Running the JGI Integration Workflow

Overview

The general stages of the workflow are:

- Loading and normalizing multi-omics datasets (e.g., transcriptomics and metabolomics)
- Integrating data (quantitative values) and metadata (sample info) between datasets
- Selecting features for analysis via statistical tests
- Building and visualizing a correlation network of integrated features
- Running multi-omics factor analysis

Note: This tutorial assumes you've already completed all stages of the workflow and environment setup detailed in *setup.pdf.*

Running the Notebook

1. You should see the JupyterLab interface rendered in your browser tab.
2. Double-click the workflow notebook `/integration_workflow.ipynb` in the left menu navigator to bring it into the workspace.
3. Double-click the configuration file in `/input_data/config/project_config.yml` to bring it into the workspace
4. Begin a workflow with all default parameters:

    A. Ensure that the kernel is "JGI Integration" by checking the top right corner of the workspace. If not, click the kernel name and select "JGI Integration" from the dropdown menu.
    B. Run all notebook cells in order, either with the "play" button in the top menu bar or with the keyboard shortcut in each cell (ctrl/cmd/shift-click).
    C. Each cell performs a workflow step (data loading, normalization, integration, correlation analysis, etc.) and prints some information to the standard output for review.
    D. Review the cell print statements to see where plots and tables are saved to the `/output_data` directory; some key results or previews are also displayed in the notebook.

    Note: At any time, instead of looking through the `/output_data` directory, you can access and view the attributes of a dataset or analysis by creating a new cell and executing the command `<object>.<attribute>`. For example, running a cell with `tx_dataset.normalized_data` will show the transcriptomics count table after all dataset normalization steps, or `mx_dataset.linked_metadata` will show the metabolomics metadata table after it has been linked to the other datasets. To see all the possible attributes that you can view for each object, run a cell with the command

`vars(<object>).keys()`. For example, `vars(analysis).keys()` or `vars(mx_dataset).keys()`.

5. Run a workflow with customized parameters:

    A. Update parameters in the `/input_data/config/project_config.yml` file (in the JupyterLab interface) using the guides provided in the `/jgi_integration/docs/*_parameters_explained.md` files.
    B. This will create a new parameter set to run the dataset processing and analysis stages on – make sure to change the `data_processing_tag` and/or `data_analysis_tag` in the config to create a new output directory to store the results and keep them distinct from previous runs.
    C. Rerun the notebook as described above from the beginning to re-load the updated configuration settings and run the full workflow.