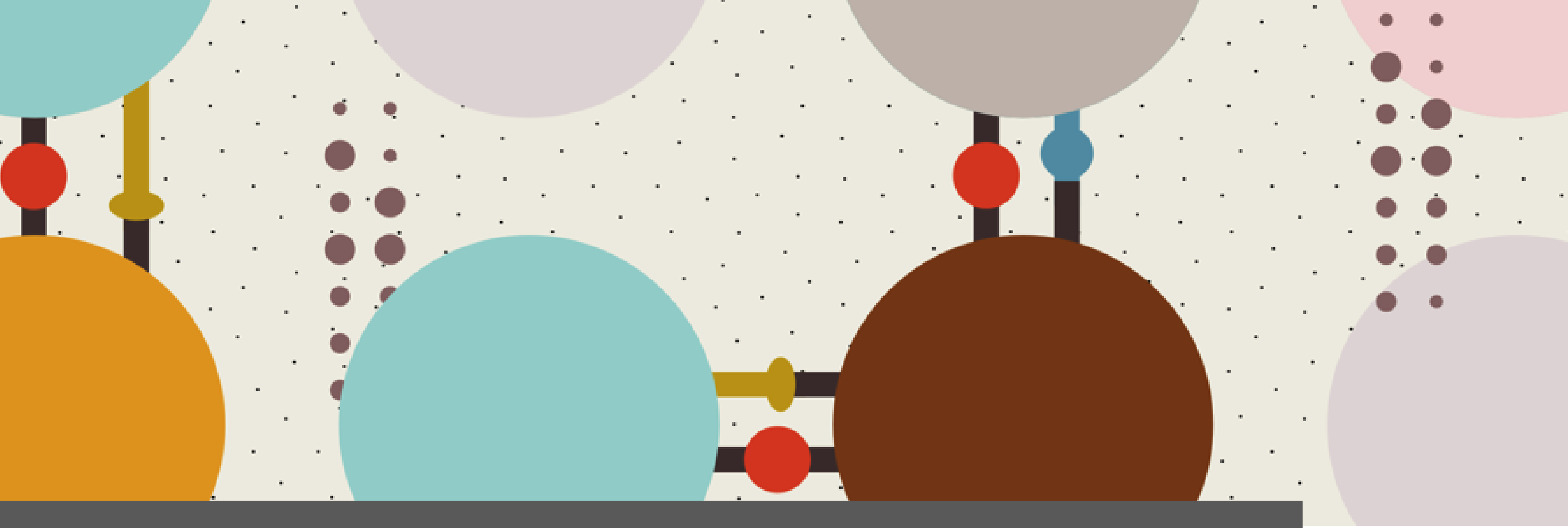


IO17 | Large Scale Bioinformatics for Immuno-Oncology

Deconvolution challenges

Francesca Finotello, Federica Eduati, and Pedro L. Fernandes

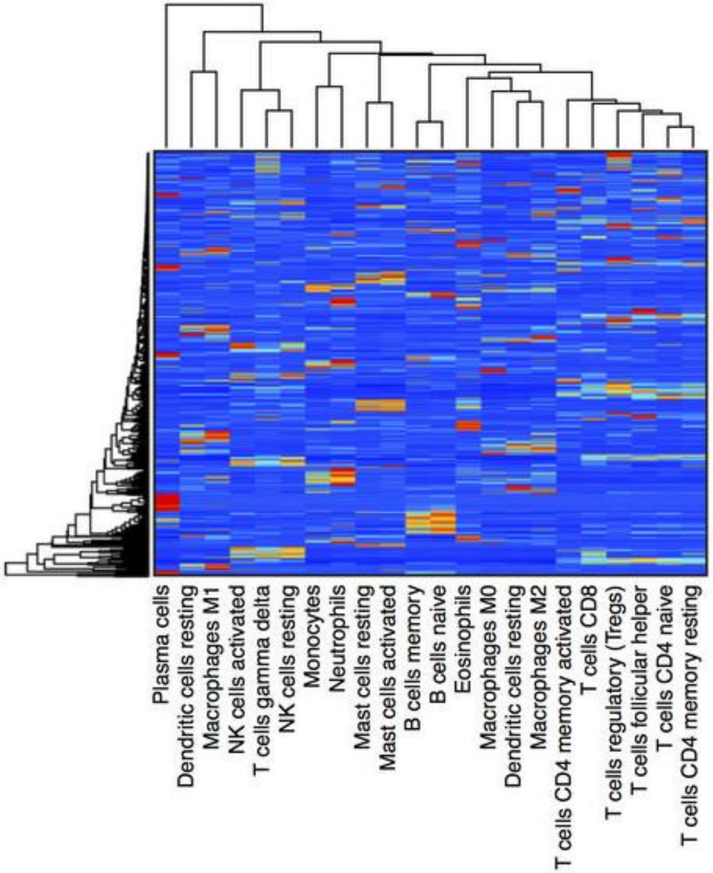
GTPB | The Gulbenkian Training Programme in Bioinformatics
Instituto Gulbenkian de Ciência, Oeiras, Portugal | Sept 19th-22nd, 2017

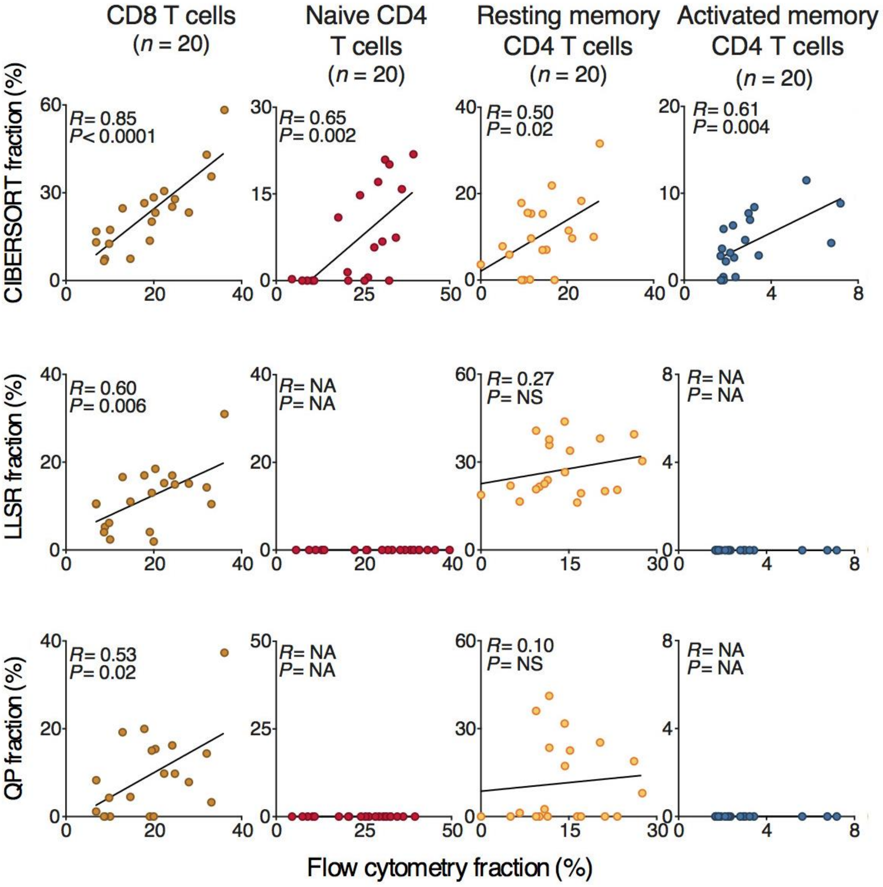
INSTITUTO
GULBENKIAN
DE CIÊNCIA

Multicollinearity

Multicollinearity: one predictor variable in a regression model can be linearly predicted from the others with a sufficient degree of accuracy

→Deconvolution of closely related cell types (e.g. T and B cell phenotyopes) is challenging!





Figures adapted from AM Newman et al, Nature methods, 2015

Data format

- High-troughput expression data can be generated with different **technologies** (e.g. Illumina HumanHT-12 beadchip microarrays or Illumina RNA-seq)

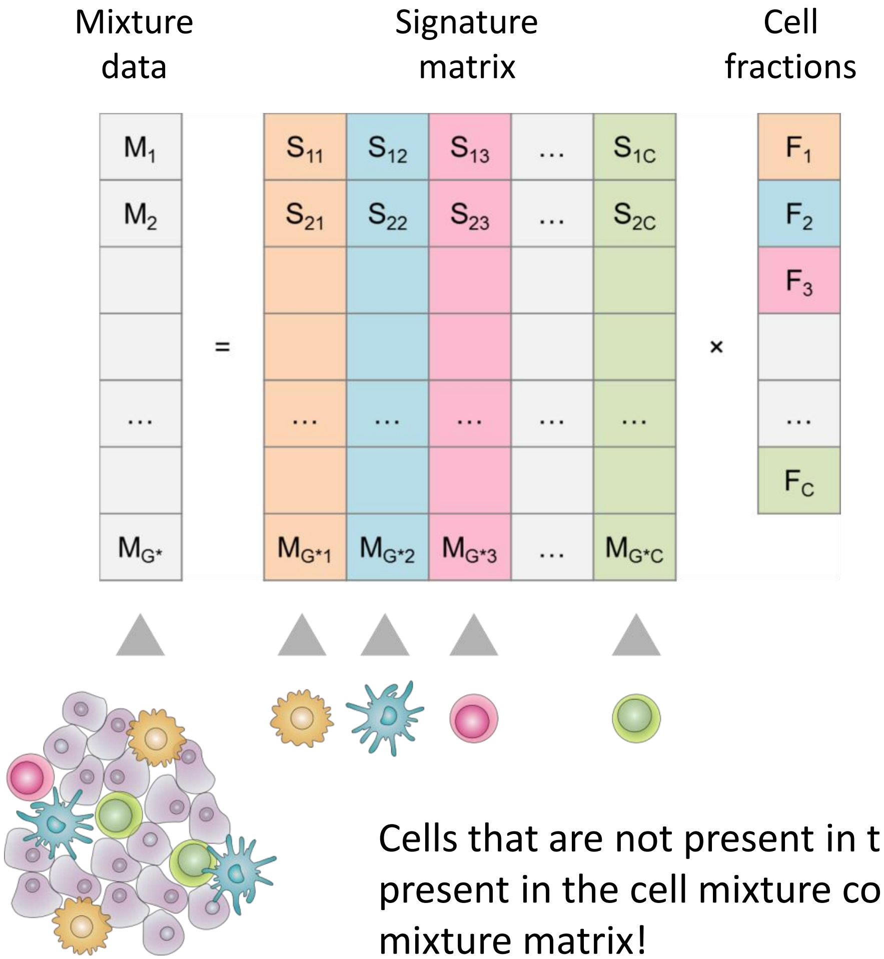
- Gene expression can be quantified using different **gene annotations** (for microarrays, also probeID → geneID). HGNC resources can be used to re-annotate gene symbols/IDs: http://www.genenames.org

- **Missing signature genes** can strongly affect deconvolution performance

- Gene expression in the signature and mixture matrix can be subjected to different **normalization** procedures: quantile, TPM, RPKM, …

- Note: deconvolution based on linear regression should be performed with expression data on the **natural scale**, not log (Y Zhong and Z Liu, Nature methods, 2012)

Unknown content



$$\begin{array}{c}M_1\\M_2\\ \\ \\ \ldots\\ \\ M_{G*}\end{array} = \begin{array}{ccccc}S_{11} & S_{12} & S_{13} & \ldots & S_{1C}\\ S_{21} & S_{22} & S_{23} & \ldots & S_{2C}\\ & & & & \\ & & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots\\ & & & & \\ M_{G*1} & M_{G*2} & M_{G*3} & \ldots & M_{G*C}\end{array} \times \begin{array}{c}F_1\\F_2\\F_3\\ \\ \ldots\\F_C\end{array}$$

$$M_1 = S_{11} F_1 + S_{12} F_2 + S_{12} F_3 + \ldots + S_{1C} F_C + S_{1X} F_X$$

$$M_2 = S_{21} F_1 + S_{22} F_2 + S_{13} F_3 + \ldots + S_{2C} F_C + S_{3X} F_X$$

…

$$M_{G*} = S_{G*1} F_1 + S_{G*2} F_2 + S_{G*3} F_3 + \ldots + S_{G*C} F_C + S_{G*X} F_X$$

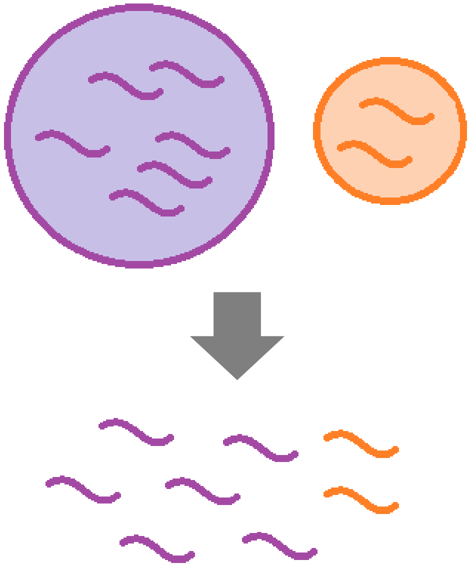
Cells that are not present in the signature matrix (e.g. tumor cells) but are present in the cell mixture contribute to the cumulative expression of the mixture matrix!

Deconvolution must be **robust to the unknown content**

Cell-specific mRNA content

Different cell types can have very different mRNA content

Cell types that contain more (less) mRNA tend to be over-estimated (under-estimated) by regression-based deconvolution methods



In **M=S**x**F, F** represents the mRNA fraction (not cell fraction) and should be scaled by a factor accounting for cell-specific differences in total mRNA content

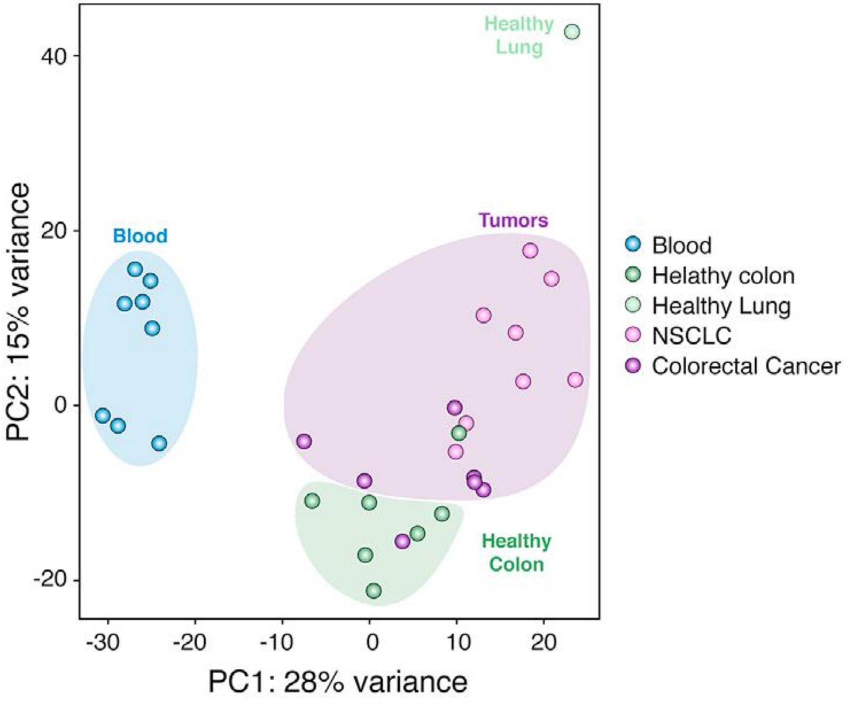
Tissue context

Most of immune-cell signatures are derived from expression data of immune cells purified/enriched from blood samples

Expression signatures of blood-derived immune cells may not represent the expression of immune cells in different tissues and disease conditions (e.g. tumor-infiltrating)

Better results can be obtained with tissue/disease-specific signatures

Regulatory CD4+ T cells



M De Simone, Immunity, 2016