Course 4 · Week 4 — Omics, fairness, reproducibility

Cheatsheet — biostats_courses

R. Heller

Bulk RNA-seq

```
library(DESeq2)
dds <- DESeqDataSetFromMatrix(counts, coldata, ~ condition)
dds <- DESeq(dds)
res <- results(dds, contrast = c("condition", "B", "A"))
```

```
library(edgeR)
dge <- DGEList(counts, group = coldata$condition)
dge <- calcNormFactors(dge)
design <- model.matrix(~ condition, coldata)
fit <- glmQLFit(estimateDisp(dge, design), design)
qlf <- glmQLFTest(fit, coef = 2)
topTags(qlf)
```

Package	Distribution	Notes
DESeq2	negative binomial	shrinks LFCs; standard for small cohorts
edgeR	negative binomial	QL-F test has good type-I control
limma-voom	mean-variance weights	fast on very large studies

Enrichment

```
library(fgsea)
fgsea(pathways, stats = stat_vector, nperm = 10000) |>
  arrange(padj)

# Over-representation with a hypergeometric test
phyper(k - 1, K, N - K, n, lower.tail = FALSE)
```

Always pre-rank by a *signed* statistic (log-FC × -log10 p), not by raw p alone.

Single-cell RNA-seq (Seurat pattern)

```
library(Seurat)
so <- CreateSeuratObject(counts) |>
  NormalizeData() |>
  FindVariableFeatures() |>
  ScaleData() |>
  RunPCA() |>
  FindNeighbors() |>
  FindClusters(resolution = 0.4) |>
  RunUMAP(dims = 1:20)
DimPlot(so)
```

FDR, knockoffs, replication

Correction	When
Bonferroni	few planned tests, strict control
Benjamini-Hochberg	many tests, accept proportion of false discoveries
Knockoffs	controlled variable selection with FDR
Permutation / SEA	small n, non-standard models

```
p.adjust(pvals, method = "BH")
```

Replication crisis shorthand: pre-register + report effect sizes + share data. Power, not p-values.

TRIPOD-AI + fairness

- TRIPOD-AI extends TRIPOD for prediction models built with ML.
- Report: data provenance, training/evaluation splits, subgroup performance, calibration, external validation.
- Fairness: stratify AUC / calibration by sex, ancestry, age bands. Disparities in calibration are often larger than in AUC.

Reproducibility at scale with `targets`

```
# _targets.R
library(targets)
tar_option_set(packages = c("tidyverse"))
list(
  tar_target(raw, read_csv("data/raw.csv")),
  tar_target(clean, clean_data(raw)),
  tar_target(fit, fit_model(clean)),
  tar_target(report, quarto::quarto_render("report.qmd"))
)
```

```
Rscript -e 'targets::tar_make()'
```

Decision rule for Week 4

- Bulk DE → DESeq2 for small samples, edgeR/limma for large.
- Many hypotheses → BH by default; knockoffs for variable selection.
- scRNA-seq → Seurat pipeline with batch correction + cluster annotation.
- ML prediction model → TRIPOD-AI + subgroup analysis.
- Any multi-step analysis → `targets` pipeline.

Common pitfalls

- Running DE on a pathway-level aggregate rather than gene-level.
- Cherry-picking a pathway database post-hoc.
- Over-clustering scRNA-seq data until “known” groups appear.
- Calling an ML model “fair” after checking only overall AUC.
- Submitting a paper without a `sessionInfo()` or a lock file.

Further reading

- Love, Huber, Anders (2014), *Moderated estimation of fold change…*
- Collins et al. (2024), *TRIPOD-AI.*
- Peng, *Reproducible Research with R.*