BIO214 Lecture 8

Bioinformatics-II

Finding Patterns in Normalized Omics Data - 2

Zhen Wei; 2023-Feb-14

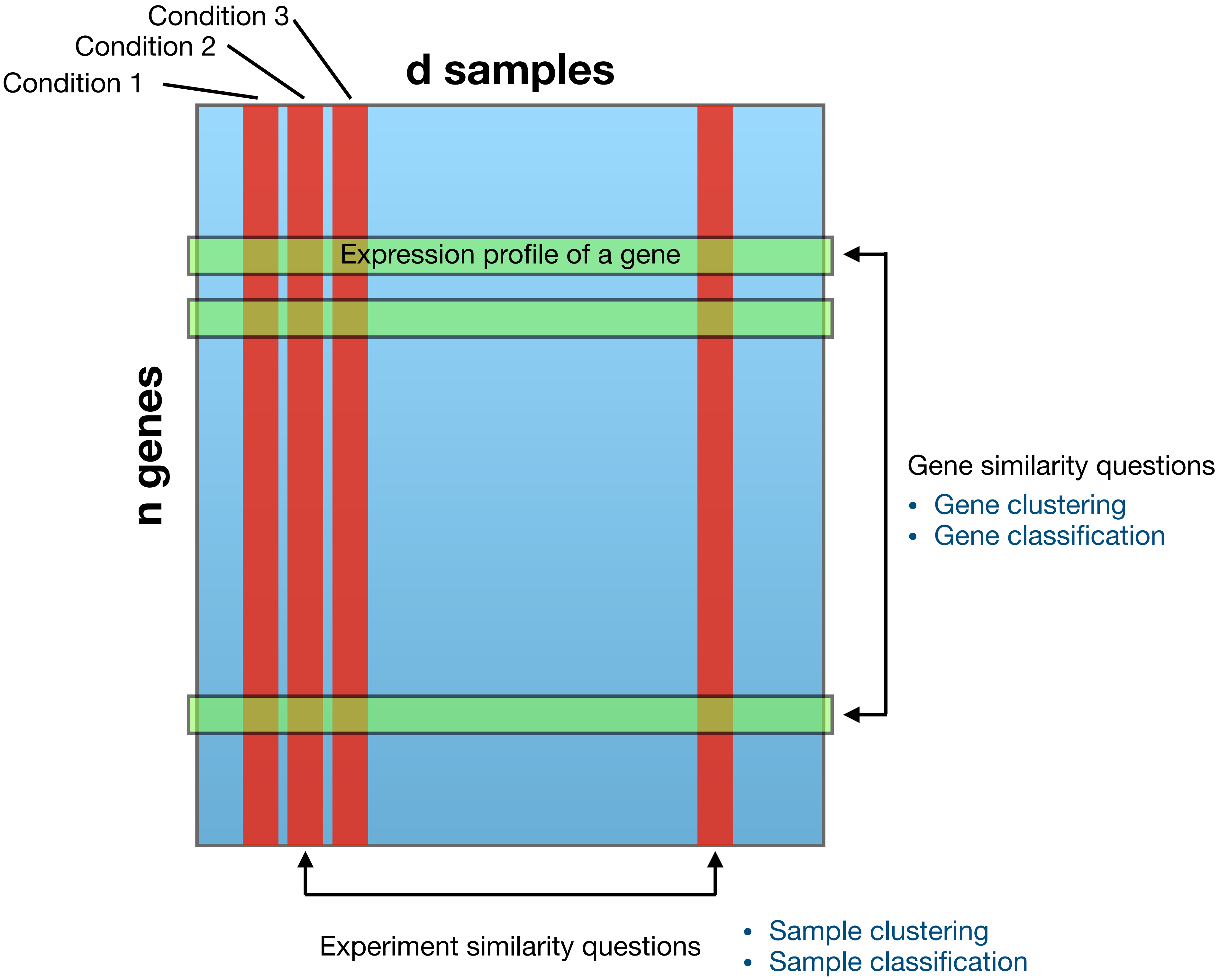
Outline

- Clustering v.s. Classification
- Hard clustering v.s. Soft clustering
- Differential expression analysis

Clustering v.s. Classification

Expression analysis data matrix

- (Normalized) measures of 20000 genes in 100s of conditions



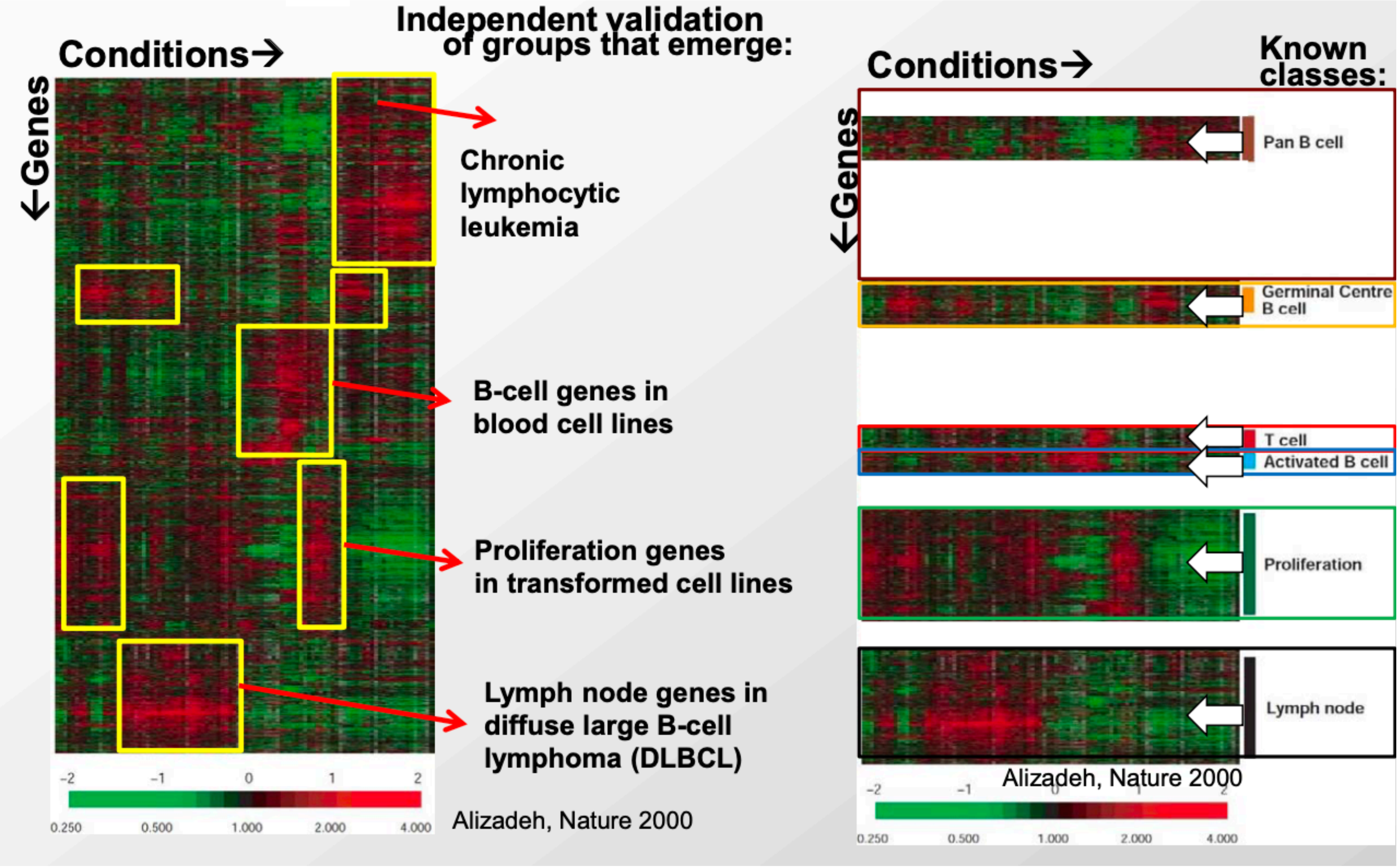
Gene similarity questions

- Gene clustering
- Gene classification

Experiment similarity questions

- Sample clustering
- Sample classification

Clustering v.s. Classification



Alizadeh, Nature 2000

Alizadeh, Nature 2000

Goal of Clustering: **Group similar items** that likely come from the same category, and in doing so **reveal hidden structure**.

- **Unsupervised learning**

Goal of Classification: Extract features from the data that best **assign new elements** to ≥1 of **well-defined classes**.

- **Supervised learning**

Clustering v.s. Classification

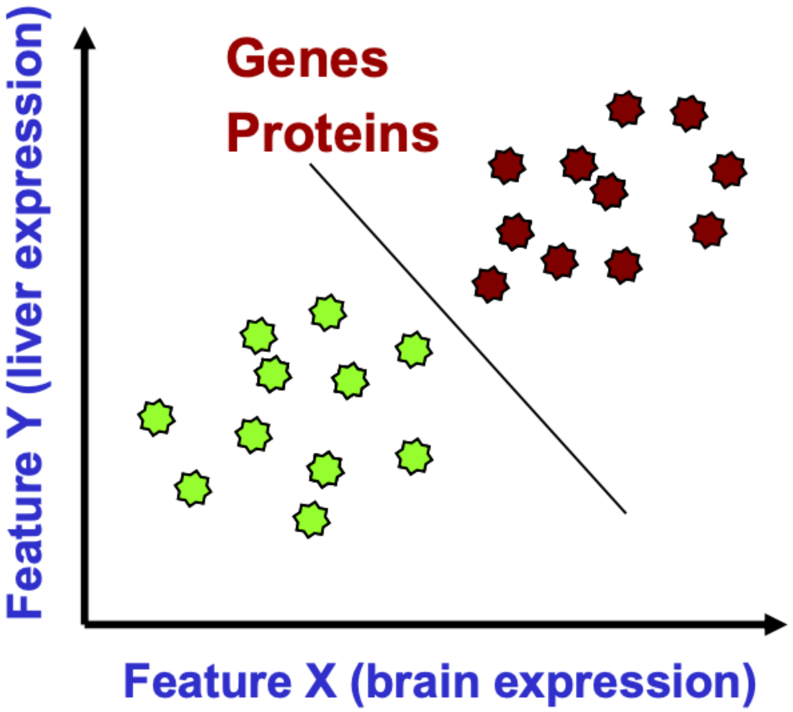
Objects characterized by one or more features

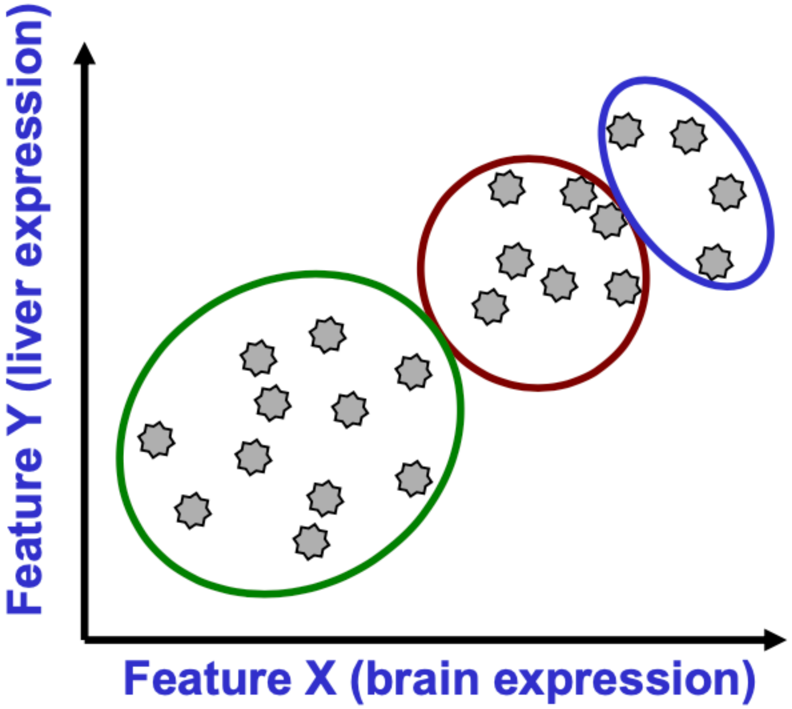
Classification (supervised learning)

- Have labels for some points
- Want a “rule” that will accurately assign labels to new points
- Sub-problem: Feature selection
- Metric: Classification accuracy

Clustering (unsupervised learning)

- No labels
- Group points into clusters based on how “near” they are to one another
- Identify structure in data
- Metric: independent validation features



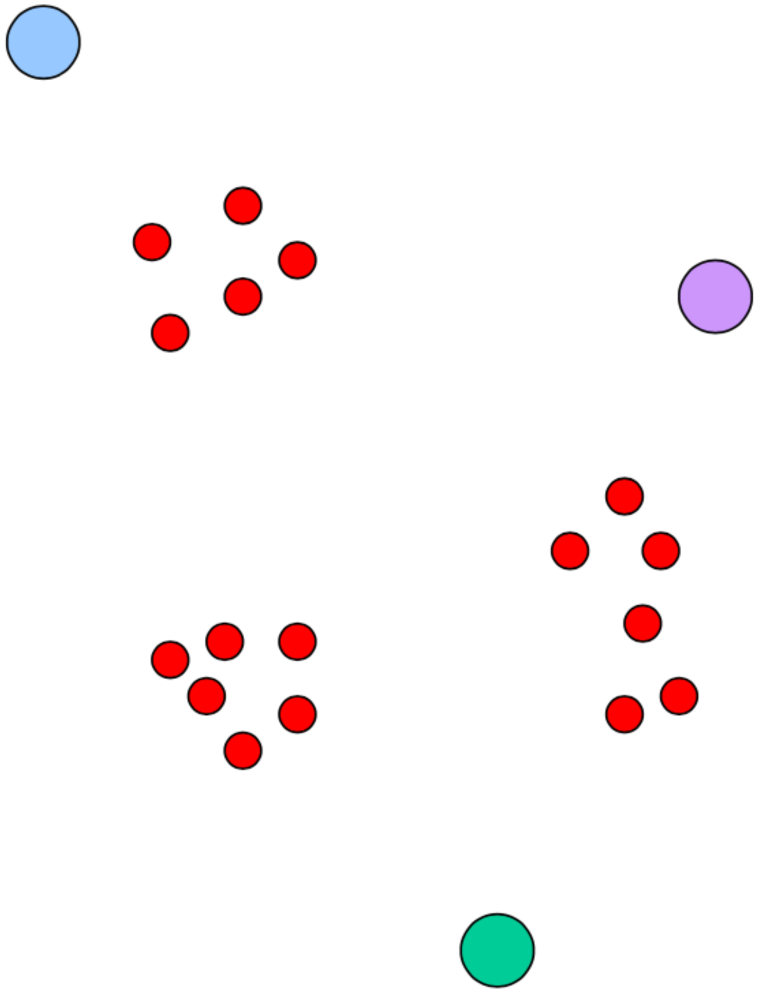


Clustering: K-Means clustering algorithm

- Randomly Initialize cluster centers
- **E step**:

  Assign data points to nearest clusters.
- **M step**:

  Recalculate cluster centers.
- Repeat... until convergence.



Clustering: K-Means clustering algorithm

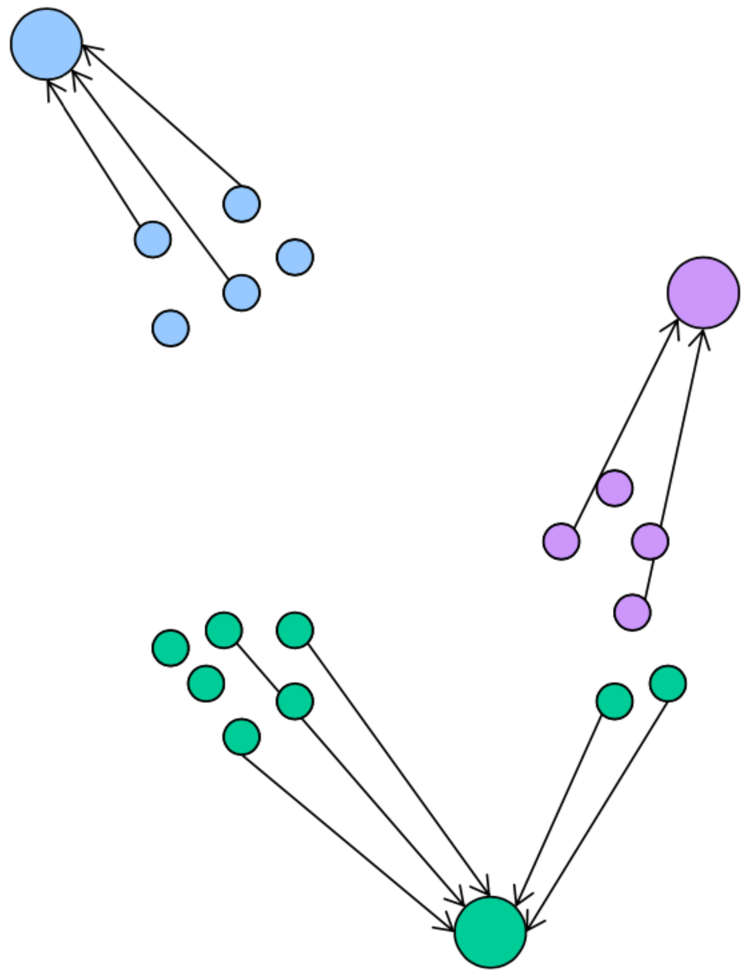
- Randomly Initialize cluster centers
- **E step**:

  Assign data points to nearest clusters.

- **M step**:

  Recalculate cluster centers.

- Repeat... until convergence.

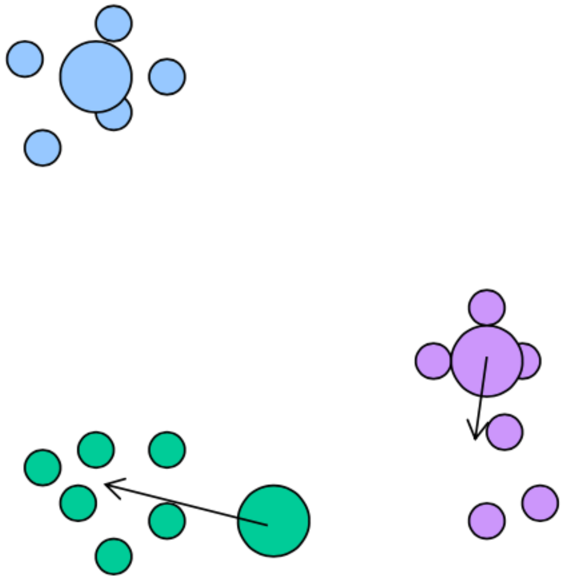


Clustering: K-Means clustering algorithm

- Randomly Initialize cluster centers
- **E step**:

  Assign data points to nearest clusters.
- **M step**:

  Recalculate cluster centers.
- Repeat... until convergence.

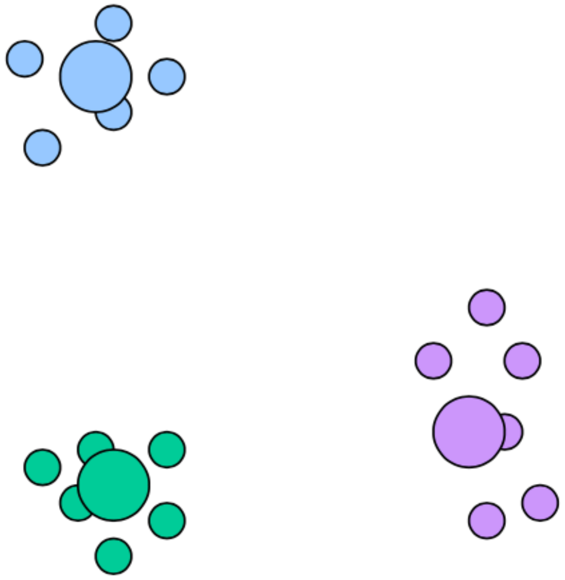


Clustering: K-Means clustering algorithm

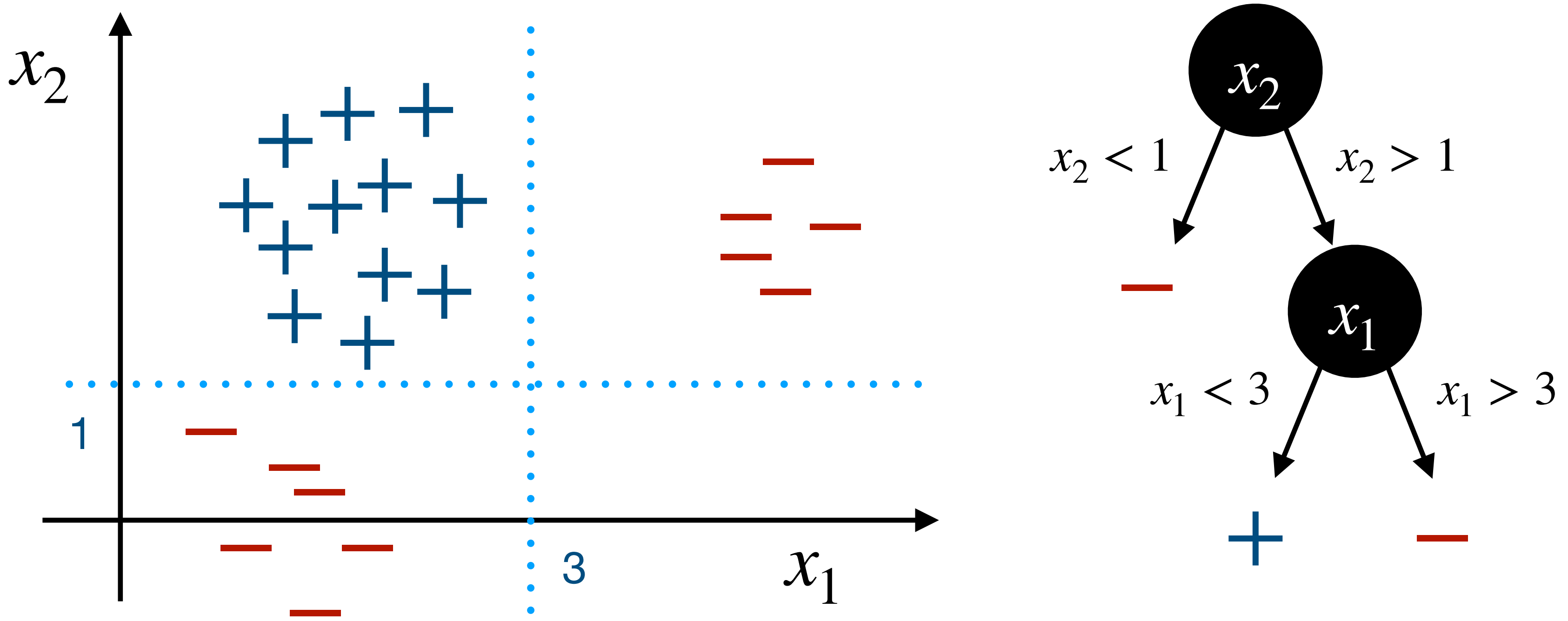
- Randomly Initialize cluster centers
- **E step**:

  Assign data points to nearest clusters.
- **M step**:

  Recalculate cluster centers.
- Repeat... until convergence.



Classification: random forest algorithm

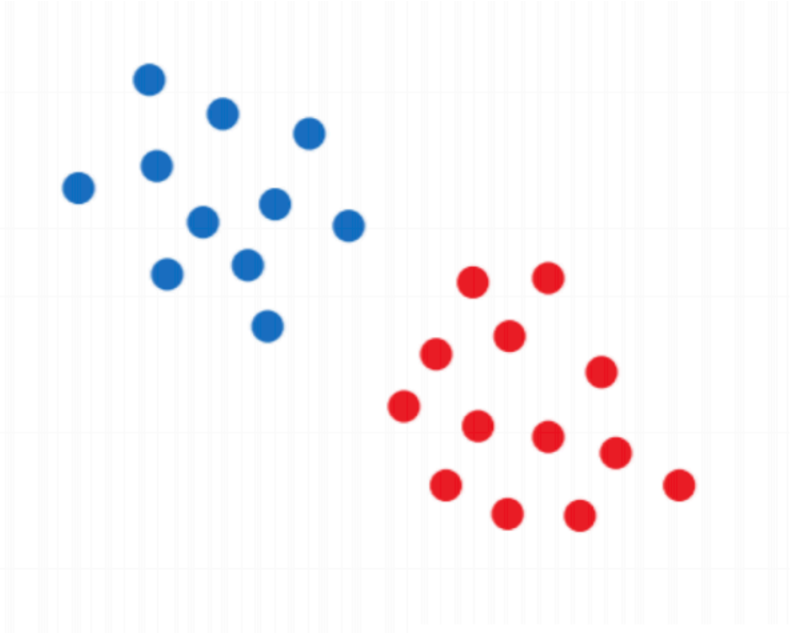


- Create N bootstrap samples, which are training sets re-sampled with replacement.
- Build a (randomized) decision tree on each bootstrap sample.
- Average the predictions made by the N randomized decision trees (averaging the predictions of multiple models is called **ensemble**)

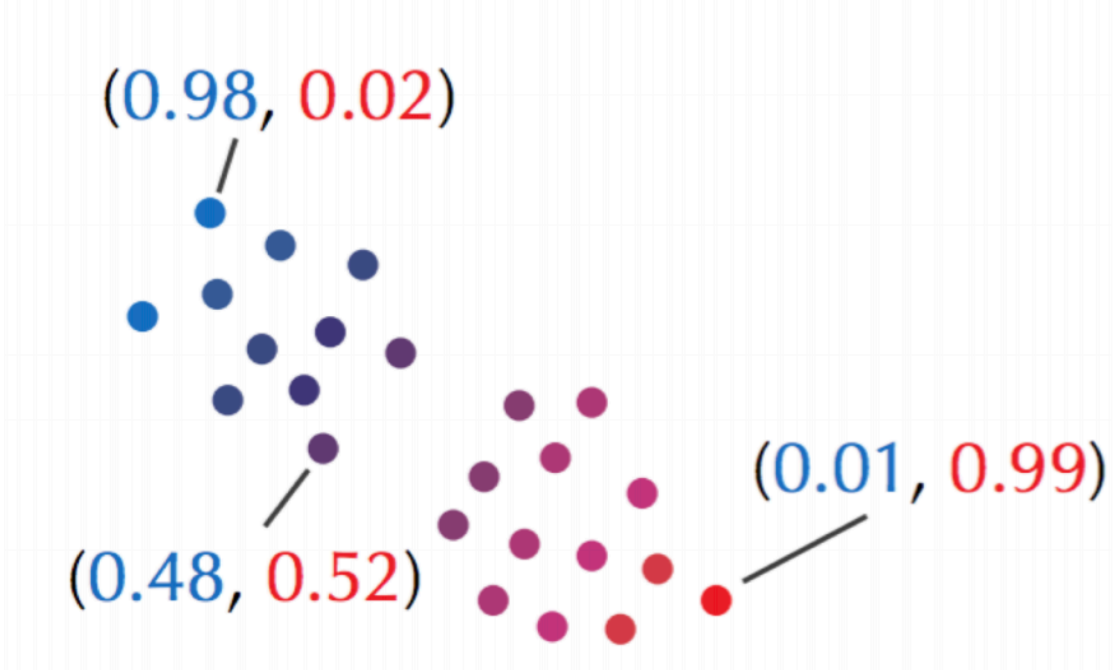
Soft clustering and its application

What about assigning clusters “softly”?

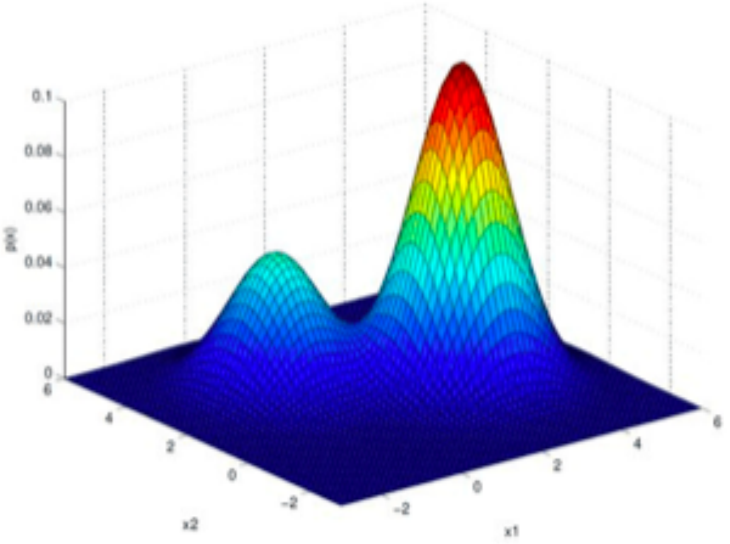
Hard clustering assignment:



Soft clustering assignment:

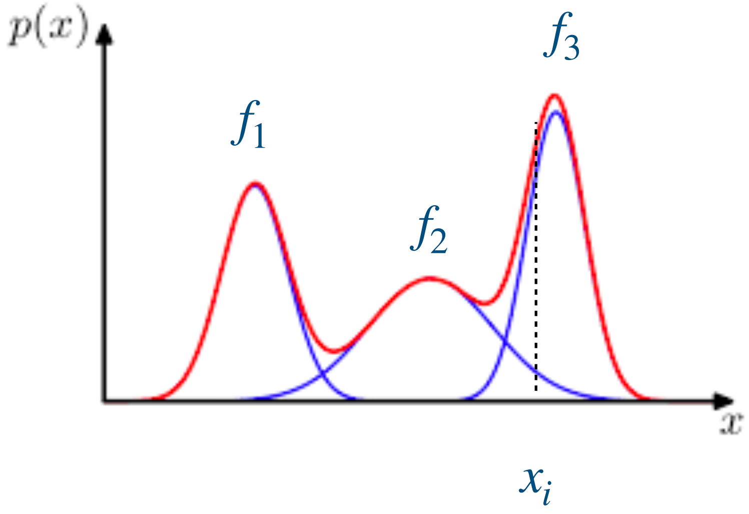


How?



GAUSSIAN
MIXTURE
MODEL

Soft clustering: Gaussian mixture model



$$\mathbb{P}(x_i \,|\, C_i = 3) = \frac{\pi_3 f_3(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i) + \pi_3 f_3(x_i)}$$

Formula for calculating the probability of point x_i assigned to the 3rd gaussian distribution, where f_k is the Gaussian pdf, π_k is the class specific weight.

- Randomly Initialize Gaussian distribution parameters (μ, σ^2).
- **E step**:

  Assign data points to each Gaussian distribution by **probabilities**.
- **M step**:

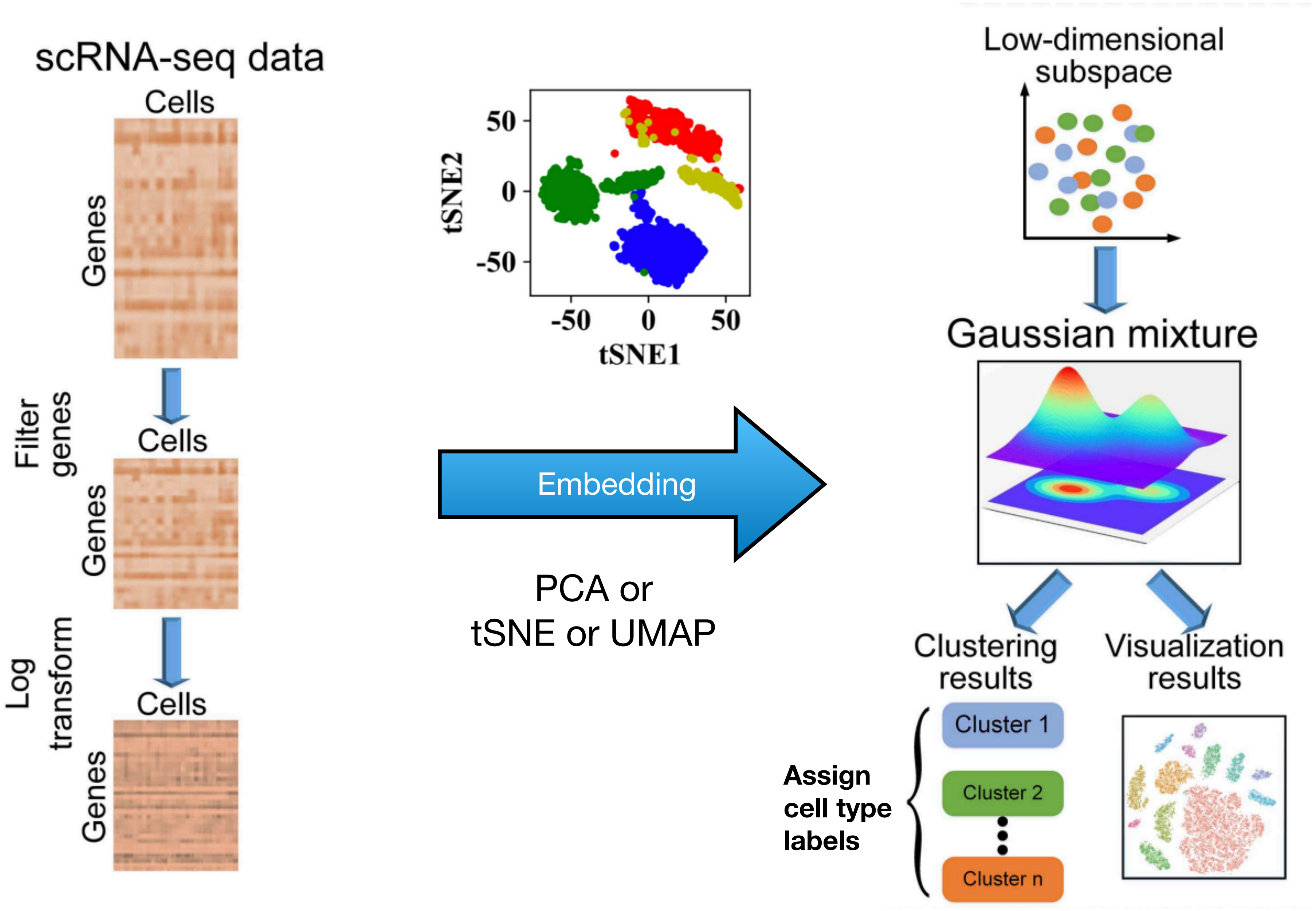
  Recalculate Gaussian distribution parameters (using weighted estimators).
- Repeat... until convergence.

Weighted mean: $\bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i}$;

Weighted variance: $s_w^2 = \frac{1}{d} \sum_i w_i (x_i - \bar{x}_w)^2$

Application of Gaussian mixture model

Cell types identification in scRNA-Seq



scRNA-seq data

Cells

Genes

Filter genes

Cells

Genes

Log transform

Cells

Genes

Embedding

PCA or
tSNE or UMAP

Low-dimensional subspace

Gaussian mixture

Clustering results

Visualization results

Assign cell type labels

Cluster 1

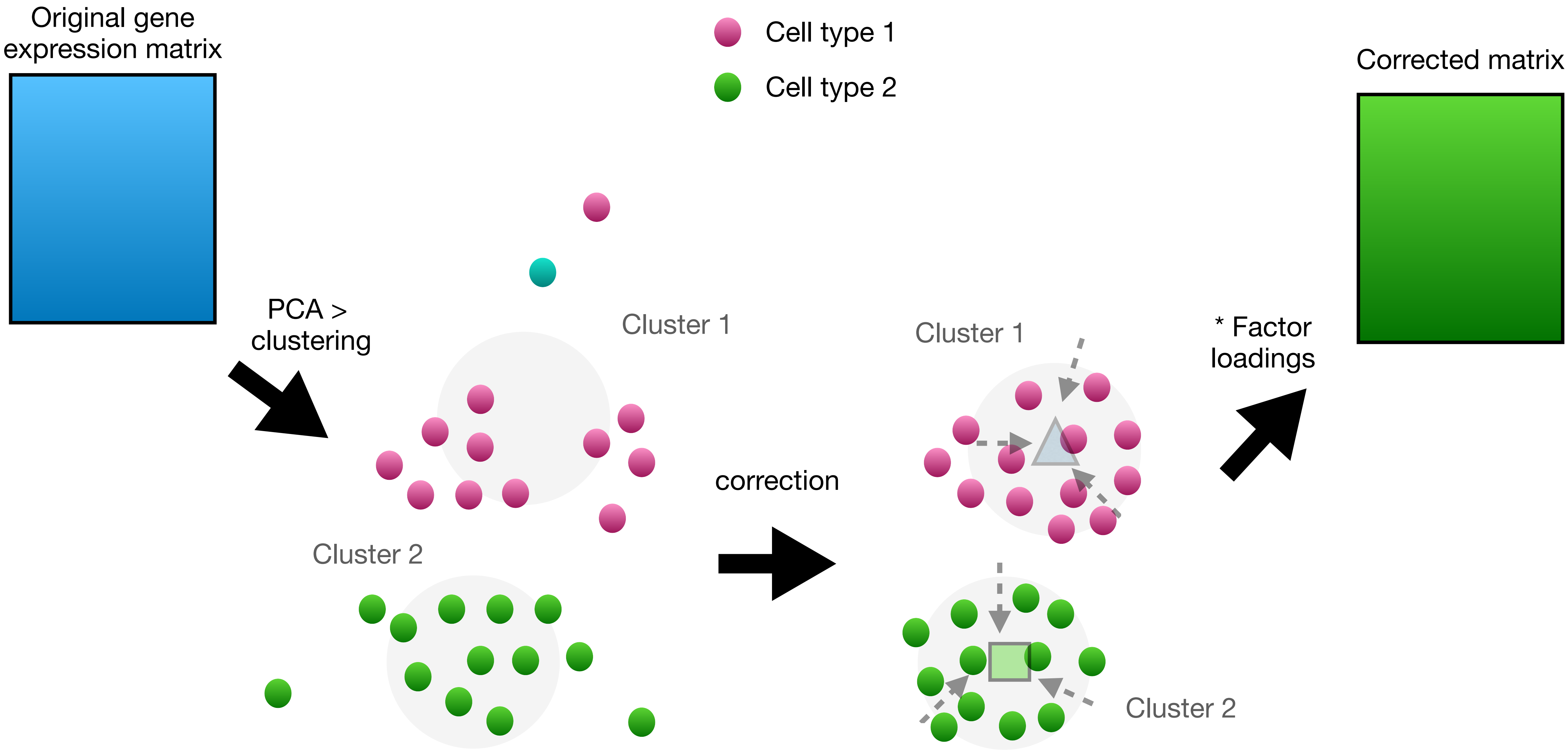
Cluster 2

Cluster n

- The dimensional reduction techniques are doing the “feature extraction” for clustering.

Batch effect correction in scRNA-Seq

Harmony

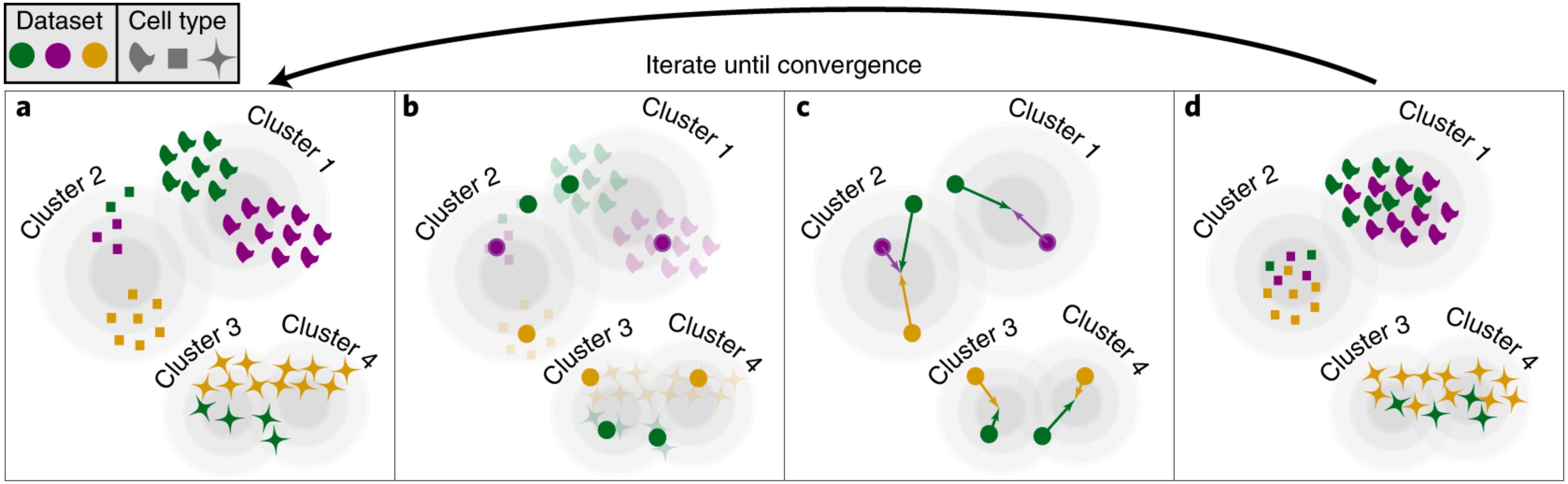


Dimensional reduced space

Correction by shrinking data points
toward clustering centroids
(in a way grouped by batches)

Korsunsky, Ilya, et al. "Fast, sensitive and accurate integration of single-cell data with Harmony." Nature methods 16.12 (2019): 1289-1296.

Harmony



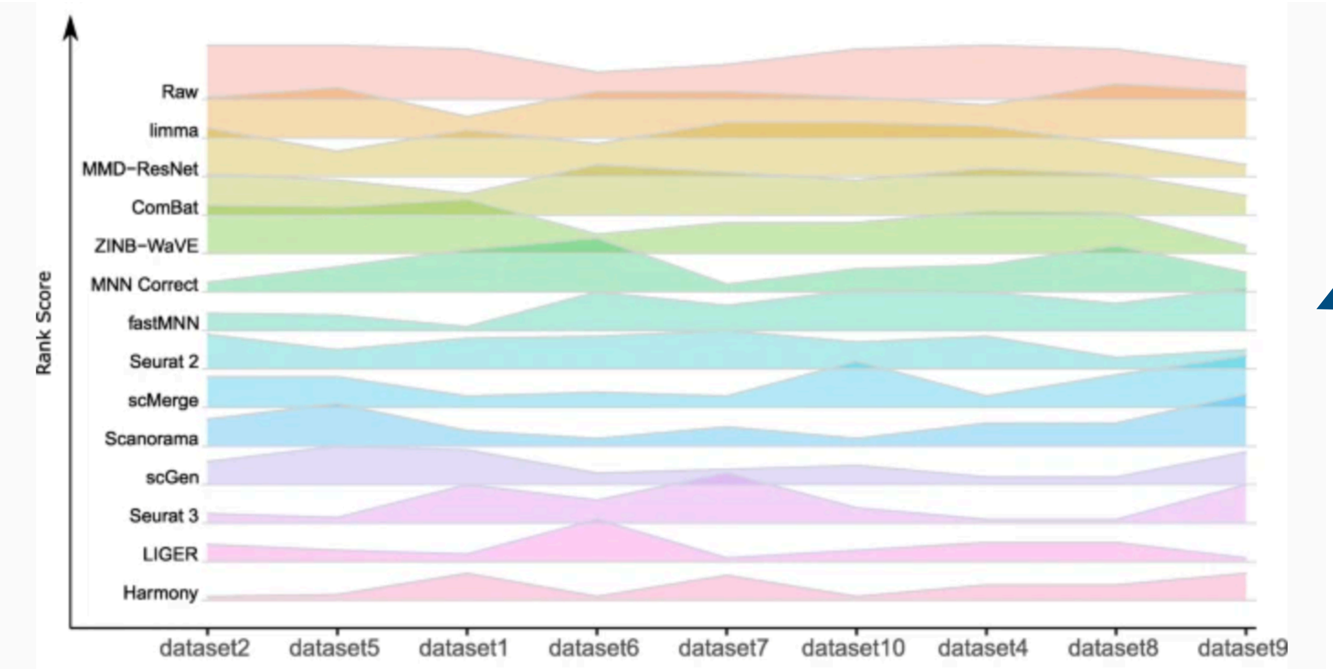
Soft assign cells to clusters, favoring mixed dataset representation

Get cluster centroids for each dataset

Get dataset correction factors for each cluster

Move cells based on soft cluster membership

Performance ranks in cell type identification



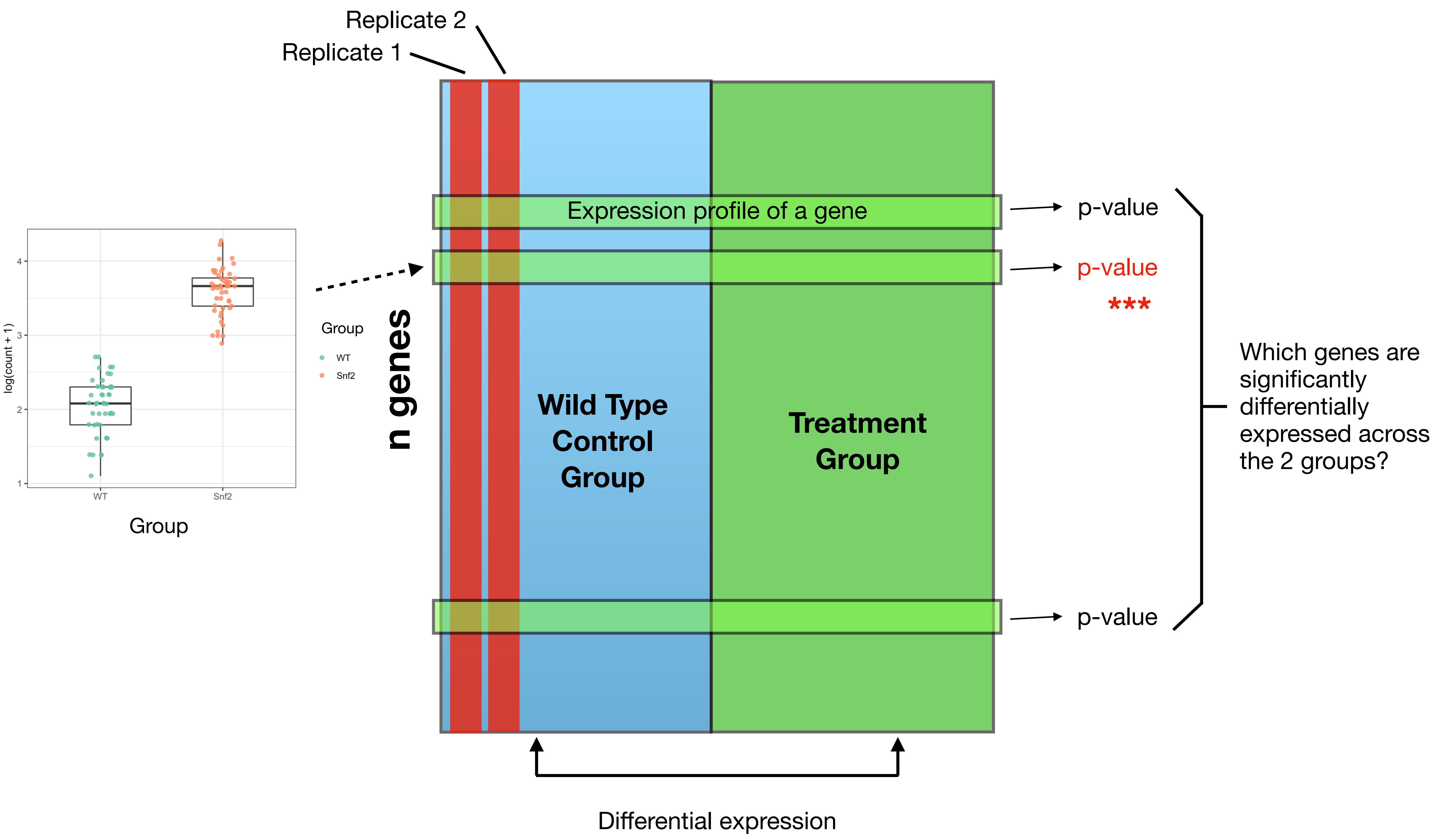
Based on 10 datasets, Harmony performed the best in cell type identification task over 14 batch effect correction methods for scRNA-Seq.

Tran, Hoa Thi Nhu, et al. "A benchmark of batch-effect correction methods for single-cell RNA sequencing data." *Genome biology* 21 (2020): 1-32.

Differential expression analysis

Differential expression analysis

m samples (biological replicates) for each group



Replicate 2

Replicate 1

n genes

Expression profile of a gene

Wild Type Control Group

Treatment Group

p-value

p-value

***

p-value

Which genes are significantly differentially expressed across the 2 groups?

Differential expression

Inference for differential expression

Challenges of data randomness

- Suppose we have 2 coins and we want to know if the probability of getting heads is different for the 2 coins.

- We first collect a data of coin tosses as the following:

	Head	Tail
Coin 1	2	2
Coin 2	3	1

2/4 is different from 3/4, but is the difference "significant" enough against the randomness?

- After collecting more outcomes, what conclusion can we draw now?

	Head	Tail
Coin 1	236	175
Coin 2	187	314

p-value: overcoming randomness

Alternative hypothesis H1:
Phead(coin1) ≠ Phead(coin2)

Null hypothesis H0:
Phead(coin 1) = Phead(coin2)

Coin 1

Coin 2

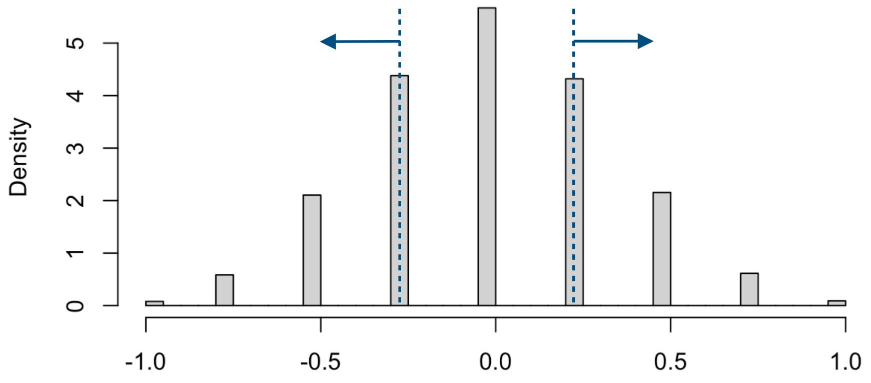
Data

	Head	Tail
Coin 1	2	2
Coin 2	3	1

Test statistics is:

2/4 - 3/4 = - 0.25

Distribution of the difference in proportions under the null assumption



p-value is:
Prob(-0.25 ≥ Null dist ≥ 0.25) = 0.715

- First, assume that there is no difference between 2 coins (null hypothesis).
- Then, calculate the probability of generating data as extreme or more extreme than the observed data under the null assumption.
- The calculated probability is called a p-value, and it can be obtained by either probabilistic modeling or simulation methods.
- If the p-value is sufficiently small (e.g. < 0.05), it indicates that the data reject the null assumption of no difference.

Statistical Modeling: rethinking of data

Well understood process in probabilistic world:

	Head	Tail
Coin 1	10	10000
Coin 2	25	15000

Can be used to represent

Data we want to model in practice:

	Read count on region x	Read count on other regions
NGS library 1	10	10000
NGS library 2	25	15000

Statistical Modeling: formulation

We could write it down in the statistical modeling terms:

count_1 ~ binomial(p = p_1, N = total reads count in library 1)

count_2 ~ binomial(p = p_2, N = total reads count in library 2)

Or equivalently:

count_1 ~ Poisson($\lambda = p_1 \times$ total reads count in library 1)

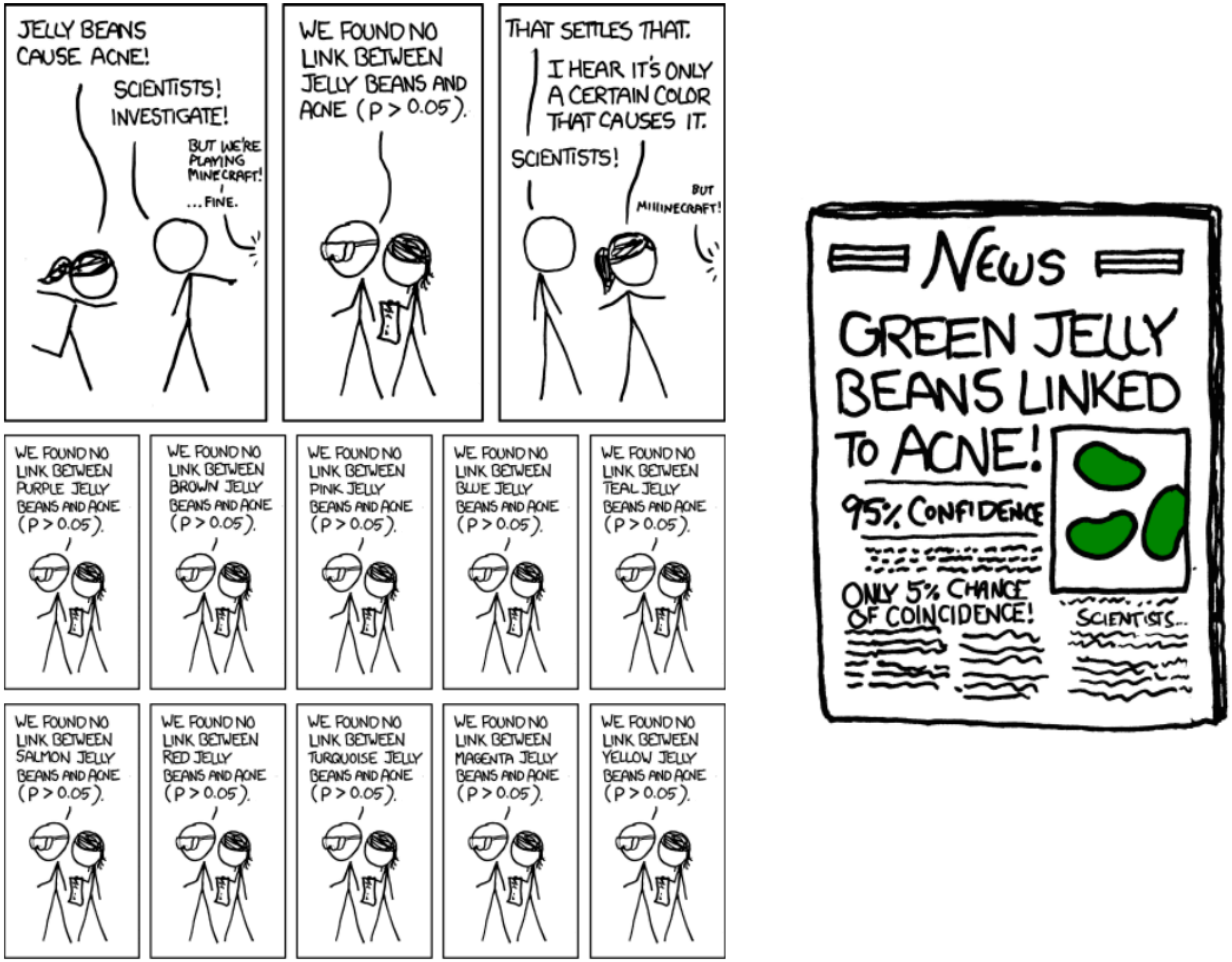
count_2 ~ Poisson($\lambda = p_2 \times$ total reads count in library 2)

$$H_0 : p_1 = p_2;\ H_1 : p_1 \neq p_2$$

- The hypothesis pair above can be evaluated by the exact test for binomial or Poisson (c-test).
- This test can be used in DGEA of RNA-Seq when there are no replicates available.
- It can also be used to determine the threshold for peak calling in CHIP-Seq using a control sample.

1st challenge of p-value: multiple hypothesis testing

Can Jelly Beans Cause Acne?


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Solution for the 1st challenge

Adjusted p-values for multiple hypothesis testing

- In differential gene expression analysis, multiple p-values are calculated over 20000 genes, therefore multiple hypothesis correction is necessary.

- Two metrics are often used:

1. **Family wised error rate (FWER)** controlled by **Bonferroni correction.**

2. **False discover rate (FDR)** controlled by **Benjamini-Hochberg correction.**

- Bonferroni corrected p-value is defined by **m × p-value**, where m is the total number of tests conducted (e.g. the # of genes in differential expression analysis).

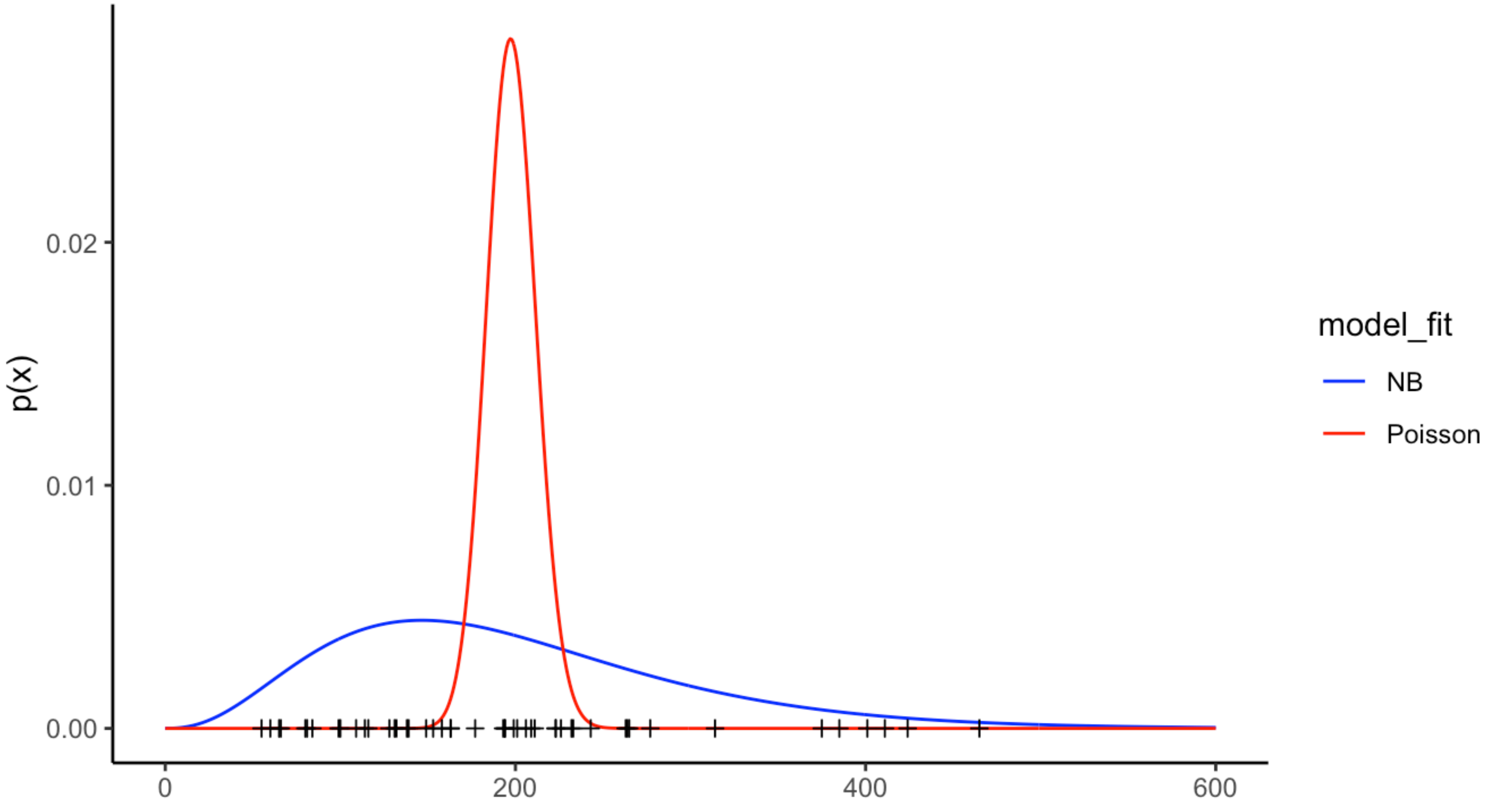
- Filtering Bonferroni corrected p-value at 0.05 ensures FWER < 0.05.

	Retain H0	**Reject H0**	
H0 True	a	b	a+b = m0
H1 True	c	d	c+d = m1
	a+c = n0	b+d = n1	a+b+c+d = m

$\text{FWER} := P(b>0|m)$

$\text{FDR} := b/n1$

2nd challenge of p-value: we may fail to define the randomness accurately



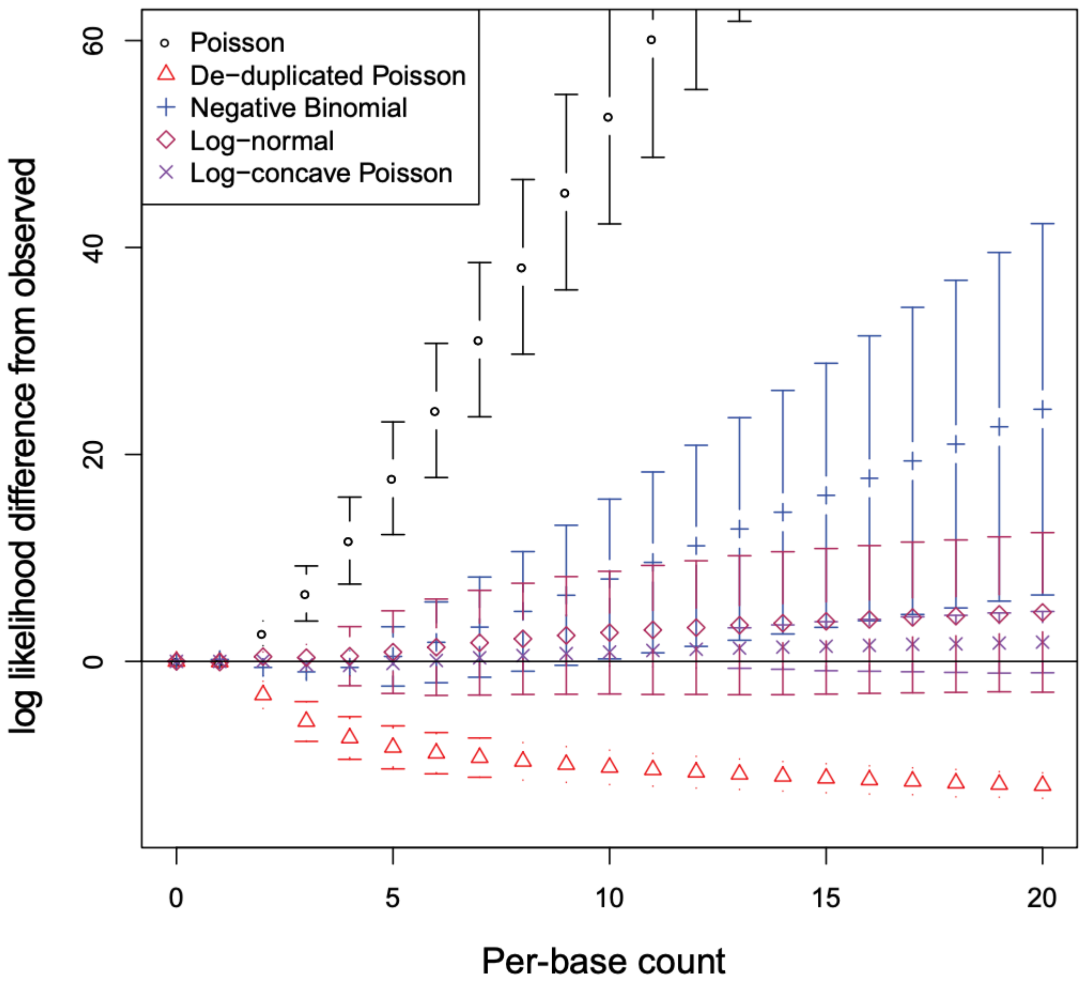
Read counts on Gene LSR1 over biological replicates

- In practice, the distribution of read counts across biological replicates follows a **negative binomial (NB) distribution** rather than a Poisson distribution.

- Many classic statistical models (e.g. Poisson/binomial models) fail to account for the over-dispersed nature of genomic count data.

Solution for the 2nd challenge

Selecting suitable statistical distribution for your data



- It is important to use a statistical model that **specify** the data, i.e. the model used should be able to generate the observed data under some parameterization.
- This can be done by examining the goodness of fits of different distribution families on the data; statistical test should be constructed using the best fitting distribution family.

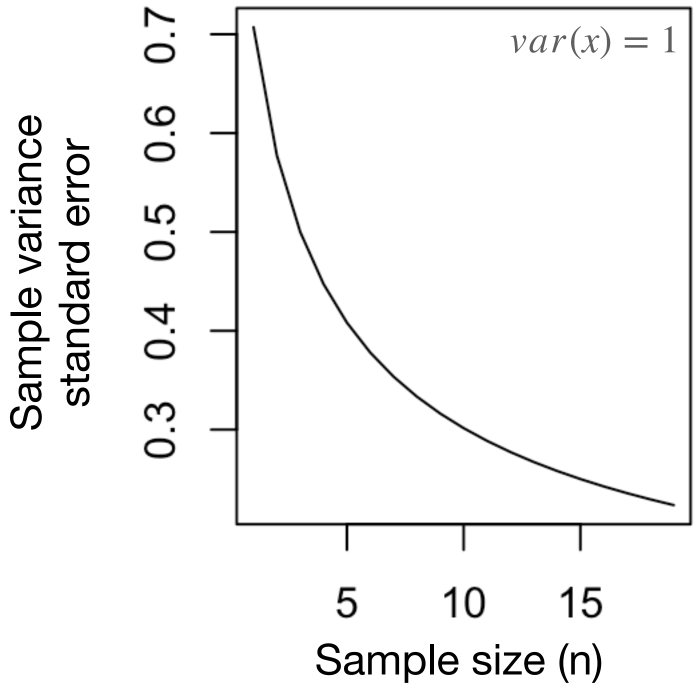
Hashimoto, Tatsunori B., Matthew D. Edwards, and David K. Gifford. "Universal count correction for high-throughput sequencing." PLoS computational biology 10.3 (2014): e1003494..

Summary of the commonly used statistical tests in genomics

Distribution family	Data type	Support	Statistical test	Application in genomics
Binomial or multinomial	Binary or categorical	$\{0,1,\cdots,n\}$	Fisher's exact test; Chi-squared test	Test for 2 X 2 contingency table; Gene set enrichment analysis; GWAS
Gaussian	Continuous	$[-\infty,+\infty]$	t-test (**limma**)	Differential expression analysis for micro-array data
Poisson	Count	$\{0,1,\cdots,+\infty\}$	Fisher's exact test; Exact binomial test	Differential analysis for NGS without biological replicates (e.g. peak calling)
Negative binomial	Count	$\{0,1,\cdots,+\infty\}$	NB test (**DESeq2, edgeR**)	Differential analysis for NGS with biological replicates (e.g. DGEA for RNA-Seq)

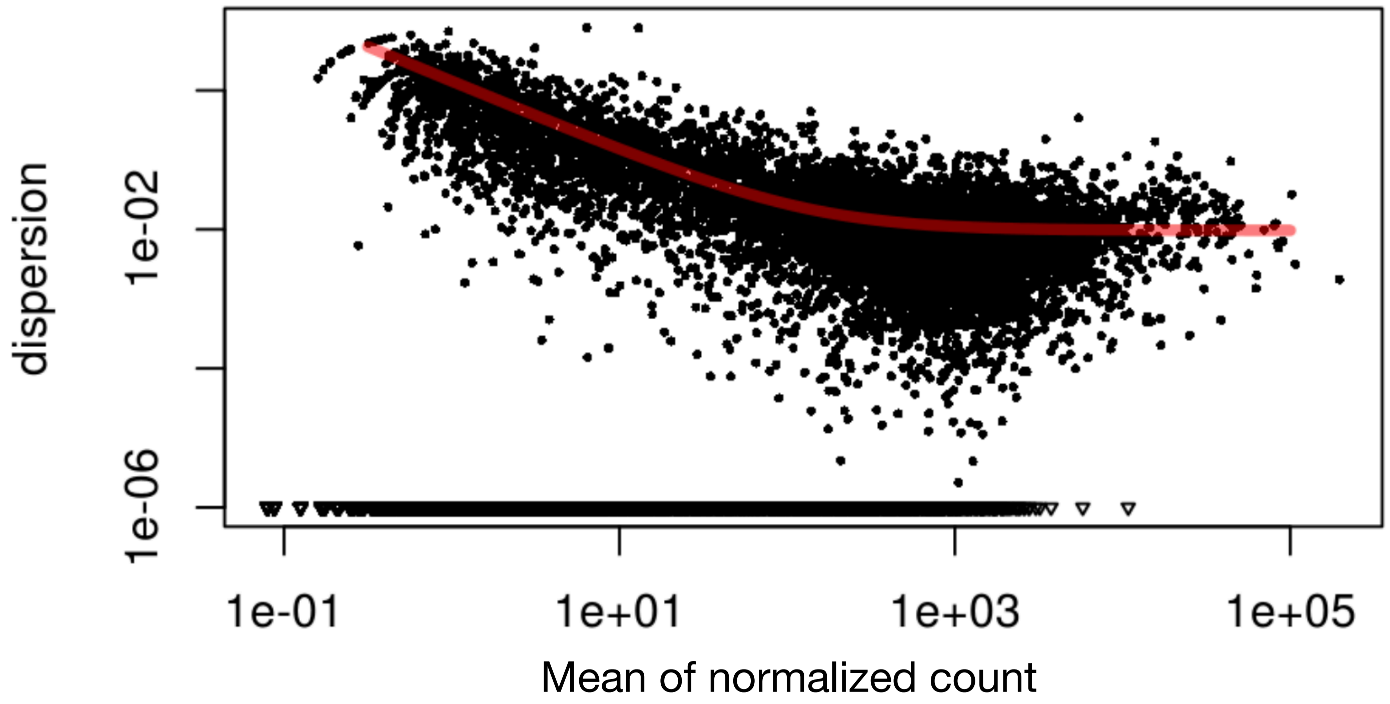
3rd challenge: limited sample size estimating gene variances

Uncertainty of variance estimation drops with sample size



Red: fitted regression curve, in which the fitted values will be used for differential testing

Black: estimated gene dispersions on limited samples



- Tests integrating multiple replicates require the estimation of dispersion parameters (e.g. Gaussian σ^2 and NB over-dispersion parameter).
- Many experiments only have 2 or 3 replicates, this is too few for accurate dispersion parameter estimation.
- One solution is to use a smooth curve to predict gene dispersions from gene means, which shares information between all genes. This approach is commonly used by DGEA packages such as Limma, EdgeR, and DESeq2.

Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." Nature Precedings (2010): 1-1.