Comprehensive benchmarking of computational deconvolution of transcriptomics data

Francisco Avila Cobos

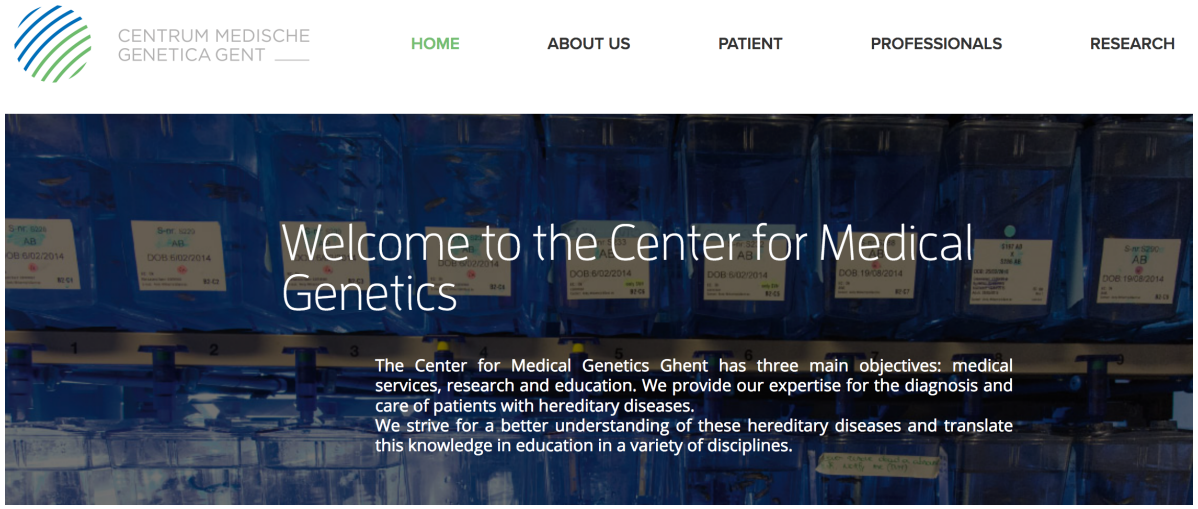
Health Data Challenge (2nd edition)

25 – 29 November, 2019

@Frank_txu

francisco.avilacobos@ugent.be

Who are we?



www.cmgg.be/en/

CANCER RESEARCH INSTITUTE GHENT

CRIG

www.crig.ugent.be

NEUROBLASTOMA **BIOINFORMATICS** **THE NON-CODING TRANSCRIPTOME**

DATA-MINING **PAN-CANCER**

Who are we?

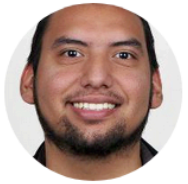
Garvan Institute of Medical Research

Single Cell and Computational Genomics

Staff

Dr Brian Gloss
Senior Research Officer

Rachael Zekanovic
Research Assistant



Jose Alquicira Hernandez
PhD Student

Dr Venessa Chin
Research Officer

Lab Head

A/Prof Joseph Powell

Francisco Cobos
Visiting Scholar

Vikkitharan Gnanasambandapillai
Research Assistant

Angela Murphy
Research Assistant (Molecular)

Dr Drew Neavin
Research Officer

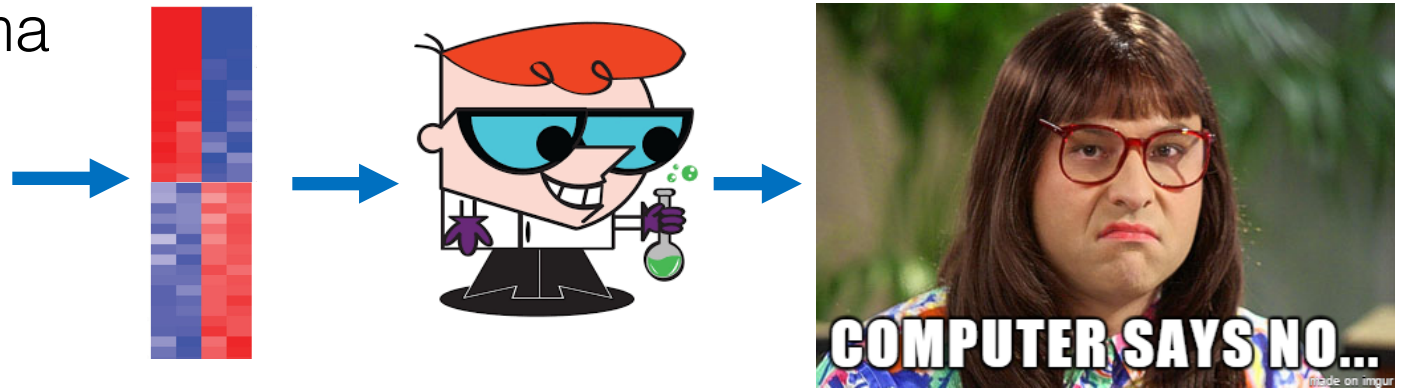
Anne Senabouth
Bioinformatician

Seyhan Yazar
Senior Research Officer

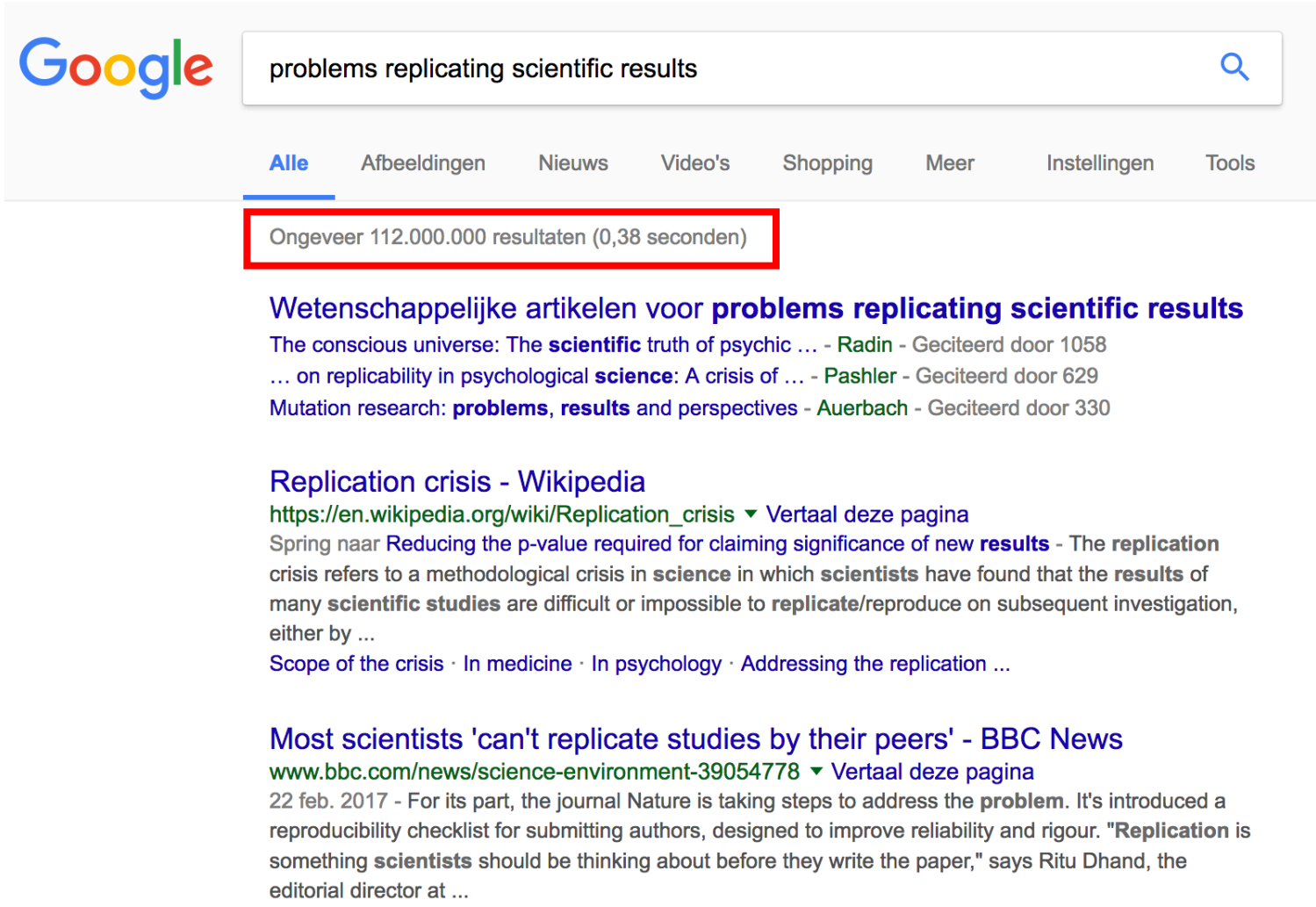
www.garvan.org.au

Even though scientific research should guarantee reproducibility and replication of any experiment…

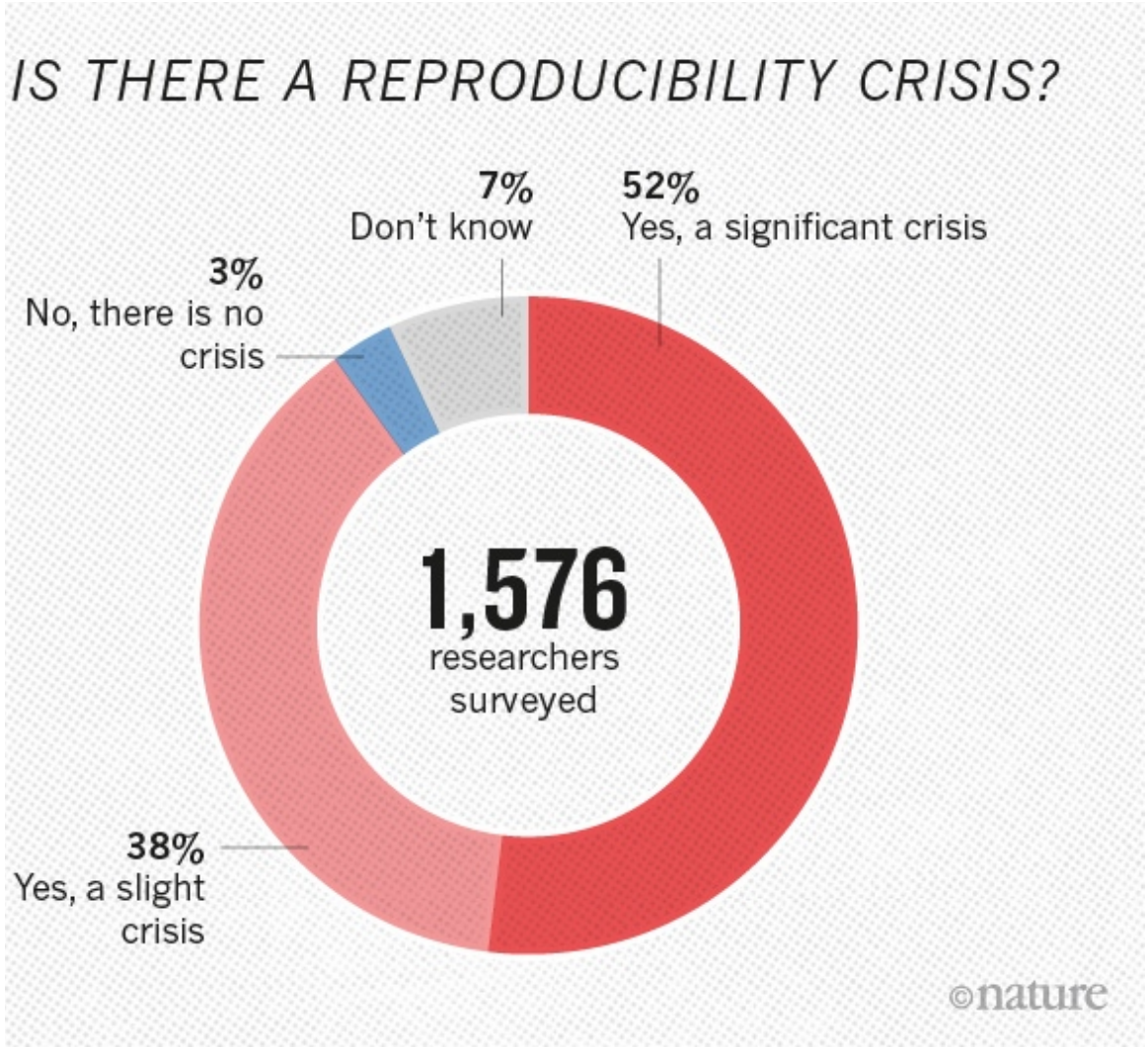
- Neuroblastoma
- ALL, AML, …
- Lung cancer
- Cancer ‘X’

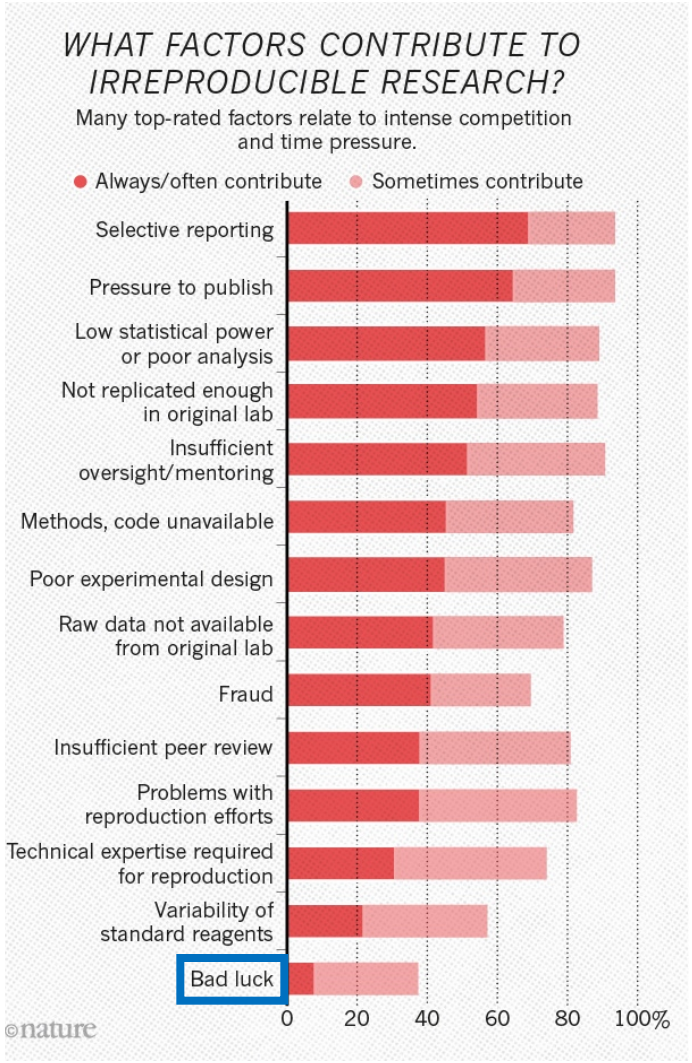


There is a clear “reproducibility crisis” in scientific research



There is a clear “reproducibility crisis” in scientific research



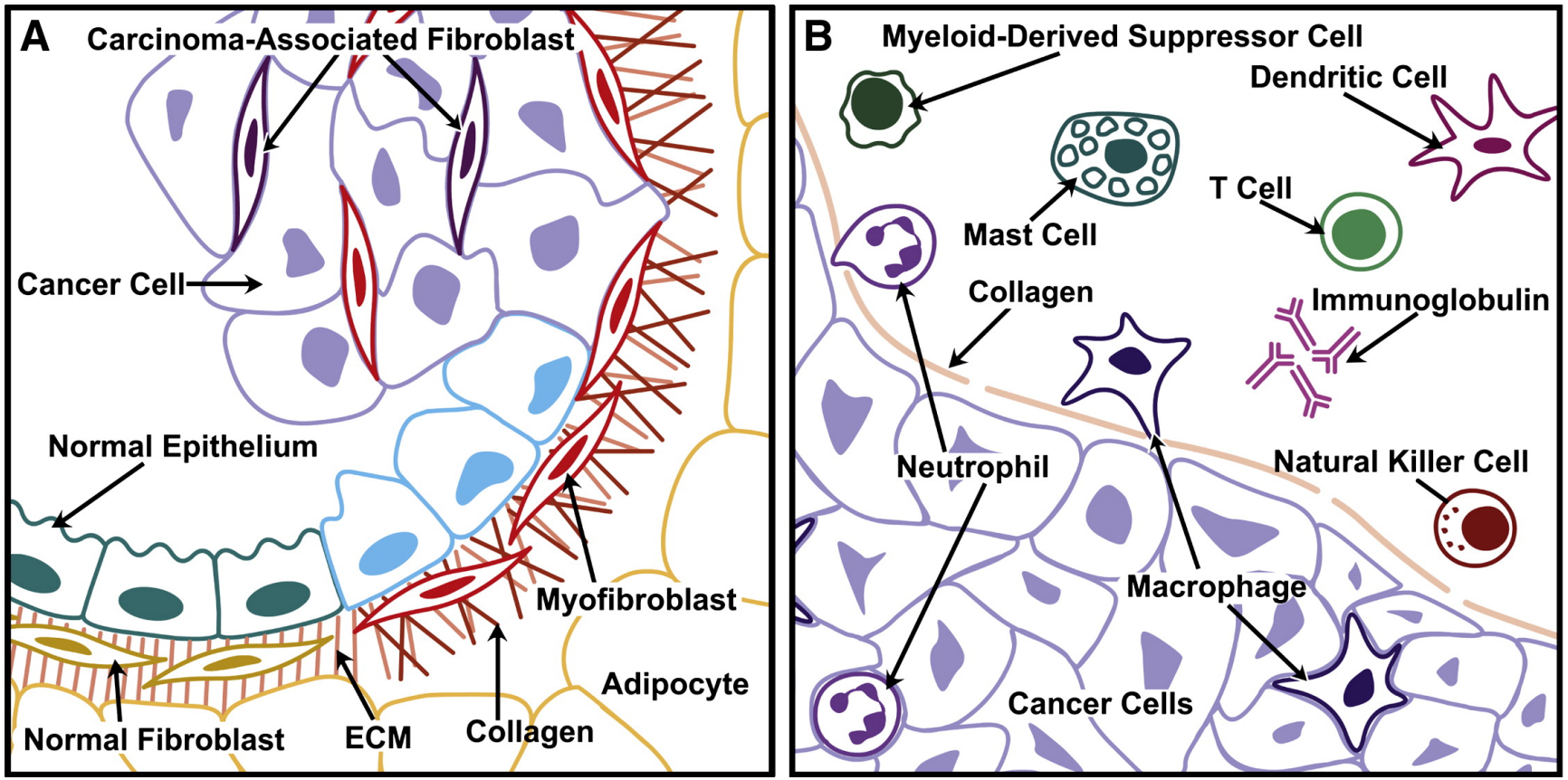


1,500 scientists lift the lid on reproducibility. Baker,M. (2016) *Nat. News*, 533,452.

There are several important factors responsible for this crisis

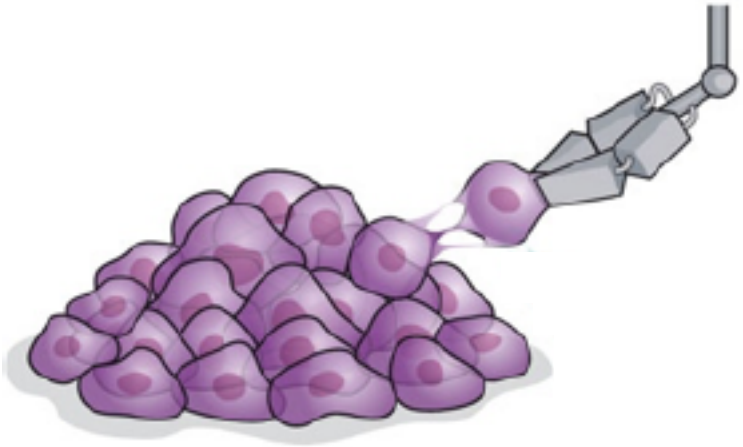
- **Sample heterogeneity**
- Insufficiently documented or incorrect data processing practices (MAQC Consortium, 2010).
- Platform-specific differences (SEQC/MAQC-III Consortium, 2014).

Tumor samples also contain a variable portion of non-malignant cells that include epithelial, stromal and infiltrating immune cells



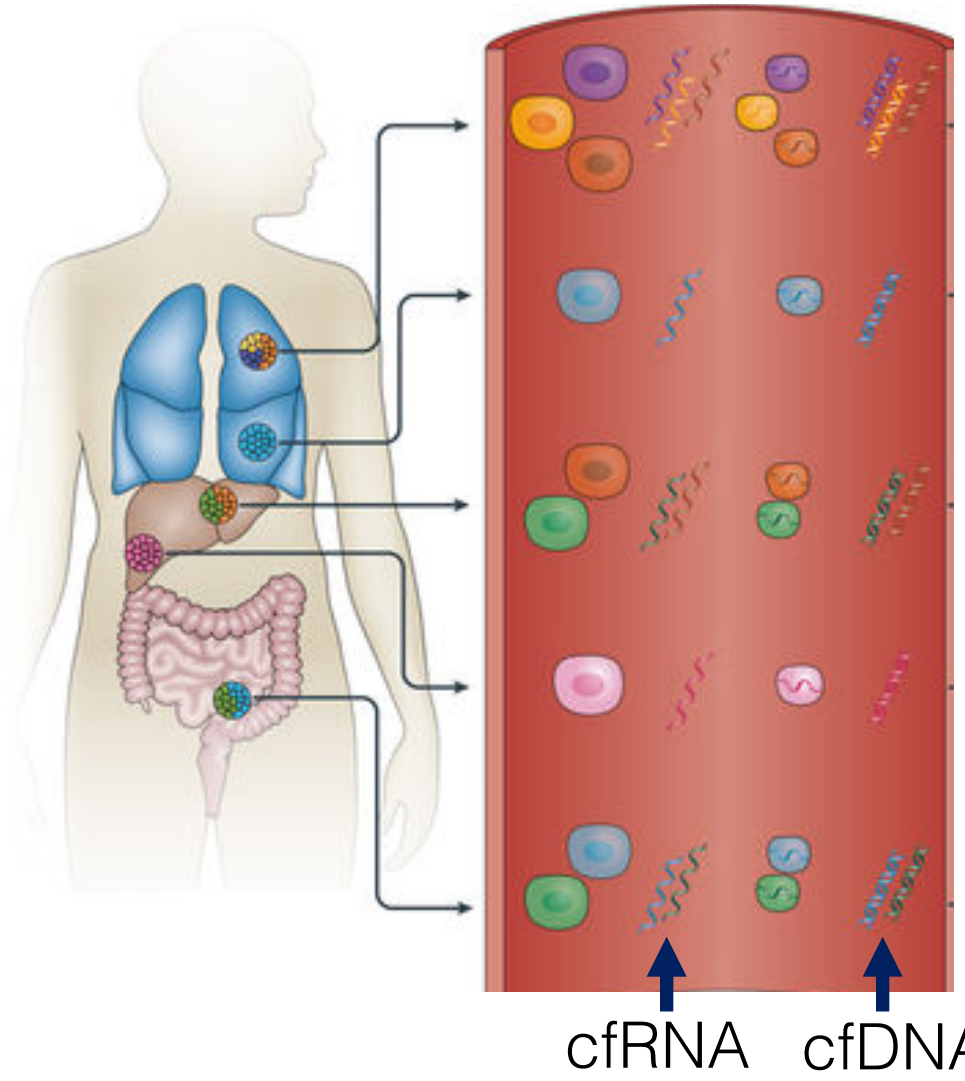
Tumors as organs: complex tissues that interface with the entire organism.
Egeblad et al., 2010. *Dev Cell*. 18(6):884-901.

Single-cell technologies allow the analysis of individual cells within heterogeneous tissues…



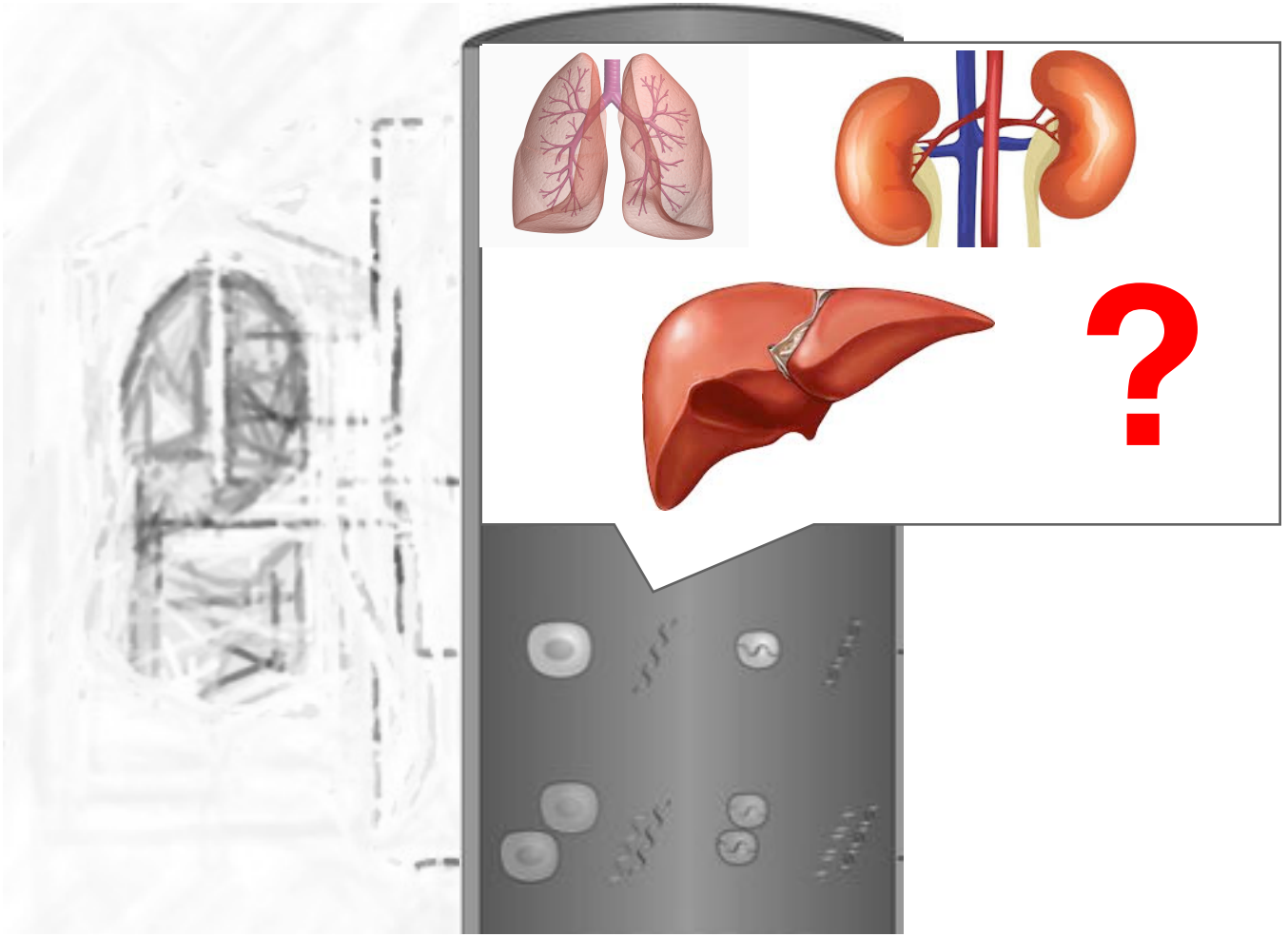
…but have labour-intensive protocols and require expensive resources, hindering its establishment in the clinic

Single-cell technology is not applicable to cell-free scenarios



Integrating liquid biopsies into the management of cancer. Siravegna et al., 2017. Nature Reviews Clinical Oncology

Single-cell technology is not applicable to cell-free scenarios



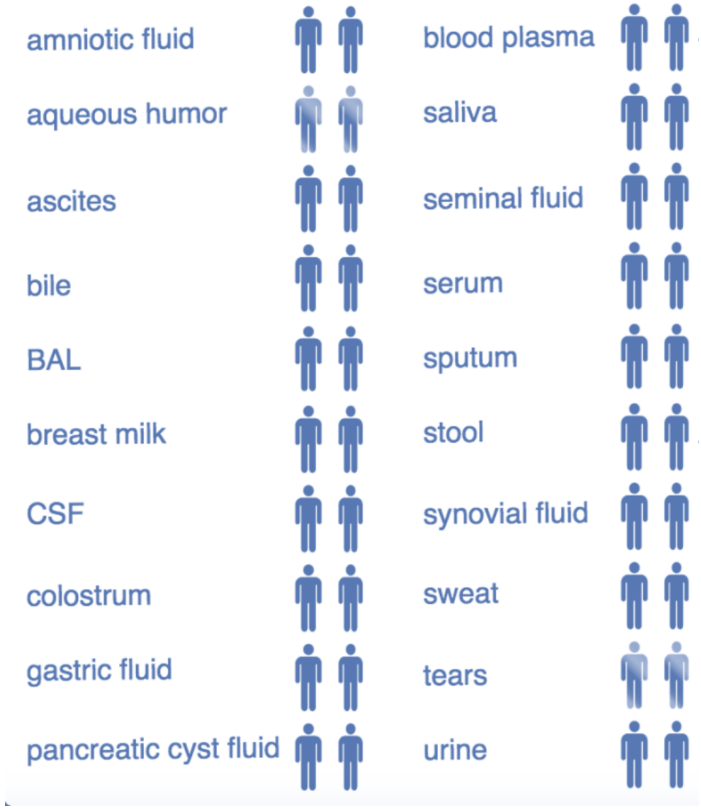
Integrating liquid biopsies into the management of cancer. Siravegna et al., 2017. Nature Reviews Clinical Oncology

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Eva Hulstaert

Charting extracellular transcriptomes in The Human Biofluid RNA Atlas

Eva Hulstaert, Annelien Morlion, Francisco Avila Cobos, Kimberly Verniers, Justine Nuytens, Eveline Vanden Eynde, Nurten Yigit, Jasper Anckaert, Anja Geerts, Pieter Hindryckx, Peggy Jacques, Guy Brusselle, Ken R. Bracke, Tania Maes, Thomas Malfait, Thierry Derveaux, Virginie Ninclaus, Caroline Van Cauwenbergh, Kristien Roelens, Ellen Roets, Dimitri Hemelsoet, Kelly Tilleman, Lieve Brochez, Scott Kuersten, Lukas Simon, Sebastian Karg, Alexandra Kautzky-Willers, Michael Leutner, Christa Nöhammer, Ondrej Slaby, Gary P. Schroth, Jo Vandesompele, Pieter Mestdagh

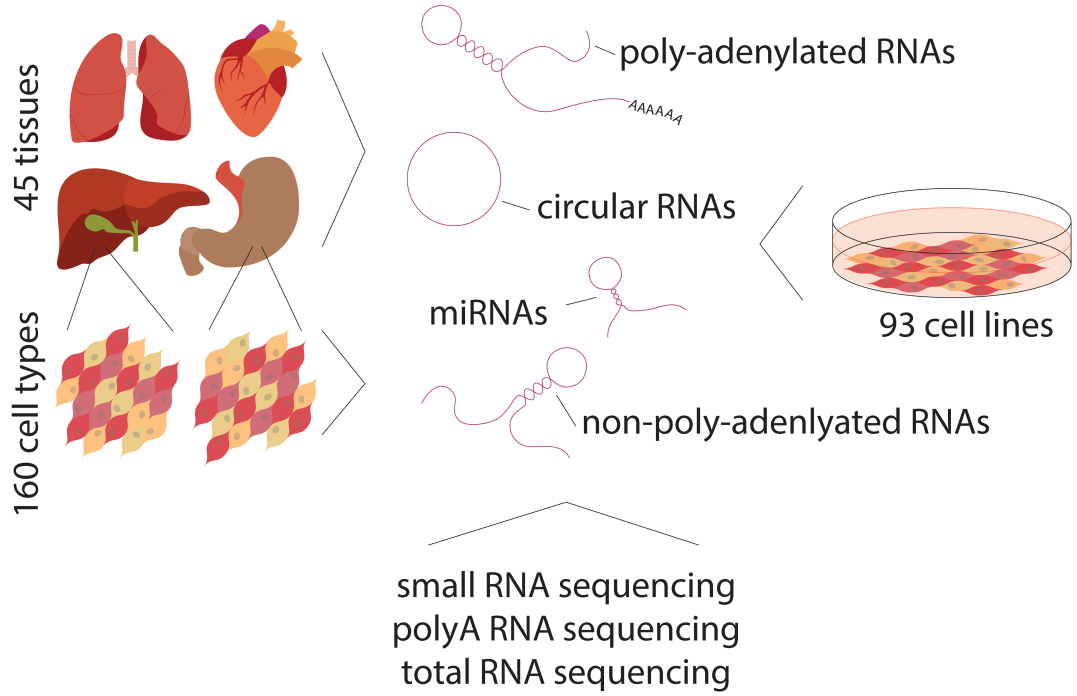


bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Lucía Lorenzi

The RNA Atlas, a single nucleotide resolution map of the human transcriptome

Lucia Lorenzi, Hua-Sheng Chiu, Francisco Avila Cobos, Stephen Gross, Pieter-Jan Volders, Robrecht Cannoodt, Justine Nuytens, Katrien Vanderheyden, Jasper Anckaert, Steve Lefever, Tine Goovaerts, Thomas Birkballe Hansen, Scott Kuersten, Nele Nijs, Tom Taghon, Karim Vermaelen, Ken R. Bracke, Yvan Saeys, Tim De Meyer, Nandan Deshpande, Govardhan Anande, Ting-Wen Chen, Marc R. Wilkins, Ashwin Unnikrishnan, Katleen De Preter, Jørgen Kjems, Jan Koster, Gary P. Schroth, Jo Vandesompele, Pavel Sumazin, Pieter Mestdagh



Computational deconvolution is the solution

DECONVOLUTION

inference of

cell type proportions

AND/OR

cell type-specific expression profiles

in heterogeneous samples

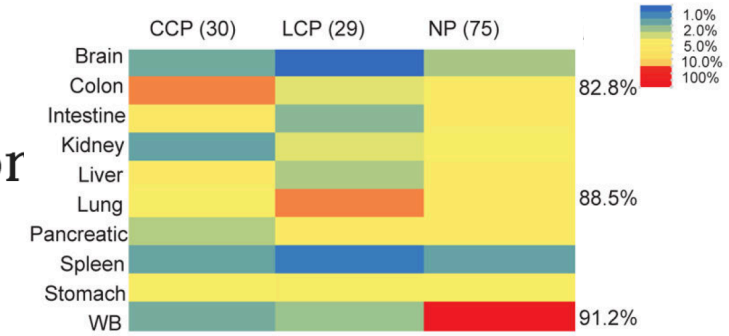
Deconvolution applied to cell-free scenarios has been mainly focused on DNA

Genome Biology

CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA

nature genetics

Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA



Although deconvolution of cfRNA also exists

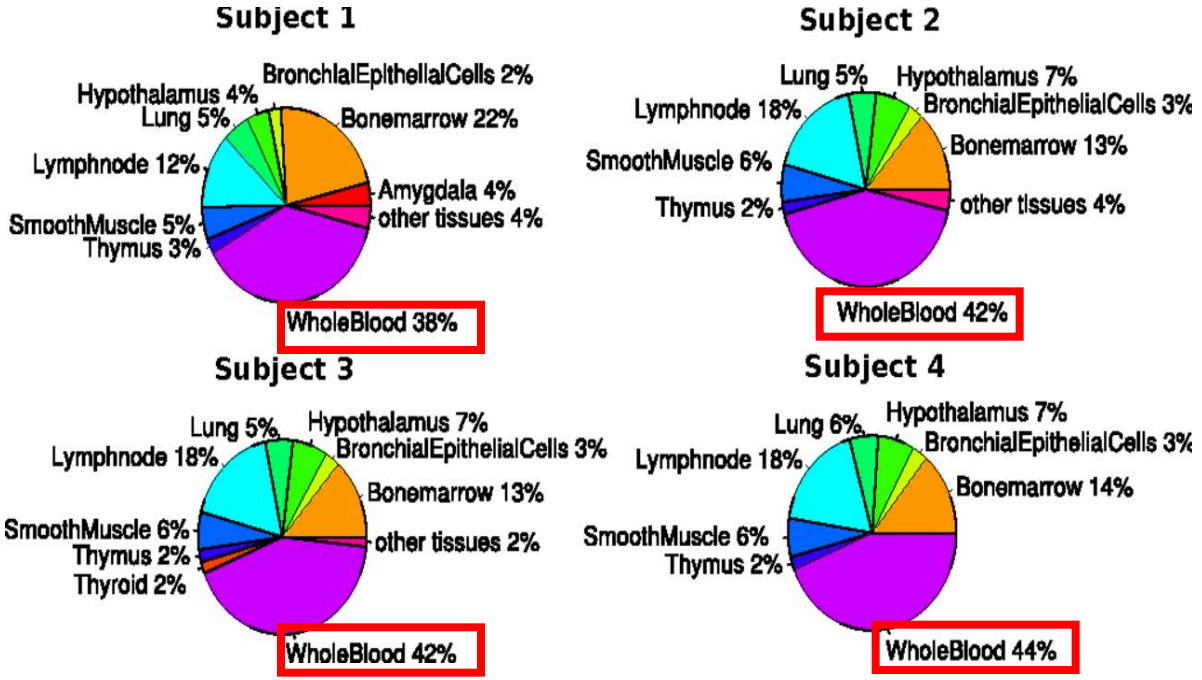
PNAS Proceedings of the National Academy of Sciences of the United States of America

Noninvasive in vivo monitoring of tissue-specific global gene expression in humans

Winston Koh, Wenying Pan, Charles Gawad, H. Christina Fan, Geoffrey A. Kerchner, Tony Wyss-Coray, Yair J. Blumenfeld, Yasser Y. El-Sayed, and Stephen R. Quake

Deconvolution of Cell-Free RNA Transcriptome Using Microarray.

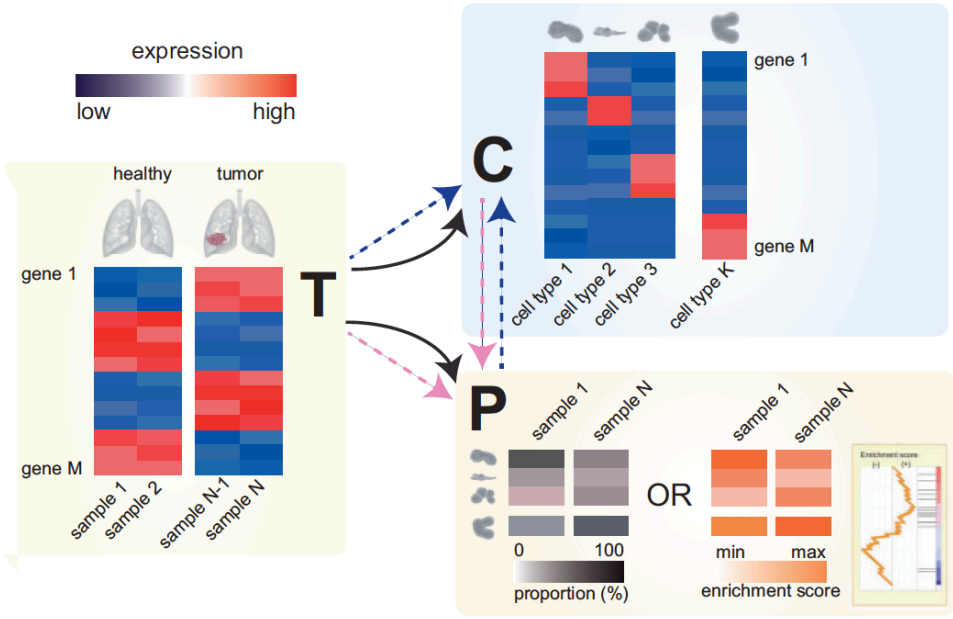
Deconvolution of a cell-free transcriptome is used to determine the relative contribution of each tissue type toward the cell-free RNA transcriptome.



Bioinformatics

Computational deconvolution of transcriptomics data from mixed cell populations

Francisco Avila Cobos[1,2,3], Jo Vandesompele[1,2,3], Pieter Mestdagh[1,2,3,†] and Katleen De Preter[1,2,3,*,†]

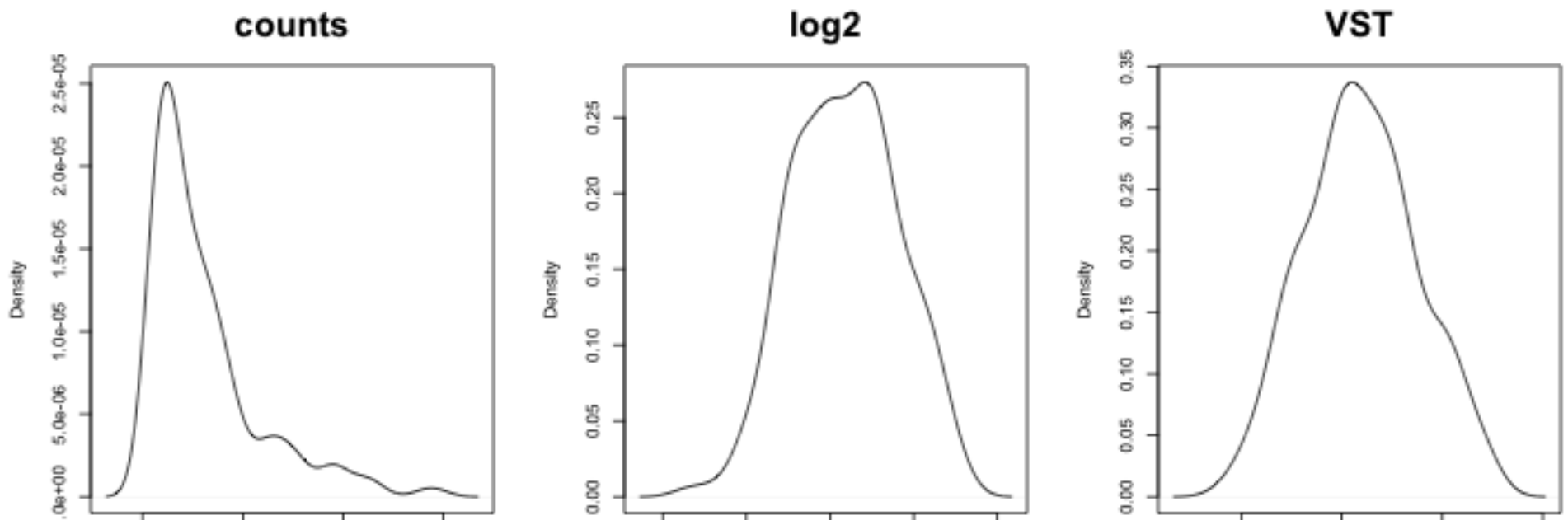


- Mathematical approaches
- Factors affecting the deconvolution efficiency:
  - Pre-processing
  - Logarithmic versus linear space

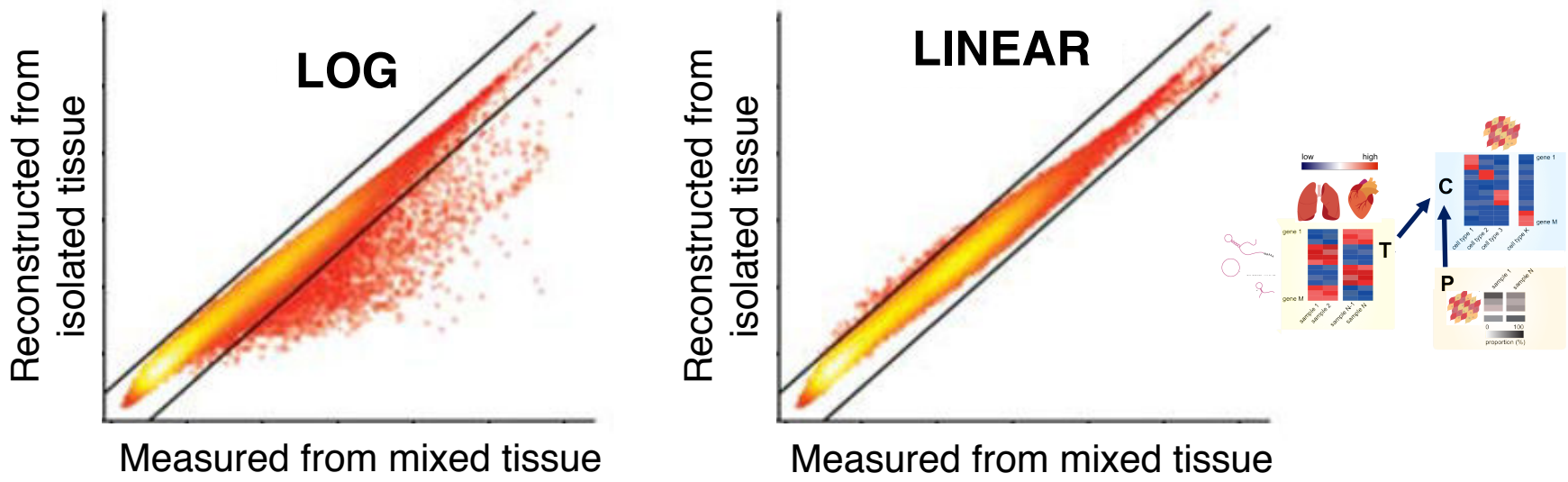
…

Although log transformation is routinely included as part of the pre-processing of omics data…

Expression data often transformed into logarithmic scale because the **statistical tests used for differential gene expression assume an underlying normal distribution**.



…deconvolution requires data in linear scale as opposed to log transformations for DGEA



- The reconstructed signal is an under-estimation of the signal measured from the mixture.
- If the data was transformed back to linear scale → accurate deconvolution

Gene expression deconvolution in linear space. Zhong,Y. and Liu,Z. (2012) Nat.Methods, 9, 8–9.

…deconvolution requires data in linear scale as opposed to log transformations for DGEA

Deconvolution is modeled by a linear equation $\mathbf{O} = \mathbf{S} \times \mathbf{W}$, where $\mathbf{O}$ is the expression data for mixed tissue samples, $\mathbf{S}$ is the tissue-specific expression profile, and $\mathbf{W}$ is the cell-type frequency matrix. If the signal is log-transformed, the linearity will no longer be preserved. The concavity feature of the log function will induce a downward bias to the reconstructed signal (Fig. 1a and Supplementary Fig. 1). Mathematically, it can be shown that the deconvolution model used on log-transformed signals is $\log(\mathbf{O}') = \log(\mathbf{S}) \times \mathbf{W}$, where $\mathbf{O}'$ is the csSAM estimate of gene-expression profiles. As $\mathbf{W}$ is a frequency matrix and its column values sum to 1, the following is true by the properties of concave functions[3]: $\log(\mathbf{S} \times \mathbf{W}) > \log(\mathbf{S}) \times \mathbf{W}$. Taking these two equations together, we can conclude that $\log(\mathbf{O}') < \log(\mathbf{S} \times \mathbf{W}) = \log(\mathbf{O})$. Thus, we proved that when log-transformed signal is used as the input for signal reconstruction, it will always yield an underestimation of the true signal. By taking an anti-log transformation, we obtained an unbiased reconstruction of the mixed tissue samples (Fig. 1b and Supplementary Fig. 2).

Gene expression deconvolution in linear space. Zhong,Y. and Liu,Z. (2012) Nat.Methods, 9, 8–9.

Comprehensive benchmarking of computational deconvolution of transcriptomics data

- What's more important: transformation, pre-processing, method?
- Are they equally important?
- Are there differences in terms of performance?
- **Pre-print will be available @ bioRxiv on December 5, 2019**

Goal: inference of cell type proportions in artificial tissues

DECONVOLUTION

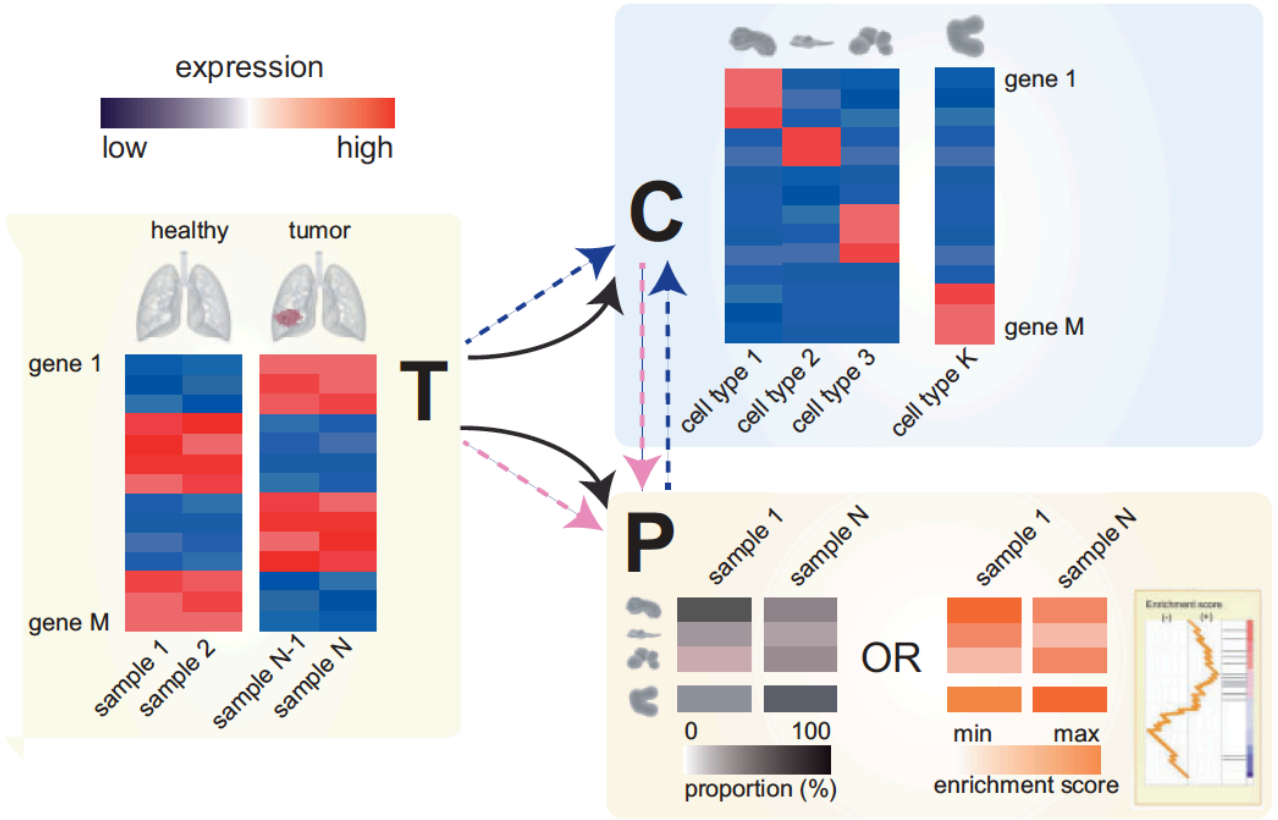
inference of

cell type proportions

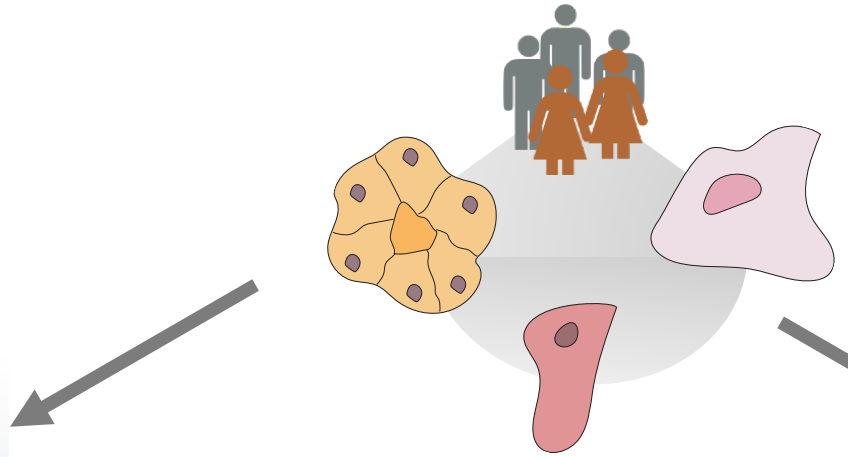
AND/OR

cell type-specific expression profiles

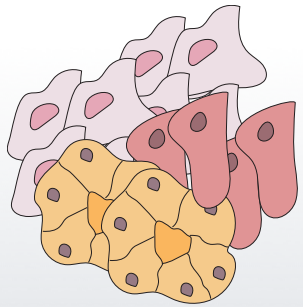
in heterogeneous samples



We take advantage of having individual cells (scRNA-seq)

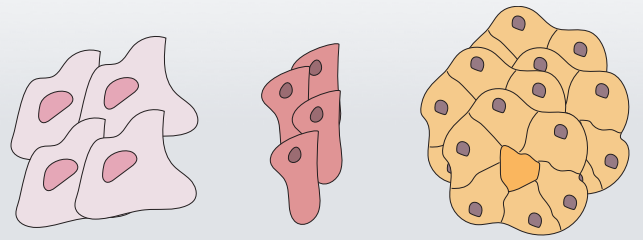


T



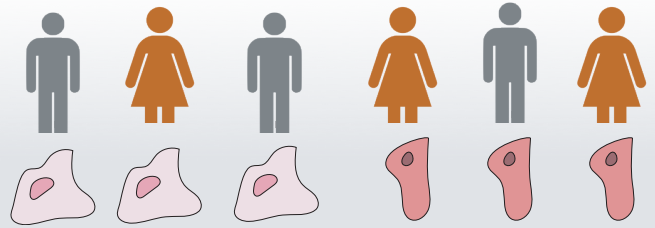
“pseudo-bulk” mixtures
(artificial tissues)

C **scenario 1**



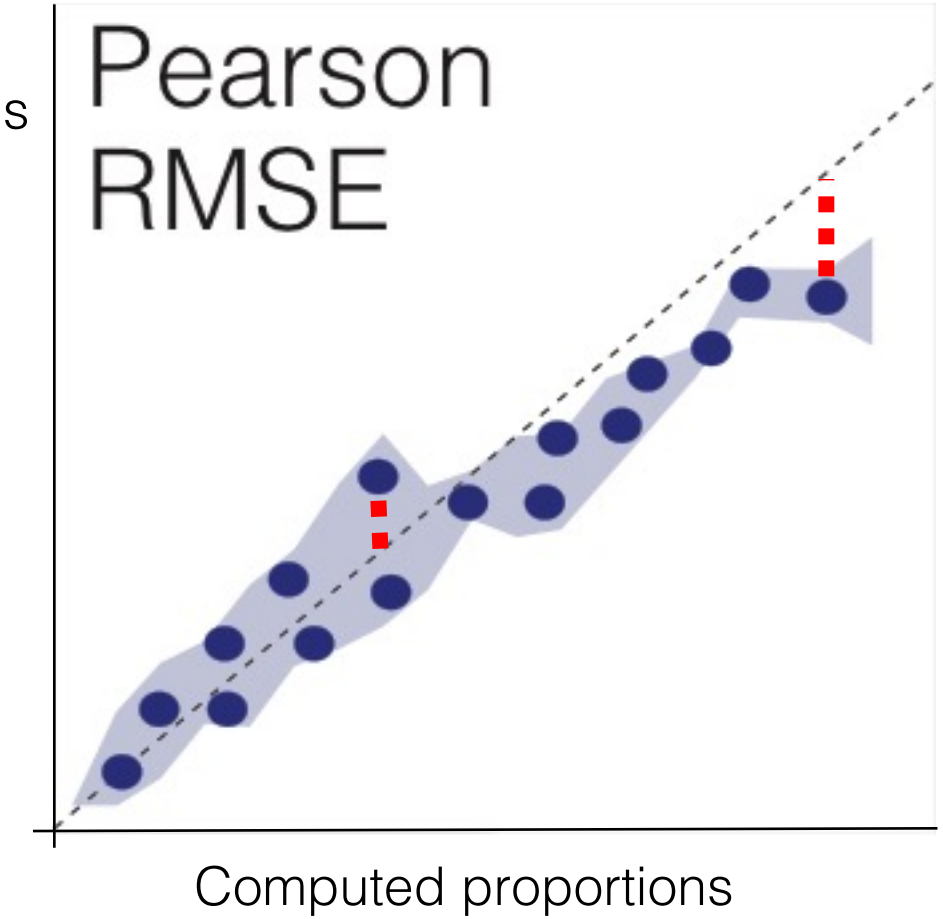
“bulk” cell-type specific profiles

C **scenario 2**



single cells from different individuals

The performance is assessed using pearson correlation and the root mean squared error (RMSE)

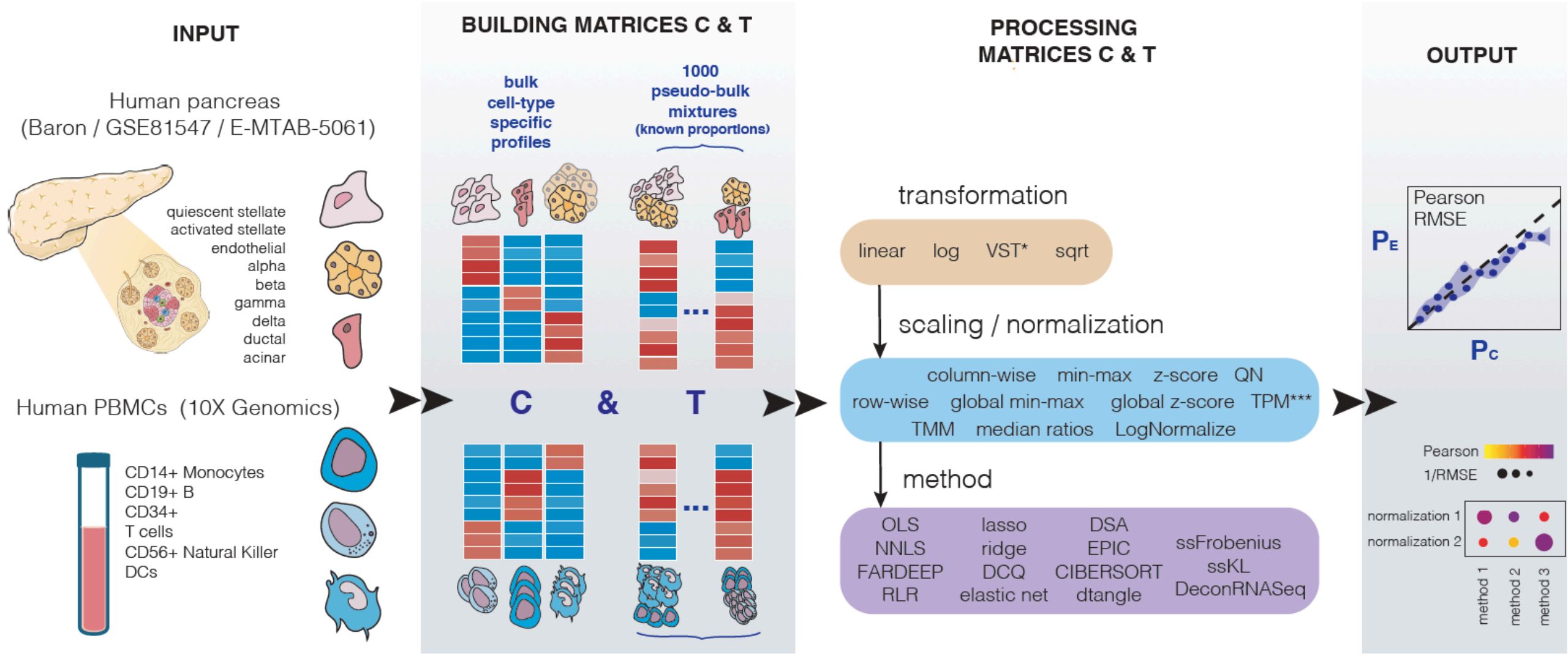


$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

expected proportions = computed proportions

Pearson correlation = 1
RMSE = 0

Scenario 1: Computational deconvolution using “bulk” RNA-seq data

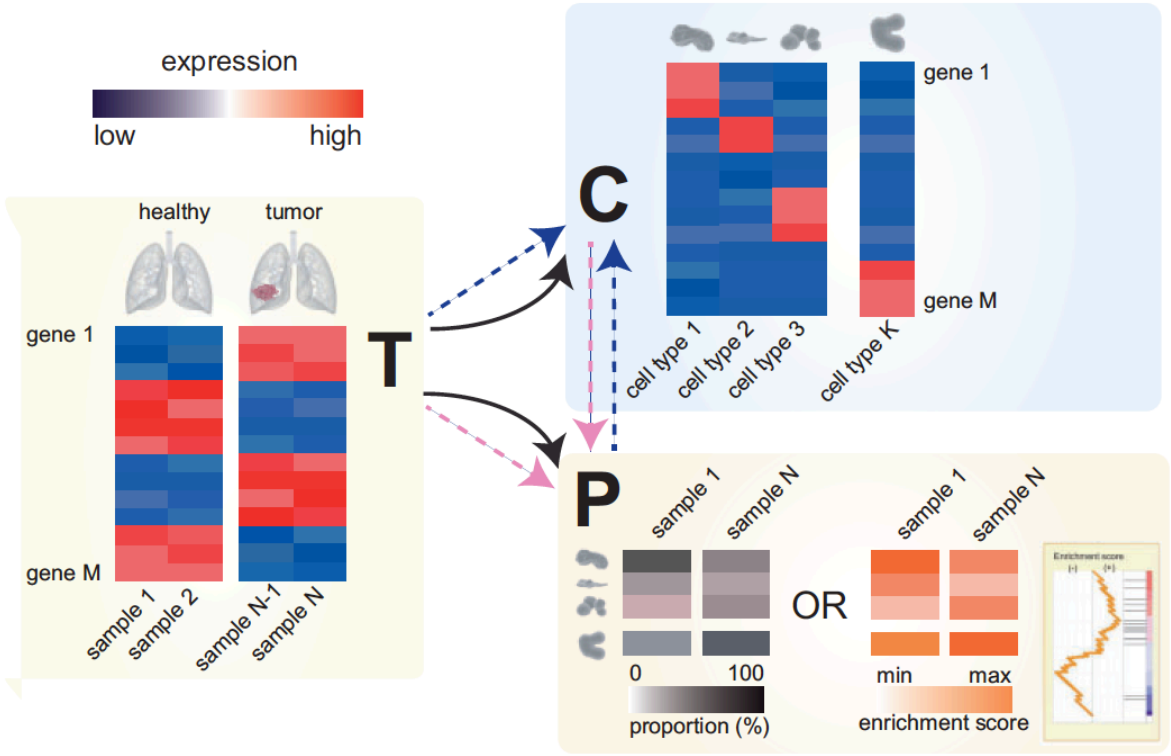


Avila Cobos *et al.*, in preparation

Mathematical approaches to solve the deconvolution problem

BULK

- Supervised:
  - a) Given T and C → **P**
    - OLS, nnls, RLR, FARDEEP, CIBERSORT, **MMAD**, DSA
  - b) Given T and P → C
    - LRCDE, **MMAD**

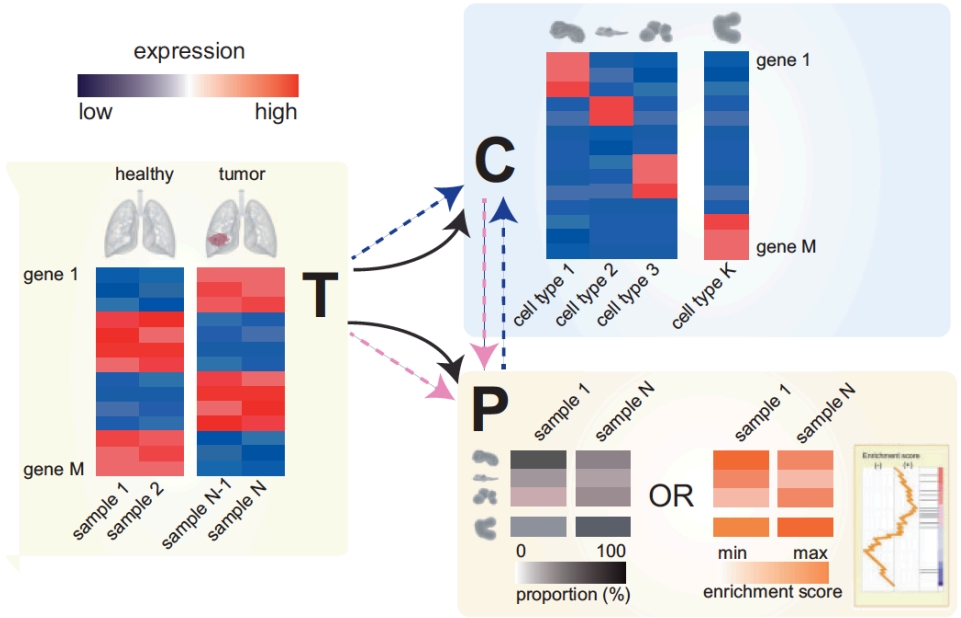


Mathematical approaches to solve the deconvolution problem

BULK

- Supervised:
  - a) Given T and C → **P**
    - OLS, nnls, RLR, FARDEEP, CIBERSORT, **MMAD**
  - b) Given T and P → C
    - LRCDE, **MMAD**
- Semi-supervised: Given T + set of markers → **P**
  - DSA, ssKL, ssFrobenius
  - WISP (NNLS).
- Unsupervised (=Complete deconvolution): Given T → C and **P**
  - **MMAD**, deconf, NMF (Virtual microdissection)
  - deconICA

Mathematical approaches to solve the deconvolution problem



$$T = C \cdot P$$

$$y = X\beta$$

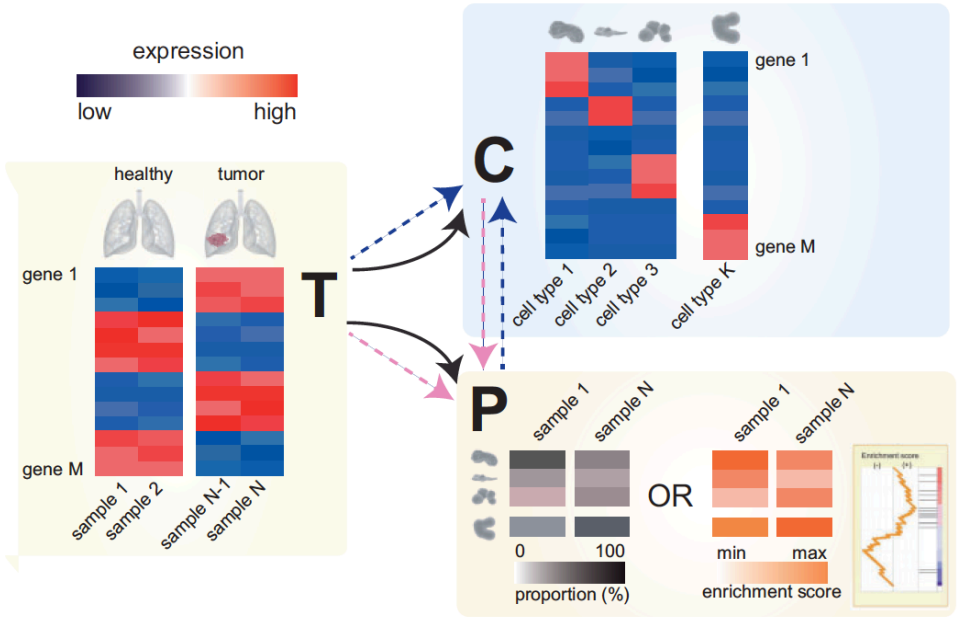
$$\min_{P(\text{or } C)} ||C \cdot P - T||^2$$

OLS: $RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{x}\beta)$

$$\hat{\beta} = (X'X)^{-1}X'y$$

- NNLS (non-negative least squares):
  - OLS + non-negativity + sum-to-one
- RLR/FARDEEP (Robust Linear Regression):
  - Outlier removal before coefficient estimation

Mathematical approaches to solve the deconvolution problem



$$T = C \cdot P$$

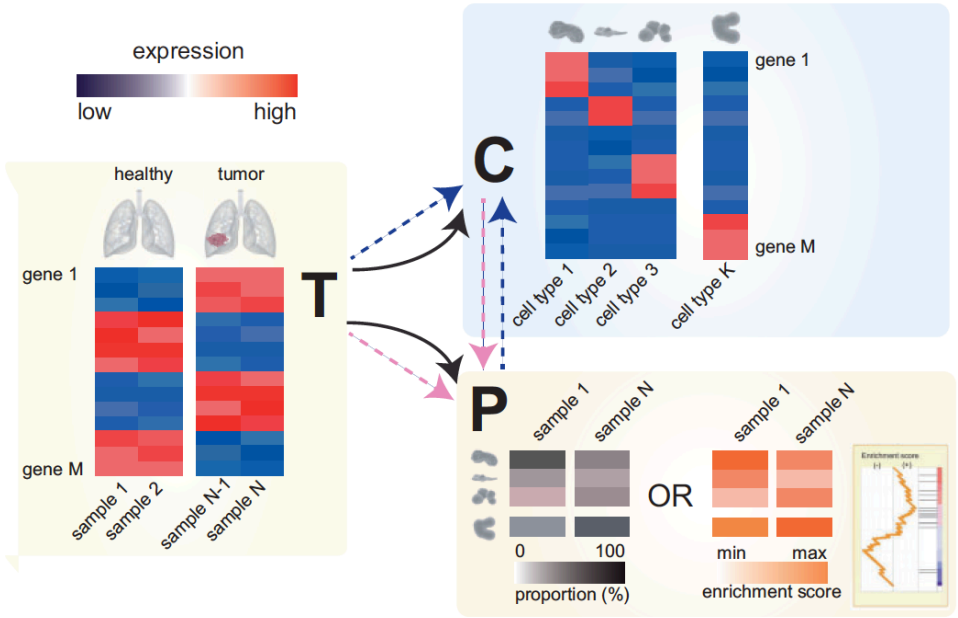
$$y = X\beta$$

$$\min_{P(\text{or } C)} ||C \cdot P - T||^2$$

- NMF
  - Random initializations of P
- ssNMF (ssKL, ssFrobenius)
  - Use marker information
- DSA:

$$\begin{pmatrix} g_{11} & 0 & \dots & 0 \\ g_{21} & 0 & \dots & 0 \\ 0 & g_{32} & \dots & 0 \\ 0 & g_{42} & \dots & 0 \\ 0 & g_{52} & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & g_{mk} \end{pmatrix} \Rightarrow \begin{pmatrix} \bar{g}_1 & 0 & \dots & 0 \\ 0 & \bar{g}_2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \bar{g}_k \end{pmatrix} \longrightarrow \text{OLS}$$

Mathematical approaches to solve the deconvolution problem



$$\mathrm{T} = \mathrm{C} \cdot \mathrm{P}$$

$$y = X\beta$$

$$\min_{\mathrm{P(or\ C)}} ||\mathrm{C} \cdot \mathrm{P} - \mathrm{T}||^2$$

- QP

$$\frac{1}{2}\|Qx - c\|^2 = \frac{1}{2}(Qx - c)^T(Qx - c) = \underbrace{\frac{1}{2}\left(x^T Q^T Q x - 2x^T Q^T c + c^T c\right)}_{\frac{1}{2}x^T Ax + q^T x}$$

- Regularization (lasso, ridge, elastic net)

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

Hastie *et al.* - The Elements of Statistical Learning (book)

Mathematical approaches to solve the deconvolution problem

- **WISP** (Weighted In Silico Pathology; Blum *et al.*, 2019):

  Transcriptomics or **methylation** data

  Two-step approach:
  1) Estimates pure population profiles based on predefined pure samples.
  2) Estimates the proportions of each pure population in a mixed sample through NNLS.

  MPM **signatures** with **three components:** epithelioid-like, sarcomatoid-like and non-tumor.
  a) for tissues (with a complex microenvironment present)
  b) for cell lines

Mathematical approaches to solve the deconvolution problem

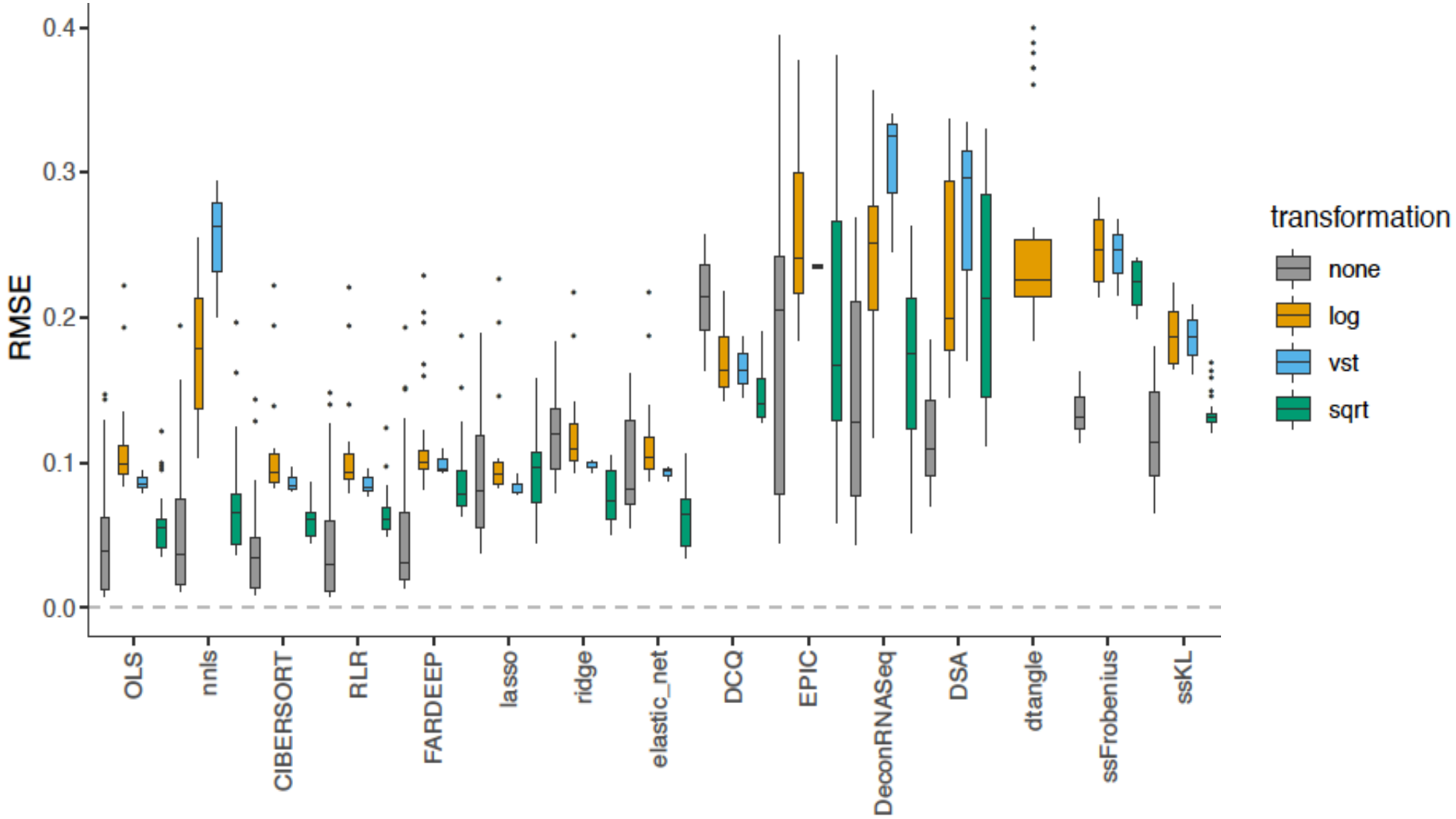
- **deconICA** (Czerwinska *et al.*):
  - Challenge: assigning components to specific biological processes, cell types and technical factors.
    - Cell-types associated with components through highest correlation.

Cluster	Compo-nent	Meaning
Immune	RIC2	B cells
	RIC25	T cells
	RIC27	B cells
	RIC28	response to wounding
	RIC37	IFN signalling pathway
	RIC57	monocytes

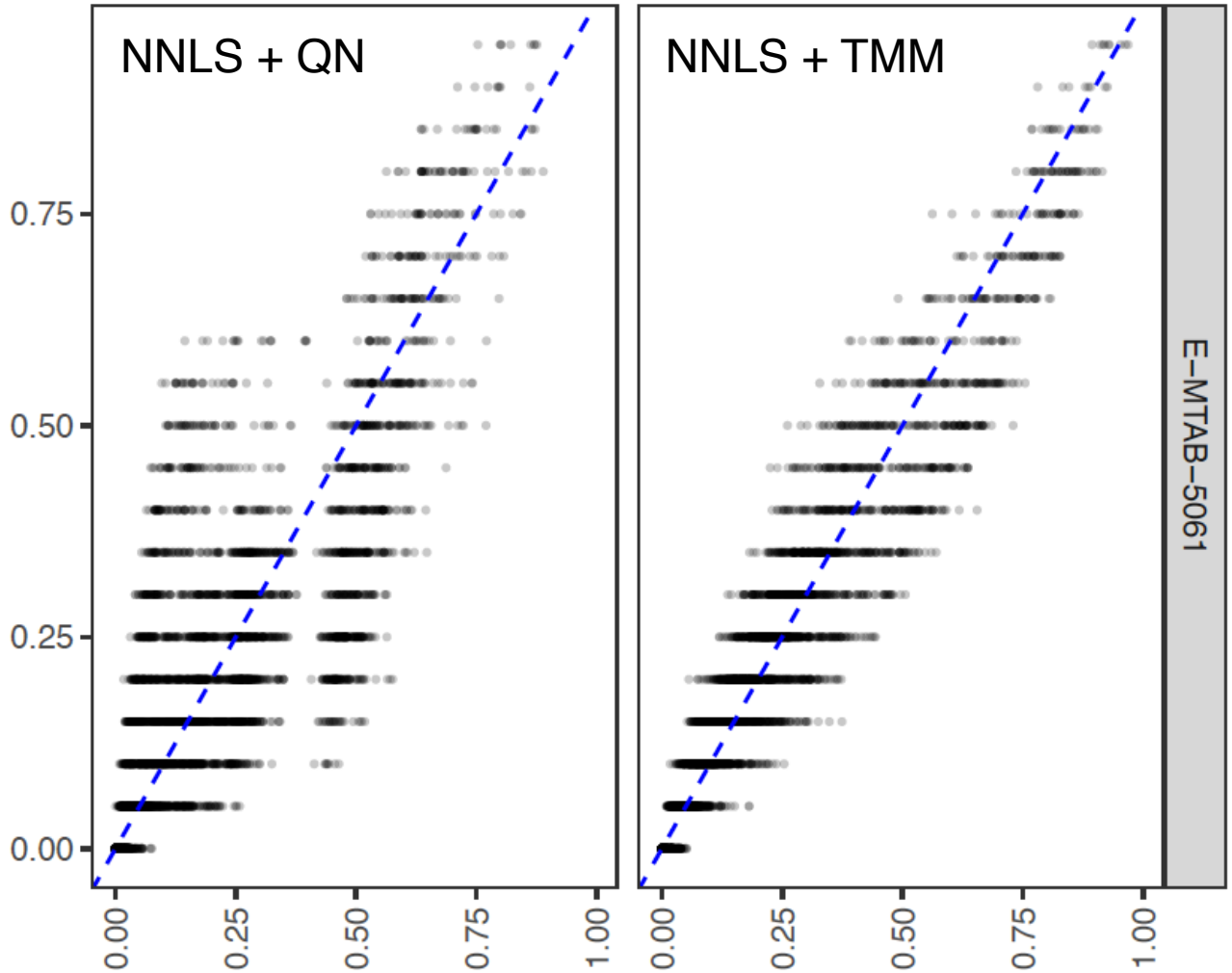
Skin-related	RIC5	epidermis development and keratinisation
	RIC7	epidermis development and keratinisation
	RIC19	epidermis development and keratinisation
	RIC31	epidermis development and keratinisation

Nazarov *et al,* 2019. BMC Medical Genomics

Data transformation has a dramatic impact on the deconvolution results

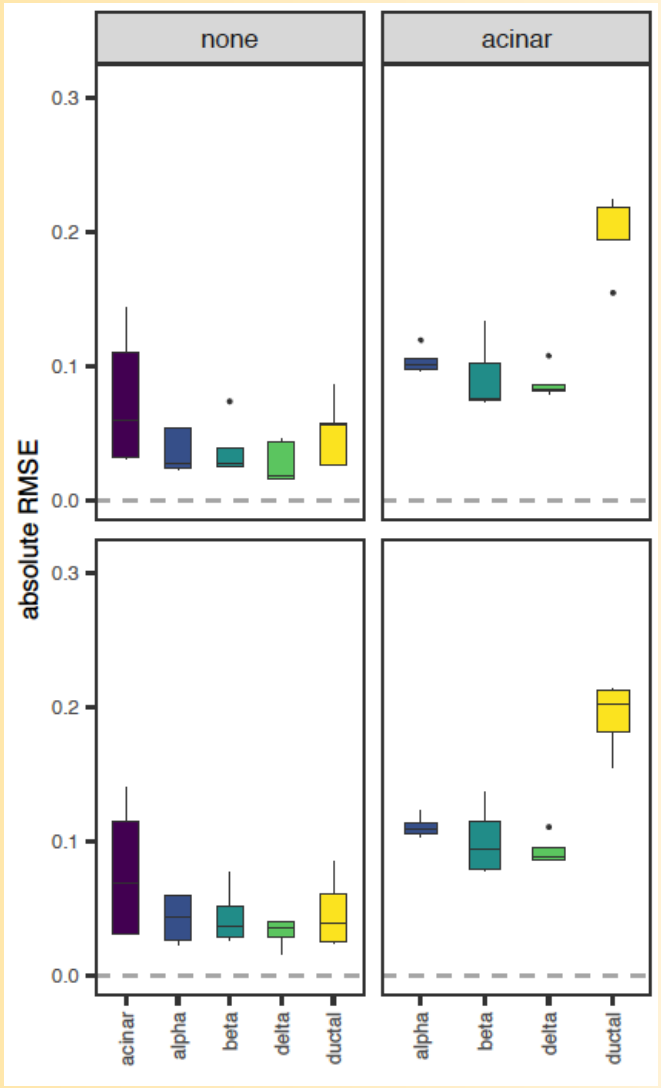


Different combinations of normalization and method reveal important differences in performance

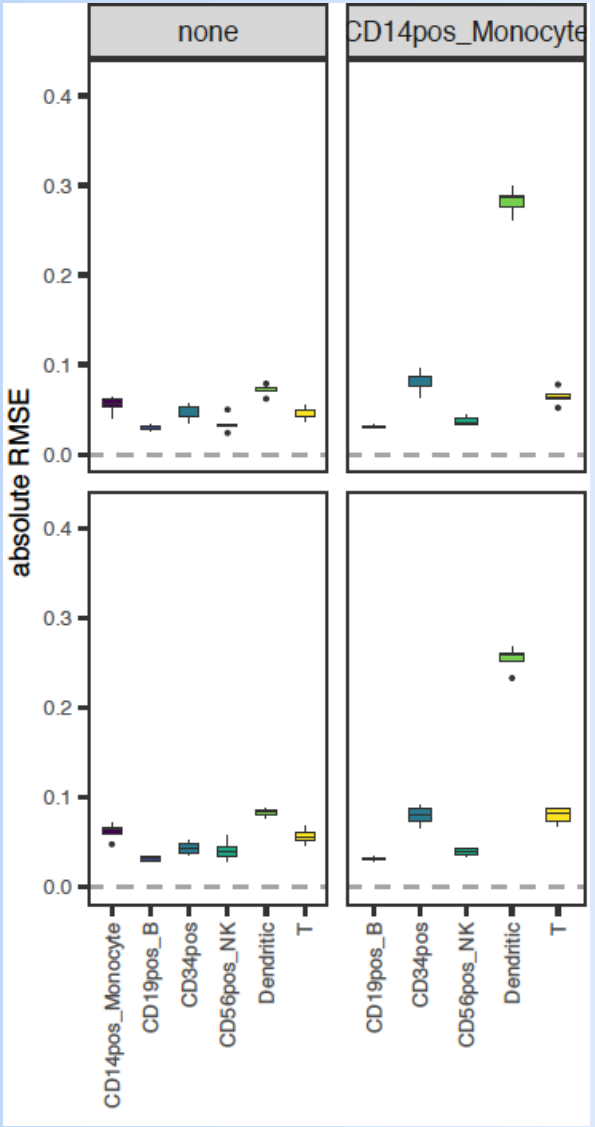


Removing cell types from the reference matrix results in substantially worse deconvolution results

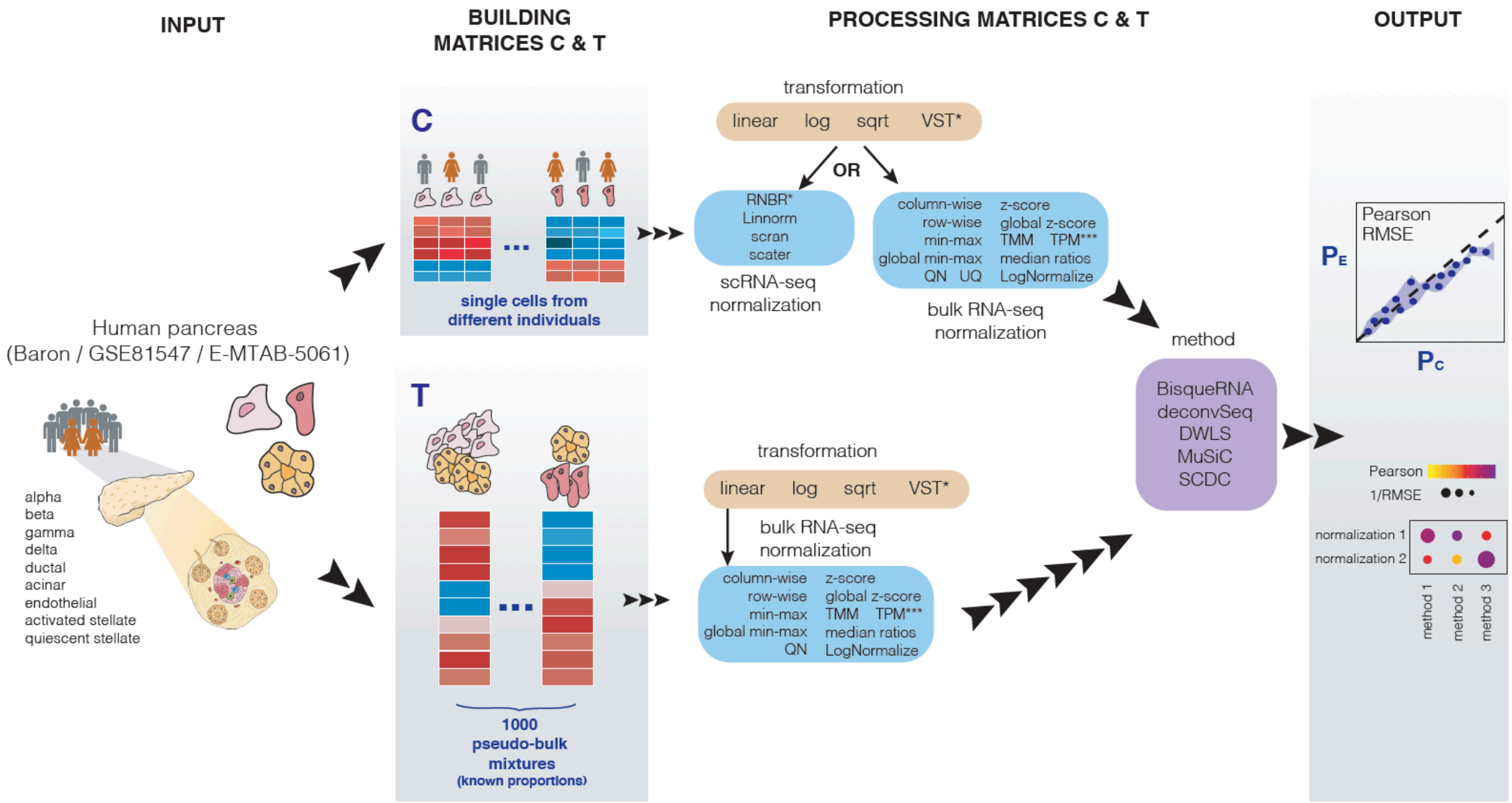
Human pancreas



PBMCs



Scenario 2: Computational deconvolution using scRNA-seq data



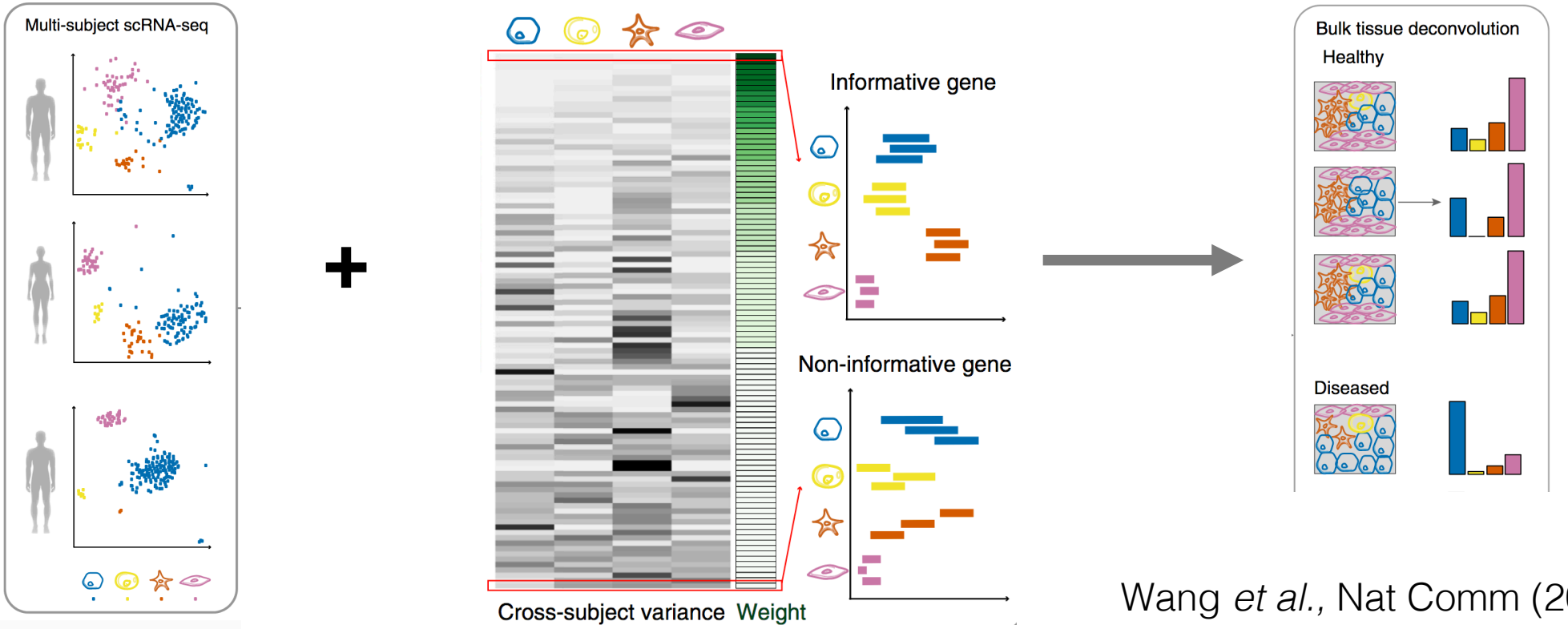
Avila Cobos *et al.*, in preparation

Mathematical approaches to solve the deconvolution problem

SINGLE-CELL

- Supervised:
  - a) Given T and C → **P**
    - DeconvSeq, **MuSiC, SCDC, DWLS**,…

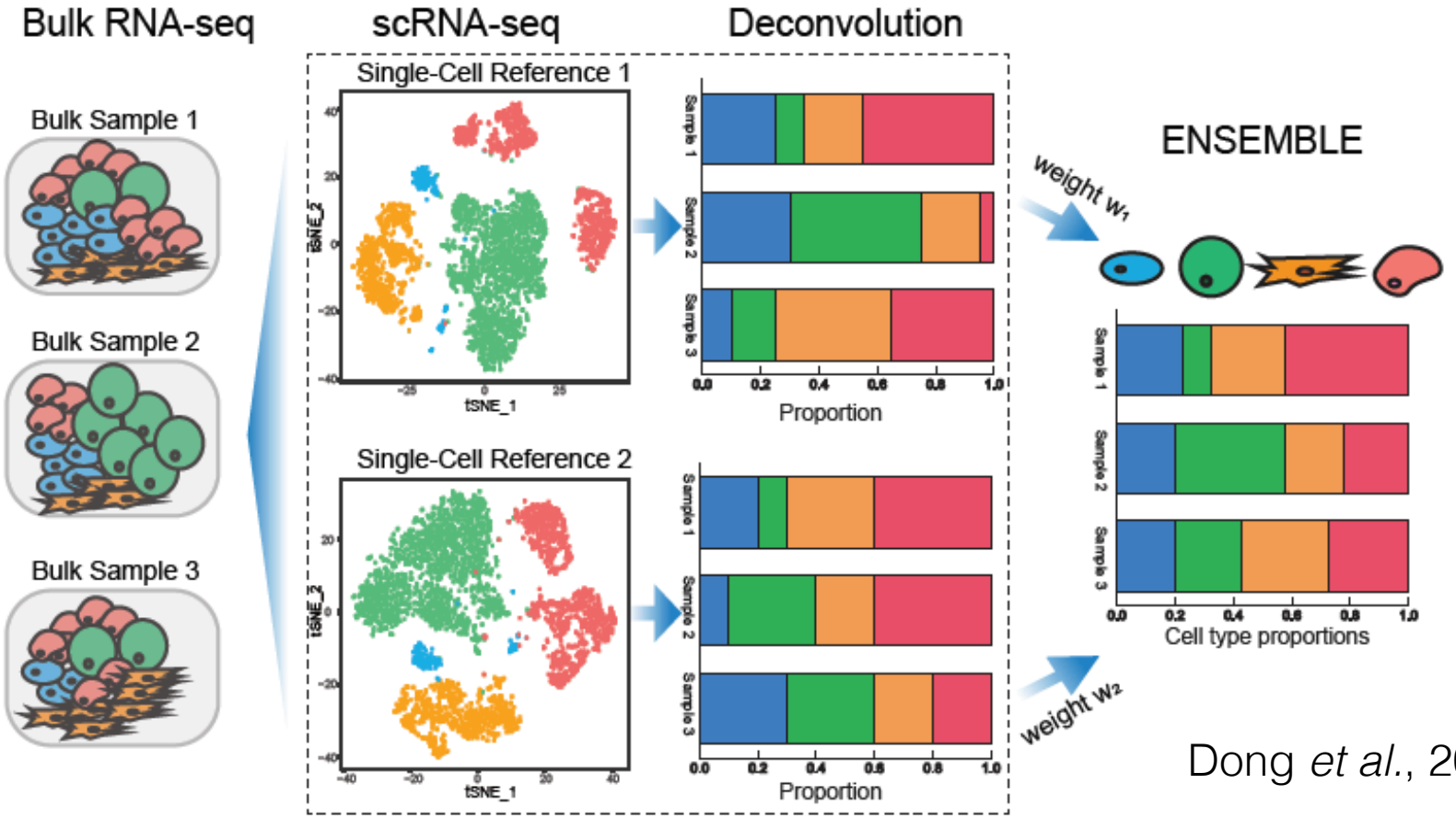
MuSiC: MUlti-Subject SIngle Cell deconvolution



Wang *et al.*, Nat Comm (2019)

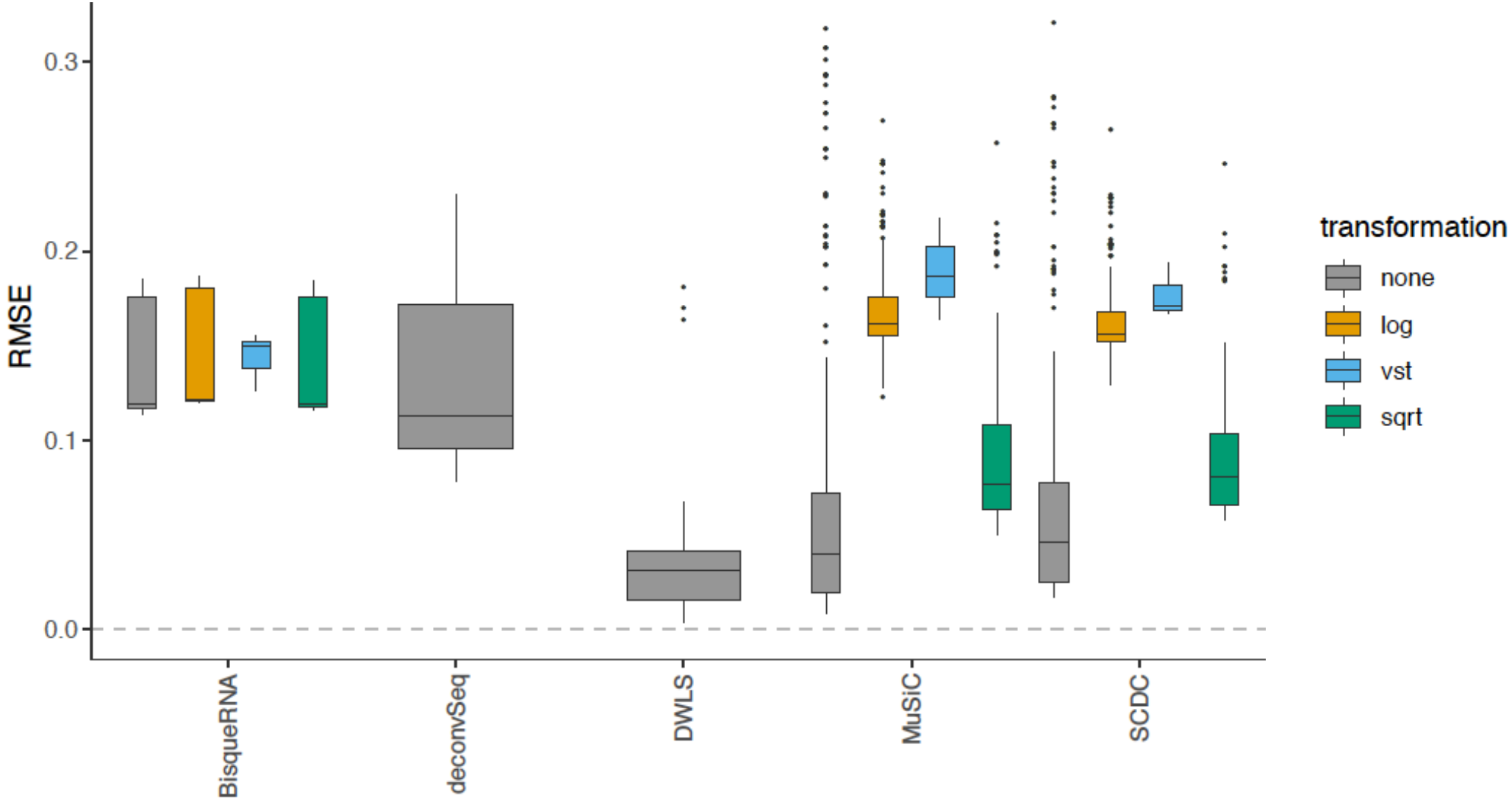
- **DWLS** (Tsoucas *et al.*, 2019): w-NNLS tweaked to adjust the contribution of each gene (e.g. avoid minimal contribution of good markers only due to low mean expression levels).

- **SCDC** (Dong *et al.*, 2019): w-NNLS + integrating multiple single-cell datasets at once while accounting for batch effects.



Dong *et al.*, 2019

Data transformation has a dramatic impact on the deconvolution results



Take-home messages

- Logarithmic transformation results in a poor performance.
- Computational deconvolution must be performed with data in linear scale.
- Different combinations of normalization and method reveal important differences in performance.
- Single-cell methods have comparable performance to the best performing bulk methods.
- Removing cell types from the reference matrix leads to worse results in both bulk and single-cell deconvolution frameworks.
- **Further reading**: Sturm *et al*. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology

Acknowledgements

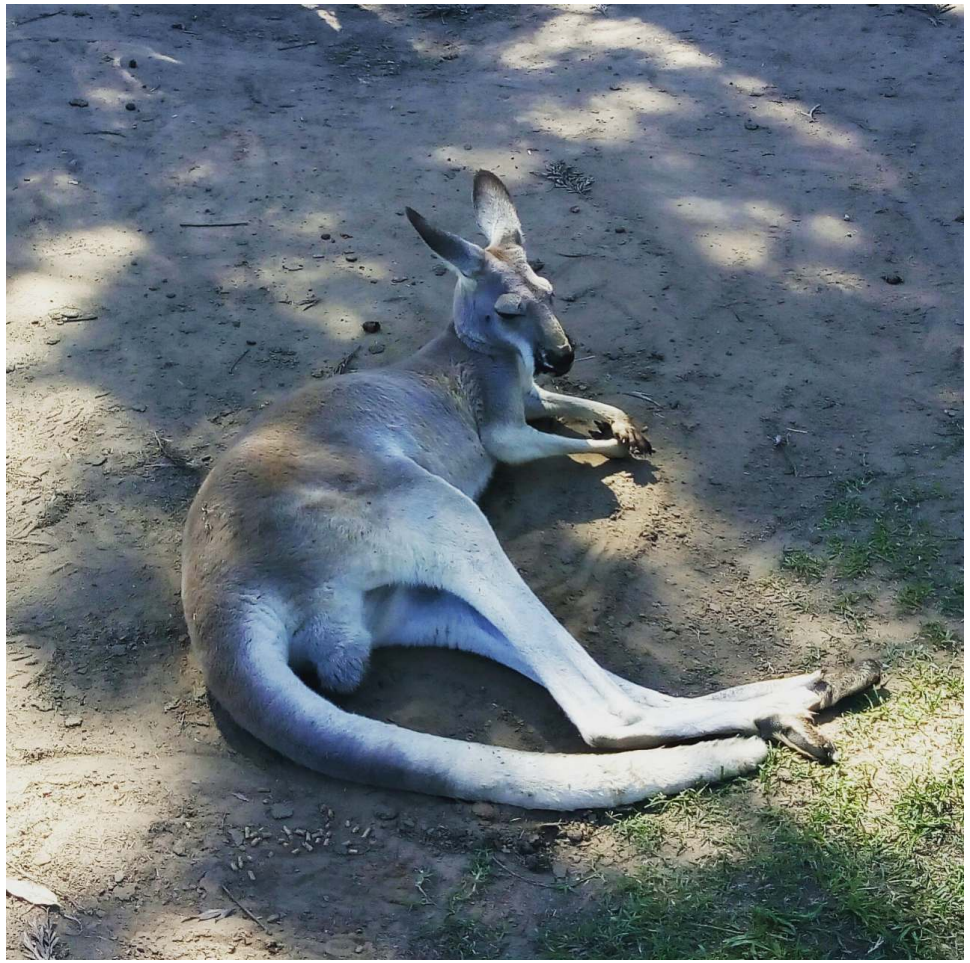
Center for Medical Genetics Ghent (CMGG)

Prof. Katleen De Preter
Prof. Pieter Mestdagh
Prof. Jo Vandesompele
Lucía Lorenzi
Eva Hulstaert

Single Cell and Computational Genomics

Prof. Joseph Powell
José Alquicira-Hernández
Evan Benn
Vikkitharan Gnanasambandapillai
Manuel Sopena Ballesteros
Derrick Lin

Questions?



@Frank_txu

francisco.avilacobos@ugent.be