DOI:10.13718/j.cnki.xsxb.2020.05.016

# 多种归一化方法对 miRNA 微阵列数据的作用分析及比较①

侯丽云, 张 旭, 吴 珍

西南大学 数学与统计学院,重庆 400715

**摘要**:通过 MA 图和箱线图比较归一化前后 miRNA 微阵列数据分布情况的变化,用 K-S 检验和均方误差来评估 6 种归一化方法的优良性. 结果显示,对于 miRNA 微阵列数据而言,局部加权回归方法和分位数归一化方法比其它方法效果更好,其中又以局部加权回归方法的效果最佳.

**关 键 词**:miRNA 微阵列数据;归一化方法;MA 图;K-S 检验;均方误差

**中图分类号**:Q522 **文献标志码**:A **文章编号**:1000-5471(2020)05-0098-05

广泛存在于真核细胞中的 miRNA 是一类长度约为 18—25 个核苷酸非编码的单链 RNA 分子,且在调控基因表达、细胞周期、生物体发育等方面起重要作用[1-2]. 为了更好研究 miRNA 与癌症的关系,需要对 miRNA 微阵列数据进行统计分析,而数据归一化是进行统计分析的必要步骤. 由于 miRNA 微阵列数据存在系统误差,所以进行归一化处理的目的就是减小系统误差. 本文就与胃癌相关的 miRNA 微阵列数据比较了 6 种不同的归一化方法. 通过绘制 MA 图与箱线图来比较归一化方法对数据分布情况的影响,并且使用 K-S 检验和均方误差来综合衡量每种归一化方法.

## 1 材料与方法

本文数据取自于 NCBI(National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28700)[3]的 GEO 数据库中与胃癌有关的数据,其中包括 22 个正常样本和 22 个胃癌样本. 实验组为胃癌样本,对照组为正常样本. 把实验组 miRNA 的表达量用 $R_i(i=1,2,3,\cdots,556)$ 表示,对照组 miRNA 的表达量用 $G_i$ 表示. 把对数强度比,即 $\log_2(R_i/G_i)$,记为 $M_i$;把平均对数强度,即 $\log_2\sqrt{R_iG_i}$,记为 $A_i$. 本文在 RStudio 中进行相关操作.

我们将比较全局归一化[4]、局部加权回归方法[5]、分位数归一化[6]、修正均值归一化[7]、方差稳定归一化[8]以及尺度归一化[9]对 miRNA 微阵列数据的影响. 本文使用 MA 图来比较 6 种归一化方法对数据分布的影响. MA 图可以清楚看到系统误差的大小. 如果 MA 图中的纵坐标 $M$ 值集中分布在 $M=0$ 附近,说明数据之间的差异较小[10]. 箱线图是利用数据中的 5 个统计量:最小值、第一、四分位数、中位数、第三、四分位数与最大值来描述数据的一种方法. 我们可以从箱线图中粗略地看出数据是否具有对称性与分布的集中或离散等信息.

本文用 K-S 检验和均方误差来验证 6 种归一化方法的优良性. K-S 检验是一种拟合优度检验. K-S 统

① 收稿日期:2019-03-14
基金项目:国家自然科学基金项目(11701471),重庆市基础科学与前沿技术研究项目(cstc2017jcyjAX0476).
作者简介:侯丽云(1994—),女,硕士研究生,主要从事生物数学研究.
通信作者:张 旭,副教授.

计量的值越小，归一化效果越好[11-12]．均方误差是偏差平方与方差之和．较小的方差和偏差值表示更好的归一化，即均方误差越小，表明归一化方法的效果越好[13]．

## 2　结　　果

### 2.1　MA 图与箱线图对归一化前后数据分布的呈现及对比

首先，给出未进行归一化的原始数据的 MA 图．图 1 中水平直线表示 M 的均值是 0.14，偏离 0．因此表明对数据进行归一化处理是必要的．

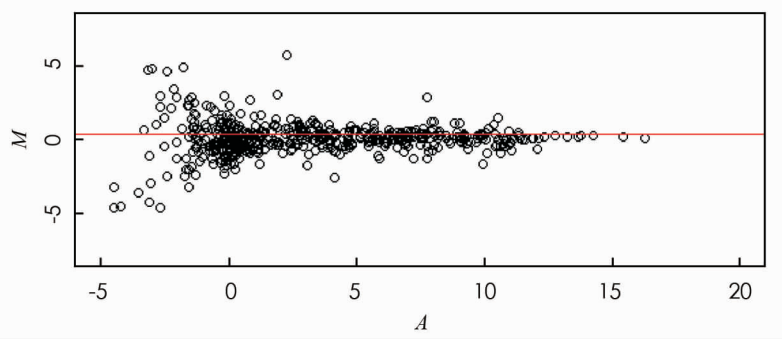


图 1　数据未进行归一化的 MA 图

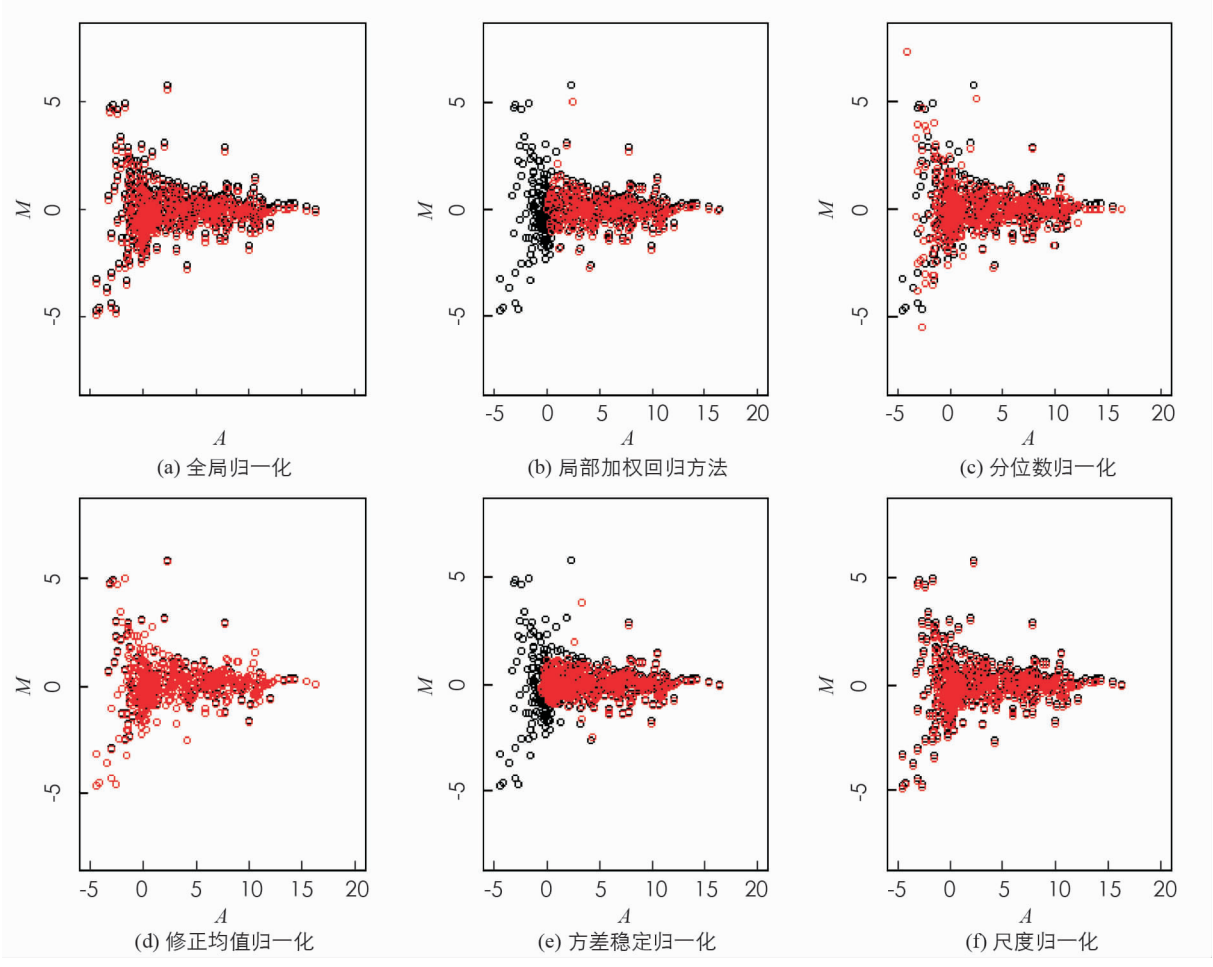


图 2　6 种归一化方法与未归一化对比的 MA 图

图 2 分别显示了通过 6 种归一化方法变换后的数据的 MA 图，其中黑色表示未归一化的数据分布图，红色表示归一化后的数据分布图．由图 2 可知：与未归一化相比，全局归一化将 $M$ 的均值变为 0；局部加权回归方法将 MA 图中的散点以 $M=0$ 为中心基本呈对称分布；分位数归一化与尺度归一化使得 $M$ 的均

值明显减小且散点分布比较对称；修正均值归一化后的MA图与未归一化的MA图无明显变化；方差稳定归一化将比较离散的点集中在0附近.

其次，利用四分位数将未归一化与归一化后的$M$值进行对比，即将$M$值分成4部分，分别用$Q_1$表示最小值与第一四分位数之间的数据(图3(a))，$Q_2$表示第一四分位数与中位数之间的数据(图3(b))，$Q_3$表示中位数与第三四分位数之间的数据(图3(c))，$Q_4$表示第三四分位数与最大值之间的数据(图3(d)). 箱线图从左到右依次是：未归一化、全局归一化、局部加权回归方法、分位数归一化、修正均值归一化、方差稳定归一化和尺度归一化. 从图3(a),(b),(c)中可以看出，归一化后和原始数据的最小值、第一四分位数、中位数、第三四分位数以及最大值之间的波动较大. 其中全局归一化方法与方差稳定归一化方法改变较大. 但是在图3(d)中，$M$值变大时，归一化后的最小值、第一四分位数、中位数、第三四分位数以及最大值均与原始数据比较接近，亦即数据越大，归一化方法对其影响越小.

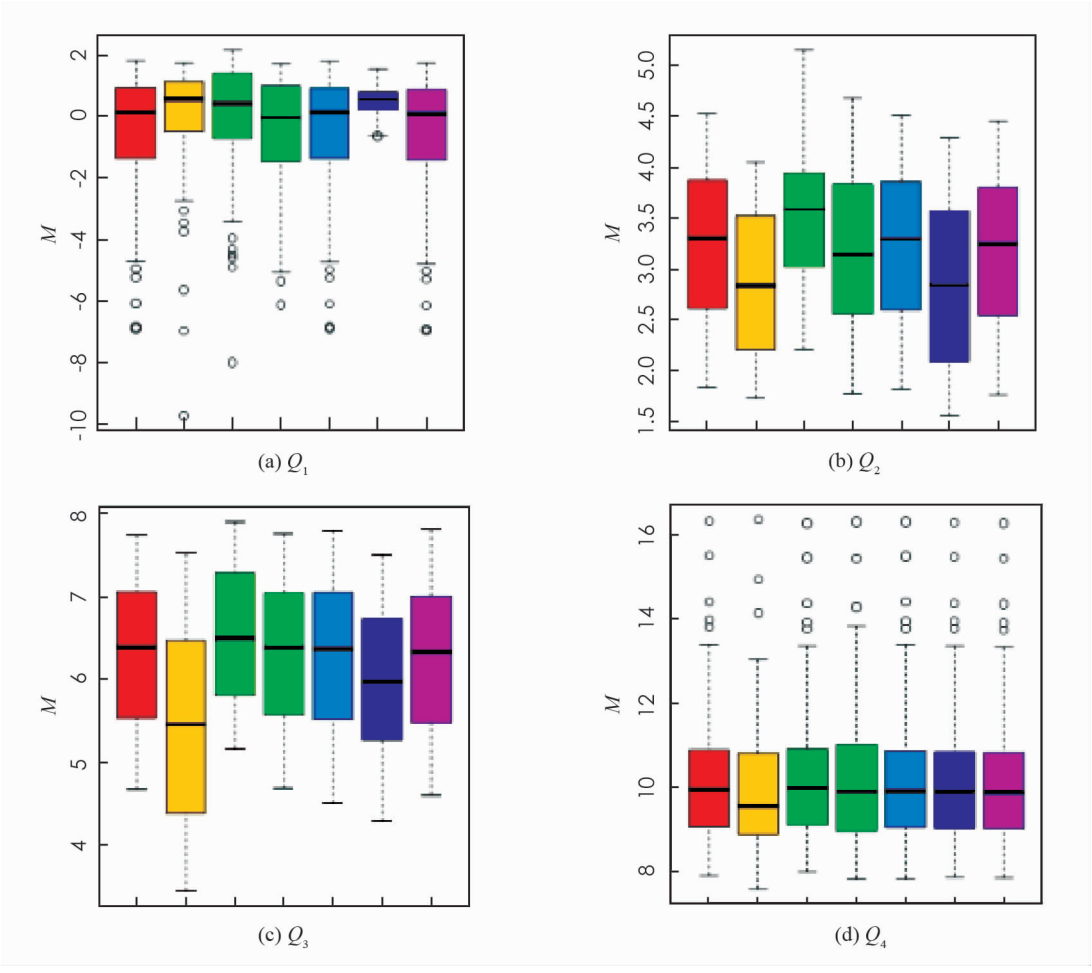


图3 归一化方法在$Q_1$,$Q_2$,$Q_3$,$Q_4$ 4个数据集上分位数对比图

## 2.2 K-S检验和均方误差对归一化方法的衡量

首先，用K-S检验来比较6种归一化方法. 对于数据GSE28700，首先计算未归一化的$M$值，再分别计算用6种归一化方法作用后的$M$值. 最后，用计算得到的$M$值进行K-S检验，得出K-S统计量值. 由图4可知，方差稳定归一化方法产生了较大的K-S统计量值. 全局归一化、修正均值归一化与尺度归一化方法产生了低于未归一化的K-S统计量值，而分位数归一化与局部加权回归方法的K-S统计量值最低. 从K-S检验的结果来看，局部加权回归方法和分位数归一化方法的K-S统计量值较小，但是相对于全局归一化、修正均值归一化以及尺度归一化这3种方法而言，K-S统计量值仅有稍微减少，优势并不显著. 因此，下面采用均方误差来进一步衡量6种归一化方法.

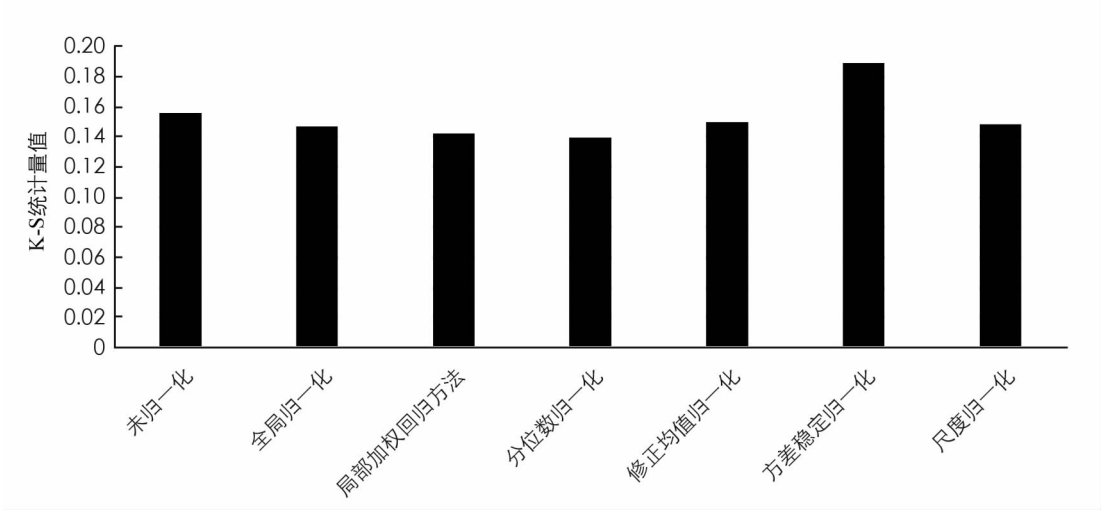


图 4　归一化前后 $M$ 值的 K-S 统计量值的对比

均方误差可以衡量不同归一化方法处理后的数据变化程度. 从图 5 可知：通过归一化方法变换之后，全局归一化方法消除了偏差，并且均方误差以及方差都有稍微的减小. 全局归一化、修正均值归一化和尺度归一化方法产生了与未归一化相近的均方误差，而修正均值归一化方法还有较大的偏差. 局部加权回归方法、分位数归一化以及方差稳定归一化方法与前面的 3 种相比，均方误差及方差值都明显减小，且偏差值都接近于 0，其中局部加权回归方法产生了最小的均方误差及方差值. 因此，从均方误差的比较结果来看，局部加权回归方法的归一化效果最好.

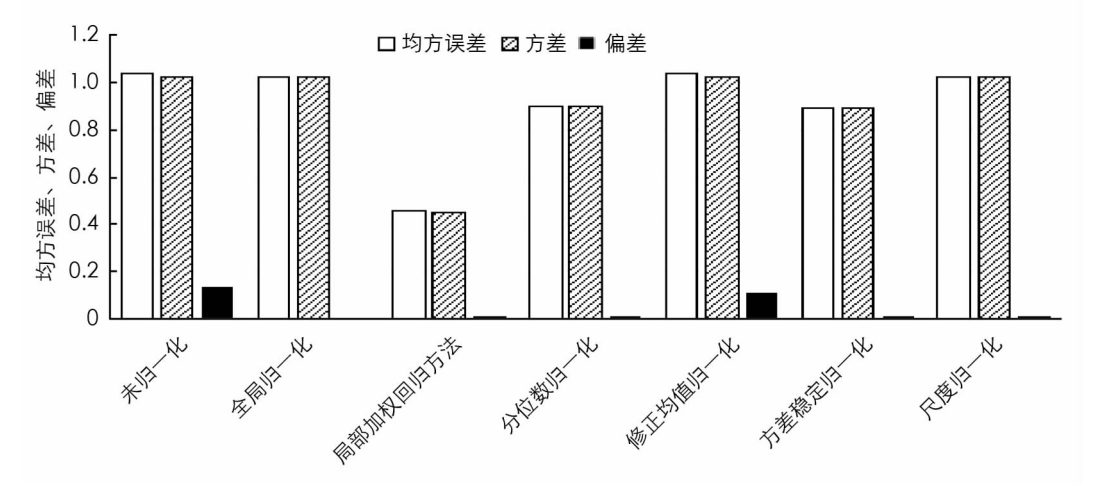


图 5　6 种归一化方法的均方误差、方差以及偏差的比较

综合上述两种归一化方法的衡量准则可知，最适合 miRNA 微阵列数据的归一化方法为局部加权回归方法.

## 3　讨　　论

数据归一化是 miRNA 微阵列数据分析中的一个关键步骤，而且 miRNA 微阵列数据对归一化方法的选择可能与 miRNA 表达的特点有关. 为了探究适合 miRNA 微阵列数据的归一化方法，我们以与胃癌相关的数据 GSE28700 为例，比较了 6 种归一化方法对数据的影响. 本文使用 MA 图和箱线图分别来比较归一化前后的数据分布情况，还使用了 K-S 检验和均方误差来衡量不同归一化方法的优良性. 综合比较 6 种归一化方法的 K-S 统计量值和均方误差值发现：对于 miRNA 微阵列数据，局部加权回归方法的归一化效果最好，其次是分位数归一化方法.

**参考文献:**

[1] PRADERVAND S, WEBER J, THOMASJ, et al. Impact of Normalization on miRNA Microarray Expression Profiling [J]. RNA, 2009, 15(3): 493-501.

[2] 胡建刚,何俊琳,黎 刚,等. 不明原因复发性自然流产胚胎绒毛组织 MiRNAs 的表达研究 [J]. 西南大学学报(自然科学版), 2010, 32(12): 56-62.

[3] CHEN C, JUAN H. The National Center for Biotechnology Information [DB/OL]. (2018-06-12)[2019-02-20]. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE28700.

[4] SMYTH G K, YANG Y H, SPEED T. Statistical Issues in cDNA Microarray Data Analysis [M]//FunctionalGenomics. NewJersey: HumanaPress, : 111-136.

[5] CLEVELAND W S. Robust Locally Weighted Regression and Smoothing Scatterplots [J]. Journal of the American Statistical Association, 1979, 74(368): 829-836.

[6] HUBER W, VON HEYDEBRECK A, SULTMANN H, et al. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression [J]. Bioinformatics, 2002, 18(Suppl 1): S96-S104.

[7] PEREIRA M B, WALLROTH M, JONSSON V, et al. Comparison of Normalization Methods for the Analysis of Metagenomic Gene Abundance Data [J]. BMC Genomics, 2018, 19: 274.

[8] HUBER W, VON HEYDEBRECK A, SUELTMANN H, et al. Parameter Estimation for the Calibration and Variance Stabilization of Microarray Data [J]. Statistical Applications in Genetics and Molecular Biology, 2003, 2(1): 1-22.

[9] SMYTH G K, SPEED T. Normalization of cDNA Microarray Data [J]. Methods, 2003, 31(4): 265-273.

[10] QUACKENBUSH J. Microarray Data Normalization and Transformation [J]. Nature Genetics, 2002, 32(S4): 496-501.

[11] WILSOND L, BUCKLEYMJ, HELLIWELLCA, et al. New Normalization Methods forcDNA Microarray Data [J]. Bioinformatics, 2003, 19(11): 1325-1332.

[12] ZHAO Y D, WANG E N, LIU H, et al. Evaluation of Normalization Methods for Two-Channel microRNA Microarrays [J]. Journal of Translational Medicine, 2010, 8(1): 69.

[13] XIONG H L, ZHANG D P, MARTYNIUK C J, et al. Using Generalized Procrustes Analysis (GPA) for Normalization of cDNA Microarray Data [J]. BMC Bioinformatics, 2008, 9(1): 25.

# Analysis and Comparison of Various Normalization Methods on Microarray Data of MiRNA

HOU Li-yun, ZHANG Xu, WU Zhen

*School of Mathematics and Statistic, Southwest University, Chongqing 400715, China*

**Abstract**: Detecting the level of miRNA in cells with microarray has become a widely used technology. There are many normalization methods for microarray of miRNA. Different normalization methods have different effects on microarray data of miRNA. In this paper, six normalization methods for microarray data of Agilent platform have been studied, including global normalization, locally weighted regression method, quantile normalization, trimmed mean method, variance stabilizing normalization and scale normalization. And the distribution changes of miRNA microarray data have been presented and compared before and after normalization by drawing MA plots and box plots. The six normalization methods have also been evaluated by Kolmogorov-Smirnov statistic and mean square error. The result shows that the locally weighted regression method and quantile normalization method are better than other methods for miRNA microarray data, and the locally weighted regression method is the best.

**Key words**: miRNA microarray data; normalization methods; MA plot; Kolmogorov-Smirnov statistic; mean square error

**责任编辑** 张 枸