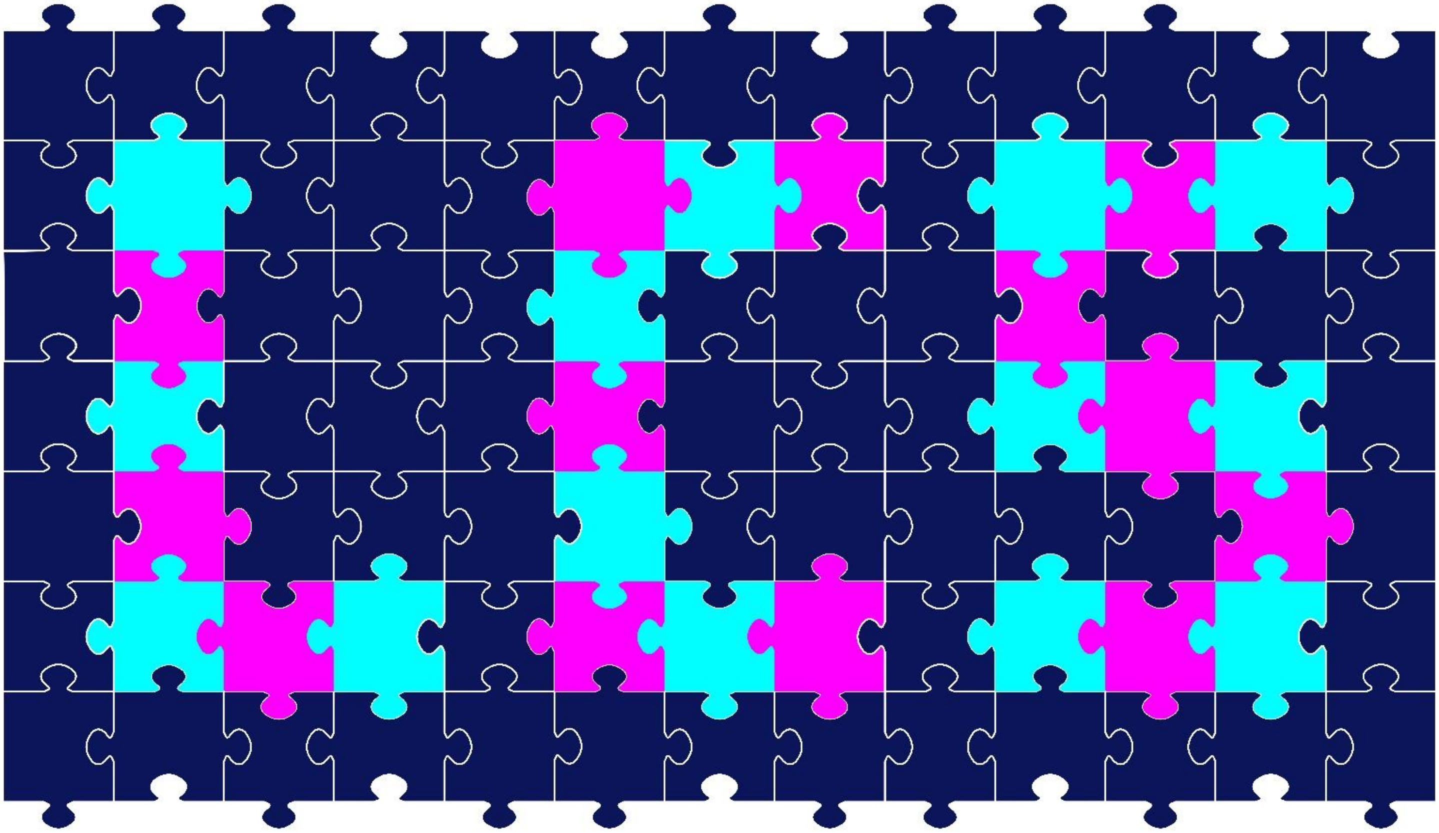


Manual (4-28-2023)
by P. Niethammer

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 standalone execution on PCs or MACs.

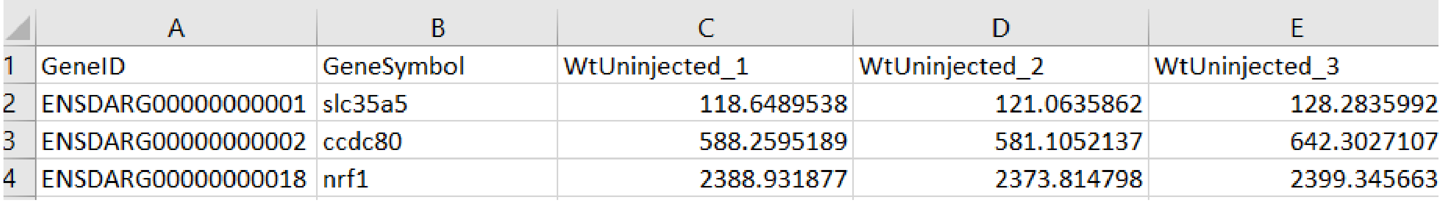
II. LCS Modules

II.1 Logical Clustering (sGCA) Module

Simple Gene Correlation Analysis (sGCA) is the heart of LCS. sGCA clusters gene expression profiles according to their correlation distance to “Ideal (logical) Phenotypes” (IPs). It thus differs from common clustering methods, such as K-means, or self-organizing maps, which cluster the gene expression profiles according to their mutual similarity. In sGCA, all cluster centers are already given through the experimental design and further assumptions are necessary. Since sGCA clusters are defined by the experimental design, they are often more readily interpretable than “unbiased” clusters, that may or may not correlate, with the experimental groups, because their “meaning” is retained. sGCA allows to logically connect and compare different transcriptomics experiments through downstream sGEA enrichment analysis (see below).

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

Load subject counts: Load Excel sheet with mRNAseq counts or comparable data (Figure 1).



	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.

Gene ID columns: By default, the ID column numbers are automatically detected, but can also be edited manually.

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 if they are not correctly detected, or if one wants to remove groups from the analysis. Or different groups can be combined into one. Format: “Group1_StartCol, Group1_EndCol; Group2_StartCol, Group2_EndCol; etc.” Column counting starts at first data column.

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. Automatically switches on if raw counts are detected.

Analyze Subject: When the button 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 analyzed subject** 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).

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 heatmap. The IP clusters are highlighted in the heatmap when the section bars are on. can be annotated with IP section bars. The filtered and sorted sGCA data are saved as sorted table. The corresponding heatmap is saved as PDF file. Individual genes of interest may be highlighted in the heatmap, together with the corresponding IPs. Individual gene expression profiles can be also summarized as boxplot.

Load analyzed subject: 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 extracted.

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.

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

Min meanbase: Set threshold for MeanBase counts. For details, please check sGCA code on GitHub.

IP section bars: Includes Ideal Phenotype Bars to indicate the logical clusters in the heatmap if **Annotation** is switched on.

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 or IPs 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.

Annotation text box: Comma separated list of gene symbols (as listed under the GeneSymbol column of an sGCA sheet) or IPs (or section bars) to be highlighted in the heatmap and boxplots.

Annotation switch: Switch on to display items in the textbox in the graphs.

Filter Analyzed Subject: When the button changes to dark blue, this executes sGCA-filtering and -plotting with the selected options.

II.3 Enrichment Analysis (sGEA) Module

sGEA determines the significance (using a Fisher's exact test) of enrichment of sGCA IPs, sGEA-, or custom marker gene sets (e.g., from scRNAseq, GO terms, etc.) in the IPs of a filtered sGCA analysis provided either by upstream sGCA analysis and filtering or loaded via the **Load filtered subject** button. Also allows to connect multiple sGCA analyses into a co-enrichment network. In this way, any number of mRNAseq experiments can be logically connected. The total gene number for the Fisher 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 subject: Load saved or generated filtered sGCA analysis. Green light indicates that filtered sGCA data is available for sGEA analysis.

Load predicate: Load a filtered sGCA, sGEA marker list, or custom gene marker list. For details on input format for custom marker lists, please check the included “zclmarkerlist” or “”zfGO” EXCEL files.

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 Subject-Predicate: When the button changes to dark blue, this executes sGEA enrichment analysis with the selected options. For details, please see MATLAB code on GitHub. sGEA heatmap are displayed with the chosen heatmap colormap. sGEA plots are not automatically saved. They can be manually saved by using the figure export options in the title menu of each figure.

Build network: Multiselect multiple filtered sGCA sheets to connect them through sGEA co-enrichment analysis into a network. Network plots are not automatically saved. They can be manually saved by using the figure export options in the title menu of each figure.

Graph layout: Select arrangement of the network graph.

Edge weight: Edge width scaling either with Fisher’s exact test p-value (Pval), or fraction of subject genes that overlap with predicate genes (e.g., 0.3 = 30%).

Plot Network: When the button changes to dark blue, this executes plotting of sGEA directed and undirected networks (MATLAB graph and digraph functions), color-coded according to experiment.

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.)

- **zfGO.xlsx:** GO term marker list assembled from zfin.gaf downloaded from http://current.geneontology.org/products/pages/downloads.html.

III.2 Acknowledgements

Helper functions: I would like to acknowledge the authors of the following useful accessory functions obtained from MATLAB File Exchange and integrated into LCS. Cite as:

- **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.).
- Tim Holy (2023). **Generate maximally perceptually-distinct colors** (https://www.mathworks.com/matlabcentral/fileexchange/29702-generate-maximally-perceptually-distinct-colors), MATLAB Central File Exchange. Retrieved April 29, 2023.

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.