

CAP 5510 - Bioinformatics

Project Proposal

Project Title

Differential Gene Expression Analysis Using RNA-seq Data

Team Members

- Shaurya Singh – 69871462
- Aryan Atre – 72162633

Abstract

We will perform an end-to-end differential gene expression (DGE) analysis comparing healthy and diseased tissue using publicly available bulk RNA-seq data. Starting from raw FASTQ files (or aligned counts when appropriate), we will build a reproducible pipeline to perform quality control, trimming, quantification (two modes: aligner-based and quasi-mapping), count aggregation, statistical DGE testing, and pathway enrichment. We will compare two quantification strategies (**STAR + featureCounts** vs **Salmon quasi-mapping + tximport**) and two differential-testing frameworks (**DESeq2** vs **edgeR/limma-voom**) to evaluate effects on detected genes, runtimes, resource usage, and biological concordance. Results will be validated by cross-referencing known disease biomarkers and by comparing results between two selected datasets. Final deliverables will include:

- A reproducible pipeline (Nextflow/nf-core or Snakemake)
- An RMarkdown notebook with DE analysis
- Figures (PCA, MA/volcano, heatmaps)
- A short report discussing biological interpretation, reproducibility, and method comparison

Plan of Action

1. What We Will Implement

- A reproducible pipeline (Nextflow + nf-core/rnaseq or a Snakemake wrapper) that runs:
 1. QC: **FastQC + MultiQC**
 2. Trimming: **TrimGalore (Cutadapt)**
 3. Quantification:
 - (A) **STAR alignment + featureCounts**
 - (B) **Salmon quasi-mapping + tximport**
 4. Generate gene count matrix and sample metadata

5. DGE analysis in R using **DESeq2** (primary) and cross-check with **edgeR/limma-voom**
 6. Downstream interpretation: PCA, heatmaps, volcano plots, and pathway/Gene Ontology enrichment (**clusterProfiler** / **Enrichr**)
- A reproducible analysis notebook (**RMarkdown & Jupyter/IPYNB**) with code and figures
 - A short report comparing methods (accuracy proxies), runtime/compute costs, and biological findings

2. Methods to Compare

Quantification / Counting Strategies

- **STAR + featureCounts** (“alignment-based”): build genome index, align reads with STAR, obtain gene counts with featureCounts. (High accuracy for spliced alignment)
- **Salmon (quasi-mapping) + tximport**: run Salmon on trimmed FASTQs, use tximport to create gene counts. (Faster, widely used for bulk RNA-seq)

Statistical Testing Frameworks

- **DESeq2 (primary)**: negative-binomial GLM framework for DGE analysis in Bioconductor
- **edgeR / limma-voom (secondary)**: confirm robustness of top results and compare gene lists/effect sizes

Comparison Metrics

- Overlap of significant genes (Venn), concordance of log2FC, top gene lists
- Biological concordance (presence of known biomarkers)
- Numeric metrics: #DE genes (FDR < 0.05), median |log2FC|, p-value distributions
- Computational metrics: runtime, memory, disk space

3. Datasets

We will analyze **two complementary public datasets** to ensure results are not dataset-specific:

- **Option A: TCGA-BRCA (tumor vs adjacent normal)**
 - Rationale: Large sample size, standardized metadata
 - Source: GDC / TCGA Data Portal
- **Option B: GEO RNA-seq dataset (e.g., GSE52194 – breast cancer)**
 - Rationale: Focused cohort with healthy vs diseased grouping
 - Source: GEO / SRA (FASTQ files available)

Additional resources (if needed):

- **GTEEx Portal** for healthy tissue reference expression
- **Download tools:** SRA Toolkit (fasterq-dump), GDC-client

4. Experiments & Measurements

Experiments

1. Run pipeline across both quantification strategies
2. Perform DE analysis with DESeq2 (primary) and edgeR/limma-voom (secondary)
3. Conduct pathway enrichment on DE genes
4. Cross-dataset comparison (TCGA vs GEO)

Metrics

- **Biological/Statistical:**
 - Number of significant genes ($\text{padj} < 0.05$)
 - Concordance (Jaccard index) of top genes
 - Effect size agreement (correlation of $\log_2\text{FC}$)
 - Overlap in enriched pathways
- **Computational:**
 - Wall-clock runtime
 - Peak memory usage
 - Disk usage
- **Quality/Mapping:**
 - Mapping rate (%) from STAR and Salmon
 - Read duplication, insert size distributions, per-base quality (FastQC/MultiQC)

5. Planned Workload Distribution

- **Shaurya Singh**
 - Pipeline setup (Nextflow/Snakemake), sample configuration
 - Run STAR + featureCounts flow, capture alignment metrics
 - Generate MultiQC reports
- **Aryan Atre**
 - Run Salmon quantification and tximport gene counts
 - Primary DE analysis in R (DESeq2)
 - Secondary checks with edgeR/limma-voom; enrichment & visualization
- **Joint Tasks**
 - Interpret results, cross-validate gene lists
 - Write final report & presentation
 - Prepare reproducible notebooks and upload configs to GitHub

6. List of papers to read (initial)

- Love, M., Huber, W., & Anders, S. (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. Genome Biology.
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). *Salmon: fast and bias-aware quantification of transcript expression*. Nature Methods.
- Dobin, A., Davis, C. A., Schlesinger, F., et al. (2013). *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics.
- nf-core/rnaseq documentation (pipeline usage)
- Bioconductor RNA-seq workflows (rnaseqGene, DESeq2 vignettes)