，而且这个结果和预期的完美符合。但是当时我还是仔细写了在官网上查询的代码，得到了这样的结果：

```
(/home/luhc/conda_envs/cellplm) luhc@a-4U-GPU-Server:~/epigenetic$ python '/home/luhc/epigenetic/searching.py'
读取本地缓存文件: illumina_450k_meta.csv...

==============================
CpG 位点注释查询结果
==============================
            CHR      MAPINFO               UCSC_RefGene_Name                  UCSC_RefGene_Group Relation_to_Island               Island_Name
IlmnID
cg00570618    8   81143035.0                             NaN                                 NaN              NaN                       NaN
cg02810967    4   17813558.0                 NCAPG;DCAF16                    Body;TSS1500          S_Shore    chr4:17811954-17813227
cg03490564    2  112811719.0                      TMEM87B                         TSS1500          N_Shore  chr2:112811809-112812134
cg03694279   10   43788993.0                             NaN                                 NaN              NaN                       NaN
cg04198091    5  139283351.0          NRG2;NRG2;NRG2;NRG2  Body;Body;Body;Body                  Island  chr5:139283350-139284282
cg05072951   11   64949398.0                      CAPN1;CAPN1          1stExon;5'UTR          S_Shore   chr11:64948715-64949060
cg07218600   13  102404067.0                 FGF14;FGF14                Body;Body                 NaN                       NaN
cg07770866    8  121821052.0                       SNTB1                      Body          N_Shelf  chr8:121823534-121824720
cg08552812    2   73612302.0                       ALMS1                    TSS1500          N_Shore   chr2:73612655-73613661
cg09360912   12    7000722.0                         NaN                       NaN           Island    chr12:7000341-7000723
cg10020333    1  101704504.0                        S1PR1                     5'UTR          S_Shore  chr1:101702445-101702745
cg11995437   19   11484869.0                    C19orf39                    TSS1500          N_Shore  chr19:11485286-11485769
cg12810503    7  150254026.0                         NaN                       NaN              NaN                       NaN
cg13157960   19   33183277.0                      NUDT19                   1stExon           Island  chr19:33182609-33183562
cg15859390    4  184908510.0                       STOX2                      Body              NaN                       NaN
cg16399859    X   19689070.0           SH3KBP1;SH3KBP1                 Body;Body              NaN                       NaN
cg17445840    3  101546486.0  FAM55C;FAM55C;NFKBIZ  3'UTR;3'UTR;TSS1500                  NaN                       NaN
cg18167921    1   68963198.0            DEPDC1;DEPDC1          TSS1500;TSS1500          S_Shore    chr1:68962283-68963006
cg18639125    2   11051728.0                     KCNF1                    TSS1500          N_Shore    chr2:11051858-11053476
cg19040026    4  103789592.0  UBE2D3;CISD2;UBE2D3;UBE2D3  Body;TSS1500;5'UTR          N_Shore  chr4:103789994-103790510
cg20128181    6  135819596.0  C6orf217;AHI1;AHI1;AHI1;AHI1  Body;TSS1500;TSS1500;TSS1500;TSS1500          S_Shore  chr6:135818501-135819160
cg20517697   14   74208225.0                C14orf43;C14orf43                5'UTR;5'UTR                 NaN                       NaN
cg20943032   17   40913525.0  LOC100190938;RAMP2;LOC100190938  TSS1500;Body;TSS1500                  Island  chr17:40912816-40913553
cg21461300   13  114830702.0                       RASA3                       Body                 NaN                       NaN
cg23269663   19   45250480.0                        BCL3                    TSS1500          N_Shore  chr19:45251975-45252330
```

图 7: 真实的甲基化修饰位点

可以说，AI 就像是手底下喜欢糊弄人的家伙，倒不是说 AI 本质就是坏的，但是其数据库的污染（太多的假演示数据）和算法为了拟人的生成式要求造就了这一点。现在的 AI 已经不乱编文献了，但是像这样的问题还是要多多小心啊！

其中影响力因素强的那几个位点似乎并没有和癌症直接的关联，第五行的 NRG2 倒是目前有证据证明其影响神经细胞的生长的分化，下面的维基百科的原文：Neuregulin 2 (NRG2) is a novel member of the neuregulin family of growth and differentiation factors. Through interaction with the ErbB family of receptors, NRG2 induces the growth and differentiation of epithelial, neuronal, glial, and other types of cells. The gene consists of 12 exons and the genomic structure is similar to that of neuregulin 1 (NRG1), another member of the neuregulin family of ligands. NRG1 and NRG2 mediate distinct biological processes by acting at different sites in tissues and eliciting different biological responses in cells. The gene is located close to the region for demyelinating Charcot-Marie-Tooth disease locus, but is not responsible for this disease. Alternative transcripts encoding distinct isoforms have been described.[5]

这也算是我本次文章的一点点小小的惊喜吧，当然从那么多位的数据中只用线性的方法找到 25 条，本身还是在逻辑上是有问题的，这点不得不承认。上面的每一个位点我并没有每一个都详细的去了解其具体的作用，如果有读者熟悉的基因，也算是向读者提供了其还有可能对于癌症影响的可能，那这篇简单的文章对我来说也算是很有意义了。

5 结语

感觉写这么多还是挺有意思的，可能以后有时间还会继续完善在这个题目。有什么建议的话，欢迎到 GitHub 上面留言。