**Breast Cancer Biomarker Discovery: Final Report**

**Authors (@slack):** Aya Elhagry (@AyaElhagry), Alaa Khaled (@Alaa043), Marwan Magdy (@Marwan16), Ahmed Abdel-Maqsoud (@Abdel-Maqsoud), Olabode Kaosara (@TheResearchNerd), Obianujunwa Martha (@obianujunwa).

**1. Introduction to Breast Cancer**

Breast cancer is among the most prevalent cancers globally and ranks as the second leading cause of cancer-related fatalities among women. It originates in the breast tissue and can metastasize to other regions of the body if left untreated. A deeper understanding of the molecular mechanisms driving breast cancer development and progression is essential for enhancing diagnosis, treatment, and patient outcomes. This project focuses on identifying potential biomarkers for breast cancer by analyzing gene expression data sourced from The Cancer Genome Atlas (TCGA). Specifically, we compare primary and recurrent tumor samples to uncover genes that may play pivotal roles in cancer recurrence or progression.

**2. Description of the Dataset and Preprocessing Steps**

**2.1 Dataset**

The dataset employed in this analysis originates from the TCGA-BRCA (Breast Cancer) project, encompassing RNA-Seq data that provides a detailed overview of gene expression profiles in breast cancer samples. Key characteristics of the dataset include:

* **Data Category**: Transcriptome Profiling
* **Data Type**: Gene Expression Quantification
* **Experimental Strategy**: RNA-Seq
* **Workflow Type**: STAR - Counts

**2.2 Preprocessing Steps**

1. **Data Download and Preparation**:
   * Utilized the TCGAbiolinks R package to query and download RNA-Seq data for breast cancer.
   * Retrieved raw gene expression counts.
2. **Missing Data Handling**:
   * Identified genes with a high proportion of missing values (threshold: 20%).
   * Removed genes exceeding this threshold and replaced remaining NA values with 0.
3. **Normalization**:
   * Applied Variance Stabilizing Transformation (VST) using the DESeq2 package.
   * This transformation helps reduce heteroscedasticity, making the data more suitable for subsequent analyses.
4. **Data Reduction**:
   * Selected 20 primary tumor samples and 20 recurrent tumor samples, yielding a total of 40 samples for analysis.

**3. Methodology for Biomarker Discovery**

**3.1 Differential Expression Analysis**

1. **Sample Selection**:
   * Chose 20 primary tumor samples and 20 recurrent tumor samples for comparative analysis.
2. **DESeq2 Analysis**:
   * Created a DESeqDataSet object with the count data and corresponding sample information.
   * Defined the design formula to contrast primary versus recurrent samples.
   * Conducted differential expression analysis using the DESeq2 methodology.
3. **Significance Criteria**:
   * An adjusted p-value < 0.05 was set as the threshold for identifying statistically significant differentially expressed genes.

**4. Results and Interpretations**

**4.1 Differential Expression Analysis Results**

The differential expression analysis revealed a collection of genes with significant expression changes between primary and recurrent tumor samples. These genes hold potential as biomarkers for breast cancer recurrence.

**4.2 Visualizations**

1. **Volcano Plot**:
   * This plot illustrates the relationship between fold change and statistical significance. Significantly differentially expressed genes are marked in red (adjusted p-value < 0.05), while non-significant genes are represented in gray.
2. **MA Plot**:
   * The MA plot displays the mean of normalized counts against the log2 fold change, aiding in the identification of genes with substantial fold changes that are statistically significant.

**5. Conclusion and Future Directions**

**5.1 Conclusion**

We successfully identified differentially expressed genes between primary and recurrent breast cancer tumors using RNA-Seq data from TCGA. These genes could serve as promising biomarkers for breast cancer recurrence and provide insights into the molecular mechanisms underlying cancer progression.

**5.2 Future Directions**

1. **Functional Enrichment Analysis**: Conduct pathway and Gene Ontology enrichment analysis to elucidate the biological processes and functions associated with the differentially expressed genes.
2. **Machine Learning Analysis**: Implement feature selection techniques and machine learning algorithms (e.g., kNN or Random Forest) to create predictive models for breast cancer recurrence based on gene expression profiles.
3. **Integration with Clinical Data**: Incorporate clinical data to explore the relationship between gene expression patterns and patient outcomes, treatment responses, or other clinical variables.
4. **Single-cell RNA-Seq Analysis**: Expand the study to include single-cell RNA-Seq data to comprehend cellular heterogeneity within tumors and identify cell-type-specific biomarkers.
5. **Multi-omics Integration**: Integrate gene expression data with other omics data types (e.g., DNA methylation, copy number variations) for a more holistic understanding of breast cancer biology.

By pursuing these directions, we can further refine our understanding of breast cancer biomarkers and potentially develop more accurate diagnostic and prognostic tools for clinical application.