DOI:10.3969/j.issn.1005-202X.2022.05.018

医学生物信息

基于部分DNA甲基化预测肿瘤浸润免疫细胞的比例

宋春晖[1,2],秦玉芳[1,2],陈明[1,2]
1.上海海洋大学信息学院,上海 201306;2.农业农村部渔业信息重点实验室,上海 201306

【摘要】了解肿瘤组织的免疫细胞比例对肿瘤患者的诊断和预后治疗具有重要意义,DNA甲基化可以作为一种细胞分子特征来量化复杂肿瘤混合物中的多种细胞类型。本研究基于非负矩阵分解算法并结合部分DNA甲基化建立肿瘤浸润免疫细胞比例预测模型。在预测肿瘤混合物中细胞类型的比例时,选择已知的对每种免疫细胞类型差异显著的DNA甲基化水平矩阵作为特征矩阵,将其和肿瘤混合物的DNA甲基化水平输入到非负矩阵分解算法进行细胞类型的比例预测。噪声和肿瘤含量不同设置下的模拟实验验证了该模型预测的免疫细胞比例与真实比例具有很高的一致性,且均方根误差较小。将该模型应用到4种真实的肿瘤样本上,得到每种样本中浸润水平最高的免疫细胞,结果与之前其他学者的生物学发现相一致。本研究的模型能准确预测组成肿瘤混合物的细胞类型比例,为肿瘤的免疫学研究提供了新的方法。

【关键词】癌症;DNA甲基化;非负矩阵分解;免疫细胞
【中图分类号】R318;R730.2 【文献标志码】A 【文章编号】1005-202X(2022)05-0627-08

Predicting the proportions of tumor-infiltrating immune cells based on partial DNA methylation

SONG Chunhui[1,2], QIN Yufang[1,2], CHEN Ming[1,2]
1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs of the People's Republic of China, Shanghai 201306, China

Abstract: Understanding the proportions of immune cells in tumor tissues has important implications for the diagnosis and prognostic treatment of tumor patients, and DNA methylation can be used as a cellular molecular signature to quantify multiple cell types in complex tumor mixtures. Herein a prediction model for the proportions of tumor-infiltrating immune cells is established based on a non-negative matrix factorization algorithm combined with partial DNA methylation. When predicting the proportions of cell types in tumor mixtures, the known DNA methylation level matrix that is significantly different for each immune cell type is selected as the feature matrix, and then the matrix and the DNA methylation level of tumor mixtures are input into the non-negative matrix factorization algorithm for predicting the proportions of cell types. Simulation experiments at different levels of noises and tumor contents verify that the proportions of immune cells predicted by the model are highly consistent with the real proportions, with small root-mean-square error. Applying the model to 4 kinds of real tumor samples yielded the highest levels of infiltrating immune cells in each sample, and the results are consistent with previous biological findings by other scholars. The proposed model can accurately predict the proportions of cell types that make up tumor mixtures, providing a new approach for tumor immunology research.

Keywords: cancer; DNA methylation; non-negative matrix factorization; immune cell

前 言

DNA甲基化状态是基因组研究中一个突出的表观遗传学标记。随着全基因组DNA甲基化数据的获得,大量研究为DNA甲基化在细胞过程和疾病如肾癌[1]、结直肠癌[2]中发挥的作用提供了证据,DNA甲基化状态为疾病的诊断和治疗提供了很好的帮助。

肿瘤微环境中存在的浸润性免疫细胞在癌症进展、患者存活和对症治疗中起重要作用[3]。组织混合物细胞成分测定的实验方法有流式细胞术和尖端单细胞技术,如Drop-seq[4]、10X基因组学和sci-RNA-seq[5],但是实验价格昂贵、劳动强度大,需要新鲜的组织,并且对细胞分离过程中的技术变化敏感。因

【收稿日期】2021-12-10
【基金项目】国家自然科学基金(61702325);国家重点研发计划项目(2018YFD0701003);上海市科技创新计划项目(20dz1203800)
【作者简介】宋春晖,硕士在读,研究方向:机器学习、生物信息,E-mail: 641811576@qq.com
【通信作者】秦玉芳,博士,副教授,研究方向:机器学习、生物信息,E-mail: yfqin@shou.edu.cn

此除了实验方法，采用计算方法预测肿瘤组织中细胞成分的含量十分必要。

基于转录组的基因表达数据对复杂混合物（包括实体瘤）进行分析进而推断细胞类型比例的方法目前有了一些研究[6-9]。Newman等[6]采用支持向量回归的方法CIBERSORT，利用微阵列数据对未知的复杂混合物进行细胞类型比例估计。基于细胞类型特异性mRNA含量的重组，EPIC考虑特征差异显著的细胞类型，从大量的肿瘤转录组表达数据中估计肿瘤和免疫细胞类型的比例[7]。然而由于临床上组织样品的化学固定[10]，通过转录组分析方法测量的RNA分子更容易降解，而DNA甲基化是更稳定的分子且具有高度的细胞类型特异性[11-12]，因此基于DNA甲基化的方法成为细胞反卷积的一种更有效的替代方法。目前，对混合物中细胞含量的估计主要分为两类，一类是基于参考的方法，Houseman等[13]将混合物样本作为组成细胞类型的DNA甲基化与比例的加权组合[14]，利用约束投影/二次规划推断混合物样本中的细胞类型比例；Teschendorff等[15]利用来自NIH表观基因组的细胞类型特异性高灵敏位点信息构建甲基化参考数据来推断全血样本中细胞类型的比例。另一类是基于无参考的方法，最初的两种方法FaST-LMM-EWASher和RefFreeEWAS分别由James等[16]和Houseman等[17]提出，随后Pavlo等[18]在此基础上，基于约束性非负矩阵分解并结合一种新的生物相关正则化函数开发了MeDeCom，并用于预测混合物中细胞类型的比例；Rahmani等[19]基于主成分分析开发了ReFACTor，该方法不需要细胞计数的先验知识，提供了细胞类型组成的改进估计。虽然原则上基于无参考的方法可以应用于任何组织，但这种方法的预测准确率较低；基于参考的方法能获得较高的细胞比例预测准确率，然而在实际临床上往往只知道一部分细胞类型的甲基化，即部分参考的情况，因此可以利用容易获得的肿瘤混合物中部分细胞类型的表观基因组信息去推断出所有组成细胞类型的比例。基于部分参考的方法已成功应用于基于转录组数据的肿瘤浸润免疫细胞的分解[20]。

本研究基于肿瘤组织中已知细胞类型的甲基化，利用非负矩阵分解的框架去估计所有细胞类型的比例，简记为MethyPR。对模拟数据的评估表明本研究的方法较现有方法在识别细胞类型比例上有了明显的提高；其次，在体外制备的混合物上验证了本研究的方法能很好地还原出所有细胞类型的比例；最后，将MethyPR应用于癌症基因组图谱（TCGA）甲基化数据，能很好地识别出癌症特异性肿瘤浸润性免疫细胞，为靶向免疫治疗的设计提供依据。

1 材料与方法

1.1 数据集获取

1.1.1 模拟混合物数据　本研究从基因表达综合数据库（GEO）下载了6种纯化免疫细胞[CD4+T细胞、CD8+T细胞、自然杀伤细胞、B细胞、单核细胞（Mon）和粒细胞（Gra）（GSE35069）]和一种乳腺癌细胞（MCF-7）（GSE44837）的甲基化谱，并按照一定比例产生免疫细胞和肿瘤细胞的混合物。

1.1.2 实验获得的真实样本数据　真实样本的甲基化数据以及组成样本的细胞类型甲基化数据来源于Onuchic等[21]的研究，具体来说，该数据集包含两个样本集和构成样本的细胞类型甲基化数据。第一个样本集由6个样本组成，分别由3个成对细胞系的组合组成（MCF-7/HMEC、MCF-7/CD8+T细胞和MCF-7/CAF），分别按75%:25%和95%:5%的比例混合组成。第二个样本集是由亚硫酸氢盐测序生成的29个乳腺肿瘤样本组成，每个样本由不同的乳腺癌细胞系、一种正常乳腺细胞系（HMEC）、一种成纤维细胞系（CAF）和一种免疫细胞（CD8+T）组成，同时使用H&E染色，估计每个样本的癌性、正常、成纤维和免疫细胞的比例。该数据集还包含了构成样本的细胞系的甲基化数据，分别是6种不同的乳腺癌细胞系（MCF-7、T47D、MDA-MB-231、MDA-MB-361、HCC1954、HCC1569）、HMEC、CAF和CD8+T细胞。

1.1.3 TCGA癌症样本数据　本研究使用GDC客户端工具从GDC数据门户（https://gdc.nci.nih.gov）下载胸腺瘤、结直肠癌、急性髓系肿瘤和弥漫性大B细胞淋巴瘤样本的Infinium HumanMethylation 450芯片的三级甲基化数据。另外，从Arneson等[22]的研究得到免疫细胞的甲基化数据，该数据集包含11种细胞类型（单核细胞、树突状细胞、巨噬细胞、中性粒细胞、嗜酸性粒细胞、调控性T细胞、幼稚T细胞、记忆T细胞、CD8+T细胞、自然杀伤细胞和B细胞）的甲基化数据（GSE35069、GSE59250、GSE71837），并且按照文献[22]的方法，采用单核细胞作为桥接细胞类型纠正批次效应，最终得到419个甲基化位点处的肿瘤浸润性免疫细胞甲基化数据。

1.2 数据预处理

对于下载得到的高维甲基化数据，首先排除甲基化位点缺失超过10%的样本以及样本缺失超过10%的甲基化位点。然后，用R包“ChAMP”包进行一系列处理：使用ChAMP.impute方法对缺失值进行填充，使用ChAMP.filter方法过滤常染色体上与性相关的和位于X、Y染色体上的位点，以避免任何与性别相关的信号，同时过滤所有与单核苷酸多态性（Single Nucleotide Polymorphism, SNP）重叠的和报

道为交叉反应的位点[23]。例如,对于模拟实验中6种免疫细胞的高维甲基化数据,排除6个在每种细胞中缺失值超过10%的位点,接下来过滤10 028个与性相关和位于X、Y染色体上的位点,使用一般的450K SNP列表过滤了59 901个SNP重叠的位点,并移除了11个交叉反应的位点,最终得到6种免疫细胞对应的412 481个位点处的甲基化水平。

为了获得对每种细胞类型差异显著的位点,对上述数据进行进一步处理,使用ChAMP.DMP方法提取细胞类型中差异显著的位点[24]。接下来,遵循文献[25]对上述得到的数据集中的每种细胞类型的甲基化水平与其他细胞类型甲基化水平进行t检验比较,得到差异显著(P<0.000 1)的甲基化位点,按照P值进行降序排列,选择出1 000个甲基化位点,并选择最终的甲基化位点为每次t检验中选择出的位点的交集。例如:对于模拟实验中的6种免疫细胞,通过"ChAMP"步骤的处理,得到25 820个对于每种细胞类型差异显著的位点,经过t检验的处理后,最终得到包含412个显著差异的位点处的甲基化水平。

1.3 模型

1.3.1 模型建立 MethyPR的模型如图1所示。$\boldsymbol{M} \in \boldsymbol{R}^{m \times n}$是来自$k$个细胞类型组成的一个行为$m$个位点,列对应$n$个样本的DNA甲基化水平矩阵。其中$M_{ij} \in \boldsymbol{M}$,表示第$i$个位点在第$j$个样本的甲基化探针占探针总数的比例,所以$0 \leqslant M_{ij} \leqslant 1$。本研究用$\boldsymbol{W} \in \boldsymbol{R}^{m \times k}$表示每个位点的细胞类型甲基化水平,即特征矩阵,其中$W_{ih} \in \boldsymbol{W}$;此外,用$\boldsymbol{H} \in \boldsymbol{R}^{k \times n}$表示$k$个细胞类型在样本中的比例矩阵,$H_{hj} \in \boldsymbol{H}$。根据文献[26],常用的DNA甲基化混合物模型是

$$M_{ij} = W_{ih} H_{hj} + \varepsilon_{ij} \tag{1}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2) \tag{2}$$

$$\forall h \forall j: H_{hj} \geqslant 0 \tag{3}$$

$$\forall j: \sum_{h=1}^{k} H_{hj} = 1 \tag{4}$$

$$\forall i \forall h: 0 \leqslant W_{ih} \leqslant 1 \tag{5}$$

其中,误差项ε_{ij}服从正态分布。式(3)和(4)中的约束要求细胞比例是正的,并在每个样本中的总和为1;式(5)中的约束要求细胞类型的甲基化水平在[0,1]范围内。

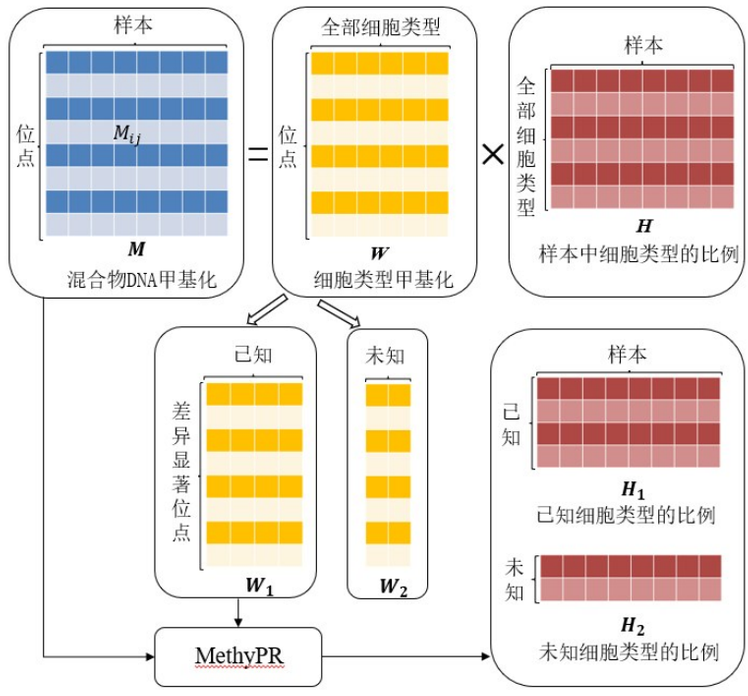


图1 MethyPR模型的框架

Figure 1 Framework of MethyPR model

在大多数情况下,矩阵$\boldsymbol{W}$的数据并不是完全已知,即研究中讨论的细胞类型部分已知的情况。此时模型可以修改为:

$$M_{ij} = \sum_{h=1}^{k_1} W_{ih} H_{hj} + \sum_{h=k_1+1}^{k} W_{ih} H_{hj} + \varepsilon_{ij} \tag{6}$$

矩阵$\boldsymbol{W}$分为$\boldsymbol{W}_1$和$\boldsymbol{W}_2$,即式(6)中的$W_{ih}(1 \leqslant h \leqslant k_1)$和$W_{ih}(k_1 + 1 \leqslant h \leqslant k)$,$\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{W}_2)$,样本细胞比例矩阵$\boldsymbol{H}$分为$\boldsymbol{H}_1$和$\boldsymbol{H}_2$,即式(6)中的$H_{hj}(1 \leqslant h \leqslant k_1)$和$H_{hj}(k_1 + 1 \leqslant h \leqslant k)$,$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \end{pmatrix}$,其中$\boldsymbol{W}_1$和$\boldsymbol{H}_1$是已知的细胞类型所对应的甲基化水平和所占的比例,$\boldsymbol{W}_2$和$\boldsymbol{H}_2$是未知细胞类型所对应的甲基化水平和所占的比例。

给定肿瘤样本的甲基化矩阵$\boldsymbol{M}$和已知细胞甲基化矩阵$\boldsymbol{W}_1$，目标是得到所有细胞类型对应的比例矩阵$\boldsymbol{H}$。这个问题可以转化为寻找合适的$\boldsymbol{H}_1$、$\boldsymbol{H}_2$和$\boldsymbol{W}_2$，使全局误差平方和最小：

$$\widehat{\boldsymbol{H}_1},\widehat{\boldsymbol{H}_2},\widehat{\boldsymbol{W}_2}=\underset{\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{W}_1}{\operatorname{argmin}}\left\|\boldsymbol{M}-(\boldsymbol{W}_1\boldsymbol{W}_2)\begin{pmatrix}\boldsymbol{H}_1\\ \boldsymbol{H}_2\end{pmatrix}\right\|_F^2 \tag{7}$$

$$\text{s.t}\quad \forall h\forall j:H_{hj}\geq 0 \tag{8}$$

$$\forall j:\sum_{h=1}^{k_1}H_{hj}\leq 1 \tag{9}$$

$$\forall j:\sum_{h=k_1+1}^{k}H_{hj}\leq 1 \tag{10}$$

$$\forall i\forall h:0\leq W_{ih}\leq 1 \tag{11}$$

其中$\|\cdot\|_F^2$是弗罗贝尼乌斯范数。

1.3.2 模型求解　采用迭代的非负矩阵分解方法来求解模型，具体流程如下：

（1）随机给定$\boldsymbol{W}_2$矩阵的初始值；

（2）$\boldsymbol{H}^t=(\boldsymbol{H}_1^t,\boldsymbol{H}_2^t)=\underset{\boldsymbol{H}_1,\boldsymbol{H}_2}{\operatorname{argmin}}\left\|\boldsymbol{M}-\boldsymbol{W}_1\boldsymbol{H}_1-\boldsymbol{W}_2^{t-1}\boldsymbol{H}_2\right\|_F^2$，$\boldsymbol{H}_1,\boldsymbol{H}_2\geqslant 0$且$\forall j:\sum_{h=1}^{k_1}H_{hj}\leqslant 1$，$\forall j:\sum_{h=k_1+1}^{k}H_{hj}\leqslant 1$；

（3）$\boldsymbol{W}_2^t=\underset{\boldsymbol{W}_2}{\operatorname{argmin}}\left\|\boldsymbol{M}-\boldsymbol{W}_1\boldsymbol{H}_1^t-\boldsymbol{W}_2\boldsymbol{H}_2^t\right\|_F^2$，$0\leqslant \boldsymbol{W}_1,\boldsymbol{W}_2\leqslant 1$；

（4）不断重复（2）和（3）的步骤，直至收敛或者达到预定的最大迭代次数。

其中，t为迭代次数。在第（1）步中，使用RPMM算法[27]来初始化未知细胞类型的甲基化水平。在运用MethyPR时要注意，该方法适用于下载的定量基因组的原始数据，如原始reads数或CpG位点甲基化水平，不适合使用logit变换的数据，因为变换会破坏数据中存在的线性关系。

1.3.3 细胞类型数量确定和模型质量评估　采用赤池信息准则（Akaike Information Criterion, AIC）度量来确定肿瘤样本中细胞类型的数量。AIC是衡量统计模型拟合优良性的一种方法，它不仅考虑了模型的拟合优度，也考虑了当模型成分数量增加时可能发生的过度拟合。由于本实验中的样本量小，因此采用适用于样本量小的改进版本AICc。AICc的公式为：

$$\text{AICc}=n\ln\left(\frac{\text{SSR}}{n}\right)+2k+\frac{2k(k+1)}{n-k-1} \tag{12}$$

其中，n表示样本量；k表示模型参数的数量（细胞类型的数量）；SSR为残差平方和。不同的k值在模型中对应不同的AICc值，在所有结果中，AICc最小时所对应的k就是最优的细胞类型数。

采用两个评价指标来评估MethyPR的性能，分别为均方根误差（Root Mean Squared Error, RMSE）和Pearson相关系数。RMSE是指对预测值与真实值差平方的平均值求平方根，这是回归问题常用的性能指标；Pearson相关系数用于衡量真实值和预测值之间的相关程度，其值在[-1, 1]之间，绝对值越接近1，相关性越强。本研究用模型预测的细胞类型比例和真实细胞比例的Pearson相关系数来衡量评估方法的精度，另一方面，使用真实细胞类型的比例和预测的细胞比例之间的RMSE作为评价指标。

2 结果与分析

2.1 模拟混合物验证

首先，在模拟混合物上对本研究提出的方法进行基准测试，模拟混合物包含6种纯化的免疫细胞（CD4+T细胞、CD8+T细胞、巨噬细胞、B细胞、单核细胞和粒细胞）和一种乳腺癌癌症细胞（MCF-7）。下载这7种细胞类型的甲基化水平之后，将肿瘤细胞以90%的含量添加到混合物中，剩余部分由6种免疫细胞类型按随机比例构成。

对于上述基准数据集，将现有的3种方法（QP[13]、CIBERSORT[6]、EDEC[21]）与本研究提出的方法进行对比。QP是基于参考的方法，直接使用组成细胞类型的参考甲基化谱，将其与混合物样本甲基化谱作为输入，利用二次规划来推断混合物样本中的细胞类型比例。CIBERSORT内核为支持向量机的一个实例（ν - SVR），将其与细胞类型差异显著的特征矩阵相结合，从而预测细胞类型的比例，其中参数ν给出了训练误差的上界和支持向量的界。EDEC假设混合物样本的对应位点甲基化谱为细胞类型特异性甲基化谱和细胞类型比例的线性组合，随机初始化细胞类型甲基化谱，通过二次规划求解约束最小二乘问题，得到估计的细胞类型比例。

已知有6种免疫细胞，本研究假设模拟混合物的细胞类型数量为7~15，将细胞数量、模拟混合物数量以及预测结果与模拟结果的误差作为AICc指标的输入，进而推断混合物中细胞类型的总数量。由图2发现，当细胞类型数量为7时AICc值最小，最小AICc值识别的细胞类型数量与我们构建模拟混合物时的数量一致。接下来，对于肿瘤含量为90%的混合物，将混合物样本和6种纯化免疫细胞经过预处理后对应的位点处的甲基化水平作为方法的输入，而乳腺癌细胞的甲基化水平未知，用4种方法分别估计了模拟混合物中6种免疫细胞类型的比例。

如图3所示，分别计算了4种方法预测的免疫细胞实际比例与预测比例的RMSE和Pearson相关系数。从图3a可以看出，本研究方法在组成混合物样本的所有免疫细胞中取得最小的RMSE分别为0.005 4、0.003 3、0.006 4、0.006 0、0.008 4、0.005 6）；图3b

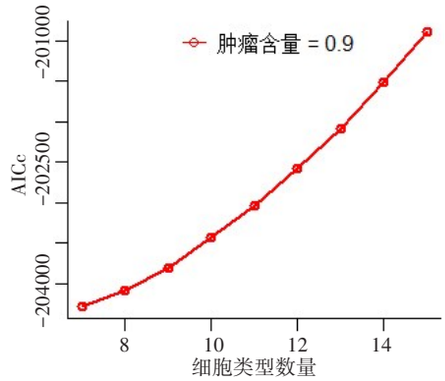
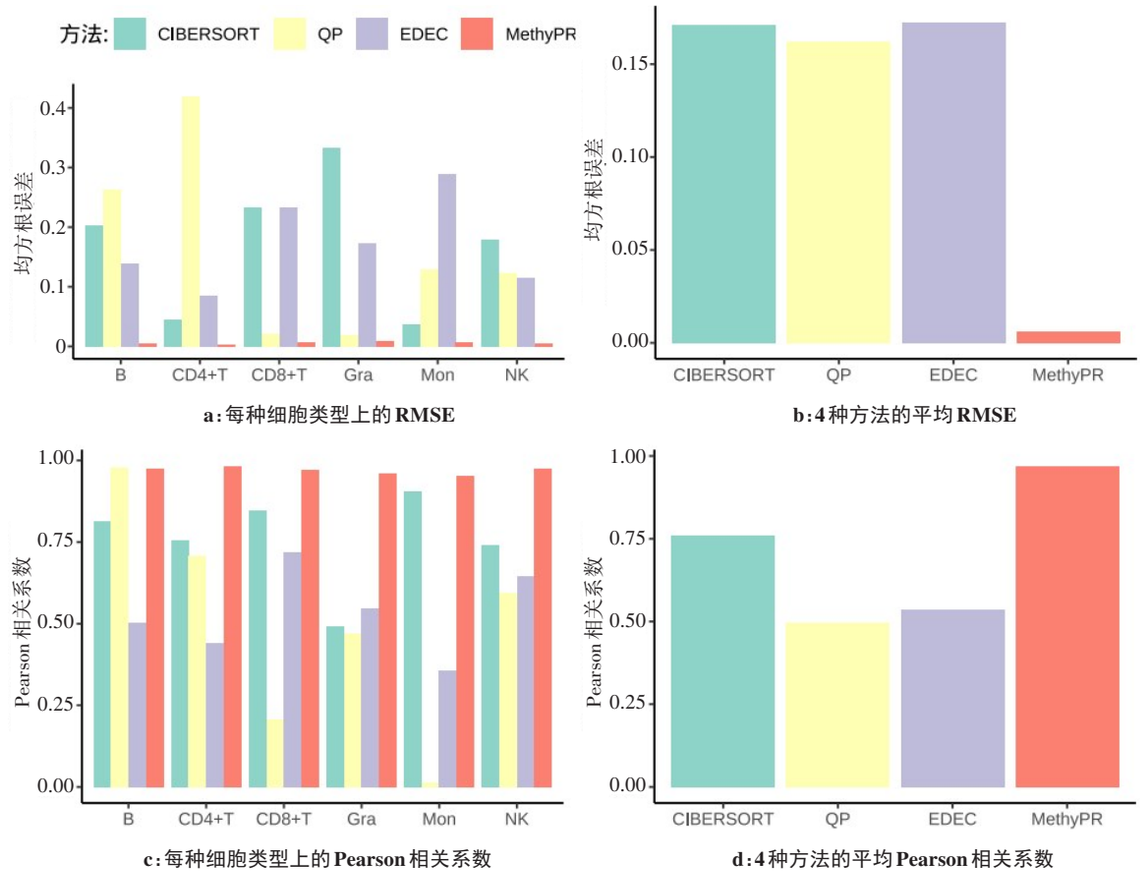


图2 识别模拟混合物中细胞类型数量的AICc线图

Figure 2 AICc line graph for identifying the number of cell types in the simulated mixture

显示了4种方法在所有免疫细胞类型上的RMSE平均值，可以观察到本研究方法明显优于其他方法。图3c和图3d为4种方法在Pearson相关系数指标下的比较，MethyPR的性能显著高于其他方法，平均Pearson相关系数指标达到0.97。

为了测试本研究方法对不同噪声水平和生成肿瘤混合物时不同细胞类型含量下的稳健性，进一步进行了以下实验。固定混合物中肿瘤含量为90%，在模拟混合物中加入不同的噪声水平，从0.1逐渐增加到0.5，比较MethyPR和其他3种方法在不同噪声数据上的预测性能，其中每次模拟重复20次。另外，固定噪声水平为0.1，将肿瘤细胞的含量从90%降低到



a：每种细胞类型上的RMSE　　b：4种方法的平均RMSE

c：每种细胞类型上的Pearson相关系数　　d：4种方法的平均Pearson相关系数

图3 MethyPR与其他算法在细胞水平上的比较

Figure 3 Comparison of MethyPR and other algorithms at the cellular level

50%（每次减少10%）添加到混合物中，混合物样本中的剩余部分由6种免疫细胞类型按随机比例构成，每次模拟同样重复20次。在这两个参数的不同设置下，估计6种免疫细胞类型的比例。

在肿瘤含量分别为50%、60%、70%、80%、90%的情况下，计算AICc值。由表1可知，混合物的细胞类型数量均在7时取得最小，这与构建混合物的数量相一致，说明采用AICc值预测肿瘤细胞类型数量得到了很好的结果。由表2可知，在不同的噪声水平下，也得到了类似的结果。

比较4种方法在不同噪声水平和不同肿瘤细胞含量下的预测性能（图4）。在不同的噪声水平下，得到细胞比例的RMSE和Pearson相关系数如图4a和图4b所示。MethyPR在不同噪声水平下均优于其他方法，对于低噪声水平的性能是最佳的，但是在高噪声下，QP的性能逐渐赶上MethyPR。从图4c可以看

表1 生成混合物中不同肿瘤细胞含量下的AICc值

Table1 AICc values of the generated mixture containing different tumor cell contents

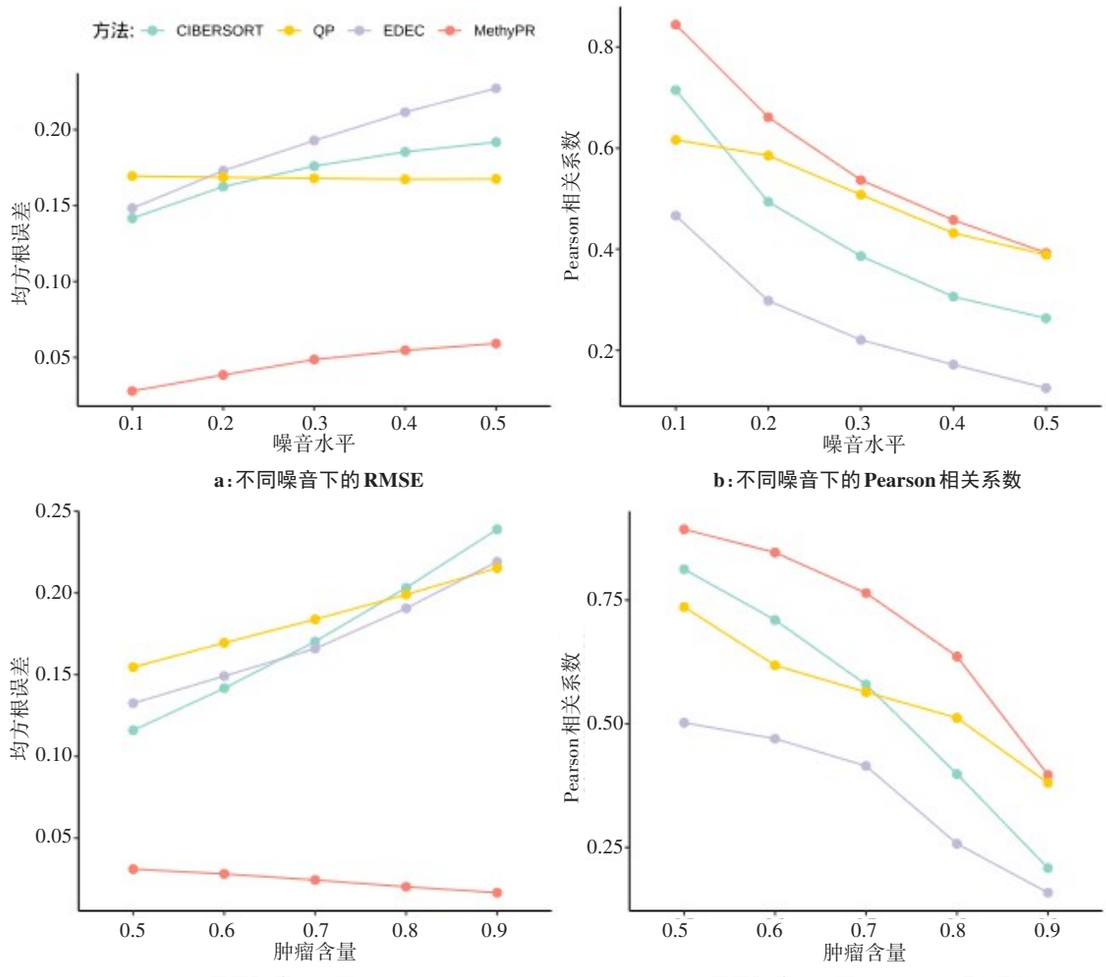
细胞类型数量	肿瘤细胞含量				
	50%	60%	70%	80%	90%
7	-219 529.3	-215 657.8	-211 864.8	-208 718.2	-204 674.4
8	-219 330.5	-215 511.7	-211 691.4	-208 500.6	-204 459.3
9	-219 071.1	-215 295.8	-211 465.5	-208 191.7	-204 147.8
10	-218 774.9	-214 998.8	-211 151.8	-207 850.2	-203 782.8
11	-218 452.5	-214 686.9	-210 814.6	-207 463.6	-203 405.3
12	-218 098.2	-214 328.3	-210 417.4	-207 020.9	-202 950.0
13	-217 661.2	-213 906.3	-209 970.0	-206 551.7	-202 437.4
14	-217 217.3	-213 461.9	-209 482.5	-205 993.5	-201 883.0
15	-216 730.5	-212 946.3	-208 965.4	-205 429.9	-201 285.0

表2 生成混合物在不同噪音水平下的AICc值

Table 2 AICc values of the generated mixture at different noise levels

细胞类型数量	噪声水平				
	0.1	0.2	0.3	0.4	0.5
7	-215 785.9	-158 235.4	-124 698.5	-101 615.16	-84 171.16
8	-215 622.7	-158 038.4	-124 426.0	-101 301.15	-83 856.15
9	-215 389.4	-157 777.0	-124 092.2	-100 935.82	-83 495.11
10	-215 092.3	-157 453.5	-123 676.5	-100 498.07	-83 025.86
11	-214 761.7	-157 064.5	-123 225.2	-99 990.76	-82 498.74
12	-214 370.4	-156 607.6	-122 712.3	-99 426.42	-81 912.15
13	-213 953.5	-156 119.4	-122 118.6	-98 813.83	-81 280.35
14	-213 485.4	-155 560.4	-121 502.1	-98 153.95	-80 612.79
15	-212 953.3	-154 940.2	-120 819.5	-97 393.01	-79 844.15

到，MethyPR在不同肿瘤细胞含量下取得稳定的较小RMSE，而其他方法随着未知肿瘤含量的增加，RMSE越来越大；图4d中，随着肿瘤细胞含量的增加，4种方法的预测性能逐渐下降，但在所有情况下，MethyPR具有最高的Pearson相关系数。



a：不同噪音下的RMSE　　b：不同噪音下的Pearson相关系数

c：不同肿瘤含量下的RMSE　　d：不同肿瘤含量下的Pearson相关系数

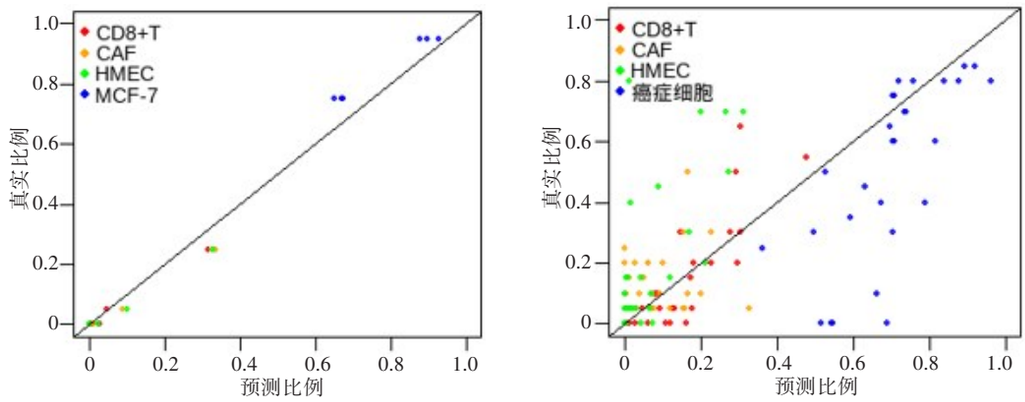
图4 不同方法在不同噪声水平和不同肿瘤细胞含量下的性能评估

Figure 4 Performance evaluations of different methods at different noise levels and different tumor cell contents

2.2 体外制备的细胞系混合物验证

在Onuchic等[21]体外制备的细胞混合物上验证本研究提出的方法。从GEO下载MCF-7、HMEC、CAF和CD8+T细胞的甲基化谱，由3对组合（MCF-7/HMEC、MCF-7/CD8+T细胞和MCF-7/CAF）各按两种比例（75%:25%和95%:5%）生成6个样本，每个样本通过靶向亚硫酸氢盐测序进行分析。采用数据预处理方法得到149个在不同乳腺癌细胞类型具有显著差异的位点并用于本研究方法。在癌症细胞MCF-7未知的情况下（图5a），MethyPR估计的免疫细胞比例和真实值之间有很强的一致性（R=0.996），同时组成混合物的MCF-7细胞比例与真实值也有很强的一致性（R=0.993）。

临床病理学家根据H&E染色对29个乳腺肿瘤样本进行细胞类型组成评估，估计癌性、正常、基质和免疫细胞的比例。本研究假定癌性细胞未知，将其余3种细胞类型差异显著的特征矩阵和29个样本的混合甲基化矩阵作为输入，用MethyPR估计29个样本中细胞的比例，观察到估计的CD8+T免疫细胞与真实免疫细胞的比例具有较高的一致性（R=0.81），而癌症细胞的相关性与之相比较低（R=0.71），见图5b。从图5可以看出，MethyPR可以准确地预测免疫细胞和剩余其他细胞类型的比例。



a：成对细胞组成的样本集预测结果　　b：H&E染色的样本集预测结果

图5 MethyPR在体外制备的混合物上的验证

Figure 5 Validation of MethyPR on the mixture prepared *in vitro*

2.3 肿瘤免疫浸润细胞成分估计

从TCGA的4种肿瘤样本中分别选取了100个样本，然后将MethyPR应用到这些数据上，估计样本中浸润免疫细胞的比例。如表3所示，不同类型的肿瘤样本表现出不同的免疫细胞浸润模式。幼稚T细胞在胸腺瘤样本中占有最高的比例（0.031），这与之前的实验研究一致[28]。在结直肠癌样本中，以T细胞为主[29]，而急性髓系肿瘤样本具有高比例的单核细胞，弥漫性大B细胞淋巴瘤样本中的B细胞比例最高，这分别与文献[21]和文献[30]结果相一致。

表3 肿瘤浸润免疫细胞在TCGA样本中的比例

Table 3 Proportions of tumor-infiltrating immune cells in TCGA samples

肿瘤类型	单核细胞	树突状细胞	巨噬细胞	中性粒细胞	嗜酸性粒细胞	调节性T细胞	幼稚T细胞	记忆T细胞	CD8+T细胞	自然杀伤细胞	B细胞
胸腺瘤	0.008	0.006	0.011	0.004	0.009	0.011	0.031	0.017	0.016	0.007	0.023
急性髓瘤	0.016	0.007	0.001	0.002	0.009	0.005	0.010	0.013	0.008	0.004	0.010
弥漫淋巴	0.027	0.016	0.020	0.015	0.008	0.018	0.019	0.039	0.025	0.010	0.053
结直肠瘤	0.012	0.006	0.043	0.013	0.005	0.013	0.005	0.027	0.006	0.014	0.019

3 结 论

本研究提出了一种使用甲基化数据对肿瘤微环境进行稳健反卷积的方法MethyPR，该方法基于非负矩阵分解方法，利用容易获得的免疫细胞类型的表观基因组信息，从DNA甲基化估计细胞组成比例。作为解决基于参考与无参考方法局限性的新方法，MethyPR能够基于部分参考数据推断混合物中的细

胞类型比例。在模拟混合物、实验混合物和真实混合物上，本研究方法都表现出良好的性能，其预测精度高，可以帮助减少估计肿瘤组成的时间和金钱成本，快速获得肿瘤混合物的免疫细胞比例，为表观基因组研究提供了新的思路。

【参考文献】

[1] Inessa S, Liudmyla T, Kateryna O, et al. Concentration and methylation of cell-free DNA from blood plasma as diagnostic markers of renal cancer[J]. Dis Markers, 2016, 2016: 3693096.

[2] Khadijeh J, Marjan A, Ali J, et al. A DNA methylation panel for high performance detection of colorectal cancer[J]. Cancer Genet, 2020, 252-253(6): 64-72.

[3] Coussens LM, Zitvogel L, Palucka AK. Neutralizing tumor-promoting chronic inflammation: a magic bullet?[J]. Science, 2013, 339(6117): 286-291.

[4] Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets[J]. Cell, 2015, 161(5): 1202-1214.

[5] Cao J, Packer JS, Ramani V, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism[J]. Science, 2017, 357(6352): 661-667.

[6] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles[J]. Nat Methods, 2015, 12(5): 453-457.

[7] Racle J, De Jonge K, Baumgaertner P, et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data [J]. eLife, 2017, 6: e26476.

[8] Altboum Z, Steuerman Y, David E, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection[J]. Mol Syst Biol, 2014, 10(2): 720.

[9] Wang Z, Cao S, Morris JS, et al. Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration[J]. iScience, 2018, 9: 451-460.

[10] Von Ahlfen S, Missel A, Bendrat K, et al. Determinants of RNA quality from FFPE Samples[J]. PLoS One, 2007, 2(12): e1261.

[11] Daugaard I, Kjeldsen TE, Hager H, et al. The influence of DNA degradation in formalin-fixed, paraffin-embedded (FFPE) tissue on locus-specific methylation assessment by MS-HRM[J]. Exp Mol Pathol, 2015, 99(3): 632-640.

[12] Baron U, Turbachova I, Hellwag A, et al. DNA methylation analysis as a tool for cell typing[J]. Epigenetics, 2006, 1(1): 59-61.

[13] Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution[J]. BMC Bioinform, 2012, 13(1): 1-16.

[14] Wiencke JK, Accomando WP, Zheng S, et al. Epigenetic biomarkers of T-cells in human glioma[J]. Epigenetics, 2012, 7(12):1391-1402.

[15] Teschendorff AE, Breeze CE, Zheng SC, et al. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies[J]. BMC Bioinform, 2017, 18(1): 105.

[16] James Z, Christoph L, David H, et al. Epigenome-wide association studies without the need for cell-type composition.[J]. Nat Methods, 2014, 11(3): 309-411.

[17] Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data[J]. Bioinformatics, 2014, 30(10):1431-1439.

[18] Pavlo L, Martin S, Gilles G, et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes[J]. Genome Biol, 2017, 18(1): 55.

[19] Rahmani E, Zaitlen N, Baran Y, et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies[J]. Nat Methods, 2016, 13(5): 443-445.

[20] Qin Y, Zhang W, Sun X, et al. Deconvolution of heterogeneous tumor samples using partial reference signals[J]. PLoS Comput Biol, 2020, 16(11): e1008452.

[21] Onuchic V, Hartmaier RJ, Boone DN, et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types[J]. Cell Rep, 2016, 17(8): 2075-2086.

[22] Arneson D, Yang X, Wang K. MethylResolver-a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents[J]. Commun Biol, 2020, 3(1): 286-291.

[23] Morris TJ, Lee MB, Feber A, et al. ChAMP: 450k chip analysis methylation pipeline[J]. Bioinformatics, 2014, 30(3): 428-430.

[24] Feng H, Zhang Y, Liu K, et al. Intrinsic gene changes determine the successful establishment of stable renal cancer cell lines from tumor tissue[J]. Int J Cancer, 2017, 140(11): 2526-2534.

[25] Abbas AR, Wolslegel K, Seshasayee D, et al. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus[J]. PLoS One, 2009, 4(7): e6098.

[26] Accomando WP, Wiencke JK, Houseman EA, et al. Quantitative reconstruction of leukocyte subsets using DNA methylation [J]. Genome Biol, 2014, 15(3): R50.

[27] Houseman EA, Christensen BC, Yeh RF, et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions[J]. BioMed Central, 2008, 9(1): 365.

[28] Philipp S, Markus H, Georgios M, et al. Paraneoplastic myasthenia gravis correlates with generation of mature naive CD4+ T cells in thymomas[J]. Blood, 2002, 100(1): 159-166.

[29] Stoll G, Zitvogel L, Kroemer G. Differences in the composition of the immune infiltrate in breast cancer, colorectal carcinoma, melanoma and non-small cell lung cancer: a microarray-based meta-analysis[J]. OncoImmunology, 2016, 5(2): e1067746.

[30] Sabattini E, Bacci F, Sagramoso C, et al. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview [J]. Pathologica, 2010, 102(3): 83-87.

（编辑：谭斯允）