

学校代码：10246

学号：09111130001



## 博士学位论文

# 基于全基因组 DNA 甲基化谱式的肿瘤诊断及 预后生物标记物研究

Cancer Biomarker Research Based on  
Genome-wide DNA methylation Profile: Diagnosis and  
Prognosis

院系：生命科学学院

专业：遗传学

姓名：郭士成

指导教师：金 力 教授

王久存 教授

熊墨森 教授

完成日期： 2014 年 12 月

# **基于全基因组 DNA 甲基化谱式的肿瘤 诊断及预后生物标记物研究**

**郭士成**

**导师：**

**金 力 教授**：复旦大学现代人类学教育部重点实验室

**王久存 教授**：复旦大学现代人类学教育部重点实验室

**指导小组：**

**熊墨森 教授**：德克萨斯大学休斯顿健康科学中心

**复旦大学生命科学学院  
现代人类学教育部重点实验室**

**献给岁月**



# 目录

中文摘要.....	2
Abstract.....	4
缩写词汇.....	7
第一章 人类基因组 DNA 甲基化概述 .....	8
1.1 人类基因组中的 DNA 甲基化.....	9
1.1.1 DNA 甲基化定义及分布特征 .....	9
1.1.2 DNA 甲基化系统的进化史 .....	11
1.1.3 DNA 甲基化系统主体蛋白 .....	12
1.1.4 DNA 甲基化与基因表达调控 .....	13
1.2 DNA 甲基化与人类发育 .....	14
1.2.1 DNA 甲基化与人类复杂疾病及肿瘤 .....	14
1.2.2 DNA 甲基化与胚胎发育 .....	16
1.2.3 DNA 甲基化在配子形成中的动态变化 .....	18
1.3 DNA 甲基化与人类群体遗传学 .....	22
1.3.1 DNA 甲基化与人类进化 .....	22
1.3.2 DNA 甲基化与减数分裂重组热点 .....	23
1.4 参考文献.....	25
第二章 DNA 甲基化标记物在肺癌诊断中的机遇和挑战 .....	39
2.1 背景.....	39
2.2 材料和方法.....	40
2.2.1 检索策略和数据提取 .....	40
2.2.2 Meta 合并比值比分析 .....	41
2.2.3 Meta-regression 分析 .....	41
2.2.4 敏感度分析 .....	42
2.2.5 发表偏倚分析 .....	42
2.2.6 综合受试者工作特征曲线.....	42

2.2.7 TCGA 数据提取和再验证.....	44
2.3 结果.....	44
2.3.1 候选文献的基本特征.....	44
2.3.2 Meta 合并分析 .....	46
2.3.3 Meta 亚组分析 .....	47
2.3.3 Meta 回归分析 .....	49
2.3.3 综合受试者工作特征曲线.....	50
2.3.4 偏倚分析和稳定性分析.....	51
2.3.5 独立 TCGA 肺癌的数据集的再验证分析 .....	55
2.4 结论.....	59
2.5 讨论.....	59
2.6 参考文献.....	62
<b>第三章 基于跨平台甲基化芯片数据建立肺癌早期诊断模型 .....</b>	<b>67</b>
3.1 研究背景.....	68
3.2 材料和方法.....	70
3.2.1 实验设计和流程.....	70
3.2.2 患者样本及 DNA.....	72
3.2.3 MSD-SNuPET：甲基化状态相关的单核苷酸引物延伸法.....	73
3.2.4 MSD-SNuPET 技术中甲基化信号的估计.....	74
3.2.5 统计分析和机器学习 .....	75
3.2.5.1 Combat 批次效应处理 .....	76
3.2.5.2 随机森林学习模型 .....	76
3.2.5.3 逻辑斯蒂回归预测模型 .....	77
3.2.5.4 支持向量机预测模型 .....	77
3.2.5.5 贝叶斯树预测模型 .....	78
3.3 结果.....	79
3.3.1 公共数据的收集和合并.....	79
3.3.2 批次效应的评估和消除 .....	79
3.3.3 筛选阶段最优预测变量选择 .....	80
3.3.4 MSD-SNuPET 对甲基化状态的验证.....	81
3.3.5 基于 DNA 甲基化的肺癌预测模型 .....	84

3.4 结论.....	85
3.5 讨论.....	85
3.6 参考文献.....	88
第四章 全基因组甲基化图谱揭示泛癌甲基化特征.....	92
4.1 背景.....	93
4.2 材料和方法.....	95
4.2.1 数据.....	95
4.2.2 方法.....	95
4.2.2.1 数据清理 (Data Cleaning) .....	95
4.2.2.2 变量相关系数抽样分布 (Variable Correlation Distribution) .....	95
4.2.2.3 差异甲基化位点分析 (Differential Methylation Loci Detection) .....	96
4.2.2.4 随机森林和变量重要性 (Random Forest and Varibale Importance) .....	96
4.2.2.5 非特异变量过滤 (Non-Specific Filtering) .....	97
4.2.2.6 特异变量过滤 (Specific Variable Filtering) .....	97
4.2.2.7 分段组合法 (Segmentation and Recombination) .....	98
4.2.2.8 主成分分析 (Principal Components Analysis, PCA) .....	98
4.2.2.9 多维尺度分析 (Multidimensional Scaling, MDS) .....	99
4.3 结果.....	101
4.3.1 全基因组甲基化位点相关系数分布.....	101
4.3.2 泛癌样本相关性分布特征.....	102
4.3.3 泛癌甲基化数据集的主成分分析.....	106
4.3.4 泛癌甲基化数据集的多维尺度分析.....	107
4.3.5 泛癌甲基化数据集的聚类分析.....	109
4.3.6 泛癌样本的差异显著甲基化位点总体特征.....	110
4.3.7 相邻 CpG 位点之间的相关性 .....	112
4.3.8 泛癌样本的随机森林预测模型.....	114
4.3.9 肺癌差异甲基化谱式特征.....	117
4.3.9.1 肺腺癌最重要的甲基化诊断位点.....	122
4.3.9.2 肺鳞癌最重要的甲基化诊断位点.....	123
4.3.10 乳腺癌最重要的甲基化诊断位点.....	124
4.3.11 结肠癌最重要的甲基化诊断位点.....	125
4.3.12 头颈部鳞状细胞癌最重要的甲基化诊断位点.....	126

4.3.13 肾透明细胞癌最重要的甲基化诊断位点.....	127
4.3.14 肾乳头状细胞癌最重要的甲基化诊断位点.....	128
4.3.15 肝癌最重要的甲基化诊断位点.....	129
4.3.16 前列腺癌最重要的甲基化诊断位点.....	130
4.3.17 甲状腺癌最重要的甲基化诊断位点.....	131
4.3.18 子宫内膜癌最重要的甲基化诊断位点.....	132
4.4 结论.....	133
4.5 讨论.....	133
4.6 参考文献.....	136
<b>第五章 甲基化特异富集测序技术鉴定甲基化肿瘤标记物.....</b>	<b>138</b>
5.1 研究背景.....	138
5.2 材料和方法.....	141
5.2.1 患者和对照样品信息.....	141
5.2.2 细胞系和正常膀胱粘膜组织.....	143
5.2.3 MBD-methylCap 测序及其分析.....	143
5.2.3.1 MACS2 对甲基化富集区域预测 .....	145
5.2.4 MSP 和 BSP 方法介绍.....	145
5.2.4.1 MSP 原理.....	146
5.2.4.2 BSP 原理.....	146
5.2.4.3 甲基化检测流程.....	147
5.2.5 多阶段标记物验证流程.....	147
5.2.6 临床相关的统计方法.....	148
5.2.5.1 单因素和多因素 Logistic 回归分析 .....	148
5.2.5.2 Kaplan-Meier 估计和 Log-Rank 检验 .....	148
5.2.5.3 Cox 比例风险模型 .....	149
5.3 结果.....	150
5.3.1 膀胱癌和正常组织的全基因组甲基化图谱 .....	150
5.3.2 BSP 方法对膀胱癌及正常组织特异甲基化谱式的验证 .....	151
5.3.3 基于尿液样本的潜在膀胱癌甲基化标志物筛选 .....	154
5.3.4 临床样本中 8 位点甲基化标记物组合的诊断价值评估 .....	155
5.3.5 甲基化诊断模型具有与膀胱镜诊断相近的诊断效果 .....	156

5.3.6 甲基化模型可用于手术切除效果有效性的评估 .....	156
5.3.7 VAX1 和 LMX1A 基因高甲基化可用于癌症复发预测.....	156
5.4 结论.....	160
5.5 讨论.....	160
5.6 附录信息.....	165
附表 1. BSP 所涉及的引物对序列.....	165
附表 2. MSP 所涉及的引物对序列.....	165
附表 3. MethylCap-seq 甲基化文库的测序基本信息 .....	168
附表 4. BSP 对甲基化文库的验证.....	169
附表 5. 不同基因组合诊断模型的 ROC 表现.....	170
附表 6. 与膀胱癌已有甲基化诊断模型比较 .....	171
5.7 参考文献.....	172
第六章 总结及展望.....	177
6.1 总结.....	177
6.2 课题的意义与不足.....	181
6.3 展望.....	181
全文图目录.....	183
全文表目录.....	185

## 中文摘要

DNA 甲基化是一种古老的、进化上十分保守的表观修饰系统。DNA 甲基化适中的可塑性使其可以承载极其丰富的信息，以应对机体复杂的功能及高度多样的生理及病理状态。DNA 甲基化具备典型的优良生物标记物的特征，可应用于疾病的风险评估、病理诊断、药物反应的预测及治疗的预后评价。然而，DNA 甲基化作为肿瘤标记物的潜力和挑战尚待系统性评估。本研究通过对已知候选基因的甲基化关联分析的再分析、全基因组甲基化芯片分析及高通量测序分析，并结合在中国人群中的实验验证，阐述了 DNA 甲基化作为肿瘤生物标记物的巨大潜力。

迄今已有大量文献报道 DNA 甲基化可以作为肿瘤的诊断或预后标记物，但其作为诊断标记物的有效性及在临床转化中的应用仍面临挑战。本论文采用 Meta 分析方法对基于候选基因研究策略的 *APC* 基因启动子区高甲基化对肺癌诊断效能进行了系统及定量的评估。随机效应模型分析显示 *APC* 基因与非小细胞肺癌（NSCLC）之间存在显著关联。单基因 *APC* 非小细胞肺癌诊断的 Summary ROC (SROC) 的 AUC 可以达到 0.64。此外，通过亚组分析和 Meta 回归分析发现肺癌诊断时的年龄、自体或异体的对照、腺癌在总样本中的比率以及所采用引物的区别是造成不同研究之间出现异质性的主要原因。而样本类型（血清或实体组织）、性别比例、TNM 早期样本比例和检测方法对 *APC* 甲基化与非小细胞肺癌的关联性无显著影响，为将 DNA 甲基化应用于临床提供了理论依据，并指出了临床应用中面临的主要挑战及注意事项。

DNA 甲基化肿瘤诊断模型面临着两个重要的难题。第一、单个基因 DNA 甲基化肿瘤诊断模型诊断的灵敏性和特异性很难达到临床应用级别，需要采用一组 DNA 甲基化的联合谱式对肿瘤进行准确的诊断，因此筛选并确定最优 DNA 甲基化组合成为首选。第二、受研究经费和临床有效资源等条件的限制，单个研究中用于建立诊断模型的样本量一般都较小，导致由此建立起的预测模型的外展性较低。为解决上述问题，作者首先收集、整理了公共数据库中所有甲基化检测平台的非小细胞肺癌相关的全基因组甲基化数据，以组建用于 NSCLC 肿瘤标记物开发的探索阶段的信息；其次，针对不同平台的甲基化芯片，采用标准化、批次效应消除等处理方法进行噪音消除。最终，本研究建立了一套最优的 5 甲基化位点组合，并在中国非小细胞肺癌样本进行了验证。结果表明 5 个甲基化标记物 (*AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1*) 可形成一个有效的标记物组

合，以实现对非小细胞肺癌的早期预测和诊断。本研究为有效利用已有研究数据、降低研究成本、进行生物标记物开发提供了行之有效的思路和策略。

全基因组 DNA 甲基化改变是所有肿瘤共通的一种表观遗传异常形式，而不同肿瘤又有其特异的甲基化异常谱式，因此理论上可采用一组特定的 DNA 甲基化组合以实现对多种肿瘤的判定。本研究收集整理了公共数据库中 11 种肿瘤 1274 个样本的全基因组甲基化芯片 HM450K 数据（泛癌甲基化数据），通过主成分及聚类分析，发现肿瘤全基因组 DNA 甲基化可全面反映不同样本之间、不同肿瘤之间的相似程度。进一步基于随机森林多类分类预测模型对泛癌数据进行分析，发现 20-50 个甲基化标记物组合即可对 11 种肿瘤的 22 种状态（11 组不同种类或亚型的肿瘤和正常）同时进行高效的预测。基于“炎症及肿瘤组织的 DNA 甲基化状态可在血浆游离 DNA 中有效呈现”的理论，本研究为基于血浆游离 DNA 甲基化进行多种肿瘤的筛查及预测的可行性提供了理论基础。

上述基于高密度 DNA 甲基化芯片技术对肿瘤标记物筛查的方法虽具有位点特异、价格低廉等优势，但在标记物筛查及诊断模型建立等早期阶段，芯片技术仍存在容易遗漏潜在重要标记物的可能。基于全基因组测序技术的甲基化特异结合-富集-深度测序方法可以弥补 DNA 甲基化芯片的这一缺点。因此，本研究开发了基于 MethylCap 结合二代测序技术的肿瘤甲基化标记物流程。以膀胱癌为研究对象，首先利用 MethylCap-Seq 建立了膀胱癌全基因组甲基化谱式，进而确立了潜在的肿瘤相关的 DNA 甲基化标记物。通过多阶段生物标记物验证流程对 104 个候选基因进行了筛选。采用小样本筛选、大样本验证、双重独立数据验证、临床双盲模拟等实验设计和分析成功建立了 5 个膀胱癌诊断标记物 (*VAX1*、*KCNV1*、*TAL1*、*PROX1* 和 *CFTR*)、2 个膀胱癌复发标记物 (*VAX1* 和 *LMX1A*) 和 2 个膀胱癌分化标记物 (*ECELI* 和 *TMEM26*)。

关键词： DNA 甲基化，肿瘤，诊断，预测，标记物，甲基化谱，聚类分析，甲基化状态依赖的单核苷酸引物延伸技术，

中图分类号： R394, Q31

## **Abstract**

DNA methylation is one of most important genetic modification systems. In perspective of evolution, DNA methylation modification is one of most conservative epigenetic modification systems, widespread from prokaryotes to mammals. DNA methylation has important structural and biological significance in the human genome. What's more, the moderate plasticity of DNA methylation keep it can carry extremely rich information storage and can correspond to the body's complex functions and highly non-redundant physiological and pathological states. Therefore, DNA methylation has excellent characteristics to be a biomarker which can be applied to the risk assessment, diagnosis, drug response prediction and prognosis evaluation of treatment. However, there is no such comprehensive study to descript the performance of the DNA methylation as a biomarker, especially in cancer. In this study, we apply Meta-analysis, genome-wide microarray technology and high-throughput sequencing techniques to evaluation the role and potential of the DNA methylation as a tumor biomarker.

The DNA methylation status has been reported to be associated with cancers in large number of association studies in the realm of diagnosis and prognosis. However, only a few diagnosis productions, especially diagnostic markers, have been released, indicating the challenge for DNA methylation diagnosis development. In the third part of the study, we try to descript the opportunity and challenge of biomarker development for cancer diagnosis based on DNA methylation, with the example of the project: quantitative assessment of the diagnostic role of APC promoter methylation in non-small cell lung cancer. The hypermethylation of the promoter region of APC was significantly associated with NSCLC based on random effected model in Meta-analysis. The AUC of the summary receiver operating characteristic curve analysis (SROC) was 0.64, indicating APC promoter DNA hypermethylation have intermediate prediction power. Subgroup analysis showed that there was a significant difference in the OR for APC methylation within the different subgroup of age at diagnosis, the primer, proportion of adenocarcinoma samples and types of the control samples, indicating these four factors were the most important sources of heterogeneity. Meta-regression

analysis confirmed two of the above four potential sources of heterogeneity, including age at diagnosis and the primers. On the other side, sample type (serum and solid tissue), gender proportion, TNM distribution in the samples and DNA methylation detection technique (MSP or QMSP) were not significantly associated with odds ratio of APC. Therefore, the methylation status of APC promoter was significantly associated with NSCLC, especially with lung adenocarcinoma. It is surprise result the AUC of the prediction model based on one gene can come up to 0.64, showing the opportunity of the DNA methylation as the cancer biomarker. However, large number of the heterogeneity provided challenge for the clinical utilization of APC methylation test.

DNA methylation was suggested as the promising biomarker for lung cancer diagnosis. However, the performance of the prediction model for lung cancer in the Pan-cancer dataset is not well. It would be caused by the small sample size. The stability of the prediction model would be very poor when the sample size is small. And almost all the DNA methylation based diagnosis models to distinguish cancer from normal were built with candidate gene strategy, which is difficult to achieve the maximum prediction accuracy. It is a great challenge to search for the optimized combination of the methylation biomarkers to obtain the maximum diagnosis performance. Therefore, the present diagnosis models need to be trained in bigger training dataset and need to apply feature selection to guarantee the accuracy and stability. In this study, we collect 3 dataset with large sample size from public database GEO as the discovery data to developed a panel of DNA methylation biomarkers with high prediction efficiency and the model was validated successfully by MSD-SNuPET (methylation status determined single nucleotide primer extension technique) in a large Chinese Han NSCLC retrospective cohort with 150 pairwise NSCLC/normal tissues. We found that high throughput DNA methylation microarray dataset followed by batch effect elimination can be a good method to discover optimized DNA methylation diagnostic panels with low cost. Methylation profiles of *AGTR1*、*GALR1*、*SLC5A8*、*ZMYND10* and *NTSR1*, could be an effective methylation-based assay for the NSCLC diagnosis.

Now, the tissues/fluids for the diagnosis of each cancer were different, therefore, the detection method and protocol were different, which increased the complexity and the cost of the cancer screening. Supposed panel of biomarkers could be used for Pan-cancer screening, the cost and time of the cancer screening would be greatly decreased. In this study, Pan-cancer dataset of 1274 cancer and adjacent normal tissues were

collected which were derived from 11 kinds of cancers based on HM450K. Cluster analysis in the individual samples level displayed the same type of tumor samples gathered together from each other while same type of normal tissues cluster together into the same group, indicating the genome-wide DNA methylation profile can reveal the primary similarity between the samples. Multi-label classification based on random forest prediction algorithm to 1274 samples was conducted with the top 25 and 75 important variables as the prediction features. The sensitivity, specificity and accuracy was 72%, 87% and 80% for the model with top 25 predictors while the sensitivity, specificity and accuracy of the model were 86%, 94% and 90% with top 75 predictors. Therefore, we can speculate that even for a variety of tumors, the diagnosis or the prediction can be conducted with no more than 100 genes as the predictor.

High-density DNA methylation microarray has advantage in biomarker development, including site-specific, low price, mature experiment and data analysis technique, however, it would miss some potentially important locus when it is not arranged in the probe. Methylation-specific enrichment combined deep sequencing technology can make up for this shortcoming. And it also ensure genome-wide search for the optimal recombination of the biomarkers for diagnosis and prognosis. This study established the genome-wide methylation profile of bladder cancer by MethylCap-Seq and detected the different methylation CpG island regions. And the prediction ability of the 104 most significant DMRs were validated in 509 independent samples with the strategy of bidirectional inversion biomarker discovery method. Ultimately, 5 diangosis biomarker (*VAX1*、*KCNVI*、*TAL1*、*PROX1* and *CFTR*) , 2 recurrence associated biomarker (*VAX1* and *LMX1A*) and 2 differentiation related biomarker (*ECEL1* and *TMEM26*) were obtained by the combination analysis with genome-wide methylation analysis, small sample screening accompanied with large sample validation, independent data validation and double-blind clinical simulation. The results strongly demonstrated that MethylCap-seq is a power method to identify DNA methylation based biomarker for complex disease, especially human cancer.

Keywords: DNA methylation, Cancer, Diagnosis, Prediction, Biomarker, MSD-SNuPET, Profile, Cluster Analysis

Chinese Library Classification: R39, Q31

## 缩写词汇

ACC	Accuracy
AUC	Area Under The Curve
BC	Bladder Cancer
BSP	Bisulfite Sequencing PCR
CpGI	CpG island
DMG	Differential Methylated Gene
FDR	False Discovery rate
GEO	Gene Expression Omnibus
HM450K	Illumina 450K Infinium Methylation BeadChip
MBD	Methylation Binding Domain
MDS	Multidimensional Scaling
MRA	Meta-regression Analysis
MSD-SNuPET	Methylation Status Determined Single Nucleotide Primer Extension
MSP	Methylation-Specific PCR
NSCLC	Non-small Cell Lung Cancer
OR	Odds Ratio
PCA	Principle Components Analysis
RF	Random Forests
ROC	Receiver Operating Characteristic Curve
Sen	Sensitivity
Spe	Specificity
SROC	Summary receiver operating characteristic curve
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TNM	TNM Classification of Malignant Tumours

## 第一章 人类基因组 DNA 甲基化概述

DNA 甲基化修饰是最古老的表现修饰系统之一，早在单细胞原核生物就已广泛出现。除了极少数物种的基因组中存在 DNA 甲基化修饰的退化外，大部分物种的 DNA 甲基化系统在长期的进化过程中都得到了保留。DNA 甲基化是表现遗传修饰系统的重要组成部分，具有重要的生物学功能。从分子生物学角度，DNA 甲基化全方位地参与中心定律过程的修饰，参与基因、非编码 RNA 等转录元件的调控，并且参与转录本的可变剪切。从细胞生物学角度，DNA 甲基化参与细胞的有丝分裂、减数分裂、凋亡等一系列重要过程。从发育生物学角度，DNA 甲基化参与精卵细胞的分化、胚胎发育的程式进行、参与个体成熟及正常的机体衰老，参与包括心血管疾病，肿瘤、自身免疫性疾病等几乎所有的复杂疾病。从进化角度，DNA 甲基化成为环境和基因相互作用的完美接口。外界环境通过影响 DNA 甲基化从而给基因组打上可以稳定遗传的信号，从而引起了拉马克学说的复兴。从群体遗传学角度，DNA 甲基化引起的定向碱基转换，从而形成特殊的群体的遗传结构的变化规律。从数量遗传角度，DNA 甲基化参与基因组遗传变异功能的修饰，从而成为寻找目前普遍存在的性状遗传力缺失的重要候选研究对象。DNA 甲基化的中等程度的可塑性介于遗传变异和环境变异之间。既能在收到一定持续刺激后改变，又能在相反刺激后恢复的特性，使得 DNA 甲基化可以承载极其丰富的信息储存，以对应其复杂的功能及机体高度非冗余的存在状态。DNA 甲基化的这种特征，使得其可以成为优良的生物标记物，应用于特性性状的早期评估、疾病的诊断，药物反应的预测，治疗的预后评价。在本章节将简单介绍人类 DNA 甲基化系统的定义、基本组成、DNA 甲基化酶、DNA 甲基化相关结合蛋白、人类基因组上的分布、进化。DNA 甲基化在正常人类机体的发育，特别是早期胚胎的发育中的作用与变化特征，DNA 甲基化与复杂疾病的关系，

以及 DNA 甲基化与人类学中的其他几个领域之间的研究交叉点，如 DNA 甲基化与群体遗传学，古 DNA 研究等。

## 1.1 人类基因组中的 DNA 甲基化

### 1.1.1 DNA 甲基化定义及分布特征

DNA 甲基化是指发生在 DNA 碱基上的甲基化修饰。在不同的物种中可以发生甲基化修饰的碱基不同。其基本分子机制是：在 DNA 甲基转移酶(DNA methyltransferase, DNMT)的作用下，以 S-腺苷甲硫氨酸(SAM)为唯一甲基供体，将甲基基团转移到基因组特定碱基的碳原子或者其他原子上[1]。

DNA 甲基化是一种最为古老的表观遗传修饰系统，从单细胞原核生物至高等哺乳动物都有该系统的存在。其主要由 DNMT 催化，以 S-腺苷甲硫氨酸(SAM)作为甲基供体而发生反应。催化反应有 3 类类型，包括将腺嘌呤 (A) 转变为 N-甲基腺嘌呤、将胞嘧啶 (C) 转变为 N-甲基胞嘧啶以及将胞嘧啶 (C) 转变为 C-甲基胞嘧啶。原核生物中广泛利用了上述 3 中修饰类型，但是在高等真核生物中只存在第 3 种类型的 DNA 修饰，即胞嘧啶(C) 转变为 5-甲基胞嘧啶(5mC)。至于目前所报道的人类的第七和第八种碱基(5-甲酰胞嘧啶，5-羧基胞嘧啶)也仅仅是 5-甲基胞嘧啶(5mC)的中间代谢物。

哺乳动物成熟体细胞中 DNA 甲基化基本上只存在 CpG 二核苷酸对上，且 70-80% 的 CpG 二核苷酸是呈甲基化状态的，另外 20% 非甲基化状态的 CpG 一般聚集成簇。除此之外，其他形式的二核苷酸对如 CHG, CHH (H 代表非 G 碱基) 也存在 DNA 甲基化形式，比如在胚胎干细胞中有 25% 的 DNA 甲基化发生在 CHG 或 CHH 位点，在卵细胞中也存在 mCHG，但在精子中没有发现 mCHG 的存在。由于 mCHG 不能被 DNMT1 识别以半保留复制的形式进行稳定遗传，所以一般情况下其含量很低。所以一定程度上 mCHG 反映了 DNMT3 进行从头甲基化时发生错误的概率[2]。

CpG 序列很少出现在人类基因组中，在现代人类基因组中其频率仅有 1% (正常期望值为 1/16) [1]，远低于基因组中的其预期频率。这些 CpG 位点绝大多数散落在基因组中，并且默认情况下是甲基化状态。但在基因组的某些区域中，CpG 序列密度很高，可以达均值的 5 倍以上，形成长度在 200bp-2K 的高 CpG 区域，称为 CpG 岛 (CpG island, CpGI)。同理，非甲基化的 CpG 一般也趋向于存在于 CpG 岛中，在其它地方出现时会由于 CpG 中的 C 易被甲基化而形成 5'-甲基胞嘧啶，脱氨基后形成胸腺嘧啶，由于 T 本身就会存在于 DNA 中，因此不易被

修复，因此这种类型的 C 在长期的基因组进化过程中逐渐会被淘汰。为此，人类基因组中的 CpG 和 CpG 岛存在如下规律：CpG 在基因组中是以岛的形式分布的。CpG 岛中的 CpG 一般是非甲基化的，特别是在生殖细胞中。由于非甲基化的 CpG 岛一般代表着开放的转录状态，所以 CpG 岛经常出现在真核生物的管家基因的表达调控区域，从而使得管家基因可以在大多数情况下处于激活的状态。同时 CpG 岛可能还具有更多的调控功能，比如有文献认为 CpG 岛可能是 DNA 复制起始位点的贡献因素之一[3]。

CpG 岛有很多定义形式，Gardiner-Garden 对 CpGI 做了最初的定义：长度 $>200\text{bp}$ 、高 GC 含量、O/E $>0.6$ ，之后又增加了长度和 GC 含量，以减少入围的外显子和寄生性 DNA 序列的比例。此外，很多统计学方法被用于 CpGI 的定义及预测，现在至少有数十种已发表的 CpGI 的预测软件或算法[4]，比如 HMM 模型等[5]。不论采用何种统计或数学模型，CpG 岛最核心的特性应该采用基因组的进化遗传特征作为金标准。从基因组或进化生物学角度，CpGI 应该是在生殖细胞中呈现去甲基化状态的 CpG 富集区域。因为甲基化的 C 更容易发生突变，所以生殖细胞为了减少传给后代的突变，要求这些区域去甲基化。

目前已发现 CpGI 与一系列基因组顺式元件具有显著的关联性，如基因组启动子区（Promoter）等。根据这些顺式元件与 CpGI 的关系，我可以推测定义 CpG 岛的参数可能在各个物种中有所不同。流行的 CpG 岛计算方法（GC 含量大于 50%，O/E 大于 0.6）只适用于人类，采用这个规则在小鼠的基因组中只发现了人类 50% 左右数量的 CpG 岛，而采用 CAP-seq 发现小鼠基本具有和人类等同数量的 CpG 岛[6]，显然这提示在采用计算生物学方法预测 CpG 岛时，不同物种需要设定不同的参数。CGI 中的 CpG 位点一般呈现去甲基化状态。启动子区域的 CGI 的甲基化与该基因的表达高度相关[7]。高度甲基化的 CGI 通过抑制转录因子与启动子结合或者招募转录抑制因子等机制抑制转录，典型的转录因子如 AP-2[8]。并且不同细胞类型具有自己独特的 CGI 甲基化的谱式，全基因组水平对精子和成熟卵细胞的 DNA 甲基化分析发现，精子中有 60 个其特异甲基化的 CGI，而卵细胞中有 900 个特异甲基化的 CGI[9]，这些特异甲基化的 CGI 是形成精子和卵细胞特异印记基因的重要原因之一。

哺乳动物基因组上的 DNA 甲基化谱式是由多方面因素共同作用形成的，比如 1) DNMT3A 和 DNMT3B 可以对 CpG 进行从头甲基化。2) 转录因子的结合可以阻止 DNMT 对周围 CpG 的甲基化，如 Sp1[10], RFX[11]。3) Tet 等 DNA 去甲基化酶可以实现 DNA 主动去甲基化。

### 1.1.2 DNA 甲基化的进化史

DNA 甲基化是最古老表观遗传修饰系统之一，从单细胞原核生物就已广泛出现。该系统使得细菌基因组中特定序列的腺嘌呤（A）或胞嘧啶（C）进行甲基化修饰，当序列相同但甲基化修饰不同的外来基因进入细菌体内时，会被细菌特有的限制性内切酶剪切，使得细菌具有了对外来基因片段的“免疫功能”[12]，同时研究也显示，细菌中的 DNA 甲基化还有“细胞周期调控”[13, 14]和“基因表达调控”[15]以及“DNA 复制突变修复”[16]等三大重要功能，同时 DNA 甲基化还影响细菌中的基因转移效率[17]以及细菌毒力有关[18, 19]。

DNA 甲基化在真核生物中同样具有“免疫功能”功能，但实现的方式和原核生物有所不同，真核生物基因组通过对外来寄生片段的高度甲基化使其表达抑制，从而间接达到免疫功能。此外 DNA 甲基化在真核生物中同样与 DNA 复制修复[20]，细胞表达调控等密切相关。由于真核生物基因组的复杂性，DNA 甲基化在原核生物基本功能的基础上，拥有了更加丰富的功能，如参与染色体构象的维持与改变，参与胚胎发育程式调控及组织分化。

Daniel Zilberman 的研究团队对多种物种基因组 DNA 甲基化的比较发现，有性繁殖的陆地植物及动物的基因组中存在着广泛存在的甲基化，然而无性生殖的单细胞动物和真菌只存在较少的甲基化或完全丧失甲基化修饰。尽管 DNA 甲基化从进化角度上看是一种古老的 DNA 修饰系统，但是由于发生甲基化的碱基会极大地增加基因突变（Mutation）的可能性，因此 DNA 甲基化系统的丧失在一些物种中也有发生。丧失 DNA 甲基化系统的物种主要存在于真菌进化的早期以及植物和动物进化的晚期。在进化过程中不同的物种在面对引入 DNA 甲基化带来的高突变性及可增加调控的复杂性之间进行取舍。对于人类，包括线粒体在内的整个人类基因组都有 DNA 甲基化系统的存在。显然人类基因组完美地应用了 DNA 甲基化所带来的调控复杂性并在一定程度上利用其完善的 DNA 修复机制充分地发挥了 DNA 甲基化调控机制。

在真菌基因组中 DNA 甲基化的水平波动较大，从 0.1-5% 都有[21, 22]，酿酒酵母 (*Saccharomyces cerevisiae*)，粟酒裂殖酵母 (*Schizosaccharomyces pombe*) 被认为不具备 DNA 甲基化修饰系统，相反丝状真菌 *Neurospora crassa* 却具有极其典型的 DNA 甲基化系统[22]。但有证据显示，粗糙脉孢菌(*Neurospora crassa*) 中 DNA 甲基化酶 dim-2 的突变能够导致全基因组 DNA 甲基化的缺失，但却不会影响其生长与有性繁殖[23]，说明虽然粗糙脉孢菌在进化上保留了 DNA 甲基

化系统，但好像也进化出了不依赖 DNA 甲基化系统的替代调控方式以维持其基本的新陈代谢。

植物的 DNA 甲基化与动物及微生物具有很大的差异。研究发现真菌和家蚕具有极为相似的甲基化调控系统。在植物中 DNA 甲基化不只可以发生在 CpG 位点上，在 CHG, CHH 中也广泛存在。植物中的 *DRM2*, *MET1* 和 *CMT3* 负责其 DNA 甲基化修饰。其中 *DRM2*, *MET1* 分布同动物中的 *DNMT3* 和 *DNMT1* 同源，*CMT3* 为植物特有。

尽管 DNA 甲基化是一种重要的基因表达调控方式，但在进化上也是一种风险系统。没有发生甲基化的胞嘧啶（C）容易自发脱氨而突变为尿嘧啶，基因组中的尿嘧啶（U）可以被碱基修复系统所识别，从而对这种突变进行校正。而 5-mC 自发脱氨后形成胸腺嘧啶（T），无法被碱基修复系统识别，造成突变发生，真核生物基因组中低频的 CpG（只有理论值的 20%）就是因为这个机制所造成的。由于上述原因，虽然从 *E.coli* 到 *H.sapience* 的整个进化阶段的物种都广泛存在 DNA 甲基化修饰系统，但仍有一些物种被认为丢失了 DNA 甲基化修饰系统。不具备 DNA 甲基化修饰系统的物种目前认为包括：面粉甲 (*Tribolium castaneum*)，酿酒酵母 (*Saccharomyces cerevisiae*)，粟酒裂殖酵母 (*Schizosaccharomyces pombe*)，线虫 (*Caenorhabditis elegans*) 及黄曲霉 (*Aspergillus flavus*) 等。其中黑腹果蝇 (*Drosophila melanogaster*) 在经过数次争议性的论证后，发现其在胚胎发育早期具有 DNA 甲基化现象。这个发现提示对于缺失 DNA 甲基化修饰系统的物种，还存在一个需要考虑的问题，即是否会在物种生命周期的特定阶段具有或缺失 DNA 甲基化修饰系统[24]。黑腹果蝇属于所谓的 *DNMT2-only* 物种，不具备 *DNMT1* 和 *DNMT3*，并且没有 5-甲基识别蛋白及其同系物的存在。因此一直以来被认为不具备 DNA 甲基化修饰系统[25]，但是最新的实验证据证明，在果蝇的发育过程中存在 DNA 甲基化现象，只不过其只存在于黑腹果蝇的早期胚胎发育阶段。并且根据这个现象还提示 *DNMT2* 也有可能具有 DNA 甲基化功能[26-28]。

### 1.1.3 DNA 甲基化系统主体蛋白

DNA 甲基化系统是由两类主体蛋白组成的，包括维持蛋白与识别蛋白。DNA 甲基化包括从头甲基化和维持甲基化。参与 DNA 甲基化的酶有 2 大家族，即 *DNMT1* (维持性) 和 *DNMT3* (3A、3B、3L) (从头甲基化) [29, 30]。*DNMT2* 一般被认为只参与 RNA 的甲基化而非 DNA 甲基化[31-33]。另外 *DNMT3L* 不具有催化功能，只具有对 *DNMT3A* 和 *DNMT3B* 的调控功能，与两者形成聚合物而发

挥作用。*DNMT1* 存在 *DNMT1B*、*DNMT1O* 和 *DNMT1P* 等多种剪切形式，广泛表达在包括静息组织在内的所有组织中。*DNMT1* 在染色体复制或修复过程中维持基因组原有的甲基化模式，在有丝分裂 G1 和 G2 期，分布于核基质中。在 S 期与增殖细胞核抗原（Proliferating Cell Nuclear Antigen, PCNA）形成复合物并定位于 DNA 复制叉上，进行 DNA 甲基化的半保留修饰[34, 35]，其表达与成体细胞增殖密切相关[36]。*DNMT3A* 和 *DNMT3B* 起源于脊椎动物产生时期附近的一次基因倍增事件，虽然它们具有较高的同源性，但是体内功能并不相同[37]。例如，哺乳动物 *DNMT3B* 甲基化染色体 DNA 的能力显著高于 *DNMT3A* 以及非哺乳动物的 *DNMT3B*[37]。双敲出 *DNMT3A* 的小鼠是胚胎致死的，但可以观察到小鼠全基因组及各种重复序列（如 LINE 及 SINE）的 DNA 低甲基化现象，而双敲除 *DNMT3A* 基因的小鼠出生后大约四周死亡，而且没有明显的全基因组 DNA 低甲基化。缺失 *DNMT3A* 或 *DNMT3L* 的男性或女性无法建立相应的印记基因或印记区域[38]。深入探讨 DNMT 分子功能对于全面理解人类全基因组甲基化谱式的特征具有重要意义，例如，最新证据显示 *DNMT3a* 的聚合方式决定从头甲基化产生的是连续还是离散的 mCpG [39]。*DNMT3L* 受 DNA 甲基化调控，造成 DNA 甲基化调控系统存在神奇的自调控系统。*DNMT3* 家族的三个蛋白 *DNMT3A*、*DNMT3B*、*DNMT3L* 在精子和卵细胞中都有表达[40, 41]，但是并不都发挥作用。在卵细胞中只有 *DNMT3A* 和 *DNMT3L* 参与基因印迹及重复序列的甲基化[42]，在精子中三个分子都发挥重要作用[43]。

相对于 DNA 甲基化系统的塑造者和维持者，基因组中还存在着对 DNA 甲基化信号进行识别的蛋白，已经有四种不同的甲基化结合蛋白（MBPs）已经被陆续发现，包括 MBD 蛋白家族（Methyl-CpG Binding Domain family proteins），BTB/POZ 锌指蛋白家族和 UHRF1 蛋白及 RBP-J 蛋白等[44]。

#### 1.1.4 DNA 甲基化与基因表达调控

DNA 甲基化密切参与基因表达的转录调控与基因组稳定性。DNA 甲基化改变了双螺旋中大沟的状态，造成相应蛋白结合力的改变。此外，值得注意的是，DNMT 对基因表达的抑制可以独立于基因启动子区的 DNA 甲基化，也就是 DNMT 的具有靶向功能[45]，可与组蛋白去乙酰化酶（Histone Deacetylase, HDAC）的相互作用直接抑制基因的表达[46]。一般而言，DNA 甲基化降低基因的表达水平是在哺乳动物中是依靠阻止基因转录起始完成的，在哺乳动物中 DNA 甲基化无法阻止转录延伸[47]，但在脉孢菌(neurospora)中其却具有这个功能[48]。哺乳

动物中的这个机制具有重要作用，它实现了对基因中可转录序列的抑制作用，同行还不干扰基因的转录。

在哺乳动物基因组中只有 5% 的碱基序列用于编码蛋白，其余 95% 的非编码 DNA 序列一方面为基因组的复杂性及稳定性提供了重要的支持的同时，必须受到严格的调控，从另外一种角度讲是基因组的很大负担。为了使得包括内含子（intron）、重复序列（repeat sequence）、转座子（transposon）在内的一系列非编码 DNA（non-coding DNA）得到长期的沉默，哺乳动物充分发挥了以 DNA 甲基化介导物形成的 Protein-DNA 沉默调控系统，从而促进基因组的稳定性。

尽管如此，值得注意的是，不同的组织类型对应着不同的甲基化谱式，并不等价于 DNA 甲基化控制着组织特异性表达的基因（Tissue specific genes），进而维持特异的组织类型。同样，也并不意味着不表达的基因都需要高甲基化去沉默，如 MYOD1 在肌肉中选择性表达，而在脑中不表达，但在脑中没有发现 *MYOD1* 的高甲基化状态[49, 50]。

## 1.2 DNA 甲基化与人类发育

### 1.2.1 DNA 甲基化与人类复杂疾病及肿瘤

DNA 甲基化是在真核生物中广泛存在的可遗传的表观修饰系统。不同于“只能读，不能写”的基因序列，DNA 甲基化修饰可以动态变化，从而决定基因的状态。DNA 甲基化是动物，植物，微生物等用来传递表观信息的重要方式之一。DNA 甲基化调控是介于稳定的基因组信息和动态变化的基因表达、转录因子结合及组蛋白修饰之间的一个表观调控方式。DNA 甲基化调控可以实现基因的长期沉默，这点可以通过生殖细胞特异的基因在体细胞中被长期关闭而证明。同时，由于其也具有一定的可塑性，因此不恰当的 DNA 甲基化改变，会造成各种病理状态如肿瘤的发生。

DNA 甲基化涉及一系列致病性生殖细胞突变及体细胞突变[51]，如 *DNMT3B* 突变造成的 ICF 综合征（Immunodeficiency，Centromeric Instability Facial Anomalies Syndrome）及 *MECP2* 突变造成的 Rett 综合征（Rett syndrome，RTT）。这些证据说明 DNA 甲基化不只在早期胚胎发育过程中发挥重要作用，相关甲基化酶和甲基化结合蛋白的突变，在出生后的发育中同样具有重要作用。

研究证实,由于甲基化的CpG的高频突变导致的碱基转换可以解释人类30%以上由于碱基转换而引起的遗传病。此外,DNA甲基化在基因转录调控上发挥着十分重要的作用,其功能异常与多种疾病的发生发展相关,如肿瘤,自身免疫性疾病,多种精神性疾病。在DNA甲基化的生物学功能日益被揭示清晰的同时,其调控转录的机理却并不是十分清晰。DNA甲基化可以改变染色质的结构,同时单分子力谱(Single Molecule Force Spectroscopy)实验证实DNA甲基化可以影响染色体双链的分离,从而调控基因的转录[52]。

肿瘤一直以来被认为是基因突变的累积造成的,但现在越来越清晰,表观修饰系统的紊乱也是肿瘤形成及发展的重要原因[53-55]。其中一个最重要的实验证据来自hochedlinger在2004年将黑色素瘤细胞的细胞核移植到健康的卵细胞中可以观察到黑色素瘤表观修饰的擦除并能发育出健康的小鼠[56]。细胞干性基因的甲基化可以使得细胞维持一定的分化状态,当这种甲基化状态失去时,细胞的干性复活,再加上抑癌基因的高甲基化造成的失活,最终使得正常细胞发生了癌变。全基因组水平对肿瘤细胞的DNA甲基化改变进行分析,将对人类彻底了解肿瘤产生巨大的推动作用。Andrew P. Feinberg在1988年其恰当地选择了结肠癌(colon cancer)作为实验材料,采用HPLC的方法利用同一个体的癌和癌旁,在消除了个体差异,没有采用细胞系从而避免了体外培养,完成了全基因组的甲基化定量,从而率先证明了肿瘤全基因组水平低甲基化[57]。DNA甲基化在肿瘤中还存在另外一个作用形式:长距离大区域的异常甲基化改变[58],这种现象也值得去关注其具体机制。

肿瘤全基因组DNA甲基化谱式的改变,总体上呈现:1)局部区域异常高甲基化和2)全基因组的异常低甲基化。抑癌基因启动子区域的异常高甲基化是肿瘤发生的重要推动因素之一。此外全基因组DNA低甲基化会导致基因间非编码序列的激活,基因组不稳定性增加。从而造成肿瘤的发生。因此DNA甲基化可以作为肿瘤诊断及预测的生物标记物。

从生物学角度来看,DNA甲基化调控编码RNA(基因)以及非编码RNA(miRNA等)的表达,因此DNA甲基化异常从理论上早于表达水平的变化,更早于组织形态学的变化,因此DNA甲基化适于作为肿瘤的早期诊断标记物。

世界范围内有相当多的一批实验室及医药诊断公司在尝试及DNA methylation应用到肿瘤的早期诊断中,在长达半个世纪的探索后,科研工作者发现在肿瘤和正常人群中发现了多大400-600个基因的甲基化具有一定的甲基化

差异，并发现了一系列具有很高应用前景的甲基化生物标记物，如 SHOX2[59-68]及 SEPT9 基因[60, 69]。

### 1.2.2 DNA 甲基化与胚胎发育

人类生命的起点是精卵结合形成的受精卵。由 ATCG 四碱基构成的人类基因组在人类生命信息的传递中行使着主要角色。但是“一维”的基因组信息仅仅只能是人类生命传承的“硬件”，早期胚胎发育及受精卵由单细胞发育成包括 220 种组织的成熟个体，基因表达调控、可变剪切、人类的衰老、各种发育疾病、机体对外间环境的反应等复杂“多维信息”无法用 ATCG 四碱基进行全面合理的解释，为此 Waddington 在 1942 年提出了表观遗传学。DNA 甲基化作为表观遗传学研究最重要的一部分[70]，是哺乳动物最常见的复制后调节方式之一[71]。DNA 甲基化修饰是细胞分化、个体发育[43]所必需的调控元件，具有重要的生物学效应[72]。哺乳动物体细胞基因组的甲基化在有丝分裂过程中可以通过类似 DNA 半保留复制的方式[73]，在维持性 DNA 甲基化酶的作用下进行稳定遗传。但是在配子形成及早期胚胎发育过程中却存在着复杂的动态重编程过程。了解这一复杂的过程对于人类遗传学，人类的进化研究至关重要。

个体的遗传信息在精卵结合后即已确定，在以后的个体发育过程中基本保持不变（淋巴细胞除外）。在受精过程中卵细胞提供了一半的基因组遗传信息、额外的表观遗传信息（如各种游离 RNA、miRNA）及卵细胞中的其他成分（如各种游离蛋白）。而正是这部分表观遗传信息及卵细胞中的其他成分，负责将来自母方及父方的遗传信息进行整合并程式地发育成一个完整的个体。表观遗传信息主要包括 DNA 甲基化，组蛋白修饰，组蛋白变异，microRNA 等，这些信息也和遗传物质一样具有稳定的遗传特性。同时其又具备有遗传物质不同的较大且规律的可塑性。例如表观信息被认为是维持特定且稳定的表达谱式的主要因素。一般认为胚胎干细胞中的表观信息具有较大的可塑性，而高度分化的细胞则具有较为严格的表观修饰。实验证实可以通过重塑细胞的表观遗传谱式可以改变其表达谱式进而改变其细胞类型。2011 年发现，Dnmt3A 也是造血干细胞分化的关键性基因，Dnmt3A 基因丢失，突变或失活会导致干细胞大量分化受阻，从而导致干细胞的大量存在和血细胞缺乏[74]。性别比例偏移在鱼类和爬行类中较为普遍，2012 年 Navarro 用欧洲黑鲈作为模型，揭示了温度影响芳香化酶 CYP19A 启动子区的 DNA 甲基化，芳香化酶（Aromatase）在非哺乳类脊椎动物中能够将雄激素转变为卵巢发育所必须的雌激素。因此高温能够导致雄性黑鲈个体的比例[75]。

大量的文献证实 DNA 甲基化在胚胎发育, 组织分化等各个方面都发挥了重要的作用, 但大量的科学探索后也发现了很多的甲基化不参与了事件, 比较重要的如DNA 甲基化不参与精子线粒体无法向后代传递机制[76]。虽然是阴性结果, 但对人类对人类的发育及进化也提供了重要的信息。

在配子形成和早期胚胎发育时期, DNA 甲基化水平会发生非常显著的变化[43]。小鼠胚胎发育过程中, 基因组发生了两次 DNA 甲基化重编程[77-80]。第 1 阶段发生在原始生殖细胞( PGCs) 迁移至生殖嵴, 进行分化和增殖的时期。此时 PGCs 经历了全基因组范围的去甲基化过程, 紧接着发生印记基因和单拷贝基因发生重新甲基化。这一阶段的重编程的特点是亲代印记基因擦除和重新建立。第 2 阶段发生在受精后及囊胚形成时期。在卵胞质作用下, 来自父本的染色体基因组在 DNA 复制开始前发生主动的去甲基化, 同时母本基因组随着细胞复制进行被动去甲基化。在附植后囊胚将在 DNMT3A 和 DNMT3B 的作用下进行重新甲基化。这种全基因组甲基化重编程的过程是使得受精卵摆脱精子和卵细胞的高度分化程度, 获得发育全能性, 进而发育成新个体所必需的。在牛受精卵胚胎中, 基因组首先发生主动去甲基化, 甲基化水平在 8 细胞期显著降低, 并持续到 16 细胞期, 然后胚胎发生重新甲基化[9, 77], 目前主要用于胚胎发育早期及配子形成的动物模型有海胆[81], 小鼠[82], 斑马鱼[83], 牛[84]等。

表观谱式决定着基因的表达, 进而决定一种细胞状态或者细胞类型[85, 86]。精细胞和卵细胞属于高度分化的细胞, 其各自具有自己独特的甲基化和组蛋白谱式[81, 87], 然而精卵结合后形成的受精卵却具有高度全能性, 所以必然要求这些表观谱式发生改变, 这个过程称为重编程[88]。

哺乳动物的全基因组 DNA 甲基化在胚胎发育过程中是一种动态变化过程。并且早期胚胎因物种不同其甲基化模式变化过程有很大的区别。鼠胚胎甲基化模式从精卵结合开始降低, 并持续直到囊胚时期[89]; 牛胚胎 DNA 甲基化模式从 2-细胞到 8-细胞不断降低, 到 16-细胞开始上升[84]。其他物种如猪和兔子在早期阶段并没有经历大规模的甲基化[90, 91]。

胚胎发育开始于精卵结合, 此时父系的前生殖核是由父系基因组缠绕着缺少 H3K9me2 和 H3K27me3 的组蛋白所构成[92-94]。但母系生殖核含有这两种组蛋白修饰[92, 95, 96]。精卵结合后, 精子染色体发生重组, 乙酰化的组蛋白取代精蛋白, 同时精子细胞全基因组 DNA 发生主动的去甲基化[97-99]。母体的 DNA 甲基化在受精后的几次有丝分裂过程中发生被动丢失, 这种丢失是由于维持 DNA 甲基化的酶(DNMT1)缺少所致[89, 100]。但并不是所有的区域都会遭受这一

波去甲基化过程，印记相关的 DNA 甲基化即属于这一类[101]。到目前为止，已证实有 17 个母系印记相关的 DNA 甲基化区域和 4 个父系系印记相关的 DNA 甲基区域[9]。这些区域的 DNA 甲基化需要一直保留以完成其印记功能。

人的受精作用在输卵管的上段完成，在输卵管中段时，开始胚胎发育。受精卵进行有丝分裂的同时沿输卵管向子宫方向下行，2-3 天到达子宫。那时的胚胎称为胚泡是由许多细胞构成的中空的小球体。此时的胚泡由两部分组成：将来形成胎盘的滋养外胚层（胚外细胞）以及发育成胎儿的内细胞团(Inner Cell mass, ICM)[102]。这两部分各自具有其特异的甲基化动态模式，胚胎细胞基因组的甲基化水平在的胚泡阶段达到最低。随后胚泡中的内细胞团发生重头甲基化，同时组蛋白也开始获得 H3K9me2 和 H3K27me3 修饰。但滋养外胚层仍保持去甲基化状态[103]，并且滋养外胚层与内细胞团也具有不同的 X 染色体失活及印记基因的调控机制[103]。

传统认为，在胚胎发育早期的各类组织分化应该是由不同区域的 DNA 差异甲基化组合进行，但现在有一些证据试图在证实，除了印迹相关 DNA 甲基化外没被擦除的甲基化的区域或基因，即卵细胞保留的甲基化区域，在胚胎发育组织分化中差异去甲基化可能也发挥重要作用[78]。

显然整个胚胎发育过程按照一种稳定的程式在进行[104]，在这个程式的维持中细胞与细胞之间的相互作用，表观遗传修饰的改变，卵细胞所携带的细胞质因子都发挥这重要的作用，显然这些因素之间存在密切的相互作用，共同维系着胚胎发育的正常运转，但是具体的作用关系，以及各因素贡献的比例都有待继续阐明。

### 1.2.3 DNA 甲基化在配子形成中的动态变化

哺乳动物配子的产生一般发生在性成熟之后，但是配子的祖先（PGC）的发育却是所有组织发育中最早进行[104]（小鼠 E6.25，人 E5W）。在小鼠中，胚胎发育的第 6.0-6.5 天外胚层的近端细胞将形成 PGC 及胚外组织[105]。大约在 E6.25 天时原肠胚开始形成，在 E6.0 时在胚胎外组织及原内胚叶的信号分子如 BMP4 和 BMP8b 出现，在这些分子的诱导下，E6.25 时会出现一种 Blimp1 表达阳性的一类细胞[106]，此时原肠胚开始形成，这个细胞在胚胎发育的第 7.5 天将增殖至 40 个左右 PGC，这些细胞构成的细胞群即为生殖细胞的原始祖先[107-109]。值得注意的是，生殖细胞的分化机制并不保守。这些 PGCs 细胞不断增殖，分化，并向生殖嵴（在 E10.5-E11.5 分化成生殖腺）方向迁移，因为 PGCs 来源于胚胎干细胞，因此在分化过程中需要丢弃到胚胎干细胞所具有的表观谱式，在 PGCs 的

迁徙过程中原有的 DNA 甲基化逐步减少，在 E12.5 天，PGCs 的甲基化水平降低至原有水平的 10% 左右[110]。同时，胚胎干细胞也在发育，此时的胚胎干细胞即将完成性别决定，其甲基化水平也有所降低，但仍维持在原有水平的 70% 以上。对于剩余的 10% 的甲基化的研究不是很透彻，这些位点或区域逃过了大规律的 DNA 去甲基化，仍保持高度甲基化状态，比如 IAPs[111]，其具体机理不是很清楚。性别决定发生后，同处于甲基化最低水平的 PGCs 开始按照不同的方式进行重新甲基化。雄性的 PGCs 在 E12.5 之后开始有丝分裂，并伴随着逐步的甲基化水平的升高，在出生时甲基化水平已达到其最高水平，并在以后的精子形成过程中保持此甲基化水平。精子在减数分裂后开始组氨酸到精氨酸的替换。组氨酸到精氨酸的替换使得精子获得了紧凑的基因组。卵细胞的发育中是否具有相同的替换，尚不清楚。组氨酸到精氨酸的替换在进化过程中并不保守，在很多动物中没有观察到这一现象的普遍发生。精子进行减数分裂时必须要求 SINE, LINE, 反转录转座子等的高甲基化，否则会造成精子减数分裂的终止[38]。但是全基因组水平丢失的成熟卵细胞并不会出现不育，同时似乎基因启动子区 DNA 甲基化并不影响卵细胞的基因表达。

哺乳动物卵巢内卵原细胞增殖及形成卵母细胞都是在胎儿出生前或出生后不久完成的。对于雌性原生殖细胞 PGCs 首先进行有丝分裂，在有丝分裂下不断成倍增加，并向生殖嵴迁徙[112]，一旦到达了生殖嵴，即为卵原细胞，卵原细胞通过有丝分裂的方式持续繁殖。这种繁殖的方式直到卵原细胞进入减数分裂时结束，卵原细胞经过分化产生初级卵母细胞，分化过程中没有发生全基因组甲基化水平的改变，仍保持 PGCs 的低位甲基化。女性胎儿发育到 5 个月时，卵巢中含有 200 万卵原细胞和 500 万初级卵母细胞，此后，大多数卵原细胞死亡，女性个体在排卵前，甚至出生后就已只存在初级卵母细胞。初生女婴的卵巢中有初级卵母细胞两百万，到性成熟时约有约 4 万个初级卵母细胞。它们都已进入了第一次减数分裂的前期 I 时期，始终保持在减数第一次分裂前期的状态不再生长发育，称之为静息期。直至出生后，初级卵母细胞开始逐渐“苏醒”，细胞质中开始积累卵黄、mRNA 和酶等物质而逐渐长大。初级卵母细胞基因组也同时开始 DNA 从头甲基化，造成甲基化水平的升高，性成熟后每 28 天左右，有一个或者两个初级卵母细胞受到体内激素刺激继续发育，初级精母细胞进入并完成第一次减数分裂为次级卵母细胞(secondary oocyte)和第一极体(polar body)。之后次级卵母细胞进入输卵管，在精子核进入次级卵母细胞之后进行第二次减数分裂。详细如图 1-1 所示。

不仅精子和卵细胞的发育过程不同，两者在其全基因组甲基化方面也有差异，包括总量和分布差异。如对三种高度重复片段在精子基因组的甲基化水平进行分析，发现 IAP 和 MUP 序列家族的甲基化在精卵基因组中基本相同，但对于 L1 重复片段，精子基因组甲基化显著高于卵细胞[113]。

在讨论配子的形成以及干细胞多能性的维持上，还不得不提及 piRNA，它是一类与 Piwi 蛋白相作用的 RNA，长度为 24-33 个核苷酸，组织特异性表达，piRC 是 piRNA 与 Piwi 亚家族蛋白结合形成 piRNA 复合物，其调控着生殖细胞和干细胞的生长发育，目前只在老鼠、果蝇、斑马鱼等哺乳动物的生殖细胞中发现了这类小分子。生殖细胞中存在的 piRNA 富集现象，并且 Miwi 突变导致的男性不育表明。有证据显示，在父系配子发育中，piwi 家族的两个成员 MILI 和 MIWI2 的缺失导致逆转录转座子及部分父系特异甲基化区域甲基化的丧失。已有证据开始证明 piRNA 是胚胎发育过程中从头甲基化的调控系统[114, 115]。至于为什么在母系中不存在这种现象，机制不详。

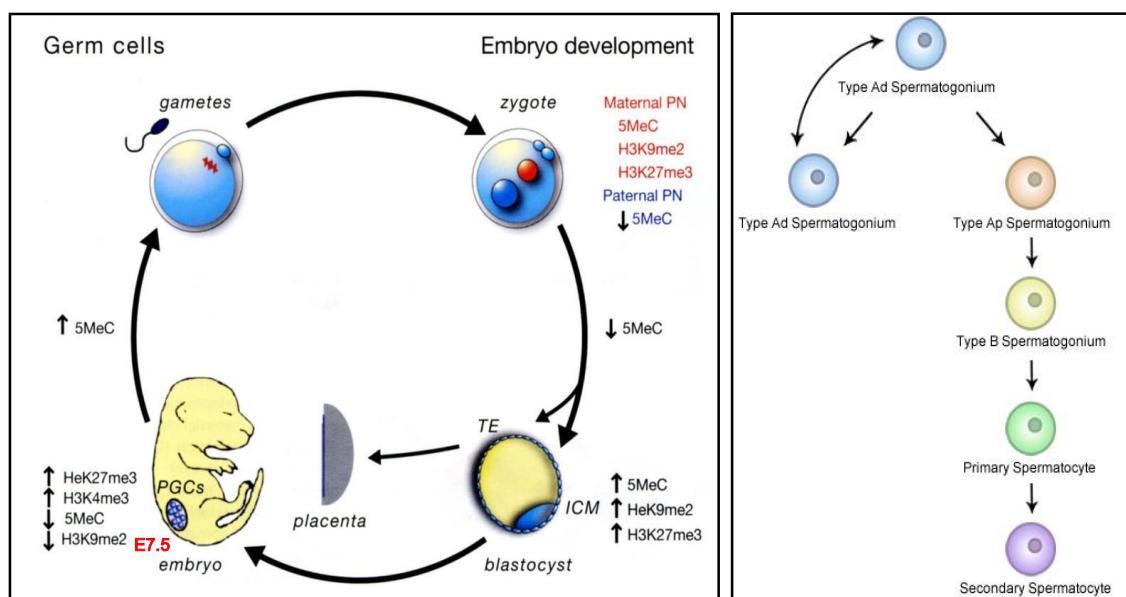


图 1-1 哺乳动物发育过程中的表观遗传重编程及精子发育过程

注：DNA 甲基化水平及各种组蛋白修饰随着配子形成及胚胎发育经历两次擦除及两次重建的动态变化。

精卵结合后，父本的单倍体细胞核由于受到精子后期染色质组氨酸到精氨酸的替换，而缺少各种组蛋白修饰，而母本的单倍体细胞核具有 H3K9me2 和 H3K27me3，相同的是两个亲本 DNA 上的甲基化开始降低，父本基因组的 DNA 甲基化在第一次有丝分裂前进行主动快速的去甲基化，而母本基因组 DNA 甲基化会缺乏 DNMT1 维持性 DNA 甲基化酶活性的情况下随有丝分裂逐步降低。之

后在囊胚期开始重建。并在配子的形成过程中再完成一次擦除与重建。TE(trophectoderm) 滋养外胚层, ICM(Inner Cell Mass)内细胞团。

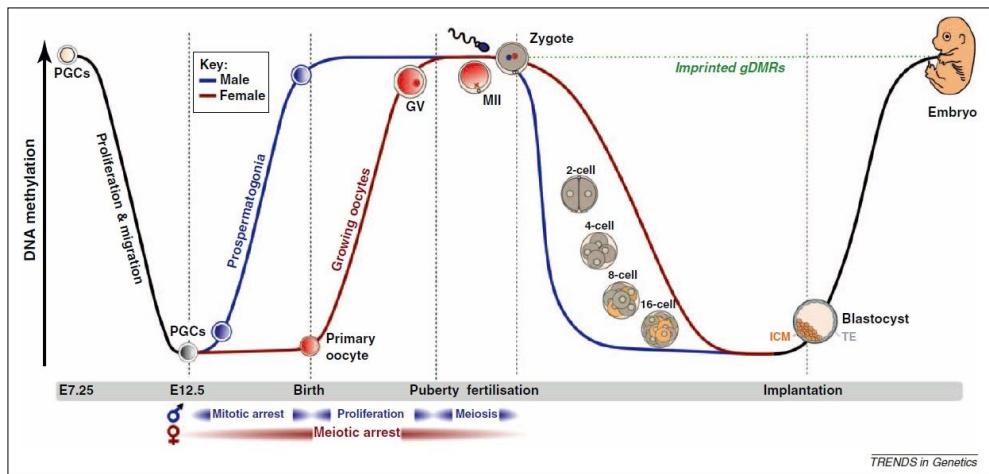


图 1-2 精子与卵细胞的发育过程中甲基化差异动态变化过程

注：本图引用自 Smallwood 等发表在 Trends in Genetics, 2012. 28(1): p. 33-42 的文章[116]。

在小鼠中，PGCs 在向生殖嵴迁移过程中不断分化，全基因组甲基化水平降低，在 E12.5 达到最低值，大约为原甲基化水平的 10% 左右。但是雄性 PGCs 的 DNA 甲基化水平迅速升高，在胎儿出生前后到达最大值并保持其甲基化水平，直到和卵细胞结合后其全基因组甲基化水平迅速再次降低。而雌性的 PGCs 在 E12.5 天时和雄性 PGCs 几乎同时到达甲基化水平最低值，之后甲基化水平基本不变，在出生后再发生和雄性 PGCs 类似的甲基化水平的升高，同时精卵结合后卵细胞为被动去甲基化，所以 DNA 甲基化水平降低的速度虽细胞传代逐渐降低。

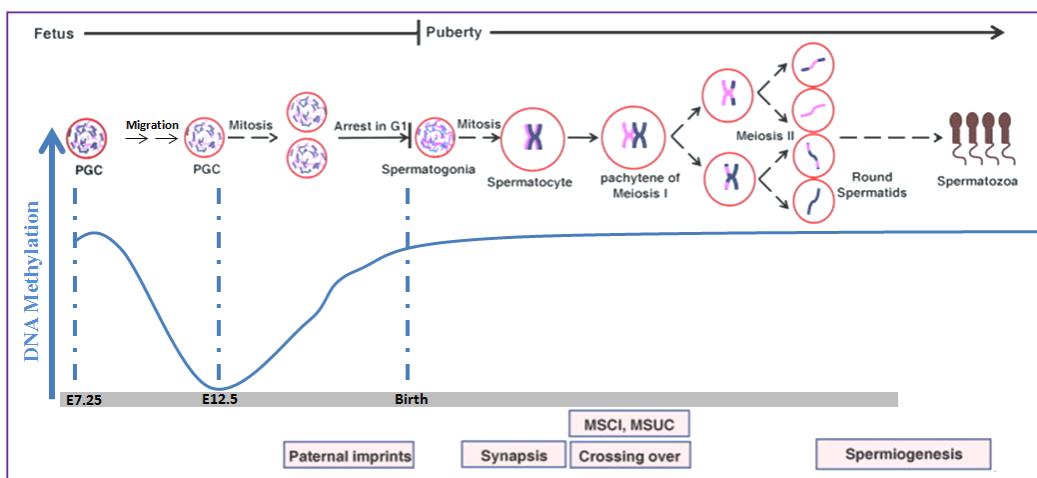


图 1-3 哺乳动物精子发育过程中的表观遗传重编程循环

注：在小鼠中，雄性 PGCs 在 E6.25 天时出现，在 E7.25 达到 40 细胞左右，并开始向生殖嵴迁移并不断分化，全基因组甲基化水平降低，在 E12.5 达到最低值，大约为原甲基化水

平的 10% 左右。之后雄性 PGCs 的 DNA 甲基化开始重建，这个阶段父系印记得以完成，在胎儿出生前后到达最大值并保持其甲基化水平[117]。

20 年前 DNA 甲基化在印记基因的形成中发挥重要作用就已经得到阐述[118]，10 年前负责从头甲基化的酶被分离[119, 120]，但是，对于卵细胞由于取材受限，直到现在还是无法透彻地阐明配子中的甲基化谱式的分子机理，但随着二代测序及单细胞基因组及表观基因组分析技术的成熟。这个领域的研究将会在未来 10 年获得巨大的突破。表观层面上对配子的形成及胚胎发育能够为，减数分裂过程中同源重组热点雌性和雄性差异等非基因序列决定的稳定打开新的视窗。配子形成是有性生殖的物种延续及进化的主要保证。深入细致地研究配子形成及胚胎发育过程中相关的表观遗传修饰机制的具体内容以及其进化机制对不育，辅助生殖及人类健康的研究具有重要的理论意义。

## 1.3 DNA 甲基化与人类群体遗传学

### 1.3.1 DNA 甲基化与人类进化

正如前文所描述的，没有甲基化的 C 容易自发脱氨而突变为尿嘧啶，但基因组中的尿嘧啶可以被碱基修复系统所识别，而 5-mC 自发脱氨后形成胸腺嘧啶，无法被碱基修复系统识别，造成突变发生，这种高频的突变造成了真核生物基因组中低频的 CpG 频率。同时为人类基因组的进化提供了动力。但是和其他类型的突变所不同的是 C 到 T 的突变为定向突变，因为对于由 DNA 甲基化造成的突变及对蛋白的影响在一定程度上可以进行预测。

2012 年 Fraser 博士采用 HM27K 芯片对源自非洲和欧洲祖先的细胞系全基因组启动子区 CpG 位点的 DNA 甲基化状态进行了检测。研究发现部分基因组中的部分位点不同人群中存在显著性差异，这部分差异性的甲基化谱式成为人群特意的甲基化谱式。在该研究中有近 1/3 的基因存在人群特异 DNA 甲基化现象。并且发现大量的甲基化位点在不同人群中具有不同的遗传模式[121]。

2013 年 Holger Heyn 博士利用 3 个不同人群：Caucasian-American, African-American 和 Han Chinese-American 对 DNA 甲基化在中性人类变异的作用进行了研究，发现 DNA 甲基化对不同种族在疾病易感性差异，药物和环境刺激的反应差异具有重要的贡献。一部分 DNA 甲基化差异可以追踪到遗传学变异从而进行解释，但是仍有近 1/3 的 DNA 甲基化差异与遗传学变异无关。暗示出了 DNA 甲基化在人类变异中占有重要的比例[121]。

2014 年 David Gokhman 及其同事通过利用甲基化及非甲基化胞嘧啶的自然降解过程，重新构建了尼安德特人与丹尼索瓦人的全部的 DNA 甲基化图谱。通过将这些图谱与现代人的甲基化图谱进行比较，2000 多个差异甲基化区域被发现，其中包括可以影响身体结构及肢体定位的 HoxD 基因簇[122]。Gokhman 观察到 HoxD 基因在现代人基因组的甲基化程度显著地比尼安德特人与丹尼索瓦人低，这意味着 DNA 甲基化的进化在一定程度上对现代人骨骼特征的进化具有重要作用。此外 Gokhman 还发现，这些差异甲基化区域的基因与很多疾病，特别是精神科及神经科疾病有关联。这说明整个现代人类的进化过程中，表观修饰的进化对于人类神经系统，脑系统可能具有重要的影响。同时也说明了遗传学层面和表观遗传学层面的协同进化共同造就了现代人类的生理及群体特征。未来的人类群体遗传学需要在表观遗传学层面展开更多的研究。

### 1.3.2 DNA 甲基化与减数分裂重组热点

在减数分裂中有一个重要的事件是同源重组。众所周知，减数分裂中的染色体重组和突变并列为产生群体水平遗传多态性的重要原因，因此，同源重组是群体遗传学的研究重点之一。同时减数分裂过程中的同源染色体重组也是染色体进行正确分离的必要条件。同源重组实现了来自于母系和父系染色体遗传物质的交换和重新组合，通过有利遗传单倍型的重组，且避免穆勒齿轮（Muller's ratchet），实现了后代对环境的适应度。因为重组会增加单倍型的增加，所以在群体水平，可以通过连锁不平衡的方法对绘制人类基因组重组区域热点。

根据目前绘制的重组区域谱氏可以发现，一方面在男性和女性中存在显著差异，同时在男性和女性的内部也存在差异，重组区域在人类基因组中不是呈现均匀分布，而是存在大量的热点和冷点区域，这种现象不仅仅在人类中得到了证实，同时在小鼠中也是如此。基因组不同区域重组频率的差异为寻找重组热点相关的基因组序列提供了研究基础。

对于减数分裂中的染色体重组热点的研究长期以来集中在寻找基因组中与重组热点相关的序列区域。近期大量的研究也确实发现了一系列与重组热点相关的序列，但是对于性别特异和近源物种特异的重组热点，无法通过简单的序列进行解释，因为性别和近源物种在基因组序列上的相似性远高于 98.6%。

传统的基于家系数据的连锁分析由于减数分裂样本的有限[123]，不能提供染色体重组事件的精细结构。对大量精子的基因分型理论上可以提供精细的染色体重组事件的精细结构，但是这只能提供男性基因组的重组热点[124]。女性染

色体的重组热点也受限于样本问题。因此，基于人群数据如 Hapmap[125, 126] 和千人基因组数据[127, 128]的重组热点成为目前减数分裂同源染色体重组热点绘制的主要方法。前期的研究显示，多种基因组序列及特征区域与重组热点相关，如 7 序列结构 5'-ATGACGT-3' 及 18 序列结构：5'-GNVTATGACGTCATNBNC-3' [129]。此外 CpG 比例, GC 含量比例, poly(A)n/poly(T)n (n>4) 结构, 距离染色粒的距离, 染色体长度等都能在一定程度上解释重组热点的变异[130]。

尽管越来越多的基因组结构和特征区域被发现与重组热点相关。但都不能对性别产生的重组热点的巨大差异进行解释。同时也不能解释人类和黑猩猩接近 98.6% 的基因组序列却对应了完全不同的重组热点[131]。此外，局部序列完全相同的个体也会不在不同的重组热点[132]，这种现象可以用重组热点不仅仅受局部区域的序列特征决定，还受远端序列的影响。但是还可以用另一种方式进行解释，即：重组热点不仅仅受遗传学序列特征的影响，还受到表观遗传修饰的调控[131-133]。这种假设在一定程度上受到一些证据的支持，比如有证据显示印记基因聚集区往往具有更高的重组率[133, 134]。目前对重组热点的遗传学已经比较透彻，已经初步确定为 PRDM9 基因调控，并有到 H3K4me3 及 H3K9ac 两种组蛋白修饰及 APO11 蛋白的参与[135, 136]。但对于男性和女性同源重组强烈的总体及分布差异的机制，以及表观遗传因素如 DNA 甲基化在其中发挥的作用是目前非常受到关注的问题[137]。

表观遗传修饰主要包括：DNA 甲基化、组蛋白变异/修饰、ncRNA 等。目前有证据显示，这三个层面的表观遗传调控都很有可能与减数分裂的重组热点相关。真核生物的染色体被有序地组装成致密的结构，这种结构具有阻碍 DNA 相关事件的功能，如抑制基因表达，抑制复制，抑制重组等事件。于此相反，当细胞需要上述事件正常进行时，则需要将染色体的构想转向允许事件发生的状态，从而开启上述一系列过程[138]。

Neumann 博士对人类基因组中两个临近重组热点 MSTM1A 和 MSTM1B 的研究发现，尽管 MSTM1A 和 MSTM1B 具有相同的局部区域序列特征(1q42.3)，但是两个位点却有着完全不同的重组特征。MSTM1B 在所有男性个体中都是活跃的重组热点，而 MSTM1A 却只在极少数个体中重组事件活跃，而在其他个体中则不是重组热点[132]。这些现象都把目标指向了是否表观遗传修饰对重组热点具有调控作用。

Sigurdsson 博士在 2009 年采用 HapMap 数据，证实了在 500kb 的尺度下，局部区域 DNA 甲基化水平和重组热点显著正相关 ( $R=0.622, P < 10^{-15}$ )。这样

首次正式了 DNA 甲基化明确参与重组热点的调控。此外，除了 DNA 甲基化对重组热点具有调控作用外，组蛋白的修饰也被报道显著也重组热点的调控有关。比如，在其他物种中已被正式，HATs 和 ADCRs 蛋白通过调控染色体的结构参与 M26 重组热点的调控[139]。在 *Saccharomyces cerevisiae* 中组蛋白甲基化酶 Set2p 和组蛋白去乙酰化酶 Rpd3p 共同参与抑制 HIS4 的重组活性[140]。H3K4Me3, H3K4Me2 和 H3K9Ac 三种组蛋白修饰酶被证实启动重组热点 PSMB9 的调控底物[141]。综上所述包括 DNA 甲基化在表观修饰可能在重组热点的调控中发挥和遗传序列同等重要的作用，更多的机制和规律需要进行详细的研究和阐述。

## 1.4 参考文献

- [1] Cooper, D.N. and M. Krawczak. *Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes* [J]. Human Genetics, 1989. **83**(2);181-188.
- [2] Sasaki, H., S. Tomizawa, H. Kobayashi, T. Watanabe, et al. *Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes* [J]. Development, 2011. **138**(5);811-820.
- [3] Delgado, S., M. Gomez, A. Bird, and F. Antequera. *Initiation of DNA replication at CpG islands in mammalian chromosomes* [J]. Embo Journal, 1998. **17**(8);2426-2435.
- [4] Zhao, Z. and L. Han. *CpG islands: algorithms and applications in methylation studies* [J]. Biochem Biophys Res Commun, 2009. **382**(4);643-645.
- [5] Wu, H., B. Caffo, H.A. Jaffee, R.A. Irizarry, et al. *Redefining CpG islands using hidden Markov models* [J]. Biostatistics, 2010. **11**(3);499-514.
- [6] Illingworth, R.S., U. Gruenewald-Schneider, S. Webb, A.R. Kerr, et al. *Orphan CpG islands identify numerous conserved promoters in the mammalian genome* [J]. Plos Genetics, 2010. **6**(9).

- [7] Lindahl, T. *DNA Methylation and Control of Gene-Expression* [J]. Nature, 1981. **290**(5805);363-364.
- [8] Comb, M. and H.M. Goodman. *CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2* [J]. Nucleic acids research, 1990. **18**(13);3975-3982.
- [9] Kelsey, G., S.A. Smallwood, S. Tomizawa, F. Krueger, et al. *Dynamic CpG island methylation landscape in oocytes and preimplantation embryos* [J]. Nature Genetics, 2011. **43**(8);811-U126.
- [10] Macleod, D., J. Charlton, J. Mullins, and A.P. Bird. *Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island* [J]. Genes & development, 1994. **8**(19);2282-2292.
- [11] Seguin-Estevez, Q., R. De Palma, M. Krawczyk, E. Leimgruber, et al. *The transcription factor RFX protects MHC class II genes against epigenetic silencing by DNA methylation* [J]. Journal of immunology, 2009. **183**(4);2545-2553.
- [12] Noyer-Weidner, M. and T.A. Trautner. *Methylation of DNA in prokaryotes* [J]. EXS, 1993. **64**;39-108.
- [13] Shapiro, L., J. Collier, and H.H. McAdams. *A DNA methylation ratchet governs progression through a bacterial cell cycle* [J]. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(43);17111-17116.
- [14] Reisenauer, A., L.S. Kahng, S. McCollum, and L. Shapiro. *Bacterial DNA methylation: a cell cycle regulator?* [J]. Journal of bacteriology, 1999. **181**(17);5135-5139.
- [15] Mahan, M.J. and D.A. Low. *DNA methylation regulates bacterial gene expression and virulence* [J]. Asm News, 2001. **67**(7);356-+.
- [16] Cooper, D.L., R.S. Lahue, and P. Modrich. *Methyl-directed mismatch repair is bidirectional* [J]. Journal of Biological Chemistry, 1993. **268**(16);11823-11829.
- [17] Wion, D., S. Allamane, P. Jourdes, D. Ratel, et al. *Bacterial DNA methylation and gene transfer efficiency* [J]. Biochem Biophys Res Commun, 2000. **276**(3);1261-1264.

- [18] Turner, K.H., I. Vallet-Gely, and S.L. Dove. *Epigenetic control of virulence gene expression in Pseudomonas aeruginosa by a LysR-type transcription regulator* [J]. Plos Genetics, 2009. **5**(12);e1000779.
- [19] Mahan, M.J., D.M. Heithoff, R.L. Sinsheimer, and D.A. Low. *An essential role for DNA adenine methylation in bacterial virulence* [J]. Science, 1999. **284**(5416);967-970.
- [20] Araujo, F.D., J.D. Knox, M. Szyf, G.B. Price, et al. *Concurrent replication and methylation at mammalian origins of replication (vol 18, pg 3475, 1998)* [J]. Molecular and cellular biology, 1999. **19**(6);4546-4546.
- [21] Antequera, F., M. Tamame, J.R. Villanueva, and T. Santos. *DNA Methylation in the Fungi* [J]. Journal of Biological Chemistry, 1984. **259**(13);8033-8036.
- [22] Binz, T., N. D'Mello, and P.A. Horgen. *A comparison of DNA methylation levels in selected isolates of higher fungi* [J]. Mycologia, 1998. **90**(5);785-790.
- [23] Selker, E.U., N.A. Tountas, S.H. Cross, B.S. Margolin, et al. *The methylated component of the Neurospora crassa genome* [J]. Nature, 2003. **422**(6934);893-897.
- [24] Liu, S.Y., J.Q. Lin, H.L. Wu, C.C. Wang, et al. *Bisulfite sequencing reveals that Aspergillus flavus holds a hollow in DNA methylation* [J]. PloS one, 2012. **7**(1);e30349.
- [25] Bird, A., P. Tate, X.S. Nan, J. Campoy, et al. *Studies of DNA methylation in animals* [J]. Journal of Cell Science, 1995;37-39.
- [26] Krauss, V. and G. Reuter. *DNA Methylation in Drosophila-A Critical Evaluation* [J]. Modifications of Nuclear DNA and Its Regulatory Proteins, 2011. **101**;177-191.
- [27] Kunert, N., J. Marhold, J. Stanke, D. Stach, et al. *A Dnmt2-like protein mediates DNA methylation in Drosophila* [J]. Development, 2003. **130**(21);5083-5090.
- [28] Lyko, F., B.H. Ramsahoye, and R. Jaenisch. *Development - DNA methylation in Drosophila melanogaster* [J]. Nature, 2000. **408**(6812);538-540.
- [29] Tajima, S., I. Suetake, F. Shinozaki, J. Miyagawa, et al. *DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction* [J]. Journal of Biological Chemistry, 2004. **279**(26);27816-27823.

- [30] Hsieh, C.L., F. Chedin, and M.R. Lieber. *The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a* [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(26);16916-16921.
- [31] Li, E., M. Okano, and S.P. Xie. *Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells* [J]. Nucleic acids research, 1998. **26**(11);2536-2540.
- [32] Tang, L.Y., M.N. Reddy, V. Rasheva, T.L. Lee, et al. *The eukaryotic DNMT2 genes encode a new class of cytosine-5 DNA methyltransferases* [J]. Journal of Biological Chemistry, 2003. **278**(36);33613-33616.
- [33] Bestor, T.H., M.G. Goll, F. Kirpekar, K.A. Maggert, et al. *Methylation of tRNA(AsP) by the DNA methyltransferase homolog Dnmt2* [J]. Science, 2006. **311**(5759);395-398.
- [34] Wigler, M.H. *The inheritance of methylation patterns in vertebrates* [J]. Cell, 1981. **24**(2);285-286.
- [35] Wigler, M., D. Levy, and M. Perucco. *The somatic replication of DNA methylation* [J]. Cell, 1981. **24**(1);33-40.
- [36] Mortusewicz, O., L. Schermelleh, J. Walter, M.C. Cardoso, et al. *Recruitment of DNA methyltransferase I to DNA repair sites* [J]. Proc Natl Acad Sci U S A, 2005. **102**(25);8905-8909.
- [37] Shen, L., G. Gao, Y. Zhang, H. Zhang, et al. *A single amino acid substitution confers enhanced methylation activity of mammalian Dnmt3b on chromatin DNA* [J]. Nucleic acids research, 2010. **38**(18);6054-6064.
- [38] Webster, K.E., M.K. O'Bryan, S. Fletcher, P.E. Crewther, et al. *Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis* [J]. Proc Natl Acad Sci U S A, 2005. **102**(11);4068-4073.
- [39] Holz-Schietinger, C., D.M. Matje, M. Flexer Harrison, and N.O. Reich. *Oligomerization of DNMT3A controls the mechanism of de novo DNA methylation* [J]. Journal of Biological Chemistry, 2011.

- [40] La Salle, S. and J.M. Trasler. *Dynamic expression of DNMT3a and DNMT3b isoforms during male germ cell development in the mouse* [J]. *Developmental Biology*, 2006. **296**(1);71-82.
- [41] Haaf, T., D. Galetzka, T. Tralau, and R. Stein. *Expression of DNMT3A transcripts and nucleolar localization of DNMT3A protein in human testicular and fibroblast cells suggest a role for de novo DNA methylation in nucleolar inactivation* [J]. *Journal of Cellular Biochemistry*, 2006. **98**(4);885-894.
- [42] Kaneda, M., R. Hirasawa, H. Chiba, M. Okano, et al. *Genetic evidence for Dnmt3a-dependent imprinting during oocyte growth obtained by conditional knockout with Zp3-Cre and complete exclusion of Dnmt3b by chimera formation* [J]. *Genes to Cells*, 2010.
- [43] Carrell, D.T. and S.S. Hammoud. *The human sperm epigenome and its potential role in embryonic development* [J]. *Molecular Human Reproduction*, 2010. **16**(1);37-47.
- [44] Bartels, S.J., C.G. Spruijt, A.B. Brinkman, P.W. Jansen, et al. *A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein* [J]. *PloS one*, 2011. **6**(10);e25884.
- [45] Rountree, M.R., K.E. Bachman, and S.B. Baylin. *DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci* [J]. *Nature Genetics*, 2000. **25**(3);269-277.
- [46] Robertson, K.D., S. Ait-Si-Ali, T. Yokochi, P.A. Wade, et al. *DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters* [J]. *Nature Genetics*, 2000. **25**(3);338-342.
- [47] Jones, P.A. *The DNA methylation paradox* [J]. *Trends in Genetics*, 1999. **15**(1);34-37.
- [48] Rountree, M.R. and E.U. Selker. *DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa* [J]. *Genes & development*, 1997. **11**(18);2383-2395.
- [49] Jones, P.A., S.M. Taylor, and V.L. Wilson. *DNA Methylation and Gene-Control* [J]. *Proceedings of the American Association for Cancer Research*, 1983. **24**(Mar);332-332.

- [50] Jones, P.A., M.J. Wolkowicz, W.M. Rideout, F.A. Gonzales, et al. *Denovo Methylation of the Myod1 Cpg Island during the Establishment of Immortal Cell-Lines* [J]. Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(16);6117-6121.
- [51] Cooper, D.N. and H. Youssoufian. *The CpG dinucleotide and human genetic disease* [J]. Human Genetics, 1988. **78**(2);151-155.
- [52] Severin, P.M., X. Zou, H.E. Gaub, and K. Schulten. *Cytosine methylation alters DNA mechanical properties* [J]. Nucleic acids research, 2011. **39**(20);8740-8751.
- [53] Jones, P.A. and P.W. Laird. *Cancer epigenetics comes of age* [J]. Nature Genetics, 1999. **21**(2);163-167.
- [54] Ting, A.H., K.M. McGarvey, and S.B. Baylin. *The cancer epigenome--components and functional correlates* [J]. Genes Dev, 2006. **20**(23);3215-3231.
- [55] Feinberg, A.P., R. Ohlsson, and S. Henikoff. *The epigenetic progenitor origin of human cancer* [J]. Nature Reviews Genetics, 2006. **7**(1);21-33.
- [56] Blelloch, R.H., K. Hochedlinger, Y. Yamada, C. Brennan, et al. *Nuclear cloning of embryonal carcinoma cells* [J]. Proc Natl Acad Sci U S A, 2004. **101**(39);13985-13990.
- [57] Feinberg, A.P., C.W. Gehrke, K.C. Kuo, and M. Ehrlich. *Reduced genomic 5-methylcytosine content in human colonic neoplasia* [J]. Cancer Research, 1988. **48**(5);1159-1161.
- [58] Clark, S.J. *Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis* [J]. Human Molecular Genetics, 2007. **16 Spec No 1**;R88-95.
- [59] Dietrich, D., C. Kneip, O. Raji, T. Liloglou, et al. *Performance evaluation of the DNA methylation biomarker SHOX2 for the aspirates* [J]. International Journal of Oncology, 2012. **40**(3);825-832.
- [60] Schneider, K.U., D. Dietrich, M. Fleischhacker, G. Leschber, et al. *Correlation of SHOX2 gene amplification and DNA methylation in lung cancer tumors* [J]. BMC Cancer, 2011. **11**;102.

- [61] Schmidt, B., V. Liebenberg, D. Dietrich, T. Schlegel, et al. *SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates* [J]. BMC Cancer, 2010. **10**;600.
- [62] Schmidt, B., V. Liebenberg, D. Dietrich, C. Kneip, et al. *Methylation of SHOX2 in Bronchial lavage - a highly specific molecular Tumor markers for Lung cancer.* [J]. Onkologie, 2010. **33**;11-11.
- [63] Field, J.K., D. Dietrich, C. Kneip, O. Raji, et al. *Shox2 DNA Methylation-a Validated Biomarker for Detecting Lung Cancer in Bronchial Aspirates* [J]. Tumor Biology, 2010. **31**;S47-S47.
- [64] Ilse, P., S. Biesterfeld, N. Pomjanski, C. Fink, et al. *SHOX2 DNA methylation is a tumour marker in pleural effusions* [J]. Cancer genomics & proteomics, 2013. **10**(5);217-223.
- [65] Darwiche, K., P. Zarogoulidis, K. Baehner, S. Welter, et al. *Assessment of SHOX2 methylation in EBUS-TBNA specimen improves accuracy in lung cancer staging* [J]. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO, 2013. **24**(11);2866-2870.
- [66] Dietrich, D., O. Hasinger, V. Liebenberg, J.K. Field, et al. *DNA methylation of the homeobox genes PITX2 and SHOX2 predicts outcome in non-small-cell lung cancer patients* [J]. Diagnostic molecular pathology : the American journal of surgical pathology, part B, 2012. **21**(2);93-104.
- [67] Dietrich, D., C. Kneip, O. Raji, T. Liloglou, et al. *Performance evaluation of the DNA methylation biomarker SHOX2 for the aid in diagnosis of lung cancer based on the analysis of bronchial aspirates* [J]. International Journal of Oncology, 2012. **40**(3);825-832.
- [68] Kneip, C., B. Schmidt, A. Seegerbarth, S. Weickmann, et al. *SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer in plasma* [J]. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer, 2011. **6**(10);1632-1638.

- [69] Kneip, C., B. Schmidt, A. Seegerbarth, S. Weickmann, et al. *SHOX2 DNA Methylation Is a Biomarker for the Diagnosis of Lung Cancer in Plasma* [J]. *J Thorac Oncol*, 2011.
- [70] Jin, B., Y. Li, and K.D. Robertson. *DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?* [J]. *Genes Cancer*, 2011. **2**(6);607-617.
- [71] Scarano, E. *DNA Methylation* [J]. *Nature*, 1973. **246**(5434);539-539.
- [72] Jones, P.A. and D. Takai. *The role of DNA methylation in mammalian epigenetics* [J]. *Science*, 2001. **293**(5532);1068-1070.
- [73] Clark, S.J., P.M. Warnecke, D. Biniszewicz, R. Jaenisch, et al. *Sequence-specific methylation of the mouse H19 gene in embryonic cells deficient in the Dnmt-1 gene* [J]. *Developmental Genetics*, 1998. **22**(2);111-121.
- [74] Challen, G.A., D. Sun, M. Jeong, M. Luo, et al. *Dnmt3a is essential for hematopoietic stem cell differentiation* [J]. *Nature Genetics*, 2011. **44**(1);23-31.
- [75] Navarro-Martin, L., J. Vinas, L. Ribas, N. Diaz, et al. *DNA methylation of the gonadal aromatase (cyp19a) promoter is involved in temperature-dependent sex ratio shifts in the European sea bass* [J]. *Plos Genetics*, 2011. **7**(12);e1002447.
- [76] Hecht, N.B., H. Liem, K.C. Kleene, R.J. Distel, et al. *Maternal Inheritance of the Mouse Mitochondrial Genome Is Not Mediated by a Loss or Gross Alteration of the Paternal Mitochondrial-DNA or by Methylation of the Oocyte Mitochondrial-DNA* [J]. *Developmental Biology*, 1984. **102**(2);452-461.
- [77] Weber, M., S. Guibert, and T. Forne. *Dynamic regulation of DNA methylation during mammalian development* [J]. *Epigenomics*, 2009. **1**(1);81-98.
- [78] Weber, M., J. Borgel, S. Guibert, Y.F. Li, et al. *Targets and dynamics of promoter DNA methylation during early mouse development* [J]. *Nature Genetics*, 2010. **42**(12);1093-U1090.
- [79] Haaf, T. *Methylation dynamics in the early mammalian embryo: Implications of genome reprogramming defects for development* [J]. *DNA Methylation: Development, Genetic Disease and Cancer*, 2006. **310**;13-22.

- [80] Brandeis, M., M. Ariel, and H. Cedar. *Dynamics of DNA Methylation during Development* [J]. Bioessays, 1993. **15**(11);709-713.
- [81] Pollock, J.M., M. Swihart, and J.H. Taylor. *Methylation of DNA in Early Development - 5-Methyl Cytosine Content of DNA in Sea-Urchin Sperm and Embryos* [J]. Nucleic acids research, 1978. **5**(12);4855-4863.
- [82] Ramsahoye, B., T. Latham, and N. Gilbert. *DNA methylation in mouse embryonic stem cells and development* [J]. Cell and Tissue Research, 2008. **331**(1);31-55.
- [83] MacKay, A.B., A.A. Mhanni, and P.H. Krone. *DNA methylation reprogramming during embryonic development in zebrafish*. [J]. Developmental Biology, 2005. **283**(2);606-606.
- [84] Fairburn, H.R., L.E. Young, and B.D. Hendrich. *Epigenetic reprogramming: how now, cloned cow?* [J]. Current Biology, 2002. **12**(2);R68-70.
- [85] Maruyama, R., S. Choudhury, A. Kowalczyk, M. Bessarabova, et al. *Epigenetic regulation of cell type-specific expression patterns in the human mammary epithelium* [J]. Plos Genetics, 2011. **7**(4);e1001369.
- [86] Koch, C.M. and W. Wagner. *Epigenetic-aging-signature to determine age in different tissues* [J]. Aging (Albany NY), 2011. **3**(10);1018-1027.
- [87] Sanford, J.P., H.J. Clark, V.M. Chapman, and J. Rossant. *Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse* [J]. Genes & development, 1987. **1**(10);1039-1046.
- [88] Reik, W., W. Dean, and J. Walter. *Epigenetic reprogramming in mammalian development* [J]. Science, 2001. **293**(5532);1089-1093.
- [89] Dean, W., F. Santos, B. Hendrich, and W. Reik. *Dynamic reprogramming of DNA methylation in the early mouse embryo* [J]. Developmental Biology, 2002. **241**(1);172-182.
- [90] Hyldig, S.M., N. Croxall, D.A. Contreras, P.D. Thomsen, et al. *Epigenetic reprogramming in the porcine germ line* [J]. Bmc Developmental Biology, 2011. **11**;11.
- [91] Sun, Q.Y., T. Chen, Y.L. Zhang, Y. Jiang, et al. *The DNA methylation events in normal and cloned rabbit embryos* [J]. Febs Letters, 2004. **578**(1-2);69-72.

- [92] Burton, A. and M.E. Torres-Padilla. *Epigenetic reprogramming and development: a unique heterochromatin organization in the preimplantation mouse embryo* [J]. Brief Funct Genomics, 2010. **9**(5-6);444-454.
- [93] Baumann, C. and R. De la Fuente. *Independent regulation of DNA methylation and histone methylation in neonatal mouse spermatogonia*. [J]. Biology of Reproduction, 2006;84-84.
- [94] Kimmins, S., M. Godmann, V. Auger, V. Ferraroni-Aguiar, et al. *Dynamic regulation of histone h3 methylation at lysine 4 in mammalian spermatogenesis* [J]. Biology of Reproduction, 2007. **77**(5);754-764.
- [95] Hemberger, M., W. Dean, and W. Reik. *Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal* [J]. Nat Rev Mol Cell Biol, 2009. **10**(8);526-537.
- [96] Wu, S.C. and Y. Zhang. *Active DNA demethylation: many roads lead to Rome* [J]. Nat Rev Mol Cell Biol, 2010. **11**(9);607-620.
- [97] Gu, T.P., F. Guo, H. Yang, H.P. Wu, et al. *The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes* [J]. Nature, 2011. **477**(7366);606-610.
- [98] Kang, Y.K., D.B. Koo, J.S. Park, Y.H. Choi, et al. *Influence of oocyte nuclei on demethylation of donor genome in cloned bovine embryos* [J]. Febs Letters, 2001. **499**(1-2);55-58.
- [99] Beaujean, N., J.E. Taylor, M. McGarry, J.O. Gardner, et al. *The effect of interspecific oocytes on demethylation of sperm DNA* [J]. Proc Natl Acad Sci U S A, 2004. **101**(20);7636-7640.
- [100] Smallwood, S.A., S. Tomizawa, F. Krueger, N. Ruf, et al. *Dynamic CpG island methylation landscape in oocytes and preimplantation embryos* [J]. Nature Genetics, 2011. **43**(8);811-814.
- [101] Monk, M. *Changes in DNA Methylation during Mouse Embryonic-Development in Relation to X-Chromosome Activity and Imprinting* [J]. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 1990. **326**(1235);299-312.

- [102] Ganienko, E.F. [Development of germ cells in the ovaries of human fetuses in the early periods] [J]. Arkh Anat Gistol Embriol, 1975. **68**(6);85-91.
- [103] Han, Y.M., Y.K. Kang, H.J. Lee, J.J. Shim, et al. Varied patterns of DNA methylation change between different satellite regions in bovine preimplantation development [J]. Molecular Reproduction and Development, 2005. **71**(1);29-35.
- [104] Surani, M.A., K. Ancelin, P. Hajkova, U.C. Lange, et al. Mechanism of mouse germ cell specification: a genetic program regulating epigenetic reprogramming [J]. Cold Spring Harb Symp Quant Biol, 2004. **69**;1-9.
- [105] Braude, P., H. Pelham, G. Flach, and R. Lobatto. Post-transcriptional control in the early mouse embryo [J]. Nature, 1979. **282**(5734);102-105.
- [106] Ohinata, Y., B. Payer, D. O'Carroll, K. Ancelin, et al. *Blimp1* is a critical determinant of the germ cell lineage in mice [J]. Nature, 2005. **436**(7048);207-213.
- [107] Lawson, K.A. and W.J. Hage. Clonal analysis of the origin of primordial germ cells in the mouse [J]. Ciba Found Symp, 1994. **182**;68-84; discussion 84-91.
- [108] Spinaci, M., P. Fantinati, S. Nicoletti, C. Cappannari, et al. Paternal chromatin remodelling in mouse oocytes following fertilization [J]. Vet Res Commun, 2003. **27 Suppl 1**;241-243.
- [109] McLay, D.W. and H.J. Clarke. Remodelling the paternal chromatin at fertilization in mammals [J]. Reproduction, 2003. **125**(5);625-633.
- [110] Jacobsen, S.E., C. Popp, W. Dean, S.H. Feng, et al. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency [J]. Nature, 2010. **463**(7284);U1101-U1126.
- [111] Hajkova, P., S. Erhardt, N. Lane, T. Haaf, et al. Epigenetic reprogramming in mouse primordial germ cells [J]. Mech Dev, 2002. **117**(1-2);15-23.
- [112] Anderson, R., T.K. Copeland, H. Scholer, J. Heasman, et al. The onset of germ cell migration in the mouse embryo [J]. Mech Dev, 2000. **91**(1-2);61-68.
- [113] Howlett, S.K. and W. Reik. Methylation levels of maternal and paternal genomes during preimplantation development [J]. Development, 1991. **113**(1);119-127.

- [114] Slotkin, R.K., M. Vaughn, F. Borges, M. Tanurdzic, et al. *Epigenetic reprogramming and small RNA silencing of transposable elements in pollen* [J]. Cell, 2009. **136**(3);461-472.
- [115] Aravin, A.A., R. Sachidanandam, D. Bourc'his, C. Schaefer, et al. *A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice* [J]. Molecular cell, 2008. **31**(6);785-799.
- [116] Smallwood, S.A. and G. Kelsey. *De novo DNA methylation: a germ cell perspective* [J]. Trends in Genetics, 2012. **28**(1);33-42.
- [117] Kota, S.K. and R. Feil. *Epigenetic transitions in germ cell development and meiosis* [J]. Developmental Cell, 2010. **19**(5);675-686.
- [118] Stoger, R., P. Kubicka, C.G. Liu, T. Kafri, et al. *Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal* [J]. Cell, 1993. **73**(1);61-71.
- [119] Li, E., M. Okano, D.W. Bell, and D.A. Haber. *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development* [J]. Cell, 1999. **99**(3);247-257.
- [120] Okano, M., D.W. Bell, D.A. Haber, and E. Li. *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development* [J]. Cell, 1999. **99**(3);247-257.
- [121] Fraser, H.B., L.L. Lam, S.M. Neumann, and M.S. Kobor. *Population-specificity of human DNA methylation* [J]. Genome biology, 2012. **13**(2);R8.
- [122] Gokhman, D., E. Lavi, K. Prufer, M.F. Fraga, et al. *Reconstructing the DNA methylation maps of the Neandertal and the Denisovan* [J]. Science, 2014. **344**(6183);523-527.
- [123] Kong, A., D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, et al. *A high-resolution recombination map of the human genome* [J]. Nature Genetics, 2002. **31**(3);241-247.
- [124] Arnheim, N., P. Calabrese, and I. Tiemann-Boege. *Mammalian meiotic recombination hot spots* [J]. Annual Review of Genetics, 2007. **41**;369-399.

- [125] Altshuler, D., L.D. Brooks, A. Chakravarti, F.S. Collins, et al. *A haplotype map of the human genome* [J]. Nature, 2005. **437**(7063);1299-1320.
- [126] Olivier, M. *A haplotype map of the human genome* [J]. Physiological Genomics, 2003. **13**(1);3-9.
- [127] [Anon]. *1000 Genomes Project announced* [J]. Journal of Investigative Medicine, 2008. **56**(4);682-683.
- [128] Stephenson, J. *1000 genomes project* [J]. Jama-Journal of the American Medical Association, 2008. **299**(7);755-755.
- [129] Steiner, W.W. and G.R. Smith. *Optimizing the nucleotide sequence of a meiotic recombination hotspot in Schizosaccharomyces pombe* [J]. Genetics, 2005. **169**(4);1973-1983.
- [130] Jensen-Seaman, M.I., T.S. Furey, B.A. Payseur, Y. Lu, et al. *Comparative recombination rates in the rat, mouse, and human genomes* [J]. Genome Res, 2004. **14**(4);528-538.
- [131] Winckler, W., S.R. Myers, D.J. Richter, R.C. Onofrio, et al. *Comparison of fine-scale recombination rates in humans and chimpanzees* [J]. Science, 2005. **308**(5718);107-111.
- [132] Neumann, R. and A.J. Jeffreys. *Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation* [J]. Human Molecular Genetics, 2006. **15**(9);1401-1411.
- [133] Sandovici, I., S. Kassovska-Bratinova, J.E. Vaughan, R. Stewart, et al. *Human imprinted chromosomal regions are historical hot-spots of recombination* [J]. Plos Genetics, 2006. **2**(7);944-954.
- [134] Lercher, M.J. and L.D. Hurst. *Imprinted chromosomal regions of the human genome have unusually high recombination rates* [J]. Genetics, 2003. **165**(3);1629-1632.
- [135] Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, et al. *PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice* [J]. Science, 2010. **327**(5967);836-840.

- [136] Parvanov, E.D., P.M. Petkov, and K. Paigen. *Prdm9 controls activation of mammalian recombination hotspots* [J]. Science, 2010. **327**(5967);835.
- [137] Bjornsson, H.T., M.I. Sigurdsson, A.V. Smith, and J.J. Jonsson. *HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination* [J]. Genome Research, 2009. **19**(4);581-589.
- [138] Lilley, D.M.J. and J.F. Pardon. *Structure and Function of Chromatin* [J]. Annual Review of Genetics, 1979. **13**;197-233.
- [139] Yamada, T., K. Mizuno, K. Hirota, N. Kon, et al. *Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot* [J]. Embo Journal, 2004. **23**(8);1792-1803.
- [140] Merkera, J.D., M. Dominska, P.W. Greenwell, E. Rinella, et al. *The histone methylase Set2p and the histone deacetylase Rpd3p repress meiotic recombination at the HIS4 meiotic recombination hotspot in Saccharomyces cerevisiae* [J]. DNA Repair, 2008. **7**(8);1298-1308.
- [141] Buard, J., P. Barthes, C. Grey, and B. de Massy. *Distinct histone modifications define initiation and repair of meiotic recombination in the mouse* [J]. Embo Journal, 2009. **28**(17);2616-2624.

## 第二章 DNA 甲基化标记物在肺癌诊断中的机遇和挑战

### 2.1 背景

肺癌是一种涉及遗传学和表观遗传学异常的复杂疾病。据美国癌症协会最新的统计结果显示，肺癌无论在男性和女性中肺癌都是癌症死亡的首要致死癌症，并且新发病例在所有肿瘤中排在第二位[1]。在肺癌中 85% 的病例为非小细胞肺癌（Non-small Cell Lung Cancer, NSCLC），主要包括腺癌（Adenocarcinoma, Ad）和鳞状细胞癌（Squamous Cell Carcinoma, Sc）。在美国，每年有超过 224,210 和 159,480 名美国人新发和死于肺癌[1]。肺癌的 5 年生存率显著依赖于肺癌初诊时的肿瘤分期。根据 SEER 的癌症统计 1975 至 2002 年的回顾分析，对于未转移（local），局部转移（regional）和远端转移（distance）肿瘤患者，其五年生存率分别为 49%，16%，2%。显然，提高肿瘤患者生存时间最有效的途径即实现肺癌的早期诊断或早期筛查[2]。DNA 甲基化是抑癌基因失活一个重要机制，因此 DNA 甲基化被认为是肿瘤诊断和筛查的有效生物标志物。此外与其他基因组或表观基因组变异相比，拥有无可比拟的优势，包括化学性质稳定、可通过远端体液对肿瘤进行监控、拥有稳定的定性及定量检测方法、检测成本低等优势[3]。目前学术界已经获得了效果比较优良的基于 DNA 甲基化对肿瘤进行早期筛查的方法[4, 5]。

本章研究以 *APC* (Adenomatous polyposis coli) 基因在肺癌诊断的研究为例子，剖析 DNA 甲基化肿瘤诊断中的潜力和挑战。*APC* 基因编码一种肿瘤抑制蛋白(Wnt 信号传导途径的拮抗剂)，参与了细胞迁移和粘附，转录激活和凋亡等过程[6]。同时，*APC* 基因参与家族性腺瘤性息肉病（一种常染色体显性疾病）发展为恶性肿瘤的过程。上述证据提示 *APC* 可能是肿瘤进展的相关基因。同时已有研究人员报道，*APC* 基因启动子甲基化能够抑制其基因表达，并且该抑制过程是通过染色质构象的改变和 CCAAT-盒结构结合转录因子与 CCAAT-盒的异常的结合所导致的[7]。*APC* 基因启动子区 DNA 甲基化与非小细胞肺癌之间的关系在过去的 20 年间已经被不断地报道，大量的研究都认为 *APC* 启动子甲基化可以作为一个有效的生物标志物用于非小细胞肺癌诊断[8, 9]。然而，事实情况是在不同的研究中，*APC* 在肿瘤中的甲基化率存在较大差异。这可能是由样本中的性别比例、年龄分布、样本遗传结构差异，DNA 甲基化检测方法等其他一系列差异

造成的。迄今为止没有尚没有 *APC* 基因甲基化对于非小细胞肺癌的诊断能力的定量评估。

在本研究中, 我们进行了一项 Meta 分析以定量评估 *APC* 甲基化对非小细胞肺癌诊断的敏感性和特异性。同时评估影响 DNA 甲基化对肿瘤进行诊断的多种异质性来源, 发掘 DNA 甲基化对肿瘤的临床推广中的各种可能存在的可能问题。基于如上目的, 基于固定效应及随机效应模型被用来估计 *APC* 基因 DNA 甲基化在肺癌中的风险比值比。Meta 回归被应用来分析和识别导致不同研究中敏感性和特异性差异的异质性来源。此外, 癌症基因组图谱计划 (TCGA) 收集了数百个非小细胞肺癌样本的各类临床和人口统计信息, 以及全基因组 DNA 甲基化微阵列数据, 此外可以用来作为没有发表偏倚的独立数据, 对 Meta 分析的结果进行再验证。在本章研究中, 我们整合采用 Meta 分析和独立无偏样本的再验证的思路, 评估 *APC* 基因启动子区 DNA 甲基化对非小细胞肺癌的诊断能力, 以期待对 *APC* 基因启动子区 DNA 甲基化与非小细胞肺癌之间的关系做出系统的结论。

## 2.2 材料和方法

### 2.2.1 检索策略和数据提取

本章研究采用对一系列文献资源数据库, 包括 Pubmed 数据库、Cochrane 图书馆、Ovid Medline 和 TMC ProSearch 数据库的检索, 以获得所有已发表过与 *APC* 基因 DNA 甲基化与非小细胞肺癌相关的文献资源。

*APC* 基因在不同的研究中有多种不同的 Symbol 比如: *BTPS2*, *DP2* 等, 为此, 本研究采用 (*APC* 或 *BTPS2* 或 *DP2* 或 *DP2.5* 或 *DP3* 或 *PPP1R46I*) AND ((lung 或 NSCLC) 和 (Tumor 或 Cancer)) 的关键词组合作为检索策略。此外, 对于每个不同的数据库, 根据其相应的数据库检索规则, 采用相应的星号 (Star, \*), 美元符 (Dollar, \$) 以及其他通配符以进行有效的文献收集。采用上述方法检索到的文献, 进一步根据标题和摘要进行筛选, 以剔除与研究目的不相符的文献来源。具体的排除及纳入原则如下:

- 1) 以下类型的研究被排除: 动物实验、病例报告、综述、Meta 分析、非病例对照研究 (只含有 case 的病例研究)、数据不足研究以及与作者联系仍无法获得全文的相关研究。
- 2) 以下类型的研究被纳入: 1) 患者被诊断患有非小细胞肺癌 (Ad 和 Sc); 2) 病例对照研究; 3) 研究必须从组织, 血液或血清 *APC* 基因启动子甲基化的数据,

3) APC 启动子区 DNA 甲基化的检测对象包括实体组织，血液，血浆或血清游离 DNA。

按照上述标准对每篇文献逐个匹配，然后对进入研究范畴的文献进行信息提取。主要提取的信息如下：第一作者、出版年份、样本量大小、肺癌诊断时年龄（平均值或中位数）、样本中性别比例（男/女，M2F）、恶性肿瘤临床 TNM（TNM Classification of Malignant Tumours）早期样本的比例（代表样本中早期非小细胞肺癌样本的比例）、出版目的名称（用于诊断或非用于诊断）、分析多个基因或单基因（在研究设计所同时检测一种或多种基因）、正常对照的类型（自体或异源的对照）。上述所有过程都由两个独立的评审进行，出现任何矛盾时，寻找另外的独立同事进行仲裁。

### 2.2.2 Meta 合并比值比分析

数据的分析和可视化采用 R 软件完成（R 版本：2.15.3）。所使用的 R 软件包主要包括：meta, metafor, mada 等。合并比值比（OR）及相应的 95% 置信区间（95% CI）用来反映关联的强度和可信范围。大多数数据从原始文献中直接进行提取，但部分指标进行了重计算以使得所有研究具有统一的变量，从而使 Meta 研究具有更缜密的分析和可信的结论。 $I^2$  统计量用来评估不同研究之间的异质性强度[10]， $I^2$  值超过 50%，卡方检验  $P \leq 0.1$  被认为所考察的研究对象之间具有显著性异质性。 $T^2$  用来确定异质性可由相应变量进行解释的程度。根据异质性统计量  $I^2$  使用 DerSimonian-Laird 随机效应模型 ( $I^2 > 50\%$ ,  $P \leq 0.1$ ) 或固定效应模型 ( $I^2 < 50\%$ ) 对数据进行合并[11]。默认的假设检验均采用双侧检验， $P \leq 0.05$  被认为具有显著性。随机效应 meta 回归用以确定解释变量对异质性解释的程度及显著性[12]。

### 2.2.3 Meta-regression 分析

Meta 回归分析（Meta-regression analysis , MRA）是基于文献对混淆因素进行定量分析的经典方法。Meta 回归分析已经在社会学，行为学和经济科学中得到普遍应用，比较重要的应用如：政策相关的参数的估计、检验经济理论、解释异质性、潜在偏差的定量评估等。从模型假设上进行分类，Meta 分析一般可以分为：简单 Meta 回归、固定效应 Meta 回归和随机效应 meta 回归。

由于本文中的研究对象间的异质性显著存在，因此我们只选择随机效应的 Meta 回归对潜在变量的异质性来源进行探讨。随机效应 Meta 回归假设 effect size 在不同研究中具有不同的估计值  $\theta$ 。

$$y_i = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \cdots + \beta_n x_{nj} + \emptyset + \epsilon_j$$

在本公式中  $\sigma_{\epsilon_j}^2$  是在第  $j$  个研究中 effect size 的方差。Effect size 在不同研究之间的方差为  $\sigma_{\emptyset}^2$ 。 $x_1 \dots x_n$  表示各个带来异质性的解释变量。

本研究中的 Meta 回归分析中包括了九个潜在的解释变量，包括对照样本类型（自体和异构），性别比例，TNM 分期早期样本的比例，平均或中位年龄（> 65 或 <= 65）、单个文献/研究中具有单个或多个目标靶点、样品类型（血清或组织样本）、甲基化检测方法（MSP 或 QMSP）、研究设计（诊断为目的或非诊断为目的）以及采用的引物。

#### 2.2.4 敏感度分析

敏感度分析来评估单个的研究对最终结果的影响程度，这可以通过对每次缺失一个个体研究观察对研究结果的稳定性或波动性进行评估。一般将缺失每一个个体研究的结果进行汇总，就可以很直观地观察到是否存在某个研究对整体实验的结果具有较大的影响。敏感度分析可以针对不同的 Meta 合并模型进行，但一般只采用假设最合理的模型，本文中只在对随机效应模型下的 meta 合并分析进行敏感性分析。

#### 2.2.5 发表偏倚分析

发表偏倚采用漏斗图与基于混合效应假设的 Egger 法进行检验。漏斗图是 effect size (X 轴) 和样本大小 (Y 轴) 的散点图。一般而言对称的倒立漏斗形状表明该研究数据集是一个比较正常的数据集。反之，如果非对称的形状出现，则表明数据存在一些问题，很有可能存在发表偏倚。如果有证据显示存在发表偏倚，则采用传统的 Meta-修剪的方法 (trim and fill) 重新估计的关联性大小。Meta-修剪方法的基本思路是通过删除引起非对称漏斗图的小样本研究对象后重新估计漏斗图的中心，然后按照对称原则寻找和之前删除研究对象相对应的可能被遗失的研究 (missing study)。然后根据修建后的数据集再次对新数据进行相应的 Meta 后续分析。

#### 2.2.6 综合受试者工作特征曲线

甲基化相关的预测模型研究与传统的单核苷酸多态性风险评估模型具有一定的差别，因为 SNP 属于离散数据类型，而 DNA 甲基化数据是连续数据类型。基于 DNA 甲基化的预测模型会涉及不同的甲基化阈值而使得预测模型的敏感性，特异性不同。在这种情况下，传统的加权平均值（平均的敏感度和特异度）不能如实地反映在预测模型在真实的临床使用过程中的准确度，因为不同的阈值标准

的可能导致灵敏性和特异性等指标出现极端扭曲的分布，被称为阈值效应[13]。为此，综合受试者工作特征曲线（Summary receiver operating characteristic curve, SROC）可用于对基于连续性临床指标的 Meta 分析数据建立的诊断模型预测能力的评估[13, 14]。

在 meta 分析中，每个研究都提供一组敏感性和特异性数值，因此可以得到每个研究的真阳性率 (TPR) 和假阳性率 (FPR) 的值。SROC 的原理即采用 Moses 和他的同事在 1993 年提出的回归模型，对所有 TPR 和 FPR 进行拟合。在 Moses 提出的回归模型中，自变量和因变量分别为 D 和 S，其表达式分别为：

$$D = \ln \frac{TPR}{1 - TPR} - \ln \frac{FPR}{1 - FPR}$$

$$S = \ln \frac{TPR}{1 - TPR} + \ln \frac{FPR}{1 - FPR}$$

在上述表达式中，D 又可以解释为诊断的  $\ln OR$ ，可以衡量诊断模型的灵敏度。S 代表诊断阈值，对应着研究所确定为病例时所采取的诊断临界值。可以发现当 S=0 时， $TPR = 1 - FPR$ ，此时的 SROC 曲线恰好是 SROC 坐标的对角线。Moses 的回归模型如下：

$$D = \alpha + \beta \times S$$

采用最大似然法可以对  $\alpha$  和  $\beta$  进行估计。

$$TPR = \frac{\exp^{\frac{a}{1-b}} \times \left(\frac{FPR}{1-FPR}\right)^{\frac{1+b}{1-b}}}{1 + \exp^{\frac{a}{1-b}} \times \left(\frac{FPR}{1-FPR}\right)^{\frac{1+b}{1-b}}}$$

$$AUC = \int_0^1 \frac{\exp^{\frac{a}{1-b}} \times \left(\frac{FPR}{1-FPR}\right)^{\frac{1+b}{1-b}}}{1 + \exp^{\frac{a}{1-b}} \times \left(\frac{FPR}{1-FPR}\right)^{\frac{1+b}{1-b}}} dx$$

被选定线性回归模型拟合 SROC 曲线上敏感性和 (1-特异性) 转化为复杂的对数变量。拟合曲线下区域面积 (AUC) 即可用来评估测试的可靠性[13]。在本研究中 SROC 下 AUC 可以衡量 APC 的甲基化对非小细胞肺癌的诊断能力的表现。并且 SROC 可以实现 APC 对多种肿瘤诊断模型可靠性的定量比较。更多关于 SROC 的资料可以参考 Jones 和 Walter 的相关研究[15, 16]。

### 2.2.7 TCGA 数据提取和再验证

从 TCGA 计划的数据库中收集包括来自 HM27K 和 HM450K 的非小细胞肺癌数据集 (<http://cancergenome.nih.gov/>)，其中共包括两套样品：1) 535 肺腺癌和 50 肺腺癌对应的正常肺组织；2) 385 肺鳞癌和 67 肺鳞癌对应的正常肺组织。HM27K 和 HM450K 的数据集共享的 25978 CpG 位点在所有样本的甲基化信号被抽提，合并，标准化处理后，用于抽提其中 APC 基因的甲基化信号。离散化处理甲基化信号时，根据传统经验，采用 beta 大于 0.3 的甲基化信号被定义为甲基化状态，小于 0.3 定义为去甲基化状态[17]。最终位于 APC 基因的启动子区的六个 CpG 位点 (cg01240931, cg15020645, cg16970232, cg20311501, cg21634602 和 cg24332422) 被作为研究对象。因为涉及到多个 CpG 位点的统计检验，多重校正采用 Benjamini 和 Hochberg 提出的 FDR ( $\alpha=0.05$ ) 进行。

## 2.3 结果

### 2.3.1 候选文献的基本特征

按照方法学部分介绍的电子检索策略，共计 506 潜在相关的文章被提取，其中来自各数据库的文献数量为：PubMed, 315 篇；SCOPUS, 112 篇；Cochrane 图书馆, 3 篇；OVID MEDLINE, 53 篇；TMC ProSearch, 23 篇。然后通过对标题、摘要、全文进一步筛选，初步确定基本的纳入文献。此外，为了实现最低的文献遗漏率，我们对符合条件的个体研究的参考文献列表也进行相应的检查。最终 17 篇与 APC 基因启动子甲基化和非小细胞肺癌之间的关系的研究被收集起来。这些文献的基本信息总结如表 2-1 和表 2-12 所示 [8, 18-33]。所有 17 篇文献都为英文文献，共计 1338 例肺癌组织/血清和 913 对应的正常组织/血清。所涉及非小细胞肺癌患者诊断年龄介于 25 到 86 岁之间。文献涉及样本的平均或中位数年龄在 53 至 67 岁之间。对于文献的研究目的，13 篇为诊断设计，另外 4 篇涉及预后，生存的研究等。17 项研究中，肿瘤样本的 TNM 分期，中早期样本的比例介于 32.1 和 100% 之间。男性个体的百分比介于 53% 和 81% 之间。对于实验方法，17 篇文章中 7 篇采用了甲基化特异性聚合酶链反应 (MSP)，而其他的用于定量 MSP (qMSP，如 Methylight, Prosequencing 等)。17 篇文章主要采用了两套甲基化检测的引物或探针。其中 7 篇文献采用了第一套引物，其在人类基因组中的位置是 Chr5:112073421-112073518；7 项研究采用了第二套引物，其在人类基因组中的位置是：Chr5:112101379-112101452；其余研究的引物无法定位。甲基化芯片的 6 个 CpG 位点没有被发现在位于上述引物序列中，但 cg20311501 位于

第二组引物的复制区域。为此可以进行适当的比较并对一些结果进行适当推论。具体如表 2-1、表 2-2 和表 2-12 所示。

表 2-1 候选文献的基本特征

Author (Published Year)	Age <sup>a</sup> (years)	Stages I %	Gender (M/F)	Patients (M+/M-)	Control (M+/M-)	ref.
Zhang et al (2011, China) <sup>b</sup>	59	32.05	29/39	44/34	10/68	[33]
Wang et al (2008, China)	-	-	17/28	19/9	1/11	[31]
Jin et al (2009, Japan)	66.7	-	17/24	27/45	22/41	[21]
Feng et al (2008, USA)	64.3	42.86	26/49	26/23	21/28	[20]
Brabender et al (2001, USA)	63.3	49.45	69/91	86/5	80/11	[19]
Virmani et al (2001, USA)	-	-	-	22/26	0/18	[30]
Yanagawa et al (2003, Japan)	67.3	66.67	18/25	28/47	36/39	[32]
Topaloglu et al (2004, USA)	-	54.84	-	17/14	5/17	[28]
Kim et al (2007, Korea)	63	56.57	64/79	48/41	33/66	[22]
Vallbohmer et al (2006, USA)	63	49.45	69/91	86/5	80/3	[29]
Lin et al (2009, China)	61.1	100.00	20/31	49/75	2/24	[23]
Shivapurkar et al (2007, USA)	-	-	-	35/5	23/17	[26]
Suzuki et al (2006, Japan)	64	34.00	33/49	53/97	3/57	[27]
Zhang et al (2011, China) <sup>b</sup>	-	-	-	54/56	5/45	[33]
Pan et al (2009, China)	53	-	17/26	40/38	0/31	[24]
Begum et al (2011, USA)	65	-	10/19	12/64	3/27	[18]
Rykova et al (2004, Russia)	NA	-	-	3/6	0/16	[25]
Usadel et al (2002, USA)	64.2	-	-	42/47	0/50	[8]

注：更加详细的内容见本章 a 年龄根据每篇文章的不同采用相应的平均或者中位数年龄。b 来自同一篇文章，同时对实体瘤和血浆进行了检测。M+, M- 分别代表甲基化阳性和阴性。

表 2-2 17 篇文章主要涉及的引物序列

Study	Forward	Reverse	Location
Pan et al (2009, China)	ACTGCCATCAACTTCCCTTG*	GACATGTGGCTGTATTGGT*	chr5:112073311-112073571
Zhang et al (2011, China) <sup>b</sup>	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Virmani et al (2001, USA)	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Kim et al (2007, Korea)	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Lin et al (2009, China)	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Shivapurkar et al (2007, USA)	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Suzuki et al (2006, Japan)	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Zhang et al (2011, China) <sup>b</sup>	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Rykova et al (2004, Russia)	CACTGCGGAGTGCGGGTC	CCGTCGGGAGCCCCGCCGA	chr5:112073421-112073518
Begum et al (2011, USA)	GGACCAGGGCGCTCCC*	GTGTGGGCGCACGTGAC*	chr5:112101379-112101452
Usadel et al (2002, USA)	GGACCAGGGCGCTCCC*	GTGTGGGCGCACGTGAC*	chr5:112101379-112101452
Jin et al (2009, Japan)	GGACCAGGGCGCTCCC*	GTGTGGGCGCACGTGAC*	chr5:112101379-112101452
Feng et al (2008, USA)	GGACCAGGGCGCTCCC*	GTGTGGGCGCACGTGAC*	chr5:112101379-112101452

Brabender et al (2001, USA)	GGACCAGGGCGCTCCC*	GTGTGGGCGCACGTGAC*	chr5:112101379-112101452
Vallbohmer et al (2006, USA)	GGACCAGGGCGCTCCC*	GTGTGGGCGCACGTGAC*	chr5:112101379-112101452

注：Wang et al (2008, China), Topaloglu et al (2004, USA) 和 Yanagawa et al (2003, Japan)的引物无法根据原文进行定位

### 2.3.2 Meta 合并分析

如图 2-1 所示，APC 的甲基化阳性在非小细胞肺癌样本中的合并 OR 采用随机效应模型和固定效应模型分别为 4.67 (95% CI: 2.66-8.22, Z = 5.35, P < 0.0001) 和 2.74 (95% CI: 1.99-3.23, Z = 8.10, P < 0.0001)，为此可以看出 APC 在肺癌组织中的呈现显著的高甲基化状态。

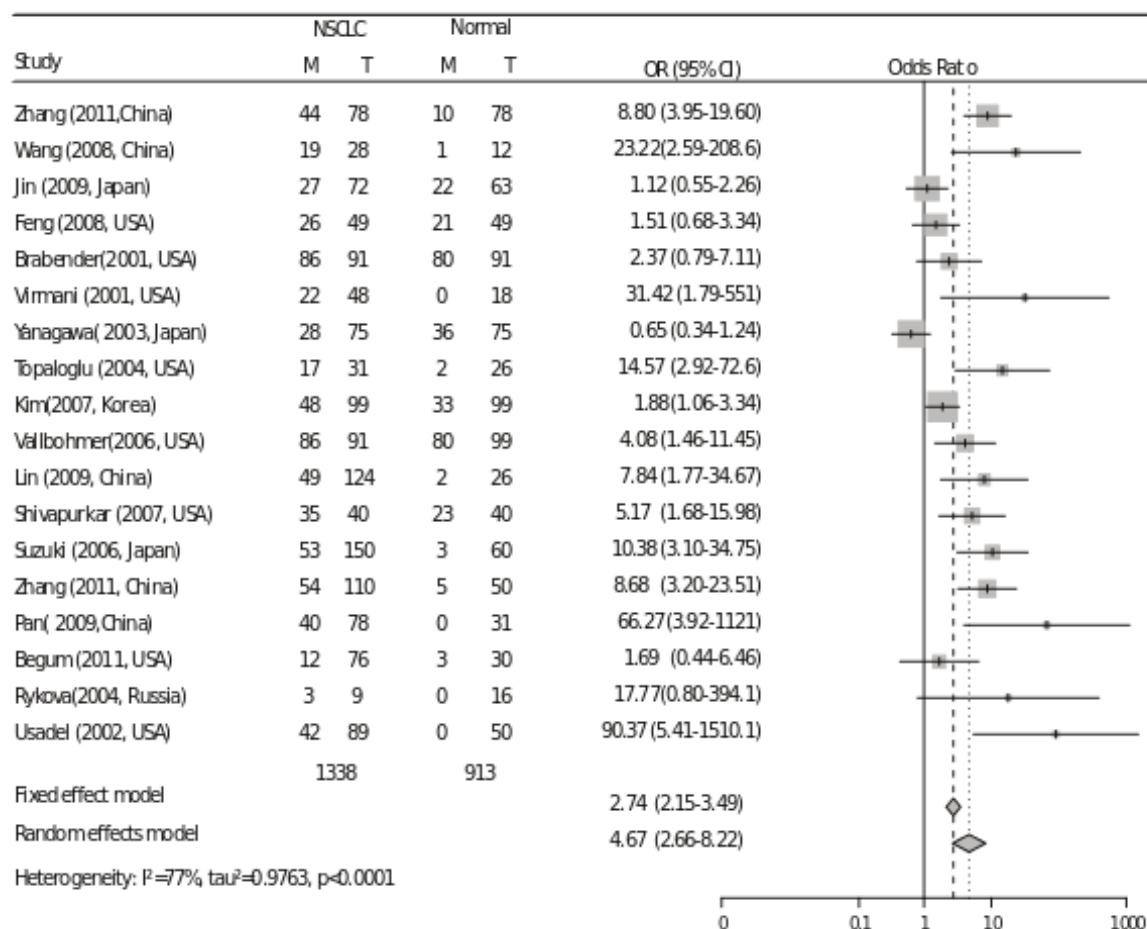


图 2-1 APC 基因 DNA 甲基化与非小细胞肺癌关联性强度的合并估计

注：采用森林图的方法展示 APC 基因 DNA 甲基化与非小细胞肺癌关联性强度的合并估计。作者，文献发表年限，实验室完成国家，肿瘤样本及正常样本甲基化阳性和阴性的样本个数。合并 OR 及其 95% 置信区间。DerSimonian-Laird 估计和 Mantel-Haenszel 分别被选择进行随机效应和固定效应模型的合并 OR 估计。

### 2.3.3 Meta 亚组分析

在不同的亚型的样本中，主要包括样品类型（组织或血清）、对照类型（自体或异体对照）、早期肿瘤比例、课题研究目的（用于诊断或无诊断）、腺癌鳞状比（Ad2Sc）、引物类别（I 和 II）等。APC 基因 DNA 甲基化与非小细胞肺癌的关联性强度具有一定的差异（如表 2-3 和表 2-5 所示）。

按照年龄中位数可以把所有研究分为高诊断年龄组亚组和低诊断年龄组亚组。分析发现高诊断年龄组亚组（0.91, 95% CI: 0.57-1.41）和低诊断年龄组（2.53-10.0 5.03, 95% CI）之间的合并 OR 值呈显著差异（ $P < 0.0001$ , 如图 2-2 的 A 图所示）。

按照研究样本中腺癌样本所占的比例，可以将所有研究分为两组：高腺癌亚组和低腺癌亚组。分析发现高腺癌亚组的合并 OR 显著高于低腺癌亚组（ $P = 0.0077$ ），这表明，APC 的甲基化可能对非小细胞肺癌中的腺癌具有更有效的诊断价值（如图 2-2 的 C 图所示）。

按照所使用的引物类型分类，显著性的 OR 差异也发现在第一套引物和第二套引物组（ $P = 0.0137$ ），这支持引物是在 APC 的甲基化诊断模型中存在较大异质性的重要来源（如图 2-2 的 D 图所示）。

在组织和血清亚组，APC 的甲基化和非小细胞肺癌之间均存在显著关联，合并 OR 分别为 3.72 和 11.54，提示血清 APC 的 DNA 甲基化可以作为用于使用组织或血清样品对非小细胞肺癌进行诊断的潜在的生物标志物。

对所采用的对照类型进行分类，可以分为异体对照研究组（其他正常个体的正常组织作为对照）和自体对照研究组（癌旁组织作为正常对照）。异体对照（ $OR = 8.33, 95\% CI: 3.77-18.39$ ）和自体对照（ $OR = 2.25, 95\% CI: 1.06-4.77$ ）亚组之间的合并 OR 具有显著的差异（ $P = 0.0187$ ）（如图 2-2 的 B 图所示）。出现这种结果一个可能的原因可能是邻近的正常标本可能已被癌细胞被轻微污染或已转化为癌前状态，为此采用癌-癌旁对照实验，有可能低估甲基化基因对肿瘤的潜在价值。

照同样本中的 TNM 分期比例的中位数分类，可以将所有研究分为高比例早期样本研究组（早期样本占主要）和低比例早期样本研究组（晚期样本占主要）。研究发现高比例 TNM 早期肿瘤样本亚组具有更高的合并 OR，表明 APC 的甲基化可在肿瘤发生的早期阶段发生或发挥作用。相似的证据在子宫内膜癌也被报道过[34]。

此外，以诊断为研究目的的亚组（OR=6.79）的合并 OR 是非以诊断为目的亚组的 2.6 倍（OR = 2.59），这可能是由于两个亚组中早期样本的比例不同而造成的（P = 0.0218，Wilcoxon 秩和检验）。

按照甲基化检测方法的不同，可以将所有研究分为两类：定性研究亚组(MSP) 和定量研究亚组(QMSP)。MSP 和 QMSP 亚组之间的合并 OR 没有发现显著性差异（P = 0.77），这表明这两种方法对甲基化检测具有等价的效果，这个也和之前的结论有相同之处[35]。

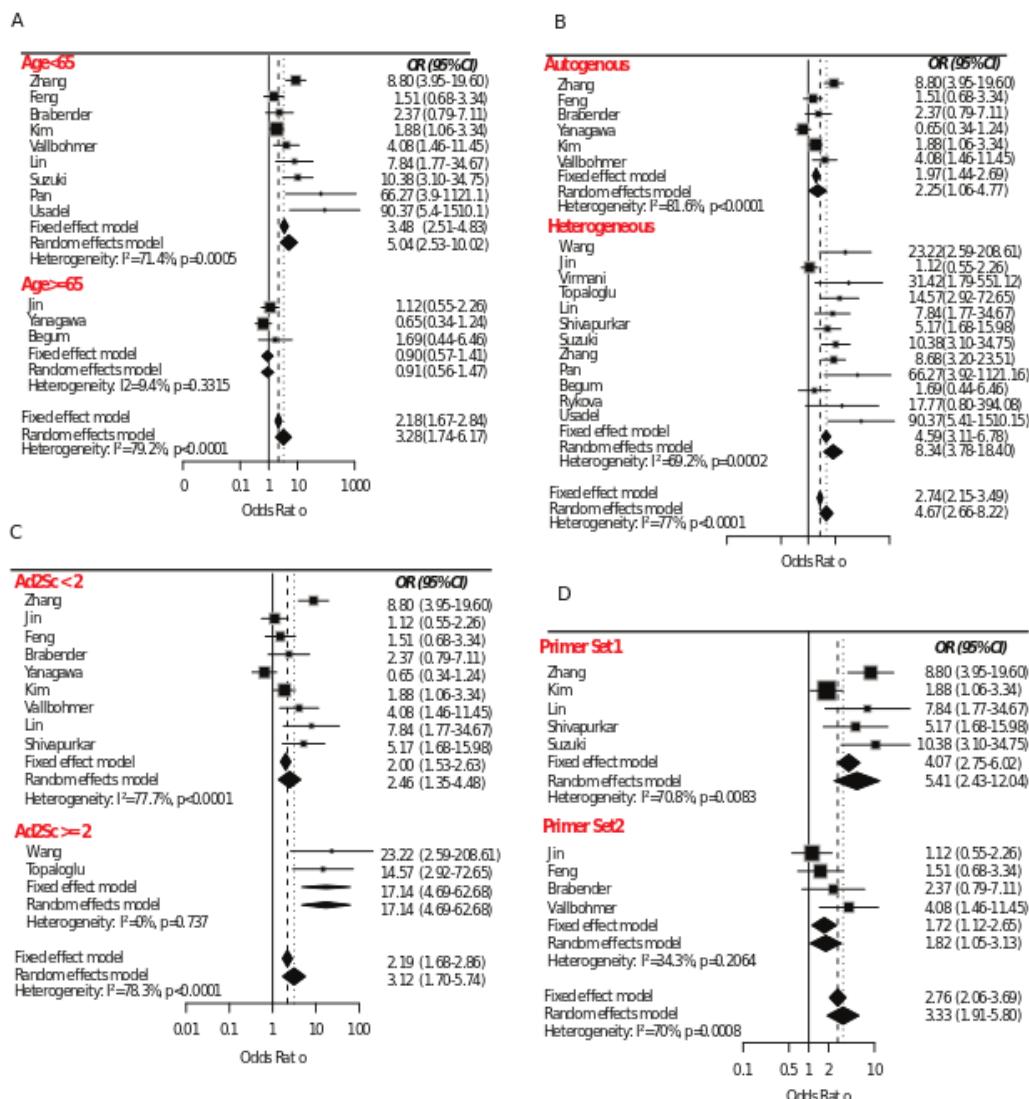


图 2-2 亚组的 Meta 分析

注：A-D 分别表示年龄，对照类型，腺癌在总样本中的比例，引物类型亚组的合并 OR 情况。E 是累计 Meta 分析结果. F 是综合受试者工作曲线特征分析。

表 2-3 混合效应模型下对主要混淆因素进行亚组分析

Subgroup	No. of Study	OR	95% CI	Q	I2	p-value
Overall	18	4.67	2.65-8.21	73.99	77.00%	
Age≤ 65	9	5.03	2.53-10.0	27.96	71.40%	
Age>65	3	0.91	0.57-1.41	2.21	9.400%	<b>&lt;0.0001</b>
Stage I>49.5%	5	4.11	1.90-8.91	12.76	68.60%	
Stage I≤ 49.5%	4	2.81	0.87-9.09	19.42	84.60%	0.5944
M2F≤ 69%	6	5.98	2.04-17.53	16.66	70.00%	
M2F> 69%	6	2.13	0.99-4.55	29.05	82.80%	0.1246
MSP	8	5.16	2.01-13.26	44.61	84.30%	
qMSP	10	4.32	2.08-8.94	29.28	69.30%	0.7685
Diagnose	13	6.79	2.99-15.44	59.54	79.80%	
Non-diagnose	5	2.59	1.33-5.05	11.56	65.40%	0.0745
Multiple targets	15	4.08	2.28-7.34	62.99	77.80%	
Single target	3	18.72	1.23-283	9.03	77.80%	0.2836
Heterogeneous	12	8.33	3.77-18.39	35.71	69.20%	
Autogenous	6	2.25	1.06-4.77	27.19	81.60%	0.0187
Serum	5	11.54	2.87-46.40	10.4	61.50%	
Tissue	13	3.72	2.03-6.78	55.18	78.30%	0.14
Ad2Sc < 2	9	2.46	1.35-4.48	35.79	77.00%	
Ad2Sc ≥ 2	2	17.1	4.68-62.7	0.11	0.000%	<b>0.0077</b>
Primer Set I	5	5.41	2.43-12.04	13.71	70.80%	
Primer Set II	4	1.82	1.05-3.13	4.57	34.30%	<b>0.0137\$</b>

注：合并模型采用随机效应模型。加粗显示的 P-value 为小于 0.05 的亚组。\$：如果该研究以血浆 DNA 甲基化为研究对象，且样本量小于 50，则这样的研究被排除在分析中，以降低偏差较强的小样本

### 2.3.3 Meta 回归分析

根据图 2-1 所示，17 项研究之间存在显著地异质性 ( $I^2 = 79.2\%$ ,  $Q = 52.78$ ,  $P < 0.0001$ )，因而我们采用 Meta 回归的方式，对其中潜在的混淆因素进行分析，发现年龄和引物组合是异质性的主要来源。合并 OR 的变化趋势与诊断时的年龄呈现显著地负相关 ( $\text{beta} = -0.3$ ,  $P = 2.0 \times 10^{-5}$ )，同时年龄变异可以解释总方差的 83.8%。这一结果与亚组分析高年龄诊断组 OR (OR = 2.24) 显着高于低年轻组 (OR = 4.65) 的 OR 是一致的。引物组也是一个重要来源异质性 ( $P = 0.05$ )，能够解释 68% 的总体异质性。其他因素如对照样本的类型，男性样本的比例，早期样本的比例，检测方法的差异均无法显著地对异质性进行解释(如表 2-4 所示)

表 2-4 基于随机效应模型的 Meta 回归对潜在的混淆因素进行分析

Subgroup	Coef. (95% CI)	P-value	$\tau^2$
Sample type	-1.03 (-2.4, 0.34)	0.14	0.90
Age	-0.3 (-0.44, -0.16)	<b>2.0×10<sup>-5</sup></b>	0.18
Proportion of Stage I	-0.01 (-0.05, 0.03)	0.608	0.79
Ratio of Male to Female	-0.69 (-8.1, 6.71)	0.855	0.98
Detection Methods	-0.09 (-1.28, 1.1)	0.88	1.11
Study Aim	-0.82 (-2.05, 0.41)	0.19	1.07
Single/Multiple Targets	1.05 (-0.71, 2.81)	0.243	1.01
Hetero/Autogenous Control	-1.25 (-2.35, -0.15)	<b>0.026</b>	0.89
Ad2Sc	0.44 (-0.56, 1.44)	0.387	0.89
Primer Set	-1.02 (-1.02, -2.02)	0.05	0.35

注：加粗显示的 P-value 为小于 0.05 的潜在异质性来源。

### 2.3.3 综合受试者工作特征曲线

在采用固定效应模型的前提下，可以对整个研究的汇总敏感性和汇总特异性进行分析，结果显示汇总敏感性和汇总特异性分别为 0.548 (95% CI: 0.42-0.67,  $p < 0.0001$ ) 和 0.78 (95% CI: 0.62-0.88,  $P < 0.0001$ )。实体组织亚组 (0.61, 95% CI: 0.45-0.75) 的诊断敏感性明显高于血清亚组 (0.396, 95% CI: 0.26-0.56)。而血清亚组的特异性(0.92, 95% CI: 0.86-0.96) 明显高于实体组织亚组(0.68, 95% CI: 0.49-0.83)。这表明这种生物标志物用兼具诊断和早期筛查的双重潜力。虽然敏感性和特异性是诊断测试或模型的重要指标，但是其受诊断阈值的影响，不恰当的阈值选择，可能会产生误导性的结果。相反，受试者工作特征 (SROC) 曲线往往能够更加客观地描述的诊断或预测模型的稳定和准确性。结果显示应用 APC 基因 DNA 甲基化对非小细胞肺癌的诊断效能的 AUC 为 0.64，这表明该预测模型具有一定的预测潜能，如

图 2-7 所示。尽管距临床诊断的标准（灵敏性和特异性>0.9）具有很大的差距。 预测模型的 AUC 在血清亚组和实体组织亚组分别为 0.67 和 0.64，这表明 APC 甲基化诊断模型对两种来源的样本呈现出类似的预测能力。

### 2.3.4 偏倚分析和稳定性分析

漏斗图分析根据所收集文献的 OR 及方差分析，显示数据集存在显著的发表偏倚（Egger 检验， $Z=4.3$ ,  $P<0.0001$ ），其中 8 个研究超过了 95% 置信区间（S 图 S1）。如图 2-3 所示。

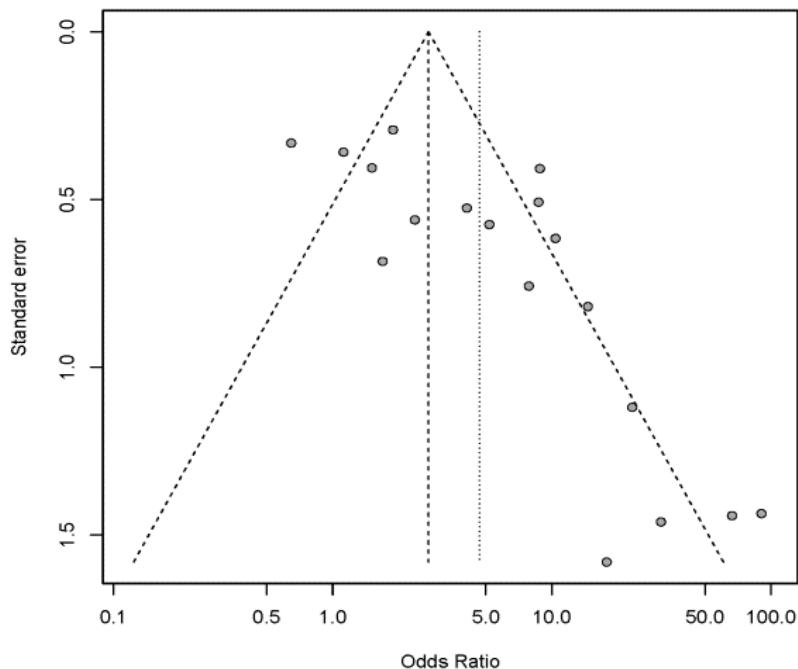


图 2-3 漏斗图分析用于展示发表偏倚情况

为了消除发表偏倚的影响，采用修剪和差补方法对随机效应模型和固定效应模型下的合并 OR 进行校正，结果显示，合并 OR 分别为 2.50 (95% CI: 1.43-4.38,  $P = 0.0013$ ) 和 2.19 (95% CI: 1.74-2.77,  $p < 0.0001$ )。这说明 APC 基因启动子区 DNA 甲基化与非小细胞肺癌之间的关联性是显著存在的。具体如图 2-4 所示。

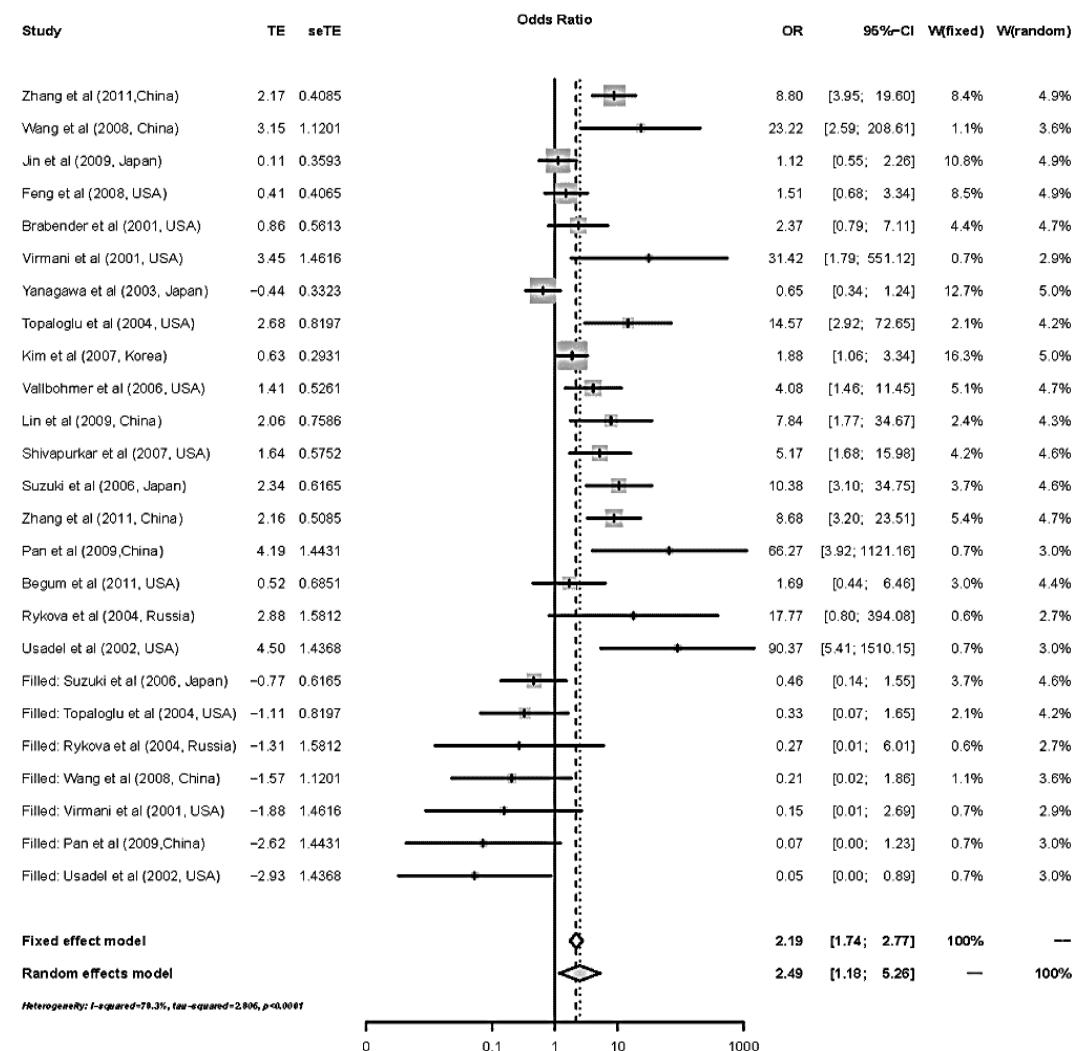


图 2-4 采用插补法对虚拟遗漏研究进行插补后再次进行合并比值比分析

在敏感性分析如下图 2-5 所示，采用随机效应模型的假设，忽略任何一个研究后的合并 OR 值的介于 4.3 (95% CI: 2.46-7.52) 和 5.27 (95% CI: 2.92-9.53) 之间。如上图所示，没有任何一项研究对整体的合并 OR 产生巨大的影响。

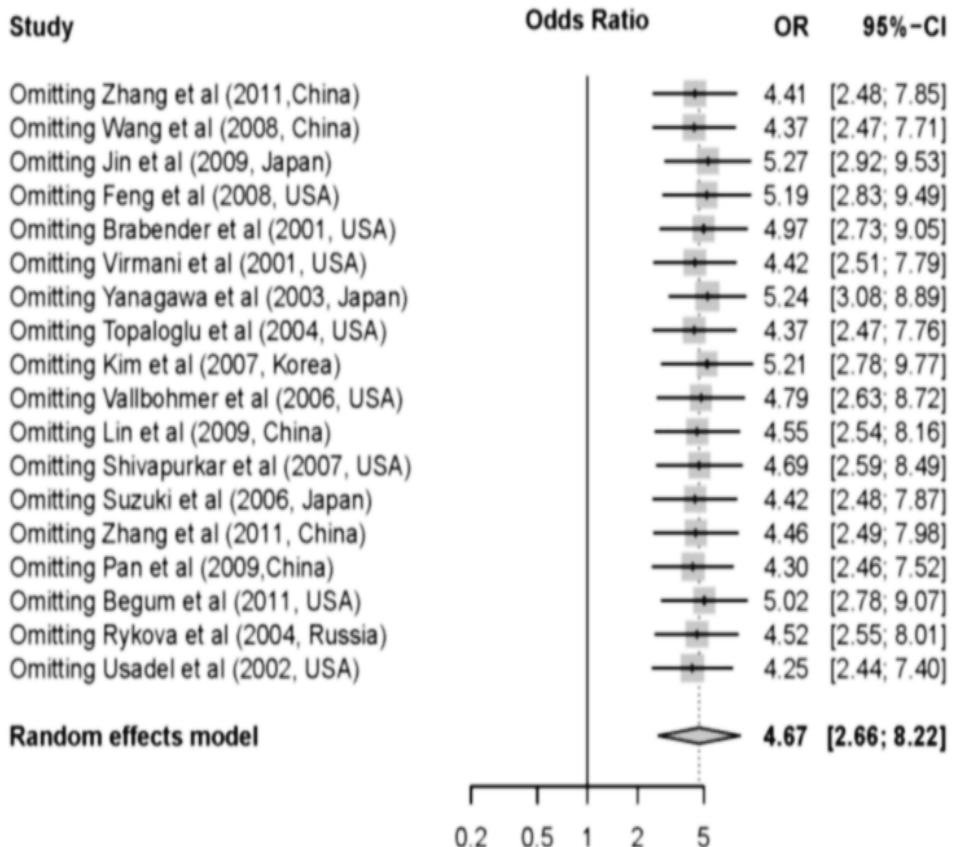


图 2-5 合并比值的敏感性分析

注：通过每次一个研究对象，分析剩余研究的合并 OR 以观察移除个别研究是否会对结果产生较大的影响。结果显示不论移除那一个具体的研究，都没有对最终的结果产生巨大的影响。

累积 Meta 分析按照随着时间的顺序，依次逐个加入文献，结果发现合并 OR 逐渐趋于稳定，表明当添加更多的文献研究后 Meta 分析更加可信，并结果逐渐处于稳定的状态。如图 2-6 所示。

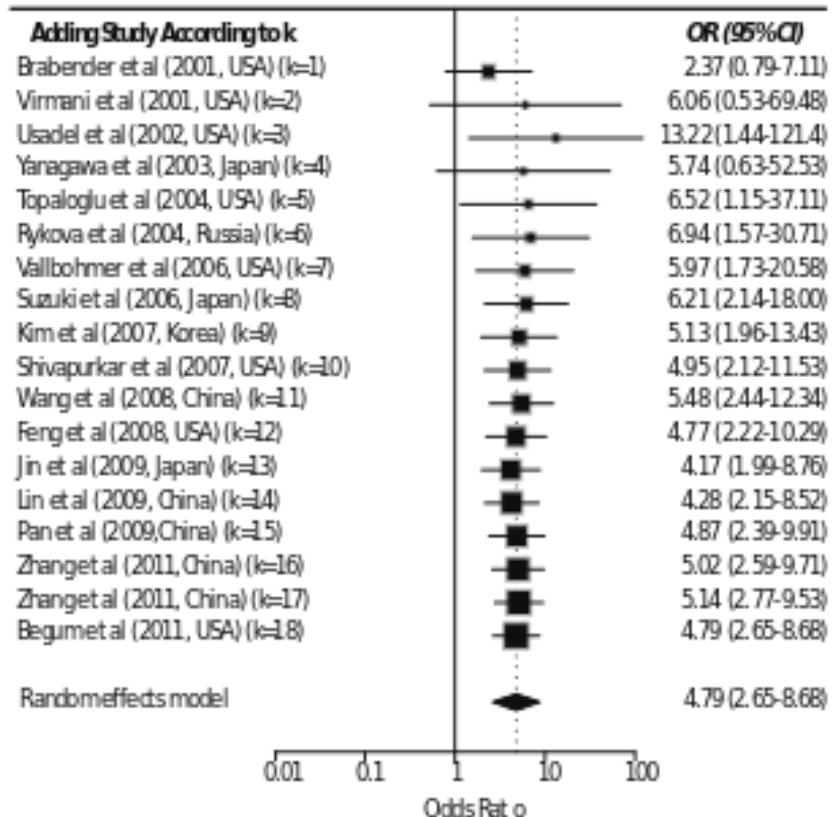


图 2-6 累计 Meta 分析显示逐渐稳定的结论

用类似的方法，我们分析 Meta 回归中对各潜在混淆因素分析时的稳定性分析。方法是在对任何一个混淆因素分析时，每次忽略其中一个研究，观察在这种情况下是否该混淆因素可以显著性地解释数据的异质性。

分析发现，如果移除 Begum et. al (2011, US) 组织类型（血清或实体组织）可以显著性地解释数据的异质性 ( $P<0.026$ )。当移除 Lin et. Al (China)，Zhang et. al( China)或 Yanagawa(Japan)其中任何一个的时候，早期肿瘤比例和研究目的都将成为异质性来源 ( $P$ -value 值分别为 0.0046, 0.029 和 0.039 分别)。这表明组织类型，早期肿瘤比例以及研究目的等因素都有可能是潜在的异质性来源的因素，需要在未来的临床推广研究中进行一定的关注。

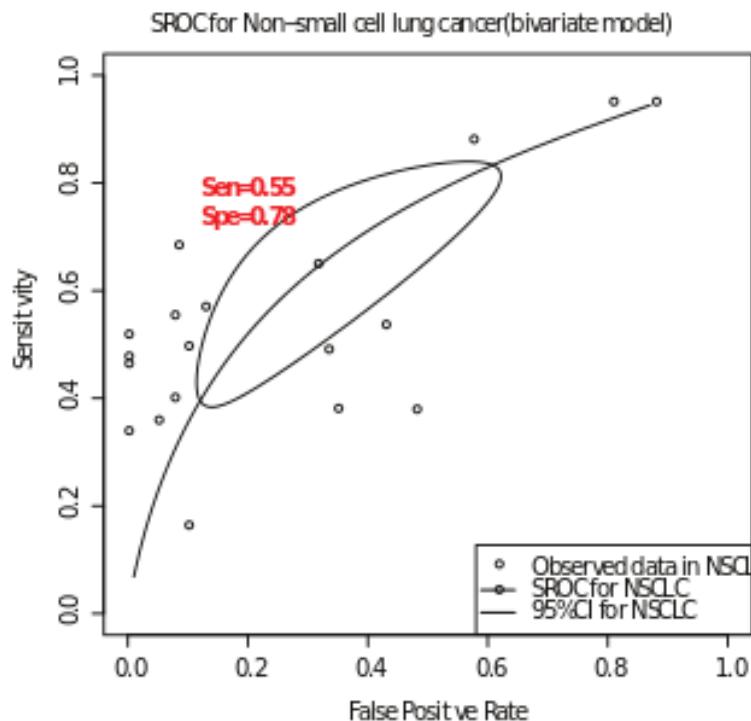


图 2-7 综合受试者工作特征曲线分析

### 2.3.5 独立 TCGA 肺癌的数据集的再验证分析

常规的 Meta 分析到上一个分析一般就结束了。为了对上述 Meta 的结论进行再次验证、或者说用其他事实进行再考察或辅助验证，我们收集的位于 TCGA 项目肺癌数据中 APC 基因的启动子区域的 DNA 甲基化数据。所涉及样本的临床资料如表 2-5 和表 2-6 所示。6 个 CpG 位点的两两甲基化状态的 Pearson 相关分析表明，除了 cg01240931 和其他 5 个位点之间的相关性较低外（平均  $R^2 < 0.45$ ），其余 5 个位点之间的甲基化状态呈高度相关的相关性（平均  $R^2 > 0.90$ ），这表明，上述 5 个位点构成了一个稳定的甲基化连锁区域（MethBlock）。为此为了保证结果的一致性，本部分研究排除 cg01240931，只研究为了甲基化连锁区域的其他五个位点。6 个 CpG 位点的详细资料如表 2-7 所示。

表 2-5 TCGA 相关样本的基本资料

	LUAD (%)	LUSC (%)
All cases	535	386
<b>Gender</b>		
Male	178 (33.3)	243 (62.9)
Female	211 (39.4)	83 (21.5)
<b>Age at diagnosis</b>		
<=65	193 (32.1)	199 (49.6)
>65	178 (49.7)	119 (37.4)
<b>Median(Range)</b>	65.3 (38-86)	67.5 (39-85)
<b>Stage at diagnosis</b>		
I	3 (0.56)	1 (0.25)
IA	96 (17.9)	55 (13.9)
IB	112 (20.9)	111 (28.0)
IIA	32 (5.98)	30 (7.58)
IIB	53 (9.90)	59 (14.9)
IIIA	58 (10.8)	41 (13.4)
IIIB	11 (20.6)	20 (5.05)
IV	22 (4.11)	4 (1.01)

表 2-6 病例对照分析中的 TCGA 相关样本的资料对比

All cases	LUAD (%)		p-value	LUSC (%)		p-value
	Case (%)	Control (%)		Case (%)	Control (%)	
Gender						
Male	178(33%)	26(46%)		243(63%)	52(74%)	
Female	211(39%)	30(54%)	0.96	83(22%)	17(24%)	0.99
Age(SD)	65.3 (9.8)	65.1 (9.8)	0.8794	67.5 (8.6)	68.3 (9.2)	0.5235

注: Age, age at diagnosis. SD, Standard Deviation

表 2-7 甲基化芯片中六个 APC 相关 Probe 的信息

CpG ID	Chromosome	MapInfo	Strand	Accession	Distance to TSS	CpG Island
cg01240931	Chr5	112101942	+	NM_000038.3	-459	FALSE
cg15020645	Chr5	112101668	+	NM_000038.3	-185	FALSE
cg16970232	Chr5	112101332	+	NT_034772.5	151	TRUE
cg20311501	Chr5	112101401	+	NT_034772.5	82	TRUE
cg21634602	Chr5	112101469	+	NM_000038.3	14	TRUE
cg24332422	Chr5	112101585	+	NM_000038.3	-102	TRUE

如表 2-6 所示, 本部分研究所涉及到的 TCGA 项目中的非小细胞肺癌样本的病例和对照样本之间不存在年龄或性别的显著差异, 数据显示这五个甲基化位点在病例和对照中都存在显著性的甲基化差异, 尤其是在腺癌的病例和对照中的比较中。

FDR 校正后的 *t* 非参检验显示在腺癌及其对照样本中所有这 5 个 CpG 位点的甲基化都存在显著差异, 而在鳞癌中只有两个 CpG 位点 (cg16970232, cg20311501) 在病例和相应用对照中存在显著差异 ( $P = 1.6 \times 10^{-6}$  和  $3.9 \times 10^{-3}$ 。如表 2-8 和表 2-9 所示)。

表 2-8 基于甲基化芯片的 APC 在肺腺癌中的差异甲基化

CpG Site	MCaM (N=535)	MCoM (N=56)	P-value <sup>¥</sup>	FDR <sup>¥</sup>	OR <sup>†</sup>	P-value <sup>†</sup>	95% CI <sup>†</sup>	AUC <sup>†</sup>
cg15020645	0.26(40.7%)	0.13(0%)	$3.5 \times 10^{-32}$	<b><math>1.0 \times 10^{-31}</math></b>	190.6	<b><math>7.7 \times 10^{-6}</math></b>	22.65-2321	0.72
cg16970232	0.3(45.2%)	0.11(0%)	$5.0 \times 10^{-38}$	<b><math>3.0 \times 10^{-37}</math></b>	108.9	<b><math>5.1 \times 10^{-6}</math></b>	17.64-1043	0.73
cg20311501	0.33(48.4%)	0.16(5.3%)	$1.4 \times 10^{-22}$	<b><math>2.1 \times 10^{-22}</math></b>	61.56	<b><math>4.96 \times 10^{-6}</math></b>	11.94-420	0.73
cg21634602	0.33(47.4%)	0.16(7.1%)	$3.6 \times 10^{-17}$	<b><math>4.3 \times 10^{-17}</math></b>	23.34	<b><math>3.6 \times 10^{-5}</math></b>	5.75-116.0	0.71
cg24332422	0.26(40.5%)	0.16(0%)	$1.0 \times 10^{-26}$	<b><math>2.0 \times 10^{-26}</math></b>	223.6	<b><math>2.81 \times 10^{-5}</math></b>	21.11-3463	0.71

注: <sup>¥</sup>基于 Wilcoxon 非参检验; <sup>†</sup>基于 Logistic 回归分析; McaM 肿瘤中的甲基化率; McoM 正常中的甲基化率

表 2-9 基于甲基化芯片的 APC 在肺鳞癌中的差异甲基化

CpG Site	MCaM (N=386)	MCoM (N=70)	P-value <sup>¥</sup>	FDR <sup>¥</sup>	OR <sup>†</sup>	P-value <sup>†</sup>	95% CI <sup>†</sup>	AUC <sup>†</sup>
cg15020645	0.13(14.77%)	0.11(0%)	0.087466	0.131199	3.16	0.406	0.28-68.72	0.61
cg16970232	0.15(18.91%)	0.09(0%)	$2.7 \times 10^{-7}$	<b><math>1.6 \times 10^{-6}</math></b>	7.54	<b>0.035</b>	1.39-64.07	0.45
cg20311501	0.18(19.95%)	0.14(0%)	0.001955	<b>0.003909</b>	2.48	0.257	0.57-13.74	0.49
cg21634602	0.16(20.47%)	0.14(7.14%)	0.222306	0.266767	1.27	0.726	0.35-5.42	0.53
cg24332422	0.16(17.36%)	0.15(0%)	0.338755	0.338755	1.6	0.656	0.23-14.30	0.52

注: <sup>¥</sup>基于 Wilcoxon 非参检验; <sup>†</sup>基于 Logistic 回归分析; McaM 肿瘤中的甲基化率; McoM 正常中的甲基化率

此外, 采用 logistic 回归分析也支持上述结果: 在腺癌中, 甲基化位点的 OR 值分别为从 23.3 到 190.6 (如表 2-8 所示), 而在鳞癌中 OR 仅仅是从 1.27 到 7.54 (如表 2-9 所示)。5 个 CpG 位点甲基化对鳞癌和腺癌的预测能力如表 2-8 和表 2-9 所示, 结果显示每个位点甲基化状态对腺癌的预测能力 (AUC: 0.71-0.73) 都高于对鳞癌的预测能力 (AUC: 0.45-0.61)。基于所有的 5 个 CpG 位点的肺癌预测模型的 AUC 值对腺癌和鳞癌分别为 0.73 和 0.60。所有上述结果表明 APC 的甲基化检测将在腺癌中有一定的预测能力。

前期的分析暗示，试验样本中腺癌和鳞癌的比例对于 APC 与非小细胞肺癌相关性的强度是有一定影响的。一般情况下，临床收集的肺癌样本中 25%-30% 是鳞癌，而 40% 是腺癌。因此，我们采用重抽样的方法从 TCGA 数据来模拟抽取不同比例的腺癌和鳞癌数据集，如 2:1, 4:3 3:4 和 1:2 的腺癌/鳞癌比例。然后来考察 APC 基因启动子区域 DNA 甲基化的非小细胞肺癌的 OR 的变化。10,000 次可重复抽样分析发现，不同的探针及不同的抽样比例下，APC 基因启动子区域 DNA 甲基化的非小细胞肺癌的 OR 存在很大的差异。不论是何种 Ad2SC 比例抽样，正如预期的那样，cg16970232 和 cg20311501 在不论腺癌还是鳞癌中总是呈现和 NSCLC 显著的相关性。因为这 2 个位点在 Ad 和 Sc 中都显著相关。而其他三个 CpG 位点与非小细胞肺癌的关联性只存在于某些特殊情况下的抽样中，这意味着这三个位点对于鉴定腺癌和鳞癌可能具有一定的帮助。此外，在 Ad2Sc 比例为 4:3 的抽样样本中，自体和异体对照两种不同情况下 5 个甲基化位点与 NSCLC 的相关性呈现出了 Meta 分析一致的结果。具体信息如表 2-10 和表 2-11 所示。

表 2-10 仿真状态下不同腺癌比例样本下的 APC 基因差异表达情况

Ad : Sc	2:1		4:3		3:4		1:2	
	CpG Site	OR	P-value	OR	P-value	OR	P-value	OR
cg15020645	7.8-173.4	0.001	11.64-56.37	<10 <sup>-4</sup>	4.86-25.90	0.02	3.84-16.44	0.572
cg16970232	17.2-186.4	<10 <sup>-4</sup>	18.99-72.49	<10 <sup>-4</sup>	14.24-50.49	<10 <sup>-4</sup>	8.9-31.73	<10 <sup>-4</sup>
cg20311501	6.4-76.6	<10 <sup>-4</sup>	8.09-36.23	<10 <sup>-4</sup>	5.84-20.38	<10 <sup>-4</sup>	3.49-13.58	<b>0.001</b>
cg21634602	3.3-31.14	0.014	3.72-12.78	<10 <sup>-4</sup>	2.78-8.30	0.022	1.74-6.33	0.582
cg24332422	4.6-203.3	0.006	9.17-48.01	<10 <sup>-4</sup>	5.47-22.77	0.012	2.79-12.21	0.546

注：上述分析结果基于配对的癌-癌旁数据；上述分析基于 10,000 次仿真分析。P-value 基于双样本 t-test。上述情况的样本量模式为：400:300、300:300、150:300、400:200。

表 2-11 仿真状态下不同腺癌比例样本下的 APC 基因肺癌关联性分析

Ad : Sc	2:1			4:3			3:4			1:2		
	CpG Site	OR <sub>a</sub>	OR <sub>h</sub>	p-value	OR <sub>a</sub>	OR <sub>h</sub>	p-value	OR <sub>a</sub>	OR <sub>h</sub>	P-value	OR <sub>a</sub>	OR <sub>h</sub>
cg15020645	572.4	43.36	<10 <sup>-4</sup>	264.6	24.41	<10 <sup>-4</sup>	0.27	3.1	<10 <sup>-4</sup>	3.26	12.6	<10 <sup>-4</sup>
<b>cg16970232</b>	149.2	59.40	<10 <sup>-4</sup>	83.3	37.84	<10 <sup>-4</sup>	14.1	16.3	<10 <sup>-4</sup>	11.42	21.5	<10 <sup>-4</sup>
cg20311501	44.6	24.88	<10 <sup>-4</sup>	24.4	17.61	<10 <sup>-4</sup>	20.3	25.5	<10 <sup>-4</sup>	3.31	10.1	<10 <sup>-4</sup>
cg21634602	23.8	9.80	<10 <sup>-4</sup>	12.6	7.53	<10 <sup>-4</sup>	6.1	12.6	<10 <sup>-4</sup>	1.88	4.8	<10 <sup>-4</sup>
cg24332422	192.7	42.02	<10 <sup>-4</sup>	64.7	24.46	<10 <sup>-4</sup>	3.3	5.84	<10 <sup>-4</sup>	2.42	11.1	<10 <sup>-4</sup>

注：a 表示自体对照模式下的分析结果；h 表示异体对照下的分析结果。上述分析基于 10,000 次仿真分析。P-value 基于双样本 t-test。样本量在上述情况模式为 300:400、400:300、560:280、200:400。

## 2.4 结论

APC 启动子的甲基化状态与非小细胞肺癌显著相关，尤其对于肺腺癌。APC 的甲基化检测具有一定的肺腺癌的临床辅助诊断价值。亚组分析显示样本诊断时年龄组成、所采用的引物、自体或异体对照、样本中腺鳞癌比例在一定程度上会给 APC 对非小细胞肺癌诊断的效能带来异质性。Meta 回归验证了其中的两个异质性来源即：诊断时年龄和所采用引物。此外，当采用结合敏感性分析的 meta 回归分析时，组织类型、早期肿瘤比例以及研究目的等因素都有可能是潜在的异质性来源的因素。

## 2.5 讨论

APC 基因已经在结直肠癌[36]被报告为一个重要的由异常 APC 的甲基化引起的抑癌基因，并且其异常甲基化现象在其他多种癌症中也有报道如：膀胱[37]、前列腺癌[38]、乳腺癌和肺癌[30]等。然而，到目前为止一直缺乏对 APC 基因 DNA 甲基化应用于非小细胞肺癌的诊断效能的定量评估。因此，本文进行了 APC 基因启动子区 DNA 甲基化对肺癌的诊断价值进行了定量的 Meta 分析。分析结果显示 APC 的甲基化与非小细胞肺癌具有显著相关性（ $OR = 4.67, P < 0.0001$ ）。考虑到发表偏倚，通过插补增加 7 个虚拟研究后，这种相关性依然显著存在（ $OR=2.49, 95\% CI: 1.18-5.26$ ），说明 APC 启动子甲基化与肺癌之间的显著关联的存在。APC 基因启动子区 DNA 甲基化对肺癌的诊断模型的灵敏性，特异性和 AUC 分别为 0.548, 0.78 和 0.64。说明 APC 的甲基化状态可以作为非小细胞肺癌诊断或辅助诊断的一个潜在的生物标志物。

基于 Meta 分析和 TCGA 数据仿真的综合分析表明，肺癌诊断时的年龄，自体或异构的对照，腺癌在总样本中的比率以及所采用的引物是造成不同研究之间出现异质性的重要原因。而样本类型（血清或实体组织），性别比例，TNM 早期样本比例和检测方法不会导致显著地异质性。复杂的异质性来源也是基于 DNA 甲基化至今没能在临床推广的重要原因。

多元 Meta 回归分析显示年龄是重要的异质性的来源之一（ $\betaeta = -0.3, P = 2.0 \times 10^{-5}$ ）。同时亚组中，数据也显示合并 OR 在较年轻的亚组（ $OR = 4.65$ ）显著高于在老年人群（ $OR = 2.24$ ），然后，采用 logistic 回归模型对年龄和甲基化的交互项进行显著性估计时，该交互项并不显著（ $P > 0.05$ ），所以可能还需要等多的实验针对这个问题进行探索。

亚组分析和 TCGA 分析都显示高 Ad2Sc 亚组的合并 OR 都高于低 Ad2Sc 组，提示 APC 的甲基化检测对腺癌具有更好的诊断性能。

自上个世纪八十年代开始，陆续开始有研究表明，肿瘤组织中的遗传和表观遗传改变，如 DNA 甲基化，也会出现在癌症患者的循环 DNA 中[39-41]。有趣的是，在本研究中，基于 Meta 分析发现，血清亚组的 APC 基因启动子区 DNA 甲基化的合并 OR 和实体组织中的合并 OR 没有显著差异，并略微比实体组织具有更大的 OR 值。这表明 APC 的甲基化检测将是一种很有前途的对非小细胞肺癌的诊断的血清标志物。

Meta 分析在分子流行病领域已广泛应用于 SNP-疾病风险的关联研究的定量评估，因为 SNP 位点有特异的基因组位置。这种分析及研究思路也逐渐被 DNA 甲基化相关的关联研究所采用，为此也正积极地推动 DNA 甲基化诊断价值的定量评估。

本研究对 DNA 甲基化分析所采用的引物进行了亚组分析，显示不同的引物对合并 OR 是具有一定影响的。但由于一个基因启动子区域往往具有上百或上千个 CpG 位点，这就造成了不同的研究可能采用了完全不同的引物序列，从而对相应的定量评估造成了一定的困难。对于引物多样性的问题，需要在今后的研究中思考一定的解决方案，从而更快的推动 DNA 甲基化研究向临床的转化。对于 APC 基因，通过我们对相关文献的阅读发现，至少存在 3 种不同的引物。为了进一步考察 APC 基因启动子不同的 CpG 位点的诊断能力，我们对 TCGA 中 5 个基因位点进行了相应的分析和比较，发现 5 个 CpG 位点的 OR 值有较大差异。亚组分析进一步表明在不同的引物组显著不同的 effect size。这些信息提醒我们，在未在的 DNA 甲基化相关研究中需要更加明确的注明信物的信息，在基因组中的位置，以方便其他临床或科研工作者对结果进行再整合或在分析。

总之，本研究通过 Meta 分析和 TCGA 独立样本的验证为 APC 启动子 DNA 甲基化对非小细胞肺癌的诊断价值提供了详细的定量评估。证明了 APC 启动子 DNA 甲基化与非小细胞肺癌之间的显著关联性。当然若要真正推动 APC 甲基化检测进入临床基因，肯定需要结合其他有效的 DNA 甲基化位点，从而使得总特异性和灵敏性都超过 90%。Meta 分析显示基于 DNA 甲基化的诊断会受到多种异质源的干扰，在临床实际应用中需要注意规避。尽管目前基于 DNA 甲基化的诊断技术还没有在临幊上得到推广，但是随着高通量 DNA 甲基化芯片的广泛应用和新一代测序技术的逐步推广，可以预测，基于 DNA 甲基化的肺癌早期筛查会有更多的突破，最终高效地应用于肺癌的诊断和早期筛查中。

表 2-12 17 篇文章的详细特征

Author (Published Year)	Sample Type	Age <sup>a</sup> (years)	Stages I %	Gender (M/F)	Patients (M+/M-)	Control (M+/M-)	Method	Aim	Multiple Target	Ad2Sc	Control Design
Zhang et al (2011, China) <sup>b</sup>	tissue	59	32.05	29/39	44/34	10/68	MSP	Diagnose	Yes	0.83	hom
Wang et al (2008, China)	tissue	-	-	17/28	19/9	1/11	qMSP	Diagnose	Yes	2.14	heter
Jin et al (2009, Japan)	tissue	66.7	-	17/24	27/45	22/41	qMSP	Non-Diag	Yes	1.87	heter
Feng et al (2008, USA)	tissue	64.3	42.86	26/49	26/23	21/28	qMSP	Diagnose	Yes	1.43	hom
Brabender et al (2001, USA)	tissue	63.3	49.45	69/91	86/5	80/11	qMSP	Non-Diag	Single	0.77	hom
Virmani et al (2001, USA)	tissue	-	-	-	22/26	0/18	MSP	Diagnose	Yes	NA	heter
Yanagawa et al (2003, Japan)	tissue	67.3	66.67	18/25	28/47	36/39	MSP	Diagnose	Yes	1.48	hom
Topaloglu et al (2004, USA)	tissue	-	54.84	-	17/14	5/17	qMSP	Diagnose	Yes	3.00	heter
Kim et al (2007, Korea)	tissue	63	56.57	64/79	48/41	33/66	MSP	Non-Diag	Yes	0.62	hom
Vallbohmer et al (2006, USA)	tissue	63	49.45	69/91	86/5	80/3	PCR	Non-Diag	Yes	0.77	hom
Lin et al (2009, China)	tissue	61.1	100.00	20/31	49/75	2/24	MSP	Diagnose	Yes	1.84	heter
Shivapurkar et al (2007, USA)	tissue	-	-	-	35/5	23/17	qMSP	Diagnose	Yes	1.22	heter
Suzuki et al (2006, Japan)	tissue	64	34.00	33/49	53/97	3/57	MSP	Non-Diag	Yes	NA	heter
Zhang et al (2011, China) <sup>b</sup>	serum	-	-	-	54/56	5/45	MSP	Diagnose	Yes	NA	heter
Pan et al (2009, China)	serum	53	-	17/26	40/38	0/31	qMSP	Diagnose	Single	NA	heter
Begum et al (2011, USA)	serum	65	-	10/19	12/64	3/27	qMSP	Diagnose	Yes	NA	heter
Rykova et al (2004, Russia)	serum	NA	-	-	3/6	0/16	MSP	Diagnose	Yes	NA	heter
Usadel et al (2002, USA)	serum	64.2	-	-	42/47	0/50	qMSP	Diagnose	Single	NA	heter

a 年龄根据每篇文章的不同采用相应的平均或者中位数年龄。b 来自同一篇文献文章同时对实体瘤和血浆进行了检测。M+, M- 分别代表甲基化阳性和阴性。Ad2Sc 表示样本腺癌和鳞癌的比值。hom, homogenous control 代表癌症组织和自体的癌旁对照; heter, heterogeneous control 代表癌症组织和非字体配对的癌旁对照。MSP 代表定性甲基化特异性 PCR; qMSP, quantitative detection method, 代表定量甲基化特异性 PCR.

## 2.6 参考文献

- [1] Siegel, R., D. Naishadham, and A. Jemal. *Cancer statistics, 2013* [J]. CA Cancer J Clin, 2013. **63**(1);11-30.
- [2] Nesbitt, J.C., J.B. Putnam, Jr., G.L. Walsh, J.A. Roth, et al. *Survival in early-stage non-small cell lung cancer* [J]. Ann Thorac Surg, 1995. **60**(2);466-472.
- [3] Gokul, G. and S. Khosla. *DNA methylation and cancer* [J]. Subcell Biochem, 2012. **61**;597-625.
- [4] Dietrich, D., O. Häsinger, V. Liebenberg, J.K. Field, et al. *DNA methylation of the homeobox genes PITX2 and SHOX2 predicts outcome in non-small-cell lung cancer patients* [J]. Diagn Mol Pathol, 2012. **21**(2);93-104.
- [5] Dietrich, D., C. Kneip, O. Raji, T. Liloglou, et al. *Performance evaluation of the DNA methylation biomarker SHOX2 for the aid in diagnosis of lung cancer based on the analysis of bronchial aspirates* [J]. Int J Oncol, 2012. **40**(3);825-832.
- [6] Fodde, R., J. Kuipers, C. Rosenberg, R. Smits, et al. *Mutations in the APC tumour suppressor gene cause chromosomal instability* [J]. Nat Cell Biol, 2001. **3**(4);433-438.
- [7] Deng, G., G.A. Song, E. Pong, M. Slezinger, et al. *Promoter methylation inhibits APC gene expression by causing changes in chromatin conformation and interfering with the binding of transcription factor CCAAT-binding factor* [J]. Cancer Res, 2004. **64**(8);2692-2698.
- [8] Usadel, H., J. Brabender, K.D. Danenberg, C. Jeronimo, et al. *Quantitative adenomatous polyposis coli promoter methylation analysis in tumor tissue, serum, and plasma DNA of patients with lung cancer* [J]. Cancer Res, 2002. **62**(2);371-375.
- [9] Tsou, J.A., J.A. Hagen, C.L. Carpenter, and I.A. Laird-Offringa. *DNA methylation analysis: a powerful new tool for lung cancer diagnosis* [J]. Oncogene, 2002. **21**(35);5450-5461.
- [10] Higgins, J.P., S.G. Thompson, J.J. Deeks, and D.G. Altman. *Measuring inconsistency in meta-analyses* [J]. BMJ, 2003. **327**(7414);557-560.
- [11] DerSimonian, R. and N. Laird. *Meta-analysis in clinical trials* [J]. Control Clin Trials, 1986. **7**(3);177-188.

- [12] Huizenga, H.M., I. Visser, and C.V. Dolan. *Testing overall and moderator effects in random effects meta-regression [J]*. Br J Math Stat Psychol, 2011. **64**(Pt 1);1-19.
- [13] Midgette, A.S., T.A. Stukel, and B. Littenberg. *A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates [J]*. Med Decis Making, 1993. **13**(3);253-257.
- [14] Jones, C.M. and T. Athanasiou. *Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests [J]*. Ann Thorac Surg, 2005. **79**(1);16-20.
- [15] Jones, C.M. and T. Athanasiou. *Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests [J]*. The Annals of thoracic surgery, 2005. **79**(1);16-20.
- [16] Walter, S.D. *Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data [J]*. Statistics in medicine, 2002. **21**(9);1237-1256.
- [17] Sproul, D., C. Nestor, J. Culley, J.H. Dickson, et al. *Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer [J]*. Proc Natl Acad Sci U S A, 2011. **108**(11);4364-4369.
- [18] Begum, S., M. Brait, S. Dasgupta, K.L. Ostrow, et al. *An epigenetic marker panel for detection of lung cancer using cell-free serum DNA [J]*. Clin Cancer Res, 2011. **17**(13);4494-4503.
- [19] Brabender, J., H. Usadel, K.D. Danenberg, R. Metzger, et al. *Adenomatous polyposis coli gene promoter hypermethylation in non-small cell lung cancer is associated with survival [J]*. Oncogene, 2001. **20**(27);3528-3532.
- [20] Feng, Q., S.E. Hawes, J.E. Stern, L. Wiens, et al. *DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients [J]*. Cancer Epidemiol Biomarkers Prev, 2008. **17**(3);645-654.
- [21] Jin, M., K. Kawakami, Y. Fukui, S. Tsukioka, et al. *Different histological types of non-small cell lung cancer have distinct folate and DNA methylation levels [J]*. Cancer Sci, 2009. **100**(12);2325-2330.

- [22] Kim, D.S., S.I. Cha, J.H. Lee, Y.M. Lee, et al. *Aberrant DNA methylation profiles of non-small cell lung cancers in a Korean population* [J]. Lung Cancer, 2007. **58**(1);1-6.
- [23] Lin, Q., J. Geng, K. Ma, J. Yu, et al. *RASSF1A, APC, ESR1, ABCB1 and HOXC9, but not p16INK4A, DAPK1, PTEN and MT1G genes were frequently methylated in the stage i non-small cell lung cancer in China* [J]. J Cancer Res Clin Oncol, 2009. **135**(12);1675-1684.
- [24] Pan, S.Y., E.F. Xie, Y.Q. Shu, L. Gao, et al. *[Methylation quantification of adenomatous polyposis coli (APC) gene promoter in plasma of lung cancer patients]* [J]. Ai Zheng, 2009. **28**(4);384-389.
- [25] Rykova, E.Y., T.E. Skvortsova, P.P. Laktionov, S.N. Tamkovich, et al. *Investigation of tumor-derived extracellular DNA in blood of cancer patients by methylation-specific PCR* [J]. Nucleosides Nucleotides Nucleic Acids, 2004. **23**(6-7);855-859.
- [26] Shivapurkar, N., V. Stastny, M. Suzuki, Wistuba, II, et al. *Application of a methylation gene panel by quantitative PCR for lung cancers* [J]. Cancer Letters, 2007. **247**(1);56-71.
- [27] Suzuki, M., H. Shigematsu, T. Iizasa, K. Hiroshima, et al. *Exclusive mutation in epidermal growth factor receptor gene, HER-2, and KRAS, and synchronous methylation of nonsmall cell lung cancer* [J]. Cancer, 2006. **106**(10);2200-2207.
- [28] Topaloglu, O., M.O. Hoque, Y. Tokumaru, J. Lee, et al. *Detection of promoter hypermethylation of multiple genes in the tumor and bronchoalveolar lavage of patients with lung cancer* [J]. Clinical Cancer Research, 2004. **10**(7);2284-2288.
- [29] Vallbohmer, D., J. Brabender, D. Yang, P.M. Schneider, et al. *DNA methyltransferases messenger RNA expression and aberrant methylation of CpG islands in non-small-cell lung cancer: association and prognostic value* [J]. Clin Lung Cancer, 2006. **8**(1);39-44.
- [30] Virmani, A.K., A. Rathi, U.G. Sathyanarayana, A. Padar, et al. *Aberrant methylation of the adenomatous polyposis coli (APC) gene promoter 1A in breast and lung carcinomas* [J]. Clin Cancer Res, 2001. **7**(7);1998-2004.

- [31] Wang, Y., D. Zhang, W. Zheng, J. Luo, et al. *Multiple gene methylation of nonsmall cell lung cancers evaluated with 3-dimensional microarray [J]*. Cancer, 2008. **112**(6):1325-1336.
- [32] Yanagawa, N., G. Tamura, H. Oizumi, N. Takahashi, et al. *Promoter hypermethylation of tumor suppressor and tumor-related genes in non-small cell lung cancers [J]*. Cancer Science, 2003. **94**(7):589-592.
- [33] Zhang, Y., R. Wang, H. Song, G. Huang, et al. *Methylation of multiple genes as a candidate biomarker in non-small cell lung cancer [J]*. Cancer Lett, 2011. **303**(1):21-28.
- [34] Ignatov, A., J. Bischoff, T. Ignatov, C. Schwarzenau, et al. *APC promoter hypermethylation is an early event in endometrial tumorigenesis [J]*. Cancer Sci, 2010. **101**(2):321-327.
- [35] Wu, T., E. Giovannucci, J. Welge, P. Mallick, et al. *Measurement of GSTP1 promoter methylation in body fluids may complement PSA screening: a meta-analysis [J]*. Br J Cancer, 2011. **105**(1):65-73.
- [36] Purnak, T., E. Ozaslan, and C. Efe. *Molecular basis of colorectal cancer [J]*. N Engl J Med, 2010. **362**(13):1246; author reply 1246-1247.
- [37] Eissa, S., M. Swellam, I.M. El-Khouly, S.K. Kassim, et al. *Aberrant methylation of RARbeta2 and APC genes in voided urine as molecular markers for early detection of bilharzial and nonbilharzial bladder cancer [J]*. Cancer Epidemiol Biomarkers Prev, 2011. **20**(8):1657-1664.
- [38] Trock, B.J., M.J. Brotzman, L.A. Mangold, J.W. Bigley, et al. *Evaluation of GSTP1 and APC methylation as indicators for repeat biopsy in a high-risk cohort of men with negative initial prostate biopsies [J]*. BJU Int, 2012. **110**(1):56-62.
- [39] Sozzi, G., D. Conte, L. Mariani, S. Lo Vullo, et al. *Analysis of circulating tumor DNA in plasma at diagnosis and during follow-up of lung cancer patients [J]*. Cancer Res, 2001. **61**(12):4675-4678.
- [40] Esteller, M., M. Sanchez-Cespedes, R. Rosell, D. Sidransky, et al. *Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients [J]*. Cancer Res, 1999. **59**(1):67-70.

[41] Jahr, S., H. Hentze, S. Englisch, D. Hardt, et al. *DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells* [J]. Cancer Res, 2001. **61**(4):1659-1665.

### 第三章 基于跨平台甲基化芯片数据建立肺癌早期诊断模型

DNA 甲基化被认为是一种非常有希望用于肺癌的辅助诊断的新兴生物标志物。文献表明肺癌和正常组织相比有大量的异常甲基化位点。差异甲基化位点数的数量级远远高于与肺癌显著相关的 SNP 数量（如本论文 4.3.9 部分）。为此，搜索最优的甲基化生物标记物组合以获得最大的诊断性能是一个巨大的挑战。此外，DNA 甲基化可以用于复杂疾病的诊断在 30 年前已经被提出，然而一直没有开发出可以用于肿瘤筛查和诊断的标记物或其组合。其原因一方面是因为上一章提到的复杂异质性来源问题，另一方面还因为之前的标记物研究的样本量不足，建立的预测模型在新的样本的扩展能力有限，此外还有肿瘤基因组异质性，不同个体可能存在不同的异常甲基化谱式，这样为建立一个统一的基于 DNA 甲基化的预测模型造成了一定的难度。最后还因为寻找最优化的诊断标记物组合具有一定的难度。为了解决这些问题，一方面我们要尽量提高肿瘤标记物开发中的样本量，此外还是建立一个最优的预测模型使之具有较好的外展性。目前的公共数据库中已经收录了大量肿瘤样本及其对照样本的全基因组 DNA 甲基化芯片数据，通过对这些数据的收集及再整合，可以有效地增加肿瘤标记物开发阶段的样本量。此外采用多种预测模型可以选择最佳的预测模型，同时结合交叉验证的方法，可以避免预测模型的过拟合，使得模型具有足够的外展型。

根据这种实验设计，我们开展了本部分的研究。本研究首先从高通量芯片数据库中收集并整合三个高通量的甲基化芯片。其次对整合数据进行标准化，批次效应消除等处理，获得了可应用于变量选择和预测模型建立的初始数据集。根据此数据集利用最优化变量选择的方法开发了一个高效的生物标志物组合，可以对非小细胞肺癌进行预测或诊断。之后，本研究利用一个中国汉族 NSCLC 回顾性队列对该标记物组合对非小细胞肺癌 (NSCLC) 的诊断效能进行了评估和验证。

在探索阶段，我们从 GEO 数据库中收集并整理了包含 458 个肺癌及正常组织的三个高通量 DNA 甲基化微阵列数据集。经过标准化，批次效应的消除和整合，我们采用 SVM (支持向量机) 的方法进行变量选择，最终确定了一个由 5 个甲基化位点组成的最优预测变量组合 (*AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1*)。在验证阶段，我们开发了基于 SNPshot 的甲基化状态确定单核苷酸引物延伸技术 (MSD-SNuPET) 对 150 对的非小细胞肺癌/正常组织个体的上述五个位点的甲基化状态进行了检测。结果显示这五个位点的甲基化在该队列人群的肿瘤组织中全部呈现异常甲基化，和探索阶段的结果吻合。之后，基于中国人群

的这五个位点的甲基化数据，我们采用 logistic 回归，支持向量机，随机森林预测及 Bayes 树的方法，结合 5 倍交叉验证的方法建立了肿瘤/正常组织预测模型，并对预测模型的灵敏性，特异性和准确性进行了评估。结果显示四个预测模型都具有很高的预测能力，贝叶斯树模型具有最优的敏感性，特异性和准确性，分别为 86.3%，95.7% 和 91%。基于 5 基因 DNA 甲基化状态(*AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1*) 及年龄，性别，吸烟史的 Logistic 模型的灵敏度，特异性，准确性，曲线下面积 (AUC) 分别为 78%，97%，87%，和 0.91。该研究表明，基于标准化处理，批次效应矫正的跨平台高通量 DNA 甲基化微阵列数据集，可以鉴定有效的肿瘤甲基化标记物。*AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1* 的联合甲基化谱可作为一种有效的非小细胞肺癌诊断模型。

### 3.1 研究背景

肺癌是一种涉及到遗传和表观遗传异常的复杂疾病，是全世界癌症死亡的主要原因[1]。肺癌中约 80% 的原发性肺癌是非小细胞肺癌 (NSCLC)，其特点是长期无症状的潜伏期，预后差。数据显示 III 和 IV 期 (晚期) 非小细胞肺癌患者的平均五年生存率只有 5%-14% 和 1%，然而早期患者 (I 和 II 期) 非小细胞肺癌 5 年生存率可以达到 50% 左右。除了早期诊断结合手术切除，目前在临幊上尚没有其他可以有效降低晚期肺癌患者生存率的治疗方法。现实临幊中肺癌的高死亡率暗示了采用影像学和细胞学检查为基础的诊断策略并没能有效地降低非小细胞肺癌患者的死亡率。而如果能够开发在早期阶段对非小细胞肺癌诊断进行诊断的技术，则可以通过手术切除的方式对肿瘤进行有效的控制，从而大大提高肺癌患者的五年生存率[2]。

遗传学及分子生物学的快速发展，特别是以全基因组学平台为基础的分子诊断技术为肿瘤的早期诊断提供了新的希望。各种遗传变异理论上都可以对疾病进行一定的预测和判定。目前已经存在一系列采用全基因组 SNP，全基因组甲基化谱式，miRNA 谱式或 mRNA 谱式对肿瘤进行不同层面的预测。在所有的遗传变异，SNP 得益于拥有近乎 100% 的检测准确性，在几乎所有的人提组织中都基本一致，且不随着年龄的改变而改变等特征，成为单基因所或寡基因遗传病并最稳定且有效的生物标志物。然而，对于复杂的疾病，特别是癌症，SNP 位点的风险预测能力是有限的。前期的研究表明基于疾病显著相关的 SNP 的非小细胞肺癌预测模型的 AUC 值仅为 0.54-0.55 [3]。不仅如此，甲状腺癌被认为是一个家族性

高风险癌各种癌症之一，我们前期的研究显示，基于疾病显著相关的 SNP 的甲状腺癌预测模型的 AUC 值也仅仅在 0.54-0.60 之间[4]，可见其他肿瘤基于 SNP 的风险评估更难获得理想的预测效果。目前，基于其他分子生物标志物如 mRNA, miRNA 和蛋白质非小细胞肺癌的诊断已经得到开发和研究，也都在一定程度上取得了不错的初步结果。然而，在过去几十年。对非小细胞肺癌的诊断准确性距离临床应用的超过 90% 的敏感性和诊断的特异性还有一定的距离。

DNA 甲基化在正常的生理和病理过程中都发挥重要的作用，涉及到基因和 miRNA 的表达调控[5]，基因的选择性剪接[6]等一系列重要的细胞生长及代谢过程，并且被证明在癌症早期阶段发挥重要作用。DNA 甲基化相对于 mRNA 具有更强的稳定性，相对于 SNP 具有更好的可塑性，相对于蛋白质其临床检测所要求的条件更为简单。DNA 甲基化可以收到外界环境的持续作用而发生改变，但一般不会发生类似 mRNA 发生类似节律性（昼夜周期）或心理学（情绪变化）等造成的瞬间变化。DNA 甲基化的特性模式需要在持续的外界作用或刺激才能形成。一旦形成又需要一定的细胞周期才能被再次改变，这种适度的可塑性恰恰满足疾病的产生过程和对疾病治疗的理念，因此 DNA 甲基化相对于 SNP [4], CNV[7] 和 microRNA[8] 具有更加优越的诊断和预后标记物的特征，因此被誉为最有前途的肿瘤诊断和预后生物标志物[9]。在过去几十年大量的非小细胞肺癌相关的异常甲基化位点已经被报道[10, 11]。尽管有几个基于 DNA 甲基化的诊断模型被建立[12]但这些研究都或多或少存在多种缺陷，如样本量小，选择的基因数低的和定性的而非定量 DNA 甲基化测量方式。这些缺陷会导致实验的可重复性较低，这也解释了为什么这些诊断模型的准确度存在较大差异甚至在其他的样本中无法重复等问题。

在本研究中，我们首先系统地从 GEO 数据库，ArrayExpress 及 TCGA 项目中整理了三个独立的高通量 DNA 甲基化数据集[13]。采用标准化和批次效应消除等方法对原始的整合数据进行降噪处理，然后从整合后的数据中采用 SVM 特征变量优化选择的方法，得到最优的 DNA 甲基化位点结合。最终 5 个甲基化位点 *AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1* 的组合被鉴定为对非小细胞肺癌的预测是的最优组合。之后，这五个位点的甲基化状态在一个由 150 对中国汉族人群 NSCLC 和其正常对照组成的回归性队列中进行了进一步的评估，为了降低对五个位点分别检测带来的批次问题，我们设计了一种新型的甲基化检测方法：甲基化状态依赖的单核苷酸引物延伸技术（MSD-SNuPET）同时在一个个体的组织中对这 5 个位点的甲基化的进行定量检测。

## 3.2 材料和方法

### 3.2.1 实验设计和流程

首先我们通过对公共高通量芯片数据库，包括 GEO 和 ArrayExpress 进行全面的搜查，以收集非小细胞肺癌相关的 DNA 甲基化微阵列数据。我们以非小细胞肺癌（NSCLC）以及 DNA 甲基化（DNA methylation）作为检索的关键词。虽然目前科研界已经有大量的针对非小细胞肺癌的高通量甲基化研究，但是只有两个 GSE 数据记录符合我们的研究目的，包括 GSE16559 和 GSE28094。GSE16559 是研究非小细胞肺癌和间皮瘤异常甲基化位点包括 57 肿瘤样本和 52 的正常组织样本。GSE28094 的目的是研究 1,628 个体不同的组织及病例状态样本的 DNA 甲基化的指纹，其中包括了 33 非小细胞肺癌和 3 肺正常组织样本。以上两种数据集都是基于 Illumina 公司 GoldenGate 的平台，其中包括 371 个基因位点的 1,536 个 CpG 位点。第三个数据集来自 TCGA 项目。TCGA 项目是目前最全面的肿瘤基因组研究工程，囊括了 38 种肿瘤全基因 SNP, CNV, DNA 甲基化，基因表达，蛋白质，miRNA 等一系列标记物的研究。此项研究中 TCGA 中包括了 262 例 NSCLC 和 51 例肺正常组织样本。TCGA 的肺癌数据集基于 Infinium 的 HM27 芯片，共覆盖 14,495 个基因和 27578 位点。GoldenGate 和 HM27K 平台共有 107 个（112 探针）重合基因。为此本研究最终包括了来自上述三个公共数据集的 458 非小细胞肺癌相关联的样本（352 例 NSCLC 和 106 正常组织）的 DNA 甲基化分析的数据。这些数据将被视为在生物标志物发现阶段的主要数据。

表 3-1 整合分析中收集的数据集

数据集	数据库编号	肺癌	正常对照	所发表杂志
Karl T. Kelsey(2009)	GSE16559	57	52	Cancer Res
Esteller M(2012)	GSE28094	33	3	Genome Res
TCGA(2014)	TCGA-2014	262	51	TCGA

当芯片数据为原始的甲基化荧光信号时，基因的甲基化水平按照如下公式进行计算：

$$\beta = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0)}$$

每一个基因在甲基化芯片中的对应有一个甲基状态指示探针区和一个非甲基化状态指示探针区，每个探针区一般都有 30 个左右的指示探针，其中 M 和

U 代表信号强度约 30 个甲基化 (M) 和非甲基化 (U) 阵列上的探针的信号值。实际在计算  $\beta$  时取这些探针信息号的最大值作为该基因的甲基化信号值。  
K 最近邻插补 (k-Nearest Neighbor imputation) 进行缺失值处理。两个微阵列平台之间共享 112 探针位点。分位数标准化方法 (quantile normalization) 应用于将不同研究中的所有数据进行标准化处理。为了进一步减少偏差, Combat 方法被用来对批次效应进行消除, 以避免存在于不同数据集的各种效应影响[14]。主成分分析 (PCA) 被用来可视化 Combat 去除批次相应的效果。基于上述数据, 差异甲基化分析, 最优预测模型对应的变量选择等分析依次被执行。差异甲基化分析采用非正太分布依赖的 Wilcoxon 符号秩检验。变量选择是由支持向量机预测模型 (SVM) 与留一交叉验证 (leave-one cross validation) 进行。为此, 最优的甲基化标记物组合得以鉴别和确定。在验证阶段(validation stage), 预测最优组合的甲基化基因的甲基化状态采用 MSD-SNuPET 技术在 150 例中国汉族人群 NSCLC 和肺正常组织中进行了检测。四种预测模型包括 Logistic 回归模型, 随机森林, 支持向量机 (SVM) 和贝叶斯树结合 5 倍交叉验证被用来验证优化得到的最后甲基化位点组合再真实数据中的预测准确度, 灵敏性和特异性。整个实验的流程如图 3-1 所示。

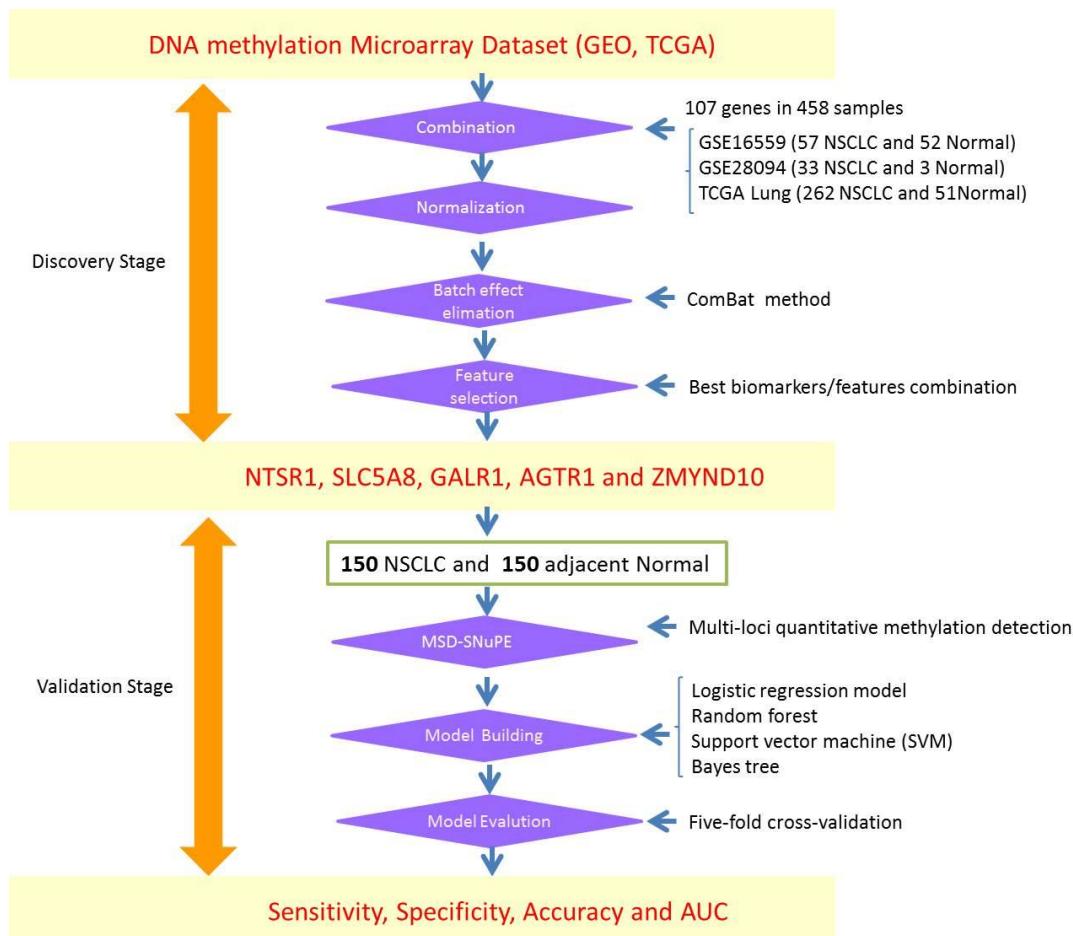


图 3-1 甲基化标记物开发流程图

注：潜在有效标记物组合采用整合分析跨平台的高通量芯片进行确定。然后在独立的汉族非小细胞患者的病例对照样本中进行确认。

### 3.2.2 患者样本及 DNA

验证研究队列中的 150 对中国汉族人群非小细胞肺癌标本和相应的正常肺组织来自 2012-2014 年间在长海医院（上海，中国）接受肺切除术的原发性非小细胞肺癌患者的捐赠。该研究得到复旦大学和长海医院的伦理审查，所有患者均签署了相应的知情同意书。受试者的肺癌样本如果包括术前放疗，化疗及辅助治疗，则排除在研究所采用的范围之外。所有的组织手术切除后立即冷冻在-80℃。组织学检查和肿瘤淋巴结转移的分类根据世界卫生组织的分类标准[15]和 AJCC 癌症分期手册第 7 版[16]进行确认。在进行 DNA 甲基化与疾病状态之间的关联时，年龄，性别，吸烟状况，组织学类型，TNM 分期和分化状态等协变量。吸烟状况采用二分类：从不吸烟和曾经吸烟。TNM 分期也采用二分类方法，包括早期（I 和 II）或晚期（III 和 IV）。样本的详细信息如表 3-2 所示。

表 3-2 本研究中非小细胞样本资料

	样本总数: 150
年龄	40 (IQR = 15-65)
性别	
男性	120
女性	30
吸烟状态	
非吸烟患者(从未吸过烟)	41
吸烟患者 (曾经吸过烟)	96
组织学分类	
肺腺癌	53
肺鳞癌	63
其他	34
肿瘤 TNM 分期	
I (IA,IB)	42 (10,32)
II (IIA,IIIB)	48 (16,32)
III (IIIA,IIIB)	46 (41,5)
IV	2
肿瘤分化程度	
中高分化	74
低分化	30

注：组织学分类的其他包括腺鳞癌，细支气管肺泡癌，粘液表皮样肿瘤和肺肉瘤样癌。TNM 分期参考第 7 版 TNM 分类标准。

### 3.2.3 MSD-SNuPET：甲基化状态相关的单核苷酸引物延伸法

DNA 提取和亚硫酸氢盐转化，进行与我们先前描述[17, 18]。甲基化状态确定单核苷酸引物延伸技术 (MSD-SNuPET) 是专为甲基化的同时在多个甲基化位点进行量化。MSD-SNuPET 技术是基于 SNPshot 技术上，以亚硫酸氢钠转换的 CpG 位点为检测对象。未甲基化的胞嘧啶在一定条件下可以被亚硫酸盐处理处理的方法转换为尿嘧啶，而同样的处理下甲基化胞嘧啶仍然保持为胞嘧啶。甲基化不同状态在亚硫酸盐处理后形成类似 SNP 的多态性恰好可以被 SNPshot 的方法通过特异性引物和 PCR 扩增进行测定，因为我们称这种方法为甲基化状态确定单核苷酸引物延伸技术 (MSD-SNuPET)。SNPshot 的方法为我们实验室 20 多年来一直采用的成熟技术，具体流程和细节可参考实验室前期一系列文章和著作[19-21]。SNPshot 技术中的引物对设计采用在线引物设计软件 primer-3.0 来设计（称为扩增引物），其被应用来扩增基因组区域包括目标 CpG 位点。等位基因特异性延伸引物被用来量化的 C 和 T 等位基因的拷贝数。引物对信息如下表。

PCR 在 10 $\mu$ L 含有 1x HotStarTaq 缓冲液, 3.0mM 的 Mg<sup>2+</sup>, 0.3mM 的 dNTP, 1 U HotStarTaq 聚合酶盒 (Qiagen Inc, 美国), 1ul DNA 模板, 并加入 1ul 混合引物。DNA 扩增在 GeneAmp PCR 系统 9700 热循环仪上进行 (Applied Biosystems 公司, 福斯特城, 加利福尼亚州, 美国), PCR 过程如下: 95°C 预变性 2 分钟, 随后 11 个循环, 94°C 变性 20 秒, 60°C 退火 40 秒, 72°C 延伸 90 秒, 循环结束后再 72°C 终延伸 2 分钟。阴性和阳性对照被包括在如上所述的 PCR 扩增的每次运行。测序引物的纯化和 SNPshot 的步骤请参考我们之前的文章[22]。DNA 测序结果的分析采用 3730 DNA 分析仪进行。GeneMapper 4.1 (Applied Biosystems 有限公司, 美国) 被用于等位基因的荧光信号判定。

### 3.2.4 MSD-SNuPET 技术中甲基化信号的估计

在 MSD-SNuPET 技术中 DNA 甲基化水平正相关于 C 等位基因 (HC) 的丰度, 同时负相关于与 T 等位基因 (HT) 的丰度。样本经甲基化处理后甲基化位点基因型为 C, 未甲基化位点基因型为 T。样本经 snapshot 方法延伸反应所得波峰高度为 HC-S, HT-S。标准样本延伸反应所得波峰高度为 HC, HT。CC 为峰型为 C 的延伸产物的浓度, CT 为峰型为 T 的延伸产物的浓度, 基因型为 C 的 PCR 产物片段所占百分比为 C%, 基因型为 T 的 PCR 产物片段所占百分比为 T %, 在同一条件下, PCR 模板浓度和 PCR 产物片段浓度正相关。因为波峰高度与该等位基因浓度正相关, 即  $HC/HT = k \cdot CC/CT$ ,  $k = HC \cdot CT / HT \cdot CC$  (在同一条件下, k 为常数)。实验中做了 11 个标准品, 其 C% 分别为 10% 20% 30% 35% 40% 50% 60% 70% 75% 80% 90%, 对标准样本进行延伸反应可以得到对应的 11 个峰高比值 (HC/HT), 以实验得到的标准样本的 HC/HT 为横坐标, 其对应的 CC/CT 为纵坐标作图, 可以得到一条相应的曲线和方程式 (从理论上来说, 得到的曲线和对应的方程式应该是线性的, 但是实际作出的曲线更接近二元方程式及对应的曲线) 及:  $y = \beta_0 x^2 + \beta_1 x$ , 其中  $\beta_0$  和  $\beta_1$  是需要进行估计的参数。根据得到的每个位点的方程式  $y = \beta_0 x^2 + \beta_1 x$  以得到每个样本理论的 C 和 T 的浓度比 (CC/CT), 其中  $X = HC-S/HT-S$ ,  $Y = CC/CT$ 。那么  $C\% = CC/(CC+CT) = (CC/CT) / (CC/CT + 1)$

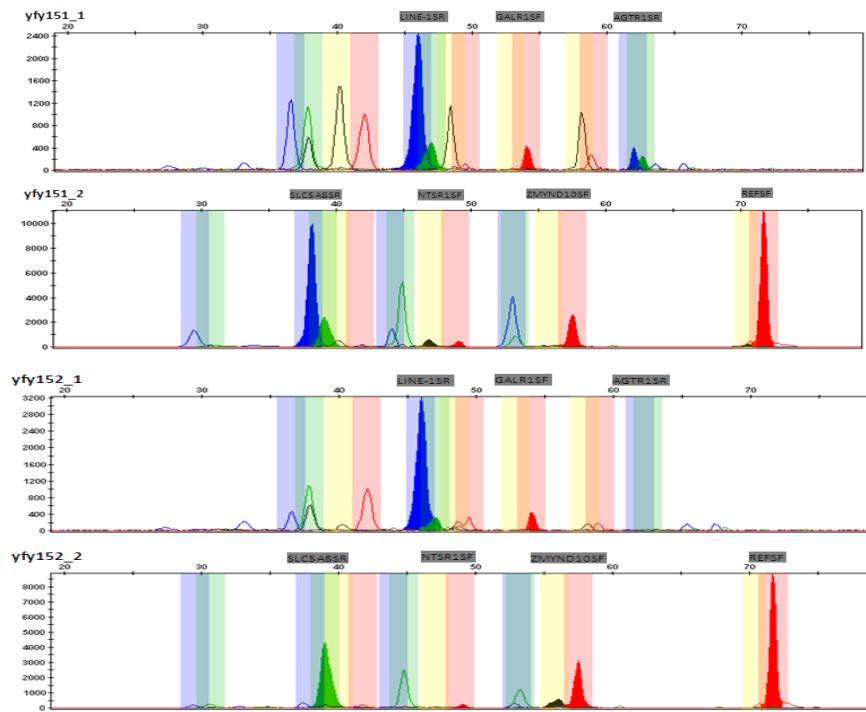


图 3-2 多位点 MSD-NEuTEP 技术分析结果示意图

如图 3-2所示，扩增产物的长度不同会导致起扩增产物在经过毛细管电泳时会出现在不同的位置，扩增产物的丰度曾可以用荧光信号的峰高进行反应。因此多个甲基化位点的信号可以在一次反应进行测定。图中的yfy151为肿瘤样本，yfy152为正常样本。yfy151-1为LINE-1, GALR1和AGTR1基因在肿瘤样本中的甲基化信号示意图，yfy151-2为SLC5A5, NTSR1, ZMYND10和隐性对照REF在肿瘤样本中的甲基化信息示意图。

此外为了保证实验数据质量，本实验还加入了一个技术控制对照和一个生物控制对照。技术控制对照采用一个测量一个非 CpG 位置的 C 位点进行质量控制，因为非 CpG 位置的 C 位点在经过烟硫酸盐处理后会 100% 转变为 T，所以甲基化信息理论值为 0；生物控制对照，采用在引物对中加入 LINE-1 的测定，根据大量的实验研究 Line-1 在肿瘤组织中呈现显著地去甲基化现象，所以理论上 LINE-1 在肿瘤组织中的甲基化应该显著低于正常组织的甲基化。

### 3.2.5 统计分析和机器学习

在用于探索的数据分析中，CpG 位点的甲基化信号在肿瘤和正常组织的分布差异采用 *t*-检验或 Wilcoxon 符号秩和检验进行测试。当甲基化信号在正常和肿瘤中的分布都符合正太分布时采用 *t*-检验的方法，否则采用 Wilcoxon 符号秩和检验的方法。FDR 校正 (False Discovery Rate) 被用于多重检验的显著性校正。

采用欧氏距离(Euclidean distance)和分区围绕中心点( Partitioning around medoids )的参数设定进行聚类分析。 Logistic 回归分析 (R, stats 包), 支持向量机 (R, e1071 分析包), 随机森林为基础的分类 (R, randomForest 包) 和贝叶斯树 (R, BayesTree 包) 被用来在非小细胞肺癌肿瘤和正常组织进行预测模型建立。所有统计分析都在 Linux 平台下的 R[23]下进行的。甲基化位点之间的潜在相互作用参考 String 9.0 提供的蛋白质-蛋白质相互作用网络[24]。

### 3.2.5.1 Combat 批次效应处理

ComBat(Combining Batches)是由 Li Cheng 和 Johnson 在 2007 年提出的, 基于经验贝叶斯方法对批次效应对不同基因带来的加性和乘性效应进行估计的一种方法。根据参数的实际分布和先验分布的不同, combat 可以采用参数或非参数方法。由于经验贝叶斯的特点, Combat 对于小样本数据的批次效应的校正具有较大优势。陈超博士在 2011 年对主要的 6 种批次效应校正方法进行了比较, 发现 Combat 是在所有方法中表现最突出的一种方法[14]。Combat 方法更多的细节可参考 Johnson 和他的同事 2007 年发表在 Biostatistics 上的论文[25]和陈超博士的学位论文[26]。Combat 的基本思想如下:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

其中  $Y_{ijg}$  是总体信号值,  $\alpha_g$  是平均信号值,  $X$  是样本状态矩阵,  $\gamma_{ig}$  和  $\delta_{ig}$  分别代表第  $i$  个批次的加性和乘性效应。 $\varepsilon_{ijg}$  是随机误差, 默认服从均值为 0, 方差为  $\sigma_g^2$ 。经过批次校正后的信号值的计算公式如下:

$$Y_{ijg}^* = \frac{Y_{ijg} - \widehat{\alpha}_g - \widehat{X\beta}_g - \widehat{\gamma}_{ig}}{\widehat{\delta}_{ig}} + \widehat{\alpha}_g + \widehat{X\beta}_g$$

其中  $\widehat{\alpha}_g$ ,  $\widehat{\beta}_g$ ,  $\widehat{\gamma}_{ig}$  和  $\widehat{\delta}_{ig}$  分别为  $\alpha_g$ ,  $\beta_g$ ,  $\gamma_{ig}$ ,  $\delta_{ig}$  的估计。

### 3.2.5.2 随机森林学习模型

随机森林 (Random forests, RF) 是由 Leo Breiman (2001) 提出一种高效的基于 tree 的机器学习方法。其综合了 Bootstrap aggregating 和 random subspace method 两种思想。随机森林通过 bootstrap 重采样技术, 从原始训练样本集 N 中有放回地重复随机抽取 K 个样本生成新的训练样本集合。此外, 随机森林同时对变量也进行了抽样, 只选取部分变量进入新的训练样本集合。根据这 K 个对样本和变量重抽样产生的训练样本集合, 生成 K 个分类树组成随机森林, 新数据的分类结果按分类树投票多少形成的分数而定。其实质是对决策树算法的一种

改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。由于抽样的原因，会有一些未被抽中的样本，这形成了所谓的袋外数据 (OOB, out of bag)，可以根据 OOB 来估计泛化误差，而不需要用交叉检验来估计。这样就意味着随机森林内嵌了 train 和 test 的过程，不需要进行外部的交叉验证。

### 3.2.5.3 逻辑斯蒂回归预测模型

逻辑斯蒂模型 (Logistic regression) 是广义线性回归中的一种，针对因变量为类别变量的情况，如果因变量为两类则对应 Binary Logistic regression；如果因变量为多类，则对应 Multinomial Logistic regression。其自变量可为定性或定量数据。对于 Binary Logistic regression 的回归模型如下：

$$p = \frac{\exp^{a+\sum \beta_i X_i}}{1 + \exp^{a+\sum \beta_i X_i}}$$

$p$ 是事件发生的概率,其值在 0 到 1 之间,值越大则发生的可能性越大。 $X$ 是影响概率分布的因子, $\alpha$ 和 $\beta$ 是需要估计的参数。对上式取自然对数，就得到了 Logistic regression 的经典表示形式：

$$\ln \frac{p}{1-p} = a + \sum \beta_i X_i$$

假设个体的属性事件为  $Y$ ，那么  $Y=1$  表示该个体或样本为肿瘤，相反  $Y=0$  表示个体或样本为正常。在本文中  $p$  表示待评估个体为肿瘤患者 ( $Y=1$ ) 的概率。 $X$  是甲基化位点，年龄，性别等 covariates。 $\alpha$  和  $\beta$  是 Logistic 回归方程的常数项和自变量系数，而自变量  $X_i$  对因变量的影响程度可以用  $\exp^{\beta_i}$  进行衡量。 $\exp^{\beta_i}$  表示肿瘤与不是肿瘤概率之比，也就是我们常说的发生肿瘤的 Odds ratio (OR)，这里表示自变量每增加一个单位，OR 就相应增加  $\exp^{\beta_i}$  个单位。

### 3.2.5.4 支持向量机预测模型

支持向量机的原理是用分类超平面将空间中两类样本点正确区分，并取得正样本与负样本到超平面的最小距离（边缘）最大化。Andrew W. Moore 给出了一种 SVM 最直观的示意图，如下图所示。

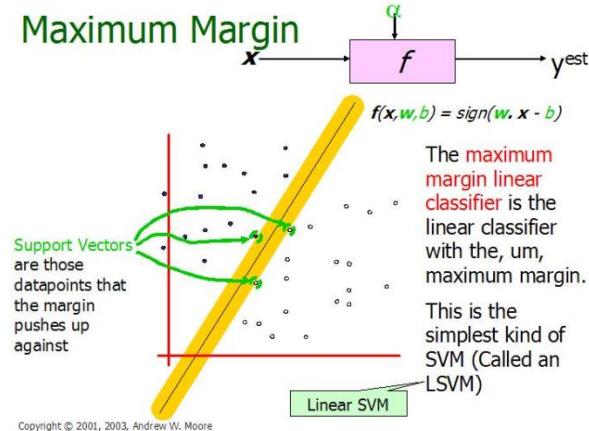


图 3-3 支持向量机示意图

注：引自于 Andrew W. Moore, 2003 用于辅助说明 SVM 的基本原理。

因此原问题转变为一个有约束非线性规划问题，我们只需要找出最大的超平面的最小距离即可对模型进行最大准确度的分类或预测。如

$$\text{margin} \equiv \underset{\mathbf{x} \in D}{\operatorname{argmin}} d(\mathbf{x}) = \underset{\mathbf{x} \in D}{\operatorname{argmin}} \frac{|\mathbf{x} \cdot \mathbf{w} + b|}{\sqrt{\sum_{i=1}^d w_i^2}}$$

上述即可 SVM 的经典表达形式。本文所采用的 SVM 算法来自 R 语言中的 e1071 程序包。

### 3.2.5.5 贝叶斯树预测模型

朴素贝叶斯树(NBTree)是决策树方法 (C4.5) 和贝叶斯思想的结合。C4.5 算法是由 Ross Quinlan 开发的用于产生决策树的算法。该算法是对 Ross Quinlan 之前开发的 ID3 算法的一个扩展。本文所采用的贝叶斯树算法来自 R 语言中的 BayesTree 程序包。

### 3.3 结果

#### 3.3.1 公共数据的收集和合并

通过对多个肿瘤高通量数据库（GEO, ArrayExpress 和 TCGA）的检索，三个独立的非小细胞肺癌甲基化相关数据集被检索到，共包括 458 芯片，其中 352 例 NSCLC 和 106 正常肺组织样本。

#### 3.3.2 批次效应的评估和消除

采用主成分分析的方法对整合后数据集进行分析时发现：第一主成分和第二主成分主要反映了数据来源的变异，即批次效应解释了数据的主要变异。如图 3-4 所示，Combat 是一个基于经验贝叶斯方法可以用来消除批次效应的算法。通过 Combat 方法对不同批次的加性及乘性效应进行校正后，可以观察到 Combat 校正后的数据，可以发现这种主要来自数据来源的差异已经被消除。

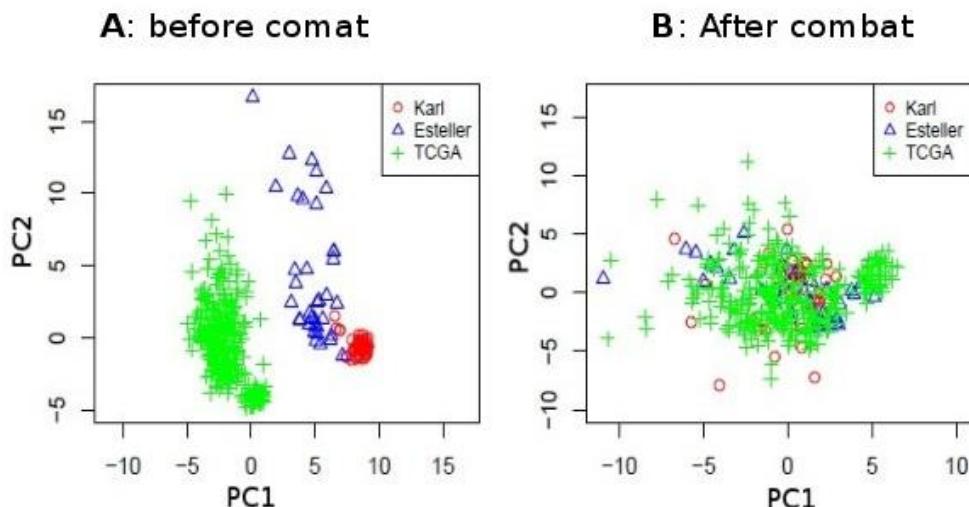


图 3-4 采用 PCA 的方法展示批次效应处理前后样本

Figure 3-4 Efficiency of the Combat treatment to eliminate the batch effect

任何对批次效应进行校正的程序，原则上只能对非生物信号进行降噪，不能使得校正前后生物或医学信号发生较大的损失。

甲基化芯片的聚类及热图分析可以用来展示各位点在肿瘤和正常样本中的甲基化特征以及样本之间的聚类关系，如图 3-5所示。图中显示了120个探针在 352 个非小细胞肺癌和106个正常肺组织中的甲基化谱式。X轴和Y轴分别表示基因和样本。其中红色样本代表肺癌样本，蓝色表示正常样本。可以看出在批次效

应校正前后，聚类分析的效果没有发生较大变化。相同类型的样本基本还是聚集在一起，生物学信息在批次效应校正前后，没有发生较大变化。

同时基因间的聚类关系也没有发生较大变化，批次效应校正前，可以观察到所有位点明显地聚集成4类基因。在批次效应校正后，依然可以观察到所有基因聚集成4类。因此，Combat算法可以在不降低生物学或医学信息的前提下，很好地消除批次效应。

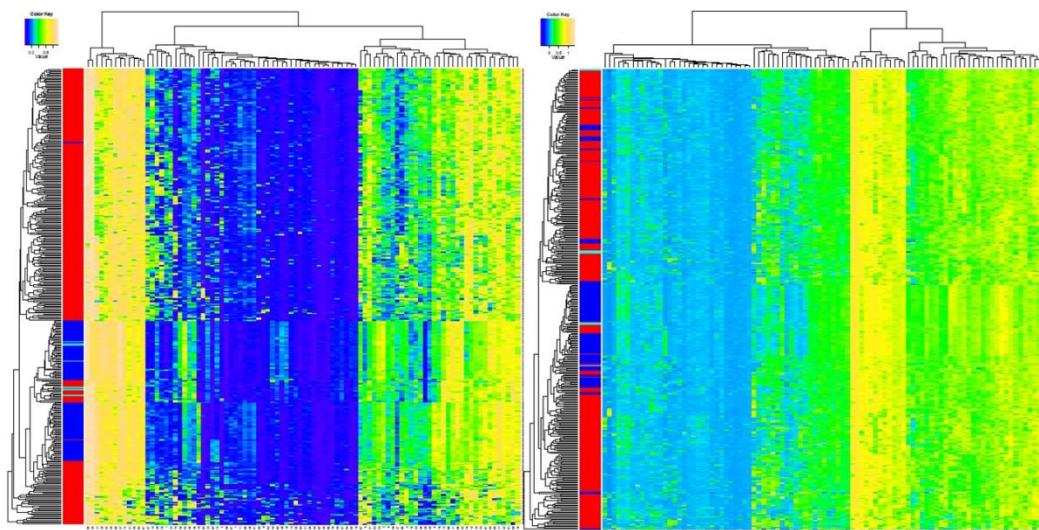


图 3-5 聚类分析展示批次效应校正前后的生物特征变化

### 3.3.3 筛选阶段最优预测变量选择

采用  $t$ -检验或 Wilcoxon 符号秩和检验对共享的 112 个 CpG 位点在肿瘤和正常组织中的差异甲基化进行检验。支持向量机 e1071 分析包自带的拟合过程启发式 Shrinking 技术进行变量选择。灵敏性，特异性和准确性用来评估的模型的预测能力，并且结合留一法交叉验证（leave-one cross-validation）评估模型的外展性。五个 CpG 位点 (*NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* 和 *ZMYND10*) 在最终被 SVM 模型选出。支持向量机预测模型在测试集中对非小细胞肺癌预测的准确度达到 98.98%。我们发现这 5 个基因均显著差异肿瘤和正常组织样本之间的甲基化。具体而言，整合后的 DNA 甲基化芯片的显示，*NTSR1* ( $P= 5.4 \times 10^{-15}$ )，*SLC5A8* ( $P= 5.9 \times 10^{-9}$ )，*GALR1* ( $P= 9.9 \times 10^{-10}$ ) 和 *AGTR1* ( $P= 6.7 \times 10^{-5}$ ) 在 NSCLC 患者显著高甲基化。而 *ZMYND10* 在 NSCLC 中显著低甲基化 ( $P= 6.2 \times 10^{-20}$ )。这些结果表明，有上述 5 个预测变量的构建的预测器对非小细胞肺癌具有很好的

预测效能。并且预测变量在癌-癌旁中都呈现差异甲基化现象，具备肺癌肿瘤标记物的基本特征。

上述五个位点在肺癌和癌旁中的差异甲基化情况如图 3-6 所示。NTSR1 在肺癌中的甲基化率为 61%，而在癌旁中的甲基化率为 32%。SLC5A8 在肺癌中的甲基化率为 27%，而在癌旁中的甲基化率为 14%。GALR1 在肺癌中的甲基化率为 47%，而在癌旁中的甲基化率为 24%。AGTR1 在肺癌中的甲基化率为 31%，而在癌旁中的甲基化率为 9%。ZMYND10 在肺癌中的甲基化率为 14%，而在癌旁中的甲基化率为 64%。五个位点均在癌组织和正常组织中均存在较大的甲基化率差异。

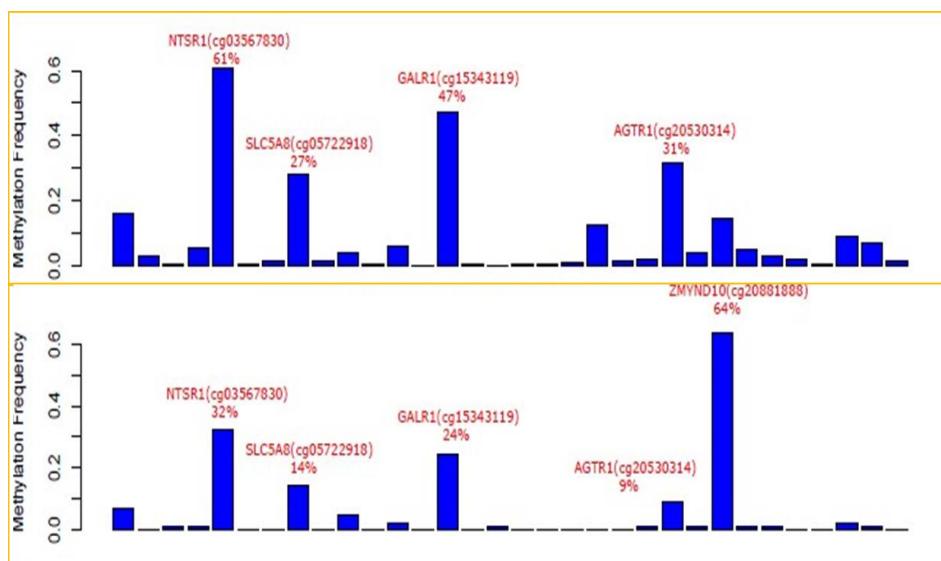


图 3-6 芯片 CpG 位点在肿瘤和正常组织中的甲基化频率比较图

图中显示了 112 个探针在 352 非小细胞肺癌和 106 正常肺组织中的甲基化频率分布情况。Beta 值小于 0.3 被定义为去甲基化状态，Beta 值大于 0.3 被定义为甲基化状态。

### 3.3.4 MSD-SNuPET 对甲基化状态的验证

为了进一步评估上述 DNA 甲基化的生物标志物对非小细胞肺癌的预测能力，我们采用 MSD-SNuPET 技术对上述五个位点在 150 对匹配的来自中国汉族人群的非小细胞肺癌和正常组织样本对上述标记物组合进行了同步检测。验证结果显示上述 5 个基因甲基化状态在 150 对非小细胞肺癌及癌旁正常组织的甲基化状态与靶点发现阶段微阵列数据的结果完全一致。这五个基因的绝对 DNA 甲基化

率在非小细胞肺癌和正常组织之间存在显著差异。如 Figure 3-7 Validation of the methylation of the candidate markers with MSD-SNuPET

图 3-7 所示。

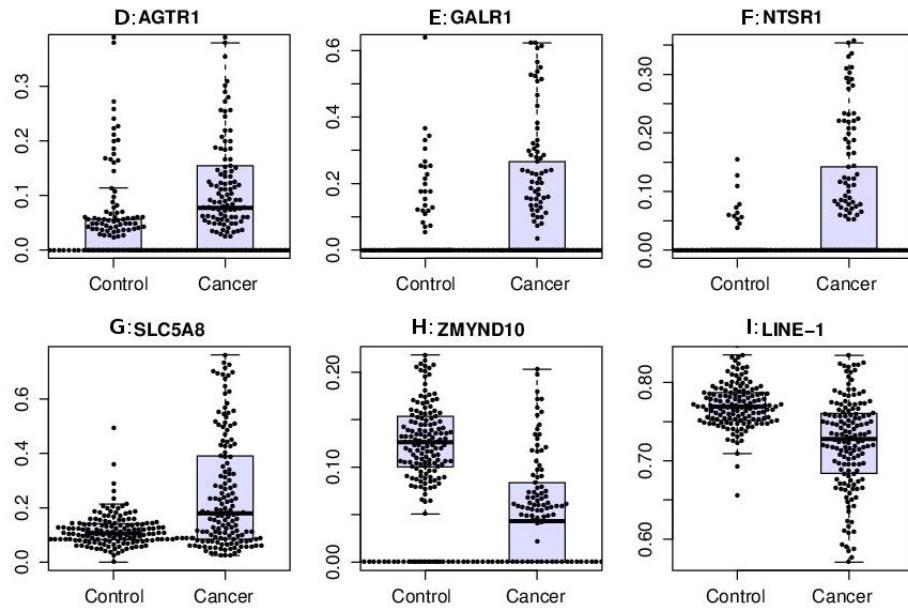


Figure 3-7 Validation of the methylation of the candidate markers with MSD-SNuPET

图 3-7 MSD-SNuPET 技术对五个甲基化标记物的甲基化状态进行检测

Figure 3-7 validation of the methylation status of the five candidate markers in an independent samples

采用 Logistic 回归分析，在校正年龄，性别和吸烟状况的情况下，*NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* 在非小细胞肺癌中呈现显著地高甲基化，P 值分别为  $5.9 \times 10^{-7}$ ,  $7.8 \times 10^{-9}$ ,  $2.3 \times 10^{-6}$ ,  $1.3 \times 10^{-6}$  而 *ZMYND10* 在 NSCLC 中呈现显著地低甲基化，P 值为  $5.2 \times 10^{-8}$  (如表 3-3 所示)。MSD-SNuPET 结果显示 *LINE-1* 的甲基化是显著降低在 NSCLC 中比正常组织 (*t*-检验， $P = 6.03 \times 10^{-14}$ )。此外，*LINE-1* 的 DNA 甲基化与性别显著相关 ( $R^2 = 0.18$ ,  $P = 0.0087$ )，这与之前的文献报道是吻合的[27, 28]，这也证明了 MSD-SNuPET 具有可靠的 DNA 检测能力。通过 logistic 回归分析，可以看出单独每个基因对非小细胞肺癌的预测能力仅仅具有适度的预测能力，灵敏度范围从 44.3% 到 73.15%；特异性从 79.59% 到 94.56%；AUC 从 0.67 到 0.80 (如表 3-3 所示)。

相关分析显示 5 个位点的甲基化状态没有显著的相关性。任何的 5 个位点的甲基化状态与年龄，吸烟，TNM 分期，肺癌分化和肺癌亚型（Ad 或 Sc）在也没有观察到显著的关联。然而，分析显示：性别和 *SLC5A8* ( $P=0.0001$ ) 及 *ZMYND10* ( $P=0.045$ ) 之间存在显著的相关性，这暗示 *SLC5A8* 和 *ZMYND10* 在 NSCLC 发生的生物学机制可能与性别有关。

从 String 9.0 蛋白质-蛋白质相互作用网络数据库显示 *NTSR1* 和 *GALR1* 存在大量的相互作用的基因。这些基因很多是癌症相关的基因，在癌症启动，进展或治疗的发挥重要角色，如 *S100A9*, *NGF*, *TAC1*, *CCK*, *FPR2*, *ADRA1B* 和 *CCL21*。String 9.0 反应的 *NTSR1* 和 *GALR1* 的基因-基因相互作用网络如图 3-8 所示。

表 3-3 肺癌差异甲基化位点

	NSCLC	Control	P-value <sup>a</sup>	$\log_{10}(\text{OR})$ (95% CI)	P-value <sup>b</sup>	Sen	Spe	AUC
<i>AGTR1</i>	12.88%	4.48%	1.06E-07	3.49 (2.08, 4.91)	1.30E-06	59.73%	79.59%	0.71
<i>GALR1</i>	18.31%	2.91%	6.58E-09	2.56 (1.5, 3.63)	2.30E-06	46.98%	85.03%	0.67
<i>NTSR1</i>	9.37%	0.56%	1.09E-09	9.02 (5.48, 12.55)	5.90E-07	44.30%	94.56%	0.70
<i>SLC5A8</i>	25.59%	11.66%	4.77E-12	3.80 (2.51, 5.09)	7.80E-09	52.35%	88.44%	0.67
<i>ZMYND10</i>	6.95%	12.82%	1.08E-07	-4.61 (-6.27, -2.95)	5.20E-08	73.15%	92.52%	0.80
<i>LINE-1</i>	72.10%	76.76%	2.39E-12	-10.3 (-13.5, -7.2)	1.80E-10	-	-	-
Reference	1.78%	1.83%	2.85E-01	-19.37 (-45.35, 6.62)	0.14	-	-	-

注：基因 DNA 甲基化与 NSCLC 的关联性采用 logistic 在校正年龄，性别，吸烟状态的情况下获得的。P-value<sup>a</sup> 和 P-value<sup>b</sup> 分别是 FDR 校正前和校正后的 P-value。Reference 位点如材料和方法部分多介绍的，为一个非 CpG 位点 C 的甲基化信号值。为了仅仅反应每个位点的预测能力，表格中的灵敏性，特异性以及 AUC 是采用 logistic 回归模型在非校正年龄，性别，吸烟状态的情况下获得的。

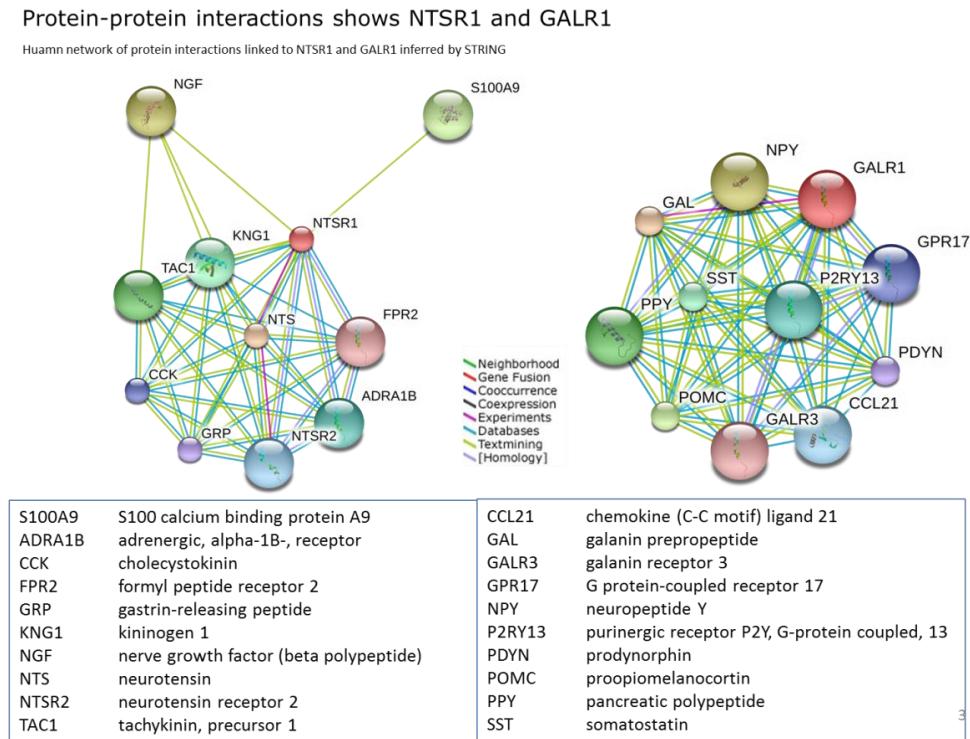


图 3-8 NTSR1 和 GALR1 相互作用的蛋白

### 3.3.5 基于 DNA 甲基化的肺癌预测模型

为了验证上述 5 个位点: *NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* 和 *ZMYND10* 对中国肺癌人群的预测能力, 基于 150 对中国人群 NSCLC -癌旁样本的四种不同的预测模型: 逻辑回归模型, 随机森林模型, 支持向量机和贝叶斯树四种方法被构建出来。五倍交叉验证用来评估模型的稳定性。预测模型中都包括年龄, 性别和吸烟状态这三个预测变量, 以增加模型实际临床应用能力。结果显示训练组和测试组在灵敏性, 特异性及准确性上没有显著的差异 (如表 3-4 所示), 说明预测模型不存在过拟合现象。此外, 四种预测模型是都具有较高的预测能力, 说明上述甲基化位点具有良好的预测潜力。具体而言, 贝叶斯树在本研究中是最强大的预测模型, 其灵敏度, 特异性和准确度分别为 86%, 96% 和 91% (如表 3-4 所示)。所有的预测模型中, Logistic 回归模型表现最差的, 即便如此, 其预测效果也十分不错, 敏感性, 特异性, 准确性和曲线下的面积 (AUC) 也可达到 78%, 97%, 87%, 和 0.906 (95% CI: 0.89-0.91), 反映出本研究中的五个甲基化位点潜在的非小细胞肺癌的诊断意义。此外, 采用贝叶斯树预测模型, 我们对吸烟和非吸烟患者, 腺癌和鳞状细胞癌患者, 早期 (I, II) 和晚期 (III, IV) 患者以及中高和差分化肺癌之间的预测能力差异进行了评估。我们发现预测模型的表现吸烟 (准确度=92.1%, 95% CI: 90.6%-93.6%) 和非吸烟 (准确度=0.939, 95%

CI: 0.935-0.943) 之间无显著差异; 在腺癌(准确度=0.82, 95% CI: 0.72-0.92) 和鳞状细胞癌(准确度=0.94, 95% CI: 0.87-0.95) 也没有显著差异; 对早期患者(准确度=0.87, 95% CI: 0.75-0.87) 和晚期患者(准确度=0.92, 95% CI: 0.82-0.92) 的预测也没有显著差异。然而分析结果显示预测模型对中高分化肺癌患者(准确度=0.9, 95% CI: 0.83-0.91) 和低分化患者(准确度=0.73, 95% CI: 0.5-0.74) 存在显著的差异( permutation test,  $P < 10^{-10}$ )。

**表 3-4 同预测模型采用五倍交叉验证的方法的表现情况**

	训练集			测试集		
	灵敏性	特异性	准确性	灵敏性	特异性	准确性
逻辑斯蒂回归	0.791	0.993	0.891	0.775	0.969	0.871
支持向量机	0.897	0.977	0.937	0.855	0.941	0.897
随机森林	0.934	0.928	0.931	0.890	0.886	0.886
贝叶斯树	0.911	0.976	0.944	0.863	0.957	0.909

### 3.4 结论

通过多平台的高通量 DNA 甲基化微阵列芯片数据的整合, 采用合适的标准和批次效应处理, 继而采用独立的临床样本进行验证可以作为一种很好的肿瘤诊断标记物开发策略。这种策略可以最大限度地利用已有的科研资料, 并且可以有效地提高标记物探索阶段的样本量, 提高肿瘤预测或诊断模型的外展性。通过预测模型的变量选择过程, 可以选择出最优的标记物组合, 结合合适的交叉验证, 从而保证预测模型的最佳预测效能并同时避免预测模型过拟合现象。本研究所开发的五个甲基化标记物: *AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1* 是一个有效的标记物组合可以实现对非小细胞肺癌诊断和预测。

### 3.5 讨论

非小细胞肺癌的早期诊断, 然后配合相应的外科手术被作为延长癌症的生存时间, 并减少非小细胞肺癌死亡率的最有效的方法。由于全局性 DNA 甲基化异常发生在肿瘤产生的早起阶段, DNA 甲基化已被认为是肿瘤诊断或早期筛查最强有力的生物标志物。在本研究中, 标记物筛选-标记物验证的两阶段生物标志物开发流程被采用, 用于开发基于 DNA 甲基化的生物标记物的非小细胞肺癌诊断模型[29]。本文从 107 个候选基因中的数据集, 采用标准化和批次效应消除处理, 开发了由五个基因 (*NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* 和 *ZMYND10*) 组成的最佳的生物标志物。然后, 采用一种多位点的 DNA 甲基化检测方法: MSD -

SNuPET，对该生物标记物组合在一个由 150 对中国汉族非小细胞肺癌和正常组织组成的回顾性队列中进行了绝对定量甲基化水平的检测。最后多个预测模型显示，这个最优的甲基化标记物组合具有很高的潜在应用价值。

经过文献检索发现，这五个生物标志物已在癌症研究中有很多报道。神经降压素受体 1 (*NTSR1*) 是一种 G 蛋白偶联受体 (GPCR)，它已被广泛地报道与肿瘤的发生，发展[30]和预后[31, 32]相关联。有证据显示，*NTSR1* 可以作为癌症的进展的生物标志物[33]这个和我们目前研究的结果相吻合。甘丙肽受体亚型 2 (*GALR1*) 可以抑制多种癌症如头颈部[34, 35]，口腔鳞状细胞癌[36]的细胞增殖。*GALR1* 基因表达失活可以通过启动子甲基化[34]引起。同时 *GALR1* 是乳腺癌的一个亚型决定基因。这表明在 *GALR1* 在癌症诊断的具有巨大的应用潜力。溶质载体家族 5 原件 8 (*SLC5A8*) 是一种肿瘤抑制基因，可以抑制结肠癌，胃癌等肿瘤的进展[37-39]。*ZMYND10* 由于错义突变及其在肺癌表达的损失的发生最近被鉴定为候选肿瘤抑制基因。

本研究验证阶段的分析是基于定量的 DNA 甲基化信号。我们还进行了按照特别规则离散化后数据的分析，比如  $\beta$  值  $<0.3$  定义为去甲基化， $\beta$  值  $> 0.8$  被定义甲基化状态， $\beta$  值在 0.3-0.8 之间是定义为半甲基化 CpG 位点[40, 41]。在这种情况下，5 个基因仍非小细胞肺癌组织和正常组织之间的显著差异性甲基。在预测模型的敏感性，特异性和准确性方面也没有显著的变化。此外，我们发现如果采用逐个加入基因的方式，诊断模型的敏感性，特异性，准确性和 AUC 也逐渐增加。如图 3-9 所示。

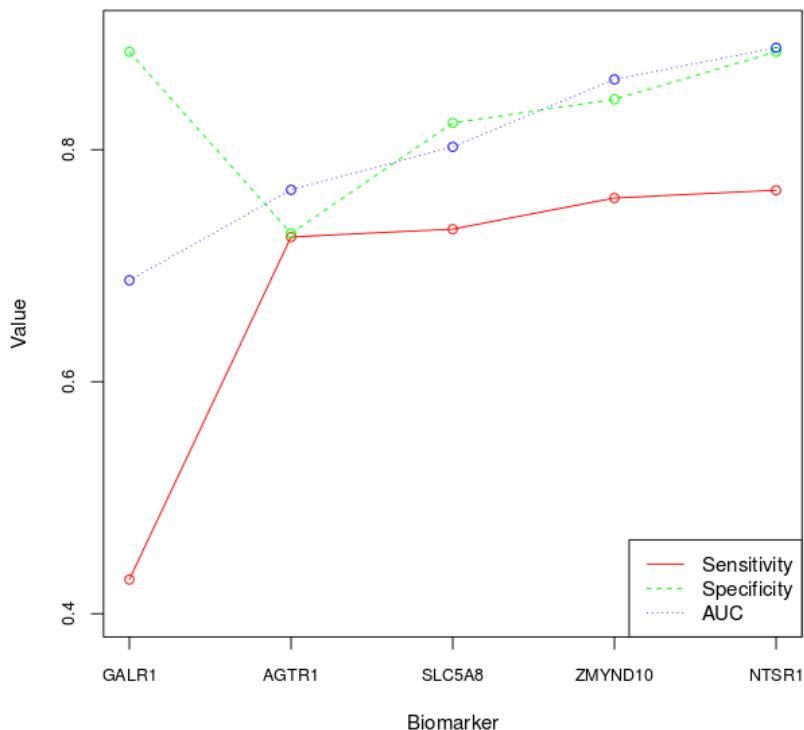


图 3-9 诊断模型随预测变量的增加的表现情况

肺癌的诊断是一个具有挑战性的问题。为了开发可用于非小细胞肺癌诊断的 DNA 甲基化的生物标志物，我们应该在全基因组范围内搜索数十或数百个基因位点的最优组合。跨多个平台的全基因组 DNA 甲基化芯片数据配合数据标准化和批次效应消除可以提供一个较大的生物标记物筛选基础。用这种方法，我们确定了 5 基因签名包括 *AGTR1*、*GALR1*、*SLC5A8*、*ZMYND10* 和 *NTSR1*，可提供高诊断的敏感性和特异性。

虽然 DNA 甲基化具有非常有前途的肿瘤诊断，筛查，预后监控能力，并且具有肿瘤特异性，最有可能成为肿瘤诊断的分子生物标记物，然而目前已有的检测技术对操作者的要求较高，流程仍然较为复杂，对 DNA 量的需求较多，只适合在科研阶段使用，开发适合临床使用的便捷、快速、重复率高、DNA 需求小的检测技术是目前 DNA 甲基化临床研究的重要任务之一。美国在这方面成果比较突出，Johns Hopkins University 目前开发了一种被称为“methylation on beads”的技术，可以实现血浆游离 DNA 甲基化状态的便捷检测，具有极高的临床应用价值[42]。

尽管如此本研究也有一些缺陷。比如对于整合的两个平台，其探针数差异较大，GoldenGate 包括 371 个基因位点的 1,536 个 CpG 位点，而 HM27 芯片

覆盖 14,495 个基因和 27578 位点。两者共享的探针数只有 112 个，大大降低了变量选择的初始选择域。随着甲基化芯片技术价格的降低，目前 HM27K 和 HM450K 芯片已经逐渐普及，届时 HM27K 和 HM450K 芯片的整合会更加显著地提高肿瘤诊断的灵敏性，特异性及准确度。

### 3.6 参考文献

- [1] Siegel, R., D. Naishadham, and A. Jemal. *Cancer statistics, 2012* [J]. CA Cancer J Clin, 2012. **62**(1);10-29.
- [2] Hankey, B.F., L.A. Ries, and B.K. Edwards. *The surveillance, epidemiology, and end results program: a national resource* [J]. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 1999. **8**(12);1117-1121.
- [3] Li, H., L. Yang, X. Zhao, J. Wang, et al. *Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model* [J]. BMC medical genetics, 2012. **13**;118.
- [4] Guo, S., Y.L. Wang, Y. Li, L. Jin, et al. *Significant SNPs have limited prediction ability for thyroid cancer* [J]. CANCER MEDICINE, 2014.
- [5] He, Y., Y. Cui, W. Wang, J. Gu, et al. *Hypomethylation of the hsa-miR-191 locus causes high expression of hsa-mir-191 and promotes the epithelial-to-mesenchymal transition in hepatocellular carcinoma* [J]. Neoplasia, 2011. **13**(9);841-853.
- [6] Flores, K., F. Wolschin, J.J. Corneveaux, A.N. Allen, et al. *Genome-wide association between DNA methylation and alternative splicing in an invertebrate* [J]. BMC Genomics, 2012. **13**;480.
- [7] Jiang, F., N.W. Todd, R. Li, H. Zhang, et al. *A panel of sputum-based genomic marker for early detection of lung cancer* [J]. Cancer prevention research, 2010. **3**(12);1571-1578.
- [8] Zhu, J. and X. Yao. *Use of DNA methylation for cancer detection: promises and challenges* [J]. The international journal of biochemistry & cell biology, 2009. **41**(1);147-154.
- [9] Laird, P.W. *The power and the promise of DNA methylation markers* [J]. Nature reviews. Cancer, 2003. **3**(4);253-266.
- [10] Zhao, Y., H. Zhou, K. Ma, J. Sun, et al. *Abnormal methylation of seven genes and their associations with clinical characteristics in early stage non-small cell lung cancer* [J]. Oncology letters, 2013. **5**(4);1211-1218.

- [11] Anglim, P.P., T.A. Alonzo, and I.A. Laird-Offringa. *DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update* [J]. Molecular cancer, 2008. **7**:81.
- [12] Nikolaidis, G., O.Y. Raji, S. Markopoulou, J.R. Gosney, et al. *DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer* [J]. Cancer research, 2012. **72**(22):5692-5701.
- [13] Edgar, R., M. Domrachev, and A.E. Lash. *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository* [J]. Nucleic Acids Research, 2002. **30**(1):207-210.
- [14] Chen, C., K. Grennan, J. Badner, D. Zhang, et al. *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods* [J]. PLoS One, 2011. **6**(2):e17238.
- [15] Gibbs, A.R. and F.B. Thunnissen. *Histological typing of lung and pleural tumours: third edition* [J]. Journal of clinical pathology, 2001. **54**(7):498-499.
- [16] Edge, S.B. and C.C. Compton. *The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM* [J]. Annals of surgical oncology, 2010. **17**(6):1471-1474.
- [17] Zhao, Y., S. Guo, J. Sun, Z. Huang, et al. *Methylcap-seq reveals novel DNA methylation markers for the diagnosis and recurrence prediction of bladder cancer in a Chinese population* [J]. PLoS One, 2012. **7**(4):e35175.
- [18] Wang, X., L. Wang, S. Guo, Y. Bao, et al. *Hypermethylation reduces expression of tumor-suppressor PLZF and regulates proliferation and apoptosis in non-small-cell lung cancers* [J]. FASEB J, 2013. **27**(10):4194-4203.
- [19] Ding, Q., Y. Hu, S. Xu, C.C. Wang, et al. *Neanderthal Origin of the Haplotypes Carrying the Functional Variant Val92Met in the MC1R in Modern Humans* [J]. Molecular biology and evolution, 2014. **31**(8):1994-2003.
- [20] Cai, X.Y., X.F. Wang, S.L. Li, J. Qian, et al. *Association of mitochondrial DNA haplogroups with exceptional longevity in a Chinese population* [J]. PLoS One, 2009. **4**(7):e6423.
- [21] 文波, *Y 染色体、mtDNA 多态性与东亚人群的遗传结构*[D], 2004, 复旦大学: 上海.
- [22] Wang, Y.L., S.H. Feng, S.C. Guo, W.J. Wei, et al. *Confirmation of papillary thyroid cancer susceptibility loci identified by genome-wide association studies of chromosomes 14q13, 9q22, 2q35 and 8p12 in a Chinese population* [J]. J Med Genet, 2013. **50**(10):689-695.
- [23] Dessaix, R.B. and C.B. Pipper. *["R"--project for statistical computing]* [J]. Ugeskr Laeger, 2008. **170**(5):328-330.
- [24] Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, et al. *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored* [J]. Nucleic acids research, 2011. **39**(Database issue):D561-568.
- [25] Johnson, W.E., C. Li, and A. Rabinovic. *Adjusting batch effects in microarray expression data using empirical Bayes methods* [J]. Biostatistics, 2007. **8**(1):118-127.
- [26] 陈超, *基因表达谱芯片校正批次效应算法的比较及网络分析在精神分裂症研究中的应用*[D], 2011, 复旦大学: 上海.

- [27] El-Maarri, O., M. Walier, F. Behne, J. van Uum, et al. *Methylation at global LINE-1 repeats in human blood are affected by gender but not by age or natural hormone cycles* [J]. PloS one, 2011. **6**(1);e16252.
- [28] El-Maarri, O., T. Becker, J. Junen, S.S. Manzoor, et al. *Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males* [J]. Human genetics, 2007. **122**(5);505-514.
- [29] Tsou, J.A., J.A. Hagen, C.L. Carpenter, and I.A. Laird-Offringa. *DNA methylation analysis: a powerful new tool for lung cancer diagnosis* [J]. Oncogene, 2002. **21**(35);5450-5461.
- [30] Heakal, Y., M.P. Woll, T. Fox, K. Seaton, et al. *Neurotensin receptor-1 inducible palmitoylation is required for efficient receptor-mediated mitogenic-signaling within structured membrane microdomains* [J]. Cancer Biology & Therapy, 2011. **12**(5);427-435.
- [31] Valerie, N.C., E.V. Casarez, J.O. Dasilva, M.E. Dunlap-Brown, et al. *Inhibition of neurotensin receptor 1 selectively sensitizes prostate cancer to ionizing radiation* [J]. Cancer Research, 2011. **71**(21);6817-6826.
- [32] Alifano, M., F. Souaze, S. Dupouy, S. Camilleri-Broet, et al. *Neurotensin receptor 1 determines the outcome of non-small cell lung cancer* [J]. Clinical Cancer Research, 2010. **16**(17);4401-4410.
- [33] Dupouy, S., N. Mourra, V.K. Doan, A. Gompel, et al. *The potential use of the neurotensin high affinity receptor 1 as a biomarker for cancer progression and as a component of personalized medicine in selective cancers* [J]. Biochimie, 2011. **93**(9);1369-1378.
- [34] Misawa, K., Y. Ueda, T. Kanazawa, Y. Misawa, et al. *Epigenetic inactivation of galanin receptor 1 in head and neck cancer* [J]. Clinical Cancer Research, 2008. **14**(23);7604-7613.
- [35] Kanazawa, T., P.K. Kommareddi, T. Iwashita, B. Kumar, et al. *Galanin receptor subtype 2 suppresses cell proliferation and induces apoptosis in p53 mutant head and neck cancer cells* [J]. Clinical Cancer Research, 2009. **15**(7);2222-2230.
- [36] Henson, B.S., R.R. Neubig, I. Jang, T. Ogawa, et al. *Galanin receptor 1 has anti-proliferative effects in oral squamous cell carcinoma* [J]. Journal of Biological Chemistry, 2005. **280**(24);22564-22571.
- [37] Park, J.Y., J.F. Helm, W. Zheng, Q.P. Ly, et al. *Silencing of the candidate tumor suppressor gene solute carrier family 5 member 8 (SLC5A8) in human pancreatic cancer* [J]. Pancreas, 2008. **36**(4);e32-39.
- [38] Ueno, M., M. Toyota, K. Akino, H. Suzuki, et al. *Aberrant methylation and histone deacetylation associated with silencing of SLC5A8 in gastric cancer* [J]. Tumour Biology, 2004. **25**(3);134-140.
- [39] Miyauchi, S., E. Gopal, Y.J. Fei, and V. Ganapathy. *Functional identification of SLC5A8, a tumor suppressor down-regulated in colon cancer, as a Na(+) -coupled transporter for short-chain fatty acids* [J]. Journal of Biological Chemistry, 2004. **279**(14);13293-13296.
- [40] Ron-Bigger, S., O. Bar-Nur, S. Isaac, M. Bocker, et al. *Aberrant epigenetic silencing of tumor suppressor genes is reversed by direct reprogramming* [J]. Stem Cells, 2010. **28**(8);1349-1354.

[41] Richter, J., O. Ammerpohl, J.I. Martin-Subero, M. Montesinos-Rongen, et al. *Array-based DNA methylation profiling of primary lymphomas of the central nervous system* [J]. Bmc Cancer, 2009. **9**:455.

[42] Guzzetta, A.A., T.R. Pisanic Li, P. Sharma, J.M. Yi, et al. *The promise of methylation on beads for cancer detection and treatment* [J]. Expert review of molecular diagnostics, 2014. **14**(7):845-852.

## 第四章 全基因组甲基化图谱揭示泛癌甲基化特征

在上个世纪八十年代，肿瘤基因组整体水平的去甲基化和局部区域的高甲基化模型被建立[1]。近半个世纪以来，表观遗传学工作者陆续发现了大量的由于启动子局部区域高甲基化造成基因表达沉默而形成的抑癌基因如 MGMT, DAPK, APC, CDH1, CDH13, GSTP1, CDKN2A 等以及启动子区域异常去甲基化造成高表达的癌基因如：WNT5A, CRIP1, S100P 等。虽然早在 2008 年全基因组甲基化单碱基测序技术(Single base-resolution methylome)已经获得突破，但是研究成本之高，使得目前的科研只能局限在对模式生物进行极少数个体的研究，比较具有代表性的如：2008 年美国科学家率先完成模式生物拟南芥的全基因组甲基化图谱绘制[2]。2010 年由中国科学家完成的人类外周血单核细胞的单碱基甲基化谱式（炎黄甲基化项目）[3]和家蚕丝腺全基因组单碱基甲基化谱式[4]。这种高质量的单碱基甲基化图谱对于基因组，表观基因组研究具有重要的价值。但是由于成本巨大，很难采用这种技术对群体遗传学，群体表观遗传学及临床医学的问题进行阐述。这样就造成了在群体水平上肿瘤基因组甲基化的精细全貌仍然比较粗糙。对于群体遗传及表观遗传学家来说，比较有优势的两种方法包括基因高密度甲基化芯片技术的甲基化精细谱式 (HM450K 芯片技术) 和基于捕获测序的精细谱式 (MethylCap-seq [5], MeDIP-seq [6], RRBS [7] 的技术）。

高密度甲基化芯片技术可以实现较大群体的甲基化谱式建立，目前比较突出的研究包括 TCGA 中 1,0681 个肿瘤及其对照样本的甲基化芯片数据[8]。由于 DNA 甲基化受到性别，年龄，环境交互等的强烈作用[9]，目前尚没有大规模的正常群体的 DNA 甲基化芯片数据，然而值得注意的是已经开始有基于 DNA 甲基化的 mGWAS 研究被陆续开展和报道如：寿命和衰老相关的全基因组表观遗传学关联研究[10]，此外 Rakyan 及其同事也对全基因组表观遗传学关联研究的理论基础进行了深入的探讨[11]，可以预期在未来的五年内，全基因组表观遗传学关联研究将遗传学及医学界大量出现。此外，全基因组水平评估多种肿瘤相似性的研究由于受到数据难以获得等因素，一直没有太多的研究对这些问题进行揭示。因此本部分旨在借助 TCGA 计划中配对肿瘤样本的全基因组甲基化芯片数据，对全基因组甲基化谱式的一系列细致的描述。并对多种肿瘤 DNA 甲基化层面的相似性及区别进行分析，同时采用随机森林预测模型的方法对基于 DNA 甲基化同时对多种肿瘤组织(泛癌组织)和正常样本进行预测的准确度进行了评估。

## 4.1 背景

对肿瘤的早期筛查，诊断或辅助诊断而言，固然可以选择利用粪便脱落细胞 DNA 甲基化对肠癌进行筛查[12]，利用支气管灌洗液 DNA 甲基化对肺癌进行筛查[13]，利用尿沉淀 DNA 甲基化对膀胱癌进行筛查[5]。但是这会增加肿瘤筛查的操作复杂性及成本，降低了 DNA 甲基化对癌症筛查的应用价值。采用血浆游离 DNA 甲基化对肿瘤进行筛查，在建立有效的筛查模型的情况下，可以一次性对所有肿瘤进行筛查，大大降低了肿瘤筛查成本，并且可以显著地降低筛查的耗时。同时对多种肿瘤进行的比较或同步分析被称为泛癌分析，而含有多种肿瘤变异信息的数据被称为泛癌数据，同时对多种肿瘤进行同步诊断可以称为泛癌诊断。虽然目前很难同时得到大量基于血浆游离 DNA 的泛癌甲基化数据。但目前可以收集到大量的实体组织的全基因组 DNA 甲基化泛癌数据。在基于实体中的全基因组甲基化谱式可以一定程度地释放到血浆中，所以可以采用实体组织的全基因组甲基化数据进行初步的分析，探讨泛癌诊断的一系列问题。

本研究收集了来自 TCGA (The Cancer Genome Atlas) 计划中 11 种肿瘤类型配对的 1274 个癌和癌旁样本的全基因组 DNA 甲基化 HM450K 芯片数据。这 11 种肿瘤包括：肾透明细胞癌、浸润性乳腺癌、甲状腺癌、头颈部鳞状细胞癌、前列腺癌、肝癌、肾透明细胞癌、肺腺癌、结肠癌、子宫内膜癌、肺鳞癌。这 11 种肿瘤涵盖了男性和女性中发病率及死亡率最高的 4 种肿瘤（如图 4-1 所示）。选择配对的样本是为了充分校正掉年龄、性别、种族等环境因素对结果的干扰。

对于多种肿瘤之间的相似性，很早就有学者提出这方面的问题，2014 年 Risbridger 博士就曾经发表相应的文章论述乳腺癌和前列腺癌的相似性远大于差异性[14]。本部分研究通过主成分分析，多维尺度分析，个体水平聚类分析，群体水平聚类分析等手段，对泛癌数据进行相似性分析，以尝试对肿瘤肿瘤相似性进行定量的估计，为临床科研提供一定的思路和借鉴。此外，目前的肿瘤诊断越来越倾向于无创的分子诊断技术。而基于血浆的肿瘤诊断被认为是最有希望的无创诊断方法，2014 年日本也斥资 79 亿日元启动了一项以血浆为介质同时诊断 13 种肿瘤的大型项目（<http://the-japan-news.com/news/article/0001506220>）。说明了科研界对多种肿瘤进行同时诊断的热情和信心。此外这种诊断技术和模型不仅可以实现对肿瘤的诊断，甚至还可以同时对多种复杂疾病进行早期筛查和风险预警。本部分研究采用随机森林预测模型探讨了利用全基因组甲基化数据对多种肿瘤和正常样本进行预测和判定的准确性评估。同时对泛癌预测模型最佳的预测位点个数进行了估计。

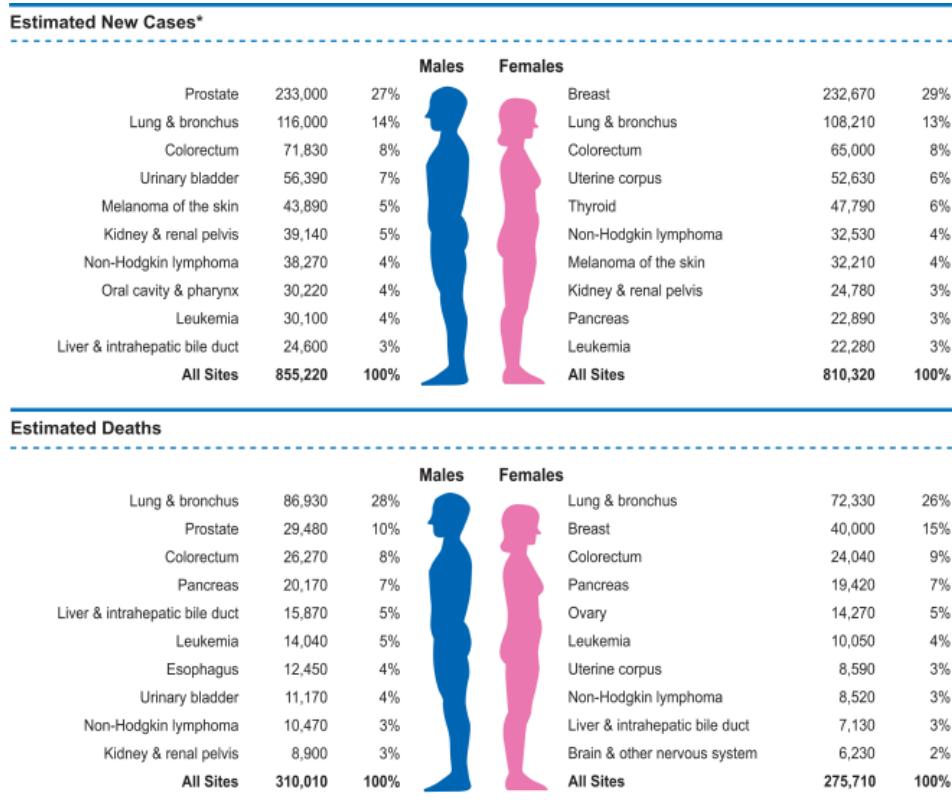


图 4-1 2014 美国肿瘤统计：发病率和死亡率前 10 名

本研究重点对 HM450K 芯片在肿瘤组织及正常组织中表现的各种特征进行分析。主要包括 HM450K 芯片位点在甲基化水平的相关性分布，肿瘤组织及正常组织全基因 CpG 位点甲基化的分布特征，基因组中 CpG 位点甲基化的连锁规律，不同肿瘤及正常样本之间的相关性，肿瘤相关差异甲基化位点，肿瘤共享异常甲基化位点及肿瘤特异甲基化位点，及采用甲基化位点对多种肿瘤进行同时预测的精度分析。最后对三种对于早期诊断具有很高需求的肺癌进行了具体的分析，对其中的差异甲基化位点，异常信号通路，肿瘤分期相关甲基化位点，主要风险因素相关甲基化位点等进行了展开。

## 4.2 材料和方法

### 4.2.1 数据

所有全基因组 DNA 甲基化数据均来自于 TCGA 计划中的甲基化芯片 HM450K 平台，11 种肿瘤 1274 个样本的全基因组甲基化数据，如表 4-所示。所有样本均为配对样本，共有 637 个肿瘤样本和 637 个对应的正常样本。HM450K 甲基化芯片上覆盖有超过 485,000 个甲基化位点。

表 4-1 配对的癌-癌旁 DNA 甲基化全基因数据信息

Symbol	Case	Control	Cancer	肿瘤名称
KIRC	160	160	Kidney renal clear cell carcinoma	肾透明细胞癌
BRCA	92	92	Breast invasive carcinoma	浸润性乳腺癌
THCA	56	56	Thyroid carcinoma	甲状腺癌
HNSC	50	50	Head and Neck squamous cell carcinoma	头颈部鳞状细胞癌
PRAD	49	49	Prostate adenocarcinoma	前列腺癌
LIHC	49	49	Liver hepatocellular carcinoma	肝癌
KIRP	45	45	Kidney renal papillary cell carcinoma	肾乳头状细胞癌
LUSC	41	41	Lung squamous cell carcinoma	肺腺癌
COAD	39	39	Colon adenocarcinoma	结肠癌
UCEC	30	30	Uterine Corpus Endometrial carcinoma	子宫内膜癌
LUAD	26	26	Lung adenocarcinoma	肺鳞癌

### 4.2.2 方法

#### 4.2.2.1 数据清理（Data Cleaning）

在本研究中，没有对含有 SNP 的甲基化探针进行了排除，以避免大量位点删除对结果造成的偏差（187,468 探针区域含有 SNP）。仅对于在所有样本中缺失值超过 30% 的探针，进行排除。其余缺失值采用 KNN 算法进行缺失值填补（R, impute 分析包）。

#### 4.2.2.2 变量相关系数抽样分布（Variable Correlation Distribution）

根据我们的经验，经过数据清理后的 HM450K 芯片数据，一般至少会保留 349,049 个变量，对这些变量的相关性分析，可以让我们对全基因组甲基化数据有个较为整理的理解。但是对 349,049 变量两两相关性分析计算量十分巨大。本研究采用抽样的方法对总体分布进行估计。采用 2,000 次随机抽样，每次抽 1,000 个变量的方法。然后对抽样本的相关系数进行相应的统计分析。

### 4.2.2.3 差异甲基化位点分析（Differential Methylation Loci Detection）

在数据清理的基础上，对芯片数据进行整体的 quantile Normalization 之后，采用 Wilcoxon signed rank test 检验对甲基化水平差异进行检验；对于非配对数据采用 Wilcoxon rank sum 检验对甲基化水平差异进行检验。分析采用双尾检验及 95% 置信区间等常规设置。

### 4.2.2.4 随机森林和变量重要性（Random Forest and Variable Importance）

高维基因组数据集，特别是以新一代高密度芯片技术和测序技术产生的全基因组数据，给传统的机器学习模型带来了巨大的挑战。高维数据同时极大地增加了变量之间的相关性及互作性，这也使得传统很多统计模型的独立不相关变量的假设不再成立，严重降低了传统统计模型的统计能力。Random forests (RF) 是由 Leo Breiman (2001) 提出一种高效的基于 tree 的机器学习方法。其综合了 Bootstrap aggregating 和 random subspace method 两种思想。RF 具有对高维高维不敏感，并且适用于变量相关或互作的数据集，随机森林可以既可以处理属性为离散值的量 (SNP)，也可以处理属性为连续值的量 (RNA, DNA 甲基化, miRNA)，为此 RF 在处理全基因组数据方面具有极高的优势。此外，样本和变量双随机的引入使得随机森林预测模型不容易陷入过拟合，同时也提高了其抗噪音能力。它能够处理高维数据，并且不用做预先得变量选择。在创建随机森林的时候，对 generalization error 使用的是无偏估计。训练速度快，可以得到。

随机森林通过 bootstrap 重采样技术，从原始数据集中有放回地重复随机抽取  $K$  ( $K <$  总样本量) 个样本生成新的训练样本数据集合。此外，随机森林算法同时也对变量 (variable) 也进行了抽样，只选取  $V$  个变量 ( $V <$  数据总维度) 进入新的训练样本集合。根据随机抽取的数据集合 (包含  $K$  个对样本和  $V$  个变量) 作为训练样本集合，生成  $K$  个分类树从而组成随机森林。对于新数据或新的 input 个体的分类结果按分类树的投票而定。随机森林算法根本上是对决策树算法的一种改进，其将多个决策树综合在一起，每棵树的训练依赖于一个独立抽取的样品，森林中的每棵树理论上具有相同的分布，因此分类误差决定于每一棵树的分类能力和它们之间的相关性。由于抽样的原因，会有一些未被抽中的样本，这形成了所谓的袋外数据 (OOB, out of bag)，可以根据 OOB 来估计

泛化误差，而不需要用交叉检验来估计。这样就意味着随机森林内嵌了 train 和 test 的过程，不需要进行外部的交叉验证。

随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元共线性不敏感，对缺失数据和非平衡的数据比较稳健。对多达几千个解释变量的作用，也可以很好地预测，显示出了其巨大的优势。基于信息增益（OOB）误分率的增加量和基于分裂时的基尼指数（GINI）下降量，随机森林可以对变量重要性进行评估，以确定每个变量在分类中的贡献度。此外，在训练过程中，能够检测到变量间的互相影响。随机森林的原理优越性促使其可以广泛应用于生物医学的各个研究领域，如疾病诊断模型建立，Biomarker 筛选，回归模型建立，遗传多态性关联研究，信号通路搜索，变量选择，上位效应检测。

#### 4.2.2.5 非特异变量过滤（Non-Specific Filtering）

RF 算法对于中小型数据集的预算时间基本可以忽略不计，但是对于高维数据，其 CPU 耗时还是非常可观的。比如对于本研究而言，1,274 个样本 12,000 变量的数据集的随机森林建模过程（1500 个子树）在 Intel(R) Xeon(R) CPU E7-4870 (Clock Speed = 2.40GHz) 需要的时间分别为 27.5 小时。非特异变量过滤（non-specific filtering）是不依赖分类信息的情况下，采用方差，变异系数等指标对变量进行筛查的过程，比如对于方差最小的 20% 的变量可以考虑不纳入到预测模型的建立中，因为方差较小，这些变量对分类过程的贡献较小，因此对预测模型的准确度影响不大，为此可以应用于 RF 分类的数据预处理阶段。

#### 4.2.2.6 特异变量过滤（Specific Variable Filtering）

特异性变量过滤考虑到分类信息，如果对于二分类型数据，可以只选取在两类样本中具有显著差异或者差异大于某个临界值的变量，而过滤掉不显著或者差异较小的变量。对于多分类数据，可以选择具有显著类别特异的变量，比如采用类别特异指数（group specificity index，GSI）过滤掉 GSI 较小的变量。

$$GSI = \frac{\sum_{j=1}^n 1 - \frac{\log_2(s(j))}{\log_2(S_{max})}}{n - 1}$$

其中 n 表示类别的个数， $s(j)$  是第 j 个类别的平均值， $S_{max}$  是平均值最高的类别对应的均值。更多地资料可以参考 Yanai 及其同事 2005 发表在 Bioinformatics 上的文章[15]和靳文菲博士的学位论文的第四章[16]。

#### 4.2.2.7 分段组合法 (Segmentation and Recombination)

非特异变量过滤或特异变量过滤法由于对变量进行了选择从而导致无法对所有 CpG 位点在分类中的重要性进行评估，为了避免这种情况，我们可以采用分段组合发进行随机森林建模。将变量分为多个组，每组可以将变量控制在 1000 个左右。1,274 个样本 5,000 变量的数据集的随机森林建模过程（1500 个子树）在 Intel(R) Xeon(R) CPU E7- 4870 (Clock Speed = 2.40GHz)需要的时间分别为 0.14 秒。然后选出每组中的最重要的 20 个变量，然后将 350 组的重要变量合并，然后对新生成的位点进行随机森林建模。采用这种方法可以对每个变量的分类重要性进行评估。

#### 4.2.2.8 主成分分析 (Principal Components Analysis, PCA)

主成分分析所要做的就是设法将原来众多具有一定相关性的变量，重新组合为一组新的相互无关的综合变量来代替原来变量。

对于一个  $n \times p$  的数据矩阵，表示该数据包括  $n$  个个体，每个个体具有  $p$  个变量，以此为  $x_1 \dots x_p$ ， $X$  表示该数据的矩阵表示。则：

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = (x_1, x_2, \dots, x_p)$$

其中：  $x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}, \quad j=1,2,\dots,p$

主成分分析就是将  $p$  个观测变量综合成为  $p$  个新的变量（综合变量），即

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \cdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases}$$

简写为：

$$F_j = \alpha_{j1}x_1 + \alpha_{j2}x_2 + \cdots + \alpha_{jp}x_p$$

$j=1,2,\dots,p$

要求模型满足以下条件：

①  $F_i, F_j$  互不相关 ( $i \neq j$ ,  $i, j = 1, 2, \dots, p$ )

②  $F_1$  的方差大于  $F_2$  的方差大于  $F_3$  的方差，依次类推

③  $a_{k1}^2 + a_{k2}^2 + \dots + a_{kp}^2 = 1$   $k = 1, 2, \dots, p$ .

此时， $F_1$  成为  $X$  矩阵的第一主成分， $F_2$  为  $X$  矩阵的第二主成分，依此类推，

有第  $p$  个主成分。主成分又叫主分量。这里  $a_{ij}$  我们称为主成分系数。

上述模型可用矩阵表示为：

$$F = AX, \text{ 其中}$$

$$F = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{pmatrix} \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

$A$  称为主成分系数矩阵。

第一主成分可以表示出  $X$  数据集最主要的方差来源。通过对前几个主成分特性的分析，可以在低维空间了解原数据集的特征属性。更多对 PCA 的讨论可以参考 Laster[17] 及 Vander[18] 对 PCA 的讨论。

#### 4.2.2.9 多维尺度分析 (Multidimensional Scaling, MDS)

多维尺度分析(multidimensional scaling, MDS) 是分析研究对象之间相似性或差异性的一种多元统计分析方法。采用 MDS 可以在低维空间(二维或三维)内对研究对象之间的距离或者相似性进行展示。MDS 利用的是成对样本间相似性，目的是利用这个信息去构建合适的低维空间，使得样本在此空间的距离和在高维空间中的样本间的相似性尽可能的保持一致。多维尺度分析和因子分析及聚类分析都有一些相似性，但也存在较大差别。多维标度仅仅需要相似

性或者距离，而不需要相关性（因子分析需要相关性）。聚类分析把观测到的特征当作分组标准，而多维标度仅仅取用感知到的差异。多维尺度分析根据数据变量形式为连续变量或类别变量，可以分为公制多维尺度分析和非公制多维尺度分析。对于非公制多维尺度分析，徐书华博士的学位论文中成功采用这种研究方法对基于基因型（genotype）的数据对人群距离（差异）进行分析，具体原理可参考其博士论文[19]中的介绍。本文中所涉及的为公制多维尺度分析，其表示如下：

假设我们有一个  $n \times p$  的数据矩阵，表示该数据包括  $n$  个个体，每个个体具有  $p$  个变量，以此为  $x_1 \dots x_p$ ， $X$  表示该数据的矩阵表示。那么  $n$  个样本两两之间的距离可以形成  $n \times n$  的距离矩阵 ( $D^X$ )。

$$D = \begin{pmatrix} \delta_{1,1} & \dots & \delta_{1,n} \\ \vdots & \ddots & \vdots \\ \delta_{n,1} & \dots & \delta_{n,n} \end{pmatrix}$$

$\delta_{i,j}$  表示  $i, j$  个体之间的距离。多维尺度分析目的就是在  $d$  维空间内 ( $d \ll p$ ) 找到  $n$  个向量  $u_1, u_2 \dots u_n$ ，使得下面的目标函数成立：

$$\min_u \sum_{n=1}^n \sum_{n=1}^n d_{i,j}^{(X)} - d_{i,j}^{(U)}$$

其中  $d_{i,j}^{(X)} = |x_i - x_j|$ ,  $d_{i,j}^{(U)} = |u_i - u_j|$ . 因此 MDS 把问题转变为在给定  $d$  的情况下上述目标函数的最优化问题。一般而言对于  $d = 2$  及  $d = 3$  便于从二维及三维对样本相似性进行观察，随着  $d$  的增大，新的  $U$  向量显然更加逼近原始向量，但也就失去了 MDS 的目的。因此在实际情况是  $d$  一般远远小于  $p$ 。更多对于多维尺度分析的介绍可以参考 Beals[20] 及 Carroll 对的 MDS 的著作[21]。

此外，决策树的理念是将数据归入不同的组中，那么同一组中的样本可以认为是比较相似的。根据这个思路可以建立起各样本间的相似矩阵。用 1-相似矩阵则可以认为是一种“距离”，利用距离就可以进行异常值检验或聚类分析。本研究中的多维尺度分析，正是选择了由随机森林模型的相似矩阵转变而来的距离矩阵。

## 4.3 结果

### 4.3.1 全基因组甲基化位点相关系数分布

对 HM450K 芯片的位点进行充分的了解有利于我们对基于这种数据对全基因组甲基化数据分析结果的解释。经过数据整合, 我们获得了一个维度为 349,049 (变量个数)  $\times$  1278 (样本量) 的高维矩阵。变量相关系数矩阵的分布可以为我们提供对变量之间相关关系的最直观感觉。结果显示变量相关性均值(Mean), 中位数(Median)和四分位差(IQR)分别为 0.063, 0.048, 0.173, 具体如图 4-2。说明强相关甲基化位点( $|\rho| > 0.5$ ) 在整个芯片覆盖的甲基化位点中不超过 0.1%, 其余甲基化位点之间是呈现中( $0.3 < |\rho| < 0.5$ ), 低程度( $0.1 < |\rho| < 0.3$ )的相关性。

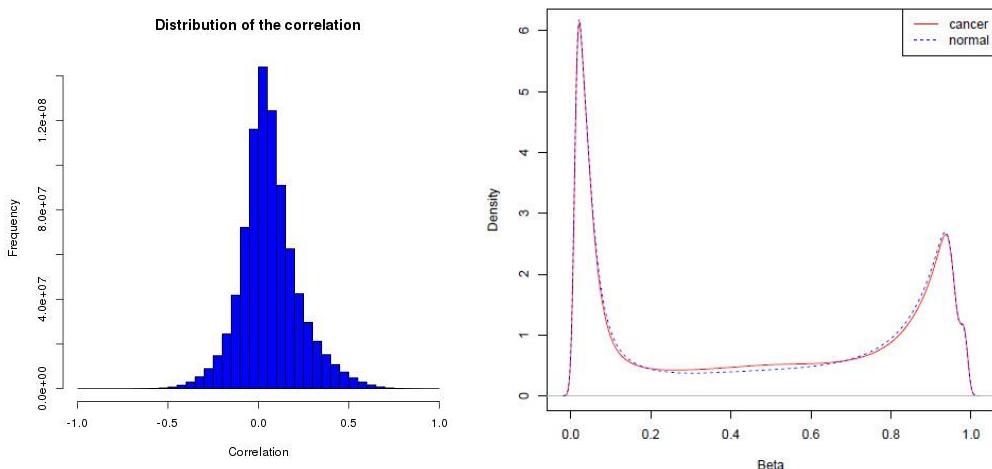


图 4-2 全基因组探针的甲基化特征

注: 在图 4-2 中, 左图为肿瘤和正常基因组 CpG 位点相关系数的抽样分布。右图为全基因组水平不同甲基化水平的频率分布。

来自 11 种肿瘤 639 对配对样本的 HM450K 芯片数据显示的不同甲基化水平的 CpG 位点的频率分布如图 4-2 的右图所示。可以观察到各水平 HM450K 芯片位点在人类肿瘤和正常基因组中呈现典型的双峰分布。并且不同程度甲基化位点的总体分布在肿瘤和正常中没有明显的差异。

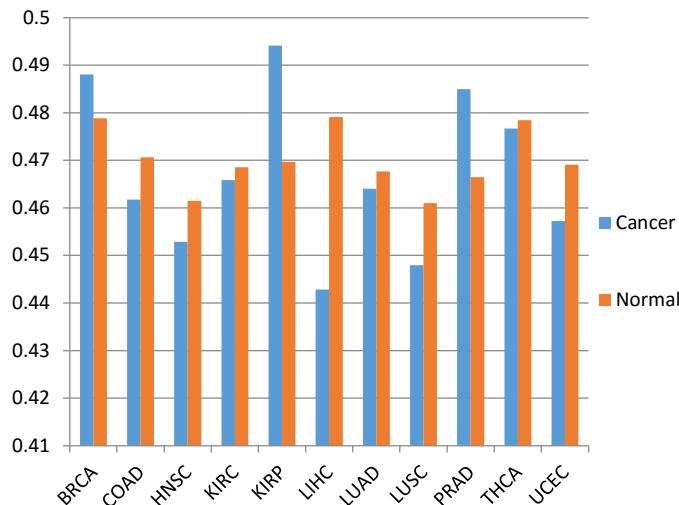


图 4-3 全基因组平均甲基化信号在肿瘤和正常中的比较

表 4-2 全基因组平均甲基化信号在肿瘤和正常中的比较

Type	Cancer	Normal
BRCA	0.4880691	0.4788181
COAD	0.4617677	0.4706255
HNSC	0.4528351	0.4614627
KIRC	0.4658208	0.4685300
KIRP	0.4941437	0.4697209
LIHC	0.4428313	0.4790895
LUAD	0.4640583	0.4676691
LUSC	0.4479694	0.4610142
PRAD	0.4849654	0.4664645
THCA	0.4766628	0.4784601
UCEC	0.4572670	0.4690640

根据已有肿瘤基因组的经验，肿瘤基因组中的甲基化呈现全基因组水平的低甲基化和局部区域的高甲基化。图 4-3 和表 4-数据表明 HM450K 芯片不能反映全基因组水平的降低（BRCA, KIRP, PRAD），其甲基化位点也不来自肿瘤局部高甲基化的区域（COAD, HNSC, KIRC, LIHC, LUAD, LUSC, THCA 和 UCEC）。

### 4.3.2 泛癌样本相关性分布特征

从样本相关系数热图矩阵（图 4-4）可以看出，1) 对于总样本而言，同一类型组织的样本之间出现高相关矩形区域，说明同一类型的组织间存在较高相

关性，如图 4-4。2) 最有正常组织，其同一类型的组织之间存在极强的强相关信号，高相关矩形区域信号非常明显。3) 对于任何组织，肿瘤样本的相关性都低于配对的正常对照之间的相关性。说明相同组织，肿瘤的异质性远高于正常样本。从图 4-4 还可以清晰看到 LUSC 和 LUAD 及 KIRC 和 KIRP 两种肿瘤存在较强的相关性，这与其相近的组织来源具有极强的吻合性。

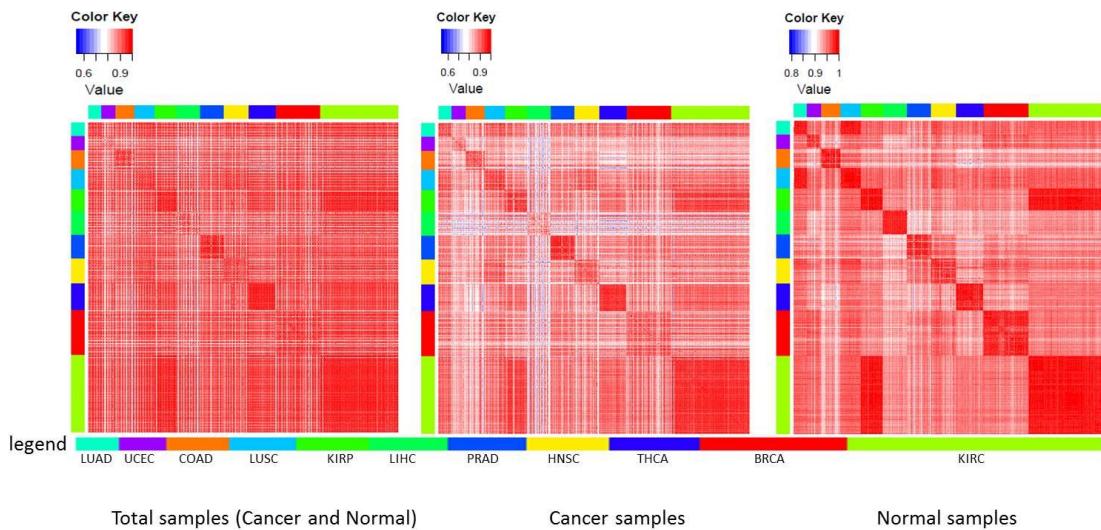


图 4-4 样本皮尔森相关性热图

如图 4-4 所示，采用热图对总样本（1274 样本量），肿瘤样本（637）和正常样本（637）的 pearson 相关系数进行展示。总样本相关系数热图中奇数行或列表示肿瘤样本，偶数行或列表示正常样本。组织顺序如图中图例所表示，从左到右，从上至下以此为：LUAD, UCEC, COAD, LUSC, KIRP, LIHC, PRAD, HNSC, THCA, BRCA, KIRC。传统观点认为肝癌是异质性比较低的一种癌症，但是肝癌样本之间的全基因组甲基化数据显示，不同肝癌个体间的甲基化相关性在所有肿瘤中是最低的( $\rho=0.84$ )。同样子宫内膜癌不同样本间的相关性也非常低( $\rho=0.87$ )，其他肿瘤类型的不同样本之间的相关性都高于 0.9。在所有肿瘤类型中 LUSC 和 LUAD 以及 KIRP 和 KIRC 呈现出比较高的相关性，其相关系数分别达到 0.9 和 0.92。对比肿瘤组织，正常组织之间的相关性及呈现出高度的相关性。全基因组甲基化数据显示相同组织类型的样本之间的相关性一般大于 0.97, LUSC 和 LUAD 以及 KIRP 和 KIRC 对应的正常组织之间的相关性均大于 0.99。

**表 4-3 肿瘤组织之间的平均相关性**

	LUAD	UCEC	COAD	LUSC	KIRP	LIHC	PRAD	HNSC	THCA	BRCA	KIRC
LUAD	<b>0.92</b>	0.85	0.86	<b>0.90</b>	0.89	0.82	0.88	0.88	0.89	0.88	0.90
UCEC	0.85	<b>0.87</b>	0.82	0.85	0.83	0.78	0.84	0.83	0.83	0.84	0.84
COAD	0.86	0.82	<b>0.90</b>	0.85	0.82	0.79	0.84	0.85	0.81	0.83	0.84
LUSC	0.90	0.85	0.85	<b>0.92</b>	0.87	0.81	0.88	0.89	0.87	0.87	0.89
KIRP	0.89	0.83	0.82	0.87	<b>0.92</b>	0.81	0.87	0.85	0.89	0.86	<b>0.92</b>
LIHC	0.82	0.78	0.79	0.81	0.81	<b>0.84</b>	0.81	0.80	0.79	0.81	0.82
PRAD	0.88	0.84	0.84	0.88	0.87	0.81	<b>0.95</b>	0.85	0.87	0.88	0.88
HNSC	0.88	0.83	0.85	0.89	0.85	0.80	0.85	<b>0.90</b>	0.84	0.86	0.87
THCA	0.89	0.83	0.81	0.87	0.89	0.79	0.87	0.84	<b>0.94</b>	0.86	0.90
BRCA	0.88	0.84	0.83	0.87	0.86	0.81	0.88	0.86	0.86	<b>0.90</b>	0.87
KIRC	0.90	0.84	0.84	0.89	0.92	0.82	0.88	0.87	0.90	0.87	<b>0.94</b>

注：泛癌全基因组甲基化数据反映的正常组织间的相关性（349,049 个甲基化位点），组织间的相关性为所有两种组织样本的两两相关性的平均。每种组织的样本数量如表 2-1。

**表 4-4 正常组织之间的平均相关性**

	LUAD	UCEC	COAD	LUSC	KIRP	LIHC	PRAD	HNSC	THCA	BRCA	KIRC
LUAD	<b>0.99</b>	0.95	0.94	<b>0.99</b>	0.96	0.95	0.94	0.96	0.95	0.96	0.96
UCEC	0.95	<b>0.97</b>	0.92	0.95	0.95	0.92	0.93	0.94	0.94	0.95	0.95
COAD	0.94	0.92	<b>0.97</b>	0.94	0.93	0.93	0.93	0.94	0.92	0.93	0.93
LUSC	0.99	0.95	0.94	<b>0.99</b>	0.96	0.95	0.94	0.96	0.95	0.96	0.96
KIRP	0.96	0.95	0.93	0.96	<b>0.99</b>	0.94	0.94	0.95	0.95	0.95	<b>0.99</b>
LIHC	0.95	0.92	0.93	0.95	0.94	<b>0.98</b>	0.92	0.93	0.93	0.93	0.94
PRAD	0.94	0.93	0.93	0.94	0.94	0.92	<b>0.97</b>	0.95	0.93	0.94	0.94
HNSC	0.96	0.94	0.94	0.96	0.95	0.93	0.95	<b>0.97</b>	0.94	0.95	0.95
THCA	0.95	0.94	0.92	0.95	0.95	0.93	0.93	0.94	<b>0.98</b>	0.94	0.95
BRCA	0.96	0.95	0.93	0.96	0.95	0.93	0.94	0.95	0.94	<b>0.98</b>	0.95
KIRC	0.96	0.95	0.93	0.96	0.99	0.94	0.94	0.95	0.95	0.95	<b>0.99</b>

注：泛癌全基因组甲基化数据反映的正常组织间的相关性（349,049 个甲基化位点），组织间的相关性为所有两种组织样本的两两相关性的平均。每种组织的样本数量如表。注意这里虽然用的行名和列名与上表相同，但是这里表示该肿瘤组织所对应的正常对照。

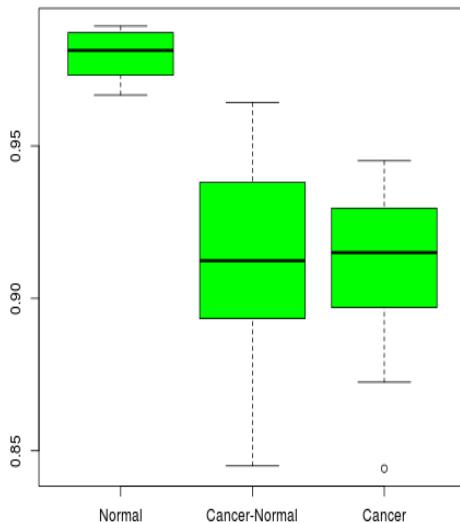


图 4-5 样本平均皮尔森相关性特征

图 4-5 采用箱线图对总样本（1274 样本量），肿瘤样本（637）和正常样本之间（637）的 Pearson 相关系数进行展示。右图展示各肿瘤组内，正常样本，正常-肿瘤配对样本及肿瘤样本间的 Pearson 相关性。结果显示肿瘤样本之间的平均相关系数为 0.9080（最小值：0.8442，最大值：0.9451），正常样本之间的平均相关系数在为 0.9797（最小值：0.9667，最大值：0.9893）。肿瘤样本见得相关性显著低于正常样本的相关性（ $t$ -test， $P = 1.059 \times 10^{-5}$ ）。此外可以发现配对的肿瘤样本和正常对照之间的相关性平均为 0.9129（最小值：0.8450，最大值：0.9643），其相关性水平基本和肿瘤样本之间的相关性持平（ $t$ -test， $P = 0.7445$ ），充分说明肿瘤之间的异质性之高。具体数值可以参考表 4-。

表 4-5 同一组织不同类型样本之间的相关性

	Normal	Cancer-Normal	Cancer
LUAD	0.9877	0.9325	0.9177
UCEC	0.9667	0.8450	0.8725
COAD	0.9749	0.8915	0.8975
LUSC	0.9893	0.9123	0.9150
KIRP	0.9874	0.9383	0.9203
LIHC	0.9813	0.8692	0.8442
PRAD	0.9715	0.9378	0.9451
HNSC	0.9708	0.8952	0.8997
THCA	0.9835	0.9643	0.9426
BRCA	0.9765	0.9081	0.8965
KIRC	0.9869	0.9474	0.9388

### 4.3.3 泛癌甲基化数据集的主成分分析

主成分分析在多元统计中具有重要的地位，PCA 可以有效地降低数据集的维度并且可以最大程度地保证数据的原有信息（方差），PCA 对于变量之间非独立的数据具有非常好的降维效果。根据我们前期的数据分析，泛癌甲基化数据集中有相当大的一部分特征变量具有中等程度的相关性。基于原始数据维度较大及特征变量具有相关性的考虑，PCA 可以简化我们对数据结构的了解。

采用主成分分析的方法可以抽提到数据集中的主要变异及分析主要差异的来源。采用主成分分析对泛癌甲基化数据（1274 样本，349,049 特征变量）的数据结构进行分析，如图 4-7 的 A 所示，我们可以看出 PCA 分析的前 10 个主成分即可解释整个数据集 55.74% 的变异。在前 50 个主成分内时，总方法随主成分的个数指数增加，之后增速逐渐放缓（如图 4-6 所示）。当取前 120 主成分时，所解释的方差占总方法的 80%。说明该数据的方差为有限规律变量，如性别，年龄，组织来源，病理状态等造成的而非大量噪音信号的叠加。通过第一主成分和第二主成分（如图 4-7 的 B 图），及第一主成分和第三主成分（如图 4-7 的 C 图）组成的平面可以看出肿瘤和正常组织在图上具有明显的分离现象。这说明肿瘤和正常的差异是数据总方差的最主要的来源。

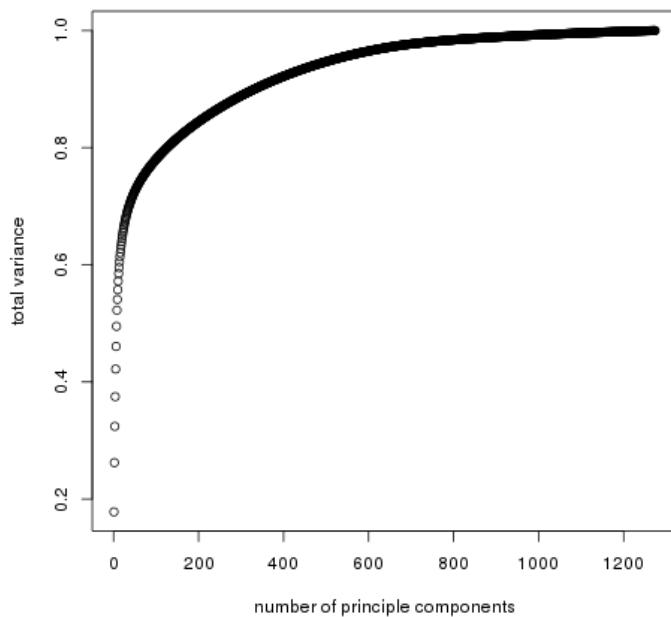


图 4-6 前 N 个主成分占总方差的比例的递增趋势

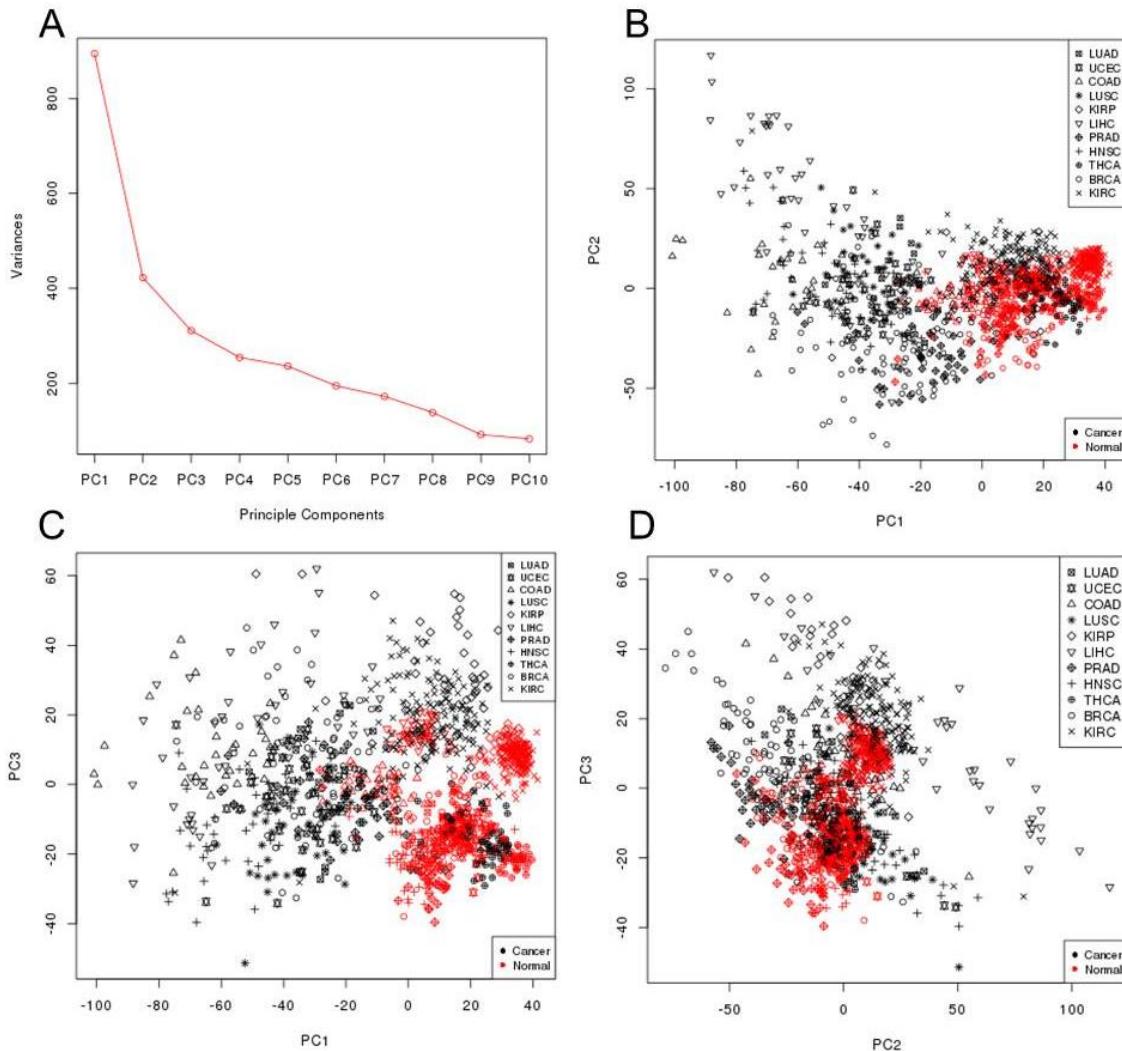


图 4-7 泛癌甲基化数据的主成分分析

#### 4.3.4 泛癌甲基化数据集的多维尺度分析

为了充分展示泛癌样本在全基因组 DNA 甲基化空间中的分布特征。我们采用了另一种空间降维的方式对泛癌样本的关系进行描述。在部分的分析中，样本之间的相似性采用随机森林模型中的两个样本之间的相似系数(两个样本分在同一类的概率)。

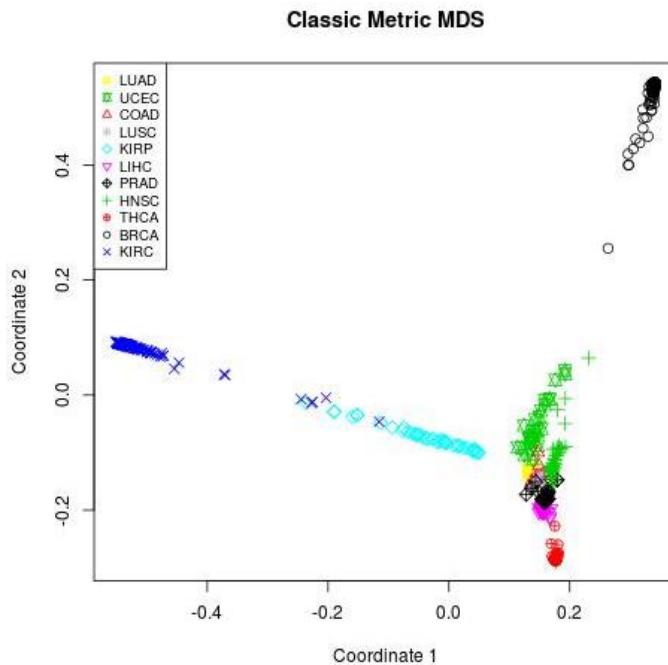


图 4-8 多维尺度分析展示癌症样本之间的距离

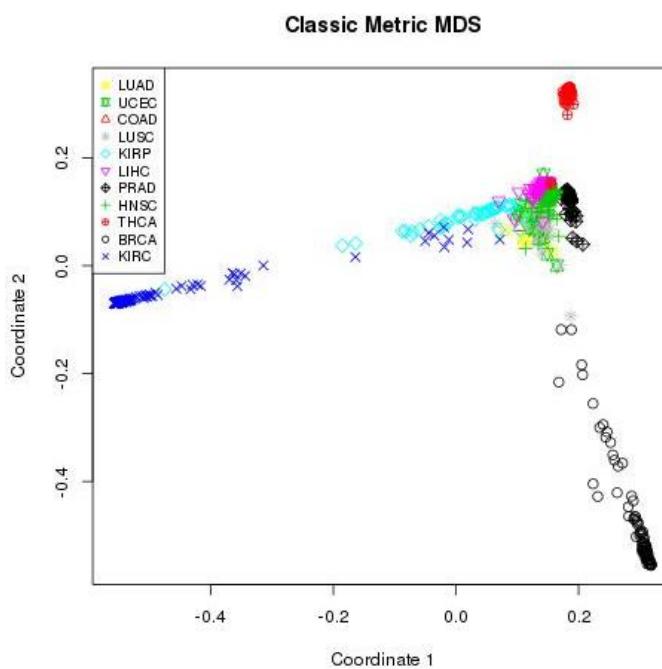


图 4-9 多维尺度分析展正常样本之间的距离

如图 4-及图 4-所示，基于 DNA 甲基化的多维尺度分析显示相同的肿瘤普遍进行聚类，不同的肿瘤具有不同的距离。比如 KIRC 和 KIRP, LUSC 和 LUAD 的距离较近。数据显示 KIRC, THCA 以及 BRCA 之间的相似性较低。通过对比肿瘤和正常样本不同组织类型之间的距离谱式，可以发现肿瘤和正常组织之间的距离谱式基本相似。

### 4.3.5 泛癌甲基化数据集的聚类分析

本部分通过两方面进行分析：个体的聚类分析和群体的聚类分析。个体的聚类分析以个体为单位。而群体的聚类是对 11 种肿瘤之间的关系进行讨论。采用 1-correlation 作为样本间距离，聚类方式采用 mcquitty 聚类法。我们完成了 1274 个样本的聚类分析，如图 4-所示。

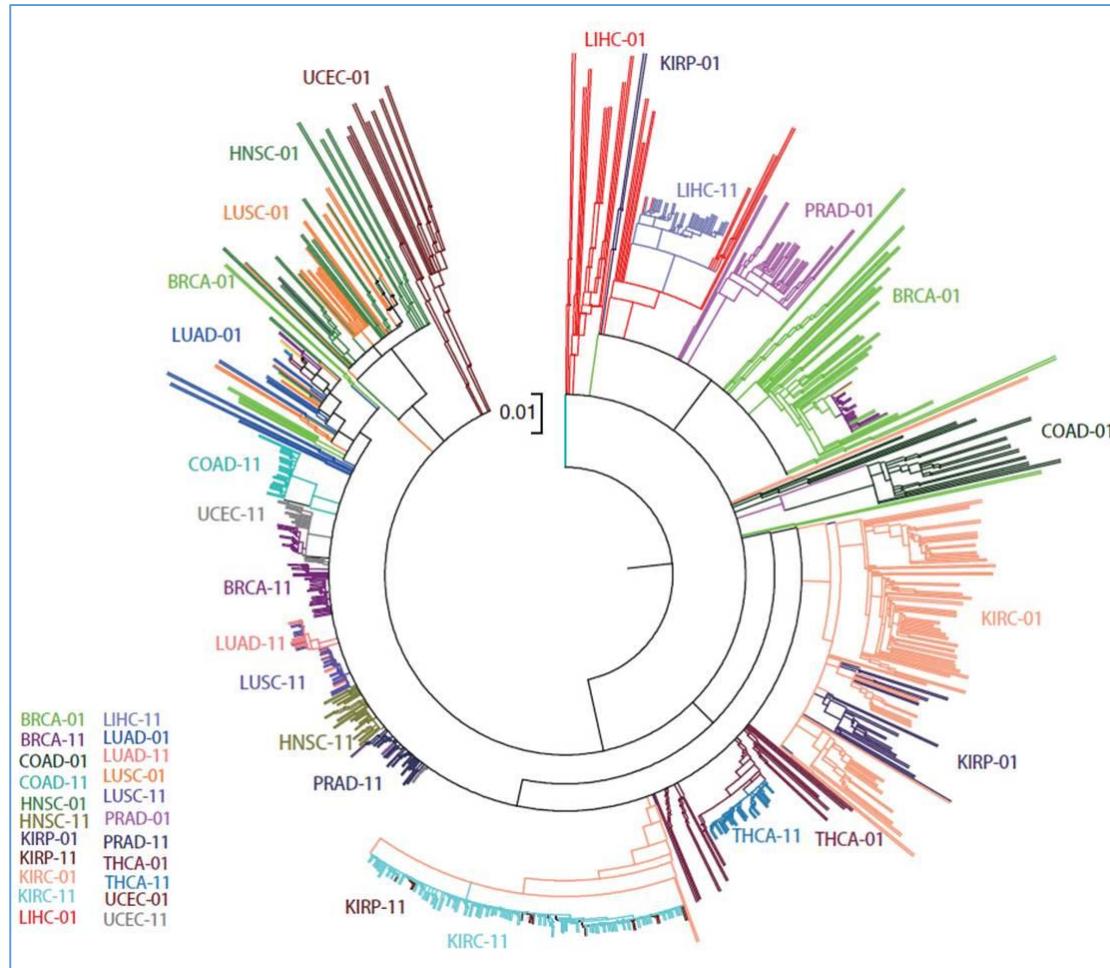


图 4-10 根据全基因组甲基化数据对 1274 样本的聚类分析

聚类分析结果显示相同类型的样本基本聚集在一起，说明全基因组 DNA 甲基化可以很好地对正常组织，各种类型的肿瘤组织进行聚类。

群体间的距离我们采用所有群体内两两样本距离的平均值作为群体距离。在本部分分析中，两两样本的距离采用：1-相关系数，群体的聚类采用所有样本（1-相关系数）的均值。算术平均的不加权对群法(UPGMA)用于聚类分析。利用上述参数和算法，我们完成了 11 类肿瘤的聚类分析，如图 4-所示。

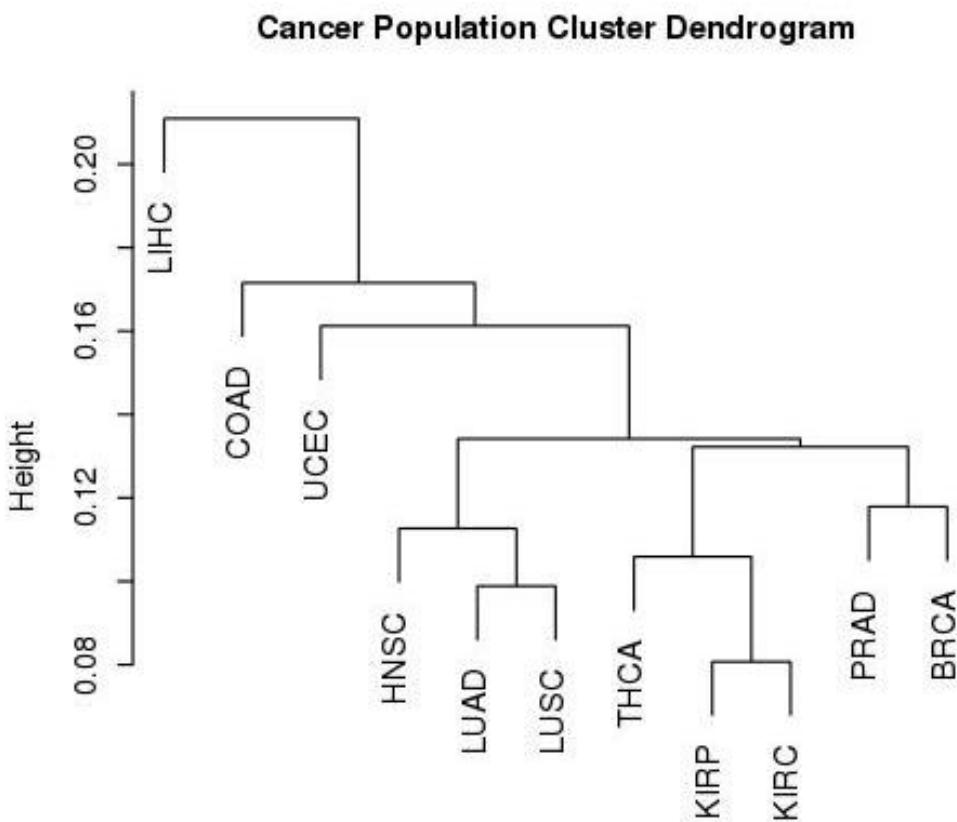


图 4-11 基于聚类分析的肿瘤组织相似性

基于肿瘤组织全基因组甲基化数据构建的肿瘤组织间近似性的树状结构。结果显示 1) 肺腺癌和肺鳞癌, 肾透明细胞癌和肾乳头状细胞癌, 前列腺癌和乳腺癌之间非常相似。2) 肝癌和肾透明细胞癌及肾乳头状细胞癌之间的相似性可能在所有癌症中最低。

#### 4.3.6 泛癌样本的差异显著甲基化位点总体特征

对 DNA 甲基化真实数据分析时发现基因甲基化在肿瘤和正常中的分布特征是比较复杂的, 有的符合正太分布, 有的不符合正太分布。采用 FDR 和 Bonferroni 两种不同的多重检验矫正方法对显著差异甲基化位点进行统计发现, 不同的肿瘤中差异甲基化位点的个数存在较大差异。

FDR 纠正过的结果显示, 所有肿瘤基因组中的都有至少 19.1% 的甲基化和正常组织存在显著差异。如果按照最严格的 Bonferroni 校正, 显著差异的 CpG 位点的比例从 1.5%-32.5% 不等。肺腺癌的 DNA 甲基化差异位点比例最低为 1.5%, 肾透明细胞癌的甲基化差异位点比例最低为 32.5%, 而肺鳞癌 (16.3%) 及肾乳头

状细胞癌（8.4%）的差异甲基化 CpG 位点比例却与之具有较大差异，说明 DNA 甲基化具有很大的潜力对不同的肿瘤亚型进行分类。具体如表 4-所示。

表 4-6 肿瘤差异基因及肿瘤特意性差异甲基化位点个数统计

Cancer	FDR-DMG (0.05)	Bonferroni-DMG	CanS-DMG	Sample size
LUAD	141558 (29.2% )	7439 (1.5% )	83 (0.0% )	26/26
UCEC	190612 (39.3% )	36724 (7.6% )	2,190 (0.5% )	30/30
COAD	194345 (40.0% )	49512 (10.2% )	3,776 (0.8% )	39/39
LUSC	257358 (53.0% )	78945 (16.3% )	6,133 ( 1.3% )	41/41
KIRP	168320 (34.7% )	40890 (8.4% )	2,909 (0.6% )	45/45
LIHC	192997 (39.7% )	47964 (9.9% )	4,261 (0.9% )	49/49
PRAD	175621 (36.2% )	62256 (12.8% )	5,295 (1.1% )	49/49
HNSC	217153 (44.7% )	56533 (11.6% )	2,003 (0.4% )	50/50
THCA	92793 (19.1% )	9273 (1.9% )	717 (0.1% )	56/56
BRCA	238149 (49.0% )	107127 (22.1% )	13,068 (2.7% )	94/94
KIRC	267168 (55.0% )	157598 (32.5% )	34,417 (7.1% )	160/160

注：FDR-DMG：基于 FDR 方法统计的差异甲基化基因个数 (0.05)。Bonferroni- DMG：基于 Bonferroni 方法统计的差异甲基化基因个数。 CanS-DMG 表示肿瘤特意的差异甲基化位点。 DMG: Differential Methylated Gene.

肿瘤相关分值(Tumor Relevant Score, TRS)在本文中被定义为该基因在多少种肿瘤被出现甲基化异常现象。分值高表示其参与多中肿瘤的发生或发展，分值低表示其特异性地在某种或某些肿瘤中存在异常甲基化。同理对于肿瘤相关分值较高的基因或位点，我们可以称为泛癌差异甲基化位点，表示其在多中肿瘤中均存在差异甲基化现象。对每个基因在 11 种肿瘤中出现显著异常甲基化出现的次数，对全基因组 DNA 甲基化位点的肿瘤相关分值的分布进行分析。经过采用配对 wilcox 检验，我们对 349,049 位点在 11 个肿瘤 1274 样本进行了差异情况分析。结果如图 4-6 所示，可以发现，在 11 种肿瘤中均出现显著性甲基化异常的位点只有 42 个。在 10 种肿瘤中均出现显著性甲基化异常的位点只有 259 个，在 9 种肿瘤中均出现显著性甲基化异常的位点只有 1051 个。

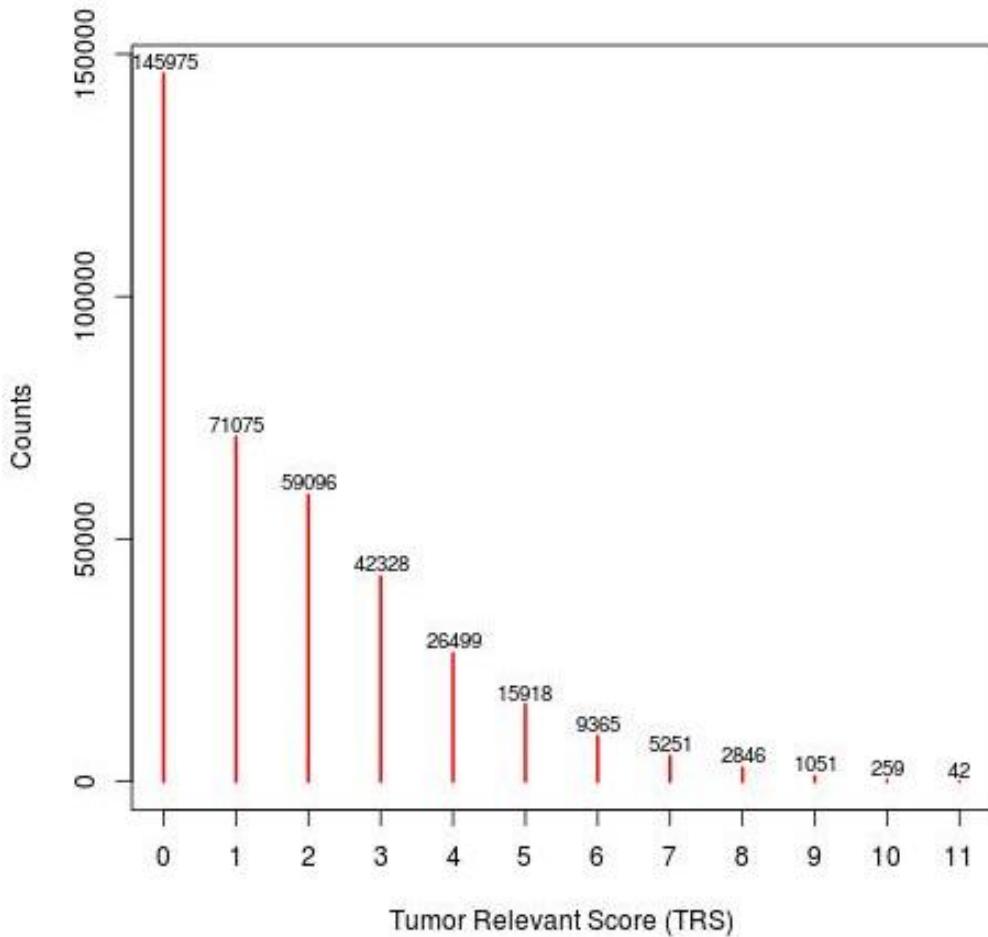


图 4-6 基因的肿瘤相关分值的分布情况

#### 4.3.7 相邻 CpG 位点之间的相关性

DNA 在遗传过程中采用自由组合和交叉互换（重组，recommbination）的方式，实现遗传信息的传承和变异。重组机制形成了群体基因组水平的连锁不平衡（LD）现象。为此我们思考 DNA 甲基化在遗传过程中是跟随着特定的重组而表现出一定的连锁现象，即甲基化在染色体上是否有类似的 LD 信号。以 Chr20 上的 875 个甲基化位点为例子，Chr20 所有位点的平均相关性为 0.160，采用 3000 随机迭代，以不同的窗口计算区域平均相关性时，我们发现当以 2 个相邻 CpG 的平均相关性为 0.448 (95% CI: 0.446-0.450)，5 个相邻 CpG 的平均相关性为 0.368 (95% CI: 0.366-0.371)，10 个相邻 CpG 的平均相关性为 0.302 (95% CI: 0.300-0.304)。CpG 位点之间的平均相关性随着距离的增加逐渐降低。如图 4-7 所示。

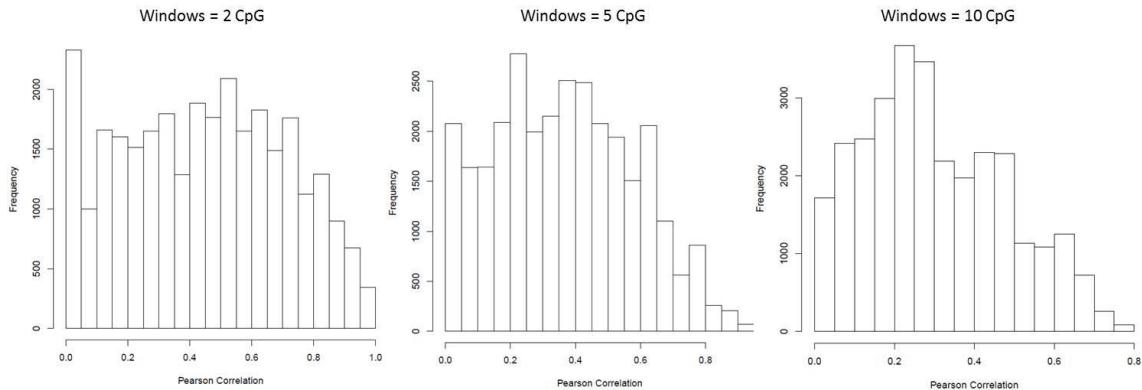


图 4-7 肿瘤和正常基因组相邻 CpG 区域平均相关系数的抽样分布

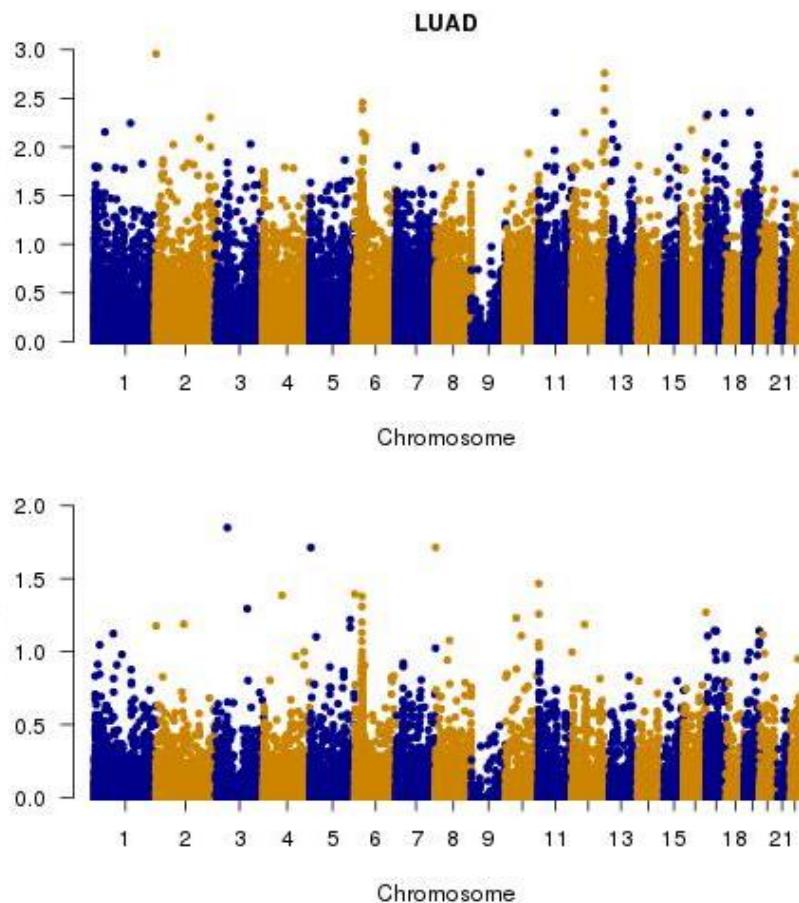


图 4-14 以肺腺癌为例展示全基因组 CpG 位点甲基化状态相关性

如图 所示,采用窗口移动法对染色体相邻位置 CpG 甲基化相关性进行分析,上图基于肿瘤数据,下图基于对应的正常数据。窗口大小为 5 个相邻探针,步长为 2 个探针数。X 轴代表染色体, Y 轴是  $-\log(1-\text{cor}/\text{cor})$ , Y 轴越大说明相邻位点之间的平均相关性越高。分析显示癌症中的相邻位点见的相关性比正常组织普遍偏高。其他 10 中肿瘤中也观察到了同样的现象,在此不在重复罗列。

### 4.3.8 泛癌样本的随机森林预测模型

对多种肿瘤进行同时预测具有重要的应用价值，特别是对于直接应用血液分子对肿瘤进行的早期筛查，诊断或辅助诊断。在随机森林预测模型中的变量（349,049）中，采用分段预测并组合初始变量选择方法，将全基因组甲基化位点按照 1000 个位点进行随机分组，选择出对每组中的最重要的 20 个变量，然后将 350 组的重要变量合并，去重复位点，得到 6631 非重复位点，采用这种逐步筛选再组合的方式，模型最终筛选到 6631 非重复位点位点。通过对 6631 个预测功能位点建立多肿瘤预测模型，如图 4-8 所示，可以发现随机森林模型可以对多肿瘤进行同时预测并具有较高的预测性能。从变量重要性的分布来看，434 个位点（变量重要性 $>0.5$ ）呈现及其重要的预测潜力。

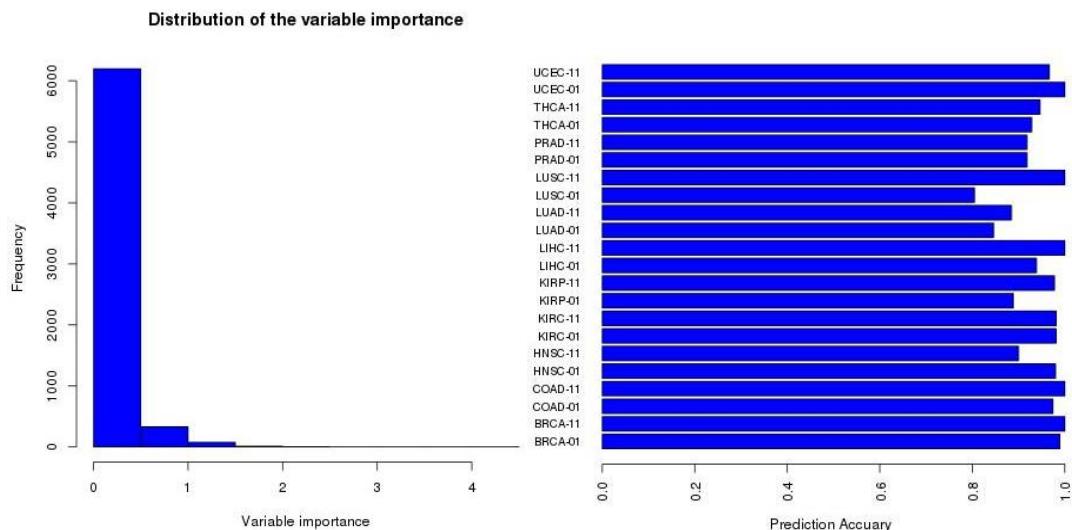


图 4-8 基于区段组合法的随机森林模型变量重要性分布及预测精度

作为比较，基于显著差异的甲基化位点的随机森林多肿瘤预测模型也按照相似的方法进行了建模。在随机森林预测模型中的变量（349,049）中有 4203 个在至少 7 种肿瘤中呈现显著性甲基化差异( $P < 1.03 \times 10^{-7}$ )。从变量重要性的分布来看，563 个位点呈现及其重要的预测潜力(变量重要性 $>0.5$ )。但是通过对比图 2-9 和图 2-10，我们能发现，采用在很多种肿瘤中异常的甲基化位点对多肿瘤进行同时预测在准确度上远远不及采用功能预测位点。如图 4-9 所示。

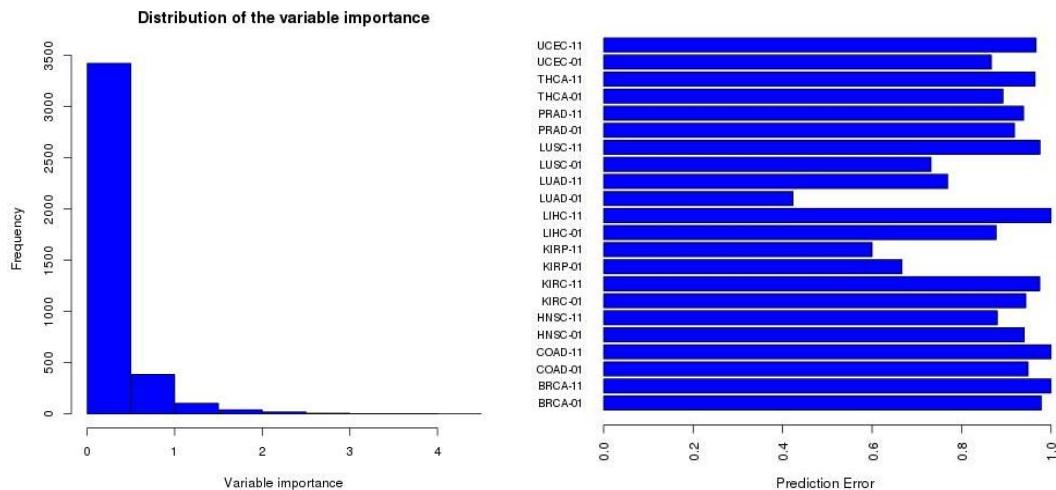


图 4-9 基于差异甲基化位点的随机森林变量重要性分布及预测精度

事实证明直接采用显著差异的位点对多肿瘤进行预测并不能使得预测模型获得最优的预测能力。我们对上述两个模型的预测变量之间的相关性的分布可以看出差异甲基化位点模型筛选出的重要变量之间具有较强的相关性，出现了明显的正负相关两个峰，其中相关性较弱的变量因为不是差异甲基化位点而未能进入预测模型。功能预测位点更多采用了中低相关性的变量，从而充分利用了变量的互相补充的信息。如图 4-10 所示。

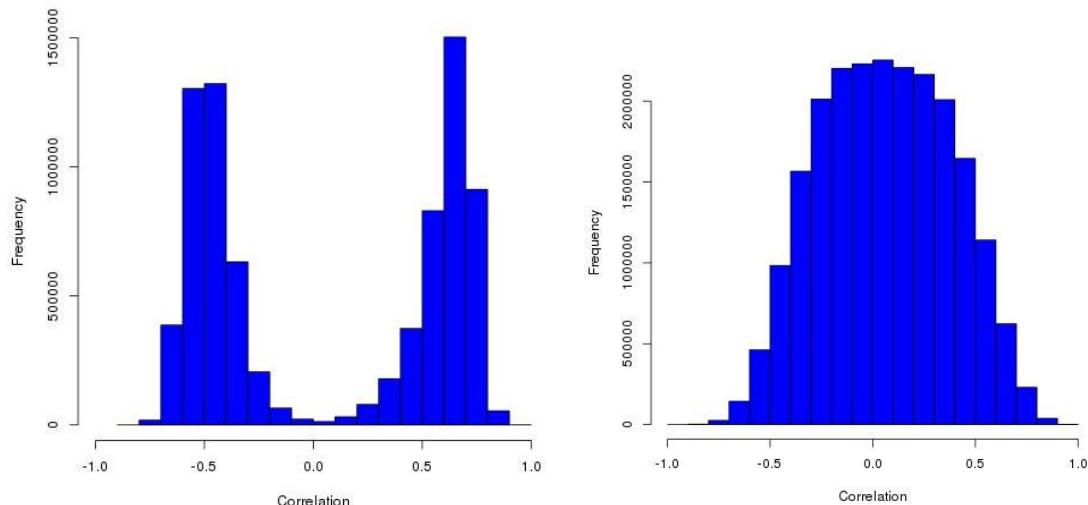


图 4-10 差异甲基化位点及预测功能位点探针的相关性

注：差异甲基化位点（左）及预测功能位点（右）探针的相关性存在显著差异。

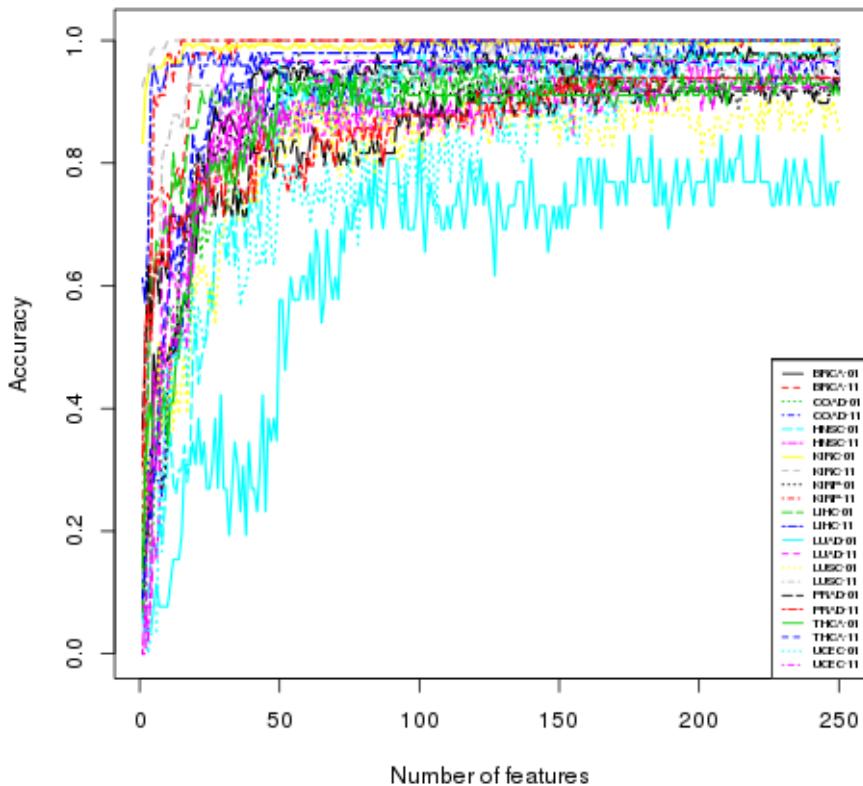


图 4-11 预测变量个数与多肿瘤预测准确度关系

预测模型基于随机森林方法。采用差异甲基化位点作为初始变量，具体为：将位点作为全基因组甲基化位点按照 1000 个位点进行随机分组，选择出对每组中的最重要的 20 个变量，然后将 350 组的重要变量合并，去重复位点，得到 3986 个非重复位点，之后对将这些位点作为变量进行随机森林方法建立预测模型，统计灵敏性，特异性，准确度及最重要预测变量。

表 4-7 采用前 25 和 75 个变量对应的预测模型表现情况

	Variable=25			Variable=75		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
BRCA	0.848	0.967	0.908	0.935	1.000	0.967
COAD	0.718	0.846	0.782	0.923	0.897	0.910
HNSC	0.580	0.720	0.650	0.940	0.900	0.920
KIRC	0.994	1.000	0.997	0.988	1.000	0.994
KIRP	0.844	1.000	0.922	0.911	1.000	0.956
LIHC	0.878	0.959	0.918	0.918	0.980	0.949
LUAD	0.269	0.731	0.500	0.577	0.846	0.712
LUSC	0.634	0.927	0.780	0.780	0.927	0.854
PRAD	0.776	0.816	0.796	0.857	0.837	0.847
THCA	0.786	0.875	0.830	0.875	0.964	0.920
UCEC	0.567	0.767	0.667	0.767	0.967	0.867

通过对图 4-11 和表 4-的观察可以看出，当采用的预测变量从 1 到 25 逐个增加时模型的预测准确度指数上升，在 Variable 为 25 个是平均的预测准确度已经达到 80%，当 Variable 为 75 个是平均的预测准确度已经达到 89.9%。值得注意的是肺腺癌的预测模型和其他肿瘤的预测模型相比，表现很差。如果不考虑肺腺癌的预测模型在 Variable 为 25 个是平均的预测准确度已经达到 82.4%。在这种情况下，当把 Variable 增加到 50 个的过程中，准确度的增加已经非常缓慢。因此我们推断。采用 DNA 甲基化作为生物标记物对本文所涉及的 11 种肿瘤：肾透明细胞癌，浸润性乳腺癌，甲状腺癌，头颈部鳞状细胞癌，前列腺癌，肝癌，肾透明细胞癌，肺腺癌，结肠癌，子宫内膜癌，肺鳞癌进行预测的最宜生物标记物组合数量可以控制在 25-50 个之间。数量过少会降低预测模型对多种肿瘤诊断的准确性，数量过多很提高检测成本。通过如果同时检测的甲基化位点过多，会增加探针之间的干扰性，降低检测的可重复性。

### 4.3.9 肺癌差异甲基化谱式特征

如图 4-12 所示，肺癌是最迫切需要早期诊断标记物的肿瘤之一。肺癌肿瘤标记物及诊断模型拥有早期诊断肿瘤可以显著性的提高肺癌患者的五年生存率。为此我们专门对肺癌的甲基化谱式情况做了一些描述。

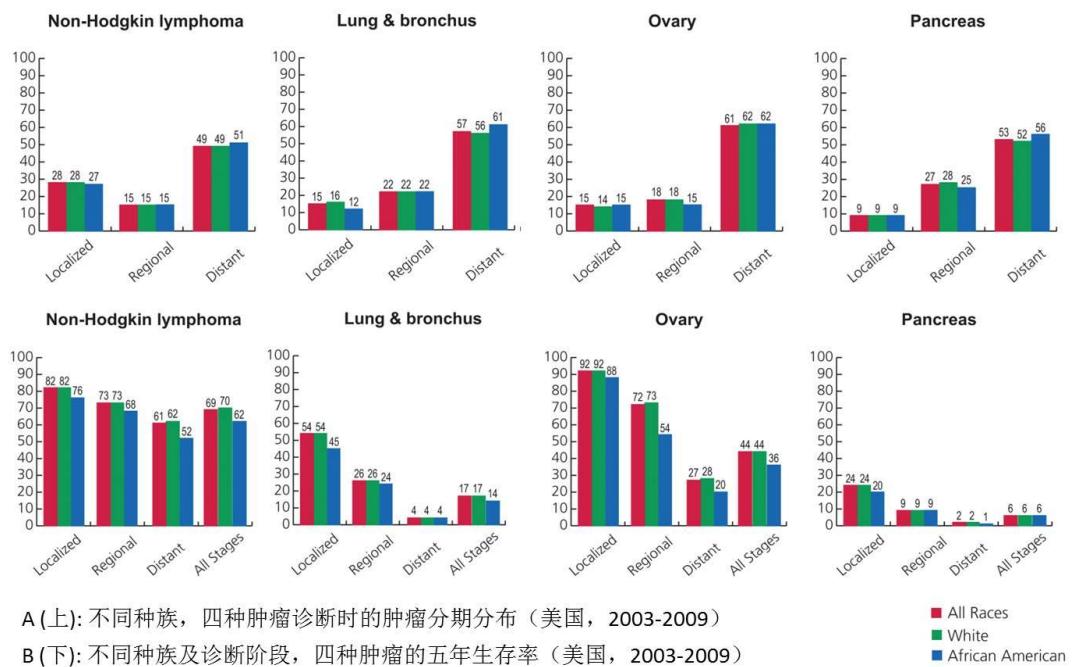


图 4-12 最迫切需要早期诊断标记物的四种肿瘤

通过对肺癌甲基化 Methylation 450K 芯片数据的整理，发现用于相关分析的数据包括： 1) 429 例肺腺癌组织及 26 例对应正常组织对照（癌旁组织）； 2)

407 例鳞癌组织及 41 例正常鳞癌组织对照（癌旁组织）。其中有癌-癌旁匹配的数据包括： 26 对肺腺癌和 41 对肺鳞癌。

首先，我们对全基因组甲基化位点与肺癌易感性的关联研究，本分析中利用配对的样本，采用 logistic regression 的方法对每个甲基化位点进行 test 以判断其是否与肺癌易感性相关。99% 置信区间和 Bonferroni 多重校正在上述检验中被采用，以降低假阳性率。通过分析，结果如

图 4- 所示。我们发现了大量的肺腺癌和鳞癌相关的甲基化位点。按照 Bonferroni 多重校正的方法， $P\text{-value}$  小于  $1.03 \times 10^{-7}$  的原则，肺腺癌和肺鳞癌中都发现了大量的显著相关的易感性甲基化位点。数据显示肺腺癌中存在 209,434 个肺癌易感的甲基化位点。肺鳞癌中有 102,536 个肺癌易感的甲基化位点。这是在没有考虑甲基化差异 effect size 的情况下得到了显著性的易感位点。如果采用更加严格的条件，要求 effect size 的差异的绝对值大于 0.3，则在肺腺癌中只有 1072 个显著位点，在肺鳞癌中只有 8510 个显著位点。

其次，我们观察肺腺癌和肺鳞癌在甲基化易感位点方面的差异，我们对每一个甲基化位点在肺腺癌和鳞癌的差异情况进行比较。在本分析中，我们采用基于配对数据的 Wilcoxon signed rank 检验对每个位点进行分析。

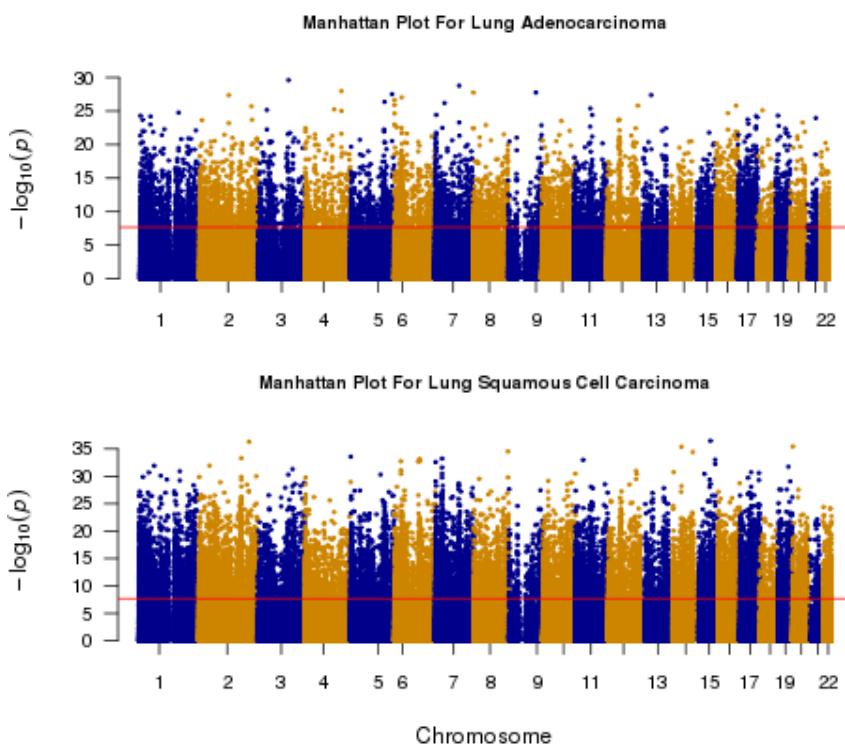


图 4-20 Manhattan 图显示肺腺癌和鳞癌异常甲基化位点

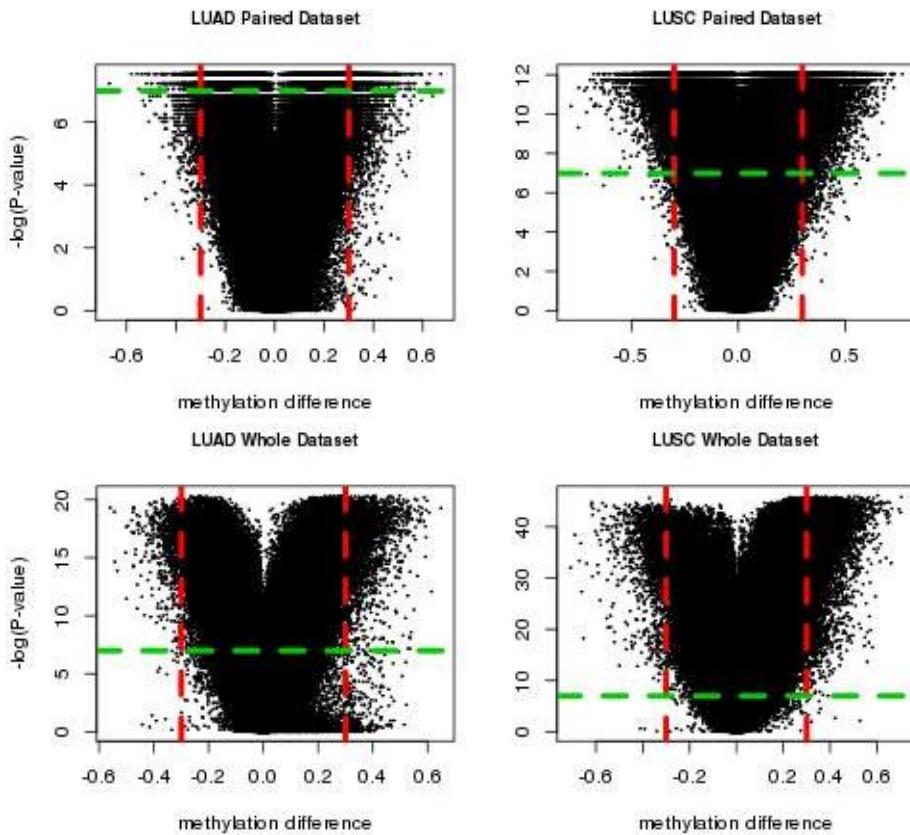


图 4-21 基于配对或非配对数据分析肺癌甲基化差异基因

注：分别基于配对 LUAD，LUSC 和全数据的 LUAD 和 LUSC 计算差异甲基化和 P-value。

基于配对样本及 Wilcoxon signed rank 检验以及 effect size 的差异的绝对值大于 0.3 的原则，我们发现在腺癌中有 1263 个差异甲基化位点，而在肺鳞癌中有 11882 个差异甲基化位点。当采用全部的数据进行分析时，我们发现在腺癌中有 3137 个差异甲基化位点，而在肺鳞癌中有 8156 个差异甲基化位点。详细如图 4-所示。根据上述分析可以看出肺腺癌和肺鳞癌在整体甲基化异常上具有较大的差异。肺鳞癌具有更多的差异甲基化位点。流行病学资料显示 75% 左右的鳞癌患者来源于吸烟人群，而只有 50% 的肺腺癌患者来源于吸烟人群。上述证据或许暗示 DNA 甲基化异常在鳞癌的发病中或许发挥更重要的作用。

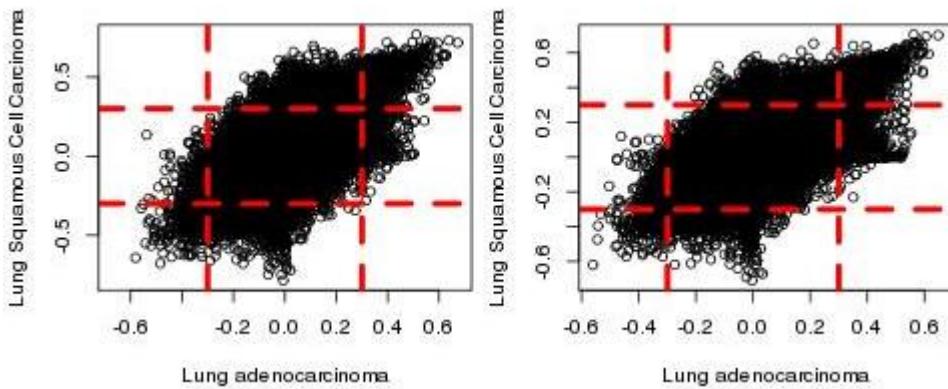


图 2-13 腺癌和肺鳞癌相同位点甲基化差异比较

注：左图基于配对数据集，右图基于全部数据集

如图 2-13 所示，基于配对数据和非配对数据对相同位点在肺腺癌和肺鳞癌中的差异情况的比较看出，大多数位点在鳞癌和腺癌中基本保持同向的变化。

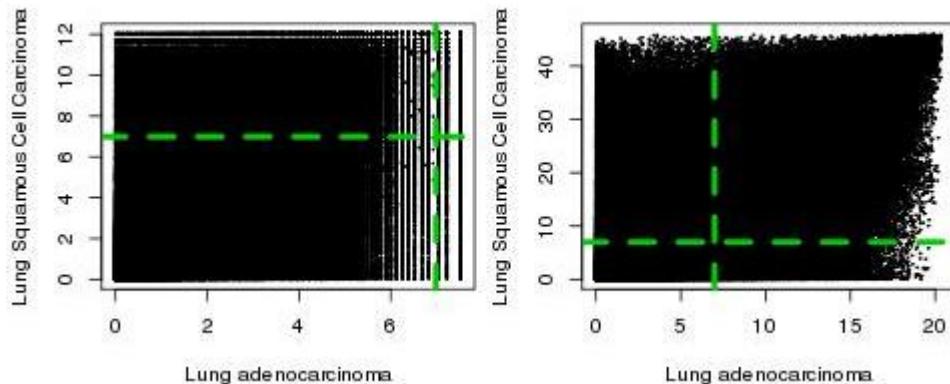


图 2-14 非腺癌和肺鳞癌相同位点甲基化差异 P-value 比较

注：左图基于配对数据集，右图基于全部数据集

所示图 2-14 所示，基于配对数据和非配对数据对相同位点在肺腺癌和肺鳞癌中的差异的 P-value 的比较看出：鳞癌比腺癌拥有更多地差异甲基化位点。之后，提取腺癌和鳞癌配对样本中均有显著差异的基因( $P < 1.03 \times 10^{-7}$ )，并设定癌症和正常对照组甲基化信号差异  $\Delta\beta > 0.3$ 。在此限制条件下 1113 个甲基化位点呈现在肺腺癌和肺鳞癌均存在病例-对照显著型差异。其中这些位点设计到 525 个基因。对这 525 个基因进行信息学分析，没有发现 KEGG 信号通路的差异，但有大量的 Gene Ontology 模块呈现显著性聚集。其中有 81 个 BP 模块 FDR 校正后的 Pvalue < 0.05，有 6 个 MF 模块 FDR 校正后的 Pvalue < 0.05。下表展示了最为显著的 30 个 Gene ontology 模块。所有 gene ontology 列表如表 4-所示。

**表 4-8 肺腺癌和鳞癌异常甲基化基因的 Gene ontology 模块**

term	Fold Enrichment	FDR
sequence-specific DNA binding	4.80	4.09E-39
transcription factor activity	3.68	5.49E-37
transcription regulator activity	2.75	5.05E-29
pattern specification process	6.39	1.18E-26
regionalization	7.17	8.89E-24
regulation of transcription, DNA-dependent	2.45	9.41E-24
regulation of RNA metabolic process	2.40	9.80E-23
embryonic morphogenesis	5.46	1.75E-22
anterior/posterior pattern formation	7.77	7.06E-19
regulation of transcription	1.97	5.30E-18
embryonic skeletal system development	10.69	1.48E-17
embryonic development ending in birth or egg hatching	4.58	1.09E-16
chordate embryonic development	4.53	4.14E-16
skeletal system morphogenesis	7.35	7.66E-13
regulation of transcription from RNA polymerase II P	2.79	2.56E-11
<b>homophilic cell adhesion</b>	<b>6.29</b>	<b>4.98E-11</b>
transcription	1.89	5.92E-11
skeletal system development	3.96	6.00E-11
neuron differentiation	3.22	3.66E-09
<b>cell fate commitment</b>	<b>5.50</b>	<b>1.13E-08</b>
positive regulation of transcription, DNA-dependent	2.90	2.48E-07
positive regulation of RNA metabolic process	2.87	3.34E-07
positive regulation of transcription from RNA polymerase II	3.17	5.61E-07
appendage morphogenesis	5.94	1.26E-06
<b>cell-cell adhesion</b>	<b>3.52</b>	<b>2.05E-06</b>
positive regulation of transcription	2.61	2.22E-06
sensory organ development	3.72	7.00E-06
embryonic limb morphogenesis	6.08	7.48E-06
<b>calcium ion binding</b>	<b>1.88</b>	<b>0.0034475</b>
transcription activator activity	2.35	0.0128422

### 4.3.9.1 肺腺癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为肺腺癌诊断标记物的如下甲基化位点。

**表 4-9 肺腺癌最重要的甲基化诊断位点**

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg10853830	2.00	2.98E-08	1.00	1.00	0.00	PITPNM1
cg02174232	1.98	2.98E-08	0.00	0.58	0.58	APEH
cg18407955	1.97	2.98E-08	1.00	0.12	0.88	PTPRN2
cg13747967	1.96	2.98E-08	0.12	1.00	0.88	CCL22
cg04461228	1.91	2.98E-08	1.00	0.08	0.92	FLJ42709
cg12419685	1.88	2.98E-08	1.00	1.00	0.00	GGT7
cg07076109	1.73	2.98E-08	1.00	1.00	0.00	FOXK1
cg26509691	1.73	2.98E-08	0.85	0.00	0.85	PITX1
cg11117099	1.73	2.98E-08	0.00	0.92	0.92	LOC100133991
cg18886436	1.73	2.98E-08	0.00	0.19	0.19	ZNF385A
cg02331025	1.73	2.98E-08	1.00	1.00	0.00	GDNF
cg00373436	1.72	2.98E-08	0.85	1.00	0.15	TAGLN3
cg27269936	1.72	2.98E-08	1.00	1.00	0.00	CAMKK2
cg17356718	1.72	2.98E-08	1.00	1.00	0.00	HDAC4
cg08328160	1.72	2.98E-08	1.00	1.00	0.00	AGER
cg07004744	1.71	2.98E-08	1.00	1.00	0.00	ERICH1
cg04762412	1.71	2.98E-08	1.00	1.00	0.00	NCOA5
cg25021051	1.71	2.98E-08	0.00	0.96	0.96	TENC1
cg02926033	1.70	2.98E-08	0.96	1.00	0.04	DPP6
cg01383890	1.69	2.98E-08	0.00	0.00	0.00	C2CD4A
cg12426196	1.68	2.98E-08	0.00	0.69	0.69	TSTD1
cg16780688	1.68	2.98E-08	0.62	1.00	0.38	MAL2
cg27149973	1.68	2.98E-08	0.12	1.00	0.88	MGAT5B
cg14231297	1.67	2.98E-08	0.96	0.00	0.96	ZSCAN18
cg13148085	1.67	2.98E-08	0.00	0.38	0.38	HELZ
cg11866303	1.66	2.98E-08	0.00	0.23	0.23	CNTD2
cg16887264	1.65	2.98E-08	0.69	0.00	0.69	POU4F2
cg05798501	1.64	2.98E-08	1.00	1.00	0.00	NOD1
cg01280080	1.63	2.98E-08	0.00	1.00	1.00	ATIC
cg03625953	1.42	2.98E-08	1.00	1.00	0.00	SMG6

### 4.3.9.2 肺鳞癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为肺鳞癌诊断标记物的如下甲基化位点。

**表 4-10 肺鳞癌最重要的甲基化诊断位点**

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg26790247	1.99	9.09E-13	1.00	0.02	0.98	ZIC4
cg20345923	1.99	9.09E-13	0.90	1.00	0.10	MYT1L
cg26951440	1.97	9.09E-13	1.00	0.17	0.83	NKX2-3
cg13153865	1.73	9.09E-13	0.00	0.00	0.00	OSGIN2
cg21039708	1.73	9.09E-13	1.00	0.00	1.00	OTX2OS1
cg06476693	1.73	9.09E-13	0.76	1.00	0.24	ADAMTS8
cg17526483	1.72	9.09E-13	1.00	0.95	0.05	EVX1
cg01772014	1.72	9.09E-13	0.00	0.71	0.71	C20orf160
cg15588997	1.72	9.09E-13	0.95	1.00	0.05	KIR3DL1
cg00157572	1.71	9.09E-13	1.00	0.02	0.98	SIX6
cg10029842	1.71	9.09E-13	0.76	0.00	0.76	LHX8
cg26612727	1.71	9.09E-13	1.00	1.00	0.00	ZPBP2
cg15798385	1.70	9.09E-13	1.00	0.39	0.61	EVX1
cg08317412	1.68	9.09E-13	0.93	1.00	0.07	PANX3
cg03464573	1.67	9.09E-13	0.95	0.00	0.95	HOXA9
cg06215569	1.65	9.09E-13	0.98	0.00	0.98	ALX3
cg09226786	1.64	9.09E-13	0.80	0.00	0.80	DLX6AS
cg15424989	1.42	9.09E-13	1.00	1.00	0.00	GATA6
cg13882284	1.42	9.09E-13	1.00	0.71	0.29	MSX1
cg02764897	1.42	9.09E-13	0.88	1.00	0.12	KRTAP13-1
cg16746576	1.42	9.09E-13	1.00	0.00	1.00	OSR2
cg08301299	1.42	9.09E-13	0.39	0.00	0.39	RNPS1
cg24826339	1.41	9.09E-13	1.00	1.00	0.00	LOC389033
cg06370771	1.41	9.09E-13	0.85	0.00	0.85	PAX7
cg00666640	1.41	9.09E-13	0.98	1.00	0.02	OR2T12
cg09924998	1.41	9.09E-13	0.07	1.00	0.93	VPS33A
cg05860723	1.41	9.09E-13	0.93	0.00	0.93	MYF6
cg24384519	1.41	9.09E-13	0.98	1.00	0.02	ADCYAP1R1
cg00298065	1.41	9.09E-13	1.00	1.00	0.00	ZNF274

### 4.3.10 乳腺癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为肿瘤诊断标记物的如下甲基化位点。

表 4-11 乳腺癌最重要的甲基化诊断位点

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg15574321	2.72	8.28E-17	1.00	0.99	0.01	WNT11
cg05937737	2.15	8.28E-17	0.99	0.16	0.83	SLC7A14
cg14394264	1.99	8.28E-17	1.00	0.91	0.09	PHF21A
cg18417954	1.99	8.28E-17	0.93	0.05	0.88	C19orf51
cg14276730	1.94	8.28E-17	1.00	0.89	0.11	CHD6
cg15873474	1.91	8.28E-17	0.78	1.00	0.22	CDH4
cg14231297	1.83	8.28E-17	0.99	0.07	0.92	ZSCAN18
cg18675097	2.12	8.56E-17	0.99	0.17	0.82	NKAPL
cg23690893	2.07	8.56E-17	1.00	0.93	0.07	TECR
cg25026529	1.95	8.56E-17	0.67	0.00	0.67	BARHL2
cg24848035	1.87	8.56E-17	0.99	0.13	0.86	RGS22
cg10659886	1.86	8.56E-17	0.97	0.24	0.73	ZSCAN18
cg25588480	1.97	8.85E-17	1.00	1.00	0.00	MINK1
cg17384889	1.95	8.85E-17	0.68	0.01	0.67	NKAPL
cg23866403	1.86	8.85E-17	1.00	0.96	0.04	LGI4
cg23610154	1.72	8.85E-17	0.89	1.00	0.11	LOC254312
cg11274778	1.71	8.85E-17	0.78	1.00	0.22	BASE
cg20748242	1.95	9.14E-17	1.00	1.00	0.00	RNF44
cg15795544	1.93	9.14E-17	1.00	1.00	0.00	SLC3A2
cg25839439	1.73	9.14E-17	1.00	0.96	0.04	CAPZB
cg18694169	1.70	9.14E-17	0.87	0.02	0.85	NKAPL
cg07204662	1.98	9.45E-17	1.00	0.99	0.01	ESRRA
cg18241962	1.73	9.45E-17	0.90	0.02	0.88	9-Sep
cg12134633	1.69	9.45E-17	0.99	0.37	0.62	SCG5
cg09069446	1.68	9.45E-17	1.00	1.00	0.00	CORO7
cg01507342	1.96	9.76E-17	0.07	0.96	0.89	PITPNC1
cg08104202	1.73	9.76E-17	0.82	0.00	0.82	C1orf114
cg21048162	1.73	1.01E-16	0.61	1.00	0.39	BATF
cg03100639	1.70	1.01E-16	0.99	0.90	0.09	TGFB1II
cg02156380	1.69	1.01E-16	0.99	0.38	0.61	KIAA1614

### 4.3.11 结肠癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为结肠癌诊断标记物的如下甲基化位点。

**表 4-1 结肠癌最重要的甲基化诊断位点**

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg18347642	1.99	3.64E-12	1.00	0.00	1.00	KCNIP4
cg24691891	1.98	3.64E-12	1.00	1.00	0.00	PHC2
cg14106046	1.98	3.64E-12	0.15	1.00	0.85	B3GNTL1
cg06952671	1.98	3.64E-12	0.97	0.00	0.97	ITGA4
cg17373442	1.96	3.64E-12	1.00	0.05	0.95	CHST2
cg09621603	1.90	3.64E-12	0.64	1.00	0.36	C10orf90
cg13405887	1.73	3.64E-12	1.00	0.03	0.97	C9orf50
cg09731694	1.73	3.64E-12	1.00	0.00	1.00	C9orf50
cg11295724	1.73	3.64E-12	0.77	1.00	0.23	SIRPD
cg22403344	1.73	3.64E-12	0.97	0.00	0.97	MAL
cg02608452	1.73	3.64E-12	0.62	1.00	0.38	KCNA4
cg00629625	1.72	3.64E-12	0.67	1.00	0.33	SLC12A5
cg06668555	1.72	3.64E-12	1.00	0.08	0.92	FAM135B
cg09316122	1.72	3.64E-12	0.97	0.00	0.97	WNT3A
cg13577076	1.72	3.64E-12	1.00	0.00	1.00	PRKAR1B
cg05445326	1.72	3.64E-12	0.72	1.00	0.28	TM4SF19
cg07709358	1.71	3.64E-12	0.77	1.00	0.23	MYOM2
cg17518550	1.70	3.64E-12	0.03	1.00	0.97	C1orf150
cg23300732	1.70	3.64E-12	1.00	0.03	0.97	STK32B
cg04215126	1.69	3.64E-12	0.18	1.00	0.82	FLJ23834
cg10451200	1.69	3.64E-12	0.56	1.00	0.44	GRM5
cg20655068	1.68	3.64E-12	0.67	1.00	0.33	NOTCH4
cg10319893	1.66	3.64E-12	0.90	0.00	0.90	C13orf36
cg12833765	1.63	3.64E-12	0.72	1.00	0.28	ANTXR1
cg09508275	1.61	3.64E-12	0.72	1.00	0.28	OR8H3
cg23859703	1.42	3.64E-12	0.95	1.00	0.05	MYOM2
cg00022911	1.41	3.64E-12	1.00	0.05	0.95	HAPLN4
cg23148701	1.41	3.64E-12	0.97	0.00	0.97	DPY19L2

### 4.3.12 头颈部鳞状细胞癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为头颈部鳞状细胞癌诊断标记物的如下甲基化位点。

表 4-13 头颈部鳞状细胞癌最重要的甲基化诊断位点

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg10655046	2.18	7.79E-10	1.00	0.68	0.32	FOXD4L1
cg12982322	2.16	7.79E-10	1.00	0.58	0.42	WT1
cg05455720	2.16	7.79E-10	0.98	0.12	0.86	MIR124-2
cg03479715	2.00	7.79E-10	0.96	0.00	0.96	FLJ43390
cg06092815	1.98	7.79E-10	1.00	0.18	0.82	SPHKAP
cg15811515	1.98	7.79E-10	1.00	0.30	0.70	CSDAP1
cg01287975	1.98	7.79E-10	1.00	0.28	0.72	TAC1
cg18840956	1.97	7.79E-10	0.88	0.00	0.88	PCSK1
cg06498267	1.96	7.79E-10	1.00	0.14	0.86	HCN1
cg00090261	1.96	7.79E-10	1.00	0.22	0.78	RGS22
cg16288089	1.96	7.79E-10	0.98	0.10	0.88	TAC1
cg01580681	1.94	7.79E-10	1.00	0.14	0.86	HAND2
cg02150135	1.88	7.79E-10	0.94	0.02	0.92	POM121L2
cg22512438	1.73	7.79E-10	0.88	0.02	0.86	TRH
cg12238343	1.73	7.79E-10	0.98	0.00	0.98	RXFP3
cg09516959	1.73	7.79E-10	1.00	0.04	0.96	CRHR2
cg13358636	1.73	7.79E-10	1.00	0.10	0.90	CNTNAP5
cg26074603	1.73	7.79E-10	1.00	0.10	0.90	KCNC2
cg14367229	1.72	7.79E-10	1.00	0.26	0.74	GSX1
cg16483466	1.72	7.79E-10	0.78	1.00	0.22	C20orf186
cg25451456	1.72	7.79E-10	0.78	1.00	0.22	OR56A3
cg06626135	1.71	7.79E-10	0.96	1.00	0.04	PXDNL
cg20051292	1.71	7.79E-10	1.00	0.06	0.94	RGS22
cg23242052	1.71	7.79E-10	0.94	0.00	0.94	EVX2
cg04476531	1.71	7.79E-10	0.78	1.00	0.22	MACROD2
cg06845853	1.71	7.79E-10	0.98	0.02	0.96	WIT1
cg10608596	1.71	7.79E-10	1.00	0.50	0.50	ZNF833
cg14833337	1.70	7.79E-10	1.00	1.00	0.00	FCRL5
cg08080489	1.70	7.79E-10	0.88	1.00	0.12	OPCML
cg00579520	1.66	7.79E-10	1.00	0.10	0.90	TMEM132C

### 4.3.13 肾透明细胞癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为肾透明细胞癌诊断标记物的如下甲基化位点。

表 4-14 肾透明细胞癌最重要的甲基化诊断位点

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg06783197	2.45	6.38E-28	0.29	1.00	0.71	PACS2
cg13289884	2.38	1.06E-26	0.03	0.94	0.92	CDKN1B
cg26986147	2.38	2.32E-27	0.11	1.00	0.89	GAL3ST1
cg07167594	2.23	5.29E-28	0.36	1.00	0.64	DGCR10
cg13749927	2.21	5.29E-28	0.00	0.05	0.05	DDB2
cg12944192	2.20	6.15E-28	0.51	1.00	0.49	KIAA0391
cg08918274	2.19	5.29E-28	0.00	0.00	0.00	VIM
cg05963604	2.19	5.60E-28	0.05	1.00	0.95	APPL1
cg06228828	2.17	5.29E-28	0.26	1.00	0.74	DLEU2
cg07809981	2.15	5.29E-28	1.00	1.00	0.00	AXIN1
ch.11.340609R	2.15	5.29E-28	0.00	0.94	0.94	TEAD1
cg26853536	2.15	8.95E-28	0.03	1.00	0.97	UBC
cg27107970	2.09	6.68E-27	0.01	0.01	0.01	C17orf70
cg16709353	2.00	5.60E-28	0.04	1.00	0.96	TPCN2
ch.17.655115F	1.99	5.39E-28	0.00	0.00	0.00	USP22
ch.15.934240F	1.99	5.39E-28	0.00	0.00	0.00	SNX1
cg07972954	1.99	4.73E-26	0.01	0.04	0.03	TUBB
ch.5.432310R	1.98	5.29E-28	0.00	0.02	0.02	MYO10
cg09320690	1.98	4.82E-26	0.01	0.00	0.01	EHD2
cg03345805	1.98	5.98E-27	0.01	0.04	0.03	INPP1
cg09461545	1.98	5.49E-28	0.44	1.00	0.56	SEL1L3
cg05615230	1.97	8.81E-27	0.01	0.13	0.13	WNK1
cg00009553	1.97	5.29E-28	0.93	1.00	0.07	CDH8
cg06913958	1.96	5.92E-28	0.00	0.66	0.66	BCL10
cg08995609	1.96	5.39E-28	0.01	0.98	0.96	RIN1
cg21330896	1.96	4.70E-27	0.03	0.76	0.73	ZNF395
cg19680850	1.95	5.29E-28	0.00	0.09	0.09	ZC3H12A
cg02117924	1.95	1.02E-27	1.00	1.00	0.00	ITGB2
cg06349174	1.95	5.39E-28	0.00	0.15	0.15	STIM1
cg22057234	1.94	5.29E-28	0.17	1.00	0.83	SPAG4
cg14570632	1.74	7.85E-28	1.00	1.00	0.00	RXRA

### 4.3.14 肾乳头状细胞癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为肾乳头状细胞癌诊断标记物的如下甲基化位点。

表 4-15 肾乳头状细胞癌最重要的甲基化诊断位点

CpGsite	RandomForest	wilcox.pvalue	CaseMR	ControlMR	DFdata	Gene Symbol
cg25986727	2.20	5.68E-14	1.00	1.00	0.00	NUMA1
cg23696248	2.00	5.68E-14	1.00	1.00	0.00	BCL3
cg23264547	1.99	1.14E-13	1.00	1.00	0.00	EXT2
cg23667868	1.99	5.68E-14	1.00	1.00	0.00	KIAA0284
cg18479961	1.98	5.68E-14	1.00	1.00	0.00	MIR671
cg24203709	1.98	5.68E-14	0.00	0.00	0.00	SPTBN5
cg01562537	1.98	5.68E-14	1.00	1.00	0.00	TSPAN18
cg09880291	1.98	5.68E-14	1.00	1.00	0.00	HOXA5
cg06746774	1.98	7.96E-13	1.00	1.00	0.00	KIAA1522
cg20251943	1.98	5.68E-14	1.00	1.00	0.00	DIP2C
cg13651876	1.97	5.68E-14	1.00	1.00	0.00	GRHPR
cg10430690	1.97	5.68E-14	1.00	1.00	0.00	KALRN
cg12085265	1.92	5.68E-14	1.00	1.00	0.00	PRKCZ
cg02276822	1.90	5.68E-14	1.00	1.00	0.00	WDR8
cg12193277	1.73	5.68E-14	1.00	1.00	0.00	CBX7
cg14224313	1.73	5.68E-14	1.00	1.00	0.00	DNMT3B
cg11908453	1.73	1.14E-13	1.00	1.00	0.00	HDLBP
cg00068750	1.73	5.68E-14	1.00	1.00	0.00	TBC1D24
cg18183961	1.73	5.68E-14	1.00	1.00	0.00	RBM47
cg04971812	1.73	5.68E-14	1.00	1.00	0.00	AKT1
cg01303141	1.73	5.68E-14	1.00	1.00	0.00	TBC1D16
cg19023977	1.73	7.39E-10	0.98	1.00	0.02	ELL2
cg11135108	1.73	5.00E-12	1.00	1.00	0.00	PBXIP1
cg03721058	1.73	5.68E-14	1.00	1.00	0.00	HYAL2
cg10336039	1.73	2.11E-11	0.98	1.00	0.02	USP36
cg21007509	1.73	5.68E-14	1.00	1.00	0.00	RAD51L1
cg03582371	1.73	5.68E-14	1.00	1.00	0.00	TBC1D16
cg05323683	1.72	5.68E-14	1.00	1.00	0.00	CDH17
cg08293531	1.72	5.68E-14	1.00	1.00	0.00	LRRK1
cg08141142	1.72	5.68E-14	1.00	1.00	0.00	MTA1

### 4.3.15 肝癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为肝癌诊断标记物的如下甲基化位点。

**表 4-16 肝癌最重要的甲基化诊断位点**

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg19611002	1.91	3.13E-13	0.53	1.00	0.47	ACTL9
cg00981877	2.10	3.55E-15	0.00	0.00	0.00	ATP9A
cg05537653	1.90	3.55E-14	1.00	1.00	0.00	C1QTNF4
cg09361748	1.73	1.78E-14	0.31	1.00	0.69	CALN1
cg04574507	1.93	2.49E-14	0.31	1.00	0.69	CD1B
cg13910460	1.70	6.75E-14	0.63	1.00	0.37	CD200
cg10487221	1.60	8.88E-14	0.37	1.00	0.63	CDH4
cg14988503	1.72	1.78E-14	0.80	0.00	0.80	CDKL2
cg11193281	1.42	1.26E-08	0.43	1.00	0.57	CHRNBT2
cg27434368	1.42	1.78E-14	1.00	1.00	0.00	CILP2
cg11629889	1.72	1.17E-13	1.00	1.00	0.00	CNTNAP1
cg13756251	1.66	1.09E-12	0.71	1.00	0.29	COL5A1
cg13494489	1.60	2.57E-10	1.00	1.00	0.00	DAB2IP
cg25678088	1.69	4.97E-14	0.98	0.78	0.20	DLX6AS
cg10659805	1.95	1.95E-13	0.82	0.00	0.82	DLX6AS
cg12643917	1.67	3.55E-14	1.00	1.00	0.00	ERI3
cg07184316	1.65	1.59E-12	0.53	1.00	0.47	FCRL3
cg18756179	1.60	1.07E-14	0.92	0.00	0.92	FLJ26850
cg11598872	1.69	3.91E-13	0.24	1.00	0.76	FLJ45079
cg08714590	1.66	7.11E-15	1.00	0.73	0.27	FZD1
cg06389444	1.96	4.97E-14	0.18	1.00	0.82	HRNBP3
cg25738340	1.69	8.88E-14	0.43	1.00	0.57	HRNBP3
cg06291867	1.68	7.11E-15	0.92	0.14	0.78	HTR7
cg20365618	1.95	2.69E-11	1.00	1.00	0.00	INPP5F
cg04455646	1.42	1.07E-14	0.00	0.12	0.12	IRAK2
cg19099050	1.68	1.17E-13	0.92	0.00	0.92	LHFPL4
cg00563352	1.41	1.95E-13	0.33	1.00	0.67	LOC285692
cg08305551	1.41	6.00E-13	0.90	0.33	0.57	MAST1
cg14128411	1.70	1.17E-13	0.49	1.00	0.51	MYT1L
cg21072795	2.48	3.55E-15	0.16	1.00	0.84	NCKAP1L

### 4.3.16 前列腺癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为前列腺癌诊断标记物的如下甲基化位点。

**表 4-17 前列腺癌最重要的甲基化诊断位点**

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg24645214	1.72	7.11E-15	0.94	0.08	0.86	RGS20
cg08862890	1.71	7.11E-15	0.98	0.10	0.88	DOCK2
cg06785746	1.69	7.11E-15	0.53	0.04	0.49	WNT3
cg17379325	1.67	7.11E-15	0.73	0.04	0.69	ATP1B2
cg02144933	1.67	7.11E-15	0.71	0.06	0.65	AOX1
cg22003435	1.41	7.11E-15	0.90	0.06	0.84	C2orf88
cg02657832	1.41	7.11E-15	0.59	0.06	0.53	AOX1
cg01965939	1.41	7.11E-15	0.86	0.04	0.82	SH3TC2
cg04380340	1.41	7.11E-15	0.84	0.06	0.78	AOX1
cg26158897	1.41	7.11E-15	0.71	0.06	0.65	ONECUT1
cg03348397	2.23	1.07E-14	0.90	0.06	0.84	RGS20
cg16317273	2.21	1.07E-14	0.88	0.08	0.80	EGFLAM
cg00498024	1.73	1.07E-14	0.90	0.06	0.84	B3GNT8
cg06244497	1.69	1.07E-14	0.88	0.06	0.82	FBXO17
cg15472092	1.42	1.07E-14	0.92	0.08	0.84	KCNH2
cg00489401	1.41	1.07E-14	0.98	0.37	0.61	FLT4
cg16260298	1.41	1.07E-14	0.92	0.08	0.84	CAV2
cg01022219	1.71	1.78E-14	0.98	0.55	0.43	C18orf1
cg06819923	1.68	1.78E-14	1.00	0.82	0.18	ZP2
cg01650776	1.67	1.78E-14	0.98	0.55	0.43	ACCN3
cg08952506	1.66	1.78E-14	0.76	0.06	0.69	AOX1
cg04448487	1.41	1.78E-14	0.88	0.06	0.82	GDAP1L1
cg01893212	1.41	1.78E-14	0.84	0.04	0.80	VWC2
cg01346501	1.41	1.78E-14	0.96	0.18	0.78	NR1D1
cg24512400	1.41	1.78E-14	0.86	0.08	0.78	KLK10
cg14650116	2.38	2.49E-14	1.00	1.00	0.00	NDOR1
cg01940855	1.73	2.49E-14	0.92	0.08	0.84	CHST11
cg13364230	1.73	2.49E-14	0.78	0.06	0.71	GJA4
cg23782734	1.71	2.49E-14	0.84	0.04	0.80	PFKP
cg20927242	1.52	2.49E-14	0.98	0.22	0.76	HLA-F

### 4.3.17 甲状腺癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为甲状腺癌诊断标记物的如下甲基化位点。

**表 4-18 甲状腺癌最重要的甲基化诊断位点**

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg20324356	1.90	8.19E-11	0.00	0.00	0.00	CARS
cg08597067	2.09	8.64E-11	1.00	0.29	0.71	ELOVL5
cg09705456	1.94	9.12E-11	1.00	1.00	0.00	MACROD1
cg04389704	1.99	1.07E-10	0.00	0.00	0.00	FBXL19
cg19440734	2.22	1.13E-10	0.34	1.00	0.66	CAMP
cg24712395	1.93	1.26E-10	1.00	1.00	0.00	CAPN12
cg01665118	1.88	1.26E-10	0.36	1.00	0.64	BTF3L1
cg27320213	1.73	1.33E-10	0.00	0.00	0.00	STAT6
cg03712476	2.14	1.40E-10	0.00	0.00	0.00	SHROOM3
cg23372723	2.38	1.56E-10	1.00	1.00	0.00	ABR
cg24469719	1.98	1.56E-10	0.84	1.00	0.16	RPN2
cg16762684	1.98	1.65E-10	0.04	0.32	0.29	MBP
cg21915100	2.39	1.74E-10	0.79	1.00	0.21	NPC2
cg22016779	1.92	1.74E-10	0.13	0.98	0.86	DNER
cg10565512	1.72	1.93E-10	0.71	1.00	0.29	FOS
cg01619562	1.92	2.04E-10	0.84	1.00	0.16	ITPK1
cg10714284	1.73	2.04E-10	0.00	0.00	0.00	HLA-DMB
cg01115923	1.96	2.15E-10	0.77	1.00	0.23	KIFC3
cg09293559	1.82	2.39E-10	0.82	1.00	0.18	ZFHX3
cg26561570	1.72	2.52E-10	1.00	1.00	0.00	FCGR3B
cg01380194	1.70	2.52E-10	0.52	1.00	0.48	ARAP1
cg03643998	1.73	2.66E-10	0.00	0.00	0.00	C1QTNF1
cg27121309	1.67	2.95E-10	0.55	1.00	0.45	ACER3
cg10944063	2.15	3.11E-10	0.43	0.96	0.54	SCTR
cg25203007	1.72	3.11E-10	0.55	1.00	0.45	GALE
cg03303857	1.98	3.28E-10	0.91	0.41	0.50	GRIK4
cg04237436	1.99	3.64E-10	0.57	1.00	0.43	TRIM44
cg09450153	1.70	3.84E-10	1.00	1.00	0.00	CREB5
cg13324103	1.90	4.26E-10	0.00	0.00	0.00	SVIL
cg04867652	1.83	4.49E-10	0.13	0.98	0.86	SEC14L1

### 4.3.18 子宫内膜癌最重要的甲基化诊断位点

我们选择了最重要的预测变量（随机森林模型）以期待获得最优的诊断/预测效果。同时保证该 CpG 位点为肿瘤差异甲基化位点，并且甲基化频率在肿瘤和配对的正常组织中的差异 $>0.6$ ，以及该 CpG 处于明确的基因区等因素，获得了最有希望成为子宫内膜癌诊断标记物的如下甲基化位点。

表 4-19 子宫内膜癌最重要的甲基化诊断位点

CpG Site	Variable Importance	P-value (wilcox)	MFC	MFN	MFG	Gene Symbol
cg14717557	1.73	1.86E-09	1.00	0.30	0.70	SIM1
cg26282792	1.73	1.86E-09	1.00	0.17	0.83	ZSCAN1
cg26728422	1.72	1.86E-09	0.00	0.40	0.40	UNKL
cg15999077	1.72	1.86E-09	0.13	1.00	0.87	SLC6A3
cg23450509	1.72	1.86E-09	0.00	1.00	1.00	CPNE4
cg19603744	1.71	1.86E-09	1.00	0.97	0.03	ZSCAN1
cg09624807	1.71	1.86E-09	1.00	1.00	0.00	NLRC5
cg18569734	1.70	1.86E-09	1.00	0.03	0.97	NR2E1
cg26790247	1.70	1.86E-09	1.00	0.03	0.97	ZIC4
cg00164678	1.69	1.86E-09	0.27	1.00	0.73	C3P1
cg20552468	1.69	1.86E-09	1.00	0.40	0.60	SLC38A10
cg02322373	1.42	1.86E-09	0.93	0.00	0.93	VAX2
cg21359747	1.42	1.86E-09	1.00	0.20	0.80	ALDH1A3
cg04865691	1.42	1.86E-09	1.00	0.07	0.93	SOX1
cg07982896	1.41	1.86E-09	0.33	1.00	0.67	CACNA1A
cg03830329	1.41	1.86E-09	1.00	0.00	1.00	SIM1
cg27441409	1.41	1.86E-09	0.33	1.00	0.67	S100A7L2
cg23290344	1.41	1.86E-09	1.00	0.10	0.90	NEFM
cg02267488	1.41	1.86E-09	0.67	1.00	0.33	DCK
cg12298598	1.41	1.86E-09	0.93	1.00	0.07	DSCAM
cg09979641	1.41	1.86E-09	0.33	1.00	0.67	KNDC1
cg14314744	1.41	1.86E-09	1.00	0.07	0.93	SIM1
cg05292478	1.41	1.86E-09	0.03	1.00	0.97	SMURF1
cg19611002	1.41	1.86E-09	0.47	1.00	0.53	ACTL9
cg27099358	1.41	1.86E-09	0.83	1.00	0.17	LOC146336
cg17953764	1.41	1.86E-09	1.00	0.70	0.30	ZAR1
cg07553761	1.41	1.86E-09	1.00	0.37	0.63	TRIM59
cg25481630	1.41	1.86E-09	0.30	1.00	0.70	PTPRN2
cg21069434	1.41	1.86E-09	0.30	1.00	0.70	MYT1L
cg07992308	1.41	1.86E-09	0.20	1.00	0.80	KCNA7

## 4.4 结论

本研究首先采用包括 11 种人类肿瘤 1274 个癌和对应癌旁组织的 HM450K 甲基化芯片数据，对人类肿瘤的甲基化谱式进行了描述。HM450K 芯片中 34 万个甲基化位点在整体数据的相关性绝大多数为中低相关，说明了甲基化位点在不同状态的样本中呈现很高的复杂度。样本相关性显示同一组织来源的样本具有较高的相关性。肿瘤样本之间的相关性远低于同组织来源的正常样本，印证了肿瘤样本相对于正常组织之间膨胀了的异质性。主成分分析显示前 10 个主成分即可解释泛癌数据（1274 样本，349,049 探针）55.74% 的总变异，当选取前 120 个主成分时，即可解释数据总变异的 80%，说明该数据的方差为有效规律变量造成的而非大量噪音信号的叠加。多维尺度分析在低维空间对样本之间的关系分析发现 DNA 甲基化可以反应肿瘤组织之间的距离（近似性），相近起源的肿瘤在低维空间距离较近。样本个体聚类分析显示相同类型的肿瘤样本相互聚集；相同类型的正常组织相互聚集，说明全基因组 DNA 甲基化谱式可以忠实地反应样本之间的相似性。肿瘤样本群体水平的聚类分析显示肺腺癌和肺鳞癌，肾透明细胞癌和肾乳头状细胞癌，前列腺癌和乳腺癌首先聚集在一起。泛癌样本差异甲基化位点分析显示只有 42 个位点在 11 种肿瘤中均呈现显著性差异。采用随机森林预测算法，我们对 11 种肿瘤和正常对照的 22 种状态的样本进行判定，结果显示采用 25 个预测变量即可实现对 22 个状态的较好预测，灵敏性，特异性和准确性分别为 72%，87% 和 80%。当预测变量数达到 75 个时预测模型的灵敏性，特异性和准确性可以到到 86%，94% 和 90%。因此可以推测即便对多种肿瘤同时预测，所需要的甲基化位点数也不会超过三位数。

## 4.5 讨论

本文从基于 DNA 甲基化芯片的泛癌数据，包括：肾透明细胞癌、浸润性乳腺癌、甲状腺癌、头颈部鳞状细胞癌、前列腺癌、肝癌、肾透明细胞癌、肺腺癌、结肠癌、子宫内膜癌、肺鳞癌，对多肿瘤之间的 DNA 甲基化特性进行了简单的分析。4.3.1 部分显示全基因组 CpG 位点之间的相关性大多数为中低相关位点，说明 DNA 甲基化甲基化信号在不同肿瘤之间具有较大的差异，从而为区分多种肿瘤提供了基础。而通过 4.3.7 部分的分析可以看出对于同一肿瘤相邻 CpG 位点之间的甲基化状态的相关性较大，并且肿瘤组织之间的相关性程度高于正常组织，提示肿瘤组织对这些 CpG 位点需要进行同向的变化，从而促进肿瘤的产生。

4.3.3 及 4.3.4 部分采用主成分分析及多维尺度分析对泛癌数据的主要方差和样本之间的关系进行了简单的分析和展示。在两个分析中都可以看出 DNA 甲基化谱式可以反应肿瘤或正常以及反应样本的组织来源。

4.3.5 采用个体及群体的方式对泛癌样本进行了聚类分析，结果显示相同肿瘤样本首先聚类，只有组织来源相近的肿瘤逐渐聚类。说明全基因组 DNA 甲基化能够反应样本的主要信息。

4.3.6 部分对泛癌数据差异甲基化位点的比例进行了分析，发现不同的肿瘤的差异甲基化比例具有较大差异。同时不同亚型了的相近肿瘤如肺腺癌和肺鳞癌等其差异甲基化比例也存在加大差异，为 DNA 甲基化对不同亚型的分类奠定了基础。值得注意的是由于不同的肿瘤具有不同的样本数，导致不同的肿瘤中对差异甲基化位点统计的 Power 有所不同。所以这部分结论尚待更进一步的分析。

4.3.8 部分采用随机森林预测模型对泛癌样本同时预测进行了探索。随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟。对于有不同级别的属性的数据，级别划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值是不可信的。因此我们同时将甲基化信号转变为高甲基化信号 ( $\text{beta} > 0.8$ )，杂合甲基化信号 ( $0.8 > \text{beta} > 0.3$ ) 和纯合低甲基化信号 ( $\text{beta} < 0.3$ )。再次分析发现，上述结论基本没有太大变化，25-50 个甲基化位点即可对上述 11 种肿瘤进行很好的预测。此外分析发现非特异变量过滤的预测效果显著高于特异性变量过滤，不论采用类别特异指数还是差异甲基化位点的方式都会降低预测的准确度。这说明在临床诊断模型/预测模型的建立过程中采用有效地变量选择方法至关重要。

4.3.9 至 4.3.18 分别对上述 11 种肿瘤的差异甲基化位点进行了列举，为临床科研工作者提供一些可以重点进行探索的甲基化位点。本部分位点具有最大的分类贡献度及最大的癌-癌旁甲基化率差异等特征，具有较高的对肿瘤进行预测，诊断的价值。

由于本部分分析完全基于全基因组芯片数据，而芯片数据探针信号之间存在一定的互相干扰。同时泛癌样本的临床及流行病学资料没有在本部分的考虑之中，比如吸烟，年龄，性别等，而这些因素对 DNA 甲基化具有显著的影响，所以本部分的一些结论需要更多更细致数据的验证。

表 4-20 高肿瘤相关分值基因 (TRS&gt;=10) 的 Gene ontology 分析

Category	Term	Count	PValue	Fold Enrichment	Benjamini
GOTERM_MF_FAT	GO:0003700~transcription factor activity	17	2.16E-05	3.38E+00	3.54E-03
GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	13	4.74E-05	4.15E+00	3.88E-03
GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	13	4.74E-05	4.15E+00	3.88E-03
GOTERM_CC_FAT	GO:0005886~plasma membrane	35	0.0001449	1.77E+00	2.16E-02
SP_PIR_KEYWORDS	glycoprotein	38	0.0001629	1.78E+00	2.59E-02
SP_PIR_KEYWORDS	calcium	13	0.0005318	3.28E+00	2.81E-02
SP_PIR_KEYWORDS	developmental protein	13	0.0004057	3.38E+00	3.21E-02
GOTERM_BP_FAT	GO:0048666~neuron development	10	5.28E-05	5.70E+00	3.58E-02
SP_PIR_KEYWORDS	cell adhesion	9	0.0010791	4.32E+00	4.25E-02

150 个 TRS>=10 的位点，对应 113 个基因。

## 4.6 参考文献

- [1] Goelz, S.E., B. Vogelstein, S.R. Hamilton, and A.P. Feinberg. *Hypomethylation of DNA from benign and malignant human colon neoplasms* [J]. Science, 1985. **228**(4696);187-190.
- [2] Cokus, S.J., S. Feng, X. Zhang, Z. Chen, et al. *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning* [J]. Nature, 2008. **452**(7184);215-219.
- [3] Li, Y., J. Zhu, G. Tian, N. Li, et al. *The DNA methylome of human peripheral blood mononuclear cells* [J]. PLoS biology, 2010. **8**(11);e1000533.
- [4] Xiang, H., J. Zhu, Q. Chen, F. Dai, et al. *Single base-resolution methylome of the silkworm reveals a sparse epigenomic map* [J]. Nature biotechnology, 2010. **28**(5);516-520.
- [5] Zhao, Y., S. Guo, J. Sun, Z. Huang, et al. *Methylcap-seq reveals novel DNA methylation markers for the diagnosis and recurrence prediction of bladder cancer in a Chinese population* [J]. PLoS One, 2012. **7**(4);e35175.
- [6] Taiwo, O., G.A. Wilson, T. Morris, S. Seisenberger, et al. *Methylome analysis using MeDIP-seq with low DNA concentrations* [J]. Nature protocols, 2012. **7**(4);617-636.
- [7] Schillebeeckx, M., A. Schrade, A.K. Lobs, M. Pihlajoki, et al. *Laser capture microdissection-reduced representation bisulfite sequencing (LCM-RRBS) maps changes in DNA methylation associated with gonadectomy-induced adrenocortical neoplasia in the mouse* [J]. Nucleic acids research, 2013. **41**(11);e116.
- [8] Cline, M.S., B. Craft, T. Swatloski, M. Goldman, et al. *Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser* [J]. Sci Rep, 2013. **3**;2652.
- [9] Szyf, M. *The early-life social environment and DNA methylation* [J]. Clin Genet, 2012. **81**(4);341-349.
- [10] Ben-Avraham, D., R.H. Muzumdar, and G. Atzmon. *Epigenetic genome-wide association methylation in aging and longevity* [J]. Epigenomics, 2012. **4**(5);503-509.
- [11] Rakyan, V.K., T.A. Down, D.J. Balding, and S. Beck. *Epigenome-wide association studies for common human diseases* [J]. Nature reviews. Genetics, 2011. **12**(8);529-541.

- [12]Hong, L. and N. Ahuja. *DNA methylation biomarkers of stool and blood for early detection of colon cancer* [J]. Genet Test Mol Biomarkers, 2013. **17**(5);401-406.
- [13]de Fraipont, F., D. Moro-Sibilot, S. Michelland, E. Brambilla, et al. *Promoter methylation of genes in bronchial lavages: a marker for early diagnosis of primary and relapsing non-small cell lung cancer?* [J]. Lung Cancer, 2005. **50**(2);199-209.
- [14]Risbridger, G.P., I.D. Davis, S.N. Birrell, and W.D. Tilley. *Breast and prostate cancer: more similar than different* [J]. Nature reviews. Cancer, 2010. **10**(3);205-212.
- [15]Yanai, I., H. Benjamin, M. Shmoish, V. Chalifa-Caspi, et al. *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification* [J]. Bioinformatics, 2005. **21**(5);650-659.
- [16]靳文菲, 混合人群和疾病基因的自然选择研究[D],2011, 中国科学院上海生命科学研究院计算生物所: 上海.
- [17]Laster, L. *Statistical background of methods of principal component analysis* [J]. Journal of periodontology, 1967. **38**(6);Suppl:649-666.
- [18]van der Voet, H. and J.P. Franke. *A discussion of principal component analysis* [J]. Journal of analytical toxicology, 1985. **9**(4);185-188.
- [19]徐书华, 高密度常染色体 SNPs 揭示的现代人群遗传结构[D], 2006, 复旦大学: 上海. p. 45-46.
- [20]Beals, R., D.H. Krantz, and A. Tversky. *Foundations of multidimensional scaling* [J]. Psychological review, 1968. **75**(2);127-142.
- [21]Carroll, J.D. and P. Arabie. *Multidimensional scaling* [J]. Annual review of psychology, 1980. **31**;607-649.

## 第五章 甲基化特异富集测序技术鉴定甲基化肿瘤标记物

虽然基于高密度 DNA 甲基化芯片技术对肿瘤标记物筛查的方法具有：位点特异，价格低廉等优势，但是在标记物筛查早期及诊断模型建立阶段，芯片技术存在容易遗漏潜在的重要的诊断标记物的问题。基于全基因组测序技术的甲基化特异结合富集深度测序方法可以弥补 DNA 甲基化芯片的这一缺点。本部分研究，我们尝试甲基化特异富集结合深度测序技术探索可以应用于中国人群的膀胱癌诊断和复发预测的新型 DNA 甲基化标志物。

目前临幊上常用的膀胱癌检测手段主要有：膀胱镜检查、尿脱落细胞形态学检查、B 超。膀胱镜检查是一种有创性检查，并且由于人为性高、检查时存在死区、具有高达 10%-40% 的误诊率[1]，而尿脱落细胞形态学检查虽然可达到较高的特异性，但是其敏感性不高，且仅对低分化的膀胱癌检出率较高[2, 3]。所以寻找可以高敏感性、高特异性地诊断早期膀胱癌、监测复发和判断预后的生物学标志物称为一个迫切的问题。

本研究尝试采用甲基化 DNA 结合结构域 (MBD) 富集技术结合第二代高通量测序(methylCap-Seq)的方法，建立常用膀胱癌全基因组 CpG 岛异常甲基化谱式。基于甲基化差异分析得到了 1627 个的位于基因启动子区域的膀胱癌显著异常 CpG 岛甲基化区域。对 1627 个谱式中差异最显著的 104 个高度甲基化的位点，按照多阶段生物标记物筛选方法，在一系列的独立的膀胱癌尿液样本中，利用 MSP 技术，以逐步缩小候选的甲基化位点，并逐步加大临床膀胱癌样本量，最终寻找到了 9 个显著的膀胱癌甲基化标记物。我们在后续的研究中仔细地分析了这 9 个甲基化标记物与膀胱癌诊断，预后复发，肿瘤分化进展的关系。

### 5.1 研究背景

膀胱癌(Bladder Cancer, BC)是泌尿系统中最常见的恶性肿瘤，如表 5-1 所示，其发病率和死亡率中在所有癌症中分别排第八位和第十位[4]。在中国，膀胱癌的发病率呈现持续上升的趋势[5]。从发病年龄进行统计可以发现膀胱癌发病率随年龄增长而增加，平均诊断年龄约为 60 岁。在性别差异方面，男性发病率是女性的 3 倍左右[6]。流行病资料显示吸烟和暴露于致癌物质是膀胱癌发病的

主要危险因素[7]。对于膀胱癌约 75-80%的新发病例发生表浅或原位癌病变阶段，其余 20-25%则处于预后较差的晚期阶段，如图 5-1 所示。然而，即使在浅表肿瘤，采用目前的治疗手段，只有 20%是可以治愈的。约 60-70%的患者会在 5 年内复发，此外其余的 10-20%会进展到转移性肿瘤。为此，对于膀胱癌病人，疾病复发的实时监测具有重要意义[8]。膀胱镜检查是最常见的诊断流程，具有较高的灵敏度（Sensitivity, SN）和特异性（Specificity, SP）。然而膀胱镜检查需要操作者的熟练的操作技巧，此外膀胱镜检查是一种侵入性（有创, invasive）的诊断方法，不恰当的膀胱镜检查反而会引起膀胱癌的扩散，从而大大地降低了其作为筛选工具的应用价值。膀胱癌临床工作者迫切需要一种早期，非侵入性的膀胱癌检测和预后实时检测工具。

**表 5-1 2014 年主要肿瘤的新发病例和死亡病例的估计(美国)**

Cancer Category	Estimated New Cases				Estimated Deaths			
	Both	Male	Female	Rank	Both	Male	Female	Rank
Breast	235030	2360	232670	1	40430	430	40000	3
Prostate	233000	233000		2	29480	29480		5
Lung & bronchus	224210	116000	108210	3	159260	86930	72330	1
Colonb	96830	48450	48380	4	50310	26270	24040	2
Skin	81220	46630	34590	5	12980	8840	4140	14
Lymphoma	79990	43340	36650	6	20170	11140	9030	8
Melanoma-skin	76100	43890	32210	7	9710	6470	3240	18
<b>Urinary bladder</b>	<b>74690</b>	<b>56390</b>	<b>18300</b>	<b>8</b>	<b>15580</b>	<b>11170</b>	<b>4410</b>	<b>10</b>
Non-Hodgkin lymphoma	70800	38270	32530	9	18990	10470	8520	9
Kidney & renal pelvis	63920	39140	24780	10	13860	8900	4960	13
Thyroid	62980	15190	47790	11	1890	830	1060	32
Uterine corpus	52630		52630	12	8590		8590	19
Leukemia	52380	30100	22280	13	24090	14040	10050	6
Pancreas	46420	23530	22890	14	39590	20170	19420	4
Oral cavity & pharynx	42440	30220	12220	15	8390	5730	2660	20
Liver	33190	24600	8590	16	23000	15870	7130	7
Myeloma	24050	13500	10550	17	11090	6110	4980	15
Stomach	22220	13730	8490	18	10990	6720	4270	16
Ovary	21980		21980	19	14270		14270	12
Acute myeloid leukemia	18860	11530	7330	20	10460	6010	4450	17
Esophagus	18170	14660	3510	21	15450	12450	3000	11

注：表格由 Cancer Statistic(2014)整理而来[9]。其发病率和死亡率中在所有癌症中分别排第八位和第十位[4]

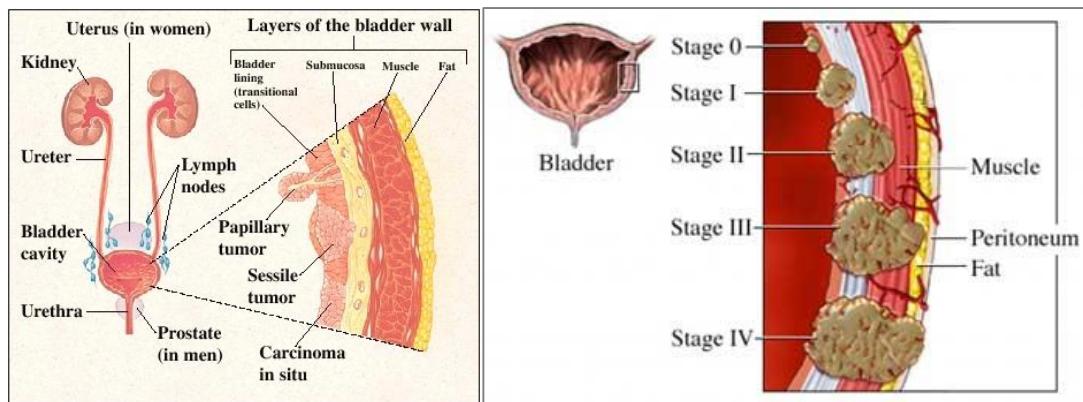


图 5-1 膀胱癌器官结构示意图

注：这两图分别介绍膀胱癌的基本情况：人体位置，膀胱器官基本结构，膀胱癌的 TNM 分期及其侵袭情况。左图来自 goodhealthhub 网站，右图来自 urologycenter 网站。

人类基因组上的表观遗传修饰，可以通过对外界环境的响应，对人类基因组信息进行相应的修饰和调控。因此表观遗传学修饰对于由于环境因素起到重要作用的复杂疾病的致病因素的研究具有重要意义。表观遗传学调控对于高等真核生物的正常生长[10]，发育[11]，衰老具有重要的调控机制。而不良外界环境的长期暴露可以导致表观遗传修饰的紊乱，进而导致很多复杂疾病的发生。通过全基因组遗传学和表观遗传学的研究，我们对于肿瘤的致病机理已经有了基本的认识。

肿瘤是一种涉及遗传学和表观遗传学异常的复杂疾病。其中遗传学和表观遗传学在肿瘤发病的不同阶段中发挥作用，并且可能在一些阶段相互作用，最终导致了肿瘤的不可逆病变[12]。DNA 甲基化是表观遗传修饰中最重要的组成部分，也是研究最为深入的一种表观修饰。肿瘤的基因组 DNA 甲基化谱式具有全局低甲基化[13, 14]和局部区域的高甲基化[15]的基本特征。最近的进展已经表明，抑癌基因的异常高甲基化是一种新兴的癌症的诊断和预后生物标记物。

在膀胱癌中一些异常甲基化的基因已被确定，同时这些标记物的性能在尿液中也得到了一系列的评估[16]。研究表明使用多个基因甲基化的标记物组合，在以组织和尿液样本中对膀胱癌进行诊断和预后预测具有很大的优势。然而当标记物组合中的基因数目过多时，又会影响其向临床转化的效率。过多的标记物会提高检测的成本，降低检测的稳定性。最佳的标记物组合应该能够实现：最少的标记物数目和最大的诊断灵敏性，特异性和准确性。采用全基因组甲基化测序技术筛选靶点，可以在一定程度上达到诊断和预后标记物的最优组合。虽然目前已经有很多基于全基因组甲基化测序的技术对膀胱癌异常甲基化标记物进行探索，但是，到目前为止尚没有研究采用全基因甲基化测序，对膀胱癌异常甲基化位点筛

查并结合临床样本，建立最优诊断和预后模型的报道。为此本文将采用这种研究策略，探索并建立膀胱癌甲基化肿瘤诊断预后标记物[17, 18]。

MethylCap-Seq 是一个最新开发的 DNA 甲基化的全基因组分析技术。这项技术采用甲基化 CpG 结合域蛋白 MeCP2 对甲基化的 DNA 片段进行捕获，然后对捕获的 DNA 片段采用二代测序的技术对甲基化的区域进行分析。采用梯度的盐浓度对包含有甲基化片段的甲基化结合蛋白的洗脱，可以获得不同甲基化程度的 DNA 片段，进而可以对甲基化的程度进行分析。对比甲基化芯片和甲基化特异性酶切技术，MethylCap-Seq 技术可以无偏地对全基因组范围内的甲基化进行分析[19]，且对比发现不论从测序覆盖度还是从测序深度，MethylCap-seq 技术都比相似技术 MeDIP-seq 或 RRBS 都要高。

在本项研究中，我们首先采用 MBD MethylCap-Seq 技术取得膀胱癌细胞系的全基因组甲基化谱及膀胱癌异常甲基化位点。之后，采用一批膀胱癌患者尿液的 DNA 样品，对最显著异常的前 100 位甲基化位点进行筛查。有潜力的标记物数目逐渐降低。最后，我们获得的一组可用于膀胱癌早期、非侵入性的检测和预后监视的 DNA 甲基化标志物组合。

## 5.2 材料和方法

### 5.2.1 患者和对照样品信息

所有的膀胱癌患者和对照均来自两所医院：复旦大学附属中山医院泌尿外科和上海市静安区石门二路街道社区卫生服务中心在 2006 年至 2009 年间收集的样本。本研究在获得复旦大学医学院机构审查委员会及中山医院的伦理审核后开展，所有病理组织，尿液及正常对照的获得均与病人签署了相关的知情同意书。共收集 212 例确诊的膀胱癌病人、同期 41 例非膀胱癌性泌尿病变和 149

例正常个体的尿液标本。为了研究甲基化对膀胱癌手术后的评估，还收集了 21 例膀胱癌患者手术前和手术后的配对尿液。此外我们还收集了 48 例患有转移性恶性膀胱肿瘤患者的尿样。

按照美国联合委员会确定的癌症指南对膀胱癌患者样本的肿瘤淋巴结转移 (TNM) 分期进行分类[21]。

样品采集和保存过程如下：50 毫升新鲜尿液样品，3,000 rpm 离心 10 分钟。弃去上清液，并且，用 1×磷酸盐缓冲盐水 (PBS) 将沉淀物洗涤一次，然后立即冷冻于 -80°C，保存备用。样本相关资料如表 5-2 所示。

表 5-2 尿液样本对于的膀胱癌病人及正常对照的临床病理资料

	Bladder cancer n = 212	Normal control n = 149	Nontumor urinary lesions n = 41	Surgery resect n = 21	Clinical cystoscope n = 48
<b>Gender</b>					
F	46	71	16	3	14
M	166	78	25	17	34
<b>Age</b>					
-30	1	8	1	0	0
31-40	1	14	4	1	2
41-50	27	29	8	4	1
51-60	41	23	6	2	4
61-70	44	26	9	4	6
71-	98	49	13	10	35
Range	29-91	22-90	16-89	35-88	31-90
Median	69	61	61	69	72
Mean±SD	66.85±12.74	59.77±17.14	60.24±17.02	65.33±14.44	68.85±13.44
<b>Grade</b>					
I	75				
II	120				
III	25				
<b>Stage</b>					
Oa	3				
I	134				
II	63				
III	7				
IV	5				
<b>Relapse</b>					
Primary	157				
Recurrency	55				
Cystitis			17		
urinary tract infection				13	

kidney stones	5
prostatitis	3
Nephritis & nephrotic syndrome	3

### 5.2.2 细胞系和正常膀胱粘膜组织

膀胱癌细胞系 T24 (ATCC 编号: HTB-4) 和 5637 (ATCC 编号: HTB-10), 购自美国典型培养物保藏中心 (ATCC, 马纳萨斯, 弗吉尼亚州)。将细胞采用含有 10% 胎牛血清 (FBS) 的 L-DMEM 培养基在 5% CO<sub>2</sub> 的条件下 37°C 下的湿润培养箱培养至指数增长期。将细胞刮收获, 将细胞沉淀物用 1×PBS 清洗两次。两个正常膀胱粘膜组织 (BM1 和 BM2) 通过健康的器官捐献者捐赠获得。

本在研究中, 我们尝试采用混合建库的方式进行混合甲基化谱式的分析, 将这两个基底细胞癌混合构建成 BCC 库, 两个正常膀胱粘膜组织混合构成 BM 库。

### 5.2.3 MBD-methylCap 测序及其分析

细胞系 DNA 的提取: 1) 当细胞满度达到 70% 以上时即可用于基因组 DNA 的抽提。2) 弃去培养液, 用适量的 PBS 洗 1-2 遍。然后用刮棒将细胞转刮下, 并移于至 1.5ml EP 管中。3) 1000rpm 离心 3 分钟, 弃去上清, 用震荡器震荡使沉淀松动。4) 然后加 400μl 含蛋白酶 K 终浓度 200 μg/ml 的细胞裂解液, 轻柔混匀至溶液澄清, 若溶液没有变澄清则需补加细胞裂解液, 然后 37°C 过夜。5) 等体积酚/氯仿抽提两次。6) 取上清加入 1/3 体积的 7.5M 醋酸铵和 3 倍体积的无水乙醇沉淀。7) 加适量 70% 酒精洗涤沉淀。最后加适量 RTE 65°C 10 分钟溶解沉淀, 更详细关于细胞及分子生物学的操作请参考徐向红博士的博士学位论文[20]。

从 5637 和 T24 细胞取等量的 DNA 相混合, 形成了 BCC 库, 并从将等量的 BM1 与 BM2 细胞 DNA 相结合, 形成 BM 库中。在 1.5mL 离心管中, 1.5 微克的混合 DNA 样本(BCC 或 BM)于 100μl TE 缓冲液中超声打断处理, 得到碱基长度为 200-300bp 的 DNA 片段。二代测序所涉及的末端修复、腺苷酸化处理和测序头的连接步骤, 如我们之前的文章所介绍[21]。采用商业化 MethylMiner™ 甲基化 DNA 富集试剂盒(Invitrogen 公司, 卡尔斯巴德, 加利福尼亚州, 美国)富集基因组中甲基化的 DNA 片段。原理图如图 5-2 所示。

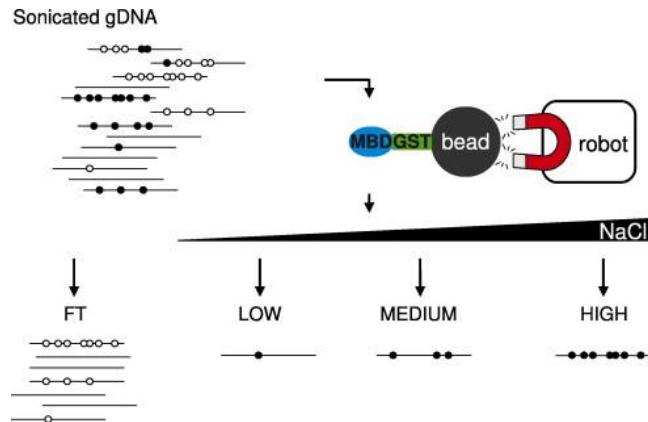


图 5-2 MBD-methylCap-seq 甲基化谱式的构建的原理简图

注：基因组 DNA 经过超声打断，MBD 可以对甲基化的 DNA 片段进行捕获。通过不同浓度的 NaCl 洗脱，可以得到不同甲基化程度的 DNA 片段。MBD 蛋白与 DNA 片段的结合力正比于 DNA 片段中甲基化的程度。低浓度 NaCl 对应低甲基化程度的 DNA 片段。高 NaCl 可以把高甲基化程度的 DNA 片段洗脱下来。该图选择自 Brinkman 博士发表在 Method 上的文章[19]。

对于每个测序文库，根据制造商的方案，每个甲基化测序库需要 1.2ug DNA 量。采用梯度 NaCl 进行洗脱，收集最后的两个洗脱组分，高浓度的洗脱物(1 M 和 2M 氯化钠)对应高度甲基化的 DNA 片段。试剂盒内置的 spike DNA 对照可用来对富集过程进行准确度监控。将回收的 DNA (纳克数量级) 采用 Qubit<sup>TM</sup> 仪器 (Invitrogen 公司) 进行定量，之后采用 12 个循环的 PCR 对回收 DNA 进行扩增，以获得足够的测序底物 (微克数量级) 进行深度测序。最后，1 微克的 PCR 产物采用基因组分析仪 II (Illumina 公司，圣地亚哥，CA) 来产生 75 碱基长的单末端测序数据，更加详细的介绍可以查看周小羽硕士的学位论文的第二部分 [22]。

Fastq 格式的测序结果由二代测序获得。然后采用 BWA 对测序数据进行 alignment 到 hg19 人类基因组的参考序列上[23]。采用 Picard 删除测序产物中的 PCR 重复。虽然有很多方法可以对捕获测序中 peak 进行识别，常用的如 MACS[24]，QuEST [25]，SISSRs [26]，PICS [27]，本文仅采用了使用最广泛的 MACS 对高甲基化区域进行定位。Hg19 基因组从 NCBI 按照每个染色体进行下载，然后合并得到人类全长的参考序列。人类基因(Ref gene)，CpG 岛等相关注释信息以 bed 格式从 UCSC 进行下载。本研究产生的甲基化测序数据已经上传到 GEO 公共数据库，以供大家分析使用(GSE33839)。

### 5.2.3.1 MACS2 对甲基化富集区域预测

MACS 被广泛应用于 ChIP-seq 对转录因子结合热点(高 reads 富集区域)的预测。MACS 方法通过两步法对结合热点进行预测，因为我们在我们的研究中该“结合热点”实际代表我们研究的高甲基化区域，为此后文中我们用高甲基化区域代替结合热点。MACS 模型假设 MBD methylCap 捕获的 DNA 片段在基因组上的分布事件服从二项分布，具体程序的执行时，由于 N(测序得到的 read 总数目)很大，P(片段落在特定染色体区域)很小，可以近似用泊松分布替代二项分布。特定位点被捕获测序的平均次数(lamda)是泊松分布的唯一参数。即假设，N 是测序得到的 read 总数目，L 是单个 read 的长度，S 是基因组的大小，X 是特定位点被测序到的次数，那么：

$$X \sim Pois(\lambda) \quad (1)$$

$$\lambda = \frac{L \times N}{S} \quad (2)$$

根据泊松分布，即可判断实际情况下每个位点是否为异常富集，及是否为异常高甲基化位点，并因此计算置信概率 P。MACS 还有其他一些优点以提高热点预测区域的准确度。比如采用二项分布对可疑 PCR 造成的重复测序进行剔除，以降低 PCR 过程导致的可能的假高甲基化区域。由于测序、mapping 过程内在的偏好性，以及不同染色质间的差异性，采用全基因组的数据对 peak 进行估计会有很大的 bias。实际上，MACS 对某个碱基进行假设检验时，并不是考虑全基因背景信息，而是只考虑该碱基附近的染色质区段，此时，上述公式中 N 表示附近特定长度区间内的 read 的数目，S 被置为该特定长度。MACS 的详细解释和使用方法请参考如下两篇文献[24, 28]。

### 5.2.4 MSP 和 BSP 方法介绍

用 MSP(methylation-specific PCR)的方法进行 DNA 甲基化的检测，可以得到细胞，组织或体液中甲基化的定性结果。该方法有很高的敏感性，能够从 10,000 个正常细胞中，检出 1-10 个肿瘤细胞[29]。亚硫酸氢盐测序 PCR(BSP)不同于 MSP，其对甲基化检测的灵敏性较低，但准确性远远高于 MSP，特别是可以对一定区域内所有位点的甲基化状态进行准确的评估。BSP 是甲基化状态测定的金标准，通过对待评估 DNA 区域的无偏扩增，从而对 DNA 区域的 DNA 甲基化状态进行准确的显示。亚硫酸氢盐转化和 PCR 分析，如先前所述进行[30]。在本研究中亚硫酸氢盐测序 PCR(BSP)和甲基化特异性 PCR(MSP)用于在全基因组甲基化谱式的验证阶段。

### 5.2.4.1 MSP 原理

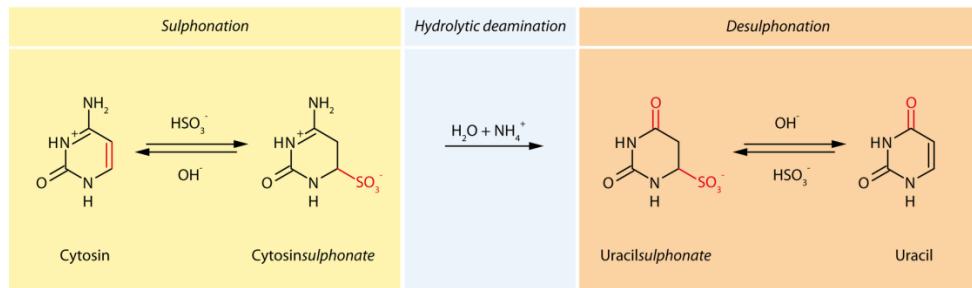


图 5-3 亚硫酸盐介导的胞嘧啶到尿嘧啶转变

注：本图摘自 Handbook of Epigenetics

如图 5-3 所示，亚硫酸盐可以在一定的条件下将胞嘧啶到尿嘧啶，而甲基化的胞嘧啶可以抵制亚硫酸盐的这种转变。因此胞嘧啶的甲基化状态可以通过检测亚硫酸盐处理后胞嘧啶的状态进行确定。如果处理后保持胞嘧啶，则表示该胞嘧啶为甲基化的胞嘧啶；而如果胞嘧啶转变为 T (PCR 过程中 U 与 A 配对，从而在 PCR 后显示为 T)，则原胞嘧啶为非甲基化的胞嘧啶。这种情况下，MSP 的研究策略是可以针对待研究的 CpG 位点，假设其为甲基化状态(M)或非甲基化状态(U)，而设计两种结合引物。利用两种引物对亚硫酸盐转变后的模板进行扩增。如果待研究的 CpG 位点为甲基化状态，则甲基化引物可以成功对模板进行扩增，去甲基化引物因为无法和模板进行互补配对，从而不能扩增；反之也是一样。为了提高 MSP 的准确性，一般要求待研究的 CpG 位点的胞嘧啶位于引物 3' 末端，以提高引物的特异性。并且要求退火温度不能过低，否则会出现假甲基化结果。MSP 一般要求扩增长度控制在 150-300bp 之间。过短不易通过凝胶电泳进行观察，过长可能造成不易扩增。

### 5.2.4.2 BSP 原理

BSP 也是建立在亚硫酸盐转变的基础之上，因为人类成熟体细胞基因组中的甲基化绝大多数发生在 CpG 二核苷酸对上。和 MSP 不同，BSP 不需要对特定位点的甲基化状态进行假设，而是选择一段不存在 CpG 位点的区域进行扩增引物的设计。对于不含有 CpG 位点的 DNA 区域，所有的胞嘧啶(C)会被转变为胸腺嘧啶(T)，从而使得具有无偏的引物可以目标区域进行扩增。通过对扩增区域的 CpG 是否转变为 TpG 就可以确定扩增内每个 CpG 位点的甲基化状态。因此 BSP 一般可以观察 100-400bp 范围内所有 CpG 位点的甲基化。

### 5.2.4.3 甲基化检测流程

BSP 和 MSP 的引物设计采用 MethPrimer 技术[31], MethPrimer 是一个在线甲基化引物设计网站。详细的引物信息如表 5-2 和表 5-3 所示。为了降低 MSP 的假阳性率, MSP 产物都经过克隆在 sanger 测序进行再验证。经过甲基转移酶 M. SSS I (NEB) 处理的 5637 和 T24 的 DNA 被用作高甲基化阳性对照。纯水被用作无模板对照, 用于对测序过程的控制。PCR 的扩增产物经过凝胶纯化并克隆到 PBS-T II 载体上(天根生化科技, 北京, 中国)。至少 5 个克隆单独测序, 以确定该对象轨迹的甲基化模式。该 BSP 甲基化百分比计算为甲基化的 CG 胞嘧啶除以在所分析的扩增子的 CG 总数量。亚硫酸氢盐测序更多细节可参考我们前期的文章[32]及马克龙博士[33]和孙晋枫硕士[34]的学位论文所述。引物由上海赛百盛生物科技有限公司合成。

#### MSP 及 BSP 的反应体系与条件

反应体系:

BSP 反应组分	体积
Template	4.0μl
10×Hot-start Taq polymerase buffer	2.0μl
10mM dNTP	0.5μl
2.5pM primers (sense+antisense)	2.0μl
Hot-start Taq polymerase	0.2μl
H2O	11.3μl
Total	20.0μl

反应条件:

step	temperature	time
Initial activation step	94°C	3 min
Denaturation	94°C	20 sec
Annealing	Varying temperature	20 sec
Extension	72°C	20 sec
cycles	38	
Final Extension	72°C	1min

### 5.2.5 多阶段标记物验证流程

多阶段标记物验证流程是一种生物标记物临床验证策略。目前的临床生物标记物开发在高密度芯片平台和高通量测序技术的推动下, 获得了前所未有的快速发展, 样本在 100-1000 左右的全基因组表达数据或全基因组甲基化芯片已经在

一些实验室出现,用于开发临床诊断标记物,临床预后标记物,药物靶标开发等。高通量及全基因组生物标记物数据有利于筛选最优的生物标记物组合以获得最佳的诊断或预测模型,大样本可以保证预测模型的稳健性和外推性。但是对于异质性较高的肿瘤诊断及预后预测等,1000的样本量可能仍然不能保证各种模型的稳健性。需要将全基因组水平获得的生物标记物在更大的样本中进行验证。在这个过程中,可以采取多阶段生物标记物验证流程,原理就是首先选择一批最有希望的生物标记物,在一批样本样本中进行测试,然后选择其中最优的一批标记物,在更大的样本量中进行测试,依次类推,随着标记物数量的减少,样本量的不断增加逐渐使得由标记物组合构成的诊断或预后模型逐渐稳定,并逐渐接近临床应用标准。

### 5.2.6 临床相关的统计方法

本研究中甲基化状态的评估设计两种数据类型: MSP 和 BSP。MSP 的结果为二分类定性结果,0 表非甲基化,1 表示甲基化。MSP 指示的甲基化与膀胱癌的相关性采用  $\chi^2$  或 Fisher 精确检验。膀胱癌预后复发状态与基因甲基化及临床变量之间的相关性由单因素和多因素 Logistic 回归分析方法进行评估。累积复发风险被定义为从膀胱癌首次诊断到肿瘤复发的时间。单变量和多变量 Cox 比例风险模型被用来评估基因甲基化及临床变量与疾病复发之间的关系。累计复发的风险曲线采用 Kaplan-Meier 法生成并由 Log-Rank 检验进行验证。所有统计分析均采用 R[35]或 SPSS 13.0 软件统计软件包(SPSS 公司,芝加哥,IL)进行。双侧检验及 P 值小于 0.05 被认为是具有显著性的统计差异。

#### 5.2.5.1 单因素和多因素 Logistic 回归分析

单因素 Logistic 回归分析用来逐个将每个甲基化标记物作为自变量,将肿瘤或对照,复发或非复发等作为变量,在年龄,性别,肿瘤分期等可能的混淆因素作为协变量的情况下,对单个甲基化标记物的 HR 进行分析。

多因素 Logistic 回归分析将所有甲基化标记物作为自变量,将肿瘤或对照,复发或非复发等作为因变量,在年龄,性别,肿瘤分期等可能的混淆因素作为协变量的情况下,同时对所有甲基化标记物的 HR 进行分析。

#### 5.2.5.2 Kaplan-Meier 估计和 Log-Rank 检验

在数据存在删失的情况下,Kaplan-Meier 是用于估计生存函数的一种非参数方法,也被称为乘积极限法(Product-Limit method)。在 Kaplan-Meier 估计模型中,生存函数被表示为一个阶梯函数,其公式如下:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leq t} \left[ 1 - \frac{d_i}{Y_i} \right], & \text{if } t \geq t_1. \end{cases}$$

$\hat{S}(t)$ 是生存函数在  $t$  时刻的估计值，在本例中代表累积复发风险值， $t_1$ 是实验的初始时刻。

Kaplan-Meier 估计的方差可以采用下面的公式进行估计：

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

Kaplan-Meier 估计使用真实数据构造生存曲线，将生存时间按照从小到大的顺序进行排列，然后在每个死亡节点上，计算初始人数、死亡人数、死亡概率、生存概率和生存率等参数。它既可以适用于小样本，又可以适用于大样本。两组样本之间的累计复发风险曲线的差异可以用 Log-Rank 检验进行统计推断。Log-Rank 检验可以用来比较两组或多组样本的生存时间，累计复发概率曲线的差异。其原假设是两组或多组样本来自同一分布。Log-Rank 比较每组中实际数量和理论数量的差异。以 2 组数据为例：

$$\chi^2_{LR} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

其中  $O_1, E_1, O_2, E_2$  分别表示第一组的复发数量及预期复发数量，第二组的复发数量及预期复发数量。因此通过自由度为  $N-1$  的  $\chi^2$  分布，就可以对两组数据的累计复发概率曲线的差异进行统计推断。其中  $N$  为数据的组数，在本例中  $N=2$ 。

### 5.2.5.3 Cox 比例风险模型

Cox 回归(cox 比例风险回归)用于评估风险函数(hazard function)与协从变量(covariates)之间的关系。Cox 模型是临床研究的经典模型，主要用于复杂疾病，慢性病的预后分析，也广泛应用于队列研究的病因分析。模型以瞬时死亡率为因变量，以多个影响因素为自变量，同时考虑到时间变量，通过回归分析拟合死亡力函数。拟合得到的模型需要进行拟合优度的检验，如果拟合力较强，则可以用来预测未来死亡率和构造生存曲线。这种方法是参数方法，因此得到参数模型可以做出比经验数据更多的分析。Cox 回归模型的数学表达如下：

$$H(t) = H_0(t) \times \exp(b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k)$$

$$\ln\left(\frac{H(t)}{H_0(t)}\right) = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

$\frac{H(t)}{H_0(t)}$ 为风险比,  $X_1 \dots X_k$ 为协变量,  $b_1 \dots b_k$ 为协变量需要估计的系数。 $\exp(b_i)$ 可以解释为瞬时相对风险。

## 5.3 结果

### 5.3.1 膀胱癌和正常组织的全基因组甲基化图谱

我们通过 MBD-methylCap 技术富集的 DNA 构建了 BCC 和 BM 全基因组 DNA 甲基化文库。对甲基化文库(MBD-富集组分)进行基于 Illumina 的 Genome Analyzer II 的高通量测序, 可以获得全基因组水平的甲基化图谱。原则上讲, 高甲基化的 DNA 片段在文库建立的过程中得到富集, 因此, 所获得的测序片段的定位与基因组中的甲基化区域对应, 测序丰度反映甲基化的程度。基于上述 BCC 和 BM 文库的深度测序产生了 6 百万左右的测序片段(reads), 每个测序片段的长度为 75 碱基(如表 5-8 所示)。平均每个库含有大约 470 百万的碱基数, 这个数量级的碱基数可使得基因组的 CGI 的覆盖深度达到 10X 左右。因此, 该数据集可以足够有效地描述全基因组甲基化图谱。

当这些测序片段 map 到基因组中时, 其分布呈现不均匀分布, 片段分布多的区域代表高甲基化区域。采用 MACS 对甲基化区域进行定位, 在 MACS 软件的分析中, 高甲基化区域被定义为一个个峰区域。本研究中我们发现在 BCC 数据中发现了 229,538 峰(甲基化区域的长度 659bp), BM 组中共发现了 210,051 峰(甲基化区域的长度 778 bp)。更加详细的信息如表 5-3 所示。

通过两组数据甲基化区域的比较我们可以获得 BCC 和 BM 的差异甲基化的区域。通过比较, 我们发现 2/3 以上的高甲基化区域是两个库所共有的。因此在后续的分析中, 这部分共享的甲基化区域将被排除在我们的研究之中。剩下的 1/3 的甲基化区域, 称之为差异甲基化区域(Differential Methylation Regions, DMR)。数据显示对于 BCC 和 BM, 分别具有 70,432 和 83,690 的 DMR, 详细信息见(表 5-8, 图 5-1 的 A 图)。

大量的 DMR 被分散在不同基因组元件中, 我们分析了 DMR 在不同基因组元件上的分布特征。分析显示 refGene 相关 DMR 在 BCC 和 BM 样本中各有 55,237 和 45,522 个, 表现出大致相近的 DMR 数量。然而, 对 CGI 的范围内的 DMR 进行统计时, 我们发现, BCC 样本中具有 21,179 个 DMR, 而在 BM 组中

只有 1,945 个 DMR，表现出了 10 倍的差异，说明 BCC 和 BM 在 CpGI 区域的 DMR 分布具有显著性的差异(表 5-8，图 5-1 的 B 图)。当我们进一步对 CpGI 和 Ref Gene 重叠元件进行分析时，我们发现 BCC 和 BM 中分别含有 4256 和 201 的 DMR。最后当限定在参考基因（ReferGene）启动子区的 CpGI 时，BCC 和 BM 组分别由 1,627 和 66 的 DMR (表 5-8，图 5-1 的 C 图)。因此，我们可以发现 BCC 和 BM 之间甲基化差异最显著的区域为参考基因启动子区的 CpGI 区域。

### 5.3.2 BSP 方法对膀胱癌及正常组织特异甲基化谱式的验证

为了验证 methyCap-Seq 所绘制的异常甲基化谱式，我们采用 BSP 技术对部分异常甲基化区域进行验证。24 个显著差异的甲基化位点被选择作为验证的对象。其中 22 个来自 BCC 特异的启动子区相关的 1627 个 DMR (17 个选自 top100 差异位点, 5 个来自 100-1627 差异位点)，另外 2 个来自 BM 特异的 66 个 DMR。经过验证发现 24 个候选的目标中有 23 个可以得到甲基化状态的吻合，符合率约 96%；有 1 个没有明显的差异甲基化特征，被判断为假阳性结果。这个结果说明本研究所建立的甲基化文库具有很高的质量（如图 5-5 及表 5-9 所示）。

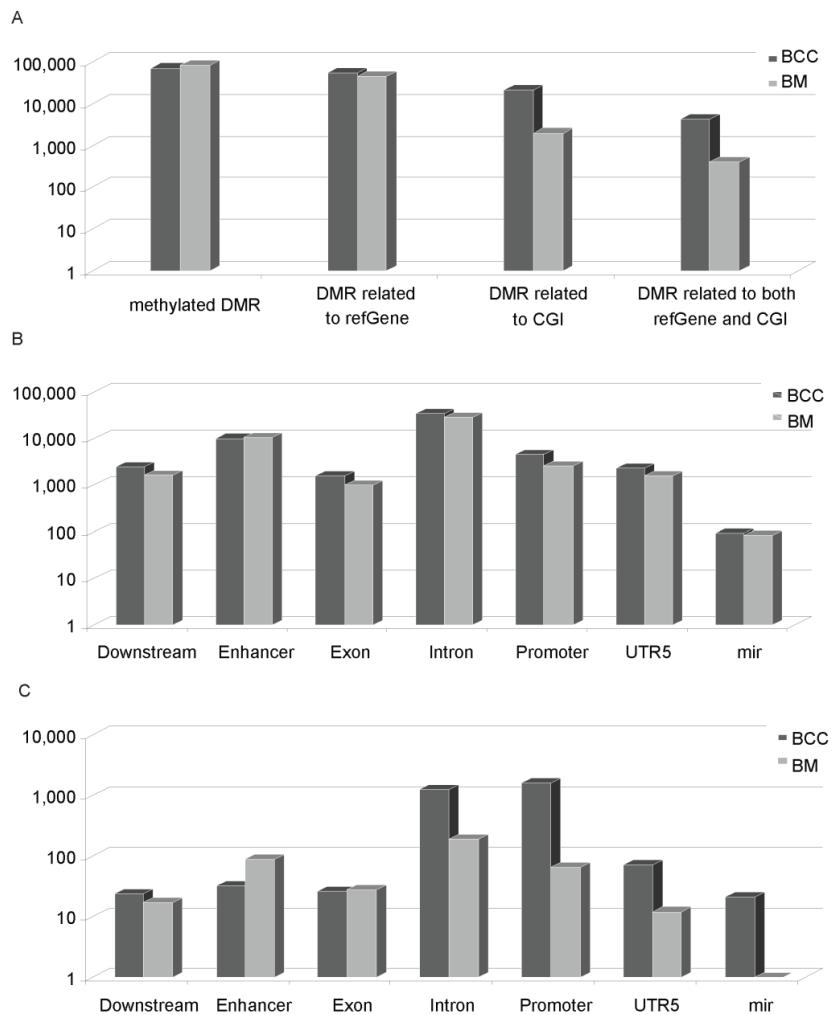


图 5-4 膀胱癌全基因组甲基化在不同基因组元件下的谱式特征

此外，我们又采取了另外两种方式对甲基化文库的质量进行评估。1)检查我们前期采用候选基因的方式发现的部分膀胱癌异常甲基化基因在本甲基化文库中的甲基化情况。通过筛选比较我们发现，前期我们发现的 21 个膀胱癌显著异常甲基化基因在本数据库中有 19 个基因得到验证，重现率为 90.5%[32]。

此外，我们还调查了另外两个研究报告的膀胱癌特异性标志物[18, 36]，结果显示 9 个文献报道的膀胱癌异常甲基化基因有 8 个位于我们的膀胱癌差异甲基化位点中。综合来看本研究所建立的 BCC 和 BM 甲基化文库可以提供较为可靠的膀胱癌异常甲基化位点。

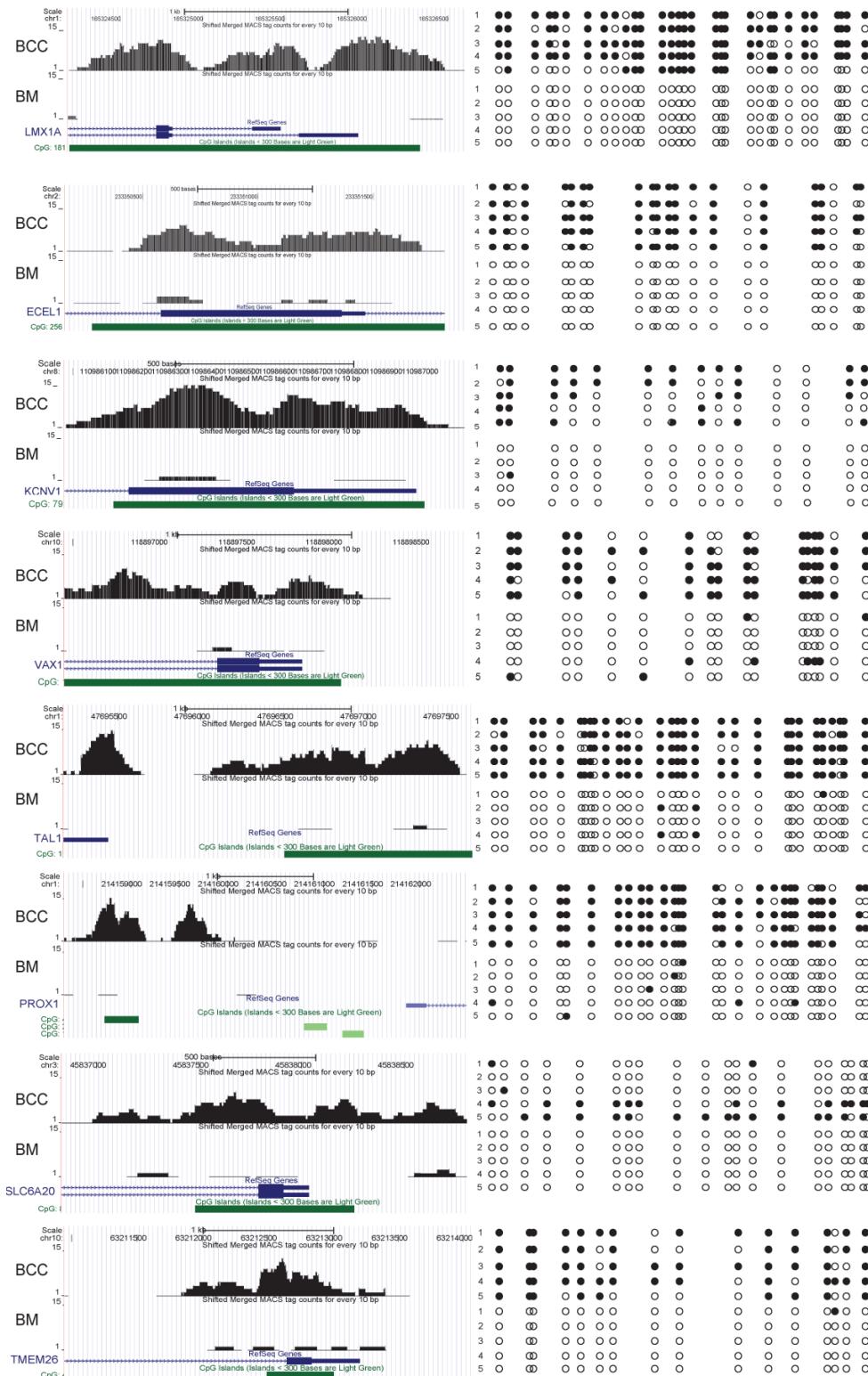


图 5-5 采用 BSP 技术对 MBD-methylCap 建立的甲基化谱式进行验证

如图 5-5 所示，左侧图展示了采用 wig 图谱表示的甲基化状态图谱，背景基因注释信息来自 UCSC 网站，将我们建立的甲基化谱图信息上传至 UCSC 数据

库后，可以实现甲基化信号与注释信息的同时展示，以便于临床科研工作的研究工作。有图是采用 BSP 技术对左边甲基化区域的验证。黑圈表示甲基化的 CpG 位点，白圈表示无甲基化的 CpG 位点。

### 5.3.3 基于尿液样本的潜在膀胱癌甲基化标志物筛选

为了筛选出可靠的具有一定临床意义的膀胱癌甲基化标记物，以形成一个有效的甲基化标记物组合而实现对膀胱癌的筛查或辅助诊断，我们采取如图 5-6 所展示的研究策略：多阶段生物标记物验证流程对甲基化文库中的膀胱癌异常甲基化靶点，采用 MSP 的方法在尿液样本中进行验证。靶点筛查的开始阶段，来自 1627 个启动子区相关的 BC 异常甲基化位点中的最显著差异的前 104 个异常甲基化位点，首先在少量的样本中，包括：2 个 BM，2 个 BC 和 8 个正常膀胱组织(BN) 中进行筛查。这里的 2 个 BM 和 2 个 BC 样本正是建立甲基化文库所使用的样本，因此这一步也是采用 MSP 技术对甲基化文库进行验证。在这一步筛查中，在 BC 中甲基化率 50%，在 BN 中甲基化率低于 25% 的差异甲基化位点可以进入下一轮筛查，不符合条件的退出筛查流程。经过这轮筛选，49 个位点最终通过。接下来，49 个甲基化位点在第二批样本中进行验证，样本包括 8 个 BN 和 18 个 BC 样本。在膀胱癌样本中甲基化率高于 17.0%，正常样本甲基化率低于 12.5% 的位点，可以进入第三轮验证。最终 8 个基因可以进入第三轮验证中，这八个基因包括：*VAX1*, *KCNV1*, *ECELI*, *TMEM26*, *TAL1*, *PROX1*, *SLC6A20* 和 *LMX*。在最后一轮筛查中 8 个异常甲基化位点在 471 样本：212 个 BC, 149 个 BN, 21 个术后样本, 48 个疑似转移性膀胱癌病人, 41 个非肿瘤性泌尿系统病变患者的尿液中进行了再验证。

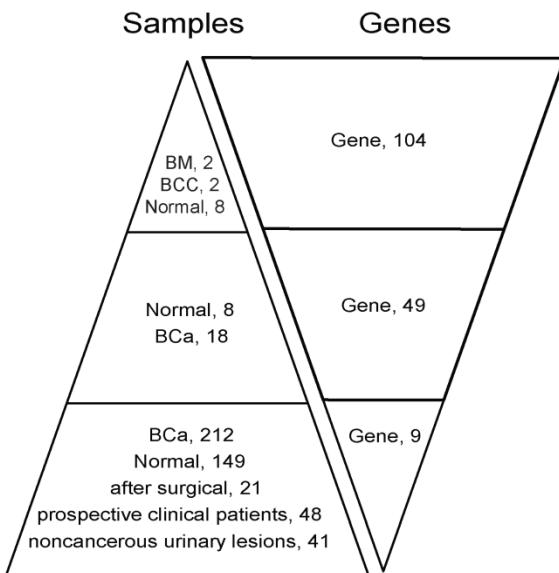


图 5-6 基于尿液样本筛选潜在的膀胱癌甲基化标志物的筛选策略

注：多阶段生物标记物验证流程：当存在大量需要验证的靶点时，首先在小样本中进行筛选，剔除不符合标准的位点，通过这种方式不减少候选基因的数量并增加筛查样本量，以提高结果的准确性。

### 5.3.4 临床样本中 8 位点甲基化标记物组合的诊断价值评估

8 个异常甲基化位点在上述最后一轮 471 样本中的验证结果在本部分进行详细介绍。8 个候选标记物位点在 212 个膀胱癌病人尿液中(BC)的甲基化率介于 9.43% 到 42.45% 之间，而在正常对照尿液中(BN)的甲基化率介于 1.34% 到 6.04%，所有的这 8 个位点在 BC 组和 BN 组中都呈现出显著性的甲基化差异( $P < 0.0001$ )，显示出这 8 个位点都是显著的膀胱癌生物标记物。

为了确保我们筛选的异常甲基化位点为肿瘤特异性异常甲基化位点，我们在样本采集时特意引入了 41 非肿瘤性泌尿病变病人的尿液样本（表 5-2 所示）。8 个基因在这批患者尿液中的甲基化频率介乎 0.00% 至 12.19%。统计显示这 8 个位点的甲基化在膀胱癌患者和非肿瘤性泌尿系统病变患者尿液中的 DNA 甲基化具有显著性差异（ $P$  均小于 0.04）。因此，这 8 个候选位点为潜在的膀胱癌诊断标记物，可以帮助对从正常对照组和良性病变组中对膀胱癌患者的识别。

考虑到肿瘤是一种多基因遗传疾病，具有复杂的亚型及不同肿瘤个体之间高度的异质性。单一的甲基化标记物很难提供足够的诊断的灵敏性和特异性[37]。因此，多个标记物组合是肿瘤筛查或诊断的必要选择[38]。我们发现采用 *VAX1*, *KCNV1*, *TAL1*, *PROX1* 四个基因组成的预测模型的灵敏性和特异性可以分别达到 88.59% 和 76.89%，详细信息如表 5-5 所示。

考虑到四个基因组合的预测标记物距临床应用仍有距离。为了进一步提高标记物组合的诊断潜力，我们另外增加了两个标记物：*CFTR* 和 *SALL3*。这两个标记物是我们前期的工作发现的膀胱癌差异甲基化位点[32]，并且也出现在本次研究出现的 1,627 个 BC 特异的 DMR 列表中。为此，我们将这两个位点的甲基化状态，在本研究的样本中进行了检测。结果发现 *CFTR* 的表现出和前期工作相似的膀胱癌诊断潜力，灵敏性和特异性分别为 52.36% 和 96.64%（如表 5-5 所示）。然而，*SALL3* 并没有重复出前期的诊断潜力，这可能是由于样本的异质性造成，因此排除在研究之外。最后，由 5 个基因（*VAX1*, *KCNV1*, *TAL1*, *PROX1* 和 *CFTR*）组成的标记物组合被用来对 BC 的诊断。5 基因诊断模型对 BC 的诊断模型的灵敏度，特异性，阳性预测值和阴性预测值分别为：88.68%，87.25%，90.82% 和 84.42%（如表 5-5 所示）。

### 5.3.5 甲基化诊断模型具有与膀胱镜诊断相近的诊断效果

对临幊上疑似病人的膀胱癌患者的准确判定是甲基化诊断模型很重要的评估指标。因此，我们采用本课题建立的 5 基因诊断模型对疑似尿路上皮恶性肿瘤病人的尿液进行检测，并对病人的疾病进行判定。在 48 名患者中，32 名 BC 疑似患者经膀胱镜检查最终证实为膀胱癌患者，这 32 个患者中有 25 个患者呈现 5 和标记物中的至少 1 个为甲基化阳性。16 例疑似患者通过膀胱镜检查未能确诊为膀胱癌，其中的 14 个甲基化标记物均为阴性。因此，5 个基因标记物组合的表现出与膀胱镜检查良好的一致性(81.25%)。

### 5.3.6 甲基化模型可用于手术切除效果有效性的评估

尿液甲基化标记物组合，理论上可以用于手术切除效果的评估。采用 PCR 技术可以对微量的甲基化信息进行监控，从而对手术效果进行评估，使得医生更加有效地对病人制定后续的治疗方案。21 个术前病人接受了 5 个标记物的检测。结果发现所有病人在 5 个基因中至少有 1 个基因表现出甲基化阳性。而术后只有 2 个(9.5%)患者还保留原有的 MSP 阳性，显示出绝大多数手术的有效性切除( $P < 0.0001$ )，也说明了为什么手术切除肿瘤可以有效地降低肿瘤的死亡率，同时也证实在尿沉淀中观察到的甲基化和相应的膀胱肿瘤负荷之间存在着密切的关系。

### 5.3.7 *VAX1* 和 *LMX1A* 基因高甲基化可用于癌症复发预测

除了研究基因的甲基化状态与肿瘤发生之间的关系。我们还研究了甲基化状态与和不同其它临床表型之间的关联性。单因素 logistic 回归分析，在校正年龄，性别，肌肉侵犯，治疗方式，肿瘤分期等因素后，9 个靶基因 (*VAX1*, *KCNV1*,

*ECE11, TMEM26, TAL1, PROX1, SLC6A20, LMX1A* 和 *CFTR*) 中 *VAX1* 和 *LMX1A* 甲基化频率在复发病例中(N=55)和原发病例(N=157)之间存在显著差异。*VAX1* 和 *LMX1A* 在复发病例中显著高于原发病例, 其 HR 分别为 2.37(95% CI, 1.27-4.44, P<0.05) 和 2.59 (95% CI, 1.01-6.65, P<0.05), 如图 5-7 所示。

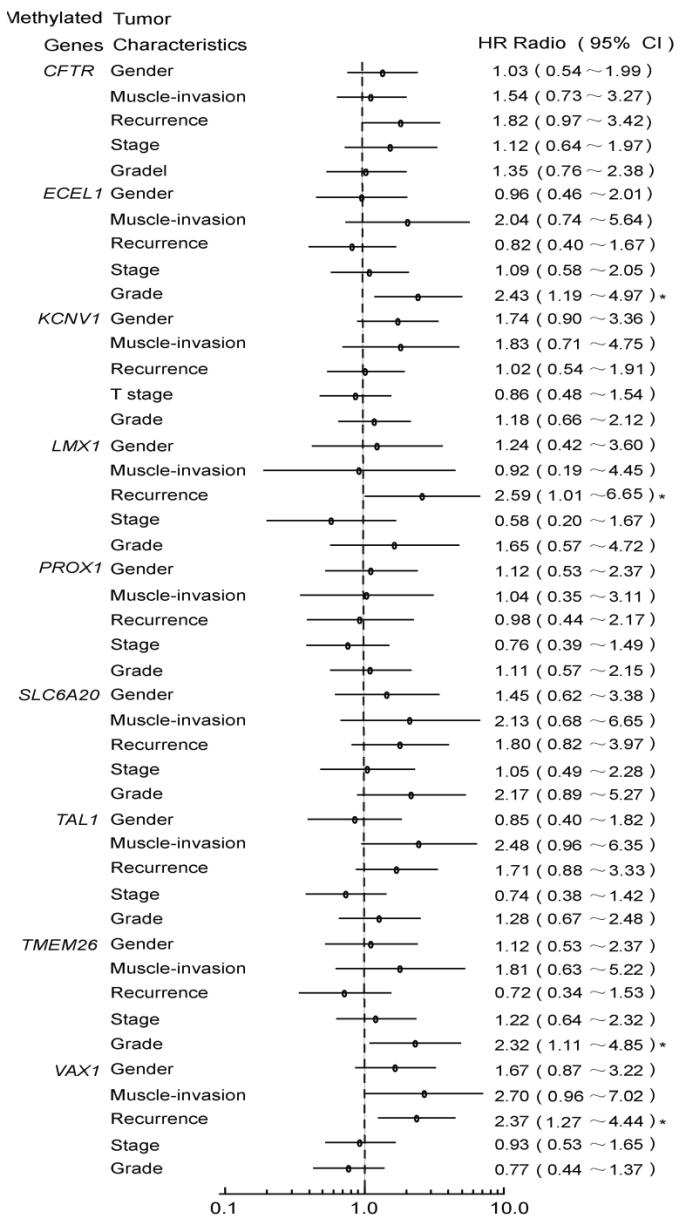


图 5-7 基因与膀胱癌临床指标之间的关系

多变量 Logistic 回归模型显示, *VAX1* 和 *LMX1A* 的 HR 分别为 2.27(95% CI 为 1.20-4.32, P = 0.047)和 2.63(95% CI, 1.01-6.85, P = 0.012)(如表 5-3 所示)。两个基因合并 HR 更是远远高于单个基因的 HR, 其 HR=4.73 (95% CI, 1.39-16.08, P = 0.013)。

表 5-3 多因素 logistic 回归估计 *LMX1A* 和 *VAX1* 的肿瘤复发危险比

Clinical variables		Number of case	Number of recurrence	(%)	HR95% CI)	P
Overall		212	55	25.94%		
<i>LMX1A</i>	Negative	192	46	23.96%		
	Positive	20	9	45.00%	2.63(1.01-6.85)	0.012
Gender	Female	46	15	32.61%		
	Male	166	40	24.10%	0.69(0.33-1.42)	0.555
Muscle invasion	Absent	178	44	32.35%		
	Present	34	11	24.72%	1.76(0.71-4.37)	0.217
Stage	0a+I	137	36	26.28%		
	II+III	75	19	25.33%	1.00(0.46-2.16)	0.920
Grade	I	73	23	31.51%		
	II+III	139	32	23.02%	0.59(0.29-1.18)	0.171
<i>VAX1</i>	Negative	122	23	18.85%		
	Positive	90	32	35.56%	2.27(1.20-4.32)	0.047
Gender	Female	46	15	32.61%		
	Male	166	40	24.10%	0.80(0.38-1.68)	0.313
Muscle invasion	Present	178	44	24.72%		
	Absent	34	11	32.35%	1.79(0.71-4.49)	0.226
Stage	0a+I	137	36	26.28%		
	II+III	75	19	25.33%	0.96(0.45-2.08)	0.998
Grade	I	73	23	31.51%		
	II+III	139	32	23.02%	0.61(0.30-1.24)	0.134
<i>LMX1A+VAX1</i>	Negative	200	48	24.00%		
	Positive	12	7	58.33%	4.73(1.39-16.08)	0.013

Hazard Ratio are reported on the basis of the multivariate logistic-regression model.

此外我们在另外一套有 145 个膀胱癌病人组成的随访数据中(108 无复发, 37 复发)也对上述结论也进行了验证, 多因素 Cox 比例风险模型显示, *VAX1* 和 *LMX1A* 的 HR 分别为 2.11(95% CI, 1.08-4.11, P = 0.029)和 3.31(95% CI, 1.27-8.59, P = 0.014), 详细信息如表 5-4 所示。两个基因合并 HR 更是远远高于单个基因的 HR, 其 HR=7.25(95% CI, 2.41-21.79, P = 0.014)。

表 5-4 基于 cox 回归的 *VAX1* 和 *LMX1a* 甲基化与预后复发的关联

Clinical variables		Number of case	Number of recurrence	(%)	Cox-ranked univariate	Cox-ranked multivariate		
					HR(95% CI)	P	HR(95% CI)	P
Overall		145	37	25.50%				
<i>VAX1</i>	Negative	93	18	19.40%				
	Positive	52	19	36.50%	1.92(1.00-3.65)	0.048	2.11(1.08-4.11)	0.029

Gender	Female	29	7	24.10%				
	Male	116	30	25.90%	0.86(0.38-1.95)	0.711	0.72(0.31-1.67)	0.44
Muscle invasion	Absent	128	33	25.80%				
	Present	17	4	23.50%	1.05(0.37-2.96)	0.928	0.93(0.30-2.85)	0.893
Treatment	TURBT+IC	78	16	20.50%				
	PC	67	21	31.30%	0.93(0.48-1.79)	0.822	0.84(0.42-1.67)	0.614
Stage	0a+I	97	26	26.80%				
	II+III	48	11	22.90%	1.14(0.56-2.30)	0.721	1.3(0.61-2.77)	0.5
Grade	I	46	10	21.70%				
	II+III	99	27	27.30%	0.77(0.37-1.60)	0.488	0.66(0.30-1.43)	0.29
<b>LMX1A</b>	Negative	136	32	23.50%				
	Positive	9	5	55.60%	3.08(1.20-7.95)	0.019	3.31(1.27-8.59)	0.014
Gender	Female	29	7	24.10%				
	Male	116	30	25.90%	0.86(0.38-1.95)	0.711	0.83(0.36-1.89)	0.649
Muscle invasion	Absent	128	33	25.80%				
	Present	17	4	23.50%	1.05(0.37-2.96)	0.928	1.09(0.35-3.36)	0.881
Treatment	TURBT+IC	78	16	20.50%				
	PC	67	21	31.30%	0.93(0.48-1.79)	0.822	0.88(0.45-1.75)	0.724
Stage	0a+I	97	26	26.80%				
	II+III	48	11	22.90%	1.14( <b>0.56</b> -2.30)	0.721	1.28(0.59-2.75)	0.529
Grade	I	46	10	21.70%				
	II+III	99	27	27.30%	0.77(0.37-1.60)	0.488	0.67(0.31-1.46)	0.312
<b>LMX1A+VAX1</b>	Negative	140	33	23.60%				
	Positive	5	4	80.00%	6.40(2.24-18.29)	0.001	7.25(2.41-21.79)	0.014

Kaplan-Meier 生存曲线显示 *VAX1* 和 *LMX1A* 基因 DNA 甲基化与膀胱癌复发的累积风险具有显著关联性( $P=0.034$  和  $P=0.013$ )，这两个基因的高甲基化状态指示不良预后。组合的 *VAX1* 和 *LMX1A* 甲基化状态和膀胱癌复发的累积风险的相关性更加显著( $P < 0.0001$ )，如图 5-8 所示。

除了 *LMX1/VAX1*，我们也评估了其他 7 个基因中任意两个基因对组合。分析发现有一些基因的甲基化组合在上述的 179 个样本中表现出与膀胱癌复发显著地相关性，但这种关联未能随访的 145 个样本的数据中得到重复，因此这些基因的异常甲基化可能是膀胱癌复发产生的结果而不是诱因。

此外我们发现，*ECEL1* 和 *TMEM26* 基因的甲基化状态与肿瘤分化程度显著相关，HR 分别为 2.43 (95% CI, 1.19-4.97,  $P=0.01$ ) 和 2.32 (95% CI, 1.11-4.85,  $P=0.03$ )。提示 *ECEL1* 和 *TMEM26* 很可能参与了膀胱癌的疾病进展过程。

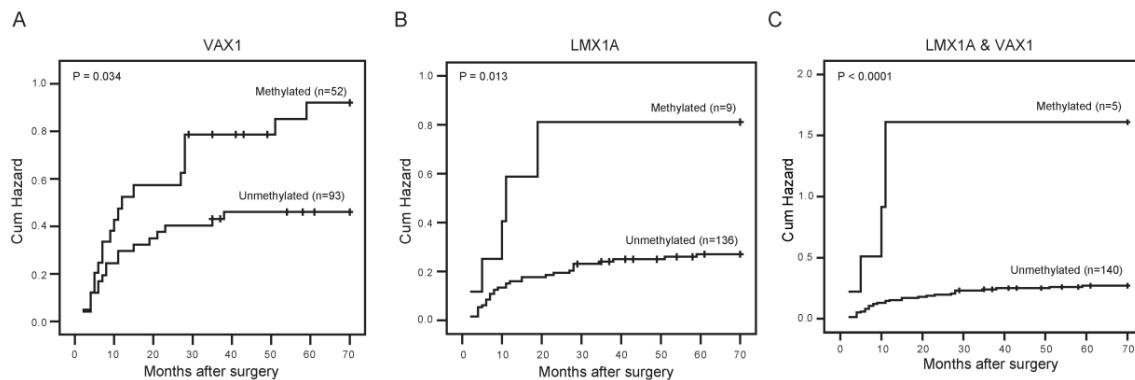


图 5-8 Kaplan-Meier 生存曲线显示甲基化与膀胱癌复发的相关性

## 5.4 结论

MBD methylCap-Seq 技术是一种稳定可靠的全基因组甲基化分析技术。通过多阶段生物标记物验证流程，我们从 104 个候选基因中，经过 509 个独立样本的验证，最终筛选出了 9 个膀胱癌相关 DNA 甲基化标记物。本研究发现，由 5 个基因(*VAX1*, *KCNV1*, *TAL1*, *PROX1* 和 *CFTR*)组成的甲基化标记物组合可以用来对 BC 进行判别。5 个基因诊断模型对 BC 的诊断的灵敏度，特异性，阳性预测值和阴性预测值分别为：88.68%，87.25%，90.82% 和 84.42%。临床双盲实验显示基于上述 5 甲基化标记物的诊断模型和膀胱镜检查具有极高的吻合度(81.25%)。两个独立的临床数据(回顾性数据和临床随访数据)都显示 *VAX1* 和 *LMX1A* 两个基因的异常甲基化与膀胱癌复发高度相关。此外，研究还发现 *ECEL1* 和 *TMEM26* 两个基因的甲基化标记物与肿瘤的分化显著相关，提示这两个基因的甲基化可能参与与肿瘤分化进展。本研究通过全基因甲基化谱式分析，小样本筛选，大样本验证，双重独立数据验证，临床双盲模拟等实验设计和分析对膀胱癌潜在的甲基化标记物进行了筛查，最终成功建立一系列具有潜在应用价值的膀胱癌甲基化标记物。

## 5.5 讨论

在本研究中，我们通过首先通过 MBD methylCap-Seq 技术建立了 BC 和 BN 全基因组水平的甲基化谱式。然后通过两组甲基化谱式的比较，分析获得了膀胱癌异常的 DNA 甲基化位点。通过大量临床样本的筛查获得了一系列针对特定临床特征(诊断，复发，肿瘤进展)的甲基化生物标记物。

根据目前的肿瘤研究,细胞系共享了大部分肿瘤实体组织的遗传学和表观遗传学谱式,包括DNA甲基化谱式。肿瘤细胞系被广泛用于肿瘤的分子生物系,细胞生物系,药物基因组学等研究模型。因此,我们考虑首先建立基于细胞系的膀胱癌全基因组甲基化图谱,从而建立膀胱癌特异性甲基化的信息。

许多基因已被报道在膀胱癌中具有异常甲基化现象。近年来,随着新的异常甲基化检测方法的出现,如甲基化微阵列芯片,已经确定了大量的肿瘤甲基化标记物,并且这些标记物具有很高的灵敏性和特异性[39-41]。这为在全基因组范围内筛选疾病相关的异常甲基化标记物提供了一个高效可靠的靶点候选库。但是基于芯片的方法具有一些非常严重的缺点,1)芯片的生物代表性严重依赖于芯片探针的选取;2)一般芯片的密度只能保证每个基因仅含有为数不多的几个位点进行代表。在本研究中,我们采用了MBD methylCap-Seq技术建立全基因组甲基化谱式,相比于芯片技术有如下几个优点:1)开放的平台,可以发现新的疾病相关甲基化位点,2)具有更高的通量,3)筛选的单位与染色体区域而非孤立的几个CpG位点,具有更高的可信度。甲基化芯片一般只重点对基因及启动子区的甲基化状态进行研究,本研究对增强子,5'UTR,外显子,内含子以及miRNA等区域的甲基化状态都进行了分析。并且我们建立的甲基化谱式经过了MSP和BSP的验证,具有很高的可重复性。

对于膀胱癌诊断,我们急需一种高灵敏度和特异性的无创诊断方法。目前采用尿沉淀DNA甲基化对膀胱癌进行辅助诊断已经得到了临床科研工作者的认可[17, 32, 36, 39, 42-45]。在这些研究中所使用的技术包括常规MSP,定量PCR,巢式MSP和MethyLight。诊断模型的灵敏性介于77%-94%,特异性介于67%-100%(如表5-11所示)。模型中生物标记物的数量介于3-11之间。我们的研究发现了一组由5个甲基化标记物组成的膀胱癌诊断模型,可以实现88.68%的灵敏性和87.25%的特异性。均衡的灵敏性和特异性是我们开发的甲基化诊断模型的优点,诊断模型的训练来自于中国人群的DNA甲基化训练集,因此目前适合在中国人群中发挥具有中等水平的辅助诊断能力。因为膀胱癌特异性甲基化标记物的诊断效能可能在不同种族人群之间的有所不同[32],所以本研究获得的诊断标记物组合对更广泛人群的试用程度需要进一步的论证。

肿瘤诊断模型的建立对样本量的要求比较高,因为肿瘤的异质性较高,发病原因复杂,亚型较多,为了使预测模型的外展性得以保证,基于DNA甲基化的肿瘤预测模型必须具备较大的样本量。此外为了使得甲基化预测模型具备一定的无创早期筛查的潜力,应该保证预测模型中正常对照的样本量,以提高特异性的

准确度。在许多早期的研究中，正常对照的样本量一般较小，从 6-20 不等[32, 39, 42-45]，这将影响到生物标志物的可靠性，及诊断模型的稳定性。最近，Reiner 等人采用了 59 例的正常对照[18]，Chung 采用了 110 例正常作为对照[36]（如表 5-11 所示）。在我们的研究中，我们采用了 212 例膀胱癌病人，149 例正常对照，根据文献检索，应属目前最大的病例对照研究。此外，我们选择了 41 例非肿瘤性泌尿系统病变患者的尿液 DNA 作为对照，以保证课题获得的甲基化标记物为肿瘤特异性标记物。膀胱癌切除手术前后的尿液沉淀 DNA 的甲基化状态，提示尿沉淀 DNA 中的甲基化谱式基本可以确定来自于膀胱癌实体组织。一个小规模的临床双盲实验显示 DNA 甲基化诊断模型与膀胱镜检查的一致性高达 81.25%。证明了该甲基化诊断模型是一个很有潜力的无创膀胱癌辅助诊断方法。

癌症复发在临幊上是降低膀胱癌病人生存率的重要因素之一。膀胱癌复发的检测主要依赖于侵入性膀胱镜检查，临幊上迫切需要找到一种无创的膀胱癌复发监控方法，从而实现对膀胱癌手术进行评估。尿沉淀 DNA 甲基化检测为膀胱癌的实时监控提供了重要的保证。本研究表明，*VAX1* 和 *LMX1A* 甲基化是膀胱癌的疾病进展及术后复发的重要指示物。具有重要的临幊应用价值。

目前，DNA 甲基化可以作为肿瘤标记物的重要生物学理论依据，包括启动子区 DNA 甲基化可以调控基因的表达。因而异常的 DNA 甲基化可能会导致抑癌基因表达沉默，或癌基因的过度激活。本研究所确认的 8 个膀胱癌异常甲基化基因均为膀胱癌中首次报道的甲基化异常标记物。但其中有 4 个基因的甲基化异常有文献报道与其他肿瘤有关。*PROX1* 和 *LMX1* 的高度甲基化及表达沉默也发生在乳腺癌及胃癌中[46, 47]。*TAL1* 的基因表达与 T 细胞急性淋巴细胞白血病相关[48]，*SLC6A20* 的异常高甲基化发生在恶性间皮瘤中[49]。此外其它 4 个基因，*VAX1*、*KCNV1*、*ECEL1* 和 *TMEM26* 是首次报道与膀胱癌相关。*VAX1* 编码一种同源结构域转录因子，其在眼睛和视交叉神经的发育中发挥重要作用[50]。*KCNV1* 基因编码一种 K<sup>+</sup>离子通道蛋白[51]。*ECEL1* 编码一种 M13 家族内肽酶[52]，而 *TMEM26* 编码进化上十分保守的跨膜蛋白，可能是参与肢体发育的分子回路中的重要一员[53]。这些靶点基因的甲基化参与膀胱癌的具体机制，还需要继续深入的肿瘤分子生物学或细胞生物学功能研究。



表 5-5 9 个候选标记物在肿瘤-正常和肿瘤-良性病变诊断能力

	Bladder cancer			Bladder cancer vs Normal control (n=149)				Bladder cancer vs Nontumor urinary lesions (n=41)			
	Sensitivity(%) (pos./total)	Specificity (%) (neg./total)	AUC(95% CI)	PPV (%)	NPV (%)	P	Specificity (%) (neg./total)	AUC(95% CI)	PPV (%)	NPV (%)	P
VAX1	42.45(90/212)	95.31(142/149)	73.3(68.0-78.6)	92.78	53.79	< 0.0001	87.81(36/41)	59.5(52.5-66.5)	94.74	21.43	0.0002
KCNV1	36.92(84/212)	93.96(140/149)	71.3(65.7-76.8)	90.32	52.24	< 0.0001	95.12(39/41)	60.5(53.6-67.5)	97.67	23.35	< 0.0001
ECELI	26.89(57/212)	97.31(145/149)	70.8(64.7-76.8)	96.55	48.33	< 0.0001	97.56(40/41)	59.3(51.7-67.0)	98.28	20.51	0.0002
TMEM26	26.42 (56/212)	96.64(144/149)	69.4(62.9-75.8)	91.80	48.00	< 0.0001	97.56(40/41)	60.3(52.5-68.0)	98.25	20.41	< 0.0001
PROX1	24.53(52/212)	98.66(147/149)	71.1(65.0-77.3)	96.30	47.88	< 0.0001	100.0(41/41)	59.1(61.2-66.9)	100.0	20.40	< 0.0001
TAL1	24.83(52/212)	98.66(147/149)	72.5(66.8-78.3)	93.44	48.51	< 0.0001	100.0(41/41)	60.4(52.9-68.0)	100.0	20.40	< 0.0001
SLC6A20	15.57(33/212)	97.89(146/149)	69.7(62.3-77.0)	91.67	44.92	< 0.0001	100.0(41/41)	59.3(50.2-68.4)	100.0	18.64	0.0039
LMX1	9.43(20/212)	98.66(147/149)	67.1(57.5-76.8)	90.91	43.36	< 0.0001	100.0(41/41)	58.8(47.5-70.1)	100.0	17.60	0.0404
<b>Panel 1</b>	76.89(163/212)	88.59(132/149)	82.8(78.5-87.1)	90.56	72.93	< 0.0001	85.36(35/41)	84.3(79.1-89.5)	96.45	41.67	< 0.0001
CFTR	52.35(111/212)	96.64(144/149)	77.4(72.6-82.1)	95.69	58.78	< 0.0001	97.56(40/41)	63.8(57.1-7.02)	99.11	28.37	< 0.0001
<b>Panel 2</b>	88.68(188/212)	87.25(130/149)	89.9(86.7-93.2)	90.82	84.42	< 0.0001	85.36(35/41)	90.0(85.9-94.2)	96.91	59.32	< 0.0001

Panel 1 包括： VAX1,KCNV1,TAL1,PPOX1

Panel 2 包括： CFTR,VAX1,KCNV1,TAL1,PPOX1

## 5.6 附录信息

**附表 1. BSP 所涉及的引物对序列**

**表 5-6 BSP 所涉及的引物对序列**

Gene	GenBank.	Primer sequences		Amplicon location relative to TSS	Size (bp)
		Sense 5'-3'	Antisense 5'-3'		
AVPRIA	NT_029419	GTTAGAGGTTATTATGGGTTGGA	AAAAAAACCAACTCACCCCTACTAAA	1759~2039	280
BEND4	NT_006238	GGATTTAGAGTGGGTGTTTTTA	TTCCCTCCCACCCCTTAATTATA	1826~2088	262
BMP7	NT_011362	TGGTTATTTGTAGGTTGGTT	CTCTAACACCCCTATTCTACTAC	577~903	326
CDO1	NT_034772	GGAGAGTTATTAAGAAAGGTGG	CTTTCTTTCCCTCTTACTCA	-332~-31	301
CNTFR	NT_008413	TTTGGGGATATTATAGGTAGATT	CCCAAAAATAACTCTTTCTC	535~805	270
DAZ3	NT_011903	TAAAGGTTAAGGTGTAAGGAAAATTAAG	ACCTACCTAAACCCAAACACTAAC	-421~-130	291
ECEL1	NT_005403	GGTTAAAGAATTAAATTGTT	TCCAATAAAATAATTATATATCC	450~611	161
HS3ST3A1	NT_010718	GAGGGTAGTAAATTGTTGGT	CCTTAACTATCACTCAAACAACC	-309~-24	285
IRSI	NT_005403	TTTGTGGTGTGGGATTATAG	CCCAACCAAATAAAACTAACCC	931~1111	180
KCNVI	NT_008046	GGATTTTTGTAGGAAGGAGA	TACCCAACCCCTCTAAAACCTTAC	589~754	165
LMX1A	NT_004487	TTAGTGATTGGAGTAGAGAGAAGTTG	CTCTACCTCCCTCCTAACCTAAA	-421~-130	284
MAFA	NT_008046	AGGGGTTTTTTAAGGTTTTTT	AACCAAATCTAACTCTTCCAAAC	1332~1604	272
NKX6_1	NT_016354	GGGTTAGGATGGATTGAG	AACTCTACTCAACCCCAAC	526~781	255
NOS1	NT_009775	GGATAGTTGGTAGTGTAAAGATT	ACTCCAACTTACAAAACCCCTAAC	865~1120	255
PGR	NT_033899	GAATTTTTTGGGATTAGGG	TTAAAAAAAAATAACTCTAAATCC	1402~1561	159
PROX1	NT_167186	AGAGAGGTTAGGAGTTGGT	AAAAACTCTCCCACCCCCAA	-3162~-2921	241
SIM2	NT_011512	TTTAAAGGAGTTTGGAAAAG	TAACCTAAAAATCAACAAAAACC	-2024~-1774	250
SLC6A20	NT_022517	AGTTGGTTAATAAGGGTGAGG	ACCCCAACCCAACTCTATAC	21~271	250
TALI	NT_032977	GGAAAATTTTAGGAGGTGATT	TCCTCCAAACTAAAAACTAAACAC	-1379~-1154	225
TBX20	NT_007819	GGTTAGGGAAAAGGGTGTATT	CCCTACTTTCTTTAAAACCTTAAAC	-913~-666	247
TLX1	NT_030059	ATTAGGGTGGTGTATGGGATT	CCTACCAAAACCATTTCAC	324~565	241
TMEM26	NT_030059	TTTAGGAAGAGTAAGAGGTTGAGTAG	AAAAAAACCACAAACAACAAAT	243~515	272
TUBB2B	NT_007592	ATTGAGGTAGGAATTGGTTTT	CTCAACAACTACAAACTAAACTCC	-1529~-1229	300
VAX1	NT_030059	ATTTTGTTTGAAGATGTAGAGTGG	ACTACAAAAACCCCATCAAAA	-2161~-1937	224

**附表 2. MSP 所涉及的引物对序列**

表 5-7 MSP 所涉及的引物对序列

Gene Name	Primer sequences		Amplicon location relative to TSS	Size (bp)
	Sense 5'-3'	Antisense 5'-3'		
<i>ACTA1</i>	GTTTTAGGTAACGCGGGTC	AAACAAATACGATTCCCGAA	-355~-190	165
<i>ADRA1A</i>	TTTTATTATTGGAATTGTTAGC	AAACTTCGAAAATTTAACCGTAA	-586~-418	168
<i>ALDHIA2</i>	GAAAGGTAGGAAAGTTAGGTTGC	GACTAAAAAAATCGAAAAATCGAA	-632~-428	204
<i>ARPC1B</i>	CGTTAGGTTGGAGTGAAATGGC	ACCGAAACGAAAAATCACGA	-1052~-840	213
<i>BARHL1</i>	GTTTTTCGGGAGAATTTCG	AAATCGAACGAATCAAAACGTC	-1990~-1789	202
<i>BDNF</i>	AAATGTGAGTTAATAGTTCGGTGC	CCCGCAAAATAAAAAACGA	-565~-407	158
<i>BEND4</i>	GGTGTCCCCACGAGGCCGT	CGTAAATTTCGAACGCGAA	1776~2075	196
<i>BHLHE23</i>	TGTTTTCGATTTTCGGTC	GCTCTAACCCCTACGTATCG	-286~-141	145
<i>BMP7</i>	AGGAGAAGTTGGTCGTCG	CCTAAACAAACGAAATACACTCG	624~842	218
<i>C10orf14</i>	AACGTGATTTAGTTGAGGGTTC	GACTTATTAAAACCTCTCCCCG	-3359~-3179	180
<i>C1QL2</i>	TTAATATTGGTTTACGTTTCGG	CAATAACCCCTACACTCACCAGAC	-1504~-1313	192
<i>C2CD4B</i>	TGTTAGTGGTTTTAGTCGTTGTC	ATACATACAAACCTCCGACGT	-136~44	180
<i>C7orf52</i>	GTTTATCGGTTAGGGGTTCG	AAACACAAAACGAAATCTCGC	-198~12	210
<i>C8orf84</i>	CGATAGGTTGGTCGTAGAAATAC	CCATAAAAACCTATAAATAACGCT	203~375	172
<i>CALCA</i>	ATTATTTTCGGTGGAACCGC	AATACGAACACGACCCCG	-1544~-1757	213
<i>CBX8</i>	TCGTTAGAGAAAGTTATTTTCGT	CAATAACCCCTAAAACCAACTCGA	-2221~-2071	151
<i>CCND1</i>	AATCGTTTTAATTTAACGAGTTGC	CGTCTATCCGACGAACCTACG	-3423~3449	176
<i>CDH8</i>	GTTTATAATGGTGGAGGGTTTC	CACCAACTACACTTAAATTACGAA	39~223	184
<i>CDO1</i>	GAGAGTTATTAAGAAAGGTGGC	AAAAAAACGTTAACCGAACGAA	-230~-32	198
<i>CHRDL2</i>	CGGGGTTAATTTGATTTTC	CAATATTCTAATTCCATCCATCTACG	-1047~-861	186
<i>CNTFR</i>	AAATTATGGGACGTCGATC	ACAAAAATCTAAACTTCAAACGC	558~699	141
<i>COBL</i>	GATGGGTCGTTAGATTTCG	AAAAAAACTCCCCGACTTCG	720~919	198
<i>COL25A1</i>	CGGTTGTCGAAGGATTTTC	AAACGAACTCCTACTTTACCTTCG	-204~3	207
<i>CTSA</i>	ATCGAAATTGTTAATGCGT	ACCTAAACCACACCTAAATCTTCG	-422~278	145
<i>CYB5R2</i>	TATTTATAAATATTCGTCGCGG	GTATCCGAATCTGACCGAT	-262~-96	166
<i>CYP24A1</i>	TATTTAGTTGGGGTTGTTTC	AAATCCTTCTACTACCTATCCG	232~482	251
<i>CYP26B1</i>	GCGTTAGTGGTTCGGAATC	ATAAAACTCCCGCTCATCGA	159~397	238
<i>DACHI</i>	ATTTAGGTTGGGAGCGT	AATATAAAAAACCGAACCTCGA	261~414	154
<i>DBC1</i>	TGAACGTCGGGTTGTTTC	GATCCCTTAAATACTCGTACGC	-305~-110	195
<i>DGKK</i>	CGGGTTTATAGTTATTCGAGC	TAAAACGACTCCAAAATTCCGT	560~770	210
<i>DLX4</i>	TTTATTAAGTGTAGGGGTGTTCGT	TACTTACTCTTCTAAATCGCCGAA	-3864~-3664	201
<i>DPY19L2P2</i>	GGGAGTGTAGTTTATTCGAGC	ACGAAAAATTCTTCAAACCTAAATCG	-150~-321	172
<i>E2F8</i>	GGGCGTAGCGTAGTAGTTGTC	CTAAAAAACCTACTTCAAATCGAA	-1949~-1745	205
<i>ECELI</i>	ATTTATCGTAGTCGCTGTTAGC	ATAACCCCTAAACCCCATCCG	-60~141	201
<i>ESX1</i>	TTTTATGTTAAGCGTTGTCGA	CTACGAAAAACACCGTACCG	-127~64	191
<i>EVX1</i>	TTTTTATACGCGTCGTTTTC	AAAACCTCCAACCTAAACTCCG	-3055~-2844	212
<i>FAM84B</i>	TAGTCGCGTTTTGGTATTTAC	GAACGATCTAATAAAATTCTGTATCG	449~684	235
<i>FEZF2</i>	TTTGCATATTGAGGAGGTTAGAC	CTACAAAACCGAACTAACCTAACG	398~595	197

<i>FGF3</i>	TTTTTGCATATTAAATTGGGTATTCT	TCTCTCTAAAACCTCTATCGACGC	-537~-374	163
<i>FOXD3</i>	TTTCGATTATCGGGGTTTC	CTCGACCTCTAACCTAACGTA	-480~-330	151
<i>GF11</i>	TATAGGTGCAAATTGAGCGTAGC	TAAACCAACCGAAACCACGA	1206~1404	198
<i>GJD3</i>	GTGTTGTATACGAATTTCGGTTC	GTTCCCTAAACTCGCTACTAACGC	895~1045	153
<i>GRID1</i>	GGACGTGAGGAGTTTTAAGC	ACTTAACGCTCTCTTAAATAAACCG	-1140~-971	170
<i>HNRNPF</i>	TTTCGTGTTTCGGGAAATC	ACAACAAAAACCAATCCTACTCG	104~280	176
<i>HOPX</i>	TTTATTGGTTTTGTGGGTC	CGAAAACCTACCGACCGTA	369~566	197
<i>HOXC4</i>	AGGTATGGAGAGGGTTAAGTCGA	AAAACCGGATTCTAAAAAAAAACG	-1974~-2150	177
<i>HS3ST3A1</i>	GAGGGTAGTAAATTGTTGGTGAC	TTAAACGACTACTACGAAACCCG	-173~-23	150
<i>IHH</i>	GGTAGTTTTGAAGCGTTC	TTACTACTACTAATAATACCGACGAC	49~246	198
<i>ISL2</i>	GATTAGATTTTAAAGGTGGAGGC	GTCGAAACGAATACAACCGAA	-1648~-1450	198
<i>JPH1</i>	AGCGTTTTTGGGAAAGGC	TAAACGACGTAACGAACGACTCG	420~590	170
<i>KCNVI</i>	CGGTTAGTTGTTAAGGCC	TACCTCCAACGACAAAACG	343~550	207
<i>LAMA1</i>	GGATTTAAGTCGGGGAGGTC	CCAACTCTAACAAAAACCGTA	-368~-154	214
<i>LHX2</i>	GACGTTTCGTCGTC	CCCCAAACCGAAATTTCG	940~1080	140
<i>LHX9</i>	TTATTAAATATTAGAACGGTATTGGATC	GAATTCATCTTCTATACTCATCACGA	-2019~-1870	150
<i>LMX1A</i>	GTCGTTTCGAGATTGTAGTCG	TAAAACGTCGGAAAAACC	-811~-653	158
<i>LOC283392</i>	TTTATTTATTTGAAGGGATGGTC	CAAAACCGCGAAACTACGAC	1498~1696	198
<i>LOC645323</i>	TTTTTGTAGAGCGTTGTTTC	CTATACACTCATTATACGCCCTCGAC	3011~3171	161
<i>MGC16275</i>	TTAGGTTGGTTGTTAGCC	ATCGAAAAACCTCGCGTC	658~866	209
<i>MGC45800</i>	TCGCGGGTTTTAGTTTTAC	AATCCTTCAAACTCCTTCGAA	1639~1800	162
<i>MRGPRF</i>	TTGTGAGTATTGTGTTGGTTAC	ATTACGACCTATCTATAACACGAA	1787~1998	211
<i>NES</i>	GACGTTTATGTGTGTTACGC	ACCTAAACCGCGCAAACCTACC	581~741	160
<i>NEUROG1</i>	GAGGAAGTCGGATAGGTATTGTC	ACAACCCCAAACTATTAAAAACGTA	147~337	190
<i>NEUROG2</i>	TTTGGTGAGTTGGCGTTTC	AATCTAATACACGATTACAAACGAC	927~1148	211
<i>NID2</i>	TGTAGGGGATTCGTTAGTTTATTAC	AACGACCGATACTATCGTCGTT	256~431	175
<i>NKX6_1</i>	CGGTATGTATATTAGGGTCGC	CTAACTAAACTACCTAACCGAACGC	585~761	176
<i>NOS1</i>	TTTCGTAAGTGGAGGTTAGGTC	AATACCAATCCCTAAAAACCGTT	964~1169	205
<i>NPPC</i>	TTATATTGGAGAGTGTAGGGGTGTC	CCGAAACGTTCTATATAACCGAA	-964~-782	182
<i>NPTX1</i>	TAGTTTTAAGTTTTGAGTCGG	GAAACTAACCCCTCCTACGAC	-1755~-1521	234
<i>ONCUT1</i>	AGTTAGTCGGTAAATATCGACGT	CGCACCTAACAAACACTCTACG	1544~1717	174
<i>OPRK1</i>	CGAGGATAGCGGTATTC	GAAAAAACACGAAAAACCGTT	397~573	176
<i>OTX2OS1</i>	ATAGGGGTCGATAGGGTTC	CCCCCAAATTAAATTAAACCGT	-673~-881	208
<i>OXTR</i>	TTTGGAGATTTCACGGACGG	GACTAACAAACCGAAACCG	-1321~-1149	173
<i>PADI2</i>	GAGATTCGGTTCGGTATTAC	AAACGCAAACACTAAACCGAC	-4~211	215
<i>PAX6</i>	ATTGTTTATGTAAAGTAGCGTCGG	AAACGAAAAAATACAAACGAACG	-529~-336	193
<i>PCSK6</i>	GGGTTTAGGGTGTAGGTTTC	CTCTACCTCGACGAACGAC	1041~1207	167
<i>PDZK1P1</i>	ATATTGTTGTTCTCGTAAGTTTCG	TCGTCCCGATAATATTACG	-2371~-2198	174
<i>PGAM2</i>	GTTTTCTGTTATCGGGGGC	AATACTAAACGAACCATCCGCA	208~423	215
<i>PGR</i>	TGGTGGTTAGCGGGGAGC	CCGCTTCTAAAAACAAACCTCG	208~424	191
<i>PHOX2A</i>	TTGGGATGCGCGGGATTTC	TAATTTAATTGAAAGTGTGCTTG	-355~-170	185
<i>MAFA</i>	TAAATATTGAAAGATCGATT	CCGCTTATACATACGTTAACG	195~467	272

<i>PROX1</i>	CGCGCGTTTATTAAAGTC	ACCCCAATAAACCTATATCAGC	-3038~ -2845	194
<i>PVT1</i>	TTTAGTAATTGGGAGGTGAAGC	CCGCCTACCTAATTCAAACGAT	-1163~ -977	187
<i>RADIL</i>	GTATAGTCGGAAAATTGAGGGTC	GATACCCGAAACCGACTAACG	315~475	160
<i>RASD1</i>	GTTGTCGTAGATGATTAGGGTAC	CTCTATACCCCTCTAAAAACGTT	692~858	166
<i>SCRT1</i>	TTAGCGTTCGTGGATACGTT	ACGCCTACCACGAAAATAA	-2015~ -1858	157
<i>SCUBE3</i>	TTGTGTCGGTGGATACGTT	AAATCGAAAAACGAAAATACGTC	-1044~ -872	170
<i>SFRP2</i>	AGTATGTCGGTTAGGGGAAGTC	AATCATAAAACAATACCACCCGAA	486~660	174
<i>SIM2</i>	GTTAGGAGCGTTGTTAGGTAC	TAAATTCTCGATATCACTACTCTG	-2001~ -1834	168
<i>SLC1A2</i>	ATTTCGTTTTAGGGATAAGGATC	AACTTACAAATAATACCGACTCGAA	989~1145	156
<i>SLC46A2</i>	TATGTTTAAATCGGTGTTACG	AATCACTCGAAAAATCACTCGAC	-5~ -182	187
<i>SLC6A20</i>	CGTTTTTCGGTTAGTGTGTC	AACAAACGTATAACGATTCCCGTA	198~356	158
<i>SLC6A3</i>	TAGGGTCGAATGGATTTCG	CGACGCACAAAACCTAAACG	-925~ -765	160
<i>SLC6A4</i>	TTAGGAGGGAGGGATTTTC	ACCCCTATATACCGTCCTATAAACG	-417~ -222	195
<i>SNX31</i>	GAAGTATTGTTGGGAATAGC	CAACTATAAAATAAAAAACCGAACG	599~748	149
<i>SP8</i>	GGTGGCGTTATTAGGTGTC	ATTTACAACAAACGTTCATACG	1538~1762	224
<i>TAL1</i>	TAGGTATAGGCGGGTTTCG	GAACGAAAATCGAAACGCT	-1574~ -1411	163
<i>TBX20</i>	CGGGAGAGGATATATTATCGC	ACTTTCTTTAAAACCTTAAACGCA	-907~ -761	146
<i>TLX1</i>	ACGGTTTGTTGTTGGTC	CCATATTACGTTATAAAAACCGA	371~554	183
<i>TMEM163</i>	ATGTAACGAGTTGGGTTTTTC	CCAACCTACCTTCAAACCGAC	488~659	172
<i>TMEM26</i>	TGGTGGTTAGCGGGGAGC	CCGCTTCTAAAAACAAACCTCG	414~611	191
<i>TOX</i>	CGATTTTTCTGGGAATGTATC	CCACGCCTAAAACTAATTAAACGTA	1314~1506	193
<i>TRIM9</i>	GGGGTATTAAATTTCGTGC	TACAACAAACGATCCCCGT	-1791~ -1642	145
<i>TUBB2B</i>	GAGAGAGATTAATTTCGTGC	GAATCATTGACCTTCTCG	484~663	179
<i>VAX1</i>	TCGTTGCGTACGTTGTTTC	ACGCCAATCTAAACTTTAAACGT	-2027~ -1797	230

附表 3. MethylCap-seq 甲基化文库的测序基本信息

表 5-8 BCC 和 BM 组的甲基化文库的测序结果基本信息

	BCC (5637+T24)	BM (BM1+BM2)
Reads	6,232,852	6,230,959
Total bases recalled	473,696,752	473,552,884
Base Q20	0.9143	0.909
Peaks ( <i>p</i> value<1e-3)	210,051	229,538
Relative high methylated DMR	70,432	83,690
DMR related to refGene	55,237	45,522
DMR related to CGI	21,179	1,945
DMR related to both refGene and CGI	4,256	401
DMR related to promoter	1,627	66

**附表 4. BSP 对甲基化文库的验证****表 5-9 BSP 对甲基化文库的结果的验证**

	-10 * lg (p-value)	mCG(CG)		mCG(CG) (%)		Consistence	
		BCC	BM	BCC	BM		
<i>DMR in BCC</i>							
1	<i>NKX6_1</i>	900.8	127/138	3/115	92.00%	2.60%	yes
2	<i>CNTFR</i>	829.3	103/132	2/110	78.00%	1.80%	yes
3	<i>BMP7</i>	748.6	75/144	0/144	52.10%	0.00%	yes
4	<i>TLX1</i>	682.7	92/156	5/156	59.00%	3.20%	yes
5	<i>PLTP</i>	621.1	47/78	5/78	60.30%	6.40%	yes
6	<i>LMX1A</i>	621.1	142/186	0/186	76.30%	0.00%	yes
7	<i>IRX</i>	576.2	84/144	3/144	58.30%	2.10%	yes
8	<i>SIM2</i>	549.5	97/114	3/114	85.10%	2.60%	yes
9	<i>TBX20</i>	498.7	63/90	1/90	70.00%	1.10%	yes
10	<i>TUBB2B</i>	455.3	111/120	6/144	92.50%	4.20%	yes
11	<i>HS3ST3A1</i>	440.1	114/168	0/140	67.90%	0.00%	yes
12	<i>AVPRIA</i>	439.8	104/130	0/130	80.00%	0.00%	yes
13	<i>BEND4</i>	425.7	136/150	5/180	90.70%	2.80%	yes
14	<i>ADRA1A</i>	424.3	90/105	1/126	85.70%	0.80%	yes
15	<i>CDO1</i>	423.8	49/165	3/165	29.70%	1.80%	yes
16	<i>KCNV1</i>	378.6	36/70	1/70	51.40%	1.40%	yes
17	<i>PROX1</i>	336.1	121/140	7/140	86.40%	5.00%	yes
18	<i>SLC6A20</i>	292.5	31/100	0/100	31.00%	0.00%	yes
19	<i>ECELI</i>	257.6	82/110	0/110	74.50%	0.00%	yes
20	<i>VAX1</i>	147.6	82/102	8/68	80.40%	11.80%	yes
21	<i>NOS1</i>	89.8	49/105	15/105	46.70%	14.30%	yes
22	<i>TAL1</i>	84.5	79/95	6/95	83.20%	6.30%	yes
<i>DMR IN BM</i>							
23	<i>IRS1</i>	197.3	0/84	51/84	0.00%	60.70%	yes
24	<i>DAZ3</i>	147.2	119/126	117/126	94.40%	92.90%	no

注: -10 \* lg (p-value), 直接来自 MACS, 用来指示差异甲基化的显著性。Consistence 表示 BSP 和 Methycap-seq 所反映的甲基化结果是否一致。

附表 5. 不同基因组合诊断模型的 ROC 表现

表 5-10 不同基因组合诊断模型的 ROC 表现

No.	Gene target sets	TP/FN	FP/TN	P n=212	SN	SP	PPV	NPV	ACC	AUC
		n=212	n=149		(%)	(%)	(%)	(%)	(%)	
1	VAX1	90/122	7/142	< 0.0001	42.45	95.30	92.78	53.79	64.27	0.689
2	VAX1 KCNV	144/68	15/134	< 0.0001	67.92	89.93	90.57	66.34	77.01	0.789
3	VAX1 KCNV PROX1	156/56	16/133	< 0.0001	73.58	89.26	90.70	70.37	80.06	0.808
4	VAX1 KCNV PROX1 TAL1	163/49	17/132	< 0.0001	76.89	88.59	90.56	72.93	81.72	0.824
5	VAX1 KCNV PROX1 TAL1 ECELI	167/45	20/129	< 0.0001	78.77	86.58	89.30	74.14	81.99	0.821
6	VAX1 KCNV PROX1 TAL1 ECELI SLC6A20	170/42	23/126	< 0.0001	80.19	84.56	88.08	75.00	81.99	0.820
7	VAX1 KCNV PROX1 TAL1 ECELI SLC6A20 TMEM26	171/41	27/122	< 0.0001	80.66	81.88	86.36	74.85	81.16	0.813
8	VAX1 KCNV PROX1 TAL1 ECELI SLC6A20 TMEM26 LMX1A	172/40	28/121	< 0.0001	81.13	81.21	86.00	75.16	81.16	0.812

## 附表 6. 与膀胱癌已有甲基化诊断模型比较

表 5-11 膀胱癌已有甲基化诊断模型比较

	Study (author, year,reference)	Target panel	Technique	Sensitivity (N)	Specificity(N)
1	Michael. Chan et al, 2002(33)	<i>DAPK, RARβ, E-cadherin, and p16</i>	MSP	90.9% (22)	76.5%(17)
2	Cristina Battagli et al, 2003 (34)	<i>APC, RASSF1A, and p14<sup>ARF</sup></i>	MSP	87% (45)	100% (6)
3	Martin Friedrich et al, 2004(35)	<i>DAPK, BCL2, and TERT</i>	MethyLight	78% (37)	100%(20)
4	Shinji Urakami et al, 2006 (36)	<i>sFRP-1, sFRP-2, sFRP-4, and sFRP-5, Wif-1, and Dkk-3</i>	Nested MSP	77.2% (24)	66.7% (20)
5	Yu et al ,2007 (17)	<i>SALL3, CFTR, ABC6, HPR1, RASSF1A, MT1A, RUNX3, ITGA4, BCL2, ALX4, MYOD1, DRM, CDH13, BMP3B, CCNA1, RPRM, MINT1, BRACA1</i>	MSP	91.7% (132)	87% (23)
6	Vera L. Costa et al, 2010 (31)	<i>GDF15, TMEFF2, and VIM</i>	MSP	94% (51)	100% (20)
7	Thomas Reinert et al, 2011(19)	<i>ZNF154, POU4F2, HOXA9, and EOMES</i>	MSP	84% (174)	96% (59)
8	Woonbok Chung et al, 2011(26)	<i>MYO3A, CA10, NKX6-2, and DBC1 or SOX11</i>	qPCR	81% (128)	97% (110)
9	This study, Yu et al, 2011	<i>CFTR, VAX1, KCNV1, PROX1, and TAL1</i>	MSP	88.68% (212)	87.25% (149)

## 5.7 参考文献

- [1] Schneeweiss, S., M. Kriegmair, and H. Stepp. *Is everything all right if nothing seems wrong? A simple method of assessing the diagnostic value of endoscopic procedures when a gold standard is absent* [J]. *J Urol*, 1999. **161**(4);1116-1119.
- [2] Volpe, A., M. Racioppi, D. D'Agostino, E. Cappa, et al. *Bladder tumor markers: a review of the literature* [J]. *Int J Biol Markers*, 2008. **23**(4);249-261.
- [3] Wawroschek, F. and P. Rathert. *Urine cytology* [J]. *Urologe A*, 1995 **34**;69 - 75.
- [4] Siegel, R., C. Desantis, and A. Jemal. *Colorectal cancer statistics, 2014* [J]. *CA: a cancer journal for clinicians*, 2014. **64**(2);104-117.
- [5] Jin, F., S.S. Devesa, W.H. Chow, W. Zheng, et al. *Cancer incidence trends in urban shanghai, 1972-1994: an update* [J]. *Int J Cancer*, 1999. **83**(4);435-440.
- [6] Kaufman, D.S., W.U. Shipley, and A.S. Feldman. *Bladder cancer* [J]. *Lancet*, 2009. **374**(9685);239-249.
- [7] Mitra, A.P. and R.J. Cote. *Molecular pathogenesis and diagnostics of bladder cancer* [J]. *Annual review of pathology*, 2009. **4**;251-285.
- [8] Yeung, C., T. Dinh, and J. Lee. *The Health Economics of Bladder Cancer: An Updated Review of the Published Literature* [J]. *PharmacoEconomics*, 2014.
- [9] Siegel, R., J. Ma, Z. Zou, and A. Jemal. *Cancer statistics, 2014* [J]. *CA: a cancer journal for clinicians*, 2014. **64**(1);9-29.
- [10] Berger, S.L. *The complex language of chromatin regulation during transcription* [J]. *Nature*, 2007. **447**(7143);407-412.
- [11] Reik, W. *Stability and flexibility of epigenetic gene regulation in mammalian development* [J]. *Nature*, 2007. **447**(7143);425-432.
- [12] Feinberg, A.P. *Phenotypic plasticity and the epigenetics of human disease* [J]. *Nature*, 2007. **447**(7143);433-440.
- [13] Baylin, S.B. and J.E. Ohm. *Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction?* [J]. *Nature reviews. Cancer*, 2006. **6**(2);107-116.

- [14] Eden, A., F. Gaudet, A. Waghmare, and R. Jaenisch. *Chromosomal instability and tumors promoted by DNA hypomethylation* [J]. Science, 2003. **300**(5618);455.
- [15] Bird, A.P. *CpG-rich islands and the function of DNA methylation* [J]. Nature, 1986. **321**(6067);209-213.
- [16] Dumache, R., D. David, A. Kaycsa, R. Minciu, et al. *Genetic and epigenetic biomarkers for early detection, therapeutic effectiveness and relapse monitoring in bladder cancer* [J]. Revista medico-chirurgicala a Societatii de Medici si Naturalisti din Iasi, 2011. **115**(1);163-167.
- [17] Wilhelm-Benartzi, C.S., D.C. Koestler, E.A. Houseman, B.C. Christensen, et al. *DNA methylation profiles delineate etiologic heterogeneity and clinically important subgroups of bladder cancer* [J]. Carcinogenesis, 2010. **31**(11);1972-1976.
- [18] Reinert, T., C. Modin, F.M. Castano, P. Lamy, et al. *Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers* [J]. Clinical cancer research : an official journal of the American Association for Cancer Research, 2011. **17**(17);5582-5592.
- [19] Brinkman, A.B., F. Simmer, K. Ma, A. Kaan, et al. *Whole-genome DNA methylation profiling using MethylCap-seq* [J]. Methods, 2010. **52**(3);232-236.
- [20] 徐向红, *EPAS1 基因转录水平调控机制的研究*[D], 2014, 复旦大学: 上海.
- [21] Zhao, Y., S. Guo, J. Sun, Z. Huang, et al. *Methylcap-seq reveals novel DNA methylation markers for the diagnosis and recurrence prediction of bladder cancer in a Chinese population* [J]. PLoS One, 2012. **7**(4);e35175.
- [22] Zhou, X., *Establishing the transgenic mouse for a liver-specific conditional Cre recombinase and lung cancer methylome*[D], 2010, Fudan University: Shanghai.
- [23] Li, H. and R. Durbin. *Fast and accurate short read alignment with Burrows-Wheeler transform* [J]. Bioinformatics, 2009. **25**(14);1754-1760.
- [24] Zhang, Y., T. Liu, C.A. Meyer, J. Eeckhoute, et al. *Model-based analysis of ChIP-Seq (MACS)* [J]. Genome Biol, 2008. **9**(9);R137.

- [25] Valouev, A., D.S. Johnson, A. Sundquist, C. Medina, et al. *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data* [J]. Nat Methods, 2008. **5**(9);829-834.
- [26] Jothi, R., S. Cuddapah, A. Barski, K. Cui, et al. *Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data* [J]. Nucleic acids research, 2008. **36**(16);5221-5231.
- [27] Zhang, X., G. Robertson, M. Krzywinski, K. Ning, et al. *PICS: probabilistic inference for ChIP-seq* [J]. Biometrics, 2011. **67**(1);151-163.
- [28] Feng, J., T. Liu, B. Qin, Y. Zhang, et al. *Identifying ChIP-seq enrichment using MACS* [J]. Nature protocols, 2012. **7**(9);1728-1740.
- [29] Cottrell, S.E. and P.W. Laird. *Sensitive detection of DNA methylation* [J]. Ann N Y Acad Sci, 2003. **983**;120-130.
- [30] Yu, J., H.Y. Zhang, Z.Z. Ma, W. Lu, et al. *Methylation profiling of twenty four genes and the concordant methylation behaviours of nineteen genes that may contribute to hepatocellular carcinogenesis* [J]. Cell Res, 2003. **13**(5);319-333.
- [31] Li, L.C. and R. Dahiya. *MethPrimer: designing primers for methylation PCRs* [J]. Bioinformatics, 2002. **18**(11);1427-1431.
- [32] Yu, J., T. Zhu, Z. Wang, H. Zhang, et al. *A novel set of DNA methylation markers in urine sediments for sensitive/specific detection of bladder cancer* [J]. Clinical cancer research : an official journal of the American Association for Cancer Research, 2007. **13**(24);7296-7304.
- [33] Ma, K., *The DNA methylation regulated microRNAs perspective of the mechanistic insights in cancer chemosensitivity: the 5-FU hepatocellular carcinoma paradigm* [D], 2011, Fudan University: Shanghai.
- [34] Sun, J., *Hypermethylated SFRP1, but none of other nine genes "informative" for the Western countries is valuable for bladder cancer detection in the mainland of China* [D], 2009, Fudan University: Shanghai.
- [35] Dessaix, R.B. and C.B. Pipper. *["R"--project for statistical computing]* [J]. Ugeskr Laeger, 2008. **170**(5);328-330.

- [36] Chung, W., J. Bondaruk, J. Jelinek, Y. Lotan, et al. *Detection of bladder cancer using novel DNA methylation biomarkers in urine sediments [J]*. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2011. **20**(7);1483-1491.
- [37] Kim, W.J. and Y.J. Kim. *Epigenetics of bladder cancer [J]*. Methods in molecular biology, 2012. **863**;111-118.
- [38] Rodriguez-Paredes, M. and M. Esteller. *Cancer epigenetics reaches mainstream oncology [J]*. Nature medicine, 2011. **17**(3);330-339.
- [39] Costa, V.L., R. Henrique, S.A. Danielsen, S. Duarte-Pereira, et al. *Three epigenetic biomarkers, GDF15, TMEFF2, and VIM, accurately predict bladder cancer from DNA-based analyses of urine samples [J]*. Clinical cancer research : an official journal of the American Association for Cancer Research, 2010. **16**(23);5842-5851.
- [40] Aleman, A., L. Adrien, L. Lopez-Serra, C. Cordon-Cardo, et al. *Identification of DNA hypermethylation of SOX9 in association with bladder cancer progression using CpG microarrays [J]*. British journal of cancer, 2008. **98**(2);466-473.
- [41] Wolff, E.M., Y. Chihara, F. Pan, D.J. Weisenberger, et al. *Unique DNA methylation patterns distinguish noninvasive and invasive urothelial cancers and establish an epigenetic field defect in premalignant tissue [J]*. Cancer research, 2010. **70**(20);8169-8178.
- [42] Urakami, S., H. Shiina, H. Enokida, T. Kawakami, et al. *Combination analysis of hypermethylated Wnt-antagonist family genes as a novel epigenetic biomarker panel for bladder cancer detection [J]*. Clinical cancer research : an official journal of the American Association for Cancer Research, 2006. **12**(7 Pt 1);2109-2116.
- [43] Friedrich, M.G., D.J. Weisenberger, J.C. Cheng, S. Chandrasoma, et al. *Detection of methylated apoptosis-associated genes in urine sediments of bladder cancer patients [J]*. Clinical cancer research : an official journal of the American Association for Cancer Research, 2004. **10**(22);7457-7465.

- [44] Battagli, C., R.G. Uzzo, E. Dulaimi, I. Ibanez de Caceres, et al. *Promoter hypermethylation of tumor suppressor genes in urine from kidney cancer patients* [J]. Cancer research, 2003. **63**(24);8695-8699.
- [45] Chan, M.W., L.W. Chan, N.L. Tang, J.H. Tong, et al. *Hypermethylation of multiple genes in tumor tissues and voided urine in urinary bladder cancer patients* [J]. Clinical cancer research : an official journal of the American Association for Cancer Research, 2002. **8**(2);464-470.
- [46] Versmold, B., J. Felsberg, T. Mikeska, D. Ehrentraut, et al. *Epigenetic silencing of the candidate tumor suppressor gene PROX1 in sporadic breast cancer* [J]. Int J Cancer, 2007. **121**(3);547-554.
- [47] Dong, W., L. Feng, Y. Xie, H. Zhang, et al. *Hypermethylation-mediated reduction of LMX1A expression in gastric cancer* [J]. Cancer science, 2011. **102**(2);361-366.
- [48] Cardoso, B.A., S.F. de Almeida, A.B. Laranjeira, M. Carmo-Fonseca, et al. *TALI/SCL is downregulated upon histone deacetylase inhibition in T-cell acute lymphoblastic leukemia cells* [J]. Leukemia, 2011. **25**(10);1578-1586.
- [49] Tsou, J.A., J.S. Galler, A. Wali, W. Ye, et al. *DNA methylation profile of 28 potential marker loci in malignant mesothelioma* [J]. Lung cancer, 2007. **58**(2);220-230.
- [50] Bharti, K., M. Gasper, S. Bertuzzi, and H. Arnheiter. *Lack of the ventral anterior homeodomain transcription factor VAX1 leads to induction of a second pituitary* [J]. Development, 2011. **138**(5);873-878.
- [51] Ebihara, M., H. Ohba, M. Kikuchi, and T. Yoshikawa. *Structural characterization and promoter analysis of human potassium channel Kv8.1 (KCNV1) gene* [J]. Gene, 2004. **325**;89-96.
- [52] Benoit, A., M.A. Vargas, L. Desgroseillers, and G. Boileau. *Endothelin-converting enzyme-like 1 (ECEL1) is present both in the plasma membrane and in the endoplasmic reticulum* [J]. The Biochemical journal, 2004. **380**(Pt 3);881-888.
- [53] Town, L., E. McGlinn, T.L. Davidson, C.M. Browne, et al. *Tmem26 is dynamically expressed during palate and limb development but is not required for embryonic survival* [J]. PLoS One, 2011. **6**(9);e25228.

## 第六章 总结及展望

### 6.1 总结

DNA 甲基化是一种古老的表现修饰系统。其在进化上十分保守，从单细胞原核生物到高等哺乳动物广泛存在。DNA 甲基化修饰在人类基因组中具有重要的结构及生物学意义，几乎参与人类生命中的所有重要进程，包括有丝分裂，减数分裂，组织形成，器官及个体发育及衰老及各种复杂疾病，特别是肿瘤。此外 DNA 甲基化参与减数分裂同源染色体交叉互换，影响碱基突变频率等，因此也是群体遗传学的重要研究对象。此外，DNA 甲基化对外界环境的响应的柔性介于基因组序列变异和细胞内游离分子（如 mRNA 和蛋白质等）之间。这种适中的可塑性使得 DNA 甲基化可以作为以环境为主要引发因素疾病的重要致病因子，易感因素或标记物。近年来大量的 GWAS 证实除了来自基因组序列的变异外，对于复杂疾病尚存在大量的遗传力丢失。甲基化能够调控基因的表达，参与染色体构象重塑，因此是疾病易感性的重要来源。相对于遗传序列，DNA 甲基化的可逆变化更为容易，大量针对 DNA 甲基化的潜在的药物作用靶点的临床实验正在开展，说明了 DNA 甲基化在复杂疾病诊断，治疗中具有重要的作用。

甲基化肿瘤标记物在不同病理状态的较大差异及快速便捷的 DNA 甲基化检测技术共同推动了 DNA 诊断标记物市场的快速发展，目前欧洲 Epigenomics 公司已经成功开发了一系列基于 DNA 甲基化标记物的肿瘤早期筛查方法，并且已经通过了美国 FDA 的一系列验证过程，进入市场的前景广阔。尽管如此，目前也仅有为数不多的几种疾病产生了一些很有希望的诊断标记物组合，这说明基于 DNA 甲基化的生物标记物，特别是诊断标记物，仍然面临着很多技术方面的挑战。

本研究通过多种 DNA 甲基化研究平台对 DNA 在肿瘤中的标记物潜力进行了探索。从样本来源角度，第二章，第三章以肺癌样本为研究模型，第五章以膀胱癌为模型，第四章以泛癌样本为模型（11 种人类肿瘤，包括：肾透明细胞癌、浸润性乳腺癌、甲状腺癌、头颈部鳞状细胞癌、前列腺癌、肝癌、肾透明细胞癌、肺腺癌、结肠癌、子宫内膜癌、肺鳞癌）。

本研究的第二章采用 Meta 分析的方法对基于候选基因研究策略的 APC 基因启动子区高甲基化对肺癌诊断效能进行了系统及定量的评估。APC 基因启动

子区 DNA 甲基化与非小细胞肺癌之间的关系在过去的 20 年间已经被广泛地报道，大量的研究都认为 APC 启动子甲基化可以作为一个的有效的生物标志物用于非小细胞肺癌诊断。然而，事实情况是在不同的研究中，APC 在肿瘤中的甲基化率存在较大差别。这可能是由样本中的性别比例，年龄分布，样本遗传结构差异，DNA 甲基化检测方法等其他一系列差异造成的。APC 基因甲基化对非小细胞肺癌的诊断能力的定量评估有利于对 DNA 甲基化生物标记物的临床表现进行细致的研究和认识。文献检索发现了截至目前为止，已经有 17 篇关于 APC 甲基化在肺癌中的甲基化状态进行了研究。Meta 合并分析显示，APC 启动子区 DNA 甲基化在非小细胞肺癌中的合并 OR 采用随机效应模型和固定效应模型分别为 4.67 (95% CI: 2.66-8.22, Z=5.35, P<0.0001) 和 2.74 (95% CI: 1.99-3.23, Z=8.10, P<0.0001)，说明 APC 在肺癌组织中的呈现显著的高甲基化状态。亚组分析显示诊断时年龄，所采用引物是异质性的重要来源，腺癌样本比例和实验设计对照样本类型中存在显著性的 OR 差异。Meta 回归分析确认了上述潜在异质性来源因素中的 2 项：亚组分析显示诊断时年龄，所采用引物。在采用固定效应模型的前提下，可以对整个研究的汇总敏感性和汇总特异性进行分析，结果显示汇总敏感性和汇总特异性分别为 0.548 (95% CI: 0.42-0.67, p<0.0001) 和 0.78 (95% CI: 0.62-0.88, P<0.0001)。实体组织亚组 (0.61, 95% CI: 0.45-0.75) 的诊断敏感性明显高于血清亚组 (0.396, 95% CI: 0.26-0.56)。而血清亚组的特异性 (0.92, 95% CI: 0.86-0.96) 明显高于实体组织亚组 (0.68, 95% CI: 0.49-0.83)。这表明这种生物标志物用兼具诊断和早期筛查的双重潜力。综合受试者工作特征曲线分析 (SROC) 显示 APC 基因 DNA 甲基化对非小细胞肺癌的诊断模型的 AUC 为 0.64。采用修剪和填充方法消除发表偏倚的影响后 APC 基因启动子区 DNA 甲基化与非小细胞肺癌之间的关联性依然显著存在。敏感性分析显示出了分析结果具有很高的可信度和稳定性。累积 Meta 分析显示 APC 启动子区 DNA 甲基化与非小细胞肺癌的关联性是一个逐渐趋于稳定的状态。此外，采用 TCGA 中的非小细胞肺癌甲基化数据对多种情况的随机模拟也支持上述的一系列结论，也支持 APC 启动子区 DNA 甲基化在腺癌及其对照中具有更大的差异。因此，APC 启动子的甲基化状态与非小细胞肺癌显著相关，尤其对于肺腺癌。APC 的甲基化检测具有一定的肺腺癌的临床辅助诊断价值。正重要的是本研究对 DNA 甲基化临床诊断的推广中需要注意的问题进行了全面分析，发现包括年龄组成，所采用的引物，自体或异体对照，样本中腺鳞癌比例等都会对 DNA 甲基化测试的结论带来显著的异质性，需要在临床科研中引起足够的重视。

本研究的第三章采用整合跨平台甲基化芯片数据结合，标准化、批次效应校正等处理建立非小细胞肺癌标记物开发的 Discovery Dataset，根据此 Discovery Dataset 建立了一套最优的 5 甲基化位点组合。进而采用中国非小细胞肺癌样本进行了验证阶。目前影响 DNA 甲基化的临床应用的一个重要因素即单个研究的样本量有限。建立在较小样本量基础上的预测模型的稳定性会大大降低（stability），从而导致预测表现欠佳。此外前期的基于 DNA 甲基化的诊断标记物都采用候选基因的方法建立的，很难得到肺癌诊断标记物的最优组合。为此，目前的 DNA 甲基化诊断模型需要以更大样本量的肺癌全基因组甲基化数据为基础，采用变量选择的方法，获得最优的肺癌诊断模型。在本文第三章的研究中，来自 GEO 数据库和 TCGA 肿瘤基因组计划的包括 458 个样本的三个肺癌高通量全基因组甲基化芯片数据被整合。经过平台探针合并，标准化处理，批次效应消除，采用支持向量机的方法进行特征变量选择，最终确定了一个由 5 个甲基化位点的最优预测变量组合 (*AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1*)。为了检验诊断标记物组合的诊断效能，采用甲基化状态确定单核苷酸引物延伸技术 (MSD-SNuPET) 同时对 150 对中国人群的非小细胞肺癌/正常组织的上述 5 位点的甲基化状态进行了检测。结果显示这 5 个位点的甲基化状态和芯片数据的结果完全吻合，在肺癌中全部呈现异常甲基化状态。说明采用标准化，批次效应矫正等处理对跨平台芯片数据进行整合可以有效地用于肿瘤异常甲基化位点的检测。采用 logistic 回归，支持向量机，随机森林预测及 Bayes 树，结合 5 倍交叉验证的方法对肺癌的预测的结果显示，Logistic 回归模型的灵敏度，特异性，准确性，曲线下面积 (AUC) 分别为 78%，97%，87% 和 0.91。贝叶斯树模型具有最高的敏感性，特异性和准确性分别达到为 86.3%，95.7% 和 91%。为此 5 位点甲基化标记物 *AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* 和 *NTSR1* 可作为非小细胞肺癌诊断的有效组合，具有重要的临床应用价值。

本研究的第四章基于对泛癌数据全基因组 DNA 甲基化芯片数据的主成分、聚类、分类分析发现肿瘤全基因组 DNA 甲基化能够全面反映肿瘤样本，肿瘤之间的相似程度。目前对多种肿瘤进行同步/同时预测已经成为临床研究的重要目标之一。本研究采用包括 11 种人类肿瘤 1274 个瘤和对应癌旁组织的 HM450K 甲基化芯片数据，对人类肿瘤的泛癌甲基化谱式进行了描述。HM450K 芯片中 34 万个甲基化位点在整体数据的相关性绝大多数为中低相关，说明了甲基化位点在不同状态的样本中呈现很高的复杂度。样本相关性分析显示相同组织来源的样本具有较高的相关性。肿瘤样本之间的相关性远低于同组织来源的正常样本，印证了肿瘤样本相对于正常组织之间膨胀了的异质性。主成分分析显示前 10 个主成

分即可解释泛癌数据（1274 样本，349,049 探针）55.74% 的总变异，当选取前 120 个主成分时，即可解释数据总变异的 80%，说明该数据的方差为有效规律变量造成的而非大量噪音信号的叠加。样本个体聚类分析显示相同类型的肿瘤样本相互聚集；相同类型的正常组织距离较近，从而相互聚集，说明全基因组 DNA 甲基化谱式可以忠实地反应样本之间的相似性。肿瘤样本群体水平的聚类分析显示肺腺癌和肺鳞癌，肾透明细胞癌和肾乳头状细胞癌，前列腺癌和乳腺癌首先聚集在一起。对相邻探针位点之间的连锁分析发现 DNA 甲基化也具有类似 SNP 的连锁现象，连锁强度和位点之间的距离成反比。并且发现肿瘤样本中相邻 CpG 位点之间的相关性高于正常。泛癌样本差异甲基化位点分析显示只有 42 个位点在 11 种肿瘤中均呈现显著性差异。采用随机森林预测算法，对 11 种肿瘤和正常对照的 22 种状态的样本进行判定，结果显示采用 25 个预测变量即可实现对 22 个状态的较好预测，灵敏性，特异性和准确性分别为 72%，87% 和 80%。当预测变量数达到 75 个时预测模型的灵敏性，特异性和准确性可以到到 86%，94% 和 90%。因此可以推测即便对多种肿瘤同时预测，所需要的甲基化位点数也不会超过三位数。

本研究的第五章采用 MethylCap 结合二代测序技术建立肿瘤标记物建立了基于膀胱癌细胞系的全基因组甲基化谱式，及一系列潜在膀胱癌相关的 DNA 甲基化标记物。继而通过多阶段生物标记物验证流程对 104 个候选基因进行了筛选。虽然基于高密度 DNA 甲基化芯片技术对肿瘤标记物筛查的方法具有：位点特异，价格低廉等优势，但是在标记物筛查早期及诊断模型建立阶段，芯片技术存在容易遗漏潜在的重要的诊断标记物的问题。甲基化特异富集结合深度测序技术可以弥补 DNA 甲基化芯片的这一缺点，同时还能保证全基因组范围内搜索最优生物标记物对诊断及预后进行评估。本研究中尝试采用甲基化 DNA 结合结构域（MBD）富集技术结合第二代高通量测序（methylCap-Seq）的方法，建立常用膀胱癌全基因组甲基化谱式，进而需要膀胱癌相关的 DNA 甲基化标记物。基于膀胱癌细胞系 5637 和 T24 的全基因组异常甲基化区域分析发现了 1627 个显著异常的启动子区 CpG 岛。对其中 24 差异位点在建库样本中检验质，发现其吻合度为 96%。MBD methylCap-Seq 技术是一种稳定可靠的全基因组甲基化分析技术。通过多阶段生物标记物验证流程，对 104 个候选基因中，经过 509 个独立样本的验证，最终筛选出了 9 个膀胱癌相关 DNA 甲基化标记物。本研究发现，由 5 个基因（*VAX1*, *KCNV1*, *TAL1*, *PROX1* 和 *CFTR*）组成的甲基化标记物组合可以用来对 BC 及正常对照进行高效的预测。5 基因诊断模型对膀胱癌诊断的灵敏度、特异性、阳性预测值和阴性预测值分别为：88.68%，87.25%，90.82% 和

84.42%。临床双盲实验显示基于上述 5 甲基化标记物的诊断模型和膀胱镜检查具有极高的吻合度（81.25%）。两个独立的临床数据（回顾性数据和临床随访数据）都显示 *VAX1* 和 *LMX1A* 两个基因的异常甲基化与膀胱癌复发高度相关。此外，研究还发现 *ECELI* 和 *TMEM26* 两个基因的甲基化标记物与肿瘤的分化显著相关，提示这两个基因的甲基化可能参与肿瘤分化进展。本研究通过全基因甲基化谱式分析，小样本筛选，大样本验证，双重独立数据验证，临床双盲模拟等实验设计和分析对膀胱癌潜在的甲基化标记物进行了筛查，成功建立一系列具有潜在应用价值的膀胱癌甲基化标记物。

## 6.2 课题的意义与不足

对肿瘤的早期诊断、肿瘤进展的实时监控、药物治疗的指导和选择、治疗预后评估都需要相应肿瘤标记物的密切参与。DNA 甲基化作为最重要的生物标记物来源，具备标记物的基本要求。本课题采用多种方法及策略包括 Meta 分析，关联分析、预后分析等对肿瘤特异性 DNA 甲基化生物标记物的开发和应用进行了探讨，对肿瘤特异性 DNA 甲基化生物标记物在临床转化中需要注意的异质性来源进行了挖掘，并证实了肿瘤特异性 DNA 甲基化生物标记物不仅可以对单个肿瘤的诊断、复发、预后具有显著的判定功能，而且还有对多种肿瘤发生进行高危预警判定的潜力。不足之处在于，本课题从根本上还处于 Proof-of-principle 的阶段，真正可以在临幊上使用的标记物对肿瘤的早期诊断、实时监控需要在无创或微创采集中所获标本（尿液、血液、支气管冲洗液等）进行验证，而肿瘤药物治疗的指导和选择、治疗预后评估则可在石蜡标本中进行，以从根本上解决上述临幊问题。

## 6.3 展望

DNA 甲基化是进化上十分保守的表观修饰系统，参与细菌的“自身免疫”，参与植物的多种胁迫反应、参与动物的全方位生理及病理代谢。因此 DNA 甲基化可以作为优良的生物标记物。特别地，DNA 甲基化标记物在肿瘤诊断、监测、药物反应预测、预后评估中具有巨大的应用价值。目前 DNA 甲基化已经逐步有中等成熟程度的单一肿瘤诊断、预后标记物正逐步等待 FDA 进行审核和批准。可以预计未来五年将会有一系列基于 DNA 甲基化的生物标记物对肿瘤的一系列阶段进行检测。然而 DNA 甲基化仍然有很多吸引众多学者的领域亟待开发。具体表现在以下几个方面。

统一的多疾病诊断标记物组合的开发会成为下一个阶段 DNA 甲基化标记物开发的趋势。基于本研究四章的泛癌研究以及大量其他的研究表明：不同恶性肿瘤之间、免疫性疾病、炎症反应等具有完全不同的 DNA 甲基化谱式，因此理论上建立以血液 DNA 甲基化检测为基础的多疾病诊断或监测标记物具有很高的可行性。建立不同组织不同生理病理状态的特征谱式，可以实现血浆 DNA 甲基化谱式对疾病位置的估计，从而彻底颠覆目前只能首先确定具体的疾病器官，然后对其进行相应的影像学、化学、分子层面的相应监测。

基因组与环境的复杂交互机制将会得到异常丰富的阐述。GWAS 技术的成熟使得众多遗传疾病的易感基因组变异位点得到揭示。然后可以观察到不论对于何种疾病，只有极少部分的遗传力得到了解释。更多遗失遗传力没有得到揭示。因为对于复杂疾病基因组变异的贡献仅仅占据疾病发病的一部分贡献。外界相关环境暴露可以解释另一部分疾病贡献，即便如此，也没有任何一种疾病可以建立 100% 的致病因素模型。究其原因则是基因组变异和环境之间的交互在疾病的发病中占有重要的贡献。众所周知，以 DNA 甲基化等为主题的表观遗传系统提供了一个环境作用基因组的界面，从而使得基因组和环境之间的交互得以建立和稳定遗传。因此采用以 SNP 为代表的 GWAS 和 DNA 甲基化为代表的 mGWAS 的联系分析，借助因果推断等统计分析，可以寻找 DNA 甲基化介导的易感 SNP 位点，从而全面揭示与环境互作进而产生特定表型的基因组变异位点。

基于 DNA 甲基化谱式的中间表型的开发也会极大地促进遗传学的研究。表型的确定和分类是遗传学研究的重要组成。一个明确的表型，特别是中间表型可以保证遗传学研究事半功倍。由于 DNA 甲基化可以指征器官，组织及个体的各种生理及病理状态，为此可以通过建立相应的 DNA 甲基化特征谱式，对不同的状态进行指证，从而建立更加丰富的表型来源，使得遗传学的关联研究、连锁分析等更具有更高的效能。

## 全文图目录

图 1-1 哺乳动物发育过程中的表观遗传重编程及精子发育过程 .....	20
图 1-2 精子与卵细胞的发育过程中甲基化差异动态变化过程 .....	21
图 1-3 哺乳动物精子发育过程中的表观遗传重编程循环 .....	21
图 2-1 APC 基因 DNA 甲基化与非小细胞肺癌关联性强度的合并估计 .....	46
图 2-2 亚组的 Meta 分析 .....	48
图 2-3 漏斗图分析用于展示发表偏倚情况 .....	51
图 2-4 采用插补法对虚拟遗漏研究进行插补后再次进行合并比值比分析 .	52
图 2-5 合并比值的敏感性分析 .....	53
图 2-6 累计 Meta 分析显示逐渐稳定的结论 .....	54
图 2-7 综合受试者工作特征曲线分析 .....	55
图 3-1 甲基化标记物开发流程图 .....	72
图 3-2 多位点 MSD-NEuTEP 技术分析结果示意图 .....	75
图 3-3 支持向量原理机示意图 .....	78
图 3-4 采用 PCA 的方法展示批次效应处理前后样本 .....	79
图 3-5 聚类分析展示批次效应校正前后的生物特征变化 .....	80
图 3-6 芯片 CpG 位点在肿瘤和正常组织中的甲基化频率比较图 .....	81
图 3-7 MSD-SNuPET 技术对五个甲基化标记物的甲基化状态进行检测 ....	82
图 3-8 NTSR1 和 GALR1 相互作用的蛋白 .....	84
图 3-9 诊断模型随预测变量的增加的表现情况 .....	87
图 4-1 2014 美国肿瘤统计：发病率和死亡率前 10 名 .....	94
图 4-2 全基因组探针的甲基化特征 .....	101
图 4-3 全基因组平均甲基化信号在肿瘤和正常中的比较 .....	102
图 4-4 样本皮尔森相关性热图 .....	103
图 4-5 样本平均皮尔森相关性特征 .....	105
图 4-6 多维尺度分析展示癌症样本之间的距离 .....	108
图 4-7 维尺度分析展正常样本之间的距离 .....	108
图 4-8 根据全基因组甲基化数据对 1274 样本的聚类分析 .....	109
图 4-9 基于聚类分析的肿瘤组织相似性 .....	110
图 4-10 基因的肿瘤相关分值的分布情况 .....	112
图 4-11 肿瘤和正常基因组相邻 CpG 区域平均相关系数的抽样分布.....	113

图 4-12 以肺腺癌为例展示全基因组 CpG 位点甲基化状态相关性.....	113
图 4-13 基于区段组合法的随机森林模型变量重要性分布及预测精度 .....	114
图 4-14 基于差异甲基化位点的随机森林变量重要性分布及预测精度 .....	115
图 4-15 差异甲基化位点及预测功能位点探针的相关性 .....	115
图 4-16 预测变量个数与多肿瘤预测准确度关系 .....	116
图 4-17 最迫切需要早期诊断标记物的四种肿瘤 .....	117
图 4-18 Manhattan 图显示肺腺癌和鳞癌异常甲基化位点 .....	118
图 4-19 基于配对或非配对数据分析肺癌甲基化差异基因 .....	119
图 4-20 非腺癌和肺鳞癌相同位点甲基化差异比较 .....	120
图 4-21 非腺癌和肺鳞癌相同位点甲基化差异 P-value 比较 .....	120
图 5-1 膀胱癌器官结构示意图 .....	140
图 5-2 MBD-methylCap-seq 甲基化谱式的构建的原理简图.....	144
图 5-3 亚硫酸盐介导的胞嘧啶到尿嘧啶转变 .....	146
图 5-4 膀胱癌全基因组甲基化在不同基因组元件下的谱式特征 .....	152
图 5-5 采用 BSP 技术对 MBD-methylCap 建立的甲基化谱式进行验证 ...	153
图 5-6 基于尿液样本筛选潜在的膀胱癌甲基化标志物的筛选策略 .....	155
图 5-7 基因与膀胱癌临床指标之间的关系 .....	157
图 5-8 Kaplan-Meier 生存曲线显示甲基化与膀胱癌复发的相关性.....	160

## 全文表目录

表 2-1 候选文献的基本特征 .....	45
表 2-2 17 篇文章主要涉及的引物序列 .....	45
表 2-3 混合效应模型下对主要混淆因素进行亚组分析 .....	49
表 2-4 基于随机效应模型的 Meta 回归对潜在的混淆因素进行分析 .....	50
表 2-5 TCGA 相关样本的基本资料 .....	56
表 2-6 病例对照分析中的 TCGA 相关样本的资料对比 .....	56
表 2-7 甲基化芯片中六个 APC 相关 Probe 的信息 .....	56
表 2-8 基于甲基化芯片的 APC 在肺腺癌中的差异甲基化 .....	57
表 2-9 基于甲基化芯片的 APC 在肺鳞癌中的差异甲基化 .....	57
表 2-10 仿真状态下不同腺癌比例样本下的 APC 基因差异表达情况 .....	58
表 2-11 仿真状态下不同腺癌比例样本下的 APC 基因肺癌关联性分析 .....	58
表 2-12 17 篇文章的详细特征 .....	61
表 3-1 整合分析中收集的数据集 .....	70
表 3-2 本研究中非小细胞样本资料 .....	73
表 3-3 肺癌差异甲基化位点 .....	83
表 3-4 同预测模型采用五倍交叉验证的方法的表现情况 .....	85
表 4-1 配对的癌-癌旁 DNA 甲基化全基因数据信息 .....	95
表 4-2 全基因组平均甲基化信号在肿瘤和正常中的比较 .....	102
表 4-3 肿瘤组织之间的平均相关性 .....	104
表 4-4 正常组织之间的平均相关性 .....	104
表 4-5 同一组织不同类型样本之间的相关性 .....	105
表 4-6 肿瘤差异基因及肿瘤特异性差异甲基化位点个数统计 .....	111
表 4-7 采用前 25 和 75 个变量对应的预测模型表现情况 .....	116
表 4-8 肺腺癌和鳞癌异常甲基化基因的 Gene ontology 模块 .....	121
表 4-9 肺腺癌最重要的甲基化诊断位点 .....	122
表 4-10 肺鳞癌最重要的甲基化诊断位点 .....	123
表 4-11 乳腺癌最重要的甲基化诊断位点 .....	124
表 4-12 结肠癌最重要的甲基化诊断位点 .....	125
表 4-13 头颈部鳞状细胞癌最重要的甲基化诊断位点 .....	126
表 4-14 肾透明细胞癌最重要的甲基化诊断位点 .....	127
表 4-15 肾乳头状细胞癌最重要的甲基化诊断位点 .....	128

---

表 4-16 肝癌最重要的甲基化诊断位点 .....	129
表 4-17 前列腺癌最重要的甲基化诊断位点 .....	130
表 4-18 甲状腺癌最重要的甲基化诊断位点 .....	131
表 4-19 子宫内膜癌最重要的甲基化诊断位点 .....	132
表 4-20 高肿瘤相关分值基因 (TRS $\geq$ 10) 的 Gene ontology 分析 .....	135
表 5-1 2014 年主要肿瘤的新发病例和死亡病例的估计(美国).....	139
表 5-2 尿液样本对于的膀胱癌病人及正常对照的临床病理资料 .....	142
表 5-3 多因素 logistic 回归估计 <i>LMX1A</i> 和 <i>VAX1</i> 的肿瘤复发危险比 .....	158
表 5-4 基于 cox 回归的 <i>VAX1</i> 和 <i>LMX1a</i> 甲基化与预后复发的关联 .....	158
表 5-5 9 个候选标记物在肿瘤-正常和肿瘤-良性病变诊断能力.....	164
表 5-6 BSP 所涉及的引物对序列 .....	165
表 5-7 MSP 所涉及的引物对序列 .....	166
表 5-8 BCC 和 BM 组的甲基化文库的测序结果基本信息 .....	168
表 5-9 BSP 对甲基化文库的结果的验证 .....	169
表 5-10 不同基因组合诊断模型的 ROC 表现 .....	170
表 5-11 膀胱癌已有甲基化诊断模型比较 .....	171

## 致谢

光阴荏苒，时光总是在不经意间从身旁匆匆流逝。此时此境，让我感到距离上次撰写本科论文宛如昨日，不知不觉中五年的博士生涯也结束了。享受着刚刚完成论文的激动，往事一幕幕浮现在眼前：从拿到推免录取通知书，到肿瘤所小白楼夜以继日的实验生活，到通过金老师的面试转入生科院，再到休斯顿的交流。每一幕都是那么让我难忘。一路走来，有太多太多的感谢致帮助过我的老师和同学。求学之路遇到你们是我此生最大的幸福。

首先，最需要感谢的是我的导师金力教授，虽然五年时间中，真正和金老师共处的时间不是很多。然而金老师所释放的正能量却感染着他的每一个学生。跟随金老师求学的这段时间让我体验到了对知识探索的满足和体验到了人际交往的魅力。五年前，我在上海市肿瘤研究所完成了本科毕业设计，课题就是基于 DNA 甲基化标记物对肝癌进行分子诊断。之后顺利地进入复旦攻读博士。博士的课题延续了本科的工作，这是一个非常接近临床的课题，思路异常简单，即找到一些最优的 DNA 甲基化标记物对肝癌进行分子诊断。对于实验科学来说，我们尝试一个又一个可能的基因，试图完成对肿瘤的完美早期诊断，以解救众多癌症患者。我相信导师的直觉，肯定存在一个合适的基因组合能够实现我们的想法。整整两年的时间，采用笨拙的 PCR 技术，我完成了近 200 个基因的尝试。那些昼夜的尝试，颇有愚公移山的精神，让我永生难忘。2011 年朱景德教授退休，我幸运地通过了金力教授的面试，成为了金力教授的博士生。刚刚进入新实验室的我，非常希望融入新的环境，希望找到自己和新学科的交叉点。我用了将近一年的时间修了遗传学专业的相关基础课和必修课，用最大的努力试图跟随金老师的步伐，进入群体遗传学。遗憾的是，一年后，我选择了放弃。群体遗传学就像一片无边的乐园，那里草长莺飞，鸟语花香。可是这片乐园前面有一座大山，当你爬不过这座山的时候，一切都遥不可及。正当我非常苦闷寻找不到出路的时候，王久存教授给我了让我永远不会忘记的建议，让我多听听各方面的报告，多寻找一些新的思路。此时恰逢一年一度的答辩季到了，陈超师兄，董华师姐，胡鹏飞师兄的博士答辩完全开启了我博士工作的新起点。通过他们的答辩，我知道了变量选择，知道了分类分析 (classification)，聚类分析 (cluster analysis) 等此刻才恍然大悟，原来博士前两年的工作正是对肿瘤的分析分析和变量选择的工作。顿时间满血复活，我对博

士课题迅速地再定位，延续原来的方向，采用统计学，流行病学和机器学习的研究方法开启了新的征程。在这短短的一年中，通过金力老师和王久存老师的帮助，我在学术上得到了一系列师兄的帮助，包括徐书华教授，何云刚研究员，王磊教授，倪挺教授，苏志熙老师，王一博士。再后来在熊墨森教授课题组访学的一年半时间里，又认识了 Frank Arnett, Joshua Akey。和他们的聊天，总能聊到金老师的故事，让我总是感到世界真的很小很小。对于一个生物背景出身的学生，甚至我的博士前两年都是以实验为主的科研，金老师提供了一个近乎完满的环境，让我逐渐转入以数据分析为主的过渡。多学科人才的汇集包括数理统计，人类进化，群体遗传，流行病学，人类学，考古学，语言学的广泛交流，成就了小平房的今天也成就了所有从这里走出来的学子。对此我只能默默地对您说：谢谢金老师，您辛苦了。掩卷沉思总能会回忆起王久存老师三年来给与的所有指导：小平房深夜修改论文，论文屡次被拒后的鼓励，鼎力支持我去美国的交流。可以说没有您如此支持，无微不至的帮助，真不知如此艰难的博士训练何以完成。在休士顿的学习经历，让我被熊墨森老师那种近乎疯狂的科研精神所感动。熊老师对我的指导来的各个方面，从生活，学习，人生参悟及科研生涯的规划。Boston 开会期间，熊老师一次又一次的引荐，让我认识了很多同行学者，与他们合作的文章都已经发表，让我充分体验到通过合作共同学习进步的效率远远高于一个人战斗。从金力教授，王久存教授和熊墨森教授的身上，我体会到了一个科研工作者应该有的执着，坚毅，勤奋和质朴。三位老师传授的具体知识，随着时间的流逝也许会过时，但这种榜样的力量将永远激励我的前行。

论文的第二章得到了徐书华老师，靳文菲师兄，鲁东胜和袁媛等的帮助。第三章得到了陈超师兄和严凤阳师弟等的大力帮助。第四章得到了王久存教授，谭立行师妹和濮伟林师弟等的帮助。第五章得到了余坚教授和赵仰星老师的大力帮助。这里还要特别感谢余坚教授，赵仰星教授，一路走来，真感谢你的不懈支持。感谢张红宇老师，是您让我再次喜欢上了编程，Perl 语言的学习让我终生受益。

除此之外我还要感谢小平房的各位给与很多帮助的老师们：谭静泽，卢大儒老师，姜正文博士，王红艳老师，李辉老师，钱吉老师，杨亚军老师，安宇老师。以及一起讨论的师兄师姐师弟师妹们，包括复旦：萧恺昌，王筱恬，郑鸿翔，陈兴栋，洪胜君，易会广，彭倩倩，潘学栋，魏馨竹，李黎明，杨敬敏，侯铮，李淑元，胡斌，楚海燕，周玮晨，王传超，马彦云，朱晓，李蕾，徐珂

琳，王盼盼，刘杰，胡雅，丁琦亮，马滕，马超，赵振宏等等。肿瘤所：朱景德教授，罗晓莹老师，顾俊老师，王伟老师，孙晋枫师姐，马克龙师兄，周小羽师姐。德克萨斯大学健康科学中心：张福涛，陈民毅，郑琰，俞斌，简学求，李乐荣，蒋俊海，林楠，徐倩，于禁，于潇，张倩，李萧，任雪晗，李东阳，马龙等。还有医学中心的：周晓东教授，何东仪教授，潘兴教授，张怡教授，殷洪山医生，夏勇医生，崔瑞琴医生，杨金龙。和你们相处的日子，让我受益匪浅，愿我们的友谊地久天长。

此外，我还要由衷地感谢我的父母，谢谢您们在上海的陪伴以及一路默默的支持和关心。最后我想把我最真诚的祝福送给我的父母，我的三位指导老师，以及所有以上老师及师兄师弟师姐师妹们，祝你们身体健康，开心快乐。

## 已发表论文

### 第一作者及并列第一作者

1. Zhao, Y., **S. Guo**, J. Sun, Z. Huang, T. Zhu, H. Zhang, J. Gu, Y. He, W. Wang, and K. Ma, *Methylcap-seq reveals novel DNA methylation markers for the diagnosis and recurrence prediction of bladder cancer in a Chinese population.* PloS one, 2012. **7**(4): p. e35175. (IF=3.5)
2. **Guo, S.**, L. Tan, W. Pu, J. Wu, K. Xu, J. Wu, Q. Li, Y. Ma, J. Xu, and L. Jin, *Quantitative assessment of the diagnostic role of APC promoter methylation in non-small cell lung cancer.* Clinical epigenetics, 2014. **6**(1): p. 5. (IF=6.2)
3. **Guo, S.**, Y.L. Wang, Y. Li, L. Jin, M. Xiong, Q.H. Ji, and J. Wang, *Significant SNPs have limited prediction ability for thyroid cancer.* CANCER MEDICINE, 2014. **3**(3): p. 731-735.
4. Wang, Y.-L., S.-H. Feng, **S.-C. Guo**, W.-J. Wei, D.-S. Li, Y. Wang, X. Wang, Z.-Y. Wang, Y.-Y. Ma, and L. Jin, *Confirmation of papillary thyroid cancer susceptibility loci identified by genome-wide association studies of chromosomes 14q13, 9q22, 2q35 and 8p12 in a Chinese population.* Journal of medical genetics, 2013. **50**(10): p. 689-695. (IF=5.6)
5. **Shicheng Guo**, Fengyang Yan, Jibin Xu, Yang Bao, Ji Zhu, Xiaotian Wang, Junjie Wu, Weilin Pu, Yan Liu, Zhengwen Jiang, Momiao Xiong, Li Jin, Jiucun Wang, *Identification and validation of the methylation biomarkers of Non-small cell lung cancer (NSCLC).* Clinical Epigenetics (2014, Accepted, IF=6.2)

### 参与的文章

6. Xiang, H., J. Zhu, Q. Chen, F. Dai, X. Li, M. Li, H. Zhang, G. Zhang, D. Li, Y. Dong, L. Zhao, Y. Lin, D. Cheng, J. Yu, J. Sun, X. Zhou, K. Ma, Y. He, Y. Zhao, **S. Guo**, M. Ye, G. Guo, Y. Li, R. Li, X. Zhang, L. Ma, K. Kristiansen, Q. Guo, J. Jiang, S. Beck, Q. Xia, W. Wang, and J. Wang, *Single base-resolution methylome of the silkworm reveals a sparse epigenomic map.* Nature biotechnology, 2010. **28**(5): p. 516-20. (IF=39.08)
7. Li, Y., J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, Y. Huang, H. Cao, X. Zhou, **S. Guo**, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck, and

- X. Zhang, *The DNA methylome of human peripheral blood mononuclear cells*. PLoS biology, 2010. **8**(11): p. e1000533. (IF=11.77)
8. Wang, X., L. Wang, **S. Guo**, Y. Bao, Y. Ma, F. Yan, K. Xu, Z. Xu, L. Jin, and D. Lu, *Hypermethylation reduces expression of tumor-suppressor PLZF and regulates proliferation and apoptosis in non-small-cell lung cancers*. The FASEB Journal, 2013. **27**(10): p. 4194-4203. (IF=5.48)
9. He, Y., Y. Cui, W. Wang, J. Gu, **S. Guo**, K. Ma, and X. Luo, *Hypomethylation of the hsa-miR-191 locus causes high expression of hsa-mir-191 and promotes the epithelial-to-mesenchymal transition in hepatocellular carcinoma*. Neoplasia, 2011. **13**(9): p. 841-IN23. (IF=5.39)
10. Wu, L., **S. Guo**, D. Yang, Y. Ma, H. Ji, Y. Chen, J. Zhang, Y. Wang, L. Jin, and J. Wang, *Copy number variations of HLA-DRB5 is associated with systemic lupus erythematosus risk in Chinese Han population*. Acta biochimica et biophysica Sinica, 2014. **46**(2): p. 155-160. (IF=2.08)
11. SONG, X., **S. GUO**, Y. CHEN, C. YANG, H. JI, F. ZHANG, Z. JIANG, Y. MA, Y. LI, and L. JIN, *Association between HLA-DQA1 gene copy number polymorphisms and susceptibility to rheumatoid arthritis in Chinese Han population*. Journal of genetics, 2014. **93**(1): p. 215.(IF=1.01)
12. He, D., J. Wang, L. Yi, X. Guo, **S. Guo**, G. Guo, W. Tu, W. Wu, L. Yang, and R. Xiao, *Association of the HLA-DRB1 with Scleroderma in Chinese Population*. PloS one, 2014. **9**(9): p. e106939. (IF=3.5)
13. Lin, S., L. Pan, **S. Guo**, J. Wu, L. Jin, J.-C. Wang, and S. Wang, *Prognostic role of microRNA-181a/b in hematological malignancies: a meta-analysis*. PloS one, 2013. **8**(3): p. e59532. (IF=3.5)
14. Wu, J., J. Liu, Y. Zhou, J. Ying, H. Zou, **S. Guo**, L. Wang, N. Zhao, J. Hu, and D. Lu, *Predictive value of XRCC1 gene polymorphisms on platinum-based chemotherapy in advanced non-small cell lung cancer patients: a systematic review and meta-analysis*. Clinical Cancer Research, 2012. **18**(14): p. 3972-3981. (IF=8.19)
15. Wang, R., J. Zhang, Y. Ma, L. Chen, **S. Guo**, X. Zhang, Y. Ma, L. Wu, X. Pei, and S. Liu, *Association study of miR-149 rs2292832 and miR-608 rs4919510 and the risk of hepatocellular*

- carcinoma in a large-scale population.* Molecular medicine reports, 2014. (IF=1.48)
16. Wu, J., J. Wu, Y. Zhou, H. Zou, **S. Guo**, J. Liu, L. Lu, and H. Xu, *Quantitative assessment of the variation in IGF2BP2 gene and type 2 diabetes risk.* Acta Diabetologica, 2011: p. 1-11. (IF=3.67)
17. Xiao, Q., S. Gao, H. Luo, W. Fan, **S. Guo**, H. Yao, S. Leng, Z. Xu, T. Tao, and X. Liu, *9q33.3, a stress-related chromosome region, contributes to reducing lung squamous cell carcinoma risk.* Journal of Thoracic Oncology, 2014. **9**(7): p. 1041-1047. (IF=5.8)
18. Pan, L.-l., Y.-m. Huang, M. Wang, X.-e. Zhuang, D.-f. Luo, **S.-c. Guo**, Z.-s. Zhang, Q. Huang, S.-l. Lin, and S.-y. Wang, *Positional cloning and next-generation sequencing identified a TGM6 mutation in a large Chinese pedigree with acute myeloid leukaemia.* European Journal of Human Genetics, 2014. (IF=4.22)
19. Huang, L., Y. Li, **S. Guo**, Y. Sun, C. Zhang, Y. Bai, S. Li, F. Yang, M. Zhao, and B. Wang, *Different Hereditary Contribution of the CFH Gene Between Polypoidal Choroidal Vasculopathy and Age-Related Macular Degeneration in Chinese Han People.* Investigative ophthalmology & visual science, 2014. **55**(4): p. 2534-2538. (IF=3.66)
20. Zhao, Y., H. Zhou, K. Ma, J. Sun, X. Feng, J. Geng, J. Gu, W. Wang, H. Zhang, **S. Guo** and Y. He, *Abnormal methylation of seven genes and their associations with clinical characteristics in early stage non-small cell lung cancer.* Oncology letters, 2013. **5**(4): p. 1211-1218. (IF=0.98)
21. Wang, J., Y. Yang, **S. Guo**, Y. Chen, C. Yang, H. Ji, X. Song, F. Zhang, Z. Jiang, and Y. Ma, *Association between copy number variations of HLA-DQA1 and ankylosing spondylitis in the Chinese Han population.* Genes and immunity, 2013. (IF=3.78)

## 审稿中的论文

1. **Shicheng Guo**, Yuan Li, Yi Wang, Haiyan Chu, Yulin Chen, Gang Guo, Wenzhen Tu, Wenyu Wu, Hejian Zou, Li Yang, Rong Xiao, Yanyun Ma, Feng zhang, Li Jin, Xiaodong Zhou, Jiucun Wang. (2014) Copy number variation of HLA-DQA1 and APOBEC3A/3B contribute to the susceptibility of systemic sclerosis in Chinese Han population. Arthritis & Rheumatology (under review, AR-14-

- 0879)
2. Nan Li, Junhai Jiang, **Shicheng Guo**, Momiao Xiong. Functional Principal Component Analysis and Randomized Sparse Clustering Algorithm for Medical Image Analysis. *Medical image analysis* (under review)
  3. Junhai Jiang, Nan Lin, **Shicheng Guo**, Jinyun Chen, Momiao Xiong. (2014), Methods for Joint Imaging and RNA-seq Data Analysis. *Proceedings of the National Academy of Sciences* (under review, MS#2014-17699)
  4. Hao Xiong, **Shicheng Guo**, Saunak Sen, Xiangning Chen.(2014), Allele-specific differential expression of rna-seq data using bivariate functional principal analysis. *Scientific Reports* (under review, SREP-14-05394A)

## 会议摘要

1. **Guo S**, Xiong M, Jin L, Wang J. (2014). Epigenetic Approaches for non-small cell lung cancer diagnosis based on DNA methylation, HGV2014, Sep 17-Sep 19, Belfast, Northern Ireland, UK (Full Scholarship)
2. Lin N, Jiang J, Guan X, Yu X, Guo S and Xiong MM. (2014). A novel method for ultrasound image analysis. NCI-NIBIB Point of Care Technologies for Cancer Conference. January 8-10, 2014, Natcher Center, NIH campus-Building 45, Bethesda, Maryland.
3. Lin N, Jiang J, **Guo S**, Yu X, Ma L and Xiong MM. (2014). Classification Analysis of Big Image Data. Statistical and Computational Theory and Methodology for Big Data Analysis. Feb 9-Feb 14, 2014, Calgary, AB Canada.
4. Yu J, Lin N, Ma L, and Xiong MM. (2014). Cloud computing for joint big genetic, epigenetic and image data analysis. Keystone Symposia: Big Data in Biology, March 23-25, 2014. Fairmont San Francisco, San Francisco, California.

## 论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名：\_\_\_\_\_ 日期：\_\_\_\_\_

## 论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名：\_\_\_\_\_ 导师签名：\_\_\_\_\_ 日期：\_\_\_\_\_

