(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119832995 A
(43)申请公布日 2025.04.15

(21)申请号 202510305223.7

(22)申请日 2025.03.14

(71)申请人 中国科学院合肥物质科学研究院
地址 230031 安徽省合肥市蜀山湖路350号

(72)发明人 王宏志 孙晓君 张帆 谷红仓

(74)专利代理机构 北京科迪生专利代理有限责任公司 11251
专利代理师 赵磊

(51)Int.Cl.
*G16B 40/20*(2019.01)

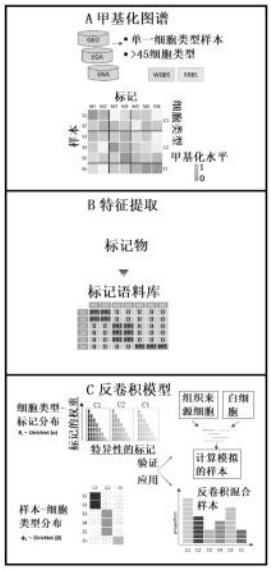
权利要求书2页 说明书5页 附图3页

(54)发明名称

一种基于主题模型的DNA甲基化测序数据反卷积方法

(57)摘要

本发明公开了一种基于主题模型的DNA甲基化测序数据反卷积方法,属于细胞类型反卷积方法技术领域。本方法通过LDA算法构建两个狄利克雷分布,进而模拟样本与细胞类型,以及细胞类型和标记区域之间的分布关系,实现可靠的细胞类型组成预测。构建METRIC的目的是解决较高稀疏性的DNA甲基化测序数据可能带来的反卷积精度降低等问题,实现高效可靠的细胞类型反卷积。




CN 119832995 A

1.一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于,包括以下步骤:

步骤一、对训练集中样本进行预处理,获取样本在每个CpG位点的甲基化水平,结合样本的细胞类型分组鉴定差异甲基化区域及差异甲基化分数;

步骤二、筛选细胞类型特异性标记区域,计算标记区域的差异非甲基化指数,构建出差异非甲基化指数训练矩阵;

步骤三、将样本视为文档,将细胞类型视为主题,将标记视为单词,利用隐含狄利克雷分配训练步骤二得到的矩阵,优化两个概率分布:样本~细胞类型分布和细胞类型~标记分布,设定推断样本和标记的细胞类型标签;

步骤四、对超参数α和β进行调优,训练预测训练集样本的细胞组分,训练完毕后,保存METRIC模型和参数;

步骤五、使用步骤四中训练好的METRIC模型进行预测,使用不包括在训练集样本中的来自组织的测序样本,与白细胞样本按照已知比例混合构建模拟测试样本,提取步骤二中鉴定得到的细胞类型特异性的标记区域数据输入模型,即可得到与训练集样本包含的细胞类型数量一致的细胞类型组分比例。

2.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤一中,训练集中的样本总数为$N=\{N_1,N_2,\cdots,N_j,\cdots,N_n\}$,$N_j(1\leqslant j\leqslant n)$,包含细胞类型种类数为$K=\{K_1,K_2,\cdots,K_i,\cdots,K_k\}$,$K_i(1\leqslant i\leqslant k)$,通过wgbstools工具鉴定识别得到差异甲基化区域,这些区域是相对于其他组别均具有特异性高甲基化或者特异性低甲基化的。

3.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤二中,综合考虑所有组别的差异高甲基化和差异低甲基化区域的数量和差异分数,组内按照差异分数从大到小排序,提取每组内的前25个差异低甲基化区域作为标记区域,得到标记区域集$T=\{T_1,T_2,\cdots,T_t,..T_s\}$,$Tt(1\leqslant t\leqslant s)$;利用wgbstools分析得到训练样本在每个标记区域的低甲基化reads占比分数U-score,然后,对于每一个细胞类型的特异性标记区域,其对应的相同细胞类型的样本的U-score乘以该标记的差异分数得到差异非甲基化指数训练矩阵。

4.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤三中,每一个样本中的细胞类型遵循多项分布:$C\sim \mathrm{Multinomial}(\theta_N)$,$C_{j,t}$表示细胞类型矩阵,对于每一个样本中的标记区域其遵循多项分布:$M\sim \mathrm{Multinomial}(\varphi_K)$,$M_{j,t}$是标记矩阵;样本j中细胞类型的分布由具有超参数α的Dirichlet分布得出:$\theta_N$~Dirichlet(α),其中超参数α控制样本中细胞类型的分布;细胞类型i中的标记区域分布由具有超参数β的Dirichlet 分布得出:$\varphi_k$~Dirichlet(β),其中超参数β控制每个细胞类型中的标记区域分布;则一个样本其LDA概率公式为:

$$p(M,C,\theta,\varphi|\alpha,\beta)=\prod_{j=1}^{N}p(\theta_j|\alpha)\prod_{i=1}^{K}p(\varphi_i|\beta)\prod_{t=1}^{T}p(C_{j,t}|\theta_j)p(M_{j,t}|\varphi_{c_{j,t}});$$

其中,M表示代样本总数,C表示细胞类型总数,$p(\theta_j|\alpha)$表示基于超参数α的样本j中细胞类型分布的概率;$p(\varphi_i|\beta)$表示基于超参数β的细胞类型i中标记物分布的概率;$p(C_{j,t}|\theta_j)$表示样本j的细胞类型分布$\theta_j$下,细胞类型$C_{j,t}$的分配给样本j的概率;$p(M_{j,t}|\varphi_{c_{j,t}})$表示在细胞类型$C_{j,t}$和对应的标记物分布$\varphi_{c_{j,t}}$下,标记$M_{j,t}$在样本j中的出现概率。

5.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤四中,步骤三中所构建的模型中的超参数α设置为auto,使其自适应数据和其他参数选择最优的值,超参数β是一个T*N的矩阵,其行与筛选得到的细胞类型特异性标记区域集对应,列与训练集样本对应,在矩阵中初始值均为0.0001,然后,对于每一个细胞类型的标记区域,其相同细胞类型样本的值设为100000;训练输出两个结果矩阵,一个是细胞类型中的标志区域分布,其描述在K个细胞类型中标记区域的分布和权重,通过统计每个训练得到的聚类簇中聚集的标记区域的特异性细胞类型来源,占比最大的特异性细胞类型就是这个聚类簇的标签,训练中每个聚类簇中均是来自相同特异性细胞类型的标记,另外一个是N个样本的K个细胞类型分布,描述在每个样本中,不同细胞类型的相对比例,对训练结果中的两个输出进行评估,用熵值衡量每个组的细胞类型单一程度,综合选择熵值最小的模型,训练后将训练好的METRIC模型和参数保存到文件中。

6.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤五中,使用不包括在训练集中的来自人类组织中的细胞类型测序样本,将其与正常人的白细胞测序样本按照设定比例进行混合,混合比例即为真实比例,记为Actual=(0%,0.01%,0.03%,0.1%,0.3%,1%,3%,10%,15%,20%,25%,30%),每个比例重复三次,从而得到一批模拟测试样本,然后,对这些模拟样本,提取步骤二中鉴定筛选得到的标记区域集区域的U-score,将数据输入到步骤四中训练好的模型中,即可预测每个模拟样本中的细胞组分,从预测结果中提取模拟样本混合的真实细胞类型对应的预测分数,记为Predict;使用皮尔逊相关性系数、决定系数、均方根误差和准确分数作为评估模型性能的指标。

7.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤五中,采用PCC来衡量两个变量之间的线性关系强度,$R^2$来量化模型对数据变异性的解释程度,以及RMSE来评估模型预测值与实际值之间的差异大小,其中,$R^2$为决定系数,RMSE为决定系数。

8.根据权利要求1所述的一种基于主题模型的DNA甲基化测序数据反卷积方法,其特征在于:在步骤五中,定义了新的指标准确分数AccuracyScore(AS)=(rank(PCC)+rank($R^2$)+rank(RMSE))/3来评价模型反卷积预测的准确性。

9.一种计算机可读存储介质,其特征在于,其上存储有可执行指令,该指令被处理器执行时使处理器实现权利要求1至8任一项所述的方法。

10.一种电子设备,其特征在于,包括:一个或多个处理器;存储器,用于存储一个或多个程序,其中,当所述一个或多个程序被所述一个或多个处理器执行时,使得所述一个或多个处理器实现权利要求1至8任一项所述的方法。

# 一种基于主题模型的DNA甲基化测序数据反卷积方法

## 技术领域

[0001] 本发明属于细胞类型反卷积方法。具体涉及一种基于主题模型的DNA甲基化测序数据反卷积方法。

## 背景技术

[0002] 人体内因细胞凋亡等原因释放到血液中的核小体大小cfDNA(Cell-free DNA)片段携带着丰富的表观遗传信息,包括片段化、组蛋白标记和甲基化模式。不同细胞类型、组织甚至癌症中能观察到群体特异性的DNA甲基化模式。利用细胞特异性或组织特异性DNA甲基化模式作为生物标志物,可构建适用于反卷积方法的参考图谱,针对DNA甲基化数据开发的有参考矩阵的反卷积方法有:MethAtlas、UXM、CelFiE和、CelFEER、cfSort等。由于基因组中的大量胞嘧啶要么没有被测序覆盖,要么覆盖度低于3×,DNA甲基化数据通常具有稀疏性。原因在于全基因组亚硫酸盐测序(Whole Genome Bisulfite Sequencing,WGBS)的高成本使得实现足够的深度变得非常困难,因此往往包含大量的CpG缺失区域,许多已发表的数据测序深度不足30×,且只有两个重复。而简化表示亚硫酸盐测序(Reduced Representation Bisulfite Sequencing,RRBS)缺乏非CpG密集区域的覆盖。类似地,自然语言处理中的文本数据经常表现出高水平的稀疏性,而潜在狄利克雷分配(Latent Dirichlet Allocation,LDA)是文本挖掘中最流行的主题建模方法之一,它对稀疏性的数据具有良好的分析表现能力,通常用于无监督的主题发现。基于LDA的方法已经应用于非DNA甲基化数据的分析。例如,基于scATAC-seq开发的cisTopic,基于空间转录组学的STRIDE、STdeconvolve,基于RNA-seq的GLDADec、GTM-decon等。尽管适用于DNA甲基化数据的细胞类型的反卷积方法已经有被提出,但是针对DNA甲基化数据的对稀疏性数据友好的反卷积方法仍未提出。

## 发明内容

[0003] 针对背景技术所述内容和相关问题,本发明提供了一种基于主题模型的DNA甲基化测序数据反卷积方法,本方法通过LDA算法构建两个狄利克雷分布,进而模拟样本与细胞类型,以及细胞类型和标记区域之间的分布关系,实现可靠的细胞类型组成预测。构建METRIC的目的是解决较高稀疏性的DNA甲基化测序数据可能带来的反卷积精度降低等问题,实现高效可靠的细胞类型反卷积。

[0004] 为了上述目的,本发明提供如下技术方案:

[0005] 一种基于主题模型的DNA甲基化测序数据反卷积方法,包括以下步骤:

[0006] 步骤一、对训练集中样本进行预处理,获取样本在每个CpG位点的甲基化水平,结合样本的细胞类型分组信息,鉴定不同组别之间的差异甲基化区域以及每个区域的差异甲基化分数。

[0007] 步骤二、利用步骤一得到的差异甲基化区域,结合其特异性得分筛选出细胞类型特异性的标记区域,计算标记区域的差异非甲基化指数(differential unmethylation

index,DUI),构建出可以表征细胞类型特异性且具有最优标记数量的差异非甲基化指数训练矩阵。

[0008] 步骤三、将样本视为文档,将细胞类型视为主题,将标记视为单词,利用隐含狄利克雷分配训练步骤二得到的矩阵,优化两个概率分布:样本~细胞类型分布和细胞类型~标记分布,设定推断样本和标记的细胞类型标签。

[0009] 步骤四、对于步骤三所构建的模型,对其中两个狄利克雷分布的超参数α和β进行调优,并设计细胞类型标签自动分配的方法,训练预测训练集样本的细胞组分。训练完毕后,保存训练好的METRIC模型和参数。

[0010] 步骤五、使用步骤四中训练好的METRIC模型进行预测,使用不包括在训练集样本中的来自组织的测序样本,与白细胞样本按照已知比例混合构建模拟测试样本,提取步骤二中鉴定得到的细胞类型特异性的标记区域数据输入模型,即可得到与训练集样本包含的细胞类型数量一致的细胞类型组分比例。

[0011] 作为优选,本发明所采用的步骤一的具体实现方式是:训练集中的样本总数为$N=\{N_1,N_2,\cdots,N_j,\cdots,N_n\}$,$N_j(1\leqslant j\leqslant n)$,包含细胞类型种类数为$K=\{K_1,K_2,\cdots,K_i,\cdots,K_k\}$,$K_i(1\leqslant i\leqslant k)$,通过wgbstools工具鉴定识别得到差异甲基化区域,这些区域是相对于其他组别均具有特异性高甲基化或者特异性低甲基化的。

[0012] 作为优选,步骤二的具体实现方式是:综合考虑所有组别的差异甲基化高甲基化和低甲基化区域的数量、差异分数等信息,组内按照差异分数从大到小排序,提取每组内的前25个差异低甲基化区域作为标记区域,得到标记区域集$T=\{T_1,T_2,\cdots,T_t,..T_s\}$,$T_t(1\leqslant t\leqslant s)$。利用wgbstools分析得到训练样本在每个标记区域的低甲基化reads占比分数U-score,然后,对于每一个细胞类型的特异性标记区域,其对应的相同细胞类型的样本的U-score乘以该标记的差异分数得到DUI,则得到差异非甲基化指数训练矩阵。

[0013] 作为优选,步骤三的具体实现方式是:每一个样本中的细胞类型遵循多项分布:$C\sim Multinomial(\theta_N)$,$C_{j,t}$表示细胞类型矩阵。对于每一个样本中的标记区域其遵循多项分布:$M\sim Multinomial(\varphi_K)$,$M_{j,t}$是标记矩阵。样本j中细胞类型的分布由具有超参数α的Dirichlet分布得出:$\theta_N\sim Dirichlet(\alpha)$,其中超参数α控制样本中细胞类型的分布。细胞类型i中的标记区域分布由具有超参数β的Dirichlet分布得出:$\varphi_k\sim Dirichlet(\beta)$,其中超参数β控制每个细胞类型中的标记区域分布。则一个样本其LDA概率公式为:

[0014] $$p(M,C,\theta,\varphi|\alpha,\beta)=\prod_{j=1}^{N}p(\theta_j|\alpha)\prod_{i=1}^{K}p(\varphi_i|\beta)\prod_{t=1}^{T}p(C_{j,t}|\theta_j)p(M_{j,t}|\varphi_{c_{j,t}})。$$

[0015] 其中,M表示代样本总数,C表示细胞类型总数,$p(\theta_j|\alpha)$表示基于超参数α的样本j中细胞类型分布的概率。$p(\varphi_i|\beta)$表示基于超参数β的细胞类型i中标记物分布的概率。$p(C_{j,t}|\theta_j)$表示样本j的细胞类型分布$\theta_j$下,细胞类型$C_{j,t}$的分配给样本j的概率。$p(M_{j,t}|\varphi_{c_{j,t}})$表示在细胞类型$C_{j,t}$和对应的标记物分布$\varphi_{c_{j,t}}$下,标记$M_{j,t}$在样本j中的出现概率。

[0016] 作为优选,步骤四的具体实现方式是:步骤三中所构建的模型中的超参数α设置为auto,使其自适应数据和其他参数选择最优的值。超参数β是一个T*N的矩阵,其行与筛选得到的细胞类型特异性标记区域集对应,列与训练集样本对应,在矩阵中初始值均为0.0001,然后,对于每一个细胞类型的标记区域,其相同细胞类型样本的值设为100000。训练输出两个结果矩阵,一个是细胞类型中的标志区域分布,其描述在K个细胞类型中标记区域的分布

和权重,通过统计每个训练得到的聚类簇中聚集的标记区域的特异性细胞类型来源,占比最大的特异性细胞类型就是这个聚类簇的标签,训练中每个聚类簇中均是来自相同特异性细胞类型的标记。另外一个是N个样本的K个细胞类型分布,描述在每个样本中,不同细胞类型的相对比例。对训练结果中的两个输出进行评估,用熵值衡量每个组的细胞类型单一程度,综合选择熵值最小的模型,训练后将训练号的METRIC模型和参数保存到文件中。

[0017] 作为优选,步骤五的具体实现方式是:使用不包括在训练集中的来自人类组织的细胞类型测序样本,将其与正常人的白细胞测序样本按照设定比例进行混合,混合比例即为真实比例,记为Actual=(0%,0.01%,0.03%,0.1%,0.3%,1%,3%,10%,15%,20%,25%,30%),每个比例重复三次,从而得到一批模拟测试样本,然后,对这些模拟样本,提取步骤二中鉴定筛选得到的标记区域集区域的U-score,将数据输入到步骤四中训练好的模型中,即可预测每个模拟样本中的细胞组分,从预测结果中提取模拟样本混合的真实细胞类型对应的预测分数,记为Predict。皮尔逊相关性系数(Pearson Correlation Coefficient, PCC)、决定系数(Coefficient of Determination, $R^2$)、均方根误差(Root Mean Square Error, RMSE)是评估统计模型性能的常用指标。本方法采用PCC来衡量两个变量之间的线性关系强度,$R^2$来量化模型对数据变异性的解释程度,以及RMSE来评估模型预测值与实际值之间的差异大小。同时,为了综合评估模型在所有指标上的表现,本方法定义了一个新的指标准确分数AccuracyScore(AS)=(rank(PCC)+rank($R^2$)+rank(RMSE))/3来评价模型反卷积预测的准确性,这些指标共同提供了对模型性能的全面评估。通过比较METRIC与其他几个方法在这四个评价指标上的综合表现,表明METRIC具有良好的反卷积性能。

[0018] 一种计算机可读存储介质,其上存储有可执行指令,该指令被处理器执行时使处理器实现上述方法。

[0019] 一种电子设备,包括:一个或多个处理器;存储器,用于存储一个或多个程序,其中,当所述一个或多个程序被所述一个或多个处理器执行时,使得所述一个或多个处理器实现上述方法。

[0020] 有益效果:

[0021] 本发明与现有基于DNA甲基化数据的反卷积方法相比,其优点在于,它是一种新的基于主题模型的反卷积方法,METRIC可以同时分析细胞组成比例和细胞特异性DNA甲基化标记物的权重,具有的较好可解释性和可靠性。对于稀疏性高的数据,具有很好的反卷积性能。

## 附图说明

[0022] 图1为METRIC模型结构示意图。

[0023] 图2为METRIC反卷积模拟测试样本结果图-乳腺上皮细胞。

[0024] 图3为METRIC反卷积模拟测试样本结果图-肝细胞。

[0025] 图4为METRIC反卷积模拟测试样本结果图-绒毛滋养细胞。

## 具体实施方式

[0026] 下面结合附图及具体实施例详细介绍本发明。但以下的实施例仅限于解释本发明,本发明的保护范围应包括权利要求的全部内容,而且通过以下实施例的叙述,本领域的

技术人员是可以完全实现本发明权利要求的全部内容。

[0027] 实施例

[0028] 下面结合附图1-2、附表1、以及实例对本发明进行阐述,此处实例仅用于解释本发明,并不限定本发明。

[0029] 图1展示了基于主题模型的DNA甲基化测序数据反卷积方法METRIC的模型结构示意图,A表示甲基化图谱,B表示特征提取,C表示反卷积模型,首先,从公共数据集收集到人类不同组织来源的多种细胞类型DNA甲基化测序数据样本,然后进行差异甲基化分析,得到细胞类型特异性的标记区域,将这些标记构建标记池,消除标记在基因组上的顺序等影响,筛选出每组前25个标记,构建差异非甲基化分数矩阵,输入到模型中进行训练。模型训练后将输出每个样本的分布,包含不同细胞类型的组成比例,以及标记区域的聚类结果,每个聚类簇表示一个细胞类型,其中每个标记具有不同的权重。然后通过反卷积按照已知比例生成模拟样本,进行模型性能验证和比较。最后,可以将METRIC应用于新的样本进行反卷积分析。

[0030] 图2、图3和图4展示了METRIC应用在三组模拟测试样本上的反卷积结果,从21个不同的公共数据集中收集到575例RRBS测序样本,包含13中不同的细胞类型,在下载这些样本的原始数据后,进行数据质控、比对到hg19参考基因组、甲基化提取等分析,然后使用这些样本进行模型训练和测试,以下是分析的具体步骤:

[0031] 步骤一、模拟测试样本生成:在所收集的RRBS测试集样本中,随机选择了三个细胞类型中的6个样本:乳腺上皮细胞*2、肝细胞*3、绒毛滋养细胞*1,将其分别与一个白细胞的RRBS测序样本进行混合,混合比例设置为0%,0.01%,0.03%,0.1%,0.3%,1%,3%,10%,15%,20%,25%,30%,记为Actual,每一种混合比例重复三次,总共合成6*12*3=216个模拟测试样本。

[0032] 步骤二、训练集数据处理:将步骤一中挑选出的6个样本去除,使用RRBS样本中剩余的样本构建训练集,对这些样本进行差异甲基化分析,鉴定出13种细胞类型特异性的DNA甲基化标记区域,按照甲基化差异水平排序后,筛选每种细胞类型前25个标记,共计得到325个标记,构成标记区域集,构建差异非甲基化指数矩阵作为模型训练输入。此外提取216个测试样本在标记区域集的非甲基化reads比例构成测试矩阵,用于模型预测。

[0033] 步骤三、训练METRIC:将步骤二中构建好的训练数据矩阵输入METRIC进行训练,此时训练的总样本数N=569,细胞类型数K=13,标记数量为T=325,训练过程中超参数α设为auto,超参数β为一个T*N的矩阵,标记样本在标记区域的值设为100000,其他值设为0.00001,passes=20,iterations=200,训练完毕,保存训练好的模型和参数。

[0034] 步骤四、使用训练好的模型进行预测:使用训练好的模型对模拟测试样本的细胞类型组成进行反卷积,得到每个样本的反卷积结果,包含每个细胞类型的组成比例。对于由乳腺上皮细胞与白细胞混合构成的模拟样本,提取乳腺上皮细胞的细胞组分预测值,构成Predict,肝细胞和绒毛滋养细胞采用相同的处理。

[0035] 步骤五、METRIC模型性能评估:对于3种细胞类型的模拟样本的反卷积结果进行评估,将2个乳腺上皮细胞的结果取平均,同理也对3个肝细胞的模拟样本的结果取平均,得到平均Predict值,计算Actual-Predict之间的三个评估分数皮尔逊相关性系数Pearson correlation coefficient,决定系数$R^2$,均方根误差RMSE,绘制Actual和Predict折线图,

并在图上标出三个系数的数值。图2表明本方法在模拟测试样本上具有良好的反卷积效果，PCC均值为0.99，$R^2$的均值为0.72，RMSE的均值为0.05。

[0036] 表1 METRIC与cfSort，UXM，MethAtlas反卷积方法的性能对比表

| 比较场景 | 模型 | 乳腺上皮细胞（Berast-Ep） | | | | 肝细胞（Hepatocyte） | | | | 绒毛滋养细胞（Trophoblast） | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PCC | $R^2$ | RMSE | AS | PCC | $R^2$ | RMSE | AS | PCC | $R^2$ | RMSE | AS |
| | METRIC | 0.998 | 0.928 | 0.029 | 4 | 0.987 | 0.303 | 0.025 | 3.33 | 0.988 | 0.943 | 0.025 | 2 |
| 场景一 | cfSort | 0.983 | 0.099 | 0.101 | 2 | 0.991 | -0.257 | 0.119 | 2.67 | - | - | - | - |
| | UXM | -0.675 | -0.542 | 0.132 | 1 | 0.962 | -0.389 | 0.125 | 1 | - | - | - | - |
| 场景二 | UXM | 0.999 | 0.829 | 0.044 | 3.67 | 0.990 | 0.504 | 0.075 | 3.33 | 0.996 | 0.420 | 0.081 | 1.33 |
| | MethAtlas | 0.998 | 0.970 | 0.018 | 4.67 | 0.993 | 0.655 | 0.062 | 4.67 | 0.996 | 0.990 | 0.010 | 2.33 |

[0037] 表1展示了其他三种反卷积方法在模拟测试样本上的反卷积比较结果。使用cfSort、MethAtlas、UXM三种方法，在两个不同的比较场景下进行模型性能评估。以下是三种方法反卷积模拟测试样本的具体步骤：

[0038] (1)使用cfSort反卷积模拟测试样本：将测试样本转换成cfSort要求的.tfrecords输入格式，使用cfSort训练好的两个模型DNN1(深度神经网络1)、DNN2(深度神经网络2)对测试样本进行反卷积，计算实际值和预测值之间的相关性指标分数：PCC(皮尔逊相关性系数)，$R^2$(决定系数)，RMSE(均方根误差)，AS值(准确度分数)。

[0039] (2)使用MethAtlas反卷积模拟测试样本：首先提取训练样本在标记区域的DNA甲基化水平，按照细胞类型分组取平均得到反卷积参考矩阵，然后使用MethAtlas进行反卷积，计算实际值和预测值之间的相关性指标分数：PCC，$R^2$，RMSE，AS。

[0040] (3)使用UXM反卷积模拟测试样本：首先，使用UXM的默认设置，去反卷积模拟测试样本。其次，使用所有训练集样本构建一个符合UXM要求的新的参考矩阵，然后再去反卷积模拟测试样本，计算实际值和预测值之间的相关性指标分数：PCC，$R^2$，RMSE，AS。

[0041] (4)按照两个预设场景对METRIC和cfSort、UXM、MethAtlas进行比较。场景一：与使用默认设置进行反卷积分析的两个工具cfSort、UXM进行比较，由于这两个方法没有将绒毛滋养细胞(Trophoblast)纳入参考矩阵，所以没有这个细胞类型的组分比例。从表1中结果可知，METRIC相比于cfSort、UXM这两个方法，具有更好的反卷积性能。场景二：使用训练数据重新构建符合UXM、MethAtlas的参考矩阵然后再对模拟测试样本进行反卷积分析，从表1中可知，METRIC相比于另外两个方法反卷积性能居中。综合各项结果表明使用本发明收集的训练队列样本进行训练矩阵构建训练的反卷积模型，能够提升反卷积准确性。

[0042] 以上所述仅是本申请的具体实施方式，使本领域技术人员能够理解或实现本申请。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所申请的原理和新颖特点相一致的最宽的范围。

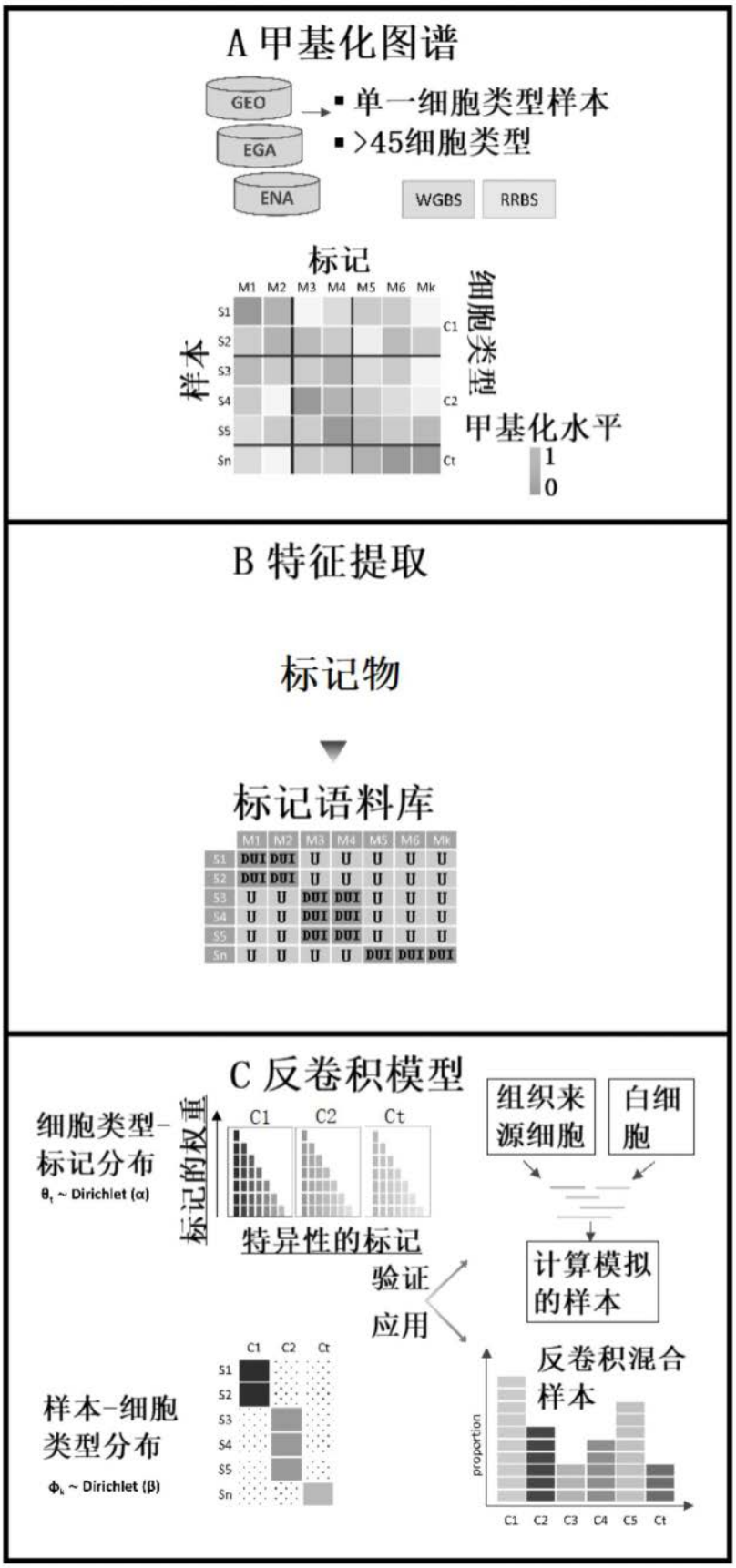

A 甲基化图谱
GEO
EGA
ENA
▪ 单一细胞类型样本
▪ >45细胞类型
WGBS
RRBS
标记
M1 M2 M3 M4 M5 M6 Mk
S1 S2 S3 S4 S5 Sn
C1 C2 Ct
样本
细胞类型
甲基化水平
1
0
B 特征提取
标记物
标记语料库
C 反卷积模型
细胞类型–标记分布
θt ~ Dirichlet (α)
标记的权重
C1 C2 Ct
特异性的标记
组织来源细胞
白细胞
计算模拟的样本
验证
应用
反卷积混合样本
proportion
C1 C2 C3 C4 C5 Ct
样本–细胞类型分布
Φk ~ Dirichlet (β)
S1 S2 S3 S4 S5 Sn

图1

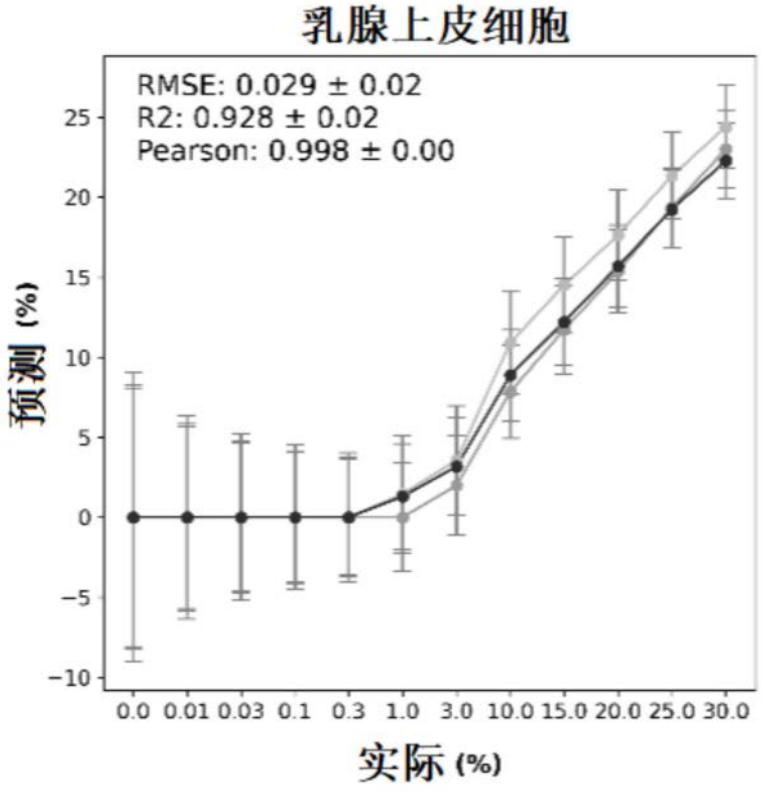

乳腺上皮细胞
RMSE: 0.029 ± 0.02
R2: 0.928 ± 0.02
Pearson: 0.998 ± 0.00
预测 (%)
25
20
15
10
5
0
−5
−10
0.0 0.01 0.03 0.1 0.3 1.0 3.0 10.0 15.0 20.0 25.0 30.0
实际 (%)

图2

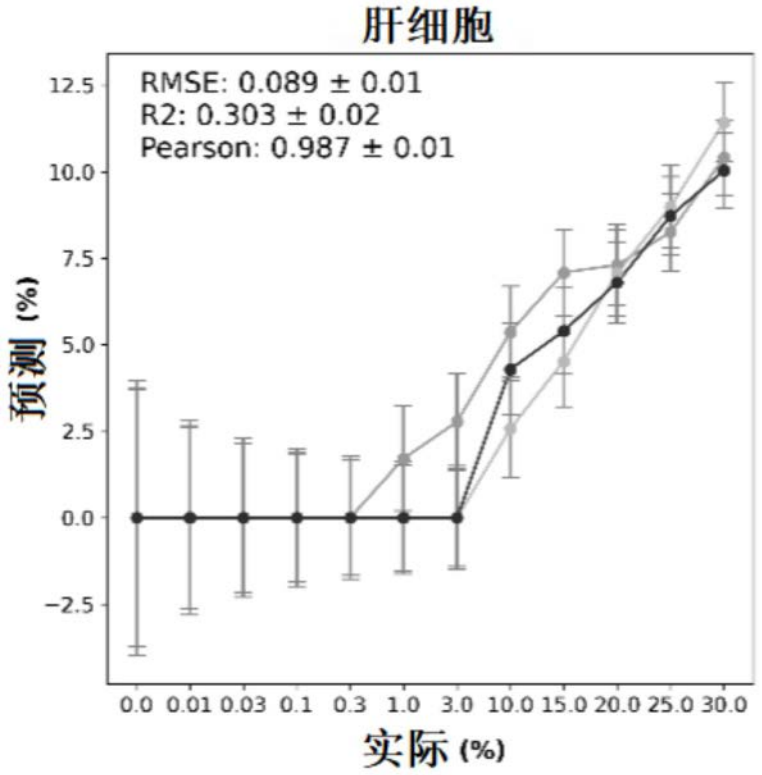

肝细胞
RMSE: 0.089 ± 0.01
R2: 0.303 ± 0.02
Pearson: 0.987 ± 0.01
预测 (%)
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
0.0 0.01 0.03 0.1 0.3 1.0 3.0 10.0 15.0 20.0 25.0 30.0
实际 (%)

图3

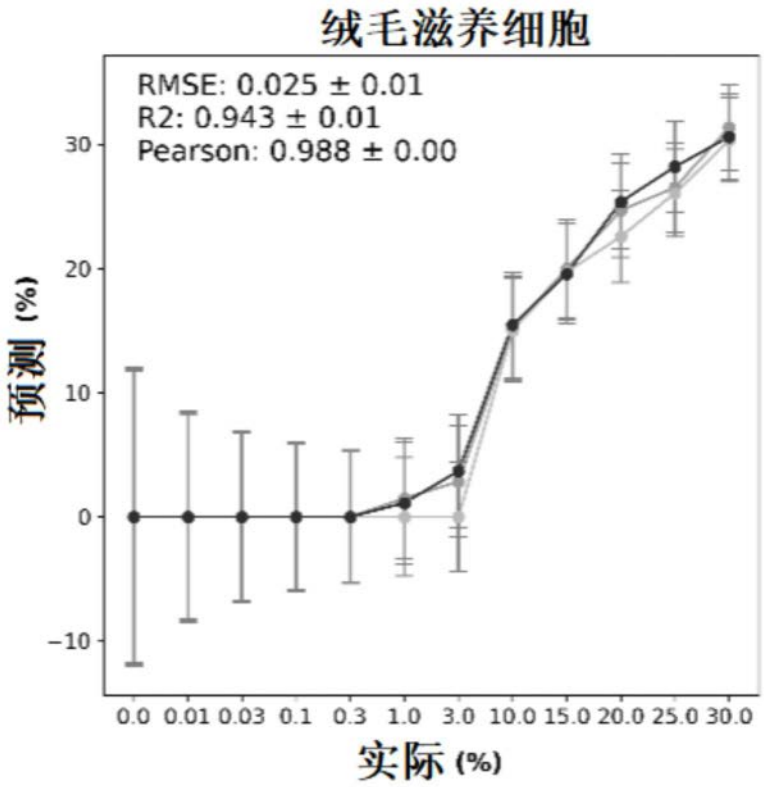

绒毛滋养细胞
RMSE: 0.025 ± 0.01
R2: 0.943 ± 0.01
Pearson: 0.988 ± 0.00
预测 (%)
30
20
10
0
−10
0.0 0.01 0.03 0.1 0.3 1.0 3.0 10.0 15.0 20.0 25.0 30.0
实际 (%)

图4