Kernel-Based Testing for Single-Cell Data

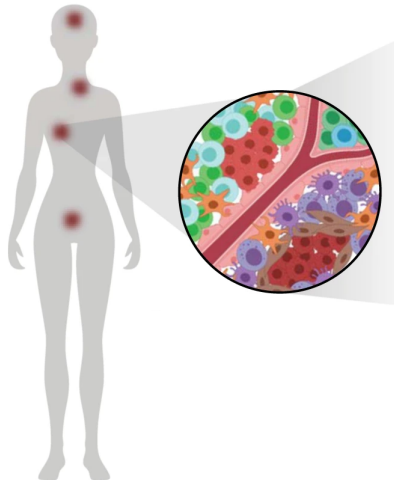
A. Ozier-Lafontaine[†], P. Lacroix[⋆], B. Samson[⋆], P. Artenseva[†,⋆],
V. Rivoirard[‡], B. Michel[†], F. Picard[⋆]

[†]Laboratoire de mathématiques Jean Leray, Nantes
[‡]Centre de Recherche en Mathématiques de la Décision, Paris Dauphine
[⋆]Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

Outline

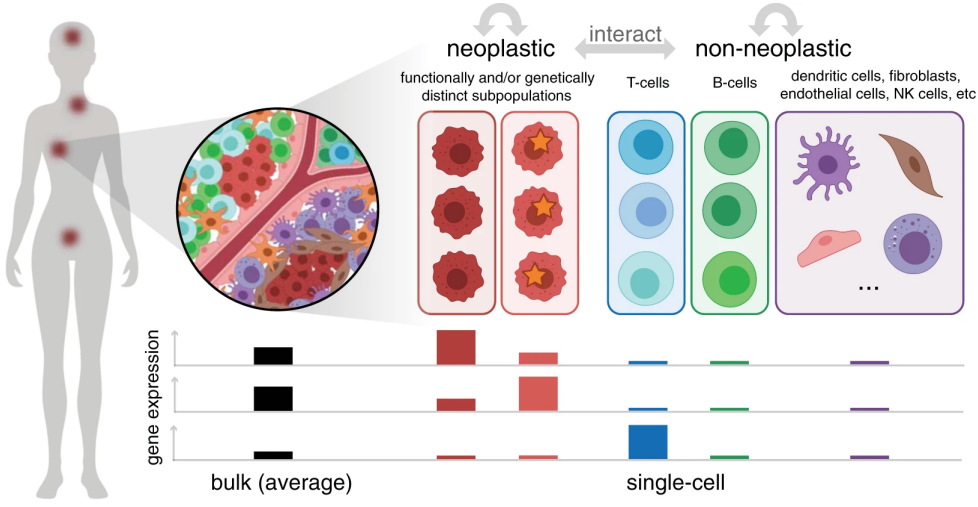
From molecular to cellular variability

- Convergence between cell biology & high-throughput sequencing
- Complexity of defining "cell-types"
- What part of the cellular variability is explained by the molecular variability ?



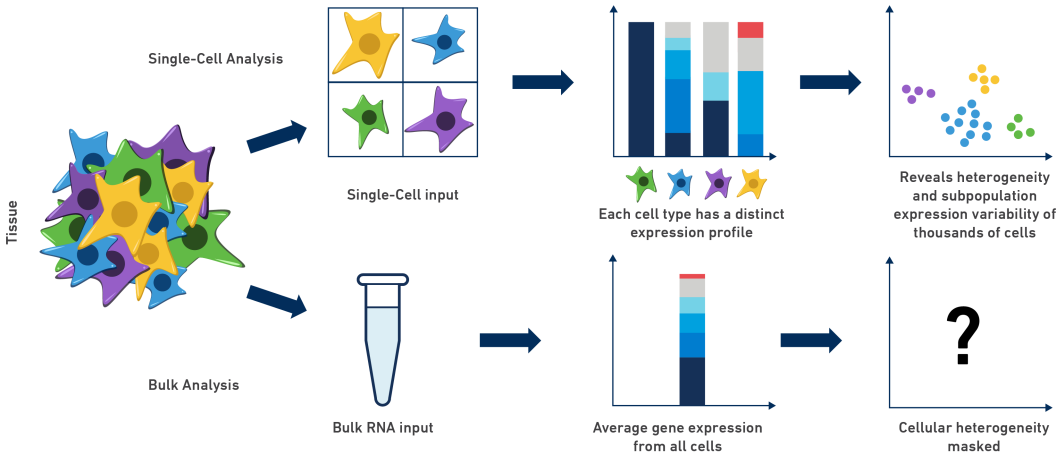
[2]

From bulk to distributions of gene expression



[2]

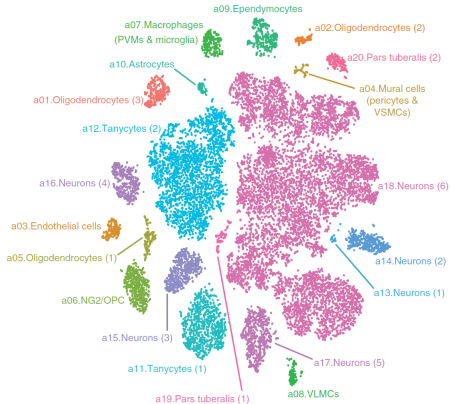
Single-Cell from a statistician's perspective



From 10X Genomics

Machine Learning challenges

- Dimension Reduction / Visualization
- Clustering cell-type discovery
- Datasets alignments
- Cell-cell communication
- Data integration
- Differential analysis

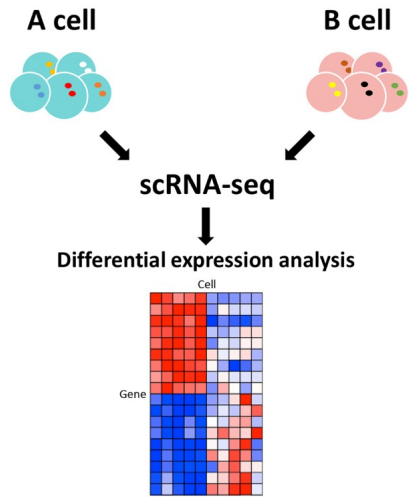


[1]

Outline

Differential Expression Analysis

- Compare the expression of each genes between 2 or more conditions
- Task: Statistical Testing
  $\rightarrow$ compute the difference
  $\rightarrow$ compute a risk
- Single-cell data $n \sim 100 - 10,000$
- How to fully exploit the potential of single-cell assays ?



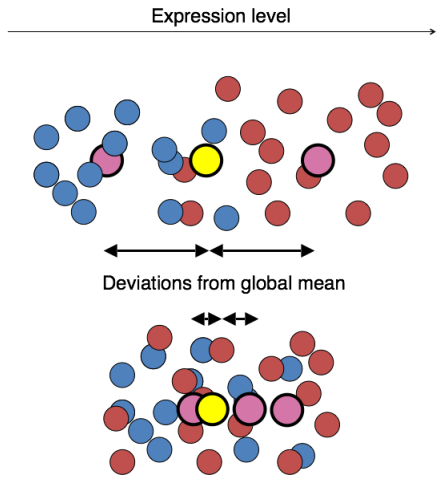
Two-sample test basic ingredients

- Consider the expression of one gene
- $X_{1,i}$ expression in condition 1 for cell i
- $X_{2,i}$ expression in condition 2 for cell i

$$\mathbb{E}(X_{1,i}) = \mu_1, \quad \mathbb{E}(X_{2,i}) = \mu_2$$

- Variability of gene expression

$$\mathbb{V}(X_{1,i}) = \mathbb{V}(X_{2,i}) = \sigma^2$$

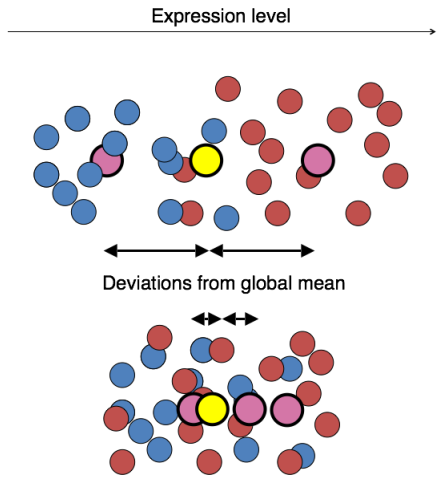


Two-sample hypothesis testing

- Statistical testing of $\mathcal{H}_0 : \left\{ \mu_1 = \mu_2 \right\}$
- Use the concept of Signal-to-Noise Ratio

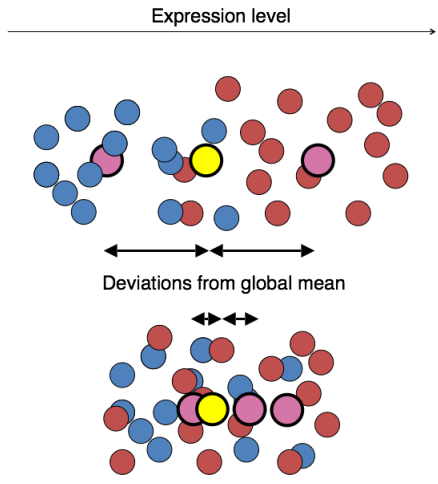
$$\text{SNR}^2 = \left(\frac{\mu_1 - \mu_2}{\sigma / \sqrt{n}} \right)^2$$

- Also called log-fold change on the log-scale



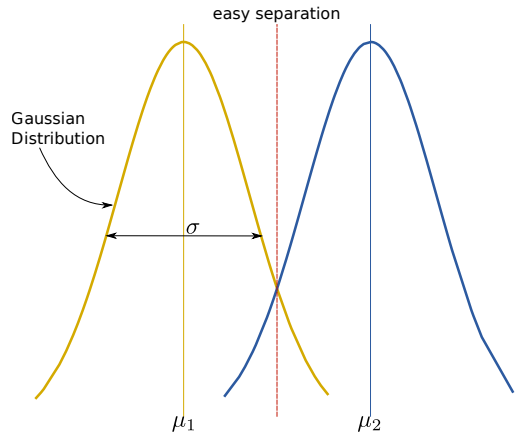
Two-sample hypothesis testing procedure

- Is the observed logFC high under the null hypothesis of no difference ?
- If High the data do not support the hypothesis $\mathcal{H}_0$
  $\rightarrow$ Reject $\mathcal{H}_0$
- Compute the *p*-value
  $\rightarrow$ proba. of observing the data if $\mathcal{H}_0$ were true
- Reject if the *p*-value $< \alpha$



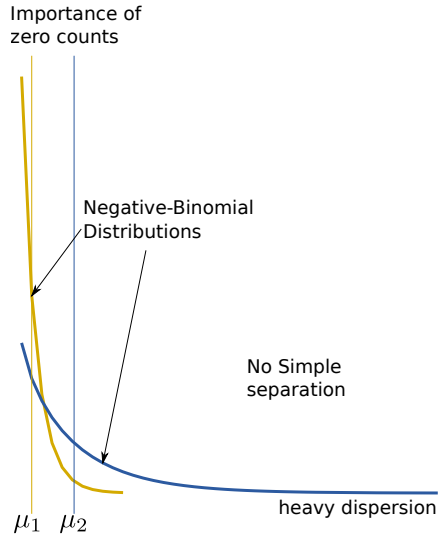
Statistical Setting: two-sample test

- logFC are valid provided μ and σ are good summaries of the information
- Easy linear separation
- Not adapted to single-cell assays

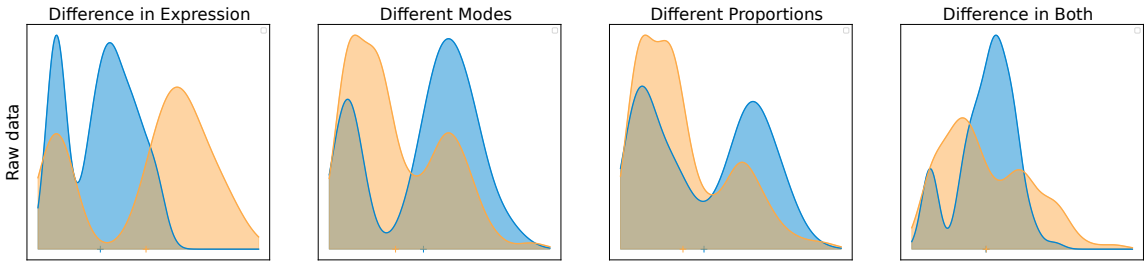


sc-RNAseq data are count data

- Specificities: discrete, zeros
- How to define the signal-to-noise ratio ?
- Standard: Negative Binomial distribution
- No simple linear separation

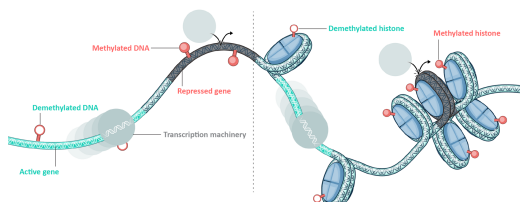


sc-RNASeq are complex distributions



What about other single-cell data ?

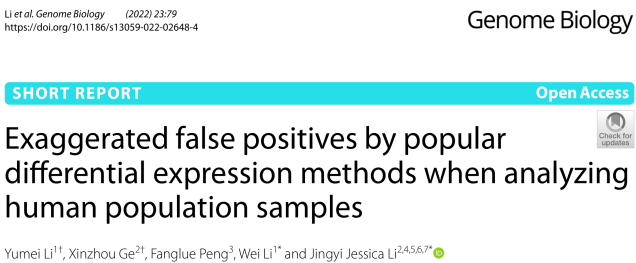
- Single-Cell ChipSeq has become popular
- Map binding sites in population of cells
- Differential Analysis is also a challenge
- Should we build a new reference model for each single-cell assay ?



https://tunetx.com/

Why is statistical modeling so important ?

- Much energy has been spent to understand the distribution of sc-RNASeq data
- Statistical testing is based on what is expected under $\mathcal{H}_0$


Li *et al. Genome Biology* (2022) 23:79
https://doi.org/10.1186/s13059-022-02648-4

Genome Biology

SHORT REPORT Open Access

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li[1†], Xinzhou Ge[2†], Fanglue Peng[3], Wei Li[1*] and Jingyi Jessica Li[2,4,5,6,7*]

$\rightarrow$ Risk: detect a difference whereas the appropriate model there would not

Take-Home Message Slide (1)

- ✓ Single-cell data are complex distributions
- ✓ the logFC may not be adapted to every situation
- ✓ Only based on summary statistics
- ✓ A dedicated framework is required to perform differential analysis based on distributions

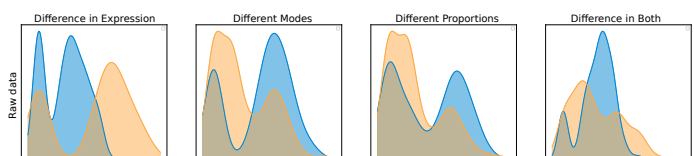
Outline

How to compare complex distributions ?

- Consider that $X_{i,1} \sim \mathbb{P}_1$, $X_{i,2} \sim \mathbb{P}_2$, such that $\mathbb{P}_1$ and $\mathbb{P}_2$ are unknown
- $\mathbb{P}_1, \mathbb{P}_2$: gene expression distribution across cells
- Single-cell differential expression can be tested using:

$$\mathcal{H}_0 : \left\{ \mathbb{P}_1 = \mathbb{P}_2 \right\}$$

- How to construct a powerful and calibrated test ?



Re-interpreting the Signal-to-Noise Ratio

- Consider the Signal to Noise Ratio for aggregated (bulk) data:

$$\text{SNR}^2 \propto \frac{\Big(\mu_1 - \mu_2\Big)^2}{\sigma^2} = \frac{\text{Distance between averaged populations}}{\text{variability}}$$

- The signal has too parts:

$$\Big(\mu_1 - \mu_2\Big)^2 = \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2$$

- Intensity of expression in each group :

$$\mu_1^2 + \mu_2^2$$

- Distance between averaged groups

$$\mu_1\mu_2$$

- Pseudo Bulk Analysis (average single cell data)

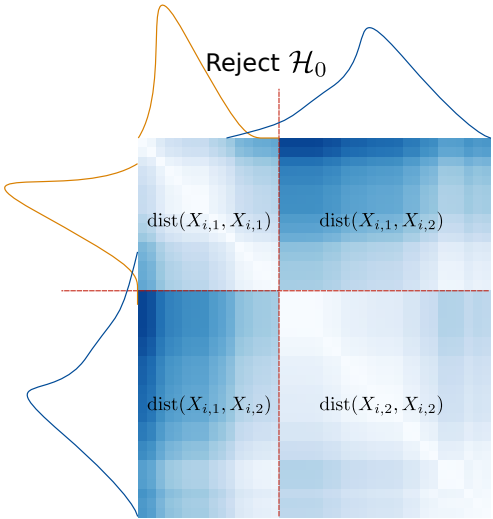
Pair-Wise Distances in Single-Cell Assays

- Provides much more information : pair-wise distances between individual cells
- Intra-condition distances

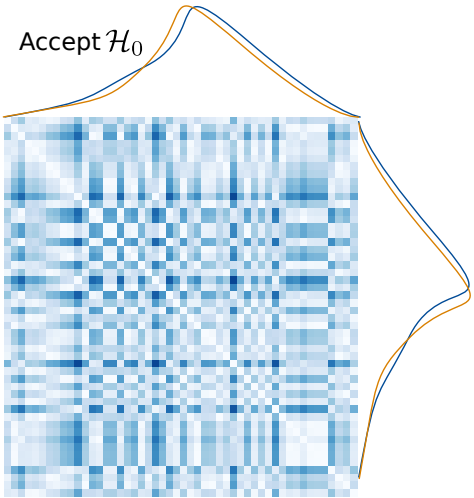
$$\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{i'=1}^{n_1}\text{dist}(X_{i,1}, X_{i',1}) \quad \text{and} \quad \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{i'=1}^{n_2}\text{dist}(X_{i,2}, X_{i',2})$$

$\rightarrow$ If small, conditions are homogeneous

Statistical Testing with pair-wise distances



$\Sigma_{\text{Within}} \ll \Sigma_{\text{Between}}$



$\Sigma_{\text{Within}} \gg \Sigma_{\text{Between}}$

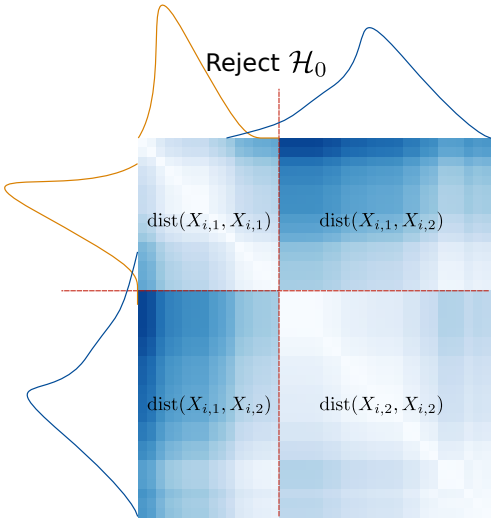
Pair-Wise Distances in Single-Cell Assays

- Provides much more information : pair-wise distances between individual cells
- Inter-condition distance

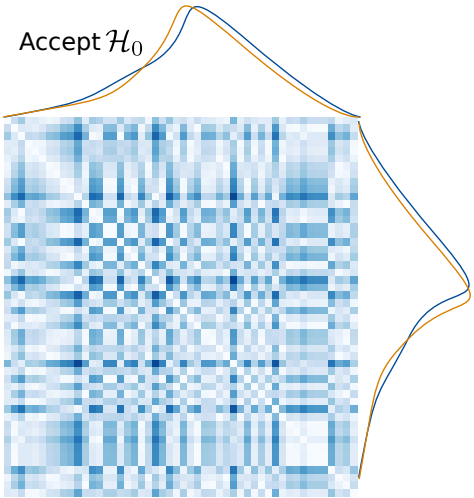
$$\frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{i'=1}^{n_2}\text{dist}(X_{i,1}, X_{i',2})$$

$\rightarrow$ If high, conditions are well separated

Statistical Testing with pair-wise distances



$\Sigma_{\text{Within}} \ll \Sigma_{\text{Between}}$



$\Sigma_{\text{Within}} \gg \Sigma_{\text{Between}}$

Intra-Inter trade-off

- Separated Conditions:

$$\Sigma_{\text{Within}} \ll \Sigma_{\text{Between}}$$

- Similar conditions :

$$\Sigma_{\text{Within}} \sim \Sigma_{\text{Between}}$$

- Construct the discriminant ratio

$$R = \Sigma_{\text{Within}}^{-1} \Sigma_{\text{Between}}$$

- Investigate the variations of the ratio under $\mathcal{H}_0$

Take-Home Message Slide (2)

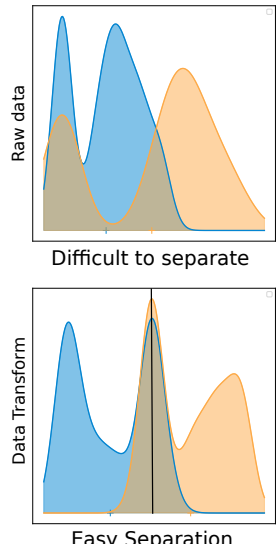
- ✓ Standard Differential Expression procedures can be applied by averaging data (pseudo bulk)
- ✓ Propose tests based on distributions comparisons
- ✓ Use pair-wise distances as a metric between distributions
- ✓ Use the discriminant ratio as a statistic

Outline

What if the distributions are difficult to discriminate ?

- sc-RNASeq distributions are complex
- Separation between condition is difficult
- Requires non-linear methods
- Could we find a transform such that they become easy to separate ?

→ This is possible thanks to kernel embedding

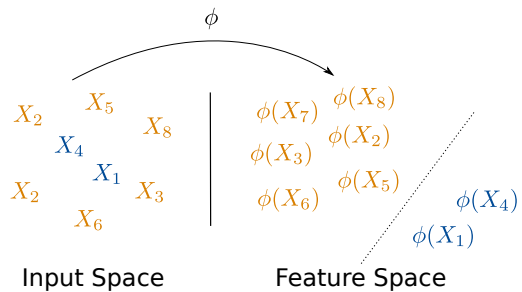


What is an embedding ?

- An embedding is a transformation of the data

$$X_i \to \phi(X_i)$$

- Easy separation after transformation
- Very popular for dimension reduction $\to$ UMAP, tSNE
- How to choose ϕ ?

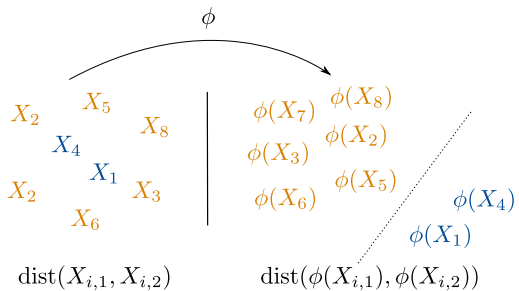


What is a kernel (in one slide)?

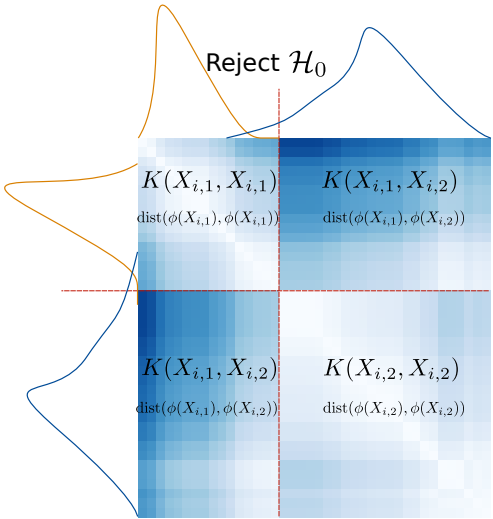
- When data are not separable
- dist. in the input space won't work
- dist. in the feature space could work !
- Kernel: distance between embeddings

$$K(X_{i,1}, X_{i,2}) = \text{dist}\Big(\phi(X_{i,1}), \phi(X_{i,2})\Big)$$

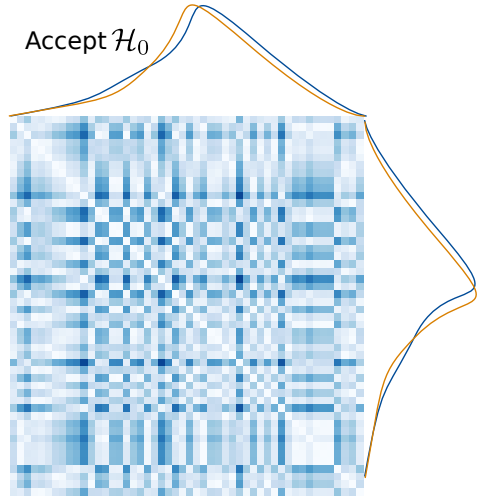
- Can work with any input data



Like a "Kernel Testing" Spirit



$\Sigma^{\phi}_{\text{Within}} \ll \Sigma^{\phi}_{\text{Between}}$



$\Sigma^{\phi}_{\text{Within}} \gg \Sigma^{\phi}_{\text{Between}}$

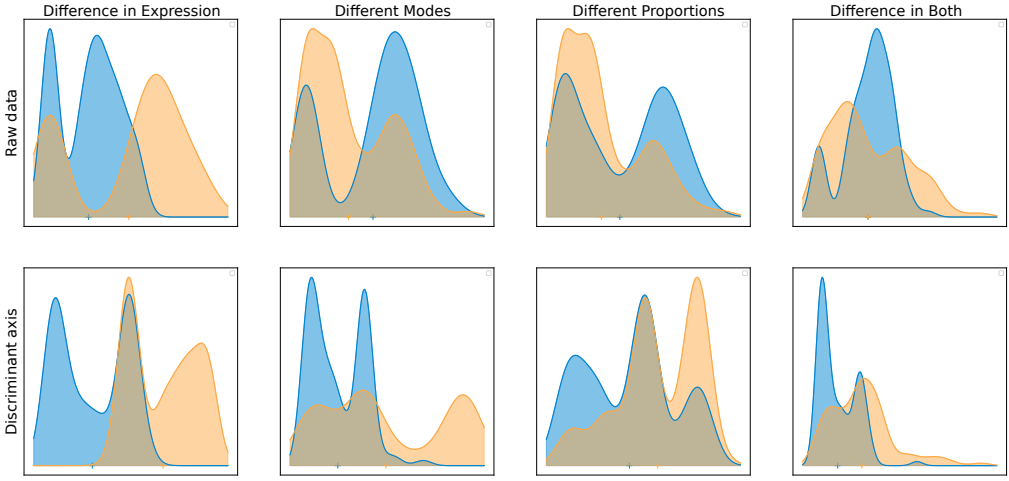
How to choose the kernel ?

- Kernel trick : no need to choose ϕ, only the kernel is necessary
- Popular kernel : Gaussian kernel

$$K(X_{i,1}, X_{i,2}) \propto \exp\left\{-\frac{1}{2}\left(\frac{X_{i,1} - X_{i,2}}{h}\right)^2\right\}$$

- Kernel trick : when you define a kernel you define the transform ϕ implicitly
- It can be considered as a non linear metric between distributions

Kernel Embedding separates complex distributions



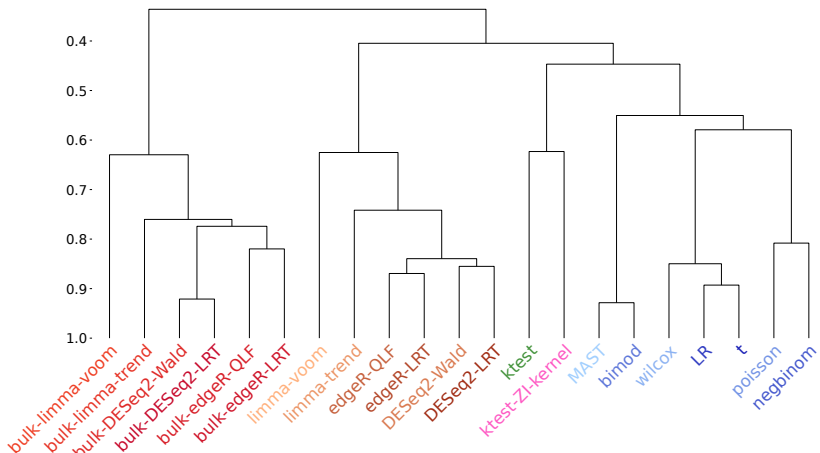
Take-Home Message Slide (3)

- ✓ Transform the data using an embedding
- ✓ Compute the pair-wise distances between embeddings
- ✓ The kernel is a non linear distance between distributions
- ✓ *A Kernel Two Sample Test* : 2012 paper, $>$ 5000 citations !

Outline

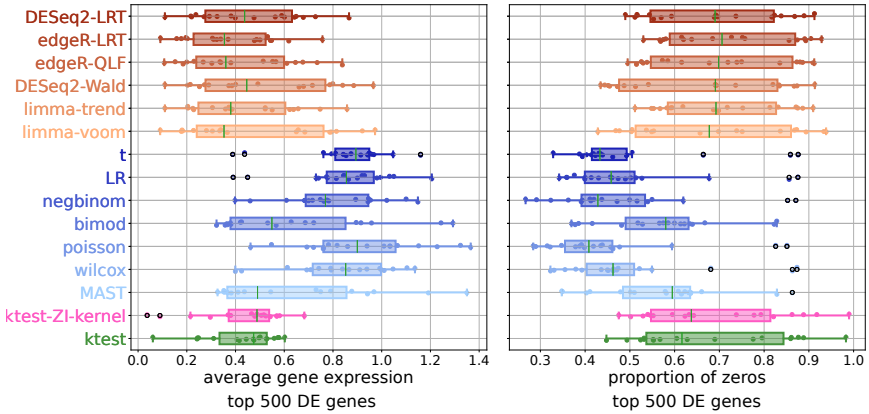
Methods comparison on experimental datasets

- 18 published datasets [3] / 20 methods
- Compare AUCCs based on reference gene lists



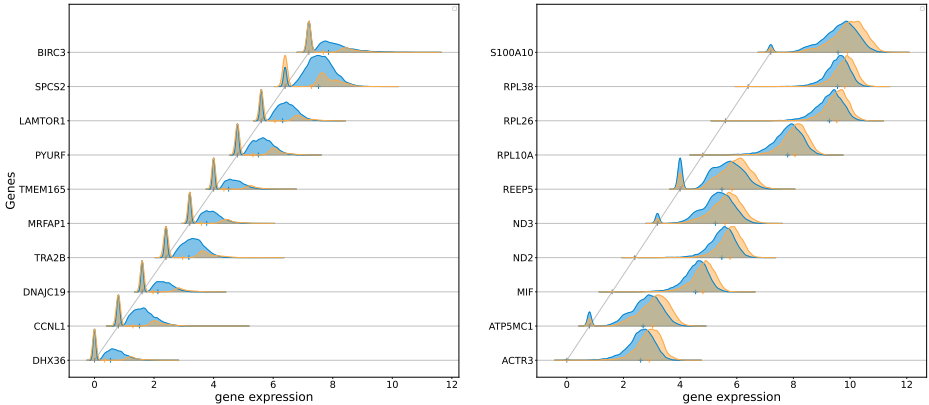
Methods comparison on experimental datasets

- 18 published datasets [3] / 20 methods
- Check the summary statistics characteristics of rejected distributions



Methods comparison on experimental datasets

- 18 published datasets [3] / 20 methods
- Check distribution forms of rejected hypothesis



Non DE in pseudo Bulk - Non DE in scDEA methods

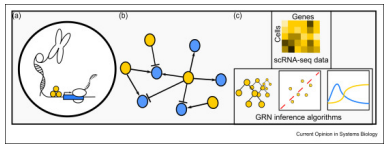
Take-Home Message Slide (4)

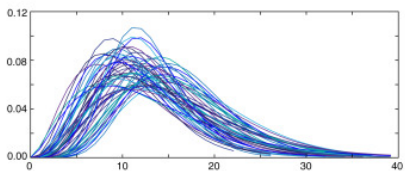
- ✓ Kernel testing is powerful and calibrated on experimental data
- ✓ Kernel testing does not share the same bias as classical DEA methods
- ✓ Kernel testing identifies complex distribution changes

Outline

Strong dependencies and lots of data

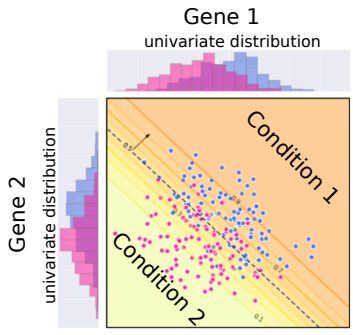
- Gene Expressions are highly dependent
- Account for GRN
- Could Differential Analysis be done on a set of genes ? whole transcriptome ?
- Kernels can be generalized to whole transcriptomes



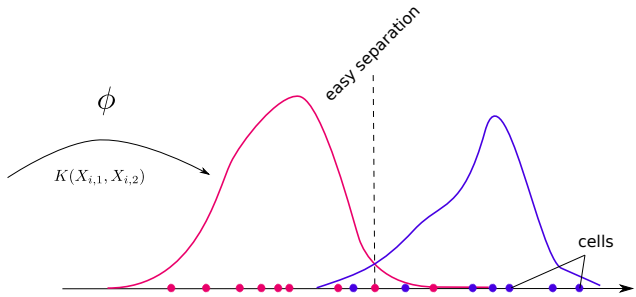


Distribution of gene expression across cells

Transcriptomic Differential Analysis



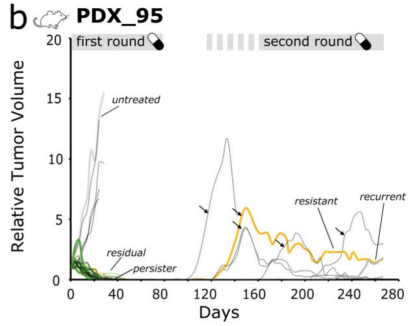
Joint distribution
Input Space

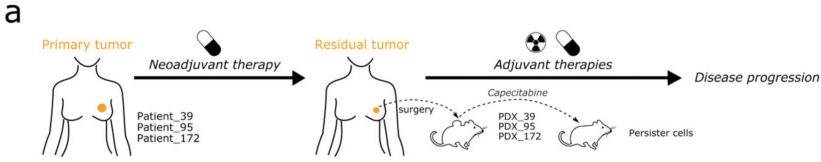


Summary Representation
Feature Space

ChemoResistance in Triple Negative Breast Cancer

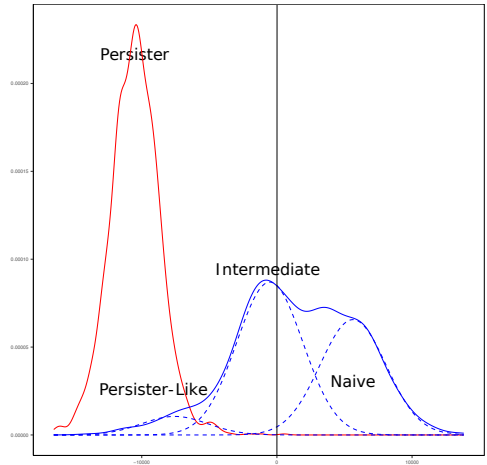
- Emergence of resistant phenotypes is a multi-step process
- After drug insult only a pool of drug-tolerant persister cells manage to tolerate the treatment and survive.
- Reservoir from which drug-resistant cells can ultimately emerge.





Kernel testing on Persister vs. Naive cells

- Persister cells survived the first treatment
- Reservoir for resistant cells
- Epigenomic data: 6376 features
- Compare untreated ($\sim$ 3000 cells) vs. persister ($\sim$ 2000 cells)
- Did we identify the reservoir of persister cells based on their epigenomic signatures ?



Summary of Whole Epigenome differences

Take-Home Message Slide (5)

- ✓ Differential Analysis can be performed on sets of genes or whole transcriptomes
- ✓ Accounts for dependencies between gene expressions
- ✓ Kernel methods can be easily adapted
- ✓ Allows the identification of sup-population of cells

Outline

What are the specific challenges ?

- Being understood by an audience of biologists
- Waiting for the editorial decision !
- Computing the risk of the procedure accurately
  🌶🌶🌶 (super)-hot topic in machine learning
- Generalize the approach to spatial transcriptomics
  🌶🌶🌶 (super)-hot topic in sc-data analysis

Check out !

- The `ktest` package ! Python-R,

  `https://github.com/AnthoOzier/ktest`

- The arxiv preprint

  `https://arxiv.org/abs/2307.08509`

References

[1] J. N. Campbell, E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, E. D. Rosen, B. B. Lowell, and L. T. Tsai. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.*, 20(3):484–496, Mar 2017.

[2] J. Fan, K. Slowikowski, and F. Zhang. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med*, 52(9):1452–1465, Sep 2020.

[3] J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. E. Matson, Q. Barraud, A. J. Levine, G. La Manno, M. A. Skinnider, and G. Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, Sept. 2021.