生物信息学

转录组

Yu Hou (侯 宇) Zhejiang University 2025

Sequencing techniques

Genomic sequencing: （DNA）

WGS, WES, Single-cell Genomic Sequencing

Transcriptomic sequencing: （RNA）

Total-RNA-seq, mRNA-seq, small RNA-seq, spatial RNA-seq

Epigenomic sequencing: （Modifications）

DNA-methylation (WGBS, RRBS, MeDIP-seq)

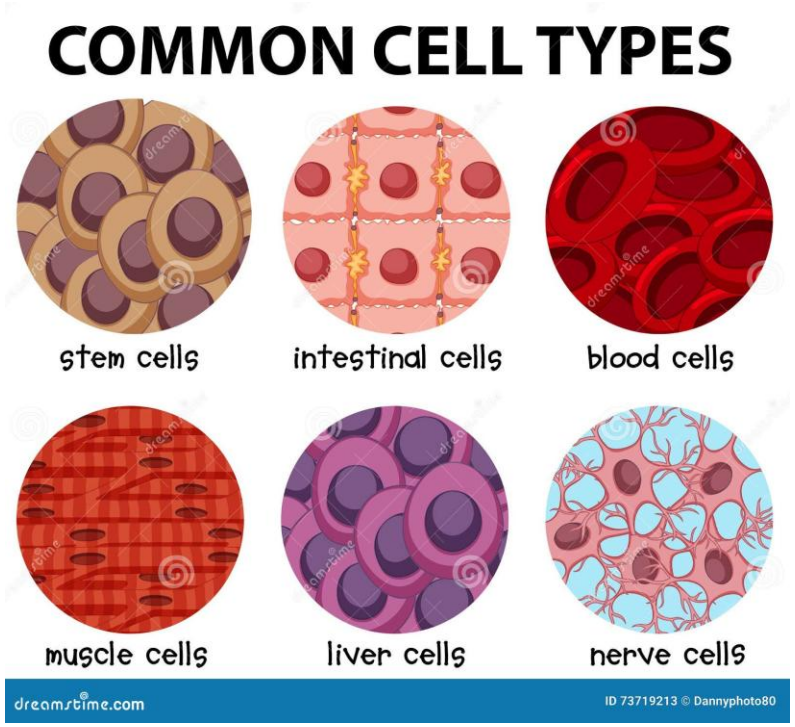
Histone-modification (ChIP-seq, CUT&RUN, CUT&Tag)

Chromatin accessibility (ATAC-Seq, DNase-seq, NOMe-seq)

RNA-seq

Within a single person, all cells share the same genomic sequence. Why do some cells become muscle cells, some cells become liver cells and verve cells.
--- **Gene expression**

Why should we do RNA-seq?

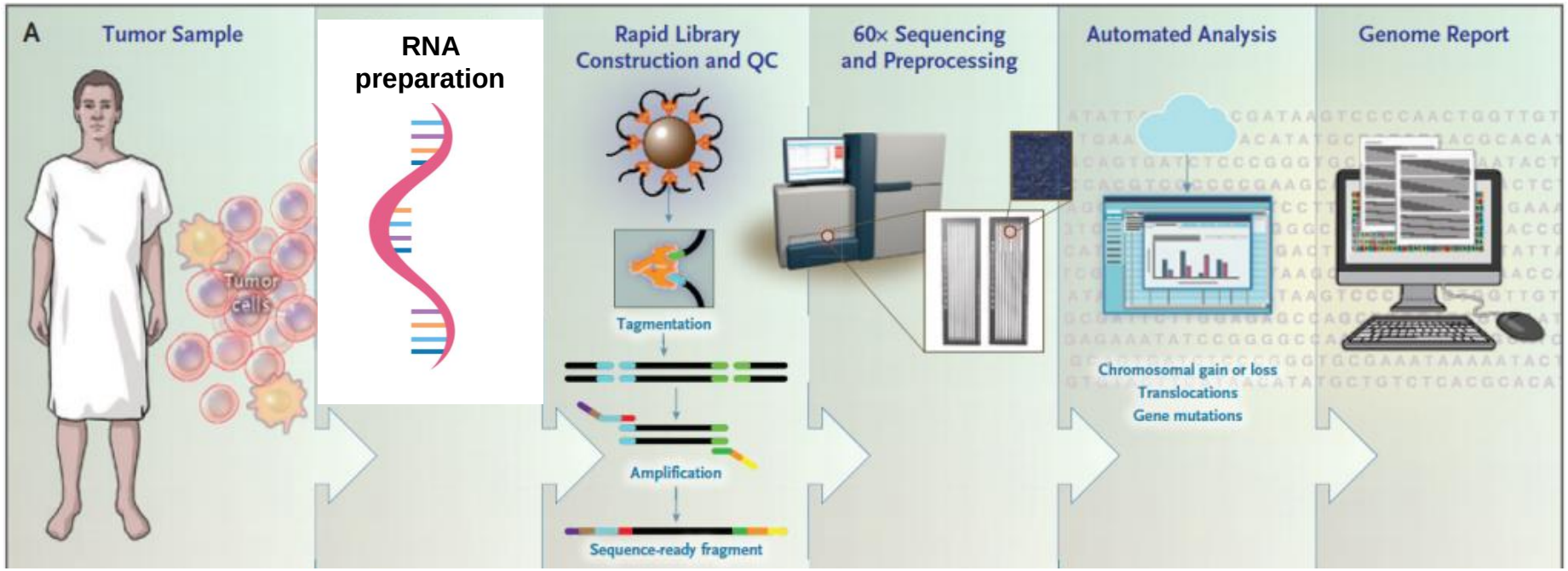


RNA sequencing (RNA-seq) is a powerful technology that has revolutionized the study of transcriptomes. It provides a comprehensive view of gene expression and RNA biology, enabling researchers to address a wide range of biological questions.

How to do RNAseq ?

RNA-seq

How to do Transcriptome (RNA) sequencing?



Step 1 Step 2 Step 3 Step 4 Step 5 Step 6

01 Sample preparation

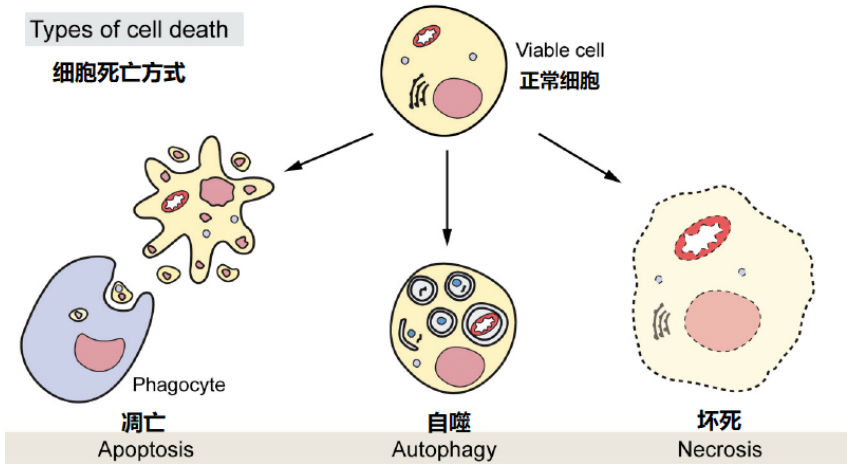
Step1: Sample preparation (Purity, Quality)

Step2: RNA extraction

Step3: Sequencing library preparation

Step4: Library sequencing

Step5: Sequencing reads analyses



RNA is unstable, avoid degradation !!

Sample Type	Ensure the correct type (e.g., blood, saliva, tissue, buccal swab) based on study needs.
Avoid Contamination	Use sterile collection tools to prevent contamination.
	Use gloves and clean workspaces to avoid contamination.
Labeling	Clearly label each sample with a unique identifier to prevent mix-ups.
Storage Conditions	Follow proper storage protocols (e.g., -80°C or liquate nitrogen for RNA samples, room temperature for stabilized saliva kits).
Avoid Degradation	Minimize exposure to heat, light, and nucleases to prevent RNA degradation.
Transport	Maintain appropriate temperature and conditions during transport to preserve RNA integrity.

02 RNA extraction

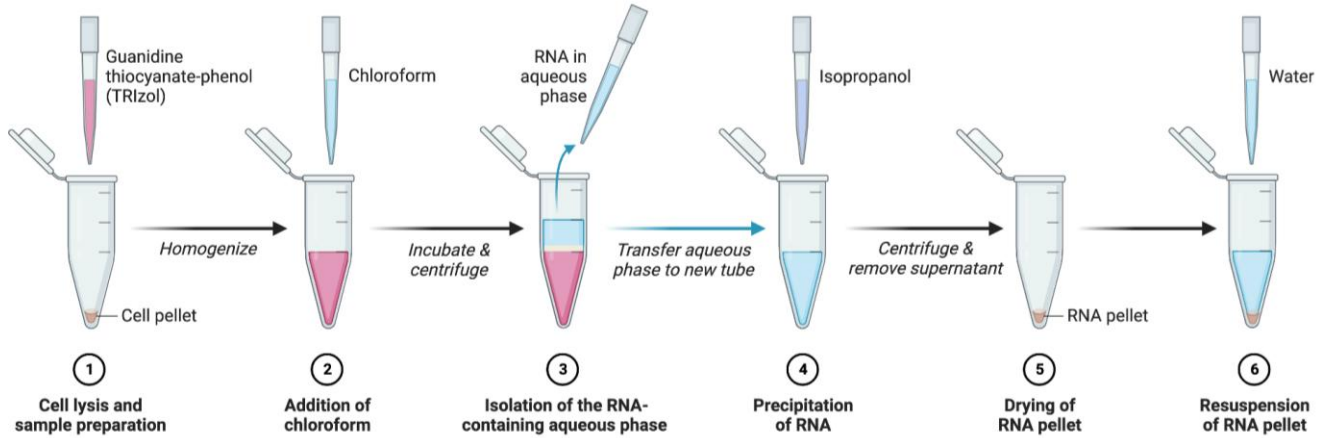
Step1: Sample preparation (Purity, Quality)

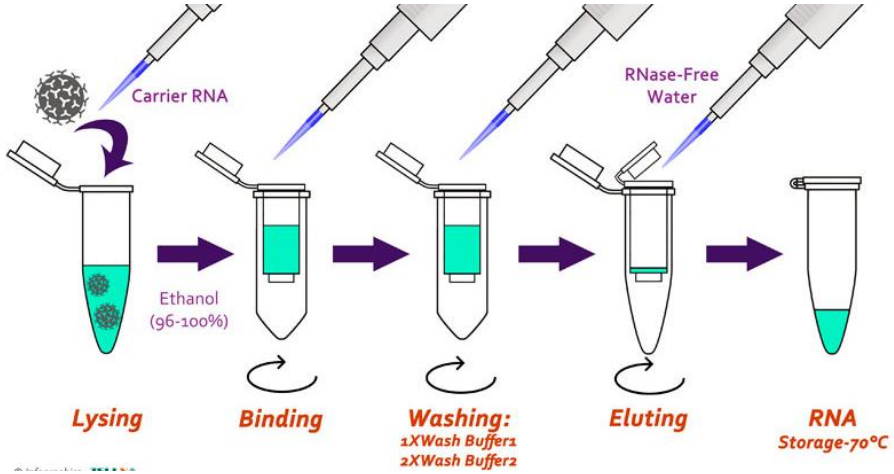
Step2: RNA extraction (Integrity, concentration, No degradation, yield)

Step3: Sequencing library preparation

Step4: Library sequencing

Step5: Sequencing reads analyses





Use **DNase treatment** to remove contaminating genomic DNA

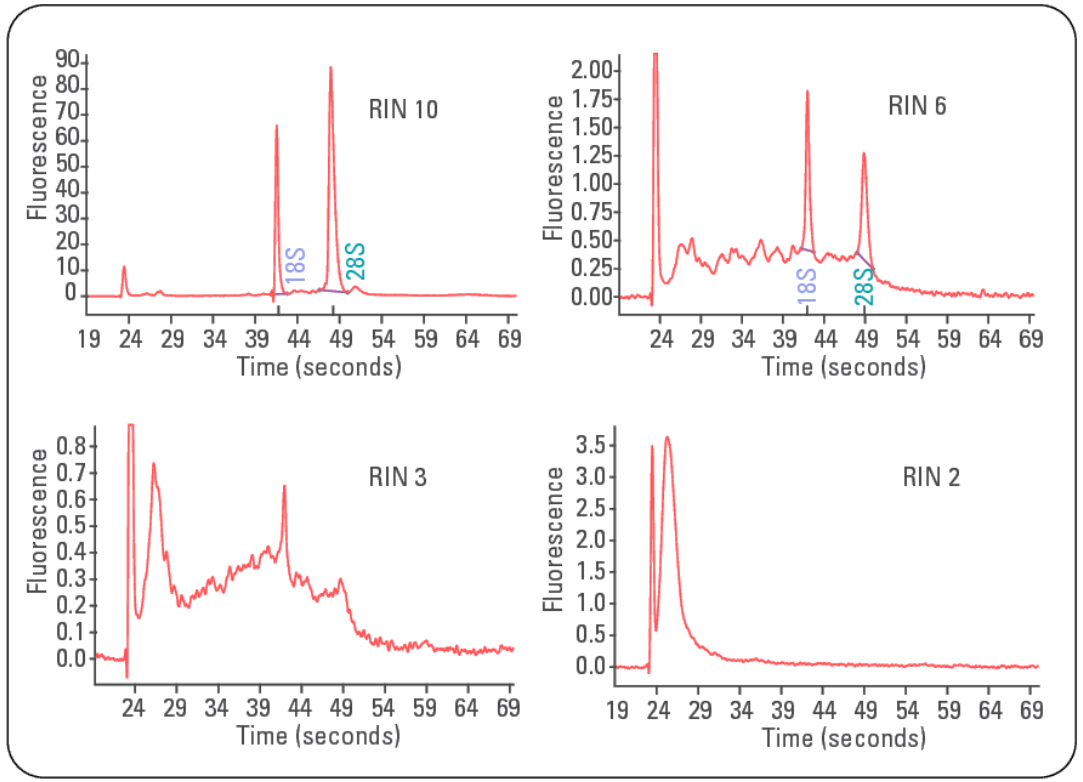
02 RNA extraction

2.1 RNA quality

RNA Integrity Number (RIN)
is a critical step in assessing the quality of RNA samples.

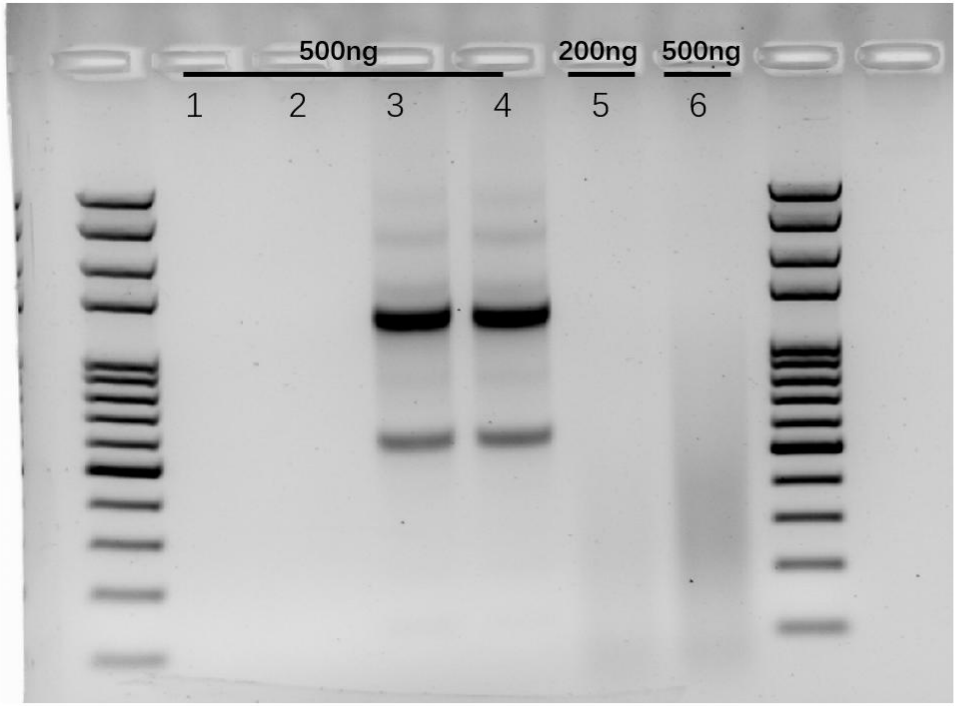
•Intact RNA samples show distinct peaks corresponding to the **28S** and **18S ribosomal RNA (rRNA)** subunits.

Interpretation of RIN Values
•**RIN ≥ 9**: Excellent RNA integrity, suitable for all downstream applications (e.g., RNA-seq, qPCR).
•**7 ≤ RIN < 9**: Good RNA integrity, suitable for most downstream applications.
•**5 ≤ RIN < 7**: Partially degraded RNA, may affect sensitive applications (e.g., RNA-seq).
•**RIN < 5**: Severely degraded RNA, not suitable for most downstream applications.

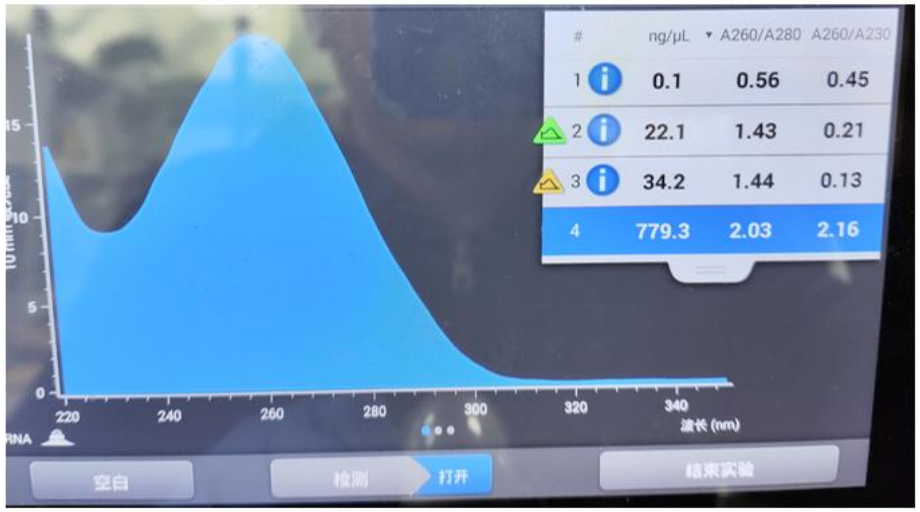


02 RNA extraction

2.1 RNA quality



1. 150W hela 2%PFA 14.5 h
2. 150W hela 2%PFA 1 h
3. 150W hela fresh
4. Hela fresh RNA frozen
5. FFPE-1
6. FFPE-2



1. 150W hela 2%PFA 14.5 h
2. 150W hela 2%PFA 1 h
3. 150W hela fresh

03 Library preparation

Step1: Sample preparation (Purity, Quality)

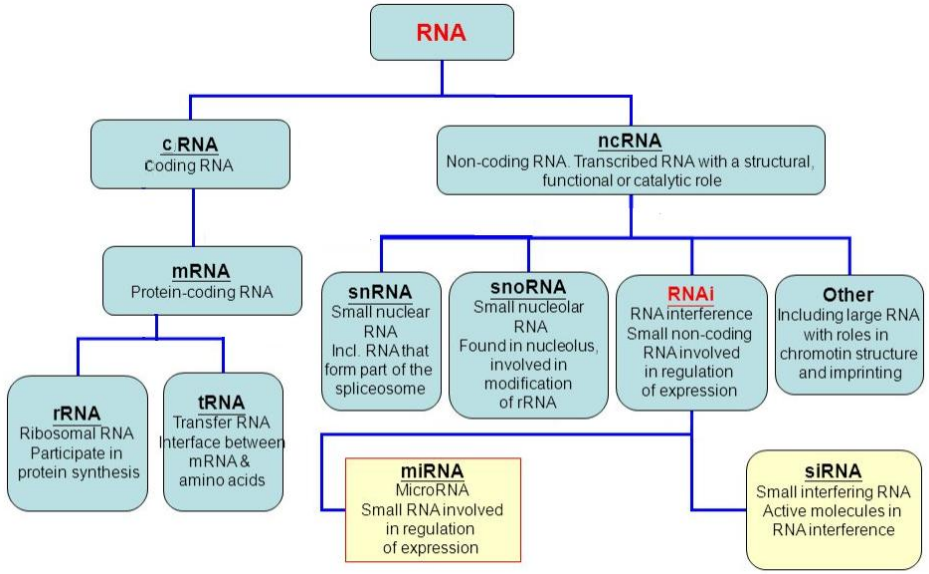
Step2: RNA extraction (Integrity, Concentration, No degradation, Yield)

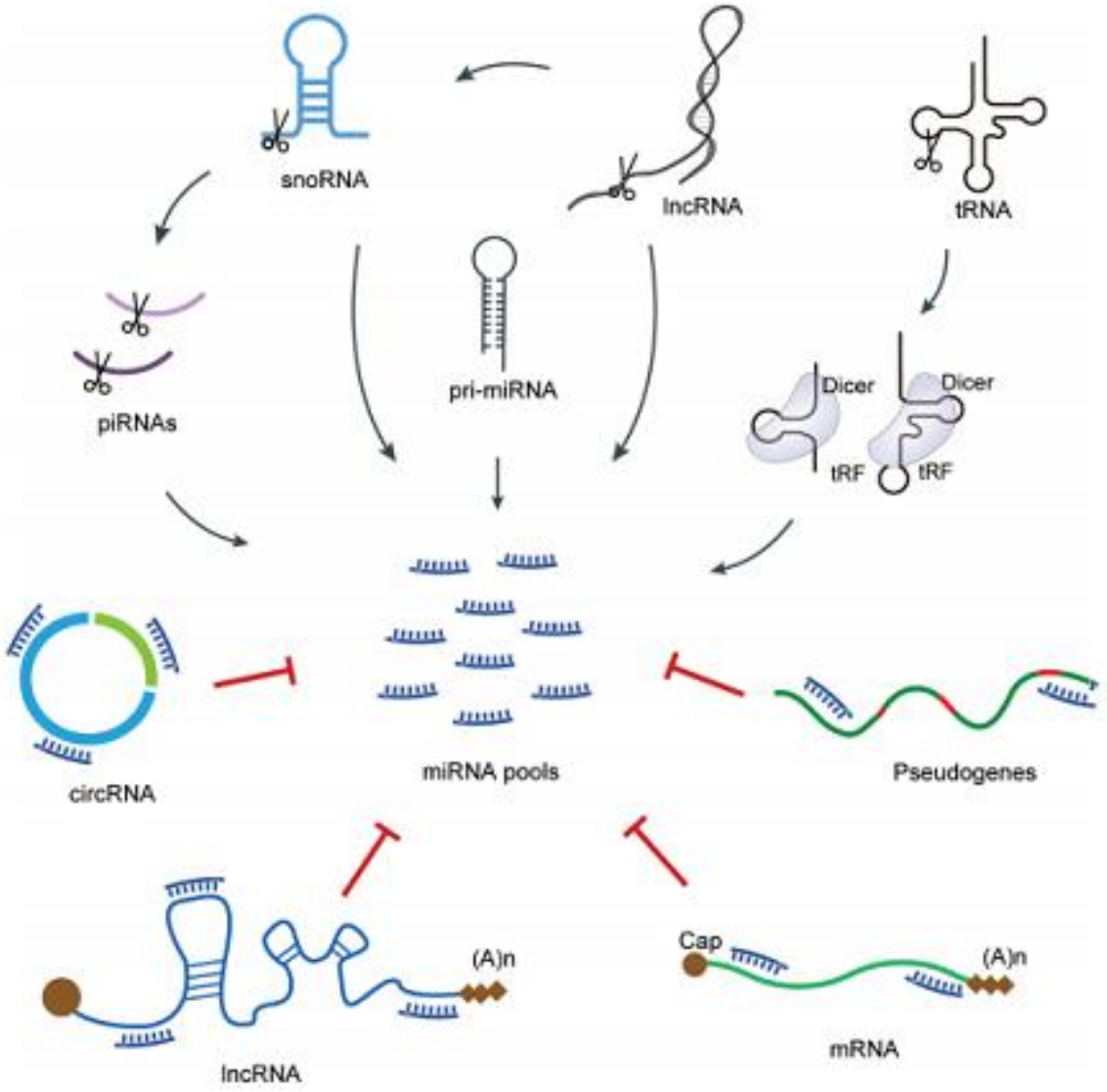
Step3: Sequencing library preparation (total RNA-seq, mRNA-seq, small RNA-seq, Ribo-seq, circRNA-seq, single-cell RNA-seq, Spatial RNA-seq)

Step4: Library sequencing

Step5: Sequencing reads analyses

03 Library preparation

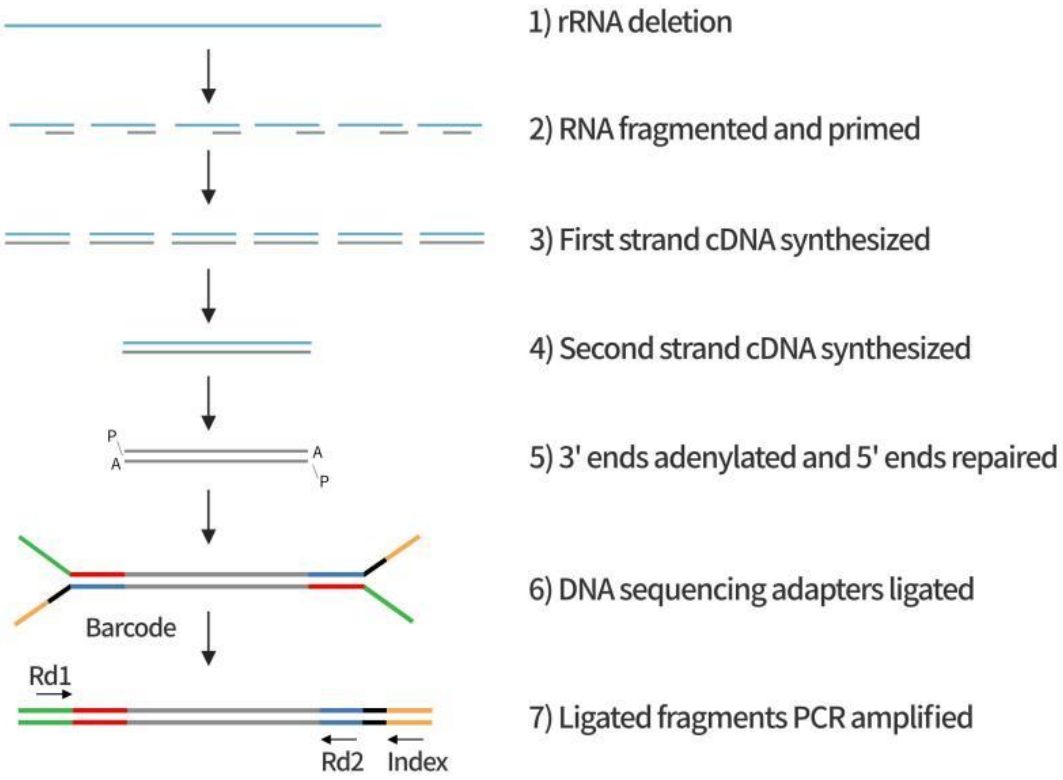




03 Library preparation

3.1 Total RNA

rRNA depletion



Advantages

✔ Captures both poly(A) and non-poly(A) RNAs (e.g., lncRNAs, circRNAs, viral RNAs).

✔ Works well with low-quality or degraded RNA samples (e.g., FFPE, plasma RNA).

Limitations

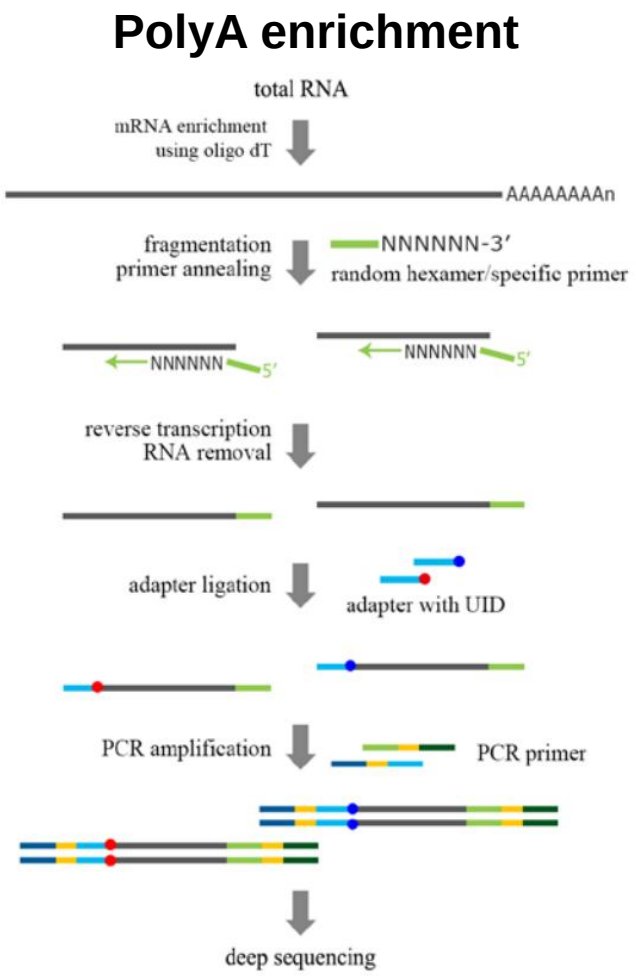
✘ More complex and costly than poly(A) selection.

✘ May retain some rRNA contamination.

Common Kits: NEBNext rRNA Depletion Kit, Illumina TruSeq Stranded Total RNA Kit

03 Library preparation

3.2 mRNA



Advantages

✔ Suitable for high-quality RNA samples.

✔ Provides a cleaner transcriptome profile (reduces rRNA and non-coding RNA contamination).

✔ Best for differential gene expression analysis.

Limitations

✘ Cannot capture non-poly(A) RNAs (e.g: long non-coding RNAs).

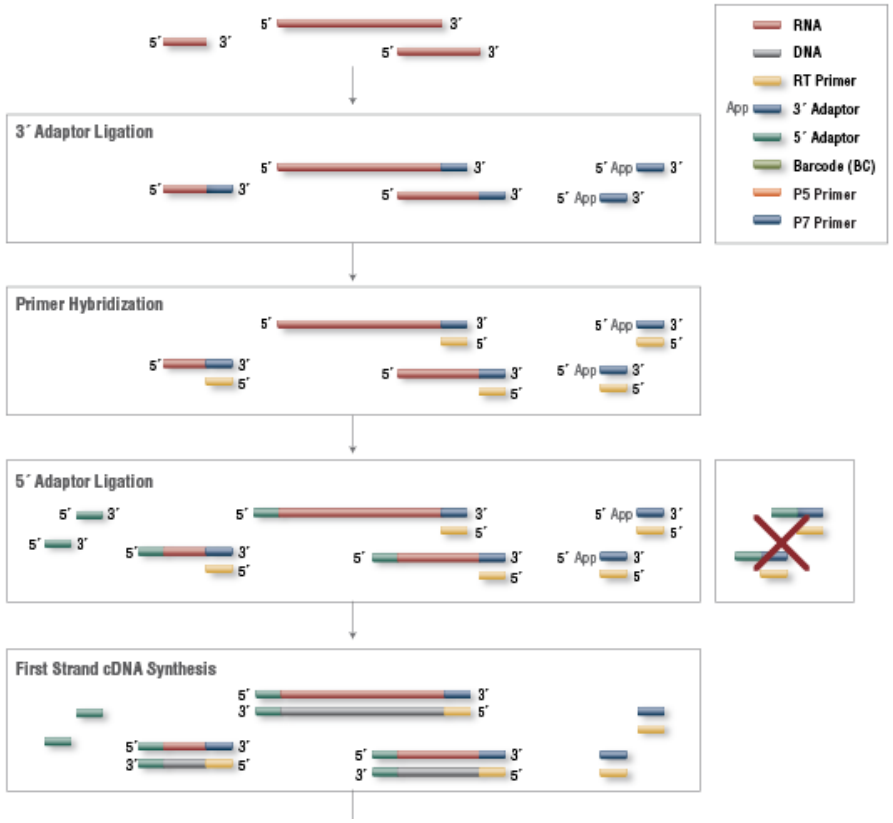
✘ Not ideal for degraded RNA (e.g., FFPE samples).

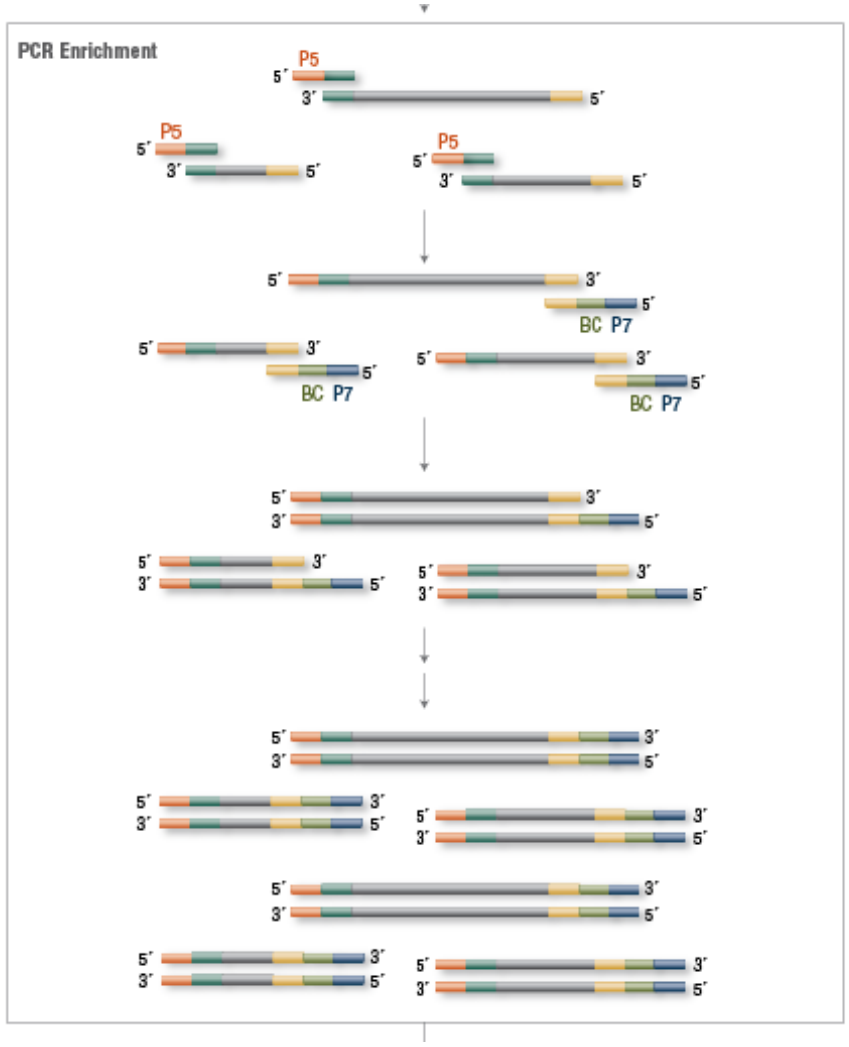
Common Kits: NEBNext Poly(A) mRNA Magnetic Isolation, Illumina TruSeq Stranded mRNA Library Prep Kit

03 Library preparation

3.3 Small RNA

microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs)



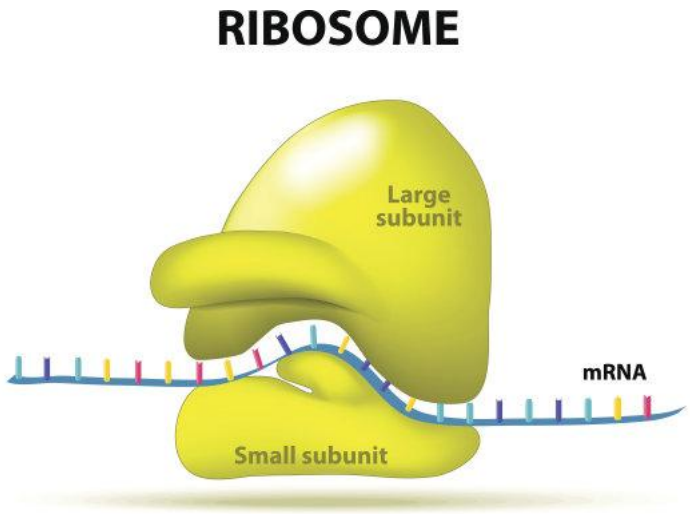


Size Selection （140-160 bp）

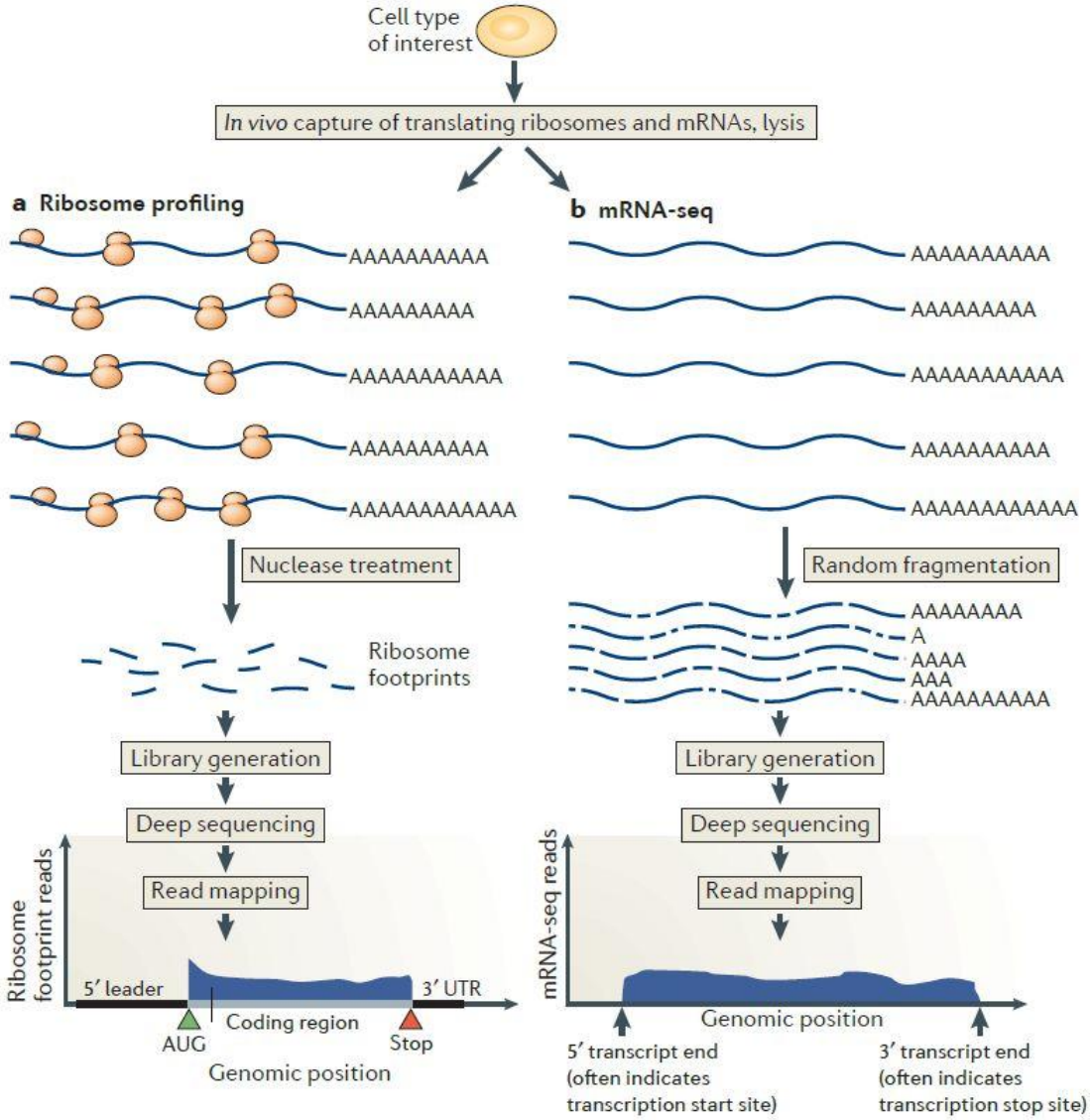
(Gel or Bead-Based Purification)

03 Library preparation

3.4 Ribo-seq



Ribosome profiling (Ribo-Seq) is a specialized RNA sequencing technique that captures actively translated mRNA fragments protected by ribosomes. This method provides insights into translation dynamics, ribosome occupancy, and protein synthesis regulation.



03 Library preparation

3.5 circ-seq: Circular RNA Sequencing

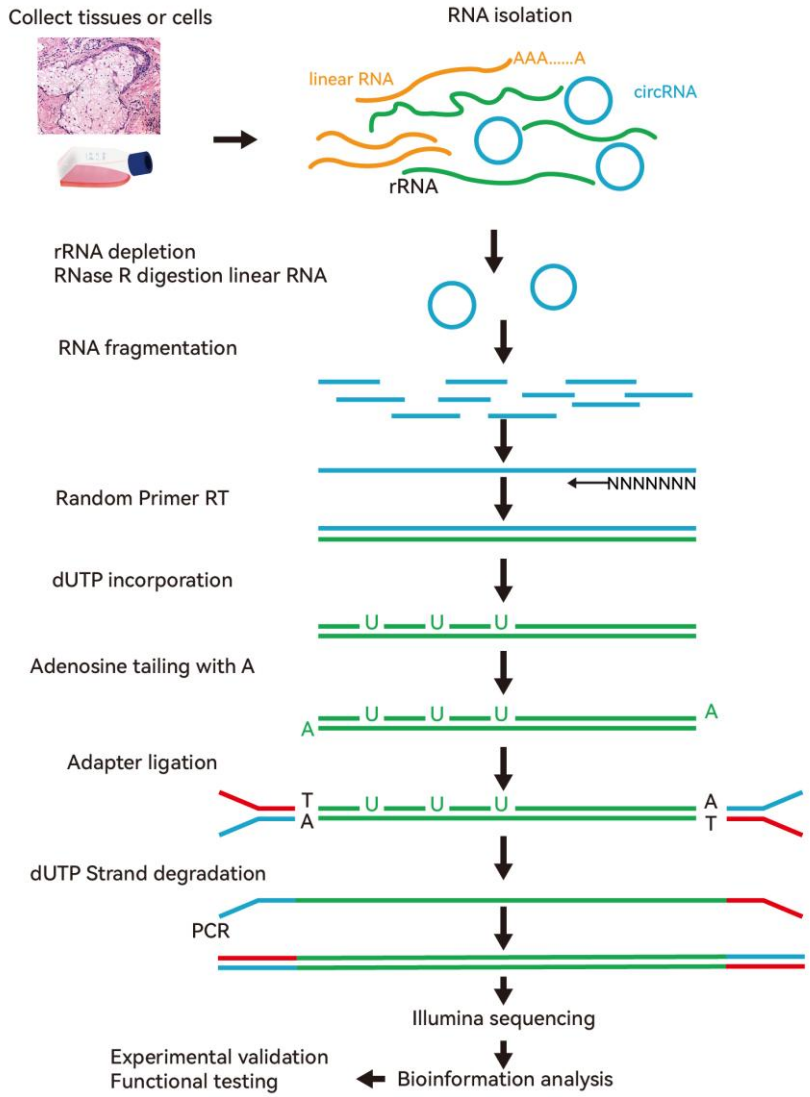
Applications of circRNA-Seq

☑ **CircRNA Identification** – Discover and characterize novel circular RNAs.

☑ **Differential Expression Analysis** – Compare circRNA expression across conditions.

☑ **CircRNA-miRNA Interaction Studies** – Identify circRNAs acting as miRNA sponges.

☑ **Cancer and Disease Biomarker Discovery** – Explore circRNA roles in diseases.



04 Library sequencing

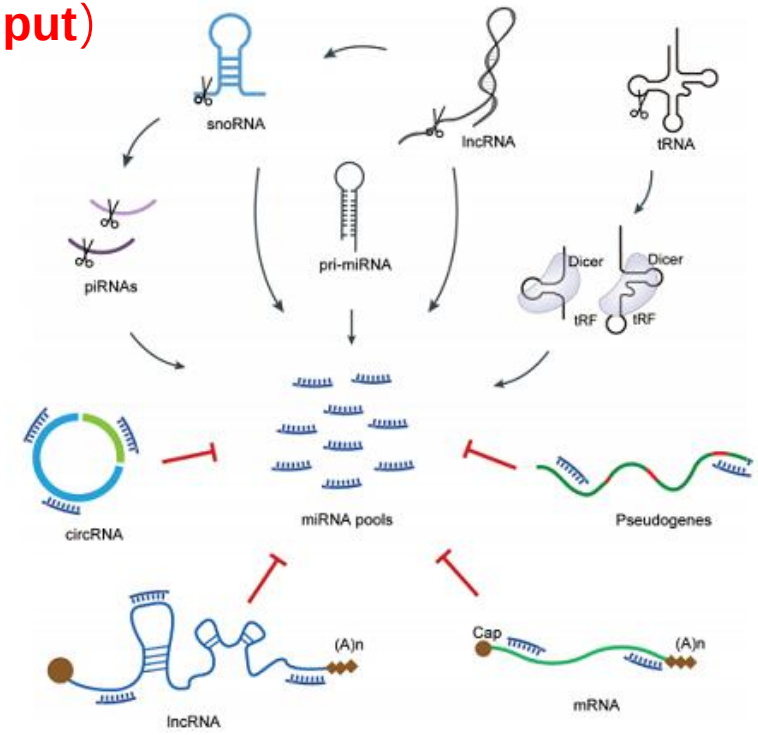
Step1: Sample preparation (Purity, Quality)

Step2: RNA extraction (Integrity, Concentration, No degradation, Yield)

Step3: Sequencing library preparation (total RNA-seq, mRNA-seq, small RNA-seq, Ribo-seq, circRNA-seq, single-cell RNA-seq)

Step4: Library sequencing (SE, PE, Read length, Data throughput)

Step5: Sequencing reads analyses



04 Library sequencing

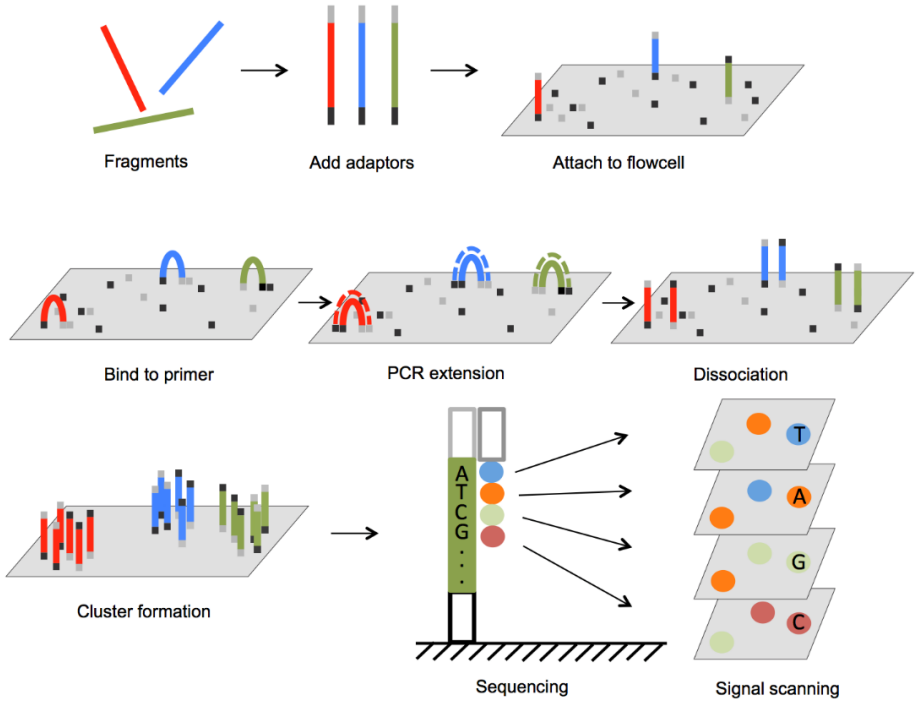
4. Library sequencing

Selection Guide:

•**SE**: Suitable for simple applications like small RNA sequencing or ChIP-seq.

•**PE**: Preferred for genome assembly, RNA-seq, whole-exome sequencing (WES), and structural variant detection.

•PE150 is the most widely used mode for WGS, RNA-seq, and WES.
•PE300 or PE600 is best for microbiome studies (16S rRNA, metagenomics).



05 Data Analyses

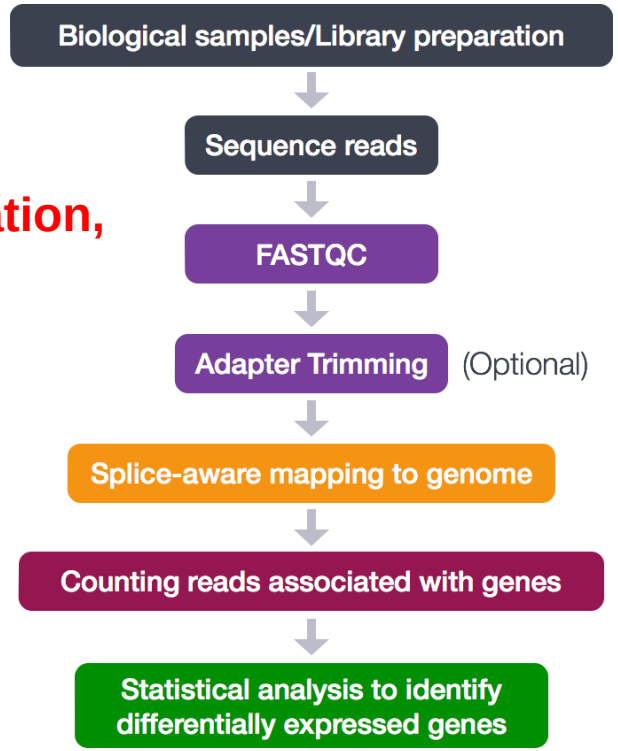
Step1: Sample preparation (Purity, Quality)

Step2: RNA extraction (Integrity, Concentration, No degradation, Yield)

Step3: Sequencing library preparation (total RNA-seq, mRNA-seq, small RNA-seq, Ribo-seq, circRNA-seq, single-cell RNA-seq)

Step4: Library sequencing (SE, PE, Read length, Data throughput)

Step5: Sequencing reads analyses (QC, Alignment, Transcript Quantification, Comparison)



05 Data Analyses

5.1. Quality Control (QC)

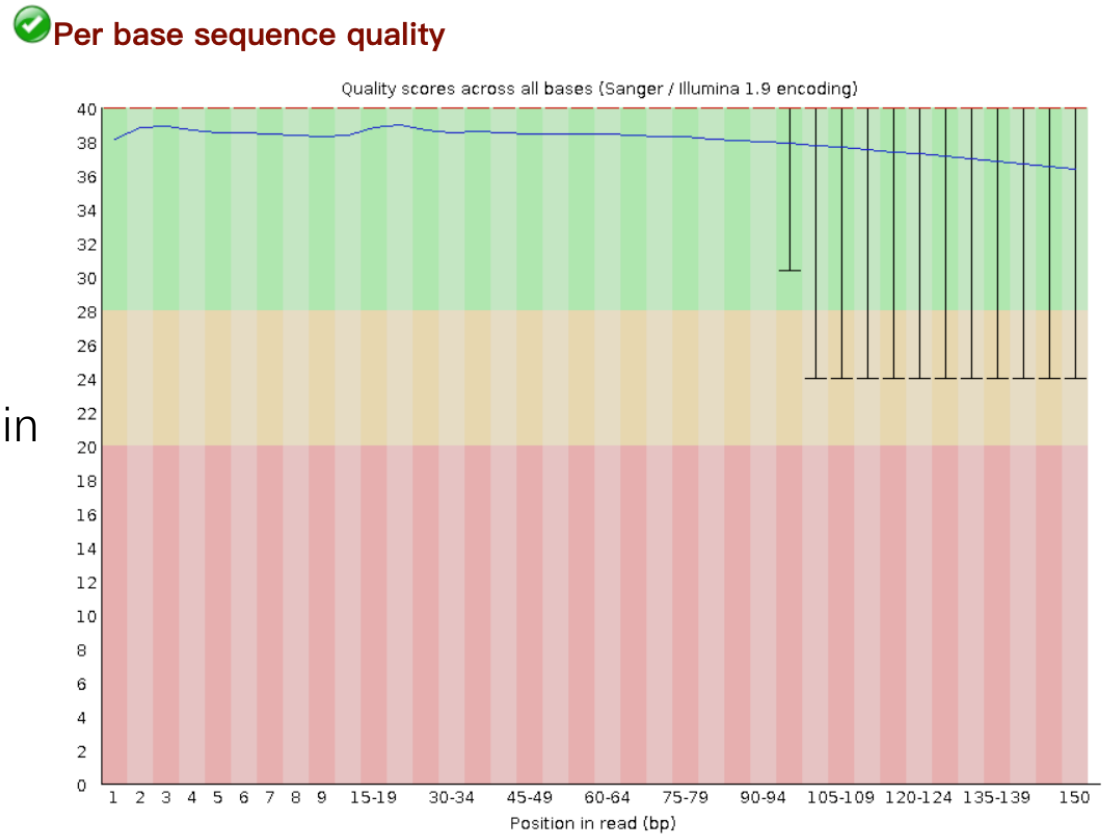
To remove low-quality reads and adapters

Tools:

- **Trim Galore / Cutadapt**: Removes adapters and low-quality bases.
- **fastp**: Performs trimming, quality filtering, and basic QC in one step.

Filtering criteria:

- Remove low-quality reads (Phred score < 20 or 30)
- Trim adapter sequences
- Discard very short reads (e.g., <30 bp)



Data Analyses

5.2. Alignment

Reads are aligned to a reference genome or transcriptome to determine their origin.

Tools:

•**STAR**: Fast and accurate splice-aware aligner.

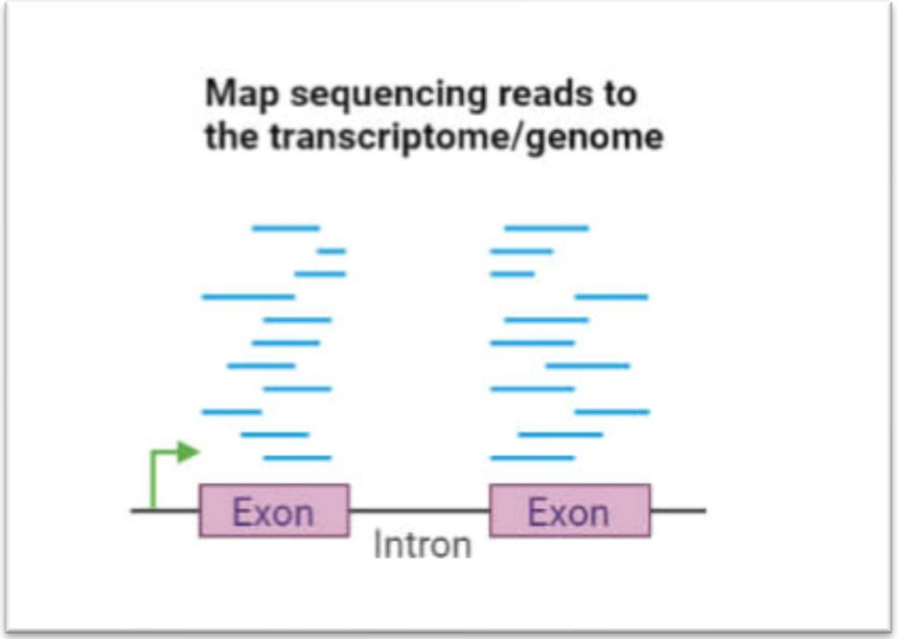
•**HISAT2**: Efficient for large-scale datasets with a small memory footprint.

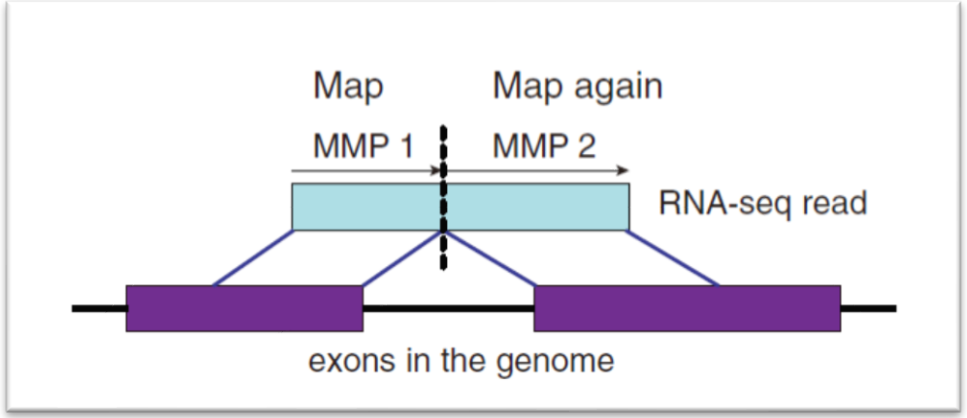
•**Salmon / Kallisto**: Pseudo-alignment for rapid transcript quantification.

Key Metrics:

•**Mapping rate**: Percentage of reads that align to the genome.

•**Multi-mapped reads**: Reads aligning to multiple locations (e.g., repetitive sequences).





5.3. Transcripts quantification and Normalization

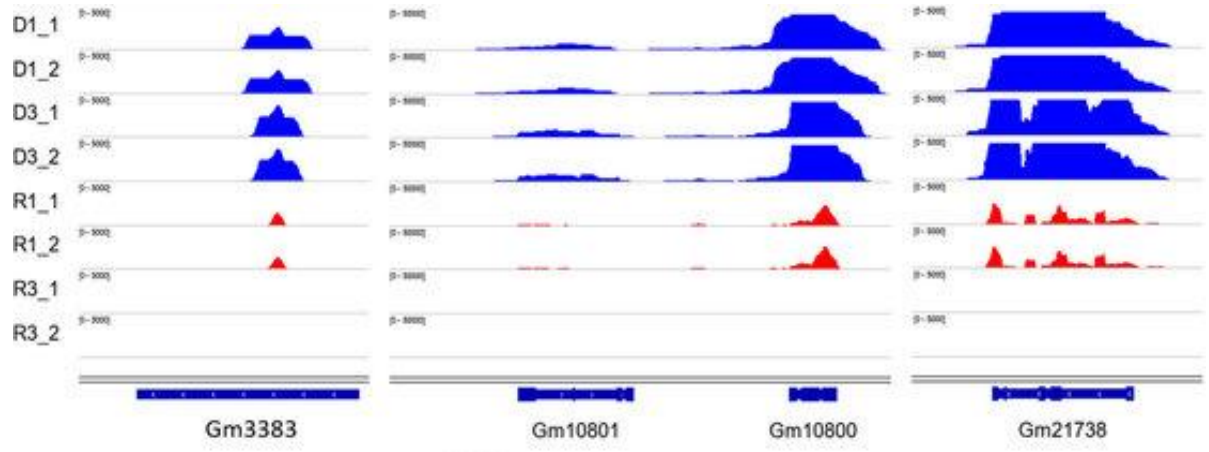
To compare gene expression across samples, normalization is required to correct for library size and sequencing depth.

Tools:

•**FeatureCounts (Subread package)**: Counts reads mapped to genes.

•**HTSeq-count**: Counts reads in exon regions for gene-level quantification.

•**Salmon/Kallisto**: Directly estimates transcript-level expression from raw reads (without alignment).



Output Format:

•**Raw counts**: Used for differential expression analysis.

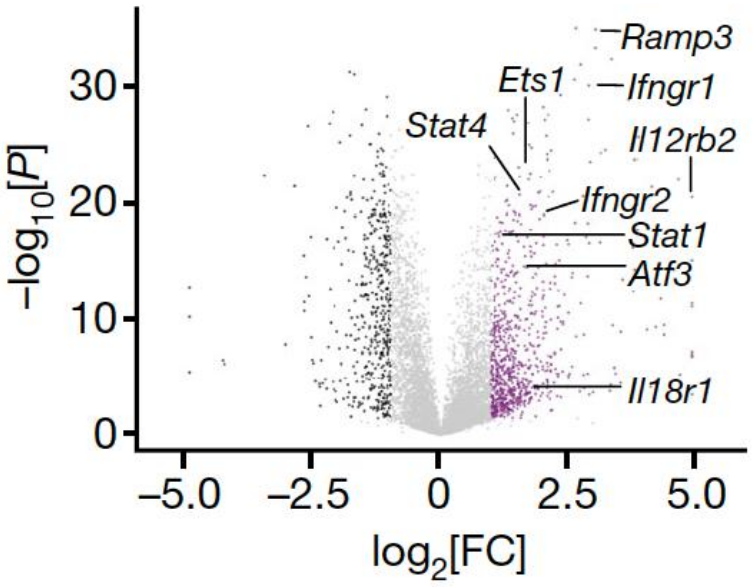
•**TPM (Transcripts Per Million reads)**: Normalized expression level.

•**FPKM/RPKM**: Normalization for gene length and library size (less preferred than TPM).

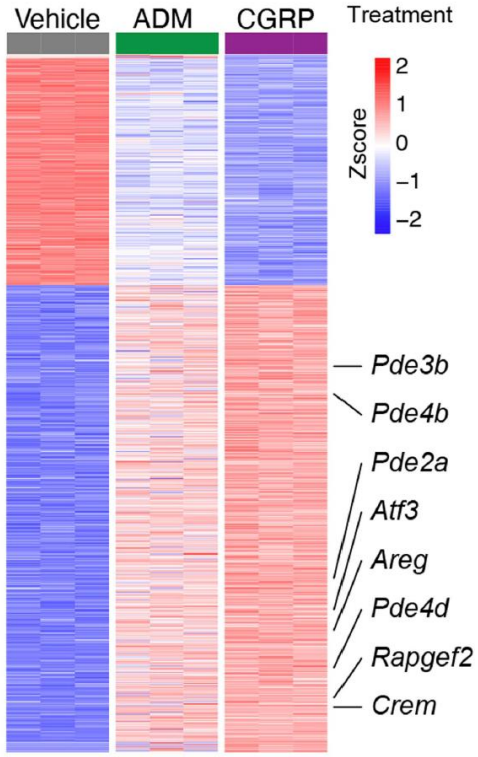
5.4. Comparison (Bulk, Fold Change & P-value)

Find the Differentially Expressed Genes (DEGs)

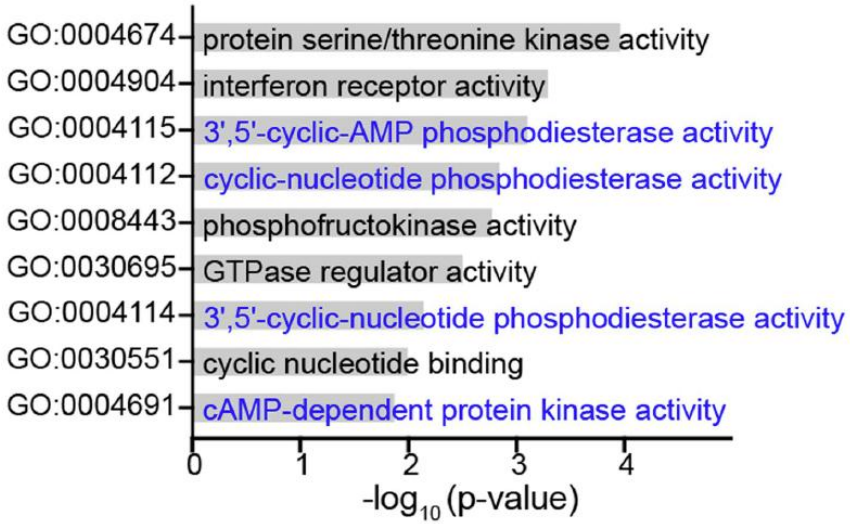
01. Volcano Plot



02. Heatmap of DEGs



03. Gene Ontology Analyses



Yu Hou et al., *Nature,* 2024

Take Home Message

RNA-seq

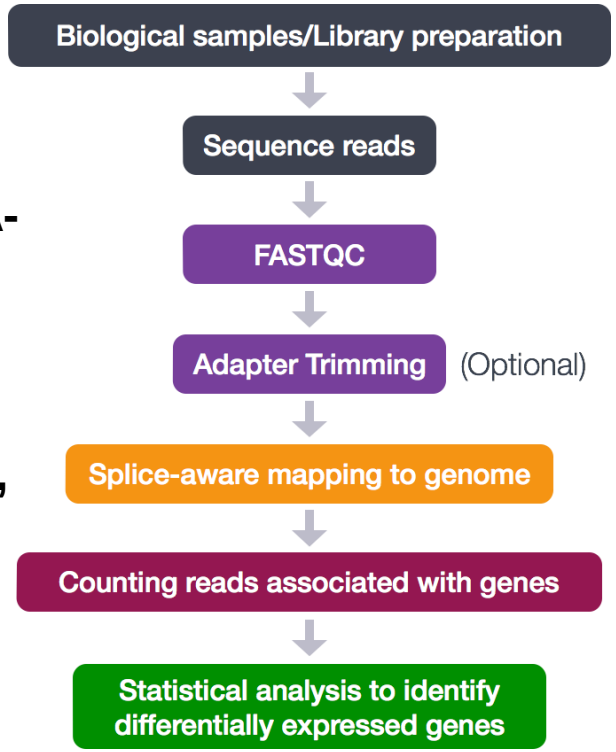
Step1: Sample preparation (Purity, Quality)

Step2: RNA extraction (Integrity, Concentration, No degradation, Yield)

Step3: Sequencing library preparation (total RNA-seq, mRNA-seq, small RNA-seq, Ribo-seq, circRNA-seq, single-cell RNA-seq)

Step4: Library sequencing (SE, PE, Read length, Data throughput)

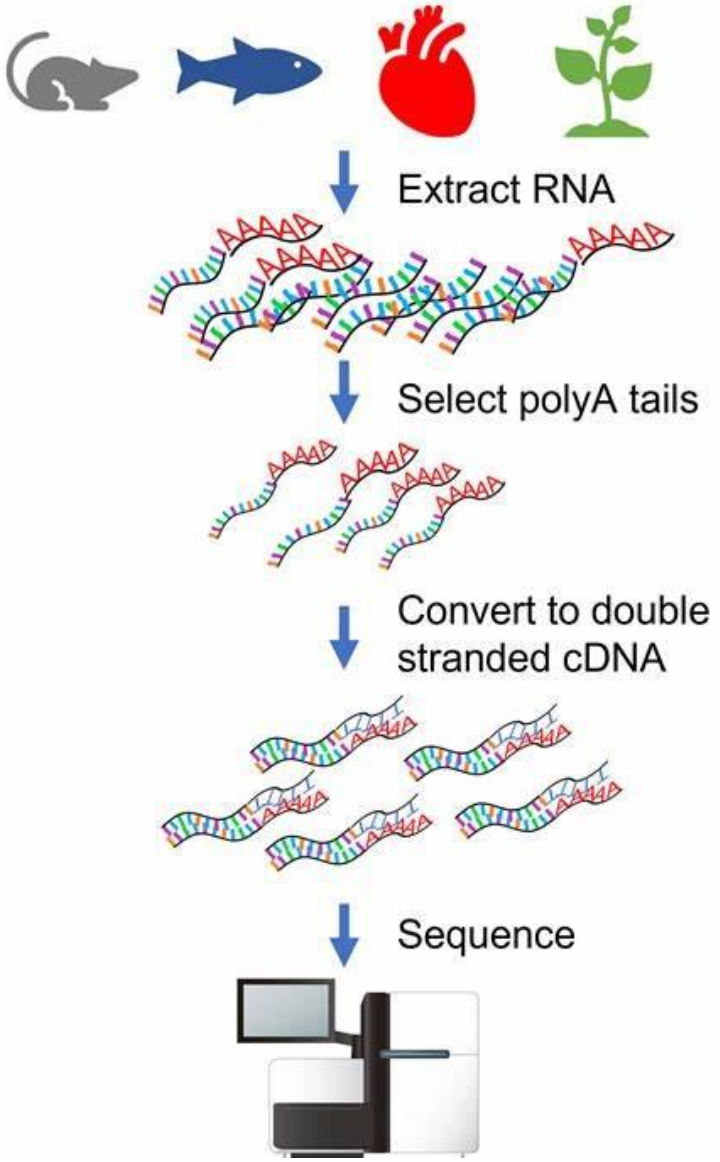
Step5: Sequencing reads analyses (QC, Alignment, Transcript Quantification, Comparison)



What can we do with RNAseq ?

RNA-seq

1. Quantify gene expression
2. Discover differential expression
3. Identify novel transcripts and splicing variants
4. Identify genetic mutations
5. Allele specific gene expression
6. Support biomarkers and drug target

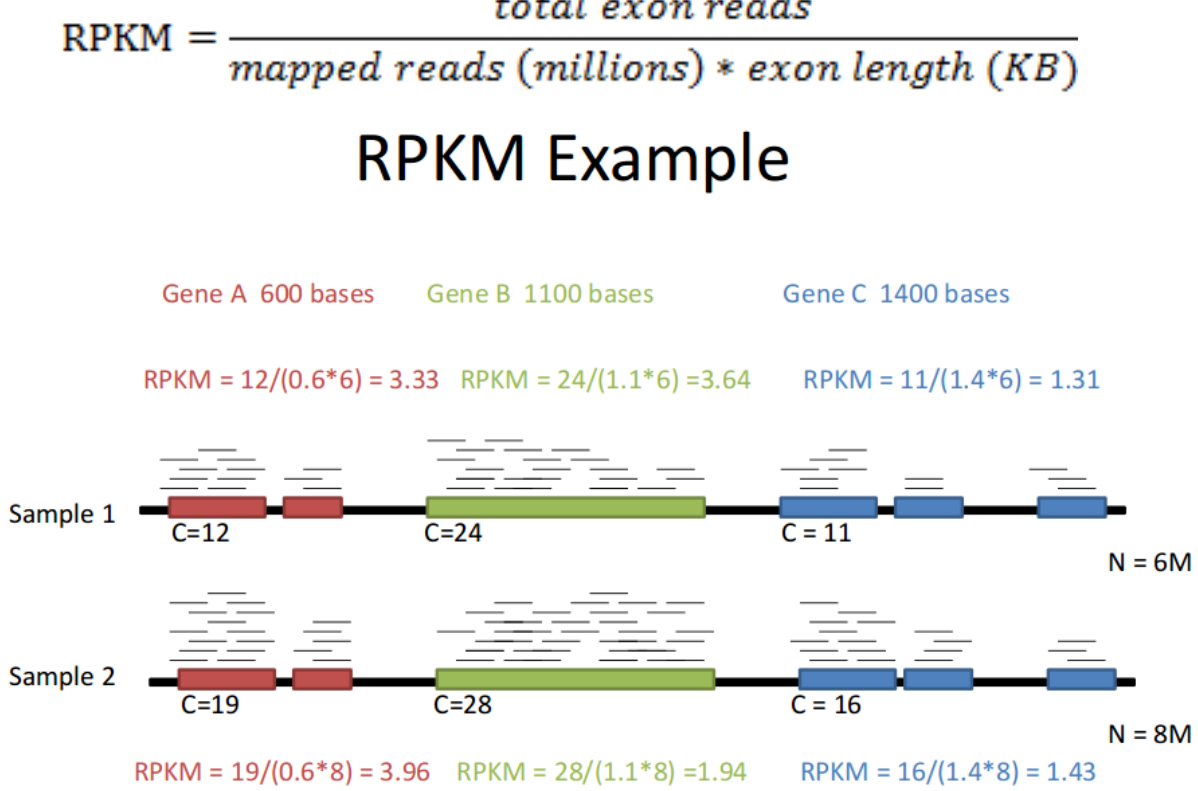


01 Quantify Gene Expression

1. Quantify Gene Expression

RNA-seq measures the expression levels of all transcripts (mRNAs, lncRNAs, etc.) in a sample, allowing researchers to identify which genes are active and how much they are being expressed.

$$\text{RPKM} = \frac{total\ exon\ reads}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$



Output Format:

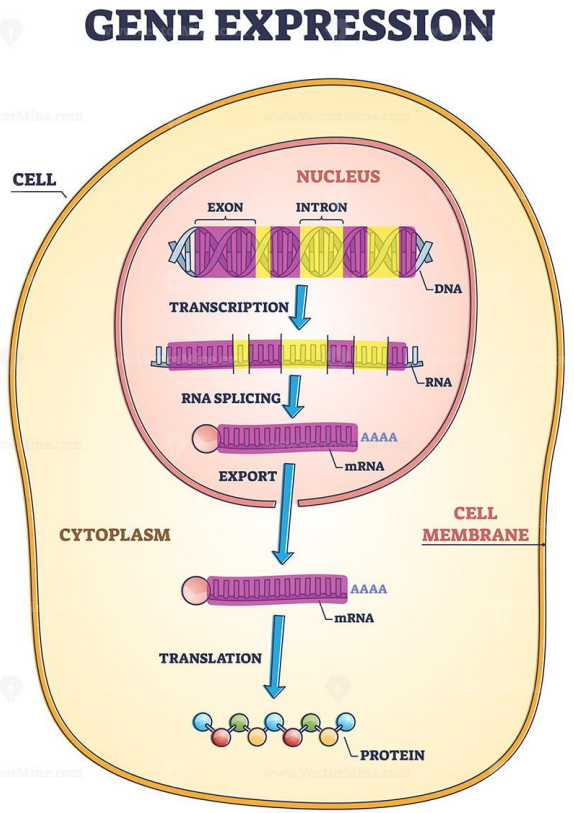
- **Raw counts**: Used for differential expression analysis.
- **TPM (Transcripts per million mapped reads)**: Normalization for library size
- **FPKM/RPKM (reads per kilo base of transcript per million mapped reads)**: Normalization for gene length and library size (less preferred than TPM).

01 Quantify Gene Expression

1.1 Gene expression

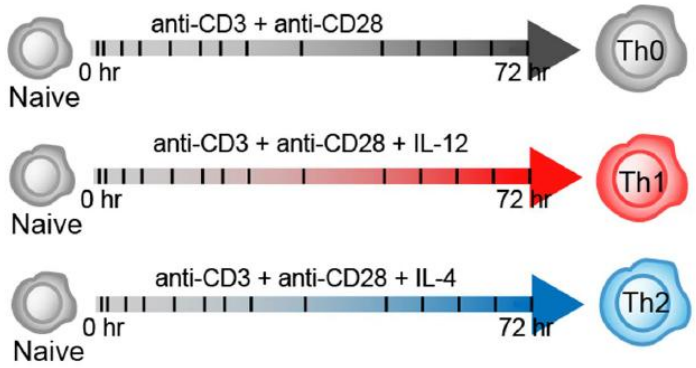
As of the latest annotations (e.g., from GENCODE, Ensembl, and RefSeq), the human genome is known to contain approximately:

- ~**20,000** protein-coding genes
- Tens of thousands of non-coding RNA genes, including:
  - Long non-coding RNAs (lncRNAs)
  - Small nuclear RNAs (snRNAs)
  - MicroRNAs (miRNAs)
  - Ribosomal RNAs (rRNAs)
  - Others

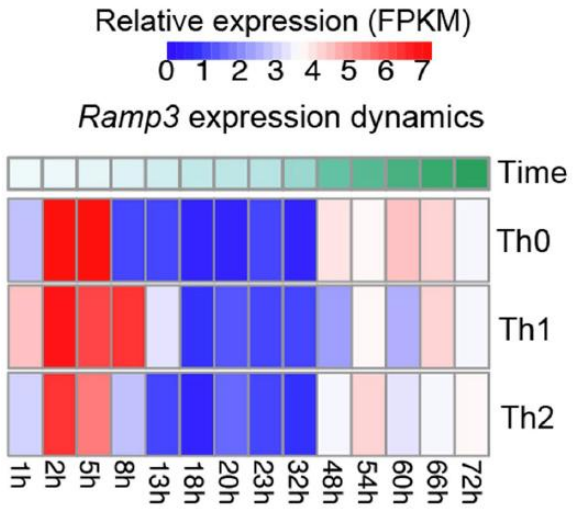


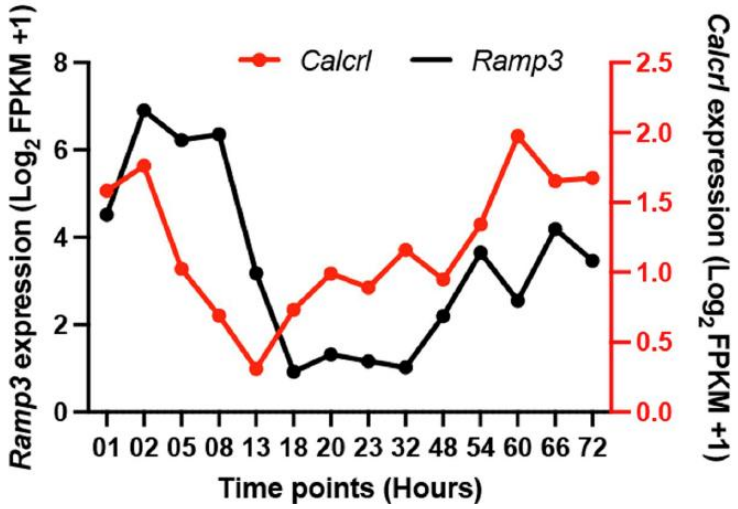
01 Quantify Gene Expression

1.2. Dynamic gene expression



Heatmap

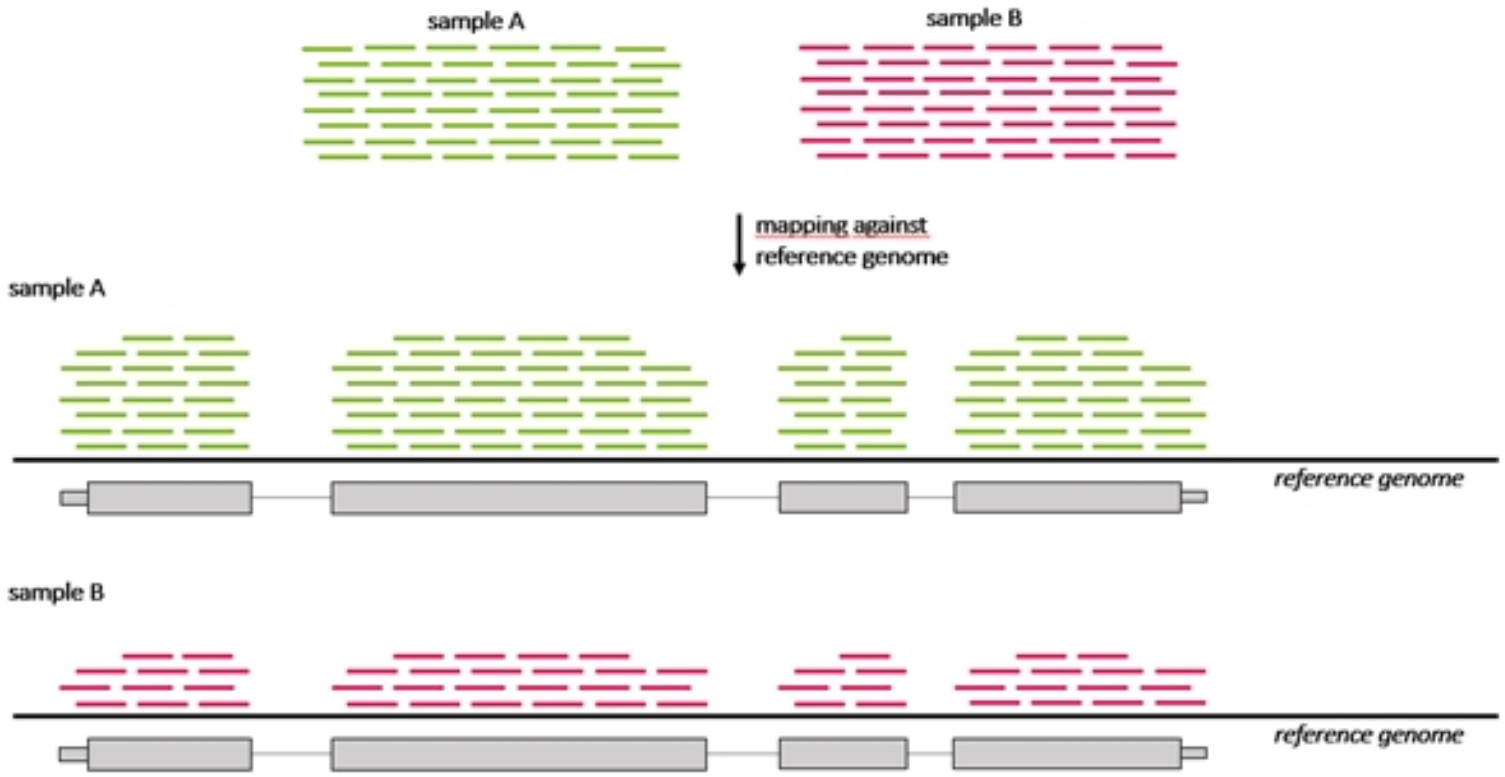




02 Differential Gene Expression

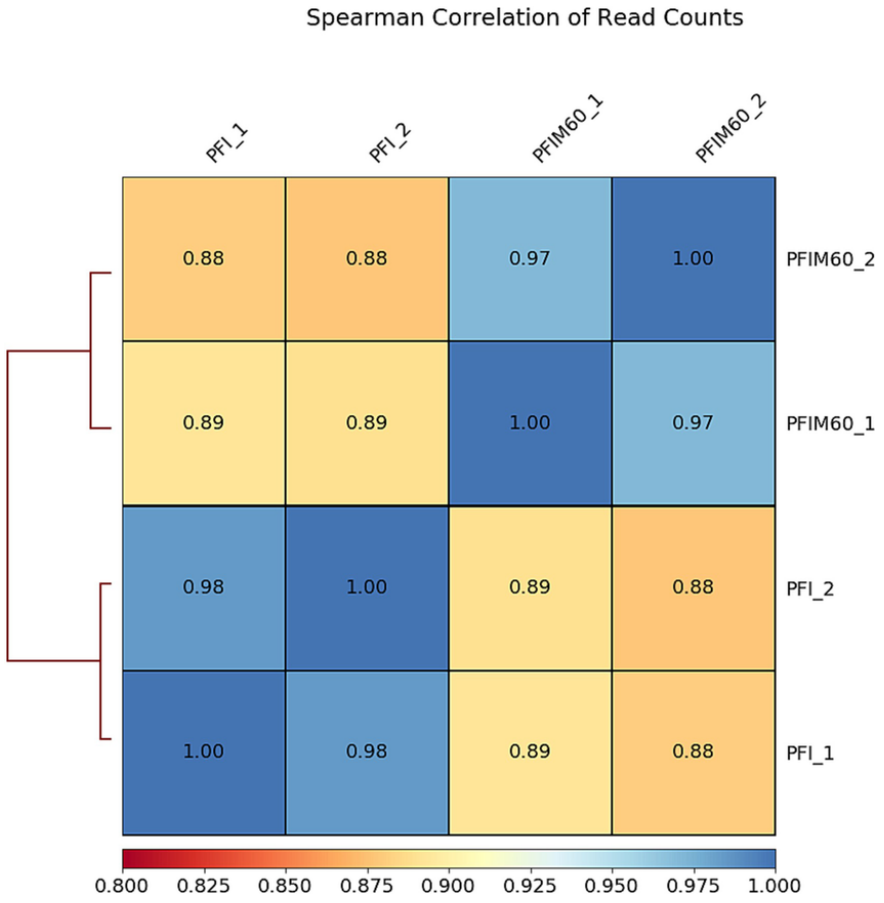
🔍 2. Discover Differential Expression

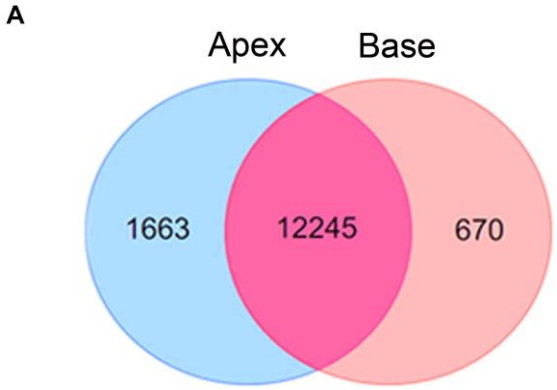
By comparing RNA-seq data between different conditions (e.g., healthy vs. diseased, treated vs. untreated), you can find **differentially expressed genes** that may be involved in specific biological processes or diseases.

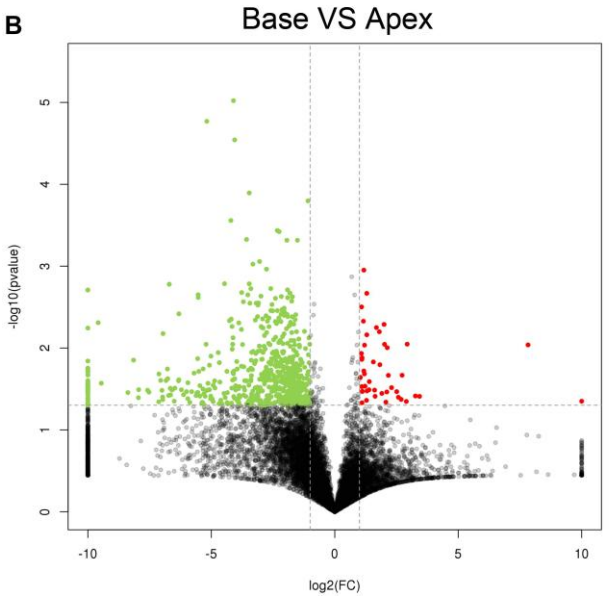


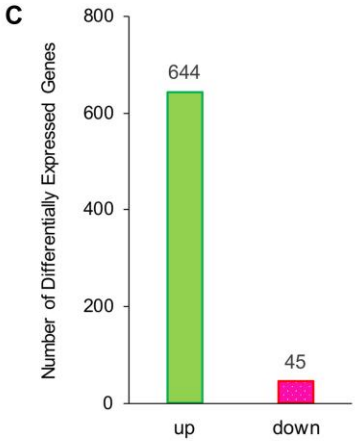
02 Differential Gene Expression

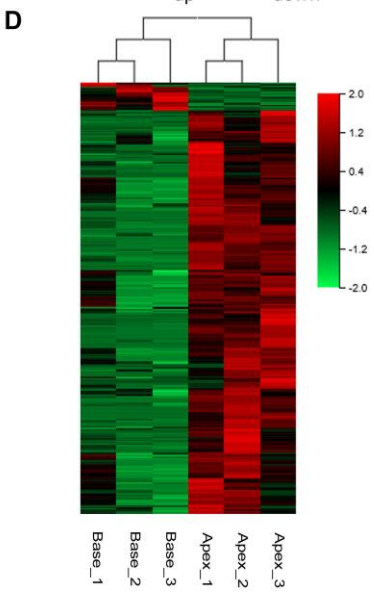
2.1. DEGs between two groups







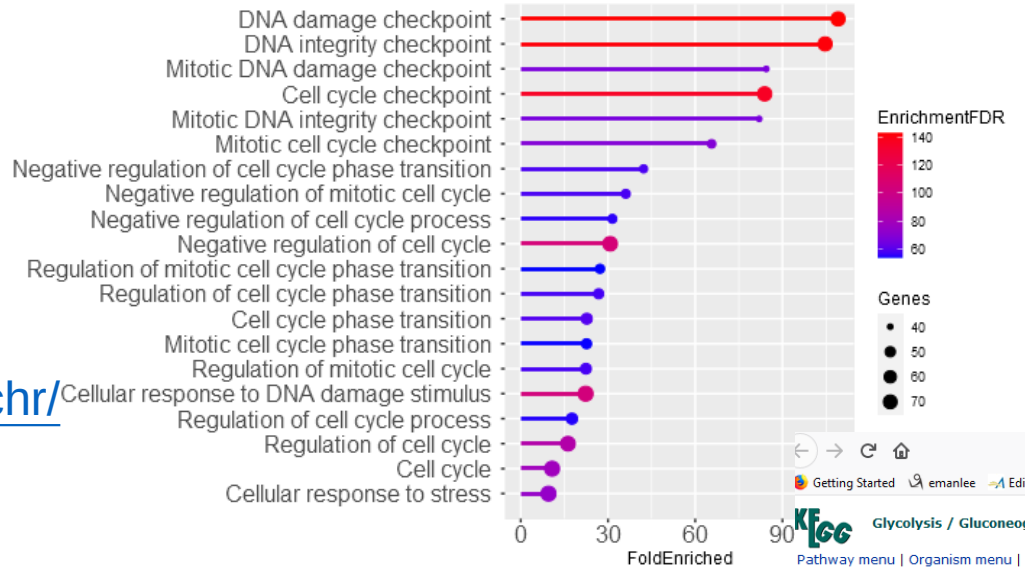


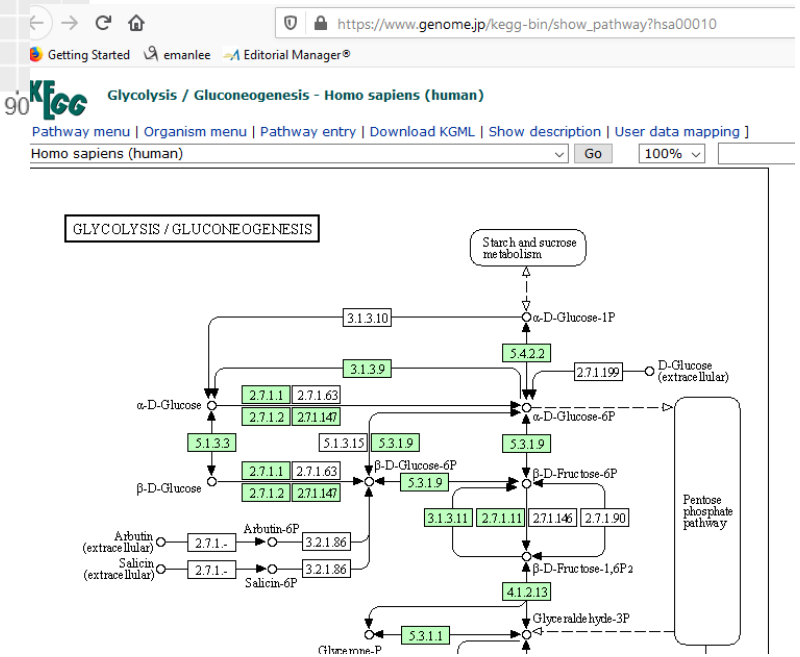


02 Differential Gene Expression

2.1. DEGs between two groups

1. **DAVID**
   https://david.ncifcrf.gov/
2. **Enrichr**
   https://maayanlab.cloud/Enrichr/
3. **Metascape**
   https://metascape.org/
4. **g:Profiler**
   https://biit.cs.ut.ee/gprofiler/
5. **WebGestalt**
   http://www.webgestalt.org/
6. **STRING**
   https://string-db.org/

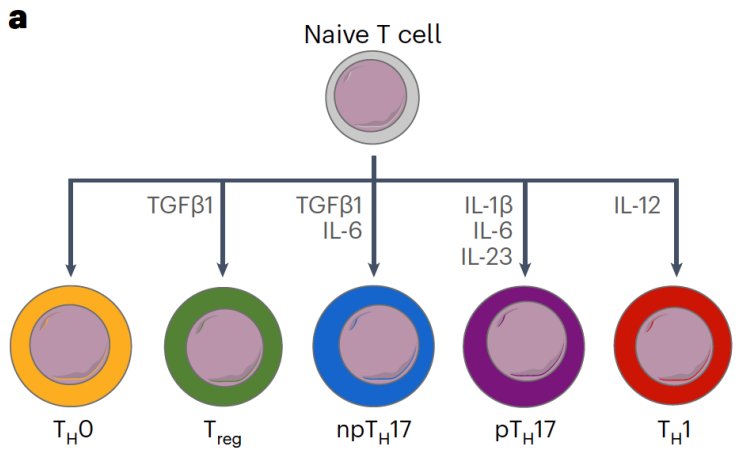


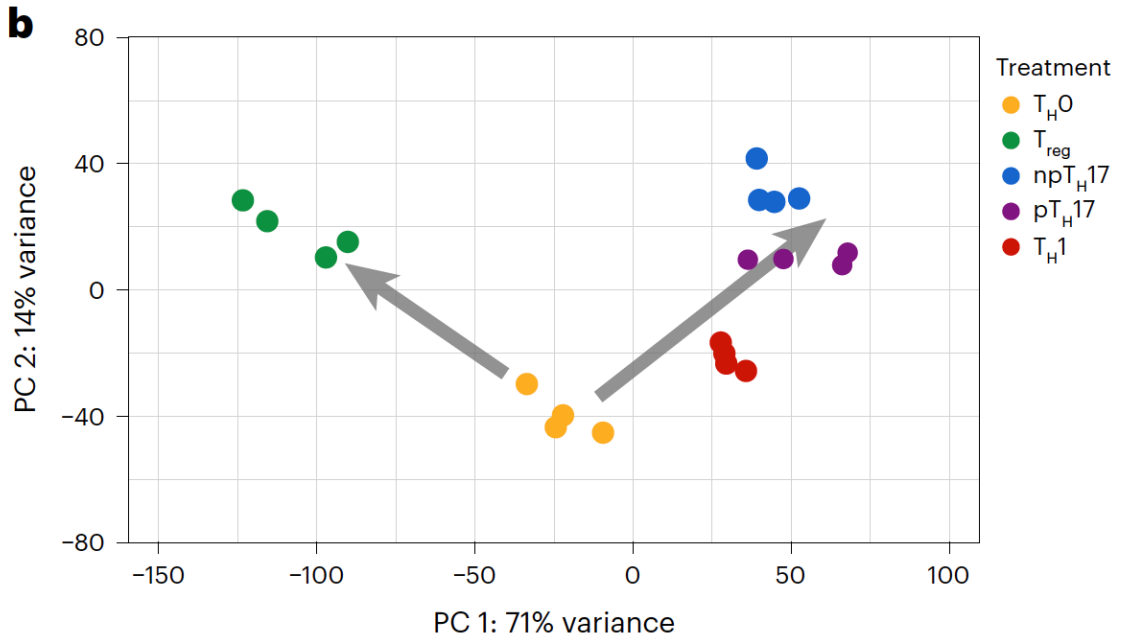


02 Differential Gene Expression

2.2. DEGs among groups

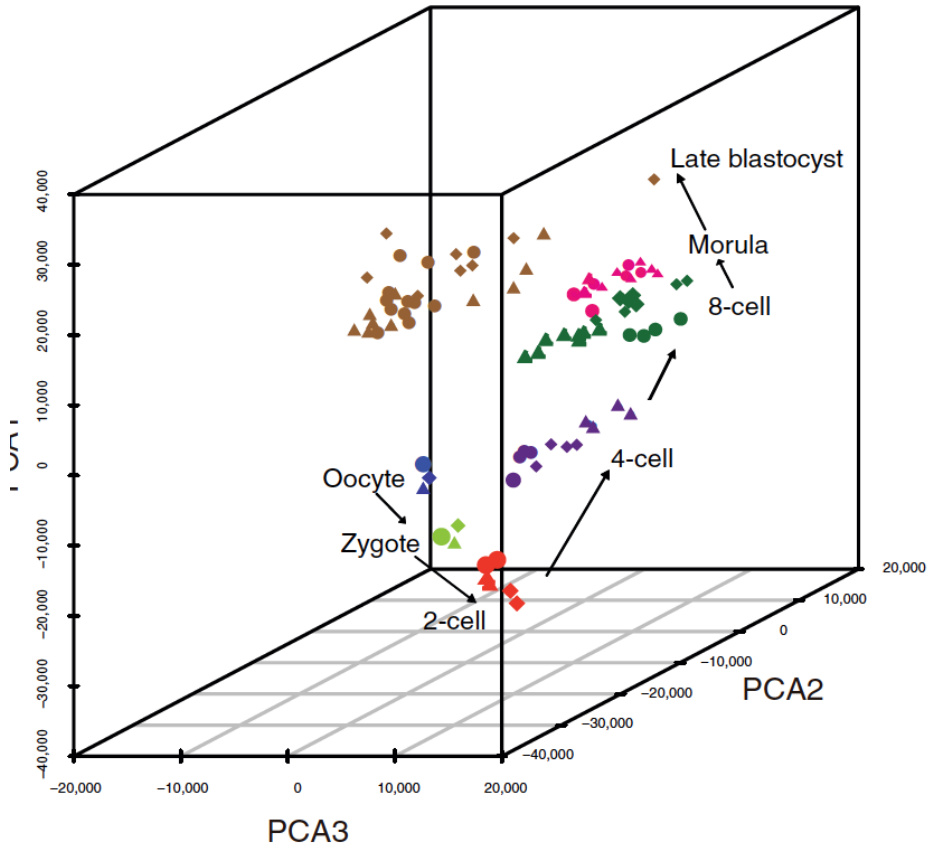
PCA: Principal Component Analysis

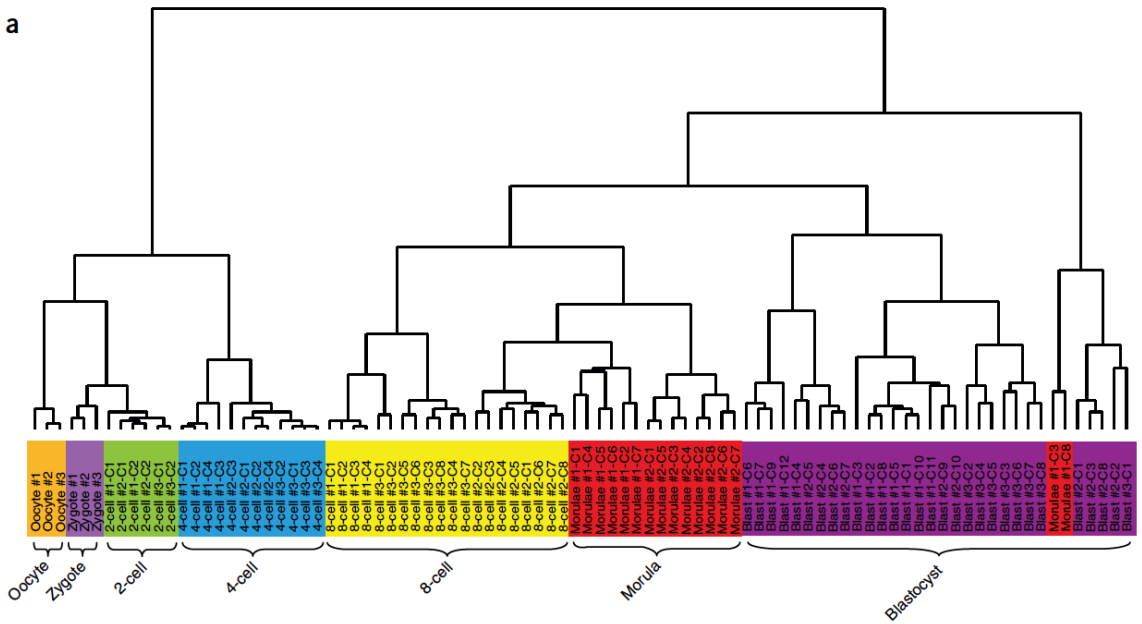




02 Differential Gene Expression

2.2. DEGs among groups

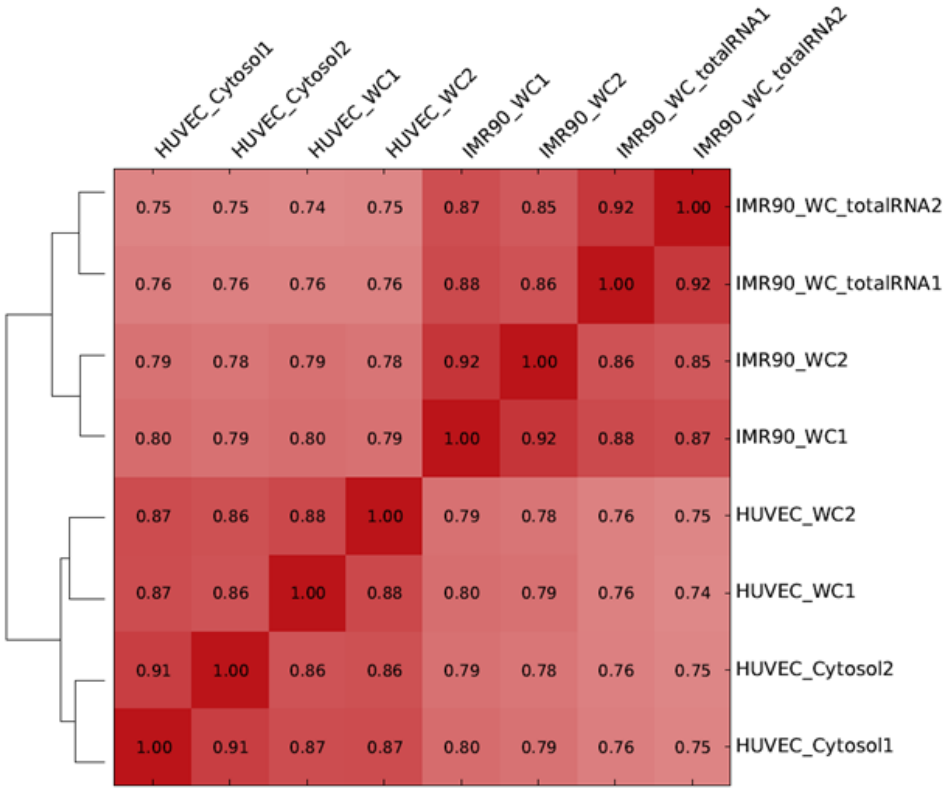


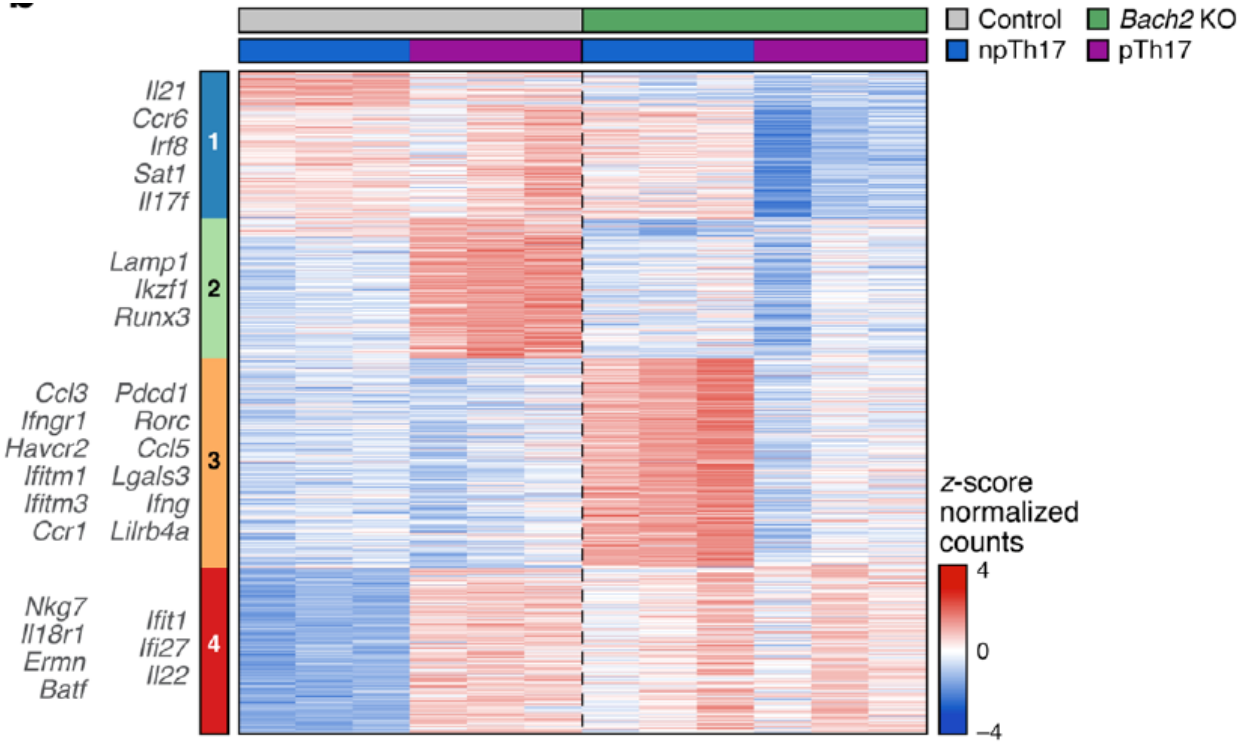


02 Differential Gene Expression

2.2. DEGs among groups

Clustering





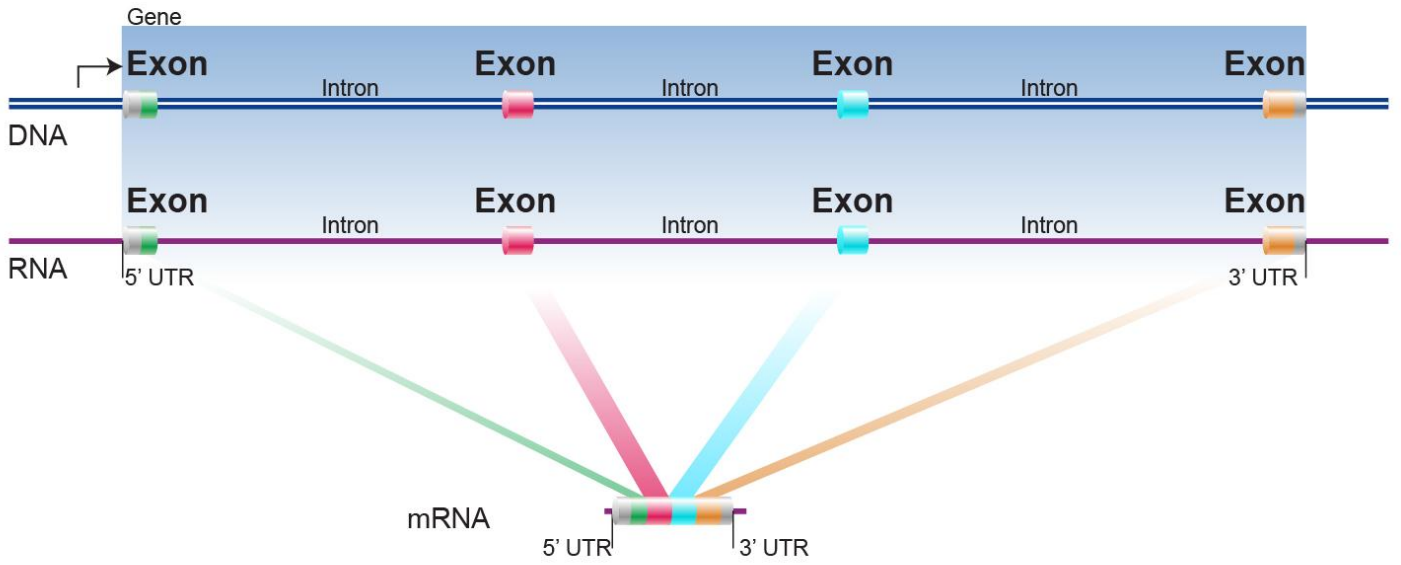
03 Identify novel transcripts

3. Identify Novel Transcripts and Splicing Variants

RNA-seq is not limited to known genes — it can detect:

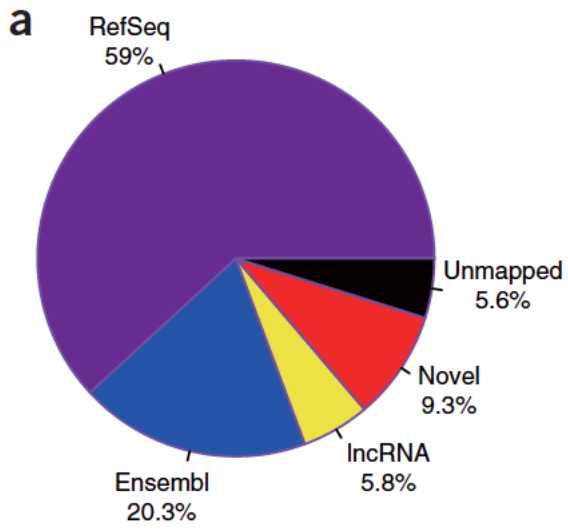
- Novel transcripts
- Alternative splicing events
- Gene fusions
- Non-coding RNAs

This helps in understanding transcriptome complexity and gene regulation.

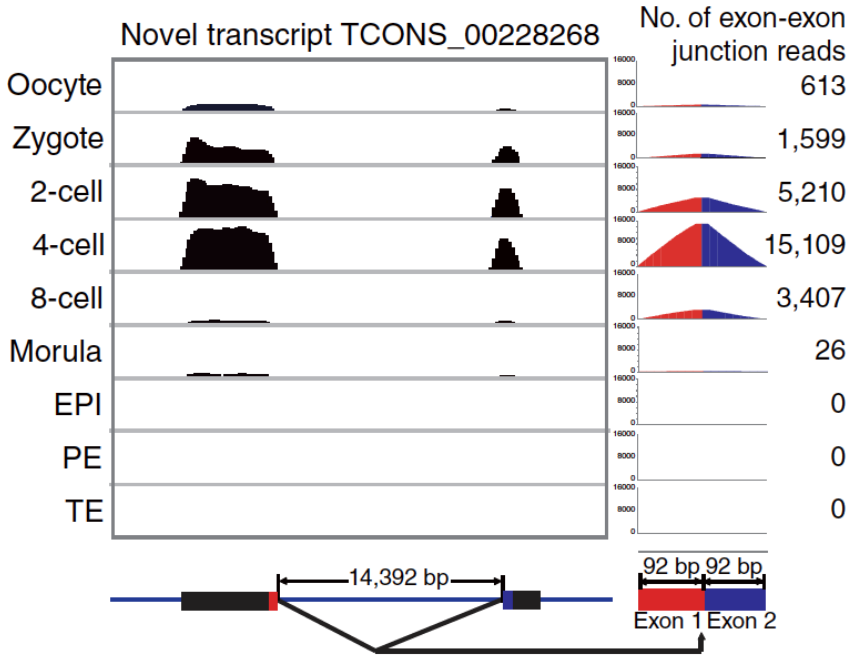


03 Identify novel transcripts

3.1. Identify novel lncRNA



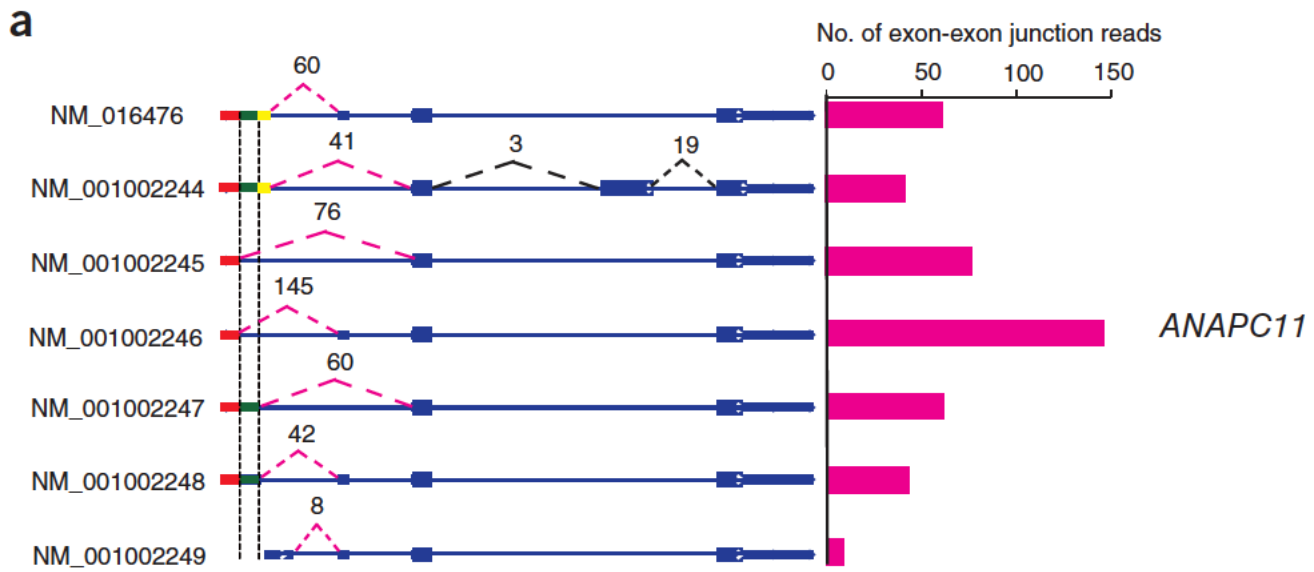
Pie chart of the percentage of reads aligned to different classes of genes.



Coverage plots of RNA-Seq reads of a novel lncRNA during preimplantation development.

03 Identify novel transcripts

3.2. Novel splicing variants



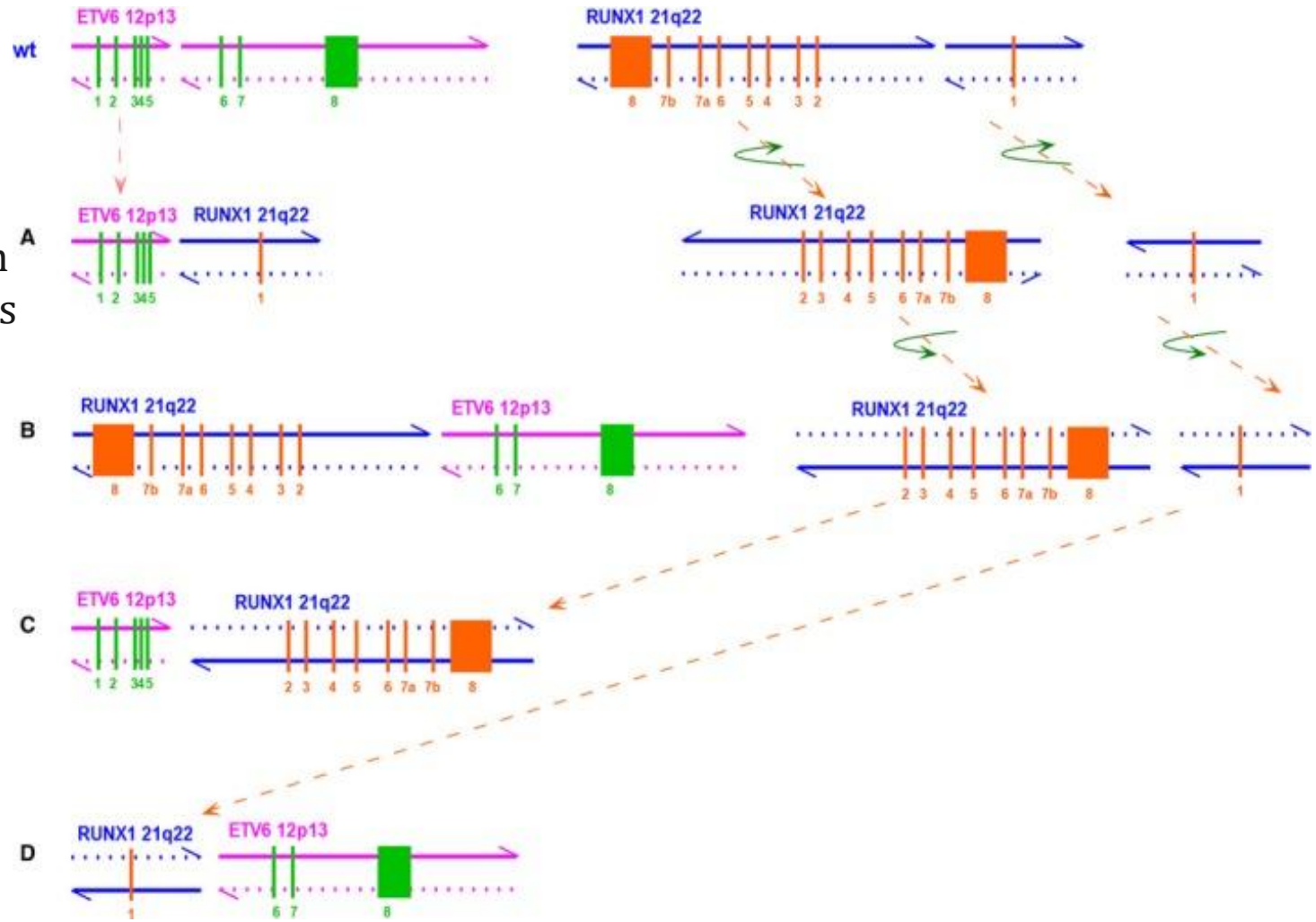
Exon-exon junction plots of all of the seven transcript variants of ANAPC11 in an individual hESC.

03 Identify novel transcripts

3.3. Gene fusion

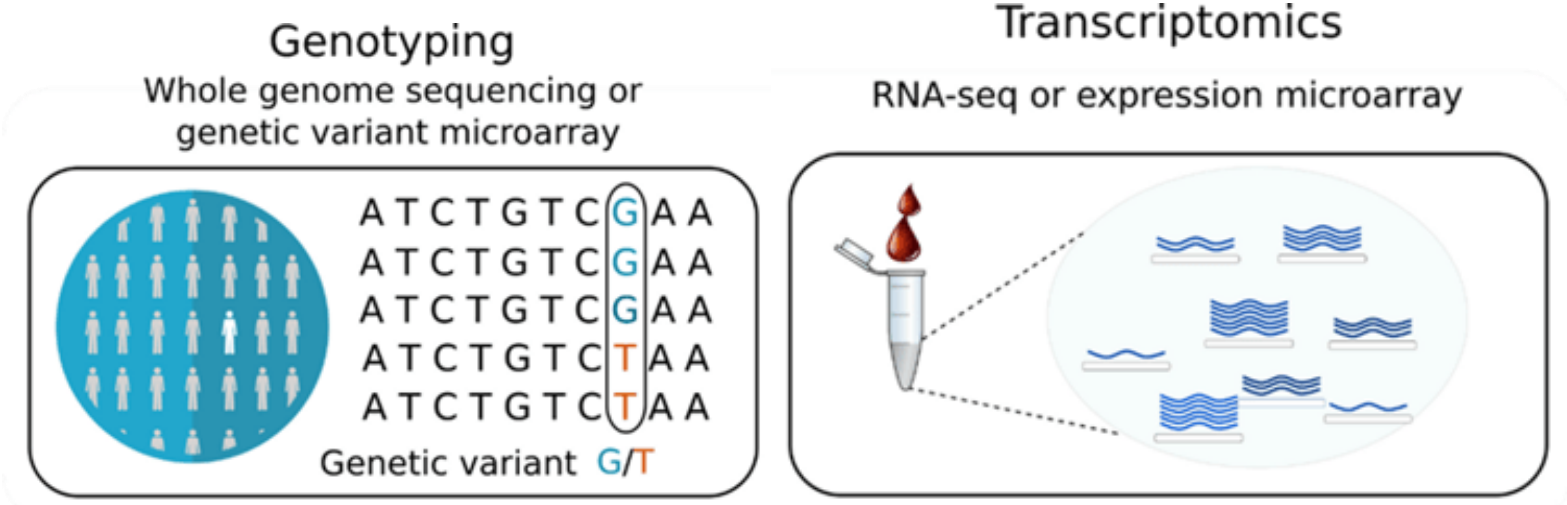
ETV6-RUNX1 fusion protein is expressed in 25% of childhood B-lineage ALL cases and is associated with favorable prognosis following conventional therapeutic strategies

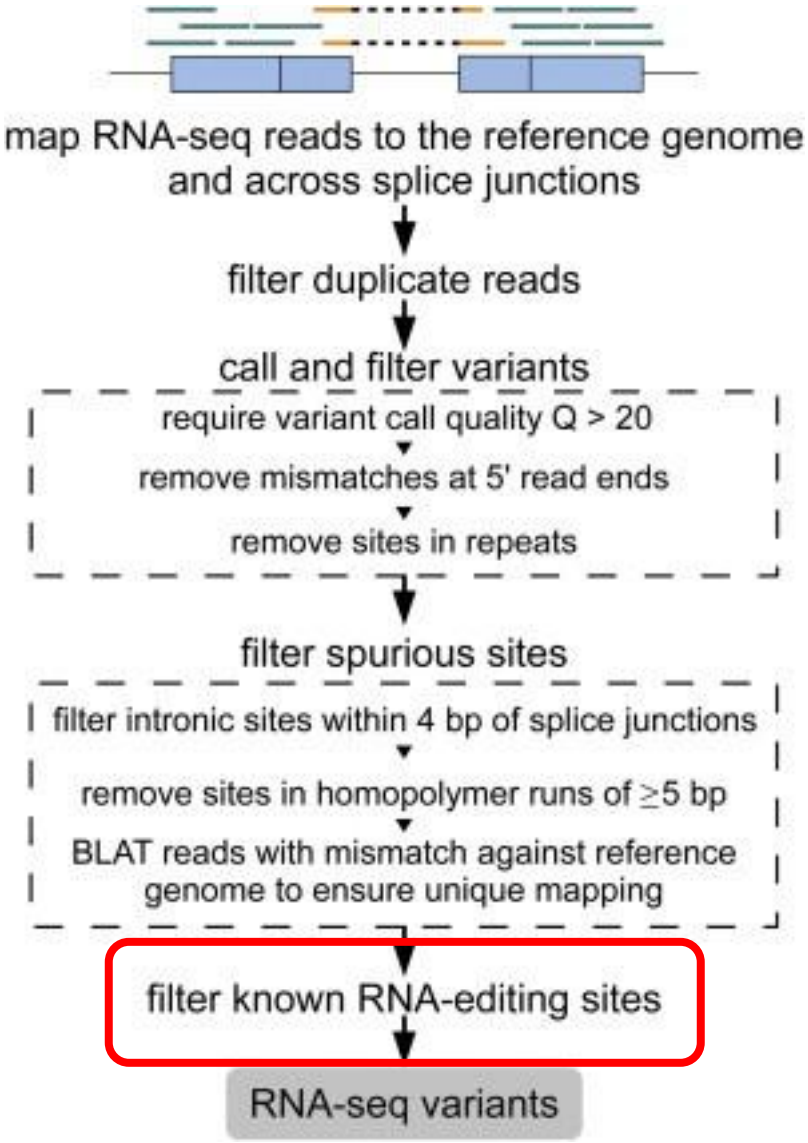
Acute lymphocytic leukemia (ALL) is a type of cancer of the blood and bone marrow



04 Identify genetic mutations

4. Identify genetic mutations

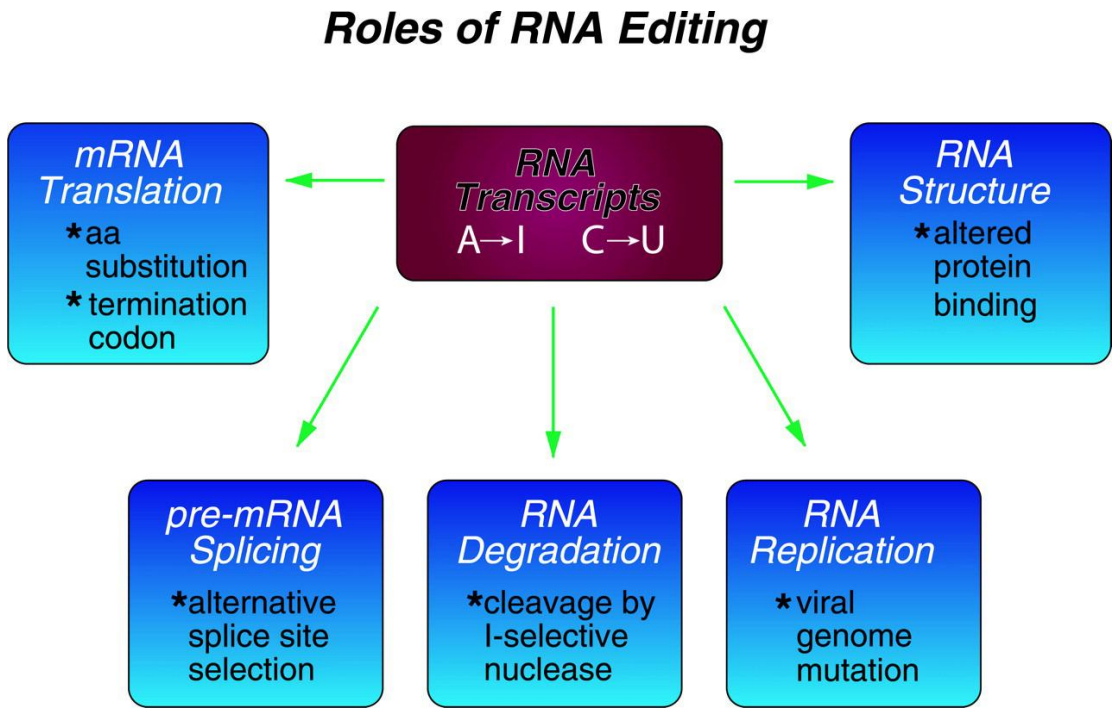




04 Identify genetic mutations

4. Identify genetic mutations

RNA editing is a post-transcriptional process where the nucleotide sequence of an RNA molecule is altered after it has been transcribed from DNA, without changing the underlying DNA sequence.
This process can change the coding potential, splicing, stability, or localization of the RNA, and therefore affect the protein it encodes — or even whether it is translated at all.



04 Identify genetic mutations

4. Identify genetic mutations

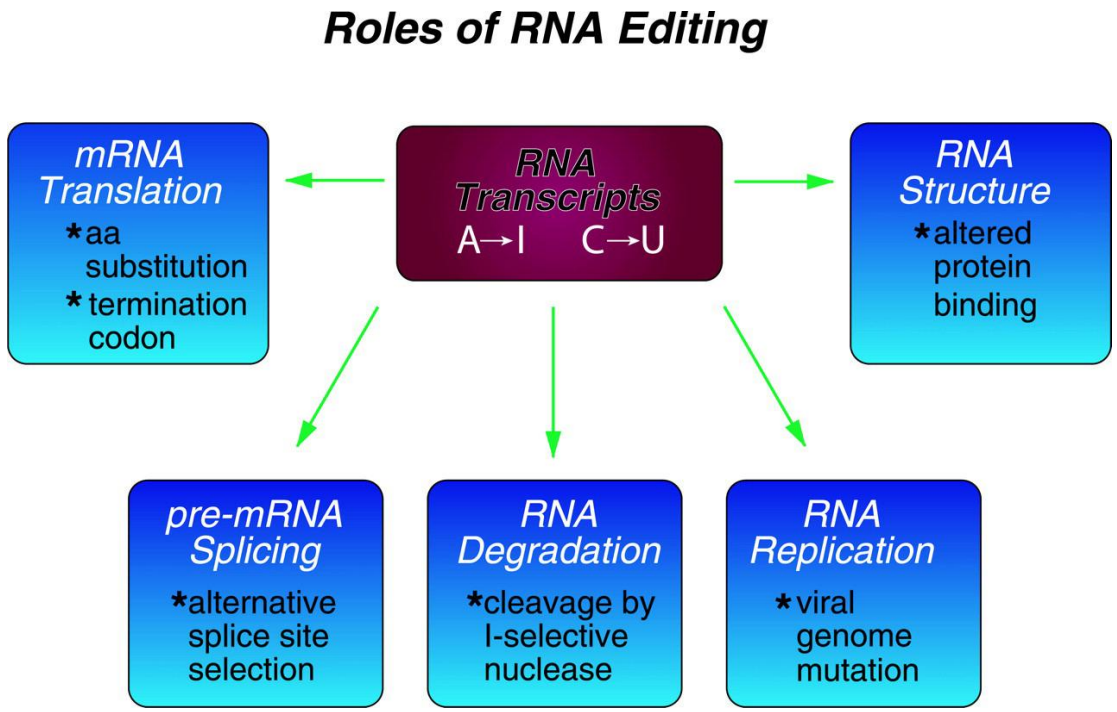
Main Types of RNA Editing (in humans):

1. A-to-I editing (Adenosine to Inosine)

- Most common type in humans
- Catalyzed by ADAR enzymes (Adenosine Deaminases Acting on RNA)
- Inosine is read as guanosine (G) by ribosomes and sequencing machinery
- Common in brain tissue, affecting neurotransmission-related genes

2. C-to-U editing (Cytidine to Uridine)

- Catalyzed by APOBEC family enzymes
- Famous example: APOB mRNA, where editing produces a stop codon, altering lipid metabolism



04 Identify genetic mutations

4. Identify genetic mutations

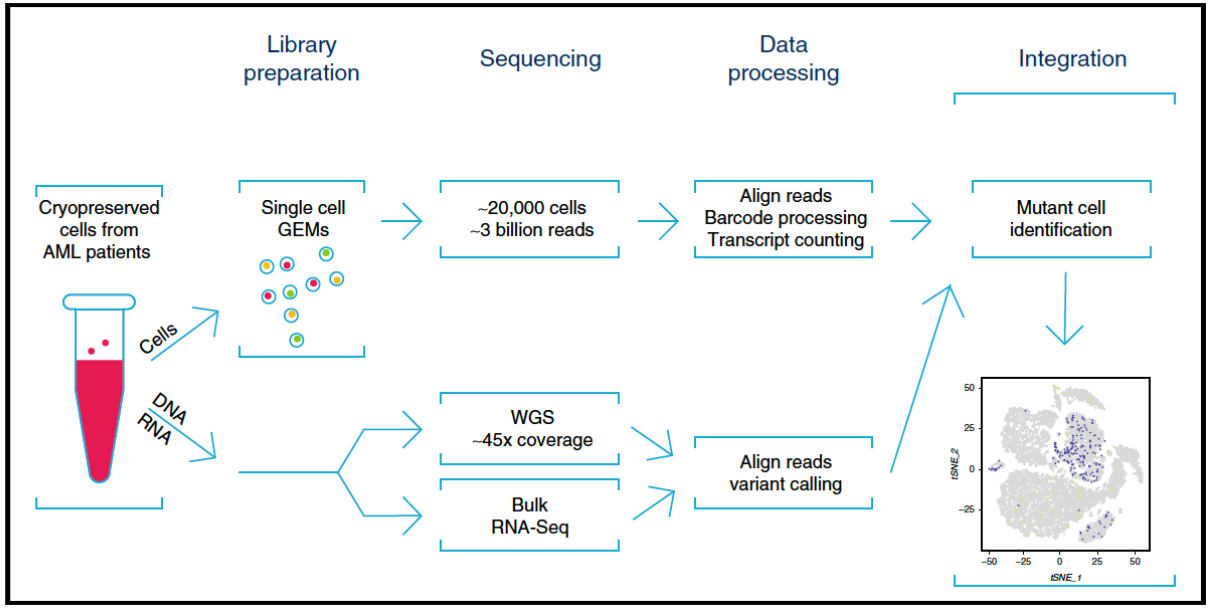
ARTICLE

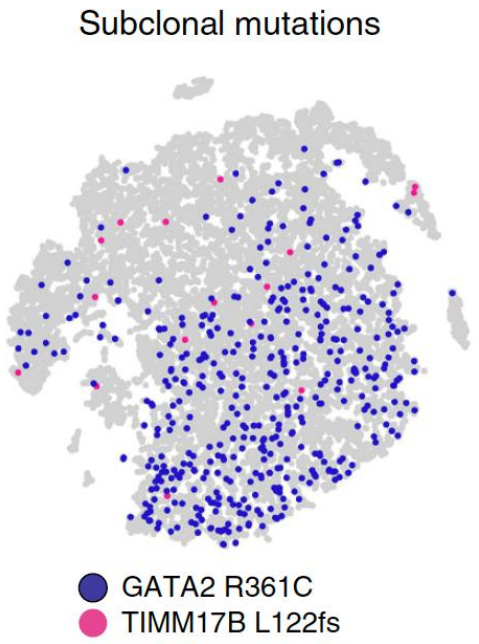
There are amendments to this paper

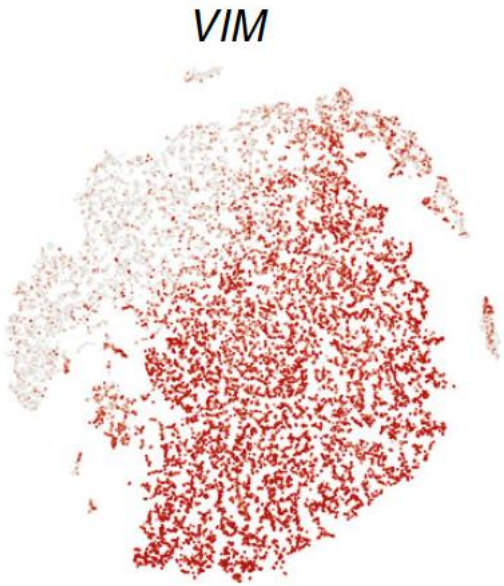
https://doi.org/10.1038/s41467-019-11591-1 OPEN

A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing

Allegra A. Petti[1,2,7], Stephen R. Williams[3,7], Christopher A. Miller[1,2], Ian T. Fiddes[3], Sridhar N. Srivatsan[1], David Y. Chen[4], Catrina C. Fronick[2], Robert S. Fulton[2], Deanna M. Church[5] & Timothy J. Ley[1,2,6]

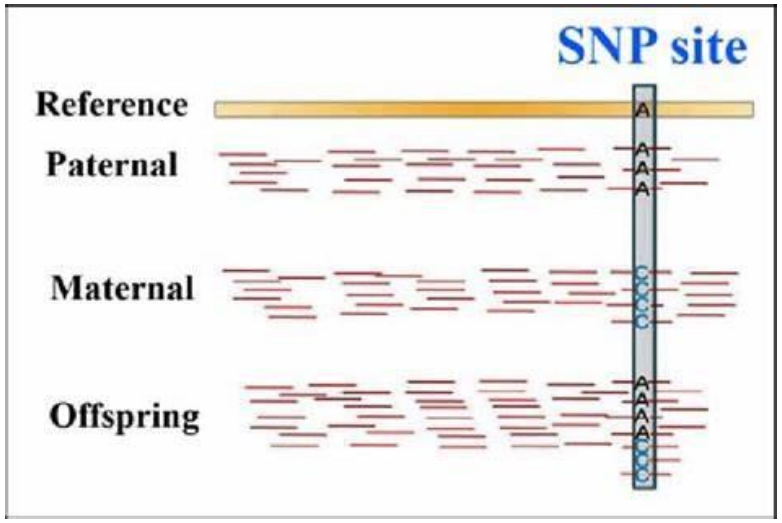


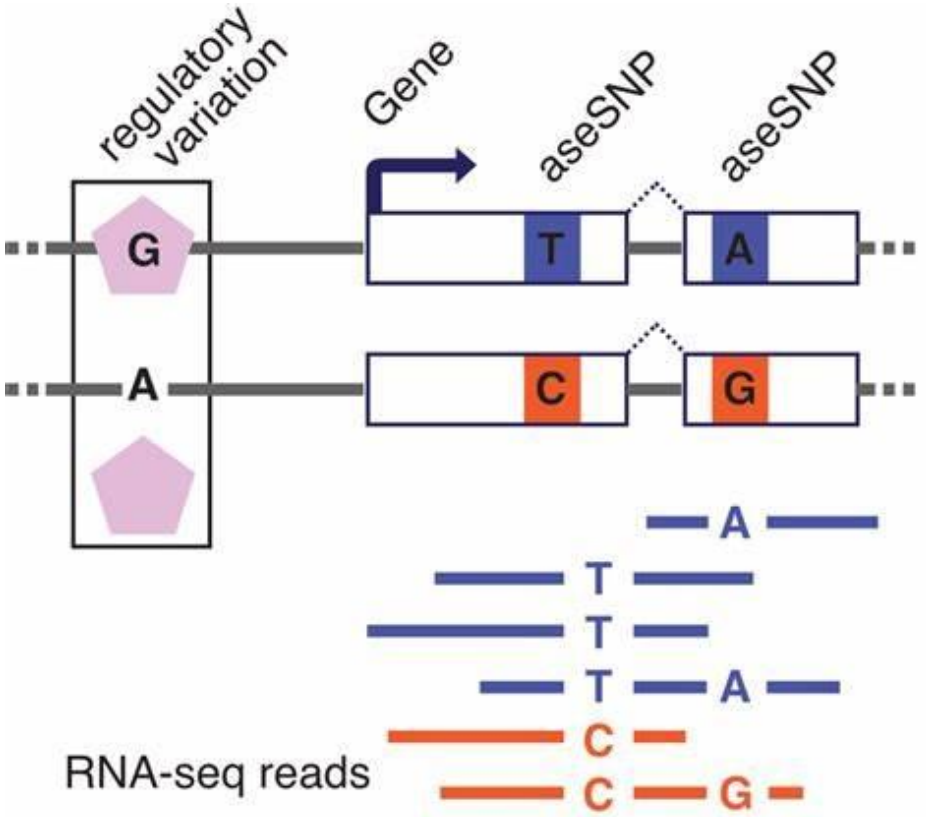




05 Allele specific gene expression

5. Allele specific gene expression





05 Allele specific gene expression

5. Allele specific gene expression

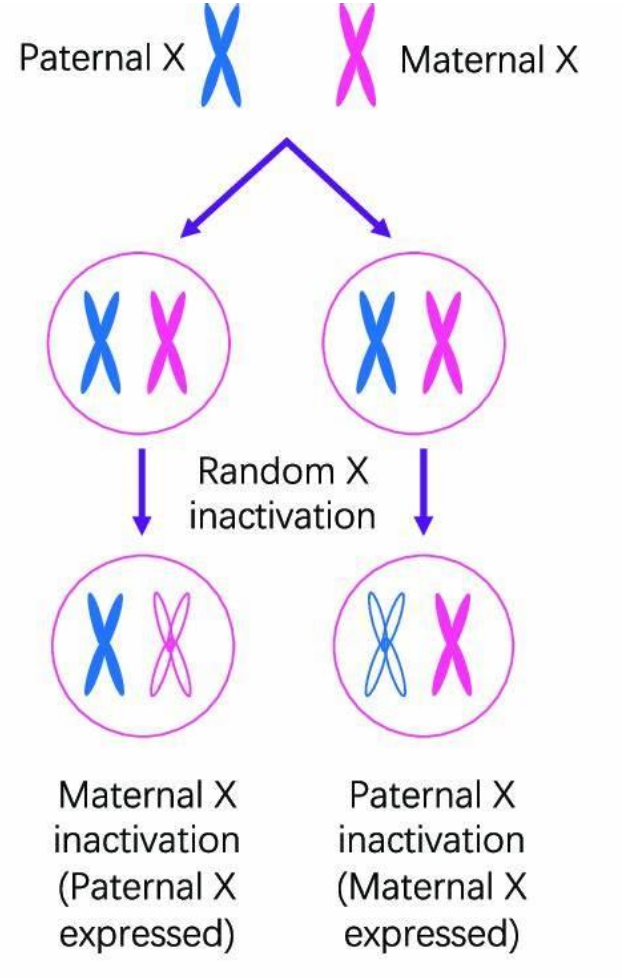
Allele specific expression

Bi-allelic

ACTGGCATTGAGCAATTCCTAGGGACC Paternal allele
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCATTTCCTAGGGACC Maternal allele
ACTGGCATTGAGCATTTCCTAGGGACC
ACTGGCATTGAGCATTTCCTAGGGACC
ACTGGCATTGAGCATTTCCTAGGGACC

Allelic imbalance

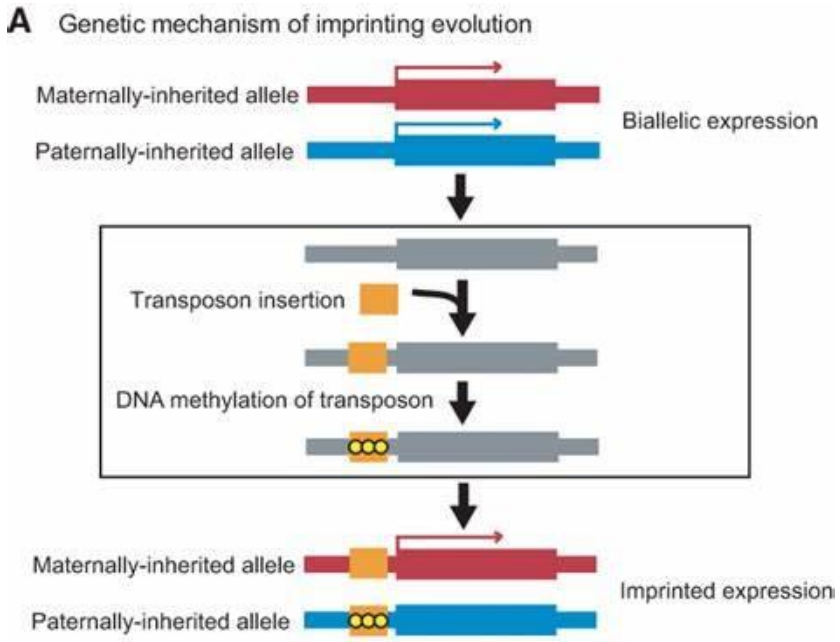
ACTGGCATTGAGCAATTCCTAGGGACC Paternal allele
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCAATTCCTAGGGACC
ACTGGCATTGAGCATTTCCTAGGGACC Maternal allele
ACTGGCATTGAGCATTTCCTAGGGACC



05 Allele specific gene expression

5. Allele specific gene expression

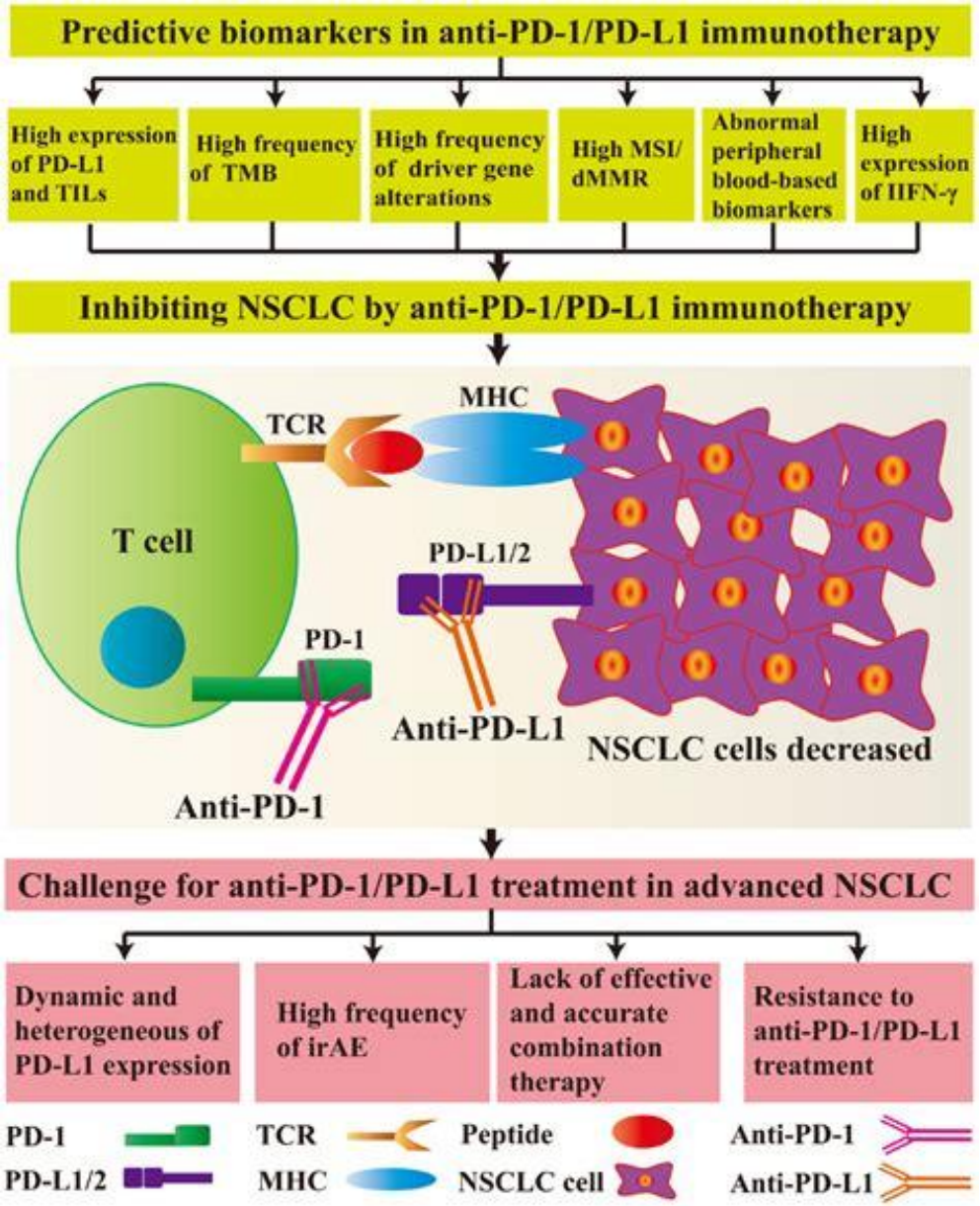
Genomic imprinting is an epigenetic phenomenon where certain genes are expressed in a parent-of-origin-specific manner. This means that the expression of an imprinted gene depends on whether it is inherited from the mother or the father.



06 Support biomarkers

6. Support Biomarker and Drug Target Discovery

Differentially expressed genes or unique expression signatures identified through RNA-seq can serve as potential **biomarkers**, **therapeutic targets**, or indicators of **drug response/resistance**.

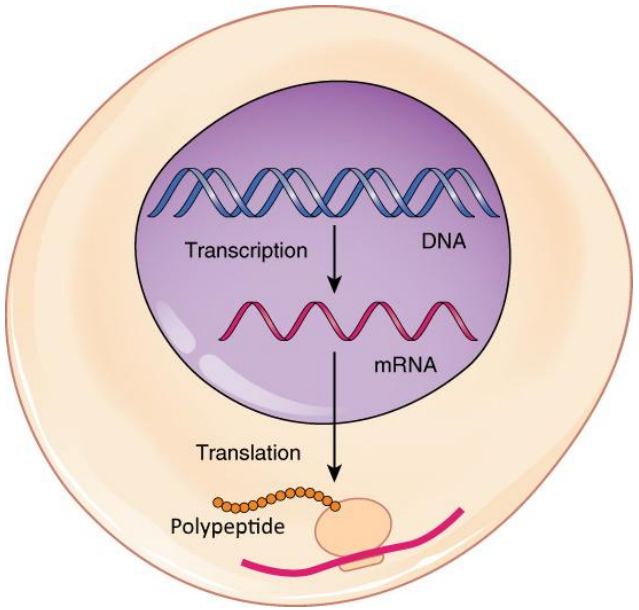


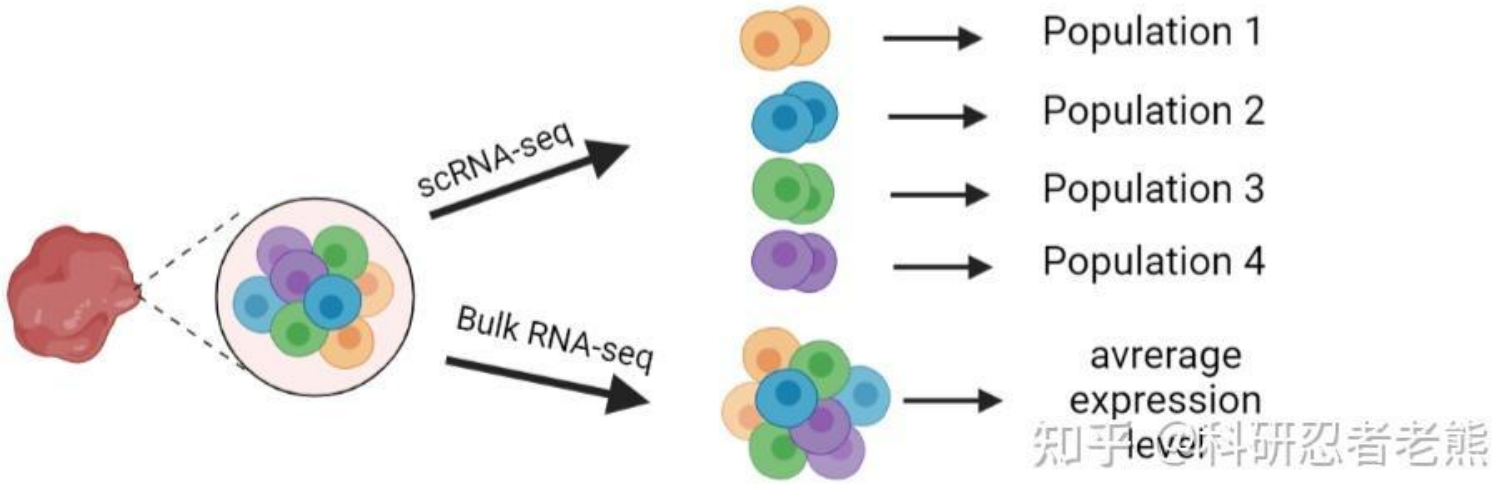
Single cell RNAseq

scRNA-seq

scRNA-seq (single-cell RNA-seq)

A single cell:
10-20 um;
~6pg gDNA；~20pg total RNA (80-85% are rRNA)

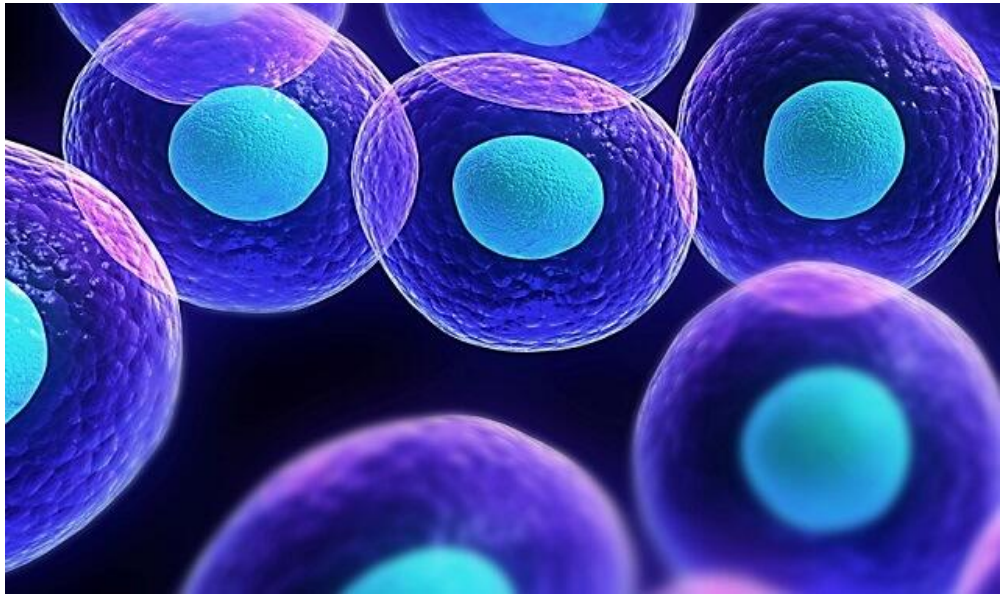




scRNA-seq

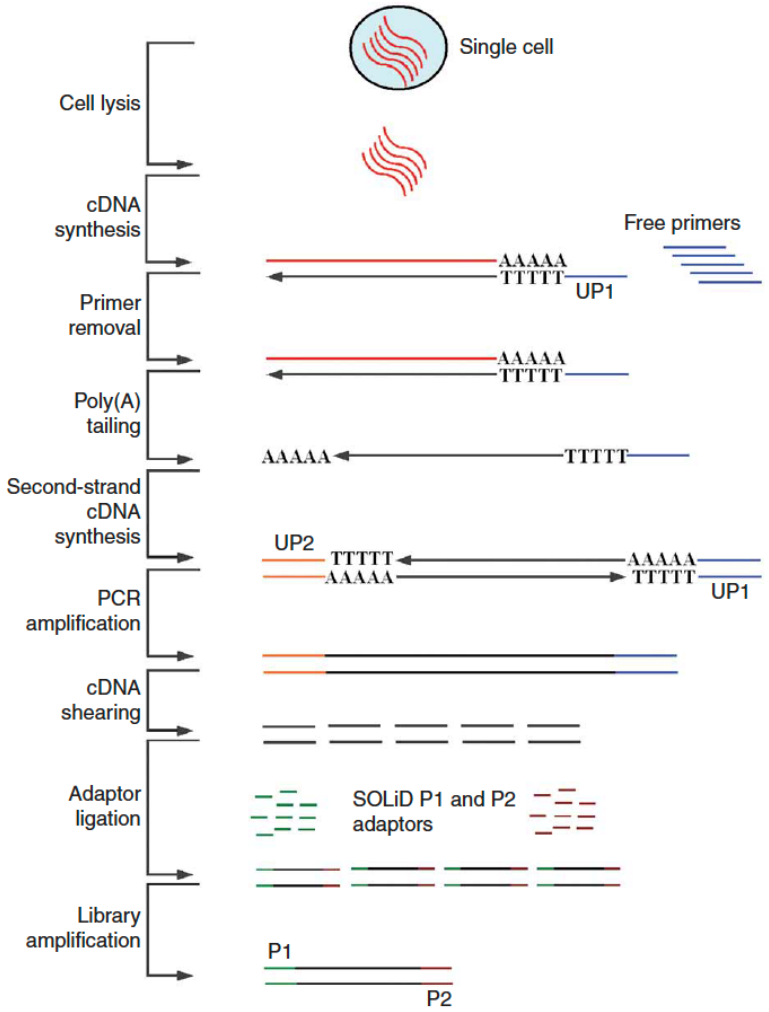
scRNA-seq (single-cell RNA-seq)

1. **Single cell RNA-seq techniques** (Tang protocol, Smart-seq2, Drop-seq, 10x genomimcs)
2. **Data analyses** (Seurat)
3. **Application of scRNA-seq** (Embryonic development, Cancer, Immune, Aging)

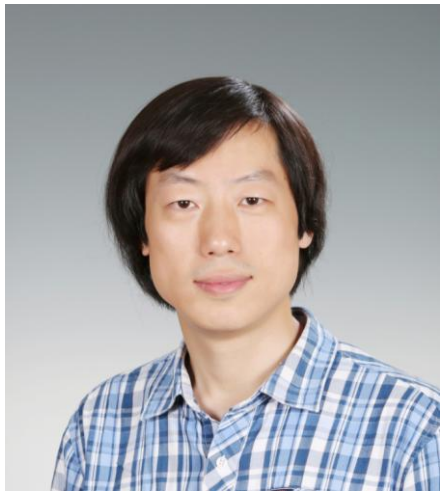


01 scRNA-seq techniques

1.1 Tang protocol



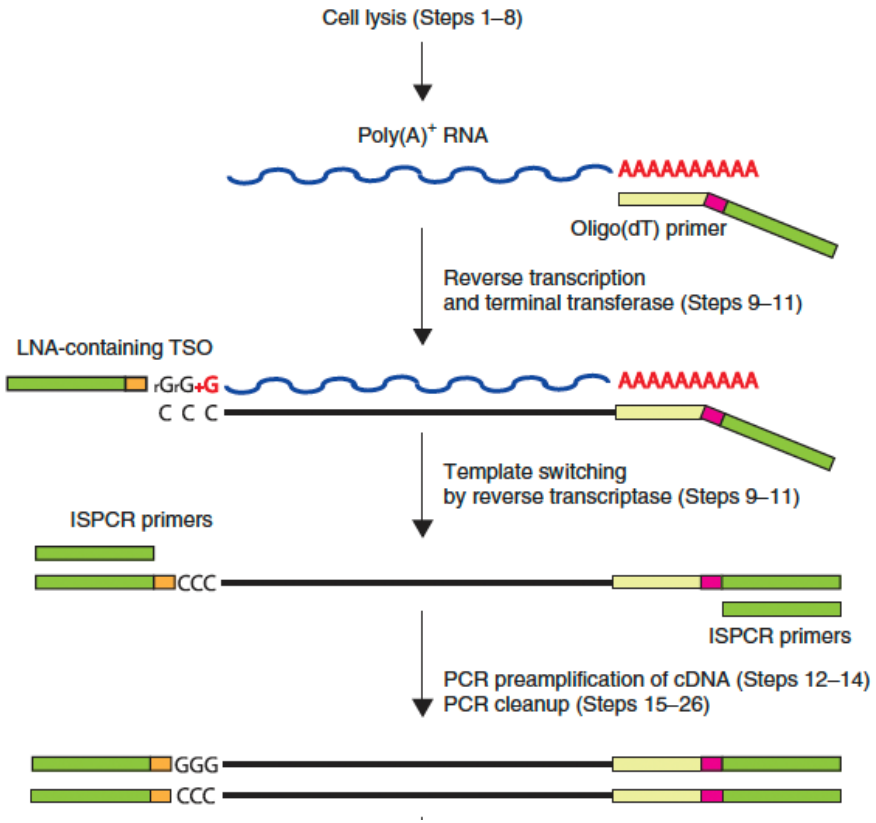
1. Amplify 5pg RNA to 1ug cDNA
2. Full length of cDNA
3. Followed by DNA library preparation
4. Individual single cell picking and amplification
5. Low throughput

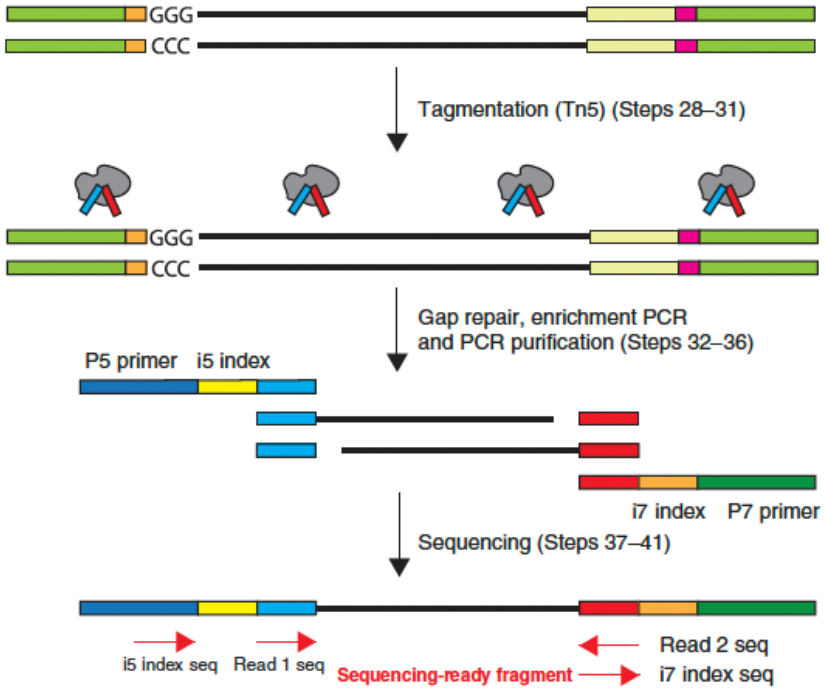


Fuchou Tang et al., *Nature Methods*, 2009

01 scRNA-seq techniques

1.2 SMART-seq



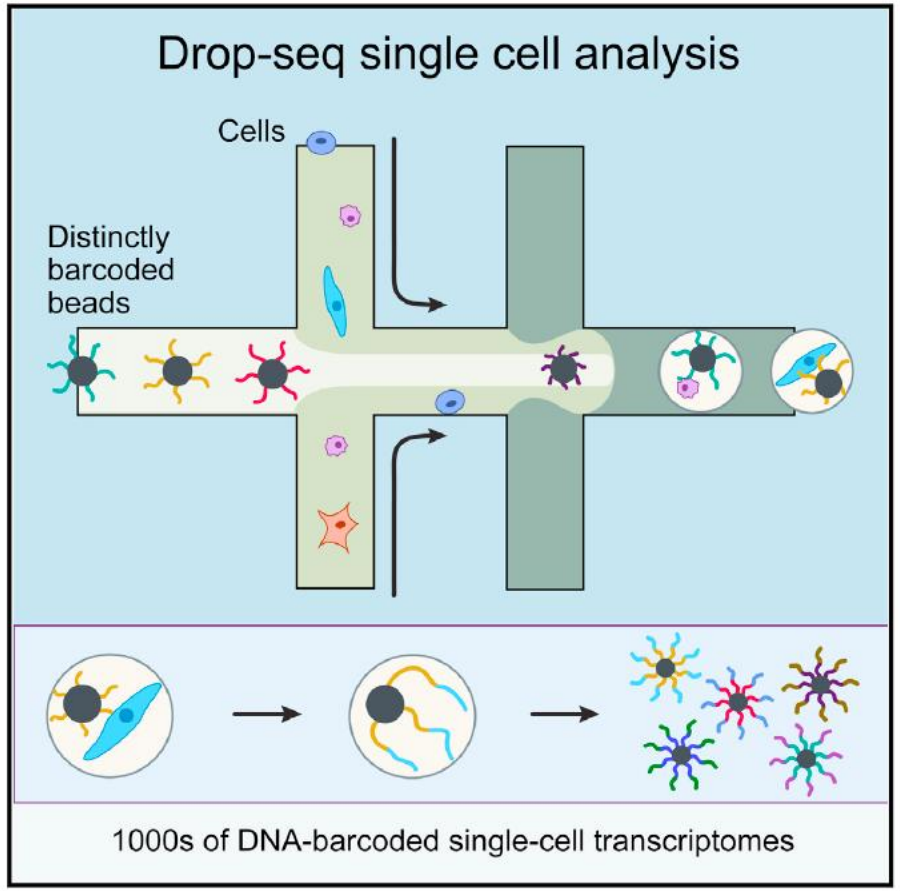


1. Full length of cDNA
2. Individual single cell picking and amplification
3. Low throughput

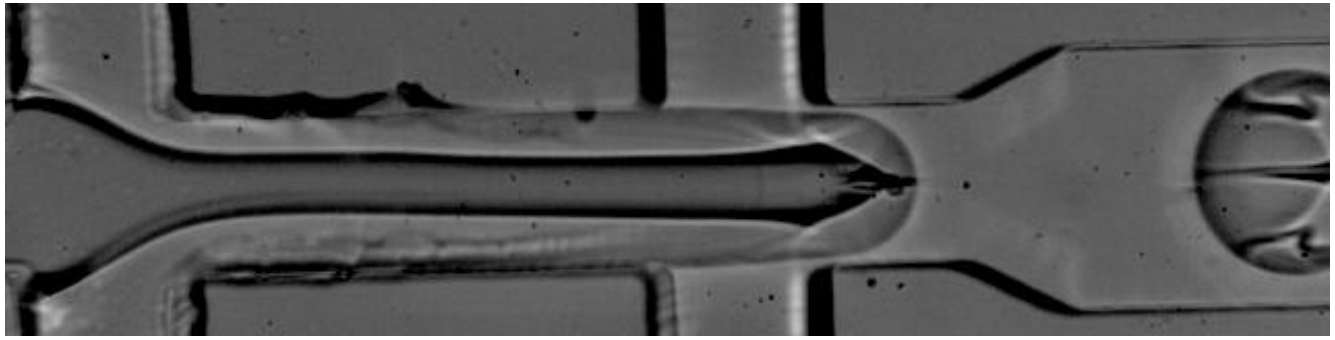
Simone Picelli et al., *Nature Protocols*, 2014

01 scRNA-seq techniques

1.3 DROP-seq



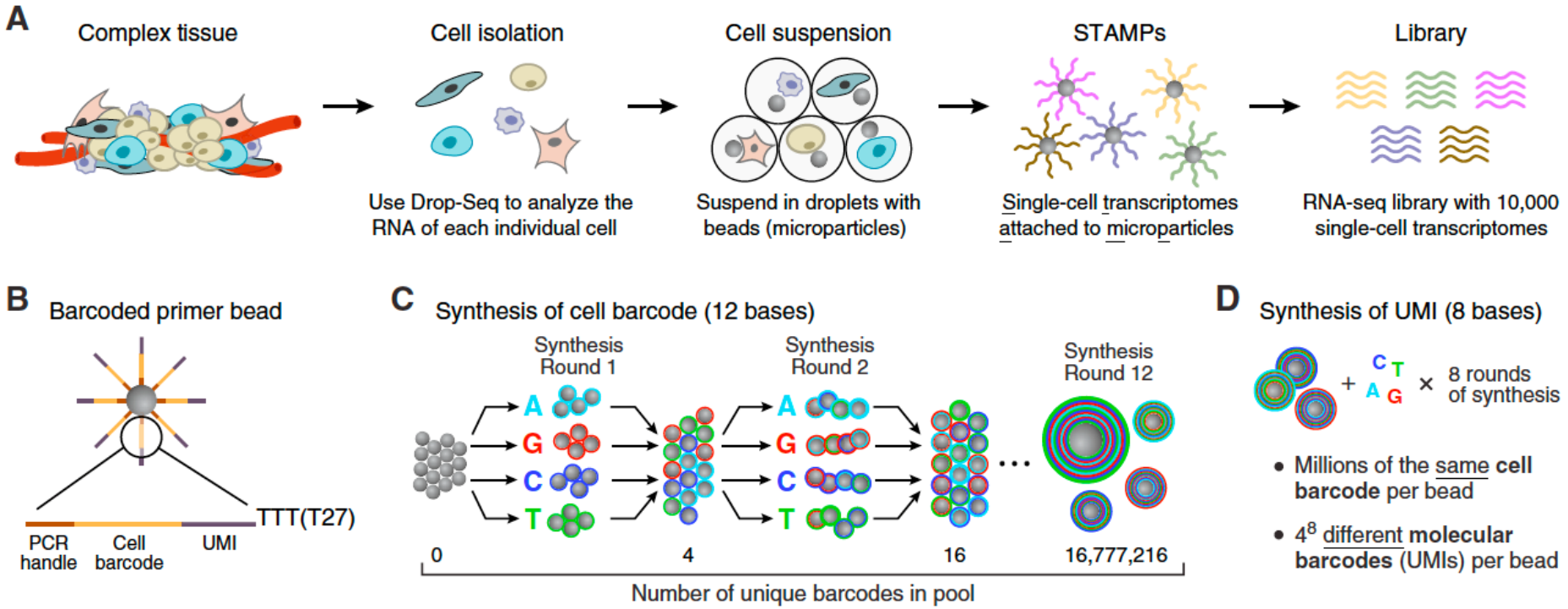
1. High throughput
2. 3 prime of cDNA
3. Low cost



Macosko et al., *Cell*, 2015

01 scRNA-seq techniques

1.3 DROP-seq



Unique Molecular Identifier (UMI)

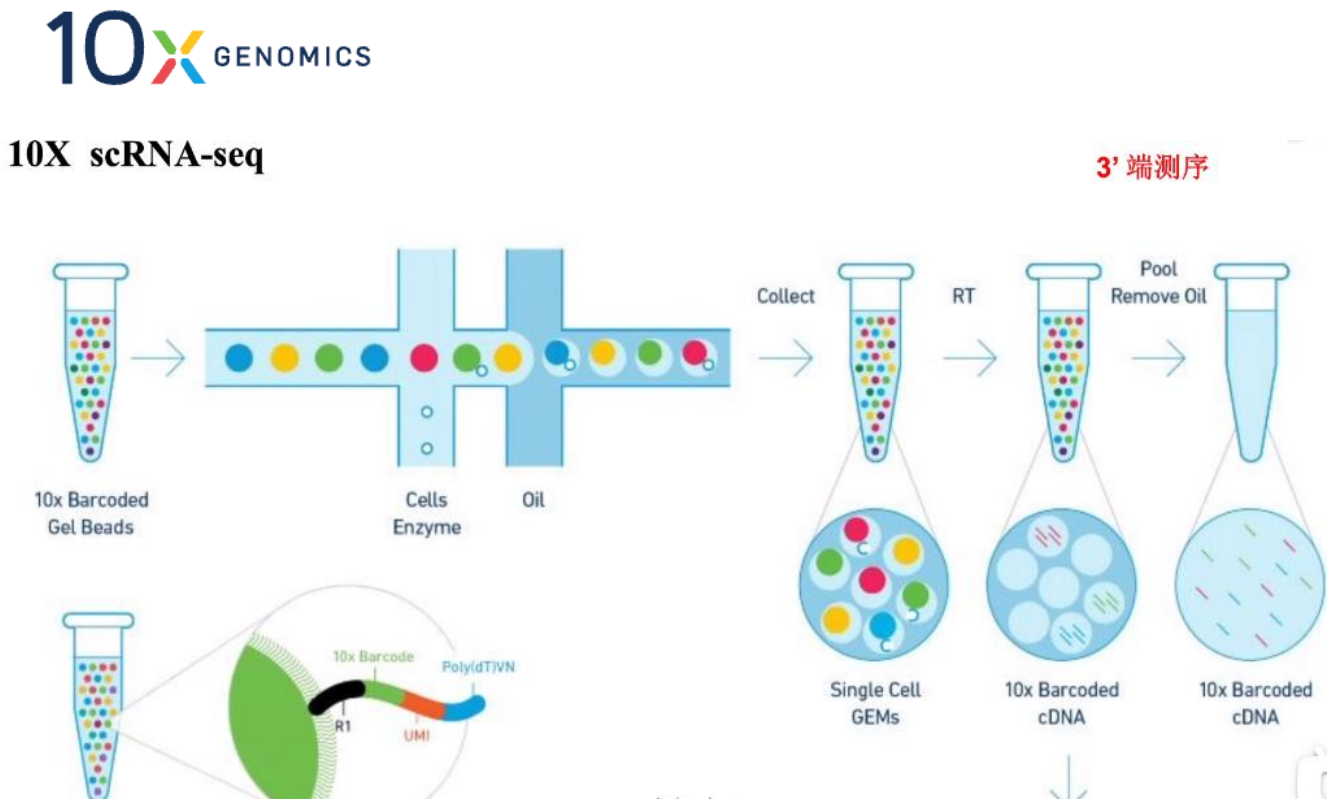
Macosko et al., *Cell*, 2015

01 scRNA-seq techniques

1.4 10x Genomics

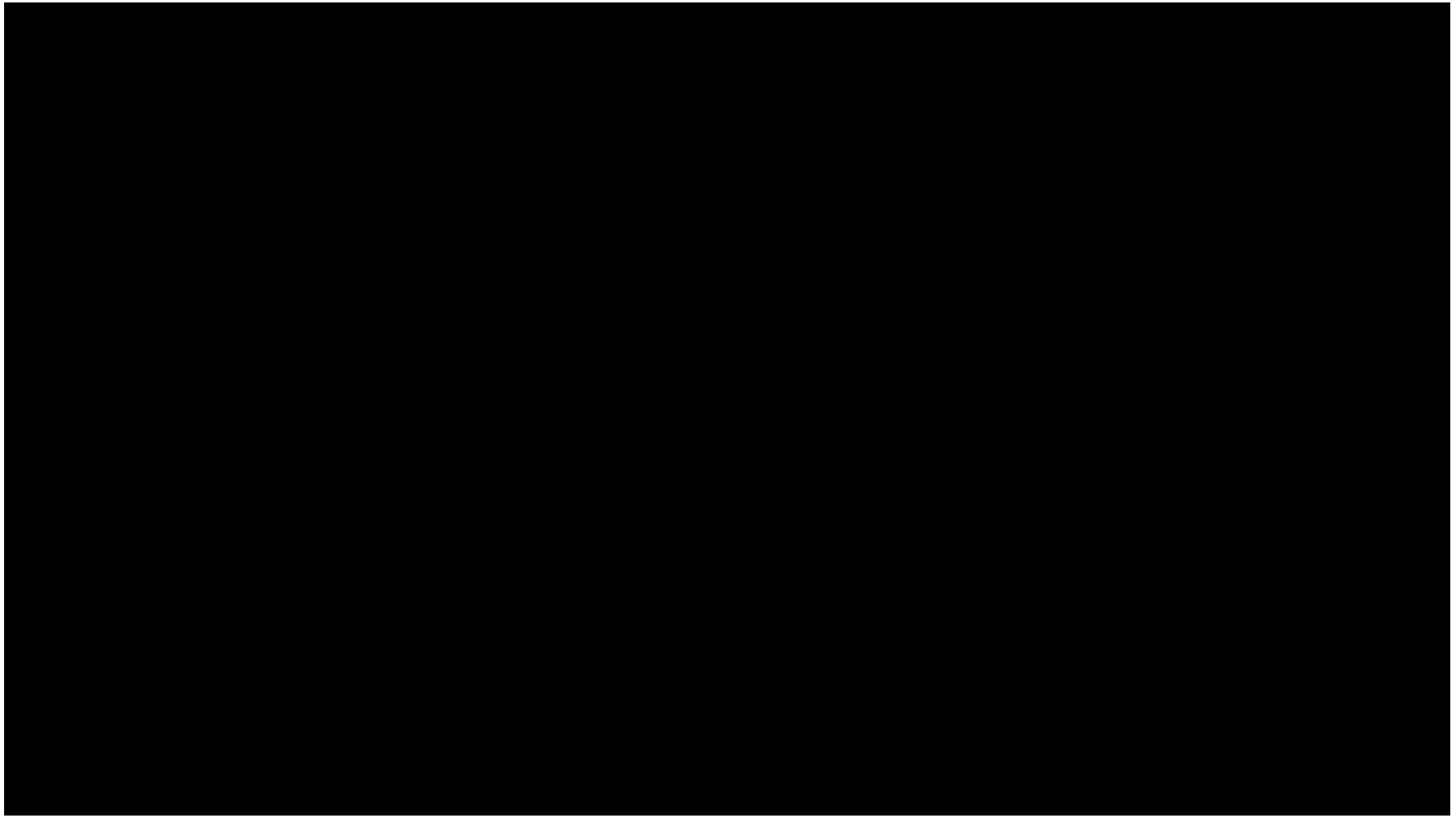
10x Genomics RNA sequencing (10x RNA-Seq) is a high-throughput single-cell transcriptomics method that allows for the profiling of thousands to millions of individual cells in a single experiment.

10x RNA-seq is one of the most widely used methods for single-cell RNA analysis, enabling breakthroughs in cell biology, disease research, and precision medicine.



01 scRNA-seq techniques

1.4 10x Genomics



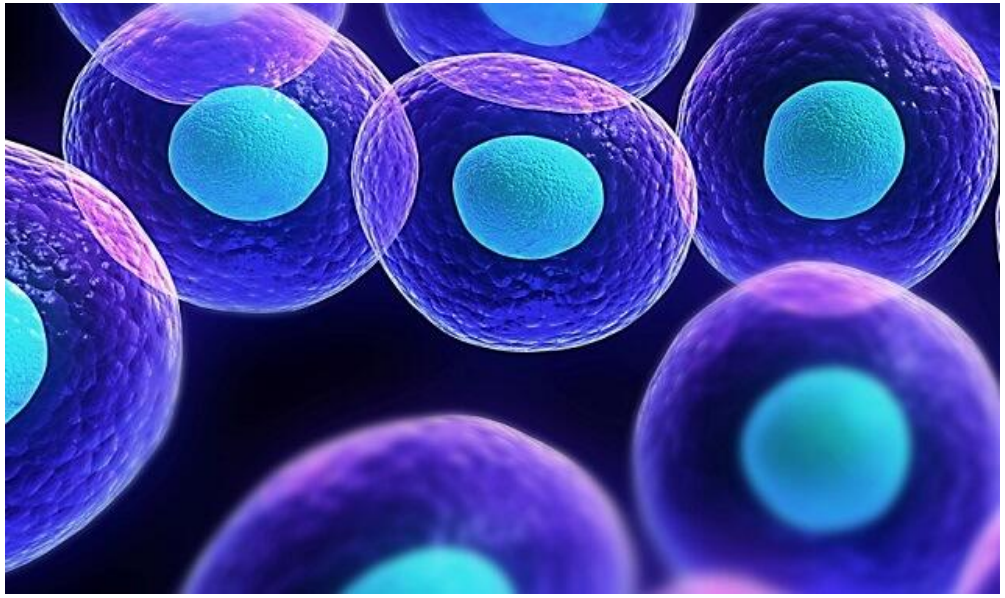
01 scRNA-seq techniques

Method	Cell Capture	Full-Length Transcript?	Throughput	Cost per Cell	Advantages	Limitations
Tang Protocol	Manual (micropipette or FACS)	☑ Yes	Low (~100s)	High	- First scRNA-seq method - Captures full-length transcripts	- Extremely low throughput - Labor-intensive - High cost
Smart-seq2	Plate-based (FACS)	☑ Yes	Low (~100s)	High	- High sensitivity - Detects full-length mRNA - Suitable for low-input	- Low throughput - Higher cost - Batch variability
Drop-seq	Droplet-based	✗ No (3′ only)	High (~10,000s)	Low	- High throughput - Low cost - Incorporates UMIs	- 3′-end only - Lower transcript coverage
10x Genomics	Droplet-based	✗ No (3′ or 5′)	Very High (>100,000s)	Medium-High	- Highly standardized - Commercial support - Supports multi-omics	- Expensive reagents - 3′/5′ only - Black-box workflow

scRNA-seq

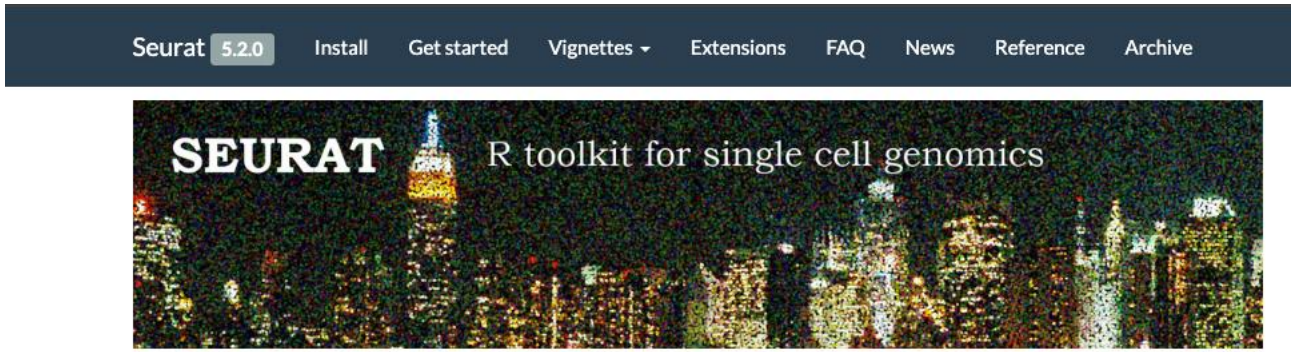
scRNA-seq (single-cell RNA-seq)

1. **Single cell RNA-seq techniques** (Tang protocol, Smart-seq2, Drop-seq, 10x genomimcs)
2. **Data analyses** (Seurat)
3. **Application of scRNA-seq** (Embryonic development, Cancer, Immune, Aging)



02 scRNA-seq Data analyses

https://satijalab.org/seurat/



Seurat v5

We are excited to release Seurat v5! To install, please follow the instructions in our install page. This update brings the following new features and functionality:

- **Integrative multimodal analysis:** The cellular transcriptome is just one aspect of cellular identity, and recent technologies enable routine profiling of chromatin accessibility, histone modifications, and protein levels from single cells. In Seurat v5, we introduce 'bridge integration', a statistical method to integrate experiments measuring different modalities (i.e. separate scRNA-seq and scATAC-seq datasets), using a separate multiomic dataset as a molecular 'bridge'. For example, we demonstrate how to map scATAC-seq datasets onto scRNA-seq datasets, to assist users in interpreting and annotating data from new modalities.

Contents

02 scRNA-seq Data analyses

Contents


```
library(dplyr)
library(Seurat)
library(patchwork)

# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "/brahms/mollag/practice/filtered_gene_bc_matrices/hg19/")
# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 200)
pbmc
```

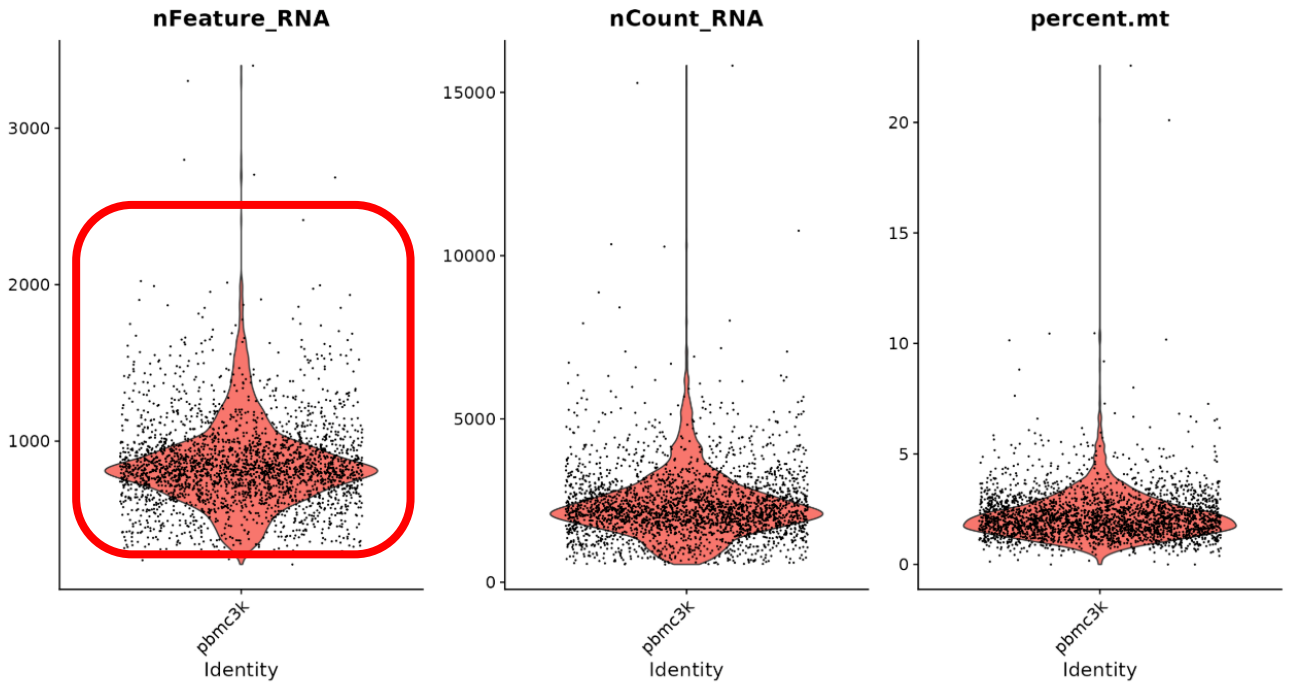
```
## An object of class Seurat
## 13714 features across 2700 samples within 1 assay
## Active assay: RNA (13714 features, 0 variable features)
##  1 layer present: counts
```

https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
# Visualize QC metrics as a violin plot
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
```




```
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
```

https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents

After removing unwanted cells from the dataset, the next step is to normalize the data. By default, we employ a global-scaling normalization method "LogNormalize" that normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result.

```
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize", scale.factor = 10000)
```

While this method of normalization is standard and widely used in scRNA-seq analysis, global-scaling relies on an assumption that each cell originally contains the same number of RNA molecules.

https://satijalab.org/seurat/

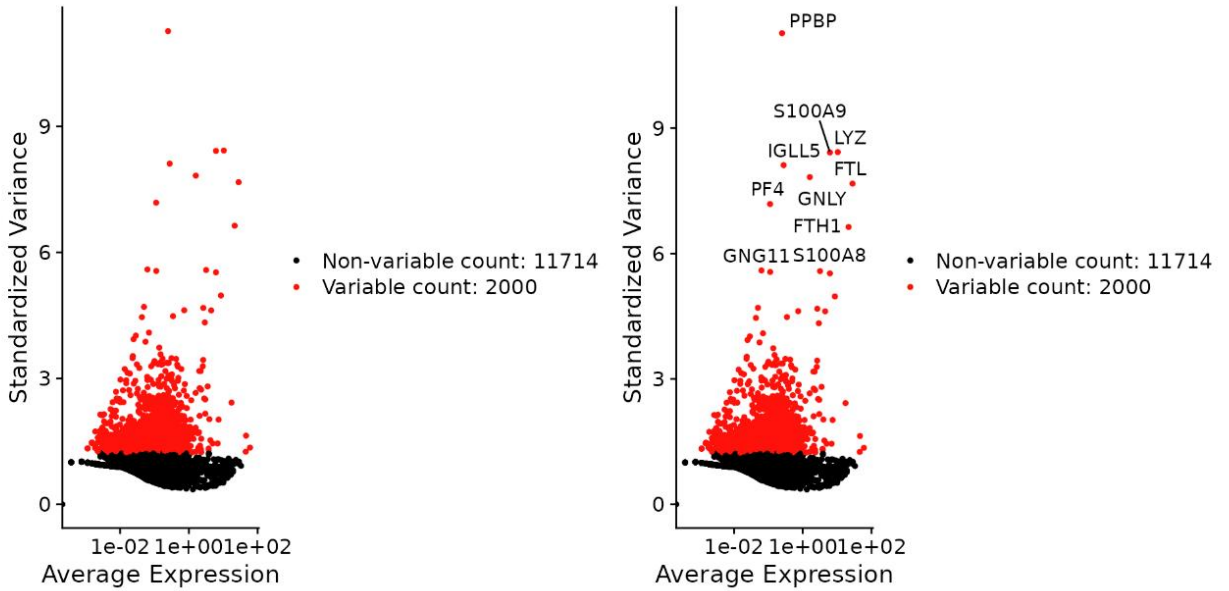
02 scRNA-seq Data analyses

Contents


```
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)

# Identify the 10 most highly variable genes
top10 <- head(VariableFeatures(pbmc), 10)

# plot variable features with and without labels
plot1 <- VariableFeaturePlot(pbmc)
plot2 <- LabelPoints(plot = plot1, points = top10, repel = TRUE)
plot1 + plot2
```



https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents

- Shifts the expression of each gene, so that the mean expression across cells is 0
- Scales the expression of each gene, so that the variance across cells is 1
    - This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate
- The results of this are stored in `pbmc[["RNA"]]$scale.data`
- By default, only variable features are scaled.
- You can specify the `features` argument to scale additional features

```
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
```

https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
```

Seurat provides several useful ways of visualizing both cells and features that define the PCA, including `VizDimReduction()`, `DimPlot()`, and `DimHeatmap()`

```
# Examine and visualize PCA results a few different ways
print(pbmc[["pca"]], dims = 1:5, nfeatures = 5)
```

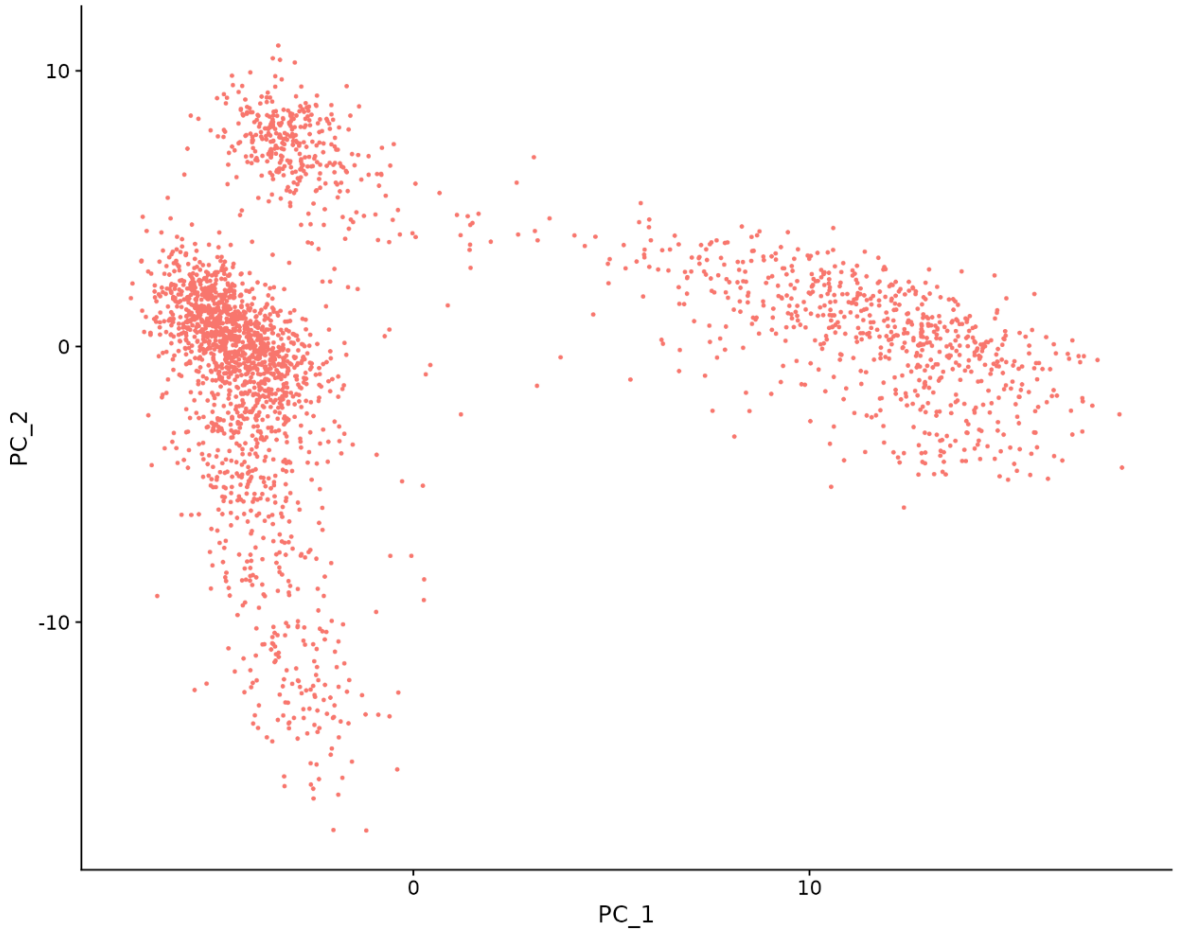
```
## PC_ 1
## Positive:  CST3, TYROBP, LST1, AIF1, FTL
## Negative:  MALAT1, LTB, IL32, IL7R, CD2
## PC_ 2
## Positive:  CD79A, MS4A1, TCL1A, HLA-DQA1, HLA-DQB1
## Negative:  NKG7, PRF1, CST7, GZMB, GZMA
## PC_ 3
## Positive:  HLA-DQA1, CD79A, CD79B, HLA-DQB1, HLA-DPB1
## Negative:  PPBP, PF4, SDPR, SPARC, GNG11
## PC_ 4
## Positive:  HLA-DQA1, CD79B, CD79A, MS4A1, HLA-DQB1
## Negative:  VIM, IL7R, S100A6, IL32, S100A8
## PC_ 5
## Positive:  GZMB, NKG7, S100A8, FGFBP2, GNLY
## Negative:  LTB, IL7R, CKB, VIM, MS4A7
```

https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
DimPlot(pbmc, reduction = "pca") + NoLegend()
```

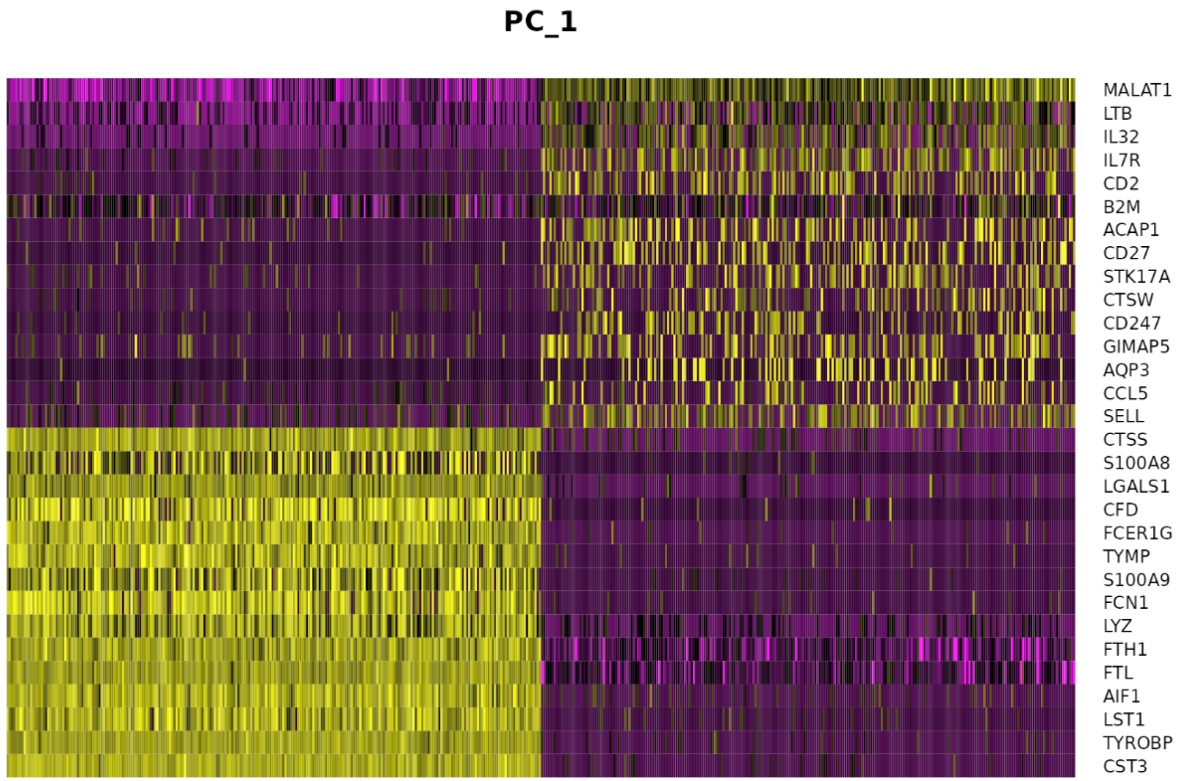


https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
DimHeatmap(pbmc, dims = 1, cells = 500, balanced = TRUE)
```



https://satijalab.org/seurat/

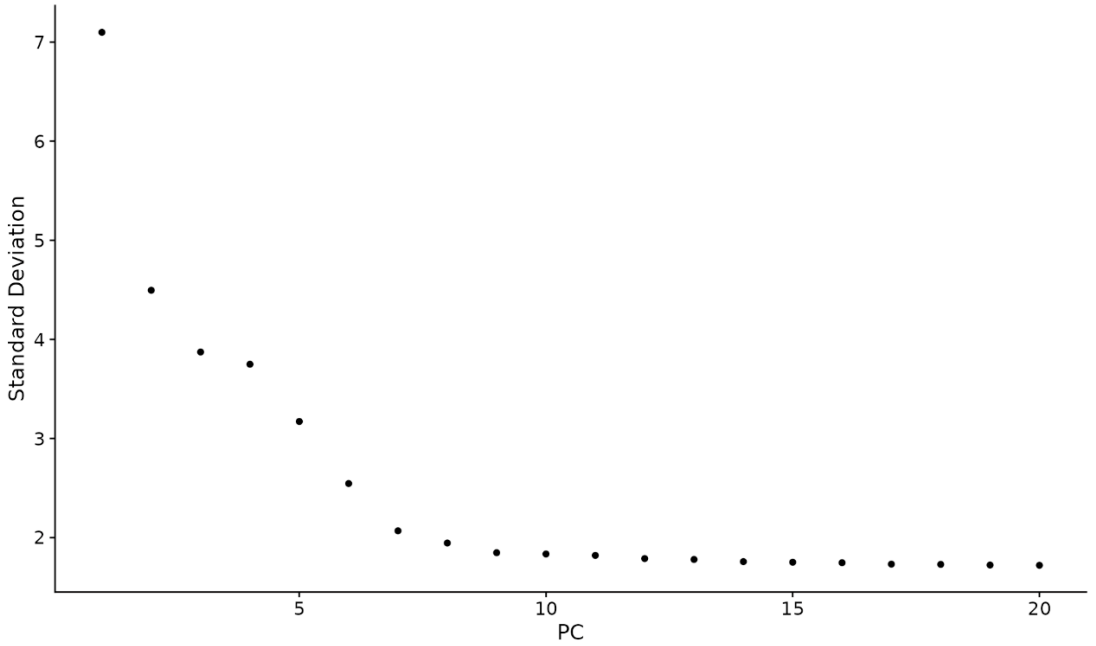
02 scRNA-seq Data analyses

Contents

To overcome the extensive technical noise in any single feature for scRNA-seq data, Seurat clusters cells based on their PCA scores, with each PC essentially representing a 'metafeature' that combines information across a correlated feature set. The top principal components therefore represent a robust compression of the dataset. However, how many components should we choose to include? 10? 20? 100?

An alternative heuristic method generates an 'Elbow plot': a ranking of principle components based on the percentage of variance explained by each one (`ElbowPlot()` function). In this example, we can observe an 'elbow' around PC9-10, suggesting that the majority of true signal is captured in the first 10 PCs.

```
ElbowPlot(pbmc)
```



https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents

This step is performed using the FindNeighbors() function, and takes as input the previously defined dimensionality of the dataset (first 10 PCs).

```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

```
## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 2638
## Number of edges: 95965
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.8723
## Number of communities: 9
## Elapsed time: 0 seconds
```

```
# Look at cluster IDs of the first 5 cells
head(Idents(pbmc), 5)
```

```
## AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1
##                2                3                2                1
## AAACCGTGTATGCG-1
##                6
## Levels: 0 1 2 3 4 5 6 7 8
```

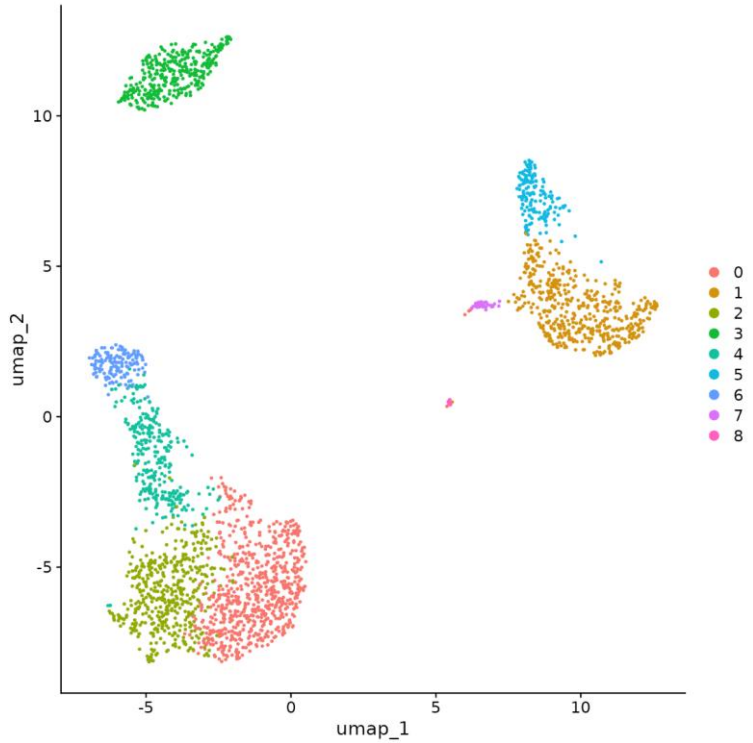
https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
pbmc <- RunUMAP(pbmc, dims = 1:10)
```

```
# note that you can set `label = TRUE` or use the LabelClusters function to help label
# individual clusters
DimPlot(pbmc, reduction = "umap")
```



https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
# find all markers of cluster 2
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2)
head(cluster2.markers, n = 5)
```

```
##              p_val avg_log2FC pct.1 pct.2    p_val_adj
## IL32 2.593535e-91  1.3221171 0.949 0.466 3.556774e-87
## LTB  7.994465e-87  1.3450377 0.981 0.644 1.096361e-82
## CD3D 3.922451e-70  1.0562099 0.922 0.433 5.379250e-66
## IL7R 1.130870e-66  1.4256944 0.748 0.327 1.550876e-62
## LDHB 4.082189e-65  0.9765875 0.953 0.614 5.598314e-61
```

```
# find all markers distinguishing cluster 5 from clusters 0 and 3
cluster5.markers <- FindMarkers(pbmc, ident.1 = 5, ident.2 = c(0, 3))
head(cluster5.markers, n = 5)
```

```
##                       p_val avg_log2FC pct.1 pct.2     p_val_adj
## FCGR3A        2.150929e-209   6.832372 0.975 0.039 2.949784e-205
## IFITM3        6.103366e-199   6.181000 0.975 0.048 8.370156e-195
## CFD           8.891428e-198   6.052575 0.938 0.037 1.219370e-193
## CD68          2.374425e-194   5.493138 0.926 0.035 3.256286e-190
## RP11-290F20.3 9.308287e-191   6.335402 0.840 0.016 1.276538e-186
```

```
# find markers for every cluster compared to all remaining cells, report only the positive
# ones
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE)
pbmc.markers %>%
    group_by(cluster) %>%
    dplyr::filter(avg_log2FC > 1)
```

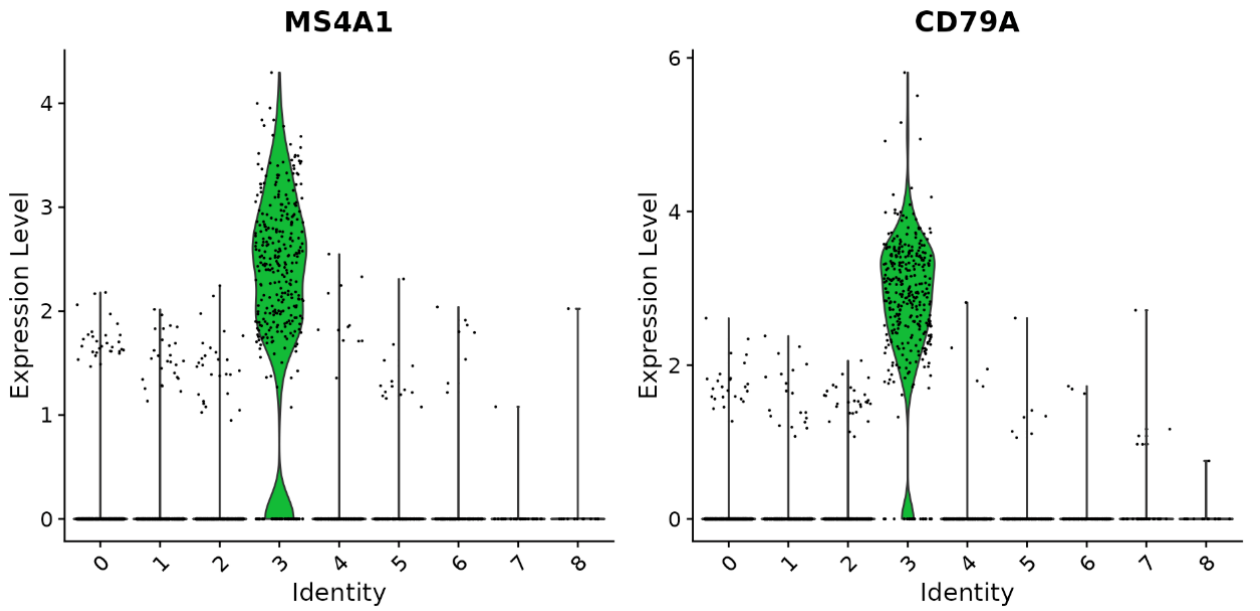
https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents

We include several tools for visualizing marker expression. `VlnPlot()` (shows expression probability distributions across clusters), and `FeaturePlot()` (visualizes feature expression on a tSNE or PCA plot) are our most commonly used visualizations. We also suggest exploring `RidgePlot()`, `CellScatter()`, and `DotPlot()` as additional methods to view your dataset.

```
VlnPlot(pbmc, features = c("MS4A1", "CD79A"))
```

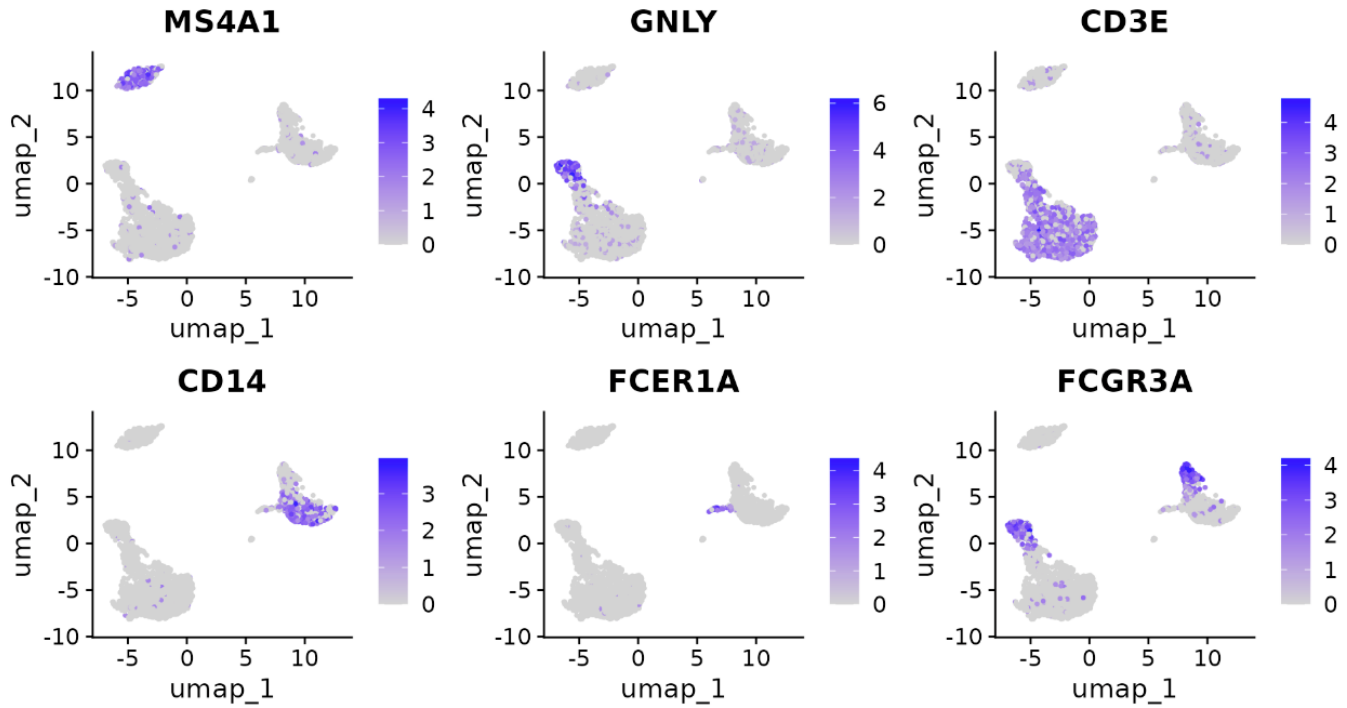


https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents


```
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP",
    "CD8A"))
```



https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents

- Setup the Seurat Object
- Standard pre-processing workflow
- Normalizing the data
- Identification of highly variable features (feature selection)
- Scaling the data
- Perform linear dimensional reduction
- Determine the 'dimensionality' of the dataset
- Cluster the cells
- Run non-linear dimensional reduction (UMAP/tSNE)
- Finding differentially expressed features (cluster biomarkers)
- Assigning cell type identity to clusters

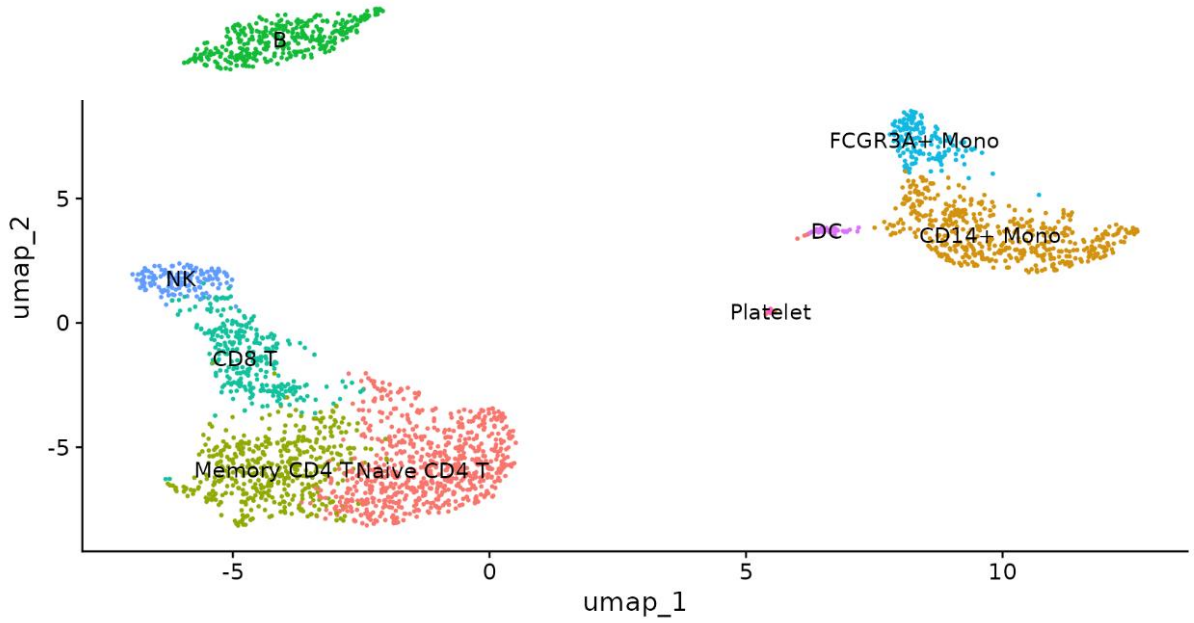
```
pbmc.markers %>%
    group_by(cluster) %>%
    dplyr::filter(avg_log2FC > 1) %>%
    slice_head(n = 10) %>%
    ungroup() -> top10
DoHeatmap(pbmc, features = top10$gene) + NoLegend()
```

https://satijalab.org/seurat/

02 scRNA-seq Data analyses

Contents

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

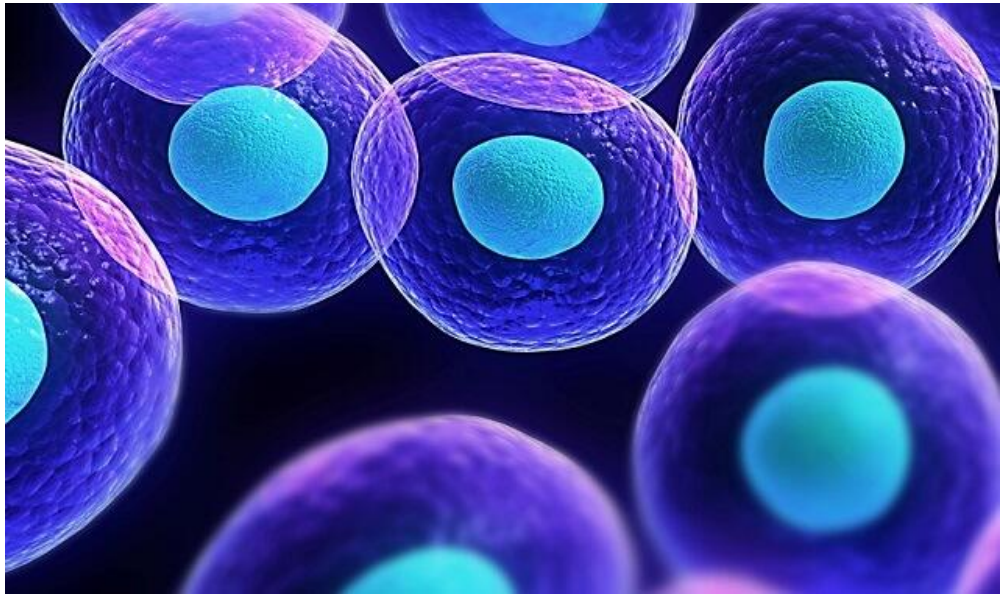


https://satijalab.org/seurat/

scRNA-seq

scRNA-seq (single-cell RNA-seq)

1. **Single cell RNA-seq techniques** (Tang protocol, Smart-seq2, Drop-seq, 10x genomimcs)
2. **Data analyses** (Seurat)
3. **Application of scRNA-seq** (Embryonic development, Viral infection, Immune, Aging)



03 Application of scRNAseq

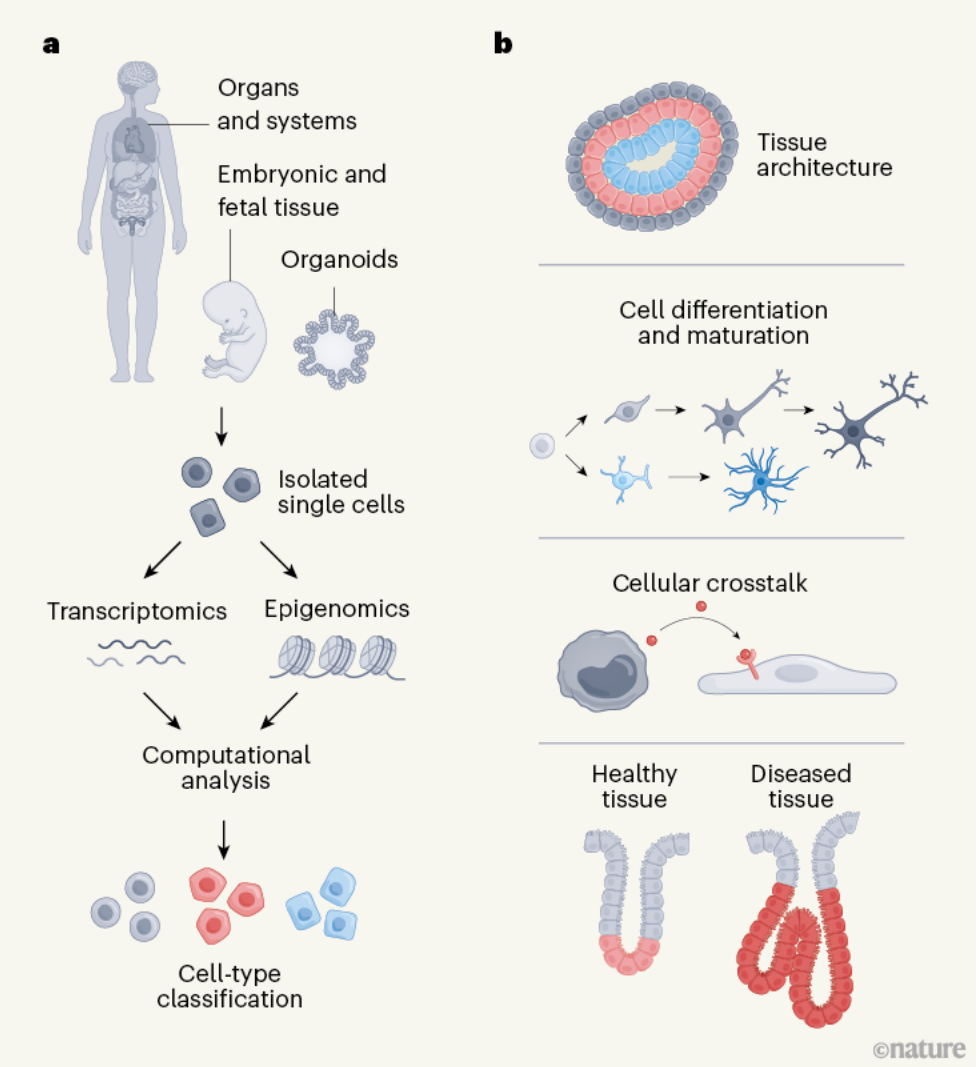
Application of scRNA-seq

THE HUMAN CELL ATLAS

人类细胞图谱（HCA）计划

HUMAN CELL ATLAS

The Human Cell Atlas is an international collaborative research consortium that is mapping all cell types in the healthy body, across time from development to adulthood, and eventually to old age. Creating this comprehensive reference map of human cells is transforming our understanding of health and disease, to drive major advances in healthcare and medicine worldwide.



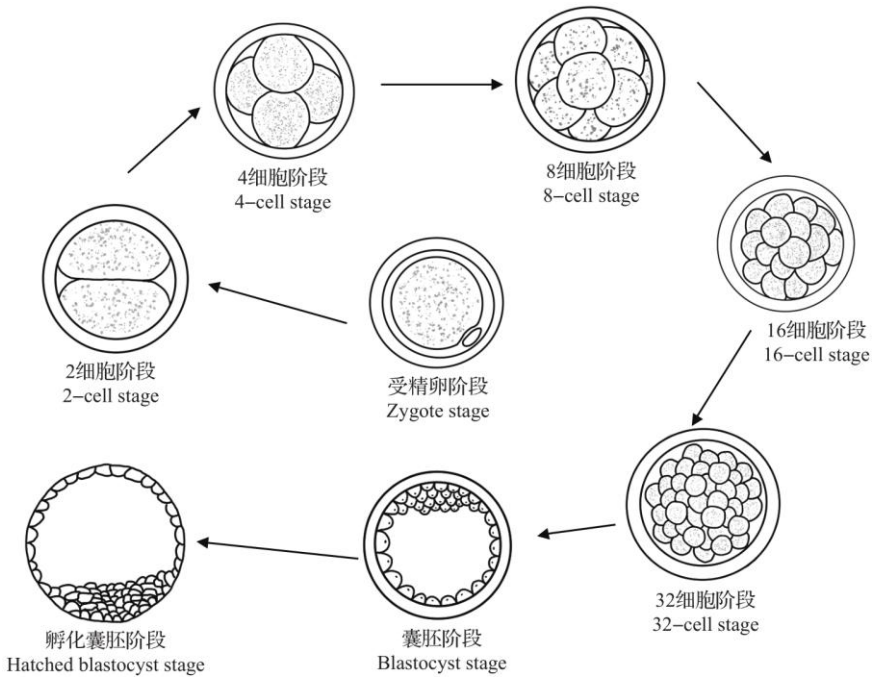
03 Application of scRNAseq

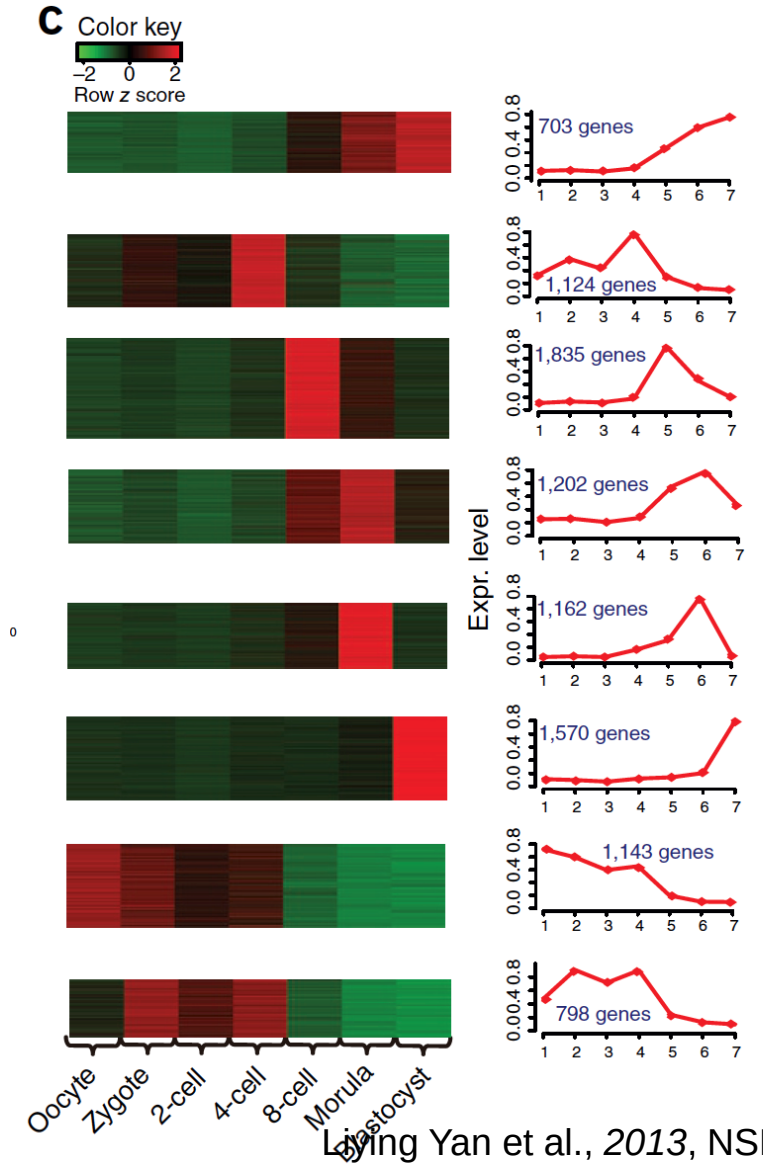
3.1 Embryonic development

RESOURCE

nature structural & molecular biology

Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells





Liying Yan et al., *2013*, NSMB

03 Application of scRNAseq

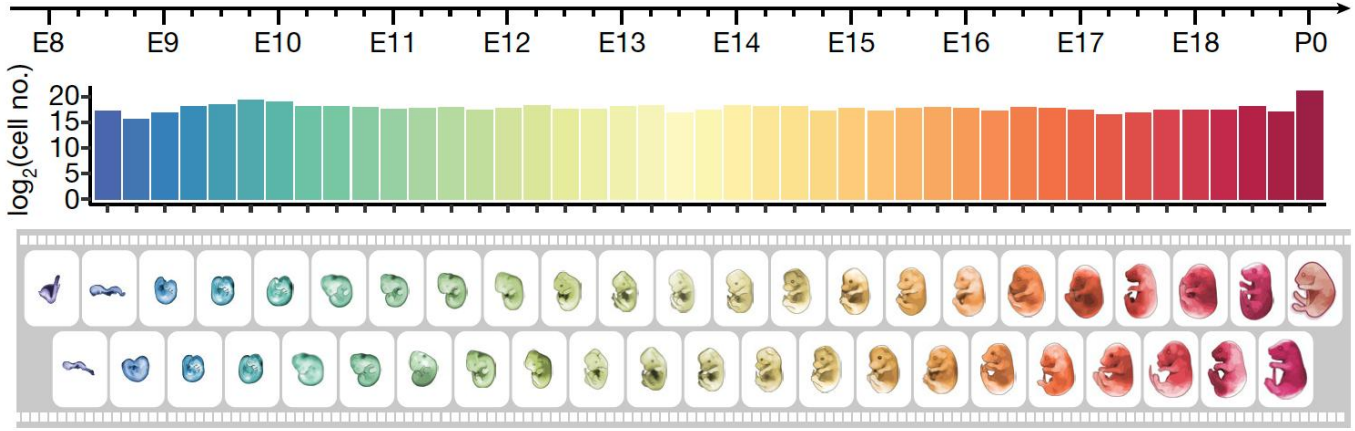
3.1 Embryonic development

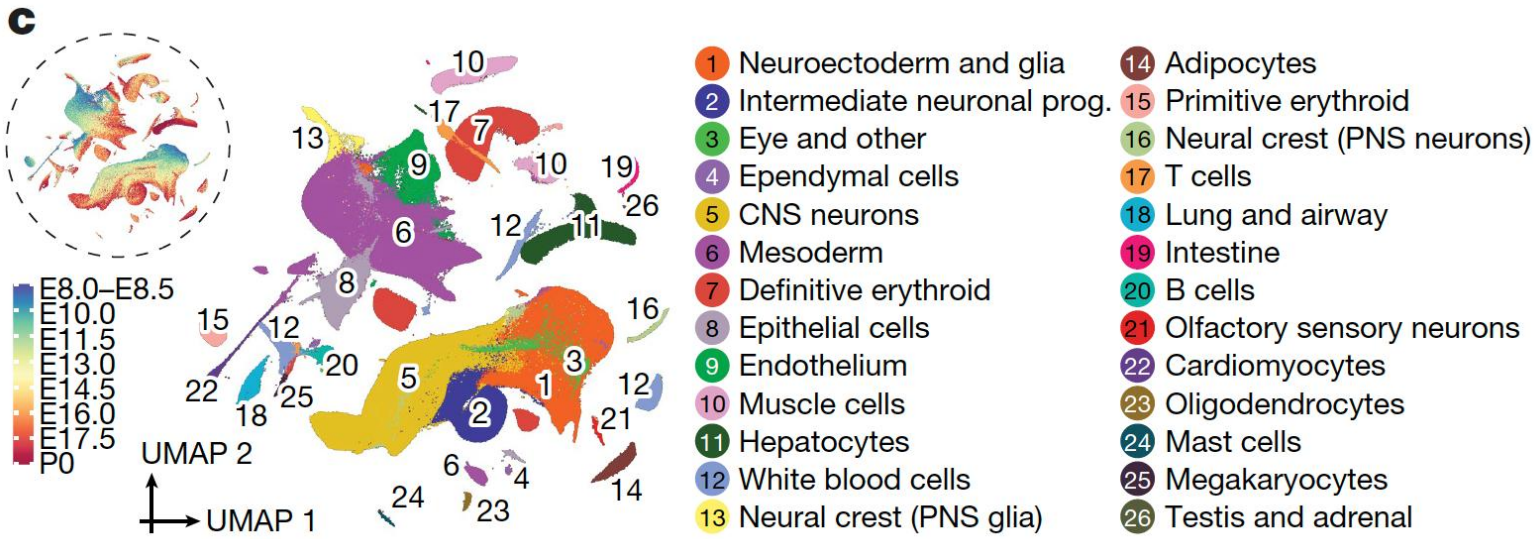
Article

A single-cell time-lapse of mouse prenatal development from gastrula to birth

https://doi.org/10.1038/s41586-024-07069-w
Received: 5 April 2023
Accepted: 15 January 2024
Published online: 14 February 2024
Open access

Chengxiang Qiu[1,18], Beth K. Martin[1,18], Ian C. Welsh[2,18], Riza M. Daza[1], Truc-Mai Le[3], Xingfan Huang[1,4], Eva K. Nichols[1], Megan L. Taylor[1,3], Olivia Fulton[1], Diana R. O'Day[3], Anne Roshella Gomes[3], Saskia Ilcisin[3], Sanjay Srivatsan[1,5], Xinxian Deng[6], Christine M. Disteche[6,7], William Stafford Noble[1,4], Nobuhiko Hamazaki[1,8], Cecilia B. Moens[9], David Kimelman[1,10], Junyue Cao[11], Alexander F. Schier[12,13], Malte Spielmann[14,15,16], Stephen A. Murray[2], Cole Trapnell[1,3,13,17] & Jay Shendure[1,3,8,13,17]





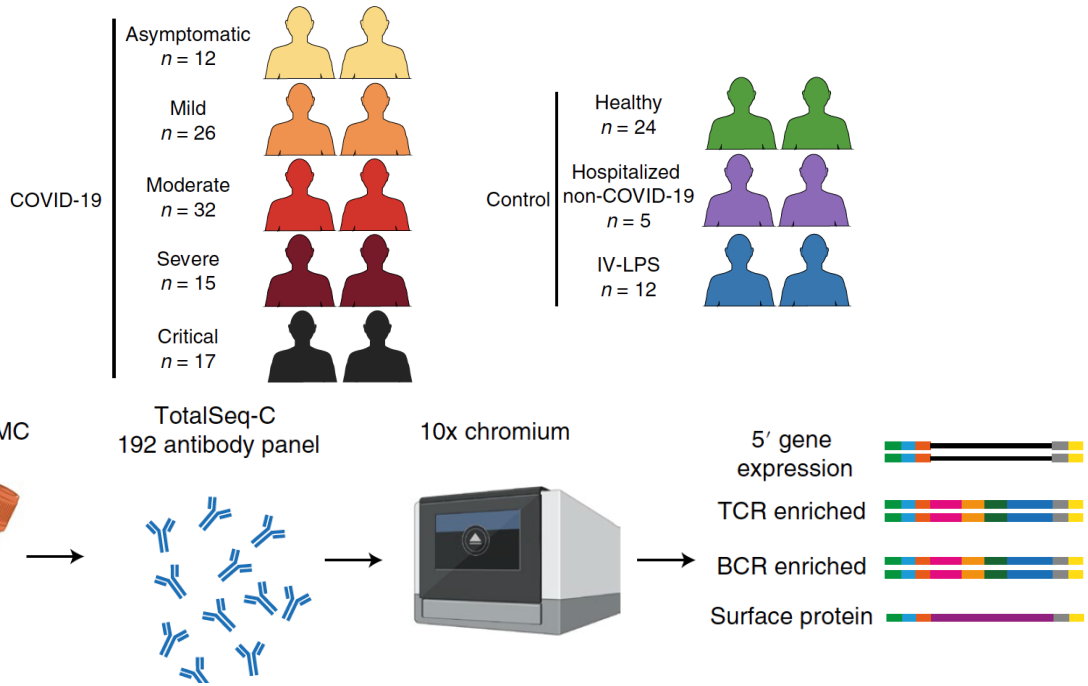
03 Application of scRNAseq

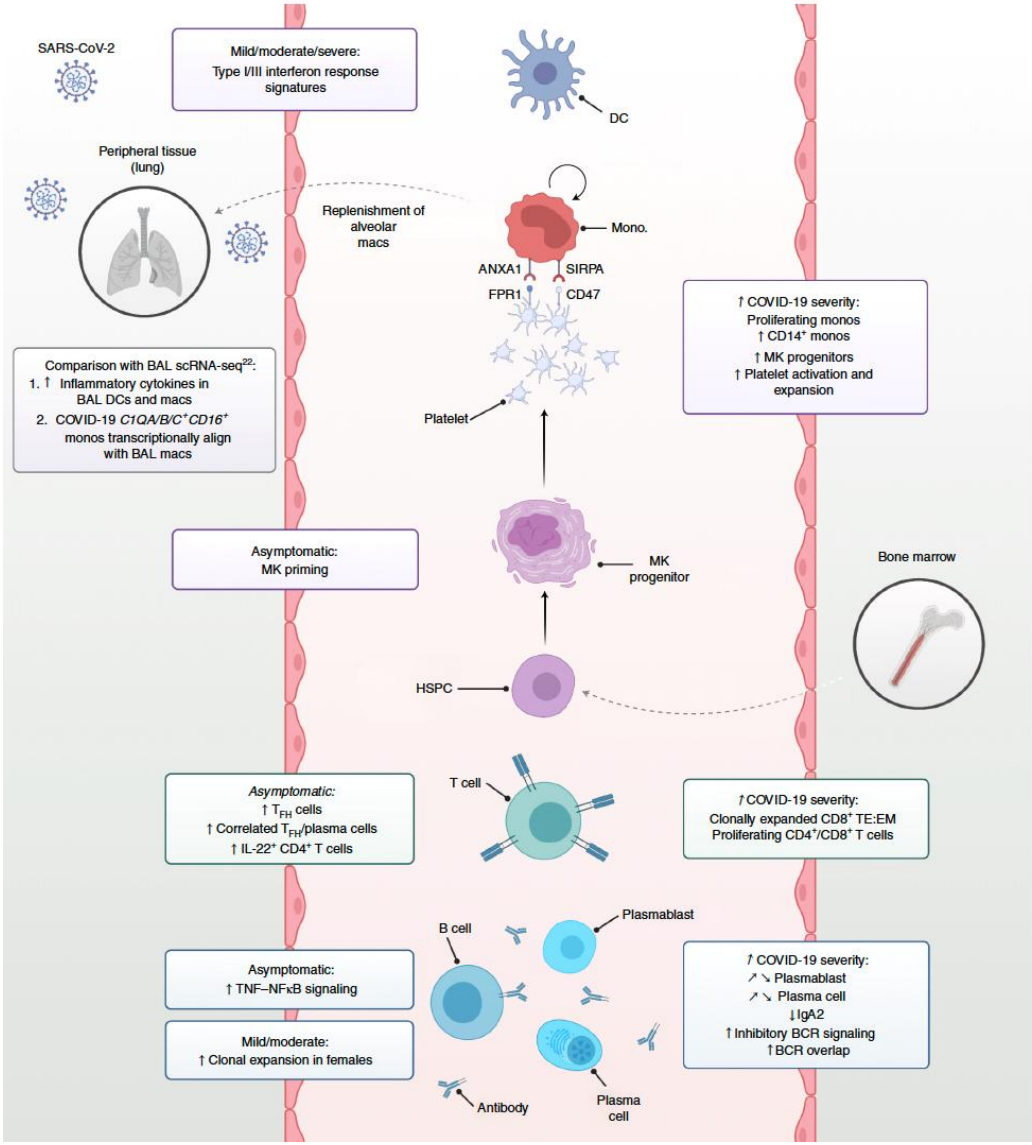
3.2 Viral infection

ARTICLES
https://doi.org/10.1038/s41591-021-01329-2
nature medicine
Check for updates

OPEN

Single-cell multi-omics analysis of the immune response in COVID-19



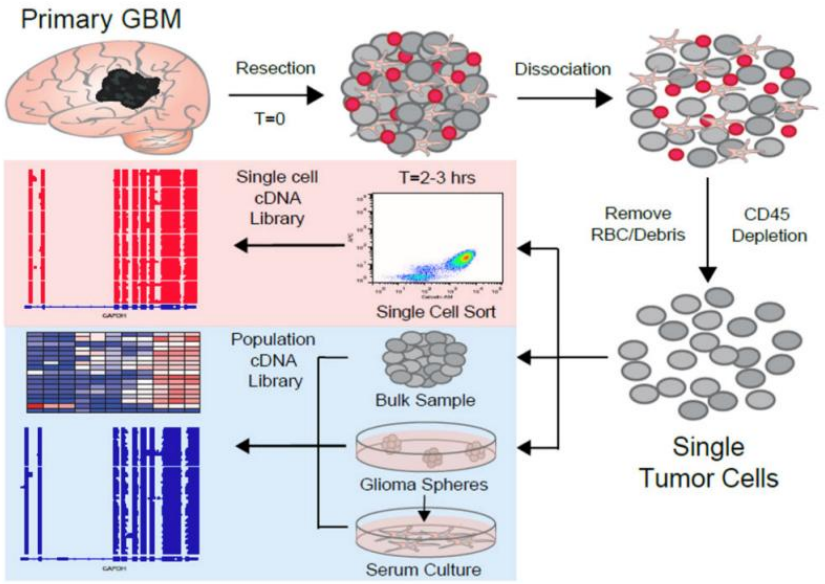


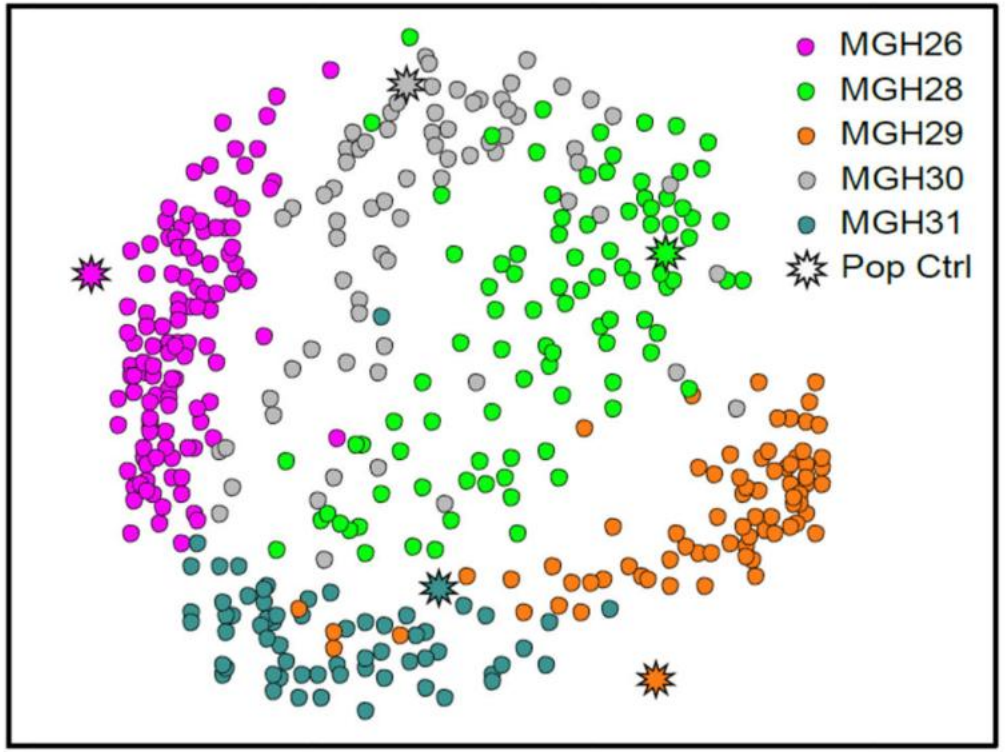
03 Application of scRNAseq

3.3 Intratumoral heterogeneity

Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma

Anoop P. Patel[1,2,3,4,†], **Itay Tirosh**[3,†], **John J. Trombetta**[3], **Alex K. Shalek**[3], **Shawn M. Gillespie**[2,3,4], **Hiroaki Wakimoto**[1], **Daniel P. Cahill**[1], **Brian V. Nahed**[1], **William T. Curry**[1], **Robert L. Martuza**[1], **David N. Louis**[2], **Orit Rozenblatt-Rosen**[3], **Mario L. Suvà**[2,3,*,‡], **Aviv Regev**[3,4,5,*,‡], and **Bradley E. Bernstein**[2,3,4,*,‡]





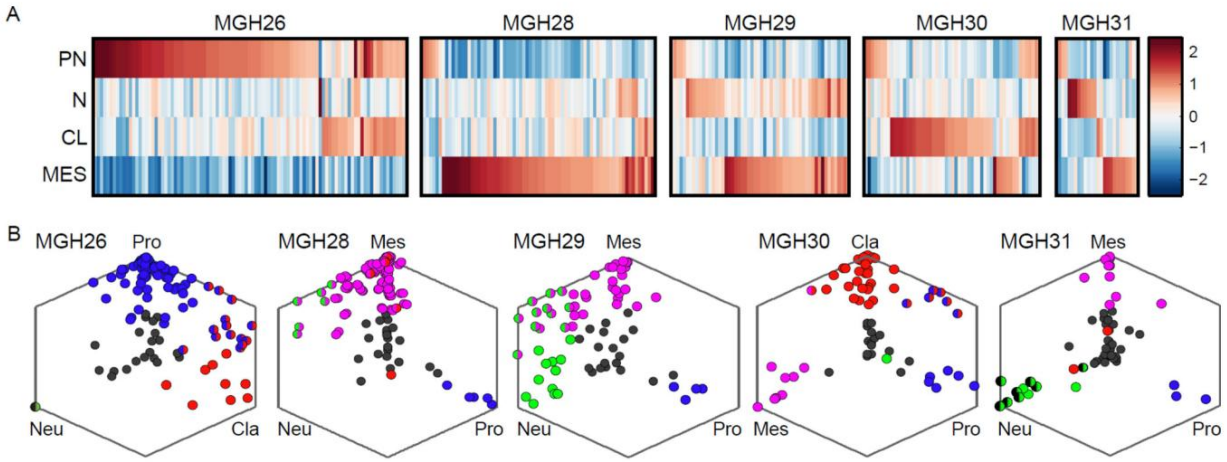
Patel et al., *2014*, Science

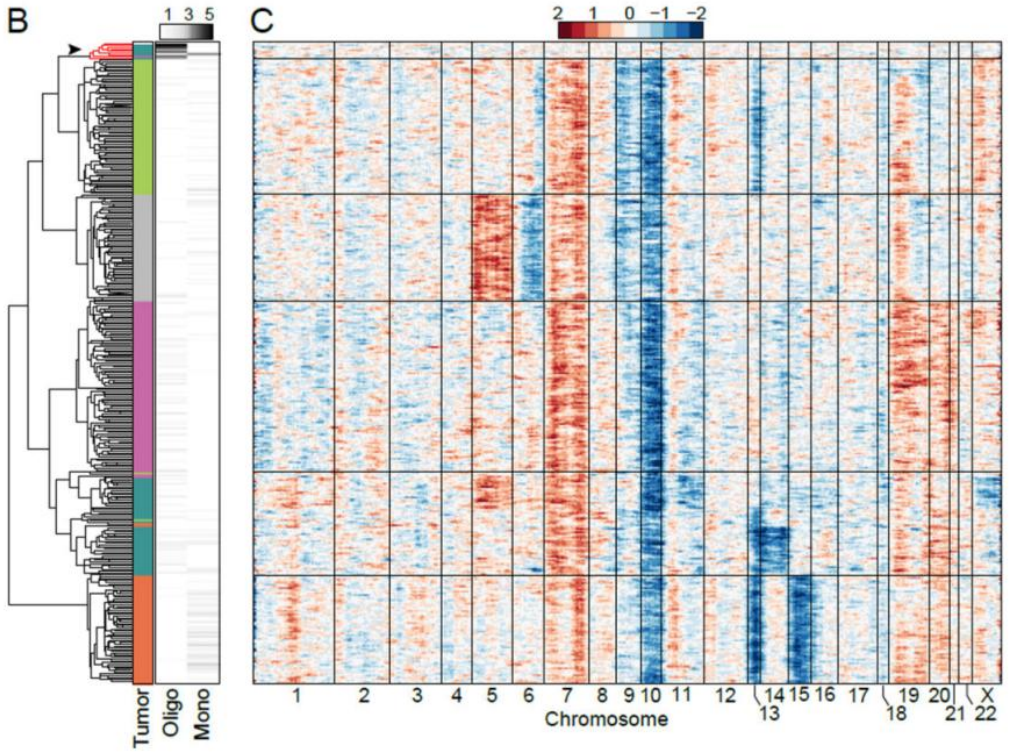
03 Application of scRNAseq

3.3 Intratumoral heterogeneity

Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma

Anoop P. Patel[1,2,3,4,†], **Itay Tirosh**[3,†], **John J. Trombetta**[3], **Alex K. Shalek**[3], **Shawn M. Gillespie**[2,3,4], **Hiroaki Wakimoto**[1], **Daniel P. Cahill**[1], **Brian V. Nahed**[1], **William T. Curry**[1], **Robert L. Martuza**[1], **David N. Louis**[2], **Orit Rozenblatt-Rosen**[3], **Mario L. Suvà**[2,3,*,‡], **Aviv Regev**[3,4,5,*,‡], and **Bradley E. Bernstein**[2,3,4,*,‡]





Patel et al., *2014*, Science

03 Application of scRNAseq

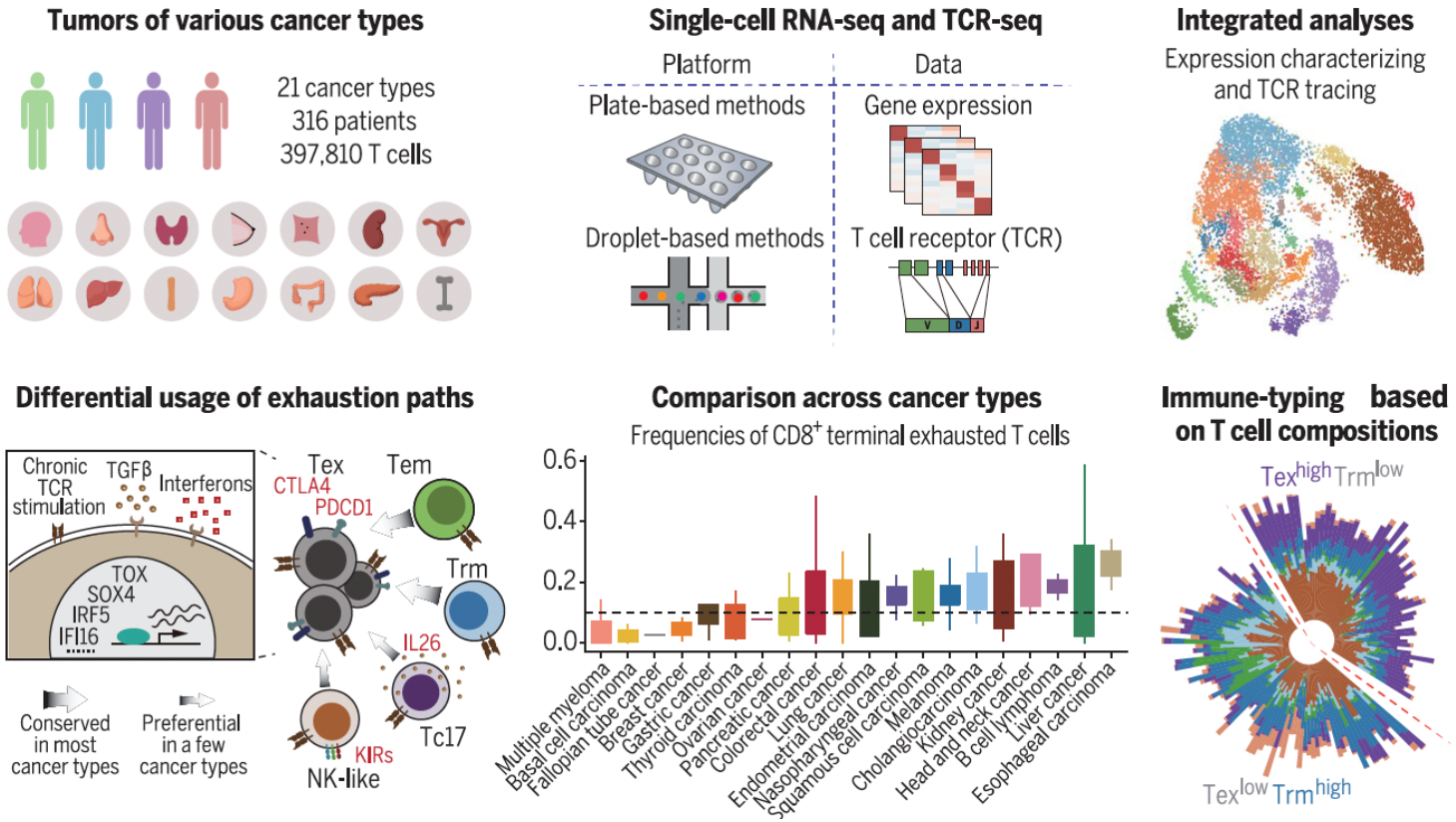
3.4 Immune

RESEARCH

RESEARCH ARTICLE

CANCER IMMUNOLOGY

Pan-cancer single-cell landscape of tumor-infiltrating T cells

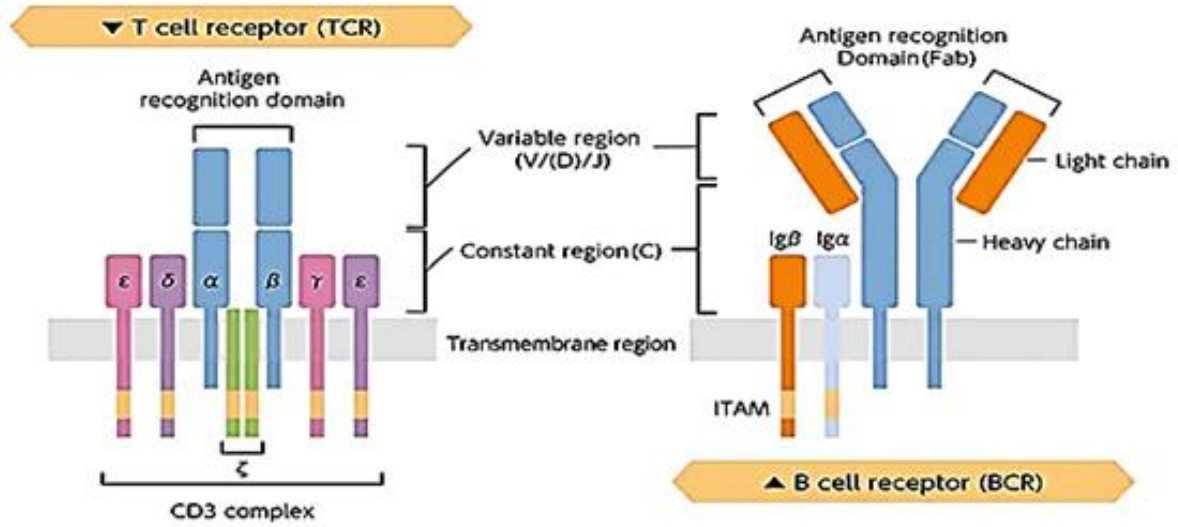


Liangtao Zheng et al., *2021*, Science

03 Application of scRNAseq

3.4 Immune

TCR and BCR are key components of adaptive immunity, enabling T and B cells to recognize antigens. Their diversity arises from **V(D)J recombination**, **somatic hypermutation (BCR only)**, and **clonal selection**. Below is a detailed comparison and analysis framework.

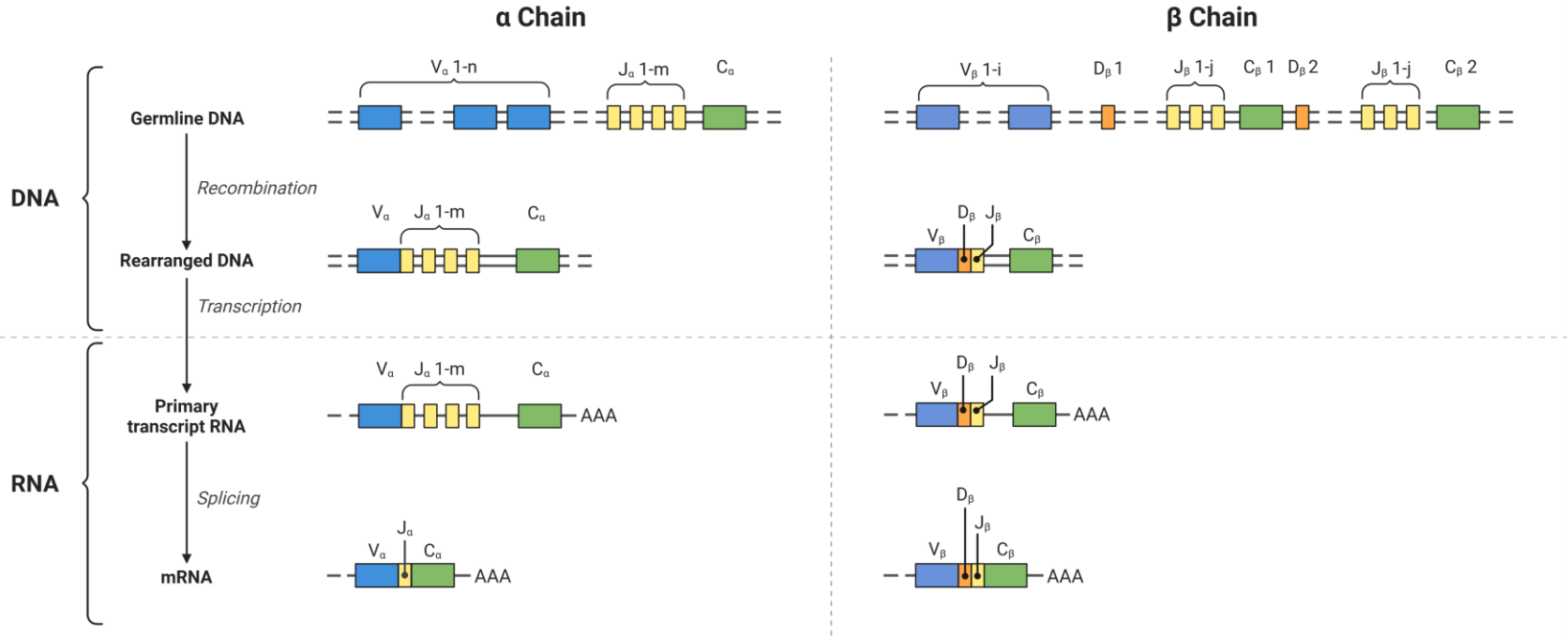


(Figure 1) The structure of TCR/BCR
TCR is composed of α chain and β chain, or δ chain and γ chain. Each TCR has variable region (V/D/J) and constant region (C). Variable region has three complementarity determining region (CDRs): CDR1 and CDR2 which recognize MHC, and CDR3 which recognizes and binds antigen. The TCR and CD3 molecules together form the TCR complex and generate the intracellular signals. BCR or immunoglobulin is composed of immunoglobulin-heavy chain (IgHC) and immunoglobulin-light chain (IgLC). There are five different isotypes for IgHC: IgA, IgD, IgE, IgG and IgM, while IgLC is classified into IgL and IgK. Fab region on heavy and light chains recognizes and binds antigen. Similar to TCR, BCR form the BCR complex with Igαβ for generating the intracellular signals.

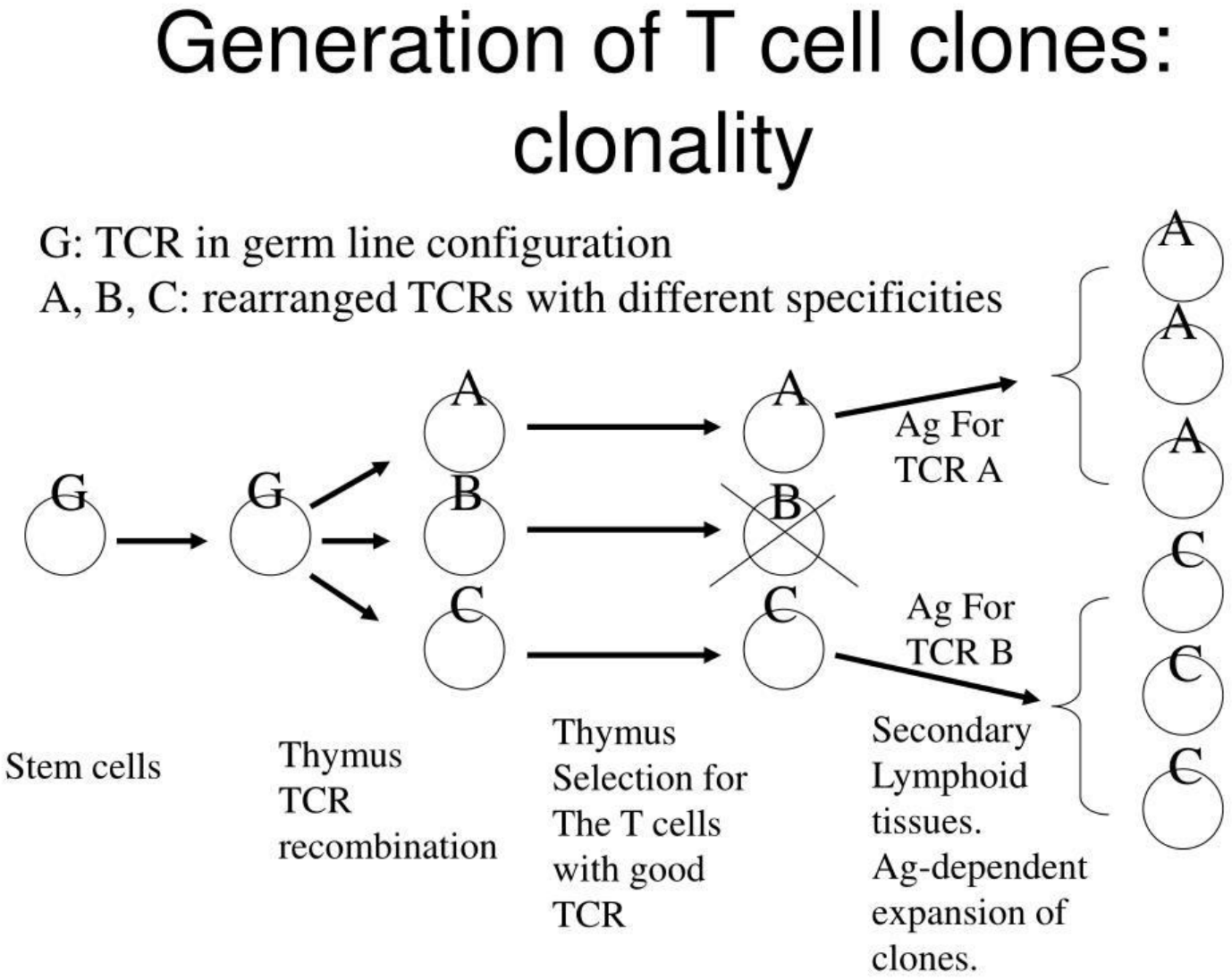
03 Application of scRNAseq

3.4 Immune

RNA sequencing (RNA-seq) can profile **T-cell receptor (TCR)** and **B-cell receptor (BCR)** repertoires, enabling insights into adaptive immune responses in cancer, autoimmune diseases, and infections.

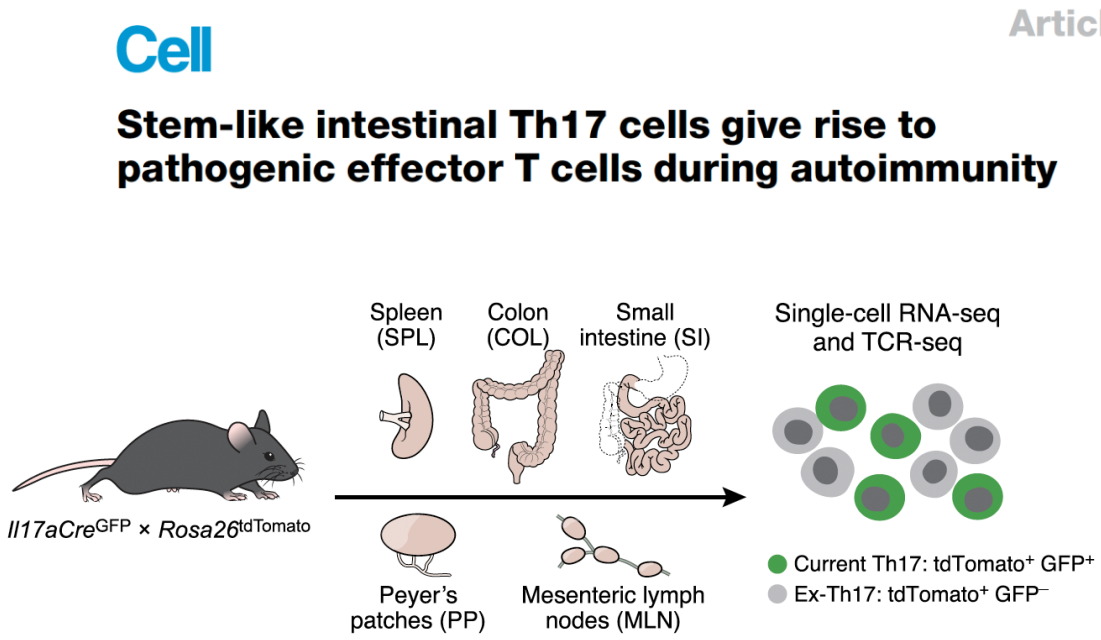


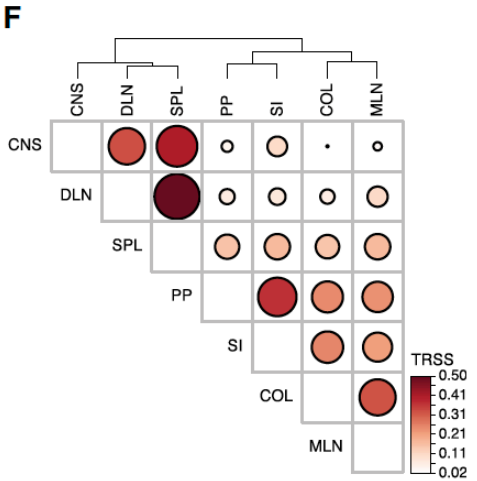
3.4 Immune

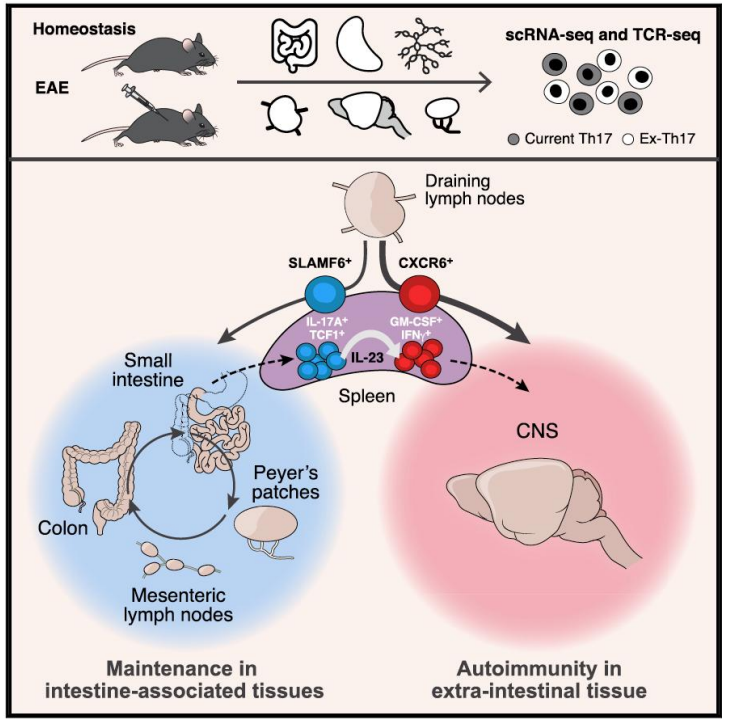

Generation of T cell clones: clonality
G: TCR in germ line configuration
A, B, C: rearranged TCRs with different specificities
G
G
A
B
C
A
B
C
Ag For TCR A
Ag For TCR B
A
A
A
C
C
C
Stem cells
Thymus TCR recombination
Thymus Selection for The T cells with good TCR
Secondary Lymphoid tissues. Ag-dependent expansion of clones.

03 Application of scRNAseq

3.4 Immune



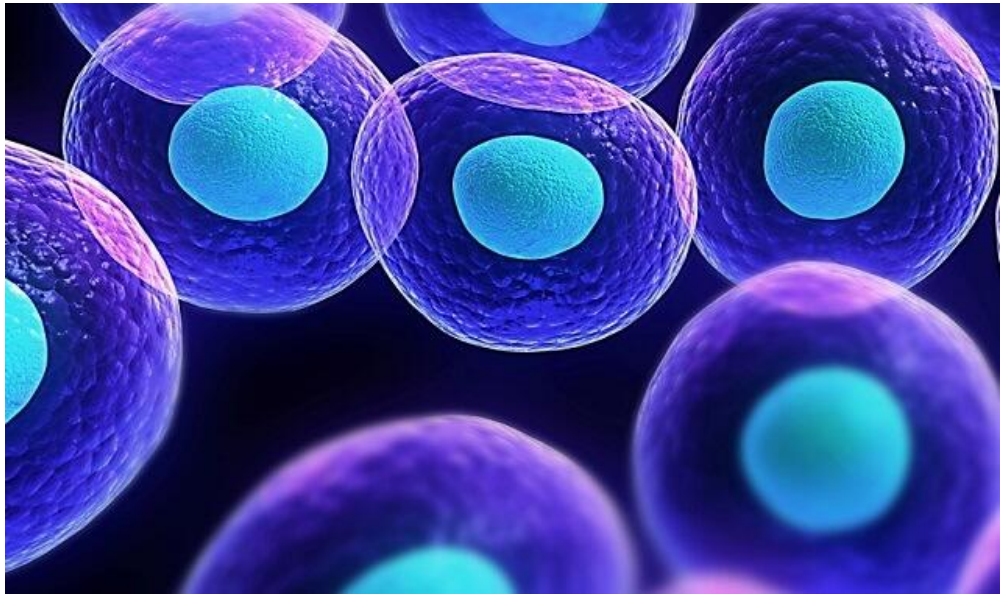




scRNA-seq

scRNA-seq (single-cell RNA-seq)

1. **Single cell RNA-seq techniques** (Tang protocol, Smart-seq2, Drop-seq, 10x genomimcs)
2. **Data analyses** (Seurat)
3. **Application of scRNA-seq** (Embryonic development, Viral infection, Immune)



Thank you

Yu Hou Zhejiang University 2025