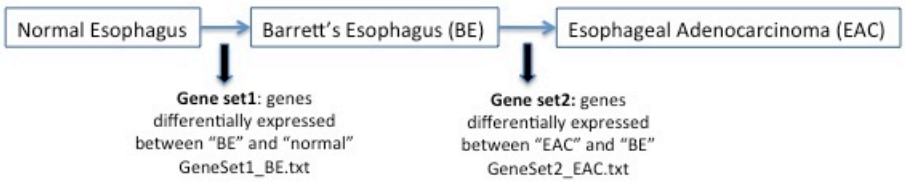
Integrated assignment
Day 5

Goal: Familiarize yourself with g:profiler and its different options and outputs as well as with Enrichment map cytoscape plugin.

Background

Esophageal adenocarcinoma (EAC) is a devastating disease with rising incidence and a 5-year survival of only 15%. The single major risk factor for development of EAC is chronic heartburn, which eventually leads to a change in the lining of the esophagus called Barrett's Esophagus (BE).

Specimens were collected from 8 patients with normal esophagus, 8 patients with Barrett's esophagus and 8 patients with EAC. RNA was extracted from these samples and expression profiling was assessed using Affymetrix HG-U133A microarray. Differentially expressed genes between BE and normal as well as EAC and BE were determined using GEO2R tool.



Our goal is to run the enrichment analysis of each gene set using g:profiler and visualize the output from g:profiler using Enrichment Map Cytoscape plugin.

PART I

1. Open g:profiler;
2. In Options check "Significant only", "No electronic GO annotations";
3. Run analysis of the genes differentially altered between BE and normal: GeneSet1_BE.txt file;
4. What is the most significant GO:term? What is the p-value for this GO:term? What type of Gene Ontology is this GO:term belongs to? (Hint: Biological process (BP) or Cellular Component (CP) or Molecular Function (MF)?)

Answer: "Extracellular membrane-bounded organelle", p-value = 2.05e-6, CC. Warning: it could be different, as databases are constantly updated.

5. Is this p-value already corrected for multiple testing? What type of correction is used for your current analysis?

Answer: Yes, by default g:profiler is using "g:SCS threshold"

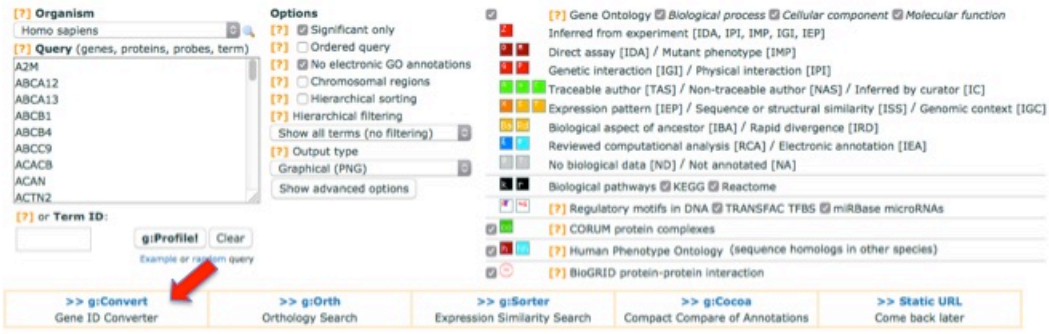
6. Re-run the analysis with “User p-value” threshold set to 0.0001. What has been changed?

Answer: The output list shortened up.

7. Some genes were automatically excluded from the analysis (Hint: look at the warning message at the bottom of the page). Why? What would you suggest to do? (Hint: use g:Convert)

Answer: in some cases the gene names are ambiguous

8. Take one of the genes with several matches in Ensembl. Use g:Convert to fix this problem:



9. Open g:profiler in a new window or tab, scroll down to your gene and find multiple Ensembl ids for it. For example, KRT8 gene is ambiguous with two ensemble ids matches (see below). It is logical to leave behind one corresponding to pseudo gene:

						WIKIGENE
135	KRT20	135.1	ENSG00000171431	KRT20	keratin 20 [Source:HGNC Symbol;Acc:HGNC:20412]	ENTREZGENE, VEGA_GENE, HGNC, WIKIGENE
136	KRT8	136.1	ENSG00000170421	KRT8	keratin 8 [Source:HGNC Symbol;Acc:HGNC:6446]	UNIPROT_GN, ENTREZGENE VEGA_GENE, HGNC, WIKIGENE
136	KRT8	136.2	ENSG00000254285	KRT8P3	keratin 8 pseudogene 3 [Source:HGNC Symbol;Acc:HGNC:31056]	UNIPROT_GN, ENTREZGENE VEGA_GENE, HGNC, WIKIGENE
137	LAMC2	137.1	ENSG00000058085	LAMC2	laminin, gamma 2 [Source:HGNC Symbol;Acc:HGNC:6493]	ENTREZGENE, VEGA_GENE, DBASS3, HGNC, WIKIGENE
138	LDOC1	138.1	ENSG00000182195	LDOC1	leucine zipper, down-regulated in cancer 1 [Source:HGNC Symbol;Acc:HGNC:6548]	ENTREZGENE, VEGA_GENE, HGNC, WIKIGENE

10. Go back to g:profiler, replace ambiguous gene name with its ensemble id and rerun the analysis. Is this gene still on the warning list? Did this correction influence the list of the most significant annotations?

    Answer: (1) No, it is not on the warning list any more. (2) The order of annotations as well as their p-values might be slightly changed.

Suggestion: in the real scientific analysis you would want to fix all warnings.

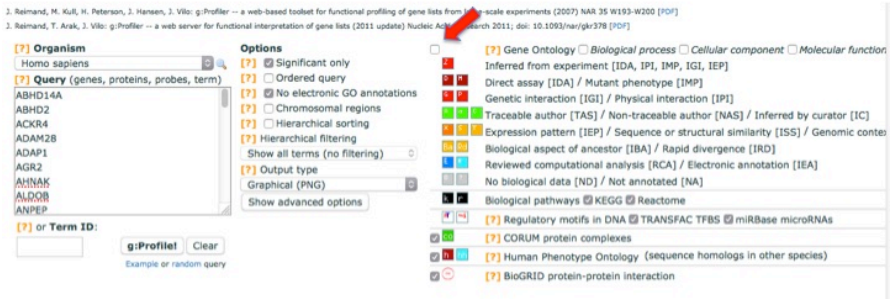
PART 2

An important feature of g:profiler is an ability to work with sorted or ranked gene lists. The head of such a list usually is more informative in determining the functional connections to GO:terms and/or pathways. Our gene list was initially sorted by the q-value based on the significant differential expression.

1. Open g:profiler in a new window and paste genes from Geneset1_BE.txt file;
2. Use options: "Significant only", "Ordered query", "No electronic GO annotations";
3. Do you see any changes in the output in comparison to the analysis of the unordered gene list (PART I)?

Answer: yes, it does, ranked gene list is significantly enriched in more GO:terms/pathways than unordered one.

4. Remove a check from "Gene Ontology" and re-run analysis only using pathways (Introduction to Pathway analysis with be done tomorrow). What pathway is the most significantly altered during progression of the normal epithelium of esophagus to Barrett's esophagus?



Answer: Basigin interactions from Reactome

5. Now we have to generate an output from the enrichment analysis and save it in appropriate format for Enrichment map app. Please, change the output type to "Generic Enrichment Map (TAB)". Use options: "Significant only", "Ordered query", "No electronic GO annotation". Re-run the analysis and download the result file: Download data in Generic Enrichment Map (GEM) format.

6. Click “show advanced options” and download g:profiler data as gmt: name. This file contains the most recent list of GO terms and pathways used by g:profiler. We will need this file for Enrichment map.

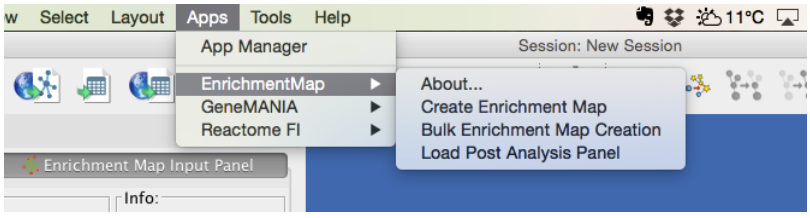
PART III

Generate and save the Generic Enrichment map for genes in GeneSet2_EAC.txt file. Use options: “Significant only”, “Ordered query”, “No electronic GO annotation”. Exclude Gene Ontology from the analysis.

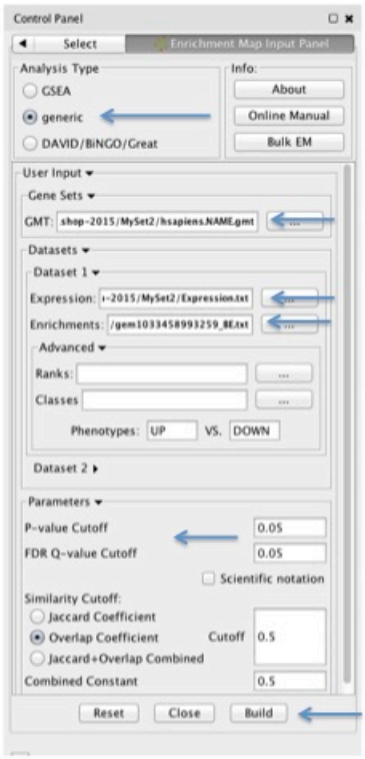
PART IV

The enrichment map cytoscape app allows users to translate large sets of enrichment results to a relatively simple network where similar GO:terms and/or pathways are clustered together.
We will use Enrichment map app to visualize the outputs from g:profiler:

1. Open Cytoscape;
2. Go: Apps -> EnrichmentMap -> Create Enrichment Map



3. Let contract Enrichment map for the pathways that were enriched by the genes differentially expressed between EAC (cancer) and BE. Upload files into app and build the map:



Analysis type is generic

Upload .gmt file: hsapiens.NAME.gmt. This input file was generated in PART II.

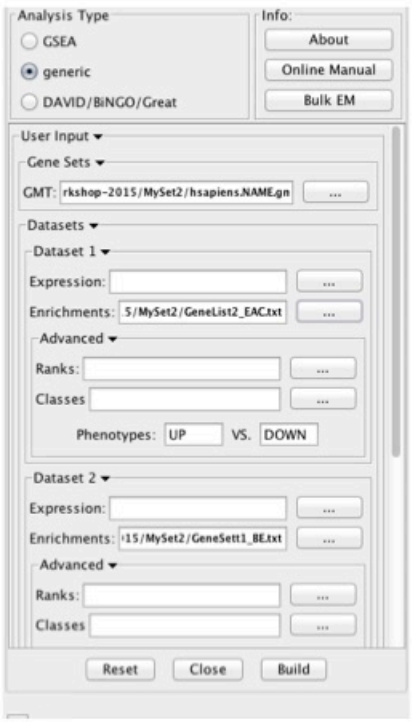
Upload Expression.txt file
Upload Enrichment file for EAC

Enter equal values for P-value and FDF Q-value cutoffs

Press Build

4. If successful, you will see a network where each node represents a pathway and edge connects pathways with shared genes. Node size is proportional to the number of genes in this pathway, intensity of the node color represents the enrichment strength and edge weight – number of genes shared between connected nodes;
5. Try different layouts. For example: Layouts -> yFiles Layouts -> Organic;
6. Change node labels from pathway ids to pathway description (Hint: go to Control Panel -> Style -> Label);
7. Using search box (upper right corner of the cytoscape) find "Epidermal thickening" node (Hint: type in search box "Epidermal"). When node is highlighted, expression profile of all genes included in this pathway appears in the "Heat Map (nodes)" viewer tab. Get familiar with the options provided by this panel. Save Expression Set;
8. Click on any edge (the line between two nodes). In the Table panel you should see a heatmap of all genes both gene sets connected by this edge have in common;
9. Select several nodes and edges. "Heat Map (nodes)" will show the union of all genes in the selected gene sets. Heat Map (edges) will show only those genes that all selected sets have in common;
10. Go to View -> Show Results Panel. Change p-value/q-value as well as similarity cutoffs and see how the network changes;

11. Compare two enrichment analyses. Upload outputs from the g:profiler for BE and for EAC sets. Press Build.



Analysis type is generic

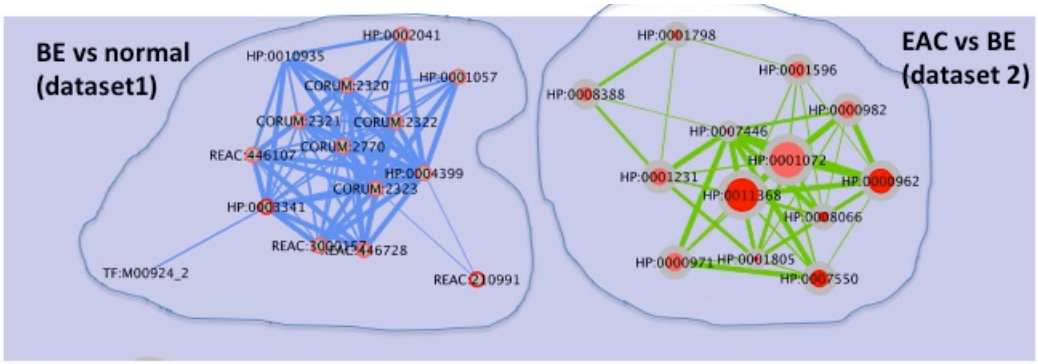
Upload .gmt file

Upload Enrichment file for EAC

Upload Enrichment file for BE

Press Build

What conclusions can you make based on these networks?



Answer: Tissue progression from normal to BE and from BE to cancer are two completely independent process (nodes are not interacting with each other). Changes in dataset 2 are more prominent (color of the nodes are brighter).