

Kernel-based perturbation testing for single-cell data

Franck Picard

Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

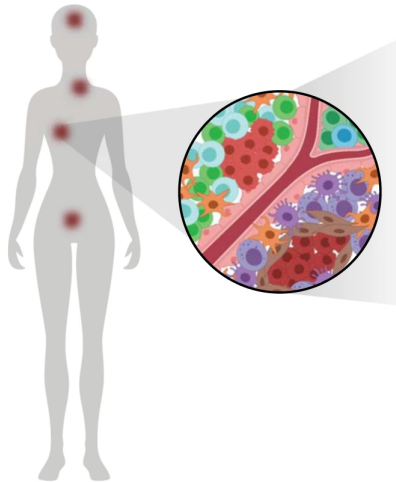


Outline

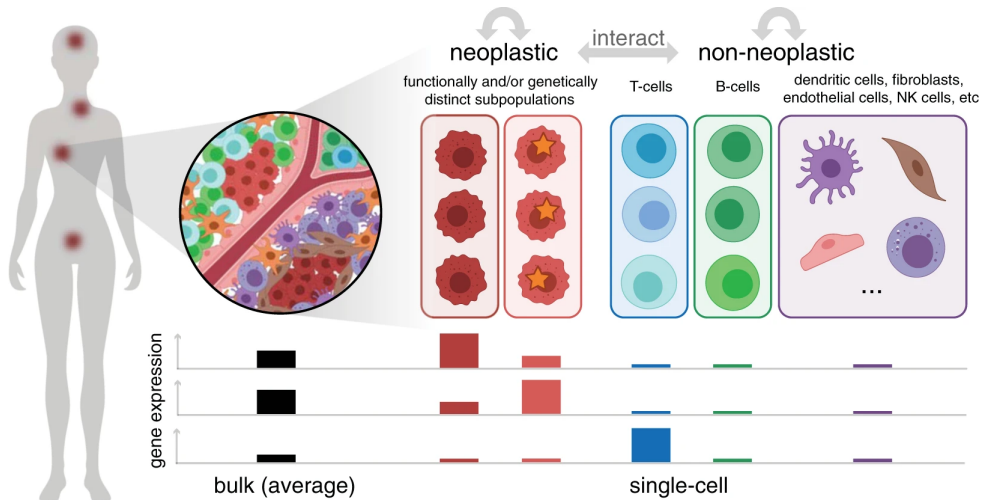
- 1. The Single-Cell Revolution**
2. Comparison of Gene Expression Distributions
3. Introduction to kernel testing
4. Discussion about methods
5. Towards perturbation analysis

The cellular scale of biology

- Cells are the basic unit of structure and function in living organisms
- Cells are characterized by their 'types' that are diverse
- Physiology emerge thanks to complex interactions between different cell-types

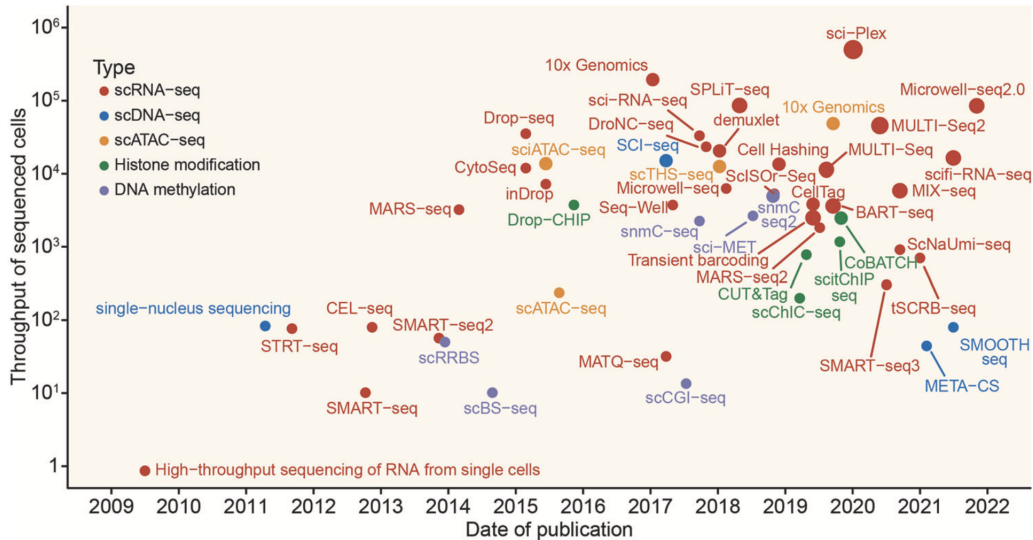


From bulk to distributions of gene expression



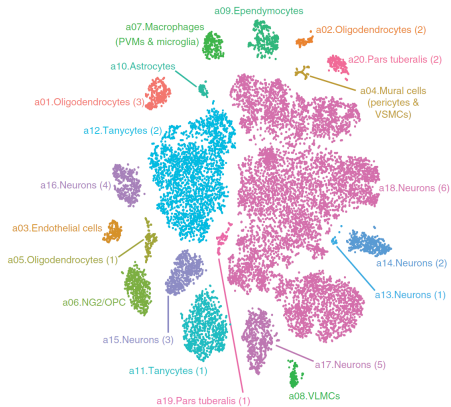
[2]

A timeline: produced data



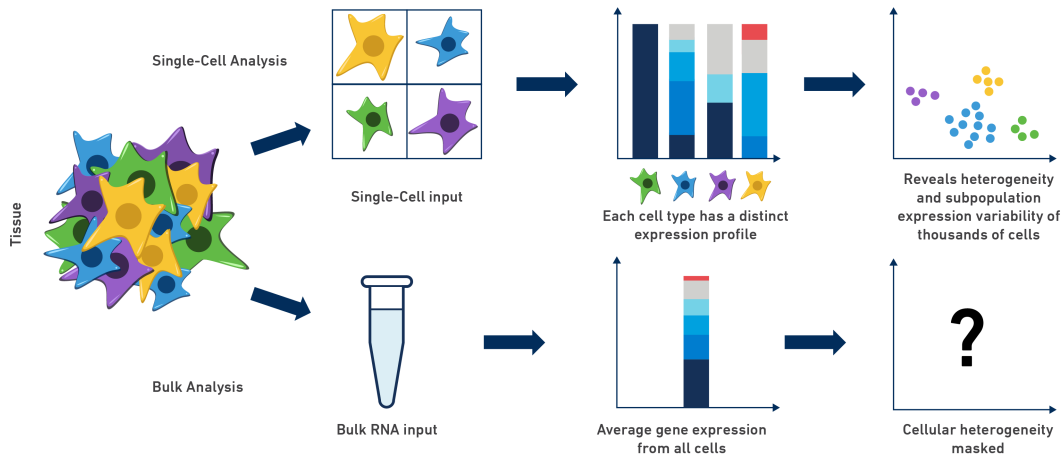
Machine Learning challenges

- Dimension Reduction / Visualization
- Clustering cell-type discovery (non supervised and semi supervised)
- Datasets alignments for non-matched samples
- Catch cells-ecosystems behaviors
- Simulation of fake data
- Data integration
- Statistical Testing (compare genes expression)



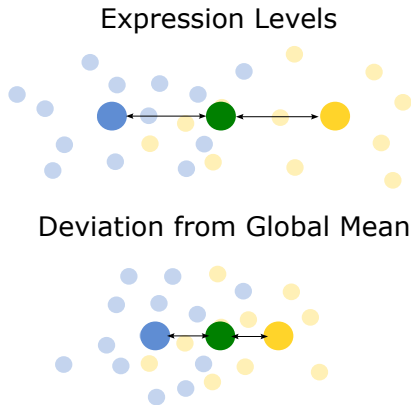
[1]

Single-Cell from a statistician's perspective



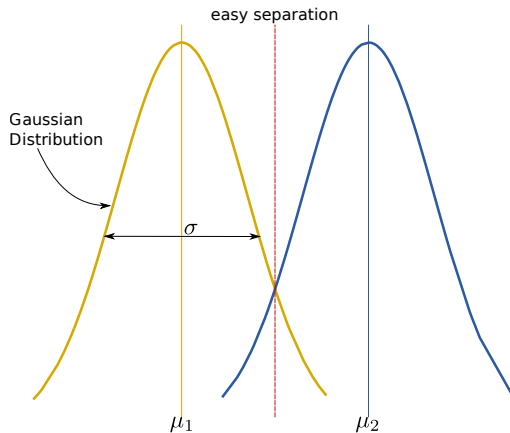
Differential Expression Analysis

- Compare the expression of each gene between conditions
- Statistical Testing
 - compute the difference
 - control type-I errors
- Single-cell data $n \sim 10^6$
- Try non-parametrics !



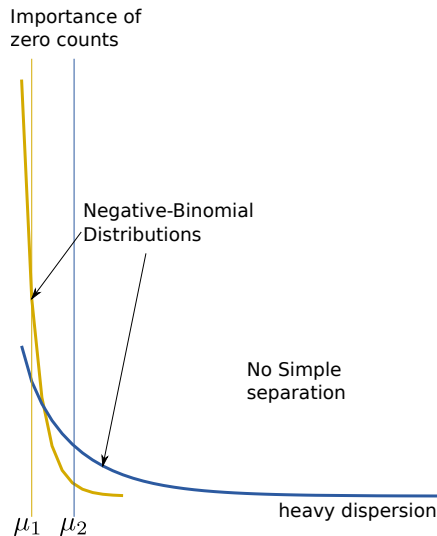
Statistical Setting: two-sample test

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



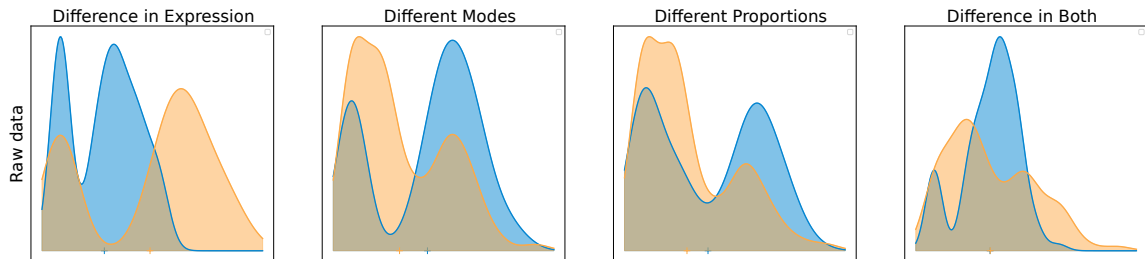
sc-RNAseq data are count data

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



sc-RNASeq are complex count distributions

Compare Gene Expression distributions \mathbb{P}_1 vs \mathbb{P}_2



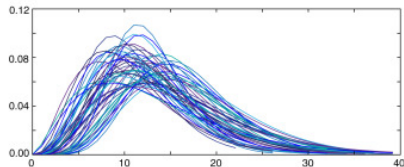
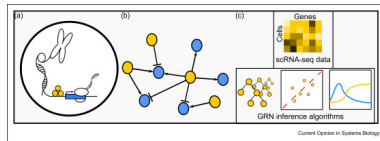
→ No simple linear separation

Strong dependencies and lots of data

- Gene Expressions are highly dependent
- Consider the multivariate model
- $\mathbf{X}_{ic} = [X_{ic}^1, \dots, X_{ic}^G]$, $\boldsymbol{\mu}_i = [\mu_i^1, \dots, \mu_i^G]$

$$\mathbb{E}(\mathbf{X}_{ic}) = \boldsymbol{\mu}_i, \quad \mathbb{V}(\mathbf{X}_{ic}) = \Sigma$$

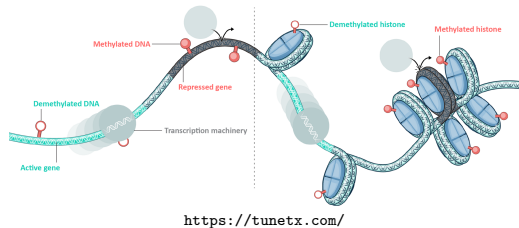
- Σ can be inferred accurately
- Powerful Linear Gene-Set Analysis



Distribution of gene expression across cells

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 ?



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 Peng³, Wei Li^{1*} and Jingyi Jessica Li^{2,4,5,6,7*} 

→ Risk: detect a difference whereas the appropriate model there would not

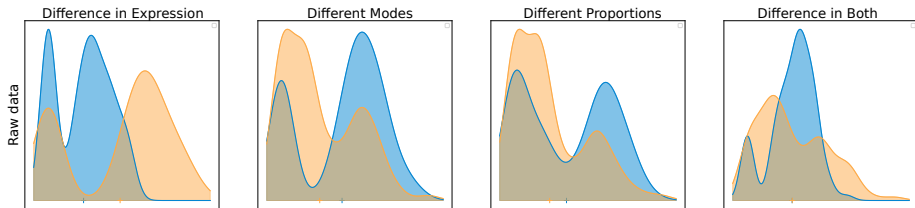
Take-Home Message Slide (1)

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

Outline

1. The Single-Cell Revolution
- 2. Comparison of Gene Expression Distributions**
3. Introduction to kernel testing
4. Discussion about methods
5. Towards perturbation analysis

Comparing Gene Expression Distributions

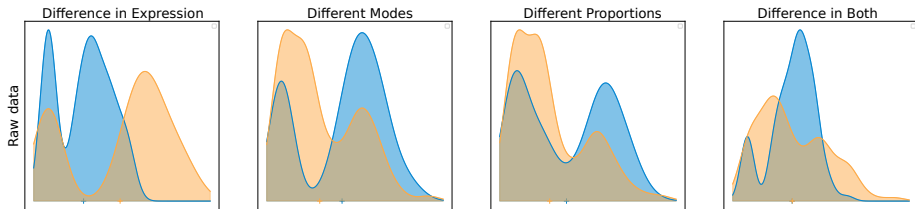


- Single-cell differential expression by distributions comparison :

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

- No simple linear separation \rightarrow SNR is not relevant anymore

Comparing Distribution Functions

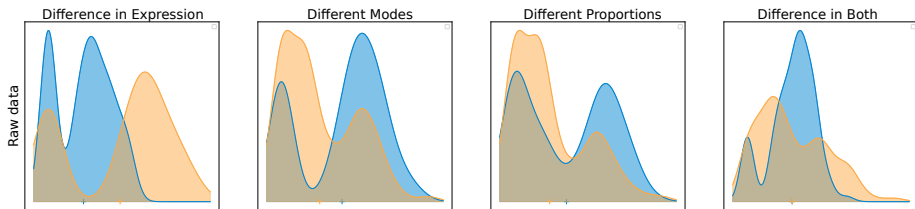


- A strategy consists in comparing cumulative distribution functions:

$$\mathcal{H}_0 : \{F_1 = F_2\}$$

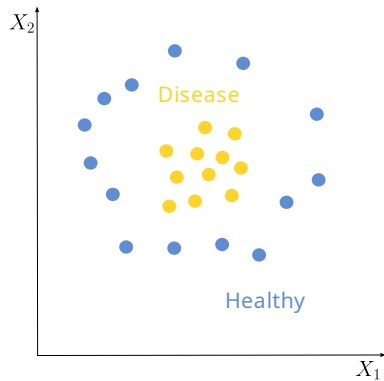
- Estimate cumulative distribution functions can be costly
- Difficult to generalize for gene sets

Comparing embedded distributions

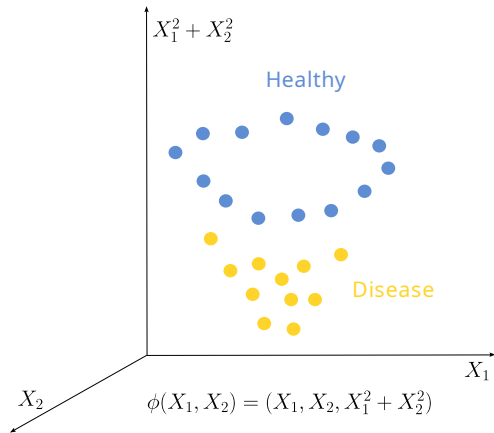


- **Idea:** transform data into a new space
- Use SNR and linear separation on the transformed data

Data transformation for better separation

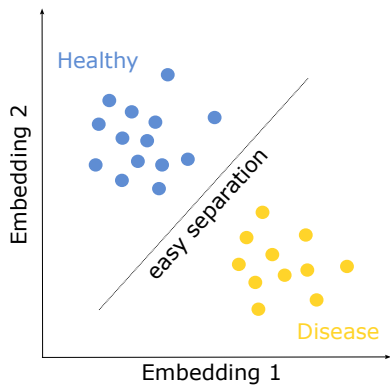


No linear separation

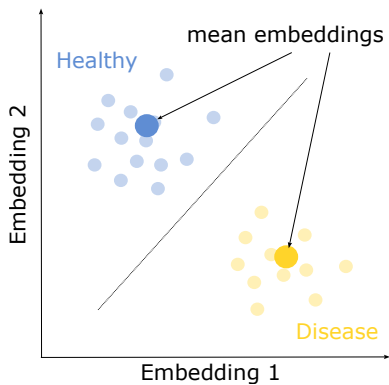


Linear separation

Rich Representations of complex data



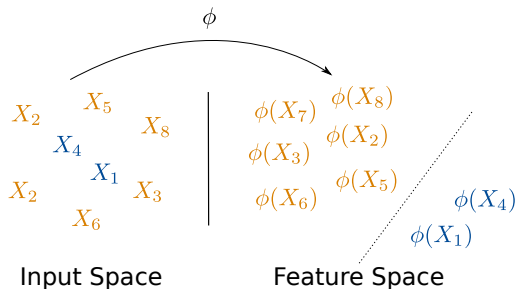
Work on joint transcriptomic embeddings



Mean embeddings by condition

What is an embedding ?

- Transform the input data $X_i \rightarrow \phi(X_i)$
- New representation (UMAP, tSNE)
- Easy separation after transformation ?
- How to choose ϕ ?

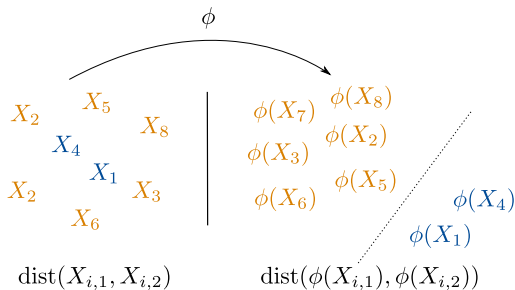


Kernel Methods provide powerful embeddings

- Similarity between data $\text{dist}(X_{i,1}, X_{i,2})$
- Similarity between embeddings

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

- Can work with any input data
- Differential analysis on embeddings



Quick intro on kernel methods

- Kernel function : \mathcal{X} a measurable space:

$$k(\bullet, \bullet) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}.$$

- $k(\bullet, \bullet)$ is a positive definite kernel iff \mathbf{K} is symmetric and positive definite.

$$\forall (x_1, \dots, x_n) \in \mathcal{X}^n, \quad \mathbf{K} = \left[k(x_i, x_j) \right]_{i,j} \in \mathcal{M}_n(\mathbb{R})^n$$

$$\forall (c_1, \dots, c_n) \in \mathbb{R}^n, \quad \sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$$

Aronszajn Theorem

- $k(\bullet, \bullet)$ is positive definite iff there exists a Hilbert space \mathcal{H}_k from $\mathcal{X} \rightarrow \mathbb{R}$ and a feature map ϕ

$$\phi : \mathcal{X} \rightarrow \mathcal{H}_k$$

$$\begin{aligned}\phi(x) &= k(x, \bullet) \\ \forall (x, x') \in \mathcal{X}^2 : k(x, x') &= \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k}.\end{aligned}$$

- \mathcal{H}_k is called a Reproducing Kernel Hilbert Space (RKHS)
- Choosing $k(\bullet, \bullet)$, determines the unique RKHS and the so-called feature map function

$$\phi : \mathcal{X} \rightarrow \mathcal{H}_k$$

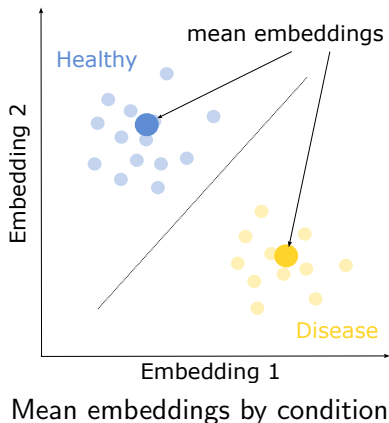
Embedding distributions

- Define the representer $\mu_{\mathbb{P}}$ of \mathbb{P} in \mathcal{H}_k , such that

$$\mathbb{P} \rightarrow \mu_{\mathbb{P}} = \int k(x, \bullet) d\mathbb{P}(x)$$

- $\mu_{\mathbb{P}}$ is called the mean embedding of distribution \mathbb{P} :

$$\mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}(\phi(X))$$



Particular case : characteristic kernel

- If $k(\bullet, \bullet)$ is characteristic, then :

$$(\mathbb{P}_1 = \mathbb{P}_2) \text{ in } \mathcal{X} \quad \Longleftrightarrow \quad (\mu_{\mathbb{P}_1} = \mu_{\mathbb{P}_2}) \text{ in } \mathcal{H}_k$$

- Come back to a test on equality of means in \mathcal{H}_k
- We will consider the Gaussian kernel:

$$k_\sigma(x, x') = \exp \left(-\frac{1}{2\sigma^2} \|x - x'\|_2^2 \right)$$

Kernel Covariance Operators

- Represent distribution beyond the mean embedding
- Quantify the variability of the embeddings
- The kernel covariance operator is the covariance of the embeddings:

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}} \left[(\phi(X) - \mu_{\mathbb{P}}) \otimes (\phi(X) - \mu_{\mathbb{P}}) \right]$$

Take-Home Message Slide (2)

- ✓ Standard Differential Expression procedures can be applied by averaging data (pseudo bulk)
- ✓ Propose tests based on distributions comparisons
- ✓ Work on the embedding of distributions using a kernel
- ✓ Describe the distributions by the mean and the covariance of the embeddings

Outline

1. The Single-Cell Revolution
2. Comparison of Gene Expression Distributions
- 3. Introduction to kernel testing**
4. Discussion about methods
5. Towards perturbation analysis

Metric between distributions

- Testing H_0 requires a metric between distributions

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

- Expected property of the metric

$$\mathbb{P}_1 = \mathbb{P}_2 \quad \Leftrightarrow \quad \mu_{\mathbb{P}_1} = \mu_{\mathbb{P}_2}.$$

- The Maximal Mean Discrepancy:

$$\text{MMD}^2(\mathbb{P}_1, \mathbb{P}_2) = \|\mu_1 - \mu_2\|_{\mathcal{H}_k}^2$$

Computing the empirical MMD

- Embed the observations in \mathcal{H}_k and define the empirical mean embeddings

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(X_{i,1}) \quad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(X_{i,2})$$

- Compute the empirical MMD as a test statistic

$$\begin{aligned} \widehat{\text{MMD}}^2 &= \|\hat{\mu}_2 - \hat{\mu}_1\|_{\mathcal{H}}^2 \\ &= \frac{1}{n_1(n_1 - 1)} \sum_{i \neq j} k(X_{i,1}, X_{j,1}) + \frac{1}{n_2(n_2 - 1)} \sum_{i \neq j} k(X_{i,2}, X_{j,2}) \\ &\quad - \frac{2}{n_1 n_2} \sum_{i,j} k(X_{i,1}, X_{j,2}) \end{aligned}$$

Interpretation : Pair-Wise kernelized Distances

- The MMD can be viewed as a testing framework based on kernelized distances
- Intra-condition distances

$$\frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_1} K(\textcolor{brown}{X}_{i,1}, \textcolor{brown}{X}_{i',1}) \quad \text{and} \quad \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{i'=1}^{n_2} K(\textcolor{blue}{X}_{i,2}, \textcolor{blue}{X}_{i',2})$$

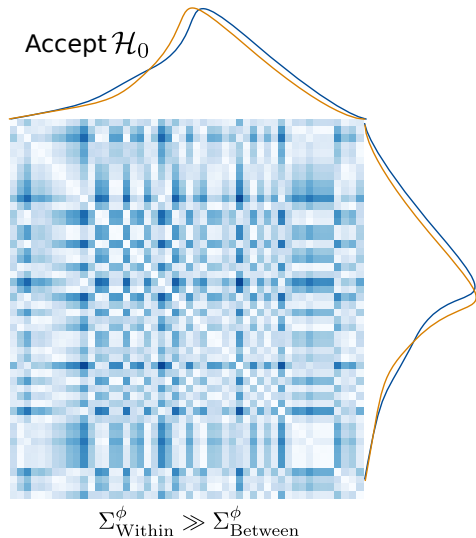
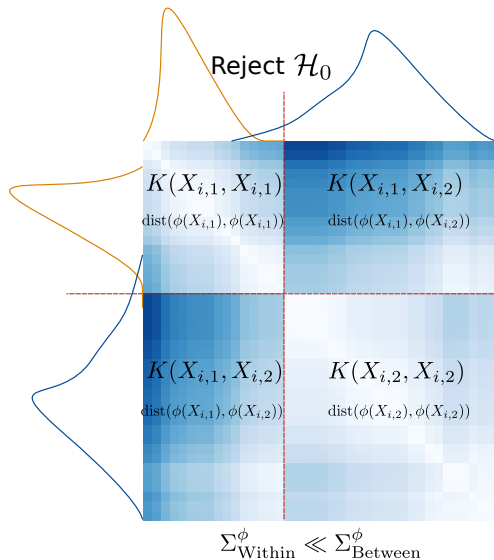
→ If small, conditions are homogeneous

- Inter-condition distance

$$\frac{1}{n_1} \frac{1}{n_2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_2} K(\textcolor{brown}{X}_{i,1}, \textcolor{blue}{X}_{i',2})$$

→ If high, conditions are well separated

Statistical Testing with pair-wise distances



Intra-Inter trade-off between embeddings variabilities

- Separated Conditions:

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

- Similar conditions :

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

- Construct the discriminant ratio

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

Definition of Intra/Inter Variance of embeddings

- The MMD is linked to the between-group covariance

$$\hat{\Sigma}_B = \frac{n_1 n_2}{n^2} (\hat{\mu}_2 - \hat{\mu}_1)^{\otimes 2}$$

- Define the within-group covariances $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$

$$\hat{\Sigma}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\phi(X_{1,i}) - \hat{\mu}_1 \right)^{\otimes 2}, \quad \hat{\Sigma}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \left(\phi(X_{2,i}) - \hat{\mu}_2 \right)^{\otimes 2}$$

$$\Sigma_W = \frac{n_1}{n} \Sigma_1 + \frac{n_2}{n} \Sigma_2$$

The Normalized MMD

- The normalized MMD statistics is

$$\begin{aligned} D^2(\mathbb{P}_1, \mathbb{P}_2) &= \frac{n_1 n_2}{n} \left\| \Sigma_W^{-\frac{1}{2}} (\mu_2 - \mu_1) \right\|_{\mathcal{H}}^2 \\ &\sim \frac{1}{n} \text{Tr} \left(\Sigma_W^{-1} \Sigma_B \right) \end{aligned}$$

- It is a kernelized discriminant ratio
- Classifier-based testing: kernel Fisher Discriminant Analysis

Statistical Challenges

- Explore the expected variations of the MMD of D^2 under $\mathbb{P}_1 = \mathbb{P}_2$.
- The target is the $(1 - \alpha)$ quantile of the distribution

$$\mathbb{P}_{H_0} \left(\widehat{\text{MMD}}^2 > q_{1-\alpha} \right) < \alpha$$

$$\mathbb{P}_{H_0} \left(\widehat{D}^2 > q_{1-\alpha} \right) < \alpha$$

- The approximate distribution can be asymptotic / non-asymptotic
- Permutation strategies are also possible to estimate $q_{1-\alpha}$

Take-Home Message Slide (3)

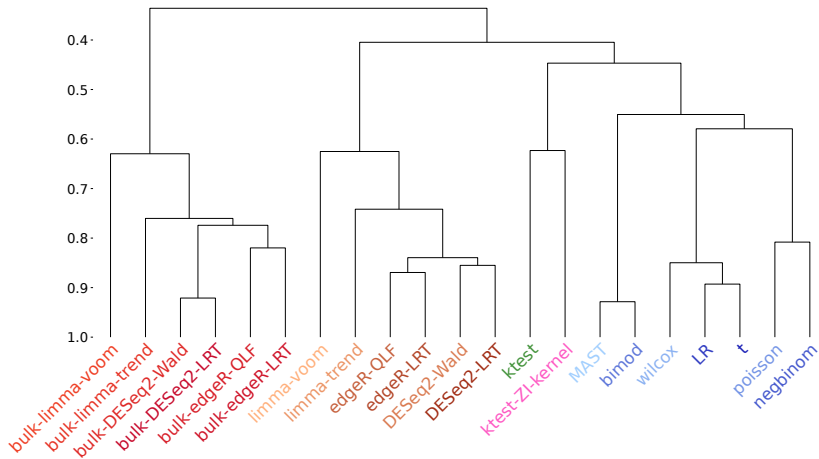
- ✓ Kernel methods can be used to define discrepancies between distributions
- ✓ Kernel tests are based on pair-wise distances between embeddings
- ✓ These distances can be normalized by embeddings variability
- ✓ pvalues can be obtained (approximations)

Outline

1. The Single-Cell Revolution
2. Comparison of Gene Expression Distributions
3. Introduction to kernel testing
- 4. Discussion about methods**
5. Towards perturbation analysis

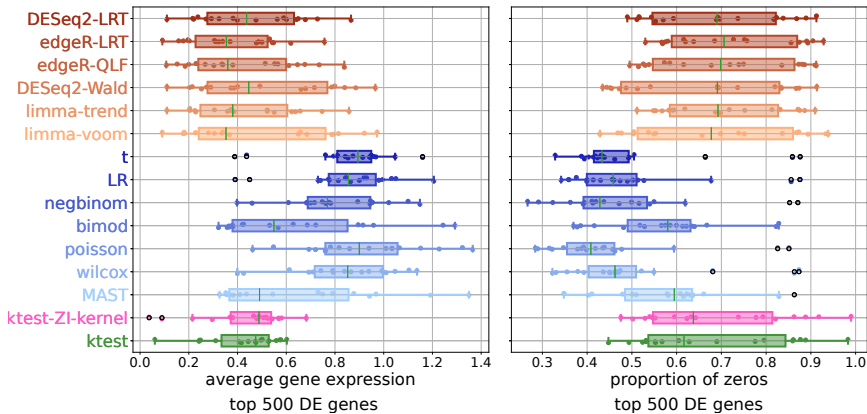
Methods comparison on experimental datasets

- 18 published datasets [4] / 20 methods
- Compare AUCCs based on reference gene lists



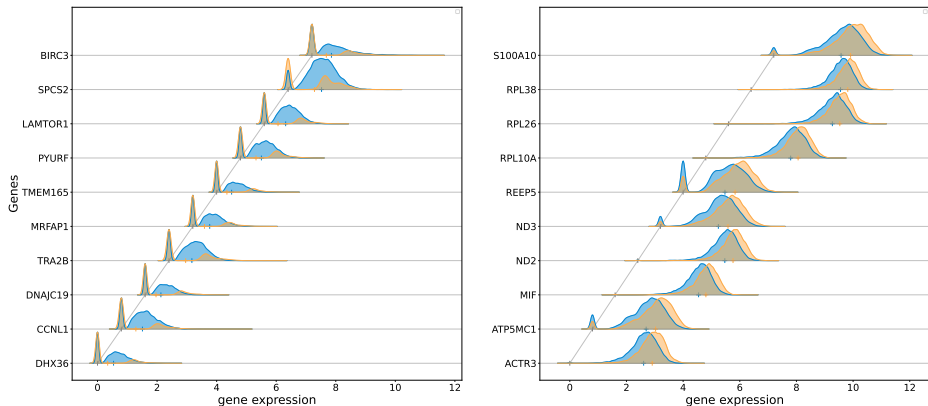
Methods comparison on experimental datasets

- 18 published datasets [4] / 20 methods
- Check the summary statistics characteristics of rejected distributions



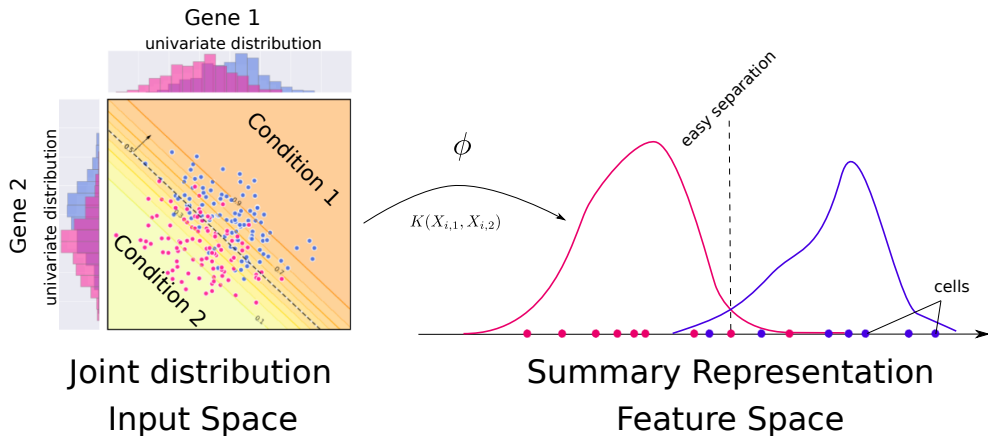
Methods comparison on experimental datasets

- 18 published datasets [4] / 20 methods
- Check distribution forms of rejected hypothesis



Non DE in pseudo Bulk - Non DE in scDEA methods

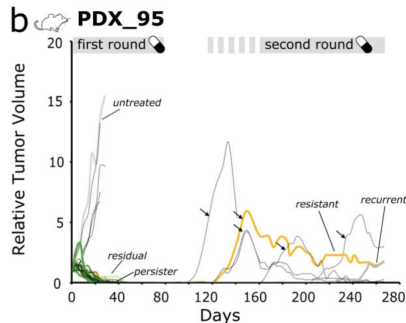
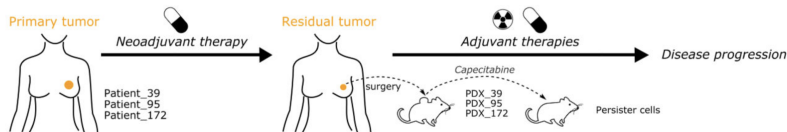
Gene-Set Differential Analysis



ChemoResistance in Triple Negative Breast Cancer

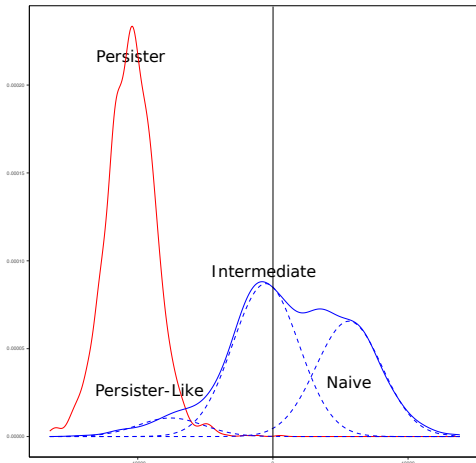
- Emergence of resistant phenotypes is a multi-step process
- After drug insult only a pool of drug-tolerant persister cells manage to tolerate the treatment and survive.
- Reservoir from which drug-resistant cells can ultimately emerge.

a



Kernel testing on Persister vs. Naive cells

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




Summary of Whole Epigenome differences

METHOD

Open Access

Kernel-based testing for single-cell differential analysis



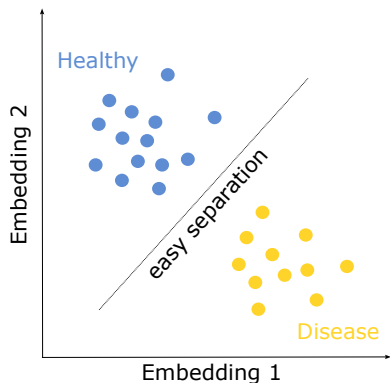
A. Ozier-Lafontaine^{1*}, C. Fourneau², G. Durif², P. Arsenteva¹, C. Vallot^{3,4}, O. Gandrillon², S. Gonin-Giraud²,
B. Michel^{1*†} and F. Picard^{2*†} 

<https://github.com/LMJL-Alea/ktest>

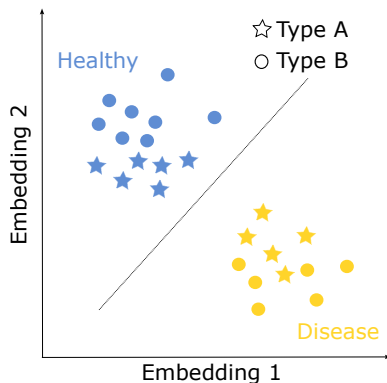
Outline

1. The Single-Cell Revolution
2. Comparison of Gene Expression Distributions
3. Introduction to kernel testing
4. Discussion about methods
- 5. Towards perturbation analysis**

From Differential Analysis to Perturbation Analysis

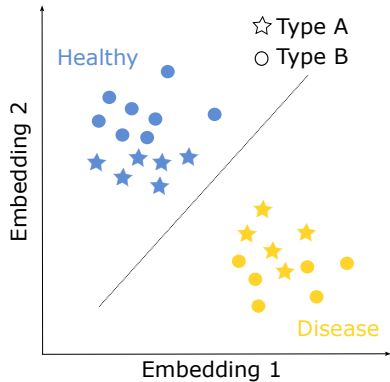


Differential Analysis of Transcriptomes

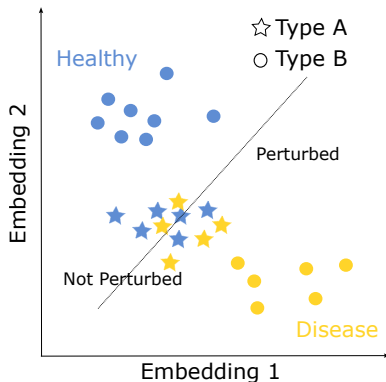


Cells are grouped in Cell Types

Detecting Perturbed Cell Types

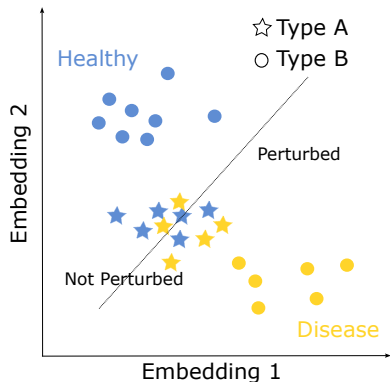


All Cell-types perturbed

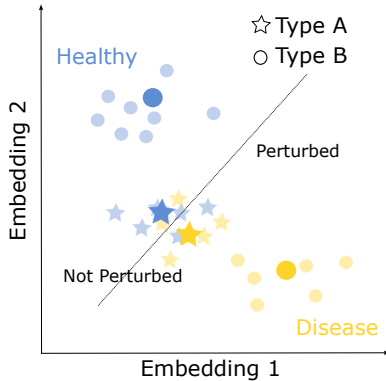


Differential Perturbation

Perturbed Mean Embeddings



Differential Perturbation



Interaction Treatment \times Cell-types

ANOVA for non-linear Embeddings

- Complex design : treatment, cell types factors

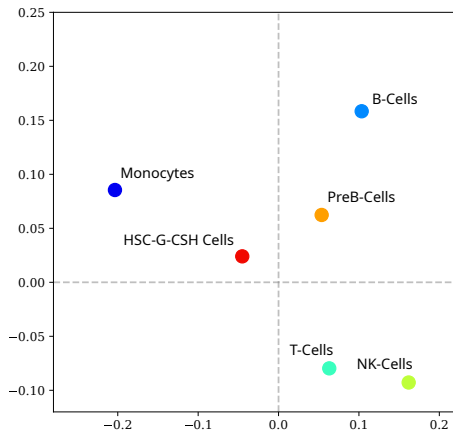
$$\phi(\text{Expression}) = \mu + \alpha_{\text{treatment}} + \beta_{\text{celltype}} + (\alpha\beta)_{\text{treatment} \times \text{celltype}} + \text{Error}$$

- Identify Perturbed cell types with the interaction terms

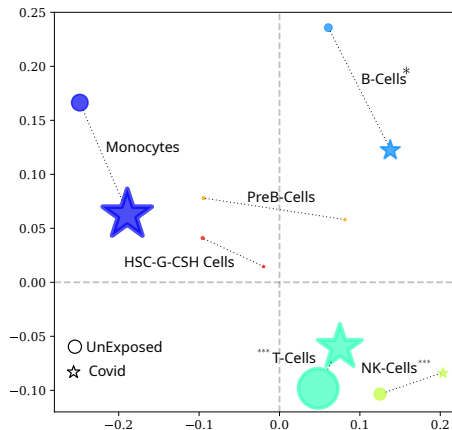
$$\mathcal{H}_0^\star : \left\{ (\alpha\beta)_{\text{Healthy} \times \star} = (\alpha\beta)_{\text{Disease} \times \star} \right\}$$

$$\mathcal{H}_0^{\circ} : \left\{ (\alpha\beta)_{\text{Healthy} \times \circ} = (\alpha\beta)_{\text{Disease} \times \circ} \right\}$$

Non-Linear perturbations following Covid Exposure

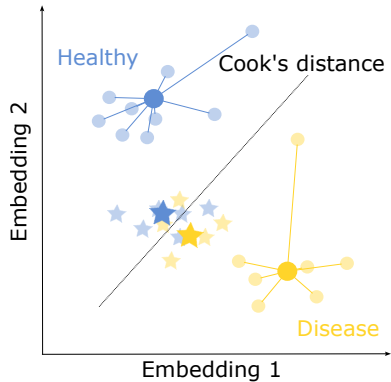


Cell-Type Effect***

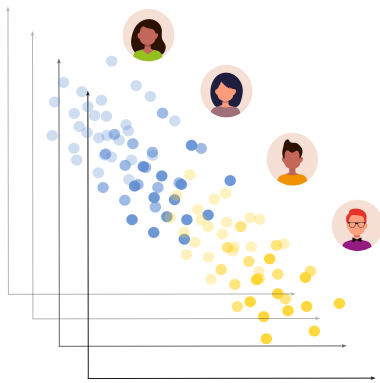


Interaction Cell-Type \times Disease***

Covid DataSet



Atypical Cells identification



Multi-patients Designs

kAOV: kernel testing for general designs

- General Model for kernel testing in any design:

$$\phi(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

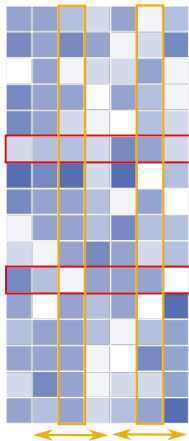
- Embedding-Based Contrast Testing:

$$\mathcal{H}_0 = \left\{ \mathbf{C}\boldsymbol{\beta} = 0 \right\}$$

- Hotelling-Lawley Trace Test (χ^2 distributed)
- Package available : <https://github.com/LMJL-Alea/kAOV>

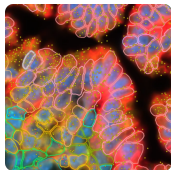
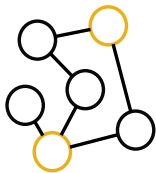
Perspectives

Genes Perturbations



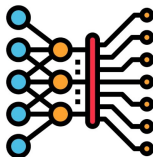
Cells perturbations

Gene Regulatory Network

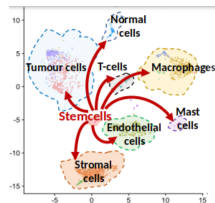
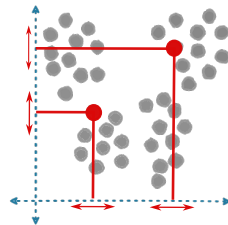


Spatial Data

Interpretable AI



Perturbation
Analysis



Tumor MicroEnvironment

Acknowledgments

- Anthony Ozier-Lafontaine, Bertrand Michel, Perrine Lacroix, Nantes University
- Polina Arsenteva, Ghislain Durif, Lucy Attwood, ENS Lyon
- Vincent Rivoirard, Dauphine University
- Philippe Bertolino, CRCL, Lyon
- PEPR Digital Health (AI4scMed), ANR

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] Q. Jia, H. Chu, Z. Jin, H. Long, and B. Zhu. High-throughput single- $\tilde{\text{Nell}}$ sequencing in cancer research. *Signal Transduction and Targeted Therapy*, 7(1), May 2022.
- [4] 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.