Table of Contents

1. The Compatibility of Browser and Operating System (OS)
2. Required Formats of the Input Files
  2.1 Flow Cytometry Data for Cell Subpopulation Identification (CSI)
  2.2 Mass Cytometry Data for Cell Subpopulation Identification (CSI)
  2.3 Flow Cytometry Data for Pseudo-time Trajectory Inference (PTI)
  2.4 Mass Cytometry Data for Pseudo-time Trajectory Inference (PTI)
3. Step-by-step Instruction on the Usage of ANPELA 2.0
  3.1 Uploading Your Data or the Sample Data Provided in ANPELA
  3.2 Feature Selection and Data Quantification Workflow (Compensation& Transformation & Normalization & Signal Clean)
  3.3 Performance Evaluation Based on Multiple Criteria
4. A Variety of Methods for Data Quantification
  4.1 Compensation Methods
  4.2 Transformation Methods
  4.3 Normalization Methods
  4.4 Signal Clean Methods

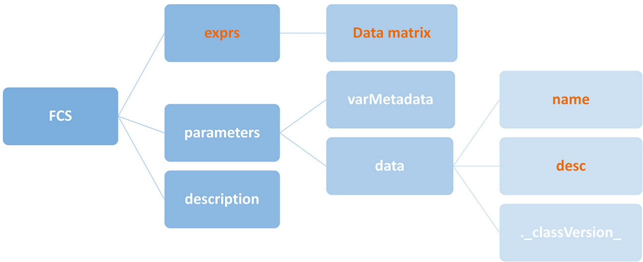
1. The Compatibility of Browser and Operating System (OS)

ANPELA is powered by *R shiny*. It is free and open to all users with no login requirement and can be readily accessed by a variety of popular web browsers and operating systems as shown below.

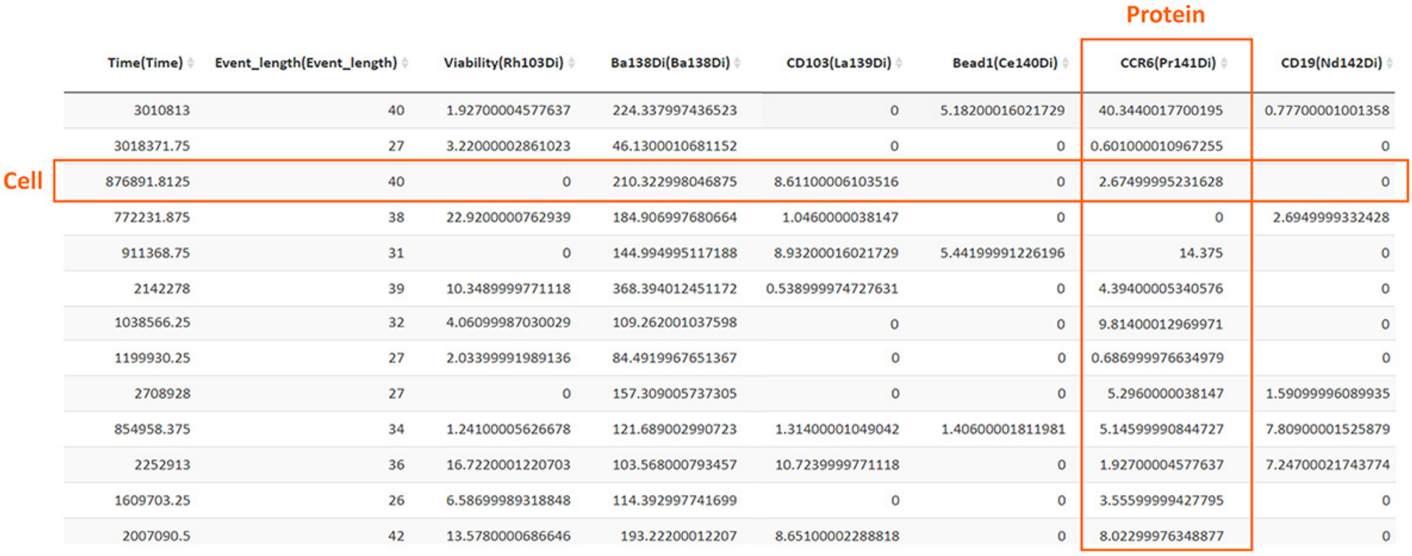
OS	Chrome	Firefox	Edge	Safari
Linux (Ubuntu-17.04)	v78.0.3904.108	v52.0.1	n/a	n/a
MacOS (v10.1)	v78	v71	n/a	v8
Windows (v10)	v78.0.3904.108	v70.0.1	v44.18362.449.0	n/a

2. Required Formats of the Input Files

In general, the file required at the beginning of ANPELA 2.0 analysis should be flow cytometry standard format (FCS). The structure of FCS are as followed, parameters needed in ANPELA workflow are shown in orange:



The data used to quantify single-cell proteomic is extracted from the "exprs" of the FCS. Column name of the data matrix was generated from the "name" and "desc" of the FCS parameters, indicating the protein and the fluorescent antibody or non-radioactive rare-earth-metal isotopes used to stain it. And each row corresponds to a single cell detected by the cytometry.


Protein
Cell

Time(Time)	Event_length(Event_length)	Viability(Rh103Di)	Ba138Di(Ba138Di)	CD103(La139Di)	Bead1(Ce140Di)	CCR6(Pr141Di)	CD19(Nd142Di)
3010813	40	1.92700004577637	224.337997436523	0	5.18200016021729	40.3440017700195	0.77700001001358
3018371.75	27	3.22000002861023	46.1300010681152	0	0	0.601000010967255	0
876891.8125	40	0	210.322998046875	8.61100006103516	0	2.67499995231628	0
772231.875	38	22.9200000762939	184.906997680664	1.0460000038147	0	0	2.6949999332428
911368.75	31	0	144.994995117188	8.93200016021729	5.44199991226196	14.375	0
2142278	39	10.3489999771118	368.394012451172	0.538999974727631	0	4.39400005340576	0
1038566.25	32	4.06099987030029	109.262001037598	0	0	9.81400012969971	0
1199930.25	27	2.03399991989136	84.4919967651367	0	0	0.686999976634979	0
2708928	27	0	157.309005737305	0	0	5.2960000038147	1.59099996089935
854958.375	34	1.24100005626678	121.689002990723	1.31400001049042	1.40600001811981	5.14599990844727	7.80900001525879
2252913	36	16.7220001220703	103.568000793457	10.7239999771118	0	1.92700004577637	7.24700021743774
1609703.25	26	6.58699989318848	114.392997741699	0	0	3.55599999427795	0
2007090.5	42	13.5780000686646	193.22200012207	8.65100002288818	0	8.02299976348877	0

2.1 Flow Cytometry Data for Cell Subpopulation Identification (CSI)

In cell subpopulation identification that based on flow cytometry data, ANPELA compares protein expression of cells under two different conditions, therefore at least two samples for each condition (four FCS files in total) are needed. A metadata csv file which matches the file name to the condition is also needed in the process, the first column is the name of the FCS (without filename extension) from the sample followed by the second column which is the condition of the sample. For compensation method expect CytoSpill additional single stained control samples are needed. Sample data of this data type can be **downloaded**.

Filename	Condition
A_sample_1	CTRL
B_sample_2	MG
C_sample_3	MG
D_sample_4	MG
E_sample_5	CTRL
F_sample_6	CTRL
G_sample_8	CTRL
H_sample_10	CTRL
I_sample_12	CTRL

2.2 Mass Cytometry Data for Cell Subpopulation Identification (CSI)

In cell subpopulation identification based on mass cytometry data (MC/CyTOF), ANPELA compares protein expression of cells under two different conditions, therefore at least two samples for each condition (four FCS files in total) are needed. A metadata which matches the file name to the condition is also needed in the process, the first column is the name of the FCS file (without filename extension) followed by the second column which is the condition of the sample. Sample data of this data type can be **downloaded**.

Filename	Condition
56_RCDII1_Biopsy	Biopsy
57_RCDII2_Biopsy	Biopsy
59_RCDII4_Biopsy	Biopsy
60_RCDII5_Biopsy	Biopsy
61_RCDII6_Biopsy	Biopsy
62_RCDII7_Biopsy	PBMC
63_RCDII1_PBMC	PBMC
64_RCDII2_PBMC	PBMC
65_RCDII4_PBMC	PBMC
66_RCDII5_PBMC	PBMC

2.3 Flow Cytometry Data for Pseudo-time Trajectory Inference (PTI)

In pseudo-time trajectory inference based on flow cytometry data, ANPELA 2.0 can generate a pseudo-progression trajectories based on samples from more then two different time point meaning that at least two FCS files (one for each time) should be uploaded by the user. A metadata csv file specifies a time for each FCS file is also needed, and the first column of the csv are file names of FCS files followed by the second column containing time points when the sample was collected. For compensation method expect CytoSpill additional single stained control samples are needed.

As for evaluation under Criterion Cd (biological meaning), an extra csv file containing the order of proteins in prior known signal transduction cascades is needed, and the format requirements for the csv are shown as the following figure, each column represents a prior known protein activation pathway, form first line to the bottom conform to the sequence of protein activation. All of the proteins in the pathway csv should be included in the marker selected in step two and named same as the markers. Sample data of this data type can be **downloaded**.

Filename	Timepoint
A_D0	0
B_D2	2
C_D4	4
D_D6	6
E_D8	8
F_D10	10

2.4 Mass Cytometry Data for Pseudo-time Trajectory Inference (PTI)

In Pseudo-time Trajectory Inference based on mass cytometry data, ANPELA 2.0 can generate a pseudo-progression trajectories based on samples from more than two different time points meaning that at least two FCS files (one for each time) should be uploaded by the user. A metadata csv file specifies a time for each FCS file is also needed, and the first column of the csv are file names of FCSs followed by the second column containing time points when the sample was collected.

As for evaluation under Criterion Cd (biological meaning), an extra csv file containing the order of the respective proteins in known signal transduction cascades is needed, and the format requirements for the csv are shown as the following figures. Each column represents a prior known protein activation pathway, from first line to the bottom conform to the sequence of protein activation. All of the proteins in the pathway csv should be included in the marker selected in step two and named as the marker. Sample data of this data type can be **downloaded**.

Filename	Timepoint
Timecourse_AllCelltypes_A01_CD4+	0
Timecourse_AllCelltypes_B01_CD4+	1
Timecourse_AllCelltypes_C01_CD4+	5
Timecourse_AllCelltypes_E01_CD4+	30
Timecourse_AllCelltypes_F01_CD4+	60
Timecourse_AllCelltypes_G01_CD4+	120
Timecourse_AllCelltypes_H01_CD4+	240

Pathway 1	Pathway 2	Pathway 3	Pathway 4	Pathway 5	Pathway 6	Pathway 7
pSHP2	pBtk	pErk	pZap70	pLat	pPlcg2	pAkt
pStat1	pAkt	pS6	pLat	pSlp76	pp38	pS6
pStat3	pS6		pSlp76	pPlcg2	pErk	
pStat5	pPlcg2		pPlcg2	pp38	pS6	
pBtk	pp38		pp38	pErk		
pAkt	pNFkB		pErk	pS6		
pPlcg2	pErk		pS6			
pS6	pS6					
pp38						
pErk						
pNFkB						

3. Step-by-step Instruction on the Usage of ANPELA 2.0

This website is free and open to all users and there is no login requirement, and can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. Quantification and comprehensive performance assessment for single-cell proteomics are started by clicking on the "Single-cell Proteomics" panel on the homepage of ANPELA 2.0. The collection of web services and the whole process provided by ANPELA 2.0 can be summarized into 3 steps: **(3.1) uploading single-cell proteomics data**, **(3.2) data quantification workflow**, and **(3.3) performance assessment** A report containing evaluation results is also generated and can be downloaded in the format of PDF, HTML and DOC. The flowchart below summarizes the whole processes in ANPELA 2.0.

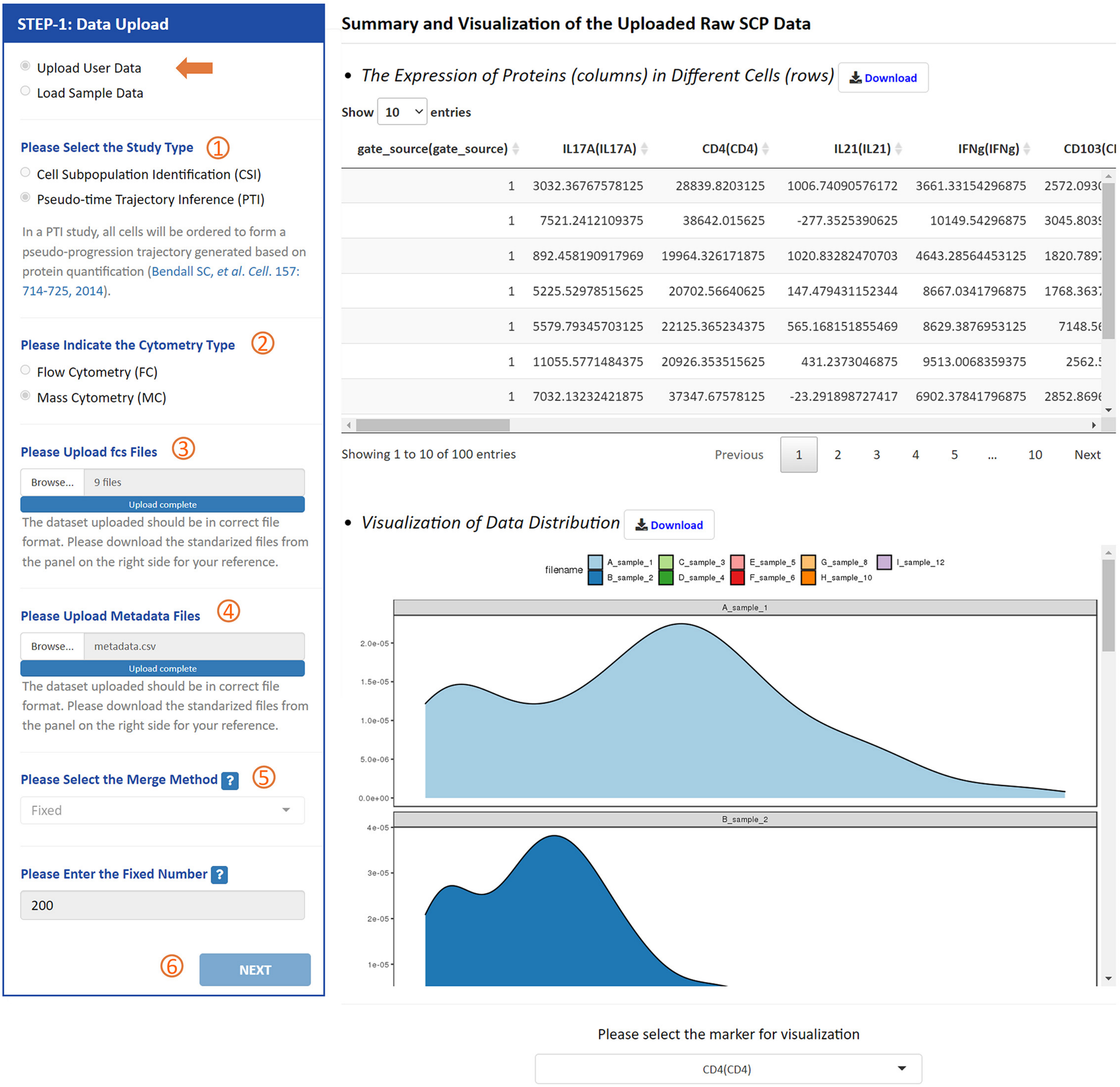
Step 1
Data Uploading
Step 2
Data Quantification
Step 3
Performance Assessment

3.1 Uploading Your Data or the Sample Data Provided in ANPELA

There are 3 radio checkboxes and a drop-down box in STEP-1 on the left side of the analysis page. Users can choose to upload their own cytometry data or to directly load sample data. The type of the study (cell subpopulation identification/pseudo-time trajectory inference) and measurement method (flow cytometry/mass cytometry) are selected in the remaining 2 radio checkboxes below.

And four different merge methods for users to choose from the drop-down box: **(1)** Ceil: Up to fixed number (specified by fixed Num) of cells are sampled without replacement from each FCS file and combined for analysis. **(2)** Fixed: A fixed number (specified by fixed Num) of cells are sampled (with replacement when the total number of cells are led than fixed Num) from each FCS files and combined for analysis. **(3)** All: All cells from each FCS file are combined for analysis and for method. **(4)** Min: The minimum number of cells among all the selected FCS files are sampled from each FCS file and combined for analysis.

Fixed number in "Ceil" or "Fixed" can be assigned by the input box below.



STEP-1: Data Upload

● Upload User Data
○ Load Sample Data

Please Select the Study Type ①

○ Cell Subpopulation Identification (CSI)
● Pseudo-time Trajectory Inference (PTI)

In a PTI study, all cells will be ordered to form a pseudo-progression trajectory generated based on protein quantification (Bendall SC, *et al*. *Cell*. 157: 714-725, 2014).

Please Indicate the Cytometry Type ②

○ Flow Cytometry (FC)
● Mass Cytometry (MC)

Please Upload fcs Files ③

Browse... 9 files

Upload complete

The dataset uploaded should be in correct file format. Please download the standarized files from the panel on the right side for your reference.

Please Upload Metadata Files ④

Browse... metadata.csv

Upload complete

The dataset uploaded should be in correct file format. Please download the standarized files from the panel on the right side for your reference.

Please Select the Merge Method ? ⑤

Fixed

Please Enter the Fixed Number ?

200

⑥ NEXT

Summary and Visualization of the Uploaded Raw SCP Data

- *The Expression of Proteins (columns) in Different Cells (rows)* Download

Show 10 entries

gate_source(gate_source)	IL17A(IL17A)	CD4(CD4)	IL21(IL21)	IFNg(IFNg)	CD103(CI
1	3032.36767578125	28839.8203125	1006.74090576172	3661.33154296875	2572.093
1	7521.2412109375	38642.015625	-277.3525390625	10149.54296875	3045.803
1	892.458190917969	19964.326171875	1020.83282470703	4643.28564453125	1820.789
1	5225.52978515625	20702.56640625	147.479431152344	8667.0341796875	1768.363
1	5579.79345703125	22125.365234375	565.168151855469	8629.3876953125	7148.5
1	11055.5771484375	20926.353515625	431.2373046875	9513.0068359375	2562.
1	7032.13232421875	37347.67578125	-23.291898727417	6902.37841796875	2852.869

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 ... 10 Next

- *Visualization of Data Distribution* Download

Please select the marker for visualization

CD4(CD4)

4 sets of sample data are also provided in this step facilitating a direct access and evaluation of ANPELA 2.0. These sample data are all benchmark datasets collected from previous articles, including **(1)** CSI-FC datasets of flow cytometry-based cell subpopulation identification which contains the blood and thymus samples from three myasthenia gravis patients and six healthy controls. **(2)** CSI-MC datasets of mass cytometry-based cell subpopulation identification which contains 6 peripheral blood mononuclear cells sample and 6 intestinal biopsies samples. **(3)** PTI-FC datasets of flow cytometry-based pseudo-time trajectory inference which contains 6 sequential time points of human embryonic stem cell line HUES9 after the induction of hematopoietic differentiation. **(4)** PTI-MC datasets of mass cytometry-based pseudo-time trajectory inference which contains the peripheral blood mononuclear cells sampled at 7 sequential time points after the activation by pVO4.

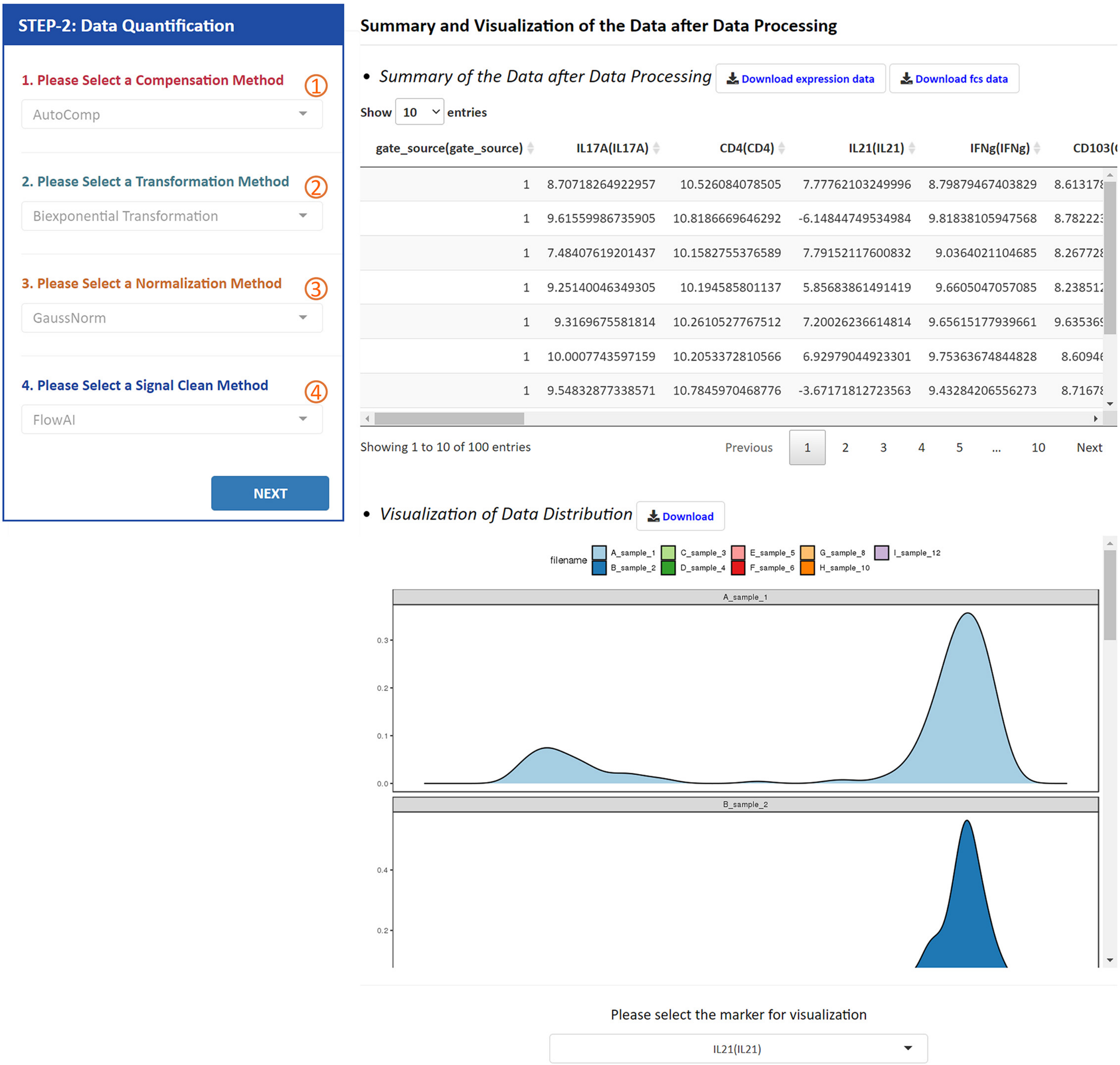
3.2 Feature Selection and Data Preprocessing (Compensation & Transformation & Normalization & Signal Clean)

Quantification of cytometry-based single-cell proteomics data requires a work flow consisting of compensation, transformation, normalization and signal clean.

(1) Data Compensation Both flow cytometry and mass cytometry suffer from signal crosstalk across detection channels which can be corrected by compensation methods that use spillover coefficients for each channel. Compensation method in ANPELA includes AutoSpill, flowCore, MetaCyto and None for flow cytometry data; CATALYST, CytoSpill and None for mass cytometry data. **(2)** Data Transformation FC and MC raw data are often characterized by skewed distribution making it hard for visualization and clustering. Therefore, transformation is adopted in ANPELA 2.0, in order to bring the expression peaks as close to a normal distribution as possible for subsequent data analysis. There are 15 methods for data transformation: Arcsinh Transformation, Asinh with Non-negative Value, Asinh with Randomized Negative Value, Biexponential Transformation, Box-Cox Transformation, FlowVS Transformation, Hyperlog Transformation, Linear Transformation, LnTransform, Log Transformation, Logicle Transformation, QuadraticTransform, ScaleTransform, TruncateTransform and None for both FC and MC data. **(3)** Data Normalization The acquisition time in FC and MC differs, from a few minutes in FC, up to a whole day in MC for barcoded samples, and therefore requires different approaches for normalizing the data. There are 4 methods for data normalization: GaussNorm, WarpSet and None for FC data; Bead-based Normalization, GaussNorm, WarpSet and None for MC data. **(4)** Singal Quality Check and Cleaning In cytometry, signal shift can be caused by several reasons, including tube clog and unstable data acquisition. Therefore, it is necessary to analyze data quality and delete invalid data through different signal-clean methods. Four signal clean methods are adopted in ANPELA 2.0: FlowAI, FlowCut and None for FC data; FlowAI, FlowClean, FlowCut and None for MC data.

A detailed explanation on each compensation, transformation, normalization, signal clean methods is provided in **Section 4** of this Manual. After selecting preferred methods, please proceed by clicking the "**PROCESS**" button, a summary of the preprocessed data will be shown on the left. The resulting data can be downloaded by

clicking the "Download" button.
After the quantification process, please select the protein marker (column name) you want for subsequent process from the drop-down list on the bottom.



STEP-2: Data Quantification

1. Please Select a Compensation Method ①

AutoComp

2. Please Select a Transformation Method ②

Biexponential Transformation

3. Please Select a Normalization Method ③

GaussNorm

4. Please Select a Signal Clean Method ④

FlowAI

NEXT

Summary and Visualization of the Data after Data Processing

- *Summary of the Data after Data Processing* Download expression data Download fcs data

Show 10 entries

gate_source(gate_source)	IL17A(IL17A)	CD4(CD4)	IL21(IL21)	IFNg(IFNg)	CD103(
1	8.70718264922957	10.526084078505	7.77762103249996	8.79879467403829	8.613178
1	9.61559986735905	10.8186669646292	-6.14844749534984	9.81838105947568	8.782223
1	7.48407619201437	10.1582755376589	7.79152117600832	9.0364021104685	8.267728
1	9.25140046349305	10.194585801137	5.85683861491419	9.6605047057085	8.238512
1	9.3169675581814	10.2610527767512	7.20026236614814	9.65615177939661	9.635369
1	10.0007743597159	10.2053372810566	6.92979044923301	9.75363674844828	8.60946
1	9.54832877338571	10.7845970468776	-3.67171812723563	9.43284206556273	8.71678

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 … 10 Next

- *Visualization of Data Distribution* Download

Please select the marker for visualization

IL21(IL21)

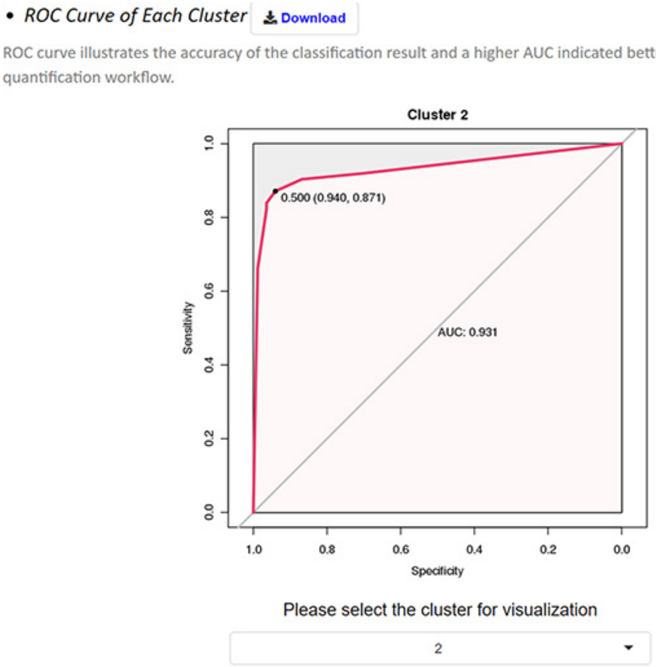
3.3 Evaluation of the Processing Performance from Multiple Perspective

In ANPELA 2.0, both cell subpopulation identification and pseudo-time trajectory inference has four well-established criteria for comprehensive evaluation on the performance of selected quantification workflow.

For **Cell Subpopulation Identification(CSI)**, those criteria includes:

Criterion Ca: ***The Classification Accuracy of Distinct Phenotypes based on Cell Subpopulations***
(Phongpreecha T, *et al. Sci Adv*. 6: eabd5575, 2020)

After cell subpopulation identification, we use KNN to classify cells within each cluster into two different conditions. F1 score or AUC are used to evaluate the accuracy between the classification result and the real condition label.



- *ROC Curve of Each Cluster* Download

ROC curve illustrates the accuracy of the classification result and a higher AUC indicated better performance of the quantification workflow.

Please select the cluster for visualization

2 ▾

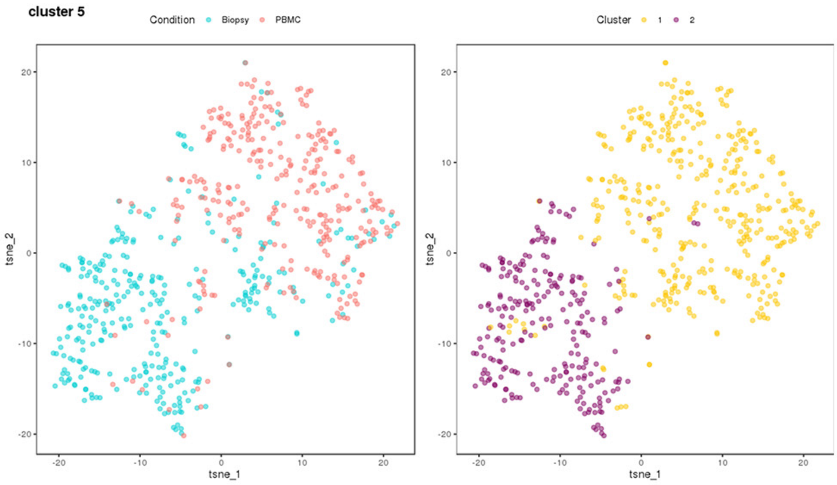
Criterion Cb: ***The Tightness of Clusters with or without Predetermined Manual Labels***

- *External Criterion "**Precision**"* (Jiang H, *et al*. *Bioinformatics*. 34: 3684-3694, 2018)

In order to assess the level of precision of the clusters among two conditions for each cell subpopulation, two well-established measures, purity and Rand index, are calculated by matching the cluster structures and the priori condition information of data.

- *Dimension Reduction Plot Colored by Sample Group and Clustering Information* Download

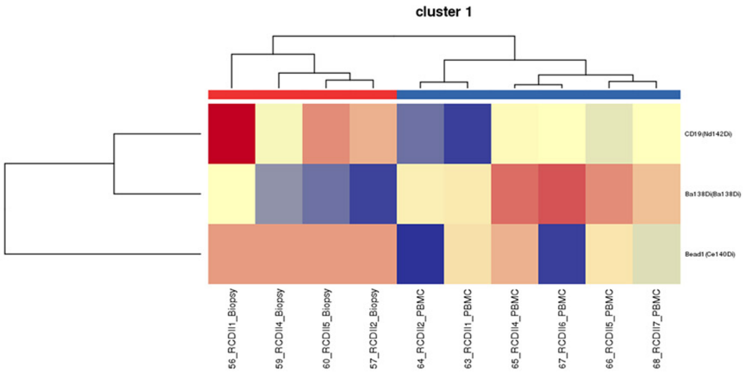
In this plot cells are stain based on the real condition and the clustered condition, a more similar distribution of colors in this two plot indicating a better precision of the quantification workflow.



Please select the cluster for visualization

5

- *Two-way Clustering Plot of Differential Proteins* Download



Please select the cluster for visualization

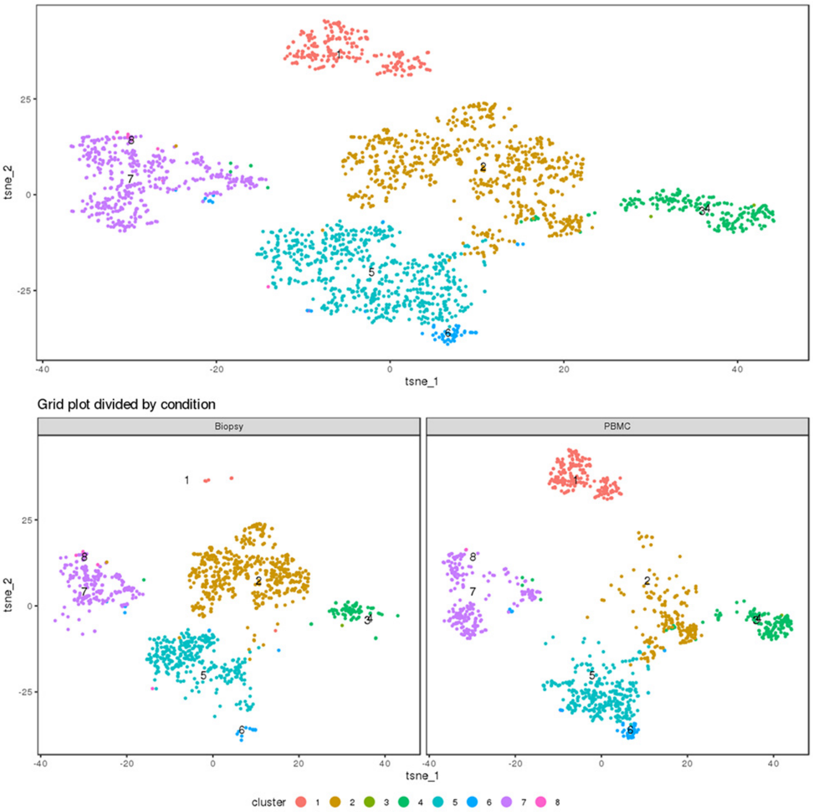
1

- ***Internal Criterion "Coherence"*** (Lee HC, *et al*. *Bioinformatics*. 33: 1689-1695, 2017)

Coherence evaluation is based on the hypothesis that an ideal clustering result should have high similarity within each cluster and high heterogeneity between clusters. Therefore, Silhouette coefficient (SC) which measures how close a datum is to its own cluster compared to the other clusters are used to evaluate the coherence. Similar measurements such as Xie-Beni index (XB), Calinski-Harabasz index (CH), Davies-Bouldin index (DB) are also adopted in ANPELA 2.0.(Lee HC, *et al*. *Bioinformatics*. 33: 1689-1695, 2017)

- *Cluster Distribution Plot* Download

The cluster distribution of each condition are shown in this plot, the density within a cluster and the distance between each clusters indicating the tightness of the quantification workflow.



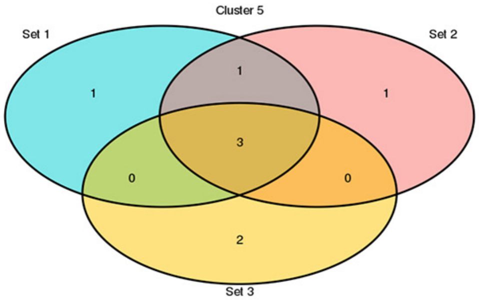
Criterion Cc: ***The Robustness of Identified Biomarkers among Randomly Sampled Subsets***

(Li B, *et al*. *Nucleic Acids Res*. 45(W1): W162-W170, 2017)

After cell subpopulation identification, each cluster is random sampled creating three subset and P-value of each protein between different conditions are used in order to find biomarker within each subset. For each cluster, the consistency score of biomarkers found within three subsets is calculated, in order to evaluate the robustness of the quantification workflow.

- *Venn Plot of Biomarkers Selected from Each Cluster* Download

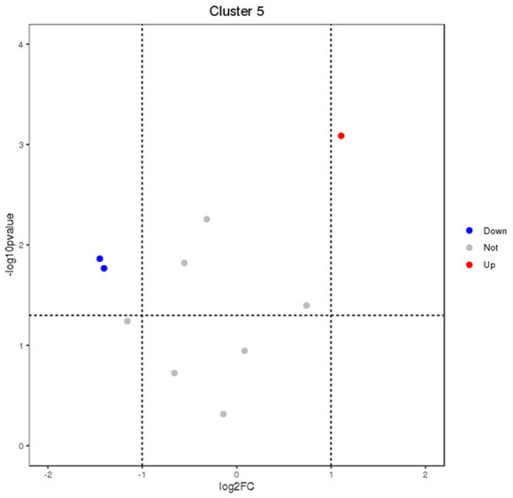
Venn diagram illustrating maker numbers and their overlaps among different subsets and more overlap indicating a better robustness of the quantification workflow.



Please select the cluster for visualization

5 ▾

- *Volcano Plot* Download



Please select the cluster for visualization

5 ▾

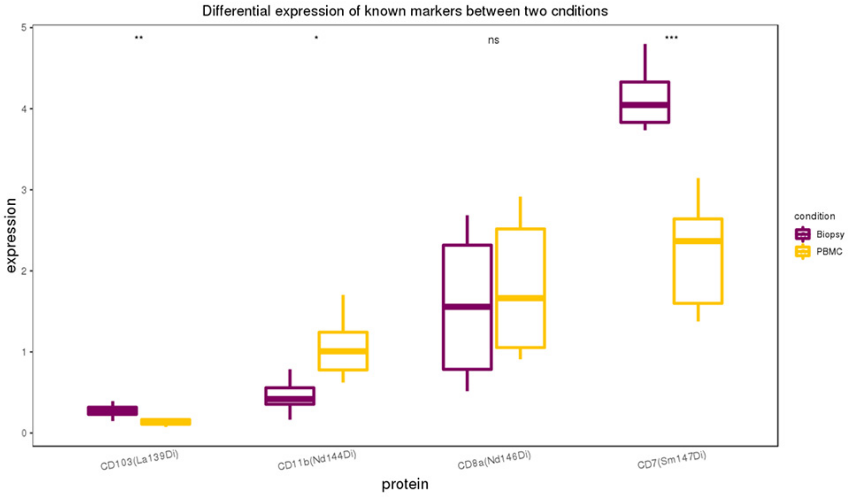
Criterion Cd: ***The Correspondence between Identified Biomarkers and Reliable Reference***

(Suwandi JS, *et al*. *J Autoimmun*. 107: 102361, 2020)

A T-test is conducted in order to find differentially expressed protein between two conditions of all sampled cells. Then the recall of the prior known biomarkers is calculated in order to evaluate the correspondence of the quantification workflow.

- *Boxplot of the Protein Abundance Variations* Download

Satars on the top indicating whether there is differential expression.



For **Pseudo-time Trajectory Inference(PTI)**, those criteria includes:

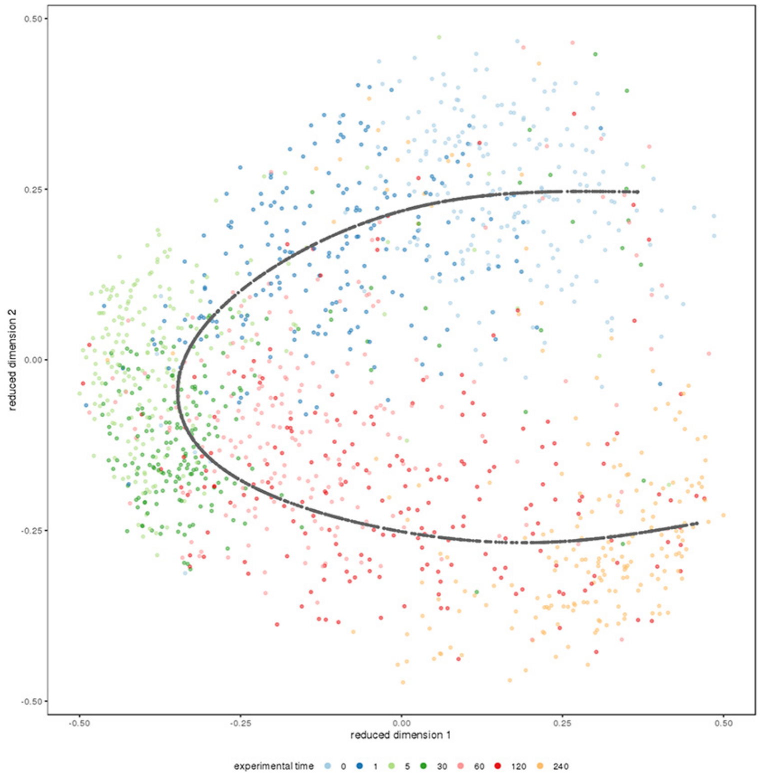
Criterion Ca: ***The Conformance of the Inferred Pseudo-time with Physical Collection Time***

(Ji Z, *et al*. *Nucleic Acids Res*. 44: e117, 2016)

The Conformance of the selected quantification workflow is assessed by comparing the inferred trajectory with the sample collected time. Specifically, we calculated the probability that the order of the two cells in the pseudo-time is consistent with the actual collection time.

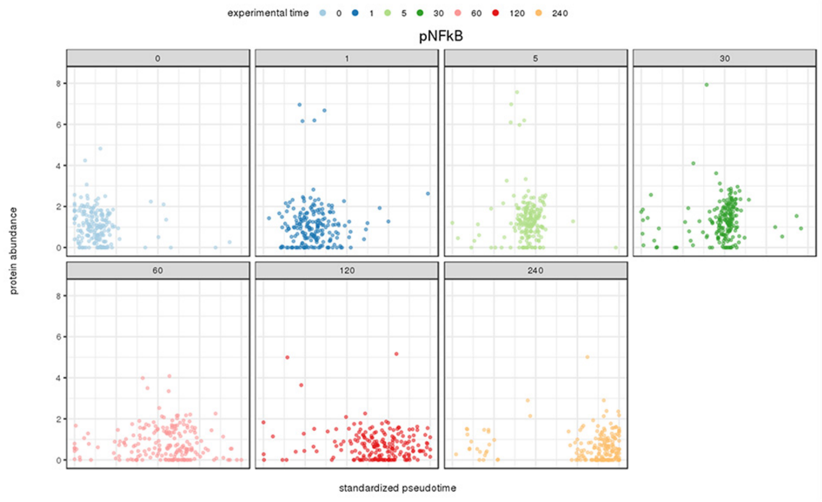
- *Trajectory with Color-coding Cells* Download

Cell dimension reduction map with inferred trajectory in which each cell is colored based on their sampled time. For a well performing workflow, the color order of the cells along the inferred trajectory should be the same as sampled order.



- *Abundances against Pseudo-time Faceted by Real Time* Download

Cells are dived by their sampled time in each plot and each plot represents the protein abundance against pseudo-time. For a well performing workflow, the pseudo-time when most of cells are at peak protein abundance should be conformed with its sampled time.



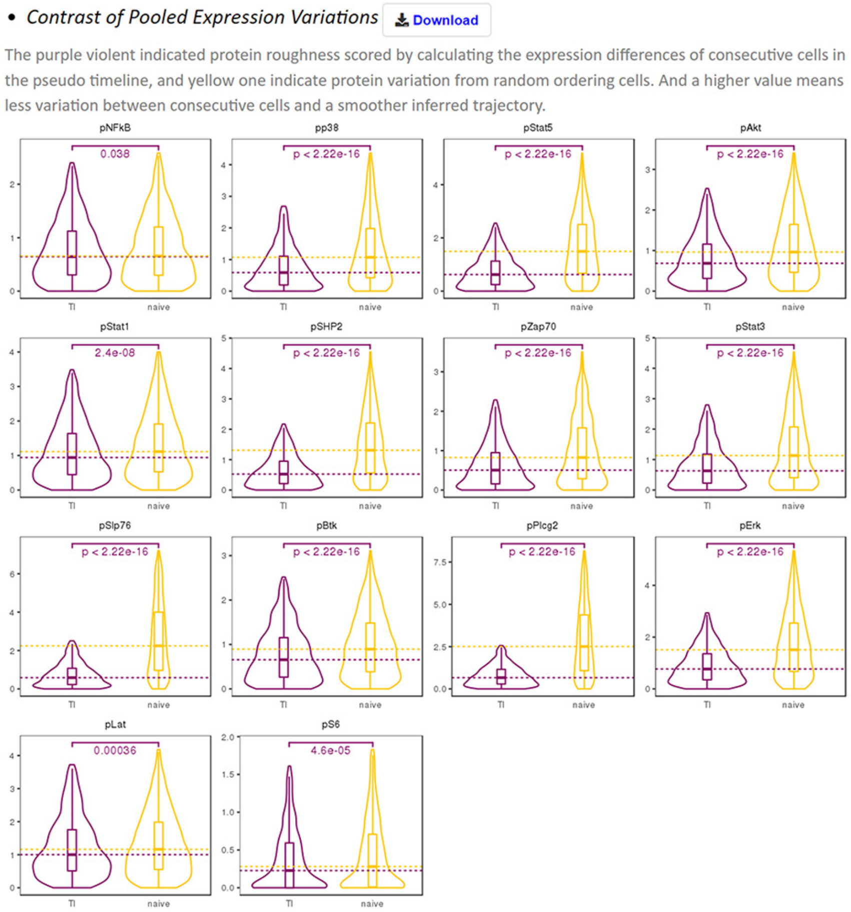
Please select the marker for visualization

pNFkB ▾

Criterion Cb: ***The Smoothness of Protein Levels through the Inferred Trajectory***

(Ahmed S, *et al*. *Bioinformatics*. 35: 47-54, 2019)

The Smoothness of the inferred trajectory for each protein is scored by calculating the expression differences of consecutive cells in the pseudo timeline. The performance of the selected quantification workflow is assessed by p-value which compares all protein smoothness scores between inferred order and random order.



- *Contrast of Pooled Expression Variations*

The purple violent indicated protein roughness scored by calculating the expression differences of consecutive cells in the pseudo timeline, and yellow one indicate protein variation from random ordering cells. And a higher value means less variation between consecutive cells and a smoother inferred trajectory.

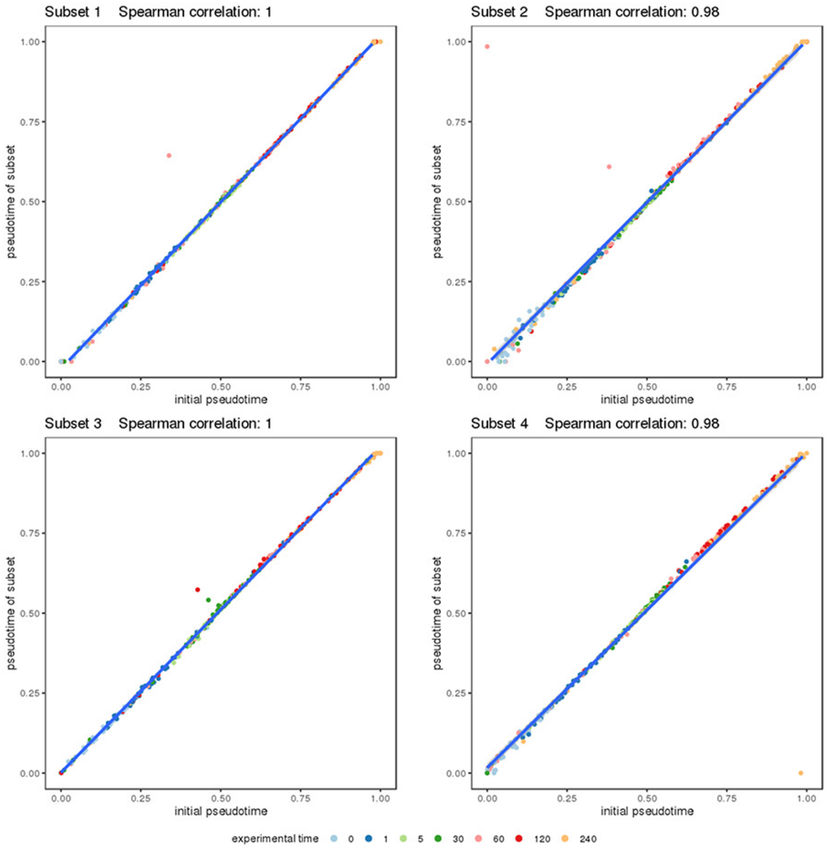
Criterion Cc: ***The Robustness of Inferred Trajectories among Randomly Sampled Subsets***

(Qiu P, *et al*. *Nat Biotechnol*. 29: 886-91, 2011)

In this criterion, four subsets are created by extracting 20% cells from the original data. The selected quantification workflow is applied on the subsets to generate four inferred trajectories. The Spearman rank correlation coefficient or Kendall rank correlation coefficient is calculated by comparing the subsets' trajectory with the original one.

- *Correlation Plot of the Inferred Pseudo-time between Original and Partial Dataset* Download

A spearman correlation plot was conducted for each randomly sampled sub-datasets. The x-coordinate indicate the pseudo-time of the selected cell in the original dataset, the y-coordinate represents the pseudo-time inferred from the sub-dataset, therefor the distribution of cells in this plot should be close to diagonal if the quantification workflow is robust.



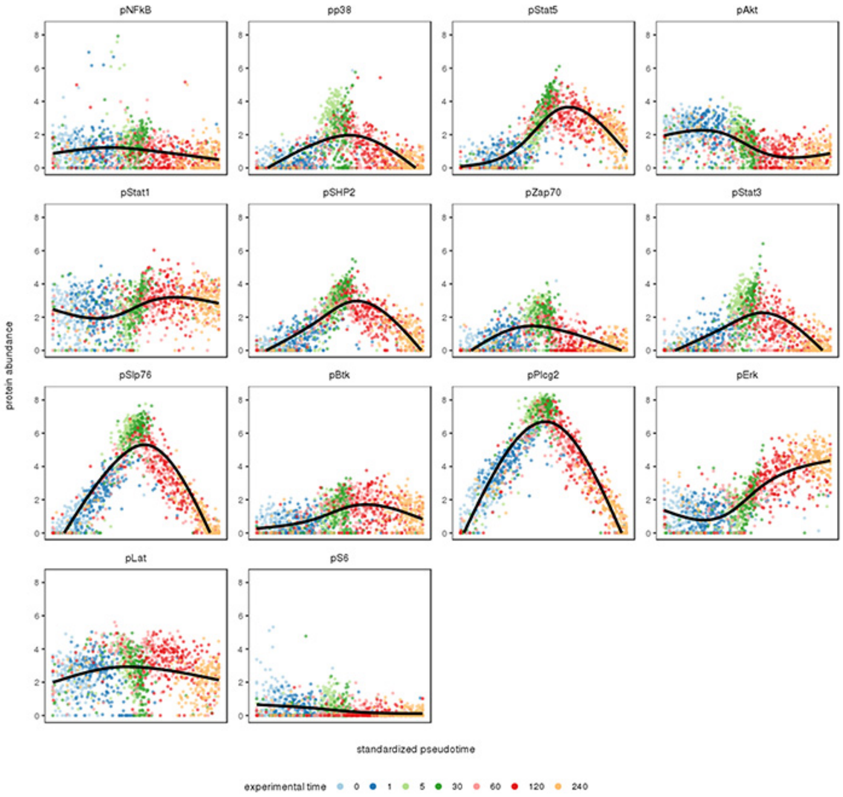
Criterion Cd: ***The Correspondence between Inferred Dynamics and Prior Biological Knowledge***

(Verrou KM, *et al*. *Cytometry A*. 97: 241-252, 2020)

In this criterion, proteins were sequenced according to the order they reach peak expression in pseudo-time. The correspondence score of the selected quantification workflow is calculated by comparing its peak expression sequence with the prior known signal transduction pathway.

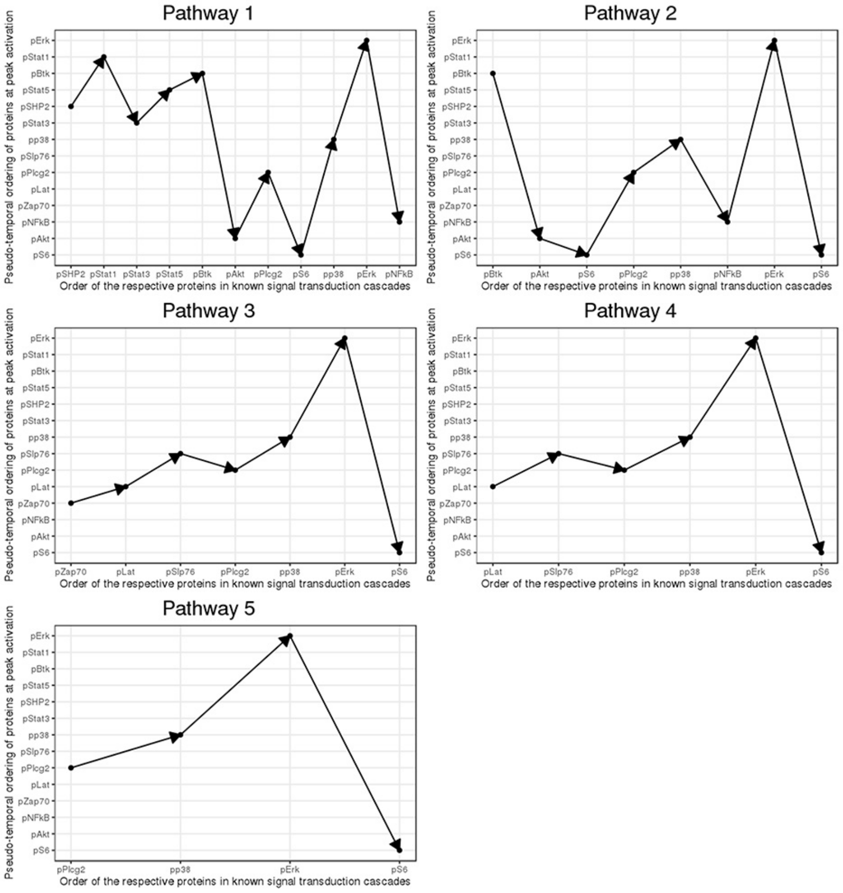
- *Protein Abundance Plot* Download

Each plot indicates a protein expression variation though the pseudo-timeline.



- *Pathway Hierarchy Correspondence Plot* Download

The vertical axis in this plot is the protein order we extract based on the time when protein reach peak expression and the vertical axis as the prior known protein activation pathway. Therefore, an upward arrow means consistent of two proteins which means a better correspondence.



4. A Variety of Methods for Data Preprocessing

4.1 Compensation Methods

- **AutoSpill**. AutoSpill uses single-color controls combine automated gating which calculate spillover matrix based on robust linear regression and iterative refinement to reduce error. (Roca CP, *et al*. *Nat Commun*. 12(1):2890, 2021).
- **CATALYST**. CATALYST is an compensation methods for mass cytometry data that can calculate a spillover matrix based on single-stain beads which are used to compensation mass cytometry data. (Helena L, *et al*. *Bioconductor*. DOI: 10.18129/B9.bioc.CATALYST).
- **CytoSpill**. By finite mixture modeling and sequential quadratic programming achieve optimal error, CytoSpill can quantifies and compensates the spillover effects in Mass cytometry data without requiring the use of single-stained controls. (Miao Q, *et al*. *Cytometry A*. 99(9):899-909, 2021).
- **FlowCore**. Compensation methods from FlowCore can provides an estimation of the spillover matrix based on single-color controls or extract pre-calculated spillover matrix from original FCS by checking valid keywords which are used to compensate the corresponding data. (Hahne F, *et al*. *BMC Bioinformatics*. 10:106, 2009).
- **MetaCyto**. MetaCyto can extract the pre-calculated spillover matrix of each FCS file and use it to compensate corresponding data. (Hu ZC, *et al*. *Cell Rep*. 24(5):1377-1388, 2018).

4.2 Transformation Methods

- **Arcsinh Transformation** . The definition of this function is currently x<-asinh(a+b*x)+c) and is used to convert to a linear valued parameter to the natural logarithm scale. By default a and b are both equal to 1 and c to 0. (Rybakowska P, *et al*. *Comput Struct Biotechnol J*. 18:874-886, 2020).
- **Asinh with Non-negative Value** . This is the suggested methods by Xshift. Before asinh transformation, a specified noise threshold (set at 1) will be subtracted from every raw value and then all the negative values will be set to zero. (Liu X, *et al*. *Genome Biol*. 20(1):297, 2019).
- **Asinh with Randomized Negative Value** . This is the suggested methods by Phenograph. Asinh with Randomized Negative Value is similar to Asinh with Non-negative Value except that negative values are randomized to a normalization distribution rather than set to zero. (Liu X, *et al*. *Genome Biol*. 20(1):297, 2019).
- **Biexponential Transformation** . Biexponential is an over-parameterized inverse of the hyperbolic sine and should be used with care as numerical inversion routines often have problems with the inversion process due to the large range of values that are essentially 0. (Hahne F, *et al*. *BMC Bioinformatics*. 10:106, 2009).
- **Box-Cox Transformation** . Box-Cox transformation is a transformation of non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. (Finak G, *et al*. *Bioconductor*. DOI: 10.18129/B9.bioc.flowTrans).
- **FlowVS Transformation** . FlowVS is a variance stabilization (VS) that removes the mean-variance correlations from cell populations identified in each fluorescence channel. flowVS transforms each channel from all samples of a data set by the inverse hyperbolic sine (asinh) transformation. For each channel, the parameters of the transformation are optimally selected by Bartlett's likelihood-ratio test so that the populations attain homogeneous variances. The optimum parameters are then used to transform the corresponding channels in every sample. (Azad A, *et al*. *BMC Bioinformatics*. 17:291, 2016).

- **Hyperlog Transformation** . The HyperLog transform is a log-like transform that admits negative, zero, and positive values. The transform is a hybrid type of transform specifically designed for compensated data. One of its parameters allows it to smoothly transition from a logarithmic to linear type of transform that is ideal for compensated data. (Bagwell CB, *et al*. *Cytometry A*. 64(1):34-42, 2005).
- **Linear Transformation** . The definition of this function is currently x <- a*x+b and is a basic transformation commonly used in preprocessing cytometry data. (Novo D, *et al*. *Cytometry A*. 73(8):685-692, 2008).
- **LnTransform** . The definition of this function is currently x<-log(x)*(r/d). The transformation would normally be used to convert to a linear valued parameter to the natural logarithm scale. Typically, r and d are both equal to 1.0 and both must be positive. (Hahne F, *et al*. *BMC Bioinformatics*. 10:106, 2009).
- **Log Transformation** . Log is one of the most commly used flow cytometry data transformation method (Arcsinh for mass cytometry data).The definition of this function is currently x<-log(x,logbase)*(r/d). The transformation would normally be used to convert to a linear valued parameter to the natural logarithm scale. Typically r and d are both equal to 1 and both must be positive. logbase = 10 corresponds to base 10 logarithm. (Schoof EM, *et al*. *Nat Commun*. 12(1):3341, 2021).
- **Logicle Transformation** . Logicle transformation creates a subset of biexponentialTransform hyperbolic sine transformation functions which represent a particular generalization of the hyperbolic sine function with one more adjustable parameter than linear or logarithmic functions. The Logicle display method provides more complete, appropriate, and readily interpretable representations of data that includes populations with low-to-zero means, including distributions resulting from fluorescence compensation procedures. (Diggins KE, *et al*. *Methods*. 82:55-63, 2015).
- **QuadraticTransform** . The definition of this function is currently x <- a*x^2 + b*x + c, and has been adopted as a transformation method within FlowCore package. (Hahne F, *et al*. *BMC Bioinformatics*. 10:106, 2009).
- **ScaleTransform** . The definition of this function is currently x = (x-a)/(b-a). The transformation would normally be used to convert to a 0-1 scale. In this case, b would be the maximum possible value and a would be the minimum possible value (Hahne F, *et al*. *BMC Bioinformatics*. 10:106, 2009).
- **TruncateTransform** . In Truncate transformation all values less than a are replaced by a. The typical use would be to replace all values less than 1 by 1 and it is often used to remove fluorescence values < 1 (Hahne F, *et al*. *BMC Bioinformatics*. 10:106, 2009).

4.3 Normalization Methods

- **Bead-based Normalization**. This method first identifies the isotope-containing bead events, converts the raw data to local medians, then the average across all files is computed, these global means is utilized to calculate the slopes for each time point and finally multiplied by all data acquired from corresponding time. (Chevrier S, *et al*. *Cell Syst*. DOI: 10.18129/B9.bioc.flowStats).
- **GaussNorm**. This method normalizes a set of flow cytometry data samples by identifying and aligning the high density regions (landmarks or peaks) for each channel. The data of each channel is shifted in such a way that the identified high density regions are moved to fixed locations called base landmarks. (Hahne F, *et al*. *Bioconductor*. 6(5):612-620.e5, 2018).
- **WarpSet**. WarpSet are normalization method from flowStats package which perform a normalization of flow cytometry data based on warping functions computed on high-density region landmarks for individual flow channels. WarpSet is based on the idea (1) High density areas represent particular sub-types of cells.(2) Markers are binary. Cells are either positive or negative for a particular marker.(3) Peaks should aline if the above statements are true. (Hahne F, *et al*. *Bioconductor*. DOI: 10.18129/B9.bioc.flowStats).

4.4 Signal Clean Methods

- **FlowAI**. FlowAI is an automatic method that check and remove suspected anomalies that derive from (i) abrupt changes in the flow rate, (ii) instability of signal acquisition and (iii) outliers in the lower limit and margin events in the upper limit of the dynamic range. (Monaco G, *et al*. *Bioinformatics*. 32(16):2473-80,2016).
- **FlowClean**. FlowClean track subset frequency changes within a sample during acquisition and reported aberrant time periods as a new parameter added to data file allowing users to exclude those events. (Fletez-Brant K, *et al*. *Cytometry A*. 89(5):461-71, 2016).
- **FlowCut**. FlowCut can identify and delete regions of low density and segments that are significantly different from the rest by calculating eight measures(mean, median, 5th, 20th, 80th and 95th percentile, second moment (variation) and third moment (skewness)) and two parameters （MaxValleyHgt and MaxPercCut）. (Justin Meskas, *et al*. *bioRxiv*. 2020).