Introduction to Statistical modelling of Count Data in R

Workshop Lead: Ryan Huang

Facilitator: Alejandro Mejia Garcia

Approximate duration: 2 hours

Prerequisites:

1. RStudio installed
2. Ability to write simple codes in R
3. Understanding of how RNA-seq count matrix was generated

Summary: (2-3 sentences summarizing the workshop)

This session will introduce participants to the statistical foundations required for analyzing RNA-seq count data. We will focus on generalized linear models (GLMs) and the principles of statistical testing that underpin differential gene expression analysis. Participants will learn how to pre-process and filter a count matrix, fit appropriate models, and interpret key statistical outputs that set the stage for downstream analyses.

Learning Objectives: (List 2-5 learning objectives participants will learn upon completion of this workshop)

1. Perform filtering and summary statistics to prepare data for differential expression analysis.
2. Explain why generalized linear models (GLMs) are appropriate for RNA-seq count.
3. Apply basic statistical tests and GLMs to RNA-seq count matrices in R
4. Diagnose and interpret statistical model outputs, including dispersion and goodness-of-fit.
5. With the statistical foundations introduced, students should be ready for the full differential gene expression analysis in Day 4

Content:

1. **Module 1: Statistical theories (60 mins)**
   a. Presentation (45 mins)
      - Characteristics of RNA-seq count data:
        (1) discreteness, overdispersion, non-normality
        (2) why Poisson regression is insufficient and how negative binomial models resolve these issues
        (3) role of variance-mean relationships in modeling RNA-seq data
      - Generalized Linear Models (GLMs) for RNA-seq

(1) Log-link functions and interpretation of coefficients

(2) Modeling dispersion and normalization factors (library size, effective length, compositional bias)

(3) Brief overview of state-of-the-art frameworks (i.e., edgeR) and how they extend GLMs

b. Hands-on activity (15 mins)

- Load pre-processed RNA-seq count matrix
- Explore distribution of counts and mean-variance trends
- Fit a Poisson and Negative Binomial GLM in R (using edgeR functions)
- Compare model fits and interpret dispersion estimates

2. Module 2: Key statistical applications for RNAseq in edgeR (60 mins)

a. Presentation (30 mins)

- Pre-processing steps before DGE:

  (1) Filtering low-count genes (e.g., edgeR's filterByExpr)

  (2) Normalization strategies (size factors, TMM, or median ratio)

  (3) Importance of reducing false positives and improving power

- Statistical testing framework for DGE using edgeR as an example:

  (1) Hypothesis testing in GLMs (Wald test vs. Likelihood Ratio Test)

  (2) Interpreting coefficients, log-fold changes, and p-values

  (3) Multiple testing correction (FDR, Benjamini–Hochberg)

b. Hands-on activity (30 mins)

- Apply gene filtering on the count matrix
- Normalize counts and inspect sample clustering (PCA/heatmap)
- Fit a GLM for a simple two-condition comparison
- Extract log-fold changes, p-values, and adjusted p-values
- Generate and understand diagnostic plots (MA plot, dispersion plot) in statistics