多条件多样本RNA-Seq数据的剪切异构体表达水平估计

张礼， 马越， 吴东洋

引用本文:
张礼, 马越, 吴东洋. 多条件多样本RNA-Seq数据的剪切异构体表达水平估计[J]. 智能系统学报, 2021, 16(6): 1126-1135.
ZHANG Li, MA Yue, WU Dongyang. Estimation of transcription variant expression level based on multi-condition multi-sample RNA-Seq data[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(6): 1126-1135.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202101028

DOI: 10.11992/tis.202101028
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210831.1315.006.html

多条件多样本 RNA-Seq 数据的剪切异构体表达水平估计

张礼[1], 马越[2], 吴东洋[1]
(1. 南京林业大学 信息科学技术学院, 江苏 南京 210016; 2. 江苏健康卫生职业学院 中西医结合学院, 江苏 南京 210018)

摘 要: 当处理多条件多样本 RNA-Seq 测序数据时, 现有方法忽略了读段分布样本之间存在高度相似性的特点。本文提出了一个基于多条件多样本 RNA-Seq 测序数据剪切异构体表达水平估计方法 MCMS-Seq。该方法建立了一个联合偏差估计模型来提取读段分布在样本之间的相似性特征, 同时考虑读段分布受全局偏差和局部偏差的影响。此外, 增加了 L_2/L_1 组稀疏约束和 L_1 稀疏约束两个正则化项, 用来体现基因和剪切异构体之间存在稀疏特性, 以及消除技术性误差和数据噪声的影响。通过多个真实数据集的验证, MCMS-Seq 方法能获得更为准确的剪切异构体表达水平, 同时也能提供更有意义的生物性解释。

关键词: 转录组测序技术; 多条件; 多样本; 剪切异构体; 表达水平估计; 稀疏特性; 读段分布偏差; 数据噪声
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2021)06−1126−10

中文引用格式: 张礼, 马越, 吴东洋. 多条件多样本 RNA-Seq 数据的剪切异构体表达水平估计 [J]. 智能系统学报, 2021, 16(6): 1126–1135.
英文引用格式: ZHANG Li, MA Yue, WU Dongyang. Estimation of transcription variant expression level based on multi-condition multi-sample RNA-Seq data[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1126–1135.

Estimation of transcription variant expression level based on multi-condition multi-sample RNA-Seq data

ZHANG Li[1], MA Yue[2], WU Dongyang[1]
(1. College of Information Science and Technology, Nanjing Forest University, Nanjing 210016, China; 2. College of Integrated Chinese and Western Medicine, Jiangsu Health Vocational College, Nanjing 210018, China)

Abstract: When analyzing multi-condition multi-sample RNA-sequencing (MCMS RNA-Seq) data, the existing methods for estimating transcription variant expression levels ignore the high similarity between read distribution samples. Thus, this study proposes a method for estimating transcription variant expression levels based on MCMS-Seq data. A joint bias estimation model was developed to extract read distribution similarity between samples, considering the influence of both global and local biases on read distribution at the same time. In addition, two regularization items, L_2/L_1 and L_1 sparse constraints, were added to reflect sparsity characteristics between genes and transcription variants and to eliminate the influence of technical errors and data noise. This method allows a more accurate estimation of transcription variant expression levels based on MCMS-Seq data and provides more meaningful biological interpretations.

Keywords: RNA-Seq; multi-condition; multi-sample; isoform; expression estimation; sparsity; read bias; data noise

收稿日期: 2021−01−18. **网络出版日期:** 2021−08−31.
基金项目: 国家自然科学基金项目 (61802193); 江苏省自然科学基金项目 (BK20170934); 南京林业大学青年科技创新基金项目 (CX2017031); 汕尾市省级科技创新战略专项资金项目 (2018D2002).
通信作者: 张礼. E-mail: lizhang@njfu.edu.cn.

选择性剪切事件是导致生物体多样性的重要原因之一。为了进一步揭示选择性剪切的内在机制, 迫切需要计算剪切异构体的表达水平。与传统的基因芯片技术相比, 高通量 RNA 测序 (RNA

sequencing, RNA-Seq) 技术具有高通量、高灵敏度、可重复性好等优势, 已成为转录组学分析的一个标准技术手段[1-5]。

RNA-Seq 测序实验获得海量读段, 将读段与参考注释序列进行匹配, 之后便可估计剪切异构体的表达水平。但是在估计剪切异构体表达水平的过程中, 面临着两个最大挑战, 即读段的多源映射和数据偏差[6-7]。研究者提出了大量剪切异构体表达水平估计方法来解决上述的问题。rSeq 方法把读段映射到外显子的过程当作一个泊松随机过程, 其泊松分布的参数对应着基因所包含剪切异构体表达水平的线性加权[8]。但是 rSeq 方法假设基因上读段分布是均匀的, 这与真实数据分布特点不一致。在真实数据中, 读段分布呈现明显的非均匀特征。读段的非均匀分布主要是由测序数据中的各种偏差造成的, 比如 GC 碱基序列偏差, 5 端和 3 端的位置偏差以及实验技术性偏差等。针对偏差所导致问题, NURD 方法考虑了全局和局部位置偏差所带来的影响[9]。POME 方法考虑了序列中碱基之间的关联性[10]。为了考虑更复杂的偏差, 大量概率生成式模型被提出, 其直接模拟读段的随机采样过程。Cufflinks 方法设计了不同的模型来消除序列偏差和位置偏差的影响, 从而更加准确地描述读段随机采样过程[11]。BitSeq 和 PBSeq 方法采用了与 Cufflinks 同样的偏差估计模型[12-13]。RSEM 方法考虑了读段匹配的不确定性因素, 并且使用了读段起始位置的经验分布来表示读段在基因上的非均匀分布特征, 但是其未考虑序列偏差这个重要因素[14]。上述方法采用不同的偏差估计模型来模拟读段的非均匀分布特征, 都能提高剪切异构体表达水平的估计准确程度。

由于数据噪声和偏差的影响, 异构体表达水平的准确性仍然有较大提高的空间[15-16]。常规的 RNA-Seq 测序实验通常会设置不同的实验条件, 比如: 同一个细胞组织下参照组和对照组, 不同时间点下胚胎发育状况等。此外为了避免实验中的技术性误差, 同一个实验条件下会进行多次重复性技术性实验。这使得一次测序实验获得的 RNA-Seq 数据集是一个多条件多样本的数据集。但是上述方法都是假设 RNA-Seq 数据集中各个样本之间是相互独立, 因此都是单独逐个处理每个数据样本。这导致样本之间的相关性没有得到充分利用。因此有少量工作开始探索联合多样本 RNA-Seq 数据进行异构体表达水平估计[17-18]。Sequgio 方法能从多样本数据中自动获取位置偏差和局部序列影响, 再通过对联合统计模型添加一个光滑的正则化项, 来控制读段在多样本的一致性[19]。MSIQ 方法考虑多样本之间的异质性所导致的结果不稳定性, 首先将同质性相近的样本归为同一组, 然后在贝叶斯框架模型下, 给同一组之内的样本赋予较高的权重, 从而获得更加鲁棒的异构体表达水平[20]。XAEM 方法采用双线性模型同时估计异构体表达水平和数据偏差, 该模型能够自动对潜在的未知偏差进行经验校正[21]。但是上述方法所处理的多样本数据, 仅仅是针对单条件下的多样本, 比如同一个组织细胞的对照组或者同一个时间点状态。当处理多条件多样本数据时, 这些方法都是假设各个条件之间不相关, 把多条件多样本数据拆分为多个单条件多样本数据集来进行异构体表达水平计算。但是基因读段分布在不同条件下同样具有高度相似性[22]。为了充分利用数据信息, PGSeq 方法采用泊松分布和伽玛分布的混合模型联合估计基因和异构体表达水平, 其伽玛分布用来模拟基因读段分布在多条件多样本下的偏差信息[23]。但 PGSeq 方法未考虑到基因和异构体表达水平之间的稀疏特性, 易受到数据噪声的影响。

基于上述问题, 本文提出了一个多条件多样本 RNA-Seq 测序数据异构体表达水平估计方法, MCMS-Seq(multi-condition multi-sample RNA-Seq)。该模型考虑了基因读段分布在不同条件下的样本具有高度相似性, 设计一个联合多条件多样本数据的偏差估计模型, 同时考虑了基因读段分布受全局偏差和局部偏差的影响。此外, MCMS-Seq 方法增加了 L_2/L_1 组稀疏约束和 L_1 稀疏约束两个正则化项, 用来体现基因和剪切异构体之间存在稀疏特性, 以及消除技术性误差和数据噪声的影响。最后, 通过 3 个多条件多样本 RNA-Seq 数据集来评估 MCMS-Seq 方法的性能。

1　MCMS-Seq 方法

1.1　MCMS-Seq 模型表示

由于选择性剪切事件在真核生物中普遍存在, 这给计算剪切异构体表达水平带来了一个最大问题, 即如何定量确定匹配到共享外显子上的读段来自哪个剪接异构体。图 1 中显示的基因包含 4 个外显子 (Exon) 和 3 个剪切异构体。其中一个外显子可以同时被多个剪切异构体共享, 比如外显子 1 被 3 个剪切异构体共享, 但是剪切异构体 2 仅共享了外显子 1 的部分序列。针对这类部分共享情况, 可将外显子 1 分割为 2 个不重叠的

外显子片段。因此该基因的 4 个外显子被分割成 7 个完全不重叠的外显子片段。映射矩阵 $\boldsymbol{A}$ 表示图 1 中剪切异构体与外显子片段的关系，其中矩阵元素 $a_{12}=1$ 表示异构体 1 包含外显子片段 2。当测序读段匹配到基因上，外显子片段上的读段数目即可被统计出来。假设某个数据集有 2 个条件每个条件包括 2 个样本，总计 4 个样本，那么图 1 中基因在不同样本中读段数据可用数据矩阵 $\boldsymbol{D}$ 表示。每一行表示该基因在一个样本中外显子片段的读段数目。

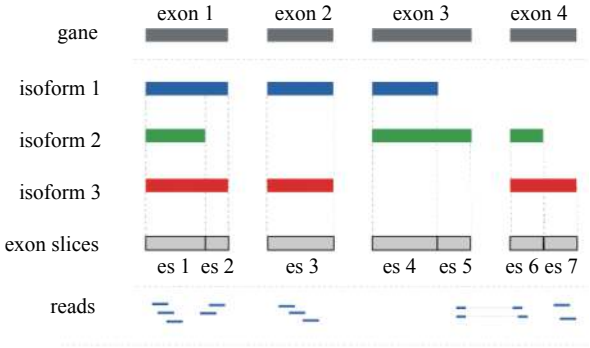

图 1　剪切异构体中外显子片段划分示例

Fig. 1　Example of exon segmentation in an isoform

$$\boldsymbol{A}=\begin{bmatrix}1&1&1&1&0&0&0\\1&0&0&1&1&1&0\\1&1&1&0&0&1&1\end{bmatrix}\quad \boldsymbol{D}=\begin{bmatrix}2&1&2&0&1&1&2\\4&1&2&0&2&2&1\\8&10&1&8&1&6&8\\9&11&1&7&0&7&8\end{bmatrix}$$

假设测序实验获得 RNA-Seq 数据包含 C 个条件，每个条件包含 N 个样本。对于基因 g，该基因包含 K 个剪切异构体和 M 个外显子片段，其与外显子片段的映射关系由映射矩阵 $\boldsymbol{A}_{M\times K}$ 表示。y_{cij} 表示基因第 j 个外显子片段在第 c 个条件中第 i 个样本上的读段数目。根据实验原理，基因外显子片段的读段数目等于共享该外显子片段的剪切异构体上读段之和，其数学模型为

$$y_{cij}=w_{ci}l_j\sum_{k=1}^{K}a_{jk}x_{cik}\tag{1}$$

式中：x_{cik} 表示基因第 k 个剪切异构体在第 c 个条件中第 i 个样本；a_{jk} 表示剪切异构体与外显子片段之间的映射关系；w_{ci} 表示第 c 个条件中第 i 个样本的读段总数；l_j 是第 j 个外显子片段的长度。

式 (1) 模型是基于基因读段是均匀分布假设的前提，但是实际数据中，基因读段分布呈现明显的非均匀特征。由于基因读段分布模式在不同条件不同样本下具有高度相似性，因此假设 b_j 表示第 j 个外显子片段的偏差权重，其值在样本之间是共享的。现将偏差 b_j 融入到式 (1) 中，得到如下模型：

$$y_{cij}=w_{ci}l_jb_j\sum_{k=1}^{K}a_{jk}x_{cik}\tag{2}$$

对于多条件多样本的 RNA-Seq 数据集，基因 g 所包含的 K 个剪切异构体的表达水平 $\boldsymbol{X}$ 可以通过回归模型计算，其公式如下：

$$\boldsymbol{X}^*=\arg\min_{x}\sum_{c=1}^{C}\sum_{j=1}^{N}\sum_{i=1}^{M}\left(\frac{y_{cij}}{w_{ci}l_jb_j}-\sum_{k=1}^{K}a_{jk}x_{cjk}\right)\tag{3}$$

所有剪切异构体在不同样本中的表达水平都要求 $x_{cjk}\geqslant 0$。为了便于理解和计算，式 (3) 可以简化为矩阵形式：

$$\boldsymbol{X}^*=\arg\min_{x}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{X}\|_F^2\tag{4}$$

式中 $\boldsymbol{D}$ 表示归一化后的数据矩阵。

一个基因虽然包含多个剪切异构体，但是在不同条件下，少数剪切异构体的表达水平决定了该基因的表达。因此基因和剪切异构体表达水平之间具有稀疏特性。通过对剪切异构体表达水平 $\boldsymbol{X}$ 增加 L_1 范数来保留稀疏特性，式 (4) 可改写为

$$\boldsymbol{X}^*=\arg\min_{x}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{X}\|_F^2+\lambda\|\boldsymbol{X}\|_1\tag{5}$$

虽然模型增加了 L_1 范数的稀疏约束，但仍然会发现出现大量低表达的剪切异构体，而这部分剪切异构体不全是真实的低表达。当一个剪切异构体在同一个条件下的所有重复样本都是低表达水平，那么可认为此剪切异构体是真实的低表达。而对于零散出现的低表达剪切异构体，则受到数据噪声和偏差的影响，不是真实的低表达。为了消除虚假的低表达剪切异构体的影响，在式 (5) 的基础上增加了 L_2/L_1 组稀疏约束得到了 MCMS-Seq 方法的最终形式：

$$\begin{aligned}&\boldsymbol{X}^*=\arg\min_{x}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{X}\|_F^2+\lambda_1\|\boldsymbol{X}\|_{2,1}+\lambda_2\|\boldsymbol{X}\|_1\\&\text{s.t. }x_{cjk}\geqslant 0\end{aligned}\tag{6}$$

式中 λ_1 和 λ_2 分别是 L_2/L_1 和 L_1 约束的系数。通过两个稀疏约束项，MCMS-Seq 方法不仅考虑了基因和剪切异构体表达水平之间的稀疏性质，同时也可以消除数据噪声和偏差对低表达剪切异构体的影响。图 2 显示了 MCMS-Seq 方法的优化问题。

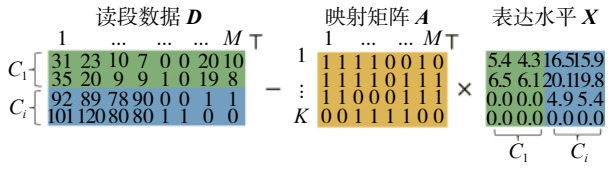

图 2　MCMS-Seq 方法的优化问题

Fig. 2　Optimization problem of MCMS-Seq method

1.2　多条件多样本偏差估计模型

在多条件多样本数据中，图 3 显示了基因的读段分布无论在不同条件下，还是在同一个条件的重

复样本中, 其分布模式具有高度相似性。MCMS-Seq 方法提出了一个基于多条件多样本的读段非均匀偏差估计模型。该偏差估计模型由两部分构成: 全局偏差 β_{global} 和局部偏差 β_{local}。全局偏差 β_{global} 的读段非均匀分布模式是从数据集中所有表达基因中获得。由于读段多源映射会影响基因读段分布, 全局偏差估计仅仅选择只包含单个剪切异构体的基因。此外, 由于低表达水平基因的不确定性, 读段计数小于 50 的基因被排除。将筛选后的基因均分为 20 个等长度的区间, 统计并归一化每个区间内读段数目。最后采用多项式回归来拟合基因每个区间上的读段数目, 得到的拟合曲线表示基因读段分布的全局偏差特征。而局部偏差 β_{local} 仅仅统计基因每个外显子片段在多条件多样本数据上的读段数目, 再进行均一化处理, 其反映了单个基因自身的读段分布特征。

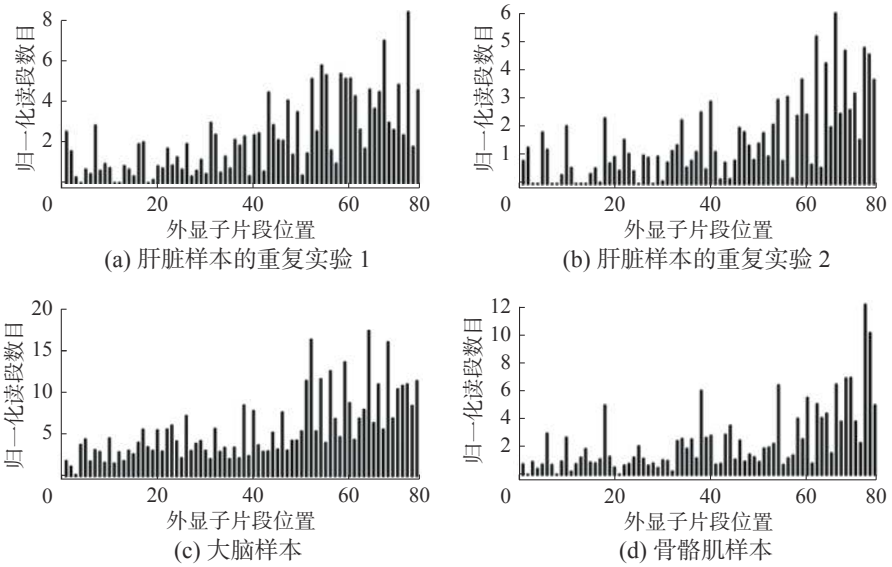

图 3　小鼠数据集中基因 Utrn 读段分布

Fig. 3　Read distributions of gene Utrn in the mouse dataset

一旦获得数据集的全局偏差曲线和单个基因的局部偏差特性, 便可以计算出基因上每个外显子片段的偏差值:

$$b_j = \alpha\beta_{\text{global}} + (1-\alpha)\beta_{\text{local}} \tag{7}$$

式中: α 是权重参数, 用来权衡全局偏差和局部偏差的影响。本文选择 $\alpha = 0.5$, 表示全局偏差和局部偏差对基因读段分布具有相同的影响[9], 不仅仅能反映读段非均匀分布在多条件多样本之间具有高度相似的特征, 同时还可以体现出每个基因独有读段分布特点。

1.3　MCMS-Seq 模型实现

MCMS-Seq 方法的实现可以分为 3 个部分: 读段数据预处理、基因偏差估计和表达水平估计。

1) 读段数据预处理, 是从匹配成功的读段数据中统计基因每个外显子片段的读段计数, 以及从注释文件中获得外显子片段和剪切异构体之间的映射关系矩阵。

2) 基因偏差估计, 是计算数据集的全局偏差和基因的局部偏差, 从而获得基因每个外显子片段的基因偏差值。

3) 剪切异构体表达水平估计, 由于模型是针对多条件多样本数据集, 同时模型包含 L_2/L_1 和 L_1 约束, MCMS-Seq 方法采用 SPAMS 优化工具箱来求解[24-25]。

MCMS-Seq 方法的详细流程如算法 1 所示, 采用 Python 和 MATLAB 混合编程实现。

算法 1　MCMS-Seq 方法

输入　多条件多样本数据, 注释文件;

输出　每个基因的剪切异构体表达水平。

1) 数据预处理: 统计外显子片段读段数目矩阵 $\boldsymbol{D}$, 构建映射关系矩阵 $\boldsymbol{A}$。

2) 基因偏差估计: 计算外显子片段偏差值。

3) 表达水平估计: 计算所有基因的 $\boldsymbol{X}^*$。

1.4　多条件多样本数据分析通道

为了方便用户使用 MCMS-Seq 方法, 本文提供了一个多条件多样本 RNA-Seq 测序数据分析通道, 如图 4 所示。当获得 RNA-Seq 测序数据样本后, 使用经典读段匹配软件 Bowtie[26], 将每个数据样本的读段匹配到参考转录组参考序列上。每个样本匹配结果作为输入数据一并输入到 MCMS-Seq 分析通道中, 从而可获得剪切异构体在不同样本中的表达水平。一旦获得剪切异构体的表达

水平, 可提供给高层次的后续分析使用。

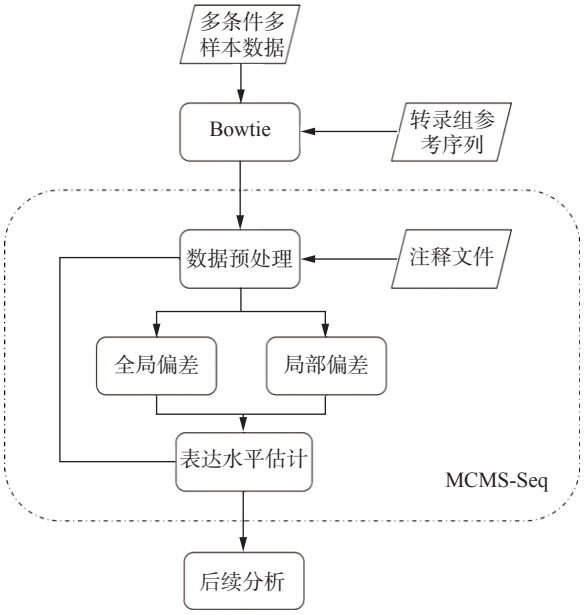

图 4　多条件多样本 RNA-Seq 数据分析通道

Fig. 4　Analysis pipeline of multi-condition multi-sample RNA-Seq data

2　实验结果与分析

本文选择了经典方法 Cufflinks(v.2.2.1) 和 PG-Seq(v.1.0), 以及最新方法 XAEM(v.0.1.1), 分别在 3 个数据集上与 MCMS-Seq 方法进行比较, 用来验证剪切异构体表达水平的性能。针对多条件多样本数据集, Cufflinks 是每个样本单独处理, 而 PGSeq、XAEM 和 MCMS-Seq 都是多个样本联合处理。

2.1　数据集

3 个多条件多样本的 RNA-Seq 数据集被用来验证 MCMS-Seq 方法估计剪切异构体表达水平的准确性。3 个数据集分别是小鼠数据集、人类大脑的 SEQC 和 MAQC-II 数据集, 它们都来自 Illumina/solexa 测序平台。

小鼠数据集包含 3 个条件, 分别是肝脏、大脑和骨骼肌 3 个组织, 其中每个组织分别包含了 2 个重复实验样本。使用 RefSeq 数据库的基因注释信息 (GRCm38/mm10), 总共包含 33 608 个剪切异构体, 主要用来验证同条件下重复样本之间剪切异构体表达水平的可重复性[27]。

MAQC(micorarray quality control) 来自美国药品监管局的生物芯片质量控制项目。该项目分为三期实施, 即 MAQC-I、MAQC-II 和 MAQC-III, 其产生的数据集被广泛应用于评估不同测序平台下不同方法的性能。本文主要利用了 MAQC-II 和 MAQC-III 两期项目提供的数据。MAQC-III 也被称为 SEQC(sequencing quality control)。SEQC 包括两个实验条件 UHRR(universal human reference rna) 和 HBRR(human brain reference RNA), 每个条件分别有 8 个重复实验样本。SEQC 数据集提供了两万多个经 qRT-PCR 实验验证的剪切异构体。与 Ensembl 注释信息 (GRCh37/hg19) 相匹配后, 最终得到 16 603 个剪切异构体。这些剪切异构体的 qRT-PCR 值被当作真实表达水平值, 可用来评估模型计算剪切异构体表达水平的准确性[28]。

基因表达水平是由其包含的剪切异构体所构成, 因此基因表达水平可用来进一步验证剪切异构体表达水平的准确性。MAQC-II 数据集同样包含 UHRR 和 HBRR 两个实验条件, 每个条件下包含 7 个重复性实验。该数据提供了 1 000 个经 qRT-PCR 实验验证的基因。根据与 Ensembl 注释信息 (GRCh37/hg19) 相匹配, 最终获得 838 个基因。这些基因的 qRT-PCR 值被当作真实基因表达水平值, 用来间接评估模型计算剪切异构体表达水平的准确程度[29]。

2.2　多条件多样本偏差估计模型验证

MCMS-Seq 方法提出了一个基于多条件多样本偏差估计模型, 同时考虑了读段分布受到全局偏差和局部偏差的影响, 用来获取读段分布在样本之间的高度相似性特征。SEQC 数据集被用来验证偏差估计模型的有效性。图 5 显示使用该模型对 SEQC 数据集的偏差估计流程。从图 5(a) 中可以看出, 在 SEQC 数据集中, 基因的读段分布呈现明显的非均匀分布特征, 特别是在基因的两端。这个现象符合基因的3′端和5′端最容易受到 RNA-Seq 测序技术影响的事实。选择基因 Cdca4 来展示估计全局偏差和局部偏差的过程。基因 Cdca4 包含 3 个剪切异构体和 5 个外显子片段, 其结构如图 5(b) 所示。图 5(c) 是通过多项式回归拟合图 5(a) 读段分布所得到 SEQC 数据集全局偏差曲线。曲线上黑点表示基因 Cdca4 外显子片段长度的比率。通过长度比率在曲线上的取值, 可得到 Cdca4 基因中每个外显子片段的全局偏差值。统计并归一化基因 Cdca4 的外显子片段在所有样本中的读段数目, 即可获得该基因的局部偏差, 如图 5(d) 所示。从图 5 中可以看出, 该基因在3′端和5′端受到的局部偏差影响要略小于全局偏差。为了进一步验证基因的局部偏差, 从 SEQC 数据集中随机选择 4 个基因: Plagl1、Eif4a、Sv2b 和 Whrn, 其分别包含 5、6、7、8 个剪切异构体。从图 6 中可以看出, 不同基因的局部偏差整体上都呈现明显非均匀分布特征, 但是单个基因

之间存在一定差异, 比如基因 Whrn 中间外显子的偏差值表现出由高到低的趋势。因此, MCMS-Seq 方法提出的多条件多样本偏差估计模型, 不仅能反映在多条件多样本数据中读段非均匀分布具有高度相似性的特征, 同时还可以体现出单个基因独有读段分布特点。

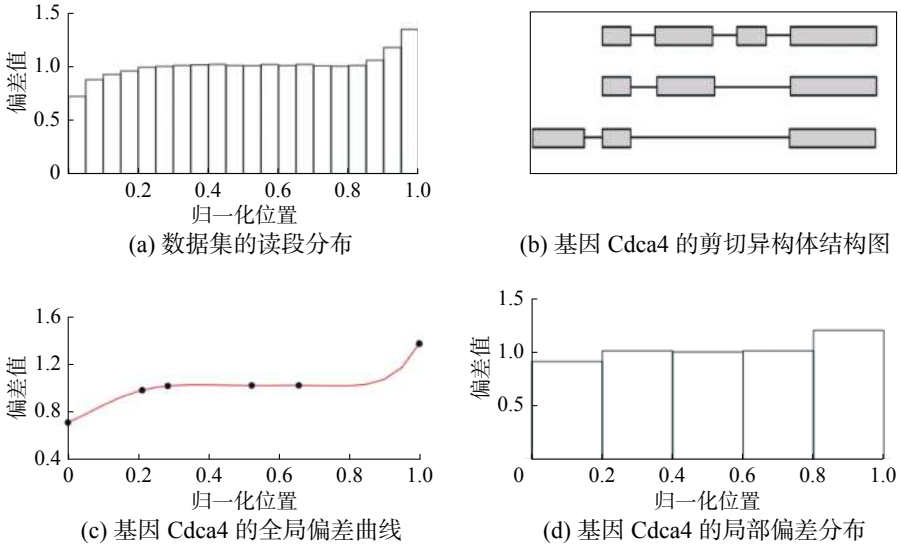

图 5　MCMS-Seq 方法的偏差估计流程

Fig. 5　Bias estimation process of the MCMS-Seq method

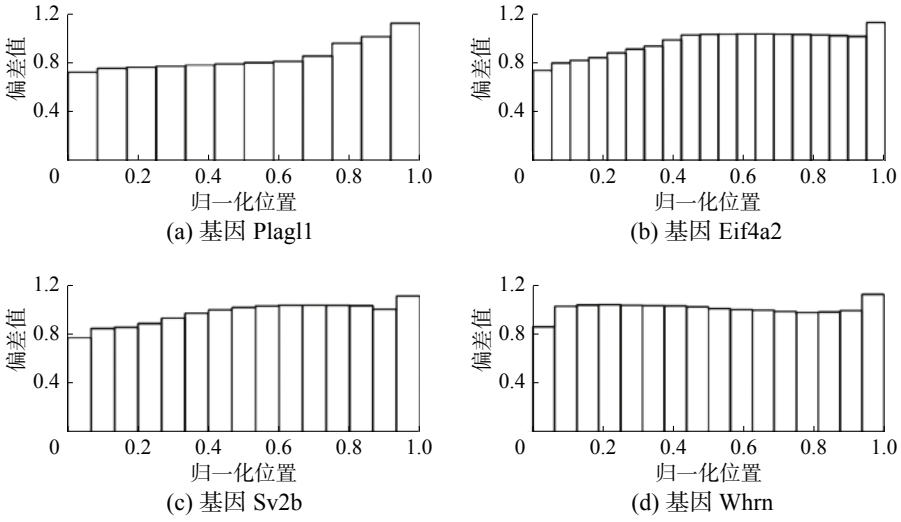

图 6　基因的局部偏差分布

Fig. 6　Local bias distribution of genes

2.3　多条件多样本下剪切异构体表达水平的验证

MCMS-Seq 方法处理多条件多样本数据集时是联合所有样本同时处理, 通过增加稀疏约束, 不仅可以消除数据噪声的影响, 同时也能体现基因和剪切异构体之间存在的稀疏特性。选择小鼠数据集的基因 Nph2 来验证, 该基因包含 3 个剪切异构体。

在小鼠数据集中, 同一个剪切异构体在同一个条件下的多个重复样本中, 其表达水平应该是相近的。若一个剪切异构体在重复样本中零散地出现低表达, 则此剪切异构体的表达水平受到数据噪声的影响。传统方法 Cufflinks 都是每个样本依次单独处理, 其表达水平值如表 1 所示。NM_001364736 表达水平在 Muscle 条件两个重复样本中就可能受到数据噪声的影响, NM_157294 在 Liver 条件下也存在同样的情况。表 2 中 XAEM 方法获得的 NM_001364736 和 NM_157294 表达水平都是极低值, 极大可能是受到数据噪声的干扰。MCMS-Seq 方法联合处理多条件多样本数据集。从表 3 中可以看出, NM_001364736 在 3 个组织条件下都未表达, NM_157294 在大脑和骨骼肌组织条件下具有真实的低表达, 而在肝脏组织条件下未表达, 能有效消除数据噪声的影响。

表 1　Cufflinks 估计基因 Nhp2 中 3 个剪切异构体表达水平

Table 1　Expression level of three isoforms in Nph2 gene estimated using cufflinks

Cufflinks	Brain.1	Brain.2	Liver.1	Liver.2	Muscle.1	Muscle.2
NM_001364736	0.3765	0.6261	0.0000	0.0000	0.0000	0.9089
NM_026631	34.6739	42.0154	23.3440	28.2050	23.9741	25.7665
NM_157294	2.6096	0.7230	0.0006	0.0000	1.0197	0.9718

表 2　XAEM 估计基因 Nhp2 中 3 个剪切异构体表达水平

Table 2　Expression level of three isoforms in Nph2 gene estimated using XAEM

XAEM	Brain.1	Brain.2	Liver.1	Liver.2	Muscle.1	Muscle.2
NM_001364736	0.0003	0.0113	0.0000	0.0000	0.0000	0.0000
NM_026631	417.2680	250.9886	155.9998	132.9998	133.9997	156.9997
NM_157294	26.7317	0.0001	0.0002	0.0003	0.0003	0.0002

表 3　MCMS-Seq 估计基因 Nhp2 中 3 个剪切异构体表达水平

Table 3　Expression level of three isoforms in Nph2 gene estimated using MCMS-Seq

MCMS-Seq	Brain.1	Brain.2	Liver.1	Liver.2	Muscle.1	Muscle.2
NM_001364736	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
NM_026631	40.4611	53.5541	28.0047	35.0090	28.5695	30.9546
NM_157294	3.4900	0.4127	0.0000	0.0000	1.1322	0.2745

此外，基因外在表现通常是由其包含的少数剪切异构体决定的，因此基因和剪切异构体之间存在稀疏特性。在表 4 中，PGSeq 方法得到的 3 个剪切异构体表达水平都存在较高的表达值，无法体现稀疏特性。而 Cufflinks 和 XAEM 受数据噪声影响，同样很难体现出该数据特性。MCMS-Seq 方法增加了 L_2/L_1 组稀疏和 L_1 稀疏约束来考虑上述生物特性。NM_026631 在所有组织条件下中都有较高的表达水平，说明基因 Nph2 的表达主要由 NM_02663 所决定。NM_001364736 在 3 个组织条件下都未表达，特别在肝脏组织条件下 NM_157294 和 NM_001364736 同时未表达，这表明基因 Nph2 在肝脏中只有剪切异构体 NM_026631 参与基因表达。因此，MCMS-Seq 方法能体现基因表达是由少数剪切异构体所决定的生物特性，提供了更好的生物可解释性。

表 4　PGSeq 估计基因 Nhp2 中 3 个剪切异构体表达水平

Table 4　Expression level of three isoforms in Nph2 gene estimated using PGSeq

PGSeq	Brain.1	Brain.2	Liver.1	Liver.2	Muscle.1	Muscle.2
NM_001364736	8.1342	8.3407	4.8802	5.0227	6.0254	5.9947
NM_026631	9.6987	10.0773	9.3418	9.6317	9.3863	9.4907
NM_157294	5.8697	3.5547	−6.3199	−6.0304	4.3553	3.0167

注：PGSeq方法提供的剪切异构体表达水平已经过对数转换

2.4　剪切异构体的可重复性验证

在多条件多样本测序实验中，同一个条件下设计多重复性样本是为了避免技术性误差所带来的影响。这使得同一个剪切异构体在同一个条件下的重复样本之间的表达水平是相近的。小鼠数据集被用来验证剪切异构体表达水平在样本之间的可重复性。采用 Person 相关系数来评估可重复性，其值越高说明能更加有效地消除技术性误差所造成的偏差。由于 RNA-Seq 测序技术得到表达水平其幅度跨度很大，Person 相关系数易受到少数高表达的剪切异构体影响。因此在计算相关系数之前，对所有剪切异构体表达水平进行对数转换，从而避免上述问题。表 5 中显示不同方法在小鼠数据集上不同条件下的相关系数值。从表中可以看出，MCMS-Seq 方法在肝脏、大脑和骨骼肌 3 个条件下都获得了比其他 3 个方法更好的结果。尽管 MCMS-Seq 方法是面向处理多条件多样本数据集，但仍然可以保证剪切异构体在同一个条件中下样本之间具有高度的可重复性。这也符合 RNA-Seq 测序实验中设计重复实验的目的。

表 5　在小鼠数据集上不同方法估计的剪切异构体表达水平在样本之间的相关系数

Table 5　Correlation coefficients between isoform expression levels estimated using various methods in the mouse dataset

方法	Cufflinks	PGSeq	XAEM	MCMS-Seq
Liver	0.950 2	0.921 5	0.962 0	0.982 4
Brain	0.947 5	0.915 3	0.966 0	0.985 4
Muscle	0.944 6	0.908 0	0.952 8	0.957 4
平均值	0.947 4	0.914 9	0.960 3	0.975 1

2.5　PCR 剪切异构体的表达水平验证

SEQC 数据集被用来验证不同方法估计剪切异构体表达水平的准确性。该数据集提供了 16 603 个经过 qRT-PCR 验证的剪切异构体, 这些剪切异构体被当作基准数据。计算不同方法得到剪切异构体表达水平与 qRT-PCR 值之间的相关系数。从表 6 中结果可以看出, MCMS-Seq 方法在 UHRR 条件上稍微优于 PGSeq 方法, 而在 HBRR 条件上获得较为明显的提升。尽管 XAEM 方法是多样本处理, 但获得最差的性能, 其可能是该方法对数据偏差考虑得不够。整体上说, MCMS-Seq 方法估计的剪切异构体表达水平能取得较为准确的结果。

表 6　在 SEQC 数据集上不同方法与 qRT-PCR 验证剪切异构体之间的相关系数

Table 6　Correlation coefficients between qRT-PCR values and isoform expression levels estimated using various methods in SEQC dataset

方法	Cufflinks	PGSeq	XAEM	MCMS-Seq
HBRR	0.733 2	0.754 0	0.728 7	0.773 5
UHRR	0.797 5	0.806 1	0.794 8	0.813 6
Mean	0.765 4	0.780 1	0.761 8	0.793 6

2.6　PCR 基因的表达水平验证

现实中包含 qRT-PCR 验证的剪切异构体数据集很少, 而基因的表达水平是由其所包含的剪切异构体所决定的, 因此可以通过验证 qRT-PCR 验证基因的表达水平来间接验证剪切异构体表达水平的准确性。MAQC-II 数据集被广泛地应用于评估不同方法估计基因表达水平的性能。MAQC-II 数据集提供了 838 个 qRT-PCR 验证的基因, 这些基因总共包含了 6 927 个剪切异构体。Cufflinks 和 PGSeq 方法提供了基因的表达水平, XAEM 和 MCMS-Seq 方法的基因表达水平由所对应的剪切异构体表达水平求和得到。表 7 显示了不同方法得到的基因表达水平与 qRT-PCR 值之间的相关系数。从表 7 中可以看出, 相比其他方法, MCMS-Seq 方法得到了更好的准确性。

表 7　在 MAQC-II 数据集上不同方法与 qRT-PCR 验证基因之间的相关系数

Table 7　Correlation coefficients between qRT-PCR values and isoform expression levels estimated using various methods in the MAQC-II dataset

方法	Cufflinks	PGSeq	XAEM	MCMS-Seq
HBRR	0.845 4	0.829 6	0.744 3	0.860 3
UHRR	0.848 3	0.830 8	0.808 1	0.868 3
Mean	0.846 9	0.830 2	0.776 2	0.864 3

2.7　稀疏参数的选择

MCMS-Seq 方法包含了 L_2/L_1 组稀疏约束和 L_1 稀疏约束两个正则化项, 不仅用来考虑基因和剪切异构体之间的稀疏特性, 同时用来消除虚假低表达剪切异构体带来的影响。在式 (6) 中, 参数 λ_1 和 λ_2 分别对应着 L_2/L_1 组稀疏约束和 L_1 稀疏约束, 其值的选择能影响到剪切异构体表达水平的准确性。当 λ_1 或 $\lambda_2 \to +\infty$ 时, 都会导致剪切异构体出现不表达情况, 区别在于, $\lambda_1 \to +\infty$ 会导致同一个剪切异构体在不同条件下都没有表达。当 $\lambda_1 \to 0$ 时, 剪切异构体的表达水平容易受到数据噪声的影响, 产生虚假的低表达。而 λ_2 减小时, 基因与剪切异构体之间的稀疏特性将减弱。选择 SEQC 数据集中 HBRR 条件来分析参数选择对剪切异构体表达水平准确性的影响。假设参数 λ_1 和 λ_2 分别选择 0.1、1、10 和 100 这 4 个值, 图 7 显示了在取不同参数值时, MCMS-Seq 方法估计的剪切异构体表达水平与 qRT-PCR 验证的剪切异构体之间的相关系数。从图 7 可以看出, 当 λ_1 和 λ_2 同时增大时, 其相关系数都显著下降, 因为有大量真正表达的剪切异构体被估计成未表达。而 λ_1 和 λ_2 在取值 1 附近能获得较为稳定的结果, 因此本文中所有实验都是设定 λ_1 和 λ_2 为 1。

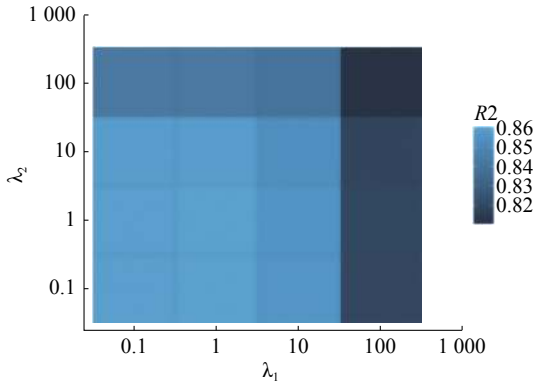

图 7　参数 λ_1 和 λ_2 对剪切异构体表达水平的影响

Fig. 7　Effect of isoform expression levels by parameters λ_1 and λ_2

3　结束语

本文提出了一个基于多条件多样本 RNA-Seq 测序数据的剪切异构体表达水平估计方法。为了考虑基因读段分布在不同条件下的高度相似性，MCMS-Seq 方法设计一个联合多条件多样本的偏差估计模型，同时考虑了基因读段分布的全局偏差和局部偏差所带来的影响。从数据分析可以看出，该偏差估计模型能较为准确地描述出基因读段非均匀分布特性。此外，MCMS-Seq 方法增加了 L_2/L_1 组稀疏约束和 L_1 稀疏约束两个正则化项，体现了基因和剪切异构体之间存在稀疏的生物特性，同时消除了技术性误差和数据噪声的影响。在小鼠数据集上，MCMS-Seq 方法估计的剪切异构体表达水平能获得更好的可重复性。通过与 SEQC 数据集中 qRT-PCR 剪切异构体和 MAQC-II 数据中 qRT-PCR 基因的验证，MCMS-Seq 方法比其他 3 个对比方法更佳的性能。

由于大量多条件多样本数据集是时序数据集，蕴含了时间信息，但是 MCMS-Seq 模型未考虑到数据中的时间信息。在未来的研究中，可以考虑在模型中融入时间信息，从而进一步提高剪切异构体的表达水平的准确性。此外，可将 MCMS-Seq 模型推广到单细胞测序数据分析，可提供更好的生物解释性。

参考文献：

[1] MARIONI J C, MASON C E, MANE S M, et al. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays[J]. Genome research, 2008, 18(9): 1509–1517.

[2] 周晓光, 任鲁风, 李运涛, 等. 下一代测序技术: 技术回顾与展望 [J]. 中国科学: 生命科学, 2010, 40(1): 23–37.
ZHOU Xiaoguang, REN Lufeng, LI Yuntao, et al. The next-generation sequencing technology: A technology review and future perspective[J]. Scientia sinica (vitae), 2010, 40(1): 23–37.

[3] 王曦, 汪小我, 王立坤, 等. 新一代高通量 RNA 测序数据的处理与分析 [J]. 生物化学与生物物理进展, 2010, 37(8): 834–846.
WANG Xi, WANG Xiaowo, WANG Likun, et al. A review on the processing and analysis of next-generation RNA-seq data[J]. Progress in biochemistry and biophysics, 2010, 37(8): 834–846.

[4] ZHANG Li, LIU Xuejun. A comprehensive review on RNA-Seq data analysis[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2016, 33(3): 339–361.

[5] MONIER B, MCDERMAID A, WANG Cankun, et al. RIS-EDA: an integrated RNA-Seq interpretation system for gene expression data analysis[J]. PLoS computational biology, 2019, 15(2): e1006792.

[6] 王凯莉, 张礼, 刘学军. 融合多平台表达数据的转录组差异表达分析 [J]. 计算机学报, 2018, 41(6): 1415–1430.
WANG Kaili, ZHANG Li, LIU Xuejun. Differential expression analysis based on integrating transcriptome expression data from multiple platforms[J]. Chinese journal of computers, 2018, 41(6): 1415–1430.

[7] 王凯莉, 张礼, 刘学军. 多实验平台下基因及异构体表达分析综述 [J]. 中国生物医学工程学报, 2017, 36(2): 211–218.
WANG Kaili, ZHANG Li, LIU Xuejun. A review of gene and isoform expression analysis across multiple experimental platforms[J]. Chinese journal of biomedical engineering, 2017, 36(2): 211–218.

[8] JIANG Hui, WONG W H. Statistical inferences for isoform expression in RNA-Seq[J]. Bioinformatics, 2009, 25(8): 1026–1032.

[9] WU Zhengpeng, WANG Xi, ZHANG Xuegong. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq[J]. Bioinformatics, 2011, 27(4): 502–508.

[10] HU Ming, ZHU Yu, TAYLOR J M G, et al. Using poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq[J]. Bioinformatics, 2012, 28(1): 63–68.

[11] TRAPNELL C, WILLIAMS B A, PERTEA G, et. al Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. Nature biotechnology, 2010, 28(5): 511–515.

[12] GLAUS P, HONKELA A, RATTRAY M. Identifying differentially expressed transcripts from RNA-Seq data with biological variation[J]. Bioinformatics, 2012, 28(13): 1721–1728.

[13] ZHANG Li, LIU Xuejun. PBSeq: modeling base-level bias to estimate gene and isoform expression for RNA-Seq data[J]. International journal of machine learning and cybernetics, 2017, 8(4): 1247–1258.

[14] LI Bo, DEWEY C N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome[J]. BMC bioinformatics, 2011, 12: 323.

[15] LI W V, LI J J. Modeling and analysis of RNA-seq data: a review from a statistical perspective[J]. Quantitative biology, 2018, 6(3): 195–209.

[16] LIU Siyun, JIANG Yuan, TAO Yu. Modelling RNA -

Seq data with a zero-inflated mixture Poisson linear model[J]. Genetic epidemiology, 2019, 43(7): 786–799.

[17] ZHANG Chi, ZHANG Baohong, LIN L L, et al. Evaluation and comparison of computational tools for RNA-Seq isoform quantification[J]. BMC genomics, 2017, 18(1): 1–11.

[18] LI Song, SABUNCIYAN S, YANG Guangyu, et al. A multi-sample approach increases the accuracy of transcript assembly[J]. Nature communications, 2019, 10(1): 1–7.

[19] SUO Chen, CALZA S, SALIM A, et al. Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data[J]. Bioinformatics, 2014, 30(4): 506–513.

[20] LI W V, ZHAO Anqi, ZHANG Shihua, et al. MSIQ: joint modeling of multiple RNA-Seq samples for accurate isoform quantification[J]. The annals of applied statistics, 2018, 12(1): 510–539.

[21] DENG Wenjiang, MOU Tian, KALARI K R, et al. Alternating EM algorithm for a bilinear model in isoform quantification from RNA-Seq data[J]. Bioinformatics, 2020, 36(3): 805–812.

[22] AGUIAR D, CHENG Lifang, DUMITRASCU B, et al. Bayesian nonparametric discovery of isoforms and individual specific quantification[J]. Nature communications, 2018, 9(1): 1–12.

[23] LIU Xuejun, ZHANG Li, CHEN Songcan. Modeling exon-specific bias distribution improves the analysis of RNA-Seq data[J]. PLoS one, 2015, 10(10): e0140032.

[24] 焦李成, 赵进, 杨淑媛, 等. 稀疏认知学习、计算与识别的研究进展 [J]. 计算机学报, 2016, 39(4): 835–852.
JIAO Licheng, ZHAO Jin, YANG Shuyuan, et al. Research advances on sparse cognitive learning, computing and recognition[J]. Chinese journal of computers, 2016, 39(4): 835–852.

[25] JENATTON R, MAIRAL J, OBOZINSKI G, et al. Proximal methods for sparse hierarchical dictionary learning[C]//Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, Israel, 2010: 487−494.

[26] LANGMEAD B, SALZBERG S L. Fast gapped-read alignment with Bowtie 2[J]. Nature methods, 2012, 9(4): 357–359.

[27] MORTAZAVI A, WILLIAMS B A, MCCUE K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq[J]. Nature methods, 2008, 5(7): 621–628.

[28] 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.

[29] BULLARD J H, PURDOM E, HANSEN K D, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments[J]. BMC bioinformatics, 2010, 11(1): 1–13.

作者简介:

张礼, 讲师, 博士, 主要研究方向为机器学习、生物信息学。

马越, 助教, 硕士, 主要研究方向为分子生物学、神经系统疾病。

吴东洋, 讲师, 博士, 主要研究方向为数据挖掘、生物信息学。