Bulk RNA-seq Differential Expression and Cell-Type Marker Identification

Objective:

Analyze bulk RNA-seq data to identify differentially expressed genes (DEGs) between conditions (e.g., healthy vs. diseased tissue). Then, use this information to infer cell types that might be present in the samples based on known marker genes.

Step-by-Step Workflow:

1. **Data Acquisition**
    - **Goal:** Obtain bulk RNA-seq data from a public repository like GEO.
    - **Process:** Choose a dataset with clear conditions (e.g., tumor vs. normal tissue) that is already preprocessed or provides read counts in a format like a count matrix.
    - **Simplification:** Download a dataset that comes with a count matrix, skipping raw data processing.
2. **Quality Control (QC)**
    - **Goal:** Ensure the quality of the data by checking library size and gene count distribution.
    - **Process:**
        - Use simple QC metrics like plotting the distribution of counts across samples.
    - **Tools:**
        - In R: `edgeR::cpm()`, `boxplot()`
        - In Python: `pandas`, `matplotlib`
    - **Simplification:** Perform basic QC with visualization (boxplots) instead of extensive filtering.
3. **Normalization**
    - **Goal:** Normalize the gene expression data to make it comparable across samples.
    - **Process:**
        - Use straightforward normalization methods like TPM (Transcripts Per Million) or library size scaling.
    - **Tools:**
        - In R: `DESeq2::rlog()` or `edgeR::calcNormFactors()`
        - In Python: `scikit-learn` for scaling, or use pre-normalized data.
    - **Simplification:** Choose a simple normalization method without complex adjustments.
4. **Differential Expression Analysis**
    - **Goal:** Identify genes that are differentially expressed between the conditions.
    - **Process:**
        - Perform differential expression analysis to find DEGs.
    - **Tools:**
        - In R: `DESeq2::DESeq()`, `edgeR::exactTest()`
        - In Python: `Scanpy` can be adapted for bulk data, or use `DESeq2` through Rpy2 in Python.

    - **Simplification:** Focus on comparing two conditions (e.g., healthy vs. diseased) to keep it straightforward.
5. **Identify Cell-Type Marker Genes**
    - **Goal:** Use the list of DEGs to identify potential cell-type markers.
    - **Process:**
        - Compare DEGs with known cell-type marker databases (e.g., CellMarker, PanglaoDB) to infer which cell types might be enriched in the samples.
    - **Tools:**
        - Online databases: CellMarker, PanglaoDB
        - In R: `clusterProfiler` for enrichment analysis
    - **Simplification:** Use online resources to match DEGs to known cell-type markers instead of manual annotation.
6. **Visualization**
    - **Goal:** Visualize the differential expression results and marker genes.
    - **Process:**
        - Generate simple plots like volcano plots, heatmaps, or bar plots of top DEGs and their associated cell types.
    - **Tools:**
        - In R: `ggplot2`, `pheatmap`
        - In Python: `matplotlib`, `seaborn`
    - **Simplification:** Focus on 1-2 visualization types for clarity.

Expected Outcomes

- Identify differentially expressed genes between the conditions.
- Infer which cell types might be contributing to the gene expression changes based on known markers.