PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2013, 40(12): 1256~1264
www.pibb.ac.cn

基于基因表达变异性的通路富集方法研究 *

贾晓东[1)**] 陈秀杰[1)**,***] 吴 欣[2)**,***] 徐建凯[1)**]
谭福建[1)] 刘香琼[1)] 刘 磊[1)] 杨瑞智[1)]
([1)] 哈尔滨医科大学生物信息科学与技术学院，哈尔滨 150086；[2)] 哈尔滨医科大学附属第三医院，哈尔滨 150086)

摘要 当前的通路富集方法主要是基于基因的表达差异，很少有方法从通路变异性(方差)角度对其富集分析．我们注意到用合适的统计量描述通路的变异性时，在疾病表型下一些通路的变异性有明显的上升或者下降．因此本研究假设：通路变异性程度在不同表型中存在差异．本文设计了 14 种描述通路变异性的统计量与检验方法，检测不同表型下变异性有差异的通路即富集通路，并将富集结果与文献检索结果进行比较，同时，分析不同芯片预处理方法对数据和结果的影响．研究结果表明：5 种预处理方法中，多阵列对数健壮算法(RMA)是数据预处理的最优方法；不同表型下通路的变异性程度存在差异；根据文献检索的通路结果，14 种基于变异性的通路富集方法中，以通路中各基因欧氏距离的方差做统计量进行 permutation 检验(方法 11)能有效识别显著通路，其富集结果优于基因集富集分析(GSEA)．综上所述，基于通路变异性的通路富集策略具有可行性，不仅对通路富集分析有一定的理论指导意义，而且为人类疾病研究提供新的视角．

关键词 变异性，通路，富集分析，预处理方法
学科分类号 R318.04，Q78
DOI: 10.3724/SP.J.1206.2012.00410

基因富集分析是应用生物信息数据库和统计工具，将目标基因富集到已知功能的生物学通路和模块上，从生物学角度深入分析疾病的发生及发病机理．基因富集分析的目的是筛选出两组或多组间表达水平有差异的基因集，即富集基因集，它是单个基因研究的自然扩展．

基于不同的算法，当前基因富集分析可以分为三类：单一富集分析(singular enrichment analysis, SEA)、基因集富集分析(gene set enrichment analysis, GSEA)与模块富集分析(modular enrichment analysis, MEA)[1]．SEA 是最传统和应用最广泛的富集分析方法，它在单基因分析选出差异表达基因列表的基础上，计算每个功能基因集被富集到的概率，富集 *P* 值的计算方法主要有 fisher's 精确检验、超几何分布和卡方检验等，常用的 DAVID[2]、GOStat[3]、EASE[4]等工具都属于此类富集算法．GSEA 主要用于分析两分组实验的表达谱数据，它不预先选出差异表达的基因，而是对所有基因按照两组间的 *t* 统计量或者基因与表型的相关性进行排序，然后用经过排序的所有基因列表计算每个功能基因集合的最大富集得分 MES，最后用 Kolmogorov-Smirnov-like、permutation、*Z*-score、*t*-Test 等检验方法计算每个功能集合的富集 *P* 值．GSEA 工具[5]、PAGE[6]、GeneTrail[7]等都属于此类富集分析方法．MEA 继承了 SEA 的主要思想，需要预先选定感兴趣的基因列表．但是在计算富集 *P* 值时考虑了节点间或者基因间的关系，因而此方法的优点是考虑了节点间或者基因间关系的生物学意义．Ontologizer[8]、topGO[9]、GENECODIS[10]等都属于此类富集分析算法．SEA 与 MEA 都需要预先选定感兴趣的基因列表，GSEA 富集分析则使用全部基因．这三种富集分析策略在富集分析之前通常要求进行单基因分

* 黑龙江省研究生创新基金资助项目(YJSCX2012-205HLJ，YJSCX2011-341HLJ，YJSCX2012-223HLJ)，黑龙江省教育厅资助项目(11541121，12531227)，哈尔滨市科技局优秀学科带头人资助项目(2010RFXXS053)，黑龙江省自然科学基金资助项目(QC2012C010)，黑龙江省卫生厅科研课题资助项目(2012-798).
** 共同第一作者.
*** 通讯联系人.
陈秀杰. Tel: 0451-86605922, E-mail: chenxiujie@ems.hrbmu.edu.cn
吴 欣. Tel: 0451-86298703, E-mail: wuxin@ems.hrbmu.edu.cn
收稿日期：2013-02-04，接受日期：2013-03-28

析，对基因进行 t 检验找出差异表达的基因，或者按照 t 统计量对各个基因进行排序等，这些单基因分析主要是从差异表达基因角度出发的.

2008 年，Ho 等[11]提出差异变异基因(differential variability(DV) gene)的概念，即在不同表型(疾病与正常)下基因表达值的方差存在差异的基因，并阐述了 DV 基因的生物学基础，而且将 DV 基因成功应用于人类疾病的研究中．Ho 等认为，一个基因的表达受一系列调控因子(如转录因子)的激活或抑制，由于调控因子的活性在个体间存在差异，使得调控因子下游的基因响应能力存在个体差异，从而导致基因在个体间的表达出现变异．因此，我们认为多个功能相互联系的基因构成了通路，而通路内多个基因的表达变异的累积便构成了整个通路的表达变异．现有的富集分析方法中罕有从变异性角度出发进行功能富集分析的．在本文中我们假设不同表型下通路的表达变异性(方差)存在差异，并提出一种基于变异性的有效通路富集方法．本研究是基于通路中所有基因，所以我们将结果与经典的 GSEA[5]方法进行了比较．在我们设计的 14 种基于通路变异性的富集方法中，跃登指数和 ROC 曲线下面积均表明第 11 种方法有明显的优越性，其富集结果优于 GSEA.

本研究从一种全新的角度来研究通路在不同表型中的变异性程度．为寻找不同表型中存在差异的通路提供新的方法，对通路富集分析有一定的理论指导意义.

1　数据与方法

1.1　数据

1.1.1　表达谱数据.

所选数据是 Stearman 等[12](2005 年)提交的人类 Pulmonary adenocarcinoma 芯片数据，GEO 编号 GDS1650, 平台为 GPL8300 [HG-U95AV2]Affymetrix Human Genome U95 Version 2 Array，含有 12 625 个探针．GDS1650 包含 20 个肿瘤组织样本和 19 个正常组织样本．GDS1650 原始 cel 文件分别用 RMA [13]、MAS5.0 [14]、GCRMA [15]、FARMS [16] 和 DFW[17] 5 种预处理方法处理，然后对预处理结果按照 Affymetrix 芯片设计原理进行探针合并．Cel 文件的预处理使用 R 工具实现，5 种预处理方法都采用默认参数，而且都对表达值做对数处理(底数为 2).

采用不同的预处理方法处理 Affymetrix 公司的 cel 文件得到的表达谱数据会有差别，进而影响富集分析的结果．为了解决这个问题，首先我们基于 5 种预处理方法处理后的 GDS1650 芯片数据评价预处理方法，其次由于 GSEA 是目前公认的全基因组富集分析算法，所以我们用 GSEA 富集结果的重合性进一步评价这 5 种预处理方法.

1.1.2　通路数据.

从 KEGG[18]上选择有节点连接关系的 169 个大通路数据，并下载其 XML 格式，从中分别提取每个通路所包含的基因 ID，并去除冗余基因 ID．这样得到 169 个只包含基因 ID 的大通路数据.

根据 **1.1.1** 所得到的表达谱数据，找到通路中每个基因 ID 所对应的表达谱数据，把同一个通路的所有表达谱数据写到同一个文件中，最后得到 169 个通路表达谱数据．对 5 种预处理方法的表达谱数据都进行此操作.

1.2　方法

1.2.1　设计统计量与定义通路变异性.

通路表达谱数据是由通路中各个基因在不同样本条件下的表达值组成的，我们设通路中基因个数为 p，样本由疾病类样本和正常类样本组成，令 i 代表表型：疾病或者正常，设表型 i 有 n 个样本．通路表达谱数据 X 为：

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1g} & \cdots & x_{1n} & x_{1(n+1)} & \cdots & x_{1(n+m)} \\ x_{21} & \cdots & x_{2g} & \cdots & x_{2n} & x_{2(n+1)} & \cdots & x_{2(n+m)} \\ & \vdots & & \vdots & & & \vdots & \\ x_{c1} & \cdots & x_{cg} & \cdots & x_{cn} & x_{c(n+1)} & \cdots & x_{c(n+m)} \\ & \vdots & & \vdots & & & \vdots & \\ x_{p1} & \cdots & x_{pg} & \cdots & x_{pn} & x_{p(n+1)} & \cdots & x_{p(n+m)} \end{pmatrix}$$

我们分别以样本和基因为单位设计描述通路变异性的统计量，以样本为单位是指对样本中所有基因表达值做相应处理后看成一个计算单位．以基因为单位是指对基因所对应的某表型下所有样本的表达值做相应处理后看成一个计算单位.

首先，设计 3 个以样本为单位的描述通路变异性的统计量．通路数据中样本 g 的各基因表达值方差为 $v_g^2=\dfrac{\sum_{i=1}^{p}(x_{ig}-\bar{x}_g)^2}{p-1}$，其中 g=[1, 2,……, n]，用表型 i 下的 n 个样本方差 v_g^2 的均值 $S_i^{2'}=\dfrac{\sum_{g=1}^{n}v_g^2}{n}$ 代表通路在 i 表型下的变异程度．通路中样本 g 的各基

因表达值到原点的欧氏距离为 $d_g=\left(\sum_{i=1}^{p} x_{ig}^2\right)^{1/2}$，用通路中 n 个样本的欧氏距离的方差 $SS_i^{2'}=\frac{\sum_{g=1}^{n}(d_g-\bar{d}_g)^2}{n-1}$ 来描述通路在 i 表型下的变异程度. 中位绝对差(*MAD*)可以很好地代替标准差(*SD*)，且计算简单对奇异值不敏感，所以选择 $MAD_i^{2'}$ 代替通路中 n 个样本的欧氏距离的方差 $SS_i^{2'}$ 来描述通路在 i 表型下的变异程度[19]. MAD 的定义为：对于一组数 $\{d_1, d_2, \cdots d_n\}$，MAD=1.4826 × median ($|d_g-z|$)，z=median(d_g)，因子 1.4826 用于调节 MAD 与 SD(标准差)的一致性.

其次，设计 3 个以基因为单位的描述通路变异性的统计量. 通路数据中基因 c 的样本表达值的方差为 $v_c^2=\frac{\sum_{i=1}^{n}(x_{ci}-\bar{x}_c)^2}{n-1}$，其中 c=[1, 2,……, p]，用通路中各基因在 i 表型下的方差均值 $S_i^2=\frac{\sum_{c=1}^{p} v_c^2}{p}$ 代表通路在 i 表型下的变异程度. 通路中基因 c 的各样本表达值到原点的欧氏距离为 $d_c=\left(\sum_{i=1}^{n} x_{ci}^2\right)^{1/2}$，用通路中 p 个基因的欧氏距离的方差 $SS_i^2=\frac{\sum_{c=1}^{p}(d_c-\bar{d}_c)^2}{p-1}$ 来描述通路在 i 表型下的变异程度. 同样用 MAD_i^2 代替通路中 p 个欧氏距离的方差 SS_i^2 来描述通路在 i 表型下的变异程度.

我们用前面介绍的 6 种统计量描述整个通路在 i 表型下的变异程度 σ_i^2，对于任何两种表型(疾病与正常)，对每个通路有双边假设：零假设 H_0，$\sigma_1^2=\sigma_2^2$；备择假设 H_1，$\sigma_1^2\neq\sigma_2^2$. 基于统计检验如果零假设被拒绝，即认为此通路是变异的. 在两种表型下，变异程度存在差异的通路就是富集通路.

1.2.2 14 种通路富集分析方法.

为了检测通路在两种表型下变异性是否存在差异，我们使用 F 检验方法. S_1^2 表示通路在疾病表型下的变异性，疾病表型样本数为 n，S_2^2 代表通路在正常表型下的变异性，正常表型样本数为 m，$F=S_1^2/S_2^2$ 服从 $F_{n-1,m-1}$ 分布，求得 P 值，进而可判断通路是否在两种表型下的变异性有差异. 中位绝对差(*MAD*)近似等于标准差(*SD*)，因此两表型下通路变异性比值 $MAD=MAD_1^2/MAD_2^2$ 也服从 $F_{n-1,m-1}$ 分布[19].

F 检验很简单且应用广泛，但是它对偏离正态分布的数据或奇异值很敏感(奇异值是样本中出现的过高或过低值). 为了克服这个缺陷，在 F 检验之前去掉计算得到的欧式距离 d_c 或者 d_g 中的奇异值，以便排除这些与大部分值相比过高或过低的值. 在本文中，我们用一个简单的 IQR 标准来检测奇异值. 对于给定的某种表型下所有欧式距离，定义 Q_1 与 Q_3 分别为 1/4 分位数和 3/4 分位数，并且 $IQ=Q_3-Q_1$. 任何欧氏距离小于 Q_1-rIQ 或大于 Q_3+rIQ 都被视为奇异值，这里 $r>0$. 奇异值被去掉后，重新确定反映不同表型下未过滤掉的欧氏距离个数. 在本文中我们用 r=1.5，因为它在实践中可以达到理想的过滤效果[11].

另外，我们设计了 4 种基于随机扰动(permutation)的检验方法，随机扰动时所用的检验统计量有两种是基于两种表型下通路变异性的比值，即 SS_1^2/SS_2^2 与 S_1^2/S_2^2. 另外两种检验统计量是基于两种表型下通路变异性的差异的，分别定义为 *SD* Diff=S_1-S_2 与 *MAD* Diff=MAD_1-MAD_2. 随机扰动(permutation)分为随机扰动类标签和随机扰动基因集两类，本文应用随机扰动类标签. 然后计算原始统计量在扰动产生的统计量中的排序，进而得到 P 值. 在我们的研究中，每个通路表达谱数据扰动 10 000 次. 14 种通路富集方法见表 1.

Table 1 Fourteen pathway enrichment methods

Method	Statistic	Test	Distribution
1	$SS_1^{2'}/SS_2^{2'}$	F	$F_{n-1,m-1}$
2	$MAD_1^{2'}/MAD_2^{2'}$	*MAD*	$F_{n-1,m-1}$
3	$SS_1^{2'}/SS_2^{2'}$	F outlier removed	$F_{n-1,m-1}$
4	$SS_1^{2'}/SS_2^{2'}$	F, permutation	Empirical
5	$S_1^{2'}/S_2^{2'}$	F, permutation	Empirical
6	$S_1'-S_2'$	*SD* Diff, permutation	Empirical
7	$MAD_1'-MAD_2'$	*MAD* Diff, permutation	Empirical
8	SS_1^2/SS_2^2	F	$F_{p-1,p-1}$
9	MAD_1^2/MAD_2^2	*MAD*	$F_{p-1,p-1}$
10	SS_1^2/SS_2^2	F outlier removed	$F_{p-1,p-1}$
11	SS_1^2/SS_2^2	F, permutation	Empirical
12	S_1^2/S_2^2	F, permutation	Empirical
13	S_1-S_2	*SD* Diff, permutation	Empirical
14	MAD_1-MAD_2	*MAD* Diff, permutation	Empirical

2 结果与分析

2.1 预处理方法的比较

2.1.1 基于预处理后的芯片数据评价预处理方法.

a．从真实的基因 - 基因互作对之间相关性的角度分析预处理方法．两个互作的蛋白质趋向于共表达，它们的基因表达数据也就有高相关性[20]．我们从人类蛋白质互作数据库(HPRD)[21]下载到所有的蛋白质 - 蛋白质互作关系对，去除自我互作数据后得到 25 723 对蛋白质 - 蛋白质(基因 - 基因)互作数据．用 RMA、MAS5.0、GCRMA、FARMS 和 DFW 5 种预处理方法得到的 GDS1650 数据计算 25 723 对基因 - 基因表达谱数据的相关系数(r)，然后分析相关系数绝对值 $|r|$ 的分布．结果显示，应用 RMA 处理 的数据，相关系数大于 0.7 的基因 - 基因互作对最多(4.57%)，并且所有基因对的平均相关系数最高(表 2)．由此看出，对于 GDS1650 数据集，RMA 方法处理的数据最适合用于研究已经确认的基因 - 基因互作对，RMA 方法得到的数据比其他方法更能体现基因 - 基因对真实的互作关系.

Table 2 The comparison of correlation coefficient ($|r|$) distribution of gene pairs

	RMA	MAS5.0	GCRMA	FARMS	DFW
The average r of all the gene pairs	0.263	0.222	0.239	0.253	0.228
The fraction of the gene pairs with $\|r\| > 0.5$	17.66%	11.68%	15.90%	16.02%	11.71%
The fraction of the gene pairs with $\|r\| > 0.6$	9.47%	5.63%	8.99%	8.57%	5.30%
The fraction of the gene pairs with $\|r\| > 0.7$	4.57%	2.42%	4.51%	3.96%	2.02%
The fraction of the gene pairs with $\|r\| > 0.8$	1.71%	0.93%	1.80%	1.36%	0.59%
The fraction of the gene pairs with $\|r\| > 0.9$	0.44%	0.25%	0.50%	0.26%	0.11%

b．从分别以基因和样本为单位计算的方差在各预处理方法下的相关性角度分析预处理方法．RMA、MAS5.0、GCRMA、FARMS 和 DFW 5 种预处理方法分别处理 GDS1650 芯片数据所得结果，分别以样本和基因为单位计算方差(v_g^2 与 v_c^2)，然后计算两两预处理方法所得结果之间的相关性．表 3 与表 4 分别显示以样本为单位和以基因为单位计算的方差在各预处理方法下的相关性．从两个表中可以看出 RMA 处理得到的数据与其他几种方法得到的数据相关性最高.

Table 3 The correlation of v_g^2 from the GDS1650 preprocessed by five different preprocessing methods

	MAS5.0	GCRMA	FARMS	DFW
RMA	0.3834	0.8290	0.8419	0.6089
MAS5.0		0.1755	0.1525	–0.1345
GCRMA			0.8751	0.6500
FARMS				0.4588

Table 4 The correlation of v_c^2 from the GDS1650 preprocessed by five different preprocessing methods

	MAS5.0	GCRMA	FARMS	DFW
RMA	0.5631	0.9350	0.9172	–0.0015
MAS5.0		0.5856	0.4673	0.0008
GCRMA			0.8608	–0.0004
FARMS				–0.0010

2.1.2 结合 GSEA 所富集到的通路结果评价预处理方法.

5 种预处理方法处理 GDS1650 原始数据得到的表达谱数据分别用 GSEA 工具富集分析，从富集结果中分别选择排在前面的 40 个通路，计算 5 种预处理方法处理所得数据经 GSEA 分析所富集到的通路的重合性(表 5)．重合性为两种方法下的通路交集个数除以 40，表 5 表明 RMA 方法与其他方法的重合性最高.

Table 5 Concordance of the 40 most enriched pathways from the analysis of GSEA using the GDS1650 preprocessed by five different preprocessing methods

	MAS5.0	GCRMA	FARMS	DFW
RMA	0.600	0.700	0.575	0.375
MAS5.0		0.600	0.375	0.300
GCRMA			0.475	0.375
FARMS				0.450

从 5 种 GSEA 结果中选择出显著的($P < 0.05$)通路，计算 5 种预处理方法下显著通路的交集个数(表 6). 从表中清楚地看出 RMA 方法下显著通路个数与其他方法的交集个数最多.

Table 6　Number of overlapping significant pathways from the analysis of GSEA using the GDS1650 preprocessed by five different preprocessing methods

	MAS5.0	GCRMA	FARMS	DFW
RMA	6	13	1	1
MAS5.0		5	0	1
GCRMA			1	1
FARMS				0

我们从两个角度三个方面分析了预处理方法对数据和富集结果的影响. 计算已知的基因 - 基因对之间的相关系数，RMA 法所得数据是最优的能反映它们之间真实互作关系的数据. 以基因和样本为单位计算的方差在各预处理方法下的相关性显示：RMA 处理得到的数据与其他几种方法得到 的数据相关性最高，计算 5 种预处理方法处理所得数据经 GSEA 分析所富集到的通路的重合性，表明 RMA 方法与其他方法的重合性最高. 综合分析结果可以得出：RMA 预处理方法处理原始的 Affymetrix 芯片数据能很好地反映基因对真实的互作关系而且数据最稳定；RMA 预处理方法处理的数据经 GSEA 富集得到的结果最优. 所以后续的分析我们应用 RMA 预处理方法处理所得的表达谱数据.

2.2　基于变异性的通路富集方法研究的可行性分析

由 **2.1** 的分析可知 RMA 法是最优的预处理方法，所以本部分我们应用 RMA 方法处理得到的芯片数据进行分析. 研究 169 个通路表达谱数据，观察通路在两种表型下变异程度是否存在差异. 我们的假设是：通路变异性程度在不同表型中存在差异. 为了证明此假设成立，分别研究了 6 种描述通路变异程度的统计量，观察不同表型下统计量是否有明显差异.

a. 通路中各样本欧氏距离的方差 $SS_i^{2'}$ (方法 1, 3, 4 中的统计量)作为描述通路变异性的统计量，观察各通路在两种表型下变异性情况(图 1).

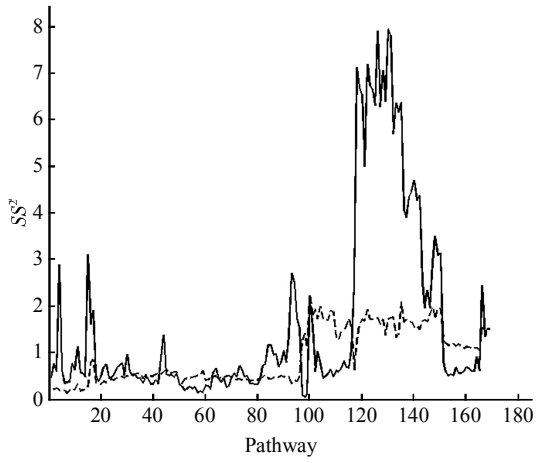

Fig. 1　The variance of each pathway between two conditions

——: Cancer; - - - : Health.

通路中各样本欧氏距离的方差作为描述通路变异性的统计量，其在两种表型下的值在一些通路中差别很明显. 从图 1 中可看出大部分通路在两种表型下变异性程度变化不大，但有一些通路有比较明显的差异.

b. 通路中各基因方差均值 S_i^2(方法 12, 13 中的统计量)作为描述通路变异性的统计量，观察各通路在两种表型下的变异性情况(图 2).

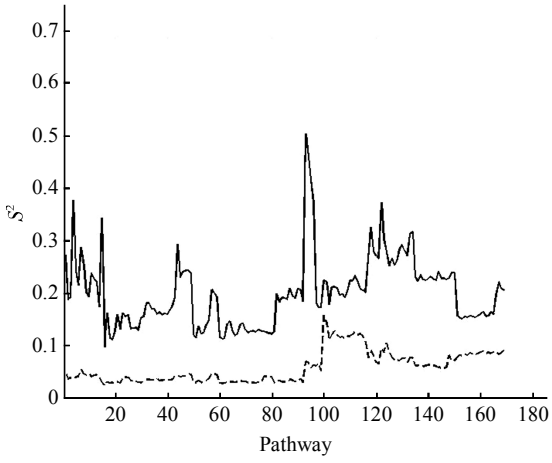

Fig. 2　The variance of each pathway between two conditions

——: Cancer; - - - : Health.

通路中各基因方差均值作为描述通路变异性的统计量，其在两种表型下的数值有明显差异(图 2). 169 个通路在两种表型下的变异程度都存在可见的差异，这验证了我们的假设．其他 4 种统计量在通路两种表型下的变异性情况见网络版附录(http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120410&flag=1).

通过 6 方面的图示可以明确得出在疾病表型下一些通路的变异性有明显的上升或者下降．我们认为在不同表型下通路的变异性程度存在差异，即我们的假设成立．在这一基础上，用 14 种通路富集方法检测通路变异性的显著情况，计算每个通路显著性 *P* 值，并分析 14 种方法的效能．

2.3 评价检验方法和统计量

我们用 14 种通路富集方法和 GSEA 对 RMA 预处理后的通路表达谱数据进行富集分析，*P* 值都用 BH 方法进行 FDR 校正[22]，0.05 为 *P* 值的阈值．为了评价所设计的基于通路变异性富集分析方法和 GSEA 法的优越性，手工检索文献，查询了 169 个通路与 Pulmonary adenocarcinoma 是否相关. 用此文献检索结果作为标准，分析我们的方法和 GSEA 结果的敏感度和特异度(表 7)．表 7 中 a 代表真阳性数，b 代表假阳性数，c 代表假阴性数，d 代表真阴性数，N 为 a、b、c、d 4 个数之和．*TP* 表示真阳性率(敏感度)$a/(a+c)$，*TN*表示真阴性率(特异度)$d/(b+d)$，*FP* 为假阳性率 $b/(b+d)$，*FN* 为假阴性率 $c/(a+c)$．从表中可以看出，在 *TP* 一列 12 和 13 方法的值很高，但是它们的 *FP* 也非常高，导致特异度很低，所以方法 12、13 不可取．其他方法诸如 1、4、7、8、9、10、14 和 GSEA 等特异度很高，但是敏感度很低，真阳性个数很少，有的甚至为 0，所以这些方法也不可取．我们希望所选方法的敏感度和特异度都在一个合理的较高水平．为了真实反映方法的分类效果，必须综合考虑敏感度和特异度两个指标．跃登指数为真阳性率与假阳性率之差，它能综合反映一个方法的敏感度和特异度[23]．跃登指数越大能直观反映出此方法的分类效果越好，方法 11 的跃登指数最高，说明方法 11 效果最好．

Table 7 Sensitivity and specificity of each method

Method	*a*	*b*	*c*	*d*	*TP*	*FP*	*FN*	*TN*	Accuracy (*a*+*d*)/*N*	Youden index (*TP*−*FP*)
1	11	16	37	105	0.229	0.132	0.771	0.868	0.686	0.097
2	21	50	27	71	0.438	0.413	0.563	0.587	0.544	0.024
3	19	59	29	62	0.396	0.488	0.604	0.512	0.479	−0.092
4	3	0	45	121	0.063	0.000	0.938	1.000	0.734	0.063
5	22	67	26	54	0.458	0.554	0.542	0.446	0.450	−0.095
6	22	67	26	54	0.458	0.554	0.542	0.446	0.450	−0.095
7	0	0	48	121	0.000	0.000	1.000	1.000	0.716	0.000
8	0	0	48	121	0.000	0.000	1.000	1.000	0.716	0.000
9	0	0	48	121	0.000	0.000	1.000	1.000	0.716	0.000
10	0	0	48	121	0.000	0.000	1.000	1.000	0.716	0.000
11	29	45	19	76	0.604	0.372	0.396	0.628	0.621	0.232
12	47	121	1	0	0.979	1.000	0.021	0.000	0.278	−0.021
13	47	121	1	0	0.979	1.000	0.021	0.000	0.278	−0.021
14	15	27	33	94	0.313	0.223	0.688	0.777	0.645	0.089
GSEA(nom-p)	9	8	39	113	0.188	0.066	0.813	0.934	0.722	0.121
GSEA(BH)	0	0	48	121	0.000	0.000	1.000	1.000	0.716	0.000

为了进一步分析各方法的富集效能，我们把文献检索的结果作为标准，对 14 种方法和 GSEA 进行了 ROC 曲线分析，ROC 曲线下的面积反应方法的分类准确性，面积越大分类效果越好．表 8 显示方法 11 在所有实验方法中最优．

Table 8 Area under ROC curve of fourteen types of methods and GSEA

Method	Area under ROC curve
1	0.6347
2	0.5579
3	0.5746
4	0.6366
5	0.5927
6	0.5972
7	0.5580
8	0.5000
9	0.5000
10	0.5000
11	0.6816
12	0.6294
13	0.5958
14	0.5641
GSEA	0.6585
GSEA(nom-p)	0.6756

3 讨 论

不同的芯片预处理方法对芯片原始数据采取不同的背景校正、探针校正和数据归一化方法，这样就导致预处理的结果数据不一致．我们通过各预处理方法所得的芯片数据评价 5 种芯片预处理方法，同时，将 5 种预处理的表达谱数据用现在最常用的 GSEA 富集方法进行分析，研究预处理方法对富集分析结果的影响．结果表明，RMA 方法预处理的数据比其他方法更能体现基因对真实的互作关系，而且所得数据与其他方法一致性最高，同时，RMA 方法所得富集结果与其他方法所得结果的重合性最高，这说明 RMA 预处理方法预处理的芯片数据最适合用于富集分析．

当前富集分析方法的思想主要基于超几何分布和 GSEA 算法．超几何分布原理是推断每个功能基因集中的差异表达基因比例是否与整个基因芯片上差异表达基因的比例相同[24]．如果功能基因集中的差异表达基因比例远远大于整个基因芯片上差异表达基因的比例，则认为某疾病与此功能基因集相关．GSEA 的思想是根据基因与表型的相关性先对基因进行排秩，然后检验功能基因集中的基因在经过排秩的基因列表中是随机排列还是主要集中在列表的顶部或者底部[25]．基于基因集的变异性(方差)进行富集分析很少有人研究过．我们首次设计了几种描述通路变异性的统计量并配以合适的检验方法，检验不同表型下基因集的变异性是否存在差异．研究结果表明不同表型下通路的变异性程度存在差异．为了评价设计的基于变异性通路富集分析方法和 GSEA 算法的富集效果，我们手工检索文献查找通路与 Pulmonary adenocarcinoma 是否相关，然后用检索结果作为标准评价这些方法．在我们设计的 14 种用于通路富集分析的方法中，通路中以各基因欧氏距离的方差 SS_i^2 做统计量进行 permutation 检验(方法 11)是最适宜的基于变异性的通路富集分析方法，其富集结果优于 GSEA．

基于变异性的通路富集分析较之传统的富集分析，具有完全不同的生物学假设，其结果可以作为后者的有效补充．例如，本文提出的方法敏感地发现 hsa04310: Wnt signaling pathway (P = 0.0019)[26]，hsa04330: Notch signaling pathway (P = 0.014)[27]，hsa04140: Regulation of autophagy (P = 0.0014)[28]等与 Pulmonary adenocarcinoma 相关，而 GSEA 方法却没有检测出这些已报道的通路，这很可能是两种方法的生物学假设不同所导致的．基于变异性的通路富集分析不仅会丰富富集分析方法，更重要的是可以为人类疾病研究提供新的研究角度．

参 考 文 献

[1] Huang D W, Sherman B T, Lempicki R A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucl Acid Res, 2009, **37**(1): 1-13

[2] Huang D W, Sherman B T, Tan Q, *et al*. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biology, 2007, **8**(9): R183

[3] Beiβbarth T, Speed T P. GOstat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics, 2004, **20**(9): 1464-1465

[4] Hosack D A, Jr D G, Sherman B T, *et al*. Identifying biological themes within lists of genes with EASE. Genome Biology, 2003, **4**(10): R70

[5] Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, **102**(43): 15545-15550

[6] Kim S Y, Volsky D J. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics, 2005, **6**(1): 144

[7] Backes C, Keller A, Kuentzer J, *et al*. GeneTrail—advanced gene set enrichment analysis. Nucl Acid Res, 2007, **35**(suppl 2), W186-192

[8] Bauer S, Grossmann S, Vingron M, *et al*. Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data

exploration. Bioinformatics, 2008, **24**(14): 1650−1651

[9] Alexa A, Rahnenfüihrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 2006, **22**(13): 1600−1607

[10] Carmona-Saez P, Chagoyen M, Tirado F, *et al*. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biology, 2007, **8**(1): R3

[11] Ho J W K, Stefani M, Remedios C G D, *et al*. Differential variability analysis of gene expression and its application to human diseases. Bioinformatics, 2008, **24** (13): i390−i398

[12] Stearman R S, Dwyer-Nield L, Zerbe L, *et al*. Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model. Am J Pathol, 2005, **167**(6): 1763−1775

[13] Irizarry R A, Bolstad B M, Collin F, *et al*. Summaries of Affymetrix GeneChip probe level data. Nucl Acid Res, 2003, **31**(4): e15

[14] Affymetrix Inc. Statistical algorithms description document. Santa Clara: Affymetrix Inc, 2002 (2013-03-28). http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

[15] Wu Z, Irizarry R A, Gentleman R, *et al*. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc, 2004, **99**(468): 909−917

[16] Hochreiter S, Clevert D A, Obermayer K. A new summarization method for Affymetrix probe level data. Bioinformatics, 2006, **22**(8): 943−949

[17] Chen Z, McGee1 M, Liu Q Z, *et al*. A distribution free summarization methods for Affymetrix GeneChip arrays. Bioinformatics, 2007, **23**(3): 321−327

[18] Ogata H, Goto S, Sato K, *et al*. KEGG: Kyoto encyclopedia of genes and gnnomes. Nucl Acid Res, 1999, **27**(1): 29−34

[19] Rousseeuw P J, Croux C. Alternatives to the median absolute deviation. J Am Stat Assoc,1993, **88**(424): 1273−1283

[20] Bhardwaj N, Lu H. Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics, 2005, **21**(11): 2730−2738

[21] Prasad T S K, Goel R, Kandasamy K, *et al*. Human protein reference database——2009 update. Nucl Acid Res, 2009, **37**(suppl 1): D767−772

[22] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B, 1995, **57**(1): 289−300

[23] 刘 杰, 林一帆, 张 沥, 等. 图象自动分析检测 MG7 抗原表达预测胃癌高危价值探讨. 中华预防医学杂志, 1996, **30**(5): 286−288
Liu J, Lin Y F, Zhang L, *et al*. Chin J Prev Med, 1996, **30** (5): 286−288

[24] Tian L, Greenberg S A, Kong S W, *et al*. Discovering statistically significant pathways in expression profiling studie. Proc Nat Acad Sci USA, 2005, **102**(38): 13544−13549

[25] Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, **102**(43): 15545−15550

[26] Sunaga N, Kohno T, Kolligs F T, *et al*. Constitutive activation of the Wnt signaling pathway by *CTNNB1* (β-catenin) mutations in a subset of human lung adenocarcinoma. Genes, Chromosomes and Cancer, 2001, **30**(3): 316−321

[27] Chen Y B, Marco M A D, Graziani I, *et al*. Oxygen concentration determines the biological effects of NOTCH-1 signaling in adenocarcinoma of the lung. Cancer Res, 2007, **67**(17): 7954−7959

[28] Chen N, Karantza-Wadsworth V. Role and regulation of autophagy in cancer. Biochim Biophys Acta - Mol Cell Res, 2009, **1793** (9): 1516−1523

A Method of Pathway Enrichment Analysis Based Gene Expression Variability*

JIA Xiao-Dong[1)**], CHEN Xiu-Jie[1)**,***], WU Xin[2)**,***], XU Jian-Kai[1)**], TAN Fu-Jian[1)], LIU Xiang-Qiong[1)], LIU Lei[1)], YANG Rui-Zhi[1)]

([1)] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin* 150086, China;

[2)] *The Third Affiliated Hosptital, Harbin Medical University, Harbin* 150086, China)

Abstract Current pathway enrichment method is mainly based on the gene that are differentially expressed, and no enrichment method considers pathway variability (variance). We observed that in the phenotype of disease, some pathways have a significant increase or decrease in variability describing appropriate statistics. Therefore, in this article, we hypothesize that the variation of single pathway is significantly different between two phenotypes. We designed fourteen types of statistics coupled with their test methods to analyze pathways variation and the pathways enrichment significance between two phenotypes, and we compared the results with those obtained by document retrieval. At the same time, the results of five different data preprocessing methods on data were investigated. The results show that RMA is stable in the five gene expression data preprocessing methods. The pathway variation is different between the two phenotypes. According to the literature research results, the permutation test coupled with the variance of Euclidean distance of each gene (the eleventh method) can identify significant pathways more efficiently than GSEA. In conclusion, pathway enrichment analysis strategy based on the pathway variation is feasible, which could be a theoretical guideline for enrichment analysis and a new biological insights of study in human diseases.

Key words variation, pathway, enrichment analysis, preprocessing method

DOI: 10.3724/SP.J.1206.2012.00410

---

*This work was supported by grants from The Master Innovation Funds of Heilongjiang Province (YJSCX2012-205HLJ, YJSCX2011-341HLJ, YJSCX2012-223HLJ), The Provincial Education Department Project of Heilongjiang (11541121,12531227), The Innovation Manpower Fund of Harbin Science and Technology Bureau (2010RFXXS053), The National Science Foundation of Heilongjiang Province (QC2012C010), Department of health of Heilongjiang Province (2012-798).

**These authors contributed equally to this work.

***Corresponding author.

CHEN Xiu-Jie. Tel: 86-451-86605922, E-mail: chenxiujie@ems.hrbmu.edu.cn

WU Xin. Tel: 86-451-86298703, E-mail: wuxin@ems.hrbmu.edu.cn

Received: February 4, 2013 Accepted: March 28, 2013

附　录

基于变异性的通路富集方法研究的可行性分析

通路中各样本方差均值 S_i^2(方法 5、6 中的统计量)作为描述通路变异性的统计量，观察各通路在两种表型下变异性情况(图 S1).

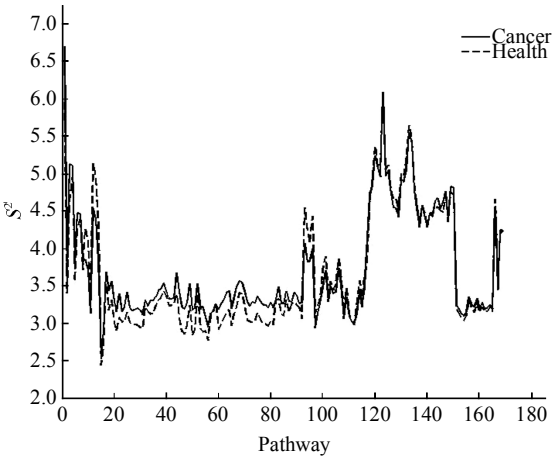

Fig. S1　The variance of each pathway between two conditions

通路中各样本欧式距离的值 MAD_i^2(方法 2、7 中的统计量)作为描述通路变异性的统计量，观察各通路在两种表型下变异性情况(图 S2).

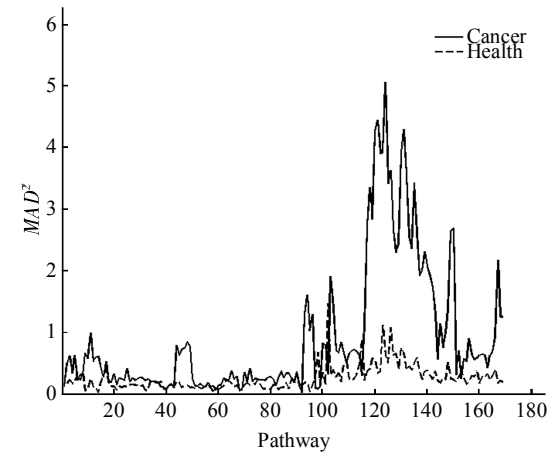

Fig. S2　The variance of each pathway between two conditions

通路中各基因欧氏距离的方差 SS_i^2(方法 8、10、11 中的统计量)作为描述通路变异性的统计量，观察各通路在两种表型下变异性情况(图 S3).

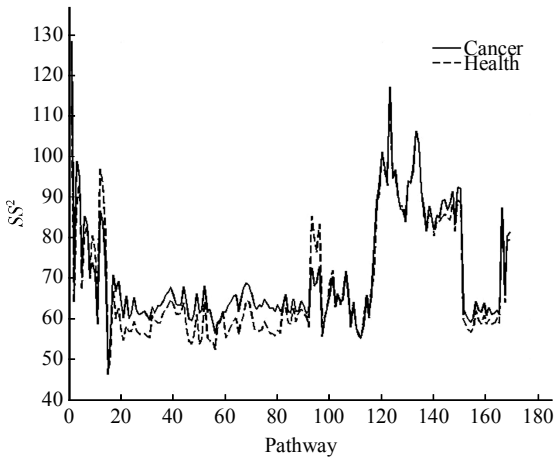

Fig. S3　The variance of each pathway between two conditions

通路中各基因欧式距离的 MAD_i^2 值(方法 9、14 中的统计量)作为描述通路变异性的统计量，观察各通路在两种表型下变异性情况(图 S4).

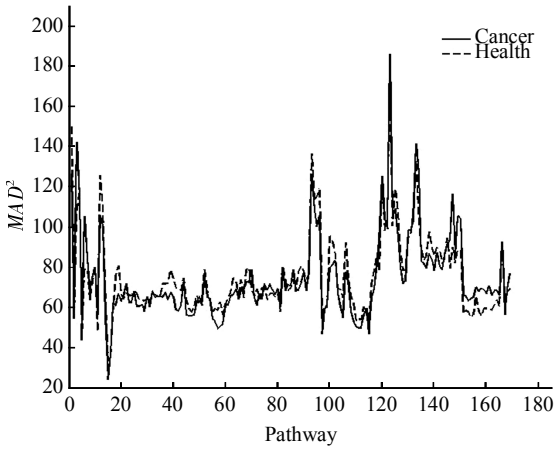

Fig. S4　The variance of each pathway between two conditions