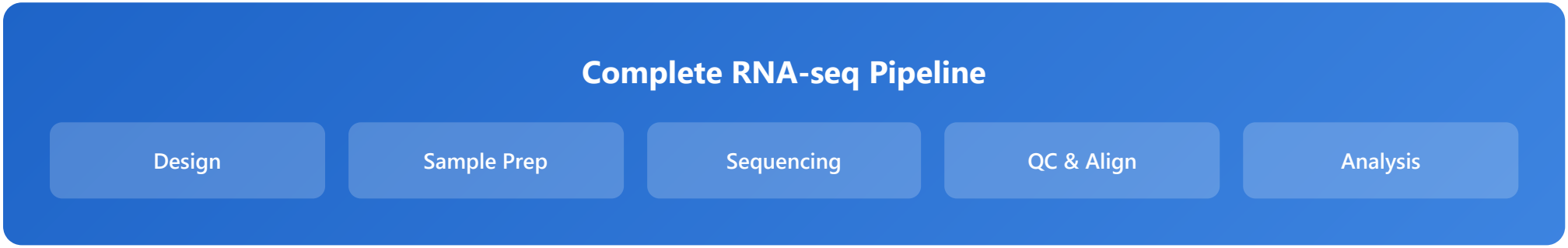
RNA-seq Workflow

Comprehensive Guide to Experimental Design and Best Practices

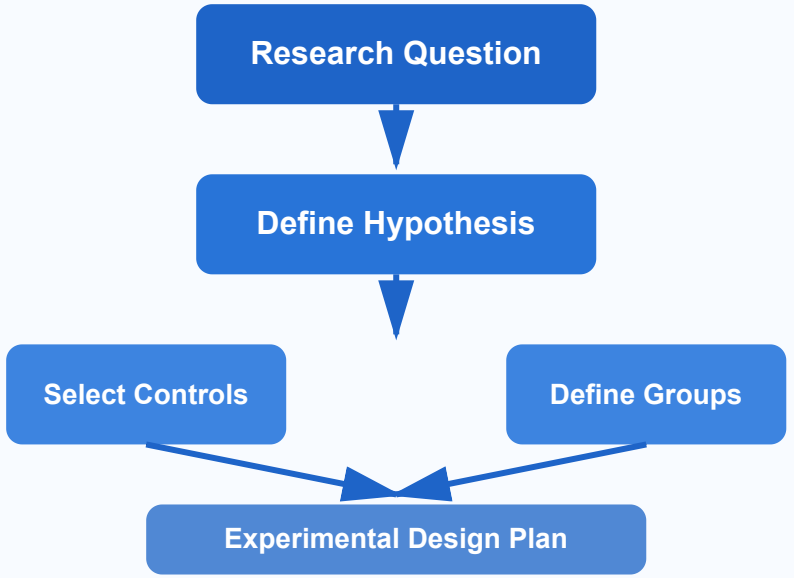


1 Experimental Design

Experimental design is the foundation of any successful RNA-seq study. A well-designed experiment ensures that the biological questions can be answered with statistical confidence while avoiding common pitfalls that can compromise data quality.

Critical considerations:

- Define clear biological hypotheses before starting
- Identify appropriate control groups
- Account for potential confounding variables
- Plan for both technical and biological variation
- Consider the statistical model for downstream analysis



Key Design Principles

- **Randomization:** Randomly assign samples to processing batches to prevent systematic bias
- **Blocking:** Group samples by known sources of variation (e.g., batch, sex, age)
- **Control Selection:** Use appropriate negative and positive controls
- **Sample Size:** Calculate required sample size based on expected effect size and variance

Common Pitfall: Starting sequencing without a clear analysis plan often leads to underpowered studies or inability to test the intended hypothesis.

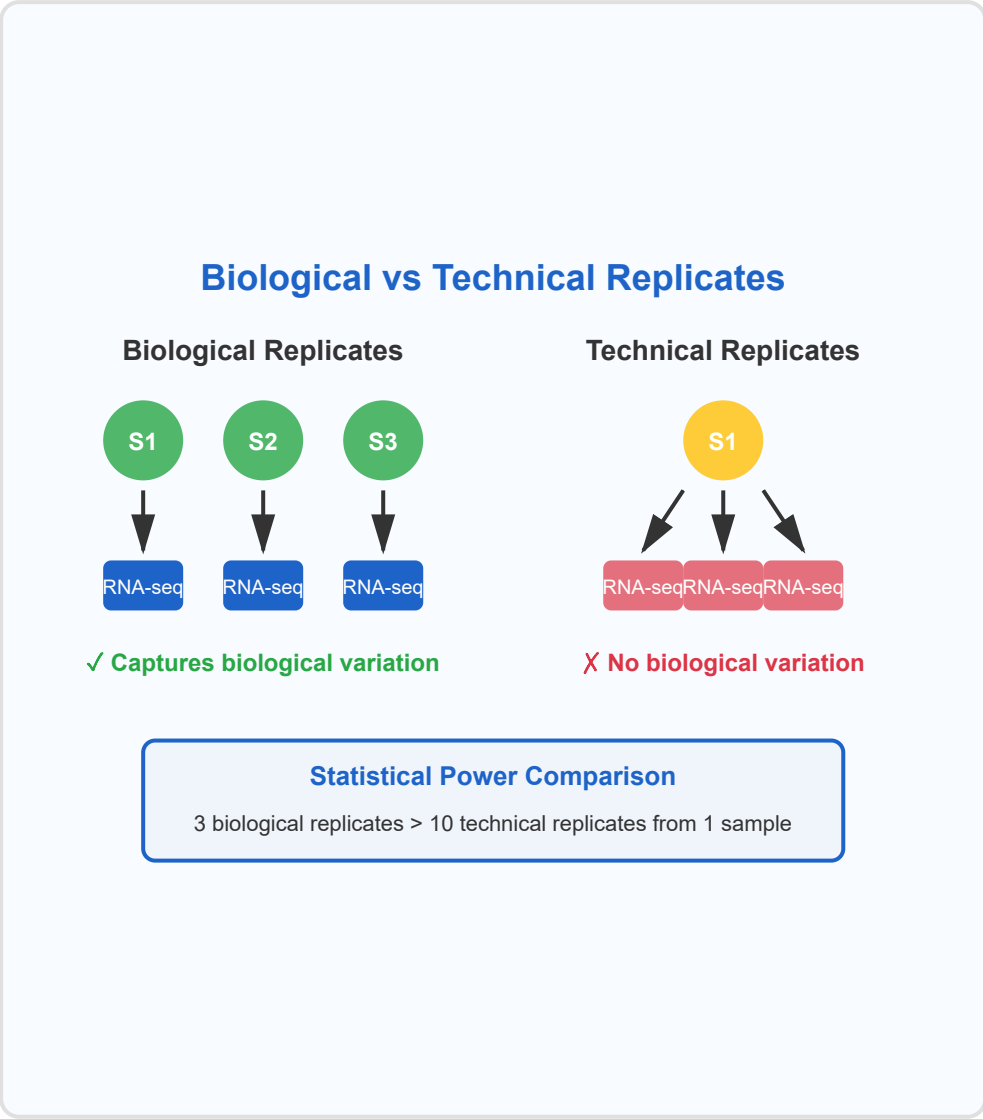
2 Replication Strategies

Replication is crucial for distinguishing true biological variation from technical noise. There are two types of replicates in RNA-seq: **technical replicates** (same sample sequenced multiple times) and **biological replicates** (independent samples from the same condition).

Biological replicates are essential because they capture the natural variation between individuals, which is typically much larger than technical variation. Modern RNA-seq platforms have low technical variation, making technical replicates less necessary.

```
Minimum Recommended: n ≥ 3 biological replicates per condition
```

For optimal statistical power, 5-6 replicates per group is recommended, especially when expecting small effect sizes or high biological variability.



Replication Best Practices

- **Prioritize biological replicates:** They provide information about true biological variation
- **Minimum of 3 replicates:** Required for basic statistical tests (t-tests, DESeq2, edgeR)
- **5-6 replicates optimal:** Provides good power to detect moderate changes (2-fold) with reasonable FDR
- **Avoid pooling:** Pooling samples reduces the ability to assess biological variation
- **Technical replicates rarely needed:** Modern platforms have <5% technical CV

Replicates	Statistical Power	Recommended For
n = 2	Insufficient - No statistical testing	Pilot studies only
n = 3	Minimal - Detects large changes only	Well-defined systems, large effects
n = 5-6	Good - Detects moderate changes (1.5-2 fold)	Most RNA-seq experiments
n ≥ 10	Excellent - Detects subtle changes	Complex studies, small effect sizes

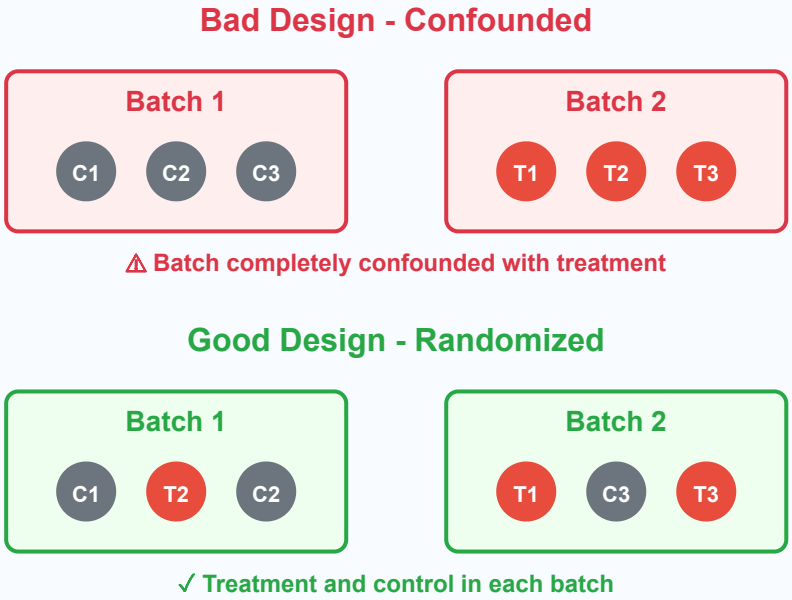
3 Batch Effect Prevention

Batch effects are systematic, non-biological differences between groups of samples processed at different times or under different conditions. They can confound biological signals and lead to false conclusions if not properly controlled.

Common sources of batch effects:

- Different processing days/times
- Different technicians or laboratories
- Different reagent lots or kits
- Different sequencing runs or flow cells
- Seasonal or environmental variations

The gold standard for preventing batch effects is **complete randomization**, where samples from all experimental groups are randomly distributed across all batches.



Batch Effect Prevention Strategies

- **Randomize sample processing:** Distribute conditions evenly across all batches
- **Process all samples together:** When possible, process all samples in a single batch

- **Block design:** If batches are unavoidable, ensure each batch contains all conditions
- **Record metadata:** Document all processing information (date, technician, kit lot, etc.)
- **Use computational correction:** Apply batch correction methods (ComBat, limma) if needed
- **Include controls:** Add control samples to each batch to monitor batch effects

Important Note: Batch effects cannot be fully corrected computationally if they are completely confounded with the biological variable of interest. Prevention through proper experimental design is crucial.

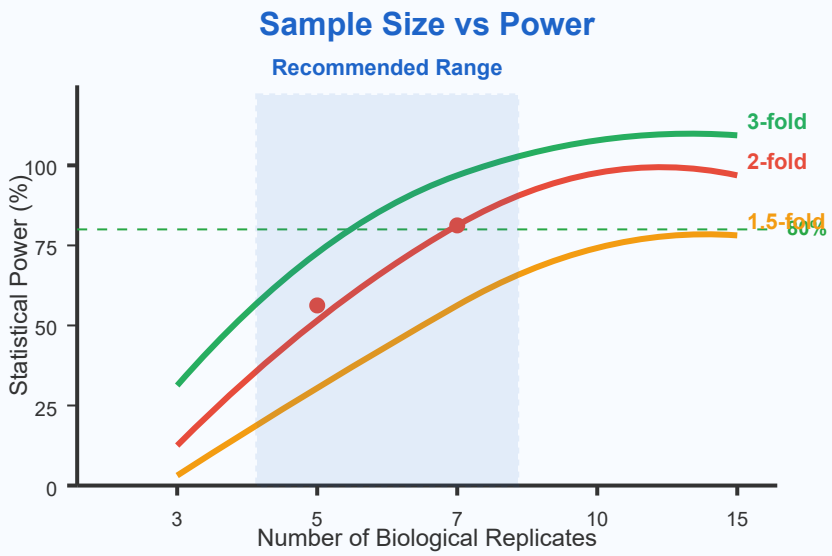
4 Power Analysis

Statistical power is the probability of detecting a true effect when it exists. Power analysis helps determine the appropriate sample size needed to detect biologically meaningful differences with acceptable confidence.

Key factors affecting power:

- **Effect size:** The magnitude of difference you want to detect (e.g., 2-fold change)
- **Biological variability:** Higher variability requires more samples
- **Sample size:** More replicates increase power
- **Significance level (α):** Typically 0.05, adjusted for multiple testing
- **Sequencing depth:** Affects detection of lowly expressed genes

```
Target Power: 80% (β = 0.20) at FDR < 0.05
```



Power Analysis Recommendations

- **Perform power analysis before starting:** Use tools like RNASeqPower, PROPER, or Scotty

- **Use pilot data when available:** Estimate variability from preliminary experiments
- **Consider effect size:** Smaller effects require more samples for detection
- **Account for multiple testing:** Adjust significance threshold for thousands of genes tested
- **Balance depth vs replicates:** More replicates usually better than extreme depth
- **Plan for dropout:** Include extra samples in case of technical failures

Fold Change	Variability (CV)	Samples Needed (80% power)
2-fold	Low (20%)	3-4 per group
2-fold	Moderate (40%)	5-6 per group
1.5-fold	Low (20%)	6-8 per group
1.5-fold	Moderate (40%)	12-15 per group

Rule of Thumb: For detecting 2-fold changes in moderately variable genes, aim for 5-6 biological replicates per condition. This provides good power while remaining cost-effective.

5 Cost Optimization

RNA-seq experiments involve significant costs, and optimizing the allocation of resources between sequencing depth and sample number is crucial for maximizing statistical power within budget constraints.

Key cost considerations:

- **Sequencing depth:** Number of reads per sample
- **Sample number:** Number of biological replicates
- **Library preparation:** Often the most expensive per-sample cost
- **Platform choice:** NovaSeq vs NextSeq vs other platforms
- **Multiplexing:** Pooling multiple samples per lane

```
Recommended: 20-30M reads per sample for standard DE analysis
```

Cost Optimization Strategies

- **Prioritize sample number:** 6 samples at 25M reads beats 3 samples at 100M reads
- **Reduce unnecessary depth:** Beyond 30M reads, returns diminish for standard DE analysis

- **Maximize multiplexing:** Pool compatible samples to reduce per-sample sequencing costs
- **Choose appropriate platform:** Match platform to depth requirements (NovaSeq for many samples)
- **Consider pilot studies:** Small pilot can inform optimal depth for main study
- **Negotiate with core facilities:** Bulk pricing for large projects

Application	Recommended Depth	Rationale
Differential Expression (Human/Mouse)	20-30M reads	Sufficient for ~15K genes with good coverage
Novel Transcript Discovery	50-100M reads	Deeper coverage needed for rare transcripts
Allele-specific Expression	60-100M reads	Higher depth for statistical confidence
Single-cell RNA-seq	50-100K reads/cell	Shallow per cell, many cells
Microbial Transcriptomics	5-10M reads	Smaller transcriptomes require less depth

Cost-Effectiveness Principle: After adequate coverage (~20M reads for mammalian samples), investing in additional biological replicates provides much better statistical power per dollar than increasing sequencing depth.

🎯 Golden Rule of RNA-seq Design

More biological replicates with moderate sequencing depth (5-6 samples × 20-30M reads)
provides superior statistical power compared to
fewer samples with extreme depth (3 samples × 100M reads)

Remember: You cannot compute your way out of poor experimental design!