IsoGeneGUI Package Vignette

Setia Pramana, Martin Otava, Dan Lin, Ziv Shkedy

February 23, 2015

1 Introduction

The IsoGene Graphical User Interface (IsoGeneGUI) is a user friendly interface of the IsoGene package which is aimed to perform analysis of dose-response studies in microarray experiments. Compared to original package, GUI is extended thoroughly by addition of methods from other packages related to gene expression analysis, namely GORIC, ORCME, ORIClust and orQA. Hence, it provides wide range of tools and covers all the range of methods currently available in are in "non-GUI" packages. The IsoGeneGUI is developed for the user with no or limited knowledge about R programming so he/she can perform the analysis of dose-response in microarray setting easily. This GUI was developed using tcl/tk package. The statistical methodologies (test statistics, etc.) used in this package are discussed by Lin et.al (2007, 2008, 2010).

2 Installation

The package can be installed directly from R environment. For proper installation, both CRAN and Bioconductor repositories need to be specified.

```
> setRepository()
> install.packages("IsoGeneGUI")
```

3 Usage

To run the package

```
> library(IsoGeneGUI)
> IsoGeneGUI()
```

4 Menus

The main window of the IsoGene-UI is presented in Figure 4. The package has five main menus: File, Analysis, Clustering, Plots and Help. In the middle of the main window there is an info box which provides information about the data (availability and summary) and the result summary of the last performed analysis.

1. File:

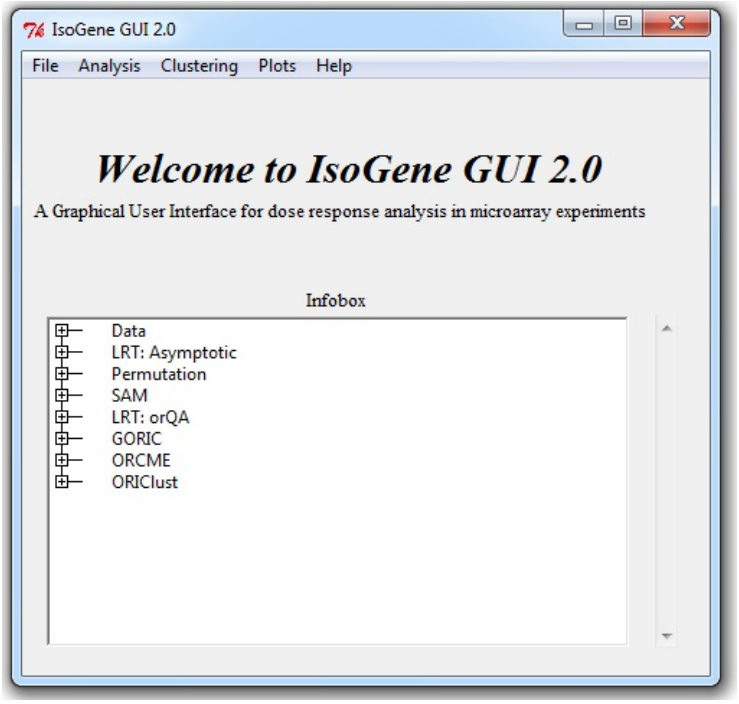

Figure 1: The main window of the ISoGeneGUI.

   (a) Open dataset
      i. R workspace (*.RData files)
      ii. Excel or Text file (*.xls or *.txt files)
   (b) Show dataset
   (c) Exit

2. Analysis:
   (a) Set seed
   (b) Likelihood Ratio Test (E2): Asymptotic
   (c) Permutation
   (d) Significant Analysis of Microarrays (SAM)
      i. SAM Permutation
      ii. SAM Analysis
   (e) Likelihood Ratio Test (E2): orQA
   (f) GORIC
3. Clustering:
   (a) ORCME
   (b) ORIClust
4. Plot:

(a) IsoPlot

(b) Permutation Plot

(c) SAM Plot

i. Plot of FDR vs. Delta

ii. Plot of number of significant genes vs. Delta

iii. Plot of number of False Positive vs. Delta

(d) User defined scatter plot

(e) Plot ORCME clusters

5. Help:

(a) IsoGene Help

(b) IsoGeneGUI Help

(c) About

5 Reading the Data

In the first step of the analysis, the data is uploaded to the package using the File menu. The package can read data in R workspace file (*.RData), Microsoft Excel and text files. Once the the data is uploaded, information about the data is presented automatically in the info box of the main window.

5.1 R workspace files

The format of the gene expression data should be a matrix or table where the columns are the arrays and the rows are the genes. The dose information should be a vector or table that contains the dose levels which corresponds to the arrays in the expression matrix/table.
The package provides an R workspace example dataset called `dopamine2`. For the dopamine2 data, the dose levels are given by

```
> dose
 [1] 0.00 0.00 0.01 0.01 0.04 0.04 0.16 0.16 0.63 0.63
[11] 2.50 2.50 0.00 0.00 0.00 0.01 0.01 0.01 0.04 0.04
[21] 0.16 0.16 0.63 0.63 2.50 2.50
```

while the expression matrix has the following structure:

```
> dopamine[1:5, 1:6]
             X1        X2         X3        X4        X5
201_at  2.579138  2.318749  2.496895  2.456772  2.479480
202_at  2.140561  2.061804  2.131749  2.107638  2.086722
203_at  6.988566  6.620562  5.764725  6.326178  7.020716
204_at 11.081855  9.974999 10.790689 10.702516 10.544664
205_at 12.104545 12.076975 11.989770 12.151120 12.118520
```

Part of the `dopamine2` can be obtained in folder "exampleData" inside the package.

The full example data in R workspace, text and Excel files can be obtained from: `http://ibiostat.be/online-resources/online-resources/isogenegui`

5.2 Open the data set

In order to upload the R workspace we choose in the file menu in Figure 4a the following sequence:

```
File > Open dataset > R workspace
```

The package will automatically refer to the example data `dopamine2` if the option to open R workspace is choosen.

5.3 Text and Excel files

Text files and Excel files can be uploaded to IsoGene-GUI as well using following sequence in the dialogue box:

```
File > Open dataset > Excel or Text files
```

The format of the Text and Excel files should be a matrix where the columns are the arrays and the rows are the genes. Optionally, the first column can contain gene names and header can be used. The dose information should be a single column containing the dose levels which corresponds to the arrays in the expression matrix/table.

6 Exploratory Data Analysis

To explore the expression of the genes, the IsoGeneGUI provides the submenu **IsoPlot** from **Plots** menu. This feature displays the data points, sample means at each dose and an isotonic regression line.

There are three input options to draw the isotonic regression plot, using gene name(s), row number(s) or using a range of row numbers. There are also three check boxes:

1. Dose as ordinal. This option will draw the plot and treat dose as ordinal scale. The default will plot with dose as a continuous variable.

2. Show isotonic regression curve for both direction. The default plot just display the isotonic trend/curve which is more likely fit the data. By checking this option, the isotonic trend for both directions will be displayed.

3. Show summary of the data. This option is to provide a short summary of each selected gene.

7 Data Analysis

The **IsoGeneGUI** package provides four options for analysis: (1) analysis with the likelihood ratio test statistic (LRT) using its exact p-values, (2) resampling based analysis (from IsoGene and orQA package), (3) significance analysis of microarrays (SAM) and (4) generalized order restricted information criterion (GORIC).

7.1 LRT Using exact p-values

To perform the exact LRT we choose the following sequence in the analysis menu:

```
Analysis > Likelihood Ratio Test (E2): Asymptotic
```

Then the main dialog box for analysis based on the LRT using exact p-values is shown in Figure 2.

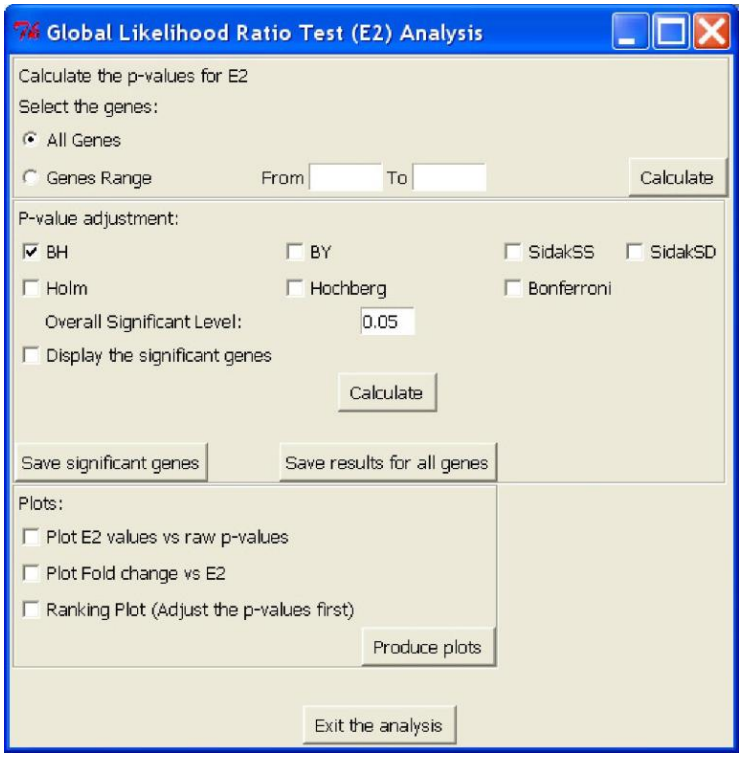

Figure 2: The main dialog box of Likelihood Ratio Test E^2 analysis.

Note that the users can choose to perform the analysis for all the genes on the array or on predefined subset. In addition, in order to adjust for multiplicity, we need to select from the menu the adjustment method to be used and the overall error rate.

7.2 Resampling Based Methods

Resampling based analysis can be performed by choosing either of following options in analysis menu:

```
Analysis > Permutation
Analysis > Likelihood Ratio Test (E2): orQA
```

first option implements function from IsoGene package and offers five different test statistics as shown in Figure 3 that presents the main menu for

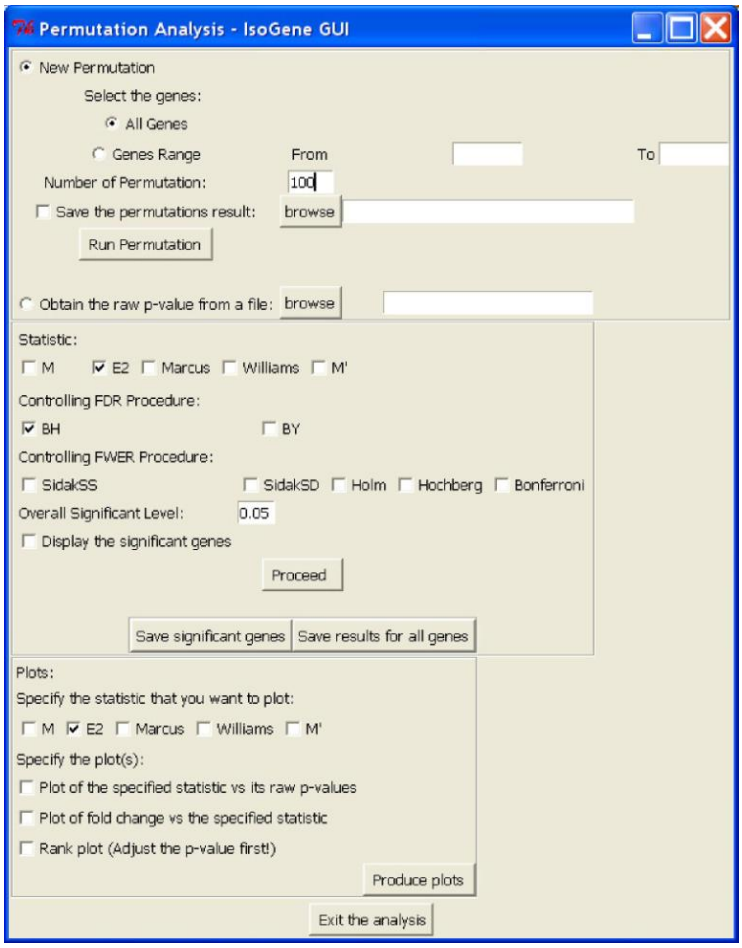

Figure 3: The main dialog box for resampling-based monotone trend test.

resampling method. Second option implements functions from package orQA and provides virtually same menu as in Figure 2. However, computation is done using permutation test. The orQA implementation is much faster than the one from IsoGene, but its restricted only to E2 statistics. Hence, iif it is only interest, we recommend orQA implementation. If you are interested in other test statistics, Permutation option is proper choice. In both options, we need to specify the number of permutations and the multiplicity adjustment(s). Similar to the E^2 menu, we can specify default graphical displays.

7.3 Significance Analysis of Microarrays (SAM)

The **IsoGene-GUI** also provides testing for the dose-response relationship under order restricted alternatives using the Significance Analysis of Microarrays procedure (SAM). To perform this analysis there are two main steps: (1) calculating the SAM regularized test statistic using permutation and (2) the SAM analysis. These two steps are performed by two sub menus in the SAM menu.

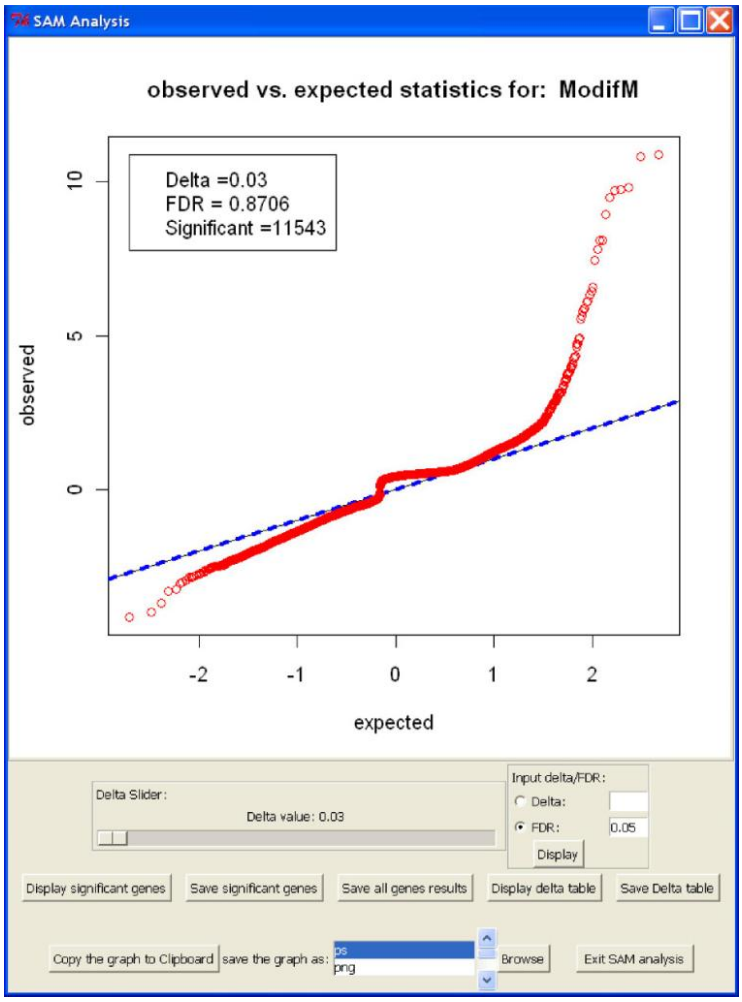

Figure 4: The main dialog box for SAM analysis.

```
Analysis > Significance Analysis of Microarrays > SAM Permutation
Analysis > Significance Analysis of Microarrays > SAM Analysis
```

Then in the first step several options of the use fudge factor are provided. User can specify to use **no Fudge Factor**, **Automatic Fudge factor** (fudge factor will be calculated using the methods described in the SAM manual), or the **fudge factor based on a specific percentile**.

After the statistics are computed or loaded from a file, we now can perform the SAM analysis. After specifying the test statistic, the main dialog box of SAM analysis, which shows the plot of the observed versus the expected test statistics and other options, will be shown (see Figure 4).

Note that, in this plot we can change the delta value (automatically using delta slider or manually using delta input box) and the FDR level.

7.4 Generalized Order Restricted Information Criterion (GORIC)

The GORIC method (Kuiper et al., 2014) focuses on model selection rather then estimates. Its implementation in the **IsoGeneGUI** package fits all possible models under monotonicity assumption and output their posterior probability based based on information criterion. The information criterion used takes into account order constraints and adjust the penalty accordingly. The posterior probability of model is computed via simulations. Therefore, the procedure is computationally intensive and in the pacakge, it is allowed to compute it for one gene at the time only. Therefore, it should be applied only on genes selected by other analyses as especiallz itneresting ones. The resulting posterior probabilities can be used for model selection purposes.

`Analysis > GORIC`

8 Clustering

The **IsoGeneGUI** package provides two methods for gene clustering according to their profiles: (1) δ-clustering method that focus on order restricted clustering with monotonicity assumption (ORCME), (2) information criterion based clustering (ORIClust).

8.1 Order Restricted Clustering for Microarray Experiments (ORCME)

Clustering analysis performed by ORCME package is based on delta-biclustering algorithm (Cheng and Church, 2000). It clusters genes according to their overall profile of isotonic means. The homogeneity of selected clusters can be influenced by appropriate parameter value. The parameter can be optimized with respect to within cluster variability and number of discovered clusters. Compared to ORIClust, the method can create more homogeneous clusters , while focusing on smaller space of possible profiles (monotone ones)

`Analysis > ORCME`

8.2 Order Restricted Information Criterion Based Clustering Algorithm (ORIClust)

The ORIClust performs one-stage or two-stage clustering algorithm based on information criterion ORICC (Liu et al., 2009). The genes are clsutered according to their profiles, while taking into account ordering of doses and shape of dose-response relationship. ORIClust provides wider range of possible profiles, such as umbrella profiles, compared to ORCME.

`Analysis > ORIClust`

9 Saving IsoGeneGUI Outputs

Each plot produced from the analysis in IsoGeneGUI can be copied into clipboard which can be easily pasted into Microsoft Office documents, such as Microsoft Word. It can be done by clicking the **Copy to Clipboard** button located in the bottom of the graph and then paste it directly to the document (or `Ctrl-V`). Furthermore, the plots can also be saved into several image formats (ps, png, jpeg, bmp, tiff). After the format is selected/highlighted from the list, the user can click **Browse** button to specify the name of the image file and also the file location.

Note that results of all genes and lists of significant genes from the three analysis above can be shown and also saved into an `R-workspace` and/or an Excel file by clicking **Save** button.

10 Complete users' manual

Users can access the IsoGeneGUI documentation online or by installing the IsoGeneGUI package locally. Then the following code can be typed at R prompt:

```
> if (interactive())
+     {
+         browseURL("http://ibiostat.be/online-resources/online-resources/isogenegui")
+     }
>
```

or alternatively:

```
> if (interactive()) {
+ library(IsoGeneGUI)
+  IsoGeneGUIHelp()
+ }
```

Users can download example dataset and a complete Users' Manual from the site as well.

References

1. Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In: *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, **1**: 93-103

2. Kuiper, R. M., Gerhard, D., and Hothorn, L. A. (2014), Identification of the Minimum Efeective Dose for Normally Distributed Endpoints Using a Model Selection Approach, *Statistics in Biopharmaceutical Research*, **6(1)**: 55-66.

3. Lin, D., Shkedy, Z., Yekutieli, D., Burzykowski, T., Göhlmann, H., De Bondt, A., Perera, T., Geerts, T. and Bijnens, L. (2007) Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resamplingbased inference. *Statistical Applications in Genetics and Molecular Biology*, **6(1)**, Article 26.

4. Lin, D., Shkedy, Z., Burzykowki, R., Ion, T., Göhlmann, H.W.H., De Bondt, A., Perera, T., Geerts, T. and Bijnens, L. (2008) An investigation on performance of Significance Analysis of Microarray (SAM) for the comparisons of several treatments with one control in the presence of small variance genes. *Biometrical Journal*, Multiple Comparison Problem, Special Issue, **50(5)**, 801–823.

5. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D. and Bijnens, L., editors (2012) *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*. Springer.

6. Liu, T., Lin, N., Shi, N. and Zhang, B. (2009), Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments. *BMC Bioinformatics*, **10**, 146