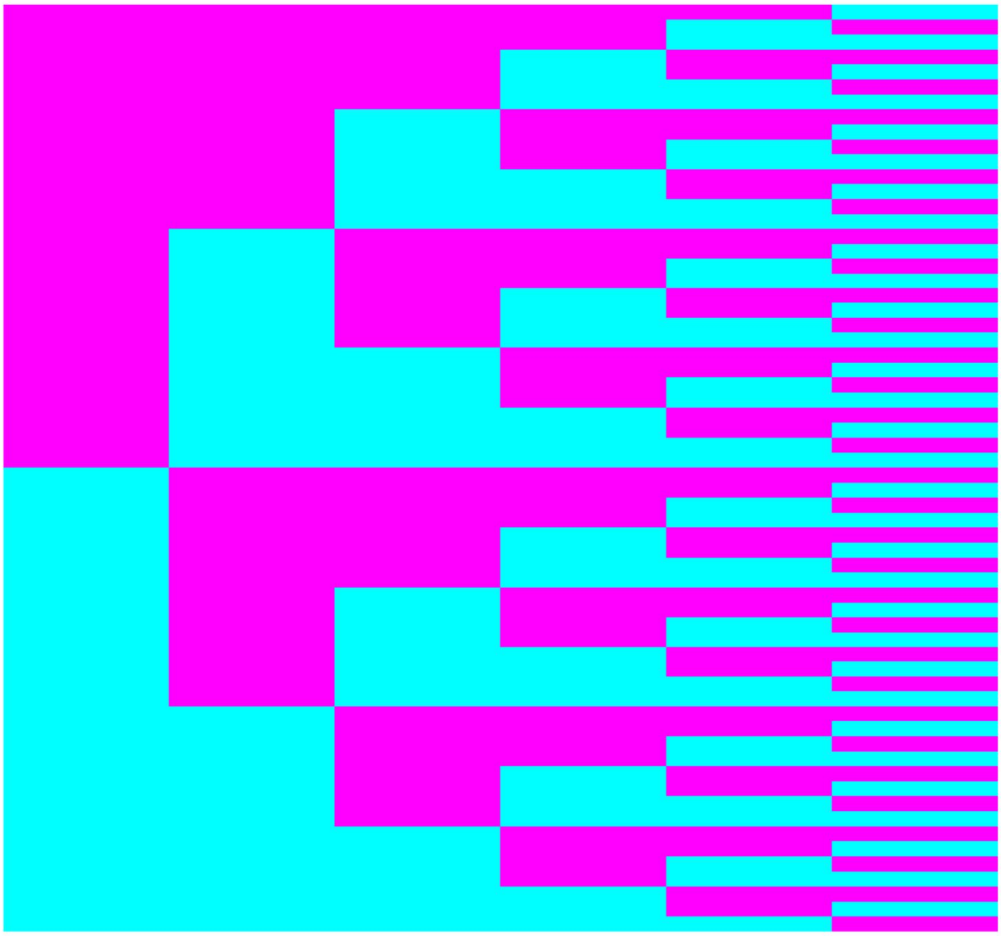
Logical Clustering Suite (v0.5)



by Philipp Niethammer (2023)

Manual (3-32-2023)

Disclaimer: This is an experimental application that did not undergo thorough testing. Therefore, interpret sGCA & sGEA results with due diligence, e.g., perform sanity checks with orthogonal methods, etc. before publishing.

Feedback: To improve the LCS application, your feedback is very welcome. Please provide a sufficient description of issues encountered by e-mail to sgcafeedback@gmail.com.

Citation: Ma Y, Hui KL, Gelashvili Z, Niethammer P. Oxoeicosanoid signaling mediates early antimicrobial defense in zebrafish. Cell Rep. 2023 Jan 31;42(1):111974. doi: 10.1016/j.celrep.2022.111974. Epub 2023 Jan 10. PMID: 36640321; PMCID: PMC9973399.

I. Installation

The LCS application is programmed and compiled in MATLAB and requires download of the MATLAB runtime environment for execution on PCs or MACs. Find the download folder, go to the subfolder “Mac” or “Win” depending on your operating systems. Double click on the “MyAppInstaller_web” application. You need to download MATLAB runtime and approve the license agreement. For further details, please check the “readme” file.

II. LCS Modules

II.1 Simple Gene Correlation Analysis (sGCA) Module

Simple Gene Correlation Analysis (sGCA) by logical clustering assumes that all experimentally meaningful cluster centers are defined by the experimental groups in the experiment, and do not have to be guessed or empirically identified. This contrasts popular clustering methods, such as Kmeans, which make users guess the most likely number of meaningful cluster centers, or density-based methods, such as DBSCAN, that can find clusters without user guesses.

Namely. these latter methods tend to be very sensitive to similarly regulated genes, even if this regulation does not correlate with the experimental groups, for example, and underlying infection signature that is not altered by the experimental conditions. Yet, those signatures that correlate with the experimental conditions are often more interesting and their cluster centers are approximately known. Simple gene correlation analysis (sGCA) assigns each gene expression profile its closest IP. It also calculates the IP-amplitude for each assigned gene expression profile as fold regulation with an adjusted P value for this regulation. By testing the closeness of each gene

	A	B	C	D	E
1	GeneID	GeneSymbol	WtUninjected_1	WtUninjected_2	WtUninjected_3
2	ENSDARG00000000001	slc35a5	118.6489538	121.0635862	128.2835992
3	ENSDARG00000000002	ccdc80	588.2595189	581.1052137	642.3027107
4	ENSDARG00000000018	nrf1	2388.931877	2373.814798	2399.345663

Figure 1. Sample Excel sheet for raw/scaled count input. LCS automatically detects numeric columns as data columns and non-numerical columns as ID columns. LCS requires at least two ID columns, one must be named “GeneID” and the other “GeneSymbol”. Without those sGCA will produce an error. The variable names must be in the format GroupName_ReplicateNumber. Based on this format the groups are automatically detected. If the variable names have a different format, sGCA might nevertheless proceed, but the axis labeling of the plots will be messy. If the counts are not already scaled (as in this example), the `Count Scaling` switch can be turned on.

expression profile to each experimentally possible ideal phenotype, sGCA is hypothesis driven and unbiased at the same time by testing all experimentally reasonable hypotheses. The more groups they are in the experimental design, the more possible IPs exist, so the longer the calculation takes. When comparing only two groups, sGCA is comparable to DESeq2.

Output directory

Select output directory for all analyses of one instance. Overrides all other output paths. Red light, no output directory selected (optional).

Load counts for sGCA

Load Excel sheet with mRNAseq counts or comparable data (Figure 1). Red light, no count sheet selected.

Gene ID columns

By default, the ID column numbers are automatically detected, but can also be edited manually. Red light, no ID column numbers indicated

Experimental groups

By default, the experimental group limits are automatically detected. Correct automatic detection of experimental group limits requires that the data column names are in the GroupName_ReplicateNumber format (Figure 1). Groups can be removed by manual editing. Format: “Group1_StartCol, Group1_EndCol; Group2_StartCol, Group2_EndCol; etc.” Column counting starts at first data column. Red light, no experimental groups indicated.

Count filter

Applies count filter to input count table before sGCA clustering. If ‘on’, the total gene # (displayed in sGEA module) is updated after sGCA completion. Uses MATLAB *genelowvalfilter*, *generangefilter*, and *genevarfilter* at default settings. Please refer to MATLAB manual for details.

Count scaling

Applies *ratio of medians normalization* to input count table before sGCA clustering. For details, please refer to sGCA code on GitHub.

Cluster Data

When button color changes to dark blue, sGCA can be executed. A progress bar will appear. It is recommended that no other actions are performed while sGCA is in progress. Once sGCA is finished and saved (if save option is selected), the light next to the button will switch to green. The sGCA analysis results automatically flow to the downstream heatmap module unless they are replaced by loading a saved sGCA analysis with the Load sGCA button (see below). Note: sGCA processing is relatively slow and processing time considerably increases with the number of experimental groups, i.e., the size of the logical permutation matrix. Consider splitting up large experiments (e.g., experiments with > 10 different experimental groups). Speed optimization may be included in future releases.

II.2 Filtered Heatmap and Boxplots Module

With this module, the sGCA analysis is filtered using the indicated thresholds. The filtered sGCA results and the corresponding group-defined logical permutation matrix are represented as heatmaps. The filtered and sorted sGCA data are saved as sorted EXCEL table. The corresponding heatmap is saved as PDF file. Individual genes of interest may be highlighted in the heatmap. Their profiles may be represented as boxplot.

Load sGCA

Load a saved sGCA analysis for downstream plotting and sGEA, which will override any upstream sGCA analysis. From the parsed sGCA analysis EXCEL and MATLAB files, the Gene ID columns, Experimental groups, and Total gene # (Fisher's exact test in sGEA, see below) are populated. To avoid potential program instability, do not manually alter these parameters. They only affect the sGCA analysis.

Min corr dist

Set correlation distance threshold between 0 (identical) and 1 (opposite). Values between 0.1-0.4 are recommended. For details, please check sGCA code on GitHub.

Min fold reg

Set minimal fold regulation threshold. For details, please check sGCA code on GitHub.

Min meanbase

Set threshold for BaseMean/MeanBase. For details, please check sGCA code on GitHub.

Max padj

Set significance threshold. For details, please check sGCA code on GitHub.

Colormap

Choose colormap for heatmap plots (i.e., logical permutation matrix, filtered sGCA, sGEA). Blue/magenta is default.

Grid

Switch on grid for heatmap of filtered sGCA.

Heatmap

Check to plot heatmap of filtered sGCA. Gene expression values are min/max scaled for each row. For details on the heatmap plot, please check the MATLAB manual. The heatmap will be saved as PDF file if the Save box is checked.

Boxplot

Generate MATLAB boxplots for Genes of interest. These plots are not automatically saved. They can be manually saved by using the figure export options in the title menu of each figure. For details on the boxplot function and its default settings, please check the MATLAB manual.

Genes of interest

Comma separated list of gene symbols (as listed under the GeneSymbol column of an sGCA sheet) to be highlighted in the sGCA heatmap and boxplots.

Filter/Plot Results

When button color changes to dark blue, this executes sGCA-filtering and -plotting with the selected options.

II.3 Simple Gene Enrichment Analysis (sGEA) Module

sGEA determines the significance (using a Fisher's exact test) of enrichment of sGCA IPs or custom marker gene sets in the IPs of a filtered sGCA analysis provided either through the upstream analysis or loaded via the

Load filtered sGCA button. The total gene number for this test is extracted from the upstream or loaded sGCA result and displayed in the Total gene # box. It can be manually edited if necessary.

Load filtered sGCA

Load saved or generated filtered sGCA analysis. Green light indicates that filtered sGCA data is available for sGEA analysis.

Load filtered sGCA|markers

Load a filtered sGCA or custom gene marker list. For details on input format for custom marker lists, please see the included "zclmarkerlist" EXCEL file.

Total gene #

This value is automatically extracted from upstream or loaded sGCA analyses and may be manually edited. It is required for the Fisher's exact test contingency table.

Connect Results

When button color changes to dark blue, this executes sGEA enrichment analysis with the selected options. For details, please see MATLAB code on GitHub.

III. Appendix

III.1 Sample Files Provided with LCS Application

- **CellRep2023_ScaledCounts.xlsx:** Scaled counts from mRNAseq wild type and *hcar1-4* deficient zebrafish larvae subjected to ear infection with *Pseudomonas aeruginosa* under standard, hypotonic bathing conditions. (***Cite:*** Ma Y, Hui KL, Gelashvili Z, Niethammer P. Oxoeicosanoid signaling mediates early antimicrobial defense in zebrafish. Cell Rep. 2023 Jan 31;42(1):111974. doi: 10.1016/j.celrep.2022.111974. Epub 2023 Jan 10. PMID: 36640321; PMCID: PMC9973399.)
- **Immunity2018_ScaledCounts.xlsx:** Scaled counts from mRNAseq wild type zebrafish larvae subjected to ear infection with *Pseudomonas aeruginosa* under standard, hypotonic and isotonic bathing conditions. (***Cite:*** Huang C, Niethammer P. Tissue Damage Signaling Is a Prerequisite for Protective Neutrophil Recruitment to Microbial Infection in Zebrafish. Immunity. 2018 May 15;48(5):1006-1013.e6. doi: 10.1016/j.immuni.2018.04.020. PMID: 29768163; PMCID: PMC6082643.)
- **zclmarkerlist.xlsx:** Cell type marker list derived from scRNAseq of adult zebrafish (Source: https://bis.zju.edu.cn/ZCL/landscape2.html). (***Cite:*** Renying Wang†, Peijing Zhang†,*, Jingjing Wang†, Lifeng Ma†, Weigao E†, Shengbao Suo†, Mengmeng Jiang†, Jiaqi Li†, Haide Chen, Huiyu Sun, Lijiang Fei, Ziming Zhou, Yincong Zhou, Yao Chen, Weiqi Zhang, Xinru Wang, Yuqing Mei, Zhongyi Sun, Chengxuan Yu, Jikai Shao, Yuting Fu, Yanyu Xiao, Fang Ye, Xing Fang, Hanyu Wu, Qile Guo, Xiunan Fang, Xia Li, Xianzhi Gao, Dan Wang, Peng-Fei Xu, Rui Zeng, Gang Xu, Lijun Zhu, Lie Wang, Jing Qu, Dan Zhang, Hongwei Ouyang, He Huang, Ming Chen, Shyh-Chang NG*, Guang-Hui Liu*, Guo-Cheng Yuan*, Guoji Guo* and Xiaoping Han*. Constrction of a cross-species cell landscape at single-cell level. Nucleic Acids Research, 2022. DOI: 10.1093/nar/gkac633. & Mengmeng Jiang†, Yanyu Xiao†, Weigao E†, Lifeng Ma†, Jingjing Wang, Haide Chen, Ce Gao, Yuan Liao, Qile Guo, Jinrong Peng*, Xiaoping Han* and Guoji Guo*. **Characterization of the Zebrafish Cell Landscape at Single-Cell Resolution**. ***Frontiers in Cell and Developmental Biology***, 2021; 9: 743421. DOI: 10.3389/fcell.2021.743421.)

III.2 Getting Started

To familiarize yourself with the Logical Clustering Suite, start with loading **CellRep2023_ScaledCounts.xlsx** into the **sGCA module**. Press Cluster Data and wait until the analysis (~10-20 min depending on machine). The

analyzed counts will flow into the **Filtered Heatmap and Boxplots module.** Press Filter/Plot Results to retrieve the heatmap of these results. Retry, this time checking Boxplot and entering "fosl1a, il1b" (two important inflammatory genes) into the Genes of interest box. Press Close all to close all open windows, then press Reset to reset the LCS. Go back to the sGCA module and load **Immunity2018_ScaledCounts.xlsx.** Press Cluster Data, and then Filter/Plot Results. The filtered sGCA data will flow into the downstream **sGEA module** as indicated by the lamp next to the Load filtered sGCA turning green. Load your previous filtered sGCA analysis via Load filtered sGCA|markers. The lamp next to this box will turn green and the Connect Results button will turn dark blue, which means that sGEA is ready for execution. Press Connect Results and wait until the sGEA heatmap appears and the lamp next to the Connect Results turns green, which means that all results have been saved (given that Save has been checked). The sGEA heatmap shows the exponent of the Fisher's exact p-value. The more significant the enrichment, the more negative the p-value exponent. Note down which clusters show the most significant overlap.

Now, open the unfiltered and filtered sGCA analysis in the output directory. The first tab in the unfiltered GCA EXCEL file shows the logical permutation matrix that you previously also generated as heatmap plot. Note down the ideal phenotype profiles that correspond to the most overlapping IPs from above. Note that the unfiltered sGCA output files start with a unique time/date-stamp (i.e., indicating when sGCA was executed), e.g., "20230330T160439sGCA". The filtered sGCA output shows the same time/date-stamp as the corresponding unfiltered sGCA file and is followed by the filter thresholds [mincorrdist minfoldreg minmeanbase maxpadj], e.g., "20230330T160439sGCA_[0.35 1.5 0.05 10]_THtable". The corresponding heatmap PDF is named "20230330T160439sGCA_[0.35 1.5 0.05 10]_THheatmap". sGEA generates its own date-timestamp. The source files for sGEA analysis are indicated in the "Info" tab of the sGEA EXCEL file. Open the first tab of sGEA output EXCEL file. You will find all enrichments ranked by significance together with a comma separated list of enriched genes. You can, e.g., copy/paste this list directly into downstream analysis applications, such as Metascape (https://metascape.org/gp/index.html#/main/step1).

Retry, loading the **zclmarkerlist.xlsx** via the via Load filtered sGCA|markers button and press Connect Results. The result may indicate possible cell types enriched in the IPs. But note, this is not a real mRNAseq deconvolution approach).

If you rename the sGCA output files (EXCEL and MATLAB), LCS is not anymore able to correctly parse them for analysis. Before you rename them, be sure that you do not need these files for further LCS analysis. Note that there are MATLAB files whose names exactly match the sGCA and sGEA EXCEL files. These files must remain in the same directory for further LCS analysis.

III.3 Accessory MATLAB File Exchange functions used by sGCA and sGEA

- **Centered colormap generator** Version 1.0.0 (1.85 KB) by Timothy Olsen. (***Cite:*** Timothy Olsen (2023). Centered colormap generator (https://www.mathworks.com/matlabcentral/fileexchange/70530-centered-colormap-generator), MATLAB Central File Exchange. Retrieved March 31, 2023.)
- **Date Vector/Number to ISO 8601 Date String** Version 2.0.1 (18.2 KB) by Stephen23. (**Cite:** Stephen23 (2023). Date Vector/Number to ISO 8601 Date String (https://www.mathworks.com/matlabcentral/fileexchange/34095-date-vector-number-to-iso-8601-date-string), MATLAB Central File Exchange. Retrieved March 31, 2023.).

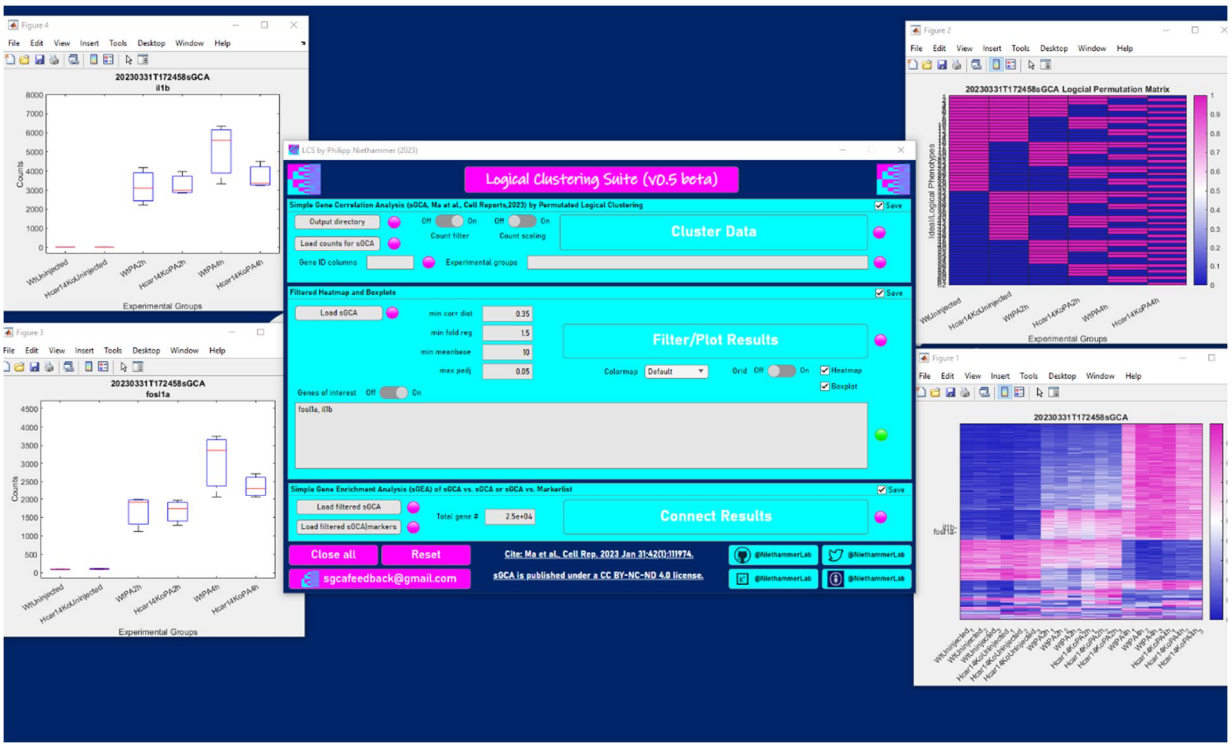
III.4 Troubleshooting and feedback

- This is an experimental application early in its development, so issues are possible, because LCS has not yet been extensively tested. Exercise due diligence when interpreting sGCA/sGEA results.

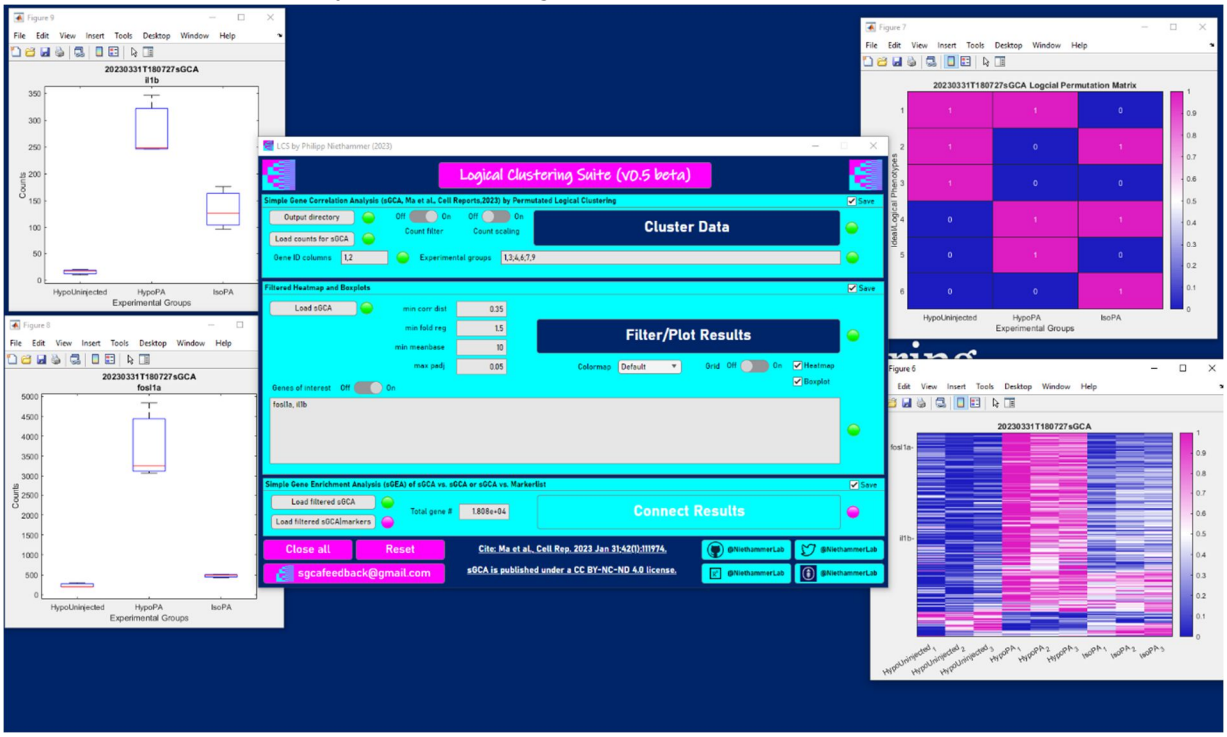
- If you notice this sign, there had been a technical issue. Clicking on the sign will restart the application. For feedback, please describe the sequence of events that caused the issue.
- The author of this app will try his best to respond to all reasonable critique and precise issue reports by improving the app accordingly. All contributions that lead to the improvement of the app will be acknowledged in the user manual of the updated version. Thank you very much in advance!

III.5 Sample screenshots

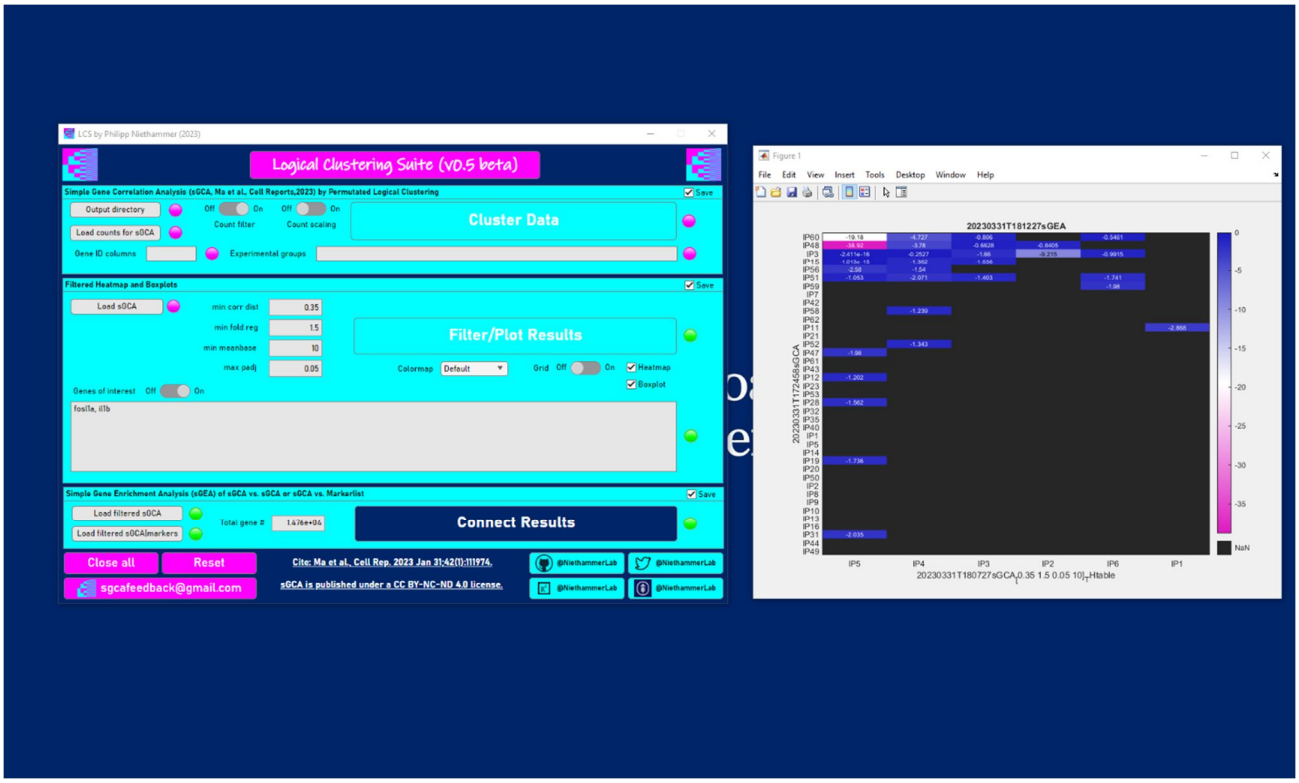
Screenshot 1. sGCA analysis of **CellRep2023_ScaledCounts.xlsx.**



Screenshot 2. sGCA analysis of **Immunity2018_ScaledCounts.xlsx.**



Screenshot 3. sGEA of **Immunity2018_ScaledCounts.xlsx** versus **CellRep2023_ScaledCounts.xlsx.**



III.6 Acknowledgements

Funding: LCS development is funded by a NIH/NIGMS grant R35 GM140883 to Philipp Niethammer.

Contributions: Philipp Niethammer conceived and coded sGCA/sGEA and the LCS.

Feedback contributions:

- **Your name and affiliations could be here. Please give me feedback to improve the app!**
- **…**