·综 述·

DOI:10.16605/j.cnki.1007-7847.2016.04.011

用于 GWAS 结果后续研究的 PBA 方法简介

张远森, 司天昭, 杨 恩*
(昆明理工大学 生命科学与技术学院, 中国云南 昆明 650500)

摘 要: 自提出全基因组关联研究(genome-wide association study, GWAS)设想以来, 在人类复杂疾病和水稻农艺性状关联研究方面, GWAS 已得到广泛运用。但作为一种典型的单标记研究方法, GWAS 不能检测小效应的遗传变异, 而稀有变异间的联合效应往往与表型密切相关, 因此, 需对 GWAS 结果进行深入的数据挖掘。基于通路的分析方法(pathway-based analysis, PBA)就是利用基因功能、生物代谢通路等相关信息建立的对 GWAS 结果进行二次挖掘的方法。该方法能从 GWAS 结果挖掘出与性状、疾病相关联的通路及具有相同功能的基因集等数据, 从而获得更多的遗传信息。现对 PBA 的出现、计算方法和相关软件进行简要综述, 以期为人们进行通路分析提供参考。

关键词: 全基因组关联分析(GWAS); 基于通路的分析方法(PBA); 单核苷酸多态性; 生物信息学; 疾病; 农艺性状
中图分类号:Q-31 **文献标识码**:A **文章编号**:1007-7847(2016)04-0345-08

A Brief Review of the PBA Methods for Follow-up Study of the GWAS Results

ZHANG Yuan-sen, SI Tian-zhao, YANG En*
(*Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming* 650500, *Yunnan, China*)

Abstract: With the idea of genome-wide association study (GWAS) presented, it has been widely used in rice agronomic traits and human complex disease. GWAS typically focuses on the analysis of single markers, and is unable to cover the relatively small effect sizes of genetic variants. However, the combined effect of rare variants is often closely associated with phenotypes. It requires in-depth data mining from GWAS results. Pathway-based analysis (PBA) method has been developed, which uses prior biological knowledge on gene function and biological metabolic pathways. By using PBA, more information about the pathway and gene sets with same functions which are associated with the diseases or traits from GWAS result could be obtained. In order to offer more reference to pathway-based analysis, the development, methods and related software of the PBA were introduced.

Key words: genome-wide association study (GWAS); pathway-based analysis (PBA); single nucleotide polymorphism (SNP); bioinformatics; disease; agronomic traits

(*Life Science Research*, 2016, 20(4): 345~352)

基于通路的分析方法(pathway-based analysis, PBA) 是为分析基因表达数据而发展起来的用于全基因组关联研究(genome-wide association study, GWAS)后续分析的一种方法, 通过结合 GWAS 结果与一个已知的分子通路来检测这个通路是否与疾病、农艺性状相关联[1, 2]。将 GWAS 所有分型的单核苷酸多态性(single nucleotide polymorphisms, SNPs)位点按照不同的生物学通路排列, 比较各个通路在病例/对照间的差异, 这就是基于通路的 GWAS 研究思路, 是对 GWAS 结果进行的二次分析, 将帮助我们从一系列的 GWAS 结果中优化出更有用的信息。本文将对 PBA 的出现、方法和相

收稿日期: 2015-12-24; **修回日期**: 2016-03-18
基金项目: 云南省昆明理工大学人才培养项目(14118480)
作者简介: 张远森(1989-), 男, 云南昭通人, 硕士研究生; * 通讯作者: 杨恩(1982-), 女, 贵州兴义人, 博士, 昆明理工大学讲师, 主要从事生物信息学研究, Tel: 0871-65920759, E-mail: enen_yang@126.com。

关软件进行简要综述, 以期为人们进行通路分析提供参考。

1 PBA 发展背景

关联分析是一种用数量性状基因座(quantitative trait locus, QTL) 来鉴定不同基因引起的遗传变异以及挖掘有效等位基因的分析方法[3]。GWAS 最先由 Risch 等[4]提出, 它是指在人类全基因组范围内找出存在的序列变异, 即单核苷酸多态性, 从中筛选出与疾病相关的 SNPs。自 2005 年 SCIENCE 杂志报道了第一项关于视网膜黄斑变性的 GWAS 研究[3]以来, GWAS 在人类复杂疾病研究方面做出了巨大贡献。Deng 等[5]对感染枯氏锥虫引起的 Chagas 心肌病做了 GWAS 研究, 两个与心肌病高度相关的 SNP 位点 (rs4149081 和 rs12582717, $P<10^{-6}$)定位于 12p12.2 染色体的 *SLCO1B1* 基因上, 并且与其他方法比, 该方法多检测出 44 个关联 SNPs。Cheng 等[6]运用 GWAS 分析发现了可能增加女性患子宫内膜癌风险的 5 个新基因区域, 使我们能进一步了解子宫内膜癌发生的遗传驱动因素。此外, 该研究还探讨了这些基因变异增加癌症风险的可能原因, 对子宫内膜癌的治疗具有一定的提示意义。Hou 等[7]对来自 ConLiGen 的 22 个基因位点做了 GWAS 研究, 发现 21 号染色体上的 4 个连接位点 SNPs (rs79663003、rs78015114、rs74795342和 rs75222709, $P<10^{-6}$) 的单基因座与 lithium 治疗反应相关。该研究确定 lithium 治疗反应的生物标志物将构成双相情感障碍临床管理的重要一步。Gage 等[8]对疾病进行了 GWAS 研究, 发现此方法可以识别能预测增加疾病风险行为的遗传变异, 可用于寻找疾病的潜在可修饰危险因素, 这样医务人员可以有针对性地进行干预,有助于识别疾病的生活方式风险。在植物研究中, GWAS 也得到了一定的运用[9]。Huang 等[10]用GWAS 方法分析了地方籼稻(*Oryza sativa indica*)的 14 种水稻农艺性状。在此基础上, 该团队通过增加材料数目及改进表型鉴定方法, 使得 GWAS 检测基因座的能力得到提升, 并鉴定出 32 个新关联位点[11]。虽然 GWAS 已经深入到动植物研究领域, 但是它仍具有一定局限性, 在 NEW ENGLAND JOURNAL OF MEDICINE 杂志上针对 GWAS 给予如下评论[12~15]: GWAS 只是基因识别过程中的第一步, 如果把复杂疾病遗传学看做拼图, 那么 GWAS 只是把菱角铺好, 还需进行深入研究。Kraft 等[13]提出了 meta 分析, 他们认为 meta 分析可以提供更稳定的风险评估及相应的临床信息。Goldstein[14]则认为 meta 分析缺乏足够的效应值来解释与评估遗传可能性及复杂疾病的相关风险。他认为是 SNPs 间的交互作用增加了复杂疾病的风险率。最后, Hirschhorn[15]在此基础上提出了基于通路的研究方法, 即 PBA。

2 PBA 方法

目前, 已有多种不同的 PBA 方法相继问世。根据输入数据的不同可以将 PBA 方法大致分为两类: 一类是“*P* 值富集方法”(*P*-value enrichment approach), 旨在判定一组特定的 SNP 位点或基因 *P* 值是否有丰富的关联信号。这种方法虽已经得到普及, 但是结果存在潜在误差[2]; 另一类是“原始基因型方法”(raw genotype approach), 使用个体水平的 SNP 作为输入数据, 从而获得基因水平和通路水平的检验统计量。该方法通常以表型数据排列, 以调整通路的统计学意义, 以致计算量巨大[2]。根据原统计假设不同可以将 PBA 方法大致分为“自包含型”(self-contained)和“竞争型”(competitive)两类[2]。自包含型 PBA 方法/工具包括 GRASS[16]和 PLINK set-test[17], 竞争型方法/工具包括 ALIGATOR[18]、i-GSEA4GWAS[19]、GenGen[20]、GESBAP[21]、GSA-SNP[22]、GSEA-SNP[23]和 SNP ratio test[24]。由于一个与性状关联的基因可以参与多个完全不同的通路, 优先关注最可能与性状相关联的通路是最可行的策略, 所以竞争型 PBA 方法被认为是更为有效的方法。而在竞争型 PBA 方法中, 基因集富集分析(gene set enrichment analysis, GSEA)类型的方法/工具(包括 i-GSEA4GWAS、GenGen、GSA-SNP 和 GSEA-SNP)是公认的最有效的方法。

由于现有的 PBA 算法在通路来源、SNP-基因图谱规则、总结基因水平的统计检验量、计算各个通路的富集分值(enrichment score, ES)以及如何评估通路的统计学意义等方面各不相同。现只将已发布的 PBA 方法进行初步分类概括, 以便参阅。

2.1 通路定义和资源

通常“通路”被用来代表一个大范围的生物学过程,包括细胞功能、新陈代谢过程、生物合成和遗传信息过程等。在细胞环境下, 一个“通路”代表一系列具有特定功能的分子行为[1]。通常而言, “通路”泛指参与同一生物学功能或过程的基因[2]。

我们在使用 PBA 分析过程中可以根据自己的要求同时参考生物学过程来定义“通路”。

Pathguide 数据库提供了大量的通路资源[2],且这些通路资源还在迅速发展。其中 Kyoto Encyclopedia of Genes and Genomes (KEGG)[25]、BioCarta 和 Gene Ontology (GO)[26]被大多数研究者所熟知,并且使用频率较高。与 GO 相似的数据库还有 Database for Annotation, Visualization and Integrated Discovery (DAVID)数据库[27]和 Protein ANalysis THrough Evolutionary Relationships (PANTHER)数据库[28]。此外,一些数据库能为 PBA 分析提供基因共表达模式或蛋白质相互作用的信息。如 Molecular Signatures Database (MSigDB)提供了由癌基因附近区域基因表达所定义的一系列基因集[29]。一些商家也提供了一些私人所有的通路数据库。还有许多,如 Science′s Signal Transduction Knowledge Environment[30]是一个适用于细胞信号通路的数据库; MetaCyc[31]是一个新陈代谢通路数据库; TRANSPATH[32]是一个转录调控数据库。当然,也有一些专用于蛋白质之间相互作用的数据库[33],如 BioGRID[34]。对于某些特殊疾病领域,当进行 PBA 分析时可以根据文献信息或生物逻辑手动收集通路数据[2]。

2.2　PBA 算法

PBA 算法由单基因分析方法发展而来,现正在不断的优化中。

常见的单基因分析方法主要有 4 个方面的缺陷,分别是: 1)多假设检验校正后,由于芯片技术固有噪声导致单个基因难于满足统计学意义阀值; 2)与 1)相反,也许会出现很多具有统计学意义的基因,但是这些基因的生物学功能并不一致; 3)可能会错过一些影响重大的通路; 4) 同一个生物系统在两个群体中进行研究时,来自两个群体的具有统计学意义的基因可能不同。为了克服这些分析挑战, Subramanian 等[29]提出了 GSEA 方法。GSEA 通过数据输入、计算 ES、评估 ES 的统计水平和多假设检验校正 4 个步骤来确定一组基因集与表型是否相关。

Zhang 等[19]对 GSEA 方法进行改进,并命名为 improved Gene Set Enrichment Analysis (i-GSEA)。i-GSEA 计算方式是在原有的 ES 基础上乘系数(基因集中显著性基因的比例/全基因组中显著性基因的比例)。i-GSEA 作为 GSEA 的一个延伸,通过执行 SNP 标签来计算在显著性比例下的基因集的 ES 值,而且其趋向于选取包含显著意义基因的基因集,这些基因集包含稀有 SNPs。与 GSEA 相比, i-GSEA 的灵敏度有所提高。

3　用于 PBA 的软件简介

在生物大数据时代背景下,利用、挖掘大数据的方法在迅速发展。GWAS 的成熟带动了相应的生物数据分析软件的发展,如: i-GSEA4GWAS、GenGen、GESBAP、GSA-SNP、GSEA-SNP、PLINK set-test 和 SNP ratio test 等。本文就几个常见的软件进行介绍。

3.1　i-GSEA4GWAS

i-GSEA4GWAS 是使用 i-GSEA 算法来鉴定 GWAS 结果中关联通路的一种在线工具(http://gsea4gwas.psych.ac.cn/)[19]。用户不需要事先注册账号就可以通过在线命令的方式免费使用 i-GSEA4GWAS。

i-GSEA4GWAS 在线工具只需要输入 SNP 数据、SNP ID 和 SNP *P* 值。程序可自动将 *P* 值转化为 log(*P* 值)而直接运用。为了确保收集的通路/基因集更广泛更有质量, i-GSEA4GWAS 在线工具采用 MSIGDB v2.5 软件提取通路信息。这些最权威的通路/基因集都来自于 KEGG、BioCarta 和 GO 在线资源的整合。此外,用户也可以使用自己定义的通路。

i-GSEA4GWSA 程序输出的是 FDR 小于 0.25 的通路。输出特征是一个 Manhattan plot, 它可以帮助用户形象地比较大规模通路的关联结果。在结果页面除了文本信息和图表外还有下载网址。此外,在程序运行前用户可以输入自己的邮箱, i-GSEA4GWSA 程序会把结果直接发送到邮箱。

Zhang 等运用 i-GSEA4GWSA 在线工具分别对 HIV-1 病毒载量 GWAS 数据[19]及双相情感障碍(bipolar disorder, BD) GWAS 结果进行二次分析[35]。结果表明:ribosomal 通路 (FDR=0.047)、st myocyte ad pathway 通路(FDR=0.040)和 1, 4, 5-三磷酸肌醇受体基因 *ITPR*1 都与 HIV-1 病毒载量相关。对双相情感障碍 GWAS 结果的分析显示: 谷胱甘肽代谢通路(FDR=0.036)和半胱氨酸内切酶活性通路(FDR=0.041)与之相关。

3.2　GenGen

基因组遗传分析(genetic genomics analysis of complex data, GenGen) 主要是针对遗传学中复杂

数据进行基因组学分析的一系列软件包, 可免费下载 (http://www.openbioinformatics.org/gengen/)[20]。GenGen 是以大类疾病调查为导向来鉴定可能的致病通路, 在复杂疾病病因学研究方面有重要意义。软件用 perl 语言书写, 但是一些程序要求使用 C 语言, 使用时需 32 位/64 位的 linux 系统, 因此许多平台不一致, 有些特殊的系统结构也不相匹配, 从而, 运行程序时需不时对原始代码进行重写。

GenGen 包含有 6 个主要的程序, 其中使用 calculate_gsea.pl 程序执行 PBA 分析。该程序着眼于 GWAS 结果中所有 SNPs: 首先检测在一个特定生物学通路中是否存在功能一致的 SNPs, 然后使用 GSEA 算法从候选通路中选出最显著的通路, 最后计算通路的统计学意义。完成整个程序需要大量的计算时间。

GenGen 程序要求 3 个主要的输入文件, 一个是关联结果文件, 一个是关联序列结果文件, 一个是 SNP-基因图谱文件或者通路定义文件。关联结果文件是由一个简单的制表符分隔为两列的文件, 第一列是 SNPs 的 ID, 第二列是对应的 *P* 值或者是 χ^2 检测值。关联序列结果文件包含所有序列的所有 SNP 位点的检验统计量, 每行含有 10 列, 前三列是对 SNP 的一般描述, 如 SNP 的 ID、所在染色体、位置, 而其余列是数据集信息, 如 SNPs 等位基因 A：B、病例组、对照组 χ^2 值、*P* 值的 χ^2 值。SNP-基因图谱文件每行有三列: SNP 的 ID、基因 ID 和 SNP 与基因的距离。通路定义文件每行就是一个通路, 前两列分别是通路 ID 和通路描述, 随后是通路中鉴定出的基因。

GenGen 程序使用竞争型统计检测方法。输出结果是每个通路的正常 *P* 值和 FDR 值。运行 GenGen 程序时可以进行参数调整, 如改变关联结果文件、改变病例对照研究的检测策略以及改变 SNP-基因图谱文件。此外, 通路注释文件可以补充, 甚至可以被取代。

Perry 等[36]运用 GenGen 程序对 2 型糖尿病 GWAS 数据进行分析, 发现在胰腺中表达的 *CCND2*、*SMAD3* 和 *PRICKLE1* 基因组成的基因集与 2 型糖尿病相关。

3.3 GESBAP

基于基因集多态性分析(gene set-based analysis of polymorphisms, GESBAP)是为解决 GWAS 在实际运用中检测方法使用受限而发展起来的一种基因集分析软件[21](用 Java 和 C++语言编写, http://bioinfo.cipf.es/gesbap/), 其支持原理是靶标功能基因受损时会影响最终的表型。GESBAP 默认的工作模式是传统的“匿名用户”, 该模式下的结果只能保留一天; 也允许用户注册账号使用, 这样能较长时间保留结果。

GESBAP 程序需要输入一个有 SNP ID 及其对应关联 *P* 值的文件。由于任何一个常见的基因标示符都可以使用 Babelomics 软件包将其转换为基因 ID[37], 所以 GESBAP 能接受多种形式的数据, 如拷贝数变异 (copy number variations, CNVs) 及其对应的关联 *P* 值文件、基因及其对应的关联 *P* 值文件。此外, GESBAP 还能接受 PLINK 程序执行关联检测获得的 *P* 值。目前, GESBAP 程序只能用于人类、小鼠和大鼠的 SNPs 分析。用户可以从 GO、KEGG 和 BioCarta 通路数据库中选择一个或者多个功能类别不同的数据库进行分析。通路中基因将由多个过滤器使用关键词和设定基因数量范围来进行筛选。

GESBAP 程序应用竞争型假设检验, 分析时将 *P* 值最低的 SNP 赋给基因, 使用基因集分析 (gene set-based analysis, GSA) 找出基因列表中具有关联意义的一组基因集, 计算 ES。这些功能基因集所对应的 *P* 值会以列表形式输出。此外, 输出的还有这些功能基因集的功能类别信息。Medina 等[21]运用 GESBAP 程序对乳腺癌 GWAS 结果进行分析发现, 与乳腺癌相关的两个通路, 即跨膜受体蛋白酪氨酸激酶信号传导通路和信号转导调节通路中都含有一种编码酪氨酸激酶受体的 *FGFR2* 基因。

3.4 GSA-SNP

GSA-SNP 是基于 JAVA 程序设计语言开发的一款软件, 是一个计算快速并且易于使用的通路分析工具(http://gsa.muldas.org.), 同时也是一个能用于病例/对照研究和数量性状研究的卓越工具[22]。该软件通过 Z 统计方法[38]、标准的 GSA 方法[39]和 GSEA 方法[28]3 个不同的基因集分析方法执行通路分析。

GSA-SNP 需要输入 marker 关联数据(marker association data), marker 关联数据主要是 SNPs 数据。GSA-SNP 程序还需要输入基因集数据(gene set data): 在默认情况下使用 GO 数据库作为通路资源。用户也可以上传 MSigDB 形式(http://www.broadinstitute.org/gsea/msigdb/)的通路定义数据。此外, 该程序能自动识别数据类型并为用户推荐

基因集分析方法和参数选择。分析过程中, 程序会根据关联信号由强到弱的顺序输出与性状相关联的基因集及其对应的 *P* 值。

Nam 等[22]运用 GSA-SNP 软件成功分析了来自 Korea Association Resource (KARE)课题[40]的 Korea 地区 8 842 个招募者的基因型数据, 研究发现 *COLLAGEN*、*GOLGI_STACK* 等 12 个基因组成的基因集与身高显著相关, 但与欧洲人相比显著性较弱[41]。

3.5 GSEA-SNP

GSEA -SNP (http://www.nr.no/pages/samba/area_emr_smbi_gseasnp.) 是为解决 GWAS 过多假阳性而发展起来的基于通路的基因集富集分析软件, 该软件需要使用 R 语言来执行程序。其结果依赖于一个假设, 即功能性 SNPs 富集在一个通路中。GSEA-SNP 有助于 SNPs 检测和通路鉴定以及了解潜在的生物学机制[23]。

GSEA-SNP 对用于基因表达数据分析的 GSEA 方法[35]做了两个方面的改进: 一是将基因 list 变为 SNP list; 二是利用基于等位基因或者是基于基因型的统计方法来判定 SNP 关联程度的大小。Neibergs 等[42]运用 GSEA-SNP 工具鉴定了牛的副结核病相关基因, 研究表明细胞运动正调控基因集(*EDN2*、*TDGF1*、*TGFB2* 和 *PIK3R1*)与副结核病相关。

3.6 PLINK

PLINK 是用 C 和 C++语言编写的一个用于全基因组关联分析的软件集 (http://pngu.mgh.harvard.edu/~purcell/plink)[17]。除了用于关联分析外, 还能用于数据管理、概要统计和种群聚类分析等。PBA 分析时主要运用的是 PLINK 中的基于基因集的关联检测方法(PLINK set-based tests)。

PLINK set-based tests 由于使用排列组合方法, 特别适用于大规模的候选基因的研究。其要求输入原始基因型数据, 使用的是自包含型假设, 分析策略是首先通过预设一个的 *P* 值范围, 然后在通路中选取具有统计意义的 *P* 值来计算每个通路的平均 ES。

Passtoors 等[43]运用 PLINK set-based 统计分析发现, 哺乳动物雷帕霉素靶蛋白(mammalian target of rapamycin, mTOR) 通路不仅与疾病调控有关,还与人类健康和长寿相关。

3.7 SNP ratio test

SNP ratio test (SRT)的基本方法是运用通路中所有显著 SNP 与非显著 SNP 的比值, 然后再以 GWAS 结果为基础通过随机排列比较这些比值的分布 (http://sourceforge.net/projects/snpratiotest)[24]。SRT 与 GSEA[35]和 PLINK[17]相似,都使用经验 *P* 值检测通路中所有关联 SNPs 富集程度。其优点在于: 1) SRT 以 PLINK 输入和 PLINK 输出形式运行, 执行起来十分简单; 2) SRT 使用一个模拟数据集来评估给定通路的显著性; 3) SRT 接受 PLINK 形式的输入文件, 同时用户也可以输入自定义的表型数据集。但是, 由于 SRT 使用通路中所有的 SNPs, 这样会扩大通路的关联度, 从而产生假阳性。此外, 由于 SRT 是在通路水平下执行的, 所以需要对通路和基因大小等因子进行调整。Dushlaine 等[24]使用 SRT 对帕金森病的 GWAS 数据进行分析, 鉴定出了更多的关联 SNPs。

需要注意的是, 由于通路间并非相互独立, 所以通路水平下的 *P* 值仍需要多假设检验进行校正, 而 SRT 在通路水平下并没有相应的多假设检验。尽管如此, 相比 SNP 水平分析, 通路水平下的 GWAS 数据分析多样性问题已经大大减少。

3.8 ALIGATOR

关联表注释者(Association LIst Go AnnoTatOR, ALIGATOR)[18]旨在鉴定潜在的致病基因集。该软件接受来自任何 GWAS 研究平台的数据, 使用竞争型假设检验, 运行 ALIGATOR 时需要输入 SNP *P* 值数据。同其他 PBA 分析软件一样, ALIGATOR 也需要进行多假设检验校正。ALIGATOR 软件的不足之处是计算量巨大, 个体基因型数据有时候不可用。Holmans 等[18]运用 ALIGATOR 方法对克罗恩疾病(Crohn disease, CD)和 BD 疾病 GWAS 数据进行分析, 表明细胞活性、转录调控和生理功能在 BD 发病机制中扮演着重要角色。

3.9 GRASS

GRASS 即关联研究中的基因集岭回归分析(gene set ridge regression in association studies, GRASS)[16]。在 PBA 分析中 GRASS 用于总结每一个基因的遗传结构, 选择一个或多个“特征 SNPs”来代表一个基因并确定基因集与性状间的关联度。GRASS 要求输入原始基因集数据, 使用自包含型原假设检测。GRASS 的使用不受 SNP 限制, 其将离基因最近的 SNP 图谱给基因, 这是GRASS 的一个显著优点。Chen 等[16]应用 GRASS 对结肠癌 GWAS 数据进行分析, 表明盐酸盐代谢通路、烟碱代谢通路和转化生长因子 β 信号通路与结肠

癌显著相关。

3.10 PlantGSEA

Plant GeneSet Enrichment Analysis Toolkit (PlantGSEA)是由中国农业大学 Su Zhen 实验室基于 GSEA 方法开发的一个主要用于植物基因集富集分析的在线工具集(http://structuralbiology.cau.edu.cn/PlantGSEA/index.php)[44]。目前 PlantGSEA 从注释系统(如 GO、KEGG)、公共数据库(如 TAIR[45]、RGAP[46])、出版文献等不同资源中收集了 20 290 个定义的基因集合, 并用于 GSEA 分析。这些基因集可以分为 GO 基因集(G1)、基于基因家族的基因集(G2)、精选基因集(G3) 和 Motif 基因集(G4)。用户只需要向 PlantGSEA 提交基因座 ID 或 Affymatrix 微阵列探针集 IDs, PlantGSEA 就可提供超几何检测、Fisher′s 检测和卡方检测 3 种统计方法进行数据处理, 其结果可以保存 3 个月。

我们使用韩斌研究员课题组对 14 个农艺性状的 GWAS 结果[10]作为研究对象, 从中选取 555 个 *P* 值小于 2E-6 的粒宽基因用于 PlantGSEA 分析, 并使用 GO 和 KEGG 基因集作为参考数据, 以全基因组水平作为背景。运行结果显示有 60 个通路与水稻粒宽性状相关联, 其中 FDR 小于 0.001 的有 21 个(表 1)。

除上述软件外, 用于 PBA 分析的软件和算法还有很多。如 Torkamani 等[47]2008 年推出的超几何分布检测 (hypergeometric test); 2010 年 De la Cruz 等[48]推出了 *P* 值联合算法(*P*-value combination approach); Tintle 等[49]于 2009 年推出了 SUMSTAT 方法; Fehringer 等[50]在原 SUMSTAT 方法的基础上进行改进推出了 mSUMSTAT 方法; Jia 等[51]介绍了 Network-assisted 方法。此外, 也有一些用于植物 GWAS 结果的总结方法在不断出现, 如 PICARA 可以从 GWAS 信号中高效地搜索出植物数量性状关联基因, 而且 Chen 等[52]已经成功将其

表 1 水稻粒宽关联基因的 PlantGSEA 分析结果

Table 1 The PlantGSEA analysis results of rice grain width agronomic traits

Gene set name (No. of genes)	Description	Category	No. of genes in overlap	*P* value	FDR
Cellular component (45 586)	GO: 0005575 cellular_component, GOslim: cellular_component	GO_CC	451	1.3E-29	6E-27
Molecular function (45 586)	GO: 0003674 molecular_function, GOslim: molecular_function	GO_MF	451	1.3E-29	2E-26
Biological process (45 586)	GO: 0008150 biological_process, GOslim: biological_process	GO_BP	451	1.3E-29	3E-26
Cell (5 550)	GO: 0005623 cell, GOslim: cellular_component	GO_CC	93	1.5E-11	2E-09
Cell part (5 550)	GO: 0044464 cell part, GOslim: cellular_component	GO_CC	93	1.5E-11	4E-08
Cellular process (7 980)	GO: 0009987 cellular process, GOslim: biological_process	GO_BP	118	4E-11	4E-08
Binding (11 705)	GO: 0005488 binding, GOslim: molecular_function	GO_MF	154	9E-11	8E-08
Catalytic activity (9 113)	GO: 0003824 catalytic activity, GOslim: molecular_function	GO_MF	127	2.6E-10	2E-07
Metabolic process (9 541)	GO: 0008152 metabolic process, GOslim: biological_process	GO_BP	126	8.6E-09	6E-06
Primary metabolic process (7 728)	GO: 0044238 primary metabolic process, GOslim: biological_process	GO_BP	107	1.6E-08	9E-06
Intracellular (3 733)	GO: 0005622 intracellular, GOslim: cellular_component	GO_CC	61	1.9E-07	2E-05
Macromolecule metabolic process (6 374)	GO: 0043170 macromolecule metabolic process, GOslim: biological_process	GO_BP	88	4.8E-07	2E-04
Localization (1 601)	GO: 0051179 localization, GOslim: biological_process	GO_BP	34	5.6E-07	2E-04
Cellular metabolic process (6 576)	GO: 0044237 cellular metabolic process, GOslim: biological_process	GO_BP	89	9.3E-07	3E-04
Establishment of localization (1 582)	GO: 0051234 establishment of localization, GOslim: biological_process	GO_BP	33	1.2E-06	3E-04
Transport (1 582)	GO: 0006810 transport, GOslim: biological_process	GO_BP	33	1.2E-06	3E-04
Organelle (2 468)	GO: 0043226 organelle, GOslim: cellular_component	GO_CC	44	1.3E-06	9E-05
Intracellular organelle (2 468)	GO: 0043229 intracellular organelle, GOslim: cellular_component	GO_CC	44	1.3E-06	9E-05
Membrane (2 063)	GO: 0016020 membrane, GOslim: cellular_component	GO_CC	38	3.4E-06	2E-04
Carbohydrate metabolic process (859)	GO: 0005975 carbohydrate metabolic process, GOslim: biological_process	GO_BP	22	3.5E-06	8E-04
Intracellular part (3 107)	GO: 0044424 intracellular part, GOslim: cellular_component	GO_CC	50	4.1E-06	2E-04

应用于玉米 GWAS 分析结果的后续研究。

4　结语

随着 GWAS 的成熟, GWAS 数据也在快速增长, 对大数据的挖掘方法层出不穷。PBA 方法发展尤为突出, 被全面应用于疾病风险相关通路的鉴定。Daneshjou 等[53]运用 PBA 方法成功鉴定与华法林代谢通路相关联的显著变异位点。证明预测服用华法林后能提高 *VKORC*1 和 *CYP*2*C*9 基因的变异能力, 降低血栓栓塞风险。Kim 等[54]对肌萎缩侧索硬化症(amyotrophic lateral sclerosis, ALS) GWAS 数据进行通路水平分析表明, ALS 与细胞组分形成(cellular component organization)和细胞微丝骨架(actin cytoskeleton)有相关性。Deelen 等[55]使用 PBA 方法相关软件 (PLINK set-based 检测、Global 检测、GRASS 检测和 SNP ratio 检测软件)对人类寿命的 GWAS 数据进行分析, 表明人类寿命与胰岛素/类胰岛素生长因子信号(insulin/insulin-like growth factor (IGF-1) signaling, IIS)通路和端粒维持(telomere maintenance, TM)通路相关。而且更多研究表明, 人类寿命与 *AKT*1 基因、*AKT*3 基因、*FOXO*4 基因、*IGF*2 基因、*INS* 基因、*PIK*3*CA* 基因、*SGK* 基因、*SGK*2 基因、*YWHAG* 基因和 *POT*1 基因相关联。我们相信 PBA 方法将被全面应用于疾病风险相关通路的鉴定, 并且将其运用于指导临床治疗也将指日可待。另外, 虽然目前 PBA 方法在植物、微生物方面的研究较少。但是随着 GWAS 的不断发展, 应用于植物或微生物 GWAS 数据挖掘的 PBA 方法或软件也将陆续被开发出来, 未来能更好地将全基因组数据运用于指导育种和优良品种培育工作。

参考文献(References):

[1] CANTOR R M, LANGE K, SINSHEIMER J S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application[J]. American Journal of Human Genetics, 2010, 86(1): 6-22.

[2] WANG K, LI M Y, HAKONARSON H. Analyzing biological pathways in genome-wide association studies[J]. Nature Reviews Genetics, 2010, 11(12): 843-854.

[3] KLEIN R J, ZEISS C, CHEW E Y, *et al*. Complement factor H polymorphism in age-related macular degeneration[J]. Science, 2005, 308(5720): 385-389.

[4] RISCH N, MERIKANGAS K. The future of genetic studies of complex human diseases[J]. Science, 1996, 273(5281): 1516-1517.

[5] DENG X T, SABINO E C, CUNHA-NETO E, *et al*. Genome wide association study (GWAS) of Chagas cardiomyopathy in *Trypanosoma cruzi* seropositive subjects[J]. PLoS One, 2013, 8(11): e79629.

[6] CHENG T H, THOMPSON D J, O'MARA T A, *et al*. Five endometrial cancer risk loci identified through genome-wide association analysis[J]. Nature Genetics, 2016, 48(6): 667-674.

[7] HOU L P, HEILBRONNER U, DEGENHARDT F, *et al*. Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study[J]. The Lancet, 2016, 387(10023): 1085-1093.

[8] GAGE S H, DAVEY SMITH G, WARE J J, *et al*. G=E: what GWAS can tell us about the environment[J]. PLoS Genetics, 2016, 12(2): e1005765.

[9] 涂雨辰, 田云, 卢向阳. 全基因组关联分析在植物中的应用[J]. 化学与生物工程(TU Yu-chen, TIAN Yun, LU Xiang-yang. Application of genome-wide association study for important traits in plant[J]. Chemistry and Bioengineering), 2013, 30(6): 7-10.

[10] HUANG X H, WEI X H, SANG T, *et al*. Genome-wide association studies of 14 agronomic traits in rice landraces[J]. Nature Genetics, 2010, 42(11): 961-967.

[11] HUANG X H, ZHAO Y, WEI X H, *et al*. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm[J]. Nature Genetics, 2012, 44(1): 32-39.

[12] HARDY J, SINGLETON A. Genome wide association studies and human disease[J]. The New England Journal of Medicine, 2009, 360(17): 1759-1768.

[13] KRAFT P, HUNTER D J. Genetic risk prediction-are we there yes? [J]. The New England Journal of Medicine, 2009, 360 (17): 1701-1703.

[14] GOLDSTEIN D B. Common genetic variation and human traits[J]. The New England Journal of Medicine, 2009, 360(17): 1696-1698.

[15] HIRSCHHORN J N. Genomewide association studies-illuminating biologic pathways[J]. The New England Journal of Medicine, 2009, 360(17): 1699-1701.

[16] CHEN L S, HUTTER C M, POTTER J D, *et al*. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data[J]. The American Journal of Human Genetics, 2010, 86(6): 860-871.

[17] PURCELL S, NEALE B, TODD-BROWN K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.

[18] HOLMANS P, GREEN E K, PAHWA J S, *et al*. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder[J]. The American Journal of Human Genetics, 2009, 85(1): 13-24.

[19] ZHANG K L, CUI S J, CHANG S H, *et al*. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study[J]. Nucleic Acids Research, 2010, 38(2): W90-W95.

[20] REICHARDT J K. GEN GEN: the genomic analysis of androgen-metabolic genes and prostate cancer as a paradigm for the dissection of complex phenotypes[J]. Frontiers in Bioscience, 1999, 4: d596-600.

[21] MEDINA I, MONTANER D, BONIFACI N, *et al*. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies[J]. Nucleic Acids Research, 2009, 37(suppl. 2): W340-W344.

[22] NAM D, KIM J, KIM S Y, *et al*. GSA-SNP: a general approach for gene set analysis of polymorphisms[J]. Nucleic Acids Research, 2010, 38(2): W749-W754.

[23] HOLDEN M, DENG S, WOJNOWSKI L, *et al*. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies[J]. Bioinformatics, 2008, 24(23): 2784-2785.

[24] DUSHLAINE C, KENNY E, HERON E A, *et al*. The SNP ratio test: pathway analysis of genome-wide association datasets[J]. Bioinformatics Applications Note, 2009, 25(20): 2762-2763.

[25] KANEHISA M, GOTO S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Research, 2000, 28(1): 27-30.

[26] ASHBURNER M, BALL C A, BLAKE J A, *et al*. Gene Ontology: tool for the unification of biology[J]. Nature Genetics, 2000, 25(1): 25-29.

[27] JR G D, SHERNAN B T, HOSACK D A, *et al*. DAVID: database for annotation, visualization, and integrated discovery[J]. Genome Biology, 2003, 4(9): R60.1-R60.11.

[28] MI H Y, DONG Q, MURUGANUJAN A, *et al*. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium[J]. Nucleic Acids Research, 2010, 38(1): D204-D210.

[29] SUBRAMANIAN A, TAMAYO P, MOOTHA V K, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proceedings of the National Academy of Sciences USA, 2005, 102(43): 15545-15550.
[30] GOUGH N R. Science's signal transduction knowledge environment: the connections maps database[J]. Annals of the New York Academy of Sciences, 2002, 971(1): 585-587.
[31] CASPI R, ALTMAN T, BILLINGTON R, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases[J]. Nucleic Acids Research, 2014, 42(D1): D459-D471.
[32] KRULL M, VOSS N, CHOI C, *et al*. TRANSPATH: an integrated database on signal transduction and a tool for array analysis[J]. Nucleic Acids Research, 2003, 31(1): 97-100.
[33] KLINGSTRON T, PLEWCAZYNSKI D. Protein-protein interaction and pathway databases, a graphical review[J]. Briefings in Bioinformatics, 2011, 12(6): 702-713.
[34] STARK C, BREITKREUTZ B J, REGULY T, *et al*. BioGRID: a general repository for interaction datasets[J]. Nucleic Acids Research, 2006, 34(12): D535-D539.
[35] ZHANG K L, ZHANG L Y, ZHANG W N, *et al*. Pathway-based analysis for genome-wide association studies of schizophrenia to provide new insight in schizophrenia study[J]. Chinese Science Bulletin, 2011, 56(32): 3398-3402.
[36] PERRY J R B, MCCARTHY M I, HATTERSLEY A T, *et al*. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach[J]. Diabetes, 2009, 58(6): 1463-1467.
[37] AL-SHAHOROUR F, CARBONELL J, MINGUEZ P, *et al*. Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments[J]. Nucleic Acids Research, 2008, 36(suppl. 2): W341-W346.
[38] KIM S Y, VOLSKY D J. PAGE: parametric analysis of gene set enrichment[J]. BioMed Central Bioinformatics, 2005, 6: 114.
[39] EFRON B, TIBSHIRANI R. On testing the significance of sets of genes[J]. Annals of Applied Statistics, 2007, 1: 107-129.
[40] CHO Y S, GO M J, KIM Y J, *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits[J]. Nature Genetics, 2009, 41(5): 527-534.
[41] SILVENTOINEN K, SAMMALISTO S, PEROLA M, *et al*. Heritability of adult body height: a comparative study of twin cohorts in eight countries[J]. Twin Research, 2003, 6(5): 399-408.
[42] NEIBERGS H L, SETTLES M L, WHITLOCK R H, *et al*. GSEA-SNP identifies genes associated with Johne's disease in cattle[J]. Mammalian Genome, 2010, 21(7-8): 419-425.
[43] PASSTOORS W M, BEEKMAN M, DEELEN J, *et al*. Gene expression analysis of mTOR pathway: association with human longevity[J]. Aging Cell, 2013, 12(1): 24-31.
[44] YI X, DU Z, SU Z. Plant GSEA: a gene set enrichment analysis toolkit for plant community[J]. Nucleic Acids Research, 2013, 41(W1): W90-W103.
[45] SWAEBRECK D, WILKS C, LAMESCH P, *et al*. The Arabidopsis Information Resource (TAIR): gene structure and function annotation[J]. Nucleic Acids Research, 2008, 36(suppl. 1): D1009-D1014.
[46] OUYANG S, ZHU W, HHAMILTON J, *et al*. The TIGR Rice Genome Annotation Resource: improvements and new features[J]. Nucleic Acids Research, 2007, 35(suppl. 1): D883-D887.
[47] TORKAMANI A, TOPOL E J, SCHORK N J. Pathway analysis of seven common diseases assessed by genome-wide association[J]. Genomics, 2008, 92(5): 265-272.
[48] DE LA CRUZ O, WEN X, KE B, *et al*. Gene, region and pathway level analyses in whole-genome studies[J]. Genetic Epidemiology, 2010, 34(3): 222-231.
[49] TINTLE N L, BORCHERS B, BROWN M, *et al*. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16[J]. BioMed Central Proceedings, 2009, 3: S96.
[50] FEHRINGER G, LIU G, BRIOLLAIS L, *et al*. Comparison of pathway analysis approaches using lung cancer GWAS data sets[J]. PLoS One, 2012, 7(2): e31816.
[51] JIA P, ZHAO Z. Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives[J]. Human Genetics, 2014, 133(2): 125-138.
[52] CHEN C, DECLERCK G, TIAN F, *et al*. PICARA, an analytical pipeline providing probabilistic inference about a priori candidates genes underlying genome-wide association QTL in plants[J]. PLoS One, 2012, 7(11): e46596.
[53] DANESHJOU R, TATONETTI N P, KARCZEWSKI K J, *et al*. Pathway analysis of genome-wide data improves warfarin dose prediction[J]. BioMed Central Genomics, 2013, 14(suppl. 3): S11.
[54] KIM N C, ANDREWS P C, ASSELBERGS F W, *et al*. Gene ontology analysis of pairwise genetic associations in two genome-wide studies of sporadic ALS[J]. Biodata Mining, 2012, 5: 9.
[55] DEELEN J, UH H W, MONAJEMI R, *et al*. Gene set analysis of GWAS data for human longevity highlights the relevance of the insulin/IGF-1 signaling and telomere maintenance pathways[J]. AGE, 2013, 35(1): 235-249.

(上接第 313 页)

[20] CHAMORRO-JORGANES A, ARLDI E, PENALVA L O, *et al*. MicroRNA-16 and microRNA-424 regulate cell-autonomous angiogenic functions in endothelial cells via targeting vascular endothelial growth factor receptor-2 and fibroblast growth factor receptor-1[J]. Arteriosclerosis, Thrombosis, and Vascular Biology, 2011, 31(11): 2595-2606.
[21] AGRA ANDRIEU N, MOTIÑO O, MAYORAL R, *et al*. Cyclooxygenase-2 is a target of microRNA-16 in human hepatoma cells[J]. PLoS One, 2012, 7(11): e50935.
[22] MITCHENER M M, HERMANSON D J, SHOCKLEY E M, *et al*. Competition and allostery govern substrate selectivity of cyclooxygenase-2[J]. Proceedings of the National Academy of Sciences USA, 2015, 112(40): 12366-12371.
[23] 徐练, 李晓龙, 刘印, 等. miR-93-5p 靶向调控 Smad5 表达抑制小鼠 C3H10T1/2 细胞成骨分化的研究[J]. 中国修复重建外科杂志(XU Lian, LI Xiao-long, LIU Yin, *et al*. miR-93-5P suppresses osteogenic differentiation of mouse C3H10T1/2 cells by targeting Smad5[J]. Chinese Journal of Reparative and Reconstructive Surgery), 2015, 29(10): 1288-1294.
[24] 陶象男, 汪忆梦, 宋传旺. miR-20b 直接靶向 3′-UTR 负性调节 VEGF 的表达[J]. 华中科技大学学报(医学版) (TAO Xiang-nan, WANG Yi-meng, SONG Chuan-wang. miR-20b directly targets 3′-untranslated region of VEGF and negatively regulates its expression[J]. Acta Medicinae Universitatis Scientiae et Technologiae Huazhong), 2016, 45(1): 22-26.
[25] BERKLEY K J. A life of pelvic pain[J]. Physiology & Behavior, 2005, 86(3): 272-280.
[26] 方远书, 何忠平, 张辉, 等. 元胡止痛胶囊的含药血清对痛经模型动物的影响[J]. 中国比较医学杂志(FANG Yuan-shu, HE Zhong-ping, ZHANG Hui, *et al*. Effect of the drug containing serum of Yuanhuzhitong capsule in animal model of dysmenorrhea[J]. Chinese Journal of Comparative Medicine), 2010, 20(8): 35-37.
[27] 刘芳, 郑翠红, 黄光英, 等. 针刺对痛经大鼠中枢及外周 β-EP 含量的影响[J]. 浙江中医杂志(LIU Fang, ZHENG Cui-hong, HUANG Guang-ying, *et al*. Effect of acupuncture on central and peripheal β-EP in rats with primary dysmenorrhea[J]. Zhejiang Journal of Traditional Chinese Medicine), 2008, 43(8): 444-446.