 

Practical session 3: Pathway analysis

Pathway analysis helps you to interpret the data in a biological meaningful context. In this session we will continue exploring the analysis of microarray gene expression data.

1. Perform Pathway analysis with analyzed microarray dataset using PathVisio
2. Perform Gene Ontology analysis with the analyzed microarray dataset using GOrilla

In this practical, we will use a gene expression dataset that studies the effects of silver nanoparticles on human intestine cell line Caco-2. The paper of the original experiment can be found [here](http://www.ncbi.nlm.nih.gov/pubmed/25997095).

Check the GEO page for more information: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62253>

**Preparation:**

PathVisio software is already installed on your practical computer. PathVisio is an open source free software available at <http://www.pathvisio.org/downloads/>. PathVisio requires a Java environment.

For this application we require the WikiPathways app, which should also be already installed. Apps can be installed in PathVisio clicking on Plugins, PluginManager, select the app and click install.

**Part 1: Pathway analysis**

**Assignment 1: Open pathway in PathVisio**

1. Start PathVisio.
2. Open the human “Statin pathway” in PathVisio.
	1. Plugins, WikiPathways, Search: Statin. Make sure to choose the human pathway.

|  |
| --- |
| **Question 1A.** Small numbers above data nodes, interactions or the info box in the top left of the pathway indicate publication references. Double click the info box in the top left (Title, Availability, Last modified, Organism) and go to the “Literature” tab. **What are the title and authors of the paper reference for this pathway?** |

|  |
| --- |
| **Question 1B.** Click on the DGAT1 gene in the top right. **With which identifier and database is this gene annotated?** (Check the “Backpage” tab on the right side).  |

Load the identifier mapping database:

1. Go to Data → Select Gene Database → Browse to “**Practical 3 data**” folder and load the **Hs\_Derby\_Ensembl\_82.bridge** file
2. Check the status bar at the bottom to see if the gene database has been loaded correctly.

|  |
| --- |
| **Question 1C.** Select the DGAT1 gene again and go to the “Backpage” tab. **Can you now also find the Ensembl identifier(s) for this gene?**  (Required for following steps! If you cant find the information ask a supervisor!) |

**Assignment 2: Data import in PathVisio**

|  |
| --- |
| **Question 2A.** Have a look at the statistical analyzed data (**comp\_Group3-Group1.txt** in **practical 3 data folder**). The first column contains the identifier of the genes. **From which of the three databases below are most of the identifiers in the dataset?** (Required for following steps!)**⎕ Ensembl****⎕ Entrez Gene****⎕ OMIM** |

Import the data as described below. Go through the different dialogs and **before you click “Finish”**, answer the questions at the end!

Description data import:

* Data → Import expression Data
* Select the nanoparticle dataset (**comp\_Group3-Group1.txt** in **practical 3 data folder**) as the input file. Everything else should be filled in automatically (see **Figure 2a**).
* In the next dialog, make sure the correct separators are used. You should see the different columns in the preview (see **Figure 2b**).

|  |  |
| --- | --- |
| Figure 2a | Figure 2b |

* Important: in the next step you need to select the column that contains the gene identifier and the database (system code) for the identifier. Select the database you chose in **question 2A** (Note: if the database is wrong your identifiers will not be recognized correctly), see **Figure 2c**.
* Now the data is imported (see **Figure 2d**). **Before clicking “Finish”** answer the questions 2B and 2C below:

|  |  |
| --- | --- |
| Figure 2c | Figure 2d |

|  |
| --- |
| **Question 2B.** **How many rows were successfully imported?** |

|  |
| --- |
| **Question 2C.** **How many identifiers were not recognized?** **What does that mean? Important**: if the number of rows is the same as the number of identifiers not recognized the data import was not done correctly - you probably didn’t select the correct database! Redo the import or ask one of the instructors for help. (Required for following steps!) |

If you click finish, you should see a default visualization on the pathway (**if all genes are gray, the data import was not successful**). If there is no pathway open, you can check the status bar at the bottom where the dataset will be listed.



**Assignment 3 Creating a basic visualization**

Follow the instruction to create a basic visualization:

1. Go to Data → Visualization Options
2. Create a new visualization named “basic visualization”



1. Select “Expression as color” and “Text label”.
2. In “Expression as color” select “Basic”.
3. Check the checkbox before “logFC” and define a new color set.



1. Select “Gradient” and define a gradient from -1 over 0 to 1 (blue - white - red) → Click OK.



|  |
| --- |
| **Question 3A.** Make a screenshot of the pathway. **What do the colors in the pathway mean biologically?** (Hint: Check the “Legend” tab on the right side).  |

|  |
| --- |
| **Question 3B.** Select the HMGCR gene (top left), go to the “Data” tab. **What is the logFC of the HMGCR gene?** |

**Assignment 4 Creating an advanced visualization**

PathVisio also allows users to visualize multiple data columns together. For that we need to create a new advanced visualization.

1. Go to Data → Visualization Options
2. Create a new visualization named “advanced visualization”



1. Select “Expression as color” and “Text label”.
2. In “Expression as color” select “Advanced”.
3. Check the checkbox before “LogFC” and define a new color set, see **Figure 3a.**
4. Select “Gradient” and define a gradient from -1 to 0 to 1 (blue - white - red) → Click OK.
5. Check the checkbox before “P.Value” and define a new color set.
6. At the bottom, click on “Add Rule”. Go to the text field next to “Rule Logic” and specify the following criteria: [P.Value] < 0.05. Then click on color and select a light green. Click OK and OK, see **Figure 3b**.

|  |  |
| --- | --- |
| Figure 3a | Figure 3b |

|  |
| --- |
| **Question 4A.** Make a screenshot of the pathway. **What do the colors in the different columns on the data nodes in the pathway mean biologically?** (Hint: Check the “Legend” tab on the right side).  |

|  |
| --- |
| **Question 4B.** **How many significant genes (P.Value < 0.05) are in the pathway?** |

**Assignment 5 Perform pathway statistics**

To identify pathways that might be affected after vaccination, you can perform pathway statistics to calculate Z-Scores for each pathway (check lecture!). PathVisio automatically ranks the pathways based on the Z-Score.

1. Go to Data → Statistics
2. First we need to define a criterion for differentially expressed genes. We are going to select those genes based on significant p-value but we are also going to make sure the change is high enough by specifying a log fold change threshold:
	1. ([LogFC] < -1 OR [LogFC] > 1) AND [P.Value] < 0.05

|  |
| --- |
| **Question 5A.** **Explain in your own words what this expression criteria means (which genes will be selected)?**([LogFC] < -1 OR [LogFC] > 1) AND [P.Value] < 0.05 |

1. Now we need to specify the pathway directory. In the Practical 3 data folder you can find the folder **Hsa pathway collection**
2. Browse to this directory and select it.
3. Then click on Calculate and wait for the result table.

|  |
| --- |
| **Question 5B.** **What are the top 5 altered pathways and what are their Z-Scores?** |

|  |
| --- |
| **Question 5C.** **How many genes of the dataset are in at least one pathway (N) and how many differentially expressed genes of the dataset are present in at least one pathway (R)?** (Check “N and R" above the result table) |

|  |
| --- |
| **Question 5D.** **What is the pathway with the lowest Z-Score? What does a low Z-Score mean biologically?** (ignore pathways with NaN) |

**Part 2: Gene Ontology analysis**

Gene Ontology (GO) provides a structured and controlled vocabulary to describe gene products in terms of their associated biological processes, cellular components and molecular functions. One of the main uses of GO is to perform functional enrichment analysis on gene sets.

For GO analysis, we need two different gene lists:

1. **Target gene set** - we select the list of changed genes in the dataset (logFC > 1 OR < -1) AND p-value < 0.05) - “**changed genes list.txt**” in the Practical 3 data folder
2. **Background set** - complete list of measured genes in the dataset - “**background gene list.txt**” in the Practical 3 data folder

**Assignment 6 Perform gene ontology analysis with GOrilla**

1. Go to <http://cbl-gorilla.cs.technion.ac.il/>.
2. Step 1: Select “Homo sapiens”
3. Step 2: Select “Two unranked lists of genes (target and background lists)”
4. Step 3: Upload files for target genes set and background set:
	1. “changed gene list.txt” and “background gene list.txt”
5. Step 4: Select “Process”
6. Click “Search Enriched GO terms”

|  |
| --- |
| **Question 6A.** You will now see the GO tree with the p-values colored on the process nodes. If you scroll down you can also see a table of enriched GO processes. **What are the top 10 GO classes?** |

|  |
| --- |
| **Question 6B.** **Can you find nanoparticle related processes** (give some example of processes that might be related to stress)?  |

|  |
| --- |
| **Question 6C.** **Do you see overlap between the pathway statistics result (Q5B) and the GO analysis result (Q6A)?** If yes, give one example. |