Task 2 - Improve the RNAseq data integration pipeline

I. Comment on the pipeline

1. Objective of the Workflow

The goal of the pipeline is to process bulk RNA-seq datasets from various sources and formats as input, automatically integrate them, and produce standardized datasets as output. These standardized datasets will enable efficient parallel comparisons and analyses of data from diverse origins.

2. Specificity of Each Dataset

The preprocessing requirements for each dataset depend on their respective formats. Below are the details for each dataset:

GSE113184_rsem_tpm_gene_name.csv

This dataset uses TPM (Transcripts Per Million), which are normalized values that account for sequencing depth and gene length.

Preprocessing required:

- This dataset includes both gene names and Ensembl IDs. Convert Ensembl IDs to gene names to align with the format of the other datasets. Note that some Ensembl IDs do not have corresponding gene names. In such cases, we will need to decide whether to exclude these genes or retain their Ensembl IDs.

GSE108322_fpkm_gene_name.csv

This dataset uses FPKM (Fragments Per Kilobase of transcript per Million mapped reads), a metric similar to TPM but less commonly used now due to inherent biases.

Preprocessing required:

- FPKM values need to be converted to TPM for compatibility with other datasets.
- Gene names are already standardized, so no further adjustments are necessary.

GSE102301_count_gene_id.csv

This dataset uses raw counts, which are not normalized.

Preprocessing required:

- Normalize the counts (e.g., using tools like DESeq2 or edgeR) to account for sequencing depth and other technical factors.
- This dataset identifies genes with their Ensembl IDs. Convert Ensembl IDs to gene names to align with the format of the other datasets. Note that some Ensembl IDs do not

have corresponding gene names. In such cases, we will need to decide whether to exclude these genes or retain their Ensembl IDs.

General Preprocessing Requirements

To ensure integration across datasets, gene names must be consistent. Conversion to a common format—gene names—is necessary. Special consideration must be given to genes without corresponding names in the Ensembl database, as this will influence the final dataset's structure. To compare expressions between different datasets, it is fundamental to have the same units of measurements for the expression level. The pipeline could also achieve this task.

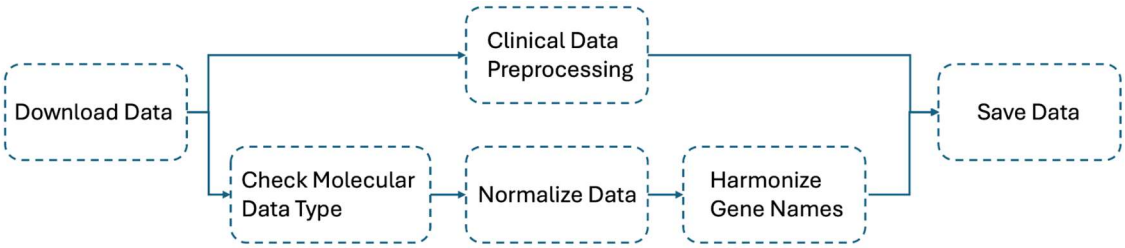
3. Functionalities of the pipeline

By analyzing the inputs and outputs of the pipeline, it appears that the pipeline currently:

- Harmonizes gene names by converting Ensembl IDs into the gene name format.
- Creates both FPKM and TPM files for datasets with raw counts.

However, it does not:

- Convert all files into a unified format (e.g., TPM) to ensure readiness for comparative analyses.



General Diagram of the pipeline

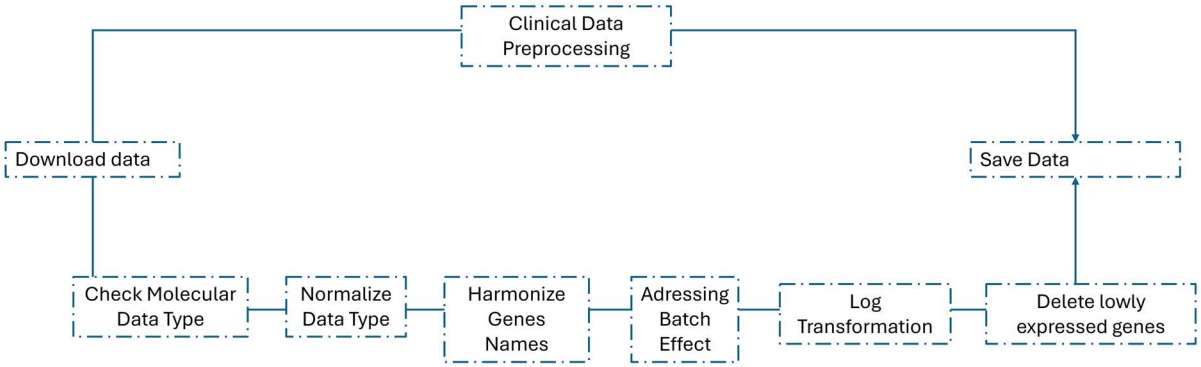
4. How could we enhance the pipeline to ensure the consistency of the data?

To enhance the pipeline and ensure data consistency, several steps can be implemented. First, all datasets should be converted to the same normalization format, such as TPM (Transcripts Per Million), to ensure comparability across datasets. This involves converting raw counts or FPKM values into TPM using standard formulas.

Next, addressing batch effects is crucial, as variations arising from differences in experimental conditions can obscure true biological signals; tools like ComBat from the sva package can harmonize data across batches. Applying a log transformation (e.g., log2(TPM + 1)) helps stabilize variance and make the data more comparable, particularly for genes with large differences in expression levels.

Finally, removing lowly expressed genes (e.g., genes with low counts, TPM, or FPKM across most samples) reduces noise and computational burden, focusing analyses on robustly detected

transcripts. Together, these steps ensure the data is normalized, corrected, and filtered for consistency, enabling reliable downstream analyses.



General Diagram of the improved pipeline

5. What challenges do you foresee?

The main challenges are as follows:

- Special attention must be given to genes without corresponding names in the Ensembl database, as these discrepancies can significantly impact the structure and usability of the final dataset.
- Batch effect correction can vary in efficiency; selecting the appropriate method and performing thorough benchmarking are crucial to achieving robust results, especially when integrating multiple diverse datasets.
- Removing lowly expressed genes requires careful testing and benchmarking to avoid discarding important information, as some low-expression genes may still hold critical biological significance.
- Taking care of the missing values in the different datasets (replacing by 0? Deleting these genes?)

II. Comment on the different output files

1. Why do we have different files for one dataset?

It appears that the GSE102301_count_gene_id.csv file was preprocessed in 3 files when it was given to the pipeline. The pipeline has done the following things :

- Creates both FPKM and TPM files and keep a Count files.
- Harmonizes gene names by converting Ensembl IDs into the gene name format in each new created dataset.

2. How would you use these different files for downstream analysis?

These RNA-seq datasets, once properly processed, could be utilized for a wide range of analyses to extract valuable biological insights. Here are some examples:

1. **Differential Expression Analysis**:
   This approach identifies genes whose expression levels vary significantly between conditions (e.g., healthy vs. diseased samples). Such analyses are essential for uncovering genes potentially implicated in disease mechanisms, therapeutic targets, or biomarkers for diagnosis and prognosis.

2. **Machine Learning Applications**:
   RNA-seq data can be used for training machine learning models to predict specific health conditions, classify diseases, or even stratify patients based on gene expression profiles. Techniques such as random forests, support vector machines (SVMs), or neural networks can leverage these data to enhance diagnostic accuracy or uncover hidden patterns.

3. **Clustering**:
   By clustering samples based on their gene expression profiles, researchers can group similar samples or identify novel subtypes within a dataset. For example, unsupervised clustering can reveal molecular subtypes of a disease, enabling personalized medicine approaches.

4. **Statistical Analysis**:
   Advanced statistical methods can be applied to study correlations between gene expression and clinical variables, such as age, disease stage, or treatment response. This can help uncover relationships between molecular data and phenotypic outcomes.

5. **Pathway and Functional Enrichment Analysis**:
   Processed RNA-seq data can be used to identify biological pathways or functional gene sets that are enriched in a given condition. This helps to contextualize differential expression results and understand the broader biological implications.

6. **Integration with Multi-omics Data**:
   RNA-seq data can be combined with other omics data types (e.g., proteomics, epigenomics) to achieve a more comprehensive understanding of biological systems and disease mechanisms.