University of Crete
School of Sciences and Engineering
Computer Science Department

SCENERY: a Web-Based Application for Network Reconstruction, Visualization and Statistical Analysis of Single-Cell Data.

Giorgos Athineou

Master of Science Thesis

Heraklion, October 2016

UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

SCENERY: a Web-Based Application for Network Reconstruction, Visualization and Statistical Analysis of Single-Cell Data.

Thesis submitted by
Giorgos Athineou
in partial fulfillment of the requirements for the
Master of Science degree in Computer Science

THESIS APPROVAL

Author: ______________________________
Giorgos Athineou

Committee approvals: ______________________________
Ioannis Tsamardinos
Associate Professor, Thesis Supervisor

____
Ioannis Tollis
Professor, Committee Member

____
Vassilis Christophides
Professor, Committee Member

Departmental approval: ______________________________
Antonios Argyros
Professor, Director of Graduate Studies

Heraklion, October 2016

Abstract

Cytometry techniques allow the quantification of the morphological characteristics and protein abundances at a single-cell level. Data collected with these techniques can be used for addressing the fascinating, yet challenging problem of reconstructing the network of protein interactions, forming signaling pathways and governing cell biological mechanisms. Network reconstruction is an established and well studied problem in the machine learning and data mining fields, with several algorithms already available. Moreover, standard statistical analysis on such data is widely used, mainly for modeling the relationship among proteins and comparing different cell populations. In this thesis, we present the first, freely available, web-oriented application, SCENERY from "Single CEll NEtwork Reconstruction sYstem", that allows scientists to rapidly apply state-of-the-art network-reconstruction methods along with standard pre-processing and statistical analysis functions on cytometry data, through advanced visualization functions.

SCENERY comes with an easy-to-use, step-wised user interface, along with an open modular architecture for ease of its extension. The functionalities of the application are illustrated and validated on data from a publicly available immunology experiment.

Περίληψη

Οι τεχνικές κυτταρομετρίας επιτρέπουν την ποσοτικοποίηση των μορφολογικών χαρακτηριστικών και της αφθονίας πρωτεΐνων σε επίπεδο ενός κυττάρου. Τα δεδομένα που συλλέγονται με αυτές τις τεχνικές μπορούν να χρησιμοποιηθούν για την αντιμετώπιση του συναρπαστικού μεν αλλά και απαιτητικού προβλήματος της ανασυγκρότησης του δικτύου των αλληλεπιδράσεων των πρωτεινών που σχηματίζει μονοπάτια σηματοδότησης και διέπει βιολογικούς μηχανισμούς του κυττάρου. Η ανακατασκευή δικτύων είναι ένα καθιερωμένο και καλά μελετημένο πρόβλημα σε τομείς όπως αυτούς της μηχανικής μάθησης και της εξόρυξης δεδομένων, με αρκετούς αλγορίθμους να είναι ήδη διαθέσιμοι. Επιπλέον, η τυπική στατιστική ανάλυση των δεδομένων αυτών χρησιμοποιείται ευρέως, κυρίως για την μοντελοποίηση της σχέσης μεταξύ των πρωτεΐνών και τη σύγκριση διαφορετικών πληθυσμών κυττάρων. Σε αυτή την εργασία, παρουσιάζουμε το SCENERY, την πρώτη ελεύθερα διαθέσιμη διαδικτυοκεντρική εφαρμογή, που επιτρέπει στους επιστήμονες να εφαρμόσουν γρήγορα και εύκολα μεθόδους, τελευταίας τεχνολογίας, για την ανασυγκρότηση δίκτυων, σε συνδιασμό με λειτουργίες προ-επεξεργασίας και στατιστικής ανάλυσης δεδομένων κυτταρομετρίας, μέσα από προηγμένες λειτουργίες απεικόνισης των δεδομένων και των αποτελεσμάτων. Το SCENERY διέπεται από μια εύκολη στη χρήση, σταδιακά διαβαθμισμένη διεπαφή χρήστη σε συνδιασμό με μια ανοικτή, αρθρωτά σχεδιασμένη αρχιτεκτονική για ευκολία της επέκτασής του. Οι λειτουργίες της εφαρμογής παρουσιάζονται και πιστοποιούνται σε δεδομένα από ένα δημόσια διαθέσιμο πείραμα στον τομέα της ανοσολογίας.

Acknowledgments

This work was funded by European Research Council (ERC) and is part of the CAUSAL-PATH - Next Generation Causal Analysis project, No 617393.

I, sincerely, thank Karen Sachs, David Gomez-Cabrero, Angelika Schmidt, Jesper Tegner and Ioannis Basdekis for their invaluable comments, suggestions and encouragement on this project. I would also like to acknowledge my professor, advisor and mentor, Prof. Ioannis Tsamardinos, who gave me the opportunity and the knowledge to work beside him, the MensXMachina group members and especially my tireless supervisors Dr. Vincenzo Lagani and Dr. Georgio Papoutsoglou, along with Dr. Sofia Triantafillou and Iordani Ksanthopoulo for their priceless time, work and effort in this project.

Last but not least, this thesis is dedicated to my family and good friends for their support and encouragement at my best and worst.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Motivation 2

1.3 The Application 2

1.4 Related Work 3

2 The SCENERY Application **4**

2.1 Software Functionality 4

2.2 Data Loading 5

2.3 Analysis Setup 7

2.4 Perform Analysis 8

3 Methods **11**

3.1 Data Visualization 11

3.2 Pre-Processing 13

3.3 Univariate Statistical Analysis 17

3.4 Network Reconstruction Analysis 20

3.4.1 (Conditional) Association Networks 23

3.4.2 Probabilistic Causal Networks 25

3.4.3 Bayesian Networks 28

4 Architecture **31**

4.1 Software Architecture 31

4.2 Modularity 34

5 Application on Immunology **37**

5.1 Definition of the Problem 37

5.2 Analyzing Cytometry Data with SCENERY 37

6 Conclusions and Future Work **40**

List of Figures

2.1 Flowchart of typical user-application interaction. In step 1 users upload data and define the study design. In step 2 they setup a computational experiment by selecting datasets and the analysis method. In step 3, users calibrate the input parameters and execute the analysis. The analysis can be reconfigured and repeated multiple times. 4
2.2 Data Loading in SCENERY. 5
2.3 Experimental design table uploading by submitting a CSV/TXT file or by filling a dynamic HTML table online, as incorporated in the first step of SCENERY's wizard. 6
2.4 Experimental design table and metadata information of the uploaded datasets, as incorporated in the first step of SCENERY's wizard. 7
2.5 Data selection and filtering, in terms of the uploaded 'Experimental Design Table' as defined in the second step of SCENERY wizard. 8
2.6 Typical layout and structure of a method in SCENERY showing the main panels and tabs as incorporated in the third step of the wizard. Here, as a use-case, we run an analysis on flow cytometry data with the MMPC NR analysis method. 9

3.1 UI of the 'Data Visualization' method, showing a histogram of a selected measurement from a mass cytometry file. 12
3.2 UI of the 'Data Visualization' method, showing an overlapping scatter-plot among three selected measurements from a mass cytometry file. . . . 12
3.3 UI of the 'Data Visualization' method, showing a scatter-plot matrix among one selected marker versus four different selected measurements from a mass cytometry file. 13
3.4 UI of the 'Compensation' pre-processing method. 14
3.5 UI of the 'Transformation' pre-processing method. 15
3.6 UI of the 'Gating' pre-processing method. 16
3.7 UI of the 'Factor Analysis' statistical analysis method. 18
3.8 UI of the 'Logistic Regression' statistical analysis method. 19
3.9 Summary of the typical network's characteristics by applying several graph theory analysis algorithms in a reconstructed association network, derived by the 'Correlation' NR method. 22

3.10 UI of the ’Correlation’ NR analysis method, showing a textual representation of the results in the ’Summary’ tab, after submitting an analysis. . . 24
3.11 UI of the ’MMPC’ NR analysis method, showing the generated reconstructed network in the ’Plots’ tab, after submitting an analysis. 25
3.12 UI of the ’PC’ NR analysis method, showing the generated reconstructed network in the ’Plots’ tab, after submitting an analysis. 26
3.13 UI of the ’FCI’ NR analysis method, showing the generated reconstructed network in the ’Plots’ tab, after submitting an analysis. 27
3.14 UI of the ’IDA’ NR analysis method, showing a textual representation of the results in the ’Summary’ tab, after submitting an analysis. 28
3.15 UI of the ’HC’ NR analysis method, showing the generated reconstructed network in the ’Plots’ tab, after submitting an analysis. 29

4.1 Overview of SCENERY’s architecture components and interactions among the different users of the application. 31
4.2 Overview of SCENERY’s Docker-based architecture moved on a cloud server. Currently on development. 33
4.3 Standardized modular architecture of the R methods in SCENERY. 34

5.1 Visualizing results in SCENERY. The retrieved reconstructed network after applying the MMPC method on selected mass cytometry data (see text for details). The analysis calibration panel is also displayed at the left of the reconstructed network, as indicative of the UI. 38
5.2 Visualizing results in SCENERY. The retrieved overlapping density plot by applying a population comparison among two donors for the protein marker p38. The univariate statistical analysis method ’Factor Analysis’ was used on two selected sub-populations from a mass cytometry public dataset (see text for details) with respect to the ’donor.id’ factor from the experimental design. The analysis calibration panel is also displayed at the left of the reconstructed network, as indicative of the UI. 39

List of Tables

3.1 Summary table of the current analysis methods in SCENERY. 30

Chapter 1

Introduction

1.1 Background

Single-cell analysis is becoming increasingly popular in biology, lately, for numerous reasons. Analyzing the mechanisms underlying single cells is mainly about studying the smallest organizational system that by definition implies life. The fact that such a small system maintains, through different organizations and over millions of years, such a high complexity and hierarchical structure imparts a fascinating and challenging aspect on the analysis of this kind of data [1], [2]. The statistical analysis of single cell data usually consists of standard pipelines that are mainly aim to a mathematical indication and understudying of the cells' biological reactions and mechanisms. This kind of analysis is mainly used for modeling the relationships among a limited set of molecular quantities, such as proteins, for identifying and comparing different cell populations and for reconstructing the structure and the properties of the network of different molecular events that forms signaling pathways and governs cell biological mechanisms.

Signaling pathways or signaling networks are well-organized chains of complex molecular events [3]. Molecular stimuli trigger these events by orderly changing the state of specific proteins ultimately perturbing the cell's metabolism, shape, gene expression, or ability to divide.

Cytometry has been traditionally used for extracting information on intra- and extracellular properties on a single cell level. The most widely used format of cytometry is flow cytometry. Flow cytometry is a robust and broadly accessible method nowadays, able to provide quantitative measurements on such sensitive macro-molecular interactions [4] by using fluorochromes in order to detect the corresponding measurements. Still, the statistical analysis and the reconstruction of signaling networks from flow cytometry measurements has not become popular, primarily due to the limited molecular quantities the method can measure. Recently, a novel format of cytometry termed Mass Cytometry was introduced revolutionizing the state of the art [4]. Mass cytometry uses a set of tagged antibodies in order to detect each molecular measurement and its inherent ability to investigate more than 30 quantities simultaneously, offers, now, the opportunity to delve deeper into the analysis of such data.

1.2 Motivation

Inducing signaling networks from data can be thought as a Network Reconstruction (NR) problem. NR methods have become increasingly popular in biology, especially for inferring gene-gene interaction networks, with numerous scientific works currently published on this subject [5], [6]. The first successful case of signaling NR in the cytometry field was achieved by Sachs and co-authors [7], followed by several applications [8], [9]. However, NR methods are not yet routinely used on single-cell cytometry data. Arguably, this is mainly due to the intrinsic complexity of the task. Attempting to reconstruct signaling pathways requires knowing in detail the semantics of the data, the peculiarities of the cytometry technology, and all available information on the specific pathway and its components. On the other hand, successfully applying NR algorithms requires mastering all the technicalities of these methods, since inaccuracies in the analysis pipeline are potentially able to invalidate all results [10]. In addition to those intrinsic problems, a horizontal factor that limits broadcasting, sharing and reusing scientific results in this domain is the reluctance towards social media [11], [12] and the limited use of online services that support sharing of credible and accurate analysis methods. These tools promote openness, hiding at the same time sensible, core aspects of an experiment of this kind (in order to prevent copyright infringement).

1.3 The Application

In this work, we present SCENERY (Single CEll NEtwork Reconstruction sYstem), a web-based application specifically devised to allow researchers to apply NR methods and standard pre-processing, statistical analysis and advanced visualization methods on single-cell cytometry data, even with limited knowledge of the technical details of these algorithms. In order to ensure a complete, efficient and robust platform for single-cell analysis, this work was focused on the development of the modular architecture and the appropriate functionality after deriving feedback from experts in various, relevant, fields such as human computer interaction, computational biology and particularly, cytometry analysis. Moreover, one of our main goals in this work was to render this type of analysis accessible, especially, in non-experts users in data analysis. This ensures an ease of selecting the appropriate pipeline and rapidly applying advanced, state-of-the-art computational methods and standard work-flows in single-cell analysis by avoiding the common and most of the times, demanding programming and algorithmic overhead associated with such types of analyses. SCENERY interface guides the user through a set of easy steps; from data loading and study design specification, to the set-up of the analysis and results visualization and sharing. Its core is built on R and its modularity grants to easily add extensions, particularly additional analysis methods and visualizations. Finally, worth noted is that the idea and a part of this work was first introduced in [13].

1.4 Related Work

To the best of our knowledge, SCENERY is the first available software of its kind. Several other applications exist for cytometry data analysis, both as stand-alone softwares (FlowJo, `www.flowjo.com`); web-services (CytoBank, [14]); R packages such as the ’flowCore’ [16], ’flowStats’ [17], ’flowViz’ [18] and ’flowWorkspace’ [19] and libraries (e.g. FlowPy [20], [15]); however none of these tools provide the user with NR functionalities.

The rest of this thesis is structured as follows. First, we provide an overview of SCENERY functionalities and current incorporated methods with indicating examples and visualizations (chapters 2, 3), and then a closer look to its internal modular architecture (chapter 4). A use case on real, publicly available data from the immunology field is then presented for better illustrating SCENERY capabilities (chapter 5). Finally, conclusions, future work and extensions of the application are discussed in chapter 6.

Chapter 2

The SCENERY Application

2.1 Software Functionality

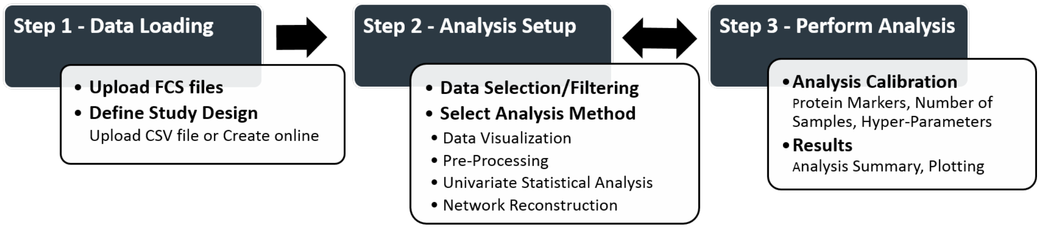
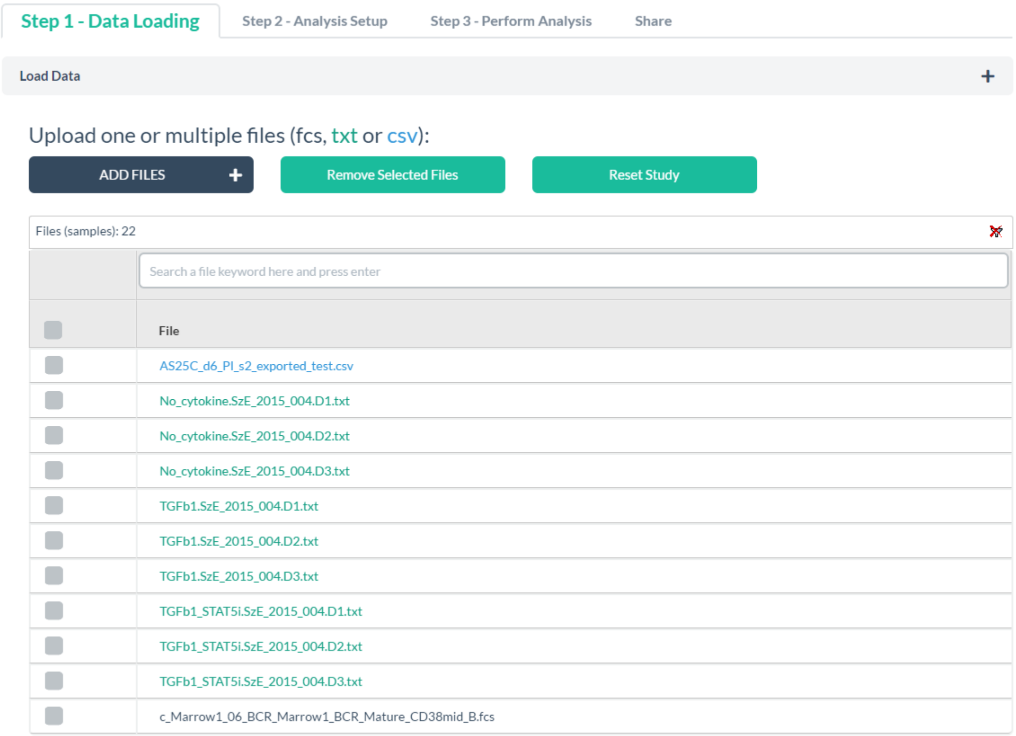


Figure 2.1: Flowchart of typical user-application interaction. In step 1 users upload data and define the study design. In step 2 they setup a computational experiment by selecting datasets and the analysis method. In step 3, users calibrate the input parameters and execute the analysis. The analysis can be reconfigured and repeated multiple times.

The functionality of SCENERY is structured by a step-wised wizard that guides the users through a specific sequence of analysis steps, as shown in Fig. 2.1. Basic aspects of the overall design and parameterized functionality of this wizard lay from the input derived from experts and bibliography for available methods [21]. This input was covered in the interface and system design, implemented on the first version of the described system, which is currently tested as for usability, user acceptance and efficiency. The user supplies the data (Step 1), defines the computational experiment (Step 2) and then sets the execution parameters (Step 3). The user may run the analysis, and redefine the execution parameters or select another analysis method until the desired outcome is achieved. This sequence of steps also serves as an educational path for less experienced users who are interested in exploring any aspect of the available analysis methods. Moreover, as the system provides dedicated services to the biologist research community, to succeed wider acceptance of the achieved results, exported data are presented in well-known and acceptable to the community presentation and formats to make them universally readable and accessible. The analysis output can be exported in various ways, mainly publication-quality figures and standard formats for graph-representation (i.e., Graph Exchange XML Format, GEXF). In the next, currently developing, version of SCENERY the user will be able to share the output of the analysis privately (via email or a repository) or publicly (via social media or blogs accounts) to a group of colleagues, for further analysis and discussion. Recent social-media citation practices indicate that scientific content is becoming more and more part of every day's conversations, thus increasing chances of citation [12]. Finally, one of the most important features of SCENERY's functionality and architecture is its modularity, described in Section 4.2, that offers the ability to the users to, easily, produce, submit and incorporate their own single-cell analysis methods, privately or publicly, under SCENERY's structure and layout.

2.2 Data Loading



Upload one or multiple files (fcs, txt or csv):

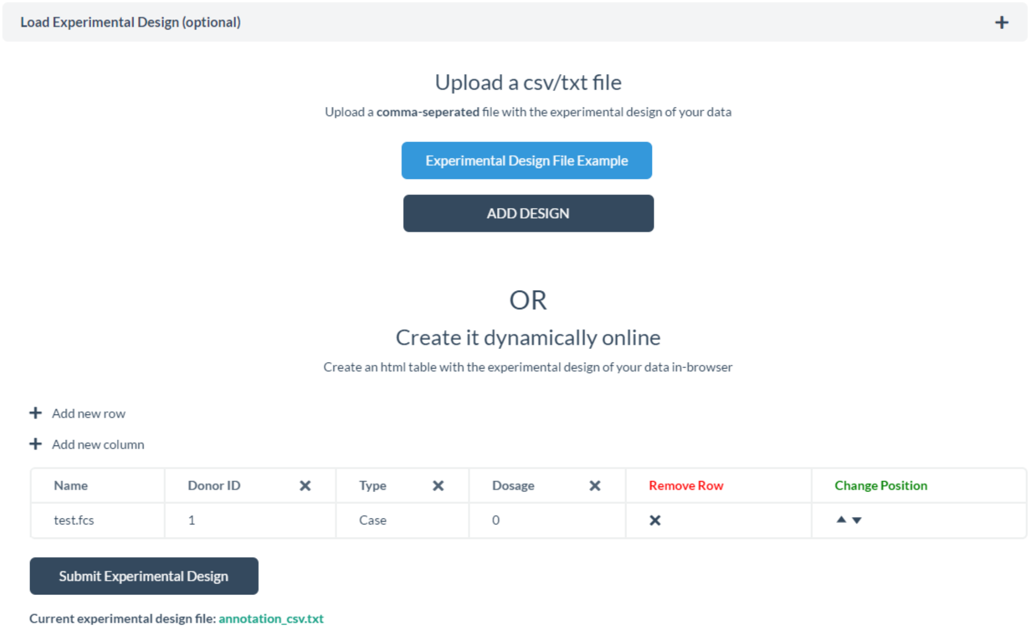
Files (samples): 22

□	File
□	AS25C_d6_PI_s2_exported_test.csv
□	No_cytokine.SzE_2015_004.D1.txt
□	No_cytokine.SzE_2015_004.D2.txt
□	No_cytokine.SzE_2015_004.D3.txt
□	TGFb1.SzE_2015_004.D1.txt
□	TGFb1.SzE_2015_004.D2.txt
□	TGFb1.SzE_2015_004.D3.txt
□	TGFb1_STAT5i.SzE_2015_004.D1.txt
□	TGFb1_STAT5i.SzE_2015_004.D2.txt
□	TGFb1_STAT5i.SzE_2015_004.D3.txt
□	c_Marrow1_06_BCR_Marrow1_BCR_Mature_CD38mid_B.fcs

Figure 2.2: Data Loading in SCENERY.

As first step, the users upload one or more data files in various formats such as TXT, CSV or, mainly, as Flow Cytometry Standard (FCS), listing to the users' account, as shown in Fig. 2.5. FCS files are universally used for storing and exchanging flow and mass cytometry data, with all major software for cytometry analysis adopting this standard. This, also allows SCENERY to import and analyze data, already, pre-processed by other applications. The main content of these single-cell data files is a dataset corresponding to the expressions of different markers (columns), such as proteins or experimental parameters, through a cell population (rows). In the rest of this thesis the corresponding uploaded files will be referred as (FCS or data) files or datasets.

Experimental Study Design FCS files contain measurements that may correspond to different samples (e.g. patients, cell types) or conditions (e.g., stimuli, inhibitor dosages). For example, an FCS file may correspond to a different patient, to a specific cell type, may be produced by a different laboratory, or under a specific experimental setting. SCENERY goes beyond traditional study design declaration, and it allows users to assert any type of metadata knowledge concerning variables, quantities, attributes, or characteristics of the samples (e.g. gender, age, etc.). Hence, any type and number of factors can be defined in a custom study design, both qualitative (e.g, cell type) and quantitative (e.g., drug dosage). This flexibility permits to accommodate virtually all possible study designs, both current and future ones. A study design can be uploaded through a CSV or a TXT file, created in advance or created online, in-browser, by filling a dynamic two dimension (2D) table with the columns corresponding to different experimental design factors and the rows corresponding to the values of the factors for each FCS file that is under the current design, respectively, as shown in Fig. 2.3.



Load Experimental Design (optional)

Upload a csv/txt file

Upload a **comma-seperated** file with the experimental design of your data

OR

Create it dynamically online

Create an html table with the experimental design of your data in-browser

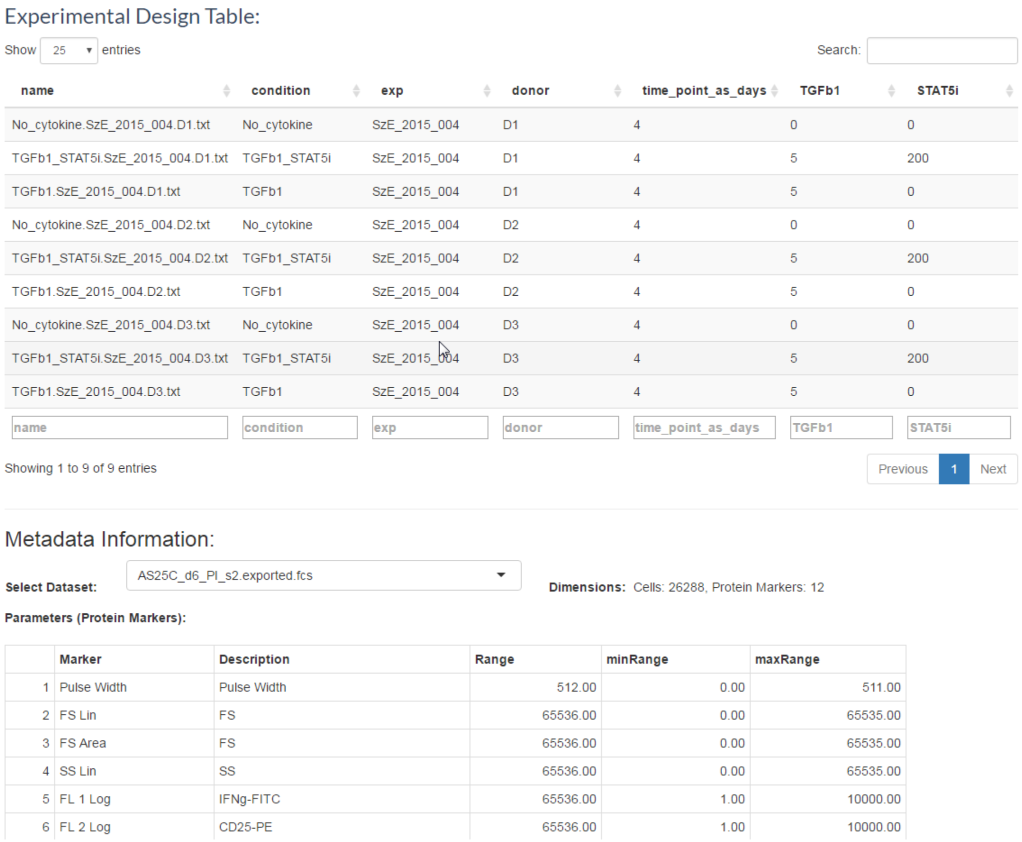
Name	Donor ID ✕	Type ✕	Dosage ✕	Remove Row	Change Position
test.fcs	1	Case	0	✕	▲▼

Current experimental design file: annotation_csv.txt

Figure 2.3: Experimental design table uploading by submitting a CSV/TXT file or by filling a dynamic HTML table online, as incorporated in the first step of SCENERY's wizard.

SCENERY automatically informs the user about the submitted study design in a 2D

HTML table format alongside with metadata information about the uploaded dataset such as the dimensions (number of cells and number of markers) and the markers' description and ranges summary (see Fig. 2.4).



Experimental Design Table:

name	condition	exp	donor	time_point_as_days	TGFb1	STAT5i
No_cytokine.SzE_2015_004.D1.txt	No_cytokine	SzE_2015_004	D1	4	0	0
TGFb1_STAT5i.SzE_2015_004.D1.txt	TGFb1_STAT5i	SzE_2015_004	D1	4	5	200
TGFb1.SzE_2015_004.D1.txt	TGFb1	SzE_2015_004	D1	4	5	0
No_cytokine.SzE_2015_004.D2.txt	No_cytokine	SzE_2015_004	D2	4	0	0
TGFb1_STAT5i.SzE_2015_004.D2.txt	TGFb1_STAT5i	SzE_2015_004	D2	4	5	200
TGFb1.SzE_2015_004.D2.txt	TGFb1	SzE_2015_004	D2	4	5	0
No_cytokine.SzE_2015_004.D3.txt	No_cytokine	SzE_2015_004	D3	4	0	0
TGFb1_STAT5i.SzE_2015_004.D3.txt	TGFb1_STAT5i	SzE_2015_004	D3	4	5	200
TGFb1.SzE_2015_004.D3.txt	TGFb1	SzE_2015_004	D3	4	5	0
name	condition	exp	donor	time_point_as_days	TGFb1	STAT5i

Showing 1 to 9 of 9 entries

Metadata Information:

Select Dataset: AS25C_d6_PI_s2.exported.fcs

Dimensions: Cells: 26288, Protein Markers: 12

Parameters (Protein Markers):

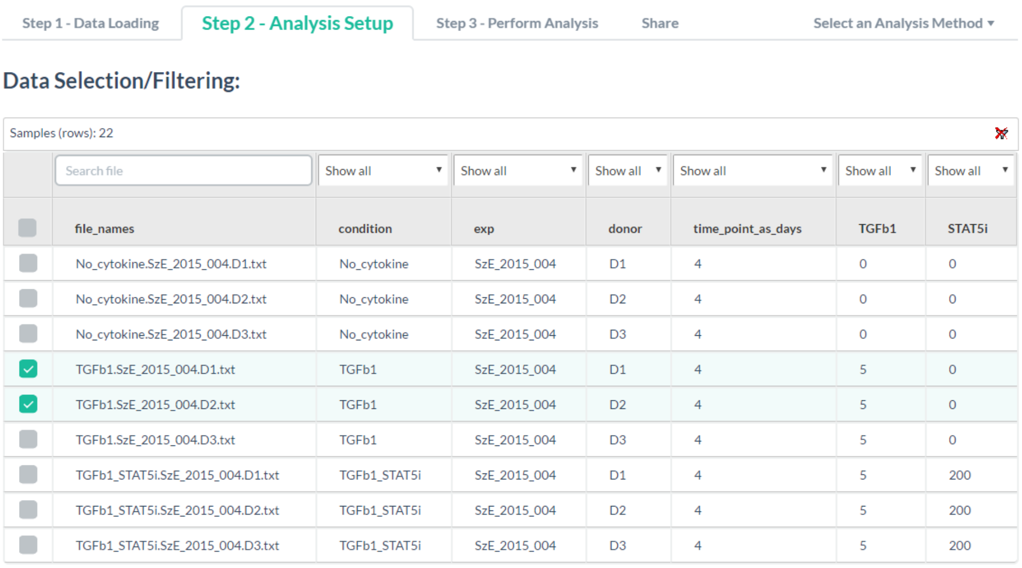
	Marker	Description	Range	minRange	maxRange
1	Pulse Width	Pulse Width	512.00	0.00	511.00
2	FS Lin	FS	65536.00	0.00	65535.00
3	FS Area	FS	65536.00	0.00	65535.00
4	SS Lin	SS	65536.00	0.00	65535.00
5	FL 1 Log	IFNg-FITC	65536.00	1.00	10000.00
6	FL 2 Log	CD25-PE	65536.00	1.00	10000.00

Figure 2.4: Experimental design table and metadata information of the uploaded datasets, as incorporated in the first step of SCENERY's wizard.

2.3 Analysis Setup

In the second step, the users set up the desired *computational experiment*. This essentially involves the selection of (a subset of) the uploaded data files, usually on the basis of factors of the study design, and the application of a single data analysis method. On that, the users can select the relevant, to the current analysis, files one by one or by filtering their selection on the basis of the study design factors (see Fig. 2.5).

The analysis methods included in the current version of SCENERY are subdivided into an, analysis independent, visualization method, standard pre-processing and statistical analysis methods and network reconstruction (NR) algorithms. The, analysis independent, visualization method allows users to visualize their data in terms of histograms or



Data Selection/Filtering:

Samples (rows): 22

□	file_names	condition	exp	donor	time_point_as_days	TGFb1	STAT5i
□	No_cytokine.SzE_2015_004.D1.txt	No_cytokine	SzE_2015_004	D1	4	0	0
□	No_cytokine.SzE_2015_004.D2.txt	No_cytokine	SzE_2015_004	D2	4	0	0
□	No_cytokine.SzE_2015_004.D3.txt	No_cytokine	SzE_2015_004	D3	4	0	0
☑	TGFb1.SzE_2015_004.D1.txt	TGFb1	SzE_2015_004	D1	4	5	0
☑	TGFb1.SzE_2015_004.D2.txt	TGFb1	SzE_2015_004	D2	4	5	0
□	TGFb1.SzE_2015_004.D3.txt	TGFb1	SzE_2015_004	D3	4	5	0
□	TGFb1_STAT5i.SzE_2015_004.D1.txt	TGFb1_STAT5i	SzE_2015_004	D1	4	5	200
□	TGFb1_STAT5i.SzE_2015_004.D2.txt	TGFb1_STAT5i	SzE_2015_004	D2	4	5	200
□	TGFb1_STAT5i.SzE_2015_004.D3.txt	TGFb1_STAT5i	SzE_2015_004	D3	4	5	200

Figure 2.5: Data selection and filtering, in terms of the uploaded ’Experimental Design Table’ as defined in the second step of SCENERY wizard.

scatter-plots. The pre-processing methods allow users to apply data transformations and standard single-cell pre-processing procedures for cytometry files, such as compensation and gating. The statistical analysis methods, currently contain standard functions such as t-test, analysis of variance and regression. Finally, the last analysis methods’ category supports a variety of methods for reconstructing different networks in terms of association, Bayesian and probabilistic causal analysis.

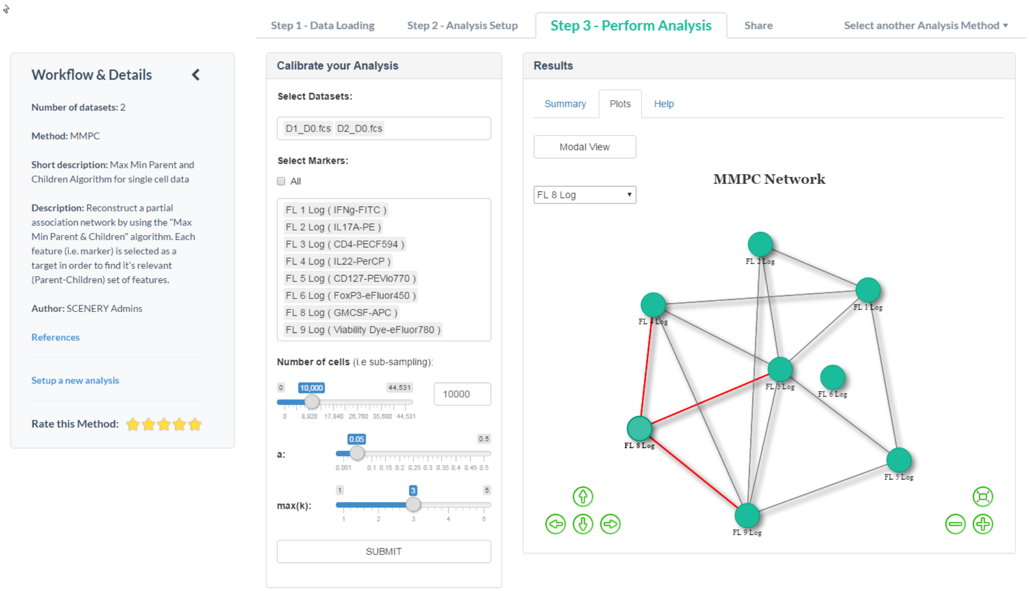
An extensive overview of the analysis methods’ categories along with the current available methods is presented on chapter 3.

2.4 Perform Analysis

Once an analysis method is selected the wizard redirects to the third step where the users will perform the analysis with the selected method and files from the second step. The ’Perform Analysis’ step is structured under a global user interface (UI) layout for each method where the main components are: A ’Workflow & Details’ panel, an ’Analysis Calibration’ panel and a ’Results’ panel, as shown in Fig. 2.6.

In the ’Workflow & Details’ panel, the users are informed for the metadata of the current analysis such as the number of datasets that are involved, the method’s description and characteristics and a rating system for grading the current method and their analysis use-case experience.

In the ’Analysis Calibration’ panel the users are provided with the analysis calibration options. Common options for all methods are deciding which markers and the number of cells to employ for the analysis. Next on this panel, the user defines the method-specific hyper-parameters, such as thresholds for statistical significance, statistical tests, scoring functions, and submits the analysis to the system (see Fig. 2.6).



Workflow & Details

Number of datasets: 2

Method: MMPC

Short description: Max Min Parent and Children Algorithm for single cell data

Description: Reconstruct a partial association network by using the "Max Min Parent & Children" algorithm. Each feature (i.e. marker) is selected as a target in order to find it's relevant (Parent-Children) set of features.

Author: SCENERY Admins

References

Setup a new analysis

Rate this Method: ★★★★★

Calibrate your Analysis

Select Datasets: D1_D0.fcs D2_D0.fcs

Select Markers:

□ All

FL 1 Log (IFNg-FITC)
FL 2 Log (IL17A-PE)
FL 3 Log (CD4-PECF594)
FL 4 Log (IL22-PerCP)
FL 5 Log (CD127-PEVio770)
FL 6 Log (FoxP3-eFluor450)
FL 8 Log (GMCSF-APC)
FL 9 Log (Viability Dye-eFluor780)

Number of cells (i.e sub-sampling): 0 – 10,000 – 44,531; 10000

0 8,920 17,840 26,760 35,680 44,531

a: 0.05 (0.001 – 0.5)

0.001 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5

max(k): 3 (1 – 5)

1 2 3 4 5

Results

Figure 2.6: Typical layout and structure of a method in SCENERY showing the main panels and tabs as incorporated in the third step of the wizard. Here, as a use-case, we run an analysis on flow cytometry data with the MMPC NR analysis method.

The analysis output is presented in the separate ’Results’ panel. This panel consists of two main sections/tabs; namely, ’Summary’ and ’Plots’. The ’Summary’ tab recapitulates the performed analysis reporting metadata information and a textual overview of the results. For example, at each statistical analysis method the appropriate summary statistics are reported and for each NR algorithm a textual representation of the reconstructed network and corresponding method summary results are given.

In the ’Plots’ tab and depending on the selected analysis method used, a separate graphical (downloadable) representation of the results is included. In most of the generated plots in this tab, extra visualization options are available, while a modal semi full-screen view is supported for a better user interaction experience. Regarding the data visualization method, different functionalities are available to the users for exploring their data in terms of histograms for standalone markers or (matrix) scatter-plots with overlapping density contour plots, on demand, for more than one selected markers. Regarding the statistical analysis methods, the results are graphically displayed by a variety of visualizations such as overlapping density plots, violin plots and scatter-plots with fitted regression lines. For the NR analysis, an interactive JavaScript-based, R-wrapped implementation for networks’ visualization (R package ’visNetwork’ [23]) is available (see Fig. 2.6). Finally and as previously said, a variety of options for exporting the generating analysis output figures, are available in multiple formats.

More information on the visualization along with the implementation and the interpretation of each analysis method are described, analytically, in the next chapter.

Chapter 3

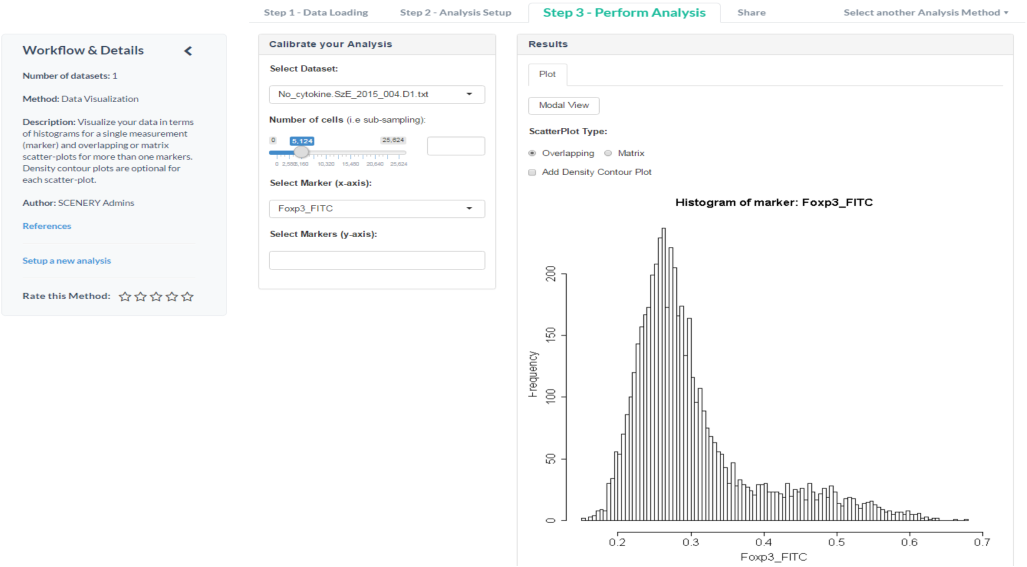
Methods

Analysis methods in SCENERY are, currently, grouped into four major categories: Data visualization, Pre-processing, Univariate statistical analysis and Network reconstruction analysis. As previously said in Section 2.4, all methods are available online under the same UI layout and structure composed by the 'Workflow & Details' panel, the 'Analysis Calibration' panel and the 'Results' panel. All analysis methods are implemented in R [22], while the R Shiny web framework is used (`shiny.rstudio.com`), [24], in order to wrap and transform the R functions into multiple interactive web applications. Below, we present analytically the methods' semantics and visualizations, the algorithms' description and implementation and use-case examples for each category. For reference, these are also summarized in table 3.1 at the end of this chapter.

3.1 Data Visualization

The data visualization category, currently contains a set of methods that allow users to directly visualize their data, even before further applying any analysis method. After discussing and getting feedback by researchers and scientists on the single-cell analysis field, it came up that this is a usual first 'pre-analysis' step, as it offers a quick remark on specific markers of the selected data and an insight on the next steps of the analysis pipeline and work-flow.

As shown in figures Fig. 3.1, Fig. 3.2 and Fig. 3.3, in the 'Analysis Calibration' panel, the users select a dataset and the number of cells involved in the analysis. Then, the users have to select the marker(s) that they want to visualize. If a single marker is selected, its histogram for the current dataset is plotted in the 'Results' panel under the 'Plot' tab (see Fig. 3.1). If two markers are selected, a scatter-plot is plotted instead. If more than two markers are selected, there are two options. Either visualize all scatter-plots, overlapping each other on a single axis by using different colors for each marker (see Fig. 3.2) or as a series of scatter-plots arranged in a matrix format (See Fig. 3.3). In addition, to better illustrate the density of each scatter-plot, overlapping density contour plots are also available, on demand. The 'ggplot2' [25] R package was used for most of the plots in this method while the generic R function 'plot' and the R function 'hist' from the built-in R package ’graphics’ [22] were used for the rest.



Number of datasets: 1

Method: Data Visualization

Description: Visualize your data in terms of histograms for a single measurement (marker) and overlapping or matrix scatter-plots for more than one markers. Density contour plots are optional for each scatter-plot.

Author: SCENERY Admins

References

Select Dataset: No_cytokine.SzE_2015_004.D1.txt

Number of cells (i.e sub-sampling): 0 5,124 25,624

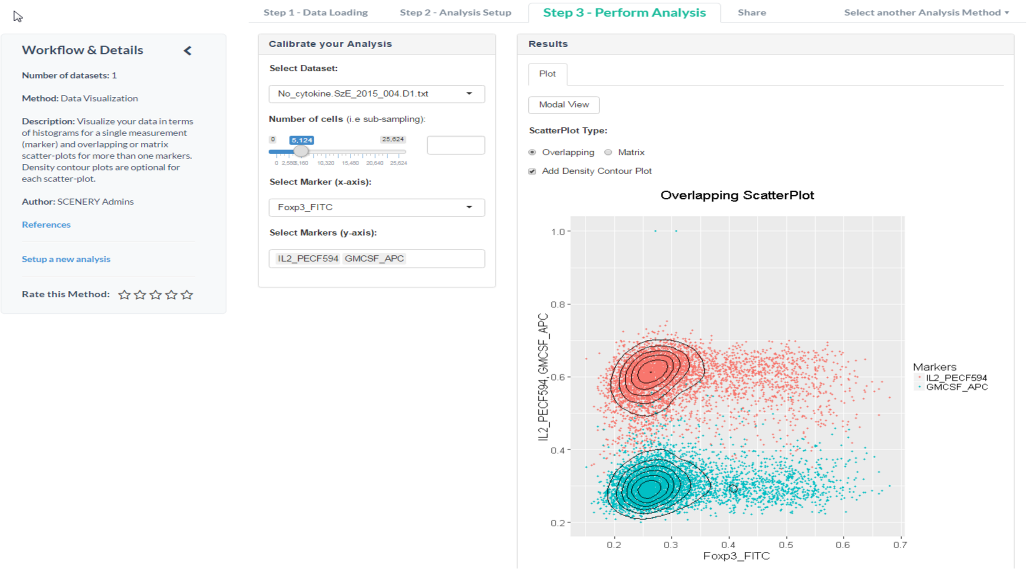
Select Marker (x-axis): Foxp3_FITC

Select Markers (y-axis):

ScatterPlot Type: ● Overlapping ○ Matrix

□ Add Density Contour Plot

Figure 3.1: UI of the ’Data Visualization’ method, showing a histogram of a selected measurement from a mass cytometry file.



Number of datasets: 1

Method: Data Visualization

Description: Visualize your data in terms of histograms for a single measurement (marker) and overlapping or matrix scatter-plots for more than one markers. Density contour plots are optional for each scatter-plot.

Author: SCENERY Admins

References

Select Dataset: No_cytokine.SzE_2015_004.D1.txt

Number of cells (i.e sub-sampling): 0 5,124 25,624

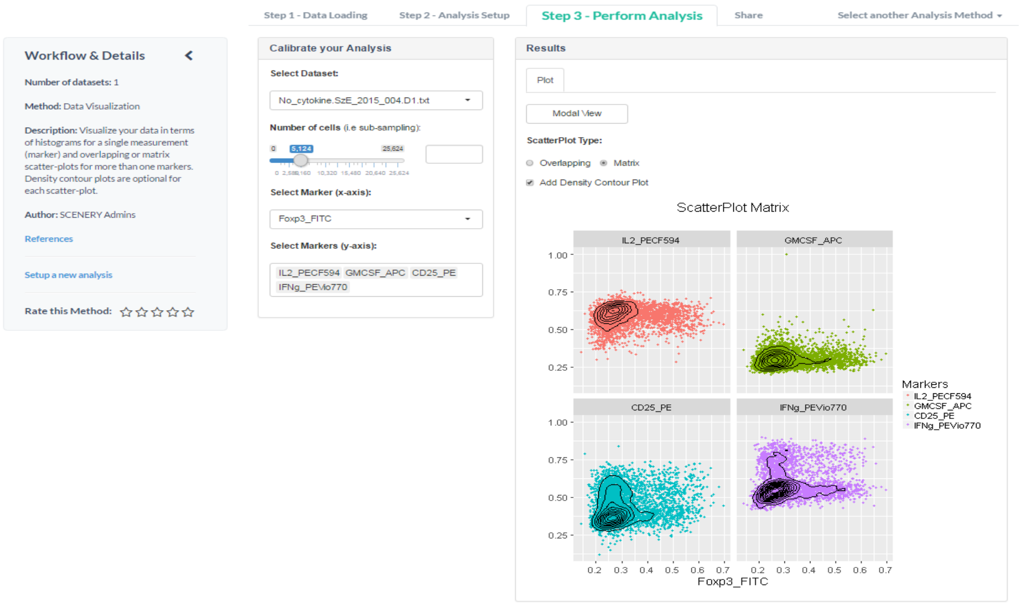
Select Marker (x-axis): Foxp3_FITC

Select Markers (y-axis): IL2_PECF594 GMCSF_APC

ScatterPlot Type: ● Overlapping ○ Matrix

☑ Add Density Contour Plot

Figure 3.2: UI of the ’Data Visualization’ method, showing an overlapping scatter-plot among three selected measurements from a mass cytometry file.



Number of datasets: 1

Method: Data Visualization

Description: Visualize your data in terms of histograms for a single measurement (marker) and overlapping or matrix scatter-plots for more than one markers. Density contour plots are optional for each scatter-plot.

Author: SCENERY Admins

Select Dataset: No_cytokine.SzE_2015_004.D1.txt

Number of cells (i.e sub-sampling): 0 5,124 25,624

Select Marker (x-axis): Foxp3_FITC

Select Markers (y-axis): IL2_PECF594 GMCSF_APC CD25_PE IFNg_PEVio770

ScatterPlot Type: ○ Overlapping ● Matrix

☑ Add Density Contour Plot

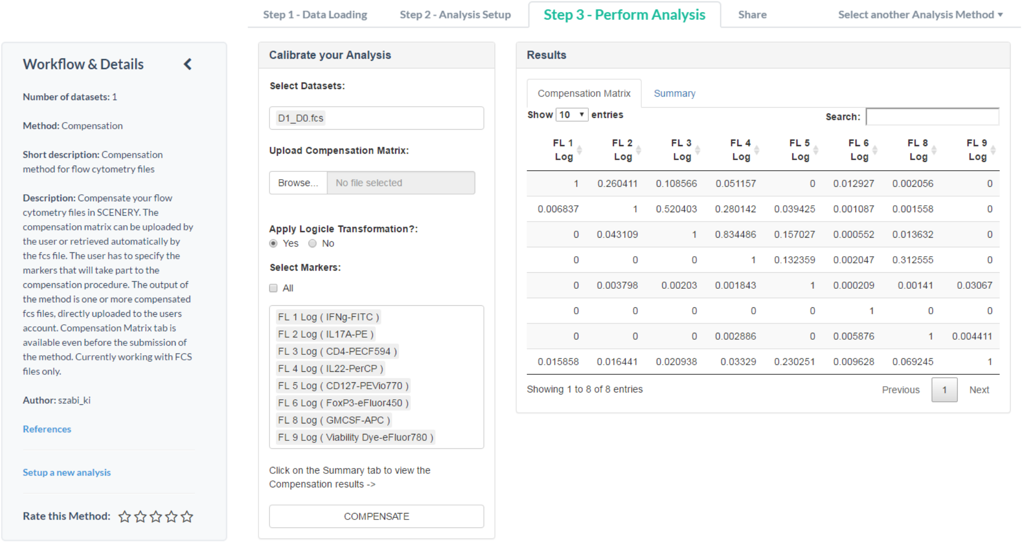
Figure 3.3: UI of the ’Data Visualization’ method, showing a scatter-plot matrix among one selected marker versus four different selected measurements from a mass cytometry file.

3.2 Pre-Processing

By selecting a method from the ’Pre-Processing’ analysis category, SCENERY users are able to apply standard single-cell pre-processing pipelines to their data, in order to better explore the data semantics and to create more appropriate samples for further statistical and probabilistic analysis. The available methods in this category are presented below:

Compensation This method allows SCENERY users to apply the, widely used, compensation procedure to raw, FCS data. This procedure, mainly, corrects the overlap among the spectra of the fluorochromes used as protein markers in flow cytometry files. More specifically, fluorescent probes are used in flow cytometry, in order to detect each marker. The emission spectra of these fluorochromes massively overlap, hence, the detectors of the machine instead of recording information coming from a single marker, they may be recording overlapping information coming from markers on neighboring channels. The main problem is that, obviously, this overlapping phenomenon (referring as spillover) impacts the accuracy and the quality of the initial data. As a solution, the amount of spillover is proved to be a linear function, so the measured average signal levels can be corrected (i.e. aligning population medians) by the compensation process. Finally, with a proper compensation setup, the pre-processed datasets will then properly visualized and analyzed [26].

The implementation that we used in SCENERY is the ’compensate’ R function from the R package 'flowCore' [16], that applies a compensation for spillover between channels by further applying on the data a spillover/compensation matrix, containing single-stained compensation controls, to one or more FCS files, assuming a simple linear combination of values.



Workflow & Details

Number of datasets: 1

Method: Compensation

Short description: Compensation method for flow cytometry files

Description: Compensate your flow cytometry files in SCENERY. The compensation matrix can be uploaded by the user or retrieved automatically by the fcs file. The user has to specify the markers that will take part to the compensation procedure. The output of the method is one or more compensated fcs files, directly uploaded to the users account. Compensation Matrix tab is available even before the submission of the method. Currently working with FCS files only.

Author: szabi_ki

References

Setup a new analysis

Rate this Method: ☆☆☆☆☆

Calibrate your Analysis

Select Datasets: D1_D0.fcs

Upload Compensation Matrix: Browse... No file selected

Apply Logicle Transformation?: ● Yes ○ No

Select Markers:

□ All

FL 1 Log (IFNg-FITC)
FL 2 Log (IL17A-PE)
FL 3 Log (CD4-PECF594)
FL 4 Log (IL22-PerCP)
FL 5 Log (CD127-PEVio770)
FL 6 Log (FoxP3-eFluor450)
FL 8 Log (GMCSF-APC)
FL 9 Log (Viability Dye-eFluor780)

Click on the Summary tab to view the Compensation results ->

COMPENSATE

Results

Compensation Matrix | Summary

Show 10 entries   Search:

FL 1 Log	FL 2 Log	FL 3 Log	FL 4 Log	FL 5 Log	FL 6 Log	FL 8 Log	FL 9 Log
1	0.260411	0.108566	0.051157	0	0.012927	0.002056	0
0.006837	1	0.520403	0.280142	0.039425	0.001087	0.001558	0
0	0.043109	1	0.834486	0.157027	0.000552	0.013632	0
0	0	0	1	0.132359	0.002047	0.312555	0
0	0.003798	0.00203	0.001843	1	0.000209	0.00141	0.03067
0	0	0	0	0	1	0	0
0	0	0	0.002886	0	0.005876	1	0.004411
0.015858	0.016441	0.020938	0.03329	0.230251	0.009628	0.069245	1

Showing 1 to 8 of 8 entries   Previous 1 Next

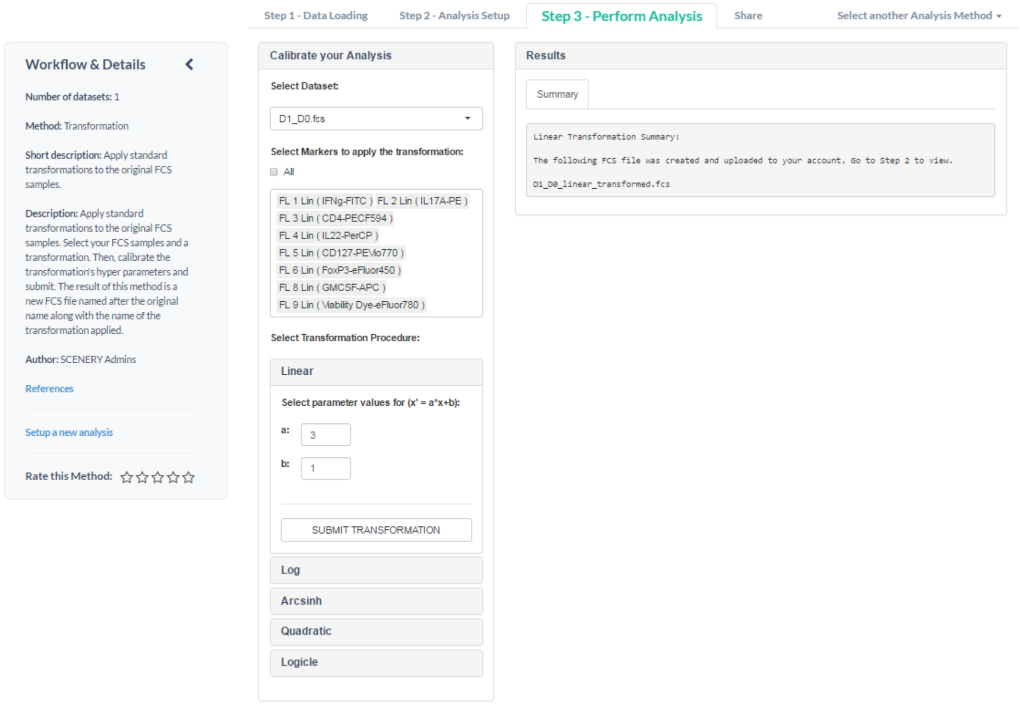
Figure 3.4: UI of the 'Compensation' pre-processing method.

As shown in Fig. 3.4 the users first select the FCS files involved in the compensation procedure (lets define the number of the files as N). Then, the system scans the files for an existing compensation matrix in the competent slot of the FCS file. If a compensation matrix is not found the user can upload a CSV file with the compensation matrix values. The compensation matrix is visualized even before the submission of the method and can be viewed in the 'Compensation Matrix' tab of the 'Results' panel. By going further in the calibration of this procedure, the users select the markers that are referenced and matched one by one with the compensation's matrix columns. An auto pre-selection of the markers is available if the matrix is automated loaded from the FCS file. Finally, a logicle transformation (see the next 'Transformation' method) is available on demand, in order to be applied on the selected markers, as well. The output of this method is N new, compensated, FCS files, directly uploaded to the users account and available for selection in the second step ('Analysis Setup') of the SCENERY wizard.

Transformation This method allows SCENERY users to apply standard transformations to the original data. The users can select multiple files to apply the current transformation automatically to all of them. The available transformations in SCENERY are, currently, the: linear, log, arcsinh, the quadratic and the logicle that are all implemented in the R package 'flowCore' [16] and are listed more specifically below:

- Linear Transformation. Creates a transformation defined by the linear transformation by the function: x’ = a*x + b.
- Log Transformation. Creates a transformation defined by the function:
  x’ = log(x, logbase)*(r/d).
- Arcsinh Transformation. Creates a transformation defined by the function:
  x’ = asinh(a+b*x)+c).
- Quadratic Transformation. Creates a transformation defined by the function:
  x’ = a * x$\hat{2}$ + b*x + c.
- Logicle Transformation. This transformation creates, automatically, a subset of the hyperbolic sine transformation functions that provides several advantages over linear/log transformations for display of flow cytometry data.

For more on these methods and their functions’ hyper parameters, please read the flowCore’s R package reference manual [16].



Number of datasets: 1

Method: Transformation

Short description: Apply standard transformations to the original FCS samples.

Description: Apply standard transformations to the original FCS samples. Select your FCS samples and a transformation. Then, calibrate the transformation's hyper parameters and submit. The result of this method is a new FCS file named after the original name along with the name of the transformation applied.

Author: SCENERY Admins

Select Dataset: D1_D0.fcs

Select Markers to apply the transformation:

□ All

FL 1 Lin (IFNg-FITC) FL 2 Lin (IL17A-PE)
FL 3 Lin (CD4-PECF594)
FL 4 Lin (IL22-PerCP)
FL 5 Lin (CD127-PEVio770)
FL 6 Lin (FoxP3-eFluor450)
FL 8 Lin (GMCSF-APC)
FL 9 Lin (Viability Dye-eFluor780)

Select Transformation Procedure:

Linear

Select parameter values for (x' = a*x+b):

a: 3

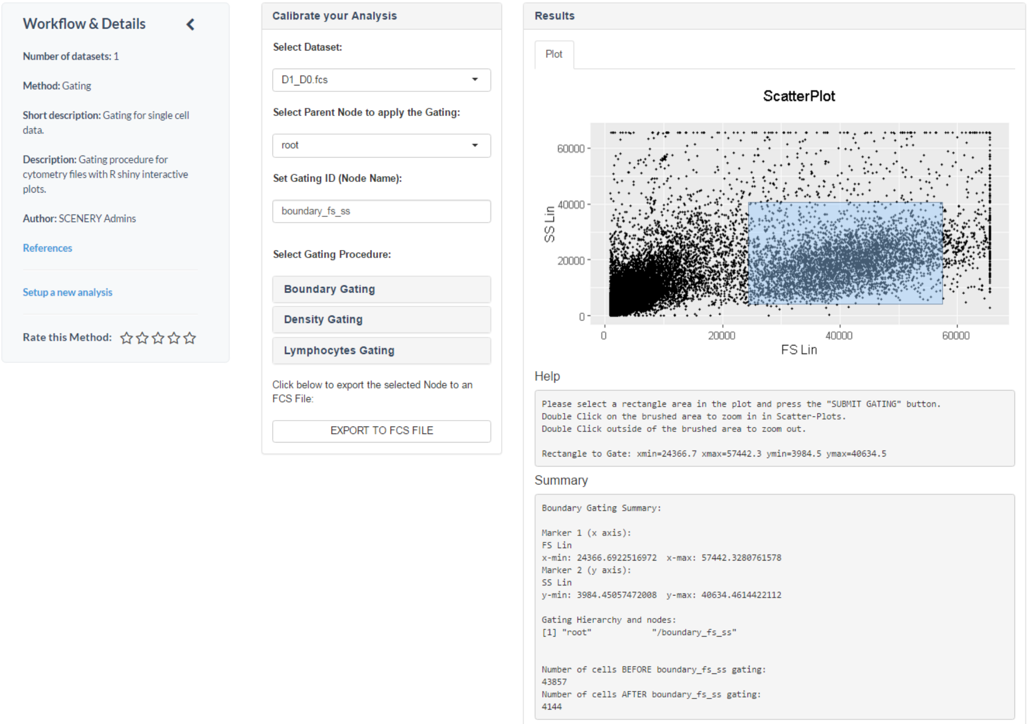
b: 1

```
Linear Transformation Summary:

The following FCS file was created and uploaded to your account. Go to Step 2 to view.

D1_D0_linear_transformed.fcs
```

Figure 3.5: UI of the ’Transformation’ pre-processing method.

Figure 3.5 illustrates this method’s UI. The users select the files involved in the transformation procedure (lets define the number of the files as N) and the selected markers to apply the transformation on. Then they select the transformation method and calibrate the current hyper parameters which are the parameters of each transformation's function. The result of this method, after submitted by the users, is the generation of N new, transformed, FCS files, directly uploaded to the users account and available for selection in the second step ('Analysis Setup') of the SCENERY wizard.

Gating This method allows SCENERY users to apply the, widely used, gating or filtering procedure for cytometry files. Gating in cytometry is the process of isolating groups of cells from the bulk measurement, based on observed cytometric events. Typically, these cell sub-populations (or gates) are manually annotated by drawing the boundaries around a set of data points. This is a hierarchical procedure, usually performed by experts, where bi-axial scatter-plots are sequentially plotted and annotated (see below for further details). It is also possible to define gates, algorithmically, by discrimination analysis (i.e. density based methods) [9]. The generated gates may then be used either for selectively gathering cell sub-populations or for segregating the cell population for further analysis.



Number of datasets: 1

Method: Gating

Short description: Gating for single cell data.

Description: Gating procedure for cytometry files with R shiny interactive plots.

Author: SCENERY Admins

Rate this Method: ☆☆☆☆☆

Select Dataset: D1_D0.fcs

Select Parent Node to apply the Gating: root

Set Gating ID (Node Name): boundary_fs_ss

Select Gating Procedure:

Click below to export the selected Node to an FCS File:

Help

```
Please select a rectangle area in the plot and press the "SUBMIT GATING" button.
Double Click on the brushed area to zoom in in Scatter-Plots.
Double Click outside of the brushed area to zoom out.

Rectangle to Gate: xmin=24366.7 xmax=57442.3 ymin=3984.5 ymax=40634.5
```

Summary

```
Boundary Gating Summary:

Marker 1 (x axis):
FS Lin
x-min: 24366.6922516972  x-max: 57442.3280761578
Marker 2 (y axis):
SS Lin
y-min: 3984.45057472008  y-max: 40634.4614422112

Gating Hierarchy and nodes:
[1] "root"            "/boundary_fs_ss"


Number of cells BEFORE boundary_fs_ss gating:
43857
Number of cells AFTER boundary_fs_ss gating:
4144
```

Figure 3.6: UI of the 'Gating' pre-processing method.

As shown in Fig. 3.6 in the 'Calibrate your Analysis' panel, the users first select the FCS file involved in the gating procedure. Then they select the desired parent node, from the generated hierarchical structure, to apply the gating and give a new name to the node that they are about to create, which will be added, automatically in the referenced hierarchical gating structure. This basically means, that after a series of gating procedures, an hierarchy of cell sub-populations (nodes) will be created and it will be valid to the users for further pre-processing by selecting one of these nodes. After selecting the node, the users are able to select to perform one, none, or several of the gating procedures, currently available in SCENERY, that is the Boundary gating, the Density gating and the Lymphocytes gating.

At the Boundary gating procedure, the users select two markers, and set the boundaries, interactively, in the scatter-plot by drawing a rectangle area. This results to the removal of the events (cells) outside of the boundaries.

At the Density gating procedure, the users select one marker and they set the gate, interactively, in the density plot by selecting the minimum and the maximum x in the x-axis. This, usually, applies to procedures such as small event's filtering or filtering of stimulation beads (debris) and cell doublets. The small event's filtering is based on a density plot of the FS-lin (Forward Scatter Linear) channel, which is proportional to cell size. Moreover, a filtering based on a density plot of the Pulse Width marker, will remove events that are probably debris (these events are very big on pulse width) or cell doublets (events very small on pulse width).

At the Lymphocytes gating the users select the Forward Scatter Linear (FS Lin) channel and the Side Scatter Linear (SS Lin) channel from the current data, and submit the gating. An automatic density-based function of elliptical shaped cell sub-populations, called 'flowGate' from the R package 'flowStats' [17], is applied while a scatter-plot visualization along with a summary of the results are available.

All interactive plots implementations are a combination of functions from the R packages 'ggplot2', graphics and shiny [25], [24]. The internal R functions involved in the Boundary and the Density gating procedures are implemented in The R package 'flowCore' [16] (i.e. the 'rectangleGate' function used in the 'Boundary Gating' procedure). See the mentioned R packages' reference manuals for more.

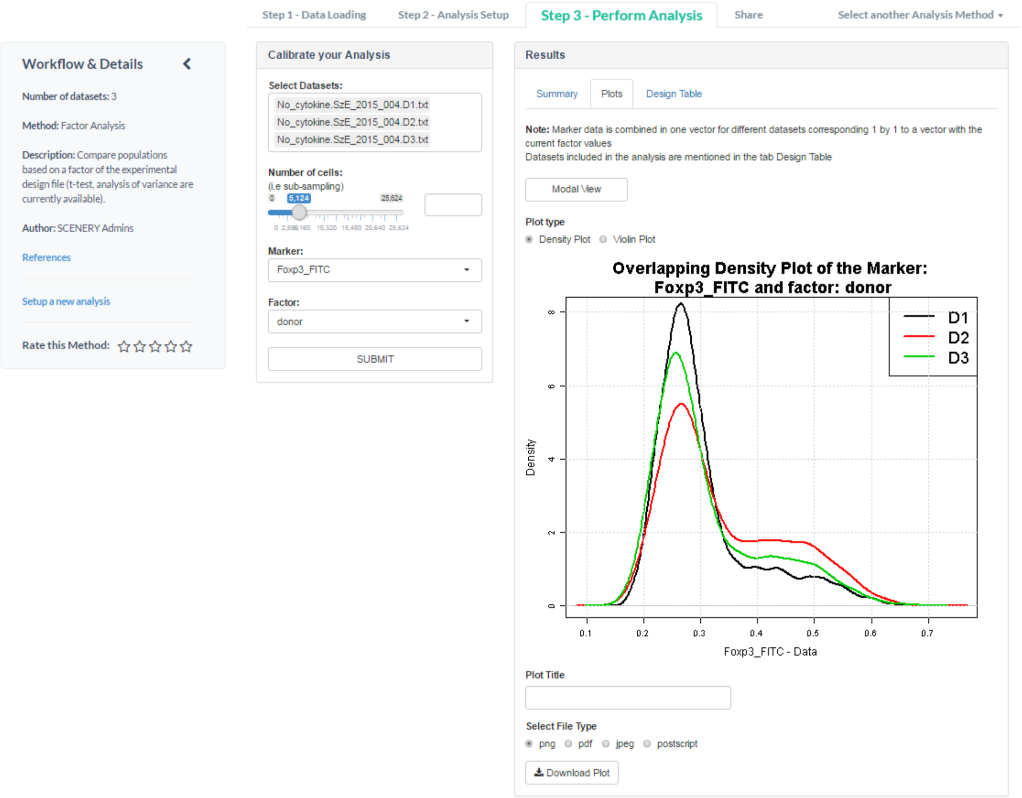
Finally, the users can select multiple sub-populations or nodes in the created gating hierarchy and export them as new FCS files, which are directly uploaded to the users account and are available for selection and further analysis in the second step ('Analysis Setup') of SCENERY's wizard.

3.3 Univariate Statistical Analysis

The univariate statistical analysis category contains methods that allow users to apply basic and state of the art statistical analysis methods, in order to model the relationship among standalone markers and study design factors, across different samples of the same study. A study design factor is selected with respect to the uploaded experimental design, as discussed in Section 2.2, and it takes a unique value for each FCS file (cell sub-population) involved in the analysis.

Factor Analysis The Factor Analysis method in SCENERY includes a t-test and an analysis of variance for deriving a population comparison across different levels of the same factor for standalone markers.

As shown in Fig. 3.7, the users have to select two or more FCS files that are under the same experimental design. Then, the users select one of the available markers and one factor, with respect to the experimental design table that is available in the 'Design Table' tab of the 'Results' panel, and submit the analysis.



Number of datasets: 3

Method: Factor Analysis

Description: Compare populations based on a factor of the experimental design file (t-test, analysis of variance are currently available).

Author: SCENERY Admins

Select Datasets:
No_cytokine.SzE_2015_004.D1.txt
No_cytokine.SzE_2015_004.D2.txt
No_cytokine.SzE_2015_004.D3.txt

Number of cells:
(i.e sub-sampling)

Marker: Foxp3_FITC

Factor: donor

Note: Marker data is combined in one vector for different datasets corresponding 1 by 1 to a vector with the current factor values
Datasets included in the analysis are mentioned in the tab Design Table

Plot type
● Density Plot ○ Violin Plot

Plot Title

Select File Type
● png ○ pdf ○ jpeg ○ postscript

Figure 3.7: UI of the 'Factor Analysis' statistical analysis method.

If the corresponding unique values of the selected factor are less than two, the population comparison can not be performed. For exactly two unique values, the Welch's two Sample t-test statistic is used [27], [28] for comparing the two different populations, while for more than two unique values of the factor, an analysis of variance [29] is performed to the corresponding populations.

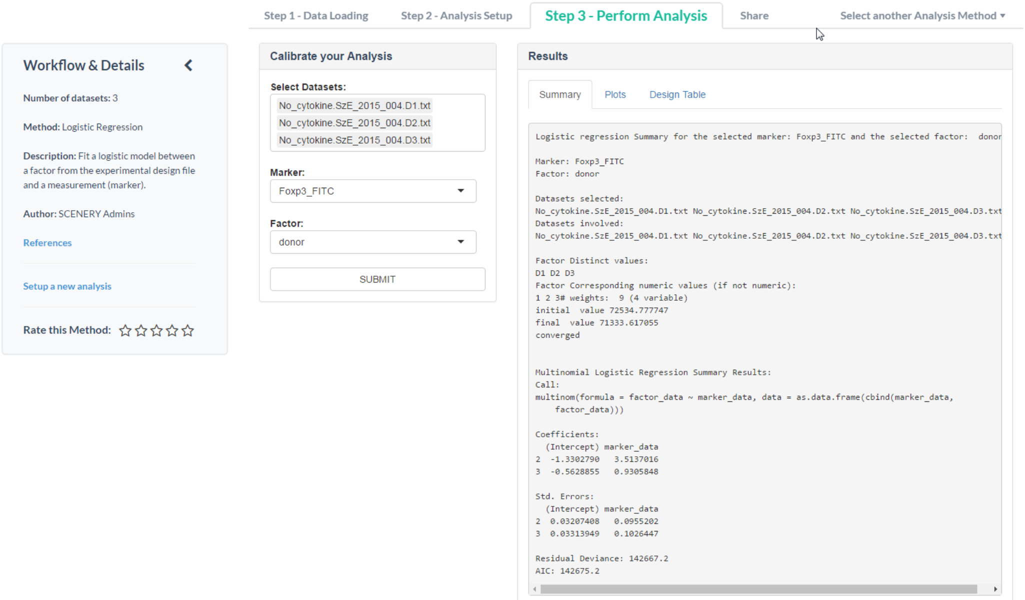
After the submission of the analysis, a textual representation of the results is available in the 'Summary' tab, which contains meta-data information about the analysis such as the involved datasets, the unique values of the selected factor and the total number of cells, along with a summary of the comparison results in terms of statistics. The 'Plots' tab of the 'Results' panel consists by an overlapping density plot of the selected marker's subpopulations retrieved with respect to the factor values and a violin plot which is similar to a box plot with a rotated kernel density plot on each side for showing the probability density of the measurement (i.e. the select marker) across the different values of the factor.

As previously said, all plots are, publication-quality figures, downloadable in multiple formats such as PNG, PDF, JPG and postscript.

For the implementation of this method, the built-in R function ’t.test’ (R package stats) was used for applying the Welch’s t-test statistic, while the built-in R function ’anova’ (R package stats) was used for the analysis of variance. For the density plot, the generic plot function of the R package graphics was used, while for the violin plot, the R function ’qplot’ of the R package ’ggplot2’ [25] was used.

Linear & Logistic Regression Univariate linear and logistic regression methods are also available in SCENERY for modeling the relationship between markers and study design’s factors, by fitting a linear model in case of a numeric factor or a logistic model in case of a categorical factor.

As shown in Fig. 3.8, the users have to select two or more FCS files that are under the same experimental design. Then, the users select one of the available markers and one factor, with respect to the experimental design table that is available in the ’Design Table’ tab of the ’Results’ panel, and submit the analysis.



Number of datasets: 3

Method: Logistic Regression

Description: Fit a logistic model between a factor from the experimental design file and a measurement (marker).

Author: SCENERY Admins

Rate this Method: ☆☆☆☆☆

Select Datasets:
No_cytokine.SzE_2015_004.D1.txt
No_cytokine.SzE_2015_004.D2.txt
No_cytokine.SzE_2015_004.D3.txt

Marker: Foxp3_FITC

Factor: donor

```
Logistic regression Summary for the selected marker: Foxp3_FITC and the selected factor:  donor

Marker: Foxp3_FITC
Factor: donor

Datasets selected:
No_cytokine.SzE_2015_004.D1.txt No_cytokine.SzE_2015_004.D2.txt No_cytokine.SzE_2015_004.D3.txt
Datasets involved:
No_cytokine.SzE_2015_004.D1.txt No_cytokine.SzE_2015_004.D2.txt No_cytokine.SzE_2015_004.D3.txt

Factor Distinct values:
D1 D2 D3
Factor Corresponding numeric values (if not numeric):
1 2 3# weights:  9 (4 variable)
initial  value 72534.777747
final  value 71333.617055
converged


Multinomial Logistic Regression Summary Results:
Call:
multinom(formula = factor_data ~ marker_data, data = as.data.frame(cbind(marker_data,
    factor_data)))

Coefficients:
  (Intercept) marker_data
2  -1.3302790   3.5137016
3  -0.5628855   0.9305848

Std. Errors:
  (Intercept) marker_data
2  0.03207408   0.0955202
3  0.03313949   0.1026447

Residual Deviance: 142667.2
AIC: 142675.2
```

Figure 3.8: UI of the ’Logistic Regression’ statistical analysis method.

If a numeric factor with more than two unique values is selected, the method fits a linear model among the different marker sub-populations and a summary of the generated results is available on the ’Summary’ tab of the ’Results’ panel. This summary contains meta-data information about the analysis and results of the fitted model in terms of the residuals, the coefficients and more.

If a categorical factor with more than two unique values, is selected, the method fits a logistic model instead. For exactly two unique values of the factor, a binomial logistic regression is fitted, while for more than two unique values, a multinomial logistic regression is fitted. A summary of the generated results is then available on the 'Summary' tab of the 'Results' panel. This summary contains meta-data information about the analysis and results of the fitted model in terms of the deviance residuals, the coefficients, the Std. errors and more.

The 'Plots' tab of the 'Results' panel consists by a scatter-plot of the marker values across the different factor values, with a fitted regression line for a visual interpretation of the relationship. As previously said, all plots are publication-quality figures, downloadable in multiple formats such as PNG, PDF, JPG and postscript.

For the implementation of this method, the built-in R function 'lm' (R package stats) was used for fitting the linear regression, while the R functions 'glm' (R package stats) and 'multinom' (R package nnet [30]) were used for the binomial and the multinomial logistic regressions, respectively. For the scatter-plot with the fitted regression line, the R function 'ggplot' of the R package 'ggplot2' [25] was used.

3.4 Network Reconstruction Analysis

This analysis category imparts a ton of innovation in this work, as SCENERY is the first free software to support a number of NR algorithms for single-cell data. In addition, the name of our system was inspired by this main functionality (SCENERY: a Single CEll NEtwork Reconstruction sYstem). All NR methods represent statistical relationships in the data as networks composed by nodes and edges. Nodes always stand for measurements (e.g., protein abundances). Edges, on the contrary, have different semantics, depending on the type of network that the method outputs.

A typical analysis with a method from this category consists of the calibration of the analysis in the 'Calibrate your Analysis' panel, a textual representation of the results (i.e. a summary of the graph in text form) along with the analysis meta-data information (Summary) which are the involved in the analysis datasets, the total number of cells and the hyper parameters values. Regarding the visual results, an interactive visualization of the reconstructed network (Plot) in the 'Results' panel is available. Each network visualization is generated by using the R package 'visNetwork' [23] with the default 'Force-directed' layout [31]. The purpose of this layout is to position the nodes of a network so that all the edges have more or less equal length and there are as few crossing edges as possible. The visualized networks' interactivity, currently, relies on features such as nodes re-positioning, zooming, and a capability of locating a network node along with its direct neighbors. In addition, every generated reconstructed network can be download as a GEXF (Graph Exchange XML Format) file. GEXF is a standard, widely used, graph-representation format and GEXF files are, usually, used as input in state-of-the-art graph visualization tools such as Gephi [32].

Moreover, a standard graph theory analysis is available in SCENERY, below every reconstructed network visualization, as shown in Fig. 3.9. This analysis, contains the basic graph characteristics from standard metrics and algorithms in the graph theory field, for further interpretation and analysis of each generated reconstructed network in aspects such as connectivity, topology, ranking of nodes given a specific metric, cliques identification and more. This functionality of SCENERY, mainly, relies on algorithms for deriving each network's density, average nodes' degree, diameter, average shortest path length, the betweenness centrality and the clustering coefficient of each node along with the average clustering coefficient of the network. More specifically:

- Density: Ratio of the current number of edges and the number of possible edges in the graph. A graph can be interpreted as dense if the number of its edges are close to the maximal number of edges (usually when density >= 0.5). So, this metric shows, intuitively, how connected is the generated network. If a network reach the maximal density, which is the value one (1), it is characterized as a complete graph where every node is connected with every other node. Finally, the density of a graph can be interpreted as an intuitive indicator of the network's connectivity.

- Average Degree: The degree of a node is the number of edges that are connected to that node. The average degree of a graph is the average number of the degree of all nodes in the current network. For directed graphs, the In-degree and the Out-degree can also defined and the average quantity of them, respectively. In-degree of a node is the number of the incoming edges to that node, while Out-degree is the number of the outgoing edges from this node. The sum of the average In-degree and the average Out-degree is the average degree of a directed graph.

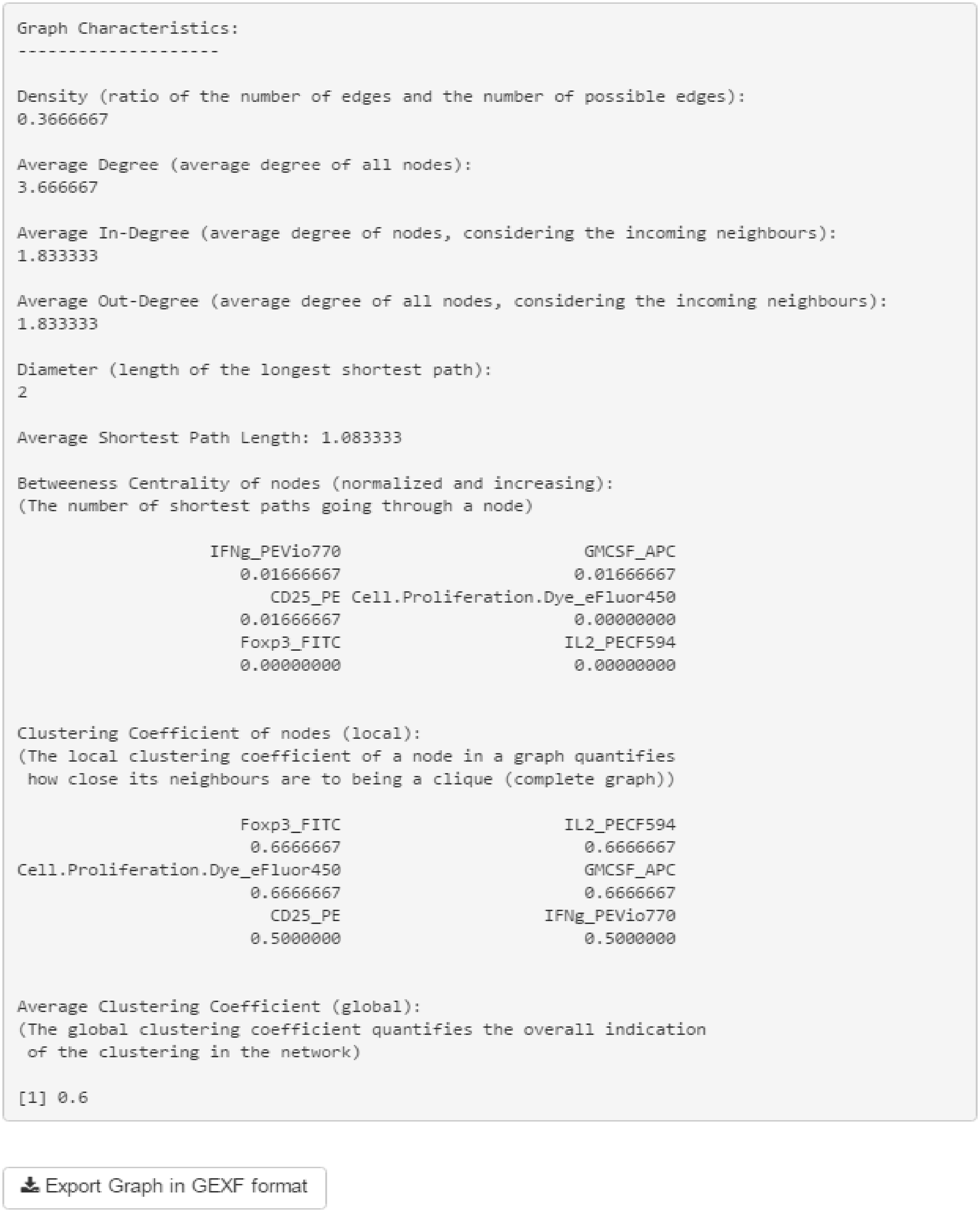
- Average Shortest Path Length: The shortest path of two nodes is defined as a path, among this nodes, such that the number of the path's edges (or the sum of the path's edges' weights) is minimized. The average shortest path length in a graph, is defined as the average number of the shortest paths' length for all possible pairs of the networks' nodes.

- Diameter: The longest shortest path in the network. It can be also interpreted as an intuitive indicator of the network's connectivity.

- Betweenness Centrality: This algorithm is an indicator of each node's centrality in the network. Betweeness centrality of a node is defined as the ratio of the number of the shortest paths, from all pairs of nodes, that pass through that node, divided by the number of all possible shortest paths. On that, we can say that a node with high betweeness centrality has a large influence on the generated network's paths, under the assumption that a transferring quantity follows the shortest paths.

- Clustering Coefficient: Measures the degree to which nodes in a network tend to cluster together. The local clustering coefficient of a node in a network quantifies how close its neighbors are to being a clique (complete graph).

- Average Clustering Coefficient: The global clustering coefficient (or transitivity) gives an indication of the clustering in the whole network. Both global and local clustering coefficient measures can be interpreted as intuitive indicators of the network's connectivity.



```
Graph Characteristics:
--------------------

Density (ratio of the number of edges and the number of possible edges):
0.3666667

Average Degree (average degree of all nodes):
3.666667

Average In-Degree (average degree of nodes, considering the incoming neighbours):
1.833333

Average Out-Degree (average degree of all nodes, considering the incoming neighbours):
1.833333

Diameter (length of the longest shortest path):
2

Average Shortest Path Length: 1.083333

Betweeness Centrality of nodes (normalized and increasing):
(The number of shortest paths going through a node)

                   IFNg_PEVio770                       GMCSF_APC
                      0.01666667                      0.01666667
                         CD25_PE Cell.Proliferation.Dye_eFluor450
                      0.01666667                      0.00000000
                      Foxp3_FITC                     IL2_PECF594
                      0.00000000                      0.00000000


Clustering Coefficient of nodes (local):
(The local clustering coefficient of a node in a graph quantifies
 how close its neighbours are to being a clique (complete graph))

                      Foxp3_FITC                     IL2_PECF594
                       0.6666667                       0.6666667
Cell.Proliferation.Dye_eFluor450                       GMCSF_APC
                       0.6666667                       0.6666667
                         CD25_PE                   IFNg_PEVio770
                       0.5000000                       0.5000000


Average Clustering Coefficient (global):
(The global clustering coefficient quantifies the overall indication
 of the clustering in the network)

[1] 0.6
```

Figure 3.9: Summary of the typical network's characteristics by applying several graph theory analysis algorithms in a reconstructed association network, derived by the 'Correlation' NR method.

For the implementation of these algorithms, the R package 'igraph' [33] was used and more specifically the R functions 'edge_density' and 'degree' were used for deriving the network's density and average degree, while the 'mean_distance' and the 'diameter' R functions, based on an unweighted breadth first search (BFS) [34], were used for generating the average shortest path length and the diameter (longest shortest path), respectively. The R function 'betweenness' was used for retrieving each node's normalized betweenness centrality by using the 'Brandes' algorithm [35]. Finally, the R function 'transitivity' was used for deriving the local clustering coefficient of each node, along with the global clustering coefficient of the network, assuming that for directed graphs the direction of the edges is ignored [36]. The results in the betweenness centrality and the clustering coefficient of the network nodes are presented decreasingly sorted. For more details on these methods, read the 'igraph' R package reference manual [33].

Finally and by returning to each NR method's analysis overview, a description of the algorithm used in each NR method along with a 'help' section are available for ease of usage.

Based on the edges' semantics criterion, we distinguished the available NR methods into the following sub-categories:

3.4.1 (Conditional) Association Networks

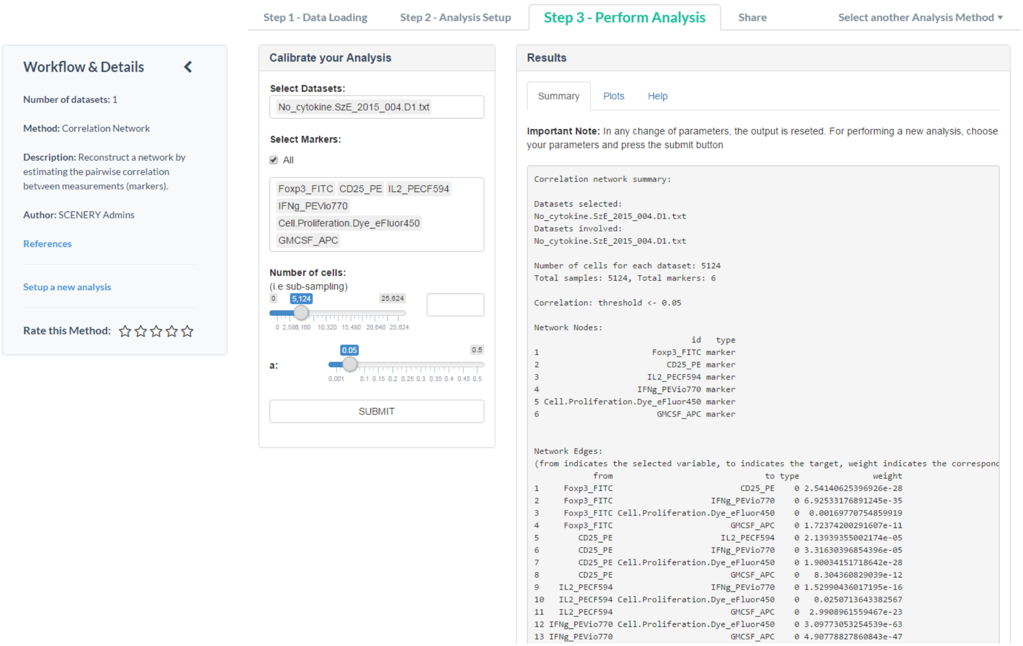
Association Networks (AN) connect two nodes with an undirected edge if the corresponding measurements are found statistically associated. Conditional Association Network (CAN) are similar to AN, but associations between nodes are computed conditioning on all (or part) of the remaining measurements. An example of such a network is shown in the 'Results' panel at the right of the Fig. 3.11.

Correlation The Correlation method, which is used for deriving ANs, is based on the Pearson's pairwise correlation coefficient between different paired measurements [37], [38]. The method's main loop tests each marker with all the rest for any statistical univariate association. Each association among two markers is represented in the network as an undirected edge.

As shown in Fig. 3.10, the users select one or more FCS files, which are pooling all together and the, involved in the analysis, markers. Then an appropriate input for taking a random sub-population (sub-sampling) is available. Finally, the users calibrate the threshold for testing the statistical significance of the generated p-value of each correlation test, and submit the analysis.

For the implementation of this method, the built-in R function 'cor.test' (R package stats) was used for generating each association's p-value.

MMPC MMPC (Max Min Parent and Children [39], [40]) employs the theoretical foundation of causal discovery to perform feature selection for a target variable, by identifying the neighborhood of the target variable (parents and children) in the graph with respect to a conditioning set. By limiting the area of interest to a single node, MMPC manages to efficiently selects the signature for the desired target even among thousands of input variables. MMPC has proven to be one of the most robust and efficient feature selection algorithms [41].



Workflow & Details

Number of datasets: 1

Method: Correlation Network

Description: Reconstruct a network by estimating the pairwise correlation between measurements (markers).

Author: SCENERY Admins

References

Setup a new analysis

Calibrate your Analysis

Select Datasets: No_cytokine.SzE_2015_004.D1.txt

Select Markers:

☑ All

Foxp3_FITC CD25_PE IL2_PECF594 IFNg_PEVio770 Cell.Proliferation.Dye_eFluor450 GMCSF_APC

Number of cells:
(i.e sub-sampling)

a:

Results

Important Note: In any change of parameters, the output is reseted. For performing a new analysis, choose your parameters and press the submit button

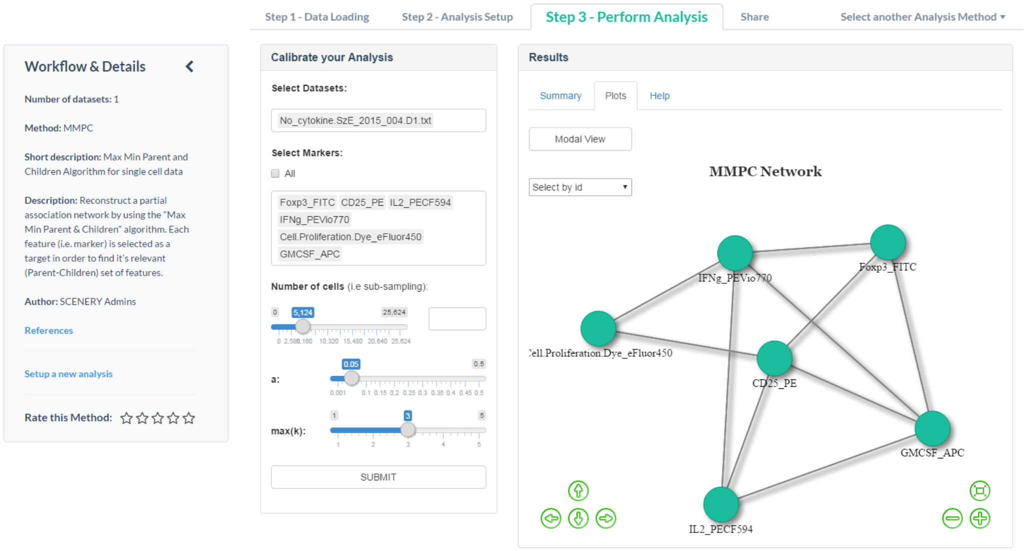
```
Correlation network summary:

Datasets selected:
No_cytokine.SzE_2015_004.D1.txt
Datasets involved:
No_cytokine.SzE_2015_004.D1.txt

Number of cells for each dataset: 5124
Total samples: 5124, Total markers: 6

Correlation: threshold <- 0.05

Network Nodes:
                                id   type
1                       Foxp3_FITC marker
2                          CD25_PE marker
3                      IL2_PECF594 marker
4                    IFNg_PEVio770 marker
5 Cell.Proliferation.Dye_eFluor450 marker
6                        GMCSF_APC marker


Network Edges:
(from indicates the selected variable, to indicates the target, weight indicates the correspond
            from                               to type               weight
1     Foxp3_FITC                          CD25_PE    0 2.54140625396926e-28
2     Foxp3_FITC                    IFNg_PEVio770    0 6.92533176891245e-35
3     Foxp3_FITC Cell.Proliferation.Dye_eFluor450    0  0.00169770754859919
4     Foxp3_FITC                        GMCSF_APC    0 1.72374200291607e-11
5        CD25_PE                      IL2_PECF594    0 2.13939355002174e-05
6        CD25_PE                    IFNg_PEVio770    0 3.31630396854396e-05
7        CD25_PE Cell.Proliferation.Dye_eFluor450    0 1.90034151718642e-28
8        CD25_PE                        GMCSF_APC    0   8.304360829039e-12
9    IL2_PECF594                    IFNg_PEVio770    0 1.52990436017195e-16
10   IL2_PECF594 Cell.Proliferation.Dye_eFluor450    0   0.0250713643382567
11   IL2_PECF594                        GMCSF_APC    0  2.9908961559467e-23
12 IFNg_PEVio770 Cell.Proliferation.Dye_eFluor450    0 3.09773053254539e-63
13 IFNg_PEVio770                        GMCSF_APC    0 4.90778827860843e-47
```

Figure 3.10: UI of the 'Correlation' NR analysis method, showing a textual representation of the results in the 'Summary' tab, after submitting an analysis.

The MMPC method in SCENERY is used for deriving CANs. The incorporated method sets each marker as a target variable in order to find it's associated parent and children set of features. Then and in order to derive the conditional association network, the method generates an undirected edge between two nodes A and B, if the corresponding measurement of A was found in the parent and children set of B and vice versa, with respect to a conditioning set. The default statistic test, currently used, in the MMPC method for testing the conditional independence among the measurements is the Fisher's Z test [42] for continuous measurements, while for future releases of SCENERY, a robust version of this test will be available [44], [45]. Moreover, future versions will contain the Spearman's correlation-based test statistic [43], as well, since measurements may not exhibit a linear relationship and the Spearman's correlation measures how well the relationship between two continuous measurements can be fit by a monotonic function [6].

As shown in Fig. 3.11, the users select one or more FCS files, which are pooling all together and the, involved in the analysis, markers. Then an appropriate input for taking a random sub-population (sub-sampling) is available. Finally, the users calibrate the MMPC hyper parameters from a range of available values. These parameters consist of a threshold for testing the conditional independence among a measurement and the target measurement given a conditioning set CS. The second hyper parameter is the max_k which is the maximum length of the conditioning set.

For the implementation of this method, the R function 'MMPC' of the R package 'MXM' [40] was used, while the R function 'testIndFisher' from the same package, was used internally for testing the conditional independence with the Fisher's Z statistic test. For more, check the MXM package's reference manual.



Step 1 - Data Loading | Step 2 - Analysis Setup | Step 3 - Perform Analysis | Share | Select another Analysis Method ▾

Workflow & Details

Number of datasets: 1

Method: MMPC

Short description: Max Min Parent and Children Algorithm for single cell data

Description: Reconstruct a partial association network by using the "Max Min Parent & Children" algorithm. Each feature (i.e. marker) is selected as a target in order to find it's relevant (Parent-Children) set of features.

Author: SCENERY Admins

References

Setup a new analysis

Rate this Method: ☆☆☆☆☆

Calibrate your Analysis

Select Datasets: No_cytokine.SzE_2015_004.D1.txt

Select Markers:

□ All

Foxp3_FITC CD25_PE IL2_PECF594 IFNg_PEVio770 Cell.Proliferation.Dye_eFluor450 GMCSF_APC

Number of cells (i.e sub-sampling): 5,124

a: 0.05

max(k): 3

SUBMIT

Results

Summary | Plots | Help

Modal View

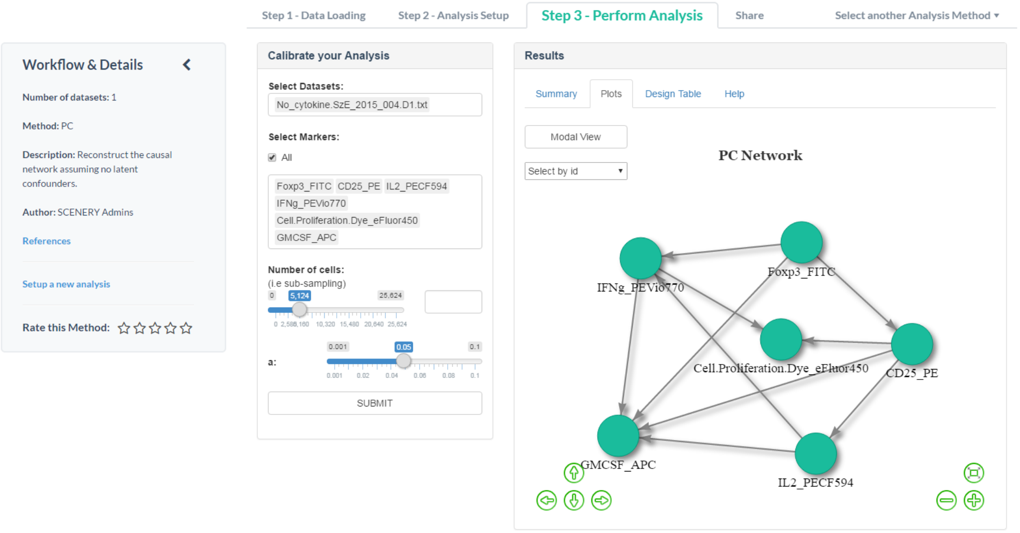
Select by id ▾

Figure 3.11: UI of the 'MMPC' NR analysis method, showing the generated reconstructed network in the 'Plots' tab, after submitting an analysis.

3.4.2 Probabilistic Causal Networks

Therefore, some algorithms output Partial DAGs (PDAGs), that use directed edges for representing causal relations, and undirected edges to represent edges whose causal direction is unclear. If hidden common confounders are also a possibility, Maximal Ancestral Graphs (MAGs) are typically used instead of Bayesian networks (BN). MAGs use directed edges to represent causal relationships, and bi-directed edges to represent confounded relationships. Again, since some causal directions are not identifiable, the algorithms usually output Partial Ancestral Graphs (PAGs) that use circle endpoints to indicate ambiguous orientations [46].

PC PC ([47]) is a landmark constraint-based algorithm, named after its inventors Peter Spirtes and Clark Glymour [48], that remains one of the most popular Bayesian-based NR algorithms. PC outputs a representative of all Bayesian networks that satisfy the conditional independencies that hold in the input dataset, under the standard causal discovery assumptions (Causal Markov Condition, Faithfulness) and the absence of latent confounders. The directed edges in the output graph can be interpreted as causal links but the direction of some edges may be undetermined, in the sense that they point one way in one directed acyclic graph (DAG) in the equivalence class, while they point the other way in another DAG in the equivalence class. So, the reconstructed network by this method can be uniquely represented by a partially directed acyclic graph (PDAG) that contains undirected and directed edges.



Workflow & Details

Number of datasets: 1

Method: PC

Description: Reconstruct the causal network assuming no latent confounders.

Author: SCENERY Admins

Calibrate your Analysis

Select Datasets: No_cytokine.SzE_2015_004.D1.txt

Select Markers: ☑ All

Foxp3_FITC CD25_PE IL2_PECF594 IFNg_PEVio770 Cell.Proliferation.Dye_eFluor450 GMCSF_APC

Number of cells: (i.e sub-sampling) 5,124

a: 0.05

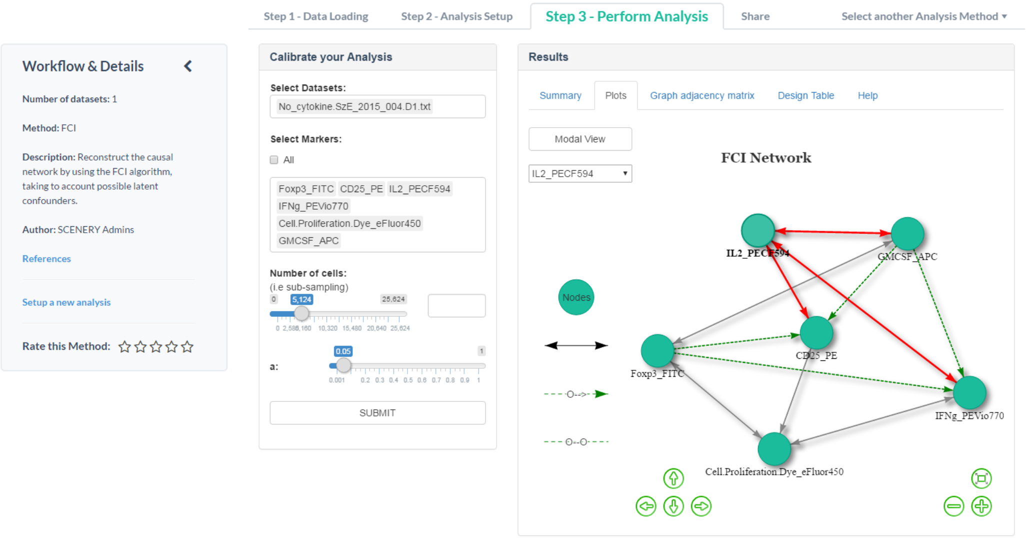
Figure 3.12: UI of the ’PC’ NR analysis method, showing the generated reconstructed network in the ’Plots’ tab, after submitting an analysis.

As shown in Fig. 3.12 and in consistency with all the NR methods in SCENERY, the users select one or more FCS files, which are pooling all together and the, involved in the analysis, markers. Then an appropriate input for taking a random sub-population (sub-sampling) is available as well. Finally, the users calibrate the significance level threshold for testing the generated p-value of each of the individual conditional independence tests, and submit the analysis.

The implementation of this method was based on the R function ’pc’ of the R package ’pcalg’ [47], while the default Gaussian based conditional independence test was used. For more, check the ’pcalg’ package’s reference manual.

FCI FCI ([48], [49], [50]), Fast Causal Inference, is one of the first algorithms that can be used to produce causal networks in the presence of latent confounders. The output graph summarizes all pairwise relationships. Each pair of variables may be connected by a causal relationship (directed edge) or be confounded by a hidden common cause (bi-directed edge). Endpoints of the edges that cannot be uniquely identified (i.e. are ambiguous in different models that fit the data equally well) are denoted by circles and the reconstructed network can be represented by a maximal ancestral graph (MAG) or by a partial ancestral graph (PAG).

As shown in Fig. 3.13 and in consistency with all the NR methods in SCENERY, the users select one or more FCS files, which are pooling all together and the, involved in the analysis, markers. Then an appropriate input for taking a random sub-population (sub-sampling) is available as well. Finally, the users calibrate the significance level threshold for testing the generated p-value of each of the individual conditional independence tests, and submit the analysis.



Number of datasets: 1

Method: FCI

Description: Reconstruct the causal network by using the FCI algorithm, taking to account possible latent confounders.

Author: SCENERY Admins

Select Datasets: No_cytokine.SzE_2015_004.D1.txt

Select Markers:

□ All

Foxp3_FITC CD25_PE IL2_PECF594 IFNg_PEVio770 Cell.Proliferation.Dye_eFluor450 GMCSF_APC

Number of cells: (i.e sub-sampling) 5,124

a: 0.05

IL2_PECF594 ▾

Figure 3.13: UI of the 'FCI' NR analysis method, showing the generated reconstructed network in the 'Plots' tab, after submitting an analysis.

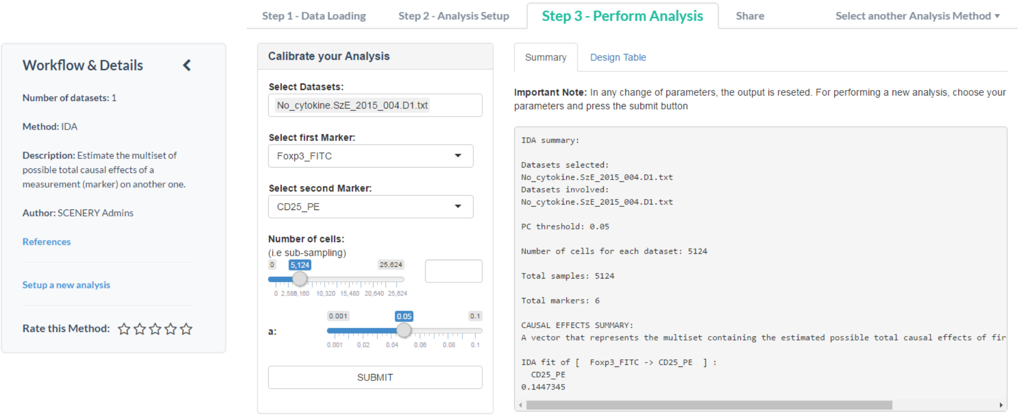
In the 'Results' panel, instead of a textual network summary, a summary of the generated R object of the FCI R function is presented. Moreover, the network adjacency matrix is available on the tab 'Graph Adjacency matrix' along with a heat-map visualization. In this adjacency matrix (adj_mat), the edges' ending points are encoded by numbers that can be summarized by: 0 = no edge, 1 = circle, 2 = arrowhead. For example, if adj_mat[i,j] = 1 and adj_mat[j,i] = 2, this represents the edge i <-o j. Finally, the interactive R based visualization (R package visNetwork) of the reconstructed network is available in the tab 'Plots'.

The implementation of this method was based on the R function 'fci' of the R package 'pcalg' [47], while the default Gaussian based conditional independence test was used. For more, check the 'pcalg' package's reference manual.

IDA The Intervention-calculus when the DAG is absent, IDA [47] algorithm, computes a lower bound for the size of the causal relationships between two variables by implicitly and efficiently enumerating the whole set of causal structures consistent with the data [51]. In this way IDA can estimate causal effects on the basis of solely observational data; the applicability of the method is limited by the assumptions of causal sufficiency (no confounders) and linearity.

In order to estimate these causal effects in SCENERY, the users select one or more FCS files, which are pooling all together and the two stand alone markers. Then an appropriate input for taking a random sub-population (sub-sampling) is available. Finally, the users calibrate the significance level threshold for testing the generated p-value of each of the individual conditional independence tests, and submit the analysis.

The result of this method is a summary of the possible causal effects, derived by the selected measurements and it is presented along with the analysis's meta-data in the 'Summary' tab of the 'Results' panel (see Fig. 3.14). This method does not plot any reconstructed network.



Number of datasets: 1

Method: IDA

Description: Estimate the multiset of possible total causal effects of a measurement (marker) on another one.

Author: SCENERY Admins

Select Datasets: No_cytokine.SzE_2015_004.D1.txt

Select first Marker: Foxp3_FITC

Select second Marker: CD25_PE

Number of cells: (i.e sub-sampling) 5,124

a: 0.05

Important Note: In any change of parameters, the output is reseted. For performing a new analysis, choose your parameters and press the submit button

```
IDA summary:

Datasets selected:
No_cytokine.SzE_2015_004.D1.txt
Datasets involved:
No_cytokine.SzE_2015_004.D1.txt

PC threshold: 0.05

Number of cells for each dataset: 5124

Total samples: 5124

Total markers: 6

CAUSAL EFFECTS SUMMARY:
A vector that represents the multiset containing the estimated possible total causal effects of fir

IDA fit of [  Foxp3_FITC -> CD25_PE  ] :
  CD25_PE
0.1447345
```

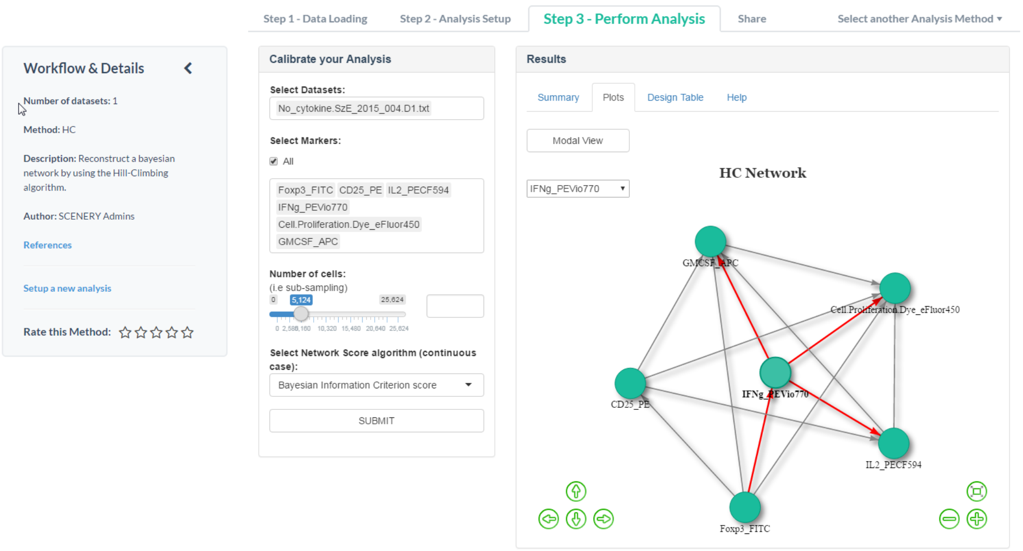
Figure 3.14: UI of the 'IDA' NR analysis method, showing a textual representation of the results in the 'Summary' tab, after submitting an analysis.

3.4.3 Bayesian Networks

Bayesian Networks (BNs) use Directed Acyclic Graphs (DAGs) for representing the multivariate distribution of the data. A common misconception is interpreting a directed edge in a BN as an indication of causal interaction. This is possible only under the standard causal discovery assumptions (Causal Markov Condition, Faithfulness), Even then, not all causal relationships are identifiable by data alone.

HC The Hill Climbing (HC, [52]) algorithm is a greedy-search algorithm that performs a heuristic search across the space of Bayesian networks that may represent the data, and returns the best candidate according to a given metric [53].

As shown in Fig. 3.15 and in consistency with the other NR methods in SCENERY, the users select one or more FCS files, which are pooling all together and the, involved in the analysis, markers. Then an appropriate input for taking a random sub-population (sub-sampling) is available. Finally, the users select the network's scoring algorithm, and submit the analysis. The available algorithms for scoring the candidate Bayesian networks, with respect to the continuous values of the measurements, are the Bayesian Information Criterion, the Multivariate Gaussian log-likelihood, the Akaike Information criterion and the Gaussian posterior density.



Workflow & Details

Number of datasets: 1

Method: HC

Description: Reconstruct a bayesian network by using the Hill-Climbing algorithm.

Author: SCENERY Admins

Rate this Method: ☆☆☆☆☆

Calibrate your Analysis

Select Datasets: No_cytokine.SzE_2015_004.D1.txt

Select Markers:

☑ All

Foxp3_FITC CD25_PE IL2_PECF594 IFNg_PEVio770 Cell.Proliferation.Dye_eFluor450 GMCSF_APC

Number of cells: (i.e sub-sampling)

0 5,124 25,624

0 2,588,160 10,320 15,480 20,640 25,624

Select Network Score algorithm (continuous case): Bayesian Information Criterion score

Results

IFNg_PEVio770 ▾

Figure 3.15: UI of the ’HC’ NR analysis method, showing the generated reconstructed network in the ’Plots’ tab, after submitting an analysis.

The implementation of this method was based on the R function ’hc’ of the R package ’bnlearn’ [52]. For more information about the implementation of the algorithm and the scoring algorithms, check the bnlearn package’s reference manual.

Method	**Analysis Category**	**Short Description**	**Result**
Data Visualization	Visualization	Visualization of single-cell measurements	Histograms, scatter-plots, density-contour plots
Transformation	Pre-Processing	Transformation procedure for cytometry files	New FCS files
Compensation	Pre-Processing	Compensation procedure for flow cytometry files	New FCS files
Gating	Pre-Processing	Gating procedure for flow cytometry files	R shiny interactive plots, new FCS files
Factor Analysis	Uni. Statistical	Population comparison based on experimental design factors (t-test, anova)	Summary statistics, density plots, violin plots
Linear Regression	Uni. Statistical	Fits a linear model between a numeric experimental design factor and a measurement	Summary statistics, scatter-plots with fitted regression lines
Logistic Regression	Uni. Statistical	Fits a logistic model between a categorical experimental design factor and a measurement	Summary statistics, scatter-plots with fitted regression lines
Correlation	NR	Reconstructs an association network	Undirected graphs
MMPC	NR	Reconstructs a conditional association network	Undirected graphs
PC	NR	Reconstructs a causal network assuming no latent confounders	Partial directed acyclic graphs
FCI	NR	Reconstructs a causal network assuming possible latent confounders	Partial ancestral graphs
IDA	NR	Estimates possible total causal effects	Causal effects summary
HC	NR	Reconstructs a Bayesian network	Directed acyclic graphs

Table 3.1: Summary table of the current analysis methods in SCENERY.

Chapter 4

Architecture

4.1 Software Architecture

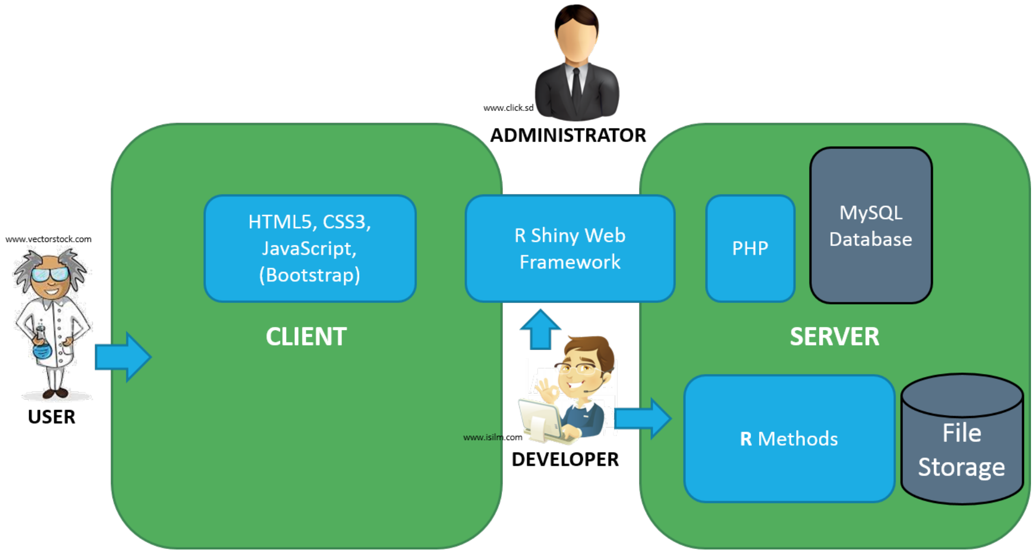


Figure 4.1: Overview of SCENERY's architecture components and interactions among the different users of the application.

SCENERY is a platform-independent web application of Client-Server architecture; built on R and PHP running on an Apache web server (`www.apache.org/`). The architecture follows the Client-Server model, where the basic idea is the partition of the User Interface (Client) from the resources/services (Server). The interface on the Client side is implemented using HTML5, CSS3 for structuring and presenting the content and JavaScript for handling light-weight tasks such as validation of forms, effects on moving elements, asynchronous communication and more. In order to alleviate the overhead associated with common tasks in web development, the Bootstrap web framework (getbootstrap.com/) is used, which is considered as the state-of-the-art web frameworks lately, while the R Shiny web framework is used (shiny.rstudio.com/), in order to wrap each method into an independent web-application and in order to allow R functions communicate among Client and Server. By going further in web frameworks, we can say that they are, usually, used in the development of dynamic websites, promoting the code-reuse idea, which is the idea of using existing software in order to build new one. They, mainly, provide templates, CSS classes and functions for structuring the content of the Client side, while libraries are also providing for session management, database access and client-server interactions (i.e. Bootstrap provides JavaScript functions while R shiny provides, mainly, R along with some JavaScript functionality). Regarding the analysis methods, they are all implemented in R and run on the Server. Additionally, PHP was used as an application skeleton/controller, for managing most of the Client-Server interaction and database operations. A MySQL database (www.mysql.com/) is used for storing users' information and history and for saving the vital entities of the application such as the users, the methods and the results. Finally, a windows-based file storage system is currently available, for storing the users' uploaded files.

Figure 4.1 illustrates the architecture of SCENERY along with the typical users of the application, which are the standard users, the administrators and the developers. Standard users interact directly with the client side of the application by using the UI. These users are mostly scientists and researchers from the single-cell analysis field. When such a user login to SCENERY, a PHP session regarding the current user's information and history is created for ease of maintaining the state of these information across different pages of the application, while HTTP cookies are used for recording the user's browsing activity for a more personalized and easy use in future sessions. On the same page, the developers of SCENERY are, at most, computer scientists that develop new R analysis methods under SCENERY's layout and structure. Finally, the SCENERY administrators maintain and develop further the system while they are responsible for the smooth operation of every analysis method, as well as, scaling up the system as it grows in terms of new users and methods. Moreover, the administrators moderate each new method submission by testing it further and incorporating it in the system.

Regarding the run of the R analysis methods on the server side, a single R shiny web application is created for each method and initialized in our system by calling the R function 'runApp' of the R package 'shiny', with the appropriate arguments. Every R shiny application should listen to a host's IPv4 address, which in our case is the IPv4 of the server that hosts SCENERY. Moreover, each application commits and listens to a unique TCP port on the server, making it available under a port-oriented URL. For example if the IPv4 address of our server is '127.0.0.1' and a method listens to the TCP port '3030', then this method is available under the URL 'http://127.0.0.1:3030'. Then, and in order to be available in SCENERY, each method's port-oriented, unique URL is displayed to the users by an HTML inline frame (iframe) which interacts with the whole system.

Finally, in order to overcome the issues of the current system's architecture, a scaling up of the application is already on development. The current issues of the SCENERY architecture are, mostly, rely on (a) the support of running an analysis method by multiple users simultaneously, (b) the limitations of the current processing power and (c) the limited file storage space. In order to overcome (a) in the current version of SCENERY, we incorporated multiple instances (currently three, under different TCP ports) for each R shiny application that corresponds to a specific method. This solution is considered as a temporary, because it is very demanding on memory and CPU resources, while it sets a limit on the number of users that simultaneously run an analysis with the same method (currently three).

On that, we are currently incorporating SCENERY to a cloud server, along with the ability of using high performance computing (HPC) and multiple virtual machines (VM) services. This will offer to our system an effective use of the required resources than the current local server, for ease of, efficiently, running each analysis process. Hence, that solved the issues (b) and (c). Moreover a Docker-based [54] idea will be applied to the internal architecture of each analysis method. More specifically, we are currently developing a 'dockerized' version of the available architecture of our R-shiny based applications, that will efficiently used in order to overcome issue (a). By going further on the 'dockerization' of our system, a single Linux-based Docker container (like a lightweight, machine independent, virtual machine) will be initialized after a user clicks to a specific analysis method and will be closed after the user exits the current method. This will ensure of running each user's analysis independently and simultaneously with other users, without any limitations. For more on the Docker technology please visit `www.docker.com`.

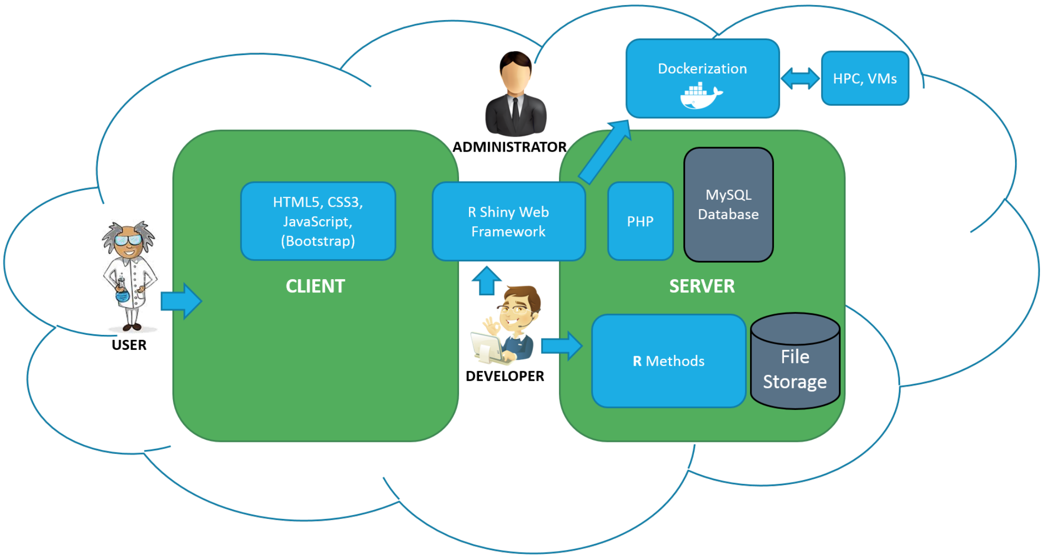


Figure 4.2: Overview of SCENERY's Docker-based architecture moved on a cloud server. Currently on development.

While the current architecture is a 'work in-progress', future version(s) will be extended with operational information built upon Design Strategy for Device Independence [55] enabling the web application to be utilized in various screen dimensions and environments of use.

4.2 Modularity

One of the main and most important features of SCENERY is its modularity. The importance of this feature is mainly relies on the aspect of extending massively the system in term of easily incorporating new, publicly available single cell analysis methods submitted by the users, transforming SCENERY to an essential tool for the cytometry community.

On that, each analysis method is provided by a single R function with a standardized signature (datasets, design table, method's options) and results' type (summary, visualization). This ensures that further analysis methods can be easily integrated within the step-wised SCENERY structure, by allowing users submitting their own NR methods as R code.

In order to incorporate new users' methods in SCENERY, an online step-wised tutorial and an HTML form are available for ease of submission.

Method Definition At first step, the users have to define their method by completing the appropriate online form. Required fields of this form are the method name, the corresponding analysis category that it belongs, the method's short and extended description and the visibility status of this method in SCENERY as public or private. Future versions of SCENERY will allow user methods to be visible in a group of users as well. Moreover, optional fields are available, regarding the method's references and a text field about the generated method's visualization manual. Finally, the author of the method is defined automatically as the user that submits the current method.

Method Standardization The next step of the submission is about the standardization of the method. Each method in SCENERY must be formatted in a standardized way, in order to be compatible with our system, as shown in Fig. 4.3 and described below.

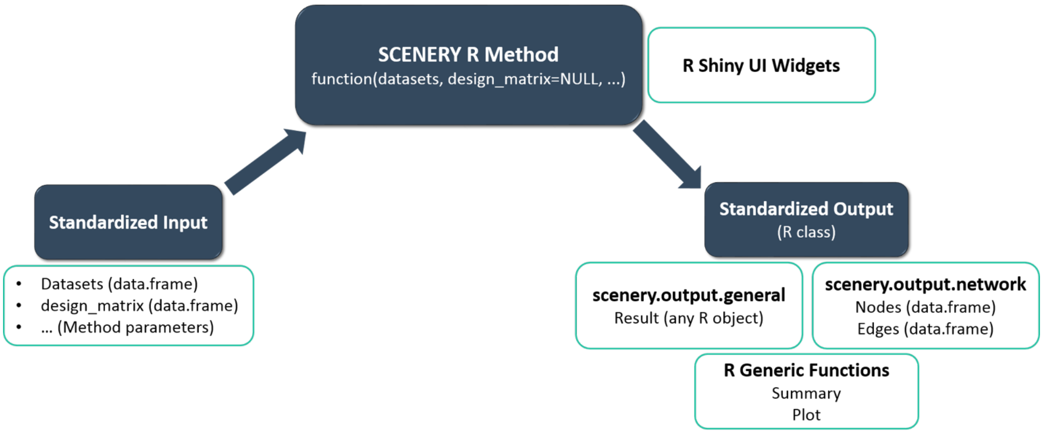


Figure 4.3: Standardized modular architecture of the R methods in SCENERY.

Standardized Input Each method is a main R function and is under a standardized input signature which is described by the following:

```
method_name <- function(datasets, design_matrix=NULL, ...)
{ ... }
```

The first required argument is called datasets and it is an R object of the class list that contains all the FCS file expressions (datasets) as data.frame objects in R. In each data.frame, the columns correspond to the markers (measurements) and the column names correspond to these marker names. The rows of each data.frame correspond to different cells, containing the expression of each measurement (i.e. marker). All the datasets involved in the method's analysis procedure must be retrieved, internally, by this argument.

The second optional argument is called design_matrix and it is under the data.frame R class, as well. This argument corresponds to the experimental design table that will be involved in the analysis. An analysis method may be independent of the experimental design so this argument is optional and its default value is set to NULL. The columns of this data.frame object correspond to the experimental design factors, as defined in Section 2.2 and the column names correspond directly to these factor names. Worth noted is that the first column corresponds to the current file name and it is named as 'Name'. The rows of this data.frame object correspond to the FCS files or samples that are involved in the experimental design, containing the values of each factor for the current sample (i.e. row).

Finally, the three dots (or ellipsis) argument corresponds to the methods' hyper parameters. The number, the names and the R classes of these arguments are decided by the method's author, without any limitations, and they have to be described analytically in the extended description of the method in the first step of the submission.

Standardized Output On the same page, each analysis method's generated results should be under a standardized output. In order to alleviate this, each methods' output should be an R object of the class scenery.output along with the generic R functions summary and plot that correspond to the 'Summary' and 'Plot' tabs, respectively, of each method's 'Result' panel in the system's UI.

More specifically, an R class called scenery.output.network (sub-class of the generic class scenery.output) is available as output for a network reconstruction method. The Slots of this class are the 'nodes' and the 'edges'. The 'nodes' slot corresponds to an R data.frame object with required named columns the 'id', a character R object that corresponds to the node names (i.e. marker names) and the 'type', a character R object for describing the type of each node (i.e. protein marker). Optional columns, defined by the method authors, are allowed as well. On the same hand, the 'edges' slot corresponds to an R data.frame object with required named columns the 'from' and 'to', character R objects corresponding to a node id and the 'type' which is a numeric R object that corresponds to each edge's ending points. Zero (0) value is used no ending point (i.e. undirected edge), one (1) value is used for an arrow head, while value 2 is used for a circle end-point and 3 for an 'X' end point. Optional named columns are currently incorporated such as the 'weight', a numeric R object for the edge's weight (i.e. a p-value or a statistic) and ’dashed’, a boolean R object corresponding to a dashed edge or not. More optional columns can be declared by the authors for extra annotation of each edge.

In addition, another R class called scenery.output.general (sub-class of the generic class scenery.output) is also available as output for other analysis methods except the NR ones. The unique slot of this class is called ’results’ and it is defined under the abstract R class-keyword "ANY" which basically means it can be under any R class that the author decides.

Each SCENERY method’s output class has to contain two polymorphic generic functions which are the R functions ’summary’ and ’plot’ that are directly correspond to the ’Summary’ and ’Plot’ tabs of the UI and they get as the only argument the, generated from the method, R scenery.output object. The ’Summary’ function prints out the textual representation of the generated method’s results as an overview by printing the key-objects of the output, needed for interpreting the results by other users, and it is required. The ’plot’ function contains the code for the graphical representation of the generated output object and it is optional. In case of a NR method, the default visNetwork-based visualization will be used. In case of other methods no default visualizations are available and they need to be incorporated by the method authors, if there are any.

Finally, R script examples and function templates are available online, for ease implementation of the new users’ methods.

In order to guarantee that a method works and its compatible with the SCENERY layout and structure, an offline validation and unit tests of the method will be performed by the system moderators once the method is submitted.

File Uploading The final part of a new method’s submission is the uploading of the required files. At first, users have to submit an R file containing the implementation of the method with at least one R function. Then, an R file with the appropriate generic functions ’summary’ and ’plot’, compatible with the method’s standardized output (see previous sub-section) has to be uploaded. R script templates are available online, containing the skeleton and appropriate comments for the definition of the requested files. Finally, a TXT file, containing the desired calibration panel R shiny widgets that will be used in the method’s UI, referenced by name for each method’s hyper parameter, is optionally requested. In case of absence of this file in the submission, default R shiny widgets will be used by the system’s administrators. On the other hand, each line of this file has to contain a hyper-parameter along with the desired R shiny widget name corresponding to the official R shiny widget gallery (`http://shiny.rstudio.com/gallery/widget-gallery.html`), separated by a comma.

Chapter 5

Application on Immunology

5.1 Definition of the Problem

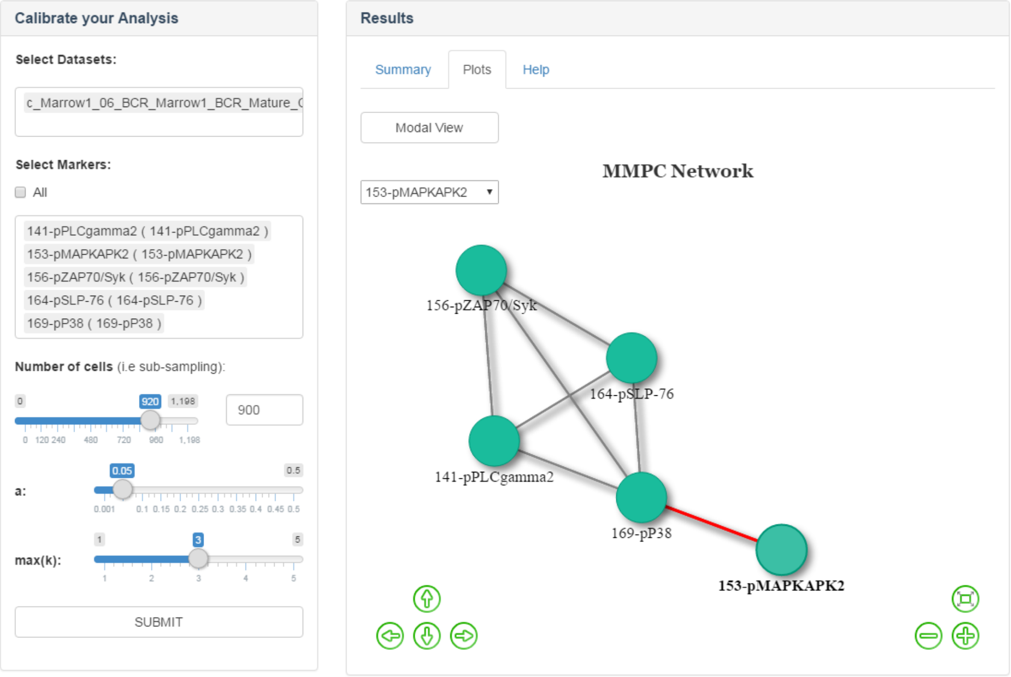
To a computer scientist, signaling networks are informal causal models for which proteins are key members. Their role is to relay the signal by switching between active and inactive states, thereby altering their function. Information is passed on sequentially from one protein to the other until the response is produced. We demonstrate an application of SCENERY by trying to reconstruct a part of a signaling pathway, using public mass cytometry data published in [56]. In the original study, the authors use mass-cytometry to measure 31 proteins related to the human hematopoietic system in two healthy bone marrow donors. Cells were stimulated with several activators to uncover distinct signaling mechanisms. Here, we use data from B-cell populations. Particularly, cells treated with stimulus of the B-cell antigen-receptor (BCR). BCR signaling is known to trigger several signaling cascades simultaneously permitting many distinct outcomes [57]. These include proliferation, survival and differentiation as well as orchestrating the generation of antibodies. Hence, this dataset provides an excellent showcase for the features and applicability of SCENERY in the research of signaling pathway networks.

Figure 5.1 illustrates the user functionality of our platform in the current use case. The user can first configure his analysis in the calibration panel shown on the left of the figure by selecting the involved in the analysis markers and setting the appropriate values to the method's hyper parameters. After submitting the analysis, the user can view the generated output of the method as a summary text or as a network visualization, as shown on the right part of the figure.

5.2 Analyzing Cytometry Data with SCENERY

In the following examples we employ a subset of proteins, known to be involved in BCR signaling, which are the spleen tyrosine kinase (SYK), the B-cell linker protein (BLNK or pSLP-76), the phospholipase C,γ2 (PLCγ2), the mitogen-activated protein 14 (p38) and the mitogen-activated protein kinase-activated protein kinase 2 (MAPKAPK2).

Figure 5.1 displays the network reconstruction analysis results as they were retrieved by running the 'MMPC' method in SCENERY. The undirected edges denote correlation between the respective protein markers in a sense that both bi-connected nodes have been selected in the Parent-Children set of each other. Starting from the top left corner, then, the reconstructed network indicates that SYK, BLNK (pSLP-76) and PLCγ2 are intercorrelated. This is true biologically because the stimulated BCR attracts and activates SYK which, in turn, attracts, interacts and phosphorylates both BLNK (pSLP-76) and PLCγ2 [57],[58]. This process is part of a complex stimulation process that ultimately activates several proteins. One of them is p38 which interacts with both PLCγ2 and BLNK (pSLP-76) in order to be activated [59]. This process is captured in SCENERY's output and is shown in the reconstructed network by the respective correlation edges. After p38 and further downstream, the reconstructed network extends to MAPKAPK2. This edge is also consistent with the literature, where MAPKAPK2 is found to be directly phosphorylated by p38 [60]. Activated MAPKAPK2 can then mediate the regulation of many biological responses including gene transcription and cell cycle control by amplifying the p38 signal [61].



Calibrate your Analysis

Select Datasets:

c_Marrow1_06_BCR_Marrow1_BCR_Mature_C

Select Markers:

□ All

141-pPLCgamma2 (141-pPLCgamma2)
153-pMAPKAPK2 (153-pMAPKAPK2)
156-pZAP70/Syk (156-pZAP70/Syk)
164-pSLP-76 (164-pSLP-76)
169-pP38 (169-pP38)

Number of cells (i.e sub-sampling): 920 (0 – 1,198) 900

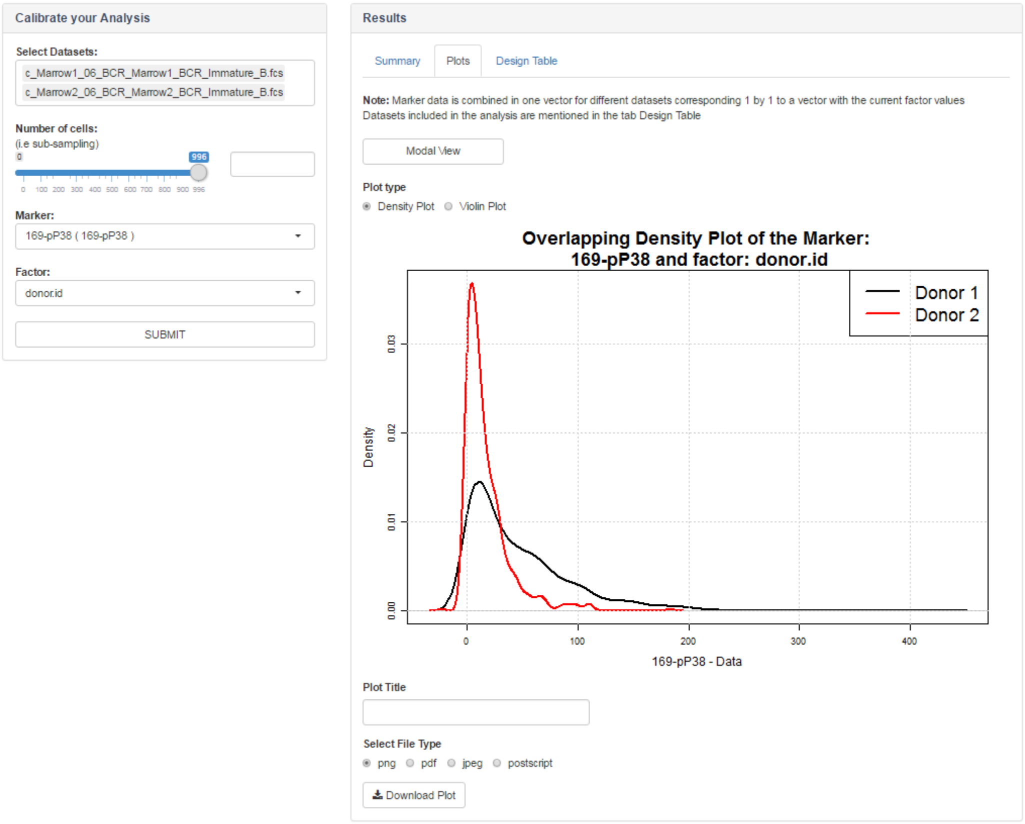
a: 0.05 (0.001 – 0.5)

max(k): 3 (1 – 5)

Results

Figure 5.1: Visualizing results in SCENERY. The retrieved reconstructed network after applying the MMPC method on selected mass cytometry data (see text for details). The analysis calibration panel is also displayed at the left of the reconstructed network, as indicative of the UI.

Moreover, Fig. 5.2 illustrates how SCENERY would visualize a population comparison result (Univariate Statistical Analysis) by the 'Factor Analysis' method. For this graph we employed data from 2 donors for the protein marker p38. At the left half of the figure, the screen-shot shows the configuration of the calibration options. At the right half, the two overlapping density plots for this specific analysis are shown. This population comparison analysis was performed on the protein marker p38 with respect to an experimental design factor denoted the donor id. Two FCS files (samples) were used and the population comparison was performed by a two-sample t-test. The summary results shown a significant difference on the compared donor populations by considering the t-test statistic summary and the generated overlapping density plot of the analysis as shown in Fig. 5.2.



Calibrate your Analysis

Select Datasets:
c_Marrow1_06_BCR_Marrow1_BCR_Immature_B.fcs
c_Marrow2_06_BCR_Marrow2_BCR_Immature_B.fcs

Number of cells:
(i.e sub-sampling)
0 – 996

Marker: 169-pP38 (169-pP38)

Factor: donor.id

Results

Note: Marker data is combined in one vector for different datasets corresponding 1 by 1 to a vector with the current factor values
Datasets included in the analysis are mentioned in the tab Design Table

Plot type
● Density Plot ○ Violin Plot

Overlapping Density Plot of the Marker:
169-pP38 and factor: donor.id

Plot Title

Select File Type
● png ○ pdf ○ jpeg ○ postscript

Figure 5.2: Visualizing results in SCENERY. The retrieved overlapping density plot by applying a population comparison among two donors for the protein marker p38. The univariate statistical analysis method 'Factor Analysis' was used on two selected subpopulations from a mass cytometry public dataset (see text for details) with respect to the 'donor.id' factor from the experimental design. The analysis calibration panel is also displayed at the left of the reconstructed network, as indicative of the UI.

Chapter 6

Conclusions and Future Work

In this work we introduced SCENERY, the first freely available web-based application for network reconstruction (NR), visualization and statistical analysis of single-cell data. SCENERY packages advanced machine-learning methods in a user-friendly environment: a wizard guides users through all phases of the complex single cell analysis effort, with an emphasis to the NR analysis, delivering a simple-to-use interface. This, mainly, allows biology researchers unfamiliar with the technical details to exploit NR methods in discovering novel signaling pathways, while it allows SCENERY to serve as an educational tool for exploring the features of single cell analysis methods.

Regarding each step of SCENERY's pipeline we indicated that uploading data in the FCS file format allows it to be used as a companion other cytometry analysis tool. We also argued about the important novelty of the software to look ahead and allow as descriptors of the experimental study design any type of factors, beyond the experimental ones (see Chapter 2). We pointed out the available standard cytometry data visualization and preprocessing functionalities (e.g. data compensation, transformation, gating); the statistical analysis methods for modeling the relationship of the experiment measurements and an analytic description of the powerful NR methods employed with numerous figures illustrating these functionalities (see Chapter 3). Moreover, a typical graph theory analysis on each reconstructed network was employed, for pointing out each network's characteristics. The outline and the typical work-flow of each customizable analysis execution was also provided, from the datasets' selection and filtering to the calibration of the analysis parameters and the interpretation of the results. Finally, we stressed out that results are given in well-known textual and graphical formats, acceptable to the biology community, with the ability of exporting the generated output in various ways, mainly publication quality figures and standard graph-representation formats.

More features will be implemented in feature releases. These include: implementation of an online archive for the users sessions; establishment of the connection with other online services for directly loading of public data (e.g. CytoBank); sharing analysis results via email or social media; and enriching SCENERY with even more analysis methods. New single cell analysis methods and UI improvements have already been considered, after getting feedback and discussing with beta-version test-users and colleagues from the cytometry field. Particularly, new analysis methods regarding NR, data dimensionality reduction (e.g. t-SNE) and visualization improvements are already on development. Regarding the future work on the visualization, new layouts and annotations of the reconstructed networks will be implemented, along with UI improvements for, massively, visualizing the appropriate quantities of different files and analysis results, organized together, for ease of interpretation and comparison. Moreover, an extension of the NR analysis to the one that will involve experimental design factors (such as disease indicator, patient ID, genre, cell-type etc) as nodes will be released in future versions, for exploring even more the reconstructed networks in terms of statistical and causal associations among measurements and such factors. Finally, public single cell data will be available online, in order for the users to explore and try out easily and immediately a variety of SCENERY features, giving also a ton of educational use to the application.

Regarding the current SCENERY architecture, as exploited in chapter 4, we indicated that SCENERY is a platform-independent web application that follows the client-server approach built, mainly, in HTML, PHP and R. We pointed out the different types of the typical users (single-cell analysis researchers, computer scientists, developers and administrators) of the application and how the interact within the SCENERY infrastructure. Moreover, we discussed about the modular design of the system, which offers the ability to further extend SCENERY on new single-cell analysis methods, submitted, as R source code, by users and developers working on the analysis of such fields, transforming SCENERY into an essential tool for the cytometry community. These custom, modular methods may be re-utilized by colleagues or members of a group specified by the end-user. Particularly and in order to develop this vital feature, we employed each analysis method in a modular way by standardizing its input and output, ensuring that further analysis methods will be easily integrated within the SCENERY layout and structure. Further more, an online step-wised tutorial and a submission form were created for letting users, easily, develop and incorporate, publicly, their methods. Regarding the scaling up of the system as a future work, we argued that a new Docker-based version of the internal SCENERY architecture is already on development, while the whole system will be hosted under a cloud server for ensuring an efficient use of the required resources. This, along with the capabilities of using HPC and VM services, will establish the running of each analysis process independently and efficiently by multiple users simultaneously, without any limitations in terms of processing power, storage space and number of users.

Finally, We showcase most of SCENERY's functionality by using test single-cell datasets in various use cases across the available analysis methods, as discussed in Chapter 3. Moreover, we illustrated and validated, through literature, the system's capabilities by using a public mass cytometry B-cell dataset published in [56], by employing a BCR signaling dataset and by demonstrated the software's applicability in the field of signaling pathway research. We illustrated in Fig. 5.2 the simplicity with which the software can represent standard statistical analysis results and in Fig. 5.1, we show how users can easily assess results from a NR analysis in SCENERY.

By concluding this thesis, our efforts towards this open-source approach hold the promise to transform SCENERY in an essential tool for the cytometry community for understanding the organization of complex cellular processes such as signaling networks.

Availability. Instructions on how to access and use SCENERY are available at `http://mensxmachina.org/en/software/`. Moreover, an installation manual of the latest version and its current dependencies, along with the system's requirements, is available, on demand, for the system's administrators.

Bibliography

[1] Anselmetti, D.: Single Cell Analysis. John Wiley & Sons, ISBN 9783527626656. (2009)

[2] Sweedler, J. V., Edgar, A.: Single Cell Analysis. Anal Bioanal Chem. 387: 1–2. (2007)

[3] Ullrich, A., Schlessinger, J.: Signal transduction by receptors with tyrosine kinase activity. Cell, 61(2):203-212 (1990)

[4] Bendall, S., Nolan, G., Roederer, M., Chattopadhyay, P.: A Deep Profiler's Guide to Cytometry. Trends Immunol., 33(7):323-332 (2012)

[5] Marbach, D., Costello, J.C., Küffner, R., et. al.: Wisdom of crowds for robust gene network inference. Nat. Methods. 2012 Jul 15; 9(8): 796–804.

[6] S. Woodhouse, V. Moignard, B. Göttgens and J. Fisher: Processing, visualising and reconstructing network models from single-cell data. Immunology and Cell Biology advance online publication, 2015; doi:10.1038/icb.2015.102

[7] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, DA., Nolan, GP.: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science, 308: 523-529 (2005)

[8] Itani, S., Ohannessian, M., Sachs, K., et. al.: Structure learning in causal cyclic networks. JMLR Workshop and Conference Proceedings (2010)

[9] Qiu, P., Simonds, E.F., Bendall, S.C., et. al.: Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotech. 29, 886–891 (2011)

[10] Lagani, V., Triantafillou, S., Ball, B., Tegner, J., and Tsamardinos, I.,: Probabilistic Computational Causal Discovery for Systems Biology. Book chapter in Uncertainty in Biology, A Computational Modeling Approach. Springer, (2015)

[11] Bik H.M., Goldstein M.C.: An Introduction to Social Media for Scientists. PLoS Biol 11(4): e1001535. doi:10.1371/journal.pbio.1001535 (2013)

[12] Priem, J. and Costello, K.L.: How and why scholars cite on Twitter. Proceedings of ASIST, 47(1), 104 (2010)

[13] Athineou, G.; Papoutsoglou, G.; Triantafullou, S.; Basdekis, I; Lagani, V.; Tsamardinos, I.: SCENERY: a Web-Based Application for Network Reconstruction and Visualization of Cytometry Data. Springer International Publishing, 10th International Conference on Practical Applications of Computational Biology & Bioinformatics, 203–211, isbn: 978-3-319-40126-3 (2016)

[14] Kotecha, N., Krutzik, PO., Irish, JM.: Web-based Analysis and Publication of Flow Cytometry Experiments. Curr. Prot. Cyt., Chapter 10, Unit10.17. (2010)

[15] Levine, JH., Simonds, EF., Bendall, SC., et. al.: Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell (2015)

[16] B. Ellis, et.al.: flowCore: Basic structures for flow cytometry data. R package version 1.38.2. `http://bioconductor.org/packages/release/bioc/html/flowCore.html` (2016).

[17] Florian, H., et.al.: flowStats: Statistical methods for the analysis of flow cytometry data. R package version 3.30.1. `http://bioconductor.org/packages/release/bioc/html/flowStats.html` (2016).

[18] B. Ellis, et.al.: flowViz: Visualization for flow cytometry. R package version 1.36.2 `http://bioconductor.org/packages/release/bioc/html/flowViz.html` (2016).

[19] Finak, G., Jiang, M.: flowWorkspace: Infrastructure for representing and interacting with the gated cytometry. R package version 3.18.10. `http://bioconductor.org/packages/release/bioc/html/flowWorkspace.html` (2011).

[20] Fischbacher, Thomas; Synatschke-Czerwonka, Franziska: FlowPy—A numerical solver for functional renormalization group equations. Computer Physics Communications, Volume 184, Issue 8, p. 1931-1945. (2013)

[21] Kelkar, S. A.: Usability and Human-Computer Interaction: A Concise Study, pp 306, (2009).

[22] R Core Team.: R: A Language and Environment for Statistical Computing R package version 0.14. `http://CRAN.R-project.org/package=shiny` (2016).

[23] Almende B.V. and Benoit Thieurmel: visNetwork: Network Visualization using 'vis.js' Library R package version 1.0.1. `https://CRAN.R-project.org/package=visNetwork` (2016).

[24] Chang, W., et.al.: shiny: Web Application Framework for R. R package version 0.14. `http://CRAN.R-project.org/package=shiny` (2016).

[25] Hadley Wickham.: ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, si 2009. `http://ggplot2.org` (2016).

[26] BD Biosences.: An Introduction to Compensation for Multicolor Assays on Digital Flow Cytometers. BD Biosciences, San Jose, CA. https://www.bdbiosciences.com/documents/Compensation_Multicolor_TechBulletin.pdf (2009).

[27] Welch, B. L.: The generalization of "Student's" problem when several different population variances are involved. Biometrika. 34 (1–2): 28–35. (1947).

[28] Welch, B. L.: On the Comparison of Several Mean Values: An Alternative Approach. Biometrika. 38: 330–336 (1951).

[29] Ronald A. Fisher.: The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Philosophical Transactions of the Royal Society of Edinburgh. (volume 52, pages 399–433) (1918)

[30] W. N. Venables and B. D. Ripley: Modern Applied Statistics with S. Springer, Fourth edition, NY (2002) ISBN 0-387-95457-0 http://www.stats.ox.ac.uk/pub/MASS4

[31] Kobourov, Stephen G.: Spring Embedders and Force-Directed Graph Drawing Algorithms. University of Arizona. (2012) https://pdfs.semanticscholar.org/46f0/fed7d54a6f2bae021d10821febbee7c229fc.pdf

[32] Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu: Gephi : An Open Source Software for Exploring and Manipulating Networks. AAAI Publications, Third International AAAI Conference on Weblogs and Social Media, retrieved 2011-11-22 (2009) https://gephi.org/

[33] Gabor Csardi and Tamas Nepusz: The igraph software package for complex network research. InterJournal, Complex Systems, 1695 (2006) http://igraph.org

[34] Lee, C. Y.: An Algorithm for Path Connections and Its Applications. IRE Transactions on Electronic Computers. (1961) http://ieeexplore.ieee.org/document/5219222/?arnumber=5219222

[35] Ulrik Brandes.: A Faster Algorithm for Betweenness Centrality. Published in Journal of Mathematical Sociology 25(2):163-177, (2001) http://algo.uni-konstanz.de/publications/b-fabc-01.pdf

[36] D. J. Watts and Steven Strogatz: Collective dynamics of 'small-world' networks. Nature. 393 (6684): 440–442 (1998)

[37] Karl Pearson: Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58 : 240–242. (1895)

[38] Fisher, R.A: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 10 (4): 507–521 (1915)

[39] Tsamardinos, I., Brown, L.E. and Aliferis, C.F.: The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Mach. Learn., 65, 31–78 (2006)

[40] Lagani, V., Athineou, G., Borboudakis, G. and Tsamardinos, I.: MXM: Discovering Multiple, Statistically-Equivalent Signatures. R package version 0.4.3. `http://CRAN.R-project.org/package=MXM` (2015).

[41] Aliferis, C.F., Statnikov, A., Tsamardinos, I., et. al.: Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. JMLR, 11:171-234 (2010)

[42] Peter Spirtes, Clark Glymour, and Richard Scheines: Causation, Prediction, and Search. The MIT Press, Cambridge, MA, USA, second edition, 2001

[43] Lee Rodgers J., and Nicewander W.A.: Thirteen ways to look at the correlation coefficient. The American Statistician 42(1): 59-66. (1988)

[44] Hampel F. R., Ronchetti E. M., Rousseeuw P. J., and Stahel W. A.: Robust statistics: the approach based on influence functions. John Wiley & Sons (1986).

[45] Shevlyakov G. and Smirnov P.: Robust Estimation of the Correlation Coefficient: An Attempt of Survey. Austrian Journal of Statistics, 40(1 & 2): 147-156 (2011)

[46] Lagani, Vincenzo; Triantafillou, Sofia; Ball, Gordon; Tegner, Jesper; Tsamardinos, Ioannis: Probabilistic computational causal discovery for systems biology. Uncertainty in Biology. (2016) `http://mensxmachina.org/en/publication-details/?docid=fca9bdd6-e1f4-3f11-adb6-356704e5078e`

[47] Kalisch, M., Maechler, M., Colombo, D., Maathuis, MH. and Buehlmann, P.: Causal Inference Using Graphical Models with the R Package pcalg. JSS, 47(11), 1-26. `http://www.jstatsoft.org/v47/i11/` (2012)

[48] P. Spirtes, C. Glymour and R. Scheines: Causation, Prediction, and Search. Springer New York, vol. 81 (1993)

[49] D. Colombo, M. H. Maathuis, M. Kalisch, T. S. Richardson: Learning high-dimensional directed acyclic graphs with latent and selection variables. Annals of Statistics 40, 294-321. (2012)

[50] J. Zhang: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, vol. 172, no. 16–17, pp. 1873–1896, (2008)

[51] M.H. Maathuis, M. Kalisch, P. Buehlmann: Estimating high-dimensional intervention effects from observational data. Annals of Statistics 37, 3133–3164. (2009)

[52] Scutari, M.: Learning Bayesian Networks with the bnlearn R Package. JSS, 35(3), 1-22. `http://www.jstatsoft.org/v35/i03/` (2010)

[53] Russell SJ, Norvig P: Artificial Intelligence: A Modern Approach. Upper Saddle Rivern (1995)

[54] Dirk Merkel: Docker: lightweight Linux containers for consistent development and deployment. Linux Journal archive. Volume 2014 Issue 239, Article No. 2 (2014) `https://www.docker.com/`

[55] Karampelas, P., Basdekis, I. and Stephanidis, C.: Web user interface design strategy: Designing for device independence. In: C. Stephanidis (ed.). Proceedings of the 13th International Conference on Human-Computer Interaction HCI International 2009 July 19-24, 2009, San Diego, CA, USA. pp. 515-524. (2009)

[56] Bendall, S.C., Simonds, E.F., et. al.: Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science. 332 (6030): 687–696. (2011)

[57] Dal Porto, J., Gauld, S., Merrell, et.al.: B cell antigen receptor signaling 101. Mol. Imm. 41:599–613 (2004)

[58] Ishiai, M., Kurosaki, M., Pappu, R., et.al.: BLNK Required for Coupling Syk to PLCg2 and Rac1-JNK in B Cells. Imm. 10: 117–125 (1999)

[59] Guinamard, R., Signoret, N., Ishiai, et.al.: B cell antigen receptor engagement inhibits stromal cell-derived factor (SDF)-1alpha chemotaxis and promotes protein kinase C (PKC)-induced internalization of CXCR4. J.Exp.Med. 189(9):1461-6 (1999)

[60] Cargnello, M., Roux, P.: Activation and Function of the MAPKs and Their Substrates, the MAPK-Activated Protein Kinases. Microbiol. Mol. Biol. Rev. 75(1):50–83 (2011)

[61] Ono, K., Han, J.: The p38 signal transduction pathway Activation and function. Cell. Signal. 12:1-13 (2000)