ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 5,Sep. 2016,pp. 965－973
DOI:10. 16337/j. 1004-9037. 2016. 05. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于模型选择的差异基因和异构体检测

王黎黎　刘学军　张　礼

(南京航空航天大学计算机科学与技术学院,南京,210016)

摘　要:基因和异构体差异表达分析是获取基因和异构体功能的重要途径,现已成为生物信息学的一个重要领域。RNA-seq是一种高通量测序技术,近年来广泛用于转录组研究。RNA-seq数据的读段多源映射现象给差异异构体检测带来挑战。针对该问题,本文采用先计算基因和异构体的表达水平,再进行差异分析的方法,以计算表达水平的PGseq模型为基础,采用贝叶斯因子方法进行模型选择,提出一个新的差异检测方法PG_bayes,解决了基因和异构体两方面的差异检测问题。将PG_bayes应用于人类和小鼠共4个真实数据集中,并与目前流行的差异检测方法进行对比。实验结果表明,PG_bayes方法在差异基因和差异异构体检测中具有较高的准确度和灵敏度,并且在差异异构体检测方面表现出优势。

关键词:RNA-seq;差异检测;多源映射;模型选择;贝叶斯因子

中图分类号:TP391. 9　　**文献标志码**:A

Differential Expression Analysis of Genes and Isoforms Based on Model Selection

Wang Lili, Liu Xuejun, Zhang Li

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

Abstract: Differential expression analysis of genes and isoforms is important in obtaining the function of genes and isoforms, thus becoming an essential research focus of bioinformatics. RNA-seq is a new experimental technique based on high-throughput sequencing and is increasingly used in transcriptome research. Read-isoform multi-mappings make it difficult to detect differential expression of isoforms. Here, we proposed a new method, called PG_bayes, to detect differential expression for both genes and isoforms. PG_bayes, based on expressions estimation method PGseq, uses a Bayes factor model selection method to detect differential expression. We applied PG_bayes to three human datasets and one mouse dataset, and compared its performance with popular alternatives. Results show that PG_bayes performs favorably in sensitivity and specificity at both gene and isoform levels.

Key words: RNA-seq; differential expression analysis; multi-mapping; model selection; Bayes factor

引　　言

RNA-seq是现代生物学中对转录组进行研究较为常规的实验方法[1],基于高通量测序技术,一次

基金项目:国家自然科学基金(61170152)资助项目;中央高校基本科研业务费专项(CXZZ11_0217)资助项目。
收稿日期:2015-07-09;**修订日期**:2016-02-20

RNA-seq 实验可以完成数以万计的读段数据的测序工作，将这些读段数据映射定位到参考基因组或转录组上可以得到其量化的表达值。在转录组学研究中，基因和异构体差异表达分析是最基本的研究目标之一。研究发现，人类一些疾病的产生往往和某些基因的表达变化有关，差异基因检测是获取基因功能的重要途径，对于不同条件下的表达数据分析，目的是识别在不同条件下表达发生变化的基因。许多基因要编码不止一种蛋白质结构，这主要通过选择性剪切[2]机制实现，选择性剪切是指一个基因的多个外显子以不同方式重新组合，产生多种异构体，进而指导合成蛋白质。据统计，人类基因组中大约有75%～95%的基因会发生选择性剪切，选择性剪切与生物组织的生长发育、生理学及疾病有关，研究发现一些非正常的波动变化与癌症有关[3]。近年来越来越多的学者关注对选择性剪切波动变化的研究，而差异异构体检测为揭示选择性剪切的变化情况提供一种可行的研究方式[4]。

目前针对 RNA-Seq 数据进行差异基因检测的方法有很多，例如基于读段数据的 DESeq[5]，baySeq[6]以及基于基因表达水平的 MMDiff[7]等，而差异异构体检测的方法相对较少，基于读段数据的方法目前都不能用于差异异构体检测。由于读段容易映射到同一个基因的多个序列相似的选择性剪切异构体上，造成多源映射，给异构体差异表达分析带来很大挑战[8]。基于读段数据的方法可以解决差异基因检测问题，因为对大多数基因，它们的读段计数都可以准确获得，但是由于多源映射，异构体上的读段计数不能确定，导致这类方法不能处理异构体的差异检测问题。为此有学者提出了基于表达水平进行差异检测的方法，这类方法先计算每个异构体的表达水平，再进行异构体差异表达分析，例如 BitSeq[9]，CuffDiff[10]等。其中 CuffDiff 利用 Cufflinks 方法计算出基因和异构体的表达水平，通过实现一个线性统计模型，利用表达水平的波动来判断是否差异表达。表达水平的准确性对后续差异检测的性能有很大影响，为了计算基因和异构体的表达水平，模拟读段分布的概率模型和模拟读段产生过程的产生式模型被提出，例如 Jiang 等[11]提出的泊松模型 rSeq，其用外显子所在异构体表达值的加权和做为泊松分布的参数，以及 Trapnell 等提出的产生式模型 Cufflinks，其使用马科夫模型学习序列偏差，模拟读段的随机采样过程。

为了解决差异异构体检测问题，本文采用了基于表达水平进行差异分析的方法。文献[12]提出了 PGseq 模型，其模拟每个外显子上的读段分布解决了读段的多源映射问题。本文在 PGseq 计算出的基因和异构体的表达水平的基础上，利用表达水平的负二项分布模型，采用基于贝叶斯因子的模型选择方法设计了 PG_bayes 方法，解决了基因和异构体两方面的差异检测问题，并采用 4 个真实的数据集验证该方法的性能。

1 本文方法

1.1 PGseq 模型

PGseq 用泊松分布模拟外显子上的读段数据，由于 RNA-seq 数据表现出比泊松分布期望的更高的变异性，即过离散[13]，PGseq 引入伽玛分布的因子模拟偏差信息，最终推导出关于基因和异构体表达水平的负二项分布模型。经验证，PGseq 模型在表达水平计算方面具有一定优势，为差异检测分析奠定了基础。

PGseq 模拟映射到每个外显子上的读段计数的分布，假设在条件 c 的样本数据中，重复实验 r 上外显子 i 的读段计数为 y_{icr}，则 $y_{icr} = w_{cr} l_i \sum_k M_{ik} t_{icrk}$，其中，$w$ 为每个重复实验的读段总数，l 为每个外显子的长度，t 为异构体上的读段计数，M 描述每个异构体的外显子构成。模型假设第 k 个异构体上第 i 个外显子的读段计数服从泊松分布，即 $t_{ickr} \sim \text{Poisson}(\alpha_{crk}\beta_i)$，其中参数 α_{crk} 是异构体 k 在条件 c 中第 r 个重复样本中的表达水平比率，设置一个隐含的随机变量 β_i 来模拟第 i 个外显子的偏差特性，解决读段分布的过离散问题，假设其服从伽玛分布，即 $\beta_i \sim \text{Gamma}(c, d)$，其中 c 为形状参数，d 为尺度参数。根据泊松分布的叠加性，基因的读段数 y_{icr} 的分布为 $y_{icr} \sim \text{Poisson}(w_{cr} l_i \beta_i \sum_k M_{irk} \alpha_{crk})$，其中 $\beta_i \sim \text{Gamma}(c, d)$。

利用极大似然估计方法分别计算出参数 α_{crk},c,d 的估计值 $\hat{\alpha}_{crk}$,$\hat{c}$,$\hat{d}$,进而推导出异构体的表达水平的负二项分布(Negative binomial distribution, NB)模型,即有

$$t_{crk}=\int P(t_{crk}\mid \hat{\alpha}_{crk}\beta_i)P(\beta_i\mid \hat{c},\hat{d})\,\mathrm{d}\beta_i \sim \mathrm{NB}\left(\hat{c},\frac{\hat{d}}{\hat{d}+\hat{\alpha}_{crk}}\right) \tag{1}$$

基因表达水平是其异构体的表达水平之和,即 $s_{cr}=\sum_k t_{crk}$,由此推导出基因表达水平负二项分布模型,有

$$s_{cr}=\int P\left(s_{cr}\mid \sum\nolimits_k \hat{\alpha}_{crk}\beta_i\right)P(\beta_i\mid \hat{c},\hat{d})\,\mathrm{d}\beta_i \sim \mathrm{NB}\left(\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{crk}}\right) \tag{2}$$

1.2 差异检测方法 PG_bayes

本文采用基于贝叶斯因子的模型选择方法,进行基因和异构体的差异检测,将该方法称为 PG_bayes。贝叶斯因子方法综合了样本信息和先验信息,充分地利用了信息,所以在假设检验问题上具有一定优势。

贝叶斯定理中指出,在给定的数据 D 下,模型 M 的后验概率为

$$P(M\mid D)=\frac{P(D\mid M)\times P(M)}{P(D)} \tag{3}$$

式中:$P(D|M)$为模型 M 的似然。贝叶斯因子方法是在原假设 M_0 和备择假设 M_1 两个模型中选择具有最大似然的模型,贝叶斯因子定义为

$$K=\frac{P(D\mid M_1)}{P(D\mid M_0)}=\frac{P(M_1\mid D)}{P(M_0\mid D)}\Big/\frac{P(M_1)}{P(M_0)} \tag{4}$$

当 $K>1$ 时,选择备择假设 M_1,否则选择原假设 M_0。

在基因的差异表达分析中,基因在不同条件的表达水平为该条件下各个重复实验的表达水平之和,即 $s_c=\sum_r s_{cr}$,由负二项分布的性质和式(2)可以推导出基因在不同条件下的表达水平分布模型,即

$$s_c \sim \mathrm{NB}\left(\sum\nolimits_r \hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{ck}}\right) \tag{5}$$

令 S_{gA} 和 S_{gB} 分别表示条件 A 和条件 B 的基因表达水平,则两个条件的表达水平的负二项分布模型分别为

$$\begin{aligned} S_{gA} &\sim \mathrm{NB}\left(\sum\nolimits_{rA}\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{cAk}}\right)\\ S_{gB} &\sim \mathrm{NB}\left(\sum\nolimits_{rB}\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{cBk}}\right) \end{aligned} \tag{6}$$

模型 M_0 为假设条件 A 和条件 B 的基因表达水平相同,即 $u_{gA}=u_{gB}$,则 $\hat{\alpha}_{cA}=\hat{\alpha}_{cB}=\hat{\alpha}_g$。而模型 M_1 为假设条件 A 和条件 B 的基因表达水平不同,即 $u_{gA}\neq u_{gB}$,两个条件下的基因表达水平存在差异。

由于 S_{gA} 和 S_{gB} 两个条件表达水平的分布独立,所以 S_{gA} 和 S_{gB} 的联合分布为二者分布的乘积,则模型 M_0 和 M_1 的似然分别为

$$P\{D\mid M_0\}=P\{S_{gA}=s_{gA},S_{gB}=s_{gB}\mid M_0\}=P\{S_{gA}=s_{gA}\mid M_0\}\times P\{S_{gB}=s_{gB}\mid M_0\} \tag{7}$$

式中:$S_{gA}\sim\mathrm{NB}\left(\sum_{rA}\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{gk}}\right)$;$S_{gB}\sim\mathrm{NB}\left(\sum_{rB}\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{gk}}\right)$。

$$P\{D\mid M_1\}=P\{S_{gA}=s_{gA},S_{gB}=s_{gB}\mid M_1\}=P\{S_{gA}=s_{gA}\mid M_1\}\times P\{S_{gB}=s_{gB}\mid M_1\} \tag{8}$$

式中:$S_{gA}\sim\mathrm{NB}\left(\sum_{rA}\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{cAk}}\right)$;$S_{gB}\sim\mathrm{NB}\left(\sum_{rB}\hat{c},\frac{\hat{d}}{\hat{d}+\sum_k \hat{\alpha}_{cBk}}\right)$。

定义贝叶斯因子为似然比

$$BF=\frac{P\{D\mid M_1\}}{P\{D\mid M_0\}} \tag{9}$$

如果 $BF>1$，则 PG_bayes 方法选择模型 M_1，认为条件 A 和条件 B 的基因表达水平存在差异；否则选择模型 M_0，即认为未发生差异表达。

对于异构体的差异表达分析，令 t_{gA} 和 t_{gB} 分别表示条件 A 和条件 B 的异构体表达水平，其负二项分布模型分别为

$$\begin{aligned} t_{gA} &\sim \mathrm{NB}\Big(\sum\nolimits_{rA}\hat{c},\frac{\hat{d}}{\hat{d}+\hat{\alpha}_{cAk}}\Big)\\ t_{gB} &\sim \mathrm{NB}\Big(\sum\nolimits_{rB}\hat{c},\frac{\hat{d}}{\hat{d}+\hat{\alpha}_{cBk}}\Big) \end{aligned} \tag{10}$$

同样按上述的基因差异表达分析方法，假设模型 M_0 为条件 A 和条件 B 的异构体表达水平相同，即 $u_{gA}=u_{gB}$，则 $\hat{\alpha}_{cAk}=\hat{\alpha}_{cBk}$；假设模型 M_1 为条件 A 和条件 B 的异构体表达水平不同，则 $\hat{\alpha}_{cAk}\neq\hat{\alpha}_{cBk}$。计算原假设 M_0 和备择假设 M_1 的似然，并由贝叶斯因子选出正确模型，进而做差异表达分析。

2 实验数据集

本文实验采用人类 SEQC 数据集、人类结肠癌数据集、人类乳腺癌数据集以及小鼠数据集来验证 PG_bayes 方法在基因和异构体两方面的差异检测性能。

2.1 SEQC 数据集

SEQC 数据集[14]是最新的标准数据集，包含了通用人类参考 RNA(Universal human reference RNA，UHRR)和人类大脑参考 RNA(Human brain reference，HBRR)两个不同条件下的双末端读段数据。两种样本数据分别进行了 8 次重复实验。该数据集提供了进行过 qRT-PCR 实验上千个基因以及上万个异构体，qRT-PCR 实验验证基因和异构体在两个条件下的表达水平是否存在明显差异。Bullard 等[15]从中筛选出 305 个基因用于差异基因检测，其中包括 218 个显著差异基因和 87 个未发生显著差异的基因。本文按照 Bullard 等提出的方法筛选出用于差异检测的异构体样本。首先从 qRT-PCR 实验结果中筛选出 3 000 个具有单一异构体表达值的样本，取 UHRR 样本和 HBRR 样本表达值的对数比。对数比的绝对值小于 0.2 的异构体认为是未发生显著差异表达，对数据比绝对值大于 2 的异构体认为是发生显著差异表达。最终筛选出 1 002 个样本数据，包括 643 个显著差异表达的异构体和 359 个未发生显著差异表达的异构体，在本文中用于异构体的差异检测实验。

2.2 人类结肠癌数据集

人类结肠癌数据集(Griffith 数据集)[16]中包含了来自具有抗药性的人类结肠癌细胞(MIP101)和不具有抗药性的人类结肠癌细胞(MIP/5-FU)两个不同条件的双末端读段数据。两种样本数据分别进行了 7 次重复实验。Yu 等[17]从 Griffith 数据集中筛选 34 个多剪切异构体基因进行差异基因检测，qRT-PCR 实验结果表明，其中包含 20 个显著差异表达基因，14 个未发生显著差异表达基因。

2.3 人类乳腺癌数据集

人类乳腺癌数据集(Human breast cancer，HBC)[18]包含了来自人类乳腺癌细胞(MCF-7)和正常细胞(HME)两个条件的样本数据。MCF-7 样本进行了 4 次重复实验，HME 样本进行了 7 次重复实验。Wang 等针对这个数据集提取 8 个异构体进行了 qRT-PCR 实验，经验证这 8 个异构体均表现为显著差异。

2.4 小鼠数据集

Winnie 等[19]采用小鼠数据集验证了小鼠神经胶质瘤与基因 F11R 有关，该数据集包含来自脑干胶

质细胞和骨髓单核细胞两个条件的样本数据,每个条件分别进行了 3 次生物重复实验。文中针对小鼠数据集做了 4 种差异基因检测实验,包括针对 RNA-Seq 数据的 Cufflinks,ALEXA-Seq 以及针对基因芯片数据的 Aroma 和 Partek,挑选出被以上 3 种及以上的方法共同判定为差异表达的 1 178 个基因。这些基因是具有高置信度的差异基因,本文将它们做为衡量差异检测方法性能的标准。

3 实验结果

本文在 4 个真实数据集上做差异基因和差异异构体检测实验,与目前流行的差异检测方法 MMDiff,Cuffdiff,BitSeq 以及 PG_exact test 做对比,通过灵敏度和准确度两个标准验证 PG_bayes 方法的性能。其中 PG_exact test 同样以 PGseq 模型为基础,它利用 PGseq 计算出的基因和异构体表达水平的分布模型,采用 PG_exact test 方法进行差异基因和异构体检测。

小鼠数据集和 HBC 数据集中只包含差异基因或差异异构体,在本文中用于验证不同差异检测方法灵敏度。根据检测方法判定差异的标准,统计正确识别的差异基因或异构体的个数,进而得到差异识别的灵敏度。SEQC 数据集和 Griffith 数据集中的基因或异构体数据分为差异表达和未发生差异表达两类,并且有 qRT-PCR 实验标准,本文针对这两个数据集上的实验结果绘制受试者工作特征(Receiver operating characteristic,ROC)曲线,验证差异检测方法的灵敏度和准确度。

3.1 差异基因检测实验

3.1.1 小鼠数据集验证结果

小鼠数据集筛选出的 1 178 个具有高置信度的差异基因用于差异基因检测,检测不同方法在识别差异基因方面的灵敏度,不同检测方法识别出差异基因的结果统计如表 1 所示。灵敏度反映了差异检测方法识别差异基因的能力,灵敏度越高,检测方法对差异基因越敏感,从实验结果上看,PG_bayes 方法对差异基因最为敏感,灵敏度达 99%,而 Cuffdiff 和 BitSeq 方法的灵敏度较低。小鼠数据集验证了 PG_bayes 方法对差异基因具有较高的灵敏度,但灵敏度高不能证明检测方法性能最优,故进一步用 SEQC 数据集和 Griffith 数据集来检验 PG_bayes 方法的准确度。

表 1 不同方法在小鼠数据集上的灵敏度

Tab. 1 Sensitivities from different methods for mouse dataset

方法	PG_bayes	PG_exact test	MMDiff	BitSeq	Cuffdiff
识别个数	1 170	1 099	1 083	1 037	1 018
灵敏度/%	99.32	93.29	91.94	88.04	86.42

3.1.2 SEQC 数据集验证结果

对于 SEQC 数据集,筛选出具有 qRT-PCR 标准的 305 个基因用于差异基因检测,验证不同差异检测方法的灵敏度和准确度。对实验结果绘制 ROC 曲线,计算曲线下面积(Area under the curve,AUC)。ROC 曲线如图 1 所示,曲线下面积如表 2 所示。

表 2 不同方法在 SEQC 数据集上的 AUC 结果

Tab. 2 Area under ROC curves from different methods for SEQC dataset

方法	PG_bayes	PG_exact test	MMDiff	BitSeq	Cuffdiff
AUC	0.965 0	0.963 7	0.955 6	0.851 0	0.807 0

图 1 和表 2 显示了不同差异检测方法的实验结果与 qRT-PCR 实验结果的一致程度。ROC 曲线下面积 AUC 越接近于 1,表明获得的结果越接近 qRT-PCR 实验结果,差异检测的准确度就越高。实验结果显示,PG_bayes 方法的准确度较高,基于 PGseq 模型 PG_exact test 方法和 MMDiff 也具有较高的准

确率，显著高于其他方法。

为了直观表现差异检测方法在 SEQC 数据集上的检测结果，以 305 个基因在两个条件下的对数表达水平为坐标画出散点图，基因分布如图 2 所示。未发生显著差异表达的基因分布在对角线上，显著差异表达的基因用“Δ”符号标识，5 种差异检测方法共同识别出的 176 个差异基因，在图中用“＋”符号标识。根据共同识别的差异基因分布情况，用直线 L 将分布图分为 2 个区间，L 以下的部分属于低表达区间，L 以上的部分属于中高表达区间。由图中可以看出，差异检测方法在低表达区间的性能较差，1/2 左右的差异基因未被任何一个方法识别，差异检测的准确度较低。而在中高表达区间，所有差异检测方法均有较高的准确率，只有个别的差异基因未被识别。由此可见，低表达区间的差异表达检测难度较高。

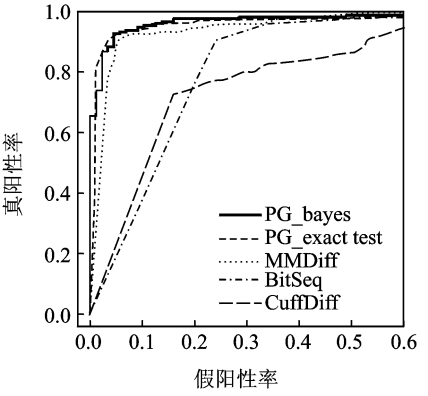

图 1 不同方法在 SEQC 数据集的 ROC 曲线

Fig. 1 ROC curves from different methods for SEQC dataset

为进一步验证 PG_bayes 方法在低表达区间上的性能，针对分布于低表达区间的 114 个基因数据做差异基因检测实验，对实验结果绘制 ROC 曲线，计算曲线下面积。ROC 曲线如图 3 所示，曲线下面积如表 3 所示。由 ROC 曲线上可以看出，PG_bayes 方法在 SEQC 数据集的低表达区间上的灵敏度明显高于其他方法。由表 3 统计的结果，PG_bayes 方法在低表达区间上的准确度也最高。由于不同差异检测方法在中高表达区间都有很高的准确度，对比 SEQC 数据集全集上的实验结果可知，差异检测方法在全集的准确度与低表达区间的准确度呈正相关，低表达区间的准确度越高，全集的准确度越高。因此，提高差异检测方法性能的关键是提高在低表达区间的灵敏度和准确度，而本文的方法在该区间的表现优于其他方法。

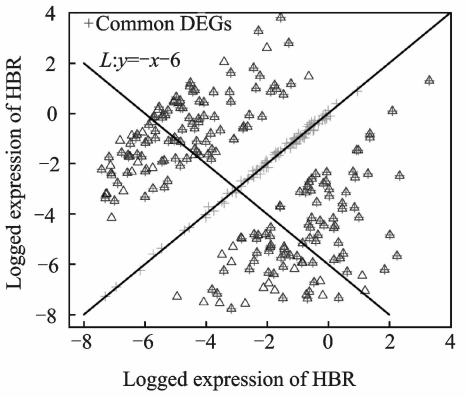

图 2 SEQC 数据集上的差异基因分布图

Fig. 2 Partition of qRT-PCR validated genes in SEQC dataset

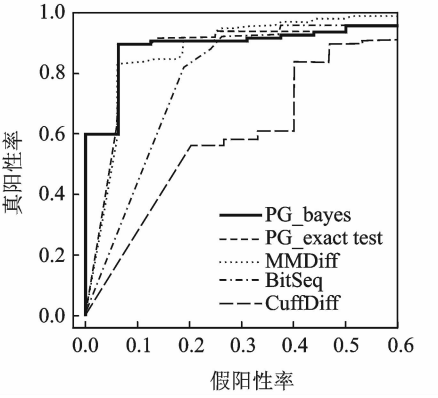

图 3 不同方法在 SEQC 数据集低表达区间上的 ROC 曲线

Fig. 3 ROC curves from different methods for SEQC datase with low expression

表 3 不同方法在 SEQC 数据集低表达区间上的 AUC 结果

Tab. 3 Area under ROC curves from different methods for SEQC dataset with low expression

方法	PG_bayes	PG_exact test	MMDiff	BitSeq	Cuffdiff
AUC	0.926 5	0.912 0	0.916 4	0.853 3	0.726 8

3.1.3 Griffith 数据集验证结果

对于 Griffith 数据集,在筛选出的有 qRT-PCR 标准的 34 个基因上进行差异基因检测,验证不同差异检测方法的灵敏度和准确度。同样与上述 4 种差异检测方法对比,绘制 ROC 曲线,比较曲线下面积。5 种方法 Griffith 数据集上的 ROC 曲线如图 4 所示,AUC 大小如表 4 所示。

从实验结果上看,在 Griffith 数据集上,PG_bayes 方法具有较高的准确度和灵敏度。与其他方法对比分析,PG_exact test 方法的准确度与 PG_bayes 方法较为接近,从 ROC 曲线上可以看出,该方法的灵敏度明显低于 PG_bayes 方法,而在 SEQC 数据集上表现出较好性能的 MMDiff 方法在该数据集上的准确度和灵敏度则较低。

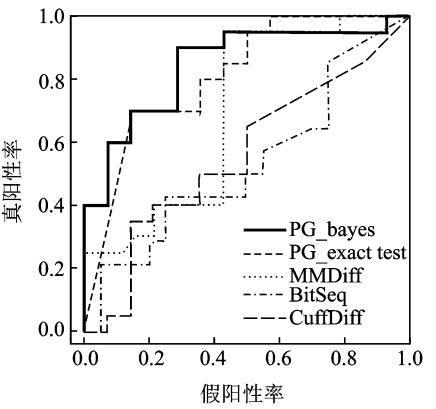

图 4 不同方法在 Griffith 数据集上的 ROC 曲线

Fig. 4 ROC curves from different methods for Griffith dataset

3.2 差异异构体检测实验

3.2.1 HBC 数据集验证结果

对于 HBC 数据集,在 qRT-PCR 实验验证的 8 个异构体上进行差异异构体检测,验证不同差异检测方法的灵敏度。异构体在两个不同条件下差异检测,实验结果如表 5 所示。其中对两个异构体表达差异显著的标为 DE(Differential expression),不显著的为 NDE(None differential expression),并且标注了表达值的变化方向,“+”表示在相应的比较条件中异构体表达上调(Up-regulation),“−”表示下调(Down-regulation)。qRT-PCR 实验结果表明这 8 个异构体均为上调的差异表达,即 MCF-7 条件下异构体的表达水平高于 HME。BitSeq 方法采用 PPLR 值衡量结果,当 PPLR>0.95 时,该方法认为异构体差异表达。PG_exact test,Cuffdiff 以及 MMDiff 方法采用 p-value 表示 DE 的显著性,PG_exact test 和 Cuffdiff 认为当 p-value<0.05 时为差异表达,而 MMDiff 认为 p-value>0.95 时为差异表达。从结果上看,PG_bayes 方法的灵敏度最高,能正确识别全部差异异构体,MMDiff 和 BitSeq 方法实验结果比较理想,均检测出 7 条差异表达的异构体,而 Cuffdiff 方法的灵敏度最低,没能识别出差异异构体。

表 4 不同方法在 Griffith 数据集上的 AUC 结果

Tab. 4 Area under ROC curves from different methods for Griffith dataset

方法	PG_bayes	PG_exact test	MMDiff	BitSeq	Cuffdiff
AUC	0.846 4	0.810 7	0.696 4	0.533 9	0.557 1

表 5 HBC 数据集的差异异构体检测结果

Tab. 5 Results of isofroms between two conditons in HBC dataset

异构体	qRT-PCR	PG_bayes	PG_exact test	MMDiff	BitSeq	Cuffdiff
uc002cvt.2	DE+	DE+(2)	DE+(0.00)	DE+(0.96)	DE+(0.99)	NDE(0.99)
uc002cvs.1	DE+	DE+(1.3)	NDE(0.46)	NDE(0.71)	NDE(0.84)	NDE(1.00)
uc002qlq.1	DE+	DE+(1.7)	DE+(0.00)	DE+(0.98)	DE+(0.99)	NDE(0.99)
uc002qlp.1	DE+	DE+(32)	DE+(0.00)	DE+(0.99)	DE+(0.98)	NDE(0.99)
uc002xmn.1	DE+	DE+(>100)	DE+(0.00)	DE+(1.00)	DE+(1.00)	NDE(1.00)
uc002xmo.1	DE+	DE+(>100)	DE+(0.00)	DE+(1.00)	DE+(1.00)	NDE(0.85)
uc003ngr.1	DE+	DE+(1.3)	NDE(0.21)	DE+(0.96)	DE+(1.00)	NDE(0.24)
uc003ngs.1	DE+	DE+(>100)	DE+(0.00)	DE+(1.00)	DE+(1.00)	NDE(0.64)
识别个数	—	8	6	7	7	0

3.2.2 SEQC 数据集验证结果

在 HBC 数据集中，验证了 PG_bayes 方法对差异异构体具有较高的灵敏度，进一步采用 SEQC 数据集来检验 PG_bayes 方法的准确度。对于 SEQC 数据集，在挑选出的有 qRT-PCR 标准的 1 002 个异构体上进行差异基因检测。同样与上述 4 种差异分析方法对比分析，并绘制 ROC 曲线，比较曲线下面积。5 种方法 SEQC 数据集上的 ROC 曲线如图 5 所示，AUC 大小如表 6 所示。

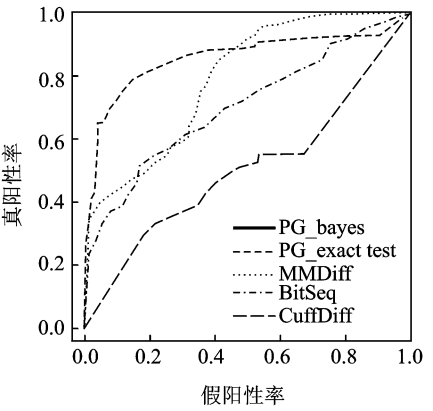

图 5 不同方法在 SEQC 数据集上的 ROC 曲线

Fig. 5 ROC curves from different methods for SEQC dataset

由于读段的多源映射，计算异构体的表达水平有一定困难，而表达水平计算结果的准确性会影响后续的差异检测，所以总体上异构体差异检测的准确度不如基因检测的高。从实验结果上看，PG_bayes 仍然表现出较好的性能，其准确率和灵敏度都很理想。与其他方法相比，PG_bayes 在差异异构体检测方面表现出优势，一方面由于 PGseq 模型在表达水平计算时很好地解决了多源映射问题，为差异检测分析奠定了基础；另一方面，与 PG_exact test 对比，PG_bayes 的灵敏度和准确度均有提升，因此贝叶斯因子方法适合基于 PGseq 模型进行差异分析。

表 6 不同方法在 SEQC 数据集上的 AUC 结果

Tab. 6 Area under ROC curves from different methods for SEQC dataset

方法	PG_bayes	PG_exact test	MMDiff	BitSeq	Cuffdiff
AUC	0.870 7	0.857 4	0.790 4	0.703 3	0.505 2

4 结束语

本文针对 RNA-seq 数据分析中的多源映射问题，以 PGseq 计算出的基因和异构体表达水平的分布模型为基础，采用贝叶斯因子方法进行模型选择，解决了差异基因和差异异构体检测问题。综合在 4 个数据集上的实验结果，PG_bayes 在差异基因检测方面具有一定优势，灵敏度高于其他方法，而且在不同数据集上都有较高的准确度。多源映射现象给异构体表达水平计算及后续的差异异构体检测带来了很大挑战，目前很多方法在差异异构体检测上的性能都较差，而 PG_bayes 方法在差异异构体检测方面表现出优势，准确度较其他方法有明显的提高，较好地解决了差异异构体检测问题。由于 PGseq 模型考虑了更多的数据信息，通过 Gamma 分布模拟数据偏差，并估计相应表达水平的不确定性，所以在后续差异表达分析中，能有效提高差异基因和异构体的识别。特别地，与同样基于 PGseq 模型 PG_exact test 方法相比，PG_bayes 方法在灵敏度和准确度方面上都有提升，表明其在转录组研究中较好的应用前景。

参考文献：

[1] Wang Z, Gerstein M, Snyder A M. RNA-Seq: A revolutionary tool for transcriptomics[J]. Nature Reviews Genetics, 2008, 10(1): 57-63.

[2] Richard H, Schulz M H, Sultan M, et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments[J]. Nucleic Acids Res, 2010, 38(10): e112.

[3] Wang L G, Xi Y X, Yu J, et al. A statistical method for the detection of alternative splicing using RNA-Seq[J]. PLoS one, 2010, 5-(1): e8529.

[4] 刘学军，李蒙，张礼. 一种针对 RNA-Seq 数据的基因异构体表达水平计算方法[J]. 中国生物医学工程学报，2013，4：454-463.

Liu Xuejun, Li Meng, Zhang Li. A method of isoform expression calculation for RNA-Seq data[J]. Chinese Journal of Biomedical Engineering,2013,7(4):454-463.

[5] Anders S, Huber W. Differential expression analysis for sequence count data[J]. Genome Biology,2010,11(10):R106.

[6] Hardcastle T J, Kelly K A. Bay-Seq: Empirical Bayesian methods for identifying differential expression in sequence count data[J]. BMC Bioinformatics,2010,11: 422-439.

[7] Turro E, Su S Y, Gonçalves Â, et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-Seq reads[J]. Genome Biol, 2011,12(2):R13.

[8] 石新新,刘学军,张礼.改进的 RNA-Seq 数据转录组表达分析研究[J].数据采集与处理,2015,30(5):1028-1035.
Shi Xinxin, Liu Xuejun, Zhang Li. Improved trancriptome expression analysis for RNA-Seq data[J]. Journal of Data Acquisition and Processing, 2015,30(5):1028-1035.

[9] Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-Seq data with biological variation [J]. Bioinformatics, 2012,28(13):1721-1728.

[10] Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks[J]. Nature Protocols, 2012,7(3):562-578.

[11] Jiang H, Wong W H. Statistical inferences for isoform expression in RNA-Seq[J]. Bioinformatics, 2009,25(8):1026-1032.

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

[13] Di Y, Schafer D W, Cumbie J S, et al. The NBP negative binomial model for assessing differential gene expression from RNA-Seq[J]. Statistical Applications in Genetics and Molecular Biology, 2011,10(1):1-28.

[14] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-Seq accuracy, reproducibility and information content by the sequencing quality control consortium[J]. Nature Biotechnology, 2014,32(9):903-914.

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

[16] Griffith M, Griffith O L, Mwenifumbo J, et al. Alternative expression analysis by RNA sequencing[J]. Nature Methods, 2010,7:843-847.

[17] Yu D, Huber W, Vitek O. Shrinkage estimation of dispersion in negative binomial models for RNA-Seq experiments with small sample size[J]. Bioinformatics, 2013,29(10):1275-1282.

[18] Wang E T, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes[J]. Nature, 2008,456(7221):470-476.

[19] Pong W W, Walker J, Wylie T, et al. F11R is a novel monocyte prognostic biomarker for malignant glioma[J]. PloS One, 2013,8(10):e77571.

作者简介:

王黎黎(1989-),女,硕士研究生,研究方向:生物信息学,E-mail: w1213231@163.com。

刘学军(1976-),女,教授,研究方向:机器学习与生物信息学。

张礼(1985-),男,博士研究生,研究方向:生物信息学。