第50卷第7期 中国科学技术大学学报 Vol. 50, No. 7
2020年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2020
文章编号:0253-2778(2020)07-1013-06

单细胞 RNA 序列数据的 PBMC 相关细胞的识别

龚乐君[1,2],周余海[1,2],程逸飞[1],高志宏[3],李华康[1,4,5]

(1. 南京邮电大学计算机学院、软件学院、网络空间安全学院,江苏南京 210023;
2. 江苏省大数据安全与智能处理重点实验室,江苏南京 210023;3. 浙江省智慧医疗工程技术研究中心,浙江温州 325035;
4. 自然资源部城市国土资源监测与仿真重点实验室,广东深圳 518034;5. 苏州派维斯信息科技有限公司,江苏苏州 215011)

摘要:细胞类型鉴定是单细胞 RNA 测序的主要任务之一.针对整个问题,提出基于随机森林的细胞类型自动识别(automatic identification of cell type based on random forest, AICTRF)方法来识别单细胞测序数据中的细胞类型,该方法使用随机森林分类模型进行训练,根据训练的模型进而预测未知的细胞类型.在人类外周血单核细胞(PBMC)测序数据集上训练了随机森林分类模型,利用该模型预测了人类 PBMC 中 B 细胞的相关亚型细胞类型.实验结果表明,该方法可以帮助相关研究人员快速而有效地自动识别单细胞测序数据中的细胞类型.

关键词:单细胞 RNA 测序数据挖掘;细胞类型;B 细胞亚型;聚类;分类
中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.07.019

引用格式:龚乐君,周余海,程逸飞,等. 单细胞 RNA 序列数据的 PBMC 相关细胞的识别[J]. 中国科学技术大学学报,2020,50(7):1013-1018.
GONG Lejun, ZHOU Shehai, CHENG Yifei, et al. Identification of PBMC-related cells of single cell RNA sequence data[J]. Journal of University of Science and Technology of China, 2020, 50(7): 1013-1018.

Identification of PBMC-related cells of single cell RNA sequence data

GONG Lejun[1,2], ZHOU Shehai[1,2], CHENG Yifei[1], GAO Zhihong[3], LI Huakang[1,4,5]

(1. *School of computer science, Nanjing university of posts and telecommunications, Nanjing* 210023, *China*;
2. *Jiangsu Key Lab of Big Data Security & Intelligent Processing, Nanjing* 210023, *China*;
3. *Zhejiang Engineering Research Center of Intelligent Medicine, Wenzhou* 325035, *China*;
4. *Key Laboratory of Urban Land Resources Monitoring and Simulation, Shenzhen* 518034, *China*;
5. *Suzhou Privacy Information Technology Company, Suzhou* 215011, *China*)

Abstract: Cell type identification is one of the main tasks of single cell RNA sequencing. This paper proposes an automatic identification of cell types based on random forest (AICTRF) method to identify cell types in single-cell sequencing data. This method uses the random forest classification model for training, and then predicts unknown cell types according to the trained model. A random forest classification model was trained on human peripheral blood mononuclear cells (PBMC) sequencing data set to predict the cell types of related subtypes of human PBMC B cells. The results show that the proposed method can help researchers automatically identify cell types in single-cell sequencing data.

Key words: scRNA-seq data mining; cell type; B cell subtype; clustering; classification

0 引言

单细胞 RNA 测序是一项突破性技术,在生物学研究领域有广泛的应用[1-3].研究表明,在研究不同生物体、组织和发育阶段的细胞异质性方面,单细胞 RNA 序列具有相当大的优势[4-7].由于大量的数据通常是在单细胞 RNA 测序实验[8]中产生的,尤其是当细胞数量很大(通常为数千)时,因此合理的对单细胞 RNA 序列分析的快速有效的计算方法非常重要.通常,分析单细胞 RNA 序列的第一步是将一个细胞指定为一种已知的细胞类型.传统方法首先以无监督的方式将细胞分组到不同的聚类中,然后找到标记基因,最后使用这些基因来手动为每个聚类分配细胞类型.这些方法需要事先知道细胞

收稿日期:2020-06-03;**修回日期**:2020-06-21
基金项目:国家自然科学基金(61502243),浙江省智慧医疗工程技术研究中心资助项目(2016E10011),中国博士后基金(2018M632349);江苏省"六大人才高峰"高层次人才项目(XYDXX-204);城市国土资源监测与仿真重点实验室开放基金(KF-2019-04-011, KF-2019-04-065);苏州市姑苏科技创业天使计划(CYTS2018233);南京邮电大学引进人才科研启动基金(NY217136)资助.
通讯作者:龚乐君(通讯作者),女,1978 年生,博士/副教授. 研究方向:数据挖掘. E-mail: glj98226@163.com

类型的标记基因,并且需要一定程度的人工来分配细胞类型,因此这个过程通常很耗时,并且需要特定领域的专业知识,对缺少相关专业背景的用户来说是一个障碍.为了克服现有方法的局限性,需要设计新的细胞类型识别方法.

细胞类型的定义可以看作机器学习中的一个分类问题.使用机器学习算法对序列数据进行分类面临一系列的挑战:第一,无监督聚类算法[9]的选择.第二,单细胞 RNA 测序数据特有的高维问题.第三,分类过程中面临的数据不平衡问题.第四,必须提供系统的方法来评估分类性能.

为了解决上述问题,本文提出了基于随机森林的细胞类型自动识别(automatic identification of cell types based on random forest,AICTRF)方法,AICTRF 可以获得每个预定义细胞类型的特征,并利用这些特征预测单个细胞的细胞类型.该方法计算效率高,不需要该领域的专业知识.AICTRF 允许用户快速识别其数据集中的细胞类型,这使得对 scRNA-seq 数据集的分析更加高效.

本文的主要贡献如下:

(Ⅰ)对人类外周血单个核细胞(PBMC)的单细胞 RNA 测序数据使用 K-means 方法[10]进行聚类分析,通过查找标志基因后,确定细胞类型,再对其中的 B 细胞进行二次聚类处理,对比标志基因得到细胞亚型;

(Ⅱ)利用主成分分析技术(PCA)[11]对单细胞 RNA 测序数据的基因维度进行降维处理;

(Ⅲ)本文提出了随机森林[12]的细胞类型自动识别(AICTRF)方法,用来对单细胞 RNA 的测序数据进行细胞类型的划分;

(Ⅳ)使用随机森林的细胞自动识别与使用支持向量机[13]、逻辑回归[14]的自动识别方法进行比较,证实使用随机森林的细胞自动识别具有更好的泛化性和准确率.

1 本文研究方法

本文的研究方法主要是通过聚类与分类从 PBMC 数据中识别 B 细胞的类型.通过聚类获取不同细胞簇的标记基因,利用标记基因获取对应于细胞簇的细胞类型注释,从中提取 B 细胞;进而注释 B 细胞中的细胞亚群.首先分别使用多种有监督的机器学习分类算法对标注的 B 细胞亚群进行分类,比较各种算法的分类性能.再选取分类性能最好的算法应用在 PBMC 的 B 细胞亚型细胞类型的识别中.这一自动化识别方法即 AICTRF.

1.1 数据集

本文使用的单细胞 RNA 序列数据来源于 PBMC 数据集(PBMC4K),该数据集由 10X Genomics[15]提供,用于构建和验证分类模型.数据集来自健康人体供体的外周血单核细胞,约有 4340 个细胞.有关该数据集的详细信息,请参见表 1.

1.2 方法框架

本文从 10X Genomics 获得原始测序数据,首先经过相应的预处理分析,包括质量控制、与参考基因组的比对、定量及单细胞基因表达矩阵的构建.然后对单细胞基因表达矩阵的数据进行聚类,通过基因的差异表达分析获得不同的标记基因簇,并通过细胞标记基因数据库[16]获得由标记基因代表的细胞类型.分别使用监督学习的支持向量机(SVM),随机森林(RF)和逻辑回归 3 种分类算法来构建分类模型,选择出其中最优的算法预测新输入的单细胞基因表达矩阵数据的细胞类型.图 1 描述了本方法的识别过程.

表 1 数据集详细信息

Tab. 1 Details of data set

Name	PBMC4k
Description	Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Estimated Number of Cells	4340
Number of Reads	379,462,522
Reads Mapped to Genome	96.1%

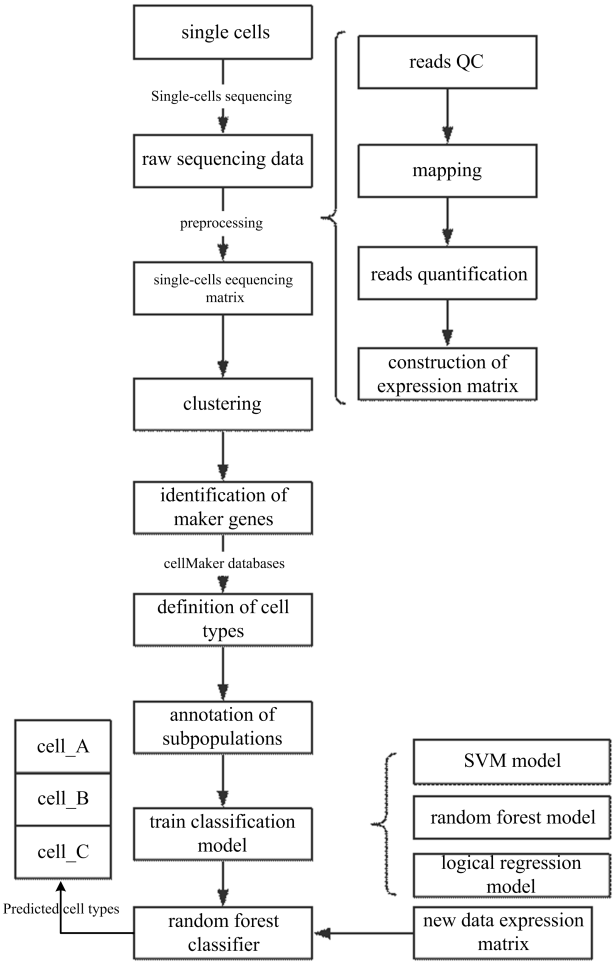

图 1 识别细胞类型的 AICTRF 框架

Fig. 1 AICTRF Framework for identifying cell types

1.3 K-means 聚类算法

K-means 算法是聚类算法.聚类算法就是根据同类别的数据在特征上具有一定的相似性,而不同

类别的数据在数据特征上具有一定的差异性. 将相似性较高的数据对象划分成一类，相似性较低的划分成不同类. K-means 算法采用欧氏距离作为判断样本相似性的标准，对于给定的样本集，根据样本之间的距离将样本集分成 k 个聚类. 欧几里得距离的计算公式为

$$\mathrm{dist}(x_i,x_j=\sqrt{(x_i-x_j)(x_i-x_j)^{\mathrm{T}}}) \quad (1)$$

式中，x 代表具有 m 个属性的行向量. 在 K-means 算法的聚类过程中，需要对同一类别中所有的数据对象求每个属性的均值，以此得到聚类的聚类中心. 聚类的中心是需要随着每次的迭代计算而更新的. 定义第 k 个群集的群集聚类中心是 Center_k，群集聚类中心更新公式为

$$\mathrm{Center}_k=\frac{1}{|C_k|}\sum_{x_i\in C_k}x_i \quad (2)$$

式中，C_k 表示第 k 个类簇，$|C_k|$ 表示第 k 个类簇中数据对象的数量，这里的求和函数 $\sum_{x_i\in C_k}x_i$ 是指类簇 C_k 中所有元素的总和.

K-means 算法使用误差平方和准则函数来更新聚类中心. 功能模型为

$$J=\sum_{k=1}^{k}\sum_{x_i\in c_k}\mathrm{dist}(x_i,\mathrm{Center}_k) \quad (3)$$

式(3)表示所有类簇中的数据与其聚类中心 Center_k 的距离之和. 其中 k 代表簇的数量. 当两次迭代之间的差值小于某个事先设定的阈值，即 $\Delta J<\delta$ 时，迭代终止，此时得到的聚类结果就是最终的聚类结果.

1.4　Random forest 分类算法

随机森林(random forest，RF)[17]是一种集成学习方法，该算法以决策树为基学习器，通过 Bagging 算法在决策树的采样过程中进行随机样本的选择，同时在训练过程中引入了随机属性的选择. Bagging 是一种并行式集成学习方法，采用自助采样法(booststrap sampling)，即对于一个包含 m 个样本的数据集 D，对其进行有放回的数据采样产生数据集 $D*$，其中进行有放回的数据采样次数为 m. 得到的最终采样结果，D 中有一部分样本会在 $D*$ 中多次出现，而另一部分样本不出现. 采样中样本始终不被采到的概率是 $(1-\frac{1}{m})^m$，取极限得到

$$\lim_{m\to\infty}(1-\frac{1}{m})^m=\frac{1}{e}\approx 0.368 \quad (4)$$

即通过自助采样，训练集 D 中约有 63.2%的样本会出现在采样集 $D*$ 中. 重复自助采样 T 次，可采样出 T 个含 m 个训练样本的采样集. 对于每个采样集都使用决策树算法来训练一个基学习器，将所有基学习器通过简单投票的方式进行最终分类结果的判别. RF 算法在训练其决策树的集学习器时，对属性的选择过程采用了随机属性选择，即先从该结点的属性集合中随机选择一个包含 k 个子属性的子集，然后再从这个子集中选择一个最优(信息熵最高)属性用于划分，一般情况下 k 的推荐值为 $\log_2 d$ (d 为当前结点属性的个数).

1.5　主成分分析

主成分分析(principal component analysis，PCA)是一种常用的降维方法. 目的是将样本点 x_i 投影在比原空间维度要低的新的空间超平面上，其在新的超平面上的投影为 $W^{\mathrm{T}}x_i$. 希望样本点的投影能尽可能分开，则应该使投影后样本点的方差最大化. 首先对所有样本进行中心化，即

$$x_i\leftarrow x_i-\frac{1}{m}\sum_{i=1}^{m}x_i \quad (5)$$

计算所有样本的协方差矩阵 XX^{T}，对协方差矩阵 XX^{T} 做特征值分解，取出最大的前 d 个特征值所对应的特征向量($w_1,w_2,\cdots,,w_d$)，最终的投影矩阵为

$$W=(w_1,w_2,\cdots,w_d) \quad (6)$$

由投影矩阵 W 就可以计算出样本点在新的超平面的投影 $W^{\mathrm{T}}x_i$，由此将原本高维的数据降低为维度为 d 的数据，从而实现降维.

2　实验与结果

2.1　从 PBMC 数据中提取 B 细胞

对于原始 PBMC 数据(总共 4340 个细胞)，使用 K-means 聚类算法($k=4$)获得图 2 的结果. 从图 2 可以看出，4 种细胞的数量相对平衡，没有细胞特别少的集群. 通过基因差异表达分析[18-19]，可以获得不同细胞群的标志基因. 不同细胞群中的基因表达热力图如图 3 所示.

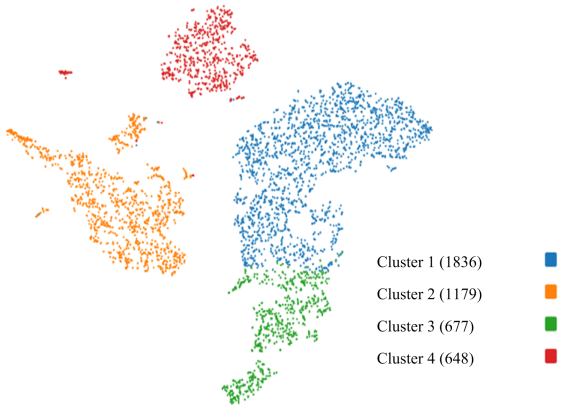

图 2　K-means($k=4$)聚类结果可视化

Fig. 2　Visualization of clustering results of K-means ($k=4$)

不同类型的细胞簇和标志基因由不同的颜色框识别，不同类型细胞簇的标志基因依次获得：簇 1(CD3E、CD3D、CD3G、IL7R)；簇 2(S100A12，CD14)；簇 3(GNLY，NKG7)；簇 4(CD79A、CD79B、MS4A1). 对应于标志基因的细胞类型可以通过 cellMarker 数据库查询得到相应的标志基因对应的细胞类型. 不同细胞簇对应的细胞类型如下表 2 所示.

表 2　对应不同细胞簇的细胞类型

Tab. 2　Cell types of different cell groups

Cell cluster	Marker gene	Cell type
Cluster1	IL7R，CD3E，CD3G，CD3D	T cell
Cluster2	S100A12，CD14	CD14+CD16+monocyte
Cluster3	GNLY，NKG7	Natural killer cell
Cluster4	CD79A，CD79B，MS4A1	B cell

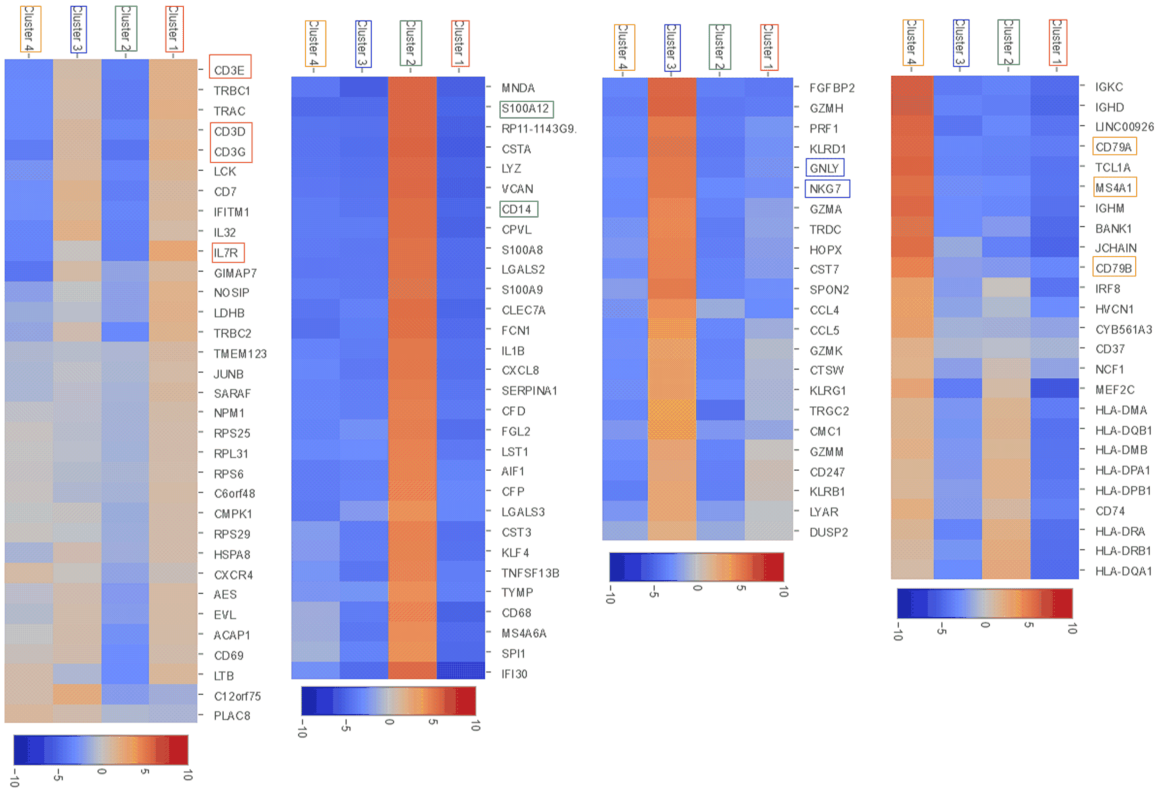

图 3　不同细胞簇的基因表达热图

Fig. 3　Heat maps of gene expression in different cell groups

部分标志基因在不同细胞簇中的分布如图 4 所示. 图 4 表明,在上面获得的标志基因代表了不同的细胞群. 该图同时也证明上文得到的标志基因在各类细胞簇中是具有代表性的.

2.2　B 细胞亚群注释

从上面获得的 B 细胞中去除一些噪声数据,获得 625 个 B 细胞. 通过 K-means 聚类得到 3 个未标记的 B 细胞亚群,分别命名为 $A1$、$A2$、$A3$ 细胞. $A1$ 细胞是浓缩的,但数量相对较少,样品数量为 17 个. $A2$ 和 $A3$ 细胞更分散、更丰富,$A2$ 细胞数为 395,$A3$ 细胞数为 213. 聚类结果如下图 5 所示,聚类结果的细胞类型注释如图 6 所示.

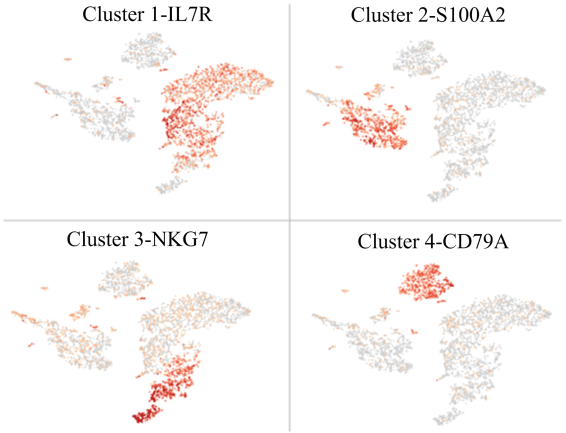

图 4　不同细胞簇中部分标记基因分布的散点图

Fig. 4　Scatter plots of partial marker genes in different cell populations

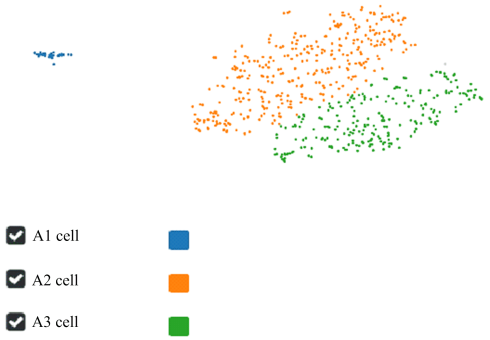

图 5　B 细胞聚类结果

Fig. 5　Clustering results of B cells

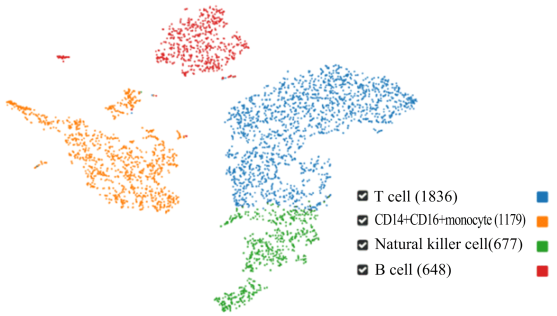

图 6　聚类结果的细胞类型注释

Fig. 6　cell type annotation of clustering results

根据基因的差异表达分析,可以获得与 3 种类型的 B 细胞亚群相对应的标志基因. 对应于 3 种细胞类型的标志基因的数量如图 7 所示. 通过查询细

胞标志基因数据库可以获得与标志基因对应的细胞类型. 查询结果如表 3 所示.

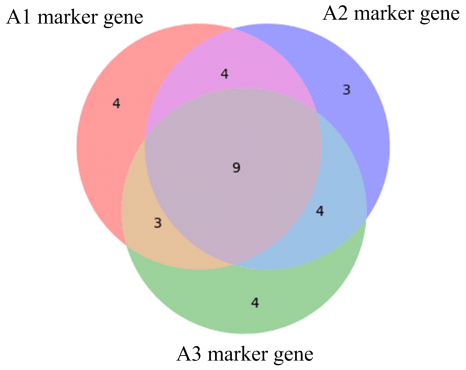

图 7　亚细胞标志基因维恩图

Fig. 7　Venn diagram of subcellular marker genes

表 3　对应于标志基因的细胞类型

Tab. 3　Corresponds to the cell type of the marker gene

Cluster	Marker gene	Cell type
*A*1 cell	MTMR11, LEPROT, C14orf2, FCMR	CD1C+_B dendritic cell
*A*2 cell	COTL1, LINC00926, HVCN1	Multilymphoid progenitor cell
*A*3 cell	CXCR4, MEF2C, RP5. 887A10. 1, CCDC24	Plasma cell

使用标志基因获得的细胞类型对 3 种类型的 B 细胞亚群进行注释，结果如图 8 所示.

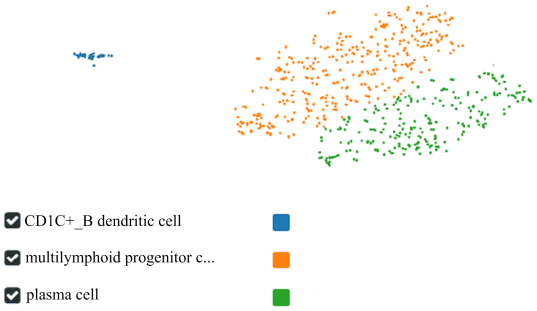

图 8　B 细胞亚群注释结果

Fig. 8　B cell subsets annotate the results

2.3　B 细胞亚群细胞分类

在使用机器学习算法对细胞进行分类的过程中，面临的第一个挑战是单细胞基因表达矩阵的高维问题. 获得的单细胞基因表达矩阵中有 5000 个基因(特征)，许多列中的基因表达值为 0. 由于基因表达值为 0 的特征可能会产生噪声，因此需要使用主成分分析(PCA)进行降维，去除基因表达矩阵中的冗余特征以减少噪声对实验结果的影响. 设置主成分分析中选取最大的前 d 个特征值的参数 d 为 500，即经过主成分分析处理后，数据维度由 5000 降低为 500. 第二个挑战是 3 种类型的 B 细胞亚群的数据不平衡. 3 种 B 细胞亚群中的细胞数分别为 17、395 和 213，使用 SMOTE 算法[20]对数据进行过采样. 过采样得到 3 种类型的 B 细胞的结果子集中的细胞数都为 395. 得到的数据集大小是 1185 个细胞，其中设置训练集和测试集的大小比例为(6∶4)，最终训练集大小是 711 个细胞，测试集大小是剩余的 474 个细胞. 本文分别采用了支持向量机(SVM)、随机森林(RF)、逻辑回归(LR)3 种分类算法，对数据集进行了训练和测试. 获得的 B 细胞亚型在五折交叉验证下的平均分类结果如下表 4 所示.

表 4　3 种细胞平均分类结果

Tab. 4　Average classification results of three cell types

Method	Accuracy	Precision	Recall	F_1_score
SVM	0. 964 13	0. 959 5	0. 958 3	0. 958 8
Random forest	0. 972 57	0. 980 2	0. 978 8	0. 979 4
Logical regression	0. 952 51	0. 912 3	0. 921 1	0. 942 3

从表 4 可以看出，随机森林细胞在五折交叉验证下的平均分类准确率为 0. 97257，精确率为 0. 9802，召回率为 0. 9788，F_1_score 为 0. 9794，高于 SVM 算法和逻辑回归的相应性能，因此我们最终选择随机森林作为 PBMC 中 B 细胞亚型细胞识别流程中的分类算法，用于识别 PBMC 的 B 细胞亚型细胞.

3　结果分析

单细胞测序提供单细胞转录物的高分辨率分析. 一般来说，单细胞序列分析的第一步是根据对标志基因的先前理解，为每个细胞指定一种细胞类型. 传统的细胞类型分配方法是先对细胞进行无监督的聚类，然后根据标志基因识别每个细胞簇的细胞类型. 然而，这种方法有几个局限性：细胞簇可能不是最佳的分离单个细胞类型，一些细胞类型可能没有标志基因. 为了更有效地识别单细胞测序中的细胞类型，本文利用随机森林对具有预定义细胞类型的细胞进行分类模型训练，并将其用于预测新数据集中的细胞类型. 与传统的无监督细胞类型识别方法相比，本文提出的 AICTRF 方法具有以下优点：①它利用所有基因来获取每种细胞类型的特征，而不是依赖有限数量的标志基因. ②不需要细胞类型标志基因的专业背景知识，传统无监督方法要求用户预先知道数据中每种细胞类型的标志基因. ③它比传统方法更有效，能够在短时间内快速让研究人员准确而高效地分辨不同的细胞类型.

4　结论

本文基于聚类和分类模型识别单细胞 RNA 序列数据的 PBMC 中的 B 细胞类型，提出了一种 AICTRF 方法自动识别单细胞测序数据中细胞类型. 该方法既利用聚类算法，又使用分类模型，能够准确预测 PBMC 数据中的 B 细胞亚型. 该方法对系统研究 scRNA-seq 数据的细胞类型注释提供参考作用，可用于快速帮助研究人员注释单细胞测序数据中的细胞类型. 预测 B 细胞中相关亚型的细胞类型的实验表明，相关细胞类型的预测准确率达到 0. 9725. 本文提供的方法可以作为现有 RNA 序列分析方法的补充，帮助相关研究人员自动识别单细

胞测序数据中的细胞类型.

参考文献(References)

[1] PAPALEXI E, SATIJA R. Single-cell RNA sequencing to explore immune cell heterogeneity[J]. Nature Reviews Immunology, 2018, 18(1): 35-45.

[2] RANTALAINEN M. Application of single-cell sequencing in human cancer[J]. Briefings in Functional Genomics, 2018, 17(4): 273-282.

[3] POTTER S S. Single-cell RNA sequencing for the study of development, physiology and disease[J]. Nature Reviews Nephrology, 2018, 14(8): 479-492.

[4] SVENSSON V, VENTOTORMO R, TEICHMANN S A, et al. Exponential scaling of single-cell RNA-seq in the past decade[J]. Nature Protocols, 2018, 13(4): 599-604.

[5] VILLANI A, SATIJA R, REYNOLDS G, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors[J]. Science, 2017: 356(6335).

[6] GRUN D, LYUBIMOVA A, KESTER L, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types[J]. Nature, 2015, 525(7568): 251-255.

[7] TIROSH I, IZAR B, PRAKADAN S, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq[J]. Science, 2016, 352(6282): 189-196.

[8] KERENSHAUL H, SPINRAD A, WEINER A, et al. A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease[J]. Cell, 2017, 169(7): 1276-1290.

[9] DUO A, ROBINSON M D, SONESON C, et al. A systematic performance evaluation of clustering methods for single-cell RNA-seq data[J]. F1000Research, 2018.

[10] KAKUSHADZE Z, YU W. * K-Means and Cluster Models for Cancer Signatures[J]. Biomolecular Detection and Quantification, 2017: 7-31.

[11] SHLENS J. A tutorial on principal component analysis[J]. arXiv: Learning, 2014.

[12] CUTLER D R, EDWARDS T C, BEARD K H, et al. Random forests for classification in ecology[J]. Ecology, 2007, 88(11): 2783-2792.

[13] CHEN P, LIN C, SCHOLKOPF B, et al. A tutorial on v-support vector machines[J]. Applied Stochastic Models in Business and Industry, 2005, 21(2): 111-136.

[14] CUCCHIARA A, HOSMER D, LEMESHOW S. Applied logistic regression[J]. Technometrics, 1992, 34(3):358.

[15] FREYTAG S, TIAN L, LONNSTEDT I, et al. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data[J]. 1000Research, 2018.

[16] ZHANG X, LAN Y, XU J, et al. CellMarker: A manually curated resource of cell markers in human and mouse[J]. Nucleic Acids Research, 2019.

[17] ZHOU Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016:78-181.

[18] ZHANG X, MALLICK H, TANG Z, et al. Negative binomial mixed models for analyzing microbiome count data[J]. BMC Bioinformatics, 2017:18(1).

[19] SHIN S, PARK J S, KIM Y, et al. Differential gene expression profile in PBMCs from subjects with AERD and ATA: A gene marker for AERD[J]. Molecular Genetics and Genomics, 2012, 287(5): 361-371.

[23] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over- sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.