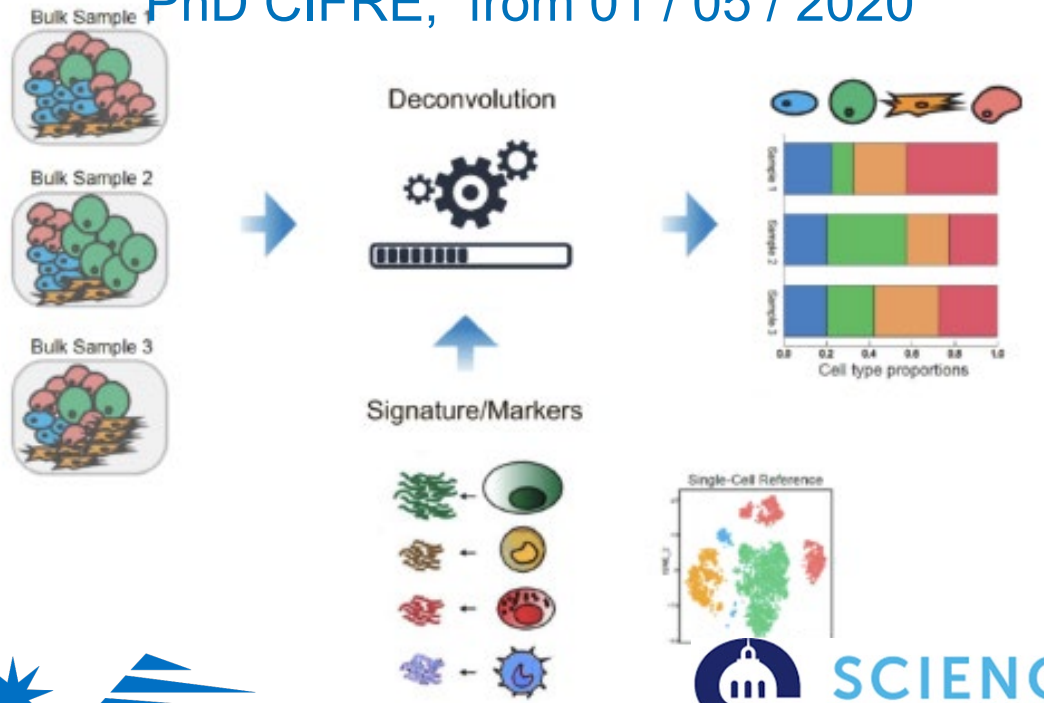
Robust deconvolution of transcriptomic samples using the gene covariance structure

Bastien CHASSAGNOL

PhD CIFRE, from 01 / 05 / 2020



Pierre-Henri WUILLEMIN

Laboratoire d'Informatique de Paris 6 (LIP6)

Etienne BECHT

Bioinformatics

Grégory NUEL

Laboratoire de Probabilités, Statistique et Modélisation

Outline

01 Analysing the biological medium

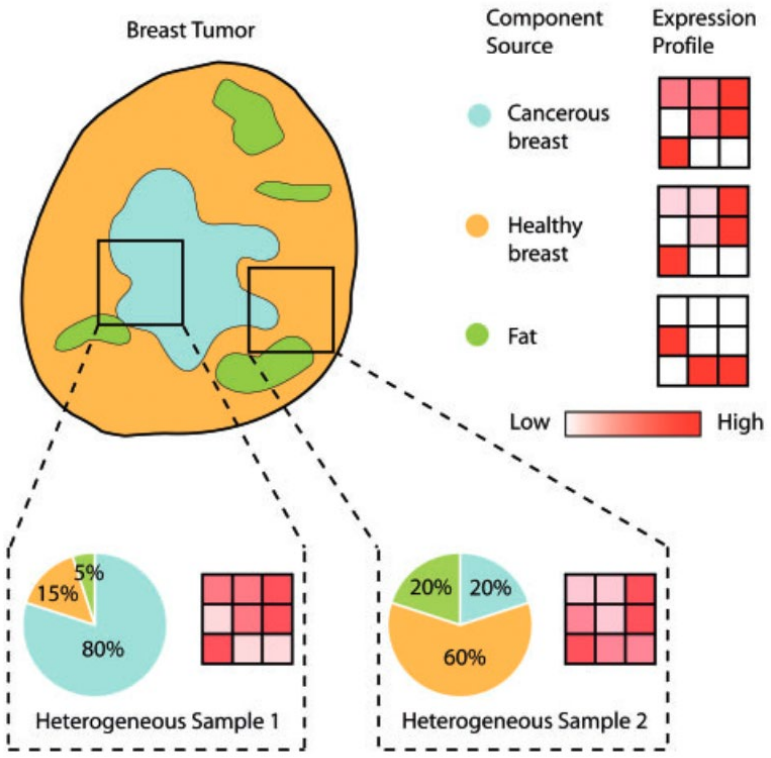
02 A definition of cellular deconvolution

03 Standard deconvolution pipeline

04 Multivariate extension to standard deconvolution algorithms

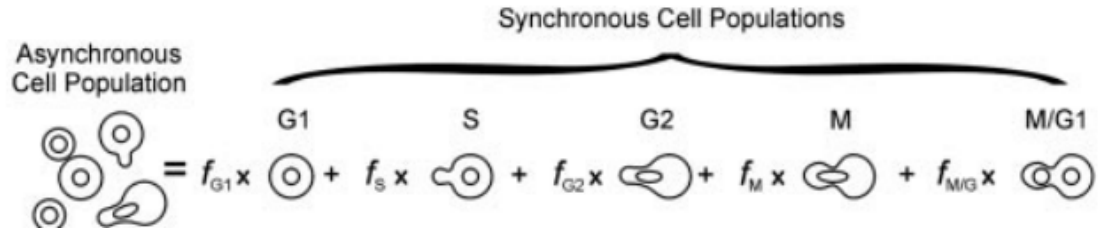
05 Numerical simulation and future development

The complexity of the biological medium



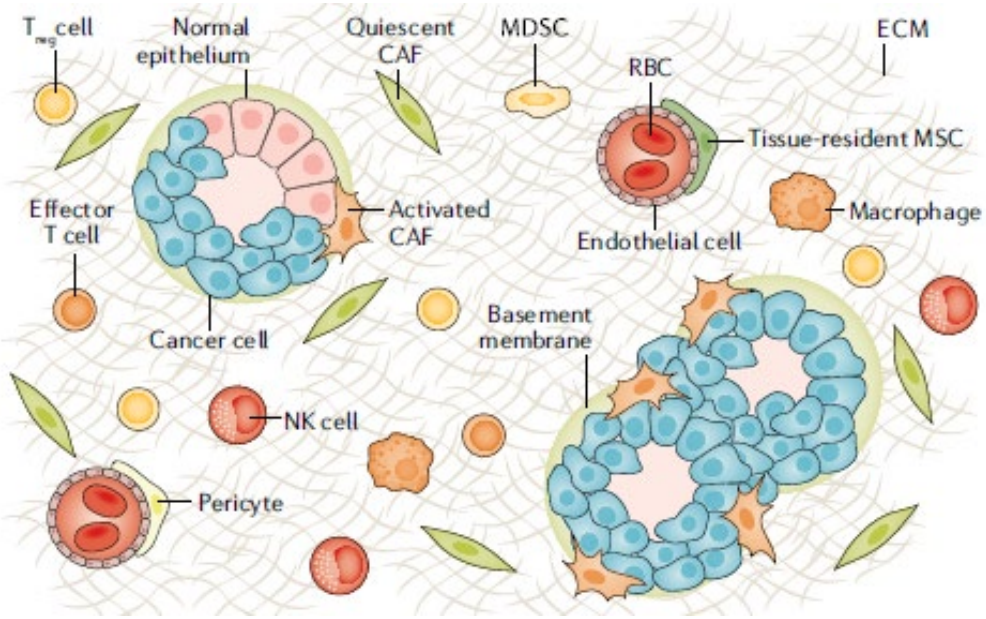
Mixture of tissues

Quon and Morris, 2009



Mixture of cell phases

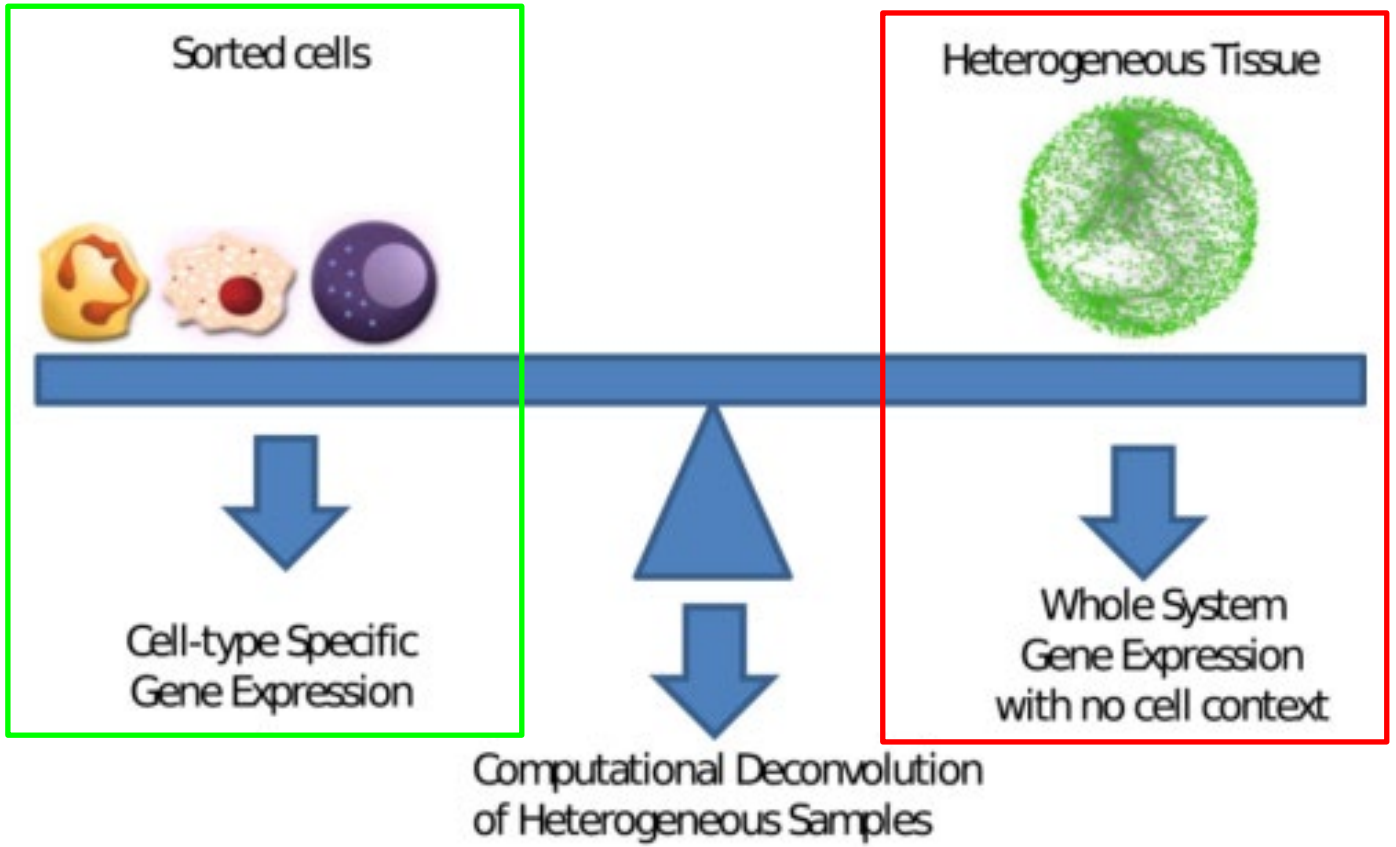
Lu et al, 2003



Mixture of cell populations

Finotello and Trajanoski 2018

Physical methods to analyse the biological medium



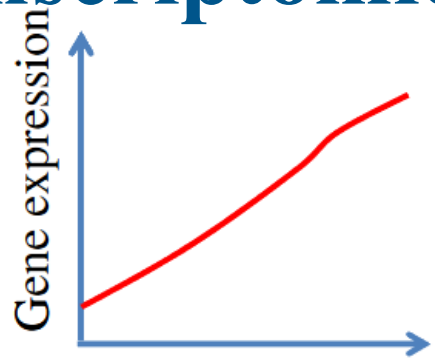
Shen-Orr et al, 2013

Before numerical deconvolution, dilemma between either characterising the individual cell populations (FACS, IHC) or getting a whole transcriptomic(RNASeq, microarray) overview.

Identify the causal transcriptomic driver

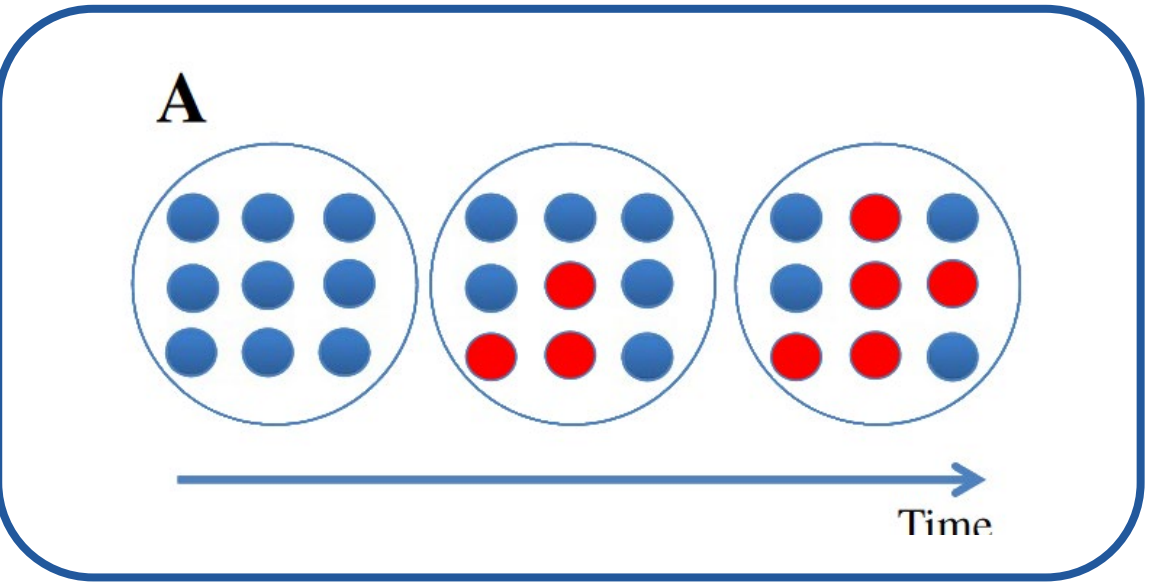
resting cell population 1

activated cell population 1

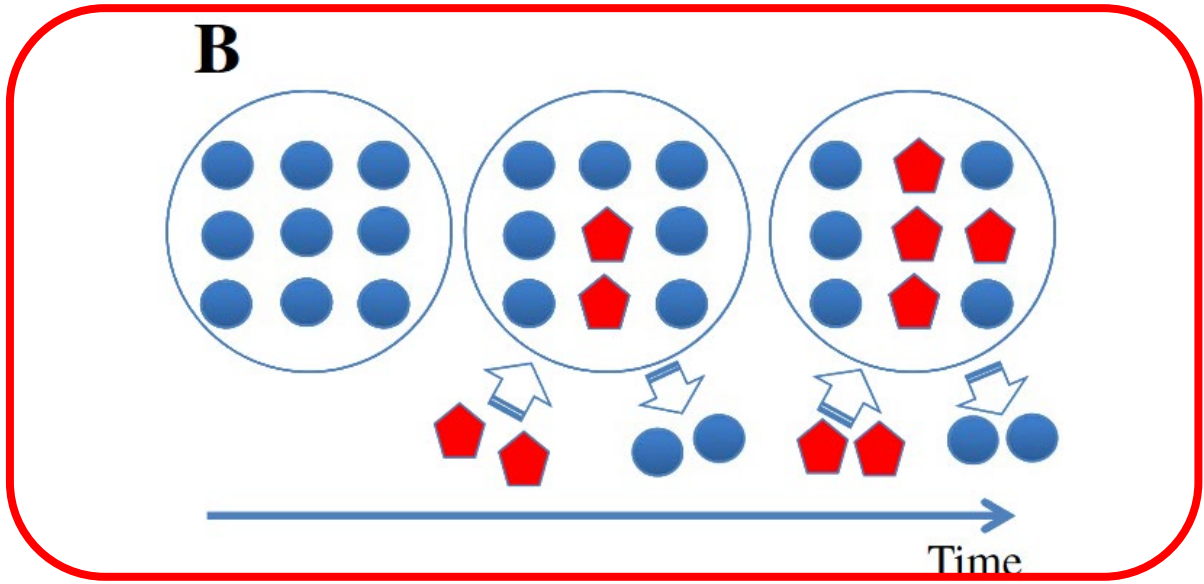


Shoemaker et al. 2012

cell population 2



Scenario A: increase of the gene expression is generated by an **activation** of cell population 1



Scenario B: the gene expression increases due to the **infiltration** of a **new** cell population 2

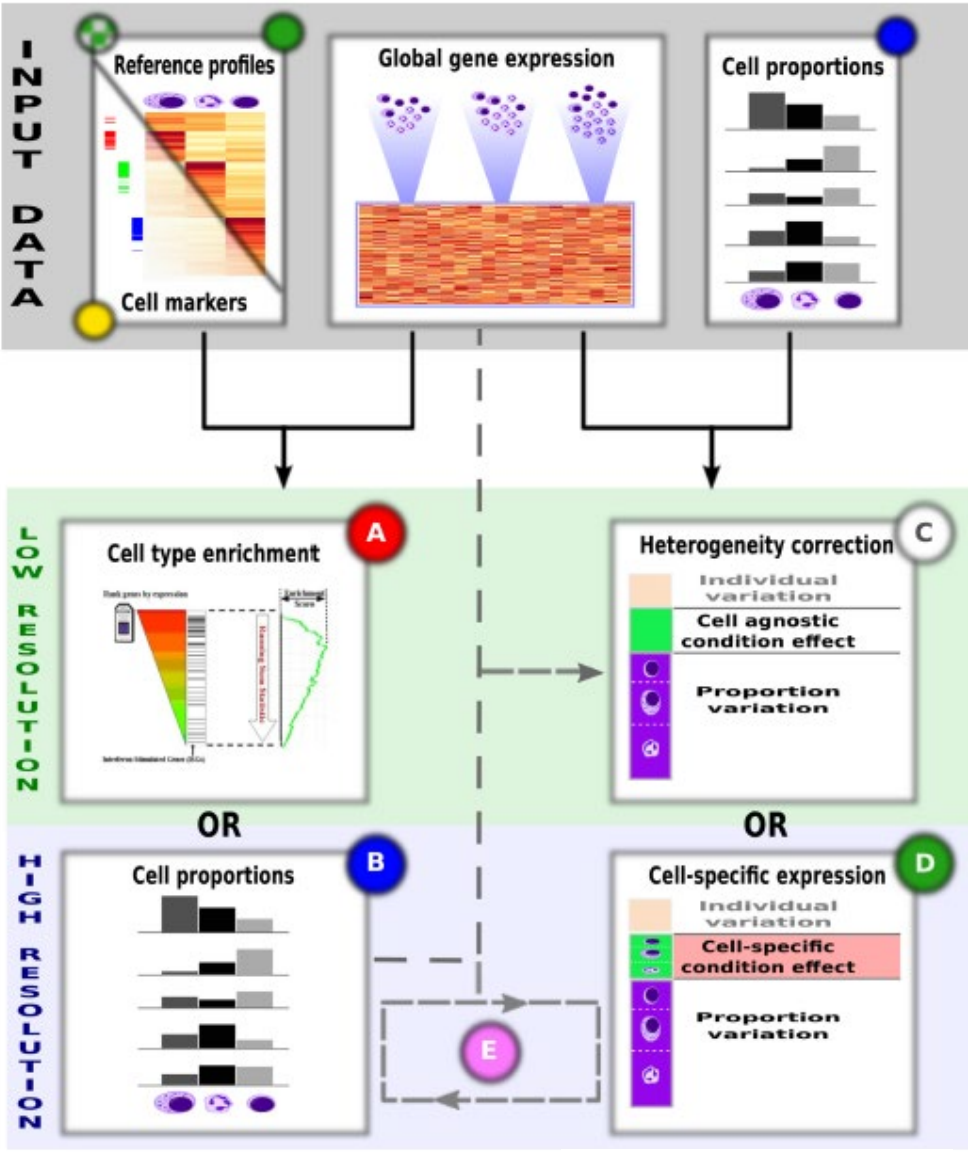
Deconvolution classes

Partial deconvolution

- Estimate the ratios *p* for all individuals with the purified cell signature **X** and bulk mixture **y.**
- Try to infer cell specific expression profiles **X** based on **p** and **y.**

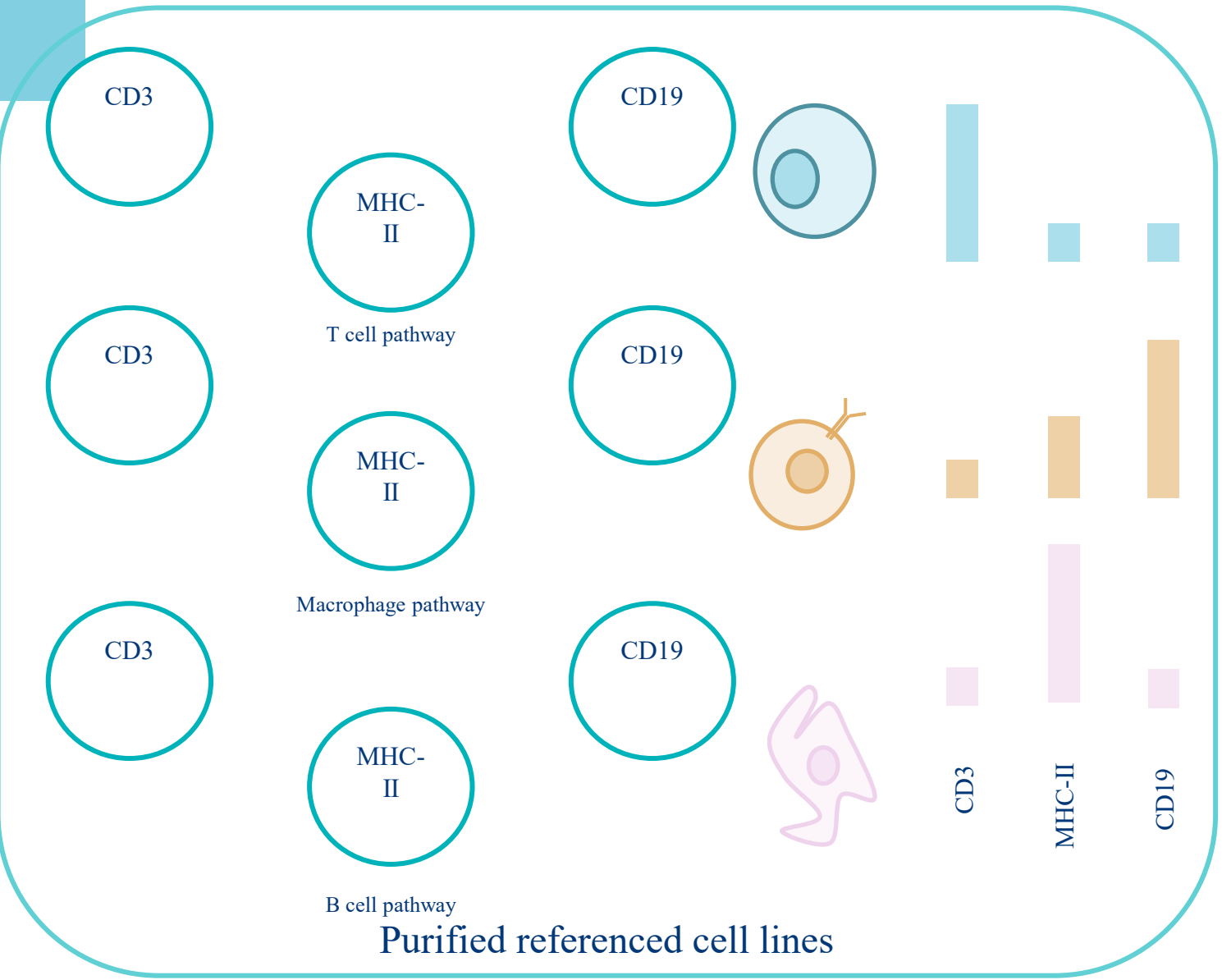
Complete deconvolution

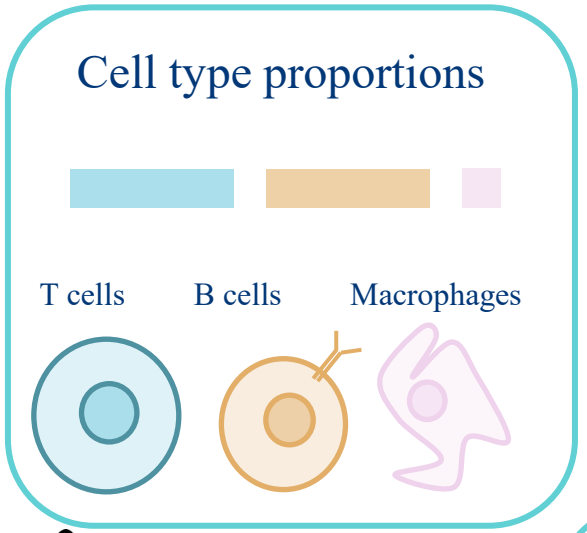
- Try to infer alternatively both **p** and **X** (unsupervised, reference-free methods). Undetermined problem without prior.

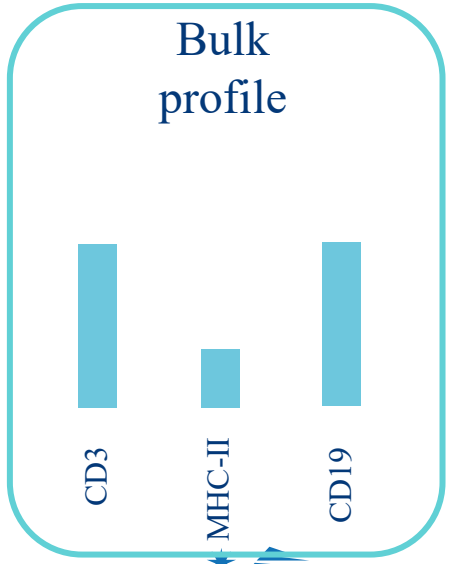


Shen-Orr et al, 2013

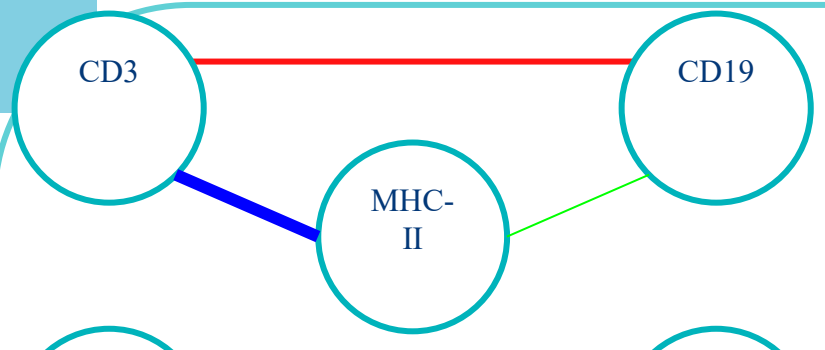
Co-regulated gene networks



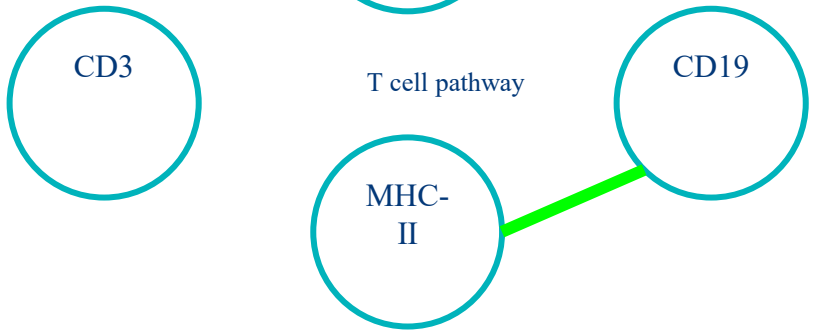




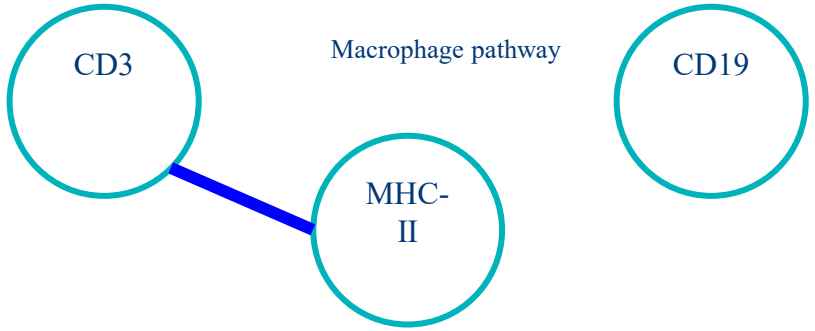
Co-regulated gene networks



T cell pathway



Macrophage pathway



B cell pathway

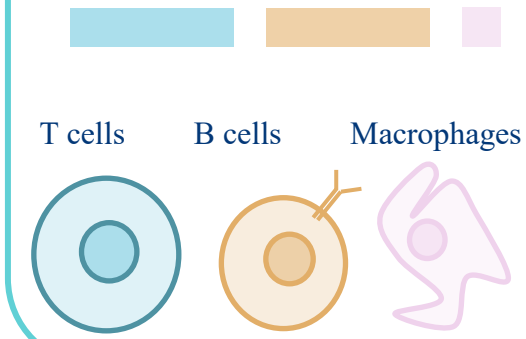
$$\begin{array}{ccc} & \text{CD19} & \text{MHC-II} \\ \begin{pmatrix} \sigma_1 & 0.5 & 0.8 \\ & \sigma_2 & 0.3 \\ & & \sigma_3 \end{pmatrix} & \begin{matrix} \text{CD3} \\ \text{CD19} \\ \text{MHC-II} \end{matrix} \end{array}$$

$$\begin{array}{ccc} & \text{CD19} & \text{MHC-II} \\ \begin{pmatrix} \sigma_1 & 0 & 0 \\ & \sigma_2 & 0.8 \\ & & \sigma_3 \end{pmatrix} & \begin{matrix} \text{CD3} \\ \text{CD19} \\ \text{MHC-II} \end{matrix} \end{array}$$

$$\begin{array}{ccc} & \text{CD19} & \text{MHC-II} \\ \begin{pmatrix} \sigma_1 & 0 & 0.8 \\ & \sigma_2 & 0 \\ & & \sigma_3 \end{pmatrix} & \begin{matrix} \text{CD3} \\ \text{CD19} \\ \text{MHC-II} \end{matrix} \end{array}$$

Purified referenced cell lines

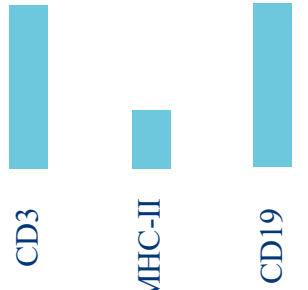
Cell type proportions



×

=

Bulk profile



General principle of cellular deconvolution

Estimate the cellular proportions

Step 1: collection and curation of datasets

Step 2: learn for each cell-type its associated characteristics

Step 3: the deconvolution algorithm itself

Step 4: biological and statistical evaluation

Purified gene expression

Cell types

T cell

B cell

Macrophage

CD3

MHC-II

CD19

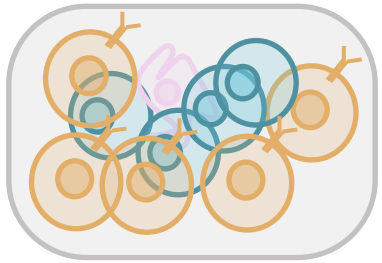
Measured transcriptomic profile

CD3

MHC-II

CD19

Associated bulk sample



Deconvolution

Cell type proportions

T cells

B cells

Macrophages

SERVIER

General principle of cellular deconvolution

Estimate the cellular proportions

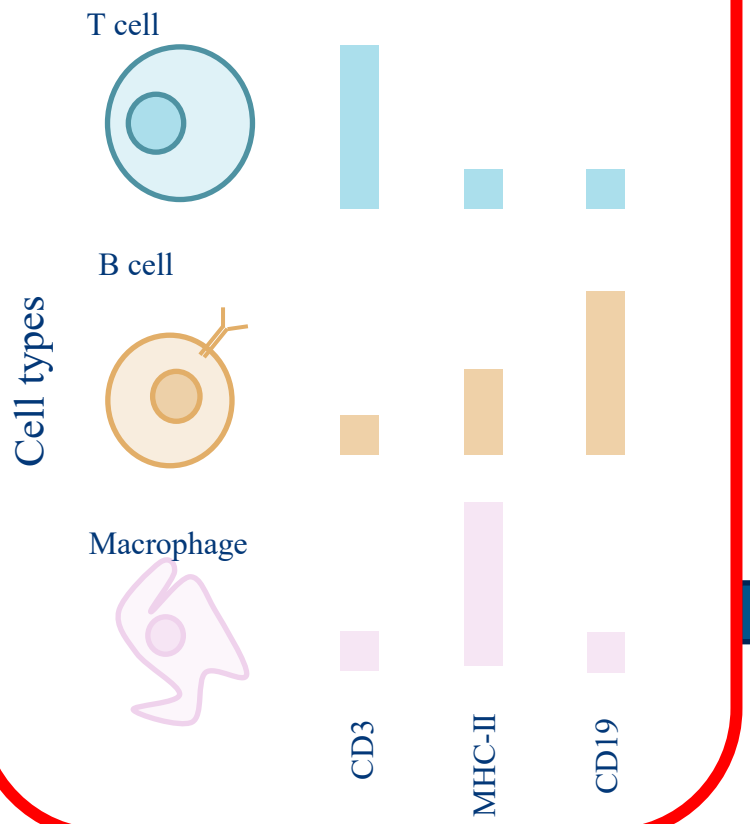
Step 1: collection and curation of datasets

Step 2: learn for each cell-type its associated characteristics

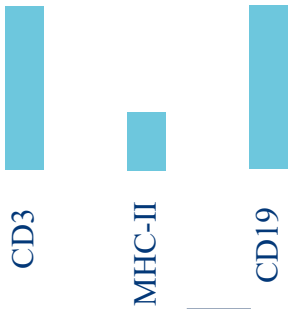
Step 3: the deconvolution algorithm itself

Step 4: biological and statistical evaluation

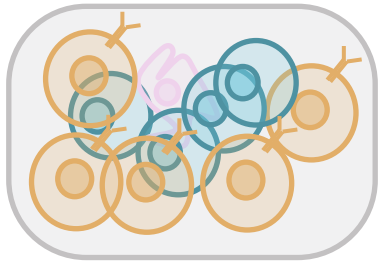
Purified gene expression



Measured transcriptomic profile

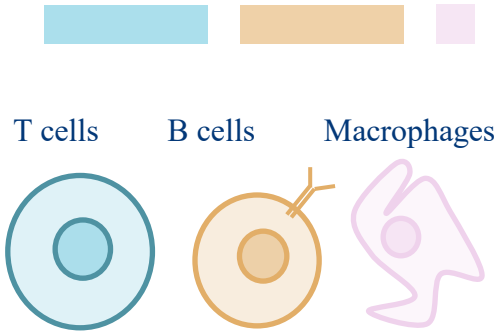


Associated bulk sample



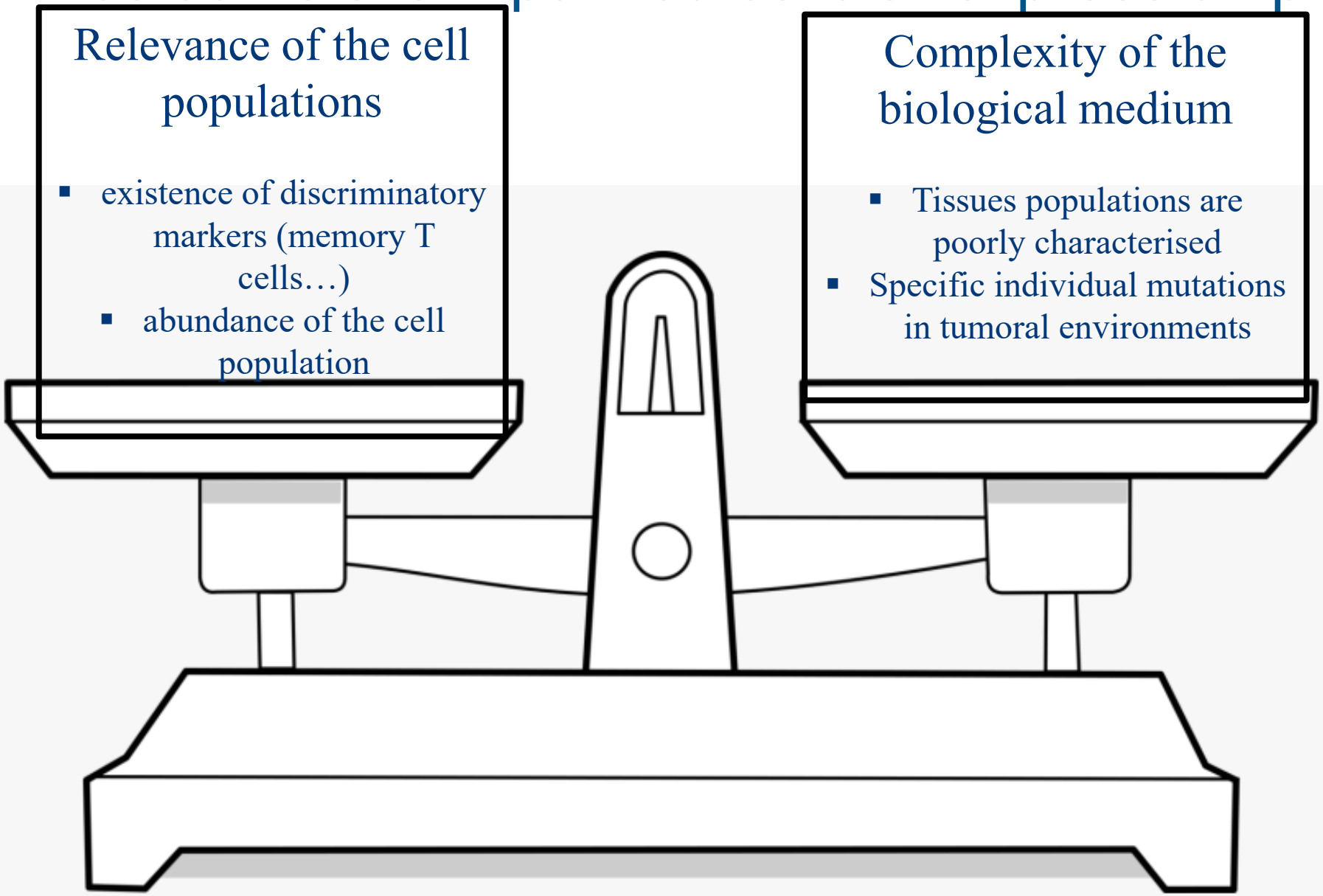
Deconvolution

Cell type proportions

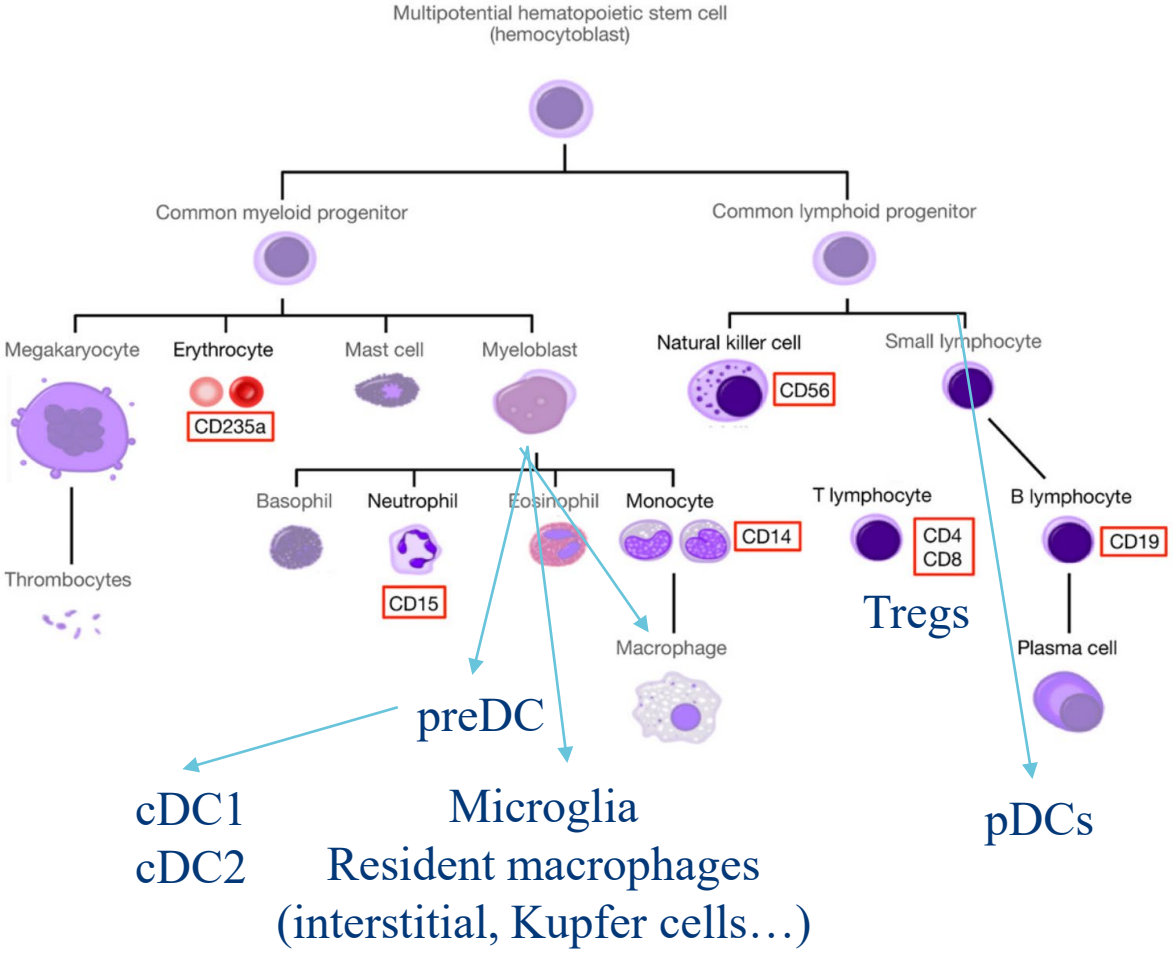


SERVIER

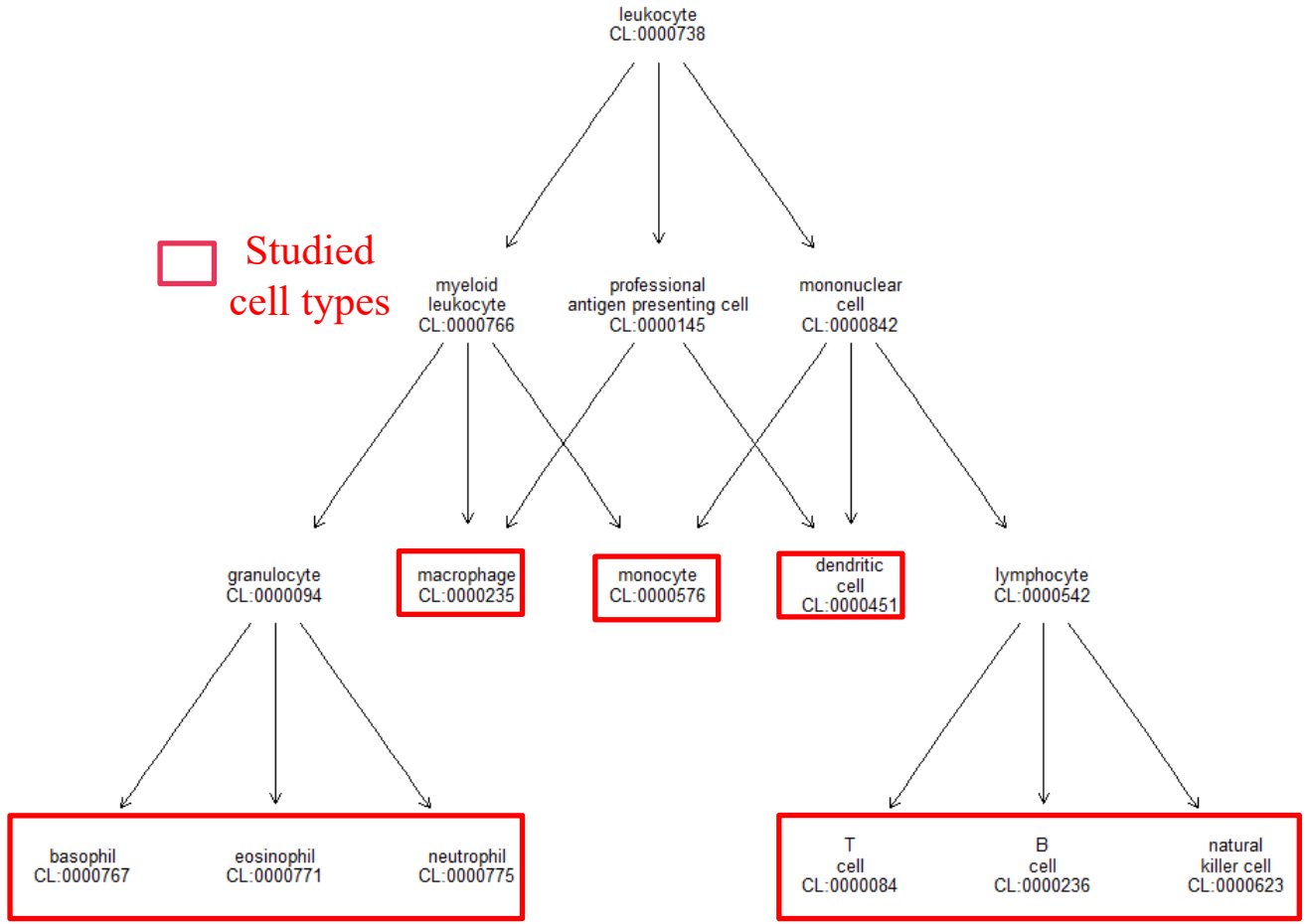
Step 1: select relevant purified cellular expression profiles



Step 1: select relevant purified cellular expression profiles



Hematopoietic stem cell lineage



Automated cell ontology using *ontoProc* (Channing, 2022) package

Step 1: selection of relevant datasets

Array accession	Cell types	Individuals	Samples	Phenotypes	Tissues	Citation
BluePrint	44	354	609	HC, tumoral	(cord) blood, thymus, bone marrow, tonsil, liver	Fernandez et al., 2016
E-MTAB-5640, the Immune Atlas	3	13	29	tumoral	kidney	Chevrier et al., 2017
ENCODE	9	13	37	HC	blood	Encode Project Consortium, 2012
GSE107011	27	13	123	HC	blood	Monaco et al, 2019
GSE137143	3	144	427	HC, auto immune	blood	Kim et al., 2021
GSE149050	4	91	223	HC, auto immune	blood	Panwar et al., 2021
GSE60424	4	20	80	HC, auto immune, Diabetes	blood	Linsley et al., 2014

7 reference RNASeq datasets of purified cell types, covering a large diversity of distinct cell populations (*75 unique cell types,* mostly immune cell types), in *8 distinct tissues* (mostly whole blood) and both *healthy, tumoral and inflammatory* conditions.

SERVIER

Step 2: build a sparse transcriptomic network

Estimate the cellular proportions

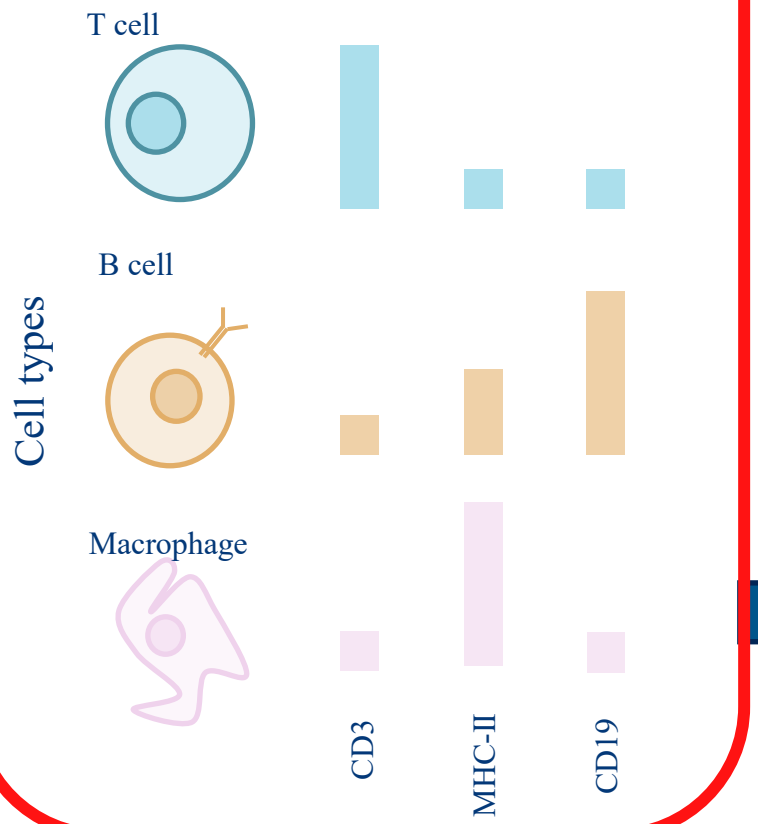
Step 1: collection and curation of datasets

Step 2: learn for each cell-type its associated characteristics

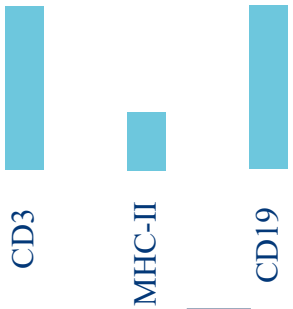
Step 3: the deconvolution algorithm itself

Step 4: biological and statistical evaluation

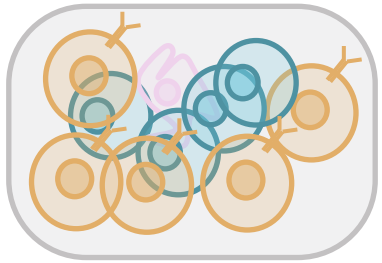
Purified gene expression



Measured transcriptomic profile

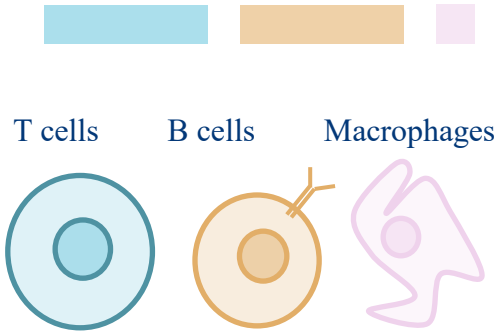


Associated bulk sample



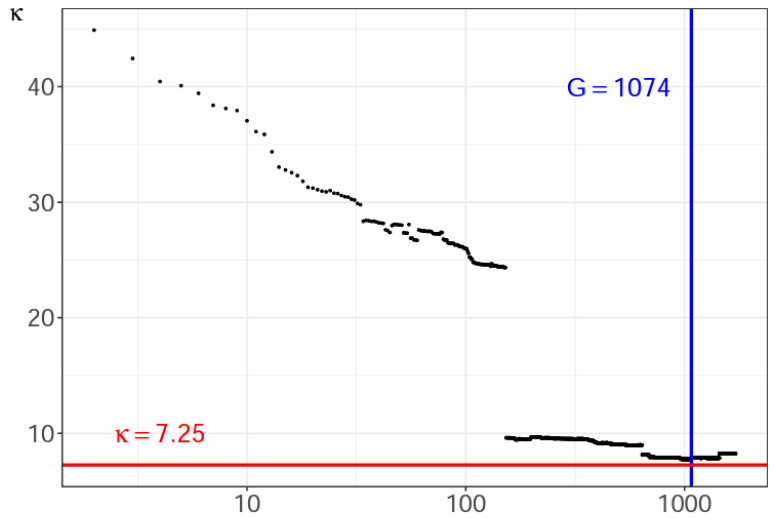
Deconvolution

Cell type proportions

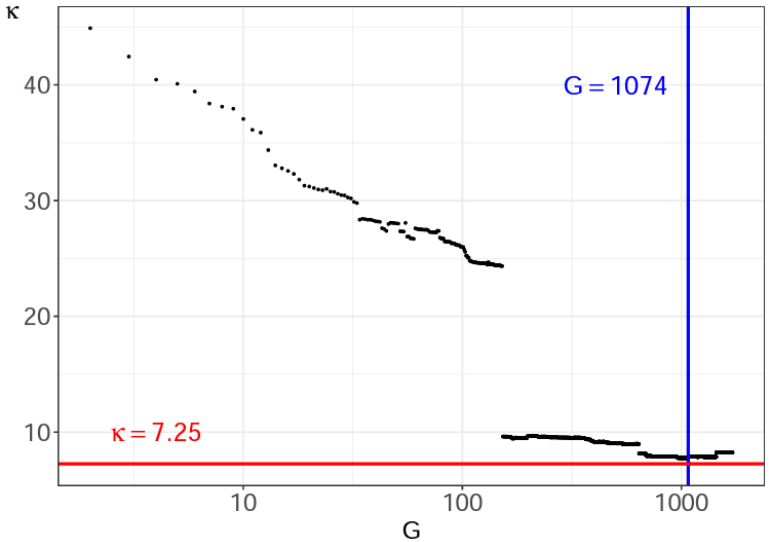


SERVIER

Step 2: learn the sparse GGM for each cell type



In (*Newman et al, 2015*), selection of the *G* genes associated to the lowest *condition number*.

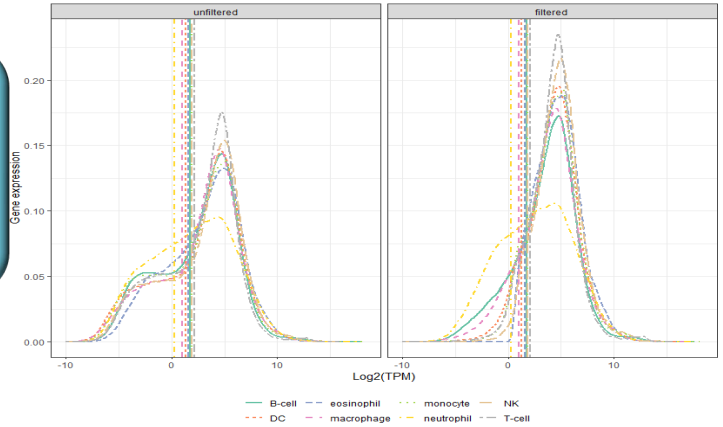


In (*Zuo et al, 2016*), use of INDEED to both learn a sparse representation, and select the most relevant genes.

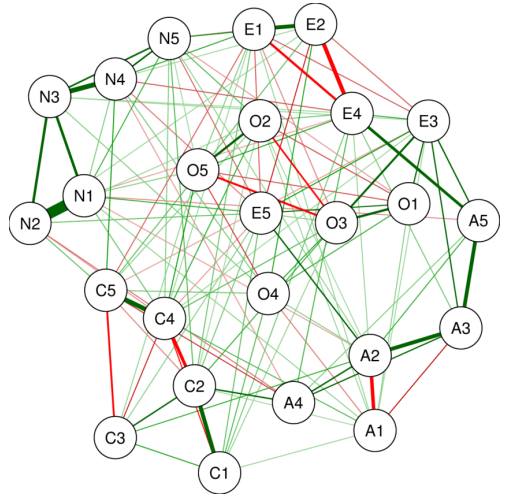
1) Filtering background noise from truly expressed signal

2) Select the most relevant genes

3) Learn a sparse representation of the interactions between the genes



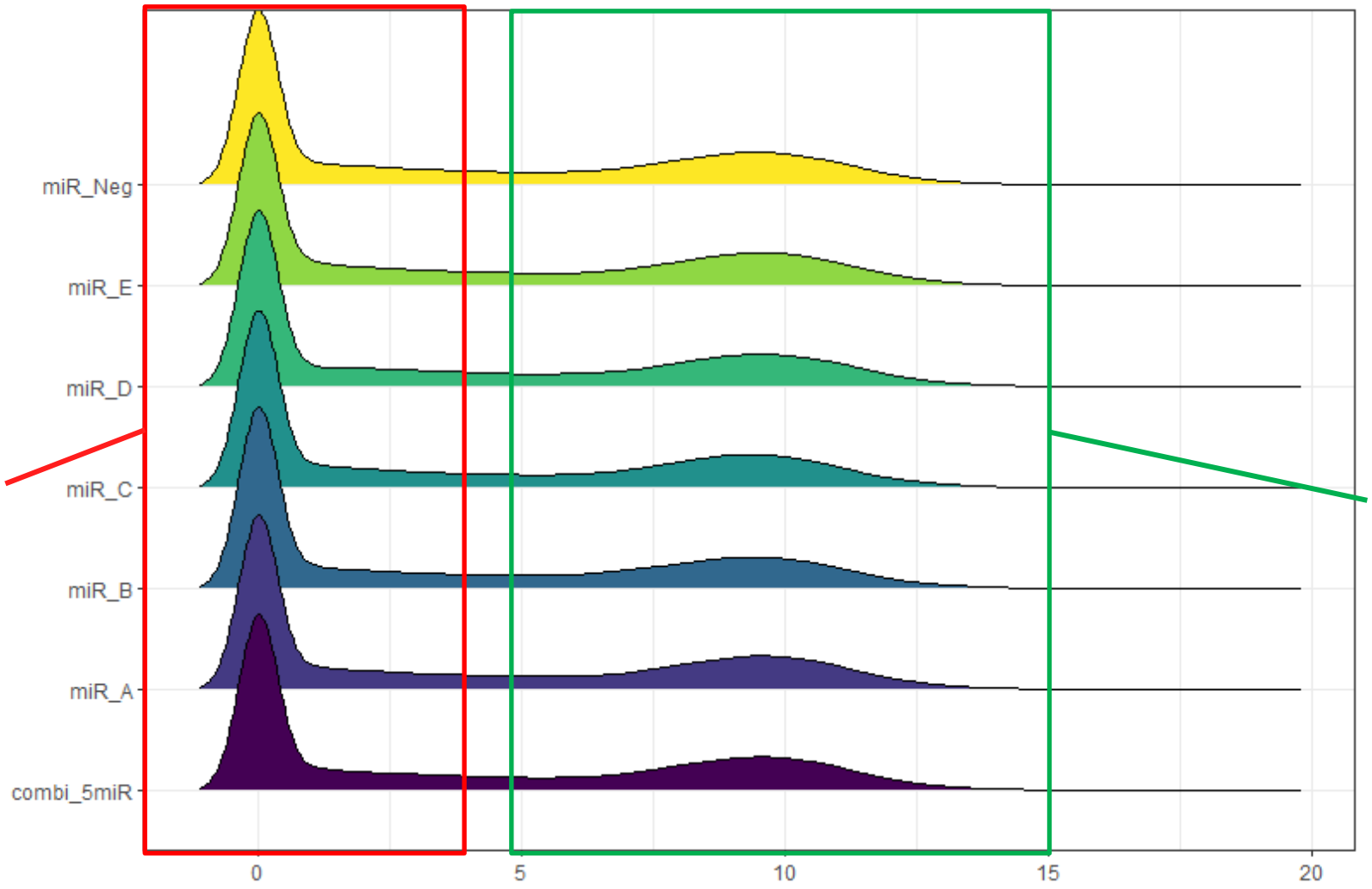
Fitted distributions before and after filtering using zFPKM (*Hart et al, 2013*) process



Nodes represent the genes, and the undirected *edges* the connections between them.

Filtering out noisy expression

First component = noise



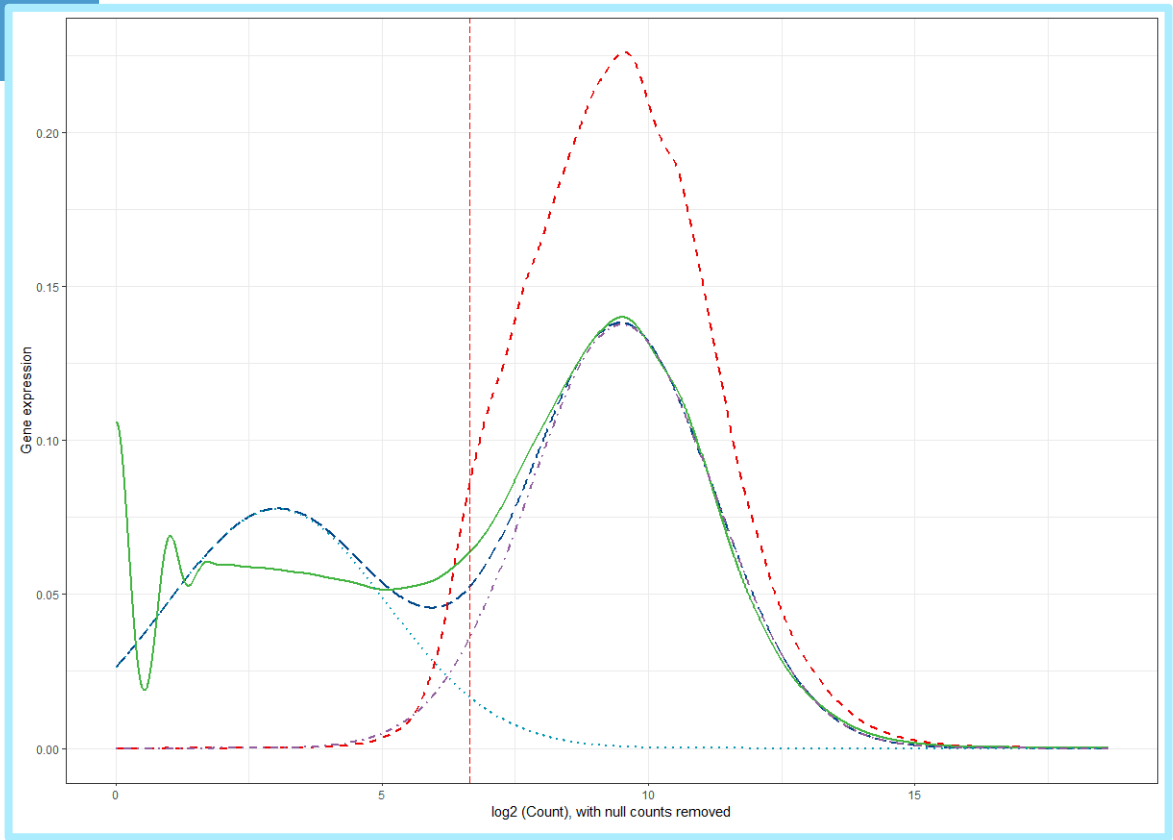
Second component = truly expressed genes

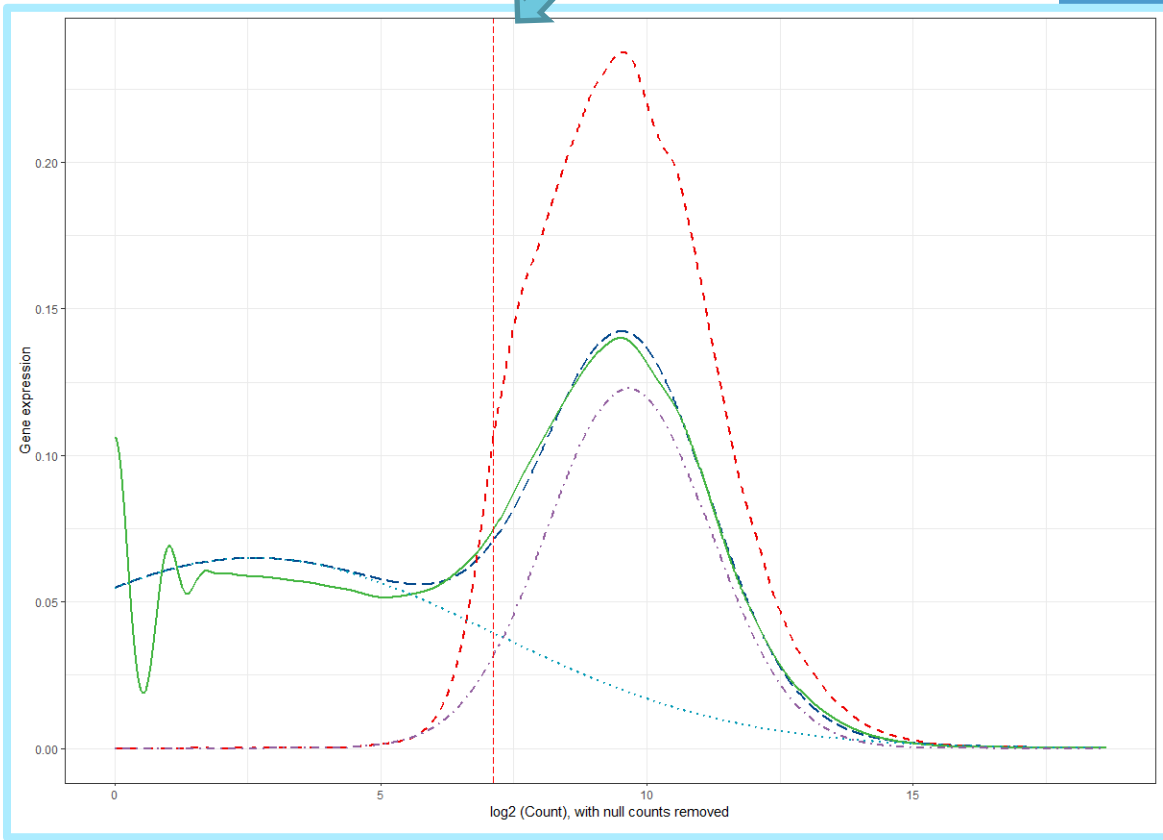
Empirical density distribution of gene counts, after TPM and *log2* normalization

Step 2: build a sparse transcriptomic network

Filtering out noisy expression

5% threshold of the expressed signal





global distribution

Fitted expressed genes

Noise signal

$$\hat{\mu} = \arg\max\left(\mathrm{KDE}(X)\right)$$

$$\hat{\sigma} = \mathrm{MAD}_n(X)\sqrt{\frac{\pi}{2}}$$

Fitted distributions before and after filtering using zFPKM process.

Fitted distributions before and after filtering with truncated Gaussian mixtures (MixNorm, Yin et al, 2020)

Multivariate Gaussian Distributions

Multivariate gaussian distribution

$$\mathbf{X}_{1:G,j} \sim \mathcal{N}_G(\mu_j, \boldsymbol{\Sigma}_j)$$

$\mu = \mathrm{E}(\mathbf{X})$ — Mean vector

$\boldsymbol{\Sigma}_{i,l} = \mathrm{Cov}(X_i, X_l), \forall 1 \leq i, l \leq G$ — Covariance matrix

$$\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{Q}_j \mathbf{D}_j \boldsymbol{Q}_j^\top$$

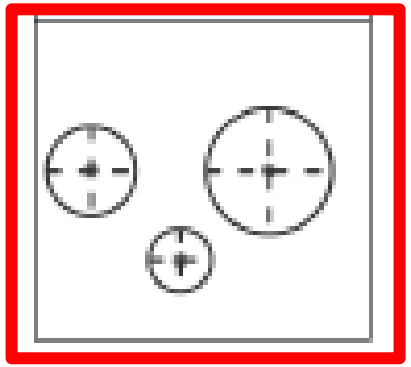
- λ_j controls the overall volume
- $\boldsymbol{Q}_j$ controls the directions
- $\boldsymbol{D}_j$ permutation matrix, control the shape

14 available parametrisations, included in 3 super-families

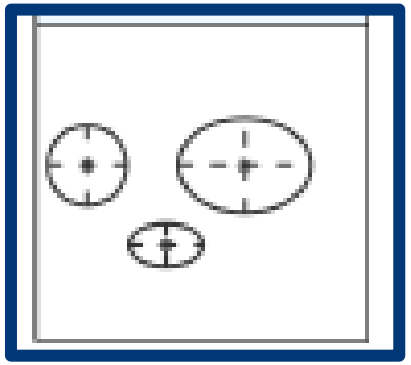
EII

volume — shape — orientation

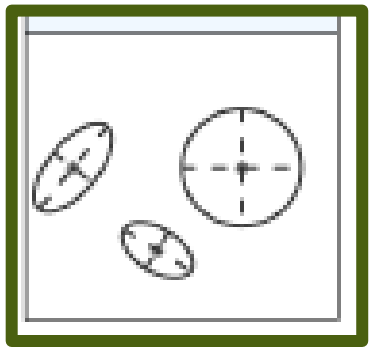
E: equal
V: variable
I: invariant (or identity)



Spherical family



Diagonal family



General (or *ellipsoidal*) family

Banfield and Raftery (1993) and Celeux and Govaert (1995)

Step 2: build a sparse transcriptomic network

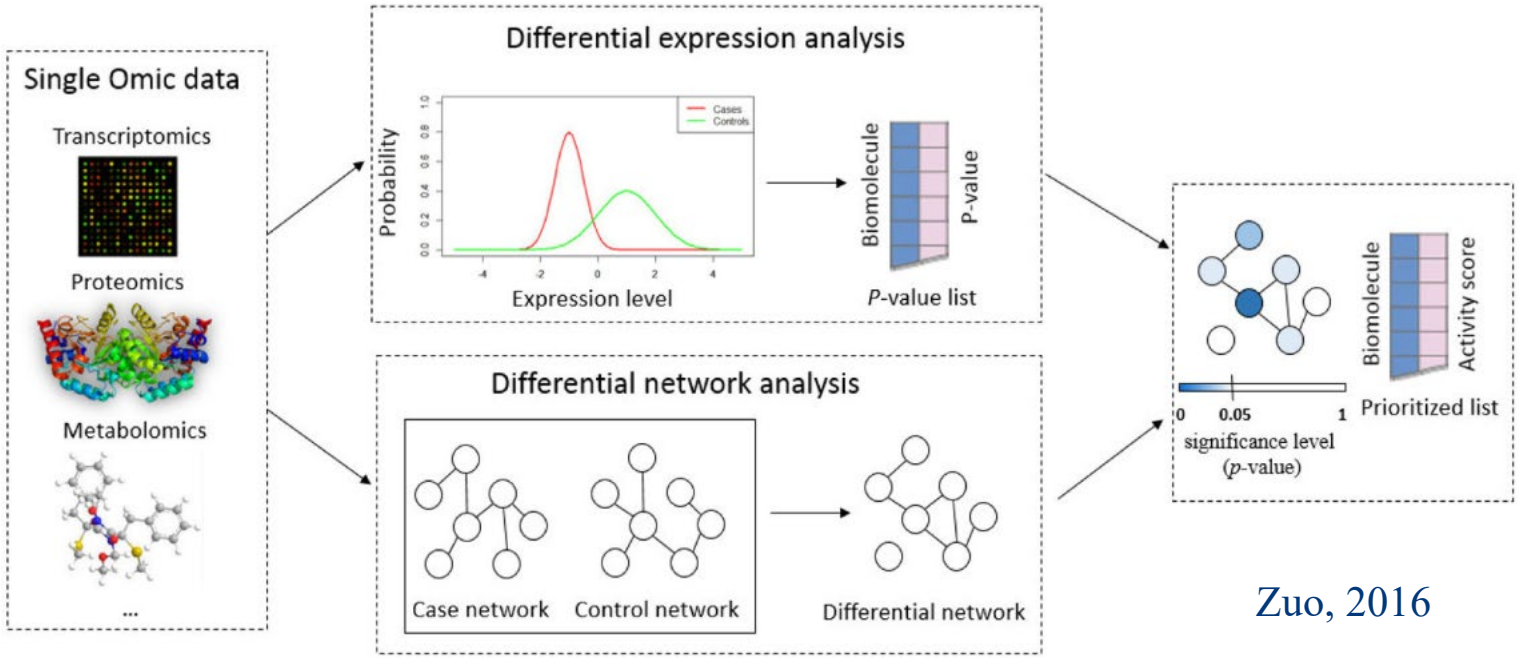
Markov networks

Multivariate Gaussian distribution $f_{\zeta_j}(\boldsymbol{X}) = \det(2\pi\boldsymbol{\Sigma}_j)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu}_j)\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_j)^\top\right), \quad \zeta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

Penalised Lasso (useful when $N < G$) (*Bickel and Levina*, 2008)

$$\boldsymbol{\Sigma}_j^{\text{Lasso}} = \arg\max_{\boldsymbol{\Sigma}_j} \ell_{\boldsymbol{\Sigma}_j}(\boldsymbol{X}_j) = \arg\max_{\boldsymbol{\Sigma}_j} \Big[\underbrace{\log\left(\det\left(\boldsymbol{\Sigma}_j^{-1}\right)\right) - \overbrace{\boldsymbol{X}_j^\top \boldsymbol{\Sigma}_j^{-1} \boldsymbol{X}}^{\text{tr}(\boldsymbol{S_j}\boldsymbol{\Sigma}_j^{-1})}}_{\text{MLE}} - \underbrace{\lambda||\boldsymbol{\Sigma}_j^{-1} * (1 - \boldsymbol{W}_j)||_1}_{\text{Penalty term}} \Big]$$

Estimate a sparse covariance structure using gLasso (*Friedman et al, 2008)* algorithm



Zuo, 2016

$$\Omega = \{\omega_{gl}, \text{ where } \Delta_{gl} = \rho_{gl}^A - \rho gl^B \neq 0\}$$

Differential network between conditions 1 and 2

- Activity scores: for each gene, sum of the z-scores of the *neighbour* differential values
- *Neighbour*: gene statistically differentially connected (permutation test) to our gene of interest

An overview of INDEED: input is transcriptomics data and the output is a prioritized ranked list gene based on the activity score defined within INDEED.

SERVIER

From GGMs to GBNs

Sparse graphical GGM

$$\Theta = (\theta_{il}, \quad (i,l) \in \{1,\dots,G\}) = \Sigma^{-1}$$

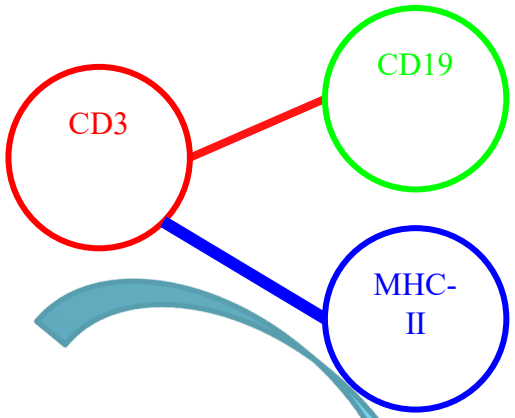
+ Inverse of the sparse covariance structure (=*precision matrix*) has nice interesting properties, after some normalization

$$\rho_{i,l|V\setminus\{i,l\}} = -\frac{\theta_{il}}{\sqrt{\theta_{ii}\theta_{ll}}}$$

$$\forall (i,l) \in V, \quad X_i \perp\!\!\!\perp X_l \Leftrightarrow \rho_{i,l|V\setminus\{i,l\}} = 0$$

- Estimator is shrunk: find asymptotically the good support (= the true zeros (*Meinshausen and Bühlmann*, 2006 // *Banerjee and other* 2007)) but the penalized estimator tends to underestimate true correlation)

$$V = \{1,\dots,G\} \qquad E = \{i,l \in V^2, i \neq l\}$$



$$\begin{array}{c} \\ \text{CD3} \\ \text{CD19} \\ \text{MHC-II} \end{array} \begin{array}{c} \begin{array}{ccc} \text{CD3} & \text{CD19} & \text{MHC-II} \end{array} \\ \begin{pmatrix} \sigma_1 & 0.5 & 0.8 \\ & \sigma_2 & 0 \\ & & \sigma_3 \end{pmatrix} \end{array}$$

Learn the MLE covariance matrix with constrained zeros

General formula from conditional distribution to global joined distribution, to retrieve the global multivariate Gaussian distribution.

Directed GBNs

1. Graph triangulation (adds a cord to any cycle above three vertices)
2. Hypothesis: from a triangular graph, able to orient the edges and learn the structure of a GBN
3. From the factorization of a GBN, learn easily the conditional distribution of each node to its parents

GBNs (Gaussian Bayesian networks)

Definition (Bayesian Network (BN))

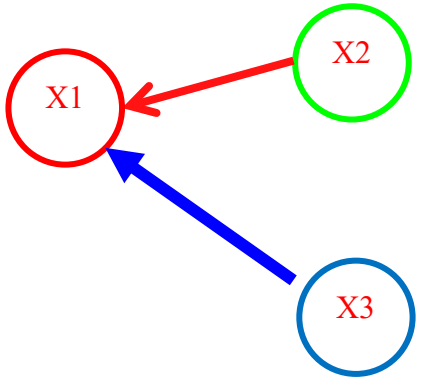
A Bayesian network is a joint distribution over a set of random variables, defined with :

- *a directed acyclic graph (DAG) G whose each node $V_i \in \mathbf{V}$ depicts a random variable X_i*
- *a global probability distribution X (with parameters Θ), admitting local factorisation for each variable X_i, depending only on its parents for each node $P\Pi_{X_i}$.*

Factorisation of the joint distribution in a BN

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P\left(X_i | \Pi_{X_i}\right) \text{ where } \Pi_{X_i} = \text{parents of } X_i$$

As we have conditionally probability distribution, such a model respects normalisation constraint : $\sum_{i=1}^{n} P(X_i, \Pi_{X_i}) = 1$. Besides, local distribution for each variable X_i is represented by a CPT.



$$\begin{array}{c} \begin{array}{ccc} \text{CD3} & \text{CD19} & \text{MHC-II} \end{array} \\ \begin{pmatrix} \sigma_1 & 0.5 & 0.8 \\ & \sigma_2 & 0 \\ & & \sigma_3 \end{pmatrix} \end{array} \begin{array}{l} \text{CD3} \\ \text{CD19} \\ \text{MHC-II} \end{array}$$

$$\mathbb{P}(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \mathbb{P}(\boldsymbol{X}_1 | \boldsymbol{X}_2, \boldsymbol{X}_3)\mathbb{P}(\boldsymbol{X}_3)\mathbb{P}(\boldsymbol{X}_2)$$

Product of Gaussian distributions

$$(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \left(\sum_{j=1}^{J} \boldsymbol{\Sigma}_j^{-1}\right)^{-1} \left(\sum_{j=1}^{J} \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j\right) \qquad \boldsymbol{\Sigma} = \left(\sum_{j=1}^{J} \boldsymbol{\Sigma}_j^{-1}\right)^{-1}$$

From conditional to joint distribution

$$\boldsymbol{X}_1 \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{X}_2 | \boldsymbol{X}_1 \sim \mathcal{N}_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{\Omega})$$

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{A}^\top \\ \boldsymbol{A}\boldsymbol{\Sigma} & \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top + \boldsymbol{\Omega} \end{pmatrix}\right)$$

Estimate the cellular proportions

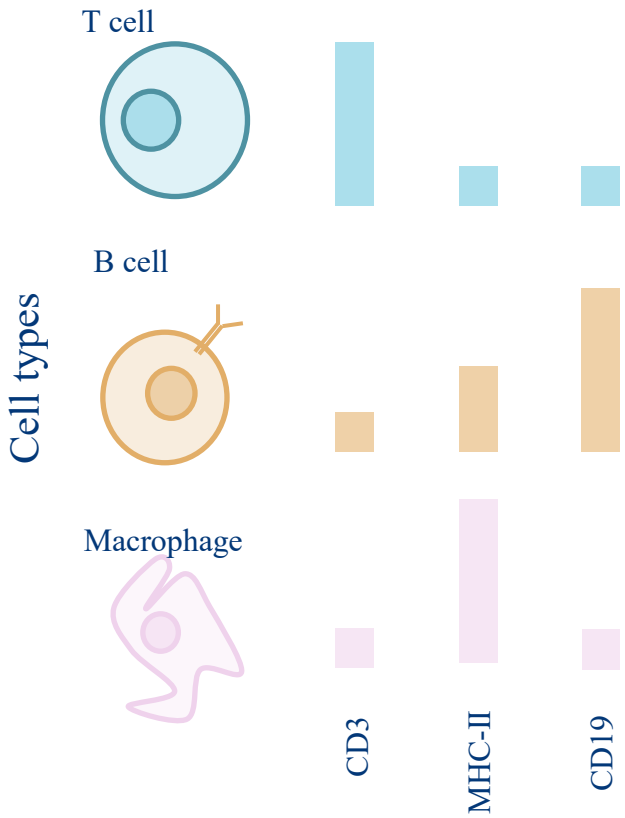
Step 1: collection and curation of datasets

Step 2: learn for each cell-type its associated characteristics

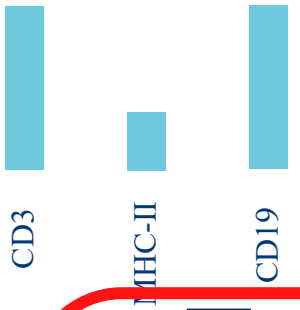
Step 3: the deconvolution algorithm itself

Step 4: biological and statistical evaluation

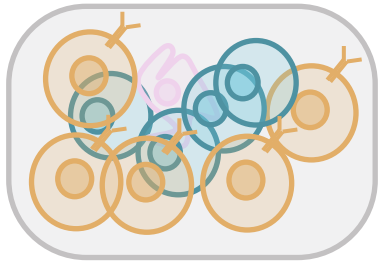
Purified gene expression

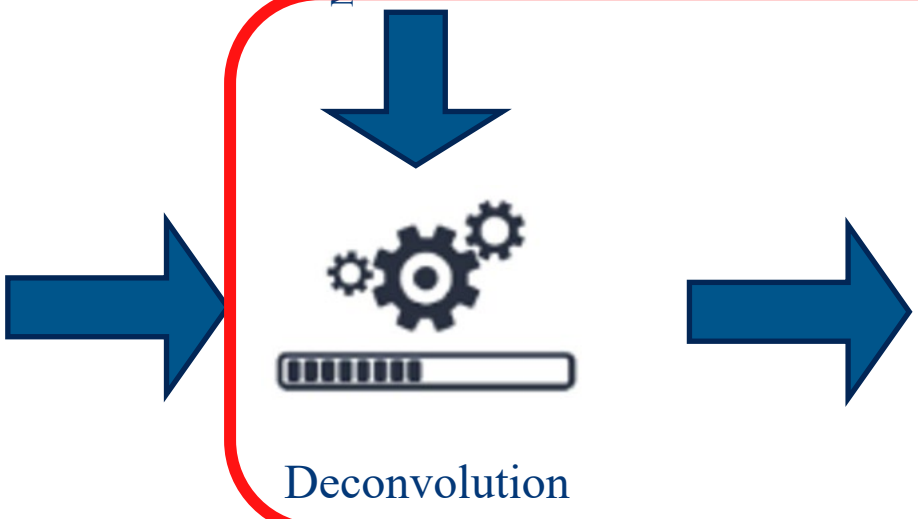


Measured transcriptomic profile

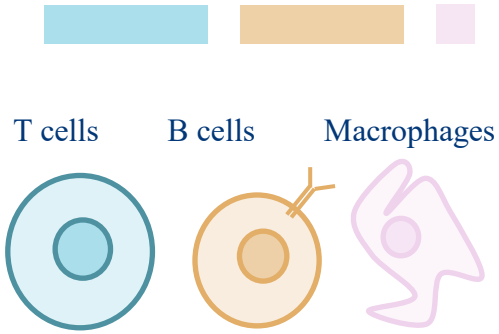


Associated bulk sample





Cell type proportions



Step 3: estimate the cell ratios

Batch effect

- **Sequencing method** (ssRNA-Seq, RNA-Seq, microarray)
- **Gene annotation and library**
- **Normalization** (TPM, CPM, raw counts, …)

Phenotypical conditions

- **Heterotopic** conditions (highly dependent on the tissue condition)
- **Tumoral environments**
- Poorly described tissues

Cellular distributions

- **Rare cell types**
- **"Spillover-effect"**
- **Model gene distributions** (truncated and discrete by nature)
- Decorrelation between cell abundance and cell transcriptome
- **Pathways**

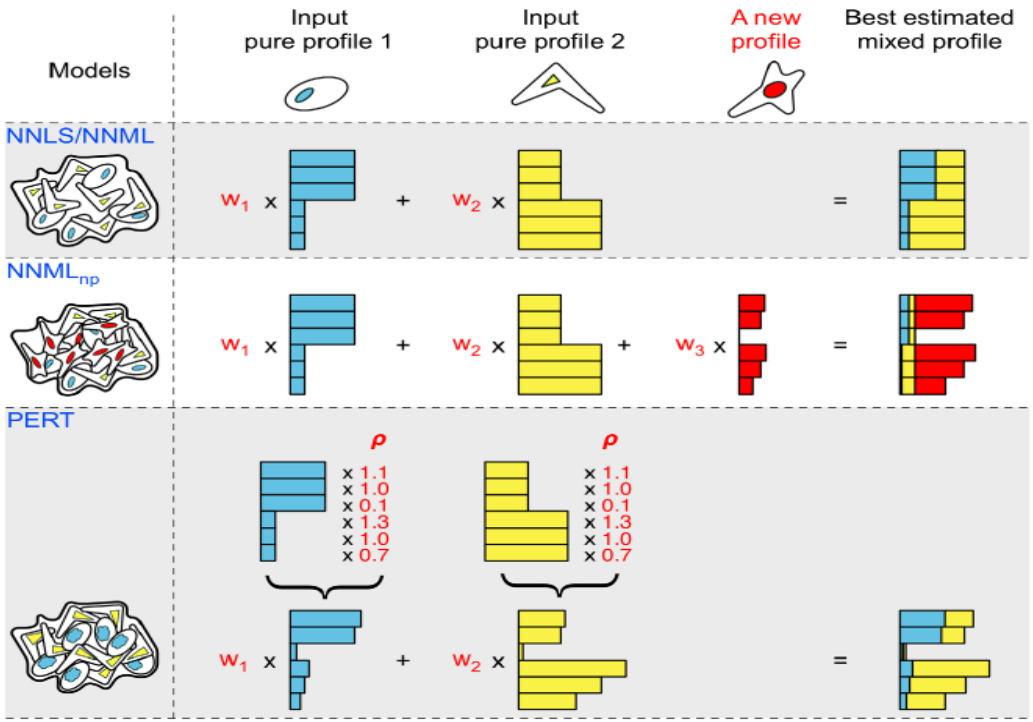
Main challenges to cope with cellular deconvolution

SERVIER

Main deconvolution categories for cellular ratios estimation

Marker-based

- Abundance scores (dtangle algorithm, …)
- Enrichment scores (GSEA, xCell, …)



Regression

- Model the distribution of gene counts as a linear combination of the individual cell types distributions
- Variants integrate robustness to outliers, "spill-over effects", …

canonical LDA model

Add an unknown noise component

Perturbation (in the same proportions) of the purified expression profiles

Qiao et al, 2012

Estimate the ratios from the reference signature and bulk mixture

Step 3: estimate the cellular ratios

$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,J} \\ \vdots & \ddots & \vdots \\ x_{G,1} & \cdots & x_{G,J} \end{pmatrix} \times \begin{pmatrix} p_{1,1} & \cdots & p_{1,N} \\ \vdots & \ddots & \vdots \\ p_{J,1} & \cdots & p_{J,N} \end{pmatrix} = \begin{pmatrix} y_{1,1} & \cdots & y_{1,N} \\ \vdots & \ddots & \vdots \\ y_{G,1} & \cdots & y_{G,N} \end{pmatrix}$$

X purified cellular profiles

Y bulk expression

p cell ratios

$$\begin{pmatrix} x_{G_1,1} & \cdots & 0 \\ 0 & x_{G_2,2} & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_{G_k,k} \end{pmatrix}$$

marker-based

$$\begin{cases} \sum_{j=1}^{J} p_j = 1 \\ \forall j \in \{1, \ldots, J\}, \quad p_j \geq 0 \end{cases}$$

Bulk expression is computed as the weighted linear average of each purified cellular expression profile.

$$\mathbf{y}_i = \mathbf{X} p_i$$

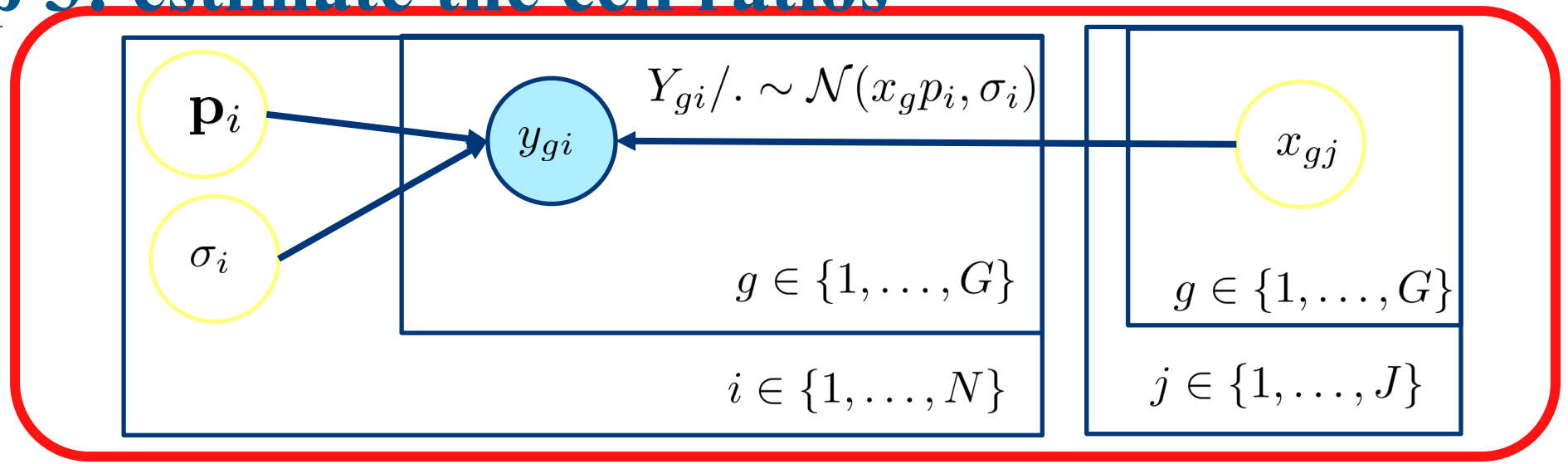
matricial form

$$y_{gi} = \sum_{j=1}^{J} x_{gj} p_{ji}$$

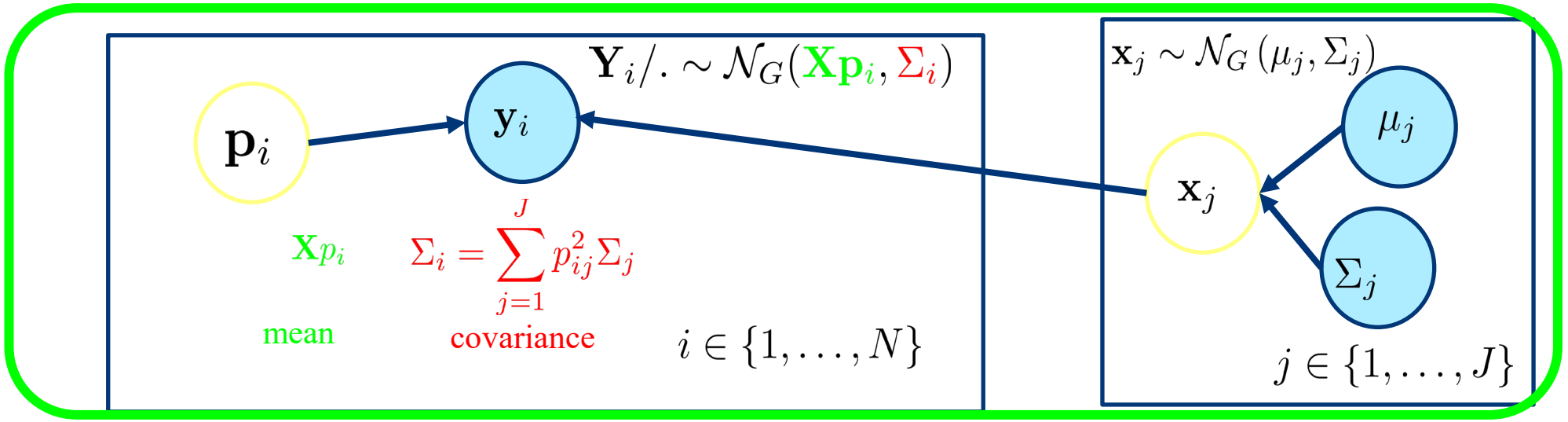
algebraic form

SERVIER

Step 3: estimate the cell ratios



Graphical model of the canonical linear regression modelling. The expression of a given gene in each cell population is supposed *fixed* and *independent* from the others.



Graphical model of our multivariate modelling: the observed variables are *stochastic*, and the genes *interplay* together.

Step 3: estimate the cell ratios

$$\hat{p}_i = \arg\min_{\hat{p}_i} ||\mathbf{X}\hat{p}_i - y_i||^2 \quad \hat{p}_i^{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}_i \quad \hat{\mathbf{p}}_{\text{MLE}} = \arg\max_{\mathbf{p}} \left(\mathbb{P}_{\mathbf{p}}(y_{1:G}|\mathbf{X}_{1:G,1:J})\right) = \arg\max_{\mathbf{p}} \left(\prod_{g=1}^{G} \mathbb{P}_{\mathbf{p}}(y_g|\mathbf{x}_{.g})\right)$$

With the Gaussian-Markov assumptions, OLS is the best *BLUE* estimator and equal to the MLE estimate.

$$\ell_{\mathbf{y}|\mathbf{X},\Sigma}(\mathbf{p}) = C + \log\left(\det\left(\sum_{j=1}^{J} p_j^2\Sigma_j\right)^{-1}\right) - \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{p})^\top\left(\sum_{j=1}^{J} p_j^2\Sigma_j\right)^{-1}(\mathbf{y}-\mathbf{X}\mathbf{p})$$

Deriving this quantity with *matrix calculus* is computable, but optimizing this quantity is intractable with non convex optimization (two local extrema, only one corresponding to the true MLE)

$$\begin{cases} p_j &= \frac{e^{p_j}}{\sum_{j=1}^{J-1} e^{p_j}+1}, j < J \\ p_J &= \frac{1}{\sum_{j=1}^{J-1} e^{p_j}+1} \end{cases}$$

- Descent-gradient based method to learn the MLE.
- Parametrization (use of exponentials) to ensure the non-negativity and sum-to-one constraints (to be compared with *Lagrangian multiplier*)

SERVIER

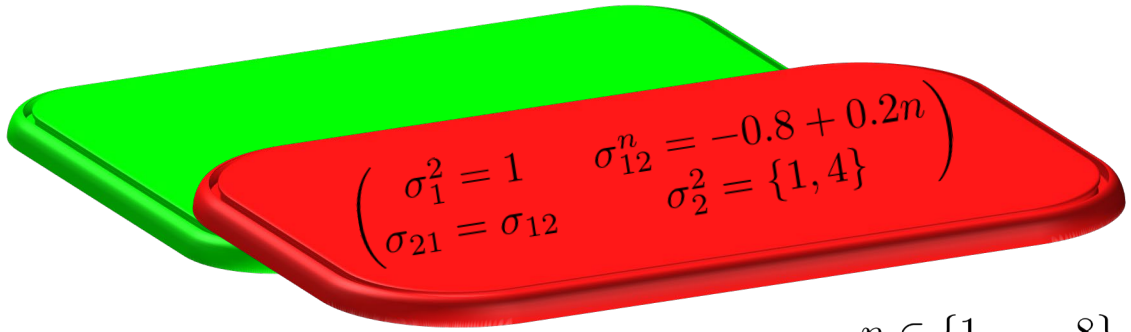
Simulation results with two genes

Generation of random purified cellular expression profile, independently for each individual and each cell population

$$\mathbf{X}_j \sim \mathcal{N}_2(\mu_j, \Sigma_j)$$

	cell type 1	cell type 2
gene 1	20	22
gene 2	22	20

$\mu_{1:2,1:2}$



$$\Sigma_{1:2,1:2,1:2} = (\Sigma_{1:2,1:2}, \Sigma_{1:2,1:2})^{n \in \{1,\dots,8\}}$$

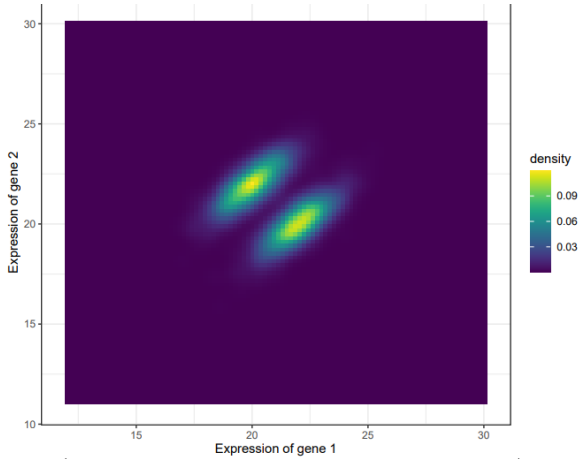
Test several levels of cell proportion disequilibrium:

- ❑ Scenario 1: $p = (0.5, 0.5)$
- ❑ Scenario 2: $p = (0.95, 0.05)$

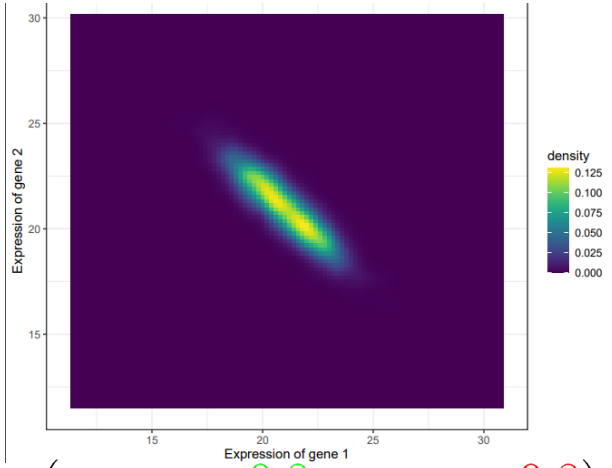
Generation of *N=2000* bulk samples ***Y***

$$\mathbf{y} \sim \mathcal{N}_2(\mathbf{X}p, \Sigma)$$

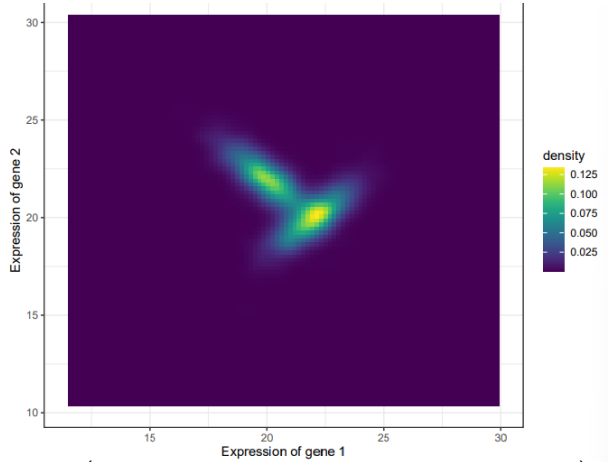
$$\begin{pmatrix} y_{1,1} = \sum_{j=1}^{2} p_j x_{1,j} & \cdots & y_{1,2000} \\ y_{1,2} = \sum_{j=1}^{2} p_j x_{1,j} & \cdots & y_{2,2000} \end{pmatrix}$$



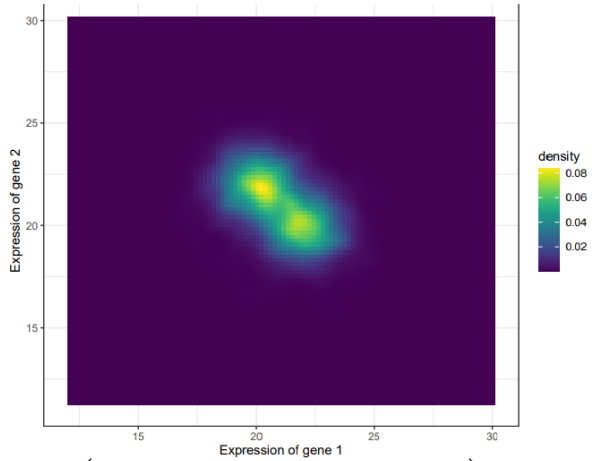
$(\sigma_{121} = 0.8, \sigma_{122} = 0.8)$



$(\sigma_{121} = -0.8, \sigma_{122} = -0.8)$

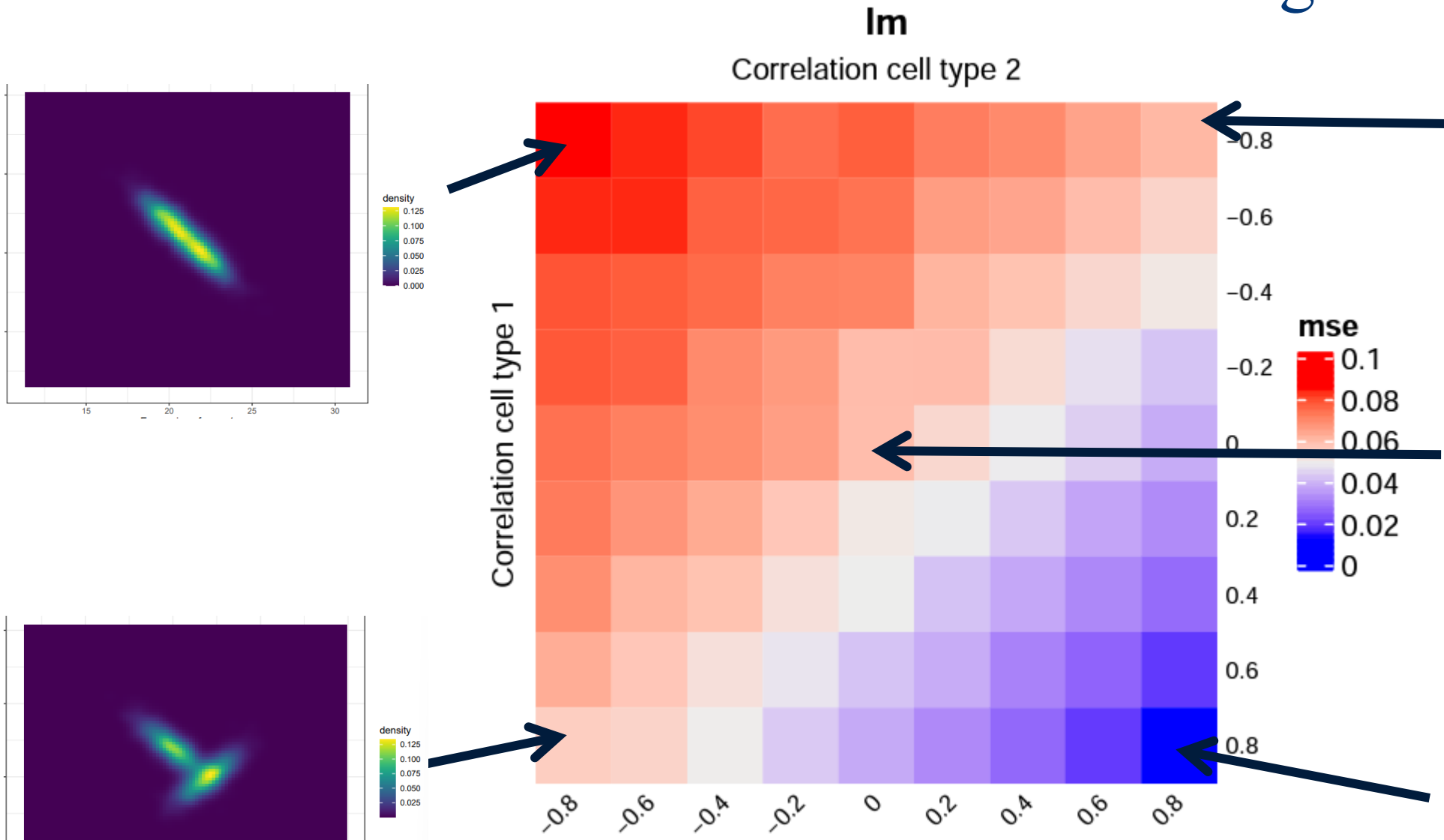


$(\sigma_{121} = -0.8, \sigma_{122} = 0.8)$



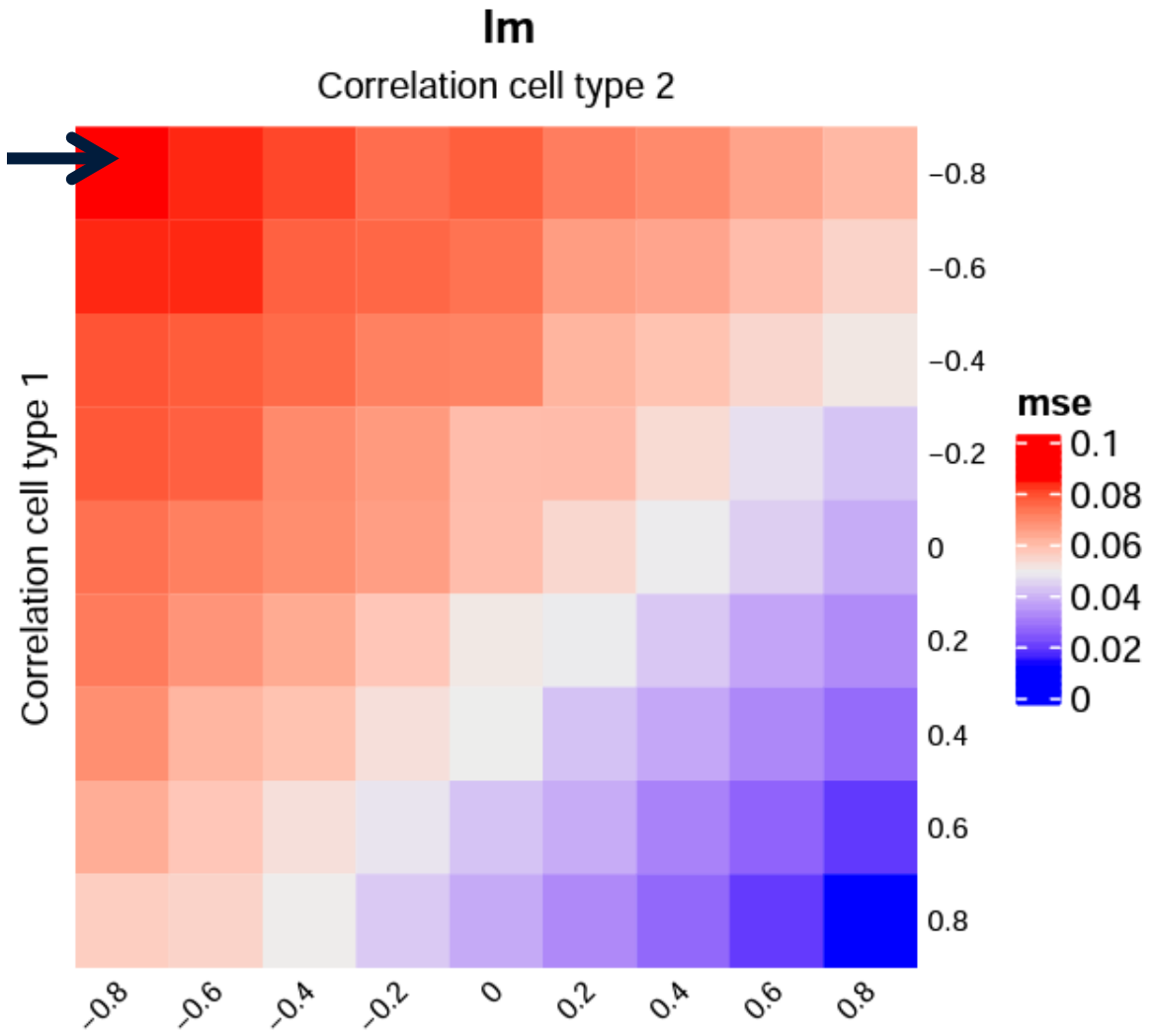
$(\sigma_{121} = 0, \sigma_{122} = 0)$

Simulation results with two genes

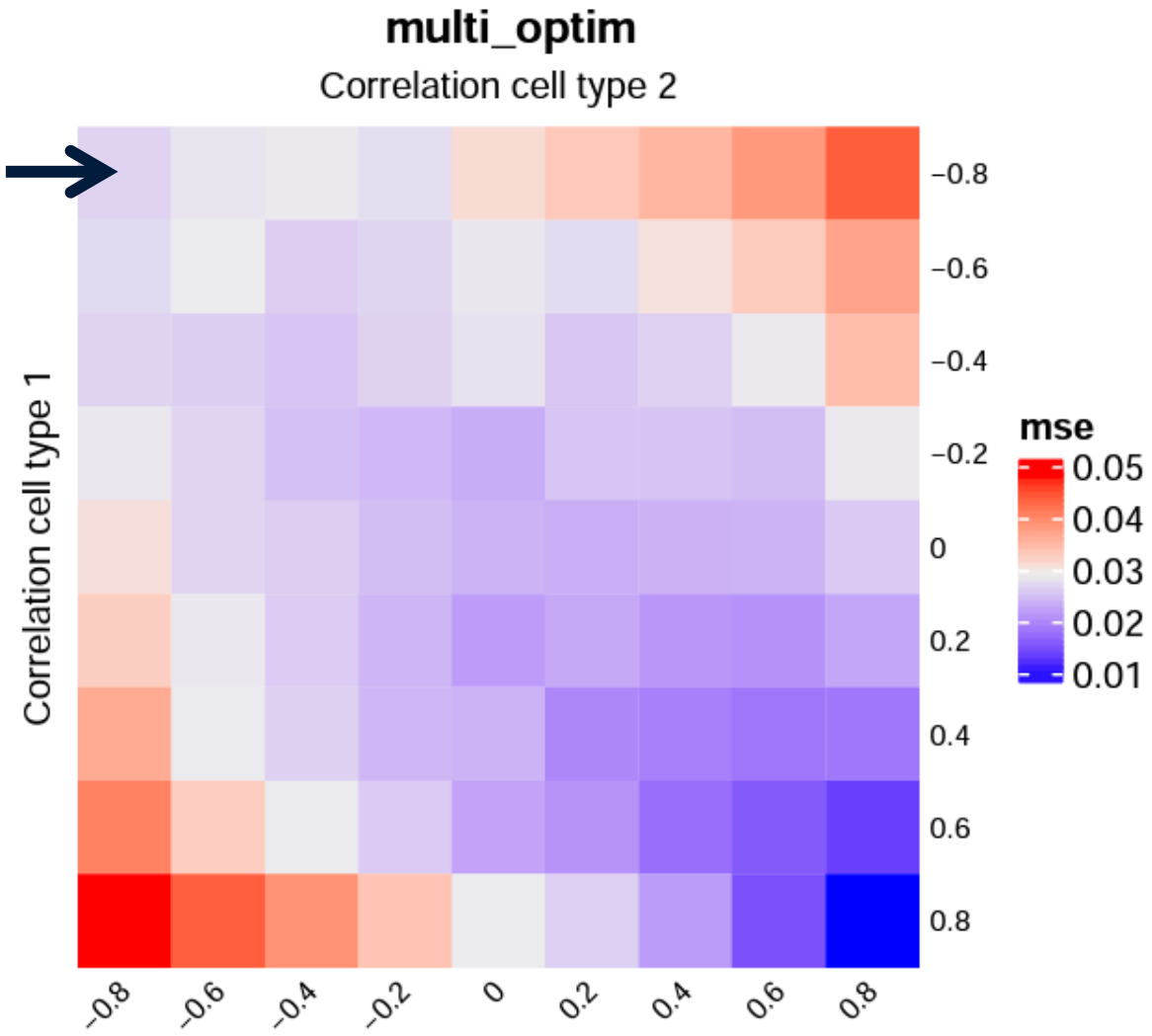


Complex Heatmap (*Gu, 2022*) of the MSE score over LLS estimated parameters, with balanced and high overlapping components.

Simulation results with two genes



Same Heatmap representation as in the previous slide



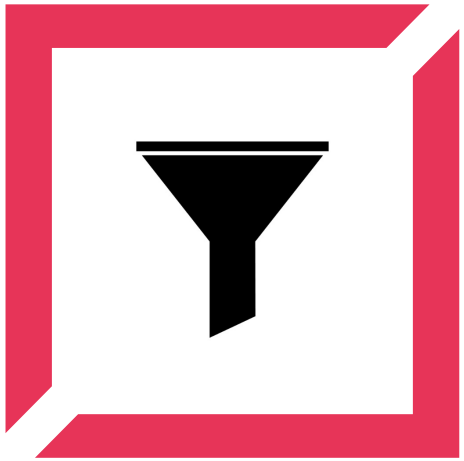
Heatmap of the MSE of the estimated ratios, but using this time the covariance information

Conclusion

Data collection

Poorly described cell populations, full exploitation of Encode and Blueprint datasets

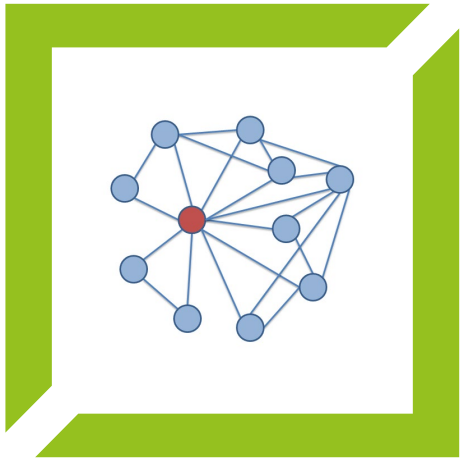
Automatic annotation and description of cellular ontology



Curation

Refine selection of relevant genes:

- Automated method for discarding background noise
- Innovative feature-selection algorithms, using both the differential expression and the covariance structure



Connectivity

Algorithm closer to biological models, accounting for the co-transcriptomic expression between the genes of the purified cell populations

Ongoing work

Statistics

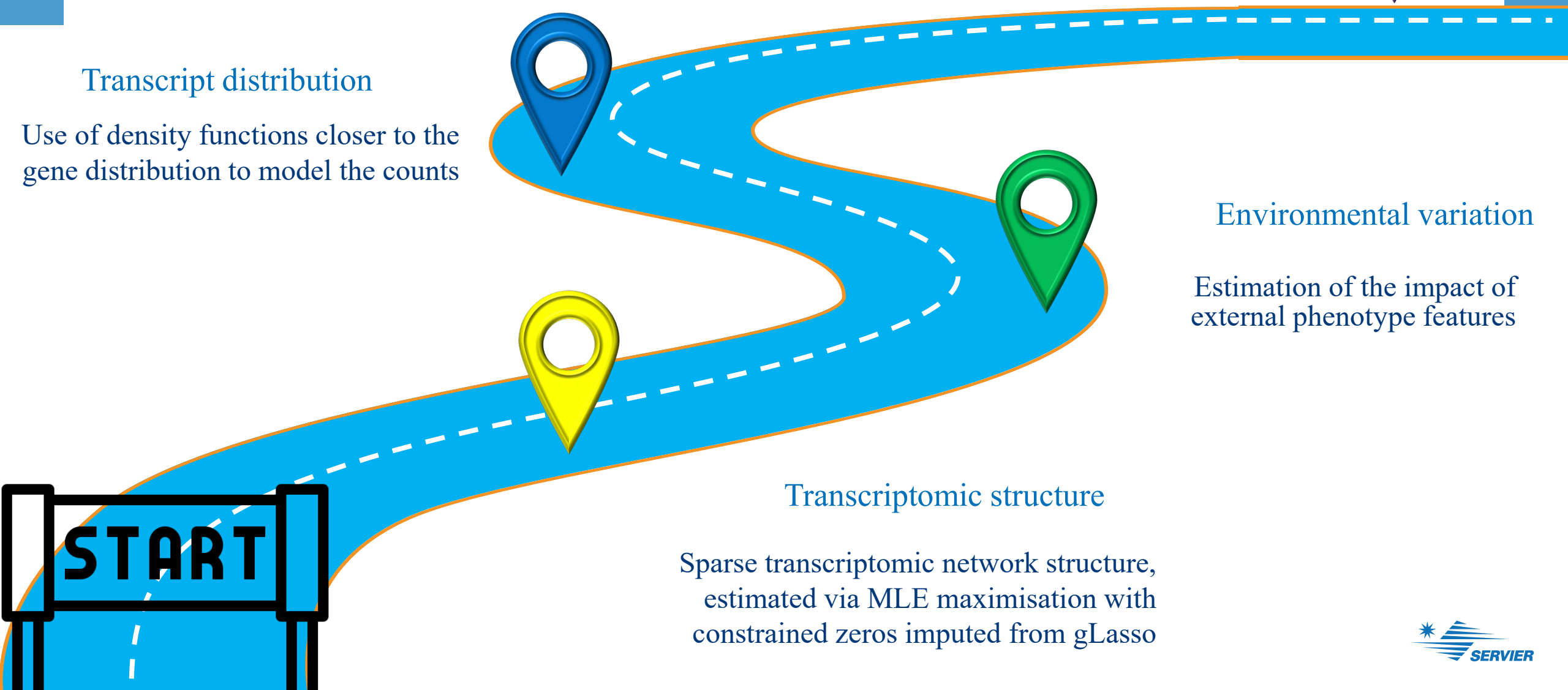
Statistical relevance of the estimates, possibly by means of a Bayesian framework.

Transcript distribution

Use of density functions closer to the gene distribution to model the counts

Environmental variation

Estimation of the impact of external phenotype features



Transcriptomic structure

Sparse transcriptomic network structure, estimated via MLE maximisation with constrained zeros imputed from gLasso

SERVIER

Acknowledgement

Thanks for your attention,

A special thought to my tutors from Sorbonne University (LPSM, LIP6) for the theoretical background and to Servier for supplying internal data and automated pipeline for the analysis of transcriptomic data.

SERVIER

References

[1] B. Panwar *et al.*, "Multi-cell type gene coexpression network analysis reveals coordinated interferon response and cross-cell type correlations in systemic lupus erythematosus," *Genome Res*, vol. 31, no. 4, pp. 659–676, Apr. 2021, doi: 10.1101/gr.265249.120.

[2] P. Lu, A. Nakorchevskiy, and E. M. Marcotte, "Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations," *PNAS*, vol. 100, no. 18, pp. 10370–10375, Sep. 2003, doi: 10.1073/pnas.1832361100.

[3] G. Quon and Q. Morris, "ISOLATE: A computational strategy for identifying the primary origin of cancers using high-throughput sequencing," *Bioinformatics*, vol. 25, no. 21, pp. 2882–2889, Nov. 2009, doi: 10.1093/bioinformatics/btp378.

[4] F. Finotello and Z. Trajanoski, "Quantifying tumor-infiltrating immune cells from transcriptomics data," *Cancer Immunol Immunother*, vol. 67, no. 7, pp. 1031–1040, Jul. 2018, doi: 10.1007/s00262-018-2150-z.

[5] F. Petitprez, C.-M. Sun, L. Lacroix, C. Sautès-Fridman, A. de Reyniès, and W. H. Fridman, "Quantitative Analyses of the Tumor Microenvironment Composition and Orientation in the Era of Precision Medicine," *Front Oncol*, vol. 8, p. 390, 2018, doi: 10.3389/fonc.2018.00390.

[6] S. S. Shen-Orr *et al.*, "Cell type–specific gene expression differences in complex tissues," *Nat Methods*, vol. 7, no. 4, 4, pp. 287–289, Apr. 2010, doi: 10.1038/nmeth.1439.

[7] J. E. Shoemaker, T. J. Lopes, S. Ghosh, Y. Matsuoka, Y. Kawaoka, and H. Kitano, "CTen: A web-based platform for identifying enriched cell types from heterogeneous microarray data," *BMC Genomics*, vol. 13, no. 1, p. 460, Sep. 2012, doi: 10.1186/1471-2164-13-460.

References

[8] S. S. Shen-Orr and R. Gaujoux, "Computational deconvolution: Extracting cell type-specific information from heterogeneous samples," *Curr Opin Immunol*, vol. 25, no. 5, pp. 571–578, Oct. 2013, doi: 10.1016/j.coi.2013.09.015.

[9] C. Fa, A.-H. J, P. J, M. P, and D. P. K, "Comprehensive benchmarking of computational deconvolution of transcriptomics data," Jan. 2020, doi: 10.1101/2020.01.10.897116.

[10] V. C. at channing.harvard.edu>, *ontoProc: Processing of ontologies of anatomy, cell lines, and so on*. Bioconductor version: Release (3.15), 2022. doi: 10.18129/B9.bioc.ontoProc.

[11] T. Hart, H. K. Komori, S. LaMere, K. Podshivalova, and D. R. Salomon, "Finding the active genes in deep RNA-seq gene expression studies," *BMC Genomics*, vol. 14, no. 1, p. 778, Nov. 2013, doi: 10.1186/1471-2164-14-778.

[12] A. Newman *et al.*, "Robust enumeration of cell subsets from tissue expression profiles," *Nature methods*, vol. 12, Mar. 2015, doi: 10.1038/nmeth.3337.

[13] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008, doi: 10.1093/biostatistics/kxm045.

[14] Y. Zuo *et al.*, "INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery," *Methods*, vol. 111, pp. 12–20, Dec. 2016, doi: 10.1016/j.ymeth.2016.08.015.

[15] Z. Gu, *ComplexHeatmap: Make Complex Heatmaps*. Bioconductor version: Release (3.15), 2022. doi: 10.18129/B9.bioc.ComplexHeatmap.

References

[16] J. M. Fernández *et al.*, "The BLUEPRINT Data Analysis Portal," *Cell Syst*, vol. 3, no. 5, pp. 491–495.e5, Nov. 2016, doi: 10.1016/j.cels.2016.10.021.

[17] S. Chevrier *et al.*, "An Immune Atlas of Clear Cell Renal Cell Carcinoma," *Cell*, vol. 169, no. 4, pp. 736–749.e18, May 2017, doi: 10.1016/j.cell.2017.04.016.

[18] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.

[19] G. Monaco *et al.*, "RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types," *Cell Reports*, vol. 26, no. 6, pp. 1627–1640.e7, Feb. 2019, doi: 10.1016/j.celrep.2019.01.041.

[20] K. Kim *et al.*, "Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis," *Brain*, vol. 144, no. 2, pp. 450–461, Mar. 2021, doi: 10.1093/brain/awaa421.

[21] P. S. Linsley, C. Speake, E. Whalen, and D. Chaussabel, "Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis," *PLoS One*, vol. 9, no. 10, p. e109760, 2014, doi: 10.1371/journal.pone.0109760.