BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 17 Single-cell RNA-seq

5 November, 2025

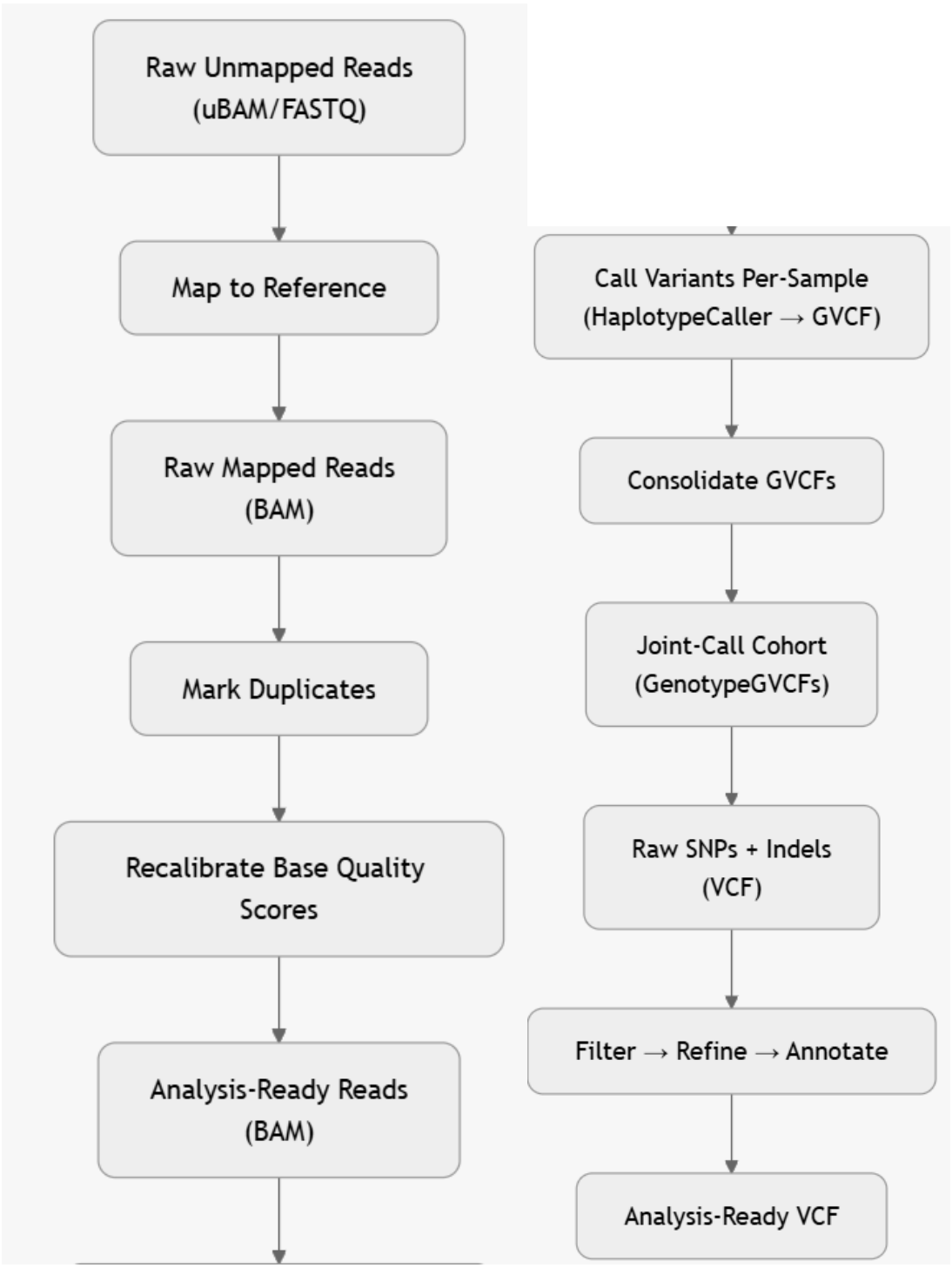
Lecturer: Yu LI

Scriber: Samuel Luo(1155213946)

Level	Study Focus
Genome	Genetic variants, SNPs, CNVs
Epigenome	DNA methylation, histone mods
Transcriptome	Gene expression, RNA-seq, fusions
Proteome	Protein expression
Metabolome	Metabolite profiling

Cancer Links:

- **Genetic variants** → Genome
- **Gene fusion** → RNA-seq
- **Abnormal expression** →
    - Genome (genetic info)
    - Epigenome (environment)
    - Transcriptome (**direct measurement**)


Raw Unmapped Reads (uBAM/FASTQ)
Map to Reference
Raw Mapped Reads (BAM)
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads (BAM)
Call Variants Per-Sample (HaplotypeCaller → GVCF)
Consolidate GVCFs
Joint-Call Cohort (GenotypeGVCFs)
Raw SNPs + Indels (VCF)
Filter → Refine → Annotate
Analysis-Ready VCF

Variant Calling

- **Why each step?** (e.g., remove duplicates)
- **Read file formats:**
    - Convert alignment → **CIGAR string**
    - Interpret **VCF record** changes
- **Factors affecting quality:**
    - Errors vs variants
    - Duplicates
    - **Depth/coverage**
    - Sequence quality

GWAS – Linking Variants to Cancer

Genome-wide association studies

- **~3.5 million SNPs**
- **Adjusted p-value** = p-value / # tests
    - Threshold: **5×10^{-8}**

Gene Fusion Detection via RNA-seq

- **Why RNA-seq?** → Captures **transcripts**, not just DNA
- **Fusion gene DNA** → abnormal **mRNA**
- **Short reads (<300 bp)** → span splice junctions
- **Long reads (>300 bp)** → span fusion junctions directly

Epigenomics: Environment Affects Expression

- **Epigenetic modifications** control gene access
- **Histone marks**, **DNA methylation**, **chromatin accessibility**

Key Assays

Assay	Purpose
ATAC-seq	Genome-wide chromatin accessibility
ChIP-seq	Protein-DNA interactions

Epigenetic Data Pipeline

1. **Read mapping** → BAM
2. **Peak calling** → BED/narrowPeak
3. **Normalization**
4. **Differential analysis**
5. **Motif enrichment**
6. **Visualization**

Why Single-cell RNA-seq?

Bulk RNA-seq Limitations

- Averages expression across **heterogeneous cell populations**
- Masks **cell-type-specific signals**
- Cannot detect **rare cell types** or **subpopulations**

Single-cell Advantages

- Higher resolution of **cellular differences**
- Understand **individual cell function** in context

- Study **tumor microenvironment**, **differentiation paths**, **drug response**

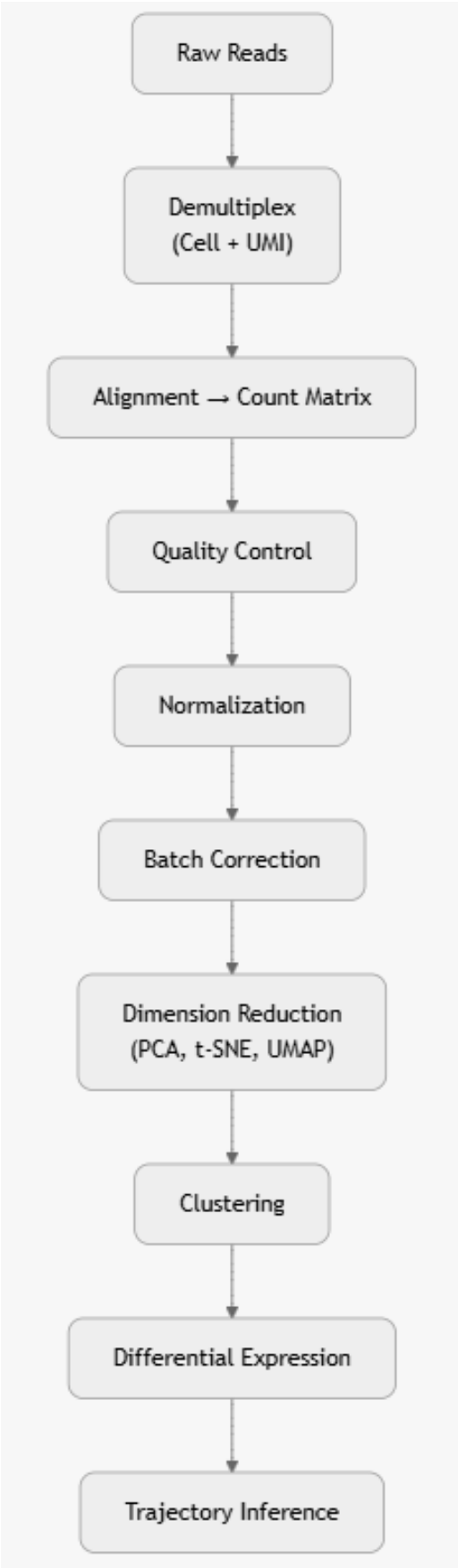
Applications:

- Cancer heterogeneity
- Neural cell classification
- Embryonic development
- Rare cell identification

Single-cell Sequencing Workflow

1. **Cell isolation** (FACS, microfluidics)
2. **Lysis & reverse transcription**
3. **cDNA amplification** (PCR or IVA)
4. **Library preparation**
5. **Sequencing** (Illumina, 10x Genomics)
6. **Barcode/UMI demultiplexing**

Single-cell Data Analytics Pipeline



14. Challenges in scRNA-seq (Pages 17-35 to 17-46)

Challenge	Cause	Solution Approach
Noise	Low input, technical variation	QC filtering (gene count, mito %)
Doublets	Two cells in one droplet	Simulate doublets → detect by similarity
Dropout	Low mRNA → failed capture	Imputation (MAGIC, scImpute)
Batch Effect	Different runs, labs	Harmony, Seurat CCA, Scanorama

Take-home:

- Understand **why** challenges occur
- Know **intuition** behind solutions
- **Technical details not required**

Gene Expression Matrix

	Gene 1	**Gene 2**	**...**	**Gene 25,000**
Cell 1	5	0	...	12
Cell 2	0	3	...	0
...	...	...	...	...
Cell 10,000	8	1	...	7

- Rows: **Cells**
- Columns: **Genes**
- Values: **UMI counts**

Quality Control Metrics

- **# genes expressed per cell**
- **Total counts per cell**
- **% mitochondrial reads** (high → dead/dying cells)

Doublet Detection

- **Scrublet, DoubletFinder**: simulate artificial doublets
- Compare real vs simulated → flag high-similarity cells

Dropout

- **Definition**: Gene expressed but **not detected**
- **Cause**: Low mRNA per cell
- **Trade-off**: More cells → more dropouts (fixed budget)
- **Solution**: Imputation (statistical/ML models)

Batch Effect Correction

- **Cause**: Non-biological factors (lab, reagent, time)
- **Methods**:
    - **Statistical**: ComBat, limma
    - **ML**: MNN, scVI, Harmony

Visualization: t-SNE

Goal: Reduce 25,000D → 2D while preserving clusters

PCA Limitation

- Linear projection → may **destroy clusters**

t-SNE Process

1. **Random initialization** in 2D
2. **Iterative attraction/repulsion**:
    - Nearby high-dim points → attract
    - Distant points → repel
3. Converge to stable layout

Advantages

- Excellent **cluster preservation**

Disadvantages

- Slow, non-deterministic, noisy, no distance preservation
- **UMAP** often preferred (faster, deterministic)