Journal of Visualized Experiments

A pathway association study tool for GWAS analyses of metabolic pathway information

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE61268R2
Full Title:	A pathway association study tool for GWAS analyses of metabolic pathway information
Section/Category:	JoVE Genetics
Keywords:	genome-wide association analysis; GWAS; metabolic pathway analysis; data interpretation; software; R; Bioconductor
Corresponding Author:	Charles Adam Thrash, Ph.D. Institute for Genomics, Biocomputing & Biotechnology Mississippi State, MS UNITED STATES
Corresponding Author's Institution:	Institute for Genomics, Biocomputing & Biotechnology
Corresponding Author E-Mail:	thrash@igbb.msstate.edu
Order of Authors:	Charles Adam Thrash, Ph.D.
	Marilyn L Warburton
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Open Access (US$4,200)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Mississippi State, Mississippi, United States of America

Cover Letter

To the Editors of JoVE;

Attached, please find our manuscript entitled “How to run the Pathway Association Study Tool and interpret the results of GWAS analyses in light of metabolic pathway information", which we submit for publication as peer reviewed Scientific Video. We have been invited to submit this paper by Lyndsay Troyer, with whom we have been working on the submission.

Genome wide association studies (GWAS) have become very common in maize, and in some cases can uncover useful variants for the study of the genetic architecture or practical improvement of a trait. In many cases, however, particularly in the case of quantitative traits, we believe that GWAS is not finding all useful information contained in the analyses. In particular, the need to set such a stringent significance threshold on results ensures that many of the results are discarded. Therefore, we created a new tool called PAST (the Pathway Association Study Tool) that assigns SNPs from a GWAS study to genes, and genes to metabolic pathways. The tool then recalculates the probability that the entire pathway is associated with the trait of interest. We have found that since we began working on this tool, a few other groups have released tools, especially for human genetics studies, but we believe no other tool is as versatile, robust, and user friendly than PAST. Lynsday suggested we could present how to use PAST in a JoVE video, and we present here the manuscript that will accompany the video, when it is finished.

We hope you agree that this manuscript is descriptive and useful. My coauthor Adam will present PAST in the video, and we hope this will allow us to reach a much wider audience with our new and potentially highly impactful tool. Thank you very much in advance for your consideration.

Sincerely,

Marilyn Warburton

TITLE:

A Pathway Association Study Tool for GWAS Analyses of Metabolic Pathway Information

AUTHORS AND AFFILIATIONS:

Adam Thrash[1], Marilyn L. Warburton[2]

[1]Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS, USA
[2]USDA-ARS Corn Host Plant Resistance Research Unit, Mississippi State, MS, USA

Corresponding Author:
Adam Thrash (thrash@igbb.msstate.edu)

Email Addresses of Co-Authors:
Marilyn L. Warburton (marilyn.warburton@usda.gov)

KEYWORDS:

genome-wide association analysis, GWAS, metabolic pathway analysis, data interpretation, software, R, Bioconductor

SUMMARY:

By running the Pathway Association Study Tool (PAST), either through the Shiny application or through the R console, researchers can gain a deeper understanding of the biological meaning of their genome-wide association study (GWAS) results by investigating the metabolic pathways involved.

ABSTRACT:

Recently, a new implementation of a previously described method for interpreting genome-wide association study (GWAS) data using metabolic pathway analysis has been developed and released. The Pathway Association Study Tool (PAST) was developed to address concerns with user-friendliness and slow-running analyses. This new user-friendly tool has been released on Bioconductor and Github. In testing, PAST ran analyses in less than one hour that previously required twenty-four or more hours. In this article, we present the protocol for using either the Shiny application or the R console to run PAST.

INTRODUCTION:

Genome-wide association studies (GWAS) are a popular method of studying complex traits and the genomic regions associated with them[1-3]. In this type of study, hundreds of thousands of single nucleotide polymorphism (SNP) markers are tested for their association with the trait, and the significance of the associations is assessed. Marker-trait associations that meet the false discovery rate (FDR) threshold (or some other type of significance threshold) are retained for the study, but true associations may be filtered out. For complex, polygenic traits, the effect of each gene might be small (and thus filtered out), and some alleles are only expressed in specific conditions that might not be present in the study[3]. Thus, while many SNPs may be retained as associated with the trait, each may have a very small effect. Too many SNP calls will be missing, and an interpretation of the biological meaning and genetic architecture of the trait may be incomplete and confusing. Metabolic pathway analysis can help to address some of these issues by focusing on the combined effects of genes grouped according to their biological function[4-6].

Several studies were completed using a previous implementation of the method described in this article. Aflatoxin accumulation[7], corn earworm resistance[8], and oil biosynthesis[9] were all studied with the previous implementation. While these analyses were successful, the process of analysis was complicated, time-consuming, and cumbersome, because the analysis tools were written in a combination of R, Perl, and Bash, and the pipeline was not automated. Because of the specialized knowledge required to modify this method for each analysis, a new method has now been developed that can be shared with other researchers.

The Pathway Association Study Tool (PAST)[10] was designed to address the shortcomings of the previous method by requiring less knowledge of programming languages and by running analyses in a shorter period. While the method was tested with maize, PAST makes no species-specific assumptions. PAST can be run through the R console, as a Shiny app, and an online version is expected to soon be available on MaizeGDB.

PROTOCOL:

1. Setup

1.1. Install R, if it is not already installed.

NOTE: PAST is written in R and, therefore, requires that its users have R installed. At the time of this writing, installing PAST directly from Bioconductor requires R3.6, but PAST can be installed from Github for users with R3.5. R installation instructions can be downloaded from the following link: https://www.r-project.org/.

1.2. Install the latest version of RStudio Desktop or update RStudio (optional).

NOTE: RStudio is a helpful environment for working with the R language. Its installation is recommended, especially for those who choose to run PAST in the command line rather than through the Shiny GUI application. RStudio and its installation instructions can be found at the following link: https://rstudio.com/products/rstudio/.

1.3. Install PAST from Bioconductor[11] by following the instructions on Bioconductor.

NOTE: Installation through Bioconductor should handle the installation of PAST’s dependencies. Additionally, PAST can be installed from Github[12], but installing from Github will not install dependencies automatically.

1.4. Install PAST Shiny (optional). Download the file “app.R” from the Releases page of the Github repository: https://github.com/IGBB/PAST/releases/, and remember where the downloaded file is located.

NOTE: PAST can be used by calling its methods directly with R, but users who are less familiar with R can run the PAST Shiny application, which provides a guided user interface. PAST Shiny is an R script available in the shiny_app branch of the PAST Github repository. PAST Shiny will attempt to install its dependencies during the first run.

1.5. Begin analysis by starting the application in one of the three ways described below.

1.5.1. PAST Shiny with RStudio

1.5.1.1. Using RStudio, create a new project in the folder where app.R is located. Click **File | New Project** and select that folder.

1.5.1.2. Once a new project has been created, open the app.R file downloaded earlier. RStudio recognizes that app.R is a Shiny app and creates a **Run App** button on the bar above the displayed source code. Click **Run App**. RStudio will then launch a window that displays the PAST Shiny application.

1.5.2. PAST Shiny with R Console

1.5.2.1. Launch R and run the following code to start the PAST Shiny application: *shiny::runApp('path/to/folder/with/shiny/app.R*'. Replace the text in quotes with the folder to which app.R was downloaded, and keep the quotes.

1.5.3. PAST without R Shiny

1.5.3.1. Run library(PAST) in an R Console to load PAST.

2. Customize Shiny analysis (optional)

2.1. Change the analysis title from “New Analysis” to something that better reflects the type of analysis being run which helps to keep track of multiple analyses (see **Figure 1**).

[Place **Figure 1** here]

2.2. Modify the number of cores and the mode. Set the number of cores to any number between 1 and the total number on the machine but be aware that devoting more resources to PAST may slow down other operations on the machine. Set the mode based on the description in section 6.

3. Load GWAS data

NOTE: Verify that the GWAS data is tab delimited. Ensure that the association file contains the following columns: trait, marker name, locus or chromosome, position on the chromosome, p-value, and R^2 value for the marker. Ensure that the effects file contains the following columns: trait, marker name, locus or chromosome, and position on the chromosome. The order of these columns is not important, as the user can specify the names of the columns when loading the data. Any additional columns are ignored. TASSEL[13] can be used to produce these files.

3.1. Load GWAS data with PAST Shiny.

3.1.1. Select an association file and an effects file by using the **Association File** and **Effects File** selection boxes. Change the column names in the **Association Column Name** and **Effects Columns Name** input boxes below the file selection boxes to reflect the column names in the data.

[Place **Figure 2** here]

3.2. Load GWAS data with PAST in the R Console.

3.2.1. Modify and run the following code:

```
gwas_data = load_GWAS_data(“path/to/association_file.tsv”, “path/to/effects_file.tsv”,
association_columns = c("Trait", "Marker", "Locus", "Site", "p", "marker_R2"), effects_columns =
c("Trait", "Marker", "Locus", "Site", "Effect")
```

NOTE: Change the paths to the actual location of the GWAS files. The values provided for association_columns and effects_columns are the default values. If the names do not match the default values, specify the column names. Otherwise, these can be omitted.

4. Load linkage disequilibrium (LD) data

NOTE: Verify that the linkage disequilibrium (LD) data is tab delimited and contains the following types of data: Locus, Position1, Site1, Position2, Site2, Distance in base pairs between Position1 and Position2, and R^2 value.

4.1. Load LD data with PAST Shiny.

4.1.1. Select the file containing LD data. Change the column names in the **LD Column Names** input boxes below the file selection box to match the column names in the LD data if necessary.

[Place **Figure 3** here]

4.2. Load LD Data with PAST in the R Console.

4.2.1. Modify and run the following code to load LD data:

```
LD = load_LD(“path/to/LD.tsv”, LD_columns = c("Locus1", "Position1", "Site1", "Position2", "Site2", "Dist_bp", "R.2")
```

NOTE: Change the path to the actual location of the LD file. The values provided for LD_columns are the default values. If the names do not match these defaults, specify the correct names of the columns; otherwise, these can be omitted.

5. Assign SNPs to genes

NOTE: Download or otherwise locate annotations in GFF format. These annotations can often be found in online databases for specific organisms. Be cautious about low quality annotations, as the quality of the annotations data will affect the quality of the pathway analysis. Confirm that the first column of these annotations (the chromosome) matches the format of the locus/chromosome in the association, effects, and LD data. For example, the annotations should not call the first chromosome “chr1” if the GWAS and LD data files call the first chromosome “1”.

5.1. Assign SNPs to genes with PAST Shiny.

NOTE: More information about determining an appropriate R^2 cutoff can be found in Tang et al.[6], in the section called “SNP to gene algorithm for the pathway analysis”.

5.1.1. Select the file containing GFF annotations. Consider what window size and R^2 cutoff are most suitable for the species being considered and modify if the defaults do not suit the uploaded data.

NOTE: Default values in PAST primarily reflect values appropriate for maize. The number of cores set at the beginning of the PAST Shiny analysis (Step 2.2) is used in this step.

[Place **Figure 4** here]

5.2. Assign SNPs to genes with PAST in the R Console.

5.2.1. Modify and run the following code to assign SNPs to genes:

```
genes = assign_SNPs_to_genes(gwas_data, LD, “path/to/annotations.gff”, c("gene"), 1000, 0.8, 2)
```

NOTE: In this sample code, several default suggestions are provided: 1000 is the size of the window around the SNP to search for genes; 0.8 is the cutoff value for R^2; 2 is the number of cores used for parallel processing. The path to the annotations should also be changed to the actual location of the annotations file.

6. Discover significant pathways

NOTE: Verify that the pathways file contains the following data in tab delimited format, with one line for every gene in each pathway: pathway ID – an identifier such as “PWY-6475-1”; pathway description – a lengthier description of what the pathways do such as “trans-lycopene biosynthesis”; gene – a gene in the pathway, which should match the names provided in the annotations. Pathway information can likely be found in online databases for specific organisms, such as MaizeGDB. The second user-specified option is the mode. “Increasing” refers to phenotypes that reflect when an increasing value of the measured trait is desirable, such as yield, while “decreasing” refers to a trait where a decrease in the measured values is beneficial, such as insect damage ratings. The significance of pathways is tested using previously described methods[4,6,14].

6.1. Discover significant pathways with PAST Shiny.

6.1.1. Select the file containing pathways data and be sure that the mode is selected in the analysis options. If necessary, change the number of genes that must be in a pathway to retain it for the analysis and the number of permutations used to create the null distribution to test significance of effect.

[Place **Figure 5** here]

NOTE: The number of cores and the mode set at the beginning of the PAST Shiny analysis (Step 2.2) is used in this step. The default number of genes is currently set at 5 genes, so pathways with fewer known genes will be removed. The user can lower this value to 4 or 3, to include shorter pathways, but doing so will risk false positive results. Increasing this value can increase the power of the analysis but will remove more pathways from the analysis. Changing the number of permutations used, increases and decreases the power of the test.

6.2. Discover significant pathways with PAST in the R Console.

6.2.1. Modify and run the following code to discover significant pathways:

```
rugplots_data <- find_pathway_significance(genes, “path/to/pathways.tsv”, 5, “increasing”, 1000, 2)
```

NOTE: In this sample code, several suggested defaults are provided. 5 is the minimum number of genes that must be in a pathway in order to keep the pathway in the analysis, increasing refers to an increasing amount of the measured trait (it is recommended that the user run both increasing and decreasing, regardless of trait; data interpretation will differ for the two, however), 1000 is the number of times to sample the effects to determine the null distribution, and 2 is the number of cores used for parallel processing. Change the path to the actual location of the pathways file.

7. View Rugplots

7.1. View Rugplots with PAST Shiny.

7.1.1. Once all inputs are uploaded and set, click **Begin Analysis**. A progress bar will appear and indicate which step of the analysis was last completed. When the analysis completes, PAST Shiny will switch to the **Results** tab. A table of results will be displayed in the left column (labeled “pathways”) and the Rugplots will be displayed in the right column (labeled “plots”).

[Place **Figure 6** here]

7.1.2. Use the slider to control the filtering parameters. When the filtering level is satisfactory, click the **Download Results** button at the bottom left to download all images and tables individually to a ZIP file that is named with the Analysis title. This ZIP file contains the filtered table, the unfiltered table, and one image per pathway in the filtered table.

[Place **Figure 7** here]

7.2. View Rugplots with PAST in the R Console

7.2.1. Modify and run the following code to save the results:

```
plot_pathways(rugplots_data, "pvalue", 0.02, "increasing", “output_folder”)
```

NOTE: In this sample code, several suggested defaults are provided. pvalue provides the data that can be used for filtering insignificant pathways after a significance threshold is chosen by the user; 0.02 is the default value used in filtering, and increasing refers to an increasing amount of the measured trait (it is recommended that the user run both increasing and decreasing, regardless of trait; data interpretation will differ for the two, however); output_folder is the folder where the images and tables will be written (this folder must exist prior to running the function). A table of filtered results, the unfiltered results, and individual images for every pathway in the filtered results are written to this folder.

REPRESENTATIVE RESULTS:

If results are not produced following a run of the PAST software tool, check to be sure that all input files are correctly formatted. A successful run using the example data in the PAST package, which are based on a maize GWAS of grain color, is shown in **Figure 8**. This table and the resulting image can be downloaded using the Download Results button. An example of the downloaded image is shown in **Figure 2**[10]. Incorrect settings might lead to results that do not make biological sense, but determining incorrectness must be up to the researcher, who should double check the validity of the chosen settings and consider all known evidence regarding the trait of interest.

Figure 9[10] shows the rugplot produced from the pathway analysis of GWAS results created with a maize panel of 288 inbred lines that had been phenotyped for grain color. This simplistic example, where the phenotypes were either “white” or “yellow”, was used because the pathway responsible for creating the bright yellow carotenoid pigments is known and should be responsible for most of the phenotype. Thus, we expected to see the trans-lycopene biosynthesis pathway (which produce carotenoids) to be significantly associated with grain color, which it is. Pathway ID and name are listed at the top of the graph. The horizontal axis of the graph ranks all genes that were included in the analysis, arranged from left to right in order of largest effect on the trait to smallest. However, only the genes in the trans-lycopene biosynthesis pathway are marked (at the top of the graph, as hatch marks, appearing in the gene rank of their effect as compared to all other genes in the analysis). There are 7 genes in this pathway. The running enrichment score (ES) is plotted along the vertical axis. The ES for each gene is added into the running total in order of effect and the total is adjusted to the number of genes analyzed. Thus, the score changes as one moves right along the horizontal axis and tends to increase as the larger effect genes are included, but at some point, the increase in the effect is smaller than the adjustment for having added another gene, and the entire score begins to decrease. The apex of the running ES line is marked with a dotted vertical line; this is the ES for the entire pathway and is used by the program to determine if the pathway is chosen and presented as a rugplot.

FIGURE LEGENDS:

Figure 1: Step 2.1.

Figure 2: Step 3.1.

Figure 3: Step 4.1.

Figure 4: Step 5.1.

Figure 5: Step 6.1.

Figure 6: Step 7.1.1.

Figure 7: Step 7.1.2.

Figure 8: Completed run of PAST Shiny.

Figure 9: Pathway image from completed run of PAST (or downloaded from Shiny). This figure has been cited from Thrash et al.[10].

DISCUSSION:

A primary goal of PAST is to bring metabolic pathway analyses of GWAS data to a wider audience, especially for non-human and non-animal organisms. Alternative methods to PAST are often command-line programs that focus on humans or animals. User-friendliness was a primary goal in the development of PAST, both in choosing to develop a Shiny application and in choosing to use R and Bioconductor to release the application. Users do not need to learn how to compile programs in order to use PAST.

As with most types of analysis software, the results of PAST are only as good as the input data; if the input data has errors or is incorrectly formatted, PAST will fail to run or produce uninformative results. Ensuring that the GWAS data, LD data, annotations, and pathways files are correctly formatted is critical to receiving correct output from PAST. PAST only analyzes bi-allelic markers and can run only one trait for each set of input data. In addition, GWAS data produced by poor genotyping or incorrect or imprecise phenotyping is not likely to produce clear or repeatable results either. PAST can aid in the biological interpretation of GWAS results but is unlikely to clarify chaotic data sets if environmental variation, experimental error, or population structure was not properly accounted for.

Users can choose to change some parameters of the analysis, both in the Shiny application and by passing those parameters to PAST's functions in the R console. These parameters can change the results reported by PAST, and users should take care when modifying these from the defaults. Because LD is measured by the users, typically using the same marker data set that was also used in the GWAS, the LD measurements are specific to the population. For all studies, especially for species other than maize, (particularly self-pollinating, polyploid, or highly heterogenous species), changes in the defaults may be warranted.

ACKNOWLEDGMENTS:
None.

DISCLOSURES:
The authors have nothing to disclose.

REFERENCES:

1. Rafalski, J. Association genetics in crop improvement. *Current Opinion in Plant Biology*. **13** (2), 174–180 (2010).
2. Yan, J., Warburton, M., Crouch, J. Association Mapping for Enhancing Maize (Zea *mays* L.) Genetic Improvement. *Crop Science*. **51** (2), 433–449 (2011).
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).
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).
5. Weng, L. et al. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*. **12** (1), 99 (2011).
6. Tang, J., Perkins, A., Williams, W., Warburton, M. Using genome-wide associations to identify metabolic pathways involved in maize aflatoxin accumulation resistance. *BMC Genomics*. **16** (1), 673 (2015).
7. Warburton, M. et al. Genome-Wide Association Mapping of Aspergillus flavus and Aflatoxin Accumulation Resistance in Maize. *Crop Science*. **55** (5), 1857–1867 (2015).
8. Warburton, M. et al. Genome-Wide Association and Metabolic Pathway Analysis of Corn Earworm Resistance in Maize. *The Plant Genome*. **11** (1), 170069 (2018).
9. Li, H., Thrash, A., Tang, J., He, L., Yan, J., Warburton, M. Leveraging GWAS data to identify metabolic pathways and networks involved in maize lipid biosynthesis. *The Plant Journal*. **98** (5),

853–863 (2019).

10. Thrash, A., Tang, J., DeOrnellis, M., Peterson, D., Warburton, M. PAST: The Pathway Association Studies Tool to Infer Biological Meaning from GWAS Datasets. *Plants*. **9** (1), 58 (2020).

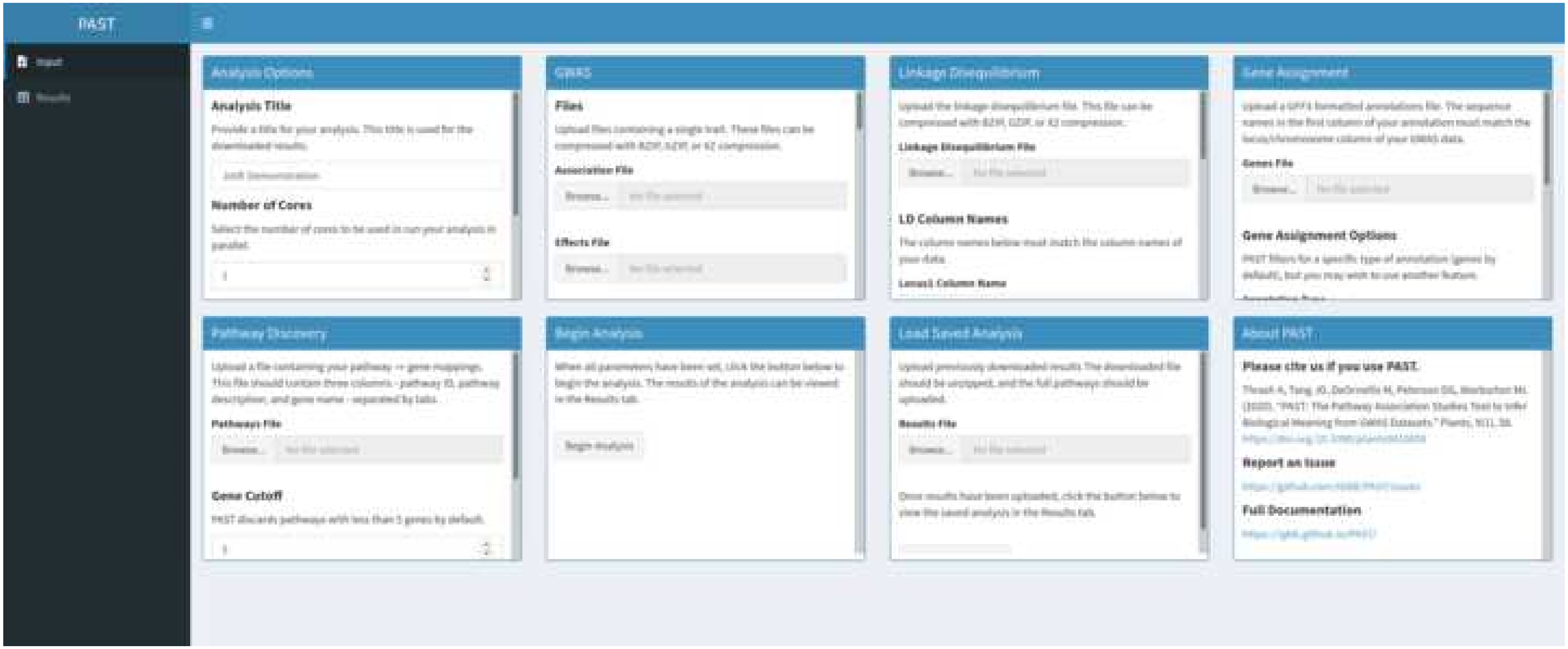
11. Adam, T., Mason, D. PAST: Pathway Association Study Tool (PAST). doi: 10.18129/B9.bioc.PAST. Bioconductor version: Release (3.10). (2020).

12. Thrash, A., DeOrnellis, M. *IGBB/PAST*. at <https://github.com/IGBB/PAST>. IGBB. (2019).

13. Bradbury, P. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*. **23** (19), 2633–2635 (2007).

14. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences U.S.A.* 102. 15545–15550 (**2005)**.

Click here to access/download;Figure;Figure_1.png



Analysis Options

Analysis Title

Provide a title for your analysis. This title is used for the downloaded results.

[illegible] Demonstration

Number of Cores

Select the number of cores to be used to run your analysis in parallel.

1

GWAS

Files

Upload files containing a single trait. These files can be compressed with BZIP, GZIP, or XZ compression.

Association File

Browse... No file selected

Effects File

Browse... No file selected

Linkage Disequilibrium

Upload the linkage disequilibrium file. This file can be compressed with BZIP, GZIP, or XZ compression.

Linkage Disequilibrium File

Browse... No file selected

LD Column Names

The column names below must match the column names of your data.

Locus1 Column Name

Gene Assignment

Upload a GFF3 formatted annotations file. The sequence names in the first column of your annotation must match the locus/chromosome column of your GWAS data.

Genes File

Browse... No file selected

Gene Assignment Options

PAST filters for a specific type of annotation (genes) by default, but you may wish to use another feature.

Pathway Discovery

Upload a file containing your pathway -> gene mappings. This file should contain three columns - pathway ID, pathway description, and gene name - separated by tabs.

Pathways File

Browse... No file selected

Gene Cutoff

PAST discards pathways with less than 5 genes by default.

5

Begin Analysis

When all parameters have been set, click the button below to begin the analysis. The results of the analysis can be viewed in the Results tab.

Begin Analysis

Load Saved Analysis

Upload previously downloaded results. The downloaded file should be unzipped, and the full pathways should be uploaded.

Results File

Browse... No file selected

Once results have been uploaded, click the button below to view the saved analysis in the Results tab.

About PAST

Please cite us if you use PAST.

Thrash A, Tang JD, DeOrnellis M, Peterson DG, Warburton ML (2020). "PAST: The Pathway Association Studies Tool to Infer Biological Meaning from GWAS Datasets." Plants, 9(1), 58. https://doi.org/10.3390/plants9010058

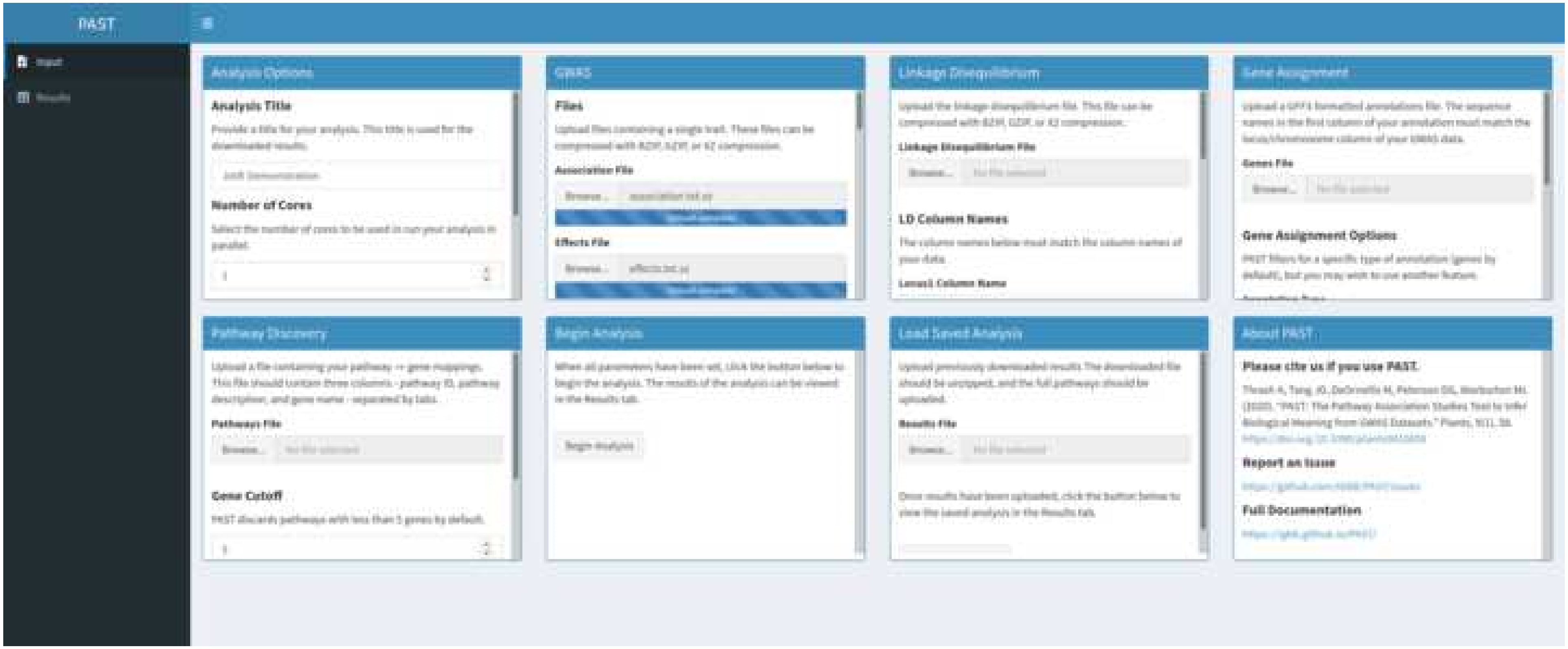
Report an Issue

https://github.com/IGBB/PAST/issues

Full Documentation

https://igbb.github.io/PAST/

Click here to access/download;Figure;Figure_2.png



PAST

Input

Results

Analysis Options

Analysis Title

Provide a title for your analysis. This title is used for the downloaded results.

Number of Cores

Select the number of cores to be used to run your analysis in parallel.

1

GWAS

Files

Upload files containing a single trait. These files can be compressed with BZIP, GZIP, or XZ compression.

Association File

Browse... association.txt.xz

Upload complete

Effects File

Browse... effects.txt.xz

Upload complete

Linkage Disequilibrium

Upload the linkage disequilibrium file. This file can be compressed with BZIP, GZIP, or XZ compression.

Linkage Disequilibrium File

Browse... No file selected

LD Column Names

The column names below must match the column names of your data.

Locus1 Column Name

Gene Assignment

Upload a GFF3 formatted annotations file. The sequence names in the first column of your annotation must match the locus/chromosome column of your GWAS data.

Genes File

Browse... No file selected

Gene Assignment Options

PAST filters for a specific type of annotation (genes by default), but you may wish to use another feature.

Annotation Type

Pathway Discovery

Upload a file containing your pathway -> gene mappings. This file should contain three columns - pathway ID, pathway description, and gene name - separated by tabs.

Pathways File

Browse... No file selected

Gene Cutoff

PAST discards pathways with less than 5 genes by default.

5

Begin Analysis

When all parameters have been set, click the button below to begin the analysis. The results of the analysis can be viewed in the Results tab.

Begin Analysis

Load Saved Analysis

Upload previously downloaded results. The downloaded file should be unzipped, and the full pathways should be uploaded.

Results File

Browse... No file selected

Once results have been uploaded, click the button below to view the saved analysis in the Results tab.

About PAST

Please cite us if you use PAST.

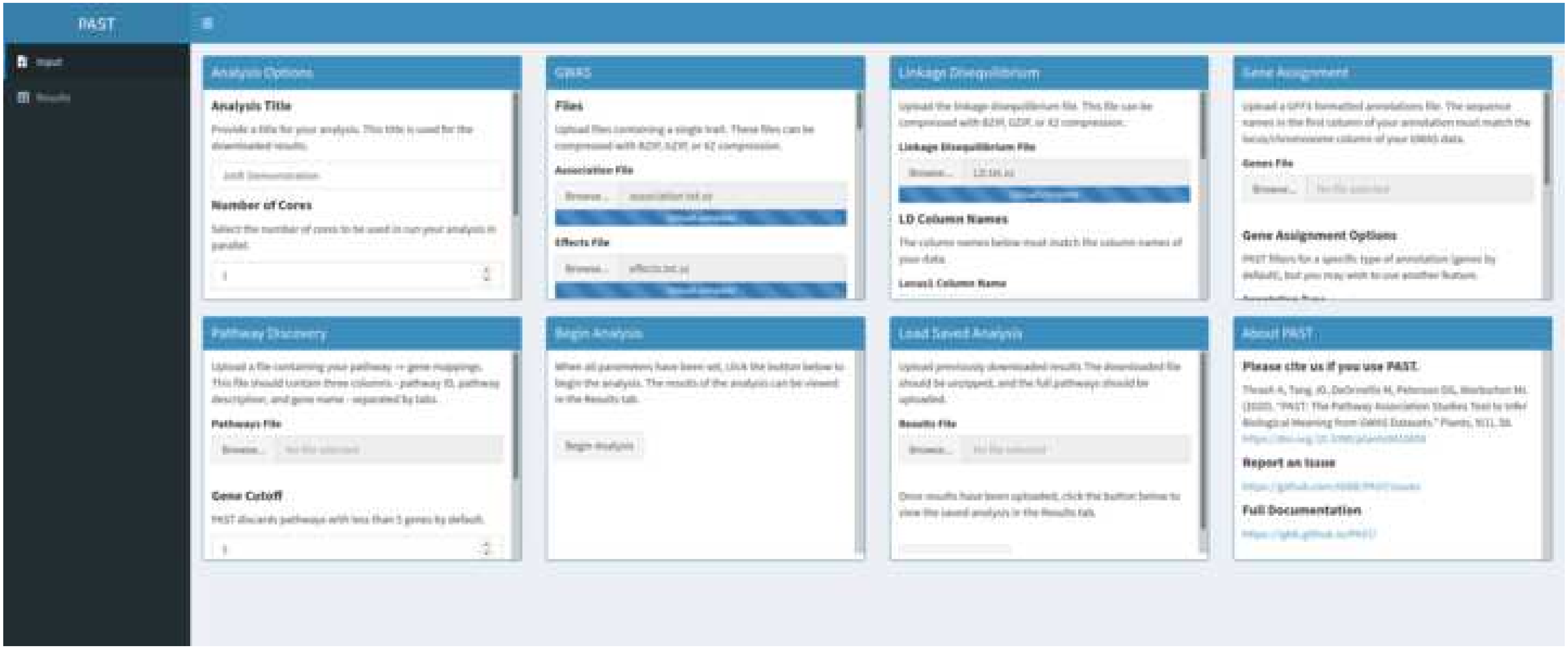
Thrash A, Tang JD, DeOrnellis M, Peterson DG, Warburton ML (2020). "PAST: The Pathway Association Studies Tool to Infer Biological Meaning from GWAS Datasets." Plants, 9(1), 58. https://doi.org/10.3390/plants9010058

Report an Issue

https://github.com/IGBB/PAST/issues

Full Documentation

https://igbb.github.io/PAST/



Analysis Options

Analysis Title

Provide a title for your analysis. This title is used for the downloaded results.

Josh Demonstration

Number of Cores

Select the number of cores to be used to run your analysis in parallel.

1

GWAS

Files

Upload files containing a single trait. These files can be compressed with BZIP, GZIP, or XZ compression.

Association File

Browse... association.txt.xz

Upload complete

Effects File

Browse... effects.txt.xz

Upload complete

Linkage Disequilibrium

Upload the linkage disequilibrium file. This file can be compressed with BZIP, GZIP, or XZ compression.

Linkage Disequilibrium File

Browse... LD.txt.xz

Upload complete

LD Column Names

The column names below must match the column names of your data.

Locus1 Column Name

Gene Assignment

Upload a GFF3 formatted annotations file. The sequence names in the first column of your annotation must match the locus/chromosome column of your GWAS data.

Genes File

Browse... No file selected

Gene Assignment Options

PAST filters for a specific type of annotation (genes by default), but you may wish to use another feature.

Annotation Type

Pathway Discovery

Upload a file containing your pathway -> gene mappings. This file should contain three columns - pathway ID, pathway description, and gene name - separated by tabs.

Pathways File

Browse... No file selected

Gene Cutoff

PAST discards pathways with less than 5 genes by default.

5

Begin Analysis

When all parameters have been set, click the button below to begin the analysis. The results of the analysis can be viewed in the Results tab.

Begin Analysis

Load Saved Analysis

Upload previously downloaded results. The downloaded file should be unzipped, and the full pathways should be uploaded.

Results File

Browse... No file selected

Once results have been uploaded, click the button below to view the saved analysis in the Results tab.

About PAST

Please cite us if you use PAST.

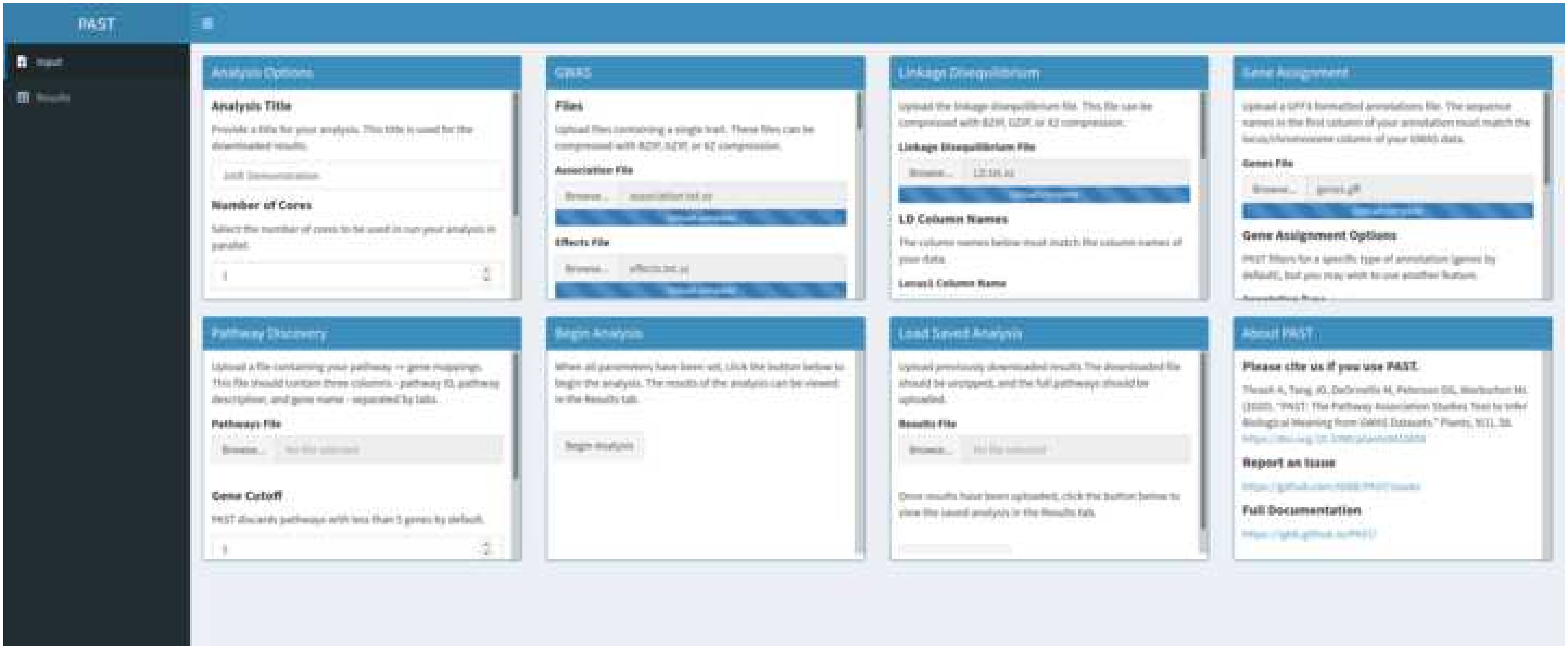
Thrash A, Tang JD, DeOrnellis M, Peterson DG, Warburton ML (2020). "PAST: The Pathway Association Studies Tool to Infer Biological Meaning from GWAS Datasets." Plants, 9(1), 58. https://doi.org/10.3390/plants9010058

Report an Issue

https://github.com/IGBB/PAST/issues

Full Documentation

https://igbb.github.io/PAST/



Analysis Options

Analysis Title

Provide a title for your analysis. This title is used for the downloaded results.

JoVE Demonstration

Number of Cores

Select the number of cores to be used to run your analysis in parallel.

1

GWAS

Files

Upload files containing a single trait. These files can be compressed with BZIP, GZIP, or XZ compression.

Association File

Browse... association.txt.xz

Upload complete

Effects File

Browse... effects.txt.xz

Upload complete

Linkage Disequilibrium

Upload the linkage disequilibrium file. This file can be compressed with BZIP, GZIP, or XZ compression.

Linkage Disequilibrium File

Browse... LD.txt.xz

Upload complete

LD Column Names

The column names below must match the column names of your data.

Locus1 Column Name

Gene Assignment

Upload a GFF3 formatted annotations file. The sequence names in the first column of your annotation must match the locus/chromosome column of your GWAS data.

Genes File

Browse... genes.gff

Upload complete

Gene Assignment Options

PAST filters for a specific type of annotation (genes by default), but you may wish to use another feature.

Annotation Type

Pathway Discovery

Upload a file containing your pathway -> gene mappings. This file should contain three columns - pathway ID, pathway description, and gene name - separated by tabs.

Pathways File

Browse... No file selected

Gene Cutoff

PAST discards pathways with less than 5 genes by default.

5

Begin Analysis

When all parameters have been set, click the button below to begin the analysis. The results of the analysis can be viewed in the Results tab.

Begin Analysis

Load Saved Analysis

Upload previously downloaded results. The downloaded file should be unzipped, and the full pathways should be uploaded.

Results File

Browse... No file selected

Once results have been uploaded, click the button below to view the saved analysis in the Results tab.

About PAST

Please cite us if you use PAST.

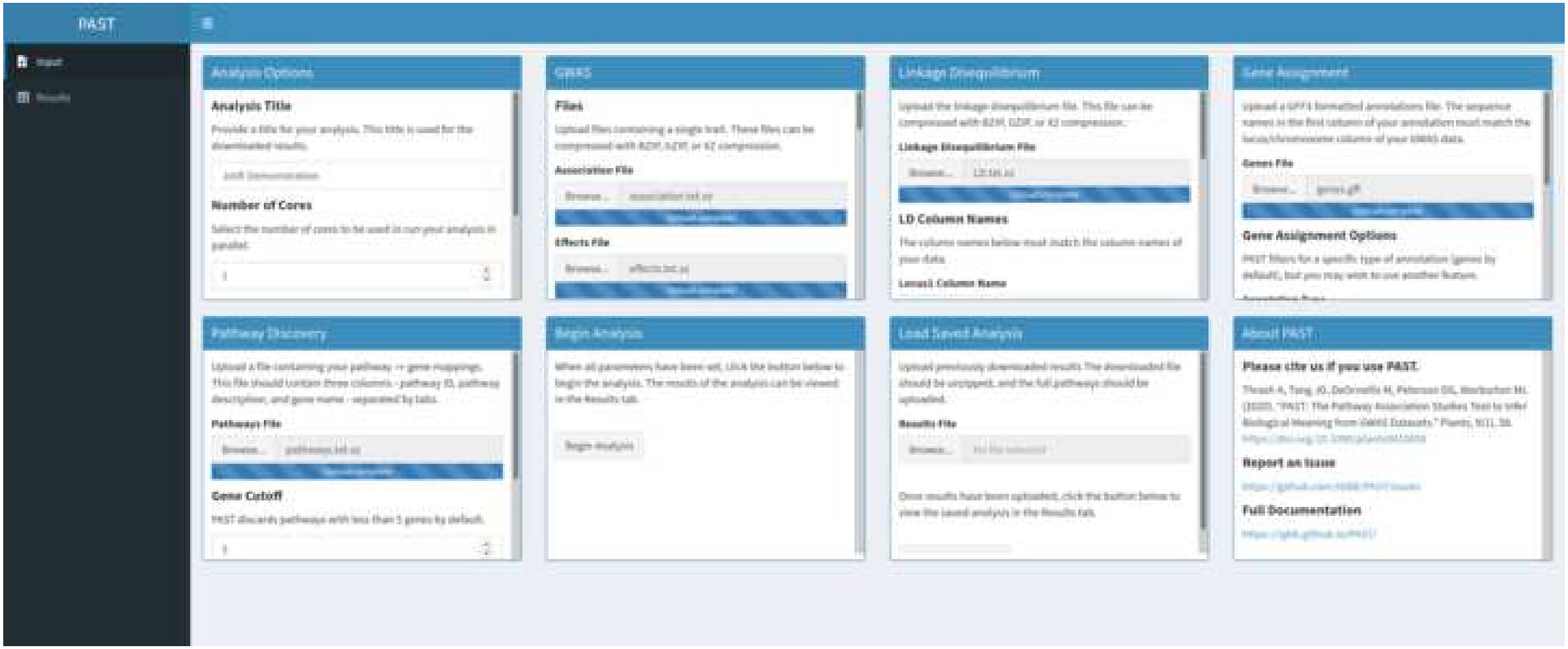
Thrash A, Tang JD, DeOrnellis M, Peterson DG, Warburton ML (2020). "PAST: The Pathway Association Studies Tool to Infer Biological Meaning from GWAS Datasets." Plants, 9(1), 58. https://doi.org/10.3390/plants9010058

Report an Issue

https://github.com/IGBB/PAST/issues

Full Documentation

https://igbb.github.io/PAST/



Analysis Options

Analysis Title

Provide a title for your analysis. This title is used for the downloaded results.

Josh Demonstration

Number of Cores

Select the number of cores to be used to run your analysis in parallel.

1

GWAS

Files

Upload files containing a single trait. These files can be compressed with BZIP, GZIP, or XZ compression.

Association File

Browse... association.txt.xz

Upload complete

Effects File

Browse... effects.txt.xz

Upload complete

Linkage Disequilibrium

Upload the linkage disequilibrium file. This file can be compressed with BZIP, GZIP, or XZ compression.

Linkage Disequilibrium File

Browse... LD.txt.xz

Upload complete

LD Column Names

The column names below must match the column names of your data.

Locus1 Column Name

Gene Assignment

Upload a GFF3 formatted annotations file. The sequence names in the first column of your annotation must match the locus/chromosome column of your GWAS data.

Genes File

Browse... genes.gff

Upload complete

Gene Assignment Options

PAST filters for a specific type of annotation (genes by default), but you may wish to use another feature.

Annotation Type

Pathway Discovery

Upload a file containing your pathway -> gene mappings. This file should contain three columns - pathway ID, pathway description, and gene name - separated by tabs.

Pathways File

Browse... pathways.txt.xz

Upload complete

Gene Cutoff

PAST discards pathways with less than 5 genes by default.

5

Begin Analysis

When all parameters have been set, click the button below to begin the analysis. The results of the analysis can be viewed in the Results tab.

Begin Analysis

Load Saved Analysis

Upload previously downloaded results. The downloaded file should be unzipped, and the full pathways should be uploaded.

Results File

Browse... No file selected

Once results have been uploaded, click the button below to view the saved analysis in the Results tab.

About PAST

Please cite us if you use PAST.

Thrash A, Tang JD, DeOrnellis M, Peterson DG, Warburton ML (2020). "PAST: The Pathway Association Studies Tool to Infer Biological Meaning from GWAS Datasets." Plants, 9(1), 58.
https://doi.org/10.3390/plants9010058

Report an Issue

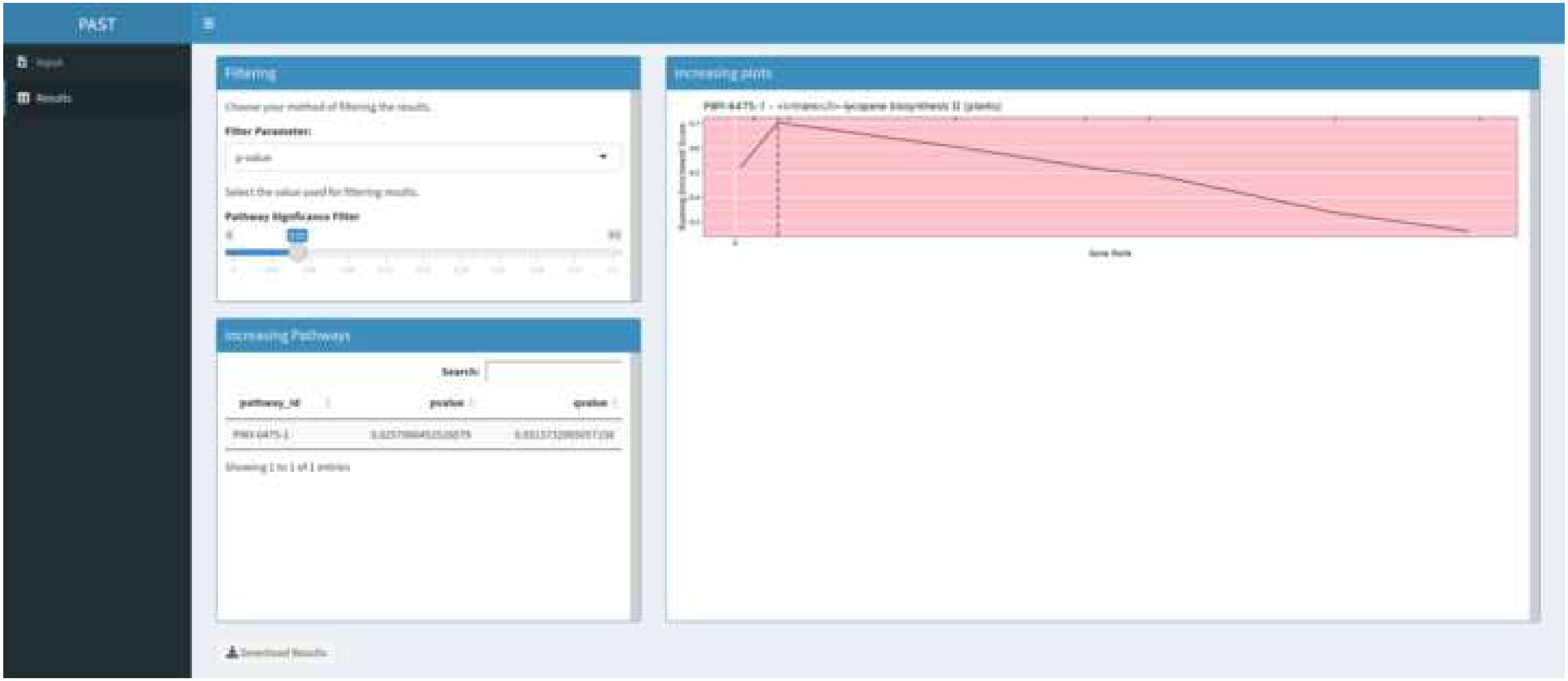
https://github.com/IGBB/PAST/issues

Full Documentation

https://igbb.github.io/PAST/

Assigned SNPs to genes 75%

Click here to access/download;Figure;Figure_7.png



Choose your method of filtering the results.

Filter Parameter:

p-value

Select the value used for filtering results.

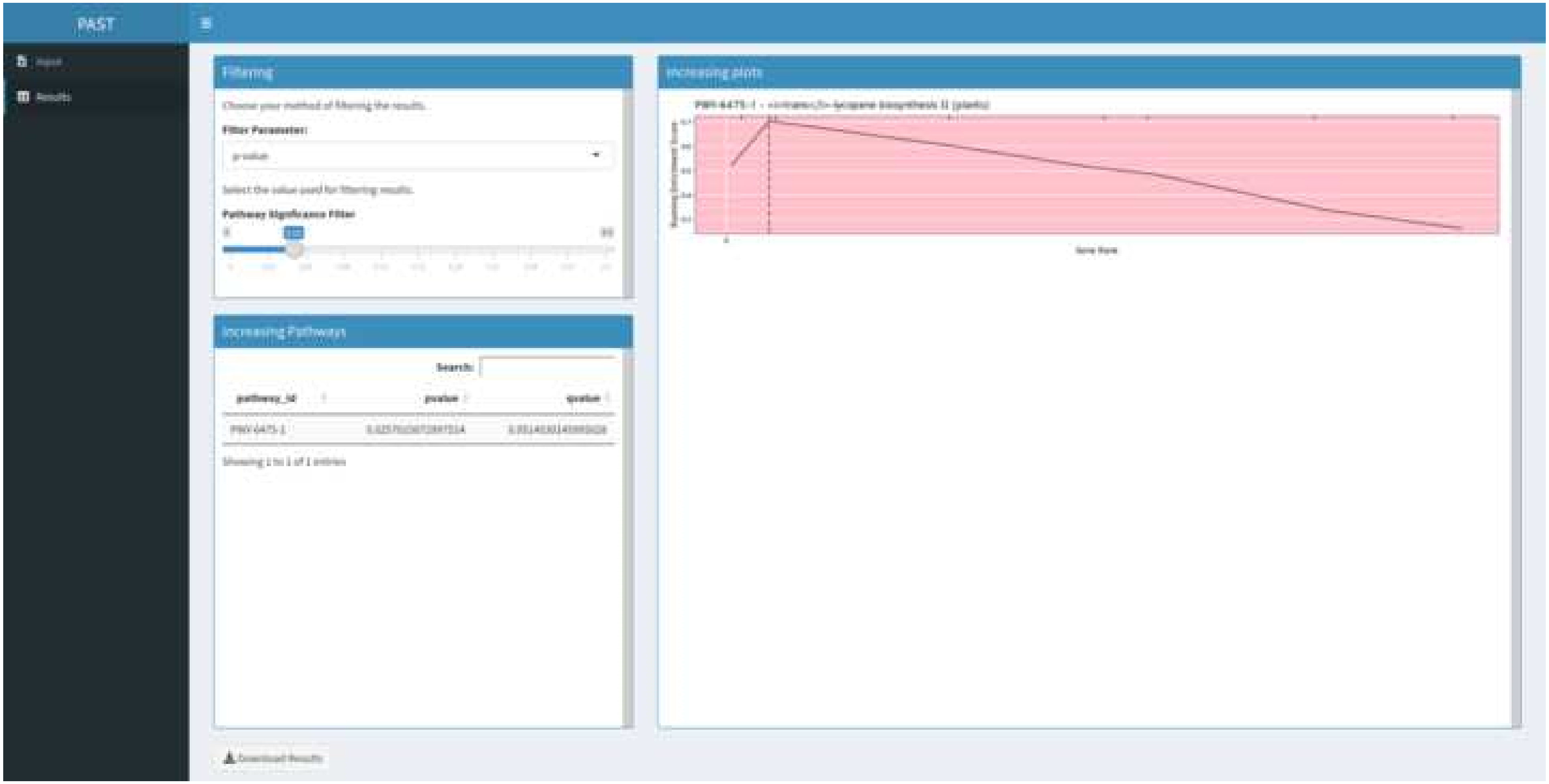
Pathway Significance Filter

PWY-6475-1 - <i>trans</i>-lycopene biosynthesis II (plants)

Search:

pathway_id	pvalue	qvalue
PWY-6475-1	0.0257066452520079	0.0313713280507238

Showing 1 to 1 of 1 entries



Filtering

Choose your method of filtering the results.

Filter Parameter:

p-value

Select the value used for filtering results.

Pathway Significance Filter

Increasing Pathways

Search:

pathway_id	pvalue	qvalue
PWY-6475-1	0.0257[illegible]	0.0514[illegible]

Showing 1 to 1 of 1 entries

Name of Material/ Equipment	Company	Catalog Number	Comments/Description
Computer	NA	NA	Any computer with 8GB RAM should be sufficien
R	R Project	NA	R 3.6 or greater is required to install from Biocon

ɲt
nductor

To Whom It May Concern:

Please find my remarks addressing editorial and reviewer concerns below. Editorial and reviewer remarks are **bold**, while my remarks are not.

Editorial Concerns

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.

Some errors were corrected in this revision.

2. Please submit each figure as a vector image file to ensure high resolution throughout production: (.psd, ai, .eps., .svg).

An SVG for Figure 2 was uploaded. Figure 1 is a screenshot, and vectorizing a raster image wouldn't do much, to my knowledge. Instead, a higher resolution PNG has been attached for Figure 1.

3. Please ensure that the references appear as the following: [Lastname, F.I., LastName, F.I., LastName, F.I. Article Title. Source. Volume (Issue), FirstPage – LastPage (YEAR).] For more than 6 authors, list only the first author then et al.

I edited the author names and order of the references.

4. Please include volume and issue numbers for all references.

More feedback would be useful on this remark. I checked the references, and the only ones without volume and issue numbers are websites.

5. Please define all abbreviations before use.

I have made edits to ensure that all abbreviations are defined before first use.

6. Please revise the table of the essential supplies, reagents, and equipment. The table should include the name, company, and catalog number of all relevant materials in separate columns in an xls/xlsx file. Please sort the Materials Table alphabetically by the name of the material.

This paper describes a software tool. There are no supplies or reagents.

7. Please revise the title for conciseness: A pathway association study tool for GWAS analyses of metabolic pathway information

The title has been changed as suggested.

8. Please ensure that all text in the protocol section is written in the imperative tense as if telling someone how to do the technique (e.g., “Do this,” “Ensure that,” etc.). The actions should be described in the imperative tense in complete sentences wherever possible. Avoid usage of phrases such as “could be,” “should be,” and “would be” throughout the Protocol. Any text that cannot be written in the imperative tense may be added as a “Note.” However, notes should be concise and used sparingly. Please include all safety procedures and use of hoods, etc.

The protocol has been edited to reflect these suggestions. Because PAST is a software tool, there are no safety procedures specific to PAST.

9. The Protocol should contain only action items that direct the reader to do something. Please move the discussion about the protocol to the Discussion.

Remarks on the protocol have either been edited to be imperative or moved.

Reviewer Concerns

--It would be good to mention that there are other tools for visualization of omics data (e.g. Pathway Collage through Pathway Tools, MetaMapp, etc.; can be useful in the context of examining differential expression), but that this one is specifically tailored towards GWAS results (or any other distinguishing features that you would like to point out).

I’ve added a note about PAST’s specific design as a GWAS tool.

--Line 40: would suggest changing 'FDR threshold' to 'FDR or another type of significance threshold', given that not all GWAS-involving analyses use FDR.

I have edited as suggested.

--Is it correct that the metabolic pathway annotations/nomenclature being used are based solely on the input file from the user? It seems so given that the tool would otherwise not be free of species-specific assumptions (as mentioned in line 59), given that it would have needed to assume a certain master set of pathways/model of plant metabolism if not accepting external pathway identifiers (as do other mapping/visualization tools). If solely based on user input, it would likely be good to mention common sources of this data (GFF annotation data and pathway information) and to encourage caution with annotations (particularly if electronically inferred). These would be in addition to the warnings already included, which state that the output is only as good as the input data (so that those warnings are not interpreted to be referring only to the genotypic and phenotypic data). It may also be good to highlight (if true) that the tool will only analyze pathways that are labeled with exactly the same identifier (e.g., if one is upstream of the other but is labeled differently, that they will be analyzed as separate pathways).

Yes, that's correct. I've added some notes about being wary of the quality of the annotations and that the gene names in the annotations should match those in the pathways file.

--Perhaps this is already planned, but it would be great in the video to show the two workflows (Shiny and R console) separately, as it is otherwise a bit tricky in written form to see the complete workflow in either case.

I think this would be the best approach as well.

--Lines 143-146: can background LD level also be specified, if examining genes in the vicinity of a significant SNP (or is this not a view that will be enabled)? Can a reference be mentioned that describes which values are standard in maize?

No, PAST doesn't yet have a way to specify background LD. That may be worth investigating as we continue to improve PAST. Tang *et al* (2015) describes how the default R^2 was decided upon.

--Line 172: it seems that there optimally would not be a default for search space window, rather would need to be the result of some analysis relevant to the data set being analyzed. That said, 1 kb is a very conservatively small window, so seems OK (just may have lots of false negatives, but seems better to aim on the too-small side than too-large).

We do expect that those who use PAST will change this parameter to reflect their knowledge of their species.

--Line 174: could the default be set to be half of the cores that the user has available? It seems that this would greatly improve the processing time (which was mentioned to be long even in this improved version), if parallel processing is already enabled.

It's technically possible, but letting the user set this on their own is better. Users should explicitly grant tools permission to use more resources; PAST has no way of knowing whether a user is running resource-intensive processes using the rest of their cores. Based on this feedback, the default has been reduced to 1 in PAST Shiny. The code examples still use 2 cores, since the user is explicitly specifying the number of cores.

--Lines 187-192: Does significance refer solely to significance of effect estimate for a SNP near a gene? And 'significant pathways' (as mentioned in lines 189-190) would then be the pathways having more than the cutoff number of genes with significant effect estimates? Or has another test been run to determine which pathways are 'significant'? Please be very specific and explicit re: which 'significance' tests are being run within this tool.

Another test has been run, but the details of this test are described in other manuscripts. I've added a line to inform readers where they can find details on how pathways are tested for significance.

--Line 218: 'associated' may not have been intended to be taken literally, but *SNPs* are 'associated' in a GWAS context, not genes or pathways (unless some of the involved significance tests explicitly test for significance of association of a pathway with marker-trait associations? Pathways are not themselves being tested in GWAS.) Perhaps another word such as 'corresponding to', to avoid confusion given that this paper somewhat operates in a GWAS context?

I changed this line to match line 201, since the parameters are the same and should be described in the same way.

--Line 243: are enrichment scores (statistically) of such a nature that they can be cleanly added? One example: PVEs are not quite additive. A bit more detail re: gene set enrichment analysis (as a method) would be useful, if that is what has been used here. (Gene set enrichment analysis was what came back when looking up 'enrichment scores' online.)

Because this method was described in the release paper for PAST, we didn't discuss it here. However, one of the citations added to address concerns with lines 187-192 does link to Subramanian et al's GSEA paper, where the details of the calculation are described.

--Lines 272-273: some mention of within-species variation might be worthwhile. For example, it has been found that LD patterns vary substantially even between the various subpopulations of maize. The same has been seen in wild vs. cultivated barley.

I’ve expanded our description to indicate that the user generates LD and thus the LD measurements are specific to the population they’re studying.

--Figure 1: Would all 'significant' pathways be depicted within this view? Could the names of the genes and/or number of SNPs in their vicinity exhibiting significant association with the trait under analysis be somehow depicted?

All significant pathways are depicted in this view, but the size of the image changes depending on how many pathways there are. Individual pathway images are written to the results folder or downloaded from PAST Shiny. The names of the genes are included in the table output that is written to the results folder or downloaded from PAST Shiny. Depicting more information becomes complicated for pathways with many genes, as the image gets quite crowded.