**Intelligent Bulk RNA-Seq Differential Gene Analysis Assistant**

**Objective**: Provide a one-stop, intelligent RNA-Seq differential expression analysis tool tailored for non-programming users.

**Project Lifecycle**

**1. Initiation**

* **Objective**: Create an intelligent RNA-Seq analysis tool for non-programmers.
* **Stakeholders**: Researchers, bioinformaticians, developers.
* **Technology Stack**: Python (backend), R (DESeq2/edgeR), Streamlit/Flask (frontend), Plotly/Seaborn (visualization).

**2. Planning**

* **Milestones**:
  1. Data upload and QC.
  2. Normalization and batch correction.
  3. PCA and differential analysis.
  4. Interactive visualizations (heatmaps, volcano plots).
* **Sprint Duration**: 1 weeks per milestone.

**Functional Modules**

**1. Data Preprocessing**

* **Upload Support**:
  + Accepts multiple input formats: CSV, TSV, Excel, or drag-and-drop file upload.
  + Validates file integrity, including gene name format, sample count, and missing values.
* **Data Quality Control (QC)**:
  + Automatically detects and filters low-expression genes.
  + Generates sample coverage distribution plots and count distribution boxplots.
* **Batch Effect Detection and Adjustment**:
  + Automatically identifies batch information (provided by users or inferred from sample grouping).
  + Provides PCA and boxplot analyses for batch effect visualization.
  + Supports batch effect correction using limma::removeBatchEffect or ComBat.

**2. Data Normalization**

* **Normalization Options**:
  + DESeq2-based normalization (size factors).
  + EdgeR-based TMM (Trimmed Mean of M-values).
* **Visualization**:
  + Generates pre- and post-normalization sample distribution plots, such as density or violin plots.

**3. PCA Analysis and Visualization**

* **Features**:
  + Draws 2D and 3D PCA plots.
  + Highlights samples by experimental group or batch.
  + Automatically flags outliers and notifies users.
* **User Interaction**:
  + Hover over points to display sample IDs and additional details.
  + Provides options to download high-resolution plots for reports.

**4. Differential Expression Analysis**

* **Methods**:
  + ANOVA or Rank-Sum test.
  + Established packages like DESeq2 and Edge (user-selectable).
  + Automatically sets thresholds for log2 fold change and adjusted p-value, adjustable by users.
* **Results Presentation**:
  + Interactive volcano plots with the ability to:
    - Select gene points to view detailed information.
    - Highlight significant genes.
  + Table of significant genes, including gene ID, log2 fold change, and adjusted p-value.

**5. Heatmap Visualization**

* **Heatmap Features**:
  + Draws heatmaps for differentially expressed genes (Top 50 or user-defined).
  + Customizable clustering methods for both samples and genes.
  + Allows display of gene names or IDs and supports exporting high-resolution images.
* **Batch Information Integration**:
  + Highlights heatmap column headers by experimental group or batch.