NAZARBAEV UNIVERSITY, INSTITUT CURIE

BIODICA

computational pipeline for Independent Component Analysis of BIg Omics Data

General documentation and user guide for the BIODICA pipeline, ver 0.9

Andrei Zinovyev , Ulykbek Kairov

10/31/2016

This documents contains the general description of the BIODICA pipeline. Terminology used us defined, description of the system's modules is provided.

Table of Contents

INTRODUCTION

Types of omics data and the problem of their efficient analysis

Large-scale projects (e.g. Cancer Genome Project, The Cancer Genome Atlas (TCGA), or the International Cancer Genome Consortium), and the efforts of individual laboratories, are generating massive amounts of high-throughput molecular data often associated with clinicopathological characteristics. Transcriptome data for tumors are the most commonly available type of large-scale molecular data, but recently molecular profiles of mutations, DNA methylation, miRNAome, proteome, phosphoproteome have been started to be available at large scale for analysis. These data remain difficult to interpret because the molecular profiles are influenced by various overlapping biological factors linked to the tumor cells or to the tumor microenvironment and non-biological factors linked to sample processing or data generation. Widely used basic statistical techniques such as hierarchical clustering or principal component analysis do not show satisfactory reproducibility from one data set to another one.

What is BIODICA?

BIODICA is a computational pipeline implemented in Java language for

(1) automating deconvolution of large omics datasets with optimization of deconvolution parameters,

(2) helping in interpretation of the results of deconvolution application by automated annotation of the compoentns using the best practices,

(3) comparing the results of deconvolution of independent datasets for distinguishing reproducible signals, universal and specific for a particular disease/data type or subtype.

Task of blind source deconvolution in application to omics data

The need to deconvolute diverse factor affecting gene expression led to the use of methods originally developed to solve the blind source separation problem (Jutten and Hérault, 1991), with the aim of recovering hidden signal sources from the observed output mixture. Independent component analysis (ICA) is one of these methods. ICA models the level of expression of each gene in a given sample as a linear weighted sum of several independent components, where each component captures the effect of one of the factors/processes. The expression data matrix is thus decomposed into a number of components, each of which is characterized by an activation pattern both across genes and across samples. The genes with the largest projection onto a component (providing the greatest contribution) are the genes the most strongly influenced by the process associated with this component. The contribution value of a sample reflects the activity of the component in this sample.

First report on application of ICA to deconvolution of signals in analysis of gene expression changes during yeast sporulation appeared in (Hori et al, 2001). In this work, ICA components, called “modes of gene expression”, reproduced manually defined biological classes of yeast genes. Since then the number of applications of matrix factorization methods in analysis of gene expression has grown very rapidly. There exist several reviews, comparing different methods and pointing out at their advantages and disadvantages [Kong et al, 2008; Gorban et al, 2008].

Brief description of the Independent Component Analysis method

In many fields of science, data can be represented as a rectangular matrix with some objects (for example, n genes) corresponding to the matrix rows and the objects’ features (for example, gene expression in m tumor biopsies) corresponding to the matrix columns. These matrices can be huge: thus, methods for revealing patterns in the distribution of the matrix element values are of extreme use. One particular approach is connected to the idea of approximating a rectangular matrix by another matrix, having much lower rank: $X \approx AS$, where X is a matrix of data of size $m \times n$, and A is a $m \times k$ matrix, $k << m$. We will call the rows of the A matrix components (m-dimensional vectors), and the columns of the S matrix projections of data vectors onto the components (a k-dimensional vector for each of n data points).

A gene (or a protein, or a DNA methylation site) can be considered as a sensor which receives regulatory signals from several sources (biological factors), see Figure 1. The biological factors can be activities of transcription factors or other various influences coming from a particular

intercellular context or from environment. The combination of factor activities regulates gene expression through a complex (and unknown) function. As the first approximation, we can assume that this function is linear:

$$\text{Expression(gene i, sample s)} = \Sigma i=1..m\ a_{Fj}^{\text{gene i}} \text{Activity}_{Fj}\text{(sample s)}.$$

A simple optimization problem $|| X - AS ||2 \rightarrow$ min with Euclidean metrics leads to the well-known Singular Value Decomposition (SVD) or, equivalently, Principal Component Analysis (PCA), two fundamental methods introduced in the data analysis in the very beginning of 20^{th} century. These methods work best in the case of multidimensional Gaussian data distribution.

By contrast, in ICA it is suggested using higher-order moments for matrix approximation, considering all Gaussian signals as noise. It was followed by other similar works and was rigorously and theoretically described in (Comon, 1994). Efficient and fast algorithms were developed for ICA (Hyvärinen et al, 2001). It was shown that the requirement of statistical independence is equivalent to maximizing non-gaussivity of data point projections onto the components, measured by some combination of higher data distribution moments (kurtosis) or other functions (negentropy, tangent function, etc.) The Gaussian signal contained in the data is usually subtracted from the data before application of ICA by data whitening such that all second moments become equal unity. Therefore, independent components can be non-orthogonal in the original space of data, which can be considered as an advantage in applications.

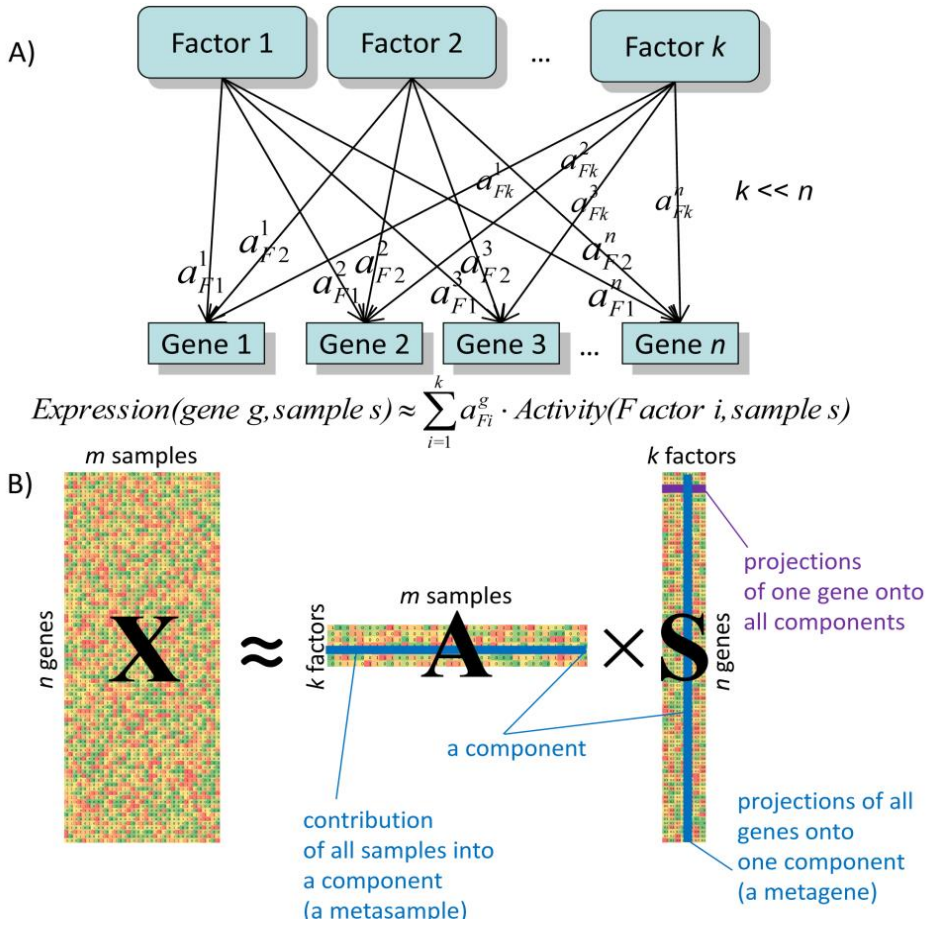

Figure 1. General principles of ICA application to gene expression data. A) Network interpretation of matrix factorization methods. In the "space of genes", each gene expression value is approximated by a weighted linear combination of few factor activities (which can be transcription factors, environmental factors, etc.). Biological samples are characterized by different factor activities, but the weights (network structure) do not depend on samples. In PCA aij implement maximum variance; in ICA aij have maximally non-gaussian distributions for each factor. B) A matrix of data X is approximated as a product of low-rank matrices A (mixing matrix) and S (score matrix). Each component corresponds to a row in A (meta-sample) and a column in S (metagene) and introduces two sets of weights for all genes and all samples. Reproduced from (Zinovyev et al, 2013).

Short term vocabulary used further in this text

Gene space and sample space

Decomposition of a gene expression data table into linear components (i.e., into the sum of matrices having rank one) can be done in practice in two ways. Being mathematically equivalent, they are distinct in implementation and interpretation, so it is important to

distinguish them. In the further we refer to the "gene space" such a space where each data point represents a gene (or a protein or a methylation site) and the coordinates of the space correspond to different samples. Oppositely, in the sample space, each point is a biological sample and each coordinate axis corresponds to a gene. Note that in many occasions, the definitions of gene and sample spaces might be opposite to the above defined, so precising the definition is always necessary.

In the further analysis we assume the analysis done in the gene space. In this case, it has the interpretation suggested above (a gene is a sensor recording signal mixtures). Application of ICA in the sample space is also meaningfull and corresponds to the application of blind source separation methods to the transposed matrix of gene expression (where rows represent samples).

Metagene

Metagene is any set of numerical weights, positive and negative, associated to (many) genes, even all genes in a genome. Each independent component is associated to one metagene, where the weights are projections of the gene onto the component.

Set of most contributing genes

In the further the set of most contributing genes refers to the gene set which have the largest (by absolute value) projections onto the component. If we represent an independent component as a metagene, then this a set of genes having the largest by absolute value weights in the metagene. The threshold for deciding the set can be adapted to a question and a task but typically it corresponds to 3 standard deviations from the mean value, which typically collects 1-2% of the total number of genes in the metagene.

Metasample

Metasample is a set of numerical weights associated to samples participating in the analysis. Each independent component is associated to a metasample, where the weights are contributions of all samples into the direction of the independent component.

Bimodality of metasample

If the weights of a metasample corresponding to an independent component forms a clear bimodal (or k-modal) distribution, then we call this component "bimodal". Bimodal components are more interesting for interpretation because they naturally define subtyping of clinical or biological samples into several groups. Example of a bimodal metasample is the signal related to the gender of a patient.

Working folder

A folder on the user's hard drive where the results of BIODICA application are stored and analysed. The folder have a specific structure described below.

Data repository

A set of folders necessary for BIODICA to work. These folders contain MATLAB executable and various databases (such as PPI databases or databases of predefined gene sets) necessary for interpreting the independent components.

A default data repository is provided together with BIODICA executable jar file, but its content and structure can be re-configured accordingly to the user's preferences using the BIODICA configuration file.

The data repository can also contain the actual data to be analyzed. During the BIODICA analysis, local copies of the data files are created in the working folder.

BIODICA SYSTEM'S ARCHITECTURE (FULL VERSION PROJECT)

Modular structure of BIODICA pipeline

BIODICA pipeline is composed of 8 main modules:

1. Big Data Management module
2. Prior Knowledge and Biological Networks module
3. Intense ICA Computation module
4. Metasample Annotation module
5. Metagene Annotation module
6. Meta-Analysis module
7. Visualization module
8. Reporting module

Each module is characterized by the required input, nature and format of files, and the output files produced which are further used in other modules.

Organization of data repository for BIODICA

The structure of BIODICA data repository is not fixed, but the recommended configuration, provided with BIODICA installation package is described below.

The default structure of the data repository is

/bin	(contains binaries assuring the work of BIODICA)
/BIODICA	(BIODICA jar file)
/fastica++	(MATLAB FastICA+ICASSO implementations adapted for using in BIODICA)
/GSEA	(GSEA binaries)
/HTML	(html files necessary for reporting module)
/others	(other usefull but not required binaries: VidaExpert, …)
/data	(data for analysis by BIODICA)
/[dataset_name]	(eg, "OVCA_TCGA")
/copynumber	(copy number profiles)
/methylome	(DNA methylation profiles)
/mutation	(mutation binary or weigthed binary tables)
/sample_info	(sample annotation files)
/transcriptome	(transcriptome tables)
/doc	(documentation files including this manual)
/knowledge	(pre-existing knowledge)
/geneproperties	(table describing some gene properties such as genomic location or CG-content)
/genesets	(pre-defined gene sets in gmt format – might be several, will be fused)
/metagenes	(pre-existing metagenes in tab-delimited or rnk formats)
/networks	(network data)
/directed	(directed networks, eg. regulatory networks)
/undirected	(undirected networks, eg. PPI networks)
/work	(working folder for BIODICA)
config	(BIODICA configuration file, located in the root folder)

Recommendations for preparing the data matrix files for using in BIODICA

For application of ICA, it is recommended to prepare the omics data matrix files accordingly to the following rules:

1) Columns are sample names, rows are genes or proteins or methylation sites, or other measured signals. First line always lists sample names, first column always contains ids of the objects (genes, proteins), the first column should have name as well.
2) Genes uses HUGOs for names
3) Each line is recommended to be centered (average equals zero) for transcriptome analysis

For automated interpretation, it is recommended to prepare the sample annotation matrix files accordingly to the following rules:

4) Columns are sample features (patient age, molecular subtype, status, etc.), rows are sample names. First line always lists feature names, first column always contains ids of the samples, the first column should have name as well.

5) IDs of the samples in the sample annotation file should exactly match the sample names in the omics data files (column names): this is important to check in advance!

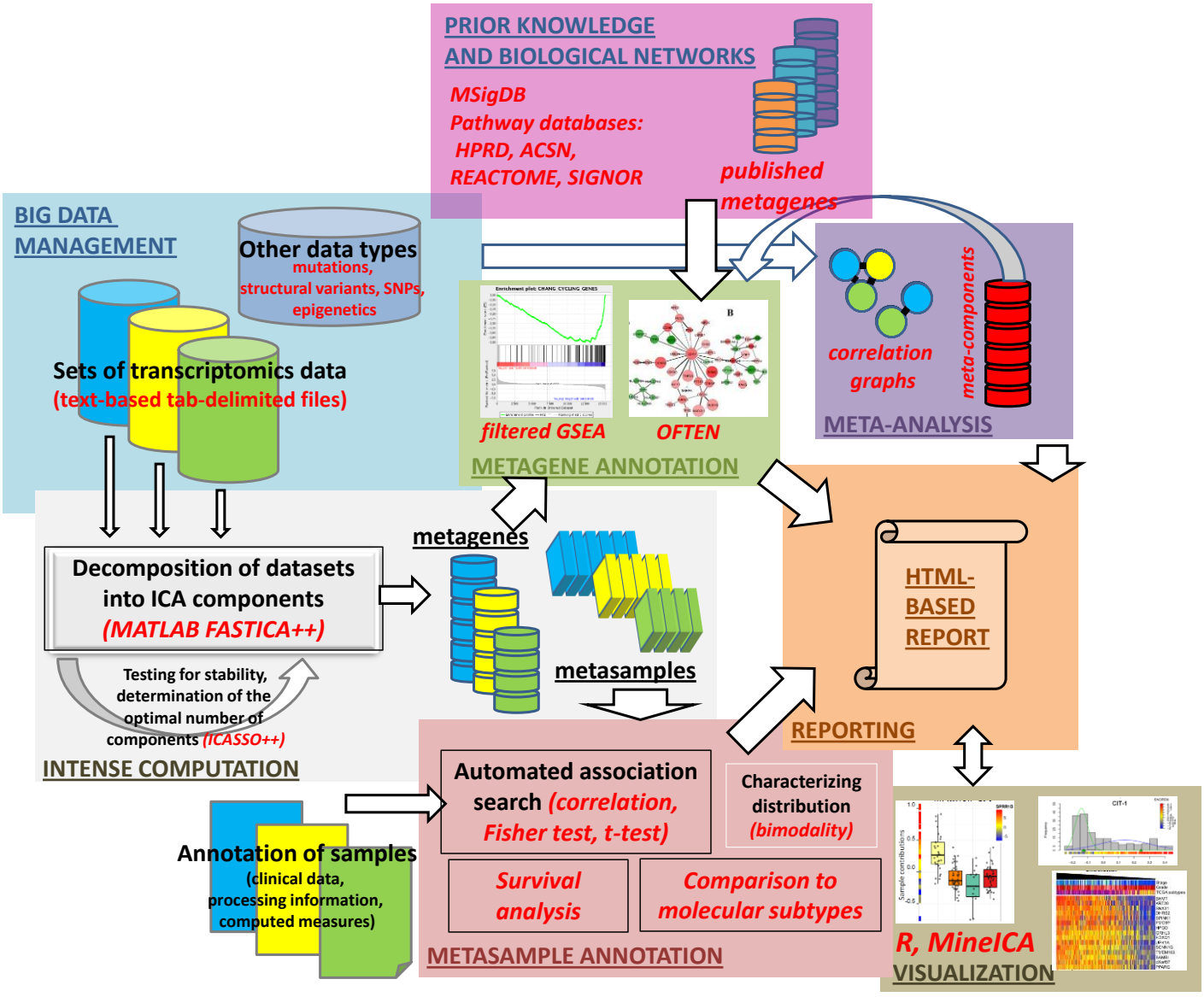

Figure 2. General architecture of the BIODICA data analysis pipeline.
Boxes of different colors separates different functional modules of the system. Described functionality corresponds to the BIODICA version 1.0.

BIODICA MODULES DESCRIPTION

Big Data Management module

This module contains the omics data themselves, clinical annotation or other sample annotation data. BIODICA contains several simple procedures for formatting of the omics data tables (such as centering, converting to log scale, extracting the numerical part, etc.) necessary for application of ICA.

Prior Knowledge and Biological Networks module

This module contains the part of data repository containing the prior knowledge and network information. BIODICA contains several simple procedures for analyzing the networks, comparing and merging them.

Intense ICA Computation module

The module contains several parts:

(1) Computation of Independent Components for a specified number of independent components using MATLAB implementation of FastICA (adapted for using in BIODICA)
(2) Estimating the stability of independent components using boot-strapping: the independent deconvolution is computed n=100 times, and then clustered. The centroid components are considered the correct ones
(3) Optimization of the number of independent component to compute based on their stability profile
(4) The results of computation are stored in the folder named <dataset_name>_ICA in the working folder.

Metasample Annotation module

Metasamples are automatically annotated using the following approaches:

1) <u>Testing the sample annotation file for putative associations with independent components</u>
    a. A folder <dataset_name>_MSAMPLE is created.
    b. File containing metasample description <dataset_name>_A.xls is copied from the ICA folder to MSAMPLE folder and merged with the sample annotation file.
    c. In the sample annotation file, all sample features are classified into numerical and categorical. The feature is considered categorical is it contains non-numerical symbols as values, or the number of distinct numerical values is

less than *minNumberOfDistinctValuesInNumericals* BIODICA parameter. All missing values (marked by "NA" or "_") are ignored.

d. If the sample feature is truly numerical then the association is computed by calculating the p-value of the Spearman correlation coefficient.
e. If the sample feature is categorical, and the number of categories is less than *MaxNumberOfCategories* parameter, then the association is computed by calculating the p-value of the Wilcoxon test between all pairs of categories and metasample weights. If in the pair, one of the categories is represented by less than *minNumberOfSamplesInCategory* samples then the test is not computed. The minimum p-value between all pair-wise comparisons is assigned to the association test between a sample feature and an independent component represented as a metasample.
f. The –log10(p-value) values are reported in the association table. Those sample features which are not associated to any independent component are not reported in the association table. A separate table is prepared with notes on each significant association, indicating which comparison gave a significant association.
g. The results are stored in <dataset_name>_A_associations.xls file in the working folder.

2) The above analysis includes association study to mutation data which is recommended to include, as binary categorical features into the sample description file.
3) If sample description contains classification of samples into molecular subtypes then a more detailed statistical testing and more specific report is produced to prove association of all metasamples with known molecular subtypes.
4) Testing Metasample for bi-modality. Several common tests are applied including computing curtosis of the metasample weights and Dip Test.
5) If survival information is provided for the set of samples, then survival analysis is done by estimating the Cox survival regression model and testing its p-value.

Metagene Annotation module

Metagenes are automatically annotated using the following approaches:

1) GSEA Pre-ranked analysis and filtering its results
   a. First, **geneset** folder in the data repository is checked for all gmt files located in it. All gmt files found are merged into one for the further GSEA analysis. If a user does not want to use some gmt files in the data repository, their extensions should be renamed (eg, to *.notuse*).

b. Each component is converted into a separate metagene and stored into a file with rnk extension, where the genes are ordered accordingly to the metagene weights.
c. For each rnk file, the GSEA process is automatically launched, the results are stored in the working folder, where new folder <dataset>_GSEA is created for this purpose.
d. The results of GSEA application are filtered for those enrichments which have at least *k* most contributing genes in the leading edge set, and which corrected for multiple testing p-value is less than p0. By default, k=5 and p0=0.01.
e. A summary html file is created named "results/results_GSEA_filtered.html". The file contains, per each independent component, a list of most frequently found genes, and links to the trustably enriched gene sets with indication of how many most contributing genes were found in the leading edge set.

2) Correlating metagenes to certain gene properties (numerical or categorical) in the way identical to the one described in the MetaSample Annotation procedure (see above).
3) Automated association of optimally functionally enriched subnetworks (OFTEN) using PPI networks from **knowledge/undirected** folder, using methodology described in (Kairov et el, 2013).
4) Studying the general properties of the pattern of metagene projection on top of the annotated human genome
5) Correlating metagenes to reference metagenes
6) Correlating metagenes to other molecular profiles (DNA methylation matched to gene promoters, gene copy number variations, etc.)
7) Correlating ICA metagenes to the metagenes computed for biological pathways through application of ROMA software (https://github.com/sysbio-curie/Roma).

Meta-Analysis module

Metaanalysis module allows comparing all metagenes computed in the working folder and calculating the recapitulating correlation graph, representing reciprocal and non-reciprocal correlation relations between components calculated for different datasets.

Data visualization module

Visualizing independent components is performed by

1) Using specific tools used to annotate metagenes and metasamples (ie, GSEA plots or survival plots).
2) Using data visualization tools developed in MineICA R package available in BioConductor
3) Using Cytoscape for visualization of graphs produced by BIODICA pipeline (eg, correlation graphs after metanalysis)

4) Using VidaExpert software used for visualization of *.dat* files produced by BIODICA pipeline (Windows only)

Reporting module

Reporting module analyses the content of working directory and creates a set of html pages for representing it. It reflects all individual dataset analyses performed as well as the results of meta-analysis application. Reporting module represents tables as sortable and interactive dynamic html representations.

EXPLOITING BIODICA IN COMMAND LINE MODE

Systems requirements:

1) Windows or Linux operating system
2) Installed Java ver 1.6 or higher
3) Installed MATLAB ver 8 or higher (no additional toolboxes are required). The matlab executable should be specified in the system's path.
4) At least 8Gb of operating memory

It is recommended to launch BIODICA with maximum amount of available memory (eg, using "-Xmx6000M" option for JVM), in order to avoid "Out of Java heap space" error message.

Creating a configuration file

The configuration file contains various options adjusting the functioning of BIODICA or changing it's default behaviour. It is recommended to keep the configuration file in the root folder of data repository.

The list of required options (paths for configuration of the BIODICA data repository):

MATLABICAFolder = [path_to_repository]\bin\fastica++

DefaultWorkFolder = [path_to_work_folder]

GeneSetFolder = [path_to_repository]\knowledge\genesets

HTMLSourceFolder = [path_to_repository]\bin\HTML

List of optional parameters changing the default BIODICA behavior (below the default values are provided)

ComputeRobustStatistics = false

MinNumberOfDistinctValuesInNumericals = 10

MinNumberOfSamplesInCategory = 3

MaxNumberOfCategories = 10

MinimumTolerableStability = 0.8

BIODICA Command line use

For listing all available options of BODICA, it should be launched without options:

java -jar BODICA.jar

For analysis of a single dataset it is necessary to provide the path to the configuration file, the data file and the output folder (by default, the working folder will be used), and list required analyses. For example,

java -jar BODICA.jar -config C:\Datas\BIODICA\config -outputfolder C:\Datas\BIODICA\work\ -datatable C:\Datas\BIODICA\data\OVCA_TCGA\transcriptome\OVCA.txt -doicamatlab 20 -dogsea 100 -dometasampleanalysis C:\Datas\BIODICA\data\ OVCA_TCGA\sample_info\OVCA.txt

Complete list of options and actions

Required options:

-config <file_name>	path to BIODICA config file
-datatable <filename>	this option is required if the analysis is done on a single datatable (not in the batch mode)

Possible options:

-outputfolder <folder_name>	changing the default work folder specified in the configuration file

Actions:

-doicamatlab <numberOfComponents>	calculate specified number of ICA components using fastica and icasso implemented in Matlab with default parameters
-dogsea <numberOfPermutations>	make GSEA analysis for all computed metagenes, using <numberOfPermutations> for assessing the p-values for the enrichment

-dometasampleanalysis <sampleAnnotationFile>	make automated analysis for associations between computed ICs and the sample annotations
-dometageneanalysis <sampleAnnotationFile>	make automated analysis for associations between computed ICs and the gene annotations

-dooften <PPI_network.xgmml> [#nstart,nstep,nend,nperm]

make automated analysis of ICA metagenes to associate them to a subnetwork in a global PPI network

-doreport <folderToCreateReport>	produce HTML report for all analyses made in the working folder in the specified folder <folderToCreateReport>
-dobbhgraph <folderWithSfiles>[#split]	compile BBH graph from a set of files ending with "_S.xls", containing the metagenes corresponding to ICs. The suffix '#split' can be added to the folder name to force decomposing each S_xls file into positive and negative distributions of metagene weights.

CASE STUDY OF USING BIODICA FOR ANALYSIS OF SEVERAL PUBLICLY AVAILABLE LARGE TRANSCRIPTOMICS DATASET

Example 1. Using BIODICA for deconvoluting a large transcriptomic dataset of TCGA ovarian cancer patient's cohort.

In this example, the deconvolution is performed on a transcriptomic table containing 413 ovarian cancer samples and 20806 genes whose expression is measured using RNA-Seq technology. The data for this example are provided in the BIODICA distribution package (/data/OVCA_TCGA/transcriptome/OVCA.txt).

Step 1. Deconvolution of the data table into pre-defined number of components (20)

Launch **_example1_OVCA_doica.bat** file which executes the following command line:

java -Xmx5000M -jar BODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -doicamatlab 20

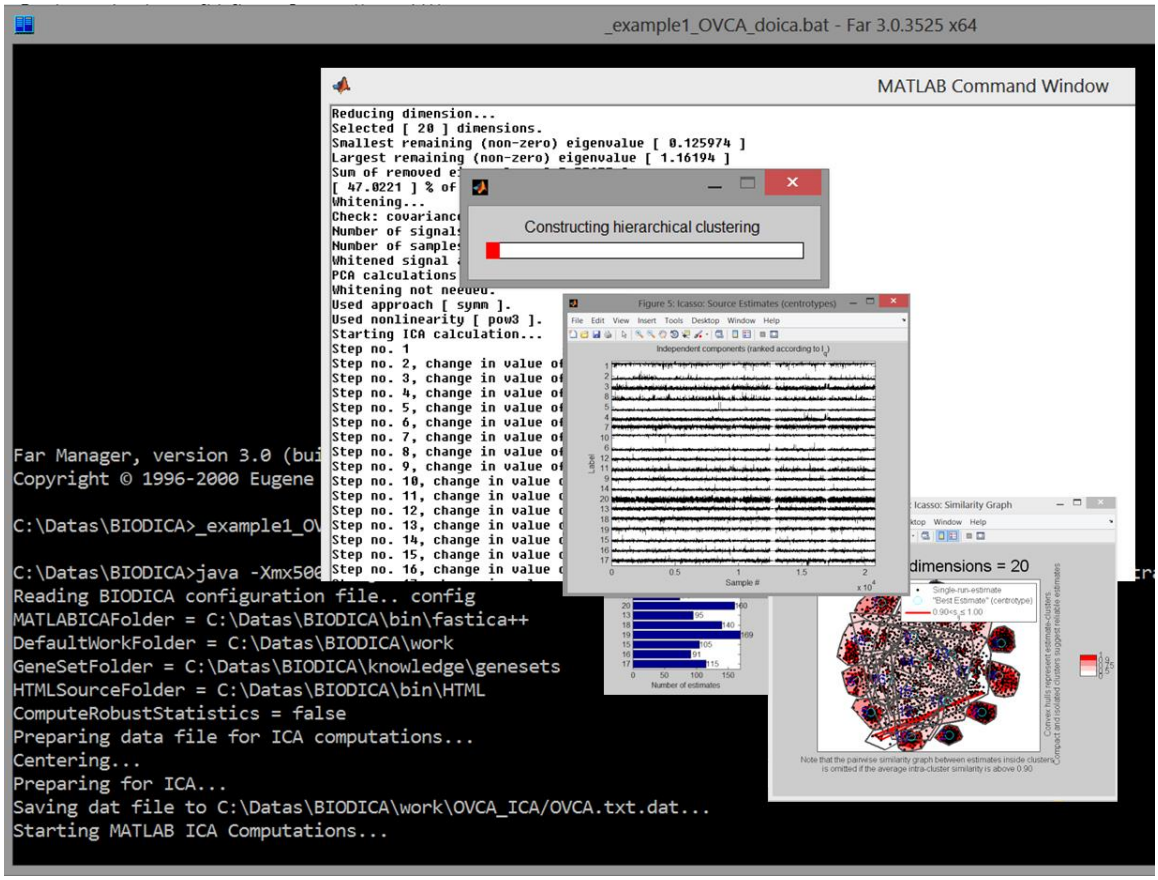

Figure 3. Screenshot of the computation of independent components in BIODICA.

Time of execution for this step on a standard laptop is few minutes depending on the computer processor type.

Step 2. Automated GSEA analysis on computed metagenes

Launch **_example1_2_OVCA_dogsea.bat** file which executes the following command line:

java -Xmx5000M -jar BODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -dogsea 100

This command merges all gmt files found in genesets folder and saves them into one gmt file **total.gmt**.

The file containing computed metagenes OVCA_ica_S.xls is decomposed into 20 rnk files containing rankings of all genes accordingly to the weight in the corresponding metagene. For each rnk file, a new GSEA analysis starts, applying 100 permutations to estimate the statistical significance of the reference gene sets.

After all analyses are completed, filtering procedure takes place. The results of filtering are stored in **/work/OVCA_GSEA/results/results_GSEA_filtered.html** file (see Figure 4).

The results of GSEA application allows associate several components to biological processes. For example, IC1 is associated to interaction between tumoral cells and extracellular matrix. IC4 and IC5 can be associated to two independent aspects of immune cell infiltration into the tumoral microenvironment. IC6 and IC20 are associated to the respiratory electron transport and translation at the same time (this phenomenon was observed already in (Biton et al, 2014)). IC7 is clearly associated to cell cycle. Interpretation of other components using GSEA is less conclusive.

At the same time IC2, IC9, IC10, IC12 and IC19 do not obtain any convincing GSEA enrichments.

Figure 4. Output of the application of GSEA analysis and filtering for OVCA dataset analysis.

This step on a standard laptop also takes only few minutes.

Step 3. Automated association search on computed metasamples

BIODICA can perform automated search for significant associations between categorical and numerical features of biological samples and the metasamples computed by ICA. In the test example this can be achieved by launching **_example1_3_OVCA_dosample.bat** file containing the following command line:

java -Xmx5000M -jar ./bin/BIODICA/BIODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -dometasampleanalysis ./data/OVCA_TCGA/sample_info/OVCA.txt

As a result of application of this procedure the folder **OVCA_MSAMPLE** will be created, and several files will be generated:

work/OVCA_MSAMPLE/OVCA_A_associations.xls

which contains the significant p-values (presented as *–log10(p-value)*) and

work/OVCA_MSAMPLE/OVCA_A_associations_info.xls

The threshold of significant *–log10(p-value)* value is defined by the **AssociationAnalysisThreshold** parameter in the *config* file.

The procedure automatically separates the sample features into categorical and numerical, accordingly to the following definition: (1) *numerical feature* values are only numericals (after removing missing data labels such as 'N/A') such that the number of distinct values is more or equal than **MinNumberOfDistinctValuesInNumericals** (parameter in the *config* file); (2) *categorical feature* contains either numericals but with the number of distinct values less than **MinNumberOfDistinctValuesInNumericals** or any text labels but with the number of distinct labels less than **MaxNumberOfCategories** (parameter in *config*). Other features (non-numerical and non-categorical) are not tested in the association study.

Currently, testing the association of a metagene with a categorical sample variable is performed by a simple t-test for all pairs of categorical labels which are represented by at least **MinNumberOfSamplesInCategory** samples. The pair of labels with the most significant p-value is reported in the resulting file (see Figure 5) for further more carefull testing.

Association between a numerical sample variable and a metagene is computed using Spearman correlation.

Both in case of categorical and numerical values a p-value for the association is computed. Only those sample variables are reported which have at least one significant association (see Figure 5).

Suspected association should be further validated using the files **OVCA_a_annot.xls** (can be open in any text editor or Excel) and **OVCA_a_annot.dat** (can be open in ViDaExpert software). In the example of OVCA dataset, the most significant associations are observed for the previous results of expression or methylation profile clustering using consensus NMF clustering by Broad data analysis pipeline (Figure 6). Only weak association has been found with one mutation (FBXW7, SCF ubiquitin protein ligase complex component) and IC6 component. Weak association has been found between IC5 and the patient age. Interestingly, moderate associtation has been found between IC6 and patien.tissue_source_site which indicates to a batch effect (especially, difference between samples collected at MD Anderson/Memorial Sloan Kettering and Roswell Park/University of Pittsburg genomic centers).

VAL	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9	IC10	IC11	IC12
patient.initial_pathologic_diagnosis_metho	_	_	_	_	_	_	_	_	_	_	_	_
patient.performance_status_scale_timing	_	_	_	_	_	pre-adjuvant therapy/other(5.7)	_	_	_	_	_	_
mRNA_cNMF	1/3(-24.6)	1/3(5.3)	_	1/2(-16.1)	3/2(11)	_	3/2(-3.5)	_	_	_	1/2(-6.9)	1/2(-6.1)
mRNA_cHierarchical	1/2(-21.9)	2/3(-4.6)	_	2/3(15.5)	1/3(-8.4)	2/3(-3.8)	_	_	1/3(4.6)	_	2/3(7.5)	1/3(3.6)
miR_cNMF	2/3(6.2)	_	_	1/3(3.4)	2/3(3.9)	_	_	1/2(4)	_	1/2(-3.7)	1/3(3.6)	1/2(3.5)
CN_cNMF	1/3(3.9)	_	_	_	_	_	1/2(-5.5)	_	_	1/2(-3.5)	1/3(5.3)	1/2(-4.3)
Methlyation_cNMF	1/2(-6.2)	_	_	1/3(-8.9)	2/3(4.1)	_	1/2(4.1)	_	_	_	1/2(-8.1)	2/3(-4.3)
RPPA_cNMF	3/2(7.7)	_	3/2(-4.1)	3/2(3.4)	1/2(3.9)	_	_	_	_	_	3/2(4.4)	_
RPPA_cHierarchical	1/2(10.9)	_	_	1/3(5.2)	1/2(3.6)	_	1/3(-4.4)	_	_	_	1/2(4.6)	_
mRNAseq_cNMF	1/2(10.3)	1/2(-6)	_	1/3(10.3)	2/3(-4.2)	_	_	1/3(3.4)	_	_	1/3(8)	1/3(6.4)
mRNAseq_cHierarchical	1/2(10.6)	1/2(-5.5)	_	1/3(10.1)	2/3(-5.8)	_	_	_	_	_	1/3(6.4)	1/3(6)
miRseq_cNMF	1/3(16)	2/3(-5.8)	_	1/3(6.4)	1/2(5.7)	_	1/2(-3.7)	_	_	_	_	1/2(-3.9)
miRseq_cHierarchical	3/4(9.2)	3/4(-5.6)	_	1/2(4.8)	_	_	_	_	_	_	1/2(4.5)	_
FBXW7	_	_	_	_	_	no/yes(4.8)	_	_	_	_	_	_
BRCA12_class	_	_	_	_	_	_	_	_	_	_	_	_
Event	_	_	_	_	_	_	_	_	_	_	_	_
patient.age_at_initial_pathologic_diagnosis	_	_	_	_	0.21	_	_	_	_	_	_	_
patient.days_to_birth	_	_	_	_	-0.22	_	_	_	_	_	_	_
patient.tissue_source_site	_	_	_	0.22	_	-0.38	_	_	_	_	_	_
miR_cHierarchical	_	0.18	_	-0.16	0.17	_	_	_	_	_	-0.17	_

VAL	IC13	IC14	IC15	IC16	IC17	IC18	IC19	IC20
patient.initial_pathologic_diagnosis_metho	_	_	_	_	tumor resection/fine needle aspi	_	_	_
patient.performance_status_scale_timing	_	_	_	_	_	_	_	_
mRNA_cNMF	_	_	_	1/2(11.2)	_	_	_	_
mRNA_cHierarchical	1/3(4.2)	_	_	2/3(-9.4)	_	_	2/3(-3.5)	_
miR_cNMF	_	_	_	1/3(-3.4)	2/3(6.1)	_	_	2/3(-3.5)
CN_cNMF	_	2/3(6.6)	1/2(4.9)	2/3(-9.2)	_	_	1/2(-3.8)	_
Methlyation_cNMF	_	_	2/3(3.6)	1/3(10.2)	_	_	_	_
RPPA_cNMF	_	_	_	_	_	_	_	_
RPPA_cHierarchical	_	_	_	_	_	_	_	_
mRNAseq_cNMF	1/3(5.1)	2/3(3.7)	_	1/3(-13.2)	1/2(-5.6)	_	1/3(-3.6)	_
mRNAseq_cHierarchical	1/3(4.3)	_	_	1/3(-10.2)	1/2(-5.1)	_	1/2(-4.6)	_
miRseq_cNMF	2/3(3.4)	_	_	_	2/3(-8.7)	_	1/3(-3.4)	_
miRseq_cHierarchical	_	1/3(6.2)	_	1/2(-4.5)	3/4(-10.6)	_	_	_
FBXW7	_	_	_	_	_	_	no/yes(8.1)	_
BRCA12_class	_	_	_	_	5/1(4.2)	_	_	_
Event	_	_	Metastasis/Progression(4.4)	_	_	_	Alive_DiseaseFree/Progression(-4.	_
patient.age_at_initial_pathologic_diagnosis	_	_	_	0.27	-0.17	_	_	_
patient.days_to_birth	_	_	_	-0.28	0.17	_	_	_
patient.tissue_source_site	_	-0.24	_	_	_	_	_	_
miR_cHierarchical	_	_	_	_	_	_	_	_

Figure 5. Sample annotation association table with detailed information. Only significant associations (with p-value<0.001) are reported. Above the gray line association with categorical sample features (evaluated by t-test) is given. The value in parentheses is the t-test value, before parentheses the most significant comparison of features is labeled (eg., Metastasis/Progression). Below grey line are correlation coefficients of IC metagenes with numerical sample features (such as patient age).

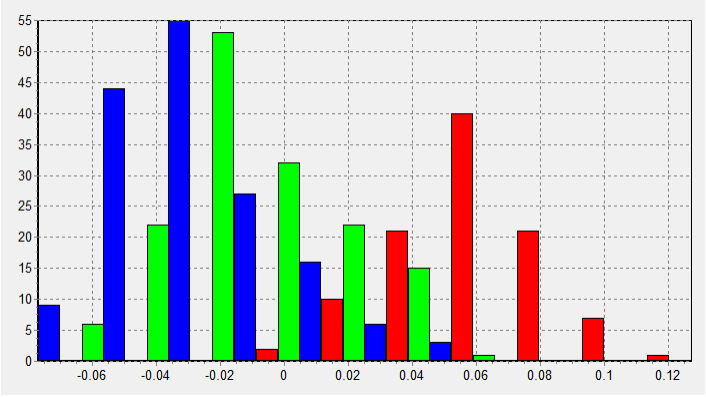

Figure 6. Association between previously computed mRNA cNMF clusters for ovarian cancer (#1 blue, #2 green, #3 red) in TCGA and the IC1 (see Figure 5).

Step 4. Automated association search on computed metagenes

Automated association search between gene properties (such as GC-content, size, type of transcript, etc.) is done using the same procedure and output as in the previous Step description. The corresponding command for this is in the **_example1_4_OVCA_dometagene.bat** file:

java -Xmx5000M -jar ./bin/BIODICA/BIODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -dometageneanalysis knowledge/geneproperties/genes.txt

In case of OVCA example, we find that IC14 is strongly associated with GC-content, and many components are slightly and strongly associated with the presence of snoRNAs transcripts in the transcriptome (Figure 7): this association is especially strong for IC3. This observation justifies excluding snoRNAs and some lincRNAs from the analysis.

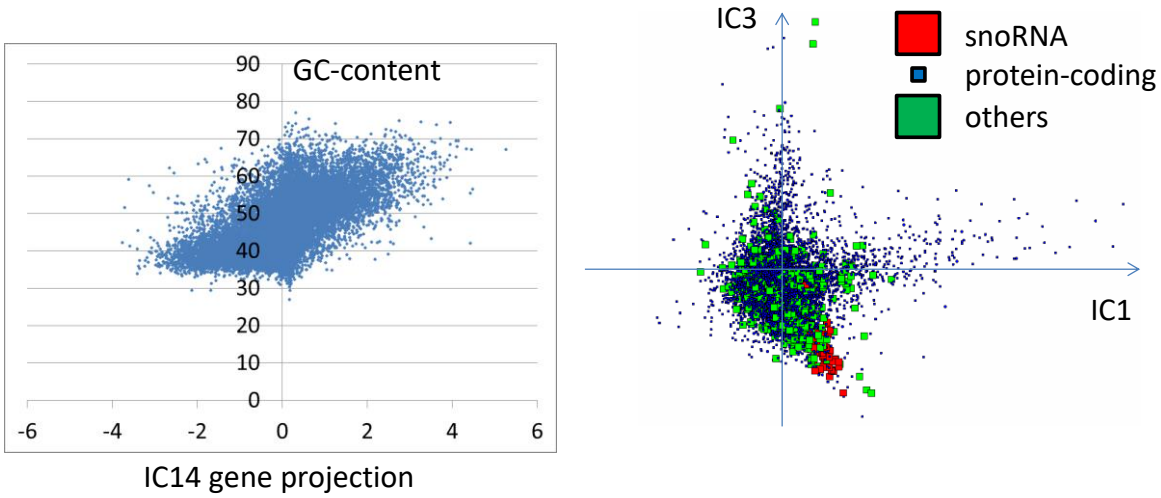

Figure 7. Association of gene properties with independent components. Left: IC14 is associated with GC-content of the genes (correlation coefficient = 0.62). Right: IC3 is associated with the presence of snoRNAs in transcriptome.

Step 5. Automated comparison of computed metagenes with previously known metagenes

Automatic comparison with previously computed metagenes can be done with the following command contained in the **_example1_5_OVCA_dometageneRNK.bat** file:

java -Xmx5000M -jar ./bin/BIODICA/BIODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -dometageneanalysis knowledge/metagenes/

This command uses a folder name as an argument for **-dometageneanalysis** action. In this case BIODICA searches for .rnk files in the specified folder and load metagenes from them. A correlation table between computed ICs and the previously defined metagenes is produced following the same specification as in the previous step. In OVCA example, we used a set of meta-components, produced by averaging independent components gene projections inside detected pseudocliques in the correlation graph from (Biton et al, 2014).

In OVCA example, the command produces **work/OVCA_MGENE/OVCA_S_associationsRNK.xls** (containing –log10(pvalue) values) and **work/OVCA_MGENE/OVCA_S_associationsRNK_info.xls** (containing the row correlation coefficients) files. The later one is shown in Figure 8.

From Figure 8, one can easily identify certain components from the decomposition. For example, IC1 is determined as other tissues contamination component, IC4 is associated to highly reproducible immune component, IC7 matches the previously identified cell cycle, IC14 matches well the GC-content-related component identified previously and as it was already detected in Figure 7. IC2 matches a pseudo-clique which was previously not interpreted and which did not contain the ovarian cancer component in the analysis from (Biton et al, 2014). This makes it interesting to further investigate. Similarly, IC3 matches a small lung cancer-specific component without clear interpretation for the moment. Interesting to notice that bladder cancer-specific pseudo-cliques and breast cancer-specific pseudo-cliques are not associated to ICs in this analysis, as expected.

VAL	INTERPRETATION	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9	IC10	IC11	IC12	IC13	IC14	IC15	IC16	IC17	IC18	IC19	IC20
BCT13_metascore	-	0.0	0.1	0.0	_	_	0.0	_	_	_	_	_	-0.1	_	_	0.0	-0.1	_	0.1	_	0.1
BCT16_metascore	-	-0.1	_	-0.1	_	-0.1	0.1	-0.1	0.1	_	_	0.1	_	0.1	0.0	0.0	_	0.1	0.0	_	-0.1
BCT17_metascore	-	_	0.1	-0.1	0.0	0.0	-0.1	-0.1	-0.1	_	-0.2	0.2	0.1	_	_	0.0	-0.1	0.1	-0.2	0.1	-0.1
BCT20_metascore	-	0.1	0.1	0.0	0.0	0.1	0.1	_	0.1	_	_	-0.1	_	_	0.0	-0.1	_	_	0.0	_	_
BCT5_metascore	-	0.2	0.0	0.0	0.1	0.0	0.2	-0.1	0.0	0.0	0.1	_	-0.1	0.1	-0.1	_	_	0.1	_	0.0	0.1
BCT7_metascore	-	0.0	_	_	-0.2	-0.1	_	0.0	_	-0.1	0.1	0.1	0.0	0.0	_	0.1	0.1	_	0.0	_	0.0
BCTR15_metascore	-	_	0.1	-0.1	0.1	_	0.1	0.0	_	0.2	-0.1	_	_	-0.1	_	0.1	0.1	_	_	_	-0.1
BCTR7_metascore	-	-0.1	0.6	_	_	-0.1	0.1	-0.1	_	0.1	_	0.0	_	0.0	_	0.0	_	_	0.1	_	0.0
CIT11_metascore	CIT-11 (unknown)	-0.1	0.0	0.0	0.0	0.0	-0.1	-0.1	_	-0.1	_	0.2	0.1	0.0	0.1	-0.1	-0.1	0.1	0.0	_	-0.1
CIT12_metascore	Myofibroblasts	0.6	0.0	-0.1	0.1	0.1	-0.1	0.0	-0.1	0.0	_	0.1	_	_	-0.1	0.1	0.1	0.1	_	_	_
CIT13_metascore	Bladder cancer pathways	_	0.1	0.0	0.0	0.1	0.1	-0.1	_	_	_	0.0	_	_	_	0.1	_	0.1	0.0	0.1	0.1
CIT14_metascore	Stress/inflamation	0.3	_	0.1	0.1	_	_	_	-0.1	0.1	-0.1	0.3	_	0.2	-0.1	-0.1	_	0.0	-0.1	0.2	_
CIT2_metascore	GC-content	0.0	_	-0.1	_	-0.1	_	_	-0.1	-0.1	-0.1	0.0	_	_	-0.7	-0.1	-0.3	-0.1	_	0.1	_
CIT20_metascore	CIT-20 (unknown)	_	_	_	_	_	-0.1	0.1	_	0.1	-0.1	0.1	0.1	0.1	_	-0.1	_	_	-0.1	_	-0.2
CIT3_metascore	Smooth_muscle	0.4	_	-0.1	0.1	0.1	0.1	-0.1	0.1	0.1	0.1	0.0	-0.1	_	-0.1	0.1	0.1	0.2	0.1	-0.1	_
CIT4_metascore	Gender	0.0	0.0	-0.4	_	-0.1	0.4	0.1	0.0	0.0	-0.1	0.1	0.0	0.0	_	-0.3	0.0	-0.1	0.0	_	0.4
CIT5_metascore	Interferon	_	_	0.1	0.3	-0.3	-0.1	0.2	0.1	_	_	0.2	0.1	0.0	-0.1	0.0	_	-0.2	0.1	_	-0.3
CIT6_metascore	Basal-like	0.0	0.0	-0.1	-0.1	0.1	-0.1	0.0	0.1	0.0	0.1	0.3	0.1	_	_	_	-0.1	_	0.0	0.0	-0.1
CIT7_metascore	Cell-cycle	0.0	-0.1	_	-0.1	_	-0.1	0.6	0.1	-0.1	-0.2	-0.1	-0.1	0.1	_	-0.1	_	-0.1	0.0	-0.1	0.0
CIT8_metascore	Lymphocytes	0.0	-0.1	_	0.7	-0.1	0.1	_	-0.1	_	_	_	0.0	0.0	_	_	-0.1	_	0.0	_	_
CIT9_metascore	Urothelial differentiation	-0.1	_	0.0	-0.1	_	0.0	_	_	_	-0.1	_	_	_	0.0	-0.1	0.0	0.1	_	0.0	0.0
CO1_metascore	-	-0.1	_	-0.1	_	_	_	-0.1	_	_	_	_	_	_	0.1	-0.1	0.0	_	_	_	_
CO16_metascore	-	0.0	_	-0.1	_	-0.1	_	-0.1	-0.1	_	-0.1	_	_	0.1	_	_	-0.1	0.1	-0.1	0.0	_
CO2_metascore	-	0.1	0.1	-0.1	-0.1	0.1	_	-0.1	0.0	0.0	0.1	0.3	0.1	_	0.1	-0.1	_	_	0.1	_	-0.1
CO9_metascore	-	_	_	_	-0.1	0.2	_	_	_	-0.1	_	_	-0.1	-0.1	_	_	0.1	0.1	_	0.1	0.1
LU17_metascore	-	-0.2	0.0	-0.2	-0.1	0.1	_	0.2	0.2	_	_	_	_	-0.1	-0.1	_	0.2	0.0	0.1	-0.1	-0.1
LU6_metascore	-	-0.2	_	0.5	-0.2	0.0	0.1	_	_	_	_	-0.2	0.0	_	0.1	-0.1	_	_	0.0	_	_
LU7_metascore	-	_	_	-0.1	-0.1	_	0.0	_	_	0.1	-0.1	-0.1	_	_	_	_	-0.1	0.0	_	_	_
LU8_metascore	-	-0.2	_	-0.1	_	0.0	_	-0.1	-0.1	_	_	0.1	_	0.1	0.0	_	-0.1	0.0	_	_	_

Figure 8. Table of comparison of computed ICs for OVCA dataset with previously computed metagenes from (Biton et al, 2014).

Steps 1-5 can be combined in one command line:

java -Xmx5000M -jar BODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -doicamatlab 20 -dogsea 100 –dometasampleanalysis data/OVCA_TCGA/sample_info/OVCA.txt -dometageneanalysis knowledge/geneproperties/genes.txt -dometageneanalysis knowledge/metagenes/

Step 6. Launching OFTEN analysis for associating independent components with PPI subnetworks

Some of the independent component can be associated to a subnetwork in a global network of pairwise interactions such as PPI network. BIODICA implements a simple OFTEN algorithm [Kairov et al, 2012] based on exploiting the percolation properties of a PPI graph and selecting the most significant largest

connected component composed of the top-contributing genes for each independent component. The analysis is done for three possible rankings defined by a component: from the positive side (PLUS), from the negative side (MINUS) and from the absolute values of gene contributions (ABS). In summary, the analysis will choose the ranking giving the most significant association with a subnetwork.

In the following example, we use HPRD PPI network as a global network of binary interactions between genes, in order to check which independent components from the computed ICA decomposition can be associated to a significant network of functionally related (interacting with each other) proteins.

After computation of independent components, OFTEN analysis can be launched by the following command line:

java -Xmx5000M -jar ./bin/BIODICA/BIODICA.jar -config config
-datatable data/OVCA_TCGA/transcriptome/OVCA.txt
-dooften knowledge/networks/undirected/hprd9_pc_clicks.xgmml#100,50,600,100

The 4 numerical values separated in the argument of the dooften action by # from the path to the PPI network in xgmml format, and by a comma from each other are the parameters of the scanning and sampling the random networks of the given size and the given connectivity distribution from. The follow in the order: *nstart* - minimal number of top genes in the ranking selected for testing, *nstep* – step with which the scanning is done, *nend* – maximal number of top genes in the ranking selected for testing, *nperm* – number of random network samples constructed in order to estimated statistical significance of the OFTEN score.

As a result of this computation, the following summary tables are produced:

OFTEN report table:

In this table, one has to read the scores and p-values of the corresponding rankings. For example, it can be read that the components IC1, IC4, IC7 have the strongest associations to a subnetwork of PPI interactions. Indeed, the previous interpretation (Step 5) already associated them to the presence of Myofibroblasts and stress, Lymphocytes and to the cell cycle correspondingly.

NAME	PLUS_GENES	PLUS_N	PLUS_SC	PLUS_PVAL	MINUS_GENES	MINUS_N	MINUS_SC	MINUS_P	ABS_GENE	ABS_N	ABS_SC	ABS_PVAL
IC1	200	55	0.388015	0	600	45	0.09003068	0	200	54	0.382692	0
IC2	100	2	0.090476	0.05	100	6	0.06357142	0	100	2	0.082609	0.05
IC3	150	2	0.044722	0.19	600	30	0.020964913	0.23	150	2	0.044324	0.18
IC4	550	202	0.49936	0	150	5	0.030229887	0.05	600	204	0.466386	0
IC5	600	42	0.08697	0	100	9	0.12145161	0	200	9	0.052975	0
IC6	500	19	0.054449	0.01	350	13	0.01442857	0.2	150	2	0.018431	0.51
IC7	100	49	0.567089	0	600	58	0.09491018	0	100	46	0.525823	0
IC8	100	2	0.021964	0.38	400	32	0.09271653	0	250	8	0.037219	0
IC9	450	17	0.04248	0	350	17	0.03911628	0.03	200	4	0.010804	0.01
IC10	550	36	0.060428	0.02	100	4	0.04343283	0	350	10	0.020946	0.05
IC11	600	58	0.107157	0	600	62	0.04686217	0.12	600	49	0.079337	0
IC12	550	26	0.034573	0.09	300	16	0.051755317	0	600	40	0.02973	0.15
IC13	500	45	0.070295	0.04	100	3	0.01835821	0.1	600	45	0.08416	0
IC14	150	2	-0.00347	0.88	600	65	0.009693592	0.45	100	2	0.002917	0.77
IC15	250	7	0.012612	0.26	400	23	0.06441861	0.01	600	41	0.037555	0.13
IC16	600	39	0.077679	0	600	75	0.09206687	0	200	5	0.022435	0
IC17	600	35	0.059855	0	600	20	0.024358975	0.06	100	2	0.020984	0
IC18	100	2	0.011515	0.17	450	32	0.09391304	0	550	17	0.029559	0
IC19	550	37	0.077768	0	150	3	0.00990099	0.17	550	32	0.070292	0
IC20	250	48	0.241558	0	450	32	0.051204383	0.05	100	16	0.1576	0

OFTEN report summary table:

The summary table represents the same information in a more compact way, selecting and reporting on only one, the most significant, ranking (PLUS, MINUS or ABS).

LABEL	SCORE	PVAL	NGENES	N	TYPE
IC1	0.388015	0	200	55	PLUS
IC2	0.090476	0.05	100	2	PLUS
IC3	0.044722	0.19	150	2	PLUS
IC4	0.49936	0	550	202	PLUS
IC5	0.121452	0	100	9	MINUS
IC6	0.054449	0.01	500	19	PLUS
IC7	0.567089	0	100	49	PLUS
IC8	0.092717	0	400	32	MINUS
IC9	0.04248	0	450	17	PLUS
IC10	0.060428	0.02	550	36	PLUS
IC11	0.107157	0	600	58	PLUS
IC12	0.051755	0	300	16	MINUS
IC13	0.08416	0	600	45	ABS
IC14	0.009694	0.45	600	65	MINUS
IC15	0.064419	0.01	400	23	MINUS
IC16	0.092067	0	600	75	MINUS
IC17	0.059855	0	600	35	PLUS
IC18	0.093913	0	450	32	MINUS
IC19	0.077768	0	550	37	PLUS
IC20	0.241558	0	250	48	PLUS

The analysis performed stores in the work folder all subnetworks associated to the components in xgmml format, which can be opened in Cytoscape environment. As an example of such a network associated to a component, let us demonstrate the subnetwork of 49 connected proteins associated to IC7 (cell cycle), see Figure 9. As one can see, the subnetwork collects the classical cell cycle genes with the major hub CDK1.

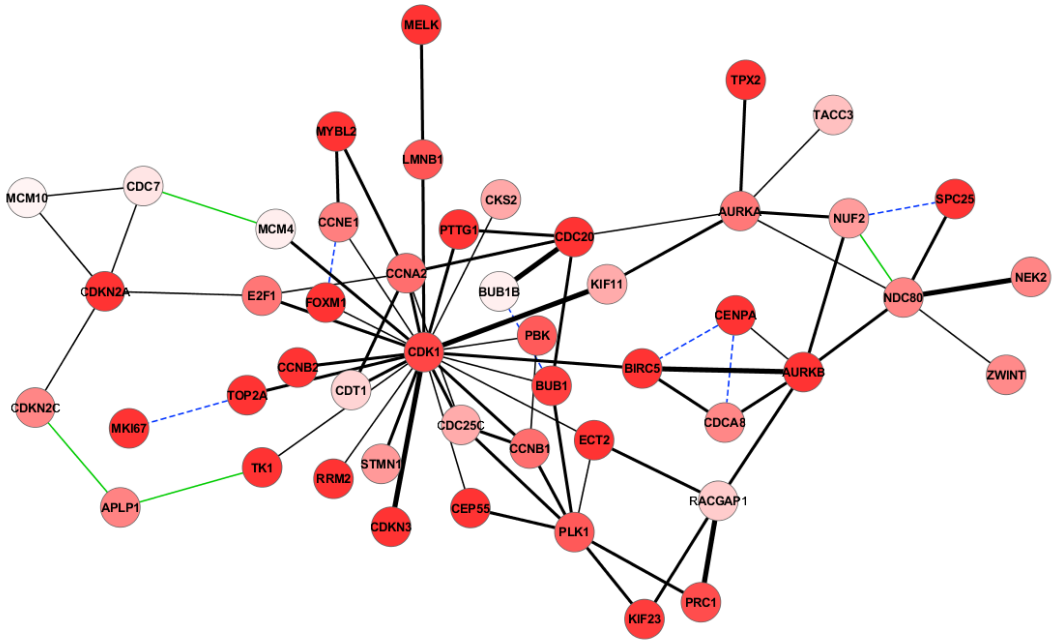

Figure 9. Proteins whose genes are top contributing to the positive side of the components IC7 associated to cell cycle and connected to each other by protein-protein interactions as described in HPRD database. The intensity of the red color shows the contribution of the gene to the IC7 component. Different colors and line types of edges corresponds to different evidences for the interaction (thick black corresponds to the most confident, dashed blue signifies co-existence in the same protein complex, green line corresponds to only one evidence of interaction from a yeast 2-hybrid large scale screening).

Step 7. Optimizing the number of components to compute

In the previous steps of this example, the number of the components was pre-defined (n=20). Here we apply the BIODICA procedure for getting an estimation of the optimal number of independent components to compute.

At first, one has to pre-compute a number of ICA analysis with various numbers of components specified:

java –Xmx8000M -jar BODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt -doicamatlab 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30

On an ordinary laptop, this computation takes approximately 3-4 hours.

Secondly, one can analyze the pre-computed decompositions in order to select the optimal number of components:

java –Xmx8000M -jar BODICA.jar -config config -datatable data/OVCA_TCGA/transcriptome/OVCA.txt –donumbercomponents stability

This produces the following report:

```
NUMBER_OF_COMPONENTS       AVERAGE_STABILITY
2      0.892  *******
3      0.942  ********      <--
4      0.822  ******
5      0.876  *******       <--
6      0.806  ******
7      0.86   *******
8      0.871  *******
9      0.884  *******       <--
10     0.858  *******
11     0.834  ******
12     0.818  ******
13     0.743  ****
14     0.719  ****
15     0.701  ****
16     0.658  ***
17     0.661  ***    <--
18     0.634  **
19     0.662  ***
20     0.669  ***    <--
21     0.643  **
22     0.598  *
23     0.594  *
24     0.561  *
25     0.535
26     0.522
27     0.511
28     0.511         <--
29     0.511
30     0.484

For the stability level: 0.9: optimal choices are  3(advised)
For the stability level: 0.8: optimal choices are  3, 5, 9(advised)
For the stability level: 0.7: optimal choices are  3, 5, 9(advised)
For the stability level: 0.6: optimal choices are  3, 5, 9, 17, 20(advised)
For the stability level: 0.8(minimum tolerable): optimal choices are  3, 5, 9(advised)
Final choice: 9 components
```

The conclusion of this analysis is that for the default and relatively conservative threshold 0.8 of minimum required (tolerable) stability of components, the advice is to take 9 components. It means that the 9 most stable components have more chances to be reproducible in other analyses. However, for more explorative threshold 0.6 one can compute 20 components, as it was done in the first steps of this example of BIODICA application.

Example 2. Computation of bi-directional best hit (BBH) correlation graph between multiple sets of metagenes

BIODICA can be used to perform metaanalysis based on ICA decompositions computed for several independent datasets and comparing the resulting metagenes with each other. The comparison that was suggested in (Biton et al, 2014) is based on detecting the best-bidirectional hits (BBH) between two sets of metagenes (analogous to evolutionary bioinformatics where BBH notion is used to define orthologous genes).

BBH is defined as the maximal correlated component from a set of metagenes A to a set B, which, at the same time, is the maximal correlated component from the set B to the set A. Correlation between metagenes is computed as Pearson correlation between the set of common genes in two metagenes.

In order to compute the BBH graph which will match the metagenes defined in several datasets, one need to decompose several datasets (e.g., gene expression datas) and store the resulting files containing metagenes in one folder. The folder should contain several files with names ending with "_S.xls" suffix (all preceding letters in the name of the file will be used to label the dataset).

As an example of application of this approach, we will test a set of 8 bladder cancer ICA decompositions and 5 breast cancer decompositions, provided as a part of BIODICA distribution package (data/ folder).

The BBH graph can be constructed using the following command line

java -Xmx5000M -jar ./bin/BIODICA/BIODICA.jar -config config –dobbhgraph data/METAANALYSIS_TEST/

This will generate a number of pairwise comparisons saved in the form of graphs as xgmml files which can be opened in Cytoscape environment. All these files will be finally assembled in the final BBH graph named ***correlation_graph_norecipedges.xgmml***.

In this example the computation involves computing 15000 correlation coefficients and this number grows quadratically with the number of datasets. Note that the command can be launched several times; in this case many pairwise comparisons will be done in parallel without interfering with each other. This can accelerate the computation of the BBH graph by distributing the computational load among several processors. Also, if the pairwise comparisons will be stored in the folder as xgmml files, then it is be possible to add novel ICA decompositions in the folder, and the command line will compute only the missing comparisons.

The result of the computation can be opened in Cytoscape through "File/Import/Network multiple types..." command. Three conditions can be used to filter the BBH graph edges, using the edge attributes: RECIPROCAL (true corresponds to BBH, empty is not BBH but simply maximal close hit), ABSCORR (absolute value of correlation between metagenes), LOG10PVAL (-log10 p-value of the

correlation coefficient; the maximum value is 16 which corresponds to the minimal p-value<10^{-16}). See an example of constructed BBH graph in Figure 10.

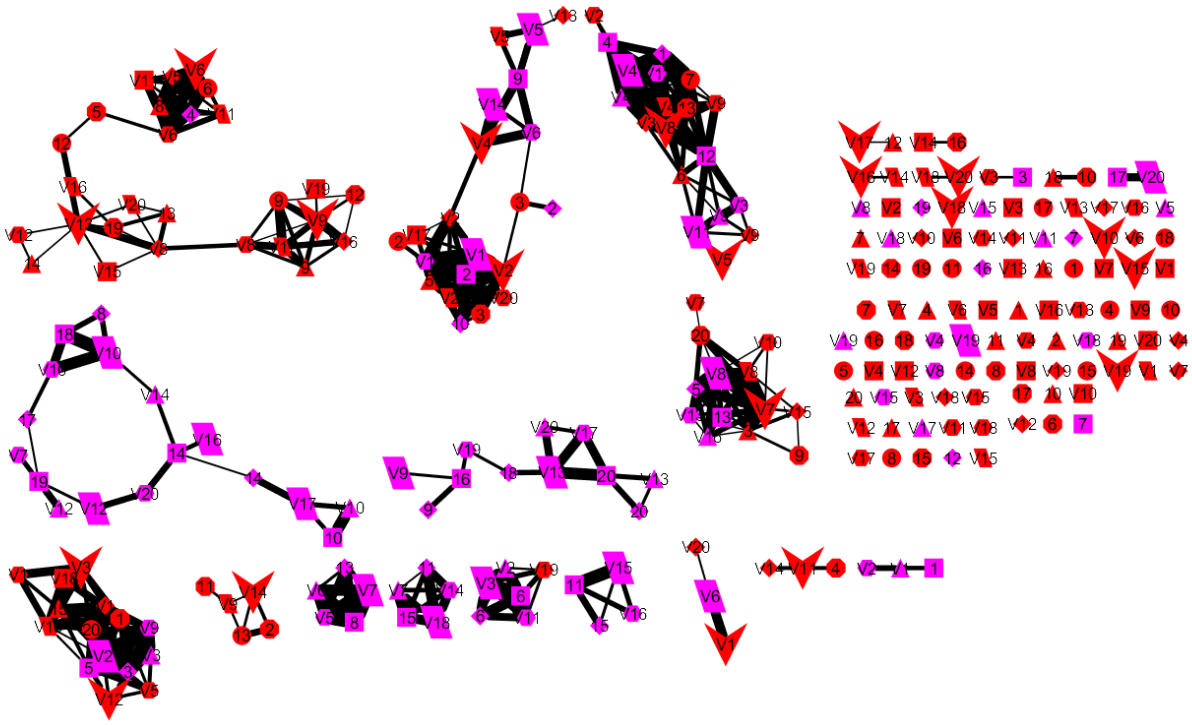

Figure 10. Best bi-directional hit (BBH) graph constructed for ICA decompositions computed for 8 bladder cancer (red color) and 5 breast cancer (purple color). The BIODICA output graph edges were filtered with the following conditions (reciprocal edge=true & -log10p-value>15.9 & absolute correlation>0.4). Communities with nodes of only one color correspond to reproducible signals specific to one cancer type, and communities mixing nodes of different colors correspond to reproducible signals common to both cancer types.

FUNCTIONS IMPLEMENTED IN BODICA ver 0.9

The following functions have been implemented in BIOIDCA version 0.9 (November 2016).

Big Data Management module

Several functions helping to prepare the dataset for analysis in BIODICA have been implemented (such as centering, double-centering, taking a log scale, imputing the missing values).

Several example datasets have been provided:

1) Ovarian TCGA transcriptomic dataset, several platforms (RNA-Seq, Affymetrix, Illumina).
2) Ovarian TCGA clinical annotations + various characteristics of tumor samples such as estimated ploidy, purity, LST (measure of large-scale genomic instability).

Prior Knowledge and Biological Networks module

The corresponding part of the data repository has been populated with

1) Reference gene sets from MSigDB; BioCarta+KEGG+Reactome+ACSN pathways
2) Reference metagenes from (Biton et al, 2014) study
3) Several types of gene-related annotations (genomic positions, GC-content)
4) PPI networks from HPRD, Reactome, ACSN databases
5) Gene regulatory networks from Signor, ACSN (binary relations).

Intense ICA Computation module

The automatic procedure for computing the given number of independent components and storing the results in the format suitable for further analysis has been implemented. Procedure for determining the optimal number of independent components to compute has been implemented.

Metasample Annotation module

Automated procedure for computing association p-values with various types of clinical information (numerical, Boolean, categorical) has been implemented.

Metagene Annotation module

Automated procedure for computing association p-values with various types of annotation fo genes (numerical, Boolean, categorical) has been implemented. It includes automated GSEA analysis of independent components, filtering the results of GSEA applications for the most significant enrichment results, representing the summary of the GSEA filtering in the form of html page. OFTEN (finding the Optimally Functionally Enriched subNetwork) analysis of the obtained components analysis has been implemented.

FUTURE DEVELOPMENT TO ACHIEVE THE VER 1.0 STATE OF BIODICA

In 2016, the development of BIODICA will be finalized to the version 1.0 with aim to accomplish the complete functionality described in Figure 2, including meta-analysis functionality, data visualization module and development of Graphical Using Interface (GUI). New original methods for interpreting metagenes and metasamples will be implemented. Connection of BIODICA reporting module to cBioPortal will be done. NaviCell Web Service will be included into the data visualization module. Version 1.0 of BIODICA will be a subject of a peer-reviewed publication in a visible systems biology journal.

To do list:

1) Make a possibility of modifying any parameter value from the values provided in the *config* file

CONTACTS

Implementation of BIODICA is a common project between Nazarbaev University (Centre for Life Sciences) and Institut Curie (Computational Systems Biology of Cancer laboratory). All quires about BIODICA state and development should be sent to

Andrei Zinovyev (http://www.ihes.fr/~zinovyev)

Ulykbek Kairov (https://www.researchgate.net/profile/Ulykbek_Kairov)

REFERENCES

1. Biton A., Bernard-Pierrot I., Lou Y., Krucker C., Chapeaublanc E., Rubio Perez C., Lopez Bigas N., Kamoun A., Neuzillet Y., Gestraud P., Grieco G., Rebouissou S., de Reynies A., Benhamou S., Lebret T., Southgate J., Barillot E., Allory Y., Zinovyev A., Radvanyi F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. 2014. Cell Reports 9(4), 1235-1245.
2. Zinovyev A., Kairov U., Karpenyuk T., Ramanculov E. Blind Source Separation Methods For Deconvolution Of Complex Signals In Cancer Biology. 2013. *Biochemical and Biophysical Research Communications* **430**(3), 1182-1187.
3. Kairov U., Karpenyuk T., Ramanculov E., Zinovyev A. Network analysis of gene lists for finding reproducible prognostic breast cancer gene signatures. 2012. *Bioinformation* **18**(6):773-776.
4. Barillot E., Calzone L., Hupe P., Vert J.-P., Zinovyev A. *Computational Systems Biology of Cancer* . Chapman & Hall, CRC Mathematical & Computational Biology, 2012, 452 p.
5. Biton A., Zinovyev A., Barillot E., Radvanyi F. MineICA: Independent component analysis of transcriptomic data. 2013. BioConductor vignette. URL: https://www.bioconductor.org/packages/release/bioc/vignettes/MineICA/inst/doc/MineICA.pdf
6. Jutten, C., and Hérault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Process. 24, 1 – 10.
7. P. Comon, Independent component analysis, a new concept?, Signal Process. 36 (1994) 287-314.
8. A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

9. A.N. Gorban, B. Kégl, D. Wunch, A. Zinovyev, Principal Manifolds for Data Visualization and Dimension Reduction, Lecture Notes in Computational Science and Engineering, Springer, Berlin-Heidelberg, 2008.
10. G. Hori, M. Inoue, S.I. Nishimura, H. Nakahara, Blind gene classification—an application of a signal separation method. Genome Informatics 12 (2001) 255–256.
11. W. Kong, C.R. Vanderburg, H. Gunshin, J.T. Rogers, X. Huang, A review of independent component analysis application to microarray gene expression data, Biotechniques 45(5) (2008) 501–520.