Epigene Labs

Lucas Bodelle

Task 2 :

Update the workflow to give more details on the processing you think is followed

Step 1: Data download

Data is downloaded from public databases or from collaborators. The data will probably be stored in a cloud, depending on the size of the dataset.

Step 2: Check molecular data type

This step checks whether the dataset corresponds to what is expected. This step is important because if the data are processed using the wrong normalization method, the data would be meaningless.

Step 3: Normalization

Once the data type has been identified, normalize with the appropriate method: tpm , fpkm etc. The results are now ready for interpretation.

Step 4: Harmonize gene names

Convert all gene indexes to the appropriate id.

How could we enhance the pipeline to ensure the consistency of the data?

- Validation method between each step. For example, for normalization, integrate another control dataset with known normalization values and check whether the output for the control is valid.

- Place the “Harmonize gene names” step before normalization. If any gene names are invalid, we'll have to re-process our dataset, especially if there's a normalization operation with the columns.

- At the “check molecular data type” stage, if the data type matches but the values are aberrant, we should ask ourselves some questions. Probably perform statistical analyses of variances and outliers

What challenges do you foresee?

- Successfully retrieve a lot of data and process it while being certain that our data is of high quality. There may be notation errors, data gaps or outliers.
- Optimize the pipeline to be able to process data faster, while limiting calculation time.
- include new data processing methods to accept other types of data

Comment on the different output files

- We have different outputs for the GSE102301 dataset, because we can calculate fpkm and tpm from a “count” input.
- The dataset with fpkm input and output GSE108322. The input contains gene names in the first column, not in the index. The input also contains gene repeats (e.g. AASDH).
- The dataset with input and output tpm GSE113184. The input has repeats in the columns (example: GSM3099302 and GSM3099302.1).

How would you use these different files for downstream analysis?

You can do several things, such as:

- Boxplot visualization
- Clustering
- Dimension reduction
- Enrichment analysis on KEGG with TPM