

基于跨细胞类型差异分析的复杂样本反卷积

田中禾, 彭 凌, 张伟伟*

绍兴文理学院数理信息学院, 浙江 绍兴

收稿日期: 2024年10月12日; 录用日期: 2024年11月5日; 发布日期: 2024年11月12日

摘 要

复杂样本的基因组数据反映了多种细胞类型表达的平均水平。然而, 细胞组成的差异会导致许多相关分析结果产生偏差。因此, 准确估计细胞组成是分析复杂样本的第一步。目前有许多计算方法已经被开发出来估计复杂样本的细胞组成, 但由于缺乏参考数据和先验信息, 它们的应用大多有限。针对此问题, 本文开发了一种基于最优特征选择的无参考反卷积算法, 该算法通过整合一种细胞类型和其他细胞类型之间的跨细胞类型差异表达分析, 以及两种细胞类型与其他细胞类型之间的跨细胞类型差异表达分析, 迭代地搜索细胞类型特异性的最优特征, 并进行细胞组成估计。模拟研究和两个真实数据集上的实证分析表明我们的算法具有更好的性能。

关键词

复杂样本, 细胞组成, 差异表达分析, 跨细胞类型, 反卷积

Deconvolution of Complex Samples Based on Cross-Cell Type Differential Analysis

Zhonghe Tian, Ling Peng, Weiwei Zhang*

School of Mathematics Information, Shaoxing University, Shaoxing Zhejiang

Received: Oct. 12th, 2024; accepted: Nov. 5th, 2024; published: Nov. 12th, 2024

Abstract

Genomic data from complex samples measure average level of multiple cell types, and differences in cell compositions lead to biased results in many relevant analyses. Therefore, accurately estimating cell compositions has been recognized as an important first step in analyzing complex samples.

*通讯作者。

文章引用: 田中禾, 彭凌, 张伟伟. 基于跨细胞类型差异分析的复杂样本反卷积[J]. 应用数学进展, 2024, 13(11): 4870-4875. DOI: 10.12677/aam.2024.1311468

Many computational methods have been developed to estimate cell compositions, but they mostly have limited applications due to a lack of reference or prior information. In this work, we develop a feature selection method for reference-free deconvolution, which iteratively searches for cell-type specific features by cross-cell type differential analysis between one cell type versus the other cell types, as well as between two cell types versus the other cell types, and performs composition estimation. Extensive simulation studies and analyses of two real datasets demonstrate the favorable performance of our proposed method.

Keywords

Complex Sample, Cell Composition, Differential Analysis, Cross-Cell Type, Deconvolution

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

复杂样本的基因组数据测量的是其含有的多种细胞类型信号的加权平均值，而不是混合物中不同细胞类型的独立测量值，因此，包含差异分析在内的复杂样本的多种相关分析工作会由于细胞混合而导致分析结果出现偏差甚至是错误[1]。此外，有研究表明细胞组成与疾病进展和治疗反应密切相关[2]。例如，神经元细胞死亡是阿尔茨海默病(AD)的显著病理特征[3]，而免疫浸润在多种癌症类型中与肿瘤进展显著相关[4]。因此，准确估计细胞组成对分析复杂样本具有重要的意义。

传统上，许多实验方法，如荧光激活细胞分选(FACS)、免疫组织化学(IHC)和单细胞 RNA 测序(single cell RNA-seq)能够提供细胞组成信息[5]。然而，这些方法由于通量有限、成本过高、操作繁琐等原因，在大规模临床应用中受限。作为替代方案，有许多被称为“反卷积”的计算方法被开发出来。目前，这些方法大致可分为两类：基于参考基矩阵的方法[6]-[9]和不依赖于参考基矩阵的方法[10]-[12]。基于参考基矩阵的方法需要将纯化组织中获得的参考面板数据作为输入，并通过如约束线性回归或支持向量回归等方法来估计细胞类型的组成[6]。研究表明，基于参考基矩阵的方法通常比不依赖于参考基矩阵的方法得到的结果更准确[6] [7]。然而，目前的参考数据仅针对少数几种组织类型(如血液、乳腺和大脑)，且仅来自少数个体，因此基于参考基矩阵的方法应用范围受到极大限制。此外，当复杂样本与参考数据在临床条件和表型特征上存在显著差异时，基于参考基矩阵的方法可能会导致细胞组成估算不准确[13]。对于没有适当参考信息的组织，不依赖于参考基矩阵的方法提供了更好的解决方案[14]。该方法基于某种类型的因子分析(如非负矩阵分解)，纯细胞类型的特征谱和细胞组成被同时进行估算[15]。然而，由于不依赖于参考基矩阵的方法中需要估计的参数众多，因此其面临着准确度较低的问题。因此，开发一种能够改进不依赖于参考基矩阵的反卷积新算法是一个值得研究的重要问题。

本文中，我们开发了一种基于跨细胞类型差异分析来筛选细胞类型特异性的特征并进行细胞组成估计。该算法通过整合一种细胞类型和其他细胞类型之间的跨细胞类型差异分析，以及两种细胞类型与其他细胞类型之间的跨细胞类型差异分析，迭代地搜索细胞类型特异性的最优特征，并进行细胞组成估计。广泛的模拟研究和两个真实数据集上的实证分析证明了我们方法的优越性。

2. 模型构建

给定一个 $m \times n$ 的矩阵 Y ，其表示 m 个特征在 n 个复杂样本上高通量实验生成的数据。我们假设这些

复杂样本是 K 种细胞类型的混合物, W 为 $m \times K$ 的矩阵, 表示 m 个细胞类型特异性特征在 K 种细胞类型上的表达谱, H 为 $K \times n$ 的矩阵, 表示细胞组成比例矩阵, H 的各元素要求非负, 并且每列元素之和为 1。假设观测数据 Y 为细胞类型特异性特征谱的线性组合, 组合系数为细胞组成比例, 即

$$Y = WH + \varepsilon$$

其中 ε 是 $m \times n$ 的误差矩阵。将我们的算法命名为 DEcd, 我们算法的目标是仅利用观测数据 Y 来估算细胞组成比例矩阵 H 。

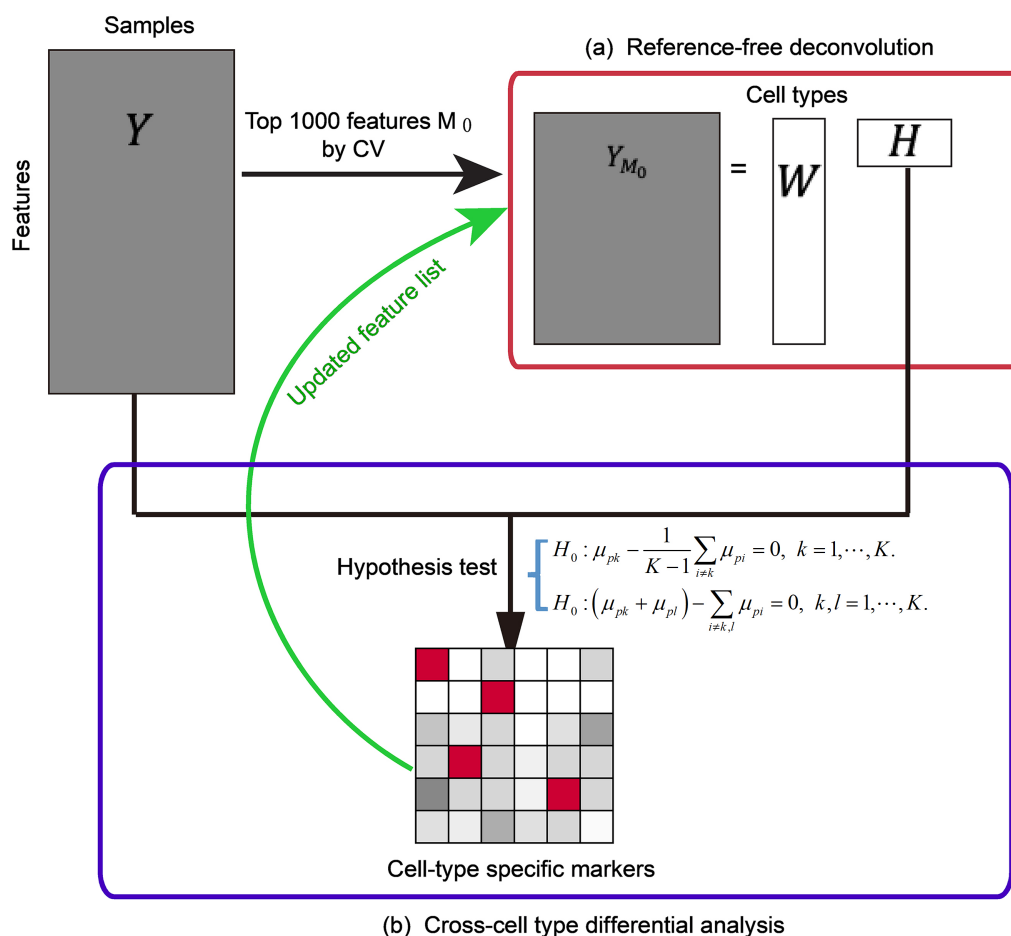


Figure 1. Workflow of the proposed algorithm DEcd

图 1. 算法 DEcd 的流程图

图 1 展示了 DEcd 的工作流程, 该流程由两个主要步骤组成: 步骤(a)从原始数据 Y 中选取变异系数最大的前 1000 个特征(M_0), 在矩阵 Y 中取 M_0 个特征, 得到矩阵 Y_{M_0} , 对 Y_{M_0} 利用 *deconf* 进行无参考反卷积, 得到比例矩阵 H 。步骤(b)利用步骤(a)估算的细胞比例矩阵 H 和原始数据矩阵 Y 进行跨细胞类型差异分析, 将一种细胞类型和其他细胞类型之间的跨细胞类型差异分析, 以及两种细胞类型与其他细胞类型之间的跨细胞类型差异分析所得到的差异基因整合起来, 将其作为新的 M_0 , 并在新一轮迭代中用于步骤(a), 该算法迭代多次后停止。根据我们的经验, 对于包含四种细胞类型和 100 个样本的基因表达数据, 30 次迭代就得到满意的结果。对于样本量较小(例如少于 50 个)或细胞类型较多(例如 6 种或更多)的研究, 则需要更多的迭代次数。在我们的软件中, 用户可以指定总迭代次数。在每次迭代中, DEcd 会计算重构

观测值 $\hat{Y}$ 与真实观测值 Y 之间的均方根误差(RMSE), 并选择对应于最小 RMSE 的迭代结果所对应的细胞组成比例矩阵 H 作为最终估计结果。

DEcd 中跨细胞类型差异分析的统计模型如下:

设所有观测数据为 $Y = [Y_1, Y_2, \dots, Y_n]^T$, μ_k 表示细胞类型 k 的表达谱。假设所有样本的细胞类型比例已知, 并将样本 s 的细胞组成比例表示为 $\theta_s = (\theta_{s1}, \theta_{s2}, \dots, \theta_{sK})$ 。因此, 观测数据 Y 可描述为一个线性模型: $E[Y] = W\beta$, 其中

$$W = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1K} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n1} & \theta_{n2} & \dots & \theta_{nK} \end{bmatrix}, \quad \beta = [\mu_1, \mu_2, \dots, \mu_K].$$

我们关注两个假设检验:

- (1) 细胞类型 k 与其他细胞类型之间的差异表达分析: $H_0: \mu_k - \frac{1}{K-1} \sum_{i \neq k} \mu_i = 0$;
- (2) 检验细胞类型 k, l 与其他细胞类型之间的差异表达分析: $H_0: (\mu_k + \mu_l) - \sum_{i \neq k, l} \mu_i = 0$ 。

第一个假设检验执行的是任一种细胞类型与其余细胞类型之间的差异分析, 第二个假设检验执行的是任两个细胞类型与其余细胞类型之间的差异分析。DEcd 分别从这两个假设检验中选择差异最显著的 500 个特征, 并且确保这两部分特征之间的重叠为 0, 然后将这两部分特征结合在一起, 在下一轮的反卷积中使用。

本文通过两个指标评估 DEcd 的性能: (1) 估计比例矩阵 $\hat{H}$ 与真实比例矩阵 H 在四种细胞类型上的平均皮尔逊相关系数 MC; (2) 估计比例矩阵 $\hat{H}$ 与真实比例矩阵 H 在四种细胞类型上的平均绝对误差 MAE。更高的 MC 值和更低的 MAE 值表明算法具有更优的性能。

3. 模拟分析及真实数据分析

模拟数据 Y 基于两个随机生成的矩阵来构造: 细胞类型特异性参考矩阵 W 和细胞组成比例矩阵 H 。比例矩阵 H 根据 Dirichlet 分布进行模拟, Y 由四种细胞类型组成时比例矩阵 H 的参数设置为(0.968, 4.706, 0.496, 0.347)。参考矩阵 W 根据基因表达综合数据库(GEO)中登录号为 GSE11058 的免疫数据集来生成。该数据集包含四种类型免疫细胞(Jurkat, LM-9, Raji, THP-1)的基因表达谱, 并且每个数据集具有来自三个样本的测量值。我们对这三个样本的基因表达值做对数变换, 计算对数变换后的每一个基因在三个样本上的均值和方差, 并用估计的均值和方差根据正态分布生成了 W 。最后, 矩阵 Y 为这两个矩阵 W 和 H 的乘积, 且加上小高斯噪声。图 2 的展示结果来自于 100 个蒙特卡罗实验结果的总结。

图 2(a)展示的是模拟结果。由于在模拟时真实的比例矩阵是生成的, 因此是已知的。我们可以计算真实比例矩阵和估计比例矩阵之间的 MC 和 MAE。更高的 MC 值和更低的 MAE 值表明算法具有更优的性能。图 2(a)展示了在迭代次数为 0 时估计比例与真实比例之间的皮尔逊相关系数(此时的估算比例是使用观测数据中变异系数最高的前 1000 个特征作为反卷积的输入)和使用 DEcd 方法进行 6 次、13 次和 20 次迭代后估计比例与真实比例在四种细胞类型上的皮尔逊相关系数。图 2(b)、图 2(c)分别显示了四种细胞类型中估计比例与真实比例之间的平均皮尔逊相关系数和平均绝对误差。“Before”时的估计细胞类型比例矩阵实际上就是迭代次数为 0 时的估计比例矩阵, 而 DEcd 的估计比例矩阵是基于 30 次迭代中 RMSE 最小的迭代对应的估计比例矩阵。显然, 在所有四种细胞类型上, 随着迭代次数的增加, 估计比例与真实比例之间的相关性总体上不断提高。与“Before”(迭代次数为 0)相比, 细胞类型比例估算精度的提升

非常显著。四种细胞类型上的平均相关性显著增加(p 值为 $3.16\text{e-}34$), 平均绝对误差显著减少(p 值为 $2.89\text{e-}34$)。

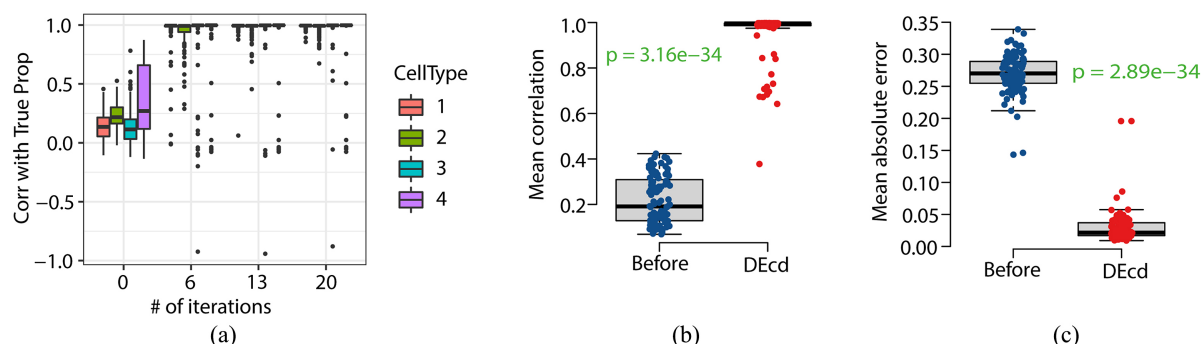


Figure 2. Performance of DEcd on synthetic mixtures. (a) Boxplots of Pearson correlations between estimated and true proportions for four cell types by number of iterations; (b) Boxplots of mean Pearson correlations between estimated and true proportions across four cell types from “Before” and DEcd; (c) Boxplots of mean absolute errors between estimated and true proportions across four cell types from “Before” and DEcd

图 2. DEcd 在合成混合物上的性能表现。(a) 不同迭代次数下, 四种细胞类型上估计比例与真实比例之间的皮尔逊相关系数的箱线图;(b) DEcd 和 “Before” 方法在四种细胞类型上的平均皮尔逊相关系数箱线图。(c) DEcd 和 “Before” 方法在四种细胞类型上的平均绝对误差箱线图

我们从基因表达综合数据库(GEO)上下载了 2 个基因表达数据集: Mouse-Mix 数据和 Immune 数据, 这两个数据具有实验算出的真实比例矩阵 H 。Mouse-Mix 数据由 3 种细胞类型组成, Immune 数据有 4 种细胞类型组成, 每一个细胞类型在每个样本上所占比例都是已知的, 即真实比例矩阵 H 是已知的。

首先, 我们利用 deconf 函数分别计算出 Mouse-Mix 和 Immune 数据的估计比例矩阵, 即图 3 中 “Before” 对应的每一个细胞类型在每一个样本上的估计比例。然后, 利用 DEcd 计算出这两个数据集上的估计比例矩阵。图 3(a)和图 3(b)显示了 Mouse-Mix 和 Immune 数据集上每种细胞类型在初始点(Before, 即迭代次数为 0)和应用 DEcd 后的估算比例与真实比例在每一个样本上的散点图, 并计算了各细胞类型的平均皮尔逊相关系数。显然, 应用 DEcd 后细胞比例矩阵估计的准确性显著提升, 平均相关系数(MC)在 Mouse-Mix 数据集上从 0.75 增加到 0.995, 在 Immune 数据集上从 0.451 增加到 0.963。

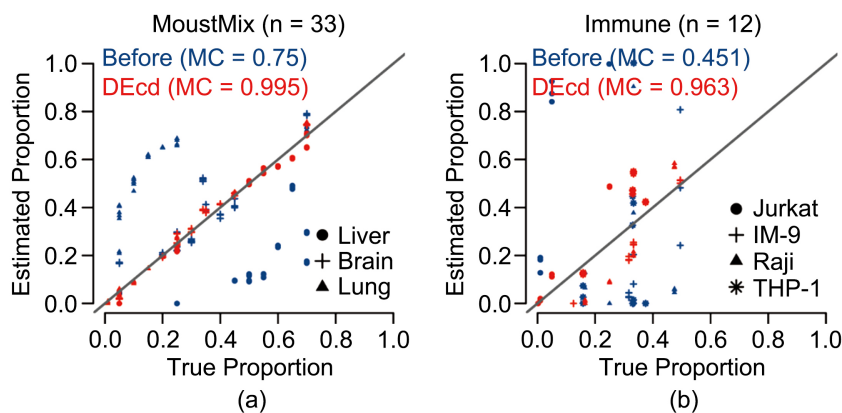


Figure 3. Scatter plots of estimated and true proportions of Mouse-Mix and Immune data for each cell type at initial point and after applying DEcd

图 3. Mouse-Mix 和 Immune 数据集上每种细胞类型在初始点和应用 DEcd 后的估计比例与真实比例的散点图

4. 结论

本文开发了一种基于最优特征选择的无参考反卷积算法 DEcd, 该算法通过跨细胞类型的差异分析, 迭代搜索细胞类型特异性的特征, 将一种细胞类型与其他细胞类型之间的差异分析和两种细胞类型与其他细胞类型之间的差异分析相结合筛选特征, 进而估算细胞组成。DEcd 完全依赖数据驱动, 不需要任何关于细胞类型或其比例的先验知识, 并且不局限于特定的无参考反卷积方法。目前, 大多数现有的不依赖参考基矩阵的反卷积算法都可以与该过程结合使用。模拟研究和两个真实数据集上的实证分析证明了 DEcd 的优越性。

基金项目

研究得到了江西省自然科学基金(20212BAB202001)的支持。

参考文献

- [1] Zhang, W., Li, Z., Wei, N., Wu, H. and Zheng, X. (2019) Detection of Differentially Methylated CpG Sites between Tumor Samples with Uneven Tumor Purities. *Bioinformatics*, **36**, 2017-2024. <https://doi.org/10.1093/bioinformatics/btz885>
- [2] Giannakopoulos, P., Herrmann, F.R., Bussi re, T., Bouras, C., K vari, E., Perl, D.P., *et al.* (2003) Tangle and Neuron Numbers, but Not Amyloid Load, Predict Cognitive Status in Alzheimer's Disease. *Neurology*, **60**, 1495-1500. <https://doi.org/10.1212/01.wnl.0000063311.58879.01>
- [3] Yang, Y., Mufson, E.J. and Herrup, K. (2003) Neuronal Cell Death Is Preceded by Cell Cycle Events at All Stages of Alzheimer's Disease. *The Journal of Neuroscience*, **23**, 2557-2563. <https://doi.org/10.1523/jneurosci.23-07-02557.2003>
- [4] Li, B., Severson, E., Pignon, J., Zhao, H., Li, T., Novak, J., *et al.* (2016) Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy. *Genome Biology*, **17**, Article No. 174. <https://doi.org/10.1186/s13059-016-1028-7>
- [5] Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., *et al.* (2019) Comprehensive Evaluation of Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology. *Bioinformatics*, **35**, i436-i445. <https://doi.org/10.1093/bioinformatics/btz363>
- [6] Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., *et al.* (2015) Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nature Methods*, **12**, 453-457. <https://doi.org/10.1038/nmeth.3337>
- [7] Teschendorff, A.E., Breeze, C.E., Zheng, S.C. and Beck, S. (2017) A Comparison of Reference-Based Algorithms for Correcting Cell-Type Heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, **18**, Article No. 105. <https://doi.org/10.1186/s12859-017-1511-5>
- [8] Hattab, M.W., Shabalin, A.A., Clark, S.L., Zhao, M., Kumar, G., Chan, R.F., *et al.* (2017) Correcting for Cell-Type Effects in DNA Methylation Studies: Reference-Based Method Outperforms Latent Variable Approaches in Empirical Studies. *Genome Biology*, **18**, Article No. 24. <https://doi.org/10.1186/s13059-017-1148-8>
- [9] Gong, T., Hartmann, N., Kohane, I.S., Brinkmann, V., Staedtler, F., Letzkus, M., *et al.* (2011) Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLOS ONE*, **6**, e27156. <https://doi.org/10.1371/journal.pone.0027156>
- [10] Houseman, E.A., Molitor, J. and Marsit, C.J. (2014) Reference-free Cell Mixture Adjustments in Analysis of DNA Methylation Data. *Bioinformatics*, **30**, 1431-1439. <https://doi.org/10.1093/bioinformatics/btu029>
- [11] Rahmani, E., Schweiger, R., Shenhav, L., Wingert, T., Hofer, I., Gabel, E., *et al.* (2018) BayesCCE: A Bayesian Framework for Estimating Cell-Type Composition from DNA Methylation without the Need for Methylation Reference. *Genome Biology*, **19**, Article No. 141. <https://doi.org/10.1186/s13059-018-1513-2>
- [12] Kang, K., Meng, Q., Shats, I., Umbach, D.M., Li, M., Li, Y., *et al.* (2019) CDSeq: A Novel Complete Deconvolution Method for Dissecting Heterogeneous Samples Using Gene Expression Data. *PLOS Computational Biology*, **15**, e1007510. <https://doi.org/10.1371/journal.pcbi.1007510>
- [13] Li, Z. and Wu, H. (2019) TOAST: Improving Reference-Free Cell Composition Estimation by Cross-Cell Type Differential Analysis. *Genome Biology*, **20**, Article No. 190. <https://doi.org/10.1186/s13059-019-1778-0>
- [14] Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., *et al.* (2017) Correcting for Cell-Type Heterogeneity in DNA Methylation: A Comprehensive Evaluation. *Nature Methods*, **14**, 218-219. <https://doi.org/10.1038/nmeth.4190>
- [15] Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T. and Marsit, C.J. (2016) Reference-Free Deconvolution of DNA Methylation Data and Mediation by Cell Composition Effects. *BMC Bioinformatics*, **17**, Article No. 259. <https://doi.org/10.1186/s12859-016-1140-4>