

# Single Cell RNA Analysis

Biocore Bootcamp



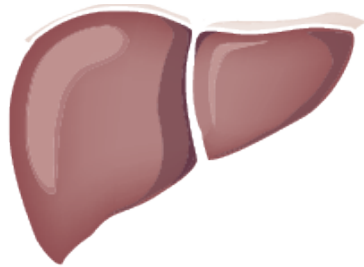
Bioinformatics Core

# Overview

- 1) Why single cell?**
- 2) inDrops Technology**
- 3) Dolphinnext pipelines**
- 4) Data Structures**

# Why Single Cell??

Solid Diverse Tissue



Minimal Prep



Isolate Cells



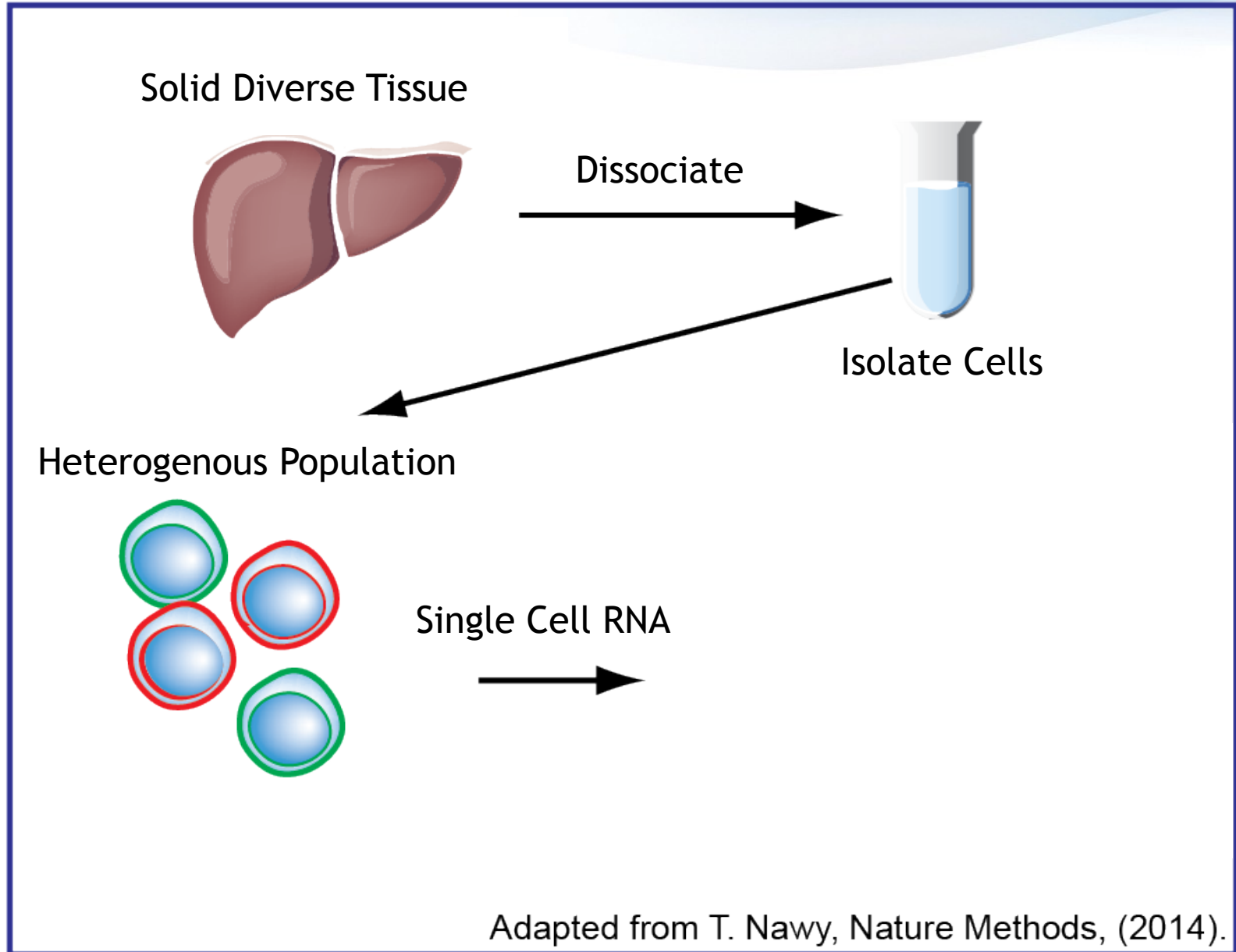
Bulk RNA Sequencing



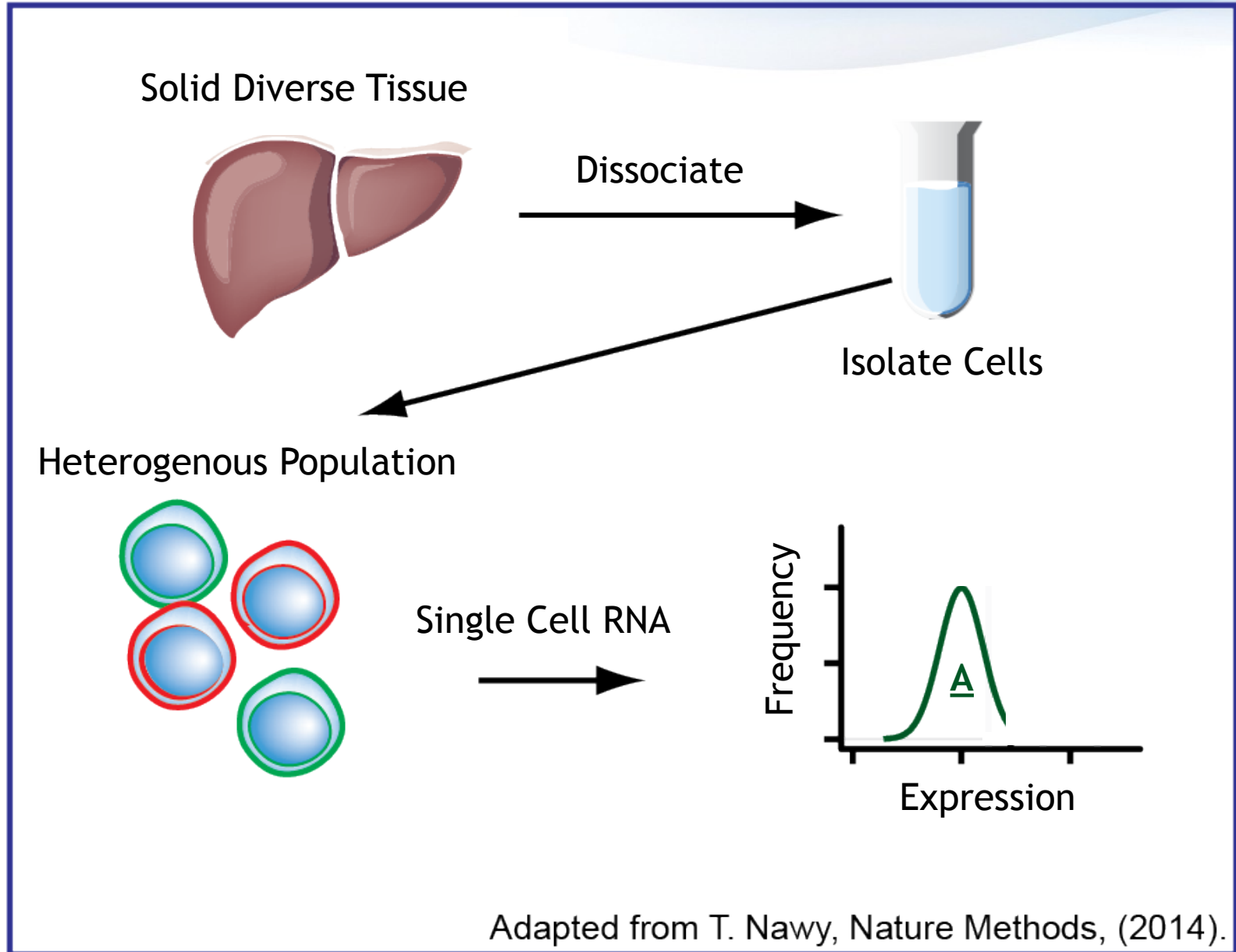
Single Gene Expression Value

Adapted from T. Nawy, Nature Methods, (2014).

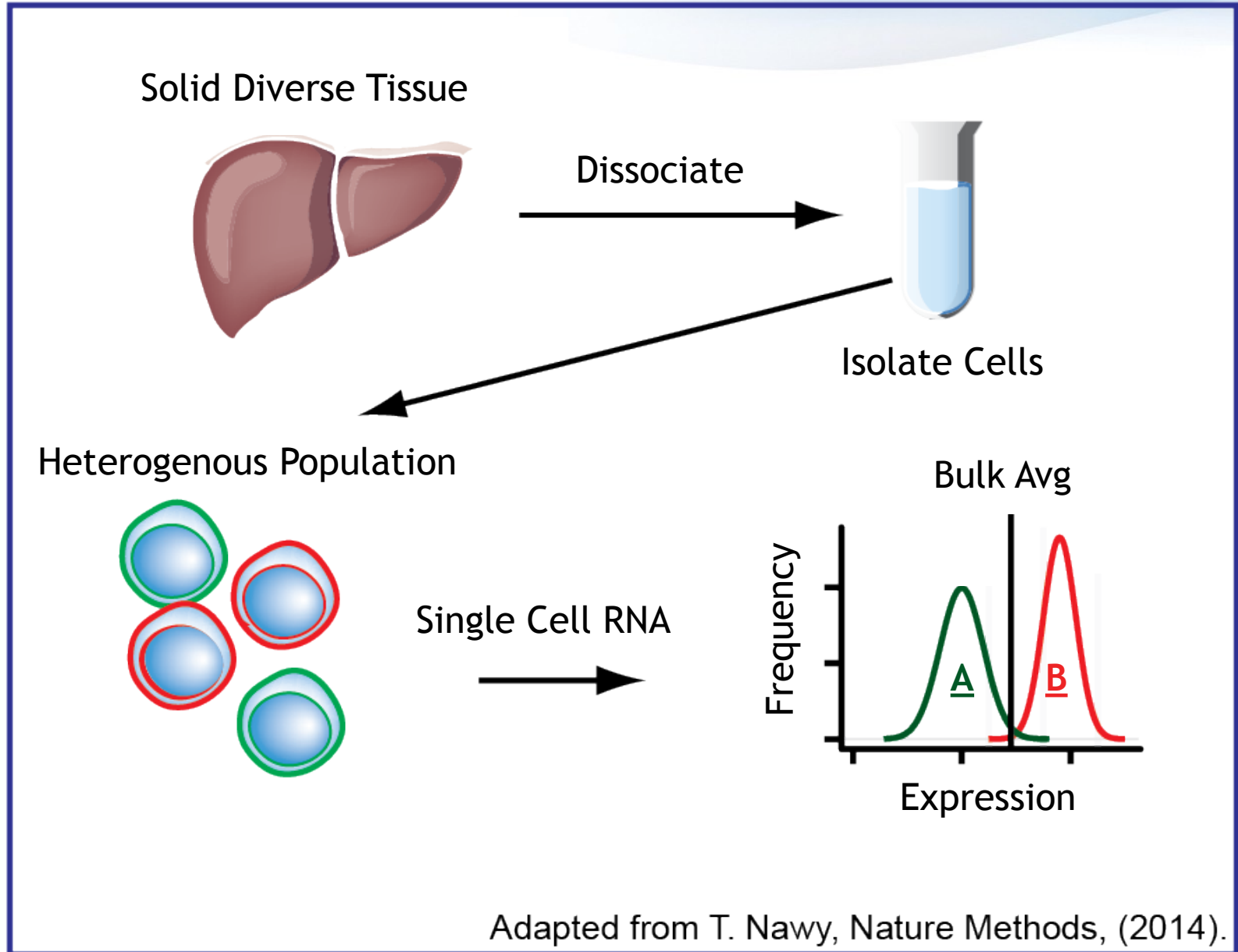
# Why Single Cell??



# Why Single Cell??



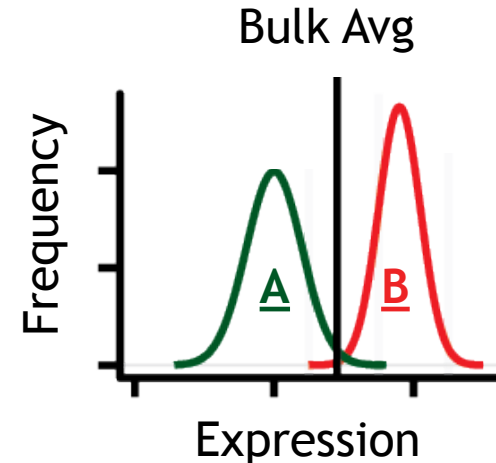
# Why Single Cell??



Adapted from T. Nawy, Nature Methods, (2014).

# Why Single Cell??

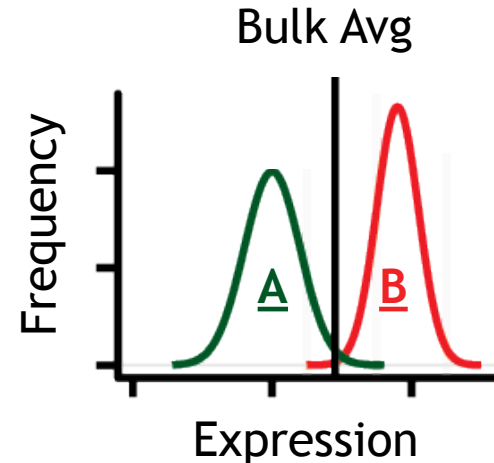
- Not all problems necessitate scRNA-seq
- It is well suited when populations are heterogeneous



Adapted from T. Nawy, Nature Methods, (2014).

# Why Single Cell??

- Not all problems necessitate scRNA-seq
- It is well suited when populations are heterogeneous
- It is a powerful tool for studying intra and inter cell type variations in gene expression
- Useful for unbiased discovery

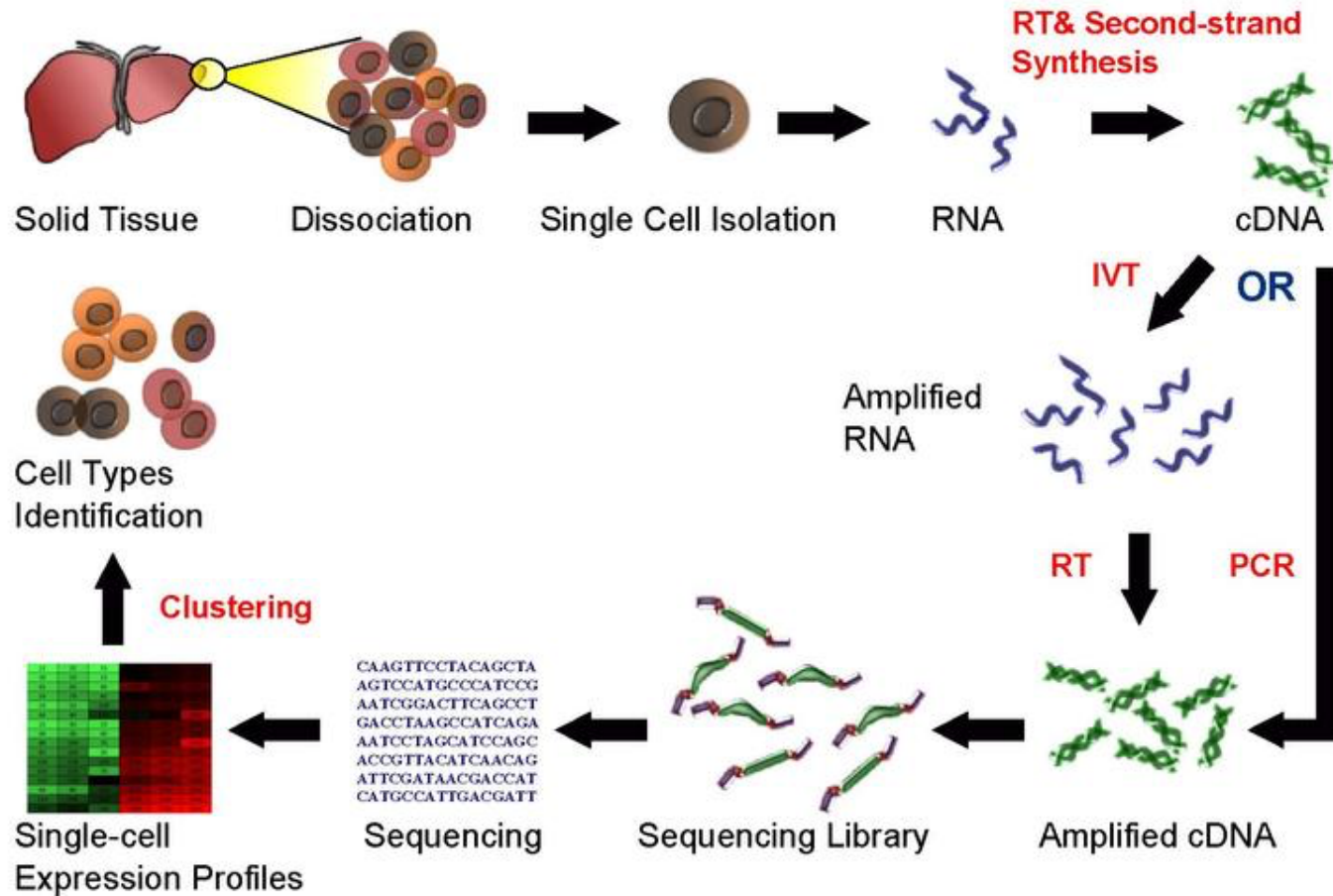


Adapted from T. Nawy, Nature Methods, (2014).



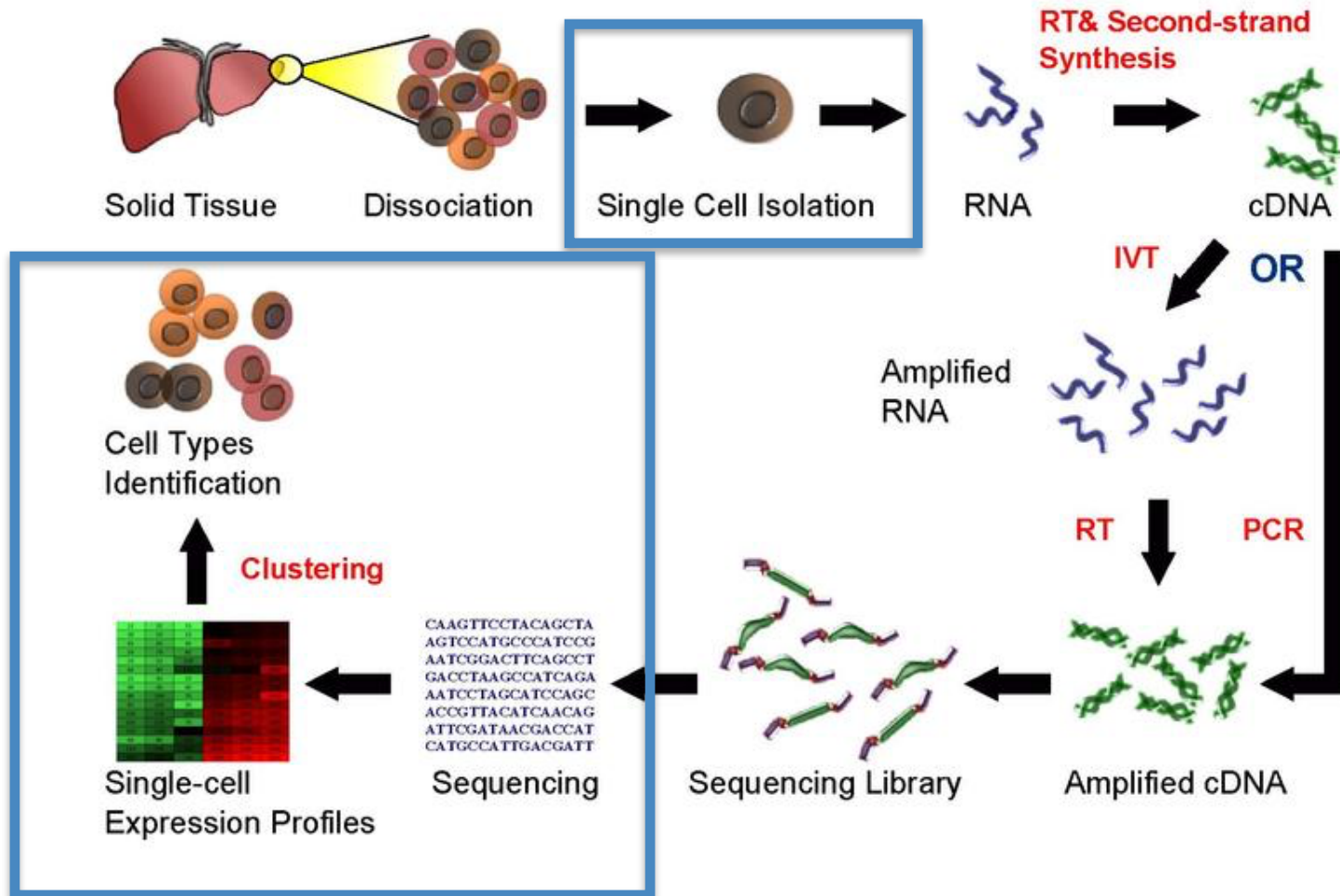
# Single Cell Workflows

## Single Cell RNA Sequencing Workflow



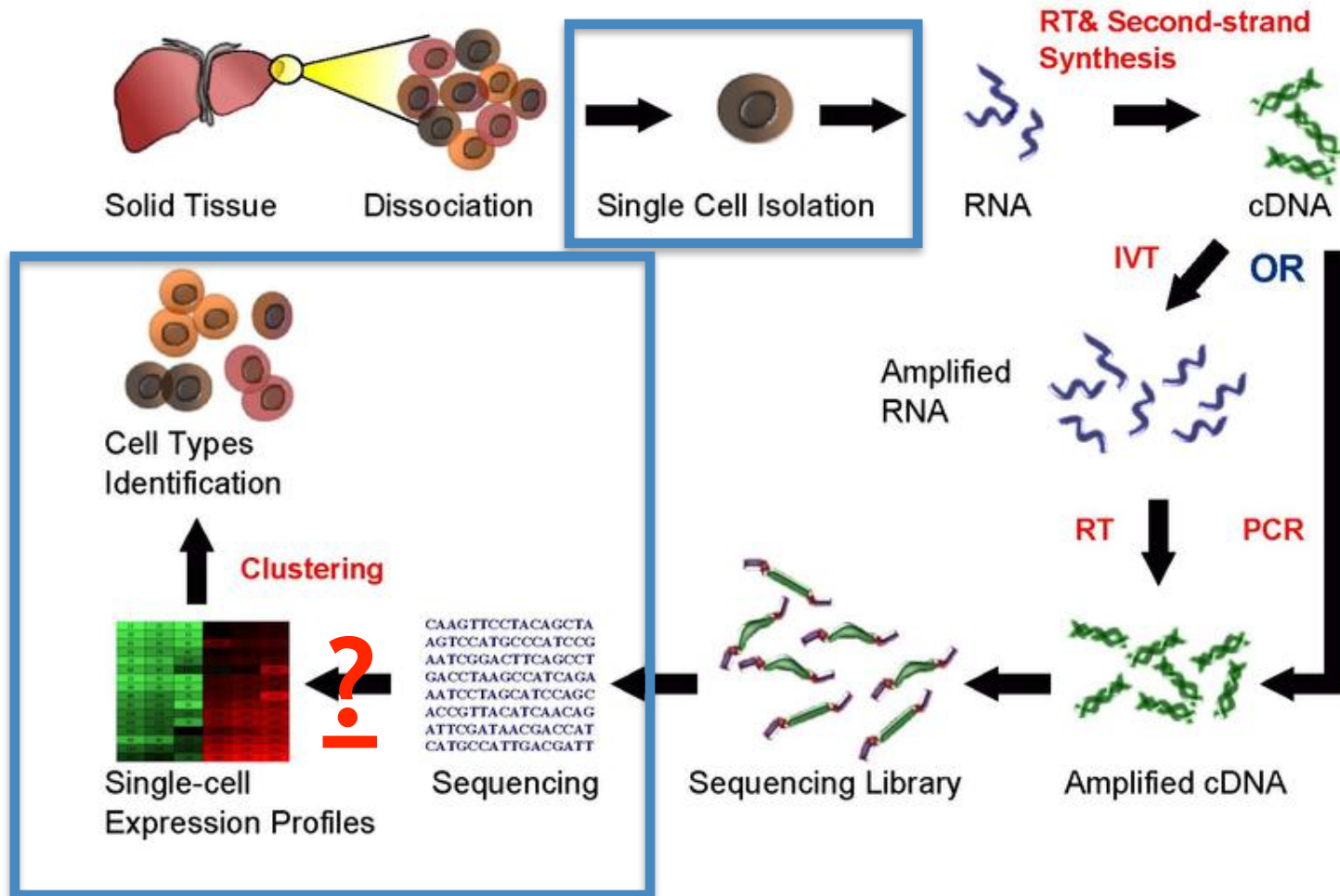
# Single Cell Workflows

## Single Cell RNA Sequencing Workflow

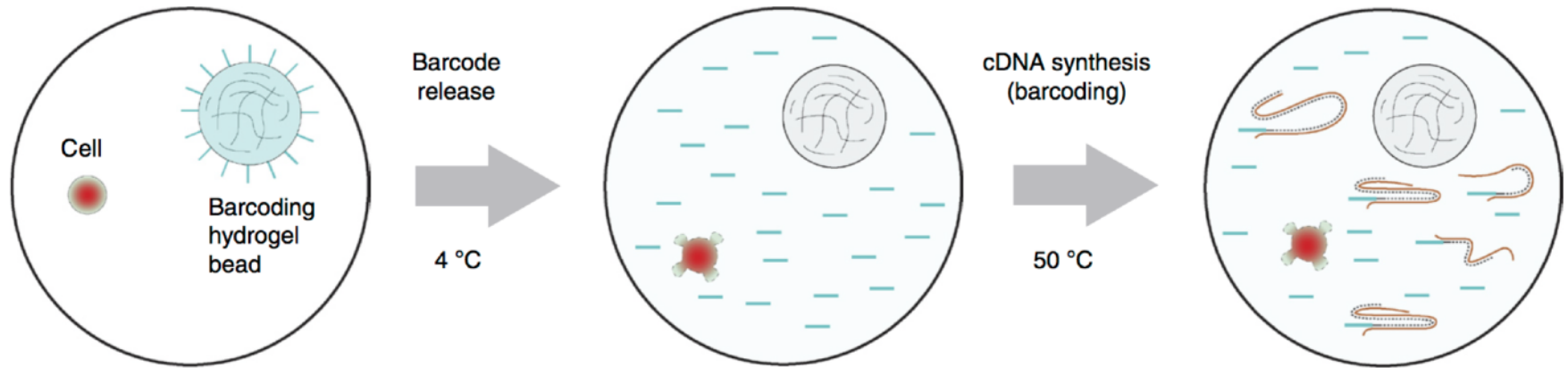


# Single Cell Workflows

## Single Cell RNA Sequencing Workflow

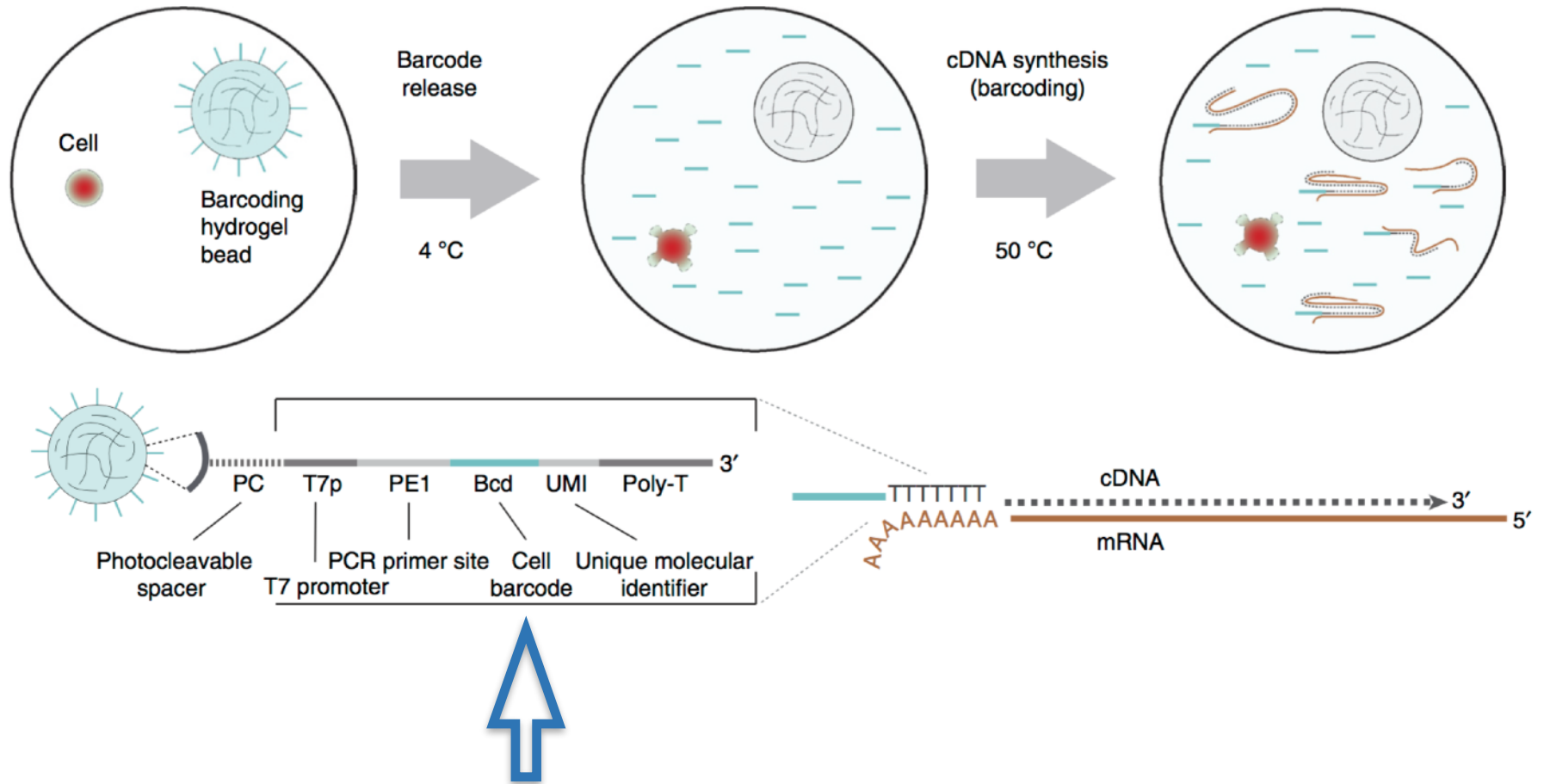


# inDrops Technology

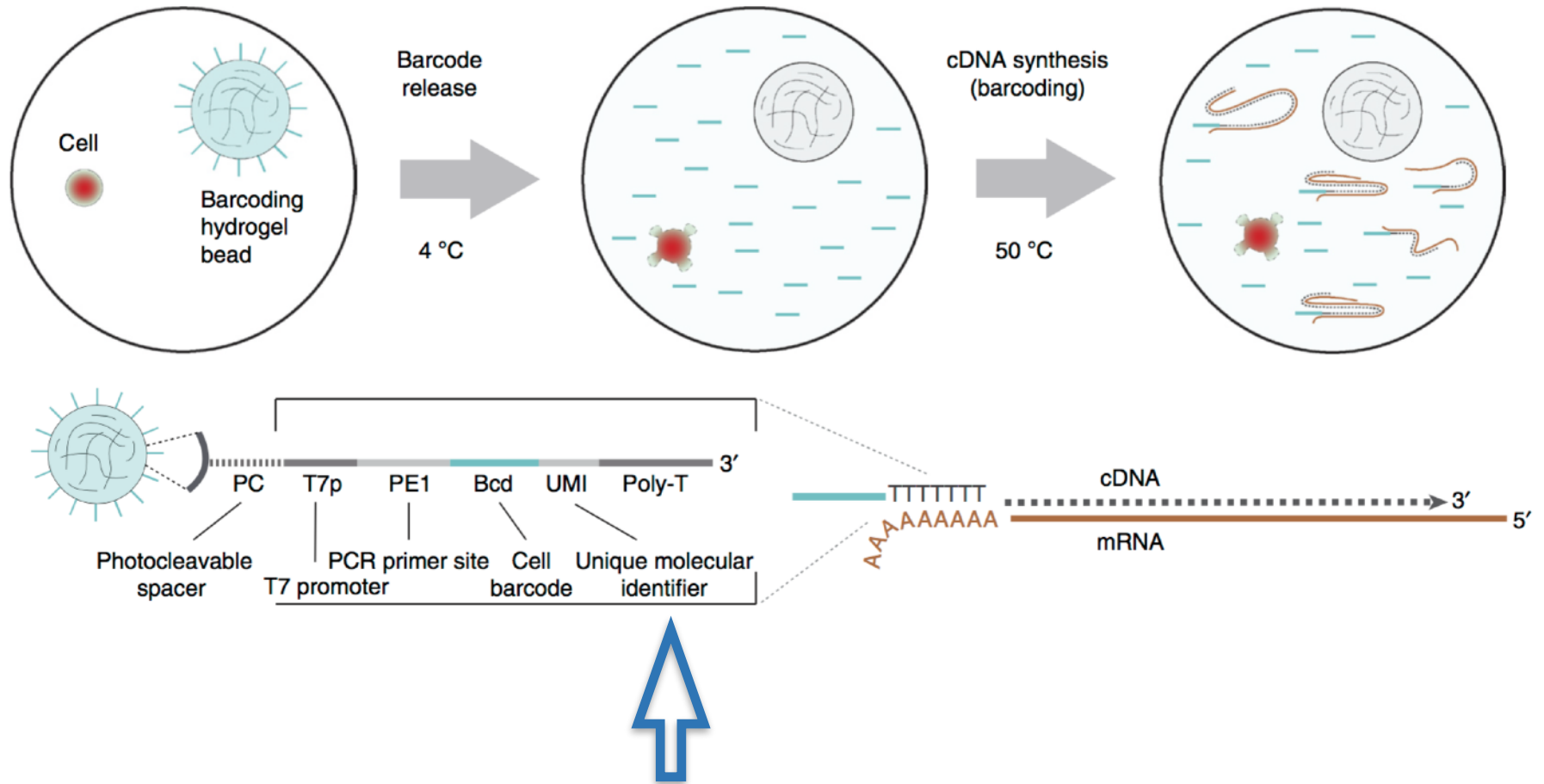


?

# inDrops Technology

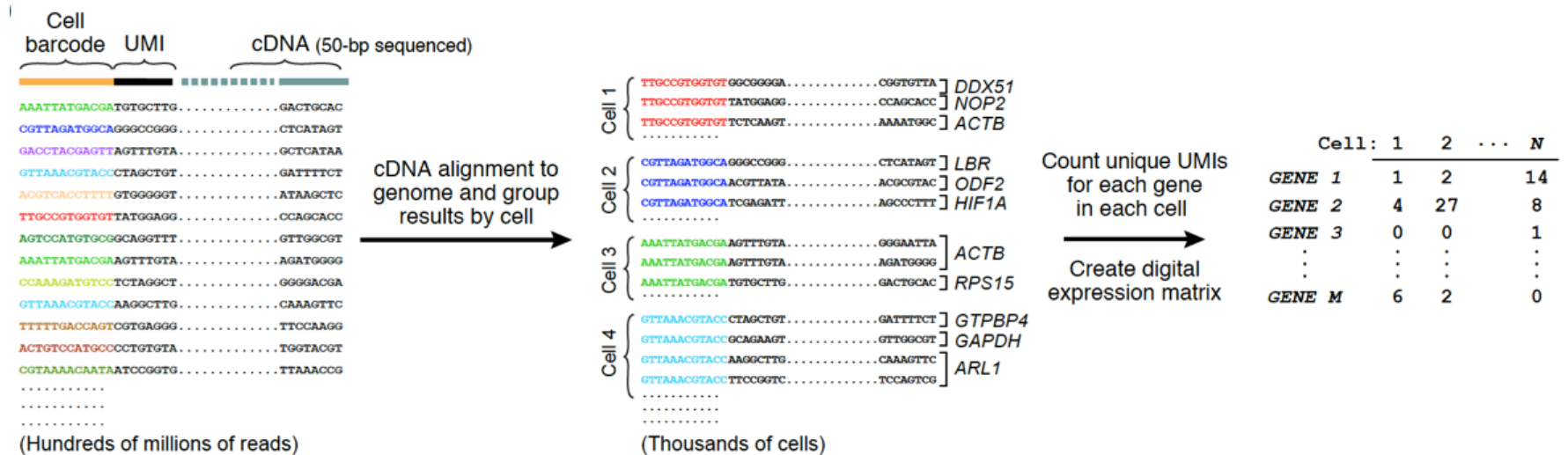


# inDrops Technology



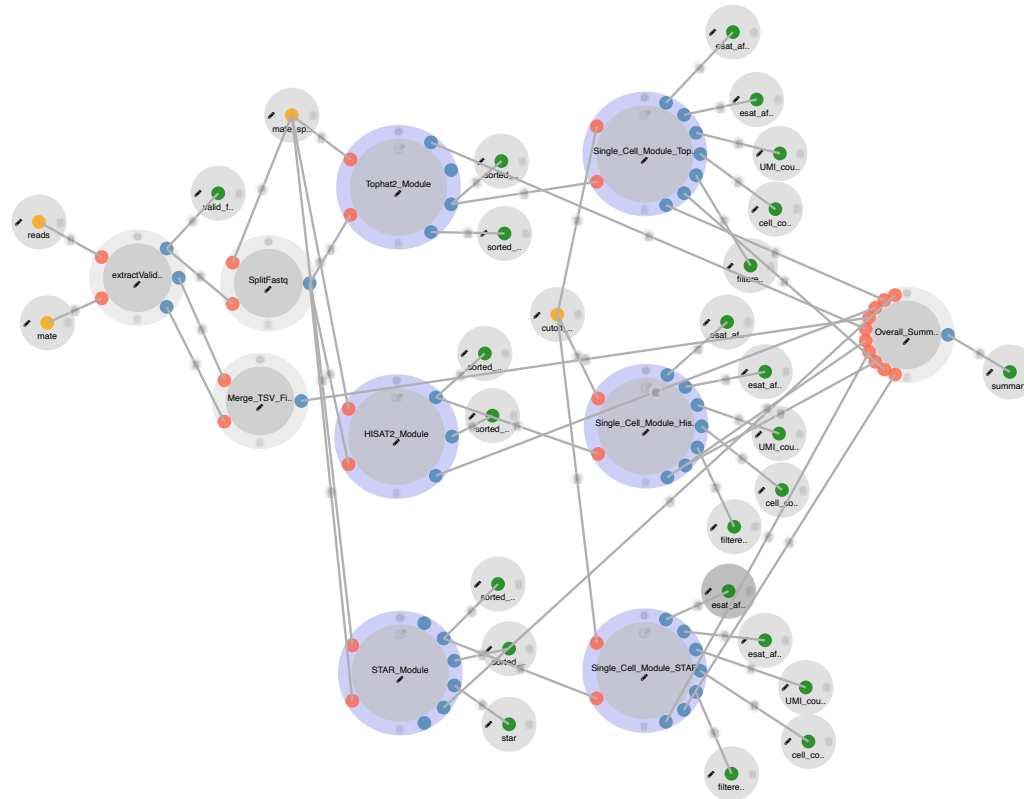


# inDrops Data Processing



# Dolphinnext Pipelines

To process the FASTQs that the instrument generates into a gene expression matrix involves many steps, which can be run as a continuous pipeline with dolphinnext

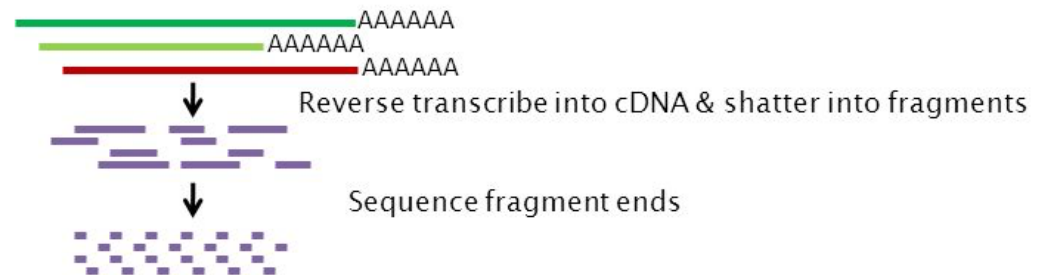




# Dolphinnext Pipelines : bcl2fastq

## 1) bcl2fastq

- valid sample barcode

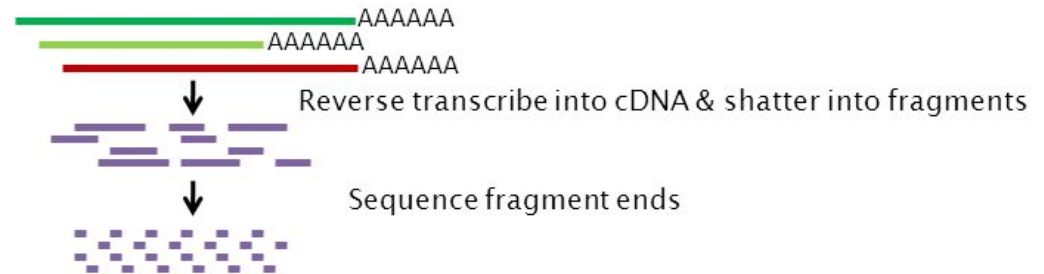


```
@NS500602:550:HM2TFBGX5:1:11101:7989:1090:CAACATTGGCCCAAT:CTGACTTA 1:N:0:TATAGACG
TTCCCNCTAGACATTTCTGTGCATAGATTTTGGTGTGTTTACATAGTCGTTATTCTG
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEE/EAAEEEEEEEEEEEEEEEEEEA6/
```

# Dolphinnext Pipelines : extract valid reads

- 1) bcl2fastq
- 2) extract reads

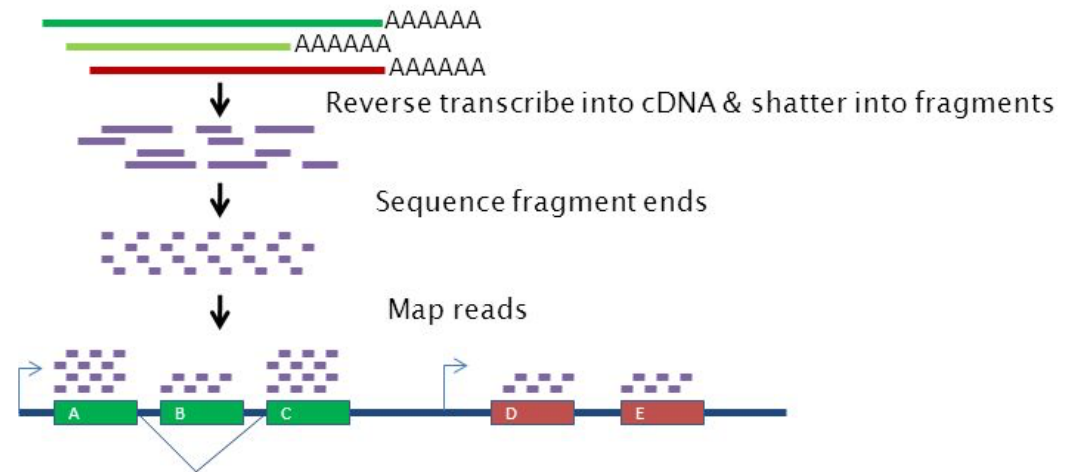
- valid cell barcode
- valid UMI



```
@NS500602:550:HM2TFBGX5:1:11101:7989:1090:CAAACATTGGCCCAAT:CTGACTTA 1:N:0:TATAGACG
TTCCCNCTAGACATTTCTGTGCATAGATTTTGGTGTGTTTACATAGTCGTTATTCTG
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEE/EAAEEEEEEEEEEEEEEEEEEA6/
```

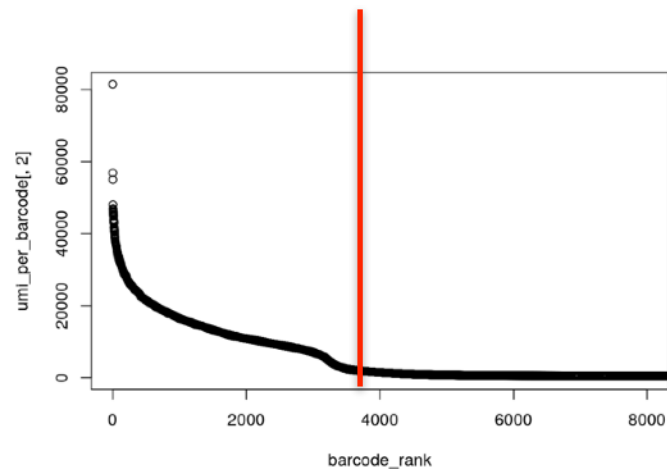
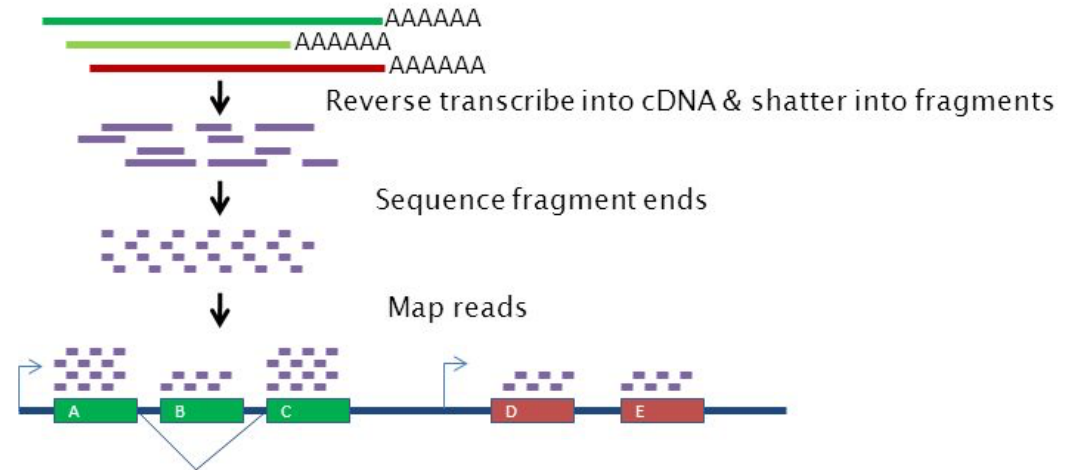
# Dolphinnext Pipelines : alignment

- 1) bcl2fastq
- 2) extract reads
- 3) align



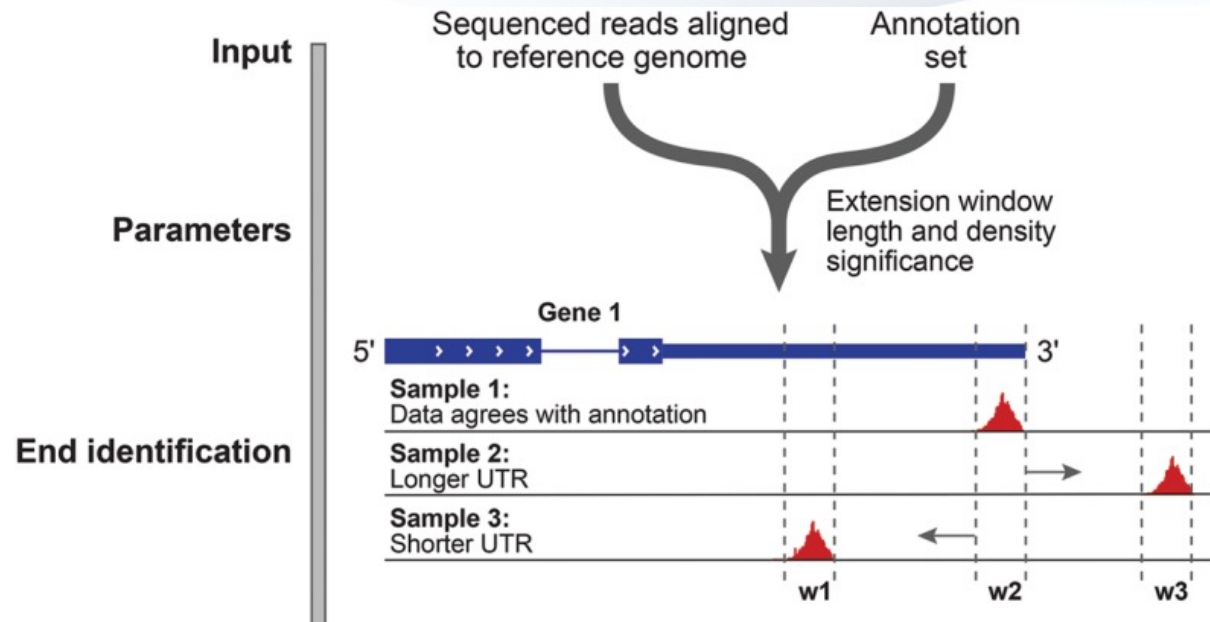
# Dolphinnext Pipelines : filtering

- 1) bcl2fastq
- 2) extract reads
- 3) align
- 4) filter
  - num reads per cell barcode



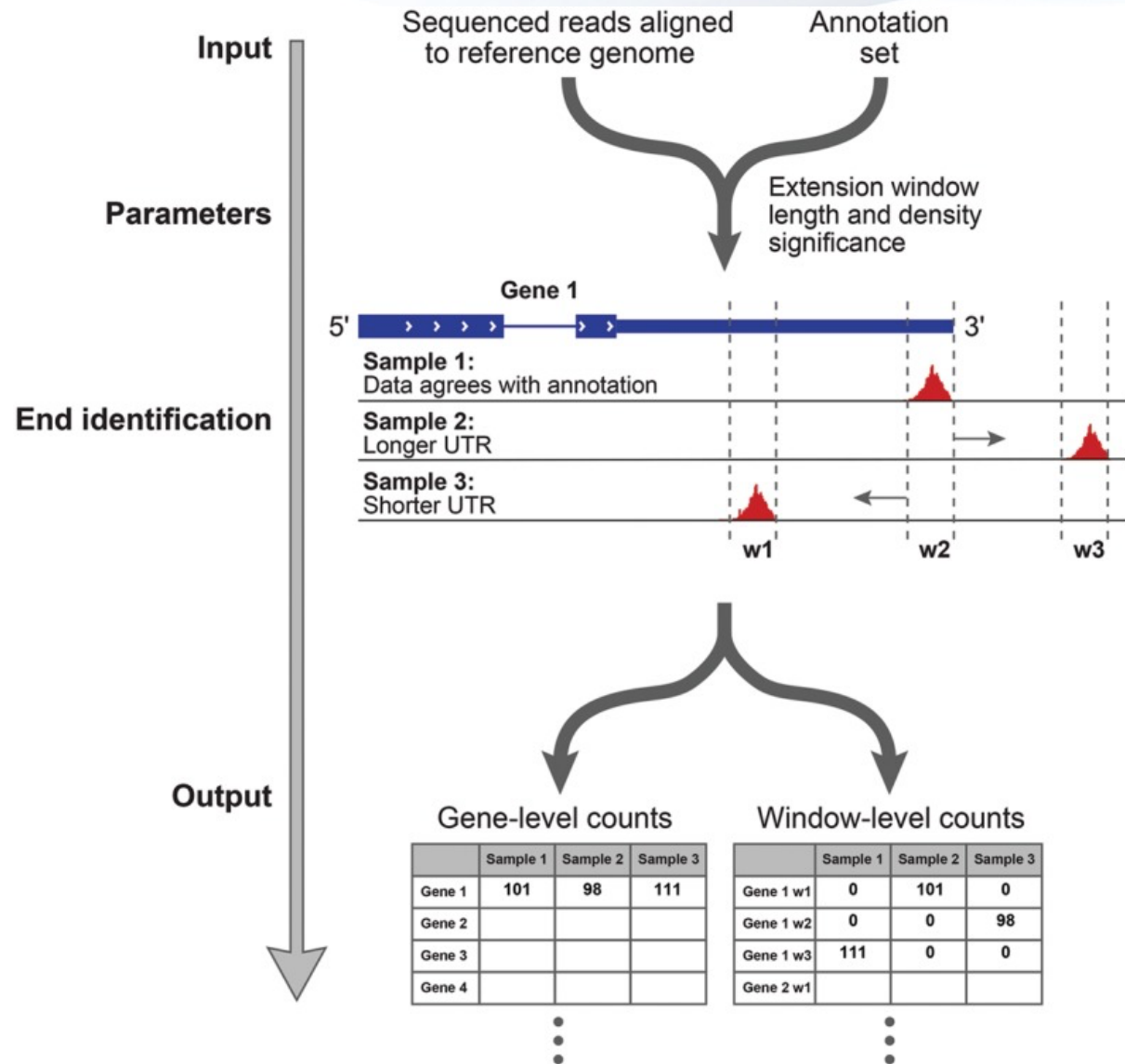
# Dolphinnext Pipelines : ESAT

- 1) bcl2fastq
- 2) extract reads
- 3) align
- 4) filter
- 5) ESAT



# Dolphinnext Pipelines : ESAT

- 1) bcl2fastq
- 2) extract reads
- 3) align
- 4) filter
- 5) ESAT



# Data Structures : UMI Table

	Cell 1	Cell 2	Cell 3	Cell 4...
Gene 1	42	43	10	9
Gene 2	25	24	2	3
Gene 3	10	9	100	98
Gene 4...	40	39	4	5
SUM	117	115	116	115

Bulk = Average of all genes across all cells

Bulk = proportion of each cell type \*  
cell types average expression profile

# Data Structures : UMI Table

	Group A			
	Cell 1	Cell 2	Cell 3	Cell 4...
Gene 1	42	43	10	9
Gene 2	25	24	2	3
Gene 3	10	9	100	98
Gene 4...	40	39	4	5
SUM	117	115	116	115
			Group B	



# Data Structures : UMI Table

+1000 Cells

		Cell 1	Cell 2	Cell 3	Cell 4...
+10000 Genes	Gene 1	42	43	10	9
	Gene 2	25	24	2	3
	Gene 3	10	9	100	98
	Gene 4...	40	39	4	5
	SUM	117	115	116	115

# Data Structures : UMI Table

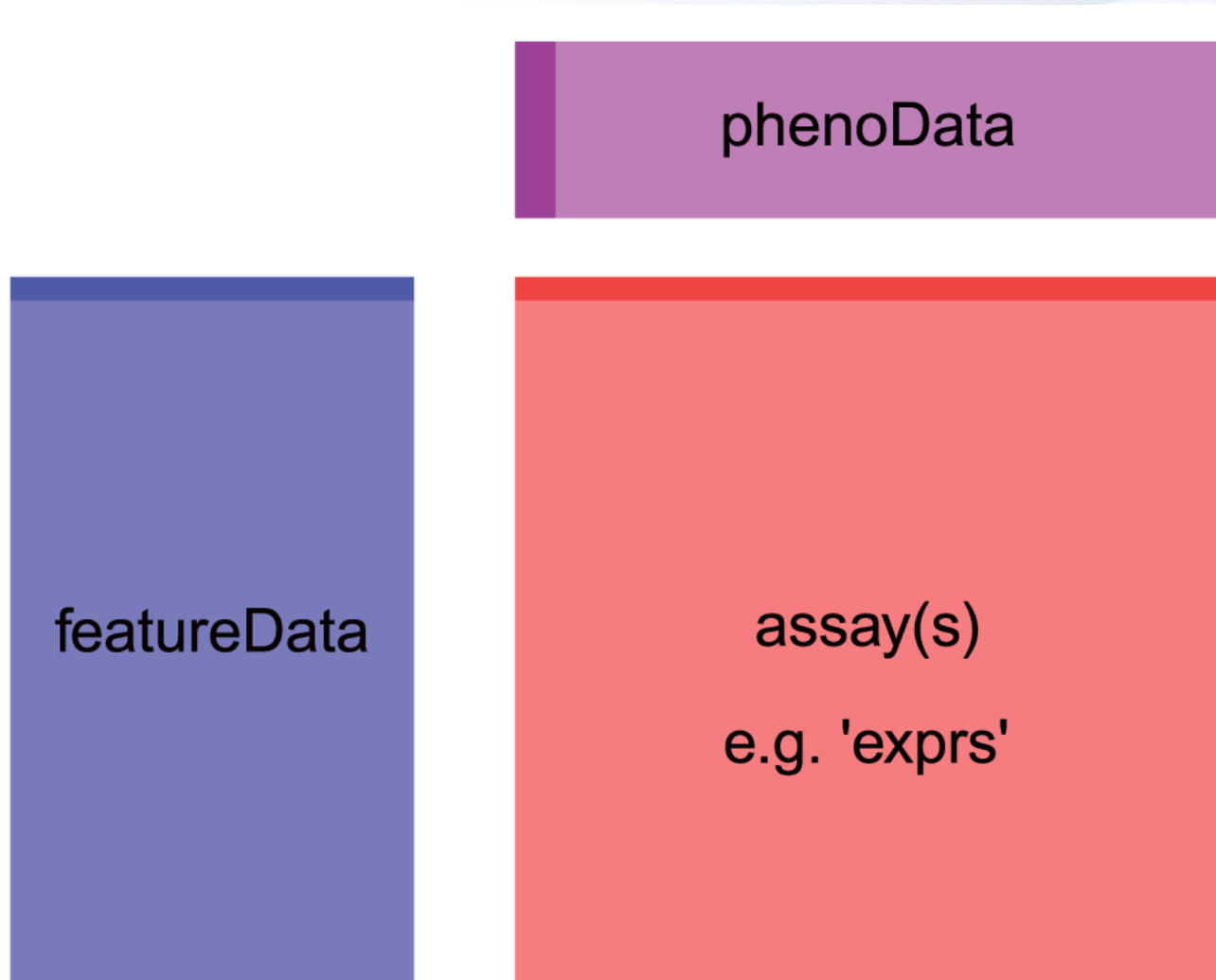
+1000 Cells

+10000 Genes

	Cell 1	Cell 2	Cell 3	Cell 4...
Gene 1	42	43	10	9
Gene 2	25	24	2	3
Gene 3	10	9	100	98
Gene 4...	40	39	4	5
SUM	117	115	116	115

**What if we want to store information about these cells or genes??**

# Data Structures : Expression Set Class



# Data Structures : exprs()

Genes	Exprs		
	0hrA_TGACGGACAAGTAATC	0hrA_ATGGGCACACCTTGCC	0hrA_TCGAAGCTGTTGCACG
	0610007P14Rik	0	2.10985244161277
	0610009B22Rik	0	1.72247211812165
	0610009O20Rik	0	0
	0610010B08Rik	0	0
	0610010F05Rik	0	0
Cells			

# Data Structures : pData()

pData

	size_factor	UMI_sum	x	y	iPC_Comp1	iPC_Comp2...	iPC_Comp12	Cluster	Timepoint
0hrA_TGACGGACAAGTAATC	1.47	1604.45	0.31	0.72	-0.02	0.01	-0.01	Cluster4	0hr
0hrA_ATGGGCACACCTTGCC	1.21	1333.71	0.93	0.33	-0.02	0.01	-0.01	Cluster3	0hr
0hrA_TCGAAGCTGTTGCACG	1.30	1102.09	0.68	0.13	-0.02	0.02	-0.00	Cluster2	0hr
0hrA_TGTTTGAGTCGGTTCG	1.63	1210.53	0.86	0.32	-0.02	0.02	-0.00	Cluster3	0hr
0hrA_TAAATAGGCACAAGGC	0.43	1714.06	0.90	0.29	-0.02	0.01	-0.01	Cluster3	0hr
0hrA_GATTAGACGGGAACCT	0.64	946.06	0.48	0.59	-0.02	0.01	0.00	Cluster1	0hr

Normalization

tSNE

PCA

Clustering Metadata

# Data Structures : fData()

fData

	C1_score	C2_score	C3_score	C1_bulk	C2_bulk	C3_bulk	0hr_score	1hr_score	4hr_score
0610007P14Rik	9704	704	2572	0.05	0.23	0.19	1187	5052	10278
0610009B22Rik	5293	642	1181	0.04	0.11	0.10	503	11045	6766
0610009O20Rik	7535	2732	7733	0.00	0.01	0.01	6310	7579	4016
0610010B08Rik	3184	5176	9845	0.00	0.01	0.00	8161	5170	3772
0610010F05Rik	75	11158	8698	0.11	0.01	0.04	10373	662	4595
0610010K14Rik	3888	5066	2634	0.07	0.10	0.12	1774	4074	10087

Cluster Markers

Aggregated Bulk

Timepoint Markers