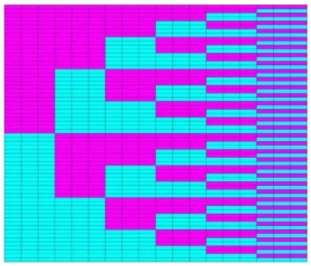
sGCA - quick application note (6-1-2022)

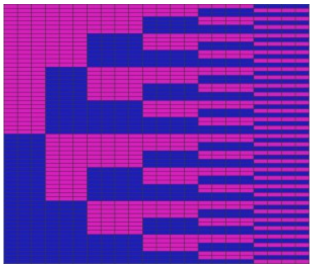
("s" in "sGCA" stands for "simple", "super", or "superfluous")



The ***sGCA*** function filters gene expression count sheets for expression profiles that correlate with experimental group-defined logical gates. Unlike regular gene network analysis (e.g., WGCNA), sGCA disregards profile correlations that are not consistent within experimental groups: For example, random viral contamination of samples would show up as strong interferon signature in regular ("unbiased") gene correlation network analysis, even if it could be considered (depending on the question) as uninteresting signal. In sGCA analysis, alike contaminant signals are less influential as long as they do not correlate with experimental groups. sGCA simplifies to DEseq2 if only two groups are compared (which can be done by picking only two groups from the counts sheets in the UI). Otherwise, sGCA runs the data through all possible (combinatory) logical gates (or "ideal phenotypes", IPs) allowed by the experimental design. As it considers all possible logical gates, sGCA can be considered "unbiased".

Input: Raw or normalized mRNAseq counts (please see powerpoint documentation on Github)

Output: (i) A time-stamped Excel file with all applied logical gates in the first tab, and gene expression profiles as well as their respective closest ideal phenotypes, correlation distance from those, fold-regulation and Padj in the second tab. (ii) MATLAB file with all workspace variables.



The accessory function ***sGCA_Heatmap*** thresholds the sGCA output for similarity to the respective logical gate (or "ideal phenotype") using (i) the MATLAB *correlation distance metric,* (ii) fold regulation between the '0' and '1' groups of the logical gate, (iii) P_{adj}, and (iv) base mean. It then plots all genes that went through the logical gates, with the most abundant IPs first (each cluster is sorted according to increasing correlation distance, decreasing fold-regulation, and increasing P_{adj}). Each gene profile is min-max normalized. There is the option to annotate the genes using a separate Excel list with gene IDs.

Input: sGCA output Excel file. Optional: Excel list with to-be-highlighted genes.

Output: (i) ranked & thresholded sGCA Excel file, (ii) Heatmap (saved as pdf), (iii) text file with all ideal phenotypes and their genes listed—to be used as multi-query for *g:Profiler* functional profiling analysis (https://biit.cs.ut.ee/gprofiler/gost).

It is possible, or even likely, that better methods to achieve the same are already existing. If so, I apologize for reinventing the wheel. After a limited search, I could not find a similar application. Thus, I wrote it myself, and now analyze all my/our mRNAseq data with it, for the better or the worse 😉.

For further details, example application & and citation (*caution: not peer-reviewed yet*):

Ma, Yanan and Hui, King Lam and Gelashvili, Zaza and Niethammer, Philipp, Oxoeicosanoid Signaling Mediates Early Antimicrobial Defense in Zebrafish. Available at SSRN: https://ssrn.com/abstract=4119004 or http://dx.doi.org/10.2139/ssrn.4119004

App & source code available on Github: https://github.com/niethamp/sGCA-Ma-et-al.-2022-

Both sGCA and sGCA_Heatmap are also available as standalone executables (you do not need MATLAB) to use them. If you have MATLAB, you can also install them as taskbar App. Be aware that sGCA is an experimental application that has not been extensively tested and documented yet. Please use it on your own risk.

Philipp