Lecture 5:

Transcriptomics and Single-Cell Analysis

From bulk to single-cell • Cell atlas projects • Resolution revolution

Introduction to Biomedical Data Science

Lecture Contents

Part 1/3:

Bulk RNA-seq

- Expression profiling
- Differential analysis
- Pathway enrichment
- Time series

RNA-seq Workflow

Experimental Design

Plan your study carefully with appropriate controls and biological questions

Replication Strategies

Biological replicates (≥3) are essential for statistical power

Batch Effect Prevention

Randomize sample processing to avoid confounding variables

Power Analysis

Determine sample size needed to detect biological effects

Cost Optimization

Balance sequencing depth and sample number for your budget

Key Recommendation

More biological replicates with moderate depth > few samples with deep sequencing

Library Preparation Methods

PolyA Selection

Enriches mRNA by capturing poly-adenylated transcripts

Ribosomal Depletion

Removes rRNA to capture all RNA types including non-coding

Strand Specificity

Preserves information about which DNA strand was transcribed

UMI Incorporation

Unique Molecular Identifiers enable accurate quantification

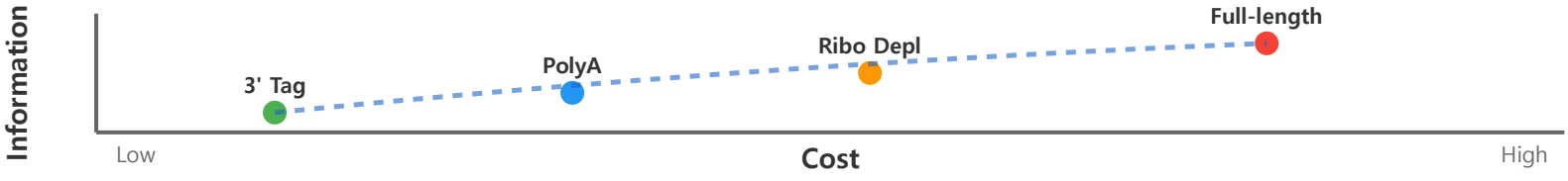
3' Tag-seq

Sequences only 3' ends - cost effective for counting

Full-length Coverage

Complete transcript coverage for isoform analysis

Trade-off: **Cost vs. Information Content**



Normalization Methods

RPKM/FPKM Issues

Reads/Fragments Per Kilobase Million - biased by composition

TPM Calculation

Transcripts Per Million - better for comparison

DESeq2 Normalization

Median-of-ratios method for differential expression

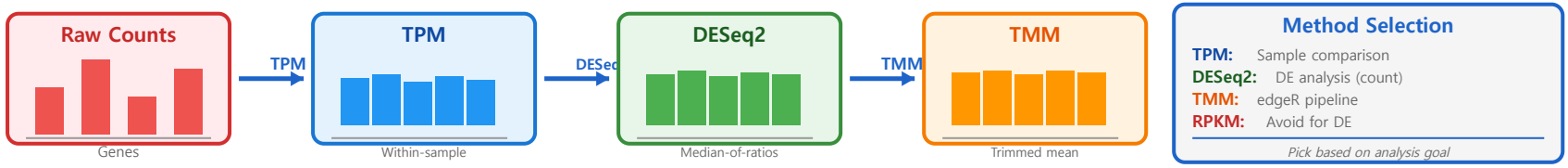
TMM Method

Trimmed Mean of M-values - robust to outliers

Batch Correction

ComBat, limma removeBatchEffect for technical variation

💡 **Choose normalization method based on your downstream analysis goals**



Differential Expression

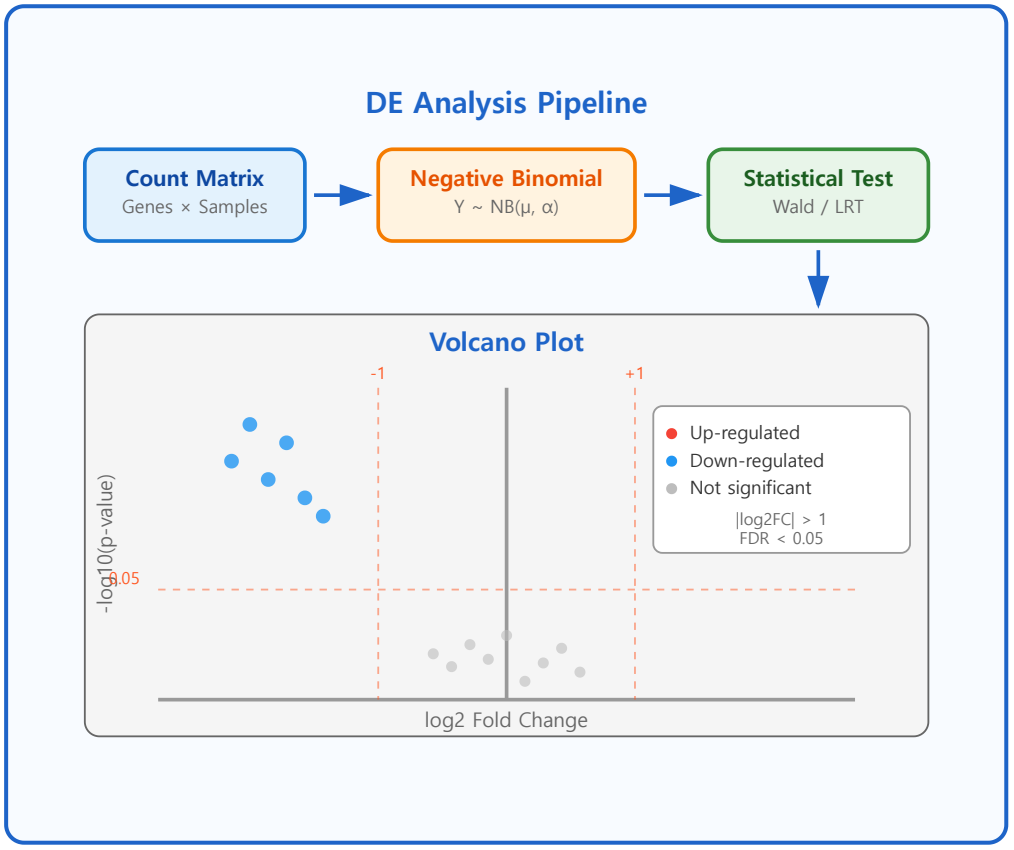
Statistical Models
Account for biological and technical variability

Negative Binomial
Models count data with overdispersion

Fold Change Thresholds
Typically |log2FC| > 1 for biological significance

FDR Control
False Discovery Rate < 0.05 for multiple testing

Volcano Plots
Visualize FC vs statistical significance



💡 Balance statistical significance with biological relevance

Statistical Testing

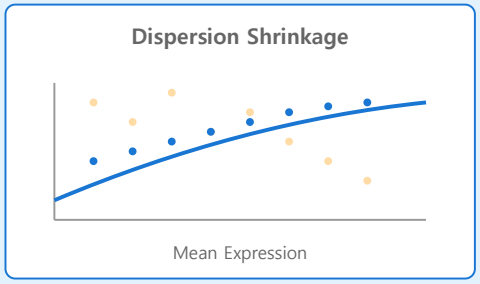
RNA-seq Statistical Methods Comparison

DESeq2

$Y \sim NB(\mu, \alpha)$

Key Features:

- Shrinkage estimation of dispersion
- Size factor normalization
- Wald test / LRT

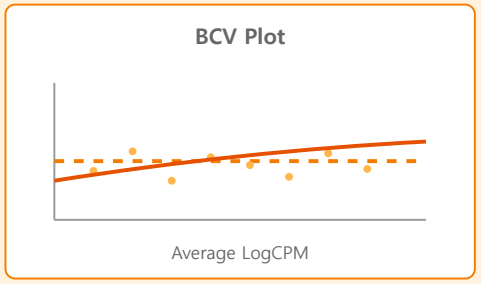


edgeR

$Y \sim NB(\mu, \varphi)$

Key Features:

- Empirical Bayes methods
- TMM normalization
- Quasi-likelihood F-test

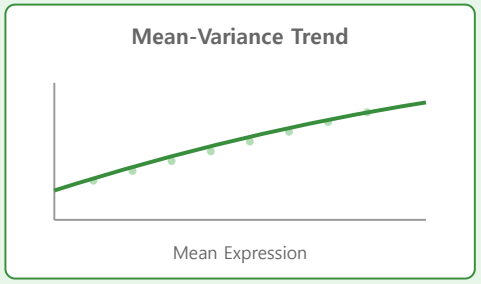


limma-voom

$\log(Y) \sim N(\mu, \sigma^2)$

Key Features:

- Transform to log-space
- Precision weights (voom)
- Linear modeling



Performance: All methods perform similarly with proper use • Choose based on experimental design and analysis goals

💡 DESeq2 and edgeR are most widely used and well-validated

Multiple Testing Correction

The Multiple Testing Problem

Without Correction

Testing 20,000 genes at α = 0.05:

Expected false positives = 20,000 × 0.05 = 1,000 genes!

19,000 True Null | 1,000 FP

With FDR Control

Benjamini-Hochberg at FDR = 0.05:

Controls proportion of false discoveries ≤ 5%

95 TP | 5 | Not called significant

Correction Methods

Bonferroni (FWER)

Formula:

p_adjusted = p × n

Very Conservative

- Controls FWER < α
- Few false positives
- Many false negatives

Benjamini-Hochberg

Formula:

p(i) ≤ (i/m) × α

Balanced

- Controls FDR < α
- More discoveries
- Standard for genomics

Q-value

Concept:

Min FDR at threshold

Adaptive

- Estimates π_0
- Data-driven
- More powerful

Permutation

Approach:

Empirical null dist.

Non-parametric

- No assumptions
- Computationally intensive
- Complex designs

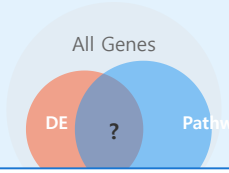
Trade-off: Strict control (Bonferroni) → Few discoveries | Relaxed control (FDR) → More discoveries with controlled error

💡 Testing 20,000 genes requires careful multiple testing correction

Pathway Analysis

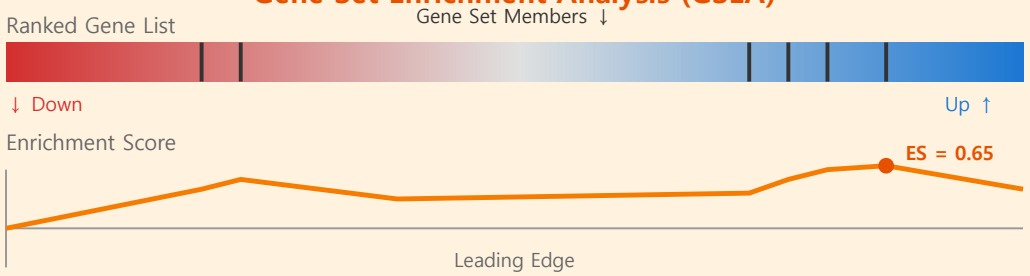
From Genes to Biological Pathways

Over-representation (ORA)



Fisher's Exact Test → p-value

Gene Set Enrichment Analysis (GSEA)



Pathway Databases

Gene Ontology (GO)

- Biological Process
- Molecular Function
- Cellular Component

Hierarchical ontology

KEGG

- Metabolic pathways
- Signaling pathways
- Disease pathways

Curated pathway maps

Reactome

- Reactions
- Complexes
- Entities

Peer-reviewed pathways

MSigDB

- Hallmark sets
- Curated sets
- Computational sets

Broad collection

ORA: Simple but threshold-dependent | GSEA: Uses all genes, more powerful | Choose database based on biological question

💡 Pathways provide biological context for gene expression changes

Part 2/3:

Single-Cell Technologies

- Technology overview
- Cell isolation
- Quality control
- Analysis challenges

scRNA-seq Overview

Evolution & Comparison of Single-Cell Technologies

Throughput vs Depth Trade-off

Cell Atlases	Development	Disease	Drug Discovery
Human Cell Atlas	Lineage tracing	Tumor heterogeneity	Target identification

💡 Revolution in understanding cellular heterogeneity

Droplet-based Methods

10X Genomics Platform

Most widely used - Chromium platform with GEMs

Drop-seq Principles

Co-encapsulation of cells and barcoded beads

InDrop Technology

Hydrogel beads with photocleavable barcodes

Barcode Design

Cell barcode + UMI for molecular counting

Doublet Detection

Computational and experimental QC for multiplets

💡 High throughput but lower sensitivity per cell

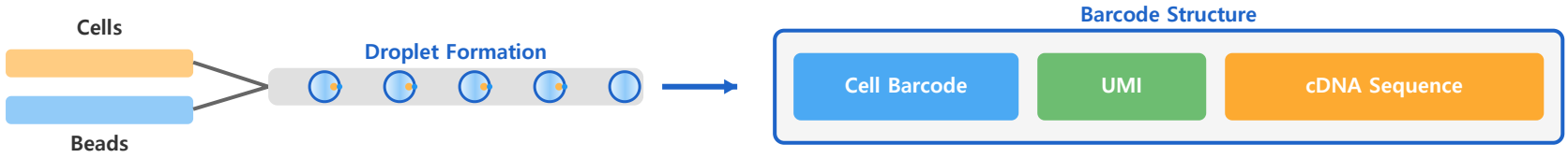


Plate-based Methods

Plate-based scRNA-seq Workflow

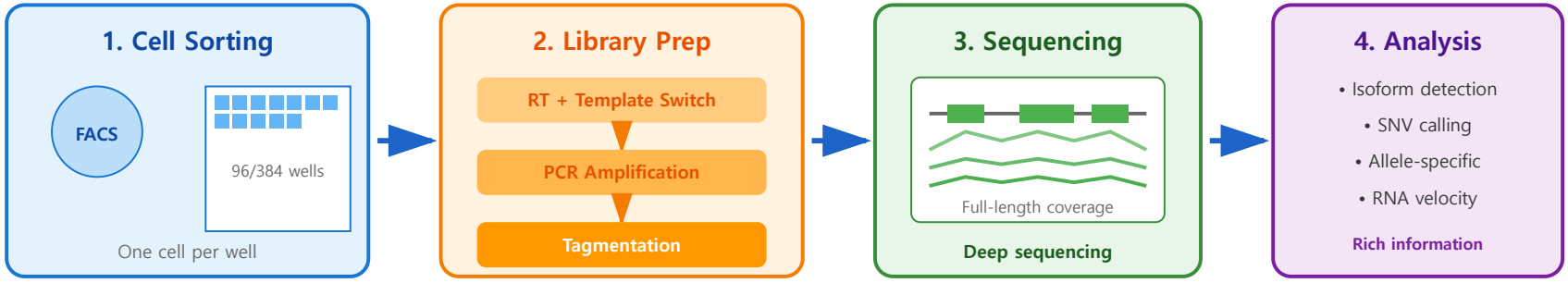


Plate-based Methods Comparison

Smart-seq2/3

- ✓ Full-length transcripts
- ✓ Highest sensitivity
- ✓ Isoform analysis
- ✗ No UMIs
- ✗ Higher cost/cell

MARS-seq

- ✓ UMI incorporation
- ✓ Automated
- ✓ 3' counting
- ✓ Cost-effective
- ✗ 3' bias

CEL-seq2

- ✓ Linear amplification
- ✓ UMIs
- ✓ Low bias
- ✓ Multiplexing
- ✗ Complex protocol

💡 Lower throughput but deeper sequencing per cell

Data Preprocessing

Cell Filtering

Remove low-quality cells and empty droplets

Gene Filtering

Exclude genes detected in too few cells

Normalization Methods

Account for sequencing depth and composition

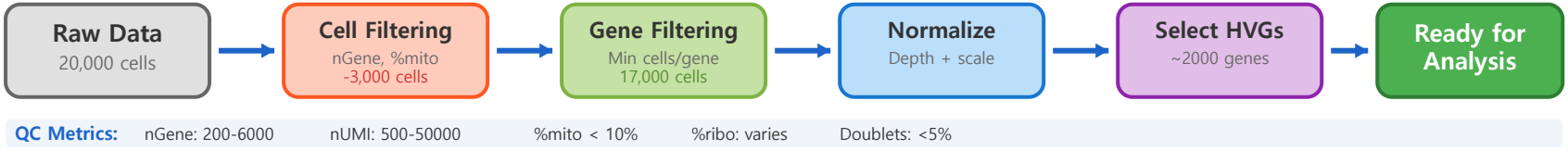
Imputation Strategies

Handle dropout events (use with caution)

Batch Effects

Technical variation from sample processing

💡 Quality control is critical for downstream analysis



Dimensionality Reduction

PCA for scRNA-seq

First step to reduce noise and computational burden

t-SNE Principles

Preserves local structure, stochastic

UMAP Advantages

Faster, preserves global + local structure

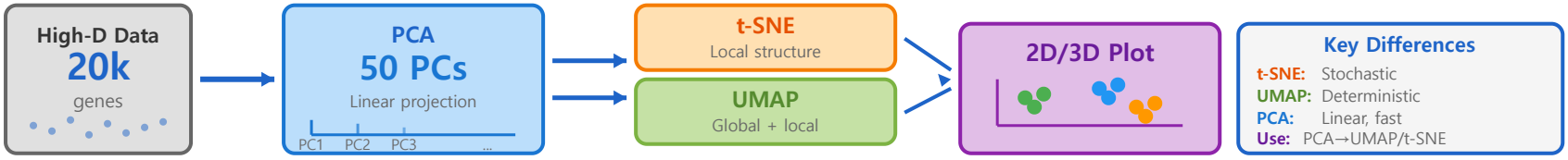
Diffusion Maps

Captures continuous trajectories

Parameter Selection

Perplexity, n_neighbors affect results

💡 **Visualization != clustering - use both appropriately**



Clustering Methods

Graph-based Clustering

Build kNN graph then find communities

Leiden Algorithm

Improved Louvain with better guarantees

K-means Adaptations

SC3 uses consensus clustering

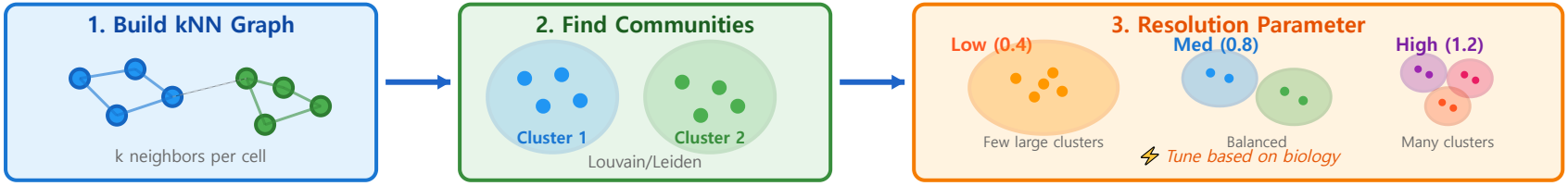
Resolution Selection

Higher resolution = more clusters

Stability Analysis

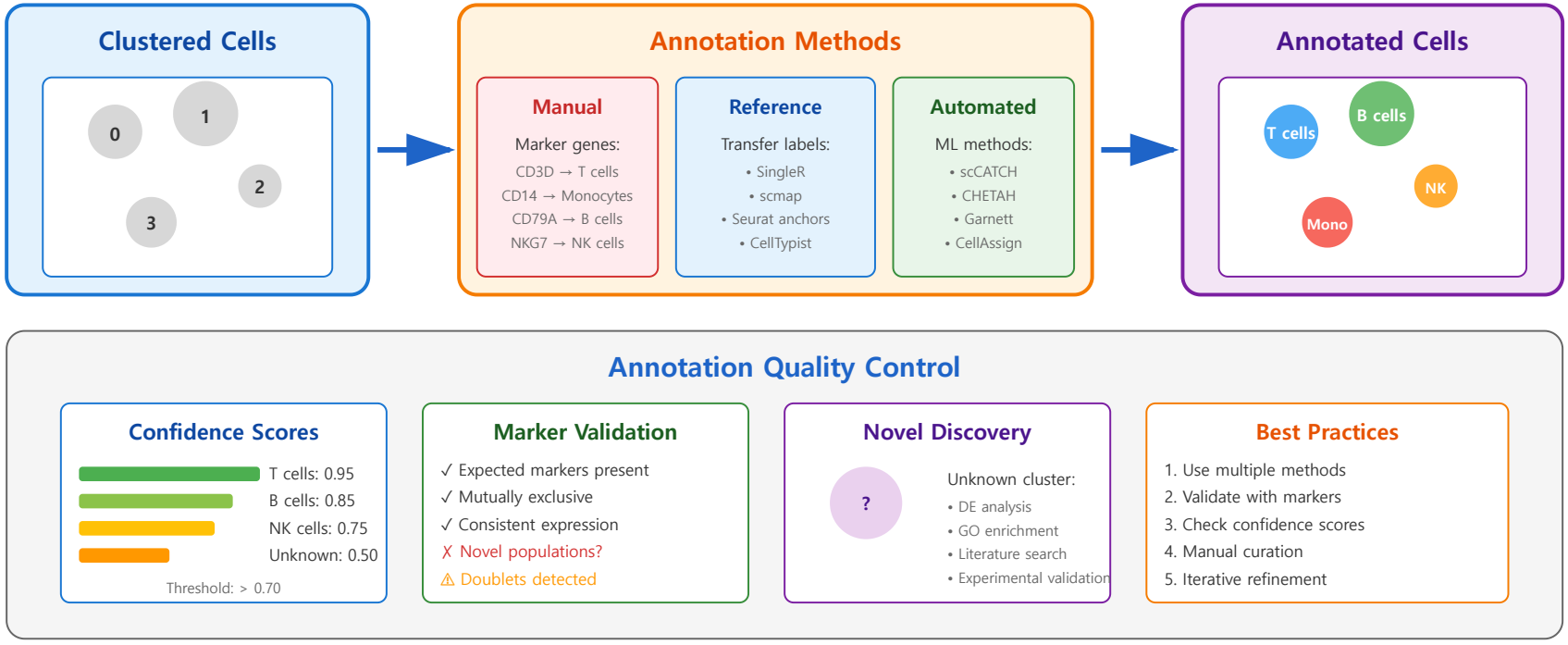
Bootstrap to assess cluster robustness

💡 No single correct clustering - depends on biological question



Cell Type Annotation

Cell Type Annotation Pipeline



💡 Combine automated tools with manual curation

Trajectory Analysis

Pseudotime Inference

Order cells along developmental paths

Branching Processes

Identify cell fate decisions

Monocle Algorithm

Reverse graph embedding

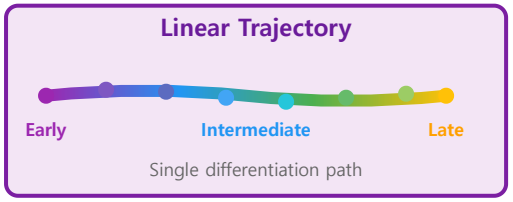
Slingshot Method

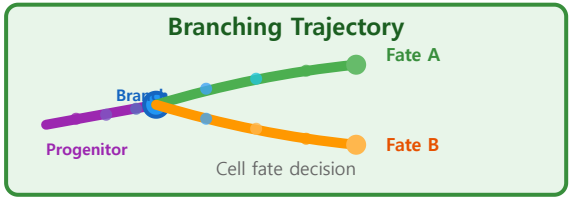
Cluster-based trajectory inference

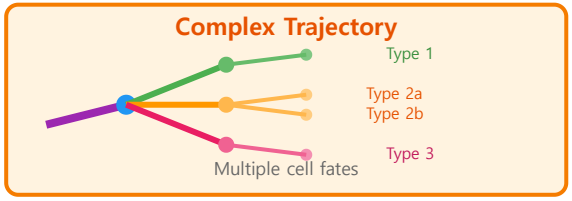
Validation Approaches

Known genes, time-series data

💡 Assumes continuous progression - verify biological relevance







Part 3/3:

Advanced Methods

- Spatial context
- Multi-modal data
- Velocity analysis
- Communication inference

Spatial Transcriptomics

Visium Technology

10X spatial - 55µm spots, whole transcriptome

MERFISH Principles

Multiplexed error-robust FISH, subcellular resolution

seqFISH Evolution

Sequential FISH with barcoding, 10,000+ genes

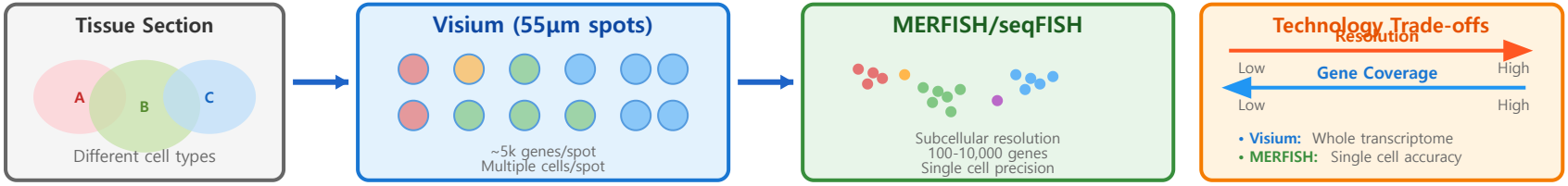
Slide-seq Methods

Bead-based spatial barcoding, 10µm resolution

Resolution Trade-offs

Gene coverage vs spatial resolution vs throughput

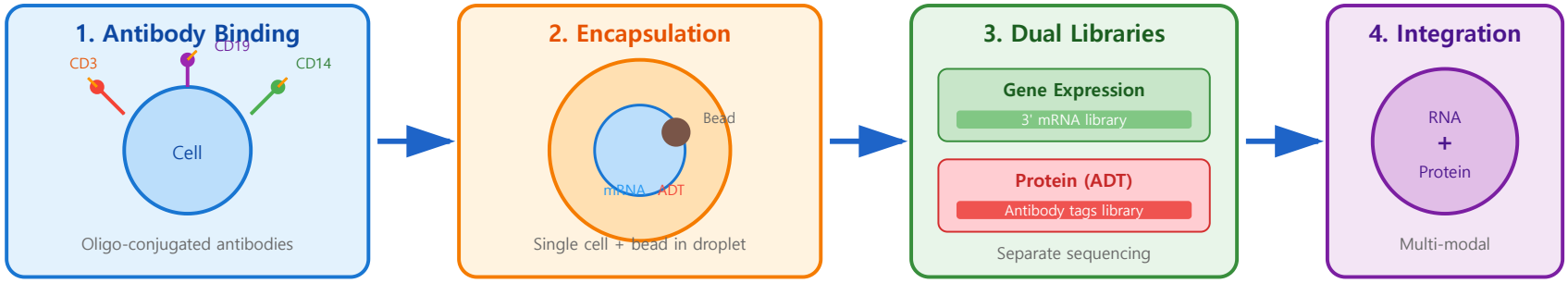
💡 **Spatial context reveals tissue architecture and cell-cell interactions**



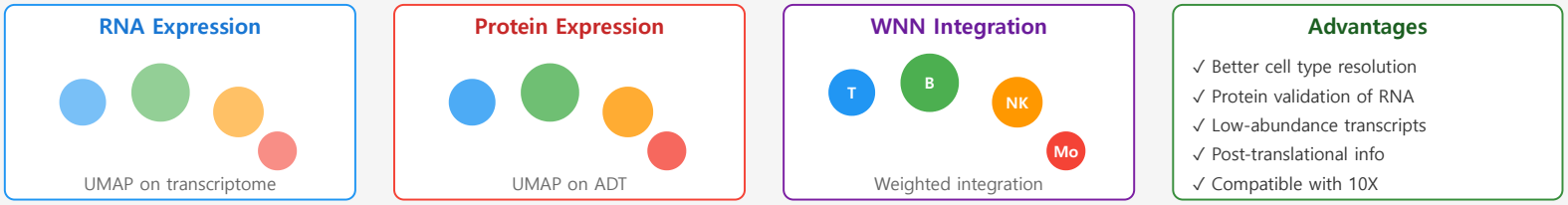
CITE-seq

Cellular Indexing of Transcriptomes and Epitopes by Sequencing

CITE-seq Mechanism



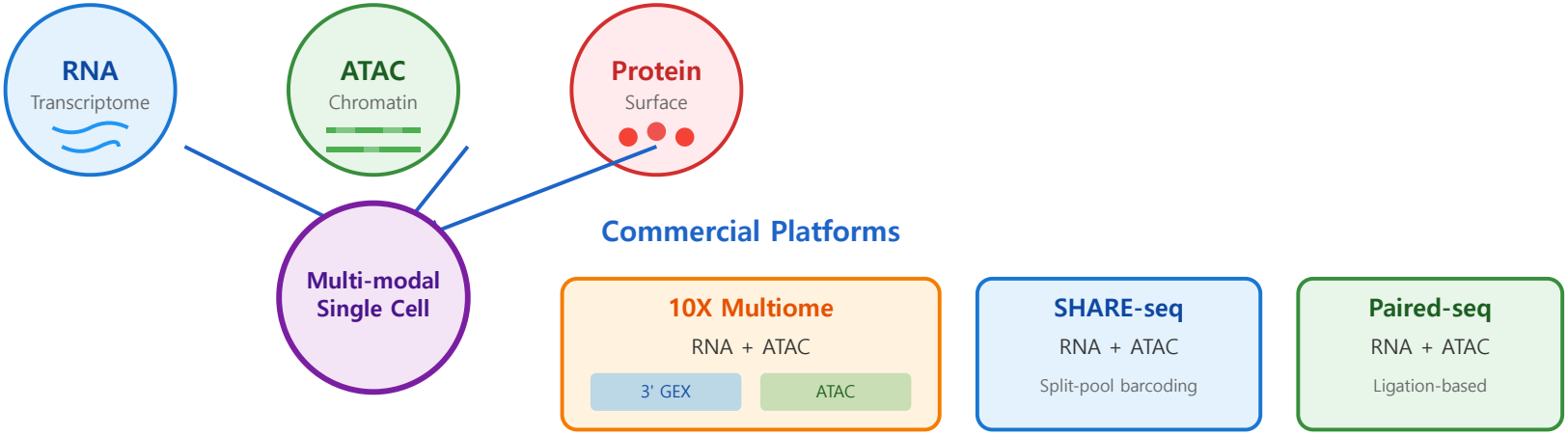
CITE-seq Data Visualization



💡 Bridges transcriptomics and proteomics at single-cell level

Multimodal Omics

Single-Cell Multimodal Technologies



Biological Insights from Multimodal Integration

💡 Multi-omics reveals regulatory mechanisms

RNA Velocity

Spliced/Unspliced Ratio

Infer transcriptional dynamics from steady-state

Velocity Estimation

Predict future cell states

Dynamic Models

Account for transcription, splicing, degradation

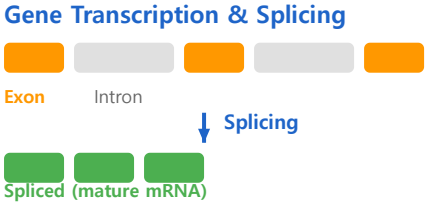
scVelo Improvements

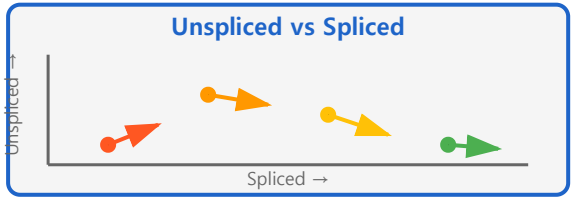
Dynamical model, latent time

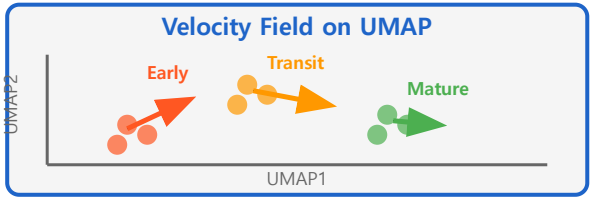
Interpretation

Direction and magnitude of cell state changes

RNA velocity adds temporal dimension to snapshots







Cell-Cell Communication

Inferring Cell-Cell Interactions from scRNA-seq

Ligand-Receptor Interaction

TNF→TNFR | IL6→IL6R | VEGFA→FLT1

Communication Inference Tools

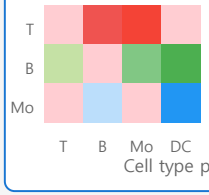
CellPhoneDB
- Statistical framework
- Permutation test
- Multi-subunit
- Spatial optional

NicheNet
- Gene regulatory
- Prior knowledge
- Target prediction
- Prioritization

CellChat
- Network analysis
- Pattern recognition
- Visualization
- Comparison

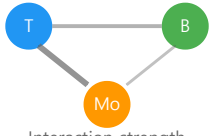
Typical Analysis Outputs

Interaction Heatmap



Cell type pairs

Network Graph

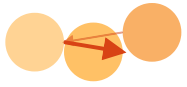


Interaction strength

Dot Plot

Size: p-value
Color: mean expr

Spatial Context



Physical proximity
→ Higher confidence

💡 Infer cellular communication from expression patterns

Batch Effect Correction

MNN Correction

Mutual nearest neighbors for batch alignment

Harmony Algorithm

Iterative clustering and correction

LIGER Integration

Integrative non-negative matrix factorization

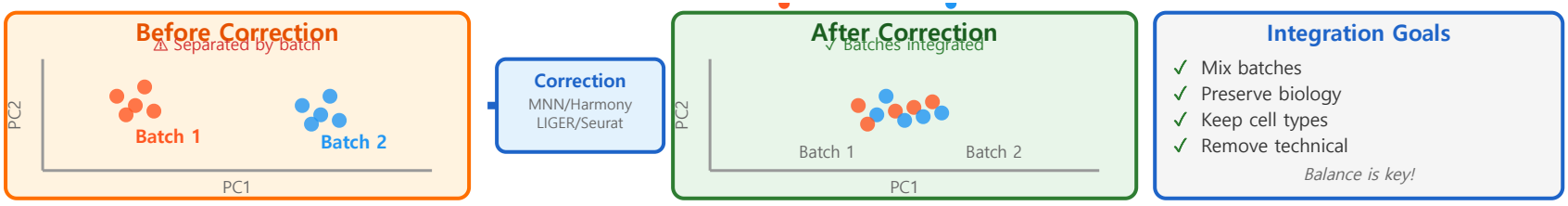
Seurat Integration

Canonical correlation analysis + anchors

Benchmark Studies

Compare methods on simulated and real data

💡 **Critical for multi-sample and multi-technology integration**



Integration Methods

Anchor-based Methods

Seurat, LIGER find correspondence between datasets

Deep Learning Approaches

scVI, scGAN learn shared latent space

Reference Building

Create comprehensive cell atlases

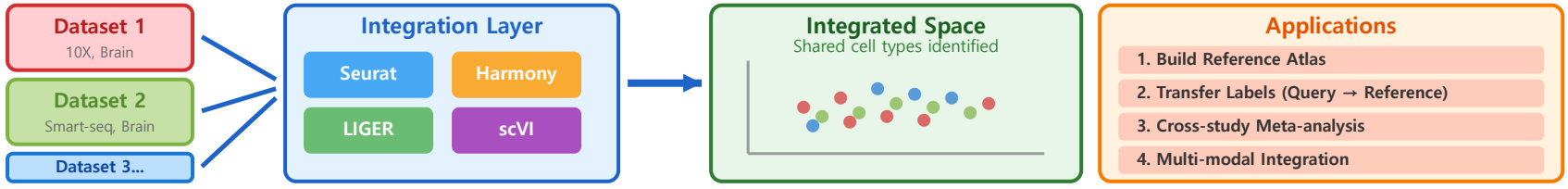
Query Mapping

Project new data onto reference

Performance Metrics

Biological conservation vs batch mixing

💡 Integration enables meta-analysis and transfer learning



Hands-on: Seurat Tutorial

Seurat v5 Standard Workflow

1. Data Loading & QC

```
# Read 10X data
data <- Read10X("filtered_feature_bc_matrix/")
seurat <- CreateSeuratObject(data, min.cells=3)
```

2. QC Filtering

```
# Calculate mitochondrial %
seurat[["percent.mt"]] <- PercentageFeatureSet(seurat, "^MT-")
seurat <- subset(seurat, nFeature_RNA > 200 & percent.mt < 10)
```

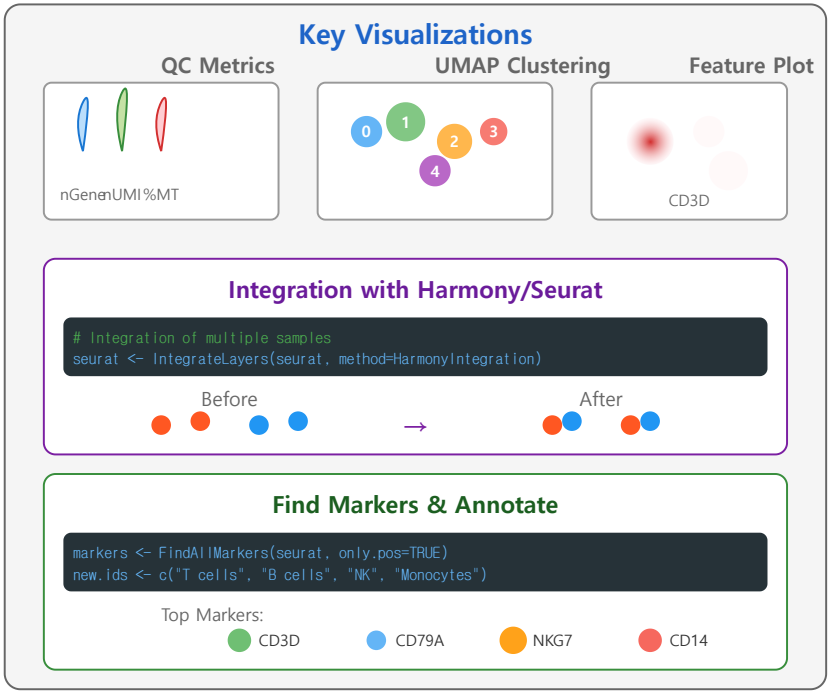
3. Normalization & Scaling

```
seurat <- NormalizeData(seurat)
seurat <- FindVariableFeatures(seurat, nfeatures=2000)
seurat <- ScaleData(seurat)
```

4. Dimension Reduction & Clustering

```
seurat <- RunPCA(seurat) %>% RunUMAP(dims=1:30)
seurat <- FindNeighbors(seurat) %>% FindClusters(res=0.5)
DimPlot(seurat, label=TRUE) + NoLegend()
```

Key Visualizations

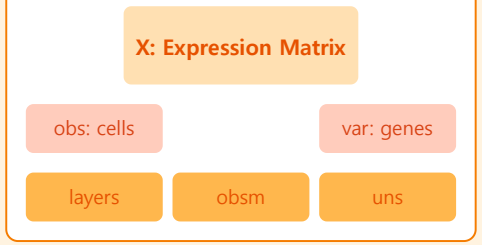


💡 Most widely used R package for scRNA-seq analysis

Hands-on: Scanpy Analysis

Scanpy: Python-based Single Cell Analysis

AnnData Structure



Standard Workflow

```python
# Import libraries
import scanpy as sc
import pandas as pd
adata = sc.read_10x_h5('file.h5')
```

```python
# QC and filter
sc.pp.calculate_qc_metrics(adata)
sc.pp.filter_cells(adata, min_genes=200)
sc.pp.filter_genes(adata, min_cells=3)
```

```python
# Normalize and find HVGs
sc.pp.normalize_total(adata)
sc.pp.log1p(adata)
sc.pp.highly_variable_genes(adata)
```

```python
# Dimension reduction
sc.tl.pca(adata)
sc.pp.neighbors(adata)
sc.tl.umap(adata)
```

Advanced Scanpy Features

Trajectory Analysis

```python
sc.tl.paga(adata)
sc.tl.dpt(adata)
```

RNA Velocity

```python
import scvelo as scv
scv.tl.velocity(adata)
```

GPU Acceleration

```python
import rapids_singlecell
# 100x speedup!
```

Large datasets

Data Integration

```python
sc.external.pp.harmony_
integrate(adata, 'batch')
```

💡 Python ecosystem with extensive documentation

Thank you!

Key Applications

- Disease studies - Cell type changes in pathology
- Development biology - Cell fate trajectories
- Drug discovery - Target identification and validation
- Clinical futures - Diagnostic and therapeutic applications