计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600057

基于 RNA-Seq 的转录组分析方法

郭茂祖[1,2]　杨　帅[1,2]　赵玲玲[3]
1 北京建筑大学电气与信息工程学院　北京 100044
2 建筑大数据智能处理方法研究北京市重点实验室　北京 100044
3 哈尔滨工业大学计算机科学与技术学院　哈尔滨 150001
(guomaozu@bucea.edu.cn)

摘　要　RNA-Seq 技术凭借测序成本低、精度高、覆盖范围广等优点,已经成为了转录组分析的重要方法,为研究基因表达模式、疾病的生物标志物探测、作物抗逆性研究和分子育种等提供了新的手段。然而,RNA-Seq 产生的海量数据也给数据分析带来了挑战,如何有效地对 RNA-Seq 数据进行处理和分析成为了生物信息学研究的热点。文中对基于 RNA-Seq 技术的转录组分析流程进行介绍,包括 RNA-Seq 数据预处理、差异表达分析和高层分析。其中,RNA-Seq 数据预处理即对原始测序数据进行质控和定量计算;差异表达分析则是对基因进行筛选,通常基于统计学或机器学习两种方法;高层分析是对差异基因进一步处理,通过富集分析等手段确定基因功能和调控网络。最后,对基于 RNA-Seq 的转录组分析方法的发展进行了探讨。

关键词:RNA-Seq;转录组分析;机器学习;差异表达分析;富集分析
中图法分类号　TP391;R318

Transcriptome Analysis Method Based on RNA-Seq

GUO Mao-zu[1,2], YANG Shuai[1,2] and ZHAO Ling-ling[3]
1 School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
2 Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing 100044, China
3 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract　RNA-Seq technology has become an important method of transcriptome analysis because of its advantages of low cost, high precision and wide coverage. It provides new +means for the study of gene expression patterns, disease biomarker detection, crop stress resistance research and molecular breeding. However, the massive data generated by RNA-Seq also brings challenges to data analysis. How to effectively process and analyze RNA-Seq data has become a hot topic in bioinformatics research. The paper introduces the transcriptome analysis process based on RNA-Seq technology, including RNA-Seq data preprocessing, differential expression analysis and high-level analysis. RNA-Seq data preprocessing is to perform quality control and quantitative calculations on the original sequencing data, and differential expression analysis is to screen genes, usually based on statistics or machine learning. High-level analysis is to further process the differential genes and determine gene function and regulatory network through enrichment analysis and other means. Finally, the development prospects of RNA-Seq-based transcriptome analysis methods are discussed.

Keywords　RNA-Seq, Transcriptome analysis, Machine learning, Differential expression analysis, Enrichment analysis

20 世纪 70 年代发表的 Sanger 法奠定了人类全基因组计划的基础[1],使得测序量提高、测序成本下降,各种基于基因测序数据的研究也自此增多,比如基于基因芯片技术的目标区域测序。然而,通过下一代测序技术进行全基因组测序工作依然需要很大的成本,而且测得大量的数据需要消耗更多的时间和资源进行数据的筛选和分析,由此产生了 RNA 测序(RNA-Seq),其通过处理转录组图谱来进行研究。

转录组分析是一个强大的工具,是更好地了解控制宿主细胞命运、发展和疾病进展的潜在途径[3],同时,它也是一种快速有效的基因组调查、大规模功能基因和分子标记鉴定的方法[4]。二十年来,转录组分析经历了基因芯片技术、基于标签的测序方法(如基因表达序列分析(SAGE)和 cap 分析基因表达(CAGE))到 RNA-Seq 的过程。相比于其他方法,RNA-Seq 不依赖于生物体基因的先验知识,可以揭示转录物之间连接的精确位置和外显子之间的连接点,发现单核苷酸多态性(Single Nucleotide Polymerphism,SNP)。已有研究表明,相比于基因芯片,RNA-Seq 增加了转录组的覆盖范围,并且为鉴定组织特异性表达以及区分密切相关的旁系同源物的表

基金项目:国家自然科学基金(61532014,61871020);北京市教委科技计划重点项目(KZ201810016019);北京市属高校高水平创新团队建设计划项目(IDHT20190506);国家重点研发计划子课题(2016YFC0901902-5);2020 年度研究生创新项目(PG202005)
This work was supported by the National Natural Science Foundation of China (61532014, 61871020), Key Project of Science and Technology Plan of Beijing Municipal Commission of Education (KZ201810016019), Beijing University High-level Innovation Team Building Plan Project (IDHT20190506), National Key R&D Program of China (2016YFC0901902-5) and 2020 Graduate Innovation Project (PG202005).
通信作者:赵玲玲(zhaoll@hit.edu.cn)

达谱提供了更高的分辨率[5]。RNA-Seq 的最大优点是它既定性又定量，因此可以测量低丰度转录物的表达水平。此外，它允许定量亚型的表达水平[6]。而且，RNA-Seq 的成本更低，具有更好的覆盖范围和分辨率，还可研究总 RNA，pre-mRNA 和非编码 RNA[7]。但是，由于 RNA-Seq 对全体 RNA 进行测序，而细胞内核糖体 RNA 和线粒体 RNA 的占比很大，导致其余 RNA 的读取数量和表达水平降低。

随着国内外众多学者研究的推进，基于 RNA 测序数据的转录组分析流程已经趋于成熟，然而，少有学者对其进行系统的总结。针对这个情况，本文对基于 RNA-Seq 的转录组分析方法进行了介绍，以加深对转录组分析的理解，同时为准备加入转录组分析研究的学者提供帮助。文章首先简单介绍了 RNA-Seq 数据的测序方法，然后对 RNA-Seq 原始数据格式、预处理方法、差异表达分析和高层分析的方法进行讨论，最后进行总结和展望。

1 RNA-Seq 数据预处理

RNA-Seq 数据包含了样本中不同基因片段的表达水平和转录丰度。对 RNA-Seq 数据进行高层分析之前，首先要对每个样本数据进行预处理，使样本间的基因表达量达到一个共同标准，如此才具有可比性。RNA-Seq 数据预处理及分析流程如图 1 所示。

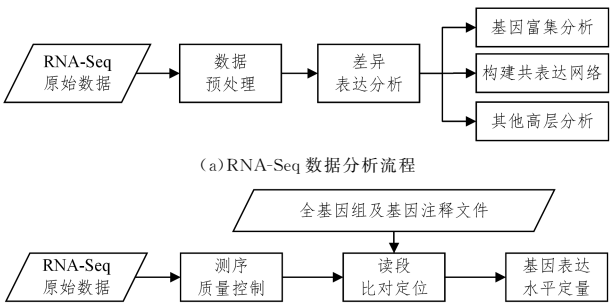

图 1 RNA-Seq 数据预处理及分析流程

Fig. 1 RNA-Seq data preprocessing and analysis process

1.1 RNA-Seq 原始数据

已有的 NGS 测序平台有 Roche 454 焦磷酸测序平台、Illumina/Solexa 合成测序平台和 ABI SOLiD 连接法测序平台等，其中以 Illumina/Solexa 平台使用最多。通过 Illumina/Solexa 测序需要经过制备 cDNA 文库、桥式 PCR 扩增和荧光信号分析等过程。

使用 Illumina/Solexa 测序平台得到的原始文件为 FASTQ 格式。如图 2 所示，每 4 行作为一个完整读段(read)，其中第一行以“@”开头，后接序列标识和对应的描述信息；第二行为该读段的碱基序列；第三行以“+”开头，后接内容与第一行相同，也可以省略标识信息；第 4 行对应第 2 行的碱基序列，是每一个碱基的测序质量得分。有关 FASTQ 格式的碱基质量得分体系等详细定义可查阅文献[8]。

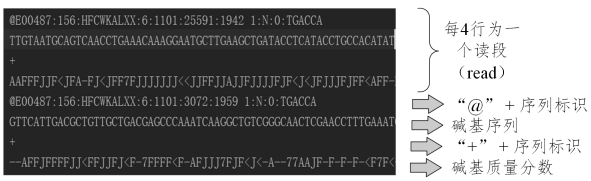

图 2 FASTQ 格式的原始测序数据示例

Fig. 2 Example of raw sequencing data in FASTQ format

1.2 测序质量控制

得到 RNA-Seq 原始数据后，首先需要对数据进行筛选，去除原始数据中测序质量比较低的读段(read)和衔接子(adaptor)序列，同时过滤掉读段长度异常的序列，其中读段的测序质量在不同的平台有不同的评分标准。目前已经开发了许多工具用于原始数据的质量控制，比如 Fastqc 和 RSeQC 等。

1.3 读段定位

原始数据经过质控后，需要将全部读段通过序列映射(mapping)定位到参考基因组，这样才能进行后续的一系列操作。序列定位的操作需要消耗大量的计算资源，因此需要高效的算法支持，主要有空位种子索引法(spaced-seed indexing)和 Burrows-Wheeler 转换(BWT)[9]。空位种子索引法的思想是将一段读段拆分成几个子段(称为种子)，然后对每个子段进行定位，并扩展到整段读段上；如果一个完整的读段能够完美定位到基因组上，那么其子段显然也能完全定位到同样的位置。BWT 的方法则是对基因组进行压缩，然后对待定位读段一次读取一个碱基并将其与压缩后的基因组进行比对，先解决一个碱基的定位，在此基础上解决两个碱基的定位，依次类推到整个读段完成定位。基于 BWT 的算法比空位种子索引法要复杂得多，然而其速度能够达到后者的 30 倍。

基于上述两种算法的程序，如 Map 和 Bowtie 最初开发出来是用于解决短读段(20～40 bp)的定位，默认允许两个错配，现在的测序数据已经能够产生超过 100 bp 的读段，相应地需要允许更多的错配。另一方面，当待比对的序列中存在碱基的插入和缺失时，上述两种算法无法取得良好的结果，于是文献[10]开发了改进的局部比对 Smith-Waterman 算法。Smith-Waterman 算法是一种动态规划算法，该算法在给定的打分模式下可以找到两个序列间最优的局部比对，算法核心是构造两个序列间的得分矩阵，并从矩阵中得分最高的元素开始回溯到得分为 0 的元素，其回溯路径揭示了两个序列间的最优比对方式。

值得注意的是，进行定位时，由于序列中存在的外显子的接合位点会影响比对效果，因此还需要利用基因组注释文件提取外显子序列来解决这个问题。此外，在排除测序错误和噪声的影响后，仍然可能存在无法定位到参考基因组的序列，这些序列可能就是潜在的未注释的新基因。因此，RNA-Seq 技术能够检测新基因，这也是其相对于基因芯片的一个优势。

原始数据完成定位后以 SAM 或者 BAM 格式保存，其中 BAM 是 SAM 文件二进制转化后的格式，能够降低存储负担。

1.4 基因表达水平定量

完成读段的比对定位后，每一个基因都有一个表达计数，这个计数就是该基因在转录样本中的表达次数，它由定位到该基因的读段数决定。显然，越长的基因，其表达计数也会更多。另外，由于测序深度的影响，以及考虑到有可能的样本间表达量的比较，一般都会对表达计数进行标准化，常用的标准化方式有计算基因的 RPKM 和 FPKM 值。计算公式为：

$$RPKM=\frac{10^6\times n_r}{L\times N} \tag{1}$$

$$FPKM=\frac{10^6\times n_f}{L\times N} \tag{2}$$

其中，n_r 和 n_f 分别为比对至目标基因的读段(read)、片段(fragment)数量；L 是目标基因的外显子长度之和除以 1 000，单位是 kb；N 是总有效比对至基因组的 read 数量。

Wagner 等人认为 RPKM 的预期含义是相对摩尔 RNA 浓度(rmc)的度量，并且表明对于每一组转录本，平均 *rmc* 是一个常数，即映射的转录本数量的倒数；并证明 RPKM 不能保证不变性，因而不能作为 rmc 的精确度量，从而提出了改进后的 TPM 标准化方法，以消除 RPKM 中的度量偏差[11]。而 FPKM 因为与 RPKM 的类似定义，尤其当原始数据是单端(SE)测序时，二者完全一致，因此也不适合作为精确度量。

2 基因差异表达分析

RNA-Seq 数据的差异表达分析是后续高层分析的基础，本节对主流的差异表达分析方法进行了介绍。

2.1 基于统计学的假设检验

一类方法是使用统计学的假设检验进行差异表达分析，常用的方法有 Fisher 精确检验[12]、似然比检验(LRT)[13]等。而当 RNA-Seq 数据中存在大量样本重复时，已有的微阵列数据分析方法同样可以用来处理 RNA-Seq 数据。

早期的研究方法是使用泊松分布对数据进行建模[14]，由于泊松分布下均值等于方差，因此可以使用负二项分布更好地拟合 RNA-Seq 数据。已有研究人员基于负二项分布开发了一些用于差异表达分析的 R 包，如 edgeR[15]和 DESeq2[16]。edgeR 最初用于两组数据的精确检验，之后通过广义线性模型进行了拓展以用于多因素设计，此外，edgeR 还可以用于分析复制水平极低的数据；DESeq2 则通过对数据施加层次模型得到关于对数倍数变化和离散的收缩估计，其分析结果具有稳定性和可重复性。然而，在较小的显著性水平下，基于负二项分布的方法会提高 I 型错误率(Type I Error Rate)[17-18]。

Fisher 精确检验基于超几何分布，是在零假设的情况下检验两类样本中数据是否有非随机相关性的统计学方法，利用列联表计算 P 值(P-value)以验证零假设。Chen 等人采用 Fisher 精确检验检测了人类口腔鳞状细胞癌中 p73，p63 和 p53 基因的差异表达[19]；Torres 和 David 等则通过限制基因集中单个基因的 p 值来防止一小部分基因决定整个基因组的统计显著性，从而改进了 Fisher 对基因集的差异表达分析[20]。

似然比检验的主要思想是检验在有、无约束条件的情况下似然函数最大值的差别，若约束条件有效，那么该条件下似然函数的最大值不会大幅减小。为了简化对上千基因进行似然比检验的计算复杂性，Hossain 等人提出了近似似然比检验(ALRT)。为了从 RNA-Seq 数据中检测差异表达的单核苷酸变体(SNV)，Fu 等提出了 MutRSeq[21]。该方法使用层次似然方法对突变事件和 RNA-Seq 数据的读段计数进行建模，并引入了基于似然比的检验统计量，不仅能够检测总体表达水平的变化，还可以检测等位基因特异性表达模式的变化。MutRSeq 还具有检测一个基因/途径中的多个突变的能力。Fu 等人还将 MutRSeq 用于乳腺癌数据集并成功检测出三阴性乳腺癌肿瘤和其他乳腺癌肿瘤亚型之间具有非同义突变的差异表达基因。此外，RNA-Seq 数据异质性的特点使得对数据的分布拟合变得困难，传统的非参数检验(如基于秩的 Wilcoxon 检验)虽然对异质性鲁棒，却会丢失数据信息[22]。Xu 等针对 RNA-Seq 数据特征，提出了一种具有均值-方差关系约束的经验似然比检验(ELTSeq)方法[22]进行差异表达分析。该方法能够保留原始数据的大量信息，并且可以处理每个组的不同程度的异质性。

2.2 结合机器学习的方法

寻找差异表达基因可以转换为提取数据特征的操作，此时每个基因的表达量(如 RPKM，TPM)都被当作一个特征。此类方法通常是在通过假设实验得到差异表达基因之后，结合 RNA-Seq 数据的标签(例如患病组和对照组)对其进行进一步的筛选，以得到范围更精确的候选基因集合。

一种筛选差异表达基因的思想是，首先确定每个差异表达基因的重要性(即在分类器中的权重)并进行排序，然后以此确定差异表达基因。Zhao 等人使用随机森林算法对差异表达 miRNA 进行排序后，通过包裹法确定了与肝癌相关性最高的 14 个 miRNA[23]。Bai 将经过 DESeq2 筛选过的差异表达基因放入 3 种机器学习模型(随机森林、支持向量机、逻辑回归)中进一步筛选肝癌中糖链相关的关键候选基因(hub-gene)[24]。

非负矩阵分解是一种无监督的、基于部分的表示和降维范式，它通过乘法更新规则将一个非负矩阵 $\boldsymbol{V}$ 分解为两个低阶非负矩阵，于 1999 年提出并首次用于面部表情识别和文本挖掘[25]。NMF 对因子分解的直观解释和对复杂性的隐式稀疏表示的优势使其能够识别突出特征，并且已有应用于差异表达分析的案例[26]。由于 NMF 不利用样本标签信息聚类，导致它的预测性能可能差于表型特征精确的直接样本分配。Wang 等人通过将 Fisher 判别准则集成到 NMF 中，即将类内散布和类间散布之间的差异作为惩罚项，发展了判别非负矩阵分解[27]；Jia 等人则将其用于 RNA-Seq 数据的差异基因排序，并通过比较实验验证了其性能[28]。DNMF 尽管对噪声鲁棒且具有学习稀疏和基于部位的表示的潜力，但其理论上仍然存在缺陷，即当收敛到固定极限点时，它不能稳定识别上调和下调的基因。

支持向量机是一种对数据进行二元分类的有监督机器学习分类器，具有稀疏性和稳健性。2002 年，Isabelle 等人提出了一种基于递归特征消除的支持向量机(SVM-RFE)方法，根据基因在分类器中的权重进行选择[29]。相比于基线方法，SVM-RFE 能够消除基因冗余，得到更紧凑的差异表达子集。Zhang 等提出了一种具有与 SVM-RFE 类似的递归策略的方法(R-SVM)[30]，其与 SVM-RFE 的预测性能相近，但 R-SVM 对噪声和异常值具有更强的鲁棒性。另外，基于 SVM 的方法在数据的分类精度上表现更好，但对差异表达基因的筛选要弱于单变量方法[30]。为了减小表达数据的高维度对分类效果的影响，Wang 等人在 SVM 方法上加入遗传算法，对初步得到的差异表达基因再次进行特征筛选，实验表明加入遗传算法对数据进行降维可以提高之后 SVM 分类器的分类准确率[31]。

3 数据高层分析

得到差异表达基因只是通过转录组分析研究目标生物问题的基础，想要揭示基因调控的分子机制，还需要从功能上对其进行研究。该部分遇到的问题与对基因芯片的研究中的问题相同，因此可以借鉴对芯片数据后续高层分析的方法。此外，对差异表达基因的功能分析，还需要结合已有的生物学知识。

基因组测序已明确表明，所有真核生物中指定核心生物学功能的基因大部分是相同的，一类共享蛋白在一种生物中的生物学作用的知识通常可以转移到其他生物中[32]。基因本体(Geneontology，GO)是一个生物信息学领域广泛使用的本体，分为细胞组分(Cellular Component)、分子功能(Molecular Function)、生物过程(Biological Process) 3 个独立的本体[32]。GO 使用有向无环图连接不同的术语，统一了不同的生物学数据库中基因和基因产物的注释，使得人们能够基于共同的生物学查询、检索基因和蛋白质。使用计算机，利用基因信息预测高层次和更复杂的细胞活动和生物体行为的想法，催生了京都基因与基因组百科全书(KEGG)，它整合了基因组、化学和系统功能信息。KEGG 作为解释生物系统的高级功能的知识库，已经广泛用于高通量测序数据的注释和整合。在生物信息学研究中，对差异表达基因进行 GO 类别富集分析和 KEGG 信号通路富集分析，能够解释基因的功能和分子调控机制。常用的富集分析工具有 David，Metascape 和 String 等。

基因共表达网络可将功能未知的基因与生物学过程关联，对候选基因进行排序或识别转录调控机制，通过 RNA-Seq 数据构建的共表达网络还可以推断非编码基因与疾病的关联[33]。共表达网络可以分为有向和无向共表达网络、权重和非权重共表达网络等，通常使用机器学习的聚类算法对相似基因进行分组。常用的是基于层次聚类的加权基因共表达网络分析(WGCNA)，已有大量实验证明 WGCNA 对多个条件下的共表达模块的识别效果良好，但是 WGCNA 的良好表现需要有关样本条件的信息。如果希望从样本组中鉴定亚组，那么双聚类算法则是更适合的方法。此外，随着 RNA-Seq 数据样本量和测序质量的不断提高，应用双聚类算法也能更准确地识别组织特异性和细胞类型特异性疾病的相关模块和调控机制。

此外，完成差异表达分析之后的数据还可以结合机器学习的方法进行特征筛选，以获得疾病早期诊断和预后的生物标志物，如 miRNA[23]，lncRNA[34]等。

结束语 近年来测序技术进展迅速，尤其是 RNA-Seq 技术的诞生，使得获取测序数据的成本越来越低，大量的测序数据为人们对生物信息学的研究提供了极大的支持。基于高通量测序技术的转录组分析可用于发现生物不同发育阶段的基因表达模式、疾病的发病机制和作物的抗逆性研究等。本文尝试对基于 RNA-Seq 的转录组分析的方法进行梳理，内容涵盖 RNA-Seq 数据格式、数据预处理方法、差异表达分析方法和高层分析方法，以期对利用 RNA-Seq 数据进行转录组分析的研究人员提供帮助。

在转录组测序方面，Roche 454 焦磷酸测序平台、Illumina/Solexa 合成测序平台和 ABI SOLiD 连接法测序平台等 NGS 测序平台已经发展出了成熟的测序技术，并被广泛应用于各种生物的测序。针对原始数据的预处理流程，学者们已开发出了各种成熟的工具，如用于测序数据的质量控制的 Fastqc 和 RSeQC，用于读段比对定位的基于不同计算机算法的 Map，Bowtie 和 Hisat2 等。针对样本间数据的表达水平定量，研究者们也已经提出了多种计算方法，如 RPKM，FPKM，TPM 等[11]，并在大量的文献中进行了应用。在基因差异表达分析方面，则有使用统计学的假设检验进行差异表达分析的 Fisher 精确检验[12]、似然比检验(LRT)[13]等方法，还有结合机器学习对差异分析结果进行进一步挖掘的方法，如使用判别非负矩阵分解算法[28]和基于递归特征消除的支持向量机(SVM-RFE)算法[29]对差异表达基因排序以发现 hub 基因。数据高层分析方面，目前已发展出了基因 GO 分析[32]、KEGG 富集分析、WGCNA 等方法用于对差异基因进行功能上的挖掘。

尽管 RNA-Seq 技术已经展现出了显著的优势，如测序成本低、测序精度高等，但是其庞大的数据量也对数据分析带来了极大的挑战。传统的统计分析方法已经无法满足转录组分析的需求，机器学习技术的发展为此提供了可能的解决途径，其中的支持向量机、聚类等算法已经广泛用于差异表达分析[30]、富集分析[23,33]等方面，结合机器学习算法进行转录组分析已经成为了一种趋势，有待于更多学者对此进行研究。

参 考 文 献

[1] SANGER F, NICKLEN S, COULSON A R. DNA sequencing with chain-terminating inhibitors [J]. Proceedings of the National Academy of Sciences, 1978, 74(12): 5463-5467.

[2] MARCEL M, MICHAEL E, ALTMAN W E, et al. Genome sequencing in microfabricated high-density picolitre reactors [J]. Nature, 2005, 437: 158-160.

[3] MUTZ K O, HEILKENBRINKER A, LÖNNE M, et al. Transcriptome analysis using next-generation sequencing [J]. Current Opinion in Biotechnology, 2013, 24(1): 22-30.

[4] MOROZOVA O, HIRST M, MARRA M A. Applications of New Sequencing Technologies for Transcriptome Analysis [J]. Annual Review of Genomics & Human Genetics, 2009, 10(1): 135-151.

[5] SEKHON R S, ROMAN B, HIRSCH C N, et al. Maize Gene Atlas Developed by RNA Sequencing and Comparative Evaluation of Transcriptomes Based on RNA Sequencing and Microarrays [J]. Plos One, 2013, 8(4): e61005.

[6] WANG Z, GERSTEIN M, SNYDER M. RNA-Seq: a revolutionary tool for transcriptomics [J]. Nature Reviews Genetics, 2010, 10(1): 57-63.

[7] KUKURBA K R, MONTGOMERY S B. RNA Sequencing and Analysis [J]. Cold Spring Harbor Protocols, 2015, 2015(11): 951.

[8] COCK P J, FIELDS C J. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants [J]. Nucleic acids research, 2010, 38(6): 1767-1771.

[9] TRAPNELL C, SALZBERG S L. How to map billions of short reads onto genomes [J]. Nature Biotechnology, 2009, 27(5): 455-457.

[10] SMITH T F, WATERMAN M S. Identification of common molecular subsequences [J]. Journal of Molecular Biology, 1981, 147(1): 195-197.

[11] WAGNER G P, KIN K, LYNCH V J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples [J]. Theory Biosci, 2012, 131(4): 281-285.

[12] HWANG S G, KIM K H, LEE B M, et al. Transcriptome analy-

sis for identifying possible gene regulations during maize root emergence and formation at the initial growth stage [J]. Genes & Genomics,2018,40(7):755-766.

[13] SHI Y,JIANG H,FRANK E S. rSeqDiff:Detecting Differential Isoform Expression from RNA-Seq Data Using Hierarchical Likelihood Ratio Test [J]. Plos One,2013,8(11):e79448.

[14] JOHN C M,CHRISTOPHER E M,SHRIKANT M M,et al. RNA-seq:an assessment of technical reproducibility and comparison with gene expression arrays [J]. Genome Research,2008,18(9):1509-1517.

[15] SMYTH G K. edgeR:a Bioconductor package for differential expression analysis of digital gene expression data [J]. Bioinformatics,2010,26(1):139.

[16] LOVE M I,HUBER W,ANDERS S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [J]. Genome Biology,2014,15(12):550.

[17] LUND S P,NETTLETON D,MCCARTHY D J,et al. Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates [J]. Statistical Applications in Genetics & Molecular Biology,2012,11(5).

[18] REEB P,JUAN S. Evaluating statistical analysismodels for RNA sequencing experiments [J]. Frontiers in Genetics,2013,4:178.

[19] CHEN Y K,HUSE S S,LIN L M. Differential expression of p53,p63 and p73 proteins in human buccal squamous-cell carcinomas [J]. Clinical Otolaryngology,2003,28(5):451-455.

[20] DAVID J,TORRES,JUDY L,et al. Self-Contained Statistical Analysis of Gene Sets[J]. Plos One,2016,11(10):e0163918.

[21] FU R,WANG P,MA W P,et al. A statistical method for detecting differentially expressed SNVs based on next-generation RNA-seq data [J]. Biometrics,2017,73(1):42-51.

[22] XU M Q,CHEN L. An empirical likelihood ratio test robust to individual heterogeneity for differential expression analysis of RNA-seq [J]. Briefings in Bioinformatics,2018,19(1):1.

[23] ZHAO X,DOU J,CAO J L,et al. Uncovering the potential differentially expressed miRNAs as diagnostic biomarkers for hepatocellular carcinoma based on machine learning in The Cancer Genome Atlas database [J]. Oncology Reports,2020,43(6):1771-1784.

[24] BAI Y F. Screening of sugar chain related genes in hepatocellular carcinoma based on network analysis and machine learning [D]. Harbin:Harbin Institute of Technology,2019.

[25] LEE D D,SEUNG H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature,1999,401(6755):788.

[26] KONG W,MOU X Y,HU X H. Exploring matrix factorization techniques for significant genes identification of Alzheimer's disease microarray gene expression data [J]. BMC Bioinformatics,2011,12(5).

[27] WANG Y,JIA Y D. Fisher non-negative matrix factorization for learning local features [C]// Asian Conference of Computer Vision. 2004:27-30.

[28] JIA Z L,ZHANG X,GUAN N Y,et al. Gene ranking of RNA-seq data via discriminant non-negative matrix factorization [J]. PloS One,2015,10(9):e0137782.

[29] GUYON I ,WESTON J,BARNHILL S,et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning,2002,46(1/2/3):389-422.

[30] ZHANG X G,LU X,SHI Q,et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data[J]. BMC Bioinformatics,2006,7(1):197.

[31] WANG W,LIU H. Genetic algorithm and support vector machine-based gene microarray analysis [J]. Journal of Clinical Rehabilitative Tissue Engineering Research,2010,14(17):3099-3103.

[32] ASHBURNER M M,BALL C A C,BLAKE J A J,et al. Gene Ontology:tool for the unification of biology. The Gene Ontology Consortium [J]. Nature Genetics,2000,25(1):25-29.

[33] SIPKO V D,URMO V,ADRIAAN V D G,et al. Gene co-expression analysis for functional classification and gene-disease predictions [J]. Briefings in Bioinformatics,2018,19(4):575-592.

[34] ZHOU M,ZHAO H Q,XU W Y,et al. Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma [J]. Molecular Cancer,2017,16(1):16.

GUO Mao-zu, born in 1966, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning, smart city, bioinformatics, etc.

ZHAO Ling-ling, born in 1980, Ph.D supervisor. Her main research interests include machine learning, smart city, bioinformatics, etc.