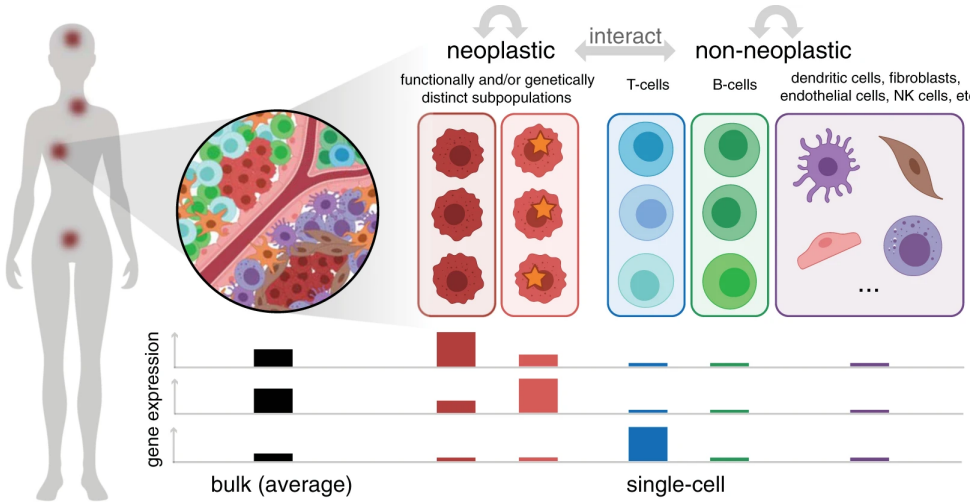
Kernel-based perturbation testing for single-cell data

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

franck.picard@ens-lyon.fr - Lyon, Oct. 2025

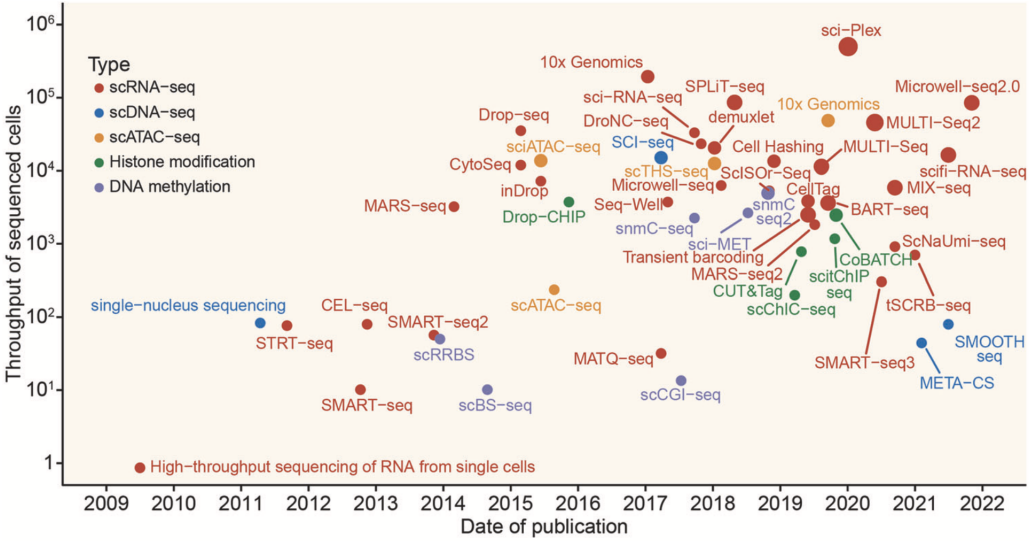
Outline

From bulk to distributions of gene expression



[2]

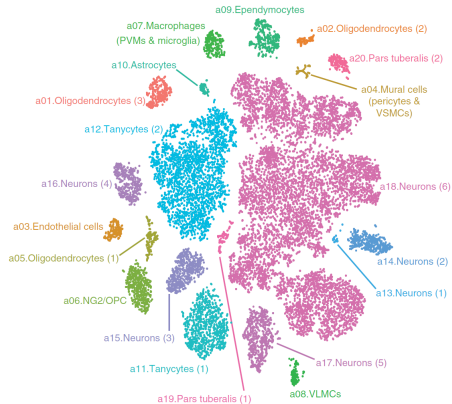
A timeline: produced data



[3]

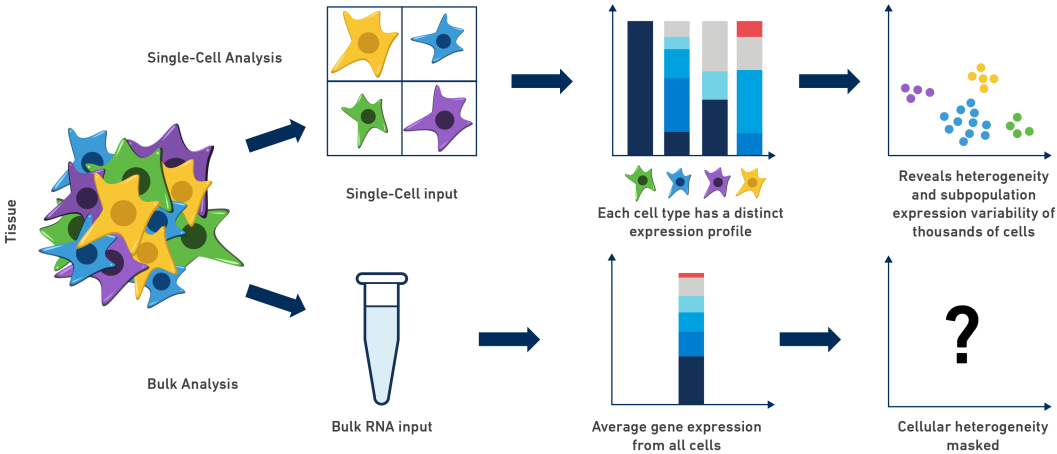
Machine Learning challenges

- Dimension Reduction / Visualization
- Clustering cell-type discovery
- Datasets alignments
- Catch cells-ecosystems behaviors
- Simulation of fake data
- Data integration
- Genes expression comparison



[1]

Single-Cell from a statistician's perspective



From 10X Genomics

Comparing Biological Conditions

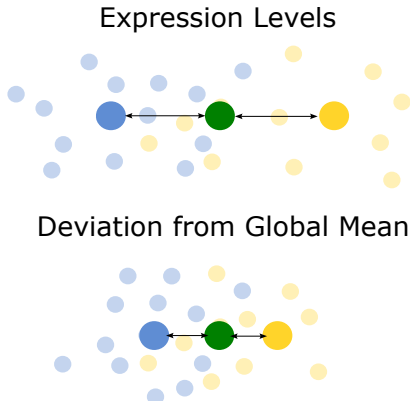
Gene-wise comparison

Statistical Testing

$\rightarrow$ Score the difference
$\rightarrow$ Control type-I errors

Single-cell data $n \sim 10^6$

Try non-parametrics !

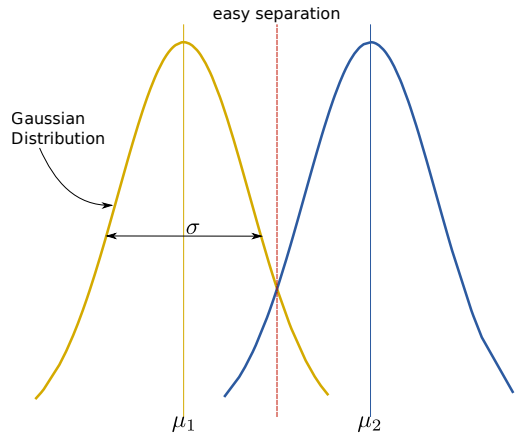


Statistical Setting: two-sample test

logFC are valid provided μ and σ are good summaries of the information

Easy linear separation

Not adapted to single-cell assays



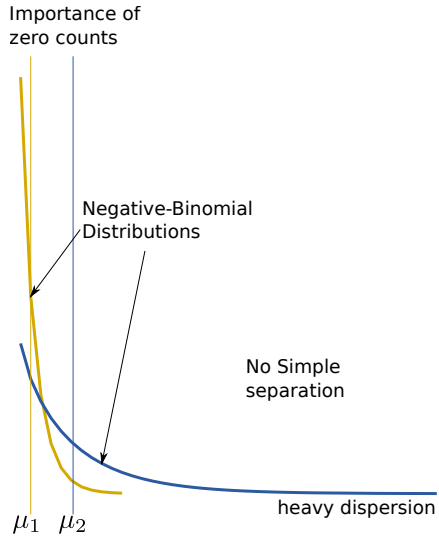
sc-RNAseq data are count data

Specificities: discrete, zeros

How to define the signal-to-noise ratio ?

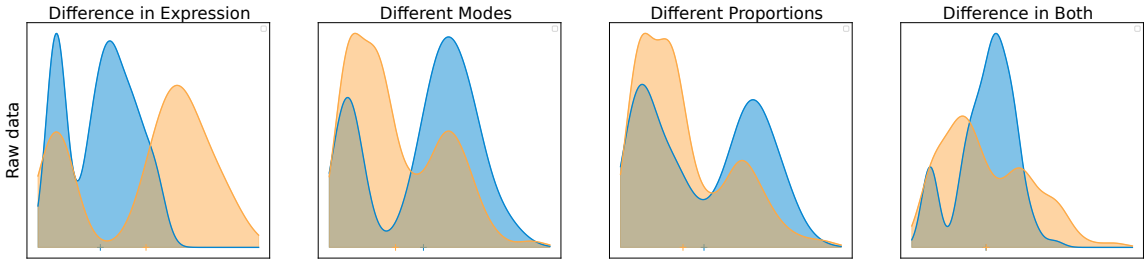
Standard: Negative Binomial distribution

Linear separation with GLM (parametric)



sc-RNASeq are complex count distributions

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



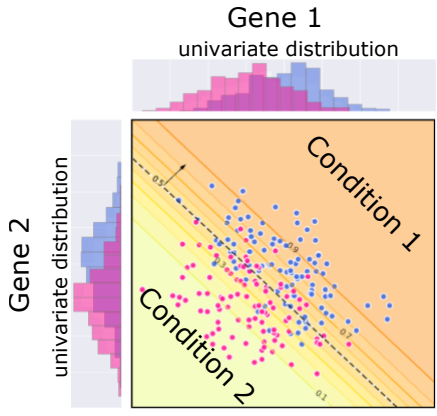
$\rightarrow$ No simple linear separation

Gene-Wise Strategy: a Good Option ?

Gene Expressions are highly dependent

Multivariate distributions

Calls for non-linear embedding

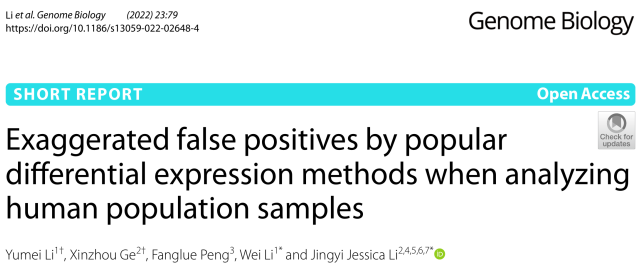


[6]

Statistical Challenge

Statistical testing is based on what is expected under $\mathcal{H}_0$

Control the random fluctuations of the embeddings under the null


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*]

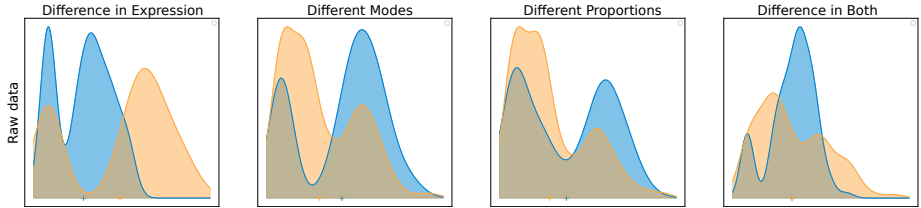
$\rightarrow$ Risk: detect a difference whereas the appropriate model there would not

Take-Home Message Slide (1)

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

Outline

Comparing Gene Expression Distributions

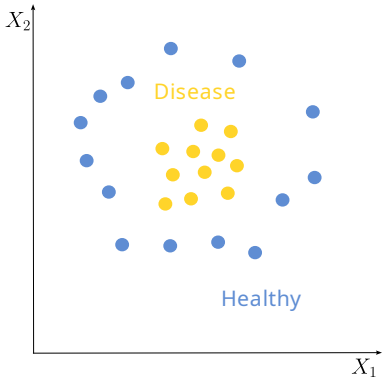


- Single-cell differential expression by distributions comparison :

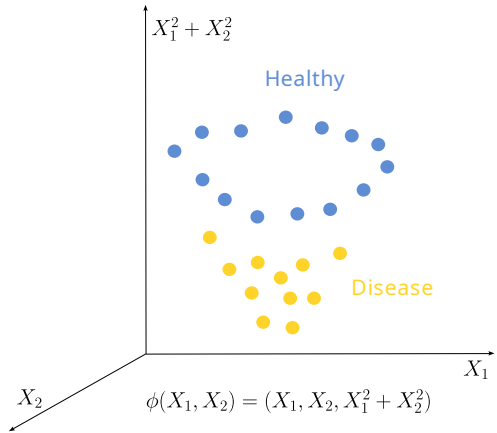
$$\mathcal{H}_0 : \left\{ \mathbb{P}_1 = \mathbb{P}_2 \right\}$$

- No simple linear separation: SNR is not relevant anymore
- **Idea**: transform data into a new space
- Use SNR and linear separation on the transformed data

Data transformation for better separation

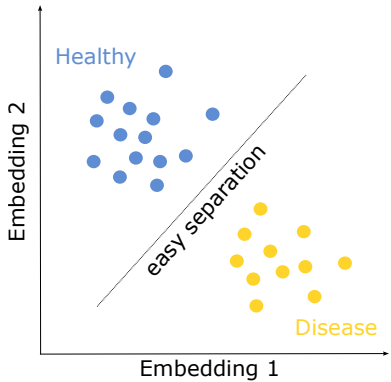


No linear separation

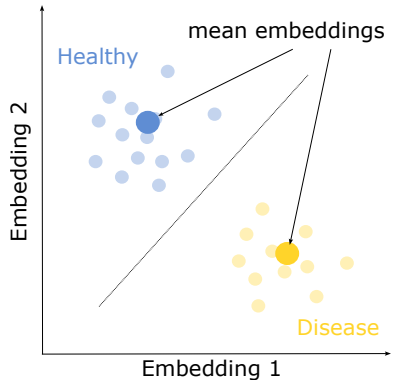


Linear separation

Rich Representations of complex data



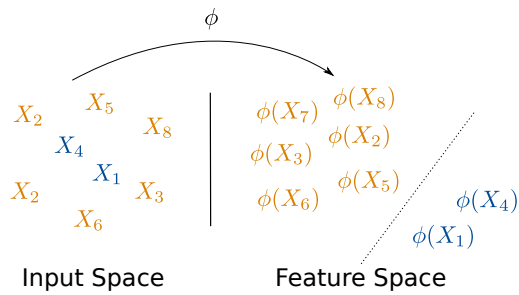
Work on joint transcriptomic embeddings



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

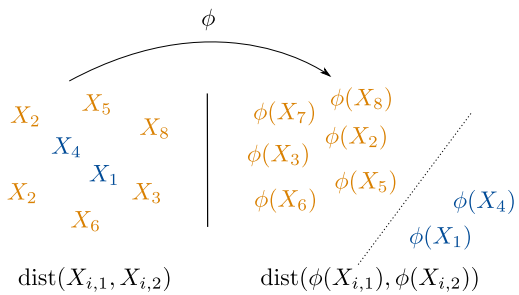
- Similarity between data

$$\text{dist}\big(X_{i,1}, X_{i,2}\big)$$

- Similarity between embeddings

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

- This is what does a kernel !



How to choose the kernel ?

- Popular kernel : Gaussian kernel (h hyperparameter)

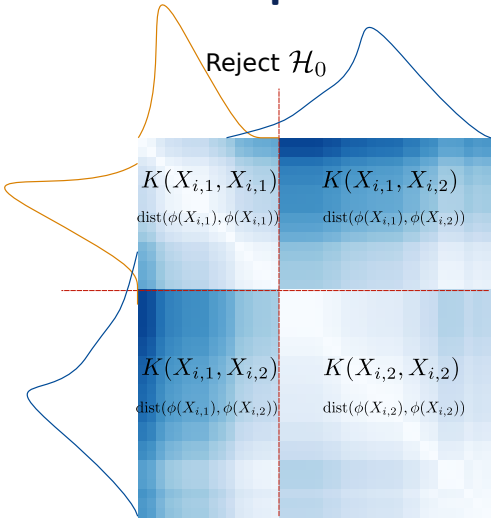
$$K(X_i, X_{i'}) \propto \exp\left\{-\frac{1}{2}\left(\frac{X_i - X_{i'}}{h}\right)^2\right\}$$

- Theory ensures it is a distance between embeddings

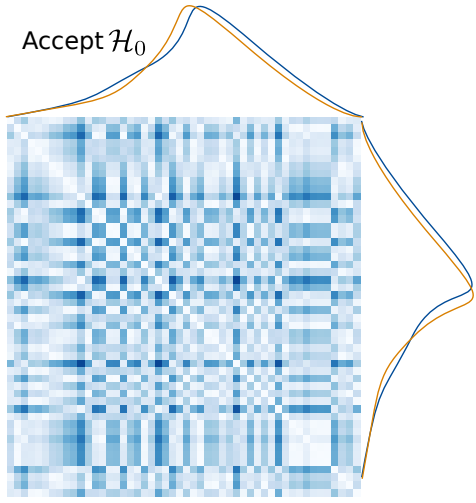
$$K(X_i, X_{i'}) = \text{dist}\Big(\phi(X_i), \phi(X_i)\Big)$$

- The embedding ϕ exists but does not need to be defined

Kernels to compare distributions



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



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

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

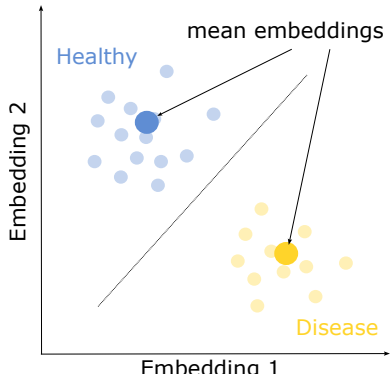
Embedding distributions

- Mean Embedding of $\mathbb{P}$:

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

- Covariance of the embeddings under $\mathbb{P}$:

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}\Big[(\phi(X) - \mu_{\mathbb{P}})^{\otimes 2}\Big]$$



Mean embeddings by condition

Metric between distributions

- Testing H_0 requires a metric between distributions

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

- 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:

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

Computing the empirical MMD

- Empirical mean embeddings and MMD

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

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

- The MMD is a testing framework based on kernelized distances

Between/Within kernel trade-off

- Intra-condition distances

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

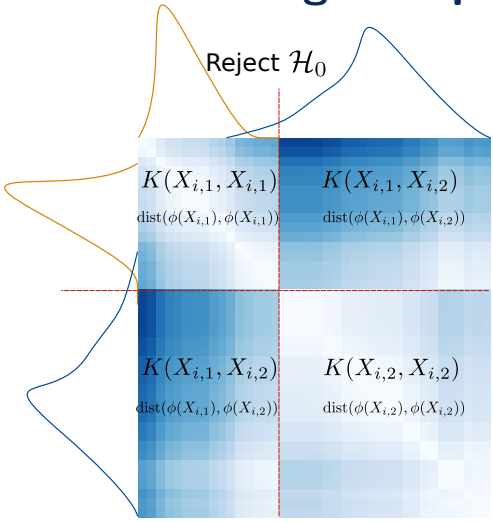
$\rightarrow$ If small, conditions are homogeneous

- Inter-condition distance

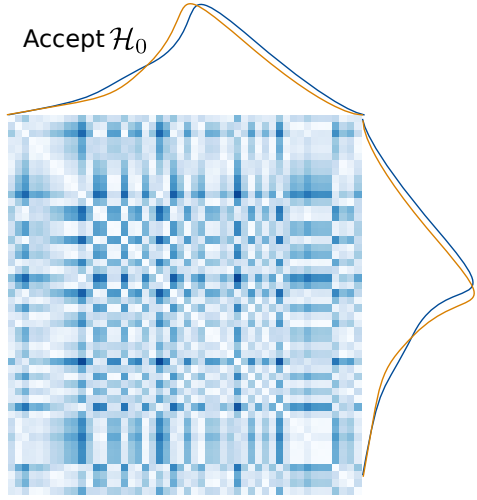
$$\frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{i'=1}^{n_2} k(X_{i,1}, X_{i',2})$$

$\rightarrow$ If high, conditions are well separated

Statistical Testing with pair-wise distances



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



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

Considering Variances

- Separated Conditions:

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

- Similar conditions :

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

- Construct the discriminant ratio between conditions:

$$\mathrm{D}^2(\mathbb{P}_1, \mathbb{P}_2) = \Sigma_{\text{Within}}^{-1} \Sigma_{\text{Between}}$$

- Similar to a 1-way-ANOVA:

$$\phi(X) \sim \texttt{Condition}$$

Definition of Intra/Inter Variance of embeddings

- The MMD is linked to the between-group covariance

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

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

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

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

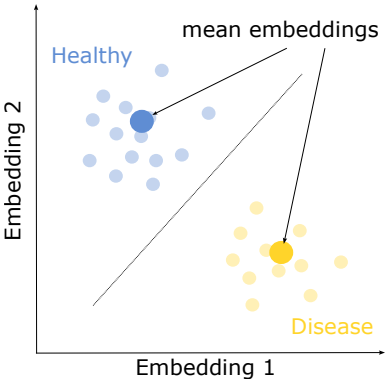
The Normalized MMD

- The normalized MMD statistics is

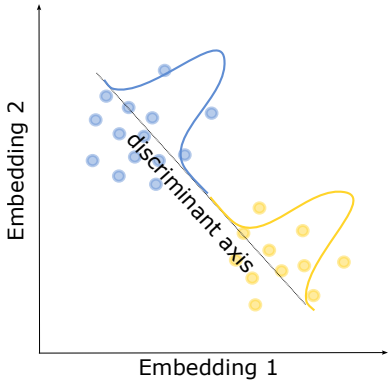
$$\begin{aligned} \mathrm{D}^2(\mathbb{P}_1, \mathbb{P}_2) &= \frac{n_1 n_2}{n} \left\| \Sigma_W^{-\frac{1}{2}} (\mu_2 - \mu_1) \right\|_{\mathcal{H}}^2 \\ &\sim \frac{1}{n} \mathrm{Tr} \left(\Sigma_W^{-1} \Sigma_B \right) \end{aligned}$$

- It is a kernelized discriminant ratio with χ^2 distribution under $\mathcal{H}_0$
- The method is based on Kernel Fisher Discriminant Analysis

From Separation to Discrimination

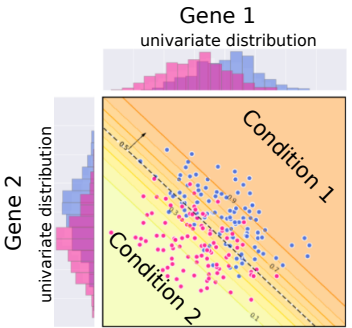


Mean embeddings by condition

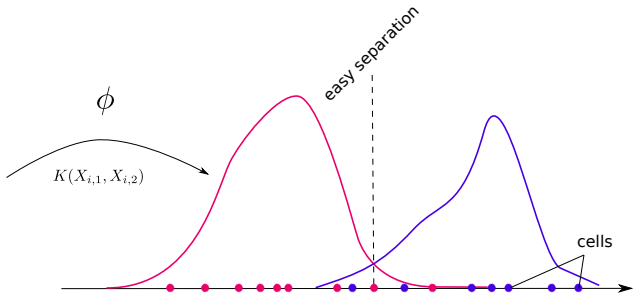


Discrimination of cell populations

Towards Classification



Joint distribution
Input Space



Summary Representation
Feature Space

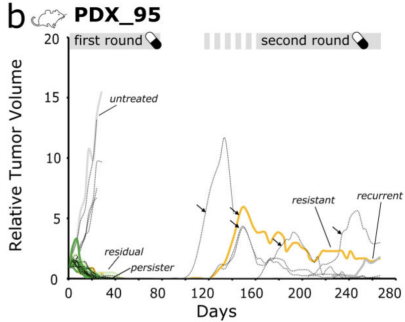
Take-Home Message Slide (3)

- ✓ 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)
- ✓ The method is based on a classifier

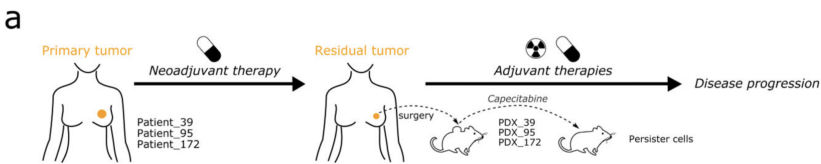
Outline

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.

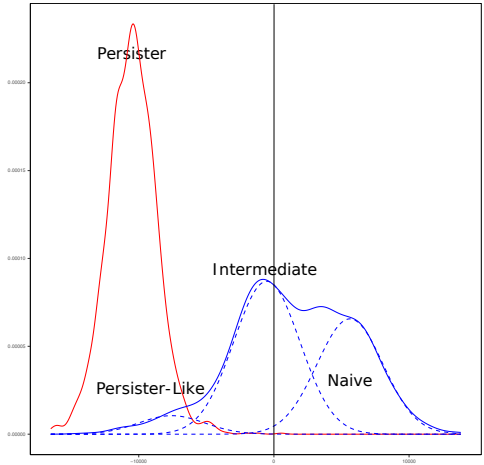


[5]



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

Ozier-Lafontaine *et al. Genome Biology* (2024) 25:114
https://doi.org/10.1186/s13059-024-03255-1

Genome Biology

METHOD

Open Access

Kernel-based testing for single-cell differential analysis

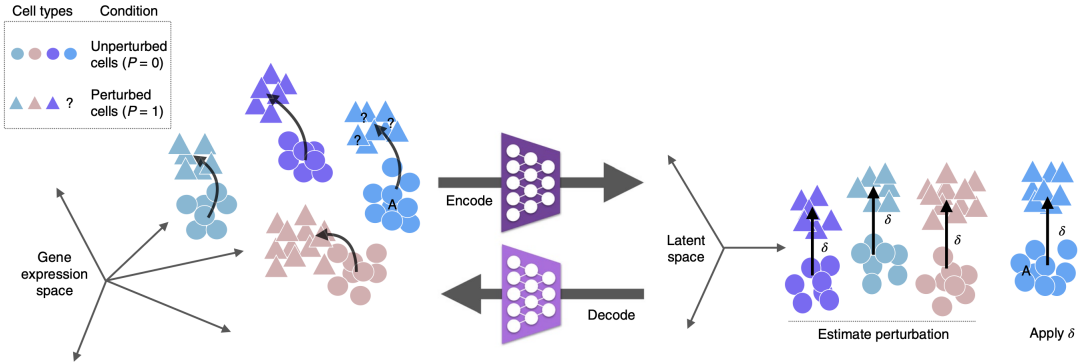
A. Ozier-Lafontaine[1*], C. Fourneaux[2], G. Durif[2], P. Arsenteva[1], C. Vallot[3,4], O. Gandrillon[2], S. Gonin-Giraud[2], B. Michel[1*†] and F. Picard[2*†]

`https://github.com/LMJL-Alea/ktest`

Outline

Context in Single-Cell Transcriptomics

- Modeling/predicting the effects of perturbations is a key task in systems biology.
- Capture the heterogeneity of cell populations using Supervised Dimension Reduction
- Latent space captures the signal to predict a cell's response to perturbation.



[4]

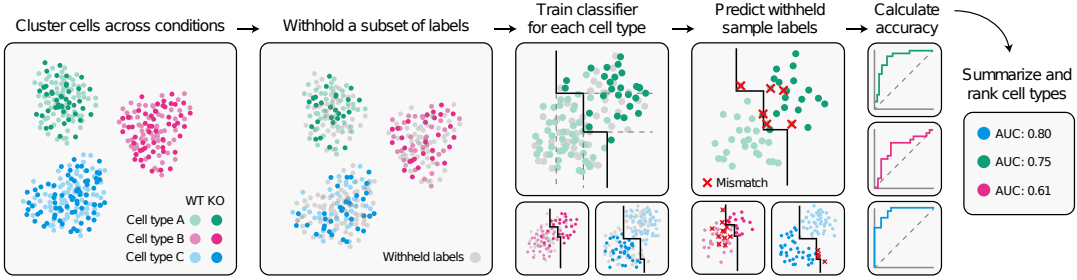
BRIEF COMMUNICATION
https://doi.org/10.1038/s41587-020-0605-1
nature biotechnology

Cell type prioritization in single-cell data

Michael A. Skinnider[1,2,9]✉, Jordan W. Squair[1,3,4,9]✉, Claudia Kathe[1,3], Mark A. Anderson[1,3], Matthieu Gautier[1,3], Kaya J. E. Matson[5], Marco Milano[1,3], Thomas H. Hutson[1,3], Quentin Barraud[1,3], Aaron A. Phillips[6], Leonard J. Foster[2,7], Gioele La Manno[1], Ariel J. Levine[5] and Grégoire Courtine[1,3,8]✉

New Paradigm

- Identify cell-types more responsive to biological perturbations
- Hypothesis: responsive cell-types should be more separable
- Cell-types are prioritized based on the area under the curve AUC



[7]

The two sides of Supervised Learning

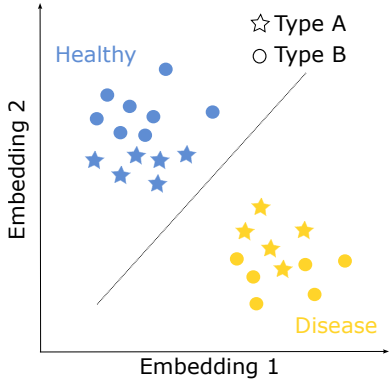
Classification

- Observe $(y_1, x_1), \ldots, (y_n, x_n)$
- Construct a predictor $f : \mathcal{X} \to \mathcal{Y}$
- Predict a new y
- $\mathcal{H}_1$ is not properly defined
- Pros: differences are not well defined
- Cons: x2 data to reach the same power

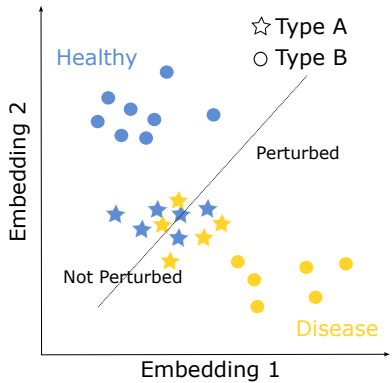
Statistical Testing

- Observe $(y_1, x_1), \ldots, (y_n, x_n)$
- Define a test statistic
- Control Type-I error
- Ensure power
- Pros: powerful if $\mathcal{H}_1$ can be defined
- Cons: when $\mathcal{H}_1$ is ill-defined

From Differential Analysis to Perturbation Analysis

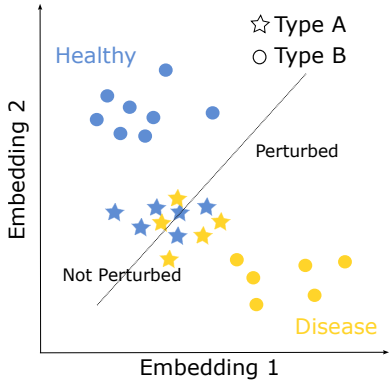


All Cell-types perturbed

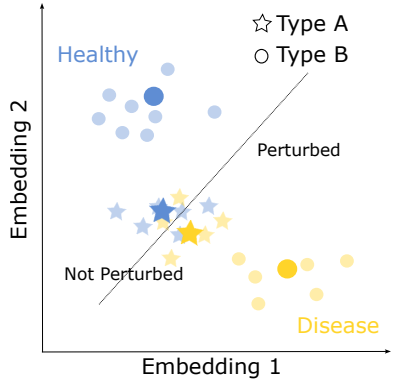


Differential Perturbation

Perturbed Mean Embeddings



Differential Perturbation



Interaction Treatment $\times$ Cell-types

ANOVA for non-linear Embeddings

- Complex design : treatment, cell types factors

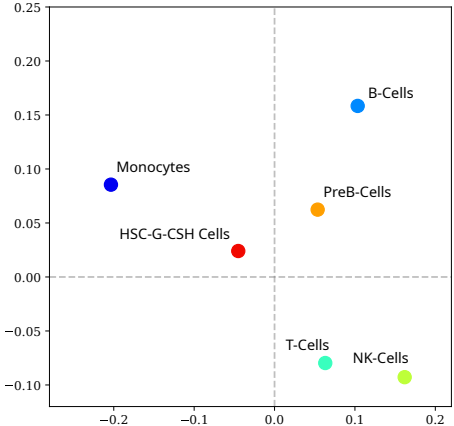
$$\phi(\text{Expression}) \quad \sim \quad \texttt{treatment} + \texttt{celltype} + \texttt{treatment} \times \texttt{celltype}$$

- Identify Perturbed cell types using contrasts

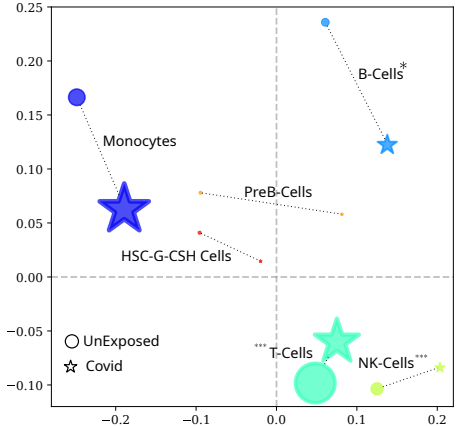
$$\mathcal{H}_0^{\bigstar} : \left\{ \texttt{Healthy} \times \bigstar = \texttt{Disease} \times \bigstar \right\}$$

$$\mathcal{H}_0^{\bigcirc} : \left\{ \texttt{Healthy} \times \bigcirc = \texttt{Disease} \times \bigcirc \right\}$$

Non-Linear perturbations following Covid Exposure

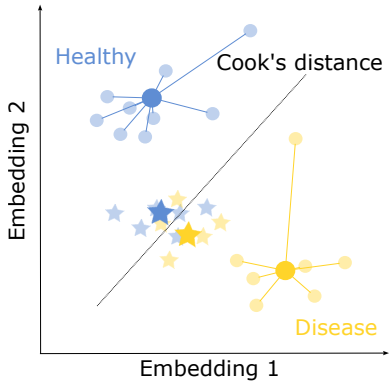


Cell-Type Effect***

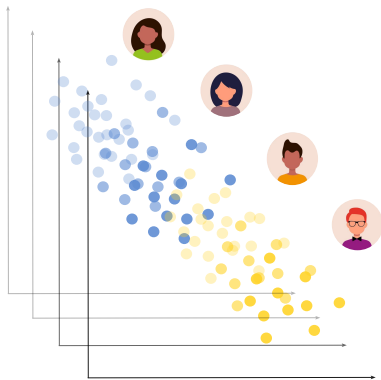


Interaction Cell-Type × Disease***

Soon Included

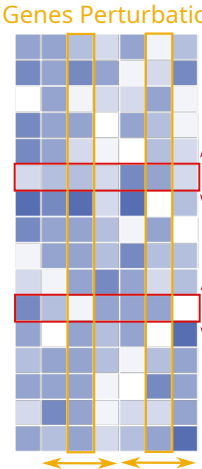


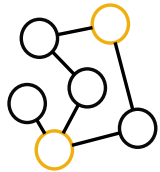
Atypical Cells identification



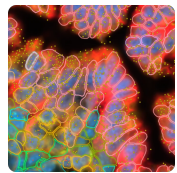
Multi-patients Designs

Perspectives



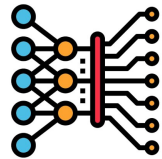


Gene Regulatory Network

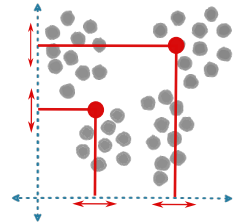


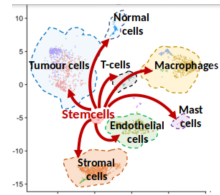
Spatial Data

Interpretable AI



Perturbation Analysis





Tumor MicroEnvironment

Perspectives

- ANOVA approach:

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

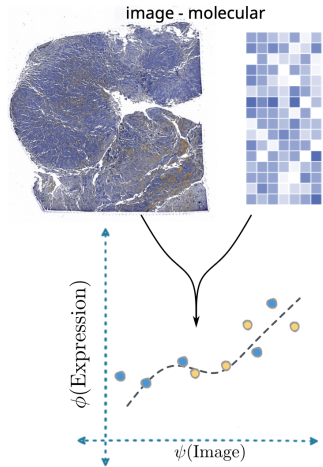
- Test Contrasts: $\mathbf{C}\boldsymbol{\beta} = 0$

- Regression Approach

$$\phi(\text{Expression}) = \psi(\text{Covariates})\boldsymbol{\beta} + E$$

- Test conditional independence: $\boldsymbol{\beta} = 0$
- Covariates can be space ! (SPARK-X !)

Data integration



Multimodal Embeddings

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-ell sequencing in cancer research. *Signal Transduction and Targeted Therapy*, 7(1), May 2022.

[4] M. Lotfollahi, F. A. Wolf, and F. J. Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, July 2019.

[5] J. Marsolier, P. Prompsy, A. Durand, A.-M. Lyne, C. Landragin, A. Trouchet, S. T. Bento, A. Eisele, S. Foulon, L. Baudre, K. Grosselin, M. Bohec, S. Baulande, A. Dahmani, L. Sourd, E. Letouzé, A.-V. Salomon, E. Marangoni, L. Perié, and C. Vallot. H3k27me3 conditions chemotolerance in triple-negative breast cancer. *Nature Genetics*, 54(4):459–468, Apr. 2022.

[6] V. Ntranos, L. Yi, P. Melsted, and L. Pachter. A discriminative learning approach to differential expression analysis for single-cell rna-seq. *Nature Methods*, 16(2):163–166, Jan. 2019.

[7] M. A. Skinnider, J. W. Squair, C. Kathe, M. A. Anderson, M. Gautier, K. J. E. Matson, M. Milano, T. H. Hutson, Q. Barraud, A. A. Phillips, L. J. Foster, G. La Manno, A. J. Levine, and G. Courtine. Cell type prioritization in single-cell data. *Nature Biotechnology*, 39(1):30–34, July 2020.