融合多平台表达数据的转录组差异表达分析

王凯莉[1)] 张 礼[2)] 刘学军[1)]

[1)](南京航空航天大学计算机科学与技术学院 南京 211106)

[2)](南京林业大学信息科学技术学院 南京 210037)

摘 要 差异表达分析是转录组研究的基本目标之一,对揭示基因功能和调控规律以及选择性剪切的波动变化具有重要作用.基因芯片与 RNA-Seq 是当前主流的测量转录组表达水平的两种实验平台,并被广泛应用于转录组差异表达分析.随着测序技术的发展,测序成本不断降低,许多研究采用多种测量平台以获得更为准确的结果.当前公共数据库中积累了大量的基因芯片和 RNA-Seq 表达数据,为多平台转录组数据分析提供了研究空间.研究表明:融合多平台表达数据能够提高差异表达分析的准确性和可靠性.大多数现有的融合多平台表达数据的差异检测研究主要对多种类型的基因芯片表达数据进行融合,较少考虑 RNA-Seq 表达数据.并且现有方法忽略了很多有用的信息,例如测量误差和重复实验产生的波动性.针对现有方法存在的问题,该文提出了融合多平台转录组数据的差异检测模型 mpDE(multi-platform Differential Expression model),寻找差异表达的基因和异构体.该模型将不同实验平台的表达数据和表达水平的技术性测量误差融入到模型中,同时考虑了同一平台在不同条件下的生物重复或技术重复的波动性,从而提高差异检测准确度.该文将 mpDE 应用到两个人类多平台表达数据集进行差异表达检测,涉及了 Affymetrix 的传统 3' 芯片、外显子芯片、HTA2.0 芯片,RNA-Seq 四种常用的转录组表达水平测量平台.该文将 mpDE 计算结果与单平台的差异检测结果和其他多平台表达数据融合算法进行了对比.实验结果表明,mpDE 获得了更为准确的差异检测结果,差异基因检测准确率与以往方法相比提高了 2%~8%;差异异构体检测准确率提高了 1%~15%.

关键词 多平台;数据融合;差异表达分析;基因芯片;RNA-Seq;生物信息学

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2018.01415

Differential Expression Analysis Based on Integrating Transcriptome Expression Data From Multiple Platforms

WANG Kai-Li[1)] ZHANG Li[2)] LIU Xue-Jun[1)]

[1)](*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106)

[2)](*College of Information Science and Technology, Nanjing Forestry University, Nanjing* 210037)

Abstract From the expression point of view, differential expression analysis is one of the basic research objectives in the transcriptome study. The differential expression analysis of genes is an important method to understand the function of gene and reveal the gene regulating mechanism. The differential expression analysis of isoforms also plays an important role in revealing the change of alternative splicing. Microarray and high-throughput sequencing technology, RNA-Seq, are the two main technologies for measuring transcriptome expression levels and widely used in differential expression analysis of transcriptome study. With the development of sequencing technology and the reduction in the cost of sequencing, many studies usually use multiple experimental platforms to obtain more accurate and reliable results in the transcriptome study. Now public repositories have accumulated a large amount of microarray and RNA-Seq expression data,

收稿日期:2016-06-03;在线出版日期:2017-04-11.本课题得到国家自然科学基金(61170152)和软件新技术与产业化协同创新中心资助.
王凯莉,女,1990 年生,硕士研究生,主要研究方向为生物信息学.E-mail:nuaa0410402@126.com.**张 礼**,男,1985 年生,博士,主要研究方向为生物信息学.**刘学军**(通信作者),女,1976 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为生物信息学、机器学习.E-mail:xuejun.liu@nuaa.edu.cn.

which provide the possibility of multi-platform expression data analysis. And some researchers have shown that integrating data from multi-platform can increase the statistical power and reliability in expression analysis of transcriptome. Although there are some achievements in the research of the differential expression analysis by integrating expression data from multiple platforms, most of the studies focus on expression data from various types of microarrays, and little work considers RNA-Seq data. Moreover, many existing methods ignore useful information, such as the technical measurement error of expression estimates under different platforms. Currently, most experiments involve biological or technical replicates for obtaining a level of uncertainty of the measured expression. Due to the experimental environment, sample preparation and other factors, the repeated experiments will produce volatility under the same experimental platform. However, most of existing methods do not consider these problems. This paper proposes a new model, mpDE (multi-platform Differential Expression model), for differential expression detection by integrating expression data from multiple platforms. The mpDE method integrates the expression data and the associated measurement error from different platforms and considers the variability of biological replicates or technical replicates under different conditions for the same platform to improve the accuracy of differential expression detection. mpDE can obtain the probability distribution of gene and isoform expression levels, which is independent of experimental platforms and repeated experiments. The independent probability distribution of gene and isoform expression levels can be used in the subsequent gene expression analysis, such as differential expression detection and clustering. This paper applied mpDE to differential expression analysis of two human multi-platform datasets which include expression data obtained from Affymetrix's traditional 3' arrays, exon arrays, HTA2.0 (Human Transcriptome Array 2.0) and RNA-Seq. The performance of mpDE is verified in terms of the accuracy of differential expression analysis of genes and isoforms and the calculation efficiency of the method. We compared the performance of mpDE with each single platform and other multi-platform expression analysis methods, MRS (Median Rank Score) combined limma and RSP (Ranked-based Semi-Parametric). Results show that mpDE is more accurate in differential expression analysis compared with other alternatives. In differential expression analysis of genes, the accuracy of mpDE is increased by 2%－8% compared with the previous approaches. In differential expression analysis of isoforms, the accuracy of mpDE is increased by 1%－15%.

Keywords multi-platform; data integration; differential expression analysis; microarray; RNA-Seq; bioinformatics

1 引 言

转录组差异表达分析是通过分析不同条件下的转录组表达数据，识别发生差异表达(Differential Expression，DE)的基因或异构体，这对揭示基因调控规律或基因选择性剪切的变化具有重要作用. 随着测序技术的发展，测序成本不断降低，许多研究选择采用多种表达水平测量平台进行转录组分析，因此公共数据库(例如 NCBI GEO[1])积累了大量的多平台表达数据. 研究表明，融合多平台数据能够增加差异检测的性能和可靠性[2-4].

基因芯片和基于高通量测序的 RNA-Seq 是转录组研究中两大高通量基因表达水平测量技术. 基因芯片实验一般是将带有荧光标记物的待测样本与芯片上的探针进行杂交. 杂交完成后，处理并扫描芯片得到芯片的荧光图像(如图 1 所示)，最终得到连续的图像灰度值即基因芯片原始数据. 探针的灰度值大小表示了样本中对应 RNA 片段的浓度. RNA-Seq 实验一般将待测样本片段化后反转录成 cDNA，通过桥式扩增等处理后进行测序，之后根据测序顺序将测序结果中同一位置的碱基连成读段(read)，

映射到基因参考序列上的读段计数反应了对应基因的表达水平，如图 2 所示. 利用基因芯片实验产生的探针灰度值以及 RNA-Seq 实验得到的离散的读段计数，均可以计算出相应基因的表达水平. 基因芯片出现时间较早，重复实验较多，而且具有成熟的数据分析方法，但是由于其固有的技术限制，导致背景噪声大而且难以消除. RNA-Seq 测序技术背景噪声小，对于低表达的基因或异构体更加敏感，但很多实验中重复实验少，降低了实验结果的可靠性. 因此如何结合这些平台的特点，系统地分析多平台表达数据并从中获取具有价值的信息是一种挑战，这也促使了许多融合、聚类等算法应运而生.

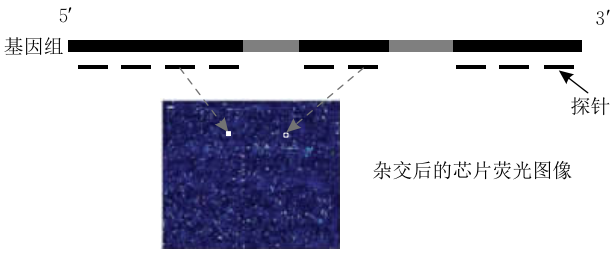


图 1　基因芯片实验的荧光图像示意图

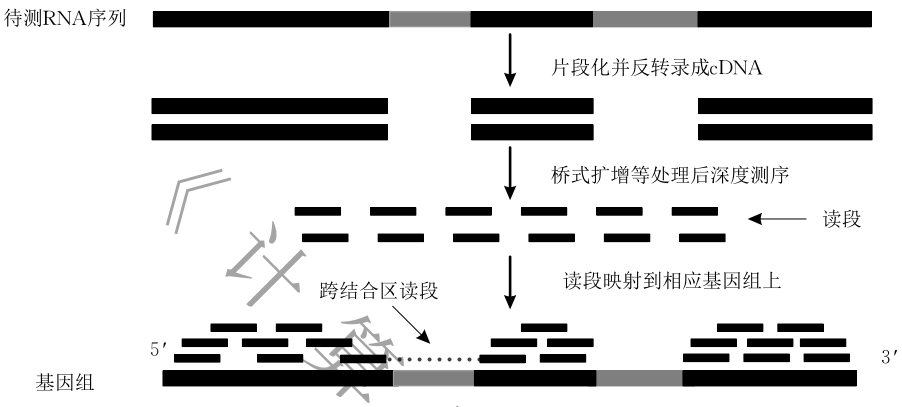


图 2　RNA-Seq 实验读段计数产生过程

目前多平台差异表达分析研究方法主要分为两类，一是采用荟萃分析(meta-analysis)进行差异检测，通过对多个同类研究结果进行合并汇总，增大样本总量，提高检测准确率和统计分析结果的一致性. 例如文献[5]通过对多种基因芯片表达数据分析，发现某些实验平台存在未测量到但可能具有研究价值的基因，因此他们采用荟萃分析框架为这些缺失的基因产生一个元分数(meta-score)，从而能够对多平台的所有基因进行后续差异表达分析. 文献[6]采用荟萃分析对多个基因芯片的骨肉瘤表达数据进行差异表达分析，揭示了差异表达基因对于骨肉瘤与正常组织之间生物变化的影响. 二是通过数据转换的方法，例如中位值排序转换方法(Median Rank Score，MRS)[7]、基于排序融合方法(Rank-Based，RB)[8]等经典方法以及较新的 vitrualArray[9]方法，将多平台表达数据合并成一个表达矩阵，然后采用相应的差异检测方法进行后续差异表达分析.

现有的多平台表达数据分析研究主要通过融合不同平台获得的基因或异构体表达水平，结合现有的差异检测方法寻找差异基因或异构体，大部分方法都是基于点估计值来计算基因或异构体的表达水平，忽略了很多有用的潜在信息，例如不同实验平台下表达水平的技术性测量误差. 另外，由于样本采集、实验环境以及实验平台等因素，导致同一测量平台在不同实验条件下会产生技术重复或生物重复的波动性，降低了差异表达分析的准确性，但是现有的多平台差异表达分析方法均未考虑这些问题. 针对这些存在的问题，本文提出了多平台转录组差异检测模型 mpDE(multi-platform Differential Expression model). 该模型基于多层贝叶斯模型，结合不同的基因表达水平测量平台的特点，对多个平台的表达水平和表达水平的技术性测量误差进行建模，同时考虑了同一测量平台在不同实验条件下的生物重复或技术重复的波动性，最终得到独立于平台、重复实验的基因或异构体表达值的概率分布. mpDE 模型采用 EM 算法和变分方法估计模型中参数的后验分布，并采用 PPLR[10]标准来判断差异表达的基因或异构体. 本文采用两个人类基因表达数据集从基因及异构体差异表达分析两方面来验证 mpDE 模型的性能，并与单平台的差异检测结果、MRS[7]结合 limma[11]方法以及较新的多平台差异基因检测模型 RSP(Ranked-based Semi-Parametric)[12]差异检测的结果进行对比.

本文第 2 节介绍相关工作并对几种融合多平台

表达数据方法进行分析;第 3 节介绍本文提出的 mpDE 模型;第 4 节通过实验将 mpDE 与单平台和其他方法进行对比;第 5 节对本文工作进行总结并展望未来的研究工作.

2 相关工作

基于排名的无参数统计方法 RankProd[13] 是较常用的荟萃分析方法之一,通过对基因在任意两个条件下的 fold-change 对基因排名并采用 p-value 和 FDR(False Discovery Rate)的方法来判断差异基因.与线性模型相比,RankProd 模型简单,对噪声具有一定的鲁棒性.另外,RankProd 方法不需要对原始的多平台表达数据进行归一化处理,通过排名克服多平台表达数据的不一致性,提高实验结果的灵敏度和可靠性.但是该方法使用固定规则的无参数模型,即采用基因在任意两个实验条件下的 fold-change 的排名规则,无法模拟出原始表达数据的结构.

在最新的研究进展中,文献[12]针对多平台表达数据,提出了基于排名的半参数荟萃分析统计模型 RSP,并将其应用到多平台差异基因检测上.RSP 模型采用 Copula 混合模型模拟数据的结构,使其既适应数据结构又对噪声具有很强的鲁棒性.首先该模型通过对基因差异的方向进行建模,模拟了不同平台中基因差异表达方向的偏好.其次,该模型考虑了基因是否差异以及差异的方向(下降规则或上升规则),这种三分量聚类框架更能够适应数据的结构.例如当下降规则差异的基因与上升规则差异的基因的比例不对称时,三分量聚类框架可以反映出这种不对称性.相反,常用的荟萃分析方法(如 RankProd)假设的拒绝域是对称的,处理这种不对称的数据时,其分析结果是不可靠的.最后 RSP 模型与尺度无关的,因此 RSP 非常适合用于不同平台的表达数据融合.但是 RSP 模型仅适用于两个平台的表达数据的融合,并且没有考虑每个条件下技术性重复或生物性重复实验的波动性.而 mpDE 模型能够融合任意多个平台的表达数据,且考虑了重复实验的波动性,应用范围更加广泛.

中位值排序转换方法(MRS)[7] 是较经典的融合多平台表达数据的方法.对原始表达数据进行去噪和归一化等预处理后,该方法首先选择重复实验最多的平台作为对照平台,计算基因在该平台所有样本中的中位值并排名;然后其他平台的基因根据表达值大小的排名替换成对照平台中相应排名的中位值;最后合并所有平台替换后的表达数据,并结合现有的差异检测方法进行后续差异表达分析,差异表达分析一般过程如图 3 所示.中位值排序转换方法能够将不同平台的表达水平归一化至相同的数值范围内,通过增加样本数量,提高后续差异表达分析的准确度,但是该方法依赖于对照平台的表达数据,其准确性对后续差异表达分析会产生一定的影响,而且该方法也改变了基因之间的相关性.

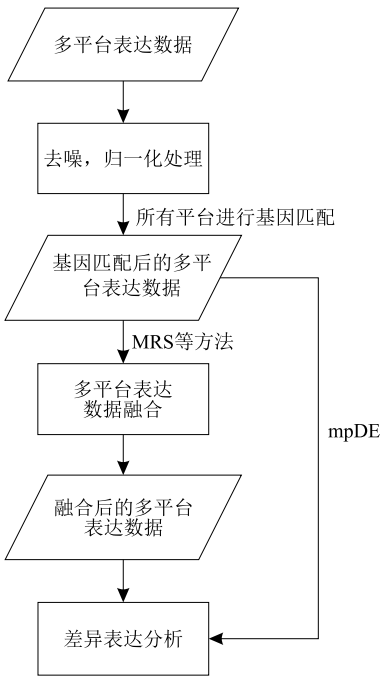


图 3 融合多平台表达数据的差异表达分析的一般过程

基于排序融合方法(RB)[8] 不需要归一化处理原始的表达数据,只需对所有平台的原始表达数据进行去噪等处理.RB 方法对所有平台的每个基因在每个样本中的表达值进行排名,然后用排名替换基因的表达值;最后直接合并所有平台替换后的表达数据,并结合无参打分方法进行差异表达分析.基于排序融合方法相对比较简单,与中位值排序转换方法相比,该方法不需要选择对照平台,但是该方法仅考虑了基因表达值的排名,忽略了基因表达值之间的差别,改变了不同平台的原始表达数据的结构.

vitrualArray[9] 是一个能够实现融合多种类型基因芯片表达数据的 R/Bioconductor 软件包.该软件包首先根据不同平台获得的探针水平表达数据以及探针与基因和异构体之间的映射关系,得到不同平台所获得的基因或异构体的表达值,然后所有平台进行基因或异构体匹配,并将匹配后的基因或异构体的表达值合并成一个表达矩阵(即虚拟阵列).最后采用 EBM(Empirical Bayes Methods)[14] 和 QD(Quantile Discretization)[7] 等方法处理表达矩阵,以减少不同类型芯片表达数据之间的影响,得到

最终的表达矩阵，并结合相应的差异检测方法寻找差异基因或异构体.

由图 3 可知，MRS、RB 与 virtualArray 需要结合相应的差异检测方法寻找差异基因或异构体，应用范围受到一定的限制，而且 MRS 和 RB 方法均改变了基因之间相关性，降低了差异分析结果的可靠性，virtualArray 仅适用于多种类型的基因芯片表达数据的融合. mpDE 模型充分利用实验数据，直接将不同平台测量到的表达值和表达值的技术性测量误差融入模型中，并且能够获得与平台无关的表达值概率分布，差异表达分析结果的可靠性更高.

3　mpDE 模型

本节首先介绍 mpDE 模型，然后给出详细的参数估计过程，最后介绍任意两个实验条件下差异基因和异构体的评估标准.

3.1　mpDE 模型描述

不同实验平台原始数据格式不同，本文的目标是设计一个通用的多平台表达数据融合模型. 所以对于不同的实验数据，我们首先采用各个平台的原始数据处理方法，如 mmgMOS[15]、GME[16]、PGSeq[17]等，得到连续的基因表达水平的概率分布，然后设计多层贝叶斯模型将不同平台获得的基因表达水平概率分布进行融合，以获得最终的平台无关的基因表达水平. 通常假设每个基因或异构体的真实对数表达值服从高斯分布[18-20]，因此 mpDE 模型假设观察到的数据 D 中每个基因或异构体的对数表达值服从高斯分布. 为了提高计算效率，该模型加入了一层隐含的变量，代表真实的对数表达值[21]. 在第 c 个实验条件下第 p 个实验平台的第 r 个重复实验中某个基因或异构体的所观察到的对数表达值 $\hat{x}_{rpc}$ 和真实的对数表达值 x_{rpc} 分布形式如下：

$$\begin{aligned}\hat{x}_{rpc} &\sim N(x_{rpc}, s^2_{rpc})\\ x_{rpc} &\sim N(\mu_{pc}, \lambda^{-1}_{pc})\end{aligned} \tag{1}$$

其中，s^2_{rpc} 表示每个基因或异构体表达值的测量误差，可由原始数据分析方法获得，比如基因芯片原始数据分析方法 RMA[22]、mmgMOS 和 GME 以及 RNA-Seq 原始数据分析方法 PGSeq 等. μ_{pc} 表示在第 c 条件下第 p 个平台测量到的基因或异构体对数表达值的均值，λ^{-1}_{pc} 表示第 c 个条件下第 p 个平台的重复实验之间的方差，模拟同一平台在不同实验条件下的生物重复或技术重复的波动性.

对于参数 μ_{pc} 和 λ_{pc}，mpDE 模型假设两者相互独立，并且假设 μ_{pc} 服从高斯分布，λ_{pc} 服从共轭的伽玛先验分布，形式如下：

$$\begin{aligned}\mu_{pc} &\sim N(\mu_c, \lambda_c^{-1})\\ \lambda_{pc} &\sim Ga(\alpha_{pc}, \beta_{pc})\end{aligned} \tag{2}$$

其中，μ_c 和 λ_c^{-1} 分别表示第 c 个实验条件下基因或异构体的平台无关的对数表达值的均值和所有重复实验之间的方差. 由于式(1)中 λ^{-1}_{pc} 已经考虑了平台有关的方差，覆盖了大多数平台有关的不确定性. 因此为了减少模型隐含变量的个数，控制模型的复杂程度，mpDE 模型假设在 λ_c 不同实验条件下共享，即 $\lambda_c^{-1}=\lambda^{-1}$，因此 μ_{pc} 可以写成如下形式：

$$\mu_{pc} \sim N(\mu_c, \lambda^{-1}) \tag{3}$$

为了方便模型求解，我们在对式(3)进行先验概率假设时考虑到共轭性，假设 μ_c 服从参数不包含任何先验信息的高斯分布，λ 服从共轭的伽玛先验分布，形式如下：

$$\begin{aligned}\mu_c &\sim N(\mu_0, \eta_0^{-1})\\ \lambda &\sim Ga(\alpha_0, \beta_0)\end{aligned} \tag{4}$$

这样 mpDE 模型的隐含变量 $\theta=\{X, U', \lambda', U, \lambda\}$，$X$ 表示所有基因或异构体的真实表达值 x_{rpc}. U' 和 λ' 分别表示所有条件下所有平台的 μ_{pc} 和 λ_{pc}. U 表示所有条件下的 μ_c. 超参数 $\phi=\{\mu_0, \eta_0, \alpha_0, \beta_0, \alpha_{pc}, \beta_{pc}\}$，超参数的个数取决于实验条件和实验平台的个数，模型概率图表示如图 4 所示.

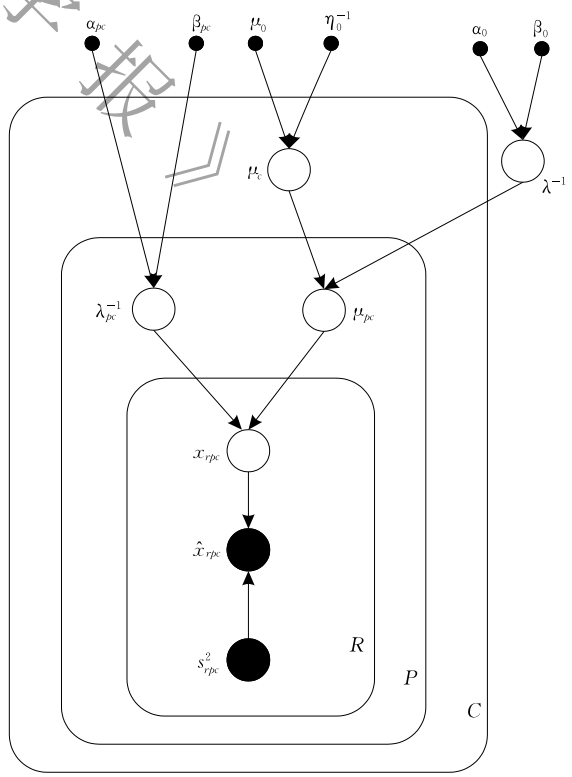


图 4　mpDE 模型概率图表示(C 代表实验条件，P 代表实验平台，R 是在条件 C 下 P 平台的重复实验)

在 mpDE 模型中,我们关心的是第 c 个实验条件下平台无关的表达值后验分布 $P(\mu_c|D,\phi)$. 假设观察到的数据之间是相互独立的,因此数据 D 的似然函数如式(5):

$$P(D|\theta)=\prod_{c=1}^{C}\prod_{p=1}^{P}\prod_{r=1}^{R}P(\hat{x}_{rpc}|x_{rpc},s_{rpc}^2)\propto\prod_{c=1}^{C}\prod_{p=1}^{P}\prod_{r=1}^{R}s_{rpc}^{-1}\exp\left(-\frac{(\hat{x}_{rpc}-x_{rpc})^2}{2s_{rpc}^2}\right) \quad (5)$$

隐含变量 θ 的先验分布如式(6):

$$\begin{aligned}P(\theta|\phi)=&P(\lambda|\alpha_0,\beta_0)\left(\prod_{c}^{C}P(\mu_c|\mu_0,\eta_0^{-1})\right)\cdot\\&\left(\prod_{c}^{C}\prod_{p}^{P}P(\lambda_{pc}|\alpha_{pc},\beta_{pc})\right)\cdot\\&\left(\prod_{c}^{C}\prod_{p}^{P}P(\mu_{pc}|\mu_c,\lambda^{-1})\right)\cdot\\&\left(\prod_{c}^{C}\prod_{p}^{P}\prod_{r}^{R}P(x_{rpc}|\mu_{pc},\lambda_{pc}^{-1})\right)\end{aligned} \quad (6)$$

早期对基因表达水平进行分析的多层贝叶斯模型[18-20]主要解决的是由于单一基因芯片实验平台重复实验次数较少,引起的在基因表达值计算中方差估计过拟合的问题. 相比以往的这些工作,mpDE 有两点改进,一是融合多个实验平台的表达数据,通过式(3)中的正态分布模拟平台无关的表达值分布,有利于充分利用实验数据,提高数据分析精度;另一个是我们先前提出的一系列原始数据分析方法,如 GME[16]、PGSeq[17]等,能够计算出不同实验平台表达水平的技术性测量误差,mpDE 考虑了这些平台相关的技术性测量误差,有利于获得真实的平台无关表达值分布.

3.2 参数估计

mpDE 模型采用 EM 算法结合变分方法优化模型中的隐含变量,通过迭代不断优化 $P(D|\phi)$ 的下界,最后估计出后验分布 $P(\theta|\phi,D)$[23]. 假设模型参数之间相互独立,则变分 EM 算法中隐含变量 θ 的后验分布 $Q(\theta)$ 被分解成式(7):

$$\begin{aligned}Q(\theta)=&Q(X)Q(U')Q(\lambda')Q(U)Q(\lambda)\\=&\prod_{r}^{R}\prod_{p}^{P}\prod_{c}^{C}Q(x_{rpc})\prod_{p}^{P}\prod_{c}^{C}Q(\mu_{pc})\cdot\\&\prod_{p}^{P}\prod_{c}^{C}Q(\lambda_{pc})\prod_{c}^{C}Q(\mu_c)Q(\lambda)\end{aligned} \quad (7)$$

在 E 步中分别对 $Q(x_{rpc})$,$Q(\mu_{pc})$,$Q(\lambda_{pc})$,$Q(\mu_c)$,$Q(\lambda)$ 进行迭代优化. 迭代开始需要初始化超参数,记为 ϕ^0,迭代过程如式(8)所示.

$$\text{E-step: } Q(\theta)^{t+1}=P(\theta|\phi^t,D)$$

$$\text{M-step: }\phi^{t+1}=\arg\max_{\phi}\int \mathrm{d}\theta Q(\theta)^{t+1}\log P(D|\theta,\phi)P(\theta|\phi) \quad (8)$$

其中 E 步中隐含变量的分布形式如式(9)

$$Q(x_{rpc})=N\left(x_{rpc};\frac{\hat{x}_{rpc}+s_{rpc}^2\langle\mu_{pc}\rangle\langle\lambda_{pc}\rangle}{1+s_{rpc}^2\langle\lambda_{pc}\rangle},\frac{s_{rpc}^2}{1+s_{rpc}^2\langle\lambda_{pc}\rangle}\right) \quad (9)$$

$$Q(\mu_{pc})=N\left(\mu_{pc};\frac{\langle\mu_c\rangle\langle\lambda\rangle+\langle\lambda_{pc}\rangle\sum_r x_{rpc}}{\langle\lambda\rangle+\sum_r\langle\lambda_{pc}\rangle},(\langle\lambda\rangle+\sum_r\langle\lambda_{pc}\rangle)^{-1}\right) \quad (10)$$

$$Q(\lambda_{pc})=Ga(\lambda_{pc};\alpha_{pc}+\sum_r\frac{1}{2},\beta_{pc}+\frac{1}{2}\sum_r\langle(x_{rpc}-\mu_{pc})^2\rangle) \quad (11)$$

$$Q(\mu_c)=N\left(\mu_c;\frac{\mu_0\eta_0+\lambda\sum_p\langle\mu_{pc}\rangle}{\eta_0+\sum_p\langle\lambda\rangle},\left(\eta_0+\sum_p\langle\lambda\rangle\right)^{-1}\right) \quad (12)$$

$$Q(\lambda)=Ga\left(\lambda;\alpha_0+\sum_{pc}\frac{1}{2},\ \beta_0+\frac{1}{2}\sum_{pc}\langle(\mu_{pc}-\mu_c)^2\rangle\right) \quad (13)$$

当 EM 算法收敛时,$Q(\mu_c)$ 近似于后验分布 $P(\mu_c|D,\phi)$,由式(12)可知每个条件下的 μ_c 相互独立,可以很容易得到每个条件下与平台无关的基因或异构体表达值的均值与方差.

3.3 差异评估

由于 $Q(\mu_c)$ 近似于后验分布 $P(\mu_c|D,\phi)$,因此得到的 $Q(\mu_c)$ 分布后,我们即可评估数据集中任意两个实验条件下,显著差异表达的基因或异构体. 假如数据集包含两个实验条件,条件 c1 的表达水平高于条件 c2,则差异表达分布 $P(\mu_{c1}>\mu_{c2}|D,\phi)$ 的概率值为

$$P(\mu_{c1}>\mu_{c2}|D,\phi)=\int_0^{+\infty}\mathrm{d}(\mu_{c1}>\mu_{c2})P(\mu_{c1}>\mu_{c2}|D,\phi) \quad (14)$$

通过式(14)计算得到的概率值被称为 PPLR 值[10]. PPLR 值与传统统计测试中显著性概率值 p-value 类似,设置一个阈值来判断上升规则差异的基因或异构体,并通过公式 $\frac{1}{2}-\left|\frac{1}{2}-PPLR\right|$ 可将 PPLR 值转化为 p-value. 而下降规则差异的基因或异构体,同样可以由式(14)识别到,只需要将积分区间改为$(-\infty,0)$即可.

4　实验数据集

4.1　MAQC 数据集

MAQC 来自美国食品药品监督管理局的生物芯片质量控制项目(Microarray Quality Control, MAQC),该项目主要用来评估不同测序平台下高质量样本的基因表达水平,项目包括 MAQC_I[24]、MAQC_II[25]和 MAQC_III(即 Sequencing Quality Control,SEQC)[26]三部分. MAQC 数据集被当作标准数据集被广泛应用于评估不同测量平台的不同方法的性能. MAQC 数据集主要采用两种人类的样本,分别是普遍参考 RNA(Universal Human Reference RNA,UHRR)和大脑参考 RNA(Human Brain Reference RNA,HBRR). 实验包括四个实验条件,实验 A 中只包含 UHRR 样本,实验 B 只包含 HBRR 样本,实验 C 和实验 D 的样本分别由 UHRR 和 HBRR 样本以 3∶1 与 1∶3 的比例混合构成. 本文采用该数据集在实验条件 A 和 B 下的 Human Genome U133 Plus 2.0 Array(HGU133), Human Exon 1.0 ST Array(HuEx1.0ST)和 Human Transcriptome Array 2.0(HTA2.0)三种芯片数据,其中 HGU133 数据和 HuEx1.0ST 数据在每个实验条件下各包含 5 个技术重复芯片,HTA2.0 数据每个实验条件下包含 3 个技术重复芯片.

另外,MAQC 数据集提供了 1000 个经过 qRT-PCR(Quantificational Real-Time PCR)实验验证的基因. 本文对 qRT-PCR 数据进行过滤处理,筛选出高置信度的差异结果,筛选结果可作为基因真实的差异表达情况. 根据文献[27]提出的 LFC(log2 fold-change)方法,计算两个实验条件下基因表达值的 LFC. LFC 的绝对值大于 2.0 的基因被认为是差异表达的基因,记为 DE. LFC 的绝对值小于 0.2 的基因则是未发生差异表达的基因,记为 non-DE. 由于 LFC 绝对值在 0.2 与 2 之间的数据置信度较低,不能确定基因是否发生差异表达,因此这里不作为差异表达分析的衡量标准. 经过筛选得到 305 个基因的高置信度的差异表达结果,其中包括 218 个 DE 基因和 87 个 non-DE 基因.

4.2　SEQC 数据集

SEQC[26]是 MAQC 项目提供的最新的实验数据集,其主要目的是评估 RNA-Seq 技术. SEQC 与 MAQC 数据集均采用 UHRR 和 HBRR 两个样本,并且分别以 3∶1 与 1∶3 的比例混合构成样本 C 和样本 D,但 SEQC 实验中采用的测序平台为 Illumina HiSeq 2000,增加了测序深度和测序长度. 本文主要采用 SEQC 数据集中实验条件 A 和 B 下 RAN-Seq 的双末端测序数据,其中在每个实验条件下包括 8 个测序通道,可当作 8 个技术重复. 另外,SEQC 数据集提供了经过 RT-qPCR(Reverse Transcription qPCR)实验验证的两万多个异构体,同样采用 LFC 方法对 RT-qPCR 数据进行筛选,筛选结果可作为异构体真实的差异情况. 经过筛选得到了 1002 个单一异构体的高置信度的差异表达结果,其中包括 643 个 DE 异构体和 359 个 non-DE 异构体.

4.3　人类骨髓数据集

文献[28]采用人类骨髓数据集(Human Bone Marrow, HBM)对基因芯片和 RNA-Seq 技术进行了对比研究,并且针对特定转录组,引入转录模式来识别跨平台的探针(集). HBM 数据集包含了来自安捷伦人类普遍参考 RNA(Agilent universal human reference RNA)和正常骨髓的 RNA(normal donor bone marrow RNA). 该数据集包含 HTA2.0 芯片和采用 RNA-Seq 技术的 Illumina HiSeq 2500 测序数据,其中 HTA2.0 数据在每个样本下各包含 2 个技术重复芯片,RNA-Seq 数据集在每个样本下只包含一个测序通道,无重复实验.

本文使用 MAQC、SEQC 数据集和 HBM 数据集验证 mpDE 模型的差异基因检测的性能,使用 MAQC 和 SEQC 数据集验证 mpDE 模型的差异异构体检测的性能,并与单平台的差异表达分析结果,MRS 和 limma 组合以及 RSP 模型进行对比. 本文中 HGU133 芯片原始数据分析方法为 mmgMOS[22], HuEx1.0ST 和 HTA2.0 芯片原始数据分析方法为 GME[16],RNA-Seq 原始数据分析方法为 PGSeq[17]. 这些方法在获得每个平台基因和异构体表达水平的同时,还得到该表达值的技术性测量误差,这些结果可作为多平台数据源直接用于 mpDE 模型. 在差异表达检测中,三种基因芯片均采用 PPLR 方法[10]寻找差异基因或异构体,RNA-Seq 采用 PG_bayes[29]方法.

5　结果与讨论

本节首先验证了 mpDE 模型计算合理性,其次验证了 mpDE 模型基因表达水平计算的准确性,并

与单平台的表达水平计算方法进行对比.然后分别从基因和异构体差异表达分析两方面验证 mpDE 模型差异检测的性能,并与单平台差异检测结果,MRS 和 limma 组合以及 RSP 模型进行对比.最后评估 mpDE 模型的时间和空间复杂度,并与 MRS 和 RSP 进行对比.

5.1 模型计算合理性验证

本文随机选取 MAQC 数据集中的一个 qRT-PCR 验证基因 ENSG00000152583,经 qRT-PCR 实验验证该基因显著差异表达,在 UHRR 条件下表达水平较低,在 HBRR 条件下表达水平较高.本文通过对比该基因的不同平台测量到的原始表达水平分布和 mpDE 模型计算获得的平台相关(μ_{pc})和平台无关(μ_c)的表达水平分布来验证模型的合理性,实验结果如图 5 所示.图 5 中的 HGU133、HuEx1.0ST、HTA2.0 和 RNA-Seq 分别表示通过基因表达水平计算获得的四个平台的原始表达数据(用细线表示);与四个平台相同线型,但是线条较粗的是 mpDE计算获得的每个条件下的平台相关基因表达水平,即 μ_{pc} 的后验分布,最粗的线表示 mpDE 计算获得的平台无关的基因表达水平,即 μ_c 的后验分布.由该图可以看出 mpDE 模型排除了技术性测量误差以及重复实验波动性的影响,通过变分 EM 算法计算出了较为合理的平台相关以及平台无关基因表达水平分布,达到了融合多平台实验数据的目的.最终获得的平台无关表达水平分布可以应用于各种基因表达后续分析中,如寻找差异基因、聚类等.本文后续部分将其应用到转录组的差异表达检测.

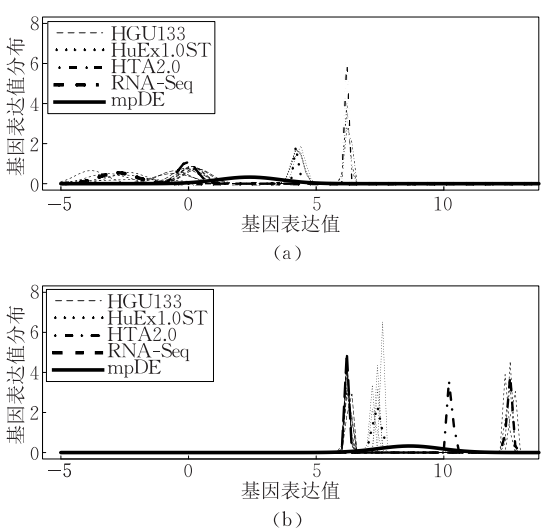


图 5 (a)和(b)分别是基因 ENSG00000152583 在 UHRR 和 HBRR 两个条件下的 mpDE 模型表达水平分布计算结果

5.2 表达水平计算准确性验证

为验证 mpDE 模型获得的平台无关的基因表达值的准确性,本文将 MAQC 数据集中经 qRT-PCR 验证的 1000 个基因与四个实验平台的注释文件进行匹配,筛选出四个实验平台共同检测到的 qRT-PCR 验证的基因有 804 个,并将 qRT-PCR 实验测量结果作为真实的基因表达水平.分别计算 UHRR 和 HBRR 两个条件下的 qRT-PCR 数据与四个平台融合后的 mpDE 计算结果之间的相关系数(squared Pearson correlation coefficient,R^2),并与单平台表达数据进行对比.相关系数越接近 1,则表示表达水平计算的准确度越高,实验结果如表 1 所示,每个条件下相关系数最高的结果用黑体和下划线标注.从表 1 可以看出,不同平台测量到的基因表达水平准确性差异较大,其中 RNA-Seq 在单平台测量结果中最为准确,而 mpDE 模型计算结果的准确性相比于三种基因芯片获得了显著提高,并且高于 RNA-Seq.由此可见融合多平台表达数据能够提高表达水平计算的准确性,有助于提高后续差异表达检测的精度.

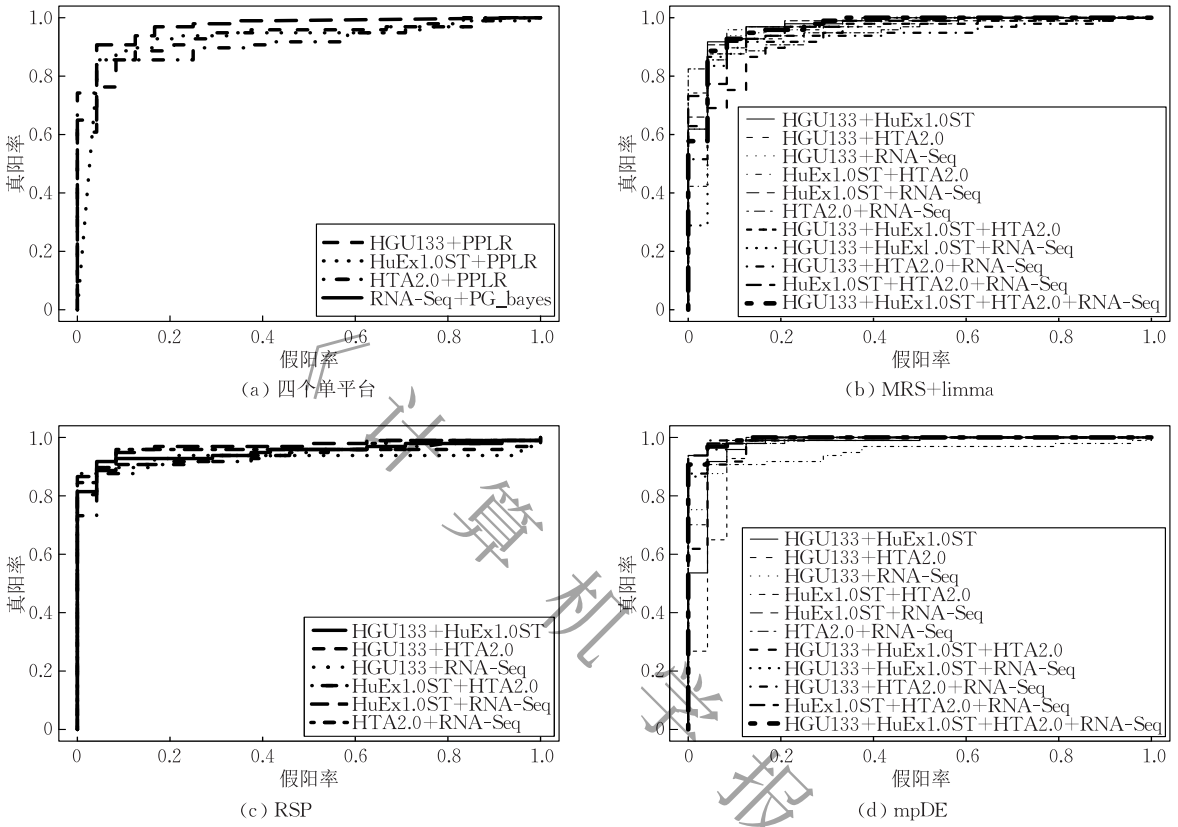
表 1 不同平台的不同原始数据分析方法以及 mpDE 获得的基因表达水平与 qRT-PCR 数据的相关系数

平台	HGU133	HuEx1.0ST	HTA2.0	RNA-Seq	mpDE
MAQC.UHRR	0.8315	0.6935	0.7204	0.8672	**0.8979**
MAQC.HBRR	0.8152	0.6800	0.7046	0.8646	**0.8858**

5.3 基因的差异表达分析

本文采用 MAQC 和 SEQC 数据集验证 mpDE 模型在差异基因检测方面的性能,并分别与单平台和融合多平台表达数据的差异检测方法进行对比.将 305 个高置信度的 qRT-PCR 验证基因与四个实验平台的注释文件进行匹配,HGU133、HuEx1.0ST、HTA2.0 芯片和 RNA-Seq 筛选出的基因数目分别是 152、305、280、270,四个平台共同检测到的 qRT-PCR 验证基因有 121 个,包括了 97 个 DE 基因和 24 个 non-DE 基因,并作为真实的基因差异情况.我们利用这 121 个基因的四个平台的表达数据验证了 mpDE 算法,对实验结果绘制接受者特征曲线(Receiver Operating Characteristic,ROC),并根据 ROC 曲线下的面积(Area Under Roc Curve,AUC)来对比评估 mpDE 模型与单平台以及融合多平台表达数据差异表达分析方法的性能.AUC 值越接近于 1,差异检测结果的准确度越高.其中,三种基因芯片均采用 PPLR 方法寻找差异基因,RNA-Seq 采

用 PG_byes 方法寻找差异基因. 多平台差异检测分别采用经典算法 MRS 和 limma 方法的组合以及最新的 RSP，实验结果如图 6 和表 2 及表 3 所示. 另外，由于 HGU133 芯片与其它三个平台测量到的基因重叠度较低，因此我们排除了 HGU133 芯片，其它三个平台共同检测到的 qRT-PCR 验证基因有 248 个，相比四个平台的 121 个共同基因具有更好的统计特性. 我们采用这 248 个基因数据进一步验证 mpDE 模型的性能，并与单平台和融合多平台表达数据的差异检测方法进行对比，实验结果分别如图 7 和表 4 及表 5 所示.



(a) 四个单平台　(b) MRS+limma

(c) RSP　(d) mpDE

图 6　MAQC 和 SEQC 数据集下 mpDE 与四个单平台及 MRS 和 limma 组合、RSP 的 121 个共同基因差异检测的 ROC 曲线

表 2　MAQC 和 SEQC 数据集下四个单平台的 121 个共同基因差异检测的 *AUC* 值

平台	HGU133	HuEx1.0ST	HTA2.0	RNA-Seq
AUC	0.9283	0.9249	0.9188	**0.9669**

表 3　MAQC和SEQC数据集下mpDE与MRS和limma组合、RSP 的 121 个共同基因差异检测的 *AUC* 值

平台+方法	MRS+limma	RSP	mpDE
HGU133+HuEx1.0ST	0.9566	0.9502	0.9712
HGU133+HTA2.0	0.9334	0.9369	0.9472
HGU133+RNA-Seq	0.9467	0.9312	0.9802
HuEx1.0ST+HTA2.0	0.9329	0.9349	0.9442
HuEx1.0ST+RNA-Seq	0.9639	**0.9691**	0.9948
HTA2.0+RNA-Seq	0.9678	0.9626	**0.9957**
HGU133+HuEx1.0ST+HTA2.0	0.9592	—	0.9764
HGU133+HuEx1.0ST+RNA-Seq	0.9665	—	0.9909
HGU133+HTA2.0+RNA-Seq	0.9622	—	0.9926
HuEx1.0ST+HTA2.0+RNA-Seq	**0.9725**	—	0.9953
All Platforms	0.9721	—	0.9944

实验结果显示，与单平台差异检测结果、MRS 和 limma 组合以及 RSP 相比，mpDE 获得了更为准确的差异检测结果. 其中，融合三种基因芯片的 mpDE 获得的差异检测准确度显著高于三种基因芯片. 由于 RNA-Seq 数据在单平台中获得了最为准确的检测结果，三种基因芯片分别与 RNA-Seq 数据融合的 mpDE 差异检测的 *AUC* 均达到了 99% 以上，表明与 qRT-PCR 数据吻合程度较高，具有较高的准确度. 另外，我们发现虽然 HuEx1.0ST 和 HTA2.0 芯片的基因表达水平与 qRT-PCR 的相关系数较低(如表 1 所示)，但是 mpDE 模型中融入了表达水平的技术性测量误差，极大地提高了差异基因检测的准确性.

对比融合多平台表达数据的差异检测结果，可以发现 mpDE 模型的准确率均高于 MRS 和 limma

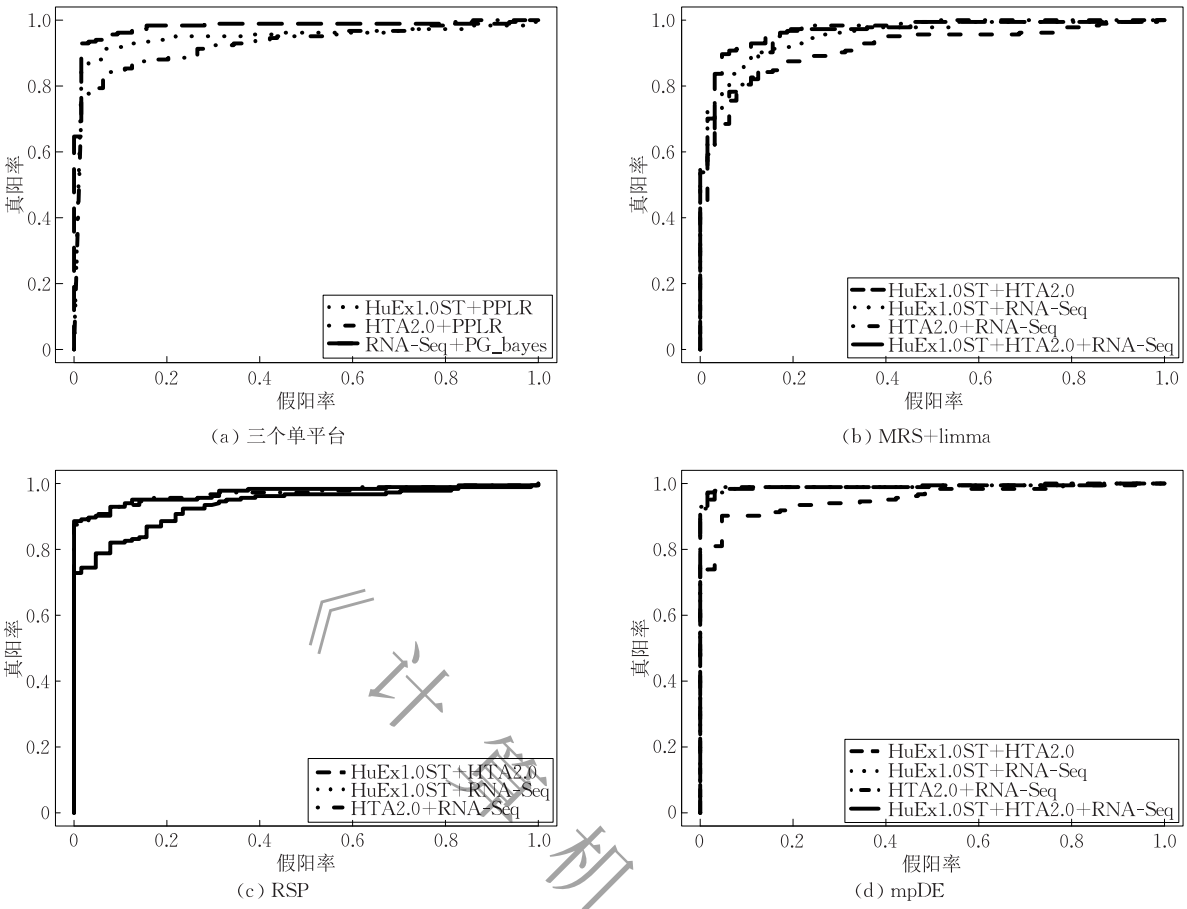


图 7 MAQC 和 SEQC 数据集下 mpDE 与三个单平台及 MRS 和 limma 组合、RSP 的 248 个共同基因差异检测的 ROC 曲线

表 4 MAQC 和 SEQC 数据集下三个单平台的 248 个共同基因差异检测的 *AUC* 值

平台	HuEx1.0ST	HTA2.0	RNA-Seq
AUC	0.9453	0.9249	**0.9784**

表 5 MAQC 和 SEQC 数据集下 mpDE 与 MRS 和 limma 组合、RSP 的 248 个共同基因差异检测的 *AUC* 值

平台	MRS+limma	RSP	mpDE
HuEx1.0ST+HTA2.0	0.9144	0.9353	0.9530
HuEx1.0ST+RNA-Seq	0.9525	**0.9702**	0.9911
HTA2.0+RNA-Seq	0.9565	0.9691	**0.9915**
HuEx1.0ST+HTA2.0+RNA-Seq	**0.9658**	—	0.9907

组合及 RSP. MRS 方法虽然能够融合多平台表达数据，但其采用排序的方法，改变了基因之间的相关性，而且对照平台的选择也会影响后续差异表达分析的准确性. RSP 模型仅能够融合两个平台的表达数据进行差异基因检测，而 mpDE 模型可以融合任意多个平台的表达数据，应用范围更广，准确率更高. 由于在真实的实验中，研究人员往往更关注假阳率较小的差异基因结果，从图 6 和图 7 中可以看出，在假阳率小于 0.1 的范围中，mpDE 模型获得的 ROC 高于单平台、MRS 和 limma 组合及 RSP 模型.

本文随机选择 MAQC 数据集中的一个经 qRT-PCR 实验验证发生差异表达的基因 ENSG00000-141570. 对于该基因，limma 方法检测到的 p-value 为 0.66，RSP 检测到的 IDR(Irreproducible Discovery Rate)[30]值为 0.23，所以 limma 和 RSP 判断该基因未发生显著差异表达. mpDE 方法获得的 PPLR 值为 0.03，判断该基因发生显著差异表达，mpDE 计算结果如图 8 所示. qRT-PCR 实验得到该基因的 LFC 值为 2.3，根据文献[27]的 LFC 筛选方法，说明该基因的差异变化相对较低，因此识别出该基因发生差异表达具有一定的难度. 而 mpDE 考虑了不同平台的技术性测量误差，以及同一平台在不同实验条件下的生物重复和技术重复的波动性，最终获得了较为准确的平台无关的表达水平和较为合理的测量误差，提高了差异表达分析的准确度和灵敏度，因此获得了较高的 *AUC* 值.

另外，本文采用 HBM 数据集进一步对比 mpDE 与单平台差异基因检测方法、MRS 和 limma 组合及 RSP 模型的性能. 由于 HBM 数据集未提供 qRT-

PCR 实验验证数据，无法提前知道哪些基因是真实发生差异表达的. 而且不同差异检测方法采用不同的统计测试来判别差异基因，它们之间无法直接比较，因此本文根据文献[17]中的验证方法，利用多个差异检测方法获得的高置信度的差异基因作为衡量标准，并将分析结果作为 MAQC 数据验证结果的一个补充. 本文首先选择每种差异检测方法的前 1000 个置信度最高的差异基因，对利用不同方法寻找到的共同差异基因用维恩图(Venn diagram)表示，如图 9 所示. 图 9 中考虑了 HTA2.0 和 RNA-Seq 两种单平台差异检测方法，以及 MRS 和 limma 组合、RSP 和 mpDE 三种融合多平台表达数据的差异检测方法. 从图中可以看出，有 121 个基因被五个方法共同判定为差异基因，具有较高的置信度，因此本文将这 121 个基因当作真正的差异基因，其他基因当作非差异基因. 通过 ROC 曲线和 *AUC* 评估不同方法差异基因检测的性能，实验获得了与 MAQC 和 SEQC 数据集一致的对比结果，实验结果如图 10 和表 6 所示. 由图 10 可以看出，mpDE 模型的 ROC 曲线最靠近纵坐标，表明 mpDE 模型具有较高的灵敏度. 由表 6 可知，mpDE 模型获得了最高的 *AUC* 值，表明 mpDE 的准确率显著高于单平台差异检测方法以及 MRS 和 limma 组合、RSP 模型.

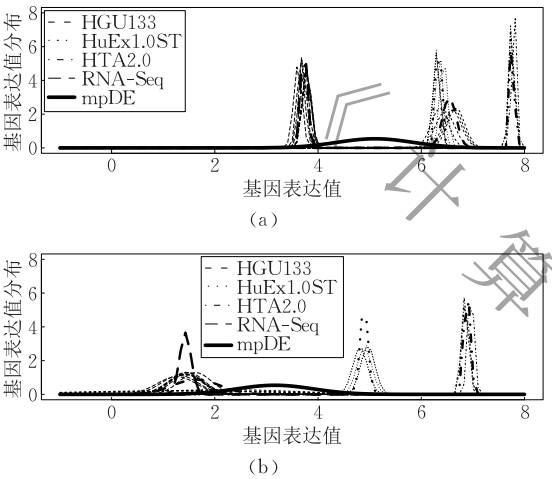


图 8　(a)和(b)分别是基因 ENSG00000141570 在 UHRR 和 HBRR 两个条件下的 mpDE 模型表达水平分布计算结果

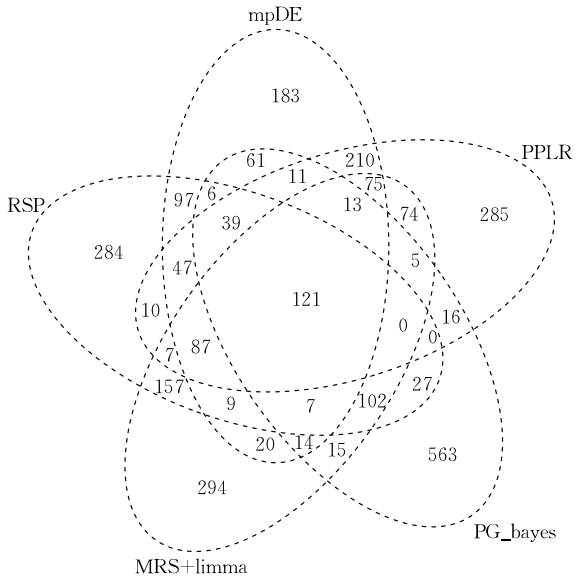


图 9　HBM 数据集下不同方法寻找到的差异基因的维恩图. 每个椭圆表示每个方法检测到的前 1000 个高置信度的差异基因. 相邻椭圆重叠区域中的数字表示椭圆对应的方法共同找到的差异基因

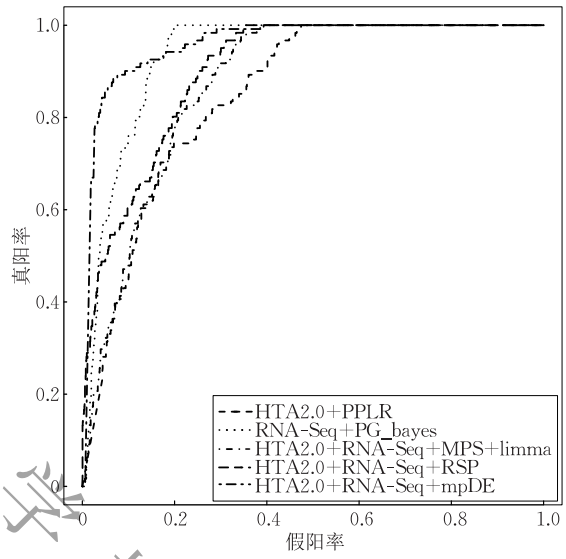


图 10　HBM 数据集下 mpDE 与两个单平台及 MRS 和 limma 组合、RSP 的差异基因检测的 ROC 曲线

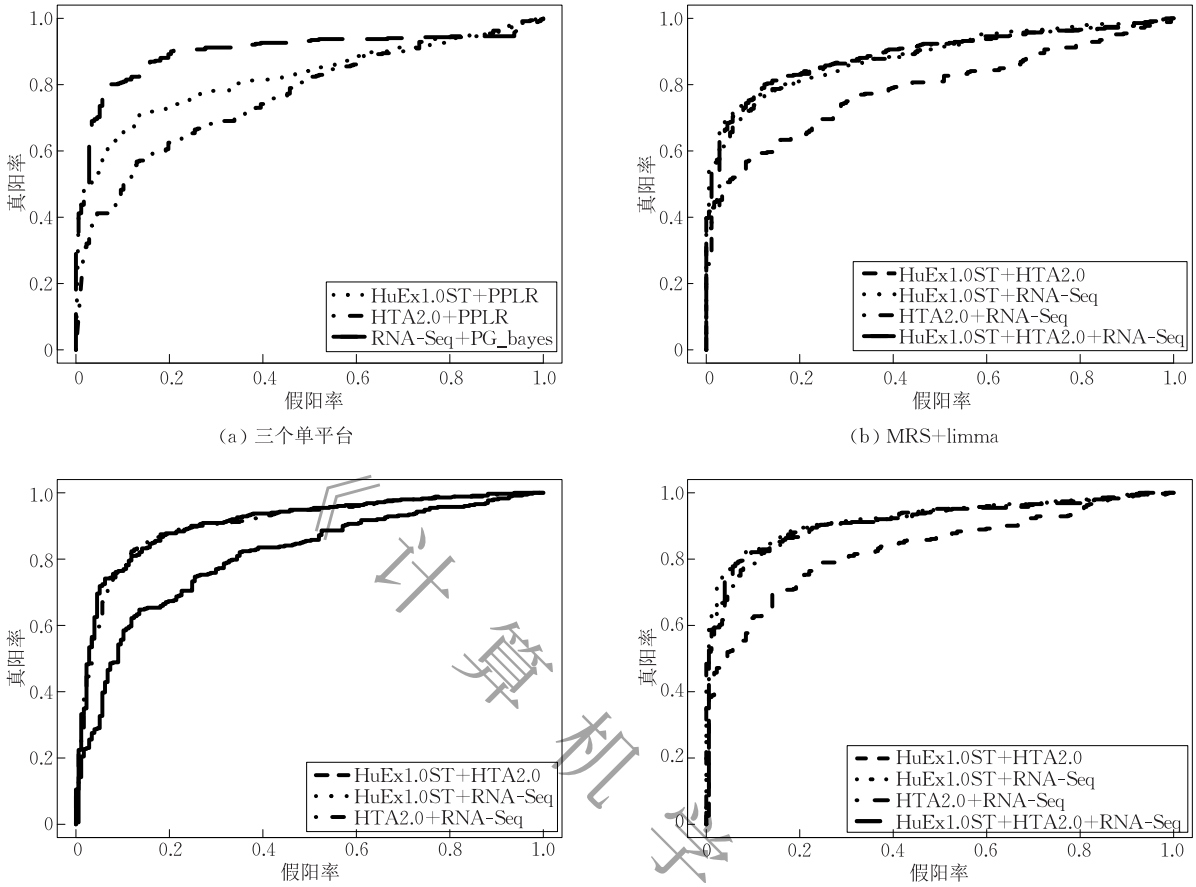
表 6　HBM 数据集下 mpDE 与两个单平台 MRS 和 limma 组合、RSP 的差异基因检测的 *AUC* 值

平台＋方法	*AUC*
HTA2.0＋PPLR	0.8499
RNA-Seq＋PG_bayes	0.9362
HTA2.0＋RNA-Seq＋MRS＋limma	0.8728
HTA2.0＋RNA-Seq＋RSP	0.9023
HTA2.0＋RNA-Seq＋mpDE	**0.9611**

5.4　异构体的差异表达分析

由于 HGU133 芯片无法测量异构体的表达水平，本文将采用 MAQC 数据集中的 HuEx1.0ST、HTA2.0 芯片以及 SEQC 数据集中的 RNA-Seq 的表达数据验证 mpDE 模型在差异异构体检测方面的性能. 本文将筛选出的 1002 个高置信度的 RT-qPCR 验证的单一异构体，与 HuEx1.0ST、HTA2.0 芯片以及 RNA-Seq 的注释文件进行匹配，三个平台共同检测到的 RT-qPCR 验证的异构体有 529 个. HuEx1.0ST 和 HTA2.0 均采用 PPLR 方法寻找

差异异构体，RNA-Seq 采用 PG_bayes 方法寻找差异异构体. 最后通过 ROC 曲线与 *AUC* 对比评估 mpDE 模型与单平台差异检测方法以及 MRS 和 limma 组合、RSP 模型的差异异构体检测的性能，实验结果如图 11 和表 7 及表 8 所示.



(a) 三个单平台　(b) MRS+limma

(c) RSP　(d) mpDE

图 11　MAQC 和 SEQC 数据集下 mpDE 与三个单平台及 MRS 和 limma 组合、RSP 的 529 异构体差异检测的 ROC 曲线

表 7　MAQC 和 SEQC 数据集下三个单平台的 529 个共同异构体差异检测的 *AUC* 值

平台	HuEx1.0ST	HTA2.0	RNA-Seq
AUC	0.8228	07659	**0.8987**

表 8　MAQC 和 SEQC 数据集的下 mpDE 与 MRS 和 limma 组合、RSP 的 529 个共同异构体差异检测的 *AUC* 值

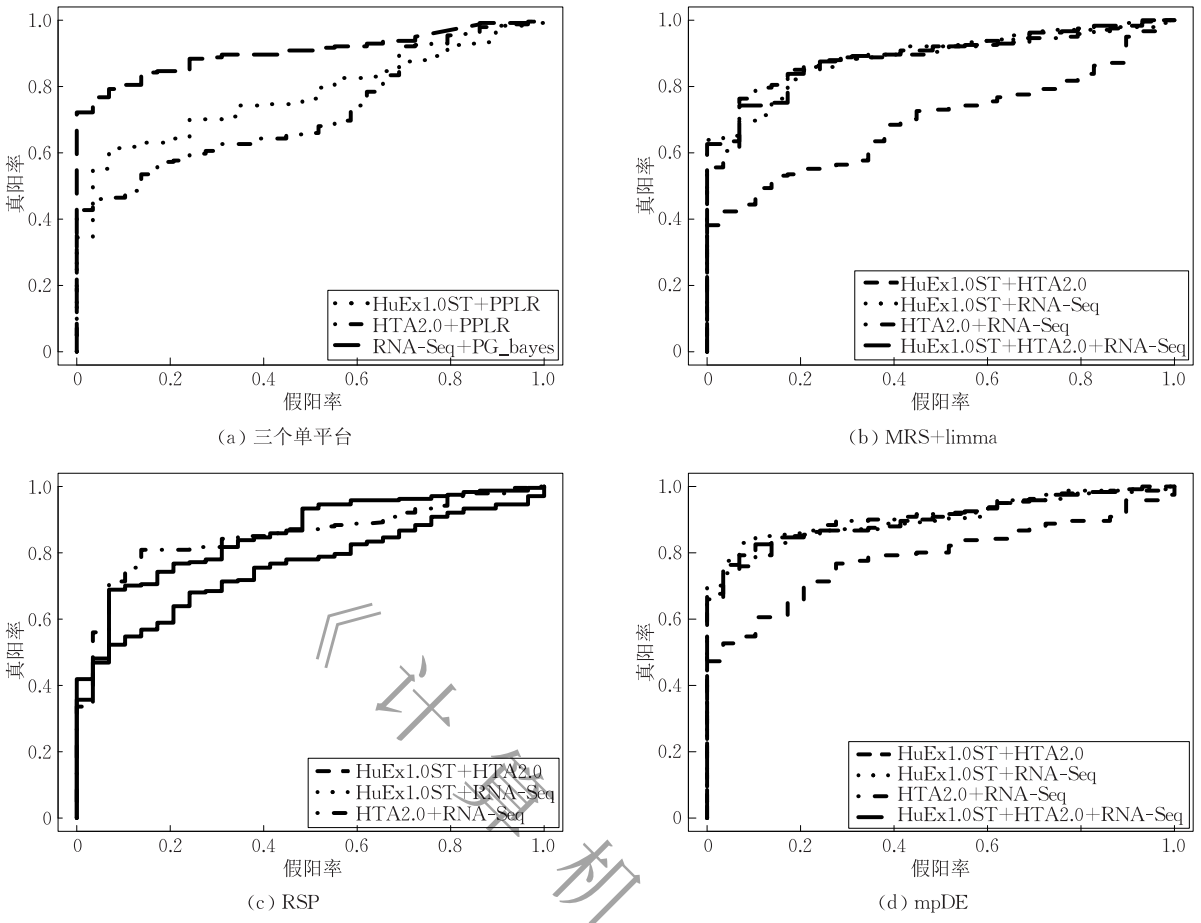
平台	MRS+limma	RSP	mpDE
HuEx1.0ST+HTA2.0	0.7881	0.8056	0.8491
HuEx1.0ST+RNA-Seq	0.8879	**0.9083**	**0.9266**
HTA2.0+RNA-Seq	0.8799	0.9049	0.9106
HuEx1.0ST+HTA2.0+RNA-Seq	**0.8913**	—	0.9021

实验结果显示，与单平台差异检测方法相比，mpDE 模型差异异构体检测的灵敏度和准确度均有所提升，其中融合 HuEx1.0ST 和 RNA-Seq 的表达数据的 mpDE 模型的差异检测准确度最高，且显著高于三个单平台差异检测结果的准确度. 对比两种融合多平台表达数据差异检测方法，mpDE 的准确率高于 MRS 和 limma 组合及 RSP 模型. 由于异构体表达水平计算难度较大，表达水平计算结果的准确性会影响异构体差异表达分析结果，所以总体上差异异构体检测的准确度低于差异基因检测，但是多平台数据融合能够一定程度地提高差异异构体检测的精度.

5.5　低表达异构体的差异表达分析

为了进一步研究 mpDE 模型在最难计算的低表达区间上差异异构体检测的性能，本文根据 RT-qPCR 测量值，将 529 个 RT-qPCR 验证的异构体分为高中低三个区间. 将 RT-qPCR 测量值小于 0.02 的异构体划分为低表达区间，共筛选出 270 个低表达异构体. 利用低表达区间的 270 个异构体数据集对比评估 mpDE 模型与单平台差异检测结果以及 MRS 和 limma 组合、RSP 模型，实验结果如图 12 和表 9 及表 10 所示.

实验结果显示在低表达区间，HuEx1.0ST 和 HTA2.0 芯片以及 RSP 模型的差异检测准确率显著下降，MRS 和 limma 组合在融合 HuEx1.0ST 和



(a) 三个单平台　(b) MRS+limma　(c) RSP　(d) mpDE

图 12　MAQC和SEQC数据集下mpDE与三个单平台及MRS和limma组合、RSP的270个低表达异构体差异检测的ROC曲线

表 9　MAQC 和 SEQC 数据集下三个单平台的 270 个低表达异构体差异检测的 *AUC* 值

平台	HuEx1.0ST	HTA2.0	RNA-Seq
AUC	0.7728	0.7221	**0.9037**

表 10　MAQC和SEQC数据集的下 mpDE 与 MRS 和 limma 组合、RSP 的 270 个低表达异构体差异检测的 *AUC* 值

平台	MRS+limma	RSP	mpDE
HuEx1.0ST+HTA2.0	0.6970	0.7578	0.8054
HuEx1.0ST+RNA-Seq	0.8894	**0.8552**	**0.9138**
HTA2.0+RNA-Seq	0.8884	0.8526	0.9057
HuEx1.0ST+HTA2.0+RNA-Seq	**0.8904**	--	0.9081

HTA2.0 芯片数据的差异表达分析中准确率有所下降. 虽然与全表达区间的差异检测相比，mpDE 的准确率有所下降，但在低表达区间仍高于单平台差异检测的准确率，并显著高于 RSP 模型. 该实验结果表明 mpDE 在低表达差异异构体检测方面显著优于 RSP 模型，并优于 MRS 和 limma 方法的组合.

5.6　评估 mpDE 的时空复杂度

为评估 mpDE 的计算时间与内存使用率，本文采用 MAQC 数据集中四个平台共同检测到的 9624 个基因进行验证，并与 MRS 以及 RSP 进行对比. 测试工作站的配置为 4 核 3.2 GHz 的 CPU 和 16 GB 内存，由于 RSP 模型仅用于融合两个平台的表达数据，因此只对比两个算法在融合两个平台表达数据方面的计算时间和内存使用率，实验结果如表 11 所示. 实验结果显示，由于 mpDE 采用了复杂概率方法，在获得计算准确性的同时牺牲了计算时间和空间的效率. 虽然 MRS 和 RSP 在计算时间和内存使用率方面比 mpDE 有一定的优势，但是 RSP 仅能用于两个平台数据的融合，应用范围有一定局限性；MRS 虽然能够应用于多平台表达数据融合，但其采用的排序策略改变了基因之间的相关性，并且对照平台的表达数据的准确性会影响后续差异检测的准确度，同时也忽略了不同平台的技术性测量误差. mpDE 不仅能够应用到多个实验平台数据的融合，并且考虑了每个平台的技术性测量误差，能够获得比较准确的计算结果. 随着高性能计算环境的日益普及，对于大规模数据集 mpDE 可以通过并行计算来提高计算速度，以进一步提高其实用性.

表 11 mpDE 与 MRS 以及 RSP 的计算时间与内存使用率的对比

平台	MRS		RSP		mpDE	
	Time/s	MEM/MB	Time/s	MEM/MB	Time/s	MEM/MB
HGU133+HuEx1.0ST	4.3	95	5.3	48	150.0	232
HGU133+HTA2.0	4.3	95	5.0	48	216.0	231
HGU133+RNA-Seq	4.3	95	50.1	48	108.0	231
HuEx1.0ST+HTA2.0	4.2	95	90.0	46	348.0	233
HuEx1.0ST+RNA-Seq	4.3	95	168.0	48	108.0	233
HTA2.0+RNA-Seq	4.0	95	13.0	48	138.0	235
HGU133+HuEx1.0ST+HTA2.0	5.4	96	—	—	150.0	234
HGU133+HuEx1.0ST+RNA-Seq	5.6	95	—	—	112.0	240
HGU133+HTA2.0+RNA-Seq	5.5	96	—	—	102.0	247
HuEx1.0ST+HTA2.0+RNA-Seq	5.5	96	—	—	113.0	243
All Platforms	6.8	101	—	—	90.0	252

6 结束语

本文针对目前融合多平台表达数据的转录组差异表达分析方法存在的问题，提出了多平台转录组差异检测模型 mpDE. 该模型利用多平台表达数据，同时考虑了技术性测量误差和同一平台在不同实验条件下产生的生物重复实验或技术重复实验导致的波动，能够更好地模拟基因或异构体真实表达值的概率分布，从而提高差异检测的准确度. 该模型获得的平台无关的表达值不仅可应用于差异表达检测，还能应用于聚类、调控网络分析等其他多平台数据的分析，因此 mpDE 模型的应用前景较为广阔.

本文采用 MAQC、SEQC 数据集和 HBM 数据集来验证 mpDE 模型的差异表达分析性能，并与单平台的差异表达分析结果以及多平台差异检测方法 MRS 和 limma 组合、RSP 进行对比. 实验结果表明，在差异基因检测中，mpDE 模型灵敏度和准确度均高于单平台差异检测结果以及 MRS 和 limma 组合、RSP 模型，而且在 HBM 数据集上其准确度显著高于 HTA2.0 芯片的单平台差异检测结果以及 MRS 和 limma 组合、RSP 的多平台检测结果. 在差异异构体检测中，mpDE 与单平台差异检测方法以及 MRS 和 limma 组合、RSP 相比，获得了较高的准确度. 在异构体低表达区间，mpDE 模型的准确率显著高于 RSP 模型. 另外，在时间和空间复杂度方面，由于 mpDE 模型采用了复杂概率方法，其计算时间和空间的效率低于 MRS 和 RSP. 在未来工作中，对于大规模数据集我们可以通过并行计算提高 mpDE 的计算速度，以进一步提高其实用性.

从整体上看，多平台数据融合模型能够显著提高单平台在差异基因检测方面的准确性，但是在差异异构体检测上，由于异构体的表达水平计算有一定难度，尤其是基因芯片采用有限数量的探针检测样本中相应转录本的丰度，而生物选择性剪切的多样性对该技术的性能构成极大挑战，导致目前差异异构体检测精度较低. 另外原始数据分析方法依赖于基因、异构体和探针的注释信息，而注释信息的完善程度对表达水平的计算精度也有一定影响. 相比之下 RNA-Seq 技术没有固定探针的限制，转录本检测的广度和深度均远超基因芯片，故在选择性剪切研究中具有更高的准确度. 由于基因芯片技术在异构体表达水平测量中的局限性，所以目前在异构体差异表达分析方面，多平台数据融合模型相比 RNA-Seq 单平台的优势不够显著. 随着芯片实验技术的提高以及注释信息的不断完善，异构体表达水平计算的准确度也将随之提高，多平台数据融合模型在异构体差异表达分析方面的准确率将会有所提升. 最后，寻找差异基因是基因表达测量实验的一个最基本的目的，我们目前的工作主要针对这个方面展开研究. 在此基础上，我们将在未来的工作中，将这些研究思路扩展到差异基因集的检测中，从系统生物学的角度提高数据分析方法的有效性.

参 考 文 献

[1] Barrett T, Troup D B, Wilhite S E, et al. NCBI GEO: Mining tens of millions of expression profiles—database and tools update. Nucleic Acids Research, 2007, 35(Suppl. 1): D760-D765

[2] Bisognin A, Coppe A, Ferrari F, et al. A-MADMAN: Annotation-based microarray data meta-analysis tool. BMC Bioinformatics, 2009, 10(1): 1-11

[3] Kim J, Patel K, Jung H, et al. AnyExpress: Integrated toolkit for analysis of cross-platform gene expression data using a fast interval matching algorithm. BMC Bioinformatics, 6 2011, 12(1): 75

[4] Chavan S S, Bauer M A, Peterson E A, et al. Towards the integration, annotation and association of historical microarray experiments with RNA-seq. BMC Bioinformatics, 2013, 14(14): 1-11

[5] Shi F, Abraham G, Leckie C, et al. Meta-analysis of gene expression microarrays with missing replicates. BMC Bioinformatics, 2011, 12(1): 1-16

[6] Yang Z, Chen Y, Fu Y, et al. Meta-analysis of differentially expressed genes in osteosarcoma based on gene expression data. BMC Medical Genetics, 2014, 15(1): 80-80

[7] Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC Bioinformatics, 2005, 6(3): 265-280

[8] Yoon Y, Lee J, Park S, et al. Direct integration of microarrays for selecting informative genes and phenotype classification. Information Sciences, 2008, 178(1): 88-105

[9] Heider A, Alt R. VirtualArray: A R/bioconductor package to merge raw data from different microarray platforms. BMC Bioinformatics, 2013, 14(1): 1-10

[10] Liu X, Milo M, Lawrence N D, et al. Probe-level measurement error improves accuracy in detecting differential gene expression. Bioinformatics, 2006, 22(17): 2107-2113

[11] Ritchie M E, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 2015, 43(7): e47

[12] Lyu Y, Li Q. A semi-parametric statistical model for integrating gene expression profiles across different platforms. BMC Bioinformatics, 2016, 17(S1): 51-60

[13] Hong F, Breitling R, Mcentee C W, et al. RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics, 2006, 22(22): 2825-2827

[14] Johnson W E, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 2007, 8(1): 118-127

[15] Liu X, Milo M, Lawrence N D, et al. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. Bioinformatics, 2005, 21(18): 3637-3644

[16] Liu X, Gao Z, Zhang L, et al. Puma 3.0: Improved uncertainty propagation methods for gene and transcript expression analysis. BMC Bioinformatics, 2013, 14(3): 1-15

[17] Liu X, Zhang L, Chen S. Modeling exon-specific bias distribution improves the analysis of RNA-Seq data. PLoS One, 2015, 10(10): e0140032

[18] Baldi P, Long A D. A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes. Bioinformatics, 2001, 17(6): 509-519

[19] Delmar P, Robin S, Daudin J J. VarMixt: Efficient variance modelling for the differential analysis of replicated gene expression data. Bioinformatics, 2005, 21(4): 502-508

[20] Krohn K, Eszlinger M, Paschke R, et al. Increased power of microarray analysis by use of an algorithm based on a multivariate procedure. Bioinformatics, 2005, 21(17): 3530-3534

[21] Sun J, Kabán A, Raychaudhury S. Robust mixtures in the presence of measurement errors//Proceedings of the 24th International Conference on Machine Learning. New York, USA, 2007: 847-854

[22] Irizarry R A, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003, 4(2): 249-264

[23] Beal M. Variational Algorithms for Approximate Bayesian Inference [Ph. D. dissertation]. University College, London 2003

[24] Shi Le-Ming, Reid L H, Jones W D, et al. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. Nature Biotechnology, 2006, 24(9): 1151-1161

[25] MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, 2010, 28(8): 827-838

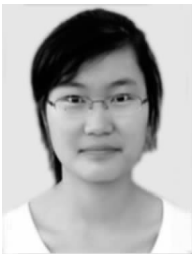
[26] Seqc/Maqc-Iii Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nature Biotechnology, 2014, 32(9): 903-914

[27] Bullard J H, Purdom E, Hansen K D, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 2010, 11(1): 1-13

[28] Yu J, Cliften P F, Juehne T I, et al. Multi-platform assessment of transcriptional profiling technologies utilizing a precise probe mapping methodology. BMC Genomics, 2015, 16(1): 1-15

[29] Wang Li-Li, Liu Xue-Jun, Zhang Li. Differential expression analysis of genes and isoforms based on model selection. Journal of Data Acquisition and Processing, 2016, 31(5): 965-973(in Chinese)
(王黎黎，刘学军，张礼. 基于模型选择的差异基因和异构体检测. 数据采集与处理，2016，31(5)：965-973)

[30] Li Q, Brown J B, Huang H, et al. Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics, 2011, 5(3): 1752-1779



WANG Kai-Li, born in 1990, M. S. candidate. Her main research interest is bioinformatics.

ZHANG Li, born in 1985, Ph. D. His main research interest is bioinformatics.

LIU Xue-Jun, born in 1976, Ph. D., professor. Her main research interests include bioinformatics and machine learning.

Background

Expression analysis is an important way for transcriptome study. Nowadays, public repositories have accumulated a large amount of microarray and RNA-Seq expression data, which provide the possibility of multi-platform expression data analysis. Some researchers have shown that integrating data from multi-platform can increase the statistical power and reliability in expression analysis of transcriptome. Microarray and RNA-Seq are the two main technologies for measuring transcriptome expression and have been widely used in expression analysis of transcriptome. Due to the inherent limitation of microarray, the background noise is large and difficult to be eliminated, while the repeated experiments for RNA-Seq data are few. In addition, the microarray original data is continuous real probe intensities, and the primary RNA-Seq data is discrete read counts. Therefore, the scales of measurements on the two platforms make them incomparable directly. Combining the characteristics of these platforms and integrating multi-platform expression data involve many intricate issues.

Currently, there are two categories of methods for integrating transcriptome expression data: one is meta-analysis, the other is integrating methods combined with some high-level analysis method, such as differential expression (DE) detection methods. However, both categories ignore useful information, such as the technical measurement error of expression measurement under different platforms. In addition, many experiments involve biological or technical replicates for obtaining the certainty of the measured expression. We aim at developing a new method to integrate expression data from multiple platforms to improve expression calculation accuracy from a single platform. We apply this method to the detection of differential expression to verify its performance. This paper proposes a new model, mpDE (multi-platform Differential Expression model), for DE detection by integrating expression data from multiple platforms. This method integrates the expression data and the associated measurement error from different platforms and considers the variability of biological or technical replicates under different conditions for the same platform. Results show that mpDE outperforms the other alternatives in terms of the accuracy of DE detection, manifesting the usefulness of our strategy for integrating expression data from multiple platforms.

This work is partly supported by the Natural Science Foundation of China (No. 61170152) and the Collaborative Innovation Center for Novel Software Technology and Industrialization.