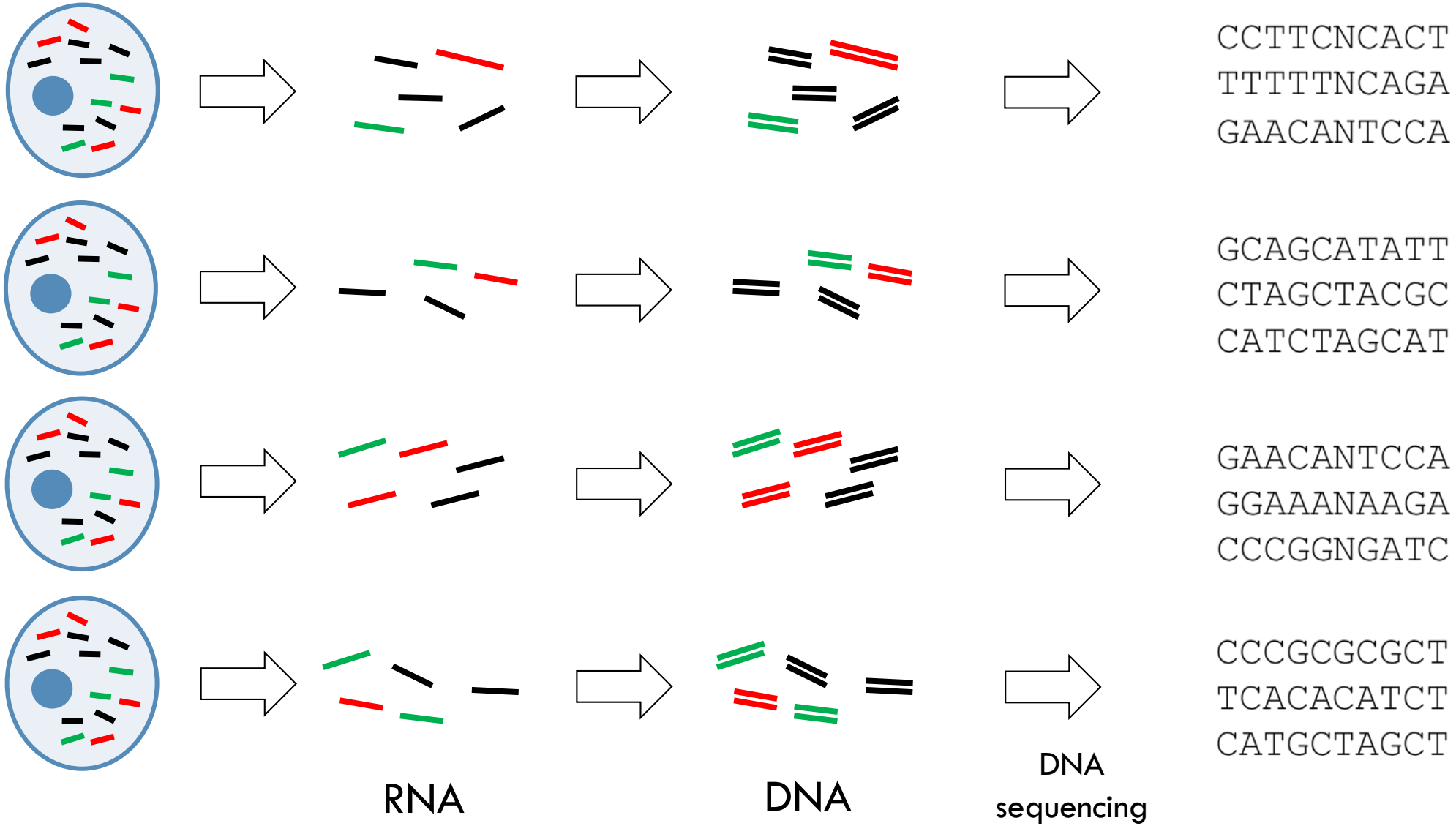
Lecture 10: Clustering Single-Cell RNA-seq data

ECE 365

Announcements:

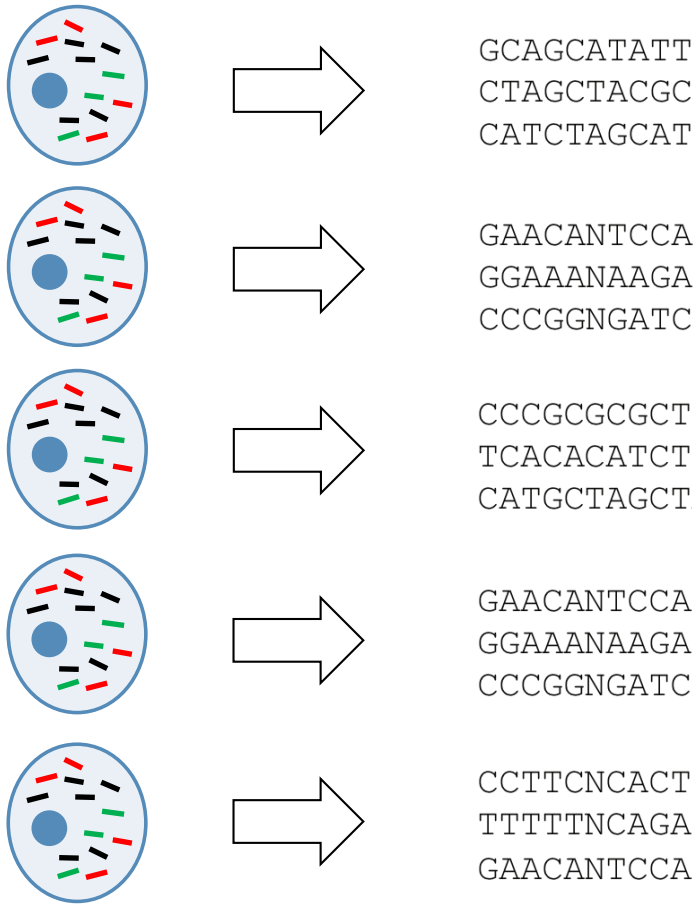
- Lab 4 (RNA-seq) due tomorrow
- Quiz 2 today

From previous lecture: Single-Cell RNA-seq



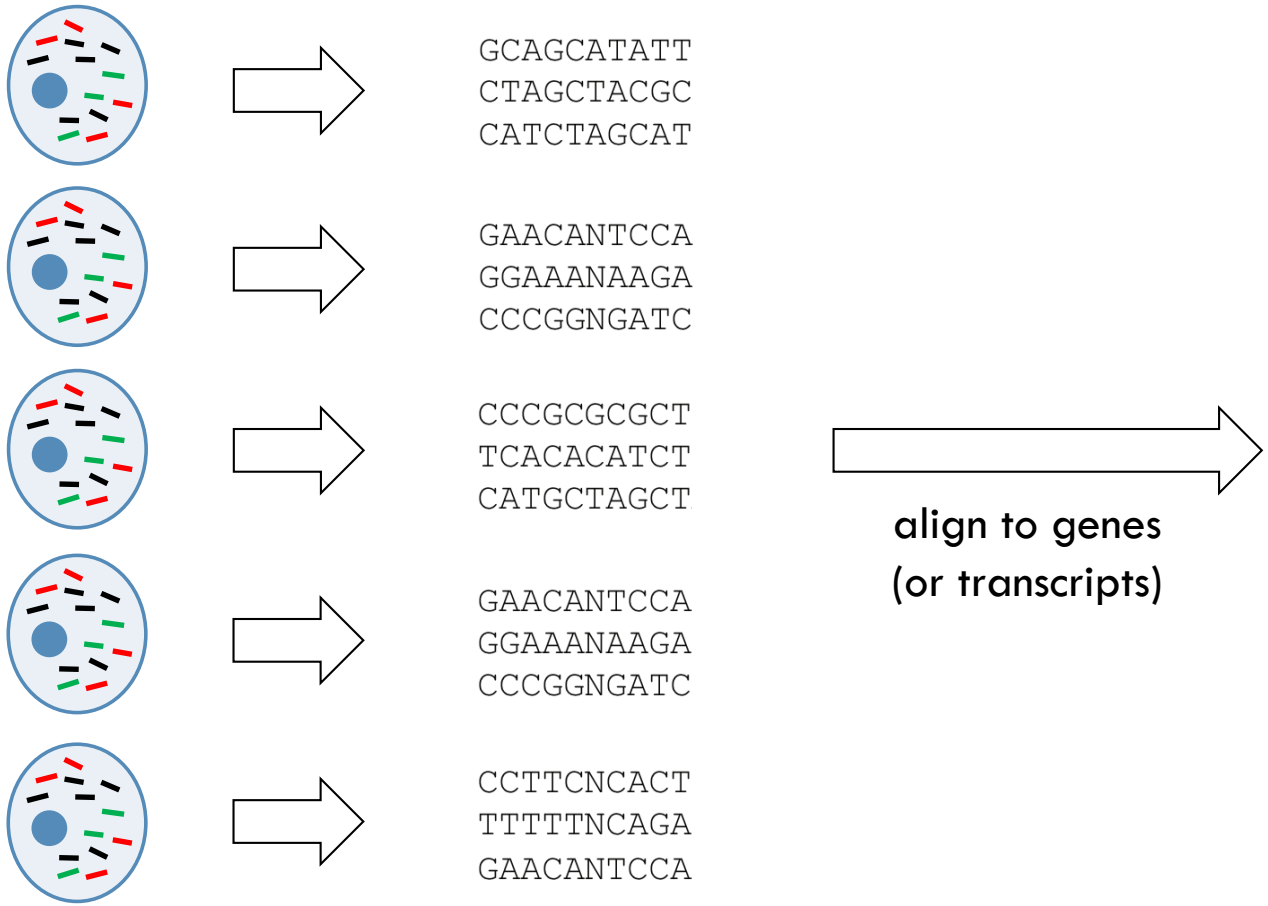
From previous lecture: Single-Cell RNA-seq

- We obtain reads for each cell



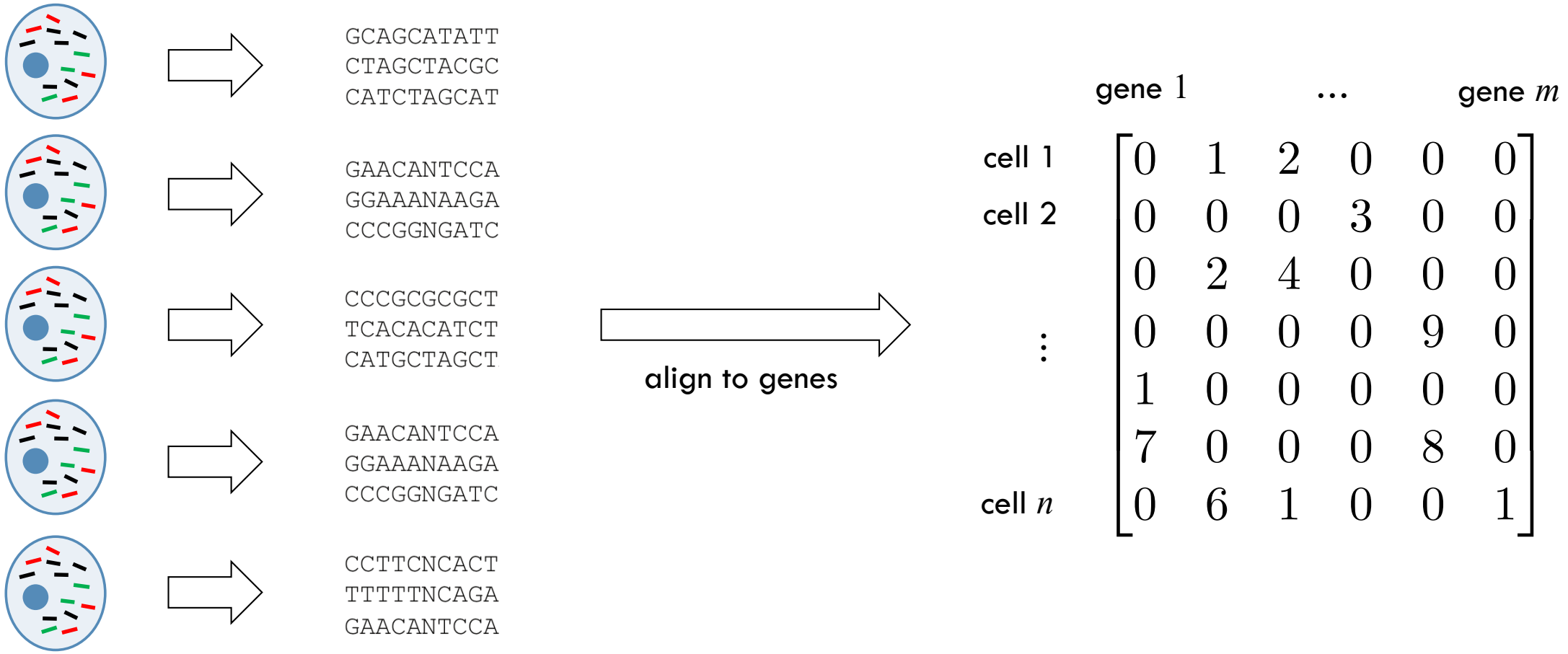
From previous lecture: Single-Cell RNA-seq

- We obtain reads for each cell



From previous lecture: Single-Cell RNA-seq

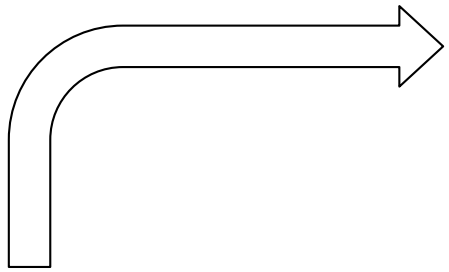
- We obtain reads for each cell

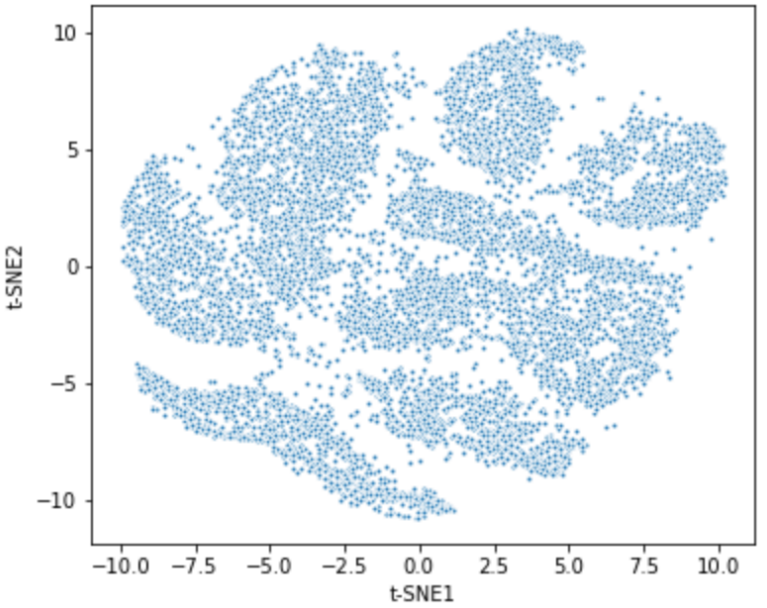


Single-Cell RNA-seq Data Analysis

	gene 1		…		gene m
cell 1	0	1	2	0	0	0
cell 2	0	0	0	3	0	0
	0	2	4	0	0	0
⋮	0	0	0	0	9	0
	1	0	0	0	0	0
	7	0	0	0	8	0
cell n	0	6	1	0	0	1

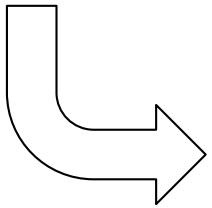
Dimensionality reduction and visualization



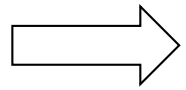


$$\ln(1+x)$$

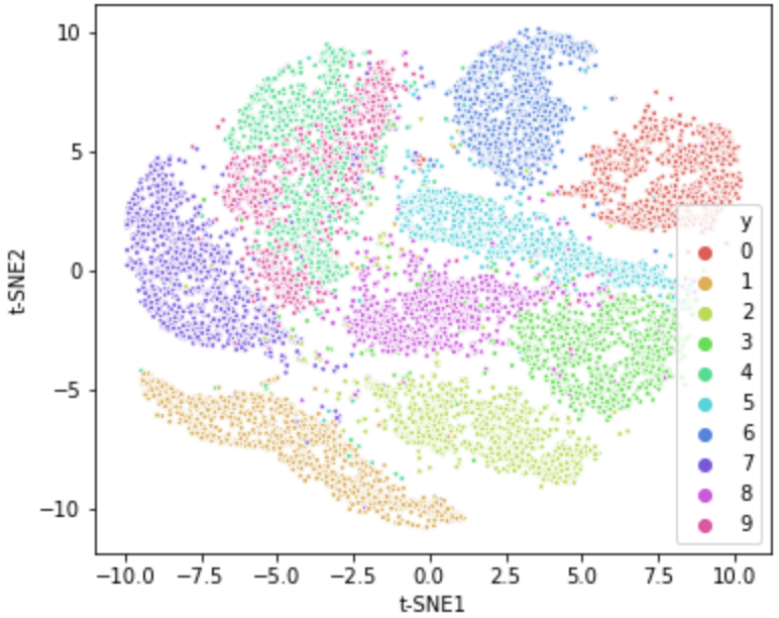
(pre-processing)



0	0.7	1.1	0	0	0
0	0	0	1.4	0	0
0	1.1	1.6	0	0	0
0	0	0	0	2.3	0
0.7	0	0	0	0	0
2.1	0	0	0	2.2	0
0	1.9	0.7	0	0	0.7



Clustering



Dimensionality reduction via PCA

Data matrix X $(n \times m)$

Perform SVD on X: $X = U_{n\times n} S_{n\times m} V^T_{m\times m}$

use this to get transformation W

$$S = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \end{bmatrix}$$

singular values

$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m$

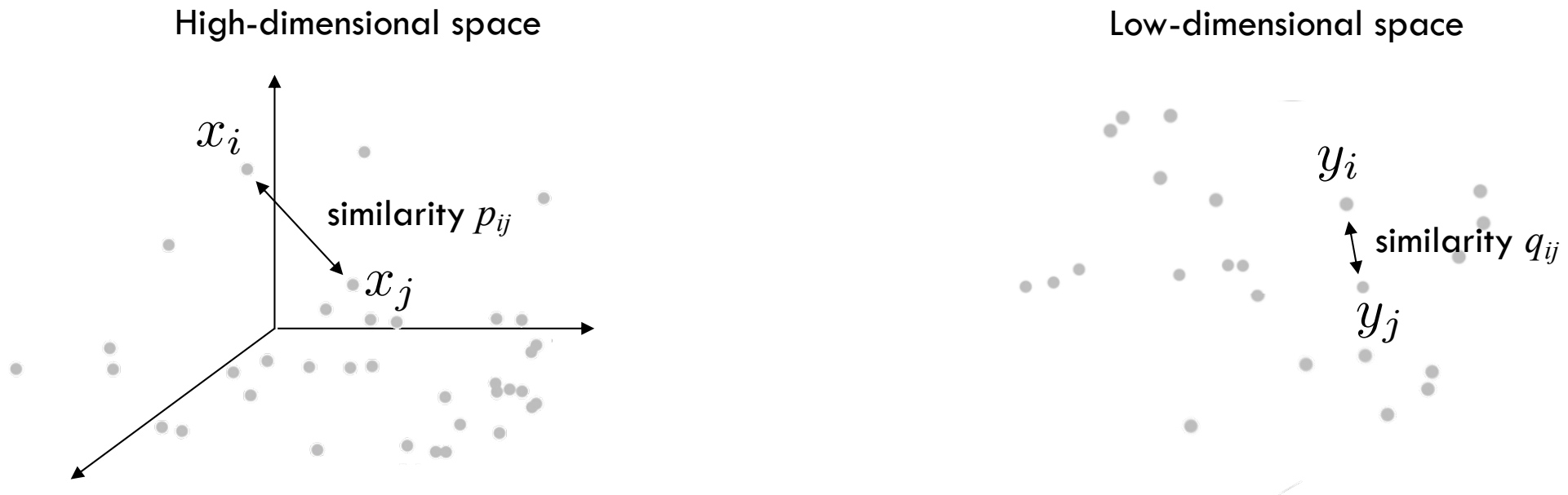
$$V^T = \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ & \vdots & \\ — & v_m & — \end{bmatrix}$$ pick top k rows $\rightarrow$ $\begin{bmatrix} — & v_1 & — \\ & \vdots & \\ — & v_k & — \end{bmatrix} = W_{k\times m}$

now we can find $\tilde{x}_i = W x_i$

$\tilde{X}^T = W X^T$

Non-linear dimensionality reduction via t-SNE

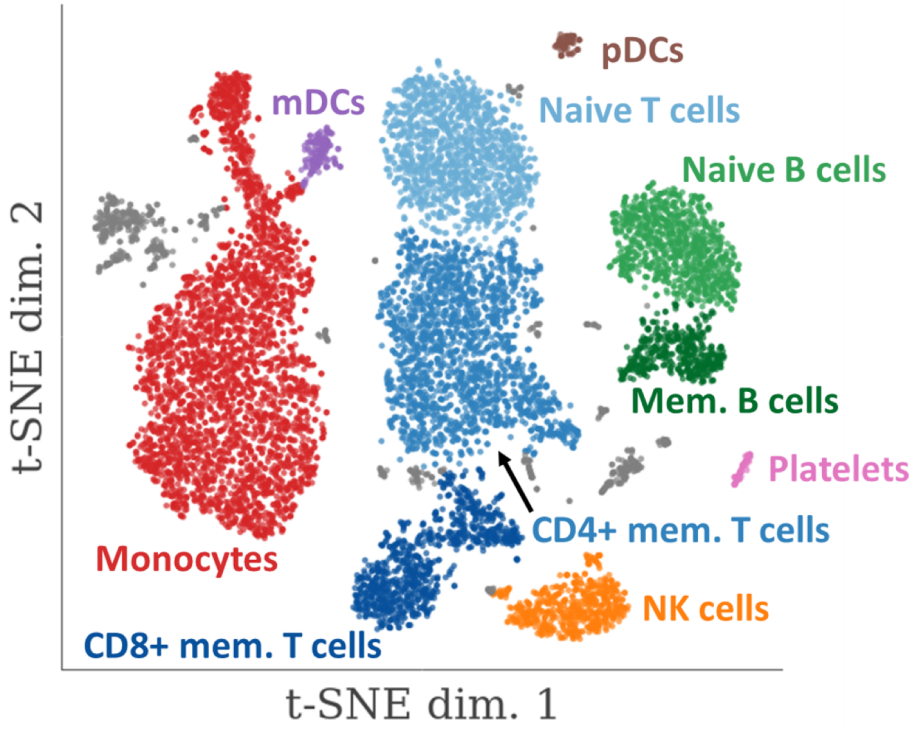
- *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE)
  - Popular *non-linear* dimensionality reduction method
  - High-level idea:



  - Use gradient descent to move y_i's to minimize $D(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

Non-linear dimensionality reduction via t-SNE

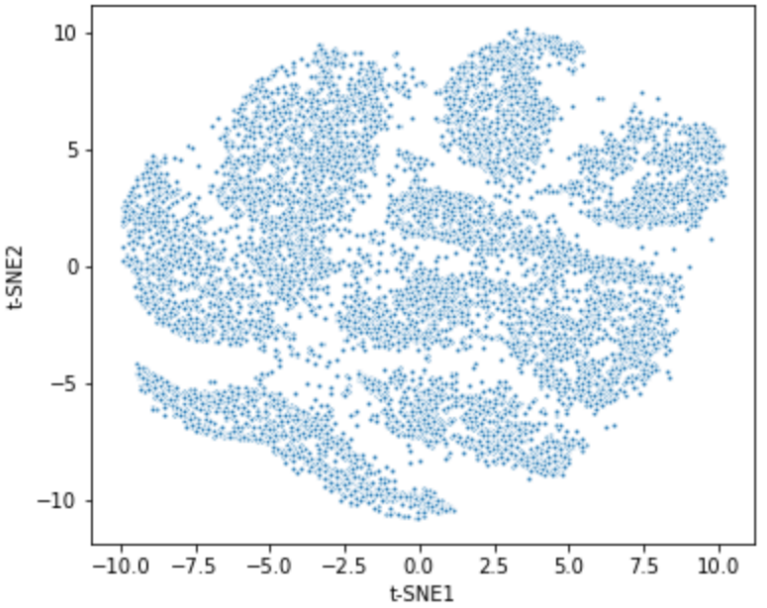
- Very popular for single-cell RNA-seq data



Single-Cell RNA-seq Data Analysis

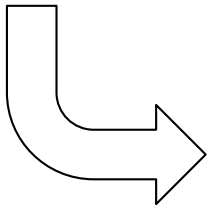
	gene 1		…		gene m
cell 1	0	1	2	0	0	0
cell 2	0	0	0	3	0	0
	0	2	4	0	0	0
⋮	0	0	0	0	9	0
	1	0	0	0	0	0
	7	0	0	0	8	0
cell n	0	6	1	0	0	1

Dimensionality reduction and visualization

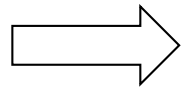


$$\ln(1+x)$$

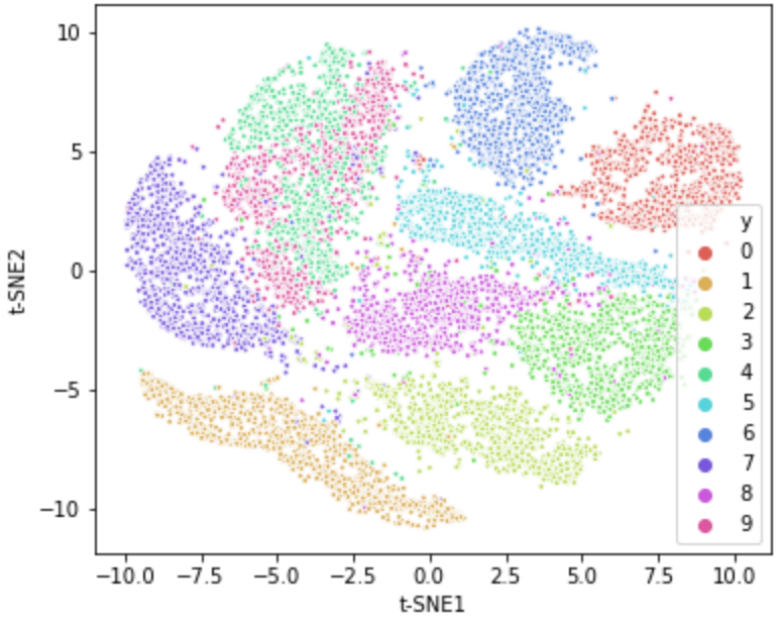
(pre-processing)



$$\begin{bmatrix} 0 & 0.7 & 1.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.4 & 0 & 0 \\ 0 & 1.1 & 1.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.3 & 0 \\ 0.7 & 0 & 0 & 0 & 0 & 0 \\ 2.1 & 0 & 0 & 0 & 2.2 & 0 \\ 0 & 1.9 & 0.7 & 0 & 0 & 0.7 \end{bmatrix}$$



Clustering



From first module: K-means clustering

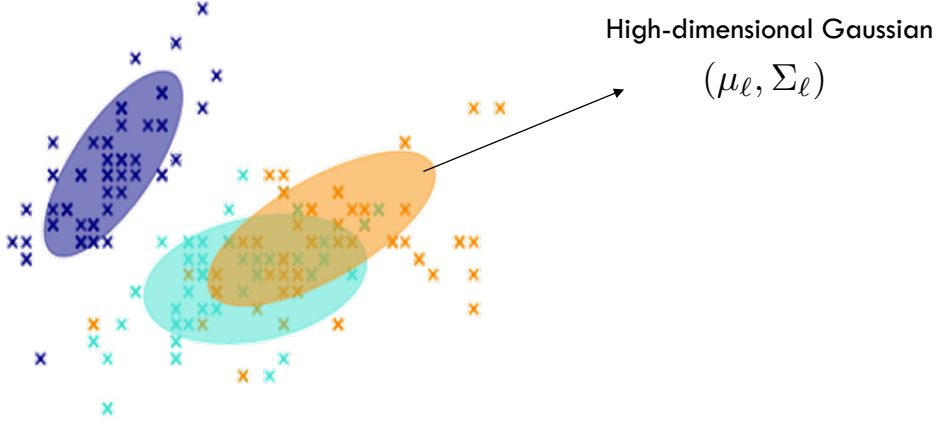
- Data points: $x_i \in R^d$ for $i = 1, \ldots, N$
- Initialize means μ_ℓ for $\ell = 1, \ldots, K$ to K random points
- Repeat until convergence:
  - Assign each x_i to closest mean
  - After all x_i's are assigned, recompute μ_ℓ for $\ell = 1, \ldots, K$ for each cluster

Gaussian Mixture Model

- Generative model: multiple high-dimensional Gaussian distributions

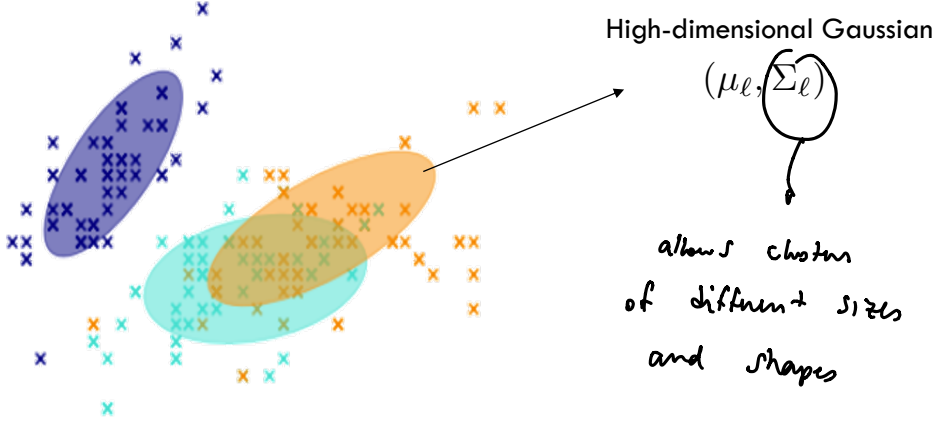
Gaussian Mixture Model

- Generative model: multiple high-dimensional Gaussian distributions



Gaussian Mixture Model

- Generative model: multiple high-dimensional Gaussian distributions



- Model parameters are estimated with the EM algorithm!

Recall: EM algorithm for RNA quantification

- Initialize: $\rho_j = \frac{1}{K}$ for $j = 1, \ldots, K$
- Repeat:

membership variable

E-step

- $Z_{ik} = \begin{cases} \frac{\rho_k}{\sum_{j \in S_i} \rho_j}, & \text{for } k \in S_i \\ 0, & \text{otherwise} \end{cases}$

M-step

- $\rho_k = \frac{\frac{\theta_k}{\ell_k}}{\sum_{j=1}^{K} \frac{\theta_j}{\ell_j}}$, where $\theta_k = \frac{1}{N} \sum_{i=1}^{N} Z_{ik}$

$\frac{\theta_k}{\ell_k}$: length transcript k

θ_k: prob. of getting a read from tr. k

EM algorithm for GMM fitting

- Initialize: (μ_ℓ, Σ_ℓ) and mixture coefficients $\theta_\ell = \frac{1}{K}$ for $\ell = 1, \ldots, K$

  fraction of each Gaussian

- Repeat:

  membership variable

  E-step

  - $Z_{ik} = \frac{\mathcal{N}(x_i|\mu_k,\Sigma_k)\cdot\rho_k}{\sum_{j=1}^{K}\mathcal{N}(x_i|\mu_j,\Sigma_j)\cdot\rho_j}$

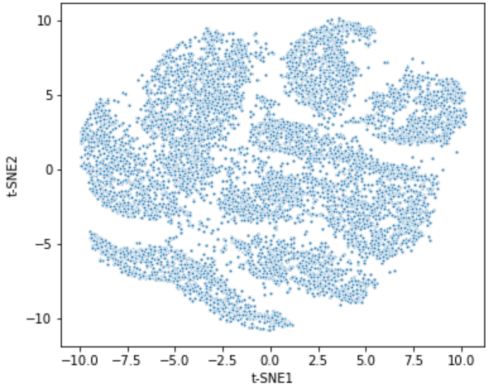
  M-step

  - $\theta_k = \frac{1}{N}\sum_{i=1}^{N} Z_{ik}, \quad \mu_k = \frac{\sum_{i=1}^{N} Z_{ik}x_i}{\sum_{i=1}^{N} Z_{ik}}, \quad \Sigma_k = \frac{\sum_{i=1}^{N} Z_{ik}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{N} Z_{ik}}$

Single-Cell RNA-seq Data Analysis

	gene 1			...		gene m
cell 1	0	1	2	0	0	0
cell 2	0	0	0	3	0	0
	0	2	4	0	0	0
⋮	0	0	0	0	9	0
	1	0	0	0	0	0
	7	0	0	0	8	0
cell n	0	6	1	0	0	1

Dimensionality reduction and visualization

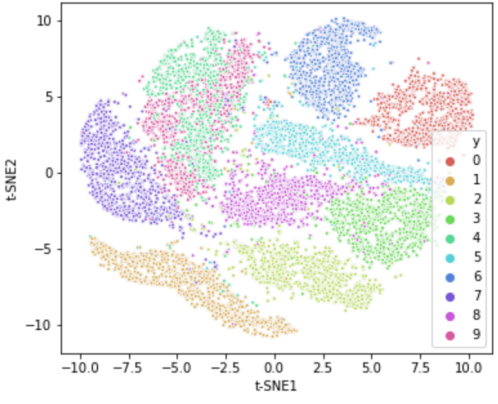


$\ln(1+x)$

(pre-processing)

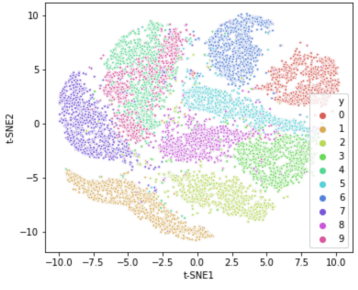
$$\begin{bmatrix} 0 & 0.7 & 1.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.4 & 0 & 0 \\ 0 & 1.1 & 1.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.3 & 0 \\ 0.7 & 0 & 0 & 0 & 0 & 0 \\ 2.1 & 0 & 0 & 0 & 2.2 & 0 \\ 0 & 1.9 & 0.7 & 0 & 0 & 0.7 \end{bmatrix}$$

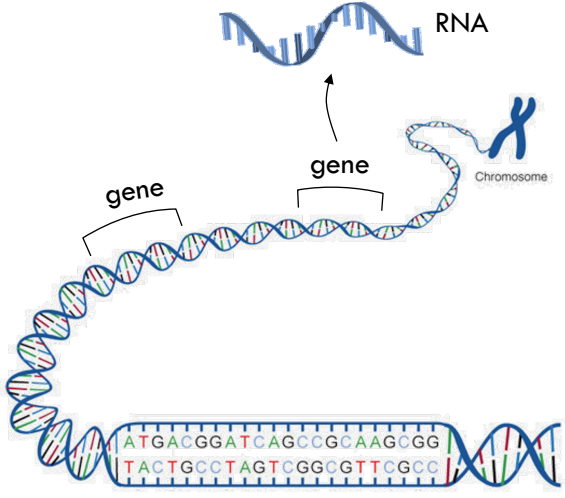
Clustering



Wrapping up

- First part: DNA sequencing data analysis
  - Sequence alignment (dynamic programming, indexing)
  - Genome-wide association studies (GWAS)
- Second part: RNA sequencing data analysis
  - RNA quantification (EM algorithm)
  - Single-cell RNA-seq (dim. reduction, clustering)





Quick review: data types

Genotype data

	SNP 1					SNP m
person 1	0	1	1	0	0	1
	1	0	2	1	0	2
	1	1	1	0	0	0
	0	0	0	2	1	1
	1	1	1	0	0	0
	0	2	0	0	1	0
person n	1	1	0	1	0	2

Bulk RNA-seq
(transcript quantification)

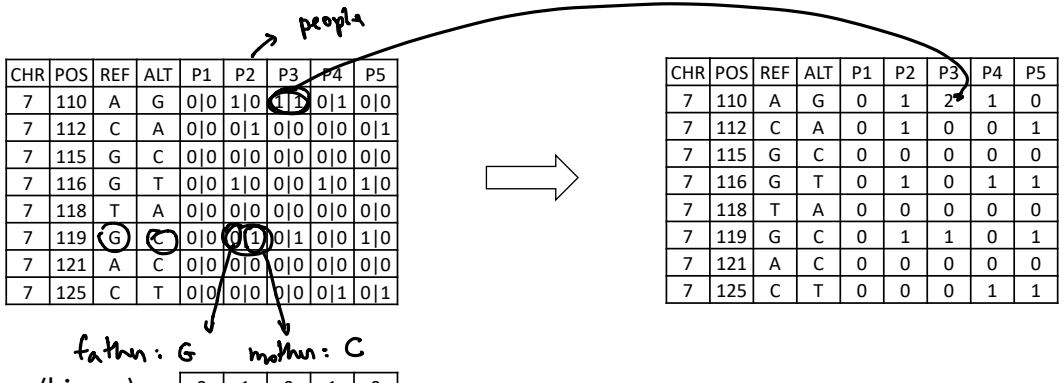
	tr. 1			⋯		tr. K
read 1	0	1	1	0	0	0
read 2	0	0	0	1	0	0
	0	1	1	0	0	0
⋮	0	0	0	0	1	0
	1	0	0	0	0	0
	1	0	0	0	1	0
read N	0	1	0	0	0	1

Single-cell RNA-seq

	gene 1			⋯		gene m
cell 1	0	1	2	0	0	0
cell 2	0	0	0	3	0	0
	0	2	4	0	0	0
⋮	0	0	0	0	9	0
	1	0	0	0	0	0
	7	0	0	0	8	0
cell n	0	6	1	0	0	1

Quick review: genotype data

- VCF file looks like this:

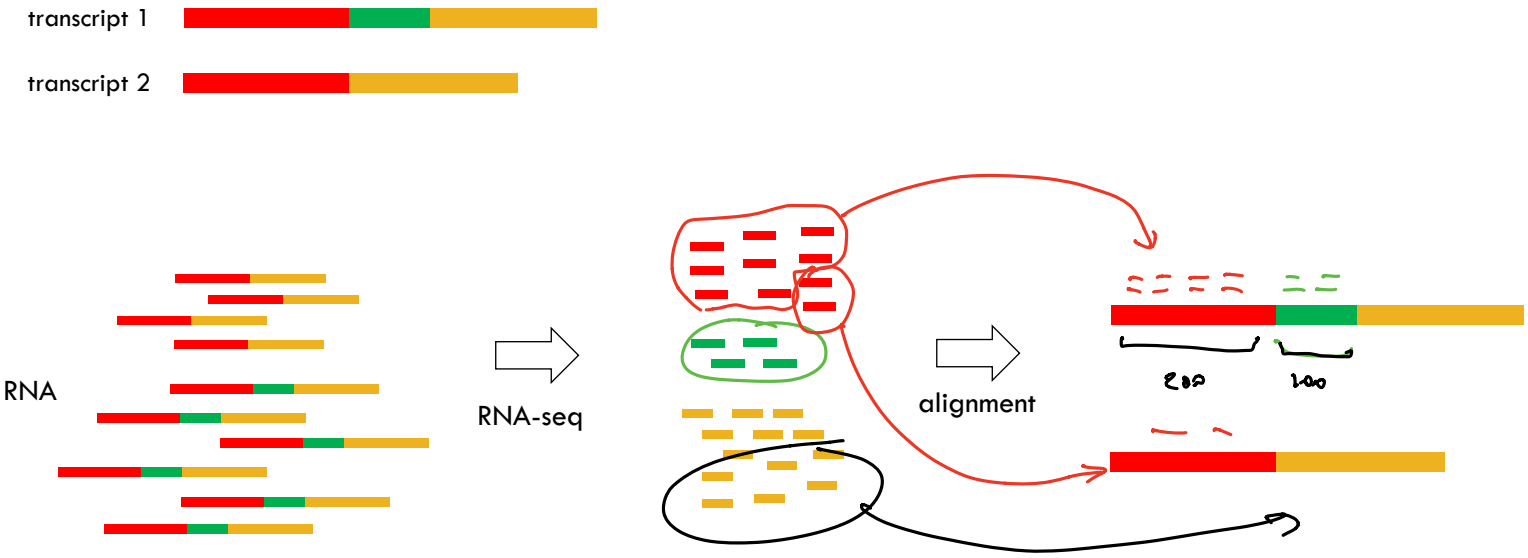


Used to obtain a GWAS risk model: $\ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_{112}x_{112} + \beta_{121}x_{121}$

log-odds-ratio for general population

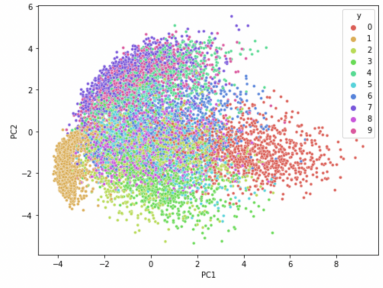
log-odds ratio

Quick review: transcript quantification

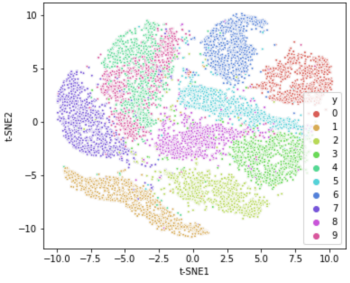


Quick review: single-cell RNA-seq

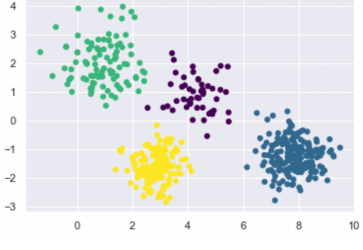
Linear dimensionality reduction (PCA)



Non-linear dimensionality reduction (t-SNE)



Clustering: K-means vs GMM



Better algorithms exist
for single-cell RNA-seq data

Wrapping up

- First part: DNA sequencing data analysis
  - Sequence alignment (dynamic programming, indexing)
  - Genome-wide association studies (GWAS)
- Second part: RNA sequencing data analysis
  - RNA quantification (EM algorithm)
  - Single-cell RNA-seq (dim. reduction, clustering)