中国科学: 生命科学 2016 年 第 46 卷 第 4 期: 363 ~ 373
SCIENTIA SINICA Vitae life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述 中国知名大学及研究院所专栏 复旦大学生命科学学院专辑

# 基因功能富集分析的研究进展

王潇[①†], 尹天舒[①†], 李柏逸[①†], 江熹霖[①], 孙慧[①], 窦亚光[①], 倪琦[①], 田卫东[①②*]

① 复旦大学生命科学学院生物统计学与计算生物学系, 上海 200438;
② 复旦大学附属儿科医院, 上海 201102
† 同等贡献
* 联系人, E-mail: weidong.tian@fudan.edu.cn

收稿日期: 2016-01-02; 接受日期: 2016-02-22
国家自然科学基金(批准号: 91231116, 31071113, 30971643)和国家重点基础研究发展计划(批准号: 2012CB316505, 2010CB529505)资助

**摘要** 基因功能的富集分析已成为高通量组学数据分析的常规手段, 对于揭示生物医学分子机制具有重要意义. 目前已有上百种基因功能富集分析的方法和工具. 根据所解决的问题和算法的原理, 这些方法可大体分为过代表分析、功能集打分、基于通路拓扑结构和基于网络拓扑结构 4 大类. 本文对这 4 大类方法的原理及其中的典型方法进行了综述, 并讨论了基因功能富集分析结果的冗余性问题及建立标准数据集的必要性.

**关键词** 组学数据, 功能富集, 冗余性, 标准数据集

随着高通量测序技术的飞速发展及相关技术的广泛应用, 生物医学相关研究领域已进入了大规模组学数据呈指数增长的后基因组时代[1]. 一方面, 这使得生物医学研究得以从单个基因的分析转变为系统水平上的研究, 对于揭示生物医学的基本分子机制具有重要推动作用. 但另一方面, 如此庞大的数据量也给信息的有效提取和分析带来了巨大的挑战. 为了从庞杂的组学数据中发掘规律, 研究者通常会对基因功能进行富集分析, 期望发现在生物学过程中起关键作用的生物通路, 从而揭示和理解生物学过程的基本分子机制. 现在, 基因功能的富集分析已成为功能组学数据分析的常规手段, 并随着高通量组学数据的发展, 如从基因芯片数据到 RNA-seq 数据的转变, 开发出了一系列相应的分析方法, 最早开发的过代表分析(over-representation analysis, ORA)仅针对一组基因, 而高通量组学数据的发展使得功能集打分(functional class scoring, FCS)应运而生, 随着对于生物学通路及复杂网络的深入完善和了解, 又相继开发了基于通路拓扑结构(pathway topology, PT)和基于网络拓扑结构(network topology, NT)的方法. 本文拟对现有的基因功能富集分析方法进行简要的总结评述, 以方便研究者了解相关领域, 并选择适合的研究工具.

## 1 基因功能富集分析的基因功能数据库和数据类型

基因功能富集分析中的基因功能指的是众多代

引用格式: 王潇, 尹天舒, 李柏逸, 等. 基因功能富集分析的研究进展. 中国科学: 生命科学, 2016, 46: 363–373
Wang X, Yin T S, Li B Y, et al. Progress in gene functional enrichment analysis. Sci Sin Vitae, 2016, 46: 363–373, doi: 10.1360/N052016-00139

 

表一定的基因功能特征和生物过程的基因功能集(gene set)[2]. 由这些基因功能集构成的常用基因功能数据库有 GO[3], 生物学通路, 包含生化反应、代谢或信号通路的 KEGG[4,5], Reactome[6], Biocarta[7]等, 整合数据库, 如 MsigDB[8]等.

在功能组学研究中, 研究者通常会获得一组他们感兴趣的基因, 如在疾病和正常组织中有显著差异表达的基因, 在药物或外界环境刺激下特定组织中表达水平有显著异常的应激基因等. 要揭示其中隐含的生物学分子机制, 研究者可针对这组感兴趣的基因, 进行基因功能的富集分析, 发现在其中有显著富集的特定生物学通路, 从而从分子机制上来解释所观察到的生物学现象. 除此以外, 高通量组学技术, 如基因表达芯片(microarray)或 RNA-seq, 可获得基因组中所有基因的表达水平. 为充分利用获得的高通量数据, 研究者也可以直接针对全基因组基因表达谱信息来进行富集分析, 从中鉴定出案例和对照状态下在研究对象中发生显著表达差异的生物通路, 从而揭示其中的生物学分子机制. 针对这些不同的数据需要开发不同的功能富集分析方法. 例如, 对于基因表达芯片和 RNA-seq, 在富集分析过程中原始数据的处理方式是不同的. 其中, 基因芯片记录的是连续的荧光信号强度值, 而 RNA-seq 记录的是 RNA 序列的读段个数[9], 需要采用不同的统计模型进行分析. 即使对同一类型数据, 基于不同的假说和统计方法, 研究者也开发出了不同的富集分析算法和模型. 现在已有上百种富集分析的方法和工具, 一方面极大地促进了研究者的科研工作进展, 另一方面也给研究者在选择合适的研究工具时带来一些困扰. 以下将针对现有方法进行分类综述, 具体方法及工具详见表 1.

## 2 基因功能富集分析方法基于算法的分类

基因功能富集分析的方法基于数据来源和算法大致可以分为 4 大类: ORA, FCS, PT, NT 的方法(图 1). 下文将对每类算法分别介绍.

### 2.1 过代表分析(ORA)方法

(1) 算法原理. 作为最早出现的一类基因功能富集方法, ORA 针对的数据是一组感兴趣的基因(基因列表), 其目的是在这组基因中发现有明显统计学上富集的基因功能集. 其基本步骤包括先将给定的基因列表与待测功能集做交集, 找出其中共同的基因并进行计数(统计值), 最后利用统计检验的方式来评估观察的计数值是否显著高于随机, 即待测功能集在基因列表中是否显著富集. 常见的统计学方法有卡方检验, Fisher 精确检验和二项分布检验[45], 而其中最为广泛使用的是 Fisher 精确检验, 即利用 2×2 的列联表, 根据超几何分布来检验基因列表中的基因在待测功能集中是否显著富集.

(2) 常用方法和工具. 目前有许多工具及数据库提供 ORA 的使用, 包括 DAVID, GOstat, GenMAPP 等. 其中 DAVID 提供的基因功能集数据库最为全面, 不仅包含大量不同物种的基因功能注释信息, 也涵盖了主流的生物通路注释库如 GO 条目和 KEGG 通路, 而且还提供了基因名称转换功能, 及良好的结果展示界面. 因而, DAVID 已成为目前应用最广泛的 ORA 分析工具.

(3) 优缺点. ORA 方法基于完备的统计学理论, 具有结果稳健、可靠的优点. 但目前常用的基于统计检验的 ORA 方法也有一定的局限性, 包括: (ⅰ) 在对基因进行计数时, 丢失了基因的表达水平或表达差异值等基因属性信息; (ⅱ) 把通路中的所有基因进行同等对待, 忽视了基因在通路内部生物学意义的不同(如调控和被调控基因的不同)及基因间复杂的相互作用; (ⅲ) 在获得感兴趣的基因时, 往往需要选取合适的阈值, 而这样有可能会丢失显著性较低但比较关键的基因, 导致检测灵敏性的降低. 为此, 人们需要开发新的富集分析方法来解决这些局限性.

### 2.2 功能集打分(FCS)方法

(1) 算法原理. 相比于针对一组感兴趣的基因通过计数来进行富集分析的 ORA 方法, 第二代功能富集分析方法 FCS 的输入数据不仅是全基因组基因, 并且还考虑到每个基因的表达水平或表达差异值等基因属性信息. 此外, ORA 的检验对象是感兴趣的基因列表与待测基因功能集的共同基因, 而 FCS 的检验对象则是待测基因功能集中的所有基因. FCS 方法的基本步骤包括: 首先根据案例和对照状态下的基因表达谱对基因组中所有基因表达水平的差异值进行打分或排序, 或直接输入排序好的基因表达谱; 其次是把待测基因功能集中的每个基因的分数通过特定的统计模型转换为待测基因功能集的分数或统计

**表 1 常用基因功能富集分析方法**

| 类型 | 方法 | 可用性 | 使用或下载网址 |
|---|---|---|---|
| ORA | DAVID[10] | 在线工具 | https://david.ncifcrf.gov |
| | GOstat[11] | 在线工具 | http://gostat.wehi.edu.au |
| | GenMAPP[12] | 在线工具 | http://www.genmapp.org |
| | GoMiner[13] | 在线工具 | http://discover.nci.nih.gov/gominer |
| | Onto-Express[14] | 在线工具 | http://vortex.cs.wayne.edu |
| FCS | GSEA[8,15] | Java 软件, R 语言包 | http://software.broadinstitute.org/gsea |
| | GSA[16] | R 语言包 | https://cran.r-project.org/web/packages/GSA/index.html |
| | PADOG[17] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/PADOG.html |
| | SAFE[18] | R 语言包 | http://www.bios.unc.edu/~fwright/SAFE |
| | Globaltest[19,20] | R 语言包 | http://www.bioconductor.org/packages/2.0/bioc/html/globaltest.html |
| | Sigpathway[21] | R 语言包 | http://bioconductor.org/packages/release/bioc/html/sigPathway.html |
| | GAGE[22] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/gage.html |
| | GSVA[23] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/GSVA.html |
| | PLAGE[24] | R 语言包 | http://dulci.biostat.duke.edu/pathways/misc.html |
| | ZSCORE[25] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/limma.html |
| | SSGSEA[26] | R 语言包 | http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/ssGSEAProjection/ |
| | MRGSE[27] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/limma.html |
| | ANCOVA[28] | R 语言包 | https://cran.r-project.org/web/packages/fANCOVA/index.html |
| | CAMERA[29] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/limma.html |
| PT | MetaCore | 商业软件 | http://www.genego.com/metacore.php |
| | Pathway-Express[30] | 在线工具, R 语言包 | vortex.cs.wayne.edu/projects.htm |
| | SPIA[31] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/SPIA.html |
| | TopoGSA[32] | 在线工具 | www.topogsa.org |
| | CePa[33] | R 语言包 | https://cran.rstudio.com/web/packages/CePa/index.html |
| | ToPASeq[34] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/ToPASeq.html |
| | NetGSA[35] | R 语言包 | https://cran.r-project.org/web/packages/netgsa/index.html |
| | DEGraph[36] | R 语言包 | www.bioconductor.org/packages/release/bioc/html/DEGraph.html |
| | BPA[37] | 软件包 | http://bumil.boun.edu.tr/bpa |
| | ACST[38] | R 语言包 | http://omictools.com/analysis-of-consistent-signal-transduction-tool |
| NT | NEA[39] | R 语言包 | https://r-forge.r-project.org/projects/nea2 |
| | EnrichNet[40] | 在线工具, R 语言包 | www.enrichnet.org |
| | GANPA[41] | R 语言包 | https://cran.r-project.org/web/packages/GANPA/index.html |
| | LEGO[42] | 在线工具, R 语言包 | lego.tianlab.cn |
| | NOA[43] | 在线工具 | app.aporc.org/NOA |
| | GOGANPA[44] | R 语言包 | https://cran.r-project.org/web/packages/GOGANPA/index.html |

值; 最后利用随机抽样获得的待测基因功能集统计值的背景分布来检验实际观测的统计值的显著水平, 并判断待测基因功能集在案例和对照实验状态下是否发生了统计上的显著变化.

(2) 常用方法和工具. GSEA 是常用的一种 FCS 方法. 其基本思路是首先基于表达差异值对全基因组基因进行排序得到基因列表, 然后检验待测基因功能集中的基因相对于随机情况而言, 是否显著地位于基因列表的顶端或底端, 即待测基因集的表达水平在案例和对照实验状态下是否发生了明显的变化. 具体而言, GSEA 首先计算了每个基因的表达水平与案例和对照两种状态下的关联系数, 并对关联系数从高到低进行了排序; 然后, 针对一特定的待测基因功能集, 根据其中每个基因的排序情况, 利用加权的近似 KS 检验, 获得待测基因功能集在排序列表中的 KS 检验值——也即待测基因功能集的统计值; 为检验观察统计值的显著性, GSEA 通过对样本的随机排列来获得统计值的背景分布, 并利用该分布来评估观察统计值的 $P$ 值. 除对样本的随机排列外, 在样本量较少的情况下, GSEA 也可用对基因的随机排列来估算待测基因功能集的显著水平.

在 FCS 方法中, 不同方法采用了不同的统计模

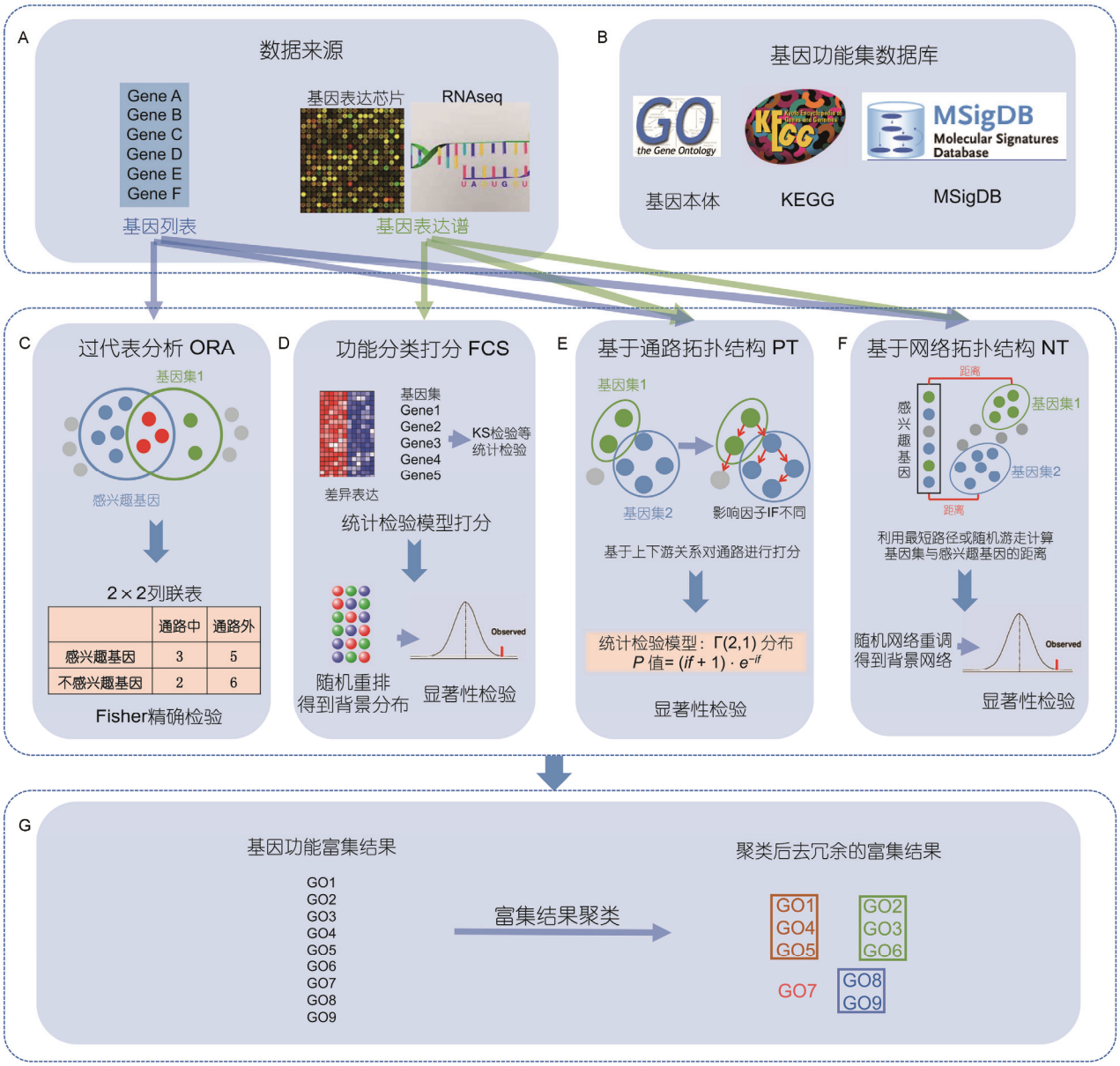


**图 1　基因功能富集分析的概况和分类**

A: 基因功能富集分析输入的数据来源包括基因列表及基因表达谱; B: 基因功能富集分析输入的基因功能注释数据库包括 GO, KEGG, MSigDB 等; C: 过代表分析 ORA, 先计数基因列表与基因功能集共同的基因, 并利用 2×2 列联表, 进行 Fisher 精确检验; D: 功能集打分 FCS, 以 GSEA 为例, 从基因表达谱中对所有基因按照与表型的差异表达程度排序, 再利用 KS 检验计算对待测基因功能集打分, 利用随机重排得到背景分布, 找出显著富集的基因功能集; E: 基于通路拓扑结构 PT, 以 Pathway-Express 为例, 根据上下游关系对通路进行打分, 上游通路影响因子较大, 对待测基因功能集根据影响因子打分, 然后根据统计检验模型计算显著程度; F: 基于网络的拓扑结构 NT, 以 NEA 为例, 将基因列表或基因表达谱和待测基因功能集放入生物网络中, 计算感兴趣的基因与待测基因功能集在网络之间的最短距离, 通过对网络重调得到背景分布, 找出显著富集的基因功能集; G: 基因功能富集分析输出的结果是排序后显著富集的基因功能集, 并对相似度较高的基因功能集进行去冗余

型来计算待测基因功能集的统计值. 如GSEA使用了加权的近似 KS 检验, GSA 利用基因的表达差异的 $t$ 值的绝对值来计算待测基因功能集的统计值, PADOG 采用基因的 $t$ 值加权平均值, SAFE 利用了

Wilcoxon rank sum 统计检验方法, 而 Global Test 则采用了经验贝叶斯广义线性模型. 另外, 在获得待测基因功能集统计值的背景分布时, 不同的 FCS 方法采用了两种主要模式来定义背景, 一类是竞争型(competitive)模式, 即将待测基因功能集外部的基因当作背景, 而另一类是自足型(self-contained)模式, 即将待测基因功能集本身当作背景. 总体来说, 自足型模式的检验功效要好于竞争型[46], 但少部分基因的显著性如果特别高的话也会造成自足型模式一定程度的过度预测. 无论是竞争型或自足型模式的 FCS 方法, 在通过随机抽样获得背景分布时, 既可以对基因进行随机排列(permutation)即基因抽样(gene sampling), 也可以对样本进行随机排列, 即样本抽样(subject sampling). 基因随机排列把每个基因独立对待, 而实际上基因之间有复杂的相互关系, 导致基因抽样的结果与实际的背景分布可能有一定的偏差. 而样本抽样可以保留基因间的相互关系, 因而抽样结果要更稳健一些. 因而, 在样本量大的情况下, 常用样本抽样; 而在样本量比较少的情况下, 则会利用基因抽样. 一般来说, 竞争型的模式通常采用基因抽样的方法, 如 Sigpathway-Q1, GAGE 等, 而自足型检验通常采用表型抽样的方法[47], 如 GSEA, Sigpathway-Q2, GSVA 等.

FCS 方法通常是把所有样本分为案例和对照两种状态, 然后来计算每个基因在两种状态下的表达差异值. 在这样的处理方式中, 单个样本中的基因表达信息没有被充分考虑. 例如, 待测基因功能集的基因有可能在一个样本中都有显著变化, 而如果放在两种状态下进行处理的话, 这种在单样本下的基因协同变化可能就无法观察到了. 为此, FCS 方法中还有一类基于单样本(single-sample, SS)的分析方法[48], 首先利用基因表达水平针对每个样本中的基因进行打分, 再利用常见的统计检验方法把样本层面的基因分数同表型联系起来. 这类方法包括 PLAGE, ZSCORE 及 SSGSEA 等. 该方法的一大优点是可以通过调整相关协变量, 相对简单地分析一些非常复杂的, 如包含时间进度的多样本设计[49].

传统的 FCS 方法主要针对基因表达芯片进行分析, 而随着高通量测序技术的发展, 已经开发出一些方法直接利用 RNA-seq 原始数据找到差异表达基因并进行功能富集, 通常使用泊松分布(poisson distribution)或负二项分布(negative binomial)找到差异表达基因, 常用方法有 edgeR[50], DESeq[51]等; 另一类方法则对原始数据进行转换后再沿用已有的基因表达芯片功能富集分析方法进行后续分析, 常用方法有 VOOM[52]等.

(3) 优缺点. 总体而言, FCS 相较于 ORA 方法在理论上有明显突破, 考虑到了基因表达值的属性信息, 而且以待测基因功能集为对象来进行检验, 也使得检验结果更加灵敏. 但 FCS 方法仍然把待测基因功能集中的每个基因作为独立的个体, 忽略了基因的生物学属性和基因间的复杂相互作用关系.

### 2.3 基于通路拓扑结构(PT)的方法

(1) 算法原理. ORA 和 FCS 方法在进行通路的富集分析时, 都将通路中的每个基因视作独立个体, 而实际上通路内的基因需要通过调控、被调控、相互作用等复杂的关系一起来影响细胞的发育、分化或疾病等生物学过程. 因而, 在进行通路的富集分析时, 尤其是基因表达的通路富集分析时, 有必要考虑到通路中基因的生物学属性. 例如, 在一个调控通路中, 上游基因的表达水平改变显然要远大于下游基因的表达水平改变对整个通路的影响. 基于通路拓扑结构的 PT 富集分析方法就是把基因在通路中的位置(上下游关系), 与其他基因的连接度和调控作用类型等信息综合在一起来评估每个基因对通路的贡献并给予相应的权重, 然后再把基因的权重整合入功能富集分析. 不同的 PT 方法在具体的权重打分时, 采用了不同的方式. 以下分别举例进行说明.

(2) 常用方法和工具. Pathway-Express 是首个引入通路拓扑结构的 PT 方法. 该方法引入了影响因子(impact factor, IF)这一概念来表征一特定通路对观察生物学现象的重要性. IF 整合了通路中显著差异表达基因数目和通路的拓扑结构, 被作为通路的最终统计量. IF 中的网络拓扑特征部分由通路中每个基因的扰动因子(perturbation factor, PF)计算得到. 一个基因的 PF 值包含了其自身和其上游基因的表达量信息. 由于通路的拓扑结构存在上下游关系, 所以上游通路中基因的差异性表达会随信号通路进行传递, 从而对整个通路的 IF 值产生较下游基因更为显著的影响. 最后, 在评估 IF 值的显著性时, 该方法采用了$\gamma$分布模型. Pathway-Express 的开发对后续研究工作有较大影响. 如 SPIA 在 Pathway-Express 的 IF 概念的

基础上, 在计算 PF 值时进一步引入了通路中每个调控关系的调控强度这一概念, 试图更加真实地反映了通路模型所包含的全部生物学信息. 除了上游基因表达量, 连接度(一个点与其他点直接相连的所有边的个数)、节点介数(所有最短路径中经过一个节点的路径的次数)等向心性参数也被引入作为表征通路的拓扑学特征. 如 TopoGSA 在比较通路间区别时, 引入了通路的向心性参数; CePa 引入了多种向心性参数并进行加权平均来计算通路的 IF 值. 现在已有一些基于 PT 算法的工具包, 如 ToPASeq 整合了包括 SPIA 方法在内的 7 种 PT 方法, 实现了 R 语言工具包, 可用来分析芯片数据及 RNA-seq 数据, 并能提供可视化展示结果.

(3) 优缺点. 总体来说, 对于研究较完善、拓扑结构完整的通路, 基于 PT 的基因功能富集算法会有更强的显著性; 由于原理上对于通路拓扑结构存在依赖性, 该类方法对于研究较少、信息不完善的通路稳健性较差, 因此目前通路注释的不完善也是限制基于 PT 的基因功能富集分析方法进一步发展的重要因素.

### 2.4 基于网络拓扑结构(NT)的方法

(1) 算法原理. PT 方法利用了通路的拓扑结构来把基因的生物学属性整合入功能的富集分析. 但目前在基因功能注释数据库中仅有 KEGG 提供了通路的拓扑结构, 而最常用的 GO 等注释数据库中基因功能集中不包含任何拓扑结构信息, 仅提供了可能属于同一通路的所有基因列表. 因而, PT 方法不能被用于 GO 通路的富集分析. 目前, 已有一些基于生物网络拓扑结构的富集分析方法, 它们利用数据库中的基因相互作用关系来间接地把基因的生物学属性整合入功能的富集分析. 这些方法的主要思路是利用现有的全基因组范围的生物网络, 如 HPRD[53], FunCoup[54], STRING[55]等, 来提取基因间的相互作用关系, 包括基因的连接度及基因在网络中的距离等, 来计算一给定的基因列表与一待测的基因功能数据集在网络中的连接关系, 从而来推测待测基因功能集是否与给定基因列表紧密相关, 如 NEA, EnrichNet 等. 另一些方法是利用网络拓扑结构来计算基因对特定生物通路的重要性并给予相应的权重, 然后再利用传统的 ORA 或 FCS 方法来评估特定生物通路的富集程度, 如 GANPA 和 LEGO 等. 还有一些方法是直接把基因列表中的功能富集问题利用网络转化为基因对的功能富集问题, 如 NOA 等. 以下分别举例进行说明.

(2) 常用方法和工具. NEA 和 EnrichNet 是两个基于网络距离的富集分析方法. 它们的主要思路都是去检验一个给定基因列表在网络中与待检测的生物通路的基因功能集相对于随机是否具有显著短的网络距离. 这两个方法的区别是 NEA 直接计算了给定基因列表与待检测基因功能集在网络中的平均连接度, 并通过对网络进行随机重调的方式, 来评估该统计量的显著程度; 而 EnrichNet 采用重启型随机游走(random walk with restart, RWR)的算法来计算给定基因列表与待检测基因功能集在网络中的距离. 然后利用随机网络与背景统计值进行比较来评估统计显著水平. 但由于网络的复杂性及对网络进行随机重调的计算效率问题, 这两个方法在实际应用过程中具有计算效率低的缺点, 而且实际测试中还发现由于对网络结构过大的依赖所造成的假阳性率高的系统偏差.

GANPA 利用了网络的拓扑结构来对通路内的基因赋予不同的权重, 用以表征该基因对通路重要性的不同. 其基本假设是如果一个通路内的基因在网络中大部分情况下仅与通路内部基因相连的话, 则该基因对通路的重要性要高于通路内部那些不仅和通路内的基因连接也和通路外部的基因连接的基因. 具体而言, GANPA 利用了超几何分布估计了一个基因在网络中与通路内部基因的连接度, 进而计算实际观察的该基因与通路内部基因的连接度与估计的连接度的差值, 用该差值来表示该基因对通路的重要性, 并作为该基因的权重. GANPA 把该基因的权重与基因表达的差异值相乘, 然后利用传统的 FCS 方法来评估一特定待测基因功能集的表达量变化的显著水平. GANPA 所用的网络是基于蛋白质互作网络、GO 的生物学过程(biological process, BP)注释和大规模基因表达芯片所构成的复杂的基因功能关联网络. 之后, GOGANPA 利用了 GO 注释构建了新的功能网络, 并可提供跨物种通用的功能富集分析. GANPA 和 GOGANPA 都是针对全基因组基因表达谱的 FCS 方法. 在这两个方法的基础上开发的 LEGO 专门针对基因列表的 ORA 分析. 与 GANPA 类似, LEGO 利用了网络的拓扑结构来给通路内部的基因赋予权重, 但与 GANPA 不同的是, LEGO 还考虑了

在网络中与通路紧密相关的邻居基因, 并也给它们赋予一定的权重. 在给定一基因列表后和一待测通路后, LEGO 把基因列表中的基因的通路特异性权重进行加权平均获得该通路的统计值; 之后, LEGO 通过基因随机排列的方法来获得该统计值的背景分布和对应的显著水平.

和以上方法不同, NOA 利用网络把一个基因列表的 ORA 分析转化为基因对的 ORA 分析. NOA 首先找到所有在网络中有连接的基因列表中的基因对, 并要求这些基因对应具有同样的功能; 然后, NOA 设计了一个完全网络作为背景网络, 利用卡方检验来检测这些特定功能的基因对是否显著高于随机.

(3) 优缺点. 总体而言, 与传统方法相比, 基于网络的基因功能富集分析方法加入了系统层面的基因重要性程度及关联信息, 使得预测结果更加准确可靠. 但是, 更多信息的加入也容易导致算法过于复杂, 计算速度较慢.

## 3 基因功能富集分析的冗余性问题

目前几乎所有的功能富集方法都是对待测基因功能集进行独立检验, 而现有的基因功能注释数据库中的基因功能集都存在一定的冗余现象, 也即基因功能集之间存在较多的共同基因, 因而也会导致富集的基因功能集之间出现冗余现象. 以 GO 数据库为例, 由于 GO 数据结构中的 GO 条目存在父子关系, 有些 GO 条目间的共同基因比较多, 使得 GO 富集结果的冗余现象尤为明显[56]. 富集结果的冗余现象对结果的解读造成一定的困扰, 难以准确揭示生物学机制. 针对这一问题, 现在已有一些初步的解决方案. 一种是在富集分析时, 不把基因功能集进行独立检验, 而是把所有基因功能集作为一个整体来进行富集分析. 如 MGSA[57]将所有的待测基因功能集作为一个整体代入贝叶斯网络进行富集分析; 由于贝叶斯网络建模时已经将基因功能集的重叠情况考虑在内, MGSA 可以避免对每个基因功能集进行独立富集分析时产生的冗余性问题. 然而, 在实际应用时, 该方法由于其较高的复杂程度导致较低的计算效率, 而检验的灵敏度较低, 因而应用不广. 另一类解决冗余性的方法是对获得的富集基因功能集进行聚类和过滤. 如 REVIGO 依赖语义相似度采用聚类算法从富集结果中众多 GO 条目里面找到最具代表性的子条目输出. LEGO 也提供了一种依赖于网络的对基因功能集的聚类-过滤(cluster and filter)方法, 首先把基因功能集按照互相之间共同基因的重叠程度构建一个网络, 再利用网络模块划分的方法得到一系列基因功能集模块, 使得每个模块内部的功能集具有较高的相似度. 这样, 在获得富集的基因功能集后, LEGO 把这些结果按照之前的聚类结果进行分类, 再选取其中最显著的基因功能集作为该模块的标志基因功能集. 还有一种解决冗余性的方法就是对基因功能集进行过滤, 降低功能集之间的相似度. 如 GO 数据库针对特定物种提供了过滤后的 GO 条目数据库——GO slims, 综合多个相似的 GO 条目从而得到少量的 GO 条目.

## 4 标准数据集和方法评估

目前研究者已开发了相当多的功能富集分析算法和工具. 面对如此多的方法, 使用者往往无从下手. 因而, 有必要建立一套合适的评价标准来对富集分析方法进行综合客观的评估, 从而有针对性地选择合适的方法.

一个理想的功能富集分析方法应该能够灵敏地检测到靶通路($P$ 值低)并且靶通路的排名(rank)比较靠前, 此外应该控制好假阳性率(false positive rate). 为此, 用于评估功能富集分析方法的标准数据集(benchmark datasets)应具有以下性质: 每个数据集应有注释的靶通路(金标准); 标准数据集中包含的数据集应具有多样性和大样本的特征: 多样性意味着每个数据集的靶通路之间的相关性较低, 大样本则要求具有一定量的数据集. 对于一个方法来说, 灵敏度与精确度不可兼得, 同时较高的灵敏度也会导致较高的假阳性问题. 因此在建立好标准数据集后, 研究者可以对不同的富集分析方法从灵敏度、精确度及特异度多个方面进行客观的比较. 为此, Tarca 等人整合了 42 个基因表达数据集来建立了一套标准数据集, 其中每个数据集都有对应的一条已知来自于 KEGG 或 Metacore 疾病数据库的靶通路. 利用该数据集, Tarca 等人对 16 种 FCS 方法进行了比较, 从靶通路在所有 KEGG 通路中 $P$ 值的大小、排名以及假阳性率 3 个方面对这些方法进行了评估. Bayerlová 等人[58], Dong 等人也利用了该标准数据集中靶通路为 KEGG 通路的 36 个基因表达数据集对 ORA, FCS, PT 及 NT

方法进行了系统的比较.

由于 ORA 方法计算简便, 耗时少, 并且仅需要输入一组基因, 因此应用范围最广, 比较适合研究人员简单初步地分析结果. FCS 方法则要求输入基因的表达谱信息, 其灵敏度、精确度均优于 ORA, 更容易检测出发生细微改变的信号. 对于 FCS 方法中的两种检验, 自足型检验的灵敏度优于竞争型检验, 因此, 如果希望最大程度地富集出更多显著的基因集, 且样本量较大时, 应使用自足型检验进行样本抽样, 这样可以最大程度地保留基因间的相互关系, 但是, 较高的灵敏度也会导致较高的假阳性率; 而当样本数较少(如对照组和实验组仅有 2~3 个样本)且基因间相互关系较弱的时候, 适合采用竞争型方法进行基因抽样, 精确度较高, 可以较为准确地富集出真正具有生物学意义的基因集. 总体而言, 传统的 ORA 或者 FCS 方法已经足够检测出显著富集的基因集. 虽然 PT 方法考虑了通路间的拓扑结构, 但是由于目前数据库中通路的拓扑结构信息不够完整, 总在不断更新, 同时不同通路的拓扑结构在不同的物种、细胞、组织、实验条件均不相同, 处理起来较为复杂, 导致 PT 方法不够灵活, 并且评估结果显示, PT 方法并不显著优于 FCS 方法, 因此实用度不高. 导致这种结果的原因是由于通路本身存在一定的冗余性问题, 因此当通路之间没有重叠基因时, 或解决 PT 方法中通路的冗余性问题, 可以一定程度提高该方法的灵敏度. NT 方法考虑了基因在生物学网络中的重要性及相互关系, 可以富集出在统计学上显著、且具有真正生物学意义的基因集, 是目前最新且主流的富集分析方法. 评估结果显示, NT 方法综合表现(灵敏度、精确度、特异度)较好, 因此在有合适的生物学网络时, 推荐使用 NT 方法.

## 5 结论

高通量实验手段的广泛应用可以得到全基因组范围内的各种组学数据, 通过统计分析方法, 根据基因所参与的生物通路的功能注释信息, 发现其中显著富集的生物学功能可从数据中揭示生物学分子机制问题, 从而服务于基础生物医学研究、应用临床医学、药物开发及个性化精准医疗等方面. 本文对基因功能富集分析方法进行了分类评述. 需要注意的是, 任何方法都没有绝对的“好坏”之分, 每个方法都有自己的优点和一定适用范围, 研究者应在对富集分析方法有一定了解的基础上, 根据研究目的和需求, 选择最为合理的方法. 此外, 本文还探讨了功能富集分析结果的冗余性问题及建立标准数据集的必要性.

## 参考文献


1 Naimi A I, Westreich D J. Big data: a revolution that will transform how we live, work, and think. Inform Commun Soc, 2013, 17: 181–183

2 Mooney M A, Nigg J T, McWeeney S K, et al. Functional and genomic context in pathway analysis of GWAS data. Trends Genet, 2014, 30: 390–400

3 Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25–29

4 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30

5 Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res, 2014, 42: D199–D205

6 Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res, 2005, 33: D428–D432

7 Nishimura D. The Computer Software Journal for Scient. BioCarta. Biotech Software & Internet Report, 2001, 2: 117–120

8 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: 15545–15550

9 Rahmatallah Y, Emmert-Streib F, Glazko G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. Brief Bioinform, 2015, in press

10 Jiao X, Sherman B T, Huang da W, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics, 2012, 28: 1805–1806

11 Beissbarth T, Speed T P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics, 2004, 20:

1464–1465
12 Doniger S W, Salomonis N, Dahlquist K D, et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol, 2003, 4: R7
13 Zeeberg B R, Feng W, Wang G, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol, 2003, 4: R28
14 Draghici S, Khatri P, Martins R P, et al. Global functional profiling of gene expression. Genomics, 2003, 81: 98–104
15 Mootha V K, Lindgren C M, Eriksson K-F, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet, 2003, 34: 267–273
16 Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat, 2007, 1: 107–129
17 Tarca A L, Draghici S, Bhatti G, et al. Down-weighting overlapping genes improves gene set analysis. BMC Bioinformatics, 2012, 13: 136
18 Barry W T, Nobel A B, Wright F A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics, 2005, 21: 1943–1949
19 Goeman J J, Van De Geer S A, De Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 2004, 20: 93–99
20 Goeman J J, Oosting J, Cleton-Jansen A-M, et al. Testing association of a pathway with survival using gene expression data. Bioinformatics, 2005, 21: 1950–1957
21 Tian L, Greenberg S A, Kong S W, et al. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA, 2005, 102: 13544–13549
22 Luo W, Friedman M S, Shedden K, et al. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics, 2009, 10: 161
23 Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics, 2013, 14: 7
24 Tomfohr J, Lu J, Kepler T B. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics, 2005, 6: 225
25 Lee E, Chuang H Y, Kim J W, et al. Inferring pathway activity toward precise disease classification. PLoS Comput Biol, 2008, 4: e1000217
26 Barbie D A, Tamayo P, Boehm J S, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature, 2009, 462: 108–112
27 Michaud J, Simpson K M, Escher R, et al. Integrative analysis of RUNX1 downstream pathways and target genes. BMC Genomics, 2008, 9: 363
28 Mansmann U, Meister R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. Methods Inf Med, 2005, 44: 449–453
29 Wu D, Smyth G K. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res, 2012, 40: e133–e133
30 Draghici S, Khatri P, Tarca A L, et al. A systems biology approach for pathway level analysis. Genome Res, 2007, 17: 1537–1545
31 Tarca A L, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. Bioinformatics, 2009, 25: 75–82
32 Glaab E, Baudot A, Krasnogor N, et al. TopoGSA: network topological gene set analysis. Bioinformatics, 2010, 26: 1271–1272
33 Gu Z, Liu J, Cao K, et al. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC Syst Biol, 2012, 6: 56
34 Ihnatova I, Budinska E. ToPASeq: an R package for topology-based pathway analysis of microarray and RNAseq data. BMC Bioinformatics, 2015, 16: 350
35 Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. J Comput Biol, 2009, 16: 407–426
36 Loi S, Haibe-Kains B, Desmedt C, et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC Genomics, 2008, 9: 1
37 Isci S, Ozturk C, Jones J, et al. Pathway analysis of high-throughput biological data within a Bayesian network framework. Bioinformatics, 2011, 27: 1667–1674
38 Mieczkowski J, Swiatek-Machado K, Kaminska B. Identification of pathway deregulation—gene expression based analysis of consistent signal transduction. PLoS One, 2012, 7: e41541
39 Alexeyenko A, Lee W, Pernemalm M, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks.

BMC Bioinformatics, 2012, 13: 226
40 Glaab E, Baudot A, Krasnogor N, et al. EnrichNet: network-based gene set enrichment analysis Bioinformatics, 2012, 28: i451–i457
41 Fang Z, Tian W, Ji H. A network-based gene-weighting approach for pathway analysis. Cell Res, 2012, 22: 565–580
42 Dong X, Hao Y, Wang X, et al. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. Sci Rep, 2016, 6: 18871
43 Wang J, Huang Q, Liu Z-P, et al. NOA: a novel Network Ontology Analysis method. Nucleic Acids Res, 2011, 39: e87
44 Chang B, Kustra R, Tian W. Functional-network-based gene set analysis using gene-ontology. PLoS One, 2013, 8: e55635
45 Khatri P, Sirota M, Butte A J. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol, 2012, 8: e1002375
46 Goeman J J, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics, 2007, 23: 980–987
47 Nam D, Kim S-Y. Gene-set approach for expression pattern analysis. Brief Bioinform, 2008, 9: 189–197
48 Tarca A L, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One, 2013, 8: e79217
49 Laukens K, Naulaerts S, Berghe W V. Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis. Proteomics, 2015, 15: 981–996
50 Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26: 139–140
51 Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol, 2010, 11: R106
52 Law C W, Chen Y, Shi W, et al. Voom precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol, 2014, 15: R29
53 Prasad T K, Goel R, Kandasamy K, et al. Human protein reference database—2009 update. Nucleic Acids Res, 2009, 37: D767–D772
54 Schmitt T, Ogris C, Sonnhammer E L. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res, 2014, 42: D380–D388
55 Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res, 2011, 39: D561–D568
56 Supek F, Bošnjak M, Škunca N, et al. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One, 2011, 6: e21800
57 Bauer S, Gagneur J, Robinson P N. GOing Bayesian: model-based gene set analysis of genome-scale data. Nucleic Acids Res, 2010, 38: 3523–3532
58 Bayerlová M, Jung K, Kramer F, et al. Comparative study on gene set and pathway topology-based enrichment methods. BMC Bioinformatics, 2015, 16: 334

# Progress in Gene Functional Enrichment Analysis

WANG Xiao[1], YIN TianShu[1], LI BoYi[1], JIANG XiLin[1], SUN Hui[1], DOU YaGuang[1], NI Qi[1] & TIAN WeiDong[1,2]

*1 Department of Biostatistics and Computational Biology, Fudan University, Shanghai 200438, China;*
*2 Children's Hospital of Fudan University, Shanghai 201102, China*

Gene functional enrichment analysis has become a common procedure in high-throughput omics data analysis and plays a vital role in revealing molecular mechanisms in biomedical sciences. Hundreds of different gene functional enrichment methods and tools have been developed. In accordance with the problems to be solved and the principle of algorithms, these methods can be approximately classified into four categories, including over-representation analysis, functional class scoring, pathway topology, and network topology. In this article, we review the principles of these four main categories and examples of commonly used approaches. We discussed the redundancy in the results of gene functional enrichment analysis and the necessity to build benchmark datasets.

**omics data, functional enrichment, redundancy, benchmark datasets**

doi: 10.1360/N052016-00139