ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 5,Sep. 2015,pp. 1028-1035
DOI:10.16337/j.1004-9037.2015.05.013

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

改进的 RNA-Seq 数据转录组表达分析研究

石新新　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: 基于高通量测序的 RNA-Seq(RNA-sequencing)是用于转录组研究的一种新技术,针对该技术在转录组表达分析研究中存在的读段多源映射和读段非均匀分布等难点,提出一个改进的转录组表达研究方法 LDASeqII(Improvement of latent Dirichlet allocation for sequencing data)。模型利用剪接异构体结构信息对参数进行约束并进行外显子读段数目归一化处理,解决了读段非均匀分布下的多源映射问题。通过引入"伪外显子"和"伪转录本"分别处理接合区读段和噪声读段。将模型应用到真实数据集上,并与原 LDASeq(Latent Dirichlet allocation for sequencing data)模型和目前流行的 Cufflinks 与 RSEM(RNA-Seq by expectation maximization)方法进行对比。结果显示,改进方法获得了更为准确的转录本及基因表达水平计算结果。

关键词: 基因表达;RNA-Seq;转录组表达;多源映射;非均匀性

中图分类号: TP391.9　　**文献标志码:** A

Improved Trancriptome Expression Analysis for RNA-Seq Data

Shi Xinxin, Liu Xuejun, Zhang Li

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

Abstract: RNA-Seq(RNA-sequencing), based on high-throughput sequencing, is a new technique for transcriptome research. Considering the difficulties in the analysis of transcript expression using RNA-Seq data, an improved method, improvement of latent dirichlet allocation for sequencing data(LDASeqⅡ) is proposed to calculate the transcript expression. To deal with multi-mappings between reads and isoforms and non-uniform distribution of reads along reference, LDASeqⅡ utilizes the known gene-isoform annotation to constrain the hyperparameters and normalizes the read counts by exon length for each individual exon. By introducing "pseudo-exon" and "pseudo-transcript", the conjunction reads and noise reads gain proper treatments. LDASeqⅡ is validated using two real datasets on gene and transcript expression calculation and compared with latent dirichlet allocation for sequencing data(LDASeq) and other two popular methods Cufflinks and RNA-Seq by expectation maximization(RSEM). The results show that LDASeqⅡ obtains more accurate transcript and gene expression measurements than other approaches.

Keywords: gene expression; RNA-Seq; transcript expression; multi-mapping; non-uniformity

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:2014-05-30;**修订日期:**2014-06-25

引 言

RNA-Seq(RNA sequencing)是基于高通量测序技术对转录组进行研究的一种新方法,具有信噪比高、分辨率高和所需样本少等优势,可以用来检测和量化任何物种的转录片段[1-2],近年来在转录组研究中得到大量应用。RNA-Seq 数据处理过程主要分为 3 个方面:(1) 将读段映射到参考基因组或转录组上;(2) 利用匹配到的读段对基因或转录组进行构建;(3) 通过计数匹配到基因或转录本上的读段数目来计算基因或转录本表达值[3]。本文主要针对最后一个步骤进行研究,即给定转录组构建、计算基因以及异构体的表达水平。

RNA-Seq 通过计数匹配到基因元件上读段数目来表示相关基因的转录水平[4],为了保持对不同基因和不同实验间估计的基因表达值的可比性,必须对基因长度和测序深度进行归一化。最常用的归一化方法是 Mortazavi 等人提出 RPKM(Reads per kilobases per million reads)方法,即每百万读段中来自该基因的读段数[5]。基因在转录过程中会形成一个或多个转录本,这些转录本共享基因的部分外显子[6],造成读段到异构体的多源映射,故只有一少部分比例的读段能够唯一确定其来源转录本[7]。因此,RPKM 方法不能直接应用于转录本表达值计算。一种解决思路为可以根据已知外显子组成和各外显子长度对转录本建立数学模型,利用基因外显子上读段数求解转录本表达值。文献[8]提出的转录本表达值计算方法和文献[9]设计的 Cufflinks 模型都采用了这种思路来实现剪接异构体的表达推断。此外,还有文献[10-11]提出的 RSEM(RNA-Seq by expectation maximization)方法,文献[12]提出的 BitSeq(Bayesian inference of transcripts from sequencing data)方法,都采取了一定手段来解决读段的多源映射问题。RNA-Seq 数据的另外一个特点是读段在参考序列上的分布有着自身的一些模式,呈现出不均匀性[13-16]。Cufflinks 方法把位置偏差和序列偏差融入到计算中,模拟了读段非均匀采样的随机性质[9]。RSEM 方法通过一个产生式概率模型模拟读段产生过程的经验分布等信息,在一定程度上消除了均匀分布的假设[10]。

针对 RNA-Seq 读段的异构体多源映射与分布不均匀两个难点,文献[17]提出了 LDASeq(Latent Dirichilet allocation for sequencing)模型来计算异构体表达水平。LDASeq 模型通过引入隐含变量,较好模拟了读段的非均匀分布和多源映射现象。后续工作中发现,该模型存在以下缺点:首先忽视了两类重要读段,即接合区读段和噪声读段对结果的重要影响。同时,对小外显子没有采取读段数目归一化且由于外显子分割成探针而带来外显子尾部不足一个探针部分信息丢失,并且没有考虑由于噪声和未知转录本的存在对已知转录本表达值的推断产生影响。针对 LDASeq 模型存在的这些问题,本文对 LDASeq 方法进行改进。改进的方法舍弃"伪探针"直接对外显子进行长度归一化,减小了探针引入所带来的信息丢失,并且增加了对接合区读段和噪声读段的处理。本文通过采用两个真实数据集对所提出的改进方法在基因和异构体表达水平计算精度上进行了验证。

1 DASeq Ⅱ 方法设计

1.1 LDASeq 模型

基于 RNA-Seq 数据和文本数据在结构上的相似性,文献[17]提出了 LDASeq 模型。LDASeq 是基于 LDA(Latent Dirichlet allocation)模型[18]的一个 3 层贝叶斯网络模型,如图 1 所示。LDASeq 模型引入固定长度的探针,将探针和单词概念对应,探针上读段的匹配个数看作单词在文档中出现的频数。单个通道下一个基因的 N 个探针上的读段对应一篇文档,若干个异构体看作文档的隐含主题,α 是一个超参数,模型中引入隐含变量 $\boldsymbol{\theta}$,$\boldsymbol{\theta}$ 是一个向量,符合参数为 α 的狄利克雷分布,$\boldsymbol{\theta}$ 的各个分量反映了异构体表达的相对强弱。模型中 $\boldsymbol{\beta}$ 是 $K\times V$ 矩阵,其中 K 表示该基因形成的剪切异构体个数,V 表示该基因含有的探针个数,第 $i(1\leqslant i\leqslant K)$ 行第 $j(1\leqslant j\leqslant V)$ 个分量表示第 i 个剪切异构体是否含有第 j 个探针,因此

每个基因对应一个 $\boldsymbol{\beta}$ 矩阵。将 $\boldsymbol{\beta}$ 矩阵进行归一化作为模型的输入，文献[17]中描述了 $\boldsymbol{\beta}$ 矩阵的产生过程。

LDASeq 模型对每个基因模拟 RNA-Seq 数据的产生过程如下：(1) 根据狄利克雷分布为每个通道产生 $\theta_l \sim \text{Dirichlet}(\alpha)$；(2) 对于通道 l，根据多项分布产生异构体 isoform_n。$\text{isoform}_n \sim \text{Multinomial}(\theta_l)$；(3) 在异构体 isoform_n 条件下，根据多项分布产生探针 probe_n，$\text{probe}_n \sim \text{Multinomial}(\boldsymbol{\beta})$。

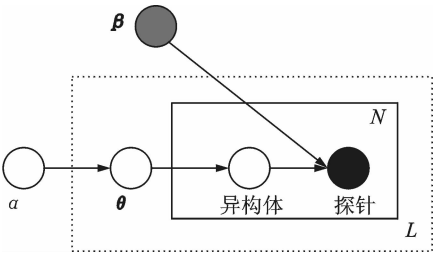

图 1 LDASeq 图模型表示

Fig. 1 Graphic model representation of LDASeq

文献[17]对该模型使用变分最大期望算法(Expectation maximization，EM)来计算。模型求解得到参数 $\boldsymbol{\theta}(\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha))$，每个分量代表着一个异构体表达比重，然后将基因每个探针上映射的读段计数按这个比重分配给对应的异构体。采用文献[5]中的 RPKM 异构体表达式计算每个异构体表达值，将一个基因所对应的所有异构体表达值求和即可得到该基因表达值[9]。

1.2 LDASeq Ⅱ 模型

针对 LDASeq 模型存在的一些缺点，本文对其进行了改进，将改进后的方法记为 LDASeq Ⅱ (Improment of latent Dirichilet allocation for sequencing data)，改进后模型如图 2 所示。与原有模型相比：(1) 舍弃原模型探针概念，将外显子与单词对应。为了减小不同外显子长度对读段数目的偏好，改进方法中将读段数目按照外显子长度归一化，即单位外显子长度上读段个数作为单词出现次数。(2) 为了处理由未知转录本和测序错误等产生的噪声读段和带有重要剪切信息的接合区读段，对每个基因所属异构体结构进行扩展，增加伪异构体和伪外显子。

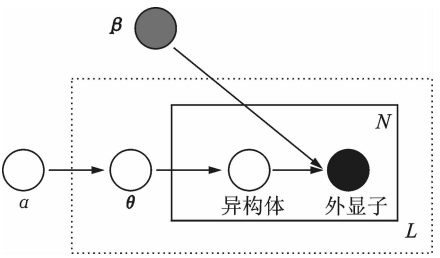

图 2 LDASeq Ⅱ 图模型表示

Fig. 2 Graphic model representation of LDASeq Ⅱ

1.2.1 伪外显子的设计

图 3 以基因 G 为例说明伪外显子构造过程。首先，扩充该基因的外显子集合。对于基因 G，原有外显子集合记为 $F=\{e_1, e_2, e_3, e_4\}$，由图 3 结构关系可知，该基因发生了选择性剪接形成两个转录本。

由转录本 1 可知,基因外显子 e_1 与外显子 e_3 之间发生选择性剪接,因此在 e_1 与 e_3 之间构造一个接合区记为 e_1-e_3,作为伪外显子加入集合 F,这样基因 G 外显子集合 $F=\{e_1, e_2, e_3, e_4, e_1\text{-}e_3\}$。由转录本 2 可知,基因在外显子 e_1 与外显子 e_2 之间、外显子 e_2 与外显子 e_4 之间发生选择性剪接,因此分别在 e_1 与 e_2 之间、e_2 与 e_4 之间构造接合区 e_1-e_2, e_2-e_4,集合 F 中不含有元素 e_1-e_2 和 e_2-e_4,因此将这两个元素作为伪外显子加入集合 F,这样集合 F 变为$\{e_1, e_2, e_3, e_4, e_1\text{-}e_3, e_1\text{-}e_2, e_2\text{-}e_4\}$。

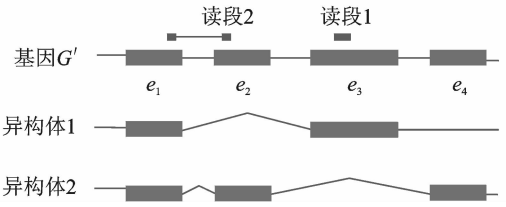

图 3 外显子和读段之间关系

Fig. 3 Relationship between exons and reads

接下来构造新的转录本与外显子映射关系。扩增后的基因外显子集合 $F=\{e_1, e_2, e_3, e_4, e_1\text{-}e_3, e_1\text{-}e_2, e_2\text{-}e_4\}$。根据注释信息可知,转录本 1 包含外显子子集记为 $f_1=\{e_1, e_3, e_1\text{-}e_3\}$,转录本 2 包含外显子子集记为 $f_2=\{e_1, e_2, e_4, e_1\text{-}e_2, e_2\text{-}e_4\}$。这样,改造后的基因 G 和转录本参考结构如图 4 所示。当统计外显子读段时,比如读段 1,完全落在外显子 e_3 上,那么将 e_3 上读段数目加 1;而对于读段 2,落在外显子 e_1 与外显子 e_2 接合处,因此,将伪外显子 e_1-e_2 上读段数加 1,其他读段作类似统计。

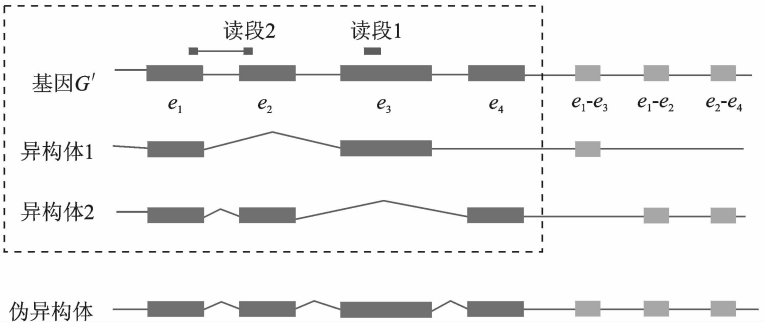

图 4 增加伪外显子、伪异构体后的基因结构图

Fig. 4 Gene structure of adding "pseudo-exon" and "pseudo-transcript"

1.2.2 伪转录本的引入

由于现有基因注释信息的不全面,没有包含那些尚未发现但已经存在的的转录本,另外,RNA-Seq 数据中还存在由于测序错误等产生的其他噪声读段。考虑到这两类噪声读段对于已知转录本表达水平的影响,试图在建模中加以矫正。通过改变每个基因所对应的 $\boldsymbol{\beta}$ 的初始值,即对于每个基因,尝试增加一条特殊的转录本参与优化,将那些噪声读段看作这条特殊转录本所产生。对于图 3 中的基因 G,模型中尝试引入了一条伪异构体带表这条基因尚未发现的剪切异构体,图 4 中即为基因 G 引入了伪异构体。这条异构体含有基因外显子集合 F 中所有的外显子。文章[17]描述了 $\boldsymbol{\beta}$ 矩阵产生规则。在对基因结构进行扩增即增加了伪外显子和伪转录本后,$\boldsymbol{\beta}$ 产生规则不变。图 4 所示扩增结构后基因对应的 $\boldsymbol{\beta}$ 矩阵初始值的产生过程

$$\begin{vmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{vmatrix} \xrightarrow{\text{归一化}} \begin{vmatrix} 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/5 & 1/5 & 0 & 1/5 & 0 & 1/5 & 1/5 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{vmatrix} \tag{1}$$

1.2.3 归一化的改进

对于图 4 中扩增结构后的基因 G'，该基因含有的 7 个外显子对应 7 个单词。为了减小由于外显子长度不同导致外显子上读段数目的偏好，将每个外显子上读段数目除以该外显子长度，取单位长度上的读段个数作为单词出现频数。对于基因 G' 原有的真实外显子，即 $\{e_1, e_2, e_3, e_4\}$，归一化长度取真实长度-读段长度，对于个别读段小于读段长度的小外显子，将上面的读段统计到相邻的接合区，无需归一化；对于新构造的伪外显子 $\{e_1\text{-}e_3, e_1\text{-}e_2, e_2\text{-}e_4\}$，归一化长度取读段长度。对于双末端读段数据，由于读段片段长度不定，实验中取其平均长度。

2 实验过程及结果

2.1 数据处理流程

由于增加了对接合区读段、噪声读段的处理，改进后方法的的数据处理过程与 LDASeq 方法[17]有所不同，整个处理过程如图 5 所示。

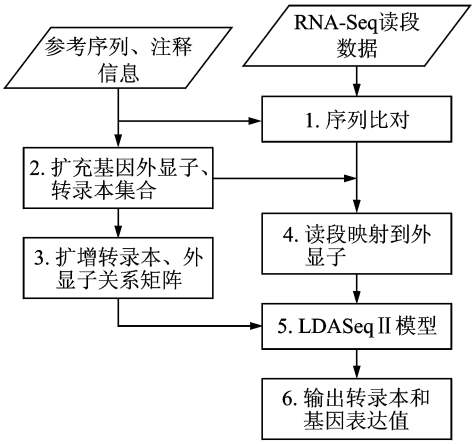

图 5 LDASeqⅡ处理流程

Fig. 5 Processing of LDASeqⅡ

处理过程为：第 1 步，用序列比对软件将读段比对到转录组上；第 2 步，根据注释信息，扩充基因外显子集合、转录本集合；第 3 步，根据扩充后的基因结构，写出基因对应的 $\boldsymbol{\beta}$ 矩阵并归一化；第 4 步，统计基因各个外显子以及伪外显子上面的读段数；第 5 步，将第 3 步和第 4 步的数据作为模型输入，计算转录本和基因表达值。

2.2 实验数据集

本文使用两个真实数据集对 LDASeqII 的计算性能进行验证，分别为人类大脑双末端数据集[19]和人类乳腺癌数据集(Human breast cancer, HBC)[20]。人类大脑数据集来自美国食品药品管理局联合全球多家研究机构举办的基因芯片质量控制(Microarray quality control, MAQC)项目[19]。该数据集包含 1 000 个经过定量反转录聚合酶链式反应(Quantificational real-time polymerase chain reaction, qRT-PCR)实验验证的基因，这些结果可以作为基因表达水平计算结果的判别标准。人类乳腺癌数据

集[20]包含两个实验条件,即乳腺癌细胞(Human breast cancer cell line, MCF-7)和正常乳腺细胞(Normal cell line, HME)。文献[21]也采用了该数据集,并提供了经过验证的 4 个多异构体基因共 8 个转录本在不同条件下的表达调控值。这些结果可以作为验证转录本表达计算性能的一个标准。

2.3 基因水平上的验证结果

在人类大脑双末端数据集上,LDASeqⅡ共计算出 740 个多异构体基因的表达值,并求出了表达值与 qRT-PCR 结果的相关系数,将结果与改进前的 LDASeq, Cufflinks 和 RSEM 这 3 种方法求出的相关系数作比较,各种方法的相关性如表 1 所示。从表中可以看出,LDASeqⅡ相比其他 3 种方法获得了最高相关系数,基因表达水平获得了较准确的计算结果。

表 1 不同方法在 MAQC 数据集上的处理结果对比

Table 1 Results of all methods using MAQC dataset

方法	与 qRT-PCR 结果相关系数
LDASeqⅡ	0.832 1
Cufflinks	0.810 8
RSEM	0.829 1
LDASeq	0.827 8

2.4 转录本水平上的验证结果

在人类乳腺癌数据集上,LDAseqⅡ, Cufflinks, RSEM 和 LDASeq 这 4 种方法分别计算,经过 qRT-PCR 实验验证的 8 个转录本在 HME 和 MCF-7 两个条件下的表达水平和表达值变化方向调控值,将结果与 qRT-PCR 结果进行对比,如表 2 所示。

表 2 人类乳腺癌数据集各方法计算结果

Table 2 Results of all methods using HBC dataset

比较分组			qRT-PCR	Cufflinks	RSEM	LDASeq	LDASeqⅡ
TRAP1	uc002cvt.2	HME vs. MCF-7	−(0.4)	−(0.6)	−(0.4)	−(0.9)	−0.9
TRAP1	uc002cvs.1	HME vs. MCF-7	−(0.5)	−(1.1)	−(0.8)	−(0.6)	−0.6
TRAP1	HME	uc002cvt.2 vs. uc002cvs.1	−(0.9)	+(4.8)	+(4.9)	−(0.8)	−0.3
TRAP1	MCF-7	uc002cvt.2 vs. uc002cvs.1	−(1.0)	+(4.3)	+(4.4)	−(0.6)	−0.04
ZNF581/0	uc002qlq.1	HME vs. MCF-7	−(0.3)	−(1.2)	−(0.9)	−(1.4)	−2.4
ZNF581/0	uc002qlp.1	HME vs. MCF-7	−(1.0)	−(1.2)	−(0.7)	−(1.4)	−1.4
ZNF581/0	HME	uc002qlq.1 vs. uc002qlp.1	+(1.2)	+(1.9)	+(1.3)	+(0.7)	+0.37
ZNF581/0	MCF-7	uc002qlq.1 vs. uc002qlp.1	+(1.0)	+(1.9)	+(1.5)	+(0.8)	+1.30
WISP2	uc002xmn.1	HME vs. MCF-7	−(5.6)	−(6.9)	−(5.4)	−(8.3)	−7.80
WISP2	uc002xmo.1	HME vs. MCF-7	−(4.5)	−(5.4)	−(4.7)	−(6.0)	−5.60
WISP2	HME	uc002xmn.1 vs. uc002xmo.1	+(0.4)	(0.0)	(0.0)	−(0.4)	+0.26
WISP2	MCF-7	uc002xmn.1 vs. uc002xmo.1	+(1.5)	+(1.5)	+(0.8)	+(1.9)	+2.40
HIST1H2BD	uc003ngr.1	HME vs. MCF-7	−(4.7)	−(3.7)	−(2.9)	−(2.4)	−1.10
HIST1H2BD	uc003ngs.1	HME vs. MCF-7	−(5.2)	−(4.2)	−(4.5)	−(3.9)	−4.30
HIST1H2BD	HME	uc003ngr.1 vs. uc003ngs.1	−(5.4)	+(2.4)	+(1.8)	−(2.4)	−0.20
HIST1H2BD	MCF-7	uc003ngr.1 vs. uc003ngs.1	−(5.9)	+(1.8)	+(0.2)	−(3.9)	−3.30
调控方向错误数				4	5	1	0
与 qRT-PCR 结果的相关系数				0.489 9	0.527 9	0.859 6	0.733 8

表 2 显示了同一转录本在不同实验条件、同一基因的不同转录本在同一实验条件下表达值的变化方向，共 16 种表达调控关系。“+”表示在相应的比较条件中转录本表达上调，“−”表示下调，括号中的数字表示调控的对数倍数。表 2 的最后两行显示了各种计算方法得到的调控方向与 qRT-PCR 结果不一致的比较次数，以及 16 个对数倍数与 qRT-PCR 结果的相关系数。调控方向反映模型在定性计算上与 qRT-PCR 结果的一致性，从表中可以看出，只有 LDASeqⅡ计算出的 16 个调控方向和 qRT-PCR 计算出的方向完全一致，而 Cufflinks，RSEM 和 LDASeq 方法的调控方向错误个数分别为 4，5 和 1。LDASeqⅡ方法在保持调控方向与 qRT-PCR 结果一致的情况下，相关系数达到 0.733 8。由此可以看出，改进后的方法即 LDAseqⅡ，在人类乳腺癌数据集上获得了相比其他方法更为准确的转录本表达水平计算结果。

3 结束语

本文针对已有的 LDASeq 模型存在的一些缺点进行改进，提出了 LDASeqⅡ方法来计算基因及转录本表达值。模型打破了读段在参考序列上均匀分布这一常用假设，通过统计基因外显子和接合区上读段数目，利用转录本的结构信息构造初始化 $\boldsymbol{\beta}$ 矩阵对模型参数进行约束，优化不同转录本下各个外显子表达强弱，最终估计每个转录本的表达比重，较好地解决了转录本表达的计算问题。LDASeqⅡ方法在 LDASeq 的基础上，改进了读段数目归一化方法，增加了对接合区读段、噪声读段的处理，实验表明，改进后的模型获得了较优的计算性能。相比其他大多数计算转录本表达值方法的创新之处在于模型中通过尝试引入未知异构体，减小了噪声读段对已知异构体表达结果的影响，可以作为提高转录本表达研究性能的一个借鉴。本文在构造特殊外显子时，假设它含有基因的所有外显子，在后续工作中，可以尝试随机分配外显子或者根据基因已知的选择性剪接信息，让模型自学习未知的转录本结构。另外，接合区读段包含选择性剪切的重要信息，通过对该区域读段的处理，可以提高异构体表达水平的计算准确性，本文方法对接合区读段单独处理，充分保留了该区域读段带有的选择性剪接信息。对于接合区长度，实验中采用一个读段长度，也可以根据接合区两端的外显子长度按比例设定接合区的长度，在后续工作中会进一步验证该方法的实际效果。

参考文献：

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics [J]. Nature Reviews Genetics, 2009, 10 (1): 57-63.

[2] Denoeud F, Aury J M, Da Silva C, et al. Annotating genomes with massive scale RNA sequencing[J]. Genome Biol, 2008, 9 (12):R175.

[3] Garber M, Grabherr M G, Guttman M, et al. Computational methods for transcriptome annotation and quantification using RNA-Seq[J]. Nature Methods, 2011, 8(6): 469-477.

[4] Marguerat S, Bahler J. RNA-seq: From technology to biology[J]. Cell Mol Life Sci, 2010, 67: 569-579.

[5] 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.

[6] Pan Q, Shai O, Lee L J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing [J]. Nature Genetices, 2008, 40(12): 1413-1415.

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

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

[9] Trapnell C, Williams B A, Pertea G, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. Nat Biotechnol, 2010(5): 511-515.

[10] Li B, Ruotti V, Stewart R M, et al. RNA-Seq gene expression estimation with read mapping uncertainty [J]. Bioinformatics, 2010, 26(4): 493-500.

[11] Li B, Dewey C N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome[J]. BMC Bioinformatics, 2011, 12: 323.

[12] Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-Seq data with biological variation [J]. Bioinformatics, 2012, 28(3): 1721-1728.

[13] Pepke S, Wold B, Mortazavi A. Computation for ChIP-Seq and RNA-Seq studies[J]. Nature Methods Supplement, 2009, 6: S22-S32.

[14] Dohm J C, Lottaz C, Borodina T, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing[J]. Nucleic Acids Res, 2008(16): e105.

[15] Li J, Jiang H, Wong W H. Modeling non-uniformity inshort-read rates in RNA-Seq data[J]. Genome Biol, 2010(5): R50.

[16] Hansen K D, Brenner S E, Dudoit S. Biases in illumina transcriptome sequencing caused by random hexamer priming[J]. Nucleic Acids Research, 2010, 38(12): e131.

[17] 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,2013,7(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.

[18] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[19] Shi L, Reid L H, Jones W D, et al. The microarry quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements[J]. Nat Biotechnol, 2006, 24(9): 1151-1161.

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

[21] Kim H, Bi Y T, Pal S, et al. IsoformEx: Isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data[J]. BMC Bioinformatics, 2011, 12: 305.

作者简介:

石新新(1989-),女,硕士研究生,研究方向:生物信息学,E-mail: shixinxin61@126.com。

刘学军(1976-),女,博士,教授,研究方向:机器学习与生物信息学。

张礼(1985-),男,博士,研究方向:生物信息学。