PAST

Pathway Association Studies Tool

Adam Thrash¹, Juliet D. Tang², Mason DeOrnellis^{3,4}, Daniel G. Peterson¹, Marilyn L. Warburton⁵

¹Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS, USA

²USDA-FS Forest Products Laboratory, Starkville, MS, USA

³Humanities and Fine Arts Division, East Mississippi Community College, Mayhew, MS, USA

⁴Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS, USA

⁵USDA-ARS Corn Host Plant Resistance Research Unit, Mississippi State, MS, USA

Abstract

Background: In recent years, a bioinformatics method for interpreting GWAS data using metabolic pathway analysis has been developed and successfully used to find significant pathways and mechanisms explaining phenotypic traits of interest in plants. However, the many scripts implementing this method were not straightforward to use, had to be customized for each project, required user supervision, and took more than 24 hours to process data. PAST (Pathway Association Study Tool), a new implementation of this method, has been developed to address these concerns.

Results: PAST is implemented as a package for the R language. Two user-interfaces are provided; PAST can be run by loading the package in R and calling its methods, or by using an R Shiny guided user interface. In testing, PAST completed analyses in approximately one hour by processing data in parallel. PAST has many user-specified options for maximum customization. PAST produces the same results as the previously developed method.

Conclusions: In order to promote a powerful new method of pathway analysis that interprets GWAS data to find biological mechanisms associated with traits of interest, we developed a more accessible and user friendly tool. This tool is more efficient and requires less knowledge of programming languages to use than previous methods. Moreover, it produces similar results in significantly less time. These attributes make PAST accessible to researchers interested in associating metabolic pathways with GWAS datasets to better understand the genetic architecture and mechanisms affecting phenotype.

Keywords

Metabolic pathway analysis, Genome-wide association study (GWAS), maize (*Zea mays* L.)

Background

Genome-wide association study (GWAS) of complex traits in maize and other crops has become very popular to identify regions of the genome that influence these traits [1, 2, 3]. In general, hundreds of thousands of single nucleotide polymorphisms (SNPs) markers are each tested using F statistics for association with the trait, which assigns a *p*-value for the SNP-trait association. Individual marker-trait

associations that meet the threshold set for the false discovery rate (FDR, the proportion of false positives among all significant results for some level α) are then studied in more detail to uncover hints as to the genetic architecture of the trait, and how best to improve it in the future. Many true associations may be missed in GWAS, however, because the threshold for FDR could be as low as α divided by the total number of SNPs being tested. In complex, polygenic traits, the effects of genes that exert only small effects on a trait may not meet the FDR threshold, especially if the effect value of the association is influenced by the environment. Additionally, alleles of many genes may be expressed only in specific genetic backgrounds and will only be useful when found in combination with the positive alleles of other genes in the same pathway [3]. These allelic combinations may not exist in the limited number of individuals in the GWAS panel. Thus, the statistical power of GWAS for detecting genes of small effect is limited by the strict levels set for FDR and by insufficient numbers of high-frequency polymorphisms found in most panels.

Metabolic pathway analysis focuses on the combined effects of many genes that are grouped according to their shared biological function [4, 5, 6]. This is a promising approach that can complement GWAS to give clues to the genetic basis of a trait. Originally developed to study differences in gene expression data in human disease studies [7], pathway analysis and association mapping have been used in medical research to find biological insights missed when focusing on only one or a few genes that have highly significant associations with a trait of interest [8, 9, 5, 10]. Pathway analysis has only just begun to be used as well in animal studies [11, 12]. In addition, biologically relevant pathways can be used to guide interpretation of large data sets produced by other high-throughput approaches like RNA sequencing, proteomics, and metabolomics.

More recently, GWAS-based metabolic pathway analysis has been used as a discovery tool to investigate the genetic basis of complex traits in plants. A pathway-based approach was used to study aflatoxin accumulation [13], corn ear worm resistance [14] and oil biosynthesis [15] in maize. Combining GWAS analysis with metabolic pathway analysis considers all genetic sequences positively associated with the trait of interest, regardless of magnitude, and jointly may highlight which sequences lead to mechanisms for crop improvement and which warrant further study and manipulation, for example, by gene editing. While combined GWAS and pathway analyses were highly successful in uncovering associated pathways, the analyses were slow and cumbersome, as the analysis tools were written in a combination of R, Perl, and Bash, and the output of each analysis was manually input into the next analysis. A single, unified and user friendly tool to accomplish this pathway analysis was lacking.

The Pathway Association Study Tool (PAST) was developed to facilitate easier and more efficient GWAS-based metabolic pathway analysis. PAST was designed for use with maize but is usable for other species as well. It tracks all SNP marker - trait associations, regardless of significance or magnitude. PAST groups SNPs into linkage blocks based on linkage disequilibrium (LD) data and identifies a tagSNP from each block. PAST then identifies genes within a user-defined distance of the tagSNPs, and transfers the attributes of the tagSNP to the gene(s), including the allele effect, R^2 and p -value of the original SNP-trait association found from the GWAS analysis. Finally, PAST uses the gene effect values to calculate an enrichment score (ES) and p -value for each pathway. PAST is easy to use as an online tool, standalone R script, or as a downloadable R Shiny application. It uses as input TASSEL [16] files that are generated as output from the General Linear or Mixed Linear Models (GLM and MLM), or files from any association analysis that has been similarly formatted, as well as genome annotations in GFF format, and a metabolic pathways file.

Implementation

PAST is implemented as an R package and is available through Bioconductor 3.9. PAST is based on a

method developed by our research group [6]. The original method was subsequently used in two other maize studies [14, 15], but required users to customize Perl and R scripts and run BASH scripts. PAST's implementation is completely in R and requires a user to install the package without needing to edit the source code. Two graphical user interfaces are available in the form of R Shiny applications. A generic version is available on CyVerse and Github, while a maize-specific version can be run on MaizeGDB [17] (explained below).

PAST processes data through four main steps. First, GWAS output data is loaded into PAST. This data comes in the form of statistics that reflect the effects of specific loci (e.g., SNPs) with a trait(s) of interest and LD data between loci. The SNPs are then associated with genes based on the allelic effects, p -values, and genomic distance between SNPs and genes. The genes and their effects are used to find significant pathways and calculate a running ES. Finally, the genes in these pathways are plotted to show a running ES for each pathway. A flowchart in Figure 1 shows the process.

Loading Data

During the process of loading data, the GWAS data is filtered to account for any non-biallelic data. Any data with more or fewer than two alleles associated with that marker is discarded. Data without an R^2 value (coefficient of determination the SNP/trait association) is removed as well, since later calculations rely on the R^2 value. The effects data is associated with the statistics data in order to collect all data about a marker into a single dataframe.

The LD data is filtered to drop rows where the loci are not the same, and then extra columns are dropped. Only data about the locus, the positions, the sites, the distance between the sites, and the r^2 value (coefficient of determination for LD) is retained. The remaining data is split into groups based on the locus.

Assigning Genes

Genes are assigned the attributes of linked SNPs according to the method described in Tang *et al* [6]. SNPs are parsed into linked groups by identifying all pairs of SNPs with LD data that exceed a set cutoff r^2 for linkage. SNP blocks occur when multiple SNPs are linked to one SNP in common. SNPs that are only linked to one other SNP are considered singly-linked SNPs, and SNPs not linked to any other SNPs are unlinked. In all cases, PAST follows an algorithm to identify one tagSNP to represent all linked SNPs in order to reduce the dimensionality of the dataset and identify which allele effect, p and R^2 to transfer to the physically linked gene(s). Unlinked SNPs are by default identified as the tagSNP. For SNPs that are linked to a single other SNP, if both have the same effect sign (positive or negative), PAST identifies the one associated with the largest effect (absolute value) as the tagSNP. If the effects are equal, the second (more downstream) SNP is used. If, however, the effect signs are different, the SNP with the lowest p -value is used. If the p -values are the same and the signs are different, the SNP is labeled as problematic, since no assignment can be made, and no tagSNP is identified; these are dropped from the analysis.

The tagSNP within blocks of SNPs is identified by first counting the number of positive and negative effects in each linkage block. If the number of positive effects is greater, then the SNP with largest positive effect is chosen. If the number of negative effects is greater, then the SNP with the largest negative effect is noted. Ties between the number of negative and positive effects are broken by checking the sign of the SNP in common defining the block. The tagSNP is then the one with the largest effect and the same sign, and it is marked to indicate the number of SNPs in the block. Once all blocks have been reduced to a single tagSNP, the tagSNP is used to locate the nearby gene(s).

Once tagSNPs have been identified, the annotation files are checked to look for genes within a physical

distance window provided by the user. The effect and the p -value of the tagSNP is transferred to the gene. The SNP-gene assignments are grouped by gene name, and if more than one SNP block or unlinked SNP was found to be linked to the same gene, each gene is tagged by counting the number of negative effect and the number of positive effect associations in the blocks linked to the same gene. If there are more negative effects, the most negative effect and p -value is assigned to the gene. If there are more positive effects, the positive effect and p -value is assigned to the gene. If there are more than one equally positive or equally negative effects, the effect with the lowest p -value is chosen and assigned to the gene. If there are an equal number of negative and positive effects, the effect with the greatest absolute value is selected. The number of linked SNPs is set to the total number of SNPs (SNPs within blocks plus blocks within genes) linked to that gene. Once all of the blocks of genes have been processed, the effects of each gene are used to find significant pathways.

Finding Significant Metabolic Pathways

Significant pathways are found by using a previously described method [4, 6, 7]. User-input determines the minimum number of genes that a pathway must contain to be retained for processing (to avoid small sample size bias), the number of times the effects data are randomly sampled with replacement to generate a null distribution of ES, and the pathways database that is being used.

For each gene effects column (observed and randomly sampled), the effects are sorted and ranked from best to worst (and whether this is in increasing or decreasing order depends on the trait under study). The ES running sum statistic increases for genes in the pathway and decreases for genes not in the pathway. The amount of increase for genes in the pathway is weighted by the absolute value of the effect. The pathway ES is the largest positive value calculated for the running sum statistic.

Pathway significance is determined by comparing the observed ES with the ES for the null distribution. The mean and standard deviation for the null distribution are used to normalize the observed ES so that z scores can be obtained. P -values are computed from the z scores using the (1-pnorm) function. Since multiple hypothesis testing is still a concern, an FDR-adjusted p -value (known as q -value) is calculated using the `qvalue` package in R [18].

Plotting

Based on user input, the pathways can be filtered for significance (either p -value or q -value), or the top n pathways can be selected. Rugplots for each pathway in the set of significant pathways are plotted as the last step. The x-axis shows the rank of each gene effect value; the y-axis shows the value of the ES running sum statistic as each consecutive gene effect value is processed. An x-intercept line indicates the highest point of the ES. Small hatch marks at the top of the image indicate the position of the effect of all genes in the pathway. An example rugplot is provided in Figure 2.

R Shiny Applications

Two versions of an R Shiny application that use PAST have been developed. These R Shiny applications provides a guided user interface that sets analysis parameters in PAST; they can also upload a saved set of results to explore again. The version available on Github and CyVerse allows a user to run a new analysis by selecting their data, annotations, and pathways depending on the species being studied. The version available on MaizeGDB [17] allows a user to upload their data and select specific versions of the maize annotation and pathways databases available on MaizeGDB. A screenshot of the generic R Shiny application is provided in Figure 3.

Results and Discussion

PAST is run by calling its functions with GWAS data from within an R script or by using an included R Shiny interface. PAST will allow a new interpretation of GWAS results, which should identify associated pathways either when one or a few genes are highly associated with the trait (these would have been identified by the GWAS analysis directly); or when many genes in the pathway are moderately associated with the trait (these would not necessarily have been identified by the GWAS analysis). Such an interpretation will add both additional results, and biological meaning to the association data, as was seen with oil biosynthesis in studies by Li et al. [19, 15]. While PAST may be useful in bringing biologically useful insights to a GWAS analysis, it will not be able to find order from a chaotic dataset if environmental variation, experimental error, or improper analysis models were used in the association analysis. For strong data sets, however, it may find pathways where GWAS found few or no significant associations which, taken in isolation, shed no real light on the genetic mechanisms underlying the traits under study. PAST may be able to overcome this limitation, and may in addition be able to identify epistatic interactions between genes in the same pathway [20].

PAST was tested using data from three previous corn GWAS on kernel color (261147 SNPs), aflatoxin resistance (261184 SNPs), and total oil production (525107 SNPs). Kernel color and aflatoxin analyses were run on a desktop computer with 32GB of memory, a 4GHz Intel Core i7 with four processors, and solid-state storage. All four processors were used when testing PAST. The kernel color test completed in 63 minutes; the aflatoxin test completed in 62 minutes. The total oil test was run on a system with the following specifications: Intel(R) Xeon(R) CPU E5-2680 v2 at 2.80 GHz with 64 GB of RAM connected to a lustre v.2.5.3 file system via an FDR InfiniBand interconnect. The analysis of the total oil trait completed in 71 minutes. Using the previous method, these analyses took 24 hours or more, depending on how attentive the user was to starting the next step in the process. The results of the analyses of all three traits were comparable when generated with PAST or with the previous method.

The use of metabolic pathway analysis to derive functional meaning from GWAS results has been used extensively in human disease studies, and methodologies and tools similar to PAST have been published for use with annotated human pathways. Some methodologies reviewed by Kwak and Pan [21] include GATES-Simes, HYST, and MAGMA. Two tools for human GWAS pathway studies have been published, including GSA-SNP2 [10] and *Pascal* (a Pathway scoring algorithm) [22]. However, these tools would need to be extensively modified to work with any set of user-supplied pathways. Some of the analysis procedures, however, should be considered in future versions of PAST. An analysis with PAST should be illuminating for any outcrossing plant species and any animal and human datasets for which annotated pathway/genome databases (or related model organism databases) and GFF annotations are available. However, for inbreeding and polyploid species, the assignment of SNPs to genes may be complicated by very long LD blocks which may contain multiple, equally linked genes, or homology to more than one genome. Tests will be run to see if these problems can be overcome.

Conclusions

In conclusion, we present PAST, a tool designed to use GWAS data to perform pathway analysis. PAST is faster and more user-friendly than previous methods, requires minimal knowledge of programming languages, and is publicly available at Github, Bioconductor, CyVerse and MaizeGDB.

Availability and requirements

Project name: PAST

Project home page: <https://github.com/IGBB/PAST/>

Shiny apps: <https://apps.maizegdb.org/PAST/>, https://apps.maizegdb.org/maizeGDB_PAST/,
<CyVerse link once we have it>

Bioconductor: <https://doi.org/doi:10.18129/B9.bioc.PAST>

Operating system(s): Platform independent

Programming language: R 3.6

Other requirements: R packages dplyr, rlang, iterators, parallel, foreach, doParallel, qvalue, rtracklayer, ggplot2, shiny, shinydashboard, gridExtra

License: GNU GPLv3

Any restrictions to use by non-academics: None

List of abbreviations

Pathway analysis study tool (PAST)

genome-wide association study (GWAS)

linkage disequilibrium (LD) data

single nucleotide polymorphism (SNP)

enrichment score (ES)

Declarations

All manuscripts must contain the following sections under the heading 'Declarations':

Ethics approval and consent to participate

NA

Consent for publication

NA

Availability of data and materials

The example data used is available on the project's GitHub, in the example folder.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded in part through USDA Agricultural Research Service Agreements 58-6066-6-046 and 58-6066-6-059, USDA National Institute of Food and Agriculture Multi-State Hatch project 17810, and National Science Foundation grant DBI-1659630.

Authors' contributions

AT contributed to the development of the project, directed development of the software, and contributed to the manuscript. JT developed the method in its original implementation, advised development of PAST, and contributed to the manuscript. MD contributed to the development of PAST, and contributed to the manuscript. DGP advised development of the project. MW provided example data, advised development, contributed to the manuscript, and directed the entire project. All authors read and edited the manuscript.

Acknowledgements

The authors gratefully acknowledge helpful suggestions from Dr. Andy Perkins and Mr. Mark Arick II; and Drs. Peter Bradbury, Brandon Monier, and Mr. Terry Casstevens for their thoughtful manuscript review and suggestions.

References

1. Rafalski, J.A.: Association genetics in crop improvement. *Current Opinion in Plant Biology* **13**(2), 174–180 (2010). doi:10.1016/j.pbi.2009.12.004. Accessed 2019-03-26TZ
2. Yan, J., Warburton, M., Crouch, J.: Association Mapping for Enhancing Maize (L.) Genetic Improvement. *Crop Science* **51**(2), 433 (2011). doi:10.2135/cropsci2010.04.0233. Accessed 2019-03-26TZ
3. Xiao, Y., Liu, H., Wu, L., Warburton, M., Yan, J.: Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant* **10**(3), 359–374 (2017). doi:10.1016/j.molp.2016.12.008. Accessed 2019-03-26TZ
4. Wang, K., Li, M., Bucan, M.: Pathway-Based Approaches for Analysis of Genomewide Association Studies. *The American Journal of Human Genetics* **81**(6), 1278–1283 (2007). doi:10.1086/522374. Accessed 2019-03-26TZ
5. Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S.G., Yu, Z., Xie, X.: SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics* **12**(1), 99 (2011). doi:10.1186/1471-2105-12-99. Accessed 2019-03-26TZ
6. Tang, J.D., Perkins, A., Williams, W.P., Warburton, M.L.: Using genome-wide associations to identify metabolic pathways involved in maize aflatoxin accumulation resistance. *BMC Genomics* **16**(1) (2015). doi:10.1186/s12864-015-1874-9. Accessed 2019-03-26TZ
7. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43), 15545–15550 (2005). doi:10.1073/pnas.0506580102. Accessed 2019-03-26TZ
8. Carlson, C.S., Eberle, M.A., Kruglyak, L., Nickerson, D.A.: Mapping complex disease loci in whole-genome association studies. *Nature* **429**(6990), 446–452 (2004). doi:10.1038/nature02623. Accessed 2019-03-26TZ
9. Torkamani, A., Topol, E.J., Schork, N.J.: Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92**(5), 265–272 (2008). doi:10.1016/j.ygeno.2008.07.011. Accessed 2019-03-26TZ

10. Yoon, S., Nguyen, H., Yoo, Y.J., Kim, J., Baik, B., Kim, S., Kim, J., Kim, S., Nam, D.: Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Research* **46**(10), 60–60 (2018). doi:10.1093/nar/gky175. Accessed 2019-03-26TZ
11. Pegolo, S., Mach, N., Ramayo-Caldas, Y., Schiavon, S., Bittante, G., Cecchinato, A.: Integration of GWAS, pathway and network analyses reveals novel mechanistic insights into the synthesis of milk proteins in dairy cows. *Scientific Reports* **8**(1), 566 (2018). doi:10.1038/s41598-017-18916-4. Accessed 2019-03-26TZ
12. Duarte, D.A.S., Newbold, C.J., Detmann, E., Silva, F.F., Freitas, P.H.F., Veroneze, R., Duarte, M.S.: Genome-wide association studies pathway-based meta-analysis for residual feed intake in beef cattle. *Animal Genetics* **50**(2), 150–153 (2019). doi:10.1111/age.12761. Accessed 2019-03-26TZ
13. Warburton, M.L., Tang, J.D., Windham, G.L., Hawkins, L.K., Murray, S.C., Xu, W., Boykin, D., Perkins, A., Williams, W.P.: Genome-Wide Association Mapping of and Aflatoxin Accumulation Resistance in Maize. *Crop Science* **55**(5), 1857 (2015). doi:10.2135/cropsci2014.06.0424. Accessed 2019-04-05TZ
14. Warburton, M.L., Womack, E.D., Tang, J.D., Thrash, A., Smith, J.S., Xu, W., Murray, S.C., Williams, W.P.: Genome-Wide Association and Metabolic Pathway Analysis of Corn Earworm Resistance in Maize. *The Plant Genome* **11**(1), 0 (2018). doi:10.3835/plantgenome2017.08.0069. Accessed 2019-03-26TZ
15. Li, H., Thrash, A., Tang, J.D., He, L., Yan, J., Warburton, M.L.: Leveraging GWAS data to identify metabolic pathways and networks involved in maize lipid biosynthesis. *The Plant Journal* (2019). doi:10.1111/tbj.14282. Accessed 2019-03-26TZ
16. Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S.: TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**(19), 2633–2635 (2007). doi:10.1093/bioinformatics/btm308. Accessed 2019-04-11TZ
17. Portwood JL, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, Schaeffer ML, et al. MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Research*. 2019;47:D1146–54.
18. Storey, J.D., Bass, A.J., Dabney, A., Robinson, D.: qvalue. publisher: Bioconductor. <http://bioconductor.org/packages/qvalue/> Accessed 2019-03-26TZ
19. Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M.L., Cheng, Y., Hao, X., Zhang, P., Zhao, J., Liu, Y., Wang, G., Li, J., Yan, J.: Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature Genetics* **45**(1), 43–50 (2013). doi:10.1038/ng.2484. Accessed 2019-03-26TZ
20. Liu, Y., Maxwell, S., Feng, T., Zhu, X., Elston, R.C., Koyutürk, M., Chance, M.R.: Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data. *BMC Systems Biology* **6**(3), 15 (2012). doi:10.1186/1752-0509-6-S3-S15. Accessed 2019-03-26TZ
21. Kwak, I.-Y., Pan, W.: Adaptive gene- and pathway-trait association testing with GWAS summary statistics. *Bioinformatics* **32**(8), 1178–1184 (2016). doi:10.1093/bioinformatics/btv719. Accessed 2019-04-05TZ
22. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., Bergmann, S.: Fast and Rigorous

Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLOS Computational Biology **12**(1), 1004714 (2016). doi:10.1371/journal.pcbi.1004714. Accessed 2019-04-05TZ

Figures, tables and additional files

Figure 1: The process used in PAST

Figure 2: Example PAST Output

Figure 3: A screenshot of the R Shiny application