QLUCORE®

Pathway analysis

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 treats 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 statistic (e.g., 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 (View > New GSEA Workbench). 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 statistical test that we would like to use to rank the genes (in this example, a two-group comparison between the E2A-PBX1 subtype and the other leukemia subtypes in the data set).
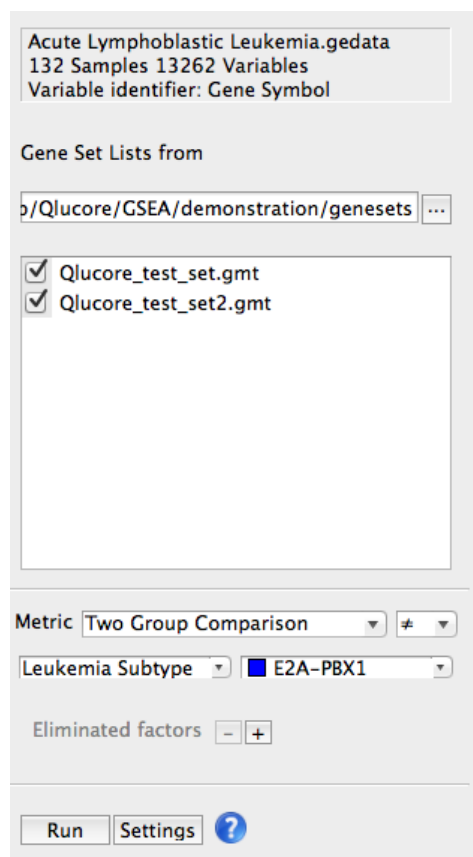

Figure 2. Selection of gene sets and statistical test.

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 sort (by normalized enrichment score) the most significantly upregulated gene sets are located in the top of the list, while the most significantly downregulated ones are located in the bottom of the list.

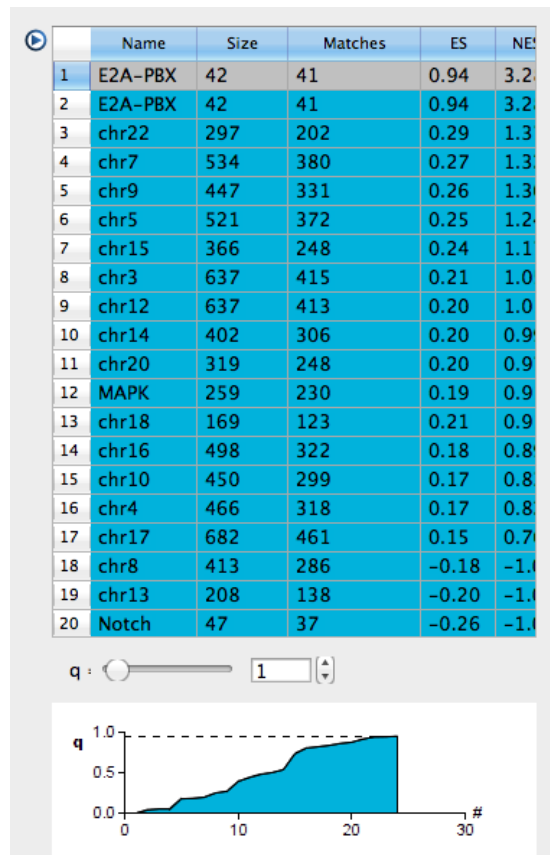|	Name	Size	Matches	ES	NE!
1	E2A–PBX	42	41	0.94	3.2
2	E2A–PBX	42	41	0.94	3.2
3	chr22	297	202	0.29	1.3
4	chr7	534	380	0.27	1.3
5	chr9	447	331	0.26	1.3
6	chr5	521	372	0.25	1.2
7	chr15	366	248	0.24	1.1
8	chr3	637	415	0.21	1.0
9	chr12	637	413	0.20	1.0
10	chr14	402	306	0.20	0.9
11	chr20	319	248	0.20	0.9
12	MAPK	259	230	0.19	0.9
13	chr18	169	123	0.21	0.9
14	chr16	498	322	0.18	0.8
15	chr10	450	299	0.17	0.8
16	chr4	466	318	0.17	0.8
17	chr17	682	461	0.15	0.7
18	chr8	413	286	−0.18	−1.0
19	chr13	208	138	−0.20	−1.0
20	Notch	47	37	−0.26	−1.0

Figure 3. The GSEA results panel. The gene sets are ranked by their normalized enrichment scores, with the most upregulated in the top and the most downregulated in the bottom.

By clicking on one of the gene sets in the list, the right panel shows the detailed results for that gene set in graphical form (Figure 4). The "Score" tab shows the distribution of gene-wise test (ranking) statistics (obtained from the statistical test chosen in the left panel), the position of the genes from the gene set in the global gene ranking, and the calculation of the enrichment score. The other tabs show a heatmap of the genes in the gene set and detailed information about the individual genes in the set. In the third panel, you can also export the results, either directly as gene lists to Qlucore Omics Explorer or as text files.
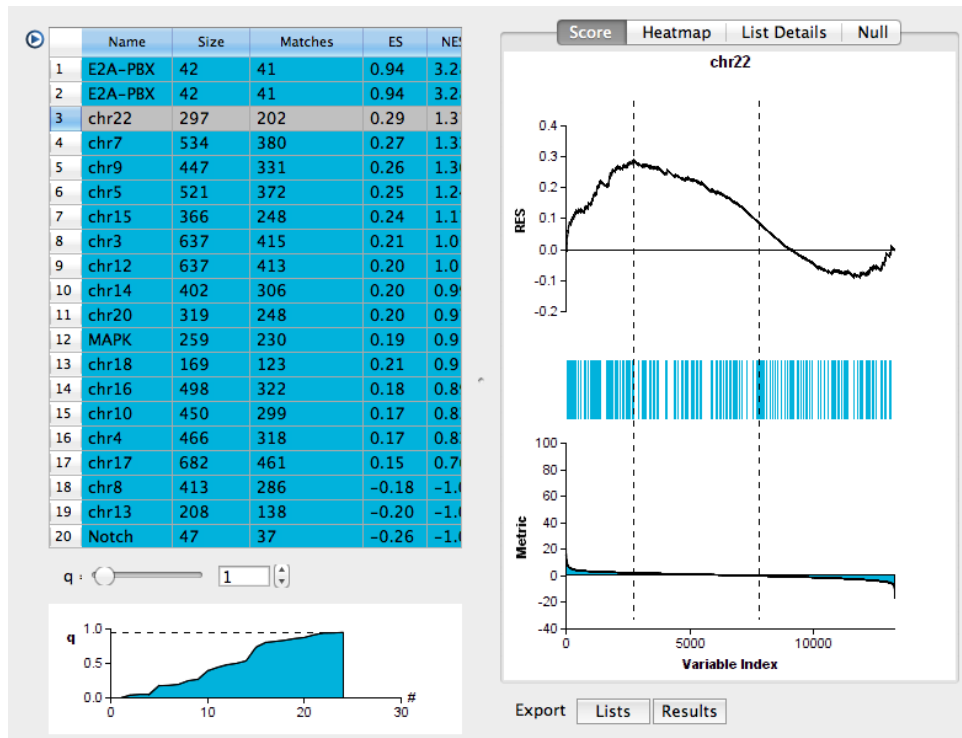
Figure 4. Graphical representation of one gene set (chr22).

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 main Omics Explorer window. We can then, for example, perform a PCA based on only the genes in this pathway. As an example, we export the top-ranked "E2A-PBX" gene set as a list. In the main Qlucore Omics Explorer window, we can then select this gene list in the "Variables" tab, and thus perform a PCA using only the 41 variables in this list. As expected, the "E2A-PBX1" subtype is clearly discriminated from the other samples based on these variables (left panel of Figure 5). We can also color the samples by their average expression level of the 41 genes in the list (right panel of Figure 5), by marking the gene list and clicking "Color Samples" in the Variables tab. Indeed, we see that the expression level of these genes is higher in the "E2A-PBX1" subtype.
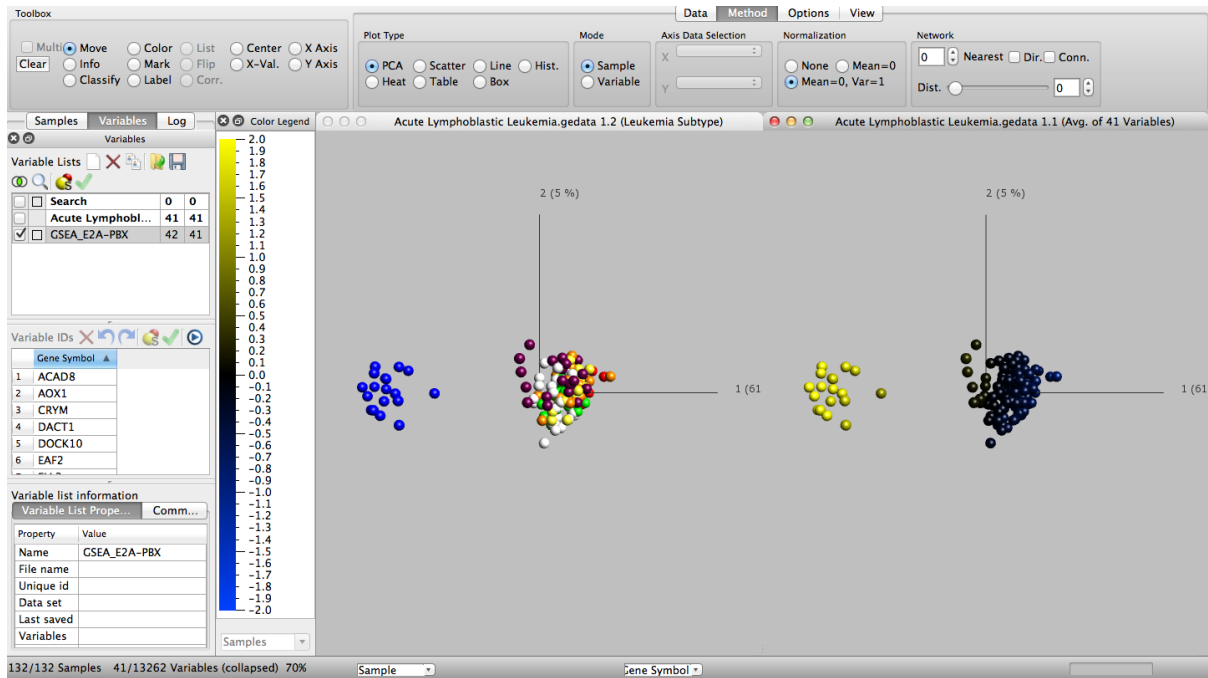
Figure 5. PCA based on only the 41 variables in the top-ranked gene list from GSEA. As expected, the "E2A-PBX1" group is clearly discriminated from the other samples.

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

Excel, Windows 2000 and Windows 7 are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Mac is a trademark of Apple. Metacore is a trademark of GeneGo. Ingeniuty is trademark of Qiagen.