

statTarget:

# Statistical Analysis of Metabolite Profile

Hemi Luan

August 19, 2016

## Contents

<b>1</b>	<b>statTarget installation guide</b>	<b>3</b>
1.1	Requirements . . . . .	3
1.2	For Mac OS users . . . . .	3
1.3	For Windows users . . . . .	3
<b>2</b>	<b>Overview of statTarget</b>	<b>4</b>
<b>3</b>	<b>statTarget in detail</b>	<b>4</b>
3.1	statTarget Shift Correction . . . . .	5
3.1.1	Parameters . . . . .	5
3.2	statTarget Statistical Analysis . . . . .	6
3.2.1	Parameters . . . . .	7

<b>4</b>	<b>References</b>	<b>8</b>
<b>5</b>	<b>Session Info</b>	<b>9</b>

# **1 statTarget installation guide**

## **1.1 Requirements**

Dependent on R ( $\geq 3.3.0$ )

## **1.2 For Mac OS users**

1 - Install XQuartz from <https://www.xquartz.org>

2 - Copy this code into R.

```
> source("https://bioconductor.org/biocLite.R")
```

```
> biocLite("impute")
```

```
> install.packages("statTarget")
```

## **1.3 For Windows users**

1 - Copy this code into R

```
> source("https://bioconductor.org/biocLite.R")
```

```
> biocLite("impute")
```

```
> install.packages("statTarget")
```

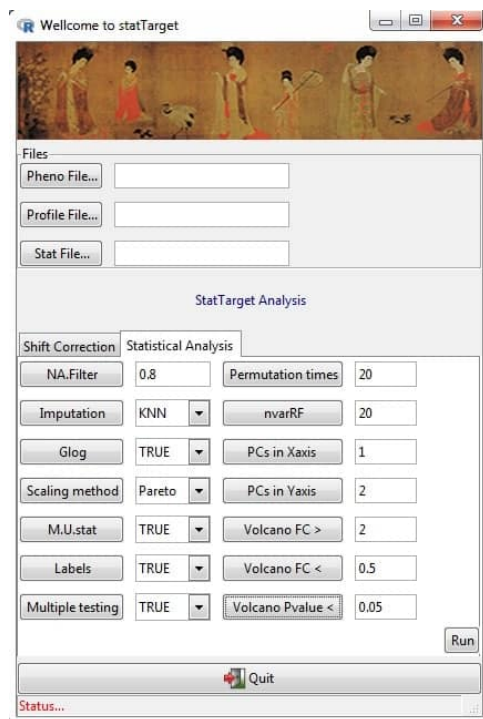


Figure 1: The main GUI of statTarget

## 2 Overview of statTarget

An easy to use tool provide a graphical user interface (Figure 1) for quality control based signal correction, integration of metabolomic data from multiple batches, and the comprehensive statistic analysis for non-targeted and targeted approaches. (URL: <https://github.com/13479776/statTarget>)

## 3 statTarget in detail

The main GUI of statTarget has two basic components. The first components is shift correction. It includes quality control-based robust LOESS signal cor-

rection (QC-RLSC) that is a widely accepted method for quality control based signal correction and integration of metabolomic data from multiple analytical batches (Dunn WB., et al. 2011; Luan H., et al. 2015). The second components is Statistical Analysis. It provides comprehensively computational and statistical methods that are commonly applied to analyze metabolomics data, and offers multiple results for biomarker discovery.

### **3.1 statTarget Shift Correction**

statTarget - Shift Correction provide QC-RLSC algorithm that fit the QC data, and each metabolites in the true sample will be normalized to the QC sample. To avoid overfitting of the observed data, LOESS based generalised cross-validation (GCV) would be automatically applied, when the QCspan was set at 0.

#### **3.1.1 Parameters**

1. Pheno File: Meta information includes the Sample name, class, batch and order. Do not change the name of each column. (a) Class: The QC should be labeled as NA. The samples' class should be matched with ordinal number, e.g., 1,2,3,4,5... (b) Order : Injection sequence. (c) Batch: The analysis blocks or batches,e.g., 1,2,3,4,5.... (d) Sample name should be consistent in Pheno file and Profile file. (See the example data)
2. Profile File: Expression data includes the sample name and expression data.(See the example data)
3. NA.Filter: Removing peaks with more than 80 percent of missing values

(NA or 0) in each group. (Default: 0.8)

4. QCspan: The smoothing parameter which controls the bias-variance tradeoff. The common range of QCspan value is from 0.2 to 0.75. If you choose a span that is too small then there will be a large variance. If the span is too large, a large bias will be produced. The default value of QCspan is set at '0', the generalised cross-validation will be performed for choosing a good value, avoiding overfitting the observed data. (Default: 0)

5. degree: Lets you specify local constant regression (i.e., the Nadaraya-Watson estimator, degree=0), local linear regression (degree=1), or local polynomial fits (degree=2). (Default: 2)

6. Imputation: The parameter for imputation method.(i.e., nearest neighbor averaging, "KNN"; minimum values for imputed variables, "min"; median values for imputed variables (Group dependent) "median". (Default: KNN)

### **3.2 statTarget Statistical Analysis**

statTarget - Statistical Analysis provide features including Data preprocessing, Data descriptions, Multivariate statistics analysis and Univariate analysis.

Data preprocessing : 80-precent rule, log transformation, KNN imputation, Median imputation and Minimum values imputation.

Data descriptions : Mean value, Median value, Sum, Quartile, Standard derivatives, etc.

Multivariate statistics analysis : PCA, PLSDA, VIP, Random forest.

Univariate analysis : Student T-test, Shapiro-Wilk normality test and Mann-

Whitney test.

Biomarkers analysis: ROC, Odd ratio.

### 3.2.1 Parameters

1. Stat File: Expression data includes the sample name, group, and expression data. The samples' group should be matched with ordinal number, e.g., 1,2,3,4,5... (See the example data)
2. NA.Filter: Removing peaks with more than 80 percent of missing values (NA or 0) in each group. (Default: 0.8)
3. Imputation: The parameter for imputation method.(i.e., nearest neighbor averaging, "KNN"; minimum values for imputed variables, "min"; median values for imputed variables (Group dependent) "median". (Default: KNN)
4. Glog: Generalised logarithm (glog) transformation for Variance stabilization (Default: TRUE)
5. Scaling Method: Scaling method before statistic analysis (PCA or PLS). Pareto can be used for specifying the Pareto scaling. Auto can be used for specifying the Auto scaling (or unit variance scaling). Vast can be used for specifying the vast scaling. Range can be used for specifying the Range scaling. (Default: Pareto)
6. M.U.Stat: Multiple statistical analysis and univariate analysis (Default: TRUE)
7. Permutation times: The number of random permutation times for PLS-DA model (Default: 20)

8. PCs in the Xaxis or Yaxis: Principal components in PCA-PLS model for the x or y-axis (Default: 1 and 2)
9. nvarRF: The number of variables in Gini plot of Randomforest model ( $\leq 100$ ). (Default: 20)
10. Labels: To show the name of sample in the Score plot. (Default: TRUE)
11. Multiple testing: This multiple testing correction via false discovery rate (FDR) estimation with Benjamini-Hochberg method. The false discovery rate for conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. (Default: TRUE)
12. Volcano FC: The up or down -regulated metabolites using Fold Changes cut off values in the Volcano plot. (Default:  $> 2$  or  $< 1.5$ )
13. Volcano Pvalue: The significance level for metabolites in the Volcano plot. (Default: 0.05)

## 4 References

- Dunn, W.B., et al., Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols* 2011, 6, 1060.
- Luan H., LC-MS-Based Urinary Metabolite Signatures in Idiopathic Parkinson's Disease. *J Proteome Res.*, 2015, 14, 467.
- Luan H., Non-targeted metabolomics and lipidomics LC-MS data from maternal plasma of 180 healthy pregnant women. *GigaScience* 2015 4:16



## 5 Session Info

- R version 3.3.0 (2016-05-03), x86\_64-apple-darwin13.4.0
- Locale:  
`en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: devtools 1.11.1, digest 0.6.9, gWidgets2 1.0-7,  
gWidgets2RGtk2 1.0-5, memoise 1.0.0, RGtk2 2.20.31, statTarget 1.2.4
- Loaded via a namespace (and not attached): BiocInstaller 1.22.3,  
cluster 2.0.4, DEoptimR 1.0-4, grid 3.3.0, impute 1.46.0, lattice 0.20-33,  
magrittr 1.5, mvtnorm 1.0-5, pcaPP 1.9-60, pdist 1.2, pls 2.5-0,  
plyr 1.8.3, pROC 1.8, randomForest 4.6-12, Rcpp 0.12.4,  
robustbase 0.92-6, roxygen2 5.0.1, rrcov 1.3-11, rstudioapi 0.5,  
stats4 3.3.0, stringi 1.0-1, stringr 1.0.0, tools 3.3.0, withr 1.0.1