QLUCORE

Pathway analysis using GSEA

TERMINOLOGY
We use **samples** to denote units such as patients, persons, study participants, animals, plants, cells,...

We use **variables** to denote quantities that have been measured for each sample, such as gene expression levels, miRNA expression levels, protein concentrations, antibody concentrations, methylation levels, answers to questions in a questionnaire, ...

Normally a **data set** is described by a matrix where the columns represent samples and the rows represent variables.

FIGURES
Qlucore Omics Explorer is available both in Windows and Mac versions. The figures in this document is based on the Mac version but all principles and controls works in a similar way on Windows.

INTRODUCTION
Pathway analysis, or gene set analysis, is a collective name for methods aimed at statistical analysis of a collection of genes, rather than single genes, in a given data set. Typically, genes are grouped together in a collection (or a gene set) if they have something in common, for example, if they are part of the same biological pathway or if they are all located close to each other along the genome. Given such a collection of genes, gene set analysis is often used to examine whether the expression levels of the genes in the collection are "collectively" significantly associated with a given covariate (for example, whether the genes in the collection are generally differentially expressed between two conditions).

DEFINING THE GENE SETS
To perform gene set/pathway analysis, two components are needed: a gene expression data set, and one or several predefined gene sets (that is, the gene sets should not be defined based on the expression values in the data set). Gene set definitions are often acquired from open online repositories such as MSigDB and Reactome, or from commercial products specialized in providing manually curated pathway information, such as GeneGo, Metacore or Ingenuity. Most commonly, gene sets or pathways are represented as an (unordered) list of the included genes, but in special cases (such as when the gene set represents a biological pathway), more elaborate representations showing also the associations among the included genes are possible.

METHODS
A large number of statistical methods for gene set analysis are available in the literature. Almost all methods treat the gene set under consideration as an unordered set of genes, and thus do not take into account any information regarding the pathway structure (even if it is available). Two methods are arguably dominating gene set analysis in practice: overrepresentation analysis via the hypergeometric test, and Gene Set Enrichment Analysis

(GSEA). The hypergeometric test evaluates whether the overlap between the gene set of interest and a set of "significant" genes (obtained e.g. by applying a statistical test on the gene expression data set) is larger than expected if the genes in the gene set were chosen randomly. In contrast, GSEA ranks all genes in the expression data set by a specified test statistic (e.g., calculated from a statistical test), and evaluates whether the genes in the gene set under consideration are enriched in the top or bottom of this ranked list (e.g., if they are generally strongly associated with a predictor) rather than distributed evenly along the list. One advantage of GSEA compared to the overrepresentation analysis is that the former does not require the user to set a significance threshold, since it considers the ranking of all the genes in the data set. The choice of significance threshold can potentially have a large impact on the results from the overrepresentation analysis.

HOW TO RUN GSEA IN QLUCORE OMICS EXPLORER

In this section we will demonstrate how to run a gene set enrichment analysis in Qlucore Omics Explorer. We will use the example data set (Acute Lymphoblastic Leukemia), which is available from the Help > Example files menu. Before we start the GSEA workbench, we need to do some preparatory work:

- How should we rank the genes? In other words, which statistical comparison are we interested in making? This is just as important for gene set analysis as for single gene analysis, the only difference is that now we will look for whole collections of genes that are significantly associated with the chosen predictor.

- Which gene sets do we want to examine? It is important that these are defined in advance, and not derived from the data set as we are analyzing. As discussed above, there are many online repositories containing predefined gene sets. Qlucore Omics Explorer comes with a few example gene set collections for demonstration purposes, but in practical applications these must usually be complemented with gene sets that are of interest for the actual question. The gene set collections can be provided in either .txt or .gmt format (see reference manual).

- Which variable identifier is used in the gene sets? Since we want to match the variables included in the gene sets with those in the data sets, we must make sure that we are using the same type of identifiers (for example; Affymetrix IDs, gene symbols or Entrez Gene IDs). If the identifiers are not the same, either the IDs in the data set or those in the gene sets must be changed. In the example gene sets in Qlucore Omics Explorer, gene symbols are used as identifiers. The default ID in the data set is the Affymetrix ID, but since the data set contains information also about the corresponding gene symbol, we can easily make the conversion in Qlucore Omics Explorer, by selecting "Gene Symbol" as the Variable Identifier under the "Data" tab. Now, since the mapping between Affymetrix IDs and gene symbols is not one-to-one (sometimes, several probe sets correspond to the same gene symbol), we should also collapse the data set to contain a single value per identifier (gene symbol). This is also done in the "Data" tab (see Figure 1).

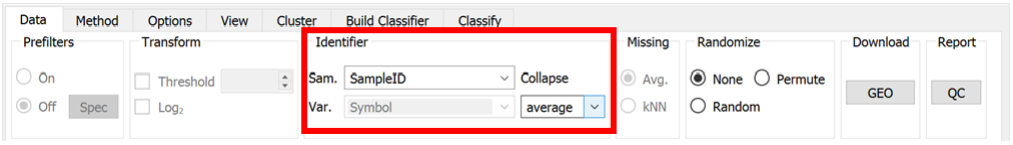


Figure 1. Change variable identifier and collapse values.

Once the data set and the gene sets are prepared, we open the GSEA Workbench from the Button in the Launch section.

In the leftmost panel, we can give the path to the directory where the gene set files are located and select the collections that we want to include in the analysis (Figure 2). In the same panel, we also specify the metric (statistical test) that we would like to use to rank the genes (in this example, a two-group comparison between the T-ALL subtype and the other leukemia subtypes in the data set).

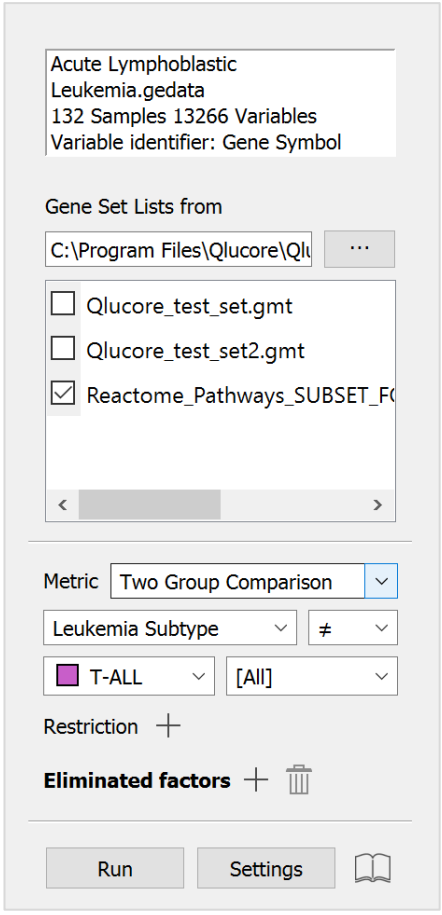
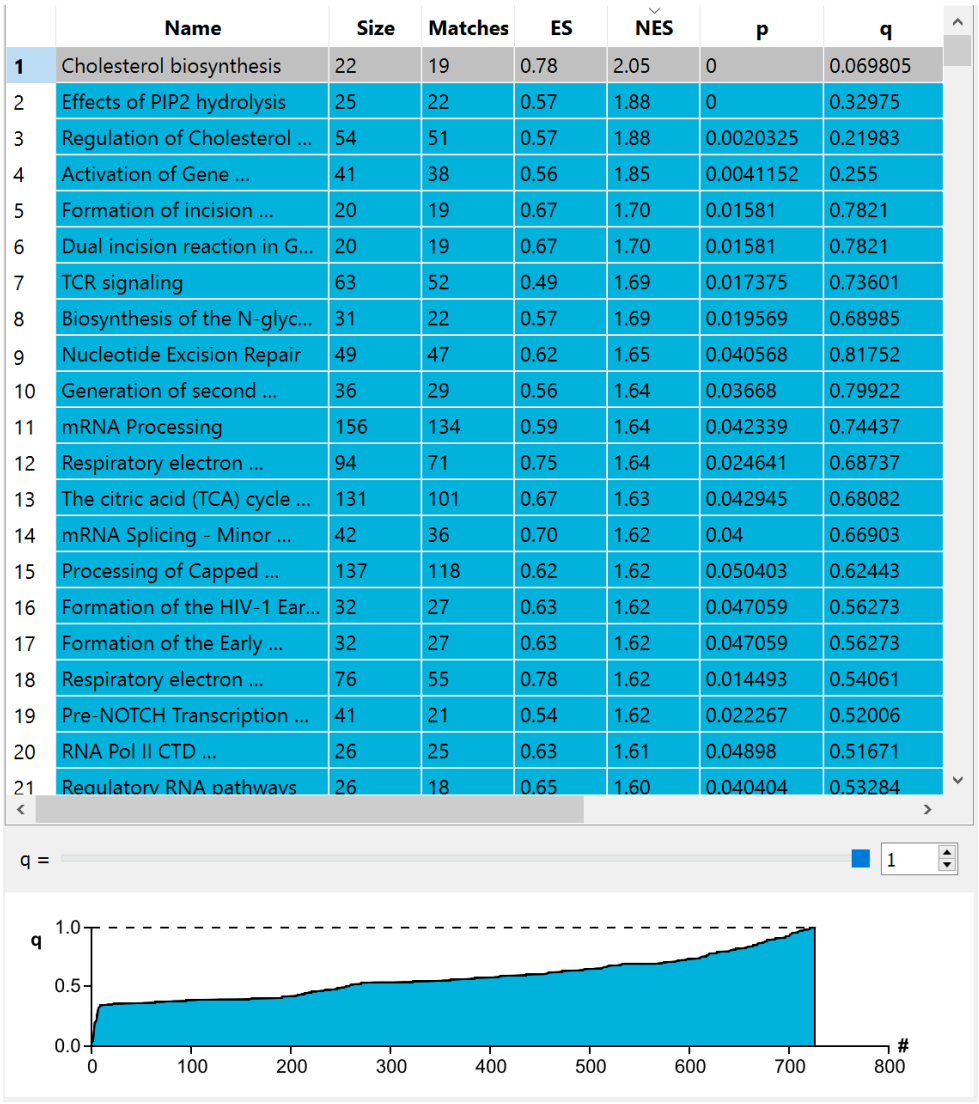


Figure 2. Selection of gene sets and statistical test.

To start the GSEA, press the "Run" button. This launches the calculations, which may take a few minutes depending on the size of the data set and the number of gene sets. Once the analysis is done you will get the results in the middle panel of the GSEA Workbench (Figure 3). This panel contains a list of the gene sets that you have tested, with associated enrichment statistics. You can sort the list by the different columns, but note that with the default sorting is normalized enrichment score. This means that gene sets with a positive enrichment score are located in the top of the list and gene sets with a negative enrichment score are located in the bottom of the list. A positive enrichment score means that a majority of the enriched genes are upregulated and a negative enrichment score means that a majority of the enriched genes are downregulated.



	Name	Size	Matches	ES	NES	p	q
1	Cholesterol biosynthesis	22	19	0.78	2.05	0	0.069805
2	Effects of PIP2 hydrolysis	25	22	0.57	1.88	0	0.32975
3	Regulation of Cholesterol ...	54	51	0.57	1.88	0.0020325	0.21983
4	Activation of Gene ...	41	38	0.56	1.85	0.0041152	0.255
5	Formation of incision ...	20	19	0.67	1.70	0.01581	0.7821
6	Dual incision reaction in G...	20	19	0.67	1.70	0.01581	0.7821
7	TCR signaling	63	52	0.49	1.69	0.017375	0.73601
8	Biosynthesis of the N-glyc...	31	22	0.57	1.69	0.019569	0.68985
9	Nucleotide Excision Repair	49	47	0.62	1.65	0.040568	0.81752
10	Generation of second ...	36	29	0.56	1.64	0.03668	0.79922
11	mRNA Processing	156	134	0.59	1.64	0.042339	0.74437
12	Respiratory electron ...	94	71	0.75	1.64	0.024641	0.68737
13	The citric acid (TCA) cycle ...	131	101	0.67	1.63	0.042945	0.68082
14	mRNA Splicing - Minor ...	42	36	0.70	1.62	0.04	0.66903
15	Processing of Capped ...	137	118	0.62	1.62	0.050403	0.62443
16	Formation of the HIV-1 Ear...	32	27	0.63	1.62	0.047059	0.56273
17	Formation of the Early ...	32	27	0.63	1.62	0.047059	0.56273
18	Respiratory electron ...	76	55	0.78	1.62	0.014493	0.54061
19	Pre-NOTCH Transcription ...	41	21	0.54	1.62	0.022267	0.52006
20	RNA Pol II CTD ...	26	25	0.63	1.61	0.04898	0.51671
21	Regulatory RNA pathways	26	18	0.65	1.60	0.040404	0.53284

Figure 3. The GSEA results panel. The gene sets are ranked by their normalized enrichment scores. The q-value indicates whether the result is statistically significant or not.

Selecting a single list in the result list view will populate the result tabs with the corresponding information. (Figure 4).

The *Score tab* displays three plots: the running enrichment score, the match positions for the genes in the selected data set and the ranking metric values for the input gene list.

The *Heatmap tab* contains a heatmap of the matching genes in the selected list for each of the samples in the input data set.

The *List Details tab* contains a table with data for the matching genes in the selected list. For each gene it shows the gene name, the ranking position of the gene into the ordered input data set, the value of the selected metric (the test statistic), the running enrichment score (RES) and whether the gene belongs to the core enrichment set, i.e. those genes that contribute to the enrichment score for the list.

Selecting multiple lists in the result list view will result in empty Score and List Details tabs, but the Heatmap tab will display the *leading edge overlap* for the selected lists, i.e. a heatmap over the percentage of the genes present in the core enrichment set of one gene set that is also present in another gene set.

To export the results of the GSEA, i.e. all plots and lists for all the selected lists, as well as an analysis summary, to file, click the Results export button. Note that the results for multiple lists can be exported at the same time.

The lists generated by GSEA can also be transferred back into Qlucore Omics Explorer by clicking the *Lists export button*. The selected lists will then appear in the variable list table in the variable panel in the main program.

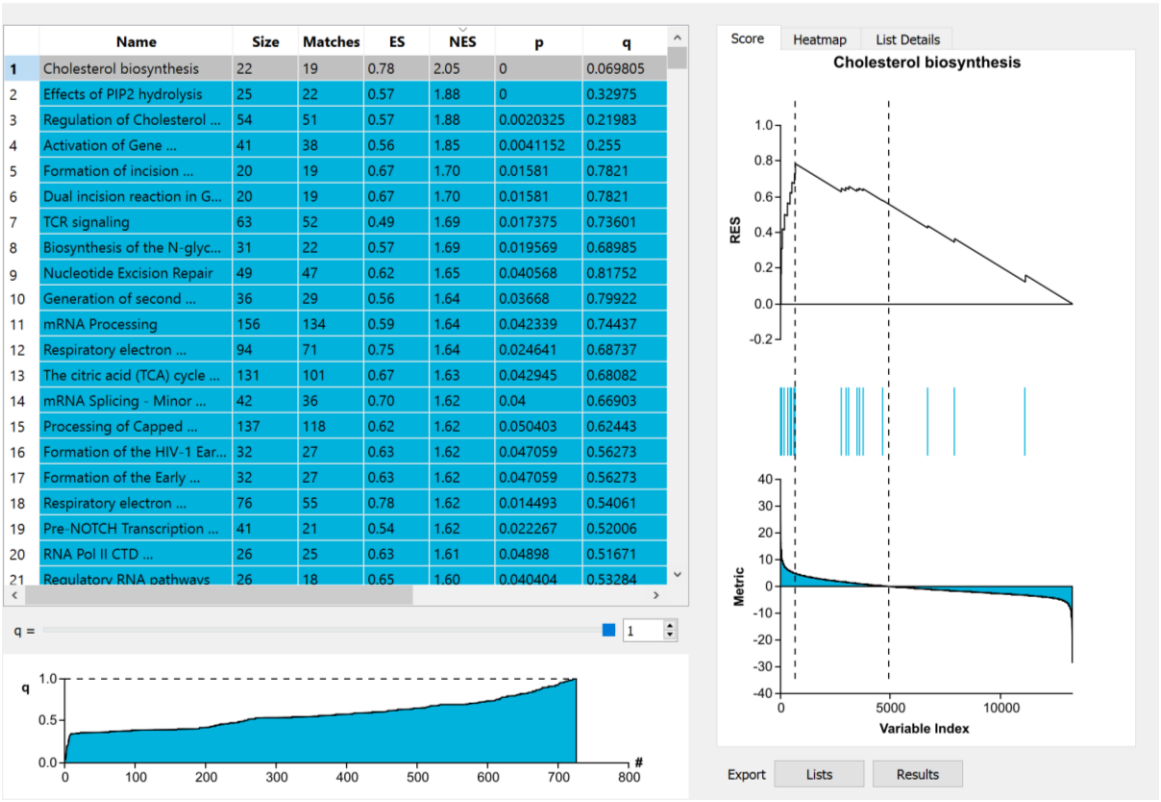


Figure 4. Graphical representation of one gene set (Cholesterol biosynthesis)

VISUALIZING AN EXPORTED GENE SET

When a gene set is exported from the GSEA Workbench using the "Lists" option, it appears as a gene list in the variable list panel in the main Omics Explorer window.

To visualize these variables we can open up a synchronized plot and change that to a variable PCA plot.
Then we filter down to the 600 best genes separating the Leukemia subtype T-ALL from the rest of the subtypes by performing a Two Group comparison (t-test) on T-ALL vs All.
In this example, we exported the top-ranked Cholesterol biosynthesis pathway as a list.

This list in the "Variables" tab can then be used to colour the remaining variables, to highlight which ones that belong to the Cholesterol biosynthesis pathway e.g. coloured yellow.

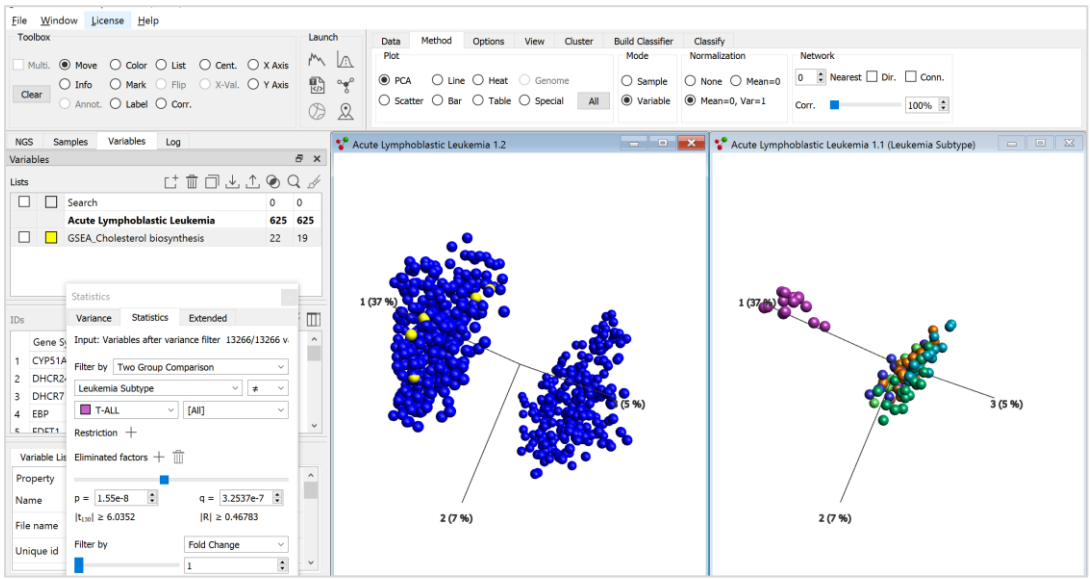


Figure 5. Variable PCA plot coloured to show genes belonging to the Cholesterol biosynthesis pathway (in yellow).

Then this list can be selected as input, by ticking the leftmost box and then only the genes belonging to the selected list will be shown in the plot.

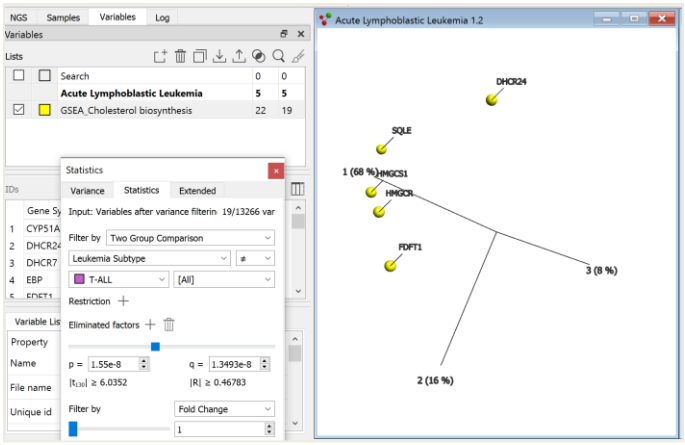


Figure 6. By selecting the list as input the variable PCA plot will only show the genes belonging to the Cholesterol biosynthesis pathway.

DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing. Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document. Qlucore Omics Explorer is only intended for research purposes.

REFERENCES

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. Proc. Natl. Acad. Sci. USA (2005) 102:15545-50.

TRADEMARK LIST