

Hands-on Tutorial: Pathway analysis in PathVisio

We will use PathVisio to visualize a lung cancer transcriptomics dataset (comparison lung cancer vs. healthy) and perform pathway statistics to find pathways that are up- or down-regulated in lung cancer.

Step 1: Open pathway in PathVisio

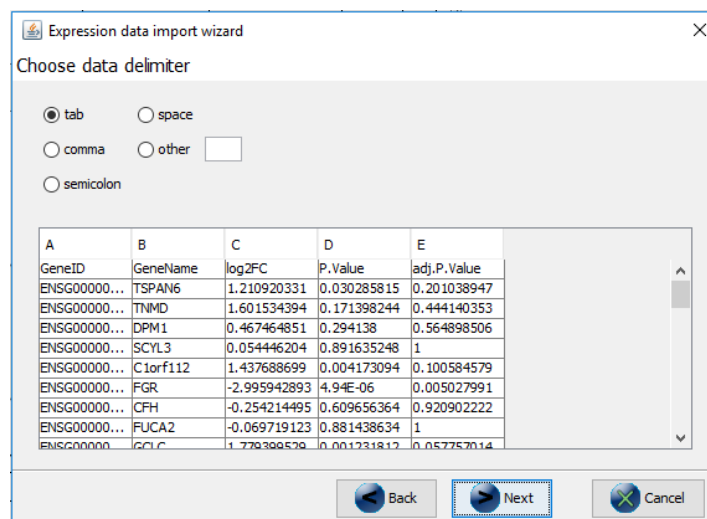
- Start PathVisio
- Open the cell cycle pathway in PathVisio (as an example pathway)
 - Go to File → Open → Browse and select **cell-cycle.gpm1** file (in data folder that you downloaded before)
- Load the human identifier mapping database from BridgeDb
 - Go to Data → Select Gene Database → Browse to **Hs_Derby_Ensembl_91.bridge**
 - Check the status bar at the bottom to see if the gene database has been loaded correctly (nothing else will change in the program)

Question 1: Computer readable annotated pathway models

- a) Click on the *GSK3B* gene in the top left. In the “Backpage” tab on the right side, you can find the annotation and cross references for the gene.
- b) In the cross references, can you find the Ensembl, Entrez Gene and HGNC identifiers for this gene?
- c) Why is it important to load the identifier mapping database?

Step 2: Import the gene expression data in PathVisio

- Go to Data → Import expression data
- Select the **lung-cancer-data.txt** file as the input file. Then click “Next”.
- Make sure the correct data delimiter is set and you see the following preview.



Question 2: Gene identifier

The first column (“GeneID”) contains the identifier of the genes. From which of the three databases below are the identifiers in the dataset? (Required for following steps!)

- Ensembl
- Entrez Gene
- OMIM

- In the next dialog, you need to define the column containing the identifier and the database used. Select the “GeneID” column for the identifier and “Use the same system code for all rows”. For that drop-down box, you should use the answer of question 2 as the “Database”. Then click “Next”. In this way, we tell the program that all identifiers in the input file are from that specific database which allows the program to do the correct identifier mapping.

Expression data import wizard

Choose column types

Select primary identifier column:

Select a column to specify system code

Use the same system code for all rows

Select the correct database here!

	A	B	C	D	E
1	GeneID	GeneName	log2FC	P.Value	adj.P.Value
2	ENSG00000...	TSPAN6	1.210920331	0.030285815	0.201038947
3	ENSG00000...	TNMD	1.601534394	0.171398244	0.444140353
4	ENSG00000...	DPM1	0.467464851	0.294138	0.564898506
5	ENSG00000...	SCYL3	0.054446204	0.891635248	1
5	ENSG00000...	C1orf112	1.437688699	0.004173094	0.100584579
7	ENSG00000...	FGR	-2.995942893	4.94E-06	0.005027991
3	ENSG00000...	CFH	-0.254214495	0.609656364	0.920902222

Back Next Cancel

- The data will now be imported. Before clicking “Finish”, check how many rows were imported successfully and how many identifiers were not recognized.

Question 3: Data import

- a) How many rows were imported successfully?
- b) How many identifiers were not recognized?

If the number of rows is the same as the number of identifiers not recognized the data import was not done correctly - you probably did not select the correct database! Redo the import or ask one of the instructors for help. (Required for following steps!)

Check before you continue!!



If you clicked finish, you should see a default visualization on the pathway (if all genes are gray, the data import was not successful → please redo the import, make sure you select the correct database in the data import, otherwise ask one of the instructors). Click on the *GSK3B* gene in the top left and check the “Data” tab on the right side → you should see that the *GSK3B* gene has a log₂FC of -0.38?



Step 3: Data visualization

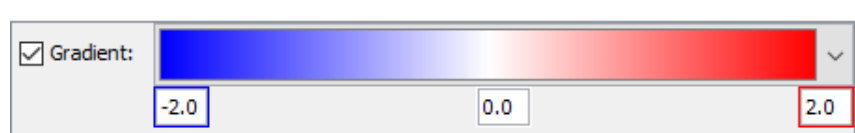
The default visualization is just a starting point to visualize your dataset; however, if you want to explore your dataset in more detail, other visualization options are available and more suitable for the data nodes (in this case gene nodes).

Question 4: Log₂FC values

- a) Can you describe in your own words, what the log₂FC means?
- b) In this example dataset, we compare primary lung cancer with normal tissue. What does a positive or negative log₂FC mean?

We will now create a visualization in PathVisio to visualize the log2FC as a gradient on the data nodes. Check out the tutorial here, if you run into any issues: <https://pathvisio.org/tutorials/visualization.html>.

- Go to Data → Visualization
- Create a new visualization named “log2FC visualization” – click on  in the top right
- Select “Text label”
- Select “Expression as color” and then the “Basic” option
- Select the checkbox before “log2FC” and define a new color set (click on  for color set)
- Select “Gradient” and define a gradient from -2 over 0 to 2 (blue – white – red) and click Ok.



Question 5: Biological interpretation

- a) As a whole, is the pathway more up- or down-regulated in lung cancer? Is this expected?
- b) Make a screenshot of the pathway. What do the colors mean?

Side note: The cell cycle pathway that we opened is just an example pathway that we can use to test the pathway visualization! For your own project, you can use any of the pathway models in the WikiPathways collection! Only use the cell cycle pathway if it is relevant for your research question!

Step 4: Pathway statistics

We now know how the cell cycle pathway is altered in lung cancer, but there could be other interesting pathways to look at. How can we find those pathways? That’s when we perform pathway statistics to find those pathways that are more altered than expected (check workshop from yesterday!).

- Go to “Data → Statistics”
- First, we want to find up-regulated pathways, so we need to define a criteria that selects all genes that are significantly changed:
$$([\log_2FC] > 1 \text{ OR } [\log_2FC < -1]) \text{ AND } [P.\text{Value}] < 0.05$$
- Then, we select the directory that contains all human pathways available in WikiPathways. We already downloaded the collection from WikiPathways and added them in the `wikipathways-20230810-gpm1-Homo_sapiens` folder. All these pathways will be tested and statistically evaluated if they are up-regulated in our dataset: click on browse to select the pathway folder.
- Then, we click on “calculate” and wait for the result table.

Question 6: Biological interpretation

- a) What are the top five up-regulated pathways and what are their Z-scores?
- b) Do you see highly ranked pathways in the result table that you expected to be upregulated in lung cancer?

Hint: you can save the result table in a txt file that can be opened in Excel.

Side note: In your own project, you can focus on any of the significantly changed pathways (Z-score > 1.96) – not only the top five. You can add the table in your results section and then focus on those pathways that relate to your research question.