shinyGISPA: A shiny tool for Gene Integrated Set Profile Analysis

Bhakti Dwivedi and Jeanne Kowalski
Winship Cancer Institute, Emory University, Atlanta, 30322, USA

Introduction

shinyGISPA is a web-based tool intended for the researchers who are interested in defining gene sets within the context of similar, a priori specified molecule profile. While the GISPA method (Kowalski et al., 2016) was developed to address genome-wide comparisons of three groups based on as few as a single sample per group, in terms of profile changes from several genomic data types (e.g., gene expression, methylation, copy number, etc.), comparisons may also be done based on a single data type. In this setting, the GISPA approach represents an extension to prior methods developed for addressing several single sample comparisons based on a single genome-wide data type. (Kowalski et al. 2004a; Kowalski et al. 2004b). The tool is developed using shiny, a web-application framework for R. Using this tool, user combine and compare several genome-wide data types from three sample classes (or groups) to find the gene sets with genomic changes specific to a sample class.

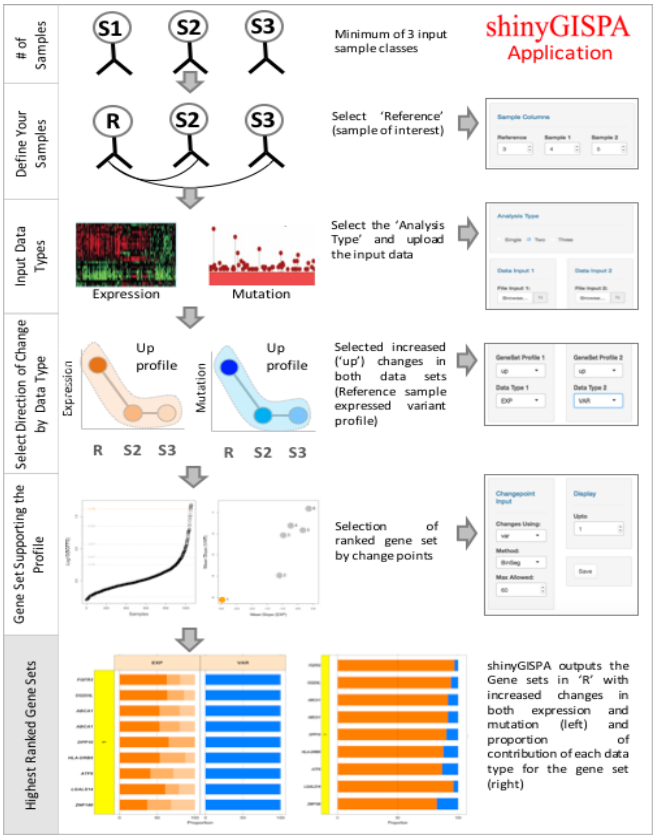


Figure 1. Schematic representation of the shinyGISPA method overview. (A) Number of samples required to run shinyGISPA. (B) Define your samples: User can define a reference sample (or sample of interest) and the remaining two comparison samples by specifying the sample columns in the input data set. (C) Input data types: User can perform either a single, two or three-feature analysis by clicking on options under 'Analysis Type'. (D) Select direction of change by data type: User uploads the input data file and profile of interest within each data type. (E) Diagnostic plots showing gene sets that support the profile of interest in the reference sample: (F) Highest Ranked Gene Sets visuals in terms of differences among samples and data types selected.

Running shinyGISPA on our private server

- **http://bbisr-tools.winship.emory.edu:3838/shinyGISPA/ (within Emory))**
- **http://shinygispa.winship.emory.edu/shinyGISPA/ (anywhere)**

Prerequisites for running on your local machine

1) Download and install R or RStudio (version 3.1.2. or later) from https://cran.r-project.organd
2) Open R and install the below required packages:
   > install.packages(c("changepoint", “colourpicker”, "data.table”, “genefilter”, “ggplot2”, “graphics”, “HH",, ”latticeExtra", “plyr”, ”scatterplot3d", “stats”, “splitstackshape”, “shiny”))

3) Install packages not available for the R version from Bioconductor
   > source("http://bioconductor.org/biocLite.R")
   > biocLite("*package.name*")

4) Users can run shinyGISPA locally using the source code available from the GitHub: https://github.com/BhaktiDwivedi/shinyGISPA, by typing the below commands in R console:
   > library(shiny)
   > runApp("shinyGISPA")

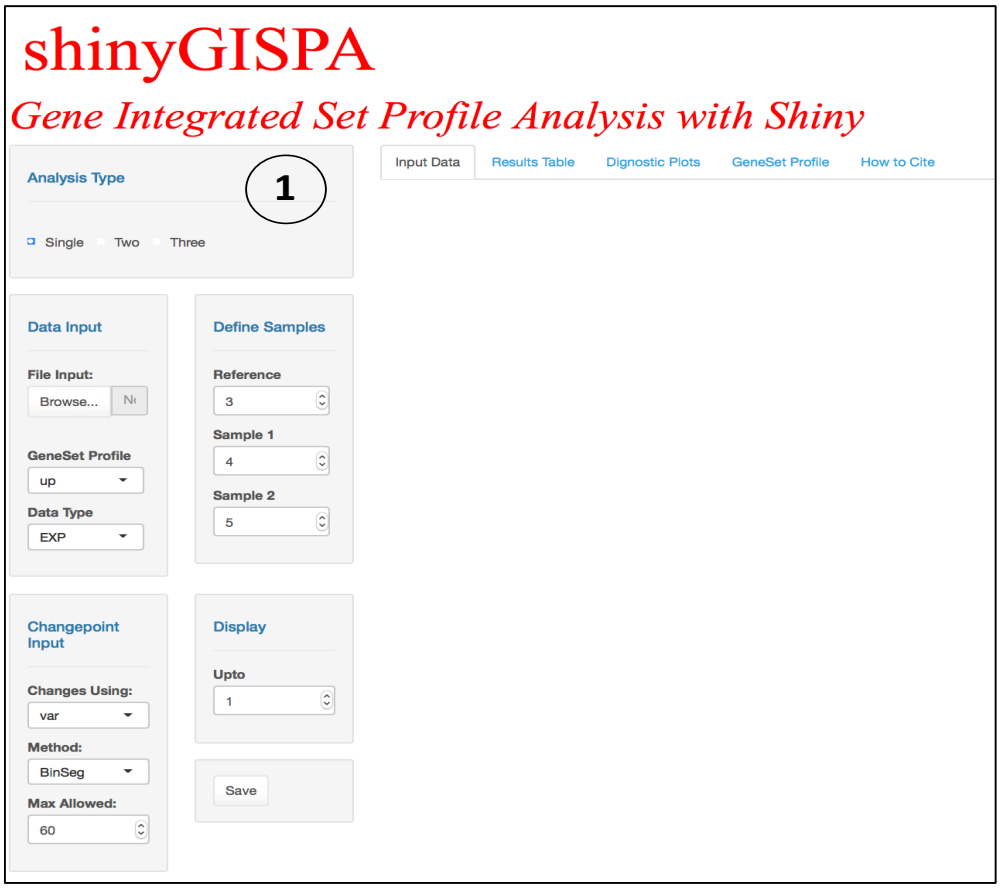
5) Users can also download and run the app from GitHub directly using:
   > shiny::runGitHub('shinyGISPA', 'BhaktiDwivedi')

6) GISPA Bio-conductor package is also available https://bioconductor.org/packages/GISPA/

Getting Started

1. Select the analysis type

Click on options under “Analysis Type” to select a single, two, or three-feature analysis. Here *feature* is defined as a specific data type (e.g., expression, methylation, somatic mutation, copy number variation).



An example of single-feature analysis is identifying gene sets with expression changes, a two-feature analysis is based on a combination of any two data types, e.g., identifying gene sets that exhibit gene expression and copy change changes, while a three-feature analysis is based on a combination of any three data types e.g., identifying gene sets that exhibit expression, copy number, and methylation changes.

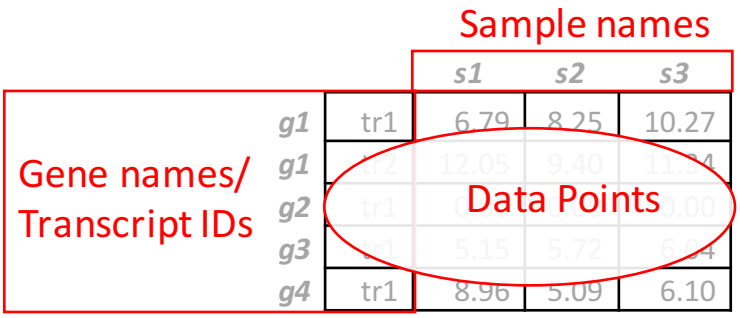
2. Upload the Data

Upload the input data file given the selected feature type from (1). User uploads the input data file and profile to define gene sets on within each data type. The input data can be genome-wide or based on prior knowledge derived from either biological processes, pathways, biomarkers discovery, or genomic analysis. Here a *profile* is a genomic change representing either increase ("up") or decrease ("down") within a specific feature or data type. The data types include expression ("EXP"), somatic mutation or variant ("VAR"), copy number change ("CNV)"), and methylation ("MET")

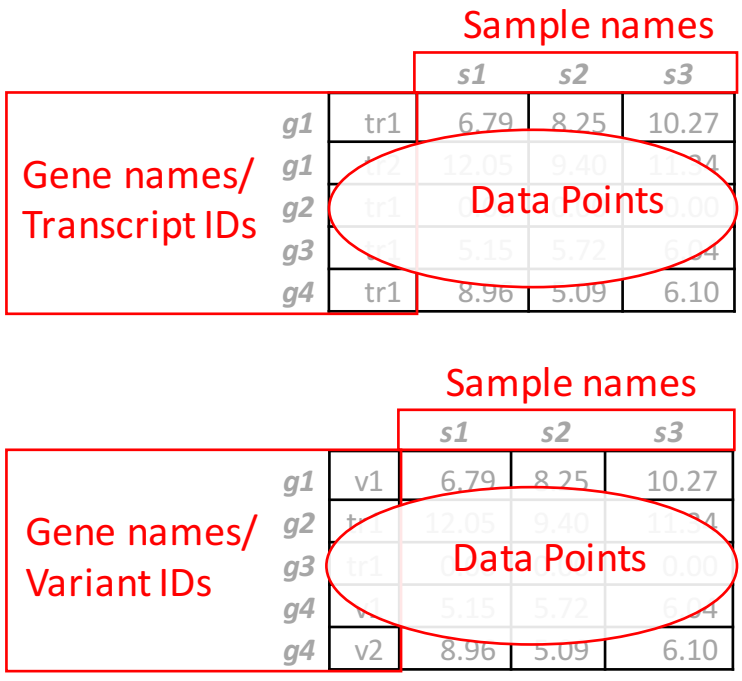
File format requirements:

- Maximum file size limit of up to 500 MB.

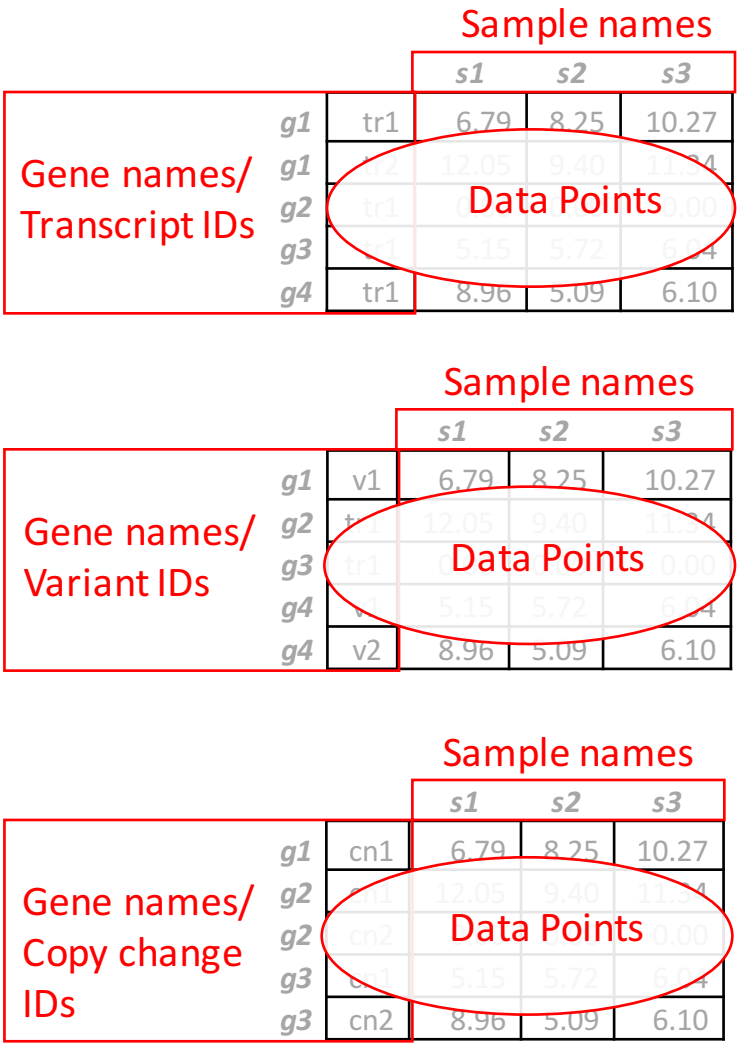
- ASCII formatted tab-delimited file, where each row represents a gene (or related gene id) and each column a sample.
- Here is an example of a user uploaded 'File Input' for single-feature analysis. First and second column correspond to gene names and gene id's (e.g., gene transcript or ensemble id) followed by the three sample classes data as shown in the screenshot below:



- Here is an example of user uploaded 'File Input 1' and File Input 2' for two data type analysis. For each input data file, first and second column correspond to gene name and gene ids (e.g., gene transcript or ensemble id) followed by the three samples data as shown in the screenshot below:



- Here is an example of user uploaded 'File Input 1', File Input 2', and File Input 3 for three-feature analysis. For each input data file, first and second column correspond to gene name and gene ids (e.g., gene transcript id, variant id, copy number segment id) followed by the three samples data as shown in the screenshot below:



- When running shinyGISPA on mutliple data types, the first column (or gene names) must overlap between the input data files. The two or three input data files must have the same exact gene names or IDs in the first column as these are used to merge the data files into a single file for analysis. The number of rows representing gene names probes, variants, or any other id may or may not be the same.
- A minimum of at least 10 genes and three sample classes are required.
- No duplicated sample (column) names or duplicated gene (row) names are allowed and analysis will be stopped.
- Gene with zero variance across all samples will be excluded from the analysis.

Example Snapshot of shinyGISPA using Two-feature analysis:

3. Define Sample classes

User defines sample classes (or groups) in the uploaded data set by specifying the sample columns corresponding to 'Reference' (sample of interest) and two other samples to compare the reference sample against.

4. Select the Changepoint Method

User can specify/modify change point detection method (Killick R, et al., 2016) to find the optimal break points within the estimated profile sample score (Kowalski, et al., 2016). The changes can be found in mean and/or variance using the user-specified method (“AMOC”, “BinSeg”, “PELT”, or “SeqNeigh”) given the allotted maximum number of change points.

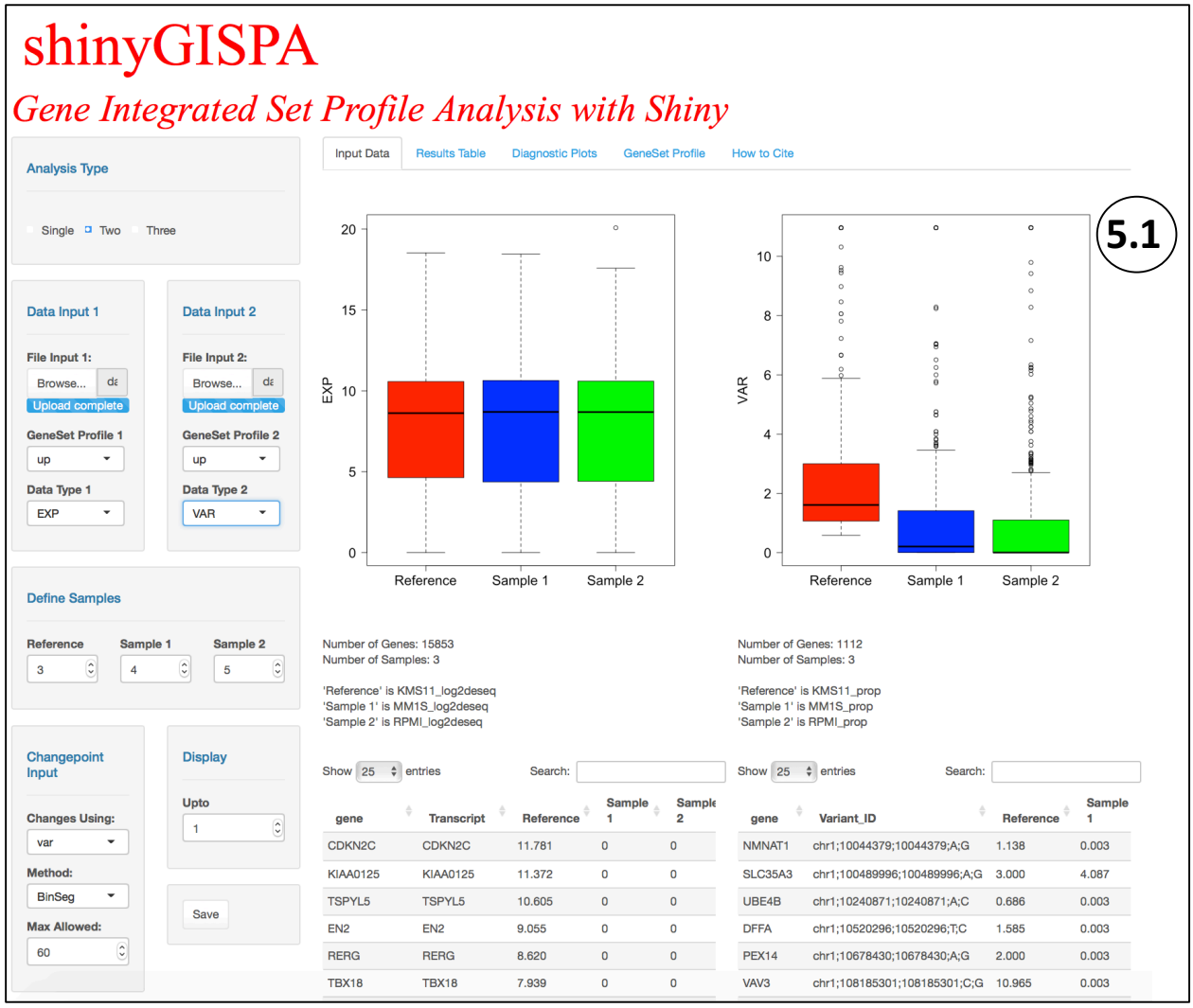
5. Result

The results are output in four separate tabs:
5.1 Input Data
5.2 Results Table
5.3 Diagnostic Plots
5.4 GeneSet Profile

5.1 Input Data

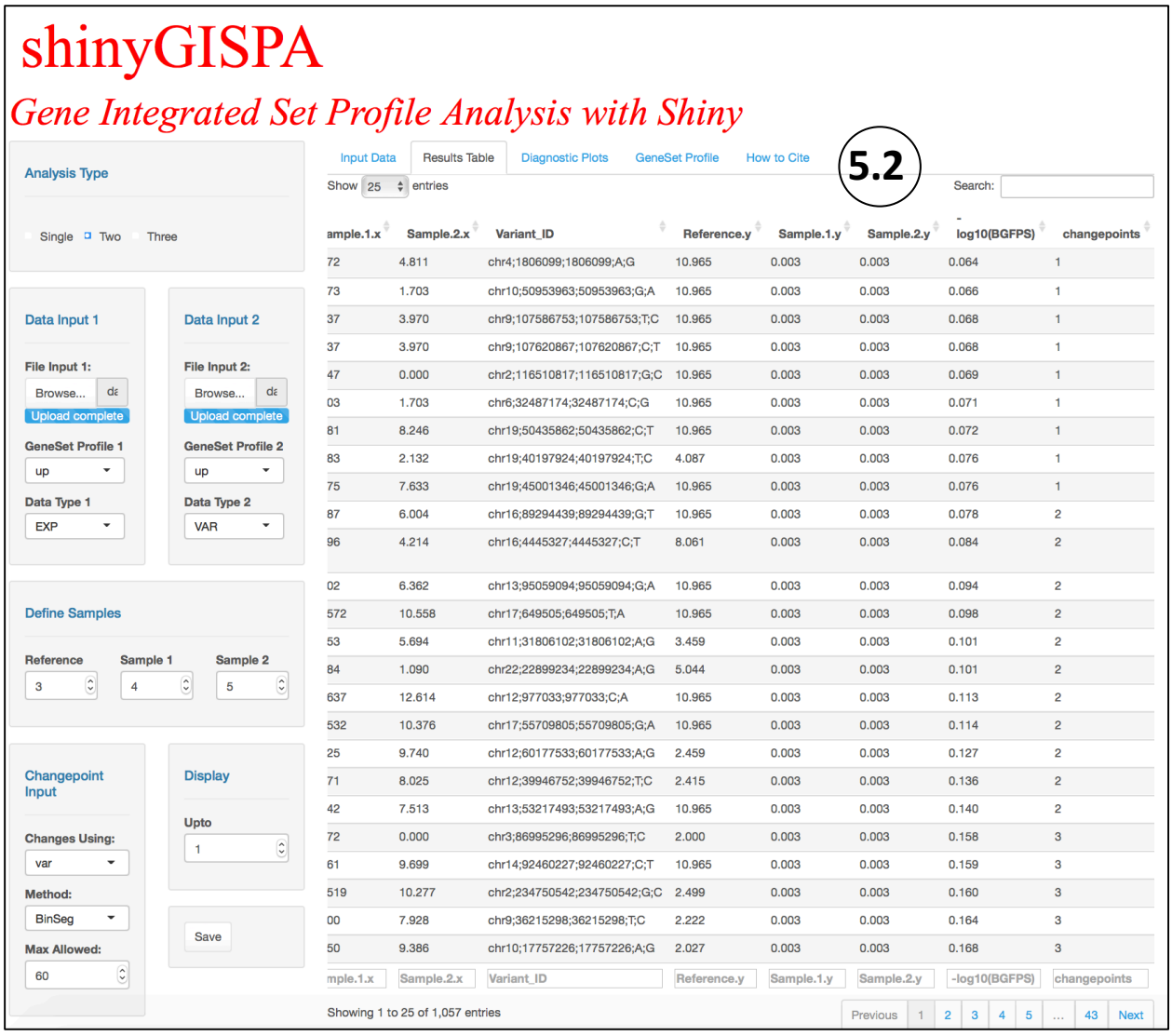
Summarizes the user input data in terms of the input number of genes (or rows), number of samples (or columns), user-defined reference and comparison samples (sample 1 and sample 2). Note that genes (or rows) with zero variance among all the three samples will be excluded

from the analysis. User can also select the color palette of choice to represent the three sample classes.



5.2 Results Table

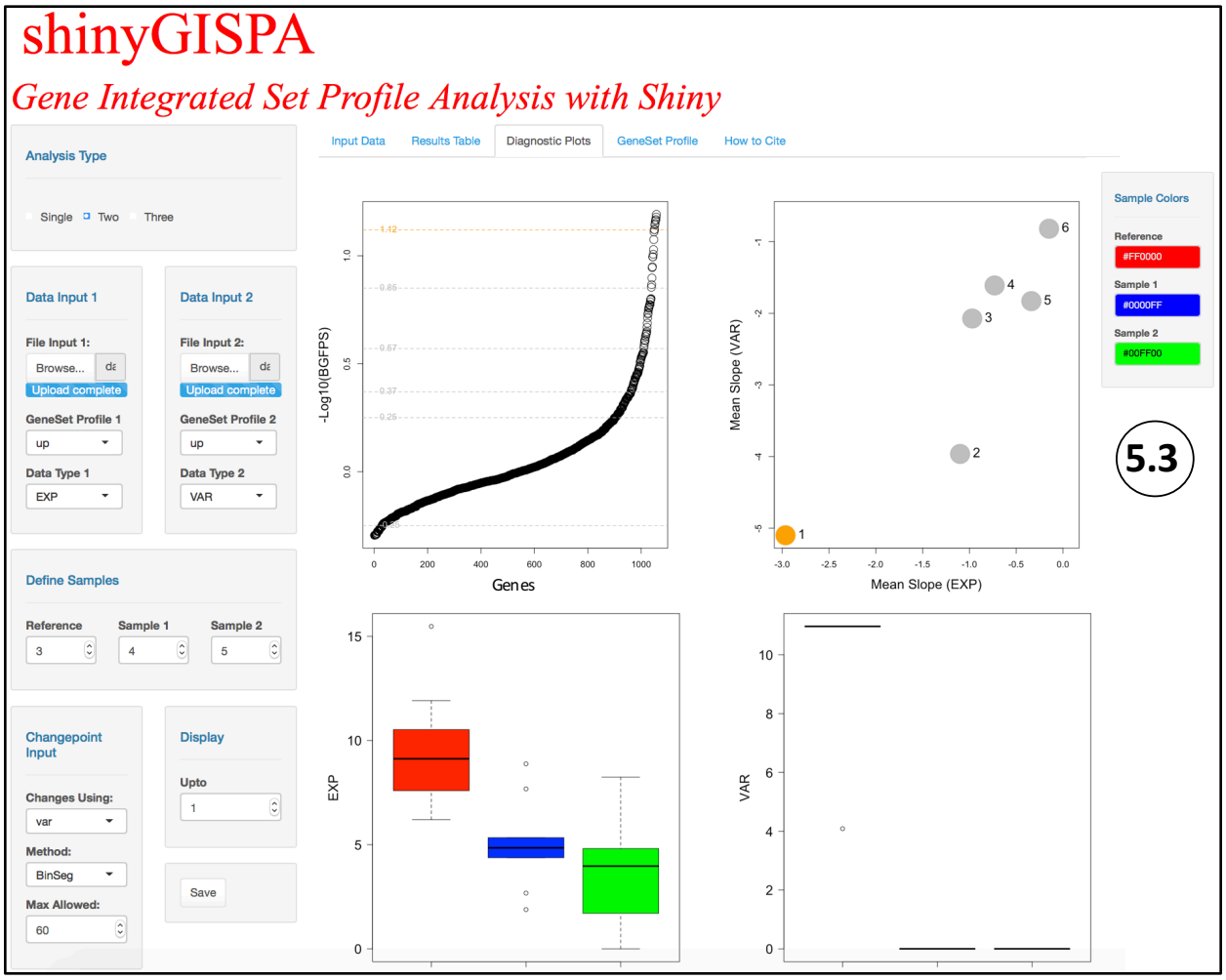
Table of ranked gene sets by their gene feature profile statistics scores for the user-selected data types and profile. The profile statistics scores for each gene are computed using the GISPA method (please see method details in Kowalski et. al., 2016). The score statistics is rank ordered by the user selected profile (e.g., up or down) for each gene. A change point model (Killick et. al., 2016) is then applied to the ranked scores to identify gene set that show similar molecular profile. Gene sets shown are grouped by change points. The results table can be searched, sorted, and filtered by any of the columns.



5.3 Diagnostic Plots

shinyGISPA generates, (1) an ordered plot of smallest (least desirable) to largest (most-desirable) between-feature profile statistic for each gene (circle) with breakpoints for cutting the data to define gene sets that support the molecular profile of interest. The yellow line is change point 1, such that the genes above it show the most support for the profile in characterizing the reference sample versus the others. (2) Diagnostic slope plot to determine the number of change points and therefore gene sets in support of the user-defined profile. Within each gene, changes in expression and methylation are summarized among the three samples by the calculation of a slope. These gene slopes are summarized among gene sets defined by each change point by taking their average; each circle represents a ranked gene

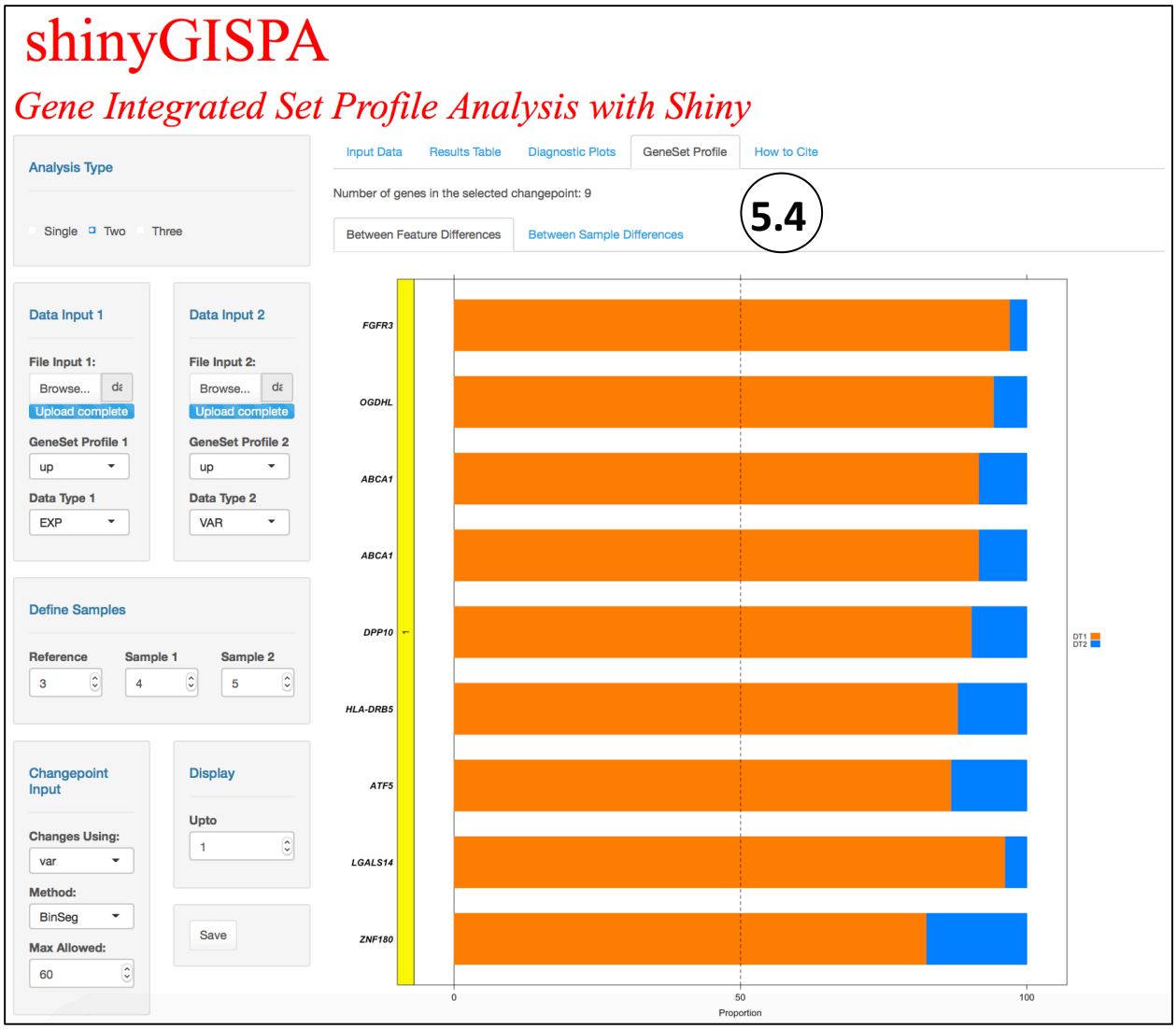
set grouped by their change points, with topmost, “best” profile (change point 1) shown with orange-filled dot. If the gene sets satisfy an increase profile of interest (i.e. increase gene expression with increase copy number), the circles tend to be at the lower-left corner. On the other hand, if the gene sets satisfy a decreased profile of interest (e.g. decrease gene expression with decrease copy number), the dots tend to be at the upper-right corner. (3) Boxplots of the distribution of each feature for genes identified from change point 1 by sample that highlight the desired profile.

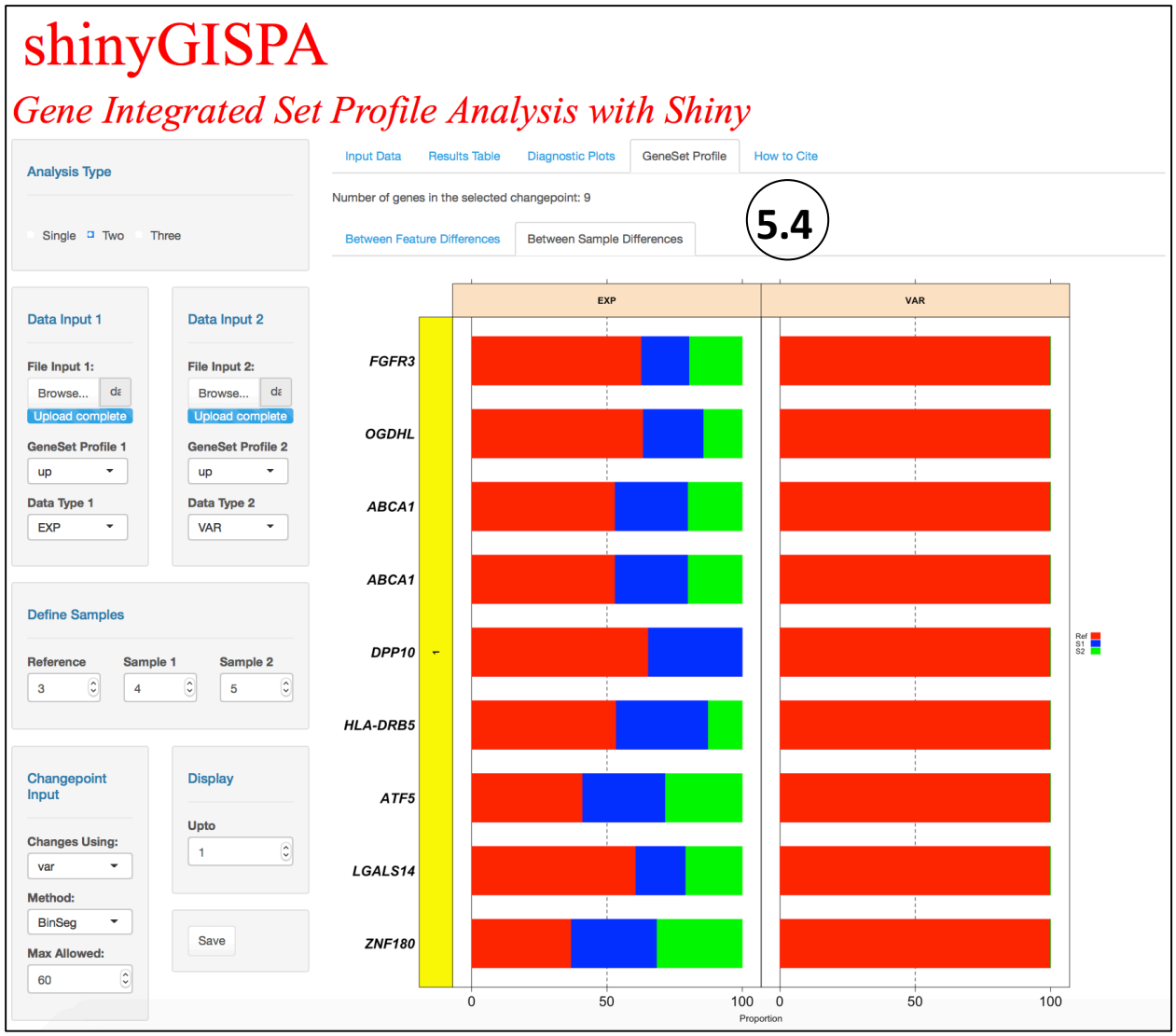


5.4 GeneSet Profile

shinyGISPA generates stacked bar plots using HH R package (Heiberger 2016) of the ranked gene sets profiles to depict their distribution based on observed input data (e.g., expression values, copy segment mean) in the reference relative to other samples. This enables the users to visualize the level-vise breakdown of each data type, whether or not gene set satisfy the profile of interest, and if not, is there a particular data type that appears to be prominent for

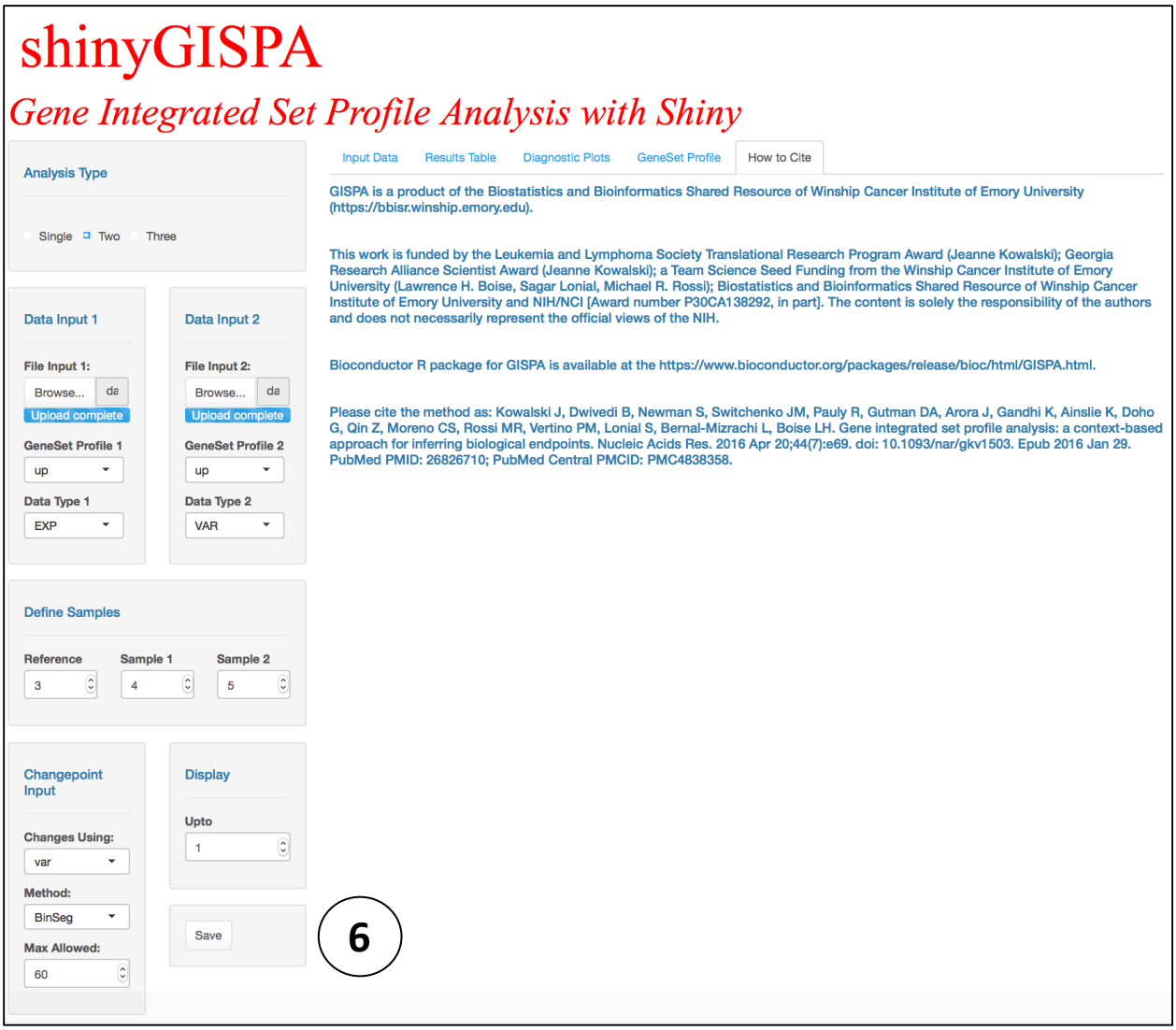
a particular gene or gene sets profile. The Between-Feature Differences represent the percent contribution from each feature or data type to the gene profile displayed, while Between Sample Differences represents the differences among the samples, i.e., the percent contribution from each sample to the summed total of each feature. User can adjust the gene labels font size, axis text size, and gaps between the plots using the respective options in the left side panel.





6 Save Results

User can download the 'Results Table" as a csv file by clicking on the "Download" button on the left panel. The pdf plots of the results shown, can be copied and saved on the local machine.



Funding

This work is funded by the Leukemia and Lymphoma Society Translational Research Program Award (to Jeanne Kowalski); Georgia Research Alliance Scientist Award (Jeanne Kowalski); a Team Science Seed Funding from the Winship Cancer Institute of Emory University (Lawrence H. Boise, Sagar Lonial, Michael R. Rossi); Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI [Award number P30CA138292, in part]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Citation

Please cite the GISPA method as: Kowalski J, Dwivedi B, Newman S, Switchenko JM, Pauly R, Gutman DA, Arora J, Gandhi K, Ainslie K, Doho G, Qin Z, Moreno CS, Rossi MR, Vertino PM, Lonial S, Bernal-Mizrachi L, Boise LH. Gene integrated set profile analysis: a context-based approach for

inferring biological endpoints. *Nucleic Acids Res.* 2016 Apr 20;44(7):e69. doi: 10.1093/nar/gkv1503. Epub 2016 Jan 29. PubMed PMID: 26826710; PubMed Central PMCID: PMC4838358.

References

1) Kowalski J, Drake C, Schwartz RH, Powell, J. Non-parametric, Hypothesis-based Analysis of Microarrays for Comparison of Several Phenotypes. (2004a). *Bioinformatics* 20: 364-373.
2) Kowalski J, Powell J. Nonparametric Inference for Stochastic Linear Hypotheses: Application to High Dimensional Data.(2004b). *Biometrika.* 91: 393-408.
3) Kowalski J, Dwivedi B, Newman S, Switchenko JM, Pauly R, Gutman DA, Arora J, Gandhi K, Ainslie K, Doho G, Qin Z, Moreno CS, Rossi MR, Vertino PM, Lonial S, Bernal-Mizrachi L, Boise LH. (2016). Gene Integrated Set Profile Analysis: A Context-Based Approach for Inferring Biological Endpoints. *Nucleic Acids Research* doi:10.1093/nar/gkv1503.
4) Killick R, Haynes K and IA, E. changepoint: An R package for changepoint analysis. *R package version 2.2.1* 2016.
5) Heiberger, R.M. HH: Statistical Analysis and Data Display: Heiberger and Holland. *R package version 3.1-32* 2016.