Normalization Across Omics: Concepts, Numbers, and Review-Proof Interpretation

Introduction

High throughput omics technologies measure molecular abundance indirectly. RNA sequencing produces read counts, while mass spectrometry produces intensity values. In both cases, raw measurements are shaped by technical effects as much as by biology. Normalization corrects these effects so that comparisons are meaningful.

Insight

Normalization defines what biological questions your data can answer.

Why Normalization Is Necessary

Raw measurements depend on:

- Sequencing depth or instrument sensitivity
- Feature length or detectability
- Dominance of a small number of features

Deep Dive

Every normalization removes one bias by assuming something else is stable.

RNA Sequencing Normalization

FPKM and RPKM

$$\text{FPKM} = \frac{\text{Fragments mapped to a gene}}{\text{Gene length in kb} \times \text{Total fragments in millions}}$$

Common Pitfall

FPKM values are not comparable across samples because the denominator depends on global expression.

TPM

TPM normalizes gene length first and rescales values so that each sample sums to one million.

Insight

TPM represents relative abundance within a sample.

TMM Normalization

TMM estimates scaling factors using trimmed log fold changes.

Deep Dive

TMM assumes most genes are not differentially expressed and that changes are balanced.

DESeq2 Size Factor Normalization

DESeq2 uses the median ratio method based on geometric means.

Insight

The median protects against highly expressed genes.

Quantile Normalization

Common Pitfall

Quantile normalization enforces identical distributions and removes real biological differences.

Proteomics Normalization

iBAQ

$$\text{iBAQ}_i = \frac{I_i}{\text{Number of theoretical peptides}_i}$$

Deep Dive

iBAQ attempts to approximate absolute protein abundance by correcting for protein length.

RIBAQ

$$\text{RIBAQ}_i = \frac{I_i}{\sum_{j=1}^{N} I_j}$$

Insight

RIBAQ is conceptually analogous to TPM.

LFQ

Deep Dive

LFQ is designed for differential testing rather than absolute quantification.

Worked Numeric Example: RNA Sequencing

Gene	Sample A	Sample B	Length (kb)
Gene1	1000	2000	1
Gene2	500	500	2
Gene3	100	100	1

Common Pitfall

A dominant gene distorts normalization based on total counts.

Worked Numeric Example: Proteomics

Protein	Sample A	Sample B	Peptides
Prot1	1,000,000	2,000,000	50
Prot2	200,000	200,000	20
Prot3	50,000	50,000	10

Insight

Relative normalization hides absolute intensity differences.

Conceptual Alignment Across Omics

Concept	RNA seq	Proteomics	Purpose
Relative abundance	TPM	RIBAQ	Visualization
Length correction	FPKM	iBAQ	Bias reduction
Differential testing	TMM, DESeq2	LFQ	Inference

Reviewer Critique Boxes

Reviewer Critique

TPM or RIBAQ values were used for statistical testing. This is not appropriate.

Reviewer Critique

The normalization assumption that most features are unchanged is not justified.

Reviewer Critique

Multi omics layers are normalized for different questions. Integration logic is unclear.

Multi Omics Integration Guidance

- Use TPM and RIBAQ only for visualization and composition trends
- Use log fold changes from DESeq2 and LFQ for integration
- Integrate ranks or pathway scores, not raw normalized values

Deep Dive

Never integrate layers normalized for different questions.

One Page Visual Cheat Sheet

Question	RNA seq	Proteomics
Relative abundance	TPM	RIBAQ
Which features change	TMM, DESeq2	LFQ
Absolute estimate	Not reliable	iBAQ
Visualization	TPM	RIBAQ
Statistics	Raw counts	LFQ

Insight

If reviewers ask about normalization, they are questioning biological validity.

Final Takeaway

Normalization is the hidden axis of interpretation. Relative normalization describes composition. Model based normalization enables inference. Multi omics integration succeeds only when layers are aligned to the same question.