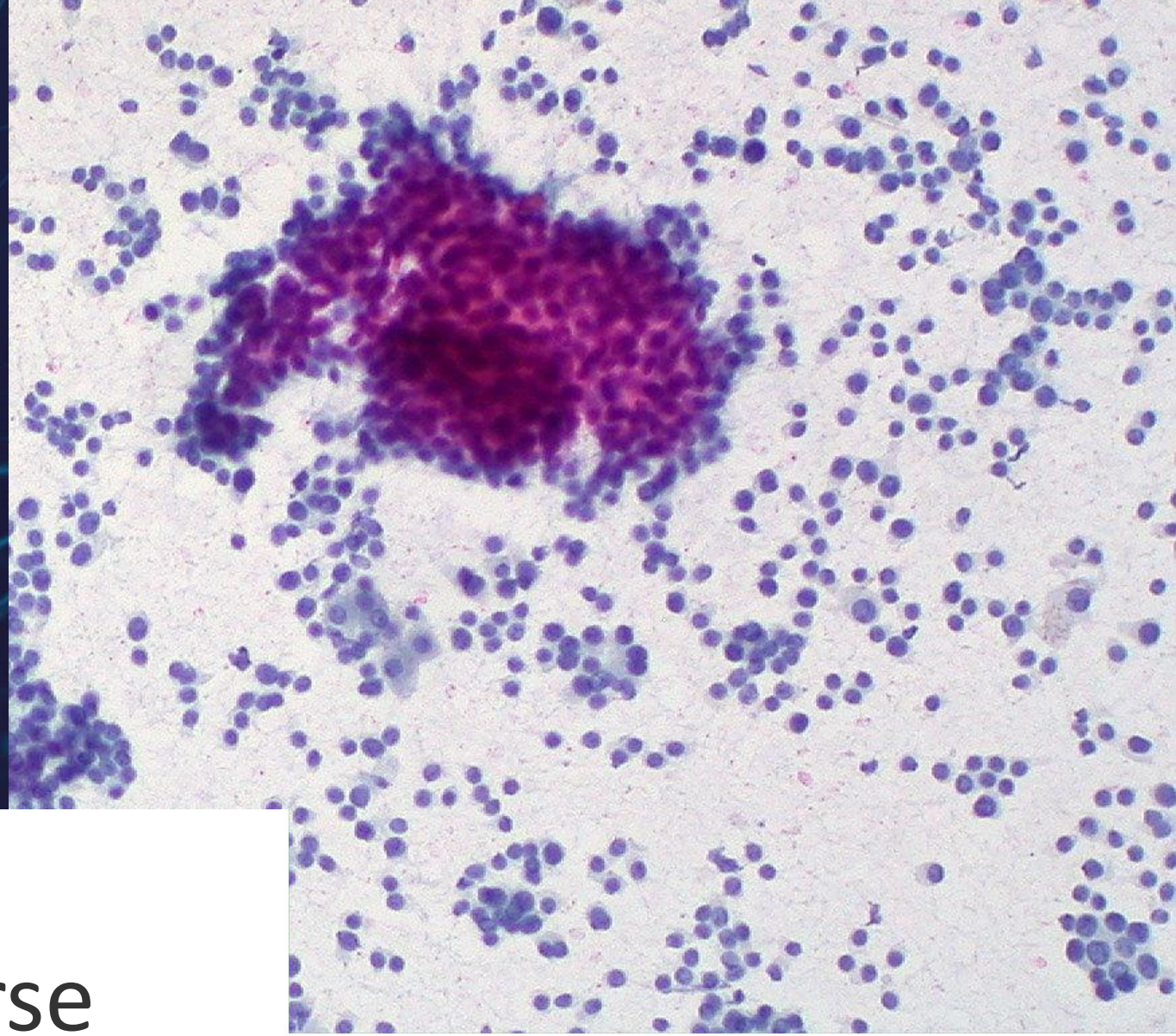


15-16 February 2021

COMETH Training course

From omics data

to tumor heterogeneity quantification

EIT Health is supported by the EIT, a body of the European Union

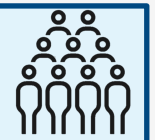
The program

DAY2

zoom

9:00 -10:00 pm LECTURE

9:00-10:00 pm Visualization and interpretation



zoom

10:00 -12:00 pm Practical work

Medical contributors	**Computational contributors**
Using COMETH web app on real datasets: small projects	Submit novel computational methods on codabench

Lunch Break

zoom

2:00-4:00 pm Practical work

2:00-2:30 pm Debriefing with slides from teams

Medical & Computational contributors

2:30-4.00 pm Focus on biological interpretation

zoom

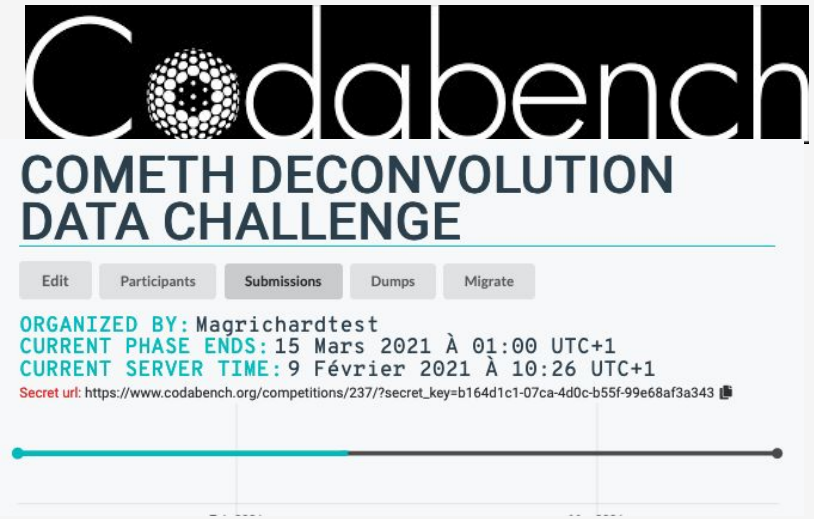
4:00-4:45 pm PRESENTATIONS

2:00-2:45 pm Results presentation & discussion

4:45 -5:00 pm CONCLUSION

EIT Health is supported by the EIT, a body of the European Union

In practical during the COMETH training



Computational group DAY 1-2

Learn how to contribute to the codabench benchmark using a toy data challenge

Cometh

Medical group DAY 1-2

Learn how to use the user-friendly COMETH web application to run methods on toy TCGA datasets

DAY 2

Learn how to biologically interpret the results of the methods

eit Health

...ch is supported by the EIT, a ...body of the European Union

16 February 2021

Bioligical interpretation

Yuna Blum and Ashwini Sharma

EIT Health is supported by the EIT, a body of the European Union

compExplore Shiny app

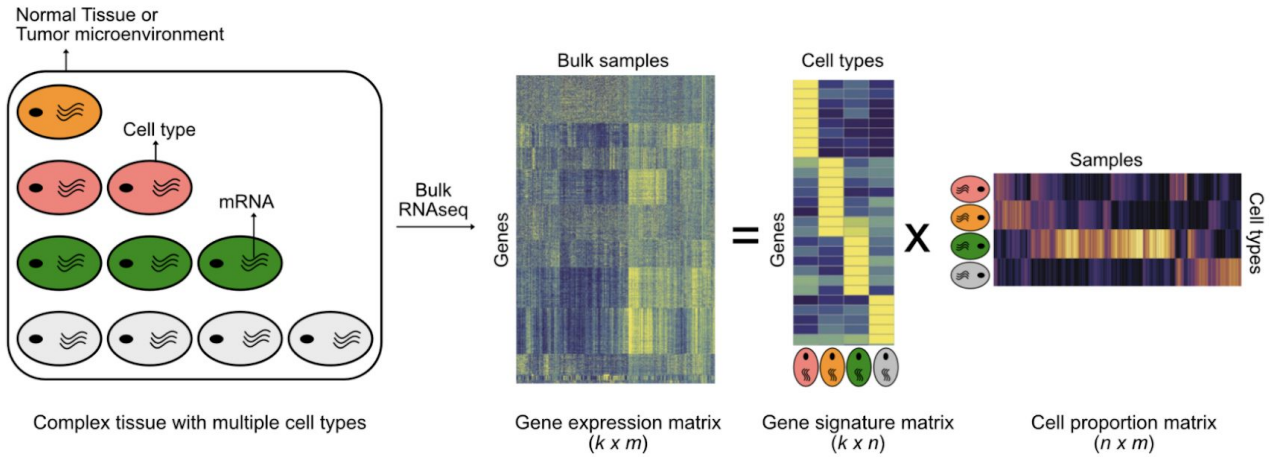
Help you in the analysis, interpretation and visualization of the results

compExplore | About | CSV-convertor | Number of CellTypes | Components enrichment | Proportion vizualisation

Different modules

compExplore

compExplore - Components explorer is a visualization tool to guide the user in the analysis and interpretation of the results from *Supervised* and *Unsupervised* gene expresssion deconvolution algorithms.



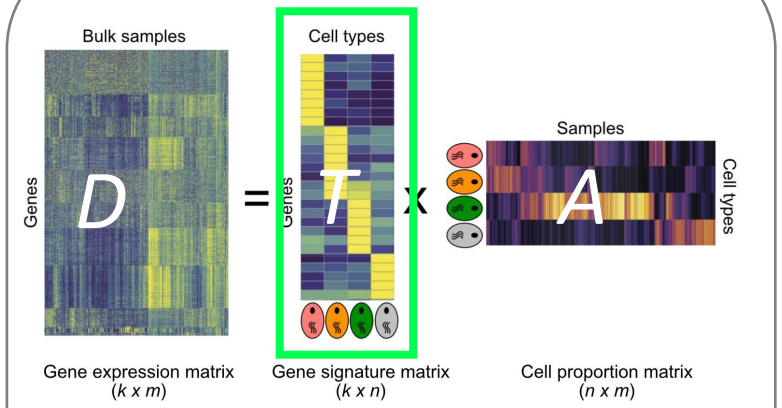
Terminology

1. Gene expression matrix - it is a *k x m* matrix with *k* rows of genes and *m* columns of samples. Each data point in this matrix represents the expression of a given gene in a given sample
2. Gene signature matrix - it is a *k x n* matrix with *n* rows of genes and *m* columns of cell fraction Each data point in this matrix represents the contribution of a gene towards a cell type
3. Cell proportion matrix - it is a *n x m* matrix with *n* rows of cell types and *m* columns of samples. Each data point in this matrix represents the proportion of a given cell type in a given sample

https://app.gebican.fr/compExplore/

Different type of computational methods

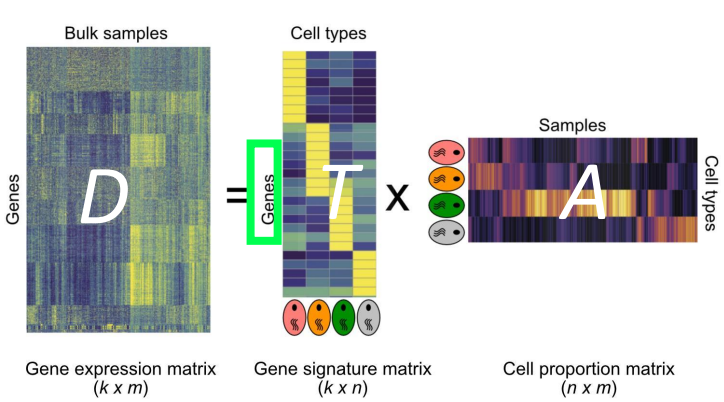
Supervised



T matrix known

Cibersort (MT8), **EPIC** (MT9), **quantiseq** (MT11)

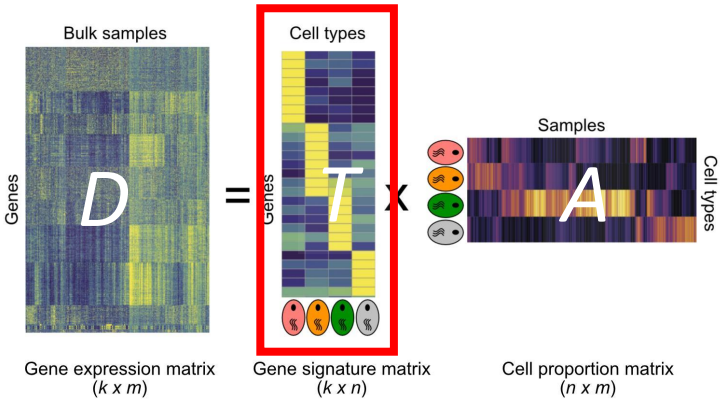
Semi-supervised



Gene markers known

cellmix (MT16, 17, 18) using cellMatch gene markers

Unsupervised



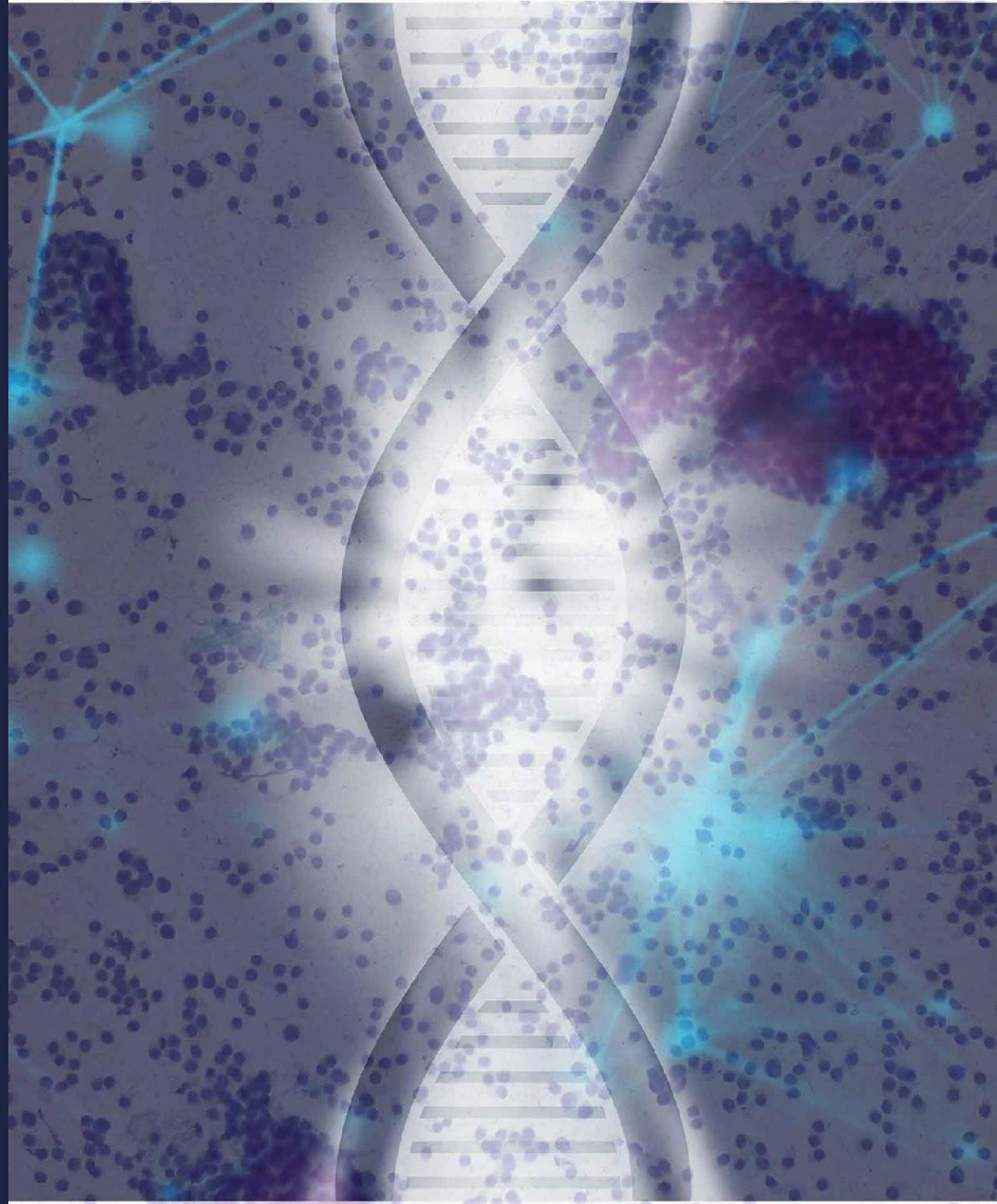
T matrix unknown

Number of component to consider (#cell types)?

Interpretation of the components

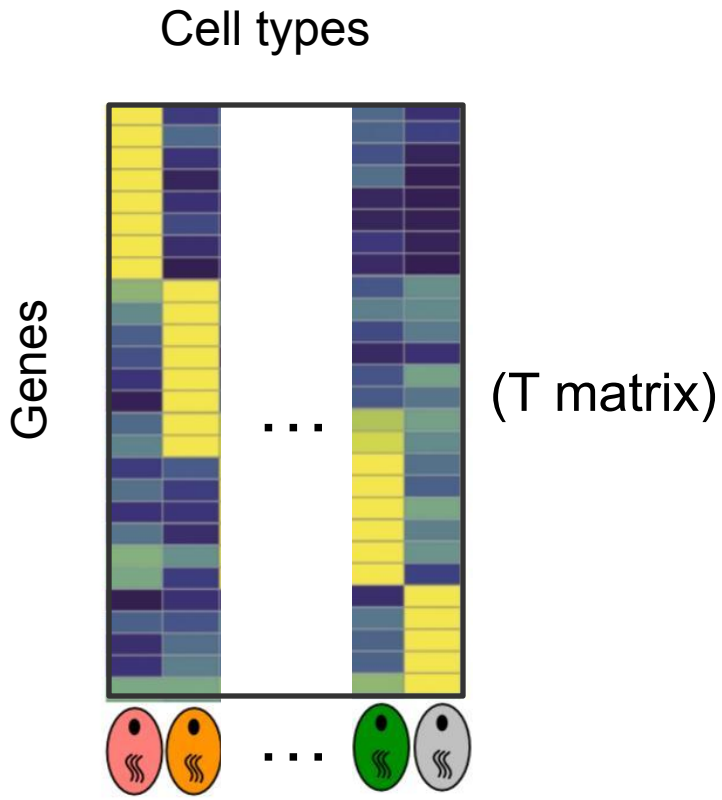
ICA with fs (MT1_ICA_fs,), **NMF with fs** (MT2_NMF_fs ,), **Edec method** (MT3_edec ,), **ICA without fs** (MT14_ICA), **NMF without fs** (MT19_NMF)

Unsupervised methods:
finding the number of k of cell types



Unsupervised

Finding the number of cell types k

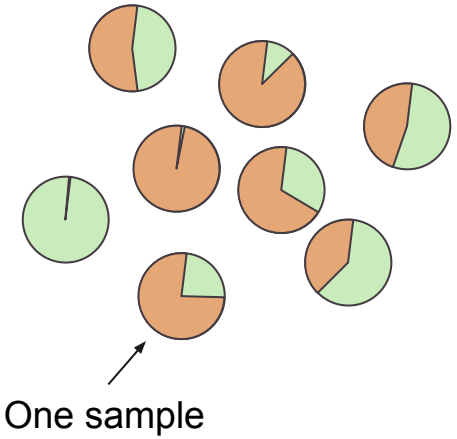


k = ?

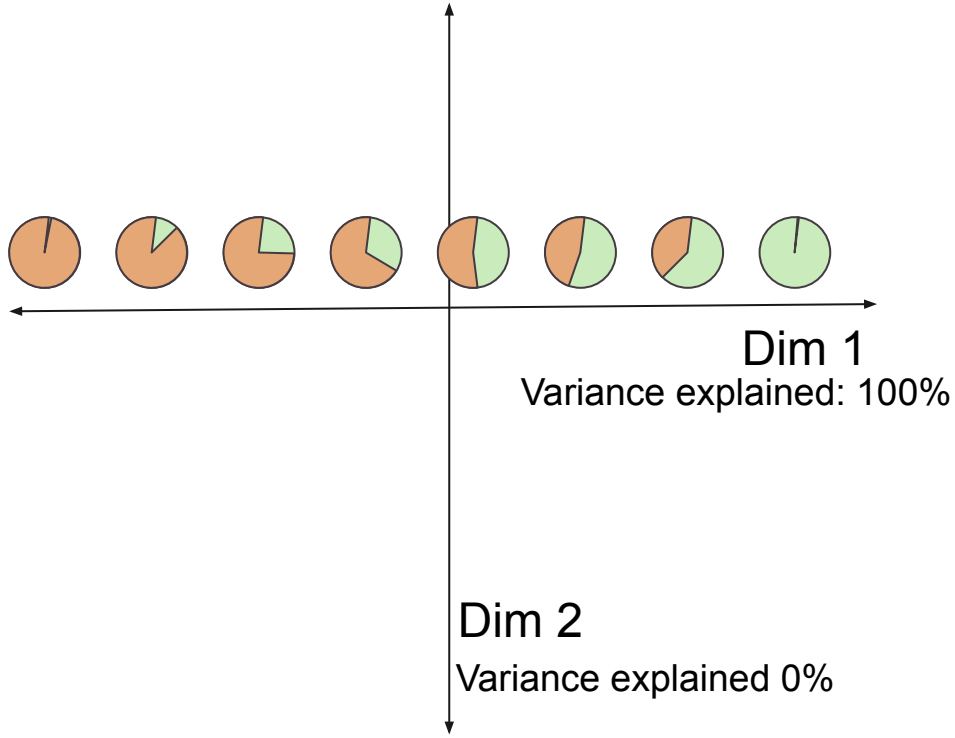
Unsupervised

Finding the number of cell types k

Samples
mixtures of 2 cell types

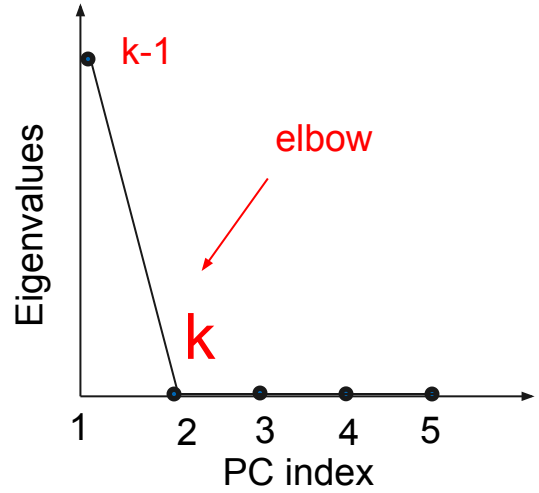


Principal Component Analysis (PCA)



Reminder
Find the axes that maximized the explained variance (inertia)
Principal components are orthogonal

Plot of eigenvalues (=scree plot)



Eigenvalues represent the variance explained

Cattel's rule: k = PCs + 1

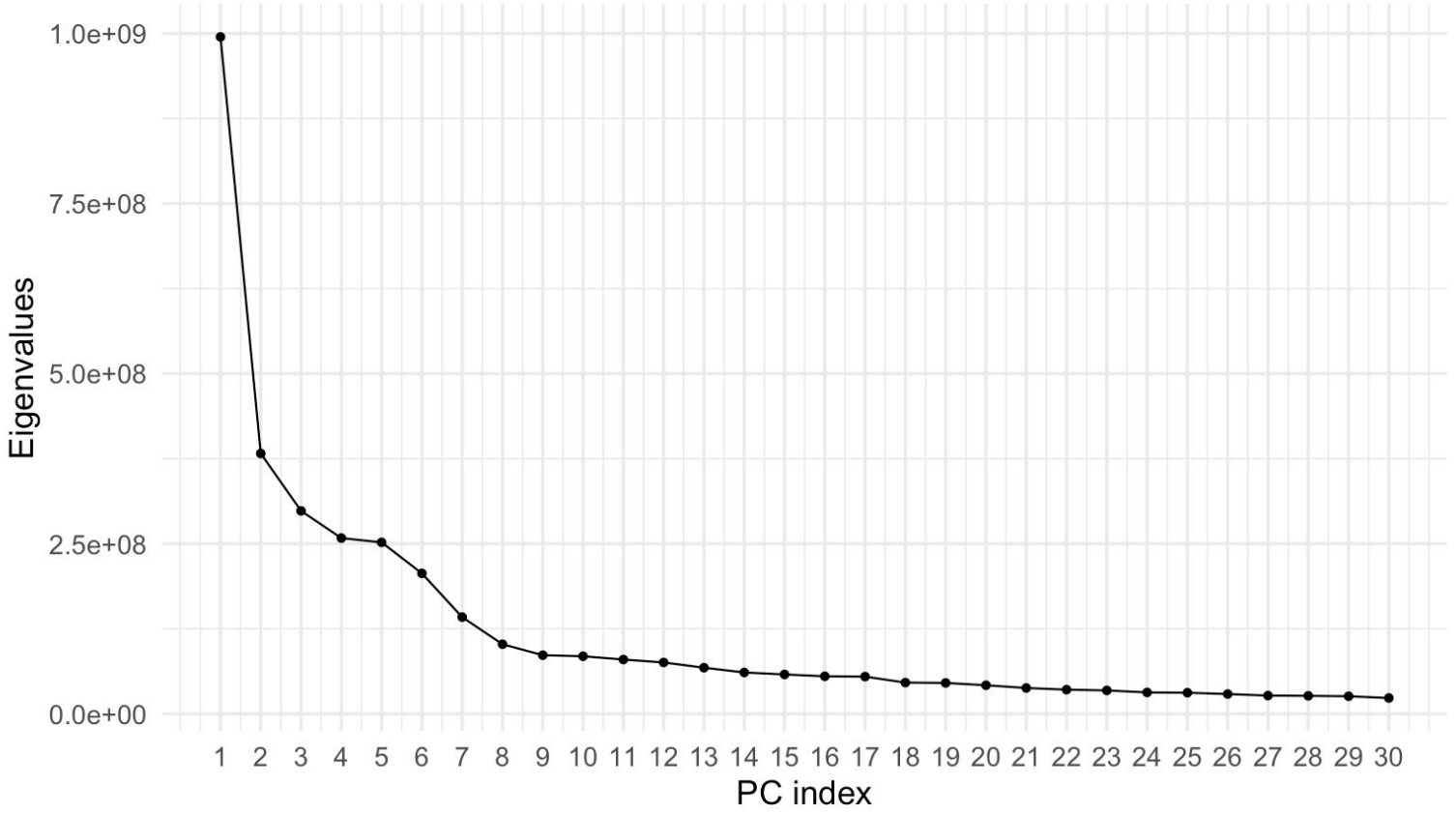
Number of relevant PCs

Unsupervised

Finding the number of cell types k

Real life

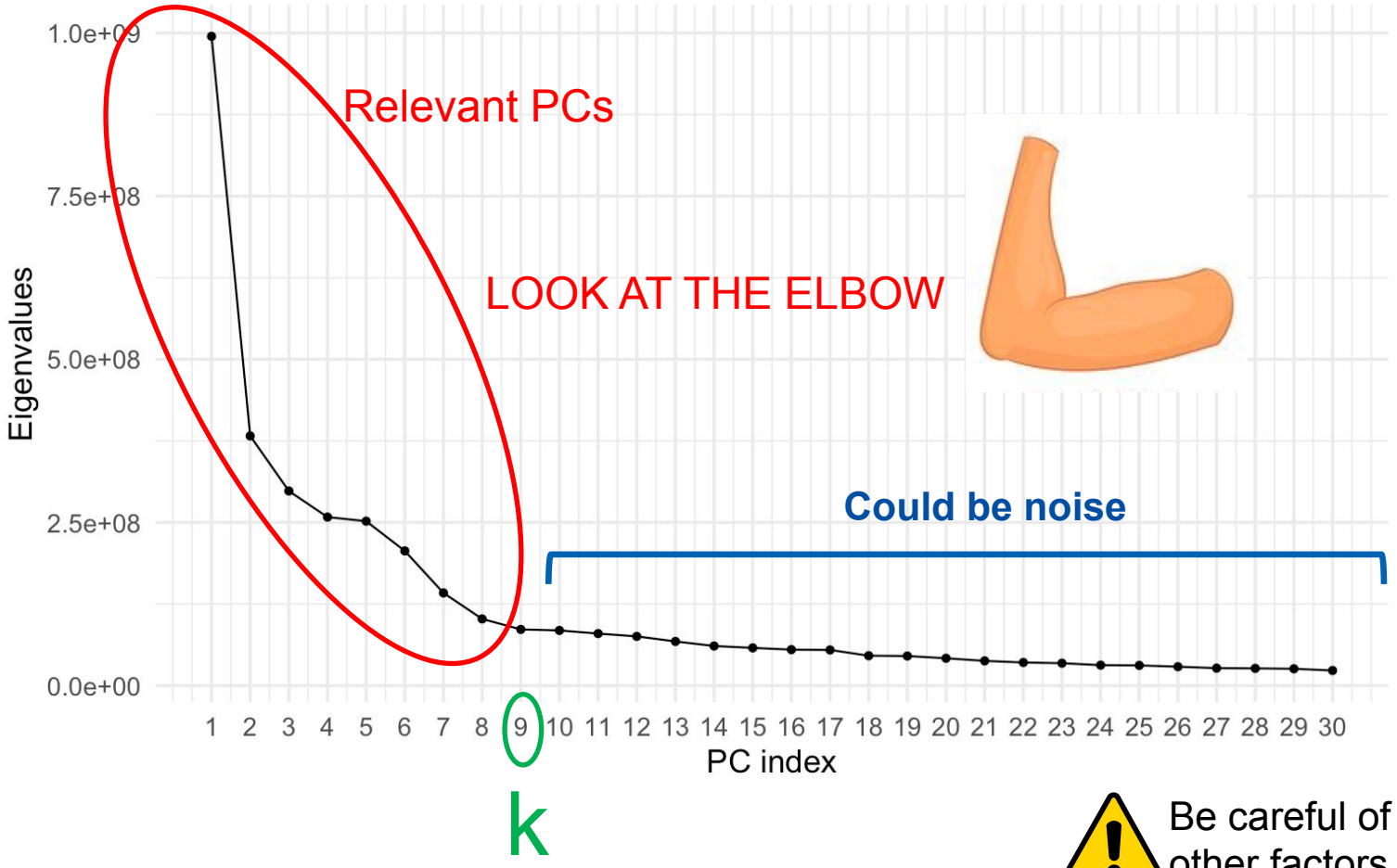
Cattel's rule: k = PCs + 1



Finding the number of cell types k

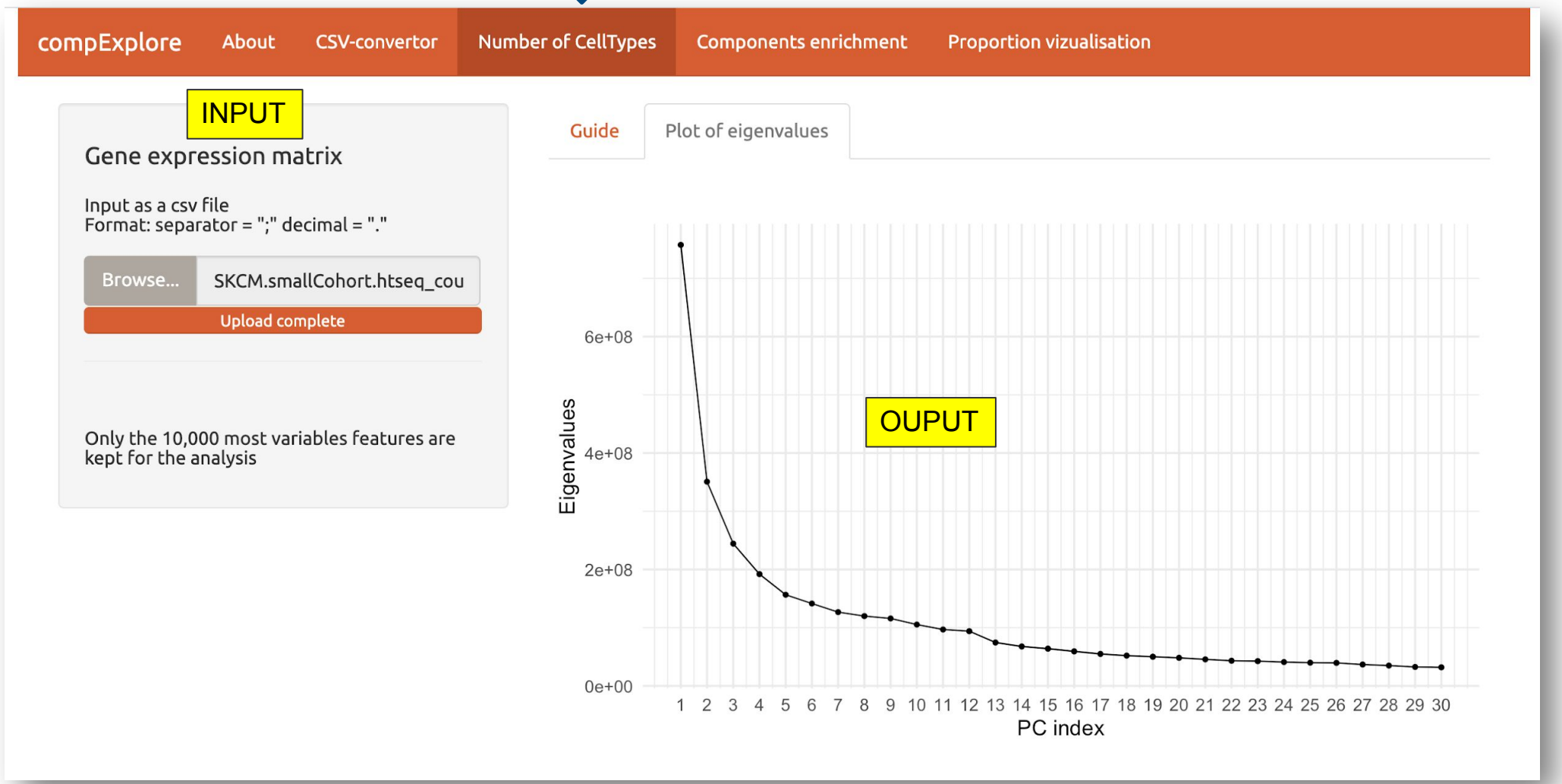
Real life

Cattel's rule: k = PCs + 1

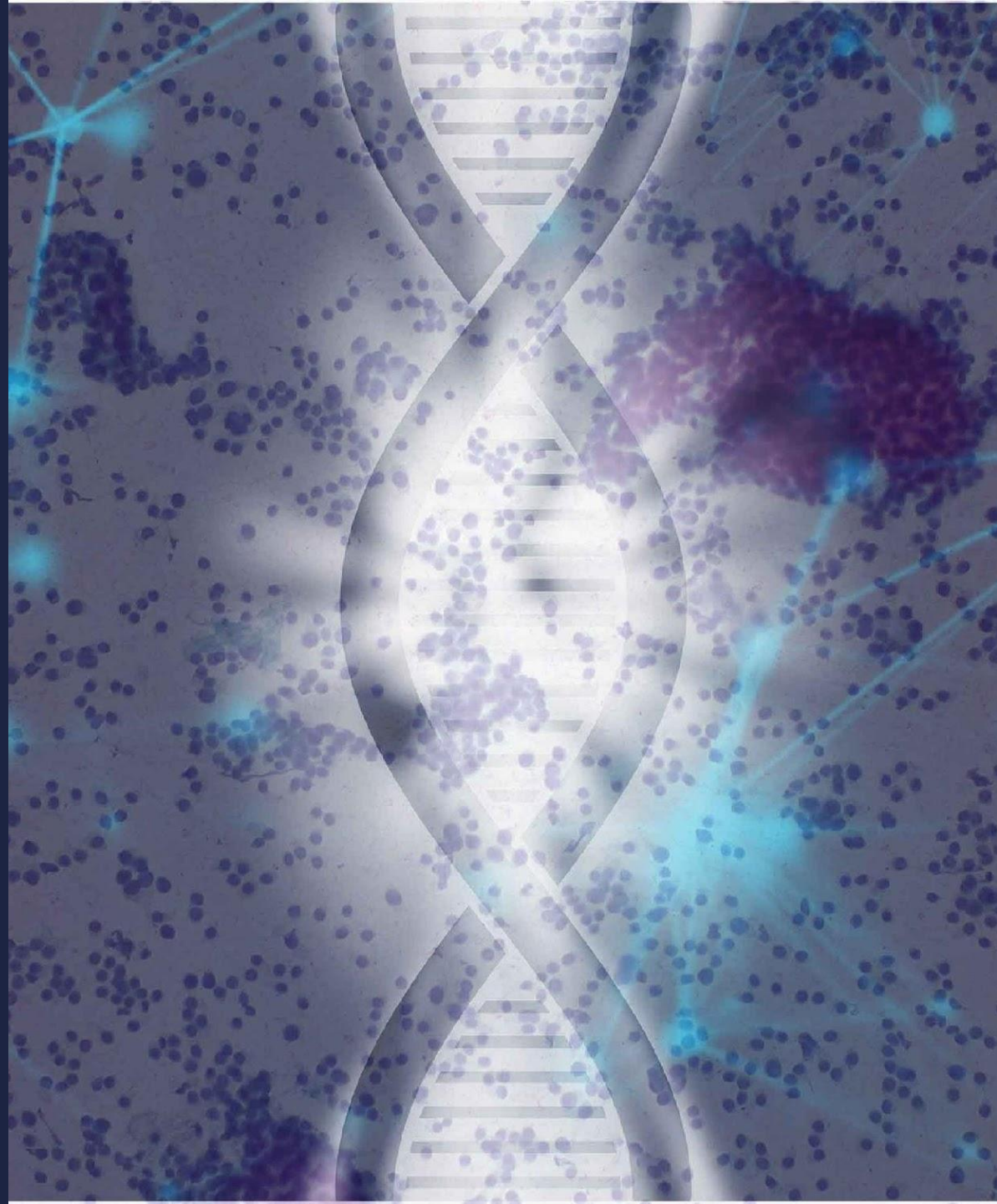


Be careful of overestimation of k due to other factors (sex, age, batch effect)

compExplore Shiny app

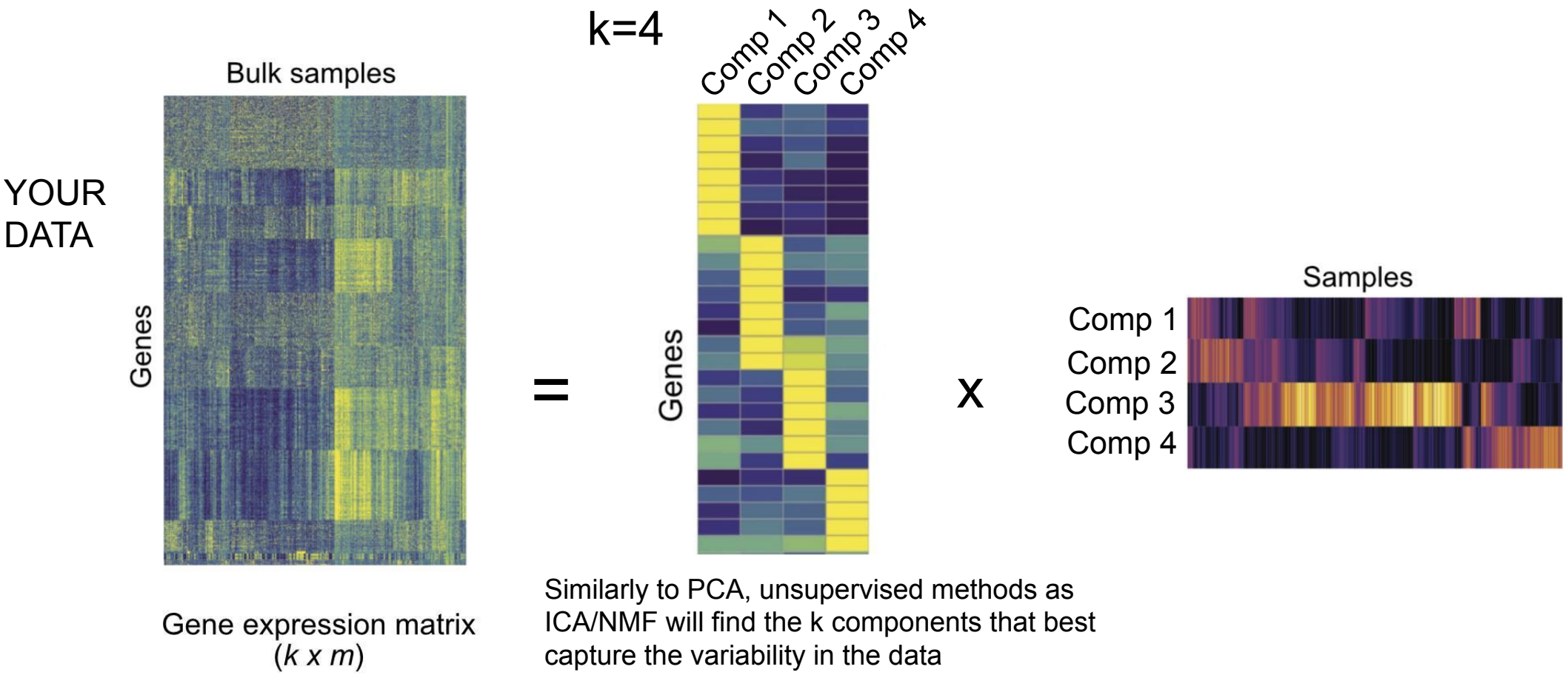


Unsupervised methods:
Interpret the components identified



Unsupervised

Interpret the components identified

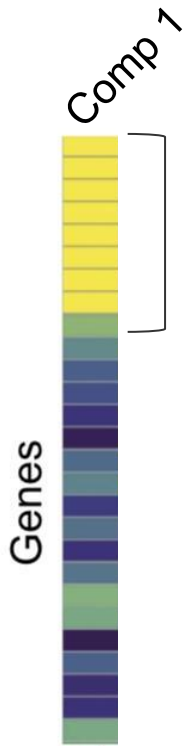


Similarly to PCA, unsupervised methods as ICA/NMF will find the k components that best capture the variability in the data

To which cell type(s) corresponds each of the components identified by unsupervised methods?

Unsupervised

Interpret the components identified



Genes with high scores on the component

Markers of a particular cell types?

Use of *CellMatch* Database

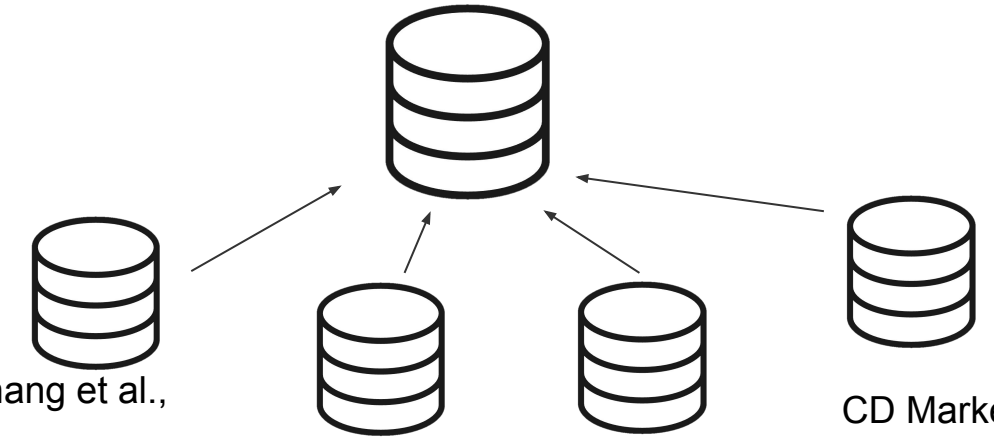
(Shao et al iScience, 2020, tool scCATCH)

Unsupervised

Interpret the components identified

Use of *CellMatch* Database

(Shao et al iScience, 2020, tool scCATCH)



CellMarker (Zhang et al., 2019b)

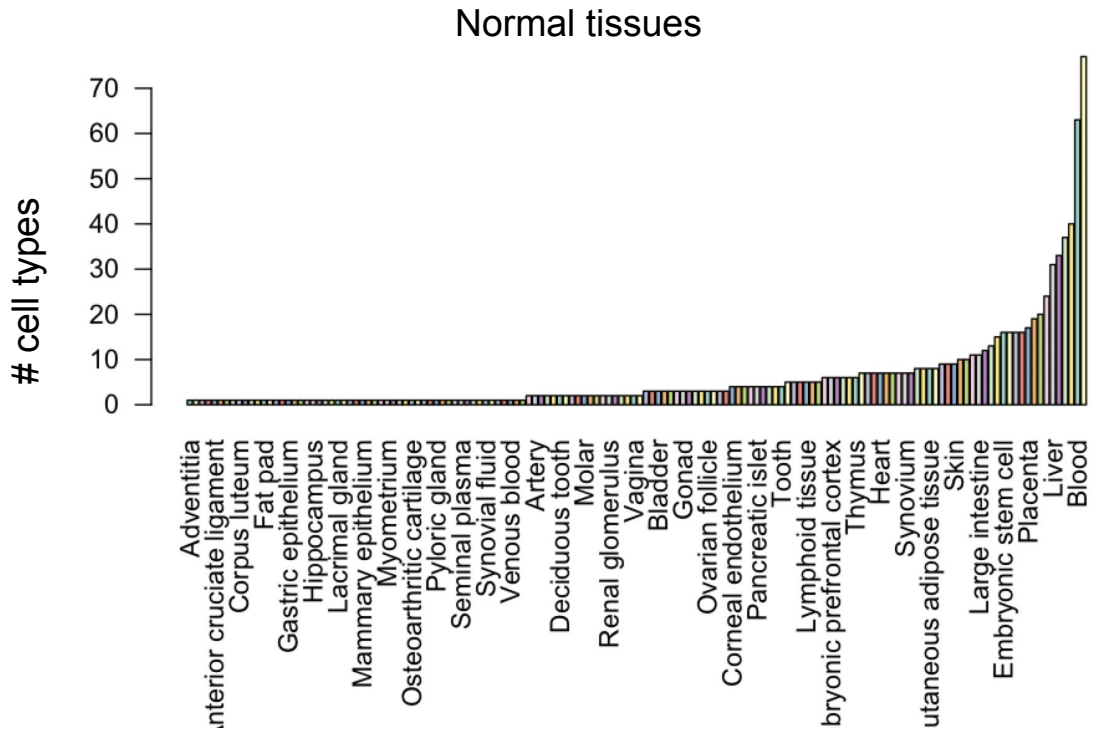
MCA (Han et al., 2018)

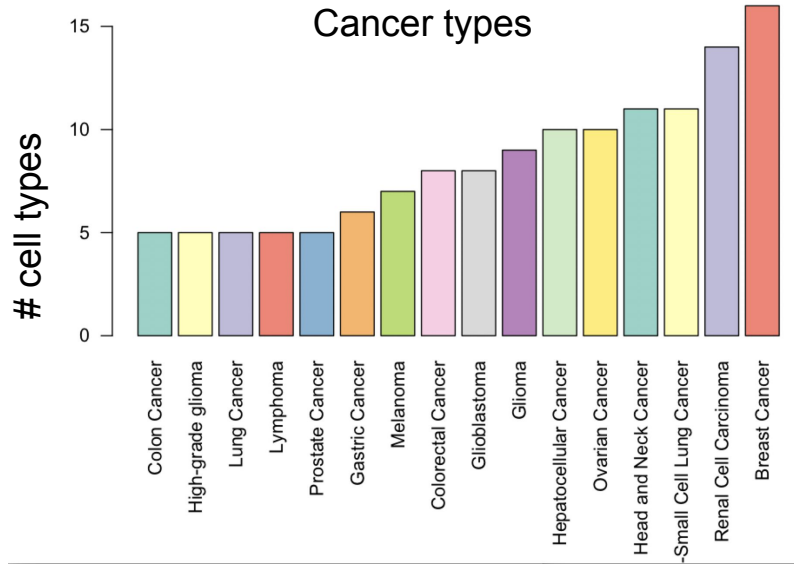
CancerSEA (Yuan et al., 2019)

CD Marker Handbook

- **33** cancer types + normal
- **150** tissues
- **412** cell types
- **12312** gene markers

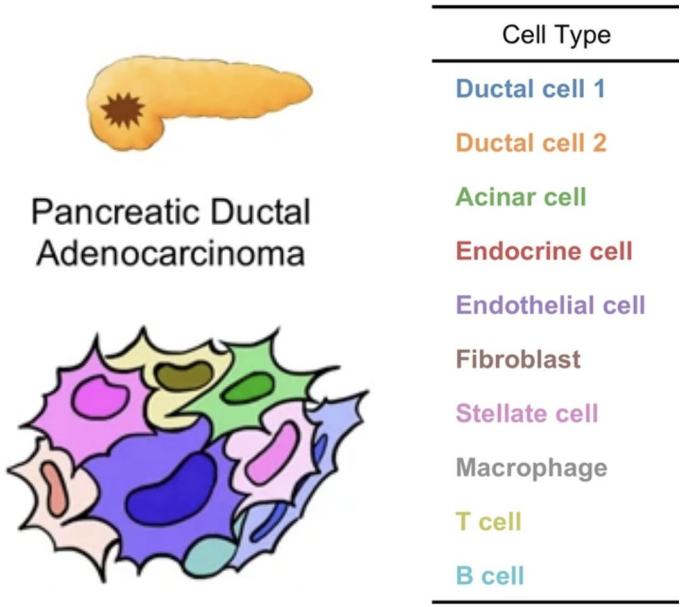
Filtering human gene markers and cancer types with at least 5 gene markers





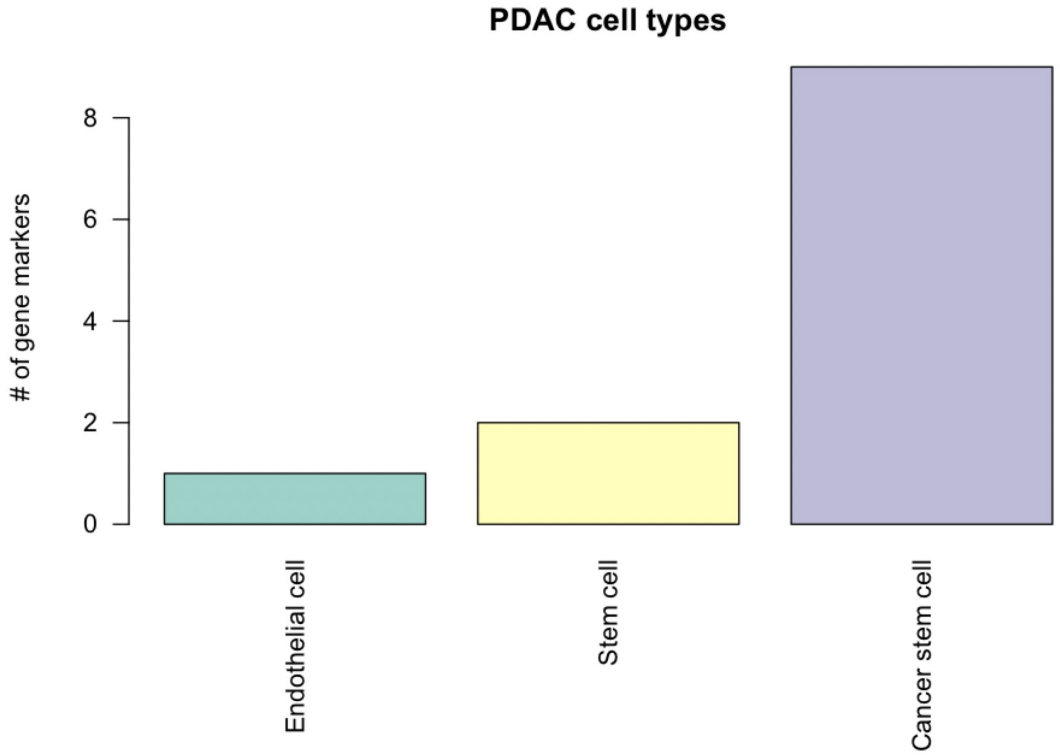
Unsupervised

Interpret the components identified



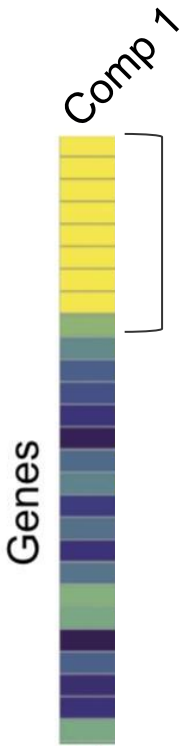
Peng et al 2019 Nature

CellMatch Database



Unsupervised

Interpret the components identified



Genes with high scores on the component

Markers of a particular cell types?

Use of *CellMatch* Database

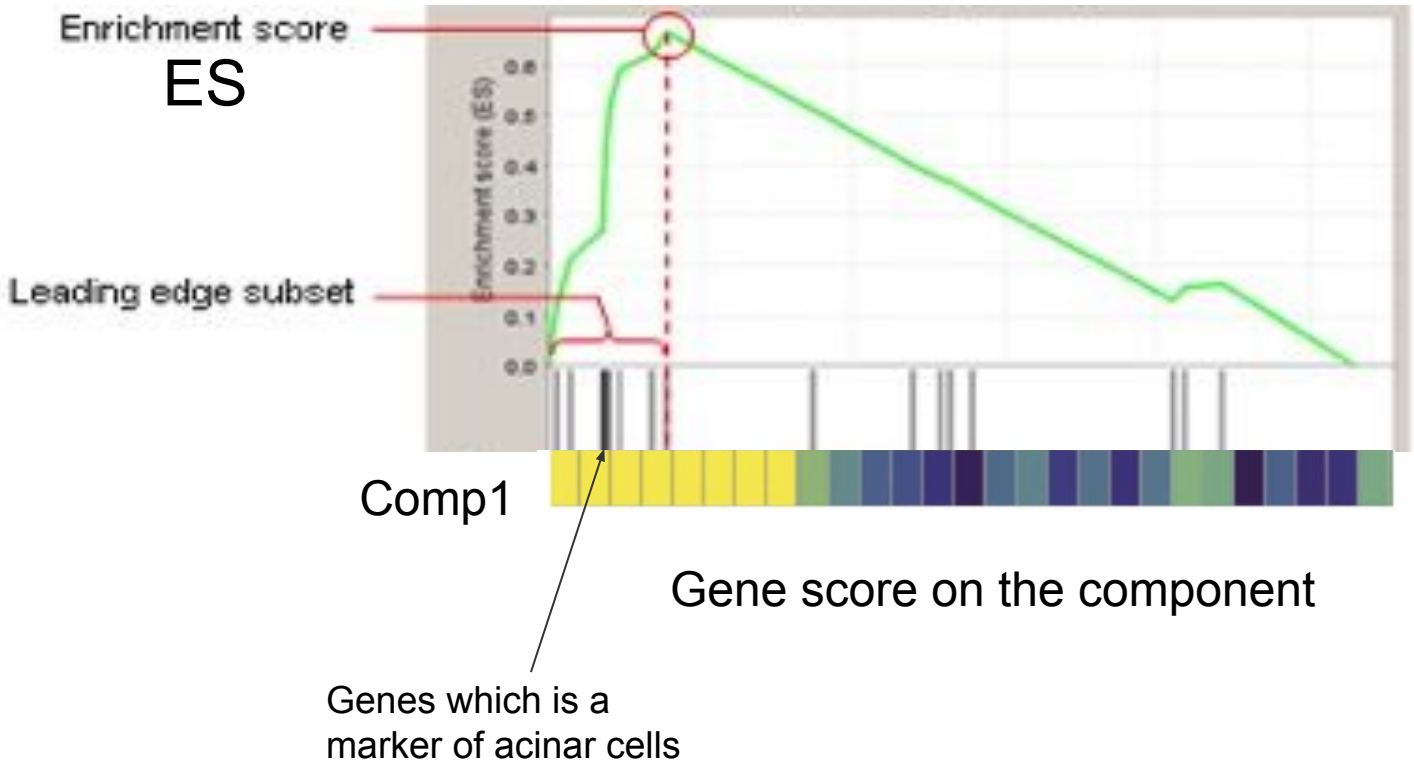
(Shao et al iScience, 2020, tool scCATCH)

Gene Set Enrichment Analysis (GSEA)

Interpret the components identified

Gene Set Enrichment Analysis (use of the fgsea R package)

Enrichment plot
For one cell type, e.g. Acinar cells



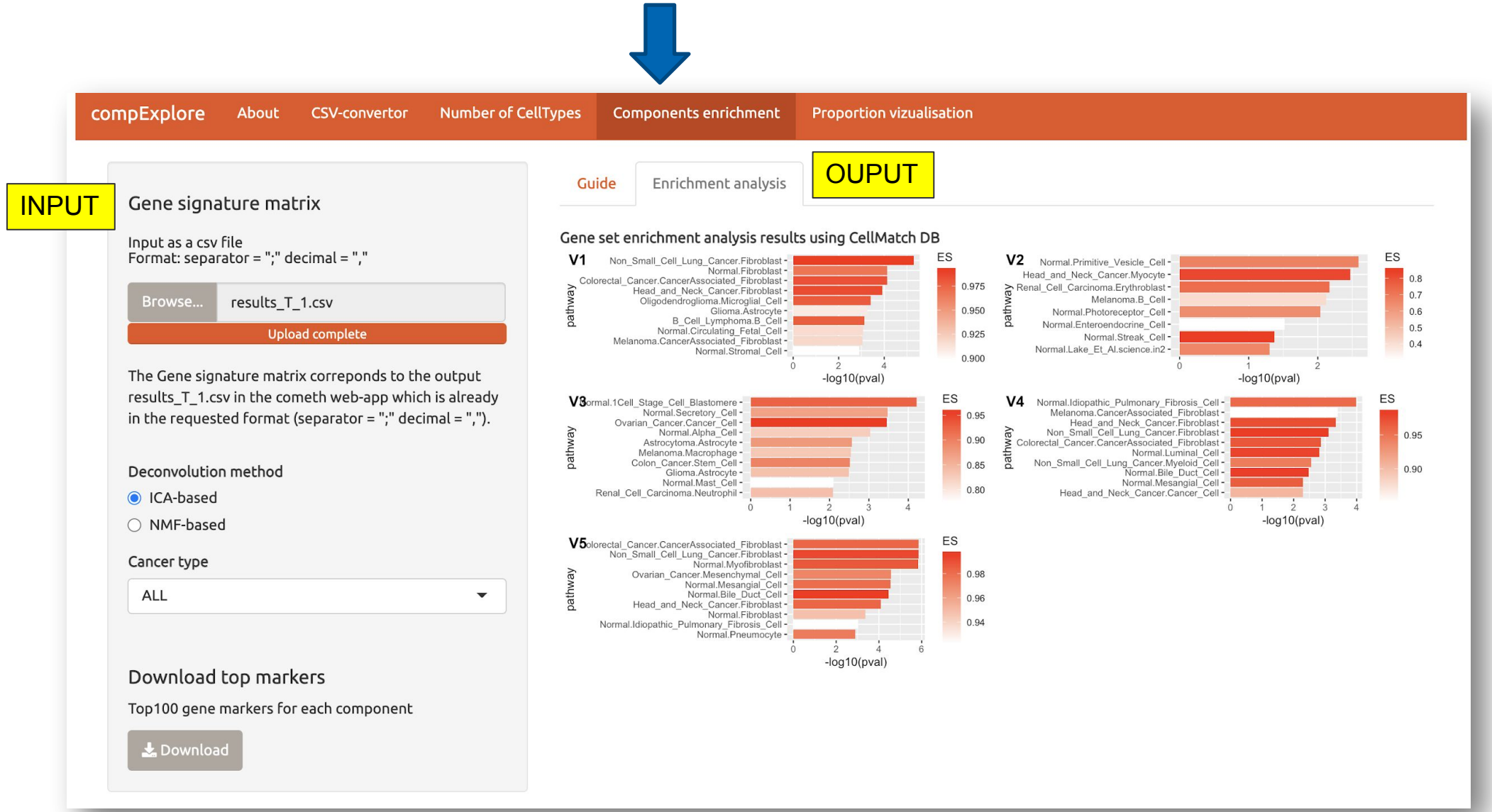
1- Order the list of genes (test statistic, p-value, here component scores...)

2- Calculation of the Enrichment Score (ES)

The algorithm scans the list: the score increases when the gene is part of the set (=cell type) and decreases otherwise. The increase and decrease values is weighted by the gene rank (for a gene set overexpressed, the increase will be higher at the beginning of the list).
The ES corresponds to the max score (absolute value).

3- Comparison of ES to a distribution of ES obtained on random data (gene permutations) . → Calculation of a p-value

compExplore Shiny app

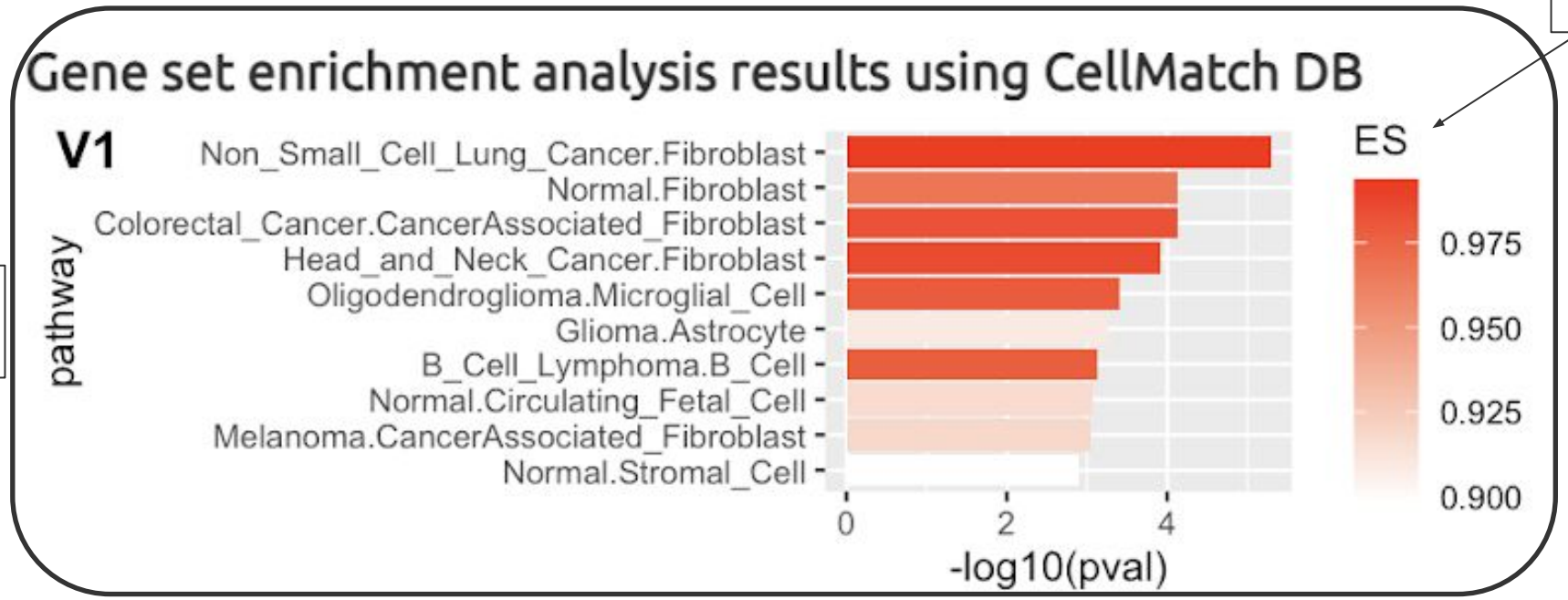


compExplore Shiny app

Example for component 1

(LUAD dataset, ICA method, k=5)

Enrichment Score



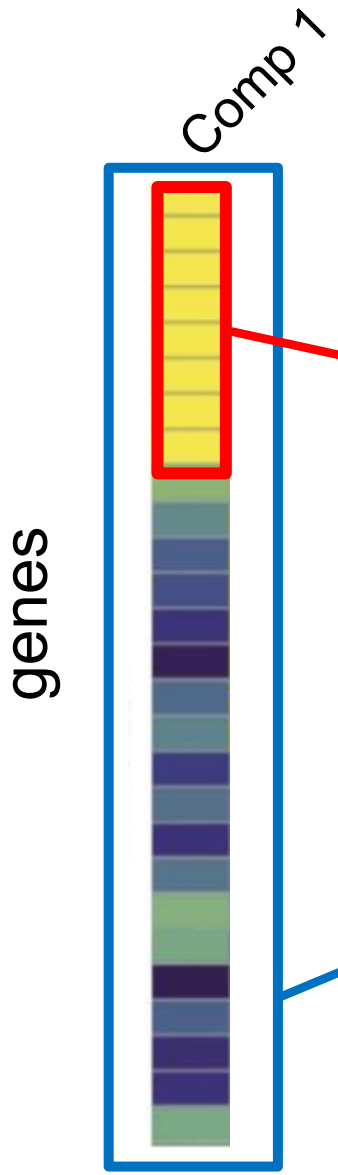
Cell types from the CellMatch DataBase

P-value of the GSEA test

--> Stromal component, Fibroblast?

Interpret the components identified

Other option



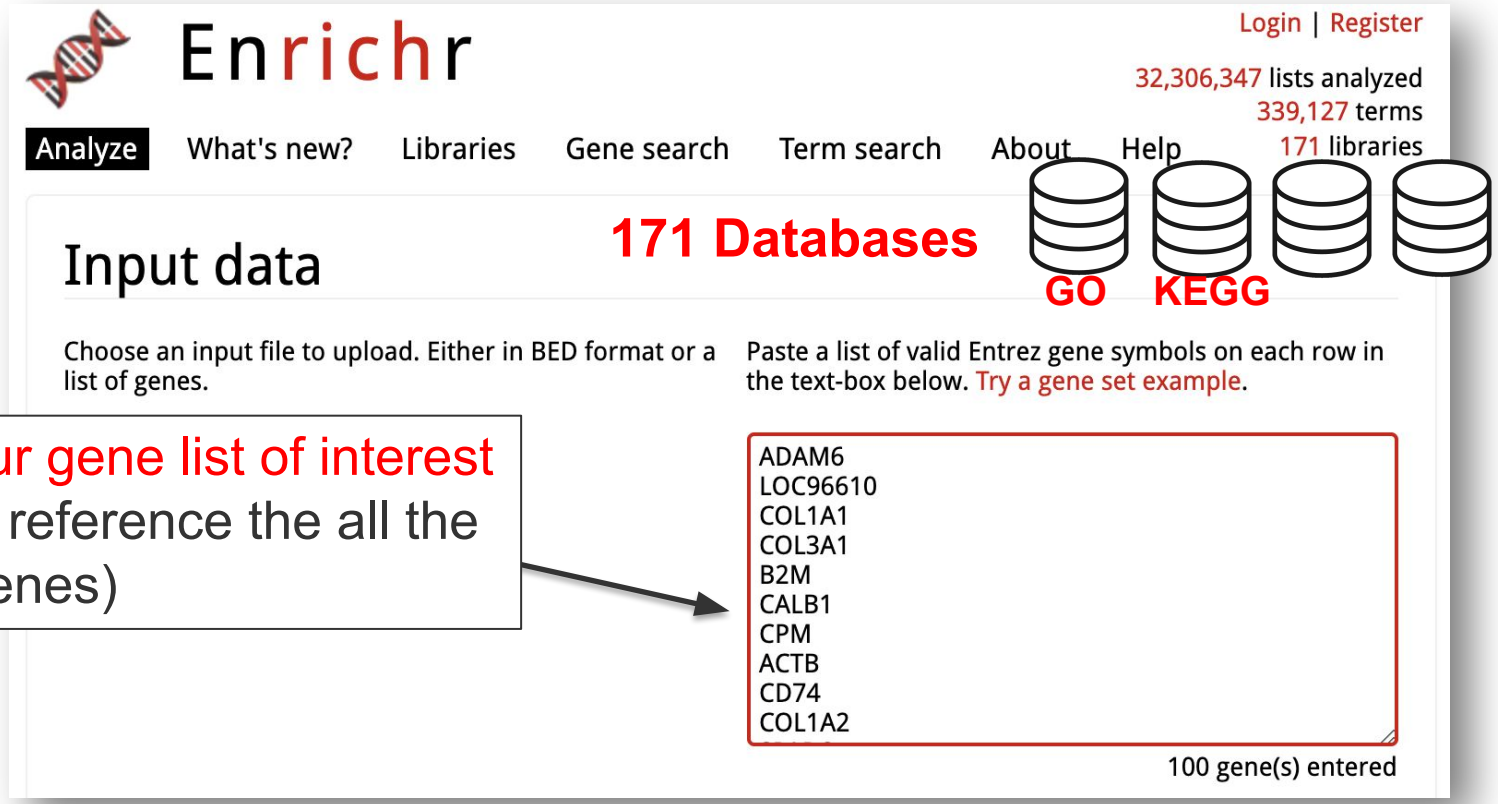
Perform an enrichment analysis
(Fisher's exact test)

Does the **gene list** contains **relatively more markers of a cell type / a pathway** than in the list of **all the genes (reference)**?

Unsupervised

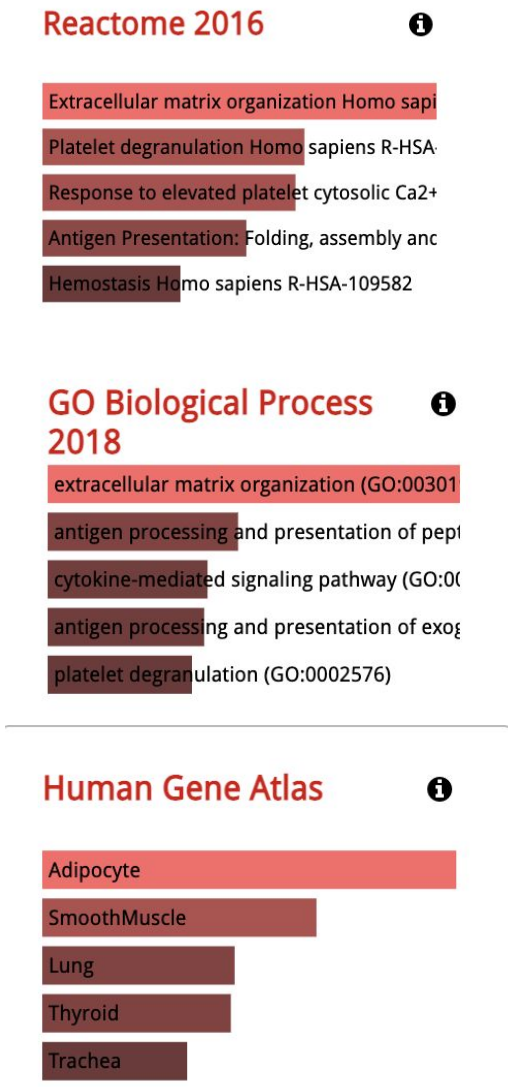
Interpret the components identified

Enrichment analysis with Enrichr



Paste your gene list of interest (takes as reference the all the human genes)

Ouput examples



Chen, E. Y., et al(2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, *14*(1), 128.

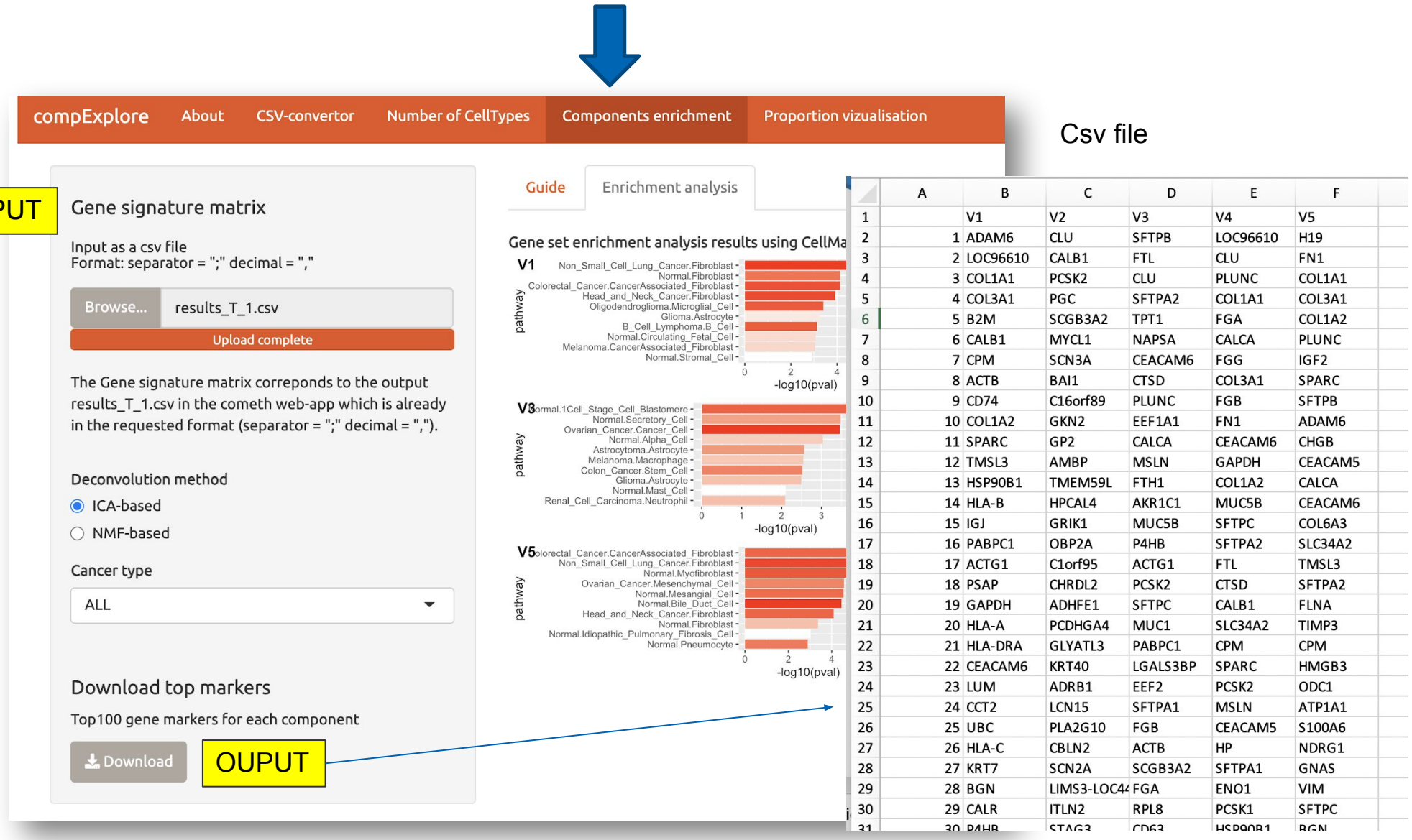
Unsupervised

Interpret the components identified

Perform enrichment analyses with other external tools

Feature/Tool	DAVID	Enrichr	ToppGene	g:profiler	clusterProfiler	Goplot	BACA	FunMappOne
KEGG pathways	✓	✓	✓	✓	✓		✓	✓
Reactome pathways	✓	✓	✓	✓	✓			✓
Gene Ontology	✓	✓	✓	✓	✓	✓	✓	✓
Graphic representation		✓	✓		✓	✓	✓	✓
Graphic user interface	✓	✓		✓				✓

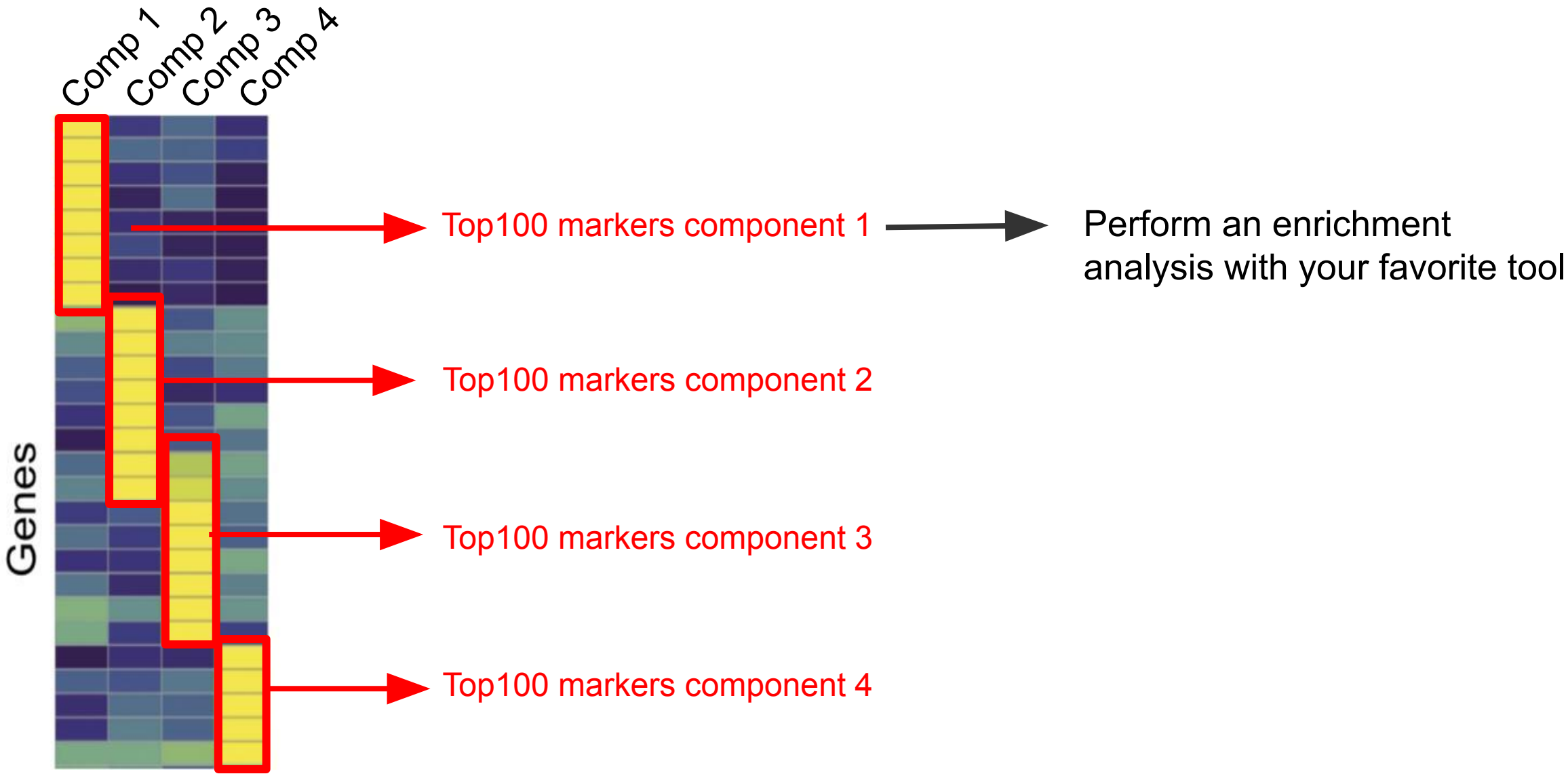
compExplore Shiny app



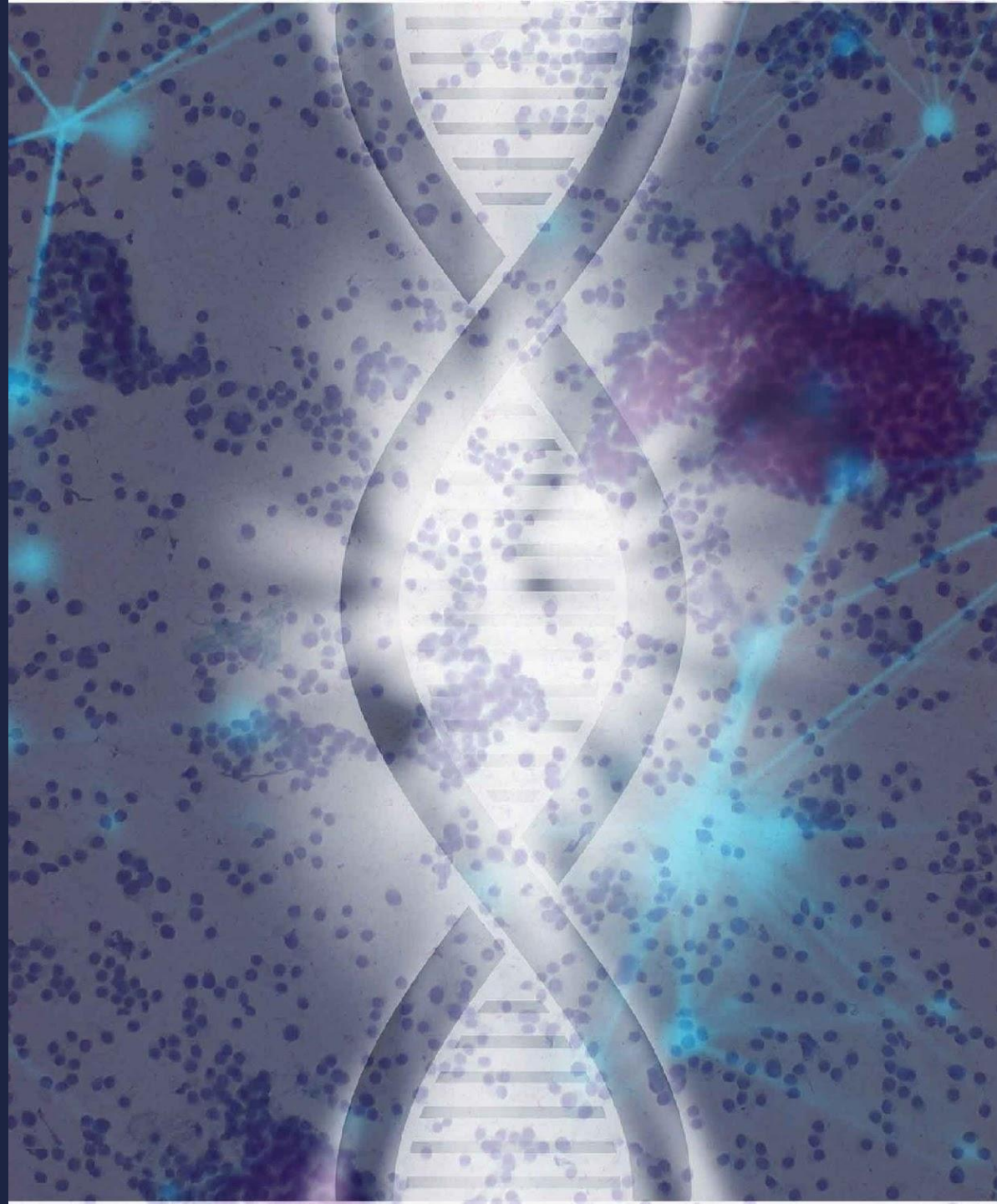
	A	B	C	D	E	F
1		V1	V2	V3	V4	V5
2	1	ADAM6	CLU	SFTPB	LOC96610	H19
3	2	LOC96610	CALB1	FTL	CLU	FN1
4	3	COL1A1	PCSK2	CLU	PLUNC	COL1A1
5	4	COL3A1	PGC	SFTPA2	COL1A1	COL3A1
6	5	B2M	SCGB3A2	TPT1	FGA	COL1A2
7	6	CALB1	MYCL1	NAPSA	CALCA	PLUNC
8	7	CPM	SCN3A	CEACAM6	FGG	IGF2
9	8	ACTB	BAI1	CTSD	COL3A1	SPARC
10	9	CD74	C16orf89	PLUNC	FGB	SFTPB
11	10	COL1A2	GKN2	EEF1A1	FN1	ADAM6
12	11	SPARC	GP2	CALCA	CEACAM6	CHGB
13	12	TMSL3	AMBP	MSLN	GAPDH	CEACAM5
14	13	HSP90B1	TMEM59L	FTH1	COL1A2	CALCA
15	14	HLA-B	HPCAL4	AKR1C1	MUC5B	CEACAM6
16	15	IGJ	GRIK1	MUC5B	SFTPC	COL6A3
17	16	PABPC1	OBP2A	P4HB	SFTPA2	SLC34A2
18	17	ACTG1	C1orf95	ACTG1	FTL	TMSL3
19	18	PSAP	CHRDL2	PCSK2	CTSD	SFTPA2
20	19	GAPDH	ADHFE1	SFTPC	CALB1	FLNA
21	20	HLA-A	PCDHGA4	MUC1	SLC34A2	TIMP3
22	21	HLA-DRA	GLYATL3	PABPC1	CPM	CPM
23	22	CEACAM6	KRT40	LGALS3BP	SPARC	HMGB3
24	23	LUM	ADRB1	EEF2	PCSK2	ODC1
25	24	CCT2	LCN15	SFTPA1	MSLN	ATP1A1
26	25	UBC	PLA2G10	FGB	CEACAM5	S100A6
27	26	HLA-C	CBLN2	ACTB	HP	NDRG1
28	27	KRT7	SCN2A	SCGB3A2	SFTPA1	GNAS
29	28	BGN	LIMS3-LOC44	FGA	ENO1	VIM
30	29	CALR	ITLN2	RPL8	PCSK1	SFTPC
31	30	P4HB	STAG3	CD63	HSP90B1	BGN

Unsupervised

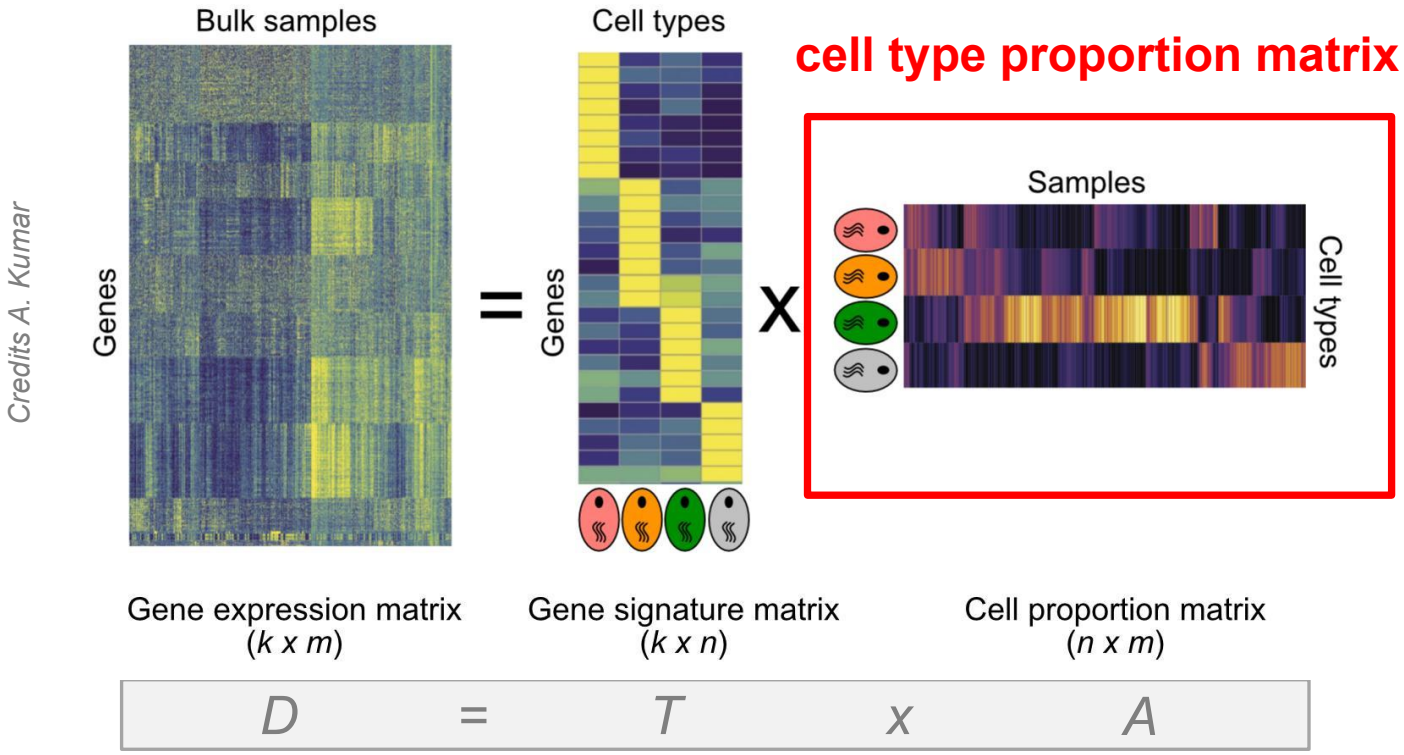
Interpret the components identified



Visualize the cell type proportion matrix



Visualize the cell type proportion matrix



- Is a given sample highly heterogeneous in its composition?
- Are all the different samples similar in their cell type composition ?

compExplore Shiny app

compExplore | About | CSV-convertor | Number of CellTypes | Components enrichment | Proportion vizualisation

INPUT

Cell proportion matrix

Input as a csv file
Format: separator = ";" decimal = ","

Browse... No file selected

The Gene proportion matrix correponds to the output results_A_1.csv in the cometh web-app which is already in the requested format (separator = ";" decimal = ",").

Select samples

No choices here yet !!

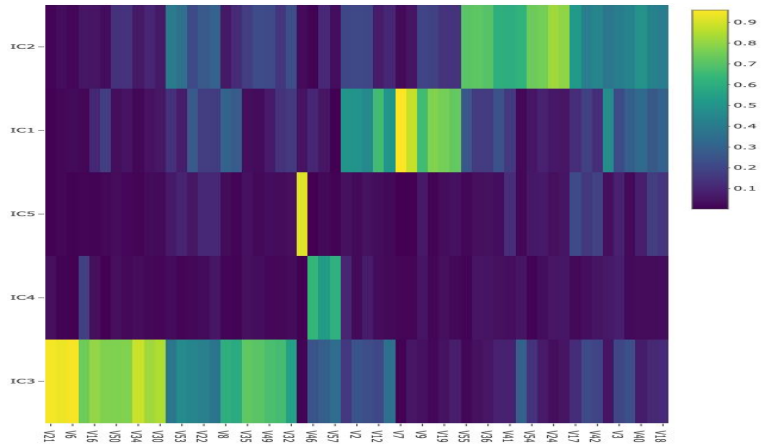
Select cell types

No choices here yet !!

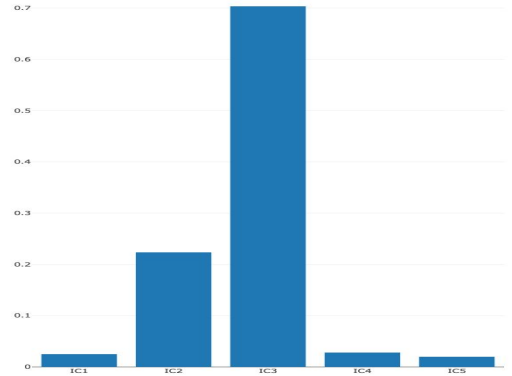
Guide | Vizualisation plots

OUPUT

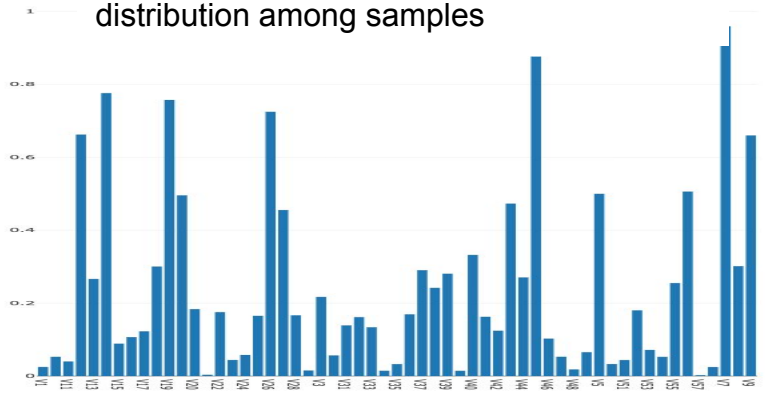
Heatmap of the proportion matrix



Focus on a selected sample:
Cell types abundance for this sample



Focus on a selected cell type:
distribution among samples



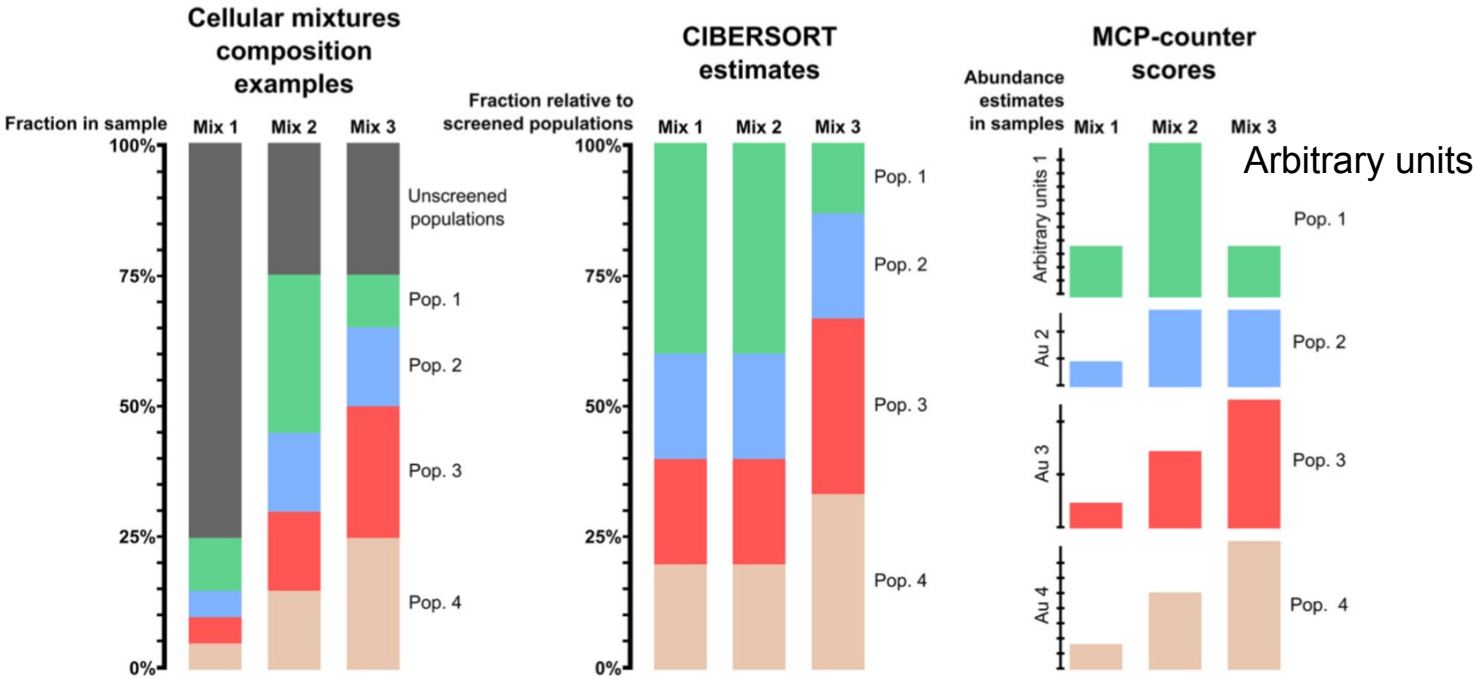
Visualize the cell type proportion matrix

(1) **CIBERSORT-ABS, EPIC and quanTIseq** can b used for both **inter- and intra- sample comparisons** i.e. comparing one cell-type withir one sample and across samples is possible

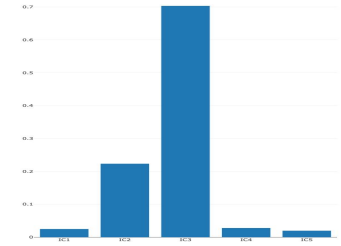
(2) **CIBERSORT** can be used only for **intra-sample comparisons** i.e. comparing different cell-types within each sample

(3) **MCP-Counter, TIMER and xCell (not provided yet in the cometh web app)** can be used only for **inter- sample comparisons** i.e. to compare one cell-type across multiple samples

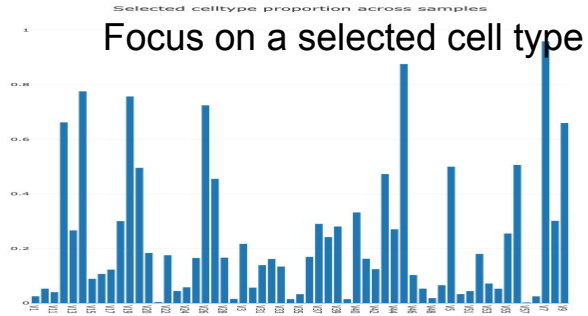
Petitprez et al., 2018 Cancer Immunol Immunother



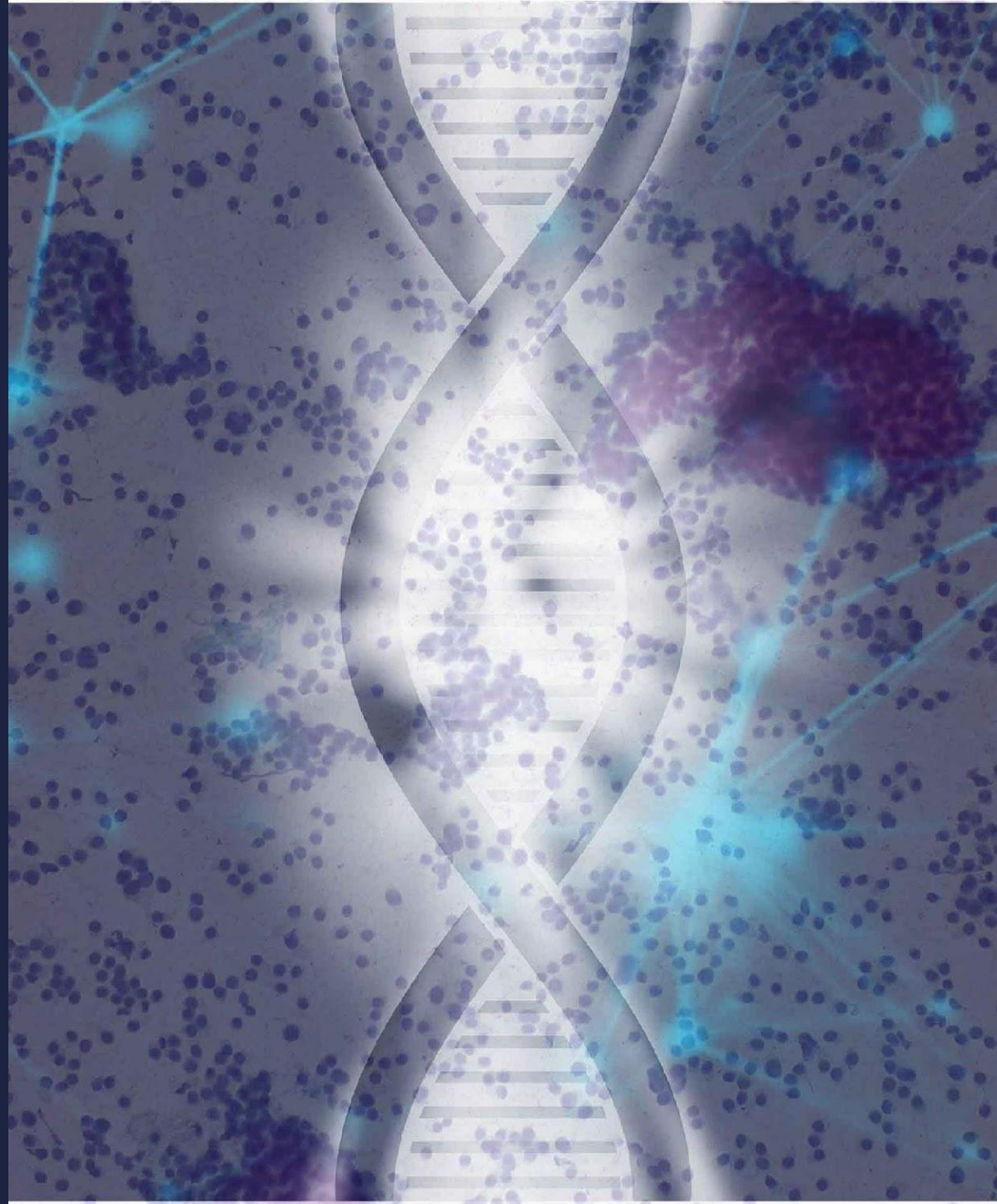
Focus on a selected sample



Focus on a selected cell type



Go further in biological interpretation



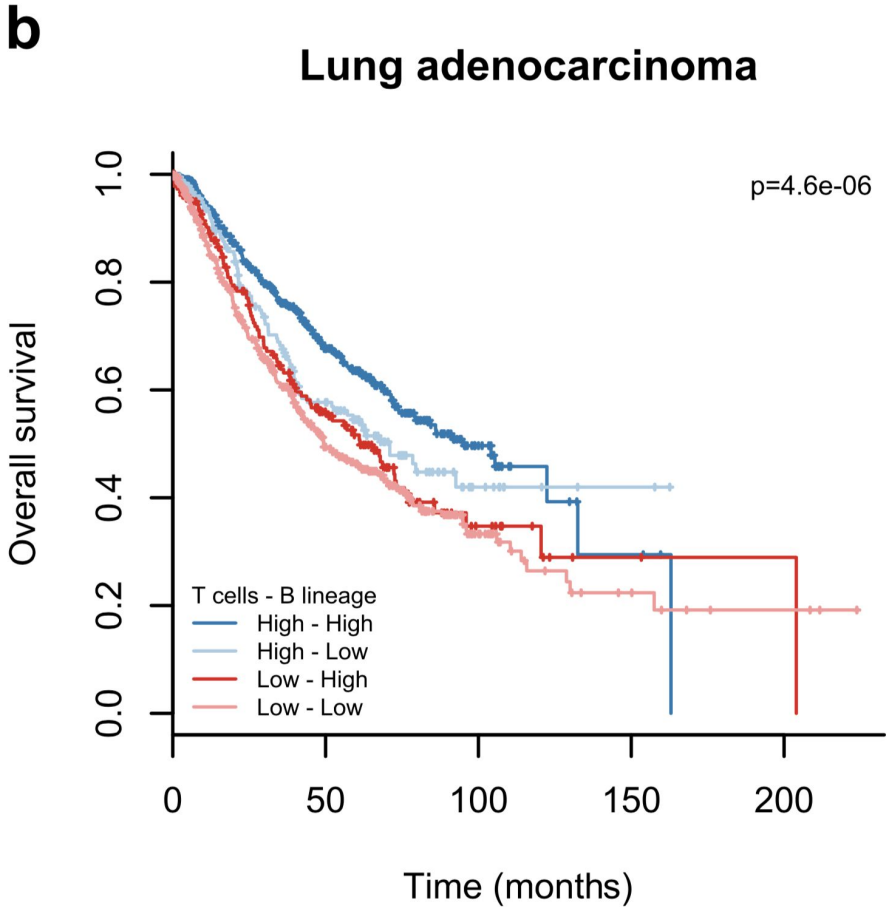
Relate cell type proportions to clinical annotations

Example of the prognosis

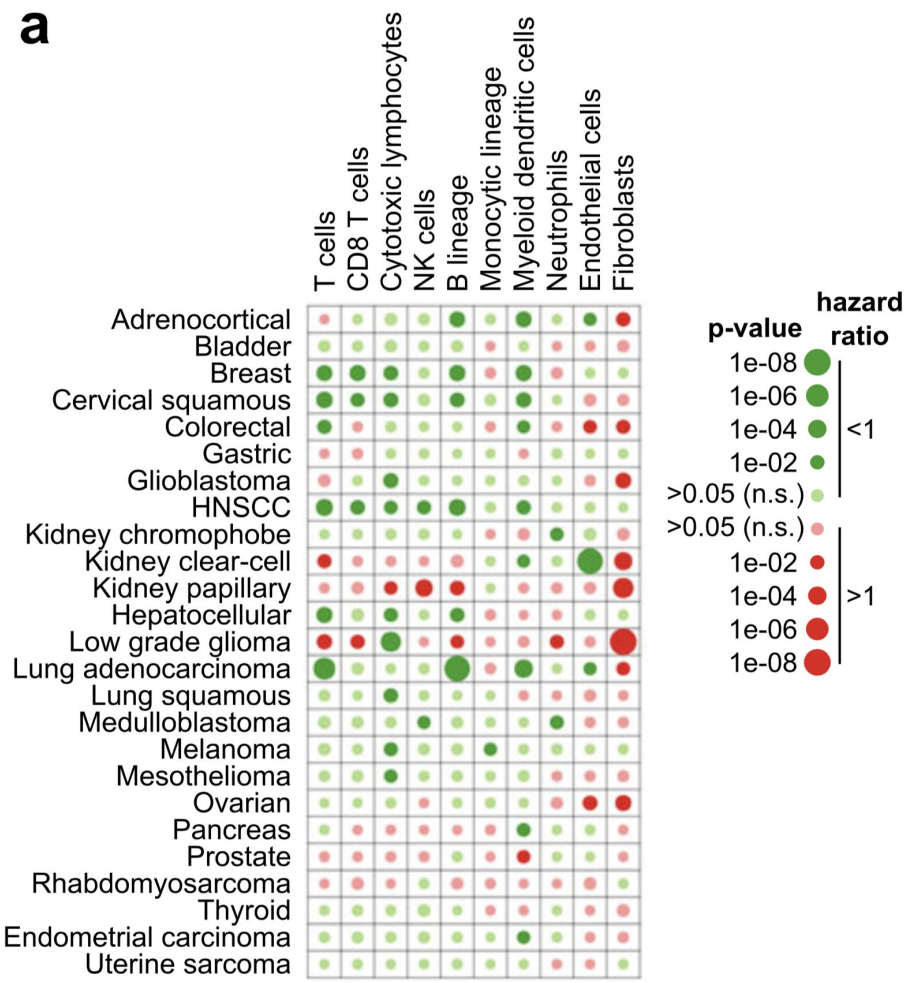
Tools

cBioPortal for Cancer Genomics

R

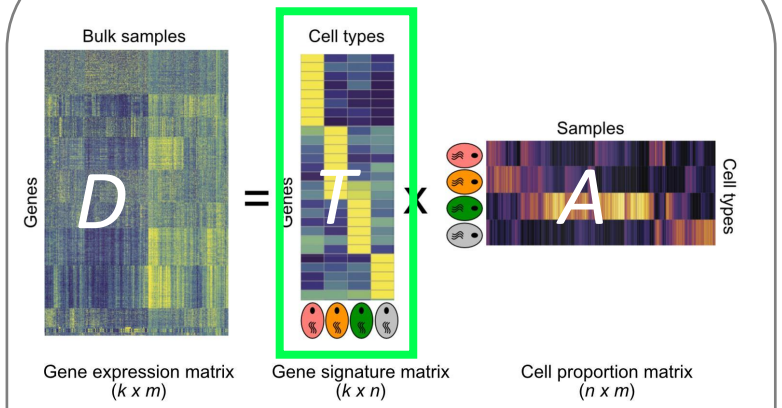


Becht et al 2016, Genome Biology



Pay attention to…

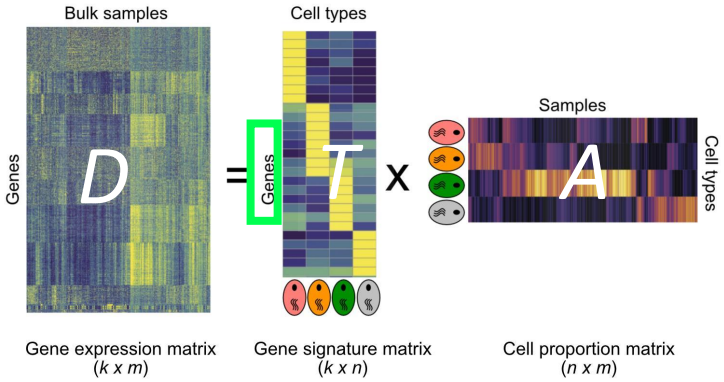
Supervised



T matrix known

- Are the cell type profiles reliable?
- Are the cell type profiles appropriate regarding the cancer types/ tissue you are looking at?

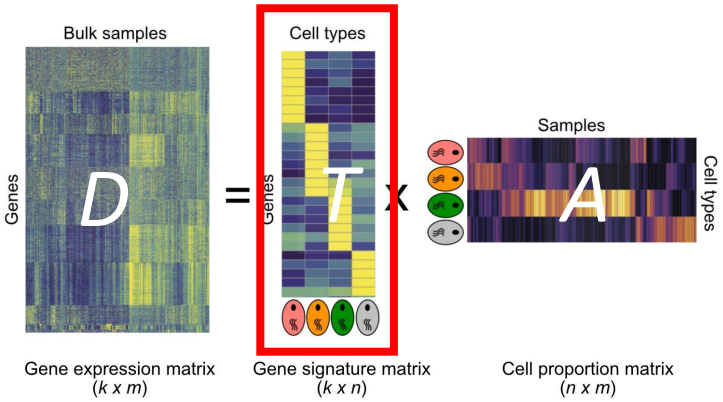
Semi-supervised



Gene markers known

- Are the gene markers reliable/robust ?
- Are gene markers appropriate regarding the cancer type/ tissue you are looking at?

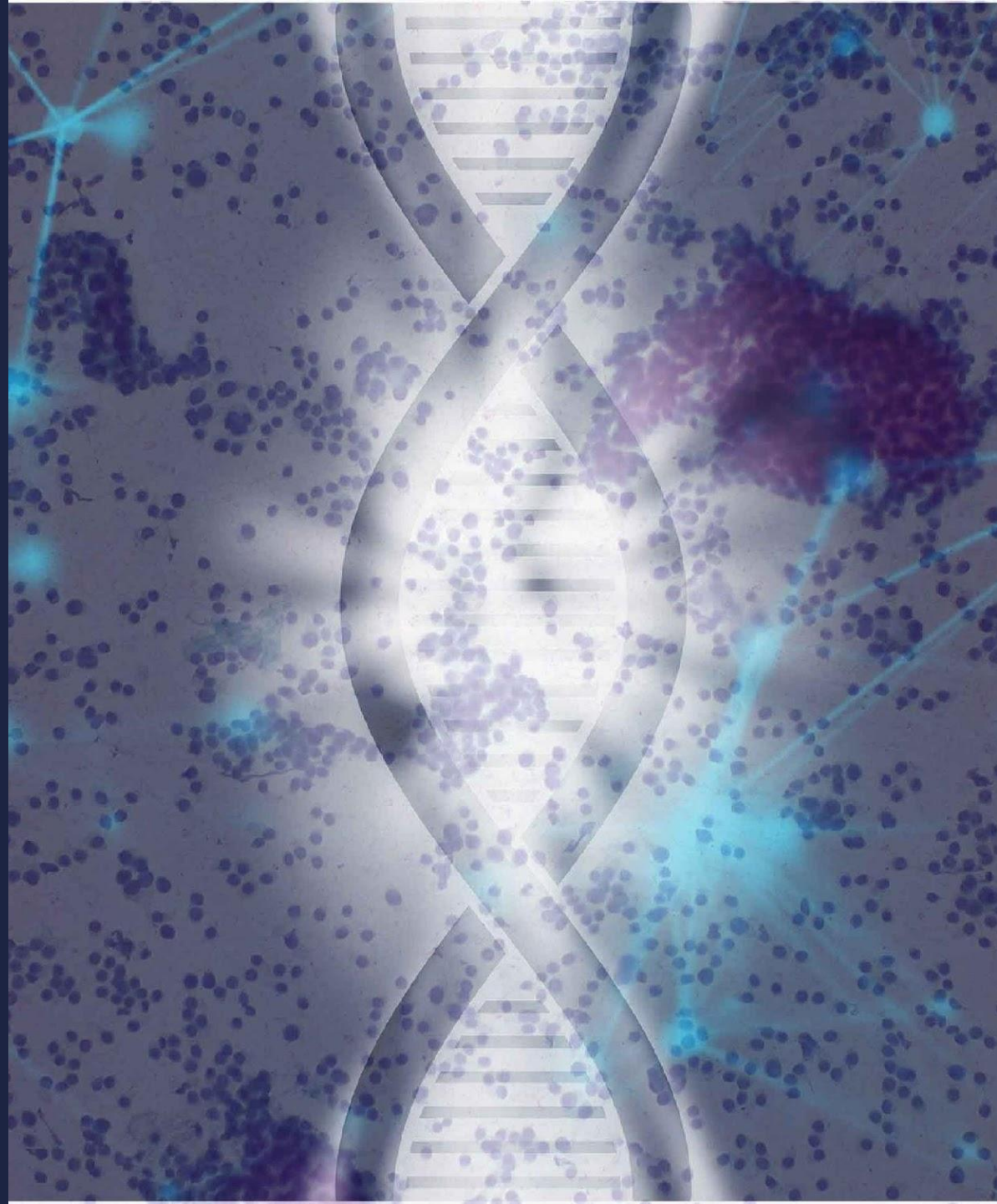
Unsupervised



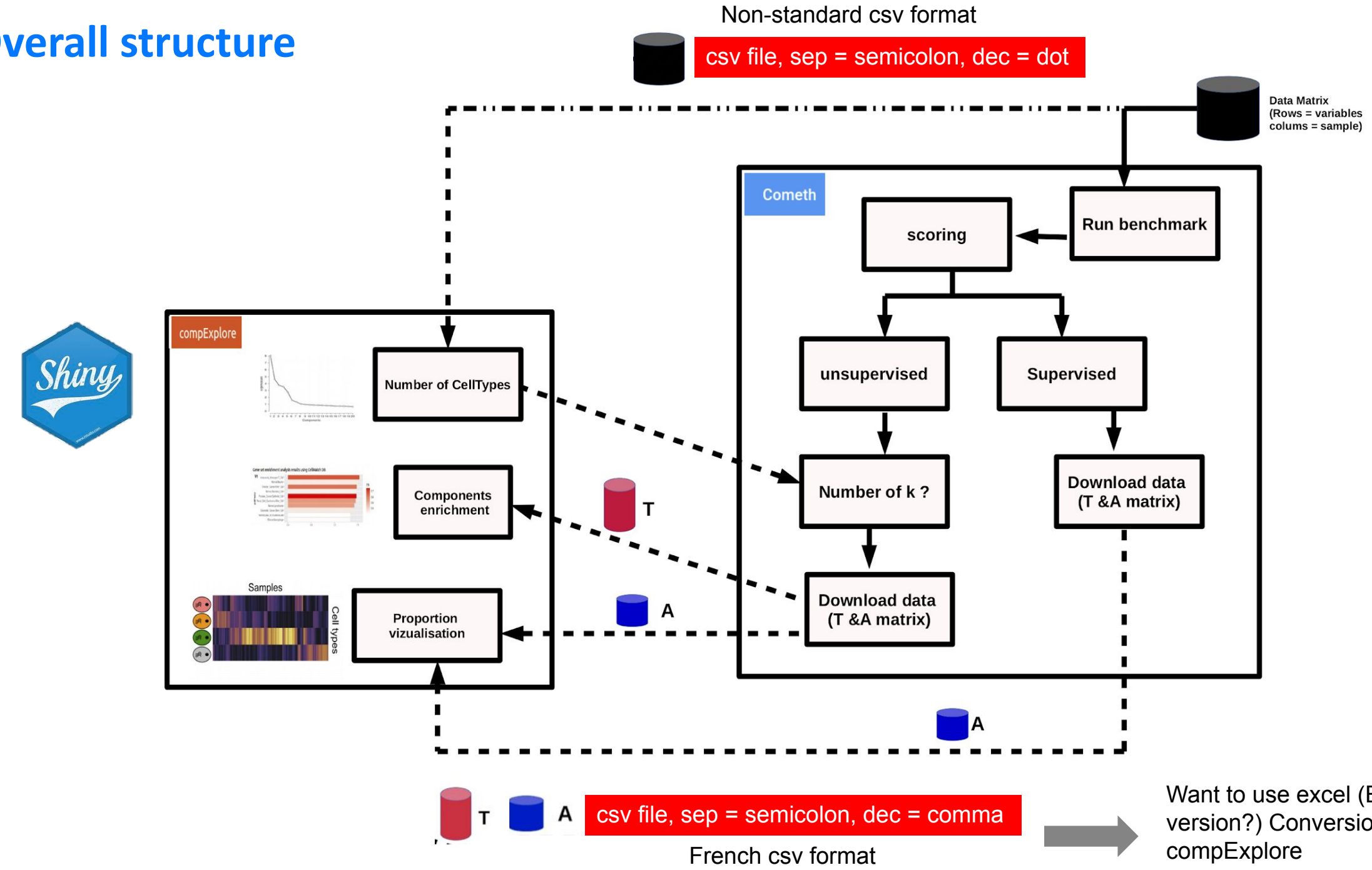
T matrix unknown

When choosing k, did I choose a good granularity ?
- Seem components be a mix between several cell types?
- Are there several components corresponding the same signal?

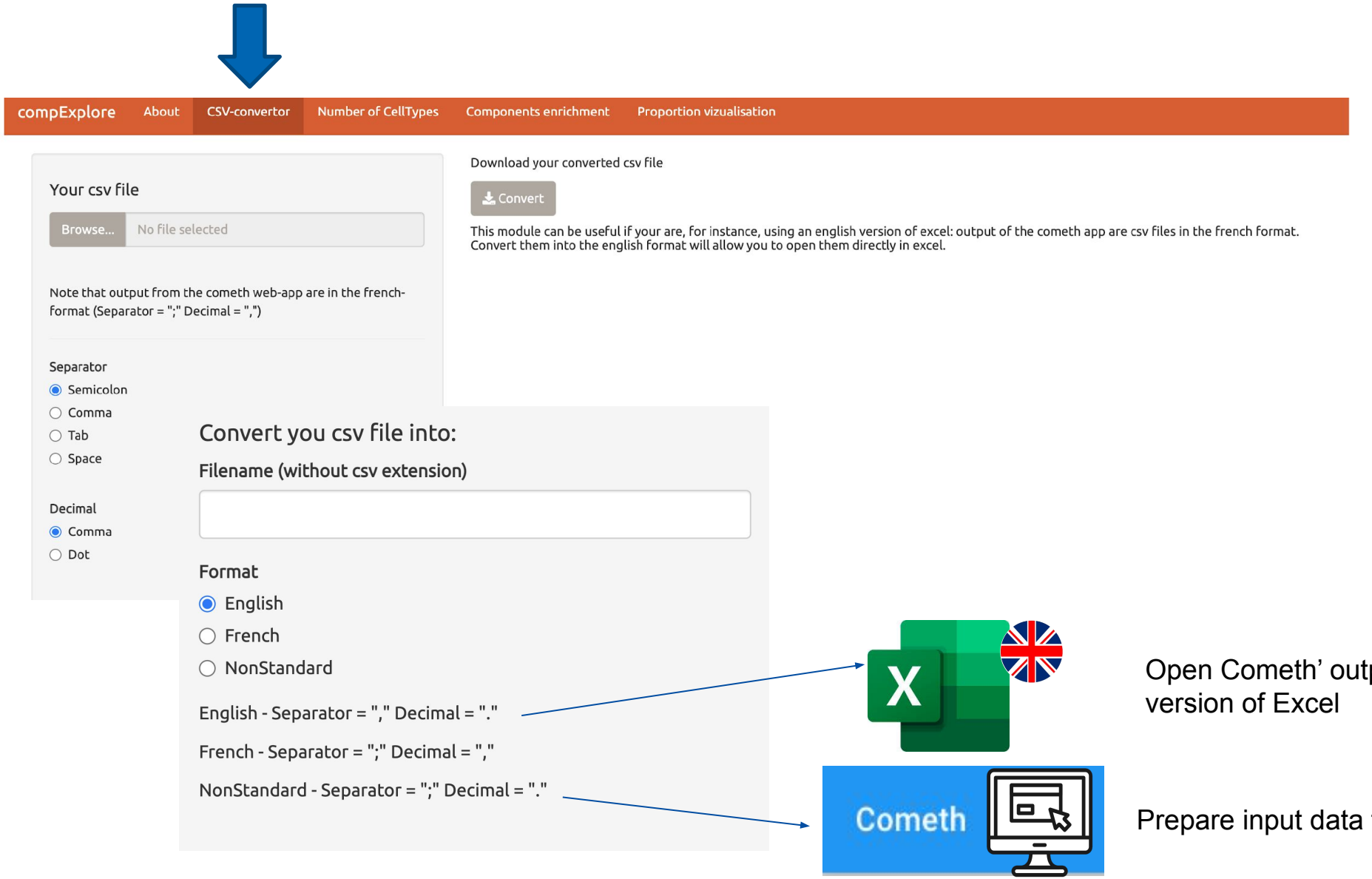
Overall structure input/output format



Overall structure



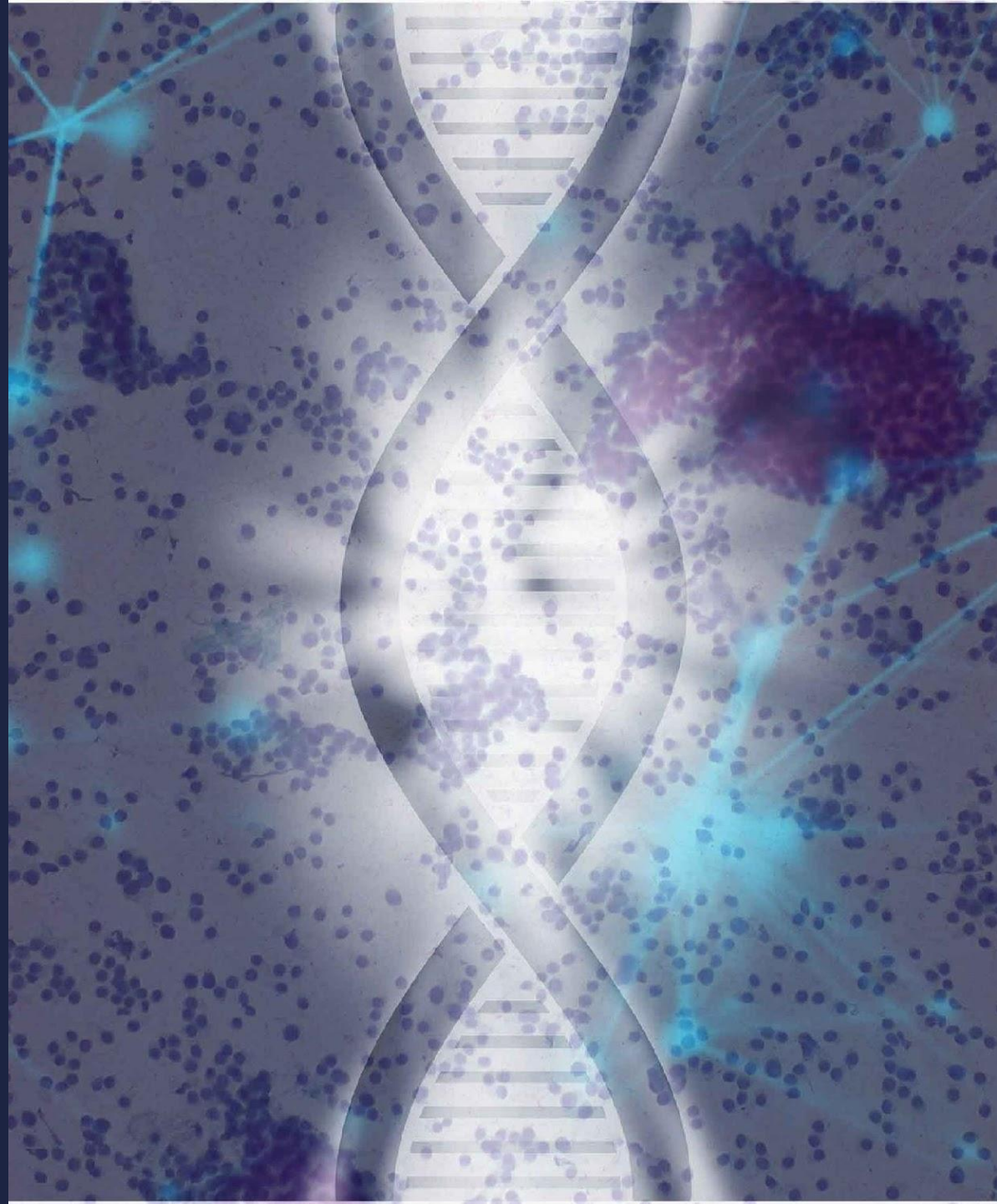
compExplore Shiny app



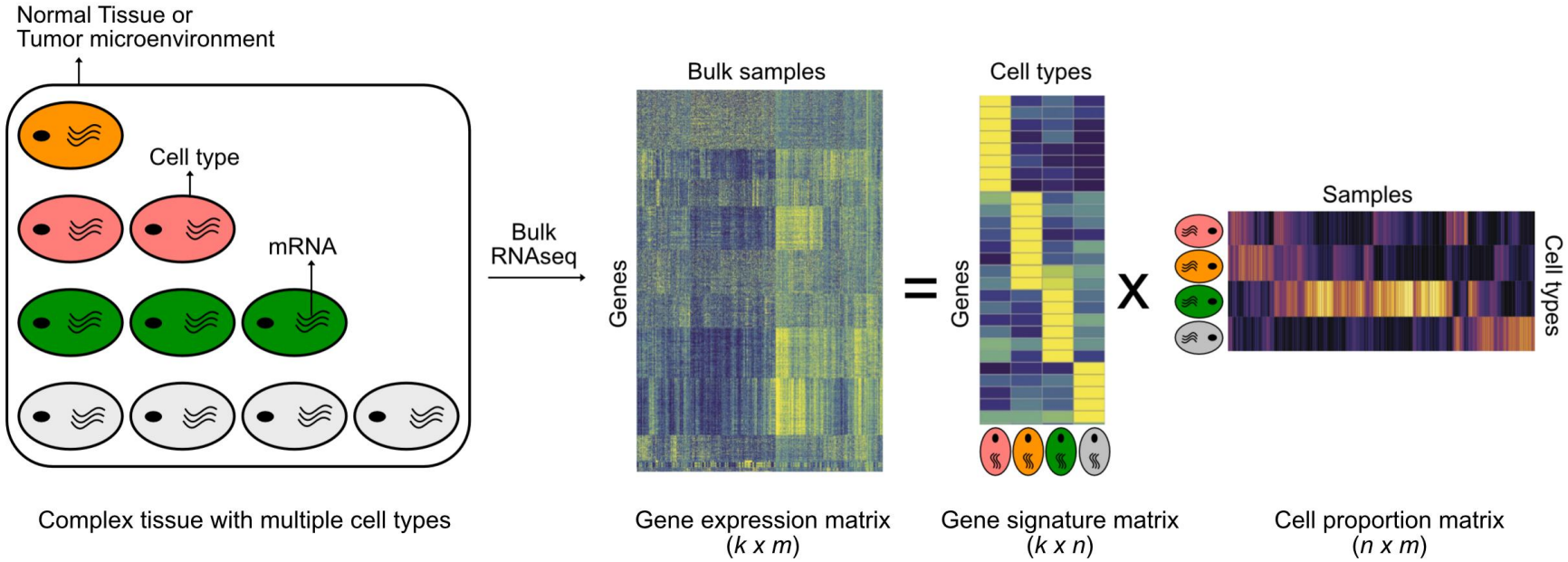
Open Cometh' outputs in english version of Excel

Prepare input data for Cometh web-app

Examples of success stories



Resolving cell types from complex tissue genomic data : RECAP



- **DeconRNASeq** (*~160* | Apr, 2013 | https://doi.org/10.1093/bioinformatics/btt090)
- **CellMix** (*~180* | Sep, 2013 | https://doi.org/10.1093/bioinformatics/btt351)
- **CIBERSORT** (*~2000* | Mar, 2015 | https://doi.org/10.1038/nmeth.3337)
- **MCP-Counter** (*~350* | Oct, 2016 | https://doi.org/10.1186/s13059-016-1070-5)
- **TIMER2.0** (*~500* | Jul, 2017 | https://doi.org/10.1093/nar/gkaa407)
- **Xcell** (*~400* | Nov, 2017 | https://doi.org/10.1186/s13059-017-1349-1)
- **EPIC** (*~100* | Nov, 2017 | https://doi.org/10.7554/eLife.26476)
- **QuantiSeq** (*~50* | May, 2019 | https://doi.org/10.1186/s13073-019-0638-6)

Table 1. Overview of cell type quantification methods providing gene signatures for immuno-oncology

Tool	Abbrev.	Type	Score	Comparisons	Algorithm	Cell types	Reference
CIBERSORT	CBS	D	Immune cell fractions, relative to total immune cell content	Intra	ν-support vector regression	22 immune cell types	Newman *et al.* (2015)
CIBERSORT abs. mode	CBA	D	Score of arbitrary units that reflects the absolute proportion of each cell type	Intra, inter	ν-support vector regression	22 immune cell types	Newman *et al.* (2015, 2018)
EPIC	EPC	D	Cell fractions, relative to all cells in sample	Intra, inter	constrained least square regression	6 immune cell types, fibroblasts, endothelial cells	Racle *et al.* (2017)
MCP-counter	MCP	M	Arbitrary units, comparable between samples	Inter	mean of marker gene expression	8 immune cell types, fibroblasts, endothelial cells	Becht *et al.* (2016)
quanTIseq	QTS	D	Cell fractions, relative to all cells in sample	Intra, inter	constrained least square regression	10 immune cell types	Finotello *et al.* (2017)
TIMER	TMR	D	Arbitrary units, comparable between samples (not different cancer types)	Inter	linear least square regression	6 immune cell types	Li *et al.* (2016)
xCell	XCL	M	Arbitrary units, comparable between samples	Inter	ssGSEA (Hänzelmann *et al.*, 2013)	64 immune and non-immune cell types	Aran *et al.* (2017)

Note: Methods can be conceptually distinguished in marker-gene-based approaches (M) and deconvolution-based approaches (D). The output scores of the methods have different properties and allow either intra-sample comparisons between cell types, inter-sample comparisons of the same cell type, or both. All methods come with a set of cell type signatures ranging from six immune cell types to 64 immune and non-immune cell types.

https://doi.org/10.1093/bioinformatics/btz363 | *Gregor Strum et.al.*

immunedeconv

Table 2. Guidelines for method selection

Cell type	Recommended methods	Overall performance	Absolute score	No background predictions
B cell	EPIC	++	++	+
	MCP-counter	++	−	−
T cell CD4+	EPIC	++	++	−
	xCell	++	−	++
T cell CD4+ non-regulatory	quanTIseq	+	++	+
	xCell	+	−	++
T cell regulatory	quanTIseq	++	++	−
	xCell	++	−	++
T cell CD8+	quanTIseq	++	++	−
	EPIC	++	++	−
	MCP-counter	++	−	−
	xCell	+	−	++
Natural Killer Cell	EPIC	++	++	+
	MCP-counter	++	−	−
Macrophage / Monocyte	xCell	−	++	
	EPIC	+	++	+
	MCP-counter	++	−	−
Cancer-associated fibroblast	EPIC	++	++	+
	MCP-counter	++	−	−
Endothelial Cell	EPIC	++	++	+
	xCell	++	−	++
Dentricic cell	None of the methods can be recommended to estimate overall DC content. MCP-counter and quanTIseq can be used to profile mDCs.			

https://doi.org/10.1093/bioinformatics/btz363 | *Gregor Strum et.al.*

immunedeconv

Article | Published: 20 March 2019

Neoantigen-directed immune escape in lung cancer evolution

Rachel Rosenthal, Elizabeth Larose Cadieux, Roberto Salgado, Maise Al Bakir, David A. Moore, Crispin T. Hiley, Tom Lund, Miljana Tanić, James L. Reading, Kroopa Joshi, Jake Y. Henry, Ehsan Ghorani, Gareth A. Wilson, Nicolai J. Birkbak, Mariam Jamal-Hanjani, Selvaraju Veeriah, Zoltan Szallasi, Sherene Loi, Matthew D. Hellmann, Andrew Feber, Benny Chain, Javier Herrero, Sergio A. Quezada, Jonas Demeulemeester, Peter Van Loo, Stephan Beck, Nicholas McGranahan ✉, Charles Swanton ✉ & The TRACERx consortium -Show fewer authors

Nature **567**, 479–485(2019) | Cite this article

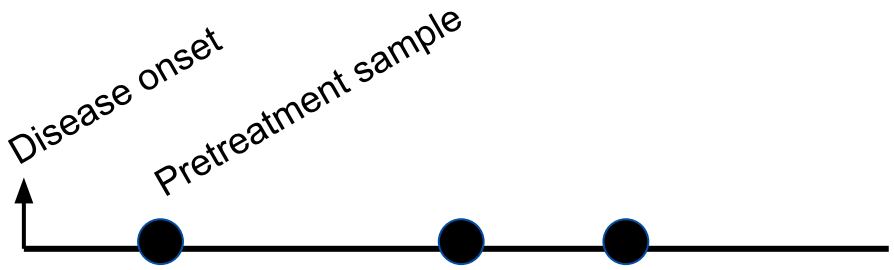
Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, University College London, London, UK

Cancer Genome Evolution Research Group, University College London Cancer Institute, University College London, London, UK

https://doi.org/10.1038/s41586-019-1032-7

nature

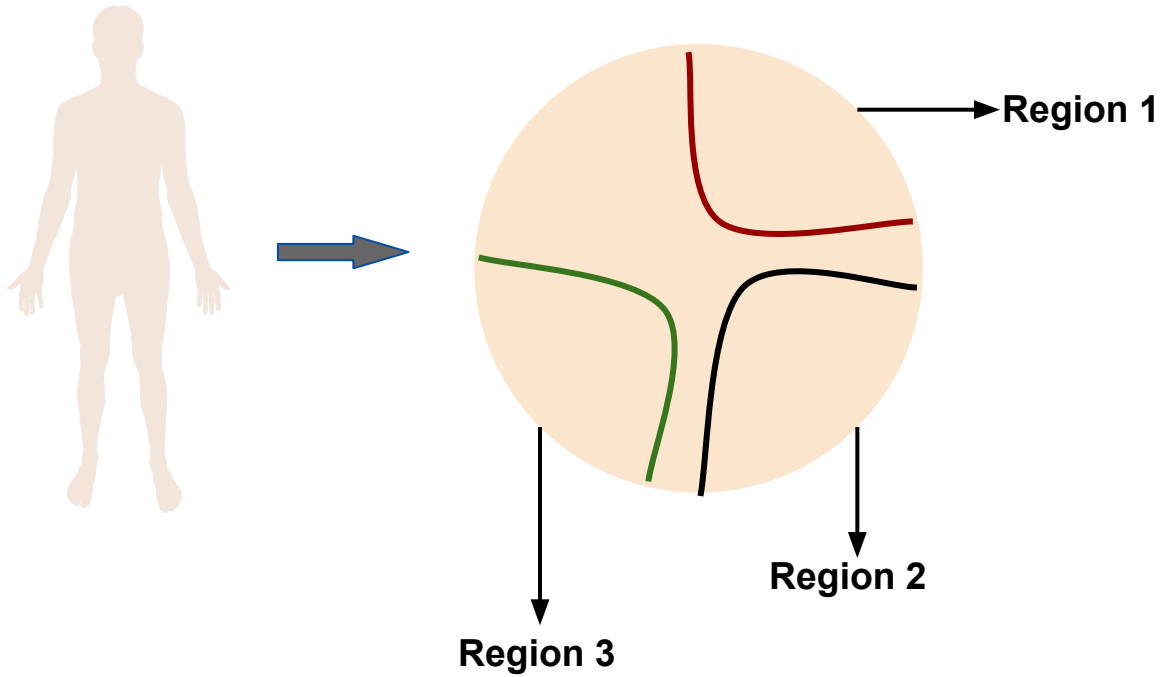
TRACERx 100 cohort

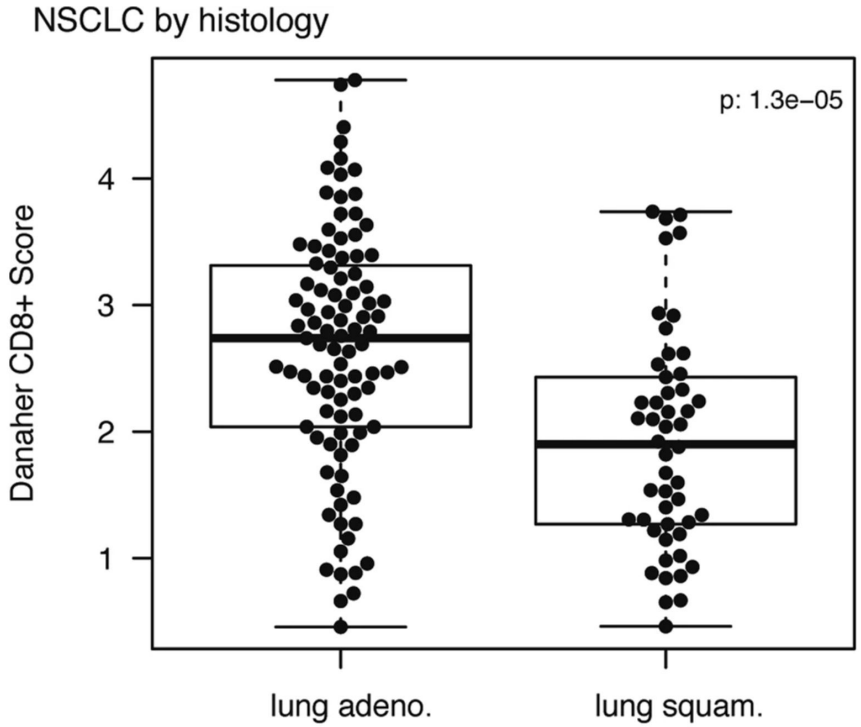


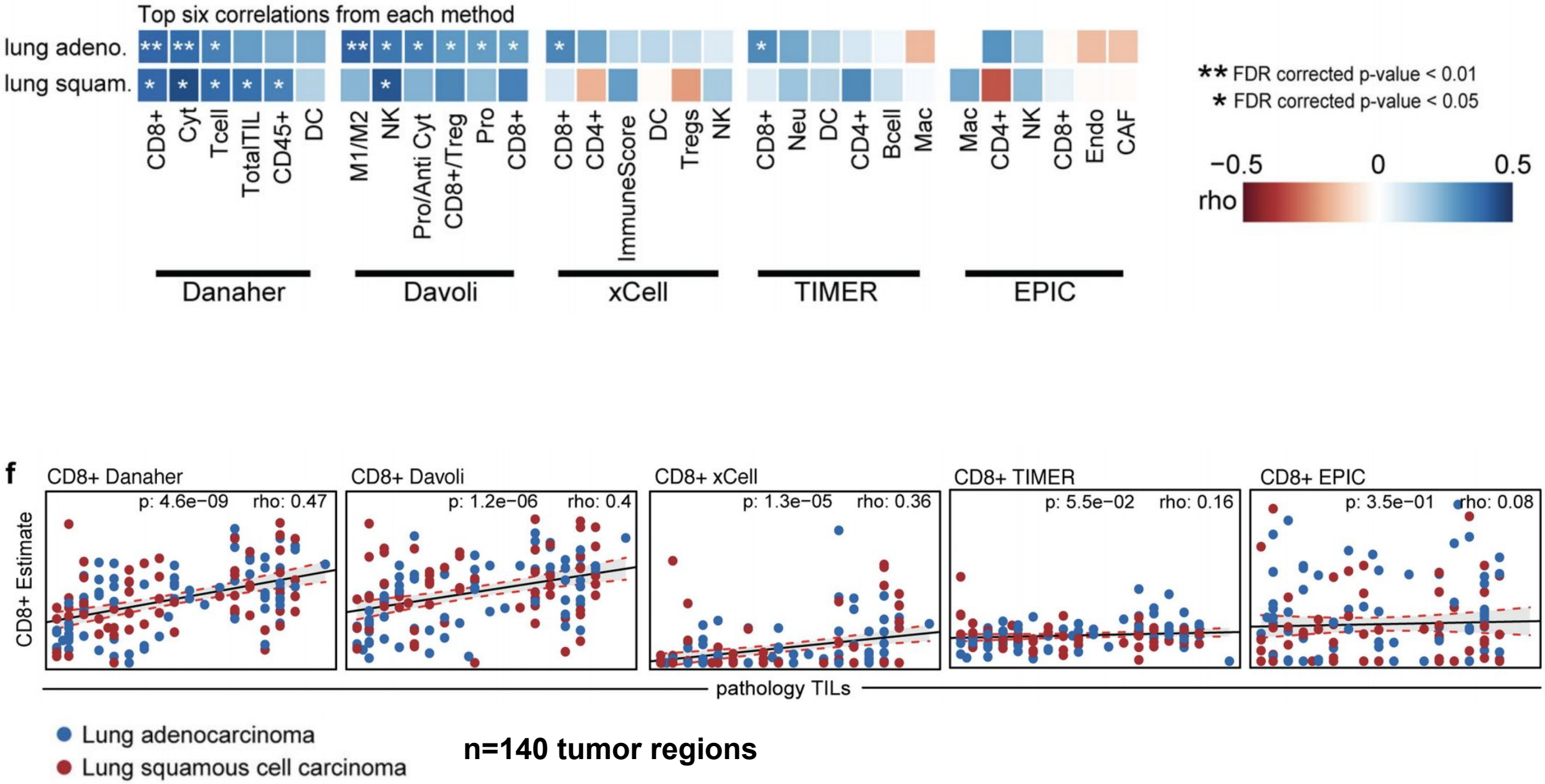
164 RNA-seq samples from **64 non-small-cell lung cancer** (NSCLC)

Tumour-infiltrating lymphocyte (TIL) histopathology estimates **(*n*=234)** from **83 NSCLC**

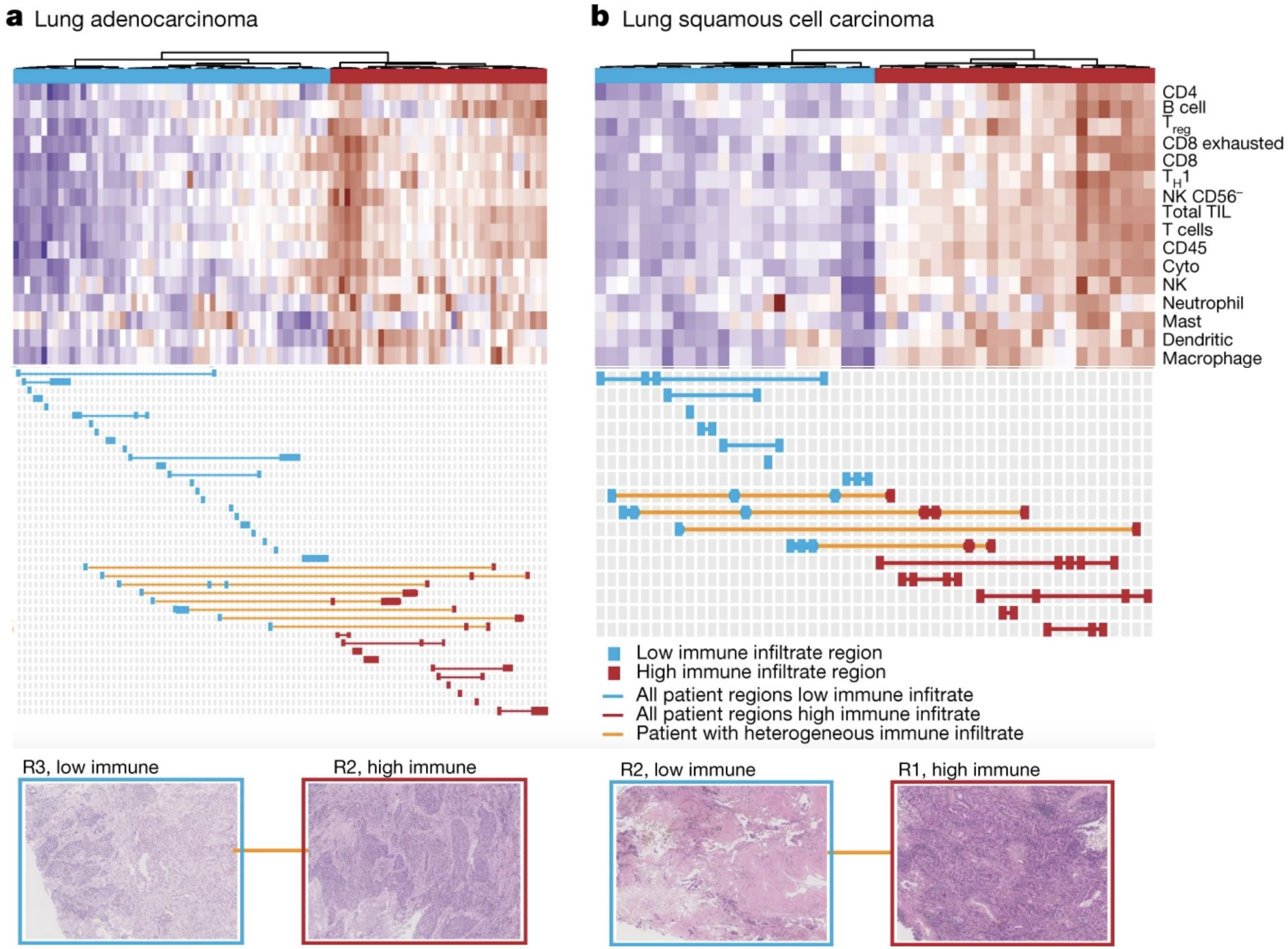
~258 tumor regions from **88** patients (TRACERx 100 cohort)

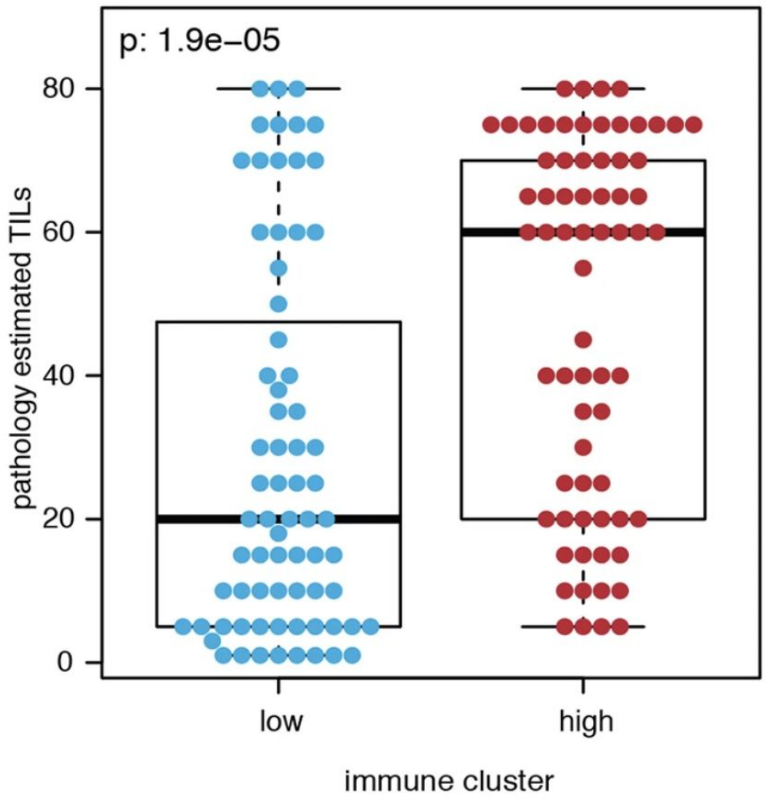


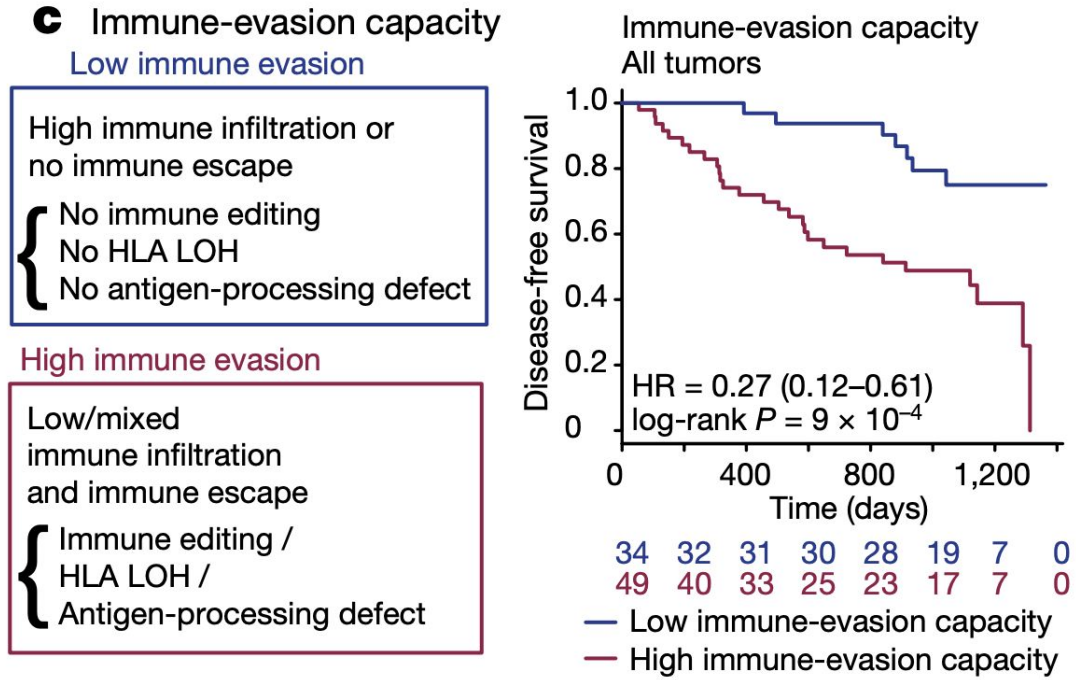




n=140 tumor regions


a Lung adenocarcinoma
b Lung squamous cell carcinoma
CD4
B cell
T reg
CD8 exhausted
CD8
T H 1
NK CD56–
Total TIL
T cells
CD45
Cyto
NK
Neutrophil
Mast
Dendritic
Macrophage
Low immune infiltrate region
High immune infiltrate region
All patient regions low immune infitrate
All patient regions high immune infitrate
Patient with heterogeneous immune infiltrate
R3, low immune
R2, high immune
R2, low immune
R1, high immune





Computational methods of supervised immune cell type enumeration can identify clinically relevant biology

Article | Published: 07 December 2020

Super enhancers define regulatory subtypes and cell identity in neuroblastoma

Moritz Gartlgruber, Ashwini Kumar Sharma, Andrés Quintero, Daniel Dreidax, Selina Jansky, Young-Gyu Park, Sina Kreth, Johanna Meder, Daria Doncevic, Paul Saary, Umut H. Toprak, Naveed Ishaque, Elena Afanasyeva, Elisa Wecht, Jan Koster, Rogier Versteeg, Thomas G. P. Grünewald, David T. W. Jones, Stefan M. Pfister, Kai-Oliver Henrich, Johan van Nes, Carl Herrmann ✉ & Frank Westermann ✉

Nature Cancer **2**, 114–128(2021) | Cite this article

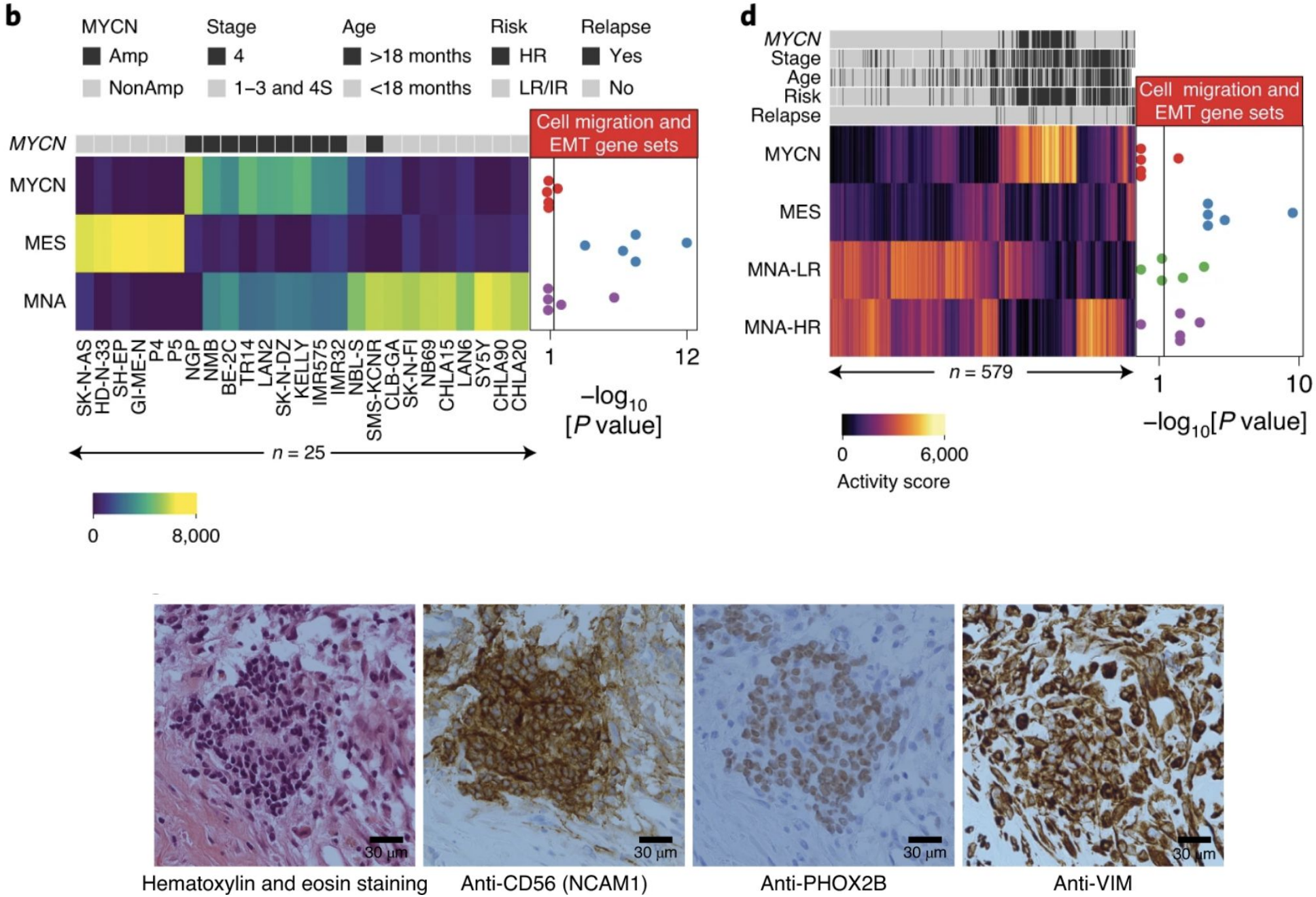
Health Data Science Unit, Medical Faculty Heidelberg and BioQuant, Heidelberg, Germany

Hopp Children's Cancer Center Heidelberg (KiTZ), Heidelberg, Germany

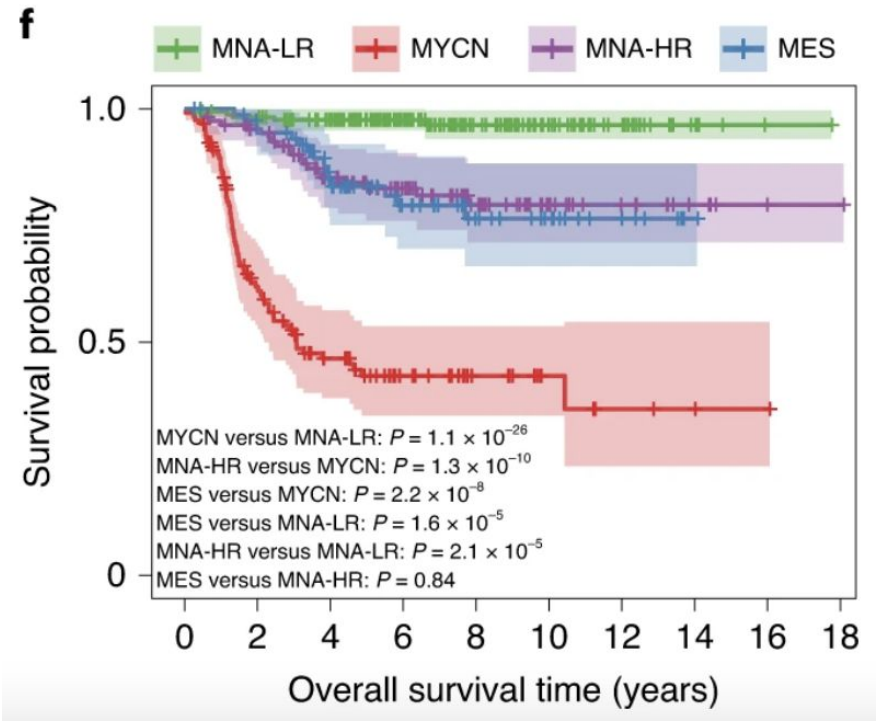
Division of Neuroblastoma Genomics, German Cancer Research Center, Heidelberg, Germany

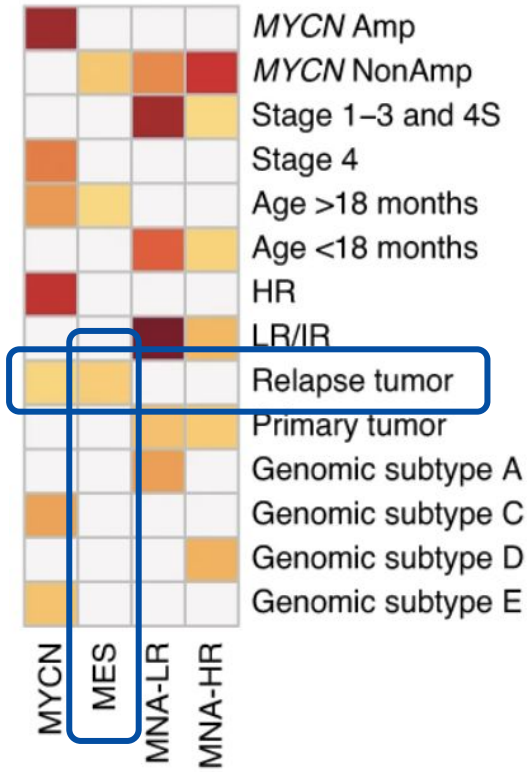
https://doi.org/10.1038/s43018-020-00145-w

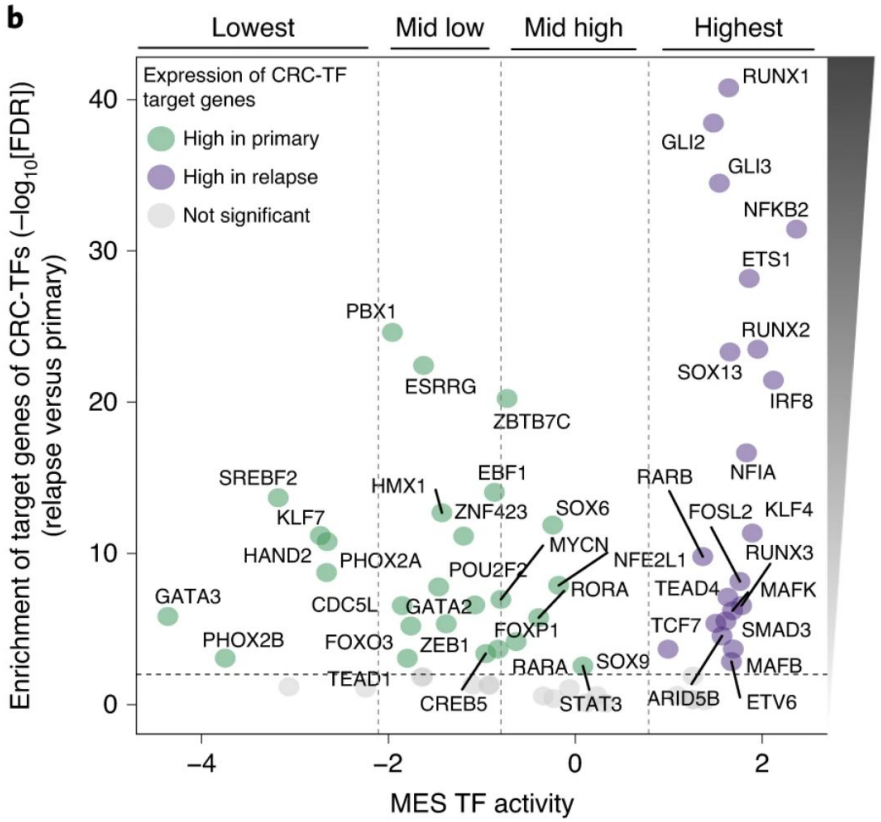
nature cancer



Hematoxylin and eosin staining | Anti-CD56 (NCAM1) | Anti-PHOX2B | Anti-VIM







Computational methods of unsupervised cell type enumeration can identify clinically and biologically relevant disease subtypes

Single cell guided deconvolution

CIBERSORTx (**CSx**)

Determining cell type abundance and expression from bulk tissues with digital cytometry

Aaron M. Newman ✉, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn & Ash A. Alizadeh ✉

Nature Biotechnology **37**, 773–782(2019) | Cite this article

Cell Population Mapping (**CPM**)

Article | Published: 18 March 2019

Cell composition analysis of bulk genomics using single-cell data

Amit Frishberg, Naama Peshes-Yaloz, Ofir Cohn, Diana Rosentul, Yael Steuerman, Liran Valadarsky, Gal Yankovitz, Michal Mandelboim, Fuad A. Iraqi, Ido Amit, Lior Mayo, Eran Bacharach ✉ & Irit Gat-Viks ✉

Nature Methods **16**, 327–332(2019) | Cite this article

Multi-subject Single Cell deconvolution (**MuSiC**)

Article | Open Access | Published: 22 January 2019

Bulk tissue cell type deconvolution with multi-subject single-cell expression reference

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R. Zhang ✉ & Mingyao Li ✉

Nature Communications **10**, Article number: 380 (2019) | Cite this article

Single cell–assisted deconvolutional DNN (**Scaden**)

Deep learning–based cell composition analysis from tissue expression profiles

Kevin Menden[1,*], **Mohamed Marouf**[2], **Sergio Oller**[2], **Anupriya Dalmia**[1], **Daniel Sumner Magruder**[2,3], **Karin Kloiber**[2], **Peter Heutink**[1] and **Stefan Bonn**[1,2,*]

[1]*German Center for Neurodegenerative Diseases, Tuebingen, Germany.*

[2]*Institute of Medical Systems Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.*

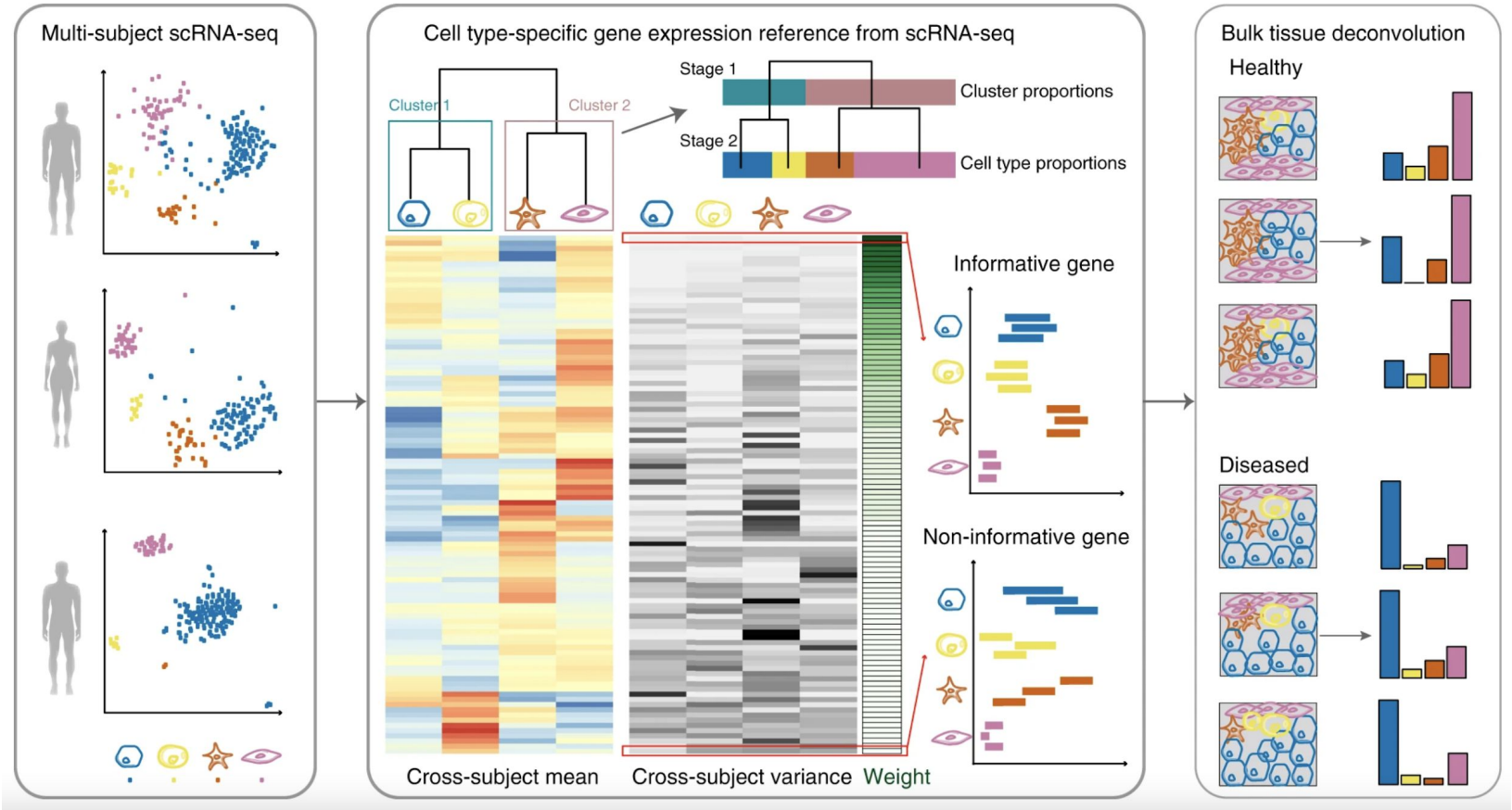
[3]*Genevention GmbH, Goettingen, Germany.*

↵*Corresponding author. Email: sbonn@uke.de (S.B.); kevin.menden@dzne.de (K.M.)

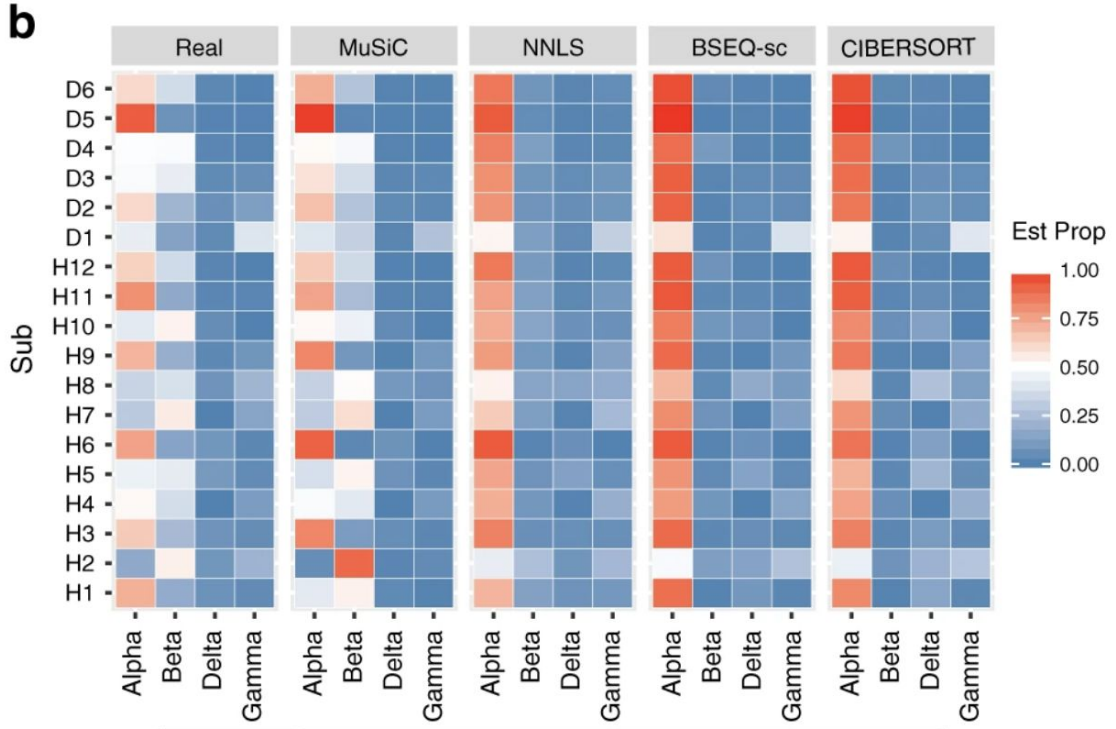
– Hide authors and affiliations

Science Advances 22 Jul 2020:
Vol. 6, no. 30, eaba2619
DOI: 10.1126/sciadv.aba2619

- Do we know all the cell types ?
- Limitations of reference marker genes
- "*You cannot find which you cannot see*"

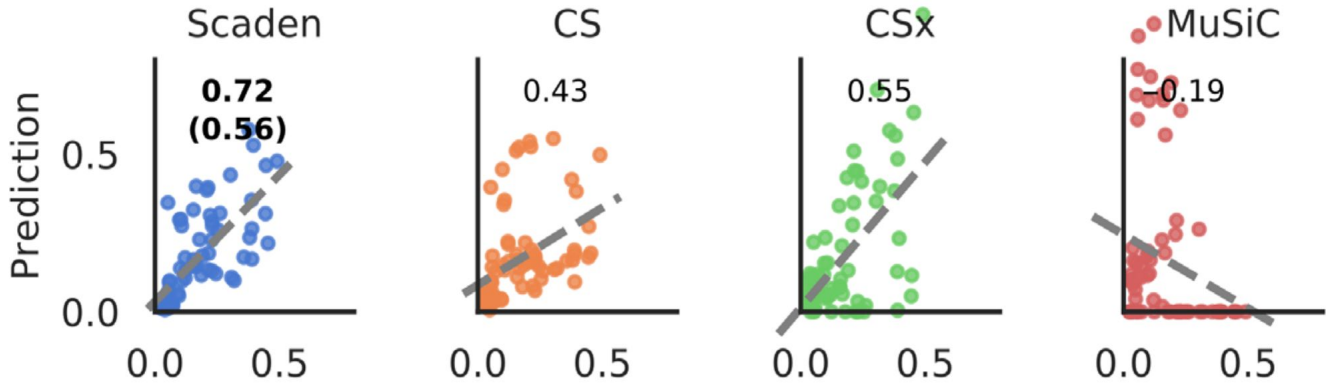


MuSiC



Method	MuSiC	NNLS	BSEQ-sc	CIBERSORT
RMSD	0.10	0.17	0.21	0.21
mAD	0.06	0.12	0.15	0.15
R	0.94	0.82	0.79	0.76

Peripheral Blood mononuclear cells



ScaDen

Brain cells

Yuna Blum, Ligue contre le Cancer

Jérôme Cros, APHP

Clémentine Decamps, Uni Grenoble Alpes

Carl Herrmann, Medical Faculty Heidelberg

Slim Karkar, Uni Grenoble Alpes

Yasmina Kermezli, Uni Grenoble Alpes

Magali Richard, Uni Grenoble Alpes

Ashwini Sharma, University Hospital Heidelberg

https://cancer-heterogeneity.github.io/cometh_training.html

www.eithealth.eu | info@eithealth.eu

EIT Health is supported by the EIT, a body of the European Union