User's Guide

Overview

This guide provides step-by-step instructions for running the provided code to analyse sex-biased gene expression across multiple tissues and visualize biological pathways using statistical and machine learning methods.

Operating Instructions

1. Setting Up Your Environment

1. **Install Python**: Ensure Python 3.x is installed on your system. Download it from Python's website.
2. **Required Libraries**: Install the following libraries using pip:

pip install pandas scipy tqdm matplotlib seaborn scikit-learn

3. **Colab Integration**: The code is designed for use in Google Colab, which provides a cloud environment for Python programming.

2. Preparing Your Data

- **File Structure**: Ensure your GTEx data files are organized as follows:

  final_project_B/Data/

  ├── Adipose_Visceral

  │ ├── mat.f.coding.Adipose - Visceral (Omentum).csv

  │ └── pheno.f.Adipose - Visceral (Omentum).csv

  ├── Muscle_Skeletal

  │ ├── mat.f.coding.Muscle - Skeletal.csv

  │ └── pheno.f.Muscle - Skeletal.csv

  ├── Nerve_Tibial

  │ ├── mat.f.coding.Nerve - Tibial.csv

  │ └── pheno.f.Nerve - Tibial.csv

  └── Liver

  ├── mat.f.coding.Liver.csv

  └── pheno.f.Liver.csv

- **Metadata**: The gene.f.csv file should contain the mapping of gene names to their descriptions.

3. Running the Code

1. **Mount Google Drive**: Ensure your data is in Google Drive and run the following:

   ```
   from google.colab import drive
   drive.mount('/content/drive')
   ```

2. **Gene Expression Analysis**:
   - The script iterates through each tissue, performs t-tests to identify sex-biased genes, and saves the significant results in tissue-specific CSV files.
   - Outputs are stored in: /content/drive/My Drive/final_project_B/Data/diff_genes_<Tissue>.csv.
3. **KEGG Pathway Visualization**:
   - Pathway visualization is generated for each tissue based on KEGG data files.
   - Charts highlight pathways with significant enrichment.
4. **Machine Learning**:
   - A Random Forest Classifier predicts gender based on gene expression data.
   - Results include:
     - Feature importance charts for top genes.
     - Classification reports.
     - Confusion matrices.
     - Saved feature importance files for each tissue.

4. Visualizations

- Pathway and gene-level insights are visualized as bar charts, saved as PNG files or shown interactively in Colab.

5. Output

- **CSV Files**: Containing significant genes for each tissue.
- **Charts**: Visual representation of KEGG pathways and top genes contributing to gender classification.