微生物学通报
Microbiology China
tongbao@im.ac.cn

May 20, 2015, 42(5): 817−825

DOI: 10.13344/j.microbiol.china.140797

研究报告

基因芯片数据预处理方法(LnMR 和 RAln)的评估和比较

郑乔舒　岳淏伟　杨云锋*

(清华大学环境学院 环境模拟与污染控制国家重点联合实验室　北京　100084)

摘　要:【目的】评估并比较两种基因芯片数据预处理方法(LnMR 和 RAln)。【方法】以西藏高寒草甸草原夏季放牧实验和中国东部农田土壤移栽与玉米种植交互作用实验的两套基因芯片数据为例，利用等级-丰度曲线、均匀度指数、单因素方差分析、Q-Q 图、α 多样性和响应比等统计方法评估预处理效果。【结果】两种方法均能有效缩小极值和信号差异，改善信号分布，减小随机误差，提高数据正态性，增强实验结果的趋势，使芯片数据更适于进一步统计分析。两种预处理方法各有特点，LnMR 法更适合检测不同处理间微生物结构差异，RAln 法可以一定程度上消除基因芯片测定的系统误差。【结论】LnMR 法和 RAln 法是两种行之有效的基因芯片预处理方法，在实际分析中研究者应根据研究需要合理选择。

关键词: 基因芯片，数据预处理，微生物功能基因，微生物生态，LnMR，RAln

Evaluation and comparison of GeoChip data pre-processing methods: LnMR and RAln

ZHENG Qiao-Shu　YUE Hao-Wei　YANG Yun-Feng*

(*State Key Joint Laboratory of Environment Simulation and Pollution Control*, *School of Environment*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: [Objective] To evaluate and compare two GeoChip data pre-processing methods, LnMR and RAln. **[Methods]** The rank-abundance curve, evenness indice, one-way ANOVA, Q-Q plot, α diversity indice and response ratio were used to evaluate the pre-processing methods of GeoChip data from two recently published studies, a summer grazing experiment in the Tibetan grassland and a field study on the mutual effects of soil transplant and maize cropping. **[Results]** Both methods are efficient in removing or diminishing extreme values, optimizing data distribution, reducing random errors, improving data normalization and manifesting experimental differences, which makes GeoChip data more suitable for further statistical analysis. In particular, LnMR is more suitable for detecting subtle differences of microbial community compositions among different treatments, whereas RAln is more efficient in removing systematic errors. **[Conclusion]** LnMR and RAln are two powerful GeoChip data pre-processing methods, and should be applied with caution.

Keywords: GeoChip, Data pre-process, Microbial functional gene, Microbial ecology, LnMR, RAln

*通讯作者：✉：yangyf@tsinghua.edu.cn

收稿日期：2014-10-16；接受日期：2014-12-29；优先数字出版日期(www.cnki.net)：2015-01-12

近年来，以高通量测序技术和基因芯片(GeoChip)技术为代表的宏基因组学(Metagenomics)技术已成为微生物生态学领域的研究热点[1]。其中，基于功能基因阵列(Functional gene array)的基因芯片技术是探究微生物群落功能组成、结构、多样性、代谢活性和动力学过程的强有力手段。它利用固定在载体表面的大量 DNA 探针(Probe)与荧光标记的目标物杂交，检测杂交信号，即可获得生物样品的功能基因信息[2]。第四代基因芯片(GeoChip 4.0)包含约 82 000 个探针，覆盖了 410 个基因家族(Gene family)的 141 995 条基因序列，可以检测目标微生物群落中与碳循环、氮循环、磷循环、硫循环、能量代谢、金属抗性/还原性、压力响应、抗生素抗性、有机污染物降解性能、噬菌体和毒性等各类功能相关的基因“丰度”[3]。基因芯片已成功应用到极端环境、污染修复、气候变化等领域的研究中[4-9]。

基因芯片等宏基因组学技术能为微生物研究提供海量数据，为探索各类环境微生物的物种和功能构成创造了可能；同时，巨大的数据量也对数据挖掘和生物信息分析方法的准确性和高效性提出更高要求。在基因芯片实验过程中，不可避免地存在各类误差。除土壤微生物群落特定功能基因含量外，样品 DNA 浓度[3,10]、杂交条件(如溶液浓度和激光激发强度)和实验操作等均会影响芯片探针与目标物杂交得到的荧光强度数值。此外，基因芯片原始数据具有以下两个特点：(1) 探针多，数据量大；(2) 各探针检测信号强度差别大，极值大。因此，在对基因芯片数据正式统计分析之前，有必要进行数据预处理，尽可能减小误差，优化数据结构，提高后续数据挖掘的准确性和效率。

对于基因芯片数据，目前存在多种预处理方法，常用的有 LnMR (Logarithmic transformation and mean ratio normalization) 和 RAln (Relative abundance normalization and logarithmic transformation)，但从未有研究给出选用这两种方法的足够理由。本研究首次系统介绍了 LnMR 法和 RAln 法，并以西藏高寒草甸草原夏季放牧实验[4,11]和我国东部农田土壤移栽与玉米种植交互作用实验[12-13]的两套土壤基因芯片数据为例，展示和比较两种方法的处理效果，论证其可行性。

1 基因芯片数据预处理方法

1.1 LnMR 法

假定实验共有 p 种处理方式，每种处理设置了 q 个重复样，即实验样品总数 $n=p\times q$；基因芯片共设置了 m 个探针。

为减小信号值，缩小信号强度差异，防止数据过大给后续计算分析造成负担，LnMR 法首先对原始数据作对数转换。对数转换是最常用的数据转换方法之一，它能使大值减小，并缩小数据差异，还能使实验结果的趋势更加明显[14-15]。将第 j 个探针在第 i 个样品中的原始信号值 S_{ij}^{0} 加上 1，再取自然对数，以保证变换结果均大于 0，得到 S_{ij}^{L}：

$$S_{ij}^{L}=\ln\left(S_{ij}^{0}+1\right)$$

实验条件的细微差别会引入误差，造成不同芯片绝对信号的差异。为提高探针信号在芯片之间的可比性，需进行数据标准化。LnMR 法采用均值标准化，即用 S_{ij}^{L} 除以其所在样品 i 的信号平均值 $\overset{m}{\underset{j=1}{Avg}}\left(S_{ij}^{L}\right)$，既减小同一样品内探针信号间方差，又便于同一探针在各芯片间的比较。计算均值时，将样品中的 0 值剔除，避免计算结果受到芯片上无信号探针的数量影响。最终得到预处理后第 j 个探针在第 i 个样品中的信号值 S_{ij} (Mean ratio)：

$$S_{ij}=\frac{S_{ij}^{L}}{\overset{m}{\underset{j=1}{Avg}}\left(S_{ij}^{L}\right)}$$

这种方法的特点是，先对探针信号作对数化处理，缩小信号峰值，后在样品内部作标准化，最大限度保持不同实验处理下芯片检测信号的差异。

1.2 RAln 法

此法先计算不同样品内各探针的检测信号在整体基因芯片数据中的相对丰度，后用对数转换缩

小信号强度。与 LnMR 法不同，RAln 法采用样品内部的总和标准化，同时考虑了不同样品间信号强度差异，可在一定程度上消除系统误差的影响。

首先，将第 j 个探针在第 i 个样品中的原始信号值 S_{ij}^0 除以样品 i 的探针信号总和，再乘以各样品探针信号总和的均值，得到相对丰度 RA_{ij}：

$$RA_{ij}=\frac{S_{ij}^0}{\sum_{j=1}^{m}S_{ij}^0}\times \overset{n}{\underset{i=1}{Avg}}\left(\sum_{j=1}^{m}S_{ij}^0\right)$$

然后对 RA_{ij} 作加 1 取对数转化，最终信号强度 S_{ij} 可表示为：

$$S_{ij}=\ln\left(RA_{ij}+1\right)$$

2 数据预处理效果分析

研究依据两套已发表的真实实验数据评估 LnMR 法和 RAln 法处理效果。西藏高寒草甸草原实验研究了 3 600 m 和 3 800 m 海拔处夏季放牧对土壤微生物功能结构的影响，3 600 g 和 3 800 g 表示放牧处理，3 600 c 和 3 800 c 代表相应的对照。在农田土壤移栽与玉米种植实验中，研究者将中国南方红壤土置换到中部和北部地区以模拟气候变冷，通过种植玉米模拟人类农业活动，S、SC 和 SN 分别表示原位、置换到中部和北部的红壤样品，m 表示种植玉米。

2.1 数据结构

由于基因芯片具有数据量大、信号差异大等特点，很难用几个描述性统计量就完整概括其数据分布；为清晰观察预处理操作对芯片信号数值的影响，现采用等级-丰度曲线(Rank-abundance curve)来描绘数据结构。等级-丰度曲线是将物种相对丰度(在研究中即为芯片探针信号值)对物种相对丰度的降序排列作图，相对丰度最大的物种记为次序 1，相对丰度次大物种排序为 2，以此类推。等级-丰度曲线可同时反映样品所含物种的丰富度和均匀度两方面特性，曲线在横轴上延伸越长，表明样品的物种组成越丰富；曲线趋势越平缓，表明样品内物种均匀度越高。利用等级-丰度曲线，能直观看出 LnMR 和 RAln 两种方法分别处理后，芯片数据整体结构的变化(图 1)。原始的芯片数据具有极大值大、较小值分布较均匀的特点。图 1 显示，两个实验的原始数据信号强度最大值都超过 1.2×10^5，但大部分探针检测信号在 10^3−10^4 之间。经过预处理后，等级丰度曲线整体趋势不变，信号极值显著减小，强度分布更加集中，数据均匀度升高。LnMR 法处理后，信号极值约 1.4，数据值集中在 0.8−1.2 之间；RAln 法处理后芯片信号值集中在 6−10，极大值不超过 13。比较原始和经过预处理后芯片数据的均匀度(表 1)也发现，增加预处理使得功能基因信号的 Pielou 均匀度和 Simpson 均匀度均显著提高(P<0.000 1)。此处采用双样本等方差的双尾 t 检验(Two-tailed Student's t-test)评估两类样本差异的显著性。

在统计分析中，极值过大往往会掩盖真实信息，对数据挖掘有负面作用：(1) 增加误差偏差，降低统计检验功效；(2) 降低数据正态性，改变犯第一类(False negative)和第二类(False positive)错误的概率；(3) 严重影响目标参数的估计[16-17]。芯片原始数据中，个别探针信号强度十几倍于平均值，无疑将"虚高"某些功能基因的相对丰度，不利于得出正确的研究结论。预处理后，芯片数据极值大幅缩小，均匀度显著提高，说明数据结构得到优化，能更好反映土壤微生物群落功能的真实情况。

值得注意的是，由于 LnMR 或 RAln 法不涉及原始数据剔除，因此两种方法对样品的功能基因丰富度均没有影响，预处理后的等级-丰度曲线在 x 轴上的延展长度不变。

2.2 误差分析

减小误差是基因芯片数据预处理的主要目标之一。前文已经提到，基因芯片各探针的信号强度差异很大，其原因是多方面的，有的差异是由实验条件差异引起(如是否夏季放牧)，称作处理效应；有的是实验过程中偶然因素干扰或操作误差所致，称作实验误差。为估计数据集中实验误差的大小，考虑统计量 S_T：

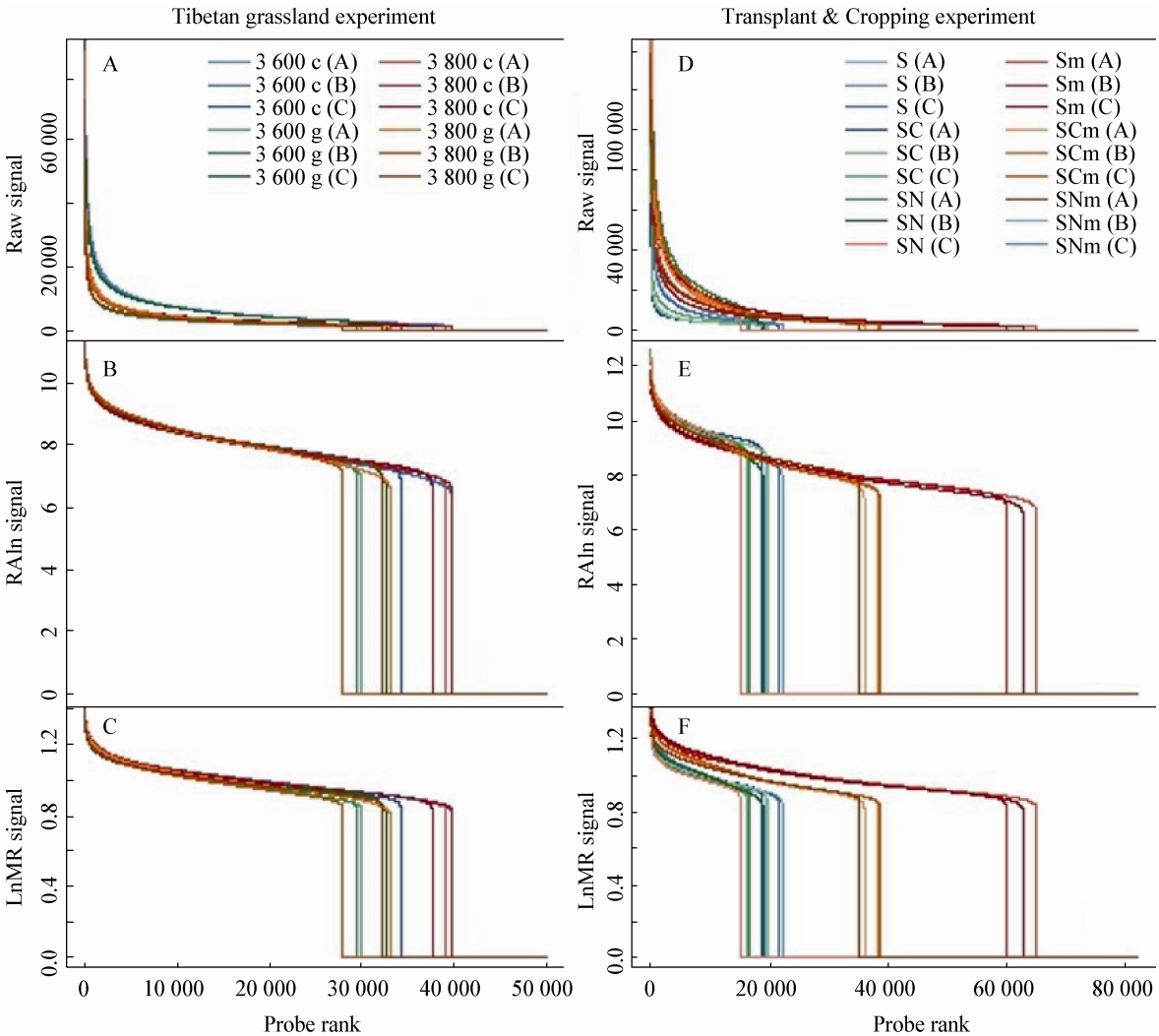

图 1 不同方法预处理后基因芯片信号的等级-丰度分布

Figure 1 The rank-abundance distribution of raw GeoChip signal using the LnMR or RAln pro-processing method

注：A、D：基因芯片原始信号；B、E：RAln 信号；C、F：LnMR 信号.

Note: A, D: Raw GeoChip signal; B, E: RAln signal; C, F: LnMR signal.

表 1 不同预处理方法对芯片数据均匀度的影响

Table 1 The evenness indice of GeoChip data with the LnMR or RAln pre-processing method

不同方法 Different methods	西藏高寒草甸草原实验 Tibetan grassland experiment		土壤移栽与玉米种植实验 Transplant & Cropping experiment	
	Pielou evenness	Simpson evenness	Pielou evenness	Simpson evenness
Raw data	0.966 9±0.002 9	0.447 1±0.022 4	0.954 0±0.009 7	0.462 8±0.069 7
LnMR	0.999 7±0	0.992 8±0.000 9	0.999 6±0.000 1	0.993 3±0.002 1
RAln	0.999 7±0.000 1	0.992 8±0.001 2	0.999 6±0.000 1	0.993 3±0.002 7

$$S_T=\sum_{i=1}^{p}\sum_{j=1}^{r}(x_{ij}-\bar{x})^2\ ,\ \ \bar{x}=\frac{1}{p\times r}\sum_{i=1}^{p}\sum_{j=1}^{r}x_{ij}$$

S_T 是所有探针信号 x_{ij} 与总平均值 $\bar{x}$ 之差的平方和，反映了芯片数据整体离散程度，称之为总离差平方和。共有 p 种实验因素，r 代表同一实验因素下芯片探针总数，为探针数 m 与重复样品数 q 的乘积 $r=q\times m$。S_T 可以分解为误差平方和 S_E 和效应平方和 S_A：

$$S_T=S_E+S_A$$

其中，S_E 是同种实验处理样品内部数据变动平方和的总和，代表了随机误差的影响；S_A 是在各实验因素 A_i 水平下样品均值 $\bar{x}_i$ 与总体均值 $\bar{x}$ 差异的总和，代表了实验变量的影响。它们的计算式分别为：

$$S_E=\sum_{i=1}^{p}\sum_{j=1}^{r}(x_{ij}-\bar{x}_i)^2\ ,\ \ \bar{x}_i=\frac{1}{r}\sum_{j=1}^{r}x_{ij}$$

$$S_A=\sum_{i=1}^{p}\sum_{j=1}^{r}(\bar{x}_i-\bar{x})^2=\sum_{i=1}^{p}r(\bar{x}_i-\bar{x})^2$$

可以证明[15,18]，S_E 和 S_A 除以各自自由度后所得均方的比值服从自由度为 $p-1$ 和 $p\times r-p$ 的 F 分布：

$$F=\frac{S_A/(p-1)}{S_E/(p\times r-p)}=\frac{MS_A}{MS_E}$$

对于同一实验，自由度只与实验条件数 p 和相同条件下探针总数 r 有关，数据是否经过预处理对自由度没有影响。由 F 的计算公式可知，F 值越大，效应误差与随机误差的比值越大，表明由实验条件差别造成的数据波动对数据偏差总和的影响越大。反之，F 值越小，说明随机误差对芯片信号总离散度的贡献越大，换言之，数据中蕴含的实验误差越大。计算芯片数据的统计量 F 其实就是进行单因素方差分析(One-way analysis of variance)的过程[18]。

计算发现，LnMR 法或 RAln 法数据拥有比原始数据更大的 F 值(表 2)，可见预处理操作能够降低基因芯片数据中的随机误差，增加效应误差占总体偏差的比重，提高实验结果的准确性和后续分析的有效性。

表 2 单因素方差分析的统计量 F
Table 2 The F-test of one-way ANOVA

不同方法 Different methods	西藏高寒草甸草原实验 Tibetan grassland experiment $F=MS_A/MS_E$	土壤移栽与玉米种植实验 Transplant & Cropping experiment $F=MS_A/MS_E$
Raw data	374 7	894.1
LnMR	539 2	5 485.0
RAln	403 6	3 944.0

2.3 数据正态性

正态分布是许多统计分析方法的理论基础，常用的分析方法如 t-检验、Pearson 相关分析和除趋势对应分析(Detrended correspondence analysis)等均要求分析指标服从正态或近似正态分布。然而，由于技术特点所限，芯片原始数据往往很难满足正态性的要求；因此，我们进行数据预处理的另一个目标是将基因芯片数据转换为正态或近似正态数据。

Q-Q 图常用来评估两个分布的一致性。以正态分布的 α 分位点为横坐标，以样品 α 分位点为纵坐标，作散点图；如果点分布在一条直线附近，说明这组数据来自某个正态分布总体。对原始、LnMR 法和 RAln 法预处理后的芯片数据作 Q-Q 图(图 2)，发现原始数据的 Q-Q 点大部分偏离直线，而经过预处理后，Q-Q 点明显趋近直线，这表明基因芯片的原始信号不满足正态分布，而 LnMR 或 RAln 变换显著提升了数据的正态性。

2.4 实验结果分析

以上分析可以看出，LnMR 和 RAln 两种方法的主要优点在于：通过调整数据结构，消除或者减少极端值的影响，使数据分布更合理。但是，如果数据预处理后的分析结果与原始数据的结果差别很大，也会引起数据处理是否正确的疑虑。

研究采用 α 多样性作为微生物群落结构评价指标，以 α 多样性在各实验条件下的变化表征微生物群落整体结构随实验变量的变化趋势。计算不同方法预处理后芯片数据的 Shannon 多样性和 Simpson

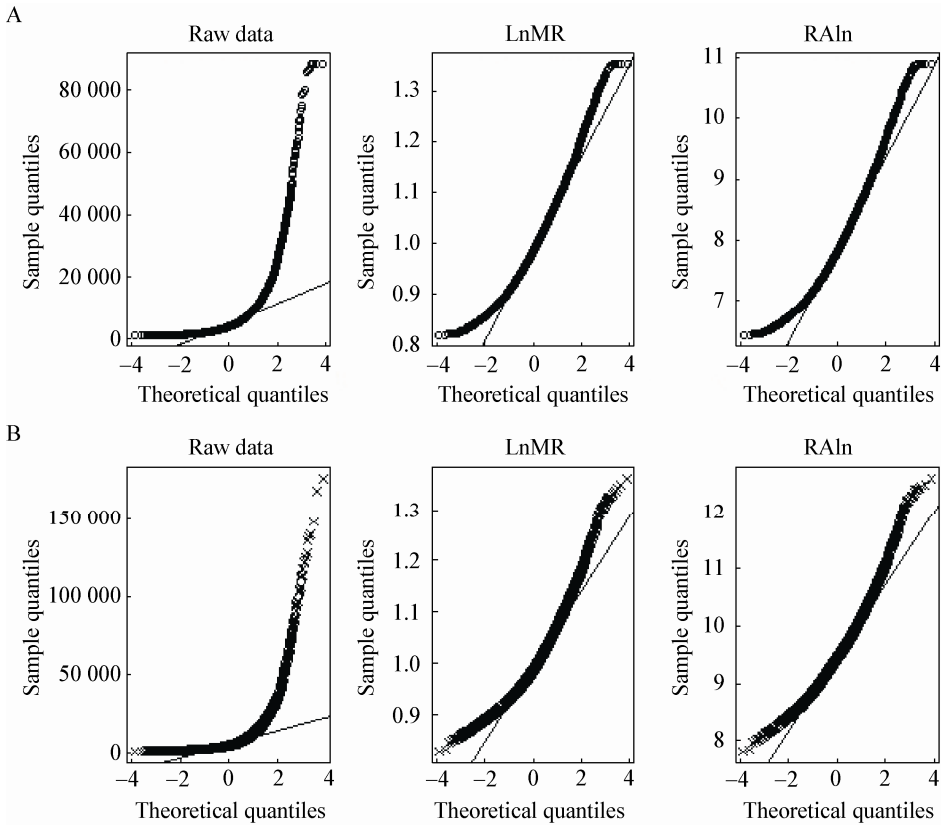

图 2 原始数据、LnMR 法和 RAln 法处理后芯片数据的正态性 Q-Q 图

Figure 2 Q-Q plots of raw GeoChip data and data with the LnMR and RAln pre-processing methods

注：A：西藏高寒草甸草原实验；B：土壤移栽与玉米种植实验.

Note: A: The Tibetan grassland experiment; B: The transplant & cropping experiment.

多样性指数(表 3)发现，西藏草原实验的原始、LnMR 法和 RAln 法数据均呈现出放牧组比对照组微生物群落多样性显著(P<0.05)或适度(P<0.1)降低的趋势；在农田土壤移栽与玉米种植实验中，种植玉米显著(P<0.01)提升了原始、LnMR 法和 RAln 法数据的多样性指数，而土壤中移和北移处理使三类数据的多样性整体降低。

在功能基因水平上，三类数据对实验处理的响应也是一致的。响应比(Response ratio)常用来量化环境变量对微生物介导的各类功能(如碳、氮循环等)的影响程度。它是实验组和对照组数据的均值之比，由两组数据的标准差计算偏差，并根据需要的显著性水平 α 求出置信区间；若置信区间完全落在坐标零点一侧，表示响应变化显著。取置信度 95%，如图 3 所示，夏季放牧降低土壤微生物基因丰度、种植玉米使微生物基因丰度增加的实验结果不因预处理操作改变。

以上分析证明了两种预处理方法能较好保持芯片信号随实验变量变化的整体趋势，保留原始数据中的目的信息，保证了研究结论的可靠性，对预处理操作会改变实验结论的担心是多余的。

在保持芯片信号随实验条件变化趋势的同时，数据预处理能够“强化”微生物群落数据对实验因子的响应。无论是放牧实验还是玉米种植实验(图 3)，数据预处理后微生物各类功能基因响应比的标准差明显缩小，置信区间完全落在零点一侧(西藏草原实验中 Bacteria phage 除外)，说明各类功能基因变化的程度更大、变化方向更集中。微生物群落 Shannon

表 3 不同预处理方法对芯片数据 Shannon 多样性和 Simpson 多样性的影响
Table 3 The Shannon index and Simpson index of GeoChip data with the LnMR or RAln pre-processing method

不同方法 Different methods		Shannon index (H')			Simpson's diversity index ($1/D$)		
		Raw data	LnMR	RAln	Raw data	LnMR	RAln
西藏高寒草甸草原实验 Tibetan grassland experiment	3 600 c	10.179 1 ±0.048 1	10.536 6 ±0.087 0	10.536 3 ±0.086 7	16 635.17 ±1 159.64	37 621.01 ±3 163.45	37 597.88 ±3 139.38
	3 600 g	9.992 5 ±0.080 5	10.326 7 ±0.057 4	10.326 9 ±0.057 0	13 769.15 ±1 735.94	30 467.39 ±1 785.59	30 473.99 ±1 764.83
	P value	0.026	0.025	0.025	0.076	0.027	0.027
	3 800 c	10.219 6 ±0.006 4	10.563 3 ±0.028 0	10.563 4 ±0.028 0	17 293.51 ±306.46	38 553.13 ±1 057.08	38 558.28 ±1 059.20
	3 800 g	10.001 6 ±0.094 6	10.335 6 ±0.092 0	10.335 8 ±0.091 9	14 120.44 ±1 541.81	30 781.35 ±2 753.33	30 794.43 ±2 752.70
	P value	0.016	0.015	0.015	0.025	0.01	0.01
土壤移栽与玉米种植实验 Transplant & Cropping experiment	S	7.272 6 ±0.059 8	7.649 0 ±0.066 6	7.649 3 ±0.066 5	911.57 ±70.94	2 094.50 ±136.65	2 095.87 ±136.58
	SC	7.260 0 ±0.105 1	7.500 3 ±0.101 2	7.500 8 ±0.101 3	963.22 ±117.65	1 810.95 ±178.83	1 812.75 ±179.22
	SN	7.151 0 ±0.097 4	7.427 2 ±0.118 0	7.427 2 ±0.118 0	935.39 ±120.68	1 684.97 ±198.07	1 684.79 ±198.10
	Sm	8.253 9 ±0.045 6	8.736 4 ±0.041 5	8.736 0 ±0.041 4	2 334.03 ±164.11	6 199.31 ±255.76	6 194.69 ±253.85
	SCm	7.763 0 ±0.029 1	8.228 2 ±0.038 1	8.227 9 ±0.038 1	1 543.69 ±24.16	3 729.68 ±140.66	3 727.40 ±140.04
	SNm	7.847 9 ±0.074 1	8.206 3 ±0.070 5	8.206 2 ±0.070 4	1 761.13 ±140.49	3 657.64 ±262.52	3 656.51 ±262.14
	S vs SC *P*	0.866	0.101	0.101	0.550	0.095	0.095
	S vs SN *P*	0.139	0.047	0.047	0.783	0.042	0.042
	S vs Sm *P*	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	SC vs SCm *P*	0.001	<0.001	<0.001	0.001	<0.001	<0.001
	SN vs SNm *P*	0.001	0.001	0.001	0.002	<0.001	<0.001

注：采用双样本等方差的双尾 *t*-检验评估样本差异的显著性.

Note: The significance of sample difference was measured by two-tailed Student's *t*-test (two sample equal variance).

多样性和 Simpson 多样性也呈现相似规律(表 3)：两种预处理数据随实验因子变化的趋势与原始数据一致，而显著性检验 *P* 值更小，说明 LnMR 法和 RAln 法都能扩大不同实验条件下芯片信号间的差异。可见，预处理确实能在一定程度上放大差异，突出微生物功能结构对环境因子变化的响应，使实验结果趋势更明显，有利于基因芯片数据挖掘。

2.5 两种方法比较

LnMR 和 RAln 两种方法的区别在于，RAln 法做数据标准化时平衡了不同实验条件下样品的芯片信号，减小由于分批测定等原因引入的基因芯片数据误差；而 LnMR 法则对各类样品分别进行标准化，保留不同实验条件下芯片检测信号的差异，以便研究者更容易观察到各实验变量对微生物功能结构的影响。换句话说，LnMR 法处理得到的数据对环境因子更加敏感。这一点可通过微生物各类功能基因对环境变量的响应比证明(图 3)。图 3 中，LnMR 法和 RAln 法处理后各功能基因信号响应变化趋势基本一致，但 LnMR 法的响应比总是比 RAln 法的响应比偏离零值更远，说明 LnMR 数据对环境因子(放牧或种植玉米)的响应更加显著。

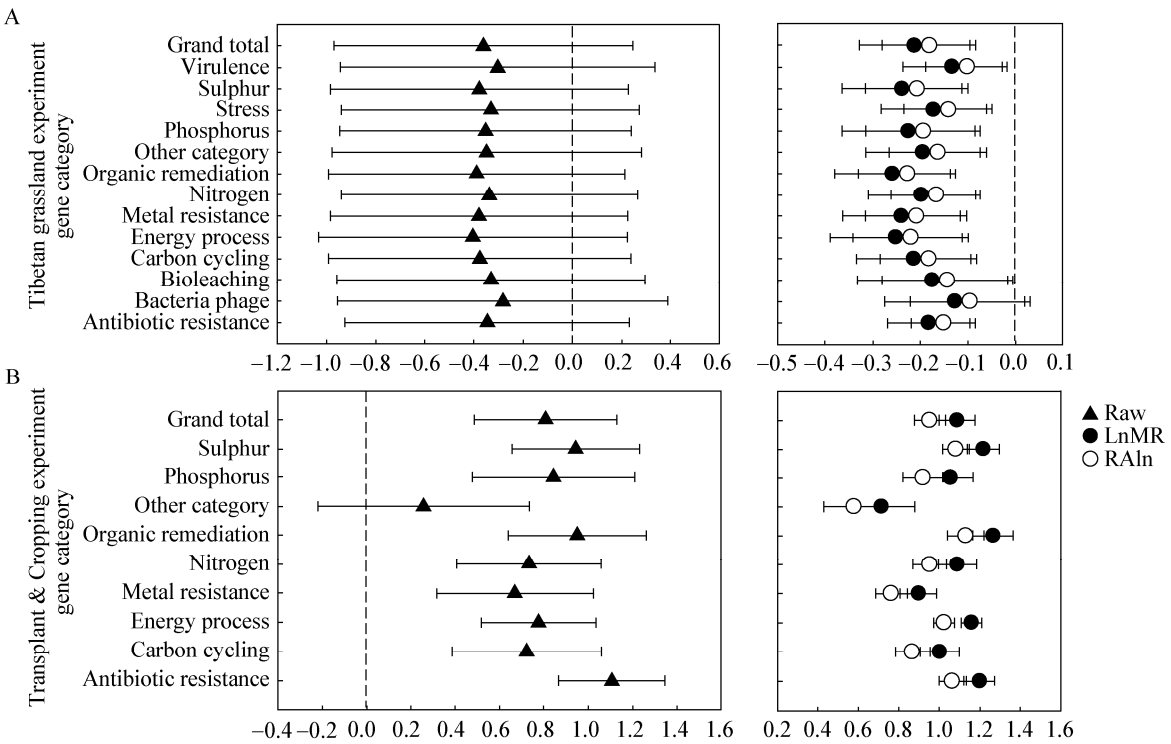

图 3 微生物各类功能基因信号的响应比(置信度 95%)

Figure 3 Response ratios of all of detected gene categories at the 95% confidence level

注：A：西藏草原 3 600 m 海拔处放牧实验；B：南方红壤原位玉米种植实验.

Note: A: The Tibetan grassland samples at the 3 600 m above sea level; B: The red soil samples from Southern China.

表 4 使用 LnMR 法和 RAln 法数据的注意事项及应用案例

Table 4 Tips in using LnMR and RAln and their applications

注意事项及应用案例 Tips and applications	LnMR	RAln
分析各类功能基因丰度 Functional gene abundance	用编码相同功能的探针的信号均值	用编码相同功能的探针的信号总和
其他注意事项 Other tips	计算各探针信号与样本均值之比时，需将 0 值剔除	如果后续统计分析对数据信号大小和正态性没有要求，预处理时可不作对数转换
应用案例 Application	气候变冷与种植玉米对红壤微生物群落的交互作用研究[12]；土壤微生物群落对大气 CO_2 浓度上升的响应研究[7]	微生物多样性沿海拔梯度研究[19]；海拔和放牧因素对西藏高寒草甸草原土壤微生物群落的影响研究[4]；草原和灌丛微生物群落功能结构对比[20]

由于 LnMR 法和 RAln 法计算原理的差异，芯片数据在使用两种方法预处理后，需要注意各自的后续分析要求。经 LnMR 法处理的芯片信号值在 1 左右，如果分析具体功能基因时把同一功能类别的基因探针按样品加和，所得信号丰度将与检测到的探针数目相近，从而浪费了探针杂交信号强度的信息。因此，利用 LnMR 法处理的芯片数据分析各类功能基因时，必须使用编码相同功能的探针的信号均值，计算时剔除 0 值。RAln 法则不存在这一问题，分析中可以使用探针信号总和。此外，还有一些注

意事项，详见表 4。在实际应用中，研究人员可参考此处列举的案例，结合自身实验目的和设计等实际情况，合理选择预处理方法。

3 结论

上述讨论说明，LnMR 和 RAln 两种方法均能够在保持数据随实验条件变化趋势的同时，有效缩小极值和信号差异，改善信号分布，减小随机误差，提高数据正态性，增强实验结果的趋势，使芯片数据更适于进一步统计分析，是行之有效的基因芯片数据预处理手段。

两种预处理方法各有特点，LnMR 法更适合检测不同处理间微生物结构差异，RAln 法可以在一定程度上消除基因芯片测定的系统误差。在实际分析中，研究者应根据自己的研究目的和实验特点合理选择。

参 考 文 献

[1] Sun X, Gao Y, Yang YF. Recent advancement in microbial environmental research using metagenomics tools[J]. Biodiversity Science, 2013, 21(4): 393-400 (in Chinese)
孙欣, 高莹, 杨云锋. 环境微生物的宏基因组学研究新进展[J]. 生物多样性, 2013, 21(4): 393-400

[2] He Z, Gentry TJ, Schadt CW, et al. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes[J]. The International Society for Microbial Ecology Journal, 2007, 1(1): 67-77

[3] Tu Q, Yu H, He Z, et al. GeoChip 4: a functional gene-array-based high-throughput environmental technology for microbial community analysis[J]. Molecular Ecology Resources, 2014, 14(5): 914-928

[4] Yang Y, Wu L, Lin Q, et al. Responses of the functional structure of soil microbial community to livestock grazing in the Tibetan alpine grassland[J]. Global Change Biology, 2013, 19(2): 637-648

[5] Wang F, Zhou H, Meng J, et al. GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(12): 4840-4845

[6] Hazen TC, Dubinsky EA, Desantis TZ, et al. Deep-sea oil plume enriches indigenous oil-degrading bacteria[J]. Science, 2010, 330(6001): 204-208

[7] He Z, Xu M, Deng Y, et al. Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO_2[J]. Ecology Letters, 2010, 13(5): 564-575

[8] Zhou J, Xue K, Xie J, et al. Microbial mediation of carbon-cycle feedbacks to climate warming[J]. Nature Climate Change, 2012, 2(2): 106-110

[9] Leigh MB, Pellizari VH, Uhlik O, et al. Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs)[J]. The International Society for Microbial Ecology Journal, 2007, 1(2): 134-148

[10] Wu L, Liu X, Schadt CW, et al. Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification[J]. Applied and Environmental Microbiology, 2006, 72(7): 4931-4941

[11] Gao Y, Wang S, Xu D, et al. GeoChip as a metagenomics tool to analyze the microbial gene diversity along an elevation gradient[J]. Genomics Data, 2014, 2: 132-134

[12] Liu S, Wang F, Xue K, et al. The interactive effects of soil transplant into colder regions and cropping on soil microbiology and biogeochemistry[J]. Environmental Microbiology, 2015, 17(3): 566-576

[13] Zhao M, Wang F, Liu S, et al. GeoChip profiling of microbial community in response to global changes simulated by soil transplant and cropping[J]. Genomics Data, 2014, 2: 166-169

[14] Zhang JT. Quantitative Ecology[M]. Beijing: Science Press, 2011: 32-35 (in Chinese)
张金屯. 数量生态学[M]. 北京: 科学出版社, 2011: 32-35

[15] Bao YK. Data Analysis Tutorials[M]. Beijing: Tsinghua University Press, 2011: 52-67,150-155 (in Chinese)
包研科. 数据分析教程[M]. 北京: 清华大学出版社, 2011: 52-67,150-155

[16] Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them)[J/OL]. Practical Assessment, Research & Evaluation, 2004[2004-03-02]. http://PAREonline.net/getvn.asp?v=9&n=6

[17] Zimmerman DW. A note on the influence of outliers on parametric and nonparametric tests[J]. The Journal of General Psychology, 1994, 121(4): 391-401

[18] Xue Y, Chen LP. Statistical Modeling and R Software[M]. Beijing: Tsinghua University Press, 2007: 336-339 (in Chinese)
薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007: 336-339

[19] Yang Y, Gao Y, Wang S, et al. The microbial gene diversity along an elevation gradient of the Tibetan grassland[J]. The International Society for Microbial Ecology Journal, 2014, 8(2): 430-440

[20] Chu H, Wang S, Yue H, et al. Contrasting soil microbial community functional structures in two major landscapes of the Tibetan alpine meadow[J]. Microbiology Open, 2014, 3(5): 585-594