外周血肿瘤特异性 DNA 甲基化位点识别方法研究

（申请清华大学工程硕士专业学位论文）

培 养 单 位：　自动化系

工 程 领 域：　控制工程

申　请　人：　刘　奇

指 导 教 师：　刘　民　教　授

联合指导教师：　周　平　高级工程师

二〇二一年五月

Research on Methods for Identifying the Tumor-specific DNA Methylation Sites in Peripheral Blood

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Liu Qi

(Control Engineering)

Thesis Supervisor: Professor Liu Min

Associate Supervisor: Senior Engineer Zhou Ping

May, 2021

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

熊智华	副教授	清华大学
李建州	高级工程师	世纪恒通科技股份有限公司

答辩委员会名单

主席	李清	研究员	清华大学
委员	陆文凯	研究员	清华大学
	熊智华	副教授	清华大学
	胡坚明	副教授	清华大学
	李建州	高级工程师	世纪恒通科技股份有限公司
秘书	赵世敏	高级工程师	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）按照上级教育主管部门督导、抽查等要求，报送相应的学位论文。

本人保证遵守上述规定。

作者签名：______________ 导师签名：______________

日　　期：______________ 日　　期：______________

摘 要

外周血肿瘤特异性 DNA 甲基化位点识别是癌症液体活检领域中的重要研究方向和热点。本文针对癌症液体活检需求，开展了外周血肿瘤特异性 DNA 甲基化位点识别方法的研究。论文取得如下主要成果：

（1）整合了 TCGA、Xena、GEO 等肿瘤相关数据库中多类肿瘤组织来源和血液来源的 DNA 甲基化数据，构建了可用于肿瘤特异性 DNA 甲基化位点识别和肿瘤组织来源预测的较大规模 DNA 甲基化数据集。

（2）针对 DNA 甲基化数据集样本数远小于特征维度的特点，提出了一种基于类别特异性过滤的 DNA 甲基化位点识别方法，该方法通过对每两个类别之间进行差异甲基化分析，进而筛选出针对每类肿瘤具有特异性的 DNA 甲基化位点。

（3）针对类别特异性过滤方法随类别增多引起的肿瘤特异性 DNA 甲基化位点筛选难的问题，提出了一种基于统计显著性水平和互信息的位点特异性衡量方法，并采用有监督分类模型进行肿瘤组织来源预测性能评估，所提出的方法可利用较少的肿瘤特异性 DNA 甲基化位点信息获得较好的肿瘤组织来源预测性能。

在多个数据集上验证了本文所提出的外周血肿瘤特异性 DNA 甲基化位点识别方法的有效性。

关键词：血液；液体活检；DNA 甲基化位点；肿瘤特异性；机器学习

Abstract

The recognition of tumor-specific DNA methylation sites in peripheral blood is an important research direction and hotspot in the field of liquid biopsy for cancers. Facing the requirements of liquid biopsy for cancers, this thesis makes the researches on methods for identifying tumor-specific DNA methylation sites in peripheral blood. The main results of the thesis are as follows:

(1) This thesis integrates DNA methylation data of multiple types of tumors and blood from some tumor-related databases such as TCGA, Xena, GEO, etc., and constructs a large-scale DNA methylation dataset that can be used for identifying tumor-specific DNA methylation sites and predicting the tissue-of-origin.

(2) Considering the characteristic that the number of samples in the DNA methylation data set is much smaller than that of features, this thesis proposes a method based on category-specific filtering for identifying DNA methylation sites. Based on this method, tumor-specific DNA methylation sites for each type of tumor can be identified through differential methylation analysis between every two classes.

(3) Focusing on the problem that it is hard to screen enough tumor-specific DNA methylation sites with the increase of tumor categories, this thesis proposes a measurement method based on statistical significance and mutual information for measuring DNA methylation sites. And supervised classification models are used to evaluate the performance of the prediction of tumor tissue-of-origin. The proposed method can use less information on tumor-specific DNA methylation sites to obtain better prediction performance of tumor tissue-of-origin.

The effectiveness of the method for identifying tumor-specific DNA methylation sites in peripheral blood is validated on multiple data sets.

Keywords: Blood; Liquid Biopsy; DNA Methylation Sites; Tumor-specific; Machine Learning

目 录

插图和附表清单

符号和缩略语说明

LP	液体活检 (Liquid biopsy)
RNA	核糖核酸 (Ribonucleic acid)
DNA	脱氧核糖核酸 (Deoxyribonucleic acid)
TME	肿瘤微环境 (Tumor microenvironment)
NGS	二代测序 (Next generation sequencing)
scRNA-seq	单细胞 RNA 测序（Single-cell RNA sequencing）
cfDNA	循环游离 DNA (Circulating cell-free DNA)
ctDNA	循环肿瘤 DNA (Circulating tumor DNA)
cfRNA	循环游离 RNA (Circulating cell-free RNA)
EVs	循环细胞外囊泡 (Circulating extracellular vesicles)
TEPs	肿瘤相关血小板 (Tumor-educated platelets)
EpCAM	上皮细胞黏附分子 (Epithelial cell adhesion molecule)
TOO	组织来源 (Tissues-of-origin)
CH	克隆性造血 (Clonal hematopoiesis)
SNV	单核苷酸变异 (Single nucleotide variants)
CNV	拷贝数变异 (Copy number variants)
DNMTs	DNA 甲基转移酶 (DNA methyltransferase)
UTR	非翻译区 (Untranslated region)
DMP	差异甲基化位点 (Differential methylation point)
EDTA	乙二胺四乙酸 (Ethylenediaminetetraacetic acid)
CNAs	拷贝数异常 (Copy number aberrations)
NSCLC	非小细胞肺癌 (Non-small cell lung Cancer)
CGI	CpG 岛 (CpG Island)
WGS	全基因组测序 (Whole genome sequencing)
WGBS	全基因组 DNA 甲基化测序 (Whole genome bisulfite sequencing)
PBMC	外周血单个核细胞 (Peripheral blood mononuclear cell)
QC	质量控制 (Quality Control)
FDR	错误发现率 (False Discovery Rate)
TCGA	癌症基因图谱 (The cancer genome atlas)
NCI	美国国家癌症研究中心 (National cancer institute)
KNN	K-近邻 (K-nearest neighbors)

SVM	支持向量机 (Support vector machine)
RF	随机森林 (Random forest)
CART	分类回归树 (Classification and regression trees)
NB	朴素贝叶斯 (Naive bayes)
LDA	Fisher 线性判别分析 (Linear discriminant analysis)
ROC	受试者工作特征曲线 (Receiver operating characteristic curve)
AUC	ROC 曲线以下的面积 (Area under the ROC curve)

第 1 章　引言

1.1　研究背景和意义

癌症是当前全人类共同面临和应对的重大疾病之一。根据美国国家癌症研究中心（National Cancer Institute，NCI）的定义，“癌症是身体内某些细胞无法停止分化而进入到周围组织的一系列疾病的集合”[1]。癌症是一个复杂的系统，可以发生在身体中的几乎所有部位，其中很多会发展成实体瘤，即由细胞和组织构成的肿块，也有一部分癌症，例如白血病，不会形成实体瘤。“种子-土壤”假说[2]广泛为人所接受，即肿瘤的发生、进展、转移等是肿瘤自身与肿瘤微环境（Tumor microenvironment，TME）相互作用的结果。

据估计全球范围内每年新增患癌人数 1800 万，新增癌症致死人数 960 万。这其中若不区分性别，肺癌都是新增患病和致死人数最多的癌种，女性乳腺癌、前列腺癌、结直肠癌的发病率也较高。就死亡率而言，结直肠癌、胃癌和肝癌位于恶性程度和致死率前列。若区分性别，肺癌和女性乳腺癌分别是对男性和女性而言最常见的癌症类型[3]。不同国家之间，例如发达国家和发展中国家，癌症的新增和死亡统计方差很大，和当地的经济发展情况、环境污染、饮食习惯、医疗水平、居民定期检查和登记情况等诸多因素有关。就中国而言，据估计全国一年新增 393 万恶性肿瘤发生病例和 234 万恶性肿瘤死亡病例，其中最常见的癌症类型分别是肺癌、胃癌、结肠癌、肝癌和女性乳腺癌。最主要的致死性癌症类型是肺癌、肝癌、胃癌、食管癌和结直肠癌[4]。此外，过去的几十年间到现在，癌症发生率和死亡率在中国呈现出持续上升的趋势[5]。严峻的癌症防治形势对居民生命健康、生活质量、经济社会发展和公共卫生系统造成了巨大的负担，对癌症进行精准诊断和治疗具有研究价值和实际意义。

近年来随着高通量检测技术，如二代测序（next-generation sequencing，NGS）、微阵列（micro array）、单细胞 RNA 测序（Single-cell RNA sequencing，scRNA-seq），的快速发展和广泛应用，相较于传统的检测和分析方法，现在人们已经有能力高效地获取海量的多组学数据。根据分子生物学中的中心法则，多组学数据主要包括基因组学、表观组学、转录组学、蛋白质组学、代谢组学和表型组学等组学数据。“精准肿瘤学”的主要目标是提升对癌症的诊断和治疗，从分子层面对癌症进行分析，加深对癌症的认识具有重要意义，也在癌症的分子诊断、预后估计、分子分型、药物靶点发现、肿瘤个性化治疗等方面产生了广泛的应用[6]，癌症相关分子标志物的识别在其中扮演着重要的角色。

1.1.1　癌症液体活检

癌症液体活检（liquid biopsy，LP）是从体液，例如血液、尿液、胸腔积液、痰液等，当中检测癌症相关的生物标志物，并加以化验分析后获取和肿瘤相关信息的方法。影像学检查是对针对实体瘤进行筛查的一种常规且无创的手段，如一项长达数年的大规模研究[7]的结果表明，采用低剂量计算机断层扫描（low-dose CT，LDCT）进行大规模筛查能让人群的肺癌死亡率显著降低。但影像学检查方法除了费用昂贵之外，还有高假阳性率的问题，容易造成过度诊断和过度治疗[8]。组织活检是现有肿瘤诊断的“金标准”，通常需要开展手术或者借助穿刺手段，从患者疑似肿瘤的部位采集组织样本，然后在实验室中对采集到的生物样本进行病理学、免疫组化等分析。组织活检采样的过程具有很强的侵入性，会对潜在患者，尤其是体质较弱的人群而言，造成生理和心理上的痛苦。此外手术或穿刺过程还有造成肿瘤转移的风险。对于一些特定的部位，例如大脑和心血管，也不便于开展组织活检。采集的样本只能反映肿瘤的局部信息。同时也不便于频繁地开展，因而还存在时效性不强、无法动态监测肿瘤进展的局限性。

癌症液体活检相较于传统癌症检测方法拥有诸多优势。基于外周血的癌症液体活检，除了操作简单、取样方便的优势之外，最主要优点在于其“无创性”与“实时性”。具体而言其几乎没有侵入性，即使对于某些不适合组织采样的部位，也能够在不同的时间点频繁开展。由于不同来源的肿瘤生物标志物都有可能汇集到检测样品中，所以一定程度上液体活检也减小了肿瘤异质性造成的影响。基于上述优势，液体活检有望在某些特定情况下成为组织活检的替代或者补充方案，在肿瘤早期诊断、进展监测、用药指导、异质性研究等方面有很多潜在的研究和应用价值。如图 1.1所示是以“液体活检”为关键词，在 PubMed①医学文献库和 Web Of Science②上检索到的近 10 年以来的相关研究随时间的变化情况。作为一项新兴的尤其是用于癌症检测的技术，癌症液体活检已得到越来越多的关注。

1.1.2　癌症液体活检生物标志物

液体活检的关键在于检测液体中的肿瘤相关生物标志物（也称为分析物），基于外周血的癌症液体活检中典型的生物标志物如图 1.2所示。外周血经过分离可以得到两部分：其一是外周血单个核细胞（peripheral blood mononuclear cell，PBMC）组分，从中可以富集分离得到循环肿瘤细胞（circulating tumor cells，CTCs）；其二是血清或血浆，可以用来检测循环游离 DNA（circulating cell-free DNA，cfDNA）、循环细胞外囊泡（circulating extracellular vesicles，EVs，其中包含外泌体 Exosomes[9]）、

① https://pubmed.ncbi.nlm.nih.gov/
② http://webofknowledge.com/

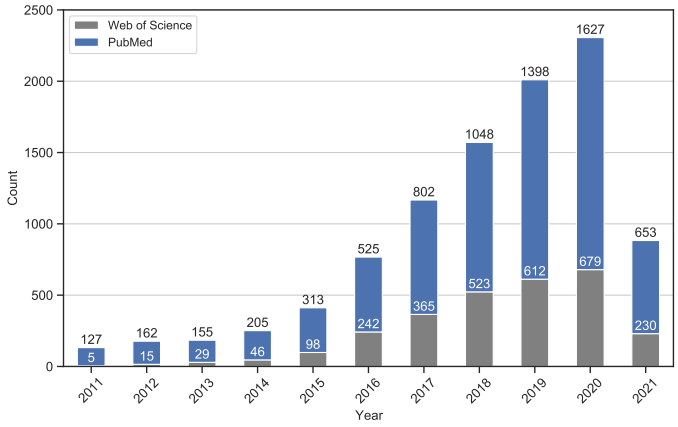


图 1.1　近 10 年来 PubMed 和 Web of Science 中液体活检相关研究数量变化（截至 2021 年 5 月 1 日）

循环游离 RNA（circulating cell-free RNA，cfRNA）、肿瘤相关血小板（tumor-educated platelets，TEPs）。在诸多癌症液体活检生物标志物分析中，又以对 CTC、cfDNA 和外泌体的检测和分析较为常见，而最终的量化在分子生物学层面均落到 DNA、RNA 和蛋白质上。

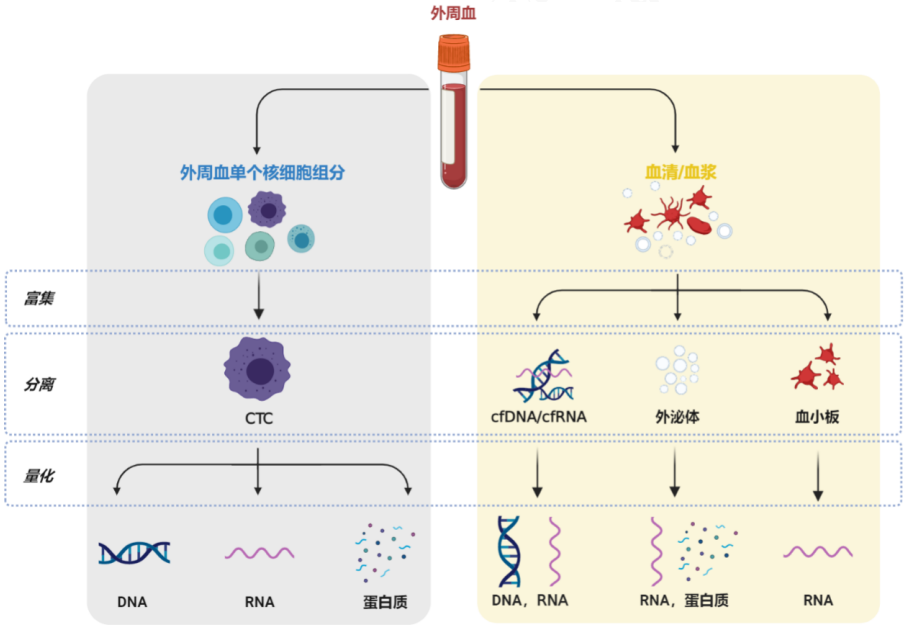


图 1.2　外周血中癌症液体活检常见生物标志物

1.1.2.1　CTC

CTC 是血液或者淋巴液中的肿瘤细胞，来源是原发或者转移部位的肿瘤组织。CTC 可能与肿瘤转移有关，对研究癌细胞免疫逃逸机制、癌细胞与免疫系统的相互作用等也具有重要意义。CTC 的分离和识别主要分为两类，依赖标记的（label-

dependent）和无需标记的（label-independent）。前者需要借助一些生物特性，例如检测上皮细胞黏附分子 (epithelial cell adhesion molecule，EpCAM)、CD45（common leukocyte antigen 45）来确定是否是 CTC；后者则通常借助一些物理特性，例如细胞尺寸、密度、可塑性等来分离提纯 CTC；也可以多种方法进行结合来检测和分离 CTC。如图 1.3所示，是通过 CTC 芯片技术从一名乳腺癌患者血液样本中捕获得到的有代表性的单个 CTC 和 CTC cluster（即聚集成团的 CTC）图像，用四种不同的染色剂进行染色观察到的 CTC 图像。

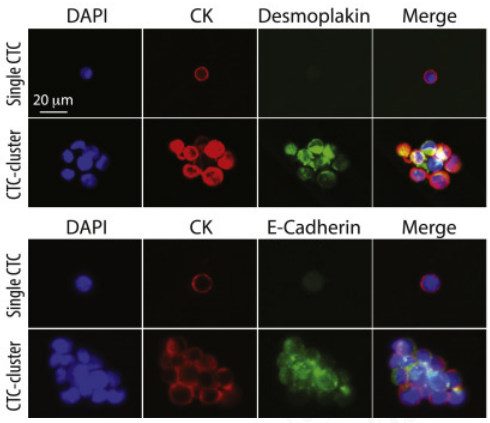


图 1.3　单个 CTC 和 CTC cluster 示意图[10]

1.1.2.2　外泌体

外泌体是由活细胞分泌的一系列大小在 30-150 纳米之间拥有膜结构的囊泡，其中包含 microRNA、mRNA、DNA 片段和蛋白质。其分泌来源包括肿瘤细胞、免疫细胞（immune cells）、间叶细胞（mesenchymal cells）在内的多种细胞[9]。癌源性外泌体可以改变邻近和远端组织细胞侵袭性，进而影响肿瘤微环境[11]。外泌体由于被双层脂质膜结构包裹，相较于其他液体活检分析物具有较好的稳定性[12]，并且其中包含核酸、蛋白质等物质，因而蕴含丰富的检测信息。外泌体作为肿瘤相关生物标志物面临的主要挑战是目前缺乏可靠和标准化的分离提纯方法。

1.1.2.3　cfDNA

cfDNA 是游离在血液中的单链或者双链 DNA 小片段，是液体活检检测的主要分析物之一，其中循环肿瘤 DNA（circulating tumor DNA，ctDNA）是 cfDNA 中来源于肿瘤的部分。ctDNA 进入血液的生物学机制目前尚无定论，关于 ctDNA 的来源，通常认为其来源于坏死或者凋亡的肿瘤细胞，另外也有可能来源于活体肿瘤细胞和 CTC[13]。根据基因组转移假设，ctDNA 可能与正常细胞转化为肿瘤细胞以及肿瘤的远端转移有关。

基于 cfDNA 的癌症液体活检，在进行分析之前需要进行 cfDNA 的富集、分离和提取。如图 1.4所示是一种典型的 cfDNA 提取流程示意图，通常需要从符合入组条件的检测对象肘部静脉采集约 10 毫升全血，转移到加入抗凝剂 EDTA（Ethylene-diaminetetraacetic acid，乙二胺四乙酸）的采样管（或者 EDTA 抗凝管）中，随后对样品进行分离得到血浆。通常血浆中的 cfDNA 含量相较于血清更高，但这往往是淋巴细胞中的 DNA 导致的，因此大多数现有的研究都是采用血浆作为分离 ctDNA 的材料。分离后的血浆可以立即或者保存于-80℃ 冰箱中直至使用核酸提取试剂盒或其他方法来提取 cfDNA。

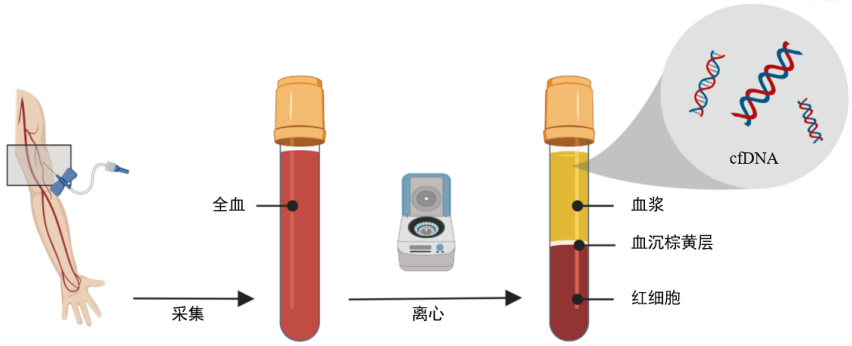


图 1.4　血液采集流程和从血浆中分离提取 cfDNA 示意图

cfDNA 中蕴含丰富的检测信息，如图 1.5所示，ctDNA 上能够检测到基因组畸变，包含基因突变（mutations）、染色体重排（chromosomal rearrangements）、拷贝数异常（copy number aberrations，CNAs）。还可以检测到其他特征，例如 DNA 片段长度、表观遗传学特征例如 DNA 甲基化（DNA methylation）等[14]。

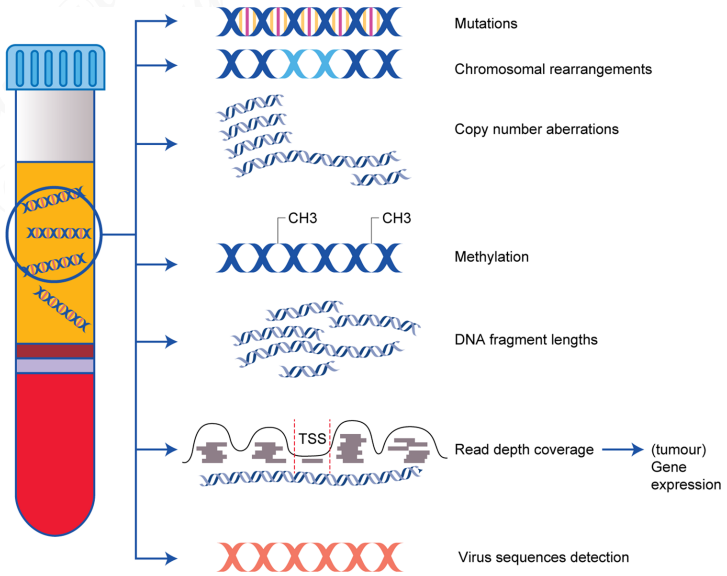


图 1.5　ctDNA 的不同检测特征[14]

1.2　国内外研究现状

现有的基于外周血的癌症液体活检主要应用于癌症的分子诊断中，其中也会用到机器学习方法。机器学习属于人工智能的范畴，是指不通过显式编程而是借助算法和统计模型从历史数据中学习并提升解决特定任务（有监督或者无监督）的一类方法。现对基于外周血的癌症液体活检用于癌症的分子诊断的相关研究进行综述，重点关注基于表观遗传学特别是 DNA 甲基化标志物识别中的统计和机器学习方法及其应用。

1.2.1　基于基因组学的癌症分子诊断

Fernandez-Cuesta L 等人的研究[15]指出可以在早期小细胞肺癌（NSCLC，Non-small cell lung Cancer）患者的 cfDNA 中检测到 TP53 基因突变，然而没有证实检测到的 TP53 基因突变就是来源于小细胞肺癌细胞。Phallen J 等人的研究[16]设计了一种称为 TEC-Seq 的方法，检测 58 个癌症相关基因，对 44 名健康人的样本分析发现其中有 16% 的 cfDNA 基因组改变来自克隆性造血（clonal hematopoiesis，CH）。对两百名结直肠癌、乳腺癌、肺癌和卵巢癌症患者的血浆样本分析发现 I 期和 II 期患者的血浆分别有 71%、59%、59% 和 68% 存在体细胞突变和患者的肿瘤信息高度一致。

约翰霍普金斯大学 Cohen 等人发表在《Science》上的研究提出了一种称为 CancerSEEK 的方法[17]，对 1005 名无转移临床诊断为卵巢癌、肝癌、食道癌、胰腺癌、胃癌、结肠癌、肺癌、乳腺癌患者血液中的循环蛋白质和 cfDNA 进行评定，并进行肿瘤检测（二分类）和组织定位（多分类）。对于癌症检测，CancerSEEK 通过优化方法从 39 种蛋白质中筛选关键特征，首先根据 Mann-Whitney-Wilcoxon test 剔除掉在正常对照组中蛋白质浓度中位数高于患癌组的蛋白质，排除了 39 种蛋白质中的 13 种，剩余 26 种。随后从 26 种蛋白质中依次剔除某种蛋白，进行逻辑回归，根据 Accuracy 降低的程度从而判断对应蛋白质特征的重要程度，从而进一步降低特征的维度，最终筛选得到 8 种类型蛋白质。共评估了 1933 种不同位置的突变（包含单碱基替换、插入、缺失），并通过线性模型计算得到 Ω 值。诊断模型采用的是逻辑回归，特征维计算得到的 8 种蛋白质浓度和 Ω 值。最终 CancerSEEK 检测特异度超过 99%，灵敏度中位数为 70%，按照癌症类型灵敏度从 69% 到 98% 不等。对于组织定位，CancerSEEK 使用相同的 9 种特征（Omega 值加上 8 种蛋白质浓度）加上患者的性别，和另外 31 种蛋白（即总共 41 维特征组成的特征向量），采用随机森林，10 折交叉验证，在通过逻辑回归在分类正确的样本上进行分类，灵敏度中位数达到 83%。CancerSEEK 对于部分癌症灵敏度依然比较低，例如肺癌而

言 top prediction 为 39%。对于早期癌症的诊断结果也较低。

斯坦福大学 Chabon 等人 2020 年发表在的《Nature》的研究[18]使用改进的 CAPP-Seq 技术来检测和分析 cfDNA，提出了一种称为 LungCLiP（lung cancer likelihood in plasma）的机器学习方法。LungCLiP 方法主要包括三个部分，一个 SNV（single nucleotide variants）模型、一个 CNV（Copy number variants）模型和最后的集成分类模型。其中 SNV 模型用来对实验组和对照组中的突变特征进行区分，采用弹性网络逻辑回归模型来识别 SNV 是否是肿瘤来源，对没有肿瘤组织样本的患者的 SNV 进行标记和打分，随后再训练又一个弹性网络逻辑回归模型从而对实验组和对照组进行分类。CNV 模型接收来对基因组区域处理得到的过滤后 CNV count、GISTIC CNV count 特征，和 Fisher's Exact Test 得到的 ***P***-value 特征，使用广义线性模型（generalized linear model）给予每个样本一个得分。SNV 模型和 CNV 模型输出的得分输入到最后的集成分类器中进行训练，分类器由五种机器学习方法（5-近邻、3-近邻、朴素贝叶斯、逻辑回归和决策树模型）构成。最终 LungCLiP 模型整合预测结果得到一个综合打分作为最后的预测指标。对于 I、II、III 期 NSCLC 的预测 AUC（Area under the ROC Curve）指标分别为 0.69（N=32）、0.71（N=9）、0.98（N=5）。LungCLiP 一定程度上验证了基于基因组进行 NSCLC 早期诊断的可行性，但参与的样本较少，且性能提升空间还很大。

目前基于癌症相关基因突变进行诊断的方法主要问题在于，存在某些 cfDNA 突变和一些非癌症因素，相关例如年龄相关的克隆性造血，会影响了检测的特异性。对于任意一种癌症很少存在有和其病理特征对应的突变，因此通过 cfDNA 的突变来确定肿瘤组织来源（tissue-of-origin，TOO）是困难的[19]。结合其余分析物的分析结果相互相验证是较好的解决方案，这其中基于表观遗传学特征尤其是 DNA 甲基化的研究取得了较多的进展。

1.2.2　基于表观遗传学的癌症分子诊断

1.2.2.1　DNA 甲基化及其生物性质

DNA 甲基化是目前研究得最为充分的表观遗传学改变，如图 1.6所示，DNA 甲基化是在指 DNA 分子中加入甲基的过程，在真核生物中通常表现为在 DNA 甲基转移酶（DNA methyltransferase，DNMTs）的催化下，DNA 的胞嘧啶核苷酸（Cytosine，缩写为 C）第 5 位碳原子所连接的氢原子，为来自 S-腺苷甲硫氨酸（SAM）的甲基所替代，转化成为 5-甲基胞嘧啶（5-Methylcytosine，缩写为 5-mC）的过程。

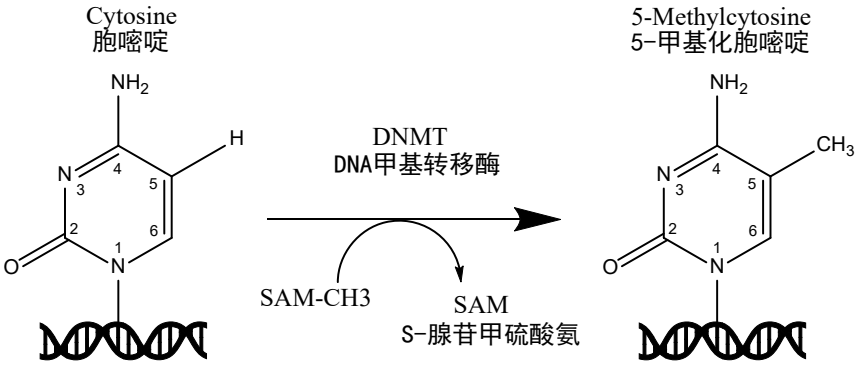


图 1.6　DNA 甲基化原理示意图

DNA 甲基化常发生于被称为 CpG 岛（CpG Island，CGI）即 DNA 上一段 CG 二核苷酸出现频率较高的区域。CpG 岛通常被定义为符合以下三个条件的区域，（1）长度大于 200bp；（2）G＋C 含量大于 50%；（3）观察到的 CpG 比例大于 0.6。根据 CpG 距离 CpG 岛的距离远近，还可以划分为 CpG Shore、CpG shelf、CpG opensea，通常分别距离 CpG 岛 2 千个碱基，2-4 千个碱基和之外的区域，如图 1.7所示。

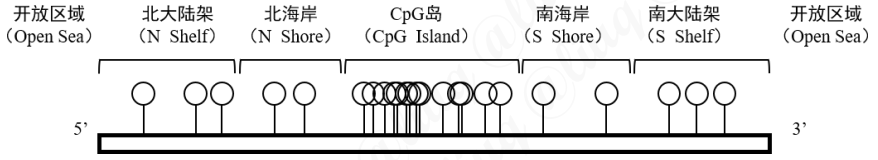


图 1.7　CpG 岛和周围的区域

根据 CpG 位点在基因组中的区域，通常关注其是否位于 TSS（transcription start sites，转录起始位点）、TSS200（距离 TSS 上游 0-200 个碱基的区域）、TSS1500（位于 TSS 上游 200-1500 个碱基的区域）、5′UTR（5′untranslated region，即 5′ 非翻译区）、1st exon（第一个外显子区域）、Gene body（基因区域）、3′UTR（3′ 非翻译区域）等位置，如图 1.8所示。

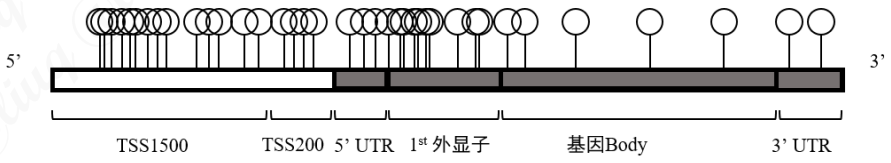


图 1.8　CpG 位于基因组区间的不同区域示意图

DNA 甲基化可以在不改变 DNA 序列原始结构的前提下改变 DNA 片段的活性，在肿瘤的发生和进展中扮演着重要的角色[20]，例如癌细胞的 DNA 甲基化发生在抑癌基因的启动子区域时该 DNA 片段的活性被改变。癌细胞通常表现有异常的 DNA 甲基化模式。DNA 甲基化是组织特异性的和癌症类型特异性的[21-22]，当 cfDNA 被死亡的细胞释放出来时，经过甲基化修饰的胞嘧啶残基不会被清除，也

因此可以用来确定 cfDNA 的肿瘤组织来源[23]。

1.2.2.2　DNA 甲基化相关研究现状

近年来，基于 cfDNA 甲基化的分子诊断正发挥着越来越重要的作用。2015 年 Yuk Ming Dennis Lo 等人发表在《PNAS》的研究[24]对血浆 cfDNA 采用全基因组甲基化测序（genome-wide bisulfite sequencing）结合生物信息解卷积处理过程得到每个样本来源于几种主要组织的相对分数，并在孕妇、癌症患者和器官移植者中进行了验证，预示着 cfDNA 在产前诊断、肿瘤检测和器官移植方面的潜力。

2017 年加州大学洛杉矶分校的 Shuli Kang 等人提出了一种基于概率模型的 CancerLocator 机器学习算法[25]。CancerLocator 方法使用全基因组 DNA 甲基化数据，能够同时进行 ctDNA 在血浆中的比例和肿瘤组织来源预测。该方法整体分为两步，第一步是特征构建和过滤，研究从 TCGA 数据库中选择了五种器官（乳腺、结肠、肾、肝和肺）对应的七种癌症 DNA 甲基化数据，从其他研究的公开数据中构建了来自健康人的 cfDNA 甲基化数据集，并定义了一系列的 CpG 位点组成的 CpG 簇（CpG cluster）将其平均甲基化水平作为特征，进而通过设定阈值进行特征选择。在每个 CpG 簇特征上，对不同类型的癌症组织和健康人血浆血浆来源样本的 DNA 甲基化水平，分别建模为不同的 beta 分布。第二步，使用选择的 CpG 簇特征及其对应 beta 分布，将病人的 cfDNA 解卷积（deconvolute）为正常血浆 cfDNA 甲基化的分布和实体瘤的 DNA 甲基化的分布，来估计 cfDNA 甲基化特征来自肿瘤的成分占比用以进行肿瘤诊断和组织定位。在仿真和真实数据集上与支持向量机（support vector machine, SVM）和随机森林（random forest，RF）的比较结果表明 CancerLocator 显示出更好的分类性能。

Farshad Nassiri 等人的发表在《Nature Medicine》研究[26]通过 cfMeDIP-seq 的方法来检测 cfDNA 中的甲基化信号，为了确定能否将神经胶质瘤与颅内肿瘤和健康对照区分开来，在 9 种不同病理的 447 个 cfMeDIP-seq 样本上，每次按照 80% 比 20% 随机划分成 50 个训练和测试集，每个训练集上使用 moderated t-test 获取前 300 个差异甲基化区域作为特征，随后训练 50 个随机森林分类器，使用十折交叉验证来优化模型；另外在 5 种不同病理的 161 个颅内肿瘤 cfMeDIP-seq 样本上采用类似的特征预处理方法，训练一系列的广义线性模型，实现了对于常见原发性颅内肿瘤的检测和区分。

2020 年名为 CCGA（The Circulating Cell-free Genome Atlas）的大型研究计划[27]得到初步结果。该计划旨在确定全基因组 cfDNA 测序与机器学习结合是否能够在较高的特异性条件下对多种类型的癌症进行检测，从而为将液体活检推广到人群癌症筛查提供参考。现阶段研究构建了 6689 名参与者的队列，包含 2482 名

超过 50 种类型癌症的患者和 4207 名健康人，cfDNA 经过靶向超过 10000 个甲基化区域的亚硫酸氢盐测序后，输入到机器学习模型中，在测试集上诊断特异度达到了 99.3%。其中 I-III 期预测灵敏度在 12 种癌症（肛门癌、膀胱癌、结肠直肠、食道、头颈癌、胆管癌、肺癌、淋巴癌、卵巢癌、胰腺癌、骨髓癌、胃癌）达到了 67.3%，在全部癌症上 I-III 期预测灵敏度达到了 43.9%。全部癌症上的灵敏度为 18%，其中 I 期为 18%，II 期为 43%，III 期为 81%，IV 期达到了 93%，整体肿瘤组织来源预测正确率达到了 93%。这一研究有望证明基于 DNA 甲基化的液体活检进行肿瘤普查将成为一种可能，然而其检测的甲基化特征、数据和机器学习算法模型均未公开，缺乏足够的透明度来进行结果的复现和检验。

中山大学的徐瑞华等人 2017 年发表在《Nature Materials》的研究[28]首先将 TCGA 肝细胞癌（hepatocellular carcinoma，HCC）组织 DNA 甲基化与健康人血液白细胞 cfDNA 甲基化数据进行比较以确定 HCC 特异性标志物，随后构建了一个由 1098 名 HCC 患者和 835 名健康人组成的研究队列，进而产生 cfDNA 甲基化数据集，使用 Lasso（least absolute shrinkage and selection operator）和随机森林算法进行标志物的选择，随后输入到逻辑回归模型中将得到的估计系数和 DNA 甲基化特征值相乘计算诊断得分，在测试集上取得了 AUC 0.944 的结果，取得了较好的预测诊断效果。与之类似，Huiyan Luo 等人 2020 年发表在《Science Translational Medicine》的研究[29]通过比较结直肠癌组织和健康人血液白细胞来确定甲基化特异性特征，构建了一个包含 801 个结肠癌病人和 1021 个健康人 cfDNA 研究队列，并使用 Lasso 和随机森林算法进行标志物的选择，随后输入到多项式逻辑回归模型得到得分来进行诊断，新方法显示出相较于传统的借助癌胚抗原（carcinoembryonic antigen，CEA）进行诊断更优的性能（AUC 0.96 超过借助 CEA 的 0.67）。

2019 年广州医科大学的梁文华等人提出一种基于 ctDNA 甲基化的方法[30]，用来诊断早期肺癌的同时也可以区分结节输入良性或者恶性，研究首先构建了有 79 个含有恶性结节和 53 个良性结节的队列，按照一比一划分训练和测试集，通过 Wilcoxon rank sun test 和 Lasso 筛选高甲基化标志物，最后使用逻辑回归模型进行诊断，在测试集上取得了 79.5% 的灵敏度和 85.2% 的特异度。

2020 年复旦大学的陈兴栋等人发表在《Nature Communications》的研究[31]设计了名为 PanSeer 的 cfDNA 甲基化的检测和诊断方法。具体而言 PanSeer 方法设计了 595 个靶向甲基化区域，随后使用 Benjamini-Hochberg 多重假设检验矫正的 t-test 来比较从 BioChain 购买得到的肿瘤组织和正常组织样本对应区域的平均甲基化分数（average methylation fraction，AMF），从中挑选在一种癌症或者多种癌症组织中具有显著差异的 DNA 甲基化区域作为标志物，并在 TCGA 数据库上验

证这些挑选出的 DNA 甲基化区域。随后构建逻辑回归模型进行分类。对于五类癌症（胃癌、食管癌、结直肠癌、肺癌和肝癌）患者，实现了在诊断后采集血样的患者中特异度为 95% 下 88% 的诊断准确率，另外有 95% 采样时无症状而随后四年内诊断为癌症的患者被 PanSeer 正确预测。

1.2.3　DNA 甲基化标志物识别

标志物选择（marker selection）是组学数据中的重要步骤。对于 DNA 甲基化数据而言，根据 CpG 位点位于基因组上的位置，以及甲基化的 CpG 位点集中的区域长度，这些标志物可以是 CpG 位点、CpG 簇[25]、CpG 单元[32]、CpG 区域[31]等。

比较不同组（即类别）之间平均值的差值是最简单同时也是最常见的差异甲基化标志物选择方法，对该结果取对数则类似于基因表达中的差异表达倍数（log fold change，logFC）的计算[33]。

Sun 等人 2015 年发表在《PNAS》上的研究[24]为了确定血浆 DNA 到组织映射的甲基化标志物，构建了来自 14 种人类组织（肝脏、肺、食道、心脏、胰腺、结肠、小肠、脂肪组织、肾上腺、大脑、T 细胞 B 细胞、中性粒细胞和胎盘) 的全基因组亚硫酸氢盐测序数据集。定义了两类甲基化标志物：第一类标志物表示某类组织特异性标志物，其在某类组织中不同于其他所有组织。具体过程是，排除胎盘，在其他 13 类组织中，若某类组织在一个基因座的甲基化密度位于全部 13 类组织的基因座平均甲基化密度均值的 3 倍标准差之外，则判定为是；第二类标志物表示在所有组织之中表现出甲基化多样性的标志物，其需要同时满足两个条件：（1）在该基因座上，最高甲基化的组织类别的甲基化密度至少比表现为最低甲基化的组织的甲基化密度高 20%;（2）每类甲基化密度除以该类的平均甲基化密度（即该类的变异系数），得到 13 种组织变异系数的标准差大于等于 0.25。随后考虑胎盘组织，选择其中的甲基化基因座标志物，要求满足该基因座上胎盘的甲基化密度位于 13 种组织的平均甲基化密度的三倍标准差之外，从而完成甲基化标志物的筛选。

Shuli Kang 等人的研究[25]构建了多种肿瘤组织和健康人血液的 DNA 甲基化 CpG 簇数据集，对于每一个 CpG 簇（即一个甲基化标志物特征）计算每种类别（即每种肿瘤和健康的血浆）的所有样本对应的平均甲基化水平，将该 CpG 簇上平均值集合的取值范围（即均值集合中的最大值与最小值的差值）作为该标志物的甲基化范围（Methylation Range，MR），MR 越高的 CpG 簇认为具有更高的差异，随后设定阈值从而进行 CpG 簇特征的筛选。Cho 等人的研究[34]从 TCGA 上获取了 345 个 CRC 肿瘤组织和配对的 38 个癌旁组织的 DNA 甲基化样本，从 GEO（GSE40279）获取 656 个外周血白细胞样本的 DNA 甲基化数据，数据均采集自

Infinium HumanMethylation450 BeadChip 平台。将 CRC 组织甲基化 β 值减去正常组织的平均 β 值，得到 $\Delta\beta$ 值。设定 $\Delta\beta$ 的绝对值需大于 0.4，筛选得到 1180 个 probe，被设为探针集合 A。类似的，将 CRC 组织甲基化的 β 值减去 GEO 血液数据集作为正常组织的甲基化 *beta* 值得到 $\Delta\beta$，采用和前者相同的方法，得到 1160 个 probe，被设为探针集合 B。随后取探针集合 A 和 B 的交集，并从中选择显著性水平 p<0.05 的前 200 个用于后续的分析。由于只考虑了均值，而忽视了 β 值在不同类别之间的异方差性，导致可能选择到整体水平很低或者很高的甲基化标志物[35]，为了避免此类问题，可以在采用其他差异分析方法后将其作为一种可选的过滤条件来使用。

基于统计的差异标志物分析方法也被用于 DNA 甲基化标志物的识别。例如 t-test（student's t-test，学生 t 检验）、秩和检验（rank-sum test，也称为 Wilcoxon test 或者 Man-Whitney U test）、方差分析（analysis of variance，ANOVA）、Fisher 精确检验（Fisher exact test）等统计方法都可以用于进行差异分析。对于 DNA 甲基化数据，虽然 t-test 的方法要求进行比较类别的特征有正态分布的假设，然而许多研究已经证明即使不遵循正态分布这一前提下，采用构建在基于正则项 t-test 的经验贝叶斯的方法例如 Limma（linear models for microarray data）[36] 也能取得很好的结果。采用例如 t-test 或者 moderated t-test 的方法对差异化的甲基化标志物进行检验后，通常按照结果显著性水平（或 p 值）进行筛选。Xu 等人的研究[28] 为了选择在肝细胞癌（hepatocellular carcinoma，HCC）组织 DNA 和健康人血液来源的 cfDNA 中有差异的甲基化标志位点，使用基于经验贝叶斯方法的 moderated t-statistics 检验[37]，经过 Benjamini-Hochberg 过程矫正错误发现率，得到矫正后的 p 值小于 0.05 的标志，从小到大排序选取其中的前 1000 个标志。后续使用无监督层次聚类方法，在该 1000 个标志对应的 DNA 甲基化特征数据上的聚类结果显示 HCC 组织和健康人血液样本被正确地区分开来。Luo 等人比较来自 TCGA CRC 组织 DNA 与来自健康人血液 DNA（GSE40279）的甲基化数据采用和前述相同的方法，得到矫正后的 p 值 <0.05 从小到大排序后取前 1000 个标志用于后续的分析[38]。

对 DNA 甲基化芯片数据进行差异化分析后，DNA 甲基化标志物还比较很多，通常也会结合机器学习等方法进一步特征降维。为了筛选出组织特异性的标志物，一种新型两层特征选择方法[39] 被提出，第一层特征选择对 miRNA 和 DNA 甲基化数据的类似，分为三个步骤：（1）对所有的正常组织，筛选出某类正常组织相对于其他类正常组织的差异表达/甲基化的 miRNAs/CpGs;（2）对同时存在癌旁和癌组织的组织类型，筛选出癌旁和癌组织中没有差异的 miRNAs/CpGs；（3）对所有的癌组织，筛选出某类癌组织相对于其他所有癌组织的差异表达/甲基化的 miRNAs/CpGs。

其中差异分析均采用 Limma 来实现。通过这三个步骤，得到的 miRNAs/CpGs 集合的交集，作为筛选出的肿瘤特异性标志物。第二层特征选择只对甲基化数据展开，采用两种方法，第一种称为 MRMD（Maximum-Relevance-Maximum-Distance）[40]，其使用皮尔逊相关性系数来衡量特征与标记的相关性，相关性越高值越大；用欧氏距离来衡量特征的冗余性，距离越大，表示冗余性越小；相关性和距离之和最大的特征被排序和作为候选特征。随后使用简单的分类器来计算所选特征的分类性能，选择性能较优的作为最终的特征。同样作为第二层特征选择方法的还有主成分分析，参数设置为 0.95，表示保留的主成分占累计总方差的 95%。这种方法在一对其列别来筛选特异性标志物的时候，实际上不对其他类别加以区分，得到的可能是对其他所有类别总体存在特异性的标志物，有可能只是对其中某一类或者几类存在特异性，而不是对其中所有类别的组织都存在特异性。另外，对于第二层甲基化数据的特征筛选，采用欧氏距离来衡量特征之间的冗余性，无法很好地衡量特征之间的差异性，存在欧氏距离很小而特征之间差异较大的情况。

1.2.4　当前研究存在的挑战

根据以上国内外研究现状可知，虽然在以 cfDNA 为生物标志物进行癌症分子诊断领域发展迅速并已经取得了许多有价值的研究成果，尤其以基于 DNA 甲基化的表观遗传学特征进行癌症分子诊断显示出巨大的前景，但依然存在着以下挑战：

（1）缺少统一的标准和规范。采用的平台和方法不统一，此外由于受到实验成本、周期等方面的限制，通常的样本规模都比较小，多则几百少则几个，没有统一的规范、流程和标准；有一部分研究设计了较为大型的研究队列，但是由于商业或者研究敏感性，不进行数据和方法公开，因而难以进行实验复现来验证方法；且通常任务较为单一，多为针对于某种特定的癌症专门进行设计，缺乏足够多的针对多类癌症的数据。

（2）数据“高维、小样本”的特点。高通量的检测方法和大 panel 的设计，导致获取到的 cfDNA 相关数据呈现典型的“$N \ll P$”（即特征数远多余样本数的 short，fat data problem）特点，对于 DNA 甲基化而言无论是采用 DNA 450K/EPIC microarray 还是全基因组甲基化测序（whole genome bisulfite sequencing，WGBS）将会得到大量的 CpG 位点的甲基化信息，其中包含着大量冗余、无关的特征乃至噪音，因而需要降低特征维度剔除掉产生干扰的噪声，留下关键的特征进行分析和其他后续处理；

（3）标志物识别和生物学意义。往往凭借先验知识和“湿实验”来手工挑选，缺乏对于标志物重要性的评估来而中挑选最优的特征来减少需要验证的关键标志物的个数；诊断的准确性特别是对于早期癌症的诊断精确度普遍还有待提升，且

应从多维度进行综合评判。

直接将含有大量的冗余乃至噪音的 DNA 甲基化数据作为机器学习模型的输入，对于模型训练和最终进行预测是不利的[41]。识别和筛选出肿瘤特异性 DNA 甲基化标志物、剔除无关变量并用于构建预测肿瘤组织来源的机器学习模型是本文的主要研究内容。

1.3　本文主要内容和结构

本文根据当前基于外周血的癌症液体活检中的实际需求，开展了外周血肿瘤特异性 DNA 甲基化位点识别方法的研究，主要工作如图 1.9所示。

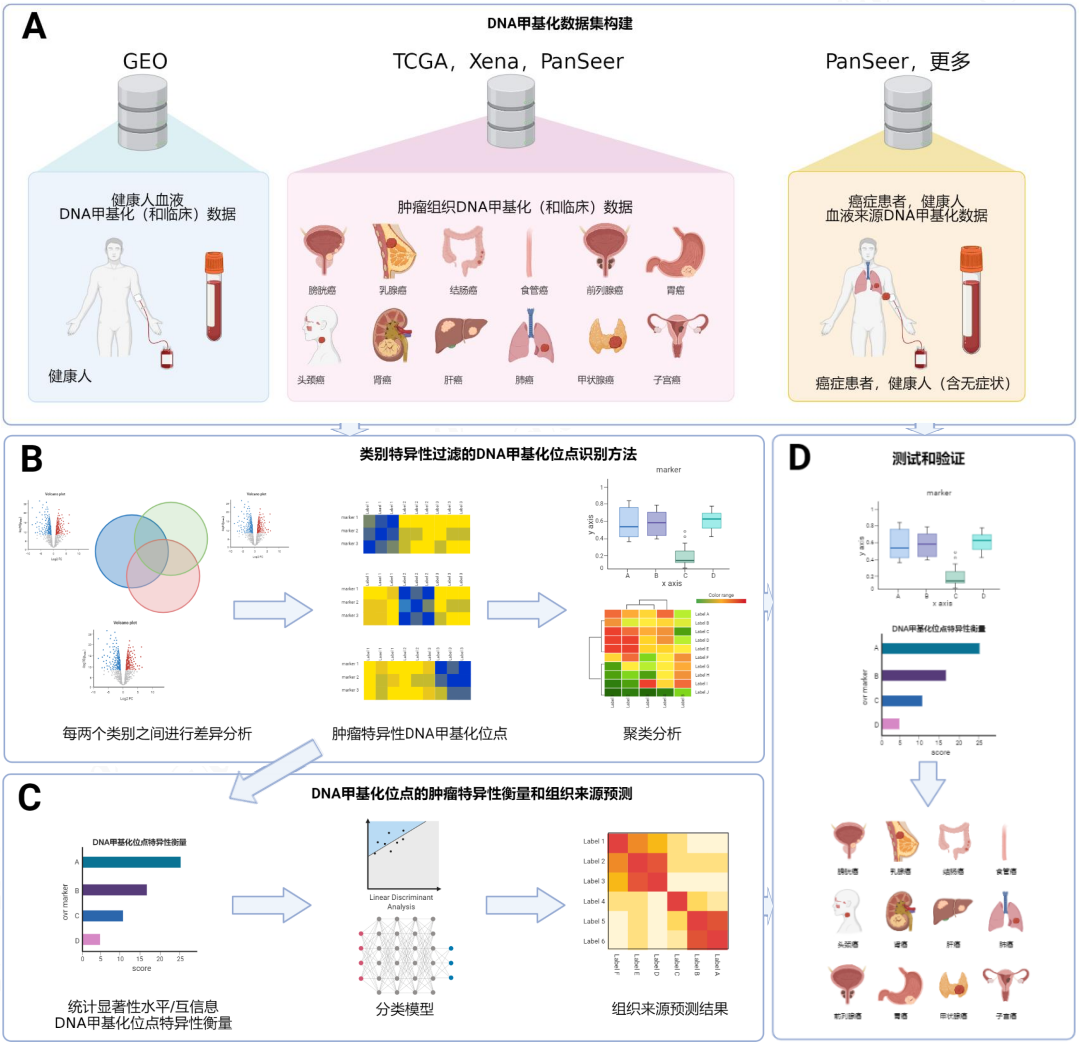


图 1.9　论文主要内容

图 1.9-A 是 DNA 相关数据集的构建和整合，用以构建能够支撑研究所需的较大样本的多癌种 DNA 甲基化数据集；在图 1.9-B 中本文提出基于类别特异性过滤

的 DNA 甲基化位点识别方法，能够从数十万的 DNA 甲基化位点中识别出多种类型肿瘤的特异性 DNA 甲基化位点，并进行了聚类分析；在图 1.9-C 中，本文对过滤得到的 DNA 甲基化位点进行特异性衡量和排序，挑选出最关键的特征，并使用并构建了机器学习模型进行肿瘤组织源预测；此外在图 1.9-D 上的实验结果证明本文方法的有效性。

本文由 5 个章节构成，各个章节的组织结构和相互关系如图 1.10所示，其中第 2 章到第 4 章是本文研究的主体内容：

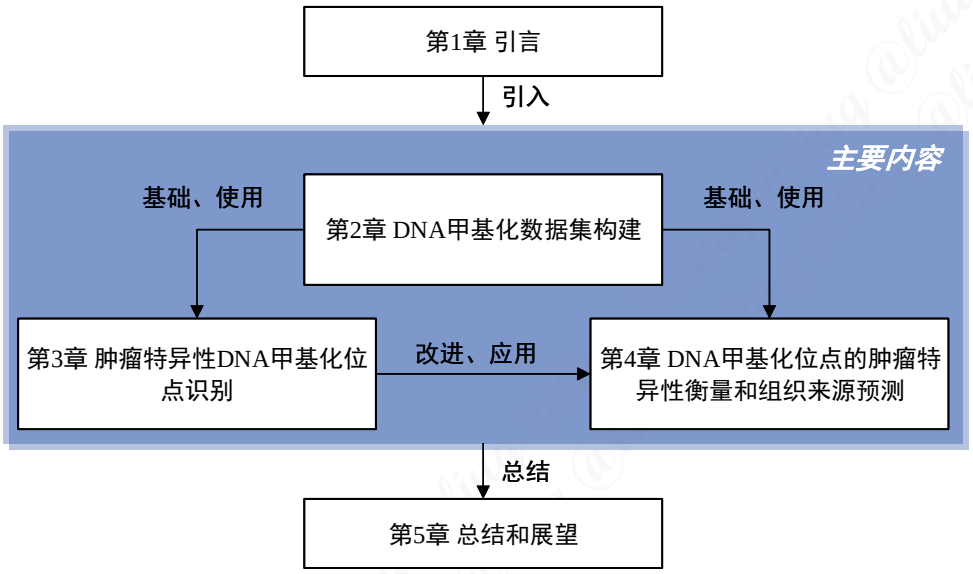


图 1.10　论文各章节安排和关系

第 1 章，引言，概述了基于外周血进行癌症液体活检的研究背景和意义，对国内外相关研究进行文献综述，总结现有研究存在的不足，阐述本文的研究动机和主要内容。

第 2 章，DNA 甲基化数据集构建。针对本文研究任务所需的多类肿瘤 DNA 甲基化数据集缺乏的问题，基于现有的公开肿瘤相关数据库和相关研究，构建了用于多癌种肿瘤特异性 DNA 甲基化位点识别和组织来源预测的肿瘤组织和外周血来源的 DNA 甲基化数据集。本章主要包含 DNA 甲基化检测方法原理的概述、DNA 甲基化数据的获取来源、预处理流程和处理后的数据集的统计分析，是本文研究的基础。

第 3 章，肿瘤特异性 DNA 甲基化位点识别。提出了基于类别特异性过滤的 DNA 甲基化位点识别方法，识别出 OvR（One vs Rest，一对多）类型的肿瘤特异性的 DNA 甲基化位点，并对识别得到的 DNA 甲基化位点特征数据进行无监督聚类分析。

第 4 章，DNA 甲基化位点的肿瘤特异性衡量和组织来源预测。针对类别数增

加可能无法得到统计意义上显著的肿瘤特异性 DNA 甲基化位点的问题，基于统计显著性水平和互信息对过滤得到的 DNA 甲基化位点进行肿瘤特异性衡量和排序，构建了多种机器学习模型进行肿瘤组织来源预测。实现了通过选取尽量少的肿瘤特异性 DNA 甲基化位点在肿瘤组织来源预测中取得较高的预测性能。

第 5 章，总结和展望。对本文的工作进行了总结，讨论了本文研究存在的不足之处，并展望了未来在 DNA 甲基化标志物识别和外周血肿瘤液体活检可以开展的工作。

第 2 章　DNA 甲基化数据集构建

2.1　DNA 甲基化检测方法概述

2.1.1　常见 DNA 甲基化检测方法

DNA 甲基化的检测和评估，按照检测区域的范围的不同，一般可分为靶向的方法和全基因组的方法，前者需在预先选定感兴趣的区域内检测 DNA 甲基化信号，而后者则可以在更大的范围检测到 DNA 甲基化信号。按照检测原理的不同，可以分为基于沉淀富集的方法、于亚硫酸氢盐转换的方法和基于限制性核酸内切酶的方法，如图 2.1所示。

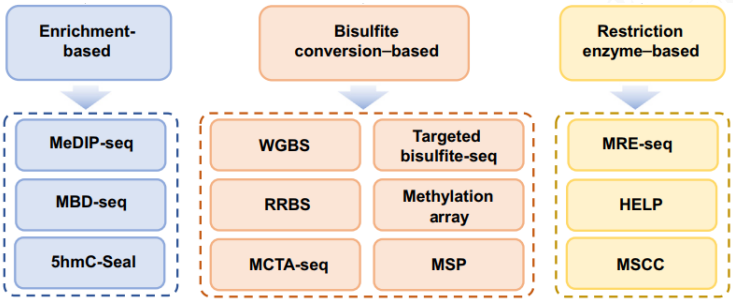


图 2.1　常见的 DNA 甲基化检测方法[42]

当前 DNA 甲基化检测的“金标准”是基于亚硫酸氢盐转换的方法。其核心是对 DNA 样品进行亚硫酸氢盐处理，使 DNA 中甲基化的胞嘧啶（C）保持不变，未甲基化的 C 则被处理成为尿嘧啶（U），从而在后续的测序或者杂交后区分开来。代表的方法有全基因组甲基化测序、RRBS（reduced-representation bisulfite sequencing）、MCTA-seq（methylated CpG tandems amplification and sequencing ），以及甲基化微芯片/微阵列。全基因组甲基化测序具有单个碱基的检测分辨率的能力和优势，包含低 CpG 密度区、基因间区域等在内的 DNA 甲基化信号都能检测出，缺点是要求的测序深度很高，成本高昂。RRBS 以及 MCTA-seq 方法相较于全基因组甲基化测序而言成本更低，所需要的 DNA 更少，但与此同时能够覆盖的区域和检测到的信息也较少。甲基化微芯片主要指当前广泛使用的两种商用 DNA 甲基化芯片 Illumina Infinium HumanMethylization450 BeadChip（HM450K）和 Infinium Methylizationpic BeadChip（HM850K），均由 Illumina 公司研制，分别覆盖约 45 万和 85 万个 CpG 位点，具有操作简便、处理规范、通量相对较高等优点。

2.1.2　DNA 甲基化芯片原理概述

基于 DNA 甲基化芯片检测方法原理同样是基于亚硫酸氢盐转换。DNA 在经过分离提取后，如图 2.2所示，首先要经过亚硫酸氢盐的处理，序列上未甲基化的 C 转化为 U，甲基化的 C 保持不变，随后在全基因组扩增（whole genome amplification，WGA）阶段，U 又被转换成为 T（胸腺嘧啶），而甲基化的 C 继续保持不变。DNA 再经过片段化以及沉淀后就与芯片上的探针进行杂交，此后 DNA 上位点的甲基化水平被转换成为光强度信号，被获取装置扫描和成像，不同位点的甲基化水平通过荧光的类型以及强度来量化和反应，后续可以进行相关的质量控制以及高层级的数据分析与应用。将处理后的样本输入到甲基化芯片进行杂交，芯片上混合使用了 I 型探针和 II 型 Bead type。

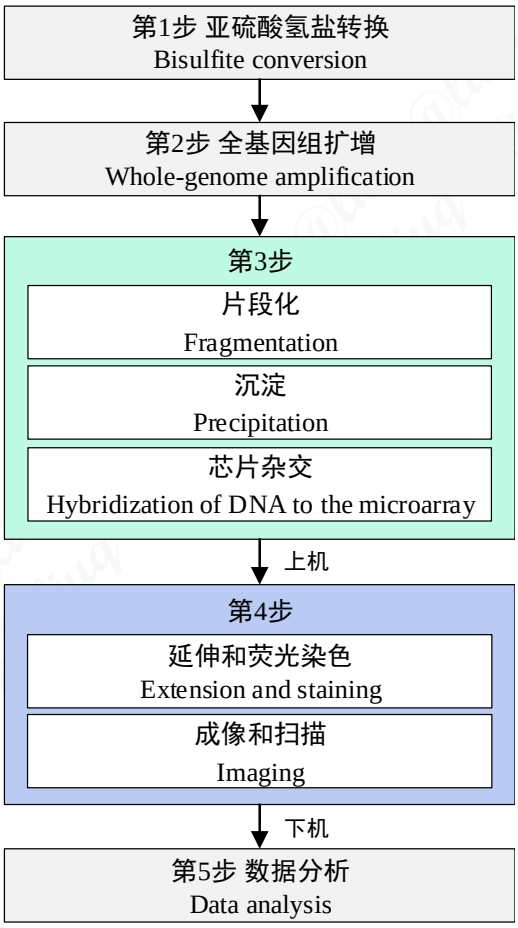


图 2.2　DNA 甲基化芯片主要处理流程

如图 2.3所示，左侧为一张 HM450K 芯片的示意图，一张芯片分为六行两列共 12 个 slide 故可以对 12 个单独的样品进行检测，每个 slide 中有数以万计的二氧化硅珠子固定在经过蚀刻的孔中，每个珠子上都附着有提前设计好的拥有多个拷贝寡核苷酸探针。DNA 甲基化芯片能够达到单个碱基的检测分辨率，如 HM450K 芯

片可以检测 485577 个 CpG 的甲基化状态，而目前较新的 HM850K/EPIC 甲基化芯片可以检测 853307 个 CpG 的甲基化状态，具有更高的基因组覆盖率。

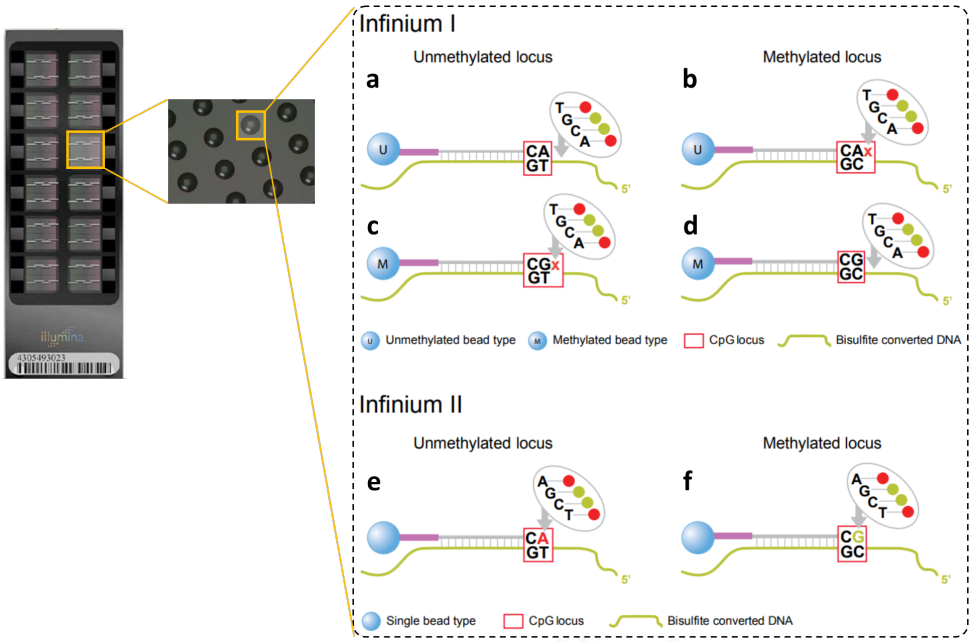


图 2.3　DNA 甲基化芯片检测原理[43]

HM450K 芯片同时使用 Infinium I 和 Infinium II 两种类型的探针来确定 CpG 的甲基化状态，其中 Infinium I 型探针对于每个 CpG 基因座都设计两种类型的珠子来检测其状态，一种用于确定甲基化状态另一种用于确定非甲基化状态，若经过转换的 DNA 样本的目标 CpG 发生甲基化，如前文所述将呈现出未转化的状态，因而会和 3′ 端以 G（鸟嘌呤）结尾的 M 类型的探针相结合形成一个 CpG（如图 2.3-d 所示）；如经过转换的 DNA 样本的目标 CpG 本身未发生甲基化，那么将会和结尾为 A 的 U 型探针相结合，从而形成 ApT 结构（如图 2.3-a 所示）；而发生甲基化的 DNA 样本无法和 U 型探针的 3′ 端末尾结合，未发生甲基化的 DNA 样本无法和 M 型探针的 3′ 端末尾结合。而后探针进行延伸从添加带有荧光标记的单核苷酸，假设靶向的 CpG 之后的所有碱基序列相同，那么两种类型的探针会发相同颜色的荧光，进而甲基化的程度就可以用发相同光的两种不同类型探针的荧光值的比例来衡量。

而 Infinium II 型探针使用一种珠子，每个探针可以允许至多三个潜在的 CpG 位点，在每个潜在的 CpG 检查位置存在一个简并碱基，在随后的延伸过程中对于 DNA 样本上甲基化的 C，位置以带有绿色荧光标记的 G 进行互补（ 2.3-f），对于未甲基化代表的 T 则以红光标记的 A（腺嘌呤）进行互补，最终通过对基因座两种荧光信号的强度来确定该位置的甲基化状态。对于任意第 i 个 CpG 位点，其甲

基化水平一般使用 β 值来衡量，具体而言 β 值是由甲基化信号 $y_{i,methy}$ 以及非甲基化信号 $y_{i,unmethy}$ 的强度比值得到的，计算公式(2.1)如下：

$$\beta_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha} \tag{2.1}$$

其中，α 是为了在当甲基化和未甲基化的探针强度都很低时，用来规范 β 值的一个偏置量，通常取 100。值统计得出的数字介于 0 和 1 之间，或 0 和 100%之间，在理想条件下，0 表示样品中 CpG 位点的所有拷贝都是完全未甲基化的，值为 1 表示该位点的每个拷贝都被甲基化[35]。

2.2　DNA 甲基化数据集构建方法

2.2.1　DNA 甲基化数据获取

2.2.1.1　TCGA 数据库

癌症基因图谱（The Cancer Genome Atlas，TCGA）计划[44]是美国国家癌症研究中心和美国国家基因组研究中心支持的一个大型国际肿瘤研究项目。该项目于 2005 年开始启动，通过高通量基因组分析和生物信息学技术得到癌症基因突变等数据，旨在提高对癌症诊断、治疗和预防的能力，已经研究了 33 种癌症，是目前世界上最全面的人类癌症临床和分子数据库。如图 2.4所示截止 2021 年 2 月，TCGA 数据库（https://portal.gdc.cancer.gov/）已包含 68 个癌症项目，涉及到的原发位置有 67 个，覆盖 84591 个样本。按照数据类型划分 TCGA 主要包括：

1. 临床诊断数据：例如年龄、性别、TNM 肿瘤分级、吸烟史、生存状况等；
2. 测序得到的读段数据，通常不对外公开；
3. 转录组数据包括 mRNA 表达和 microRNA 表达两大类，其中 mRNA 表达量数据由 mRNA 芯片或 RNA-Sequencing 测得，microRNA 表达量数据由 microRNA 芯片或 microRNA Sequencing 测序得到；
4. 拷贝数变异数据：由单核苷酸异质性（single nucleotide polymorphism，SNP）芯片测序得到的肿瘤对比正常组织染色体；
5. 单核苷酸异质性数据：肿瘤组织测序数据相对于参考基因序列得到的核苷酸变化，包含核苷酸的插入缺失等基因突变变异情况；
6. DNA 甲基化数据：主要通过 DNA 甲基化芯片获取；
7. 生物样本数据：通常是对于癌组织以及癌旁组织的染色切片图像。

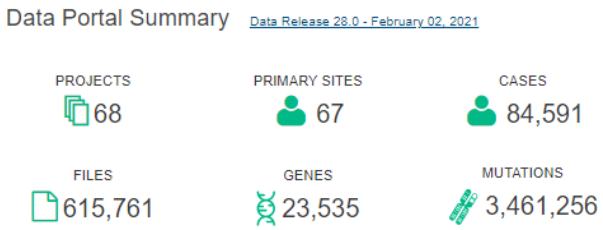


图 2.4　TCGA 数据库现有数据集统计（截至 2021 年 2 月 2 日）

2.2.1.2　Xena 数据库

UCSC Xena①[45]是一个多组学和临床/表型数据浏览和分析在线平台，由加州大学圣克鲁兹分校基因组研究所构建和维护。如图 2.5所示，Xena 整合了包括 TCGA、GDC 在内的多种肿瘤相关公共数据库中的多组学数据，涉及 50 种癌症类型。Xena 主要分为两部分，第一个部分是基于 Web 用于交互的 Xena Browser，为研究者提供了较为方便的多组学数据探索接口、基础的分析工具与可视化功能；另一个部分是后端 Xena Hubs，支持研究者上传自己的私有数据，以及对受限的数据进行保护。

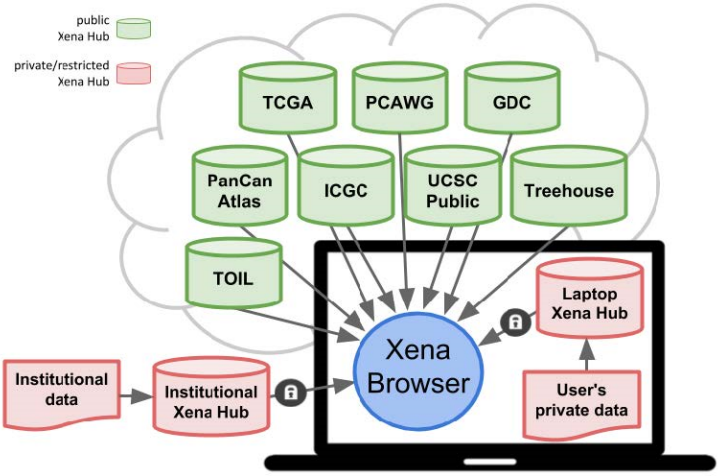


图 2.5　Xena 数据库概览[46]

2.2.1.3　GEO 数据库

GEO②（Gene Expression Omnibus）数据库[47]由美国国家生物技术和信息中心（National Center for Biotechnology Information，BCIB）建立于 2000 年，最初专门用来存储基因表达数据，现已发展成为包括 DNA 甲基化数据在内的一个综合数据库。其中的数据主要由研究人员按照规范进行上传，平台提供存储、检索以及部分基础的分析功能。是重要的生物信息学数据获取来源。

① http://xena.ucsc.edu/
② http://www.ncbi.nlm.nih.gov/geo/

2.2.1.4　PanSeer 肿瘤组织和血液 DNA 甲基化数据集

复旦大学陈兴栋等人的研究[31]构建了 PanSeerDNA 甲基化数据，其主要包含两部分来源，一部分是来自 BioChain①肿瘤和癌旁组织的 DNA 甲基化样本，样本数量如图 2.6所示，共有肺癌、胃癌、乳腺癌、结肠癌四种类型癌症的组织样本共 192 个，每类样本数均为 48，其中癌组织和癌旁组织样本数分别为 40 和 8，共有 160 个肿瘤组织样本和 32 个癌旁组织样本。

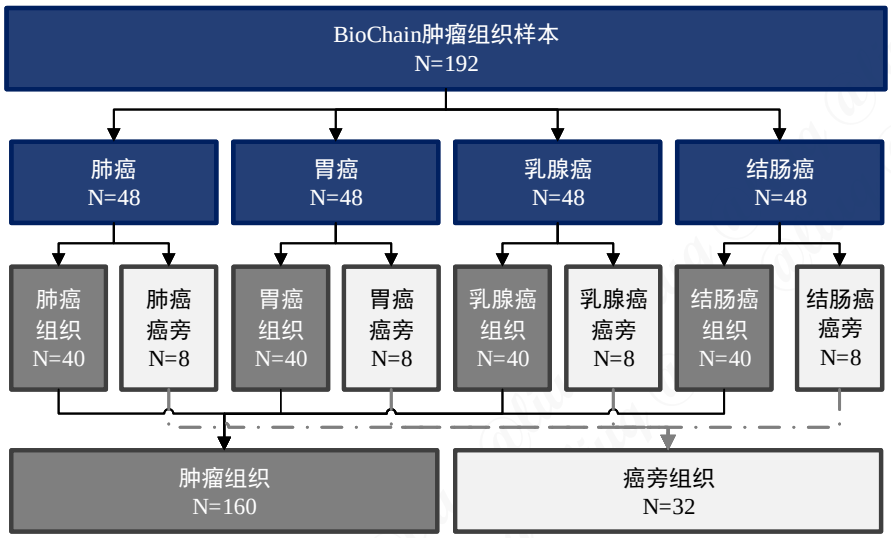


图 2.6　PanSeer 中来自 BioChain 肿瘤组织和正常组织样本的数量

另一部分是来自一项由复旦大学牵头的“泰州纵向研究（Taizhou Longitudinal Study，TZL）”，所有血浆样本均由泰州健康科学中心（Taizhou Institute of Science and Health，TSH）采集，如图 2.7所示，共有 223 个血浆样本采集自 5 种癌症患者，其中肺癌（N=56）、胃癌（N=69）、肝癌（N=23）、结肠癌（N=7）、食管癌（N=68），即构成确诊后采血数据集；另外来自 605 个采样时无症状的人群，其中 191 名在随后四年的随访中被诊断出患有癌症，构成确诊前采血的数据集，肺癌（N=47）、胃癌（N=35）、肝癌（N=29）、结肠癌（N=35）、食管癌（N=45）；剩余采集自无症状者的血浆样本被认定来自健康（N=414）人群。

研究[31]研发的 PanSeer DNA 甲基化检测分析方法设计了 607 个 DNA 甲基化区域作为特征，特征值的含义为平均甲基化分数（Average Methylation Fraction，AMF），对任意一个靶向的区域 R_k，总共覆盖长度为 M 个 CpG 位点，则该区域上的 AMF 计算如公式(2.2)所示：

① https://www.biochain.com/

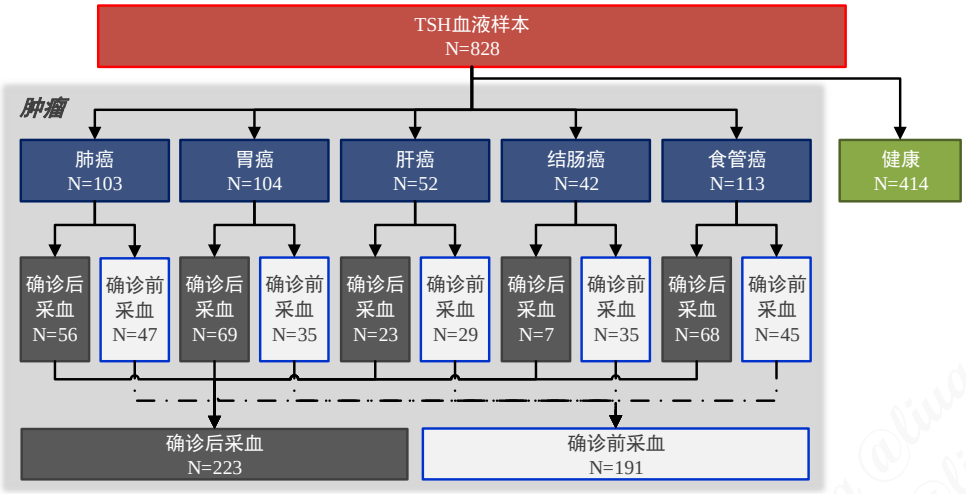


图 2.7　PanSeer 中来自肿瘤患者和健康人血浆样本的数量

$$AMF_{R_k} = \frac{\sum_i^M N_{C,i}}{\sum_i^M \left(N_{C,i} + N_{T,i}\right)} \tag{2.2}$$

其中 $N_{C,i}$ 表示通过 PanSeer 检测方法在 CpG 位点 i 上所有 read 上得到的胞嘧啶（C）计数值，公式(2.2)的分子部分表示整个区域 k 上所有胞嘧啶（C）的个数，$N_{T,i}$ 表示 CpG 位点 i 上所有 read 上得到的胸腺嘧啶嘧啶（T）计数值，分母部分表示整个区域上 C 和 T 的个数总和。

2.2.2　450K DNA 甲基化芯片数据预处理

对于 Illumina HumanMethylation450 BeadChip 芯片而言，对应于 485577 个 CpG 位点，从 TCGA 或 Xena 可以获取肿瘤和癌旁组织 DNA 甲基化芯片的 β 值数据，对于 GEO 数据库以及部分文献中的 DNA 甲基化 450K 芯片数据，得到的是下机后未经处理的 IDAT 文件，对应于前文所述 Infinium I 和 II 得到的光强信号值，本文使用 minif[48]对 IDAT 文件进行处理，从而得到 β 值数据。随后对于得到的 450K DNA 甲基化 β 值数据，本文的按照如图 2.8所示的环节进行数据预处理，这里的预处理从生物信息学的角度可称之为“质量控制”（quality control，QC），是一系列处理流程的集合，旨在尽可能减少因为生物实验处理而引入的变量，并保留生物样本本身所具有的变量。

2.2.2.1　缺失值处理

DNA 甲基化数据的缺失值处理常见有两种策略，其一是直接对含有缺失值的样本或者特征进行删除；其二是进行缺失值填充，利用 K 最近邻（K-nearest neighbors，KNN）算法进行缺失值填充在 DNA 甲基化数据处理中较为常见[49]，对

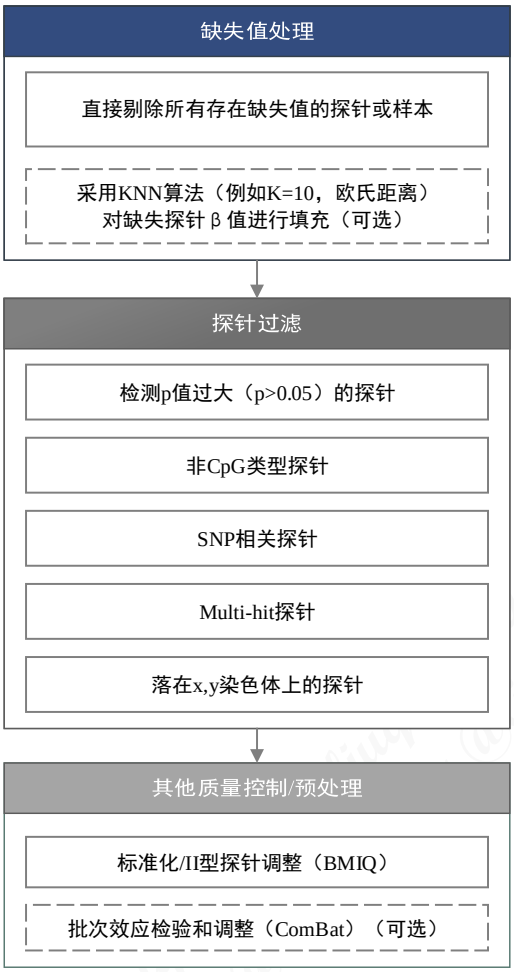


图 2.8　DNA 甲基化 450K 芯片数据预处理流程和方法

于任意一个含有缺失值的特征向量 u，计算到该特征不缺失的列（样本）上欧氏距离最近的 K 个特征，见公式(2.3)。

$$||u-v||_2=\left(\sum|(u_i-v_i)|^2\right)^{1/2} \tag{2.3}$$

即对于候选的候选邻近特征 v，也有可能在某些样本上存在缺失，本文只在不缺失的样本上计算欧氏距离的平均值，计算得到 K 个最近的特征后，将含有缺失值的探针上的缺失项，用 K 个近邻对应位置的均值作为填充值。

由于本文构建的 DNA 甲基化数据集较大，特征维度较高，为了保证数据的准确性同时便于计算，本文直接删除含有缺失值所在的特征行，对于 450K 芯片数据而言，即直接剔除包含缺失值的 CpG 位点。

2.2.2.2　探针过滤

探针过滤包含多个过滤条件和步骤，首先是过滤未能成功杂交的探针，即过滤掉检测 *p* 值（detection *p*-value）过大（*p*>0.05）的探针，该检测 *p* 值保存于下机

形成的 IDAT 文件中，过大通常意味着检测的信号和背景噪声没有差异，另外还要过滤掉和少于三个珠子结合的探针，以及所有非 CpG 类型的探针。在所有数据集上本文根据文献[50]过滤掉所有单核苷酸异质性（SNP）相关的探针；过滤掉文献[51]提及的 multi-hit 探针，另外过滤掉所有落在性染色体上（即 X，Y 染色体）的探针。具体实现上，本文使用 ChAMP[52] R 包进行处理。

2.2.2.3　标准化

如前文所述 HM450K 芯片使用两种类型的探针（即 Infinium I 和 Infinium II 型）来检测 DNA 甲基化信号，即本文得到的 450K DNA 甲基化芯片的特征包含两种类型，这会导致 II 型探针相较于 I 型探针范围缩小，本文希望两种类型探针的信号强度分布大致一样，即对 II 型探针的分布进行调整或者成为标准化（normalization），调整的方法有 SWAN[53]，BMIQ（Beta MIxture Quantile dilation）[54]，PBC[55]等，本文使用 BMIQ 来对所有经过前面处理环节的所有甲基化数据集的 II 型探针进行调整。实现上，本文使用 wateRmelon[56] R 包进行处理。

2.3　数据集构建结果分析

2.3.1　TCGA 肿瘤组织 DNA 甲基化芯片数据集

本文以 TCGA 中选择 14 类较为常见的癌症作为研究对象，癌症类别和含义如表 2.1所示。从 UCSC Xena①[46]下载对应样本整合后的 DNA 甲基化数据，其均采用 Illumina Infinium HumanMethylation450 BeadChip 平台收集得到。探针映射到基因组上的注释信息采用 GEO GPL13534②注释文件，并从 TCGA③中检索和获取每个样本对应的临床数据。

表 2.1　从 TCGA 中检索和选择的含有 DNA 甲基化数据的 14 类癌症类型

名称缩写	英文名称	中文全称
BLCA	Bladder urothelial carcinoma	膀胱尿路上皮癌
BRCA	Breast invasive carcinoma	乳腺浸润性导管癌
COAD	Colon adenocarcinoma	结肠腺癌
ESCA	Esophageal carcinoma	食管癌
HNSC	Head and Neck squamous cell carcinoma	头颈部鳞状细胞癌
KIRC	Kidney renal clear cell carcinoma	肾透明细胞癌

① http://xena.ucsc.edu/
② https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534
③ https://portal.gdc.cancer.gov

表 2.1　从 TCGA 中检索和选择的含有 DNA 甲基化数据的 14 类癌症类型（续）

名称缩写	英文名称	中文全称
KIRP	Kidney renal papillary cell carcinoma	肾乳头状细胞癌
LIHC	Liver hepatocellular carcinoma	肝细胞癌
LUAD	Lung adenocarcinoma	肺腺癌
LUSC	Lung squamous cell carcinoma	肺鳞癌
PRAD	Prostate adenocarcinoma	前列腺腺癌
STAD	Stomach adenocarcinoma	胃腺癌
THCA	Thyroid carcinoma	甲状腺癌
UCEC	Uterine Corpus Endometrial Carcinoma	子宫内膜癌

获得的 TCGA 肿瘤组织样本如表 2.2所示，原始的每类癌症样本既包括癌组织（含转移组织）也包括正常的癌旁组织的 DNA 甲基化数据。本文只选择其中样本类型为原发癌组织的样本，根据 TCGA 样本和 ID 的对应规则即 TCGA 条形码，本文对每个组织案例只选择其 vial 为 A 的样本，即一个案例对应于一个组织样本，其余的样本以及对应临床数据缺失的样本都被排除。最终得到 14 类癌症共计 5765 个原发肿瘤组织来源样本的 DNA 甲基化数据。

表 2.2　来自 TCGA 的 14 类癌症 450KDNA 甲基化芯片数据样本类型和计数

	癌组织			癌旁组织		
癌种	原发	转移	排除	癌旁	排除	总计
BLCA	409	1	3	21	0	434
BRCA	775	5	10	81	17	888
COAD	293	1	5	38	0	337
ESCA	174	1	11	16	0	202
HNSC	523	2	5	45	5	580
KIRC	316	0	4	160	0	480
KIRP	274	0	2	45	0	321
LIHC	375	0	4	50	0	429
LUAD	455	0	5	32	0	492
LUSC	366	0	6	41	2	415
PRAD	484	1	14	49	1	549

表 2.2　TCGA 14 类肿瘤组织样本计数（续）

	癌组织			癌旁组织		
癌种	原发	转移	排除	癌旁	排除	总计
STAD	393	0	3	2	0	398
THCA	502	8	5	54	2	571
UCEC	426	0	6	46	0	478
合计	5765	19	83	680	27	6574

2.3.2　GSE40279 健康人血液 cfDNA 甲基化芯片数据集

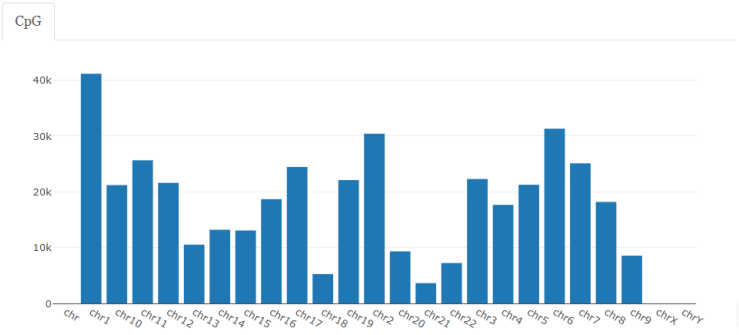
本文借鉴研究徐瑞华、罗慧妍和 S Kang 等人的研究[25,28-29]中的做法来构建能够进行液体活检的多类癌症数据集。其中健康的血液数据集来自加州大学圣迭戈分校 Gregory Hannum 等人 2013 年发表在《Molecular Cell》的一项研究（GSE40279），该数据集的获取的实验平台和 TCGA 中 DNA 甲基化数据一致，得到了 656 个来自健康人群的样本，临床信息如表 2.3所示。

表 2.3　GSE40279 数据集临床信息统计

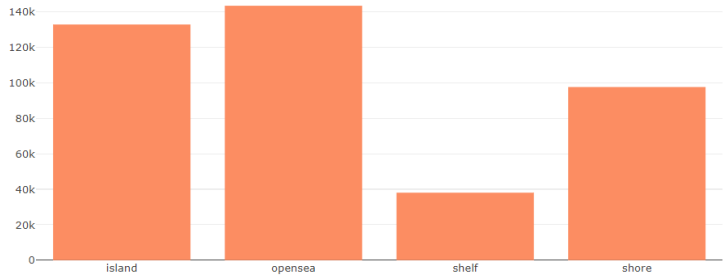
临床和病理特征	取值	计数
Age	N (Nmiss)	656（0）
	Mean±SD	64.04±14.74
	Min-Max	19-101
	Median (Q1-Q3)	54-75
Gender	Male (%)	338 (48.5%)
	Female (%)	318 (51.5%)
Source	Boston	35 (5.3%)
	UCSD	304 (43.6%)
	USC	139 (21.2)
	Utah	178 (27.1)
Ethnic	Caucasian - European	426 (64.9%)
	Hispanic - Mexican	230 (35.1%)
Tissue	whole blood	656 (100%)

如图 2.9所示是 GSE40279 经过过滤后剩余的探针在基因组上的分布情况，其中图 2.9(a)表示过滤后的探针在染色体上的分布统计情况，图 2.9(b)表示过滤后的

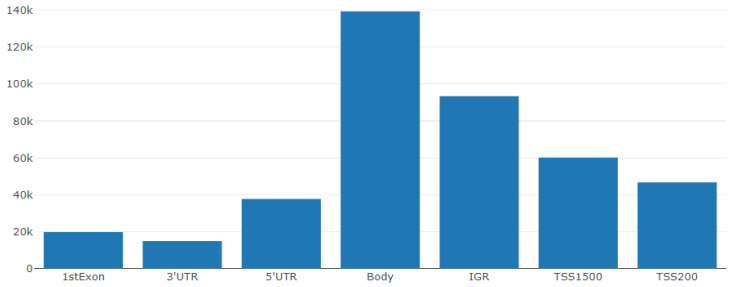
探针和 GpG 岛的相对位置关系统计结果，图 2.9(c)表示过滤后的探针和基因的相对位置关系统计结果。对于其他 450K DNA 甲基化数据本文可以进行相同的统计分析。



(a) 在染色体上的分布统计直方图



(b) 和 GpG 岛的相对位置关系统计直方图



(c) 和基因的相对位置关系统计直方图

图 2.9　在 GSE40279 数据集上进行探针过滤后的 CpG 位置统计

每类样本采用分层抽样的方法，按照训练集: 测试集为 6:4 的比例随机采样构建训练集和测试集，并作为统计和机器学习模型的输入。每类（癌症和健康）的样本个数如图 2.10所示，总共包含 14 类癌症以及健康人血液来源样本共 6421 个，其中训练集样本 3858 个，测试样本 2563 个，其中最多的是乳腺癌（BRCA），最少的是食管癌（ESCA），其余癌症和健康类别样本数量介于二者之间。

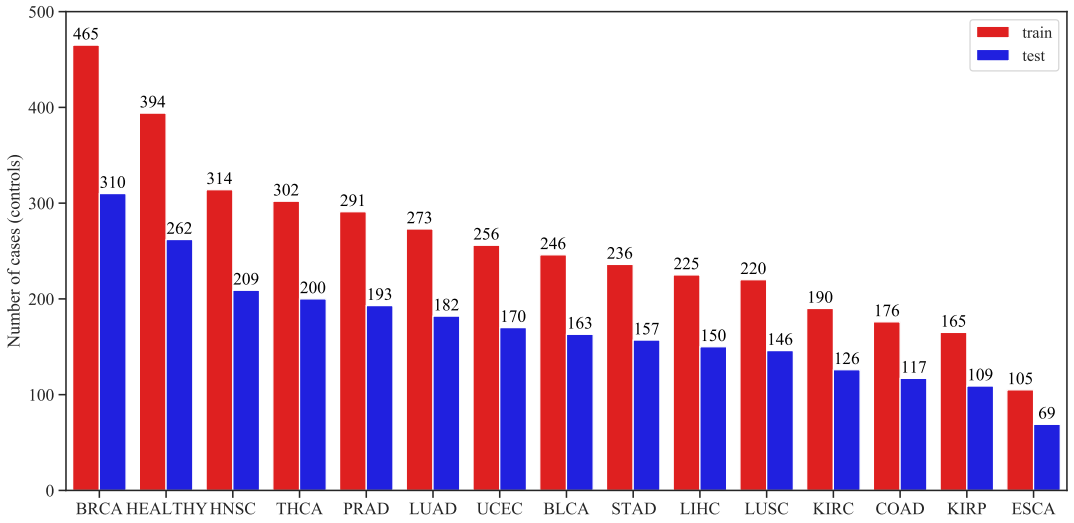


图 2.10　T14B1 数据集训练和测试集样本数量统计

对 TCGA 肿瘤组织以及健康人血液来源（GSE40279）450K 芯片 DNA 甲基化数据采用前文所述相同的预处理方法，对于每类处理过后的特征，本文取其公共的部分随后进行合并。如表 2.4所示，是 450K 芯片数据探针数预处理前后变化。

表 2.4　450K DNA 甲基化芯片数据探针数在预处理前后变化

	探针数		
类别	预处理前	预处理后	排除
HEALTHY	473034	412039	60995
BLCA	485577	349611	135966
BRCA	485577	333366	152211
COAD	485577	346304	139273
ESCA	485577	348874	136703
HNSC	485577	343854	141723
KIRC	485577	351083	134494
KIRP	485577	352717	132860
LIHC	485577	340795	144782
LUAD	485577	347994	137583
LUSC	485577	349188	136389
PRAD	485577	350949	134628
STAD	485577	341397	144180
THCA	485577	341190	144387

表 2.4　450K DNA 甲基化芯片数据探针数在预处理前后变化（续）

类别	探针数		
	预处理前	预处理后	排除
UCEC	485577	346551	139026

由于预处理后各个类别剩余的探针类型和数量不等，公共的探针数与类别数有关，为了用于后续的肿瘤来源预测、肿瘤诊断等实验，本文基于预处理后的各类别 DNA 甲基化数据，创建了不同类型的数据集。

表 2.5　本文创建的 450K DNA 甲基化芯片数据集

数据集	类别		公共探针数	样本数		
	名称	个数		总数	训练	测试
T14B1	HEALTHY, BLCA, BRCA, COAD, ESCA, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, STAD, THCA, UCEC	15	283044	6421	3858	2563
T14	BLCA, BRCA, COAD, ESCA, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, STAD, THCA, UCEC	14	283105	5765	3464	2301
T4B1	HEALTHY, COAD, LIHC, LUAD, STAD	5	320237	2172	1304	868
L2B1	HEALTHY, LUAD, LUSC	3	342942	1477	887	590

如表 2.5所示，是创建完成的四个数据集对应的类别、样本以及探针信息，其中 T14B1 表示 14 类肿瘤组织和健康人血液 DNA 甲基化数据集，T14 表示 14 类肿瘤组织来源 DNA 甲基化数据集，T4B1 是中国最常见的四类常见癌症（结肠癌、肝癌、肺腺癌、胃癌）组织和健康人血液来源 DNA 甲基化数据集，L2B1 则为两类肺癌（肺腺癌、肺鳞癌）组织和健康人血液来源 DNA 甲基化数据集。T14B1 数据集样本总数最多达到了 6421 个，其中 3858 个样本作为训练集，2563 个作为测试集。样本最少的 L2B1 数据集，样本总数为 1477，训练集样本为 887，测试集样

本为 590。对于特征（公共探针）而言，其中 L2B1 的特征维度达到了 342942 维，T14B1 数据集特征维度为 283044 维，其他数据集的介于二者之间。

2.3.3　PanSeer DNA 甲基化数据集

由于原始的 PanSeer 中组织来源和血液来源的样本类别不完全一致，为了满足识别肿瘤特异性 DNA 甲基化位点以及后续的肿瘤组织来源预测的需求，本文只选择其中来自 BioChain 以及 TSH 公共的三类癌症即肺癌、胃癌和结肠癌患者血液和健康人血液来源 DNA 甲基化标志物数据。对 BioChain 数据而言，本文只选择这三类癌症其中的癌组织的部分，对 TSH 数据而言本文选用这三类癌症中确诊后采血以及健康人血液来源的样本。

如表2.6所示，共得到组织来源 DNA 甲基化样本 120 个，肺癌、胃癌、结肠癌均为 40 个样本，血液来源的样本共 546 个，其中健康人血液来源样本 414 个，癌症样本 132 个，癌症样本中肺癌、胃癌、结肠癌数量分别为 56、69 和 7 个。类似对于 DNA 甲基化芯片数据的处理，原始的 607 维特征经过预处理后剩余 477 维。将 TSH 来源的三类癌症和健康人血液的 DNA 甲基化样本按照 6 比 4 划分成训练集和测试集，各个样本数量如表 2.7所示。

表 2.6　从 PanSeer 中选用的 DNA 甲基化数据集样本类别和计数

	来源		
类别	TSH	BioChain	总计
Healthy	414	0	414
Stomach	69	40	109
Lung	56	40	96
Colon	7	40	47
总计	546	120	666

表 2.7　PanSeer 三类常见癌症和健康数据集实验样本计数

类别	训练	测试	总计
Healthy	248	166	414
Stomach	41	28	69
Lung	34	22	56
Colon	4	3	7
总计	327	219	546

2.4　本章小结

本章对 DNA 甲基化检测方法进行了概述，阐述了实验所采用的 DNA 甲基化芯片的原理、数据的获取来源、数据预处理方法，对肿瘤组织来源以及健康人血液来源的 DNA 甲基化数据进行了整合，构建了较大规模的 DNA 甲基化数据集，为后续分析环节奠定了基础。

第 3 章　肿瘤特异性 DNA 甲基化位点识别

3.1　本章引言

对 DNA 甲基化芯片得到的高维数据，从中选择关键的肿瘤特异性 DNA 甲基化位点，对肿瘤的精准诊断具有重要意义。关于 DNA 甲基化标志物的选择，通常的做法是进行差异甲基化分析，从中选择差异 DNA 甲基化标志物。DNA 甲基化标志物不限于是 CpG 位点，也可以是差异甲基化区域、区块、单元等。从计算角度看，从大量的 DNA 甲基化位点中筛选得到肿瘤特异性的 DNA 甲基化位点候选集以供临床验证，可将其视为一个特征降维过程，该过程中若不改变 DNA 甲基化位点的特征表示，则属于特征选择的范畴。特征选择本质上是从一个特征集合中选择其子集，具有可解释性的优势。特征选择方法整体分为三类：1）过滤法，通常按照特征的发散性或者特征之间的相关性来对各个特征进行打分，通过设定阈值大小或者待选择阈值的个数进行特征选择；2）包装法，通常根据目标函数例如预测评价指标，每次从特征空间中选择一部分子集或者排除若干特征；3）嵌入法，过程和过滤法类似，区别在于嵌入法通过某些机器学习模型进行训练得到各个特征的权重系数，根据系数大小来选择特征。

3.2　基于类别特异性过滤的 DNA 甲基化位点识别方法

3.2.1　方法整体框架

如图 3.1所示是本文提出的基于类别特异性过滤的 DNA 甲基化位点识别方法的整体框架。最左侧以四种不同颜色的方块来表示多类 DNA 甲基化数据，不同类别两两之间进行基于统计的差异 DNA 甲基化位点分析。

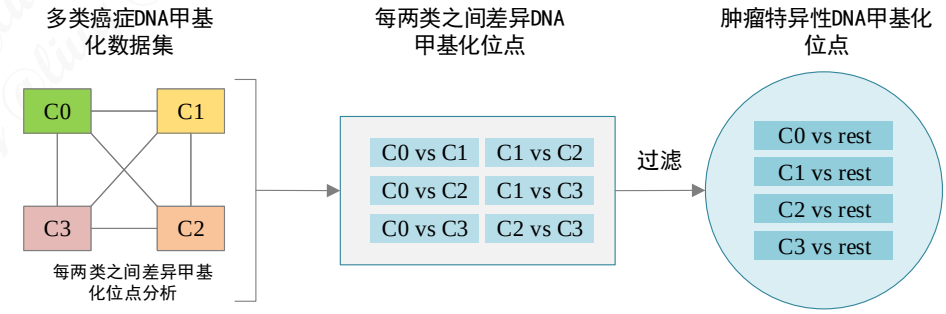


图 3.1　类别特异性过滤 DNA 甲基化位点识别方法整体框架

假定参与分析的类别数为 N，将得到 C_N^2 差异 DNA 甲基化位点分析的结果，

随后对计算得到的差异 DNA 甲基化位点进行过滤，得到每一类相对于其他类存在特异性的 DNA 甲基化位点集合，方法原理详情见后文。

3.2.1.1　符号表示

设预处理后的 DNA 甲基化 β 值矩阵或其他对 DNA 甲基化进行量化（例如 M 值[35]）得到的特征矩阵为 D。如图 3.2所示，每一行表示一维 DNA 甲基化特征，每一列表示一个样本/观测值，则有 $D=\{(x_i,y_i)\}, i\in(1,2,\cdots,m)$，其中 X 和 y 分别表示特征矩阵和对应样本的标记，$X\in\mathbb{R}^{p\times m}$ 表示共有 p 维的特征和 m 个样本。x_i 和 y_i 表示第 i 个样本和对应的标记，$x_i\in\mathbb{R}^{p\times 1}$，$y=\{y_1,y_2,\cdots,y_N\}$，即所有的样本的真实标记类别数为 N，图 3.2中以不同的颜色进行表示，若无特殊说明本文所用符号含义和图中一致。

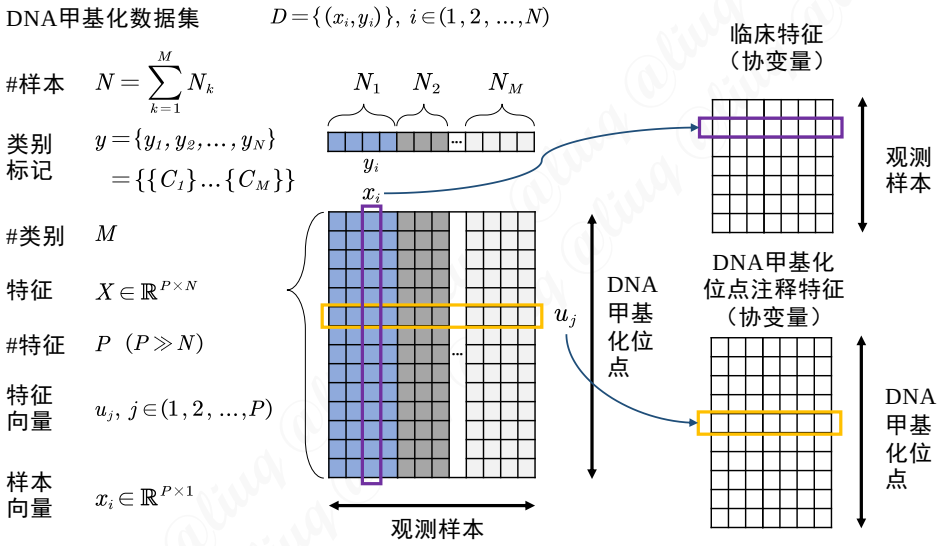


图 3.2　本文采用的 DNA 甲基化数据集相关符号表示

3.2.1.2　Welch's t-test

Welch's t-test 是 Student's t-test 的变种，也叫方差不齐 t-test，用于检验两个样本集具有相同的均值的假设，即两组之间的平均值差异是否在统计学上具有显著性，结果通常用 p 值来衡量。设定一个阈值 α，例如 0.05 或者 0.01 等，p 值低于 α 的认为其是统计意义上显著的值得进一步探究，而大于 α 的则认为不具有统计意义上的显著性。Welch's t-test 首先选择零假设（H0）和备择假设（H1），在零假设下统计量 t 近似符合 Student's t-test：

$$H_0:\mu_1=\mu_2 \tag{3.1}$$

$$H_1:\mu_1\neq\mu_2 \tag{3.2}$$

μ_i 表示类别为 i 的样本集对应的总体的均值。统计量 t 计算公式如(3.3)：

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{3.3}$$

其中 $\bar{X}_i$，s_i^2 和 N_i 分别表示类别为 i 的样本集的均值、方差和样本个数，统计量 t 的自由度估计如下：

$$df \approx \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}} \tag{3.4}$$

其中 X_i 表示样本集 i 的均值，N_i 是样本集 i 的观测值的数量，s_i 表示样本集 i 的方差即：

$$s_i = \sqrt{\frac{1}{N_i - 1} \sum_{t=1}^{N_i} \left(x_t - \bar{X}_i\right)^2} \tag{3.5}$$

使用 Welch’s t-test 在 PanSeer 数据集上进行差异分析，计算每个 DNA 甲基化标志物的统计量和得到的 p 值。

3.2.1.3　Benjamini-Hochberg 矫正

假设从 P 个 DNA 甲基化位点中挑选出 N_α 个特征，需要进行多重假设检验，但是每次统计推断都有可能产生假阳性率，其等于所设定的阈值 α，在所有选择得到认为有统计显著性的特征中，实际上并不显著所占的比例称之为错误发现率（False Discovery Rate, FDR）。令通过假设检验得到全部 P 个 DNA 甲基化位点的 p 值构成样本集 $\{p_i\}, i = 1, 2, \cdots, P$，显著性水平取 α 时，得到 N_α 个有显著性的特征，则这 N_α 个特征的错误发现率为：

$$FDR = 1 - \prod_{j=1}^{N_\alpha} \left(1 - p_j\right) \tag{3.6}$$

通过 Benjamini-Hochberg 过程来矫正得到的 p 值。该过程可分为两步，首先对所有 N_p 个 p 值从大到小排序，随后计算每个矫正过后的 p 值结果如下：

$$q_i = p_i \times \frac{N_p}{rank(p_i)} \tag{3.7}$$

其中 $rank(p_i)$ 表示 p_i 的秩，即在整个排序后 p 值序列中的位置顺序，效果是对于最大的 p 值和原来保持一致，其余 p 值进行了放大，从而在 α 阈值条件下过滤得到 FDR 更小的特征样本集。

3.2.1.4　Limma

Limma[36]是广泛应用于芯片数据差异分析的 R 程序包，不仅仅用于基因表达的分析，也可以对 DNA 甲基化进行差异分析。本文使用其来对 450K DNA 甲基化芯片数据进行两类之间的差异甲基化分析。Limma 对每个 DNA 甲基化位点都建立一个线性模型，随后通过经验贝叶斯（empirical bayesian）来从所有的 DNA 甲基化位点中估计每个 DNA 甲基化位点的后验方差，进而计算得到经过多重假设检验的 moderated t-statistics 和对应的 *p* 值，进而进行后续的差异 DNA 甲基化位点分析。

经过差异 DNA 甲基化分析计算癌症/健康类别的 DNA 甲基化位点统计量，结合 DNA 甲基化位点所在的基因组注释文件，可以将其作为 DNA 甲基化位点的特征。令参与 DNA 甲基化差异分析的类别分别用 T 和 C 表示，得到 DNA 甲基化位点特征如表3.1所示，可以将其中的一部分先验知识作为过滤条件，整合到本文的方法中用来筛选差异 DNA 甲基化位点。

表 3.1　借助 Limma 对 450K 芯片进行差异分析后结合基因注释信息得到的 DNA 甲基化位点的相关特征

缩写	名称和含义
T_AVG	类别 T 的平均甲基化值
C_AVG	类别 C 的平均甲基化值
P.Value	raw *p*-value，未经过矫正的 *p* 值
adj.P.Val	adjusted *p*-value，矫正后的 *p* 值（也称为 *q*-value）
t	moderated t-statistic 统计量
logFC	log fold change，对数差异表达倍数，和 deltaBeta 含义相同
deltaBeta	T_AVG-C_AVG 的结果，含义和 logFC 相同
AveExpr	该位点上的平均甲基化值
CHR	chromosome，染色体
cgi	CpG 岛的相对位置
MAPINFO	和标准基因组比对后的位置
gene	基因，位于哪个基因上
feature	和基因的相对位置，例如落到 1stExon

3.2.2　方法原理

差异 DNA 甲基化分析（Differential methylation analysis）通过比较两个属于不同的组（即类别）的样本集分布的差异来识别 DNA 甲基化标志物。当推广到多类癌症时，本文借鉴支持向量机进行多分类的思想，将其归纳为识别组间（between-group）和一对多（one-vs-rest，OvR）两种类型的 DNA 甲基化标志物，前者为能够对区分所有类别的 DNA 甲基化标志物，而后者是能够区分某个类别相对于其他类别的 DNA 甲基化标志物，下面对方法原理进行阐述。

目前针对两类数据的差异化分析通常利用统计假设检验的方法来获取具有显著差异（$p \leqslant \alpha$）的前 K 个标志物组成的集合 $M_\alpha = \{m_k \mid 1 \leqslant k \leqslant K\}$。扩展到 N 类数据上，能够得到 C_N^2 组两两类别之间差异显著的标志物集合 $\left\{ M_\alpha^{ij} \mid 1 \leqslant i \leqslant j \leqslant N \right\}$。为了构造能够同时区分多个类别的分类器，通常需要将两两类别之间不同的标志物集合进行整合得到统一的标志物集合。一个直观的整合方法是取所有 C_N^2 个标志物集合的交集：

$$\tilde{M}_\alpha = \bigcap_{i=1}^{N-1} \bigcap_{j=i+1}^{N} M_\alpha^{ij} \tag{3.8}$$

交集的含义在于得到的标志物集合 $\widetilde{M}$ 中，每个标志物 m_k 都具有区分所有类别的潜力。换句话说，不同类别在标志物这个特征维度上具有差异显著的数据分布，如图 3.3所示。对于后续多分类任务而言，这显然是一个理论上合理且理想的标志物集合。然而，数据分布的上下界是确定的，因此可容纳的具有显著差异（例如 $p \leqslant 0.05$）的类别数有限。当类别数持续增长时，想要找到能够显著区分所有类别的标志物就会变得愈发困难，不得不降低“显著区分”的标准（例如 $p \leqslant 0.1$），如此寻找到的标志物便会损失其区分所有类别的准确度。

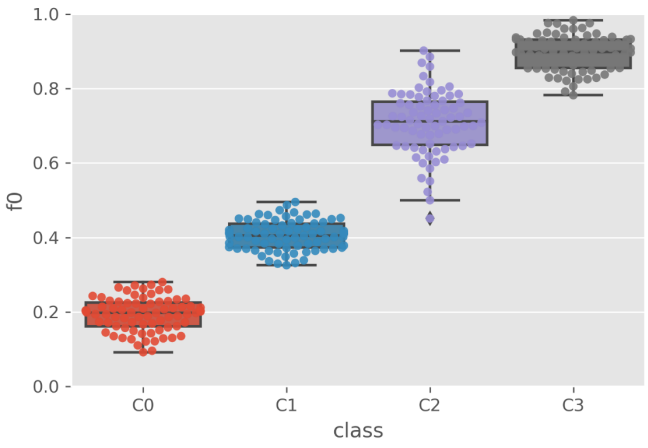


图 3.3　Between-group 类型 DNA 甲基化位点特征示意图

取交集得到的标志物虽然在一定范围内有助于后续的分类任务，但其本身并不是具有最佳类别特异性的标志物。针对第 n 类数据，更希望找到该类与其他所有类别间均具有显著差异的标志物 m_k^n，如图 3.4所示。

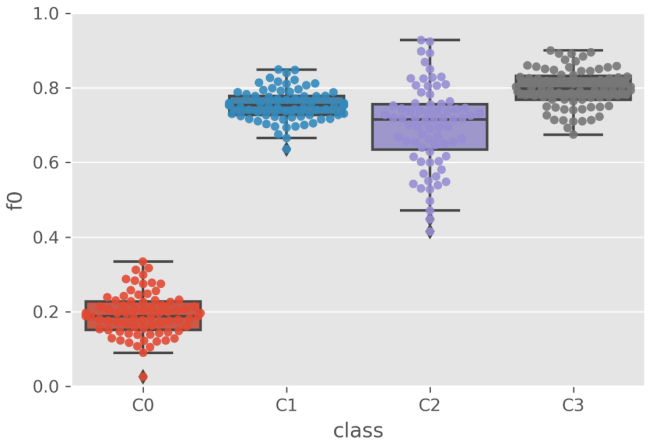


图 3.4　One-vs-rest 类型 DNA 甲基化位点特征示意图

具体而言，要求在 m_k^n 上，类别 n 与其他任何一个类别的数据分布都具有显著差异，而其他类别之间并没有显著差异：

$$p_{nj} \leqslant \alpha_1 \cap p_{ij} \geqslant \alpha_2, \forall i, j \neq n, 1 \leqslant i < j \leqslant N \tag{3.9}$$

其中，p_{nj} 和 p_{ij} 分别表示第 n 类与第 j 类数据和第 i 类和第 j 类间差异化检验的显著性水平，α_1 和 α_2 均表示显著水平阈值。α_1 约束了第 n 类与其他类别之间的差异水平，值越小表明第 n 类与其他类别的数据分布差异越大。α_2 约束了其他任意两类之间的差异水平，值越大表明其他任意两类之间的差异水平越小。由公式(3.8)得到的标志物是第 n 类的特异性标志物，即可以用来区分第 n 类和其他 $N-1$ 个类别。

公式(3.8)展示的计算方式并不直观，接下来以 3 类（A, B, C）为例介绍标志物的选取方式，如图 3.5所示。值得注意的是，该示意图需要保证 $\alpha_1 = \alpha_2$ 的情况下才成立，当两者不相同时以三类为例则对应有六个圆，另外圆的大小和集合的大小不成比例：

假定三个圆分别表示类别两两之间具有显著差异水平 α 的标志物集合 $M_\alpha^{AB}, M_\alpha^{AC}, M_\alpha^{BC}$。对于第 A 类而言，特异性标志物的筛选方式可以由以下公式得到：

$$M_{\alpha_1,\alpha_2}^{A} = M_{\alpha_1}^{AB} \cap M_{\alpha_1}^{AC} - M_{\alpha_2}^{BC} \tag{3.10}$$

该集合为图中标注为红色和绿色重叠后排除蓝色区域部分。同理，B 类和 C 类的

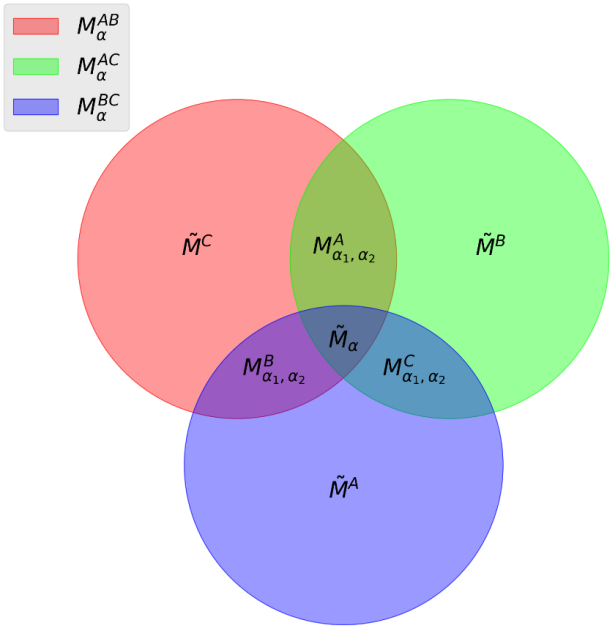


图 3.5　肿瘤特异性 DNA 甲基化位点集合韦恩图表示（以三类为例）

类别标志物集合分别是图中红色与蓝色重叠排除绿色、蓝色与绿色重叠排除红色的部分。更进一步地分析发现，中间三个圆相交的部分就是前文中取交集的过滤 between-group 类型标志物方法。三个集合中没有任何相交的部分表示在相应显著水平下仅能区分两个类别的标志物集合，三个集合以外的部分是不能对任意两个类别进行有效区分的所有特征位点集合。公式(3.10)得到的特异性标志物集合同样可以通过排除掉交集标志物集合得到，扩展到 N 类，对于第 c 类的特异性标志物集合需排除更多区域可由公式(3.12)得到：

$$\boldsymbol{M}_{\alpha_1,\alpha_2} = \bigcup_{c=1}^{N} \boldsymbol{M}_{\alpha_1,\alpha_2}^{c} \tag{3.11}$$

$$\boldsymbol{M}_{\alpha_1,\alpha_2}^{c} = \bigcap_{j=1,j\neq c}^{N} \boldsymbol{M}_{\alpha_1}^{cj} - \bigcup_{i=1,i\neq c}^{N-1} \bigcup_{j=i+1,j\neq c}^{N} \boldsymbol{M}_{\alpha_2}^{ij} \tag{3.12}$$

3.3　实验设计与结果分析

3.3.1　实验设计

3.3.1.1　层次聚类算法

通过层次聚类可以发现找寻的 DNA 甲基化位点和样本的层次关系，层次聚类是一种基于相似性的无监督聚类算法，相似度通常指的是距离。本章采用自下而

上的方式来构建层次聚类树，位于聚类树底部的是经过类别特异性过滤的 DNA 甲基化位点识别方法过滤得到的 DNA 甲基化位点数据，树的顶层是聚类得到的根节点，通过观察在样本上层次聚类结果，可以直观地判断过滤得到的 DNA 甲基化位点对于不同肿瘤的区分性。

算法 3.1 层次聚类

1. 计算所有 n 个肿瘤或健康人 DNA 甲基化样本点相互之间的相似性（采用欧氏距离）
2. 初始化 n 个聚类簇，即一个样本点作为一个单独的聚类簇
3. 合并聚类簇间距最小的两个 cluster，采用 ward 方法进行 cluster 之间距离的计算，并构建一个新的 cluster
4. 计算新的 cluster 与当前各 cluster 的距离，若 cluster 为 1，终止计算；否则转到 3

3.3.1.2　K 均值聚类算法

K 均值聚类算法广泛应用于无监督聚类分析，K 均值聚类有很多变种，本章采用经典的 K 均值聚类算法进行 DNA 甲基化位点的聚类分析，具体算法流程如下：

算法 3.2 K 均值聚类

1. 随机选取 K 个样本点作为初始化 cluster 中心
2. 对样本进行聚类。计算每个样本点到 K 个 cluster 中心的距离，并将其指派到距离最近的 cluster 中心，从而构成聚类结果
3. 对当前每个 cluster 中心计算其中样本的均值，作为新的 cluster 中心
4. 如果迭代收敛或符合停止条件，则返回当前聚类结果；否则放回步骤 2

本文设定聚类算法得到的样本簇标签数量和样本真实的类别数量相同，使用 Kuhn-Munkres 算法[57]求出簇标号到真实标记之间的最大匹配，从而将簇标号映射到与之对应到类别标记上，进而转换为分类指标来衡量识别到的肿瘤特异性 DNA 甲基化位点区分不同癌症的能力。

3.3.2　结果分析

3.3.2.1　DNA 甲基化芯片数据集上实验结果

如图 3.6所示，是在 L2B1 数据集上以参数 α=1e-90，过滤得到的 24 个 between-group 类型肿瘤特异性 DNA 甲基化位点的层次聚类结果，上方颜色长条代表不同样本对应的类别，横轴表示不同的探针（每个探针和一个 CpG 位点对应），聚类结果显示出较为明显的三类聚类模式，能对两种肺癌亚型和健康样本进行区分。改变过滤 α，过滤得到更多的 between-group 类型肿瘤特异性 DNA 甲基化位点。类似的在 T4B1 数据集上，α=1e-20，过滤得到 28 个 between-group 类型肿瘤特异性 DNA 甲基化位点，层次聚类结果如下图所示，虽然仍能看出 5 组不同的 DNA 甲基化模式但效果不明显。

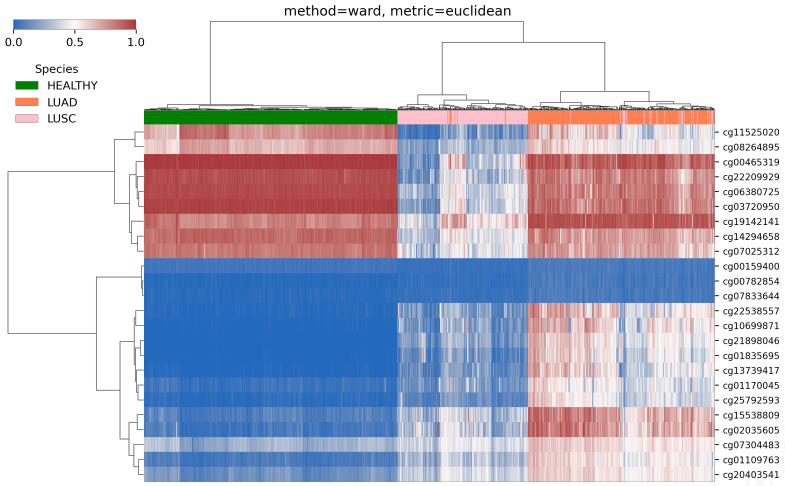


图 3.6　在 L2B1 训练集上过滤得到的 24 个 between-group 类型肿瘤特异性 DNA 甲基化位点和特征矩阵层次聚类结果

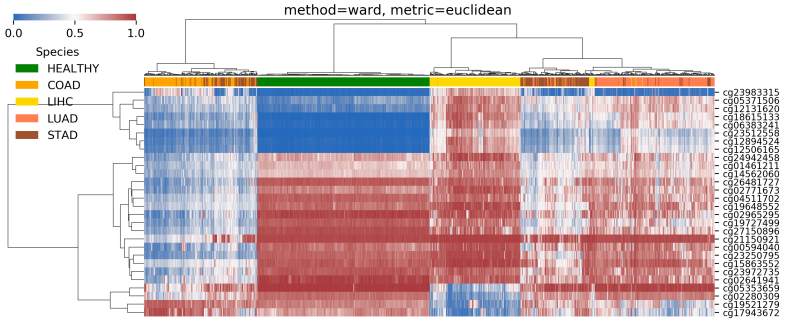


图 3.7　在 T4B1 训练集上过滤得到的 28 个 between-group 类型肿瘤特异性 DNA 甲基化位点和特征矩阵层次聚类结果

对于 between-group 类型肿瘤特异性 DNA 甲基化位点的识别，当类别增多的时候，会面临过滤条件过于严苛而无法筛选到符合统计假设检验的 DNA 位点，实验结果如表 3.2，可以看出在相同统计显著性的约束条件下，随着类别的增加 between-group 类型肿瘤特异 DNA 甲基化位点数量逐渐减少，从而验证了之前的分析。

表 3.2　随着类别的增加 between-group 类型肿瘤特异 DNA 甲基化位点数量逐渐减少

α	L2B1	T4B1	T14B1
0.05	247565	55757	0
0.01	221788	33521	0
0.005	212730	27597	0
0.001	194832	18586	0
1.00E-05	156195	7546	0

表 3.2　随着类别的增加 between-group 类型肿瘤特异 DNA 甲基化位点数量逐渐减少（续）

α	L2B1	T4B1	T14B1
1.00E-10	98921	1251	0
1.00E-20	44521	28	0
1.00E-30	21100	1	0
1.00E-40	9551	0	0
1.00E-50	3764	0	0
1.00E-60	1191	0	0
1.00E-70	322	0	0
1.00E-80	77	0	0
1.00E-90	24	0	0
1.00E-100	4	0	0

对于 one-vs-rest 类型的 DNA 甲基化位点的识别，在 L2B1 数据集上，设定先验条件针对当前某类特异性的特征和其他类别之间的 β 值平均值只差不小于 0.1，通过调整 α_1 和 α_2 参数，得到特异性 DNA 甲基化位点的数量如表3.3所示，表头中的类别表示针对某类癌症（或健康具有特异性）：

表 3.3　T2B1 数据集上，调整 α_1 和 α_2 为不同的值过滤得到的 one-vs-rest 类型的类别 DNA 甲基化位点个数

1	α_1	α_2	健康	肺腺癌	肺鳞癌
2	0.01	0.9	1856	140	84
3	0.01	0.85	2773	204	125
4	0.01	0.8	3672	287	169
5	0.01	0.1	21358	1750	1217
6	0.01	0.01	32335	2649	1855
7	0.01	0.05	25171	2038	1452
8	0.01	0.001	39990	3304	2377
9	0.05	0.9	1856	140	84
10	0.05	0.85	2773	204	125
11	0.05	0.8	3672	287	169
12	0.05	0.1	21358	1750	1217
13	0.05	0.01	32335	2649	1855

表 3.3　T2B1 数据集上，调整 α_1 和 α_2 为不同的值过滤得到的 one-vs-rest 类型的类别 DNA 甲基化位点个数（续）

1	α_1	α_2	健康	肺腺癌	肺鳞癌
14	0.05	0.05	25171	2038	1452
15	0.05	0.001	39990	3304	2377

在 L2B1 数据集上，满足先验知识 $|\Delta\beta| = \left|\bar{\beta}_{current} - \bar{\beta}_{Rest}\right| > 0.1$ 的前提下，控制 $\alpha_1 = 0.01, \alpha_2 = 0.9$ 计算得到的针对健康、肺腺癌、肺鳞癌类型标志物数量分别为 1856、140 和 84 个。然后从中随机采样进行可视化，得到的 DNA 甲基化位点的分布结果如图 3.8所示，可以看出选取的位点符合将当前类和其他类进行区分的预期。

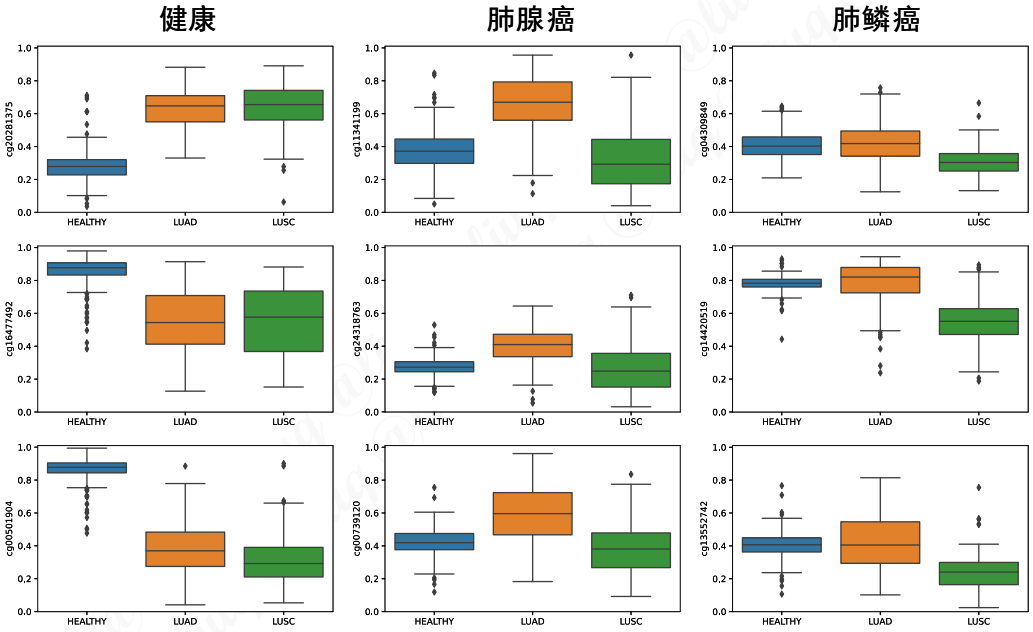


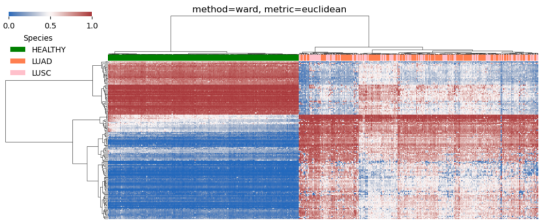
图 3.8　在 L2B1 数据集上随机采样得到的 one-vs-rest 类型 DNA 甲基化位点分布箱线图

从每一类 one-vs-rest 类型 DNA 甲基化位点随机挑选 K 个，使用 K 均值算法进行聚类，聚类簇数量设置为与真实类别数量相等。进而在测试集上预测的簇标记通过 Kuhn-Munkres 匹配算法将其和真实类标记对应起来，从而得到分类结果。如表3.4是 K 取 10 得到的测试集上的分类结果，初步验证本文过滤得到的 DNA 甲基化标志物对于类别预测具有较好的效果。

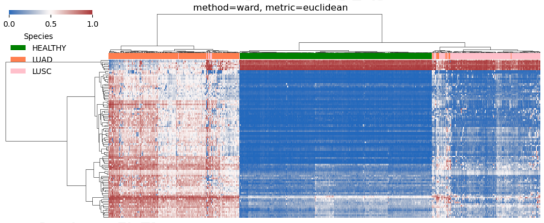
表 3.4　采用随机选取的 OvR 特征聚类后簇标号匹配得到的分类预测结果

	# 样本	灵敏度	特异度	ACC	Precision	F1	ACC (macro)
健康	262	1.000	0.966		0.960	0.979	
肺腺癌	182	0.863	0.949	0.907	0.882	0.872	0.886
肺鳞癌	146	0.795	0.948		0.835	0.814	

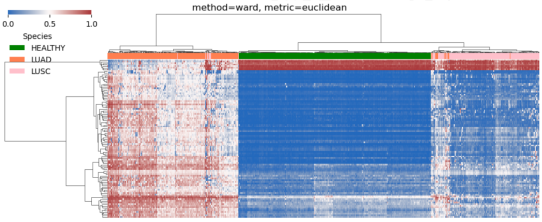
将每类肿瘤特异性 DNA 甲基化位点进行合并，再次进行层次聚类，如图 3.9所示，聚类结果也显示出来明显的三组模式。对其中的 DNA 甲基化位点进行采样，发现能够过滤出符合预期的 one-vs-rest 类型肿瘤特异性 DNA 甲基化位点，表现为在当前类和其他类之间分布差异大，而其他类别之间差异程度较小。



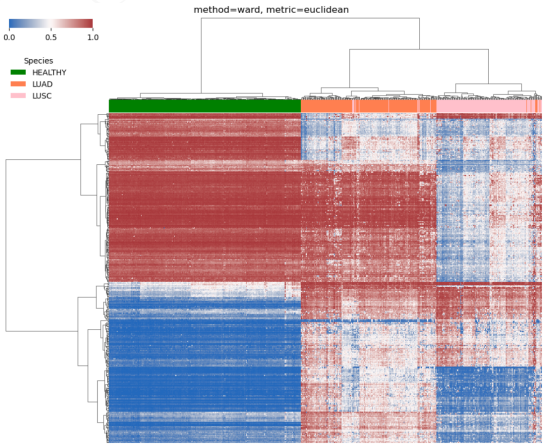
(a) 针对第 1 类（健康）



(b) 针对第 2 类（肺腺癌）



(c) 针对第 3 类（肺鳞癌）



(d) 三类 OvR 类型特征进行合并后的层次聚类结果

图 3.9　在 L2B1 数据集上肿瘤特异性 DNA 甲基化位点特征矩阵层次聚类结果

3.3.2.2　PanSeer DNA 甲基化数据集上实验结果

由上一章可知，健康人血液来源的样本和三类癌症组织来源的样本数分别为 248 和 120（肺癌、胃癌、结肠癌的数量均为 40）。在每个 DNA 甲基化标志物 j 上，每两类样本集合之间进行 Welch't t-test，总共进行 $C_4^2 = 6$ 次比较的结果，随后基于 Welch't t-test 的差异分析结果使用的方法来识别两类 DNA 甲基化位点。

通过设置参数 α_1 和 α_2 来筛选 DNA 甲基化标志物的数量，如图 3.10是在不同参数下识别得到的各类 DNA 甲基化标志物的数量变化图，每个散点图中横轴和纵轴分别表示参数 α_1 和 α_2 设置为不同梯度时的负对数值，图中圆的大小表示识别的对应类别的 DNA 甲基化标志物的个数。

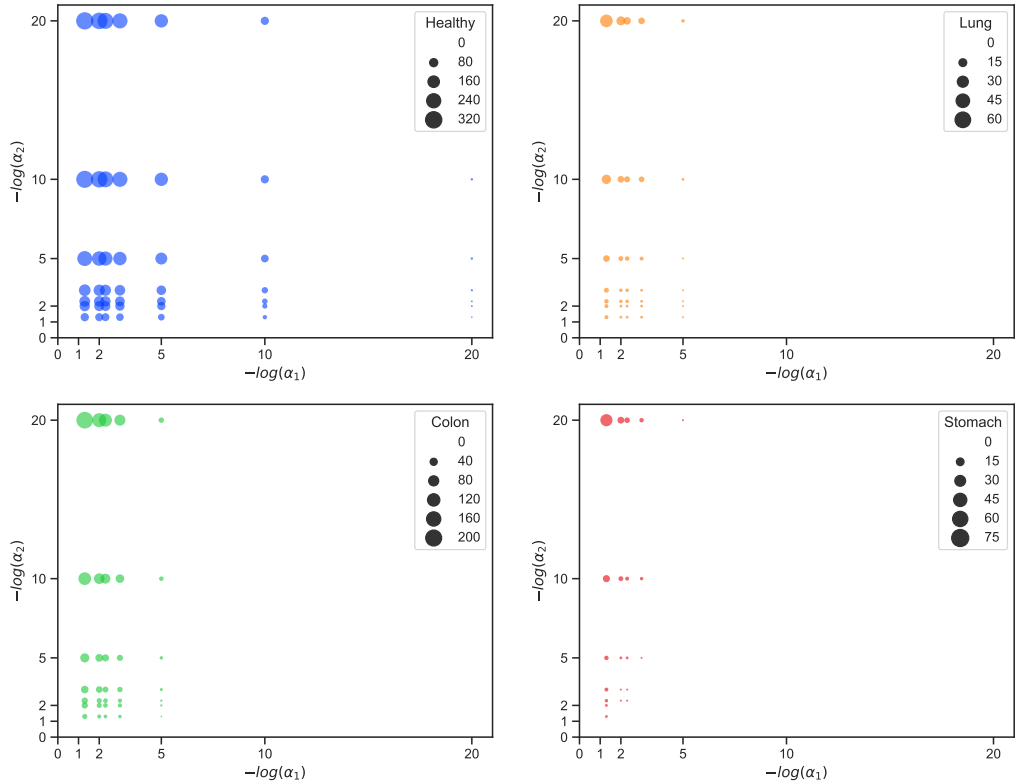


图 3.10　在 PanSeer 三类常见癌症数据集上 OvR 类型 DNA 甲基化标志物数量随参数变化图

DNA 甲基化标志物个数随着 α_1 值的减小即限制当前类和其他类的差异越显著而减少。保持 α_1 不变，随 α_2 的减小即其余类别之间的差异性可以很大甚至几乎不做受限制，DNA 甲基化标志物个数逐渐增多。同一参数下，不同类别特异性 DNA 甲基化位点的数量不同，例如参数设置为 $\alpha_1 = 0.001, \alpha_2 = 0.001$，针对健康、肺癌、结肠癌、胃癌的 DNA 甲基化标志物个数分别为 279，17，104，10 个。

如图 3.11所示是 PanSeer 三类常见癌症数据集上，采样识别得到的 OvR 类型 DNA 甲基化标志物特征分布箱线图，与之对应的在 TSH 液体活检数据训练集上的结果如图 3.12(a)。结果显示通过在健康人血液和肿瘤组织样本来源的 DNA 甲基化标志物，在实际的血液数据上特征可能不那么明显，这可能和样本的数量和样本来源和相应的处理方式不同有关。

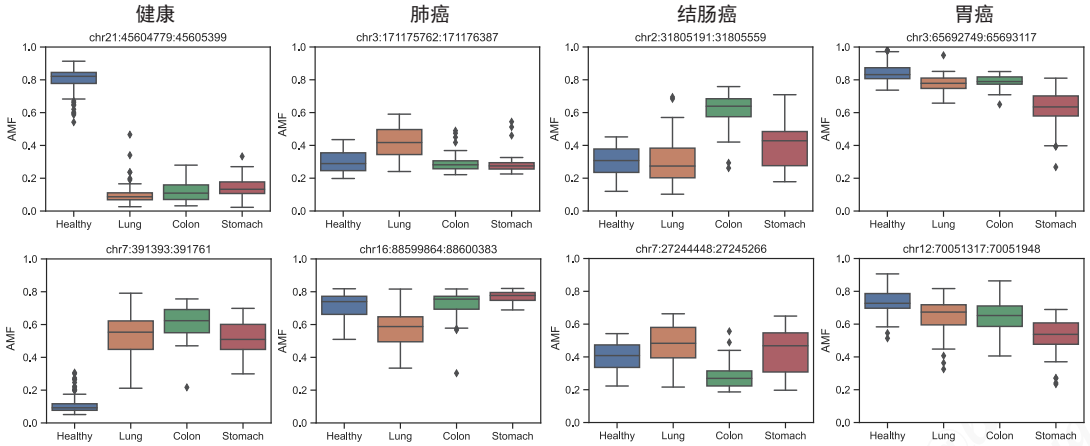
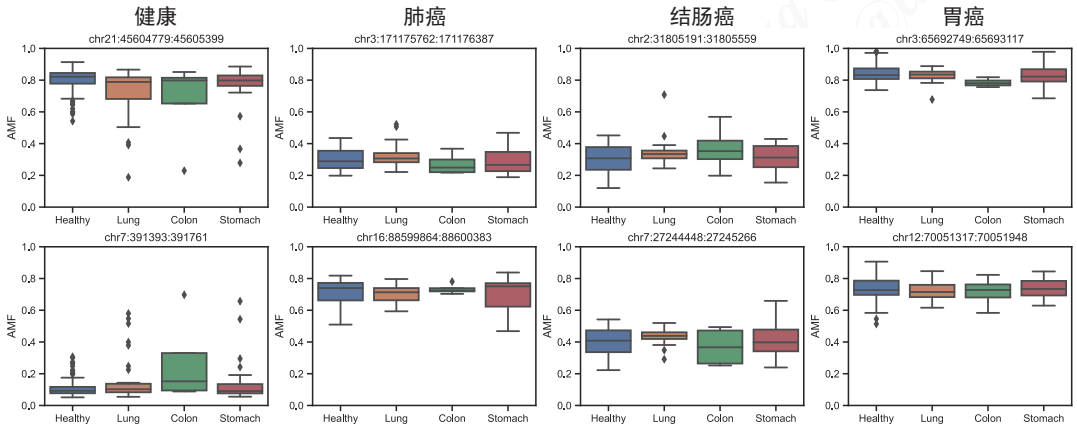
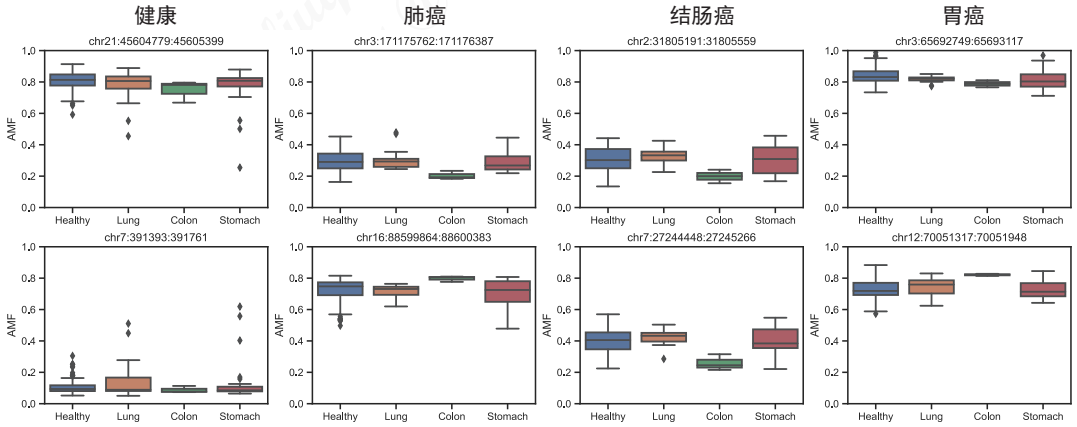


图 3.11　在 PanSeer 三类常见癌症组织和健康人血液来源 DNA 甲基化数据集上识别出的 DNA 甲基化标志物随机采样特征分布箱线图



(a) 训练集上 DNA 甲基化标志物特征分布箱线图



(b) 测试集上 DNA 甲基化标志物特征分布箱线图

图 3.12　在 PanSeer 三类常见癌症患者和健康人血液来源 cfDNA 甲基化数据集上识别出的 DNA 甲基化标志物随机采样特征分布箱线图

3.4　本章小结

本章提出了基于类别特异性过滤的 DNA 甲基化位点识别方法，并在不同的数据集上开展了不同参数设置对过滤结果影响的实验，在对应的测试集上对过滤得到的 DNA 甲基化位点进行特征可视化和层次聚类的实验。实验结果表明，本章的方法在能够识别出某一类肿瘤相对于其他类别存在显著差异的肿瘤特异性 DNA 甲基化位点。

第 4 章　DNA 甲基化位点的肿瘤特异性衡量和组织来源预测

4.1　本章引言

从系统生物学的角度来看，借助统计推断、机器学习、网络构建等方法，可以获得多种类型和层次的分子标志物，可以是单个标志物，例如对于诊断最有帮助的变异基因，也可以是多标志物之间构成的边（edge）、基序（motif）或模块（module），例如差异相关成对出现的差异表达基因[58]，或是多标志物相互关联构成的静态网络，和随着时间变化的动态网络标志物，如图 4.1所示。

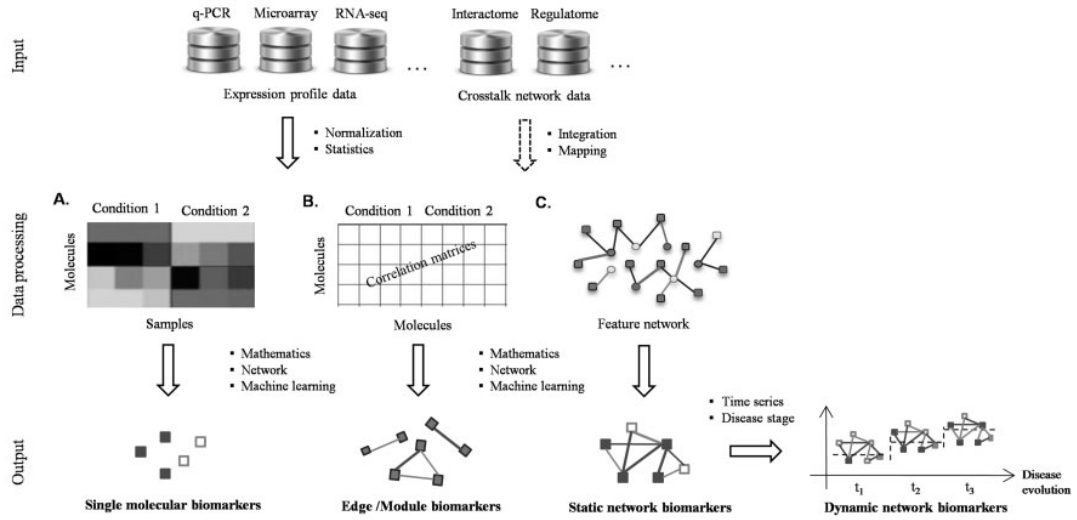


图 4.1　分子标志物在系统生物学上的类型和层次[59]

前一章本文提出了基于类别特异性过滤的 DNA 甲基化位点识别方法，该方法在类别较少时能够识别出符合预期的类型的 DNA 甲基化位点，当类别增多时，在满足统计显著性约束条件（例如 $\alpha < 0.05$）的前提下可能无法识别到足够多的肿瘤特异性 DNA 甲基化位点。当类别较少时，在相同的统计显著性约束条件下过滤得到的标志物数量可能又太多，对于统计显著性水平均很高（表现为 p 值都很小且接近于 0）的 DNA 甲基化位点，无法较好地评判特征之间的相对重要程度。本章针对上述问题对前一章的方法进行改进，对过滤得到的肿瘤特异性 DNA 甲基化位点进行特异性衡量，并采用有监督分类模型进行肿瘤组织来源预测，进而对预测性能进行评估。

4.2　基于统计显著性水平和互信息的位点特异性衡量方法

4.2.1　方法总体框架

如图 4.2所示是本章方法的总体框架，对于采用基于类别特异性过滤的 DNA 甲基化位点识别方法得到的肿瘤特异性 DNA 甲基化位点，本章基于统计显著性水平和互信息对每类过滤出的 DNA 甲基化位点进行打分和排序，随后选择每类 DNA 甲基化位点中排名靠前的前 K 个 DNA 甲基化位点作为特征，输入到有监督多分类模型中训练，最终进行用训练好的模型对测试集样本进行组织来源预测和性能评估。

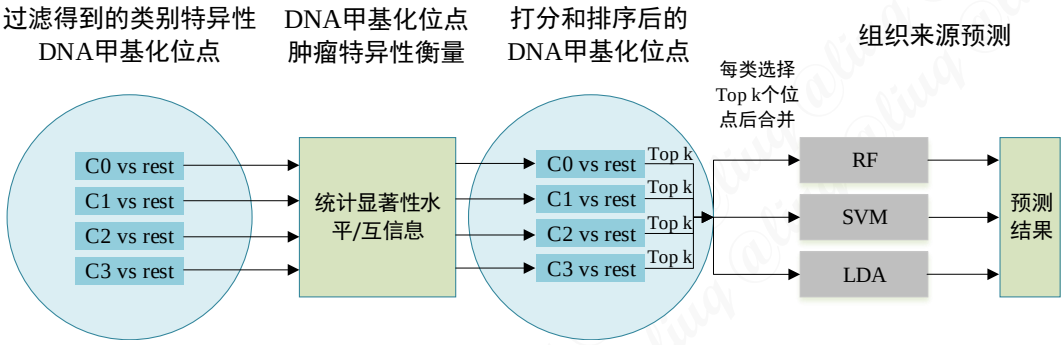


图 4.2　DNA 甲基化位点的肿瘤特异性衡量和组织来源预测方法总体框架

4.2.2　方法原理

本文设计了几种用于肿瘤特异性 DNA 甲基化位点特异性衡量的特征打分和排序规则，名称和含义如表4.2所示。

表 4.1　DNA 甲基化位点特异性衡量的特征打分和排序规则

缩写	名称和含义
MI_ovr	当前类和其他类特征的互信息
Chi2_ovr	当前类和其他类的卡方检验统计量
ANOVA_ovr	当前类和其他类特征的方差分析 F 统计量
adjP_ovr	当前类和其他类特征的矫正后 p 值的均值
P_B2T	当前癌症同血液差异分析矫正后的 p 值

互信息（Mutual information，MI）是一种基于信息论用来衡量两个随机变量的相关程度的统计度量，在本研究中被用来进行肿瘤特异性 DNA 甲基化位点的打分和选择。给定两个随机变量 X 和 Y，它们都是离散的，计算公式如下：

$$I(X,Y)=\sum_{x\epsilon X}\sum_{y\epsilon Y}p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \tag{4.1}$$

其中 $p(x)$ 和 $p(y)$ 分别是 X 和 Y 的边缘概率密度函数，$p(x,y)$ 表示 X 和 Y 的联合概率密度函数。对于 X 和 Y 都是连续的情况下，通常的做法是将连续的变量离散化到一些列的分组（bins）中，随后在离散条件下来估计 MI，然而要较好得估计 MI 需要足够多的分组数，但这样做会损失分辨率[60]。Kraskov A 等人的研究[61]提出一种考虑数据点之间间距和最近邻的方法来计算都是连续数据条件下的 MI。

针对 DNA 甲基化数据的特征选择而言，特征 β 值都是连续的，而样本对应的是否患有癌症和癌症类型信息是离散的（discrete），更狭义上讲是分类型（categorical）变量。为了解决这一问题，本章引入改进的 MI 估计方法[62]，并将计算得到的 MI 作为特征打分、排序和选择的依据，下面对该方法原理进行概述。

设一个连续型随机变量 X，和离散型随机变量 Y，对于每个样本点 $P_i=(x_i,y_i)$，基于其在连续变量 x 上的最近邻来计算一个数 I_i。首先在点 i 附近离散值都为 y_i 的数据点集合 N_{y_i} 中。基于某些特定的距离度量找到 K 个最近邻，设到第 K 个最近邻的距离为 d，随后在整个数据集中统计在距离 d 内中所有点的个数（包含第 K 个最近邻在内）记为 m_i，基于 N_{y_i} 和 m_i 可以计算出分数 I_i：

$$I_i=\psi(N)-\psi\left(N_{y_i}\right)+\psi(k)-\psi\left(m_i\right) \tag{4.2}$$

其中 $\psi(\cdot)$ 表示双伽玛（digamma）函数，基于公式(4.2)在所有点上对 I_i 取平均，得到结果：

$$I(X,Y)=\langle I_i\rangle=\psi(N)-\left\langle\psi\left(N_x\right)\right\rangle+\psi(k)-\langle\psi(m)\rangle \tag{4.3}$$

其中 $\langle\cdot\rangle$ 表示平均值，K 的需要指定大小，K 越大采样误差就越小。具体在计算针对某类 C_i 的肿瘤特异性 DNA 甲基化位点的 MI_ovr 时，会将其他类别标记统一置为 Rest，当前类别标记保持不变，从而进行运算并将得到的 MI 作为肿瘤特异性 DNA 甲基化位点衡量。

F 统计量也叫方差分析（ANOVA），可以针对某一具有多种水平的因素进行方差检验，在本研究中也可以用来对 DNA 甲基化位点进行打分，具体而言对于来自于 M 类样本的某个 DNA 甲基化位点，本章首先建立假设：

$$H_0：\mu_1=\mu_2=\cdots=\mu_M \tag{4.4}$$

$$H_1：\mu_1,\mu_2,\cdots\mu_M\text{不全相等} \tag{4.5}$$

随后根据表4.2，依次计算总体方差的组间估计和组内方差估计，最后计算出 F 统计量，和对应的 p-value。其中 $\bar{x}$ 表示任意一个 DNA 甲基化位点特征向量的均值。在计算 ANOVA_ovr 时，本文同样将其他类的标记设置为 Rest，而当前类保持不变，因此总的类别数 M 即为 2。

表 4.2　F 统计量计算方法

方差来源	平方和	自由度	均方	F 统计量
组内	$SSW=\sum_{j=1}^{M}\sum_{k=1}^{N_j}\left(x_{jk}-\bar{x}_j\right)^2$	$df_{\mathrm{w}}=M-1$	$MSW=\frac{SSW}{df_{\mathrm{w}}}$	$F=\frac{MSB}{MSW}$
组间	$SSB=\sum_{j=1}^{M}(\bar{x}_j-\bar{x})^2$	$df_{\mathrm{b}}=N-M$	$MSW=\frac{SSB}{df_b}$	
总和	$SST=\sum_{j=1}^{n}\left(\bar{x}_j-\bar{x}\right)^2$	$df_{\mathrm{t}}=\mathrm{N}-1$		

卡方检验是常用来检验两个变量是否相关的一种假设检验方法，其中零假设 H_0 用于检验的变量之间不具有相关性，备择假设则和 H_0 相反。本章将卡方检验统计量用来衡量 DNA 甲基化位点特征同类别标记的相关性，作为 DNA 甲基化位点的肿瘤特异性打分和排序规则之一。对于每个针对当前类 C_k 的 DNA 甲基化位点 f_i 的特征向量 u_i，本章在计算 Chi2_ovr 时，将当前类保持不变，其余类统一为 Rest 标记，并进行 u_i 和标记向量 y 之间的 Chi2 检验，按照公式(4.6)得到 Chi2 统计量。

$$\chi^2=\sum\frac{(O-E)^2}{E} \tag{4.6}$$

其中 O 和 E 分别表示真实样本集取不同值时的观测值和期望，计算方法如表4.3所示。由于本文 one-vs-rest 的处理方式，所以针对每种肿瘤特异性 DNA 甲基化位点而言，每次进行 Chi2 检验真实标记数均为 2，对应的 Chi2 检验自由度为 1。

表 4.3　对 DNA 甲基化位点的卡方检验打分计算

	f_i	
类别	O	E
One	$\sum_{i=1,y_i=\mathrm{One}}^{N_{\mathrm{One}}}x_i$	$\frac{N_{\mathrm{One}}}{N}\sum_{i=1}^{N}x_i$
Rest	$\sum_{i=1,y_i=\mathrm{Rest}}^{N_{\mathrm{Rest}}}x_i$	$\frac{N_{\mathrm{Rest}}}{N}\sum_{i=1}^{N}x_i$

本章从原样本集中选择经过打分和排序后的肿瘤特异性 DNA 甲基化位点作为特征的子集，采用六种广泛使用的机器学习模型作为候选分类器。为了比较和检验不同分类器的预测性能，在训练集上采用 10 折交叉验证，从中挑选较好的分类器作为肿瘤来源预测任务的模型。随后在整个训练集上进行训练，并在测试集上测试肿瘤来源预测模型的分类性能，计算出相应的预测指标。

4.3　实验设计与结果分析

4.3.1　实验设计

现有的基于 DNA 甲基化来进行标志物选择方法没有统一的规范[41]，根据前一章提到的相关工作，本章总结了现有的部分打分规则，如下表所示：

表 4.4　现有常用的肿瘤特异性 DNA 标志物挑选准则总结

方法	解释	来源
$MAD = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - m(X)\rvert$	平均绝对偏差（Mean absolute deviation，MAD），其中 $m(X)$ 表示中心偏离测度，通常取均值。	[63]
$SPEC$	Spectral feature selection 谱特征选择[64]。一种无监督的特征选择方法。	[63]
$MR = \max\{\bar{x}_j\} - \min\{\bar{x}_j\} \geqslant t$	肿瘤甲基化程度。对某个 DNA 甲基化区域而言其特征平均值的最大值减去最小值。	[25]
(1) $FDR(q-value) < 0.01$ (2) $\lvert\Delta\beta\rvert = \lvert\bar{\beta}_T - \bar{\beta}_B\rvert > 0.25$	假设检验（Limma 模型）统计显著性	[27]
(1) $p-value < 0.05$ (2) $Top\ 1000$	假设检验（Moderated t-statistics）统计显著性	[28],[29]
$(a)\ \lvert\bar{x}_j - \bar{x}\rvert \geqslant 3SD, j \in \{1,2,\cdots N\}$ $(b)\ SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{n-1}}$ $(A)\ \frac{\max\{\bar{x}_j\} - \min\{\bar{x}_j\}}{\min\{\bar{x}_j\}} > 20\%$ $(B)\ CV = \frac{SD}{\bar{x}} > 0.25$	作者定义的两类 DNA 甲基计划位点，其中类型 I 满足条件（a）、（b）；类型 II 满足条件（A）、（B）	[24]

本文对 DNA 甲基化位点肿瘤特异性衡量方法所得到的排名靠前的 DNA 甲基化位点与现有方法挑选出的 DNA 甲基化位点进行分析，并通过有监督分类模型进行肿瘤组织来源预测性能评估和比较。为了选择合适组织来源预测模型分类模型，本章使用了六种机器学习模型，分别为 KNN（K 近邻）、SVM（支持向量机分类模型）、RF（随机森林）、CART（决策树）、NB（朴素贝叶斯）、LDA（Fisher 线性判别），采用 scikit-learn 实现，均选择默认参数。本章放宽前一张章中的过滤条件，只要求当前类和其他类存在显著差异，而不要求其他类别来自同一分布，即

α_2 为近似于 0 的数值。对过滤出的 DNA 甲基化位点采用前文所述的打分规则进行打分，按照得分从高往低排序，选择得分靠前的 DNA 甲基化位点，最终用于肿瘤的组织来源预测。

4.3.1.1　评价指标

混淆矩阵是常见用于已知样本真实标记和预测标记来评价预测结果的可视化方法，如图 4.3所示，对于类别总数为 M 的混淆矩阵令其为 $CM = (c_{i,j})$，矩阵的每一列代表一个类的实例预测，而每一行表示一个样本的实际类别标记，其中的每个元素 $c_{i,j}$ 表示实际为 i 类的样本被预测为 j 类的数量。对于任意某个类别 k 而言，样本的预测结果有四种情况：

真阳性（True Positive，TP）：被正确判断的阳性样本，样本数为 $c_{k,k}$

真阴性（True Negative，TN）：被正确判断的阴性样本，数量为 $\sum_{i\neq k,j\neq k} c_{i,j}$

假阳性（False Positive，FP）：预测的阳性样本中属于是误判的，数量为 $\sum_{i\neq k,j=k} c_{i,j}$

假阴性（False Negative，FN）：预测的阴性样本中属于是误判的，数量为 $\sum_{i=k,j\neq k} c_{i,j}$

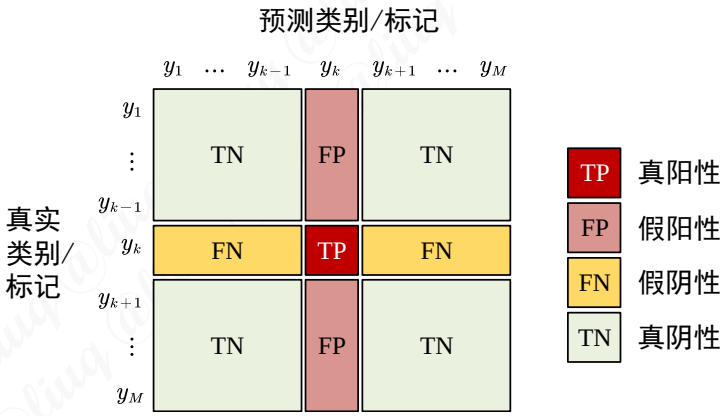


图 4.3　混淆矩阵示意图

对于每个单独的类别 k，采用的评价指标如下：

召回率，也叫灵敏度，衡量对于某个真实的类预测的好坏，例如实际患有某类癌症的样本中多少比例被正确预测的，计算公式如(4.7)。

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{4.7}$$

精确度，用来衡量对于预判的某个给定的类别 k，其中多少比例是被正确预测的，公式如(4.8)：

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \tag{4.8}$$

F1 指标，是召回率和精确度的调和平均值，用于衡量预测结果具有较好的鲁棒性，公式如(4.9)：

$$\text{F1}_k = \frac{2 * TP_k}{2 * TP_k + FN_k + FP_k} \tag{4.9}$$

马修斯相关性系数（Matthews Correlation Coefficient，MCC），计算公式如(4.10)：

$$\text{MCC}_k = \frac{TP_k * TN_k - FP_k * FN_k}{\sqrt{(TP_k + FP_k)(TP_k + FN_k)(TN_k + FP_k)(TN_k + FN_k)}} \tag{4.10}$$

对于总体的预测结果，评价指标采用总体准确率，公式如(4.11)：

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^{N} c_{i,i}}{\sum_{i=1}^{N} \sum_{i=1}^{N} c_{i,j}} \tag{4.11}$$

对于多类也可以计算其召回率，主要有 macro 和 micro 两种方式，分别为公式(4.13)和公式(4.12)：

$$\frac{\text{Recall}_{macro} = \sum_{i=1}^{M} \text{Recall}_i}{M} \tag{4.12}$$

$$\text{Recall}_{micro} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} TP_i + \sum_{i=1}^{M} FN_i} \tag{4.13}$$

4.3.2　结果分析

4.3.2.1　组织来源预测模型选择结果

如图 4.4所示，在 T14B1 数据集上的训练集上，采用本文提出的肿瘤特异性 DNA 甲基化位点识别方法加上 adjPval_ovr 打分，在 T14B1 训练集上对多种组织来源预测模型进行 10 折交叉验证，得到的分类 Accuracy 性能随着选取 DNA 甲基化位点个数的变化图。具体而言纵轴表示分类 Accuracy，横轴表示 K（图中范围取 0 到 100），即按照每类和其他类的 p 值的均值打分各取前 K 个然后合并，总共的特征数为 $15 \times K$（总的特征数范围在 0 到 1500）个，分别绘制了六类分类模型上得到的组织来源预测结果。图中用六种不同的颜色和对应的六种不同形状的点以来区分六种不同的分类模型，图上每条曲线上的一个点表示对应的分类模型和 K 的条件下，十次随机十折交叉验证的在验证集上的平均结果。如图 4.4左侧可以看出，随着 K 值的增加，即所能利用的特征信息的增加，各个分类模型的分类性能

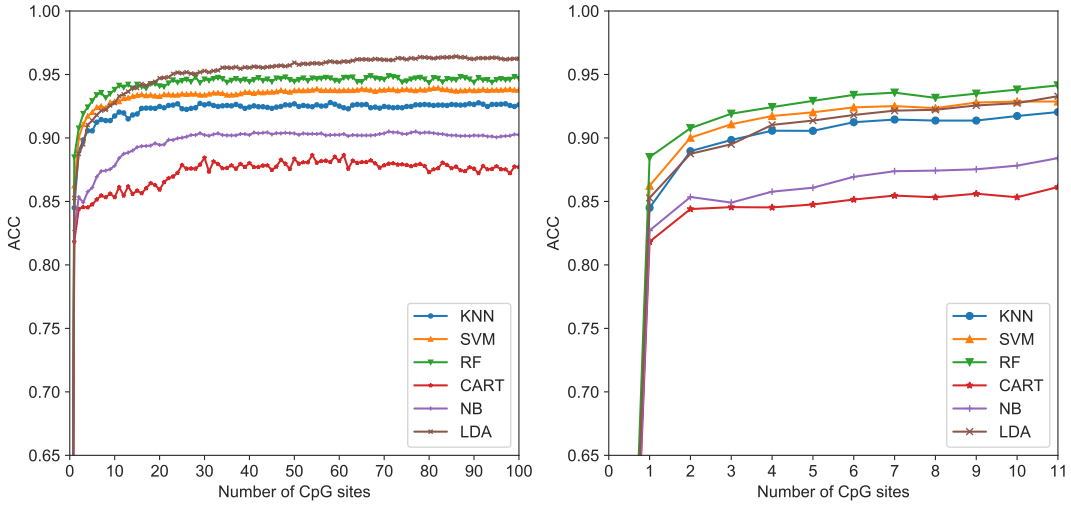


图 4.4　以 adjPval_ovr 打分选择的 DNA 甲基化位点在 T14B1 训练集上进行多种组织来源预测模型交叉验证的 ACC 结果比较

总体呈现上升趋势，当 K 较大时，例如超过 50（即每类选取打分前 50 个 DNA 甲基位点）时，各个分类模型的性能几乎区域稳定不再改变，其中 LAD，RF，SVM 分类模型性能的表现较好。当 K 较小时各个分类模型的分类性能变化趋势不尽相同，从起点和上升幅度来看，RF、SVM、LAD 可能是较好的模型类器。

类似的作为对比，在 T14B1 数据集上本章保持其他实验设置相同，只在特征选择上不采用本文提出的肿瘤特异性 DNA 甲基化位点识别方法，采用 MAD 作为打分规则，从高到低选择靠前的特征，如图 4.5所示，纵轴依然表示分类 Accuracy 性能，横轴依然表示 K 对应的选择的特征是 MAD 打分从高到低前 $15 \times K$ 个，从而保持和本章之前的实验选择特征数量上一致，图中不同颜色的曲线和点的含义和前文描述相同。

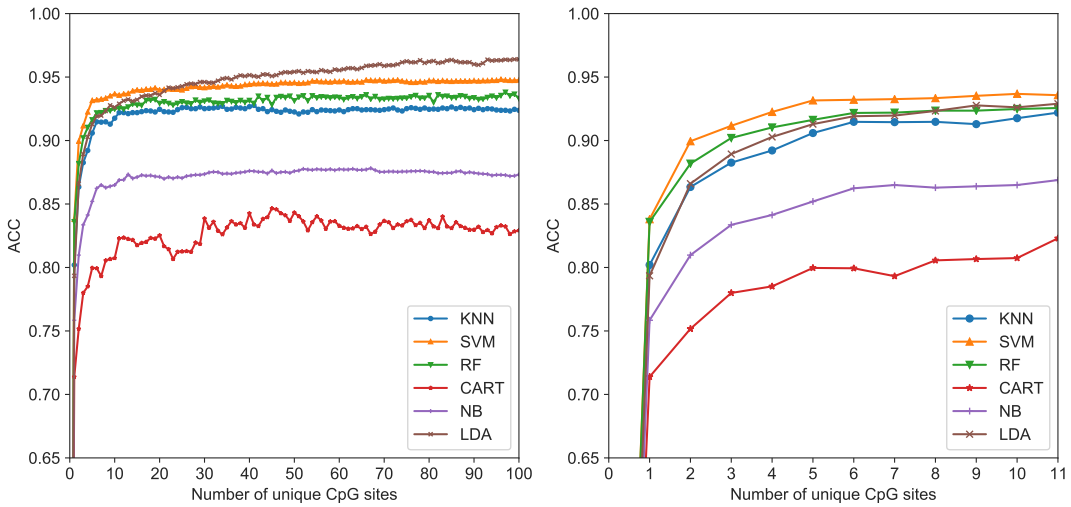


图 4.5　以 MAD 打分选择的 DNA 甲基化位点在 T14B1 训练集上进行多种组织来源预测模型交叉验证的 ACC 结果比较

采用 MAD 打分选择的 DNA 甲基化位点进行组织来源预测地分类结果显示，采用 SVM、RF、LDA 分类模型的分类 Accuracy 性能要优于 KNN、CART 和 NB。综合上述结果，本章采用较优的三类分类模型作为组织来源预测模型。

4.3.2.2　单个分类模型上特征打分规则比较

本章的标志物识别方法设计了四种特征打分规则。如图 4.6所示，纵轴表示使用随机森林分类模型在测试集上得到的 Accuracy 分类指标，越大表明组织来源预测结果越好，横轴表示对每一类而言按照特征打分规则取前 K 个。左图 K 的范围从 1 到 50，右图相较于左图的区别在于，右图采用的 K 范围缩小至 1 到 10。对于整个 T14B1 数据集，共 15 类，不考虑重复总的 DNA 甲基化位点，$K=1$ 即每类取 1 个 DNA 甲基化位点，则选择的 DNA 甲基化位点总数有 15 个，对应的有，当 $K=50$ 即每类取前 50 个 DNA 甲基化位点时，位点总数为 750 个。图中每一个点表示在对应 K 和采取对应的打分规则的条件下在测试集上均采用随机森林进行组织来源预测所得到的 Accuracy 结果，不同形状和颜色来区分本章采用四种不同的打分。

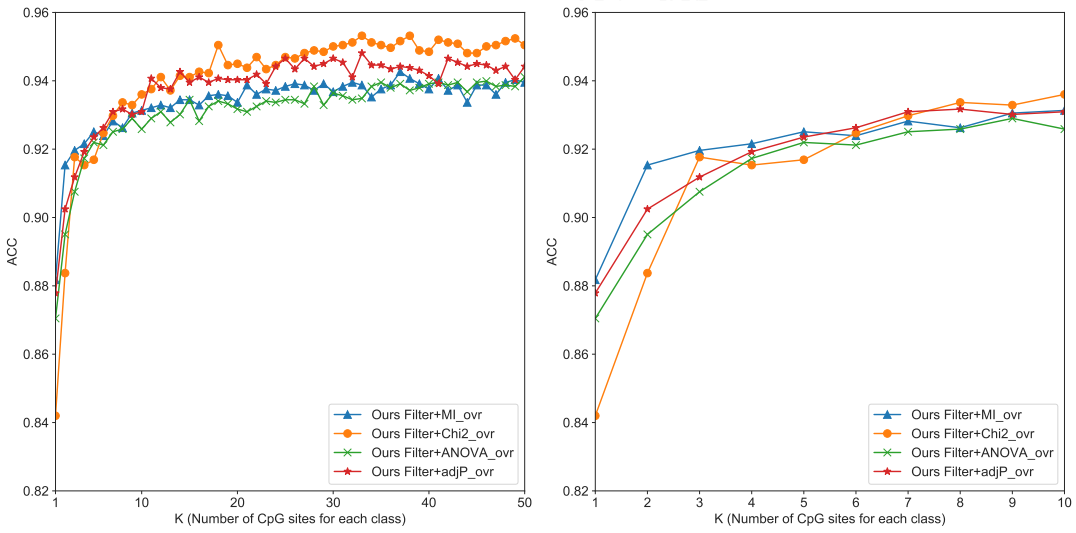


图 4.6　不同打分规则选择的 DNA 甲基化位点以 RF 为组织来源预测模型在 T14B1 测试集上的 ACC 结果比较

另外，本文希望使用尽量少的特征就能得到较好的分类性能。如图 4.6左侧所示，随着 K 值的增加，总体上的分类性能呈上升趋势。如图 4.6右侧所示，当 K 较小（$K<6$）的情况下，采用本章的方法并运用互信息作为特征打分规则，能够得到较好的分类效果。例如当 K 取 1 时，即每类取 1 个特征，T14B1 共十五类的测试集上的分类 Accuracy 就能够得到 0.88，$K=2$ 时 Accuracy 能够达得 0.915，较第二名的采用当前类和其他类的所有 p 值的均值作为打分得到的 0.902 高出 0.01。

对应到现实世界中，本章可以对很少量的特征进行实验验证，从而减少实验成本。如图 4.6左图所示，当 K 较大，即本章能够验证的特征较多的情况下，采用当前类和其余类进行卡方检验统计量作为打分规则在测试集上得到的 Accuracy 性能较高。根据实验结果，本章选择互信息和卡方检验统计量分为作为需要检验的 DNA 甲基化位点个数较少和较多情况下的打分规则。

本章还比较了采用互信息作为特征打分规则和现有 DNA 甲基化位点选择方法在测试集上组织来源预测结果。如图 4.7所示，左侧为采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合互信息作为特征打分方法，随机森林作为组织来源预测模型，在测试集上的组织来源预测结果随选取的 K 的变化图。图中红、绿、蓝色带星号虚线表示本章用来比较的方法，其中 MAD 直接在所有特征上进行打分。当本章的方法每类取前 K 个特征时，MAD 方法取分数在前 $15 \times K$ 个特征，即保持特征维数和本章采取的方法一致，SPEC 的处理同理。P_B2T 表示在每个特征上进行健康人血液来源样本同每类肿瘤组织进行比较，即本章的方法取前 K 个特征时，P_B2T 取 14 次比较，每次比较取前 K 个 p 值最小的特征，即总共 $14 \times K$ 个特征，相比本章的方法在相同的 K 取值的条件下，总的特征数要少 K 个。

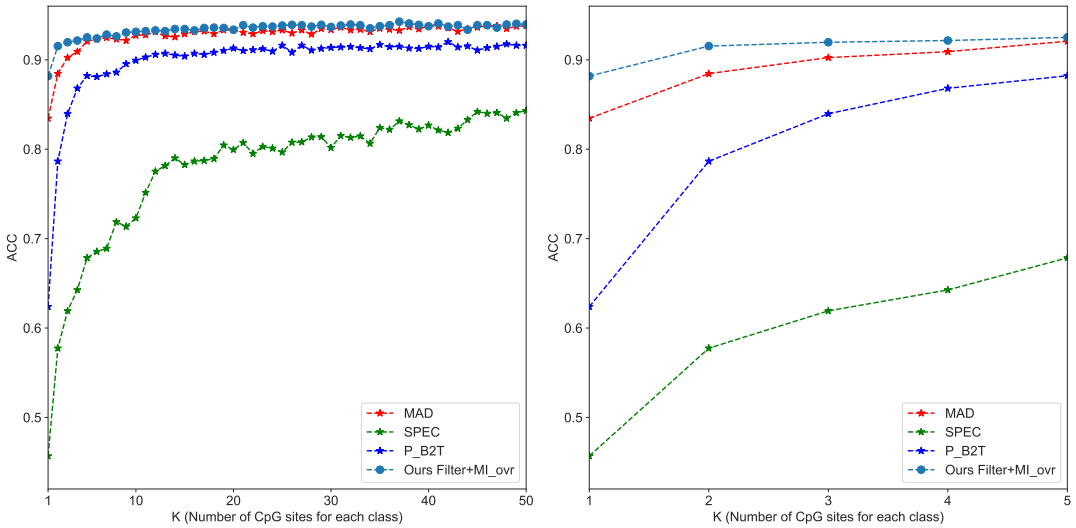


图 4.7　K 取值较小时本章处理方法与现有方法以 RF 为组织来源预测模型在 T14B1 测试集上的 ACC 结果比较

根据本章前文所述得出的结论，本章的方法结合互信息作为特征打分方法当 K 较小时也能够得到较好的分类性能。如图 4.7右图所示，是 K 小于 6 时采用相同的分类模型在测试集上采用本章的方法和打分规则同另外三种方法的比较结果，结果印证了本章此前的结论，即本章的方法分类性能优于另外三种方法。并且尤其是 K 取值为 1 时，本章的方法（0.88）同第二名（0.83）方法的相比要高出 5 个百分点，显示本章的方法能在尽可能少的特征数的前提下，实现较优的分类性能。

如图 4.8所示，左侧为本章的方法结合卡方检验作为特征打分规在 K 取 1 到 50 是的分类性能，根据从而说明当 K 较大即所能够验证的特征数量较多时，根据前文采用则能够取得较好的分类性能，右图是在 K 为 6 到 50 上的结果，显示出本章的方法在该范围内均优于另外三种方法。

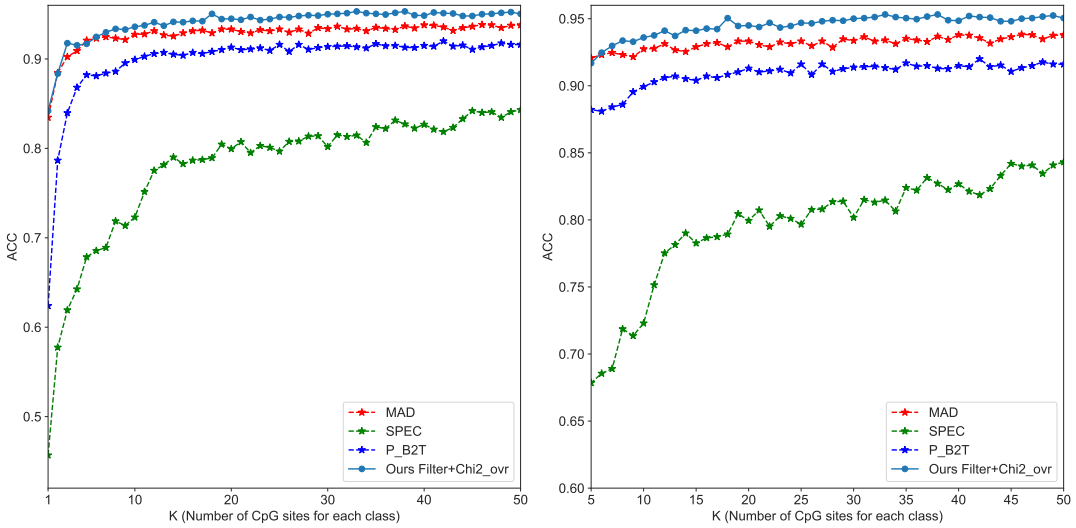


图 4.8　K 取值较大时本章处理方法与现有方法以 RF 为组织来源预测模型在 T14B1 测试集上的 ACC 结果比较

4.3.2.3　多个分类模型上特征打分规则比较

为了减小组织来源预测模型分类模型带来的分类性能差异，本章将在三种分类（RF、SVM、LDA）上效果较好的模型的结果取平均，作为最终的分类结果，来比较不同打分方法。如图 4.9所示，是采用本文提出的肿瘤特异性 DNA 甲基化位点识别方法和不同的特征打分规则，在测试集上得到的三种模型平均后的分类性能随着 K 的变化情况，其中纵轴表示采用前面提到的三种分类模型平均之后的 15 类（14 种癌症和健康）分类 Accuracy 性能，上方两图的纵轴方位均为 0.75 到 0.96 之间，下方两图纵轴范围 0.9 到 0.96 之间。横轴表示不同的 K 范围，表示针对每类选择前 K 个特征合并，图 4.9左上从 K 的范围从 1 到 50，即选择的 DNA 甲基化位点总数以总类别数（15）为间隔从 15 到 750 之间，另外三张图是左上图的局部放大，其中右上图的 K 的范围从 1 到 6，左下和右下图的方位 K 的范围分别从 6 到 20，和 20 到 50。图中以四种不同颜色表示采用本章过滤方法和三种模型平均的前提下采用的四种打分规则，图中每一个点均表示在测试集上使用三种不同的分类模型测试结果的平均值。

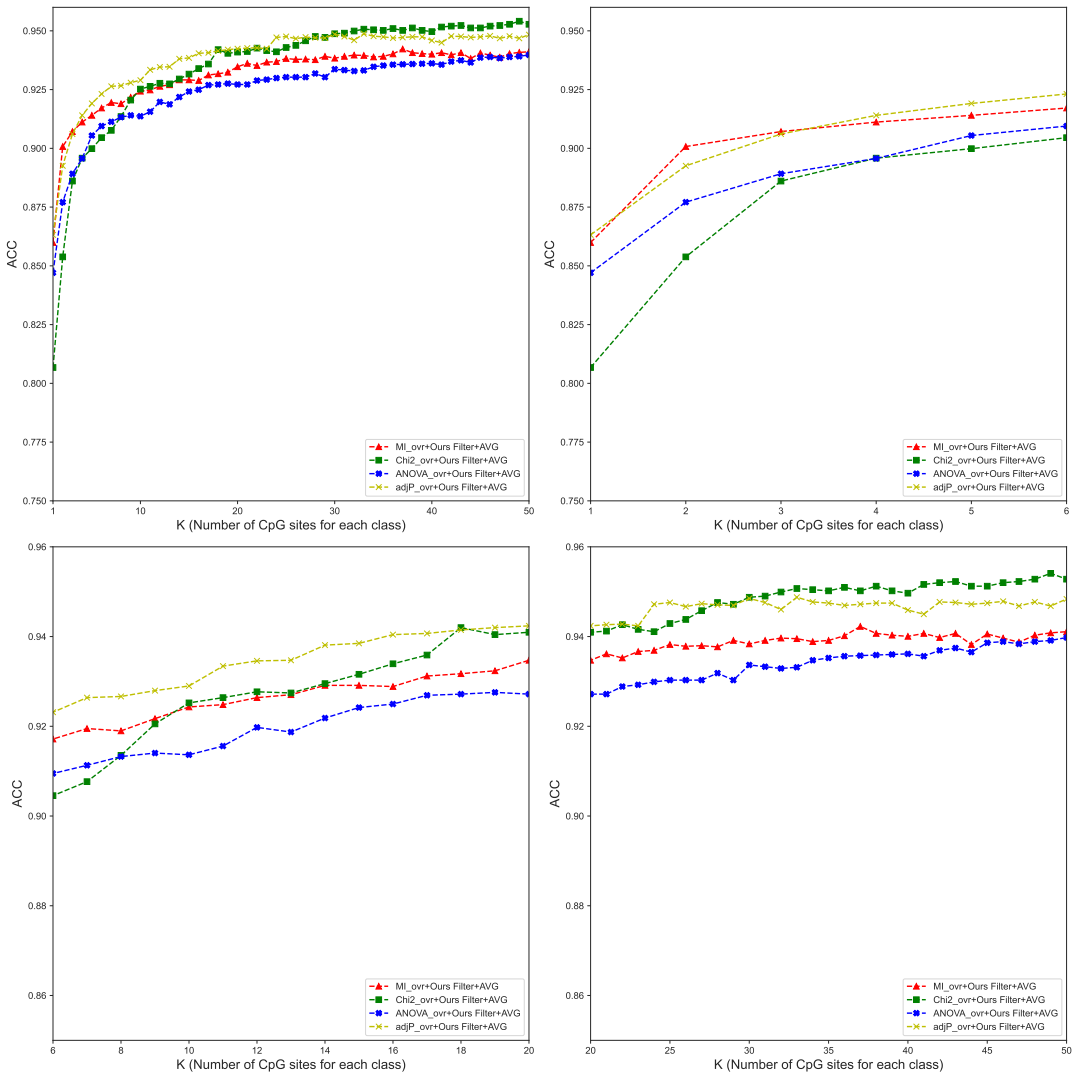


图 4.9　采用不同打分规则和三种分类模型在 T14B1 测试集上进行组织来源预测的平均 ACC 结果比较

通过观察图 4.9的起点和变化趋势，当 K 较小的情况下，采用 MI_ovr 和 adjP_ovr 特征打分得到的分类性能较优。当 K 较大时 adjP_ovr 也能够得到较好的分类性能，Chi2_ovr 特征打分曲线的变化幅度较大，前期性能较差后期和 adjP_ovr 特征打分方法得到的分类性能接近。考虑到本章希望采用尽可能少的特征数来达到较好的分类性能，得出采用 MI_ovr 和 adjP_ovr 这两种特征打分方式能够得到较好的预测性能。

本章将上述两种打分对应预测性能和现有的三种特征选择方法得到的预测性能进行了比较。如图 4.10是采用以 adjP_ovr 为特征打分标准和另外三种方法在测试集上进行比较的结果，其中纵轴和横轴的含义与之前相同，图中以不同的颜色表示所采用的特征打分方式类型，左图中每个点也表示在三种分类模型上的预测结果的平均值。可以看出本章的方法在测试集上得到的分类 Accuracy 性能同现有的

以 MAD 作为打分标准选择特征得到的分类性能接近，且明显优于另外两种（SPEC 特征打分和 P_B2T 的处理方式）特征打分方法所得到的结果。

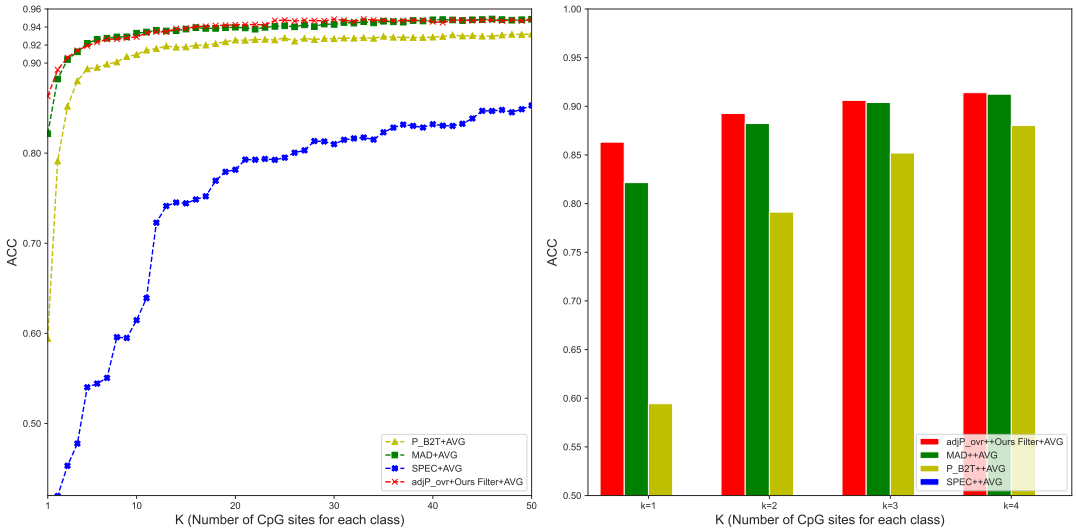


图 4.10　使用 adjP_ovr 和其他打分方法在 T14B1 测试集上三种分类模型进行组织来源预测的平均 ACC 结果比较

值得注意的是，当 K 的值很小时，即表示选取最优的特征时，如图 4.10右侧柱状图所示，采用本章的处理方式并以 adjP_ovr 特征打分的结果要优于另外现有的三种处理方法，例如当 K 取 1 时，本章的方法比第二好的 Accuracy 方法性能要高出 5%。如图 4.11所示，采用本章的处理方法结合 MI_ovr 特征打分的效果有也相同的结论。

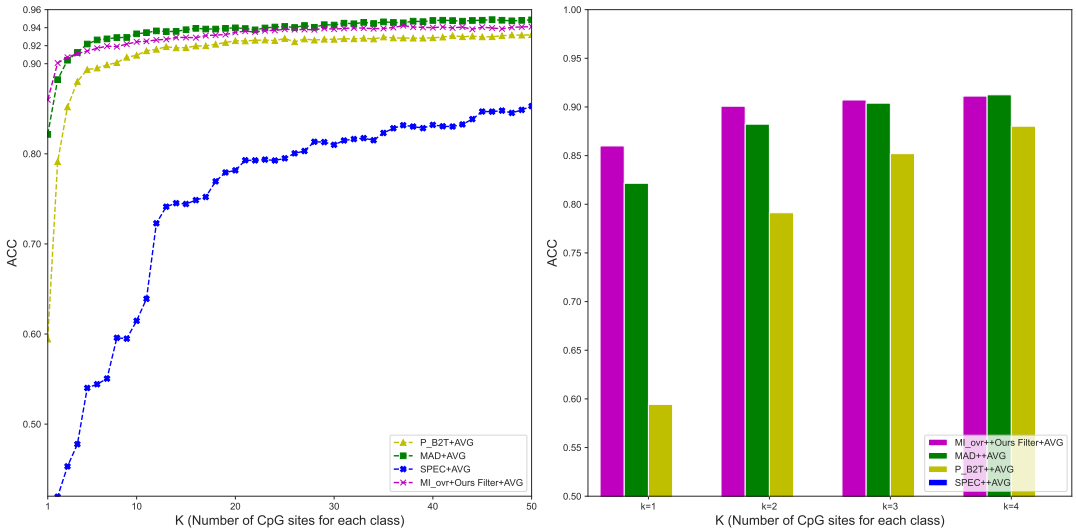
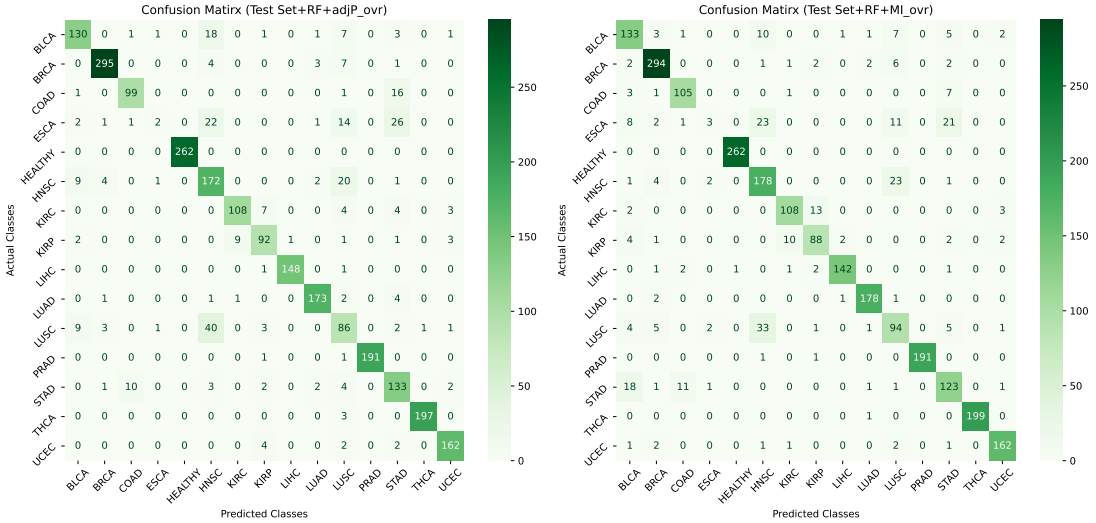
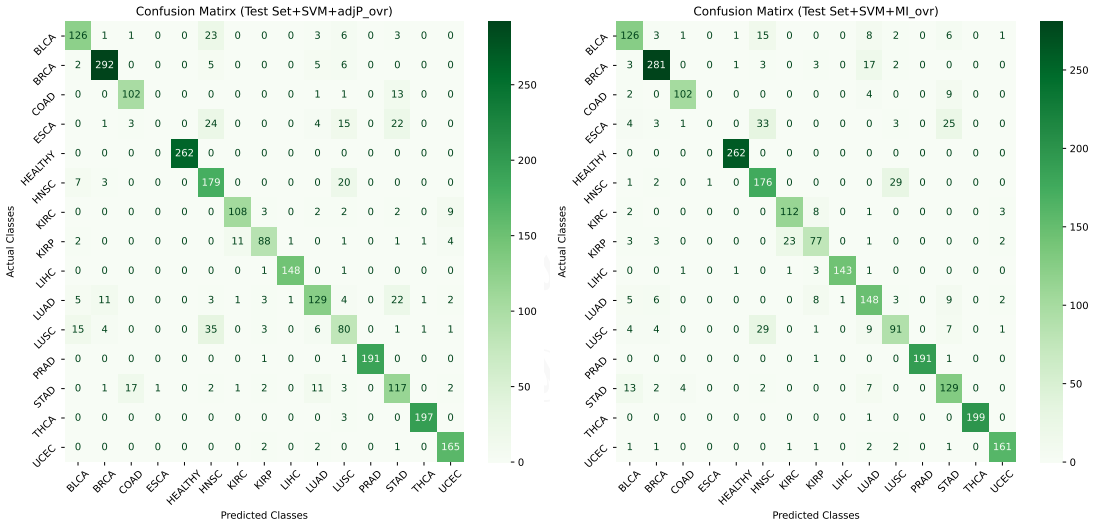


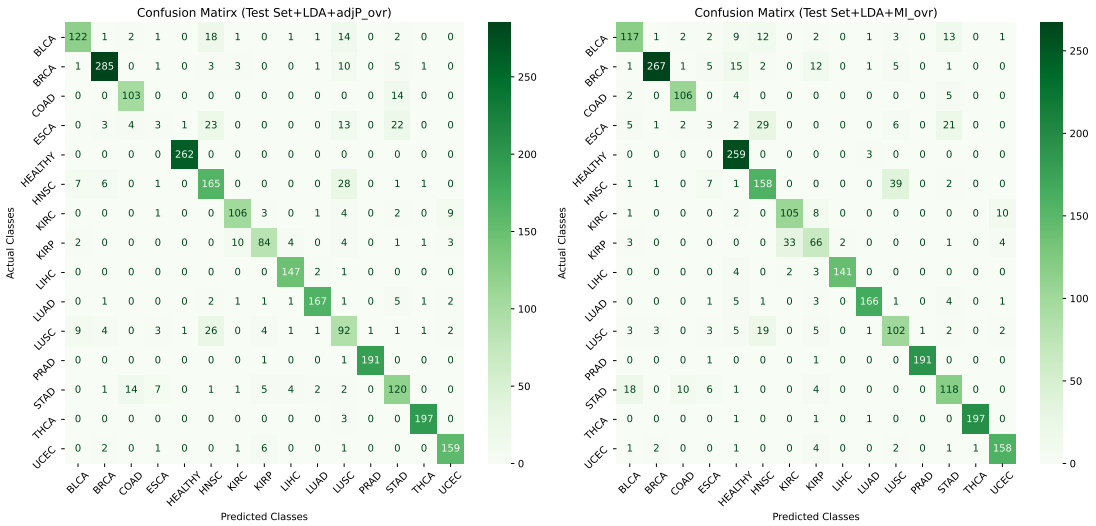
图 4.11　使用 MI_ovr 和其他打分方法在 T14B1 测试集上三种分类模型进行组织来源预测的平均 ACC 结果比较



(a) RF+adjP_ovr 预测结果　　(b) RF+MI_ovr



(c) SVM+adjP_ovr 预测结果　　(d) SVM+MI_ovr 预测结果



(e) LDA+adjP_ovr 预测结果　　(f) LDA+MI_ovr 预测结果

图 4.12　K=1 时在 T14B1 测试集上采用不同打分和三种组织来源预测模型得到的混淆矩阵结果

当 K 取 1 时，每类肿瘤和健康类别的组织来源预测结果混淆矩阵如图 4.12所示，根据混淆矩阵计算出评价指标，并用来衡量本文方法进行肿瘤诊断和肿瘤组织来源预测的效果。

如表 4.5所示，是使用 MI_ovr 进行评分，每类选择最优秀的 1 个特征（即 k=1，共 15 个特征）在测试集上使用 RF 进行肿瘤组织来源预测计算得到的结果，与之相比采用 MAD 评分选取前 15 个 DNA 甲基化位点并采用 RF 分类器进行肿瘤组织来源预测的结果如表4.6所示，结果表明本文的方法使用极少量的肿瘤特异性标志物就能够实现很高的诊断性能，和针对多种癌症在保证很高特异度的前提下，取得较好的诊断灵敏度。并且和现有方法比较，本文识别出的针对肿瘤类别的 DNA 甲基化位点，具有显著的肿瘤特异性。

表 4.5　T14B1 数据集上使用 MI_ovr 进行评分每类选择最优的 1 个特征在测试集上 RF 的分类性能

							总体		
类别	N	灵敏度	特异度	F1	精确率	MCC	ACC	灵敏度（Macro）	灵敏度（Micro）
BLCA	163	0.816	0.982	0.785	0.756	0.770			
BRCA	310	0.948	0.990	0.939	0.930	0.931			
COAD	117	0.897	0.994	0.886	0.875	0.881			
ESCA	69	0.043	0.998	0.078	0.375	0.120			
HEALTHY	262	1.000	1.000	0.998	0.996	0.998			
HNSC	209	0.852	0.971	0.781	0.721	0.763			
KIRC	126	0.857	0.994	0.871	0.885	0.865			
KIRP	109	0.807	0.992	0.815	0.822	0.807	0.882	0.834	0.882
LIHC	150	0.947	0.998	0.959	0.973	0.957			
LUAD	182	0.978	0.997	0.973	0.967	0.971			
LUSC	146	0.644	0.979	0.646	0.648	0.625			
PRAD	193	0.990	1.000	0.995	1.000	0.994			
STAD	157	0.783	0.981	0.757	0.732	0.741			
THCA	200	0.995	1.000	0.997	1.000	0.997			
UCEC	170	0.953	0.996	0.950	0.947	0.947			

表 4.6　使用 MAD 评分最优的 15 个特征在 T14B1 测试集上 RF 的分类性能

类别	N	灵敏度	特异度	F1	精确率	MCC	总体		
							ACC	灵敏度（Macro）	灵敏度（Micro）
BLCA	163	0.675	0.974	0.655	0.636	0.631			
BRCA	310	0.942	0.984	0.915	0.890	0.904			
COAD	117	0.838	0.990	0.817	0.797	0.808			
ESCA	69	0.043	0.997	0.075	0.273	0.100			
HEALTHY	262	1.000	1.000	1.000	1.000	1.000			
HNSC	209	0.804	0.970	0.752	0.706	0.730			
KIRC	126	0.841	0.992	0.845	0.848	0.837			
KIRP	109	0.743	0.992	0.771	0.802	0.762	0.835	0.784	0.835
LIHC	150	0.913	0.997	0.929	0.945	0.925			
LUAD	182	0.808	0.991	0.838	0.870	0.826			
LUSC	146	0.548	0.981	0.586	0.630	0.564			
PRAD	193	0.959	0.997	0.959	0.959	0.955			
STAD	157	0.726	0.971	0.669	0.620	0.648			
THCA	200	0.990	1.000	0.995	1.000	0.995			
UCEC	170	0.929	0.988	0.888	0.849	0.880			

4.3.2.4　标志物比较和生物学解释

通过前文在多模型上特征打分规则比较实验，采用本文的基于类别特异性过滤的 DNA 甲基化位点识别方法和 adjP_ovr 和 MI_ovr 打分方法选择的 DNA 甲基化位点，和用作对比的方法中最优的 MAD 作为特征打分规则选择得到的 DNA 甲基化位点，总体上在测试集上进行肿瘤组织来源预测的性能近似，并且在选取较少的 DNA 甲基化位点时，采用本章的处理方法能够得到较优的预测性能。

为了检验本章的方法识别出的标志物的特异性，对 K 为 1 情况对识别得到的标志物的分布进行可视化，如图 4.13所示是采用 MAD 作为打分方式取得的得分最高的前 15 个 DNA 甲基化位点在不同类别上的分布箱线图，左上角是 MAD 打分最高的 DNA 甲基化位点特征值对应在各类上的分布箱线图，从左到右从上到下 DNA 甲基化位点特征的 MAD 打分依次减小。每一张图对应一个识别到的 DNA 甲基化位点，其纵轴表示 DNA 甲基化位点 DNA 甲基化位点对应的 DNA 甲基化

β 值，横轴表示不同的类别，从左到右依次是健康（血液）、膀胱癌、乳腺癌、结肠癌、食管癌、头颈癌、肾癌、肝癌、肺腺癌、肺鳞癌、前列腺癌、胃癌、甲状腺癌、子宫内膜癌。可视化结果显示，采用 MAD 打分得到的标志物在 DNA 甲基化位点特征值在各个类别之间的分布直观上来看具有较为明显的差异，但是特定的癌症或者健康样本而言，无法直观确定其是否是肿瘤特异性的。

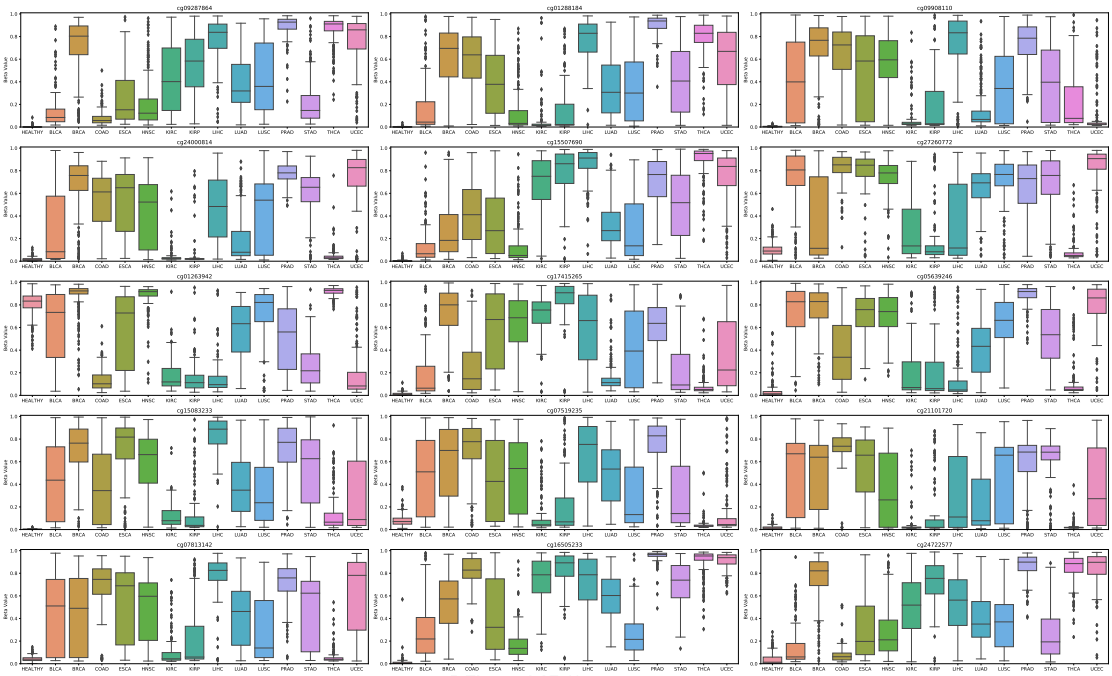


图 4.13　T14B1 测试集上 MAD 打分挑选出的前 15 个 DNA 甲基化位点特征分布箱线图

如图 4.14和 4.15所示，分别是采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合 adjP_ovr 和 MI_ovr 打分识别得到的 $K=1$（针对每类取得分最优的）DNA 甲基化位点特征在不同类别上的分布箱线图，从左到右从上到下分别是针对健康和 14 类癌症（顺序和前文相同）的 DNA 甲基化位点其特征值的可视化结果，横轴和纵轴的含义与图 4.13相同。如图 4.14和 4.15所示，可以直观地看出，对于每一类识别的 DNA 甲基化位点在其对应的类别的特征值的分布和其他类别存在显著差异，由于本章将其他类别之间的差异性过滤条件设置为接近于零的数，导致部分特征其他类别之间的差异性也比较大，但通过特征打分方式，最终本章还是能够识别出某一类和其他类别之间差异性较大，而其他类别之间差异性较小的 DNA 甲基化位点，说明本文提出的方法能保证总体分类性能较优的前提下，识别出位点在测试集上也能得到针对不同肿瘤类型的特异性位点，与此同时本文方法不改变特征原本的含义且具有良好的可解释性。

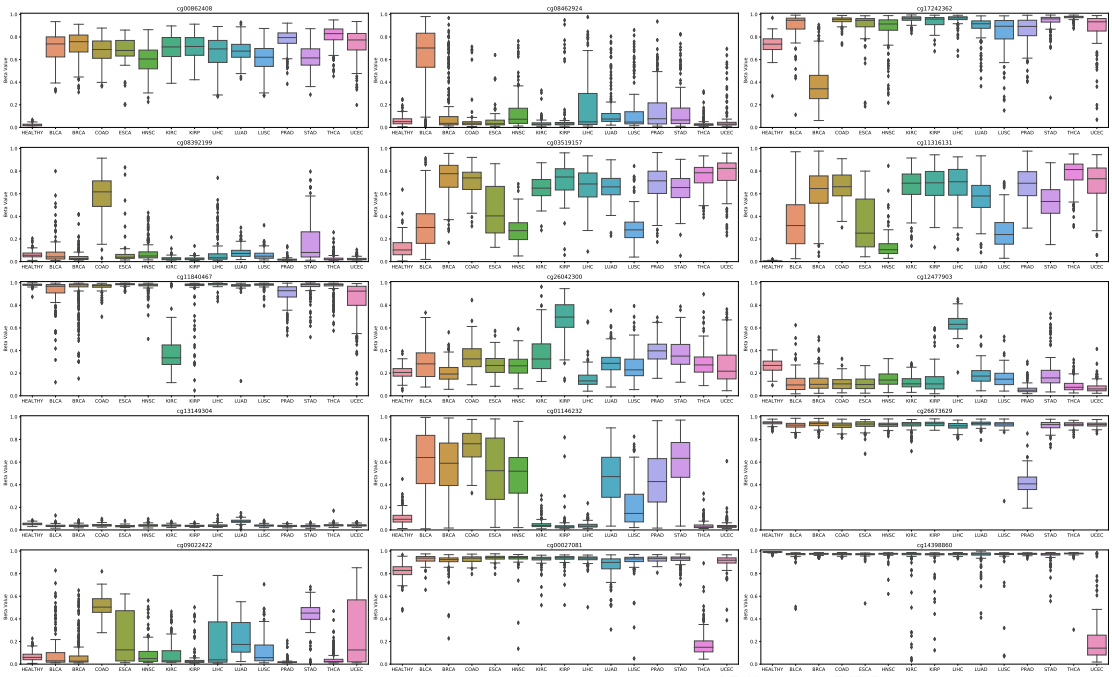


图 4.14　T14B1 测试集上采用 adjP_ovr 分数选择的前 15 个（每类取前 1 个）肿瘤特异性 DNA 甲基化位点特征分布箱线图

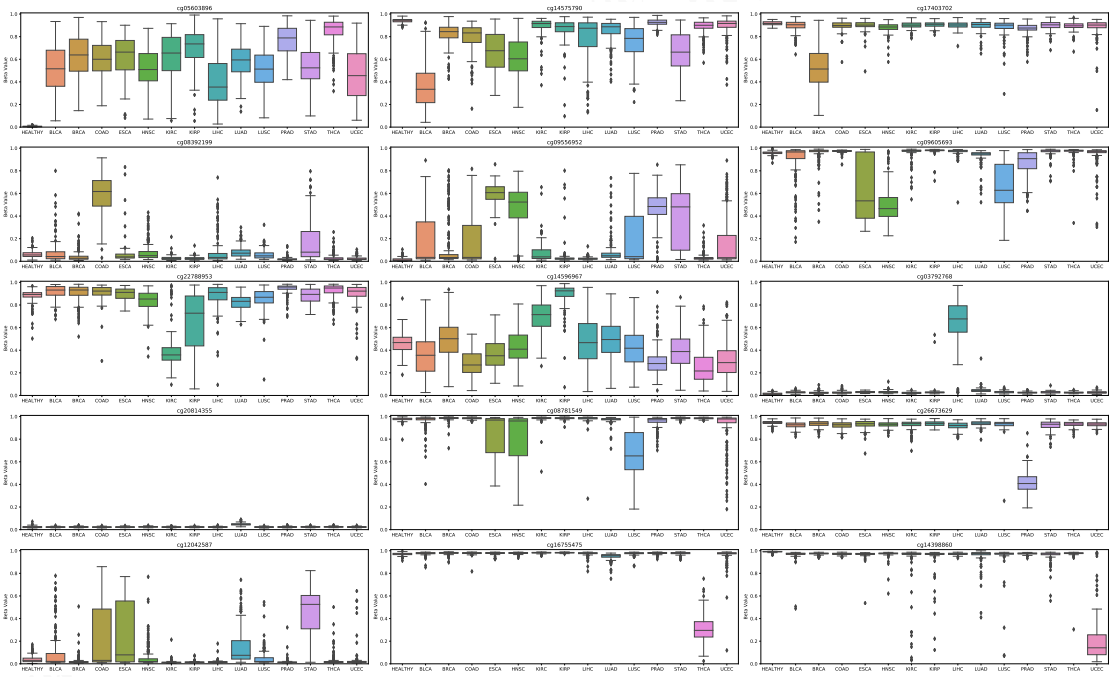


图 4.15　T14B1 测试集上采用 MI_ovr 分数选择的前 15 个（每类取前 1 个）肿瘤特异性 DNA 甲基化位点特征分布箱线图

虽然从计算层面，采用本章的方法取得了较优的分类性能，但还欠缺对应的生物学解释，本章引入 450K 芯片探针设计的注释文件引入到识别得到的 DNA 甲基化位点中，查询得到识别出的位点对应在基因组上的位置和功能信息，如表 4.7是采

用采用 MAD 打分选择出的前 15 个 DNA 甲基化位点的基因注释信息。表 4.8则是采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合 adjP_ovr 打分选择出每类最优的 DNA 甲基化位点的基因注释信息，表 4.9表示采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合 MI_ovr 打分选择出每类最优的 DNA 甲基化位点的基因注释信息，有待开展真实的临床实验来检验对应的位点的生物学意义和诊断的应用价值。

表 4.7　采用 MAD 打分选择出的前 15 个 DNA 甲基化位点的基因注释信息

位点	排序	染色体	基因	和基因相对位置	和 CpG 岛相对位置	调控功能
cg09287864	1	7				未预测细胞类型特异性
cg01288184	2	18	CABLES1	Body		未预测
cg09908110	3	11	MACROD1	Body		未预测细胞类型特异性
cg24000814	4	3			Island	未预测细胞类型特异性
cg15507690	5	3			S_Shelf	未预测
cg27260772	6	6	TFAP2B	Body	Island	
cg01263942	7	10	DIP2C C10orf108	Body TSS200		
cg17415265	8	17				未预测
cg05639246	9	6			S_Shelf	
cg15083233	10	9	PALM2-AKAP2 AKAP2	Body TSS1500	N_Shore	启动子相关
cg07519235	11	16	GPRC5B	5’UTR	Island	
cg21101720	12	17	ANKRD13B	Body	Island	启动子相关
cg07813142	13	2	SP5	Body	Island	
cg16505233	14	3	EDEM1	Body		未预测
cg24722577	15	5				未预测

表 4.8　采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合 adjP_ovr 打分选择出每类最优的肿瘤特异性 DNA 甲基化位点的基因注释信息

位点	特异性	染色体	基因	和基因相对位置	和 CpG 岛相对位置	调控功能
cg00862408	HEALTHY	11	TBC1D10C	Body	Island	启动子相关

表 4.8　采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合 adjP_ovr 打分选择出每类最优的 DNA 甲基化位点的基因注释信息（续）

位点	特异性	染色体	基因	和基因相对位置	和 CpG 岛相对位置	调控功能
cg08462924	BLCA	1			Island	
cg17242362	BRCA	7	ATXN7L1	Body		
cg08392199	COAD	5	LIFR	5'UTR	Island	
cg03519157	ESCA	11	DGKZ	5'UTR 1stExon Body	Island	
cg11316131	HNSC	5	MGAT1	1stExon 5'UTR TSS1500	S_Shore	启动子相关
cg11840467	KIRC	13	ATP11A	Body	S_Shore	
cg26042300	KIRP	1	SMAP2	3'UTR		
cg12477903	LIHC	5	F12	Body	Island	启动子相关
cg13149304	LUAD	21	GABPA ATP5J	5'UTR 1stExon	Island	启动子相关
cg01146232	LUSC	16	SALL1	TSS1500 1stExon	Island	
cg26673629	PRAD	6	SLC22A23	3'UTR		
cg09022422	STAD	3	SLC6A11	TSS200	Island	
cg00027081	THCA	2	TSSC1	Body	S_Shelf	
cg14398860	UCEC	10	INPP5A	Body		

表 4.9　采用 MI_ovr 打分选择出每类最优的肿瘤特异性 DNA 甲基化标位点的基因注释信息

位点	特异性	染色体	基因	和基因相对位置	和 CpG 岛相对位置	调控功能
cg05603896	HEALTHY	6	SCML4	TSS200		启动子相关
cg14575790	BLCA	10				
cg17403702	BRCA	11	ARFIP2	Body	N_Shelf	
cg08392199	COAD	5	LIFR	5'UTR	Island	
cg09556952	ESCA	21	SIM2	Body	Island	
cg09605693	HNSC	8	TSNARE1	5'UTR		
cg22788953	KIRC	7	TTYH3	Body		

表 4.9　采用本文的肿瘤特异性 DNA 甲基化位点识别方法结合 MI_ovr 打分选择出每类最优的 DNA 甲基化位点的基因注释信息（续）

位点	特异性	染色体	基因	和基因相对位置	和 CpG 岛相对位置	调控功能
cg14596967	KIRP	11	ZBTB16	Body	S_Shelf	
cg03792768	LIHC	3	BDH1	5’UTR	Island	启动子相关
cg20814355	LUAD	7	OGDH	TSS200	Island	启动子相关
cg08781549	LUSC	2	HDAC4	3’UTR		
cg26673629	PRAD	6	SLC22A23	3’UTR		
cg12042587	STAD	5	GHR	TSS200	N_Shore	
cg16755475	THCA	16	ZNF500	Body	S_Shore	
cg14398860	UCEC	10	INPP5A	Body		

4.4　本章小结

本章在前一章的基础上，针对癌症类型数量增多导致的肿瘤特异性 DNA 甲基化位点筛选难的问题，放宽了肿瘤特异性 DNA 甲基化位点过滤条件，增加了对于 DNA 甲基化位点的肿瘤特异性衡量特征打分规则，并采用有监督机器学习分类模型进行肿瘤组织来源。实验结果表明，本章的方法能够使用尽量少的 DNA 甲基化位点得到优于现有方法的肿瘤组织来源预测性能。另外，还结合基因注释文件对 DNA 甲基化位点进行了特征分析。

第 5 章　总结和展望

5.1　总结

借助外周血进行基于 DNA 甲基化的癌症液体活检现已成为癌症诊断领域中的研究热点之一，高通量检测技术的进步使获取较大规模 DNA 甲基化数据成为现实，研究肿瘤特异性 DNA 甲基化位点识别方法具有理论意义和应用价值。本文根据当前外周血癌症液体活检中的实际需求，开展了肿瘤特异性 DNA 甲基化位点识别方法的研究，主要研究成果如下：

（1）DNA 甲基化数据集构建

本文针对现有外周血肿瘤特异性 DNA 甲基化位点识别相关研究，缺少用于多类癌症的肿瘤特异性 DNA 甲基化位点识别数据集这一问题，考虑癌症液体活检的泛癌诊断需求，整合了来自 TCGA、GEO、Xena 等公开数据库中的 450K DNA 甲基化芯片数据，构建了较大规模 DNA 甲基化数据集。从上述公开数据库中获取了膀胱癌、乳腺癌、结肠癌、食管癌、头颈癌、肾癌、肝癌、肺癌、前列腺癌、胃癌、甲状腺癌、子宫癌等 14 类肿瘤原发组织以及健康人血液来源的 450K DNA 甲基化芯片数据以及对应的临床数据。同时，对每类肿瘤组织和健康人血液来源的 DNA 甲基化数据集，进行了缺失值的处理、探针过滤、标准化、训练测试集划分等数据预处理工作。在此基础上，构建了可用于识别 14 类癌症的肿瘤特异性 DNA 甲基化位点的多种 DNA 甲基化数据集。另外，还基于 PanSeer 研究，构建了源自三类癌症（胃癌、肺癌、结肠癌）患者和健康人的中国人全血 DNA 甲基化数据集。

（2）肿瘤特异性 DNA 甲基化位点识别

针对样本数远小于特征维数背景下的肿瘤特异性 DNA 甲基化位点识别问题，本文提出了基于类别特异性过滤的 DNA 甲基化位点识别方法。该方法对每个 DNA 甲基化位点，在多类癌症组织和健康人血液来源的样本集之间两两进行差异甲基化分析，在 DNA 甲基化芯片数据集的训练集上采用 Limma 模型进行差异甲基化位点分析，在 PanSeer 数据集的训练集上采用 Welch’s t-test 方法进行差异甲基化位点分析，通过设置统计显著性水平和先验知识等过滤条件，得到每两类之间差异甲基化位点。随后，根据计算得到的 DNA 甲基化位点集合的类别特异性，以及相互之间的集合运算关系，过滤得到某一类别相对其他类别具有显著差异，而其他类别之间差异较小的 DNA 甲基化位点。在不同数据集上的实验结果表明，所提出的方法能够识别出某一类肿瘤相对于其他类别存在显著差异的肿瘤特异性 DNA 甲基化位点。

（3）DNA 甲基化位点的肿瘤特异性衡量和组织来源预测

本文针对类别数增加所引起的肿瘤特异性 DNA 甲基化位点筛选困难的问题，提出了一种基于统计显著性水平和互信息的 DNA 甲基化位点肿瘤特异性衡量方法。该方法在之前工作的基础上，放宽肿瘤特异性 DNA 甲基化位点过滤条件，只要求当前类别和其他类别存在显著差异，而不要求其他类别来自同一分布。设计了基于统计显著性水平和互信息的 DNA 甲基化位点的肿瘤特异性衡量打分规则，得分计算遵循一对多的方式。在训练集上对过滤得到的肿瘤特异性 DNA 甲基化位点进行特征打分和排序，随后选取每类打分靠前的肿瘤特异性 DNA 甲基化位点进行合并，作为随机森林、支持向量机、Fisher 线性判别分析等机器学习模型的输入，分别对上述模型进行训练，用于在测试集上对肿瘤组织来源进行预测，并将上述模型预测值取平均得到最终的预测结果。实验结果表明该方法能够利用尽量少的肿瘤特异性 DNA 甲基化位点取得较优的肿瘤组织来源预测性能。此外，对于识别得到的肿瘤特异性 DNA 甲基化位点，本文结合基因注释文件进行了特征分析，并将其映射到基因组上对应的位置和基因以进行生物学解释。

5.2　展望

本文在肿瘤特异性 DNA 甲基化位点识别方法研究上取得了一些成果，但在 DNA 甲基化标志物的丰富性以及有效性验证等方面还存在不足。未来可根据外周血癌症液体活检的实际需求，从肿瘤特异性 DNA 甲基化位点识别问题所涉及的更复杂的 DNA 甲基化标志物的识别、标志物有效性评价和验证、多组学/多模态数据整合分析等方面，开展进一步研究。

1）更复杂的肿瘤特异性 DNA 甲基化标志物识别

本文所提出的肿瘤特异性 DNA 甲基化位点识别方法是针对 DNA 甲基化位点的，即研究对象是 DNA 上的单个碱基。虽然本文的数据来源也可以是 DNA 甲基化区域特征数据，但在处理时需要将每个 DNA 甲基化区域视作一个 DNA 甲基化位点。未来可以开展更复杂的肿瘤特异性 DNA 甲基化标志物识别方法的研究，可基于现有的差异 DNA 甲基化区域分析方法，检测由多个相邻的 DNA 甲基化位点构成的区域。随后，从得到的差异 DNA 甲基化区域集合中识别出具有肿瘤特异性的 DNA 甲基化区域。也可以基于无监督聚类分析，将具有相似模式的 DNA 甲基化位点聚类成为 DNA 甲基化区域，然后识别出其中具有肿瘤特异性的标志物。更进一步，可以识别由多个 DNA 甲基化区域构成以及存在相互作用的肿瘤特异性 DNA 甲基化标志物，并利用生物信息学等方法，研究标志物的生物学功能以及对于肿瘤相关表型的影响，以更好地实现癌症的早期诊断与动态检测。

2）标志物有效性评价和验证

本文提出的针对多类癌症的肿瘤特异性 DNA 甲基化位点识别方法，需要以肿瘤组织 DNA 甲基化数据作为参考，前提假设是甲基化的 cfDNA 中有一部分来自肿瘤组织，并且来自肿瘤组织的那部分 DNA 的甲基化模式进入血液后保持稳定。此外，还假设肿瘤特异性 DNA 甲基化位点的甲基化模式相较于血液中其他部分 cfDNA 的甲基化模式存在较大差异。而当前关于液体活检中 cfDNA 的来源和体内甲基化模式的变化机制还有待更进一步的研究来揭示。虽然本文的方法在 TCGA、GEO 等公共数据库中的肿瘤患者组织来源和健康人血液来源的 450K DNA 甲基化芯片数据集上取得了较好的实验结果，但该方法在实际应用中受到人种、性别、年龄、吸烟史、遗传因素等样本信息的影响，可能导致预测偏差过大的问题，因而需要进一步研究这些样本信息对 DNA 甲基化的影响。另外，从实际源自中国人临床的 PanSeer 液体活检 DNA 甲基化数据集上的实验结果来看，本文方法以肿瘤组织作为参考所识别的肿瘤特异性 DNA 甲基化位点，在有限的肿瘤患者和健康人血液来源的 cfDNA 甲基化数据集上并没有表现出很明显的肿瘤特异性，还应在更多不同肿瘤患者和健康人血液来源的 cfDNA 数据集上进行更充分的验证。此外，本文的 DNA 甲基化位点的肿瘤特异性衡量方法也有改进的空间，例如增加更多领域知识、结合 AUC 等指标进行评价等，从而设计出和生物学验证更加符合的 DNA 甲基化标志物评价指标。

3）多组学/多模态数据整合分析

本文的研究目前是基于 DNA 甲基化这一表观遗传学特征，未来可以从以下几点进行多组学/多模态数据整合分析，以进一步提升癌症液体活检的效果：（1）对液体活检生物标志物进行充分的数据挖掘，例如针对 cfDNA 同时分析其中的 DNA 甲基化、单核苷酸异质性、拷贝数变异等特征，构建统计模型、层次模型、集成学习模型等计算模型，从多角度对分析结果进行相互比对和印证，从而增加对癌症液体活检生物标志物的认知。（2）检测多种癌症相关液体活检生物标志物，并进行量化和多组学整合分析，例如对 DNA 甲基化和基因表达的关系进行研究。面向癌症液体活检开展多组学整合分析会涉及多组学数据的获取、处理和分析等技术难题，需要进行深入研究。（3）可以将影像学，例如 PET-CT 技术等，和液体活检技术结合。现有的液体活检进行组织来源预测后，还需要借助影像学手段检验预测结果，未来可以同时采集医学影像和液体活检的数据，并构建综合诊断和分析模型，以实现癌症的精准诊断。

参考文献

[1] Institute N C. What is cancer?[EB/OL]. 09/17/2007 - 08:00[2021-05-20]. https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[2] Paget S. The distribution of secondary growths in cancer of the breast.[J]. The Lancet, 1889, 133(3421): 571-573.

[3] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2018, 68(6): 394-424.

[4] Rs Z, Kx S, Sw Z, et al. Report of cancer epidemiology in china, 2015[J/OL]. Zhonghua Zhong liu za zhi [Chinese Journal of Oncology], 2019, 41(1): 19-28. DOI: 10.3760/cma.j.issn.0253-3766.2019.01.005.

[5] Wei W, Zeng H, Zheng R, et al. Cancer registration in china and its role in cancer prevention and control[J/OL]. The Lancet Oncology, 2020, 21(7): e342-e349. DOI: 10.1016/S1470-2045(20)30073-5.

[6] Heitzer E, Haque I S, Roberts C E S, et al. Current and future perspectives of liquid biopsies in genomics-driven oncology[J/OL]. Nature Reviews Genetics, 2019, 20(2): 71-88. DOI: 10.1038/s41576-018-0071-5.

[7] Team N L S T R. Reduced lung-cancer mortality with low-dose computed tomographic screening[J/OL]. New England Journal of Medicine, 2011, 365(5): 395-409. https://doi.org/10.1056/NEJMoa1102873.

[8] Nelson H D, O'Meara E S, Kerlikowske K, et al. Factors associated with rates of false-positive and false-negative results from digital mammography screening: An analysis of registry data [J/OL]. Annals of Internal Medicine, 2016, 164(4): 226-235. DOI: 10.7326/M15-0971.

[9] Tkach M, Théry C. Communication by extracellular vesicles: Where we are and where we need to go[J/OL]. Cell, 2016, 164(6): 1226-1232. DOI: 10.1016/j.cell.2016.01.043.

[10] Aceto N, Bardia A, Miyamoto D, et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis[J/OL]. Cell, 2014, 158(5): 1110-1122. https://www.sciencedirect.com/science/article/pii/S0092867414009271. DOI: https://doi.org/10.1016/j.cell.2014.07.013.

[11] Soung Y H, Ford S, Zhang V, et al. Exosomes in cancer diagnostics[J/OL]. Cancers, 2017, 9(1). https://www.mdpi.com/2072-6694/9/1/8. DOI: 10.3390/cancers9010008.

[12] Zhou B, Xu K, Zheng X, et al. Application of exosomes as liquid biopsy in clinical diagnosis [J/OL]. Signal Transduction and Targeted Therapy, 2020, 5(1): 1-14. DOI: 10.1038/s41392-020-00258-9.

[13] Akca H, Demiray A, Yaren A, et al. Utility of serum dna and pyrosequencing for the detection of egfr mutations in non-small cell lung cancer[J/OL]. Cancer Genetics, 2013, 206(3): 73-80. DOI: 10.1016/j.cancergen.2013.01.005.

[14] Keller L, Belloum Y, Wikman H, et al. Clinical relevance of blood-based ctdna analysis: mutation detection and beyond[J/OL]. British Journal of Cancer, 2021, 124(2): 345-358. DOI: 10.1038/s41416-020-01047-5.

[15] Fernandez-Cuesta L, Perdomo S, Avogbe P H, et al. Identification of circulating tumor dna for the early detection of small-cell lung cancer[J/OL]. EBioMedicine, 2016, 10: 117-123. DOI: 10.1016/j.ebiom.2016.06.032.

[16] Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor dna[J/OL]. Science Translational Medicine, 2017, 9(403). DOI: 10.1126/scitranslmed.aan2415.

[17] Cohen J D, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test[J/OL]. Science, 2018, 359(6378): 926-930. https://science.sciencemag.org/content/359/6378/926. DOI: 10.1126/science.aar3247.

[18] Chabon J J, Hamilton E G, Kurtz D M, et al. Integrating genomic features for non-invasive early lung cancer detection[J]. Nature, 2020, 580(7802): 245-251.

[19] Cescon D W, Bratman S V, Chan S M, et al. Circulating tumor dna and liquid biopsy in oncology [J/OL]. Nature Cancer, 2020, 1(3): 276-290. DOI: 10.1038/s43018-020-0043-5.

[20] Esteller M. Cancer epigenomics: Dna methylomes and histone-modification maps[J/OL]. Nature Reviews Genetics, 2007, 8(4): 286-298. DOI: 10.1038/nrg2005.

[21] Varley K E, Gertz J, Bowling K M, et al. Dynamic dna methylation across diverse human cell lines and tissues[J]. Genome research, 2013, 23(3): 555-567.

[22] Schultz M D, He Y, Whitaker J W, et al. Human body epigenome maps reveal noncanonical dna methylation variation[J]. Nature, 2015, 523(7559): 212-216.

[23] Peng X, Li H D, Wu F X, et al. Identifying the tissues-of-origin of circulating cell-free DNAs is a promising way in noninvasive diagnostics[J/OL]. Briefings in Bioinformatics, 2020. https://doi.org/10.1093/bib/bbaa060.

[24] Sun K, Jiang P, Chan K C A, et al. Plasma dna tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments[J/OL]. Proceedings of the National Academy of Sciences, 2015, 112(40): E5503-E5512. DOI: 10.1073/pnas.1508736112.

[25] Kang S, Li Q, Chen Q, et al. Cancerlocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free dna[J]. Genome biology, 2017, 18(1): 1-12.

[26] Nassiri F, Chakravarthy A, Feng S, et al. Detection and discrimination of intracranial tumors using plasma cell-free dna methylomes[J/OL]. Nature Medicine, 2020, 26(7): 1044-1047. DOI: 10.1038/s41591-020-0932-2.

[27] Shen S Y, Singhania R, Fehringer G, et al. Sensitive tumour detection and classification using plasma cell-free dna methylomes[J]. Nature, 2018, 563(7732): 579-583.

[28] Xu R h, Wei W, Krawczyk M, et al. Circulating tumour dna methylation markers for diagnosis and prognosis of hepatocellular carcinoma[J]. Nature materials, 2017, 16(11): 1155-1161.

[29] Luo H, Zhao Q, Wei W, et al. Circulating tumor dna methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer[J]. Science translational medicine, 2020, 12(524).

[30] Liang W, Zhao Y, Huang W, et al. Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted dna methylation sequencing of circulating tumor dna (ctdna)[J]. Theranostics, 2019, 9(7): 2056.

[31] Chen X, Gole J, Gore A, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test[J]. Nature communications, 2020, 11(1): 1-10.

[32] Sun K, Jiang P, Chan K C A, et al. Plasma dna tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments[J/OL]. Proceedings of the National Academy of Sciences, 2015, 112(40): E5503-E5512. https://www.pnas.org/content/112/40/E5503. DOI: 10.1073/pnas.1508736112.

[33] Teschendorff A E, Relton C L. Statistical and integrative system-level analysis of dna methylation data[J]. Nature Reviews Genetics, 2018, 19(3): 129.

[34] Cho N Y, Park J W, Wen X, et al. Blood-based detection of colorectal cancer using cancer-specific dna methylation markers[J/OL]. Diagnostics, 2021, 11(1): 51. DOI: 10.3390/diagnostics11010051.

[35] Du P, Zhang X, Huang C C, et al. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis[J]. BMC bioinformatics, 2010, 11(1): 1-9.

[36] Ritchie M E, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies[J/OL]. Nucleic Acids Research, 2015, 43(7): e47-e47. https://doi.org/10.1093/nar/gkv007.

[37] Smyth G K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments[J]. Statistical applications in genetics and molecular biology, 2004, 3(1).

[38] Luo H, Zhao Q, Wei W, et al. Circulating tumor dna methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer[J/OL]. Science Translational Medicine, 2020, 12(524). https://stm.sciencemag.org/content/12/524/eaax7533. DOI: 10.1126/scitranslmed.aax7533.

[39] Tang W, Wan S, Yang Z, et al. Tumor origin detection with tissue-specific miRNA and DNA methylation markers[J/OL]. Bioinformatics, 2017, 34(3): 398-406. https://doi.org/10.1093/bioinformatics/btx622.

[40] Zou Q, Zeng J, Cao L, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification[J/OL]. Neurocomputing, 2016, 173: 346-354. https://www.sciencedirect.com/science/article/pii/S0925231215012801. DOI: https://doi.org/10.1016/j.neucom.2014.12.123.

[41] Feng H, Jin P, Wu H. Disease prediction by cell-free DNA methylation[J/OL]. Briefings in Bioinformatics, 2018, 20(2): 585-597. https://doi.org/10.1093/bib/bby029.

[42] Luo H, Wei W, Ye Z, et al. Liquid biopsy of methylation biomarkers in cell-free dna[J/OL]. Trends in Molecular Medicine, 2021. DOI: 10.1016/j.molmed.2020.12.011.

[43] Infinium humanmethylation450 beadchip[J]. 4.

[44] Weinstein J N, Collisson E A, Mills G B, et al. The cancer genome atlas pan-cancer analysis project[J]. Nature genetics, 2013, 45(10): 1113-1120.

[45] Goldman M J, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the xena platform[J]. Nature biotechnology, 2020, 38(6): 675-678.

[46] Goldman M, Craft B, Hastie M, et al. The ucsc xena platform for public and private cancer genomics data visualization and interpretation[J/OL]. bioRxiv, 2019. https://www.biorxiv.org/content/early/2019/09/26/326470. DOI: 10.1101/326470.

[47] Barrett T, Wilhite S E, Ledoux P, et al. Ncbi geo: archive for functional genomics data sets—update[J]. Nucleic acids research, 2012, 41(D1): D991-D995.

[48] Aryee M J, Jaffe A E, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays[J]. Bioinformatics, 2014, 30 (10): 1363-1369.

[49] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for dna microarrays[J]. Bioinformatics, 2001, 17(6): 520-525.

[50] Zhou W, Laird P W, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes[J/OL]. Nucleic Acids Research, 2016, 45(4): e22-e22. https://doi.org/10.1093/nar/gkw967.

[51] Nordlund J, Bäcklin C L, Wahlberg P, et al. Genome-wide signatures of differential dna methylation in pediatric acute lymphoblastic leukemia[J]. Genome biology, 2013, 14(9): 1-15.

[52] Tian Y, Morris T J, Webster A P, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips[J/OL]. Bioinformatics, 2017, 33(24): 3982-3984. https://doi.org/10.1093/bioinformatics/btx513.

[53] Maksimovic J, Gordon L, Oshlack A. Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips[J]. Genome biology, 2012, 13(6): 1-12.

[54] Dedeurwaerder S, Defrance M, Calonne E, et al. Evaluation of the infinium methylation 450k technology[J]. Epigenomics, 2011, 3(6): 771-784.

[55] Teschendorff A E, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data[J]. Bioinformatics, 2013, 29(2): 189-196.

[56] Pidsley, Ruth, Wong Y, et al. A data-driven approach to preprocessing illumina 450k methylation array data[J/OL]. BMC Genomics, 2013, 14: 293. DOI: 10.1186/1471-2164-14-293.

[57] Kuhn H W. The hungarian method for the assignment problem[J]. Naval research logistics quarterly, 1955, 2(1-2): 83-97.

[58] Geman D, d'Avignon C, Naiman D Q, et al. Classifying gene expression profiles from pairwise mrna comparisons[J]. Statistical applications in genetics and molecular biology, 2004, 3(1).

[59] Lin Y, Qian F, Shen L, et al. Computer-aided biomarker discovery for precision medicine: data resources, models and applications[J/OL]. Briefings in Bioinformatics, 2017, 20(3): 952-975. https://doi.org/10.1093/bib/bbx158.

[60] Cover T M. Elements of information theory[M]. John Wiley & Sons, 1999.

[61] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information[J/OL]. Physical Review E, 2004, 69(6): 066138[2021-03-13]. https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

[62] Ross B C. Mutual information between discrete and continuous data sets[J/OL]. PLOS ONE, 2014, 9(2): 1-5. https://doi.org/10.1371/journal.pone.0087357.

[63] Levy J J, Titus A J, Salas L A, et al. PyMethylProcess—convenient high-throughput preprocessing workflow for DNA methylation data[J/OL]. Bioinformatics, 2019, 35(24): 5379-5381. https://doi.org/10.1093/bioinformatics/btz594.

[64] Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning[C/OL]// ICML '07: Proceedings of the 24th International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2007: 1151–1157. https://doi.org/10.1145/1273496.1273641.

致　谢

感谢导师刘民教授的悉心指导，他使我得以窥见宏大科学世界之一角，是我科研生涯的引路人，他的“高标准、严要求”也将使我受益终生。感谢企业导师周平高级工程师，为我提供了宝贵的实践机会，使我得以深入当地大数据企业一线，积累了宝贵的实践经历和人生经验。

感谢清华大学自动化系智能生物制药与生物治疗研究中心的程振老师和董明宇老师，他们勇于钻研和求真务实的科研精神让我获益匪浅。感谢实验室周晓、宋新芳、杨威杨博士生，与他们交流探讨总能让我获益颇多。

感谢清华大学自动化系贵州大数据实践项目相关老师和工作人员的辛勤付出和巨大投入，也祝项目越办越好。感谢清华大学深圳国际研究生院，在深期间提供了良好的学习和生活条件。感谢贵州省大数据局、贵阳市高新区以及实践基地相关人员，在黔期间提供了生活和学习上的诸多便利。感谢深数据硕 18 班集体，一起度过了难忘的三年研究生时光。感谢所有支持和帮助过我的人。

感谢我的家人，一直无条件在背后支持着我，让我的求学生涯没有后顾之忧。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：＿＿＿＿＿＿　日　期：＿＿＿＿＿＿

个人简历、在学期间完成的相关学术成果

个人简历

1995 年 8 月 23 日出生于湖北省蕲春县。

2014 年 9 月考入中国矿业大学计算机学院电子信息科学与技术专业，2018 年 6 月本科毕业并获得工学学士学位。

2018 年 9 月免试进入清华大学自动化系攻读控制工程专业工程硕士至今。

在学期间完成的相关学术成果

参与的科研项目：

[1] 国家智能制造专项项目“中信重工特种机器人制造智能化工厂”（2016ZXFM02005）

[2] 清华-贵州大数据研究生企业实践项目“医学图像器官语义分割自动辅助标注系统”，获优秀研究生奖

指导教师学术评语

外周血肿瘤特异性 DNA 甲基化位点识别是癌症液体活检领域中的重要研究方向。该论文针对癌症液体活检需求，对外周血肿瘤特异性 DNA 甲基化位点识别方法进行了深入研究，论文具有明确的工程应用背景，研究工作具有理论意义和应用价值。论文在对国内外相关领域的研究状况进行较全面综述的基础上，取得如下主要成果：

（1）构建了可用于肿瘤特异性 DNA 甲基化位点识别和肿瘤组织来源预测的较大规模 DNA 甲基化数据集；

（2）针对 DNA 甲基化数据集呈现的样本数远小于特征维度的特点，提出了一种基于类别特异性过滤的 DNA 甲基化位点识别方法；

（3）针对已有类别特异性过滤方法所存在的类别增多会导致肿瘤特异性 DNA 甲基化位点筛选难的问题，提出了一种基于统计显著性水平和互信息的位点特异性衡量方法，并采用有监督分类模型进行肿瘤组织来源预测性能评估。在多个数据集上验证了本文所提出的外周血肿瘤特异性 DNA 甲基化位点识别方法的有效性。

论文工作表明，作者已掌握控制工程专业坚实的基础理论和系统的专门知识，具有从事科学研究和担负专门技术工作的能力。论文结构合理、论述清楚、语言通顺，论文达到了工程硕士学位论文要求，同意进行硕士学位论文答辩，并建议授予刘奇工程硕士学位。

联合指导教师学术评语

该论文在现有国内外研究现状的基础上，对外周血肿瘤特异性DNA甲基化位点识别方法进行了研究，主要工作如下：

(1) 构建了较大规模的用于识别肿瘤特异性DNA甲基化位点的多癌种数据集；

(2) 提出了一种新型的肿瘤特异性DNA甲基化位点识别方法，该方法基于统计假设检验和DNA甲基化差异分析，能从数十万的DNA甲基化位点中有效地识别出多类癌症的肿瘤特异性的DNA甲基化位点；

(3) 针对肿瘤类别数增多导致的肿瘤特异性DNA甲基化位点识别困难的问题，提出了一种基于统计显著性水平和互信息的DNA甲基化位点特异性衡量方法，并采用机器学习多分类模型来对进行肿瘤组织来源预测，并比较了不同方法的预测性能。在多个数据集上的实验结果显示了论文中方法的有效性。

论文表明作者已经掌握了相关理论知识和研究能力，达到了工程硕士的学位要求，同意刘奇参与硕士学位论文答辩。

答辩委员会决议书

论文研究外周血肿瘤特异性 DNA 甲基化位点识别方法，选题具有理论意义与应用价值。

论文取得了以下主要研究成果：

1．构建了可用于多类肿瘤特异性 DNA 甲基化位点识别和肿瘤组织来源预测的 DNA 甲基化数据集；

2．提出了一种基于类别特异性过滤的 DNA 甲基化位点识别方法，筛选出针对每类肿瘤具有特异性的 DNA 甲基化位点；

3. 提出了一种基于统计显著性水平和互信息的位点特异性衡量方法，进行肿瘤组织来源预测性能评估；

论文条理清晰，阐述清楚，写作规范。

论文工作表明作者掌握了本领域坚实的基础理论和系统的专门知识，具有独立承担技术工作的能力。答辩过程中叙述清楚，回答问题满意。经答辩委员会无记名投票表决，5 人一致同意通过论文答辩，并建议授予刘奇同学工程硕士学位。