Background JGI Integration Workflow

Background

Integrating metabolomics and transcriptomics data is a powerful approach in biological research because it provides a comprehensive understanding of cell or organismal processes by linking gene expression to metabolic activity. Transcriptomics reveals the abundance of RNA transcripts, reflecting gene expression levels, while metabolomics offers insights into the end products of metabolic pathways, including small molecules and metabolites. By combining these two layers of biological information, you can identify how (or if) expression activity relates to metabolic changes, uncover the biochemical pathways that underlie complex phenotypes or experimental perturbations, and find networks of quantitatively correlated RNAs and metabolites.

Introduction

This document provides step-by-step instructions for running the JGI Integration Workflow. It will walk you through running the workflow notebook for multi-omics data integration, customizing parameters, understanding input/output locations, and interpreting the main analysis results.

Data Input and Results Output

Input Data

- **Raw data:** JGI provides pre-processed transcriptomics (RNA counts) and metabolomics (metabolite peak height) data and metadata (generated from user-supplied information). These files are located in `/input_data/raw_data/` and are read into the workflow in the background and passed to the data processing steps.
- **Configuration file:** The workflow configuration is in `/input_data/config/project_config.yml`. This file controls all parameters for data processing and analysis stages.
- **Link script:** This python script is used to link the multi-omics datasets together sample-wise (e.g., by creating a one-to-one relationship between a transcriptomics sample and a metabolomics sample). It is a custom script produced by JGI analysts that uses experimental variables and user-supplied metadata to link the dataset samples and should not be altered.

Output Data

- **Processed data and analysis results:** All outputs are stored under the directory specified in the project config file (`user_settings -> project_name`). The output

files are located in `/output_data/`. To aid in multiple runs of the workflow using different parameters, there is a "tag" that can be used to customize the name of output folders for different data processing and analysis runs.

```
└── project_name/
    └── Dataset_Processing--TAG1/
        ├── Analysis—TAG0/
        │   ├── feature_network/
        │   ├── mofa/
        ├── metabolomics/
        │   ├── plots/
        └── transcriptomics/
```

Analysis and processing folders:

Results are organized by tags (see data_processing_tag and data_analysis_tag in your config).

Example output structure:

Results Interpretation

Correlation networks. A full or bipartite network is constructed from the integrated data. Before conducting network analysis, features without significant variation across groups in your experimental design (based on a non-parametric Kruskal-Wallis test) are dropped from the analysis because they are not meaningful to correlation analysis and large graphs are very difficult to interpret without sub-setting the data in a biologically or technically meaningful way. Only features which pass this test - up to a maximum of the most variable 5,000 combined transcripts and metabolites due to graph complexity - are plotted on the correlation network. We construct both a full (all edges and nodes) and a bipartite network (only edges which connect different 'omics data nodes), and for the bipartite network we also construct submodules (discrete sets of connected nodes). The networks are provided in .graphml format for uploading to CytoScape or a similar tool, and the node and edge files of each graph are provided for custom analysis. More details on interpreting the integrated network results can be found below.

MAGI2 (Metabolite Annotation and Gene Integration). MAGI2 analysis uses a biochemical reaction network to calculate a score for metabolite-gene associations. The score emphasizes consensus between metabolites and genes through biochemical reactions and is used to link features from metabolomics with features from transcriptomics via a shared biological pathway. Our workflow uses the genome or metagenome from which the transcriptome data was derived along with the untargeted metabolomics data as input to MAGI2, resulting in a results table that associates and scores RNAs and metabolites alongside links to external database sources. More details on interpreting MAGI2 results can be found below. The source code for MAGI2 can be

found here [https://github.com/biorack/magi] and the publication describing the MAGI algorithm can be found here [doi:10.1021/acschembio.8b01107].

MOFA (Multi-omics Factor Analysis). MOFA2 is a modelling tool that helps integrate and interpret ‘omics datasets in an unsupervised fashion. The software produces factors that are analogous to principal component analysis (PCA) axes but tailored for use on multi-omics datasets. It takes two or more data matrices with differing structure (e.g., RNA counts and metabolite peak heights) that have the same or at least overlapping samples and infers “interpretable low-dimensional representation in terms of a few latent factors”. More details on interpreting MOFA2 results can be found below. The source code for MOFA2 can be found here [https://github.com/bioFAM/MOFA2] and the publication describing the MOFA2 algorithm can be found here [doi:10.1101/2020.11.03.366674].