FatiGO exercises

Exercise 1: Functional characterization of genes that carry interesting variants obtained from WES experiments.

Objectives: Functional characterization of genes selected from variant discovery experiments in retina dystrophy.

Whole exome sequencing (WES) is useful in de novo variant discovery. After analyze some patients, we get a group of genes with some characterized mutations and through functional enrichment we get additional information that allows improve the understanding of that process they regulate.

Data: 73 genes with variants of subjects with retina dystrophy. File: dystrophy.

Workflow:

1. Open the file “dystrophy.txt” in a text editor and inspect its content.
2. Upload the file in Babelomics through the *Upload* button. We have to specify the *data type*: “Id list (Id)”.
3. Select the functional enrichment analysis tool in the menu “Functional / Single Enrichment: FatiGO”.
4. Firstly, we compare our list against the rest of genome. We have to select our data and organism (human).
5. It´s possible perform simultaneously several enrichment analysis using different databases. In this exercise, we select biological process, molecular functions and cellular components from Gene Ontology.
6. We name the job and execute the task. It may expend about 20 minutes in complete the job.

Questions:

1. How many GO terms are significative? (Biological process, molecular functions and cellular components)
2. How many using FDR < 0.005 as threshold?
3. We are interested in term “retinoid binding” (molecular function). In our list, which genes are enriched with that function?
4. What means sign and value of logarithm of Odds Ratio?
5. How do we interpret the graphic that appears under the significant results?

Exercise 2: Functional characterization of genes up and down expressed in RNAseq experiment.

Goals: To characterize functionally two gene sets that are up or down expressed from RNAseq data analysis where we compare the expression level of two experimental groups: motor VS apoptosis.

Data: We have 105 up-expressed genes in apoptosis (and down-expressed in motor) and 124 genes up-expressed in motor group. They are included in files “apoptosis.txt” and “motor.txt”.

Workflow:

1. Open both files in a text editor and inspect their content.
2. There shouldn't have common genes between them. Check it quickly using http://bioinfogp.cnb.csic.es/tools/venny/ or bash scripting.
3. From *Venny*, why *Venny* numbers doesn't match with the previous numbers (105 and 124)? Do repeated genes alter the functional enrichment results?
4. Upload both files into Babelomics through the *Upload* button. We have to specify the *data type*: "Id list (Id)".
5. Select the functional enrichment analysis tool in the menu "Functional / Single Enrichment: FatiGO".
6. We are interested in several scenarios. Use molecular functions from Gene Ontology.
   a. Functional characterization of up-expressed genes in apoptosis group.
   b. Functional characterization of up-expressed genes in motor group.
   c. Functional characterization of both groups, checking those functions enriched in one group and not in the other. What analysis do we should use? Be careful with repeated genes!

Questions:

1. How many GO term are significant for each situation?
2. We are interested in GO term "Regulation of protein modification process" (molecular function). What gene set has this function over represented?
3. What means sign and value of logarithm of Odds Ratio?
4. What are the differences between p-values and adjusted p-values in significance?
5. How do we interpret the graphic that appears under the significant results?

Exercise 3: Analysis of protein-protein interaction networks of down-expressed genes.

Goals: To characterize functionally one gene set obtained from down-expressed genes comparing young and elder people fibroblasts in a RNAseq experiment.

Functional enrichment of those genes provides extra information that helps to understand those processes where they act and identify new candidates.

Data: After differential expression analysis we identify 47 down-expressed genes (elder population) in fibroblasts from connective tissue. Included in file "fibro.txt".

Workflow:

1. Open the file "fibro.txt" in a text editor and inspect its content.
2. Upload the file in Babelomics through the *Upload* button. We have to specify the *data type*: "Id list (Id)".
3. Select the functional enrichment analysis tool in the menu "Functional / Single Enrichment: FatiGO".
4. Firstly, we compare our list against the rest of genome. We have to select our data and organism (human).
5. It´s possible perform simultaneously several enrichment analysis using different databases. Firstly, we select biological process, and specify those functions annotated with 5 – 300 genes (filtering general and most specific functions).
6. We name the job and execute the task. It may expend about 20 minutes in complete the job.

Questions:

1. How many GO terms are significative?
2. We are interested in term “microtubule organizing center organization (GO:0031023)” In our list, which genes are enriched with that function?
3. What means statistical values that appears in this GO term?
4. How do we interpret the graphic that appears under the significant results? What means unpainted terms? And color graduation?

Exercise 4: Analysis of melanoma differentially expressed genes.

Goals: To characterize functionally one gene set of over-expressed genes comparing melanoma cell lines and melanocytes in a microarray experiment.

Data: After differential expression analysis we select 996 over-expressed genes (melanoma samples) using FDR < 0.05 as threshold, and 4022 with FDR < 0.25. Those data are included in files “melanomaVSmelanocyte_17k.fati996.txt” and “melanomaVSmelanocyte_17k.fati4022.txt”.

Workflow:

1. Open the files in a text editor and inspect its content.
2. Upload the file in Babelomics through the *Upload* button. We have to specify the *data type*: “Id list (Id)”.
3. Select the functional enrichment analysis tool in the menu “Functional / Single Enrichment: FatiGO”.
4. Firstly, we compare our list against the rest of genome. We have to select our data and organism (human).
5. It´s possible perform simultaneously several enrichment analysis using different databases. We select biological process, molecular functions and cellular components from GO.
6. We name the job and execute the task. It may expend about 20 minutes in complete the job. Repeat the analysis with the second file.

Questions:

1. How many GO terms are significant for each file.
2. We are interested in compare both results. Do we get the same results? Do they have the same statistical significance? Why?

GSA Logistic Model exercises:

Exercise 1: Functional characterization of genes in a transcriptomic study (RNAseq technology).

Objective: Detect gene sets with correlated expression and common functional annotations from RNAseq transcriptomic analysis comparing 2 groups: Falconi anemia and healthy population.

Data: We have data of 2726 genes ranked by their differential expression (sick VS control population in file "FA_diffExp.statistic.txt".

Workflow:

1. Open the file "FA_diffExp.statistic.txt" in a text editor and inspect its content.
2. Upload the file in Babelomics through the *Upload* button. We have to specify the *data type*: "Id list (Ranked)".
3. Select the functional enrichment analysis tool in the menu "Functional / Gene Set Enrichment: Logistic Model".
4. It is possible perform simultaneously several enrichment analysis using different databases. Firstly we use cellular components from Gene Ontology.
5. We name the job and execute the task. It may expend about 20 minutes in complete the job.

Questions:

1. How many GO terms are significant?
2. What means sign and value of logarithm of Odds Ratio?
3. What is the difference between p-value and adjusted p-value in significance?
4. How do we interpret the graphic that appears under the significant results? What means the color and intensity of the nodes? Is there any difference between top and bottom nodes in the graph?
5. Modify some nodes and save the graph in vector format (SVG)

Exercise 2: Functional characterization of genes in transcriptomic RNAseq study.

Objective: To detect gene sets with correlated expression and common functional annotations from RNAseq transcriptomic analysis comparing 2 treatments: treatment with fluorouridine VS established treatment.

Data: We select top 1000 differential expressed genes included in data file "fluorouridine.txt".

Workflow:

1. Open the file "fluorouridine.txt" in a text editor and inspect its content.
2. Upload the file in Babelomics through the *Upload* button. We have to specify the *data type*: "Id list (Ranked)".
3. Select the functional enrichment analysis tool in the menu "Functional / Gene Set Enrichment: Logistic Model".
4. It is possible perform simultaneously several enrichment analysis using different databases. Firstly we use cellular components from Gene Ontology.
5. We name the job and execute the task. It may expend about 20 minutes in complete the job.

Questions:

1. How many GO terms are significant?
2. We are interested in GO term "Golgi lumen". What genes are enriched in this function?
3. What means sign and value of logarithm of Odds Ratio?
4. What is the difference between p-value and adjusted p-value in significance?
5. How do we interpret the graphic that appears under the significant results? What means the color and intensity of the nodes? Is there any difference between top and bottom nodes in the graph?
6. Modify some nodes and save the graph in vector format (SVG)

Exercise 3: Functional characterization of genes in transcriptomic microarray study.

Objective: To detect gene sets with correlated expression and common functional annotations from RNAseq transcriptomic analysis comparing expression from melanoma cell lines and melanocytes.

Data: We use ranked expressed genes included in file "melanomaVSmelanocyte_17k.txt".

Workflow:

1. Open the file "melanomaVSmelanocyte_17k.txt" in a text editor and inspect its content.
2. Upload the file in Babelomics through the *Upload* button. We have to specify the *data type*: "Id list (Ranked)".
3. Select the functional enrichment analysis tool in the menu "Functional / Gene Set Enrichment: Logistic Model".
4. It is possible perform simultaneously several enrichment analysis using different databases.
5. We name the job and execute the task. It may expend about 20 minutes in complete the job.

Questions:

1. How many GO terms are significant?
2. What means sign and value of logarithm of Odds Ratio?
3. What is the difference between p-value and adjusted p-value in significance?
4. How do we interpret the graphic that appears under the significant results? What means the color and intensity of the nodes? Is there any difference between top and bottom nodes in the graph?
5. Modify some nodes and save the graph in vector format (SVG)