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: Bulk RNA-seq Analysis

Part 2: Single-Cell Technologies

Part 3: Advanced Methods and Integration

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

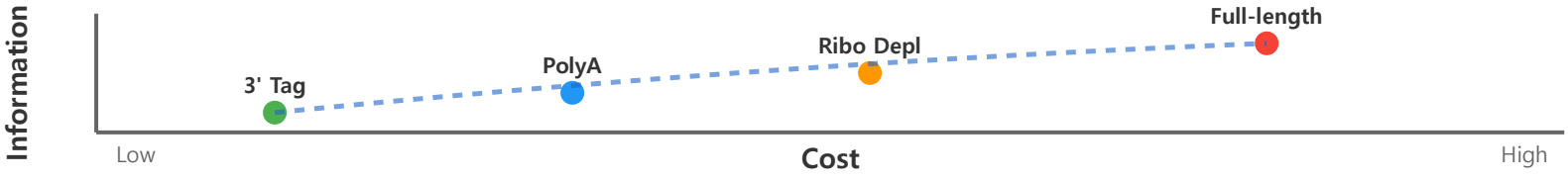
3' Tag-seq

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

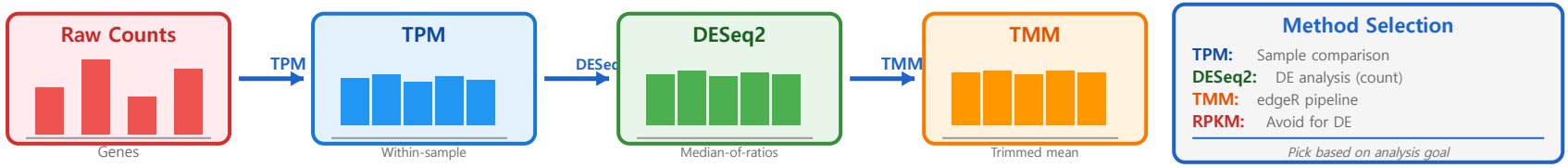
TMM Method

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

DESeq2 Framework

Generalized linear model with negative binomial

edgeR Approach

Empirical Bayes methods for dispersion estimation

Limma-voom

Transform counts for linear modeling

Nonparametric Methods

Rank-based tests when assumptions violated

Benchmarking Results

Performance varies by experimental design

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

Multiple Testing Correction

FDR vs FWER

False Discovery Rate vs Family-Wise Error Rate

Benjamini-Hochberg

Controls FDR - less conservative than Bonferroni

Q-value Estimation

Minimum FDR at which a test is called significant

Permutation Tests

Empirical null distribution for complex designs

Power Considerations

More tests = need more samples or larger effects

💡 **Testing 20,000 genes requires careful multiple testing correction**

Pathway Analysis

Gene Set Enrichment

Identify coordinated changes in functional groups

Over-representation

Fisher's exact test for enrichment in DE genes

GSEA Algorithm

Uses full ranked gene list, not just DE genes

Pathway Databases

GO, KEGG, Reactome, MSigDB

Network Analysis

Protein-protein interactions and regulatory networks

💡 **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

Historical Development

From Tang et al. 2009 to modern high-throughput platforms

Technology Comparison

Droplet vs plate-based methods

Throughput vs Depth

10X: 10K cells/run, Smart-seq: 100-1000 genes more

Cost per Cell

$0.05-1.00 depending on platform and scale

Applications

Cell atlases, development, disease, drug screening

💡 **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

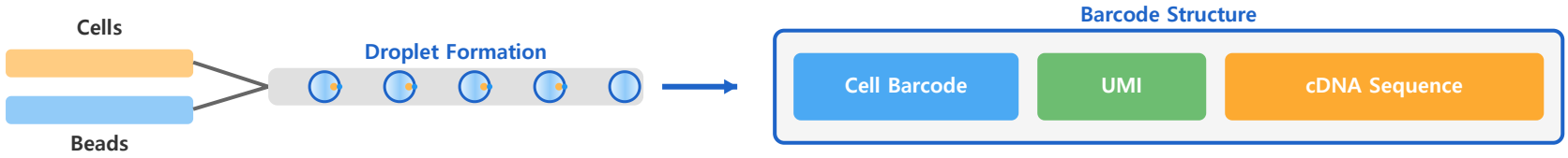


Plate-based Methods

Smart-seq Protocols

Full-length transcripts with high sensitivity

MARS-seq

Automated plate-based with UMIs

CEL-seq

Linear amplification with in vitro transcription

Full-length Advantages

Isoform analysis and better gene coverage

Automation

Liquid handling robots for reproducibility

💡 **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

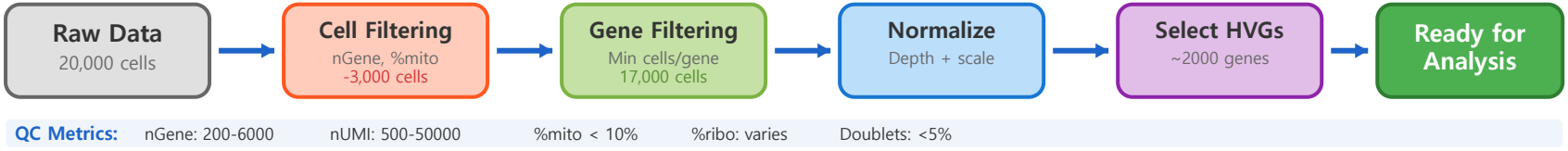
Imputation Strategies

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

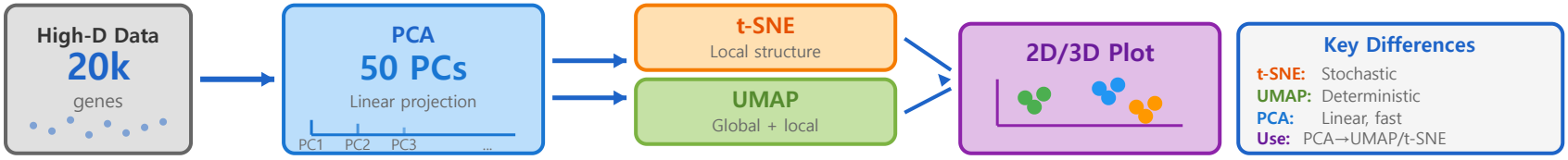
Diffusion Maps

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

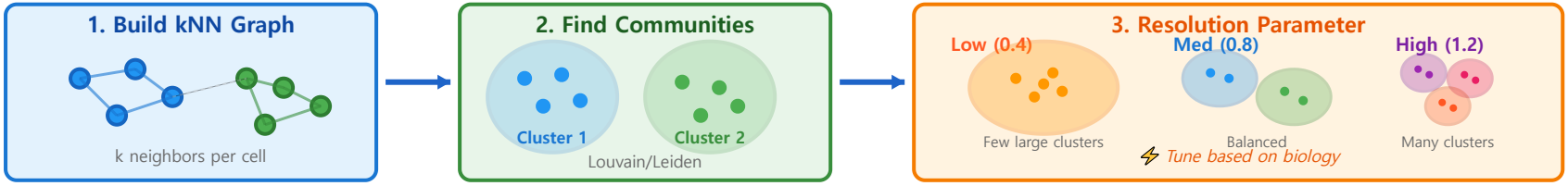
Resolution Selection

Higher resolution = more clusters

Stability Analysis

Bootstrap to assess cluster robustness

💡 No single correct clustering - depends on biological question



Cell Type Annotation

Marker Genes

Known markers from literature and databases

Reference Mapping

Transfer labels from annotated datasets

Automated Methods

SingleR, scmap, CellTypist

Confidence Scores

Assess certainty of annotations

Novel Cell Types

Identify and characterize unknown populations

💡 **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

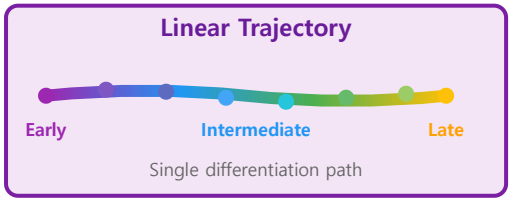
Slingshot Method

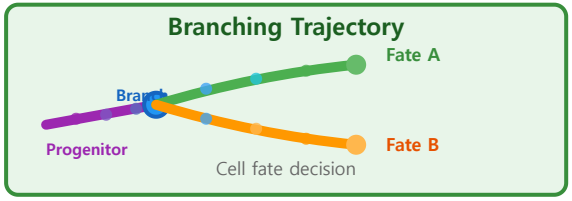
Cluster-based trajectory inference

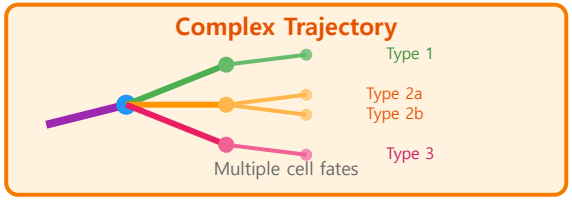
Validation Approaches

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

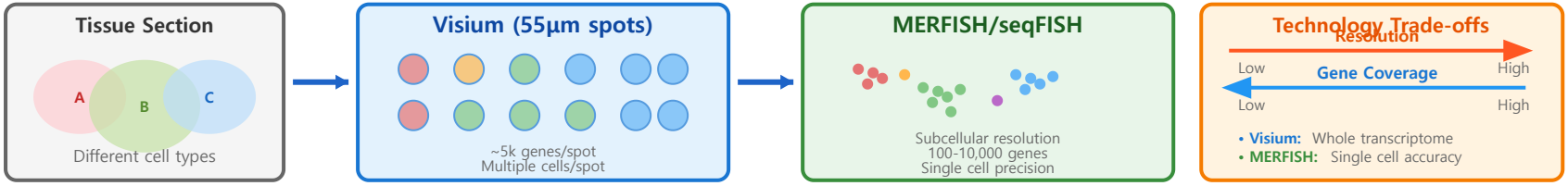
Slide-seq Methods

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

Antibody-derived Tags

Oligonucleotide-conjugated antibodies

Protein Quantification

Surface protein expression alongside RNA

ADT Processing

Separate library for antibody-derived tags

Multi-modal Integration

Weighted nearest neighbor analysis

Panel Design

Select antibodies for biological questions

💡 **Bridges transcriptomics and proteomics at single-cell level**

Multimodal Omics

SHARE-seq

Simultaneous RNA and chromatin accessibility

Paired-seq

Co-assay of transcriptome and accessible chromatin

10X Multiome

Commercial GEX + ATAC in same cells

Integration Challenges

Different data types, scales, and sparsity

Biological Insights

Link regulatory elements to gene expression

💡 **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

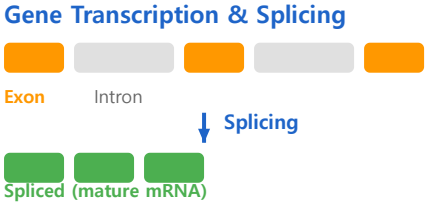
scVelo Improvements

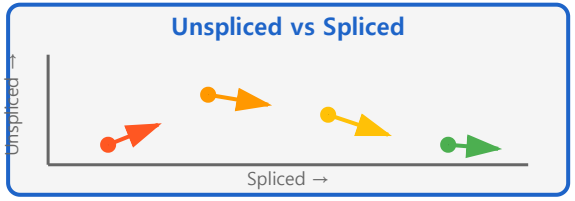
Dynamical model, latent time

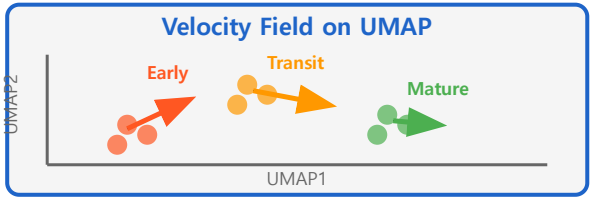
Interpretation

Direction and magnitude of cell state changes

RNA velocity adds temporal dimension to snapshots







Cell-Cell Communication

Ligand-Receptor Databases

CellTalkDB, CellPhoneDB, NicheNet

CellPhoneDB

Statistical framework for interaction significance

NicheNet

Predict ligands affecting target genes

Spatial Considerations

Physical proximity in spatial data

Validation Methods

Experimental verification with perturbations

💡 **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

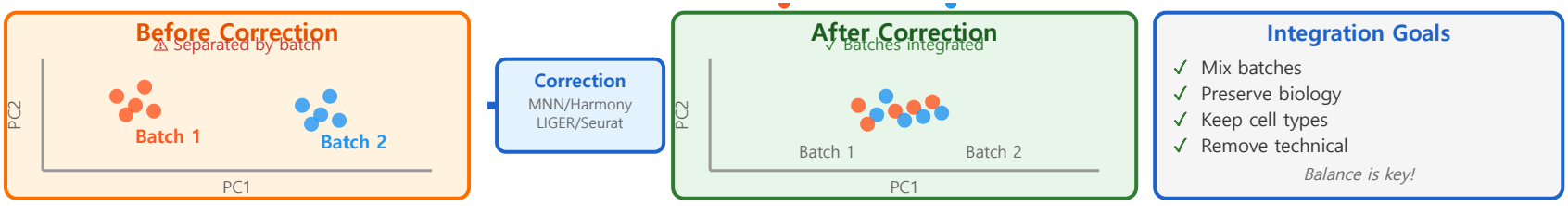
Seurat Integration

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

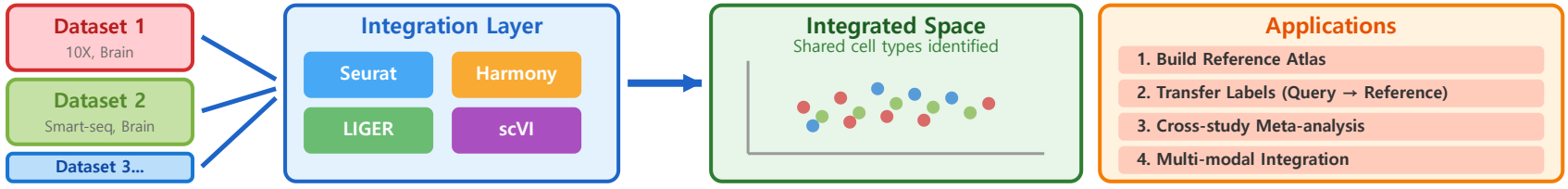
Query Mapping

Project new data onto reference

Performance Metrics

Biological conservation vs batch mixing

💡 Integration enables meta-analysis and transfer learning



Hands-on: Seurat Tutorial

Data Loading

Read 10X data, create Seurat object

QC and Filtering

Mitochondrial %, nFeature, nCount filtering

Standard Workflow

Normalize → Scale → PCA → UMAP → Cluster

Integration Example

Integrate multiple samples or conditions

Visualization

FeaturePlot, DotPlot, VlnPlot

💡 **Most widely used R package for scRNA-seq analysis**

Hands-on: Scanpy Analysis

AnnData Structure

Python alternative - obs, var, X, layers

Python Workflow

Filter → Normalize → HVG → PCA → Leiden

Advanced Analyses

Trajectory, velocity, spatial

GPU Acceleration

RAPIDS integration for large datasets

Interoperability

Convert between Seurat and AnnData

💡 **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